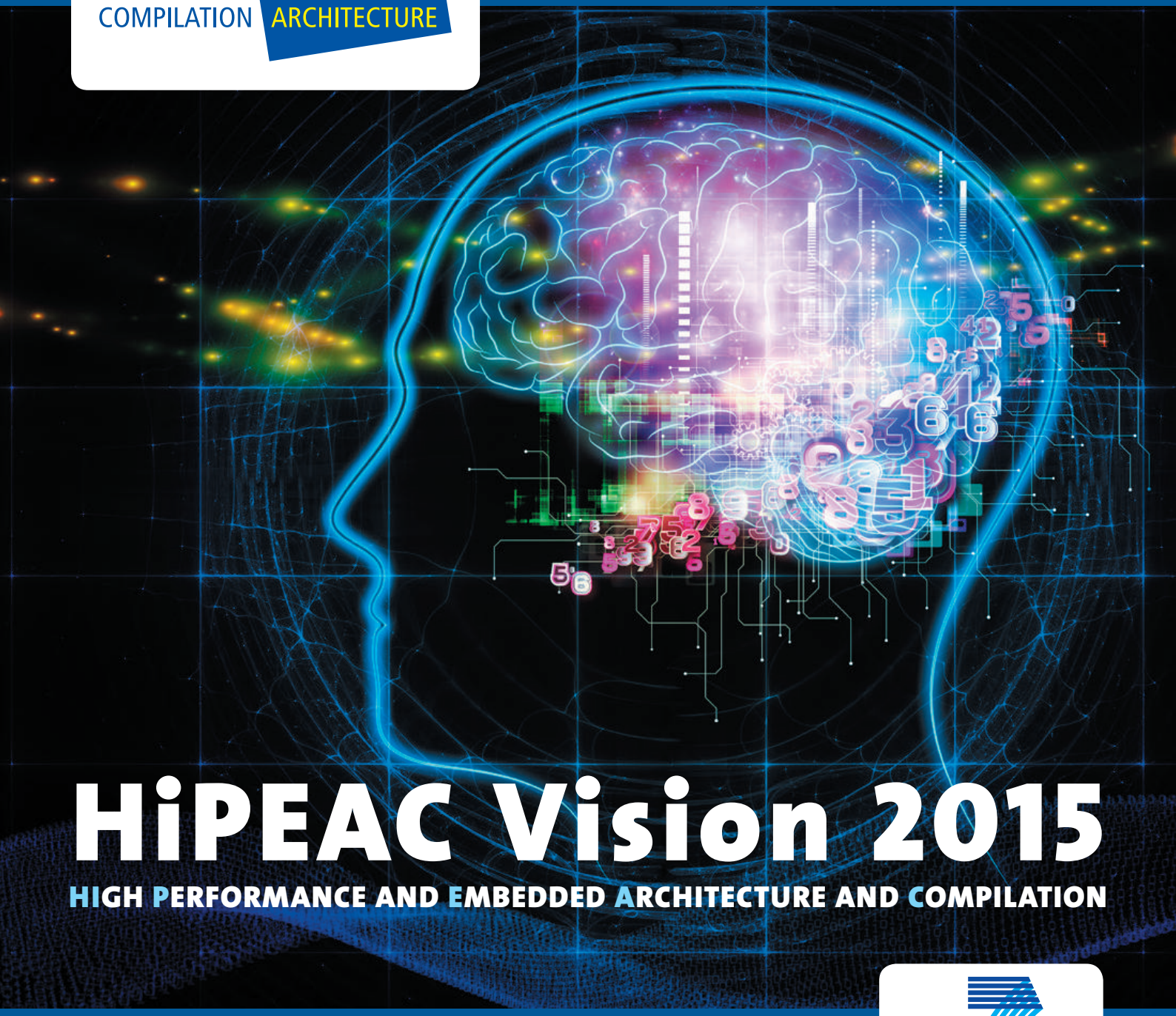
HiPEAC
COMPILATION ARCHITECTURE



HiPEAC Vision 2015

HIGH PERFORMANCE AND EMBEDDED ARCHITECTURE AND COMPILATION

Editorial board: Marc Duranton, Koen De Bosschere, Albert Cohen, Jonas Maebe, Harm Munk

SEVENTH FRAMEWORK PROGRAMME

This document was produced as a deliverable of the FP7 HiPEAC Network of Excellence under grant agreement 287759.
January 2015

The editorial board is indebted to Dr Max Lemke and to Dr Sandro D'Elia of the Complex Systems and Advanced Computing unit of the Directorate-General for Communication Networks, Content and Technology of the European Commission for their active support to this work.

CONTENTS

EXECUTIVE SUMMARY

If there is one thing that characterizes the development of information technology, it is the unrelenting, always accelerating change. Computing systems keep pervading society deeper and deeper, and have acquired a strong foothold in everyday life. The smartphone and tablet revolution, which was enabled by the wide availability of Internet access, has made life without information technology unthinkable. With the Internet of Things on the doorstep, we are facing the challenge of the exponential growth of the amount of online data. With the advent of Cyber-Physical Systems, computing systems are not just monitoring, but actively controlling parts of the world around us. Finding one's way in the growing amount of data and determining a course of action requires cognitive systems that can make decisions based on available information much faster than humans can.

Yet, some of the key technologies that have fuelled the exponential growth of computing systems seem to be heading for a wall dictated by physical limits, creating serious technological challenges for computing power, storage capacity and communication speed: we are nearing the end of the information technology world as we know it. Society has become almost completely dependent on information technology. In reaction to several breaches of trust and integrity, society is increasingly concerned about privacy and security. To address these issues, we need a holistic view of the complete system, both of hardware and software. Designing trustworthy systems will require better security tools. Europe, with its strong presence in software security and with one leading processor architecture is well positioned to play a prominent role here.

Information technology is accelerating the replacement of jobs, particularly middle class ones, by computers and robots. Depending on society's ability to create new jobs as has happened in the past, this may either result in a shift to different jobs or in a growing unemployment rate. However, as compared to the past, the pace of technological development is still accelerating, allowing society less and less time to adapt. This will require considerable flexibility from the educational system, policy makers, legal frameworks and unions, to name just a few.

The human race is increasingly depleting Earth's resources. The production of ICT devices, for example, has an almost insatiable hunger for rare elements. On the other hand, information technology also, increasingly, has the ability to optimize the usage of scarce resources, ranging from real-time tailoring of energy production to consumption, to recycling and the reusing of equipment. Further unleashing the power of information technology on these processes may become possible, again, through the Internet of Things.

Finally, Cyber-Physical Systems and robotic systems, whereby humans, computers and the physical world interact, will enable assisting or replacing human actors in dangerous, tedious, or dirty jobs. These systems still need a considerable amount of research and development, might rely on cloud resources, and will necessitate measures to protect the privacy and security of their users.

CHALLENGES AND ACTIONS

DEPENDABILITY, SECURITY

Cyber-Physical Systems and the Internet of Things require the highest possible levels of security, safety and reliability for their adoption. In particular, security has to become one of the primary design features of whole systems. Systems have to be dependable by design.

MANAGING SYSTEM COMPLEXITY

In information science, we are entering the era of systems of systems, with the accompanying exponential growth in complexity. We need new system paradigms, such as reactive systems and feedback systems to manage this increase in overall complexity.

POWER AND ENERGY EFFICIENCY

We need to overcome energy as the limiting factor for increasing system performance. Instead of over-designing systems and targeting best effort, we need to design methodologies and tools that allow for systems to scale to the desired Quality of Service. We need to investigate methods that allow for information flow and optimization across system layers. Instead of being locked-in silicon-based Von-Neumann architectures, we need to investigate other architectures based on alternative technologies, such as neuromorphic architectures, Bayesian systems, etc.

ENTANGLEMENT BETWEEN THE PHYSICAL AND VIRTUAL WORLD

Information systems will sense, monitor, and even control the physical world via Cyber-Physical Systems and the Internet of Things. Dealing with large amounts of unstructured data will require new approaches to bridge the gap between the physical world and computer systems. As these systems also exercise control over the physical world, safety is a fundamental concern and becomes a primary design constraint.

MULTIDISCIPLINARY

Information systems have moved out of the realm of computer science and moved into the hands of experts in other disciplines. Yet, many systems still need to be adapted to the human-in-the-loop paradigm, letting knowledgeable experts concentrate on the what-to-solve instead of the how-to-solve. This requires a multidisciplinary approach to application development.

HOLISTIC

Systems are becoming distributed systems-of-systems of heterogeneous architectures with a wide variety of applications and communication infrastructure. To make effective and especially efficient use of such complex architectures, we have to develop holistic approaches to system design, allowing for non-functional information flow to enable (dynamic) optimizations.

TECHNOLOGICAL EVOLUTION

Several technologies that have fuelled the exponential performance growth of computer systems in the past decades are facing physical walls. We need to develop new, disruptive technologies for efficient data storage, for information processing, and for communication, to sustain this exponential growth. They will, however, need quite some time to become mainstream, so we need to prepare for a period with performance stagnation.

ON POLICY RECOMMENDATIONS

RESEARCH PROJECT CONTENT

Experience from participants of past and current European projects indicates a concern over the protection of intellectual property rights in projects. Based on those concerns, we recommend careful examination of IP rights in projects, especially in the light of the business models of the project participants. Additionally, IP rights containment is less problematic with a limited number of project participants. Therefore we recommend that some projects can have a limited number of participants, possibly even bilateral concerning IP rights.

Many participants recommend focusing projects on topics that seek to develop basic, non-discriminative technology, or projects that pursue technology development that is too expensive to be carried out by a single company.

PROJECT IMPACT

The HiPEAC community is typically involved in development at the bottom of the pyramid, enabling the application of devices and technologies in new end products. Many factors determine the success of these end products, and it is for that reason often difficult or even impossible to assess the impact at the project's end.

As the instruments and the organization of H2020 deviate considerably from the previous Framework Program, the community is asking for increased guidance to optimize the use of the H2020 instruments and funding.

PROJECT VISIBILITY

Compared to physics and astronomy, European projects in the HiPEAC community are virtually invisible to the general public. This, again, is a direct consequence of the position of the very specialized subjects of these projects near the base of the product development pyramid. To change this situation, we recommend instituting contests, both on a large and on a small scale. These contests should revolve around inspiring, high-visibility subjects, where the large-scale contests typically involve conglomerates of participants, and the small-scale contests would attract smaller groups of people.

In addition, we recommend the use of pre-commercial procurement and requirements regulations on consumer products to increase the visibility and use of the results of European projects.

INTRODUCTION

THE END OF THE WORLD AS WE KNOW IT

This edition of the "HiPEAC vision" is special for two reasons:

- It is the last one in its current form, as the HiPEAC Network of Excellence will end in December 2015 and the "Network of Excellence" instrument is not part of Horizon 2020.
- For the first time, we have noticed that the community really *starts looking for disruptive solutions, and that incrementally improving current technologies is considered inadequate to address the challenges that the computing community faces*.

This view is shared by industry, at least in the US: Hewlett-Packard is announcing "the Machine", a new architecture using small cores, photonics and memristors *[Machine]*, while IBM recently announced its brain-inspired computer and is investing $3B in new technologies. According to the IBM announcement *[IBMpr]*, "*the $3 billion over the next 5 years will be used in two broad research and early stage development programs to push the limits of chip technology needed to meet the emerging demands of cloud computing and Big Data systems... Beyond 7 nanometers, the challenges dramatically increase, requiring a new kind of material to power systems of the future, and new computing platforms to solve problems that are unsolvable or difficult to solve today. Potential alternatives include new materials such as carbon nanotubes, and non-traditional computational approaches such as neuromorphic computing, cognitive computing, machine learning techniques, and the science behind quantum computing.*"

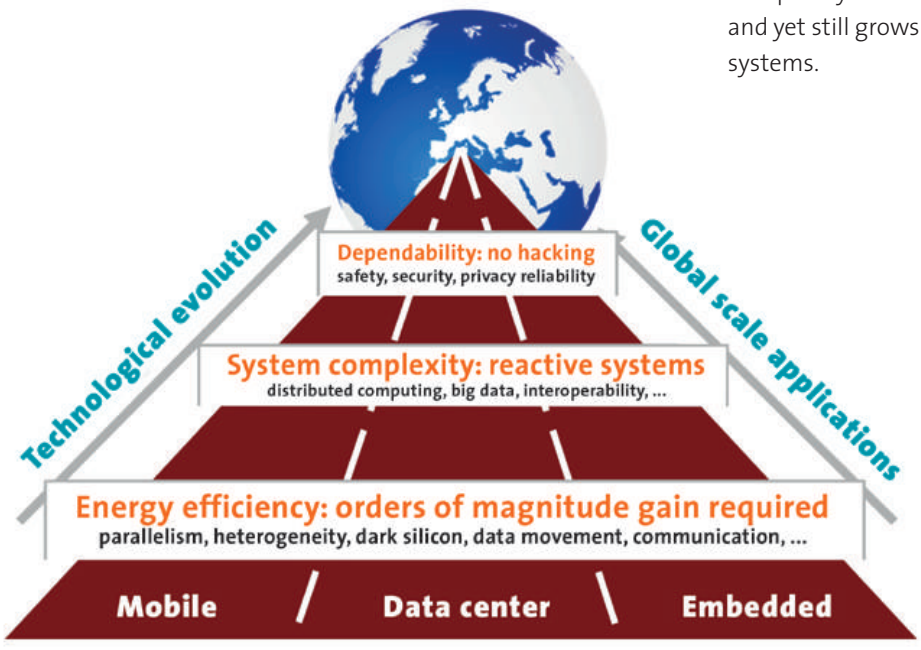
Google and Facebook are hiring neural network specialists and Google has bought several companies in the fields of robotics and artificial intelligence. Even Qualcomm is interested in robotics *[Qualcomm]*. Similarly, Dropbox bought KBVT, a company specialized in image recognition.

Even if there is a clear trend towards looking for potentially disruptive solutions, they do not offer short-term solutions. First, there is no consensus regarding which technology is most promising, and second, they will take years to become mainstream. We nevertheless have to explore various potential solutions and (industrial) champions to push them into reality.

Secondly, the major challenges identified by the previous vision document are still present, and they have become increasingly important. Even if this document is self-contained, we encourage the reader to browse the previous HiPEAC vision document *[Roadm13]* for the rationale of some challenges.

In a nutshell, the main message of the last HiPEAC vision document was that current technologies will prevail in the short term and will be improved while trying to cope with the challenges that are shared by embedded, mobile, server and HPC domains. These include, notably:

- Energy and power dissipation: the newest technology nodes made things even worse;
- Dependability, which affects security, safety and privacy, is a major concern. The revelations by E. Snowden clearly reinforced the need to design systems for security and privacy;
- Complexity is reaching a level where it is nearly unmanageable, and yet still grows due to applications that build on systems of systems.



The main topics of the HiPEAC vision 2012 are still valid, only amplified

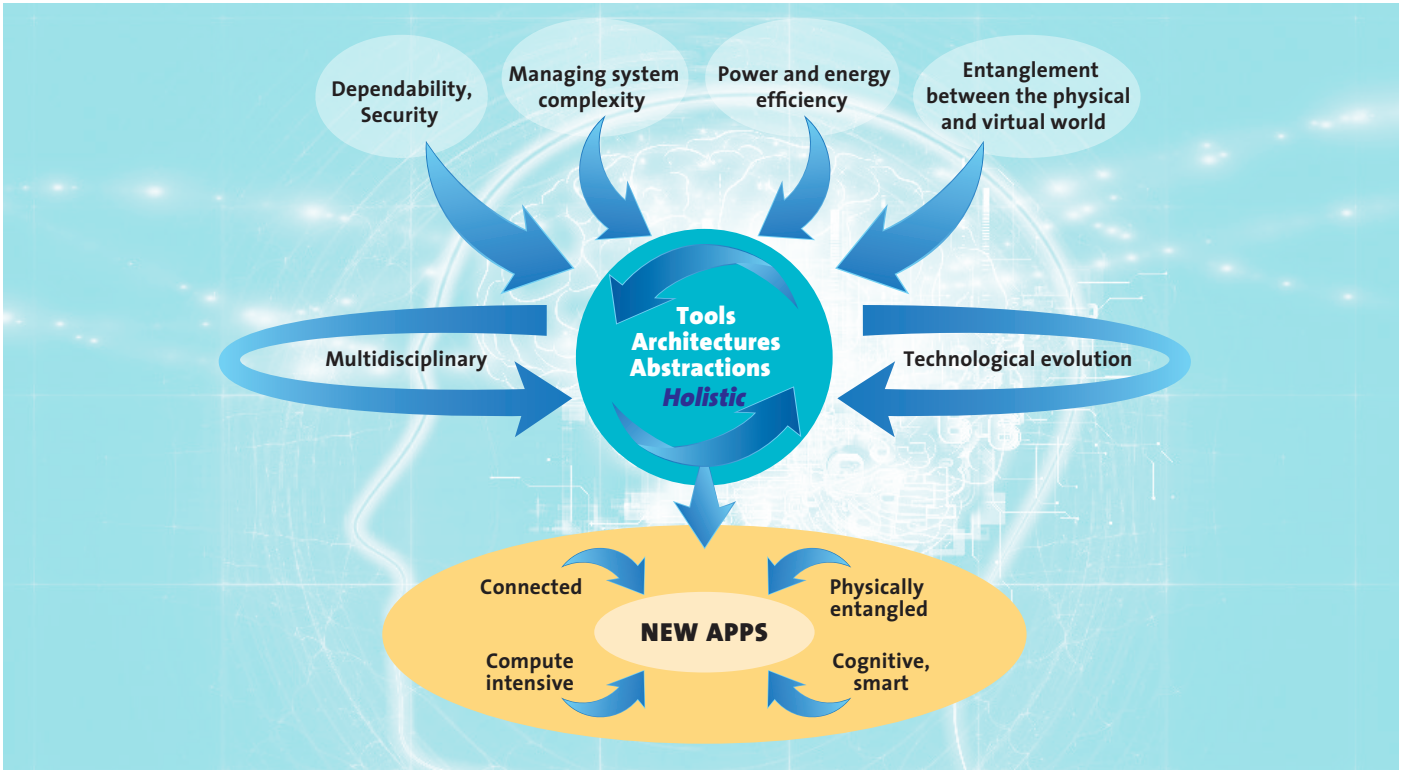
ABOUT CHALLENGES, CONTEXT AND ACTIONS

This section lists the recommendations of the HiPEAC Vision 2015. They are aimed at a wide network, including: the HiPEAC community, the European commission, Europe, academics and industry. Some solutions contribute to addressing multiple challenges and hence may be mentioned more than once. Also note that the current challenges and recommendations do not replace those of the previous HiPEAC vision document, but instead *complement* them.

The ultimate goal of our community is to develop Tools, Architectures and Abstractions to build the next generation of killer applications. These applications will be characterized by four elements:

- They will be compute-intensive, i.e. they will require efficient hardware and software components, irrespective of their application domain: embedded, mobile or data center.
- They will be connected to other systems, wired or wireless, either always or intermittently online. In many cases they will be globally interconnected via the Internet.
- They will be entangled physically, which means that they will not only be able to observe the physical environment that they are operating in, but also to control it. They will become part of our environment.
- They will be smart, able to interpret data from the physical world even if that data is noisy, incomplete, analog, remote, etc.

All future killer applications will have these four characteristics, albeit not all to the same extent.



Our community has to provide the next generation of tools, architectures and abstractions required to build these killer applications efficiently and correctly. Building them will require taking into account several non-functional requirements such as energy, time, security and reliability. New computational models will be needed in some cases, such as neuromorphic architectures, Bayesian computing, pseudo-quantum computing, and statistical/probabilistic computing.

Potential ways to tackle these challenges are to:

- Develop approaches (= architectures, tools, abstractions) that take into account **non-functional information** (such as temperature, energy consumption and management, wear, ageing, errors) at all levels of the applications, making it possible to make decisions throughout all levels of the stack rather than only at the lower levels. This should be performed in a way that ensures a high level of interoperability (thereby developing – de-facto – standards) and security (keeping in mind potential misuse of this information).
- Develop methodologies that enable **combining multiple computation paradigms** in a single system (e.g. Von-Neumann, streaming, distributed, reactive, neuromorphic, Bayesian computing, pseudo-quantum computing, statistical/probabilistic computing). In particular, these methodologies must ensure quality, testability and reliability of the results of these systems.
- Further develop the path of **reactive systems**, in particular by applying knowledge from the cybernetics domain concerning the use of **feedback loops** to stabilize dynamic, complex systems
- Develop formalisms, methodologies and tools for "*adequate precision computing*", or more generally to deal with a "**desired level of Quality of Service**": tools that take into account power, security, and time, which use the design-by-contract paradigm, which can reduce over-specification, and which can be used both with predictable and with reactive systems.
- Develop approaches that enable domain specialists to express only **what** needs to be done, while tools and computers take care of **how** to transform this knowledge into an efficient computing system representation and execution.
- Further develop design space exploration methods and tools.
- Develop approaches to validate complex systems composed of black or grey box components.

Due to the complexity and the large variety of problems, a single tool from one provider will certainly not solve the problem. Instead, we need a coherent framework of interoperable and complementary tools from multiple sources. For example, Eclipse *[Eclipse]* has created such an ecosystem.

In order to provide these tools, architectures and abstractions, we will have to deal with seven challenges, four primary and three secondary.

2.1. PRIMARY CHALLENGES

DEPENDABILITY BY DESIGN

Challenge:

Design systems with dependability in mind starting from the earliest stages of application development, up to and including the maintenance stage. We call this challenge "**Dependability by design**". It applies from top to bottom to both software and hardware and to all aspects of dependability: safety, reliability, security ... We especially wish to emphasise that security is a high priority concern that is commonly overlooked during all stages of system design and implementation.

Solutions:

- Create user awareness regarding security, educate users and developers.
- Make the complete system transparent and **auditable**: use "open" cores, "open" software.
- Incorporate hardware support for security in all platforms, even in micro-controllers.
- Create hardware and software to support **data processing in untrusted environments** (e.g. developing homomorphic encryption schemes).
- Add support for expressing security concerns or guarantees in mainstream programming languages.
- Develop tools to automatically protect code and to check for security leaks.
- Improve low power advanced encryption schemes and (co-) processors.
- Use multiple protection domains, sandboxing, virtualisation.
- Develop data anonymisers.

Remarks:

- Secure methods are a necessary condition for protecting privacy. A code of conduct about the use of private data should be recognised and enforced.
- Security is an element of guaranteed functional correctness.
- There can be no safety or real-time guarantees without security.
- Encryption could be legally enforced for certain services on the Internet and for certain kinds of applications.

MANAGING SYSTEM COMPLEXITY

Challenge:

Manage the complexity of systems that are increasing in size, distribution, parallelism, compute power, communication, connectivity, and amount of processed data.

Solutions:

- Develop new advanced Models of Computation, Abstractions, and Computing Paradigms to solve the software crisis.
- Explore the path of **reactive systems**, in particular by applying knowledge from the cybernetics domain concerning the use of **feedback loops** to stabilise dynamic complex system.
- Develop tools to help developers cope with the complexity of systems, especially with non-functional requirements (timing, power, reliability ...).
- Further develop design space exploration tools that help selecting an optimum in a multi-dimensional solution space.
- Develop tools and approaches to validate complex systems composed of black or grey box components.
- Develop tools to design and test the next generation of hardware platforms, accelerators, 2.5D and 3D integrated systems, trusted modules ...
- Enable expressing the concurrency in an application (the "what") to improve mapping applications to parallel or distributed systems: compilers and tools are becoming better than humans (at least in specific domains) at these kinds of transformations (the "how"). This could lead to more declarative programming as opposed to imperative programming.

Remarks:

- Developing reactive systems will require a new set of advanced tools, development frameworks, and hardware support.
- Standardisation is an important element of solving complexity. Europe should make more efforts to be part of standardisation efforts in this area.
- When designing advanced Models of Computation, Abstractions and Computing Paradigms, the huge body of legacy systems must also be taken into account.

ENERGY EFFICIENCY

Challenge:

Energy is currently *the* limiting factor for performance. If we want to further increase the performance of systems, we need to find ways to reduce their energy consumption. E.g., sensors that are part of the Internet-of-Things may have to operate by scavenging their power from the environment, significantly limiting their maximum power consumption.

Solutions:

- Move away from **over-specifications** and best effort approaches: adapt the hardware to the actually required QoS. Make system performance scale to demand, develop systems that make trade-offs between QoS, precision, required amount of data, energy.
- Develop formalisms, methodologies and tools for *"adequate precision computing"*, or more generally to deal with **"desired level of Quality of Service"**: tools that take into account power, security, time, which use the design-by-contract paradigm, which can reduce over-specifications, and which can work both on predictable and on reactive systems.
- Develop accelerators for specific tasks, but not limited to number crunching: for machine learning, for cognitive computing, for data mining and other data center workloads. These accelerators must be accompanied by associated formalisms (languages) and tools so that they can be used easily and efficiently.
- Ensure interoperability not only across systems, but also enable **cross layer information flows** within and between virtual systems and APIs. The resulting holistic overview will enable global system and cross-system level optimisation schemes, especially for managing energy.
- Develop novel, energy-efficient, **non-Von-Neumann** architectures (neuromorphic architectures, Bayesian, stochastic). Their efficient implementation may require the application of non-Si technology.
- Exploit new low power/persistent memory technologies to build power-efficient memory hierarchies, or to revisit the memory hierarchy.
- Develop system architectures that limit data movement by processing the data where it is generated/stored as much as possible. This is called "Near-Data Computing", whereby computing resources are deployed in the vicinity of the big data.
- Create an ecosystem around 2.5D design with "chiplets" to stimulate the development of heterogeneous, flexible, energy-efficient system architectures.

Remarks:

- Optimisation should not only concern computing engines, but computing and communication. Having local processing will increase the energy for the compute part, but will reduce the amount of data to be transmitted.

ENTANGLEMENT BETWEEN THE PHYSICAL AND VIRTUAL WORLD

Challenge:

The virtual, digital world and the real, physical world become entangled via the Internet-of-Things and in Cyber-Physical Systems (CPS). An increasing number of computers directly observes the physical world, and also controls that world. The challenge is to correctly observe the physical world, make decisions and react appropriately. This requires an interdisciplinary approach.

Solutions:

- **Cognitive computing** bridges the gap between the natural, unstructured, non-formalised analog world and the digital computing realm. New approaches for computing systems will be required to encompass the requirements of cognitive systems.
- The design of CPS requires advanced knowledge of mechanical, chemical, biological, and control processes. This requires interdisciplinary teams or computing researchers with a broader training.
- Safety is a very important design criterion for systems that interact with the physical world (autopilots, pacemakers, medical scanners ...). While traditional software programs cause harm when they stop working, crashing CPS systems can result in disasters. **Safety by construction** should be mandatory in such situations.
- Develop tools to help developers cope with the complexity of systems, especially with non-functional requirements (timing, power, reliability ...).

