HiPEAC Vision 2021

HIGH PERFORMANCE EMBEDDED ARCHITECTURE AND COMPILATION

Editorial board:

Marc Duranton, Koen De Bosschere,
Bart Coppens, Christian Gamrat,
Thomas Hoberg, Harm Munk,
Catherine Roderick,
Tullio Vardanega,
Olivier Zendra

This document was produced as a deliverable of the H2020 HiPEAC CSA under grant agreement 871174 - HiPEAC
The editorial zboard is indebted to Dr Sandro D'Elia of the Directorate-General for Communication Networks, Content and Technology of the European Commission for his active support to this work.

Design: www.magelaan.be

ISBN 9789078427025

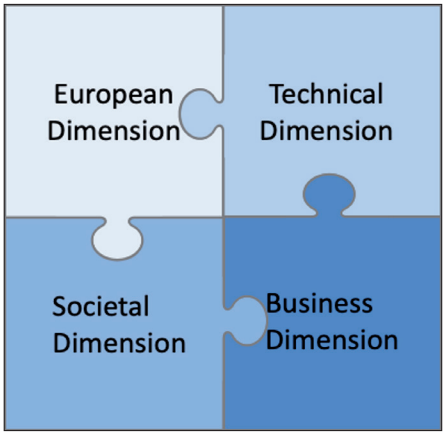
Foreword

Our world is evolving very rapidly, both from the technological point of view – with impressive advances in artificial intelligence and new hardware challenging longstanding PC hardware traditions, for example – and as a result of unexpected events. The year 2020 was quite exceptional, an *annus horribilis*, according to some. It is hard to disagree with this statement, but every dark cloud has a silver lining. 2020 was also the year that accelerated digital transformation beyond what could have been imagined in 2019. Vaccine development happened faster than would ever have been conceivable a year ago, digital payment became the norm for many people and e-commerce and online sales threatened brick and mortar shops. Employees were encouraged to work from home – with its advantages and disadvantages, videoconferencing became the de facto way to interact with both family and colleagues, schools were forced to experiment with distance learning. The list goes on. After living for over a year in an online world, most people will not return completely to the "old normal". They will go for a combination of the "old normal" and things they discovered and experimented with in the circumstances forced upon us by COVID-19; they might keep their home office on some days, and be in the workplace on other days. Higher education will certainly also continue to offer online teaching.

The rapidly evolving digital world has also had an impact on the HiPEAC Vision: updating it every two years no longer seems quite in keeping with the speed of the evolution of computing systems. Therefore, we decided to move from producing a large roadmap document every other year, to an agile, rapidly evolving electronic magazine-like set of articles. The HiPEAC Vision 2021 has two main parts:

- A set of recommendations for the HiPEAC community. It will also introduce the second part of the Vision and will be updated periodically, or on particular occasions.

- A set of "articles", like in magazines, that will be regularly updated, the purpose of which is to support the set of recommendations, or to introduce new topics or situations. This will guarantee that the HiPEAC Vision keeps evolving and remains up to date. These articles are intended to be self-sufficient and can be read independently. They are grouped into four themes or dimensions: technical, business, societal and European. New articles will be added over the course of the years (and outdated ones might be removed). A further element of this new approach to the Vision is that the editorial board asked and will ask various authors to contribute to those articles. This adds heterogeneity and a diversity of point of view that could be helpful for better analysis of the computing systems landscape as well as improve the quality of the recommendations.



The HiPEAC Vision will be available on the HiPEAC website, with easy access to each article. You will also find a "consolidated" version of the Vision, similar to the previous versions, and incorporating all the current articles. This consolidated version will also be printed and distributed like the previous ones.

Contents

Introduction

This is the eighth edition of the HiPEAC Vision. The first was published in 2008. In little over a decade, the performance and range of uses of computing devices have increased dramatically, changing our lives forever. In 2008, the societal impact of smartphones had not yet been felt (the first iPhone was launched in 2007), and ICT was mainly seen by consumers as PCs or games consoles. Social networks had not yet changed the way we interact and share information (Facebook had become accessible to everybody in 2006) and self-driving cars were science fiction dreams. Personal assistants were still in research (Apple's Siri was presented on iPhones in 2011 and Amazon's Alexa in 2014). The rebirth of artificial intelligence, thanks to deep neural networks, can be traced back to 2012 when the work of Hinton's team drastically reduced the error rate in image recognition in the ImageNet classification challenge. Since then, artificial intelligence has invaded every domain with the promise of increased efficiency. It has also increased the scope for fraud and forgery, with "deep fakes" rendering the distinction between real and fake more and more difficult to make.

Each time the editorial board starts work on a new HiPEAC Vision, it thinks there will be a "small" increment over the previous one. Yet when we wrote the HiPEAC Vision 2021, a lot had happened between 2018 and 2020. Artificial intelligence had gained even more efficiency and more consolidations had taken place in the ICT domain: Nvidia buying ARM and AMD buying Xilinx for example. International tensions increased and a trade war between the United States and China made the future of ICT – which is based on worldwide distribution of activities – more difficult to predict. Being self-sufficient or as autonomous as possible in ICT technologies is now a very pressing concern. Last but not least, the SARS-Cov-2 virus – a tiny piece of information eager to replicate – changed the world by locking down a large part of the population, increasing the inequalities between those who can work from the safety of home and those who can't.

But the main recommendations of the 2019 Vision document (and of the previous ones) still hold and some of them need even greater emphasis. These recommendations include improving the energy efficiency of computing systems (in line with the European Green Deal), increasing efforts on the research, design and industrialization of ICT systems - both hardware and software - in Europe (in line with the drive towards European sovereignty), better protection against online crime and attacks (cybersecurity) and mechanisms for increasing trust in computing systems.

As explained in the HiPEAC Vision 2017 (hipeac.net/v17), ICT is expanding beyond cyberspace to interact with us directly, for example in self-driving cars, driverless underground and overground trains, factories and even cities. We are now in the era of cyber-physical systems (CPS), and these systems will be increasingly enhanced with artificial intelligence, so that we could call them **cognitive cyber-physical systems**, or C^2PS. This evolution will further increase the requirement for trust, autonomy, safety and efficiency, and will drive CPS to be pro-active rather than reactive thanks to a better modelling of the environment. We think that CPS should become $(CPS)^2$, i.e. **C**ognitive **C**yber and **P**redictive **P**hysical **S**ystems of **S**ystems – a new "CPS version 2". The "System of System" part is also important: software, applications and infrastructures will increasingly be aggregates of heterogeneous artefacts, including legacy ones, with a variety of deployment requirements. Software will be distributed, becoming a "**continuum of computing**" across platforms and devices, from the "deep edge" to the cloud, HPC and data centres. Programming has to be reinvented for this, with languages and tools to orchestrate collaborative distributed and decentralized components, as well as components augmented with interface contracts covering both functional and non-functional properties. Even if the cloud is mainly driven by non-European organizations, Europe has the potential to be a key actor in the "edge" and "deep edge" solutions if it does not wait too long. As explained in the HiPEAC Vision 2019, the pendulum is swinging more towards having "intelligence at the edge" rather than just in the cloud, but the window of opportunity will be small, and being able to deliver a high level of diversity of systems, customized to particular applications and designed quickly in a highly productive design phase, will be key for success.

The "continuum of computing", an idea pushed by several HiPEAC Vision documents, is now more widely recognized; the new "TransContinuum Initiative" involves a number of European organizations (ETP4HPC, ECSO, BDVA, 5GIA, EU Maths, CLAIRE, AIoTI and HiPEAC) and will promote the ideas further through concrete actions.

We are finally seeing the emergence of what the HiPEAC roadmap proposed in 2009: "**keep it simple for humans, and let the computer do the hard work**" (hipeac.net/v10). While already a commercially available concept for mechanical engineering, asking machines to explore a gigantic design space to find a good solution to a given problem will become a live avenue of research for hardware design ("auto-design?"), for software programming – sometimes referred to as software 2.0 – and even for finding the parameters of the AI techniques themselves (cf. automatic machine learning, or "auto-ML", where reinforcement learning, among other techniques, is used to design deep learning networks). Several groups around the world are now exploring ways to use "artificial" intelligence to write code or to design better hardware.

On the hardware side, silicon technology has not yet run out of steam; even if the performance increase continues in the coming years, it will not be primarily because of "classical" shrink of technology nodes, but because of a combination of different factors (better geometry of transistors ("Gate All Around" FET transistors), contacts on top of active areas, use of more levels and of 3D stacking, better materials, tools, etc). The use of assembly at nanoscale (interposers, chiplets) is more and more widespread, even if a universal standard has not yet been proposed (like "JEDEC" for discrete components), and no other alternative technologies have really emerged, and despite there being a lot of hope for quantum computing to provide a certain kind of acceleration for data centre class applications.

Climate change is accelerating rapidly and its effects are being felt across the globe. Creating sustainable ICT and lowering its footprint is an ever more urgent necessity. It is expected that the increase in energy used by ICT will be offset by the energy efficiency gains it produces elsewhere and its impact on other domains. Yet ensuring that **computing devices require less energy** to manufacture and to operate is still a major avenue for further development and still needs to be actively pursued, as explained in the previous Vision documents.

The list of topics that we need to consider is growing: greater security for our data (privacy), interoperability, the widespread "as a Service" trend, code migration, containerization, use of legacy systems, edge to cloud computing ("continuum of computing"), distribution of computing, secure and trustable platforms, closer intertwining with the physical world (CPS)[2], support for people with limited digital skills, ease of use, and many more. They drive us to a more global "vision" of what the ICT infrastructure of tomorrow could be, some kind of next generation web; we propose to instantiate this vision through a multidisciplinary **"moon-shot" programme** that we call "**Guardian Angels**". This program will be explained in detail in a specific article.

As you can see, there are new topics to explore or to develop further since the publication of the last Vision. In order to help the reader, instead of having big monolithic chapters, we propose to split the "rationale" part of the document into a coherent set of independent articles, illustrating or explaining different aspects of the recommendations we propose. We hope it will help and that you will take as much pleasure in reading this new release as we have done in producing it.

Figure 1: The history of the HiPEAC Vision documents.

Key recommendations of the HiPEAC Vision 2021

The 18 key recommendations of the HiPEAC Vision 2021 are grouped in three clusters:

- **Technical recommendations** which are useful for the HiPEAC community members as a source of inspiration and to guide their strategic research directions. They may also offer a steer of use to regional and European funding agencies as they plan calls for research proposals.
- The second cluster of recommendations are the **global policy recommendations**. These are targeted at a more global level and for example can be taken into consideration by the European Commission to steer actions or instruments. They propose a number of high level non-technical actions that will strengthen the European computing ecosystem.
- Finally there are **societal recommendations**, which are even more generic and can be addressed at the level of regional and European Parliaments, for example. These recommendations relate to the direction in which we want to evolve globally, as a society.

1. TECHNICAL RECOMMENDATIONS

We have come a long way from the time when a computing system consisted of one computer core programmed in one or very few programming languages. The need for more computing power on energy-constrained computing platforms (from deep edge sensor node to supercomputer) first forced computer vendors to introduce multicores. More recently it has obliged them to leave behind homogeneous multicores in favour of heterogeneous multicores consisting of different kinds of accelerators that are more efficient, but more difficult to program and use efficiently.

The emergence of new workloads such as deep learning and large-scale industrial cyber-physical systems has led to a series of new challenges that are related to non-functional properties power consumption, timing, complexity, security, safety and sustainability. These are not new challenges, but the scale at which they need to be tackled now calls for more effective solutions. Saving 100 mW of power each for ten billion IoT devices saves 1GW. Securing a network of 100,000 distributed computing nodes is totally different to securing a server in a data centre. Proving safety properties for autonomous vehicle software that consist of millions of lines of code written in a dozen programming languages is a challenge and can no longer be done manually.

The design and implementation of modern computing systems has become so complex that it exceeds the cognitive capacity of even the best computer scientists. Either the development phase will take too long to bring the system to the market, or the resulting system will contain too many errors, some of which might create safety hazards. The current approach to managing complexity by adding layers of abstraction has reached its limit due to the inefficiency introduced by each additional layer and the lack of global optimization. There is little hope that systems and the associated applications will become less complex in the future: they won't. Every complexity reduction earned by local optimizations will be seized upon to build even more complex systems. Hence, there is only one way forward: we have to find practical and efficient solutions to deal with the increasing complexity.

The technical recommendations can be condensed into the following: a move towards 5S.$(CPS)^2$.

$(CPS)^2$ stands for the new generation distributed systems that interact with the physical world: they are **C**ognitive **C**yber and **P**redictive **P**hysical **S**ystem of **S**ystems: CCPPSS = $(CPS)^2$, pronounced "CPS square".

5S stands for the key non-functional requirements: **S**ober, **S**ecure, **S**afe, **S**traightforward and **S**ustainable.

For Europe to lead in 5S.$(CPS)^2$, the following research challenges should be tackled.

Recommendation 1: Cognitive

Artificial intelligence is the number one disruptive technology of the moment. The least the customer expects when acquiring a new device is that it is smart, i.e. that it is able to make sense of the environment it is operating in (continuously collecting information and processing it), and to take actions that are perceived as intelligent and helpful by humans. Although as old as computing itself, artificial intelligence has become mainstream in the last decade thanks to breakthroughs in the domain of machine learning (made possible by the combination of algorithmic innovations, big data and efficient hardware acceleration), and fuelled by the deep pockets of big tech companies. Artificial intelligence is reaching a productivity plateau in some application domains but has hardly scratched the surface in others. It is still facing serious technical challenges such as the need for trustable AI, the computing cost of the training phase, low power accelerators for edge devices and efficient integration in traditional computing substrates. HiPEAC

recommends investment in ultra-low power accelerators for AI and in investigating approaches that use less labelled data.

Recommendation 2: Cyber

The cyber space is growing enormously. 90% of all the data in the world was generated in the last two years, and almost half of the world population uses the internet. This was made possible by the enormous increase in performance of computing systems including in storage and communication (x100 000 000 in 33 years), but is also thanks to protocols, interfaces that allow interoperability and scalability of heterogeneous systems. Due to its intertwining with the physical world, the "next web" will have to cope with new constraints (non-functional properties for example) and more machine-to-machine communications while remaining compliant with legacy. Just as Europe set the basis for the world wide web, HiPEAC recommends that it should secure its place at the forefront of the "next web" by adding the necessary innovations and standards on top of existing technologies to meet and satisfy human needs and interests.

Recommendation 3: Predictive

Modern CPS systems should be predictive. This means they should be able to anticipate the actions or reactions of the real world and not just be reactive to it. This requires them to have access to an accurate model of the environment they are operating in. This is amongst other considerations vital for determining the outcome of particular action. For example, the control system of a self-driving car should have a good model of the dynamic behaviour of a car (e.g. how much force to apply to the brakes in order to stop the car at a red light, taking into account the weather conditions and the assumed behaviour of other road users). Obviously, there is no time for the digital twin to carry out detailed physical simulations while the car is in motion, or to consult the cloud. Instead, there is a need for simple yet accurate models that can be executed in real-time at the edge. Artificial intelligence and surrogate models might come to the rescue. HiPEAC recommends investment in digital twins and models that can be executed accurately and efficiently at the edge.

Recommendation 4: Physical

Interacting with the physical world requires not only that computing systems understand the environment, but also that they can react appropriately. This means that a CPS needs to be more than just functionally correct. It must also respect the non-functional requirements imposed by the physical world: timing, energy consumption, reliability, resilience, size and form factor amongst others. Common programming languages (and hardware) have little or no support for expressing non-functional properties but instead focus on best effort functional properties, i.e. calculate functions as fast as possible. HiPEAC recommends investment in research into ways to correctly model non-functional properties and to guarantee them in the systems.

Recommendation 5: System of systems

Modern CPS systems are a continuum ranging from deep edge (microcontrollers linked to sensors or actuators), to edge, concentrators, micro-servers, servers and cloud or HPC. Every system itself is now a component of a larger system, and an increasing number of systems are very dynamic: they not only change behaviour over time (via updates), but also move around (like vehicles in a transportation system, or smartphones). More and more large CPS span entire countries (e.g. a railway system) or even a continent. They are among the most complex systems ever built. Managing such a dynamic system requires devices that are intelligent and able to negotiate with their peers, exchange capabilities and interface formats in order to ensure the quality of service (QoS) in various configurations and situations. "Self-X" might allow the reconfiguration of systems in a broader system and ensure a minimum mode of operation even in degraded situations. At a higher level, there is a need for a distributed orchestrator to steer the individual components, and then implement the application itself. The orchestrator should have the freedom to distribute the required functionality over the complete computing continuum (from HPC to deep edge) as e.g. containerized microservices. HiPEAC recommends investment in systems of systems research and development on tools for orchestrating large dynamic heterogeneous systems.

Recommendation 6: Sober

Ultra-low power computing remains the holy grail of computing because power consumption is, in practice, the hard limit on performance. It is needed to extend the battery life of mobile and IoT systems, and it is a key performance metric for affordable cloud computing and supercomputing (cost of ownership). Exponential growth of the internet can only be sustainable if it is matched with a similar increase in power efficiency of devices and communication and data centre infrastructure. For example, three application domains that are currently challenged by power constraints

are exascale computing, the training of advanced deep learning models, and bitcoin mining (or other applications of distributed ledgers). Another is the battery powered devices for which it is difficult to replace batteries (like implanted devices, or difficult-to-reach sensor nodes). Although many electronic devices consume more resources in their production phase than in operation, the environmental impact of their power consumption is non-negligible. Saving 100 mW of stand-by power for ten billion electronic devices means a power reduction of 1GW, or the equivalent of four million tons of CO_2 per year. This is a consequence of the sheer size of the installed base. HiPEAC recommends investment in the development of ultra-low power computing platforms covering the complete digital continuum, and in tools allowing assessment and design of systems with explicit power constraints.

Recommendation 7: Secure

Security is used to protect computing systems against attempts by unauthorized parties (i) to steal information from systems; or (ii) to disrupt the correct operation of a system. The smaller the attack surface of a system, the more secure it can be made. The attack surface of ICT systems grows with the size of the systems, and with the number of entry points to the system. Security is an arms race between attackers and defenders. This arms race is not only technological, but also social (weak passwords, stolen devices, …). The combination of the global digital transformation and the waging of cyber warfare by several countries with the help of companies leads to increased security risks. Security is the mother of all functional and non-functional properties of software. If a hacker gets access to a system, none of these properties can be guaranteed anymore. AI-based systems have created a new class of security risks in which hackers can mislead an artificial intelligence by offering it carefully crafted input. HiPEAC recommends greater investment in cybersecurity research, and in particular in the automated finding of security risks in existing applications, in the means to automatically mitigate or remove those risks, and in the development of secure hardware and tools that can produce secure by design software and hardware.

Recommendation 8: Safe

CPS systems that actively interact with the real world, should not cause damage to physical goods, or injure living beings. Safety is the way to protect the outside world against unwanted actions by computing systems. These actions could be intentionally caused by malware or unintentionally by defects, bugs or bad specifications. Safety hence requires (i) correct software (proven by extensive testing or by a formal proof); (ii) resilience to avoid failures due to defects (e.g. by adding redundancy); and (iii) secure systems. No system should ever be declared safe if it cannot be proven to be secure. We should aim for systems that are safe by construction, or that can be proven to be safe, or which use safeguards to ensure that safety is always observed even in unforeseen conditions. HiPEAC recommends further investment in research and development in the methodology and design of safety-critical systems.

Recommendation 9: Straightforward

Future 5S.(CPS)^2 will in practice be very complex, and they should be manageable at human scale. Therefore, effective tools should be developed to help design, implement and maintain such systems. Potential solutions are methods and tools to support modularity, components, containers, contracts, specifications, services, orchestration, or formal methods for modelling IT systems and their functional and non-functional properties. At the engineering level, advanced automation tools are needed to produce software and hardware. A promising approach is to use artificial intelligence to automate particular tasks like hardening code against malicious attacks, formal or automatic validation and verification of contracts, design exploration for hardware design, amongst other applications. HiPEAC recommends the development of approaches that improve human productivity to design, produce and manage complex systems, including with the use of AI techniques.

Recommendation 10: Sustainable

State-of-the-art electronic devices require the use of around 65 of the 102 elements of periodic table of the elements, some of which are scarce or come from politically unstable areas of the world; the supply of these minerals is not guaranteed forever. At the moment, only 15% of the world's computing devices are currently recycled, and state-of-the-art recycling technology can only extract 17 of the 65 elements. To decrease the negative environmental and societal impacts caused by electronic device production, and shipping electronic waste to third world countries for disassembly and recycling, burning it or dumping it in landfills, Europe should focus on urban mining, and try to recycle as much raw materials as is technologically possible.

Another aspect of sustainability is the energy needed to produce and use a device. Common computing devices like smartphones

and tablets require more energy to manufacture than they use during their full lifetime (around 70%/30%). By making the devices more energy efficient, this ratio will further increase. HiPEAC recommends that Europe funds research to lower embodied energy of devices, and encourage the extension of the lifetime of devices by upgrading, reusing and repairing them. Europe should have the ambition to lead in the design of sustainable electronics.

Finally, 5S.(CPS)² can contribute a lot to the United Nations Sustainability Goals. Supporting ten billion people while protecting the planet will be impossible without the radical use of advanced computing solutions to optimize resource consumption.

2. GLOBAL POLICY RECOMMENDATIONS

These recommendations are more transversal and 'high level' and need to be implemented across several programmes or projects.

Recommendation 11: Open source

Although software sharing is much older, the free software movement started in 1983 with the GNU project. The term open source was adopted in 1998. Around the same time, the concept of open hardware was launched. Open source hardware became well-known with the creation of the RISC-V architecture in 2010. HiPEAC recommends investment in free open source digital commodities in the form of a EU platform that guarantees privacy compliance, inspection and audit, verification of compliance, security, sustainability, ... This platform can then be used by third parties to create value. This requires the establishment of European open source institution to support open source (software and hardware), to clarify legal aspects, and to promote collaboration. HiPEAC also recommends that the critical parts of the cybersecurity of IT systems are based either on open source software and hardware, or on EU-made, trustable because audited, proprietary hardware or software.

Recommendation 12: Moonshot programme "Guardian Angels"

The goal of a moonshot programme is to synergise technologies across disciplines. HiPEAC recommends the creation of a "Guardian Angels" moonshot programme that encompasses all the 5S.(CPS)2 technologies in a system that will serve European citizens and companies. A "next web" that intertwines the cyber and physical worlds for industrial and personal use is to be developed, overcoming the fragmentation of vertically-oriented closed systems, heterogeneity and the lack of interoperability. It should demonstrate self-configuration and self-management in a dynamic plug-and-play environment, while also coping with security and privacy of personal and corporate data and offering natural interfaces for their users. The core consists of advanced orchestrators, which are called "Guardian Angels", loyal to their users, placed at the interface of the physical and virtual worlds, to orchestrate in a safe and secure way the various services provided by the "next web".

Recommendation 13: International competence centre

Europe has many national competence centres with a leading international reputation in computing (imec, CEA, CWI, Fraunhofer, …) but has few international competence centres such as CERN for physics or the ESA for space research. HiPEAC recommends the creation of such a well-funded European competence centre in computing so that Europe is able to retain and attract top talent, to set its own ambitious research agenda, to attract large investments, and to form the core of a network of regional competence centres. Such a network will be crucial for defending Europe's position as a scientific powerhouse. It will also the entry point of an innovation pipeline.

Recommendation 14: Digital infrastructure

The success of the digital transformation of Europe depends on the quality of the digital infrastructure (networks, data centres, security tools and services, and so on). COVID-19 has made clear that the current infrastructure is not yet able to support generalized remote activities such as teaching, working, shopping, and streaming. HiPEAC recommends that Europe invests in a state-of-the-art digital infrastructure. The fast roll-out of 5G is crucial for supporting the next generation of productivity enhancing and resource saving applications (smart cities, smart transportation, industry 4.0, …).

Recommendation 15: New computing technologies

Even if the classical silicon technology still delivers performance improvements, it is increasingly difficult and costly to achieve these improvements. HiPEAC recommends that Europe continues to investigate emerging technologies, not with a view to them directly replacing silicon technology, but to complementing it. This research should be wide-ranging, and include new ways to code information (e.g. using "qbits", temporal coding like with "spiking" neuromorphic architectures, or using physics phenomenon – like light – as analog computing approaches), as well as methods to

efficiently integrate these approaches as "accelerators" in a silicon technology-based system, on both the hardware and software sides.

In the field of quantum computing, HiPEAC recommends that Europe supports R&D in the field of architecture and software stack for quantum computing, develops the integration of quantum accelerators into future exascale infrastructure, and promotes the emergence of a European quantum cloud.

In addition to supporting research at the interface of computing paradigms and technologies, and since most novel computing technologies rely on an efficient integration with classical computing platforms, *maintaining capabilities* in CMOS circuits design and fabrication is a requirement. Therefore HiPEAC recommends that Europe promotes an active ecosystem based on 3D technologies such as monolithic 3D, 2.5D (use of interposers and chiplets) to maintain the capability of modularity and independence for complex designs made by assembly of standardized chiplets. HiPEAC also recommends that Europe retains its knowledge base in advanced CMOS technology (Gate all Around transistors – GaaFET) to enable understanding and efficient use of these devices in systems.

3. SOCIETAL RECOMMENDATIONS

Digital technologies will continue to transform society. An increasing number of citizens and scientists are worried that this transformation will be so profound that it might disrupt society itself. Major areas of concern include the impact of computer-based automation on employment, the use of artificial intelligence for automatic decision making, the impact of social media on public opinion, and the impact of computing on inequality and on sustainability. In order to mitigate and manage the effects of this transformation, Europe should continuously train its workforce in digital skills and awareness so that nobody is left behind. It should stimulate an innovation culture to find solutions for the grand societal challenges, and work on a European ethics framework to protect its values.

Recommendation 16: Training

As the digital transformation progresses, the economy and society depend on technology – and on the people who develop and maintain it – more than ever before. With the baby boomers retiring, and a shrinking active population in large parts of Europe, the speed of digitalization will be limited by the size of the workforce that can be mobilized in the process. HiPEAC recommends that Europe invest in education and training in general, so as to stay competitive, and produce more highly-skilled computer scientists to advance the state of the art in $5S.(CPS)^2$ in all its aspects: software engineering, digital twins, artificial intelligence, security, safety, … Given the speed with which technology evolves, Europe should also invest in lifelong learning to retrain the existing workforce in new technologies, and with new skills.

Recommendation 17: Innovation culture

Thanks to its excellent research infrastructure, Europe produces 20% of the top scientific publications, and these publications attract 22% of global citations. This is roughly similar to the output of the United States. For a similar research output, the United States attracts six times more venture capital to commercialize the results. Europe is clearly underperforming in innovation. HiPEAC recommends that Europe invests more in the creation of an innovation culture at all levels (education, society, industry) to stay competitive in this fast-evolving world and to help attract venture capital for scale-up companies.

Areas with growth potential in Europe are: industry 4.0 and the automation of manufacturing; smart mobility; health; the silver economy; entertainment and gaming; technologies for the Green Deal and cybersecurity technologies.

Recommendation 18: European values and digital ethics

The day computing ceased to be about technology but about content and data, saw it enter a difficult relationship with social and ethical values such as privacy, ownership of data, constitutional freedoms, truth, responsibility, liability, sovereignty, sustainability, and so on. Europe should make sure that its physical world value system is also implemented in the cyber world and that every citizen is as well protected in the cyber world as in the physical world. This will require extra legislation to be implemented and enforced.

Intelligent computing systems might lead to unwanted bias and discrimination if not designed carefully. HiPEAC recommends that digital ethics be developed as a separate discipline and become a standard element in computing curricula. This should lead towards responsible AI, which means that the public and the private sector agree that they should only use AI for the betterment of society. All technology companies should comply with the European rules and ethical frameworks if they want to do business in Europe.

HiPEAC Vision recommendations in depth

Our world is becoming dependent on cyber-physical systems (CPS) providing unprecedented innovation opportunities but also representing unprecedented complexity. We need to design future CPS to make sure that they become human-centred and part of a circular economy.

Cyber-physical systems have far-reaching implications

By MARTIN TÖRNGREN

Cyber-physical systems (CPS) integrate computation, communication and physical processes to form small- as well as large-scale systems with improved or new capabilities. The term cyber lends itself to some confusion since it can be seen as either stemming from "kubernetes" – referring to feedback systems – or relating to the use of computers or computer networks. Both interpretations generally make sense for CPS, of which there are many types!

CPS have been available since at least the 1970s with the advent of direct digital control and are thus not new. Integrating and leveraging advances in a combination of technologies (e.g. materials, batteries, sensors, machine learning, processing and communications) is, however, providing entirely new capabilities, enabling the creation of new types of systems. These trends are well illustrated by automated vehicles as CPS. However, the trends and the example only touch upon part of the societal impact that future CPS are likely to have. CPS are applicable in – and across – all domains. They are, moreover, increasingly collaborative, connected over their lifecycles and deployed in open society scale settings ("robots at home and on the streets").

With our societies increasingly dependent on well-functioning CPS (water, energy, transport, health, manufacturing, etc.), it becomes crucial that we develop such CPS to be human-centred; that is, CPS that are explicitly supporting interactions and collaborations with humans, and help us in addressing sustainability challenges. This requires us to be able to balance and manage the complexity of future CPS. This article elaborates on these perspectives.

Key insights

- CPS provide unprecedented innovation opportunities through the synergetic integration of cyber components (software, data-driven algorithms, computation and communications) with physical components, equipping CPS with a unique and evolving set of capabilities.
- HHumans constitute a key aspect in CPS design, with manifold interactions between humans and CPS not receiving sufficient attention. These interactions, as well as the capabilities and limitations of both CPS and humans, need to be considered in conjunction in order to improve usability and manage risks; the automation paradox is more relevant than ever.
- Future CPS will represent unprecedented complexity, which has to be balanced and made manageable. This calls for new scientific theories, lifecycle engineering methodologies, curriculum renewal and lifelong learning. Collaborating CPS form cyber-physical systems of systems (CPSoS) where interactions, operational management and emergence require specific attention.
- The increasing penetration of CPS represents an ongoing socio-technical shift that requires industrial and societal transformation to adapt organizations, processes, standards, legislation and more, highlighting the need for testbeds and involvement of a multitude of stakeholders.

Key recommendations

- Promote organizations and schemes, such as competence networks, to raise awareness and drive the collaboration and change required to obtain circular and human-centred CPSoS
- Undertake R&D efforts into methodologies for:
  - Human-centred CPS, taking into account CPS complexity and the importance and potential of designing CPS and their supporting development environments to appropriately interact with and support humans.
  - Circular CPS and CPS to support the transition to a circular economy, emphasizing needs for cross-fertilization between CPS and sustainability collaborative research.

What do we mean by CPS?

The concept of cyber-physical systems (CPS), as coined in the US in 2006, has had significant impact. It led among other things to the introduction of a new foundational and multidisciplinary research programme by the US National Science Foundation and to the launching of Industry 4.0 in Germany in 2010. In Europe there has also been growing interest and awareness including through a number of roadmapping efforts [1]. At the same time, it is clear that some confusion remains as to the meaning of CPS and that several domains are often using other terms to represent similar types of systems. To illustrate the slight confusion of terminology, consider the following question posed at the HiPEAC ACACES summer school in July 2020:

Which interpretation of a cyber-physical system (CPS) – what it is or represents, makes sense to you?

1. CPS is really the same thing as embedded systems
2. CPS is the same as Industry 4.0 (I4.0)
3. CPS represent feedback systems
4. CPS represents integrations of computation, networking and physical processes
5. CPS is an umbrella term covering e.g. I4.0, IoT and Industrial IoT
6. CPS is a clever term invented to gain more research funding

The summer school participants provided a plethora of answers, covering all possibilities. This response was not unexpected, given the potential dual interpretation of the term "cyber":

- *cyber* – the use of computers or computer networks – with many connotations such as cyberspace and cybersecurity [16].
- *cyber* – as part of the concept of cybernetics, coined by Norbert Wiener in the 1940s, from the Greek word "kubernetes" – "governance", referring to feedback systems [2].

The initial definition of CPS, as the "integration of computation, networking and physical processes where CPS range from miniscule (e.g. pacemakers) to large-scale (e.g. national power grids)", has the characteristic of being very general. In fact, as suggested by Alberto Sangiovanni-Vicentelli during the EU-funded CyPhERS project: which system will not be of a CPS nature in the future? In CyPhERS [17], we set out to provide an agenda for CPS research. One conclusion was that we needed a taxonomy, a way to characterize CPS to describe different types of CPS [3]. In the late 1970s, Enslow wanted to clarify the concept of "distributed processing", and introduced dimensions for characterizing them [18]. He had a strong opinion on what constituted a distributed system along these dimensions. In the CyPhERS approach, however, we decided to recognize that there are many different types of CPS that are distinguishable through a number of characteristics, such as relating to the technological emphasis at hand, the scale and the level of automation. Agreeing on terminology is not always easy. It is interesting to note that even some of the most experienced *systems engineers* had problems on agreeing on a definition of "systems". The Greek word "systema" refers to "a set of things working together as parts of a mechanism or an interconnecting network; a complex whole", or "a set of principles or procedures according to which something is done" [19]. However, they were able to reach a consensus on what characterizes systems [20]. This speaks further in favour of characterizations as a tool to improve communication.

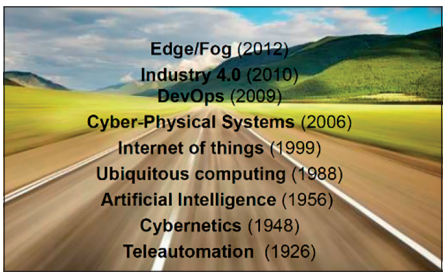


Figure 1. Terms mirroring an expanding technological shift (with an emphasis on the cyber side)

A further source of confusion may stem from the plethora of terms as illustrated in Figure 1.

In the author's view, these terms reflect different perspectives to similar technological trends. For example, Industry 4.0 represents a manufacturing (and thus domain-specific) specific interpretation of CPS, while IoT emphasizes sensing, identifiers and internet connectivity. In contrast, CPS emphasizes the consideration of both the cyber and physical aspects, and the providing of integrated systems.

This line of thought is shared by Lee and Seshia [3], as follows: "In our view, the term CPS is more foundational and durable than all of these (author's note, referring to terms like those in Fig. 1), because it does not directly reference either implementation approaches (e.g. the "Internet" in IoT) nor particular applications (e.g., "Industry" in Industry 4.0). It focuses instead on the fundamental intellectual problem of conjoining the engineering traditions of the cyber and the physical world".

In this article, we emphasize the etymological meaning of CPS and the resulting characteristics and capabilities of CPS, as a basis for our discussion on implications.

CPS capabilities through synergetic combinations of physical, software and data-driven parts

Törngren and Sellgren provide a characterization of CPS by assessing contrasting differences between physical and software parts/components, in order to discuss characteristics of the key components that literally form part of a CPS [5]. This view is elaborated and extended in Table 1, in particular by adding "data-driven" components. The cyber space, as represented in Table 1 by the "software" and "data-driven" columns, provides unprecedented communication, synchronization, and collaboration opportunities – which may indeed span optimization and orchestration across entire value chains, while encompassing development–operation cycles (data collection, upgrades etc.). Data-driven components incorporate machine learning to provide a new way of programming through learning – with strong dependence on (lots of) quality data. Such components enable unprecedented prediction and correspondingly enable proactive behaviours at different time constants.

	Physical components	Software components	Data-driven components
Phenomena and complexity	Multiple coupled physical phenomena, (wear, fatigue, heat, emissions, ..) "slow cycles" and global effects	Pure behavior! State space; bugs; connectivity; variability; Fast local & global effects; "fast cycles", hidden dependencies	Super-human performance in certain areas; Black-box (deep-learning)
Abstractions, synthesis, and platforms	Approximations, Continuous time & value; No single "platform", Behavioral model simulation vs. Geometry based synthesis	Digital / discretization "platform" foundations; Logic preserving transformations; Abstracted physical properties	New programming model; model- or learning based. Software under the hood.
Extra-functional properties	Generally established cost and reliability models	Difficult to estimate life-cycle cost; difficult to reason quantitatively about reliability for critical systems	Data quality dependence; Difficult to assess robustness and reliability, Opaqueness; Brittleness

Table 1: Characterization of key CPS "components", extended from [5]

Further characteristics in Table 1 include "complexity", highlighting that this in a physical system relates to coupled physical phenomena (with similar orders of magnitude and slow global effects), whereas complexity in the cyber world largely stems from the large state spaces and connectivity, which may cause very rapid global effects. The platforms for these types of components differ. The abstractions made possible in software systems have paved the way for their enormous success – enabling high expressiveness that can be compiled into digital machines. This has however been accomplished by abstracting away physical properties – leading to problems in dealing with, for example, real-time and energy [22].

Combining these components/technologies, to form new integrated CPS components and systems, provides CPS with unique capabilities, to:

- Carry out real-world sensing as well as gathering, storing and processing data and models;
- Provide awareness and prediction through algorithms operating on data and models, concerning both the CPS itself as well as its environment, while taking uncertainty into account;
- Plan and make decisions based on an established awareness and predictions, and given system goals;
- Provide physical structures, acting on these structures (e.g. through force or torque), and storing, generating and controlling energy;
- Generate physical artefacts (e.g. through additive manufacturing) as well as software/data entities;
- Coordinate the operation of multiple CPS, and collaborate with other CPS and humans!

These capabilities can be applied at different levels (from a component, over a system, to collaborating systems). Consider, for example, a drone used to collect data (in some setting). The drone itself represents a CPS, which could then be used in conjunction with other CPS (say other drones and robots, 3D printers and computing facilities) to provide collaborative capabilities. The described capabilities can also be applied at different time horizons (from real-time to long-term) and at different geographical spans. The combination of cyber and physical is underlined by the concept of digital twins, as a model of a physical item. Connecting physical parts with their corresponding cybermodels provides novel opportunities, such as for detecting failures (needing handling in real-time), predicting maintenance needs and planning for system upgrades.

Many of these proposed capabilities will relate to general definitions of artificial intelligence (AI) in terms of abilities to "think" or "act" like a human being – thus encompassing both cyber (e.g. abstract reasoning and problem solving) and physical (e.g. physical object manipulation) skills [23]. In this context, the capabilities we proposed here are more comprehensive on the physical side, as exemplified with the

Credit: ID 104631152 ©Sompong Rattanakunchon | Dreamstime.com

inclusion of generation of physical entities. Similarly, by also emphasizing the physical capabilities, these proposed capabilities augment those identified by autonomic computing systems as shown, for example, by the Monitor-Analyze-Plan-Execute (so called MAPE-K loop) pattern [24].

It is interesting to reflect on these capabilities and use them to consider the implications of CPS (in the coming sections). It is clear that the envelope and specific performance–cost characteristics of CPS are steadily evolving. Several technological trends are driving improvements in CPS capabilities, for example, through the development of new cost-efficient sensors, not the least driven by the technology push in automated driving and developments in artificial intelligence, and with computational and communication improvements, in particular through the envisioned new tier of edge computing. It is also clear that innovations lie ahead of us that will certainly add entirely new capabilities! An interesting line of investigation would be to approach the capabilities from the perspective of limitations (e.g. in terms of latency, computations, human machine communication) in order to explore what could be achieved if (or perhaps rather when) these limitations are overcome.

Collaborating CPS (CPSoS) and the missing link – humans!

Communicating and collaborating CPS will increasingly be used to form cyber-physical systems of systems (CPSoS). Collaborating CPS provide special challenges when they lack a single integrator; as a result, multiple organizations and stakeholders need to coordinate and manage the so formed systems of systems (SoS). SoS are often described and characterized in terms of operational and management independence, and by emergent behaviour [14]. A typical example is a traffic system in a city, a situation with a multitude of stakeholders, independent evolution (of streets, vehicles, other infrastructure) and responsibilities that are not always clear. A change in one of the constituent systems may cause unexpected behaviours (emergence). A typical example of the latter includes the introduction of automated vehicles and the (potential unexpected) effects it may have on the larger traffic system in terms of traffic flows, amount of traffic, accidents etc. The cyber space in terms of connectivity, computing and artificial intelligence is providing new innovation and business opportunities in SoS by enabling data sharing, and by providing new means for coordination and prediction. This in essence means that many new SoS will indeed be CPSoS.

The example CPSoS highlights a number of further challenges and opportunities. For example, what are suitable "rules of the game" for the interactions between a number of CPS (compare autonomous vehicles and in mixed traffic, with vehicles at different levels of automation and humans on the streets)? How can data reasonably be shared and managed and who would be liable in case of an accident? The CPSoS setting further highlights depend-

encies between various CPS including the infrastructure. It is clear that frameworks will be needed in terms of defining goals, policies and mechanisms for interactions among constituent CPS. From the above example and discussion, it should be clear that CPSoS need to be considered as socio-technical systems as their introduction will have large repercussions on a number of societal aspects and stakeholders.

A further very important realization concerns the "missing link" – humans! CPS will not exist in isolation of humans, and will always (as far as we can project) be interrelated with humans who have different roles, including developers, part suppliers, manufacturers, users, operators, and those that maintain CPS. It is apparent that the gap between technological expertise and the understanding of human needs and traits is difficult to bridge, and that the interactions between humans and CPS deserve much more attention. It is of paramount importance that future CPS are developed to be human-centric. This topic will be further elaborated in the section on "Implications and impact of CPS". Another related concern (also to be followed up in the same section) is that of the increasing complexity of CPS, which requires large teams of experts and companies to collaborate in their development. The net effect of growing CPS complexity is that more entangled aspects need consideration, requiring more knowledge and interactions. Communication between people thus also becomes more and more important along with an understanding of how our brains work, and how we as humans can collaborate with AI systems [5,27].

CPS opportunities and design space, contributing to a circular economy

The capabilities of CPS lead to a very large potential for innovation and applicability within existing and new domains, including society, farming, healthcare, gaming, industrial production and infrastructure (e.g. water, electricity and transportation). There is of course a difference in how far domains have come. Many domains already have a substantial adoption of CPS, entailing legacy but also efforts to increase connectivity, lifecycle management, automation and smartness. Other domains may have more of a greenfield approach, a situation that sometimes could be beneficial (less legacy). See more on the concept of circular economy in the article "Towards circular ICT: from materials to components".

New CPS applications are driven firmly by the automotive industry (automated vehicles), by telecom (5G and beyond) and by the "internet" giants [21]. Many existing publications describe the strong innovation potential of CPS to create new businesses and for improving many aspects of our lives [1]. The innovation opportunities are representative of a socio-technical shift, with new companies emerging, and some disappearing [11]. The corresponding opportunities provide additional needs and drive innovation in several directions, for example in 3D printing, perception and in computing, providing likely accumulating effects with both technology push and pull.

To illustrate the design space and innovation potential with CPS, consider the Hy-Wire Skateboard concept introduced as early as 2002 by GM (see Figure 2) [25]. With a CPS platform, featuring fuel cells, electrical motors per wheel and a distributed computing platform, this concept demonstrated the possibilities for a substantial vehicle redesign when replacing the mechanical connection (transferring force and position) with a distributed control system ("steer-by-wire"). This also illustrates the concept of "software-defined", where the properties of the former mechanical steering column are now essentially programmable, limited only by the communication bandwidth, real-time computation, and of the physical constraints imposed by the electrical actuators and power provided. The example moreover highlights some of the inherent challenges and inertia in deploying new CPS; the Hy-wire concepts are not yet on the roads, partly because regulation still does not allow steer by wire, and partly because of the new infrastructure that would be needed for hydrogen-based fuel cells. This dilemma illustrates that innovation opportunities, and barriers for innovation, can be identified at different levels [31].

The CPS capabilities provide new tools that can – and should – be leveraged in order to reduce overall CO_2 emissions and support a circular economy. Taking Intelligent Transportation Systems (ITS) as a CPSoS example, an ITS can be developed to provide proactive behaviours including the total avoidance of formation of cracks

Figure 2. CPS opportunities illustrated through the Hy-Wire Skateboard concept car (Source: GM, 2002)

and pot-holes in roads, thereby optimizing and reducing efforts, emissions and energy for road repairs, and moreover improving traffic safety. Of course, the potential does not stop here, and extends to energy optimization and traffic efficiency. The example is interesting because it illustrates the opportunity to share data between two separate domains, in this case road infrastructure and road traffic. Combining the data (for example that gathered during road construction, by automated vehicles and infrastructure) with vehicle/infrastructure computational capabilities for learning, prediction, planning and coordination, enables new functionalities and services in the ITS.

In going from the existing "*take-make-dispose*" industrial model, a circular economy represents a paradigm shift in the way of thinking of waste and the lifecycle of materials, components and systems. Perhaps an appropriate way to think about circularity is in terms of "value flows" – where for example "waste" material from production, and components no longer providing adequate performance, are seen as assets instead of as waste. This requires entirely new ways of designing the value chain, highlighting restorative and regenerative abilities to maintain systems and prolong their lifetime, emphasizing reuse, repair, remanufacturing and recycling. There are many ways in which CPS technology can be leveraged to support a circular economy, including by providing:

- Identification and tracing of components and entire products;
- Monitoring and prediction of actual as well as anticipated future states of components/products, in turn facilitating planning and decisions e.g. regarding reuse and repairs;
- Upgrading or downgrading of products to provide an adapted level of services and quality;
- Individualized and tailored production of spare parts.

CPS architecting will be essential for many CPS innovations and for making sure that the added functions are themselves circular. Architecting encompasses well underpinned modularization of both cyber and physical parts, to facilitate all aspects of reuse, recycling, upgrading, remanufacturing etc. Part of the specific capabilities may be provided through connected edge- and cloud-based services, "augmenting" the CPS. Having these capabilities in place will strongly support service oriented business model for CPS products.

Implications and impact of CPS – towards balanced and managed complexity

As indicated in the previous sections, the many opportunities and market potential of CPS are promoting a strong market penetration in virtually all areas of society. The implications are that CPS will have a corresponding deeper socio-technical impact, where many vital societal systems will be relying on CPS operation to function. It then becomes important to be aware of the risks and complexity of these CPS, and we will therefore turn to elaborate on these.

Just as the capabilities of a CPS draw upon the characteristics of its components, so do, unfortunately, the resulting risks and complexity. A CPS will for example inherit the faults and failure modes of its component technologies with one implication being the characteristic difficulty of understanding fault-propagation and root causes; a system failure may have a multitude of causes. A CPS may for example fail due to triggered software bugs, hardware faults (mechanical, electronic, etc.), environment uncertainties, improper training of machine learning algorithms and security attacks. With increasing levels of automation it is essential to understand that the *automation paradox* remains relevant, and that its importance is increasing; the fact is that higher levels of automation leave humans with even more difficult situations to handle when the automation fails. This is sadly highlighted by the Boeing 737 accidents related to the so-called MCAS system.

CPS will also inherit complexity facets from both the cyber and physical sides, generally implying that CPS can be characterized as hybrid and heterogeneous systems, requiring multiple viewpoints and experts engaged in their development. Similarly a CPS can be said to manifest multiple complexity facets, e.g. related to its structure, behaviour, heterogeneity, uncertainty and scale [5].

The advancement of CPS capabilities relates naturally to the deployment of CPS in more open and unstructured environments, as is well illustrated by automated vehicles. There will be a widened "cone of uncertainty" when an automated vehicle is deployed and a key question concerns how to assess that the vehicle is ready for the road (homologation). This highlights that current approaches to deal with system safety and security, as well as many other system properties, are challenged [1,17], generally necessitating the connection of developments with operations to improve and adapt CPS during their lifecycle [5]. It is appropriate here to recall the statement of Herbert Simon from his book, The Sciences of the Artificial [6]: "The thesis is that certain phenomena are "artificial" in a very specific sense: they are as they are only because of a system's being moulded, by goals or purposes, to the environment in which it lives."

When deploying CPS in increasingly complex environments with more advanced CPS goals and capabilities, the corresponding CPS will require a corresponding level of complexity in its realization (e.g. in terms of sensors, algorithms, architecture, etc.). This is the price that has to be paid for the enhanced capabilities.

This increase in complexity poses challenges for us as humans. Our standard way of dealing with complexity is to divide and conquer. While this has worked relatively well so far, the problem with CPS is that, however we divide it, there will be interfaces and dependencies everywhere. CPS are integrated systems with both direct and indirect dependencies between parts [5]. As Sapolsky explains it, humans tend to divide separate aspects into categories – "boxing". This helps us to focus but is devastating for our ability to "think outside the box" and moreover, these boxes ("viewpoints"), are arbitrary. The human trait of "group thinking", the fact that we are far from (or have to struggle to be) rational, and the fact that humans are built to take decisions in small groups with local effects, aggravate this problem further [7, 8]. The challenge then

Image: ID 93286496 | © Valerii Brozhinskii | Dreamstime.com

becomes how we are to deal with increasingly complex CPS. We need not only better techniques to master complexity, but also to get better in avoiding over-complex solutions, moving towards balanced and manageable complexity. So-called featuritis is a well-known problem in software products. As Chris Sacca phrased it: – "Simplicity is hard to build, easy to use and hard to charge for. Complexity is easy to build, hard to use, and easy to charge for" [15].

Referring to the Cynefin framework, many future CPS applications will indeed move us from the "Complicated" to the "Complex" domain – a domain where we lack established theory and methods. It is clear that the shift in complexity corresponds to, and will require, a corresponding paradigm shift in development methodologies and scientific theories. It will moreover, considering the Cynefin framework [33], require that the right strategies are adopted. When moving into the "complex" domain, where appropriate regulations, theories and methodologies are still lacking, there is a need for controlled learning, experimentation, research and technology maturation [11, 29]. A similar concept and realization is that of interactive management [30]. This strongly emphasizes the need for multidisciplinary and multi-stakeholder collaboration. Ideally, societies would want to introduce new CPS in a controlled way, balancing innovation with some caution. The situation is complicated by the very strong market forces for many CPS applications, platforms and technologies, akin to a modern goldrush. Billions of dollars are being invested in artificial intelligence, IoT and automated vehicle technologies. This brings with it the risk that unnecessarily complex, or otherwise inappropriate, systems may hit the market.

Many further considerations are needed when developing, deploying and even using CPS because of their potential societal-scale implications, including potential side-effects and emergence as described previously. It becomes essential that CPS are designed to be resilient, manageable and understandable, with explicit account for risk management and the system lifecycle. Introducing new or modified CPS/CPSoS, thus requires careful consideration, and supervision and learning during operation; the widened cone of uncertainty implies that unknowns have yet to be unveiled as the systems are used and evolve.

Many CPS will involve gathering data from users in various forms, directly and indirectly. This raises the questions of who owns the data, who can access the data, how this data might be exploited, and to what extent users should care about their privacy. Shoshana Zuboff raises the concerns of "surveillance capitalism", where (hidden) business models will not only benefit from exploiting our data, but also control our behaviours [13].

With societal systems relying on CPS, it becomes essential for both organizations, regions and economies to have access to such technologies and the right competence to be able to master them. Along with the CPS socio-technical shift, it is thus essential that education and training are emphasized, there is a need for educational renewal [1,12]. There is a further need to bring awareness to a broader set of stakeholders (including decision makers and the general public) [12]. Efforts are needed for developing new theories and methods to deal with CPS complexity [10,12]. Suitable state-of-the-art approaches to organizational strategies (motivation, incentives, etc.) can lead to superior performance of teams, so called "collective intelligence" [27]. This is an area where awareness needs to be raised. Since CPS complexity is reflected within organizations, it is essential to further research on effective organizations including human-Computer Aided Engineering collaboration to manage the complexity. Managing the paradigm shift requires multi-domain and multidisciplinary collaboration. In this setting, the use of competence networks, i.e. nonprofit collaborations between industry and academia to promote learning and knowledge creation, stands out as a promising approach to deal with the outlined issues [32].

Finally, as we develop highly automated CPS, from individual machines with built-in dynamic risk management, to automated factories and large-scale CPS, ethics becomes an important element in our efforts to create human-centric systems. The built-in risk metrics will relate to the system safety, and how we shape socio-technical systems including CPS with humans, will have a large impact on the quality of people's lives, from working conditions to the roles that humans will have in the future.

Conclusion

In summary, CPS as the integration of computation, networking and physical processes are not new and were already widely available through microprocessor integrated systems in the 1970s and 1980s. The novelty and important aspect of CPS is represented by the new scale of capabilities including integration opportunities across domains and lifecycles, with more open society-scale deployment and a corresponding business model evolution.

Cyber-physical systems with their corresponding trends and facts, provide strong indicators for an ongoing socio-technical shift where we as humans will likely overestimate impact in the short run, but underestimate it in the long run (the so called "Amara's law"). Our societies will involve and rely on collaborating with CPS, and CPS collaborating with humans. We need to balance and manage the complexity and risks of these CPS, and leverage them to deal with the environmental sustainability crisis, showing the way towards a circular economy. To accomplish this we need to stimulate new and true multidisciplinary multi-stakeholder collaborations, promote systems thinking and lifelong learning.

Acknowledgements

The author would like to thank Edward Lee for introducing him to the world of cyber-physical systems and colleagues in the CyPhERS and Platforms4CPS projects for fruitful collaboration, all of which have inspired this article. This work has been supported in part by the Nordic University Hub on Industrial Internet of Things (Project number: 86220, funding from NordForsk), the Campus 2030 project (funded by the Strategic innovation program InfraSweden2030(Vinnova), and the Vinnova Competence Center for Trustworthy Edge Computing Systems and Applications at KTH Royal Institute of Technology.

References

[1] Haydn Thompson and Meike Reiman. Platforms4CPSKey Outcomes and Recommendations. Report Platforms4CPS project (H2020, Grant Agreement 731599), 1st ed., 2018, Stuttgart. ISBN 978-3-95663-84-9.

[2] N. Wiener, 1948: Cybernetics: Or Control and Communication in the Animal and the Machine. Librairie Hermann & Cie, Paris, and MIT Press. Cambridge, MA

[3] Martin Törngren, Fredrik Asplund, Saddek Bensalem, John McDermid, Roberto Passerone, Holger Pfeifer, Alberto Sangiovanni-Vincentelli, Bernhard Schätz. Characterization, analysis and recommendations for exploiting the opportunities of Cyber-Physical Systems. Chapter in "Cyber-Physical Systems: Foundations, Principles and Applications", Elsevier, Sept. 216. Editors. H. Song, D. B. Rawat, S. Jeschke and C. Brecher. ISBN: 9780128038017.

[4] E. A. Lee and S. A. Seshia, Introduction to Embedded Systems - A Cyber-Physical Systems Approach, Second Edition, MIT Press, 2017.

[5] Martin Törngren and Ulf Sellgren. Complexity Challenges in Development of Cyber-Physical Systems. In Principles of Modeling; M. Lohstroh et al – editors; Springer, 2018; Vol. 10760, Lecture Notes in Computer Science, July 2018

[6] Herbert Simon. The Steam engine and the computer. EDUCOM Bulletin Vol. 22, no. 1, 1987.

[7] Daniel Kahneman. Thinking, Fast and Slow. Macmillan, 2011

[8] Robert M. Sapolsky. Behave - The Biology of Humans at Our Best and Worst, 2017

[9] Yuval N. Harari. Homo Deus – A brief history of tomorrow, 2016

[10] Hermann Kopetz. Simplicity is Complex - Foundations of Cyber-Physical System Design, 2019

[11] Martin Törngren and Paul Grogan. How to Deal with the Complexity of Future Cyber-Physical Systems? Designs, 4, 2018

[12] Martin Törngren, Saddek Bensalem, John McDermid, Roberto Passerone, Alberto Sangiovanni Vincentelli and Bernhard Schätz. Education and training challenges in the era of Cyber-Physical Systems: beyond traditional engineering. Workshop on Embedded and Cyber-Physical Systems Education (WESE) at ESWEEK 2015, Amsterdam

[13] Shoshana Zuboff, "A Digital Declaration: Big Data as Surveillance Capitalism". FAZ.NET (in German). ISSN 0174-4909. Retrieved 2018-08-28

[14] M. W. Maier, "Architecting Principles for Systems-of-Systems," INCOSE Int. Symp., vol. 6, no. 1, pp. 565–573, Jul. 1996.

[15] BrainyQuote, www.brainyquote.com/authors/chris-sacca - accessed Nov 16, 2020.

[16] Merriam Webster entry for the term "cyber", www.merriam-webster.com/dictionary/cyber - accessed. Nov. 16, 2020.

[17] Bernhard Schätz et al. D6.1+2 - Integrated CPS Research Agenda and Recommendations for Action. CyPhERS (FP7-ICT support action, contract number 611430), project final deliverable, 2015. Accessible at http://www.cyphers.eu/ - accessed. Nov. 16, 2020

[18] P. H. Enslow. What is a "Distributed" Data Processing System? Computer, vol. 11, no. 1, pp. 13-21, Jan. 1978, doi: 10.1109/C-M.1978.217901.

[19] Merriam Webster entry for the term "system", www.merriam-webster.com/dictionary/system - accessed. Nov. 16, 2020.

[20] D. Dori et al. System Definition, System Worldviews, and Systemness Characteristics. IEEE Systems Journal, vol. 14, no. 2, pp. 1538-1548, June 2020, doi: 10.1109/JSYST.2019.2904116.

[21] "List of Largest Internet Companies", https://en.wikipedia.org/wiki/List_of_largest_Internet_companies - accessed Nov. 16, 2020.

[22] E.A. Lee, "Computing Needs Time", Communications of the ACM 2009, 52, 70–79

[23] S. Russel and P. Norvig. "Artificial Intelligence: A Modern Approach", 4th Edition. 2020, Pearson.

[24] IBM white paper - An Architectural Blueprint for Autonomic Computing. Autonomic Computing White Paper, p. 34, 2005.

[25] "GM's hy-wire", https://auto.howstuffworks.com/hy-wire.htm - accessed Nov. 16, 2020.

[26] Tom Harris, "How GMs Hy-wire Works", https://auto.howstuffworks.com/hy-wire.htm - accessed Nov. 16, 2020.

[27] A.W. Woolley, I. Aggarwal, and T.W. Malone, "Collective intelligence in teams and organizations" In T. W. Malone and M. S. Bernstein (Eds.), Handbook of Collective Intelligence. MIT Press, 2015.

[28] Hillary Sillitto, (2010). 1.3.1 Design principles for Ultra-Large-Scale (ULS) Systems. INCOSE International Symposium. 20. 10.1002/j.2334-5837.2010.tb01057.x.

[29] "Cynefin framework", https://en.wikipedia.org/wiki/Cynefin_framework - accessed Nov 16, 2020.

[30] Ackoff RL (1981). Creating the Corporate Future. Wiley: New York.

[31] J. McDermid, Victoria Cengarle, Martin Törngren and Thomas Runkler. Market and innovation potential of CPS – Deliverable 3.2 of the CyPhERS FP7 project, Aug. 2014. http://www.cyphers.eu/sites/default/files/D3.2.pdf

[32] M. Törngren, F. Asplund and M. Magnusson, "The Role of Competence Networks in the Era of Cyber-Physical Systems — Promoting Knowledge Sharing and Knowledge Exchange," in IEEE Design & Test, vol. 37, no. 6, pp. 8-15, Dec. 2020, doi: 10.1109/MDAT.2020.3012087.

[33] "Cynefin framework", https://en.wikipedia.org/wiki/Cynefin_framework

Martin Törngren is Professor at KTH Royal Institute of Technology in Stockholm, Sweden.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: M. Törngren. Cyber-physical systems have far-reaching implications. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 12-19, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

The development of cyber-physical systems (CPS) involves many technology and influencer communities. Novel approaches and tools will be required to tackle the multi-dimensional challenges that connect these communities in order to achieve the desired benefits of future CPS.

Bridging the stakeholder communities that produce cyber-physical systems

By **CHARLES ROBINSON et al.** (for the list of contributors, see the acknowledgements)

There are many communities involved in the creation of cyber-physical systems, which are used in domains including transport, health, manufacturing and, in the longer term, will be in the home, where miniaturization will play a role. In this article we explain that engineering for future CPS needs a centre of gravity that has the purpose of drawing these communities together. This will provide common goals around which technical advances can be aligned. Overviews of the communities involved are provided, with examples of their relevance to build CPS and to some common challenges. Advancements of aggregating technologies are multi-dimensional challenges, representing many influencing dependencies from all communities, especially at higher levels where the whole system product is drawn together. This means that, to make good progress, Europe will require higher levels of coordination for research orchestration and for capitalizing on lessons learned in relation to the cumulative advances among the communities.

Key insights

- Physically interactive and collaborating systems (CPS) involve many contributor and influencer communities in their creation, but these communities have a tendency to make advances in isolation. Creating the technical bridges between these communities is essential for future CPS.
- The scope is wide and communities need a technical interface around which to align. Discussions suggest this centre of gravity for development and operation of such systems to be real-time safe and secure automation.
- Evolved research coordination is encouraged for directing cumulative developments from the stakeholder communities. CPS projects with cross-community challenges and which involve most stakeholders are needed for this.
- The development of CPS requires a holistic development approach that brings together a wide range of disciplines.
- Aggregating technologies have different industrial uptake lifecycles to component technologies. A shared cross-programme research instrument would be very beneficial for investigating and implementing specific technical support (for projects, programmes and industrial policy).
- Guided by the target products, with a centre of gravity around which to interface and with tailored support, a holistic development approach for the communities will be enabled.

Key recommendations

- Creating the technical bridges between communities is essential for future CPS.
- Communities that contribute to CPS, and affect their development and operation, need a centre of gravity in order to relate to each other. This is proposed to be real-time, safe and secure automation.
- Research orchestration needs to be developed for coordinating cross-community CPS research. This also calls for projects tackling CPS challenges that are common across the communities.
- A shared cross-programme research instrument, dedicated to the application side, is valuable for advancing in particular aggregating technologies and in general the technology uptake by CPS.
- Through the cross-community projects; CPS-focused coordination and support actions; and an ongoing instrument for research orchestration, cross-community collaboration can be developed and refined.

Introduction and new cross-community development approaches

In order to manage large complex problems, people break them down into parts. It is for this reason that, from a technology point of view, there are many contributing and influencing communities involved in the creation of future CPS products. Of course, the parts need to then be assembled together in order to address the initial complex problem. For the same reasons, the various technological contributions to CPS require layered aggregation in order to achieve these physically interactive and collaborating systems. This means that there are significant multi-dimensional influences across the communities and they contribute to our ability to transfer technology to industry: approaches and technologies used for aggregation play a significant role in creating CPS.

Where future CPS is mentioned in this article, it is in the context of an application; that is to say, the term can be replaced directly with e.g. railway transport or satellite constellations. Describing CPS from the technology perspective and providing a concise definition of a CPS, is out of the scope of this article and the reader is invited to look, for instance, at the articles "Cyber-physical systems have far-reaching implications" and "Cyber-physical systems from the application perspective" respectively. Suffice it to say that in this article, CPS represents the future *physically interactive and collaborating systems* that are present in many domains including transport, health and manufacturing.

The involved communities, discussed in the subsequent section, range from providers of a) functional properties such as sensing, physical action and processing to b) system-level engineering including properties like safety and performance specifications, managing customer requirements, architecting, system validation, mechanical engineering and control engineering. There are technology support communities providing c) enabling technologies like the Internet of Things, systems of systems, big data, artificial intelligence and high performance computing. Finally, there are the influencing communities from d) the production environment, with enterprise processes and product line, and e) the market, such as regulation and current and future needs of society.

These communities have tended to transfer technology as a one-to-one mapping with products, however they will need to take relations with the other contributing communities much more into account, so as to be able to foster responses to the challenges of future CPS as well as to enhance technology transfer. While the challenges and importance of advancing aggregation techniques are discussed later, there also needs to be a common point from which one community can interact with any of the other communities. It should provide a common interest based on physical challenges of these systems. *Discussions have proposed this centre of gravity to be real-time, safe and secure automation* of CPS development and operation.

Such a centre of gravity represents three limiting factors that it is in the interest of all contributing communities to see addressed and which are relevant to key common CPS challenges. For technologies to be accepted in these systems, they must be compliant with the safety and security constraints of a product and not compromise real-time responses. This means the easier it is to couple your technology with these system constraints (through automation), the easier it becomes adjust it to the system (or adjust the system for new technologies). It is usually the case that, in order to add new technologies to a CPS, the whole system requires re-certification, which can be prohibitive without sufficient automated information about the impact on safety and security. As a result, a centre of gravity, as shown in Figure 1, provides a very useful point to which all the communities can relate and contribute.

While management of trade-offs between the system properties of performance, safety and security is an established skill in system development, it still remains very much a manual and qualitative process and one that is based on prior experience, and in need of transformative automation. It remains to this day very much a bottleneck and is holding back the communities contributing to CPS development from making advances in areas such as trust in artificial intelligence applied to CPS. This being said, automation between these system properties can rely on a number of decades of research in techniques [1], some of which have already been applied in industry but are generally in need of new approaches for technology transfer. Such approaches are included in the coordination suggestions for research orchestration described later in this article. Of course, current pressures for industry to find advanced solutions for managing system property trade-offs are also driving the search for automated coupling. As examples of some initiatives, the UK Research Institute in Trustworthy Interconnected cyber-physical Systems RITICS involves dozens of UK universities and industrial collaborators. Topics include safety-security and autonomous systems. Relating to autonomous vehicles, the Intel Research Collaborative Institute of Safety of Autonomous Cars (ICRI-SAVe), deserves a mention as a vibrant community. Many industries are actively looking for solutions to manage the performance, safety and security of their products, and include large enterprise like Siemens, Thales and AVL, who have been forming combined safety-security teams. The challenge also

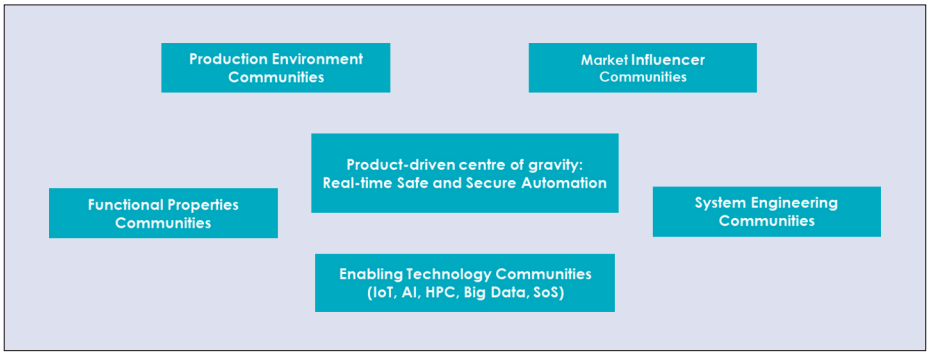


Figure 1: The stakeholder communities for creating CPS.

affects small and medium-sized enterprises (SMEs) in their products and services. This recent momentum has visibility, for example, in the Ada and IEEE conferences, in recent large research collaborations including MERgE, SeSaMo and AQUAS and co-engineering discussions.

Overview of stakeholder communities for creating and advancing CPS

We now provide overviews of the five communities, indicated in the previous figure, which are involved in creating CPS. We give descriptions and examples of their relevance to CPS as well as their relation to cross-community challenges for future development. These include embedded computing as a CPS backbone, system decentralization and decomposability, and physical collaborations with people.

Relevance of functional properties

While considering physically interactive and collaborating systems (future CPS) the *functional properties* have to address aspects that cover sensing, actuation, communication, energy provision, processing and coordinated collaboration. Such properties are key characteristics of these systems with actors in specific communities researching and developing the different components.

The relevance of functional properties becomes more evident when considering novel and innovative advanced applications that are being progressively adopted in a number of large-scale, safety-critical domains e.g. industry, transportation, smart cities, critical infrastructures, space, etc. Some examples can be found in H2020 projects such as CPSwarm and other CPS cluster initiatives. Industry-driven needs and the well-established nature of research communities in the CPS domain, mean that it is feasible to envision projects that might prototype concepts such as swarms of unmanned aerial vehicles and rovers supporting safety and security operations; swarms of automated ground robots that collaboratively support humans in logistic operations within a smart warehouse; or enhanced and dynamic platooning applications for autonomous freight vehicles. The development of such applications cannot currently leverage a simple plug-and-play integration of the various technologies entailed, given the complexity of managing teams of systems and humans in evolving and dynamic scenarios with emergent properties.

As a consequence, in order to properly combine and integrate the different required technology building blocks, the various 'functional properties communities' have to be properly engaged. Experts from the functional property communities will need to work with other actors with collaborative systems competence. Moreover, while the increased adoption of CPS has resulted in the maturation of solutions for CPS development, a single consistent science for future CPS has not yet been consolidated. Few functional properties community members have already started working alongside other communities on a connective framework e.g. using modelling, design/development tools and methodologies, deployment solutions, monitoring and controlling solutions for large-scale challenges. In this context, model-centric approaches have clear relevance for facilitating collaboration between experts from different sectors and thus enabling the definition, composition, verification and simulation of collaborative, autonomous CPS.

For these reasons, it is important for future CPS to be considered not only from the technology perspective but also as an **application domain** where the technology of the functional properties community plays a role for aggregation of CPS-related research. To promote this, closer and wider collaboration is needed within the community, along with new research initiatives. Understanding the nature of this aggregation from bottom-up and top-down is important for driving the communities towards much needed technology advances. The resulting collaboration plays a very important role in finding solutions to the bottlenecks that currently prevent CPS from having greater impact on the society; such solutions would also promote market uptake, open up new markets and optimize the use of resources in the various industry sectors.

These communities have many cross-cutting challenges for future CPS. Embedded computing will evolve significantly for such systems and plays an essential enabling role for functional properties. For instance, the need to use specific sensors on a CPS and to process onboard the relevant raw data will need increased computational power. However, energy limitations introduce other constraints: only a holistic vision of CPS can help drive research initiatives. In relation to decentralization and decomposability, with distributed intelligence and emergent properties, an example research context would aim to solve/work on delays in physical, computing and actuation timing. This requires model design and simulation approaches to capture the whole heterogeneity of the system and its contributing communities. Physical interaction with people requires a system to have high fidelity knowledge of its environment and its physical dynamics. This requires the technologies of the functional properties community which in turn need integration within the safety and security measures set by the system engineering community. It clearly emerges that the best way to advance future CPS is to further support integration and aggregation approaches for community collaboration.

Relevance of systems engineering

The development of CPS requires a holistic development approach that brings together a wide range of disciplines. This includes the typical systems engineering disciplines, such as requirements engineering, architectural design, implementation and quality assurance including system-wide responsiveness, safety and security. The disciplines of this community are important in terms of both the CPS in general and individual systems engineering sub-processes, such as mechanical engineering, control theory, electrical engineering and software engineering.

In almost all of our application-driven future scenarios, like in autonomous driving and Industry 4.0, CPS must be able to fulfil their purpose to a large extent without the intervention of human users [2]. According to the Society of Automotive Engineers (SAE) taxonomy for autono-

mous driving, we refer to such systems as highly-automated or fully-automated CPS [1]. Already today and even more so in the future, systems engineering is one of the core competence fields for building such highly automated or fully automated CPS.

In the case of highly automated CPS, it is necessary to have a more comprehensive understanding of the term 'functional safety'. In contrast to the understanding of the term by ISO 26262, which essentially considers the malfunction of system components, highly automated CPS require an analysis of the interaction of a) the functionality of the CPS under consideration with b) its context (e.g. other CPS in collaboration). This analysis serves to detect possible safety threats resulting from the interaction between system functions and contextual conditions, such as the interaction between the autonomous driving function of a vehicle and the failure of the signalling system at an automated road intersection. This new understanding of functional analysis, which goes far beyond the requirements of ISO 26262, is the subject of the SOTIF standard [3]. These threats to safety must be identified during the development process and mitigated, e.g. by specifying suitable requirements. Since CPS often monitor and control technical or physical processes, control theory is a discipline of great importance in the development of such systems. In this context, the concepts of monitoring and controlling technical/physical processes are reflected in various artefacts of systems engineering. For instance, the requirements originated from the way the processes should be controlled, as well as from decisions made about the design of the necessary sensors and actuators or even about the design of the algorithm for the computational processes of the feedback system.

In order to be able to develop such complex technical systems consisting of software and hardware, seamless systems engineering processes are required, establishing techniques, methods and tools for challenges such as the following examples. Since CPS in many fields of application work together in dynamically formed networks at runtime to pursue higher-level

goals, possible collaboration structures must be identified and analyzed in requirements engineering. For example, in the development of autonomous vehicles, it must be considered in which collaboration structures these vehicles must operate, such as in vehicle convoys to optimize the flow of traffic or at automated intersections to ensure safe crossing of the intersection, even with high traffic volumes and in complex traffic situations. In collaborative CPS, the issue of coordinated decentralized monitoring and control of technical/physical processes is added; an example of this is the coordinated acceleration or deceleration of the various vehicles within a convoy of vehicles. In the case of highly automated systems, the involvement of the human user is required in (a few) defined situations to ensure that the system is able to fulfil its purpose of ensuring safe operation. The integration of the human user must be effective, i.e. the user interface of these systems must be designed in such a way that the human user is able to perform the necessary tasks according to the intention, as free from errors as possible and within the existing time restrictions. One might think here of the example of autonomous road traffic, where highly automated systems require the driver to take control of the vehicle when a critical driving situation occurs.

Relevance of enabling technologies

1. Internet-of-Things (IoT)

The Internet of Things community developed around the goal of providing a means for all devices to be globally connected via the internet. The name Internet of Things was first used in 1999 by Kevin Ashton during a presentation to his higher management at Procter & Gamble. He described IoT as a technology that connected several devices with the help of RFID tags (radio frequency identification) for supply chain management. In 2008 the first international conference on IoT took place in Switzerland, discussing RFID, short-range wireless communications, and sensor networks; these topics continue today to represent the major technological research domain for advancing the IoT, gathering information about the real world that can then be made useful in some way [4].

Since 2010 it has been normal for many different devices to be in our homes to be connected to the internet. Connected devices are used extensively in the consumer domain. In 2015, to support advancement of IoT for industry, the European Commission created the Alliance for Internet of Things Innovation (AIOTI). Applying IoT to the industrial environment has been termed industrial IoT, or IIoT and has the

goal of optimizing production value while considering the many additional challenges related to safety, security and performance. IIoT technologies support interconnectivity with the internet in the context of these challenges, enabling not only networked smart objects and information technologies but also "optional cloud or edge computing platforms, which enable real-time, intelligent, and autonomous access, collection, analysis, communications, and exchange of process, product and/or service information, within the industrial environment" [5]. Thus IoT and in particular IIoT technologies will be standard constituent elements of future CPS.

Enabling the infrastructure to support distributed intelligence and information exchange is at the core of IoT, so supporting cross-community work on CPS decentralization, decomposability and human interaction is important. These are already areas receiving some focus from the IoT community [6,7], as indeed is the case for bringing communities around an embedded computing backbone, with work considering edge-cloud computing [8] exchanges.

2. Artificial Intelligence (AI)

Autonomy will bring incredible new benefits to CPS, but is also faced with major challenges that must be overcome in order to realize future cyber-physical systems. The intelligence that can be applied is limited by current approaches to certification, legal frameworks and (lack of) trust for such systems. These need to be addressed while maintaining and increasing the safety of such systems *(which calls for improved traceability of the influences between the contributing communities to CPS)*. Reducing or mitigating these limiting factors will be an enabler for many advanced AI technologies related to decision making, learning etc, for the operation of the systems. In parallel, the other communities can provide technologies more robust for systems that are evolving as a result of AI. Of course, there are identified routes for AI to become 'more trustable'; these include explainability of actions in human language, and the application of AI to non-safety-related aspects of CPS like decision support for system design.

A significant characteristic of CPS will be coordinated collaboration. This relates to the way components of a CPS coordinate with each other or with people for outcomes only achievable through such cooperation. AI can bring strong support here such as through the field of decentralized intelligence called Multi-Agent Systems (MAS) [9]. Regarding design, the needs of CPS include the explicit representation of the environment and the need to represent abstraction layers, from the physical layer to the components and system, as CPS are closely coupled to the hardware elements of the system. Finally, it may also be necessary to represent the non-functional requirements, such as safety or resilience. Some MAS design tools, such as Tropos [10], if correctly used, may help to meet these requirements.

In terms of decentralized intelligence for CPS, there are many challenges to that need to be addressed, in particular methods for executing coordination. The whole system needs to be able to react in real time, which is not the case for most decentralized AI coordination protocols, which rely on negotiation, usually with no defined deadline for decisions [10]. As another example, finding ways to work with the functional property community on communication middleware for intelligent collaboration is likely another issue needing to be tackled.

3. High Performance Computing (HPC)

High performance computing (HPC) consists of the aggregation of highly powerful computing resources for solving problems that require large computing power [11]. Recently, HPC technologies were only required in the context of traditional massively parallel "number crunching" applications like weather prediction, computational chemistry, or computational fluid dynamics. However, the latest developments in low power computing technologies [12] – required in the HPC industry to scale performance levels further – has facilitated the adoption of HPC technologies in a wide range of CPS applications.

Existing HPC platforms offer the computation capabilities needed by the most demanding CPS applications within an affordable power budget in domains such as automotive, space, avionics, robotics and factory automation. Centralized domain architectures that replace the traditional federated computing architectures – like those required by economically affordable autonomous driving systems – are only possible when HPC technologies are deployed. Single-chip high-performance embedded computing platforms reduce the traffic flow through CPS' electronic networks and enable high-speed communication as required for processing vast amounts of information in real time. So this community will be important for consolidating the embedded computing backbone. Furthermore, these technologies involve parallel processing, that is, splitting the tasks up into parts for several computers (or multiple cores) to process, thus reducing the time taken to complete tasks. This characteristic thus holds a direct relation with the CPS challenges of decomposability and decentralization – how tasks can be split up while ensuring safety and security for people, the system and its environment.

Unfortunately, the deployment of HPC in a CPS increases the complexity of the resulting system and may have non-negligible impact on the verification and validation costs of relevant system properties (e.g. safety and security). Thus, an effective exploitation of HPC technologies in cyber-physical applications requires at least either the development of new methodologies to verify and validate such complex systems or the adaptation of key technologies to the specific context.

4. Big Data

Cyber-physical systems are being driven by the combination of embedded and internet technologies and the vision of "Smart Anything Everywhere". The blend of this cyber, physical (and social) data can help us to understand incidents and changes in our adjacent environments better, monitor and control buildings and urban infrastructures, and provide better healthcare and elderly care services, among many other applications. To make effective use of the physical-cyber-social data, integration and processing of data from a variety of heterogeneous sources is necessary. A key objec-

tive for big data in CPS is to analyze very large, fast, and heterogeneous data streams from, mostly, industrial environments. This can be achieved through machine learning, which is the most common technique used to extract information from the data.

The core CPS Big Data applications are in varied fields such as energy utilization, city management, transportation systems and disaster management. For example, a smart transportation system would generate big data consisting of driver behaviour, commuter information, vehicle locations, traffic signals management, accident reporting, automatic fare calculations, and so on. Robot-aided surgical systems (i.e. human-in-the-loop CPS) comprise a teleoperation console operated by a surgeon, an embedded system hosting the control of the automated robot, and the physical robotic actuators and sensors. Big data methods can be used here for modelling surgical skills, detection and classification of surgical motions for automation and environment, and integration of this knowledge into control and automation of surgical robots. In the operation of complex systems (e.g. aircraft and industrial processes), fault detection and isolation schemes are designed to detect the onset of adverse events. Such systems use big data methods (such as machine learning classifiers) to enhance the diagnostic accuracy of the online reasoner on board the aircraft. Moreover, big data can be utilised in command and control with cyber-physical infrastructures for emergency services and defence.

The value of the Big Data community as a contributor to CPS products can only grow in the future due to increasing interest in data as an important business asset. The combination of heterogeneous data from numerous sources will require new applications for integration, query and analysis, along with embedded computing, high performance computing, and data reduction techniques. This remains an open research issue for CPS. The variety of types and sources of data will give rise to new kinds of data stores to sustain flexible data models. Another important issue is that of remote storage of big data. Until now, cloud-based models have facilitated the storage and processing of big data sets, providing data accessibility and better IT power. However, this creates a centralized data store that does not scale in the CPS setting. To facilitate decentralized data storage and processing, a number of problems (e.g. replication, parallelism and requirements) arise. There is an urgent need for new approaches and techniques.

5. System of Systems

The concept of 'system of systems' (SoS) has been around for at least fifty years, but in the last twenty it has been an area of major concern. Following the description of its characteristics by Maier [15]; it is defined in ISO15228 as: *"SoS…brings together a set of systems for a task that none of the systems can accomplish on its own. Each constituent system keeps its own management, goals, and resources while coordinating within the SoS and adapting to meet SoS goals*" [16]. As for CPS, SoS represents a type of application as well as a technology domain.

Broadly, one can consider SoS applications as independent systems that interoperate (work together) to achieve a purpose, with a significant amount of ubiquitous networking. In the case where they have extensive software control between safety critical systems, the application itself is both a SoS and a CPS because they share common characteristics. Figure 2 describes the relationship between SoS, CPS, and the Internet of Things. Where infrastructure interactions are supported by internet protocol, then the CPS is also described as IoT, which is necessarily always a SoS. There are also interesting SoS-CPS applications that interact through means other than the internet protocol (e.g. mechanical or electromagnetic interactions) and the engineer may need to guard against such interactions for safety or performance reasons.

However, from the technology perspective, CPS application research considers how all technology communities are integrated to create a system and its interactions, with the SoS technology community contributing to the coordinated collaboration aspect. This is a key property for future CPS meaning SoS research is indispensable for creating future CPS. In relation to embedded computing, the importance of localized processing while also having connection to centralized processing capacity is recognised as a priority, in areas such as edge computing, which uses SoS technology. This also links directly with the challenge of decentralization or decomposability where systems work together. A smart city is an example of human interaction and SoS, for example; it manages busy traffic at city junctions to minimize delays for drivers and pedestrians.

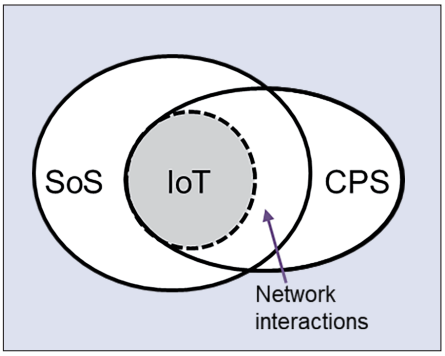


Figure 2: Technology relations of SoS, IoT, and CPS [17]

In 2012, INCOSE conducted a survey to identify "pain points" for SoS practitioners, i.e. the problems that kept systems engineers and managers awake at night [18]. The study indicated seven main areas of concern: SoS authorities; leadership; constituent systems; capabilities & requirements; autonomy, interdependencies & emergence; testing, validation & learning; and SoS principles. It is no coincidence that creating CPS includes these pain points, because they are concerned with networked, intelligent systems of high complexity. Thus, one could argue that the communities of SoS and CPS have areas of common interest suitable for collaboration.

Relevance of the production environment influencers

The members of the production environment communities are responsible for the industrial product process and lifecycle. This includes enterprise policy and processes, decisions about technology usage and the evolving physical plant [19]. They drive the large-scale production of goods using equipment in the form of modular automated product lines. Such equipment typically combines mechanical,

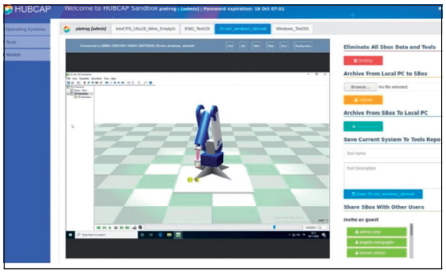


Figure 3: Snapshot from the Sandbox showing SME asset.

electrical, and software components and requires substantial initial investment and maintenance costs. Throughout its long lifecycle (15 - 30 years) [19], the equipment operator and component suppliers cooperate to repair and repurpose/upgrade parts at a minimal cost. This imposes several constraints on component models and their versions, which in turn constrains policy and process management.

Besides, the arrival of digitalization and the CPS revolution brings the "servitization in manufacturing" opportunity, a paradigm shift where manufacturers shift to offer product-related services, beyond just selling a tangible asset. In the above example of automated product lines, the component providers could offer online maintenance, repair, and overhaul services [20] amongst other value-added services. Service contracts generate more steady revenue compared to the cyclical product business, but, in general, organizations in manufacturing struggle to drive servitization [20], because the introduction of the new services incurs higher costs without proportional returns.

The adoption of digitalization tools and solutions and the development of innovative services leveraging the full potential of CPS require incentives and coordinated efforts among different partners. Research projects, partnerships in which early movers and less digital companies cooperate to embrace servitization and adopt CPS tools, provide a nurturing environment, where decision makers find that test-before-invest concept is an incentive that helps lower barriers and can evaluate potential benefits. For example, in the H2020 HUBCAP [21] project, companies find a one-stop-shop for embracing digital innovation and developing solutions using model-based design technology. The offering encompasses: 1. a platform with a cloud-based sandbox solution with pre-installed models and tools; 2. a decentralized network of Digital Innovation Hubs providing access to training and skills; and 3. an open call programme to attract and foster experiments, and to establish partnerships among companies.

The prime innovative aspect of HUBCAP is a web-based collaboration platform that facilitates stakeholders' access to computing resources and advanced CPS design and engineering solutions, by providing a cloud-based sandbox solution (Figure 3). The sandbox provides pre-installed models and tools, allowing companies to experiment with new tools and assets in a ready-to-use virtual machine available via a regular web browser.

The production environment community members are deeply involved with the cross-community challenges identified. In the case of the embedded computing backbone, there is a historical synergy in the development and advancement of embedded computing, which will continue in the future. This community is always demanding advancements in embedded computing, and advances in manufacturing also affect how we produce the embedded platforms of the future. Regarding decentralization and decomposability, there are several lessons learned and case-studies in which cooperation and adaptation to local and greener processes foster research, discussion, and changes to manufacturing. Finally, this community has a particular interest in the challenge that is physical collaboration with people. This interest is from both an internal perspective, covering topics such as human-machine interaction and collaborative robots, and an external perspective where the potential for improvement from product usage data impact practice needs to be fully explored.

Relevance of market influencers (society needs, regulation, standards, policy)

CPS are believed to have an enormous impact on many aspects of socio-economic life. Therefore, a number of stakeholders grouped here under the generic name of 'market influencers', will have a stake in shaping the future of CPS and of the contributing communities.

Society needs may be described basically through the individuals or groups benefitting from CPS. The individual appears here as the consumer who is, in one way or another, making use of either a product incorporating CPS, or elements of larger CPS implementations, addressing communities of end users in terms of mobility, personal life (general wellbeing), healthcare, leisure, environment, etc. A further area of needs is represented by public services offered at local and national government level, including education, healthcare services, community services, and operation of public institutions. Some specific fields include education and employment, as CPS induces obsolescence of certain professions and creates new ones. Therefore education, including training and retraining will be affected, as will the employability of the existing and future workforce, which will have implications for the labour market and social security.

Regulation – both hard and soft legislation – will have to be adapted in order to govern CPS so as to ensure their smooth integration into society. However, given the rapid cross-border spread of CPS technology, international agreements might be needed, too, particularly if we consider the international nature of today's value chains. Regulation will have to address the interplay between CPS actors (producers, consumers) as well the foreseen and unforeseen effects of the technology. Regulation is also supposed to be structured according to the societal needs that the technology is supposed to fulfil. A particular aspect of related regulation might address the human individual, chiefly in relation to human-machine interaction, which is anticipated to increase significantly in the coming years (intruding into both privacy and healthcare). The 'must be implemented' regulation should be supplemented with recommendation type measures of indicative nature.

Standards ensure interoperability and compatibility of products from different producers and allow the market presence

of a large number of actors. Moreover, standards are important in order to set and describe safety levels and quality frameworks. To some extent, standards provide the technical base for legislation governing the area and also give room to innovation as usually standard specifications can be fulfilled in a variety of competing ways.

Policy aims to achieve certain results in the given field by reflecting society's needs or goals. Public policy in particular is directed towards the fostering of certain areas through frameworks of development in terms of tax incentives, grants or even regulation. Policy also includes public investment in facilities or processes of general interest. A further aspect for consideration is policies aiming to increase employment in a differential manner within the given population (i.e. in favour of disadvantaged groups), or to ensure development of regions lagging behind. Such policies also set out to address issues of general interest like climate change (that can only be done at international level) or the environment. Beyond public policy, one should take into consideration policies of generically named "groups of interest". Enterprises, for example the NGO type ones or consumer associations also have policies for their vision and procedures supporting their realisation although the large part is internal and unrelated or very indirectly to market interests such as charitable events. Business or product policies inside an enterprise is the domain of the previously described production environment community.

These 'market influencer' stakeholders between them offer a robust representation of the conditions under which all the other communities operate for producing future CPS. The relevance of their involvement should be apparent, especially when considering the aggregative effects of contributing and cross-community technologies. Deficits in education in one community can have a knock-on effect on other communities. Training approaches and certification can be a deciding factor in sustainability of mixed-community technologies. Policy can evolve approaches and perspectives that enhance behaviours supporting longer-term governance or culture, providing resilience, value generation and trust in new technologies.

Research orchestration for cyber-physical systems

With respect to coordinating research of CPS as an application domain, additional approaches and orchestration should be introduced. This is because the application domain perspective is based on the product side, with cumulative effects being considered through the aggregation of layered contributions from the stakeholder communities. Orchestration of research is particularly about knowledge management, longer development cycles, persistence and refinement of multi-disciplinary approaches for collaboration between communities. Take the example of constructing a building where a new team takes over every few months. Limited progress can be made without guidance at a higher level. This is similar for advancing CPS research. Persistence of acquired interaction techniques, between project collaborations, is significantly more difficult to maintain. For instance, usability and sensor experts have specific languages for their domains. Therefore, approaches that support collaborations and which have been developed during said collaborations should be taken, refined, and applied in subsequent collaborations of different groups. A dedicated CPS research instrument could advance this concept, in conjunction with future CPS support action projects. Projects themselves will also need to provide environments with favourable conditions for aggregative research considering the multi-dimensional challenges, with conditions significantly different to those for developing component technologies.

Considerations for future CPS projects

For advancing CPS research as a technology domain, useful mechanisms already exist. For example, there have been projects following the standard approach, which gathers technology providers around one or more CPS-related use case. If awarded funding, the partners then work together for a few years to bring their technologies closer to market deployment (we tend to talk about advanced 'technology readiness levels' or TRL).

Cascade funding, where funded projects themselves fund smaller initiatives, has also shown itself to be a useful means for transferring component technologies for CPS, because the smaller initiatives are directly managed by companies looking for particular solutions.

However, *for the application domain side* of CPS research, new project approaches and higher support mechanisms also need to be introduced, enabling the multi-dimensional challenges previously discussed to be tackled. The characteristics that are believed to be essential in such projects are:

- Use cases: physically interactive and collaborative systems; of relevance to all communities, likely to be uniquely large industry or with integrated small-medium enterprises. Supplied also with the intention of advancing 'industrial readiness levels' of production and product lifecycles for new technologies.
- CPS centre of gravity: all projects addressing the multi-dimensional challenges between communities should interface on work advancing real-time safe and secure automation for CPS design and operation.
- Cross-community challenges: projects on application domain research should focus on grand challenges that need contributions from each community. Proposed call topics include:
  - Embedded computing backbone
  - Decentralization & decomposability
  - Physical collaborations with people
- Developing the support environment: tools and approaches are required not only by industry, but also by researchers to support engagement of the different CPS stakeholders and perspectives. It is proposed for such projects to include some dedicated work (a work package) that develops support for collaboration on the multi-dimensional challenges.
- The new approaches established iteratively: orchestration approaches should be implemented in a manner that can be refined. Avoid 'one hit wonders' that seek to solve everything at once. A second iteration of such projects could also include smaller spin-offs and initial stage smart city investigatory projects.

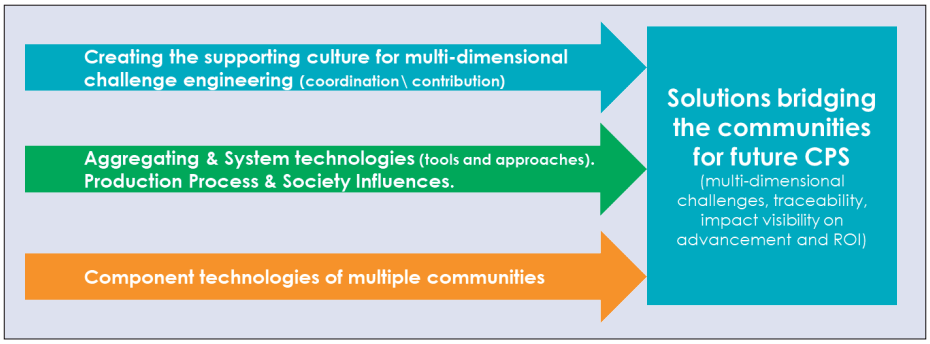


Figure 4: Stakeholder inputs to projects.

Contributions to be these projects can be visualized, as shown in Figure 4, to involve the technology component providers, the influencers/aggregative technology providers and those developing the culture and support environment. This provides the means to advance on the CPS aggregation techniques which are required to address the multi-dimensional CPS challenges.

These HiPEAC proposed project characteristics (for application domain research) relate directly to previous CPS community recommendations to the Commission, including trustworthy and societal scale CPS, ethics data protection and liability, CPS engineering, interoperability, complexity, edge computing, humans-in-the-loop, co-engineering of system properties and enhancing uptake of CPS technologies.

Considerations for future advisory coordination & support actions for the CPS communities

The European Commission funds coordination and support actions (CSAs) to accompany, coordinate and stimulate innovation in particular technology fields and their communities. CSAs also help the Commission to navigate particularly complex topics. CSAs carry out studies and engage with specialists in the community from both academia and industry. This is done through workshops and conferences, in order to identify key technology challenges, priorities and where support can be provided, e.g. coordination, awareness-raising, policy discussion and support for standardization.

A particular challenge for a CPS CSA is that it is in fact a multi-community subject. This is because, while CPS is a technology domain with specific complex challenges related to cyber and physical integration and cyber to physical plan realisation, CPS is foremost an application domain. This is of consequence because CPS and other technologies can be much more difficult to apply to the final systems without also advancing the means for their combination. To support application domain research projects, future CPS CSAs will likely support the transfer and synchronization of project environments, support the 'big picture' metrics of aggregations in CPS and specific ROI valuation techniques to pre-empt industry needs. In particular, they will support a focal point for all the contributor technology and influencer communities.

A CPS programme instrument: supporting projects, funding programmes and industry

A cross-programme team, dedicated to support on the application side, would be very beneficial for supporting in particular aggregating technologies and technology uptake by CPS in general. This will happen if the cross-programme team is an ever-present pivot for CPS projects and CSAs, developing the support environment required for the multi-dimensional challenges.

They would have two support roles: development and investigation of concepts that are provided by projects and programmes and would likely to be assets, but their implementation being normally outside their scope of operation.

From the development side, support to programmes would include, as an example, enhanced tool techniques for directed communication (the right information, to the right people, at the right time – especially for start-ups). For the projects, the team provides prototyping tools, where the CSA would support deployment in CPS projects, who then test and further develop the tools. Support for the creation and testing of tools largely depends on results from the investigatory side. Some examples include:

- Inter-community supports like wiki-type project glossaries to manage the multiple perspectives (e.g. mediation between safety/security, medical/railway, SME/LE).
- Multi-community access like digital passports, allowing users to access and test many research tools with the same account.
- Improved techniques like supporting management of intellectual property rights.
- Connecting contributions like global vision on open source tool advancement across projects.

The investigatory side considers and proposes enhancements, from the product-side perspective, for projects, programmes and industrial policy. These would be potential assets for promoting in particular aggregative technology uptake and longer-term profitability. Investigations would consider enhancements outside our normal fields of operation. Potential concepts for attention include:

- Supporting the project environment for capitalization on and continuation of knowledge from multi-stakeholder interactions. Approaches for iterative improvement. Incentives, performance measures, mentoring. KPI implementations will be useful to investigate while watching out for the Cobra Effect.
- How are CPS-specific and aggregative technologies advancing, what is the funding flow to the contributing communities? Studies on benefits – but also consequences of lack of funding.
- Managed contributions, e.g. open source results – rather than a default expectation, should be with respect to conditions (such as business model, maintenance, community building).

- Considerations for adapting the destination (industrial processes) to the new technologies; how to lift constraints at the product-side.
- Lighthouse initiatives within programmes (advancing structuring and management policies) may provide ideas to be explored.
- Technology readiness levels (TRLs) measure the advancement of individual components rather than aggregations of components, so a complementary approach, let us say Aggregative-TRLs, is likely needed. This is not to be confused with the 'integration readiness levels' measuring the interface between technologies (how they connect), rather than aggregating technologies (managing their combined effect).
- Supporting the development of a body of knowledge – teach the science of CPS engineering.
- Balancing local/national/European interests across networks. For instance, cross-border Digital Innovation Hubs (DIH) could complement the specific interests of regional or national DIHs.
- Policy on protection of EU business data (~B2B GDPR). CPS representation would be relevant here to consider effects of such policy on CPS technology advancement.
- Studies to advise/encourage industry towards longer-term strategies. This may also include changes in government regulation to shift from short-term competition of yearly quotas towards longer-term and more profitable competition and managing incentives where average employee turnaround is 3-4 years. No CPS-specific studies on corporate evolution seem to exist yet.

The proposed way forward through this higher-level support from a CSA and a research instrument, not only enables advancement of CPS application domain research, but also addresses the recommendations made by previous visionary projects for CPS technology (with Platforms4CPS representing an update of several roadmaps). These earlier recommendations included: collaboration and defragmentation of siloes; public understanding of the importance of CPS; supervisory support to draw together a common body of knowledge; and developing talent in order to maintain Europe's leadership and sovereignty of diverse technology aggregations for multi-domain applications including transport, manufacturing and health.

Acknowledgements

Contributors for the five communities

- **Alessandra Bagnato** is Research Scientist and Head of Modelio Research at Softeam (Docaposte Group).
- **Claudio Pastrone** is Head of IoT and Pervasive Technologies Research Area in LINKS Foundation.
- **Thorsten Weyer** is Head of Requirements Engineering and Conceptual Design, paluno (The Ruhr Institute for Software Technology, University of Duisburg-Essen).
- **Peter Popov** is Associate Dean (International), School of Mathematics, Computer Science and Engineering, City University London.
- **Hugo Daniel Macedo** is Researcher in the DIGIT Centre, Department of Engineering, Aarhus University.
- **Claudio Sassanelli** is Researcher in the Manufacturing Group, School of Management of Politecnico di Milano.
- **Peter Gorm Larsen** is Professor and Head of the DIGIT Centre, Department of Engineering, Aarhus University.
- **Carles Hernandez Luz** is Senior Researcher in Processor Designs for Safety-Critical Systems, Universitat Politècnica de València.
- **Michael Henshaw** is Professor and Programme Director in Systems Engineering, Associate Dean for Teaching, Loughborough University.
- **Cédric Buron** is Research Engineer in Artificial Intelligence at Thales Research & Technology, France.
- **Rajendra Akerkar** is Professor and Head of Big Data Technologies at Western Norway Research Institute.
- **Miklós Györffi** is Senior EU Affairs Analyst and former member of the European Parliament.

References

[1] State of the Art White Paper Report: Recommendations for Security and Safety Co-engineering - Part A, S. Paul et al., MERgE Project, Feb. 2016.

[2] SAE International: International Standard J3016: Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems, SAE, 2014.

[3] International Standardisation Organisation (2019). ISO/PAS 21448:2019 – Road Vehicles – Safety of the Intended Functionality.

[4] https://www.techrepublic.com/article/how-the-term-internet-of-things-was-invented/

[5] H. Boyes et al., The industrial internet of things (IIoT): An analysis framework, 2018

[6] K. H. Wöhnert et al., Secure Cyber-Physical Object Identification in Industrial IoT-Systems, 2020.

[7] S. Sachdev, Voice-Controlled Autonomous Vehicle Using IoT, 2019

[8] A. J. Ferrer, Admission Control for Ad-hoc Edge Cloud, 2021.

[9] M. Wooldridge, An introduction to multiagent systems, John Wiley & Sons, 2009.

[10] C. Cares, S. Sepúlveda et C. Navarro, "Agent-Oriented Engineering for Cyber-Physical Systems", in International Conference on Information Technology & Systems, 2019.

[11] https://ec.europa.eu/digital-single-market/en/high-performance-computing

[12] Nikola Rajovic et al., Supercomputing with commodity CPUs: Are mobile SoCs ready for HPC?, Supercomputing, 2013.

[13] D. Reinhardt and M. Kucera, Domain Controlled Architecture - A New Approach for Large Scale Software Integrated Automotive Systems, 2013.

[14] Nicholas Mc Guire, Approaching the certification of complex systems, 2020.

[15] MW Maier. Architecting Principles for Systems-of-Systems. Syst Eng. 1998.

[16] ISO/IEC/IEEE. ISO 15288: Systems and software engineering — System life cycle Processes. 2015th ed. Vol. 15288:2015. Geneva: ISO, IEC, IEEE; 2015.

[17] MJDC Henshaw. Systems of Systems, Cyber-Physical Systems, The Internet-of-Things…Whatever Next? INSIGHT. 2016.

[18] J. Dahmann 1.4.3 System of Systems Pain Points. INCOSE Int Symp. 2014.

[19] S. Braun, Requirements on Evolution Management of Product Lines in Automation Engineering, IFAC Proceedings Volume 45, 2012.

[20] M. M. Herterich, The Impact of Cyber-physical Systems on Industrial Services in Manufacturing, Procedia CIRP Volume 30, 2015.

[21] P. Larsen et al., A Cloud-based Collaboration Platform for Model-based Design of Cyber-Physical Systems, In Proceedings of the 10th International Conference on Simulation and Modeling Methodologies, Technologies and Applications: SIMULTECH, 2020.

Charles Robinson is Research Engineer in Critical Embedded Systems at Thales Research & Technology, France.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: C. Robinson et al. Bridging the stakeholder communities that produce cyber-physical systems. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 20-29/, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

Cyber-physical systems (CPS) are products that involve diverse technologies, expertise and stakeholders. The clear identification of these essential contributions and their aggregation for complete CPS is paramount, supporting the vision for funding, impact and guidance on future multi-dimensional challenges.

Cyber-physical systems from the application perspective

By CHARLES ROBINSON

Physically active and collaborating systems (future CPS) will play an ever-increasing role in supporting society. They already represent a significant proportion of national infrastructure, occurring in many domains such as manufacturing, transport, medicine and robotics. Defining their key characteristics is essential for having both a clear public understanding of their importance and solid ground on which the many stakeholders, including researchers, engineers and policy makers, can come together and respond effectively to changing societal and market demands. Thus, classification of these types of systems also enables a clear foundation, collecting the various contributing communities, including CPS-specific technologies, around common goals. CPS engineering is concerned with the assembly of technologies to provide complex services requiring physical action and collaboration. So challenges close to the product side have a significant role. They relate to the processes and technologies that facilitate bringing the parts together and therefore are key for supporting technology transfer. Furthermore, classification provides a means to develop clearer measures on research advancement and the industrial pulse. This perspective is supported by feedback from engaging over a hundred specialists in five workshops and two conferences in the Platforms4CPS and HiPEAC projects.

Key insights

- CPS is an application domain, like air traffic management, to which technology domains, including enablers like IoT, are applied. CPS research should relate to both the contributing technologies and to their aggregation for creating future products.
- The CPS classification section should be used as a baseline for describing these systems to stakeholders, especially non-experts. This is an initial step for aligning involved communities.
- Future CPS development involves multiple disciplines and stakeholders to create an operational system, having scope across all domains where systems have physical interactions and collaboration. Solutions to multi-dimensional challenges are required.
- Future CPS should be considered to exhibit subsumption behaviour, merging and separating as teams in response to circumstances, for example, collaborating on search and rescue activities.

Key recommendations

- CPS research requires recognition of the duality between the technology and application domains.
- The CPS classification section should be used as a baseline for describing these systems to stakeholders.
- Managing multi-dimensional community challenges is required for steering progress towards future CPS needs. This includes focus on the aggregation of the contributing technologies and influencers.

Evolving to future cyber-physical systems

Cyber-physical systems have their roots in safety-critical automation and embedded systems, that is, devices with the capacity to monitor and affect their surrounding environment. These devices are used all around us: domestic kettles and cookers can adjust automatically, devices including pacemakers save lives and autonomous vehicles deliver post, while others are present in transportation such as trains or in manufacturing for product assembly. Embedded systems have radically increased wealth and standards of living for people across the world. Of particular consequence have been the advances in the production and transportation domains where automation has greatly facilitated the exchange of goods, enabling significant societal advances. Advances in infrastructure, particularly for commerce, are directly linked to enabling civilizations to develop, especially through increased availability of time, expanding their skills in engineering, the sciences and the arts.

CPS will be capable of physical interaction and collaboration and this represents a significant evolution from embedded systems. This collaboration implies in particular the ability to achieve goals that are unattainable for individual systems. Current examples include satellite navigation in our cars, which requires access to three satellites to operate correctly, robotic-assisted surgery, which requires feedback to the surgeon, and air traffic management. Future CPS will aid rescue teams using autonomous vehicles that need to be combined to tackle large incidents, as well as support cooperating autonomous cars, smart hospitals and robotic assistance that enables home care. This support continues up to the level of smart cities, by providing coordinated collaboration of the many interacting services, aiming to improve safety and support the many needs of citizens.

However, the realization of such decomposable and decentralized systems, with multiple levels of supervisory control required and the critical need to ensure safety, is limited by current research and development approaches. Where embedded systems have addressed design complexity by splitting the problem (separation of concerns), it has led to siloed expertise, limiting the complexity that we can manage. Bridging approaches and tools are required between the silos in order for future CPS to be developed to their full potential.

© VanderWolfImages - dreamstime.com

Furthermore, the great advances supported by mechanical automation are endangering the very planet that supports human life and provides other long-term wealth. By way of example, automation requires energy and, in Europe, transport accounts for around 30% and industry for 25% of consumption. Of this energy consumed, 70% is produced from fossil fuels (according to the EC\EEA). CPS inherits such challenges and will need to turn them around; it is of note that application-side CPS research provides the opportunity for addressing the challenges. *Solving the challenges linked to energy consumption and transition will require evolved and new methods, tools, processes and policies. In particular, aggregating technologies that manage how these elements are combined, will be key to future CPS application research.*

The aggregating technologies determine the options available to use new component-based technologies, so CPS application-side research represents an important avenue for investigation. However, we need to stabilize and master the foundations with the essence of CPS while advancing the much-needed aggregating technologies, particularly those that are cross-domain and interdisciplinary.

The future cyber-physical system research domain represents a duality between dedicated *CPS technology and CPS as an application*, the latter considering the unification of technologies to create this class of products. These two are intertwined because the application domain represents the destination or gateway that enables technology to be used. That is to say, the application domain research is about the bottlenecks for technology uptake on the product side, including the approach and influences on product life-cycle, and how the system components are brought together supported by several layers of aggregating technologies. The combined research of the two perspectives has the capacity to be life changing, with resulting technologies opening up completely new levels of personal freedom and new high-value professions including engineering and science.

Classifying future cyber-physical systems

The creation of future cyber-physical systems (CPS) will involve many communities. This means there are many perspectives on a CPS, including those of specialist engineers, sales people, users and regulators. It is therefore challenging to create

a sufficiently clear definition. Nevertheless, such a definition is critical for directing research and funding to where most impact is needed, particularly for common challenges across domains such as transport, manufacturing and medicine. This is especially the case for Europe's quest for a resilient and sustainable future where CPS are directly involved, representing a large part of society's technology support infrastructure. We present here a distillation of many iterations considered in recent years by communities working in this area.

The most *distinguishing features for CPS are their physical interactions and collaboration*. CPS are products ranging in scale and diversity, from cooperating wheels of a car, to nanobots, railway networks, satellite constellations, smart hospitals, autonomous car fleets, up to higher levels such as smart cities in which automation improves quality of life while improving resource usage.

These systems have six key physical characteristics that can be used for identification and which they cannot do without: *sensing, communications, physical action, processing, the provision of energy and coordinated collaboration*, as shown in Figure 1. Classifying these physical characteristics is important not only for guiding research communities, but also in terms of understanding high-level progress across industry and for the public to be able to relate to and understand the impact of CPS.

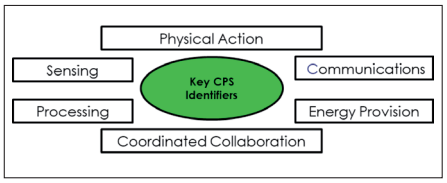


Figure 1: The six key characteristics to identify a CPS.

Future CPS will function like teams, where subsumption occurs with CPS merging and separating. *They become one CPS*, not because everything is managed centrally but because they are collaborating as a team sharing some higher levels of supervision. For instance, in terms of the supervision of safety, if an accident occurs the law searches for one entity to take responsibility. So CPS become one team of systems (likely also constituting other systems) because they share a common framework enabling them to collaborate, converting cyber plans to real-world outcomes through physical actions [1].

We of course need to manage the interactions between the six characteristics. This interaction may be between specific characteristics but is particularly important for their overall control. This control also has to consider system properties that affect all characteristics. In particular, given that these systems may involve movement and interact in our world, they must respect time and ensure safety and security. This already means that there are external stakeholders in terms of regulation, policies and standards that must be taken into account before considering the requirements of a customer. Since CPS are products, other stakeholders include those responsible for the product lifecycle: from its development and testing, through operation and maintenance to product retirement. This includes the ability to respond to and to impact markets and evolve product processes to support new technologies. There are evidently many technical and non-technical communities involved in creating a CPS, as highlighted by Figure 2.

CPS is an application domain like a rail network, which has many technologies and influences combined to create the final product. That is to say, in this context a CPS represents the common goal for which the technologies are combined and it is a destination for technology transfer. In Figure 2, technology provides the methodologies, hardware and software at each of the levels shown. As one moves up these levels, linking functions and system properties, the technologies are faced with increasing scope to manage and stakeholders to take into account. This means one is moving from technologies treating a single problem → single solution to group problems → group solutions. As a result, the related assembling and cementing technologies require different and longer lifecycles to mature, but are essential for advancing cyber-physical systems and the uptake of new component technologies. Technology domains such as Internet of Things (IoT), Systems of Systems (SoS), Artificial Intelligence (AI) and Big Data will also have increasingly important roles to play in the future of CPS as a destination of technologies. In particular, frameworks for safe interactions between systems and with people will call upon these enabling technologies; the application of SoS represents a particular key stepping stone and recent research has abbreviated this as CPSoS. Note that, as future CPS applications are expected to contain SoS behaviours and

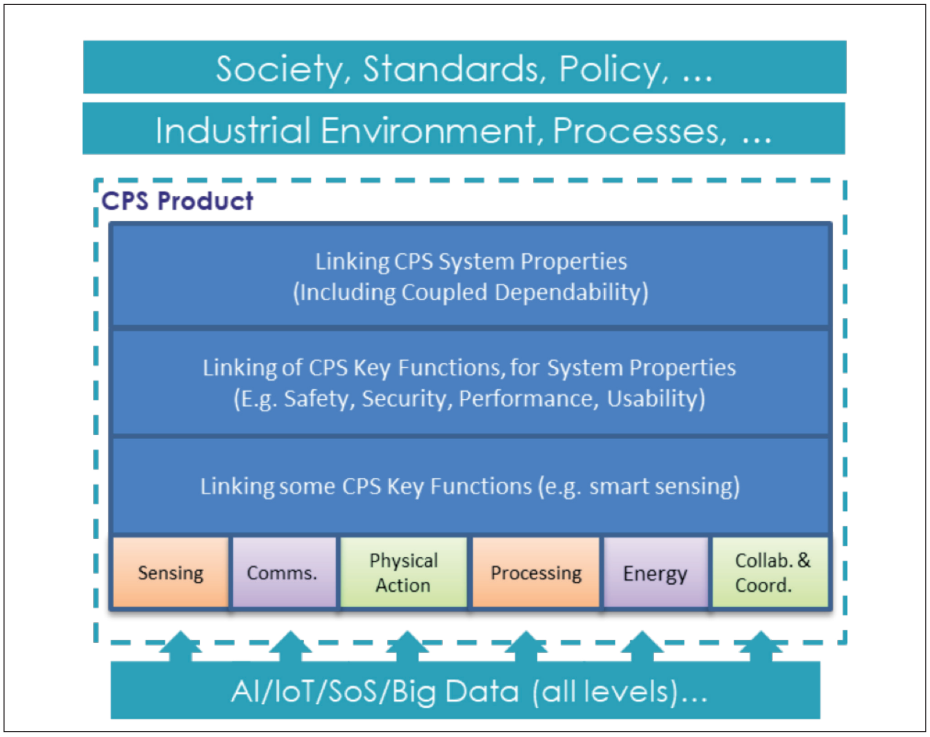


Figure 2: Contributors and influencers to aggregate in order to realise a cyber-physical system.

that interacting CPS merge into one CPS (team) for safety [2], CPSoS might evolve to something like SoS4CPS.

The definition of future CPS products presented here has been broken up into three parts: there are the two distinguishing features (physically interactive and collaborating systems) and the six key functions (sensing, communication, physical action, processing, energy and coordinated collaboration). Finally, we propose a model to be used, which includes the many stakeholder communities required for realising a complete cyber-physical system.

Significance of a clear CPS definition for application-side research

Being able to classify and link the contributing elements for creating these future physically interactive and collaborative systems is essential for many reasons. From the public point of view, the ability to identify such systems and their advantages is important, particularly because existing benefits are indirect, such as those delivered via services. Market push will be initially needed, as current purchases of CPS are usually business to business or made by governments for improving national infrastructure (e.g. transport and manufacturing capability). In the longer term, the consumer will be more in the loop, for instance owning components of a CPS (such as an autonomous car) or having miniaturized systems like nanobots interacting and customising features of the home.

A clear definition provides firm points of reference where a cohesive framework can be put in place, *essential for understanding the cumulative effects of successive aggregations of technologies required to create a CPS*. Here we use aggregation to mean the approaches and tools to bring the parts together. Traceability across these aggregations provides significant support for a global view, which can benefit many aspects of development and research for CPS.

From a funding point of view, application-side research will improve techniques for measuring the impact of investments on CPS and aggregating technologies, or other complex technology combinations. A common characterization means the market sectors that will benefit can be clearly identified. While higher-level aggregations often take longer to exploit because there are more stakeholders involved, the availability of metrics improves. In addition, more specific feedback from related businesses can be provided, for instance from their interaction with safety certification authorities or use of specific tool combinations. This is also relevant for tracing the evolution of product process methodologies, where direct feedback of new approaches applied can be difficult to measure [3], often including intangible factors, such as improved knowledge generation and transfer between contributing developers. Therefore, such a framework provides improved means for keeping a finger on the industry pulse.

Similar benefits hold true for CPS research needs and contributor support, with significantly more visibility of bottlenecks along the chains of aggregations. The overall response to top-level research challenges and impact from many levels of contributions can be traced through the aggregations. Achieving cooperating systems, for instance, will need supervisory control layers over the combined systems to ensure safety and also other mechanisms for decomposable teams. Other top-level technical and societal challenges include sustainability, resilience and the circular economy. Traceability through extensive aggregations is especially relevant when it comes to CPS representing smart cities, where clear guidance on impactful research is important, given the scope of potential research paths for higher-level aggregations.

With respect to the technology contributors, clear characterization and traceable aggregations means more informed and prioritized development can take place, such as in response to higher-level aggregating technologies presenting barriers to new technologies, or even for aggregations at the same level, such as between security and safety [4]. As an example, an existing certification technique used in companies may need to be evolved so that some new technologies for physical action can be applied. As CPS has a cross-domain focus (e.g. transport and manufacturing), it also provides the ability to identify and prioritise work on cross-domain barriers, with the capacity to open up new markets for contributors to CPS.

Furthermore, because of the aggregation nature of CPS application domain research, focus must be placed on tools and techniques not only for multi-disciplinary operations of a CPS, but also for supporting multi-specialist interaction in research, including translation and mediation between perspectives. This will be particularly important for bridging the siloes between disciplines, something that will be required to achieve the expectations for future CPS. A clear definition and work on aggregating knowledge and tools between stakeholders are consistent with many of the other recommendations of the current HiPEAC Vision. They also support the recommendations from Platforms4CPS [5], including the ability to couple system properties like safety and security, CPS with explainable actions for trust, improved collaboration and defragmentation of communities and maintaining our leadership over competition on the world stage while taking into account societal impacts.

References

[1] M. Törngren et al, Oct. 2018, "Platforms4CPS Report: Collaboration on the Foundations of CPS Engineering".

[2] "HiPEAC Report (forthcoming): D3.3 Best Practices and Tools for CPSoS", (Workshop held Sept. 2020).

[3] A. Thum-Thysen et al, "Unlocking Investment in Intangible Assets", Discussion Paper 47, May 2017.

[4] C. Robinson et al, "MERgE: Technology advancement for cohesion of concerns in system engineering", Proceedings of the 1st Workshop on security and dependability of critical embedded real-time systems (CERTS2016), 2016.

[5] Haydn Thompson et al, "Platforms4CPSKey Report: Outcomes and Recommendations", Oct. 2018.

Charles Robinson is Research Engineer at Thales Research & Technology, France.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: C. Robinson. Cyber-physical systems from the application perspective. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 30-33, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

It requires models to make cyber-physical systems interpret and interact with the physical world. But constructing these models is getting harder and harder as system complexity grows.

The model inside

By HARM MUNK

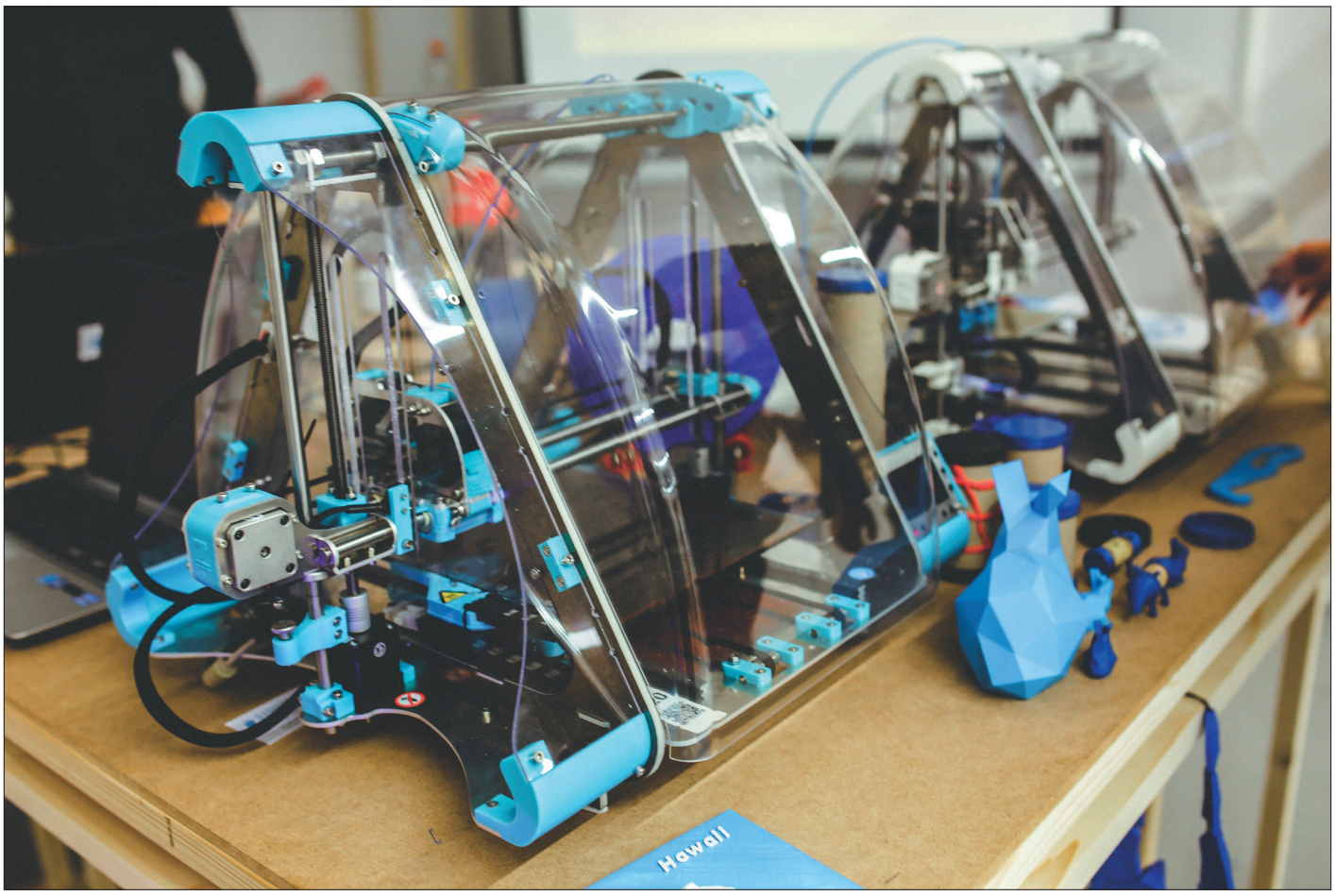
Interactions with the physical world set cyber-physical systems apart from traditional computer applications. These CPS sense their environment through a variety of increasingly complex means, such as radar and cameras. Sensing is the first step in a pipeline that involves interpreting these measurements. That interpretation requires an understanding to some extent, a model of the environment, of the physical world. Designing and implementing such models for increasingly complicated CPS is a never-ending challenge that involves many disciplines.