Remarks:

- CPS that make autonomous decisions like autopilots and, by extension, also the cognitive systems they rely on, will require legal and ethical frameworks in which they can make decisions. Technical solutions that can enforce these limits will be required once society lays down the rules.
- Current CPS interact with their environment, but 3D printers can even create new objects in the physical world that are conceived in cyberspace.

2.2. SECONDARY CHALLENGES

MULTIDISCIPLINARY

Challenge:

Future applications will be so complex (compute intensive, connected, sensorial, cognitive) that they will require a multidisciplinary approach. Some components will not even be IT-components (human body, combustion or electric engine, part of a factory, robot ...). This will require a multidisciplinary view incorporating aspects from physics, medicine, biology and psychology. A challenge is to open up the IT tools to the experts in these other domains. For example, data analytics is the domain of statisticians as opposed to computer scientists, pace makers are programmed by cardiologists, and so on. We need direct, close collaboration between experts from different domains to design the appropriate hardware and software.

Solutions:

- Make information science tools easier to use by people that are not IT-specialists. One approach could be using high-level languages with large libraries such as Python and Matlab instead of C++.
- Develop approaches that enable domain specialists to express only **what** needs to be done, while tools and computers take care of **how** to transform this knowledge into an efficient computing system representation.
- Take the **human factor** into account at all levels: from interfaces to reaction time, from what is really required from a human point of view to what is acceptable in terms of cost, lifetime, reliability, safety, privacy.
- Make computer science education more multidisciplinary, exposing students to core knowledge from other science disciplines in the context of their computer science education: biology, physics, basic engineering ... and design.

Remarks:

- One possible explanation for the success of vertical companies (e.g. Apple, Samsung, Google) is that their employees have many different backgrounds (not just computer science) due to the large scope of activities of those companies, and they can leverage this broad knowledge base while working towards common goal.
- The success of "interpreted" languages such as Python, Mathlab is certainly due to their immediate result, without the edit-compile-execute-debug cycle and their adaptation to specific domains by easy addition of libraries.

HOLISTIC

Challenge:

Future applications will be distributed, requiring different and multiple, interconnected computing engines, using different software and hardware stacks. As a result, their optimisation will require a holistic approach. Every component will only be a small part of the solution, and even if each component is optimised individually their combination will be not operate optimally because local optimisations may impose penalties on other components. Additionally, the extra compatibility layers and translations required to make all components interoperable will also have their cost. Therefore, we need close collaboration to design the appropriate hardware and software, and a holistic view of the complete system in order to make it efficient. Due to the combinatorial explosion of possible combinations of different structures and devices, only dynamic real-time adaptation and optimisation will be have a chance at succeeding.

Solutions:

- Develop approaches (= architectures, tools, abstractions) that take into account **non-functional information** (such as temperature, energy consumption and management, wear, ageing, errors) at all levels of the applications, allowing for making decisions throughout all levels of the stack rather than only at the lower levels. This should be performed ensuring a high level of interoperability (thereby developing – de-facto – standards) and security (keeping in mind potential misuse of this information). In many cases, a **dynamic** run-time approach may be the only viable approach.
- Further develop design space exploration tools that allow mapping the holistic complexity of a system onto simple decisions.
- Develop tools and approaches to validate complex systems composed of black or grey-box components (hardware and software).

Remarks:

- Creating a European ecosystem in which the participants can complement each other's technologies and solutions will require a strong willingness to break out of the existing "ivory tower" environments.

TECHNOLOGICAL EVOLUTION

Challenge:

Information systems technology hardware is tightly coupled to exponentially growing aspects, such as integration density (Moore's law), communication bandwidth (Nielsen's law), storage (Kryder's Law), and also to a decreasing device cost (transistors). These "laws" are running into physical limits that slow down the expected performance increases. The challenge is to break through these walls. Fortunately, new technologies can help us overcome the upcoming roadblocks that hamper existing technologies. They will, however, need significant development before they can become mainstream, which may result in a period of performance stagnation.

Solutions:

- For new storage technologies: new non-volatile memories.
- For new compute technologies: carbon nanotubes, graphene. For some applications: neural networks based on new devices ("synapstors") ...
- For communication technologies: photonics and 5G.
- To cope with the increasing cost of developing ASICs with the latest technology node: assembling chiplets using various technologies on an interposer (active or passive, using silicon or organic materials). This may allow for design diversity in order to reduce costs, by reserving the most advanced nodes for the high performance parts of the systems (compute chiplets).
- Another approach to keep design diversity are new (coarse grain) reconfigurable fabrics that take advantage of new technologies.

Remarks:

- New technologies will have a drastic impact on system architecture. For example, the availability of small and fast non-volatile memories could change the memory hierarchy. New reconfigurable devices using mixed technologies (silicon and non-volatile) could reach higher performance and thereby become suitable for wider markets. Even if the technologies are immature, computer scientists and architects should already analyse the impact they may have on existing system architecture.

MATURITY LEVEL

All of the above-mentioned recommendations should be followed as soon as possible, but their current maturity level is quite diverse. Below is a subjective classification of their maturity, indicating whether they may become mainstream in the short (3 years), mid (6 years) or long (after 2020) term.

When will the following recommendations become mainstream (in short term (S), mid term (M) or long term (L))?

Challenge 1: dependability

S Solution 1: Create user awareness on security, educate users
M Solution 2: Make the complete system transparent and auditable
M Solution 3: Create hardware support for security in all platforms
M Solution 4: Create hardware and software to support data processing in untrusted environments
L Solution 5: Create support for security in mainstream programming languages
M Solution 6: Develop tools to automatically protect code and to check for security leaks
S Solution 7: Improve low power advanced encryption schemes and (co-)processors
S Solution 8: Use multiple protection domains, sandboxing, virtualization
M Solution 9: Develop data anonymizers

Challenge 2: Managing system complexity

L Solution 1: Develop new advanced Models of Computation, Abstractions, and Computing Paradigms
M Solution 2: Explore the path of reactive systems and learn from cybernetics using feedback
M Solution 3: Develop tools to assist developers cope with the complexity of systems
M Solution 4: Further develop design space exploration tools that help selecting an optimum on a multi-dimensional solution space
M Solution 5: Develop tools and approaches to validate complex systems composed of black or grey boxes components
S Solution 6: Develop tools to design and test the next generation of hardware platforms, accelerators...
M Solution 7: Expressing the concurrency in an application (the "what") is also a good approach for mapping applications to parallel or distributed systems

Challenge 3: Energy Efficiency

M Solution 1: Move away from the over-specifications and best effort approach
M Solution 2: Develop formalisms, methodologies, and tools for adequate precision computing or at large coping with the desired level of Quality of Service
S Solution 3: Develop accelerators for specific tasks
L Solution 4: Ensure interoperability between systems, and cross layer information flow in virtual systems and APIs
L Solution 5: Exploit new low power/persistent memory technologies to build power efficient memory hierarchies or revisit the memory hierarchy
M Solution 6: Develop system architectures that limit data movement by processing the data where it is generated/stored as much as possible
M Solution 7: Create an ecosystem around 2.5D design with "chiplets"

Challenge 4: Entanglement between the physical and virtual world

M Solution 1: Cognitive computing bridges the gap between the natural, unstructured, non-formalized analog world and the digital computing realm
S Solution 2: The design of Cyber-Physical Systems requires interdisciplinary teams or computing researchers with a broader training
L Solution 3: Safety needs to be guaranteed by construction, as a key design criterion for systems that control the physical world
L Solution 4: Reinforce the effort on tool development, and on architectures to efficiently cope with the non-functional requirements, especially time

Challenge 5: Multidisciplinary research

L Solution 1: Make programmable information science tools easier to use by people that are not IT-specialists
L Solution 2: Let the specialist only express what needs to be done, not how
S Solution 3: Take the human factor into account at all levels
M Solution 4: Make computer science education more multidisciplinary

Challenge 6: Holistic approaches

M Solution 1: Develop approaches that take into account non-functional information at all levels of the applications
L Solution 2: Further develop design space exploration tools that allow mapping the holistic complexity of a system onto simple decisions
M Solution 3: Develop tools and approaches to validate complex systems composed of black or grey-box components

Challenge 7: Technological evolution

S Solution 1: For new storage technologies: new non-volatile memories
M Solution 2: For new compute technologies: carbon nanotubes, graphene; neural networks based on new devices
S Solution 3: For new communication technologies: photonics
S Solution 4: Assembling chiplets using various technologies on an interposer
M Solution 5: Coarse grain reconfigurable fabrics that take advantage of new technologies

2.3. POLICY RECOMMENDATIONS

ON THE RESEARCH CONTENT OF EU PROJECTS

Participants from medium to large companies in European projects indicate that they join EU projects either to evaluate innovations they would not evaluate otherwise, or to enlarge their ecosystem (by cooperating with new partners), *but not to develop core-differentiating technologies*. This is understandable, because core and innovative technologies enable companies to differentiate themselves and hence are generally not developed in an open, multi-partner structure. They only consider it advantageous to participate in projects to develop infrastructure technologies that are fundamental, that require large networking effects to provide optimal value, or that are too expensive to be developed independently. The situation is different for start-ups and SMEs, but they, too, share a fear of key assets and IPR leaking. This explains why European projects seldom have immediate impact, and in the best case lead to commercial products in an indirect way.

Some conclusions can be drawn from this observation:

- Intellectual Property Rights should be carefully analysed in projects, especially regarding selecting licenses that do not conflict with the business models of the partners involved.
- Bilateral collaborations are easier for deeper collaboration than multipartite collaborations. Therefore, it might be interesting to either have projects with a limited number of partners, or to cluster projects with various sets of bilateral collaborations.
- Basic and non-discriminative technologies, and technologies that are too expensive to be developed by a single company, are good examples of topics that European projects should target.

ON THE IMPACT OF EU PROJECTS

Since EU projects in computing often result in technological improvements of tools and architectures, it is very tricky to determine their impact in terms of increased turnover, market share, etc. What is the impact of a compiler that generates 10% faster code? Will it lead to additional compiler licenses being sold? Will it trigger sales of more consumer devices containing the faster code that was produced by the compiler? Or will it simply result in energy savings due to less power-hungry hardware requirements?

For commercial tools, there is the additional difficulty that sometimes free alternatives exist and are considered "good enough" for the job. It is very hard to compete with a free alternative. The development of enabling technologies for which there is no large market (e.g. compilers, HPC tools) will always require public funding. Otherwise, this expertise will completely disappear from Europe.

Since the computing systems community produces architectures and tools for computing, as opposed to consumer products, it always depends on the success of consumer products using these technologies and tools to grow and to have economic impact. Intel became big thanks to the success of the personal computer; ARM became big thanks to the success of mobile devices. Hence, the impact of research performed in computing systems projects not only depends on the technical quality of the results, but also on many external factors that cannot be controlled by the partners of a project.

The instruments and the organisation of H2020 are quite different from the previous framework program. The community is asking for further guidance on how to make optimal use of these instruments and research funds.

ON THE VISIBILITY OF EU PROJECTS

Quite often, the general public is not aware of the research results of EU projects. One of the reasons is that computer scientists often have great difficulty explaining their technical challenges and solutions to laypeople. Compare this to the search for the Higgs Boson. Physicists are apparently able to generate public excitement about their work. We do not have a similar overarching, high level, and yet a concrete challenge in the HiPEAC domain.

A solution that was mentioned several times during the consultations is to launch contests or competitions with a clear and practical goal, similar to the DARPA challenges. This idea could be developed on either a large or a small scale:

- On a large scale, we could create "Grand Challenges" with very focused and quantifiable goals. For example, the DARPA challenges *[DARPA]* were a clear enabler to the self-driving vehicle that was still considered to be a dream at the beginning of HiPEAC (10 years ago). Google capitalised on the results to create the first, real self-driving car. Having competing teams creates a positive, fun and inspiring atmosphere leading to fast progress in high technology readiness level (TRL) domains by validating the various solutions on real world cases, rather than under laboratory conditions. At the end of the contest, the teams with the best results can join forces to merge the strong points of their solutions, shortening the time to market. Financing can be in the form of a prize, or partial help to get the prototype produced. The publicity and public involvement are also very beneficial side effects. However, it is difficult to find challenges that are feasible in a short time frame... and fun. Challenges should be progressive, with quantifiable steps, building from the results of the previous one (like how the DARPA grand challenge increased the complexity for autonomous vehicles from one challenge to the other). The solar race is another example of such a challenge.
- On a smaller scale, we could organise a set of small, very focused contests or "mini-benchmarks" to push innovation. Such challenges or contests exist for developing games. Here also, the difficulty will be to find feasible targets that motivate people, that are fun and that can deliver interesting improvements for a technology or application. A potential and fun challenge could be to design a drone that stays in the air for 24h and can carry a payload of 200g. Year after year, the challenge can be made harder: stay in the air and stream video at a particular quality, perform some local processing, carry out

surveillance tasks independently, fly in real weather conditions. Such a challenge will definitely lead to technological innovations, and inspire thousands of youngsters. It might lead to hundreds of spin-off projects.

- On the other side of the spectrum, the EU could also use pre-commercial procurement and regulations to stimulate innovation. Examples of regulations are the requirements to have universal USB plug for chargers of mobile phones, to have an e-call chip in cars, energy efficiency labelling for refrigerators and other household appliances ... Similar regulations could be made with regard to minimal safety and security in commercially sold software, data security, privacy guarantees etc. Pre-commercial procurement is trickier at the EU-level as the European Commission is not a big high-end equipment buyer. Here, local governments could play a larger role, especially because their budgets are much greater than the EU's. Potential examples are electric cars, self-driving cars, fully electric flying engine, battery-operated handheld supercomputer, low power data centers ...

RATIONALE: CHANGE KEEPS ACCELERATING

πάντα χωρεῖ καὶ οὐδὲν μένει –
Everything changes and nothing remains still
Heraclitus of Ephesus – c. 535- c. 475 BCE

The only constant we observe is change, and this change keeps accelerating. Although most of the observations from the previous roadmap still hold, some new evolutions have taken place. In this part of the roadmap we describe the evolutions in society, the market and technology that impact the computing systems community.

3.1. SOCIETY

Over the past years, we have seen a growing societal concern about privacy and security. We also observe opposite reactions to information technology: on one side, there is an increasing concern that information technology might be destroying jobs, and that the production of electronics and the business models of electronics companies are negatively impacting the environment. On the other side, there is a belief that information technology is an essential enabling technology to solve societal challenges, and that it will allow us to improve the use of natural resources such as food, water and energy. There is also enthusiastic demand for the latest smartphone or game. In some countries, people prefer to spend less on food so they can afford a (smart) phone. The rapid improvement rate in terms of features, novelties and performance from past decades seems so natural that it is now taken for granted, and society may not be ready for technological stagnation and increased costs.

MISUSE OF INFORMATION TECHNOLOGY MIGHT DESTROY OUR PRIVACY

Edward Snowden's disclosure of the activities of the NSA, GHCQ and other national intelligence agencies not only caused a lot of upheaval inside and outside the US. Internet privacy issues before May 2013 were largely ignored, or at least shrugged off. After Snowden, they came to the foreground; it was as if the Internet had lost its innocence, triggering worldwide discussions on Internet privacy issues in the ensuing months.

This so-called "Snowden effect" quickly went beyond the activities of government agencies to encompass corporate data governance as well. Before, Internet privacy discussions focused on malicious and criminal activities. In the wake of Snowden, however, plenty of examples surfaced showing not only what data are kept (passive), but also how companies deduce information from, for example, the surfing behaviour of users (active).

Also last year, the Heartbleed Bug (a serious vulnerability in the popular OpenSSL cryptographic software library) was a wakeup call for the world that software is inherently vulnerable, and that propriety as well as open source software can be attacked. With the advent of the Internet of Things, these discussions take a new twist. Having your computer broken into and personal data stolen or destroyed by a computer virus is definitely unpleasant. If a critical safety device gets hacked, it could be dangerous or lethal. However, the Internet of Things not only promises to connect computers, tablets and smartphones, but virtually every electronic device: electronic locks in our home, climate control systems, fridges, dishwashers, etc. Google is introducing a driverless car, capable of driving with no one behind the wheel, or even without a steering wheel at all, which gets part of its data from remote servers. All of this data may end up in the cloud, allowing not only the owner to access it, but potentially everyone, including people with less noble intentions. As a result, someone might check the thermostat in your home to see if anybody is in. Or someone might take over the car that is driving you home, possibly with deadly results! The truth is that almost anything that contains a processor can be hacked, especially if it is connected to a network. Without security, no guarantees can be given concerning safety, privacy, timing correctness, or even functional correctness. The BadUSB research *[BadUSB]* shows that even a simple USB device can be turned into a malware device and, once infected, computers and their USB peripherals can never be trusted again.

Oxford defines the Internet of Things (IoT) as "a proposed development of the Internet in which everyday objects have network connectivity, allowing them to send and receive data." The OWASP Internet of Things Top 10 is a project designed to help vendors who are interested in making common appliances and gadgets network/Internet accessible. The project walks through the top ten security problems that are seen with IoT devices, and how to prevent them.

The OWASP Internet of Things Top 10 - 2014 is as follows:

1. Insecure Web Interface
2. Insufficient Authentication/Authorisation
3. Insecure Network Services
4. Lack of Transport Encryption
5. Privacy Concerns
6. Insecure Cloud Interface
7. Insecure Mobile Interface
8. Insufficient Security Configurability
9. Insecure Software/Firmware
10. Poor Physical Security

From https://www.owasp.org/index.php/OWASP_Internet_of_Things_Top_Ten_Project#tab=OWASP_Internet_of_Things_Top_10_for_2014.

Controlling who will have access to the data of the Internet of Things realm is a major challenge. If users are not confident about the security and privacy of the data, the forecasted explosion of connected devices may not happen. Providers of hardware are aware of this threat, and will have to secure even their low-end micro-controllers.

Another, often underestimated, problem is the firmware upgrade process: for Internet of Things devices, this should be done automatically without help of the user (having tens of devices requesting the user to download an update will rapidly become intractable), but in a safe and secure manner from a trusted delivery source. However, these devices, which will often be based on low power/low performance micro-controllers, currently don't have the cryptographic capabilities of more powerful devices. Even in case of adequately powerful devices, such as smartphones, the majority do not run the latest version of the operating system, either due to negligence of the user, lack of support from the manufacturer, or obsolescence.

Similar problems exist in the PC domain, where Microsoft stopped supporting security patches for Windows XP when more than 25% of the computers were still running this OS *[Dailytech]*. In theory, the problem of obsolescence is less likely to occur with free software, as in that case anyone can continue maintaining it for as long as they like. That only holds if there is an active community supporting the software though, and such communities can dwindle quickly if the software's main contributors withdraw.

The Snowden effect also erodes the trust that users have in major providers of hardware and software, and the recent judgments requiring US companies to deliver private data even if they are stored offshore *[Overseasdata]* is reinforcing this effect. This could lead towards a global shift to non-US based solutions for hardware, software and services. The clandestine mass electronic surveillance data mining program called PRISM (or SIGAD US-984XN) launched in 2007 by the National Security Agency (NSA) with participation of the British GCHQ [PRISM], showed an alleged involvement of all major US companies such as Microsoft, Yahoo, Google, Facebook, Youtube, AOL, Apple. iOS devices (like others) are suspected of having backdoors [Backdoors].

The revelations also quickly erode trust in the providers of hardware and software at governmental and company levels. This generates a push towards hardware and software solutions that can be audited easily, in order to check that extra "functionality" has not been added.

Many Internet companies' business models are based on the exploitation of private data – a free service in exchange for private data. This is often attractive for users – especially the young ones – who only see the "free" aspect. The end-user should, however, be informed of the exact conditions of this "contract". Initiatives like the "Declaration of the Digital Human Rights" demonstrate the growing concern for this kind of issues (see inset).

Not only end-users are victims of this business of collecting data, even big companies can be trapped. As explained in [Android], "Google made the consumer electronics industry an offer it couldn't refuse: We'll give you an operating-system platform that lets you make an almost-as-good-as-an-iPhone, so you can make profit margins almost-as-good-as-Apple's. Android was modern and it was "free", and manufacturers could tailor it." However, now, in mature markets, it is very difficult for companies to make profits with Android phones – except Chinese companies as the market in China is still booming, or at the expense of large marketing budgets. Google continues to use the Android platform to collect more and more data.

The Forum d'Avignon proposes [ddhr] a

PRELIMINARY DECLARATION OF THE DIGITAL HUMAN RIGHTS

1. Every human being's personal data, in particular digital data, conveys information on his cultural values and private life. Personal data cannot be reduced to a commodity.
2. The reasonable exploitation of data is an opportunity for the development of research and the pursuit of the general interest. It must be governed by a universal code of ethics that protects each individual's dignity, privacy and creative works, and the diversity of opinions.
3. Everyone has the right to have respect for his dignity, private life and creative works, and shall not be discriminated against on the basis of access to his personal data and the use made thereof. No private or public entity may use personal data to manipulate access to information, freedom of opinion or democratic procedures.

4. Everyone has the right to inspect and control his personal data, including that resulting from his behaviour and objects connected to him. Everyone has the right to the confidentiality of his personal data, and to the protection of his anonymity when he so requests.
5. Any exploitation of the data or creative works of any individual requires his free, prior, informed, time-limited and reversible consent.
6. The users of personal data, whatever their level of accountability, including states, public and private authorities, companies and individuals, shall show total transparency in the collection and use of any individual's data, and shall facilitate each individual's access to his data, as well as its traceability, confidentiality and security.
7. Open research and innovation, based on the sharing, subject to consent, of any individual's anonymised data, with respect for his dignity and for cultural diversity, are in the general interest.
8. Cooperation between civil society and businesses is required to put the human being back at the center of a trustworthy society, aided by the reasonable use of disclosed and inferred personal data.

Over the coming years, more and more attention will have to be paid to Internet privacy and security, both to safeguard the democratic integrity of society and to protect economic interests. As argued above, this has to be at both the software and the hardware level. Europe, with its software security know-how (Gemalto is a global leader in digital security, Nagra is a global leader in DRM solutions, AES was designed in Europe ...), its strong tradition in software verification and the availability of the ARM architecture, has an opportunity to play a prominent role in security. It must however act quickly and strengthen this important field. *[Mann14]*

GOVERNMENTS WANT TO CONTROL THEIR (AND OTHERS) INFORMATION TECHNOLOGY

In the past, world domination could be accomplished by ruling the seas, or by having the largest or best-equipped army, or by having the most economic power. It seems that over the last decade, the balance is shifting towards controlling information and the technology that forms the backbone of the information society. Now that an increasing part of our life is happening online, it is clear that governments want to be present in the virtual world too.

Recent examples of explicit actions to control information include the ban on Twitter in several countries in the Middle-East, Russia and China, various European initiatives to instate website blocking lists, and data retention initiatives all over the world. There are also reported cases of censorship of Facebook, political pressure on credit card processors to stop accepting payments to WikiLeaks, the SWIFT and Passenger Name Records agreements with the US, the difficulty with implementing the "right to be forgotten" ... And, as mentioned earlier, US-based companies can now be forced to disclose private information about their customers, even if this information is stored outside the United States. On top of all this, governments naturally also make use of information that people publicly share about themselves, or about others (possibly without their consent or knowledge).

A more extreme case of controlling information is cyber war. Increasingly, governments are investing in cyber war technology – in order to protect their own critical digital assets from being spied on, and in several cases also to spy on other countries, or even to disrupt/destroy their infrastructure. Stuxnet is one example of this.

It is easier to control all of this information if one also controls the infrastructure and fundamental software. At this time, US-based companies are in control of the majority of the infrastructure, operating systems and information in Western Countries. China and Russia are trying to regain independence in this area. On the hardware side, a number of government-initiated foreign initiatives are already unfolding, or have even been ongoing for almost a decade. China started the development of a MIPS-based processor, the Godson, in 2001, with multi-core chips expected early 2015. China started this development to obtain independence from the US-dominated CPU-market.

Recently, the Chinese firm Hi-Silicon unveiled an ARM 64-bit based 16-core computer board targeted at datacenter applications, also showing a desire to become independent. In the wake of the Snowden disclosures, Russia announced its plans *[Baikal]* to replace all Intel/AMD-based government desktops and servers by an ARM-based architecture named Baikal to be developed in the next year, and with HPC solutions by T-Platforms. These developments serve to show that there is a growing movement away from well-established computing platforms such as Intel's and Microsoft's. The well-established platforms react by taking privacy more seriously: *[ios-8data]* and *[Androiddata]*.

No state-funded or EU-funded initiatives exist in Europe, yet. The opening up of the CPU-market is, however, an opportunity for Europe to jump in, as it clearly shows that information technology is not tightly bound to one computing platform anymore. Open architectures, where the code can be reviewed and the design audited, may play a major role in this climate. It is already the case with the ARM architecture in the embedded/mobile domain, but for example IBM with its OpenPower initiative *[OpenPower]* is also following this trend, and MiPS is now part of Imagination Technologies in UK. The RISC-V Instruction Set Architecture (ISA), from the University of California at Berkeley, is a standard open architecture that can be used by industry *[Riscv]*. Finally, the Leon (SPARC instruction set) is also an example where people have access to the netlist of the implemented core *[Gaisler]*.

INFORMATION TECHNOLOGY MIGHT EVENTUALLY DESTROY MORE JOBS THAN IT CREATES

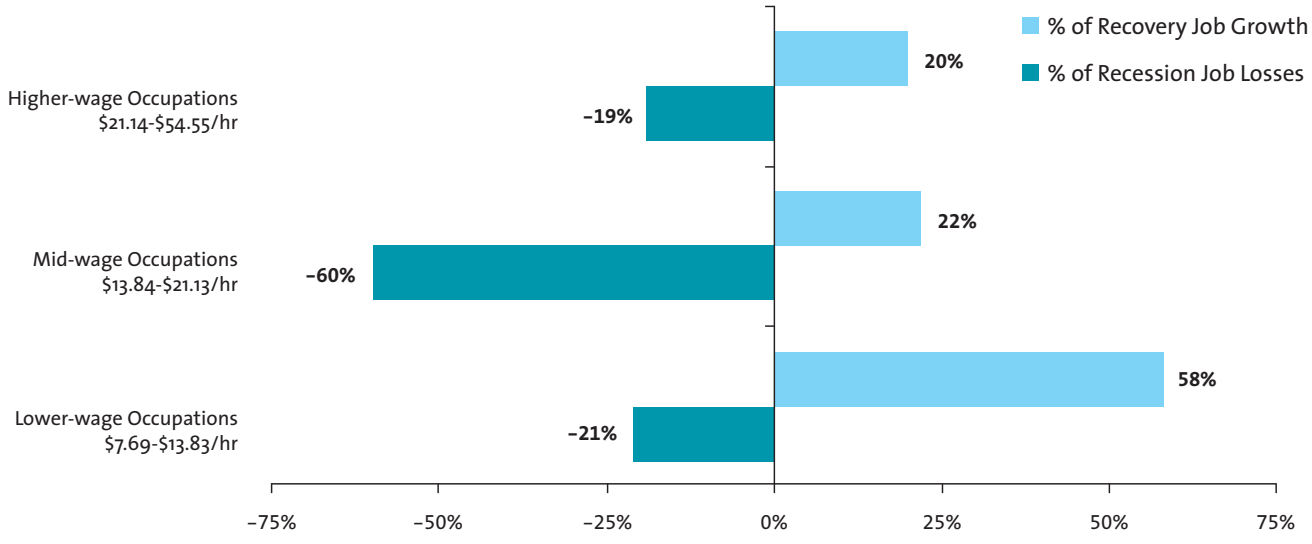
According to Ray Kurzweil, now director of engineering at Google, computers will be able to think for themselves by 2029 and the *singularity* might happen around 2045. It was Vernor Vinge who introduced the concept of the technological *singularity* in 1983. It refers to the point at which machines and technology evolve faster than can be understood by humans, and that at some point cognitive systems will become more intelligent than humans. Beyond this point, no vision or roadmap document can be drawn up anymore, at least not by humans. Some consider the fourth paradigm of science (data science, big data analytics) as evidence for this evolution: scientific discovery that would be next to impossible without computers.

The singularity would create unprecedented disruptions in society, such as massive unemployment, sometimes also called "the end of work" *[EndofWork1, EndofWork2]*. The reasoning is that by 2045, most of the work performed by humans today will be carried out by computers and robots. Non-believers claim that this has already happened a couple of times in human history (most notably during the first industrial revolution), but that society has always been able to adapt by creating new types of jobs *[Bess14]*. This statement is definitely true, and many of today's jobs did not even exist 10 or 20 years ago. The biggest difference with previous revolutions, however, is that the current evolution is much faster, and society does not seem to adapt fast enough. Society adapts at the rate of generations of people, not generations of Moore's law.

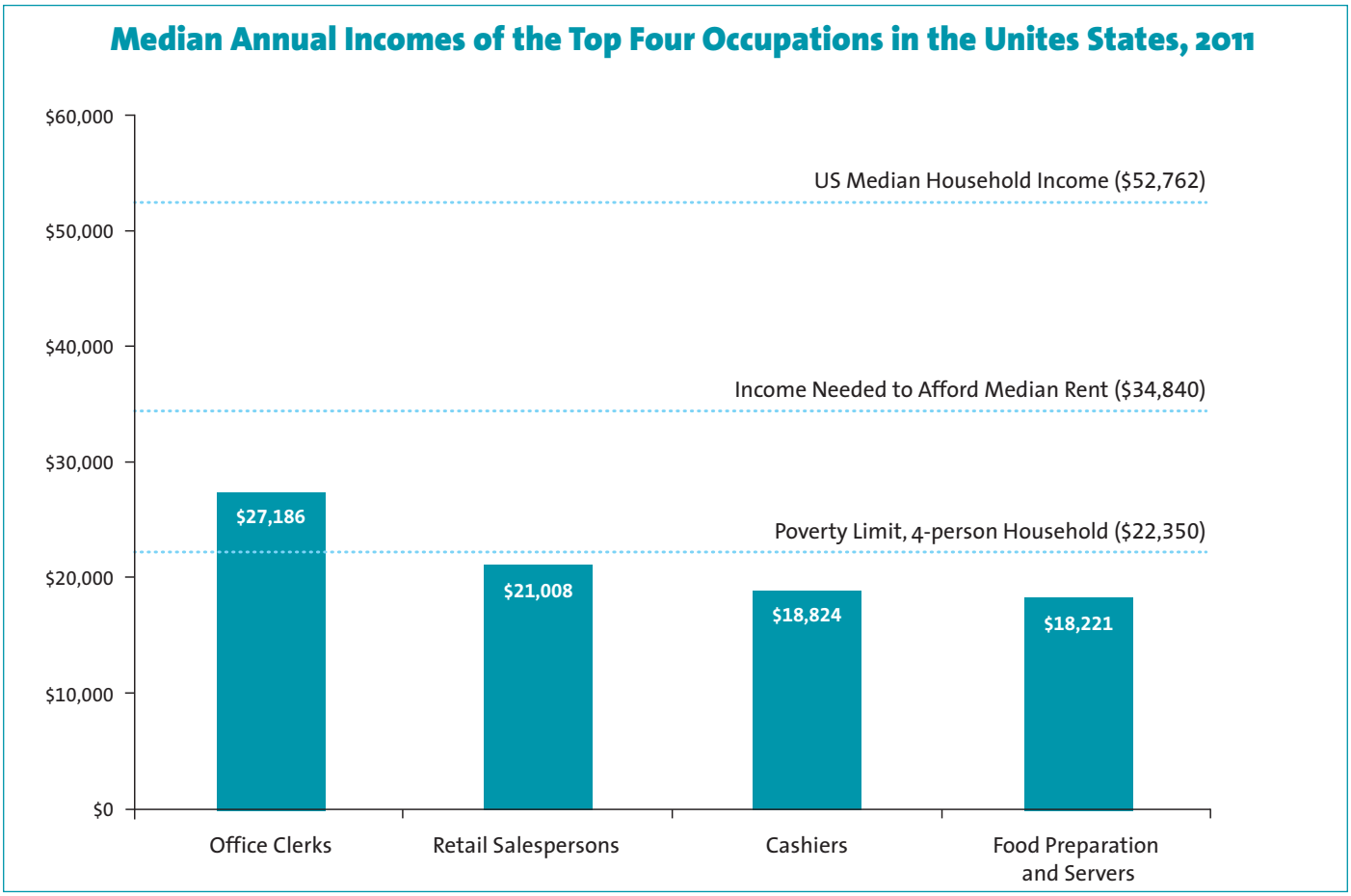
An analysis of the effect of the 2008 global economic recession *[NELP12][PLUM13]* reveals more evidence pointing in the direction of the end of work. After the recession of 2008-2009, the partial recovery from 2010 to 2012 in the USA not only failed to recreate the same number of jobs, but the jobs it created were lower paid jobs. E.g.: when a bookshop closes, a well-educated bookshop owner loses his or her job, leading to more sales for web shops like Amazon. Consequently, Amazon needs more people in its warehouses and shipping companies like FedEx need more personnel to deliver books. Several 'good' jobs are replaced by one or more unskilled and low-paid jobs. When Uber *[Uber]* arrives in a city, some taxi drivers (who already have a lower-wage job) are replaced by private drivers. The self-driving car may create even more havoc.

On top of this, the income from these lower wage occupations is often below the poverty limit, leading to a growing class of *working poor*.

Job Losses and Gains during the Recession and Recovery, 2008-2012



Source: National Employment Law Project



Sources: U.S. Bureau of Labor Statistics Occupational Employment and Wage Estimates, May 2011; US Department of Health and Human Services Poverty Guidelines, 2011; American Community Survey 5 year Estimates, 2011. Note: Median monthly rent for the U.S. in 2011 was $871; affordable rent is defined as nor more than 30% of income.

Some people welcome the "end of work", because it can lead to shorter working weeks and more leisure time, while letting the computers and robots do the work. However, reality might be different. Economy favours the combination of full time employees and unemployed people instead of part time jobs for everyone. Changing this will require drastic societal and mindset changes.

Some analysts predict *[NISE13]* that in the long term 50-75% of people will be unemployed simply because there is too little work left for which the economy needs humans, and therefore jobs will either be very specialised or will require hardly any skills. This would mean that the number of middle class jobs will keep shrinking and that income inequality will continue to rise. Not all experts share this pessimistic vision though, as covered later on.

All experts do agree, however, that the current educational system is not adequately preparing a generation to deal with this future. Similarly, our political and economic institutions are poorly equipped to handle these radical changes *[FoJ14]*.

The limiting factor is not technological advancement, but society's ability to absorb them. It is clear that education, our political systems, unions, markets, economic incentives, welfare systems and legal frameworks are currently too rigid to deal with the speed of technological innovations in a adequate way. It is a major challenge for society to cope with this change.

INFORMATION TECHNOLOGY WILL HAVE AN INCREASING IMPACT ON THE ENVIRONMENT

According to The Global Footprint Network, the 2014 Earth overshoot day was on 19 August, which means that by August 19 all resources that planet Earth naturally produces in one year (wood, fish, animals ...) were consumed by that day. Earth overshoot day arrives a few days earlier every year *[Footprint]*.

The total power consumption by information technology infrastructure grows around 7% per year *[VanH14]*, i.e. doubling every decade, consuming a total of 900 TWh in 2012, or 4.6% of the worldwide electrical energy production. With the further proliferation of information technology (Internet of Things, the introduction of 4G ...), there is no evidence to assume that this trend is going to change in the coming years.

ENERGY DISSIPATED FOR A CALL: WHAT IS BEST, LANDLINE OR MOBILE?

A landline telephone, when in use for voice communication, consumes about 7.5 Watt. This power is furnished completely by the central switch office.

A mobile phone's transmitter consumes from 0.5 to 3 Watt. When not in use, the mobile phone is also continuously listening for, and making its presence known to the base station, consuming at least some power. It turns out that, if we would charge our mobile phones continuously, it would consume about 0.5 W. Note that in a landline, when the phone is on the hook and not ringing, it does not consume any power because it is electrically disconnected: a connection is made when the receiver is picked up from the hook, or when the bell starts ringing. So, when comparing mobile phone connections to landline connections in terms of power consumed by the phone device, we are comparing 7.5 W to 0.5 W.

Of more importance is the amount of power used globally for mobile devices, compared to the total power consumption of the ICT ecosystem. The number of mobile phones is expected to exceed the world population this year [DTC13]. Then, the total power consumption of all the mobile phones in the world, assuming there are 7 billion people on Earth, is 3.5 GW. The total amount of power consumed globally by end user devices (PCs, TVs, mobile devices) is estimated to lie between 50 and 140 GW. The total amount of power used by the whole ICT ecosystem, including data centers, communication infrastructure, and ICT device production, is estimated between 1250 and 3200 GW. (The uncertainty comes from the many assumptions made, and the ongoing shift of TV into the digital ecosystem.) Compared to these numbers, mobile devices account for a mere 0.3% to 0.1% of the total digital ecosystem's power consumption. (See [MPM13])

Another trend is so-called planned obsolescence *[Obsolescence]*, i.e. devices that are designed to last only for a certain period of time, devices that are designed in such a way that they cannot be repaired or are very difficult to repair, and devices that won't be serviced after a certain time. The smartphone business is infamous for its planned obsolescence. Given the impact of the production of mobile devices on the environment (both in terms of rare Earth materials *[Shei14]* and energy required *[Ragh11]*), planned obsolescence for mobile devices is very detrimental to the environment.

Code bloat is one of the causes of device obsolescence. It has many causes and it is not a new issue. With the resource constraints of mobile devices and the end of the free lunch offered by Moore's law, the incentive to do better is growing rapidly, however. For example, double-digit energy savings could be expected on popular websites if simple measures were taken to contain code bloat *[Aron12]*.

The production of electronic devices also has a geopolitical aspect, because many of the required minerals are scarce and have to be mined in geographical areas that are politically unstable or with bad working conditions. There is a growing awareness in technological companies for this aspect of their business *[Gunt14]*.

There are some smaller initiatives that produce mobile devices of which certain minerals are mined under acceptable conditions in a certifiable way. One example is the Fairphone *[Fairphone]*. Google's project Ara *[ProjectAra]* is an attempt to reduce the electronic waste by making a modular phone, and being able to replace/upgrade/repair individual components.

Urban mining techniques can increase the availability of raw materials. Such techniques extract minerals from electronic waste instead of from ore *[UrbanMining]*. One metric ton of electronic waste can contain up to 300-400 gram of gold, which is up to 50x richer than ore. Urban mining is one of the key technologies to make the electronics industry more sustainable.

INFORMATION TECHNOLOGY WILL BRING IMPROVED EFFICIENCY TO SOLVE SOCIETAL CHALLENGES

While one could get the impression from the previous sections that information technology has only negative impacts (which could lead to some new form of Luddism), there are of course also many positive potential developments. For example, a recent report entitled "AI, Robotics, and the Future of Jobs" *[FoJ14]* shows that while half of the interviewed experts envision a future in which robots and digital agents have displaced significant numbers of both blue- and white-collar workers, the other half expect that technology will *not* destroy more jobs than it creates by 2025. They believe that, as in the past, through human ingenuity we will create new jobs and industries. Furthermore, Information Technology is also expected to be key to improve our use of natural resources.

In order to determine how exactly technology can help us, we first have to look at the problems we face. Western society is confronted with many challenges, most notably: energy, efficient and sustainable mobility, affordable health care, protection of the environment, productivity in a globalised economy, safety and security, and the ageing population. One class of approaches to tackle these challenges encompasses (global) optimisation of processes in order to reduce the consumption of resources, increasing automation, and more machine intelligence. All of these require the availability of powerful computing platforms and the processing of huge amounts of data.

More concretely, we can look at the smart grid to increase energy efficiency and production/consumption mismatches, at cognitive computing for increasing the productivity of many workers, and at the Internet of Things and the smart city for multiple challenges. In these approaches, the computer will increasingly become the mediator between the physical world (through sensors and actuators), humans (via a series of advanced user

interfaces) and other information processing applications (often in the cloud).

Companies such as General Electric and Cisco Systems estimate that by the end of this decade at least a trillion sensors will be deployed to build a pervasive Internet of Things, representing a market of about $15 trillion by 2020 *[Cole14]*. However not everyone agrees with these numbers *[Gans14]*. It is also believed that the first real push towards the Internet of Things, or the Internet of Machines, will be in the industrial domain rather than in the consumer domain: Internet of Things, together with big-data and data analytics, has a big potential to improve processes and thereby save billions of euros, enough to motivate companies to invest in developing and implementing the technologies. The Internet of Things will probably first be used to improve productivity, inventory and parcel tracking, and for improving the efficiency of current industrial processes and production. At a later stage it may also reach consumers.

Interoperability and ecosystem creation will, however, be easier in an industrial context. Big consumer companies would like to keep their consumers locked in and will therefore not readily support interoperability with devices from companies that are not part of their ecosystem.

Continuous tracking of operational parameters, correlated with the environment and other interrelated systems, will allow for further improvement of the efficiency of systems. For example, real-time analytics of the flight parameters of planes combined with weather information allow for decreasing fuel consumption and increasing the planes' lifetimes. Cars communicating with each other and with traffic infrastructure will enable optimisation of travel time and energy consumption, and reduce accidents. Through the collection of trusted data, avoiding fake or biased information, real time data mining and analytics can become key elements to reduce our footprint on the environment by improving the efficiency of complex systems.

Due to population growth and the effects of climate change, providing a sufficient amount of healthy food and water to humanity will become more challenging. Complex networks of sensors, computers and actuators could monitor water and food production, reduce waste and improve productivity.

Table 1: Industrial Internet: The Power of 1 Percent

What if... Potential Performance Gains in Key Sectors			
Industry	**Segment**	**Type of Savings**	**Estimated Value over 15 Years** (Billion nominal US dollars)
Aviation	Commercial	1% Fuel Savings	**$30B**
Power	Gas-fired Generation	1% Fuel Savings	**$66B**
Healthcare	System-wide	1% Reduction in System Inefficiency	**$63B**
Rail	Freight	1% Reduction in System Inefficiency	**$27B**
Oil & Gas	Exploration & Development	1% Reduction in Capital Expenditures	**$90B**

Note: Illustrative examples based on potential one percent savings applied across specific global industry sectors.
Source: GE estimates

From http://www.ge.com/docs/chapters/Industrial_Internet.pdf

FULFILLING HUMAN NEEDS THANKS TO INFORMATION TECHNOLOGY

The fulfillment of human needs will be the driver for future successful innovations in the domain of computing and communication.

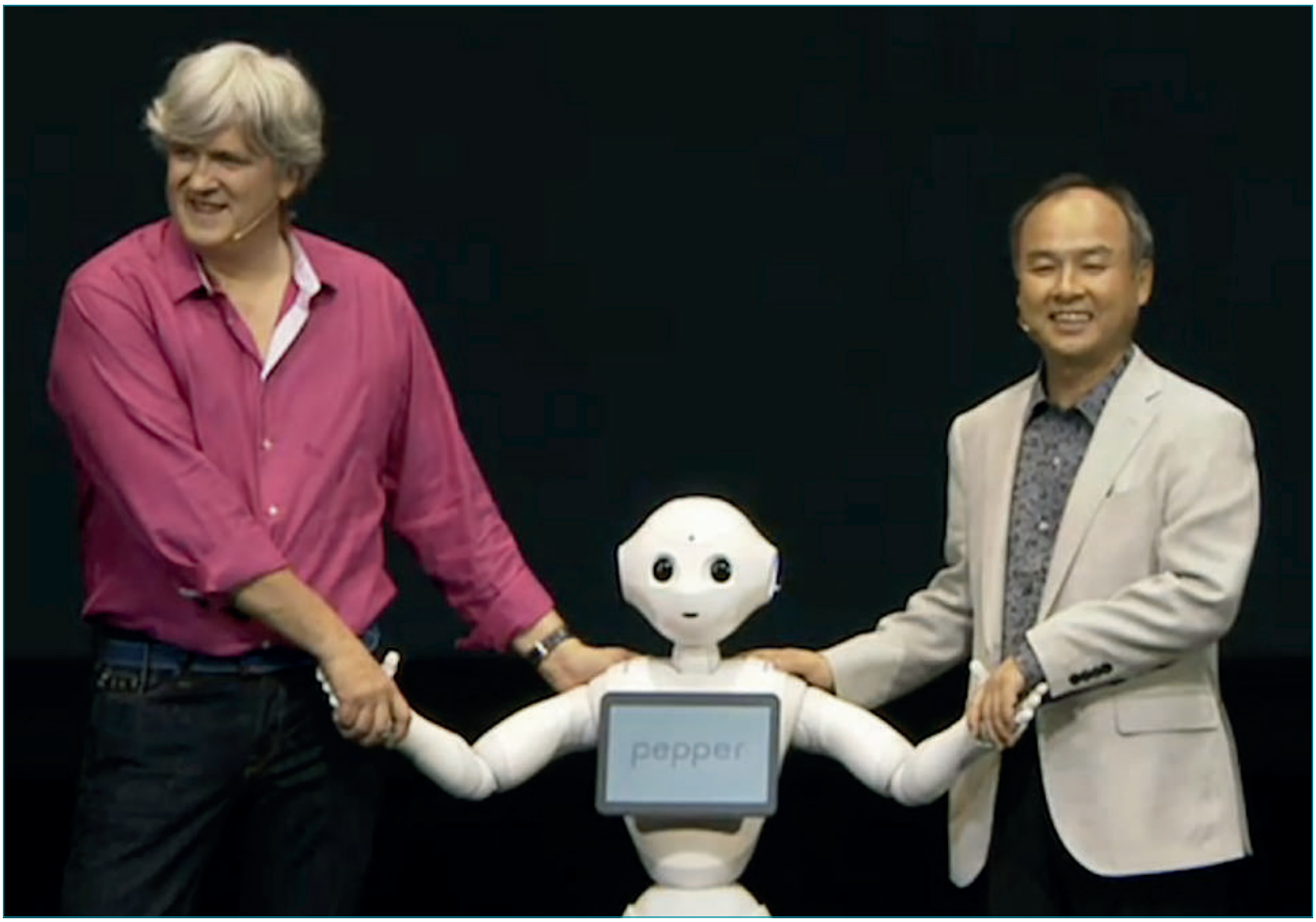
Communication is a basic human need, and the boom of mobile phones was driven by this need. In poor countries, people even limit their food consumption in order to save for a mobile phone. Socialisation was the key to the success of social networks. Our need for entertainment drives the games market, with more and more visually attractive and interactive artificial worlds, driving the need for ever increasing computing and graphics power.

Another recurrent demand is the ability to have an aid that can help with tedious tasks. The trend to develop Cyber-Physical Systems and of robots, i.e. devices interacting with the physical world, may soon drive a new market. As a result, the next big wave of technological innovations might come from robotics, and here as well, we see attempts by different countries and companies to try to dominate this emerging market.

The physical shape of these systems and robots may vary according to the local culture: Western countries have the so called "Frankenstein" syndrome (a fear of human-like machines), while in Asia, and particularly in Japan, humanoid robots are more likely to find social acceptance. "Robots" will probably have various forms through which they can integrate in their environments. The self-driving car and the smart home are examples of such devices. As far as the EU is concerned, we currently seem to be focused mainly on industrial robotics and automation with the "Factories of the Future" programme, smart factories, and "Industry 4.0" in Germany.

In Japan, Japanese Prime Minister Shinzo Abe announced a policy to revive the Japanese economy by tripling the robot market to ¥2.4 trillion ($24 billion). He aims to replace Japan's ageing workforce by robots and has expressed a desire to "set up a council on making a robotic revolution a reality in order to aid Japan's growth". The policy, dubbed the "robot revolution", will also see Japan host a Robot Olympics alongside the 2020 Tokyo Olympic Games.

Similarly, the South Korean government has announced a 2.6 trillion won ($2.69 billion) investment in its local robotics industry, in a move set to benefit industrial giants such as LG, Samsung and Hyundai. A key focus of the new South Korean policy will be the development of robots for disaster recovery and rescue operations. By 2019, the country hopes to develop a "healthcare town" where robots assist senior citizens, as well as a "robot business zone" where manufacturers can test smart factory technologies *[SKinvestm]*.



The Pepper "consumer" robot, from Softbank/Aldebaran: Bruno Maisonnier from Aldebaran and Masayoshi Son from SoftBank with Pepper in Tokyo.

In the domain of consumer electronics, home robots may not be that far off anymore. The recent announcement of the Pepper robot by the Japanese Softbank (the 2nd telecom operator in Japan) and the French Aldebaran companies *[Aldebaran]* might mark the start of affordable home robots. On sale in Japan by beginning of 2015 at the cost of ¥198000 (about 2000), it is already on display in the Softbank phone shops.
In the US, Google has bought eight leaders in robotics and AI including Schaft Inc., Industrial Perception Inc., Redwood Robotics, Meka Robotics, Holomini, Bot & Dolly, Boston Dynamics and DeepMind Technologies. The purpose of these acquisitions is not yet clear, but it is a definite sign that robots are a hot topic. Be it in the disembodied form of dispersed Cyber-Physical Systems, or in the concrete form of humanoid or other shapes, robotics is shaping up to be a big challenge for 2020. It may even become the next big market after PCs and mobile/smartphones.

In any case, the more we will be surrounded by robots to assist us, the more information about our lives will be disclosed to the companies running these robots if they are not autonomous and have to communicate with services administered by third parties. Google may very well know where all its driverless cars (and its owners) are – globally. The household robot will know when you wake up and go to bed, when you have your meals, etc. Companies and countries are both working hard to get access to all this information and to use it. Privacy-by-design techniques that inherently anonymise all aggregated data work via federations of open servers, and are optimised for divulging as little personal information as possible, and should be investigated in order to make these technologies compatible with professional secrecy, the need of a functioning democracy for informal (political and other) confidentiality, and the right to a private life. The best solution would be to have fully autonomous systems that only use local resources, without mandatory connectivity. However, at first "robots" will probably need to be connected to face the world. For example, the "Robo brain" *[Robobrain]* is a project that collects information from Internet and accumulates it in databases that could be easily accessed by robots.

STAR TREK: AFTER THE COMMUNICATOR, THE REPLICATOR

The Sci-Fi series Star Trek depicts a 23rd century where humanity travels through space. Even if the vision of the 60's about the far future still shows futuristic approaches (like travelling faster than the speed of light), some of the devices are now (nearly) real thanks to advances in computing systems. The communicator has similar functionality as (or even less than) a modern smartphone. Voice recognition and voice synthesis, including computers answering verbally to complex questions, exist today in tools like Apple Siri, Microsoft Cortana and Google. IBM's Watson was even able to win Jeopardy.
The next device that might become reality soon is the replicator, in the form of 3D scanners and 3D printers.
3D printing devices might create a new form of domestic industry: the investment for such a device has come down a lot in recent years and they are now roughly twice as expensive as a good quality laptop. The application areas are diverse, however, ranging from specialised plastic parts to applications in food preparation, requiring additional software to assist in design and production.
"Repair cafés", where you can get spare parts and repair help for broken devices, are popping up in many places and assist with creating a more sustainable economy. Augmented reality and computer-aided manuals, or even robots, could help to explain and show how to repair, while 3D printers could duplicate the broken parts.
In hospitals, 3D printing is already used to make prosthetics, and the same goes for dentists. (Giant) 3D printers are already building houses.
A next step will be to use 3D printers to make food (there are already some experiments to make pizzas *[Wong14]*), after which the food replicator from Star trek will no longer be just a fantasy.

THE NEXT CONSUMER KEY PRODUCTS

Robotics research, which in a way is part of cyber-physical research, is a preeminently multidisciplinary field, bringing together in one system information and communication technology, sensor technology, mechanical engineering, and knowledge of human-machine interaction. Information technology fulfills the key role to process all sensory data and to subsequently control the robotic system's physical interaction with its surroundings, including humans. Research on merging data from different sensors and external information sources will enable robots to perform specialised tasks required to take care of people, or manage processes. Since robots directly and **physically interact** with humans in this scenario, data integrity and data protection is extremely important.

THE ROBOT REVOLUTION IS IMMINENT

At the dawn of the information technology revolution, many people, including experts from industry and members of the public, ridiculed the idea of computers at home. Nowadays, almost every single electrical appliance includes information technology, ranging from very complex to relatively simple. We expect the same will hold for the future of robotics: nowadays, only specific tasks are envisaged for specialised robots. In the next ten years, however, we expect robotic technology to enter the house through appliances, making everyday life easier, but probably unrecognisable as robotic devices. Autonomous cars will be mainstream. Robots could be the answer to helping elderly people stay at home for a longer time, and they also are very useful to help nurses in hospitals or even at home for household duties.

Apart from the resulting impact on the industry, we must also be prepared for the societal impact as explained above.

3.2. MARKET

VERTICALISATION IS PROGRESSING

About 20 years ago, Europe had several vertical companies that had multiple activities and that were involved in many application domains. Companies like Siemens, Philips and Thomson covered domains from basic silicon technology up to the end product. Technologies were shared between application domains. Then, about 10 years ago, the more constrained economics, combined with the growing pressure on shareholder value, called for more "horizontal" companies, i.e. companies focused in those domains in which they excel. This led to the spin-off of semiconductor divisions (NXP from Philips, Infineon from Siemens, STMicroelectronics from Thomson) amongst other restructuring. Now Europe is composed of many specialised companies, each of which focuses on their own knowhow and their part of the value chain. They are squeezed between their providers and their customers, who in turn also try to maximise their margins and thereby put pressure on other players lower or higher in the value chain.

Over the last decade, especially in the domain of consumer electronics, companies that controlled a large part of the value chain (from technology, to hardware, to software, to devices, to services, to end users' systems) have gained dominance. Thanks to their diversified activities, they also weathered the current economic crisis fairly well. They achieved this by creating complete ecosystems and locking in the end users. By controlling the end-user data, they have access to a gold mine of useful information that allows them to further improve their ecosystem.