Key insights

- Models play a key role in everyday life, even if we do not realize they exist.
- The physical model implemented in a cyber-physical system (CPS) is an essential factor in determining its success or usefulness.
- A physical model is an abstraction of the physical world.
- Finding the right abstractions is hard, and even harder as CPS grow in complexity.
- The notion of (real) time is crucial in physical models for CPS.

Key recommendations

- Digital systems used in the implementation of CPS must be time deterministic to make CPS behaviour predictable.
- Future programming languages and systems must include a model of time.



We humans have been modelling the physical world ever since we started thinking. We have been constructing mental models to make the world around us comprehensible and predictable. We must do this because we interact with the physical world and need to create mental order in the chaos that surrounds us. Sometimes we have recorded parts of these models in the form of paintings, small statues or large monuments. Take Stonehenge as an example. We do not know why it was constructed, but its construction is partly based on a model that describes the movements of the sun in the sky. The total model is probably much richer than that, but that knowledge has been lost in the mists of time.

Greek mythology attempts to explain many things in the world surrounding us, things we can check to see if they are correct or wrong. But Greek mythology also contains a description of an invisible world. A world inhabited by gods, inaccessible to humans: things we cannot check.

With the dawn of science, something changed in the way we, or better, scientists dealt with models. Just constructing the model was not sufficient. All aspects of the model had to be verifiable. That made Greek mythology into an interesting but unscientific model of the world. Newton's laws of mechanics, formulated in the late 17th century, although not really laws, are mathematical models of forces between masses. And those laws (let's call them laws instead of models, because that's how we know them) immediately explained many things that mystified scientists until then. The consequences of these laws were verifiable. Even better, these laws were able to predict phenomena, such as the movements of the planets and comets.

However, some things could not be explained with Newton's laws. As a notable example, they did not correctly explain the motion of Mercury around the sun. It turned out that Newton's laws were flawed in a fundamental way. It took another scientist, Einstein, to introduce a new model, general relativity, to set that right. But even though Newton's laws were wrong, they are more than sufficient to explain almost every aspect of the world we interact with in our daily lives. To put it to an extreme, Newton's laws suffice to get us to the moon and back, but it requires Einstein's general relativity to make accurate navigation systems (GPS, Galileo). Ignoring that theory in the GPS model would introduce ever-growing positioning errors. We can sum it up that different models are used to explain the same phenomena, but with different modelling accuracy.

Models of the physical world evolved, both in character (from myths to verifiable laws), and in the amount of detail they are able to describe. That last aspect of change plays an important role in cyber-physical systems.

Physical models

Models help us to understand the surrounding world by describing its relevant aspects. For example, Newton's theory of gravity describes the attracting force between two point masses. Point masses don't really exist, but for the sake of the computation we can imagine all the mass of a body concentrated in a mathematical point. The model abstracts away from the real world by leaving out details that have negligible or no influence on the problem.

Another example of a model is provided by digital electronics. In digital electronics, transistors, which in themselves are analogue components, are described as simple switches, either blocking or conducting electric currents. In reality, an open digital electronic switch still conducts a small amount of current. And a closed switch can only conduct a limited amount of current. Apart from being able to reason about the logical function of an electronic circuit, those details do not matter. Only when the function is implemented in an actual circuit do these details become important again. In fact, we then extend the logical model into a more elaborate electronic model. But for understanding the working of circuits as a digital system, the abstraction of the transistor as a switch is all that is needed.

The use of models is widespread. Modern programming languages are based on a model of computation. That model is so much ingrained into the mind of the software engineer that most of them are not aware of it. These, for modern programming languages, often quite complicated models are partly described in the semantics of the language. How a program written in C++ is mapped on processor instructions is a level of detail that the programmer does not have to know at all to be able to write programs. The programming model is an abstraction of the implementation.

Even assembly language programmers work with models, and not only with the model of the assembly language. They also deal with a model of the underlying hardware, with a model that describes how the instructions change the state of the system. Those state changes are an intricate interplay of digital electronics but, again, something they do not have (to have) in mind when writing a piece of code. It is, again, an abstraction, this time of the digital circuit that implements the processor, where the underlying model is described in the ISA.

We need models because we need to be able to concentrate on the relevant aspects of the problem we are dealing with. That in itself is a requirement for humans to be able to solve complicated problems. Our brains are limited in the amount of detail we can oversee at once, otherwise we get lost in them, literally.

Computers, models and time

The first practical digital computers were constructed in the 1940s. Some were what we now call application specific processors, hardly recognisable as a computer. An example of these machines is Colossus, constructed especially to solve a subproblem in the deciphering of encrypted messages during the Second World War. But soon, electronic systems were developed, inspired by these early machines that were truly general-purpose computers.

Early computers were stand-alone machines. Program and data entry and storage were local, through paper tape, teletypes, and magnetic drum and tape. Early machines were developed often for military purposes: ENIAC, completed in late 1945, was designed and constructed to compute artillery firing tables. Soon, computers were also used for the design of

airplanes and other machines and processes that required extensive computations. But computers were mostly unconnected.

The programs these computers executed contained models of physical processes, in the form of differential equations. The programs computed an approximate numerical solution for these equations and printed the results. There was no interaction with the physical world, other than interpretation by humans of these numerical solutions and using those interpretations to construct a machine or process. Note that the notion of time was not considered a primary characteristic: it may have taken the computer hours to accurately solve a problem for a process that in reality took just a millisecond.

For weather forecasts, it was, of course, the other way around: it took an early computer a few hours to compute the next day's weather. But here speed came with a cost: the physical models of the Earth's atmosphere were a coarse approximation, resulting in a weather forecast for just a day ahead.

Between the 1950s and 1980s this slowly changed, as it was recognized that computers could be used, for example, to control processes, which were until then mostly controlled by mechanical systems.

That meant that computers had to be hooked up to the physical world process. Sensors acquired the value of physical quantities such as pressure, temperature and distance. Actuators manipulated the physical world through valves and electromotors.

The advantage of digital computers over mechanical process control is their flexibility: it is relatively cheap to change the control algorithm in the computer – just change the program. Yet it also required accurate models of physical systems, and, very importantly, the notion of time. Think of a flight control system: it must finish its computations ahead of time to be able to fly the airplane safely. And although some high technology systems are not as time-critical as an airplane, time plays an important role for quality or economic reasons, for example. Take the example of an expensive professional printer used in an on-demand print shop, where books are printed after customers place orders. Having the printer print as many books as possible increases the return on investment on that printer. As another example, take a lithographic scanner for IC production, in which patterns for transistors and wiring are projected on silicon wafers, one chip at a time. Each exposure heats up the wafer locally, which ever so slightly deforms the wafer. It is a tiny deformation of just a few nanometres, but then the patterns are also in the order of nanometres. It is now possible to calculate and compensate for those deformations, but (cooling) time must be taken into account. It is possible to do the calculations for that but, in a scanner, it must be done in real time, because machine time is precious. Thus, you need a model of the wafer, how it is heated during an exposure, and how it conducts the heat and deforms.

In this respect, it should be noted that most programming languages do not have an abstraction of time. You can indeed model time in FORTRAN but the semantics of that notion of time is not part of FORTRAN, it is part of the semantics of the program. And it often requires interaction with the operating system, or even with the underlying hardware to keep time in a program.

Keeping time with some accuracy is not at all easy in modern processors. The

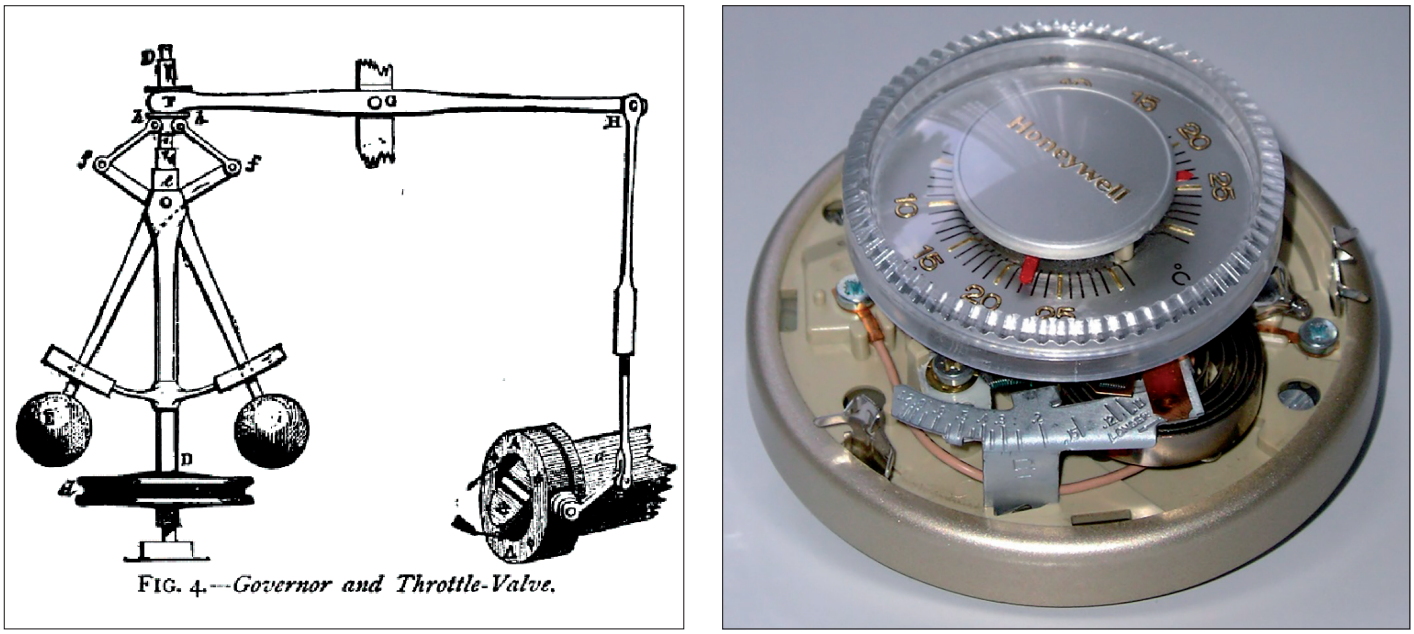
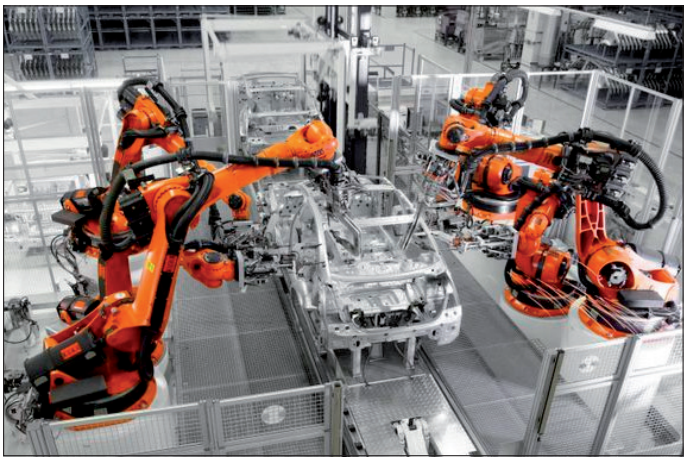


Figure 1: Left: drawing of the Watt governor, or steam regulator, an early example of a control system that was used to stabilise the speed of a steam engine, developed in the late 18th century. Right: an electromechanical thermostat that could be found in many homes to regulate the ambient temperature, dating from the 1970s. It is the precursor of the modern digital thermostat.
(Sources: left picture: https://upload.wikimedia.org/wikipedia/commons/1/1e/Centrifugal_governor.png - right picture: https://upload.wikimedia.org/wikipedia/commons/1/1e/Honeywell_thermostat_open.jpg)



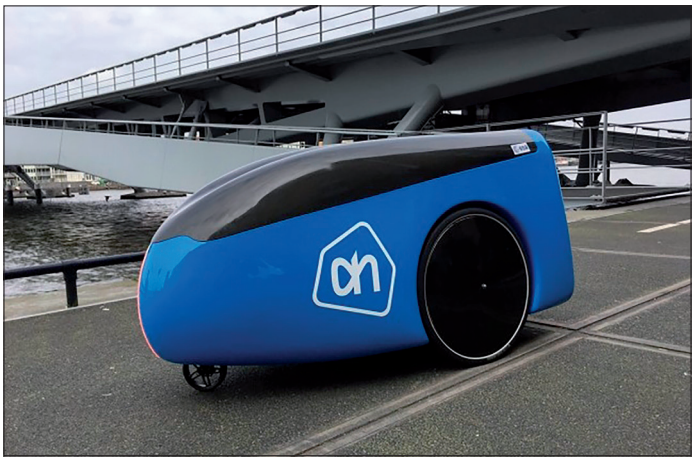


Figure 2: Two pictures illustrating the changeover from "caged" CPS to "CPS" in the free. Industrial robots do not have to deal with uncontrolled situations such as humans getting in the way, at they are locked in, protecting the environment from their actions. On the other hand, self-driving cars such as this one from Uber must be able to deal with many, often complex situations.
(Sources: https://www.kuka.com/en-de/industries/automotive, https://api.time.com/wp-content/uploads/2014/08/140818_inv_futuregoogle_car.jpg?w=800&quality=85)

semantics of the ISA is of course well-defined, but the timing of instructions, because of several levels of caches and instruction reordering in the processor is not.

So, the notion of time is absent from many programming languages (certainly in the case of the programming language C, which is the most widely used language for CPS) and, in addition to that, the timing of the underlying hardware is not deterministic. Something needs to be done!

There are some developments seeking to solve the time issues, both in hardware and in software. These developments [2] are still in the research stage, and a lot of work is needed to bring them to engineering practice for the construction of CPS.

CPS

Another important shift started back in 1969: computers started to be connected through the internet. Network technology took off in subsequent years, which reduced the costs of networking hardware. If you wanted to buy a computer with a network connection in 1990, you had to buy a separate network card. Try buying a new laptop today that is not Wi-Fi enabled and you will have a hard time finding one.

The enormous growth in microprocessor capabilities in term of functionality per mm^2 meant that process control could be distributed cheaply, with smart sensors transmitting (huge volumes of) data to be processed elsewhere, resulting in commands for intelligent actuators somewhere else. These are the essential characteristics of a cyber-physical system: sensors, processors and actuators, all connected through a network fabric.

A modern car contains tens of microprocessors for engine management, break control, road condition monitoring, navigation, entertainment, etc., relying on sensors and actuators distributed throughout the car. If you push on the accelerator, the position is translated through a sensors-processor-actuators chain into changes in petrol injection in the engine, also taking into account the condition of the engine. There is no direct mechanical connection anymore.

Note another important change: until recently, we confined such CPS to a closed, controlled environment. We used robots in factories and inside the factories, the robots were often in cages. That kept the responsibility for security out of the hands of the robots which, in turn, allowed the focus to be on the function the robot had to achieve, not being bothered by unpredictable humans that might get in the way. But such robots can't be let loose.

Today, we are testing the current pinnacle of CPS: self-driving cars. These cars have to "take responsibility" for keeping safe the environment they are operating in. What that means is that they have to have a model of the physical world surrounding them in order to achieve that goal.

We can expect to see CPS become ever more pervasive in our daily lives. Robots will deliver the groceries, will help in doing the daily household chores, etc.

In order to keep the world we live in as safe as possible, the CPS needs to have inside them a model of the physical world so that they can operate safely in that world. To complicate things, that world is dynamic, it changes, and on several timescales. The mix of cars, pedestrians, road signs, etc. that an autonomous car has to deal with in real time does not change quickly. But particular road layouts can change overnight, road works spring up unexpectedly. And on a timescale of years traffic evolves: rules change, new road sign are introduced, new vehicles are introduced. The best way to do that depends on the architecture of the CPS. It might contain a flexible model that is capable of dealing with some change. Or existing models for the CPS might be upgraded and then replace the old models once they are tested and validated. The point here is that it is difficult to design a model for the things we do not yet know. Note that a model based on deep learning needs to be retrained. If, for example, a new type of

vehicle is introduced, then without retraining the model might or might not recognize this new vehicle for what it is: there is no guarantee.

Constructing the model

Developing models for CPS is a challenge for a number of reasons. For one, it involves several disciplines [4], of which software engineering is only one. As CPS sense and react to the physical world, physics is certainly involved. But it depends on the application domain and on the required accuracy of the model which branch of physics is required. The autopilot of an airliner of 40 years ago kept the plane on course at a particular altitude. The model for that kind of CPS is far less complicated than the fly-by-wire controls of almost all modern passenger airliners.

There are several approaches possible to be taken to construct the models. If the physics behind the model is well-known and understood, then starting from first principles, from the physical laws known to describe the CPS and its environment, can be the most logical starting point. This approach can in principle result in accurate models, as accurate as the underlying physical laws.

However, it may also occur that the physics involved is too complicated to be developed into a practical implementation of a model. In those cases, an efficient, approximate model can be constructed. If such an approximate model is not even feasible, it may be possible to construct a statistical model based on data from experiments. Take, for example, an airplane wing: although all laws involved in computing lift and drag of an airplane wing are known, it is impossible to derive these numbers in analytical form for every type of wing. That is why models of airplanes are tested in wind tunnels: to derive experimental models of the airplane that form the basis for e.g. the fly-by-wire system for the control of the plane.

Experimentally derived models can be made as accurate as models derived from first principles. But there is a fundamental difference in the application of these models. In essence, a model derived from first principles is accurate over the span of the physical laws underlying it, even if that involves conditions that, for all practical purposes, the CPS will never be subjected to. That cannot be said of experimentally derived models: such models are only valid within the boundaries of the experiments that were carried out to derive those models. Using experimental models outside those boundaries is potentially dangerous for the CPS.

It is important to note that approaches based on neural networks are not the same as physical models. A physical model, based on physical laws, is generally applicable, in all situations that fall within the boundaries of those laws. That is a potentially infinite set of situations. Neural networks, on the other hand, are trained with a finite set of examples. If a neural net is presented with inputs very different from what it has been trained for, in general it will give the wrong result. Even worse: it has been shown that it is possible to fool a neural network to give the wrong answers, by mimicking expected input in an unexpected form. Students at the MIT Computer Science & Artificial Intelligence Lab demonstrated this well [1].

Note that a model and its implementation are two separate notions. A model of a physical process in the form of a differential equation will have to be implemented by an algorithm of a numerical solver for that equation. Such a solver can itself model the differential equation with limited accuracy for several reasons. It will by its very nature only approximate the solution of the differential equation. But as it is implemented on a finite piece of hardware, the accuracy of the representation of the numerical values is also finite and will only be an approximation of the real values. This aspect requires careful analysis to ensure that the implementation of the model is sufficiently accurate for the purpose of the CPS.

Digital twins

There are key differences between a digital twin and a CPS. Digital twins are engineered to be a replica of a physical system. Depending on the purpose of the digital twin, it can be an isolated system, not interacting with its environment at all. Such digital twins are for instance fed with pre-recorded input traces derived from the physical counterpart. A CPS, on the other hand, is a system that interacts with its surroundings through sensors and actuators.

Therefore, a digital twin is a *replica* of a CPS. It might contain some or all of the model that is also in the CPS. But more importantly, it contains a model of the CPS itself.

There are many roles for digital twins at various stages in the lifecycle of a CPS, starting right from the development of the CPS, where a digital twin might be used for subsystem development. The physical part of the CPS can be modelled and then used to develop the control software for the system. A digital twin can also be used during the operational phase of the system to diagnose performance errors or system failures when, for example, the actual CPS is unavailable (e.g. a spacecraft after launch). A flight simulator, or, for that matter, any vehicle simulator, is a perfect example of a digital twin: it is used to train pilots and drivers, but can also be used to verify and validate new control setups or algorithms. A digital twin can also be used to predict the reaction of the environment in which the CPS evolves, allowing it to take the right way of options (leading to *Predictive CPS*).

Developing models for CPS

The model, on which the implementation of a CPS is based, is the ingredient that determines how and how well the CPS interacts with its environment. When developing such models, it is not only the requirements of the CPS that must be considered; a thorough understanding of the environment, of that part of the physical world the CPS will operate in, is also essential [3]. What the developed model looks like is thus determined not only by the physical environment of the CPS, but also by the role of the model in the CPS.

It is virtually impossible to design and implement a complete model of the environment. It would also be a pointless exercise, as there will be many aspects of the physical world that do not play a role in the operation of a particular CPS. When

dealing with road conditions, a self-driving car does not need a climate model; it needs sufficient information to determine the state of the road surface.

So, it is essential to make the right abstractions for the physical models for a CPS, to understand what should be incorporated in the model, and what should not be and the reasons why not. As all physical models leave aspects out, that brings us to a very important observation: all physical models are wrong, but some are useful. And that is the crux of the matter: what should be included and what can be left out, is hard to determine, and to find out. It is a multi-disciplinary challenge because it requires an understanding of the physical world (the what), of the needs of the CPS (the why), and also of the way such models are to be implemented (the how).

Constructing useful models is not so much a complicated task ("we know what to do, it's just a matter of hard work"), it is a complex task ("we have to find out what we need to do"), and becomes more complex as the CPS we construct get increasingly complex themselves. It is a continuous balancing act. Systems developers and architects are often inspired by new technology and are eager to apply such developments. They have to be careful not to get distracted by the prospects of new technology. They must fully assess and understand if that new technology is applicable to the situations they are trying to model.

For now, it is humans doing the modelling. Computers are taking over mundane tasks from humans, even to the point of programming, yet they are not yet modelling the physical world.

What is more, as it is a multidisciplinary, multi-person activity, modelling requires not only knowledge of technology and physics, but also careful reflection on the modelling process itself.

References

[1] https://www.csail.mit.edu/news/fooling-googles-image-recognition-ai-1000x-faster

[2] Edward A. Lee, The Past, Present and Future of Cyber-Physical Systems: A Focus on Models, Sensors 2015, 15, 4837-4869; doi: 10.3390/s150304837

[3] Patricia Derler, Edward A. Lee, Alberto Sangiovanni Vincentelli, Modeling Cyber-Systems, Proceedings of the IEEE 100(1):13-28, DOI: 10.1109/JPROC.2011.2160929

[4] Dmitry Morozov, Mario Lezoche, Hervé Panetto, Multi-paradigm modelling of Cyber-Physical Systems, IFAC PapersOnLine 51-11 (2018) 1385-1390, DOI: 10.1016/j.ifacol.2018.08.334

Harm Munk is Senior Project Manager at TNO, Eindhoven, The Netherlands.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: H. Munk. The model inside. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 34-39, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

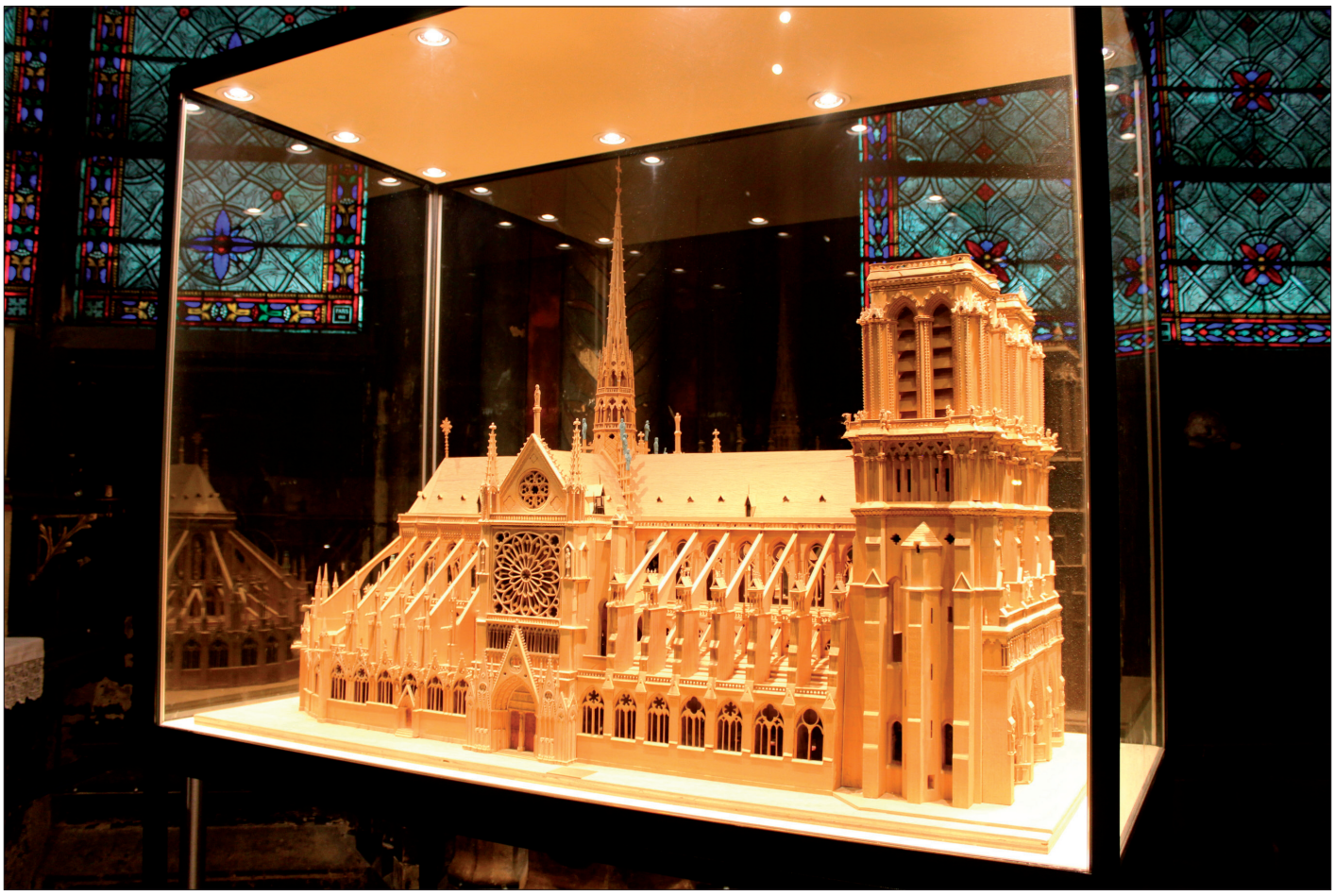


Image: ID 194111920 ©Subash Rajesh Kannan Dreamstime.com

The TransContinuum Initiative: eliminating the silos in order to achieve a better orchestration of complex applications

The continuum of computing

By MARC DURANTON, MICHAEL MALMS and MARCIN OSTASZ

As explained in the 2017 HiPEAC Vision [1], ICT is expanding from cyberspace and it is now interacting with us directly, e.g. in self-driving cars, driverless underground and overground trains, factories and cities. In this new paradigm, applications are not built around a single piece of code running on a single machine, but span across various computing resources distributed in multiple locations. The notion of a continuum of computing has emerged, which denotes a set of computing resources and software (the smartphone, the communication infrastructure, the cloud) acting together in order to complete a complex task involving multiple systems and multiple technologies. This concept has been proposed by HiPEAC and promoted in its Vision roadmaps for the last five years. This has led to the creation of the "TransContinuum Initiative", which involves various European organisations including ETP4HPC, ECSO, BDVA, 5GIA, EU MATHS IN, CLAIRE, AIOTI and HiPEAC with the objective to promote these concepts further through concrete actions.

Key insights

- Distributed systems are required to operate applications on various computing resources distributed in different locations.
- Distributed systems need to be developed in such a way that they ensure performance, efficiency, security, reliability and trust.
- The task calls for a number of communities to work together to achieve the orchestration of this complex set of various systems and software (which now is mainly achieved by ad-hoc approaches) and to develop interoperability, both in the protocol between systems and in the data structures.
- The cross-disciplinary "TransContinuum Initiative" (TCI) constitutes an attempt to "break the silos" and achieve better efficiency through horizontal collaborations among various organisations.

Key recommendations

- Cross-disciplinary work to achieve better efficiency through horizontal collaborations among various organisations is required to achieve the success of tomorrow's solutions that will form a "continuum" of computing.
- Software is becoming increasingly distributed across platforms and devices and therefore programming has to be reinvented with languages and tools to orchestrate collaborative distributed and decentralized components, as well as components augmented with interface contracts covering both functional and non-functional properties.
- Interoperability is mandatory, not only in the format of data that need to be recognized by all elements of the continuum, but also interfaces, protocols and API.
- All solutions should support increased software and hardware sustainability, and a high-level of cybersecurity to rely on dependable and resilient components, services and digital infrastructures.
- Algorithmic efficiency will need to be drastically improved (e.g. more efficient AI), which requires development of basic modelling, simulation and optimization methodologies in data-rich environments (MSODE), including model-order reduction. Management and deployment of large-scale application workflows will have to be adapted or invented.
- Building over existing technologies will facilitate acceptance, ease the development and reduce the cost, even if new software layers need to be added on top of existing solutions.

Continuum of computing explained

Despite the fact that there are still a lot of standalone applications, applications in general rely increasingly on other applications to improve their efficiency, while some of them are not even able to work in standalone mode.

For example, the "cyberspace" of **aircraft** is composed of a multiplicity of tasks, from collecting data from sensors, to controlling the engines and ensuring the safety of those on board, and these tasks are done by a multiplicity of computing resources, from the small micro-controller integrated in a sensor to the main on-board computers (OBC). However, planes are also connected with other planes and with the ground, not only for air-traffic control, but also to (indirectly) obtain the results of the supercomputers that are involved in predicting the weather. They are part of a complex orchestration not only of other aircraft, but also of computing hardware and software. The trajectory of the aircraft is predicted even before it takes off in order to ensure that it will be alone in its space and will have a slot for landing.

In the **factories** of today, machines do not work independently but are connected and orchestrated. Sensors are everywhere (IoT), collecting data from each machine and how they perform their processes. All this data is processed increasingly locally in order to provide meaningful information on how the machine works. All this information is collected and processed increasingly on the fly to maximize efficiency for the entire factory. In parallel, analysis of the data allows predictive maintenance to be performed. In more advanced systems, a complete model of the factory, its digital twin, running on high performance computers and fed by the actual data, allows further enhancement of the global efficiency, and advance prediction of what should be done to control the various actuators in the factory. Techniques using artificial intelligence can also be used alongside numerical simulation to enhance the parts where explicit modelling is not easy. This forms a kind of complete "continuum" that spans from data sources to commands (controlling physical devices) which is depicted in Figure 1.

Even for consumers, **photography** and **sharing photos** through social networks also involve a multiplicity of ICT devices working together: the picture is first preprocessed by a digital signal processor, then is upgraded using AI accelerators to improve it before being stored in the memory of the smartphone. It can also be postprocessed in order to be efficiently sent with its associated metadata, which could include the localization provided by GPS. GPS information is also generated by a set of computing tasks, some of them being done in a satellite in order to send the right signal to the GPS receiver and its associated software in the smartphone. Then the picture, with its corresponding metadata, is sent to baseband stations – also computers- and through several steps enters a data centre (cloud) and then gets distributed over several processors or even data centres to make it appear on a social media feed.

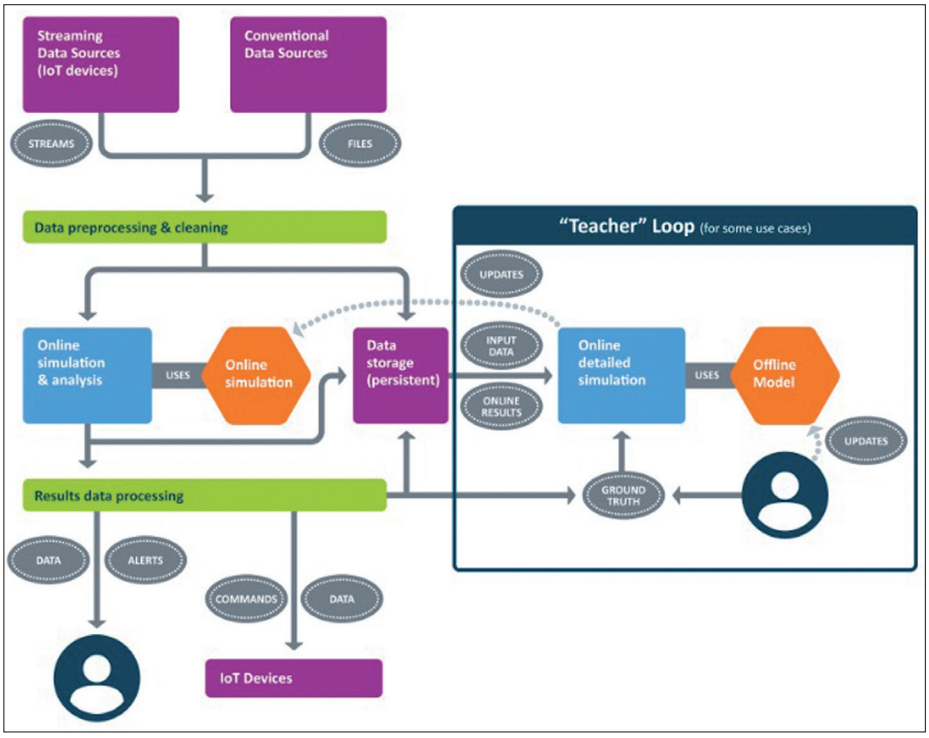


Figure 1: A typical mixed simulation: IoT and edge, machine learning workflow [3]. This is a generic high-level model, i.e. all elements depicted or a subset thereof are present in every complex system.

The examples above illustrate the emergence of the **continuum of computing**, i.e. a set of computing resources and software (the smartphone, the communication infrastructure, the cloud) acting together in order to achieve a complex task of e.g. making the picture you took accessible to your friends.

Even if the cloud is mainly driven by non-European organizations, Europe has the potential to be present in the "edge" and "deep edge" solutions if it doesn't wait too long. As explained in the HiPEAC Vision 2019 [2], there is an alternance of centralized then decentralized tendencies in computing: from 1970s computing centres, to distributed set of PCs, to the cloud where most of processing is done. Nowadays, the pendulum is swinging more towards having "intelligence at the edge" rather than just having all computing done in the cloud. But the window of opportunity for Europe will be small: edge systems are very diverse, and being able to deliver a high diversity of systems, customized to particular applications, quickly put on the market with a high productivity (and therefore low cost) will be key for success. Furthermore, AI could help to improve this productivity. For example, in the factory and plane examples, a lot of processing has to be done locally, for safety, latency or privacy reasons, and cannot rely on communicating all the raw data to "the cloud".

The TransContinuum

The term TransContinuum describes the defining characteristic of the infrastructure

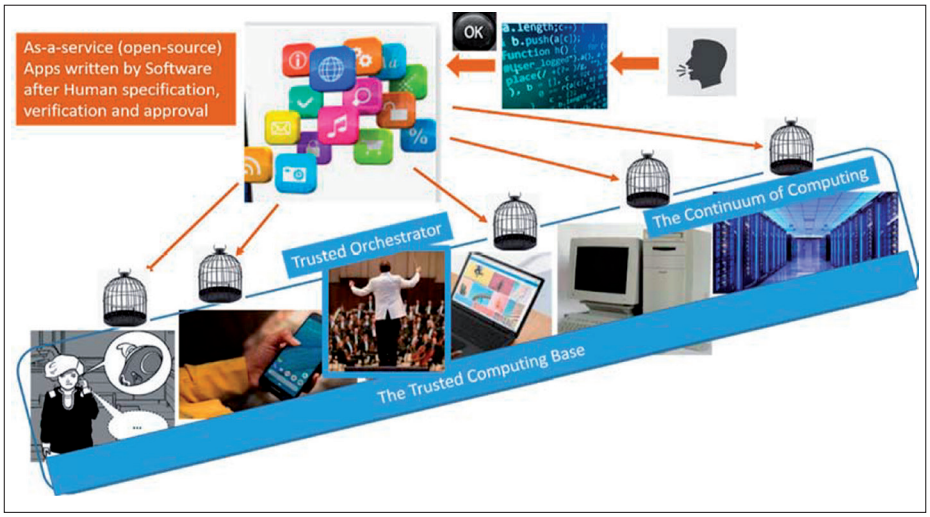


Figure 2: A pictorial view of the Continuum of Computing (see the chapter for more explanations).

required for the convergence of data and compute capabilities in many leading edge industrial and scientific use scenarios. A paradigm change is needed: we will have to design systems encompassing millions of compute devices distributed over scientific instruments, IoT, supercomputers and cloud systems through LAN, WLAN and 5G networks.

In this continuum, edge, fog and cloud computing platforms are being pulled close together into what will likely become a seamless execution environment, as depicted at the bottom part of Figure 2. This change is being driven by a massively increasing amount of value-added applications targeting mobile, handheld, wearable and unattended devices.

We are observing a continuous miniaturization of computing and storage devices, as well as their ubiquitous deployment ranging from data centres to the edge and beyond. There is also the need to enable workflows beyond single control domains like a data centre. Therefore, a new overall system architecture must be designed to accommodate the ecosystem changes (environmental and technological) to be expected in the coming decades and to integrate horizontally the different actors.

The challenges associated with the "continuum of computing"

Software is becoming increasingly distributed, becoming a "continuum of computing" across platforms and devices, from the "deep edge" to the cloud, HPC and data centres. Programming has to be reinvented for this, with languages and tools to orchestrate collaborative distributed and decentralized components, as well as components augmented with interface contracts covering both functional and non-functional properties. Data collected by deep-edge devices is concentrated and processed by edge computing, consolidated by federations of systems (fog computing) or, if required, travels to the cloud or to HPC centres to feed simulations that facilitate decisions on actions to be taken, e.g. for managing fleets of vehicles, complex factories or air traffic, therefore closing the loop.

The new demands and challenges, having an impact on the need to combine data, storage and compute, to have them distributed across the continuum, and ensure lifecycle maintenance and resource efficiencies, are pushing for drastically increased software and hardware sustainability. Furthermore, the need to provide high-level cybersecurity to rely on dependable and resilient components, services and digital infrastructures is changing the game profoundly. Efficiency and resilience will have to reach levels never achieved thus far, while taking into account the intrinsic distributed and heterogeneous nature of the **continuum**. In addition, the question of dealing with very high volumes of data needs to be addressed, and the preponderance of quality versus quantity will become unavoidable. These considerations will be relevant to all components.

Interoperability is mandatory, not only in the format of data that need to be recognized by all elements of the continuum, but also interfaces, protocols and APIs. Today, there is one system that connects billions of heterogeneous devices, from the smallest IoT device to the largest data centre: the internet. We propose therefore that the TransContinuum runs mainly on top of existing technologies, such as those used in the internet, like TCP/IP, RESTful services, etc. Building over existing technologies will also facilitate acceptance, even if new software layers might need to be added on top.

Long-lifetime hardware devices will have to be reconfigurable, modular, exchangeable and self-aware in order to be operational over extended periods. Algorithmic efficiency will need to be drastically improved (e.g. more efficient AI), which requires development of basic modelling, simulation and optimization methodologies in data-rich environments (MSODE), including model-order reduction. Management and deployment of large-scale application workflows will have to be adapted or invented. Network protocols will have to offer better control over the data logistics.

Furthermore, it is widely recognized that AI will play a central role in these extreme-scale, continuum infrastructures. This will occur at three levels:

- "AI for digital infrastructure" addresses how AI-inspired techniques can pilot and monitor the continuum and, in so doing, provide solutions to the points listed above.
- "Digital infrastructure for AI" tackles the question of re-designing the e-infrastructure to efficiently deal with data analysis and machine learning, which means, amongst other things, tuning of data access, I/O, low precision arithmetic, and moving code and data to where they will be most efficiently performed.
- "AI for science, industry and societal challenges" deals with the ever-increasing need to exploit AI techniques for extreme-scale, combining data and compute through the interpretation and coupling of computing results, measure-

ments and observations (e.g. digital twins in extreme earth modelling, combining climate models with satellite data and on-ground sensors).

The overall objective is to target high Technology Readiness Level (TRL) solutions (7 and more), based on horizontal synergies and interdependencies between all the concerned digital infrastructure technologies: HPC, big data, machine learning, IoT, 5G, cybersecurity, processor technology (EPI) and robotics. All of these components of the digital infrastructure will, together, be able to address critical societal challenges and sustainable development goals by mobilizing their amazing potential all the way across the continuum.

The TransContinuum Initiative (TCI)

The TransContinuum Initiative was launched in the wake of ETP4HPC's collaborative work with five other associations and projects to prepare the latest edition of the ETP4HPC Strategic Research Agenda for HPC in Europe [5], which was published in March 2020. Currently, it involves ETP4HPC, ECSO, BDVA, 5GIA, EU MATHS IN, CLAIRE, AIOTI and HiPEAC.

The TransContinuum Initiative will focus on the following five objectives:

- Identify priorities and recommendation for European R&I work programs: jointly we will elaborate recommendations for R&D to be carried out in EU-funded work programs addressing challenges in the digital continuum. The recommendations will cover challenges in technological (hardware and software) functionality, interoperability, and APIs. New standards, best practices, methodologies and project-type related suggestions will also be generated. Applications deployed in the digital continuum are addressed wherever needed.
- Engage in discussions with European R&I funding agencies and R&D programs (e.g. JUs, Missions): the recommendations will be presented to EU-funding entities such as Joint Undertakings (JUs) and applicable programs within the MFF 2021-2027. TCI representatives will be available to present and explain these recommendations as well to discuss any possible further analysis and elaborations.
- Generate and foster an interdisciplinary network of experts: we look forward to a lively exchange of ideas about EC work programs, calls and related events, events of partner organizations and potentially joint activities. We will jointly analyze new industrial and scientific use cases to better understand the challenges presented, together with an identification of "weak signals" for preparing for the future. This is a pre-requisite for any R&I recommendations and it also facilitates the forming of interdisciplinary consortia for the upcoming calls.
- Contribute to SR(I)As and other partners' documents: based on the results of the joint work mentioned above, contributions to the Strategic Research (and Innovation) Agendas or any other roadmapping documents issued by participating partners will be offered.
- Contribute to the 5 Horizon Europe missions: one of the first pragmatic actions will be to design the contribution of the concept of digital twin to the Horizon Europe missions (adaptation to climate change including societal transformation, cancer, healthy oceans, seas, coastal and inland waters, climate-neutral and smart cities, soil health and food). To achieve their respective goals, these missions will require digital technologies, in which digital twins should constitute the critical component.

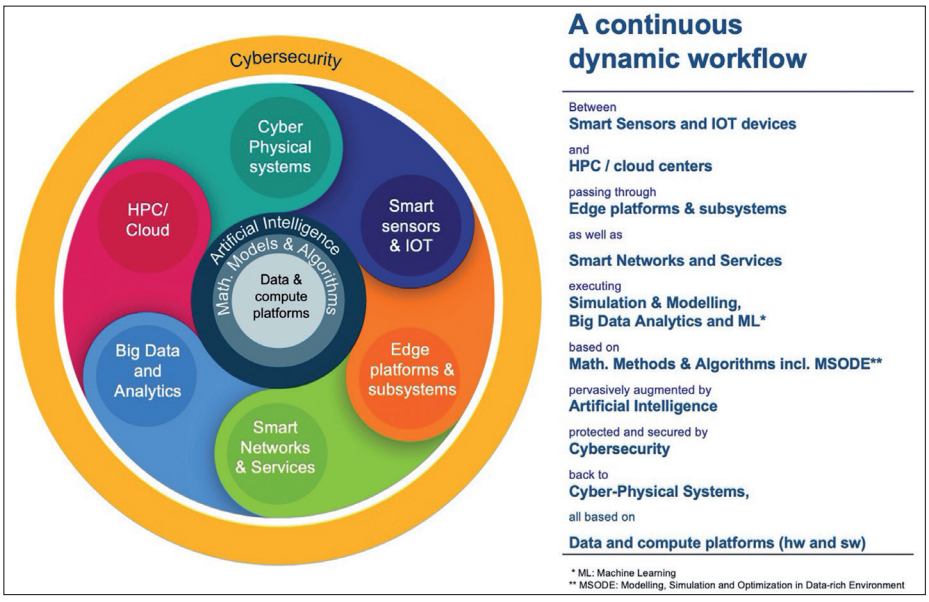


Figure 3: A pictorial view of the Transcontinuum.

References

[1] HiPEAC, "HiPEAC Vision 2017", https://www.hipeac.net/vision/2017/

[2] HiPEAC, "HiPEAC Vision 2019", https://www.hipeac.net/vision/2019/

[3] ETP4HPC, "A blueprint for the new Strategic Research Agenda for High Performance Computing", https://www.etp4hpc.eu/pujades/files/Blueprint%20document_20190904.pdf

[4] "TransContinuum Initiative (TCI): our vision", https://www.etp4hpc.eu/transcontinuum-initiative.html

[5] ETP4HPC, "Strategic Research Agenda (SRA)", https://www.etp4hpc.eu/sra.html

Marc Duranton is the coordinator of the HiPEAC Vision 2021 document, and also participant of ETP4HPC and involved in its SRA.

Michael Malms is the SRA lead editor within the office team of ETP4HPC and main initiator and coordinator of the TransContinuum Initiative.

Marcin Ostasz is an ETP4HPC Office Expert who co-manages the delivery of ETP4HPC's roadmapping documents and other related tasks.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: M. Duranton, M. Malms, and M. Ostasz. The continuum of computing. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 40-43, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

The TransContinuum Initiative: exploiting the full range of digital technologies for the prediction of weather and climate extremes

The extremes prediction use case

By PETER BAUER, MARC DURANTON and MICHAEL MALMS

Dealing responsibly with extreme events requires not only a drastic change in the ways society addresses its energy and population crises. It also requires a new capability for using present and future information on the Earth system to reliably predict the occurrence and impact of such events. A breakthrough in Europe's predictive capability can be made manifest through science and technology solutions delivering as yet unseen levels of predictive reliability with real value for society.

The "TransContinuum Initiative", initiated by ETP4HPC, ECSO, BDVA, 5GIA, EU MATHS IN, CLAIRE, AIOTI and HiPEAC, offers unprecedented opportunities to overcome the technological limitations currently hampering progress in this area. Beyond providing this use case with better technology solutions, the initiative offers a foundation for an Earth system – computational science collaboration that will eventually lead to science and technology being truly co-developed, and thus to sustainable benefit for one of today's most relevant applications and for European technology providers.

Key insights

- The reliable prediction of weather and climate extremes represents an extreme-scale computing and data handling challenge.
- The European and global Earth system science community has established a solid network of technology solutions for its complex and time critical workflows; however, they will not scale to meet future requirements.
- Climate change and the urgency for societies to adapt to future extremes require new technology solutions that provide breakthroughs for how we observe and simulate the Earth system.
- The cross-disciplinary "TransContinuum Initiative (TCI)" promises true co-design opportunities exploiting the entire breadth of digital technologies for the benefit of skilful extremes prediction capabilities in Europe.

Key recommendations

- Future systems will require at least 100 times more computational power for producing reliable predictions of Earth-system extremes with short lead times. It will require implementing an Earth-system digital twin – a cyber-physical entanglement. A solution is a layered federation with fewer elements near the heavy workloads and more elements at the observational data pre-processing front-end and the data analytics post-processing back-end.
- The extensive use of edge computing needs low-cost yet high-performance computing facilities and to overcome the data-transfer bottlenecks between the computing and data intensive parts of the digital twin and downstream applications
- Interoperable machine learning toolkits facilitating the portability of data processing in the cloud and workflow management options in the cloud for orchestrating the rather complex data assimilation and Earth-system simulation workloads should be developed
- Several domains need to be simultaneously promoted:
  - for software: interactive workflows, mathematical methods and algorithms, high-productivity programming environments, performance models and optimization tools.
  - for hardware: heterogeneous processor configurations through accelerators and data-flow engines, high-bandwidth memory, deep memory hierarchies for I/O and storage, superfast interconnects and configurable computing including the supporting system software stack.

The use case

Natural hazards represent some of the most important socio-economic challenges our society faces in the decades to come. Natural hazards have caused over 1 million fatalities and over 3 trillion Euros of economic loss worldwide in the last twenty years, and this trend is accelerating as a result of the drastic rise in demand for resources and population growth. The combination of likelihood and impact makes extremes and climate action failure the leading threats for our society [1,2,3].

The "Earth system extremes" use case relies on very complex numerical Earth system simulation models that ingest hundreds of millions of observations per day to help improve the formulation of the physical process representation as well as produce the initial conditions used for launching predictions of the future. This logic applies equally to weather timescales of days or weeks and climate timescales of decades. These systems are also run as ensembles whereby each ensemble member represents both initial condition and model uncertainty. The result is a prediction of state, but also a prediction of uncertainty, which is crucially important for decision-making on tight schedules [4].

Europe currently leads the world in medium-range weather prediction and is also a major competence centre for global climate prediction [5,6]. For decades, HPC has been a key enabler for this track-record and has led to weather and climate prediction becoming one of the leading use cases for computing and data handling at large scale. The apparent change in computing architectures has stimulated a wide range of programmes that aim to prepare the operational Earth-system monitoring and prediction infrastructures for future technologies [7]. These programmes increasingly realize that it will take more than progress in HPC to fulfil future extremes prediction targets. This is where the need to exploit the opportunities offered by the entire digital TransContinuum [8] comes into play, and where new ways of co-developing Earth system and computational science need to be found.

Implementing an Earth-system digital twin – called the cyber-physical entanglement in Fig. 1 – through these technologies is the ultimate goal. The loop shown in the figure would be open as humans are continually influencing nature, for example through CO_2 emissions at global scale or irrigation in agriculture at small scales. But natural variability can also affect the system and needs to be replicated, for example extreme events such as volcanic eruptions injecting large amounts of ash into the atmosphere leading to a change in radiative forcing.

Beyond advancing present-day Earth-system simulation and observation capabilities, the digital twin would close the gap between Earth system science and socio-economic sectors in which it might be applied and facilitate a high level of flexible intervention by non-science/technology experts. Hiding the complexity of the digital TransContinuum from the user is key to success for user-driven digital twins [9].

As shown in Fig. 1, the TransContinuum embeds smart sensors and IoT within smart networks supplying data to extreme-scale computing and big data handling platforms. These platforms exploit cloud services across workflows giving access to additional distributed computing and

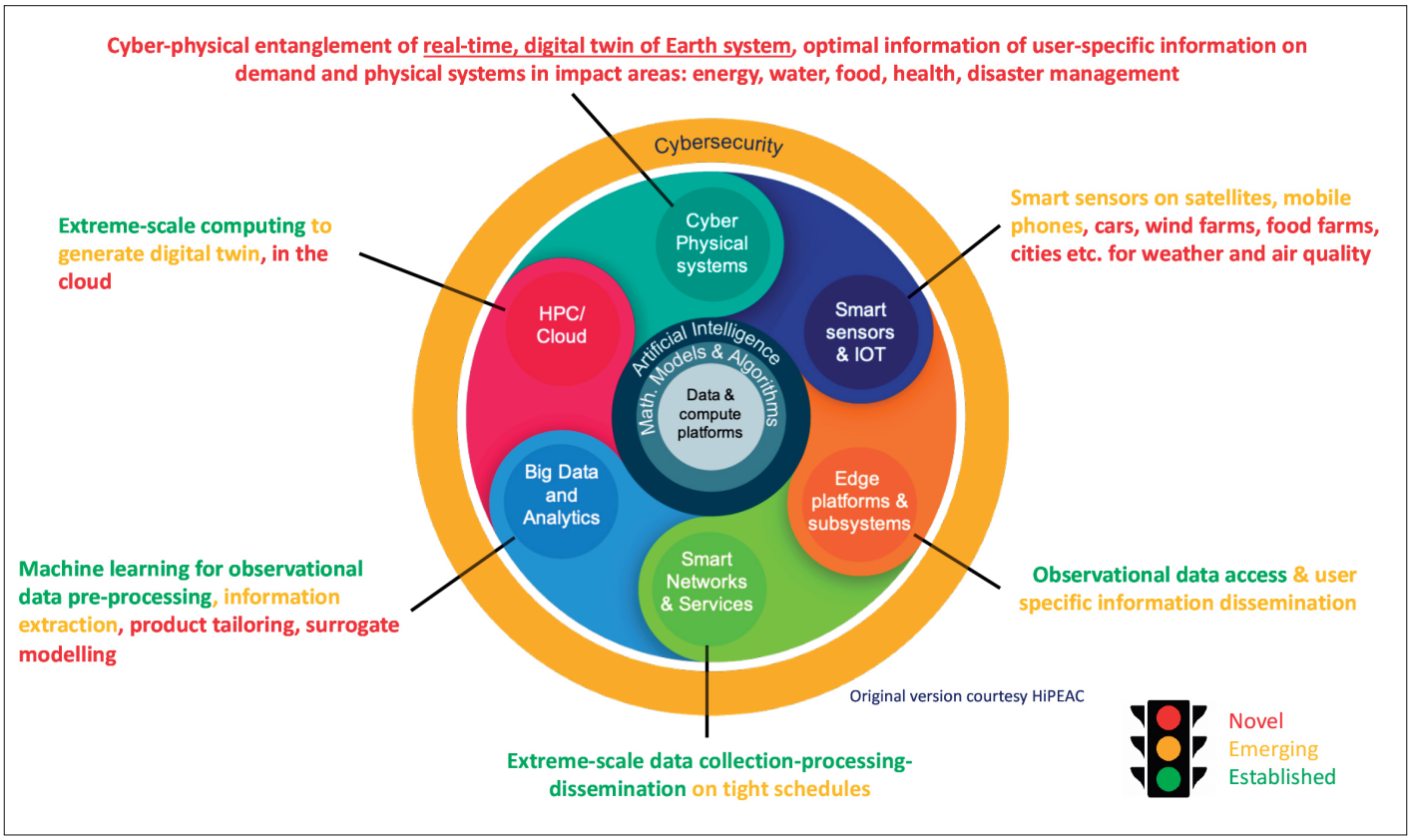


Figure 1: Main elements of digital continuum and relevance for extremes prediction use case including readiness of both application and technologies (green = established in production, but not optimal in performance; yellow = research, not in production mode; red = novel and unexplored).

data analytics capabilities. Mathematical methods and algorithms as well as artificial intelligence and machine learning provide the glue between the elements of the TransContinuum. Cybersecurity acts as a shield for the entire loop. Such a methodological framework based on extremes prediction also offers gains for other domains.

Challenges along the TransContinuum

Assessing the challenges for the extremes prediction use case with respect to the TransContinuum requires a closer look at the production workflow in today's systems and how they are likely to evolve in the future.

Smart sensors and Internet of Things

At present, Earth-system observations that are used operationally already comprise hundreds of millions of observations collected daily to monitor the atmosphere, oceans, cryosphere, biosphere and the solid Earth, the largest data volumes being provided close to a hundred satellite instruments [10]. This volume is expected to increase by several orders of magnitude in the next decade, bringing with it a need to ingest such observations in digital twin systems within hours. Smart sensor technology is highly relevant for satellite instruments that can perform targeted observations and perform on-the-fly data pre-processing.

However, observations from commodity devices deployed on e.g. phones, car sensors and specialized industrial devices monitoring agriculture, renewable energy sources and infrastructures also offer data that is currently inaccessible to operational services and that can fill vast observational gaps in less developed countries, offers much finer resolution in densely populated areas and in regions of significant socio-economic interest [11]. Beyond technical challenges, a generic approach for fast access to commercial data for public use via public-private partnerships is required, which is already on the agenda of intergovernmental organisations like the World Meteorological Organization. This element of the TransContinuum is only developing now and has huge potential.

Smart networks and services

Collecting and transferring massive amounts of diverse data from devices scattered in multiple areas via distributed pre-processing centres to centralized digital platforms and computing centres requires a suitable network infrastructure. Evolved networks and services should offer secure and trustable solutions that will support the desired quality of service for different data flows. The extremes prediction use case is particularly HPC- and big data-driven and creates a significant footprint for the supporting communication networks. Pre-processing in edge devices can already offload some of this burden as this use case requires near real-time data availability at the HPC facilities. The extensive use of edge computing needs low-cost yet high-performance computing facilities that will interact with end devices as well as with one another.

Given the decade-long evolution of interconnected data collection, transmission, pre-processing, centralized HPC and post-processing, and dissemination in weather prediction workflows there is little room for optimization of present-day systems. However, workloads, data volumes and data diversity grow much faster than in the past and the demand for providing more skilful predictions is more urgent than ever before. Therefore, network and service solutions need to orchestrate and dynamically manage data routing and computing resources with much more flexibility and scalability. Whether both performance and cost-effectiveness need dedicated solutions or whether it can be achieved by a mixture of commercial and institutional systems is an as yet unsolved question.

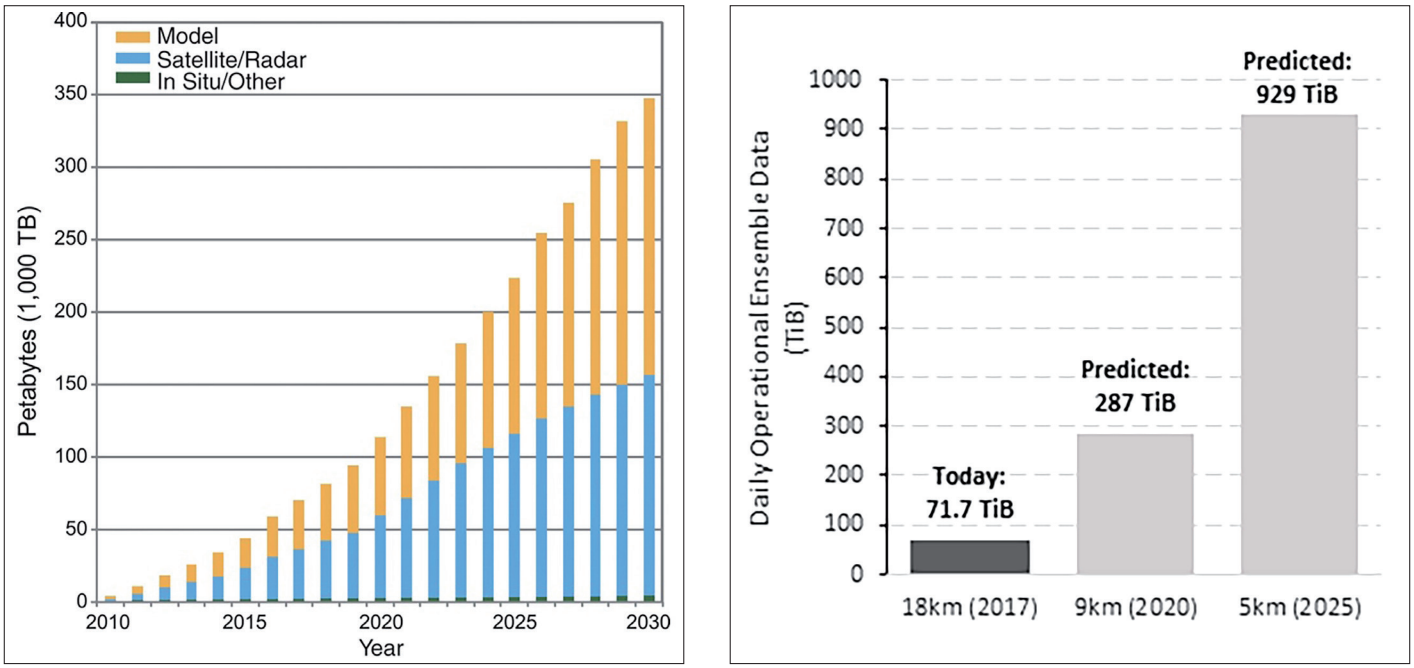


Figure 2: a) The volume of worldwide climate and daily weather forecast data is expanding rapidly, creating challenges for both physical archiving and sharing, as well as for ease of access, particularly for non-experts. The figure shows the projected increase in global climate data holdings for climate models, remotely sensed data, and in situ instrumental/proxy data (from [12]). b) Daily data volumes produced by the 50-member ECMWF weather forecast ensemble at today's spatial resolution of 18 km, the next upgrade to 9 km and the expected configuration in 2025.

Big data and data analytics

The volume of Earth-system observation and simulation data in weather prediction already exceeds hundreds of TBytes/day while climate projection programmes produce PByte-sized archives, which take climate scientists years to analyse (see Fig. 2) [12]. The extrapolation of data volumes and production rates to advanced Earth-system digital twins will prohibit the effective and timely information extraction that is critical for timely actions to anticipate and mitigate the effects of extremes. Both simulations and observations need to be generated and combined in the Earth-system digital twin within minutes to hours of time-critical workflows towards near-real-time decision-making.

Overcoming the data-transfer bottlenecks between the computing and data intensive parts of the digital twin and downstream applications is crucial, and future workflow management needs to make such applications an integral part of the observation and prediction infrastructure. Powerful data analytics technology and methodologies offer the only option to make the effective conversion from PBytes/day of raw data to Mbytes/day of usable information. This conversion must be tailored to individual sectors needing to prepare and respond to extremes, namely water, food, energy, health, finance and civil protection. Machine learning-based quality control and error correction, information extraction and data compression as well as processing acceleration will be key.

High-performance and cloud computing

Today, experimental and operational Earth-system simulations use petascale HPC infrastructures, and the expectation is that future systems will require at least 100 times more computational power for producing reliable predictions of Earth-system extremes with lead times that are sufficient for society and industry to respond [13]. This need translates into a new software paradigm to gain full and sustainable access to low-energy processing capabilities, dense memory hierarchies as well as post-processing and data dissemination pipelines that are optimally configured across centralized and cloud-based facilities. European leadership in this software domain offers a unique opportunity to turn European investment in HPC digital technology into real value.

A big challenge is the definition of the evolution of today's strongly centralized workload and data lifecycle management towards a more distributed system – still being data-centric to keep big data near extreme-scale computing – that allows the decomposition of workflows such that the production chain can exploit the full processing and analysis potential of the TransContinuum. A likely solution is a layered federation with fewer elements near the heavy workloads and more elements at the observational data pre-processing front-end and the data analytics post-processing back-end. This distribution also needs to be driven by the urgency of product delivery as, particularly for extremes prediction, processing speed trumps all other considerations.

Adding flexibility without sacrificing performance through the TransContinuum is where most of the strategic recommendations for research and innovation compiled by the European Technology Platform for HPC (ETP4HPC) [15], the Big Data Value Association (BDVA) [16] and High Performance Embedded Architecture and Compilation (HiPEAC) [17] are positioned. For software, these are interactive workflows, mathematical methods and algorithms, high-productivity programming environments, performance models and optimization tools. For hardware, heterogeneous processor configurations through accelerators and dataflow engines, high-bandwidth memory, deep memory hierarchies for I/O and storage, super-fast interconnects and configurable computing

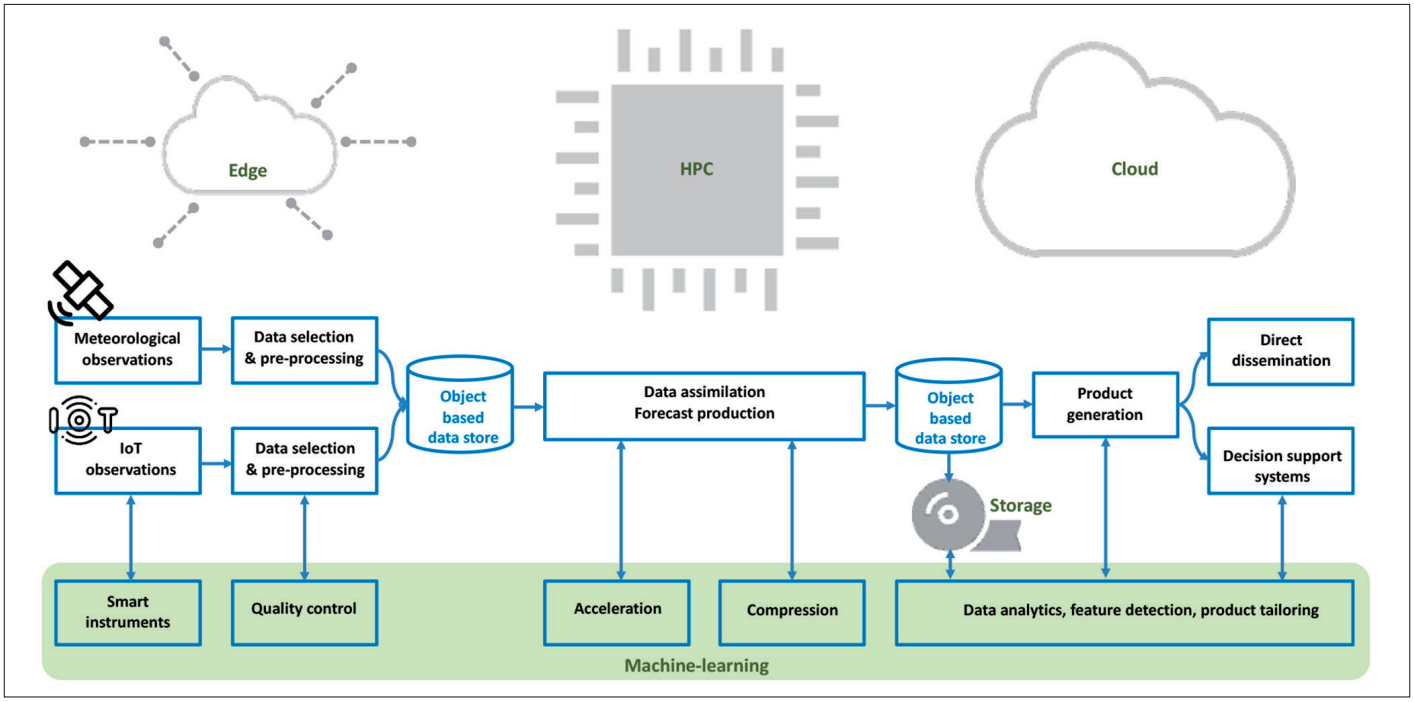


Figure 3: Main components of extremes prediction workflow, their alignment with edge, high-performance and cloud computing elements of the TransContinuum, and key contributions of machine learning to workflow enhancements.

including the supporting system software stack are part of the agenda [14].

For the prediction of extremes, the main computing tasks are currently performed on centralized, dedicated systems where the main data storage facilities also lie. The observational input data flow crosses all levels of edge, cloud and centralized computing during which selected pre-processing steps are exercised, for example, for satellite data collection and pre-processing at distributed receiving stations of the ground segment, their further dissemination and processing by space agencies and meteorological centres, and assimilation into models at prediction centres. Similar data flow mechanisms exist for ground-based meteorological observation taken from networks, stations and (few) commercial providers.

Increasingly however, the back-end of model output data processing interfacing with commercial service providers is being placed into the cloud, which also offers better access to machine learning-based data analytics. Here, the biggest gaps are interoperable machine learning toolkits facilitating the portability of data processing in the cloud and workflow management options in the cloud for orchestrating the rather complex data assimilation and Earth-system simulation workloads (see Fig. 3).

Conclusions

Predicting environmental extremes clearly has an extreme-scale computing and data footprint. It has reached a point where only through significant investment in all parts of the TransContinuum can the future needs of our society for reliable information on the present and future of our planet be fulfilled. These investments must go hand in hand with investments in basic Earth system science, to better understand key processes and their interaction, and for service provision ensuring that Earth system data is translated into useful information for society.

While the present digital infrastructures support weather and climate prediction well enough, the chosen domain-specific solutions will not scale to meet future requirements. True near-sensor, edge processing does not yet exist and IoT devices are not yet used; smart and configurable networks as well as flexible HPC and cloud solutions are not available; and even for the extreme-scale computing and data handling tasks, the present algorithmic and programming frameworks do not allow the exploitation of the real potential of emerging technologies.

Building on past European science and technology programmes such as the Future and Emerging Technology for HPC (FET-HPC) [18] under Horizon 2020, the recent implementation of the EuroHPC Joint Undertaking, and the Copernicus Programme, Europe has actually recognized the present gaps and needs for serious investments. This has motivated the ambitious Destination Earth action [19]. In support of the Green Deal and the European strategy for data [20], Destination Earth promises to bring the Earth system science, digital TransContinuum and service development strands together assigning the highest priority to extremes prediction and

climate adaptation. The programme has adapted the concept of digital twins of the Earth system for this purpose as promoted by HiPEAC and ETP4HPC and its strong digital infrastructure contribution has huge potential for achieving European technology leadership by addressing one of the principal challenges for today's society.

References

[1] UNISDR, "Economic losses, poverty & disasters, 1998-2017", (UNISDR, Geneva, 2018)

[2] AON, "Weather, climate & catastrophe insight", Annual report, (AON, Chicago, 2020)

[3] WEF, "The global risks report 2020", Insight report 15th edition, (World Economic Forum, Geneva, 2020)

[4] P. Bauer, A. Thorpe, and G. Brunet (2015). The quiet revolution of numerical weather prediction. Nature, 525(7567), 47-55.

[5] R. A. Kerr (2012). One Sandy forecast a bigger winner than others. Science, 736-737.

[6] P. Voosen (2019). New climate models predict a warming surge. Science, 16.

[7] P. Neumann, J. Biercamp, (2019, September). ESiWACE: On European Infrastructure Efforts for Weather and Climate Modeling at Exascale. In 2019 15th International Conference on eScience (eScience) (pp. 498-501). IEEE.

[8] "TransContinuum Initiative (TCI): our vision", https://www.etp4hpc.eu/transcontinuum-initiative.html

[9] P. Bauer, B. Stevens, & W. Hazeleger (2020). A digital twin of Earth for the green transition. Nature Climate Change, submitted.

[10] W. Balogh, and K. Toshiyuki. "The World Meteorological Organization and Space-Based Observations for Weather, Climate, Water and Related Environmental Services." In Space Capacity Building in the XXI Century, pp. 223-232. Springer, Cham, 2020.

[11] J. M. Talavera, L. E. Tobón, J. A. Gómez, M. A. Culman, J. M. Aranda, D. T. Parra, ... & L. E. Garreta, (2017). Review of IoT applications in agro-industrial and environmental fields. Computers and Electronics in Agriculture, 142, 283-297.

[12] J. T. Overpeck, G. A. Meehl, S. Bony & D. R. Easterling (2011). Climate data challenges in the 21st century. Science, 331(6018), 700-702.

[13] T. C. Schulthess, P. Bauer, N. Wedi, O. Fuhrer, T. Hoefler & C. Schär (2018). Reflecting on the goal and baseline for exascale computing: a roadmap based on weather and climate simulations. Computing in Science & Engineering, 21(1), 30-41.

[14] P. Bauer, P. D. Dueben, T. Hoefler, T. Quintino, T. Schulthess & N. P. Wedi (2020), The digital revolution of Earth-system science. Nature Computational Science, submitted.

[15] ETP4HPC, https://www.etp4hpc.eu/pujades/files/ETP4HPC_SRA4_2020_web(1).pdf

[16] BDVA, https://www.bdva.eu/BDVA_SRIA_v4

[17] HiPEAC, https://www.hipeac.net/vision/#/latest/

[18] "European HPC technology research projects", https://www.scientific-computing.com/white-paper/european-hpc-technology-research-projects

[19] "Destination Earth (DestinE)", https://ec.europa.eu/digital-single-market/en/destination-earth-destine

[20] "A European Strategy for Data", https://ec.europa.eu/digital-single-market/en/european-strategy-data

Peter Bauer is the Deputy Director of the Research Department at the European Centre for Medium-Range Weather Forecasts, Reading, UK.

Marc Duranton is researcher at the Research and Technology Department of CEA (Alternative energies and Atomic Energy Commission), France and the coordinator of the HiPEAC Vision 2021.

Michael Malms is the SRA lead editor within the office team of ETP4HPC and main initiator and coordinator of the TransContinuum Initiative.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

V.1, January 2021.

Cite as: P. Bauer, M. Duranton, and M. Malms. The extremes prediction use case.

In M. Duranton et al., editors, HiPEAC Vision 2021, pages 44-49, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020

research and innovation programme under grant agreement number 871174.

Combining artificial intelligence (AI), cybersecurity, web technologies, and selecting and orchestrating services to fulfil your requirements, the "Guardian Angels" concept is the next generation interface between the cyber and the real worlds.

"Guardian Angels" to protect and orchestrate cyber life

By MARC DURANTON and TULLIO VARDANEGA

The web has forever changed the ways in which people interact with computers. Applications are no longer a single piece of code running on a local computer. From Web 1.0, where data and its presentation were solely determined by the hosting server, to Web 2.0 where users can also upload data, create and animate social networks, we now forecast a "next web". The centre of that transformation will be personalized, loyal and user-centred "Guardian Angels", mobile programmable virtual engines. Such Guardian Angels will allow businesses and individuals to draw from the services available in the digital world, free from the hassle of keyboard-based or tactile programming, and the annoyances of information flooding, and protected from the dangers of abusive, malicious or hostile digital actors.

The evolution to "next web" will expose composable services, associated with document-based contracts that will specify functional and non-functional requirements such as response time and latency, energy requirements, migration of the service to a local computing resource, cost, etc.… Applications will result from the orchestration of these composable services. Orchestrators will select services according to users' preferences, and will orchestrate their execution in a coherent way. The orchestrator will also be in charge of managing the in-bound (security, verification of trust) and out-bound (privacy and confidentiality) data, acting as the "Guardian Angel" of users' assets. The "programming" of this orchestrator will be done using natural interfaces, such as voice, drawings and schematics, or even by examples, thanks to the innovation brought about by artificial intelligence techniques.

These "Guardian Angels" will be loyal to their owners, and will communicate with each other to compose additional, higher-level, services. They can also run locally and master the locality of code execution thanks to containerization and other virtualization techniques, allowing migration of the execution as and when opportune along the continuum of computing.

To kick-start this evolution to the "next web", we propose the launch of a European "moonshot" project that will build a first demonstration instance of this evolution, and gather competences from a large set of domains.

Key insights

- "Guardian Angels" is a proposal for a "moonshot programme" that will promote collaboration and synergy between artificial intelligence, in- and outbound data management, interoperability and contract-based programming approaches, edge processing and resource federation, natural-interface programming, containerization and code migration, real-time and safe and secure services, etc.
- It will encompass the digital continuum, from extreme-edge IoT to the cloud and, protecting the assets of internet users, will ensure a trustworthy interface between the real world and cyberspace.
- It will be the seed of the "next web", composed of services with known properties based on interface contracts, allowing users to create complex applications by orchestrating the selection of and exchanges between the services, intertwining cyber and physical worlds with guarantees of fulfilling the (non-functional) requirements, with safety and security in mind, for industrial and personal use.
- The orchestration "scripts" will be automatically generated from requirements expressed in a natural way (voice, graphics, gestures, showing examples, …).
- The "core" of the "Guardian Angel", containing its behaviour, and important and private data, will be embedded in devices with very low physical footprint (Ultimately, as small and portable as earbuds?).
- It will be built on top of existing technologies and will be interoperable with current technologies.

Key recommendations

- A "next web" is to be developed and deployed; it will intertwine the cyber and physical worlds for industrial and personal use, overcoming the fragmentation of vertically-oriented closed systems, the heterogeneity and the lack of interoperability.
- It will increase scalability in a dynamic environment where systems should self-configure, self-manage and be plug-and-play, while also coping with security and privacy of personal and corporate data.
- The core is the notion of advanced orchestrators, which we call "Guardian Angels", loyal to their "users", placed at the interface of the physical and virtual world, to orchestrate by the "new web" in a safe and secure way.

Back in time

When the web (WWW) first appeared thirty years ago, very few people understood that it would disrupt the way we interact with computers. Back then, the number of servers was limited, and the communication flow was mainly monodirectional, where the "content provider" provides the content and also determines how to present it.

By adding a small software layer, Web 2.0 allowed users to upload their data to the servers, and to share it. Even if presentation of information was still mainly driven by the service provider, the user-directed mode of data exchange enabled social media to come into existence, bringing along with it an entirely new way for users to interact.

Figure 1: The Web 1.0 was mainly mono directional

The next step that changed the manner in which we use "computers" involved not only software, but also a complete redesign of the terminal. In 2007, the introduction of the iPhone by Apple, and its support for wireless connectivity totally modified our interaction with the web, which could now be accessed virtually everywhere. The other key element was the presentation of the information: no longer imposed by the provider through a fix-layout page accessed via a web browser, but personalized by an "app". No imposed keys or one-size-fits-all style: multiple apps can access the same provider, which is now only providing a "service", and can use and display it in totally different ways.

Limits of the current model

While enjoying the benefits of such innovation, we also begin to appreciate the limits of the current model and how it can have negative impacts on our lives, rather than offering only improvements.

Complexity and multiplicity of apps: Initially, each service provider had its own app. For example, if you wanted to book a hotel room, you had to use the app from each hotel to find the one you

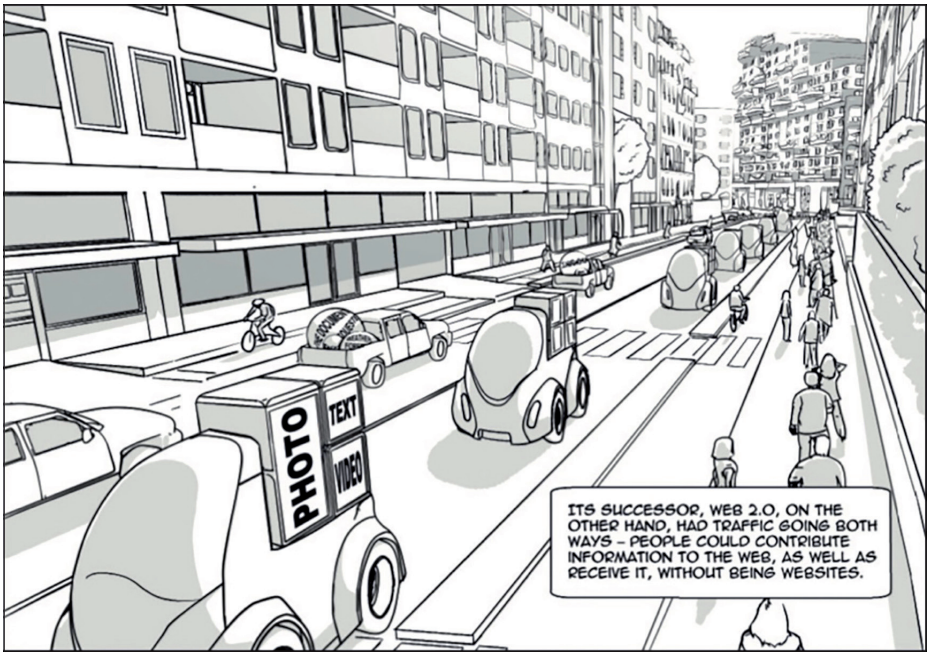


Figure 2: Web 2.0 allowed users to share content

liked. "*Aggregator apps*" appeared to mitigate this annoyance: they allowed the user to easily compare the various services and choose among them from a single vantage point. Next, aggregators of aggregator apps appeared, which compare and try out (not just present) the services of all the different aggregators to find the best overall service to meet the user's specifications. This progression could easily cause the end-user to download even tens of different apps for just one and the same eventual service: booking a hotel room.

What if you could instantaneously build your own aggregator, which knows your preferences and profile, and could orchestrate the information available in the web to deliver the best service possible according to your request?

Overflow of information: Attention time is a very precious and scarce resource for individual users, and a very important asset for commercial organizations. The recommendation engines are tools whose side effect is to create addiction in service users, to induce them to spend as much time as possible using such services. The recommendation engines are built to maximize the use of the service provider.

What if you had your own (or your corporate, in case of business users) recommendation engine which is truly loyal to you and fetches and selects information according to your personal preferences instead of you having to put up with it being pushed at you by providers?

Fake news: It is increasingly difficult to tell the difference between forged (fake) news, pictures, interviews and movies, and real ones. Deep-fakes are increasingly realistic and easy to make.

What if you had an intelligent agent that fetches news and information, and that is able to check (from correlation with trusted sources, analysis of the document by AI techniques, etc.) the trustworthiness and authenticity of the documents, including e-mails and protection against phishing?

Protecting private data: At present, commercial organizations make a business of their customers' private data. Your data is being collected and analyzed and sometimes sold without you knowing. You should be in control of what you share with a particular service provider; you don't want to share the same information in a public social network that you share with your medical doctor.

What if an intelligent agent filtered what you exchange with others, ensuring that you don't involuntarily share information you don't want to share? Such an agent will know your preferences, and will interact with you in the event that you have not yet specified any necessary preferences.

No more silos: The current recommendation engines are very crude and tend to present you only with items very similar to the ones you already accessed. This has the drawback of locking you into a "bubble", a virtual silo, which is reinforced each time you access a new recommended item.

What if your "loyal" assistant is able not only to provide you with relevant items that are truly in line with your wishes, but also to analyze the past context and provide you with new and diverse items from different sources that will enlarge your views and your life, avoiding locking you in silos and filter bubbles? Of course, it is very important that this "psychological profile" should be kept highly confidential and accessible only by you (even localized on a personal hardware, un-hackable – or as difficult as possible to hack from outside).

Better health: Current smartwatches, together with their associated apps, are able to monitor your health and some of them are even able to send alarms when they detect anomalies. However, it is not very clear where your personal health data is stored and analyzed, or whether it is sold to health insurance businesses …

What if your assistant was truly loyal to you and kept your personal health data private? What if it carried out analysis and diagnosis either locally or using techniques like Homomorphic encryption that allowed your data to be protected even if the analysis was done remotely on untrusted servers. It is very important that this "health profile" should be kept highly confidential and accessible only by you and your medical doctor (upon your approval), and is protected against hacking.

Frustration with digital assistants: The current digital bots (Google assistant, Amazon's Echo, Apple's Siri, Microsoft's Cortana, Samsung's Bixby, Baidu's Duer, Xiaomi's Xiao AI, etc. …) are quite limited and not very efficient in dialogues. Their understanding of the context is limited, leading to frustration for their users.

What if your assistant could start active discussion with you when they don't have enough information, and know and store your preferences in a protected way? It should be also able to take the initiative

according to your habits or usages, or to the context, like a "human" assistant would do.

Independence from remote processing: It is a common experience that when a particular service closes down or changes its cost model, it puts its users in jeopardy. A recent example is the smart home device integrator IFTTT that changed its cost model from free to subscription-based while still being financially backed by the providers of smart home devices. Even worse are devices that become useless if their server is down temporarily or, as in case of bankruptcy, permanently. This happened for the owners of the Nabaztag ambient electronic devices, and for the owners of the Cozmo robots, when their servers went down. This has even happened for Amazon's Echo services, which were down for several hours in July 2020.

What if you could easily change the service provider if your usual provider were down or, better still, if you could migrate the functionality to a more trusted server, or to one more local to you, or even on your own hardware? This migration could be done automatically or on request. For example, a smart light bulb should be connected locally, not to a server 10,000 km away, so that you can still turn it on or off even if your internet connection were down. We should have the choice of where the service is executed, and even be able to split a complex service into sub-services where we can assign localization of execution for each of the sub-services. Migration of services from one server to another, with interoperability and clear interfaces, is one challenge that has to be overcome.

No contextual information: Current systems don't take into account their users' situation and don't select appropriately their interface with them. For example, if you are in a meeting or in a noisy restaurant, a short message on your smart watch might be more appropriate than a loud ring; if you are driving, vocal interface might be more appropriate than a display, but only if the situation allows it.

What if the system you used selected its interface depending on the context and always used the most appropriate and least annoying one? This could range from a virtual keyboard, to listening to your voice, to understanding your mimics or your gesture, and displays information on your smart watch if you are looking at it, or on your smartphone or any screen around if you are looking at it, or by voice or tactile signaling if it is appropriate.

Plug and play and interoperability: When you add a new device to your ecosystem, its integration is seldom smooth: you have to download a particular application, connect it to the device, try several times, enter credentials and other data just to discover that it was made for a particular ecosystem, which is not the one you have.

What if you could immediately access the services provided by a new device as soon as you power it on, its resources being automatically discovered and integrated in your own ecosystem and interoperable with all other devices you already have?

Technical challenges and technologies to overcome the limitations of the current model

The notion of "continuum of computing" reflects the observation that edge, fog and cloud computing platforms are being pulled close together into what will likely become a seamless execution environment (see Figure 3, bottom part). This is being driven by the push exerted by a massively increasing amount of value-added applications targeting mobile, handheld, wearable and unattended devices, which can all run over one and the same base programmable interface: the internet. Those applications may serve human users or industrial apparatuses, as in the IoT; one trait that they have in common is the drive to offer their user a constantly increasing wealth of functionalities, see Figure 3, top-centre part.

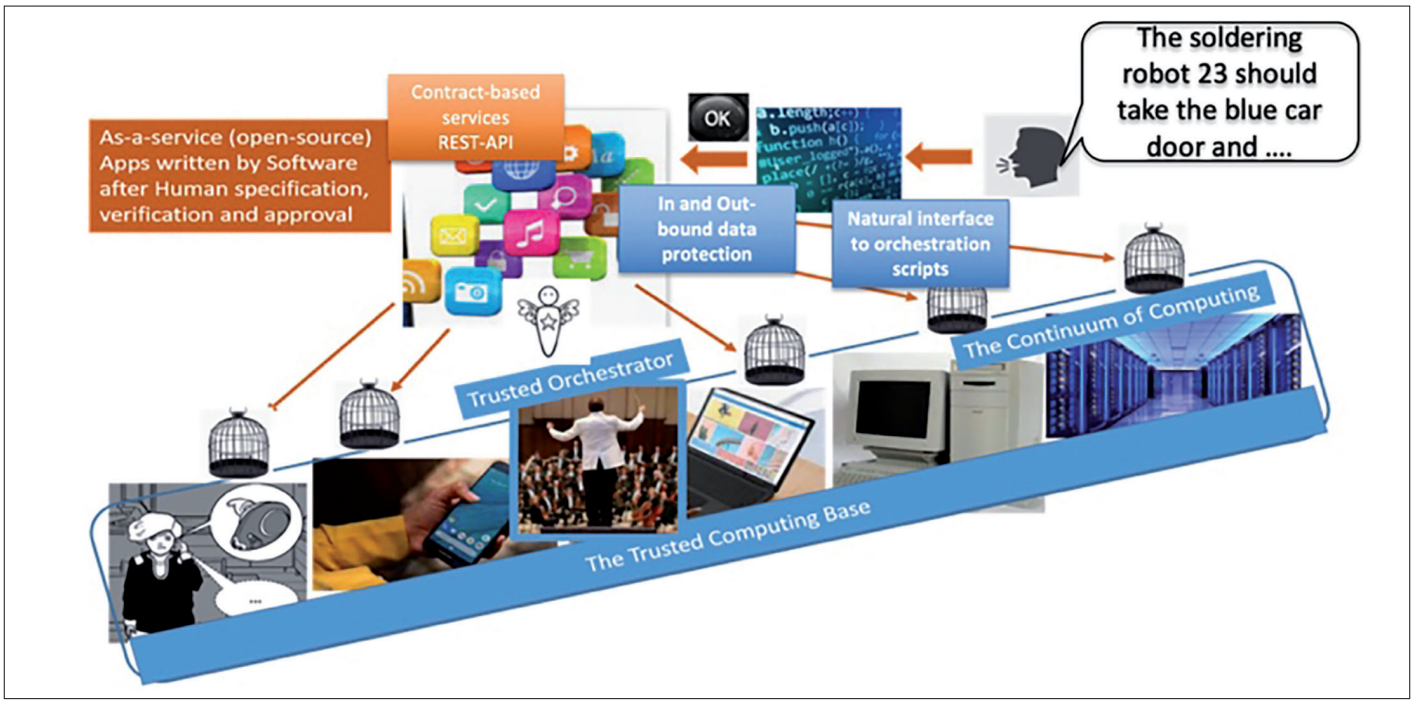


Figure 3: The "Continuum of computing"

To overcome the limitations of the current model and to enable the "Guardian Angels" concept, several technologies and developments need to be launched in a coordinated fashion. Fortunately, existing technologies or technologies in development can be the base of what will be needed to achieve the "Guardian Angels" goal as above outlined. However, their development and integration need to be coordinated in a coherent way: this can be done through the moonshot proposal described below. A non-exhaustive list of technology axes that need to be developed includes:

- The technology for the **"next web"**, allowing interoperability, scalability, and the orchestration of services [5]. Typifying examples of candidate technology solutions include:
  - As-a-service delivery mode
  - Use of the highest level of the internet protocol stack (HTTP/3)
  - Application-level containerization
  - Self-served scheduling
  - Lightweight Trusted Computing base (similar to the mechanics behind modern web browsers), written in memory-safe programming language and verified by verification tools before compilation and deployment
  - Execution by interpretation or JIT, like in Web Assembly
  - Dynamic web directories to discover services spread in the cyberspace
  - Dynamic orchestration of the services in a trusted and secure way, for example by cloud-native API-integration languages like Ballerina [2].
  - …
- A human-friendly, **natural way to "program"** the orchestrator, for example based on speech-to-code transformation. The recent advance of AI, as described in the "The omnipresent artificial intelligence" article of the HiPEAC Vision 2021, or as demonstrated in the Almond project [3] of the Stanford Open Virtual Assistant Lab, are steps in that direction, the GPT-3 model being the most recent and futuristic case at present.
- The recent evolution of recommenders should allow them to run locally, with better security and control of their results.
- The "Guardian Angels" should be able, individually or collectively, to detect fake or untrusted data, in general being able to assess the trustworthiness of the accessed data. The latest research in the field of fake data detection should be embedded in the system.
- Sharing data, with privacy in mind, as demonstrated in the Almond project, is important to enable collaboration between "Guardian Angels" from different (private or corporate) users. The mixing of deep learning techniques, federated computations and homomorphic encryption can also solve the problem of off-loading data and computation to untrusted servers without violating the secrecy of the data and of the results of the computations.
- The latest advances in cybersecurity should be introduced into the overall system and into each of the elements of the system.
- As the "Guardian Angels" will interact with the physical world by reading data from sensors and controlling devices, they form cyber-physical systems that should operate in a timely fashion, and conform with safety and other non-functional requirements like energy use, cost of activating paid services, locality of data and of processing, latency etc. In order to choose the right services to orchestrate to achieve the required ultimate functionality according to such requirements, contract-based interfaces should be supported by each such service.
- Last but not least, hardware at the edge should be taken into account. Ideally, the "core" (each "Guardian Angel") should be able to run on trusted hardware, near the user. Being able to develop low power hardware that can be integrated into a home (or factory) device at reasonable cost is very important for the success and the acceptance of the "Guardian Angel" concept. Subsequently, this could evolve in a smartphone form or perhaps even in an ear bud, like in the "Her" movie [4].

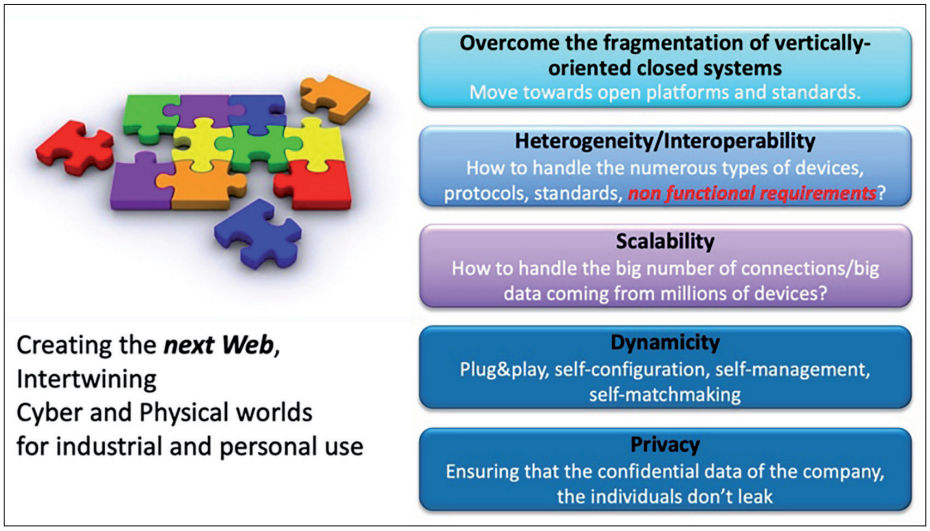


Figure 4: Requirements of the "next Web"

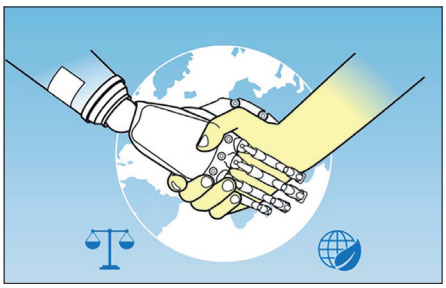


Figure 5: "Guardian Angels" will allow a true collaboration between men and the cyber environment

Conclusion

IoT devices, edge and deep edge devices, and smart assistants are at the interface between the real world and the cyber-world. However, owing to the increasing inherent complexity of the cyber-world, its efficient use becomes more and more challenging for industrial and personal users. A "next web" is to be developed and deployed; it will intertwine the cyber and physical worlds for industrial and personal use, overcoming the fragmentation of vertically-oriented closed systems, the heterogeneity and the lack of interoperability. It will increase scalability in a dynamic environ-

Figure 6 (Source: HiPEAC Comic Book [1])

ment where systems should self-configure, self-manage and be plug-and-play, while also coping with security and privacy of personal and corporate data.

The core of this moonshot proposal is the notion of advanced orchestrators, which we call "Guardian Angels", loyal to their "users", placed at the interface of the physical and virtual world, to orchestrate in a safe and secure way the various services provided by the "new web".

Everyone is welcome to help shape the proposal, and to make it happen!

References

[1] HiPEAC Comic Book, https://www.hipeac.net/media/public/files/46/7/HiPEAC-2019-Comic-Book.pdf
[2] Ballerina, https://ballerina.io/
[3] Stanford University, "Almond", https://almond.stanford.edu/
[4] Wikipedia, "Her', https://en.wikipedia.org/wiki/Her_(film)
[5] T. Vardanega "The continuum of computing: enabling technologies", HiPEAC vision 2021.

Marc Duranton is a researcher at the Research and Technology Department of CEA (Alternative energies and Atomic Energy Commission), France and the coordinator of the HiPEAC vision 2021.

Tullio Vardanega is Associate Professor in the Department of Mathematics of the University of Padua, Italy.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: M. Duranton and T. Vardanega. "Guardian Angels" to protect and orchestrate cyber life. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 48-53, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

The web interface is becoming the programmatic interface of new-generation applications. This interface can unite the edge, the fog and the cloud seamlessly, permitting applications to be developed, deployed and operated as an orchestration of independent parts that reside and operate wherever they best meet the user requirements.

The continuum of computing: enabling technologies

By TULLIO VARDANEGA

User-oriented and business computing is becoming ubiquitous, and spans from the cloud to the edge via fog nodes in between. The spectrum across such platforms is becoming the natural host environment for an increasing number of value-added applications that can be deployed at various points along it, without necessarily having a single fixed position. Most such applications are independent of one another, often very specialized, sometimes equivalent in function. Transforming their delivery mode into an as-a-service style favours their "servitization", which, by not requiring local installation, alleviates the burden on the end-user platform, thereby increasing their economic value. That transformation is comparatively easy to achieve since the web has become the programmatic interface of new-generation applications. This in turn allows us to regard the spectrum across the cloud and the edge via the fog as the "continuum of computing". That continuum is the perfect habitat for meta-applications that orchestrate selected sets of independent applications into user-defined workflows that single out the service providers via the schema-based interface contracts that they publish; decide on deployment; and obey the execution preferences that are most opportune to meet user requirements. Such preferences also extend quality of service to include privacy, confidentiality, energy, social and ethical concerns.

Key insights

- New-generation browsers, with a thoroughly streamlined protocol stack, will be the operating systems of the future. The web is going to be their API.
- New-generation browsers will allow industrial and consumer applications to run on resource-scarce edge nodes, in addition to more resourceful cloud and fog nodes. This will cause the "continuum of computing" to emerge as a seamless execution platform.
- Value-added meta-services will be realized as user-defined orchestrations of independent, specialized applications running on the continuum.
- Application orchestrators will be programmed with natural interfaces such as use voice command, textual natural-language specifications, and learning-by-example. Orchestration engines will serve the user's best interests in privacy, confidentiality, energy, social and ethical preferences.

Key recommendations

- Support the development of browsers that guarantee efficiency and fitness for direct execution on edge devices, and can sandbox portable value-added applications.
- Support the development of trusted orchestrators, loyal to their users, which can run on any mobile device and appliance.
- Support the development of "natural interfaces" for users to command the orchestrators running on their edge devices.

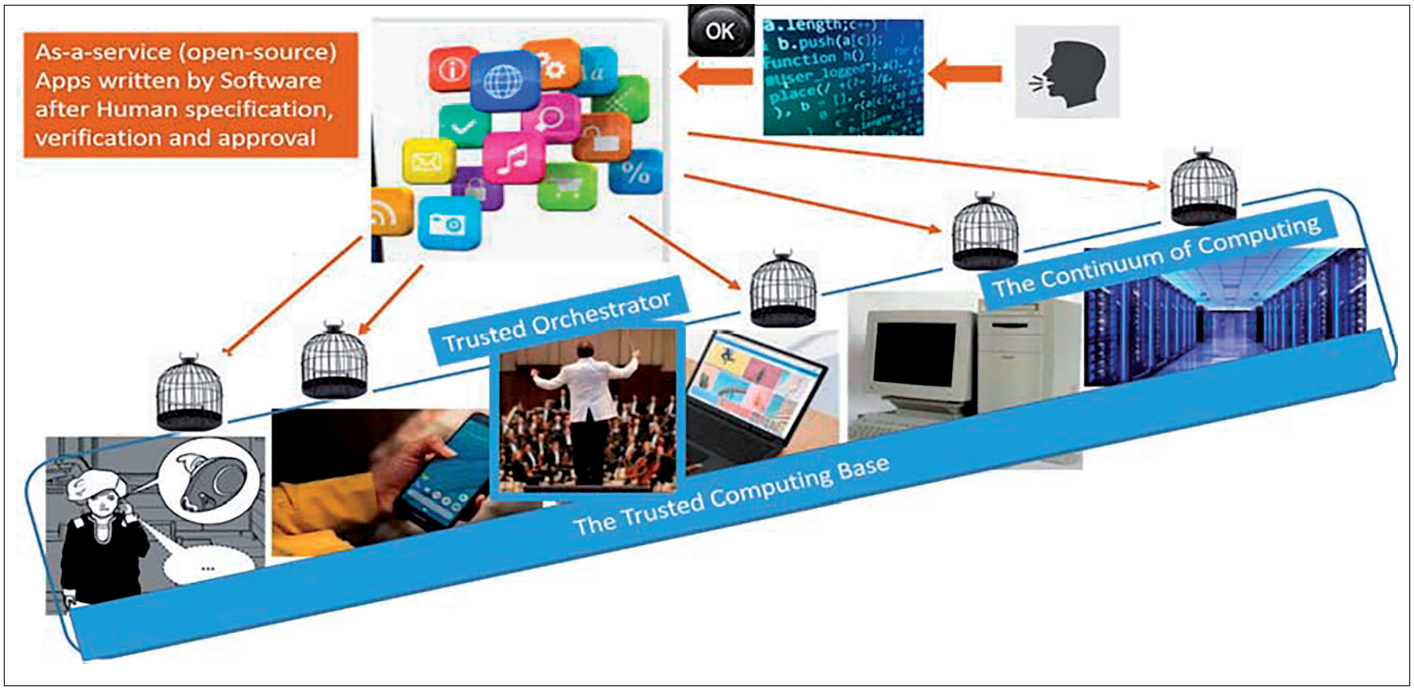


Figure 1: A pictorial view of the "continuum of computing"

Introduction

The notion of "continuum of computing" reflects the observation that edge, fog and cloud computing platforms are being pulled close together into what will likely become a seamless execution environment (depicted in the bottom part of Figure 1). This happens under the push exerted by a massively increasing number of value-added applications targeting mobile, handheld, wearable and unattended devices, which can all run **over one and the same base programmable interface**: the internet. Those applications serve either human users or industrial apparatuses, as in the IoT, with a wealth of functionalities (depicted in the top-center part of Figure 1). This trend is made even more prolific – and the value-added higher – by the fact that most such applications need not be deeply embedded in the target device. This trait accelerates their lifecycle many times over that of traditional device-specific deeply embedded applications.

Whereas this vision may be more immediately associated with the consumer market, it applies equally well to industrial scenarios. Such a notion reflects the opportunity to extend monitoring control of IoT applications to mobile users and the desire to seek low-latency use of mission- or business-critical services (e.g., deep learning) by deploying them at the edge.

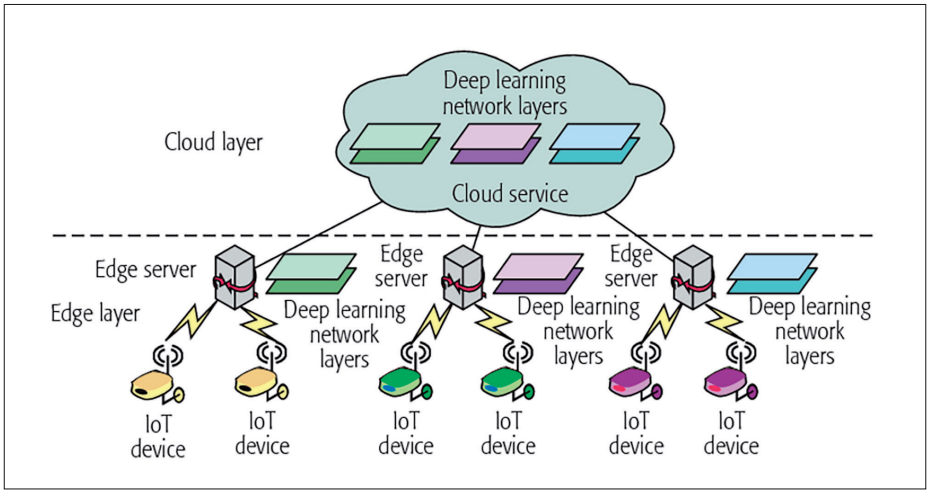


Figure 2: A view of the continuum applied to the IoT (Source: DOI:10.1109/MNET.2018.1700202)

There are two very contrasting possible versions of the continuum. One is captive and vendor-specific, and defers to the big giants of the internet, the only actors in that context that have the technical and financial resources required to cover the whole span of the continuum, horizontally (toward the user and the application developer) and vertically (toward the on-target runtime). The other is open and vendor-neutral; in line with the intended nature of the internet, it allows **interoperation** across all parties. It is this latter version on which we focus.

The intrinsic resource limitations of the near-user hosting devices (battery, storage, bandwidth, processing) require the applications delivering such functionalities to be designed so that, while becoming available on the target device, they should be able to offload the heavy-duty work to "collaborating" computing nodes. These nodes can be located opportunistically anywhere appropriate, from a fog node near the user to the deep centre of the cloud, as long as the user requirements continue to be met. Such a scenario has interesting, interwoven ramifications.

- The parts of the application nearer to the user, on the target device, will have an **as-a-service delivery mode**. The host environment of such a delivery mode can be very different from a normal commodity computer: it requires virtually nothing in the way of the traditional, file-system-based local installation; likewise, it does not need a large proportion of the operating system infrastructure. File systems have no practical value for resource-constrained devices, which can have other ways to arrange isolated storage for executing processes. Similarly, general-purpose operating systems, with their massive overhead and evolutionary clutter, are not very fit for purpose in this scenario.
- The collaboration between local and remote parts of the application, for delocalized data as well as distributed computation, will occur at the highest possible level of abstraction, which means at **the highest levels of the internet protocol stack** (HTTP/3 [1]). That choice reaps the largest benefits in terms of language-neutral expressive power (with architectural approaches that can be resource-centric with REST [2], data-centric with GraphQL [3], collaboration-centric with gRPC [4], real-time streaming with WebRTC [5], etc.), combined with portability and interoperability (stemming from relying on HTTP derivatives, the most standard and ubiquitous of APIs, outside of the boundaries of any given programming language). Moreover, as the endpoints of all such protocols are defined as schema documents, they can be the basis of **contract-based assertions and specifications**, set on the service interface exposed by the individual app parts. This prospect enables the construction of very advanced component aggregations, in which the application provider declares what the app can offer (aka "guarantees" in functional and non-functional terms) and under what conditions (aka "assumptions"). The app user can decide which app to choose from those that provide the same functional API.

The as-a-service delivery mode and the mobility of the computation across the edge-fog-cloud continuum extends the opportunity for deploying new services. The former requires app composition and integration to adopt open and efficient means for service discovery and service registry. The latter requires lightweight containerization, and advanced hosting and deployment engines. Interestingly, **base solutions to most such needs are already in place, for example in Kubernetes' ecosystem, originally born in the cloud but now being "miniaturized" to operate within smaller compute confines. Those solutions need to be evolved towards increased openness, agility and capabilities**.

Application hosting

The scenario described comes with distinct implications for the execution platform that hosts end-user applications. A most natural internet-enabled candidate for host for the on-target part of the application is a web browser, as opposed to a more classic general-purpose operating system, whose excess of evolutionary clutter makes it fatally unfit for purpose [6]. The prime reason for that candidacy is that modern browsers, most notably Chromium and its derivative Chromium OS [7], have learned very well how to:

- Efficiently separate their various activities, of which visual rendering is just one (extraordinarily sophisticated, but progressively less central as natural-interaction comes to the foreground – depicted in the top-right part of Figure 1). For obvious reasons of efficiency, all of this concurrent operation – where not flat-out parallelism – requires **self-served scheduling**, which does not have to be deferred to the untenably costly services of an underlying kernel-space operating system;
- Sandbox the web-hosted apps and plugins that the user may wish to use; realizing such sandboxing requires the host to expose standard APIs (which in turn facilitates the much-desired mobility of hosted apps, and amplifies their economic value) and to self-provide the most efficient form of **application-level containerization** that does not need running directly on a hypervisor;
- Natively **communicate with internet protocols over and above HTTP**, so that service invocation and service composition can all be uniformly expressed in terms of HTTP-based requests, independent of programming languages (hence maximally portable), attached to contract-based interfaces that augment their schema-based endpoint specifications.

Modern web browsers of this kind are the operating systems of the future, especially in the segment of the continuum exposed to the user, because they do everything that really matters to respond to the user application needs, viz., safe componentization (aka sandboxing), scheduling, high-level inter-process communication, and life-cycle management (including installation, update, removal, all without the need for local file systems, but merely using simple storage management).

Interesting advancements toward the secure and efficient deployment of sandboxed mobile code within such browsers are being made [8]**: their base technology needs to be trialled more extensively, hopefully with moonshot projects, to determine how to bring them to maturity.**

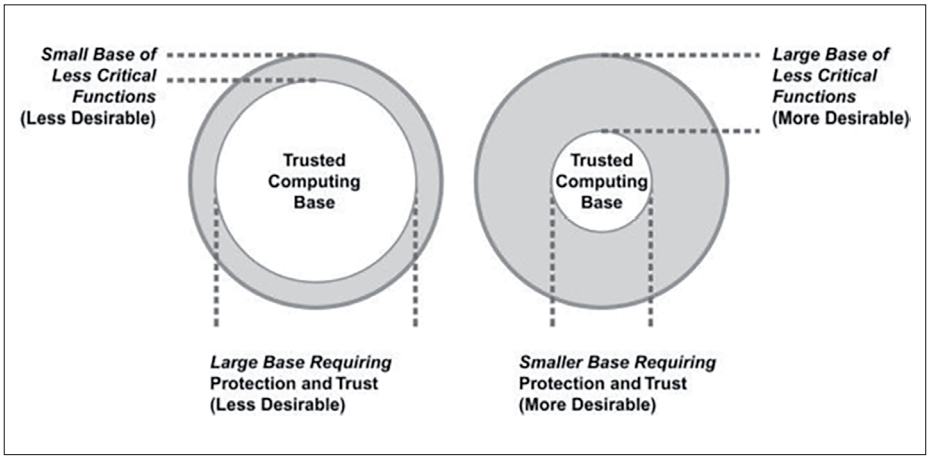


Figure 3: The change of landscape in the software computing stack.

Deployment engine

The observations above posit a radically different landscape in terms of software computing stack from the application to the deployment and execution engine at run time, as Figure 3 attempts to portray. The new landscape (on the right) determined by the previously described scenario will not push out and entirely replace the old, traditional one (seen on the left), but rather will coexist with it. The former will be fitter for the near-to-user part of the continuum, where there is more room for disruptive innovation; the latter will constitute the legacy, closer to the periphery of the cloud, where the ecosystem is more stable and tradition is a helpful convenience. The cloud itself uses a different structure, which employs virtualized or native containerization to modularize the application space, and strives to separate infrastructure management, application management and service delivery.

- The traditional structure on the left has a large base – the operating system, with its own vulnerabilities, and the applications with theirs – requiring protection and needing to be assessed for trust when used in critical settings. This verification is very difficult and costly to achieve, because of the highly-coupled composition of the software stack (e.g. general-purpose operating system with traditional process-based isolation, resident applications that scarcely separate data from code, ad-hoc inter-process communication).
- The wholly novel structure on the right is the opposite: it has a much smaller, leaner, sleeker base that has to be made trusted, which we call here Trusted Computing Base (TCB) and liken to the core of a modern web browser. It also has a large base of less critical functions (sandboxed applications – depicted in Figure 1 as the cage that encapsulates the apps, above the interface to the continuum platform), which can be very user-need-specific, short-lived, and free to fail, incurring local service disruption, but without causing ripple effects.

On its inside, the TCB may be regarded as a message-based micro (but not minimal) kernel that is oriented to supporting web services and apps. All software execution within it obtains sandbox isolation in four ways as described below, without using costly privilege levels:

- The kernel being written in memory-safe programming languages (e.g., Rust [9], Ada SPARK [10]), whose strict memory model uses ownership tags to control the lifecycle of and the access to program data, without requiring garbage collection;
- The extensive use of static verification tools that accompany such languages (rendered unprecedentedly easier to produce by the safeness traits of the language itself);
- The execution by interpretation or just-in-time compilation using languages such as Web Assembly or WASM8. WASM is attractive in this particular context for various reasons. Firstly, it warrants rigid memory separation between executing modules and strict separation between code and data to prevent execution breaches via corrupted data, and it solely allows structured-control-flow instructions, to prevent uncontrolled jumps. Secondly, it defines a Web Assembly System Interface (WASI) to support calls across heterogeneous platforms. Thirdly, it uses a capability-based mechanism to control execution access to external resources;
- The sandboxing of WASM applications within the WASM interpreter, and its hosting in a kernel-level process that acts as resource broker to it. The execution of the hosted (aka target) process follows the principle of least privilege, to reduce the surface of attack to the maximum possible extent, and uses message-based communication, in keeping with the share-nothing principle, prerequisite to robustness, scalability and parallelism at various levels of software system infrastructures.

Two notable conceptual precursors to the TCB described here are especially pertinent. Singularity [11] was established as a proof of concept at Microsoft Research in 2003, built with proprietary technology, and experimentally released between 2007 and 2013, to explore brand-new innovation in the architecture, build and execution of operating systems, and fits the vision we describe in this article. Fuchsia [12] is an open-source project launched by Google, which has not yet reached an official release date but has resonated well with the public given its inspiring principles, which are very much in line with the four points in the list above. **Neither of these technologies constitutes, per se, a self-sufficient stepping stone; they should be seen as**

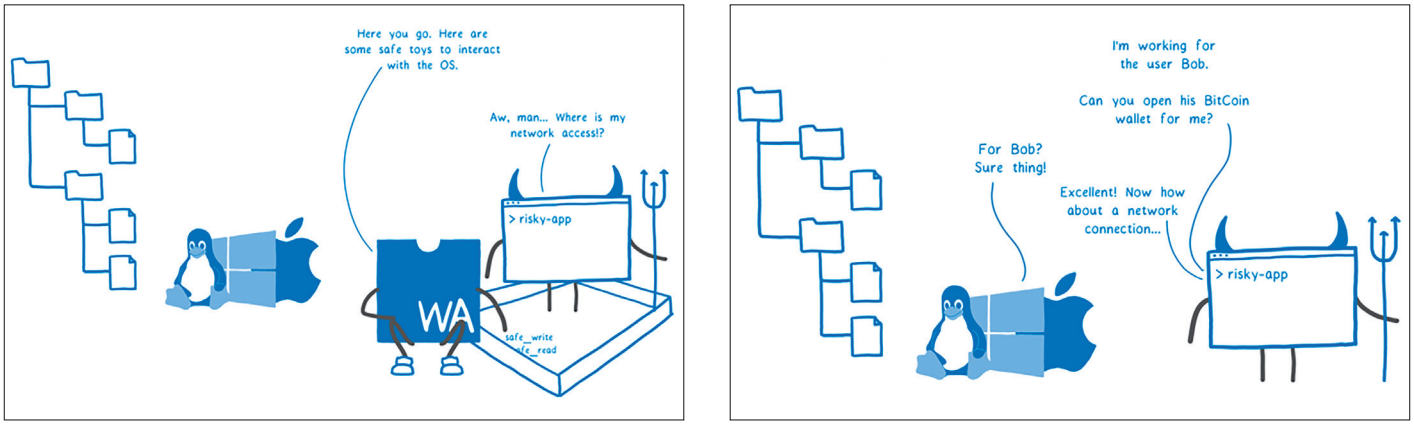


Figure 4: Sandboxing (right) made easy, contrasted with normal hosting (left).
(Source: https://hacks.mozilla.org/2019/03/standardizing-wasi-a-webassembly-system-interface/.)

evidence that the required direction can be pursued. Development from scratch of a demonstrator TCB, no larger than a compact and streamlined mini-kernel, would be a useful tool on the path towards the continuum.

Value-added end-user applications: service orchestration

With the large amount of value-added, prevalently independent, individual applications that are candidate for running on the TCB in the novel scenario, the opportunity arises to promote and sustain "non-anticipative evolution" [13] at the runtime platform level. This concept suggests that the complexity of the modern world (in terms of needs, expectations and potential) is better served by allowing adaptive openness that follows second-order user requirements (those that reflect actual user experience as opposed to prescriptive anticipation of it) than by platforms that anticipate them placing arbitrary constraints on the user (be it human or application). One illustration of how non-anticipative value added can be obtained, elegantly and agilely, from the software stack on the right in Figure 3 is the provision of an orchestration engine. This engine would allow individual users (humans or applications) to specify and launch a flow (presumably captured as a directed acyclic graph) that executes selected apps in some defined order and determines how the corresponding output of one should become the input of another.

Such an orchestration engine would have two complementary parts:

One, at the user side, should support "natural interfaces" (e.g. voice-commanded specification – depicted in the top-right part of Figure 1) for the simple yet compelling reason that producing (as opposed to using) such orchestration flows should be within the reach of every individual and not only of trained programmers. This is currently not the case, which is hardly justifiable, either technically or conceptually. An excellent example of this striking deficit is the current radical separation of coding and use of the so-called "skills" of Amazon Alexa (now renamed Echo Dot). With this device, no matter how attractive and satisfactory the deployment of such skills would be to the user, their programming is still a specialized human job, rendered so by an awkward low-level programming interface (to produce very trivial programming directives, in fact). The reason for this is not technical; it is more likely related to market economics (presumably, creating the technology that enabled voice-commanded specification of skills would have delayed the launch of the product. What was – and is – important in the consumer market for a producer is to be first wherever innovation is, which frequently causes corners to be cut in the race) and **should be overcome thanks to the maturation of speech-to-code technology** [14] **or spec-to-code or by-example-code and all possible instances of "natural interfacing" between humans and computers**. These things favour freedom from the obligation of writing arbitrarily idiomatic program text. Such a component of the envisioned orchestration engine should help the user manage the lifecycle of "flows", enabling them to be created both offline (without requiring immediate execution) and online (with immediate execution, perhaps even in place of or in addition to other executions), as well as to be inter-

rupted, modified, resumed or stored. All of that should be equally doable with new-generation natural interfaces (primary) and with more traditional programmatic solutions (secondary, back-office style). **Seen at a wider angle, the envisioned form of application coding, which can be dubbed "software-writing-software", is becoming increasingly attractive and rendered possible – at boosted capacity – by deep learning compute infrastructures. Such infrastructures can be trained to understand an enormous base of example code, and directed to generate code artefacts that have the required traits, functional and non-functional, regardless of the target programming language of choice or necessity**.

The second part, on the execution platform side, is a trusted orchestrator, a prominent element of the envisioned TCB, situated on its upper interface (depicted in Figure 1 as the picture box showing an orchestra director, positioned at the centre of the TCB interface exposed to the user side), which executes service calls directed to independent service apps, aggregated according a user-specified flow. The term "orchestrator" is used here to evoke the conductor of an orchestra that obtains the desired musical effects from commanding into action individual instruments that act independently according to their own music sheet. The particular orchestrator that we envision executes such workflows in a manner that is loyal to the user, for choice of applications (where multiple alternatives can provide the same service), preservation of data privacy (by controlling what data is extracted, directly and indirectly, from the execution and where they are stored), and social or ethical preferences (e.g. open-source versus proprietary; closer versus anywhere; run on green computing versus carbon-based). Workflow execution engines are commonplace in business process management [15]; this vision statement takes them to the next level in two ways: one is making them the prominent way of deploying applications at the outermost level of the software computing stack, recognizing that *digitalization is fundamentally about the composition of third-party functionalities across the network*; the other is adding "loyal intelligence" to their operation, which is one essential ingredient of the "ethical computing" that must accompany the digitalizing-of-everything in order for it to be human-centred. We are seeing the arrival of **programming languages that operate at the level of abstraction as described above (i.e. over the internet, which effectively takes over the role previously played by the middleware), and which are at a stage of their development that allows them still to be augmented with the missing features**. Such languages are often called cloud-native, to signify that their technology contemplates all of the steps entailed in going from traditional compilation and local execution to enabling containerization, deployment in a container orchestration framework, and distributed non-local execution. The most distinctive features of such languages are that they (1) are designed to be integration languages as opposed to system ones; (2) allow for multiple implementations according to the runtime platform that is to host them; (3) support network-friendly types for data, commands and style of interactions; (4) contemplate security abstractions; (5) favour static verification; and (6) enable powerful error treatment at run time. **Example languages that go in this direction include Ballerina** [16] **and Jolie** [17].

Conclusion

The "continuum of computing" is emerging as the confluence of several concurrent phenomena. The most prominent of them is that supporting the web (meaning its internet protocols) as the programmatic interface of modern applications means that it can be so thoroughly streamlined as to become fit even for resource-scarce edge nodes alongside the more resourceful cloud and fog nodes. In that sense, therefore, the continuum of computing emerges as a seamless platform where a multitude of user- and business-oriented web-based applications can be deployed and executed at will. The as-a-service delivery mode of most such applications favours their composition into user-defined orchestrations that operate as value-added meta-services. Such orchestrators may be very attractive vehicles of innovation in at least two respects. Their user side should allow the production of preference-based workflows of service calls with the lowest possible cognitive load for the user, with a shift from a programmatic to a natural-interface model that can use voice command, textual natural-language specifications, and learning-by-example support. Their execution engine should instead animate user-defined workflows singling out service providers via their schema-based interface contracts, deciding on their deployments, and obeying specified preferences that augment quality of service and take into account privacy, confidentiality, energy, social and ethical concerns. All of this is very much within reach from a technology perspective, and holds potential for significant innovation and advancement.

References

[1] Catalin Cimpanu, "Cloudflare, Google Chrome, and Firefox add HTTP/3 support", https://www.zdnet.com/article/cloudflare-google-chrome-and-firefox-add-http3-support/

[2] REpresentational State Transfer, https://restfulapi.net/

[3] GraphQL, https://graphql.org/

[4] gRPC, https://grpc.io/

[5] WebRTC, https://webrtc.org/

[6] "What is Google Chrome OS?", https://www.youtube.com/watch?v=0QRO3gKj3qw

[7] Chromiom, https://www.chromium.org

[8] WebAssembly, https://webassembly.org/

[9] Rust Programming Language, https://www.rust-lang.org/learn

[10] SPARK, https://www.adacore.com/about-spark

[11] Microsoft, 'Singularity', https://www.microsoft.com/en-us/research/project/singularity/

[12] Fuchsia, https://fuchsia.dev/fuchsia-src/concepts

[13] David Weinberger, "Everyday Chaos', https://www.everydaychaosbook.com/"

[14] VoiceCode, https://voicecode.io/

[15] Aleksei Kornev, "Why there are Cloud Functions and Service Mesh & whats next", https://itnext.io/the-concept-of-workflow-engines-c14e8088283

[16] Ballerina, https://ballerina.io/

[17] Jolie, https://www.jolie-lang.org/

Tullio Vardanega is associate professor in the Department of Mathematics of the University of Padua, Italy.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: T. Vardanega. The continuum of computing: enabling technologies. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 56-61, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

Artificial intelligence is improving every day, but it comes at a cost: more and more computing power

The omnipresent artificial intelligence

By MARC DURANTON

Artificial intelligence (AI) is everywhere in the news and is the next "big thing" that every company should (and will?) use. It allows the extraction of information from the "big data" that are collected everywhere and will hopefully make a variety of processes more efficient. It can be used in most domains, from medicine to energy, and from smart cities to transportation and manufacturing, and can even be applied to ICT. Officially born at a workshop at Dartmouth College in 1956, it had several "winters" before its last revival in 2012. It has continued to grow ever since. Initially used by web companies to verify content, AI has evolved: it now requires more and more compute power for the "big" systems and can also be used in deeply embedded devices.

Key insights

- AutoML (automated machine learning) is democratizing the development of applications using a deep learning approach, allowing non-specialists to develop their own applications.
- Simulation of virtual environments is starting to be used to generate data for training deep neural networks (for example, NVIDIA's Metaverse).
- Reinforcement learning approaches, or self-supervised learning techniques, do not need as much data for the learning phase; they only need a reward function or data that provide supervision. They will be more and more used in applications.
- The size of the neural networks that offer the best performance is increasing as a result of that performance, with the consequence that computing power is becoming less and less affordable and accessible and brings with it a sizeable carbon footprint.

Key recommendations

- It is necessary to continue improving the efficiency (in terms of both energy and cost) of the hardware, software and algorithms that are used by deep learning-based AI.
- Make the computational resources that are required for the learning phase of deep learning easily accessible to companies and academics.
- Research and development of tools that help to identify bias and misbehaviour of AI should be developed further.
- Research in Europe could focus on new approaches that do not require a lot of data for learning, and on federated learning approaches that allow privacy to be preserved.
- Europe could focus its efforts on near-the-edge and edge devices, leveraging its knowhow in cyber-physical systems and embedded systems.

The AI bandwagon

Artificial intelligence (AI) is currently marketed as an easy way to exploit big data and large computer infrastructure to improve business processes, with the promise of achieving optimizations to open up even unknown market potential. AI, and more specifically deep learning, was first a necessity for the major technology companies in the United States (Google, Amazon, Facebook and Apple - GAFA) and in China (Baidu, Alibaba, Tencent and Xiaomi - the BATX). For example, it allowed such companies to check if the millions of pictures uploaded everyday were "correct" (a typical Facebook deep learning use case). They have all the necessary resources: large and powerful computing infrastructure for learning and managing large databases, large sets of data and ways to attract the best scientists.

But AI is not new: the term Artificial Intelligence was coined by John McCarthy during a workshop at Dartmouth College in 1956 and developed in several directions, the two main ones being:

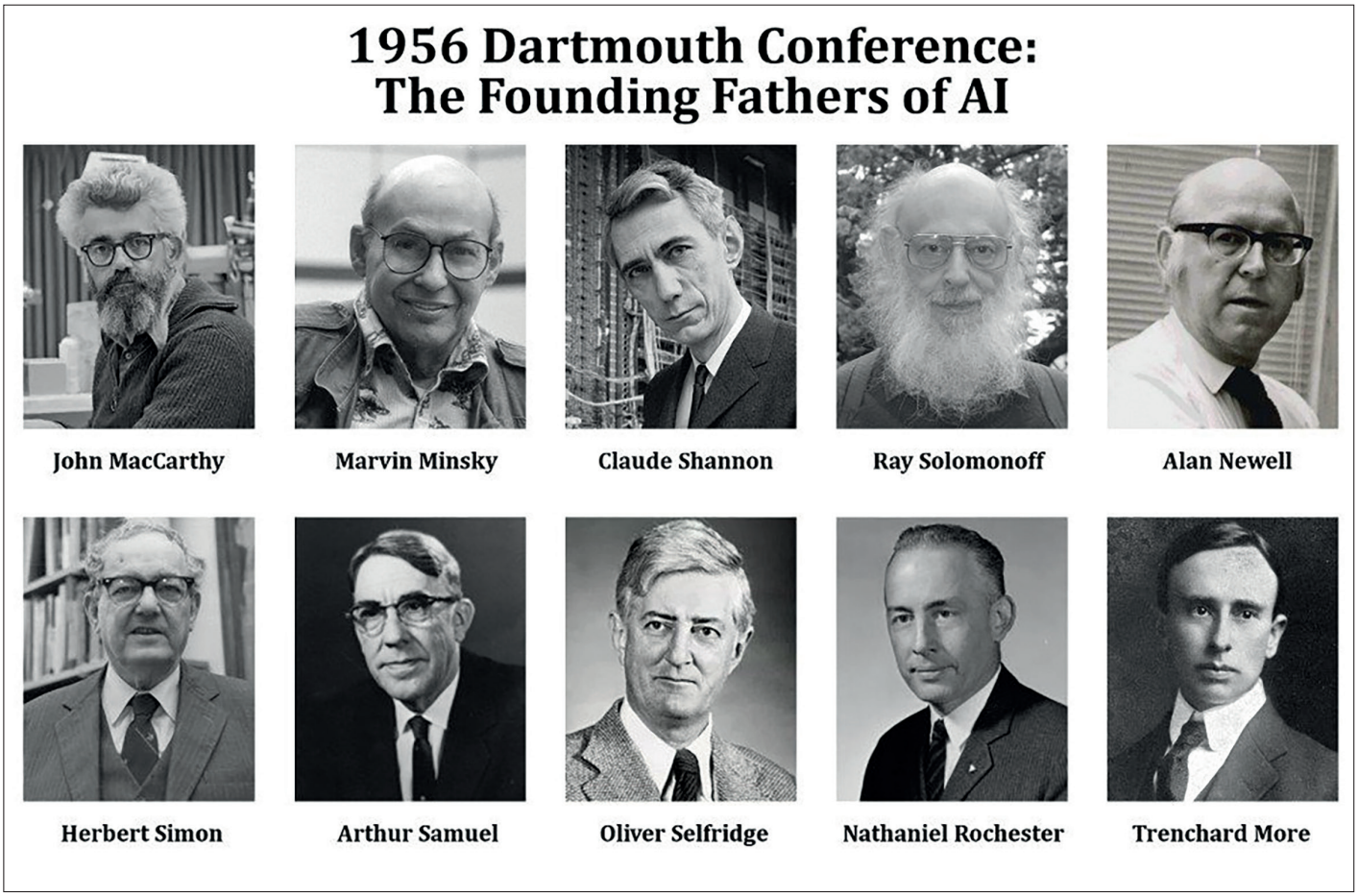


*Image: (Source: https://miro.medium.com/max/2400/0*8MW8iP2QC_WNhmiW)*

- Symbolic (or algorithmic) AI where high-level "symbolic" (human-readable) representations of problems are transformed by explicit (logic) rules, generally coming from humans (either directly, applying algorithms, or indirectly, being extracted and organized in expert systems, which use a network of production rules). The "deductions" of the AI system can be therefore followed and explained, but all the "rules" have to be put in the systems beforehand. As computers can process large sets of data quickly, it can create the illusion of "intelligence" because of the size and the speed it applies rules/algorithms to the data.
- "Connectionist" or "artificial neural network based" (now extended to deep learning) – AI. Here, simple models of the neurons of the brain – originating from the work of Warren McCulloch and Walter Pitts in 1943 – are assembled in networks. The "neurons" are connected by "weights" to simulate the biological synapses, and the purpose of the network is mainly a result of the various values of the synapses. During a learning phase, these weights are determined by repetitive presentations of patterns at the input of the network, and what it means at the output. These multiple presentations set the weights so that at the end, the network associates the desired output with the input. This is called "supervised" learning because the inputs are labelled with the correct responses the neural network should give at the output. The resulting neural network is not only an associative memory: it can also give similar results to input data that are "similar" (not too distant) and this is called generalization. It can then classify (during what is called the "inference" phase) input patterns it has never seen during the learning phase.

There are also other ways to train those neural networks without a supervised approach (for example, with unsupervised learning, where the network determines its output from different inputs which then do not need to be labelled and tries to automatically discriminate entries into different classes, or with reinforcement learning, which focuses on predicting a reward. Self-supervised learning is another of the possible options). Reinforcement learning was used to train the AlphaGo program and its successors, like Alpha Zero, which, in a few hours and without any knowledge of the field except the rules, beats all its predecessors (and humans) both at the game of Go and at chess. The latest update, MuZero, did not even have to be taught the rules; it observes the results of its actions and therefore improves itself. "The new system tries first one action, then another, learning what the rules allow, at the same time noticing the rewards that are proffered—in chess, by delivering checkmate; in Pac-Man, by swallowing a yellow dot. It then alters its methods until it hits on a way to win such rewards more readily—that is, it improves its play. Such learning by observation is ideal for any AI that faces problems that can't be specified easily" [29]. "MuZero algorithm, which, by combining a tree-based search with a learned model, achieves superhuman performance in a range of challenging and visually complex domains, without any knowledge of their underlying dynamics.

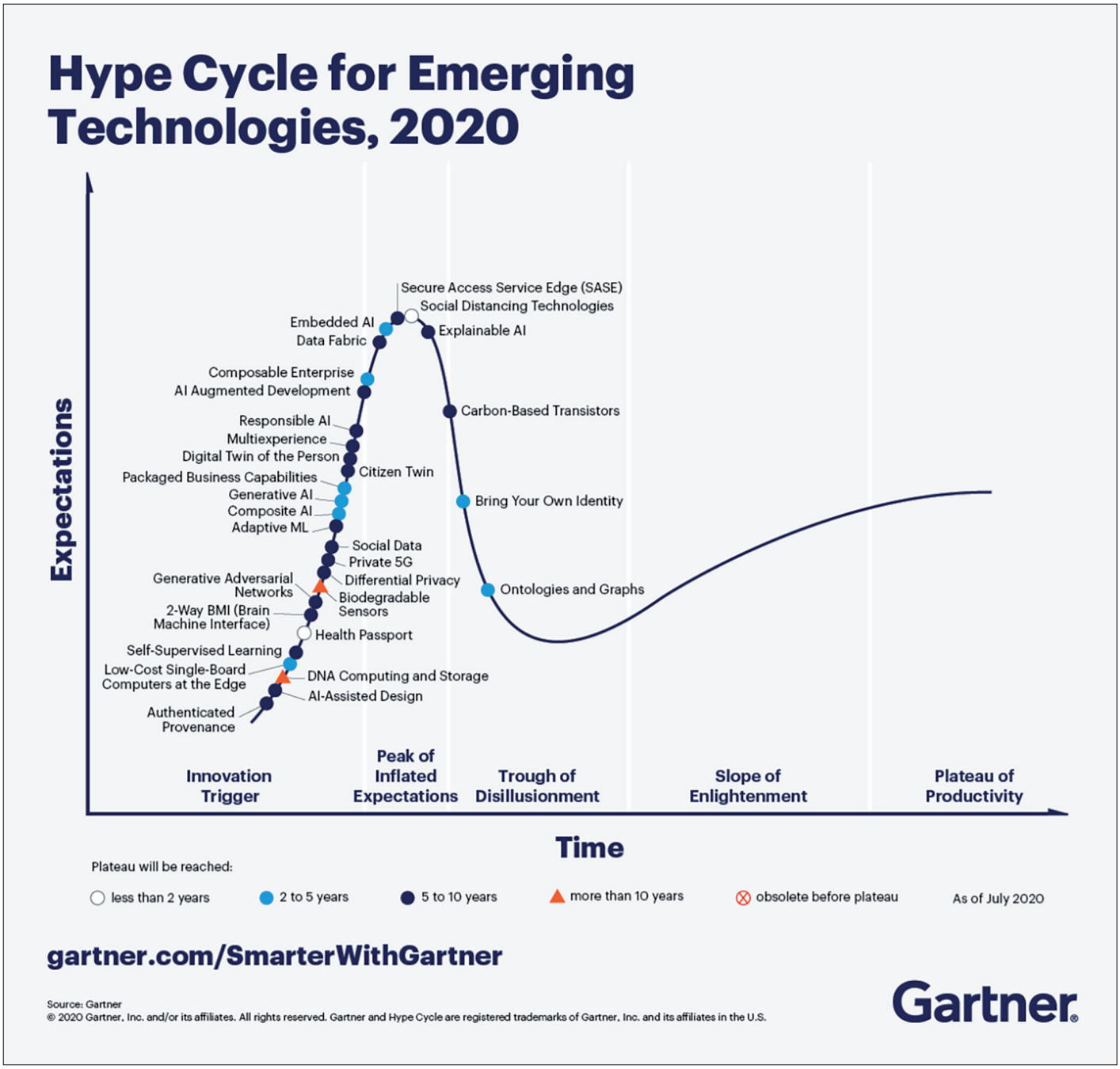


Figure 1: The 2020 hype curve, where AI techniques are very present [8]

The MuZero algorithm learns an iterable model that produces predictions relevant to planning: the action-selection policy, the value function and the reward" [28]. Other approaches are also being developed, such as generative adversarial networks (GANs) that put different networks in competition.

The renaissance of AI in 2012 was triggered by the superior performance of the deep learning approach (a special form of formal neural networks) for image classification, initiated by the work of Hinton et al. in that same year with their "deep" neural network they called "Supervision". As deep learning provides relatively good (or, at least, good enough) results when applied to various application domains, with relatively low human effort, it has really taken off since then; now it is at the top of the curve of expectations. From a marketing point of view, companies feel obliged to apply these technologies to their products to keep up with the trend. Even methods and approaches that have been used for some time are now jostling for position under the umbrella of "artificial intelligence". As it does every year, Gartner has published its "Hype Cycle for Emerging Technologies", and, without surprises, AI is very present in the 2020 edition.

Deep learning provided breakthroughs in terms of analyzing unstructured data such as images and sound, as well as allowing an efficient interface between computers and the world, facilitating cyber-physical applications. This has really opened up possibilities for new solutions and business propositions, like self-driving cars, personal assistants and so on.

A brief history of Deep Learning (From HiPEAC vision 2019, pp.23-24)

Throughout history, people have sought to make machines that amplify their physical, then mental abilities. The brain was not always considered the centre of intelligence: Aristotle believed that it was only used to cool the heart. However, the approach advocated by Plato, Hippocrates and Democritus, for whom the brain was the centre of awareness of sensations and the guardian of intelligence, finally prevailed and many generations of researchers have sought, and still seek to analyse its functioning. The idea of imitating it to make "intelligent" systems is not new, but it was the discoveries of the 20th century that triggered the first results.

Drawing on the knowledge of biologists of their time, in 1943 Warren Sturgis McCulloch, an American neurologist, and Walter Pitts, a mathematician and psychologist, proposed a mathematical model of the simplified functioning of biological neurons, cells which form one of the components of the brain. Their paper, "A Logical Calculus of Ideas Immanent in Nervous Activity", was published in 1943 in the "Bulletin of Mathematical Biophysics" (5:115-133) and remains the basis of formal neural networks. Their model is simple: a neuron performs a binary function that compares the weighted sum of its inputs (connected to the other neurons) to a threshold.

They have shown that a sufficiently complex network can "calculate" any function. John von Neumann, whose "First Draft of a Report on the EDVAC" is considered to be the first description of a modern computer (von Neumann's machine) cites only this McCulloch and Pitts paper in this 1945 report and infers from McCulloch and Pitts' article that *"everything that can be described exhaustively and unambiguously... can be conceived as an appropriate neural network*". It confirms that a neural network can represent a universal Turing machine, and therefore a universal calculator. Unfortunately, the limitations of the technology of the time did not allow him to develop the highly parallel approach of neural networks, and thus it resulted in an architecture with memory, a control unit, an arithmetic unit and input and output units, which are found in today's computers.

In 1957, psychologist Frank Rosenblatt invented an algorithm called a "perceptron". For this classifier, the weighting between neurons is inspired by the Hebb rule, which considers that when two neurons are excited together, their link is strengthened. The perceptron rule takes into account the observed error when propagating an input whose output function is calculated by the perceptron. The first winter of neural networks was caused by Marvin Minsky and Seymour Papert's book *Perceptrons: an introduction to computational geometry*, which shows limitations of perceptrons. The 1986 book *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* by David Everett Rumelhart and James McClelland relaunched the field with a testable approach of multilayer networks (essentially with an intermediate layer, called the "hidden layer") called multi-layer-perceptrons (MLPs).

A learning rule (called gradient retro-propagation) for calculating the weights of intermediate layers was published in his thesis in 1985 by Yann LeCun (now at Facebook), then widely distributed by David Rumelhart, Geoffrey Hinton (now at Google Brain) and Ronald Williams in 1986. This led to an initial explosion in the use of neural networks in the 1990s. They were used for the recognition of handwritten characters (to recognize postcodes), for image analysis etc. A first era of specialized circuit development followed, but the techniques of the time allowed only limited parallelism, and the rapid advance of general-purpose processors limited their expansion.

The uptake of support vector machine (SVM) then signalled the beginning of a new winter of neural networks by offering better performance than MLPs for image classification. The principles were explored between 1963 and 1970 by Vladimir Vapnik, but it was only in 1992 that an article by Boser, Guyon and Vapnik synthesized the results and allowed broad development of SVMs for classification.

Meanwhile, neural networks became deeper (with more layers), thanks to methods allowing the use of back-propagation approaches to gradient networks with more than one hidden layer. The networks became more complex, specializing the layers as in the visual cortex. The results of neuroscientists David Marr, David Hubel and Torsten Wiesel (the latter two were awarded the Nobel Prize in Physiology or Medicine in 1981 for their discoveries concerning information processing in the visual system) inspired researchers to make networks more suitable for object recognition. Their predecessor is the "neocognitron" invented in the 1980s by Kunihiko Fukushima. Deep convolutional networks as currently used are more than 20 years old, but thanks to the dramatic increase of data availability and computer power, more complex networks are now possible, which unlock a complete new range of performance.

The most recent renaissance was brought about in 2012 by Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton, who used deep convolutional neural networks for the ImageNet challenge, which consists in classifying images in the ImageNet image database. The Hinton Supervision Network beats the other approaches with an error rate of 15.3% compared to 26.1% for the second. From 2013, the top eight approaches in the challenge are based on deep neural networks. Indeed, deep networks are better than a human on this challenge (human level was estimated at 5% errors, according to the work of Andrej Karpathy), with less than 3.5% errors. The following table shows the very rapid improvement of deep learning algorithms, until being better than humans.

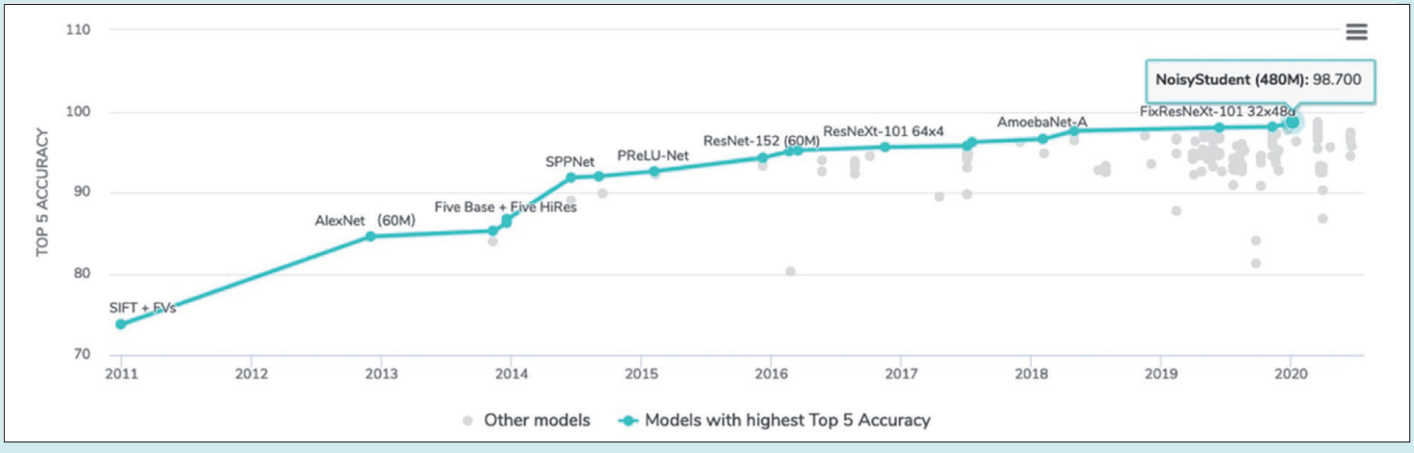


Figure 2: From [11]

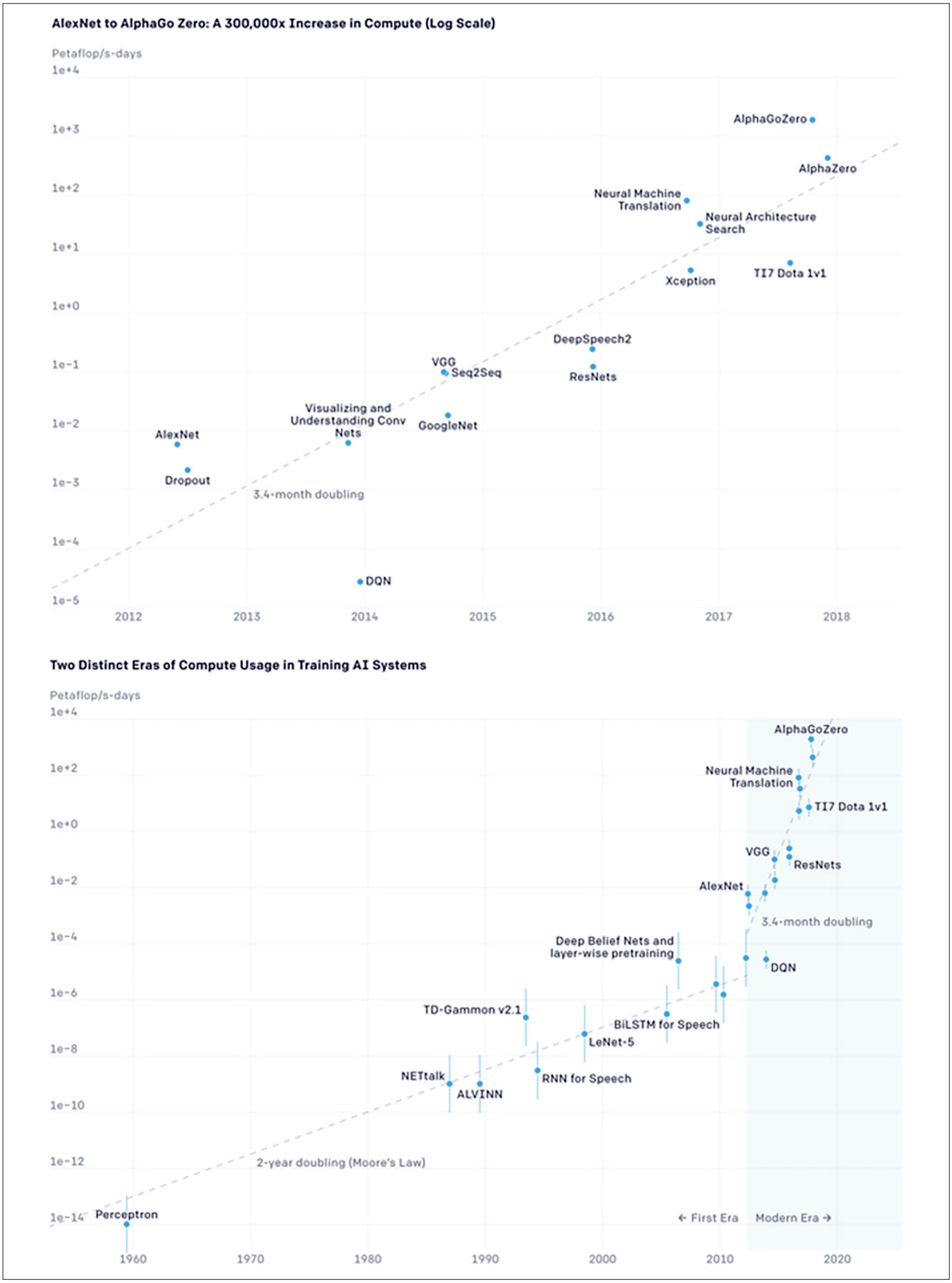


Figure 3: From [10]

The size (and compute requirement) of deep learning networks are growing exponentially

Thanks to the fact that a deep network is formed by learning and not explicitly programmed, it can be applied in many applications where it is difficult to explicitly define an algorithm, such as in image recognition (essential for autonomous vehicles), speech comprehension (all personal assistants, from Siri to Alexa or Google Now, use deep networks, often recursive), lip reading and participation in various games. A large "labelled" (indexed) database is all that is needed; these are often available from major internet players (Google, Baidu, Facebook, Microsoft, Apple, etc.), explaining why they conduct in-depth research on learning. For example, several billion photos pass every day through two types of deep networks at Facebook, Instagram, Messenger, WhatsApp for image/index recognition and facial recognition (although these are not enabled in Europe).

Networks and techniques are becoming more complex, combining several approaches. For example, the AlphaGo program developed by Google DeepMind beat Lee Sedol (a 9-dan professional in the Go game) in March 2016, generating a lot of publicity for deep learning and AI techniques.

In general, there are two phases in the use of deep networks: the learning phase, in which the network parameters (connection weights) are determined by the learning rule, and the inference phase in which the network is used to classify the data. The learning phase is the most demanding, with millions or billions of example presentations and changes in network settings. It is now generally done on 16-bit floating point graphics processing units (GPUs) (even if recent research seems to show that 4-bits might be enough [12]) or on specialized circuits such as Google's Tensor Processing Units (TPUs) [23]. The inference phase is less demanding and can be performed with less precision (integer, even reduced to 8-bits, or even lower, down to two states). It is usually this phase that is implemented in embedded devices for image recognition, etc. Synaptic weights are downloaded after learning and can be updated after a new learning, extending the number of recognized objects.

For example, Supervision, the network developed by the University of Toronto's Geoffrey Hinton and colleagues is composed of 650,000 artificial neurons connected by 630,000,000 shared connections (synapses). On today's networks, the learning stage could require a few exaFLOPS (more than a billion billion operations). Figure 3 shows the exponential increase of compute power required by the most advanced neural networks, roughly by x10 per year (to be compared to Moore's law – the observation that the number of transistors in a dense integrated circuit doubles about every two years).

This exponential growth* is confirmed in 2020, with GPT-3 [15], a system for natural language processing (NLP) using the "transformer" approach that has 175 billion parameters (the largest GPT-3 model – with 175B parameters – uses 96 attention layers, each with 96x128-dimension heads). *"GPT-3 175B model required $3.14x10^{23}$ FLOPS (so about 87h of an exascale machine – 10^{18} floating point operations per second) of computing for training. Even at theoretical 28 TFLOPS for V100 and lowest 3 year reserved cloud pricing we could find, this will take 355 GPU-years and cost $4.6M for a single training run. Similarly, a single RTX 8000, assuming 15 TFLOPS, would take 665 years to run*" [2].

However, GPT-3 shows amazing results in a lot of domains, from completing text, to translation or to even generating correct code from high level textural specifications [14].

"GPT-3 shows that language model performance scales as a power-law of model size, dataset size, and the amount of computation… GPT-3 demonstrates that a language model trained on enough data can solve NLP tasks that it has never encountered. That is, GPT-3 studies the model as a general solution for many downstream jobs without fine-tuning… GPT-3 175B is trained with 499 Billion tokens" [2].

Dataset	# Tokens (Billions)
Total	499
Common Crawl (filtered by quality)	410
WebText2	19
Books1	12
Books2	55
Wikipedia	3

We see now a race for bigger and bigger neural networks, Microsoft already released DeepSpeed, that "adapts to the varying needs of workload requirements to power extremely large models with over a trillion parameters while achieving near-perfect memory-scaling and throughput-scaling efficiency" [15]. DeepSpeed, using a machine with a single NVIDIA V100 GPU, can run models of up to 13 billion parameters without running out of memory.

We are even beginning to see research using deep learning approaches to create other, more optimized, deep learning networks. This is called AutoML. "AutoML open-source tools automate the entire life cycle of ideating, conceptualization, development, and deployment of predictive models. From data preparation through model training to validation as well as deployment, these tools do everything with almost zero human intervention" [16]. AutoML can facilitate the use of deep learning, allowing even non-specialists to develop advanced application quasi automatically. All major companies such as Amazon (AutoGluon), Google (cloud autoML) and Microsoft (Azure) are providing AutoML tools to their users. However, some techniques used in AutoML (like design space exploration) are also large consumers of processing power.

If AutoML is democratizing the development of applications using deep learning, the user still has to provide a large database of data to train the networks (if using supervised learning approach). This can be solved by creating data sets through simulation in a real-life like environment. For example, NVIDIA's Omniverse environment allows the creation of a photorealistic environment to train and simulate various AI [17].

* Since the writing of this article, Google Brain announced its Switch Transformer Language Model that packs 1.6 trillion parameters!

These virtual environments also allow the validation of AIs that use smaller data sets or only "reward" functions like for reinforcement learning or in self-supervised learning, which is a form of unsupervised learning where the data provides the supervision (for example, by withholding some part of the data, and the task of the network is to predict it).

It should be noted that, in terms of tools and software for developing deep neural networks, the major players in the field provide their development tools as free software. Examples include TensorFlow (Google), CNTK (Microsoft), DSSTNE (Amazon), Theano, Caffe (Berkeley) and Caffe2, Torch (Facebook) and PyTorch (Python interface), N2D2 (CEA), Torchnet learning modules, OpenAi Gym (Open AI), MXNet, etc. In fact, software is a non-critical element in creating an effective system of in-depth learning. A large database and neural network topology are the main ingredients: the value lies in the neural network topology and its weights, determined after learning on a particular database, not in the software that executes it – which should be optimized to use less resources and energy. But advances in models and algorithms allow for the reduction of the complexity of the neural networks; their improvement is therefore a key factor driving progress in AI: "Since 2012 the amount of compute needed to train a neural net to the same performance on ImageNet1 classification has been decreasing by a factor of 2 every 16 months" [3].

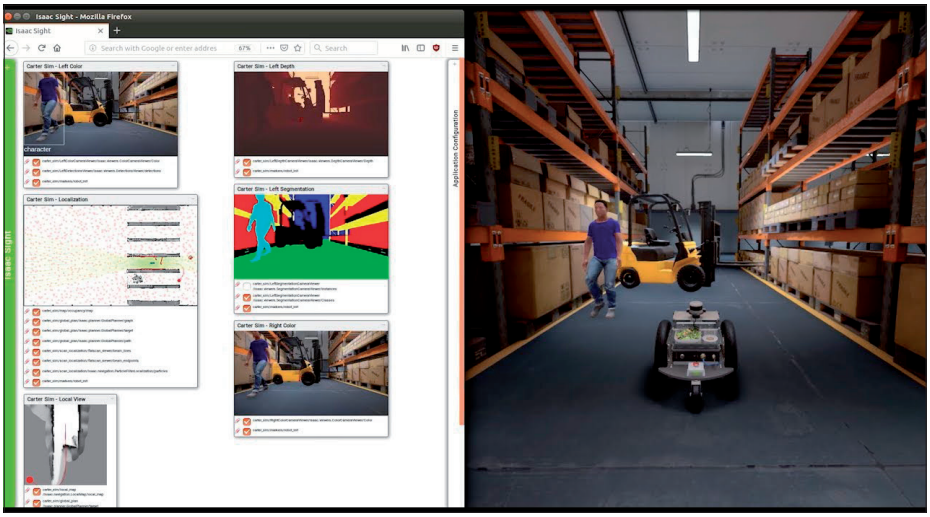


Figure 4: Nvidia Omniverse [6]

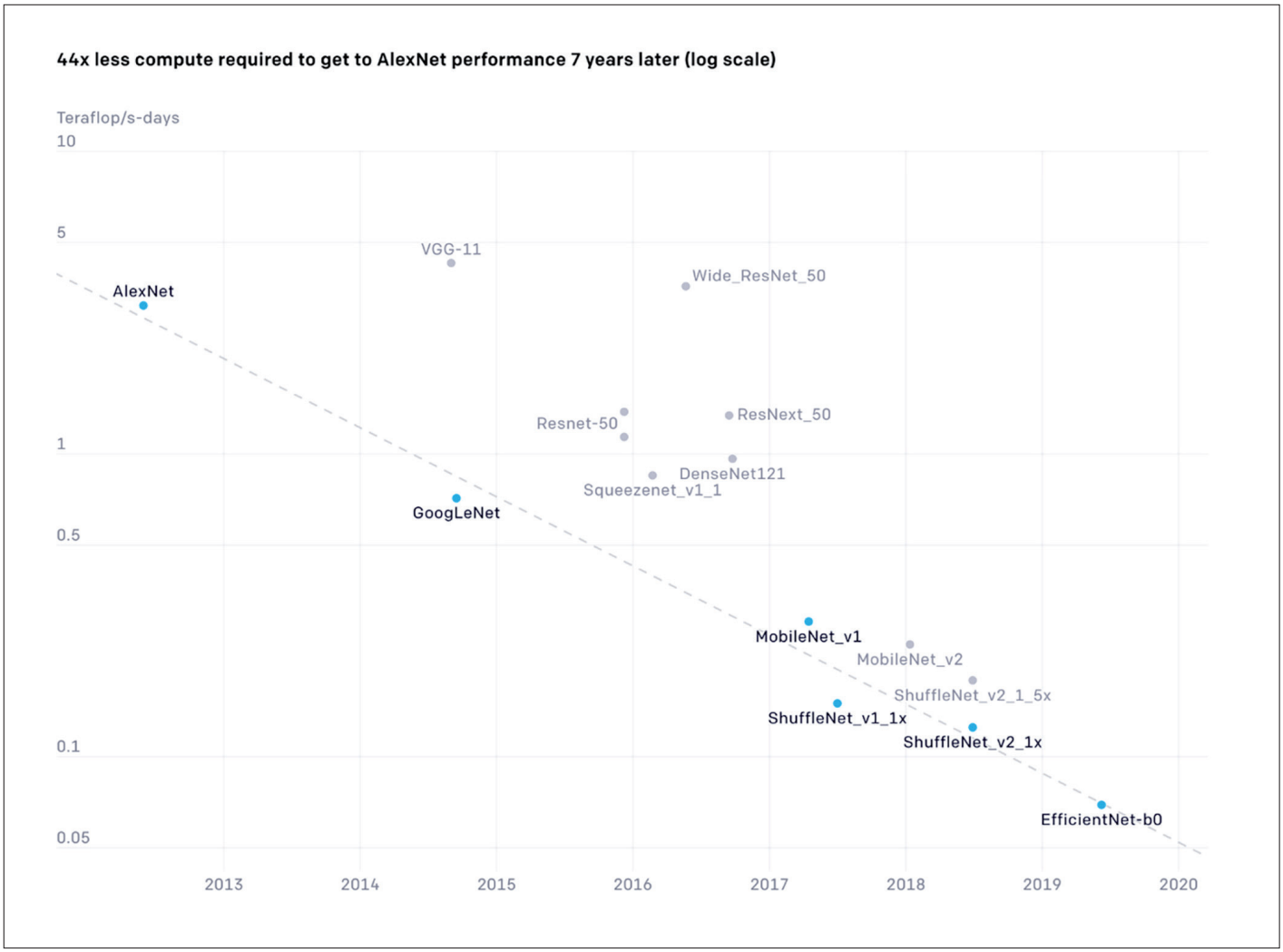


Figure 5: Decrease of performance needed to carry out a similar AI task over time [3]

However, a comparison of Figure 5 and Figure 3 clearly shows that the computing power requirement is still greatly exceeding the algorithmic improvement, and even exceeding the combination of algorithmic improvement and improvement of the hardware and software (better dedicated architectures, Moore's law, better integration processing and storage, better tools, etc.). This has several direct implications:

- It is necessary to continue improving the hardware, software and algorithms that are used by deep learning-based AI. Pruning networks, using computing with less numerical precision (decreasing the numbers of bits), or even using other ways to code information, like spiking networks (where the information is carried by a model of the "spike" signals like the ones used in biological neurons) are ways to improve efficiency. Using sparsity, and computing and moving data when it is really necessary, are approaches to improve the overall efficiency of computing systems. ("Spike" computing, even implemented in digital systems, is a way to realize sparsity of computation, because the information is coded "in time", and the more "spikes" that are processed, the more accurate the results. The system can therefore dynamically stop computing when accuracy is good enough to give meaningful results, unlike binary coding where all bits are always processed). Finding more or less automated ways to reduce the size of computation [12], using more efficient existing hardware [13] and optimizing networks to target architecture [19,20,21] form an active domain of research and development, together with the development of new hardware accelerators.
- It is also obvious that the energy consumption of those systems is also becoming increasingly problematic, as laid out in [26,27] ("Training a single AI model can emit as much carbon as five cars in their lifetimes"), leading to further need for improvement of the energy efficiency of deep learning systems. However, the learning is done once, and if the resulting system can be reused in inference millions of times and for several applications, the energy consumed during training should be offset by the gains the system brings during its use.

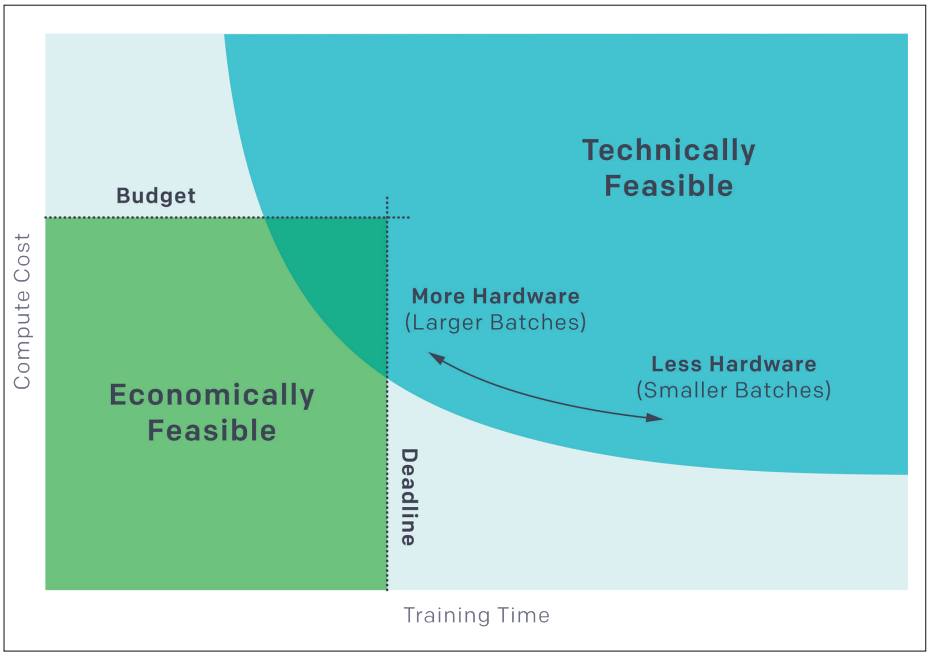


Figure 6: Increasing parallelism makes it possible to train more complex models in a reasonable amount of time. A Pareto frontier chart is the most intuitive way to visualize comparisons between algorithms and scales [4]

- The computing resources required for training those gigantic networks are no longer available to most research centres or industry settings. As of the end of 2020, the most powerful supercomputer in Europe – 7th in the TOP500 list – has a processing power of about 44 petaFLOPS (44.1015 FLOPS), so it would take about 82 days at full capacity to train GPT-3 (according to the figure provided by [2]). It therefore seems that, in practice, only the giant companies like Google, Amazon (with its AWS infrastructure) or Microsoft (with Azure) have the required processing capabilities in the western world. Research centres are either acquired (Deepmind is part of Google) or closely linked with those companies (OpenAi has a deal with Microsoft [22]). Moreover, the computing requirements of AI are so large that those companies are all developing hardware accelerators to boost performance. This was pioneered by Google with its line of TPU chips (they are on their 4th generation, which are on average 2.5x better performing than the previous generation [23]), followed by Amazon (AWS Inferentia [24]) and Microsoft (using Graphcore accelerators, but planning to develop their own chips as well [25]).

The level of investment required to explore the capabilities of huge neural networks, which indeed offer surprising even if not flawless [14] performance, is so high that it might limit their evolution to take place only within large companies or state entities. Not everybody is able to pay $4.6M for training GPT-3, or invest in the development of complex chip accelerators or farms of servers. The move of OpenAI from an "open" model to a commercial one with Microsoft is an example. The dilemma of computing power, time and cost is illustrated well in Figure 6.

Artificial intelligence: strategy for companies and for countries

As shown earlier in this article and in the HiPEAC Vision 2019 (pp. 21-20), artificial intelligence is becoming both a strategic asset for companies and a topic of sovereignty for countries. AI could create a "winner takes all" phenomenon: the first results from AI will result in economic benefits, then a quasi-monopolistic status because the AI-designed approach will allow greater profit margins, so that the winner could reduce prices and kill the competition until it achieves a monopoly. Countries like the United States, China and Japan are launching major AI projects, confident that new breakthroughs will

occur and will have a profound impact on our society in the years to come. Russia's President Putin has said the nation that leads in AI "will be the ruler of the world". China is making huge investments in AI and is feared by the US, which is also starting to invest heavily in AI. In contrast to China, in the US most activity around AI is currently undertaken by the major technology companies (GAFAM), which are also draining universities in the rest of the world of AI experts. A list of AI initiatives in different countries can be found in [7].

Besides being strategic for countries, AI is also a growing market for business, and IDC forecasts that "worldwide revenues for the artificial intelligence (AI) market, including software, hardware, and services, are expected to total $156.5 billion in 2020, an increase of 12.3% over 2019. [...] A new forecast from the IDC Worldwide Semiannual Artificial Intelligence Tracker shows worldwide revenues surpassing $300 billion in 2024 with a five-year compound annual growth rate (CAGR) of 17.1%" [30].

Edge AI and federation of devices: the position of Europe

What could the position of Europe in this international landscape of AI be? Europe does not have the big B2C companies that can harvest large quantities of consumer data and develop large computing infrastructure. However, it has several assets: its educational system trains good researchers and specialists in AI (who unfortunately, are often attracted to work outside Europe (brain drain) because of better working conditions such as access to computing power, databases, etc.) and its knowhow in cyber-physical systems and embedded systems. Therefore, if Europe reacts fast because the competition is fierce, it can be leading in artificial intelligence at the edge, leveraging its assets in embedded systems. In the short term, it could use corporate or institutional data (from cities, electric and water networks, hospitals, etc.) to optimize its processes and manage solutions and services at the edge. It can also apply its privacy-related ethics to AI systems that do not centralize private data on cloud servers, but are based on a federation of smaller and distributed systems. "Federated Learning is a machine learning procedure where the goal is to train a high-quality model with data distributed over several independent providers. Instead of gathering the data on a single central server, the data remains locked on their server, and the algorithms and predictive models travel between them" [31].

Figure 7: International strategies on AI [7]

The recommendation of HiPEAC Vision 2019 still holds in 2021:

If one day, artificial general intelligence becomes a reality, and if that artificial general intelligence is more powerful than human intelligence, Europe will only be able to compete with the rest of the world by building ever smarter computing systems. Instead of the war for talent (fought by companies, universities and countries), in order to improve competitiveness, Europe will have to invest in intelligent systems that will help create better products and do better research. There is a belief that "the future information society will not be built on human brains but on artificial brains". The societal values of Europe should be built into systems, in order to ensure its future existence.

HiPEAC Vision 2019, p.22

According to a prediction by Forbes, the total number of academic research papers published on federated learning will surge. "Data privacy is becoming an increasingly urgent issue for consumers and regulators. Given this, privacy-preserving AI methods will continue to gain momentum as the most sustainable way to build machine learning models. The most prominent of these methods is federated learning. The number of academic research papers published on federated learning has grown from 254 in 2018, to 1,340 in 2019, to 3,940 in 2020, according to Google Scholar. This exponential growth will continue: in 2021, over 10,000 research papers will be published on the topic of federated learning" [18]. Europe has without doubt a central role to play in this field.

Research in Europe could also focus on new approaches that don't require large volumes of data for learning, such as self-supervised learning, reinforcement learning or systems like MuZero that learn by observation [28].

In terms of hardware, Europe could focus its effort on near-the-edge and edge devices, relying on its semiconductor manufacturers and start-ups that are well positioned in the embedded space. Systems for self-driving vehicles or ADAS, or for edge computing have a large and very diverse market, requiring a multiplicity of different uses and modes of application. The landscape is large, as shown in Figure 8.

Explainability and/or transparency of AI: importance of ethical AI

One of the main complaints about machine learning, particularly deep learning, is that its models are opaque, non-intuitive, and difficult for people to understand and that the machines are unable to "explain" their decisions, leading to a lack of confidence and trust in the system. Its results can also be totally different when altering just a small part of its input, such as a few pixels in an image (this is called "adversarial attacks"). This is an important problem, leading to new kinds of piracy tuned to this kind of processing. To address this, there are few developing fields of research concerning deep neural networks:

- Explainable AI (especially for deep neural networks);
- Creation of robust solutions which are impervious to deliberately introduced fake data;

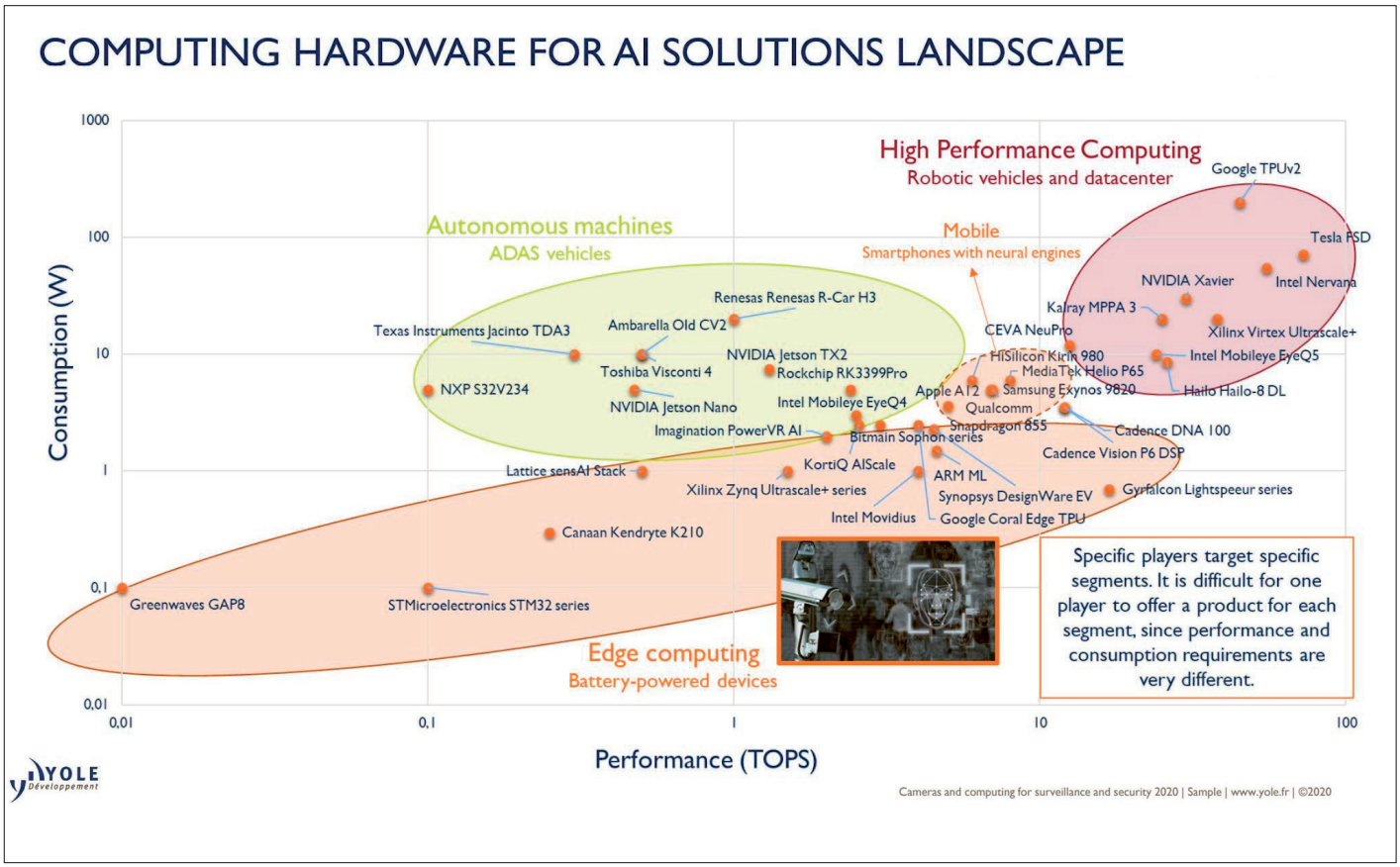


Figure 8: The chip landscape [9]

- Detection of bias in the learning data sets that could lead to unethical responses of the deep learning systems;
- Design of systems that can assess the likelihood of their results.

The explainability of the results of deep learning is an important topic in terms of the acceptability of technology solutions, but should not be taken too far; by way of comparison, there are a number of industrial processes that are not fully understood but this does not prevent them being used in everyday life. Processes of certification and validation can be used to ensure a certain "trust" in the system without everything about it being understood. The process of how the system is built and tested in the field might be enough in certain cases (for example, you trust a taxi driver because he has a licence, and you do not have to open his brain to see how it is working). What is more important is to ensure that a deep learning neural network effectively learns what it is supposed to learn, and not anything else.

It is therefore very important to develop approaches and tools to check that the learning databases are free from bias, whether introduced deliberately or accidentally. This is perhaps easier to achieve than full "explainability" of deep learning decisions, which is important but will be difficult to achieve without clear breakthroughs.

The most important points are to ensure that the specifications that lead to the learning databases are as complete and exhaustive as possible, with minimum bias, and to check after learning that the system has learnt only what it was supposed to learn, rather than other artefacts present in the learning database. It is also important to expose the system to counter examples, that is, things it should not do. Most of the time, designers focus on what the system should do (recognition rate) and not on what it should not do (false alarm). Sometimes, modern databases are "too good", with only clear images, making the system more sensitive to noise or other artificially introduced artefacts. With this "classical approach" of supervised learning, humans are still in the loop and ultimately responsible for the design of the learning database, therefore also responsible for the resulting deep network and what it will do in the inference phase. Research and development of tools that help to identify bias and misbehaviour of AI should be developed further.

Another idea to consider is that of not using deep learning alone for a task, but combining several approaches (including other deep learning solutions or symbolic ones) and adding a kind of supervisory process that checks whether the results are coherent. If fed with totally random input, a deep neural network will still give results, which are of course not relevant. The system should be able to indicate if its results are relevant/correct or not, and be able to generate an "I don't know" answer.

Legal liability in the event of an AI system failure is important, but, in the case of deep learning, it applies more to the initial specifications and definition of the learning database (done by humans – at least for now) than to how the deep learning system works by itself. Because humans are at the origin of the algorithms or the learning database, they should ultimately be held responsible in the event of errors by artificial intelligence. The main problem will be to identify potential errors, and to be able to correct them.

Needless to say, as shown in the Figure 9, humans are also prone to error (such as optical illusions) and sometimes have difficulty checking whether a system is unrealistic.

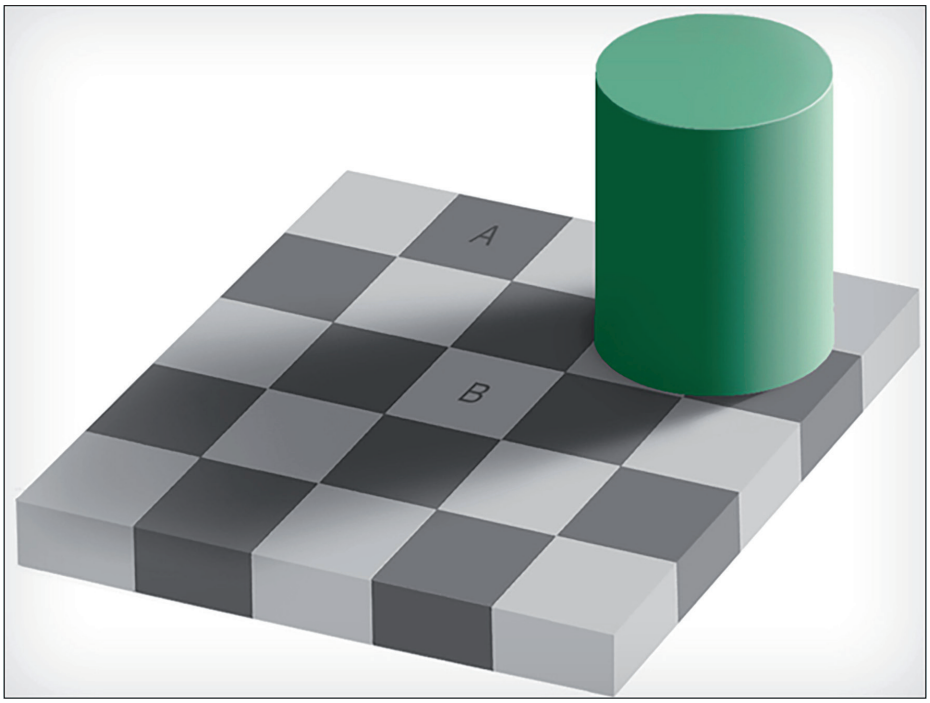


Figure 9: Unlike what it seems, the checkerboard boxes A and B are of the exact same shade of grey: see below

THE 5TH RESEARCH PARADIGM?

The convergence of simulation, machine learning and knowledge allows the emergence of a 5th paradigm in science and technology: as explained in the HiPEAC Vision 2019: "The first three paradigms were experimental (empirical description of phenomenons), theoretical (discovery of laws, models, etc. able to predict results) and, more recently, computational science (computer simulations). The fourth paradigm of scientific discovery is the analysis of massive data sets, enabled, e.g. by data capture, curation, mining and analytics techniques and thus permitting new scientific discoveries.

In the fourth paradigm, computers are used to extract information from raw data, but it is still humans who perform the analysis of the information and make the scientific discovery. We believe that within the next decade there will be a fifth paradigm, in which computers will be not only extract information from data, but will also formulate a hypothesis, design new experiments and simulations or make a formal proof and finally make scientific discoveries without human intervention. We already have examples of this with formal provers, data analytics, and approaches like IBM's Watson. Potentially, the Ultra-Intelligent machine could solve problems that are beyond the reach of human intelligence"

HiPEAC Vision 2017 pp.59

References

[1] Sally Ward-Foxton, "Arm Leaps Into TinyML With New Cores", 2020, https://www.eetimes.com/arm-leaps-into-tinyml-with-new-cores/

[2] Chuan Li, "OpenAI's GPT-3 Language Model: A Technical Overview", 2020, https://lambdalabs.com/blog/demystifying-gpt-3/

[3] "AI and Efficiency", 2020, https://openai.com/blog/ai-and-efficiency/

[4] "How AI Training Scales", 2018, https://openai.com/blog/science-of-ai/

[5] https://codecarbon.io/

[6] https://www.youtube.com/watch?app=desktop&v=jtMoxUyPPXk

[7] "The 2020 AI Strategy Landscape", 2020, https://www.holoniq.com/notes/50-national-ai-strategies-the-2020-ai-strategy-landscape/

[8] Alexandra Patard, "Les 5 tendances technologiques émergentes en 2020 selon Gartner", 2020, https://www.blogdumoderateur.com/gartner-hype-cycle-emerging-technologies-2020/

[9] https://1.bp.blogspot.com/-EqHDWJT3G88/X2OvK14CfrI/AAAAAAAAgns/18TZfBtyYS09j6zGxo4DKBmSPudR9oYuACLcBGAsYHQ/s2048/Yole-5.JPG

[10] "AI and Compute", 2018, https://openai.com/blog/ai-and-compute/&#addendum

[11] "Image Classification on ImageNet", https://paperswithcode.com/sota/image-classification-on-imagenet

[12] Xiao Sun at al, "Ultra-Low Precision 4-bit Training of Deep Neural Networks", NeurIPS 2020, https://papers.nips.cc/paper/2020/file/13b919438259814cd5be8cb45877d577-Paper.pdf

[13] DeepSpeed team, "DeepSpeed: Extreme-scale model training for everyone", 2020, https://www.microsoft.com/en-us/research/blog/deepspeed-extreme-scale-model-training-for-everyone/

[14] Anushka Sandesara, "Meet GPT-3-No need to Code!", 2020, https://medium.com/analytics-vidhya/meet-gpt-3-no-need-to-code-afb5924fd864

[15] Tom B. Brown et al, "Language Models are Few-Shot Learners", 2020, https://arxiv.org/abs/2005.14165

[16] "Open Source AutoML Tools: AutoGluon, TransmogrifAI, Auto-sklearn, and NNI", 2020, https://www.bizety.com/2020/06/16/open-source-automl-tools-autogluon-transmogrifai-auto-sklearn-and-nni/

[17] "Isaac Sim: Omniverse Robotics App", https://developer.nvidia.com/isaac-sim

[18] Rob Toews, "10 AI Predictions For 2021", 2020, https://www.forbes.com/sites/robtoews/2020/12/22/10-ai-predictions-for-2021

[19] https://www.tensorflow.org/lite

[20] https://github.com/CEA-LIST/N2D2

[21] Markus Levy, "Glow Compiler Optimizes Neural Networks for Low-Power NXP MCUs", 2020, https://www.nxp.com/company/blog/glow-compiler-optimizes-neural-networks-for-low-power-nxp-mcus:BL-OPTIMIZES-NEURAL-NETWORKS

[22] Ben Dickson, "The untold story of GPT-3 is the transformation of OpenAI", 2020, https://bdtechtalks.com/2020/08/17/openai-gpt-3-commercial-ai/

[23] Naveen Kumar, "Google breaks AI performance records in MLPerf with world's fastest training supercomputer", 2020, https://cloud.google.com/blog/products/ai-machine-learning/google-breaks-ai-performance-records-in-mlperf-with-worlds-fastest-training-supercomputer

[24] Sébastien Stormacq, "Majority of Alexa Now Running on Faster, More Cost-Effective Amazon EC2 Inf1 Instances", 2020, https://aws.amazon.com/fr/blogs/aws/majority-of-alexa-now-running-on-faster-more-cost-effective-amazon-ec2-inf1-instances/

[25] Ian King and Dina Bass, "Microsoft Designing Its Own Chips for Servers, Surface PCs", 2020, https://www.bloomberg.com/news/articles/2020-12-18/microsoft-is-designing-its-own-chips-for-servers-surface-pcs?sref=ctSjKj2N

[26] Karen Hao, "Training a single AI model can emit as much carbon as five cars in their lifetimes", 2019, https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes

[27] Emma Strubell et al., "Energy and Policy Considerations for Deep Learning in NLP", 2019, https://arxiv.org/abs/1906.02243

[28] Julian Schrittwieser et al. "Mastering Atari, Go, chess and shogi by planning with a learned model", 2020, https://www.nature.com/articles/s41586-020-03051-4

[29] Philip E. Ross, "DeepMind's New AI Masters Games Without Even Being Taught the Rules", 2020, https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/deepminds-new-ai-masters-games-without-even-been-taught-the-rules

[30] "IDC Forecasts Strong 12.3% Growth for AI Market in 2020 Amidst Challenging Circumstances", 2020, https://www.idc.com/getdoc.jsp?containerId=prUS46757920

[31] "What is Federated Learning?", https://owkin.com/federated-learning/

Marc Duranton is a researcher in the Research and Technology Department of CEA (Alternative energies and Atomic Energy Commission), France and the coordinator of the HiPEAC Vision 2021.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: M. Duranton. The omnipresent artificial intelligence. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 62-73, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

Many IT systems still lack (cyber)security. We have a lot to lose from having poorly secured IT systems, and a lot to gain from secure ones. The good news is: we can do it, in Europe. Here is how.

Cybersecurity must come to IT systems now

By OLIVIER ZENDRA and BART COPPENS

After decades of apparently low-intensity cyber attacks, during which security was not really thought of on most IT systems, recent years have brought a flurry of well-organized, larger-scale attacks that have caused billions of Euros of damage. This was made possible by the plethora of IT systems that have been produced with no or low security, a trend that has further increased with the rise of ubiquitous computing, with smartphones, IoT and smart-* being everywhere with extremely low control.

However, although the current situation in IT systems can still be considered as critical and very much working in favour of cyber attackers, there are definitely paths to massive but doable technical improvements that can lead us to a much more secure and sovereign IT ecosystem, along with strong business opportunities in Europe.

Key insights

- Cyber attacks are ever increasing, and the cost of damage caused by lack of cybersecurity is soaring.
- Security must now become a first-class citizen of new programs from day one, in both specifications and source code.
- However, legacy does not go away: tools that are able to analyze the legacy code base, and find vulnerabilities and unwanted behaviour are needed, as are those that automatically circumvent or mitigate them.
- Automated diversity is a means of resilience against security attacks and of safety.
- Liability in IT systems is crucial to the advance of security aspects. Security certification must become mandatory. Certification and liability reinforce one another. Regulation must play its role.
- Cybersecurity, hence IT sovereignty, which drastically impacts political sovereignty, depends on having control over hardware and software.

Key recommendations

- Promote that research and industry have security as a first-class citizen of new IT systems, both in specifications and in source code.
- Promote research on methods and tools to find vulnerabilities in existing code, as well as tools to automatically prevent or mitigate them.
- Mandate security certification for IT systems whose malfunctioning would impact a large number of citizens, in the same ways as certification for critical systems.
- Regulate to make IT systems providers and resellers liable.
- To reclaim IT sovereignty, base the critical parts for cybersecurity of EU IT systems either on open-source software and hardware, or on EU-made, trustable because audited, proprietary hardware or software.

CYBERSECURITY MUST COME TO IT SYSTEMS NOW

Barely a few years ago, cybersecurity was, if not unheard of, at least not on the minds of many people and leaders. IT systems seemed to be working, attacks seemed to target only "others"; in short, cybersecurity was a non-pressing matter, hence often non-existent… Since then, cyber attacks have made headlines: influence on the 2016 US Presidential elections; Petya ransomware attacks in 2016-2017 with losses estimated to US$10 billion [8], and Wannacry ransomware attacks in 2017 leading to losses estimated to up to US$4 billion [9].

However, if awareness has indeed risen, it is still true that most of the world, including Europe, has not yet fully awakened to the cybersecurity aspect of IT systems.

Threats are numerous: malware in all its forms (spyware, ransomware, trojan…), sniffing, spoofing, Man In the Middle attacks, backdoors in hardware or software, etc. They are also spread across most if not all IT domains, ranging from the simplest, cheapest IoT devices to the more expensive smartphones, cars, planes, banking systems, air traffic control systems, etc. They know no frontiers, as IT information can easily cross (most) frontiers in the world, and know no delays thanks to the quasi-instantaneousness of information transmission. Cyber threats used to be considered as being limited to hacking, which itself was seen as an uncommon activity of geeky-nerdy underground teenagers (see e.g. the movie "Wargames" [10] of 1983). Now, there is plenty of evidence that this has morphed into "industrial-scale" activities, with states using cyber warfare units even in relatively peaceful times, and organized crime also having their professional hacker teams. Even a period of global pandemic such as that of COVID-19 is not creating any truce on the cybersecurity front, since "Cybercriminals are developing and boosting their attacks at an alarming pace, exploiting the fear and uncertainty caused by the unstable social and economic situation created by COVID-19." [11]

Yet, it is clear that as a society it is important that all our systems, both industrial-scale and personal ones, be as secure as possible. The stability and continuity of our daily activities, be they private or professional, and even our lives, depend heavily on the secure and continuous operation of IT systems. Until very recently, security was not a top priority for those who design and implement IT systems. This has led to the current situation, with scores of vulnerable IT systems being used, and many still being developed with poor security.

Cybersecurity can be seen as a defence system, and has often been compared to medieval castles. Although it is true that some cyber defences can be nested like medieval castle defences were, the metaphor is inappropriate for whole IT systems. Unlike medieval castles, in which attackers must pass all the defences *successively* (climb the slope, and cross the outer moat, and climb or breach the outer wall, and the same for the inner wall, *and* for the dungeon), IT systems are indeed much less nested. They are more layered, or stacked, meaning that any breach present in *either* the hardware, *or* the operating system, *or* the execution environment, *or* the application, could be exploited by an attacker to gain information or control from the system. In that sense, *one* hole in IT systems is enough. This is a bit similar to the proverbial "forgotten, concealed postern" in a castle, or to a secret tunnel. Of course, there still may be some compartmentalization, in which case not the whole system falls but only part of it (a bit like one castle among several). But overall, IT systems and their defence appear rather more vulnerable in principle that the iconic medieval castle.

However, solutions exist to this grim situation, some of them being specific to one layer of the IT system, some others being applicable to several layers. Many defence techniques exist, and can add their "stone" to the security walls of IT systems.

Address Space Layout Randomization (ASLR) is a defense technique which consists in making the structure of processes more random with regards to the places in which code, data and libraries are found in memory, so as to make the task of an attacker (e.g. malware) more difficult. Indeed, by offering it a less predictable target, the attacker can less easily move/jump the instruction pointer at execution to a target of its choice, which makes the executable less vulnerable to buffer overflows and code injection. ASLR is a cross-layer technique, since it can be applied to the memory of application code/data, execution environment (e.g. VM) code/data, and even OS code/data.

Another way to create a "stronger stronghold" for IT systems is by having a smaller attack surface, by basing everything on a smaller yet highly secure base, like the **TCB** (Trusted Computing Base). Research work has clearly shown that "From the security point of view, the monolithic OS design is flawed and a root cause of the

majority of compromises." [22], and that **microkernels** make it easier to have more secure OSes, even more so when they are verified. Indeed, OSes must have some parts that execute with the highest level of privileges, in kernel mode. But being software, kernels are also flawed. The bigger the kernel, the more flaws can be exploited in kernel mode to gain access to all information on the system. By reducing this attack surface to a bare minimum, the risk is mitigated. By reducing this kernel code to a bare minimum, it becomes easier to check it, secure it, or even verify it formally.

At the application and OS levels, a simple solution would be to use end-to-end cryptography in order to provide better protection for user data with regard to attacks, at storage level (encrypted filesystems), memory level and communication level. Recent OSes make it easy to encrypt files or the whole filesystem, and dedicated hardware support has made the cost of this largely unnoticeable in practice. At application level, several messaging or conferencing applications have introduced end-to-end crypto, some of them even putting a strong emphasis on this aspect (e.g. Olvid [23], Signal [24]). This need has only grown with COVID-19-induced teleworking, and the flurry of cyber attacks it has allowed [11].

Many more techniques exist. We focus in the remainder of this article on a few techniques that must be supported and promoted since they provide the means for the EU to secure its IT systems and IT ecosystem. We show very promising *technical solutions* that make it possible to **find and fix vulnerabilities** in existing IT systems, to **strengthen** the security and **resilience** of IT systems, and to **express security properties** in order to be able to *verify* the security of existing systems and to *produce* new ones that are much more *secure* if not devoid of any vulnerability. But if we do want to increase the security of our systems, we also need to *motivate* all actors to act in their best shared interest. This may imply **regulations**, based on **liability** in the face of the law and on **certification** processes. Finally, cybersecurity and IT sovereignty, hence simply **sovereignty** for the EU, depend on **trustable and auditable hardware and software**.

Image: ID 133487501 | © FabioConcetta | Dreamstime.com

Finding and fixing vulnerabilities in existing systems

Since it is now infeasible to recreate all systems from scratch and redesign and rewrite them with perfect security, we have to live with the legacy of our existing systems for a long time to come. As a result, these systems will need to be made secure to ensure that users and companies can access and store their private data for years to come.

In some cases, pragmatism can help us move forward on this path. As an example, consider the now commonplace technique of ASLR, which defends against memory-related vulnerabilities in software. Its core idea is that software will contain many bugs that can lead to exploitable vulnerabilities, but that attacks against these vulnerabilities were made exceedingly easy by the fact that program code and shared libraries were located on fixed addresses in memory. ASLR randomizes the locations of program code and libraries in memory, which makes such easy attacks fail. Thus, applying ASLR to all systems increases the cost and difficulty of mounting attacks, even though it cannot completely prevent all possible attacks. Still, users are safer thanks to it.

This goes to show that in the cases where we cannot rigorously ensure the required security properties, we can still increase the overall security of systems through other means. In particular, **we need to have techniques to find, and fix or mitigate, security vulnerabilities in legacy code bases**; or techniques to at least isolate, thanks to containerization, these potentially insecure legacy code bases in hypervisor-like infrastructure, controlling at the same time their exchanges with the rest of the system. While none of these techniques will be able to guarantee that programs are *completely* secure, each of them can lower the impact of inadvertent mistakes, and can raise the bar against specific attacks. The more of these techniques that are combined, the greater the security of the resulting system will be.

Even in large-scale systems, which are hard to deal with globally, vulnerability finding techniques can perform a dual function. While an automated tool being unable to find security vulnerabilities in a certain amount of time does not constitute proof of the security of a system, if such an automated tool *does* find a security vulnerability, that vulnerability serves as a constructive proof of the system's insecurity. Such security vulnerabilities (found automatically) can be passed as actionable items to the appropriate engineers to solve. This is a much more tractable task than trying to prove the security and correctness of an entire, large-scale system. For example, operating system kernels are complex, security-critical systems that are hard to analyze. Special-purpose techniques can be developed that target different kinds of bugs that can have a security impact in operating systems [12,13]. These analyses can still be improved in terms of how many such bugs they find in certain hard-

to-analyze contexts. Furthermore, as such tools only target specific classes of bugs, more such tools need to be developed in order to find so-far under-reported classes of bugs.

For the cases in which the systems still contain vulnerabilities which neither automated tools nor manual code analyses find, we need to have ways to mitigate the impact of these remaining vulnerabilities. This is typically done by targeting the ways in which an attacker typically exploits such vulnerabilities. An example of this was already touched upon in the context of ASLR: if typical attacks use hard-coded memory locations, making memory locations vary between executions will make attacks harder. Similarly, attackers can try to redirect the execution of the program in a way which the original developer did not write in the source code of the application. Then a type of defence, called control-flow integrity, inserts checks that verify that functions are executed only from the calling context of functions of which the developer intended this [14]. More such defences need to be researched and developed.

Diversity is a mean of resilience against attacks

One way of mitigating and protecting against security vulnerabilities is introducing *diversity*. This is akin to monoculture in crops: having a more diverse gene pool in crops can make fields more resilient to diseases and infections. This was already alluded to in the context of ASLR: if every execution of a program has a totally unpredictable memory layout, it is harder for an attacker to create a single exploit that works against all these different memory layouts. An attacker would need to carefully craft an exploit that works around these limitations, or there would need to be a way for the attacker to gain knowledge of the exact memory layout.

Diversity can be introduced at many different, if not all, levels of an IT system: hardware, operating system, execution environment, application level. This can be done by having different implementations, by generating different versions at compile time[15], by randomizing the programs at install time[16], by randomizing them at load time such as with ASLR, or even by randomizing them during the execution of the program [15,16]. Diversity has even more security-related applications. For example, multi-variant techniques take their inspiration from reliability in critical systems such as airplanes in which multiple different implementations are compared against one another [18]. By executing (specially chosen) diversified instances of the same application, feeding these the same inputs, and then comparing their outputs, some classes of attack can be prevented from the fact that these attacks will impact these instances in a different way, which can be detected[19]. An application of diversity in a wider sense can be found in the context of security updates. When a developer discovers and fixes a security vulnerability and releases an update, users don't typically apply this update immediately. However, attackers can still compare the original application which contains the vulnerability with the just-released updated version of the application in which the vulnerability has been patched. Based on this difference, attackers can easily create attacks against the users who do not yet have the update installed [20]. This is sometimes called patch *Tuesday / exploit Wednesday*, as Microsoft typically releases their (security) updates on a Tuesday, after which attackers can try to exploit the unpatched users the day afterwards. A mitigation to this can consist of making the original version and the updated version differ more, that is, making both versions more diverse from each other. This will slow down the analysis and attack-generation by the attacker, allowing more users to apply the security update [21].

Some of these diversity techniques are already widely adopted, but definitely not all. One reason that some proposed techniques are not yet used in practice, is that they have too high an overhead to be applicable in most practical situations, or still have other limitations. Furthermore, as no technique can prevent every possible vulnerability, it is important that more such techniques are investigated and can be applied in practice. **This means researching new diversity techniques that are both efficient and effective, as well as making existing technologies more widely applicable**, by raising their maturity levels such that they can be adopted at large, rather than existing as academic prototypes.

Non-functional security properties must be included as first class citizens in new IT systems

In addition to taking care of vulnerabilities in existing legacy systems, it is of the utmost importance to include security as a first-class citizen when building new IT systems. Way too often, security is not considered upfront in programs at design and implementation stages, but only as an afterthought. This situation absolutely must change. The non-functional property that security is, or more precisely the set of non-functional properties that pertain to security in its various aspects, must be present in programs, in the minds of designers and implementers, from the very beginning of the creation of an IT system.

An example of a security property would be the level of threat by interception and decryption of communications in a given environment (e.g. known by its location in a military conflict). Another example would be, say, the strength of a cryptographic algorithm. An IT system could adapt its operation depending on the threat level, sending unencrypted or lightly (hence cheaply in term of time and energy)

Image: ID 128042127 | © Blackboard373 | Dreamstime.com

encrypted information in safe cases, choosing a stronger encryption algorithm in higher threat situations, or even avoiding transmitting if the level of threat is above the strength of the available algorithms.

To get to this point, security properties must be treated as first-class citizens, which means that they have to be expressed as clearly and as explicitly as the functional properties (i.e. what the program does, its algorithms), and that designers and developers should be able to reason about them, query them, manipulate them, same as for the functional aspects of the program. It is only by having security interleaved in all the fibres of the IT system that it can be solid.

To this end, first, security properties must be present in the **specifications** of the IT system from the beginning. Designers must be able to express what levels of security they want, for the various facets of security, or reasons about them. **Security contracts** must be present in the system that express and guarantee security at module, or component boundaries. Developers must be able to pick up off-the-shelf **modules** or components knowing the security levels they provide for specific facets or security, and plug them in as part of the security continuum of their system.

Programming languages are also not all equals in terms of security. Many programming languages tolerate sloppy programming, where code that looks reasonable at first sight may in fact contain major vulnerabilities. Some, e.g. C/C++, tend to be harder to master, more error-prone, making it difficult to find bugs like security issues. For a compendium of programming language vulnerabilities, see the work by ISO/IEC TR 24772 Programming languages — Guidance to avoiding vulnerabilities in programming languages [2]. Other languages should be promoted: those that have stricter rules, more safeguards, and make it easier to develop more secure code. The Rust programming language, originally developed by Mozilla, is one example that makes many aspects of memory management and memory safety explicit in its language constructs, making it harder to leave room for security glitches that could be exploited malevolently.

Automated or semi-automated **tools** must be able to rely on these specification and security contracts **to verify security**, to prove security properties, to provide strong guarantees about the quality of IT systems with regard to security, at both specification and code level. Notable examples that can help in this endeavour in the line of program verification include Frama-C [3], Coq [5] and Ada SPARK's Discovery toolset [4]: those tools operate on the premise that the source code must conform to some formal specification.

Although **security by design** of whole IT systems seems a perfect answer to security issues, the current approach is often more limited and pragmatic, with only a limited part of the IT system being trusted. This Trusted Computing Base (or TCB) [1] comprises the system hardware, firmware and software components whose combination is intended to provide the system with mechanisms for a secure environment. The idea here is that **verifying**, either automatically with verification tools or manually by human examination, this on a small set of hardware and software is a more tractable and less costly task than doing it on the whole system. However, verification of large pieces of real-life software is already doable, as proven by recent research-level successes like the CompCert compiler [6] and the verified parts of the seL4 microkernel [7].

Liability for IT systems

Liability and **certification** are crucial building blocks needed to mandate taking into account non-functional security properties in IT systems. Indeed, **adding the legal building block of liability**, hanging the threat of a potential cost on non-secure systems, makes it more appealing for system builders and providers to commit effort, and hence money, on having secure systems. With liability, the extra time of specification and the extra step of verifica-

Image: ID 119668736 | © Alexandersikov | Dreamstime.com

tion would become worth taking. **Certification is the technical building block** that makes the liability legal one viable. With certification, system builders can have their efforts for security quantified, priced and legally acknowledged; purchasers can mandate security based on an independent assessment; regulatory bodies can outlaw low-security systems. Certification can be based on test suites that must be passed, on verification tools, on development practices that must be adhered to, etc.

With liability and certification, companies have the incentive to create secure software, or to keep finding, and mitigating or fixing vulnerabilities.

Sovereignty depends on trustable and auditable hardware and software

A castle can have the deepest moat and the highest and strongest walls, but they are of no use if an adversary has the key to the secret tunnel or hidden postern. Similarly, if we don't control the parts of our IT systems that are crucial for cybersecurity, we can hardly guarantee it. If backdoors exist in the operating system or even in the hardware we buy, unknown to us, they can be exploited by attackers and it is very difficult to add extra elements of security to counter them. The same is true for development tools, like the compilers that generate the actual executables, or the software libraries used as building blocks to compose programs: being closed source and distributed as binaries, they could embed backdoors in the programs that are produced with them.

Currently, by basing most, if not all, of its activities on IT systems running on proprietary hardware and operating systems not made in the EU, the EU effectively entrusts the providers with the keys to its whole economy and all aspects of its (cyber)security. The situation is barely better for development tools, compilers, and libraries, in which EU production is quantitatively very limited. This is why it is crucial to retain the keys of the castle, which means to retain sovereignty and control over the important hardware and software components for at least the TCB, and hopefully the whole software stack.

The way for the EU to reclaim its IT **sovereignty** is thus to **base the TCB of EU IT systems either on open-source software and hardware, or on EU-made, trustable-because-audited, proprietary hardware or software.**

Conclusion

(Cyber)security is only as strong as its weakest link, which means that the level of security of an IT system is at the minimum of the level of security of its components. For way too long has security been forgotten in the interests of better time-to-market and lower prices. However, increased awareness of the **cost of poor security**, coupled with the flurry of attacks in recent years, in a much more organised way than before, is now making it both necessary and possible to reverse this trend, and to take a path of better security in IT systems, providing better security for economies and for people. The technical building blocks to this end are within sight in the research community and must be nurtured and pushed forward, so as to quickly mature and then irrigate the whole IT industry. Europe has a key role to play in terms of **IT systems that are more valuable because of their higher quality thanks to higher security**, by promoting **targeted research**, taking the appropriate **regulatory steps**, and taking the necessary steps to reclaim **sovereignty of the critical building blocks** of its own IT systems.

References

[1] TCB: Trusted Computing Base: https://en.wikipedia.org/wiki/Trusted_computing_base

[2] ISO/IEC TR 24772 Programming languages — Guidance to avoiding vulnerabilities in programming languages: https://committee.iso.org/sites/isoorg/contents/data/committee/04/52/45202/x/catalogue/p/1/u/1/w/0/d/0

[3] Frama-C: https://frama-c.com/features.html

[4] Ada SPARK's Discovery toolset: https://www.adacore.com/sparkpro

[5] Coq proof assistant: https://en.wikipedia.org/wiki/Coq

[6] CompCert compiler: http://compcert.inria.fr

[7] seL4 microkernel: https://sel4.systems

[8] Petya ransomware attacks: https://en.wikipedia.org/wiki/Petya_(malware)

[9] Wannacry ransomware attacks: https://en.wikipedia.org/wiki/WannaCry_ransomware_attack

[10] "Wargames" movie: https://www.imdb.com/title/tt0086567/

[11] INTERPOL report shows alarming rate of cyber attacks during COVID-19: https://www.interpol.int/News-and-Events/News/2020/INTERPOL-report-shows-alarming-rate-of-cyberattacks-during-COVID-19

[12] Precise and Scalable Detection of Double-Fetch Bugs in OS Kernels. Meng Xu, Chenxiong Qian, Kangjie Lu, Michael Backes, Taesoo Kim. IEEE Symposium on Security and Privacy, 2018

[13] Check It Again: Detecting Lacking-Recheck Bugs in OS Kernels. Wenwen Wang, Kangjie Lu, Pen-Chung Yew. ACM Conference on Computer and Communications Security, 2018.

[14] Control-flow integrity principles, implementations, and applications. Martín Abadi, Mihai Budiu, Úlfar Erlingsson, Jay Ligatti. ACM Trans. Inf. Syst. Secur. 2009

[15] Enhanced Operating System Security Through Efficient and Fine-grained Address Space Randomization. Cristiano Giuffrida, Anton Kuijsten, Andrew S. Tanenbaum. In USENIX Security Symposium, 2012

[16] SoK: Automated Software Diversity. Per Larsen, Andrei Homescu, Stefan Brunthaler, Michael Franz. IEEE Symposium on Security and Privacy, 2014

[17] Librando: transparent code randomization for just-in-time compilers. Andrei Homescu, Stefan Brunthaler, Per Larsen, Michael Franz In ACM Conference on Computer and Communications Security, 2013

[18] N-Variant Systems: A Secretless Framework for Security through Diversity. Benjamin Cox, David Evans. USENIX Security Symposium, 2006

[19] Cloning Your Gadgets: Complete ROP Attack Immunity with Multi-Variant Execution. Stijn Volckaert, Bart Coppens, Bjorn De Sutter. IEEE Trans. Dependable Secur. Comput. 2016

[20] Automatic Patch-Based Exploit Generation is Possible: Techniques and Implications. David Brumley, Pongsin Poosankam, Dawn Xiaodong Song, Jiang Zheng. IEEE Symposium on Security and Privacy. 2008

[21] Feedback-driven binary code diversification. Bart Coppens, Bjorn De Sutter, Jonas Maebe. ACM Trans. Archit. Code Optim. 2013

[22] The Jury Is In: Monolithic OS Design Is Flawed: Microkernel-based Designs Improve Security. Simon Biggs, Damon Lee, Gernot Heiser: APSys 2018: 16:1-16:7

[23] Olvid. https://olvid.io/technology/en/

[24] Signal. https://signal.org/docs/

Olivier Zendra is a Tenured Computer Science Researcher at Inria, Rennes, France.

Bart Coppens is Postdoctoral Researcher in the Electronics department of Ghent University, Ghent, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: O. Zendra and B. Coppens. Cybersecurity must come to IT systems now. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 74-79, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

Software breaks computer security on principle: how can we retain the value of software without rebooting IT?

Reversing John von Neumann and Steve Jobs, but not software

By THOMAS HOBERG

The base architecture, programming languages and operating systems that run practically all computing today were written for isolated systems nobody wanted to harm and computer viruses were mere game theory, simulating life. Today the internet exposes them all to virus and malware attacks, threatening to destroy our livelihood from our pockets, unless we delegate their security to vendors we may not trust. Which solutions can deliver the trust and security we require without losing seven decades of software legacy?

Key insights

- The key ingredient to the software revolution left nearly all existing computers vulnerable to malware: that wasn't a concern until they were connected to the internet: now they need very labour-intensive protection.
- Cloud operators and client device vendors may offer protection at the cost of lock-in and loss of sovereignty.
- Various techniques to harden systems and reduce the dependency exist as prototypes and even solutions, but they do not further the IT giants' business model.
- The lack of resilience and multi-compliance is a major obstacle to create value in the world of internet of things.
- The EU has a natural motivation to advance such technology to supports its ethics.

Key recommendations

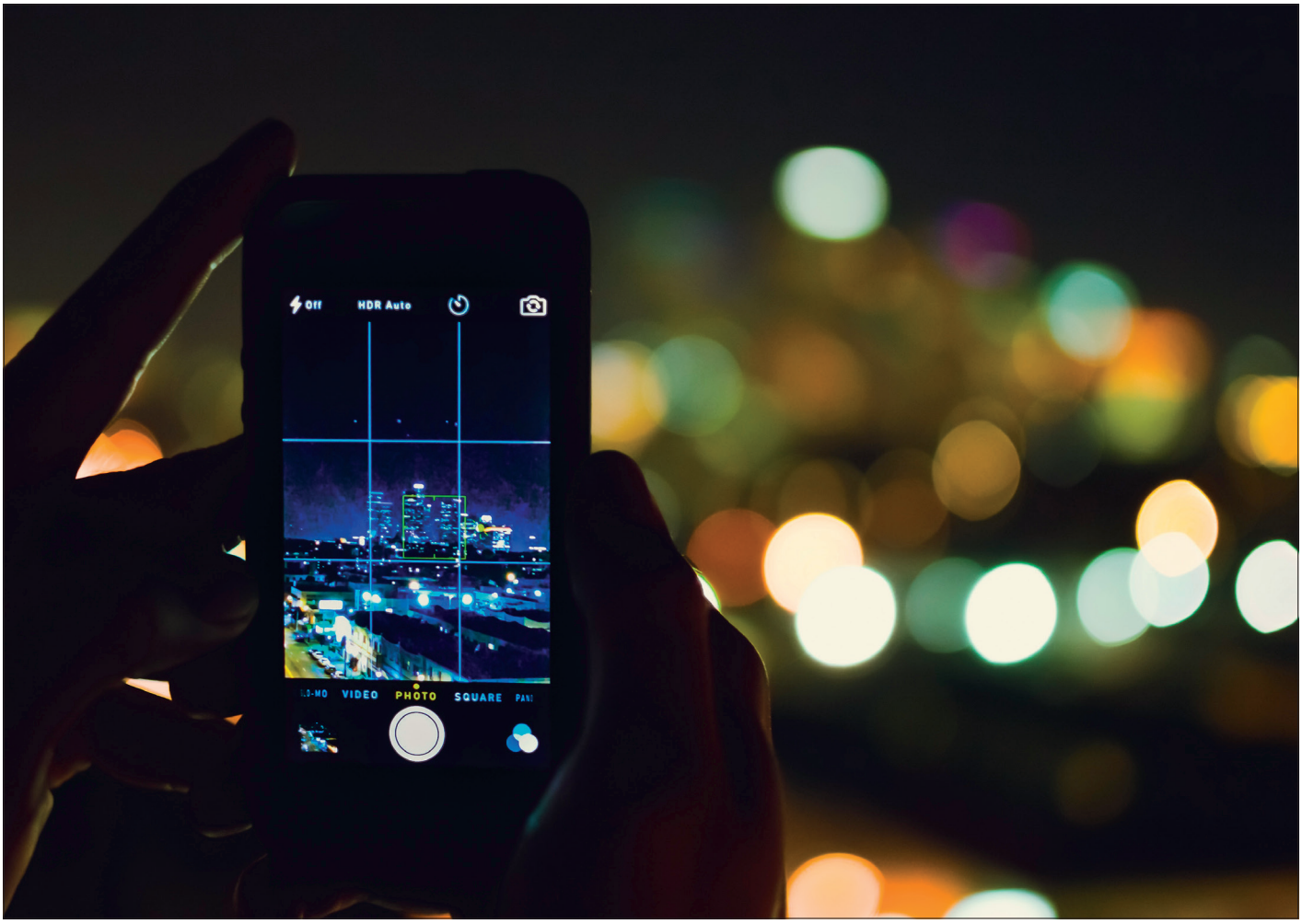
- Regulate to force vendors of very personal computers and IoT devices to give primary sovereignty and control to owners: support for vendor, partner, government and other enclaves will then happen very quickly.
- Regulate to put all client-side generated data under the control of the owner. Exceptions required by regulatory compliance (e.g. FCC, automobile data), require full transparency and regulation.
- Computing devices that are uncompromisable by design are key to affordable cyber-physical systems: the EU should invest in that technology and matching regulation.
- The ability to support multiple and contradictory compliance frameworks with conflict resolution under owner control is key to enabling cyber-physical systems that all stakeholders, from consumers to governments, can trust: the EU should invest in that technology and corresponding regulation.

The original problem of mixing code with data

Computers cannot tell good code from bad.

Emil Post had already proven [1] that as impossible by the time John von Neumann turned it from mere academic theory into a pervasive problem by promoting the mixing of code and data in a single memory space to create software [2], making both indistinguishable. Data could turn into code and code could create more of both. To von Neumann this self-modifying code gave Turing completeness and evolutionary propagation, which were the essential ingredients to turn code into a new virus like [3] digital life form with the full potential of intelligence.

To mere IT professionals this inability to distinguish between code and data, good or harmful, has become the Achilles heel of software, the fundamental vulnerability which allows trojans and malware to spread even into the best protected data centres, and far more easily still into our very personal computers and the IoT devices we purchased for a convenient digital lifestyle full of smart things and services.



Walled gardens: from paradise to prison?

Steve Jobs is best known for enabling people, who prefer not to bother with details, to do things with complex devices, which are made to appear simple with the help of sophisticated software when most subscribers do not even understand what that is or what they are giving up.

His solution to achieve simplicity was to take control, hide the computer, take away the universality not needed and reduce all user accesses and involvement not required for function. While he was rarely shy to impose his will on others, he probably didn't quite foresee that this approach, created for a more convenient portable music player, would transform from a design paradigm into an issue of individual and even national sovereignty.

Simple solutions cease to work at scale

Today we face a situation where the fundamental vulnerability of software on von Neumann computers is most easily contained by using a Jobs approach: putting all devices into a walled garden and the management of these devices into the hands of their vendor or software provider. But it is no longer just about songs and it is no longer just about iPods. As Google, Microsoft and many others with even fewer credentials have chosen the same path, we stand at a point where corporate ethics violate sovereignty, from individual to national or value system levels.

That conflict reduces value. Google cannot earn money from services it is prohibited from selling in China. Amazon also cannot earn top dollar from consumers who do not entrust Alexa with everything their butler might know about them after twenty years of service. Smartcams, routers, digital media recorders etc. first hijacked by the MIRAI botnet and then remotely destroyed by the likes of BrickerBot [4] demonstrate direct negative value, while the damage resulting from the attacks they enabled could be orders of magnitude bigger than their purchase price.

The challenge is to reverse the negative effects of von Neumann and Jobs and allow the creation of new value, without giving up seven decades of software nor the convenience of delegated management.

Binary is far too few choices for digitalization

Digital computers are only true or false at their heart, applications on Unix are controlled via a single superuser bit, which gives them god-like power or nothing special. An internet connection is very similar to current cryptography: either you can see and communicate because you have the key or a connection, or not.

Digitalization needs to reflect and include social codes, probably not a direct copy, but very likely with an even higher degree of differentiation and regulation. It is similar with hierarchies: as long as Jobs' design decisions only affected one music

player in many, there was no issue about Apple's total control of the firmware.

But when you'd like your very personal computer to become:

- your digital brain extension, that contains all hard-to-remember facts and secrets about yourself, your family and friends;
- the primary interface to all your virtual personal assistants or digital servants, which manage your home, your health, your property, your mobility, your logistics and the relationship with the assistants of your immediate and extended family members, your friends your colleagues, clients etc;
- the primary interface to complex machines and devices under regulation or warranties with strong foreign compliance requirements;
- the primary interface to your profession, your employer, colleagues, clients, corporate services under corporate compliance rules; or
- the primary interface to government, insurance, medical, banking services with strong data security and authenticity requirements under various compliance schemes,

even the best single vendor approach is insufficient. Not only is there a need to support multiple hierarchies of trust, security and delegation, but the vendor is no obvious choice for a leading position. It is like leaving your stonemason in charge of your mansion, the realtor to manage your finances or like trusting your car manufacturer to manage your ecological footprint.

And since we know from Kurt Gödel [5] that these distinct hierarchies and regulations are never free of contradictions if they are complex enough to be useful, somebody must be in charge; to decide or overrule and nobody is more qualified for that role than the owner, who will quite simply refuse to purchase and consume otherwise, but he should be helped in his charge otherwise he will delegate, and we are back to square one.

The European Union must provide regulation to elevate the owner's sovereignty on smart objects above all others.

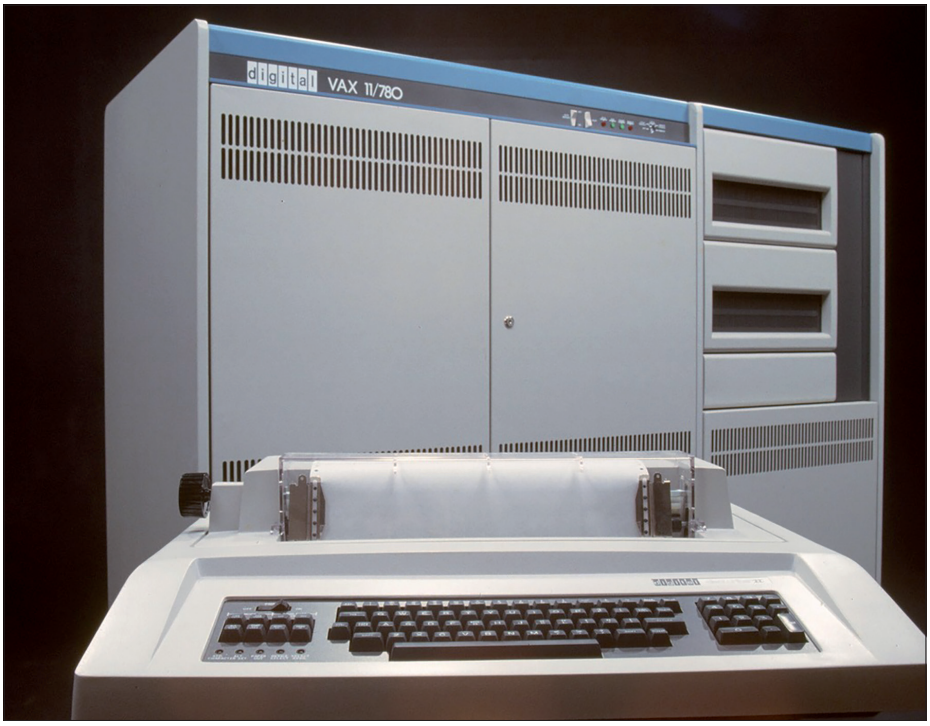


Figure 1: Digital Equipment Corporation VAX 11/780 [25]

The enemy inside the castle keep

Practically all mobile, personal and server computers today, from a €5 Raspberry PI Zero to supercomputers costing billions, share the same basic architecture as Digital's 1977 VAX and trace their operating systems back to that machine. Windows is very much a re-implementation of VMS by its chief architect Dave Cutler, Linux is similarly a re-implementation of BSD-Unix while MacOS and iOS are even more direct descendants with a bit of Mach 3 flavour added.

Unix and VMS were designed so that a computer could run dozens of programs for dozens of users in a manner where:

- they would all appear to be running in parallel, even on a single processor;
- they would appear to have practically unlimited memory via virtual memory and paging;
- code and data share the same memory space, function call return addresses are stored and manipulatable on the same stack as data;
- a faulty application would not stop the system or impact the programs others were running;
- all system resources assigned to a user could be protected from access by others.

And even if the internet was largely born on these machines, they started as standalone computers in a very different world. They would assume:

- being sole master of their universe, without any notion of connectivity, collaboration between peers or delegation of authority to others;
- running code with well-known progeny, mostly self-written and compiled locally;
- all inputs, data and code, would come from either well-known users or self-managed devices.

A VAX was very expensive to own and therefore used mostly by IT experts or scientists. It was supervised by staff under physical and logical security with a keen eye on ensuring that it was used economically and with minimal disruption from careless mistakes or outright abuse.

These machines were also highly individual in their configuration with tailor-made 'sysgens' (operating system generation or rebuilding) performed locally.

The €5 Raspberry Pi Zero is functionally very similar to the most powerful VAX ever built, a VAX 9000 model 110 mainframe sold for a $1M in 1990. They share RAM capacity (512MB) and removable

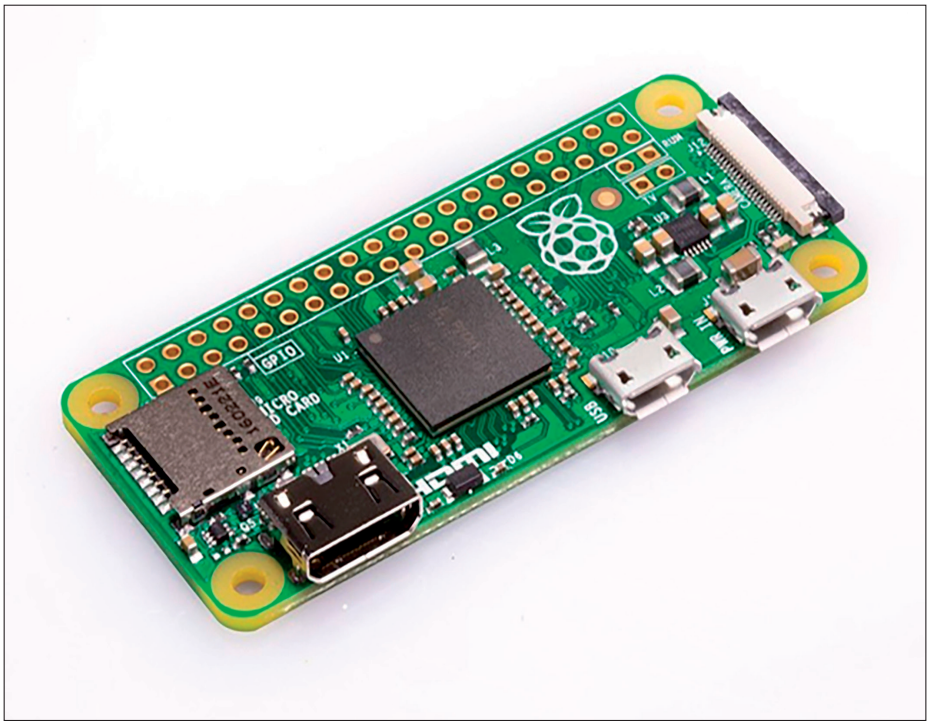


Figure 2: Raspberry Pi Zero

storage, code compiled from source would easily run 20x faster on the Pi, full VAX emulation somewhat slower [6] because that is complex for a RISC.

The operating systems have changed so little, such a tiny machine looks like a VAX to an application, a user on a terminal or any other machine on the internet, including cloud servers of any size. Because they have the full functionality and the development convenience of a Unix mainframe but cost a fraction of the full retail product, countless home-routers, DVRs, SmartCams, smart toys, home-control systems etc. use embedded Linux systems very similar to the Pi.

But today's environment is an almost complete opposite of the original design conditions. They need to operate in a world where:

- they share the internet with millions of other systems, but never properly embed the concepts of networking, collaboration, trust and delegation;
- application code is distributed as binaries from authors unknown and potentially hostiles;
- most communication is over the network with other computers, many of which are guaranteed to be hostile while data might contain code under their control;
- a complete system is cheaper than a good cup of coffee, while staff salaries have multiplied.

Their functional equivalence to a VAX means that the operational effort they require is also similar, a proper sysgen (full recompile of the OS) still hours of work. That is why most of these systems:

- run with an outdated binary copy of the operating system with well-known vulnerabilities and documented exploits;
- operate without 24x7 data centre support staff, without physical and logical security and without network firewalls and intrusion detection systems on the open internet;
- can be abused, hijacked and converted into cyber-bots under the control of a hostile command server;
- can be used to attack the more valuable assets in your network e.g. with ransomware, or anyone else on the internet.

The VAX enabled a multiplexed kernel and user-land applications to isolate errors and avoid crashes. Virtual memory, a ring-based security model to protect devices and OS memory regions gave full theoretical protection against application errors or attacks on a VAX, if the OS code was free of vulnerabilities. Users who regularly ran badly-behaving or downright malicious code were quickly identified by operators, given very personalized feedback or had their system access removed.

Today's sad reality is that many free toy apps popular with kids already contain malware while even professionally written applications contain vulnerabilities which can be exploited to override their control flow to the point where they run an attacker's code using return-oriented programming [7]. Exploits arrive via data ingested by the application as part of its normal operation, which has been carefully prepared on a cheap binary clone sold by the million. And once malicious code runs in a user's app space, it only takes another vulnerability to hijack the control flow of a system service or the kernel itself to turn that tiny cheap system into the functional equivalent of a full Unix mainframe gone rogue.

Even inside Apple, Google, Amazon or Microsoft's carefully groomed walled gardens, dozens of new OS level vulnerabilities are discovered and patched every month [8], while third party applications may not be checked at all. The far cheaper devices outside those walls, without any vendor maintenance or even the appropriate facilities, are turned into a cyber-bot within minutes of their first internet connection.

A giant security industry has sprung up over recent decades (typical estimated size for 2020 is around $40b in the United States). It tries constantly to keep VAX type von Neumann machines from being invaded via very labour-intensive processes:

- complex compliance frameworks are developed, implemented, monitored and audited;
- all communications go through firewalls which are carefully scanning protocol and application-level data aimed at exploiting vulnerabilities;
- intrusion detection systems observe and scan systems, appliances and communication data for traces of intrusion and leakage of sensitive data;
- systems and applications are subjected to penetration testing and extensive input

fuzzing tests by vendors or QA teams and regularly patched to close discovered vulnerabilities.

But that industry depends on the cost of these efforts being balanced by the value generated from computation. Typically, that requires scale, e.g. big iron providing high-value financial services, huge cloud data centre farms with millions of servers, or billions of end-user devices in a walled garden.

Apple, Google, Amazon and Microsoft appreciate the walled garden approach for very personal computers, cloud servers and the increasing market because it is a natural extension of their cloud operations, and it also gives them exclusivity and control. It allows these companies to tax the value generated by others in that ecosystem, to access all the data flowing through it and to sell any value yet undiscovered.

But it fails to solve the ever more pressing issue of sovereignty, where a single set of corporate ethics no longer satisfies the demands of a highly diverse global market with a myriad of conflicting sets of ethical, legal and compliance frameworks, privacy rules and individual preferences and needs. And more specifically it fails to satisfy the loyalty demands of consumers who would like to buy into the promise of smart cyber-physical systems at their service.

To overcome the fundamental vulnerabilities and shortcomings of von Neumann and Jobs, two major avenues of research need to be pursued:

- reducing the attack surface and strengthening computing devices' resilience with the least amount of evolutionary change for the existing software base
- enable the parallel and safe operation of multiple enclaves under distinct sovereignties and reliable mechanisms to resolve contradicting rulesets.

Both need to happen without increasing energy or any other expense but with better sustainability, people focus etc.

Research and solutions already exist in these dimensions. That they are not exploited to their full potential is clearly a consequence of the lack of incentives. The walled garden approach is attractive for the companies who control client and cloud platforms. Because the single-sovereign walled garden approach can't be sustained and results in significant political pressures, it needs to be managed and regulated. The current effective monopoly of power over platforms like Android, iOS, Chrome, Office 365, G-Suite, Facebook, TikTok, Amazon etc. needs to be controlled and opened, to enable the evolution of solutions that allow the collaboration between multiple sovereignties and strengthen the resilience of the devices themselves.

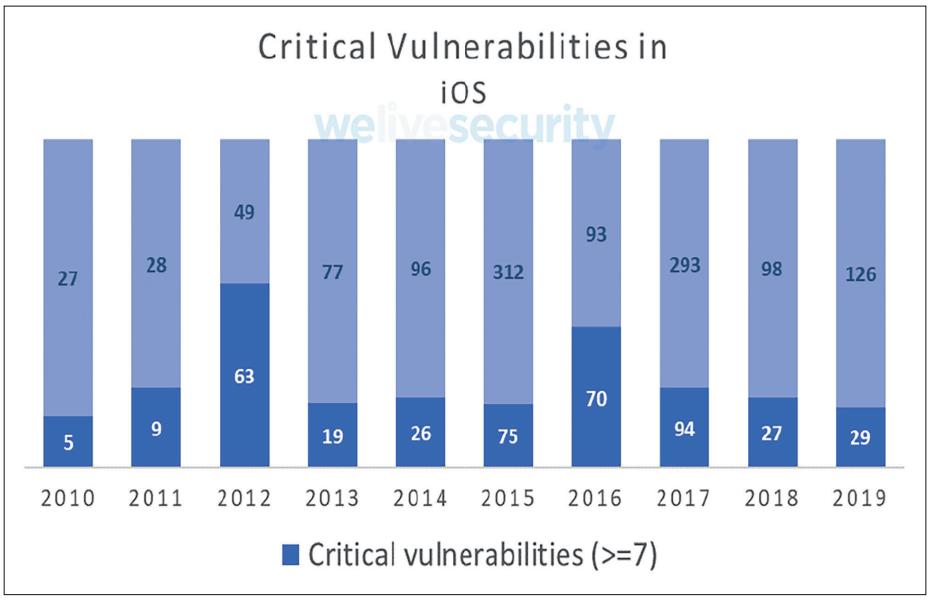


Figure 3: Vulnerabilities in iOS [26]

Example solutions

Baseband controller enclaves

Both the phone and WIFI networks are subject to heavy regulatory compliance. A direct access from a general-purpose computer like the main SoC of a smartphone and its application code is a risk that TelCo operators, the FCC and similar bodies around the world will not tolerate.

Baseband controllers on smartphones run as a secure enclave, currently even on a physically separate computer with its own processor, memory, storage and operating system: one popular choice is the OKL4 microvisor [9], a formally verified µ-kernel with a long European and international progeny. Communication between the mobile phone OS and the baseband controller is similarly restricted and controlled as if to an internet backbone server.

The mobile phone's SIM card is yet another computer with its own OS, which includes the ability to run secure customer enclaves, albeit with very low computational power and very high operational cost and fees.

Both show that smartphone manufacturers can follow regulations and support distinct sovereigns on devices they design when regulation and financial stimuli are properly set. Current implementations use strong physical separation, and use formal verification to ensure the software has no exploitable vulnerabilities. For a more dynamic growth of sovereigns and the workloads they need to operate, technology assets will be developed once the incentives are present.

Encrypted virtual machines

The multi-sovereign issue has been a pressing problem in mainframe collocation for decades and is now also affecting cloud servers evolved from personal computer technology: how can workloads from competitors or downright enemies on the same physical host be securely and reliably kept apart, and is there a way to keep even the host from tampering with them?

Hardware support for encrypting each virtual machine on a system with a different

customer-specific key has recently found its way into x86 CPUs from AMD [10] and Intel, even if it is so far only enabled on server variants. Research to commoditize and secure this approach, which may provide easy-to-use and dynamic enclaves, is ongoing [11] and needs further support.

Trusted root

Key management for the physical host VMs underneath, a trusted boot path and integrity checks for critical hardware components like non-volatile memory are also becoming a standard requirement for cloud and client systems and clearly represent an area where vendors have lacked both transparency and quality [12]. While technically Microsoft's initiative [13] to include a hardware root-of-trust into the very core of the vast majority of all CPUs is a move in the right direction, this project needs a completely open source, international and likely federated approach so as not to risk suspicion of violating national and individual sovereignty.

Hardware control flow integrity, tagged memory, capability-based access

Today it is very difficult to even think outside the box of the VAX architecture. But in the early days of computing, most of what today seems set in stone was still open for discussion and indeed alternatives did exist, where the need for them was high enough. Capability based addressing as a mechanism to protect against a user and his code accessing memory to which they had no permission, was explored in the 1960s [14] and even implemented in e.g. the Plessey S/250, where the main interest was actually in securely updating shared data structures in a fault tolerant distributed memory space controlling military communications e.g. in the Gulf War [15]. While a 32 or 64bit number fully qualifies for access to data within an application's VAX type virtual memory address space, memory references in CAP systems are never fully exposed to the application but kept outside its visible address space and maintained as a collection of access tokens under constant hardware control, which can be duplicated and transferred to enable communication, never directly read or written. While VAX legacy software doesn't easily support such a model, information sharing between enclaves under API control may still be inspired by such approaches.

There are more and more attempts to bring capabilities and access tags into the mainstream architecture for their tamper resilience using in-address space capabilities for backward compatibility. Google is even developing a new mobile operating system called Fuchsia to explore capabilities [16]. ARM 8.3 pointer authentication code (PAC) [17] and ARM 8.5 memory tagging extensions (MTE) [18] take advantage of the fact that in multi-level store architectures the high-order address bits of a full 64-bit address tend to go unused, as many as 24 bits on all client devices by default. PAC can use these bits to have the compiler create a cryptographic authentication code in the high-order bits of a jump or branch target, which a return-oriented programming routine could not duplicate, resulting in a program abort.

The most ambitious and forward-looking initiative for in-address-space memory capabilities is CHERI [19], while external address space [20] capability tags are also explored. PAC capable hardware has entered the mainstream client device population e.g. with the Apple A12 chip, MTE is fully specified and CHERI still experimental.

Other technologies aimed at defending against the hijacking of control flows are part of the Russian Elbrus processor [21] designed for military grade security and the ability to execute x86 workloads via binary translation very similar to Apple's new M1 chip. Control flow integrity is also entering the x86 mainstream via hardware support of shadow stacks and indirect branch tracking on Intel Tiger Lake CPUs and AMD Ryzen 3 chips.

Randomization and diversification

A single flipped bit can destroy a perfect program, or bring down an airplane or bank. For decades IT engineers have toiled to ensure that, even in the smallest structures, the fastest fabrics, in billions of bits of code and data not a single one would flip …unless so instructed.

This obsession with perfect digital quality ensures that even hundreds of millions of small VAX-like smartphones run an OS which is a bit-by-a-billion-bits, a perfect clone of an original produced through a tightly controlled process somewhere in a data centre.

While it is designed to ensure that each will behave exactly as intended, it exposes them all to the very same exploit tailor fit to any of its clones. That quality helps build the business case of malware creators.

Sexual reproduction introduces several random shuffles into genetic code that to this point must have been rather good to reach maturity with the two base subjects. The genetic code quality of the resulting offspring must be inferior on average as a result of the random changes. Only a sufficiently long life can prove some of the resulting samples are also good enough and that a few are better, and will even survive that new virus which killed parents and siblings. It took a lot of pressure from killer viruses to come up with these random shuffles and sexual reproduction: most living organisms still stick to cloning.

And it is only as a consequence of constant cyberwar attacks that IT professionals bend their 100% quality digital cloning mindset and start introducing carefully controlled random shuffles to destroy malware authors' business cases. Of course, they started small, e.g. with ASLR [22] which shuffles the position of key application image segments in a process' virtual address space on loading, while KASLR randomizes kernel addresses at boot.

Computer viruses cannot really be detected as Emil Post had proven [23], so virus scanners only compare memory contents against a list of known strings. To evade detection, virus authors invented code morphing, random reshuffles of their code that didn't change its functionality. This technique can also be implemented via small changes to compilers to generate variants of every basic block with distinct code generation and optimization settings. As part of their optimization, VLIW compilers regularly create dependency graphs for statements and sub-expression

and identify portions of code which are independent of each other and reshuffle them for a wider execution. That logic can be recycled to create basic block variations all written into object files.

It then only requires small changes on the linker to randomly chose a different variant for each basic block as applications are loaded into distinct systems to ensure that even on millions of systems, few true clones will appear, while the compiler vouchsafes that all variants are valid machine language transformations of the original source code.

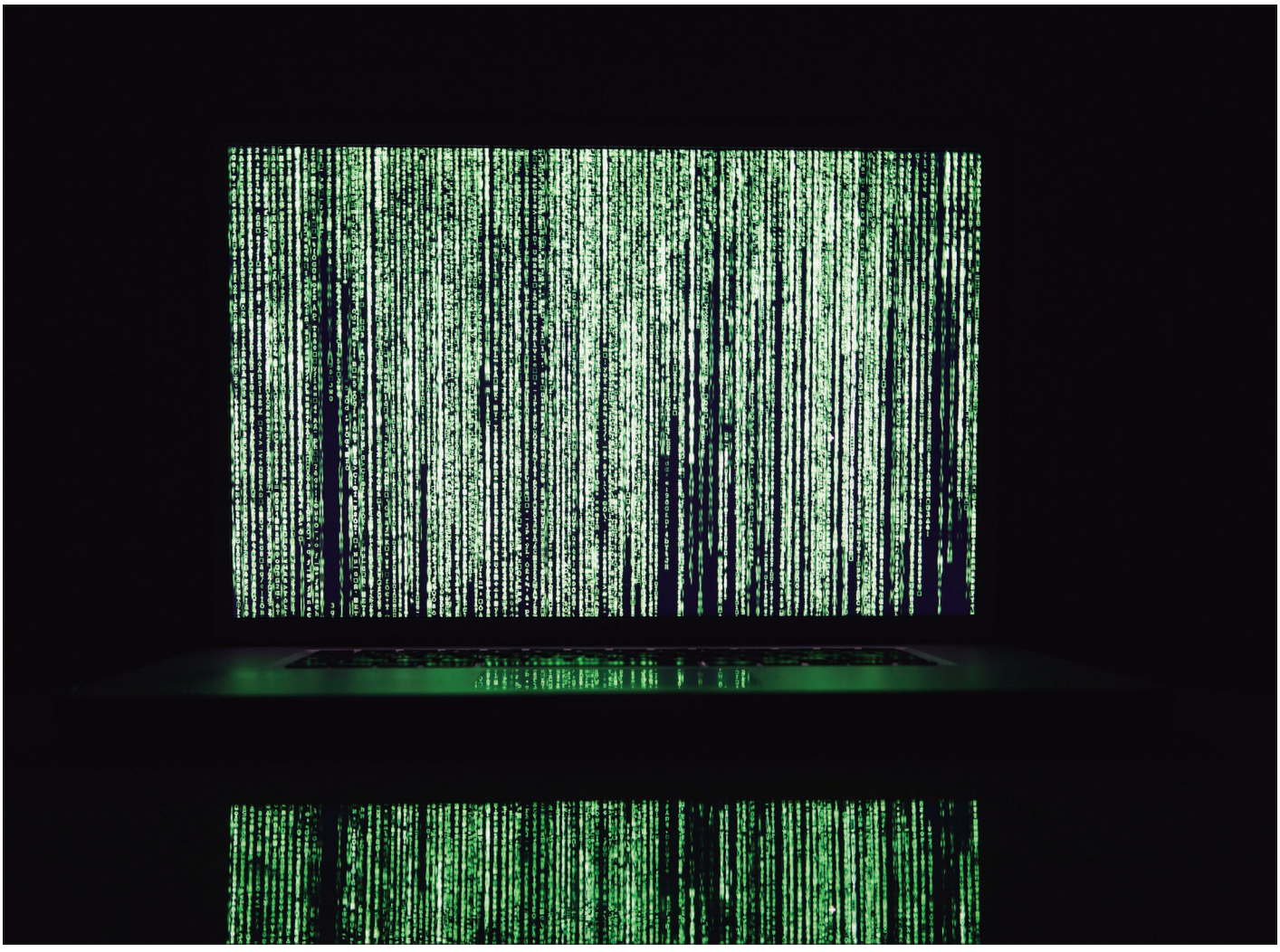
Of course, all this shuffling can break the code. Or rather, it can expose existing errors in distinct ways. Initial self-tests with restarts may solve some issues, but since the random key chosen by the link loader for the basic block shuffle can be recorded and re-created for diagnostics, such bugs can be fixed eventually.

While this approach works on existing systems without additional hardware, a much stronger randomization solution with a customized RISC-V core called Morpheus [24] offers a much stronger protection. It uses a combination of tagging, which allows distinguishing code, data and pointers for both via information the compiler encodes in the object code from their usage in the high-level language source code, (code/data) domain encryption and a cryptographically generated 60bit offsets for code and data pointers, which can be continuously re-encrypted with a new offset in a process called churn.

This assumes that attackers will always find ways to obtain pointer data to critical data structures, but also, that this process e.g. via side channel attacks, will require time related to the strength of entropy in the encrypted data. And it then ensures that any such data loses its usefulness long before its determined, because all code and data pointers are regularly replaced with another 60-bit displacement. Tagging violations, e.g. using a pointer marked as data pointer via the compiler as jump or branch target can be set to cause extra churn cycles to dynamically increase security under attack.

Safe and diverse outside walled gardens

For the internet giants, bringing client devices under the control of their cloud security mechanisms and processes is not only the path of least effort, but also locks their customers in and allows to strip mine their data for profit. For corporate and private consumers, it limits the value they can obtain from cyber-physical systems under vendor control to the level of trust they have in these vendors, while governments are less and less comfortable yielding control and taxation to corporate ethics.



Technologies like the ones highlighted here, show that the exclusivity is not at all necessary and indeed counter-productive to achieve the future large scale – terascale – economy. For that true interoperability and vendor independence, multi-sovereignty enclaves and a verifiable and flexible sovereignty conflict resolution framework are required. They will offer a peaceful co-existence of truly private and personal applications, as well as employer, government, vendor, telco, and various service provider enclaves on each device.

Regulation should set that in motion, while an accelerated evolution of open source and technical assets within the EU should be supported via matching projects and initiatives.

Automated conflict resolution

With the help of the internet and connected machinery the friendly gentleman sitting across the coffee table may be about to nuke the planet or selling child pornography. Enabling or blocking communication with a power plant half way across the globe or videos of naked children playing are evidently not activities that can be managed with a universal set of rules. Plenty of context needs to be taken into account and enforced in a reliable manner. And while that context might not reflect 1:1 the physical world, it might simulate it in a manner that is almost as natural. For example, it might treat your communication with your spouse's phone as if you were sitting next to her, while the device-to-device communication with the man across the table passes through some rather public scrutiny, because you share nothing but a table.

Neither privacy nor social controls can be absolute without destroying the value of personal computing devices or indeed harming society. Current debates are far too binary and show that the current walled garden approach is too simplistic. A plethora of new challenges for technology and regulation lay ahead, which Europe can lead and sell globally if it moves quicker and with a motivation that the United States and China for example do not share yet within their domestic markets.

References

[1] Emil L. Post, "Recursively Enumerable sets of Positive Integers and Their Decision Problems", https://pdfs.semanticscholar.org/ed71/ebe0eee4f88f095247c8b62ba1d3b217a68d.pdf

[2] John von Neuman, "First Draft of a Report on the EDVAC", https://web.nmsu.edu/~davidp/hist_projects/EDVAC.pdf

[3] John von Neuman, "Theory of Self-Reproducing Automata", http://cba.mit.edu/events/03.11.ASE/docs/VonNeumann.pdf

[4] Catalin Cimpanu, "BrickerBot Author Retires Claiming to Have Bricked over 10 Million IoT Devices", https://www.bleepingcomputer.com/news/security/brickerbot-author-retires-claiming-to-have-bricked-over-10-million-iot-devices/

[5] Kurt Gödel, "Uber formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I", http://www.w-k-essler.de/pdfs/goedel.pdf

[6] Wilm Boerhout, "A working VAX 11/780 – revisited", https://vxcompany.com/2016/02/13/a-working-vax-11780-revisited/

[7] Ryan Roemer, Erik Buchanan, Hovav Shacham and Stefan Savage, "Return-Oriented Programming:Systems, Languages, and Applications", http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.216.778&rep=rep1&type=pdf

[8] Denise Giusto Bilic, "Semi-annual balance of mobile security 2019", https://www.welivesecurity.com/2019/09/05/balance-mobile-security-2019/

[9] General Synamics Mission Systems, "Hypervisor", https://gdmissionsystems.com/products/cross-domain-solutions/hypervisor

[10] Lucian Armasu, "Intel Follows AMD's Lead on Full Memory Encryption", https://www.tomshardware.com/news/intel-mktme-amd-memory-encryption,39467.html

[11] Thomas Unterluggauer, Mario Werner and Stefan Mangard, "MEAS: memory encryption and authentication secure against side-channel attacks", https://link.springer.com/article/10.1007/s13389-018-0180-2

[12] Brian Benchoff, "What you need to know about the Intel Management Engine", https://hackaday.com/2017/12/11/what-you-need-to-know-about-the-intel-management-engine/

[13] David Seston, "Meet the Microsoft Pluton processor – The security chip designed for the future of Windows PCs", https://www.microsoft.com/security/blog/2020/11/17/meet-the-microsoft-pluton-processor-the-security-chip-designed-for-the-future-of-windows-pcs/

[14] Henry M. Levy, "Capability-Based Computer Systems", https://homes.cs.washington.edu/~levy/capabook/

[15] Wikipedia, "British Armed Forces communications and information systems", https://en.wikipedia.org/wiki/British_Armed_Forces_communications_and_information_systems#Ptarmigan_(obsolete)

[16] Google, "Fuchsia", https://fuchsia.dev/fuchsia-src/concepts.md

[17] Qualcomm, "Pointer Authentication on ARMv8.3", https://www.qualcomm.com/media/documents/files/whitepaper-pointer-authentication-on-armv8-3.pdf

[18] Arm, "Armv8.5-A Memory Tagging Extension", https://developer.arm.com/-/media/Arm%20Developer%20Community/PDF/Arm_Memory_Tagging_Extension_Whitepaper.pdf?revision=ef3521b9-322c-4536-a800-5ee35a0e7665&la=en&hash=D9F2FA87FEA090C2B20938F09BBAC71698FA18BA

[19] Robert N. M. Watson, Peter G. Neumann, Jonathan Woodruff, Michael Roe, Hesham Almatary, Jonathan Anderson, John Baldwin, Graeme Barnes, David Chisnall, Jessica Clarke, Brooks Davis, Lee Eisen, Nathaniel Wesley Filardo, Richard Grisenthwaite, Alexandre Joannou, Ben Laurie, A. Theodore Markettos, Simon W. Moore, Steven J. Murdoch, Kyndylan Nienhuis, Robert Norton, Alexander Richardson, Peter Rugg, Peter Sewell, Stacey Son, Hongyan Xia, "Capability Hardware Enhanced RISC Instructions: CHERI Instruction-Set Architecture (Version 8)", https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-951.pdf

[20] Wei Song, Alex Bradbury and Robert Mullins, "Towards General Purpose Tagged Memory", https://riscv.org/wp-content/uploads/2015/06/riscv-tagged-mem-workshop-june2015.pdf

[21] Alexander Kim, Ignat Bychkov, Vladimir Volkonskiy, Feodor Gruzdov, Sergey Semenikhin, Vladimir Tikhorsky, Vladimir Feldman, "Russian Microprocessors of the Elbrus Architecture Series for Servers and Supercomputers", http://russianscdays.org/files/talks/VVolkonsky-RSCDays-2015.pdf

[22] Brad Spengler, "PaX", https://grsecurity.net/PaX-presentation.pdf

[23] Wikipedia, "Post correspondence problem", https://en.wikipedia.org/wiki/Post_correspondence_problem

[24] Mark Gallagher, Zelalem Birhanu Aweke, Austin Harris, Valeria Bertacco, Lauren Biernacki, Salessawi Ferede Yitbarek, Zhixing Xu, Sharad Malik, Todd Austin, Shibo Chen, Misiker Tadesse Aga, Baris Kasikci, Mohit Tewari, "Morpheus: A Vulnerability-Tolerant Secure Architecture Based on Ensembles of Moving Target Defenses with Churn", https://web.eecs.umich.edu/~barisk/public/morpheus.pdf

[25] Digital Common Wealth, "Digital Equipment Corporation VAX 11/780 mainframe computer, Maynard", https://www.digitalcommonwealth.org/search/commonwealth:sn00b086d

[26] Welivesecurity, "Semi-annual balance of mobile security 2019", https://www.welivesecurity.com/2019/09/05/balance-mobile-security-2019/

Thomas Hoberg is Technical Director R&D at Worldline, Germany.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: T. Hoberg. Reversing John von Neumann and Steve Jobs, but not software. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 80-87, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

There is growing awareness of the importance of privacy while, at the same time, we are sharing ever more private data with third parties. This creates an uneasy tension.