Google, starting as a search engine, now collects a lot of information about users and their behaviour by tracking their web activities, by analysing their free mail service and by locating them thanks to Android phones, using Google's operating system. Google is not a pure software company anymore; it is building its own data centers and has tablets and phones with their own brand name, even if the design and construction was sub-contracted. It invests in new generations of devices – such as wearable ones – with smart watches and augmented reality tools.

Apple is also enlarging its part in the value chain by designing its own processors. By controlling the hardware and the software, Apple can have optimised solutions without having to pay for extra layers of interoperability. The software is tuned to the hardware and vice versa, allowing for reduction of the amount of memory required, for example, thereby saving costs. Amazon, Facebook and Samsung also have been trying to grow their share in the complete value chain, from basic technology to devices and retail shops.

WHAT IF.... A VERTICALLY INTEGRATED COMPANY HAD A MOOD SENSOR?

Imagine a company that provides what you see (streaming video or audio), and that also can determine exactly where you are (by your smartphone), what you are doing (via smart sensors) and can even analyse your movements, your pulse and the resistance of your skin. Next, this company can correlate all of this data and construct an emotional profile while you are watching a movie or listening to music. It could then tell you exactly what you like, even if you do not realise it yourself. It could also generate the appropriate environment according to your mood or with whom you are spending time. It could give you recommendations to avoid stress – and in some ways manipulate you in a very subtle manner. Wonderful or scary?

To better control their value chain, de-verticalised European companies should work together to establish a "virtual" vertical company. In the domain of airplanes, Airbus is quite successful in creating a multi-partner verticality. The space domain also has a strong ecosystem. In the domain of processor IP, ARM is one successful example of a company creating the right conditions for a virtual vertical ecosystem, worldwide.

However, two key elements that explain the success of vertically integrated companies are that they do not have to pay the cost of interfaces, compatibility layers, or standards to link the various pieces together, and that they can optimise (costs) globally, which is more efficient than trying to optimise all parts independently. In a virtual vertical ecosystem, interfaces are high level because low-level features can be proprietary and can disclose part of the IPR of the companies. Additionally, each company tries to maximise its own profit, which is sometimes counter-productive at the global level.

In the domains covered by HiPEAC, it would be interesting to see if one or two strong leaders could emerge and crystallise a coherent ecosystem around them (a sort of "Airbus" of computing

systems). This might be possible as the consumer market is slowly moving away from traditional PCs, smartphones and tablets towards "things" interconnected, with new constrains of low power, safety, security and deeply embedded in the physical world.

At least as important as having a couple of tent-pole companies is the ecosystem that they create to operate in. This ecosystem consists of suppliers, service companies ... That same ecosystem also attracts start-up companies that develop technology that might eventually be used by the tent-pole company. The perspective of becoming a supplier, or even of being bought by the large company, attracts venture capitalists interested in investing in such technology startups, looking for quick profits. It is far more attractive for venture capitalists to invest in companies that already have a potential exit plan. That many European technology startups are eventually acquired by non-European companies is a result of the fact that there are very few European companies interested in buying out such startups. There are counterexamples though: Gemalto recently acquired the US-based Safenet. Sysgo was acquired by Thales in 2012. Gaisler Research was acquired by the US-based Aeroflex in 2008, and then bought back in 2014 by the UK company Cobham plc.

Tent-pole companies are not created overnight, but they grow vertically and horizontally in time. Apple started as a desktop computer company and gradually expanded into other markets. Amazon started selling books, and is now a major player in the cloud business. These companies re-invent themselves regularly. Companies like Apple reinvented mobile telephony; Google is reinventing car-based mobility. The European carmakers could also have expanded into different sectors, but they did not. There are success stories though. Nokia successfully transformed from a paper mill and rubber factory to a cable factory, a telephone switch company and a car telephony company, to finally become the global market leader for mobile phones that it was for more than 10 years. Unfortunately, it didn't reinvent itself in time to cope with the emergence of smartphones. Europe could use more companies that want to become global market leaders.

THE MARKET CARES ABOUT APPLICATIONS AND SERVICES, NOT ABOUT PLATFORMS

The consumer market is driven by applications the general public is willing to pay for, or in other words that provide answers to basic human needs, like the need to communicate, to move, to stay healthy, to relax ... Consumers are attracted by solutions that are cheaper, more powerful, or that offer new functionalities, save time, etc. If a product or a service is cool or prestigious, some consumers might be willing to pay a premium for it. Electronic consumer goods are generally priced below $250, which seems to be a magical upper boundary for gifts, or even impulse purchases.

The professional market is driven by applications that generate more business income or save on the cost side, i.e. that have an impact on the profitability of the business. Investments must have a positive return. In the coming years, considerable savings will be possible by further eliminating paper as an information carrier and by making this information available on mobile platforms, instead. There are still huge administrative optimisation opportunities in governments, in the insurance sector, in healthcare. Another optimisation opportunity is interconnecting all isolated industrial automation systems, only 20% of which are currently IP-based *[Sege14]*. This will be a basic requirement for factories of the future, keeping in mind that they will also have to be secured.

Very few end users are knowledgeable about how modern appliances work. They generally do not care about the hardware platform, the technology used, or whether it is open source or not. They are mostly interested in the experience offered by a particular application or service and at what cost. That means that companies that are active in the development of platforms and tools critically depend on killer applications that happen to use their technology in order to be successful.

A structural weakness of the computing systems area is that it depends on the success of others to become successful. Another structural weakness is that computing platforms are rarely changed (which is why they are created in the first place). That means that a newly developed platform (be it hardware, software, programming language, tools ...) must be much better than the state of the art in order to convince companies to start using it, and it will often have to support legacy systems. It is not a coincidence that we are still using processors of which the architecture was designed 20 years ago. The same holds for programming languages.

NO MORE CLOUDLESS DAYS

Due to the large amount of data generated by users (pictures, movies ...), and the need to be online all the time and to share information (social networks), we moved from stand-alone desktop computers to mobile devices connected to the cloud. In the cloud, data is stored in remote servers, processed by companies, and can be exchanged and addressed by multiple terminals of various types (from computers to smartphones, tablets and wearables). Current computing and storage clouds, both for private and for business users, are mainly hosted by large companies like Google, Amazon, Microsoft and DropBox. This allows such companies to tune their hardware and software stacks to customer and periodic usage patterns, hardware availability, and overall strategic directions of their own business. The advent of federated clouds, mesh-distributed environments (Internet of Things), and increasing storage and broadband capacities available to home users is changing this, however, and is already opening up the "post-cloud" era. The growing awareness that this data is often abused by spy agencies, private companies and malevolent hackers does not help.

As a result, the massive unified data stores that are currently in these large data centers will become again fragmented over countless smaller systems, from Cyber Physical Systems (sensor data) to private people's personal NAS/SAN devices. The question

HOMOMORPHIC ENCRYPTION

The design of efficient and secure encryption schemes with the performance of general computations in the encrypted domain has been one of the holy grails of the cryptographic community. Despite numerous partial answers, the problem of designing such a powerful primitive has remained open until the theoretical breakthrough of the fully homomorphic encryption (FHE) scheme published by Gentry in the late 2000s. Since then, progress has been fast-paced, and it now seems that practical homomorphic encryption-based computing will become a reality in the near future *[Agui13]*.

then becomes: how do we ensure optimal processing, distribution and safeguarding of distributed data?

Computing started with big centralised mainframes that users accessed via dumb terminals. This was followed by the decentralised era of the PCs when computing became cheap. Similarly, the current trend to store all data in big data centers, which corresponds to the current optimum as economic model, might change if cheap and high volume storage becomes available. Such a trend may also be fuelled by the fear that the data can be exploited without our consent.

New disruptive technologies, like non-volatile storage, can easily change the landscape. What if we can have several Exabytes of storage for a cheap price in 10 cm^3? Perhaps users and companies would prefer to store their own data in a device they own and of which they know the location. In that case, only more advanced functionality will be distributed and will only have access to the data (or metadata) it needs. Such individual data stores that are globally coupled are called federated, or distributed, clouds.

While remote applications will need data from these private data stores in order to perform the requested tasks, they should not get access to all raw data. After all, in that case we would again be stuck with the problem of data confidentiality as it exists today with unified clouds. Instead, they should only be provided with the information required to perform the task. This information could moreover be anonymised, or be limited to statistical or other metadata, thereby abstracting the real information from the user. Therefore, reliable anonymisation and anonymising statistical abstraction of information is a necessary feature for the concept of federated clouds to take off.

Another emerging approach is to send encrypted data to the remote application. The application then performs its operations without ever decrypting the data. As a result, the application never knows the actual data nor the meaning of the results it computes. This process is called homomorphic encryption. It is the ultimate solution for keeping data private, but it runs completely counter to the current business model of companies like Facebook and Google. After all, they are built on gathering and reselling as much information about their users as possible.

BIG DATA: A FEW FIGURES

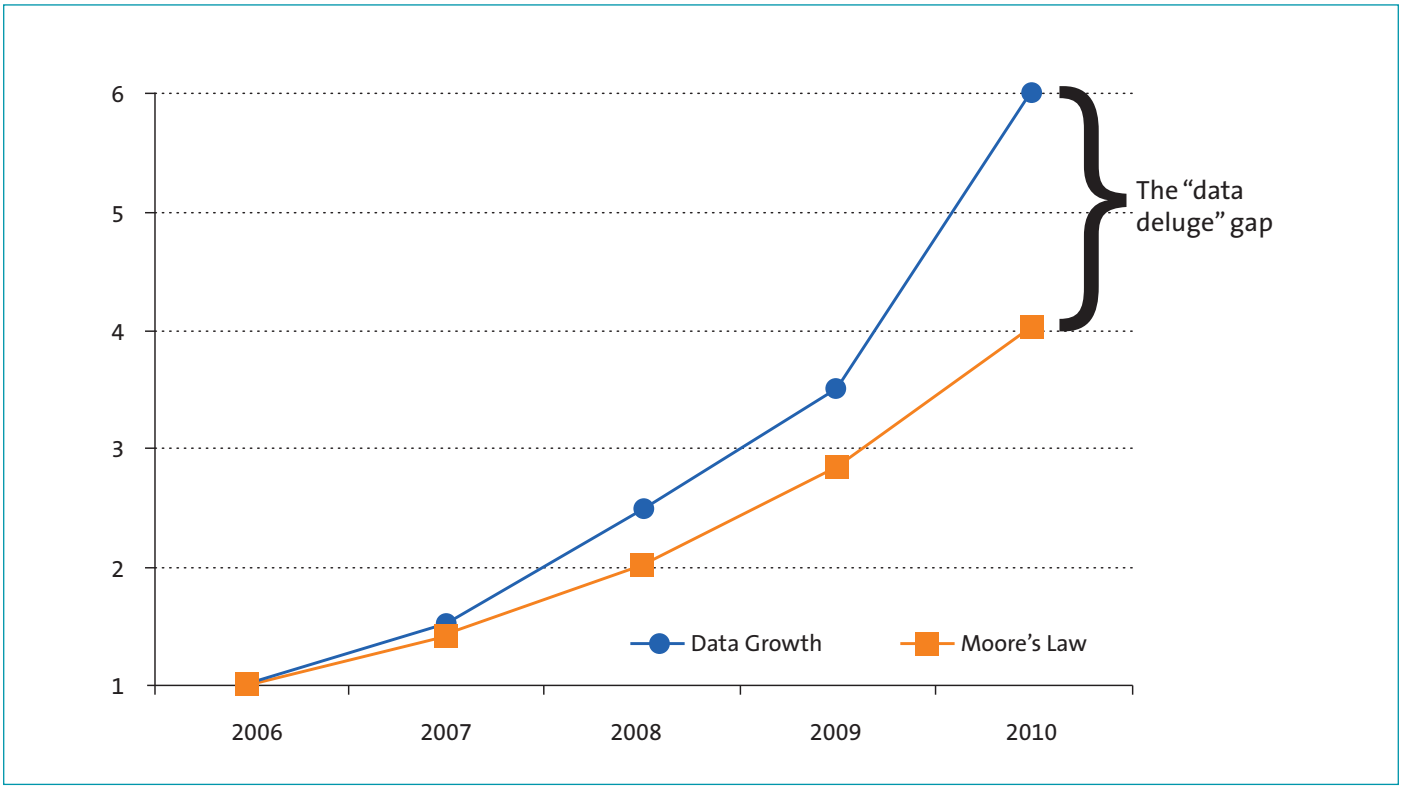
The term "Data Deluge" was coined in 2003 [Hey03] in the context of scientific data management to describe the massive growth in the data volume generated in research (and by scientific instruments), which was rapidly dwarfing all the data previously collected in the history of research. Since then, the ability to generate vast quantities of data has outpaced the infrastructure and support tools. This is true for scientific data, digital media (audio and video), commercial transactions, social networks, legal and medical records, digital libraries, and so on. The need to analyse, organise, and sustain "big data" is one of the highest priorities in information technology across disciplines, organisations and geographies [E10].

In 2010 the world generated over 1.2 Zettabytes (1021 bytes) of new data, 50% more than everything produced in human history before that year. To put this in perspective, 120 Terabytes of new data was generated in the time it took to read the previous sentence. For example, Microsoft Update and Windows Update push out more than a Petabyte of updates monthly (it should be noted that this is essentially a broadcast: the updates are new data, but all of the individual copies aren't). A social network like Facebook produces more than 10TB of data per day, with Twitter not far behind (7 TB/day); each of the 4.6B mobile phones and 30B RFID tags produces several events per second that need to be stored, processed and analysed. Likewise, the 2B Internet users also generate a variety of events that can have important value in areas like statistics, demographics or marketing. And the 50B connected devices expected by the year 2020 will cause all of the previously mentioned figures to balloon even further. Domains like gaming and other virtual worlds or augmented reality are also turning into massive data management problems.

In the scientific community, discovery has become a data-driven process, which represents a relatively new fourth paradigm in science [Hey09], next to the empirical, theoretical and computational models. The problem is that with current increases in computation, it may take over a decade to gain some understanding of what has already been archived from the most important scientific experiments. All fields of science (astronomy, physics, energy, medicine, drug discovery, climate, public health, etc.) are completely swamped with data. This requires major breakthroughs in repositories, storage, and computing architectures, topics that are all central to the HiPEAC mission.

If we compare this growth with Moore's law (transistor density doubling every two years, see below), it is clear that data show a higher exponential growth than computation capacity, and this unprecedented trend is forcing us to reevaluate how we work with data in computer systems.

[Hey03]



Data growth vs. Moore's Law trends in the last 5 years. Data "deluge" means that we are heading towards a world where we will have more data available than we can process.

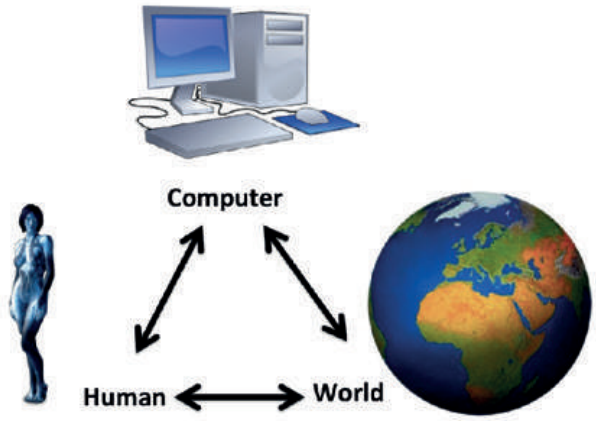
COMPUTING BUILDS ON INTERACTION + REACTION

It is possible to categorise the interactions between humans and computers in 3 main stages:

- **The keyboard era:** during this era, the interfaces between the user and the computer were mainly keyboard, mouse and screen. The performance model was best effort, which means that the result was computed as fast as possible and was presented to the user as soon as it was available.
- **The touchscreen era:** in most cases, the fixed keyboard and the buttons vanished and were replaced by touch sensitive locations on a screen. The hardware is now a white (in fact, black) sheet where various applications create the user interface they need. This is still the predominant interfacing style today. It is so natural that even small children are able to interact with mobile devices. Voice interface is emerging as well, with acceptable voice recognition and synthesis (Siri for Apple, Google, Cortana for Microsoft). Smaller devices can be equipped with a jog dial, like the Apple watch.
- **The cyber era:** We are entering an era where interaction between computing resources and humans will be multi-modal, and the computer will be directly interfacing with sensors and actuators. Therefore, its response time is constrained by external physical phenomena. Voice recognition and synthesis, artificial vision, sensing and processing natural data are the new challenges for computers. Results of computations are not only displayed, but can also directly control actuators, movements of a car or a plane, or regulate the power or temperature in your house. Google glass is the first example of a device that belongs to this class. Since the 70's, research has been ongoing to connect computers directly to the brain. Most of this research has been targeted at replacing lost sensory perception in humans, like cochlear implants for the deaf, and cameras for the blind. More recently, muscle control is under investigation with some success. From a market point of view, this field is too small to drive the market: the number of people with cochlear implants was estimated to be 220,000 worldwide in 2010, cochlear implants being the most widely applied directly connected devices. However, the gaming industry has been watching brain control with increasing interest.

In this new era, there are interactions between humans and computers, computers and the world (and, as before, between humans and the world even if, in some cases, the world is artificial – virtual reality). Of course, humans will keep interacting with each other, and computers will increasingly interact with each other as well. This is called the internet of machines, also called the industrial internet *[IndIntern]*.

Machines talking to machines, the integration of sensors and actuators in the world and new natural interfaces to communicate between humans and computers open totally new markets.



Now computers are interacting in a natural way with humans, and are interfaced with the reality

Examples are: driverless cars and autopilots of all kinds, factories of the future and 3D printing, virtual doctors (where a doctor can remotely check body sensors). These might become the killer applications of the next two decades. They will be characterised by the fact that they will react with humans, the physical world and the cloud.

This will translate into an increasingly diverse and ubiquitous computing and networking fabric (the Internet of Things), and will build on the conjunction of needs traditionally associated with distributed computing, embedded systems and control – so-called Cyber-Physical Systems – and increasingly, High-Performance Computing as well. They will feature a growing complexity, intrinsic reactivity, and the importance of non-functional properties: "*real-time computing is all about reactive computing. See how reality changes and react to it*" (Arvind, MIT). Since the physical world and nature's ways are all about feedback and reactivity, emerging computing systems should embrace these on a par with traditional computation and data communication, storage and processing.

The year 2014 is the 50th anniversary of the death (and the 120th anniversary of the birth) of Norbert Wiener, the "father" of cybernetics, a formalisation of the notion of feedback with implications for engineering, systems control, computer science, biology, neuroscience, philosophy, and the organisation of society *[NWiener]*. Feedback and reactive systems solve the complexity of analysing the inner details of the behaviour of a system by controlling its results in a closed signalling loop. For example, it is impossible to calculate beforehand all the commands required to control a flight from, say, Brussels to San Francisco: too many parameters are unknown, the equation or their influence too complex. But, in practice, by constantly calculating the "simple" difference between the current trajectory and the objective, and dynamically correcting it, it is possible to reach the objective without knowing all the equations and parameters that will influence the flight. A "simple" control rule in a "circular causal" relationship allows to master the complexity of calculating a flight. When computing systems were limited to the digital "world", they were quite simple, and their behaviour could be predicted beforehand "at compile time". When they interacted with each other (parallel or distributed systems), their behaviour could be theoretically predicted, but the complexity was so high that this is not done in practice, leading to more bugs and unforeseen effects. Now that computing systems have to cope more strongly with non-functional properties (e.g. execution time, reliability ...) and interact with the physical world, which is not (entirely) modelled by equations, it is very difficult to predict the right operation beforehand or at "compile time". Therefore, it is only by using principles from cybernetics that the systems will remain within the limits of the operational parameters or "meta-rules". Stronger links with control theory and reactive systems will be necessary to ensure that computing systems will work correctly, leading certainly to new programming and validation concepts.

A report to the US president *[Sciencedaily14]* explains that many techniques already exist for operating adaptive/autonomous devices, but their verification is hard, and new research is required. We believe that solutions inspired by cybernetic concepts may help progress in this field.

The design of such systems is squeezed between the hammer of dependability, performance, power and energy efficiency, and the anvil of cost (programmability, validation and verification, deployment, maintenance, complexity). Traditional, low-level approaches to parallel software development are already plagued by data races, Heisenbugs, the difficulty to reason about relaxed consistency models, time unpredictability, non-composability, and unscalable verification. Directions of research and emerging solutions exist to raise the abstraction level, and to develop dependable, reusable, and efficient parallel implementations. There is no alternative but for these kinds of solutions to mature, and relieve the designers and developers of real systems. New hardware architectures, software stacks and programming paradigms may be needed. Reactivity on a par with computation, optimisation for the current use case together with enforcement of non-functional properties (time-predictability, fault-tolerance) and specialisation should become mainstream. The design of reactive and dynamically adapting systems will push for rule-based or cognitive approaches, rather than explicit control and programming.

The Internet of Things also pushes for reconciling computation and control in computing systems. The convergence of challenges, technology, and markets for high-performance consumer and mobile devices has already taken place. The increasingly data-intensive and computational nature of Cyber-Physical Systems is now pushing for embedded control systems to run on complex parallel hardware.

The ubiquity of safety, security and dependability requirements meets cost efficiency concerns. Correct-by-construction methodologies have been successful for the design and validation of safety-critical embedded systems: these include being able to program, test, verify, compile a single source code, using an abstract model for verification and an executable model for simulation, relying on proven software layers with explicit

management of non-functional properties. Such approaches should be considered for more complex, mixed-critical and even non-critical computing systems alike. Long-term research is needed, as well as research evaluating the maturity of correct-by-construction methodologies with respect to time-to-market, *time-on-market*, and total cost of ownership.

One problem with real-time systems is that the current processor architectures were fundamentally designed to abstract the notion of time, thereby improving best-effort performance rather than predictability and the ability to cope with strict time deadlines. Caches, branch prediction and speculative execution add complexity when trying to determine the Worst Case Execution Time (WCET) of single core processors. Real-time is a major challenge that worsens with the emergence of multi-core systems. In domains where certification is mandatory and the lifetime of computing hardware is long, this becomes a major problem because there are fewer and fewer "simple" and predictable processors. Users are obliged to buy multicores and use only one core, even disabling caches and other features that add unpredictability. New computing architectures, designed from the start with predictability in mind, could really help the emergence of trustable systems for cyber-physical applications.

COMPUTING BECOMES INCREASINGLY COGNITIVE

With Cyber-Physical Systems and robots, computing systems are no longer simply a screen and a keyboard, but they interact directly with the physical world. Software and applications also have to evolve from just number crunching or ASCII code rearrangement, to higher levels of data processing: so-called "*cognitive computing*".

THE "HIPEAC" DEFINITIONS

Smart sensors: Combination of sensor, processing and communication modules. Aims at extracting data from the real world and efficiently communicating it.

Internet of Things = Internet of machines = industrial Internet: machines communicating with machines without humans (ff "classical internet" where humans are at the source or reception of information).

Cognitive computing: "translator" between real and cyber world. Aims at extracting information from data. Uses resources of the cyber-world to generate data impacting the real world. It interprets real world data and drives data mining and data analytics in the cyberworld.

Cyber-Physical systems: embedded intelligent ICT systems that are interconnected, interdependent, collaborative, autonomous and that provide computing and communication, monitoring/control of physical components/processes.

THE IBM DEFINITION OF COGNITIVE COMPUTING

Cognitive computing systems learn and interact naturally with people to extend what neither humans nor machines could do on their own. They help human experts to make better decisions by penetrating the complexity of Big Data.

Artificial intelligence meets business intelligence

Big Data growth is accelerating as more of the world's activity is expressed digitally. Not only is it increasing in volume, but also in speed, variety and uncertainty. Most data now comes in unstructured forms such as video, images, symbols and natural language - a new computing model is needed in order for businesses to process and make sense of it, and enhance and extend the expertise of humans. Rather than being programmed to anticipate every possible answer or action needed to perform a function or set of tasks, cognitive computing systems are trained to use artificial intelligence (AI) and machine learning algorithms to sense, predict, infer and, in some ways, think.

Systems with domain expertise

Cognitive computing systems get better over time as they build knowledge and get to know a domain - its language and terminology, its processes and its preferred methods of interacting. Unlike expert systems of the past, which required rules to be hard coded into a system by a human expert, cognitive computers can process natural language and unstructured data and learn by experience, much in the same way as humans do. While they'll have deep domain expertise, instead of replacing human experts, cognitive computers will act as a decision support system and will help them make better decisions based on the best available data, whether it is in healthcare, finance or customer service.

Humans and machines working together

In traditional AI, humans are not part of the equation, yet in cognitive computing, humans and machines work together. To enable a natural interaction between them, cognitive computing systems use image and speech recognition as their eyes and ears to understand the world and interact more seamlessly with humans. It provides a feedback loop for machines and humans to learn from and teach one another. By using visual analytics and data visualisation techniques, cognitive computers can display data in a visually compelling way that enlightens humans and helps them to make decisions based on data.

From http://researchweb.watson.ibm.com/cognitive-computing/index.shtml#fbid=mRF37M4r_BX.

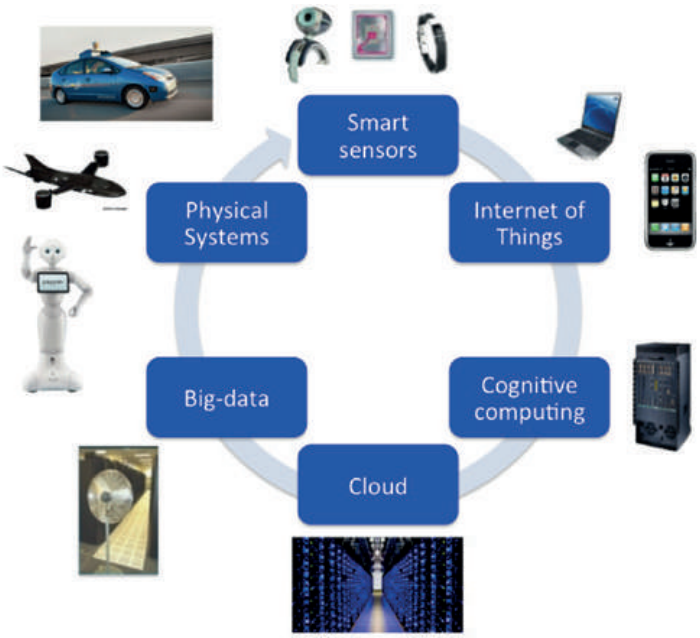
Cognitive computing promises a whole new range of killer applications for both consumer and professional markets. Potential applications are real-time speech-to-speech translation, self-driving cars, real-time image and video interpretation, cloud-based call centers, truly-interactive computer-based tutoring, online health monitoring, humanoid robots ... Companies like

Google and IBM are already investing heavily in this direction. The flip side of the coin is that cognitive computing will also destroy many middle class jobs: translators, teachers, health workers, drivers, call center operators ...

During the last two years, two fascinating examples of cognitive computing have been produced.

The first one is the self-driving car. Prototypes of self-driving cars are now able to navigate autonomously through different environments, and are improving at a fast rate. The most visible case of a self-driving car is the Google car (there is even a prototype that does not have a steering wheel anymore), but all car manufacturers are experimenting with self-driving in particular situations (self-parking, driving on a highway, in traffic jams ...) *[Self-driving]*. Carlos Ghosn (the CEO of Nissan and Renault) says Robocar sales could start in 2018 *[Robocars]* and that "*The problem isn't technology, it's legislation, and the whole question of responsibility that goes with these cars moving around ... and especially who is responsible once there is no longer anyone inside*".

More and more countries are allowing controlled experiments on the roads. Self-driving is a potentially disruptive technology which will completely change the way we think about mobility, transportation, the organisation of cities, etc. It will also severely impact employment for people working in transportation, a major source of global employment.



Entanglement between the physical and virtual world: The virtual, digital world and the real, physical world are being connected in the Internet of Things and in Cyber-Physical Systems. Cognitive computing is making the interface, often driving big-data analytics and data mining.

And, as C. Ghosn said, it is also a case of technology being ahead of the law in many areas.

First of all, there is the question of liability in case of a failure of the system. In the current legal framework the driver is liable for damage caused by an accident. Who will be liable if the driver was not controlling the car at the time of the accident, or if there was nobody at all in the car?

A second, more disturbing question is about the decision-making capability of the car's software in case of a dilemma. If a car has the choice between saving the car passenger's life and saving a pedestrian's life, which decision will it make in a split second? There is no doubt that the car can be programmed to give priority to protecting either the passenger or the pedestrian. Will it be possible to give more priority to the passengers than to the pedestrians by paying more? Will the driving skills of a car depend on its price or on the quality of its sensors?

As we are in the field of automated vehicles, another domain that will emerge with disruptive effects is drones. They are or will be used in agriculture, in energy for monitoring pipelines and oil rigs, in construction for inspecting bridges and other constructions, in the military for carrying weapons or inspecting areas (reconnaissance), in business to deliver goods as demonstrated by Amazon and Google, in law enforcement and emergency services to observe and inspect dangerous situations, in filming, in security and surveillance, in news gathering, in entertainment, etc.

As explained in *[Drones]*, drones represent an interesting convergence of three fast growing technology areas:

1. The Internet of Everything: Drones will be a key part of our trillion-sensor future, transporting a variety of sensors (thermal imaging, pressure, audio, radiation, chemical, biological and imaging) and will be connected to the Internet. They will communicate with each other and with operators.
2. Advanced Battery Technology: Increases in energy density (kilowatt-hours per kilogram) will allow drones to operate for longer periods of time. Additionally, solar panel technology enables high-altitude drones to fly for weeks without landing.
3. Automation Software and Artificial Intelligence: Hundreds of teams around the world are working on automation systems that a) make drones easier for untrained users to fly, but more importantly, b) allow drones to fly and operate autonomously.

DRONE TECHNOLOGY

The billion-fold improvement we've seen between 1980 and 2010 is what makes today's drones possible, specifically in four areas:
GPS: In 1981, the first commercial GPS receiver weighed 50 pounds and cost over $100K. Today, GPS comes on a 0.3 gram chip for less than $5.
IMU: An Inertial Measurement Unit (IMU) measures a drone's velocity, orientation and accelerations. In the 1960s an IMU (think Apollo program) weighed over 22.6 kg and cost millions. Today it's a couple of chips for $1 on your phone.
Digital Cameras: In 1976, Kodak's first digital camera shot at 0.1 megapixels, weighed 1.7 kg and cost over $10,000. Today's digital cameras are a billion-fold better (1000x resolution, 1000x smaller and 100x cheaper).
Computers & Wireless Communication (Wi-Fi, Bluetooth): No question here. Computers and wireless price-performance became a billion times better between 1980 and today.

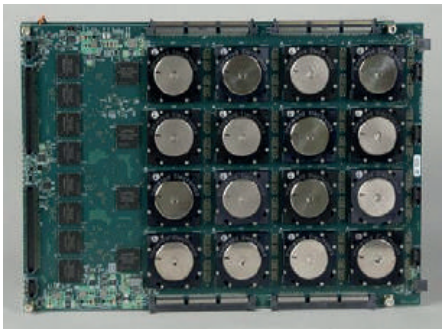
From http://singularityhub.com/2014/08/11/top-10-reasons-drones-are-disruptive/.

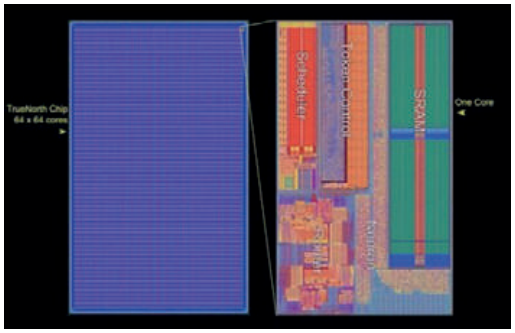
Even if US companies are the ones most visible in the field of cognitive computing, European companies are not lagging behind. All major European car manufacturers have self-driving cars, and, for example, Mercedes has demonstrated a 100km ride with an autonomous car between Mannheim to Pforzhein *[Mercedes]*. Companies like Thalès are also active in this field.
A second impressive example of cognitive computing is Watson from IBM *[IBMWatson]*. Watson is able to analyse unstructured data like video, images, symbols and natural language and builds up a database with domain expertise, like a self-learning expert system, and thereby becomes a powerful decision support system. Watson won the popular TV-game Jeopardy from human Jeopardy champions, and is now being trained with medical knowledge *[Watsonmedic]* and as a call center operator. Watson is another potential disruptive technology with the potential to displace a huge range of white-collar jobs in medicine, finance, translation, education and even engineering. To save money, insurance companies might demand patients in the future to first consult Dr. Watson before visiting a human doctor, or governments might require schools to use cognitive computing in education in order to save resources.

As for the self-driving car, there is also the question of the liability in the case of Watson. What if Watson suggests potential solutions that eventually turn out to be unfeasible, but take a lot of time and resources to investigate? Or what if it decides to switch off power in a part of the smart grid to protect the complete grid, thereby possibly endangering the lives of some people? The European project RoboLaw had a deliverable constituting a "guideline on Regulating Robotics" *[Robolaw]*, which is an in-depth analysis of the ethical and legal issues raised by robotic applications.
Cognitive computing is a new kind of workload, which will require hardware platforms and programming paradigms other than classical systems optimised for number crunching. This will lead to new accelerators for machine learning and "Big Data" analytics. For example, IBM recently announced its SyNAPSE chip, a brain inspired computer architecture powered by 1 million neurons and 256 million synapses. It is the largest chip IBM has ever built at 5.4 billion transistors and consists of 4096 neurosynaptic cores. This architecture is meant to solve a wide class of problems from vision, audition and multi-sensory fusion at very low power.
Other companies have also taken an interest in cognitive computing:

- Facebook works with Yann LeCun, a specialist in neural networks
- Google hired Geoffrey Hinton, a pioneer in Neural networks, and his team along with a few other experts
- Qualcomm developed the Zeroth neuromorphic chip
- Apple is interested in Neural Networks for speech recognition
- Twitter hired specialists in neural networks.

Several other companies also heavily invest in cognitive computing, which shows that they all believe that cognitive computing will be one of the future killer applications of computing.
It is clear that society will have to think seriously about an ethical framework for decision making by computers. A step towards the stories from Isaac Asimov and his "Three Laws of Robotics", which may not seem too Sci-Fi in 2020...
In the domain of High-frequency trading *[Highfreqtrading]*, where computers buy and sell financial instruments in seconds or fractions of a second, we have already experienced the effects of uncontrolled reaction due to the collective actions of computers: the May 6, 2010 Flash Crash *[FlashCrash]* was largely caused by overreactions of computers. Here as well, global rules or laws should be defined to limit such feedback loops leading to a crash.



IBM SyNAPSE neuro-synaptic chip (Source: http://www.research.ibm.com/articles/brain-chip.shtml)

DARK DATA

According EMC/IDC the digital universe will grow from 10 000 EB in 2014 to 40 000 EB in 2020 [Ailamaki14]. We have reached the point where sensors can produce much more data than can ever be processed, and we have to ask the question: how much data do we really need to answer a particular question? Do we need to know the key weather parameters of every square kilometre each minute to make a good weather forecast, or is it sufficient to have just 0.1% of this information, but well-chosen in space and in time? If Moore's law is really about to end, the future challenge (before new technology will re-enable the growth in performance for less power and area) will no longer be how to do more with more computing power, but how to do more with the same computing power. The big data community teaches us that a very limited amount of the stored data is actually used:

- 80% of the data processing jobs only use 1-8% of the stored data. 90% use less than 10% of the data stored.
- 80% of the data is used within 3 hours, and 95% of the data used is less than one day old (Yahoo!).

Hence the interesting data often is new data that is accessed within 24 hours after it was generated. All other data may never be used, and is sometimes called dark data. It consumes resources, but does not lead to new insights. The big challenge is to filter the data when it is generated. At CERN, the 1-2 MB data generated per second is reduced to 300 bytes that will eventually be stored. Some of the early filtering is done in hardware, while the final sifting is performed by software.

Essentially, the process consists of finding ways to record the information that will be needed in the future, and to design algorithms that can efficiently search the resulting data sets for answers. Experts in statistics typically perform this task, but in the future domain experts, with the help of cognitive computing, should be able to care of this by themselves. While computer scientists are not directly involved in this process, they are the ones who provide the cognitive computing environment to the domain experts, again demonstrating the importance of multi-disciplinary cooperation. The resulting environments need to strike a balance between required processing power (with the time and energy aspects), storage requirements, and quality of results, which requires input from all involved parties.

Another problem facing big data is that it is often not known beforehand how the data will be mined. As a result, a sufficiently general yet also efficient way to store the data must be devised, to enable dynamically structuring the data while it is processed. To reduce storage and communication requirements, it will be necessary to process data at all levels, including during production, and to store only potentially useful information.

HIGH PERFORMANCE COMPUTING: FROM BIG-DATA TO EXAFLOP COMPUTING

Science is entering its 4th paradigm: data intensive discovery. After being essentially based on Theory, Empirical science and Simulation, science is now using data analytics (cognitive computing) on instrument data, sensor data, but also on simulation data. Simulations are therefore becoming more and more important for science and for industry, avoiding spending millions in experiments. High performance computing is an enabling technology for complex simulations. There is a clear link between Big Data and Big Compute: Extreme Scale Computing is an enabling factor for timely and increasingly complex processing of also increasingly larger Big Data sets. It is expected that, in genomics, the data volume increases to 10 PB in 2021, in High Energy Physics (Large Hadron Collider) about 15 PB of data are generated per year, and climate forecasting will induce data sets of hundreds of EB [Exascale].

For significant advances in various research domains, the objective is to reach an exascale machine (able to perform the equivalent of 10^{18} floating point operations per second). It is very challenging, and the deadline is shifting over time. Five years ago, it was predicted for 2018, by now it has shifted to 2022-2024.

The challenges to overcome in order to reach exascale computing are enormous. We have to achieve 10^{18} operations per second and store 10^{18} bytes in a space with a power envelope that is not larger than today's petascale systems (200 cabinets, and 20 MW power). It will have 500 to 1000 times more performance than a Petaflops system and have 1 billion degrees of concurrency. Apart from the hardware design challenges which basically require that the current generation of computing systems has to be made two orders of magnitude smaller, more powerful and more energy-efficient, the application challenges are also non-trivial. First of all, there are only a limited number of applications that need or can afford exascale computing. There are often recurring complex calculations where the time of delivery is important. Examples are: space weather simulation (the result is not very useful if it is only ready a few days after a solar storm hits the earth), weather forecasts in general, (personalized) drug design (patients cannot wait years for a drug). However, an important side effect of exascale research is that petascale systems will become much more affordable, and that terascale systems might fit in a smartphone.

In order to reach exascale performance, many challenges must be overcome, and we will not be able overcome them with a "business-as-usual" evolutionary approach. Instead we will need real breakthroughs in these four areas:

- Parallelism / concurrency: adequate abstractions and programming models to express algorithms that exploit a billion-fold parallelism in a cost-effective and efficient way, do not exist. More research is required to design such abstractions and programming models.

- Reliability/resiliency: a computing system consisting of a billion components will never be fault free. Continuously, there will be failing components. The software will have to be made resilient, and deal with failing hardware components.
- Energy efficiency: given the end of Dennard's scaling, the challenge to reduce the power consumption by two orders of magnitude for the same performance (or have two orders of magnitude more performance for the same power envelope of today's petascale systems) is huge. We might have to turn to radical new solutions in order to reach this goal.
- Memory/Storage. The memory and storage requirements for exascale systems, and the associated energy costs are not affordable, given the current technologies. New memory technologies might be required to make them affordable.

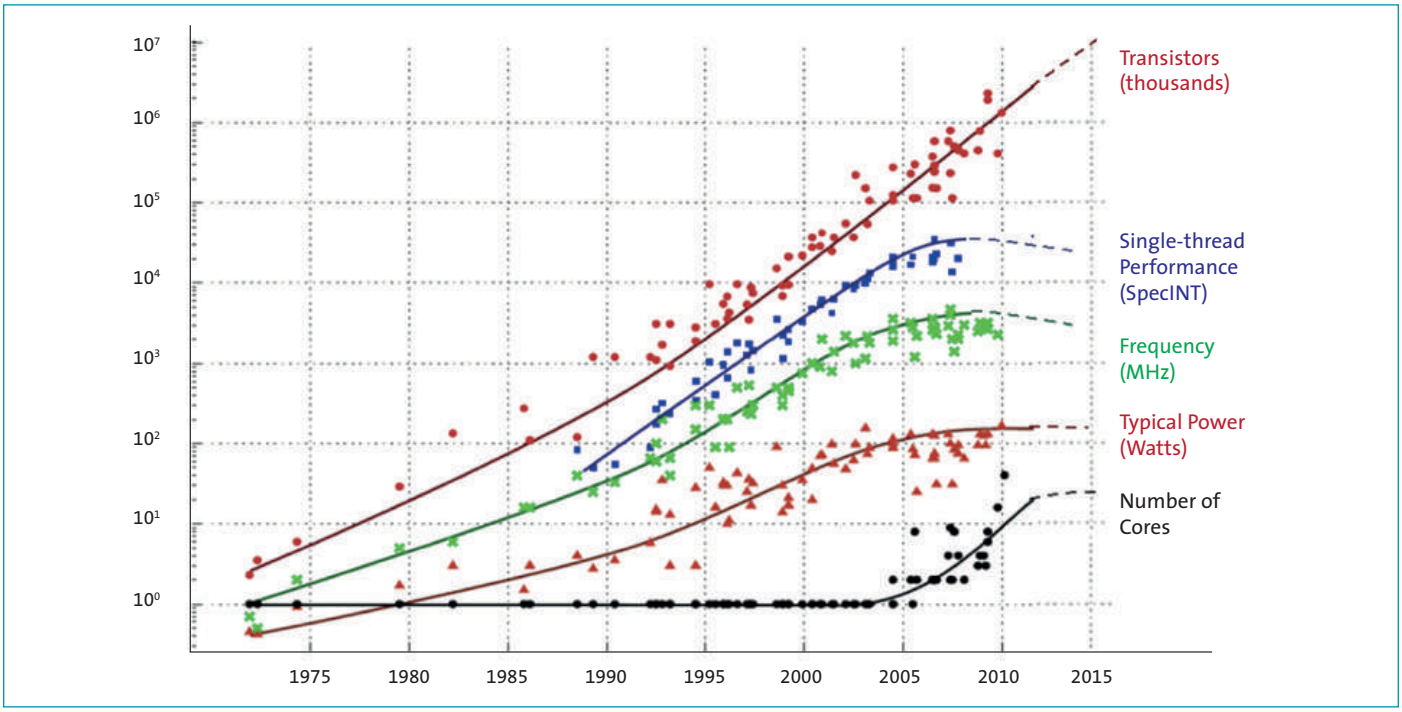
For more insight into HPC, the European Technology Platform for High Performance Computing (ETP4HPC) has released a vision paper [ETP4HPC] that details the challenges, the technological objectives and recommendations to reach exascale computing in Europe.

3.3. TECHNOLOGY

SILICON-BASED TECHNOLOGY: MORE AND MORE ROADBLOCKS

END OF DENNARD'S SCALING STILL A KEY LIMITING FACTOR: POWER AND ENERGY

In order to continuously increase the performance of devices, the race for ever smaller transistors is still ongoing. While Moore's law was claimed to have ended several times during the last decades, it still holds today and the number of transistors per square millimetre is increasing with each new technology node. Of course, physical limits will be reached, but technologists already think they will find a solution to finally reach sub-10 nm technology, which will give us a few extra years. What has changed in the last decade is the "law" of Dennard *[Dennard]*. In the early technology nodes, going from one node to the next allowed for a nearly doubling of the transistor frequency, and, by reducing the voltage, power density remained nearly constant.
With the end of Dennard's scaling, going from one node to the next still increases the density of transistors, but their maximum frequency is roughly the same and the voltage does not decrease accordingly. As a result, the power density increases now with every new technology node. The biggest challenge therefore now consists of reducing power consumption and energy dissipation per mm^2. This leads to the concept of "dark silicon"*[Darksilicon]*, where you can have a lot of transistors on a chip, but you will not be able to use them all simultaneously or the device will burn/melt. The end of Dennard scaling is one of the key elements driving the need for lower energy consumption, as highlighted in the previous HiPEAC vision document. Dark Silicon is also an important concern for server chips [Darksiliconserver].
Furthermore, due to reduced on-chip feature sizes, power leakage also increases. The reason is that transistors cannot be completely "switched off" anymore, which significantly increases their standby current. Hence, the ratio of leakage versus active current will increase, further degrading the systems' energy efficiency. The pattern sizes will become so small that signal loss in on-chip interconnects due to high wire resistance will no longer be negligible, increasing inefficiency even more. All of these factors aggravate the "dark silicon" phenomenon.



From C. Moore (original data by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten)

INCREASE OF THE COST PER TRANSISTOR

A new phenomenon appears when we approach the sub-10 nm nodes: for the first time in the history of silicon manufacturing, the cost per transistor may increase from one technology node to the next *[Merr12, Orba1403]*. If this happens, it will have a drastic impact on the economy of semiconductor manufacturing: not only will the most advanced technology nodes require gigantic investments, but for a similar design, the final cost will be higher. Since Moore's law is basically a law of economics, moving to the next technology node will no longer be driven by cost reduction, but only by the increase in transistors it will provide.

Only top end high-performance systems, or systems where the cost will be amortised by very large quantities, will in practice use the new technology nodes. Global players like Intel seem to be less affected by the cost per transistor *[Orba1412]*.

TICK-TOCK

Tick-tock is a model adopted by chip manufacturer Intel Corporation since 2007 to follow every micro architectural change with a die shrink of the process technology. Every "tick" is a shrinking of process technology of the previous micro-architecture (and sometimes introducing new instructions, as with Broadwell) and every "tock" is a new micro-architecture. Every 12 to 18 months, a tick or a tock is expected [Tick-Tock].

With the increased cost of design and of transistors, the die shrink by moving to a new technology node (the "tick") might neither bring an increase in performance (and may even induce more leakage power) nor a decrease in cost, so only the architecture changes (the "tock") may lead to improved performance, perhaps no longer systematically followed by a technology shrink. However, for Intel transistor cost is not yet a problem [Merr14].

PATTERNING, ADDING MORE CONSTRAINTS ON DESIGN

Reaching the ultra-small dimensions of the new masks requires the use of EUV *[EUV]*. This technique will require changes to most of the design methodology and will impose a geometric structure on the design: in order to reach the sub-10 nm range, which is an order of magnitude smaller than the radiation wavelength used to "draw" the patterns, interferometry processes are used. These processes, however, can no longer draw patterns on the mask and hence not on the silicon either. The highest density will only be attainable through the use of regular structures. This process is called multi-patterning. Having more regular structures at the physical level might increase the importance of computer architectures with very regular structures at small dimensions, perhaps like arrays of very small or simple cores, or FPGAs, which can be highly regular by nature (although this also needs to be apparent at the nanostructure).

HOMOGENEITY + DARK SILICON = END OF MANY-CORES?

The dark silicon problem will further limit the development of homogeneous many-cores: in a design where all cores are identical, moving to the next technology node allows for an increased number of cores, thanks to the still ongoing Moore's law. However, the power dissipation and the "Dark silicon" phenomenon will limit the number of cores that can be simultaneously used: it would be useless to add more cores, since only a fraction can be used at any one time anyway. It is therefore preferable to stay with a less advanced technology level where the dark silicon problem is not as high for homogeneous many-core systems. Even the cost factor of going to the new technology node with a similar number of cores will have to be seriously considered due to the increased cost per transistor.

A way to avoid the "Dark Silicon" problem is to use heterogeneous and therefore more specialised cores: depending on the task currently performed, the set of cores or accelerators that are optimised for the tasks will be activated. As they are specialised, they will also be more energy-efficient than a general purpose core to solve the ongoing tasks.

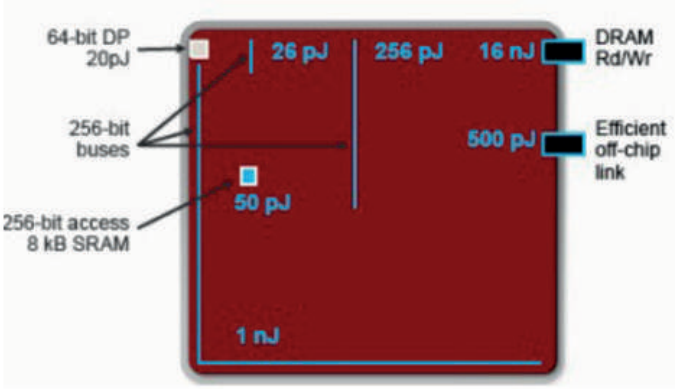
Another way to specialise is to use reconfigurable devices coupled with processors: currently, most FPGA vendors have solutions that merge FPGA arrays and classical processors. On such systems, the FPGA can be configured to be a coprocessor with a higher efficiency than a full software solution running on a general purpose processor. In this context, Microsoft has announced an approach to boost their Bing search engine using an FPGA solution *[fpga]*, and nearly at the same time Intel unveiled a Xeon chip with integrated FPGA, which offers a 20x performance boost for some tasks *[Xeonchip]*. If the market is large enough to cover the development cost, the FPGA configuration can be turned into a real optimised ASIC. New techniques to increase FPGA efficiency in terms of area and power may also emerge in the coming years. Examples include using non-volatile memory to store the configuration, using 3D stacking or interposers (which are already used in high-end FPGAs to increase yield — a big chip with a low yield is replaced by several smaller chips that each have a far higher yield), or new hybrid architectures that straddle the boundary between many-core systems and FPGAs (coarser grained FPGA).

COMMUNICATION COST DWARFS COMPUTATION COST

As explained in the previous HiPEAC vision document, communication and data movement take far more energy than processing the data themselves. Moving a 64 bit word from a chip to external memory requires roughly 3 orders of magnitude more energy than a 64 bit operation.

The interconnect is a very important source of power consumption, according to Bill Dally, quoting data from nVIDIA's 28nm chips (see figure). For a floating point operation, he said, performing the computation costs 20pJ and for an integer operation, the corresponding number is 1pJ. However, getting the operands for this computation from local memory (situated 1mm away) consumes 26pJ. If the operands need to be fetched from the other end of the die, the cost rises to 1nJ, while if the operands need to be read from DRAM, it rises further to 16nJ. *The energy required for computation is significantly smaller than the energy needed for interconnects!* This trend is expected to get worse with scaling, according to Dally.

From http://www.monolithic3d.com/blog/the-dally-nvidia-stanford-prescription-for-exascale-computing.



Computation cost is significantly lower than communication cost in 28nm nVIDIA chips (Source: W. Dally).

Therefore, it is of paramount importance that we rethink the memory hierarchy and reduce the data transfers. This will also avoid data duplication. Currently, data that is stored in external memory (and that also might be in several locations on the network or on disks), is duplicated in all cache levels (up to 3) and in the processor registers. At the hardware level, new kinds of memories, that are as fast as SRAM, have a low footprint as low as DRAM, and are non-volatile like Flash memories or hard drives, could enable us to complete the rethinking of the system and memory hierarchy.

For some applications, (simple) processing can be done directly inside the memory and thereby save energy. New, non-Von-Neumann computing paradigms, can move into the spotlight because of their inherent low number of data transfers and associated energy requirements. Neural Networks or other data flow approaches require only local operations on a "flow" of information that is transformed during its way through the system.

Even in conventional systems, the software tools should try to limit the data transfers, at the cost of re-computing some operations. We should move away from the compute centric way of thinking towards a data centric view: "*The memory is the computer*" *[Farab14CSW]*. Due to architectural improvements, the number of operations per cycle is increasing for processor cores, but the amount of data available in time by a instruction remains roughly constant, which results in major problems keeping the core fed with data and instructions. Major changes will be required, but the huge investment in the current architectures and memories will prevent disruptive changes. The memory market went through a large consolidation phase, leaving only three major memory manufacturers: Micron, Samsung and SK Hynix. Non-volatile memories might be the disruptive element that will reshuffle the market, but it will take time and money to become competitive with the heavily optimised existing technologies.

PROPOSED COURSE OF ACTIONS

To overcome the above-mentioned technological roadblocks, the HiPEAC community identified a number of technical solutions that should be investigated. The following sections will describe a few of them.

QUALITY OF EXPERIENCE WILL ALLOW FURTHER IMPROVEMENT OF LOW POWER COMPUTING ARCHITECTURES

The Internet of Things is predicted to become a driver for the next wave of volume growth in the semiconductor market. Ultra-low power wireless sensor nodes will be one of the dominant enabling technologies behind the Internet of Things vision and a great opportunity for the European industry to combine forces and win globally. Europe, with its complete ecosystem of IP providers, semiconductor manufacturers, middleware and applications, has the potential to provide innovative solutions by cross-layer optimisation across its ecosystem. To meet the ultra-low power challenge, new innovative solutions have to emerge, such as scalability in Quality of Service that can be further enhanced by combining circuit architecture, run-time adaptivity and in-situ performance monitors.

This opens up opportunities for new HW/SW computing paradigms that use design methodologies based on error tolerance and resilience techniques that employ graceful degradation for Quality of Service, speculative timing, and non-linear voltage scaling at reduced design margins. Correctly adapting to real application needs and hardware performance are key to achieving further system energy reductions. For example, it is not always necessary to use floating point or even 32-bit coding for some data. Operations do not always need to be accurate to the last bit.

If we relax the precision constraints, then the hardware can be simplified or can work with less energy. For example, the Razor *[Razor]* principle is to tune the supply voltage by monitoring the error rate during circuit operation, thereby eliminating the need for voltage margins, which allows for substantial energy savings (up to 64.2%) with little impact on performance due to error recovery (less than 3%).