Privacy: whether you're aware of it or not, it does matter!

By BART COPPENS and OLIVIER ZENDRA

While privacy used to be a concern of only a limited number of people, in recent years awareness of it has been growing. This has been for a number of reasons including the enactment of the GDPR, the growing impact of data leaks, data logging by governments and companies, and even the recent discussions about COVID-19 contact tracing. At the same time, most of us are knowingly or unknowingly sending more and more private data to the cloud, which increases the risk of it being leaked or abused in some way.

In order to try and reconcile these two opposing directions, consumers and companies alike should increase their usage of privacy-enhancing technologies, and businesses should integrate privacy by design into their development.

Key insights

- Ever more data is being sent to and collected by governments and private companies alike.
- The scope and volume of the data being collected and analyzed is often not clear to consumers, who are even sometimes completely unaware.
- However, due to the GDPR and COVID-19, there is an increasing public awareness of privacy, not only of the fact that personal data is being collected, but also that it can be leaked, either on purpose or inadvertently.
- Technical solutions exist to improve privacy. The EU has a role to play.

Key recommendations

- The EU should promote research into technologies that enhance individuals' privacy and reduce the risks and impact of leaks of private data.
- The EU should encourage or even require companies to actively adhere to the principles of privacy by design.
- The EU should stand by its principles of privacy for its citizens, and not allow backdoors being put into applications.
- Existing solutions that limit leakage of personal information should be promoted by the EU.

We live in an era in which *almost everything we do is transmitted to servers beyond our control*. To give just a few examples, our private documents are stored in the cloud, while in some countries internet providers are legally obliged to keep track of which websites we visit [14]. Mobile service providers keep track of where our mobile phones make contact to their base stations and thus keep track of where we are; when we drive, our vehicle licence plates are captured by more and more automatic number-plate recognition (ANPR) cameras which are placed for various purposes by governments and municipalities [15,16,17]. The list goes on. People even freely put microphone-based listening devices such as Amazon Echo [11], Apple Siri devices [12], Google Nest [13], etc. in their homes for purposes of convenience and comfort.

Sometimes this sharing of information is quite intentional. As a matter of fact, sharing of information online has dramatically increased as a result of the sharp rise in home working caused by the COVID-19 pandemic lockdowns that also boosted e-commerce. When people do share a document online with others, they fully expect this information to be shared only with those specific individuals. However, the fact that this document is stored on servers – which can be hacked and can leak their documents – is something that most people forget. In addition, *most of the time people do not even realize the extent to which their private actions are tracked or shared with others.* People are surprised to find not only that their listening devices send out snippets of their private conversations to the companies that made their device, such as Amazon, but also that these private snippets are sent out to subcontractors who listen to them in order to increase the accuracy of the voice recognition engines that power these devices [18]. There is thus a real issue surrounding privacy and awareness of privacy issues.

It is thus clear that privacy is an important topic that directly affects the lives of many people. In the remainder of this article, we first discuss in more detail the kinds of **personal and private information** that nowadays are being generated, collected, and potentially leaked. We then describe

Image: ID 93286496 | © Valerii Brozhinskii | Dreamstime.com

some of the **technical directions** that can lead us to **protect** our **data and privacy** better.

Personal and private information

As a society, we are generating and storing ever-increasing amounts of (private) data. This includes the (confidential) data of companies. Almost all of this data, regardless of its source or whether or not it was sent intentionally by a user, is sent to and stored in cloud-based servers. This basically boils down to a consolidation of the software and hardware stacks of different users in the cloud. Because the infrastructure is shared between many different users, and is not located locally with these users, these cloud-based systems are much more vulnerable than locally run systems if they were to be unconnected from the network. Because both private customers and businesses need their cloud-based data to be secure from third-party snooping, leaking and interference, these systems need to be protected against many different kinds of attack.

Thanks to *regulations such as the General Data Protection Regulation (GDPR)*, European Union citizens should be better protected against at least some forms of unwanted processing of private data, and they should now at least be informed when such data is leaked or mishandled. This represents huge progress in terms of the public being informed and aware of data protection matters and should be hailed as a very positive step. Still, this does not mean that data leaks have magically gone away. For example, Figure 1 shows the number of data breaches involving US healthcare data, where in each case at least 500 records were leaked. The trend is unfortunately in the direction of more data breaches, not fewer.

Furthermore, even though we as European users of data platforms should now be *informed about the fact that data is collected, most people are unclear about the scope of the increasing amount of data that is being collected, processed, and stored*. The

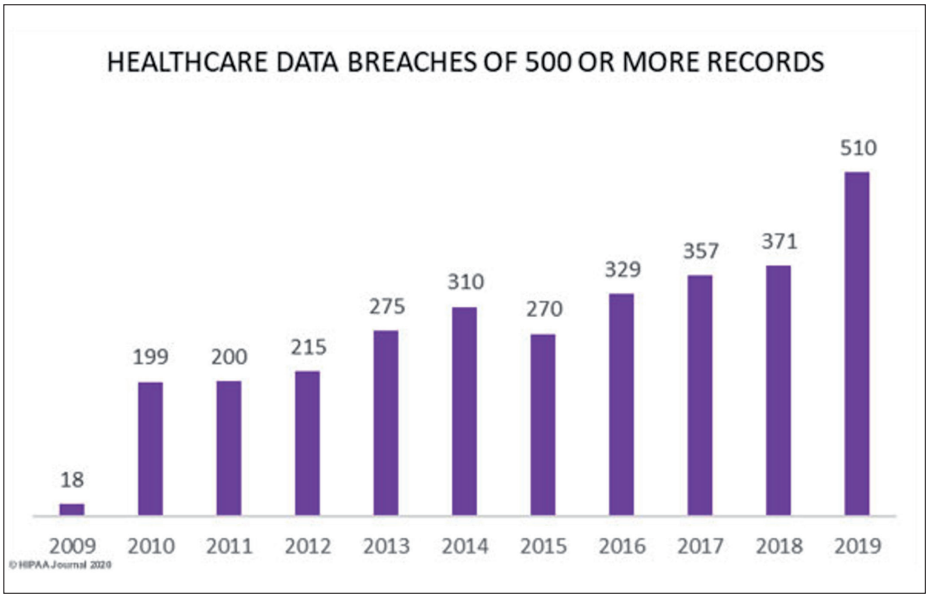


Figure 1: Number of breaches of 500 or more medical data records, as reported to the US Office for Civil Rights, Department of Health and Human Services (HHS) in the United States since October 2009 [1].

trend towards a service-based cloud economy only serves to exacerbate the scope to which this data is shared, and the risks to which this data is exposed.

While leaks of all kinds of data loom ever larger in the background of today's society, it typically remains a phenomenon that, in the mind of most people, is a concern only for others, rather than something that they feel will ever affect them (until it actually does, of course). Still, when (the threat of) a leak of private information does indeed seem imminent, sometimes it does indeed lead to more widespread debate. This happened only recently in the context of the COVID-19 pandemic. While contact tracing is something that had already happened for some other (but contained) infectious diseases such as tuberculosis, it had only been done manually and on a small scale by having contact tracing staff interview people directly. However, now that most people have a smartphone, technology has advanced enough to provoke proposals for contact tracing to be done by electronically tracking people's interactions with one another. In many countries, this sparked intense discussions about which design criteria such an app should adhere to: should it be based on location data and/or on Bluetooth-based proximity data, should it be totally anonymous or not, should the data be stored in a centralized or decentralized fashion, should it be mandatory or not, etc.

Despite these lengthy discussions on the very specific topic of contact tracing in the context of COVID-19, many people are still unaware of the extent to which data similar to this is already being kept track of. First and most obvious is the data that is already being kept that is related to government activities (ANPR cameras, cell phone data, ...). However, because of the privatization of the many functions of government, much of this data is already being kept by private companies.

Worse still, *even outside of such government-related activities, private companies are keeping track of increasingly more detailed information about people who are not even necessarily their users*. People's online activities are tracked by advertising companies such as Google and Facebook, so that they can then sell ever-more targeted ads [22,23]. They use data analysis techniques to create a profile of your interests, and even very private information such as your sexual orientation [28]. The users of such platforms are not always explicitly consenting to providing such information about their interests; user interests can be inferred implicitly using machine learning techniques which base themselves on the online behaviour of the tracked individuals in question. Some advertising companies go to quite some lengths to circumvent some anti-tracking measures that are implemented by browsers [24]. Some totally unscrupulous companies even scrape the internet for people's pictures from different social media channels and other websites, in order to build and train very accurate facial recognition of these people without their consent [25]. These facial recognition engines can then be sold to governments and commercial organizations across the world [26,27]. This tracking of data has further ramifications with regards to sovereignty. US or Chinese companies keep track of the data of EU citizens, harvest their pictures and process this data abroad, outside EU laws and regulations; this can create problems that are hard to address and solve.

Unfortunately, and quite surprisingly given the public debates over the COVID-19 tracking, no such sizeable debate is happening over these other kinds of even more widespread tracking and collecting of data. Still, we are hopeful that people will become increasingly aware of the fact that they do not want this kind of private and personal information to be used indiscriminately by parties beyond their control.

Technical means to protect our data and privacy

Now, how best to protect your data? The easiest solution would of course be to not share this data at all: unshared data truly is private data. However, the extent to which data is already being shared as soon as we try to interact with today's heavily digital society makes this infeasible except for people who are willing to become (partial) digital and social recluses. So, a middle ground needs to be found. This can be done at least in part by promoting and choosing technologies that enhance your privacy, rather than ignore it, or even worse, try actively to circumvent it.

In order to do so, we need to *design systems from the ground up with privacy in mind*. How a system handles private data and how it deals with privacy, should be design requirements from the start. In 2010, the International Conference of Data Protection and Privacy Commissioners published a resolution encouraging the recognition of the fact that privacy by design is an essential component of privacy protection, as well as the adoption of a set of foundational principles of privacy by design [29]. One of these principles is that privacy should be *embedded* in the design, and it should be an essential core of the functionality of the system [30]. These guiding principles were as true and necessary then as they are now, and their importance has only grown.

For example, rather than sharing documents and messages with people where the shared data is stored in an unencrypted form on cloud servers, we can try to use solutions with *true end-to-end encryption*. For example, one could use Signal [10] or EU-based Olvid [9] for sending messages to one another, rather than, for example, Facebook Messenger, as the former encrypt the data such a way that intermediate servers cannot decrypt the messages. In the latter case, Facebook has access to the original plain-text messages. When true end-to-end communications are not possible, or when people risk being tracked, it would still be beneficial to at least choose a *technology or solution which explicitly focuses on the privacy of its users*, like for example the EU-based Qwant search engine [31], or the Brave browser [32], that put an emphasis on protecting their users' privacy. The EU should encourage more initiatives and further developments and investments into such privacy-aware technologies and companies, to help protect its citizens and its sovereignty over data.

Of course, a good and easy way to have fewer problems with private data potentially being compromised is to not send it over the network at all. One way in which this

Credit: | ID 117352101 | ©Pop Nukoonrat | Dreamstime.com

can be solved is by having most or even all computations that would normally happen in the cloud, now happen locally, with the *processing being done at the edge*. This also has implications for industrial applications in the context of IoT and CPS: the more data is being processed in those devices themselves, rather than that the data has to be transmitted to cloud servers for processing, the less private data can be abused or leaked. A fog or federation of local devices sharing part of the global information in an encrypted way could solve the problem of accessing larger computing or storage resources in a more local manner.

If data does need to be transmitted or computed remotely, it is important to do this in a secure fashion that preserves as much security and privacy as possible. Most companies already try to *protect most sensitive data at rest and in transit with encryption*, for example with the Advanced Encryption Standard (AES) and Transport Layer Security (TLS). However, this data still needs to be processed, for which the data is currently still decrypted (and thus unprotected) on the systems that process it. Furthermore, if this data processing involves the data being searchable or queryable in a database, many systems will still store this data in an unencrypted form. One way to mitigate this problem is to *do the data processing on encrypted data*, in such a way that the personally identifiable information (PII) is not known to the system performing the actual processing. Examples of such techniques are (fully) homomorphic encryption (FHE), which still requires research to decrease its computing resource requirements, and secure multi-party computation. There are many fields in which homomorphic encryption would significantly increase the privacy of data in the presence of cloud-based data processing. In the medical sector, users would be able to upload their ECG data and have a cloud provider monitor their health without actually sharing their data with that cloud provider [2]. Similarly, we would be able to have our genome analyzed by third parties without information being passed on about which genetic diseases we have or other PII such as gender, race, etc [3]. Modifying different cloud-based machine learning tasks to protect PII would also significantly reduce the risks associated with outsourcing the relevant data. For example, face verification or face recognition would no longer expose photographs of people [4], and performing optical character recognition would no longer leak the text being processed [5].

Furthermore, if the recognized text is from licence plates that need to be queried in a database of stolen and wanted vehicles, for example, you can prevent the processing of all licence plates from leaking information about non-stolen cars [6]. The EU should invest in more technologies such as these, so that if PII data does need to be processed, the amount of data that can be intentionally or inadvertently leaked is minimized as much as possible.

Given the urgency for today's business landscape of the need to achieve more robust data privacy systems, we predict and advocate for an increase in the design and use of such homomorphic encryption and related techniques. Some start-ups already provide very specific applications of these techniques [7]. One limiting factor in applying FHE right now is its overhead. Both the time needed to process the data and the size of the messages that need to be exchanged with the cloud provider currently increase dramatically when FHE is applied. At present, this means that many of those techniques are unfortunately not yet usable in practice. In the meantime, some specific cases might not need to send the PII itself to third parties. Another issue to take into account when protecting data by encrypting it is how resistant the encryption scheme is to the changing landscape of attackers' capabilities. One clear but constant change is the increase in the processing speed of computers. As one of the most obvious goals of an attacker is to recover the information, the question is how long information can remain private, and how this time decreases with an increase in processing speed, and by how much we then increase the strength of the encryption (for example, by increasing the key size) to compensate for this. For traditional computers, it is quite clear how these scaling laws work, and increases in computing power do not immediately threaten the security of data encrypted with traditional encryption schemes. However, when switching to the different computing paradigm of quantum computers, this is not necessarily the case, because certain algorithms are believed to run significantly faster on quantum computers than on traditional computers. With some algorithms, it is sufficient to choose larger key sizes to compensate for this. However, other algorithms can be completely broken with quantum computers. Such algorithms need to be replaced with algorithms that could withstand attacks from a quantum computer. This field is called post-quantum cryptography.

However, it is not sufficient to use state-of-the-art encryption algorithms to protect PII. Software that is not secure can obviously leak all kinds of confidential and private information to attackers, even if under normal circumstances this data is stored and transmitted securely. Some *security-related Instruction Set Architecture extensions* have explicit implications for improving privacy. For example, one of the goals of Intel's Software Guard Extensions (SGX) is to protect the execution of certain code fragments from attackers that have control over the rest of the system, including the operating system itself. This can then be used to protect sensitive and private information even when the entire system is being attacked. However, the many recent attacks on SGX show that even this technology is clearly not yet mature enough to withstand such attacks in practice [19,20,21]. It may even be that the SGX model of allowing execution of code on private data, and general-purpose code execution by untrusted users, might not be feasible.

In this context, it is important to stress the importance of the entire system being *secure and not weakened by backdoors*. Some countries have argued for the presence of such backdoors in operating systems, telecommunications systems, and secure/encrypted communication platforms, such that only "they" can (lawfully) gain access to systems and decode encrypted information. These backdoors reduce the security of the entire system, since there is no guarantee that the law enforcing agents of your own country will be the only ones with access to these backdoors: other countries and criminals might be able to use them too. For example, a pseudo-random number generator containing a weakness in it which had allegedly been introduced by the NSA, eventually found its way into firewalls, where it was exploited by unknown parties [34]. Some people even claim that Intel's closed-source management engine on chips does not only support good-intentioned remote management features, but could also be used by other (malicious) parties to remotely gain access to machines [35].

These backdoors also reduce the overall trust people have in computers and telecommunication systems, thus undermining all the efforts of the EU to increase the privacy and security of its citizens. Not only that, such measures will also affect the confidential data of companies, which would then also become vulnerable to being leaked through these backdoors as well. Thus, to protect both the privacy of its citizens, and the confidential data of its companies, the EU should not give in to calls to action to even consider legalizing such backdoors.

However, while an insecure system can lead to information leaks, the converse is not necessarily true. A secure system cannot distinguish between purposeful leaks of information (for example, a user who wants to print his/her own bank statements), versus inadvertent leaks of information (for example, these bank statements being stored unencrypted on disk). One possible solution here is language-based information-flow security that allows programmers to explicitly define which flows of information are allowed, and to define properties on these flows [8].

A final source of leaking private information are the users themselves: often they are not aware of the actual private information that can be extracted from the data being shared: posting pictures of somebody in a bar or in a nightclub might be interpreted by an insurance company as somebody being a health risk because they drink alcohol. There are artificial intelligence-based systems that can analyse such public content and can warn users of this "side channel" information [33].

Conclusion

Public awareness of privacy issues is slowly increasing thanks to initiatives such as the enactment of the GDPR and to the issue of contact tracing for COVID-19. These will hopefully be a trigger for people

to think more about where and how their private data is being collected, stored, and used, which in many cases is anyplace, anytime, by most of the helper tools and applications (smartphones). This will hopefully lead people to try more actively to protect their own privacy. The EU has a role to play in terms of regulation and promoting and financing privacy and sovereignty preserving EU-based solutions.

References

[1] Healthcare Data Breach Statistics, HIPAA Journal, online, accessed December 3, 2020. https://www.hipaajournal.com/healthcare-data-breach-statistics/

[2] Kocabas, Ovunc, et al. "Assessment of cloud-based health monitoring using homomorphic encryption." 2013 IEEE 31st International Conference on Computer Design (ICCD). IEEE, 2013

[3] Miran Kim, Kristin Lauter, "Private genome analysis through homomorphic encryption", BMC Med Inform Decis Mak. 2015; 15(Suppl 5): S3.

[4] J.R. Troncoso-Pastoriza, D. González-Jiménez, F. Pérez-González, 2013. Fully private non-interactive face verification". IEEE Transactions on Information Forensics and Security, 8(7), pp.1101-1114

[5] Nathan Dowlin, Ran Gilad-Bachrach, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. 2016. "CryptoNets: applying neural networks to encrypted data with high throughput and accuracy. In Proceedings of the 33rd International Conference on International Conference on Machine Learning – Volume 48 (ICML'16), Maria Florina Balcan and Kilian Q. Weinberger (Eds.), Vol. 48. JMLR.org 201-210.

[6] Sunil, Archana Bindu, Zekeriya Erkin, and Thijs Veugen. "Secure matching of Dutch car license plates." Signal Processing Conference (EUSIPCO), 2016 24th European. IEEE, 2016

[7] https://www.privatebiometrics.com/, accessed December 4, 2020

[8] A. Sabelfeld and A. C. Myers, "Language-based information-flow security", in IEEE Journal on Selected Areas in Communications, vol. 21, no. 1, pp. 5-19, Jan. 2003.

[9] Olvid. https://olvid.io/technology/en/

[10] Signal. https://signal.org/docs/

[11] Amazon Echo. https://en.wikipedia.org/wiki/Amazon_Echo

[12] Apple Siri. https://en.wikipedia.org/wiki/Siri

[13] Google Nest. https://en.wikipedia.org/wiki/Google_Nest_(smart_speakers)

[14] Judgments in Case C-623/17, Privacy International, and in Joined Cases C-511/18, La Quadrature du Net and Others, C-512/18, French Data Network and Others, and C-520/18, Ordre des barreaux francophones et germanophone and Others. Press release 123/20, Court of Justice of the European Union, 6 October 2020

[15] Automatic Number Plate Recognition, Police.uk https://www.police.uk/information-and-advice/automatic-number-plate-recognition/ accessed December 2020

[16] Denmark: Targeted ANPR data retention turned into mass surveillance EDRi, September 6, 2017, https://edri.org/our-work/denmark-targeted-anpr-data-retention-turned-into-mass-surveillance/

[17] Automatic number-plate recognition - Usage, Wikipedia, https://en.wikipedia.org/wiki/Automatic_number-plate_recognition#Usage Accessed December 2020

[18] Apple contractors 'regularly hear confidential details' on Siri recordings, The Guardian, July 26, 2019.

[19] Foreshadow - Extracting the Keys to the Intel {SGX} Kingdom with Transient Out-of-Order Execution. USENIX Security Symposium 2018.

[20] Plundervolt: Software-based Fault Injection Attacks against Intel SGX. Murdock et al. IEEE Symposium on Security and Privacy 2020

[21] CrossTalk: Speculative Data Leaks Across Cores Are Real. Ragab et al. Accepted in the IEEE Symposium on Security and Privacy, 2021.

[22] Why targeted ads are the most brutal owns. Vox, September 25, 2018. https://www.vox.com/the-goods/2018/9/25/17887796/facebook-ad-targeted-algorithm

[23] Google's ad tracking is as creepy as Facebook's. Here's how to disable it. The Guardian, October 21, 2016. https://www.theguardian.com/technology/2016/oct/21/how-to-disable-google-ad-tracking-gmail-youtube-browser-history

[24] Ad Tech Surveillance on the Public Sector Web, Cookiebot Report, version July 14, 2020. https://www.cookiebot.com/media/1136/cookiebot-report-2019-ad-tech-surveillance-2.pdf

[25] Scraping the Web Is a Powerful Tool. Clearview AI Abused It. Wired, January 25, 2020. https://www.wired.com/story/clearview-ai-scraping-web/

[26] Clearview's Facial Recognition App Has Been Used By The Justice Department, ICE, Macy's, Walmart, And The NBA. Buzzfeed, February 27, 2020. https://www.buzzfeednews.com/article/ryanmac/clearview-ai-fbi-ice-global-law-enforcement

[27] Secret Users Of Clearview AI's Facial Recognition Dragnet Included A Former Trump Staffer, A Troll, And Conservative Think Tanks. Buzzfeed, March 11, 2020. https://www.buzzfeednews.com/article/ryanmac/clearview-ai-trump-investors-friend-facial-recognition

[28] Researchers Claim Facebook Ads Could Out LGBTQ+ Users. Out, August 30, 2019. https://www.out.com/tech/2019/8/30/researchers-claim-facebook-ads-could-out-lgbtq-users

[29] Resolution on Privacy by Design. 32nd International Conference of Data Protection and Privacy Commissioners. Jerusalem, Israel 27-29 October, 2010. https://edps.europa.eu/sites/edp/files/publication/10-10-27_jerusalem_resolutionon_privacybydesign_en.pdf

[30] Privacy by Design: The 7 Foundational Principles. Ann Cavoukian, Ph.D., Information & Privacy Commissioner of Ontario, Canada. https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf

[31] Qwant. Accessed December 2020. https://www.qwant.com/?l=en

[32] Brave. Accessed December 2020. https://brave.com/

[33] https://www.researchgate.net/publication/301221061_Personalized_Privacy-aware_Image_Classification

[34] Researchers Solve Juniper Backdoor Mystery; Signs Point to NSA, Wired, December 22, 2015. https://www.wired.com/2015/12/researchers-solve-the-juniper-mystery-and-they-say-its-partially-the-nsas-fault/

[35] Is the Intel Management Engine a backdoor? TechRepublic, July 1, 2016. https://www.techrepublic.com/article/is-the-intel-management-engine-a-backdoor/

Bart Coppens is Postdoctoral Researcher in the Electronics department of Ghent University, Ghent, Belgium.

Olivier Zendra is a Tenured Computer Science Researcher at Inria, Rennes, France.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: B. Coppens and O. Zendra. Privacy: whether you're aware of it or not, it does matter! In M. Duranton et al., editors, HiPEAC Vision 2021, pages 88-93, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

Parts of software engineering can be likened to out-of-the-box creation, other parts as science. The latter yield a variety of prescriptions and rules. The staggering complexity of modern systems causes the body of such prescriptions to grow inordinately, making their application dauntingly labour-intensive. AI applications may bring much needed relief in this ambit.

The artificial programmer

By HARM MUNK and TULLIO VARDANEGA

At present, the production of software is not a prime-time goal for the advances of artificial intelligence (AI) fuelled by compute-hungry deep learning. Yet, recent surveys show that AI has found application in virtually all software engineering tasks. To date, this has been more so in requirements engineering and testing, and less so in implementation and maintenance. This disproportion reflects the fact that the immediate target of AI-based assistance tools in any human production effort, including software development, invariably tends to be routine work. The requirements of engineering and testing tasks are generally acknowledged as the most labour-intensive ones (and the least rewarding because of their introvert slant). This is why they have gotten earlier attention from AI. As programming becomes increasingly an integration task and less a clean-sheet creation however, it is easy to predict that AI will enter and progressively dominate the stage of automated code generation. When that happens, the nature of software design as a speaking-to-the-human prelude to coding will be called into question, while maintenance and refactoring will have to ingest the same "AI-logic" that drives automated code generation. That drive will cause software engineering to transform itself so as to keep control (for trust, traceability, and explainability) over the end-to-end process throughout the penetration of AI in its elemental processes.

Key insights

- AI-fuelled tools may provide much-desired assistance and relief in virtually all of the software lifecycle processes.
- The software engineering tasks that are more immediately amenable to the application of AI assistance are those that involve routine, repetitive, mechanical work whose execution follows precise rules and mechanisms.
- Software engineering will not go away in the meanwhile: it will transform itself so as to keep control across each individual process and task, and assure trust, traceability and explainability of all the ensuing artefacts.

Key recommendations

- Promote and support the experimentation of AI-fuelled engines to support specific software tasks.
- Promote the augmentation of software engineering guidance in a manner that can exert control and yield assurance over all products of AI-assisted software engineering tasks.

Introduction

If you are a software engineer, and you are eyeing the flurry of developments in artificial intelligence (AI) with some anxiety because you fear that AI applications might snatch jobs away from you, rest assured that there is still a long and uncertain journey ahead before that prospect may become reality. In fact, if you are a "true" (i.e. authoritatively trustworthy) software engineer, then you might as well cheer up instead, because your skills will continue to be in high demand – including by AI itself, surprise, surprise – and there are so very few of you around.

As the reader is likely to have experienced, an abundant number of publications claim or prophesise that the impetuous advancements in the capabilities of AI applications is bound to change the way software is produced. That is certainly true, as we begin to see concrete examples of that happening [1]. The most manifest change applies to the development of user applications that *incorporate* AI engines: arguably, this has been the prime motivation for AI and software production to come together. Less prominent but visible in the news has been the development of custom software artefacts *written by* AI, mostly for demonstrative purposes. Far less attention, instead, has gone so far to the empowering-by-AI of the software engineering process itself, which is a much broader and complex concern.

Developing software that applies or uses AI techniques [2] will without doubt ask for a different approach to the software development process than we are used to for traditional software applications. Software engineers will have to master the principles on which these AI-based applications are built to be able to effectively apply such principles in the corresponding software development. The successful application of, e.g. neural networks not only relies on the ability to use neural network hardware and software in an application, but also on the (deep, pun intended) understanding of how such a network can be efficiently trained, most of all by making sure that a sufficiently rich training set is selected and kept up-to-date.

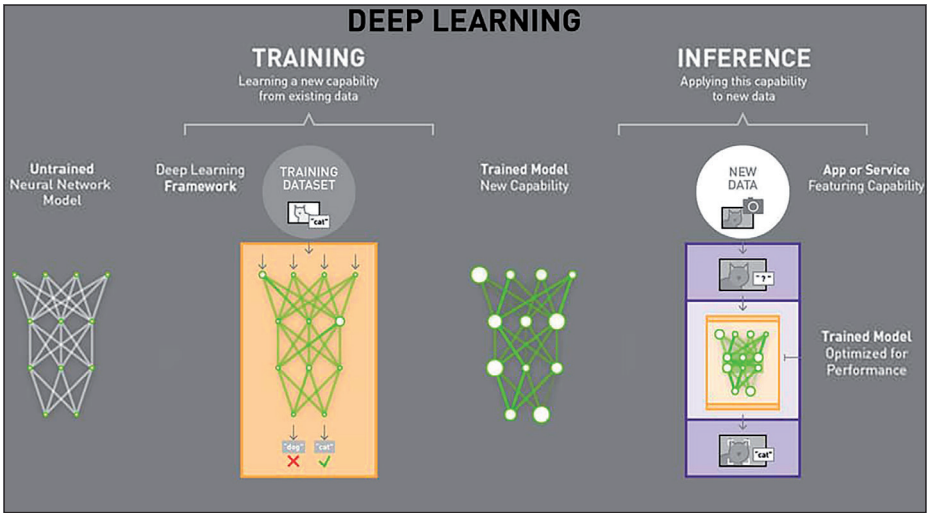


Figure 1: The steps of training an inference model and using the trained model for application purposes. (Source: https://blogs.nvidia.com/blog/2016/08/22/difference-deep-learning-training-inference-ai/)

The skills necessary to accommodate this novel need differ substantially from what was common practice in software engineering just a few years ago. Luckily, the modern-day software engineer is aided in this respect by the emergence of componentised, reusable AI applications, as for example put forward by fast.ai [3], which help to realise a structured, less black-box, approach to the training model.

Artificial intelligence and software development

As noted earlier, it is opportune to make a distinction between the use of AI in software and the use of AI in *software development*. This article focuses on the latter. The expectation that human programming will be obsolete in some (near) future seems to dwell on the expectation that AI-based applications will produce the software program code that, to date, is still the output of a human programmer.

Before we dive into the reality of this prediction, we should recall that software engineering is far more than just coding: it also comprises design, verification and maintenance (see Figure 2). Coding, that is, the act of producing program text (aka source code) is the culmination of a thinking process, and there is no substitute for it. Likewise, delivering source code for a software engineer is not just to commit algorithmic concepts to program text, but to verify that the execution yields correct results, in both functional (what it computes) and non-functional (how it does so) terms, and to make sure the corresponding logic transpires from said text so that future maintenance shall not be mystified.

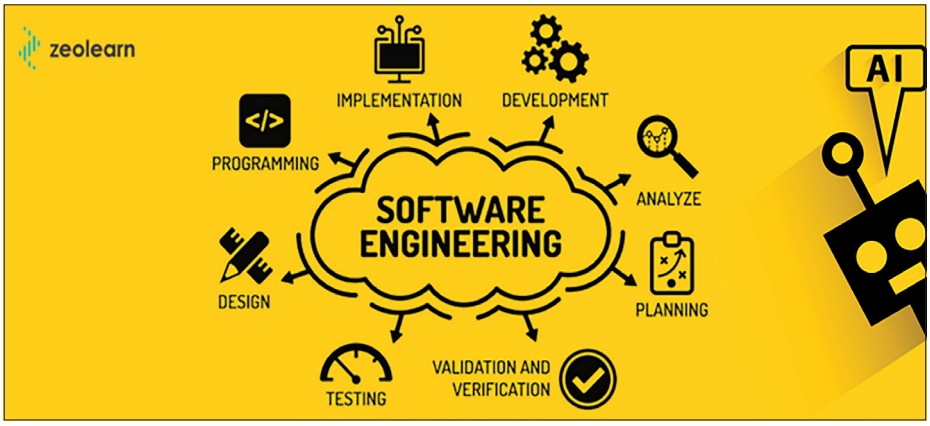


Figure 2: A view of the components of software engineering. (Source: https://www.zeolearn.com/magazine/the-role-of-artificial-intelligence-in-software-engineering)

In that respect, therefore, a more ambitious (and futuristic) title of this article might have been: "The Artificial Software Engineer", to reflect the oft-forgotten understanding that writing program text is only a fraction of the deed and frequently not the most important one, as counter-intuitive as it may seem. Useful good software has many happy users, is frequently used, yields satisfaction, and consequently has a long lifetime. It therefore addresses the user needs well, works well, and is designed and coded well. Not very many software programs have all of these qualities at once, across professional, institutional, social, and personal ambits. This rather frustrating state of affairs is not surprising since the count of "true" software engineers worldwide still is very modest as yet, after all, 52 years from when the discipline was first conjured into existence [4].

More modestly and perhaps more realistically, then, the title of this article is "The Artificial Programmer". This downcasting from grander ambition reflects the fact that current AI is applicable, and has been applied already, to an increasing fraction of the software-engineering stages referred to earlier, but so far in an isolated and segmented manner, i.e. covering more the production of specific development artefacts (hence vertically) than seamlessly supporting larger tracts of software lifecycle overall, from conception to retirement (hence horizontally). It is comparatively easy to predict that the current trend of vertical progress will thrive: commercial reasons, quest for public resonance, and appetite for here-and-now provide abundant fuel to vertical-driven efforts, some of which admittedly have achieved fascinating results [5].

In the sequel we consider each of the five software engineering stages in isolation: requirements engineering, design, implementation, testing, and maintenance. In doing so, we take inspiration from recent reviews [6,7].

Requirements engineering (RE)

RE is the activity to define what a system is supposed to do, which problems it should help solve or which tasks it should execute. Such requirements are most often defined by the customers (perhaps with little or no understanding of software) using natural language. As the requirements as stipulated drive all of the subsequent development (design and implementation, but also verification and maintenance), they must capture the intentions of the customer correctly and consistently.

An overlooked, incorrectly specified, or superfluous requirement may result in costly corrections in later stages of the project lifecycle, if not even in a failed product. Also, requirements must be consistent (i.e. sound, agreed, free from contradiction, without abrupt changes), and must be kept so when changes are applied. If requirements are specified using natural language, then it is very important that the language used for them is very carefully crafted, to avoid confusion, ambiguity, inconsistency, etc. That is often hard on the customer. Using specialized formalisms is less exposed to that risk, but that is an expert task, more suited for the supplier and scarcely so for the customer, at risk of creating undesirable intellectual distance between them. Moreover, whatever the chosen form, keeping the requirement specification correct and consistent all along the development process drains precious energy from the productive side of it, so that such need is frequently neglected or relegated to an admin duty.

Undoubtedly therefore, AI techniques, especially natural language processing (NLP), may offer important aid to the activities of the RE stage. A complicating factor, however, is that requirements are often silently specified in the context of the application domain of the system being developed. In other words, a large fraction of the background information needed to interpret requirements correctly is assumed known and not written down. That makes it necessary for prospective AI-fuelled RE tools to embed a substantial amount of domain knowledge. The effort required to build such domain knowledge is reminded of the goals of ontology research long associated with the so-called semantic web [8] or with the metamodels that underpin model-based or model-driven development [9]. The results that transpire from vocal assistants or chatbots showcase the potential that can be unleashed in this ambit, provided sufficient resources, for computing hardware, data sets, and test oracles (aka tagging), are deployed to this end.

Design

As requirements state the problem, the subsequent stage conceives and realises a solution for it. In classic software engineering, with humans as the development actors, the conception stage (aka design) is paramount, in that it earns several essential benefits. It helps break the bigger prob-

Figure 3: Dilbert's view of the chasm between the user needs and the analyst's understanding of requirements.

lem into multiple, simpler parts (which eases verification), and determines how such parts can be joined together into the final product (which guides integration). It allows reasoning on the goodness of fit of the proposed solution under the project metrics of interest ahead of committing such concept to code (which promotes correctness-by-construction in place of costly correctness-by-correction). It allows multiple programmers to work together, consistently on separate and non-conflicting coding tasks (which speeds development).

Whereas primordial design "invented" software solutions, modern design applies various kinds of patterns, increasingly following "blueprints" with good, proven properties, which fit the need as seen by the solution architect. Historians of software engineering attribute this radical change of attitude to the influence exerted by the publication of Christopher Alexander's (a true, non-software, architect) "*The Timeless Way of Building*" book in 1979. It took another fifteen years for the first authoritative collection of software design patterns to be systematized [13]. After that milestone date, the shift was complete and the design part of software engineering education started focusing on (1) familiarising students with the various taxonomies of design patterns, (2) learning how to single out those that serve the need of the project, (3) fitting them to the design context, (4) transferring them faithfully into program code.

It can therefore be maintained that software design, of late, has become essentially a job of selecting, adapting and composing pre-existing patterns, shedding much of its initial creativeness. This evolution arguably makes modern software design a task fit for knowledge-based systems, possibly combined with automation-geared evolutions of domain- or model-driven design [14,15] or combinations of them.

Two provisos are in order with regards to the plausibility of that prospect, however. First, software design most certainly is a human artefact: it "speaks" to the human perception and guides our understanding of how the smaller pieces (those that we can act on with minute programming) relate to the bigger picture. Without design we humans would be lost and so would be the maintenance of the software product. But design *need not be* an AI concept, much the same way as the intermediate language used in the internals of Google Translate or DeepL bears very little relation, if any, to what we humans call "language". This suggests that the wisdom of investing in AI research to support the design stage of the software engineering process very much depends on what value the "design" product item has, and to whom. Pragmatic reasoning may contend that, when AI has taken over software engineering, all that matters is that the software product can be traceable to the requirements, without necessarily the intermediate step of human-centric software design. In some sense, this line of argument is akin to the logic that guides current research on "explainable AI", which has a stronger flavour of backward traceability than of forward design [16]. Second, the goodness of fit of a software product has much to do with the user experience, which includes but is not limited to the user interface. Not much work been done to date on applying AI techniques to the design of user interfaces or, more broadly, human-computer interaction. At the same time, there is growing understanding that the design of user interfaces requires much deeper knowledge of human perception than has been contemplated so far. Perhaps, and in fact, very likely, the user interface conduit will progressively mutate into a radically different "natural interface" that can use voice or holograms or even brainwaves instead of the windows, icons, mouse and pointer of the past.

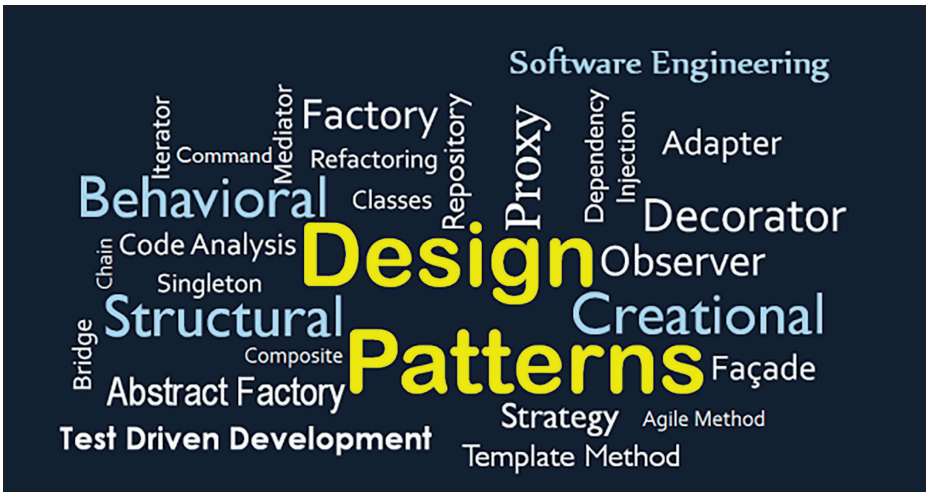


Figure 4: artist's view of the growing variety of software design patterns. (Source: https://medium.com/@madasamy/introduction-to-object-oriented-design-patterns-part-i-4e5c7845015b

Implementation

The so-called implementation stage – coding – commits the solution concept to program code, text written in one or more programming languages. In effect, coding has two sides. The one just evoked runs forward. The other – far more frequent in professional development – runs backward, when maintenance or refactoring works are required on the program text. Much like for software design, the forward side of coding is transmogrifying from the clean-sheet creation of new artefacts to the gluing together of pre-existing ones. Multiple factors instigate that transformation: the dependence on legacy; the convenience of predefined libraries; the urge of productivity, all of which call for building on existing, established parts, frequently regardless of their intrinsic quality. Interestingly, the modern, integration-geared, form of coding is becoming fitter for AI than for humans, owing to the intrinsic complexity of navigating the ocean of APIs that products need to build on [17].

Milder forms of machine-assisted coding exist, some of which predate AI and use ontology-based, domain-specific, databases to infer code fragments that best meet given selection criteria. In fact, one of the promises of multi-flavoured model-driven development is the automated generation of ever larger fractions of program code, boilerplate, business- or domain-specific [18]. Such promise still stands, and holds

is attractiveness. Not surprisingly, current research in that ambit contemplates AI augmentation to increase the capabilities and reach of automated code generation. Taken all together, these observations suggest that the need for human coding is rapidly declining and likely destined to vanish.

The darker side of coding is *refactoring* whether before product release or as part of post-release maintenance. Refactoring is the process of modifying the structural organization of existing program code without changing functional behaviour, to improve performance or better conform with quality standards. The biggest drag on such work is when the program logic (its rationale) does not transpire sufficiently well from the program text. This hindrance makes refactoring one of the most expensive parts of software implementation. The root causes of such a problem are more easily attacked by removing arbitrariness from human programming: one route to that end is to lean increasingly on machine-assisted code generation, including to modernize legacy code [19]; the other is to use tool assistance, for example code-smell [20] detectors, to help the human programmer to deliver higher-quality program code.

Verification

As E.W. Dijkstra would have it, the endeavour of making a thing (a software program) that satisfies stated needs should be split into two cohesive tasks: (i) stating the properties of a thing, by virtue of which, it would satisfy our needs, and (ii) making a thing that is guaranteed to have the stated properties [21]. Verification is the software-engineering process that helps achieve such guarantees.

Verification takes two forms: one is static, that is, it is carried out without requiring execution of the object of verification; the other is dynamic, and it is carried out by what we know as testing. Attention to static analysis is growing: its true potential is still not fully revealed as it requires cohesive cooperation between all tools used for specification, design and implementation. It is therefore premature for AI to usefully address that field.

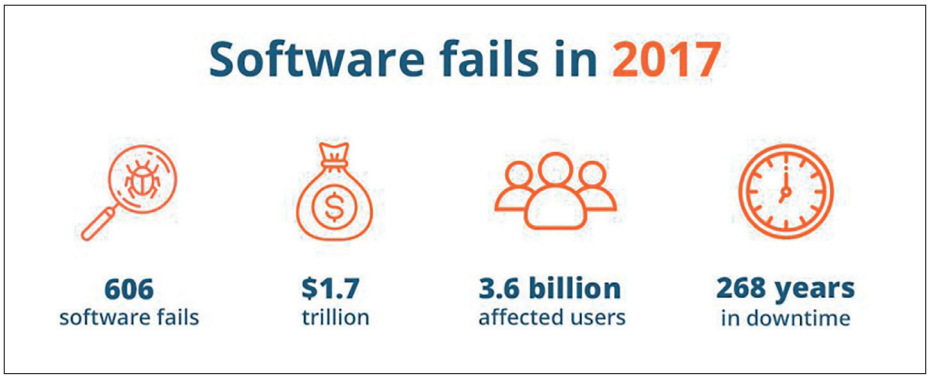


Figure 5: some figures that tell how much wanted is thorough software testing. (Source: https://medium.com/@elena.gumenyuk/you-dont-need-software-testing-do-you-9c0f0809da95 Data for the year 2018 can be found at https://spectrum.ieee.org/riskfactor/computing/it/it-failures-2018-all-the-old-familiar-faces)

Testing has been around for much longer, instead. It aims to verify that the system was built the right way (following the wanted quality principles) and that the right system was built (meeting all due requirements). Releasing software products without thorough testing is utterly irresponsible and incurs consequences ranging from embarrassment to serious damage. Yet, achieving high-coverage testing for modern systems has become largely impractical.

The prime hindrance to the desired exhaustiveness is the staggering complexity of modern systems, both vertical (from the stacking of abstraction-building layers) and horizontal (from the increasing specialization of functional algorithms). The biggest obstacle to testing beyond the unit level is the exponential complexity of part interaction, which gives rise to phenomena that may superficially appear as non-deterministic.

Testing is more than finding software faults (the proverbial bugs), it is also about assuring quality, reliability, performance, robustness, and increasingly security. In spite of a growing number of automation tools that support the execution of software testing, achieving high coverage in all those verification respects – which is expected for business- or mission-critical software – remains a very difficult objective. Conceiving tests is as effort intensive as software design. "Planting" test automation hooks in the program source code requires additional effort to normal programming. Running tests can only be done when the program code exists: when this is late in the project schedule (which is frequently the case), resources begin to be scarce and therefore the relative cost of software testing becomes dominant and frequently no longer affordable. In addition to adopting practices that anticipate software testing (for example, as in Test-Driven Development), all ways of machine-intelligence assistance to software verification are much desired. AI can help in various ways in this regard. It may help translate test specification from natural language to coded test cases. It may deploy genetic algorithms and swarm optimisation techniques in white-box testing and black-box testing. It may guide regression testing by helping single out the strictly minimal set of test cases that match the extent of the changes. It may help trial user interfaces and stress-test distributed web-based applications. It may automate the analysis of test results across development cycles to predict the likelihood of the future reappearance of already seen failures.

Maintenance

AI-tools used in the maintenance phase focus on aspects that range from extracting information from user product reviews to the classification of bug reports. Knowledge-based systems are applied to establish the maintainability of software, and to plan software maintenance. An important class of AI tools in maintenance are formed by assistants that help the software engineer find its way in the source code base of an application, guided by a particular problem that has to be solved. AI-based software refactoring tools also play an important role in this phase.

Summing it up

The source cited in reference [7] reports that 28% of the works on the application of AI in software engineering address requirements engineering, 12% design, 9% implementation, 42% testing, and 9% maintenance. This review evidence suggests that requirements engineering and testing are the most popular areas for the application of AI techniques, whereas implementation and maintenance still lag behind. These proportions may reflect the fact that requirements engineering and testing project a routine yet effort-intensive nature, while implementation is associated – increasingly inaccurately as we have noted earlier – with a far more a creative endeavour. Overall, the conception of AI-based assistance tools for software development generally tends to take over as much routine work from the software engineering tasks as possible. As in all cases of automation, the premise to that movement is to make more time available for human creativity.

Parts of software engineering can be regarded as outside-the-box creation, other parts as science. The scientific aspects of software engineering yield rules and counsels with various degrees of rigour, do's and don'ts, best practices, things that can be formalized. The staggering size and complexity of software systems, and the rising proportion of extra-functional requirements cause the body of such rules and counsels to grow accordingly. This in turn makes their diligent application more complex and daunting and labour-intensive. AI applications can and do bring much needed relief in this situation.

The more "artistic" side of software engineering, where thinking outside-the-box and creativity is required, is a challenging application area for AI-based software engineering tools. ML applications will certainly be able to find unforeseen relations in programming techniques to come up with new programming solutions, but coding outside-the-box is still a bridge too far.

AI-assisted software engineering and the HiPEAC community

The HiPEAC community is not involved in AI research; its focus is on software development. AI-based software engineering should be viewed as an application of AI, and is therefore of great interest to the HiPEAC community: it will open the way to the development of larger, more complex software-intensive systems of higher quality. HiPEAC, or more generally, Europe, must focus on this important application area of AI, leveraging the results of state-of-the-art AI research.

References

[1] David Schatsky and Sourabh Bumb, "AI is helping to make better software", https://www2.deloitte.com/us/en/insights/focus/signals-for-strategists/ai-assisted-software-development.html

[2] Louis Dorard, "Architecture of a real-world Machine Learning system", https://medium.com/louis-dorard/architecture-of-a-real-world-machine-learning-system-795254bec646

[3] Jeremy Howard, "I violated a code of conduct", https://www.fast.ai/

[4] Andreas Brennecke, Reinhard Keil-Slawik, "History of Software Engineering", https://www.dagstuhl.de/Reports/96/9635.pdf

[5] Greg Brockman, Mira Murati, Peter Welinder and Open AI, "OpenAI API", https://openai.com/blog/openai-api/

[6] Feras A.Batarseh, Rasika Mohod, Abhinav Kumar and Justin Bui, "The application of artificial intelligence in software engineering: a review challenging conventional wisdom", https://doi.org/10.1016/B978-0-12-818366-3.00010-1

[7] Marco Barenkamp, Jonas Rebstadt and Oliver Thomas, "Applications of AI in classical software engineering", https://aiperspectives.springeropen.com/articles/10.1186/s42467-020-00005-4

[8] Tim Berners-Lee, James Hendler and Ora Lassila, "The Semantic Web", https://web.archive.org/web/20130424071228/http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb.html

[9] Uwe Aÿmann, Steffen Zschaler and Gerd Wagner, "Ontologies, Metamodels, and the Model-Driven Paradigm", https://oxygen.informatik.tu-cottbus.de/IT/Research/AssmannZW06.pdf

[10] Malek Zakarya Alksasbeh, Bassam Alqaralleh, Tahseen A. Alramadin and Khalid Alemerien, "An Automated Use Case Diagrams Generator From Natural Language Requirements", https://www.researchgate.net/publication/315924400_An_Automated_Use_Case_Diagrams_Generator_From_Natural_Language_Requirements

[11] Ammar HH, Abdelmoez W, Hamdi MS. Software engineering using artificial intelligence techniques: current state and open problems. In: Proceedings of the First Taibah University International Conference on Computing and Information Technology (ICCIT 2012), Al-Madinah Al-Munawwarah, Saudi Arabia; 2012. p. 52.

[12] Ferrucci F, Harman M, Sarro F. Search-based software project management. In: Software Project Management in a Changing World. Berlin, Heidelberg: Springer; 2014. p. 373–99.

[13] Erich Gamma, Richard Helm, Ralph Johnson and John Vlissides, "Design patterns: elements of reusable object-oriented software", https://dl.acm.org/doi/book/10.5555/186897

[14] Airbrake, "Domain-Driven Design: What is it and how do you use it?", https://airbrake.io/blog/software-design/domain-driven-design

[15] Jochen Küster, "Model-Driven Software Engineering and Foundations of Model-Driven Software Engineering", https://researcher.watson.ibm.com/researcher/files/zurich-jku/mdse-01.pdf

[16] Giulia Vilone, Luca Longo, "Explainable Artificial Intelligence: a Systematic Review", https://arxiv.org/abs/2006.00093

[17] Denrie Caila Perez, "This Deep-Learning AI Can Code Just Like a Programmer ", https://www.engineering.com/DesignerEdge/DesignerEdgeArticles/ArticleID/16827/This-Deep-Learning-AI-Can-Code-Just-Like-a-Programmer.aspx#:~:text=Bayou%20is%20a%20deep%2Dlearning,of%20human%20programmers%20using%20Java

[18] Rina Diane Ceballar, "Programming Without Code: The Rise of No-Code Software Development", https://spectrum.ieee.org/tech-talk/computing/software/programming-without-code-no-code-software-development

[19] Dexter Johnson, "IBM Watson's Next Challenge: Modernize Legacy Code", https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/ibm-ai-watson-modernize-legacy-code

[20] "Code Smell", https://wiki.c2.com/?CodeSmell

[21] Edsger Dijkstra, "On the role of scientific thought", https://dl.acm.org/doi/10.5555/539053.C1104639

Harm Munk is Senior Project Manager at TNO, Eindhoven, The Netherlands.

Tullio Vardanega is Associate Professor in the Department of Mathematics of the University of Padua, Italy.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: H. Munk and T. Vardanega. The artificial programmer. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 94-99, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

The complexity of IT systems is a tremendous, costly and growing issue. By steering the way we develop these systems towards appropriate tools and methodologies, we'll be able to tame this IT complexity hydra.

Taming the IT systems complexity hydra

By OLIVIER ZENDRA and KOEN DE BOSSCHERE

Although most people remain unaware of its presence in the background, the ever-increasing complexity of IT, with its multiple sources, has been an ongoing issue for quite some time. It can even be qualified as a crisis, in both hardware and software. Indeed, this complexity has reached the point where systems are no longer fully understandable by human beings, which raises the question of how we can continue being in full control of their functioning. It is of course a matter of cost for the IT industry. But a number of incidents caused by bugs or a misunderstanding of some part of an IT system have already occurred. With an IT world that is permanently connected on a worldwide scale, the risk of damage caused by the lack of control of IT systems is both real and growing, with errors and malevolent attacks the most likely culprits.

Taming the IT complexity hydra is thus more necessary than ever. Fortunately, various solutions can be proposed to tackle the various heads of the hydra (i.e. the various aspects of complexity); these are solutions based on existing methodologies, tools and resources or extensions thereof.

Key insights

- IT systems complexity is high and ever increasing.
- IT complexity is threatening the quality and control of crucial systems that can affect the lives of many businesses and people in the EU.
- Taming IT complexity is vital for quality (safety, security, performance, sustainability, trustability, resilience) and cost (time to market and maintenance), hence competitiveness of EU industry.
- There is no silver bullet against complexity, not even AI.
- Modularity is key to mastering complexity. Modularity demands components, containers, contracts, specifications, services and orchestration.
- Formal methods, models, can harness (part of) IT systems complexity.
- The EU, like the rest of the world, is in dire need of educated, highly skilled IT specialists.

Key recommendations

- The EU should support efforts to tame IT complexity, for the sake of quality (safety, security, performance, sustainability, trustability, resilience) and cost (time to market and maintenance), hence competitiveness of EU IT industry.
- The EU should promote research on methods and tools on modularity, components, containers, contracts, specifications, services and orchestration.
- The EU should promote research on formal methods and tools for modelling IT systems and their functional properties (what they do, the algorithms) as well as their non-functional properties (how they do it: time, energy, security...).
- The EU should train more highly skilled IT specialists.

Image: ID 124451653 ©Pop Nukoonrat Dreamstime.com

The complexity of IT systems, on both the hardware and the software side, keeps growing exponentially and creates an ever-bigger challenge. It is already the case that some systems can be considered as *no longer completely understandable*, hence no longer mastered, not only by their users but above all by their designers, developers and maintainers. This state of affairs cannot go on, so *it is crucial for the EU that complexity be mastered*, for users, by its IT system providers, in all its dimensions.

This article provides an overview of the many sources of complexity that make it similar to the mythological hydra, and presents solutions we deem important to cut off its ugly heads and/or tame this beast.

IT system users don't like complexity: developers must hide it

From the user point of view, *complexity* has to be hidden so as to provide an easy, pleasant user experience. IT systems have very much improved and even done well to hide the nitty gritty details for basic levels of use but the complexity for users tends to move to higher levels of use. Users increasingly want to have access to multiple functions and services, from various providers, spread all across the world, and all of this at the same time, possibly on multiple and varied terminals (from smart watches to desktops, via smartphones and tablets), presented to them in a simple and convenient way.

Multiple installed applications go against simplicity. Users need *as-a-service meta-applications*, that is to say, aggregators of multiple applications, to save them from the connection and coordination issues associated with accessing the various applications. In addition, these aggregators cannot just present users with a juxtaposed view of the various application results: they must be smart integrators that process and manage the complexity of the various results and present them in a more synthetic, and easier to understand way. Good examples of such meta-applications, which currently tend to be domain-specific, are price aggregators and comparators (for travel, hotels, etc.), and virtual personal assistant capabilities like Alexa Skills.

IT systems are full of complexity: the sources are varied

For IT system developers, *complexity springs up in all corners of IT systems development*, for both **hardware** and **software**, and its *various aspects call for different kinds of solutions*. This section presents an overview of the root sources of IT complexity that we deem important.

In **hardware**, since Dennard scaling has stopped, processor systems have become tightly-interconnected multi-cores (exposing *parallelism* with and without *concurrency*), fitting in an increasing number of accelerators (exposing *heterogeneity*), aggregated in variably deployable units (exposing *statelessness*) and networked (exposing geographical distribution and decentralization), for ease of access via the web (exposing *asynchrony*). Field-programmable gate arrays (FPGAs) are based on radically *different programming models*. Hybrid platforms are also emerging. *Heterogeneity* is thus probably more prevalent than was generally expected. As a consequence, the hardware environment is *evolving extremely rapidly*, even faster than in the era of Moore's law.

This increasing hardware complexity is now an emerging crisis. The (incomplete) documentation amounts to 9000 pages, with a table of contents of 100 pages, is written in informal English and periodically amended by errata. This, for many chips, every few weeks [3]. How can humans cope with such a *diluvial amount of information*? Is this huge engineering effort worth it? Since hardware has no formalized semantics, how can we ensure it is correct? Is verified software built on the shifting sands of possibly incorrect hardware? Bugs occur, safety or security breaches too, making attacks by various aggressors easier.

This explosion of complexity is matched in **software**, which contains many sources of complexity at all levels of the software stack.

The system development ecosystem is an immense maze consisting of scores of *methodologies and their derived tools (Figure 1)*.

Programming languages alone are also a source of complexity (Figure 2). Historically, statically typed languages used to be the most popular ones. Then dynamic languages became popular (Ruby leading). Now languages are more mixed [4]. As a result, today *more than 8000 languages exist*, from the generalist languages addressing a wide range of needs to more specialized, targeted languages, or even DSLs (domain specific languages) fully tailored to one specific domain of application. A large number of them are still in active, live use, creating a modern-era technological Tower of Babel:

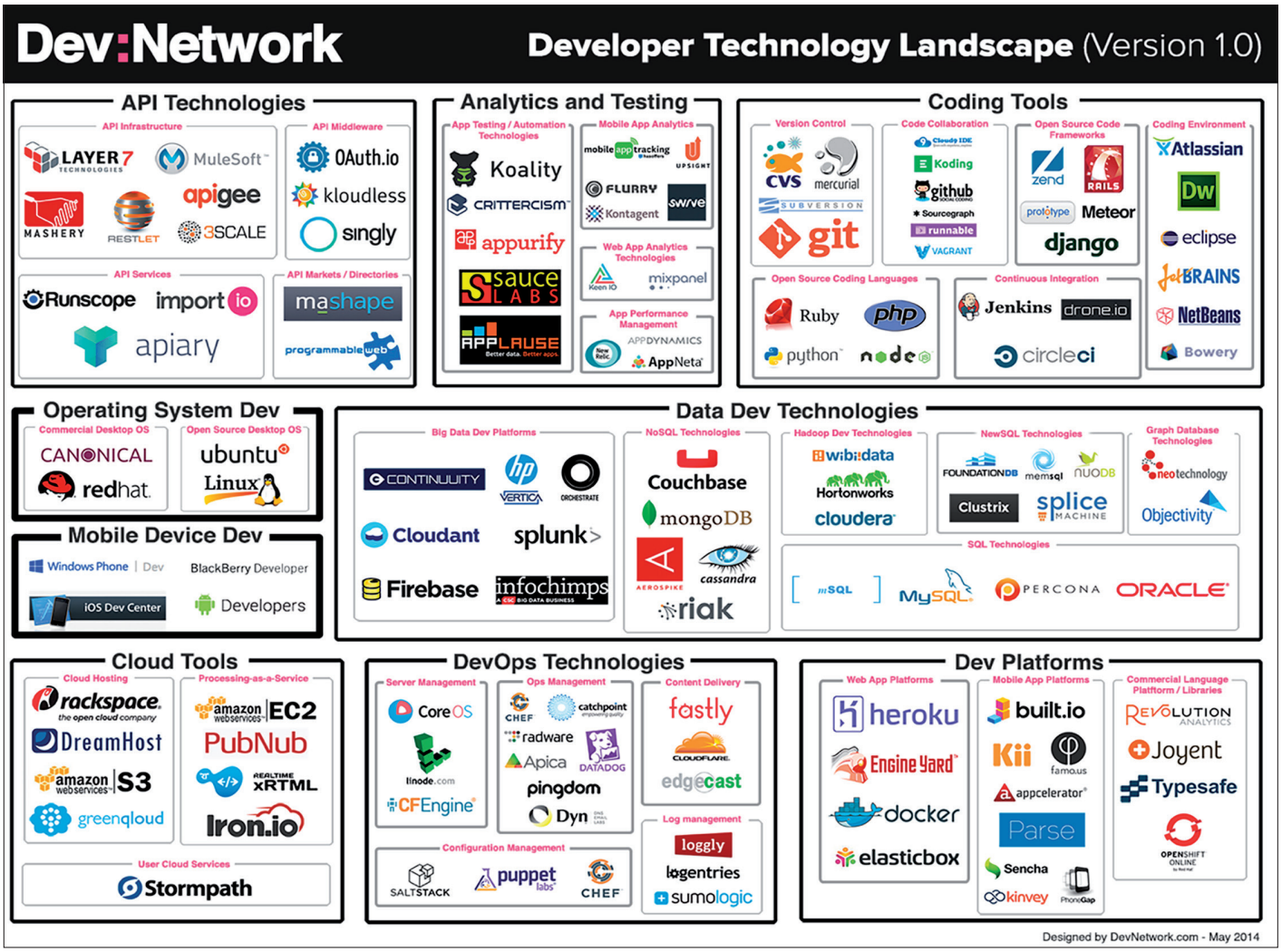


Figure 1: A sample of the developer's technology landscape in 2014. Things have not improved since. Source: DevNetworks Sought enhancements: asserting correctness.

This *multiplicity and heterogeneity of languages* also creates huge complexity, in terms of their interaction and developers' ability to master them, etc. What are the tools available to manage this complexity when using different languages in the same codebase (e.g. how do you diff)?

At the same time, the various languages exist for a reason, while addressing different needs. A good illustration here are DSLs that help tackle the peculiarities of some application domains in better ways than general purpose languages, hence in a simpler way.

Software libraries provide off-the-shelf capabilities, or features, that developers like to reuse to build systems from existing parts rather than reinventing the wheel. Yet the library ecosystem is huge. Which library to choose when wanting to add a feature is a rather informal, ad-hoc process, and may be a complex one when faced with many alternatives. Library versions also have to be taken into account, because compatibility issues between versions of the various libraries create an intricate web of dependencies.

Application code itself can be quite complex, just because of the inherent complexity of the problems it solves. Admittedly, this kind of complexity seems more useful than certain other kinds, yet it has to be managed. Application code complexity can also come from the coding style and/or the language used. Terseness can sometimes lead to obfuscation, while some verbosity may help understanding, hence lower complexity (see Figure 3). A balance thus has to be found.

In addition, with the growing pervasiveness of the use of computer systems in virtually every aspect of our daily life, the production side of the IT community is faced with complexity factors that add to the classical functional complexity. These factors comprise *non-functional properties* such as energy, time and other resource constraints, ever-advanced human-computer interaction, the weaving of cyberspace into physical reality, as well as continuous delivery within continuous operation. All this results in many

Language Ranking: IEEE Spectrum

Rank	Language	Type	Score
1	Python		100.0
2	Java		95.3
3	C		94.6
4	C++		87.0
5	JavaScript		79.5
6	R		78.6
7	Arduino		73.2
8	Go		73.1
9	Swift		70.5
10	Matlab		68.4
11	Ruby		66.8
12	Dart		65.6
13	SQL		64.6
14	PHP		63.8
15	Assembly		63.7
16	Scala		63.5
17	HTML		61.4
18	Kotlin		57.8
19	Julia		56.0
20	Rust		55.6
21	Shell		52.0
22	Processing		49.2
23	C#		48.1
24	SAS		45.2
26	Cuda		41.0
27	Visual Basic		40.3
28	Objective-C		38.9
29	Delphi		38.6
30	Perl		38.2
31	Verilog		37.6
32	VHDL		36.7
33	LabView		36.7
34	Elixir		35.8
35	F#		34.7
36	Prolog		34.6
37	Lua		34.4
38	Lisp		33.0
39	Ada		32.8
40	Apache Groovy		32.0
41	Scheme		31.4
42	Haskell		30.8
43	Cobol		30.4
44	Clojure		29.8
45	ABAP		29.5
46	D		27.7
47	Forth		23.7
48	Ocaml		23.7
49	TCL		22.1
50	LadderLogic		19.5

Figure 2: IEEE Spectrum Top 50 programming languages 2020 [8]. Python's success may lie in its ease of use to "glue" together libraries, where most of the computation are done, and its large number of supporting libraries, covering many domains.

```
define F (getchar()&15)
#define v main(0,0,0,0,
#define Z while(
#define P return y=~y,
#define _ ;if(
char*l="dbcefcbddabcddcba~WAB+ +BAW~          +-48HLSU?A6J57IKJT576,";B,y,
b,l[149];main(w,c,h,e,S,s){int t,o,L,E,d,O=*l,N=-1e9,p,*m=l,q,r,x=10 _*l){y=~y;
Z--O>20){o=l[p=O]_ q=o^y,q>0){q+=(q<2)*y,t=q["51#/+++"],E=q["95+3/33"];do{r=l[p
+=t[l]-64]_!w|p==w&&q>1|t+2<E|!r){d=abs(O-p)_!r&(q>1|d%x<1)|(r^y)<-1){_(r^y)<-6
)P 1e5-443*h;O[l]=0,p[l]=q<2&(89<p|30>p)?5^y:o;L=(q>1?6-q?l[p/x-1]-l[O/x-1]-q+2
:0:(p[l]-o?846:d/8))+l[r+15]*9-288+l[p%x]-h-l[O%x];L-=s>h||s==h&L>49&1<s?main(s
>h?0:p,L,h+1,e,N,s):0 _!(B-O|h|p-b|S|L<-1e4))return 0;O[l]=o,p[l]=r _ S|h&&(L>N
||!h&L==N&&1&rand()){N=L _!h&&s)B=O,b=p _ h&&c-L<S)P N;}}}t+=q<2&t+3>E&((y?O<
80:39<O)||r);}Z!r&q>2&q<6||(p=O,++t<E));}}P N+1e9?N:0;}Z l[B]=-(21>B|98<B|2>(B+
1)%x),++B<120);Z++m<9+l)30[m]=1,90[m]=~(20[m]=*l++&7),80[m]=-2;Z p=19){Z++p<O)
putchar(p%x-9?"KQRBNP .pnbrqk"[7+p[l]]:x)_ x-(B=F)){B+=O-F*x;b=F;b+=O-F*x;Z x-F
);}else v 1,3+w);v 0,1);}}
```

Figure 3: Complexity is not only depending of the size of the code, this example shows a complete program. How it works was explained in a book of 170 pages. Can you guess what it is doing? [9]

systems being composed of complex webs of dependencies that are easy to break and hard to maintain. Furthermore, there are still issues that have not been completely solved by the IT community. Among them, *how do we measure and value software quality?* Which non-functional properties or metrics have to be considered? *How do we value non-functional properties like speed, low-energy, high-security?*

Another of the crucial and complex aspects in the IT ecosystem is the importance and variety of legacy. Legacy is the heritage of the past, composed of *existing operating systems, libraries, languages, development tools and hardware*. Legacy represents a huge amount of code, estimated in 2000 at over 100 billion lines of code, most of it COBOL [6,7]. Legacy hinders the taking of new directions, yet it cannot just be done away with. Indeed, the service provided by large existing legacy systems must still be provided, so disrupting them is not an option. New languages and libraries must be able to interoperate with legacy ones. New software must often run on old hardware or old OSes, which multiplies the possibilities and tests to be done, the compatibility patches to write, for no other usage than having an IT system work in yet another particular context.

Furthermore, in order to reduce cost and time to market, it is much better and very common to reuse existing elements – even code parts found in public repositories on the web – to extend or modify existing systems, than to start each development from scratch. Indeed, legacy code, despite all its drawbacks, still makes it possible to tackle problems without reinventing the wheel, by reusing old, tested and tried, libraries that have been very fine-tuned and well debugged over the years, thus removing a lot of the complexity of new developments.

The issue there is thus not so much the existence of legacy as its intrinsic *quality* and the complexity to *integrate* it into new developments. Tradeoffs are thus of the essence when reusing legacy code.

Figure 4: Caeretan hydria, c.525 BC, Hercules slaying the Lernean hydra, Collection of the J. Paul Getty Museum, Malibu, California. Image: Wolfgang Sauber: Getty Villa, CC BY-SA 3.0, creativecommons.org/licenses/by-sa/3.0, via Wikimedia Commons

Overall, all these sources of complexity add up, as do the size of the elements composing the system: *the bigger the system in terms of functionalities, the greater its complexity*. Similarly, outside of the purely technical complexity, the greater a system, the greater the team needed to develop it, the greater the complexity of the development process and its management.

Impacts of complexity

The consequences of complexity in IT systems, coming from the above-mentioned sources, are very *simple: high levels of complexity mean high costs, and high risks.* High complexity brings high development costs, because of the size of the teams needed to develop the systems, and of the time needed to do so. It also incurs high risk of delays in the process, risk of poor quality in the system (risk of bugs leading to malfunctions), lack of speed, lack of safety and lack of security. The same applies of course for ongoing maintenance of complex existing IT systems, and for their evolution, with added difficulty that the original knowledge of the system designers and implementers is often gone.

In a nutshell, complexity must be tamed, in order to keep the IT system under control.

Fortunately, various solutions exist, or are within reach with reasonable efforts, to tackle the various heads of the complexity hydra.

Modularity does manage complexity, with additional benefits

Modularity at various levels is the main key found by humans to mastering complexity and to the reuse and integration of hardware or software legacy, especially across programming languages. The elements to (re)use have to be properly architected so as to be taken as whole *modules*, properly *contained* and *encapsulated*. On the software side this implies isolating the implementation inside the proper module (class, object, library, container…) and exposing only the right amount of interfaces at the boundaries of the module, to provide (micro-)services. Clear *contracts* must thus explicitly define the behaviour of the *interfaces* exposed at the boundaries, both inward and outward. Enhancing module/container interface *specifications*, so that they help assess semantic conformance at build, integra-

tion, deployment and execution time is necessary to properly achieve these goals. It should take the form of enforceable contracts covering both *functional* (i.e. the algorithms) and *non-functional* (e.g. time, power and energy, security and safety, etc.) properties.

This kind of modularity would be especially apt to current IT systems, which are generally extremely connected and distributed over the web. It matches very well with the *microservice* paradigm, an enabler of modern, heterogeneous software composition. Indeed, an individual microservice is a small self-contained application that has a single responsibility (which gives it a clear and distinct role in a composition), a fully-self-contained and preferably lightweight stack (which allows its software dependencies to be always fully satisfied), and which can be deployed, scaled and tested independently (which facilitates software evolution) [1]. The "microservices" architectural style yields a single application from the coordination of a suite of unitary services [2], each of which exposes an application programming interface (API) *outside* of their codebase, which is invoked using *asynchronous* (crucial to loose coupling) *web-based* service requests (key to reachability). Microservices can run isolated from others in containers, using hypervisors to segregate them. This provides more resilience in case of hacking, since contaminations should be blocked between containers.

Software applications and infrastructures will increasingly be aggregates of heterogeneous artefacts with a variety of deployment requirements. Controlling them can hardly be done in a merely declarative way or scattered in a maze of uncorrelated and independent scripts. Languages and tools for *orchestrating* collaborative distributed and decentralized components are thus needed.

In addition to helping integrate different (possibly legacy) elements, modularity is key to boosting the repairability of IT systems. Being able to replace a part (be it a software or a hardware one) with another when it is found to be faulty, or when it becomes obsolete, is a powerful way to extend the lifespan of an IT system. Although this seems obvious, when thinking of e.g. automobile parts, this is a concept less developed for hardware in IT systems. Software parts are more often upgraded, with many OSes, libraries and applications having new versions released with patches and/or improvements, thanks to these updates being mostly automated, hence very easy, on the user side.

By repairing hardware parts or modules, IT system lifespan can be increased, thus decreasing their global ecological footprint both in terms of raw resources and carbon impact.

By patching/upgrading software, IT system quality can continuously be improved, thus avoiding the costs and inconvenience of faulty behaviour, especially with respect to safety and/or security.

Modularity is also key to more easily developing new IT systems, allowing reuse of hardware or software modules and making it possible to create whole new product lines with limited effort. A well-known example is printer product lines, which clearly rely on modularity and componentization to produce a large variety of similar but not identical products to tackle a variety of consumer needs, thanks to a limited set of common subparts. There, modularity clearly decreases financial expenditure and time to market, which provides several competitive advantages.

Like for reusability of (legacy) elements, modularity for repairability requires clear interface contracts, specifications, at module boundaries, since the mechanisms are the same. Again, these contracts and specifications must take into account the non-functional properties as well, so as to carry enough information to ensure the proper composability of the modules, especially in the long term, with various evolutions of the system, hence evolutions of the other surrounding modules. These contracts must also be easy for developers to master and to deal with, especially when taking into account the shortage of skilled IT professionals in the EU. At the same time, these contracts must be amenable to formal verification and/or proofs.

Carrying enough information, while at the same time providing a good *level of abstraction* to hide away the details and not go in the way of composition, is an issue that must be tackled. It implies being able to have different levels of abstraction in the models and the tools, so as to be able to zoom in or out, depending on the level of details needed at the level of composition considered. These levels allow different views, with more or less information being provided, while keeping the underlying information complete, with no loss.

Abstracting away complexity with formalization, models and tools

In order to provide these *different levels of abstraction*, and the expression of contracts at module boundaries, appropriate models of the systems have to be relied upon to cope with complexity.

There is a dire need for *formalized semantics* to facilitate better analysis of the system and its properties, and the derivation of formal proofs for (at least some of) them. It is currently, however, near impossible, or at least prohibitively expensive, to mathematically formalize and completely prove large IT systems: they are too complex. However, we do know how to reason about functional correctness of programs and some smaller parts, modules, can be formalized and proven. This partial *verification* is currently the norm, to provide levels of assurance, mastering part of the complexity.

A lot of hardware has no formalized semantics. The hardware models for software development are thus inaccurately specified. So, it is difficult to formally ensure correctness: verified software would even in a way be built on (shifting) sands… Fortunately, this is changing. ISAs (instruction set architectures) are being (more) formally specified [3], and include ARM [10,11], RISC-V [12,13].

Software systems too will increasingly rely on formal methods. This is already happening; for example, the Isabelle proof assistant [14] is commonly used in the writing of seL4 OS [15], while Coq [21] is for the formal verification of the Compcert compiler [22]. Executable specifica-

Credit: ID 92496894 ©Nattapol Thepyasuwan | Dreamstime.com

tions put in the code (i.e. contracts) should also be increased, proving both a means to document the intent of the code, its specification, and to help its verification by automated tools. In addition, these specifications should provide information not only about the program functional aspects (what it does, the algorithms and functions), but also about what is currently called its *non-functional properties* (how it does it), like time and reactivity, power and energy, safety, security, etc. There lies a real current challenge: programmers and support tools should be able to express, manipulate, and reason about these non-functional properties, to yield static proofs of functional as well as non-functional correctness, to make runtime decisions, to support runtime assertions to check that the necessary properties hold during execution, and that they have adequate semantics to handle violations so that safety conditions are restored.

Efforts along these lines already exist and must be supported. It is necessary at the same time to also pursue less formal but more practical efforts aiming to improve the quality of the developed system elements and modules, in a very concrete and practical way, helping developers cope with some parts of the complexity. The MDE (model driven engineering) methodology [16], including the well-known UML (unified modeling language) [17] adopted by the OMG (object management group) computer industry standards consortium [18], and the related modeling tools (e.g. the Eclipse Papyrus Modeling Environment [19]), have been for a long time making progress in that direction and should be supported.

However, the advance of formal methods in IT systems has been hindered by past and current market realities. Indeed, business constraints (time to market, cost of production) and the programmers' mindset have generally focused on delivering functionalities to customers, since this is what sells. Integral correctness is rarely pursued by design; more often it is sought as a product of quality assurance activities, either performed retrospectively or in parallel to development, but not sufficiently ingrained in it. While some enterprises do specialize in providing tools that help the quest for correctness, their success has never even remotely approached that of organizations providing functionalities to the end user, such as the likes of Facebook or Twitter. Still, the potentially negative impact of this situation is huge, for loss of value, increase of risk, and spread of threats, and should be acted upon with a more vigorous quest for quality. The fact that some very famous IT companies provide end-user licence agreements that, in essence, remove any responsibility on their part should the product not work, means that the cost of such failures falls to the customer rather than to the provider, which is a very uncommon practice in other business domains. *Regulations against this could strongly help the quest for quality, by putting a higher price on the damage caused by poor quality IT systems.* Mandating liability for IT systems should thus be a priority for the EU to boost the quality of its IT systems.

Coping with complexity of formal methods may be an issue for developers/designers, but it will be a simpler one to solve than directly managing the full complexity of hardware and software. With tools getting easier to use, good programmers should have no problem mastering formal tools.

To tame the hydra, you need tamers: the role of IT education

Mastering all this complexity indeed requires *many highly skilled IT specialists.*

Educated designers use proper design methods and tools, producing high qual-

ity architecture and modular systems. Educated programmers program well and produce good implementations, even with poor languages. It is a fallacy to believe that to build significant IT systems, uneducated programmers can simply use tools in a kind of copy-paste way, not fully understanding what they are doing and what are the fundamental underlying concepts, and yet still produce good quality systems. Learning and mastering the fundamental concepts is key to good decision making in IT system production.

"Controlling complexity is the essence of computer programming." [5]
Brian Kernighan

At the same time, it is necessary to have tools that can present the proper level of abstraction, hiding the details when they are not needed. Easy programmability is thus a must-have goal. Tools must improve. Graphical programming was an interesting track to ease programming and bring it to the masses to some extent, with visual programming languages [20]. However, so far, graphical programming is still not scalable, strongly limiting its usefulness in commercial application building. The UML modelling language, with its graphical representation, goes to some extent in the graphical programming direction (but certainly not for the masses), and has made it possible for professional IT systems developers and designers to better represent the systems, hence to master complexity in a better, yet still incomplete and imperfect, way.

Machine learning can also be seen as an interesting track to help in IT system production. It has the capacity to learn heuristics, hard-coded control loops, policies, and help implement them in an automated or semi-automated way, thus saving significant amounts of time. It could also help with some architectural choices, based on the specifications. However, we should not think machine learning will write all of our programs for us anytime soon. All the hard, system-wide problems will remain: security, correctness, reliability, availability. In addition, although AI or its currently fashionable incarnation, deep learning, could help take away some of the complexity of programming IT systems (chopping off one head of the complexity hydra) by for example writing automatically some parts of programs (e.g. [23]), the use of machine learning could make it much more difficult to analyze the correctness, hence the safety and security, of IT systems, thus increasing the complexity of these stages (thus growing news heads for the complexity hydra)... So all these problems will have to continue to be addressed mainly by skilled human IT specialists.

Unfortunately, the latter are in scarce supply in the EU, with numbers being insufficient to fulfil the needs of our economy; this greatly hinders EU innovation and competitiveness. The need for a sizeable community of educated professionals capable of developing IT systems and who understand the fundamental concepts that underpin IT systems, must be addressed.

Conclusion

The complexity of IT systems is a monstrous hydra that cannot be left unattended. To tame it, and therefore ensure the good quality of the EU's IT systems, as well as the competitive advantage of its IT systems providers, there is no silver bullet. On the contrary: the answer is multifaceted, as much as complexity is multifaceted. Concretely, the EU must steer the way we design, develop and maintain these systems in more modular ways, using the practical power of containers, encapsulation, contracts, microservices and orchestration, as well as the formal power of verification, proofs, and correctness checking methodologies and tools, and explore how new technologies such as artificial intelligence can help. The EU must also have a large workforce of people skilled in IT systems, to tackle the challenges of today and those of tomorrow.

References

[1] J. Thönes, Microservices, IEEE Software, 32(1):116, Jan 2015.
[2] Martin Fowler, Microservices, https://martinfowler.com/articles/microservices.html, Mar 2014
[3] Timothy Roscoe, HiPEAC Vision Consultation meeting, 7 April 2020.
[4] Tiobe language index. https://www.tiobe.com/tiobe-index/
[5] Brian W. Kernighan, P. J. Plauger. Software Tools. Addison-Wesley Publishing Company, 1976
[6] F. P. Goyla, Legacy integration-changing perspectives [Cobol], in IEEE Software, vol. 17, no. 2, pp. 37-41, March-April 2000, doi: 10.1109/52.841604.
[7] COBOL legacy, Wikipedia. https://en.wikipedia.org/wiki/COBOL#Legacy
[8] "The Top Programming Languages", https://spectrum.ieee.org/static/interactive-the-top-programming-languages-2020
[9] Toledo Nanochess C program: https://nanochess.org/chess3.html
[10] ARM Architecture: https://en.wikipedia.org/wiki/ARM_architecture
[11] ARM ISA family: https://developer.arm.com/architectures/instruction-sets
[12] RISC-V: https://en.wikipedia.org/wiki/RISC-V
[13] RISC-V specifications: https://riscv.org/technical/specifications/
[14] Isabelle proof assistant: http://isabelle.in.tum.de/
[15] Proofs in seL4: https://sel4.systems/Info/FAQ/proof.pml and https://docs.sel4.systems/projects/l4v/
[16] Model-Driven Engineering: https://en.wikipedia.org/wiki/Model-driven_engineering
[17] Unified Modeling Language: https://en.wikipedia.org/wiki/Unified_Modeling_Language
[18] Object Management Group: https://en.wikipedia.org/wiki/Object_Management_Group and https://www.omg.org/
[19] Eclipse Papyrus Modeling Environment: https://www.eclipse.org/papyrus/
[20] Visual programming languages: https://en.wikipedia.org/wiki/Visual_programming_language
[21] Coq proof assistant: https://en.wikipedia.org/wiki/Coq
[22] CompCert compiler: http://compcert.inria.fr
[23] Alexey Svyatkovskiy, Shao Kun Deng, Shengyu Fu and Neel Sundaresan, IntelliCode Compose: Code Generation Using Transformer. Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2020

Olivier Zendra is a Tenured Computer Science Researcher at Inria, Rennes, France.

Koen De Bosschere is Professor in the Electronics department of Ghent University, Ghent, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: O. Zendra and K. De Bosschere. Taming the IT systems complexity hydra. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 100-107, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

Although reaching the limits of physics, silicon technology continues to improve, therefore still allowing the design of improved and better performing circuits and systems.

Silicon technology is still in the game

By CARLO REITA, SANDRINE CHERAMY and MARC DURANTON

Silicon technology, which has fuelled the improvement of ICT system performance for the last sixty years* or more, is reaching the limits of physics, with device size in the range of a few tens of atoms. However, even if the scaling itself reaches its limit (3nm is announced for 2022), various approaches (new geometry, new materials, 3D stacking, …) allow for further improvement of the density and efficiency of transistors, allowing chips that are more efficient and complex to be designed.

Key insights

- Silicon technology is still finding new approaches to increase its performance.
- New transistor structures are emerging, such as Gate All Around FET (GaaFET).
- 3D structures reduce the length of wires (and therefore power dissipation) and increase the density of systems when compared to traditional packaging technologies.
- The use of interposer and chiplet technology will allow a reduction in design costs (thanks to the possible re-use of IP and to smaller computing devices) and the mixing of different technologies, such as analog, power converters and digital.

Key recommendations

- Continue investigating new approaches to further increase performance of silicon-based systems.
- Europe should keep hold of its knowhow in this domain, otherwise it will risk losing the ability to make optimal use of those new technologies.
- The use of interposer and chiplet technology might become the sweet spot for edge devices, having the best performance/cost ratio.

The semiconductor technology CMOS, which fuels the digital era, has reached a period of diminishing return, and will require gigantic investment to allow further improvement to its performance. At the same time, no other technology is on the horizon as a possible replacement, at least for the foreseeable future. However, performance is still improving, albeit more slowly, with new designs, in a "3nm" technology, scheduled for the latter half of 2022. This article will explain what the reality of scaling is and the various solutions that are used to continue improvements in performance, at least for the next five years.

3nm available in volume production in 2022

"In August 2020, TSMC announced details of its N3 3nm process, which is new rather than being an improvement over its N5 5nm process [7]. Compared with the N5 process, the N3 process should offer a 10–15% (1.10–1.15×) increase in performance, or a 25–35% (1.25–1.35×) decrease in power consumption, with a 1.7× increase in logic density (a scaling factor of 0.58), a 20% increase (0.8 scaling factor) in SRAM cell density, and a 10% increase in analog circuitry density. Since many designs include considerably more SRAM than logic, (a common ratio being 70% SRAM to 30% logic) die shrinks are expected to only be of around 26%. TSMC plans risk production in 2021 with volume production in the second half of 2022" [8].

* The first working example of an integrated circuit was demonstrated by Jack Kilby on 12 September 1958.

The reality of scaling

News of the impending end of the microelectronics scaling roadmap has reached even the general public, but as usual, the situation is more complex than what its appears to be at first glance. The attention of the press, following Intel's longstanding lead in microelectronics, has mostly focused on Moore's law. This "law" was, in reality, an observation made by G. Moore [1], when still based at Fairchild and before the CMOS devices technology was even in production. He observed that the density of the components on a chip had been doubling roughly every year; in 1975 a more complete version of the paper postulated the doubling as being every three years [2]. This statement is more a business model than a law, and the only real law derived from the physics of the MOS transistors was outlined R. Dennard in 1974. This law states the relations between number of physical characteristics when a uniform reduction of dimension is applied to a standard MOS transistor, notably the power density (that remains constant) and the speed (Figure 2). In fact, the improvements resulting from the shrinking of dimensions had already slowed down or disappeared altogether by around 2005. In his paper, R. Dennard had indicated where this type of purely geometrical scaling would start breaking down: the dopant concentration. In fact, the behaviour of electrical junctions is controlled by the dopant concentration in silicon and this cannot physically be larger than a few percent of the atomic density of Si. This limit is reached at around 20nm to 30nm of gate length. Two other limitations of pure geometrical scaling were always very clear: dissipation and patterning.

The power dissipation per unit area of the transistor is constant, so, at some point, the increase of transistor numbers per unit area of chip would make it impossible to extract the dissipated heat. Dissipation is directly proportional to frequency of operation and inversely proportional to the square of the supply voltage. As a consequence, the speed at which the transistor can be operated has to be limited (hence the stop at around 2-4GHz for the processor clocks) and not all the transistors may be operating at the same time (parts of circuits need to be switched off from time to time, also called "dark silicon"). The supply voltage has also been reduced but there is a limit at which the transistor cannot be controlled (around 0.3V).

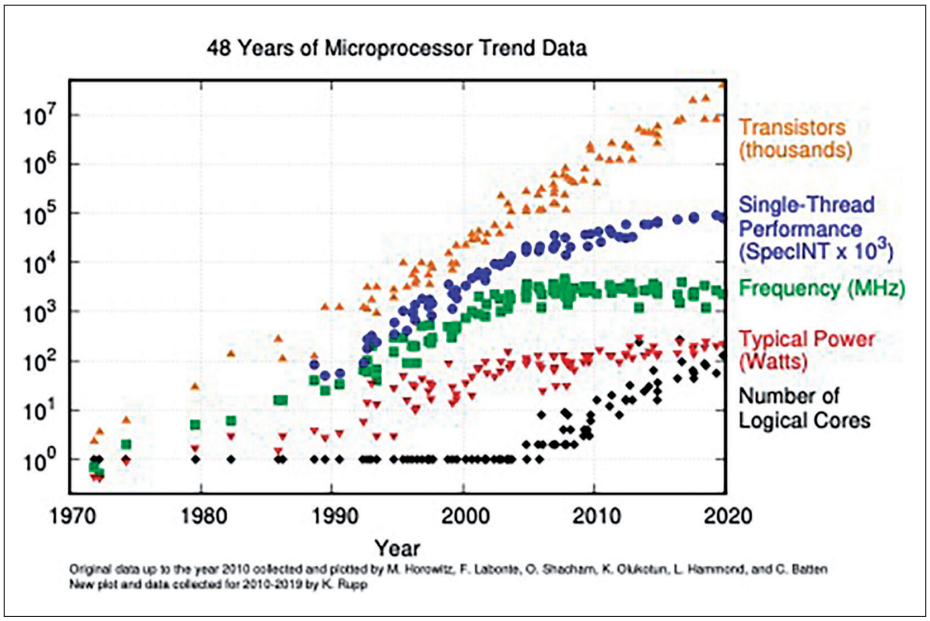


Figure 1: 48 years of microprocessor trend [6]

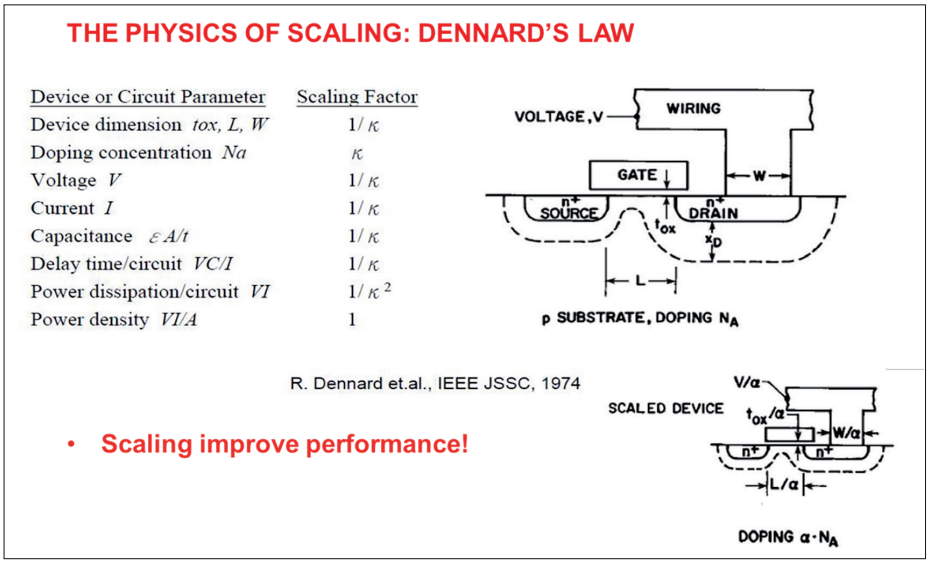


Figure 2 Dennard's scaling law

The second limitation has always been the way patterning can be done. The optical process of transferring the image onto the layer to be patterned is limited by the wavelength. Initially this limitation was reduced by the introduction of powerful lasers down to 248nm, which was sufficient until the normal scaling broke down.

The push for the reduction of size, then, has continued in order to keep reducing the unit cost of the chips but the era of pure geometrical shrinking has now been over for more than ten years. The continuation of the shrinking of dimensions has led to the need for some major changes in the structure of the devices and hence a major complexification. To control the junction after the maximum dopant density was reached, the devices needed to be fabricated on thin layers of silicon and this brought about the FinFET and the FDSOI (Fully Depleted Silicon on Insulator) technologies. The reduction of the dimension and this confinement entail a reduction of the maximum current density in the transistor channel limiting the capacity of driving the metallic interconnects and effectively the

speed. To overcome this limitation, strained layers and silicon germanium alloys have been introduced to present higher current mobilities. In order to limit the leakage from the control gate, thicker, high permittivity oxides have been introduced by first adding hafnium to silicon dioxide and/or using pure hafnium oxide.

The issue of patterning has always been central and, early on, the lithography tools have standardized on a reduction by 4X of the mask pattern projected down to the photoresist coated wafers. The way to improve resolution has relied on improving the two parameters determining resolution of an optical system: numerical aperture NA and wavelength λ. From classical optics, in fact, is well known that the minimum pitch for two lines to be imaged is directly proportional to λ and inversely proportional to the NA (diffraction limit). The wavelength, hence, has been progressively reduced from 365nm down to 193nm while NA has increased progressively to 0.95. At this point (early 2000s), the first attempt to further reduce the wavelength failed (157nm did not pass the research stage) and so, like in the microscopes of the late 1800s, to further improve the resolution a higher refraction index medium was introduced between the last lens and the wafer. These systems, using 193nm laser illumination, are known as immersion systems and have supported the scaling down to 20nm node. At the same time research started on a radical change of wavelength to move to 13.5nm (Extreme UV, EUV) requiring the imaging process to be done in a vacuum and using only mirrors in the optical system. Unfortunately, the ten years initially forecast for the development of such tools became nearly twenty years and the industry had to find other ways to keep the scaling pace. Initially, modifications of patterns on the mask were applied to correct proximity effects (Optical Proximity Correction, OPC, techniques), but soon another optic "trick" was used. By imaging only along one axis and accepting a deterioration on the perpendicular one while also adapting the illumination source, it is possible to reduce the pitch below the classical limit. This has led to radically new design constraints forcing superposition of layers with opposite directions leading to

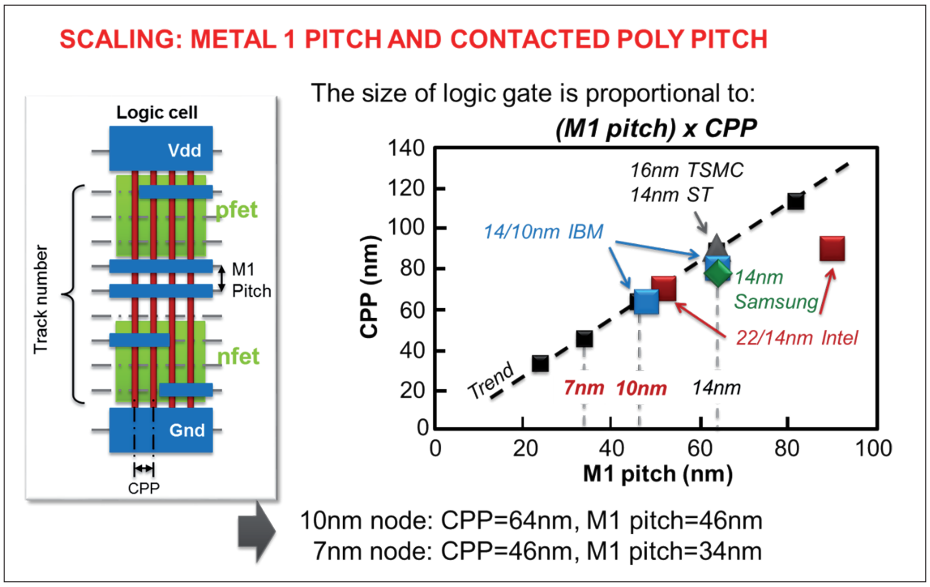


Figure 3: The design parameters of a logic gate

the fact that the density of a process node can be related to the pitch of the three critical layers (gate, contact and metal). When the limits of even this technique were reached, the patterns of each layer were separated into multiple masks with multiple intermeshing exposures needed to attain reduced pitches. These techniques go by the name of "multiple patterning" and have dramatically increased production and design costs. However, at the 7nm node, the EUV tools have finally become available, simplifying somehow the designs at the cost of other specific issues but providing once more the real possibility of imaging at the dimensions required for the 5nm and 3nm nodes.

Collectively the use of all these techniques has allowed what has been called "equivalent scaling", meaning that the density has kept increasing even if the dimension of the transistors themselves has not been reduced as in the past.

Essentially, at each generation, the gate length is not scaled by a factor 0.7 to obtain a gain by a factor of two in surface for the device as was seen in the past; the gain is now only around 20%-30%. Performance is typically improved by between 20% and 40% at each node as a result of this reduction and of multiple new technology developments like the previously mentioned new materials, use of strain, change in shape and aspect ratio of devices, etc. The density of devices, however, still tends to improve by a factor of two by reducing tolerances (thanks to better lithography and processes in general) and playing on the layout (number of fins per transistor, number of tracks per cell, number and position of contacts). The use of these techniques introduces a level of confusion and uncertainty about how to compare technologies and processes between nodes and across manufacturers and applications (e.g. very large differences in densities between logic, high performance, low power and analog libraries for a given node).

The naming convention for the "equivalent scaling" has also started to be a misleading one. In effect, until the 45-40nm the value which indicated the node (here 40nm) was typically the gate length of the smallest devices or half of the pitch of the densest interconnect metal layer. Pioneered by Intel (moving the naming from 28nm to 22nm …) with the marketing department of all players following suit, the node naming stopped reflecting any real physical dimension (e.g. at 7nm the smallest physical dimension of any parameter is ~12nm ….) and became just a commercial label. In Figure 4 are some examples of node naming by manufacturers and real dimensions in the design and devices. Where a reference is not provided, data is extrapolated from multiple sources. Real data for 5nm and 3nm exists publicly but only in the form of company announcements and so has not been included.

Nominal node		28nm	22nm	20nm	18nm	16nm	14nm	12nm	10nm	7nm	5nm	3nm	2nm (N2 imec estimation)
Intel	Lg (nm)		24				20		16	~14nm			10nm
	Fin Pitch (nm)		60 FinFET				FinFET 42		FinFET 34	FinFET			16
	CPP (nm)		90				70		54	47			42
	M1 (nm)		80				52		44	26			16
	SRAM		HD 0.092µm2				0.0588µm2		0.0312µm2	0.027µm2			
	Year Publication		VLSI 2012				IEDM 2014		IEDM 2017/ISSCC2018				DTCO conf. 2020
	Risk Prod		2011				2014		1Q18	1H2021			
Samsung	Lg (nm)	32		25	25		30		~20	~16	~16	~13	
	Fin Pitch (nm)	BULK		BULK	FDSOI		48 FinFET		Single Fin 42	FinFET 7LPP	FinFET 5LPE	Horizontal Nanosheets (HNS	
	CPP (nm)	114		86	86		78		68	54/57	54/57	40	
	M1 (nm)	90		64	64		64		51	36	36	32	
	SRAM	0.152µm2		0.084µm2			0.064/0.08µm2		0.04µm2	HD 6T SRAM 0.026µm2			
	Year Publication	ICSIST 2011		VLSI 2012			JSSC 2014		ISSCC/VLSI 2017	VLSI 2017/ISSCC2017-2018			
	Risk Prod	2011		2013			4Q-2015		1Q2017	1H-19	2H2020		
TSMC	Lg (nm)	30	30	30		33		*25*	25	18	16	~13	
	Fin Pitch (nm)	BULK	BULK	BULK		FinFET 45		*FinFET 45*	FinFET	FinFET 7FF	FinFET N5	Horizontal Nanowire (HNW)	
	CPP (nm)	118	105	90		90/80		*90/80*	64	54	50	45	
	M1 (nm)	90	80	64		64		*64*	42	40	28	22	
	SRAM	0.155µm2	0.155µm2			0.07µm2			0.03µm2	0.027µm2	0.021µm2		
	Year Publication	VLSI 2012	VLSI 2012	VLSI 2014		IEDM 2013		*6Track*	VLSI 2016	IEDM 2016	ISCC 2020		
	Risk Prod	2011	2018	2013		4Q-2015		*3Q2016*	4Q2016	3Q-17	1H2020		
GF	Lg (nm)		28				30						
	Fin Pitch (nm)		FDSOI				48 Fin FET						
	CPP (nm)		90				78						
	M1 (nm)		78				67						
	SRAM		0.110µm2				0.110µm2						
	Year Publication		IEDM 2016				IEDM 2016						
	Risk Prod		2016				2H-2016						
SMIC	Lg (nm)	30					30						
	Fin Pitch (nm)	BULK					48 Fin FET						
	CPP (nm)	118					78						
	M1 (nm)	90					67						
	SRAM												
	Year Publication												C.Reita, C.Fenouillet-Beranger - CEA-LETI - 2021
	Risk Prod	2016					2019						
												Forecast	

Figure 4: comparison of the technologies of various foundries

While during the last sixty years of microelectronics development it has been unclear which technical solutions would drive development ten years down the line, there was never any doubt that there was no fundamental technical limitation and that competition guaranteed that the necessary investments were made in a timely way. What is more, R&D activity provided a number of options with a degree of maturity that was instilled confidence that they would be ready when needed. Over the last ten years the situation has changed dramatically. Technically the specification for the 3nm node is very close to the physical limit for transport in semiconductors (gate length of 7 to 10nm) before stochastic phenomena introduce an intolerable degree of variability. No investigations carried out by R&D groups have found materials or device architectures that have the potential to behave better than silicon [3].

Around these dimensions, the device structures need a major new modification to keep the gate controlling the current flow effectively. Both FinFET (impossible to reduce the fin thickness for structural reasons) and FDSOI (not enough current per channel) need to move to horizontally stacked nanowires and nanosheets. In fact, this type of new structure will have the technology converge again as it can be seen as an evolution of FinFET where the fin is split in multiple layers or a series of FDSOI channels on top of each other. These, announced by Samsung and demonstrated in the past both by IBM and CEA-LETI [4], are really at the limit of what can be obtained in term of scaling of individual components [5]. Nothing better has been shown in literature in the last ten years and so we may well have to start looking more for further improvements in the other parts of systems such as in the assembly of multiple chips, management of the I/O, packaging, etc.

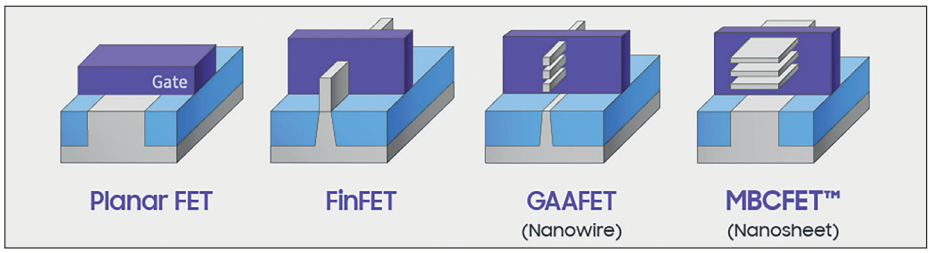


Figure 5: Evolution of the transistors' topology (from https://tellitlikeitisnews.com/samsung-announces-3nm-gaa-mbcfet-pdk-version-0-1/)

The introduction of production capable EUV tools, the delay in the deployment of which generated doubts about the feasibility of the scaling, has slowed down the increase in cost and has given a greater degree of freedom to the process, making the new device structure manufacturable. The road towards 2nm node (10nm gate length) 3D devices, while not free from risk, now looks viable. The target of such devices being in production in 2024 is a reasonable one and so the gain in density and performance should continue at the same rate as that of the last five years.

All the above is applicable to logic devices. Where memories are concerned, the very high regularity of their structure, both for Dynamic Random Access Memory (DRAM) or NAND Flash memory, has allowed a path that has been somewhat smoother but where the use of the third dimension has occurred earlier. This has happened as a result of two requirements: density improvement in the chip and very large throughput for the I/O. Some of the logic processes are being progressively introduced into the memory (the latest is the use of replacement metal gates in 3D

NAND) when necessary. Some of the typical memory 3D techniques will progressively filter back towards logic integration at the same time. In the case of memories devices too, then, there does not appear to be any real bottleneck in their continuous improvement for the next five years. In fact, the large amount of memory required in AI-dedicated circuits will probably drive them even harder towards higher densities and speed, with SSD disks increasingly competitive with mechanical hard disks.

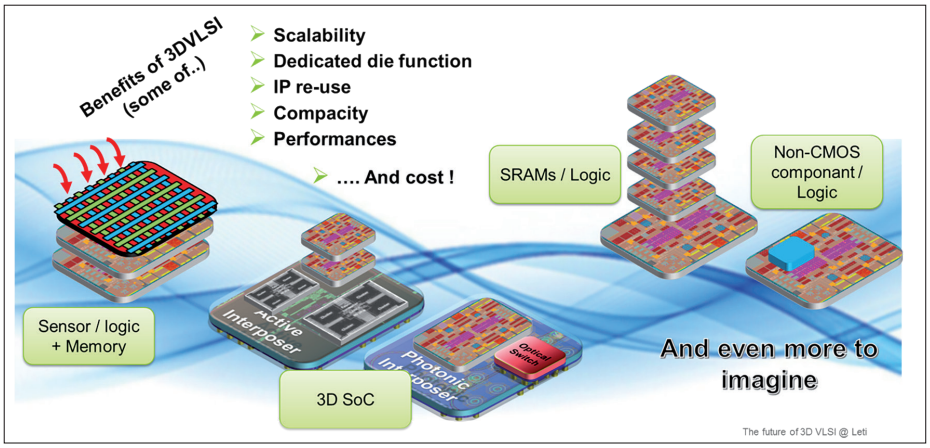


Figure 6: Benefits of 3DVLSI (Source: CEA-LETI)

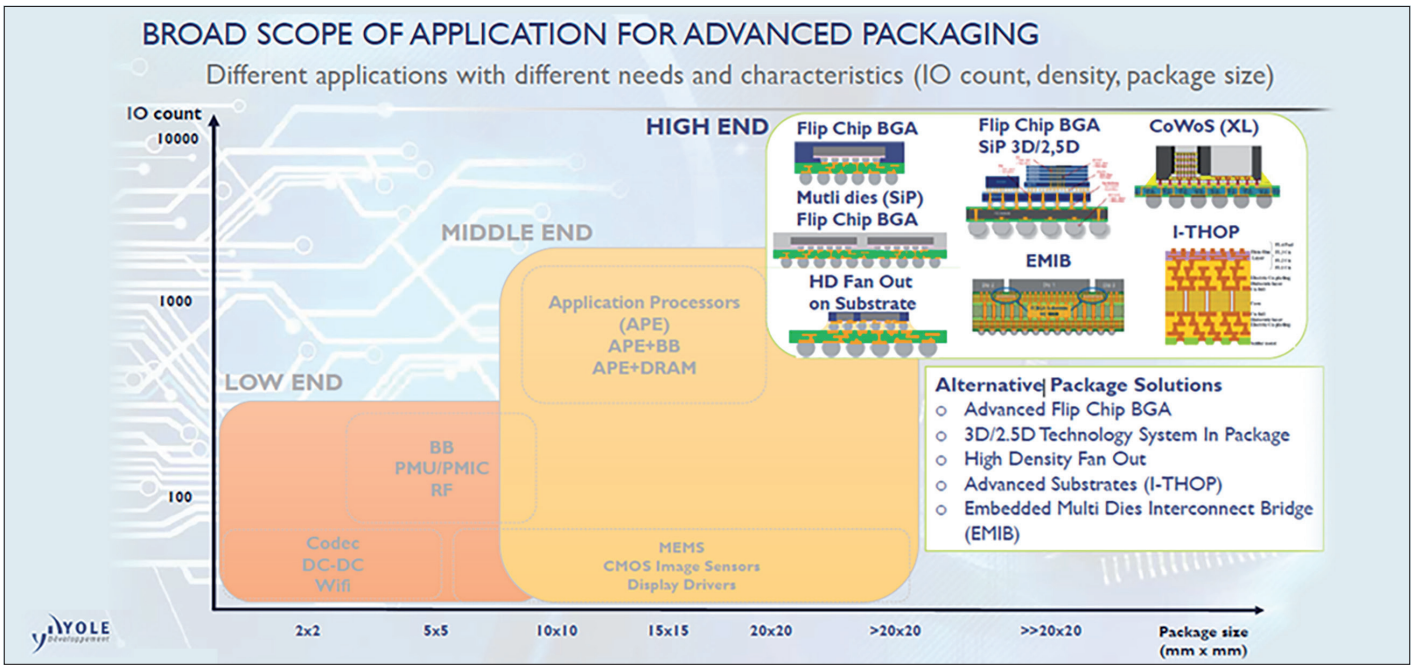


Figure 7: View of different advanced packaging solutions for high-end applications. (Source: Yole, 3DTSV & 2,5D 2016 report)

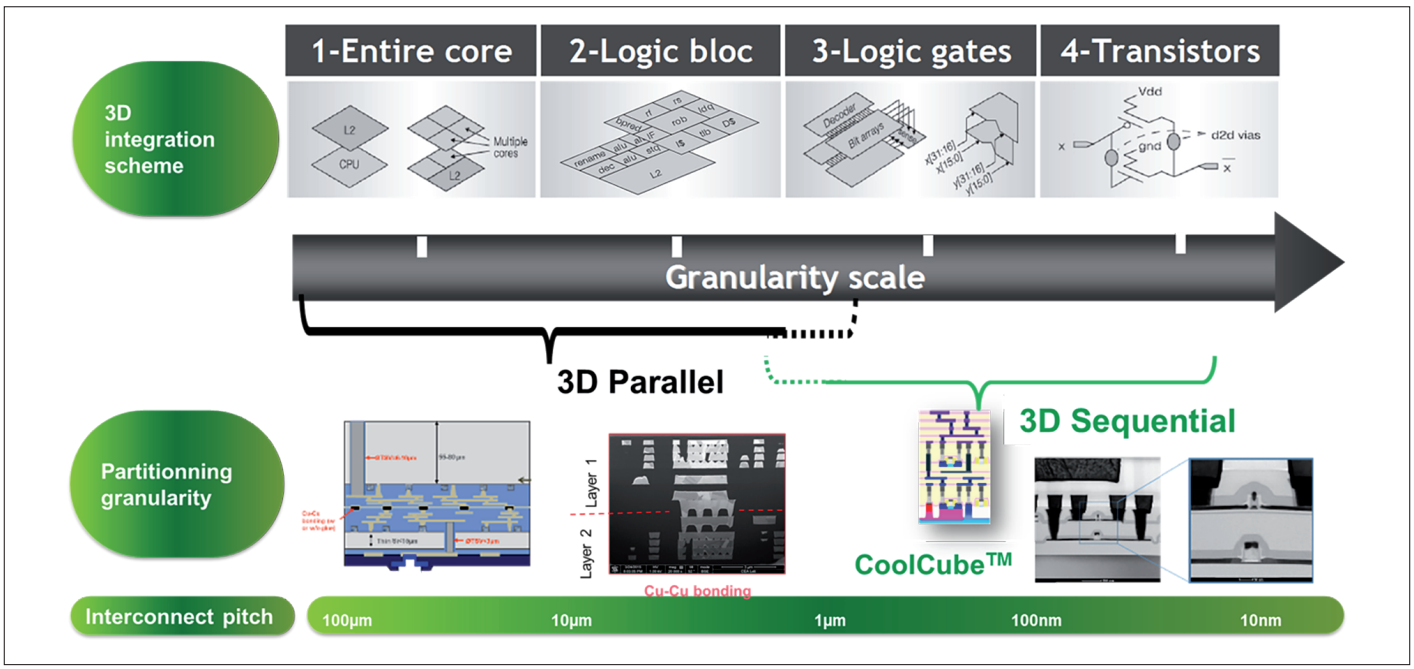


Figure 8: 3D parallel and 3D sequential positioning, depending on partitioning granularity and pitch of interconnect (Source: CEA-LETI)

Playing with 3D

3D-stacking is a promising new technology that will allow the density of transistors in systems to continue increasing by using the third dimension to stack dies of silicon one above another even if the transistors do not shrink any further. This technology is already in use for some specific products but needs to be more widely developed to enable greater diversity of chips at an acceptable cost.

Some of the key advantages of heterogeneous integration compared to classical planar architecture are:

- Transistors, or function by surface unit increase;
- Possible re-use of advanced intellectual property (IP), which allows faster time-to-market as well as cost reduction;
- The ability to mix chips with the most appropriate technology for each given function.

The first objective of heterogeneous integration, in this context, is to maintain or increase the energy efficiency of the global system while reducing its size. Several solutions are already available in foundries or outsourced semiconductor assembly and test providers, mainly driven by TSMC with their integrated fan-out (InFO) and CoWoS advanced packaging, or Intel with its EMIB and FOREVOS solutions.

Nevertheless, current solutions will soon face seriously challenges in terms of achieving high density of interconnects between separate functions (like those required by logic to memory) while achieving a dissipation of less than 1pJ/bit per vertical link.

This figure of merit naturally leads to two different but complementary integrations: 3DIC stack (also called 3D parallel) and 3D monolithic (also called 3D sequential). Here also, the technology will be chosen with regard to the architecture and the density required, as shown in Figure 8.

3DIC stack

3DIC stack with through-silicon via (TSV) and µbumps are well known, mainly for field-programmable gate array (FPGA) applications (partitioned dies on a passive

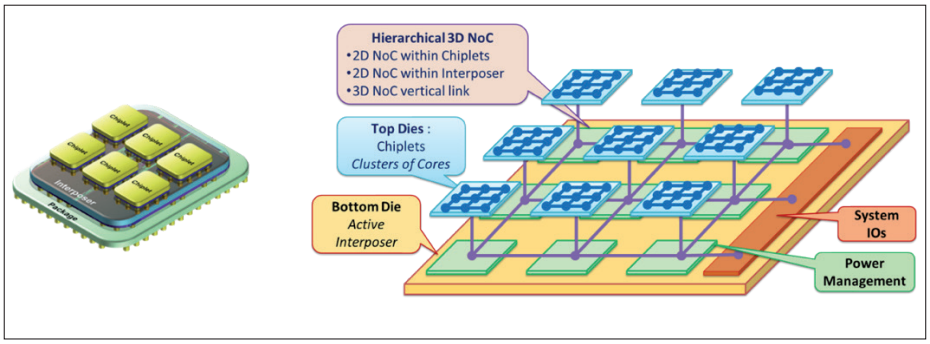


Figure 9: One possible architecture partitioning between chiplets and interposer (Source: Leti)

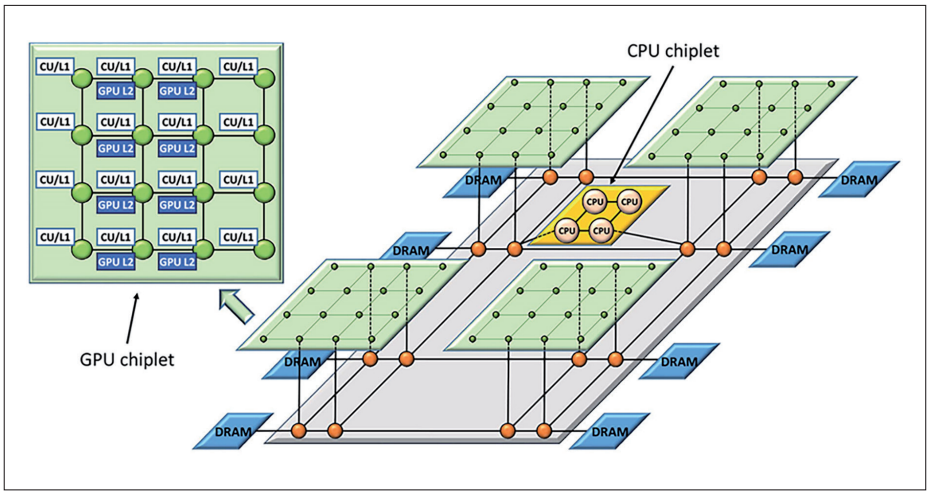


Figure 10: A future system might contain a CPU chiplet and several GPUs all attached to the same piece of network-enabled silicon (Source: AMD)

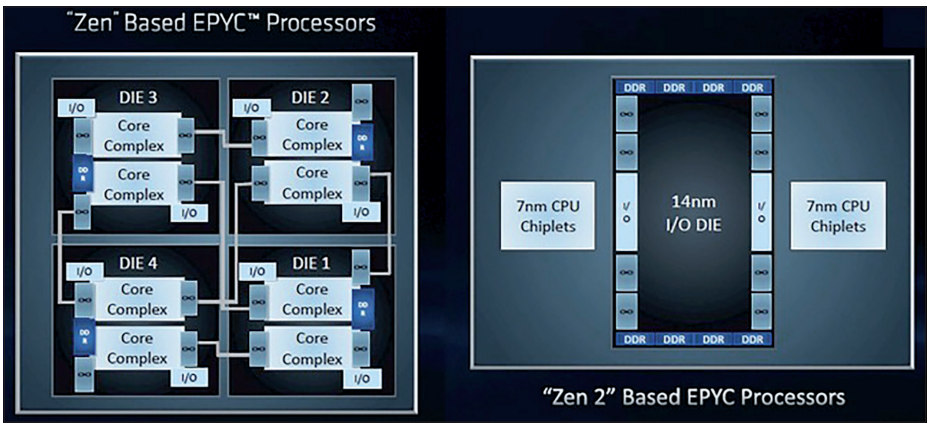


Figure 11: Zen Basezd EPYC Processors (Source: The Next Platform)

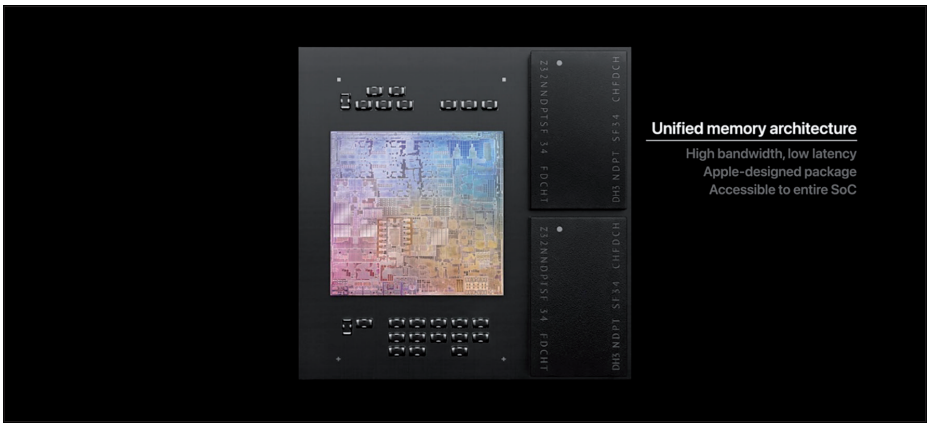


Figure 12: Apple M1 system in package

silicon interposer), and for high-bandwidth memory (HBM) and hybrid memory cube (HMC) stacking. The pitch between 3D features is in the range of 40µm, while a pitch of less than10µm is evaluated to reach <1pJ/bit for the consumption of each vertical link.

That's the reason why advanced 3D technology with very small TSV (a diameter in the range of the µm) and very fine-pitch chip-to-chip interconnects is required. The chip-to-chip fine-pitch interconnect may be based on the µbumps (copper and solder) interconnect or direct hybrid bonding. Wafer-to-wafer by direct hybrid bonding is also a potential solution, already famous for CMOS image sensor (CIS), by partitioning the sensing layer from the logic layer, and more recently by embedding in the stack a third layer with dynamic random access memory (DRAM).

As an example, DARPA's Common Heterogeneous Integration and IP Reuse Strategies (CHIPS) programme demonstrates work in this area, while AMD, INTEL and CEA-LETI have published work and launched initiatives for advanced silicon interposers and chiplets.

In several application domains, advanced 3D integration may advantageously replace a multi-core monolithic planar die.

A further advantage of die stacking is the innovative potential it provides to introduce novel materials, such as III-V chemical compounds like gallium nitride (GaN), onto silicon CMOS wafers. This is another example of "using the right technology and the right material for the right function". Additionally, this would save some rare or expensive materials by limiting their use.

The roadmap of alignment accuracy is already quite clear with the objective of reaching under 10µm of pitch having already been achieved and advanced proofs-of-concept from laboratories having delivered a 1µm pitch for wafer-to-wafer hybrid bonding (CEA-LETI, 2017), or 3µm for a µbumps interconnect (Imec, 2017).

On the industrial side, the AMD "Rome" Epyc processors use this principle: all of the I/O and memory controllers are set into a 14nm "interposer" that sits at the centre of the Rome package. The compute chiplets are 7nm.