The power consumption of CPUs and memory systems has traditionally been constrained by the need for strict correctness guarantees. However, recent work has shown that many modern applications do not require perfect correctness *[Samp11]*. An image renderer, for example, can tolerate occasional pixel errors without compromising its overall Quality of Service. Selectively-reliable hardware can save large amounts of energy with only slight sacrifices to Quality of Service (from *[Sampa]*).

Other approaches are possible, but their common idea is to allow a trade-off between precision requirements and energy. The quality of service of the device should be enough to satisfy the application or the customer, but not exceed margins that will lead to more energy use. This approach is often called "approximate computing" *[Approximate]*, but we prefer to call it "*Adequate Precision Computing*". This term better expresses the fact that the computations are tuned to the actual needs.

As mentioned in previous parts of this document, a lot of data that will be processed comes from the natural world. As a result, exact processing with a lot of accuracy is not always necessary. Several approaches and new technologies can be used to reduce the power, or increase the density, of processing engines, taking advantage of "less accurate" processing. For example, probabilistic CMOS, pioneered by K. Palem *[Cheemo5]*, consists of using standard CMOS, but lowering the voltage of the transistors. As a consequence, the energy is significantly reduced, but the transistor provides the correct output only with a certain probability. Large classes of important algorithms (e.g., optimisation algorithms) are compatible with such a probabilistic computing medium. The first prototypes show very promising energy gains.

"Independent from the specific application a system will be used for, all state-of-art compute architectures are based on power/performance trade-offs. In fact, it is unconceivable to think that any kind of competitive compute solution without power management could be marketed in the entire application field. However, the specific power management techniques and computing architecture optimisation could differ between individual market segments and could exploit the application-specific use-case and signal processing conditions.

The transition from dedicated to multi-purpose mobile devices and the typical service aspects of end-node enabled applications are changing how the perceived quality of a consumer electronics product can be evaluated. A change in paradigm from Quality of Service to quality of experience offers a chance to incorporate soft factors into the quality measure. This opens up opportunities for new hardware computing paradigms that use design methodologies based on error tolerance and resilient techniques through gracefully degraded Quality of Service, speculative timing, and non-linear voltage scaling at reduced design margins. Scalability in Quality of Service can be further enabled by a combination of a circuit architecture that enables run-time adaptivity and in-situ performance monitors. Essentially, scalable architectures can offer significant power and throughput gains if computational errors are treated as a new source of noise and if the processor's throughput is allowed to have stochastic behaviour depending upon the executed instructions.

Furthermore, specifically for the sensor signal processing functions in the end-nodes, computational or accuracy requirements can be compromised for further energy efficiency improvements. These techniques include exploiting the signal information redundancy, and deploying radical new concepts such as content-aware stochastic and approximate compute architectures."

Dr. Hans Rijns, CTO NXP Semiconductors and Prof. Jose Pineda de Gyvez, Fellow NXP Semiconductors

Even if the hardware is at the core of the energy reduction, a complete system view is required in order to efficiently use the added hardware features. Some of the challenges of Adequate Precision Computing are:

- How to specify the precision needed by the application, or for individual parts of the application?
- How to establish a productive HW/SW interface to control precision? The user or programmer should be able to easily provide relevant information that allows the toolchain to determine the precision requirement for all parts of the applications, and the trade-offs.
- New application algorithms will be required in order to efficiently cope with adequate precision computing.
- The adequate precision system should save power and area, even in case of reconfiguration to full precision, or at least not add too much overhead for moving from full to adequate precision.

Adequate Precision Computing will avoid designing the hardware for the worst case, and should allow for systems with one-sigma designs (68% of the products manufactured are statistically expected to be free of defects). As with reactive systems, adding measurements and feedback will allow a one-sigma design to operate within its QoS requirements.

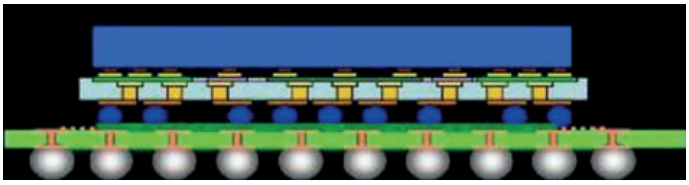
The scientific community shows a growing interest in approximate computing, and several research groups in the world are working on it *[Esma12ACM]*.

The idea of providing exactly the required level of accuracy or performance for the task in a controlled way (not in best effort mode), could also help when introducing new computational models like Deep Neural Networks, Bayesian computing, statistical/probabilistic models, reservoir computing, pseudo-quantum computing (à la D-Wave *[dwave]*), spike computations, etc.

ENTERING THE THIRD DIMENSION

A logical continuation of Moore's law is to stack dies and interconnect them vertically. Entering the third dimension may bring solutions for the challenges of power (by reducing the length of interconnections), diversity (making a "System on Interposer" composed of functions), and cost for building highly complex monolithic SoCs (by manufacturing each component using the optimal technology level).

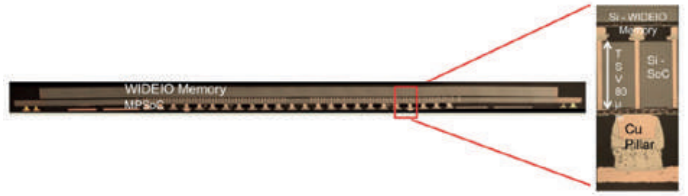
Several techniques to enable high-bandwidth interconnects between dies already exist, such as Cu-Cu interconnections, "through-silicon vias" (TSVs) and optical connections.



Multi-die stacking using Copper-pillars and TSVs (Courtesy of STMicroelectronics & CEA-Leti)

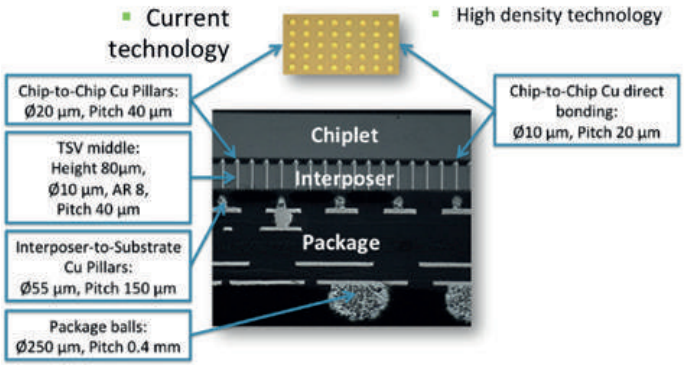
Die stacking creates several opportunities:

- It enables us to build new composite architectures by physically placing separately manufactured components very close together through stacking. For example, we can place memories and processors (for many-core architectures) or sensors and processing (example: intelligent retinas) on top of each other. Directly stacking the memory above the processor increases the energy efficiency by reducing the interconnect losses and, due to the higher number of interconnect points, can increase the bandwidth between the memory and the processor. For example, using WIDE I/O memories, the energy efficiency could be 4 times higher than using LPDDR2/3. Wide I/O 2 (JESD229-2), Hybrid Memory Cube (HMC), High Bandwidth Memory (HBM) are the new standards for 3D stacked memories [Cadence].



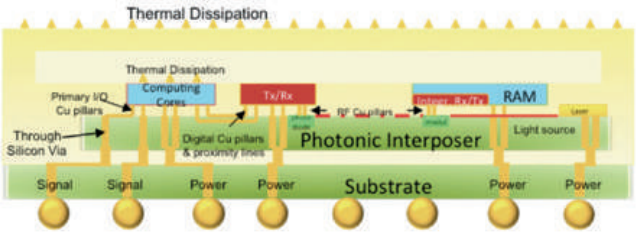
A Wide I/O memory stacked on top of a MPSoC (courtesy of CEA-Leti)

- It allows us to combine different technologies in one package, meaning that not all the dies in a package need to be produced in the same technology node. This can extend the lifetime of existing fabs by, for example, producing a sensor chip in an older technology and mounting it on top of a modern processor die.



Examples of 3D interconnect technology (Courtesy of CEA-Leti)

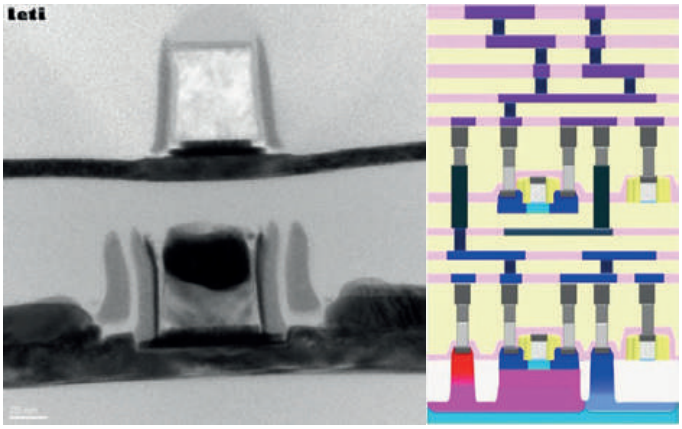
- Through different combinations of dies, we may be able to regain the chip diversity lost due to the increasing costs of chip design, and particularly to the cost of semiconductor chip fabrication plants, which doubles every four years (Rock's law). By reusing dies in different packages, the die volume of one design will increase, thereby lowering the cost, while simultaneously providing for more differentiation through different stacking combinations.
- Silicon interposers are also promising for the integration of silicon photonics, thereby enabling optical interconnects between systems or dies, with the potential advantages of lower energy for communication and higher bandwidth.



Artist view of a compute module using photonic interconnect between chiplets

3D Stacking also has its own set of challenges that must be overcome to make it economically viable at a large scale. These hurdles include reliably connecting the dies, dissipating the heat from the different dies, and distributing the functionality among the different dies. To increase yield, technologies such as die-to-wafer bonding or die-to-die bonding have to be improved, compared to the current wafer-to-wafer approach which is easier to achieve but which requires the yield of each wafer to be very high.

Another promising technology is monolithic 3D stacking, where transistors or nano-devices are built on top of each other, for example N transistors over P transistors for building CMOS logic with a smaller footprint.



Example of M3D, monolithic 3D devices, where transistors are built on top of each other (Courtesy of CEA-Leti)

Before the 3D stacking technology matures, systems using silicon interposers (also called "2.5D" solutions) will become mainstream. These systems use a silicon or organic substrate and a metal interconnect to create a highly integrated circuit board that can connect multiple dies together. These dies can be directly connected to both sides of the interposer (using a Cu-Cu interconnect for example, or using TSVs if necessary). Interposer technology is less demanding than full 3D stacking and can even be manufactured in existing fabs. Nevertheless, they enable high levels of integration at reasonable costs and can even be active, incorporating I/O or converter functions. An alternative solution is proposed by Intel with its "Embedded Multi-Die Interconnect Bridge" *[EMDIB]*.

2.5D WITH INTERPOSERS AND INTERNET OF THINGS: AN ESCAPE POD FOR EUROPE?

Development of very large and complex Systems on Chip (SoC) in the latest technology node requires huge investments and a worldwide market to amortise the costs. As a result, only SoCs for large markets, such as smartphones, and general-purpose processors will leverage those new nodes. This will also have the effect of drastically decreasing the diversity of designs. It is unlikely that European foundries of silicon manufacturers will be able to compete with the already established leaders in that field.

Fortunately, the new market for Internet of Things devices will drive device diversification, with each device specialised for its market niche. Some of these devices will not require the most advanced technology nodes at all, but in other cases a combination of newer and older technologies may prove sufficient and efficient.

Dies produced using different technology levels can be combined through the use of interposers into so-called chiplets. In this process, we may be able to regain the chip diversity lost due to the increasing costs of chip design. This also drives new challenges in standardisation, modelling, and programming for producing these complex systems in a package. Advanced chiplets, consolidating only the processor, its near memory (caches) and its interconnect, can be assembled on various interposers so that they can be integrated into a large range of systems. These may include mobile devices and ECUs (processing modules for automotive), micro-servers and data servers, and even HPC systems. Such wide applicability would allow us to amortise the cost of development of the chiplet over a large range of applications.

This constructive approach could ease the incorporation of novel IP blocks, accelerators or interconnects, like GPUs, non-volatile memories and photonic interconnects. Furthermore, a logical continuation of Moore's law is to stack dies and interconnect them vertically. This makes it possible to adapt architectures by physically reducing the distance between different components through stacking. For example, placing a memory on top of a processor could increase bandwidth while reducing energy needed for communication.

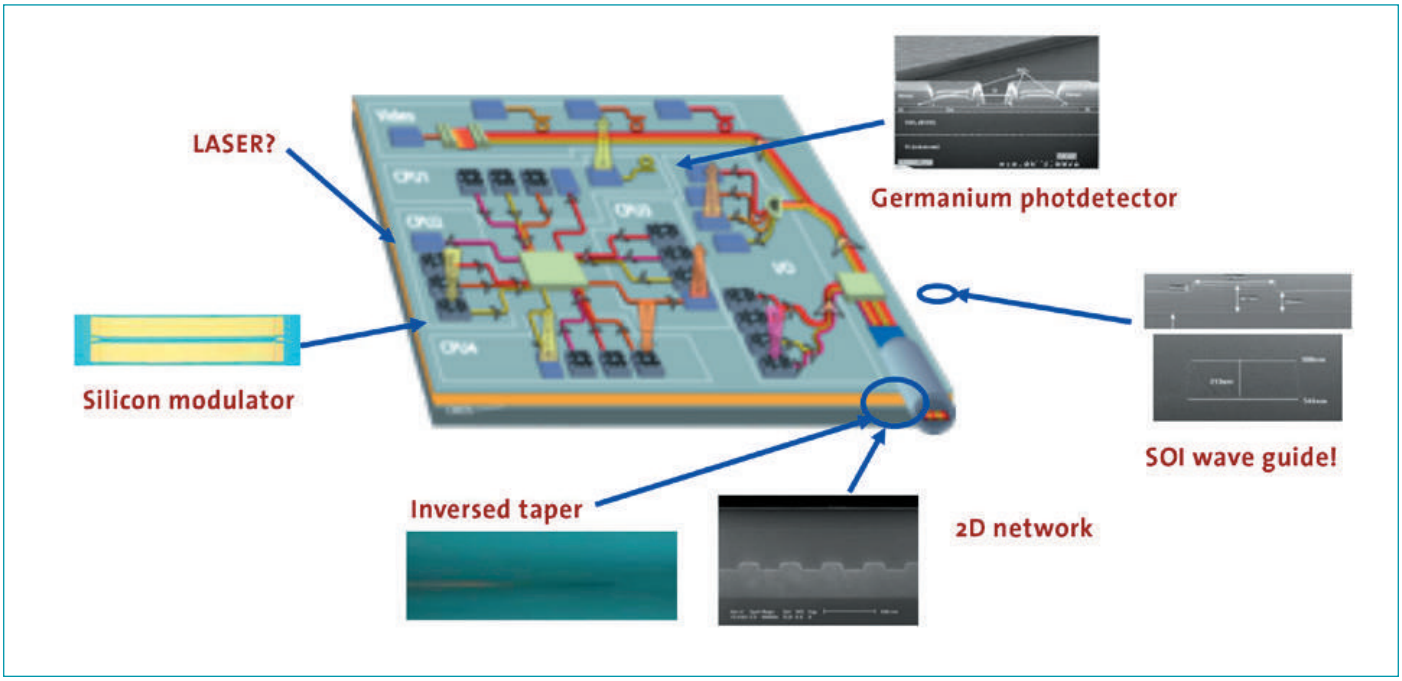
SILICON PHOTONICS

Current optical interconnects are generally used at the systems integration level or beyond, such as optical fibers connecting Internet devices across countries or oceans, and rack connections in data centers. Research into a scaled down version of this concept, called silicon photonics, promises to lower communication energy cost, higher bandwidth and low manufacturing cost in existing fabs. The technology is compatible with existing silicon technology. Transforming the electrical signal into an optical one can be power consuming, especially if serial conversion is required, since this leads to very high frequencies. Therefore, optical systems that support multiple wavelengths may be preferable from an energy point of view.

SUMMARY OF ADVANTAGES OF SILICON PHOTONICS FOR COMPUTING SYSTEMS

CMOS photonics: integration of a photonic layer with electronic circuits

- Use of standard tools and foundry, wafer scale co-integration
- Low manufacturing cost
- Lower energy (~100 fJ/bit), (wire: ~1 pJ/mm) -> less heating
- High bandwidth (10 Gbps), Low latency (~10 ps/mm)
- High integration
- Can also be used for on-chip clock distribution (using e.g. passive SOI interposer)
- More reliable / low error rate



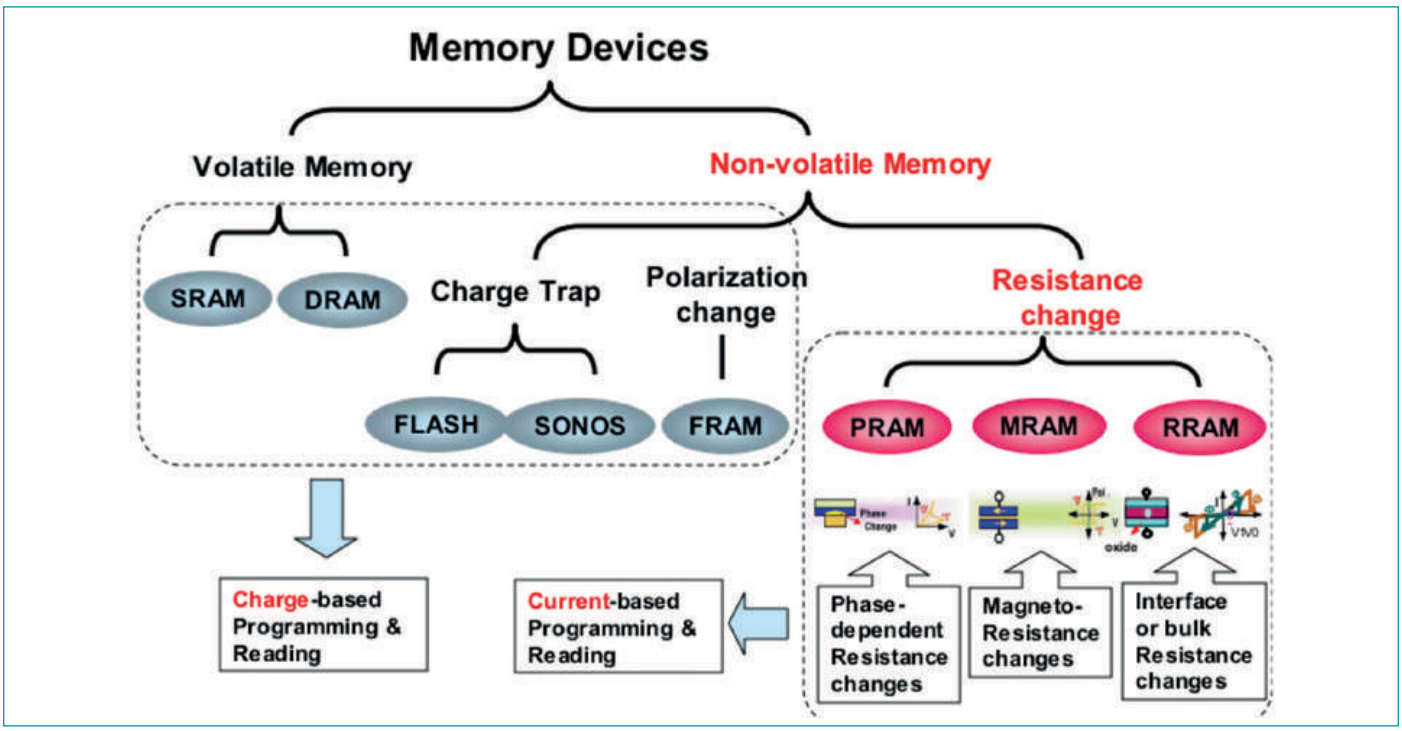
Example of photonic elements

EMERGING MEMORY TECHNOLOGIES

Non-volatile memories offer a number of very interesting properties that can be used in future computing systems. These novel memory technologies can be classified based on the physical processes implementing the data storage and recall process. The following figure shows a high level taxonomy.

For the reader interested in this burgeoning topic, a valuable review on emerging memory technologies can be found in *[Pershin11]*.

Of all technologies depicted below, MRAM and PRAM devices are the most advanced from an industrial point of view. MRAM was first introduced in 2004 by Freescale, and since then the technology has been refined to the point where Everspin was able to introduce a 64 Mb ST-MRAM chip. PRAM is currently in active development and industrial products are on the way with Micron just announcing a 1Gb PRAM module for mobile applications. Resistive ram (RRAM) comes in many technological flavors such as Conductive-Bridge or Oxide Ram; they are actively developed by industrial companies such as Adesto technologies



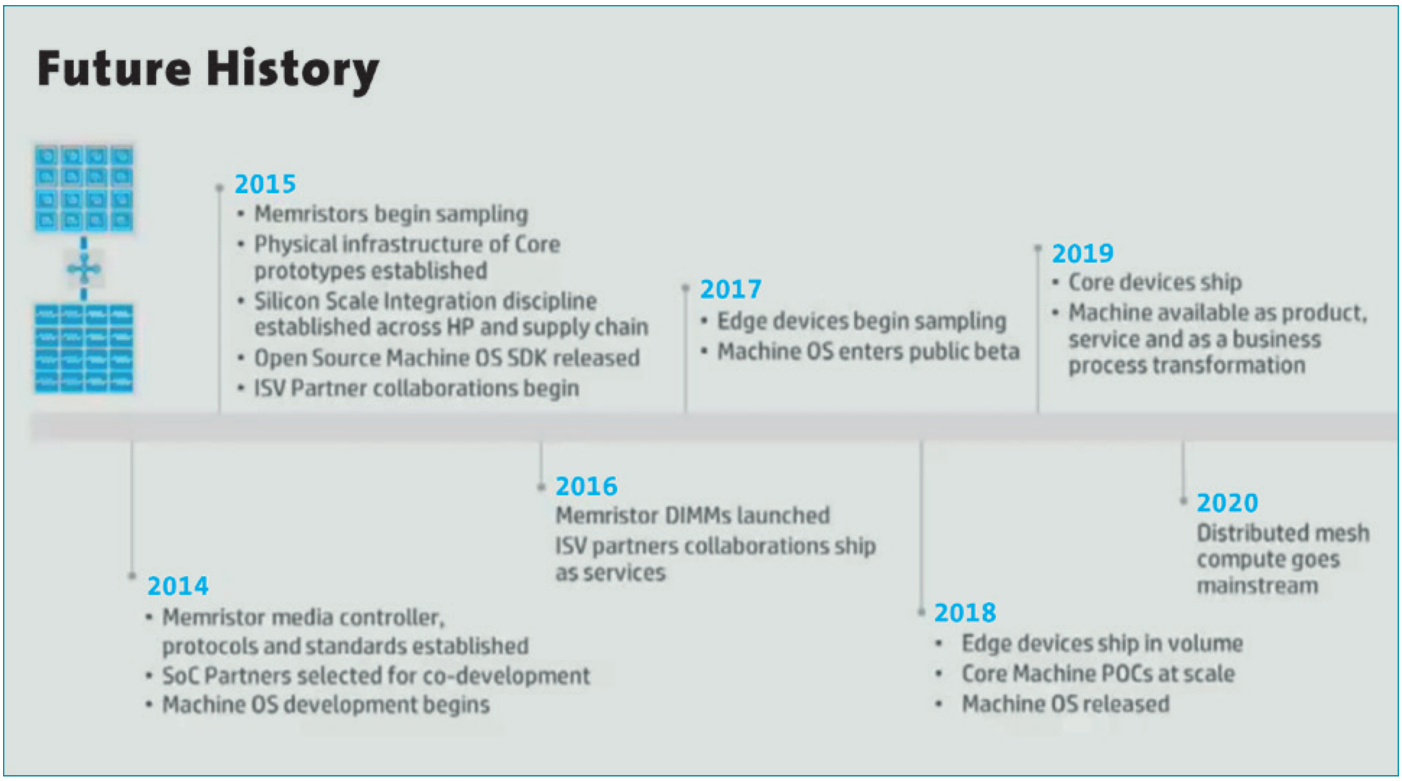
A high level taxonomy of memory device technologies

or Hynix and may prove to be the best option in the future for very dense arrays of non-volatile memory.

Since some novel resistive devices are dipoles (i.e. they are two-terminal devices), for which resistance can be controlled by applying voltage levels, the idea to organise them in crossbars comes naturally with the expectation of getting ultra-dense arrays of non-volatile memory bits. Although using the crossbar scheme with devices of nano-metric sizes can provide dramatic memory densities, doing so requires many problems to be solved: select logic, sneak paths and process variability, to name a few. However, a convincing demonstration of a 2Mb resistive device crossbar without selection logic has been unveiled recently by Hynix and HP [Lee12]. This work shows that putting those devices to practical use in digital design would be possible at the price of rethinking the whole architecture of the memory plane.

Up until now, practical demonstrations have targeted either reconfigurable logic or full arrays of memories. Another very promising application of non-volatile memories could be to use them in the design of processor register files with the aim of building instant-on/instant-off CPUs. HP recently announced a new computing architecture, called "the Machine" [Machine], which uses memristors for storage and photonics for the interconnect. The architecture of "The Machine" is not disclosed yet, but the memory hierarchy might be revisited due to the promises of new storage devices.

Non-volatile memories will also have an impact on software, and more particularly on the OS. HP calls for the participation of academia in the development of a new Operating System for its architecture using non-volatile memories. According to the HP roadmap, memristor memories (in the form of DIMMs) will be launched in 2016. Similarly, Toshiba plans to switch to 3D ReRAM in 2020: "But in and after 2020, we will need a new memory having a different mechanism. ReRAM (resistive random-access memory) and ion (magnetic wall displacement type) memory are candidates. We are also considering the manufacture of those new memories by stacking layers three-dimensionally, and they can possibly be combined with scaling beyond 10nm." [Toshiba]



From HP keynote [HPkeynote], showing their roadmap on "the Machine"

ELECTRONS FOR COMPUTING, PHOTONS FOR COMMUNICATION, IONS FOR STORAGE AT NET ZERO ENERGY

Besides the current technology for computing nodes, evolving with FinFETs or FDSOI, which rely on electrons, two other elements are key enablers for future computing systems: photonics for interconnects and non-volatile RAMs for storage. Light must ensure a 30x increase in bandwidth at 10% of the power budget of copper.

The current problem of photonics, with "ring resonators", is that energy is consumed even when the device is not used. This is an important fixed cost. Therefore the optical link must be shared (communications multiplexed on the same channel). Bandwidth to memory is very important, and it is a good field for improvement, but there is the problem of inheritance for the DRAM industry, with a low possibility of impacting it with new architectures, except when it will hit a wall. This will then be a good time for emerging technologies such as the NVRAMS and memristors. Non-volatile memory, with its property of addressing at the byte level, will induce new approaches to define efficient file systems that are not based on blocks of data. This byte access will also change the structure of the data cache. However, the memristors will not resolve the access latency problem (but perhaps will decrease the requirement for L2 caches). On the other hand, for reasons of low dissipation, compute nodes will work near the threshold voltage with working frequencies in the sub-GHz range. Simultaneous multi-threading can be used to hide memory latency. Memristors can be used to save the state of the processor when switching tasks. It would take about 100 threads to hide the latency of memory accesses, which is accessible for data server applications.

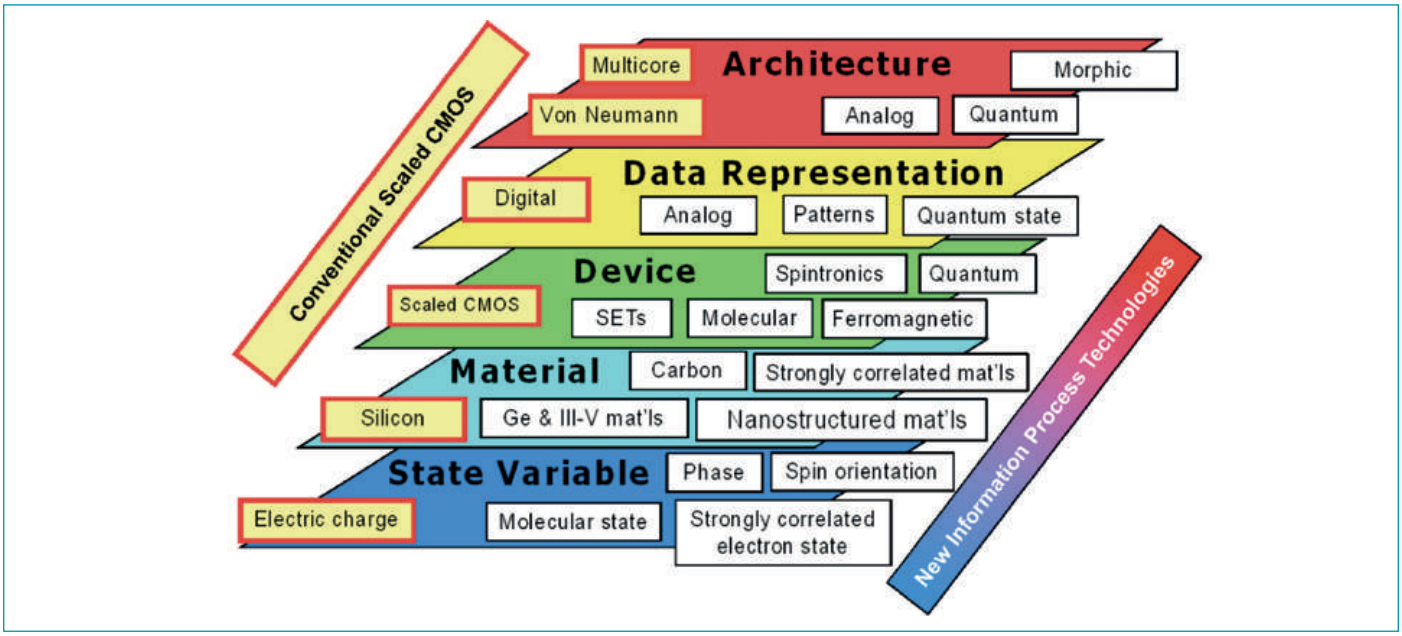
from Partha Ranganathan [Partha12]

In reconfigurable logic circuits, non-volatile memories can be used to implement lookup tables or to implement switchbox matrices, as demonstrated on a small scale by HP *[Xia09]*. This work was also one of the first demonstrations of the co-integration of titanium dioxide memristors on top of an industry CMOS process. More original approaches to implement sequential or stateful logic circuits using memresistive devices have been proposed *[Borgh10]*.

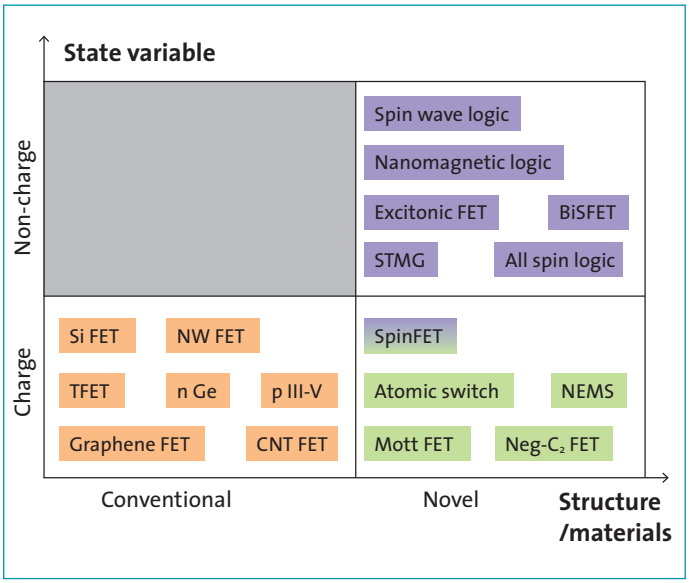
For digital applications, the key selling points of these novel technologies is their potential to yield ultra-dense arrays of non-volatile devices with possible CMOS-compatible processes. Using non-volatile memories at key points in a computing architecture can dramatically reduce its overall power consumption. New concepts for FPGAs or CGRAs (Coarse Grain Reconfigurable Architectures) will likely emerge, thanks to the properties of those new memories.

NEW TECHNOLOGIES FOR COMPUTING

The ITRS 2013 document on emerging research devices *[ITRS2013]* gives a taxonomy for emerging information processing devices and is a very complete document on new technologies that can be used for building new computing devices. We encourage the reader who wants to have more details on emerging technologies to read it.



A high level taxonomy of memory device technologies



Taxonomy for emerging information processing devices (from [ITRS2013]).

Research on novel computing architectures for use in such devices has already started around the world. The STARNet initiative in the US *[STARNet]* groups 40 universities and 108 industry associate personnel. It directly attacks the problem of the next generation of computing systems.

Two potentially disruptive technologies in this context are graphene and graphyne *[Schirber]* (Nobel Prize in Physics, 2010) transistors, which seem capable of increasing their clock frequency beyond the capabilities of silicon transistors (in the 100 GHz to THz range). Currently such transistors are significantly bigger than silicon transistors and only limited circuits have been implemented. Their current application scope is mainly fast and low-power analog signal processing, but research on graphene and graphyne transistors is still in its infancy. The computing industry may shift back to the frequency race instead of parallelism, if complex devices running with low power at 100 GHz become possible *[Manchester]*.

Many new technologies, such as wave computing, ambipolar devices, spintronic, plasmons, synaptors ..., will have properties that are very different from those of our current transistors. As a result, the most effective and efficient computational models may also be quite different. We should therefore look into alternative models such as Neural Networks, Bayesian computing, probabilistic, reservoir computing, pseudo-quantum computing (à la D-Wave), spikes computations, swarm computing, and amorphous computing.

NON-VON-NEUMANN ARCHITECTURES

Classical computing models use an explicit declaration of how to perform tasks. This is typical *imperative programming*, using the *Von Neumann model*. We can, however, think of paradigms where, instead of instructing the machine on how to perform its tasks, we only specify the goal(s) or the objectives, or teach the machine by examples. This category encompasses *declarative programming*, like database query languages (e.g. SQL), regular expressions, logic programming, and functional programming, but also other approaches like Neural Networks. These approaches are promising to cope with the complexity of programming large-scale parallel and/distributed systems. Most of them can easily be mapped to parallel systems.

Gaining inspiration from the brain is one way to improve our computing devices and to progress beyond Moore's law. As high-performance applications are increasingly about intelligent processing, the application scope of neural networks becomes very significant. More importantly, as the number of faulty components increases, hardware neural networks provide accelerators that are intrinsically capable of tolerating defects without identifying or disabling faulty components, but simply by retraining the network. They will not replace a complete system, but can be used as accelerators for "*cognitive tasks*". They could decrease the computational load of classical systems in an energy-efficient approach, mainly if implemented with new technologies.

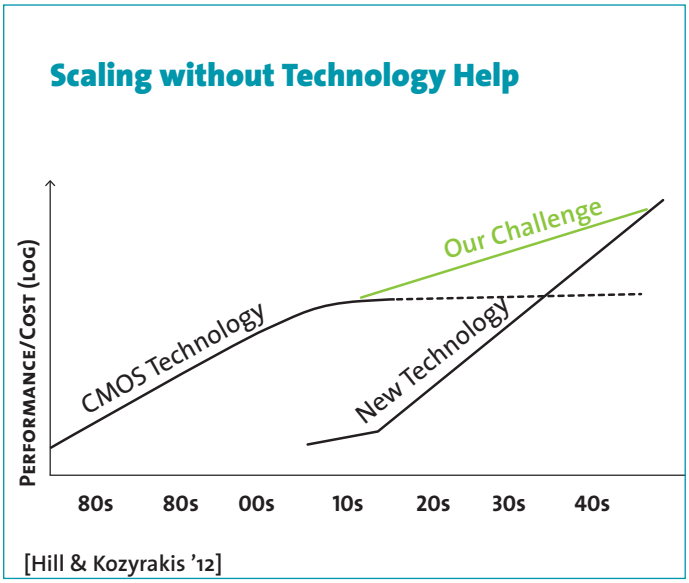
The relation between the hysteretic voltage-current diagram of memristive technologies and that of an artificial synapse has been put forward by several authors: Leon Chua in his seminal paper on the memristor *[Chua71]* and more recently the HP labs team *[Strukov08]*. Furthermore, it has been proposed that by combining the symmetric thresholds of a memristive device together with a spike-based coding, the induced resistance change could result in the implementation of a learning scheme known as STDP (Spike Timing Dependent Plasticity) *[Snider08]*. An important body of work has been triggered in an effort to implement dense arrays of synapse-like devices using emerging memory technologies. Similarly, organic devices have been demonstrated to have synaptic properties *[Alibart10]* and recently phase change memory was proposed for the same purpose *[Kuzum11]*. Beyond the fact that the STDP learning strategy implies a paradigm shift in information coding (from state-based to event-based), it promises easier implementations of very dense arrays and more energy-efficient systems for qualified classes of applications (e.g. sensory data processing, vision).

It is interesting to note that memory technologies such as PRAM, MRAM and RRAM offer many opportunities in other fields. As described above, they triggered the exploration of new digital processor architectures, but also of disruptive computing paradigms and coding (spikes).

In the context of spike-based coding and the related computing paradigms, big research initiatives in the field of neuromorphic computing are currently under way. Examples include the DARPA-funded SYNAPSE project coordinated by IBM *[SyNAPSE]*, and in Europe several ICT projects with FP6-FACETS, FP7-BrainScale, and parts of the Human Brain Flagship project (HBP).

POTENTIAL PERFORMANCE STAGNATION IN THE COMING YEARS?

As new technologies that can take over from silicon technology for increasing performance are not yet ready, and since the current silicon-based developments may further reduce the performance increase over the coming years, we can foresee a few years (nearly a decade) of performance stagnation.



[Hill & Kozyrakis '12]

The challenge will be to keep pushing performance improvements using current technologies, while selecting and increasing the maturity of new technologies.

STORAGE

Today, bits on a hard disk are stored in sets of magnetic grains. A magnetic grain is about 8nm, and it cannot be made much smaller because super-paramagnetism will cause random flips of the magnetic grains under the influence of temperature. One stored bit consists of 20-30 grains and has a bit width of 75nm and a bit length of 14nm. The number of grains cannot be reduced much if we want to keep a sufficient signal to noise ratio. Therefore, the maximal density of perpendicular recording is about 1 Tb/in^2. Today, hard disks with a density of 1 Tb/in^2 are commercially available, e.g. the Seagate Enterprise Capacity 6TB. Today's bits on a hard disk are therefore of the same order of size as the size of a transistor on a chip. The bit density can be further increased by reducing the bit (= track) width. The idea is that a track is written full-width, but the next track partially overwrites the previously written track (just like shingles on a roof, hence the name "shingled magnetic recording"). The remaining strip of the track is wide enough to be read, but we cannot write on it anymore without destroying the data in the neighbouring tracks. This leads to disks where data must be stored in bands. These hard disks have to be used like solid-state disks; bands must be written sequentially and cannot be changed, they can only be overwritten. Shingled magnetic recording can lead to a further density increase of about 25%. *[SMR]*

Beyond the limits of perpendicular recording, other approaches are needed *[PLUM11]*. One class of approach is energy-assisted magnetic recording, of which heat-assisted magnetic recording is the best known. It uses heat in combination with a magnetic field to record the bits. This, however, requires that a heat spot must be localised on a single track and that the rise and fall times of the media must be in the sub-nanosecond range. Designing such a head is challenging.

Another approach is to make use of patterned media. In patterned media, each bit is recorded on a small island of magnetic material, surrounded by a non-magnetic material. In this case, a bit can be made as small as a single magnetic grain (instead of 20-30 grains for perpendicular recording). In order to reach 1 Tb/in^2 in patterned media, we need to etch islands of 12 nm, which is beyond the resolution of current lithographic systems. That means that patterned media will have to rely on self-ordering. Densities of up to 50 Tb/in^2 seems to be theoretically possible with patterned media, if combined with heat-assisted magnetic recording. However, today, bit patterned media is not yet ready for the market.

Solid state drives seem to be the replacement for hard disks, but they are not:

- First al all, they are still one order of magnitude more expensive than hard disk storage ($0.50/GB vs. $0.05/GB).
- Secondly, there are simply not enough fabs to produce even a fraction of the magnetic storage capacity that is currently produced on an annual basis. So, even if the entire industry would decide to massively move to SSDs, there is not enough capacity to produce them.
- Third, write endurance is much lower than for hard disks. It used to be 100 000 cycles for SLC (cells storing one bit), and 10 000 cycles for MLC (cells storing multiple bits), but that was for 50 nm technology. In today's technology, write endurance for MLC is between 1000 and 3000 cycles, and in the next technology node, it is expected to be reduced even further.
- Fourth, data retention for flash is 10-20 years for new flash cells. For used cells (5 000-10 000 write cycles) data retention at 50°C is 6-12 months. Even reading stresses the cells in the same block, which means that SSDs that are written only once but read many times, will eventually start losing their data as well.

That means that SSDs are neither affordable nor reliable enough for data storage, and further shrinking beyond 20nm will make it even worse. At best, SSDs could be used as a cache, not as stable storage.

Possible alternatives for flash are Phase Change Memory, CBRAM, MRAM, RRAM, MRAM, FeRAM, Racetrack memory, NRAM, millipede memory (see section above on non-volatile memories) ... Of all these alternatives, Phase Change Memory seems to be the most promising candidate.

Features	Parameter	eFlash	TASMRAM	PCM
Technology	Storage	Charge	Magnetics	Phase Change
	Node	90nm	90nm	90nm
Write Access	Prog/Erase time (ns	5000	60	70
	Current (mA)	4	3	-
Read Access	Read time (ns)	35 to 50	15	135
	Current (mA)	2	2	-
Call Size	Per cell (um2)	0.166	0.133	Equ. to Flash
Retention	Retention (y)	10	20	-
	Endurance (cycles)	400k	> 10^16	> 10^6
Temperature	Max (C°)	125	250	85
Max Capacity		64Mbit	32Mbit	128Mbit
Complexity	Xtra mask layers	7	2	

A comparison of non-volatile memory technologies. Source: Crocus Technology [Crocus].

It can endure 10 million write cycles, and it is currently about 16x faster than MLC SSD for reads and 3.5x slower than MLC SSD for writes. The main showstopper is however the amount of power needed to change the state of the stored information. 440 µJ to program a 4 KB block, 630µJ to reset of 4 KB block. That means that a transfer rate of 100 MB/s will require about 10W just for storing the bits. That, in turn, means that PCM devices are currently limited by the amount of power they can get on the chip and dissipate. Other technologies are even less mature *[Tso14]*.

The conclusion is that magnetic storage seems to have hit physical limits, and that the solid state replacements like flash or phase change memory are not yet ready to take over its role. At the same time, storage space needs to grow exponentially due to the data deluge. This will have profound consequences for the future evolution of computing. Not only are we able to generate more data than we can process, but also more than we can store in an affordable way.

Today, there is no clear alternative for magnetic storage, and since any good alternative will require years to be made ready for the market and to become mainstream, we will probably experience a slowdown of Kryder's law for storage (on top of a similar slowdown for Moore's law for semiconductors). We are being slowed down by the challenges of reducing the physical size of stored bits (due to the limits of lithography), by the limits of reliability at small feature sizes, and by the recording power needed by alternative technologies.

At the same time, non-volatile memory will definitely find its way into existing computing systems and, there, it will lead to the need for revisiting software stacks. Likely areas of interest include: virtual memory, distributed shared memory, and relaxed memory models without coherence. Increasingly cheaper persistence should push for dramatic changes in I/O design and interfaces, hybrid memory architectures, data-centric execution models, and tying data and computations.

COMMUNICATION

It is clear that computing and communication cannot be treated independently. Examples stressing this interdependence go from sensor networks over mobile phones to cloud computing. Therefore, it is recommended that at least requirements from and interface to communication systems should be taken into account.

Even if the "always connected" motto is interesting and allows for adding more features to systems, it is not always possible "to be connected" and the system should cope with this. Chrome OS initially always had to be connected, but users rapidly figured out that this was not realistic, so editing text, writing and reading emails soon became possible even offline. Self-driving cars or autonomous devices should have enhanced features when connected, but they should still be able to perform most of their task when offline. Some areas will have no or bad reception of wireless signals, even via satellites. It will be expensive to cover all underground areas with full speed connectivity. Furthermore, wired or wireless communication can be intercepted or jammed.

Having all your data in the cloud is convenient because it allows accessing it from anywhere and from several different devices, if the cloud is reachable. Having a local backup or cache at home, however, will not only reduce the bandwidth to the cloud, but also provides access to the data in case of network or cloud problems. Recent outages of Google, Facebook and some major wireless phone companies show these kind of problems do happen in practice.

OPTICAL COMMUNICATION

The cloud-computing infrastructure is almost completely based on opto-electronical wiring. Where in the past fiber was mostly used for communication between data centers, in recent years it has also found its way into those data centers, replacing copper wiring. Current optical networks reach speeds up to 100 Gigabit/s per channel. Using WDM (Wavelength Division Multiplexing, a technique to send multiple wavelengths over the same fiber) bandwidths over 1 Terabit/s per fiber connection are possible.

Communication between data centers is mainly based on single-mode, 10 to 25 Gb/s per channel fiber technology, through the use of WDM in combination with advanced modulation techniques to achieve higher bandwidths. This technology is limited by several factors:

- Speed: dispersion in the fibers causes bit smearing. Dispersion is the effect that different colours of light have different speeds in glass. This effect becomes noticeable over long distances even with the small bandwidth of the light used for one channel. It can be solved through optical de-dispersion, or through signal processing at the fiber ends. Using signal processing to counter the dispersion effects couples communication speed improvements to Moore's law.
- Power: high-light intensities from high-power lasers cause non-linear signal distortion effects in fibers.

These limitations come from the combination of the medium (quartz glass) and the used wavelengths, and cannot be solved easily. Only disruptive technologies can bring exponential growth again, but nothing seems to be on the horizon. For now, 100 Gb/s technology per channel is in the labs, and will probably be used in the field several years from now. Further developments in WDM, advanced modulation techniques and enhanced protocols will further increase these speeds to 1 Tb/s in the near future.

WDM techniques scale communication speeds linearly: adding an extra channel of 25 Gb/s adds just 25 Gb/s bandwidth to a single fiber. Chip speed limitations apply equally to processors and to communication chips, limiting either speed, or requiring comparatively large power consumption. As a result, moving large amounts of data around over long distances is becoming expensive both in terms of time and of energy. One technique to address this issue is keeping the raw data close to the initial processing elements, and only communicating processed data (which is hopefully smaller).

Communication between systems, and between racks with systems, is based on both single-mode and multi-mode fiber technology. Multi-mode fiber technology is cheaper (relaxed mechanical precision on the opto-electronic and inter optic interfaces), but limited in distance (on the order of 100m) and speed (more dispersion than single mode fibers, 10 Gb/s is common per channel).

With the latest advancements, single-mode and multi-mode are becoming competitive. However, once built, the communication backbone of a data center is more or less fixed, locking the owner into a particular technology for several years. For example, Google uses multi-mode technology, while Facebook uses single-mode technology in new data centers. In intra-datacenter technology, a speed up of a factor of ten should be technically feasible now, but could require heavy investments.

Optical communication technology is also being developed for creating connections within systems, between components. Such components that directly connect the processing devices to optics are coming to market already, a trend that is referred to as "moving the (optical) connector from the edge of the PCB to the middle". This evolution also makes opto-electronic chip interconnects possible. HPC profits the most from this development; for other applications, copper interconnection is still more cost-effective. Finally, the application of opto-electronics communication is also moving inside the chip, but this is still a research topic.

COPPER-BASED COMMUNICATION

Optical communication is the communication technology of the future. However, over the past decades huge investments have been made in the copper-based communication infrastructure: telephone lines and coaxial cables for cable television. Replacing these copper-based networks by fibers is so costly that several techniques have been developed to increase the capacity of copper wires.

Two-wire telephone lines were designed to carry a signal with a bandwidth of 3.4 kHz (voice). The length of these cables is usually in the order of 5 km maximum. In other words, most telephones are within 5 km from a local switch office. Over the past decades, techniques have been developed to reach bandwidths up to several tens of MHz (30 MHz for VDSL2) over these connections, allowing speeds of up to 100 Mb/s. These speeds can only be achieved up to a limited distance though, usually 500m. Some experimental techniques have already been shown to reach data rates of 300 to 800 Mb/s over 400 meters in laboratory experiments.

Combining fiber technology and high speed two-wire transmission techniques allows for the installation of high speed networks at reduced costs, by bringing the optical communication end point as close as possible to the end user, and using VDSL-techniques to cover the last few 100 meters.

Coaxial cable television networks offer higher bandwidth by design, ranging from 750 MHz to 1 GHz. These networks were originally designed to offer one-way communication to deliver high quality television signals to the home, but have been equipped for two-way digital communication. The topology differs from the topology of telephone lines in the last segment to the end-user: telephone lines have a point-to-point connection between switch office and end-user, while in coaxial cable networks one cable segment services several subscribers. When used for two-way Internet communication, this means that in coaxial cable networks subscribers share bandwidth on a cable segment. However, line segment length is usually not limited due to the installation of amplifiers at certain points.

In the past, these technologies have allowed for what is dubbed Nielsen's law of Internet bandwidth: user's bandwidth grows by 50% per year *[Nielsen]*. However, we should bear in mind that Nielsen's law follows Moore's law, as it is mainly thanks to digital signal processing techniques that we have been able to exploit the available bandwidth of these signal carriers.

Telephone subscriber line and coaxial cable television networks, which together form the most pervasive copper-based communication networks, in combination with fiber optical communication technology have enabled us to quickly build high-bandwidth communication networks. This approach allows for a gradual transition to a global optical fiber network, and for achieving the broadband objectives of the Digital Agenda for Europe: "...download rates of 30 Mb/s for all of its citizens and at least 50% of European households subscribing to Internet connections above 100 Mb/s by 2020." *[whitepaper], [FTTH]*

WIRELESS COMMUNICATION

Connecting to the cloud and the Internet of Things is, and will be, done in several ways. Next to wired connections (copper or fiber), wireless connections also play a very important role. Mobile devices must use WiFi and 3G/4G cellular networks to connect to the Internet. For shorter distances, Bluetooth, ZigBee, and wireless-USB are used. Wireless communication is not only used for mobile devices, but also where wires are difficult to install or undesirable. As an example, some experimental satellites use Bluetooth communication instead of wires to interconnect various subsystems, avoiding potentially damageable wires or fibers *[delfi-c3]*.

The bandwidth of wireless connections depends directly on the frequency used: the higher the frequency, the higher the bandwidth (Nyquist limit). Unfortunately, the range of wireless connections varies inversely with the frequency. E.g., 5 GHz Wifi, which can reach speeds up to 6.93 Mb/s, can only do so within line of sight connections. Moreover, in wireless communication all devices in range can receive each other, which may cause interference. The radio spectrum is becoming increasingly crowded, requiring strict regulations for its various users to minimise interference and maximise bandwidth.

A solution could be the further development of Ultra Wide Band (UWB) technology. This technology uses a large part of the spectrum ("large" is defined as at least 500 MHz bandwidth, or more than 20% of the base frequency), as opposed to the narrow frequency channels used in almost all current wireless connections. As the communication is spread over the assigned band, it appears to other devices as low-power noise, resulting in much reduced interference. Cognitive radio is another promising form of radio communication. In this case, the transceiver systems actively search for the best wireless channel available to ensure optimal use of the spectrum, dynamically switching between channels when necessary.

Ad hoc storage and forward networks can be based on multiple wireless communication standards. This technology combined with UWB protocols holds the promise to connect all sorts of devices to the Internet of Things, especially for deploying devices in remote areas.

In recent years, wireless interconnects on chip have been proposed *[Gang11]*. Although not a truly wireless radio link, this way of interconnecting parts of a chip can potentially reduce the power required for communications by orders of magnitude, while outperforming wired networks-on-chip in throughput and latency. This technique is in an experimental stage, and requires more research.

With the advent of wireless communication, the environment is increasingly filled with electromagnetic energy. Energy densities have now become sufficiently high to allow for energy scavenging from the electromagnetic field to drive low power devices. This technique could enable large node sensor networks, where the sensor devices extract energy from the electromagnetic field to drive the sensor and communicate with the network *[TECS07]*. This does mean that transmitters may have to become more and more powerful or have more processing to compensate for such harvesting though, otherwise their range will decrease.

WIRELESS COMMUNICATION AND THE INTERNET OF THINGS

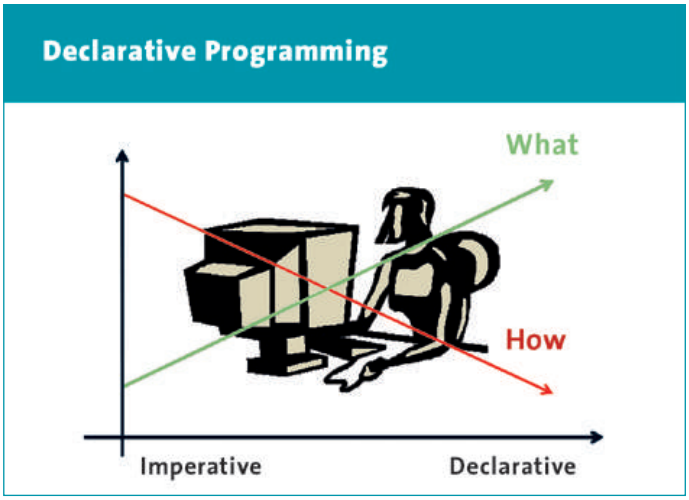
The current state of communication technology appears to be ready for the Internet of Things: there is sufficient technology available or under development to offer the required bandwidth, both for connecting sensors and actuators, and for connecting users, wired or wireless. However, low power communication systems still need to be further improved, mainly in the standby mode. Zero power standby is a must for the Internet of Things, where devices will not communicate constantly but from time to time, either on request of a master device or when they have relevant data to transmit. Having systems that consume zero power while waiting for a wake-up (radio) signal is important for low power. Similar requirements are for remote-controlled consumer devices, while they are turned off, which accounts for 5–10% of total household energy consumption.