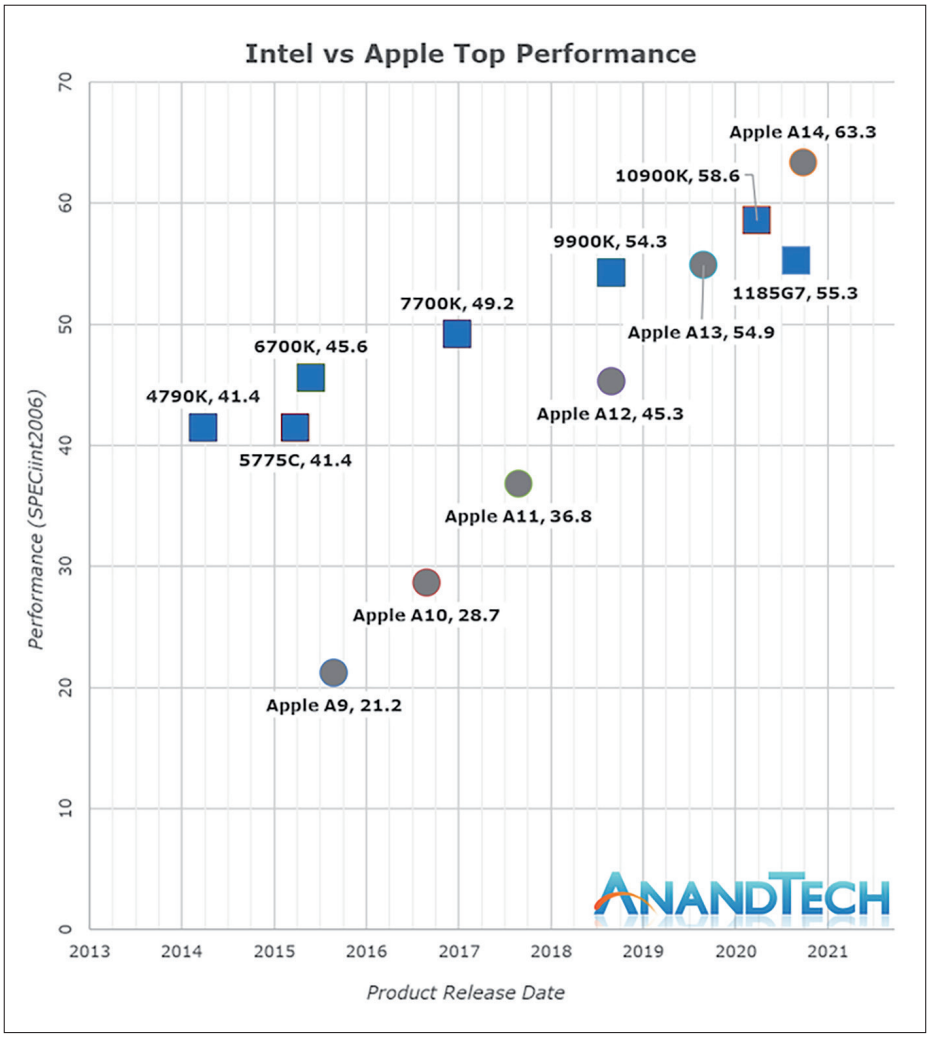


Figure 13: According to Anandtech [9] "Whilst in the past 5 years Intel has managed to increase their best single-thread performance by about 28%, Apple has managed to improve their designs by 198%, or 2.98x (let's call it 3x) the performance of the Apple A9 of late 2015".

The Apple M1, designed with TSMC 5nm technology, uses a similar approach, with the main core and the memory on the same organic interposer [9]. This allows the dissipation of energy to be reduced while moving data from/to memory, and the size of the communication busses to be increased. Together with a carefully designed memory infrastructure, the use of a multiplicity of coprocessors and processors with large L2 caches and wider microarchitecture, featuring an 8-wide decode block, Apple's high performance cores are currently by far the most widely commercialized design in the industry [8].

3D sequential

3D sequential, which consists of the wafer-level manufacturing of a low-temperature device layer on top of a standard Front-End-Of-Line, introduces the notion of very high density on the transistor-to-transistor interconnect, opening up a range of new architectures (sensor on top of CMOS, low-energy CMOS such as FDSOI on top of high performance FinFET devices etc..).

The alignment of the two layers is performed by lithography (rather than by bonding) which means a possible sub-10nm alignment accuracy between the layers.

The various 3D integrations outlined above are not at the same level of maturity. While wafer-to-wafer stack is already

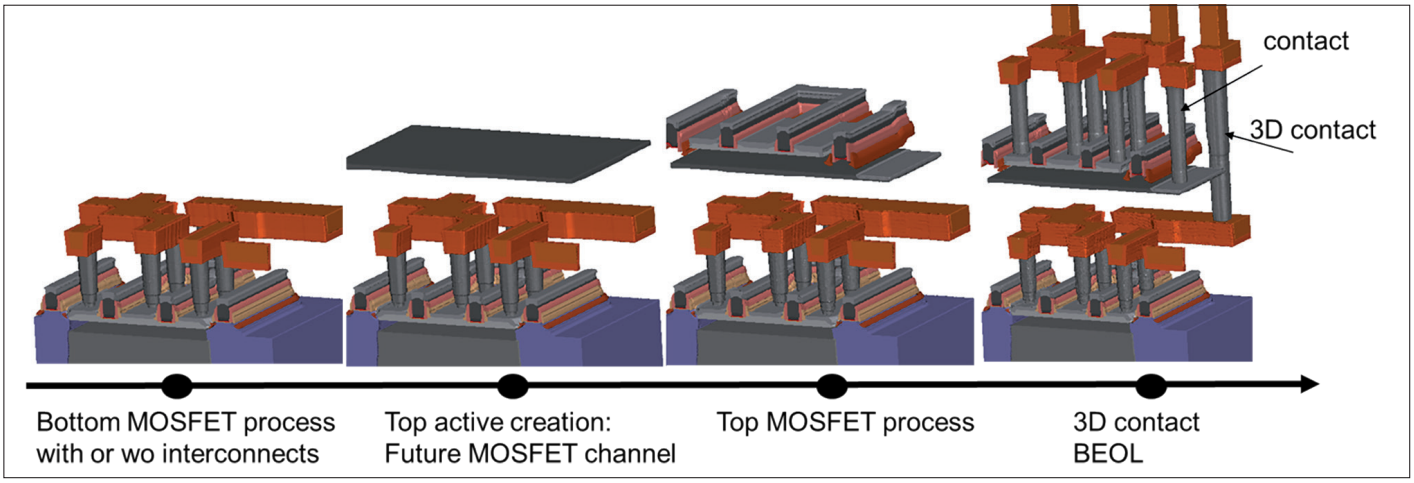


Figure 14: Principle of 3D sequential integration (Source: CEA-LETI)

in production in a ≈5µm pitch for image sensors, other solutions may require a learning curve of five to 15 years depending on the technology. In addition to integration challenges, design and computer aided design (CAD) tools need some disruptive innovations in order to fully exploit such technologies.

To conclude, the evolution of transistor and 3D technologies is mandatory for the pursuit of solutions to computing's endless requirement for increasing performances (more compute, more memory and still higher efficiency). While the number of actors active in advanced node development is very limited (three are still in the race, TSMC being perhaps the only one in near future), heterogeneous integration will provide to end-users both the possibility of differentiation and another wave of innovation. Europe should find, or find again, an active role in this domain, leveraging its successes in the 3D arena.

References

[1] Gordon Moore, "Cramming More Components onto Integrated Circuits," Electronics Magazine Vol. 38, No. 8 (April 19, 1965).

[2] Gordon Moore, "Progress in Digital Integrated Electronics" IEEE, IEDM Tech Digest (1975) pp.11-13

[3] D.E. Nikonov, I.A. Young, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits 2015

[4] Scotten Jones, "IEDM 2017 – LETI Gate-All-Around Stacked-Nanowires", SemiWiki, 12 Feb 2018, https://www.semiwiki.com/forum/content/7282-iedm-2017-LETI-gate-all-around-stacked-nanowires.html

[5] Scotten Jones, "7nm, 5nm and 3nm Logic, current and projected processes", https://www.semiwiki.com/forum/content/7544-7nm-5nm-3nm-logic-current-projected-processes.html

[6] https://zenodo.org/record/3947824#.X_TdU-DjJTY

[7] https://www.anandtech.com/show/14666/tsmc-3nm-euv-development-progress-going-well-early-customers-engaged

[8] https://en.wikipedia.org/wiki/3_nm_process

[9] https://www.anandtech.com/show/16226/apple-silicon-m1-a14-deep-dive

Carlo Reita is a researcher at the Research and Technology Department of CEA (Alternative energies and Atomic Energy Commission), France and tasked with AI and Digital technology roadmapping.

Severine Cheramy is a researcher at the Research and Technology Department of CEA (Alternative energies and Atomic Energy Commission), France.

Marc Duranton is a researcher at the Research and Technology Department of CEA (Alternative energies and Atomic Energy Commission), France and the coordinator of the HiPEAC Vision 2021.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: C. Reita, S. Cheramy and M. Duranton. Silicon technology is still in the game. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 108-115, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

Although no one can yet forecast if or when it will become an operational technology, quantum computing R&D efforts will nevertheless have a profound impact on the future of computing as it is boosting research in important topics such as hybrid architectures, and algorithmic and computational complexity. It also forces us to re-think the very nature of information processing.

Towards operational quantum computing? Or: thinking beyond qubits

By Christian Gamrat

The paper published in Nature by the Google team in late 2019 took the computing community by storm. However, if quantum supremacy - or more realistically, operational quantum computing - ever becomes a reality, it will probably not be the result of a milestone event. It could materialize in very specific fields, mostly unnoticed, and will be very progressive. The fact is that quantum computing must first address huge engineering challenges, in both hardware systems and software. At the same time, classical and algorithmic computing are progressing, as is our understanding of application needs and computational complexity.

Quantum supremacy is a moving target. It moves according to progress in non-quantum information processing and our better understanding of when and how to tackle hard problems. Quantum computing is and will remain a hybrid computing approach in which advances in the fields of quantum, classical and artificial intelligence (AI) computing will enrich each other. This is why the system and software stacks making up the future quantum computer will be an important area for future developments. It is probably the right time to think beyond the qubit.

Key insights

- No matter what, a quantum computer will be a cloud-based hybrid classical/quantum computer. HPC infrastructures will be the ideal playground for developing such a hybrid system stack.
- A consequence of the "cloudification" of quantum computing could trigger a move towards a European quantum cloud, similar to EBRAINS [21], as seen in the Human Brain project.
- R&D in system architecture and software stack for quantum computing (QC) needs to be pushed. With European assets in the system, software and HPC technologies this could be a good opportunity for Europe.
- The fields of quantum computing and AI are becoming closer. This might appear to be the result of a buzzword strategy but there are real scientific and technological reasons behind the scenes.
- Quantum computing sheds new light on computational complexity theory. It appears to be the ideal tool to rejuvenate basic research in the field of information theory.
- Fears emerged that the longer-term research needed to tackle QC might be negatively impacted by new funding strategies in the aftermath of COVID-19, but the potential impact of QC will probably prevail.
- Startups in quantum computing are emerging. Europe is on the right path.

Key recommendations

- Support R&D in the field of architecture and software stacks for quantum computing.
- Develop the integration of quantum accelerators into future exascale infrastructure.
- Promote the development of a European quantum cloud.

Quantum computing is a rapidly developing field. It is currently raising a lot of hopes and creating a lot of noise both in the world of research and in the world of industry. In principle, the realization of machines applying quantum principles could make it possible to deal with problems that are difficult or impossible to tackle with conventional computers. The applications fields that would benefit the most from quantum computing are those with problems incurring an exponentially large number of variables with respect to the size of the problem. Such problems are often found in chemistry, pharmacology, physics, cryptography, optimization and machine learning.

However, the design of a quantum computer capable of operating efficiently on these problems remains a long-term prospect. At the hardware level, the realization of the indispensable qubits with potential for sufficient fidelity and scalability is a major research field with several technologies being actively investigated. Superconducting qubits have pioneered the development of quantum computing and have been demonstrated [1]. They are now playing a role in the development of a spin qubit technology based on silicon semiconductors [2]. At the application and algorithmic level, there is a broad range of work on various platforms. There is clearly a gap between experimental efforts at the hardware level and more theoretical research at the algorithmic level.

In a paper published in Nature, a Google/NASA team claimed to have reached quantum supremacy [3]. Quantum supremacy is defined as "the potential ability of quantum devices to solve problems that classical computers practically cannot" and was initially introduced by John Preskill [4].The Google/NASA team actually achieved this but on a rather limited problem. In fact, they used a 54 superconducting qubits chip (the "Sycamore chip") with nearest neighbours' interactions to run a circuit that generated near perfect random sequences of bits. It's just like rolling a dice with several millions faces instead of just six. Indeed, to simulate a near perfect distribution over rolling such a dice with a powerful classical computer it could take a fair amount of time. In the paper they estimate that it would take 10,000 years if run on the best Google servers and require a few PW (10^{15} Watts) of electrical power. They claimed the Sycamore chip can do the same computation (generate a sequence of several million random numbers) in 600 seconds using as little as 10 kW of power. That's a really impressive result! Does this mean that quantum computing is ready, that the race is over, and we shall all start ditching our good old classical computers? Not so fast.

The fact is, if some sort of quantum advantage was demonstrated, it was on a rather useless application. Nevertheless, the scientific achievements were important as the paper demonstrated that errors do not depend on entanglement and computational complexity, that no new decoherence physics was detected (an indication that error correction algorithms should be working) and that quantum computing works on a rather large system. If this paper did not show that quantum computing is ready for real applications, it showed that its basic principles seem to be valid.

Qubit technologies update

Besides progress in qubit technologies, one of the main questions that need to be tackled on the road to operational QC is that of scalability. Depending on the qubit implementation technology, the challenges may vary but one that sticks is the problem of noise. The term noise refers to the impact of various factors leading to a loss of precision of the qubits states (Figure 1). Those factors can result from fabrication imperfections, sensitivity to parasitic signals, crosstalk, material degradation, etc. Digital computing technologies such as CMOS are facing many of the same problems; after all, a CMOS transistor also suffers from noise, parasitic interferences and crosstalk but the difference is that the classical binary coding scheme provides an efficient shield against those underlying sources of noise.

Hybrid HPC quantum and the quantum cloud

A quantum computer cannot work without a classical Von-Neumann type computer! Indeed, if quantum information processing can bring benefits in terms of parallelization of massive calculations (superposition of states), it will not bring anything significant in logic and control operations. Moreover, the problems we are interested in solving along with their potential solutions exist in our very classical world. Therefore, a quantum computer is and will remain a hybrid-computing

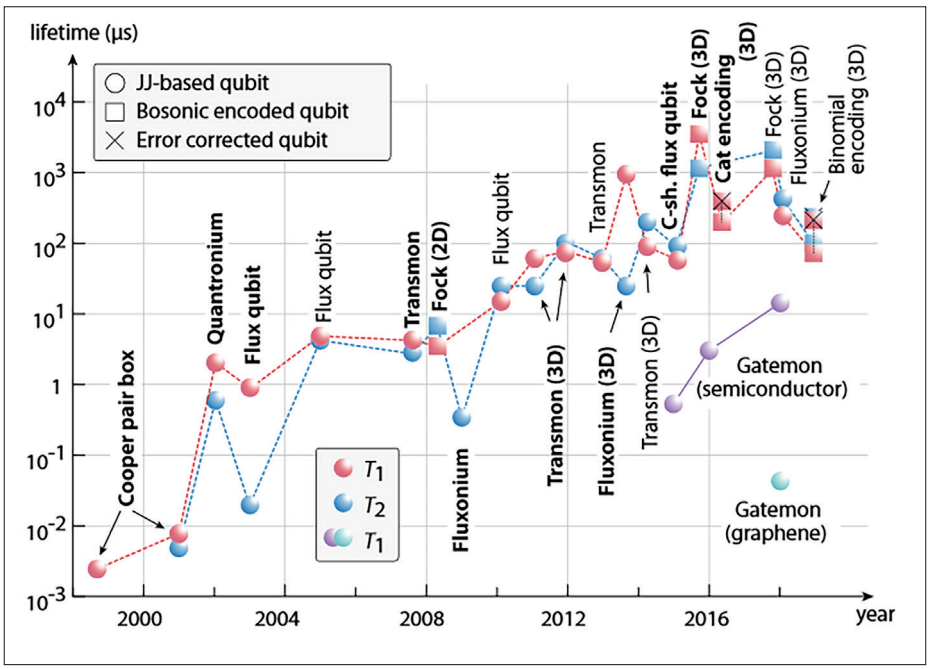


Figure1: Evolution of lifetime and coherence time of superconducting qubits. From Kjaergaard et al.

machine with a "boring" classical computer doing the input/output interface and probably a big bulk of the data-processing logic. We have then the concept of a quantum accelerator core that, just like a GPU core, can speed up graphical computations or, like an FPGA, can speed up specific tasks hand-in-hand with their host processor (Figure 2).

Most of the application domains that would benefit from quantum computing involve extremely large sets of variables. These domains include chemistry, physics, cryptography and machine learning. Such applications can only be processed using large computing infrastructures. A typical example being artificial intelligence in which it is common to have embedded processors (in a portable device, in a car) running the inference phase (e.g. pattern recognition/classification) while the computationally heavy machine learning phase is offloaded onto a HPC server. Large computing platforms are therefore the primary target for the integration of quantum accelerators and the development of hybrid quantum-classical programming methodologies. This is for example what the Jülich UNified Infrastructure for Quantum computing (JUNIQ) is currently introducing [22] (Figure 3). JUNIQ will integrate quantum computers (e.g. ATOS QLM) and quantum annealers (e.g. D-Wave) in the form of quantum-classical hybrid computing systems into the modular HPC environment of the Jülich Supercomputing Centre.

On another level, QuTech in the Netherlands (TNO and TU-Delft) has introduced a quantum computing platform to help run and evaluate quantum algorithms on a variety of simulators or hardware [5, 23] (Figure 4). With this platform the focus is put on the ease of access to a quantum simulation environment for experimenting with quantum processing on both simulated and real hardware platforms.

Those European based QC platforms represent a bold move towards offering state of the art quantum hardware and quantum programming tools to both industry and academic users in addition to the online platforms already proposed by the likes of Google (Cirq) [6], IBM (QisKit), DWave,

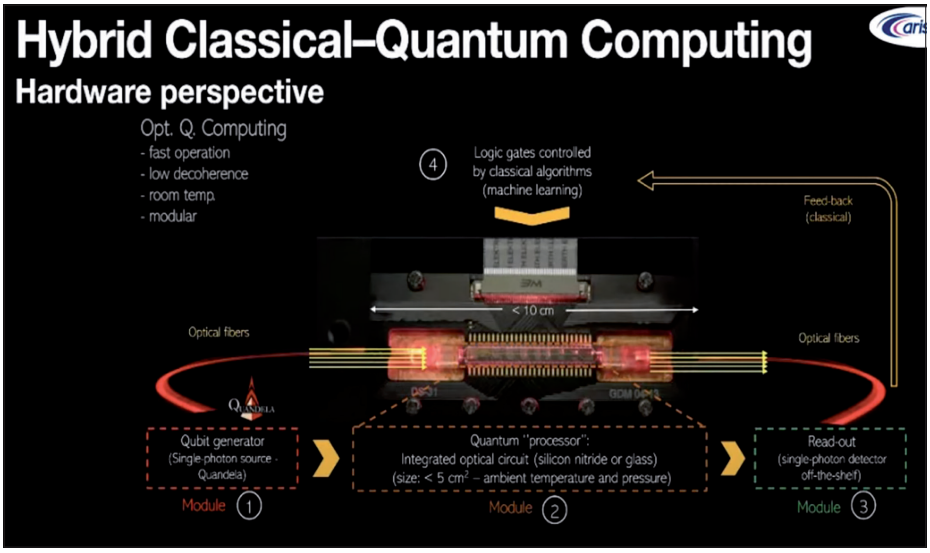


Figure 2: Perspective of a Hybrid Classical-Quantum computing architecture based on single photon source qubits as proposed by the Quandela startup company (https://www.nature.com/articles/s41467-020-19341-4)

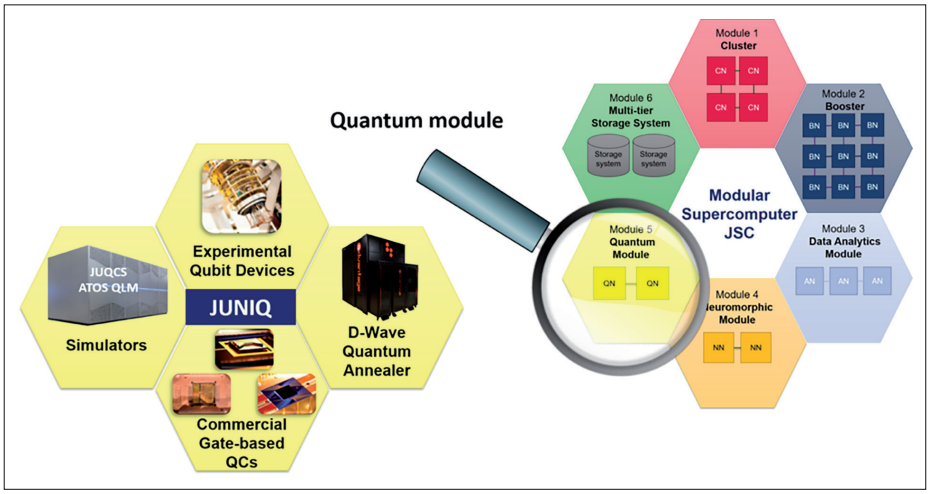


Figure 3: JUNIQ is a uniform quantum computing Platform as a Service (QC-PaaS) proposed by the Jülich Supercomputer center.

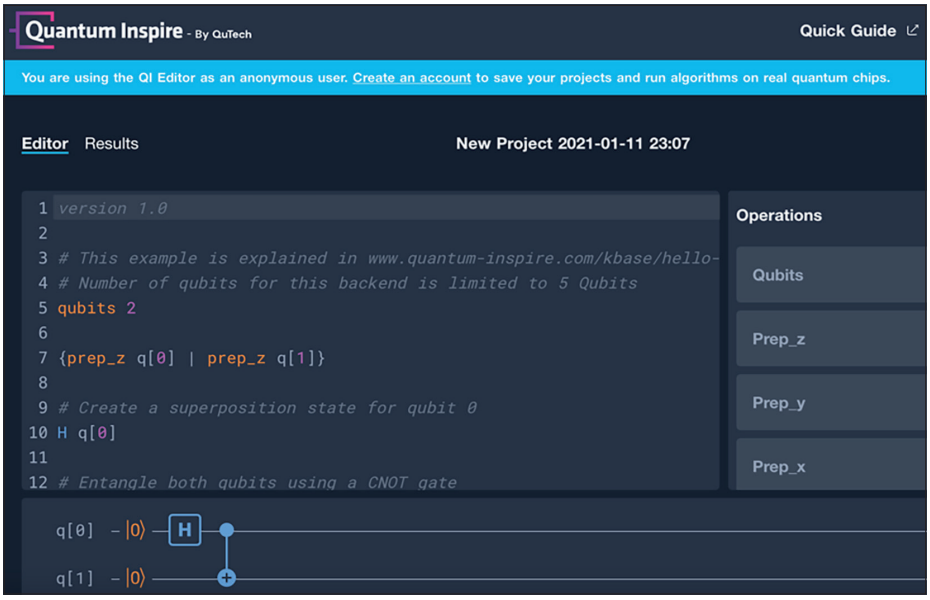


Figure 4: The QI Editor from Quantum Inspire allows users to write, run and test quantum algorithms online.

IONQ or Microsoft (QDK). The advent of European cloud-based quantum computing platforms might also mitigate some confidentiality and sovereignty concerns. The recent announcement by ATOS of the quantum metrics reference "Q-Score" [24], which aims to benchmark and measure the capacity of quantum processors to solve real problems will also help experimentation with the effectiveness of quantum computing.

Toward a system and software stack for quantum computing

As the trend towards integrating quantum accelerated processors with classical (probably CMOS based) computers accelerates, the need arises to develop specific system architectures and software layers that will orchestrate the execution of both classical and quantum circuits required to run real applications.

However, to achieve this goal, it is necessary to address the huge gap between the current experimental efforts at the qubits level and the more theoretical research at the quantum algorithmic level. In order to bridge the gap, we need to design an architecture (hardware, software and system) that will allow the programming and the execution of applications on the available quantum technology.

This architecture will make it possible to link a "software stack" made up of languages, compilers and other tools for the description of applications and their algorithms with an "execution stack" responsible for the orchestration of the physical circuits. If the software/hardware duality is at the heart of modern digital computing systems, it will also be the case for quantum computers in addition to the classical/quantum duality. Therefore, as said earlier, a quantum computer will be composed of a set of quantum operators acting on qubits tightly coupled with a set of classical processors realizing a hybrid computing architecture. Just like in modern microprocessors, the efficiency of a future quantum computer will depend, for a large part, on the design of the hybrid interfaces: software/hardware, classical/quantum. The key role of the software and architecture stack for quantum computing has been highlighted in pioneering work by the University of Delft [7] (Figure 5).

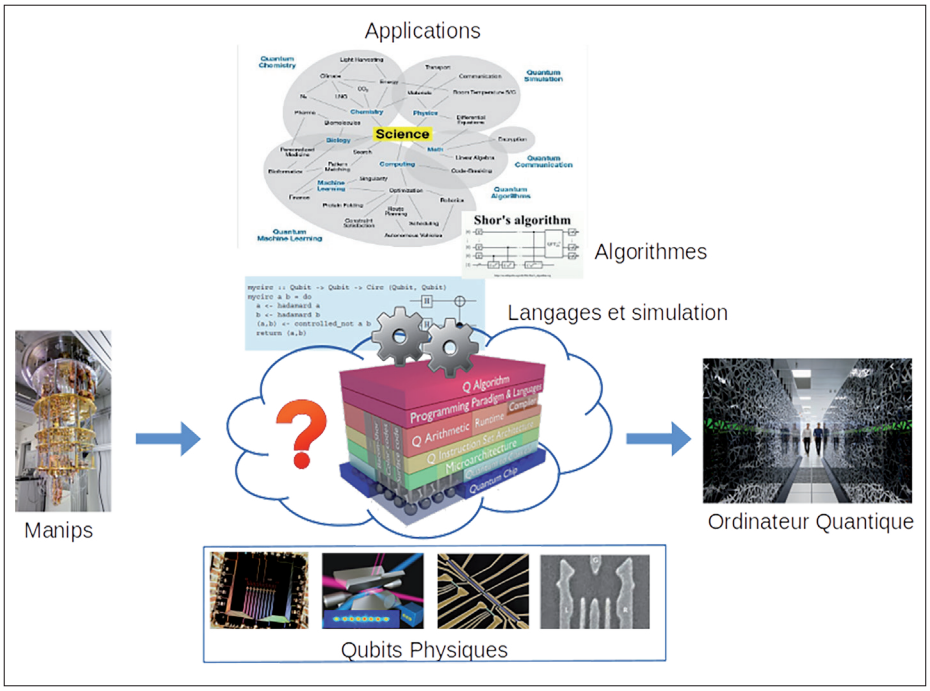


Figure 5: Bridging the gap between physical qubit technologies and applications and experiments on operational quantum computers, through the development of the system and software stack (part from @fu2016).

Although software platforms for hybrid programming have recently been proposed [6] (in particular in the field of machine learning), they are still either too abstract or too proprietary to allow for a simple interface with the physical qubits currently under development. However, they represent a starting point that would be useful as a basis for a coordinated R&D effort and help bridge the gap between physical qubits and high-level quantum algorithms. Software and system research is often overlooked and considered as a "simple" engineering endeavour. It is however the quality of the system stack that will make the difference between an experiment with qubits and an operational quantum computer. We believe that, together with a strong position in cloud-based quantum infrastructure, a healthy and well-supported research base in the field of system architecture and software stack has the potential to put Europe at the forefront of quantum computing.

Quantum computing boosts research in computational complexity

A recent work [8] shed further light on quantum supremacy by introducing theoretical elements showing that such a supremacy could be challenged on a classical laptop computer if limitations of real quantum computers are taken into account. The paper acknowledges the fact that perfect quantum computers are exponentially difficult to simulate on a classical computer. However real quantum computers are not and will likely never be perfect and will suffer from various sources of decoherence and imprecision. If those limiting factors are considered when designing an algorithm for a classical computer, then the authors [8] shows that the simulation time can be drastically reduced. This remark on simulation time of a real quantum computer stresses the importance of computational complexity (Figure 6).

Paradoxically we might face a situation in which we solve most QC challenges, build the QC, program the QC and finally end up in a situation where the horizon of "quantum supremacy" has been pushed away by better knowledge of and progress made in problem solving.

Research in quantum can also have good side effects: low temperature electronics for controlling and interfacing qubits can lead to cryogenic computing, and research in quantum algorithms can also improve classical algorithms: as an example, the young Ewin Tang has proven that classical computers can solve the "recommendation problem" nearly as fast as quantum comput-

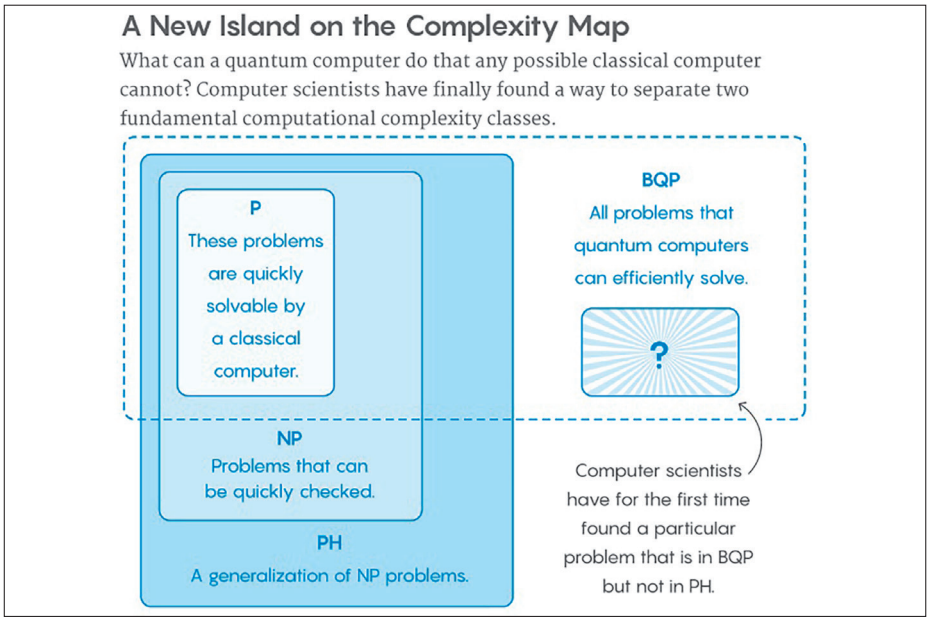


Figure 6: Overview of complexity classes, including Bounded-error Quantum Polynomial time (BQP)

ers. The recommendation problem relates to how services like Amazon and Netflix determine which products you might like to try. Computer scientists had considered it to be one of the best examples of a problem that's exponentially faster to solve on quantum computers — like Kerenidis and Prakash's algorithm [9], Tang's algorithm ran in polylogarithmic time (meaning the computational time scaled with the logarithm of characteristics like the number of users and products in the data set) and was exponentially faster than any previously known classical algorithm [10].

This example shows that regardless of the actual availability of QC hardware, rethinking algorithms and programming in the light of quantum information brings a new vision on the actual limits of computer problem solving. Quantum supremacy is a moving target.

The QC and AI entanglement

Quantum computing and artificial intelligence are often cited like a winning combination. A number of recent scientific works emphasized the interactions between these two technologies [11]. It is hard to tell if this will become a reality or if this is just the result of a buzzwords strategy.

Machine learning (ML), which is at the heart of the current incarnation of AI, is a multi-parametric optimization problem and it can in principle benefit from advances in QC or quantum annealing. At the same time, progress in ML and AI can also offer benefits to QC. It is particularly true in the field of quantum state measurements. Machine learning-based measurement protocols have shown interesting results with both supervised [12] and unsupervised [13] learning methods.

Google has released an open source library aimed at bridging the gap between AI and quantum computing: "TensorFlow Quantum" [25]. In fact TensorFlow Quantum (TFQ) is the result of a merging of the TensorFlow library (TF) and the Google Quantum circuit simulator CIRQ [6].

Will quantum computing boosted by AI be able to actually help solve real problems in a foreseeable future? In some ways, AI techniques with their abilities to capture hidden mechanisms might be part of the path to a solution. However, even AI is not magic, and it has been shown many times that in order to be efficient, AI methods need a little help from experts. That is to say: some physical knowledge (in our case quantum mechanics) has to be added into the overall equation.

Like we said earlier, it was an obvious observation that a quantum computer will be a hybrid classical-quantum machine, but we also come to realize that QC might even be more of a hybrid than that! It is quite possible that quantum computing could only be achieved as the result of a close coupling between quantum and machine learning techniques and algorithms: the quantum-AI hybridization.

Are we heading towards a QC winter ?

Recently, several news items have highlighted the idea that investments in QC could be significantly affected in comparison to recent times [14-17] (Figure 7). Another indicator is the fact that most of the over-hyped announcements by IBM, Google, Intel and others are still very far from being ready for real applications. Finally, the announcement of John Martinis' resignation from Google [18] might have raised suspicion about the confidence that the company behind the 2019 Quantum Supremacy announcement has in the rapid development of QC. Officially, Martinis said he had resigned because Google was putting him off the executive role in its Quantum research unit [19]. However, it is not difficult to understand that, as a respected scientist, Martinis' agenda for QC is probably different from Google's corporate agenda. Indeed, claiming "quantum supremacy" in Nature does not mean that an operational quantum computer will soon be ready for real applications. There were actually two possible readings in Google's paper: on one side the corporate reading for the investors, mainly the title and abstract, expecting rapid return on investments. On the other side was the scientific reading for the academics, who saw a very nice experiment on fundamental (albeit not directly applicable) aspects of quantum computing. Obviously, the time frame of both readings does not stick to the same roadmap.

QC startups in Europe

A number of startup companies in the field of quantum computing have emerged over the last few years [20] (Table 1). Even if a few are just consultancy companies, a number of them are developing and marketing hardware, software or communication products. Their activities range from development of single photon sources that could fuel quantum networks or photonic based quantum computing (Qandela) to designing complete quantum

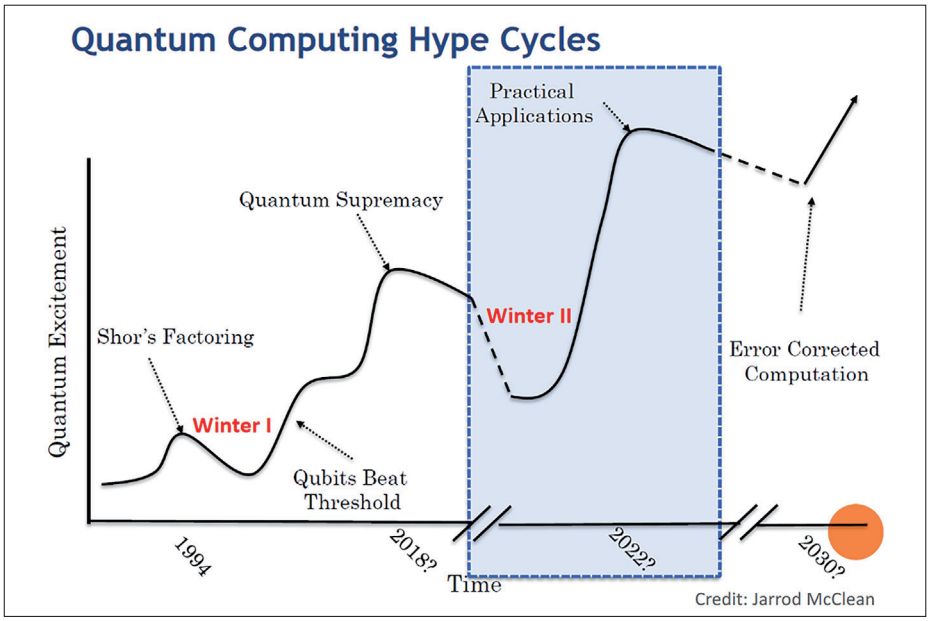


Figure 7: Quantum computing hype cycle https://extremecomputingtraining.anl.gov/files/2019/08/ATPESC_2019_Dinner_Talk_9_8-8_Alexeev-QC_Trends.pdf

processors with atomic arrays (PasQal). Some are working on developing software IPs for enterprise and industrial customers (AppliedQubit), while others are developing quantum software platforms for drug discovery (ApexQubit), or working in the field of quantum-secured communications (ID Quantique) and quantum network technologies and distributed quantum computing (VeriQloud).

	Hardware	Software	Communication
France	5	2	2
Germany	2	3	3
Greece		1	
Italy		1	
Norway		1	
Poland		2	
Spain	2	4	
Sweden	1		
Switzerland	2		1
The Netherlands	4	2	
UK	7	6	11
Other	3	5	1
Total	**26**	**27**	**18**

Table 1: The QC startup landscape Europe (from https://quantumcomputingreport.com/privatestartup/)

References

[1] D. Vion et al., "Manipulating the Quantum State of an Electrical Circuit," Science, vol. 296, no. 5569, pp. 886–889, May 2002, doi: 10.1126/science.1069372.

[2] R. Maurand et al., "A CMOS silicon spin qubit," Nat. Commun., vol. 7, p. 13575, Nov. 2016, doi: 10.1038/ncomms13575.

[3] F. Arute et al., "Quantum supremacy using a programmable superconducting processor," Nature, vol. 574, no. 7779, pp. 505–510, Oct. 2019, doi: 10.1038/s41586-019-1666-5.

[4] J. Preskill, "Quantum computing and the entanglement frontier," ArXiv12035813 Cond-Mat Physicsquant-Ph, Mar. 2012, Accessed: Nov. 11, 2018. [Online]. Available: http://arxiv.org/abs/1203.5813.

[5] TNO, "Minister Ingrid van Engelshoven and European Commissioner Mariya Gabriel launch Europe's first quantum computer in the cloud: Quantum Inspire," TNO, Apr. 2020. /en/about-tno/news/2020/4/europe-s-first-quantum-computer-in-the-cloud-quantum-inspire/ (accessed Jan. 12, 2021).

[6] M. Broughton et al., "TensorFlow Quantum: A Software Framework for Quantum Machine Learning," ArXiv200302989 Cond-Mat Physicsquant-Ph, Mar. 2020, Accessed: Mar. 10, 2020. [Online]. Available: http://arxiv.org/abs/2003.02989.

[7] X. Fu et al., "A Heterogeneous Quantum Computer Architecture," in Proceedings of the ACM International Conference on Computing Frontiers, New York, NY, USA, 2016, pp. 323–330, doi: 10.1145/2903150.2906827.

[8] Y. Zhou, E. M. Stoudenmire, and X. Waintal, "What Limits the Simulation of Quantum Computers?," Phys. Rev. X, vol. 10, no. 4, p. 041038, Nov. 2020, doi: 10.1103/PhysRevX.10.041038.

[9] I. Kerenidis and A. Prakash, "Quantum gradient descent for linear systems and least squares," ArXiv170404992 Quant-Ph, Apr. 2017, Accessed: Aug. 02, 2018. [Online]. Available: http://arxiv.org/abs/1704.04992.

[10] E. Tang, "A quantum-inspired classical algorithm for recommendation systems," ArXiv180704271 Quant-Ph, Jul. 2018, Accessed: Aug. 02, 2018. [Online]. Available: http://arxiv.org/abs/1807.04271.

[11] R. G. Melko, G. Carleo, J. Carrasquilla, and J. I. Cirac, "Restricted Boltzmann machines in quantum physics," Nat. Phys., vol. 15, no. 9, pp. 887–892, Sep. 2019, doi: 10.1038/s41567-019-0545-1.

[12] V. Havlíček et al., "Supervised learning with quantum-enhanced feature spaces," Nature, vol. 567, no. 7747, pp. 209–212, Mar. 2019, doi: 10.1038/s41586-019-0980-2.

[13] K. A. McKiernan, E. Davis, M. S. Alam, and C. Rigetti, "Automated quantum programming via reinforcement learning for combinatorial optimization," ArXiv190808054 Quant-Ph, Aug. 2019, Accessed: Nov. 30, 2019. [Online]. Available: http://arxiv.org/abs/1908.08054.

[14] E. Gent, "Investment in Quantum Computing Is Booming—But Will a Quantum Winter Follow?," Singularity Hub, Oct. 14, 2019. https://singularityhub.com/2019/10/14/investment-in-quantum-computing-is-booming-but-will-a-quantum-winter-follow/ (accessed Jan. 12, 2021).

[15] M. Swayne, "Quantum Winter? Is the Red Hot Quantum Startup Space About to Turn Ice Cold?," The Quantum Daily, Mar. 09, 2020. https://thequantumdaily.com/2020/03/09/quantum-winter-is-the-red-hot-quantum-startup-space-about-to-turn-ice-cold/ (accessed Jan. 12, 2021).

[16] InsideQuantumTechnology, "'Quantum Winter' a Concern," Inside Quantum Technology, 2020. https://www.insidequantumtechnology.com/news/quantum-winter-concern/ (accessed Jan. 12, 2021).

[17] "Quantum Computing in 2021," Strategic Finance. https://sfmagazine.com/technotes/december-2020-quantum-computing-in-2021/ (accessed Jan. 12, 2021).

[18] Wired, "Google's Head of Quantum Computing Hardware Resigns," Wired, Apr. 2020.

[19] M. I. and Strategy, "Google's Top Quantum Scientist Explains In Detail Why He Resigned," Forbes, Apr. 2020. https://www.forbes.com/sites/moorinsights/2020/04/30/googles-top-quantum-scientist-explains-in-detail-why-he-resigned/ (accessed Jan. 12, 2021).

[20] "Quantum gold rush: the private funding pouring into quantum start-ups." https://www.nature.com/articles/d41586-019-02935-4 (accessed Jan. 12, 2021).

[21] "EBRAINS is powering a new era in Brain Research", https://ebrains.eu, (accessed Jan. 12, 2021).

[22] "JUNIQ - Jülich UNified Infrastructure for Quantum computing", https://www.fz-juelich.de/ias/jsc/EN/Expertise/JUNIQ/_node.html, (accessed Jan. 12, 2021).

[23] "The multi hardware Quantum Technology platform", https://www.quantum-inspire.com, (accessed Jan. 12, 2021).

[24] "Q-score: Measure what truly matters", https://atos.net/en/solutions/q-score, (accessed Jan. 12, 2021).

[25] "TensorFlow Quantum is a library for hybrid quantum-classical machine learning"https://www.tensorflow.org/quantum, (accessed Jan. 12, 2021).

Christian Gamrat is a researcher at the Research and Technology Department of CEA (Alternative energies and Atomic Energy Commission), France.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: C. Gamrat. Towards operational quantum computing? Or: thinking beyond qubits. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 116-121, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

Information and communication technologies are the fuel of the digital society. This article reviews the big challenges to make ICT (more) circular.

Towards circular ICT: from materials to components

By THOMAS ERNST and JEAN-PIERRE RASKIN

Europe is facing one of the biggest challenges of its history: that of preserving a viable environment for the decades to come. On human consciousness of its environmental impact, Prof Bartlet says "The greatest shortcoming of the human race is our inability to understand the exponential function" [1]. ICT was, and still is, driven by exponential functions (Moore's law, Ops/Watt, data exchanges volume worldwide, …). But each exponential has its sustainability limits which are often reached sooner than expected. In ICT history to date, when those limits are reached, they are overcome by a technology breakthrough. By way of example, there was the CMOS (Complementary Metal-Oxide-Semiconductor) technology, a product of laboratory curiosity of the 1980s when bipolar transistor power consumption was too high. New integrated system (multi-core processor) and device (SOI, FinFET,…) architectures entered the market when energy consumption limits were reached by CPUs; CMOS is still today the much cheaper and reliably power efficient technology for the development and growth of the Internet of Things. Awareness of the rapid proliferation of wirelessly connected objects around us gives rise to questions and fears among the public. There is a need for more transparency on the environmental footprint of ICT and a need to establish an ambitious roadmap, at research and industrial levels, for moving into a virtuous cycle of eco-innovation.

In this article, we propose pathways to maintain or even reduce global (exponential) use of energy and materials for ICT.

Key insights

- ICT is energy-intensive and contributes approximately 4% of greenhouse gas emissions. Approximately half of the emissions from ICT comes from their production.
- ICT is based on a large number of critical materials and Europe imports nearly all of them.
- Recycling of ICT is technically difficult, consumes energy, generates pollution, and is not currently profitable because of the low price of primary raw materials. Only 10-15% of electronic waste is recycled.
- A product-centric economy favours the "produce, consume and waste" vision and thus leads to a planned obsolescence of digital objects and infrastructures. The service-centred model would be more in line with the principles of the circular economy.
- The rebound effect negates efficiency gains in ICT systems.
- It is time to design differently, to design within limits.

Key recommendations

- Integrate circular economy concepts, eco-design and full lifecycle assessment (including fabrication, use and end-of-first-life phases) at the early stage of research and development of new ICT technologies.
- Avoid toxic and substitute critical materials to protect environmental and human health as well as ease the recycling of materials at the end of the device or component life.
- Use secondary (recycled) materials from the ICT industry, for instance, the use of recycled cobalt from batteries.
- Develop economically viable and environmentally friendly recycling processes bespoke and dedicated to ICT.
- Develop bio-based materials and greener chemistry.
- Integrate all the externalities related to the development and the exploitation of a technology.
- Assess technology by a transdisciplinary panel throughout its lifecycle.
- Expand the lifetime of devices through better design, by both enhancing the intrinsic durability of components and adopting a modular approach in which replacement of faulty or obsolete components is made easy.
- Invest in additive fabrication processes for ultra-small circuit design instead of today's subtractive processes.
- Draft guidelines and legal policies to enforce supply chain transparency for all imported and non-imported materials and components.

An increasing impact of ICT growth on natural resources

Behind our screens, we have the impression of living in a dematerialized world, where everything is fast, clean, and reconfigurable. The reality is different. The digital society we live in has never been so energy- and material-intensive and this is leading to increasing pressure on natural resources, ecosystems and climate. Today, ICT is a growing economic activity, similarly to transport, energy production, manufacturing and agriculture. Information and communication technologies consume around 5% of the world's electricity production and are responsible for around 4% of greenhouse gases, a level equivalent to air transport. Our smartphone contains electronic circuits that require more than sixty different materials. We are talking about virtually every element on Mendeleev's periodic table except radioactive materials. At the end of their life, the recycling rate of electronic equipment is very low (less than 15%).

It is extremely difficult to separate the sixty materials that make up electronic circuits.

Umicore, one of the most advanced companies in the field of electronic materials recycling, manages to extract 17 elements out of the 60. Material recovery from obsolete equipment is not profitable given the low cost of raw materials imported from countries south of the equator. We do not pay for the environmental and social costs. This leads to a double penalty for the Southern countries: they suffer from environmental pollution (loss of biodiversity, pollution of the air and groundwater, etc.) during the extraction of raw materials as well as receive 75% of our electronic waste.

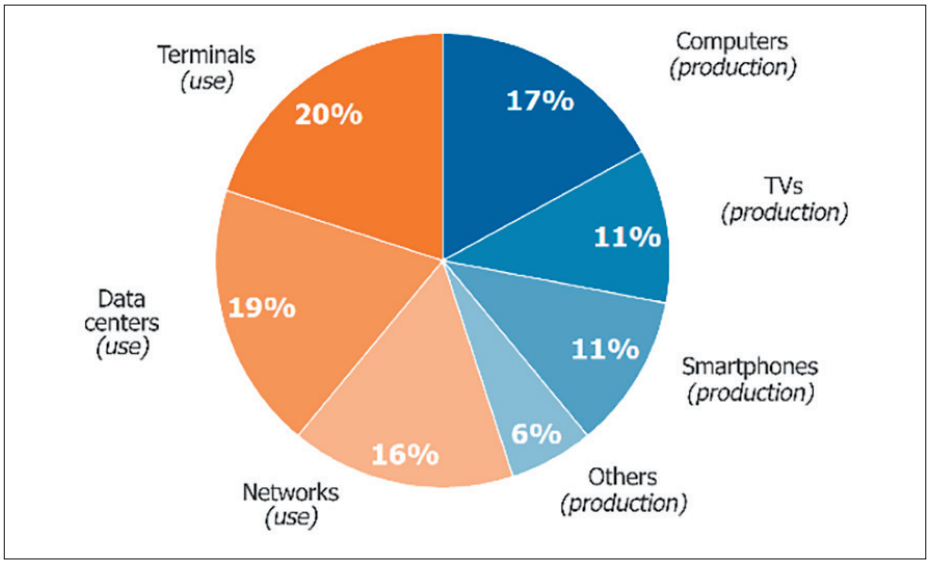


Figure 1: Estimation of distribution of the energy consumption of digital technologies for production (45%) and use (55%). (Source: the shift project 2018)

Toward a low waste production

Efforts are being made to reduce power consumption and industrial waste during the manufacturing of electronic devices and their components. For example, low temperature processes, reduction of heat dissipation in and from the oven, and reduction and recycling of chemicals and water cleaning, are evaluated for each step of the manufacturing of nanoelectronics devices. However, beyond resource efficiency and recycling waste, the transition towards a cost-effective circular economy needs to be implemented and the general design methodologies of materials for sustainable development proposed by Ashby [19] adapted to specific ICT domains such as nanoelectronics.

Some companies are beginning to adopt the lifecycle approach including, for example, in the production and recycling of the ultra-pure water needed [2] for the microelectronics industry. There are active research programmes, at the equipment makers level and in research labs, seeking to reuse exhausted gas or fluids within the fab [11], and also to develop much more efficient materials deposition techniques.

There is still room for improvement and a high potential for reduction in use of raw materials. While progress is continu-

ally being made towards greener chemistry, this approach is still far from being environmentally friendly and etches away a significant fraction of materials, as well as requiring very high consumption of water for rinsing. The reduction of toxic chemicals such as solvents in the microelectronics foundries must be a priority. Tomorrow, bio-based [12] chemicals and materials may be used to reduce significantly the use of solvents and chemicals in lithography, as shown in preliminary results [13]; more research efforts are required. Bio-based materials are also being investigated for use in packaging in the ICT domain.

More circularity between companies must be facilitated by the coherent installation of industrial activities in the same area, thus developing eco-systems in which the by-products or wastes from one industry could be the supply material for another.

For instance, the hafnium required in CMOS production is a by-product of ultrapure zirconium used by the nuclear industry and produced mainly in France and the United States [14].

To further reduce the waste of energy and materials, the ICT industry must adopt a holistic approach to developing sustainable products. Several initiatives already exist in the private sector. As an example, we can point out the significant and long-term efforts by ST-Microelectronics to evaluate (by lifecycle assessment) the carbon footprint of their microcontrollers [15] and to establish a clear material declaration availed online.

To envision a more sustainable future, Europe must:

- Take clear actions to make the ICT supply chain more transparent;
- Make lifecycle assessment and declarations of materials systematic (including for imported products) with shared methodology worldwide;
- Implement a clear and ambitious plan to maximize product lifetime and anticipate its end of life.

This will encourage both research and industry sectors to innovate for the best of all.

Zoom on minerals

The electronic industry needs a wide variety of minerals. For example, since the 1970s the silicon-based Complementary Metal-Oxide-Semiconductor (CMOS) Field Effect Transistor (FET) has been the mainstream technology for most transistor applications, thus making possible today's digital economy. Over the years, the number of elements exploited in their manufacture has increased greatly (Figure 2), especially since 2000 with the implementation of high-k dielectrics and metal gate stacks which are essential to minimize short channel effects and gate leakage current of short transistors (today gate length shorter than 20 nm).

A growing awareness of the limited nature of the supplies of some elements that have specialized and important uses is reflected in the proliferation of terms to describe them and the ores from which they are derived, including 'gateway minerals' and 'critical' and 'endangered' elements. Some countries have adopted policies recognizing the high strategic importance of some of these for their physical and economic security.

In 2010, 14 elements were considered as critical by the European Commission (EC) according for both their strategic importance for future technology and their scarcity, while in 2017 the number rose to 17, including metals (such as tungsten and some rare earth elements), and also other raw materials such as phosphates, natural graphite and magnesite[3].

Modern devices and systems rely heavily on a high degree of control of material properties and a mastery of manufacturing techniques and, to date, the ICT industry has been remarkably successful in fulfilling these needs.

The manufacturing process is a 'top-down' or 'subtractive' one based on UV photolithography, etching and many sequential, highly organized and efficient steps of chemical and physical treatment of the chip, layer after layer.

Although silicon is abundant on Earth, the sand used to produce high-purity silicon is already being viewed as a scarce commodity due to its multiple additional uses in concrete, asphalt and glass.

Work is being undertaken on substitution or decreased use of toxic, hazardous and critical raw materials. For currently crucial elements such as indium, ruthenium, platinum, gallium, arsenic and gold, new technologies and materials are being

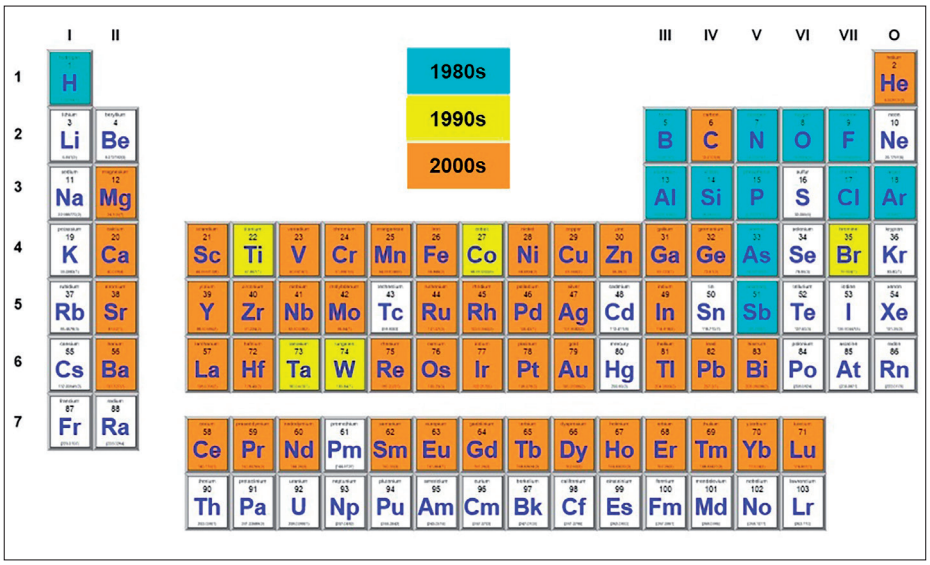
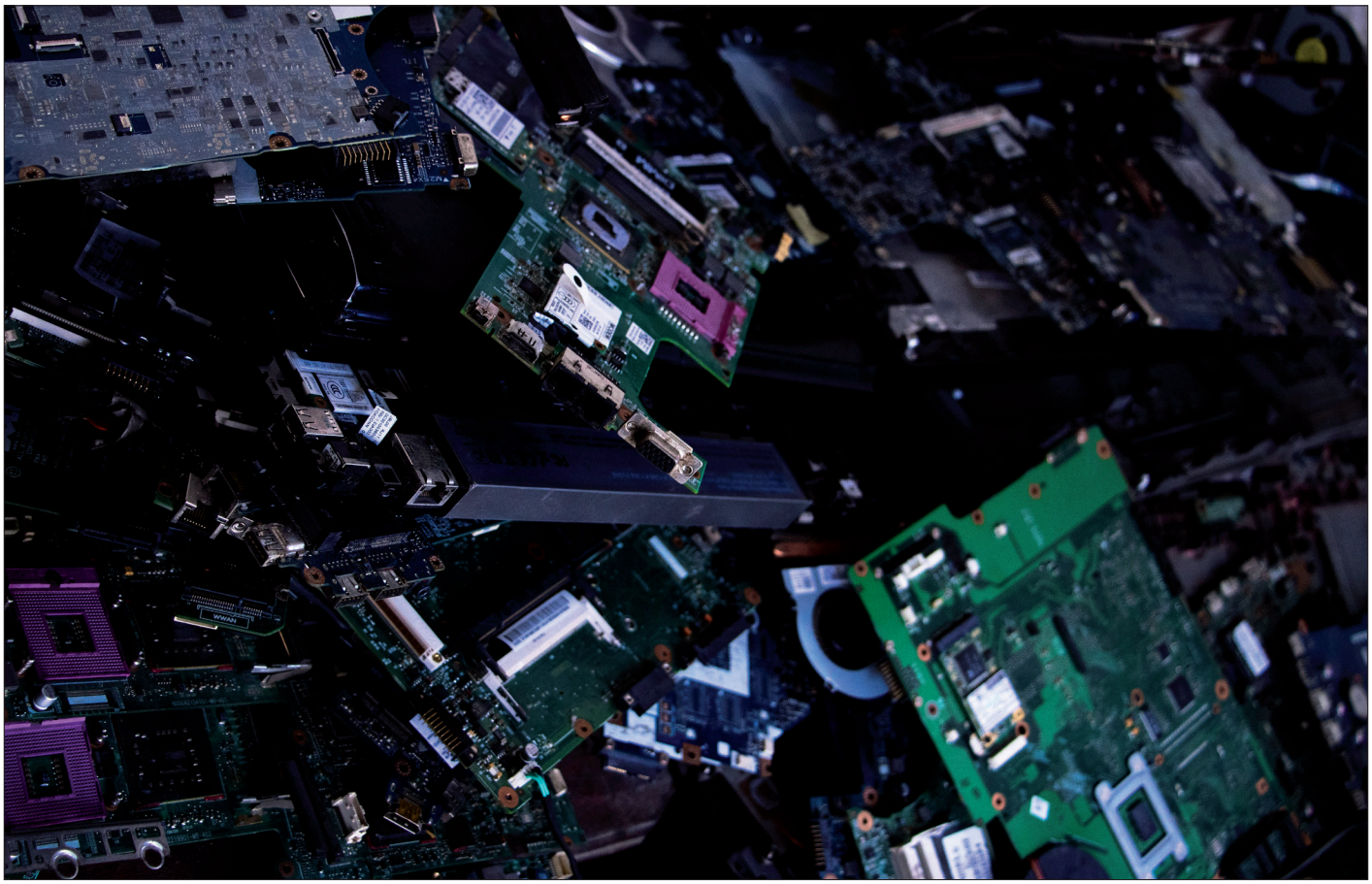


Figure 2. Introduction of elements in the manufacture of CMOS transistors: Complementary Metal-Oxide-Semiconductor (CMOS) transistors mainly involved silicon, oxygen, boron, phosphorus and integrated circuit interconnections were made of aluminium in the 1980s. There were relatively few changes in the 1990s, but a large diversity of elements was introduced in early 2000 and many IC interconnections were switched to copper.



investigated with a view to replacing them or drastically limiting their use in some critical devices (e.g. in sensors, memories, optoelectronics and spintronics).

Other examples of the move towards sustainable electronics include the avoidance of lead in micro-components like actuators included in cell phones, use of 2D mono-atomic or ultra-thin atomic-deposition layers to reduce the use of some active materials by a factor of up to 106, and use of silicon-based substrates such as Silicon-on-Insulator (SOI), instead of materials made from combinations of Group III and Group V elements, for radiofrequency (RF) technologies.

The scaling down of dimensions of high-tech devices in recent decades and the multiplication of materials in the components – some of them in extremely small quantities of a few micrograms – lead to new challenges in recycling. The need for large amounts of power and the use of aggressive acids and solvents can make recycling of such electronics impractical.

Approaches to increasing the sustainability of microelectronic devices must include expanding their lifetimes through better design, by both enhancing the intrinsic durability of components and adopting a modular approach in which replacement of faulty or obsolete components is made easy.

These approaches can draw on the experience of, for example, some European microelectronics manufacturers and R&D laboratories (e.g. On-Semi, X-FAB, Infineon, ST-Microelectronics, NXP) that are designing or fabricating highly reliable components for automotive, energy management and security applications.

Challenging the 'top-down' or subtractive approach, 'bottom-up', additive fabrication processes may offer a new paradigm for ultra-small circuit design. This could involve 3D printing and consist in layer-on-layer depositing of 2D materials such as graphene, molybdenum sulphide and hexagonal boron nitride [4]. Nanomaterials offer potential alternatives to the widely used indium tin oxide transparent conducting layers. Proposed alternatives include a transparent plastic sheet on which nanocarbons have been deposited, or which contains a loose conducting nanomesh [5]. It is expected that technology based on nanostructures will require much fewer raw materials that any traditional approach. However, recycling nanomaterials at the end of the life of the device will bring new, challenging problems [6].

Steps towards the ambitious goal of achieving sustainability of the material basis of the digital society will require concerted action covering a range of interlocked approaches. These will address the entire lifecycle of not only the digital devices themselves but also the services that support them, paying attention to their energy and environmental footprints as well as economy and efficiency in the utilization of resources.

Research on substitution of critical materials with more abundant ones or which leads to breakthroughs to reduce significantly their use, for instance by localising active atoms where needed, will open up the path to more responsible electronics.

Waste hierarchies and multiple Rs

The concept of waste hierarchies has become a central feature of the 3R (reduce, reuse, recycle) and circular economy approaches, in which materials recovered across the lifecycle are re-applied at the highest possible level of utility and with minimization of waste or environmental damage. The potential of the approach is evidenced by the virtual disappearance of waste landfilling in, for example, the Netherlands [7].

However, the need to take a comprehensive view of the entire global system of recycling is illustrated by the case of microelectronic devices. Around 50 million tonnes of electronic waste, or e-waste, is being thrown away each year. A large amount of e-waste (40% of that produced by Canada, the EU and the United States) is shipped in containers to countries in Africa and Asia, where people manually dismantle the appliances and burn them in the open, producing dangerous levels of hazardous substances which severely impact their health.

Over the last few decades, the number of Rs has grown and at least eleven have been cited in relation to the operation of circular chemistry. Their relevance to the challenge of sustainability of the material basis of the digital society is outlined in Table 1.

Table 1: Levels of resource hierarchy and relevance to sustainability of the digital society

Level	Level definition	Explanation and application to sustainability of the digital society
1	Reject	Rejecting use of a material or process, whether to conserve scarce resources, reduce energy demand or avoid serious pollution, is considered the highest hierarchy level. A broad approach to sustainability will require questions such as: "do we really want this component, product or process at all?" and "can this substance be replaced by one that is more available, less toxic, or more readily recyclable?" • EU directives restrict the use of harmful substances, such as lead in solders used in microelectronics circuitry. • It has been shown to be possible to replace rare-earth metals used in electronic devices with combinations of more common elements [8]. • Harmful organic solvents can be substituted with more eco-friendly ones [9].
2	Reduce	Overall reductions in quantities of materials and energy consumed can be achieved with a combination of strategies including product simplification (e.g. fewer materials), miniaturization of the constitutive transistors, interconnects, electrical components, improved battery performance and lower energy consumption, increased product life span and reduced distance for transport. • 3D printing and other additive processes are being introduced in microelectronics to lower consumption of raw materials and chemicals for etching. • Graphene is being explored [16] as a replacement for transparent conductive films such as indium tin oxide widely used in touchscreen applications. • Efforts have been made to replace or at least to reduce the amounts of 'III-V' materials such as InP, GaAs, GaN with high carrier mobility, which are used in high frequency devices. Thanks to the Silicon-on-Insulator technology introduced to the market in 2012, more than 95% of RF switches used in mobile devices are now silicon- and not III-V-based. Furthermore, III-V localisation by local wafer bonding (for instance CEA-Leti [17]) or local epitaxial growth (for instance IMEC [18]), only where this material is needed, drastically reduces the amount used.
3	Reuse	Reuse enables an object to remain in service for a long period of time. The library is a very widely known example. • Second-hand shops are a familiar way of enabling reuse of items for their original purpose, including laptops, tablets, smartphones, and many other electronics appliances. • Leasing of ICT equipment (e.g. smartphones provided on a use/return basis as part of a mobile network contract) can also be a channel for reuse, depending on the approach of the leasing company. For example, nearly all the internet provider companies do not sell their WiFi box to clients but rent them. The company stays responsible for the box. They take care of its maintenance and upgrade. Thanks to this move to the economics model of functionality instead of ownership, the lifetime of each electronics component is greatly extended.
4	Redistribute	Redistribution is also concerned with reusing resources, but involves transport over longer distances – for example, to bring a second-hand product to a new market. • Rates of mobile phone usage have risen very steeply in Africa in recent years, with Nigeria providing the biggest African market for second-hand smartphones.

5	Repair	Repairing and maintenance are generally the least resource-intensive solutions to extend the life of devices that are damaged or cease to perform effectively. Unfortunately, however, the way that most ICT devices are manufactured does not facilitate their repair. Moreover, access to spare parts is often impossible or only guaranteed for a short period of time. Once a device needs repair, critical factors are: a good diagnosis of the problem and easily removable elements. There is growing interest from both consumers and some producers in repair of ICT devices. • Citizens' initiatives such as 'repair cafés' flourish in many countries, providing locations where specialist tools are available and expert advice may be offered by volunteers. Thanks to the internet there are also several websites which bring people together to form a community for helping each other repair things, for example iFixit. • Consumer pressure for the redesign of electronic products to promote repair rather than disposal of broken items has been growing and has been reflected in the EU's circular economy approach. • Fairphone, the world's first ethical modular smartphone on the market, has been designed for a high level of reparability. Most of the spare parts can be ordered online and the product webpage has many tutorials to assist consumers to repair or upgrade their device.
6	Refurbish	For electronic materials, refurbishment means combining repaired and redistributed products. It may also involve updating a product to current standards. Availability of components and accessibility for repair and replacement are critical. • Reconditioned cellular phones are sometimes returned to the market by companies that initially produced them and recovered them after a period of use. In 2019, Fairphone launched its scheme "Refurbished phones give valuable resources a new life". These phones have the same high-quality standards as a brand new Fairphone and the same two-year warranty — but a lower price (half the price of a brand new one).
7	Repurpose	Repurposing is the updating or adaptation of a product such that it can be used to serve a new function or within another context. • There are initiatives such as Puzzlephone from which, at the end of a smartphone life, parts can be used in a new context or application such as in a computer. The repurposing of an object will be favoured if retrieval of its functional parts is possible.
8	Remanufacture	Remanufacture is generally a more thorough process of disassembling a product, replacing worn and broken parts with new ones and reassembling it. • An assessment of remanufacturing of end-of-life computers identified potential to enhance resource conservation and prevent natural resource degradation, and that remanufactured computers could be technically, environmentally, economically and socially feasible if there is an adequate supply of quality cores, involvement of highly skilled workers, incorporation of a standardization process, and the use of advanced machines tools.
9	Recycle	As well as the simple reuse of materials, recycling also encompasses the recovery of component materials to act as a feedstock for new material or device production. • Electronic scrap is a highly complex and heterogeneous group of materials. Only a few precious elements present in integrated circuits are recovered by current recycling techniques because they require a large amount of energy and the use of highly polluting chemicals, while the freshly mined minerals are quite cheap. However, Umicore refines e-scrap containing precious metals such as silver, gold, palladium and copper. • The problem of flame-retardant additives needs to be solved in order to recover plastic components of e-waste. • Better technology is needed to improve the purity of recycled metals such as noble metals used in digital equipment. • More sustainable chemical and biological processes may be applicable to both synthesis and recycling of organic components such as polymers.
10	Recover	Recovery involves retrieving the lowest forms of energy or feedstock for energy production from a material. The actual material is broken down and cannot be recycled further. • There is a huge number of waste-to-energy plants in the world. Burning waste and particularly e-waste produces, as well as CO_2, toxic smoke and fine particles which are difficult to prevent from being released into the atmosphere. The efficiency of these waste-to-energy plants and their impacts on human health and the environment remain challenging. • Production of hydrogen as a fuel from printed circuit boards by steam gasification has been examined [10].
11	Return	Most hierarchies call this landfill. However, this stage is more than that – it encompasses the return (sometimes in chemically modified form) not only of solids, but also liquids and gases, back to the environment after use. • More than 70% of e-waste makes its way into landfill, leading to potentially dangerous contamination of groundwater.

Image: ID 25538663 | ©Dwfotos | Dreamstime.com

It is emphasized that attention to the Rs alone will not solve the problems of waste and resource depletion in the next couple of decades, but will help to extend the timeframe of a transition period during which optimal solutions to sustainable sourcing, production, use and recycling can be developed. In terms of the mining of the currently required minerals, diversified approaches must be developed during this transition period. These include at least three complementary avenues:

1. Develop further mining of current mine waste in order to use the maximum amount possible of the extracted rocks, minerals and chemical elements. This will require the best knowledge of each deposit, information/teaching to mining companies and, where possible, new regulations.
2. Encourage refining of by-products in the mining of main deposits of ores (i.e. low-abundance chemical elements associated with a much more abundant one). For example, potentially valuable levels of gallium and scandium are associated with ores of aluminium. Similarly, useful quantities of copper, antimony, silver, arsenic and tellurium may be recovered from gold deposits. However, at the same time, there is a need to develop independent supplies of technologically important minerals, beyond their 'companionality' as a by-product of one or more host metals from geologic ores, to ensure reliability of supplies.
3. Promote tolerance of new, responsible mining of metals in countries or regions which have become averse to mining. This will require education, with frank and balanced information provision and discussion that engages civil society, scientists and engineers, government and the media. The discussions must include assessment of the true price of raw materials, including the ethical and environmental prices (which is not the case at the moment), and would lead to increased costs of raw material in the coming years, with concomitant increased costs of the associated products. Taking account of the "true" price would, in turn, help engender a redefining and reordering of priorities for production and consumption.

Acknowledgement

The present article summarizes the main discussion outcomes from a panel of experts in the fields of geology, materials science, micro and nanoelectronics fabrication process, electronics circuits design, electronics packaging, supply chain management, and IT systems: Mathilde Billaud (Fraunhofer), David Bol (UCLouvain), Thierry Baron (CNRS), Patrice Christmann (BRGM), Marie Garcia-Bardon (imec), Tapani Jokinen (Fraunhofer), Francois Martin (CEA-Leti), Bertrand Parvais (imec), Karine Samuel (UGA), Lutz Stobbe (Fraunhofer), Olivier Vergeynst (Green IT).

References

[1] Al Bartlett, https://www.albartlett.org/

[2] W. Den, C.-H. Chen, Y.-C. Luo. "Revisiting the water-use efficiency performance for microelectronics manufacturing facilities: Using Taiwan's Science Parks as a case study". Water-Energy Nexus 1 (2018) 116-133. https://doi.org/10.1016/j.wen.2018.12.002.

[3] "List of Critical Raw Materials for the EU". Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the 2017 European Commission, Brussels, COM/2017/0490 final, 2017. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52017DC0490 (accessed 29 August 2020).

[4] X. Ling, Y. Lin, Q. Ma, Z. Wang, Y. Song, L. Yu, S. Huang, W. Fang, X. Zhang, A.L. Hsu, Y. Bie, Y. Lee, Y. Zhu, L. Wu, J. Li, P. Jarillo-Herrero, M. Dresselhaus, T. Palacios, J. Kong. "Parallel stitching of 2D materials". Adv. Mater. 28 (2016) 2322-2329. https://doi.org/10.1002/adma.201505070.

[5] A. Khan, S. Lee, T. Jang, Z. Xiong, C. Zhang, J. Tang, L.J. Guo, W.D. Li. "High-performance flexible transparent electrode with an embedded metal mesh fabricated by cost-effective solution process". Small 12(22) (2016) 3021-3030. https://doi.org/10.1002/smll.201600309.

[6] S.A. Younis, E.M. El-Fawal, "P. Serp. Nano-wastes and the environment: Potential challenges and opportunities of nano-waste management paradigm for greener nanotechnologies", in: C. Hussain (Ed.), Handbook of Environmental Materials Management. Springer, Cham, 2018. https://doi.org/10.1007/978-3-319-58538-3_53-1.

[7] A. Lansink. "Challenging changes – Connecting waste hierarchy and circular economy". LEA, Nijmegen ISBN/EAN 978-90-821783-5-7, 2017. https://www.challengingchanges.org/the-book/ (accessed 29 August 2020).

[8] M. Irving. "Common element combos could replace rare-Earth metals in electronics". New Atlas 5 July 2019. https://newatlas.com/common-elements-replace-rare-earth-metals-electronics/60447/ (accessed 29 August 2020).

[9] F. Pena-Pereira, A. Kloskowski, J. Namiešnik. "Perspectives on the replacement of harmful organic solvents in analytical methodologies: a framework toward the implementation of a generation of eco-friendly alternatives". Green Chemistry 17 (2015) 3687-3705. https://doi.org/10.1039/C5GC00611B.

[10] J.A. Salbidegoitia, E.G. Fuentes-Ordóñez, M.P. González-Marcos, J.R. González-Velasco, T.K. Bhaskar. "Steam gasification of printed circuit board from e-waste: Effect of coexisting nickel to hydrogen production". Fuel Processing Technol. 133 (2015) 69-74. https://doi.org/10.1016/j.fuproc.2015.01.006.

[11] "CVD-semens monosilane reactor process with complete utilization of feed gases and total recycle", patent US 8,657,958 B2, 2014

[12] M. Caillau et al, "Sub-micron lines patterning into silica using water developable chitosan bioresist films for eco-friendly positive tone e-beam and UV lithography", SPIE Advanced Lithography, 2018, San Jose, California

[13] Mathieu Caillau, Pierre Crémillieu, Emmanuelle Laurenceau, Yann Chevolot and Jean-Louis Leclercq, "Fifty nanometer lines patterned into silica using water developable chitosan bioresist and electron beam lithography", https://doi.org/10.1116/1.4996870

[14] Framatome, "Fuel Business Unit Jarrie", https://www.framatome.com/EN/businessnews-142/framatome-fuel-business-unit--jarrie.html

[15] ST, "Footprint of a Microcontroller", https://www.st.com/content/st_com/en/about/st_approach_to_sustainability/sustainability-priorities/sustainable-tec hnology/eco-design/footprint-of-a-microcontroller.html

[16] Dexter Johnston, "The Market for Nanomaterial Solutions for ITO Replacement Gets Crowded", https://spectrum.ieee.org/nanoclast/semiconductors/nanotechnology/nanomaterial-solutions-for-ito-replacement-gets-crowded

[17] Bertrand Szelag, Karim Hassan, Laetitia Adelmini, Elodie Ghegin, Philippe Rodriguez, Fabrice Nemouchi, Pierre Brianceau, Elisa Vermande, Antoine Schembri, David Carrara, Pierrick Cavalié, Florent Franchin, Christophe Jany, Segolene Olivier, "Hybrid III–V/Silicon Technology for Laser Integration on a 200-mm Fully CMOS-Compatible Silicon Photonics Platform", https://doi.org/10.1109/JSTQE.2019.2904445

[18] Alan Y. Liu, John Bowers, "Photonic Integration With Epitaxial III–V on Silicon" https://doi.org/10.1109/JSTQE.2018.2854542

[19] Ashby, "M. Materials and Sustainable Development". Butterworth-Heinemann (2015). https://www.elsevier.com/books/materials-and-sustainable-development/ashby/978-0-08-100176-9 (accessed 29 August 2020).

Thomas Ernst is Scientific Director at Leti, CEA tech, France

Jean-Pierre Raskin is Professor at the Université catholique de Louvain, Belgium

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: T. Ernst and J-P. Raskin. Towards circular ICT: from materials to components. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 122-129, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

The impact of computing on all aspects of life is tremendous, and artificial intelligence will have an even bigger impact. We can no longer imagine a life without computing. As usual, there are positive and negative effects.

AI for a better society

By KOEN DE BOSSCHERE

The impact of computing on society is big, and all-pervasive. Computing has both positive impacts (easier access to information, more transparency, increased productivity, …) and negative impacts (growing inequality, fake news, privacy erosion, …). The challenge is to maximize the benefits while mitigating the negative consequences. In any case, artificial intelligence (AI) will lead to change, and change always causes resistance, but we cannot (in a competitive world) turn back the clock and go back to a time when there was no internet and no AI. The global challenges of the 21st century will not be solved without the help of artificial intelligence because the problems are so complex that they cannot be solved without advanced computing. We cannot make a world of ten billion people sustainable without AI. Eventually we will have to evolve towards responsible AI.

Key insights

- Algorithms are smarter than people and today they know more about us than we know about ourselves. Their negative aspects are often a reflection of our own. We have to protect society from the negative effects such as fake news, data leaks and privacy erosion.
- Computing accelerates societal change and this creates resistance. There is, however, no way back in an open and competitive world. The only way forward is to make sure that nobody is left behind, that the systems will cope with our ethical requirements and that everybody can benefit from the changes.
- AI and the internet consume a lot of power, but also help us save energy by optimizing processes.

Key recommendations

- We cannot make a world of ten billion people sustainable without advanced computing to limit the ecological footprint of such a population. In order to save the planet, we will have to invest more than ever in computing and artificial intelligence.
- AI can be used for good and for bad. We should evolve towards responsible and ethical AI, which means that the public and the private sector agree that we should only use it for the betterment of society. This implies that they should move away from purely economic criteria when making decisions.
- Computing systems should be made loyal to their users, not to the companies that provide devices and services.
- Computing system could also protect and help users by advising and informing them about the data (both inbound and outbound) that they exchange.

When Google was founded in September 1998, and the PageRank algorithm hit the world, people fell immediately in love with the search engine that seemed to be able to guess what a user was looking for. Surprisingly, the only thing it needed was a single search box, no long list of search options, check boxes, etc. Furthermore, it never disappointed the user. If somebody entered a URL instead of a keyword, Google just displayed the webpage. If a mathematical expression was entered, Google evaluated it. It corrected spelling errors, it automatically converted currencies and units, it also checked for synonyms. Such was the powerfulness of its offering, it immediately made other indexing websites obsolete. Google became the access point for the internet. Today, we expect search engines to read our mind, and immediately show what we are looking for, be it the closest restaurant, the cheapest online shop selling a particular product, driving directions... you name it, Google finds it.

With artificial intelligence, big data analytics, deep learning, and huge computing resources, platforms like Google became almost omniscient and able to serve us almost exactly the information we wanted. The younger generations cannot imagine how much energy it took in the 20th century to find reliable and recent information. Information in books and encyclopaedias was basically obsolete shortly after they were printed. Today, even small children can find the information they need and it has made some skills like searching in an alphabetically ordered list almost obsolete. It is fair to say that search engines have completely changed the way in which we deal with information and, in the process, they have made information available to all, and made society more transparent.

The technology that makes search engines so powerful has been adopted by social media platforms to show the relevant messages on personalized timelines, by news agencies to compile a personalized digest of the latest news, by dating apps to show matches one might be interested in, by streaming platforms to show the content somebody might like (the recommender system algorithms). And on top of all this, most of these websites deliver all these services for free on condition that they can show us some adverts. But that seems harmless because printed newspapers have adverts too, don't they? Not exactly.

How users became the product

Few people fully understand internet companies' business models. Facebook is a free platform with around 2.5 billion active users. In 2019, its revenue was 70 billion USD – an average of 28 USD per user. So, that is the average value in 2019 of the seemingly worthless information we share on our Facebook accounts. Facebook's real customers are the companies and organizations paying for marketing campaigns. The goal of a marketing campaign is to change the behaviour of the target group (for example by convincing them to buy a particular product, to sign up for a service or to vote for a political party).

For companies like Facebook or Google, the users are the product, and as any other company, Facebook and Google try to optimize their product (i.e. us!) to the needs of their customers. The perfect product is a user who spends a lot of time on the platform and reacts in ways intended by the (paying) customers (i.e. buying goods and services, voting [16] and so on). The more information the platform has about its users (the queries we enter, the links we follow, the pages we spend time on), the more targeted and the more effective the marketing campaigns can be made, and the more the platform can charge for them. The longer a user spends on the platform, the more advertisements can be shown, and the bigger the revenue will be. The more features the platform offers (face recognition, language translation, video, games, and so on), the more the users will enjoy the platform, the more time they will spend on it, and the more frequently they will return. There are good reasons why Google goes to great lengths to offer a wide variety of services. They want to be a one stop shop.

Attention is a valuable resource

There is an arms race between (social) media companies for the attention of the user. Unfortunately for them, a user cannot spend more than 24 hours a day in front of a screen. All these companies are thus competing against each other to get a greater share of users' attention. Platforms deliberately use mechanisms to make them addictive, or at least habitual. These include likes, automatic notifications, clickbait and scoring. This has been called brain hacking [1].

Addicted users come back frequently, which translates into higher revenue. Finally, the number of users has to grow fast for start-up internet companies and this influences the content. On one hand, platforms try to ensure that nobody will be offended by content on the platform, so they censor all content that might be considered inappropriate to valuable groups of users. Censoring is tricky as it starts from a world view of what is acceptable and what is not, especially when it comes to political statements, religious views or sex. On the other hand, viral (including outrageous) content is welcomed because it means that more people spend more time on the platform, and hence generate extra revenue. These platforms have also become (unintended) instruments to promote the values of the large user groups (e.g. American or Chinese values).

And they are successful in gaining user attention: in the younger generations, social media has almost completely outcompeted traditional media like television and newspapers [2]. In their competition for more attention, social media platforms are also monopolizing people's time, in both their professional and private lives. Active professionals believe they have to have a presence on social media, and to amass large numbers of followers. This leads to loss of productivity and mental absence at meetings, etc. In many people's private lives, screens have replaced face-to-face interactions at home (especially in 2020), at the dining table, at the pub, in restaurants and on public transportation. This leads to a phenomenon known as "phubbing", or phone snubbing: checking your smartphone during social events instead of giving your full attention to the people who are physically there [3].

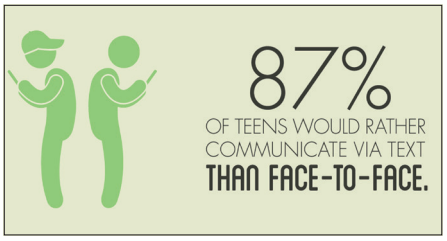


Figure 1: Phubbing (Source: stopphubbing.com)

The final frontier is competing with sleep. Studies show that millions of people suffer from sleep deprivation resulting from excessive use of smartphones and tablets [4].

Social media create echo chambers

What also sets social media apart from traditional media is that traditional media broadcast their messages publicly so that everybody can receive them and, ideally, learn about the arguments of a range of stakeholders by watching their channels. In contrast, the combination of advanced big data analytics and significant computing power hosted in large data centres has enabled social media platforms to create a personalized experience for each individual user. That means that every user gets to see a different stream of messages and that users cannot see the message streams of other users. Users can share messages in their own network, but since networks tend to be clustered, users tend to see more of the same messages rather than different points of view.

In so doing, social networks create information silos or filter bubbles and act as echo chambers which reinforce the values of the members of the network. "Wrong" posts will not garner a large number of "likes" and will quickly disappear from timelines. Hence, it is very difficult for information in one information silo to make it into another. The following figure illustrates three different communities living in Israel: pro-Palestinian, pro-Israel and religious/Muslim. There are very few links between the pro-Palestinian and pro-Israel communities. Most links are shared via the religious/Muslim community. There is little chance that messages from the pro-Israel network will ever make it into the pro-Palestinian network and vice versa.

What is worrying is that a handful of private global companies and their proprietary algorithms decide who gets to see what, in which order, and when. They can even gradually modify the user's preferences by proposing only a limited set of items and removing items that are old, in low demand or not in accordance with the ideas of the providers, for example. In the past, opinion-shaping messages came in hard copies, which were harder to remove – it was necessary to physically find them in people's house and burn them, as in *Fahrenheit 451* – than digital media on private servers and streamed to people who are not using local backups.

All this means that social media companies are in a sense helping to create a worldview per user, formed by purely business decisions – i.e. decisions that will optimize the profitability of the company – mostly unregulated by governments. The fact that traditional media such as newspapers and television news have declined in popularity among "digital natives" strengthens the impact of social media on the world view of young people. This explains to a certain extent why traditional media outlets anticipated neither Brexit nor the election of Donald Trump. They were simply unaware of messages shared in circles they did not belong to [5]. The fact that a significant number of American Trump supporters, conservatives and right-wing extremists recently moved to the Parler social networking platform is a sign that they are not interested in anything but their own messages.

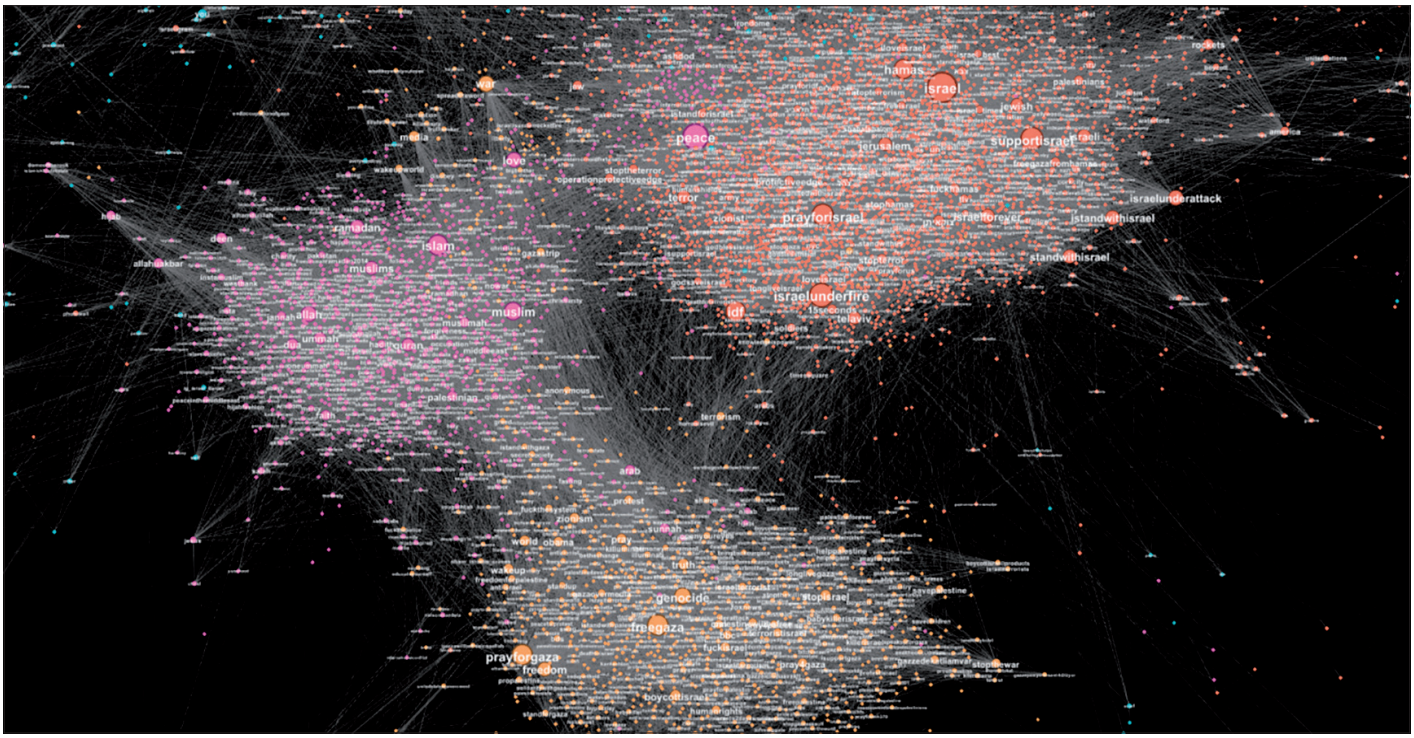


Figure 2: Israel, Gaza, War & Data – Social networks and the art of personalizing propaganda (Source: [Gilad Lotan, "Israel, Gaza, War & Data: Social networks and the art of personalizing propaganda", 2014])

The internet leads to privacy erosion

There are multiple definitions of privacy. In the 19th century, privacy was defined as the "right to be left alone". A more modern definition is that privacy is the "control one has over the information about oneself". It is necessary for doctors to maintain medical records about their patients, but nobody expects the doctor to share this information with third parties (medical privacy) unless this were to be required for medical treatment. We expect the same behaviour from financial institutions (financial privacy), websites (internet privacy) and voting systems (political privacy). We do not expect an email service to use the content of our messages to influence the advertisements we see on websites, or a booking website to use the type of rental car we prefer to result in us seeing advertisements for that particular type of car.

Gathering information about users is crucial to the business model of internet companies. That is why many websites nudge users to complete their profiles, thereby collecting additional monetizable information. Some companies, like the now notorious Cambridge Analytica, have made a business model out of collecting information, analyzing it, and selling it to whoever is willing to pay for it.

Many people are largely unaware about the cost of convenience in terms of lost privacy; or if they are aware, they are willing to give up some of their privacy in return for convenience:

- Booking websites collect details about every single trip their users book. This is crucial marketing information for hotels, airlines and rental car companies.
- Streaming music applications have data on when and where users listen to music, as well as what their musical preferences are. The better streaming music providers can profile their users, the better suggestions they can make and the more frequently and longer people use the service.
- Companies selling e-books know the identity of every single reader of a book, when they are reading a book, which parts they actually read and so on. In a sense, they know what a buyer learned from the books they bought. The more they know, the better suggestions they can make; it is not difficult to guess the interests of somebody buying books on classic cars, cookery, political history, or travel guides, for example. By (not) making particular suggestions, they can even steer what their users read and even think.
- Social media networks monitor all the private details users share with their most intimate friends, and use this data to infer information (for example, that the person feels depressed), in order to send them targeted advertisements they know work well (such as make-up or medication for those suffering from depression). Their aim is not to help people, but to sell and to influence. The people in social media control rooms are not medical staff; they do not have to comply with professional codes and they do not care about whether the advertised drugs are effective or safe.
- News websites track which articles users read, and adapt their content offering (news and advertisements) to their interests. They basically decide what their users will read, which might lead to a biased perception of the world. In the US, Democrats and Republicans live in two different news universes, leading to mutual demonization of the other side.
- Satellite navigation systems detect where the navigation system (and, by extension, probably its owner too) is at any time. It is comparable to being shadowed by somebody wherever you go.
- Voice-controlled devices keep track of what goes on in a house or office, and they can be hacked to eavesdrop on conversations. Few people would appreciate a stranger sitting in their house all the time.

In addition to the examples above, people are already under surveillance for a large part of the day, through access control systems in companies and hotels, numerous cameras in public places, licence-plate recognition, Google Street View filming the street, tourists taking pictures with people in the background and posting them on social media, and so on. Most people do not protest about this surveillance because they believe that it helps the government to enhance their safety and prevent terrorist attacks. Surveillance of people is an old practice, but it was limited to selected individuals (for spying, criminals, etc), but algorithms and computing systems now allow for mass surveillance at a low cost.

Irrespective of the application, the fact is that (i) all our actions in *cyber space*, and an increasing number of actions in physical space are being recorded and stored in huge databases; (ii) that an increasing number of such databases are being linked (often through acquisition, or by linking government databases to facilitate e-government); and (iii) that that there is no guarantee that this data is only used for the purpose it was collected for. As these data collections are grown organically, where independent individuals decided to include/exclude particular information, the databases can be biased and, if used by algorithms, can come up with biased conclusions.

People are also often not really aware of the impact of the information they share; for example, a picture taken in a bar could mark you as a sociable person, but also perhaps as a drinker, which may interest your health insurance company. Computing systems could help inform people of the risks of their data exchanges.

It is clear that there is an urgent need for increased regulation. Privacy should be better protected, and there should also be more guarantees for unbiased database contents used for machine learning. People also deserve the right to be forgotten.

Fake information is part of the DNA of the internet

Whereas traditional media have built-in filters that require journalists to verify their sources, there is no such thing in social media. Anybody can post anything, and as soon as it passes social media companies' decency filters, it becomes public. The social media reviewers censor particular content (child abuse, sexual content, hate speech, …) but not fake information. The higher the number of people reading and liking the fake information, the better it is for the business results of the platform. In response to public concern over the spread of fake news and hate speech on social media, major companies such as Facebook have employed editors to monitor the content, and take it down, or add

a warning. Several baseless claims on election fraud by Donald Trump were labelled as such by social media in 2020. This is however not done with the same scrutiny for all other messages, and it can only be done if factually wrong. In other cases, it is often just a question of opinions, which cannot be kept out of social media.

Over the last few years, there has been a surge in false or misleading information such as fake news, fake science and deep fake videos. Fake information is information that is presented as a reliable piece of information, but is either completely made up or highly misleading. Such messages are like hoaxes on steroids. Popular genres are the launch of conspiracy theories (e.g. the widespread QAnon conspiracy in the United States), and the spreading of pseudo-science (such as the dangers of vaccination). The motives of people spreading such information range from making money (mostly from advertisements alongside stories that go viral) to political objectives (influencing elections, creating unrest, destabilizing societies).

The most recent technical evolution of fake information is the so-called deep fakes, a successful application of face swapping technology to video. Originally designed to put the face of celebrities on pornography actors in action, the technology has been used to create credible fake interviews [6]. For the naïve viewer, these interviews are hard to distinguish from the real thing and can thus be misleading, as well as a misrepresentation of the views or ideas of the 'interviewee'. Deep fake also applies to real-time voice substitution, leading one to believe that he or she is speaking to a known person.

Figure 3: How to spot fake news (Source: IFLA)

The algorithms are smarter than people

The effects of digital technology on humans has been studied extensively, and there are both positive and negative effects. Customers have access to online information, they can make online appointments and buy goods and services without having to queue, physical meetings can be replaced by virtual meetings, collaboration tools allow people to work together efficiently and form the basis of the paperless office. On a personal level, it is now easier to keep in touch with friends and family members via social media. Many disabled and older people can also participate in social networks because their participation is not constrained by their limited mobility; this, in turn, helps them maintain or develop cognitive abilities. Thanks to video conference software, companies, governments and schools could continue (some of) their activities online during COVID-19 restrictions.

However, there are also some side effects. In some cases, people have become dependent on their smartphones. The smartphone does to the brain what using a lift, rather than the stairs, does to the body. Rather than memorizing information, people constantly refer to the internet, which can lead to digital amnesia [7]. Skills like mental arithmetic, memorizing numbers (mathematical constants, phone numbers), searching in a sorted list, and driving without a navigation system are disappearing in young people.

Perhaps even more disturbing is the fact that the web is full of texts that fit on just one or two screens, and that this has been linked to losing the ability of "deep reading", that is to say, the ability to focus on

a long text for an extended period of time. Research suggests that the disappearance of this skill, which is needed to read a book or to study, [8] can lead to lower academic performance.

Information technology has made sharing information so easy and cheap that it has become endemic. Many modern workers receive hundreds of messages per day; reading and responding to these messages takes up a significant part of their time, without being explicitly mentioned in their job description. Processing emails has become a struggle, putting people's bodies in fight mode for extended periods of time, and leading to exhaustion, burnout and faster ageing [9].

There is plenty of evidence that the use of technology has an impact of the amount of sleep we get. A 2015 survey showed that it was the sleep of young adults that was impacted most by technology. More recent studies show that that the problem is at least as severe in teenagers, [4] who practise late-night socializing, called vamping, which, in some extreme cases, takes place at any time of the night. Teenagers need around nine hours of sleep, but in 2015, 43% of US adolescents reported less than seven hours on most nights, which means that half of teenagers in the country are seriously sleep deprived, with 18-year-olds being the worst affected. Causes of disturbed sleep include (i) the use of social media which is both mentally and emotionally stimulating, and (ii) the blue light emitted by smartphones and tablets which simulates daylight, inhibiting the brain's production of melatonin, the hormone that regulates sleep.

Slowly, awareness about the negative effects of heavy smartphone usage is growing and even technology companies have started to offer tools to measure or restrict screen time, such as Apple's Screen Time and Google's Digital Wellbeing.

A number of former employees at the larger internet companies have started regretting what they built [10]. Some of them founded the Center for Humane Technology (http://humanetech.com) and give advice on how to take back control. The most extreme suggestion is to go "cold turkey" and delete all one's social media accounts. It has been claimed that this simple action will increase productivity, reduce stress and improve overall wellbeing. Some companies have introduced a policy not to allow their workers on the corporate network to check emails outside working hours. Sometimes it is useful to observe what insiders do; a number of high-profile executives at internet companies have admitted that they put serious restrictions on the use of social media and mobile devices by their own children.

However, at the same time, many schools are intensifying the use of technology as part of the learning process. This includes introducing "massive online open courses" (MOOCs) and flipped classroom courses, by using learning platforms that need to be used by children and students for their homework in the evening. According to an OECD study [11], the results are mixed at best. Students who use computers moderately at school tend to have somewhat better learning outcomes than students who use computers rarely. But students who use computers very frequently at school do a lot worse in most learning outcomes, even after accounting for social background and student demographics. The COVID-19 lockdown with distance learning will enable us to gauge the impact of intense computer use on learning. Time will tell whether the benefits of technology outweigh the side effects on children's development.

Computing transforms the job market

Computing, by definition, has an impact on the job market. The introduction of automation destroys jobs, creates new ones and changes the content of the remaining jobs. This has always been the case, ever since automation was invented. The key question many people have been focusing on is whether the current wave of automation fuelled by artificial intelligence and robotics will create more or fewer jobs than it destroys.

As of today, there are no signs that there are fewer jobs than, for example, twenty years ago, but the jobs have changed, and the effect of this change seems to be more

Figure 4: Cartoon by Jeff Stahler

inequality, a shrinking middle class and the emergence of a dual economy [12]. There is however a clear shift of jobs. The jobs that are most vulnerable to be destroyed are routine jobs, both manual (e.g. factory workers) and cognitive (e.g. accountants, or radiologists). The jobs that are created are non-routine jobs, i.e. jobs that require problem solving skills, creativity, entrepreneurship. In the near future, jobs in the event and hospitality sector might disappear when companies decide that some virtual events are as good as physical ones in the post-COVID-19 times, cutting down on travel. Some companies with a substantial part of the work force working from home are starting to wonder whether they still need huge office buildings. This will impact the real estate market. If people keep working a couple of days per week from home, catering and taxi services in business districts will also suffer. This is an indirect consequence of the use of technology in businesses.

Machine learning has a growing ecological cost

As the data sets used to train deep neural networks keep growing, energy consumption grows too. One study reports a 300,000-fold increase in power consumption for this purpose between 2012 and 2018. That same study reports that training one model with 175 billion parameters of a GPT-3 language model requires 28 000 GPU-days and has a carbon footprint of

85 tons. This is the emissions equivalent of one car driving 700,000 km [13]. The estimate for the power consumption of the AlphaGo Zero software is around 200 MWh, or a carbon footprint of 136 tons.

This is the cost of training one model. Fortunately, these large models (such as GPT-3) are rather effective in different use cases, decreasing the need for extra training. But if the model has to run on different platforms, different models might have to be trained. For applications in which the data set is changing regularly (like face recognition, traffic sign recognition, …) but also to fix bugs, the models might have to be retrained regularly, fortunately incrementally if the algorithm allows it (e.g. using transfer learning). It is clear that the environmental cost of training the models is no longer negligible, but it might be offset by the benefit of the resulting application, which can be distributed in large numbers (for example in smartphones) and perhaps also used for applications that offer energy savings.

Although the energy consumption mentioned seems to be huge, one has to put it in perspective. The 85 tons of CO_2 for the GPT-3 model is equivalent to the yearly emission of 'only' four American citizens. At global scale it is the equivalent of the average yearly emission of 21 people. This is, however, not an argument to leave everything as it is. Every ton of CO_2 that can be avoided, should be avoided.

Since the carbon intensity per MWh varies wildly between countries from almost zero in countries with an abundance of hydropower or nuclear power to tenfold or more in countries with 100% fossil fuel production, the carbon footprint of a model can be reduced by running it in a place with low carbon intensity. There are also carbon intensity fluctuations during the day and the year. By avoiding the periods of the day in which fossil fuel power plants have to support the electricity grid, the carbon footprint can be reduced. At the hardware level, low power customized accelerators should be used to reduce power consumption. And finally, at the algorithmic level, new training algorithms and models could be used like the once-for-all models that can be customized for different platforms instead of retraining them [14].

We also have to compare the computational cost to the value of the model. If a model would help reduce the energy needed for heating and cooling, or for transportation by only 1% – it is definitely worth it. Another example is agriculture, which is a big source of greenhouse gasses. We will need AI to feed the world's population while sustaining the planet. By making better use of natural resources, by creating new forms of agriculture in vertical farms, by growing meat in labs, … we will be able to reduce the ecological footprint of the world population. In these cases, the ecological benefits should of course also be bigger than the ecological costs.

According to the World Economic Forum, there are six environmental challenges where AI could be part of the solution, rather than part of the problem [15].

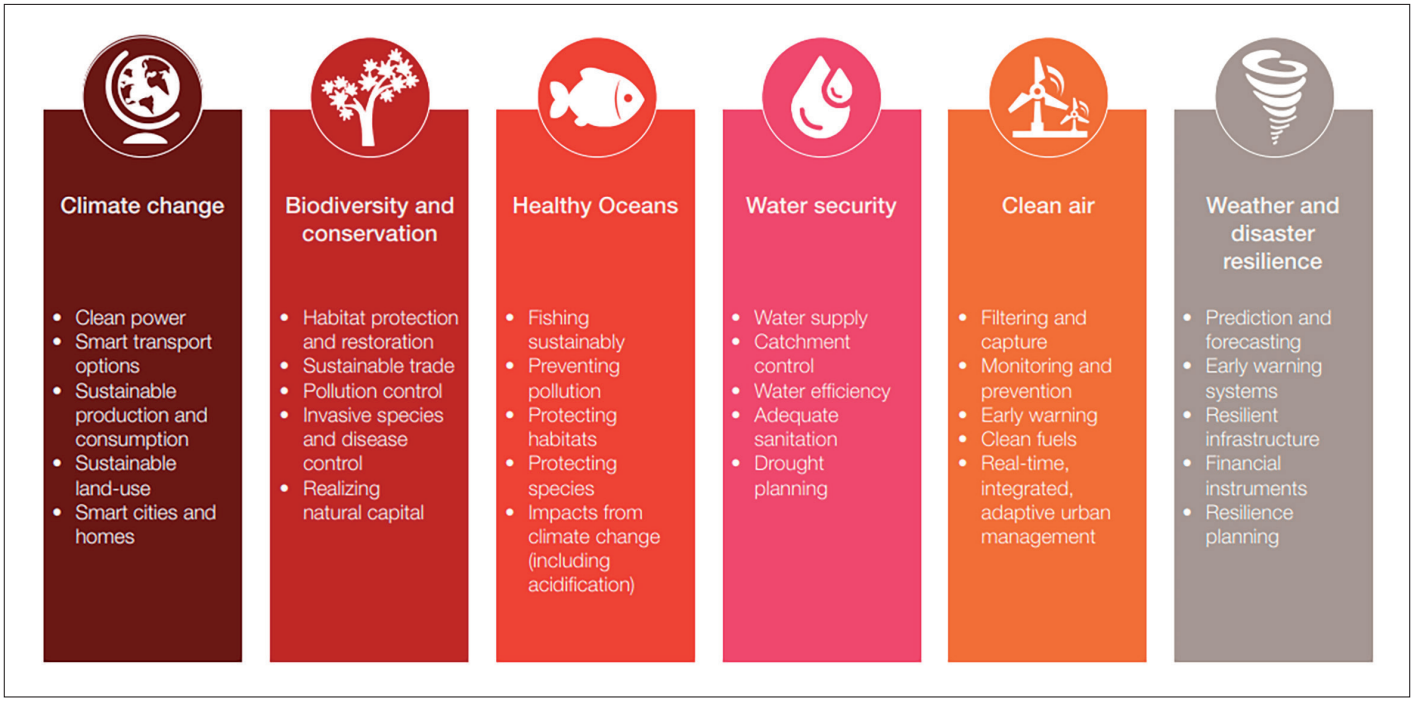


Figure 5: Priority action areas for addressing Earth challenge areas [15]

Figure 6: AI risks [15]

Unfortunately, there are also six risks.

Its conclusion is that we should work on 'responsible AI' on top of the classical criteria like safety, ethics, value and governance. AI should be used not to speed up the destruction of the Earth's ecosystem by facilitating more efficient extraction of its natural resources, but to preserve the ecosystem. This will require leadership from the public and private sector. Some companies are already taking the initiative in this direction [17].

References

[1] Richard Freed, "The Tech Industry's War on Kids: How psychology is being used as a weapon against children", 2018, https://medium.com/@richardnfreed/the-tech-industrys-psychological-war-on-kids-c452870464ce

[2] Jean Twenge, Gabrielle Martin, and Brian Spitzberg, "Trends in U.S. Adolescents' Media Use, 1976-2016: The Rise of Digital Media, the Decline of TV, and the (Near) Demise of Print", 2018. Psychology of Popular Media Culture, https://www.apa.org/pubs/journals/releases/ppm-ppm0000203.pdf

[3] Kimberley Holland, "How to Identify and Manage Phubbing", 2018, https://www.healthline.com/health/phubbing

[4] Jean Twenge, "Analysis: Teens are sleeping less. Why? Smartphones", PBS, 19 Oct 2017, https://www.pbs.org/newshour/science/analysis-teens-are-sleeping-less-why-smartphones

[5] Samidh Chakrabarti, "Hard Questions: What Effect Does Social Media Have on Democracy?", https://newsroom.fb.com/news/2018/01/effect-social-media-democracy/

[6] James Vincent, "Watch Jordan Peele use AI to make Barack Obama deliver a PSA about fake news", The Verge, 17 Apr 2018, https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed

[7] S. Coughlan, "Digital dependence 'eroding human memory'", BBC News, October 2015. http://www.bbc.com/news/education-34454264

[8] Nicholas Carr, "The Shallows: What the Internet Is Doing to Our Brains", 2010, Norton & Companym

[9] David Robson, "The reasons why exhaustion and burnout are so common", http://www.bbc.com/future/story/20160721-thereasons-why-exhaustion-and-burnout-are-so-common

[10] Noah Kulwin, "The Internet Apologizes", The New York magazine, 2018, http://nymag.com/selectall/2018/04/an-apology-forthe-internet-from-the-people-who-built-it.html

[11] "Students, Computers and Learning: Making the connection", OECD Publishing, 2015, https://read.oecd-ilibrary.org/education/students-computers-and-learning_9789264239555-en

[12] Peter Temin, "The Vanishing Middle Class", 2017, MIT Press

[13] Lasse F. Wolff Anthony, Benjamin Kanding, Raghavendra Selvan, "Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models", https://arxiv.org/pdf/2007.03051.pdf

[14] Rob Matheson, "Reducing the carbon footprint of artificial intelligence", 2020, https://news.mit.edu/2020/artificial-intelligence-ai-carbon-footprint-0423

[15] "Harnessing Artificial Intelligence for the Earth", World Economic Forum, 2018, http://www3.weforum.org/docs/Harnessing_Artificial_Intelligence_for_the_Earth_report_2018.pdf

[16] Julia Carrie Wong, "The Cambridge Analytica scandal changed the world – but it didn't change Facebook", https://www.theguardian.com/technology/2019/mar/17/the-cambridge-analytica-scandal-changed-the-world-but-it-didnt-change-facebook

[17] David Hagenbuch, "The 4 Pillars of Ethical Enterprises", https://www.entrepreneur.com/article/240035

Koen De Bosschere is Professor in the Electronics department of Ghent University, Ghent, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: K. De Bosschere. AI for a better society. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 130-137, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

Fighting the COVID-19 pandemic is a good rehearsal for tackling the global challenges the world will have to face in the 21st century. If we are smart, we can learn a lot from it.

COVID-19 is more than a pandemic

By KOEN DE BOSSCHERE

Small crises lead to small changes; large crises lead to large changes. COVID-19 is a large crisis but, at this moment in time, it is hard to tell what its long-lasting impacts on society and the economy will be. One thing is sure: we will all make a distinction between the time before Corona and the time after Corona, and the new normal is very likely to be different from the old normal, no matter how hard people try to go back to the latter. Historically, it might become as impactful as the fall of the Berlin Wall, or 9/11.

Although it is still premature to talk about the new normal, it is good to start thinking about it by analyzing how society is dealing with COVID-19 at the moment, and by looking at the trends and trying to imagine how they will evolve in the near future.

Key insights

- Thanks to all the previous efforts to digitize society and the economy, western countries were ready to go online quickly and without major disruption to essential services. A stable broadband internet connection at home has proven to be as essential as water and electricity.
- COVID-19 speeds up innovation and digitization, which creates huge opportunities for the computing industry in the coming years: more online services and automation in all economic sectors, more remote working, more e-commerce.
- COVID-19 is an opportunity to rethink and restructure the economy, and to prepare it for the challenges ahead of us. The post-COVID world will be different from the pre-COVID world: politically, economically, and societally. COVID-19 is a once in a lifetime opportunity to redesign the world we live in.

Key recommendations

- Europe should use the momentum generated by the COVID-19 pandemic to speed up digitalization of Europe by investing in broadband access for all, 5G, e-commerce, smart transportation, smart cities, …

When COVID-19 emerged in late 2019, most people could not imagine the impact it was going to have on the world. In the space of just three months, it spread to every continent, initially leading to the isolation of individuals, followed by a lockdown of cities and eventually a lockdown of most European countries in Spring and Autumn 2020. Only essential workers were allowed to go to work. All non-essential businesses had to close and large gatherings, including activities in schools and higher education institutions, were forbidden for months. Private mobility was severely restricted. Cities and countries came to a standstill.

Most European countries succeeded in reducing case numbers enough to allow them to cautiously reopen the economy by June, in time for the holiday season. However, only weeks after the start of the season, the number of cases started rising again, eventually leading to a resurgence (the 'second wave') in autumn. It then became clear that without the vaccination of the global population, the virus would continue to spread, leading to massive economic and human losses. Until a vaccine becomes available, the only option left is to learn to live with the virus.

For most people, more than anything else, the virus has impacted their lifestyle: less social interaction, less travel, fewer events, less personal freedom. Many people have struggled to believe that this is happening to them in the 21st century, and still hope that it is just an ordinary nightmare. Unfortunately, it is a reality, and it is both sobering and humbling.

The lockdown is without any doubt the biggest economic and social experiment of the early 21st century. A lockdown was completely unimaginable just a year before the pandemic struck. Such an experiment creates unprecedented learning opportunities: what went well, what went wrong, who was forgotten, how should we prepare for a recovery, what is the role of science in dealing with a global crisis, and how can we take advantage of this situation to do better in the future?

The situation is often compared with the Spanish flu of 1918-1919, which

Image: © DisobeyArt – Shutterstock.com

infected about 500 million people, killing 50 million. It is also compared with the Black Death of 1347-1351, which led to an estimated death toll of 75-200 million people. Both pandemics led not only to a lot of human suffering but also to profound and lasting societal and economic change. The Black Death dramatically improved the standard of living of the masses (due to a shortage in the labour force, they could negotiate better conditions), and it eventually marked the end of feudalism in Europe and the onset of the early Renaissance. The Spanish flu in time led to public healthcare systems and socialized medicine and healthcare [1, 2].

COVID-19 exposes existing strengths and weaknesses

Crises always put stress on systems, and, when a system is stressed, its strengths and weaknesses are exposed. Some wealthy countries did not have the courage or hesitated to make fast draconian decisions at the beginning of the pandemic with the hope of protecting their (large) economies as much as possible. In some cases, their hesitation led to additional economic and human devastation. The virus treats rich and poor countries alike. Many poor countries suffered a lot due to a lack of resources, but some of them dealt remarkably well with the pandemic. The resilience they need to survive in a normal situation might have helped them deal with the COVID-19 virus and its effects.

One notable case is the United States, where the presidential election campaign coincided with and became, in a sense, part of the pandemic, heavily politicizing measures to fight the virus. At state level, policy was determined not only by public health considerations but also by electoral considerations, and by the choice between *people first* and *economy first*. In some states, this approach has led to very high case counts, a severely stressed healthcare system and high mortality. It is surprising that the richest and smartest country in the world did not perform better, and was not able to reduce the case count far enough to allow a safe reopening of the economy. Indeed, four months was not even sufficient for the country to set up an effective nationwide testing system, while other western countries and some developing countries managed to set up a testing system in just a couple of months.

In Europe, over 30 sovereign countries had to fight the pandemic in their own territory, with their own resources. The public healthcare systems of most European countries did not collapse in the spring and were able to give care to the patients who needed it (but still with a high death toll). The United Kingdom and Sweden were two special cases: the UK was late to announce a lockdown, and Sweden decided to let the virus spread in its population in an unsuccessful effort to protect its economy. Future analysis will show what the best approach was. Surprisingly and unfortunately, no leadership or substantial help came at European level until mid-2020, when a recovery deal was approved.

The United States and Europe both suffered a serious second wave while many Asian countries were able to prevent it. For the first wave, one could argue that the virus was new, and that we were not prepared, but that argument does not hold for the second wave. The second wave was accurately predicted by the scientists, and still it caught us off guard. This brings back reminiscences of the 2008 financial crisis that we could not prevent and did not manage very well. It shows that the West is

not always superior to the rest of the world. Both events are damaging for the reputation of the United States and Europe, especially in Asia and in the Pacific. One day, these countries might question the preeminence of the United States and Europe in some areas and start wondering why, for example, the head of the World Bank is always an American and the IMF always European. Hence COVID-19 might eventually change the world order (a bit) [13].

There is a clear parallel between the second wave and climate change [8]: scientists have been warning politicians of its dangers for years and their models are quite accurate. It is a process with feedback loops, long delays and a potential tipping point. It is a global shift that requires urgent action and international collaboration to stop it, and the short-term solutions seem to hurt the economy. Hopefully, we will learn from COVID-19 how to best tackle that challenge [9].

The impact of COVID-19 on the economy

Given the duration of the pandemic, and the impact on society and the economy, it will take time to recover from the recession that follows the lockdown. Many economists were hoping that, after a short pandemic, the economy would experience a quick V-shaped recovery. Unfortunately, the longer the pandemic lasts, the less likely it is that we will see a fast recovery across all sectors of the economy. There is little hope that we will be back to normal before there is a vaccine that leads to herd immunity of the population. Vaccines will be available early 2021, but there will be an inevitable lag before a sufficiently large portion of the world population is inoculated.

The economic impact, however, depends on the sector. Important economic sectors like energy, construction and manufacturing were shut down for a relatively short time, and are recovering fast. Other sectors including the events sector (trade fairs, cultural events, sports), tourism and the hospitality sector, will not fully recover until an effective COVID-19 cure is found. Until then, these businesses will have to scale down, and many of them might go out of business, leading to considerable unemployment. According to the Federal Reserve Chairman Jerome Powell, the economy as we knew it might be over because the pandemic has accelerated the introduction of technology, e-commerce, telework and automation, and the lower-paid workers will be impacted more than higher-paid workers. It might take years to for the market to adjust [12].

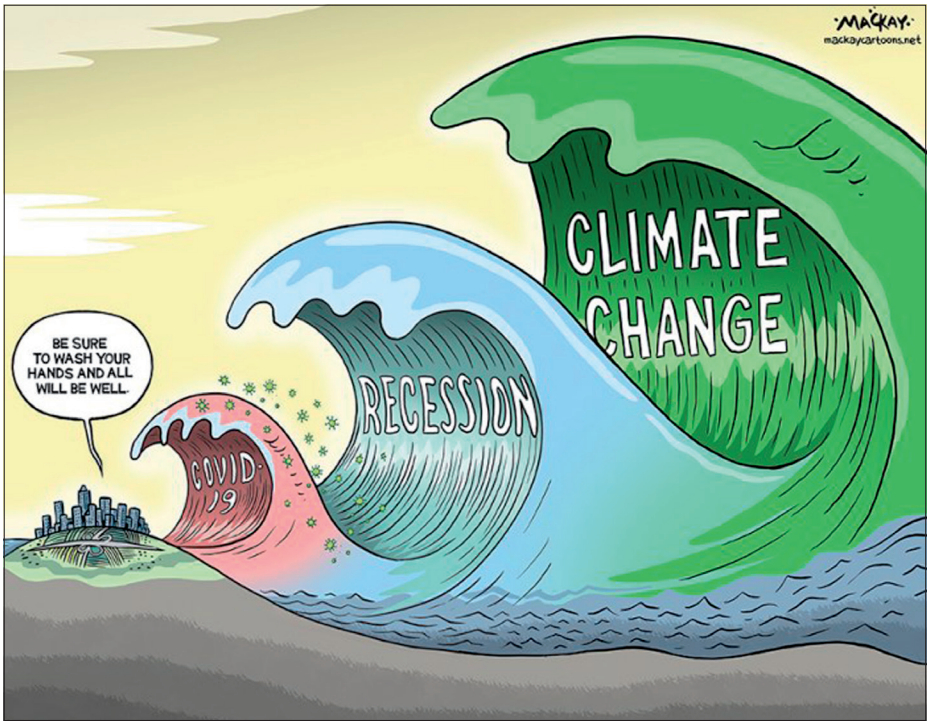


Figure 1: COVID-19 will not be the last global crisis (Source: https://mackaycartoons.net)

A shrinking economy leads to a shrinking tax revenue, and lower public spending (or increased debts). This is a double whammy: after all the extra expenses incurred to fight the virus, countries will need more money to support their economy and their social security (health care, unemployment benefits) at the very time that the budget is decreasing. Together with the rising costs associated with an ageing population and investments to fight climate change, this turns the aftermath of the pandemic into a perfect storm for years to come. In that budget situation, countries might become more self-centric, focused on their own problems, and less willing to help each other, or refuse to contribute to international organizations. Early signs of this are the growing number of countries that are leaving international agreements, or lose interest in international leadership. COVID-19 was the first pandemic for which the G7 could not agree on a common text because one country wanted to refer to COVID-19 as the "China virus" [7]. COVID-19 has already had an impact on the progress of the sustainable development goals [3].

At the same time, this situation might also be an opportunity to rethink the economy. Mass tourism has become a burden in many places, and air travel is one of the causes of climate change. Ecologists have long been advocating short food supply chains. Some countries tried hard to bring back manufacturing from the low wage countries. Donald Trump accelerated this process at a global scale with his "America first" stance.

Some economists are advocating new economic models like the doughnut economy [4], or other economic models that look not only at growth but also at other metrics to assess progress. All these models encourage governments to no longer invest in old industries with a large carbon footprint, but instead to invest in a green recovery [3] by supporting companies that work on sustainable solutions. Europe and Canada seem to be eager to take that path [11].

If we act smartly, the COVID-19 recession offers a unique opportunity to replace lost jobs with more sustainable jobs that compensate for the losses. Unfortunately, this will take some time.

Sovereignty

COVID-19 has shown a weakness of the global market. The international supply chains on which many countries rely for essential goods are not guaranteed to work well in the case of a major international crisis. In normal circumstances, if demand goes up, production is increased, and everything goes back to normal. This works well in the absence of natural disasters, export restrictions, war, etc.

The COVID-19 crisis has shown how dependent all countries are on the rest of the world. Most countries were not able to quickly produce the personal protective equipment (PPE) and the testing equipment they needed. Over the course of recent decades, many countries gave up manufacturing low cost commodity products, because they could be bought at a cheaper price in low-wage countries. At the same time, stockpiles were reduced because money could be saved by ordering in the moment. This works fine in a normal situation, but not in a global pandemic where a large part of the planet is in lockdown, and where the whole world is simultaneously and frantically trying to buy the same products (ranging from facemasks to ventilators to toilet paper). Setting up large-scale local manufacturing in a crisis situation is not simple and takes time. Even after six months, some countries were still struggling to get enough PPE for hospitals and nursing homes. This lack of PPE has led directly to the loss of thousands of healthcare workers across the world.

There is a lesson to be learned from this experience. Optimizing supply chains by buying from the cheapest supplier, and eliminating all buffers to save money, kills resilience. Even without man-made export restrictions, there can be situations in which supply chains are broken by external causes (natural disasters, war, global hoarding/stockpiling). For essential goods and raw materials, it is wise to always have at least one local source. It is worrisome that, today, some essential life-saving drugs are produced in a very small number of countries (90% of all penicillin is produced in China; Europe no longer produces paracetamol [7]. Europe might consider as a future requirement that essential goods like commodity medical equipment and life-saving drugs that are admitted to the European market are also (partially) produced in Europe.

Broken supply chains eventually get restored, and all the stakeholders involved work together to restore them as soon as possible. Export and import restrictions are totally different. They are political (e.g. to protect a country's own industry, or as a retaliation measure), which means that the industrial stakeholders involved can do little to lift the restrictions (especially if they were imposed by a different country). Only the politicians can do this, and the interests of a handful of companies might not weigh enough to change relations between countries.

The growing interest in sovereignty will unavoidably lead to further de-globalization, which has been an ongoing process over the last decade [7].

COVID-19 accelerates digital transformation

The sudden lockdowns in many countries forced companies and organizations to figure out how they could continue their operations under the restrictions and stay-at-home orders. Projects that would normally take months or years to implement were implemented in days, weeks or months. Examples of transformations that happened almost overnight are:

- Cashless payments. Today an increasing number of businesses simply refuse to accept cash. Such a transition would normally take years to complete in a cash-based economy, and require extra legislation. It is not likely that cash will make a massive comeback after COVID-19.
- Online teaching and testing. This was completely uncharted territory for most primary and secondary schools, but it has now become common in many places. For higher education, online teaching was not new, but the speed at which it was scaled up is unseen. It is clear that after the schools go back to normal, online teaching will be retained for some activities.
- Digital signing. Wet signatures on paper are impractical under stringent stay-at-home orders. It did not take long for digital signatures to be widely accepted for most documents that need to be signed. It is unlikely that wet signatures will come back. Governments should instead work on the infrastructure to support digital signatures.
- E-commerce. Companies that invested in an online shop before COVID-19 did not have to completely close their business during the lockdown and could continue to serve their customers. Major online

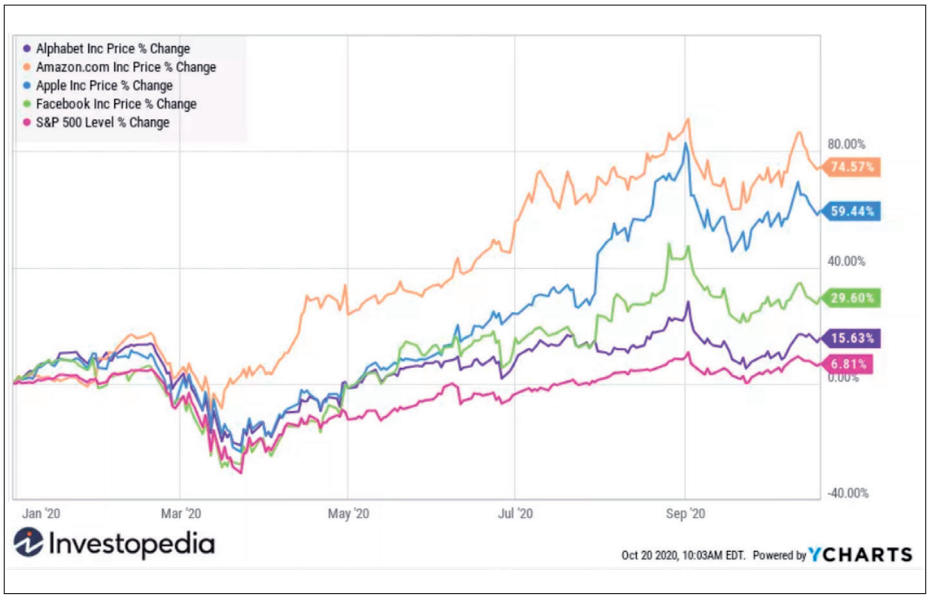


Figure 2: Evolution of the shares of big tech companies since the lockdown in mid-March.

shops experienced a large increase in trading, as did the shipping and courier companies. The fact that Amazon became one of the most valuable companies on the planet is no surprise (Figure 2).

- Remote working. Stay-at-home orders led to unemployment for people working in non-essential jobs that could not be done remotely and to home working for the rest. Whereas office work used to be the norm, and remote working the exception (often granted to the employee as a favour), telework has now become the new normal, and office work the exception, in many companies. Employers' attitudes towards home working have changed, and many employees have discovered its advantages (no time lost to commuting, more flexibility). Office work will return, but it will not replace all telework. This evolution will also impact the real estate market. Many people are looking for an apartment or a house with an extra room for a home office.
- Online meetings. Many people have discovered online video meetings during the lockdown period and have learned that they can be quite efficient for certain types of discussion. They also discovered the disadvantages: less opportunity for informal contact, less information about other participants from body language, more fatigue and exhaustion. Nevertheless, after COVID-19, online meetings will stay with us. Even medical doctors discovered that they could diagnose some of their patients remotely via a video chat – completely eliminating the risk of infection.

It is almost a miracle that this transition was able to happen so fast with the lockdown and stay-at-home-orders in place. This was only possible thanks to the fact that all the underlying technologies and infrastructure were already in place, and were ready to be scaled up via huge global data centres. Industry was ready for it; it only had to push the button. In the early days, there were some bandwidth and stability problems, but most of them disappeared very quickly. Lockdown and social distancing also created opportunities for innovative digital startups, which saw a surge in the adoption of their solutions, like companies producing rapid tests, vaccines, solutions for efficiently disinfecting objects, models to predict the evolution of the pandemic, …. It is fair to say that COVID-19 has spurred innovation, and that it is an accelerator for the adoption of digital technologies. It also unveiled the weaknesses of the adopted solutions. Since the COVID-19 situation will last for a year or longer, several of the ad hoc solutions will become permanent and there is a huge opportunity for the computing industry to launch new solutions.

Image: © DisobeyArt – Shutterstock.com

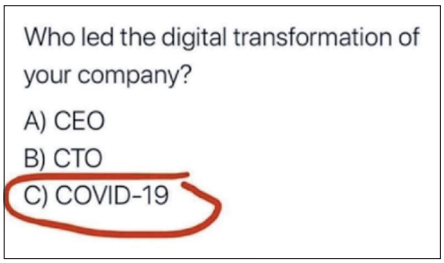


Figure 3: COVID-19 was in 2020 a main driver for digital transformation (Source: https://journal.laurea.fi/welcome-our-new-digital-transformation-officer-COVID-19/)

COVID-19 increases inequality

COVID-19 has impacted the less educated more than it has the well-educated and has affected females more than males. The reason is clear: those with lower levels of education more often carry out jobs that were affected by the lockdown (retail, hospitality) or carry out essential jobs (food preparation, cleaning, nursing, transport), which exposed them more to the virus (and to the costs of a treatment in the event of infection). Those with higher levels of education are more likely to have office jobs that can be done from home (teachers, managers, accountants), and they live in better homes, making it easier for them to stay at home and avoid infection. A stay-at-home order is simpler in a suburban house with a garden, a pool and a broadband internet connection than in a one-bedroom apartment in a city. To make things worse, a less educated person who lost their job will find it more difficult to find a new one, further increasing inequality. This is different from the situation after the Black Death, where there was much more work to be done than there were available workers.

But there is more. The impact of COVID-19 also has a varying impact on the different generations.

- Children will forever remember the year that they did not have to go to school. Children from underprivileged families were severely hit (no computer at home for online lessons, no daily hot meal at school, overcrowded homes with domestic violence, in some cases). This might have a lasting impact on their further development. Children from middle-class families might experience a far less negative impact from COVID-19.

Figure 4: Levels are historically high.

- University students (generation Z or zoomers) were suddenly studying at an online university, and many decided to move back to their parents' home after having lived for a couple of years in a university residence. This has social implications for both the students and their parents. Students found it more difficult to find student jobs (which had previously often been in bars and restaurants) to help pay for their study. Students graduating in 2020 and 2021 might end up with a higher student debt and might have a harder time finding a first job.
- Working parents with young children (the millennials) had a hard time too. They had to find a solution for their children who could not attend school, and were expected to help them with their schoolwork. At the same time, they also had to work full time, either at the workplace or at home.
- Working or recently-retired people in the 50-70 age group (baby-boomers) were probably hit the least. They do not have to combine caring for children with a job or retirement activities, and their age group was less vulnerable than the over-70s.

Depending on how hard the post COVID-19 recession hits and how long it takes to overcome it, the intergenerational solidarity between the boomers and the zoomers might be severely tested. The baby boom generation is now 55+, has built up some wealth and has recently retired or is looking forward to retirement. Many of them were not hit hard by COVID-19 or the recession. The zoomers, however, are now entering higher education. They will enter a job market that might not offer them full employment and many of them will carry a student debt. They will have to pay taxes to pay back the public debt incurred during the pandemic, for climate action and to support the ageing population (boomers and older). Debt levels have now surpassed the global debt level seen at the end of the Second World War (expressed at % of GDP, see Figure 4) [10].

When the boomers pass away, their wealth will be transferred to the zoomers' parents (millennials), not to the zoomers. So, COVID-19 might lead to serious discussions about new forms of intergenerational solidarity [5, 6].

Conclusion

COVID-19 is more than a pandemic; it has the potential to change the world as we know it. We will have to make the right decisions, and change the world for the better: more sustainable, more equal, more diverse. It is important to imagine the world you want in 2040, and then to work towards it, one step at a time. The world needs visionaries to show the rest of the world the possibilities.

References

[1] Kate Whiting, "A science journalist explains how the Spanish flu changed the world", https://www.weforum.org/agenda/2020/04/COVID-19-how-spanish-flu-changed-world/.

[2] Lawrence Wright, "How Pandemics Wreak Havoc – and open minds", https://www.newyorker.com/magazine/2020/07/20/how-pandemics-wreak-havoc-and-open-minds

[3] David Watts, "Global sustainable development in the aftermath of the COVID-19 pandemic", https://ieep.eu/news/global-sustainable-development-in-the-aftermath-of-the-COVID-19-pandemic

[4] Kate Raworth, "Meet the doughnut: the new economic model that could help end inequality", https://www.weforum.org/agenda/2017/04/the-new-economic-model-that-could-end-inequality-doughnut/

[5] Dave Lee, "The Recessionals: Why COVID-19 is another cruel setback for millennials", https://www.straitstimes.com/opinion/the-recessionals-why-COVID-19-is-another-cruel-setback-for-millennials

[6] Kendra Cherry, "How Different Generations Are Responding to COVID-19", https://www.verywellmind.com/how-different-generations-are-responding-to-COVID-19-4802517

[7] Josep Borrell, "The post-coronavirus world is already here", https://www.ecfr.eu/publications/summary/the_post_coronavirus_world_is_already_here

[8] Dania Eel Akkawi, "Climate Change and COVID-19: There Is More Than One Curve to Flatten", https://www.thecairoreview.com/midan/climate-change-and-COVID-19-there-is-more-than-one-curve-to-flatten

[9] Bill Gates, "COVID-19 is awful. Climate change could be worse", https://www.gatesnotes.com/Energy/Climate-and-COVID-19

[10] Martin Wolf, "The threat of long economic Covid looms", https://www.ft.com/content/f9a0c784-712e-4bf9-b994-55f8d63316d9

[11] "A European Green Deal", https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en

[12] "The economy as we knew it might be over, Fed Chairman says", https://edition.cnn.com/2020/11/12/economy/economy-after-covid-powell/index.html

[13] Sven Biscop, "Can corona cure our superiority complex?", https://www.egmontinstitute.be/can-corona-cure-our-superiority-complex/

Koen De Bosschere is Professor in the Electronics department of Ghent University, Ghent, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: K. De Bosschere. COVID-19 is more than a pandemic. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 138-143, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

Digitalization has expanded to the point where it is an actor in every aspect of human life, and an important factor in our ability to maintain peace and democracy. If we identify and manage properly the challenges this poses, we stand to gain greatly.

The impact of technological evolution… on humans and societies

By THOMAS HOBERG

As information technology has transformed from a faster way to calculate numbers to the default way of professional and personal interaction, it has become mostly about content, where common ground for cultures, nations, religions and even couples can be much harder to find. What started as niche local products has become highly political, merely by scaling to billions of users, a natural threat to any leader who cannot control it, and to consumers who cannot escape it. Yet for sustained economic growth, we need to keep scaling it, without destroying its value or the planet.

Key insights

- The IT industry has reached the size of humanity and is exhausting its capacity for content consumption. Significant and sustainable growth could be created through products and services enabled for operation through smart virtual assistants, if their owners can trust them to remain loyal to them.
- The West believes opinions should spread freely, but technology be controlled; the East believes technology should be freely shared, but the spread of opinions needs tight control. As the majority of all human interaction becomes digital, cloud giants need to transform from private corporate ethics to comply with the huge diversity of ground rules.
- Content which self-optimizes for consumer consumption, naturally becomes more addictive, less objective and risks breaking established social code.
- Culture, traditions, beliefs and legal frameworks are the social code that keep societies working with minimal friction between humans. When humanity moves the majority of its interactions to the digital space, pieces of software used by the majority of users also become social code, simply because of the scale of its applications and the embedded decision making it contains. The governance used for the previous social code must then be extended to such software.
- We are in a new type of cold war, attempting destabilization and control over foreign nations via control over technology to achieve dominance, while the planet requires global collaboration.

Key recommendations

- Extending the value of a multitude of devices via computational intelligence without compromising their utility by increased individual need for attention, requires that the individual devices be holistically managed by no more than one virtual personal assistant per major domain.
- Support sustainable value of devices and virtual personal assistants via open standards, open source assets and regulation.
- Develop, support and mandate the use of modular and multiple compliance and ethics frameworks for code and content so distinct enclaves for e.g. owner, corporate, parent, vendor, government or party service provider can co-exist safely, without compromising owner sovereignty in the EU.
- Mandate the decoupling of corporate ethics from current web giant products and services within the EU and value system allies.

Image: ID 172194124 | ©Elnur | D reamstime.com

The challenge of sustainable economic growth at planetary scale

A smartphone is one of the most complex industrial products that we have today, yet everybody who can obtain value from owning one, already has one. And most already spend as much time with the device or on a content platform as they can afford to. There is no incremental value in having twice or ten times as many phones, multiplying your social media channels or the time you spend with either—or their permutations; we are stuck at the scale of humanity.

And it's not that different for any other complex industrial product, be it a car, a climate control system or a washing machine: the potential demand no longer has orders of magnitude of growth remaining, and the actual demand is increasingly satisfied by producers outside of Europe.

To retain significant economic potential, Europe needs to overcome this barrier of human scale and find products or services scaling several orders of magnitude beyond. These must still be valuable to human consumers to buy, but without placing undue strain on consumer budget or attention and without endangering Europe's stringent goals on sustainability and citizen benefits.

The vision of cyber-physical systems (CPS) or an internet of things (IoT) was created as one avenue to achieve this seemingly impossible mission, of providing economic growth far beyond human scale in the world of IT [1].

IoT for growth beyond human scale

The vision of IoT is to embed digital controls into products as we replace them, and to employ application program interfaces (APIs) to match them with services. We then design these services to increase the level of convenience, e.g. through autonomy, seamless integration or additional control, depending on personal preference. And for Europe we need to ensure that neither sustainability nor civil liberties suffer, while our products are globally more competitive because of it.

Current approaches, whether they are called Siri, Alexa, Cortana, Google Assistant or WeChat fall seriously short, because they don't put these virtues up front. Yet they have created valuable concepts and open source assets, which Europe must quickly expand to build a powerhouse of product and service offerings. These offerings need to be sustainable not only environmentally, but also as a long term service offering to consumers, who can trust that their investment in adapting habits to match this mesh of assistants and devices will bring life-long benefits that will not simply disappear after a few months or years, as vendors merge or fail, and will not turn them into a victims of addiction, price-hikes or ransomware.

Europe is in a much better position to deliver sustainable user benefits than US or Chinese internet giants, whose culture and business models do not match. But as

both struggle to grow beyond their home territories, they will catch up quickly. We have a chance, but we need to make haste, because sustainable growth seems difficult to find elsewhere.

Scale vs. resilience in a world torn between globalization and isolationism

Because scale in IT is so important, it is naturally global, unless it's constrained artificially. Historically, some governments have tried to protect their local IT industry from competition via tariffs and import restrictions; likewise, there was a long history of export restrictions during the Cold War. But generally, the US IT industry has done rather well at using its Second World War-induced lead and huge domestic market to outpace competitors everywhere on the planet, as long as the competition was limited to technology.

With the rise of the internet and digitalization, IT also became about content, and content contains culture, beliefs, rules and values. These are areas where countries like China or Russia, Iran or Cuba obviously beg to differ: even Europe finds itself diverted and divided over the threat of a naked nipple vs. a concealed gun or free vs. hate speech.

China decided early on that the free flow of content and culture needed tight controls, while it found it harder to accept that the free exchange of technology or intellectual assets should be restricted. This East-West double barrier created a walled garden, which allowed China's domestic digital industry to flourish under careful government control to a scale where it stands to outgrow that of the United States—and with better future projections.

At this point, the United States is trying to turn the tide, leveraging every bit of control it has left, with Europe and most other nations on the planet stuck somewhere in the middle. The European Union was created with the vision of uniting its member states with their rather large diversity in culture, values and legislation into a single economy much bigger than any nation alone. Europe's clear choice was to take a running leap to expand this experience of managing diversity and modular regulation to the rest of the world: it had painfully learned not to impose religion or national ideology on neighbours and not to mistake isolationism for autonomy and thus risk insignificance.

Today we experience how long-standing international agreements are broken for national political expedience; we live in a world of constant cyberwarfare that can flare into physical trade wars, or worse, at any moment. Since Europe cannot control how the major nations swing between protectionism and free trade, it needs to operate with a strategy that works as well as possible with both, a multi-lateral open global market and deep trade and cyberwar trenches.

Within the domain of IT, digital and open source assets are more naturally global while physical and proprietary ones are more likely under the control of companies which can be coerced by governments. As a consequence, Europe must use, encourage, protect, produce and spread open source assets and avoid foreign proprietary ones to ensure sovereignty. And one of the most critical challenges here is making open source financially viable.

It should be noted that for the discussion of 'open source', we treat code and data as synonymous, not limited to computer source code, but including intellectual property assets, hardware description language code, datasets, recordings, images, video, books, designs, specifications, protocols, rule-sets, neural network training results etc.; in short, any digital content or binary encoded immaterial asset that is made available for copying and modification.

Europe needs to ensure that its citizens, its industries and its governments have complete access to all the cyber-physical systems components that they might need or want at prices and effort levels that are globally competitive.

The aspects of empowerment according to consumers, vendors and governments

Computers, and especially the 'very personal computers' somehow still called smartphones, have become a bit of everything: a tool to the point of a digital brain extension or an internet limb, and a digital serf, secretary or butler we can hardly do without.

They extend our senses and limbs, our mouth and individual reach from arm's length to around the globe and to everyone else likewise equipped; they give us telekinetic and telepathic powers over everything or everyone connected to the Internet. Their potential to empower provides a maximum of value to their owner, yet could put a nuclear launch button into the hands of a misguided kid, or grant titan mind changing powers to moral dwarfs: what consumers consider value, governments consider a potential threat while internet giants simply treat it as their own.

The original personal computer was regarded much like any other valuable tool or costly appliance at home or in the office: nobody but the owner could decide how and by whom it was used, and the attempt by a third party to access its data or use it without permission was considered hacking and a crime. The far more personal computers in our pockets currently receive a treatment that would have been classified as hacking thirty years ago, yet are claimed as a new normal by those with vested interests. GAFA and BATX have long sold devices and applications that might be considered foreign/antagonistic informers/denunciators and influencers/manipulators, when owners deserve undivided loyalty. And even democratic governments, who should be fighting such misuse, seem more interested in turning phones owned by citizens into agents of state and guardians of public health e.g. to fight the COVID-19 pandemic.

In this triangle of owners, vendors and governments, only the owner is the one who chooses to invest based on obtainable value, but the other two seem invested at reducing choice. Since devices and services that are disloyal to their owners

Image: ID 48605501 | ©Denisismagilov | Dreamstime.com

and consumers attain little or even negative value to the one who pays, this trend chokes off value creation, before it can really take off.

Europe should neither prescribe how other geographies and value systems operate this most personal computing space, nor suffer foreign intervention: when relationships become digital, the law on the ground also needs to govern clouds above. Current early implementations of smart assistants like Amazon's Alexa or Apple's Siri tie corporate ethics and US regulation to technology and consumers to the device vendor: both are very much limiting their usefulness to the lowest common denominator internationally. If digital assistants are to become as smart and resourceful as a well-trained butler thirty years with the family, that would require a level of discretion and loyalty to their owners, which is impossible to sustain with current business and operating models.

Europe's unique opportunity here lies in the fact that it doesn't have an industry that depends on strip mining consumers' data or governments that depend on censoring public opinion. We can use the global power of open source and enforced compliance within our borders to ensure that EU companies can concentrate on improving our digital servants' talents and skills, in our homes, workplaces and places of leisure, in agriculture and the great outdoors, as well as in the public spaces in between.

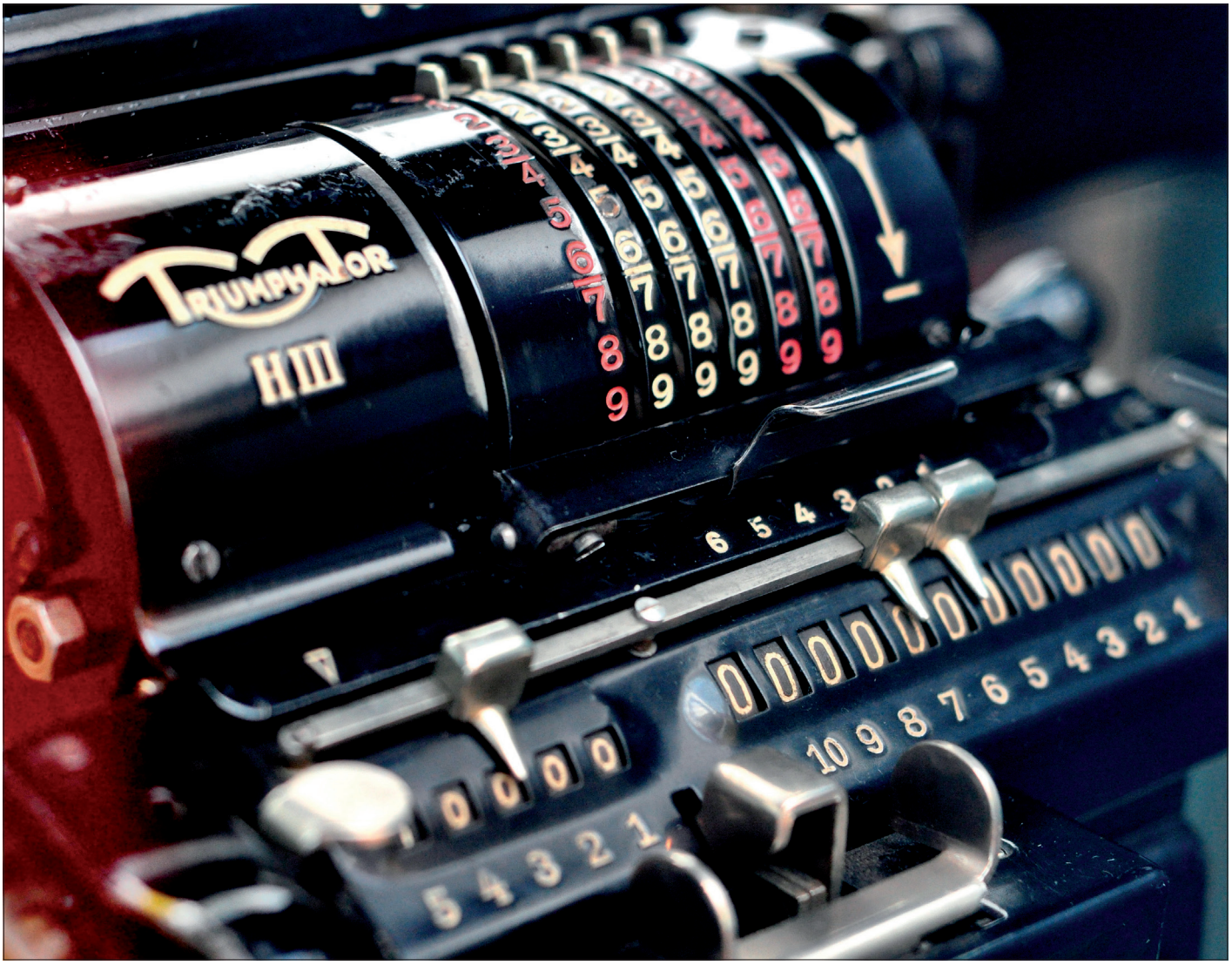
Only without vendor lock-in, using open standards, proven interoperability and open source assets, can devices and servants be operated at an ecological optimum and sustained over the lifetime of a consumer. If we can ensure that at least within the borders of our laws we can interact with our digital servants under legal privileges similar to those we enjoy with our spouses, only the technical challenges of seamless interaction or digital augmentation remain to be solved; and those are significant enough.

The battle for attention and addiction

Facebook, Google and in fact every form of wooing or advertisement, is about influencing our prejudices in favour of their sponsors. A front-door salesman would grow hungry, hoarse or just tired after meeting a few potential customers a day. Yet cloud servers can whisper into millions of minds at once, have infinite patience and perseverance, no conscience for lack of conscious and, with the help of deep analysis of clicks, likes and purchases of social peers, evolve content into the very gestalt consumers like best, often past the threshold of addiction, creating a feedback bubble that is difficult to escape.

Within these automatically individualized bubbles self-reinforcement leads to radicalization of opinion and expression that can easily lead to monstrous content. And the social media CEOs who merely intended 'to connect people' [2], suddenly find themselves accused for the damage that creates while the effort for clearing out data along a diversity of guidelines that matches their user base, threatens their bottom line.

Intended or not by the providers, the evolution of social media and streaming video will seek to keep the exclusive attention of a consumer, because it's much more expensive to gain back from someone who strayed elsewhere. With the help of decades of cognitive science and the



almost limitless economy of scale that software provides, endless scrolls of auto-start content are generated and carefully self-tune to a victim's preference, a permanent cliffhanger, aimed to wring peak purchasing budget and attention from consumers.

As they scale to billions of clients, their gravitational influencing pull eclipses that of major nations, upsetting their leaders, who with a focus on ground and borders, didn't realize clouds could do that.

The serious game of making rules or sovereignty in a digitalized society

Little is so important for the evolution of higher life forms as playing: you can observe it with the small kitten of the big cats or your own kids who have evolved to take it dead seriously. Where playful games may mostly hone individual behaviour for the big cats beyond instinct, with herd animals like humans it evolved into social code, which some have lifted to the level of Gods, others turned into legislation: the main difference between games, ideology, laws or religion isn't as much in the code, as in the scale of its application… which of course depends on the quality of that code. Some games are so successful that they become serious enough to write down their rules. And increasingly we have computers execute them.

Computers started as a faster replacement for mechanical calculators operating on minimally contentious numbers while the first nodes of the internet connected nerds interested in little more than technology. Without measurable impact on society, there was no need for regulation. Today 'Very Personal Computers' and the internet are moving past the inflection point where the majority of all human created content and interaction is not just passing through them, but actively enabled, blocked, reinforced or transformed by code and AI models applying bias, rules and regulation across every hop, from edge to core and back into our brains.

What started as a playfully nerdy version of a college yearbook (Facebook), a better way to rank the relevance of Internet pages (Google), a more convenient way to sell music on a mobile player (iPhone), or an attempt to reduce IT hardware spends off season (Amazon Cloud) has become political, simply because their code is now applied at planet scale throughout the Internet, with little regard to existing borders and regulations. China was one of the first major nations to erect internet border controls against content but, as the digitalization of societies progresses, the code at the core becomes political—even for individuals—because of its vast scope and scale of application.

Politicians are 'influencers' by default and are ever more interested in controlling digital platforms within what they consider their domain. As US web giants search for growth planet wide and functionally ever deeper, this increases the attack surface, while political leaders outside the United States reduce their tolerance to US companies exerting influence, even if those never wanted the political ballast: "just connecting people" becomes political aggression when you turn more minds than an opponent.

And because the direction and impact of digitalization is clearly before us, regulation needs to switch from reactive to proactive and start to plan with a desired target state, best inferred from what already exists outside the digital domain.

To continue demanding that cyberspace and the clouds have a distinct set of rules [3], values or morals from the physical world on the ground is, at best, foolish. It is more likely just intended to protect vested interests; ground staked during the early Wild West of the internet.

Any sovereign nation, and in fact every smaller social aggregate with its own rules, only has two choices going forward: regulate any code that reaches a scale or scope big enough to change society, or lose sovereignty to that code.

When an operating system like Android or a browser like Chrome is used by practically everyone for practically every digital interaction, or an Office suite like Microsoft's handles the vast majority of all structured documents in current use, billions of users send and receive trillions of messages a week with a single app, changing a single line of code becomes political, a developer SCRUM meeting a session of parliament, an application redesign a change of constitution.

The separation of powers and the delegation of legislative and executive functions to choice or chosen individuals built societies and gave birth to nations. In the case of code that has amassed enough traction to change society, or by chance becomes the sole player at a social inflection point, it is no longer just an issue of lacking competition: it risks violating sovereignty. Such code needs to be taken from the very hands and corporations that created it, its fate no longer determined by the CEO but the sovereign, in every form of government.

Global information technology in a divided planet

While technology seems less political than content does, the current trade wars end a rather free global flow of technological ideas and products seen during recent decades. As China is challenging Western digital industry leadership, the United States has resorted to blocking its manufacturing supplies to protect markets it considers its own. Where previously the IT industry was consolidating towards one global player for each of the many puzzle pieces required to build chips and systems, suddenly nobody is left with the ability to produce or maintain the full stack, without at least some ingredients from beyond the trenches that sprung up everywhere.

All of these technology assets, whether they are used for physical manufacturing or digital replication, could be redeveloped within each block, given enough time and budget. But they would initially be inferior in quality and price, and not competitive once a barrier was lifted again. China has made it national policy to develop such a full stack; outside the United States, nobody else currently wants to make that sacrifice.

Whilst an ideal solution might be to give each nation on Earth a critical piece of the IT stack and thus assure global cooperation, criticality changes over time and, in any case, the biggest players have their sights set on dominance.

Europe, and whoever else currently holds a critical piece, needs to guard and aim to expand their assets judiciously in order to retain a minority vote and influence in this struggle by allying with other significant economic blocks. At the same time, Europe must modularize its technology and supply chains to minimize impact as barriers shift and change. All players still need consumers to feed their economies, even if the growth potential of the Chinese domestic market alone is bigger than that of any other newly isolated block.

Open source is a great tool to level an uneven playing field and much easier to move and copy than silicon foundries. But like everything else invented or picked up by humankind, it's also turned into a weapon. More and more it is used to level the competitor's garden into a barren wasteland without any revenue to sustain him.

A weapon's preferred use is actually negotiation and that is what Europe must do, understanding that things can only get more difficult if our economy and thus the strength of our arguments get left behind. This pandemic, the next, climate change and pollution will eventually have everyone against a wall, with only two options: successful negotiation or mutual suffering.

References

[1] Alexandre Menard, "How can we recognize the real power of the Internet of Things?", https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/how-can-we-recognize-the-real-power-of-the-internet-of-things

[2] "Mark Zuckerberg asks governments to help control internet content", https://www.bbc.com/news/world-us-canada-47762091

[3] Ben W. Heineman, "The Conflict Between a Corporation's Global Standards and National Law", https://business-ethics.com/2016/04/05/1514-the-conflict-between-a-corporations-global-standards-and-national-law/

Thomas Hoberg is Technical Director R&D at Worldline, Germany.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: T. Hoberg. The impact of technological evolution… on humans and societies. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 144-149, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

The current education system was designed in the 19th century to support the first industrial revolution. Does it still meet the needs of the 21st century?

Rethinking education

By KOEN DE BOSSCHERE and TULLIO VARDANEGA

The education system as we know it was designed in the 19th century to prepare the workforce for the needs of industry and government, as understood at the time. The basic concept has not fundamentally changed in the last 150 years. It worked well, provided that (i) people had lifelong jobs for which they could be trained at a young age, and (ii) knowledge was mostly shared via specialized books and libraries and changed slowly. Today, (i) schools can no longer prepare students for a lifelong career, and (ii) the ease of internet access and use has shattered schools' strict monopoly on knowledge dissemination. This has profound consequences for the future of education, which should be focused less on delivering static degrees of competence, and more on flexibility, self-directed learning, entrepreneurship and innovation. At the same time, while much younger in age, Computer Science and Computing Engineering education also have to change in order to encompass more profound awareness of the changes that informatics and digitalization are having on the very fabric of our society.

Key insights

- The current education system was designed the serve 19th century industry and society
- Schools have lost their monopoly on teaching; the internet has become one big international school: cheap, open to everybody, 24/7

Key recommendations

- Modern education should be a mix of formal and informal learning, coaching students to become T-shaped professionals who serve the economy and society in the 21st century
- Universities should have two additional important roles: fostering lifelong learning and supporting mechanisms for regional entrepreneurial ecosystems
- It is time to develop digital ethics as a separate discipline and to integrate it into computing curricula.

Compulsory education has a long history. In Europe, it started in the 16th century with the protestant reformation in Germany in which Martin Luther called for compulsory schooling, in order to make sure that Protestants could read the German translation of the bible by themselves. In the 17th century, several protestant territories in Europe and the North American colonies implemented compulsory education for both boys and girls. The first state-wide compulsory education system was installed in Prussia in 1763 (for all children age 5-13). Around the same time, Prussia also invested in the creation of modern universities that used German as the language of instruction, chose rationalism over religious orthodoxy, applied new modes of teaching, and gave a lot of freedom to the professors (academic staff) who could spend part of their time on research. The universities became the centre of German enlightenment in the 17th and 18th centuries [1].

Unleashing the intellectual power of the country's smartest people led to an unprecedented series of leaders in all areas of human cognition: Bach, Mozart, Schubert, Mendel, Freud, Engels, Marx, Kant, Nietzsche, Bonhoeffer, Ratzinger, Euler, Gödel, Gauss, von Liebig, Kekulé, Koch, Clausius, Boltzman, Hertz, von Helmholtz, Röntgen, Planck, Einstein, ... This intellectual development also led to many technological innovations [2].

In the 19th century, compulsory education became mainstream in most western countries and was universal by the time of the Second World War. Compulsory education was no longer inspired by religion, but by the needs of the industrial revolution: the availability of lots of workers with the right skill set. This explains why the modern compulsory education system is organized like a factory assembly line: children all enter at the same age and, year-by-year, they learn a standard set of facts, skills, competences and attitudes. Children and adolescents are also disciplined as if they worked in a factory: schools run according to a strict daily and weekly schedule; arriving late is not permitted. There is also quality control: if they fail the test for a particular year, they have to retake the year, or

switch to a different track. Some countries organize standardized tests at crucial transition points in the name of 'guaranteeing' quality. Out of such schools comes a steady stream of standardized and government-approved workers: doctors, lawyers, teachers, nurses, construction workers, mechanics, ICT workers, ….

In every country, public education is a very important department of government because it is not only large-scale but also politically sensitive, as this is the place where the future generations are being formed, and where the future of society is shaped. The focus often depends on the government that is in power: more or less focus on nation building, on integration of under-represented groups, on STEM education, on religious or cultural formation, on entrepreneurship or on excellence, for example. Furthermore, many countries honour the freedom of education which means that parents have the right to have their children educated in line with their personal views (political, religious, social, language, …) and without intervention of the nation state (i.e. via private schools or home schooling). In the 20th century, a number of governments experienced difficulties because they could not agree on education policy.

At the transition between the 20th and 21st centuries, the education model has changed minimally: (i) compulsory education has been extended in most countries (lowering the age of entry and raising the age of exit to 18), and (ii) the body of knowledge designed for learning has been continuously updated. In an evolution that started after the Second World War, many students started higher education because secondary education was no longer considered a sufficient basis for a prosperous career. Today, for many young adults, the school career ends when they obtain a Bachelors or Masters degree. Fifty years ago, an increasing number of graduates with a Masters degree started to undertake a PhD, leading to increasing numbers of PhD degrees. This model has served society and industry very well over the last 150 years, allowing countries to exploit well the intellectual potential of the population.

Schools are no longer the only option for education

Schools focus on formal learning: the teachers explain the material, and the pupils prove that they understand it via testing. Eventually, pupils obtain a diploma that proves that they completed a particular study programme. Schools have a historical monopoly on formal learning but, with the rise of the internet, they now have strong competition [8,9].

The under-18s of today do a lot of non-formal (structured learning outside school, e.g. learning to play a musical instrument) and informal learning (in daily life) about their interests, hobbies and the world around them. A significant amount of non-formal and informal learning happens on the internet, which offers many opportunities for self-directed or incidental learning. The material is often presented in a way that is both very attractive (short movies, animations, games, demos, …) and fun, and does not feel like learning.

In Europe, the complete educational track from preschool to final graduation is

funded by the government (at least for the public schools). Governments start paying from the age of 2-3 years when children enter preschool, until they obtain their highest degree: up to 20 years in the case of a Masters degree. Such a period equates to almost 50% of the span of a full career; it hence represents a huge societal investment in every single newborn. Some countries encourage universities to charge tuition fees to cover part of their costs. In countries with high tuition fees, it is common to finance a higher education degree with a student loan that has to be paid back. The offering of the internet is available 24/7, and is very cheap compared to the cost of schools. It is an attractive option for people who cannot afford the traditional option. The offering is overwhelming in its breadth and range, and is growing daily.

Future education will be a mix of formal, non-formal and informal learning

Today's globalized world is called VUCA: volatile, uncertain, complex and ambiguous [13]. Change is accelerating; there are no longer lifelong guarantees. Unfortunately, schools are all but VUCA: they are very structured and predictable, they simplify things and avoid ambiguity, they offer a very protected environment for students and seldom require them to leave their comfort zone. This is quite different from the real world.

Despite all the efforts to make them non-VUCA, schools do not work equally well for all pupils. It goes too fast for some, too slow for others. Some have problems with the strict daily and weekly routine, and some encounter difficulties when they hit puberty. The one-size-fits-all model is not the best choice for all children. Thanks to freedom of education, parents who can afford to do so can decide to send their children to a private school, where they can get a more a personalized education.

In the real world, the evolution of science and technology goes at breakneck speed, and the knowledge base delivered by school education is rapidly and increasingly proving insufficient. What is certain is that it will not serve a complete career. Today, the half-life time of knowledge in some disciplines is less than ten years [3]. Much of the information that people learned in school twenty years ago has been refuted by new scientific insights. As a result, learning cannot stop with graduation. Adults will have to further develop their knowledge and competences throughout their lifetime.

Obviously, we cannot send adults back to school to update their knowledge after a number of years. Instead, we must work with non-formal learning and informal learning. But then the question is: why wait until graduation to start lifelong learning? And why not start much sooner with non-formal and informal learning? In other words: is twenty years of uninterrupted formal education the best preparation for a career in the 21th century? It probably isn't and, if it isn't, which things should be part of formal education, and which could be part of the non-formal and informal education of children?

Formal education might be the best option to learn the basics of the established disciplines: mathematics, physics, biology, chemistry, languages, history, culture, economy, …. When studying the basics, study programmes should not focus too much on teaching solutions (which are by definition changing), but instead focus on reasoning and on the fundamental principles of the discipline, which have a much longer half-life. Furthermore, it is the fundamental principles that are needed to develop outside-the-box solutions in the future. Learning about the solutions can be done more easily in a non-formal or informal learning setting by reading, watching documentaries, and undertaking internships and voluntary work, often on a need-to-know basis in the context of project work. Making students partially responsible for their own education will lead to them being responsible for it after graduation too.

At the competence level, study programmes should focus on the eight key competences for lifelong learning as adopted by the European Parliament in 2018 (Figure 1) [4].

Notable in that recommendation is the focus on science, technology, engineering and mathematical competences, combined with digital skills as key competences for all

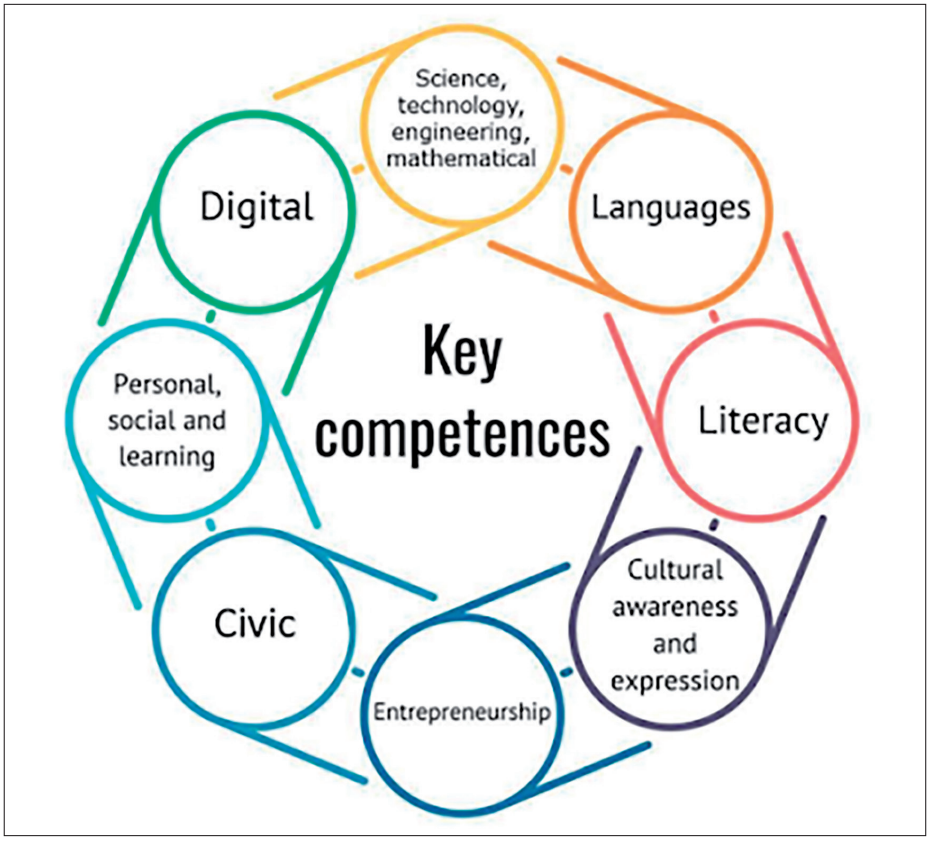


Figure 1: Key competences for lifelong learning [4]

citizens in Europe. The focus on entrepreneurship in combination with the soft skills of personal, social and learning competences is intended to make Europe more competitive. The combination of the 'cultural awareness and expression' competence with civic competences should provide all Europeans with a common framework for values, democracy, globalization and multiculturalism. Finally, literacy and (foreign) languages and culture are important as a means to learn, listen and express ideas. These eight key competences are essential for personal fulfilment and development, employment, social inclusion and active citizenship. They break with two legacy traditions that have burdened formal education worldwide since the 20th century: the dichotomy between the humanities and the sciences, and the dichotomy between pure and applied training [5].

Towards a new role for universities

There has been a constant evolution in the role of higher education. Originally, universities were pure teaching (viz. transmissive) institutions in which professors passed on to the next generation the accumulated knowledge of the previous generations.

Later, professors were encouraged also to do personal research, and to create new knowledge. With the appearance of research universities, all professors were required to be active researchers. This requirement went hand-in-hand with the development of a number of research degrees (Masters and Doctorates). Research universities have contributed enormously to the development of the modern 20th century world. Many of the things that we take for granted today have been developed at (or in collaboration with) university laboratories. The availability of large numbers of graduates with research degrees has also led to a professionalization of industry and government agencies.

More recently, universities have been encouraged to broaden the I-shaped profiles of their graduates into T-shaped profiles (Figure 2) [6].

This change means that students should have a broad base of general supporting knowledge and skills, supplemented with deep knowledge and skills in one or more areas. In the broad base, the student must learn complex problem solving, critical thinking, creativity, people management, coordination with others, emotional intelligence, judgement and decision-making, service orientation, negotiation and cognitive flexibility [7]. These are the competences that set humans apart from computers and robots. The deep knowledge and skills element must encourage the student to learn how to push forward the state of the art in a subject, and to create new knowledge and to innovate. The harder students are pushed to stretch themselves in the deep elements, the more they will learn, and better placed they will be for the challenges of the 21st century. One thing is certain: there is little value in receiving specialist training only to end up doing routine tasks. Such jobs are disappearing because they are the easiest and fastest to automate. T-shaped education offers better guarantees for a life of self-fulfilment and wellbeing.

T-shaped profiles can also help to reduce the shortage of ICT workers in Europe (which is estimated at 1 million). It are not only computing specialists that stand to broaden and update their competences; other disciplines could also broaden their graduates' skill sets by providing them with a basic understanding of computing, big data analytics and artificial intelligence, on a par with the prescribed basics of sciences, history and foreign languages. Such graduates could easily contribute to the ICT sector in roles that are technically less demanding. There will be a need for a broad range of graduate profiles because society increasingly depends on ICT (for digitizing industry, securing ICT systems, designing smart grids for the transport of renewable energy, development of precision agriculture to reduce the use of pesticides and irrigation, etc.). This measure would also improve the gender balance of the ICT sector in general.

Another recent evolution is that universities are encouraged to actively monetize their research via IP portfolios and creation of spin-offs. All universities now have a technology transfer office tasked with helping researchers to protect their intellectual property and to exploit it, either via an agreement with existing companies, or via the creation of a spin-off company.

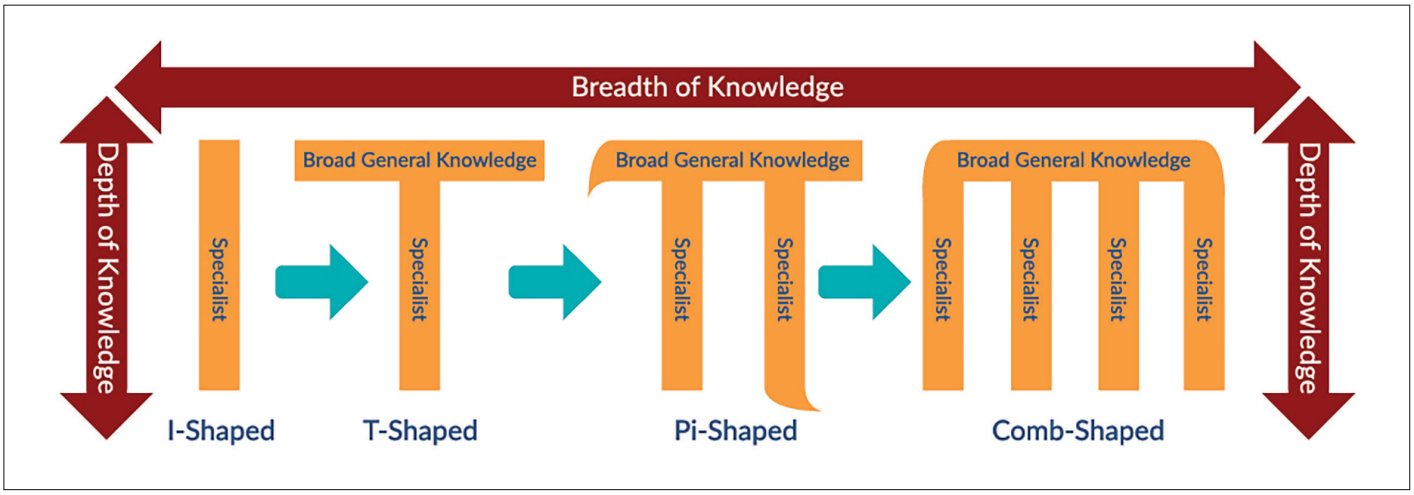


Figure 2: Broadening I-shaped to T-shaped profiles [6]

In some places, this evolution has led to two additional changes:

- The active promotion of an entrepreneurial mindset in students, via curricular activities that stimulate them to leave their comfort zone, and try to think more like an entrepreneur.
- Universities becoming an active partner in the regional entrepreneurial ecosystems, to support and develop them by creating spin-offs and by training the next generation of entrepreneurs through study programmes. This activity also creates attractive jobs for its highly specialized graduates (often with PhDs). This job market can slow down the brain drain to other parts of the world, and guarantee that Europe maintains a critical mass of expertise in important sectors.

This is a transition through which traditional research universities become entrepreneurial universities that produce not only graduates and research results, but also innovations and economic prosperity. This is the response of higher education to the needs of the post-industrial economy.

The need for digital ethics

Humans have lived through pre-history (where all was verbal and the clan was the information agent) to history (where all that someone has been able and allowed to write is written down: this is where the state is the information agent), to hyper-history [10] (where machines automatically produce the majority of data and have become the primary information agent).

In its acting as an information agent in hyper-history, ICT does more than actors did in history: it "wraps" the world in a digital envelope designed to facilitate computerized operation within it. We humans are being progressively absorbed into that envelope of which we are less and less in control. A few examples illustrate this situation.

The combined effect of exponential increase in processing power and equally exponential decrease in processing cost is the generation of an exponential amount of digital data. Exponential phenomena are problematic: when you begin to notice them, it is because they have become sufficiently large to catch the eye. Yet it then only takes a few more steps for them to ramp to gigantic proportions, entirely beyond control. The quantity of data being produced today already vastly exceeds the available storage capacity, and this will only worsen in the future. This means that massive amounts of data should be deleted daily or never even recorded. Deciding what to delete or what to skip is not a merely technical decision if that data is subsequently used to train AI systems, which, increasingly, is tantamount to "writing history", because deletion or skipping may cause harmful bias in the learned inferences.

Digitalization has broken the link between Place and Law. Geographical space and cyberspace do not coincide: the old legislative foundations no longer apply and the spreading of digitized information in cyberspace is very hard to control. Not surprisingly, therefore, some legislators react to this situation by pushing forward the notion of "data sovereignty", which enforces the rights of the state (Place) over the data originating within its boundaries. Examples of such efforts are the GDPR and the recently invalidated EU-US Privacy Shield [11]. This move, however, goes counter to the original fabric of the internet, the web and the logic of their pipelines. Adopting sovereignty-driven norms calls for a rethink of the architecture of the internet and of the web above it.

Current ethical frameworks often assume that there are humans, or at least living biological beings, involved in an action. When two computers attack each other with malware, there are no living beings involved in the action, but malicious people are at the origin of the problem. This is why digitalization requires us to reconsider existing ethical frameworks and to adapt them for the various world envelopes that ICT keeps creating. There is a need to develop digital ethics as a discipline (similar to e.g. bio-ethics or medical ethics). The fact that some universities are setting up chairs and institutes in digital ethics is a promising evolution.

We are at a point at which computer engineers who develop hardware and software for decision-making systems should be more acutely aware of how ethics "flows" in their systems, e.g. the fact that a seemingly insignificant decision to store only relevant events in a database and leave others out might eventually lead to a bias in systems that use the database. Unfortunately, however, current computer science and computer engineering education scarcely extends to understanding the workings of ethics. We therefore believe it is necessary to include the basics of digital ethics in computing curricula.

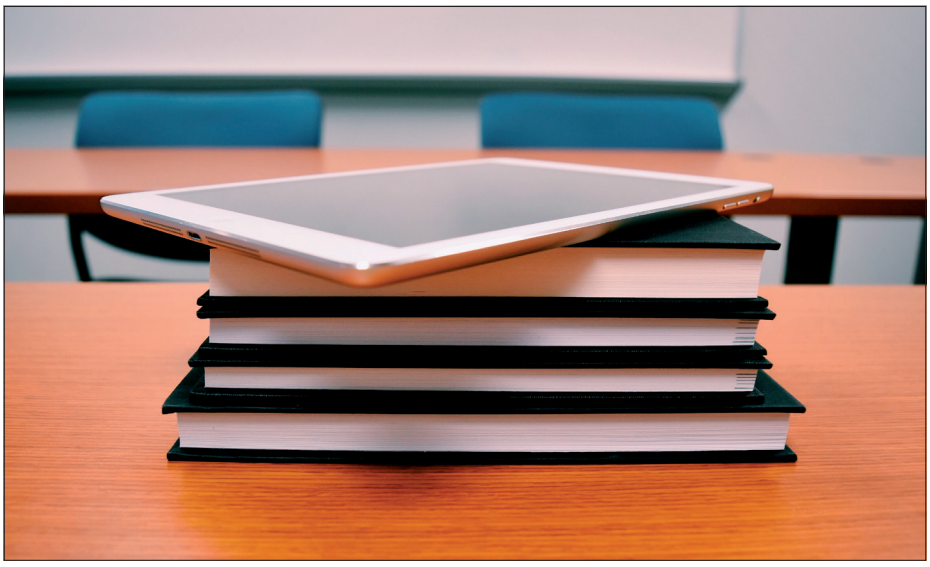
This objective will very unlikely be achieved by adding siloed Ethics, Philosophy or Law optional courses within computer science (CS) and continuing education curricula. Such courses risk being isolated and not "germinating" in the students' mind. It is reported [12] that top-quality schools have begun recruiting philosophers in their CS departments and had them contribute to the design the syllabi of curricular courses. This is a promising approach that Europe should explore further.

Conclusion

Education has always been important. Excellent education is key to solving the global challenges of the 21st century. Without a well-trained workforce, Europe will not be able to compete with the rest of the world. Human minds are the most important resource that we have in Europe.

Europe was the birthplace of modern education, and it has one of the best education systems in the world: free and excellent education for all children up to the age of 18, and affordable higher education. Thanks to its education system, Europe has one of the best-trained and most professional workforces of the world. The education system is also an enabler of social mobility. It is an asset that we should cherish, protect and make future-proof. This requires us to adapt it to the changing reality and evolving needs, and making it capable of driving future progress instead of undergoing it. It is the only way to stay on top in a knowledge-based global economy.

It is time to develop digital ethics as a separate discipline and to integrate it into computing curricula.



References

[1] "Martin Luther University of Halle-Wittenberg", https://en.wikipedia.org/wiki/Martin_Luther_University_of_Halle-Wittenberg

[2] Peter Watson, "The German Genius", https://www.amazon.com/German-Genius-Renaissance-Scientific-Revolution/dp/0060760230

[3] Samuel Arbesman, "Half-life facts", https://www.wired.com/2013/03/half-life-of-facts/

[4] "EU Key Competences for Lifelong Learning 2018", http://keepcalmandteachenglish.blogspot.com/2018/03/eu-key-competences-for-lifelong.html

[5] World Economic Forum, "The Future of Jobs", http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf

[6] Jayne Groll, "From I-Shaped to T-Shaped – Why DevOps Professionals Need to be Multi-Skilled", https://devopsinstitute.com/2017/11/15/from-i-shaped-to-t-shaped-why-devops-professionals-need-to-be-multi-skilled/

[7] World Economic Forum, "The 10 skills you need to thrive in the Fourth Industrial Revolution", https://www.weforum.org/agenda/2016/01/the-10-skills-you-need-to-thrive-in-the-fourth-industrial-revolution/

[8] Neil Selwyn, "The Internet and Education", https://www.bbvaopenmind.com/en/articles/the-internet-and-education/

[9] D. Schugurensky, "The forms of informal learning: towards a conceptualization of the field", https://pdfs.semanticscholar.org/6315/0f9e5376503715b1c0175f2a5354dd78fbcf.pdf

[10] L. Floridi, "Ethics & Social Science: Ethics in the Age of Information", https://www.youtube.com/watch?v=lLH70qkROWQ

[11] V. Manancourt, "EU court ruling strikes hammer blow to transatlantic data flows", https://www.politico.eu/article/eu-court-ruling-strikes-hammer-blow-to-transatlantic-data-flows/

[12] M. Wagner, "An AI and Computer Science Dilemma: Could I? Should I?", https://informatics.tuwien.ac.at/news/1896?utm_campaign=feed&utm_term=news

[13] "Volatility, uncertainty, complexity and ambiguity", https://en.wikipedia.org/wiki/Volatility,_uncertainty,_complexity_and_ambiguity

Koen De Bosschere is Professor in the Electronics department of Ghent University, Ghent, Belgium.

Tullio Vardanega is Associate Professor in the Department of Mathematics of the University of Padua, Italy.

This document is part of the HiPEAC Vision available at hipeac.net/vision.
This is release v.1, January 2021.
Cite as: K. De Bosschere and T. Vardanega. Rethinking education. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 150-155, Jan 2021.
The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

The three largest economies in the world are the United States, the European Union and China. They are roughly based on three different philosophies: companies-first, people-first and government-first.

Europe should be the humans-first continent

By KOEN DE BOSSCHERE

The three strongest economic blocks are home to three different political systems: capitalism and freedom for the United States, social democracy in Europe and a government-controlled regime in China. Another way to characterize them is companies-first, people-first and government-first. The United States and China are competing for global dominance economically, politically, militarily, scientifically, in cyberspace, ... What is the role of Europe in this epic battle?

Key insights

- The integration efforts of the European Union have created one of the three largest economies in the world, enabling EU leaders to play an important role in international decisions.
- The European Green Deal is not only a growth strategy for the European economy, but also for the European computing industry. Sustainability requires a lot of monitoring and optimization to save resources, and this will always require some form of computing.
- Europe has the potential to create and maintain the best trained workforce in the world.

Key recommendations

- Europe should continue to invest in education and training and stay competitive with the rest of the world.
- Europe should create a solid digital ethics framework to assess the introduction of new technologies in Europe.

The United States actively promoted its values after the Second World War, and took a role of global leadership during the Cold War. Neither Europe nor China were at that time promoting their own values on a global scale. China was, for a long time, a closed society with its own internal issues. It was only when globalization started in the 1990s that it became the factory of the world and experienced unprecedented economic growth. Now that it has joined the top three world economies, it wants to play a more prominent role in various ways: economically, politically, militarily, scientifically, in cyber space, … It is rapidly expanding its influence in each of these domains. China is building political relationships with many countries in the southern hemisphere, and is building a global transportation system (the belt and road initiative). China is currently working diligently on a new world order, in which it plans to play a prominent role.

After having been centre stage of two world wars, Europe needed time to let the wounds of war heal, and it had to deal with the effects of decolonization, the Cold War in its backyard, the fall of the iron curtain and, finally, the reunification of the continent. Now that the healing process has come to an end in large parts of Europe, the time has come for it to speak for itself, to promote its values and to claim leadership in areas where it is strong. If it fails to do so, it might eventually become crushed between two superpowers, or it might become an appendix of the Eurasian continent, and just one of the world's many cultural holiday destinations.

Since its creation in 1993, the European Union has steadily been working on its integration: development of the internal single market, European Union law, a common visa policy in the Schengen area, a common currency in the Eurozone. All these accomplishments required time, and delicate political agreements. Only the older generation still remembers how divided Europe was when it came out of the Second World War: one currency per country, no single market, the iron curtain and the threat of the Cold War, compulsory military service for young men, several European countries run by dictators. A lot has changed for the better.

In the last decade the European Union has developed a common Foreign and Security Policy, and attempts to speak with one voice to hold more weight in matters of external relations and defence. It now has diplomatic missions all over the world, and at the United Nations, the G7 and G20. Important decisions are however not made by a president, or by a parliament, but by 27 member states, and require lengthy negotiations. Europe is today an economic union, but not yet a political one. The lack of political union became clear during the debt crisis in parts of the Eurozone, in the

migration crisis at the southern borders of Europe, and in the discussions about the COVID-19 relief fund. Due to its large diversity (religion, language, culture, living standards, …), political debate in the EU is not going to disappear anytime soon. Fortunately, there is a lot of diplomatic talent in Europe, and integration moves forward continuously, albeit slowly at times.

The European Union was unanimously awarded the Nobel Peace Prize in 2012 "for over six decades having contributed to the advancement of peace and reconciliation, democracy and human rights in Europe". This is an important accomplishment after around fifty wars of varying size in Europe since 1800. Europe is today one of the best places to live in the world, and it remains an attractive place for new EU candidate countries, and for over a million people a year who seek to enter Europa and to start a new life here.

At the same time, there are also many Eurosceptics living in the union. They mostly complain about the loss of sovereignty and control and about the fact that integration moves too fast. The UK became the first major country to leave the European Union. The negotiations leading to the withdrawal agreement made clear how strong the economic integration is in practice, and how difficult and costly it is to leave the Union (even without being part of the Eurozone or the Schengen Area). European politicians will learn a lot from this experience, and will hopefully be better prepared for the other major challenges like immigration, growing inequality, ageing population and climate change.

Notwithstanding all the political difficulties, Europe is gradually gaining global influence, and it tries to retain a seat at the table on international decisions. Well-known are the antitrust cases against large international corporations like Microsoft (2004-2008: €2 billion), Intel (2009: €1 billion) and Google (2017: €2.4 billion) [1].

Another illustration of EU influence is the introduction of the General Data Protection Regulation (GDPR) which put control over personal data back where it belongs, namely in the hands of the citizen. The GDPR turned out to be rather powerful because it did not just have an impact in Europe: it forced everybody who wants to do business in Europe to comply. Thanks to the GDPR, we can now opt-out of unsolicited mails with a single click. Of course, individual European countries would never have been in a position to impose such a legal framework, but the European Union could. But there are more domains in which Europe could take a leading role. Why not try to protect Europeans from disinformation, fake news and hate messages on the internet?

Towards a 100% sustainable Europe

For many years, sustainability was only relevant if it did not conflict with profitability. Engineering schools taught that the products made from raw materials were of a better quality than products made of recycled materials. Business schools taught that it was better for the economy to replace a consumer product than to repair it. After the publication of the United Nations Sustainable Development Goals, an increasing number of people started to realize that sustainability is no longer a "nice to have" but a necessity for maintaining our living standards, and those of the generations to come. The brilliance of the SDGs lies in the fact that they span 17 domains, and most people will find at least a couple of domains that they care about (climate, poverty, biodiversity, gender equality, justice, …). The SDGs have created a framework, which is, for the first time, taken seriously by politicians, universities, and businesses. It is no longer acceptable to ignore sustainability.

Economists have investigated how to create business models for a sustainable economy and discovered that e.g. decarbonizing the economy is a once-in-a-lifetime investment opportunity because renew-

able energy is getting cheaper by the day, while fossil fuels are getting more expensive (harder to exploit, more investments needed to protect the environment, …). Furthermore, fossil fuel reserves are finite, and will eventually have to be replaced by renewable energy sources. That process will require a lot of innovation, technology and investment. Holding off on making the transition will only serve to give a head start to other countries developing the technology before us, and to help them become global leaders in sustainable technology.

Therefore, Europe did the right thing by resolutely promoting the European Green Deal. There is no time to lose if we want to stay relevant in the future. By working on the European Green Deal, we will be able not only to reduce Europe's ecological footprint but also to develop sustainable technologies that we can sell across the world. Rather than hurting the economy, implementation of the Green Deal will help it to transition to a more sustainable model. Europe and the EU member states will have to help businesses to make this transition.

The European Green Deal is an opportunity for the European computing industry. Sustainable processes are optimized processes (less energy consumption, less waste, …) and such optimizations are only possible with the help of lots of computing, CPS, IoT, …. Hence, the European Green Deal is a growth strategy not only for the economy as a whole but also for the computing industry specifically, and we should therefore embrace it.

Towards excellent education for all

Europe has an excellent education system. Higher education is more affordable than in the United States and, in the top one hundred best universities worldwide in the 2020 Times "Higher Education Ranking", Europe has 37 institutions (North America has 45, and Asia 18) [2]. Unfortunately, since 2016, Europe has lost five universities in the top one hundred. Two places went to the United States and three went to Asia; European universities are experiencing competition from the rest of the world. Another noteworthy observation is that the majority of the 37 European universities are located in countries which have historically been predominantly protestant.

Country	#100	#50
United Kingdom	11	7
Germany	8	3
The Netherlands	7	0
Switzerland	4	2
France	3	1
Sweden	1	1
Belgium	1	1
Finland	1	0
Denmark	1	0
Total	**37**	**15**

In terms of absolute numbers of tertiary graduates, Europe and the United States are stagnating, while China has grown from two million per year to 15 million per year over a period of just 15 years [3]. In terms of relative numbers, China now has more tertiary graduates per 1,000 of the population than Europe (Figures 1, 2). It is clear that China is preparing its future as a scientific powerhouse and this workforce will fuel and accelerate its innovation potential in the coming years. Europe should invest more in tertiary education because it will need the innovation potential of these people in order to stay competitive on the world stage.

European universities produce on average more PhDs per 1,000 of the population than the United States or China, even in science

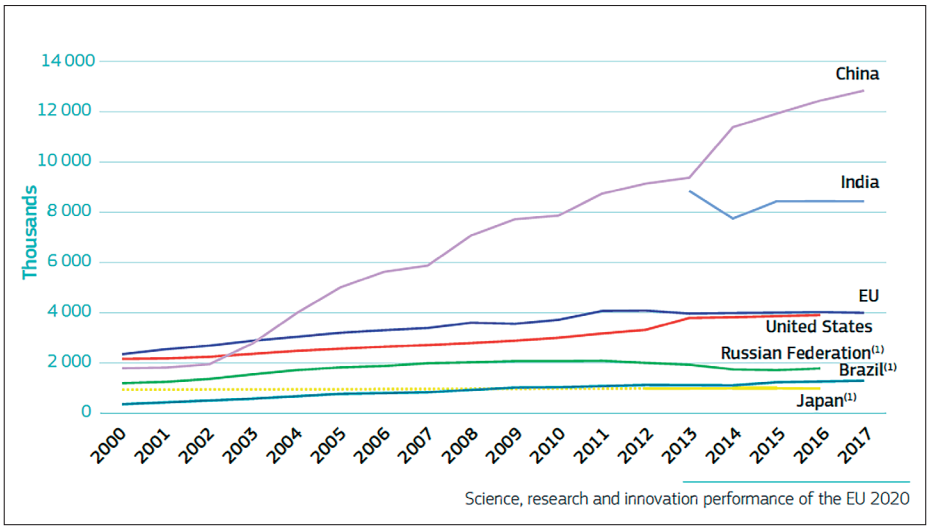


Figure 1: Total number of tertiary graduates, 2000-2017

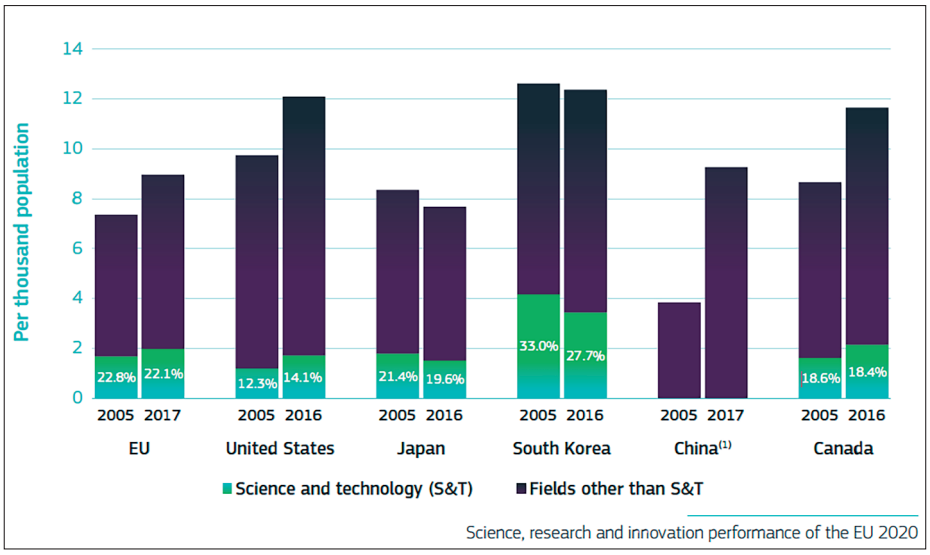


Figure 2: Tertiary graduates per thousand population broken down by science and technology and other fields, 2005 and 2017

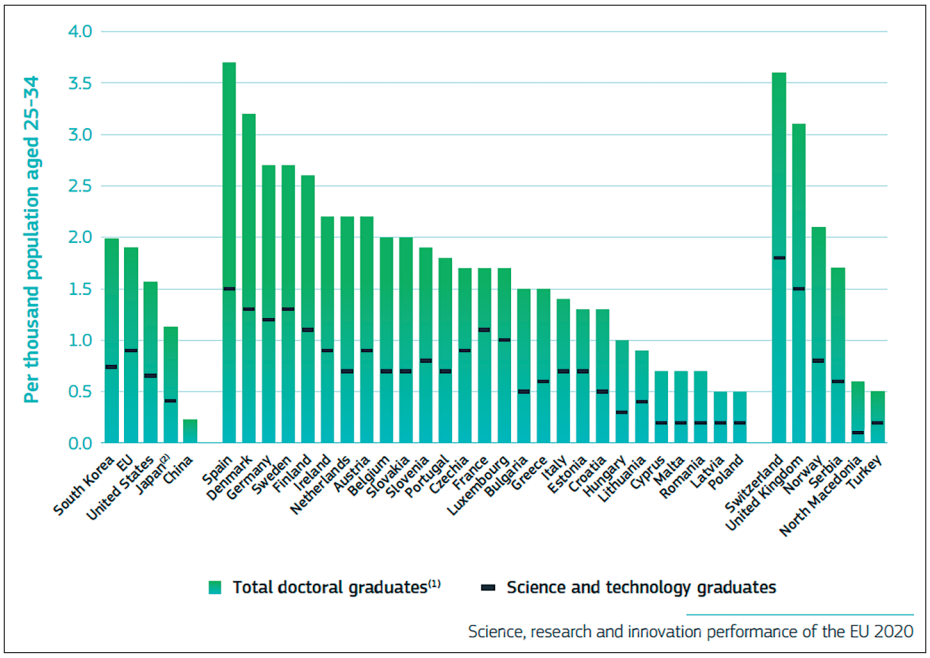


Figure 3: New doctoral graduates per thousand population aged 25-34, 2017

and technology. Many individual European countries do better than the United States. This is therefore a clear strength (Figure 3).

Therefore, it is important for Europe to keep investing in education. Before the industrial revolution, the landmass of a country was an indicator for its economic production (because agriculture requires land); after the industrial revolution, it was the size of the population that determined the scale of industrial production, and the size of the army that could be mobilized. Nowadays, it is the number of well-trained knowledge-based workers that have the biggest impact on the economy: innovators, scientists, entrepreneurs, … Europe will never be able to beat China in terms of the absolute number of graduates, but it could do better than the United States because it has a larger population. A better trained workforce is good not only for the economy: people with higher levels of education have a better income (all higher education degrees lead to a higher than median income in the United States, and lower than median unemployment rate; the situation in Europe is similar) [4]. They also have better health and longer life expectancy, care more about sustainability, are politically more engaged, and participate more in cultural events. Investment in raising the education level of a population pays for itself even in the short term. However, Europe should be aware of the "brain drain": highly educated people often move to the United States for better working conditions and salaries, and they are highly appreciated because of their good education, especially in the domain of AI.

Create a solid digital ethics framework

Computing has become such a powerful commodity that we should start thinking about whether everything that can be done should be done. Decades ago, similar questions led to the establishment of disciplines like medical ethics, bio-ethics, business ethics, military ethics and so forth. It is now time to invest in digital ethics as a discipline, and to make sure that all professionals in computing receive basic training in it. The creation of cyber armies in many countries might also call for some form of regulation.

Digital ethics is not a new concept; in fact, it was first touched upon in the mid-1940s by Norbert Wiener, who coined the term cybernetics in his book "Cybernetics: or control and communication in the animal and the machine" (1948). At the time, it was not taken very seriously by the scientific community. The last decade has, however, witnessed a sharp increase in interest in digital ethics.

The key problem is that, whereas in the past computers did the calculations and humans made the decisions based on the calculations, this is not longer entirely the case. Decision makers are assumed to have an ethical framework to guide them in the decision-making process. With artificial intelligence, computers not only do the calculations, but also make the decisions, indirectly guided by humans during the building of the learning database in the case of deep learning, the most common implementation of AI. At the moment, these are small decisions but the expectation is that they might also be asked to make important, even life-changing decisions in the near future, as in the example of a self-driving car. This means that the people who make the decision-making algorithms have to follow an ethical framework to guide the decision-making process. Not all computer scientists have developed an ethical framework comparable to that of people involved in building public policy, for example. In some cases, they are not even to be blamed, as the bias comes from the data that was used to train the system (also compiled by humans), and not from their code.

Some universities have established centres for digital ethics (for example, the Digital Ethics Lab of the Oxford Internet Institute [7] founded in 2017, and Center for Digital Ethics and Policy of The Loyola University of Chicago [8]). As mentioned in "Rethinking Education", modules on digital ethics are being introduced into several computer science courses in order to ensure that graduates have a basic understanding of the ethical aspects of their profession.

There seems to be more interest in digital ethics in Europe than in the United States or in China, so this is an opportunity to lead. A solid ethical framework, based on the most recent scientific insights should be developed and used as a touchstone when introducing new technologies in Europe, irrespective of whether the technology was developed in Europe or not. Europe can lead in this area, like it is leading in the protection of the privacy of its citizens with the GDPR.

European Commission priorities

The European Commission has set forward six priorities for 2019-2024 [6].

Image: ID 44916222 ©Björkdahl Per | Dreamstime.com

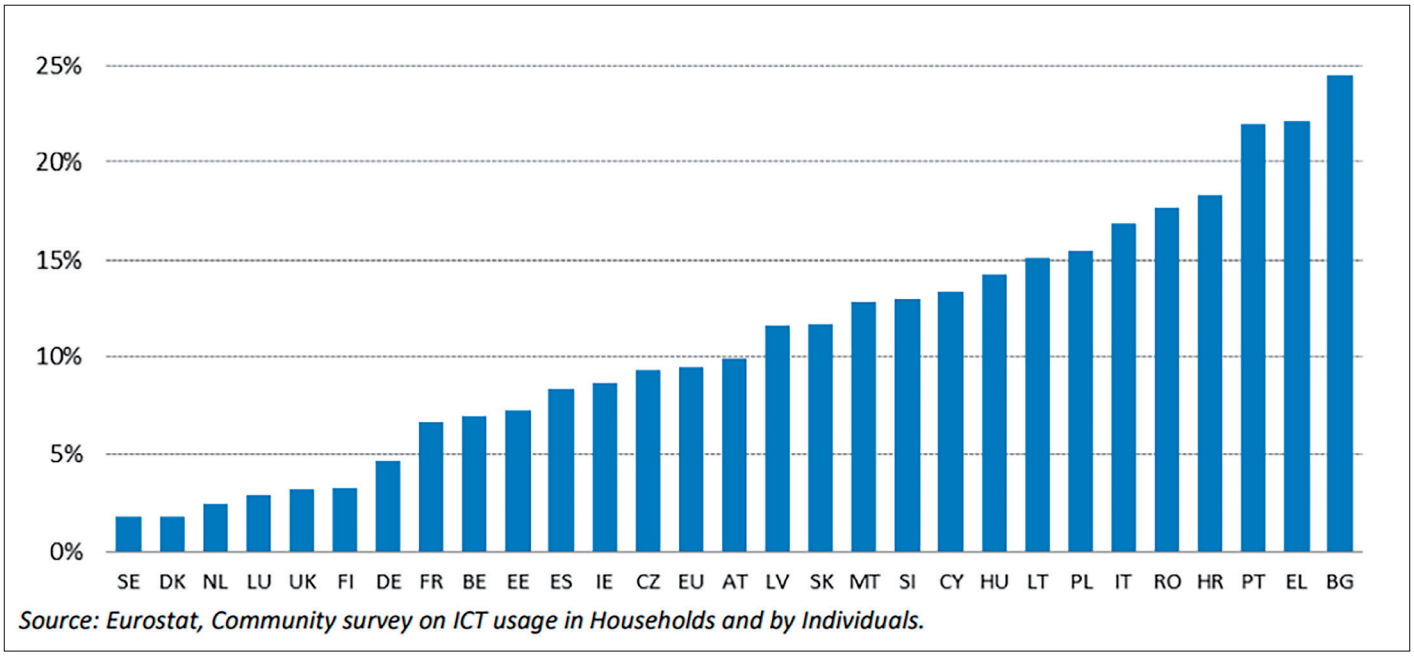


Figure 4: People who never used the internet (% of individuals), 2019

A European Green Deal: Europe aims to be the first climate-neutral continent by becoming a modern, resource-efficient economy. At the same time, it also wants the Deal to become the growth strategy of the European economy, and the source of millions of jobs in the sustainable economy.

A Europe fit for the digital age: The EU's digital strategy will empower people with a new generation of technologies. COVID-19 has shown how critical the digital infrastructure is, and how important it is for digital participation to be inclusive. Affordable broadband internet access for all is essential. Without access to the internet, citizens can no longer fully participate in society [5]. In some countries there is still a long way to go (Figure 4).

An economy that works for people: The EU must create a more attractive investment environment, and growth that creates quality jobs, especially for young people and small businesses. To create more and better jobs all over Europe, we need more startup companies and SMEs. Europe should not promote international companies with zero hour contracts but insist that all the jobs created in Europe are decent jobs. There should be no working poor.

A stronger Europe in the world: The EU will strengthen its voice in the world by championing multilateralism and a rules-based global order as advocated in this contribution.

Promoting our European way of life: Europe must protect the rule of law if it is to stand up for justice and the EU's core values as advocated in this contribution.

A new push for European democracy: We need to give Europeans a bigger say and protect our democracy from external interference such as disinformation and online hate messages. Democracy is best protected by investing in solid education for young people, by offering them a decent job, and by fighting disinformation and fake news on the internet.

Several of these priorities boil down to caring for everybody, and especially for those with the least resources at their disposal. It is important that nobody is left behind, that inequality does not lead to a polarized society, and that people are protected from the adverse effects of technology.

References

[1] "Top 5 antitrust fines handed out by EU", https://www.euractiv.com/section/competition/news/top-5-antitrust-fines-handed-out-by-eu/

[2] "World University Rankings", https://www.timeshighereducation.com/world-university-rankings/2020/

[3] "Science, research and innovation performance of the EU 2020", https://ec.europa.eu/info/publications/science-research-and-innovation-performance-eu-2020_en

[4] Elka Torpey, "Measuring the value of education", https://www.bls.gov/careeroutlook/2018/data-on-display/education-pays.htm

[5] "The Digital Economy and Society Index", https://ec.europa.eu/digital-single-market/en/desi

[6] "The European Commission's priorities 2019 – 2024", https://ec.europa.eu/info/strategy/priorities-2019-2024_en

[7] Oxford Internet Institute, Digital Ethics Lab, https://www.oii.ox.ac.uk/research/digital-ethics-lab/

[8] Center for Digital Ethics, "Welcome to the Center for Digital Ethics & Policy", https://www.digitalethics.org/

Koen De Bosschere is Professor in the Electronics department of Ghent University, Ghent, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: K. De Bosschere. Europe should be the humans first continent. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 156-161, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

Europe is currently no superpower in computing, but it is never too late to try to become one.

The position of Europe in the world

By KOEN DE BOSSCHERE

In order to set out a strategy for the future, it is important to know one's strengths and weaknesses, and it also helps to see opportunities, and to prepare for threats. Although Europe is currently no superpower in computing or in B2C ICT business, it is never too late to become one. For that to happen, it is important to understand the obstacles that make it difficult to grow global companies, and to develop a policy to remove them.

In this article, we present a SWOT (strengths, weaknesses, opportunities, threats) analysis of the European computing systems ecosystem. We make a distinction between three stakeholders: (i) publicly funded universities and research institutions ("Science and Technology"), (ii) the computing industry and its market, and (iii) the local and European governments responsible for creating an environment in which research, innovation and commercialization can take place ("Policy and Government"). Most data in this article comes from "Science, research and innovation performance of the EU 2020" [1].

We start with the strengths and weaknesses.

	Strengths	Weaknesses
Science and Technology	• High-quality education • Excellent research • World leader in lithography for semiconductor manufacturing	• Weak academia-industry link • Strong in research, but not in commercialization
Industry and Market	• Second largest market in the world • Stronger in systems than in components	• EU ICT contributes less to GDP than in other advanced countries • Lack of venture capital culture • Lack of advanced foundries
Policy and Government	• Common market • Decent public funding level of research	• Lack of ICT workers • Fragmentation of funding

Key insights

- Europe's global economic impact is dwindling because other regions are growing faster due to their demographics, rapid economic development, or abundance of natural resources. The biggest long-term challenge for Europe is to sustain economic growth with a shrinking active population and growing costs for social security. Maintaining the current standard of living will require a highly-educated and productive workforce.
- It is important for Europe to realize that computing is a key enabling technology of strategic importance because it is at the basis of all modern smart products and services. Europe should never lose the capacity to build its own computing solutions.
- Europe is a scientific powerhouse, but it fails to monetize some of its research results due to the lack of entrepreneurial talent, venture capital and a large enough ICT workforce.

Key recommendations

- The creation of well-funded international competence centres will help to retain and attract top talent, and to stay at the forefront of new digital technologies.
- Europe should continue to invest heavily in research and innovation, and in a more entrepreneurial Europe that generates lots of start-up, scale-up and global companies. Europe needs more venture capital to support the growth of scale-ups.
- Europe should invest in future-proof areas: the silver economy to support the ageing population (home automation, health, entertainment, …), technologies for sustainability (low power, recycling, …). Technologically, hardware accelerators, and artificial intelligence are key elements of future computing systems.

High-quality education

As mentioned in the article on "Europe should be the humans-first continent", Europe has an excellent and affordable educational system from preschool to university. According to the 2020 Times "Higher Education Ranking" [2], more than one third of the top 100 universities are located in Europe, including three in the top 10. The United States dominates the top 10 and the top 100, but all international rankings put Europe as the dominant continent in the top 500. This shows that Europe has a very solid higher education system.

Participation in higher education is growing in Europe, but it is still behind that of South Korea and the United States. Participation levels in China are catching up very quickly (Figure 1).

The growing participation unfortunately does not lead to more graduates; it is just enough to compensate for the shrinking population in the age band associated with higher education. A positive evolution is however that the share of science and technology graduates is the second highest in the world (after South Korea), and that Europe produces the second largest number of PhDs per population (Figure 2, again, after South Korea).

Given demographic evolutions, Europe will not be able to match the number of higher education graduates of China in the future. There are only three ways to increase the number of graduates.

- Try to further increase participation but there is not an unlimited number of students that are qualified to attend higher education (Europe sets 40% as a target; some countries already reach this level).
- Try to increase the number of graduates via lifelong learning. That means that workers are (re)trained while working, or between two jobs.
- A last option is to try to attract more foreign students/graduates, especially those who have plans to stay in Europe after graduation. Given the fact that almost every country in the world is trying to stop brain drain, and has created incentives to bring successful expats back to their home country, the impact of recruiting overseas students is also limited, and the numbers will always be lower than the number of local students.

In Europe, higher education is mostly government-funded, making it affordable for most young people. In Europe, the students pay on average less than 10% of the real costs of tuition fees while in the United States and the UK it can be up to 30%. European universities generally do not have access to the huge endowments of some US universities.

Excellent research

European universities produce significantly more PhD degrees per 1,000 of the population than American or Chinese universities. The majority of European countries perform better than the United States, even in science and technology.

During the last 20 years, Europe has maintained its global share of scientific publications, while the United States has seen a steady decline, and China has shown spectacular growth (Figure 3). Europe kept not only its share of scientific publications, but also its share of the 10% most highly cited publications. The US is also losing its share of highly cited publications. The same trend is visible for the top 1% highly cited publications. It is remarkable that the sum of highly cited publications from the United States and China has been almost

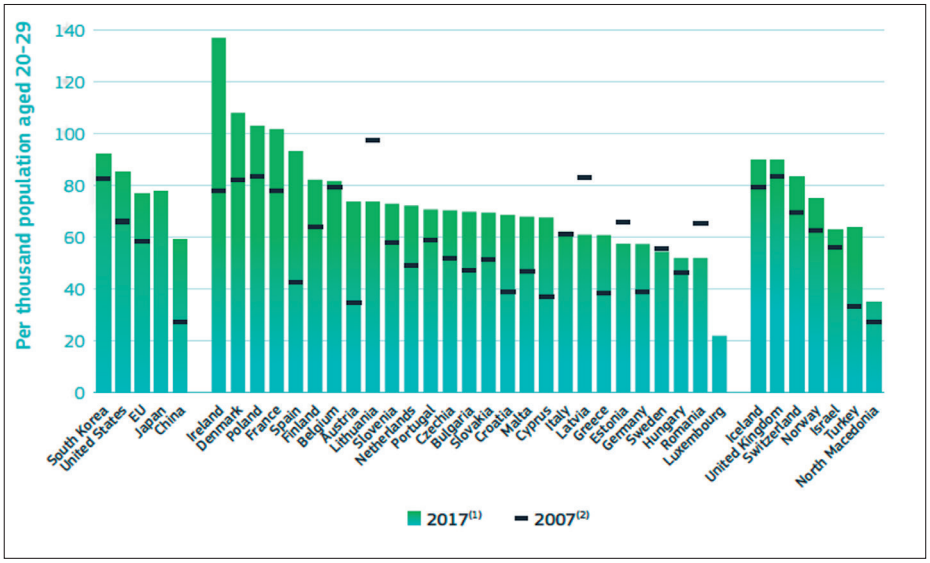


Figure 1: New graduates from tertiary education per 1000 population aged 20-29, 2007 and 2017 (Source: DG Research and Innovation)

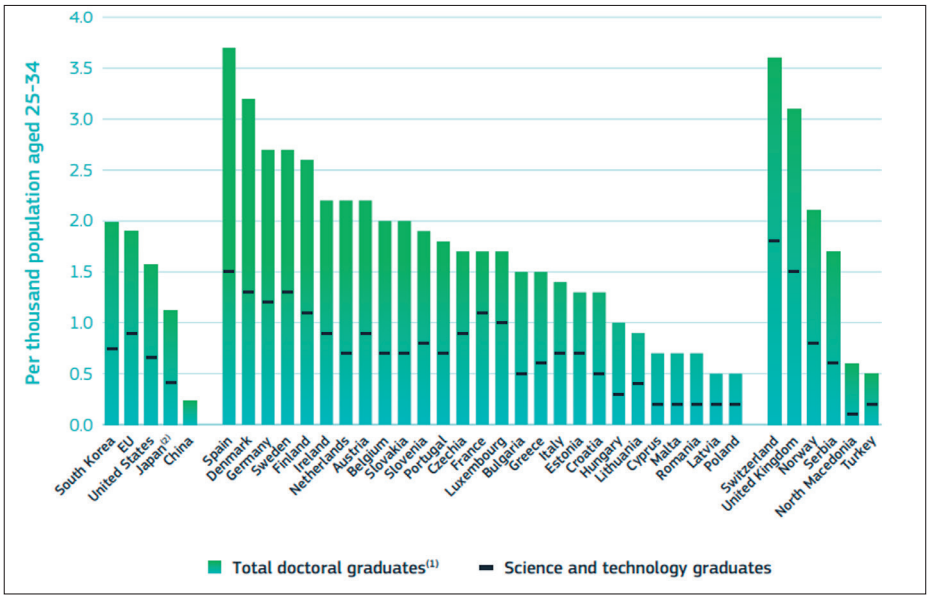


Figure 2: New doctoral graduates per 1000 population aged 25-34, 2017 (Source: DG Research and Innovation)

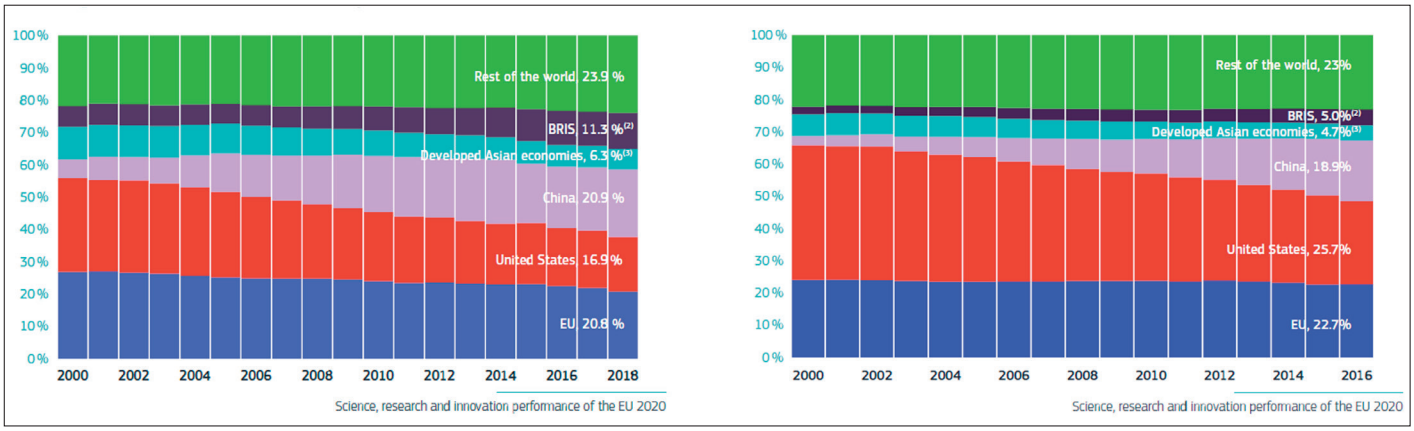


Figure 3: World share in scientific publications and 10% highly cited scientific publications (Source: DG Research and Innovation)

constant since 2000, which seems to indicate that the Chinese presence is growing at the expense of the US. Could there be a brain drain from the US to China: Chinese graduates from world-class American universities who return to China, and start a research career at home? If this trend continues, China will become the leader in publications and citations by the end of this decade. Hopefully, Europe will be able to defend its position in second place. As there is a correlation between the amount of public research funding and the number of highly cited publications, it seems crucial not to cut down on public research funding. The United States and Japan have been cutting public research funding in the last decade, and their numbers of publications and citations have followed the same path. Under the Trump administration, foreign students were discouraged to study

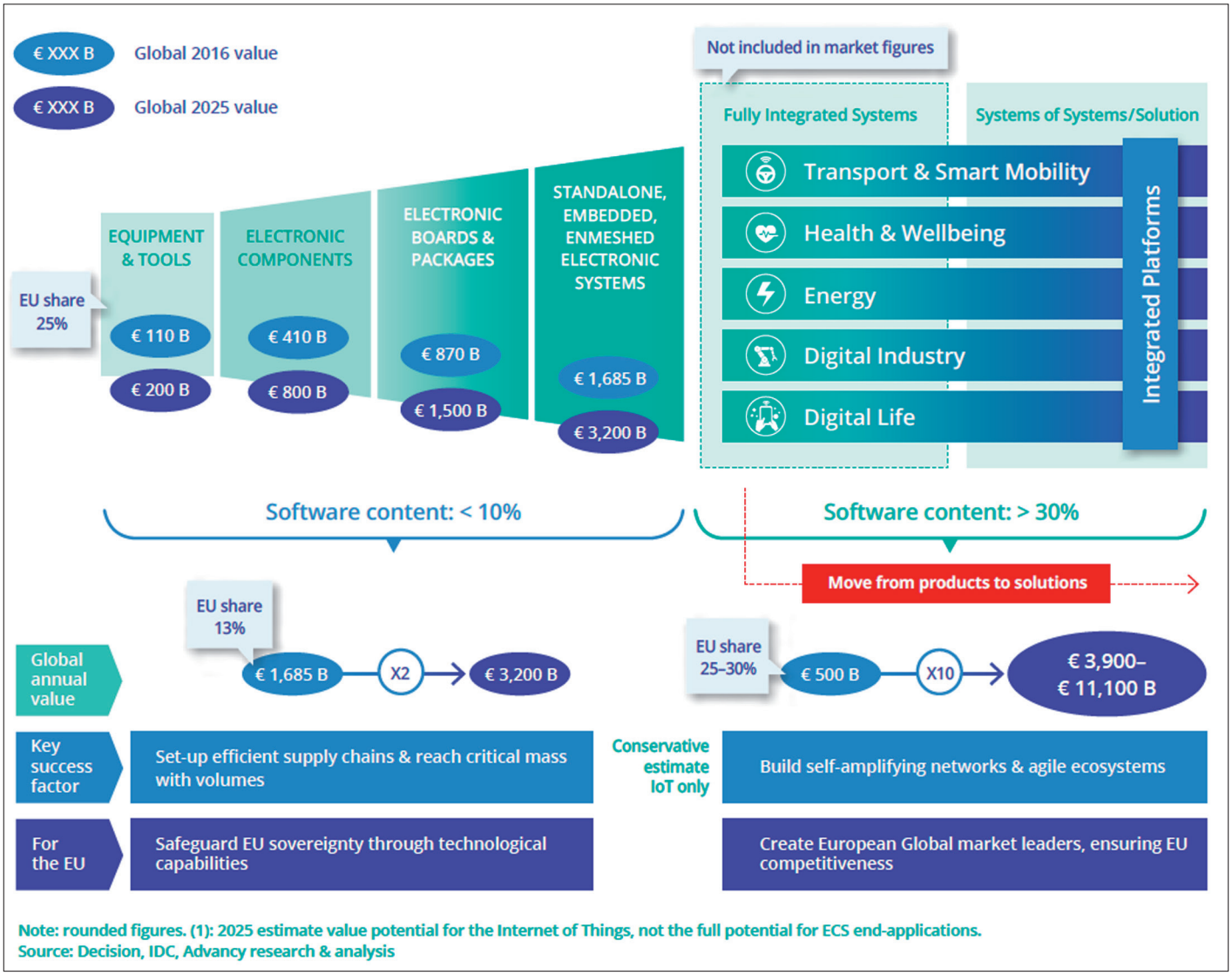


Figure 4: Embedded application domain markets (Source: ECS)

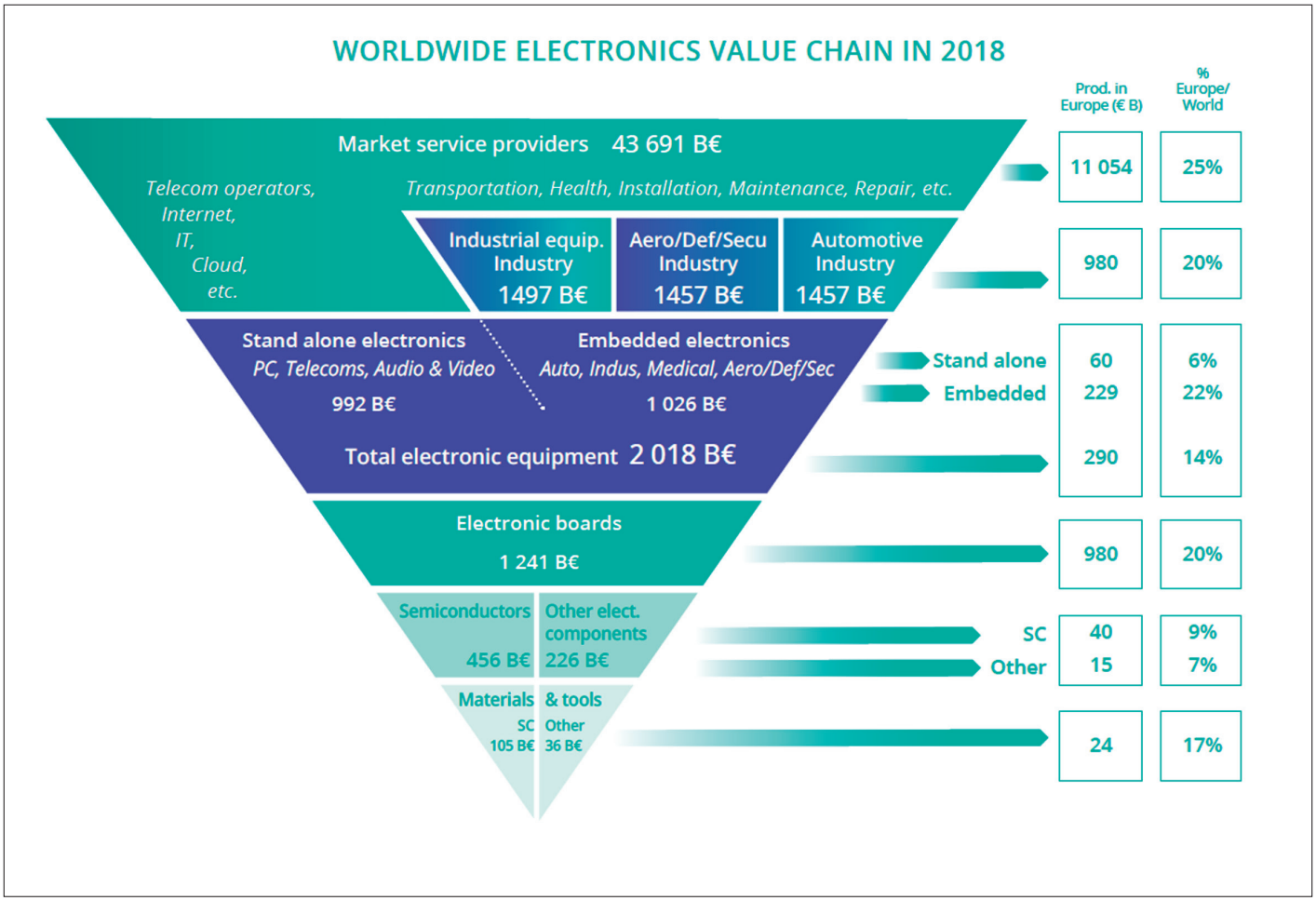


Figure 5: Worldwide electronics value chain in 2018 (Source: ECS)

in the United States, and currently the COVID-19 pandemic keeps them at home. Given the dependence of US research institutions on foreign talent, this might accelerate the decline in the number of publications, and citations.

World leader in lithography for semiconductor manufacturing

Europe has several research institutes and companies that are key players in semiconductor technology development (including ASML for making advanced EUV lithography machines, CEA, imec and Fraunhofer). They are Europe's biggest asset when it comes to the further development of CMOS-technology, and their expertise might be crucial to the development of post-CMOS technology.

Having the knowledge is also very important in order to understand the technology and therefore being able to use it efficiently in products. Curiously, since Global Foundries decided to step out of the race, Europe no longer has any advanced fabs. TSMC recently decided to build a fab in Arizona [10]. Europe might create incentives to attract advanced fabs too.

Second largest market in the world

According to the International Monetary Fund, Europe (EU-28) has the second largest GDP in the world, and the second highest GDP/capita.

Country	GDP in billion USD (2018)	GDP/capita USD (2018)
USA	20 513	62 571
EU	18 769	43 120
China	13 891	9 580

European businesses have access to a large internal market, with significant potential for growth in the new member states. Having access to a large internal market (like China and Europe) is an important advantage in times of troubled international trade relations. However, the European market is very fragmented due the diversity of regulations and languages across countries.

Stronger in systems than in components

In major embedded application domains, Europe is a global leader (Figure 5). According to the "Strategic Research Agenda for Electronic Components and Systems 2020" [3],Europe produces 25-30% of the global annual value at the system level, compared to 13% at the component level (Figure 4). The system level is expected to grow tenfold between 2016 and 2025 while the component level will only double. The system level is a clear strength in Europe, and a strength that we should exploit. The sector that is particularly strong is transportation (automotive, air, rail).

Common market

At the policy level, one of the strengths is the common market, and the fact that Europe can act as one economic block in

global trade negotiations. Individual countries do not have to negotiate individual agreements. However, there is still a long way to go before Europe becomes a fully integrated market with one set of laws, one currency and one tax system. The difference in minimum wages across Europe shows how pronounced the difference between countries is (Figure 6).

Strong public funding

Europe has a variety of research funding instruments, complementing national funding instruments. The research and innovation programmes of the European Commission help to stimulate research collaboration. European Research Council instruments support research excellence; the flagship programmes aim to create critical mass in key research areas; the European Institute of Technology aims to stimulate research and innovation; and joint undertakings like ECSEL aim to pool local and European funding to encourage research and innovation.

The total amount of public funding available makes Europe a good place to carry out R&D (at 0.7% of GDP). Worldwide, Europe is in second place after South Korea (Figure 7).

However, the relatively high amount of public funding across the EU does not compensate for the low R&D investments by industry (see weaknesses). When considered as a whole, Europe is dramatically lagging behind other parts of the world. The aim for Europe is to spend 3% of GDP on R&D, but it is still far away from that target (Figure 8).

The intensity of R&D translates into the number of researchers employed. Although Europe produces a higher number of PhD graduates per 1,000 of the population than any other continent, this does not lead to more employed researchers and almost half of them are employed by the public sector (Figure 9).

The total picture of R&D intensity is depicted in Figure 10. Asian countries appear to be preparing for the future. Their R&D intensity (apart from that of Japan) is growing as least as fast as the average growth in Europe.

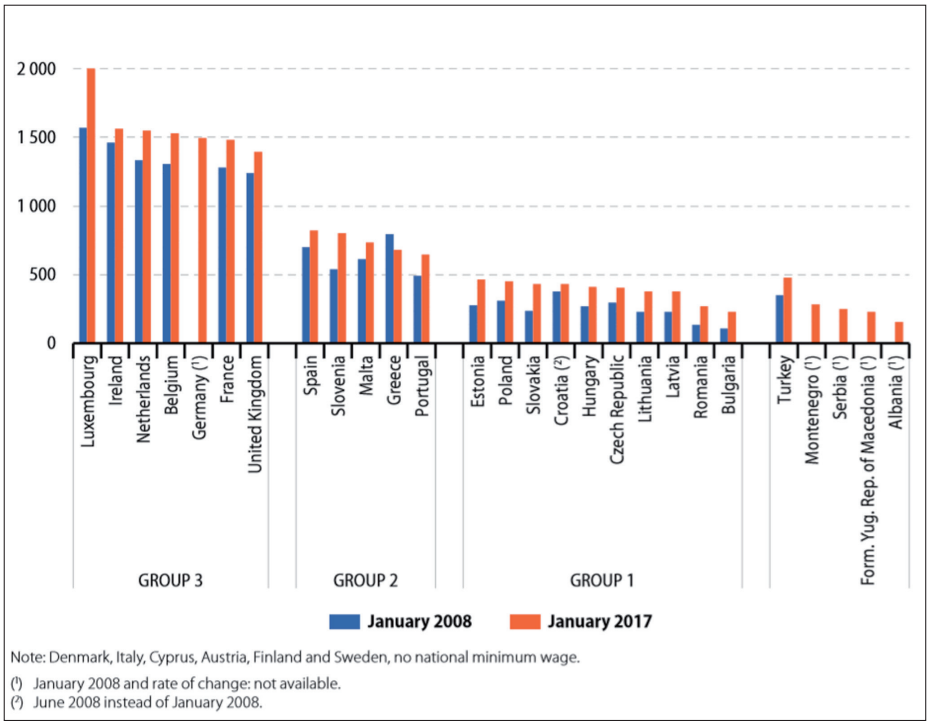


Figure 6: Minimum wages, January 2008 and 2017 (EUR per month) (Source: Eurostat)

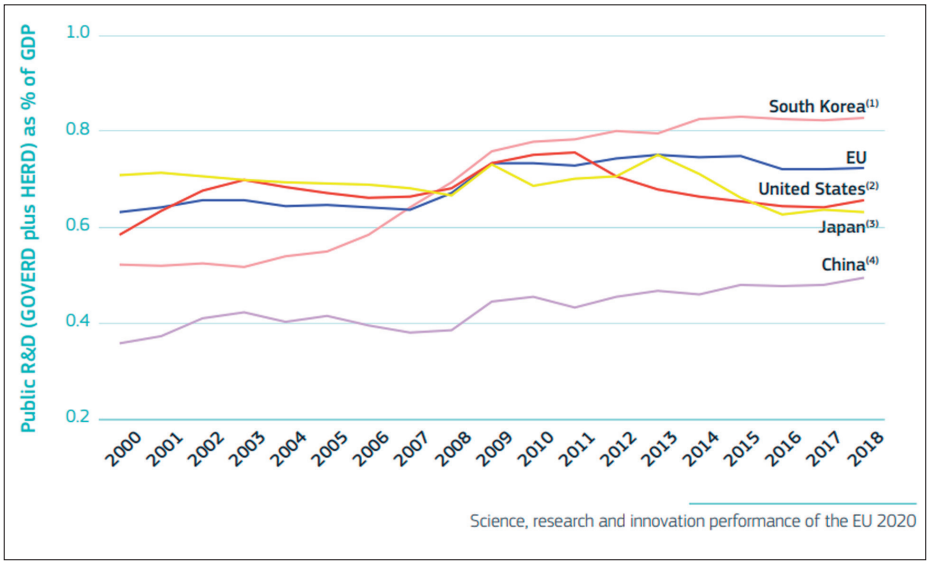


Figure 7: Evolution of public R&D intensity 2000-2018 (Source: DG Research and Innovation)

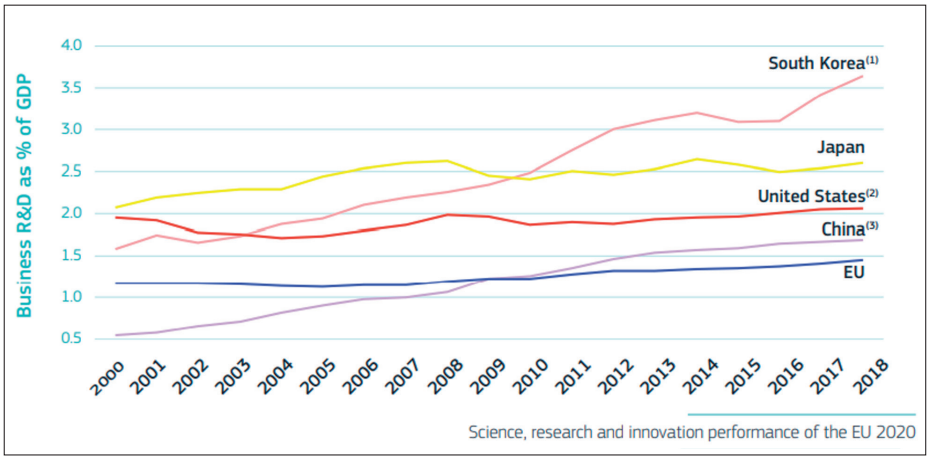


Figure 8: Evolution of business R&D intensity 2000-2018 (Source: DG Research and Innovation)

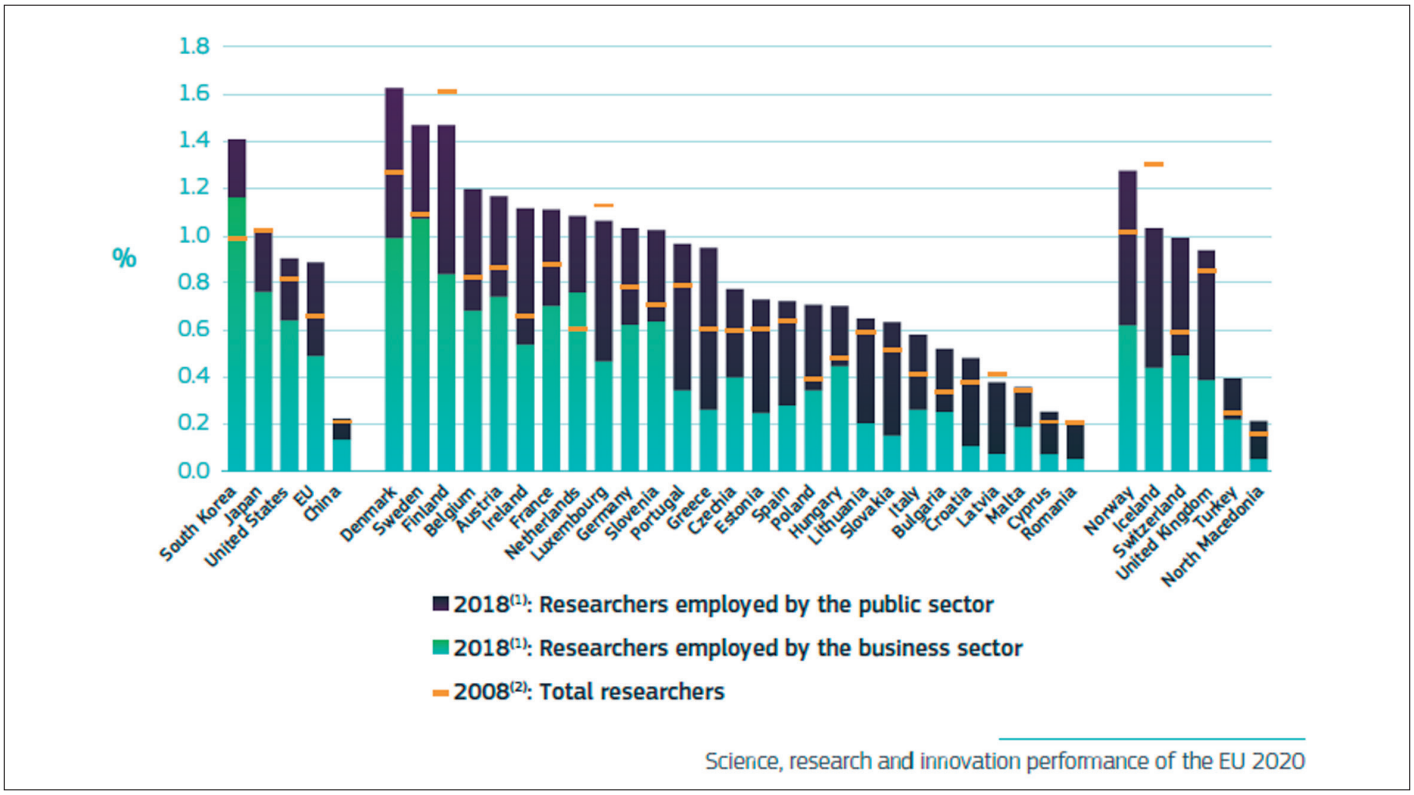


Figure 9: Total researchers (FTE) as % of total employment 2000-2018 (Source: DG Research and Innovation)

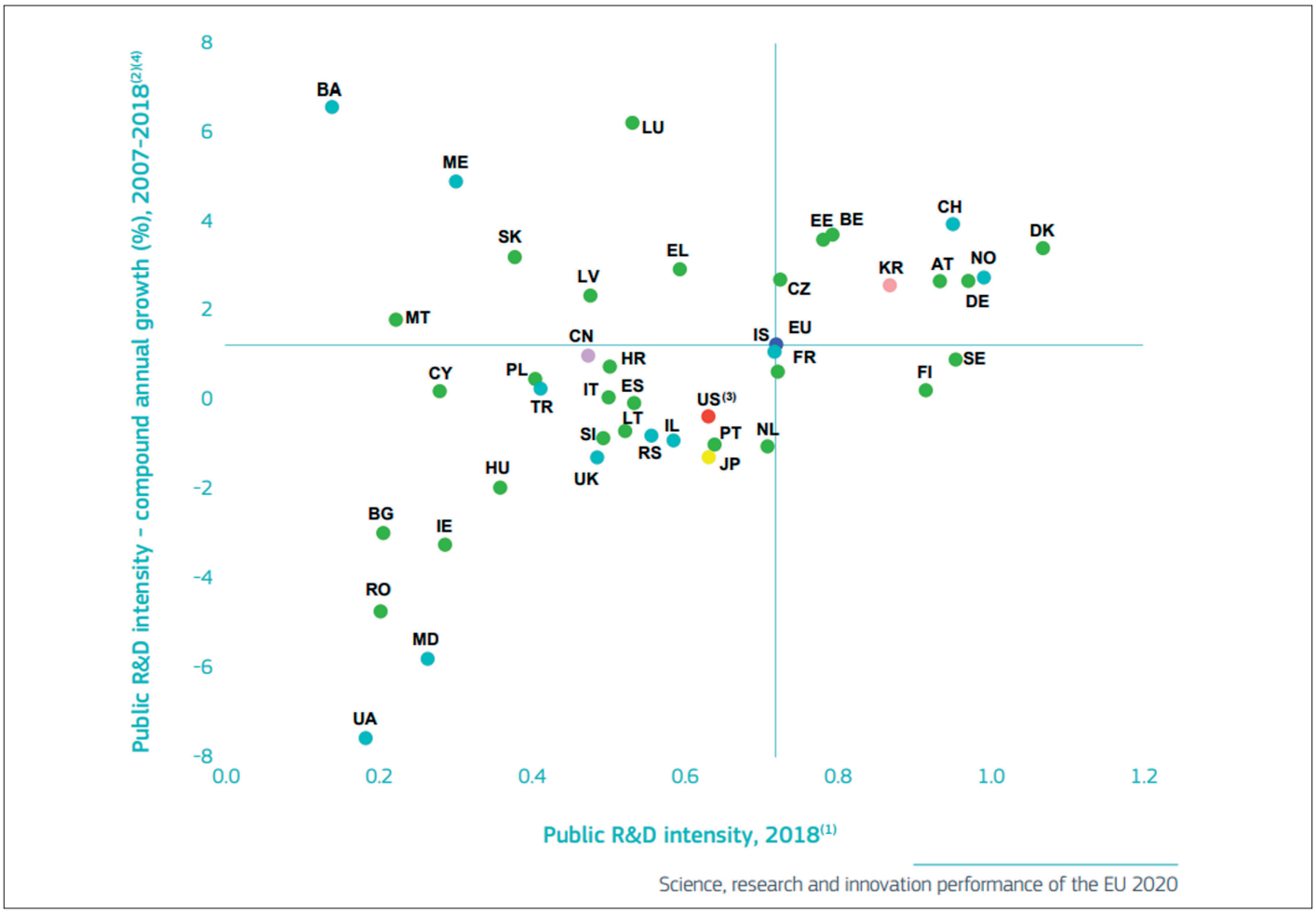


Figure 10: Public R&D intensity, 2018 and compound annual growht (%) 2007-2018 (Source: DG Research and Innovation)

Weaknesses

Weak academia-industry link

Compared to the United States, the collaboration between academia and industry (quantified as the number of joint scientific publications) is on average weaker, but there are large regional differences. The situation is fortunately improving (Figure 11).

Strong in research, but not in commercialization

Europe is lagging behind Japan with respect to the innovation output indicator (based on four components: patents, employment in knowledge-intensive activities, trade in knowledge-based goods and services, and the innovativeness of high-growth enterprises). There are however large differences in innovation performance between member states. The EU as a whole has recently caught up with the US (Figure 12).

In recent years, seven European cities have emerged as start-up ecosystems in the global top 30. Regional and European authorities are currently investing heavily in the creation and the support of start-up ecosystems in all European urban areas (Figure 13).

EU ICT contributes less to GDP than in other advanced countries

The European ICT-industry contributes around 4% to GDP (and is decreasing), compared to more than 5% in competing geographies [4] (Figure 14). One explanation is that Europe lacks GAFAM (Google, Apple, Facebook, Amazon, Microsoft) or BATX (Baidu, Alibaba, Tencent, Xiaomi), and other major ICT companies like HP, Dell and IBM, and the ecosystem supporting them. One exception is Ireland, which is home to the European headquarters of several major ICT companies. Figure 15 shows what a missed opportunity this is. The lack of such companies is a structural weakness which also limits the innovation potential for the ICT sector (the smaller the sector, the fewer the resources available to invest in research and development). The lack of such large corporations can be explained by the lack of venture capital (VC) culture in Europe. In order for companies to grow to be worth to US $ 50-100 million, they have to enter non-European capital

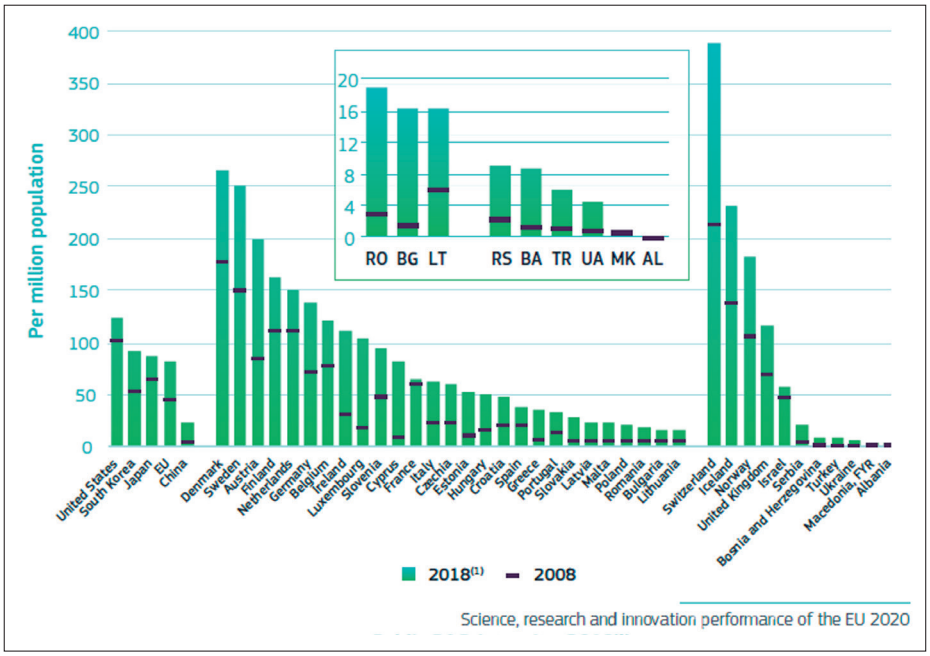


Figure 11: Public-private co-authored scientific publications per million population 2008 and 2018 (Source: DG Research and Innovation)

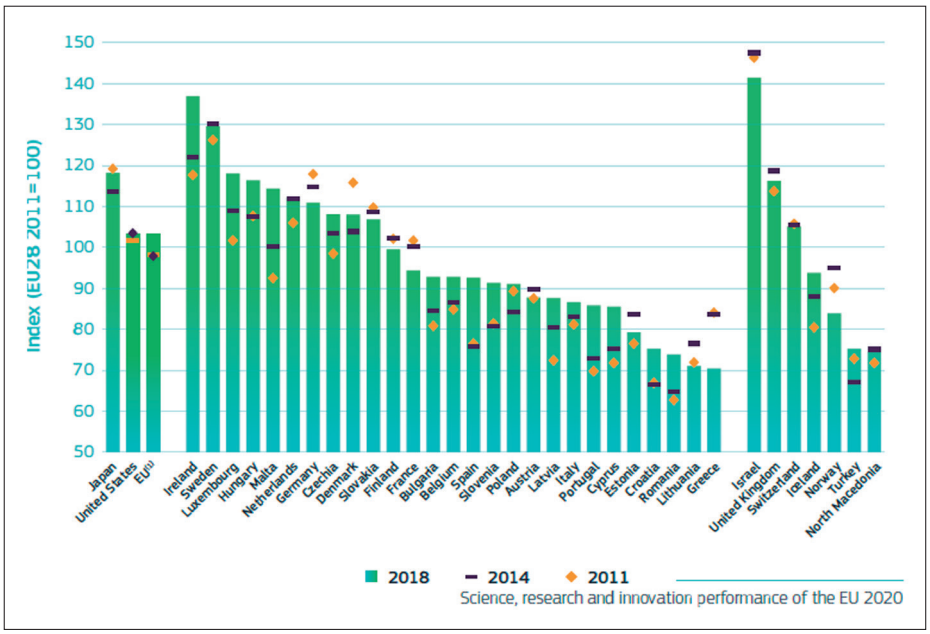


Figure 12: Innovation output indicator (EU2011=100), 2011, 2014 and 2018 (Source: DG Research and Innovation)

markets like the United States or China. The US market is very competitive and sophisticated, and Asian markets are even more challenging. Even growing within Europe has its challenges, because Europe is not a single entity; it is composed of a plurality of markets, languages, laws, cultures and so on. Therefore, it is difficult for a company to address the whole of Europe without extra work to adapt to each country. As an example, voice assistants appear later in non-English speaking countries due to the additional effort required to adapt them to different languages. Neither US or Chinese companies face such challenges. That is one of the explanations why European VCs are more cautious; they doubt whether many companies have the potential to successfully enter markets outside Europe.

Employment in ICT manufacturing is very low in the US and in the EU. China, Japan and South Korea are the ICT factories of the world. The US and the EU are strong in services, and on a par with South Korea and Japan (Figure 16).

Figure 13: 2019 Global startup ecosystem ranking (Source: DG Research and Innovation)

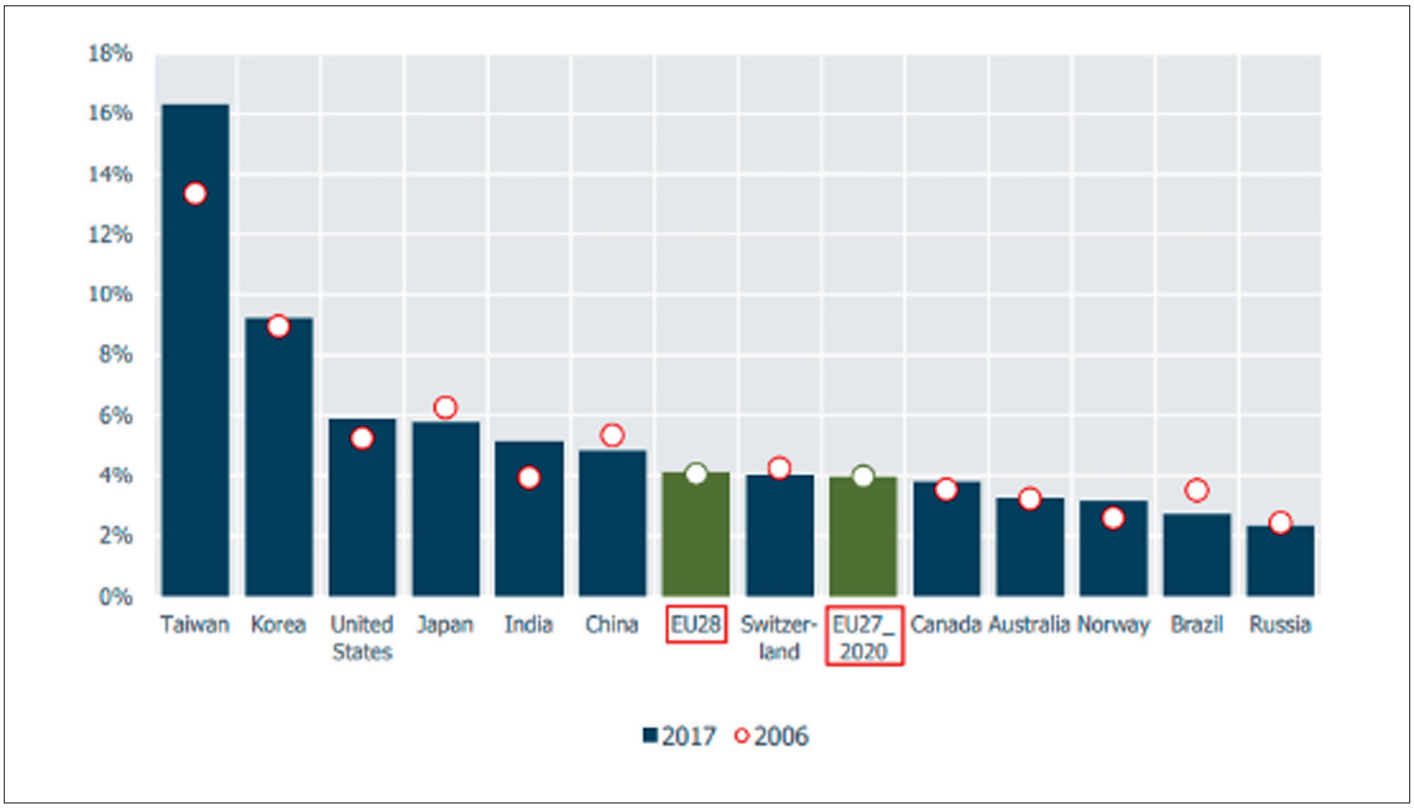


Figure 14: ICT sector value added share of GDP 2006-2017

Figure 15: Largest global companies in 2010 and in 2020.

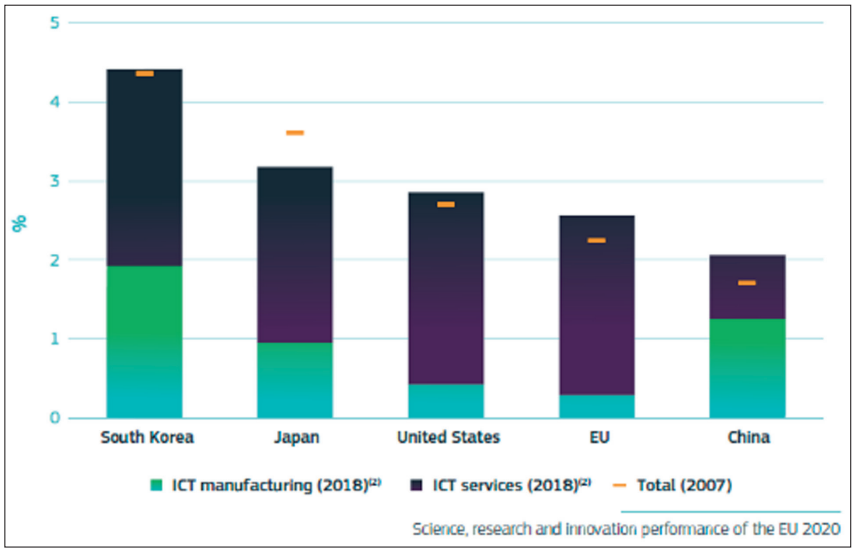


Figure 16: Employment in ICT as % of total employment broken down by manufacturing and services globally 2007-2018

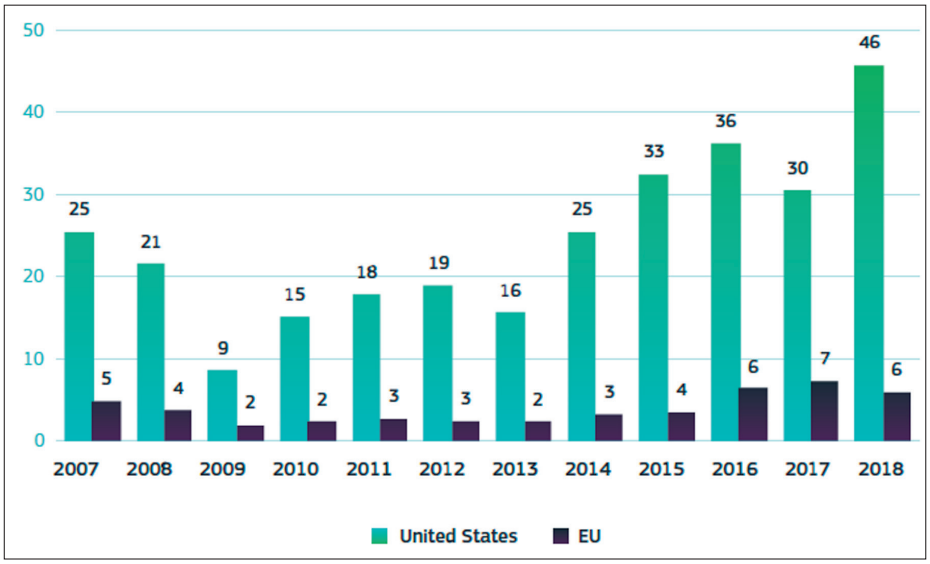


Figure 18: Venture capitalist funds raised (billion euro) in the EU and in the United States 2007-2018 (Source: DG Research and Innovation)

Figure 19: Venture capital average fund size (EUR million) in the EU and in the USA, 2007-2018 (Source: DG Research and Innovation)

The fact that Europe lacks major ICT companies has far-reaching consequences: it also means that venture capitalists are less eager to invest in European start-ups and scale-ups because there are fewer companies that might be able to acquire them. Companies that do grow significantly are often acquired by non-European companies: Nokia was acquired by Microsoft, Movidius by Intel, ARM by Softbank and more recently by NVIDIA, for example. Fortunately, there are also counterexamples like Sysgo, which was acquired by Thales.

Non-European business leaders like Elon Musk, Tim Cook, the Google founders and Masayoshi Son seem to have a clearer vision of the future than their European counterparts, which they promote actively in the media. Very few people know the CEO of major European computing companies like Infineon, Ericson and STMicroelectronics, who lack the "rock star status" associated with their international counterparts.

Lack of VC culture

More generally, Europe lacks a VC culture and, in this metric, the gap between the United States and Europe could not be bigger (Figures 18,19). This observation, in combination with the large number of young start-up companies, is problematic. It means that they have to fight hard to get the funding to become a scale-up company.

Lack of advanced foundries

There used to be foundries in Europe, but they were acquired by non-European companies and disappeared (Figure 20). The fact that Europe depends on foreign foundries means that it has to import most of its semiconductors. The leading foundries are not located in low-wage countries, meaning that they did not leave Europe due to labour costs. Given the fact that Europe is a world leader in the development of the technology used in foundries (CEA, imec, Fraunhofer, ASML), it is surprising that no large foundries are left in Europe and that Global Foundries some years ago decided to stop the development of 7nm technology and instead make its 14/12 nm FinFET platform more relevant to its customers. One explanation is that European countries did not aggressively invest in new foundries (as was the case in South Korea and in Taiwan), and that European VCs are not interested in foundries (while they are in the United States). Another is that the European customers of the foundries like STM, NXP and Infineon are making products that do not require advanced processes because their market is microcontrollers instead of microprocessors, analog devices versus memories.

Lack of ICT workers

Europe lacks hundreds of thousands of ICT workers. Most European countries are witnessing positive growth in the number of graduates overall, but the number of graduates does not yet match demand (Figure 21). It would seem that Europe is not succeeding in convincing large enough numbers of high school students to start a career in the ICT sector. This is unfortunate because the competitiveness of this sector in Europe will depend on the size of its workforce in order to innovate in big data analytics, artificial intelligence, robotics and so on.

Importing well-trained foreign workers to Europe en masse to help mitigate the shortage is not an effective solution. First of all, Europe needs more than one million ICT workers in the next decade. Secondly, most countries try hard to keep their local talent. Finally, Europe has become less inviting to immigrants during the last decade. To complicate things further, foreign ICT workers will be attracted by well-paid jobs in the major innovation hubs, and it will be more difficult to convince them to accept a job in smaller cities, or in poorer countries. The only long-term and sustainable solution is to invest heavily in the technical education of local people.

Fragmentation of funding

The public funding system in Europe is highly fragmented. There are national funds, regional funds and European funds. There are funding instruments for applied research, for innovation, and for fundamental research. There are individual grants and collaborative research grants. A particular research proposal could fit multiple funding instruments and calls. Sometimes a research proposal can only be funded if different agencies agree to each fund a part of the proposal. On top of this, the success rate for research proposals is sometimes lower than 10%.

Within a funding agency, different committees deal with particular topics, which makes multidisciplinary project proposals very hard to get funded because committees tend to give priority to the proposals that belong to the core of a domain, leading to *lower acceptance rates for interdisciplinary projects*. It is therefore very hard for technologies that are common to several application domains, such as research in computing hardware and software, to be funded for their own intrinsic development. Instead, they must piggy-back on more domain-specific

1Q20 Top 10 Semiconductor Sales Leaders ($M, Including Foundries)

1Q20 Rank	1Q19 Rank	Company	Headquarters	1Q19 Total IC	1Q19 Total O-S-D	1Q19 Total Semi	1Q20 Total IC	1Q20 Total O-S-D	1Q20 Total Semi	1Q20/1Q19 % Change
1	1	Intel	U.S.	15,799	0	15,799	19,508	0	19,508	23%
2	2	Samsung	South Korea	11,992	875	12,867	13,939	858	14,797	15%
3	3	TSMC (1)	Taiwan	7,096	0	7,096	10,319	0	10,319	45%
4	4	SK Hynix	South Korea	5,903	120	6,023	5,829	210	6,039	0%
5	5	Micron	U.S.	5,465	0	5,465	4,795	0	4,795	-12%
6	6	Broadcom Inc. (2)	U.S.	3,764	419	4,183	3,700	410	4,110	-2%
7	7	Qualcomm (2)	U.S.	3,753	0	3,753	4,050	0	4,050	8%
8	8	TI	U.S.	3,199	208	3,407	2,974	190	3,164	-7%
9	11	Nvidia (2)	U.S.	2,215	0	2,215	3,035	0	3,035	37%
10	15	HiSilicon (2)	China	1,735	0	1,735	2,670	0	2,670	54%
—	—	Top-10 Total		60,921	1,622	62,543	70,819	1,668	72,487	16%

(1) Foundry (2) Fabless
Source: Company reports, IC Insights' *Strategic Reviews* database

Figure 20: Top 10 semiconductors sales leaders (Source: IC Insights)

Figure 21: Graduated in the field of ICT per thousand population aged 25-34, 2017 and compound annual growth, 2010-17 (Source: DG Research and Innovation)

project calls. The organizational structure of the funding agencies thus ends up constraining the research work that can be proposed in one single project. The design of a novel, secure, cloud-based IoT solution will cut across the topics of at least three units of DG CONNECT. The fact that European Regional Development Fund has also started to be used to fund research only adds to the complexity.

Finally, there are the European state-aid rules that significantly add to the complexity.

Opportunities

Fortunately, there are also opportunities.

	Opportunities	Threats
Science and Technology	• The end of Moore's law	• Economic stagnation • Brain drain
Industry and Market	• Embedded systems, IoT, CPS, edge intelligence	• Saturating markets • Computing initiatives in countries such as China, Russia and Japan
Policy and Government	• Solutions for societal challenges	• Political instability

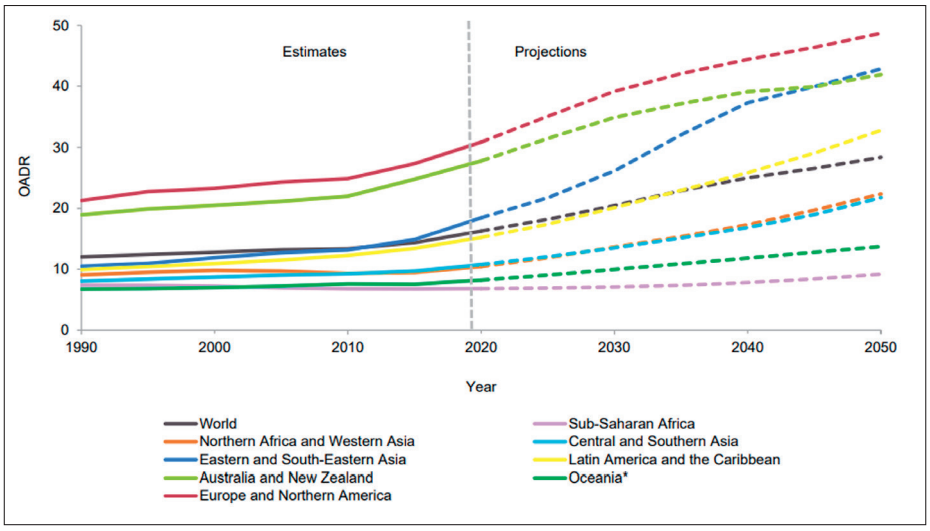


Figure 22: Ratio of people aged 65+ per 100 people of working age (old age dependency ratio) (Source: United Nations (2019))

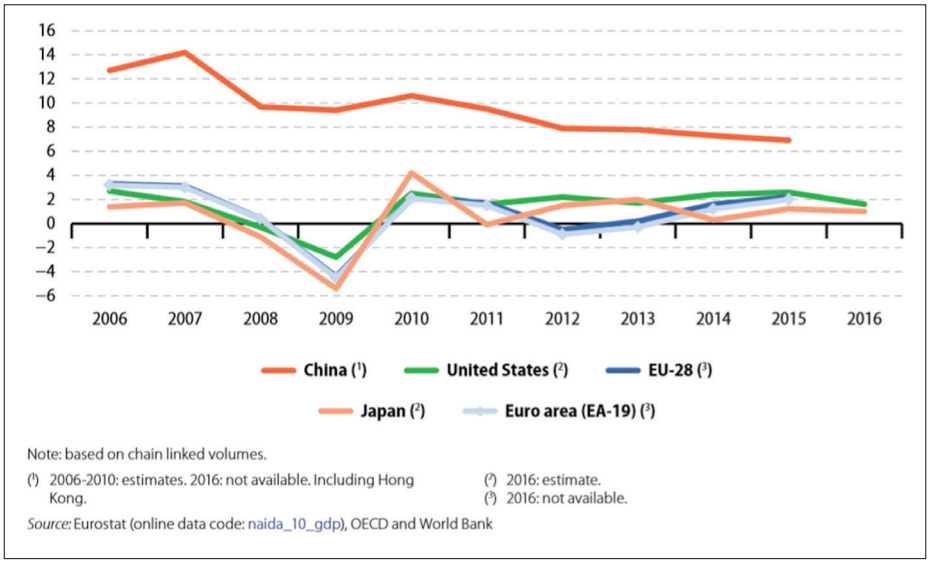


Figure 23: Real GDP growth 2006-2016 (% change compares with previous years) (Source: Eurostat)

The end of Moore's law

The increase of the sequential performance of a processor at the pace of Moore's law already ended a decade ago (end of Dennard's scaling); parallelism kicked in to keep performance increasing in lockstep with the number of transistors and cores, and now accelerators are the preferred technique to further improve performance, but the parallelism and the heterogeneity add a lot more complexity for software developers.

The design of accelerators marks a new era of architectural research to devise clever solutions to improve performance per Watt. There is, however, room (and also a need) for more disruptive solutions, possibly replacing the (rather inefficient) von Neumann architecture with other computing paradigms, but only if it is done with a high level of efficiency and in a short period of time. Progress in artificial intelligence could help in the efficient design of new systems.

Designing a new accelerator or launching a new technology is however a daunting task as the proposed solution has to be better than current solutions, and also needs to have a roadmap in order to keep the lead.

Embedded systems, IoT, cyber-physical systems (CPS), edge intelligence

The number one market opportunity in computing systems is the strongly growing market of embedded systems (including the IoT, CPS, edge intelligence and the digitization of European industry). Europe has the second largest economy in the world, it has a number of world-class players producing the key enabling technology for advanced embedded systems, and it has strong transportation and health industries. Furthermore, there are no non-European dominant companies like Google, Apple, Facebook, Amazon or Microsoft (GAFAM) in this space yet. The stars of the CPS era will probably not be the same as those of the internet era (which are different from those of the mainframe era). Could the company dominating computing in 2030 be European? The only way to win this race is to create as many innovative start-ups as possible, support them to scale up, and hope that they will become world leaders.

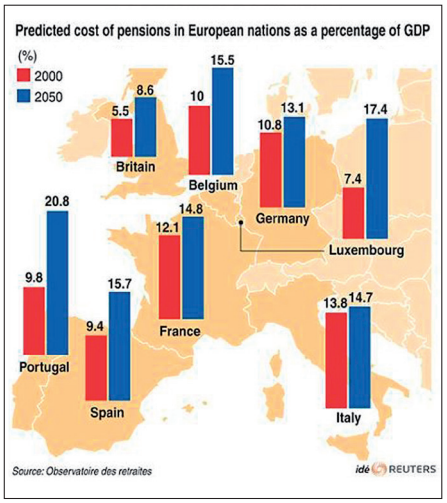


Figure 24: The cost of pensions in Europe [5]

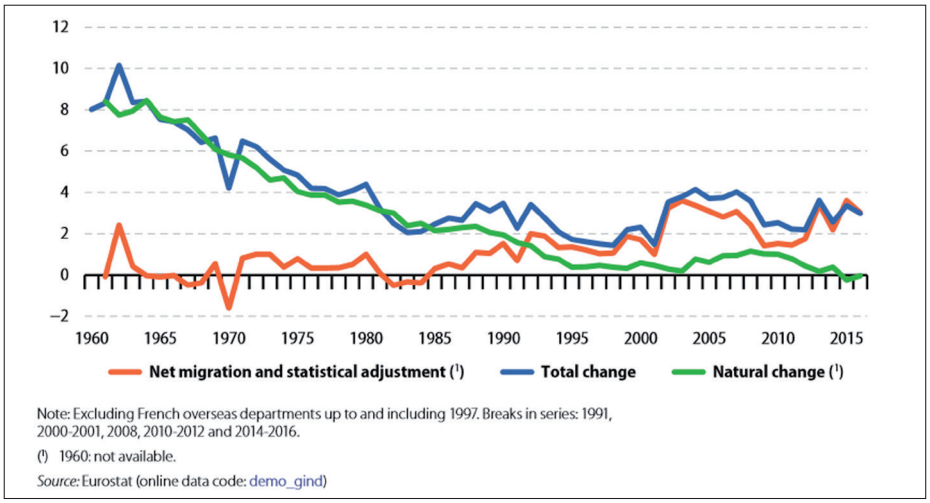


Figure 25: Population change by component (annual crude rates), EU-28, 1960-2016 (per 1.000 persons) (Source: Eurostat)

Solutions for societal challenges

Societal challenges form a huge opportunity for the European computing industry. Europe is the region with the highest number of people aged 60 or older [9]. The old age dependency ratio (OADR) is the number of 65+ people per 100 people of working age (24-64). That ratio will increase dramatically over the next 30 years (Figure 22). That means that Europe will have to search for solutions for the ageing population first. Since the rest of the world will face the same challenges in the future, Europe has an opportunity to develop and commercialize services and products for the silver economy first and sell them to the rest of the world.

The same reasoning holds for the environment. The European population (together with the US) has one of the largest ecological footprints in the world. Reducing the ecological footprint will become one of most important global challenges of the rest of the century. The European Green Deal will help European scientists and industry to find solutions for footprint reduction that can be used and applied across the world. This is a once in a lifetime opportunity.

Threats

Economic stagnation

Europe has been characterized by low economic growth in the past decade (Figure 23).

This situation has recently been aggravated by the economic impact of the COVID-19 pandemic. Around the time when the impact of the pandemic might become less severe, part of Europe risks experiencing the negative impact of Brexit. In the background, there is the increasing cost of supporting the ageing population. The cost of pensions will continue to grow until 2040 when "baby boomers" will have reached their life expectancy (Figure 24).

All the above puts stress on businesses, governments and people. Given the fact that European research is primarily funded by public money, this situation might lead to cuts in R&D budgets, especially for long-term curiosity-driven research.

Brain drain

There is a lot of public attention given to the topic of immigration in Europe. It is indeed the case that immigration has increased since the fall of the Berlin Wall in 1989 and is now a major source of population growth in Europe.

This graph, however, masks the fact that net immigration is the difference between immigration and emigration (Figure 25). Emigration usually takes place from economically weaker countries toward economically stronger countries: from the Middle East and North Africa to Europe, from Eastern and Southern Europe to North-Western Europe, but also from North-Western Europe to the United States and other rich countries in the world.

In computing, there seems to be a brain drain from Europe to the US. Top researchers and ambitious entrepreneurs are attracted by the merit-based American society and top salaries for high potential in both academia and industry. Large multinational ICT companies are attractive employers for young European talent eager to travel the world and make a fast career. If they do not want to move, US-based companies acquire European companies in order to have access to their talent or to open subsidiaries in Europe. This is a less visible form of brain drain. Particularly in machine learning, there has been a very strong pull on the top talent in Europe by companies like Facebook and Google.

Europe should create large and well-funded competence centres to retain European talent, and to attract excellent workers from abroad. CERN is a good example of such a competence centre, attracting talent from all over the world. The proposals for pan-European centres in artificial intelligence [6] and cybersecurity [7] launched recently will hopefully help fulfil this need.

Saturating markets

The market for new desktop computers and laptops is shrinking, and the market for smartphones is shrinking too (after having cannibalized the markets of other devices like navigation systems, cameras,

Image: ID 168883350 ©Cineberg Ug | Dreamstime.com

music and video players). The reason is that we have reached human scale: most people in the western world already have all the devices they need, and the features of most devices have stabilized, eliminating the need to replace devices to get access to more features. The COVID-19 pandemic and the requirement to work and study from home might have created a short peak in the demand for devices needed to telework (mobile devices, headsets, webcams), but this is not a long-term trend. Sustainability requirements encourage people to use devices until their end-of-life or have them repaired if they are not yet end-of-life. This will further reduce the demand in the longer term. Fewer sales means fewer resources to spend on the development of new devices and new features.

Computing initiatives in countries such as China, Russia and Japan

A threat to the European computing industry is the rapid development of the computing industry in China, Russia and Japan. Many countries understand that computing is a key enabling technology of strategic importance, and are investing in their own research, products and companies. If Europe fails to do the same, it might eventually become dependent on technology that is designed, developed, produced and controlled outside Europe. The same holds for cybersecurity solutions.

The fastest growing country of the moment is China. There are several sectors where it has the ambition to become a world leader (artificial intelligence [11] and renewable energy being just two examples). This is evident from the quickly growing number of patent applications by Chinese companies.

The ambition of China to become the frontrunner in artificial intelligence was made very clear in 2017 in their Next Generation Artificial Intelligence Development Plan [8]. It states: "*… by 2030, China's AI theories, technologies, and applications should achieve world-leading levels, making China the world's primary AI innovation centre, achieving visible results in intelligent economy and intelligent society applications, and laying an important foundation for becoming a leading innovation-style nation and an economic power*". China created a five-year AI talent training program, and invested more than US$2 billion in a huge AI industrial park in the suburbs of Beijing. The presence of Baidu, Alibaba and Tencent (BAT) is an asset in developing advanced AI applications. The fact that the Trump administration has put restrictions on Chinese companies will not change this.

Political instability

Another threat is the political instability that Europe and the rest of the world are currently experiencing. Brexit, the inability to form stable governments in some countries, the COVID-19 pandemic and the refugee crisis are influencing business and consumer confidence. In some countries, there is a trend towards more authoritarian regimes that want to turn back some liberal civil rights.

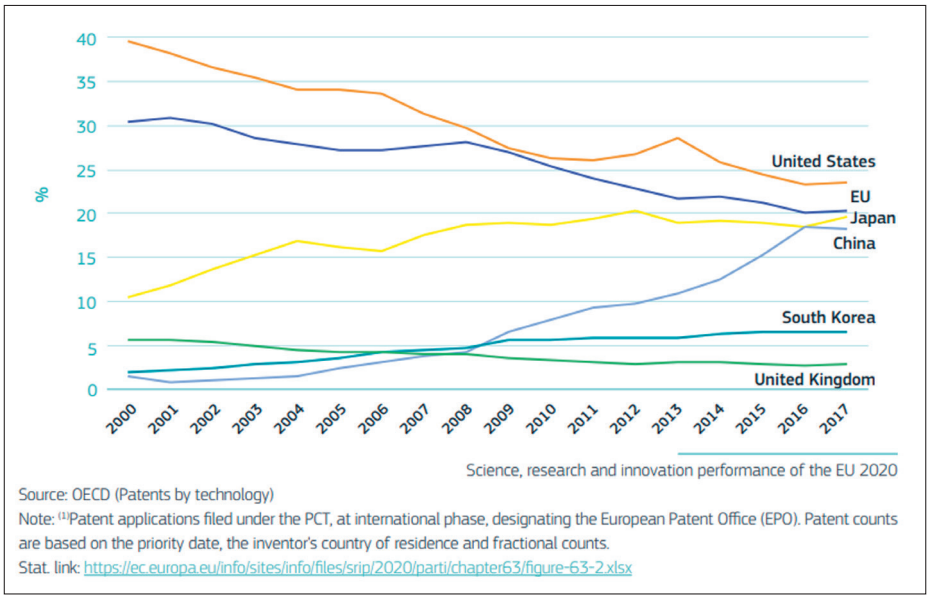


Figure 26: World share (%) of PCT patent applications 2000-2017 (Source: OECD (Patents by technology))

References

[1] European Commission, "Science, research and innovation performance in the EU", https://ec.europa.eu/info/publications/science-research-and-innovation-performance-eu-2020_en

[2] "World University Rankings 2020", https://www.timeshighereducation.com/world-university-rankings/2020/

[3] "Artemis Strategic Research Agenda", https://artemis-ia.eu/documents.html

[4] European Commission, "Science, Research and Innovation Performance of the EU 2020", https://ec.europa.eu/info/sites/info/files/srip/2020/ec_rtd_srip-2020-report.pdf

[5] J. Salmon, "Great pensions divide: Private sector staff must put in a third of their pay to match state worker benefits", https://www.dailymail.co.uk/news/article-2009669/Great-pensions-divide-Private-sector-staff-pay-match-state-worker-benefits.html

[6] CLAIRE, "Confederation of Laboratories for Artifical Intelligence Research in Europe", accessed december 2020, https://claire-ai.org

[7] ECSO, "European Cyber Security Organisation", accessed december 2018, https://ecs-org.eu

[8] R. Creemers, "A Next Generation Artificial Intelligence Development Plan, China Copyright and Media", 1 Aug 2017, https://chinacopyrightandmedia.wordpress.com/2017/07/20/a-next-generation-artificial-intelligence-development-plan

[9] UN, "World Population Ageing 2019" https://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2019-Highlights.pdf

[10] S. Pham, "Taiwan chip maker TSMC's $12 billion Arizona factory could give the US an edge in manufacturing", https://edition.cnn.com/2020/05/15/tech/tsmc-arizona-chip-factory-intl-hnk/index.html

[11] F. Westerheide, "China – The First Artificial Intelligence Superpower", https://www.forbes.com/sites/cognitiveworld/2020/01/14/china-artificial-intelligence-superpower/?sh=5c6dc28f2f05

Koen De Bosschere is a Professor in the Electronics department of Ghent University, Ghent, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: K. De Bosschere. The position of Europe in the world. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 162-175, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

Seventy years after humans invented software, humanity realizes that it is no longer under its control but has entered a symbiotic co-evolution.

Extreme reuse: the only future any code can afford

BY THOMAS HOBERG

Software engineers tend to see themselves as designers or architects, true creators of things new, perhaps even optimal. That may have been true for the early pioneers, who created algorithms that many considered "art". But to understand what is happening in the IT industry, we should perhaps regard ourselves as mutating the code base of that species of software, with which we have entered into symbiosis. We are a bit like a virus, an agent of evolution, which has software adapt to a changing world.

Key insights

- While the cost of storing code has fallen exponentially for decades, the cost of creating good new code remains dramatically high: because nobody can afford clean design, code only evolves and includes potentially dangerous atavisms or leftovers.
- The energy investment vs. the value of computation becomes so dominant, that both start controlling the execution and the quality of the results opportunistically.
- The complexity of systems follows the complexity of society and its digitalization: specialization and differentiation deepen, aggregates evolve, interdependence and fragility increase as a consequence of economic optimization.
- The evolution, success, sustainability and life span of large bodies of software and programming languages critically depend on the quality and the investment of the communities around it. Those communities may require significant initial and long-term support to achieve and sustain critical size.
- Revolutionary from scratch designs only tend to work in isolation or on a green field. Where there is an existing competitor, they tend to fail even if that has significant defects. But when the pressure is high enough, that may evolve so rapidly, it looks completely new.
- There is no natural process to clean up code no longer used, not in genetics, not in software, not in social code. That creates significant vulnerabilities.
- Cost reduction pushes legacy code into IoT devices creating critical risks that we need novel ways to manage.

Key recommendations

- Accept that some legacy code will live in niches much longer than anyone planned for. It requires dedicated support in education, but may return value in excess of that cost.
- Only failure, potentially catastrophic, will clean up unmaintained legacy code: regulate and support the creation of evolutionary pressure beforehand to reduce the risk.
- Ensure regulation to assure a minimum of quality where safety and resilience are needed to counteract the overwhelming temptation of cutting & pasting legacy code into use cases and domains it was never designed for.
- Invest into defensive technology for the trusted base like control flow hardening, randomization, SoC IP blocks for application level firewalls.
- Where bad legacy code keeps getting recycled, support process and assets so that manufacturers of even the cheapest IoT devices can be convinced to switch to an EU-maintained high quality base.
- Support research in tools that help with refactoring and black-boxing legacy code, translating it into safer languages, or help with annotation for automated testing and entry-exit checks at black-box borders.
- Support a world-leading community of compiler maintainers capable of extending mainline compilers like LLVM or specialty compilers like SPARK with resilience features e.g. for control-flow integrity, memory capabilities or tagging support.
- Support legacy code and documentation data bases with smarter tooling to find code similarities and discover build tool dependencies.
- Support research into value-driven computing, which adapts its processing dynamically to the value of input data, as a key paradigm extension for sustainability.

Some history

During World War II critical mass was reached for the atom bomb and the computer era:

- Alan Turing proved that even the most complex mathematical problem could be divided into a series of rather primitive; logic operations given sufficient memory
- Electronic implementations of that logic blasted the speed barriers of mechanics away;
- John von Neuman proposed encoding the sequence of operations in the same memory as data to create software.

Earlier computers had to be physically rebuilt for each new task, now software enabled evolution with code: existing solutions could be cut & pasted whenever and wherever a similar problem needed solving and re-used with a minimum of incremental adaptation.

It enabled Grace Hopper to design high-level languages which allowed an algorithm to be expressed at a human level of abstraction while the computer itself would then translate it into the primitive machine instructions for execution. High-level language code is easier to write, read, understand and re-use by people and allows users to build the most complex applications from small well-designed and tested building blocks. The same principle applies to hardware, which combines libraries of primitive logic functions into layers and stacks of higher-level aggregates to implement the most complex chips, while it is implemented in a hardware description language first and "compiled" into silicon after.

The invention of integrated circuits created aggregates of transistor logic on silicon using photo-lithography reaching several billion gates over four decades on a single chip today, shrinking the etched structures from 10µm in 1971 to 10nm in 2016 or by a factor of 1000:1 along both sides, while the yield in transistors exploded in squares.

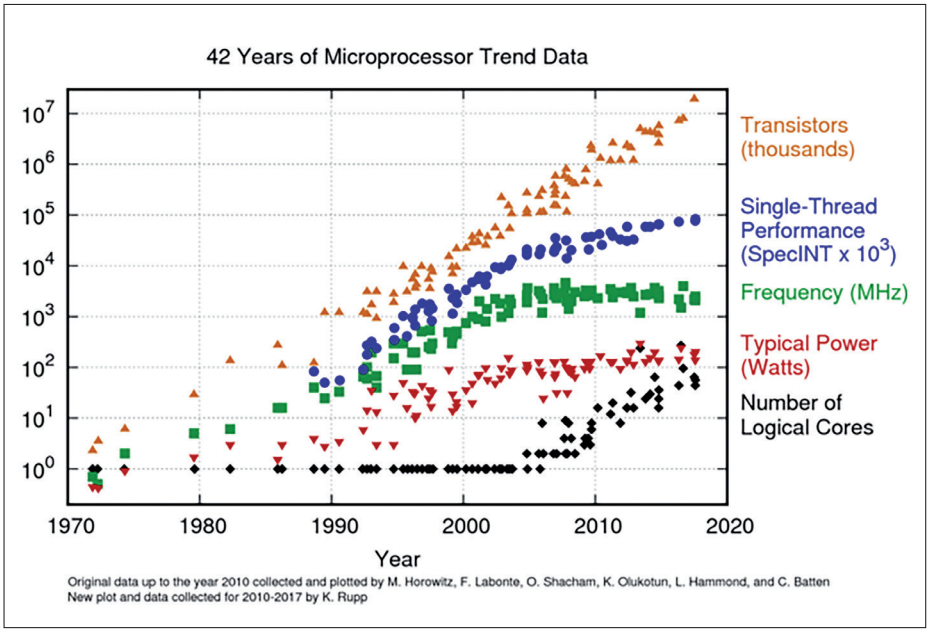
The incredibly rich functionality of current hardware and software is enabled by mind-boggling complexity and only possible because each design iteration was evolutionary, re-using and incrementally expanding the functional base for the last seven decades. The functional expansion of code, its ability to do things quicker than mechanics and much more complex than discrete electronics had it replace both, because process shrinks eliminated the markup of physical complexity while software enabled economy of scale for hardware more generic than discrete solutions.

But it is also based on an architecture which solves even the largest problem by breaking it down into a sequence of very small steps, executed by a general-purpose CPU on a memory space that is shared between code and data.

Initially operations on larger numbers had to be broken down into a sequence of operations on individual digits, just like we learned to do sums in elementary school. Likewise, a single high-level language operation, such as a multiplication, would be broken down into a sequence of machine level operations by a compiler or internally by the CPU itself. That didn't matter, because the competition was manual or mechanical and electronical computers were getting faster very quickly. Cycle times shrunk x4000 in the three formative decades of personal computers ((1972-2002 740KHz (8008) to 3GHz (Pentium IV)), while the work completed per cycle also increased via the widening of the architecture from 4/8-bit to 64-bit and by pipelining complex instructions to the point where all operations on scalar data types finish, on average, within a single cycle. The combined effect was a scalar speed boost of x500000 for a line of Cobol code written in 1972 and running in 2002.

Computers became so fast so quickly, that their human operators could not keep up and had to be replaced by a piece of software called the operating system to organize batches of jobs. They were so fast, that they could be multiplexed among several, even hundreds of users, each of whom would seem to have the computer to themselves. In the art and science of software engineering generations of students were trained to break every problem into small reusable sequences of easy-to-understand code. Practically all current operating systems, programming languages and the vast majority of existing code assets were created on this architecture where the central processing unit was so fast, it was cut into tiny time slices to share among many workloads.

But at the dawn of the new millennium that changed fundamentally when silicon physics hit the "Gigahertz Wall (or alternatively as the breakdown of the Dennard scaling)" [1] and a sharp uptake in energy consumption and heat erected vertical barriers to economical operation beyond



Source: https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/

that point. The physical and economical barriers of process shrinks rise somewhat slower, but the effect of both on the economy of software is much bigger.

For a brief moment two decades ago, high-level languages and CPUs met on equal footing, as every statement expressed in a high-level language could be translated into a machine level equivalent delivering a result every cycle at the maximum speed of silicon physics. At that point, the architecture of CPUs could no longer be improved without sacrificing general purpose and high-level languages. As sequential operations on scalar data types could no longer be sped up, going parallel was the only way forward.

Special purpose vector extensions were developed for scientific computing as early as 1960 [2] and have since even trickled down into smartphones. A single such instruction can easily require several pages of high-level code to replicate functionally for a scalar CPU. In fact, programmers would have little choice but to write such code, both for portability and because "high-level" languages offer no vector intrinsics: while the programmer may be well aware of a hardware's vector facilities and tries to organize his scalar loops for easy unrolling, he relies on the compiler and sometimes on annotations (much a secondary programming language hidden in comments, that a normal or non-vectorizing compiler would ignore) to turn his loops into vector operations for the expected speedup. Where the nature of the problem permits and the pressure on speed outweighs the pain of writing and testing such code, it is used as long as it delivers advantages.

With scalar CPU power exhausted, the transistor windfall from process shrinks led into a veritable 'core explosion', 4 to 64 cores between 2006 and 2019 on a desktop budget, even smartphones regularly have 8 or more CPU cores today, while the cores in the graphics processing part of their system on chip (SoC) count in thousands. The processing units of these GPUs trade a much simpler per-core design for vastly greater numbers on the same silicon die area. As a consequence, they operate at perhaps 1/10 comparable CPU speed per core with traditional sequential code, but when you're able to put a 1000 of them, each with 8-64x (Depending on precision: 8x for FP64, 64x for INT8 are typical) vector extensions, to operate on a problem in closely synchronized lock-step, they deliver 100x CPU (or 800-6400x scalar speed) performance at similar transistor and electricity budgets, but not with Cobol source code.

The last two decades have seen an almost complete inversion of how we need to operate and program computers. For half a century processor speeds increased at a pace, where iteration was discarded as a limitation. Instead operating systems focused on loading as many parallel tasks as possible on systems to avoid idle time, while programmers made iteration and induction the most important ingredients of programming languages and algorithms to create the highest quality and reusable code.

Today any single social media application, HPC workload, ERP system, even a leading smartphone app need to marshal thousands if not billions of heterogeneous computing cores to satisfy or convince their user base within the tiny time and energy budget these applications can afford without losing user attention or to a competing service.

Imagine cooking for Christmas or building a brick and mortar house: the first few helpers offer quick gains but a hundred or a thousand extra hands require re-inventing

Image: ID 29575936 | ©Dmytro Tolokonov | Dreamstime.com

the process or rewriting the software from scratch, to avoid logjam. As a result, the expressive power of high-level languages today sits far below hardware level instructions, and their essential contribution to the economy of software, portability and reusability, is lost.

This affects seven decades of software heritage, almost every line of code ever written in a high-level language. Loops, once a clear indication a programmer had been able to abstract and generalize a problem, now just prove a workload was neglected to be parallelized. Code is no longer both, the universal inductive proof of mathematical correctness as well as the solution to the problem. Instead it has become a bundle of marching orders organizing wave-fonts of handlers on a specific machine with custom instructions, where neither the approach nor the purpose can even be guessed back from reading the source.

Code gets ever more expensive to create and cheaper to copy

The first computers had so precious little memory, even code was external, requiring physical re-assembly to change the computational sequence. Only critical reference data was kept in RAM and all other inputs were fed from punch cards or ticker tapes, while output went straight to paper output or again punch cards/ticker tapes for reuse. It was cheaper and larger memory that enabled the sequencing logic to be represented as code in the same memory and thus enabled the software revolution. Even so, only the most abstracted data – typically numbers – were processed in the old days. But with the cost of memory falling exponentially, data grew similarly, while code growth was much more linear, if that.

Creating code to perform a task remains expensive; correct, efficient, resilient, secure and trustable code much more so. But the encoding requires miniscule amounts of memory, especially if you compare it to data like video. Programmers may write a hundred lines of code per day, much smaller than a single video frame, millions of which anyone with a smartphone can produce with little more effort than a stiff arm.

What makes coding affordable is reuse: programmers no longer start on a green field but can copy seven decades of problem-solving know-how with a keystroke or mouse click, before they write the first new line of code. With that they resemble life on Earth, where everything that lives does that very same thing: first the full genetic repository gets copied and then is used to run life of the organism. A few mutations will get transcribed into the genome of future generations and selection then picks a winner on each cycle; code grows and differentiates depending on the habitat and the speed varies with selection pressure. That genetic repository is full of solutions that worked at least at one point in time of that species' evolution, but in the case of the human genome, perhaps as little as 10% is still in active use today.

The code base rarely shrinks, neither in IT nor in nature, because the slight increase in memory space is cheaper than cleaning up code. It's only when disused code contains vulnerabilities to foreign virus code attacks that it becomes a liability and ceases to be passed on to a next generation.

The extremely high cost of producing new code is typically offset by its very wide, if not global, distribution. But that usually just works out, when you reach that scale. Shrink-wrapped commercial software rode on the wave of personal computers selling millions of copies, at a time when code size still mattered and distribution could be controlled. The internet and compute density improvements make it possible to pretty much store all applications ever produced for a field of interest on a normal PC resulting in tumbling revenues from commercial software. The vendors who survived have learned to create lock-in and addiction, subsidizing convenience and a low initial effort or price and charging when the effort of changing habits seems more costly than paying up to the vendor.

From fixed function or purpose to meta programming

Originally every program was written for a single purpose, even if iterating over many different input data sets. As systems became bigger and more complex, they were soon split into specialized reusable pieces, copied and re-assembled for similar and other tasks with minimal extensions. Today we have ERP systems, constructed much like the factory floors and supply chains they operate, or interactive software instruments that allows us to create or assemble at a higher level with them, meta design or programming tools. They could just be rich documents and spreadsheets with formulas or macros, databases with queries and triggers, CAD software with parametric design or physics simulation plug-ins far more costly than the base application, websites with much more code than content, or it could be complex animations, simulations or game worlds of galactic proportions using several layers of automation and different ones for outside appearance, content behaviour and the interactions with hundreds of thousands of participants in a single game world.

Because "interactive" means human and manual, wherever a critical mass accumulates, APIs handle-bars are added to enable further automation, perhaps even with AI and we see a trend for orchestration languages, some very domain-specific, to code and operate at this meta-level… which may then go through a similar process on the next layer.

The affordability of clouds comes from the degree of automation they achieve with their many-layered global operating system. Facebook is extreme, a single application controlling millions of servers and capturing billions of users and their devices, using perhaps dozens of meta programming levels. It is interactive the other way around, with consumers reacting to machine inputs generated from code and AI bots.

Evolution at very different paces

But while new types of code and systems are developed, they don't universally replace all of the old ones. They solve arising challenges and may use new approaches to do so. They cause a Cambrian explosion of software and hardware species where the evolutionary pressure is fiercest, but other areas, even important ones, see hardly any change. It follows the same patterns as nature, where we still see the most primitive single cell life forms dating back to

the origins of life, past coelacanths little changed over the last 100 million years to the wide variety of Darwin finches [3], who had to exploit every Galapagos island niche, but whose homeland only emerged from the sea as recently as 3 million years ago.

Software follows the rewards

The mobile phone apps market resembles Galapagos in the sense that both Android and iOS rose from the sea with only the bare minimum of programming tools necessary to launch life, but a huge potential market. To tap into it, programmers had little choice but to learn the language chosen by the eco system creator, Objective-C for iOS, a variant of Java for Android, and dive deeply into the respective APIs, libraries and toolchains. Cobol and Fortran never made it across the internet ocean from server continents to the mobile island even if their operating systems build on Unix roots.

The frantic pace of mobile app development created such high stress on both Java and Objective-C, that new programming languages sprung from the need to code more efficiently, but still perfectly molded into the dominating paradigm: Kotlin [4] is undistinguishable from Java to Android, once the compiled code arrives as an app. Swift [5] likewise makes a huge productivity difference for the application developer, but none whatsoever for the iOS platform once compiled.

Programmers flock to these languages because in the mobile apps space, few things are as crucial as being the first to offer a new service and the most aggressive to add new variants and features, outpacing any competitor who ends up asphyxiated from lack of download revenues on pages 2 to 400 of the app store, where few potential clients ever wander.

The Go [6] language was created at Google as an alternative to C and to avoid C++ mainly because with those the repetitive parsing of include files slowed the recompilation of Google's giant and constantly changing code base adding multi-processing and networking support within the language. Principle use cases are scale-out orchestration code like Docker and Kubernetes, or most of what runs the server side of the Internet.

The Rust [7] language was created by Mozilla for the opposite end, with the principal focus on safe multi-core programming within a single application, the browser in this case. The browser has become the single most important application on all client platforms and a hotly contended battleground for software supremacy. In both cases the single original use case was under such intense evolutionary pressure for safe efficiency because of the giant scale of its application (millions of servers, billions of users), that it seemed to justify a new language to ease the pain of constant refactoring of the most critical parts alone. Both languages are also seeing adoption outside those initial use cases, but without that pressure and importance, sticking with C or its derivatives would have been the lesser effort and path most likely taken.

SPARK [8] was defined as a formally verified subset of ADA for use in flight or train control systems, where formal proof methods were required for verifiably safe operation, much smaller bodies of code than a browser or cloud orchestration system. But where Rust aims at avoiding crashing thousands of browsers every second with monthly releases, SPARK's aim is to avoid thousands of planes carrying hundreds of passengers at every moment ever crashing over two or three decades of operation and software release cycles somewhat less agile.

At the same time, we see extremely specialized software packages for oil field extraction or car crash simulation that carries a price tag of millions for a single installation, so nobody bothers adapting it to run more efficiently using the lower-cost accelerator capabilities of modern hardware: with only dozens or thousands of installations on the planet, the cost of refactoring the code easily exceeds the price tags of the biggest traditional scale-up hardware.

Value-driven computation

IT has always been cost-conscious if only because early computers were as expensive as spaceships. Yet where until recently IT calculated its business case within a relatively static Total Cost of Ownership (TCO) model outside an applications view, the huge scale-out applications Facebook, Google, Amazon and their Chinese counterparts are running need to calculate their profit and loss balance continually themselves. They sell or use the power to change consumers' minds: their business model could be described as venture analytics or electron capitalism in the sense that they constantly need to weigh the energy-dominated cost of their computations against the value they generate. And that value is much more limited than the computing capacity available. They are paid by producers of goods and services either to influence consumer purchasing decisions directly via advertisement or indirectly via analytics and recommendations, but marketing and sales cost can only be a part of the final purchase price, typically small.

Perhaps the amount of data will continue to explode, but there is no reason why consumer purchasing power would follow that trend. It may actually suffer from human workforces competing ever more globally and with robots and AIs, neither of which earn salaries to feed back into economy. Consumers need to pay for necessities first and can be influenced mostly on how they spend what remains. The internet giants hold the steering wheel for content, they can also step on the accelerator, analyze deeper or increase the advertisement ratio, but human consumers are a very slippery road and the giants' business model can lose traction earlier than they expect.

As Google, Amazon and Facebook start stepping on each other's toes in search for growth, they need to become smarter in where they spend energy on analyzing ever bigger amounts of data. Whenever they receive a bit of data, they spend electrons and decide to discard, harvest, forward, correlate, analyze, store, trust, augment, predict, multicast the input in order to generate more value than they invest, while of course the value and the effort are very

dynamic and the consumer still can resist even an optimal recommendation.

Classic computing is built on binary, true or false, correct or invalid. Non-repeatability, or that a piece of code might return distinct results for the same inputs is considered a failure: few of us would accept a bank statement balance preceded by "roughly estimated at…". Yet with Google searches or Facebook recommendations, not many will complain about missing a match or two while sifting through the millions of responses returned within the blink of an eye (while the fact that certain matches keep getting listed even after data owners had filed for delisting is becoming a growing legal issue). The paradigm that the quality of the result very much depends on the effort invested, and that the effort depends on the potential value, was crucial for evolution of the web giants under economic pressure and is sure to find willing adopters elsewhere.

Value-driven architectures come from the need to no longer just compute or analyze every bit of input given, but to continuously decide on if and how to proceed on data, based on a hunch of its value, tuned in carefully designed feedback loops. From *computing on-demand* it turns into *computing on value opportunity* and the competitive pressure is on extending the depth and the quality of the analysis without raising the cost, much—if not most—of which is energy. And because faster clocks or more instructions per cycle of general-purpose compute has become impossible to obtain economically, it means new hardware and software approaches are required and invented, but always at low budgets and just enough quality.

IoT is similar to the GAFAM business model in that the main enabler and driver is generating more value from data than expended via energy and all other cost and risks. And while most IoT-related services will have their counterparts in the cloud and in perhaps various stations along the path from the edge to the centre, the cost distribution, life cycle requirements may be as different as the things turned smart: some will enjoy a surplus of harvestable energy or sit right on the powerlines as smart meters, others may need to operate on a single charge for life. And since value and opportunity can be highly dynamic, a device needs to be aware of them. A smart fire alarm will try to conserve energy most of the time, but choose to expend a life's worth in high-frequency reports when it's burning, while activity in hybrid solar and battery powered edge devices follows the sun.

The only constants are likely to be severe price pressure and that they'll rapidly evolve with all essential fidelity wherever competition is most ferocious; they'll need to take baby steps, as internet giants and IoT can ill afford to stumble in the presence of alternatives, but at extreme speeds to stay ahead.

Value-driven architectures for the Internet of Things must satisfy essential rules and this is the most important one: *while digitalization may expand in number of nodes involved or in the depth of their processing, newer generations of technology cannot cost more than what they replace.*

IoT is also quickly becoming an intrinsic part of the GAFAM business model, and they are the first to try bend that rule: Android devices, smart speakers and household appliances as well as browsers are turned into both sensor data gatherers and the first levels of compute and analytics platforms. Instead of just pushing raw data into their data centres, where they need to pay for energy, web giants try to push as much of the computing workload back towards the edge or client devices. From their standpoint that maintains the primary rule of IoT, because they can go wider and deeper without paying extra: not only are consumers becoming the product [9], now they even pay for being sold.

Designing software is social and human

Programming is a very social activity, even if the actual coding typically requires a high degree of concentration somewhat incompatible with chatter. It starts with the fact that we learn first and best through imitation, just as we learned to speak by listening to our mothers. Reading code that is both well-made and easy to read helps develop the skill, and it also helps reuse the code.

What you consider easy to read, depends very much on what you have read before and how much time you spend both reading and coding. Cobol [10] reads almost like plain English and was designed so even a non-programmer would be able to make out what it is doing. But it can be as exhausting as some of the old mathematical proofs in the original Greek; the language may be an obstacle, sure, but it's the lack of a concise mathematical notation (and perhaps a picture), the sheer verbosity required to describe the very abstract, which makes them extra difficult to understand.

APL (A Programming Language) [11] has always represented the opposite corner, offering an extremely concise way of encoding program logic using mathematical symbols instead of English, but to the uninitiated it looks like Klingon. Both languages have enjoyed huge success and for a long time, but never with the same crowds.

But just as surely, even large communities will eventually die, after plenty of small ones have long gone. Sumerian [12], ancient Greek and Latin were commonly spoken only for centuries, but they survived millennia as classic languages for literature, religion and business as long as the content encoded with them remained relevant enough to motivate new disciples to learn and exchange. Cobol, Fortran [13], C [14] and vast bodies of commercial and scientific code written in them may survive in niche communities much longer than expected, but just as with publications in Sumerian, Attic Greek and Latin, the rate of content creation is slowing down significantly, while new dialects and variants continue to flourish best where crowds of enthusiastic users make for fertile ground. RatFor, B, D, Algol [15] and various 4GLs share their fate with Etruscan, Hittite, or Aztec, who despite historical importance and fame, failed to achieve long-term success. When the code or the knowledge encoded with them can't be black-boxed as a service like sages in a temple, it would need to be re-invented.

Some software dinosaurs will stay with us, many died unnoticed

Life on Earth wasn't designed, which shows in skeletons. The first type, 'exo' or outer skeleton got species moving much faster than mere skin, both to eat and to avoid being eaten, creating huge success and many variants, but also a serious size limitation, especially on land. The size of a specimen is determined as it transitions from the pupa or nymph form into an adult and that's why even the biggest locust won't ever reach the size of a dinosaur, even if size, as dinosaurs proved, can be a giant advantage. But since nobody could afford to throw all that genetic code away and start fresh from a single cell design, some individual popped an organ or a limb outside the shell and made that stick. It even turned it into a full new growth path, where all of the mobility parts had much more room to expand, even to the size of a brontosaurus, which still started with a single cell for every individual, but included genetic code that had long outgrown the limits of an exoskeleton. The 'endo' or inner skeleton was an evolutionary approach, that preserved working code and concentrated change on where it delivered the biggest competitive benefits. We still use exoskeletal code for our head and around our heart, because it doesn't obstruct critical growth there: brains are big enough at birth and ribs are the genius compromise somewhere between exo and endo.

The code that implements the TCP/IP [16] protocol inside an operating system was written to handle the thousands of things that could go wrong, when a data packet was send from the University of Hawaii to Stanford via dozens of relays over radio in a storm. All of these potential pitfalls are still checked, even when the simulated network between two virtual machines is in fact the single physical memory bus they share for code and data on the host they run on, … or when that tempestuous radio network is in fact a fabric [17] Google purpose built for their data centres with extreme reliability and transparent fault tolerance. At the size of Google with millions of physical machines, this needless overhead became so painful that its engineers decided to cut from the code all errors that cannot occur in its fabrics or use cases. Google is now reported to have 100 different implementations matched and directly linked to each application [18], completely circumventing the Linux TCP/IP kernel code, which is still there and even used for booting the machines.

Maintaining such a great number of implementation variants can become quite an effort in itself. But a clean rearchitecting or modularization that is generic enough to pass back to the Linux community for maintenance is also a significant task and a process far too lengthy to perform at the speed Google requires. Generally, code evolution is fastest with the cloud giants, but will still just favour the least effort required to obtain the critical competitive advantage.

In some cases, projects to completely refactor a software stack still occur and succeed, most visibly with browsers, which are among the most important pieces of code in existence, by their concurrent rate of execution with billions of users.

In other cases, their success is rather more limited as with OpenSSL/LibreSSL/BoringSSL [19]. OpenSSL was discovered to have significant cryptographic weaknesses, but the code base was such a huge pile of code from so many sources with questionable quality, that two groups felt it was necessary to restart from scratch. Neither have succeeded to deliver a fully functional replacement so the race is still open, as it will remain in many similar scenarios.

Dinosaurs are still with us: we call them birds. Like browsers they have changed their feathers to the point we don't even realize their legacy roots. Yet while browsers only came to be three decades ago, some Cobol applications continue to run as-a-service in black boxes after decades, simply because there was too little benefit for the cost of a new implementation. Every major ecosystem change will kill entire bodies of software like dinosaurs and much code will disappear, rarely with major impact. Some will remain with little change and some will change so completely its roots are hard to recognize. In every case, engineers will gauge the benefits and risks of putting a bypass or patch on a hotspot vs. a wider or complete refactoring, with the decision and results depending on the context and the ecosystem around it.

Linux started as a Unix clone, but the pressure on its code quality is so high, it is seriously considering Rust code in critical device drivers when C and Unix were considered symbiotic for decades. And while reimplementation of a full OS like the Linux kernel in formally verified SPARK for airplane or automotive use seems beyond imagination, a refactoring of tiny hypervisors like OKL4 [20]/Sel4 [21], which is already formally verified if in C, is much more likely and could coexist with e.g. a Linux system inside security enclaves (it is used on iPhones), as management engines or run it as guests.

There is no stopping software until it dies or kills

The giggling smart pet we give to our children, the smart baby monitor or light bulb, the last fridge, coffee or washing machines we bought were sold as an improvement over their predecessors, which still used embedded controllers, discrete electronics or even mechanical cores to achieve their function. Their internet connectivity and augmented services were advertised as extra value while in fact all cleverness lies in the employment of a full functional equivalent of Unix mainframe at their core as a cost cutting measure.

That mainframe has shrunk from a full room to less than a fingernail, yet may run much faster with a single battery for life and sell for pennies rather than millions, but by the million as a result. The extremely rich functional base of a full open source Linux software stack allows the vendor to implement the entire functional logic of all these devices in perhaps a couple dozen lines of PHP code by a college student as class assignment. Previously a qualified specialist might have been required and had to use a commercial development

environment for a proprietary embedded OS; similarly, an electrical engineer and various prototype iterations for the discrete electronics or a complete separate assembly line for the mechanical solution might have been needed.

But that mainframe and its software were designed to run accounting and even a moon mission. It was written to correctly process all permutations of well-defined inputs which paid for the expensive mainframe. It wasn't designed to withstand the near infinite cross product of the complementary inputs, almost guaranteed to break it in a manner that can be observed on a digital clone and thus turned into an exploit. It wasn't written to survive immediate attacks from the industrialized cyber-armies, which will convert everything with a known vulnerability into another botnet member. It could still be safe, if it ran well maintained in a secured data centre, behind protective layers of firewalls and intrusion detection systems, instead of your home WiFi. To the MIRAI botnet [20] and for most ransomware trojans these micro mainframes are fully functional giants and offer the additional advantage of well documented vulnerabilities easy to exploit and never closed for lack of an update channel. The potential physical and financial damage from legacy code running on shiny new gadgets vastly outstrips any potential benefit, yet the relentless push for cost reduction and novelty will drive code into your pacemaker, your car, your home, in short everywhere it was never intended to be by those who wrote the vast majority of it, which just got carelessly copied by someone not interested in the implication.

The efforts to contain that risk can't really be bigger than the value gained by expanding the reach of IoT. It is for this reason that finding cheap replicable countermeasures is one of the highest priorities for any vendor or government trying to create value from information technology today. We highlight some technologies in "Reversing John von Neumann and Steve Jobs, but not software".

References

[1] Robert H. Dennard, Fritz H. Ghaensslen, Hwa-Nien Yu, V. Leo Rideout, Ernest Bassous, Andre R. Leblanc, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions", https://www.ece.ucsb.edu/courses/ECE225/225_W07Banerjee/reference/Dennard.pdf

[2] Richard W. Hamming, Adward A. Feigenbaum, "Computer Structures: Readings and Examples", Illiac IV, http://gordonbell.azurewebsites.net/cgb%20files/computer%20structures%20readings%20and%20examples%201971.pdf

[3] Galapagos Conservation Trust, "Darwin's Finches", https://galapagosconservation.org.uk/wildlife/darwins-finches/

[4] "Why teach Kotlin", https://kotlinlang.org/education/why-teach-kotlin.html

[5] John Timmer, "A fast look at Swift, Apple's new programming language", https://arstechnica.com/gadgets/2014/06/a-fast-look-at-swift-apples-new-programming-language/

[6] Russ Cox, "Eleven Years of Go", https://blog.golang.org/11years

[7] "Rust, A language empowering everyone to build reliable and efficient software", https://www.rust-lang.org/

[8] SPARK Ada, https://www.adaic.org/advantages/spark-ada/

[9] Jason Fitzpatrick, "If You're Not Paying for It; You're the Product", https://lifehacker.com/5697167/if-youre-not-paying-for-it-youre-the-product

[10] R.W. Bemer, "A View of the History of COBOL", http://archive.computerhistory.org/resources/text/Knuth_Don_X4100/PDF_index/k-8-pdf/k-8-u2776-Honeywell-mag-History-Cobol.pdf

[11] Rajat Acharya, Nitin F. Pereira, "APL Programming Language", http://courses.cs.vt.edu/~cs5314/Lang-Paper-Presentation/Papers/HoldPapers/APL.pdf

[12] Martin Worthington, "Ancient language of Babylonia is brought back to life", https://www.alumni.cam.ac.uk/news/ancient-language-of-babylonia-is-brought-back-to-life

[13] John Backus, "The History of Fortran I, II and III", http://www.softwarepreservation.org/projects/FORTRAN/paper/p25-backus.pdf

[14] Brian W. Kernighan, Dennis M. Ritchie, "The C Programming Language First Edition", https://archive.org/details/TheCProgrammingLanguageFirstEdition/page/n7/mode/2up

[15] "The Algol Programming Language", https://web.archive.org/web/20161006113915/http://groups.engin.umd.umich.edu/CIS/course.des/cis400/algol/algol.html

[16] Lawrence G. Roberts, "The evolution of Packet Switching", https://web.archive.org/web/20181231092936/http://www.ismlab.usf.edu/dcom/Ch10_Roberts_EvolutionPacketSwitching_IEEE_1978.pdf

[17] Arjun Singh, Joon Ong, Amit Agarwal, Glen Anderson, Ashby Armistead, Roy Bannon, Seb Boving, Gaurav Desai, Bob Felderman, Paulie Germano, Anand Kanagala, Jeff Provost, Jason Simmons, Eiichi Tanda, Jim Wanderer, Urs Hölzle, Stephen Stuart, and Amin Vahdat, "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network", https://dl.acm.org/doi/pdf/10.1145/2829988.2787508

[18] H.K. Jerry Chu, Yuna Liu, "Use Space TCP – Getting LKL Read for the Prime Time", https://storage.googleapis.com/pub-tools-public-publication-data/pdf/6ece8c450ee375eb31b2795a2a227a5d8c3598b7.pdf

[19] Alessandro Ghedini, "Make SSL boring again", https://blog.cloudflare.com/make-ssl-boring-again

[20] General Dynamics, "Hypervisor", https://gdmissionsystems.com/products/cross-domain-solutions/hypervisor

[21] "The seL4® Microkernel", https://sel4.systems/

[22] Paras Jha, "Who is Anna-Senpai, the Mirai Worm Author?", https://krebsonsecurity.com/2017/01/who-is-anna-senpai-the-mirai-worm-author/

Thomas Hoberg is Technical Director R&D at Worldline, Germany.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: T. Hoberg. Extreme reuse: the only future any code can afford. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 176-183, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

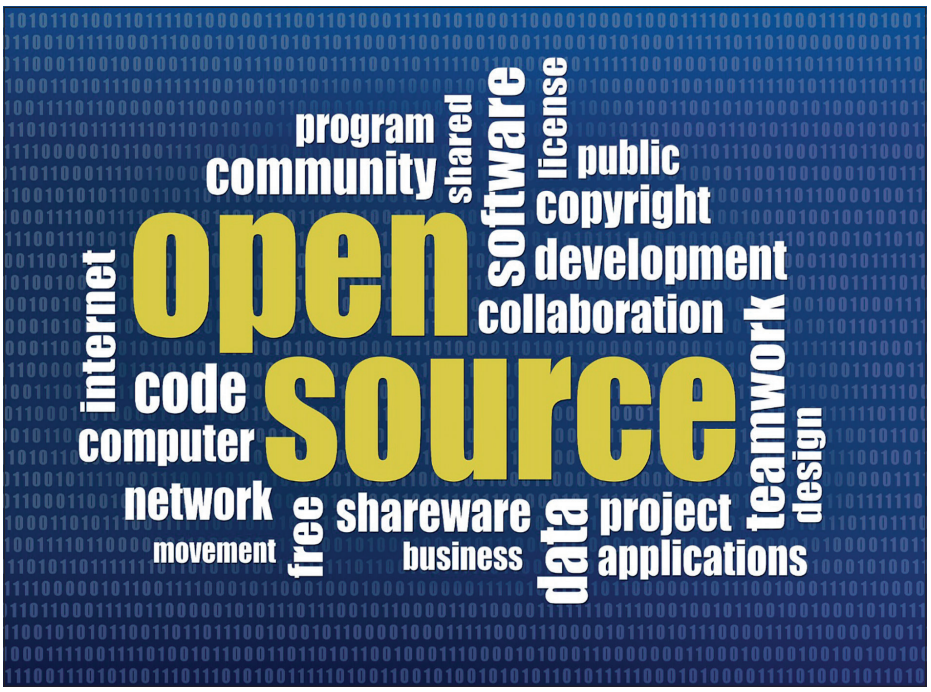
Excellent quality open source code can be extremely successful with the proper social code to support it

Open source code and content

By THOMAS HOBERG

Open source has become the dominant form of executed code and content; it is freely shared without costly licences and is available in greater quantity than what is commercially created. Behind both code and content is the fundamental social urge of humanity, to always offer something with minimal prompting. To create the most excellent open source software, it takes sophisticated social code; extremely well curated, wide and deep bodies of knowledge like Wikipedia and the big data economy treasure troves of social media.

As digitalization diffuses the majority of human interaction for the majority of people, the role of the social code around open source becomes much more important and far more political as it applies to, and even steers, more and more of us. This makes it a central topic for social, political and economic sciences as well as computer science.



Source: ID 40010486 Marek Uliasz - dreamstime.com

Key insights

- A million people each contributing just a little created Linux and Wikipedia. Getting a million people to agree on a common cause requires social code few can create.
- Open source has become as diverse as it has become huge and the difference between code and content is disappearing.
- Software giants have turned to eliminating competitors via zero-revenue open source on one hand, while holding customers captive with proprietary vertical code on the other.
- Open source software is resilient even in times of conflict.
- Sharing knowledge, experience and open source with less bias helps to create trust and reduce tension everywhere, even at political level.

Key recommendations

- Europe needs to provide freely the critical open software components that have the extra qualities it wants: privacy compliance, sustainability, support for multiple sovereignties. Compromising on EU values must not be cheaper than using what the EU provides to enable open competition for value created on top of this platform.
- Digital commodities so universally used that they become the public space, need to be transferred to public control and therefore regulated, maintained and taxed if necessary.
- Digital commodities need to be maintained as free open source; components can be given to companies much like public tenders for other infrastructure.
- Europe needs the ability to maintain a fork or EU branch for digital commodities. That takes critical mass and that critical mass needs to be maintained in government, military and education, who must favour those solutions over proprietary ones.
- Evolution of commodities is a strategic asset and needs to play to the ground rules, with diversity being a necessary inclusion.

Open collaboration is human

The overwhelming success of the human species comes from its ability to turn collaboration into an advantage far more important than agility, efficiency of metabolism, keenness of senses, strength, or resilience. The ability to learn, retain and communicate impressions, opinions and knowledge to our kin, more recently across time and around the globe, has allowed us to profit far beyond what any individual could ever learn alone. At its base is an altruistic readiness to share experience with minimal prompting, perhaps an outgrowth of the parental instinct, without which complex organisms could not evolve. As imitation and the following of shared advice turn into success, they transmute from chatter into knowledge that is passed on as tales and gets encoded into legends which form the fabric of societies. Abstraction turns legends into research, which engineers can build into industries, creating entirely new sources of knowledge and bodies of science, a collective genius beyond individual imagination, which is now augmented, accelerated, propagated and harvested by code running on computers.

Greed is just as human as altruism; we're much less likely to gossip about a patch of juicy berries or a hidden cookie jar, when we know we'll be hungry again tomorrow. Withholding knowledge or tradecraft comes naturally, especially when social distance and competition for scarce resources increase. Unilateral knowledge on how to start fire, plant crops and create superior tools, products and weapons decided the fate of our ancestors vs. their opponents, but only until the competition caught up.

As societies evolved and functional differentiation deepened, some turned into artisans, artists, sages, engineers and scientists, entirely invested in advancing the state of the art. As the specialists increasingly relied on sponsors for their livelihood, they employed their growing affluence to support the keeping of accounts and records, history and knowledge; scribes became a privileged specialty, writing an artform. Copying a handwritten book was quite laborious and only permitted under tight control, with gifts of equal value expected in return. The copy was therefore almost as valuable as the original. All that changed when Johannes Gutenberg introduced mass production via printing, and knowledge and art could be pirated with ease and at little cost. It has reached a new quality with the proliferation of the internet and digitalization, where all codified knowledge, art and content can be accessed and copied at negligible cost, while creation remains as costly as ever.

Societies faced a dilemma: on one hand the creators needed reliable remuneration to tie their fate into furthering the state of the art. On the other hand, too tight a control on the distribution of knowledge and art would clearly curtail an individual's ability to improve himself and society as a whole. Striking the proper balance, promoting fair use while discouraging abuse, is already a complex topic at local scale; across borders and value systems, it has become a major part of politics. Today's trade wars are as much about control over immaterial assets as they are about physical resources.

Open source and intellectual property are political

Richard Stallman's main motivation to create the Free Software movement lay in the belief that locking away software source code was as immoral as the enclosure of common land in feudal England, where land ownership came to be reinterpreted by nobility in a manner that deprived commoners of the traditional usages of the area around their villages they had enjoyed since times immemorial. Resources traditionally shared by all who worked the ground gradually came under the exclusive control of a few, thus forcing villagers into contracts to work for hire or even indentures and emigration to avoid starvation.

To Stallman, source code represented human knowledge on how to solve problems, to be shared among all those who are interested and contribute, because exclusive ownership stifles the progress of society. The GNU General Public License, GPL or CopyLeft was designed to ensure that all knowledge embodied in software source code published under it, would always be free to copy and use, without any individual, company or government gaining exclusive control.

On the other hand, the English land grab also enabled a far more efficient agriculture through scale effects and created the human and capital surplus that enabled industry: pooling manpower and money under the control of a few jumpstarted the industrial revolution of which IT is a part. Patents, trademarks and copyright allowed for the steering of vast pools of talent and resources, under the direction of a few visionaries, to accelerate the development of entire industries, while hunger, sickness, war and dire needs provided the basic motivation to spur leaps of imagination and progress.

Driven to their extremes, both variants (free sharing of all things; everything under the control of a few) seem similarly capable of producing drastic innovation as well as debilitating stagnation, with the pendulum switching direction at different paces for different markets, populations, countries and value systems, causing stress, relief, strife, détente or simply change… otherwise known as progress.

Open source and competition

A very young Linus Torvalds was inspired to create his own operating system in 1991, when he read in a manual for his Intel 80386 CPU that a complete task switch—the very core functionality of a multi-tasking OS like Unix—could be achieved with a single jump instruction to a task state segment: it seemed to fit the crucial part of an OS on as little as a single page of code (currently more than half a million pages for Linux). He overlooked that this single instruction took hundreds of micro-programmed CPU cycles to execute and the resulting performance was so atrocious [1] that everybody replaced it with full routines that performed much better. His OS also had a file system so primitive that it was practically unusable. The significant advantage was that its source code, which he was proud not to use, came with a book by Andrew Tannenbaum. A book that he seems not have finished reading, because it carefully explained why monolithic kernels, like the one Torvalds was creating, were already outdated.

At the time there were two major Unix variants readily available for 32-Bit PCs: AT&T licensed System V.R3 and 386BSD, the latter of which was even open source. Both were mature, stable and performant ports from the ubiquitous VAX architecture, ensuring a rich software and user base.

Linux should have died in its crib against such competition. What changed the world of IT was that Torvalds posted his source code on the internet under the very same licence as Richard Stallman's GNU-C compiler that he had been using, incorporating its libraries and tools.

Very soon, people who actually knew how to write an operating system replaced his code. And with hundreds of developers contributing, many of them from front line research in academia, and much bigger numbers tracing bottlenecks, misfeatures or outright bugs with the help of the source code, the quality of Linux eventually eclipsed all commercial Unix variants and open source BSD. One key catalyst for open source code Linux was the social code of the GNU CopyLeft, and its diode effect, which assured that all effort given freely in contributions could not be claimed back or misappropriated by anyone, but would remain open source perpetually. Another key ingredient was Linus Torvalds's quality of judgement, from his ability to recognize and his readiness to accept superior code, to the manner in which he established community governance, or even adapted his personal code of conduct therein.

Linux and Wikipedia are the best-known examples of open collaboration initiatives, where humans are willing to invest effort into a universal body of encoded knowledge, because it feels intuitively beneficial to most, or because they become intellectually convinced it is a better approach than withholding code or content for coin. They also prove that the quality of the product, in terms of reach, functionality, resilience, diversity and completeness, can exceed what even the largest commercial or public sector entities might achieve under exclusivity, even when each individual contribution is relatively small. Where the user base or the ecosystem expands to billions, the cumulative effect of small contributions and broad oversight for digital content become so big that it creates gravity at planetary scale. And the quality of the social code, in terms of group culture, regulation, governance and exemplary personal conduct, is a critical part of the success.

Facebook, Twitter, TikTok and WeChat show that humans will even collaborate when it is far less clear that they are to benefit the most from the effort they invest. Here the platform operators claim ownership over extracts, aggregates, insights and knowledge derived from the content their users 'share': they then sell to the highest bidder.

Open source: catalyst or bulldozer?

Superior quality, open source and zero licence cost make it difficult to compete with a commercial offering based on charging for closed source proprietary code. Putting out a begging bowl and programming like a Buddhist monk for a prayer, holds little appeal to engineer parents saving for their offspring's university tuition fees, so employers have sought elements of exclusivity to exact payment from customers. Some sell support services, warranties and certifications, others sell exclusive automation or value-added services and wherever competitors face off on level ground, a new type of price war has developed: zero

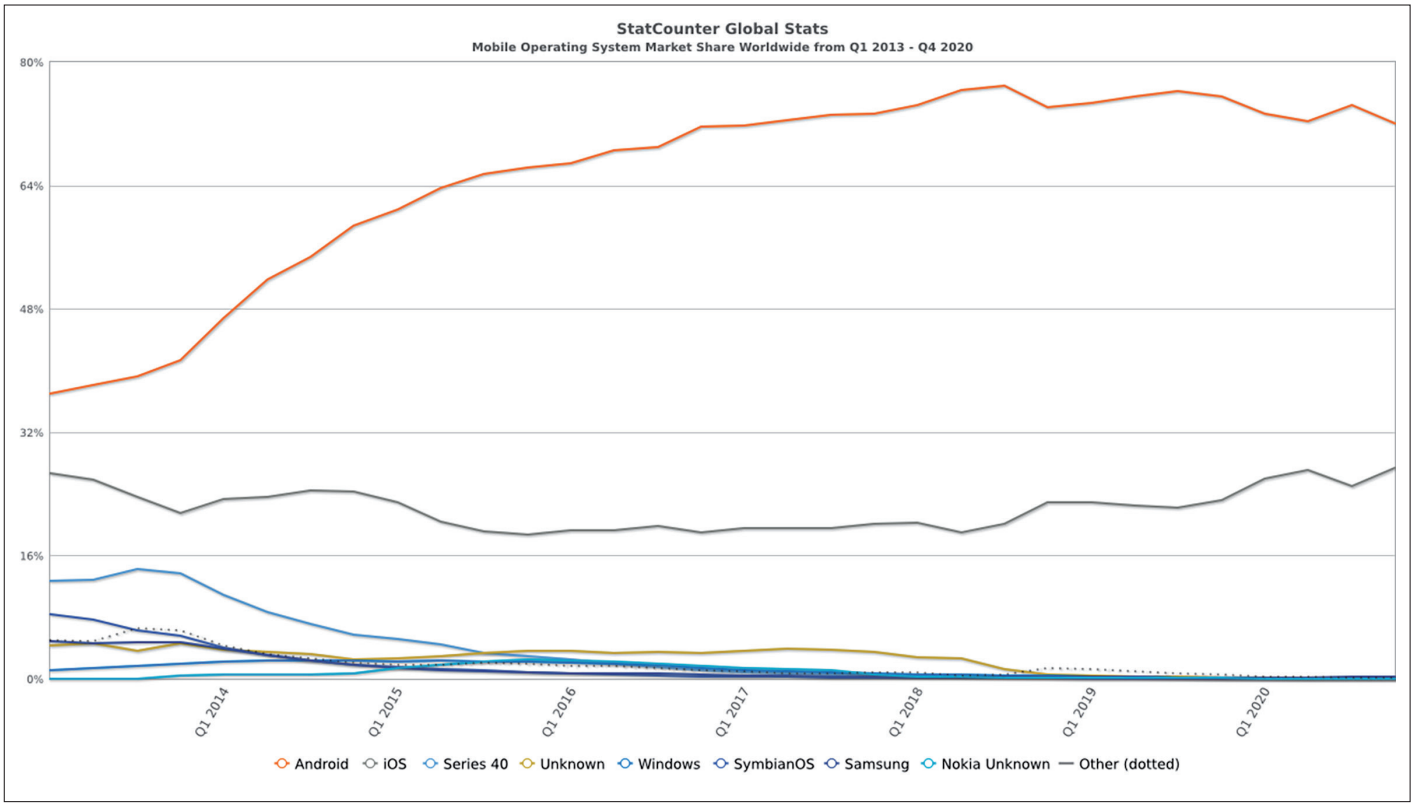


Figure 1: Mobile Operating System Global Market Share Q1 2013 – Q4 2020

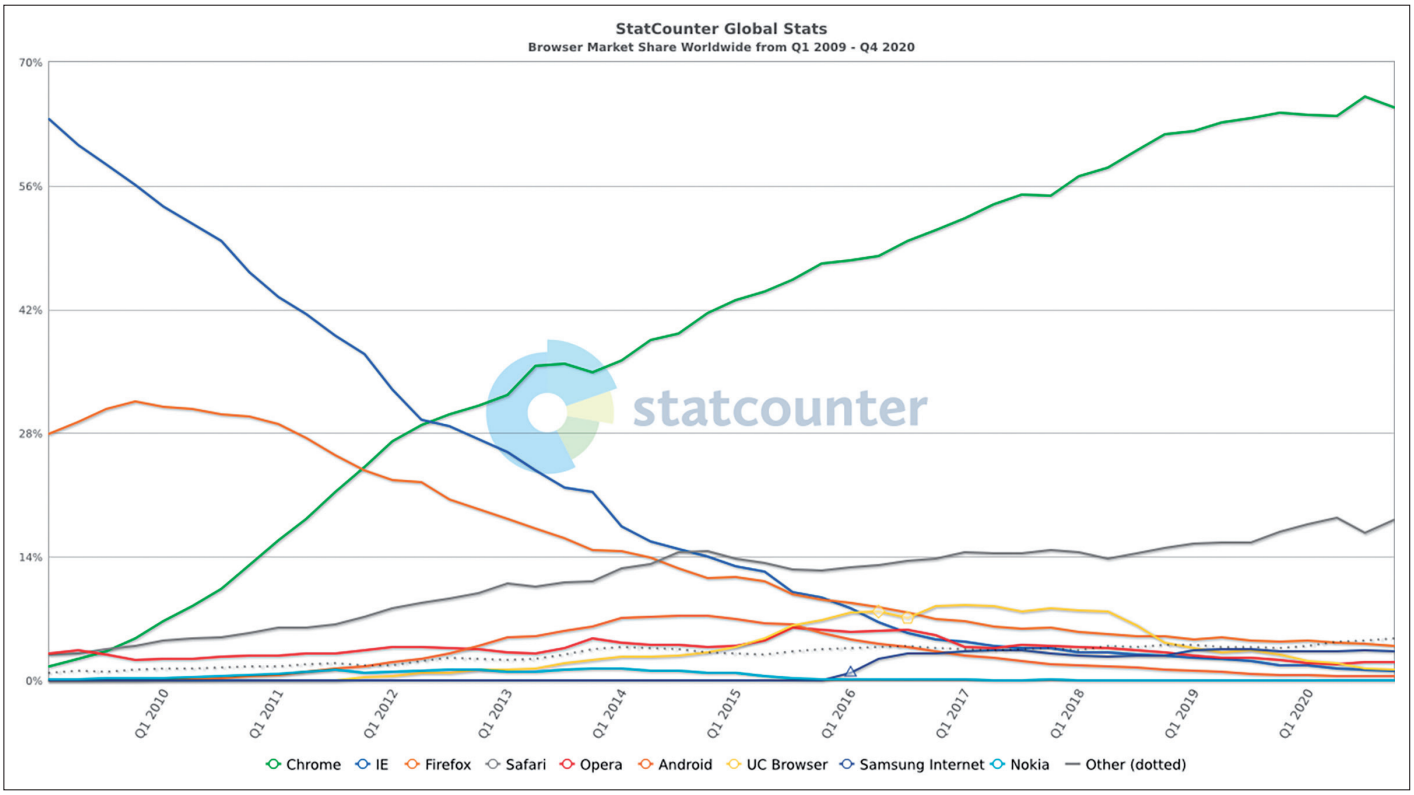


Figure 2: Global Browser Market Share Q1 2009 – Q4 2020

against zero, open source against more open source, even zero against contra revenues.

In the traditional markets of yore, where people created higher value products from purchased raw materials through work, fair competition tended to reach a balance at subsistence levels, because undercutting one another leads to debt and starvation. Guilds then fixed prices a little higher to benefit their members, a practice more widely accepted then. In software production per-item cost is zero, but your ability to charge for exclusive options rises sharply, when there is little competition left for the base product. It leads to IT companies providing a commodity core without fees or even with active subsidies, to stave off competition and achieve a practical monopoly, and then charge significantly for the value-added part. Without charge there is no price fixing involved for the base product, and in that manner, they so far avoid anti-trust legislation, which has not kept up with the immaterial world of software.

Internet giants have made this approach a core aspect of their business. Their giant user base and economy of scale has allowed them to reach a level of operational efficiency that is hard to match on-premise. And while they offer open source software components that allow workloads to onboard and take advantage of these operational gains, they do not repeat that with their secret sauce, those components which permit them to operate in such a manner: their objective is to lure customers into their proprietary clouds, not to seed competition.

To ensure Android would gain against Apple's mobile platform, Google took free and open source to battle on the client side, at a time when it seemed positioned to achieve with mobile phones what Microsoft had attained with personal computers. Covering all angles, Google made its smartphone platform: (i) open source; (ii) portable; (iii) free of charge; and (iv) irresistible, with popular apps like Maps, Mail, YouTube and Chrome included; and (v) it invested heavily into a developer ecosystem and support. Its goal was to undercut and outgrow all proprietary and licence revenue-dependent competition, and thus gain control over the platform the majority of users would use to interact with the digital world. It then could strip mine user data on the handsets for sellable insights in order to pay for the platform support.

Its gamble paid off: Google owns the global mobile base platform [2], apart from the iPhone enclave. It managed the same with the browser [3]: again, apart from Apple's walled garden and a Firefox holdout, it very much owns the main interface to the internet via Chrome, giving away the best open source browser for free, so nobody will be tempted to run the client side of the internet on an alternative.

In order to also crack the personal computer market with its more affluent enterprise user base and more valuable structured office data, Google created Chromebooks as a lighter, cheaper and secured PC platform, to run little more than a Google Chrome browser on an open source Linux kernel. Meanwhile, its web-based Workspace [4] acts as a cloud-based Microsoft Office replacement, with online collaboration benefits added. It removed dependence on decades of PC technology, Microsoft Windows and Office at once and enabled enterprise, government and educational use at a fraction of the opera-

tional cost. But it also created a dependency on Google cloud services, which are proprietary closed source, and for which it charges very selectively, ensuring minimal onboarding cost for habit-forming educational markets, where it has gained 60% in the United States. This dual pronged attack represented so big a threat to Microsoft that it transformed that company.

While Microsoft is following up on Google's approach to gain revenue from insights into user data, it can't easily sacrifice the revenue it obtains from the licensing closed source software. That is why Microsoft is working on moving all customers from local Office installations and document storage into their Azure cloud via Office 365. The company markets scale benefits but gains data control and plans its future on its ability to transition its exclusive hold on user habits formed during past decades to an exclusive hold on Azure for future generations. Microsoft offers Windows, Office 365 and Teams to educators at fully comprehensive prices, where open source and free licence variants lose public tenders, because their operational costs alone can be higher than a lock-in Azure offer. It balances its books by charging employers whose graduate employees are set in their habits.

Both Google and Microsoft maintain significant bodies of code and shoulder hefty expenses to operate clouds, which to them are very down-to-earth data centres. They need to develop, market, provide, operate and support it all; they give services away free to first-time users. Users' exclusive dependency on their cloud gives them dual benefits:

- It allows them to charge for profit, once customers are caught in a web of convenience or dependency;
- They can monetize all knowledge extracted and harvested from documents stored in their respective cloud with the help of AIs.

Microsoft's exclusive licence on OpenAI's GPT-3 [5] shows its determination and direction vis-à-vis Google's DeepMind acquisition. For every component and service that both offer to consumers, quality open source product alternatives are available for free, that neither create an exclusive dependency for the back-end, nor strip mine user data for valuable knowledge. However, these alternatives require infrastructure and manpower to set up, operate and maintain from the very start. While they can offer privacy by design, and may turn out cheaper and without the risk of foreign control in the long run, the current practice of cloud giants of selling below price to "influencers" makes them look less attractive, especially when sweeping pandemic-induced changes require a fast set up.

Europe's ethical demands for open, heterogeneous and sustainable IT products and services, which put citizens with their civil liberties and privacy first, also implies a hefty burden of compliance. Similar to how US giants seed their clouds by subsidizing browsers, Web-frameworks, mobile platforms and onboarding tools, the EU must level the playing field by putting in place:

- Assets enabling that EU-generated services and products can be sold domestically and abroad with the overhead for compliance overheads removed;
- Assets enabling foreign vendors to add EU compliance to their products and services with minimal overhead;
- Taxation on the value generated from inputs via the European use base, to pay for the above efforts;
- Regulation to ensure that compliance for users within Europe's borders can be enforced and taxes can be appropriately levied across all frontiers;
- Regulation to ensure that devices and services who exceed a threshold of popularity or population must use open APIs and be made interoperable, so they need not be tied to a single vendor.

If Europe wants IoT or CPS devices to operate with embedded software stacks that employ industry's best practice during generation and are continuously updated via security patches, it may need to supply an easy and convenient equivalent of an 'IoT Chrome', so that Chinese manufacturers looking for the cheapest solution prefer using the free EU offer instead of something they had a freelance student assemble years ago on a penny budget.

If Europe wants IoT or CPS devices from different vendors to work with virtual digital assistants on any cloud without the servant selling its loyalty to the highest bidder, the EU needs to put in place quite a few free building blocks which offset the overhead in development cost and complexity. It must then use regulation to ensure that, at least for the internal market, Alexa, Siri or Google's assistant don't profit from shortcuts in privacy or security.

Open source suffers the same fate as any other tool devised by humans. The first stick was used to dig out a nourishing root but, perhaps a minute later, it was also used to decide who would eat it. While open source was created with the idealistic motive to create fairness by maximizing knowledge sharing, one of its major use cases today is to remove fairness from corporate battles.

Why Europe needs to enforce and support open source

When code becomes used by the majority of people most of the time, every automated decision it makes becomes a political choice. When such decisions are even potentially made without the sanction or indeed against the local laws, culture or ethics, actions should be taken to ensure that the code or application follows rules, quite independently of how the code actually performs, or its original purpose and scope.

Google, Facebook, WeChat, WhatsApp, Tiktok or Twitter were not originally designed for the transformative power on society that they hold today, just like the first editions of MS-DOS, Unix, Android or Mosaic never aimed to run the planet.

Today, a ban on Google's Search, Android and Chrome or Microsoft's Windows and Office could completely cripple Europe, while similar bans on Facebook, WhatsApp and Twitter might just do the opposite. With decisions taken recently by the US government, what seemed like a remote possibility a few years ago has become a real threat everyone needs to plan for. It also became the urgently required wake-up call that the natural threshold of intervention against the 'legislative power'

of web-giant code was actually crossed a long time ago.

Defensive measures: reducing the attack surface

Proprietary closed source products from companies which can be coerced by governments outside the EU and which cannot be easily substituted, must be regarded as a threat to European sovereignty.

It is not much better if the products are open source, but encumbered by licences or intellectual property restrictions that prohibit their continued use in case of trade conflicts. If those restrictions cannot be negotiated or regulated away, substitution is required.

Social media or content platforms with algorithmic transformation of content need to be matched to political borders and either opened for regulation or blocked. Privacy can be viewed as a very local censorship bias and achieved when regulation works.

Enabling measures: increasing market potential

When Europe wants to impose restrictions on how web-giants strip mine all data that flows through their platform for sellable insights, those might counter by reducing their services and stopping the distribution of the client-side assets. Android, the Google Search/Play Store/Chrome/Maps/Mail etc. might have to be replaced by local alternatives as they are in China already and for Huawei especially, which may cause significant inconvenience for the existing European user base, but no essential loss.

If Europe prepares itself to take up the slack, it doesn't have to start from scratch: for Linux, Android, Office and most other "digital essentials", most assets are open source. Even if their vendors stopped their distribution in the EU, they could be substituted with funding perhaps not too different to what European customers currently pay in licensing fees.

It also becomes an opportunity to break up the inherent link between the corporate business model and corporate ethics, and to insert modularity and compatibility with greater variances in local ethics and content regulation. It is a two-sided sword, which can deliver far better control for achieving civil liberties and privacy for EU consumers, but would also enable more undemocratic government controls elsewhere. Europe could contribute to a truly global code base, if it is prepared to completely separate out business model and ethics to make them modular and configurable.

There is no reason why even China, Iran, Cuba or North Korea might not want to collaborate with the EU on a single code base for all digital essentials, if everybody gets what they want, nothing of what they don't and can rest assured that nobody can hijack the project or critical parts of it. Even Google, Facebook and the UK might join back in eventually, when the governance is done right.

Perhaps most importantly, providing a PC, server, smartphone and IoT platform independent from domination via foreign governments enables the market of trillions of smart device instances. In such a market, cyber-physical systems that have all the virtues and qualities Europe demands can be created: full sovereignty for the owners of virtual personal assistants, their knowledge base and the devices they manage, with all the necessary loyalty, discretion, sustainability and security to give them significant value.

Open source needs more support

Wikipedia [6] is a prime example of open source content creation and curation that manages to do without paying the vast majority of its contributors; while the quality and value thus generated is hard to gauge, unless it is monetized via its use as carefully curated input for the latest machine learning models.

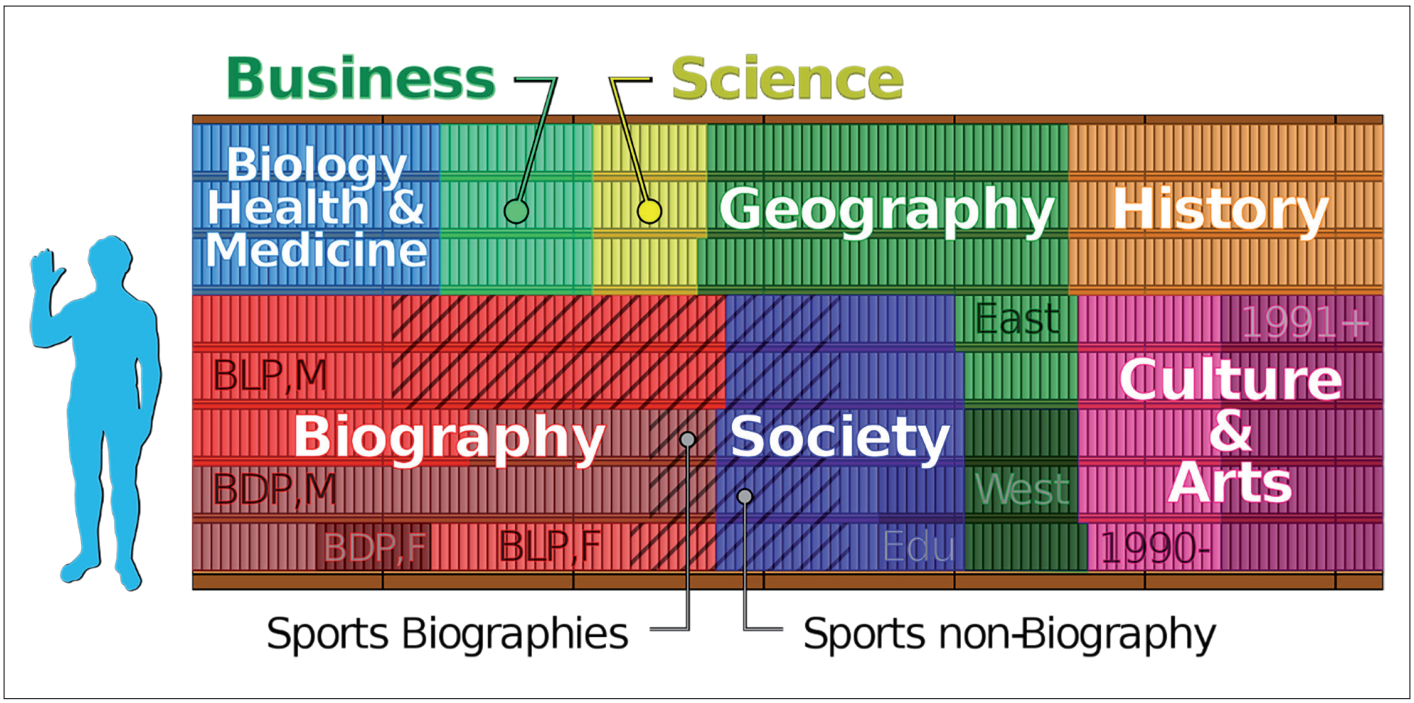


Figure 3: Wikipedia topic breakdown 2016 [7]

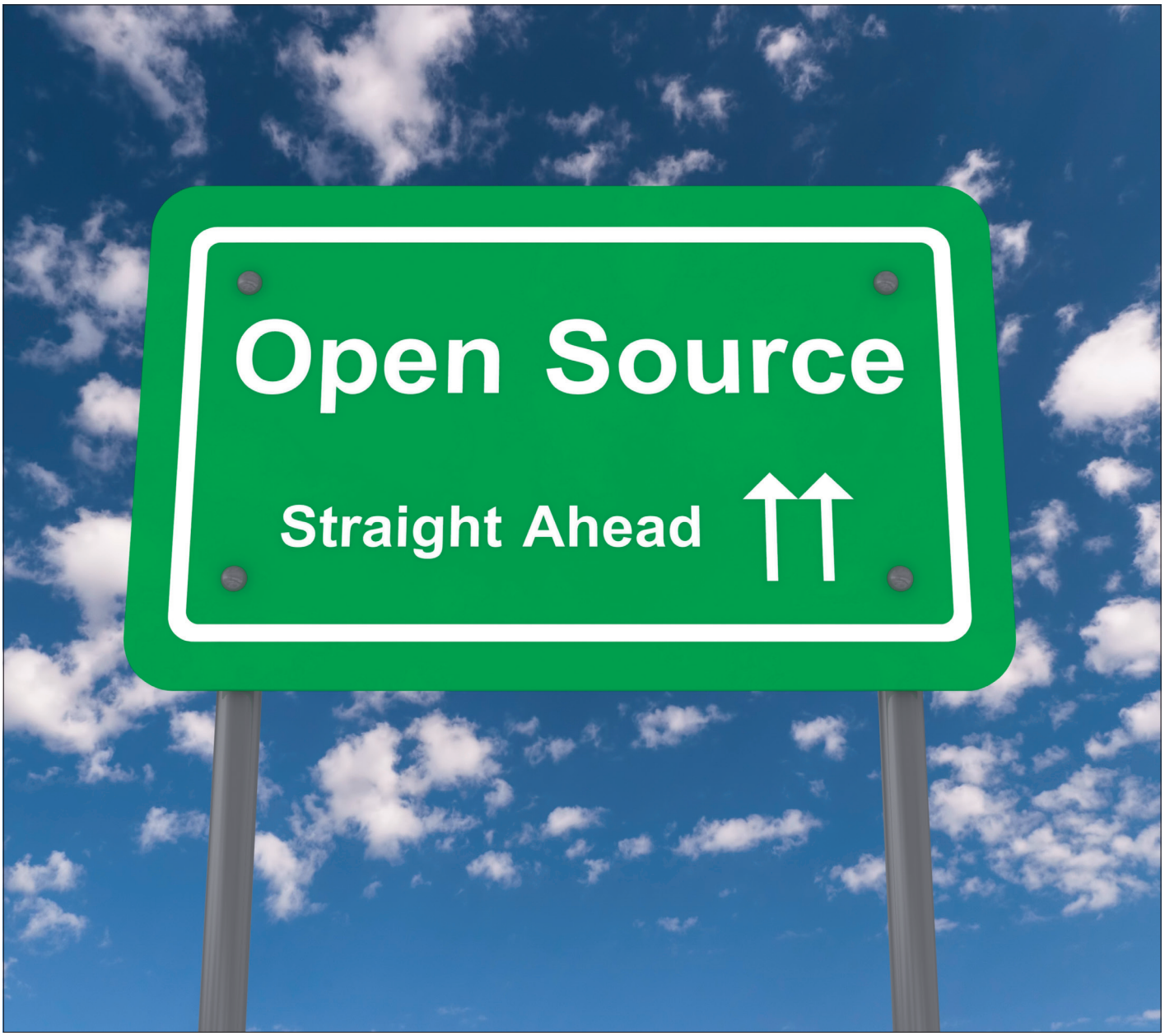


Image: ID 35101825 | © Adonis1969 | Dreamstime.com

More traditional open source programming language code creation and curation is a more specialized skill, the main line of work for many contributors, who need income to continue.

However, since the resulting bodies of encoded knowledge are free to use for everyone, the exclusivity typically required to exact payment is missing and, when programmers and their families go hungry, their contributions stop. If Europe wants open source assets to level the playing field against competition, it needs to find more ways to provide a living for those who contribute to this digital commodities infrastructure.

The earliest forms of government likely arose from trading: swapping wares, because both sides had a surplus of one good while they needed another. Setting up a market, creating space, roads, rules and their enforcement was an effort paid for in taxes.

While taxes were levied at every bridge during the middle ages, once realms and roads connected and news spread how roads and bridges should be built and markets operated, it became clear that there was little benefit to having a cut-throat competition between roads, bridges and markets right next to each other. That is to say, an aggregated or even centralized government-controlled infrastructure gave better commodity benefits and enabled the competition to move up and into new green fields of innovation.

Wherever one or very few competitors are left with little differentiation but scale, these tend to invest their profits more into solidifying their stranglehold than into improving their services. At that point, the majority service vendor effectively assumes a government function and that's where a government that wants to retain its sovereignty needs to step in.

The digital commodities required to simply participate in a digital world can be

considered the equivalent of what governments deliver in the physical world. While we pay for own shoes (smartphones) and cars (laptops), governments provide sidewalks and roads (OS, browser), regulation and enforcement, financed via taxes. Competition can still have a significant place there, just like road-building or other major infrastructure projects are segmented and publicly tendered.

Modern cities invest in much more than just basic commodities like roads, marketplaces and water supply. They feature public transportation, educational institutions, incubators, even public internet sometimes, because they believe that cut-throat competition for these commodities doesn't really help to achieve their real target: the best infrastructure to support a wide diversity and depth of new businesses.

Qualities of open source

Open source has evolved and diversified in recent decades: there are many variants and business models, but all of them rely on rules being followed, and social code and governance being properly executed as well. Such cooperation is natural to humans, but we switch into competition when threatened or when we judge the gains high enough. This leads to cyberwar that carries significant risks. But at least it tends to halt code evolution only at the point where collaboration was lost. All code and content distributed up to that point remains available to everyone with a local copy, because there is no centralized control. Proprietary software and exclusive services are much easier to choke off completely, with crippling and devastating effects.

Open source provides tangible incentives for international collaboration, and more efficient and stable algorithms benefit everyone who needs them. The more widespread the use of certain pieces of code, the more everyone depends on them being stable and efficient, and the more likely they are to motivate collaboration even across value divides. While open source doesn't automatically deserve trust, it is even more difficult to trust closed source. Where the two compete, open source is likely to win. Where code is really critical, either because it is used everywhere (e.g. browser or shared library) or where its use is critical for security, open source enables such a scrutiny and widespread willingness to improve it, that it out-evolves proprietary solutions. It should be noted that open source alone doesn't guarantee quality: practically all student homework today is hosted on open repositories from the very start of their learning curve. And in an environment where the rate of adoption by other developers for a new niche or use case can seem more important than its readiness for production, quality can suffer.

While we used to view code as something technical or mathematical at the dawn of computer science, code and data have been somewhat irreversibly mingled very early on and today quite simply are encoded knowledge and content which co-evolve with the human species. Being open about knowledge and sharing freely typically helps growth when there is little competition or resource constraints. When scarcity has competition heat up, closing down can help win a battle or a war, and accelerate the development of a few vs. a bulk left behind, at the risk of losing all. Neither strategy guarantees long-term success: we recommend switching intelligently and swiftly between the two approaches.

References

[1] "Evolution of the x86 context switch in Linux", https://www.maizure.org/projects/evolution_x86_context_switch_linux/

[2] IDC, "Smartphone Market Share", https://www.idc.com/promo/smartphone-market-share/os

[3] Statista, "Global market share held by leading desktop internet browsers from January 2015 to June 2020", https://www.statista.com/statistics/544400/market-share-of-internet-browsers-desktop/

[4] Wikipedia, "Google Workspace", https://en.wikipedia.org/wiki/Google_Workspace

[5] Ben Dickson, "The implications of Microsoft's exclusive GPT-3 license", 2020, https://bdtechtalks.com/2020/09/24/microsoft-openai-gpt-3-license/

[6] Jonathan Band and Jonathan Gerafi, "Wikipedia's economic value", http://infojustice.org/wp-content/uploads/2013/10/band-gerafi10032013.pdf

[7] "Wikipedia:Size of Wikipedia", https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

Thomas Hoberg is Technical Director R&D at Worldline, Germany.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: T. Hoberg. Open source code and content. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 184-191, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

In the last five years, open source hardware has moved from being a little-known niche activity to become an essential research vehicle and has even established itself in commercial business plans.

Open source hardware is here to stay

By FRANK K. GÜRKAYNAK

When you only look at the files and descriptions needed, at first sight there is very little difference between developing an operating system (software) and an integrated circuit (hardware). Considering how successful and widespread open source software is, it might even be surprising to see that it took open source hardware a quarter of century to become relevant. However, in 2020, open source hardware has firmly established itself and continues to find a broader user base not only in research but in industry as well.

In this article, I will try to highlight the differences between open source hardware and software development and trace its development over time.

Key insights

- Open source hardware is the key ingredient to allow agile co-operation of partners both in academia and industry, which is essential for modern integrated circuit design.
- While fundamentally similar to open source software, hardware and integrated circuit design in particular involve more stakeholders that need to be aligned.
- There is an opportunity to accelerate the acceptance of open source hardware, which will lead to a competitive advantage for early adopters.

Key recommendations

- Establish a European institution to support open source hardware activities.
- Increased support for integrated circuit design activities in Europe is essential.
- More work is needed to clarify legal aspects and allow closer co-operation in Europe.

I am old enough to remember the beginning of the Open Source Software movement. I was able to experience it first-hand as the Free Software Foundation, the GNU Public License, Linux and many others evolved from humble beginnings and essentially changed how software is developed and used. Roughly a quarter of a century later, what I see now is very reminiscent of those times: we have started to change the way we design and use hardware, as open source principles that have become so common in software are being applied to hardware as well.

But before we go any deeper, what exactly do we mean by hardware and how do we differentiate it from software? After all, hardware is a very broad term; brewing beer, building 3D printed medical equipment, designing printed circuit boards as well as implementing integrated circuits could all be seen as hardware design. Since my specialization is in integrated circuit (IC) design, I will concentrate on open source hardware (OSH) for computing hardware. Personally, I like the following definition by Richard Stallman of Free Software Foundation (FSF) fame [1].

"Software is the operational part of a device that can be copied and changed in a computer; hardware is the operational part that can't be."

Still the distinction is not so easy, especially because recent developments would allow a designer to develop hardware using a subset of a conventional language like C, then use a high-level synthesis (HLS) compiler to translate it into a hardware description language and apply it to a field programmable gate array (FPGA), a process that looks deceptively similar to developing a software application using an embedded platform like Raspberry Pi. This *semblance* has led to many discussions and misunderstandings in recent years. Therefore, in this article I will explicitly concentrate on an even more *restrictive aspect* of OSH, where we make use of it to design ICs that can be used to build better computing systems.

Integrated circuit design differs significantly from software development

No matter how you look at it, getting an IC manufactured is quite different from developing software. First of all, an IC is a physical component; it has to be manufactured through a very complicated process that takes weeks in dedicated factories. These so called *fabs* are operated by technology providers, like TSMC, Intel, Samsung, GlobalFoundries and UMC to name just the major players, which have invested billions of dollars in infrastructure to be able to manufacture ICs which today can have tens or even hundreds of billions of components. I tell my students that making a modern 7nm chip is technologically more complex than sending man to the Moon. It may be an exaggeration, but it is not that far off the mark. IC design involves a very substantial one-time (or non-recurring) upfront cost just to get going, and modern large ICs can only justify this cost through large production volume. Of course, there are *cheaper* ICs that are not that complex, but the fact remains that IC design involves working together with a technology provider as well as substantial investment.

It should be no surprise that designing something so complex also involves a wide range of dedicated software, which is collectively known as electronic design automation (EDA) tools. Over the years, following the pace dictated by Moore's law, ICs grew exponentially in complexity, and the tools had to be developed to keep pace with and manage this growth. Today three major companies (Cadence, Synopsys and Mentor (Siemens)) dominate the EDA tool market. Any serious IC design relies on these commercial tools, which come with significant licensing costs.

The key to managing the complexity is modularity and a substantial part of modern IC design relies on pre-designed and validated sub-systems that are made available through third party providers. They can be as simple as standard cell libraries that contain simple Boolean logic gates, I/O drivers, memories, clocking, interconnection solutions, and even processor cores that can then be combined to make a system-on-chip (SoC).

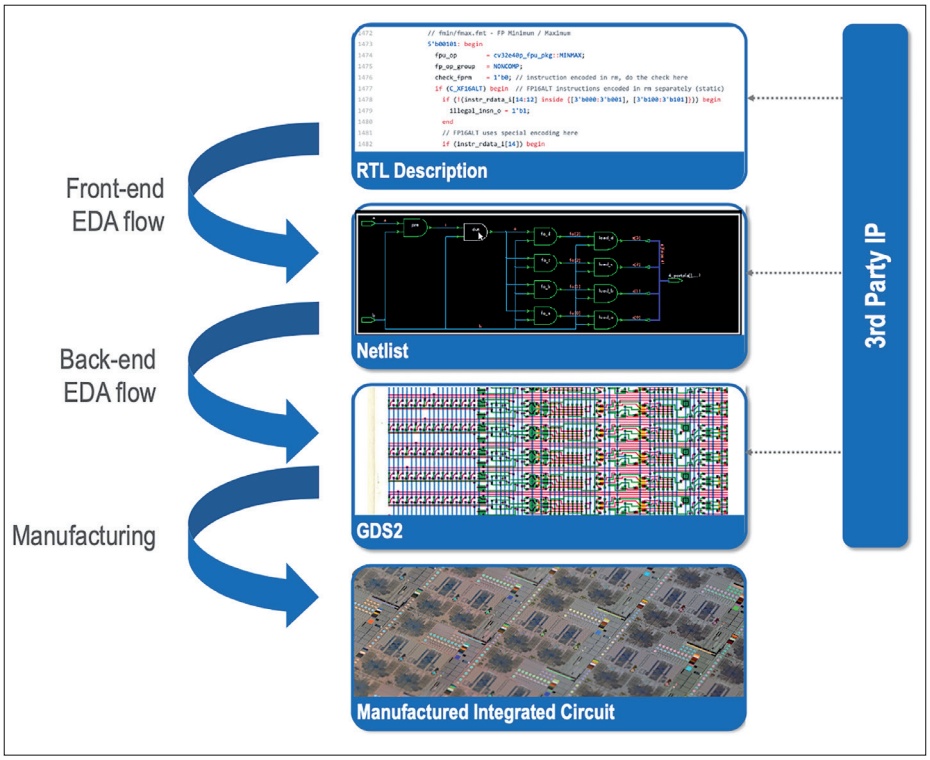


Figure 1: Steps of manufacturing a typical IC, showing interaction of third-party IP, technology and EDA tool providers

As a result, the challenges facing open source hardware are not only the fact that you cannot "copy and change it in a computer" as Stallman stated, but also that there are multiple entities with different commercial interests involved in the process. There are both obstacles and opportunities for open source at all these levels, but the process is taking longer as these different entities (technology, EDA, IP providers) have different goals and concerns. It is important to understand these relationships to develop more sustainable solutions for open source hardware.

There is a second challenge, which is directly related to the complexity of hardware design. There are simply much fewer IC designers than software developers, which reduces the pool of people that can provide open source solutions. This may have contributed to the fact that it took some time for OSH to establish itself but, in the last five years, OSH has started making some serious noise and it will continue to do so.

How does open source factor in?

Even with all the complexities involved, in the end, IC manufacturing relies on a number of computer-generated files that can be stored, transmitted and manipulated electronically. The ultimate description is probably the physical blueprint of the IC, the so-called physical layout captured in a format known as GDS2. This is essentially the only information that the technology provider needs to develop your circuit. Such a file can only be generated with the support of all three entities I have described above. The physical layout discloses information on the capabilities of the technology provider, and is only distributed under strict confidentiality agreements. A physical description that is designed properly needs the support of several EDA tools, and the vendors take issue when such descriptions are made available as it could mean fewer customers will need their tools. This is a particular issue in the case of universities and research institutions that pay only a fraction of the actual licensing costs to the EDA tool vendors. As mentioned, the physical description will most probably feature several different pieces of IP from third parties who may object to their IP being openly distributed as part of a larger

design. At the moment, making GDS2 files openly available is still an issue, although in 2020 we saw the first steps to address this problem with the efforts of Google and eFabless in connection with the 130nm Skywater technology [2].

One level below this abstraction is what we call a netlist, a circuit mapped into readily available components of a basic library such as AND, OR gates and Flip-Flops as well as some common blocks like ADCs, PLLs, memory macros etc. Such a netlist has physical properties, you know basically how large the circuit is, how fast it can operate and estimate its power consumption as the functionality is mapped to a technology-specific library. Similar to GDS2, releasing netlists mapped to libraries gives rise to similar problems: third party providers give access to their libraries under strict NDAs and EDA tool companies are not happy to see the output of their tools released.

At least for digital designs, there is an even higher level where we describe the functional behaviour of the circuit using dedicated languages such as SystemVerilog, VHDL, Bluespec or Chisel. In this form that we call RTL, the complete IC description, together with supporting information for verification, is available. However, it needs a front-end design flow to first produce the netlist and then the back-end design flow to generate the GDS2 which can be used to manufacture the integrated circuit. It is exactly these RTL descriptions that fuel current open source hardware success. At this abstraction level, there are still no EDA tools directly involved, and no technology-specific information is disclosed. If you do not look closely, a SystemVerilog description of a processor will look quite similar to the C++ code of a display driver. The difference of course is that good open source RTL descriptions are those that have some pedigree: they have been used as part of actual implementations and working integrated circuits. As an example, there are already forty highly-successful integrated circuits that have been manufactured based on the Parallel Ultra-Low-Power (PULP) platform [3], an extensive collection of optimized implementations of energy-efficient RISC-V based computing systems in SystemVerilog by ETH Zürich and the University of Bologna.

Eventually the success of open source RTL descriptions will also pave the way for open source releases in lower levels of abstraction, as there is nothing that fundamentally limits the distribution of open source GDS2 files once companies embrace open source principles.

How did it all start?

Open source hardware is enjoying a fair amount of the spotlight at the moment, but there were many products with OSH components long before people started to take notice. In the beginning, similar to open source software, most of the contributions came from volunteers and enthusiasts, people that were both passionate and had time on their hands. One of the most well-known early repositories was accessible under Opencores.org (the current www page is not maintained by the same group that originated it). While there were many smaller and simpler projects, as early as 2000, one of their key projects was OpenRISC [4] an open source processor which found serious use in many applications. In fact, the early versions of our PULP platform [3] used customized versions of the OpenRISC. The user group around OpenRISC later ended up founding the Free and Open Source Silicon (FOSSi) Foundation [5] and organized a small meeting called ORConf in 2012. The first three editions attracted only a small group, but I can safely say that OrConf 2015 in Geneva was a key event in OSH history. If you take a look, you will identify most of the key people active in OSH today among the 100+ attendees. The key change was the involvement of major academic groups in these meetings. In addition to our group in ETH Zürich and the University of Bologna, the University of Cambridge, IIT Madras, TU-Munich and UC Berkeley were present at OrConf 2015, held at the premises of CERN. This clearly marked a change in the OSH world, as the initial OSH volunteers were now joined by well-known research centres and universities.

While members of universities were also contributing to early work on open source hardware, involvement at the institutional level allowed longer term projects, and more people to work on them. This also had a direct effect on the output: larger and better supported projects started to became available.

At around the same time RISC-V started to have a noticeable impact. Developed by UC Berkeley, on its own RISC-V is not directly OSH. But the instruction set architecture (ISA) provided a contract between the software and the underlying hardware that was fresh, clean and was made openly available. A well accepted ISA is impor-

Figure 2: Group photo of OrConf2015 in CERN, Geneva (from https://orconf.org/2015/). One of the key milestones of OSH.

tant to allow both the supporting software (compilers, libraries, operating systems) and hardware to be developed independently. It is important to note that RISC-V was not the first open ISA. Open SPARC and the aforementioned OpenRISC were available long before RISC-V but, while the other two still continue to exist, RISC-V has enjoyed far more success. A large part of this success lies in the work put in by the RISC-V Foundation [6], which nurtured the ISA and was able to attract many high-profile companies to support the effort.

If OrCONF 2015 in Geneva was the coming out party of larger universities joining the OSH movement, 2019 (and perhaps the RISC-V Workshop we organized in Zurich) marked the time when OSH received serious industrial backing. Not one, but two non-profit organizations, the Chips Alliance [7] (Google, Western Digital, SiFive) and OpenHW group [8] (NXP, Thales, Silicon Labs) as well as the OpenTitan [9] initiative by Google and LowRISC were announced in 2019. All three of these efforts committed significant resources to develop, curate and improve open source hardware. Today, practically all major companies have significant involvement in these organizations and many see real benefits from using OSH.

How does the public, academia and industry benefit from open source hardware?

In all reality, if OSH did not have real benefits, it would not survive on its principles alone. In my opinion, the main reason why it took so long for OSH to have an impact has more to do with Moore's law and the associated growth in complexity in integrated circuits. Twenty-five years ago, the development of an efficient 32-bit microprocessor was the pinnacle of integrated circuit design. The Intel Pentium MMX from this era had the complexity of about 1 million gates and could be clocked at 233 MHz. Today Masters students at ETH Zürich where I work, regularly design integrated circuits that contain several processors. What was once considered special has now become a commodity, a building-block to make larger and more capable systems. Having such basic building blocks freely available as open source has been a very attractive proposition for many companies. In fact, Greenwaves, a startup from Grenoble, was able to base about 90% of their GAP8 (and follow-up GAP9) [10] IoT processors on OSH that the PULP project made available. This allowed the company to concentrate their efforts on differentiating their product by adding specialized accelerators and modifications. Especially for SMEs, the cost saving associated with procuring a proven processor infrastructure and peripherals can be substantial, and opens a faster path for innovation.

For those of us in the research field, the main enabler and driver has been the ability to co-operate with both industry and other academic partners freely. Modern integrated circuits have become so large and complex that creating innovations in this field is virtually impossible if you need to do everything on your own. The ability to share and co-operate with partners is essential if projects are to have impact. Without lengthy discussions on NDAs and access regulations, we can quickly get started on projects at a scale that is relevant and concentrate on innovations instead of spending inordinate amounts of time on what is essentially commodity infrastructure. Simply put, the ability to share our work with different partners has become part of our research infrastructure, which is greatly simplified under an open source model.

We have also come to realize that we get more co-operation opportunities and more dissemination as a direct result of our open source activities. There are also some additional benefits that get often overlooked. The ability to have fair and well-controlled benchmarks in integrated circuit design and the widespread use of OSH designs in training and teaching activities even by commercial EDA companies are just two examples.

In a time when people are increasingly worried about privacy, and popular attacks like Spectre and Meltdown have grabbed the headlines, having access to the complete inner workings of the hardware for public scrutiny presents yet another opportunity for OSH. Just having access to the RTL description is certainly not sufficient, but removes an important hurdle for an auditable system. In a time when digital systems control large aspects of our life, getting more secure and auditable solutions will be

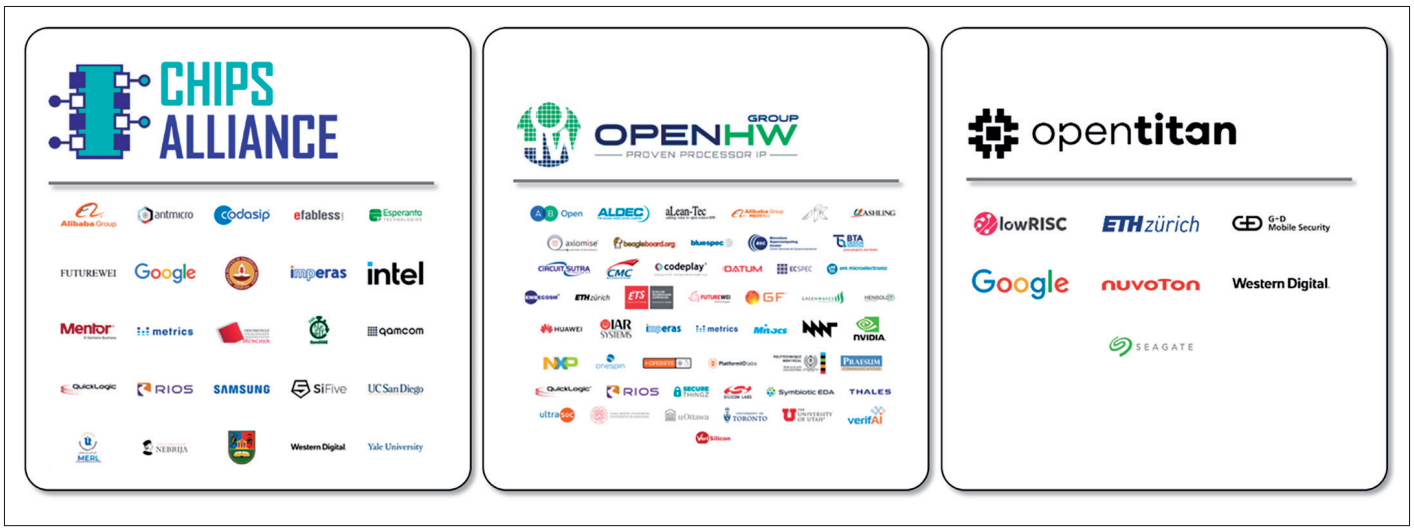
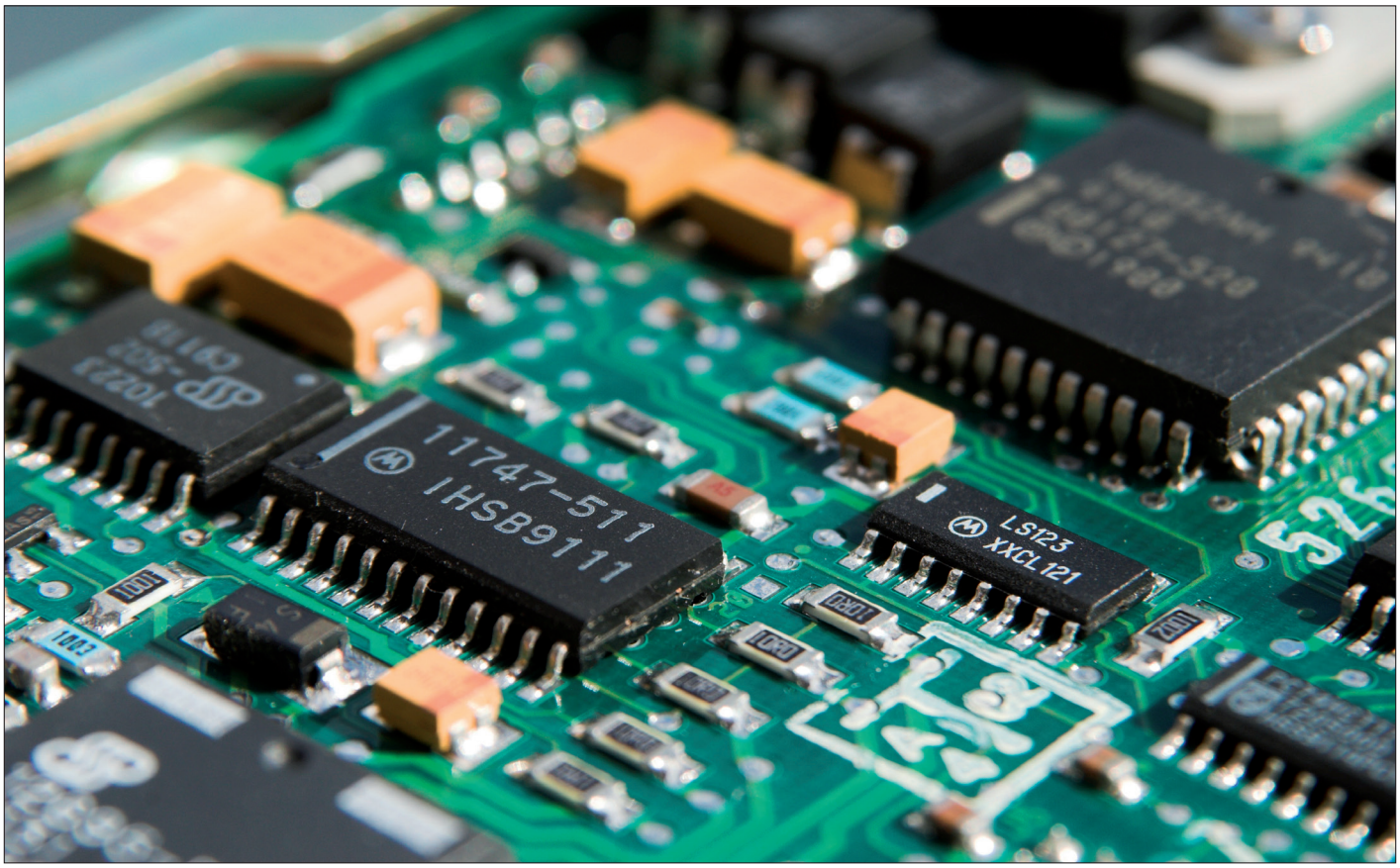


Figure 3: Major industry backed open source initiatives and their backers as of November 2020.



one of the most important contributions of open source hardware to society. There are already several well-funded projects, like the Posh Open Source Hardware [11] from DARPA, to address these concerns.

A few words on open source licences for hardware

As OSH becomes increasingly popular, the licensing aspects have also come under scrutiny. I had to learn it the hard way that working with open source licensing is a bit more involved than simply keeping all the rights for yourself. Our first OSH release was delayed by several months until we could understand and sort out all the issues.

The first point is to understand that there are fundamentally two separate families of open source licence. What we call permissive licences (Apache, MIT, BSD) basically allow your users to take what you have provided, use it, modify it and even sell it. They do not even have to tell you what they are doing with it. Most annoyingly, they can take what you have started and, when they make something better out of it, they do not have to share it with anyone else. Particularly at the beginning of the open source movement, this was seen as a major problem, and so called reciprocal licences were developed (GPL, LGPL). This second family of licence asks the user to make systems built using what they have received openly available under the very same licence.

Traditional open source contributors and volunteers prefer and advocate the reciprocal licences as much as possible. On the other hand, industrial users have to be careful to be able to protect the extent that open source components they use penetrate the overall ownership and rights of their products and therefore will generally only work with permissively licensed OSH. What makes everything more difficult is that solutions in software that set practical boundaries on how far the influence of the reciprocal licence will reach (like the Lesser GPL license, LGPL) do not translate well to OSH. Recent efforts by CERN on the Open Hardware License [12] have been an attempt to bring more clarity. The fact remains that, until they are challenged in court, we will not know for sure how well licences used for OSH will hold up. For example, is Apache good enough as a permissive OSH licensee or do you need the additional clarification that the Solderpad [13] licence brings on top of Apache? Will companies be more susceptible to patent lawsuits if they use OSH, or will the combined strength of the industrial interest groups like RISC-V International, OpenHW Group and Chips Alliance that support these OSH be sufficient to deter such suits? There is still a lot to learn in the coming years.

What can Europe do to lead in this area?

There is no denying that Europe has a keen interest in OSH activities and, as a result, it has become one of the most important players in the OSH movement, the recent move of RISC-V international to Switzerland being just another example of this trend.

In simple terms, OSH allows more people to work and innovate on IC design, and it is important to support these efforts and encourage the re-use of common building blocks to develop designs of much higher complexity and significance. A key

issue is supporting OSH at lower abstraction levels, in addition to RTL descriptions, as well as paving the way to distribute ready-to-manufacture GDS2 files as well. This is especially important for analog components that need to be designed specifically for a given technology as well as providing OSH components that have already been manufactured and proven to work as advertised. As described earlier, this aspect is still facing some challenges from stakeholders (technology providers, EDA tool companies, third party IP providers) that are comfortable in their established practices and realize that additional effort will be needed on their part for a change. Europe is home to several companies that are involved as stakeholders, and active encouragement to support OSH activities will accelerate these changes.

An important arena in which OSH is expected to play a key role is the realization that technologies that everyday life increasingly relies on (computers, data centres, communication infrastructure) are being developed, manufactured and also controlled by a very limited number of companies (and countries). Recent efforts of the European Processor Initiative [14], which has significant contributions from OSH, is part of the push for digital sovereignty for Europe. It is clear that individually, the member states and their research centres, universities and companies will have a hard time competing with established powerhouses in IC design unless they are able to pool their resources effectively and work in close co-operation. OSH can be an effective tool to facilitate just such a co-operation, but more work is needed to establish it within Europe.

The Europractice service [15] has been the key enabler facilitating access to both EDA tools and IC manufacturing services for SMEs and academia for more than two decades. An obvious step would be to bolster and extend these services in such a way as to allow members to be more active in OSH. Such a European institution (Europractice-OSH if you will) could take a leading role in opening discussions with stakeholders and creating an environment that not only provides an infrastructure for sharing OSH but also helps to establish legal framework and clarify licensing discussions around OSH usage for member states.

When it comes to designing high-performance ICs, especially for computing hardware, it is very important to realize that these are very costly projects, due not only to personnel costs but also to those associated with manufacturing. These include EDA tools and the necessary third-party IP, even when significant elements of it are being realized using OSH. If Europe wants to take a role of leadership in OSH, it also needs to support activities for supporting the manufacturing of designed ICs. Most of the current funding schemes are not compatible with the costs of modern IC manufacturing. The aforementioned institution could also serve in this capacity, as an interface to negotiate third-party IP for use in European-sponsored projects, educate European decision makers on the costs and feasibility of such projects and provide the necessary technical and legal framework to allow project partners access to and to share jointly developed projects. Note that Europractice already provides excellent service to fabrication (through Europractice-IC) and EDA tools (through Europractice Software Service), which represent two of the three stakeholders identified; expanding this service to allow design enablement through OSH principles seems like a logical next step.

It is important to note that commercial entities (big or small) will equally benefit from a more dynamic environment enabled by a wider influx of OSH in IC design. The entry barriers to SMEs designing their own ICs will be reduced, resulting in more designs that will be manufactured, requiring additional EDA licences, and increased need for both open source and commercial third-party components as well as services and businesses around these opportunities. While there is a good chance that these changes will happen organically over time in line with market demands, there are opportunities to accelerate this process within Europe, to allow the Union to move further ahead through government support to improve the acceptance of open source hardware among all stakeholders.

Conclusion

Within the last five years OSH has already made a significant impact, which is only going to increase as more and more stakeholders realize that the opportunities that it presents outweigh the concerns they have over quality and potential loss of revenue. This is not to suggest that all future ICs will be 100% open source but, as the open source software example has shown us, for components that everyone needs (think of GCC, Linux), taking advantage of the collective experience and effort of an open source approach allows everyone to benefit from solid building blocks and concentrate their energy into further innovation.

References

[1] Richard M. Stallman, "Free Hardware and Free Hardware Designs", https://www.gnu.org/philosophy/free-hardware-designs.en.html

[2] Github, "SkyWater Open Source PDK", https://github.com/google/skywater-pdk

[3] PULP Platform, https://pulp-platform.org

[4] OpenRISC, https://openrisc.io/

[5] The Free and Open Source Silicon Foundation, https://fossi-foundation.org

[6] RISC-V, https://riscv.org

[7] Chips Alliance, https://chips-alliance.org

[8] OpenHW Group, https://openhwgroup.org

[9] OpenTitan, https://opentitan.org/

[10] Greenwaves Technologies, https://greenwaves-technologies.com/gap8_gap9/

[11] DARPA, "Posh Open Source Hardware", https://www.darpa.mil/program/posh-open-source-hardware

[12] CERN Open Hardware Licence, https://ohwr.org/cernohl

[13] Solderpad Hardware License, http://solderpad.org/

[14] European Processor Initiative, https://www.european-processor-initiative.eu/

[15] EuroPractice, http://www.europractice.stfc.ac.uk/europractice_com/

Frank K. Gürkaynak is a Senior Scientist in the Digital Circuits of Systems Group of ETH Zürich, in Switzerland. He has been part of the open source PULP platform effort since its beginning in 2013.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: F. Gürkaynak. Open source hardware is here to stay. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 192-197, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

While digital transformation is widespread in industry, healthcare has only started to follow a similar path in the last few years, with COVID-19 leading to rapid acceleration. With the rise of the Internet of Medical Things, integrated care networks and connected healthcare, new opportunities emerge.

Is healthcare ready for a digital future?

By GUYLIAN STEVENS, KOEN DE BOSSCHERE and PASCAL VERDONCK

The rise of integrated healthcare networks, connected healthcare and the Internet of Medical Things (IoMT) has led to a huge and rapid expansion in the volume of data generated. The current fragmented structure of the healthcare landscape is not fit to manage, or make the most of, this vast amount of information. In a world where data has become currency and sensors are continuously generating new data, we are no longer able to process this continuous inflow. Therefore, a computational approach to analyzing and visualizing this data is needed to prevent healthcare systems and providers from drowning in an overwhelming lake of 'useless' data.

Novel microchip developments are finding their way into healthcare. Whereas until now, mostly wearables have been in use, new developments are now opening up the path to in vivo sensing devices, "insideables". This will lead to personalized healthcare and a resulting explosion of the volume of data. Is healthcare ready for this?

Key insights

- Healthcare is evolving from a centralized institution-based structure into a decentralized network-based structure founded upon increasing value for both patients and healthcare workers. This leads to the emergence of **value-based**, **integrated** and **connected** care.
- The Internet of Things, integrated care networks and connected care have led to the creation of an overwhelming volume of data, causing an overload for both clinicians and decision makers. **Collection, transmission, protection** and **processing** are the four key components of this data network. AI-based algorithms will be necessary to process this data into a form useful for clinical, financial and ethical decisions. A clear vision on how to store and analyze this data – ranging from edge devices to data centres – needs to be developed.

Key recommendations

- Computational healthcare will include a wide variety of domains: ranging from artificial intelligence in **decision support systems** over **computer-aided diagnosis** and data collection, transmission and security of wearables and insideables to manufacturing applications in **3D printing** and **computational modelling**.
- Future healthcare will require advanced cyber-physical systems with high quality of service and low latency.
- Insideables and bioelectric devices will require ultra low-power computing. nteroperability that allows security and privacy to be preserved will be a key requirement of the move to the cloud.

From healthcare to health

In 2010, the European Commission defined the primary goal of healthcare policy as that of maximizing the health of the population within the limits of the available resources, and within an ethical framework built on "equity and solidarity principles". This goal encompasses three pillars of major importance: maximize health, with limited resources, in an ethical framework. Over the last decade the goal to maximize health has not changed, however the limitations and the framework surrounding this goal have changed significantly [1].

Various frameworks aiming to redefine healthcare have been developed. A clear shift from individual and sporadic one-on-one care to a network-based continuous care is emerging. An early example was initiated by the ministry of health of Singapore. It describes three "moves" to implement better care (Figure 1).

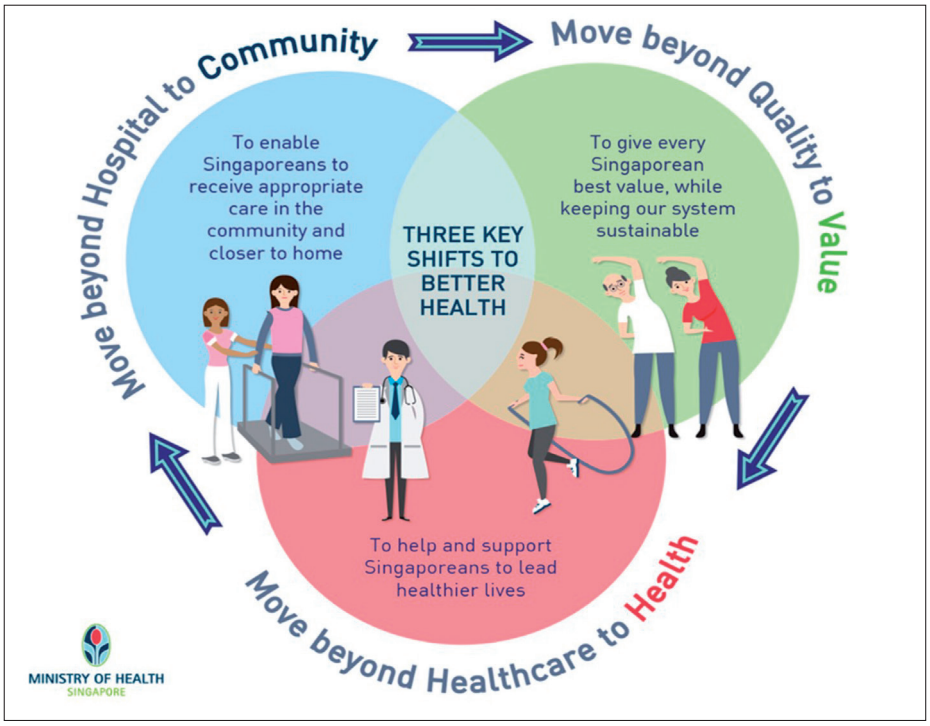


Figure 1: Framework healthcare Singapore [2]

The first move goes beyond the concept of hospital to the concept of community. This allows patients to receive good and appropriate care within their own community and thus closer to home. This does not mean that the hospitals become redundant, but that they organize care close to or in the comfort of the patients' home. This could include telemedicine, chronic care management, patient remote monitoring, patient self-management with or without coaching, etc. This is not only more comfortable and more affordable for the patient: it is also essential in areas where people do not have easy access to advanced medicine and hospital care. The second move is to go from quality to value, and to offer every patient the best possible value he or she can get. The last move is from healthcare to health, meaning caregivers would not simply act in the healthcare process but in fact intervene in the prevention and pre-care process to enable better and healthier living conditions for their previous, current and future patients [2].

This move from healthcare to health shows similarities with the health continuum plan proposed by Philips. This continuum describes how clinicians should not only focus on diagnosis and treatment but also get involved across the patient care spectrum, in healthier living, prevention and home care. This starts from the idea that citizens are in a continuous loop.

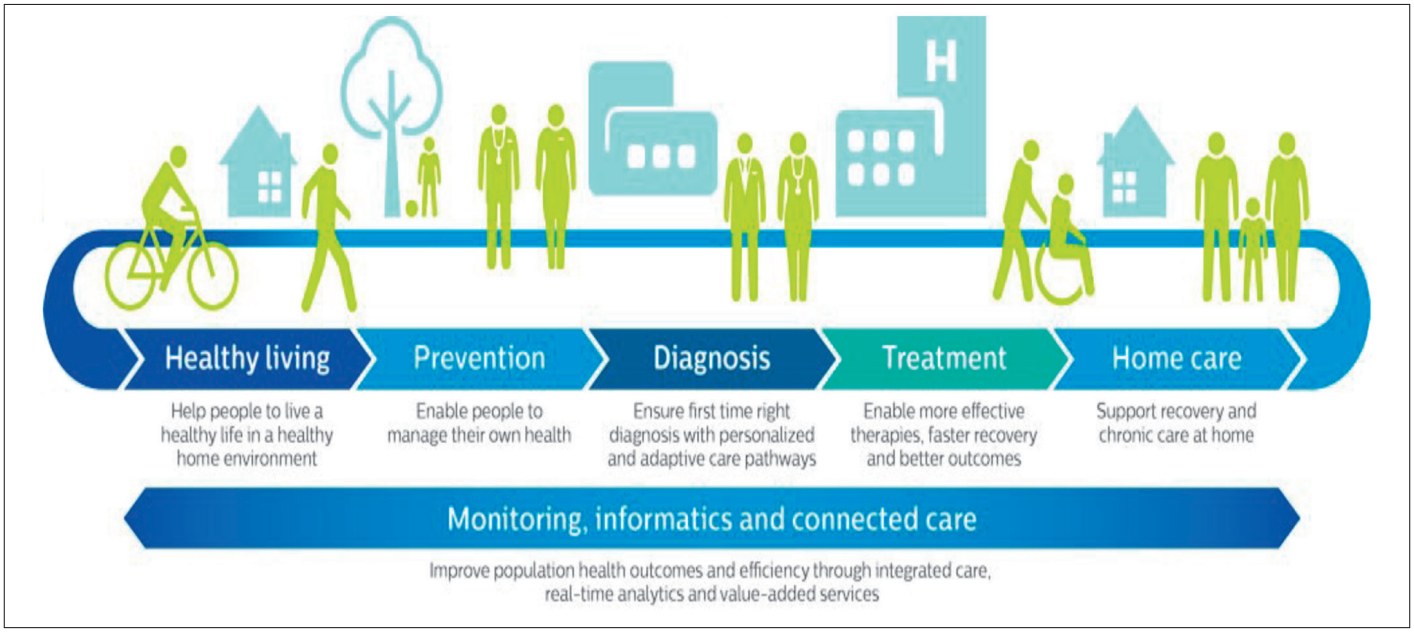


Figure 2: Healthcare continuum Phillips [2]

When citizens are treated and receive home care, they are treated as patients. Once they are fully recovered, they go back to being citizens. However, chances are high that if no change in their way of life is initiated, they will quickly return to being patients in the diagnosis phase. As citizens are either future or past patients, it is logical that healthcare practitioners are connected to both citizens and patients. As can be seen in Figure 2, this injects into the cycle several additional actors and information that can only be coordinated well if a foundation of monitoring, informatics and connected care-giving is established [3].

Towards value-based, integrated and connected care

Within these frameworks, three healthcare concepts play an important role: value-based healthcare, integrated healthcare and connected care. The idea of value-based healthcare is that the goal and purpose of healthcare in general is to improve the value for patients. Value is the product of the patient's personal experience with his or her outcome divided by the cost to achieve this outcome.

As M. E. Porter said, the only way to unite the interests of all participants in the healthcare system is by using value as a goal. Value is created by caring for a patient's medical condition over the full cycle of care. The improvement of the outcomes are the most powerful single lever in order to reduce costs and improve value [4].

The second concept is **integrated care**. Integrated care brings together inputs, delivery, management and organization of services related to diagnosis, treatment, care, rehabilitation and health promotion. Nowadays, a hospital or an individual healthcare professional can no longer operate in isolation. All of these individuals are integrated into one or more healthcare networks. Once a patient consults a healthcare worker of a certain network, he or she automatically becomes part of this network, of course with the liberty to leave or change as it pleases [5,6].

The last concept is **connected care**. We are increasingly confronted with smart environments. Wearables, smart cameras, ambient technology, connected equipment, etc. are embedded in a world where the Internet of (Medical) Things is growing exponentially and everything and everyone is becoming connected. This connection can be seen at different scales, going from on-body to in-home to community, in-clinic or in-hospital settings. When a patient is treated with a certain device or undergoes a certain examination, digital data will be generated, stored and used by other devices, people, institutions or environments in order to improve the value for the patient [7].

Towards a data framework turning data into value

Value, integration and connection are based on data. Without data that measures a patient's outcomes and costs, the value to optimize the healthcare process cannot be calculated. Without efficient data exchange, healthcare networks described above cannot operate and healthcare staff would once again work on their own isolated islands, leading to redundant examinations and consultations.

The healthcare of the future will generate massive volumes of data. Connected devices, healthcare workers and patients

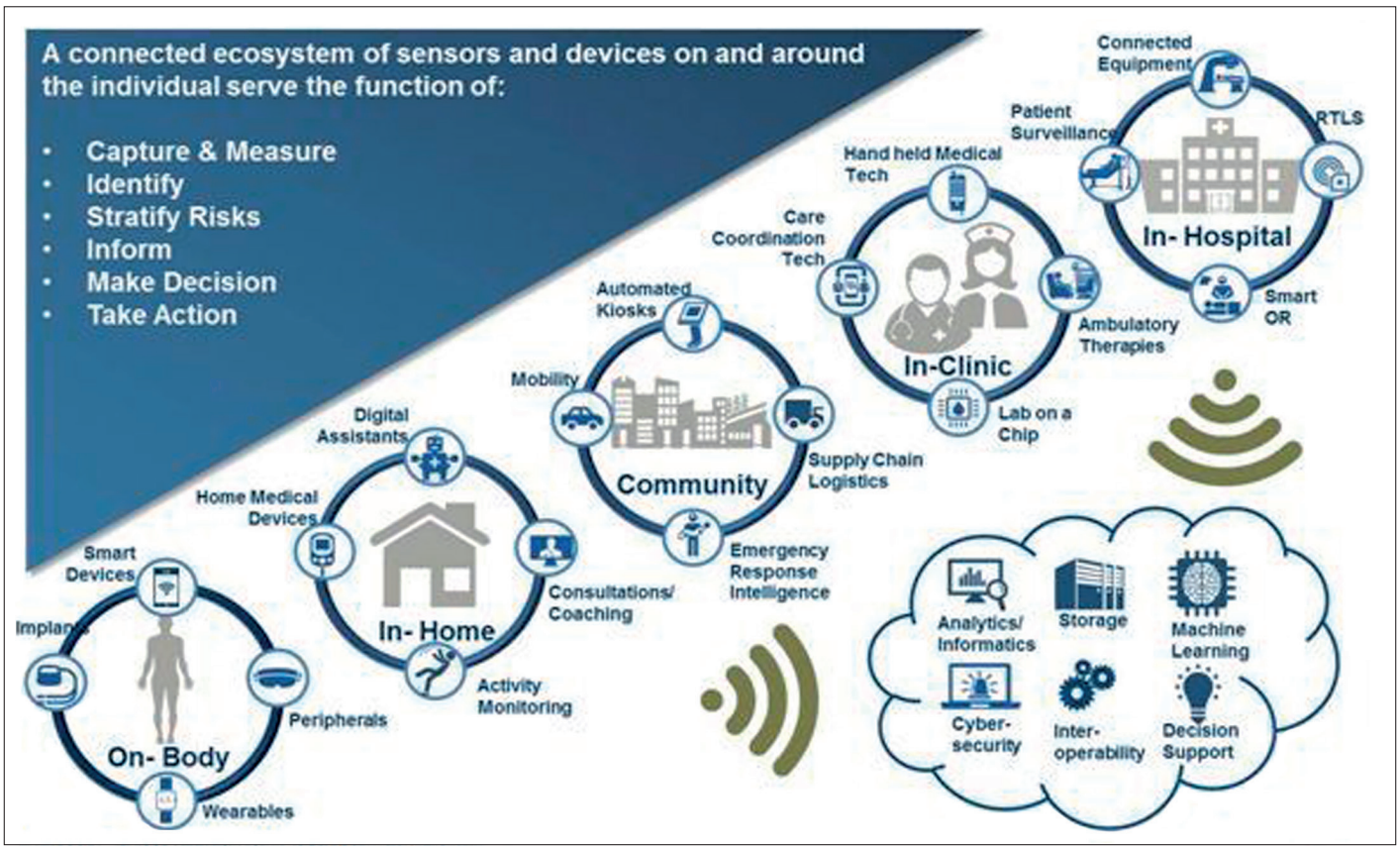


Figure 2: Connected care in 2025 by Frost & Sullivan

will create a lake of data. Sensors will perform the data collection from health, activity, location, emotions, parameters, etc. They will be the well of the data lake. One should not only think of wearables or insideables that monitor parameters of the patient but also the external ambient sensors that measure humidity, temperature, weather, pollution, etc. Measuring the living environment of a patient is something that is already done with a technique called ambient intelligence. Ambient intelligence creates a digital environment that is aware of the individual's presence and context, and is sensitive, adaptive, and responsive to their needs, habits, gestures and emotions. Combining these measurements with the patient-specific data allows for a more patient-specific approach. This data, used as **currency** in the network, can take many forms, ranging from personal to population or environmental data and will be transferred between a variety of interested parties [10].

Different communication protocols can establish a connectivity (**data transmission**) between all those different parties in this integrated and connected care network: Wi-Fi, Bluetooth, cellular (3G, 4G, 5G), LoRa, etc. Of course, merely collecting and transporting data will not be enough. This continuous flow of data collection will outgrow the storage capabilities of modern-day healthcare infrastructures, and the IoMT and decentralized healthcare will only increase further the demand for storage space.

Physical examinations will be partly substituted by a flow of digital information. Such a huge volume of data could easily overwhelm healthcare workers, leading to a situation in which value and patient outcomes no longer improve. The amount of non-processed data might lead to a decrease in efficiency and poor use of time, and thus also to a worsening of patient outcomes.

Therefore, next to collection, currency and transmission, **intelligence** needs to be introduced into the data framework. This intelligence, when done by computers is called artificial intelligence (AI) and can be a great asset to healthcare teams. AI is able to analyze data in order to provide insights that can inform the actions of healthcare staff, decision makers or the patients themselves. Signals that would otherwise go unnoticed or reacted upon too late can now be detected by the AI to trigger alarms directly to the responsible party.

The use of this artificial intelligence in healthcare can be put into three categories: knowledge-based decision support systems, data-driven clinical decision support systems and computer-aided diagnosis.

This third category has been particularly rich in development of new applications within the last couple of years. For example, the University of Pittsburgh has developed an algorithm that can identify prostate cancer with near-perfect accuracy. During tests, a sensitivity of 98% and specificity of 97% was demonstrated [26]. Another example is found in the domain of deep learning and neural networks. A collaboration between MIT and the Massachusetts General Hospital trained neural networks to automatically administer anesthetics during a surgery, based on the monitoring of the patient's vital parameters [27]. The universities of Melbourne and Otago have trained a neural network to predict the risk of cardiovascular diseases by gathering information from the retina of the eye [28]. As one can see, without this computer-aided intelligence, the data would still be available but would still require processing by individuals. With the current volume and complexity of such data, this wouldn't be achievable in practice.

Last but not least, an important aspect of data is **security**. With the new GDPR regulations of 2018, data protection has gained a tremendous importance over the last two years. This, in combination with a huge volume of generated data, leads to the necessity of rethinking the concept of data protection as a whole. Figure 4 represents the five discussed aspects of data: collection, currency, transmission and intelligence all protected within a shell of security. Each of these elements is essential and the loss or weakness of one would undermine the potential and capabilities of the others.

Nowadays, data is typically spread over different entities: central servers of hospitals, decentralized personal computers of

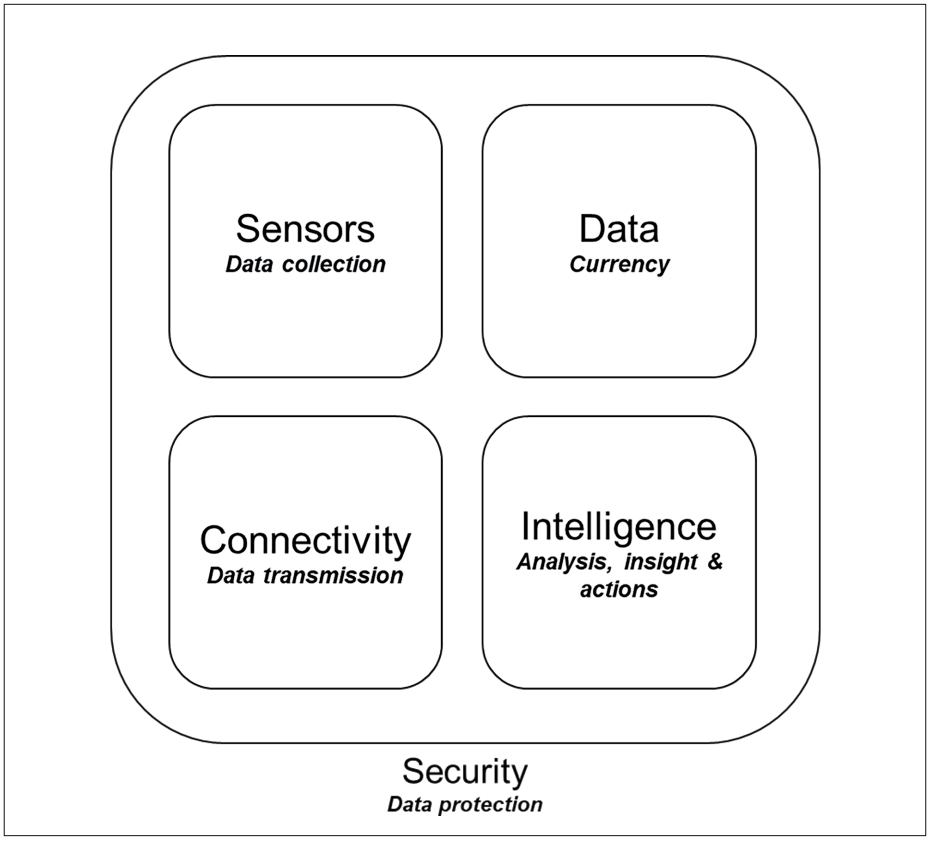


Figure 4: Aspects of the data framework

general practitioners, and centralized cloud servers of medical and non-medical apps. This is an obstacle to the introduction of connected and integrated healthcare. New concepts of connecting all these data sources in a decentralized network of data will be required. Technology like blockchain can control who has access to which data and could be the foundation on which to create these decentralized networks. Full Homomorphic Encryption allows the processing of information in encrypted form on untrusted servers. The possibility to use this data in a secured and effective way will lead to an increase of the future value of healthcare. As EY [8] reported in 2019, the future value will be equal to innovation to the power of data. Hence, without data, innovation will not create any value or vice versa.

From hospital servers to the cloud

As described above, healthcare is evolving into an integrated, connected care network. There are many different players involved, including healthcare practitioners, hospitals, governments and technology providers, all of whom play a role and have a stake in creating optimum value. The fundamentals of this evolution will be collection, transmission, sharing, analyzing and protection of data.

Data analysis will be the greatest challenge in the short term. The need to process this continuous dataflow will force healthcare systems to innovate and upgrade their IT infrastructure continuously, with more computational power, more storage and more network capacity. It will be necessary to transform massive amounts of data into information. To achieve this, a digital transformation is already taking place in healthcare. Computers are used to store and organize electronic medical records, for diagnosis, to administer medication, in operating rooms, for off-site care, for telemedicine, for machine learning, for clinical decision support systems, for clinical assisted diagnosis systems and so on.

Regarding the processing of the data, the way the data is stored leads to three processing techniques. When all data is stored locally in a database server, High Performance Computing (HPC) and quantum computers can be used to perform computationally intensive tasks and large analyses. At the moment, HPC systems are used in a good number of healthcare applications ranging from computational biology and chronic disease recognition to big data analyses and artificial intelligence. Quantum computing can open up new possibilities in healthcare in terms of drug design, in silico clinical trials with virtual humans simulated "live", real-time whole genome sequencing and analytics, the achievement of predictive health, or the security of medical data via quantum uncertainty.

Hospital servers are cautiously being replaced by cloud servers. Today, healthcare decision-makers remain skeptical about cloud computing due to its vulnerability to hacking, tampering and data leaks. However, with the rise of the IoMT, apps, wearables and ambient intelligence, cloud computing in healthcare will have to become mainstream, and technical and legal solutions will need to be found so that necessary levels of trust in the cloud can be achieved.

Figure 5 shows a conceptual scheme of the computing techniques of present and future.

Robots, AI and 3D-printing

Another major application of computing in healthcare is robots. Robots can be employed in several parts of the health continuum (prevention, treatment, diagnosis, aftercare). A primary application would of course be in the operating theatre (treatment). In surgery, many operations are already carried out by robots, operated by a specialized surgeon. Such robots are safety-critical cyber-physical systems with high reliability requirements. When used for telesurgery, the robot and the surgeon have to be connected via a fast and reliable low-latency network. Based on data from surgical robots, (parts of) some operations might in the future be carried out by autonomous robots.

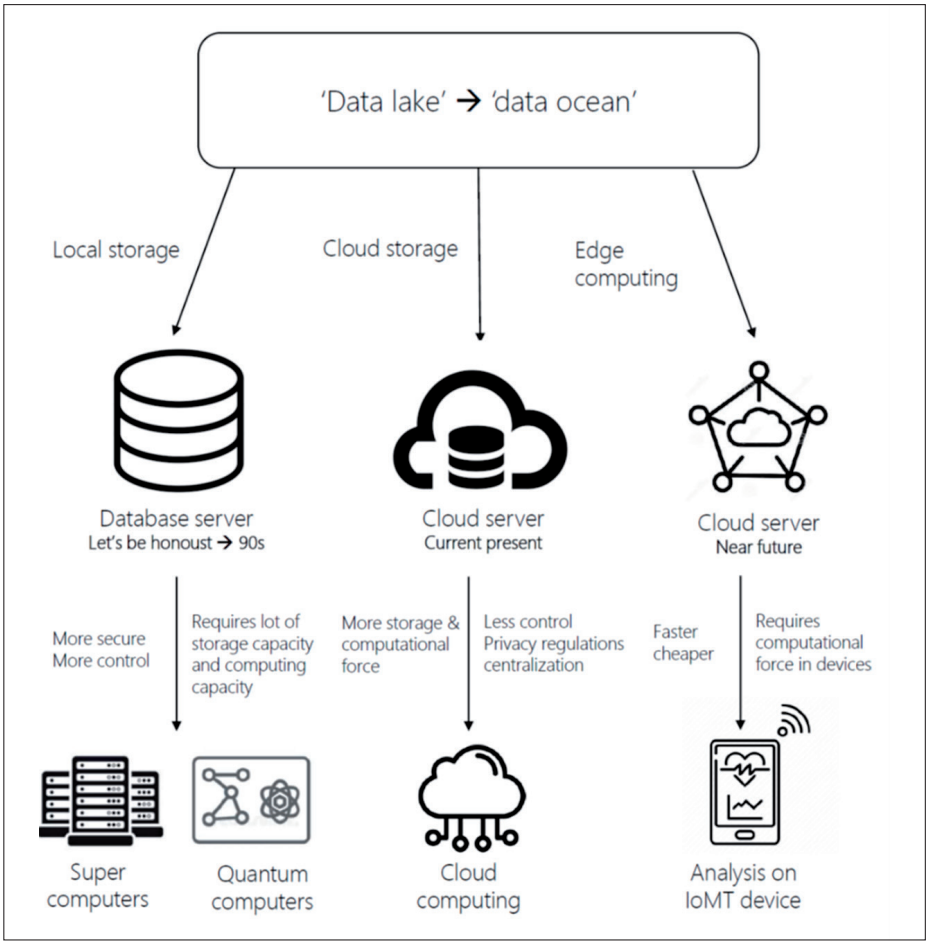


Figure 5: Concept scheme of computational data flows in the present and future healthcare institutions

However, the application of robots in healthcare goes beyond the operating theatre: think of their usability as a chat function, in which they can support patients by addressing misunderstandings and concerns about a procedure, delivering information in a responsive, conversational way over email or text. The general idea is that in the future, **these talking or texting algorithms might become the first contact point for primary care**. Patients will not automatically get in touch with physicians or nurses or any medical professional with every single health-related question; instead, they will turn to chatbots first. This chatbot function can be performed by an online platform or executed more physically in the form of a real humanoid robot, giving company to children or elderly people who would benefit most from this (prevention and aftercare) [20].

Other applications in which computers and robotics have proven to be great assets are logistics, assistive technologies (daily activity tasks, mobility, smart prostheses, etc.), revalidation (physical as neurological) and training or education.

Medical diagnosis is another field in which computers can play an important role. Take the IBM Watson technology: it was able to pass first year medical school exams but failed in the more practical tasks of the following years. Furthermore, Watson for Oncology did not deliver on its promises because the data it had to work with contained too much ambiguity. Watson for Genomics seems to work at least as well as humans when searching for mutations and proposing relevant drugs and clinical trials. In some cases, it has spotted important mutations overlooked by a human doctor. But there is no proof yet that it led to a better outcome. Watson is however very much appreciated as a smart librarian suggest the relevant studies [31].

In radiology, clinical decision support systems can support radiologists with routine tasks, improving their productivity and increasing the quality of their work, leading to better patient outcomes. The software will not take over the radiologists' jobs but will support them in such a way that they can focus more on the specific cases (diagnosis) [22].

A last important innovative domain within computer technology in healthcare is situated in the field of 3D printing. The combination of this additive manufacturing technique with software that can transform medical images into computer models (e.g. Mimics, 3-matic of Materialise) opens up a new range of possibilities within different hospital departments. Nowadays, 3D printing is used for computer-assisted and image-based patient-specific implants in orthopedics and maxiofacial surgery as well as for didactic communication between radiologists and organ specialists. Therapeutic 3D printing of cells with the use of bio printers for use in vivo and for regenerative medicine are still a dream for the hospital of the future [29, 30].

Microchips and insideables

Development of new materials and new chip manufacturing techniques have led to very small microchips with promising applications on and inside the human body. One of the first widely adopted applications is the use of microfluidics and with it the creation of lab-on-a-chip (LoC) devices. LoC allows the integration of several laboratory operations such as PCR (polymerase chain reaction) and DNA sequencing into a single chip on a very small scale [11]. This technology has seen spectacular growth over the last few years with applications including diagnosis of infectious diseases, handheld diagnostic devices and detection of analytes. In a recent project, researchers have developed a smartphone attachment that can detect multiple infectious diseases in a few minutes from a single drop of blood. Detection zones in a tiny cartridge present in the phone detect antibodies in the blood that enters the cartridge, thus detecting a disease. This device was field-tested by researchers in community clinics in Rwanda and was used to screen 96 patients for HIV and syphilis. The results had an accuracy level of 96% in detecting various infections when compared to those of standard lab tests [12, 13].

Another application of microchips is a human computer tag, called the VeriChip. It can be compared to a kind of wireless barcode or dog tag for humans. The VeriChip is a radio-frequency identification (RFID) tag produced commercially for implantation in human beings. Its proposed uses include identification of medical patients, physical access control, contactless retail payment, and even the tracing of kidnapping victims. In healthcare settings, the VeriChip can help identify a "Jane Doe" or "John Doe," that is, an incapacitated or disoriented patient whose identity is difficult to establish [14].

Of course, the functionality of microchips as being small computers for insideables brings with it a new dimension for consideration, which spans safety, reliability, efficiency, etc. Requirements can be functional (e.g. sensors, computational unit, telemetry), environmental (e.g. biocompatible, biostable, no biofoulding, limited heating, etc.) specific for device-tissue interaction (e.g. miniaturization, biomimetic, flexible, stretchable) or risk assessment-based (e.g. redundancy in hardware and software). Regarding the safety issues, biocompatibility and biostability are the key focus. This requires new techniques for chip production that will allow safe contact between the chip and biological tissues depending on the application. Innovative manufacturing techniques in chemical, physical and atomic layer deposition need to be supported. They need to be smaller, integrate more sensors and process the measurements in a meaningful stream of data – in other words, they need to become smart. This requirement will of course go hand in hand with higher energy consumption. To counter this problem, innovation in batteries (e.g., lithium-iodine, lithium-manganese dioxide), remote powering (e.g., telemetry, inductive coupling, ultrasound powering) and internal energy harvesting from sources in the body (e.g., temperature differences, bio-fuel cells, piezo-electric conversions) are the main focus. In Belgium, imec is currently one of the more advanced developers in medical microchip technology [25].

Imec and Ghent University have collaborated on the production of smart contact lenses developing a lens that can mimic the human iris to combat eye deficiencies (Figure 6). The iris aperture is tunable

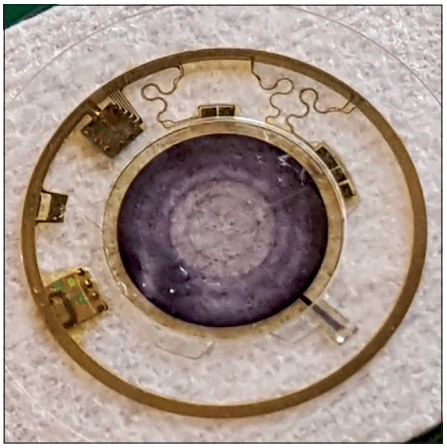


Figure 6: Smart contact lens by imec and Ghent University

through concentric rings on an integrated liquid crystal display (LCD). The smart contact lens is designed to operate for an entire day thanks to an ultra-low power design, offering a practical solution for people who suffer from deficiencies of the eye such as aniridia, high order aberrations like keratoconus, and light sensitivity or photophobia. The artificial iris lens is capable of dynamically changing the pupil size, bringing back two levels of functionality of the eye: light adaptation and expanded depth-of-focus [15].

Last but not least, neural probes seem to be the major future application for microchips in healthcare. This so-called bioelectronics will open up a completely new dimension of computing: the human-computer interface. Whereas, at the moment, people decide on how a computer device should react, neural probes will enable computers, and thus also robotics, to react directly to the neurological behaviour of the human. This can open up possibilities in terms of robotic prosthetics, allowing paralyzed people to speak again or control their automated wheelchairs just with their thoughts. Besides assistance, neural probes might also contribute to brain mapping, nerve stimulation, restoration of neuronal functions, and investigation of brain disorders [16, 17, 18]. In this domain, imec has created a neural recording and stimulation probe, implanted into the peripheral nerves. This research was carried out in collaboration with DARPA to research the opportunity of controllable prostheses for army amputees. The technology works with an in vivo CMOS chip and computational electronics that regulate the external control unit of the bionic prosthesis [15].

Market

The medical technology market in Europe was worth €120 billion in 2018, which makes it the second biggest global market (27%) after the United States (47%). Germany, France and the UK own more than 50% of the European market. The biggest export market is the US (41%) followed by China (10%). The total employment in the medical technology industry in Europe is almost 700,000. Europe has 27 000 medical technology companies, 95% of which are SMEs. The medical technology industry is a research-intensive industry, which reinvests 6-10% of its annual sales in research. With a global market share of 27%, medical technology is an important economic sector in Europe, and a position that is worth defending. Given the rapid digitization of medical technology, it is also an important market for the European computing industry.

New competitors have recently arrived on the medical technology market. Big tech companies like Google and Apple are investing heavily in the healthcare market. They currently limit themselves to consumer-based non-clinical applications and wearables. It is clear that they are focussing on the medical market too. Apple has already launched the Apple Health record and equipped their Apple watch with a single ECG measuring sensor. With its health kit, the company is able to report trends taken from other health apps and wearables. Furthermore, they have created a platform on which third parties can create pre-approved medical apps that can simply connect to the Apple health ecosystem. This will allow healthcare providers and insurance companies to create custom apps to use in their practices [23].

Google is also investing in the healthcare market. A recent joint venture between Google and Johnson & Johnson has led to Verb Surgical Inc. Verily (an Alphabet company) brings its expertise in data analytics, visualization and machine learning to the Verb Platform. Johnson & Johnson and Ethicon bring their deep knowledge of surgery and expertise in surgical instrumentation to this partnership. As they say themselves: "In the future, our actions will connect surgeons to an end-to-end platform for surgery, including pre-operative planning, intra-operative decision making and post-operative care." [24] Besides this venture, Google has bought Fitbit in order to launch new platforms and wearables onto the medical market as they are aiming for FDA approval in due course. It has also created a platform to support healthcare application builders in the development and implementation of new healthcare apps.

Roadblocks: smart infrastructure and new competencies

The implementation of integrated and connected healthcare will require an infrastructure upgrade for fast and stable connectivity through the widespread implementation of 5G wireless networks. The speed and bandwidth of this network will allow the creation of a speedway for data transmission and analysis. At the moment, 5G is being rolled out in Europe. A stable network with very high levels of coverage will be essential, along with suitable solutions to the possible issues of data security and privacy, for the next steps in the future of healthcare to be taken.

Given that cloud storage decreases control over stored personal and medical data, it will be vital to check if providers are completely compliant with the GDPR. Beside safe storage and transfer of data, the interoperability between the devices generating this data and the electronic patient records will also be of major importance. Without this integration, healthcare professionals will hesitate to use this large source of information and data will be lost.

Another hurdle is the urgent need for new competencies in healthcare leadership in general, in particular in hospitals. There is a growing gap between healthcare management competencies and capabilities that are acquired, and those that are now needed in the changing technological landscape. New computer technology can be embraced in order to reduce expenditure and help healthcare provision to adapt to changing social values [19].

Conclusion

In healthcare, data has become the currency of healthcare professionals, patients, governments, hospitals and decision makers. In the bigger picture of integrated and connected care, value-based healthcare can only be achieved if this data can be collected, transmitted and analyzed. In order to do this the computational landscape is shifting from a local data storage model to a cloud storage model. However, due to the fear of tampering, hacking and loss of data, this transition is moving slowly when compared to other domains. The greatest challenge will be the transformation of data into actionable information. Artificial intelligence, combined with increasing decentralized computational power, will play a key role in this. With current and future developments of insideables and bioelectronics, the data stream will only increase. This requires support and encouragement for continuous innovation in computational healthcare.

References

[1] Council of the European Union, "Employment, Social Policy, Health and Consumer Affairs", https://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/lsa/118254.pdf

[2] Ministry of Health Singapore, "Updates on COVID-19", https://www.moh.gov.sg/

[3] Philips, "Our Strategic Focus", https://www.philips.com/a-w/about/company/our-strategy/our-strategic-focus.html

[4] ICHOM 2019 Conference, https://www.ichom.org/events/conference-2019/

[5] Oliver Gröne and Mila Garcia-Barbero, "Integrated Care", https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1525335/pdf/ijic2001-200121.pdf

[6] J. Ribera, "Hospital of the Future", https://media.iese.edu/research/pdfs/ST-0388-E.pdf (p24)

[7] Frost & Sullivan, "Healthcare world in 2025", https://store.frost.com/vision-2025-the-future-of-healthcare.html

[8] Ernst & Young, "Annual Report 2019", https://www.ey.com/nl_nl/jaarverslag/2019

[9] Philippe Duluc, Pierre-Antoinne Harraud, Ewan Munro and Joaquin Keller, "How quantum technology will revolutionize Healthcare", https://atos.net/en/lp/ascent-magazine/how-quantum-technology-will-revolutionize-healthcare

[10] Albert Haque, Arnold Milstein and Li Fei-Fei, "Illuminating the dark spaces of healthcare with ambient intelligence", https://www.nature.com/articles/s41586-020-2669-y

[11] Burak Yılmaz and Fazilet Yılmaz, "Lab-on-a-Chip Technology and Its Applications", https://www.sciencedirect.com/science/article/pii/B9780128046593000087

[12] Frederick Balagaddé, "Where the chips fall – lab-on-chip diagnostic", http://www.medicaldevice-developments.com/features/featurewhere-the-chips-fall-lab-on-chip-diagnostics-4212684

[13] Melissa J. MacPherson and Mayoorendra Ravichandiran, "Lab-on-a-chip technology: the future of point-of-care diagnostic ability", http://www.uwomj.com/wp-content/uploads/2011/08/Macpherson.pdf

[14] John Halamka, Ari Juels, Adam Stubblefield and Jonathan Westhues, "The Security Implications of VeriChip Cloning", https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1656959/

[15] Imec and Ghent University, "Imec and Ghent University Present a Smart Contact Lens Mimicking the Human Iris to Combat Eye Deficiencies" https://www.imec-int.com/en/press/imec-and-ghent-university-present-smart-contact-lens-mimicking-human-iris-combat-eye

[16] Imec, "Imec releases neuropixels neural probe to the global Neuroscience Community" https://www.imec-int.com/en/articles/imec-releases-neuropixels-neural-probe-to-the-global-neuroscience-community

[17] Jari Scheirlincks, Artificial intelligence: "Brain Computer Interface (BCI)", https://www.slideshare.net/JariScheirlinckx/artificial-intelligence-brain-computer-interface-bci

[18] Geon Kook, Sung Woo Lee, Hee Chul Lee, Il-Joo Cho and Hyunjoo Jenny Lee1, "Neural Probes for Chronic Applications", https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6190051/

[19] Future Healthcare Management in Europe, https://fhme.eithealth.eu/

[20] The Medical Futurist, "Health Chatbot", https://medicalfuturist.com/top-12-health-chatbots/

[21] Andrew Nusca, "After a year of medical school, IBM's Watson passes first milestone", https://www.zdnet.com/article/after-a-year-of-medical-school-ibms-watson-passes-first-milestone/

[22] Ory Six, "The ultimate guide to AI in radiology", https://www.quantib.com/the-ultimate-guide-to-ai-in-radiology#how-can-AI-help-the-radiologist-help-patients

[23] Apple, "Healthcare", https://www.apple.com/healthcare/

[24] Verb Surgical, https://www.verbsurgical.com/about/

[25] Centre for Microsystems technology (CMST) Imex associated lab at gent university, Prof. Maaike Op de Beeck, ARFA BIMI, Section 7,8

[26] University of Pittsburg, "Artificial Intelligence Identifies Prostate Cancer With Near-Perfect Accuracy", https://scitechdaily.com/artificial-intelligence-identifies-prostate-cancer-with-near-perfect-accuracy/

[27] Charlie Osborne, "Neural network trained to control anesthetic doses, keep patients under during surgery", https://www.zdnet.com/article/neural-network-trained-to-control-anaesthetic-doses-keep-patients-under-during-surgery/

[28] Aimee Chanthadavong, "Australian and New Zealand scientists use AI to predict heart disease risk", https://www.zdnet.com/article/australian-and-new-zealand-scientists-use-ai-to-predict-heart-disease-risk/

[29] Lawrence Livermore National Laboratory and Duke University, "Research team pairs 3D bioprinting and computer modeling to examine cancer spread in blood vessels", https://www.llnl.gov/news/research-team-pairs-3d-bioprinting-and-computer-modeling-examine-cancer-spread-blood-vessels

[30] Susan Decker, "Future of 3D Printing Is in U.S. and Europe Patenting", https://www.bloomberg.com/news/articles/2020-07-14/future-of-3d-printing-is-in-u-s-and-europe-patent-study-shows?sref=db2f3qgr

[31] Eliza Strickland, "How IBM Watson Overpromised and Underdelivered on AI Health Care", https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care

Guylian Stevens is involved in digital transformation at the Maria Middelares General Hospital in Ghent, Belgium.

Koen De Bosschere is Professor in the Electronics department of Ghent University, Ghent, Belgium.

Pascal Verdonck is Professor in the Electronics department of Ghent University, Ghent, Belgium.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: G. Stevens, K. De Bosschere, and P. Verdonck. Is healthcare ready for a digital future? In M. Duranton et al., editors, HiPEAC Vision 2021, pages 198-205, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

Dematerialization and the as-a-service (XaaS) business model are a strong business trend. They are key to reducing the carbon footprint of the economy.

Everything as a service

By KOEN DE BOSSCHERE and MARC DURANTON

Over the last thirty years, almost everything that can be represented by bits and bytes, has been digitized: music, movies, photos, books, news... the list goes on. This digitization has led to full dematerialization of production, transmission and consumption. Thanks to increased compute power, large and affordable digital storage capacity, and fast networks, the digital solution has not only replaced the physical one but is, in many cases, better. This trend has led to disruption in several economic sectors, which have had to reinvent their business models so as to transition from selling physical goods to selling a service. COVID-19 is leading to an accelerated digital transformation, which will lead to further disruption and more as-a-service business models.

Key insights

- Digitization has transformed the entire media industry. Consumers and distribution platforms became content creators too.
- The as-a-service business model changed users' world view. Ownership is being gradually replaced by "24/7 access" and renting. This leads to a rapid dematerialization of the economy, and might have impact on the long-term existence of creations.
- The impact of dematerialized services like streaming, videoconferencing and cloud gaming on the environment is moderate, which possibly makes them ecologically less damaging than travelling to a movie theatre, a meeting or a gaming event.

Key recommendations

- Keep investing in ultra-low power computing technology (data centres, networks, devices) so as to reduce the carbon footprint of digital services, and offset the environmental impact of their exponential growth.
- If the above recommendation is executed, keep investing in further dematerialization of services by improving existing solutions and creating new services.
- Create digital libraries and archives, in order to preserve digital-only creations.
- Ensure that European ethics are thoroughly taken into account by content providers.

Very few people in 1982 realized that the introduction of the compact disc (CD) was the start of a new trend (digitization of analog information) that was going to disrupt whole industries. For the customers of 1982, it was just a convenient and higher quality carrier of music.

In 1988, Fujifilm introduced the first fully digital camera, able to store up to ten photographs on a memory card. This represented another major analog-to-digital transition. The first camera phone was the Kyocera Visual Phone VP-210 in 1999. By 2010 all smartphones could record and play media, and resolution and storage capacity were no longer a serious constraint for most users. In the end, it took almost thirty years to evolve from the first digital music player to a powerful multimedia device in pocket format that people are ready to spend a couple of hundred Euros per year on, and that they carry with them at all times. Today, for millions of people it is the last device they see at night, and the first they see in the morning.

Digitization and the as-a-service business model leads to disruption

In the process of digitization, the business model of the content providers has also changed. Instead of selling physical content (like they did in the times of the CD), they started selling digital content that could be downloaded. When the networks became better, they moved to a subscription model for a streaming service where the user has full access to millions of songs for a flat monthly rate of less than €10, or less than the cost of one CD per month. Subscriptions to video streaming services are of the same ilk. Per streaming service, there are several plans, and access to premium content requires a higher monthly subscription rate. Some also have a free plan where content is regularly interrupted by adverts, which is called the freemium business model (from free to premium). This model has the advantage for the provider to lock-in the subscribers and to provide more stable revenues. As, most of the time, the digital contents have DRM which can be revoked at any time, the owners don't own the content, and their subscription is only a rental licence. For the first time in history, intellectual content can be erased with one

click (it happened in 2009, when Amazon remotely deleted some digital editions of the books of George Orwell – including "1984" – from the Kindle devices of readers who had bought them) [7].

Streaming services have put music shops and video rental services out of business in a very short time (Figure 1). Users now have all the content they can dream of available without having to invest in a collection, and music is available on every platform at an affordable price. The choice of what is available now depends on the choices of the provider; for example, some classic movies can't be found at all on platforms that keep only what is fashionable and immediately profitable. The distributors no longer have to invest in the production and distribution of physical media and – very importantly – have full access to the behaviour of their customers. In the last decade, some (especially in the video streaming sector) have started producing their own content and have become very successful. In 2019, Netflix spent $15 billion on content creation and will become the second largest entertainment producer of 2020 (Figure 2). It earned no less than 24 Oscar nominations in 2020. One of Netflix's key assets is that it has access to the behaviour of its viewers and can tailor its offering to their interests and preferences. It is a dominant market leader, making it more difficult for smaller players to survive. The "recommender systems" use artificial intelligence to best profile the users. The side-effect is that they can easily lock the users in their preferences, and might be, in malevolent hands, a way to manipulate people, based on their individual behaviour – customized manipulation; and victims may well be unaware that this is happening.

More generally, the digitization of media (audio, video, photos, newspapers, books, games, ...) has profoundly changed society. The impact of this technology may be compared to the introduction of the printing press in Europe by Johannes Gutenberg. Today everything from books to audio recordings and movies can be duplicated forever without loss of quality, in no time, and at an extremely low cost. Thanks to the internet and the cloud, available content can be accessed from nearly anywhere in the world – if not blocked by regional licences and DRMs. This was made possible thanks to the giant leaps made in performance in processing, storage and digital communications, fuelled by Moore's law. Modern streaming media companies are made possible by reliable broadband access and huge data centres distributed across the globe.

But there is more. The printing press enabled more people to print and distribute their ideas in printed format. Digital media enables everyone with a smart phone to produce and distribute audio, video, photos, text, games, … We are all now prosumers who produce and consume at the same time. YouTube is the market leader for video sharing, Instagram for

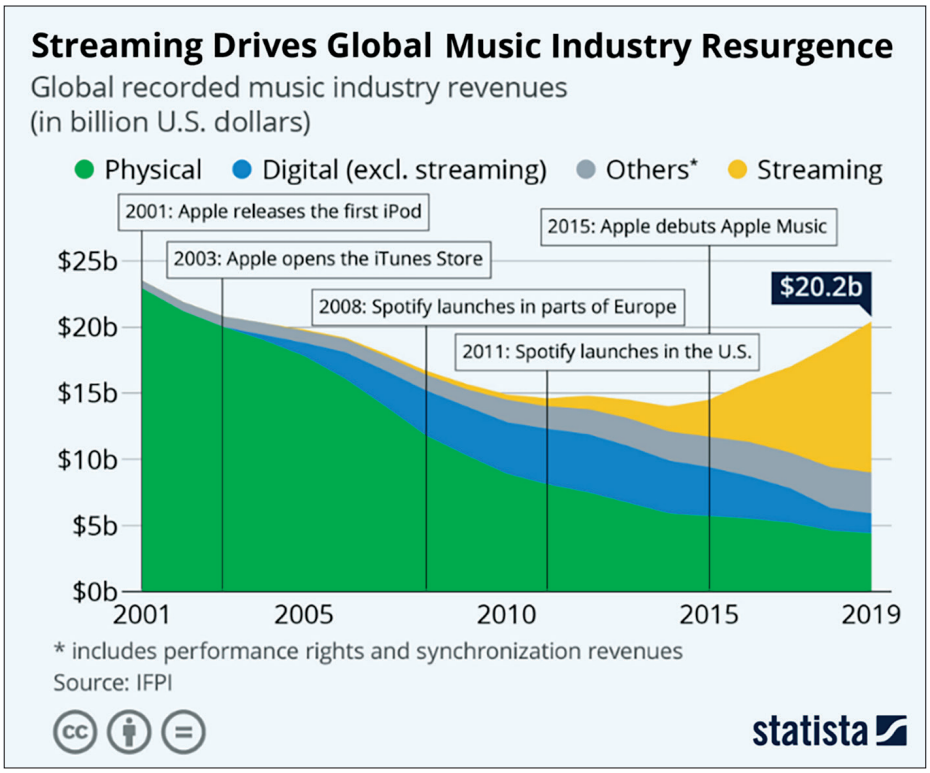


Figure 1: The fast growing music streaming industry brings music industry the revenue back to levels seen in the 2000s [1]

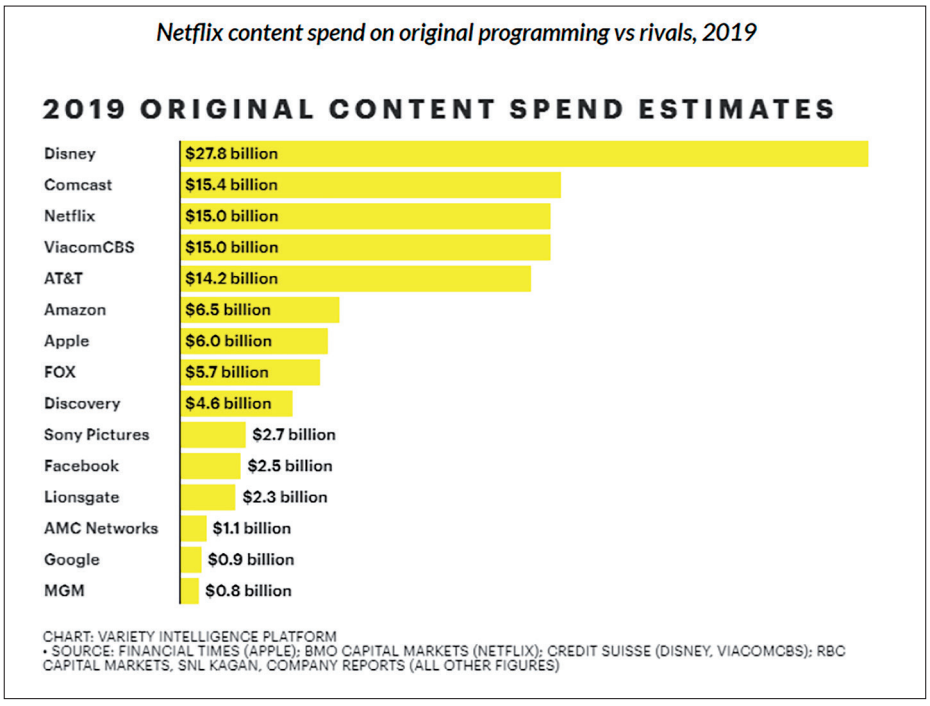


Figure 2: Content creation spending [2]

photo sharing. But this time there is no editorial board to control the quality or the veracity of the content. Facebook and YouTube use algorithms based on artificial intelligence to check the most obvious characteristics of the content and identify pornography or violence for example, but recent history has shown that this is not a perfect system and that it can also incur problems. These social media applications and their massive data sets are a key driver for the development of modern AI.

The consumer market has reacted to this evolution by offering a wide range of devices to improve the media experience (headsets, headphones, loudspeakers, large screens, …). The computing part is almost invisible: a smartphone, tablet or smartphone suffices to carry out the necessary processing.

Traditional printed media like newspapers and magazines have followed the trend. They cannot survive without a digital channel. In the beginning, the content was merely a digital copy of the printed version, but that no longer suffices. Today, media outlets start from the digital content that feed into a news website, offering video, audio, blogs, vlogs, digital puzzles, etc. They are processing information as fast as the television networks: in real time with breaking news and updates. The printed version follows the next day. The benefits for the consumer are clear: digital delivery is more convenient, faster, available 24/7 on the platform of choice, cheaper and takes up less space. Since it is dematerialized, it is available worldwide, and converts the use of physical resources (paper, ink, fuel for transportation) into the use of energy. However, it is much harder to come up with a profitable business model: many users do not want to pay for premium content they can find for free elsewhere on the internet, and the income from ads is also shrinking because many advertisers prefer the larger platforms like Google or Facebook. Printed newspapers are struggling to survive and are primarily bought by those over fifty. The younger generation gets 53% of its news from news websites and social media; their smartphone, rather than a television, radio or printed newspaper, is their window to the world (Figure 3). The availability of large amounts of content also entails that users spend less time on each one, switching from one to another, at the expense of paying close attention. This drives content to be short, superficial and very appealing, if not sensational, rather than to offer any in-depth analysis.

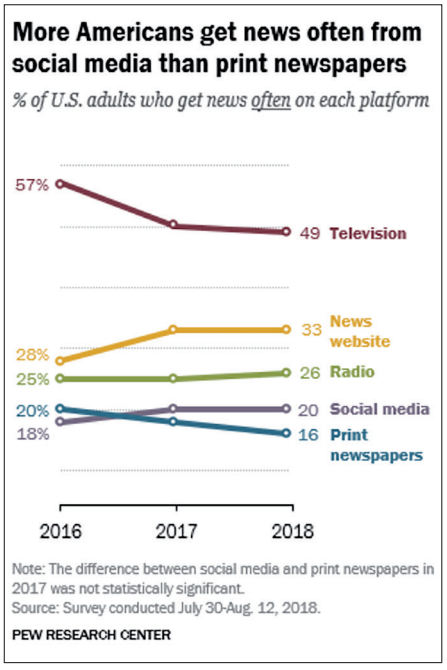


Figure 3: Platform usage by US adults [3]

The as-a-service economy changes people's world view

This technology is also having an impact on how people view the world. With things increasingly becoming non-physical, younger generations have developed a different view of possession. Having 24/7 access to information in the cloud is perceived as being as good as possessing it, even for emotional assets like family pictures and movies; many people no longer care to keep such treasures safe at home. In fact, there is no point in seeking permanent possession of a physical good that has a digital equivalent (a video, a picture, a book, etc.) or whose availability can be summoned instantly (a shared car, look-up in a dictionary or knowledge base, etc.). The physical good occupies physical space, which is a scarce resource for many, and implies an upfront investment to acquire it and, in some cases, cost of maintenance and care. Owned goods tend to become rapidly obsolete, rarely acquiring value in the process. The digital equivalent has none of those limitations but has one single, vital, prerequisite: connectivity for the users, and storage for the providers.

This observation is at the heart of the *as-a-service economy* [4], which is sure to expand far beyond the cloud as we know it and enter our everyday lives through the simple appendix of a connected device. The as-a-service economy materializes in apps that, once installed in the user's device, form a gateway to a gigantic and ever-growing wealth of potentially cooperating services. Access to an almost unlimited amount of information also alters the value that is attributed to that information. When people had to go to the effort of visiting a shop or library, for example, to access a song, a movie, a newspaper article or a book, information had value. Today, with near-free access to almost limitless content and information, their perceived value is less and people discard them more easily. When buying a newspaper, people will normally read at least part of it. With a subscription, more articles will go unread, and people are not per se listening to more music than they did in the past because they find it tiring to discover it. Most websites therefore have recommender systems that suggest a small range of selected items to the user, hoping he or she will click on them, like them… and keep paying the subscription.

COVID-19 is currently the digital transformation officer of the world

In 2020, COVID-19 has spurred the transition to the as-a-service economy. At record pace, yet more physical objects and activities have been digitized.

- Money: most people have now learned how to pay cashless (and sometimes even contactless) in shops. It is easy, more hygienic and more convenient. There is little chance that cash payments will make a great comeback in the future.
- Physical meetings: everybody has learned in 2020 that in order to meet, there is no need to move a brain in a body with a car to another location. A meeting can happen in the dematerialized cyberspace instead of in a physical meeting room. Although most of us are well aware of the limitations, we also appreciate the advantages (no need to move to another location, switching between meetings takes

less than one minute, it is very convenient to share documents, virtual meetings are easier to schedule, everybody is more equal – there are no reserved seats at the table, it is easier to do some other work if an item on the agenda does not require one's attention, some people even manage to be in two meetings simultaneously,…). There are also clear disadvantages: most people find virtual meetings more exhausting than physical meetings, it is less effective for meetings in which the participants do not know each other, scheduling a meeting between distant timezones is a challenge. It is however clear that a number of physical meetings will stay permanently substituted by virtual meetings in the future.

- Schools: although not ideal for compulsory education (from preschool to the end of high school), tele-teaching is suitable for knowledge transfer in higher education and lifelong learning. This is an example of the dematerialization of a classroom. It is however not a good way to teach skills or attitudes and it is definitely not good on a social level, to make friends or build a network, which is also an important goal of higher education. The fact that it is not good for compulsory education has to do with the fact that schools do more than teaching. They are also very important for children's social and emotional development, in some cases playing a role in protecting them (by bringing them in a safe environment, offering them a healthy meal, …).

Users are also discovering the impact of this new economy: if they have an internet outage, or the provider of a service is down, they are stuck without any options until the service starts again. Some people are discovering that physical books are more reliable than services that can be disrupted.

Maybe surprisingly, dematerialization does not only happen with goods and services that can be digitized: it also happens for material goods that are getting smaller and lighter, and hence require less (and lighter) physical material to manufacture [5].

The impact of streaming on the environment

Some people are worried about the climate impact of the use of digital technologies, fuelled for example by a report by the *Shift Project* indicating that watching one hour of NetFlix generated 3.2 kg CO_2. Although clearly too high because this number was based on old assumptions and on an error in which bit rate was confused with byte rate, there is a lot of controversy about this kind of study because of the huge economic interests involved (big tech + entertainment). This study generated a worldwide discussion on the climate impact of streaming. A follow-up study by George Kamiya from the International Energy Agency led to much lower numbers, but was also based on assumptions (only the operational costs were taken into account, not the investments in devices and data centres) and best estimates. The conclusions of this study are [6]:

- The energy efficiency of computing doubles every 2.7 years, and it doubles every two years for transmission networks.

It is important to base an analysis on recent data. Using five year old data will lead to an error of 400%;

- The energy consumption of hyperscale data centres has remained flat at 1% of global electricity use, while the traffic has tripled, and the workload has doubled. The increased demand has apparently been offset by efficiency improvements;
- The biggest energy cost of viewing streamed content is the viewing device itself. A smartphone is five times more efficient than a laptop, and 100 times more efficient than a 50-inch TV screen;
- According to George Kamiya, one hour of Netflix streaming consumes the amount of energy (in Wh) as shown in Figure 4. The total energy consumption is, in the first place, determined by the energy consumption of the device (smaller is better), and the resolution (smaller is better too). Depending on the mix of devices for Netflix, this leads to an average of 76.9 Wh, of which 70% is used in the device, 25% in the transmission, and 5% in the data centre;
- Expressed in terms of emissions, one hour of streaming is equivalent to 35.6 g CO_2 (based on the global average CO_2 intensity of 463 g/kWh in 2019). This is the equivalent of driving 250m (and eight times less on a smartphone). Only 30% is due to the streaming itself; the remaining 70% would still be used whatever the device was used for (like watching a broadcasted movie, or playing a game). These numbers are confirmed by other recent studies [8,9];
- The numbers also depend heavily on the origin of the electricity, and on the time of the day the streaming happens. In France, it would only be 4g CO_2 or driving 28m;
- These numbers will keep falling in the coming years due to efficiency gains in data centres, transmission networks and devices, and the decarbonization of electricity generation.

Energy consumption (Wh)	TV	Laptop	Smartphone	Average
	(Wifi, 4K)	(Wifi, HD)	(4G, Auto)	-
Data centre	13.9	6	0.5	3.7
Transmission	18.8	18.3	8.5	17.7
Device	120	22	1.2	55.5
Total	**152.7**	**46.3**	**10.2**	**76.9**

Figure 4: Energy consumption of 1h streaming on different platforms

The conclusion is that it is very difficult to draw a definitive conclusion on such studies, in part because there are a range of financial and commercial implications of their results. But, if we follow George Kamiya's study, the carbon footprint of streaming a movie is moderate, and definitely much lower than taking a car to go to a movie theatre, and the gap between the two will continue to widen in the future. Given the similarity between streaming video content and videoconferencing systems, meeting virtually will always consume less energy than attending a physical meeting. The same holds for cloud gaming and streaming content including gaming.

However, there is also Jevon's paradox that states that more efficient use of a resource can lead to lower cost and therefore increased demand, undoing the effects of the efficiency gains. The streaming market seems to grow faster than the efficiency gains in computing.

Conclusion

Dematerialization has been ongoing for the last forty years, and there is no reason why it would or should stop now. Dematerialized services might consume fewer resources than physical ones, they are cheaper and are available 24/7, yet there are many concrete challenges to work on. The environmental footprint (also in terms of energy) should be further reduced and existing solutions should be improved; for example, the tools for virtual meetings, lectures and conferences that do not yet offer the immersive experience that physical events offer.

On top of that, there are still huge opportunities for big virtual events, for tourism, and for museum visits. How cool would it be to 'visit' a city or a museum with an interactive video guide who shows you all the interesting places or objects, and where you can determine how the tour will evolve? This could happen at home, but perhaps also in a virtual tourism facility with a fully immersive experience including the sounds, the smells, the burning sun, and maybe a meal with local food afterwards.

Finally, we should also take into account the possible negative impact of digital-only information, such as the loss of ownership of content, which comes as a result of the content providers deciding what to offer.

References

[1] Felix Richter, "Streaming Drives Global Music Industry Resurgence", 2020, https://www.statista.com/chart/4713/global-recorded-music-industry-revenues/

[2] Mansoor Iqbal, "Netflix Revenue and Usage Statistics (2020)", 2020, https://www.businessofapps.com/data/netflix-statistics/

[3] Elisa Shearer, "Social media outpaces print newspapers in the U.S. as a news source", 2018, https://www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/

[4] "Servitization in Industry", Gunter Lay (editor), 2014, Springer, ISBN 978-3-319-06934-0, doi: 10.1007/978-3-319-06935-7.

[5] "Dematerialization", https://en.wikipedia.org/wiki/Dematerialization_(economics)#

[6] George Kamiya, "Factcheck: What is the carbon footprint of streaming video on Netflix?", 2020, https://www.carbonbrief.org/factcheck-what-is-the-carbon-footprint-of-streaming-video-on-netflix

[7] Brad Stone, "Amazon Erases Orwell Books From Kindle", 2009, https://www.nytimes.com/2009/07/18/technology/companies/18amazon.html

[8] "The carbon footprint of (nearly) everything", https://www.viessmann.co.uk/company/blog/the-carbon-footprint-of-nearly-everything

[9] Louis-Philippe P.-V.P. Clément, Quentin E.S. Jacquemotte, Lorenz M. Hilty, Sources of variation in life cycle assessments of smartphones and tablet computers, Environmental Impact Assessment Review, Volume 84, 2020.

Koen De Bosschere is Professor in the Electronics department of Ghent University, Ghent, Belgium.

Marc Duranton is Researcher at the Research and Technology Department of CEA (Alternative energies and Atomic Energy Commission), France and the coordinator of the HiPEAC Vision 2021.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: K. De Bosschere and M. Duranton. Everything as a service. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 206-211, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

Nothing is as seriously important as playing, not just in IT.

Gaming trends

By THOMAS HOBERG

As with children, the future of information technology depends on the quality of its playing: this lies in experimentation, which is at the heart of engineering and science.

Key insights

- Computer gaming with its huge user base and unlimited appetite for realism has become the most important driver for innovation in information technology.
- Games push every frontier: increased realism, larger worlds, bigger crowds, lower cost, greater network bandwidth and lower network latencies.
- Planet-sized distributed game worlds enable critical global digital twins for virtual personal assistants (VPA).
- The battle of digital distribution platforms calls for regulation, splitting roles to avoid strongarming consumers.

Key recommendations

- Take gaming seriously: it is the main motor of innovation in every field of IT.
- Support serious applications for gaming technology: autonomous vehicles and machinery as well as virtual personal assistants require digital twins in an open world game at planet size.
- Ensure the EU have as full a set of IP assets for game technology as possible: treat it as critical infrastructure and encourage the development of open source components to enable heterogenous open worlds.
- Regulate to separate gaming platforms, separate game studios, stores, servers and draw ground geography borders into the clouds to avoid monopolization and ensure consumers won't lose access to purchased games or are forced into subscriptions.
- Regulate that purchased game titles can be freely moved between platforms and devices and that behind a 'buy' button is a real purchase.

Gaming at the core of IT evolution

Without acting out our ingrained playfulness we could not make sense of our senses, nor articulate our limbs let alone a thought. Life is a constant dance in our brain, what we sense, what we do, and how that changes our perception. When kittens play "catch-a-tail" or babies "grab-taste-drop" they develop the most important skills required for long-term survival through systematic experimentation and training. The world is really a game in our head, reality what we share with those we choose to believe, parents or peers, charlatans or scientists.

Experimentation with careful analysis and abstraction allowed engineers and scientists to develop models that lend themselves to extrapolation and recombination, so results could be projected: not every bridge or cathedral had to be fully tested before the final construction, once the statics and materials were understood. From the very beginning, computers have been used to fill the gaps between experiments and project the properties of experiments now executed much quicker as simulations. In fact, computer chips themselves are among the most complex structures, constructed and simulated in code-only for most of their design, aiming for immediate usability of the first production run. Such pure digital design, where most if not all design stages are made on computers, became possible mostly because of a virtuous circle so typical in information technology:

- Computers became cheap enough to experiment and play with;
- Some of these games were inspiring enough to increase the user base;
- The additional user base paid for an improved implementation;
- Improved implementations found professional applications.

The first hobby computer flight simulator from 1980 on the Apple][computer

Figure 1: Early version of the flight simulator. (Source: https://upload.wikimedia.org/wikipedia/en/5/5f/Sublogic_Flight_Simulator_II.png)

offered little more than the main dials of an aircraft and an extremely rough sketch of major landmarks and the landing strips as coloured dots on 280x192 pixels, but the moving contours provided a seasoned flyer or passenger with enough visual cues to imagine moving through the air in a plane. The 1 MHz 8-Bit CPU not only had to manage a physical model of an aircraft and the effect of its control surfaces on the flight path, but it had to project a vector representation of the terrain onto the bitmap display and draw each resulting pixel. The human mind did most of the work of interpolation and imagination into a glide path, which worked incredibly well and it is hard to exaggerate the inspirational effect it and its successors had on the computer industry.

Good looking vs. looking good enough, another uncanny valley

Early games resemble stories and books where the listener was responsible for translating a description of a scene into an image in the mind. Realism had to be limited to what computers could support and users could afford; real-time response was much more important. Comic-like two-dimensional side scrolling games gained colour, detail and resolution, but 3D solutions targeted the professional market, where very boring surfaces with flat colours might actually be preferred in CAD.

A quantum jump in game quality was achieved in 1995 when texture mapping accelerators became affordable, which would project an image bitmap on an otherwise rather flat polygon mesh composed of triangles. Ironically this happened just after Wolfenstein, Doom and Quake from ID-Software had popularized 3D games which had been extensively optimized to run with software-only rendering. Since it could be performed for every triangle in an embarrassingly parallel manner without interdependence, dedicated hardware could perform this much faster than any CPU via iteration, even if those raced from 60-3800MHz in the ten years that followed. It enabled a new species of games with a third dimension, mazes and worlds much easier to get lost in and hooked on.

A naturalistic rendering of a real scene would require not only infinitesimal detail in data and model, it also requires tracing the path of each ray of light as it reflects, scatters or is finally absorbed. It explodes in complexity with image resolution and dynamic range and puts a linear dependency on each ray with a variable number of iterations breaking the latency requirements of games. It became the preferred path for digital art and computer generated films, which could much less afford to sacrifice image quality and therefore run their final renders for weeks or months on giant computer farms for an hour of video.

For many iterations of graphic processing units (GPUs), the only path to better-looking graphics was to improve the cheating around the processing of textures from a fixed function process to something more adaptive. For example, instead of breaking up larger triangles into smaller ones to create finer contours, bump maps were introduced, which add a z-axis offset much like colour to a texture to create dents and elevations, or add logic to fuzzy or sharpen edges and texture details to avoid edge and texture flicker as surfaces moved much closer or further away. It ultimately led to an architecture that used code routines to do all the texture processing instead of fixed function ASIC blocks and thus gave birth to today's GPUs, which are fully programmable and have transformed HPC and machine learning. They have even picked up some limited ray tracing functionality recently and became flexible enough to accelerate digital film rendering, but nowhere near real-time frame rates. It is a bit like zooming in on a Mandelbrot set: the computational gap between fidelity and frame rates only seems to increase, the closer you get.

Something similar is now happening with the simulation model behind the scenes.

Most of the early 3D games were quite literally set in stone. Stone walls helped to limit the computational cost of visual depth while their immutable nature allowed eliminating many inside surfaces and polygons. It also made for minimal communication bandwidth requirements for multi-player games, as a static world would not alter, only dynamic object and player information needed to be exchanged. Even today creating the most realistic looking maps, objects and characters requires artistic design and manual optimizations, resulting in a game where the missions are rigged and consist in little more than finding one of the programmed solutions.

Another genre of games comes from simulation backgrounds; Conway's game of Life or Dyment/Ahl's Hamurabi [1] have

Figure 2: An Autodesk 3DS Max example render using a photorealistic render to showcase the technology's importance in architecture (Source: https://static.chaosgroup.com/gallery_images/images/000/000/159/gallery_image/conor-harll-design-architecture-vray-3ds-max.jpg?1518068742)

inspired generations, The Sims [2] and the Anno series of games are still being continued. These sandbox games typically contain randomly generated open world maps made of grid or block elements, where parametric ruleset govern inter-element relations and operations both during initial creation and during all in-game interactions. The code base, both for behaviour and for visualization, can be very complex, while the data structures for element individualization are kept small to allow for large worlds and massive numbers of players sharing them on a server.

For lack of individual artist design, these games may offer a less naturalistic visual appeal, but much better scalability and extensibility. Minecraft with its Lego esthetics is a prime example and with it carries the ability to become a full parametric CAD design tool, physical experiment simulator or urban planning tool [3] via code extension to the engine and block elements.

Complex buildings like Notre Dame [4] have been replicated and the BlockBy-Block [5] initiative, which uses Minecraft for community driven neighborhood improvement, hints at full digital twins of real-world environments, which even at Minecraft granularity and abstraction wouldn't easily fit our planet on any single computer.

Massive multiplayer games like EVE online [6] feature over 60 000 planets and millions of players inhabiting a single universe. While a lot of the empty space between planets can be fit into a game simulation cube data structure, once players approach a planet's surface or just interact with their spaceships, the level of visual and logical detail to model the parallel universe has no intrinsic limit, only technical barriers. If thousands of players then engage in battle as they did in the "Bloodbath of B-R5RB [7]", the question arises: how does one distribute the model, the interactions and the visualization?

For high-quality lag-less visualization the detailed 3D model should be local to the PC or console to render via the GPU, especially for virtual reality. But where exactly are the other thousands of players moving at rocket speed and where would you aim to hit the enemy? To avoid shooting at ghosts or being killed by an enemy never seen, the detailed logical model of the game universe needs to be kept centrally, updated and perhaps even rendered simultaneously via cloud GPUs, to minimize latency-induced position skew. With the ability to use game engine data for motion prediction to create latency optimized video encoders as fixed function IP blocks on GPUs, the transmission time for GPU generated video content at a given resolution and fidelity now has a manageable, even adjustable upper limit. But the transmission time for a model of empty space vs. thousands of players and their ships in a single location could vary from milliseconds to way beyond tolerable. Much better than Minecraft granularity, physics at the quality of a car crash simulation, and a game world the size of galaxies make Facebook look rather tiny, and that application runs distributed on millions of servers. Representing space, objects, people and managing their interactions takes more than scaling up an ERP architecture. It requires scale-out and smart partitioning at a level that requires innovating from ground up, implementing a fabric that forwards information by running spatial algebra in hardware on the data plane, not just a spatial operating system [8]. Compared to galaxy sized world models involving immensely complex faster-than-light autonomous vehicles, Pokémon Go's creatures only spawn locally to interact with human players on foot. But it also showcases how games drive this type of innovation towards a global digital twin of everything, the digital replica of our planet that our virtual personal assistants need to link into to work most effectively, to reliably avoid undesired conflicts while encouraging only the interactions we find valuable, which makes it a key topic for EU funded research and open source assets.

The money angle

When it comes to the cost of entertainment, the live performance of a 19th century opera or ballet may top the ranks

Figure 3: Minecraft (Source: https://en.wikipedia.org/wiki/Minecraft#/media/File:Minecraft_cover.png)

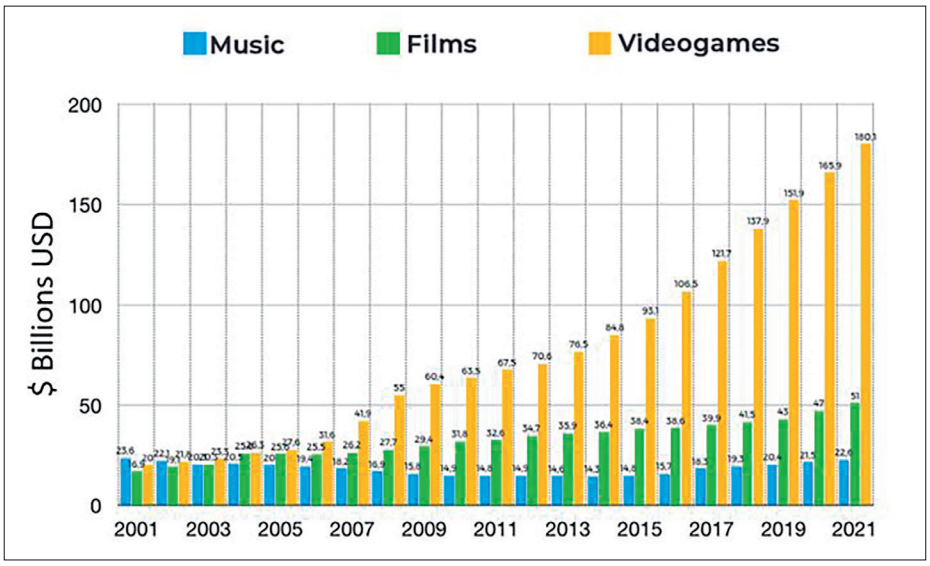


Figure 4: Revenue of music, film and game industry (Source: https://twitter.com/intangiblecoins/status/1154403586417274880/photo/1)

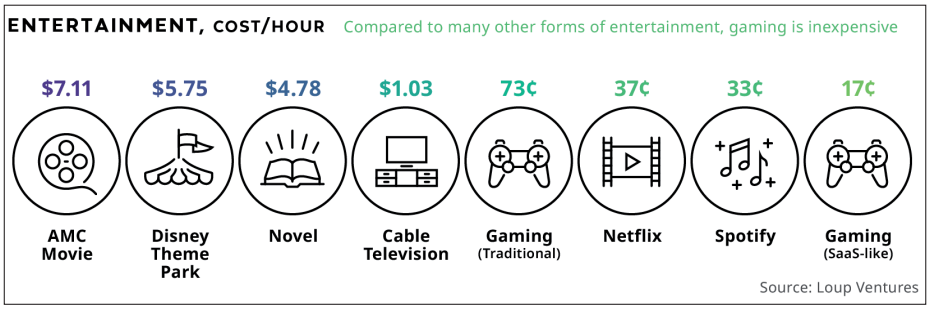


Figure 5: Entertainment cost/hour (Source: https://www.visualcapitalist.com/multi-billion-dollar-console-gaming-market/)

of cost per hour: Hundreds of highly skilled professionals are working together to create an ephemeral performance only royal coffers could originally bring to stage. Television added orders of magnitude to the audience cutting the effective cost; recordings and films again offer dramatic cost reductions to being entertained, allowing to vastly expand the stage setting into a full movie with a cast that may count in hundreds, true or now increasingly simulated masses. But gaming replaces the highly skilled and most beautiful (or scary) expensive actors by code, both slashing the cost and adding a degree of interaction, film and television cannot offer. While AIs are busily being developed to add beauty to average looking actors or drama to uninspired games, we can see how the gaming industry has been eclipsing music and film for more than a decade now, without any sign of slowing down.

And the main reason behind that trend is that the computer automated characters in gaming are even cheaper than the human lay actors in a reality show.

Within the gaming industry we see the cost trend dominating in a similar manner, where the cheapest end-user device, the smartphone every consumer with money to spend already has, receives the biggest number of low-threshold small-value purchases, which cumulate to a trillion dollar business, leaving PCs and consoles behind.

Entry level smartphones are already as capable as a PlayStation 3, Nintendo WII or an Xbox 360, while even mid-range smartphones snap at the heels of the PlayStation 4 or Xbox One, which are currently being replaced. But phones, tablets, Chromebooks or pretty much any device with a screen are increasingly used as an end point for remote rendering cloud gaming, which extends the advantage of unbeatable device cost even into the domain of otherwise unaffordable visual experience and has everyone in big tech entering into cloud gaming.

Issues of immaterial possession

Massive gamer populations sharing a single game world is not the only reason cloud-based game rendering has been tried and tried again, starting in 2009 with OnLive [9], purchased by Sony who used the technology for their PlayStation Now offering. More recently practically all web

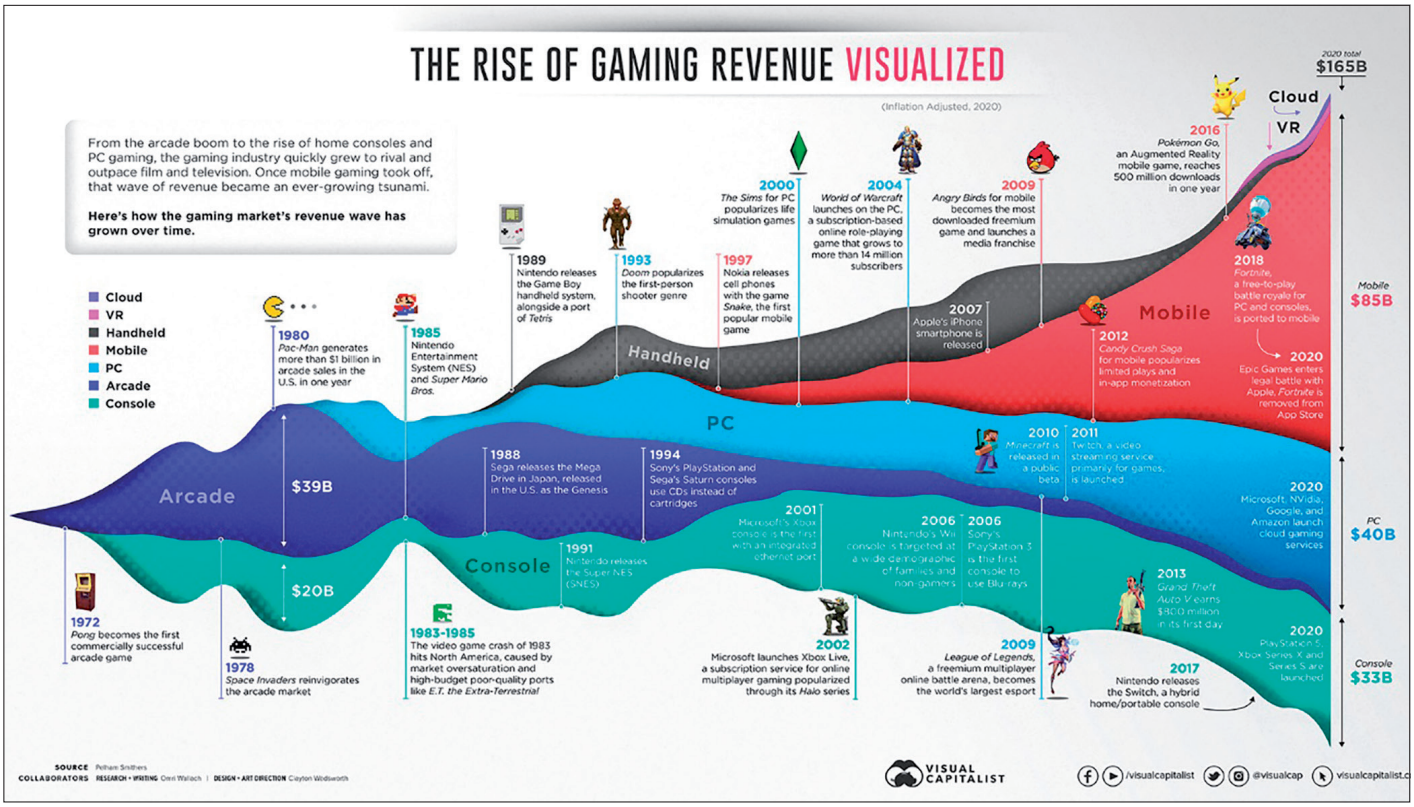


Figure 6: Gaming revenue by platform with major events (Source: https://www.visualcapitalist.com/wp-content/uploads/2020/11/history-of-gaming-by-revenue-share-full-size.html)

giants have joined the fray, Google with Stadia, Amazon with Luna, Microsoft with xCloud as well as console or hardware vendors like Sony, Nvidia and Apple.

There could be potential consumer benefits, like the ability to play on any screen from mobile phone, tablet, lightweight laptops to the biggest TV or projector screens. Likewise, games can be carried from home to friends, on trips or rides without actually taking disks, consoles or a full gaming rig along. Such a rig can easily cost thousands of Euros, if you want something up-to-date on a PC, consoles still cost hundreds and every few years they become obsolete. If you want to play with your kids or if those come with friends, the multipliers become a challenge.

Blockbuster games not only cost hundreds of millions of real money to develop and are bought by hundreds of millions of gamers, they are also reaching hundreds of gigabytes in size, which makes casually gaming with several titles seem very attractive in a cloud even with fibre connections. The latest incarnation of Microsoft's flight simulator had many buyers wait days for the first minimal download to complete before they could launch the game and a plane, but the quality is visibly improved from forty years ago.

But one of the biggest concerns with computer games has always been that kids love to share. The earliest Atari home computers and most game consoles started with ROM cartridges for games few kids could copy easily, while taking them to a friend helped spread the ecosystem. When personal computers made it to every desktop with floppy disk, CD and DVD media, their lack of copy protection in the base platforms spawned a copy protection industry, each with its own copy protection scheme, some of which included a rootkit [10]. Popular game consoles from Sega and Sony switched to CD and DVD media for low cost distribution and did include various technologies [11] to prevent copying disks, while later consoles like Sony's PlayStation and Microsoft Xbox would include cryptographically protected boot ROMs and hypervisors, also very common in mobile phones and now making it into future personal computers [12].

Valve Software started out as a game developer, then started to develop Steam in 2003, originally only as an internet platform to distribute updates for their own games. Since 2009 Valve enabled publishers using their platform to deliver downloaded games encrypted under a user-specific key, which didn't inhibit copying or moving games, but required at least temporary online interactions to enable game start, essentially a DRM so minimally intrusive yet effective it gained market dominance for game distribution mostly by consumer choice. Today it effectively is the app store for personal computer games, it charges the same 30% revenue share Apple and Google demand for their walled gardens, but leads

Figure 7: A cockpit snapshot from Microsoft's 2020 Flight Simulator (Source: Xbox.com)

mostly by customer convenience, allowing users to play games with a maximum of flexibility, local streaming and cloud-based game state replication across machines and even operating systems. That bothers practically everyone at Microsoft and Apple, who would rather prefer complete control over what they consider their own as well as the largest game studios, who would prefer to keep those 30% and have set up similar distribution platforms for their own titles [13] and initiated a price war [14], which may turn out very unfortunate for consumers:

While Steam and the other platforms promote the illusion of a true game purchase, the licence to the game is effectively tied to the platform and lost if it should go down. Competitive pricing would seem to benefit consumers at first glance, but turns platform choice into online gambling

EPIC is using its power as the producer of one of the most important game engines [15] to tilt the platform battle in their favour, much like Microsoft is flexing its Xbox and Windows muscle and Google and Apple work hard to retain their exclusive hold on Android and iOS. This is clearly not in the interest of those consumers, who would prefer to retain and run their purchased games and the non-game applications independent from the distribution platform and across all base architectures supported.

Separating the sale and the rent for the cloud hosting of game distributables and game saves with the right to move both between providers could be a sensible entry point for EU consumer protection regulation.

Nvidia's GeForce NOW is quite different from the other big entrants into the cloud remote gaming market in that it is a pure rental for the hardware. It doesn't include a store but instead validates game licences against the popular personal computer distribution platforms and ideally works with a cached solid-state copy to reduce game startup times.

Unfortunately, many publishers refuse to let their titles run on GeForce NOW, routinely without comment but evidently not with consumer's best interest at heart.

Google Stadia, Amazon Luna and Microsoft xCloud may offer the same titles you already own for your PC, but not only do you have to purchase them again to run there, you stand to lose them should you terminate your subscription. All of them share that the visual quality, resolution and reactivity is still far below what their marketing materials suggest, while Steam remote gaming in a home LAN proves that it must be a network issue.

True digital goods like e-books or classic audio recordings from Karajan and his Berlin Philharmonics neither suffer decay nor require significant upgrades year after year. They could be inherited for generations and provide enjoyment without any additional expense or revenue. Purchase and ownership by consumers mean far too fickle revenue streams for digital content providers, which surely explains why we see them converging on subscription models for books, music, films, games, Office suites etc. and the delivery of each. The business model for the vendors is so compelling, it is hard to predict how long consumers will retain a choice. And since the reduction of consumer choice provides significant bottom line benefits, addiction is a weapon that choses itself at every opportunity and requires active countermeasures imposed and controlled via regulation.

One bright thing to look forward to: vendors will still want to sell exclusive shiny hardware to go with that subscription.

References

[1] Wikipedia, "Hamurabi", https://en.wikipedia.org/wiki/Hamurabi_(video_game)

[2] Wikipedia, "The Sims", https://en.wikipedia.org/wiki/The_Sims

[3] Fruzsina Eordogh, "Minecraft Partners With United Nations For Urban Planning", https://readwrite.com/2012/09/06/minecraft-partners-with-united-nations-for-urban-planning/

[4] Minecraft, "Notre Dame de Paris", https://www.planetminecraft.com/project/notre-dame-de-paris-4108833/

[5] Block by Block, https://www.blockbyblock.org/

[6] MMO Populations, "Eve Online", https://mmo-population.com/r/Eve/

[7] Wikipedia, "Bloodbath of B-R5RB", https://en.wikipedia.org/wiki/Bloodbath_of_B-R5RB

[8] Improbable, "Multiplayer Netowrking", https://improbable.io/multiplayer-networking

[9] Wikipedia, "OnLive", https://en.wikipedia.org/wiki/OnLive

[10] Wikipedia, "SecuROM", https://en.wikipedia.org/wiki/SecuROM

[11] Wikipedia, "List of Compact Disc and DVD copy protection schemes", https://en.wikipedia.org/wiki/List_of_Compact_Disc_and_DVD_copy_protection_schemes

[12] Microsoft, "Meet the Microsoft Pluton processor – The security chip designed for the future of Windows PCs", https://www.microsoft.com/security/blog/2020/11/17/meet-the-microsoft-pluton-processor-the-security-chip-designed-for-the-future-of-windows-pcs/

[13] EPIC, Ubisoft, Electronic Arts, but also Microsoft

[14] Nick Statt, "Epic vs. Steam: The Console War Reimagined on the PC", https://www.theverge.com/2019/4/16/18334865/epic-games-store-versus-steam-valve-pc-gaming-console-war-reimagined

[15] Unreal Engine, https://www.unrealengine.com/en-US/

Thomas Hoberg is Technical Director R&D at Worldline, Germany.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: T. Hoberg. Gaming trends. In M. Duranton et al., editors, HiPEAC Vision 2021, pages 212-217, Jan 2021.

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.

Glossary

5GIA	5G Industry Association
ACAS	Airborne Collision Avoidance Systems
ADA	ADA Programming language
ADAS	Advanced Driver-Assistance Systems
AES	Advanced Encryption Standard
AGI	Artificial General Intelligence
AI	Artificial Intelligence
AIOTI	The Alliance for the Internet of Things Innovation
ALU	Arithmetic Logic Unit
ANPR	Automatic Number Place Recognition
ANT	Multicast wireless sensor network technology, designed by ANT Wireless
API	Application Programming Interface
APL	A Programming Language
ARM	Adanced RISC Machines
ASIC	Application-Specific Integrated Circuit
ASLR	Address Space Layout Randomization
ASML	ASML Holding N.V. is a Dutch company and currently the largest supplier in the world of photo-lithography systems for the semiconductor industry.
Auto-ML	Techniques to design the meta-parameters associated with deep learning networks
AWS	Amazon Web Services
B2B	Business-to-Business
B2C	Business-to-Consumer
BATX	Baidu, Alibaba, Tencent, Xiaomi
Bayesian computing	Bayesian computing refers to computational methods that are based on Bayesian (probabilistic) statistics.
BDVA	Big Data Value Association
Big data	Complex and exceedingly large data sets
BLE	Bluetooth Low Energy
BSD	Berkeley Software Distribution
C2PS	Cognitive Cyber-Physical Systems
C3PS	Connected Cognitive Cyber-Physical Systems
CAD	Computer-Aided Design
CAGR	Compound annual growth rate is a specific business and investing term for the smoothed annualised gain of an investment over a given time period.
CBRAM	Conductive-Bridging RAM
CC	Creative Commons
CD	Compact Disc
CERN	Conseil européen pour la recherche nucléaire, European Organization for Nuclear Research
CGRA	Coarse-Grained Reconfigurable Architecture
CIS	CMOS Image Sensor
CLAIRE	Confederation of Laboratories for Artificial Intelligence Research in Europe
Cloud computing	Cloud computing is a paradigm whereby computing power is abstracted as a virtual service over a network. Executed tasks are transparently distributed.
CMOS	Complementary Metal–Oxide–Semiconductor is a common technology for constructing integrated circuits. CMOS technology is used in microprocessors, microcontrollers, static RAM, and other digital logic circuits.
CNTK	Microsoft Cognitive Toolkit
COBOL	Programming Language
COVID	Corona Virus Disease
CPS	Cyber-Physical Systems combine computing resources and sensors/actuators that directly interact with and influence the real world. Robotics is one of the primary fields that works on such systems.
CPU	Central Processing Unit
CRISPR	CRISPR gene editing is a genetic engineering technique in molecular biology by which the genomes of living organisms may be modified. It is based on a simplified version of the bacterial CRISPR-Cas9 antiviral defense system.
CS	Computer Science
CSA	Coordination and Support Action, type of EU project
CWI	Centrum Wiskunde & Informatica is the national research institute for mathematics and computer science in the Netherlands
DAB	Digital Audio Broadcasting
DARPA	Defense Advanced Research Projects Agency
Data analytics	Data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights.
DCC	Digital Cassette
Deep learning	A class of machine learning techniques characterised by having deeply stacked layers that process the input.

DG Directorate General

DIH Digital Innovation Hubs

DIY Do-it-yourself

DLX An ISA developed at Berkeley

DMD Digital Micro-Mirror Devices

DNA Deoxyribonucleic Acid

DRAM Dynamic RAM

DRM Digital Rights Management

DSL Domain Specific Language

DSSTNE Deep Scalable Sparse Tensor Network Engine

DVD Digital Versatile Disc or Digital Video Disc

DVR Digital Video Recorder

ECG Electrocardiogram

ECMWF European Centre for Medium-Range Weather Forecasts

ECS Electronic Components and Systems

ECSO European Cyber Security Organisation

EDA Electronic Design Automation

Edge computing
Edge Computing is pushing the frontier of computing applications, data, and services away from centralized nodes to the logical extremes of a network. It enables analytics and knowledge generation to occur at the source of the data.

EMIB Embedded Multi-Die Interconnect Bridge

ENIAC Electronic Numerical Integrator and Computer, the first programmable, electronic, general-purpose digital computer

EPI European Processor Initiative

ERI Electronics Resurgence Initiative

ERP Enterprise resource planning

EU MATHS IN
European Service Network of Mathematics for Industry and Innovation

EUV Extreme ultraviolet lithography is a next-generation lithography technology using an extreme ultraviolet (EUV) wavelength, currently expected to be 13.5 nm.

FCC Federal Communications Commission, an independent agency of the United States government that regulates communications by radio, television, wire, satellite, and cable across the United States.

FDA U.S. Food and Drug Administration

FDSOI Fully Depleted Silicon On Insulator (MOSFETs). For a FDSOI MOSFET the sandwiched p-type film between the gate oxide (GOX) and buried oxide (BOX) is very thin so that the depletion region covers the whole film. In FDSOI the front gate (GOX) supports less depletion charges than the bulk transistors so an increase in inversion charges occurs resulting in higher switching speeds. Other drawbacks in bulk MOSFETs, like threshold voltage roll off, higher sub-threshold slop body effect, etc. are reduced in FDSOI since the source and drain electric fields cannot interfere, due to the BOX (adapted from Wikipedia).

FEOL Front-End-Of-Line, is the first step in fabricating an IC, in which devices on the wafer (such as transistors, resistors, etc.) are formed.

FET Field-effect transistor

FHE Fully Homomorphic Encryption, a form of encryption that allows computations being performed on data without having the key to decrypt that data

FinFET A FinFET is a nonplanar, double-gate transistor built on an SOI substrate.... The distinguishing characteristic of the FinFET is that the conducting channel is wrapped by a thin silicon 'fin', which forms the body of the device. In the technical literature, FinFET is used somewhat generically to describe any fin-based, multigate transistor architecture regardless of number of gates (from Wikipedia).

FMCG Fast-moving consumer goods

Fog computing
Fog computing is an architecture that uses one or more end-user clients or near-user edge devices to carry out a substantial amount of storage (rather than stored primarily in cloud data centres), communication (rather than routed over the internet backbone), control, configuration, measurement and management.

FPGA Field-Programmable Gate Array

FSF Free Software Foundation

FTE Full Time Equivalent

GaAs Gallium Arsenide

GaaFET Gate-all-around FET

GAFA Google, Amazon, Facebook and Apple

GAFAM Google, Apple, Facebook, Amazon and Microsoft

GaN Gallium Nitride

GCC The GNU Compiler Collection

GDP Gross Domestic Product

GDPR General Data Protection Regulation (EU) 2016/679

Generative Design
Generative design is a technology that starts with your design goals and then explores all of the possible permutations of a solution to find the best option. Using cloud computing, generative design software quickly cycles through thousands—or even millions—of design choices, testing configurations and learning from each iteration what works and what doesn't. The process lets designers generate brand new options, beyond what a human alone could create, to arrive at the most effective design

GM General Motors

GNU GNU's Not Unix

GPL GNU General Public License

GPS Global Positioning System

GPU A Graphics Processing Unit refers to the processing units on video cards. In recent years, these have evolved into massively parallel execution engines for floating point vector operations, reaching performance peaks of several gigaflops.

HBM High-Bandwidth Memory

HD Replace by HDD in service paper

HDD Hard Disk Drive

HHS Department of Health and Human Services

HiPEAC The European Network of Excellence on High Performance and Embedded Architecture and Compilation coordinates research, facilitates collaboration and networking, and stimulates commercialization in the areas of computer hardware and software research.

HIV Human Immunodeficiency Virus

HLS High-Level Synthesis

HMC Hybrid Memory Cube

Homomorphic encryption
Homomorphic systems send encrypted data to an application (generally executed on a remote server) and lets an application perform its operations without ever decrypting the data. As a result the application never knows the actual data, nor the results it computes.

HPC High Performance Computing

HPDA High-Performance Data Analytics

HTML HyperText Markup Language

HTTP HyperText Transfer Protocol

IARPA Intelligence Advanced Research Projects Activity

IC Integrated Circuit

ICT Information & Communication Technology is a generic term used to refer to all areas of technology related to computing and telecommunications.

IDC International Data Corporation, a premier market intelligence provider

IDM Integrated Device Manufacturer

IFTTT "If This, Then That", company providing a software platform that connects apps, devices and services from different developers in order to trigger one or more automations involving those apps, devices and services

IGZO Indium-Gallium-Zinc-Oxide

III-V Chemical compounds with at least one group III element and at least one group V element.

IMF International Monetary Fund

Imperative programming
Imperative programming is a programming paradigm that describes computation in terms of statements that change a program state. In much the same way that the imperative tense in natural languages expresses commands to take action, imperative programs define sequences of commands for the computer to perform. The opposite concept is declarative programming.

InFO Integrated Fan-Out (InFO) advanced packaging technology from Taiwan Semiconductor Manufacturing Company (TSMC).

Internet of Things
The Internet of Things (IoT) is a computing concept that describes a future where everyday physical objects will be connected to the internet and will be able to identify themselves to other devices.

IO Input Output

IP 1. Internet Protocol 2. Intellectual property

IP block Intellectual property block, is a reusable unit of logic, cell, or chip layout design that is the intellectual property of one party. IP cores may be licensed to another party or can be owned and used by a single party alone. IP blocks can be used as building blocks within ASIC chip designs or FPGA logic designs.

IPCC Intergovernmental Panel on Climate Change

ISA An Instruction Set Architecture is the definition of the machine instructions that can be executed by a particular family of processors.

ISO From the greek ίσος = equal, International Organization for Standardization

ISS Integrated Smart Systems

IT Information Technology

ITAR International Traffic in Arms Regulations

ITS Intelligent Transportation Systems

JIT Just-in-Time

JVM Java Virtual Machine

KASLR Kernel Address Space Layout Randomization

KPI Key Performance Indicator

LAN Local Area Network

LCD Liquid Crystal Display

LE Large Enterprise

LED Light Emitting Diode

LGPL GNU Lesser General Public License

LIDAR Light Detection And Ranging is a technology that measures distance by illuminating a target with a laser.

LLVM The LLVM Project is a collection of modular and reusable compiler and toolchain technologies.

LTE Long Term Evolution, a standard for mobile internet communications

LTPS Low Temperature Polycrystalline Silicon

MAPE Monitor-Analyze-Plan-Execute

MAS Multi-Agent Systems

MCAS Maneuvering Characteristics Augmentation System, in airplanes

MCU Micro Controller Unit

MDE Model Driven Engineering

MEMS Micro-Electrical-Mechanical Systems

MFF Multiannual Financial Framework

MFM Magnetic Force Micrograph

MIPS Microprocessor without Interlocked Pipeline Stages, a RISC ISA

MIT Massachusetts Institute of Technology

ML Machine Learning

MMX Matrix Math Extensions, instruction set extensions for the Intel Processor

MNIST A large database of handwritten digits

MOS Metal-Oxide-Semiconductor

MPU Micro Processor Unit

MRAM Magnetic RAM

MRI Magnetic Resonance Imaging

MS-DOS Microsoft Disk Operating System

MSODE Modelling, Simulation and Optimization Methodologies in Data-rich Environments

MTE Memory Tagging Extensions

MTJ Magnetic Tunnel Junction

NA Numerical Aperture

NAND NOT-AND, a type of logic gate

NAS Network attached storage

NES Nintendo Entertainment System

Neural networks

Neural networks are computational entities that operate in a way that is inspired by how neurons and synapses in an organic brain are believed to function. They need to be trained for a particular application, during which their internal structure is modified until they provide adequately accurate responses for given inputs.

Neuromorphic

Analog, digital, or mixed-mode analogue/digital VLSI and software systems that implement models of neural systems.

NFC Near Field Communication

NGO Non-Governmental Organizations

NIST National Institute of Standards and Technology

NLP Natural Language Processing

NML Nanomagnet Logic Quantum Cellular Automata

NoC Network-on-Chip

NOR NOT-OR, a type of logic gate

NRE Non-Recurring Engineering costs refer to one-time costs incurred for the design of a new chip, computer program or other creation, as opposed to marginal costs that are incurred per produced unit.

NSA National Security Agency

NVM Non-Volatile Memory

OADR Old Age Dependency Ratio

OCP Open Compute Project

OECD Organisation for Economic Co-operation and Development

OLED Organic Light Emitting Diode

OMG Object Management Group

OPC Optical Proximity Correction

Open source Projects (software, schematics, etc.) in which the relevant source files are distributed to end users. Depending on the type of license, users can also be allowed to modify and redistribute these projects.

Operational research

Mathematical study of making decisions

OPU Optical Processing Unit, produced by Lighton

OPV Organic Photovoltaics

ORNL Oak Ridge National Laboratory

OS Operating system

OSAT Outsourced Semiconductor Assembly & Test, companies performing IC packaging and testing

OSH Open Source Hardware

Ox RAM Oxide based RAM

PAC Pointer Authentication Code

PC Personal Computer

PCB Printed Circuit Board

PCM Phase Change Memories

PCR Polymerase Chain Reaction

PCRAM Phase Change Memories

PCT Patent Cooperation Treaty

PDMS PolyDimethylSiloxane

PHP A programming language.

PII Personally Identifiable Information

Post-quantum cryptography

Field of study in which cryptography is made secure in the presence of quantum computers

PPE Personal Protective Equipment

Programming model

A programming model is a collection of technologies and semantic rules that enable the expression of algorithms in an efficient way. Often, such programming models are geared towards a particular application domain, such as parallel programming, real-time systems or image processing.

Pseudo-quantum computing

Pseudo-quantum computing is a term used to refer to machines that allegedly are quantum computers, but that in practice have not actually been proven to be faster than regular computers executing very optimized algorithms.

PULP Parallel Ultra-Low-Power

Python A programming language

QA Quality Assurance

QoS Quality of Service

RAM Random-Access Memory

RE Requirements Engineering

Reservoir computing

Reservoir Computing is similar to neural networks, but rather than modifying the internal structure during the training phase, the way to interpret the output is adjusted until the desired accuracy has been obtained.

REST Representational State Transfer. A paradigm for transferring, accessing, and manipulating textual data in a stateless manner.

RF Radio Frequency

RFID Radio-Frequency Identification is the use of a wireless non-contact system that uses radio-frequency electromagnetic fields to transfer data from a tag attached to an object, for the purposes of automatic identification and tracking.

RISC Reduced Instruction Set Computing, a type of Instruction Set Architecture generally characterised by a simple and general design rather than having a large set of instructions, many of which are complex or specialized.

RISC-V	An open RISC Instruction Set Architecture, developed at UC Berkeley
RNA	Ribonucleic Acid
ROI	Return on Investment
ROM	Read-Only Memory
RSA	A cryptographic algorithm, named after its inventors Ron Rivest, Adi Shamir and Len Adleman
RTL	Register-transfer level
SAE	Society of Automotive Engineers
SAN	Storage area network, a dedicated network that connects a set of storage devices that are able to share low-level data with each other.
SCRUM	Scrum is an agile framework for developing, delivering, and sustaining complex software products.
Secure multi-party computation	A computation in which several parties compute the result of a function on different inputs together, while at the same time keeping these different inputs secret from each other.
SEM	Scanning Electron Microscope
SGX	SGX Software Guard Extensions, an extension to Intel's x86 ISA
Si	Silicon
SIM	Subscriber Identity Module
SIMD	Single Instruction, Multiple Data
SME	Small and Medium-sized Enterprise, a company of up to 250 employees.
SoC	A System on Chip refers to integrating all components required for the operation of an entire system, such as processors, memory, and radio, on a single chip.
SOI	Silicon-on-Insulator
SOTIF	Safety Of The Intended Functionality (ISO/PAS 21448)
SPARC	Scalable Processor Architecture, a RISC ISA developed by Sun Microsystems.
SPARK	A formally defined computer programming language based on the Ada programming language.
Spike computations	A programming model where large collections of devices, modelled after neurons, interact through the transmission of spike signals.
SRAM	Static RAM
SSD	Solid State Disk
STDP	Spike-Timing-Dependent Plasticity is a biological process that adjusts the strength of connections between neurons in the brain. The process adjusts the connection strengths based on the relative timing of a particular neuron's input and output action potentials (or spikes).
STEM	Science, Technology, Engineering and Mathematics
Streaming analytics	Streaming analytics, also called event stream processing, is the analysis of large, in-motion data called event streams. The growing number of connected devices – the Internet of Things – will exponentially increase the volume of events that surround business activity. The more data is generated, the greater the potential benefits from streaming analytics.
SVM	Support Vector Machine
SWOT	Strengths, Weaknesses, Opportunities, Threats
TCB	Trusted Computing Base
TCI	TransContinuum Initiative
TCO	Total Cost of Ownership
TCP	Transmission Control Protocol
TFET	Tunnel FET
TFLOPS	TeraFLOPs, 10^{12} floating-point operations per second
TFT	Thin-Film Transistor
TLS	Transport Layer Security
TOF	Time-of-Flight
TPU	Tensor Processing Unit
TRL	Technology Readiness Level
TSMC	Taiwan Semiconductor Manufacturing Company
TSV	Through Silicon Via, a (vertical) electrical interconnect that goes through a silicon die or wafer ("via" = vertical interconnect access)
TSX	Transactional Synchronization Extensions, an extension to Intel's x86 ISA
TV	Television
UC	University of California
UML	Unified Modelling Language is a general-purpose, developmental, modelling language in the field of software engineering that is intended to provide a standard way to visualize the design of a system.
URL	Uniform Resource Locator
US/USA	United States (of America)
USB	Universal Serial Bus
USD	US Dollar
UTP	Unshielded Twisted Pair
UV	Ultra Violet
VAX	Virtual Address Extension
VC	Venture Capital
VHDL	VHSIC (Very High Speed Integrated Circuit) Hardware Description Language
VLIW	Very Long Instruction Word
VLSI	Very-large-scale integration is the process of creating integrated circuits by combining thousands of transistors into a single chip.
VM	Virtual Machine
VMS	Virtual Memory System
VPA	Virtual Personal Assistant
VUCA	Volatile, Uncertain, Complex and Ambiguous
WASI	Web Assembly System Interface
WASM	Web Assembly
WIFI	Wireless Fidelity
WLAN	Wireless Local Area Network
WSN	Wireless Sensor Network
WWW	World Wide Web
XAI	Explainable Artificial Intelligence

Process

The HiPEAC Vision has been a biennial document that presents the trends that have an impact on the High Performance and Embedded Architecture and Compilation community. It evolves with this edition into a more agile document consisting of standalone, independent articles that underpin its recommendations. The articles and recommendations are based on information collected through a number of channels.

- Meetings with teachers and industrial partners at the ACACES 2019 Summer School;
- A survey circulated to all HiPEAC members, and which received 35 responses;
- Several consultation meetings:
  - CPS workshop organised by Charles Robinson, 16 January 2020, Brussels
  - Consultation meeting "How can we move towards sustainable electronics and energy efficient devices?", 10 March 2020
  - Consultation meeting on "What will software be like in 2030? What role will humans have in software programming?", 7 April 2020
  - Editorial board meeting, 16 April 2020, virtual
  - Consultation meeting on "Trustable computing. What do we need to arrive at trustworthy hardware and software?", 22 April 2020
  - Editorial board meeting, 27 April 2020, virtual
  - Consultation Meeting on "What will Open Source be like in 2030?", 28 May 2020
  - Editorial board meeting, 17 June 2020, virtual
  - Editorial board meeting, 1 July 2020, virtual
  - Editorial board meeting, 25 August 2020, virtual
  - SoS workshop organised by Charles Robinson, 10 September 2020, virtual
  - Editorial board meeting, 7 October 2020, virtual
  - Editorial board meeting, 5 November 2020, virtual
  - Editorial board meeting, 19 November 2020, virtual
  - Editorial board meeting, 27 November 2020, virtual
  - Editorial board meeting, 4 December 2020, virtual
  - Editorial board meeting, 18 December 2020, virtual;
- Exchanges with other organizations, such as ECS and ETP4HPC;
- Feedback from presentation of the HiPEAC Vision 2019 at several conferences and workshops (ISQED2019, Artemis Technology conference 2019, First Open Virtual Assistant Workshop, EFECS, HiPEAC 2020, and IoT and Edge Computing Future Directions for Europe workshop, for example) and from exchanges with a number of units of the DG Connect.

The document is called a 'Vision' because it is the result of the interpretation of the trends and directions as seen by the HiPEAC community. As HiPEAC has no direct power to enforce the recommendations, the timeline associated with the potential implementation of the recommendations is uncertain; this is why the document is not a roadmap per se.

Marc Duranton running a Technology Trends workshop at EFECS 2019

Acknowledgements

This document is based on the valuable input of HiPEAC members. The editorial board, composed of Marc Duranton (CEA), Koen De Bosschere (Ghent University), Bart Coppens (Ghent University), Christian Gamrat (CEA), Cath Roderick, Harm Munk (TNO), Thomas Hoberg (ATOS), Tullio Vardanega (University of Padua) and Olivier Zendra (INRIA), would like to thank all the authors of the individual articles, and Vicky Wandels (Ghent University), Eneko Illarramendi (Ghent University) and Jennifer Sartor (Ghent University), for their useful comments and their support. The editorial board would also particularly like to thank the team from ETP4HPC (Michael Malms, Marcin Ostasz, Peter Bauer) for the fruitful discussions and their contribution to the new structure of the Vision.

Key recommendations of the HiPEAC Vision 2021

We have come a long way from the time when a computing system consisted of one computer core programmed in one or very few programming languages. The need for more computing power on energy constrained computing platforms (from deep edge sensor node to supercomputer) first forced computer vendors to introduce multicores. More recently it has obliged them to leave behind homogeneous multicores in favour of heterogeneous multicores consisting of different kinds of accelerators that are more efficient, but more difficult to program and use efficiently.

The emergence of new workloads such as deep learning and large-scale industrial cyber-physical systems has led to a series of new challenges that are related to non-functional properties including power consumption, timing, complexity, security, safety and sustainability. The design and implementation of modern computing systems has become so complex that it exceeds the cognitive capacity of even the best computer scientists.

Therefore, HiPEAC recommends that we move towards 5S.(CPS)^2:

Cognitive Cyber and Predictive Physical System of Systems
that are
Sober, Secure, Safe, Straightforward and Sustainable.

The recommendations of this 2021 Vision are:

Cognitive: HiPEAC recommends investing in ultra-low power accelerators for AI and in the investigation of approaches that use less labelled data.

Cyber: Just as Europe set the basis for the world wide web, HiPEAC recommends that it should secure its place at the forefront of the "next web" by adding the necessary innovations and standards on top of existing technologies to meet and satisfy human needs and interests.

Predictive: HiPEAC recommends investment in digital twins and models that can be executed accurately and efficiently at the edge.

Physical: HiPEAC recommends investment in research into ways to correctly model non-functional properties and to guarantee them in the systems.

System of systems: HiPEAC recommends investment in systems of systems research and in development of tools for orchestrating large dynamic heterogeneous systems.

Sober: HiPEAC recommends investment in the development of ultra-low power computing platforms covering the complete digital continuum, and in tools allowing assessment and design of systems with explicit power constraints.

Secure: HiPEAC recommends that more investments be made in cybersecurity research, and in particular in the automated finding of security risks, in the means to automatically mitigate or remove those risks, and in the development of secure hardware and tools that can produce secure by design software and hardware.

Safe: HiPEAC recommends further investment in research and development in the methodology and design of safety-critical systems.

Straightforward: HiPEAC recommends the development of approaches that improve human productivity to design, produce and manage complex systems, including with the use of AI techniques.

Sustainable: HiPEAC recommends that Europe funds research to lower embodied energy of devices, and encourage the extension of the lifetime of devices by upgrading, reusing and repairing them. Europe should have the ambition to lead in the design of sustainable electronics.

Open source: HiPEAC recommends investment in free open source digital commodities in the form of an EU platform that can be used by third parties to create value. This requires the establishment of a dedicated European institute or hub to support open source.

New computing technologies: HiPEAC recommends that Europe continues to investigate emerging technologies, not with a view to them directly replacing silicon technology, but to complementing it. This research should be wide-ranging, and include new ways to code information (e.g. using "qbits", or temporal coding like with "spiking" neuromorphic architectures, or using physics phenomenon – like light – as analog computing approaches), as well as methods to efficiently integrate these approaches as "accelerators" in a silicon technology-based system, on both the hardware and software sides.

Moonshot programme "Guardian Angels": HiPEAC recommends the creation of a "Guardian Angels" moon-shot programme that encompasses all the 5S.(CPS)^2 technologies into a system that will serve European citizens and companies.

International competence centre: HiPEAC recommends the creation of a well-funded European competence centre in computing so that Europe is able to retain and attract top talent, to set its own ambitious research agenda, and to defend its position as a scientific powerhouse in computing.

Digital infrastructure: HiPEAC recommends European investment in a state-of-the-art digital infrastructure.

Training: HiPEAC recommends European investment in education and lifelong learning, with a view to producing more highly-skilled computer scientists to advance the state of the art in 5S.(CPS)^2 in all its aspects.

Innovation culture: HiPEAC recommends greater European investment in the creation of an innovation culture at all levels (education, society, industry) to stay competitive and to help attract venture capital for startups and scaleups.

European values and digital ethics: HiPEAC recommends that every individual be guaranteed the same protection in the cyber world as they are in the physical world. Digital ethics should become part of business practice.