Mobile communication providers have invested several times over the past 20 years in concessions and new equipment to move to the newest generation of mobile technologies. Competition can be fierce, eroding margins. With the advent of the Internet of Things, extra investment is required, not only in new base station equipment, but even more so in bandwidth on the wired side, which connects the base stations to the Internet. Because of the smaller margins, providers may take a while to make the necessary investments, resulting in a mobile infrastructure with marginal bandwidth at first *[FTH12]*.

FACING A NEW SOFTWARE CRISIS AND ITS COURSE OF ACTION

The major cause of the software crisis is that the machines have become several orders of magnitude more powerful! To put it quite bluntly: as long as there were no machines, programming was no problem at all; when we had a few weak computers, programming became a mild problem, and now that we have gigantic computers, programming has become an equally gigantic problem.

— Edsger Dijkstra, The Humble Programmer [Dijkstra 72],



THE PRODUCTIVITY CHALLENGE

The "software crisis" was a term used in the early days of computing science to describe the difficulty of writing useful and efficient computer programs in the required time. The software crisis was caused by the rapid increase in computer power and the complexity of the problems that could be tackled *[Softwarecrisis]*. Successful methodologies and abstractions have been designed over the years to improve software quality and development productivity.

The resulting progress in productivity threatens to be nullified by today's evolutions. Computing systems grow in complexity, diversity and ubiquity, and also become massively distributed, which all translates into additional challenges for software developers. Modern devices are subject to operating constraints from originally separate domains: the seamless interaction with other devices and the environment (Cyber-Physical Systems), rich human-computer interfaces with constantly connected, high-bandwidth and cognitive processing, life-long and dynamic adaptation, mission-critical and non-functional requirements, all involving hardware-accelerated data- and compute-intensive components.

Research and experience in programming languages has progressively introduced more declarative and less explicit syntaxes and semantics: from functional or equational abstractions to managed memory abstractions, data structures, combinators, polymorphism (or genericity) and other forms of code reuse. These successes do not alleviate the need for low-level software stacks and languages, but they continue to offer productivity increases and greater accessibility to programmers. This trend will continue, and should be encouraged further in the domains of parallel programming (current abstractions remain too limited, or too low-level) and of interaction with the environment (sensors, control). In particular, event-driven and low-level actor-oriented programming should progressively yield to higher-level reactive approaches. Low-level APIs should not be exposed to end-users, even the ones solving computational tasks or dealing with complex non-functional requirements. Better abstractions are definitely needed to continue this trend towards more declarative design and programming, together with built-in methods for abstraction-penalty removal.

To support the necessary abstractions, the need for better tools increases with growing system complexity, stronger operating requirements, and expectations of continuous performance improvements.

More than ever, mature, interoperable tools are needed, from compilers and functional debuggers to integrated development environments, design-space exploration, monitoring and performance debugging, modelling and verification of functional and non-functional properties, testing, verification, and certification. Without advances and investments in tools, the cost of leveraging the advances of heterogeneous parallelism is too great for most applications, due to the current immaturity of the development flows and techniques; the modelling, programming, validation and certification of Cyber-Physical Systems will remain out of reach for most engineers.

Progress in tools are required, otherwise the elastic distribution and scaling of data-intensive applications will be restricted to a narrow range of high-value markets and vertically integrated businesses, with limited algorithmic diversity.

Unfortunately, the lack of sustained investments in development tools has been a recurrent weakness of our organisations.

Tools are capital-intensive, high risk, and often plagued with technical immaturity and unrealised hopes and in a vicious circle induced by the weakness of the tool market, very low on the return on investment. A notable exception is the CAD market in VLSI design and test. Niche markets also exist, thriving on the necessary regulation in safety-critical application domains. Outside these, the free and open source model has been very effective at gathering the investment of large companies, engaging with developer communities in a mutualised effort. Nevertheless, for disruptive technologies to emerge on the market, this funding model is not sufficient: success stories in tools, whether free or proprietary, depend on the engagement of their users, and sustained funding from businesses depending on these tools.

To wrap-up, the productivity challenge will not be met without capital investments from users of productivity tools. These investments are also necessary to achieve any significant leverage on standardisation bodies, and to guarantee inter-operability and long-term maintenance.

THE CORRECTNESS CHALLENGE

In programming language history, the portability of performance has always been secondary to functional portability, also known as correctness. This paradigm has to be dramatically revisited, however, as the correctness of Cyber-Physical Systems directly relates to execution time, energy consumption and resource usage. In other words, performance in the broad sense is now part of the specification; it is a requirement and no longer an objective function. This is very well known in embedded system design, but is now becoming applicable to a much larger share of the software industry.

Furthermore, this increased sensitivity to timing and quantified resource usage comes at the worst possible time, as the successful methods demonstrated in embedded system design are precisely the ones that are put in jeopardy by the technological and architectural evolutions. One example of such issues is timing unpredictability, from the process level up to the chaotic interactions of bus-based cache coherence. The success of heterogeneous hardware even breaks functional portability: most accelerators feature ISAs departing from the Von Neumann model (including GPUs and FPGAs), and approaches like CUDA and hardware-oriented dialects of procedural languages proudly break decades of progress in software engineering (in modularity, reusability, portability). In the last decade, hardware has evolved much faster than software. We are in dire need of better programming abstractions and of automated or semi-automated tools to lower these abstractions to target platforms.

Let us consider a few examples:

1. The map-reduce paradigm of APIs like Hadoop thrives on classical functional abstractions for collective operations, such as map and fold, combined with the referential transparency and layout abstraction of functional data structures. However, map-reduce remains limited to certain classes of algorithms and (purposely) limited in the expression of parallel patterns. For example, parallel patterns with dependent tasks are capable of expressing more communication-efficient computation pipelines and algorithms *[StarSs, OpenStream]*. The strength of such approaches is to enable the software engineers to reason about "what" rather than "how" to do things. It also pushes for the adoption of domain-specific approaches and languages in production environments, and for the maturation of development tools supporting these approaches.
2. Other paradigms such as reactive programming also deserve better abstractions, given the intrinsically distributed, reactive nature of modern applications (reactive to communications, the environment, and through human interfaces). While reactive language designs exist, the common practice in reactive application development remains very low level, with callbacks and explicit event handling, as is typically found in game engines, graphical user interfaces, web services, SystemC. More abstraction and automation will not only improve productivity, it will open new areas of scalability, resource optimisation and portability. Reactive programming is not declarative programming, and functional reactive programming languages are complementary to existing languages (C, C++). Algorithmic programming will remain, but reactive thinking is currently missing.

   Cyber-Physical Systems also push for programming languages where time is a first class citizen, including time-triggered semantics, reaction to timed events, watchdogs, controlled execution time of specific reactions (best and worst case), support for comprisable analyses of execution time, and compiler optimisations for the worst-case execution time rather than for the common case. With the convergence of interactive, embedded, and computational applications and systems, these time-friendly features must coexist with the functional abstractions, modularity and time-independence of general-purpose languages.

   Finally, with ubiquitous connectivity, dynamic features, computational components collaborating with secure and mission-critical ones, programming languages and tools need to take security and safety properties seriously. They need to provide means to enforce functional and non-functional isolation, to implement effective countermeasures and intrusion detection mechanisms, and to be traceable for certification and analysis purposes. Understanding what this convergence of concerns and semantical properties really means remains essentially open, at this point. Note that as for any software component with a critical aspect, specific regulations may be needed to guarantee the soundness of the certification processes and to guarantee a minimum level of transparency in the implementation for independent analysis.
3. We may also want to reconsider the meaning of numerical correctness. Sometimes approximate results are as good as highly precise results. The current momentum of approximate computing research is behind low-level optimisations that trade precision for performance and efficiency, from mixed-precision numerical methods to micro-architectural proposals *[Esma12ACM]*. There are, however, orders of magnitude to gain by working on the specification and algorithms in tandem, crosscutting the application stack with a multi-disciplinary approach. A holistic design approach is also necessary to enable high energy savings: software engineers are currently "blissfully unaware" of energy issues, but they will need more "energy transparency" from the language to the hardware-software interface in order to design efficient systems in the future. Abstractions are again needed, as well as tools implementing the required automated procedures for static and reactive adaptation.

Adequate precision programming may also be one of the approaches that can be used to cope with timing and energy requirements; in particular, it may be easier to design a safe system and to prove its safety by making it capable of reacting to a wide range of timing variations, possibly with associated graceful degradation of the Quality of Service, rather than attempting to control all possible variations at design and static analysis time.

THE PERFORMANCE CHALLENGE

Because performance is not portable, developers need to take into account the target platform and its execution environment. They always did, and this was highlighted by Dijkstra as one of the initial causes of the software crisis *[Dijkstra72]*. What we did to address the initial software crisis was to start caring more about humans and less about machines. Powerful abstractions were designed to make developers more productive, relying on automation (compilers, runtime execution environments and operating systems) to bridge the gap with the low-level hardware interfaces. Methodologies, languages and tools to tackle complexity have been hugely successful, leading to the thriving software-dominated technological world we live in today. Most programming abstractions assume a Von Neumann architecture, possibly threaded to some degree.

These abstractions map relatively well on general purpose processors, but their suitability for advanced modern platforms (heterogeneous, reconfigurable, distributed, ...) is being challenged. For example, object-oriented programming principles thrive in a threaded Von Neumann environment, but are almost entirely absent from hardware accelerated, massively parallel or specialised hardware. The lack of static typing, binding, and referential transparency in object oriented methodologies is a no-go for a parallelisation of compilers and for restricted, specialised hardware. Furthermore, the best established software engineering practices can even be counter-productive when performance is a non-functional requirement: cross-component optimisation, inlining and specialisation break portability, modularity and code reuse, if not automated and made transparent to the application developer.

Right now, however, decades of progress in programming languages, software engineering and education seems to be nullified because of hardware-software interface disruptions. As a result, the crisis also looms in the interaction between development teams with different expertise and procedures. It equally emerges from the interaction between diverse programming languages and runtime systems (e.g., MPI + OpenMP + CUDA, task parallel runtimes + virtualisation). The clever automated techniques implemented in compilers and execution environments are not designed to deal with disruptive changes in the hardware-software interface.

Automated code generation techniques are needed. These techniques should provide abstraction without performance penalty: a well-known programming language dilemma that constantly needs to be revisited, following hardware evolutions. In practice, the kind of modularised, well-defined components that may be good for software engineers are totally orthogonal to the components needed for parallelisation and efficient resource usage. A real challenge is to allow developers to continue to modularise programs with a code reuse and productivity mindset, without making it impossible for automated tools to implement cross-module optimisations and thoroughly repartition the application for efficient execution on a heterogeneous, parallel target.

This kind of decoupling of programmer abstractions from platform-dependent mapping is exactly what Model-Driven Engineering (MDE) research aimed to demonstrate for many years. The lack of precise semantics of general-purpose modelling frameworks unfortunately made this impossible. We should address this by learning from rigorous formalisms and tools, such as synchronous languages used in the correct-by-construction design of safety-critical systems *[Lustre]*.

THE DATA CHALLENGE

The IT world becomes increasingly data-centric, while memory, storage, and interconnect technologies reach scaling limits. This paradox will induce massive changes in the software stack and consolidation in the industry *[Farab14Conf]*. At the same time, non-volatile memory will definitely find its way in existing computing systems, and they will also push to revisit software stacks. Likely areas of interest include virtual memory, distributed shared memory, relaxed memory models without coherence, data-centric execution models tying data and computations. Increasingly cheaper persistence should push for dramatic changes in I/O design and interfaces, and in hybrid memory architectures. On the other hand, the uncertain industrial maturity of new memory and communication technologies will delay their concrete impact. Researchers and engineers will have to live through a world of fragmented and trial-and-error adoption of these technologies, creating scientific and business opportunities, but under a tight energy cap. Research and innovation should break out of the incremental refinement and tuning of existing system architectures and layers. Technology reaching a scaling plateau will push for more efficient, leaner and specialised solutions all over the software stack.

In a data-dominated computing landscape, one goal is to enable high-performance for scale-out, heterogeneous parallel and distributed systems without sacrificing programmer productivity. Applications must be made adaptable to runtime environment changes, and to evolutions in computing system architectures. This ambitious goal depends on the ability to make applications portable and elastic. This is especially important in environments such as mobile devices where power constraints can force the application to migrate to low-power cores or cloud services, where the amount of resources fluctuates depending on the workload. Another scenario is where we move computational tasks closer to high-bandwidth sensors to reduce the

communication cost and to enable upstream data integration (e.g., cognitive cameras, augmented reality).

These evolutions motivate research and innovation in the area of process and system virtualisation, just-in-time compilation, binary-level optimisation, dynamic orchestration, fault tolerance and monitoring, and programming models with a global address space for elastic computing.

The data challenge is also characterised by the emergence of new application domains and compute-intensive problems, pushing the limits of existing tools and abstractions for high-performance computing. There is already a need for high-performance libraries and domain-specific languages to support these new applications and their key algorithms. In these fields, security and data integrity are cross-cutting concerns, interacting with all aspects of hardware, system, and development methodologies.

Such complex applications will require the collaboration of domain experts. For example, the design of advanced user interfaces can benefit from a close interaction with people having backgrounds in ergonomics and in behavioural and medical sciences. The safe interaction with the physical world through sensors and actuators requires a good knowledge of system theory and signal processing. Applications for health monitoring will naturally require the help of medical professionals, etc.

THE HOLISTIC CHALLENGE

The past decades have seen growing, but fragmented development ecosystems, with standardised interfaces to connect system parts. Each ecosystem has its own set of optimisations, but this setup, as well as a more integrated approach has its limitations:

- Optimising systems locally is insufficient: individual systems in general will not be sufficiently powerful to perform all necessary calculations (e.g. sensor networks). On the other hand, some form of local preprocessing will always be required as otherwise the amount of data that needs to be transferred will overwhelm the communication links, or at least consume inordinate amounts of bandwidth and power.
- Global optimisations similar to the integration performed by current cloud providers will be infeasible due to the fragmented nature of the systems in terms of ownership, control, security and privacy concerns and the proprietary architecture of the devices. Virtualisation has a lot of nice properties, but it also hides too much of how the system works in view of global optimisation.

We will have to come up with a new holistic approach that deals with all of these concerns in order to be able to improve the efficiency of large-scale distributed systems.

A basic requirement for a holistic approach is a large degree of interoperability between all systems, so that the optimisations can enlist the cooperation of as many involved systems as possible, to an as large extent as possible. This calls for an increased standardisation effort.

Within single systems, we need APIs that enable cross-layer optimisations. Only looking at the hardware or software is not enough. For example, software, including the cross-system optimisation layer, may need to know the relative power usage of a particular kind of processing versus transmitting data in order to determine the most efficient way to proceed. In particular, this means that the software layer needs access to detailed probes at the hardware level that provide it with information about power usage. This information cannot be statically encoded in the software, not only because of portability concerns but also in the face of increased hardware ageing effects that change its properties, and because the dynamic nature of the cloud makes energy consumption difficult to predict.

The increased amount of cooperative coordination-related communication among systems results in extra security concerns: systems have to protect themselves against both attacks on their own integrity and being induced into unwittingly attacking other systems (DoS). They also have to balance the optimal functioning of the network as a whole with their own QoS requirements and resource capabilities.

DOMAIN-SPECIFIC LANGUAGES

Domain-Specific Languages (DSLs) are one important approach to reconcile productivity and performance, and to facilitate the interaction between design and development teams with different competences and practices. The DSL approach has been around for some time, with great successes like SQL and XML-based languages for data bases and web services, and many shortcomings when it comes with adoption, with the maturity of the tool chains to construct DSLs and to use them in complex projects, and with the abstraction penalty of managed languages (dynamic typing, concurrent garbage collectors). The appeal for DSLs is strong in signal and image processing *[SPIRAL]* and extends in many areas, including big data analytics [Blainey]. But the construction of effective domain-specific frameworks remains a challenge *[Halide]*, and domain knowledge and optimisations only address part of the challenge. Worse, the interaction between components designed with different domain languages leads to a crisis in its own right: complex software written in tens of programming languages leads to highly inefficient software stacks and break cross-layer optimisations. Furthermore, it is also impacting programmer productivity because there are very few development tools that can effectively deal with cross language issues (optimisation, debugging, interoperability ...). *[Chafi]*

STANDARDS

Although the programming languages landscape could change in the future, it is relevant to remark that the situation has remained quite stable over the past years with only minor fluctuations *[IEEE Spectrum]*. In particular, the conventional solution of designing a new programming language has consistently failed to solve the existing problems and achieving mainstream use. In contrast, the approach of constructively engaging in the evolution of already existing mainstream programming languages proven as useful in industry, seems to be much more cost-effective and seems to be a more pragmatic path to follow. Companies will not take risks in investing in a language that will not "survive" in 20 years, unless it is much better than the existing ones and it is backward compatible (legacy).

Programming language and tool design needs to consider multiple dimensions, including stability over time, market trends, adoption and education. Europe is currently failing to take advantage of an opportunity to influence the future of software development practices, despite the numerous successes of European-based research in the programming language field. The evolution of the most widely used programming languages, such as Ada, C, C++ and JavaScript, are controlled by formal standards bodies, but European participation is limited. These committees deal with highly technical issues and involve world-renowned engineers and scientists. European experts including domain experts from the embedded systems area, data centers and mobile computing should participate much more actively in standardisation committees. Supporting tool developments in these areas is also a precondition. Major industrial users of programming languages and tools should take their full responsibility and contribute to the soundness of the software development and tool ecosystem. More than ever, the tool industry is looking for stronger partnerships with application and technology providers.

STATIC VERSUS DYNAMIC

We may want to revisit the classic tradeoff between static and dynamic optimisation. As the granularity of the operations grows in time and/or space, dynamic decisions in runtime execution environments tend to perform better than static decisions in a compiler. This tradeoff is a common source of interesting language-compiler-runtime co-design problems in task-parallel languages. It is important to reconsider existing solutions in light of two important evolutions: (1) memory access and communications operations are orders of magnitude more energy-hungry than plain computations on modern hardware, and (2) the diversity in ISAs, micro-architectures and hardware-software interfaces is increasing after decades of business-driven concentration. Optimisation does not mean offline static only: configuration-time, just-in-time and online optimisations can involve aggressive compilation, with performance motivations (specialisation, adaptation to the target) and not only for portability (unlike, e.g., Java). While virtualisation is generally seen as a necessary evil by performance-motivated engineers it can also create more freedom for delayed optimisation that takes advantage of precise information from the running context. This is obvious in the design of CUDA and OpenCL, for example, and led to the inception of the split compilation paradigm *[Diou10], [Cohe10], [Nuzm11]*. It will likely spread to other domains, supporting the seamless and performance-portable migration of code across a variety of distributed devices.

THE POSITION OF EUROPE

In this section, we bring a SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis of the European computing systems community.

4.1. STRENGTHS

STRONG EMBEDDED ECOSYSTEM

The European computing industry has a strong embedded ecosystem spanning the entire spectrum from low power VLSI technologies to consumer products. Companies like ARM and Imagination Technologies are leaders in providing semiconductor processing elements and IP for embedded systems. IMEC and CEA-Leti are leading the development of semiconductor technology, while ASML is the leader in photolithography equipment. Large end-user European companies have a strong market presence internationally in areas like automotive (Volkswagen, Renault-Nissan, Peugeot-Citroën, Fiat, Daimler, BMW, Volvo), aerospace and defense (Airbus, Dassault, Thales, Saab, Barco), telecommunications infrastructure (Ericsson, Intracom), telecommunications operators (Deutsche Telekom, Telefónica, Orange), system integrators (Thales, Siemens), Semiconductor companies (Nokia, NXP, ST, Infineon) and software services (SAP). These companies are all globally competitive and work at the forefront of system design and implementation.

These larger players also rely on a thriving community of SMEs that strengthen the technical and innovative offers in the market. This strong embedded ecosystem creates a fertile environment for innovation, better integration and worldwide leadership. The SME ecosystem also spans the complete spectrum.

Europe has many fabless semiconductor companies, such as Kalray, Recore Systems, Think Silicon, Clearspeed Technology and NovoCore Ltd. These are supplemented by companies that build tools and carry out consulting for these tools like ACE, Maxeler Technologies, Vector Fabrics, Codeplay, CriticalBlue, Ylichron and Leaff Engineering. Some companies are active in the real-time and safety critical domain, such as OpenSynergy, Sysgo, Rapita Systems and Yogitech; others do formal verification, like Monoidics, or security, like Herta and INVIA.

We have also seen a decline of the previously big and vertically integrated European companies like Philips, Siemens and Thomson. Unlike companies like Samsung, they "de-verticalised" by spinning out their different business, for example their semiconductor businesses. The resulting companies, NXP and Infineon, reduced the scope of their activities and their ranking dropped amongst major semiconductor players.

PUBLIC FUNDING FOR R&D AND TECHNOLOGY TRANSFER

Europe benefits from a large pan-European centralised research program through the Framework Programmes for Research and Technological Development. If used properly, these serve as a powerful tool to direct community-wide research agendas and create a strong incentive for diverse groups of companies and research institutions to work together across political and cultural divides.

The value of the resulting consortia generally does not lie in bringing a particular new product or technology to market in the context of a project, but rather in investigating the potential of a new technology in terms of applications, business opportunities for individual partners, and creating new ecosystems and markets.

EXPLORATORY WORK, BASIC RESEARCH

Basic research and exploratory work are high-risk activities for industry, and also harder to fund for universities than is more applied research. At the same time, fundamental innovation is one of the activities with the highest possible return on investment, due to its potential to create whole new markets and product niches. It is therefore of paramount importance to strongly support such undertakings, and public funding is an excellent way to lower the financial risk by steering them towards selected domains and research directions.

BOOTSTRAPPING NEW BUSINESS ECOSYSTEMS

A product by itself seldom creates a new, large market. The real value of many products lies in the fact that they can be used in many different ways, combined with many different other products, integrated into existing workflows. These so-called network effects, whereby value increases through third party products and services making use of and inter-operating with the original product, are particularly important in today's connected world.

Enabling such network effects requires setting up ecosystems, which often require expertise from many different horizontally specialised players: hardware design, development tools, programming models, drivers for third party hardware, operating system support, etc. Publicly funded projects are excellent drivers to bring all these players together and help funding the necessary support and integration. The resulting environments, often at least partly open, encourage the creation of startups by offering a stable foundation to build on.

ONE OF THE BIGGEST MARKETS

According to the World Bank, the world GDP in 2013 is distributed over the major geographical areas as follows:

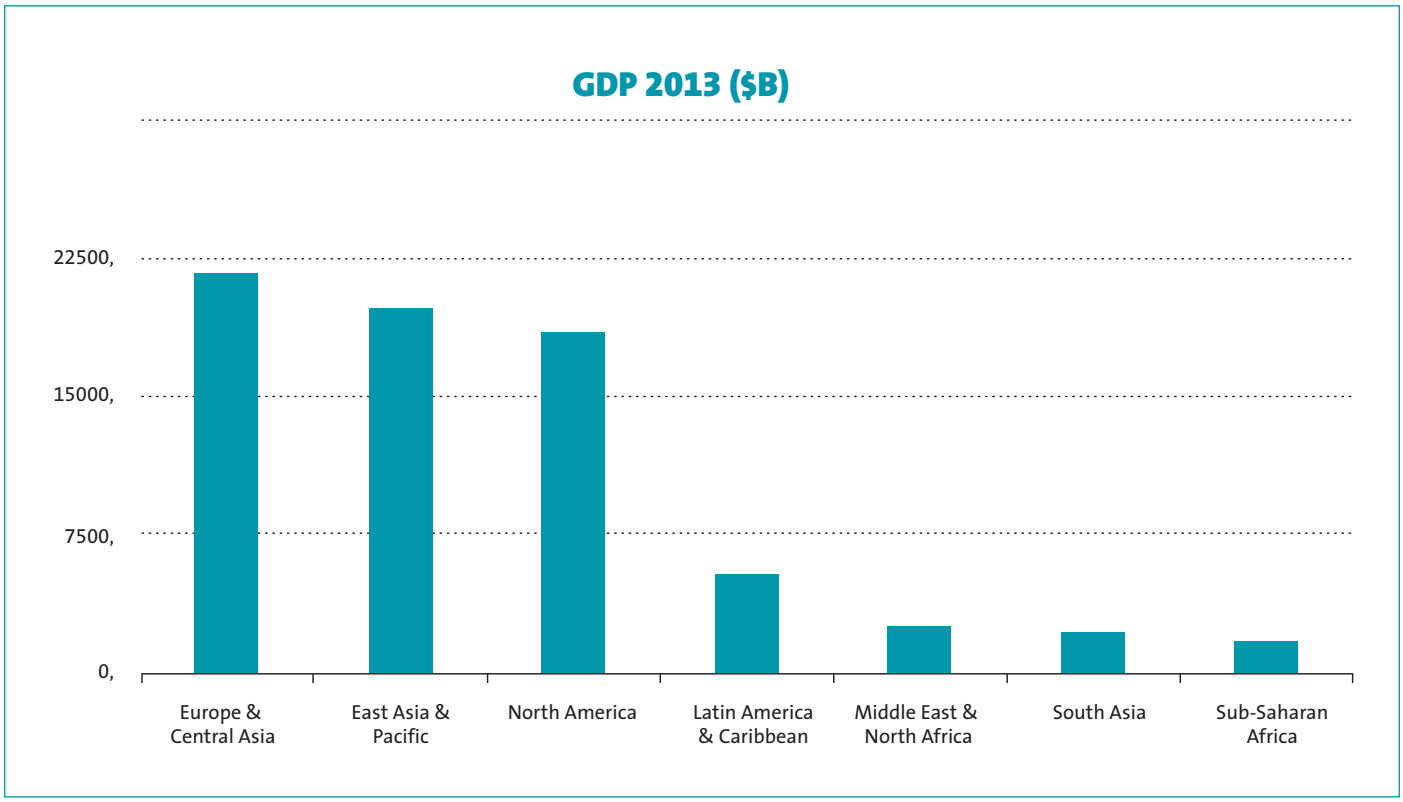
From this table it is clear that Europe and Central Asia are the largest economies in the world with a growth potential in recently joined member states, and in large neighbouring markets (Central Asian States, Middle East, Africa). With a high GDP per capita, there is a lot of money to spend on innovative products and services.

GOOD EDUCATION

From an educational perspective, 205 European universities are rated among the top 500 universities in the Shanghai University ranking *[ARWU2014]*. This is more than any other continent in the world.

Region	Top 500
Americas	177
Europe	205
Asia/Oceania	113
Africa	5
Total	500

Europe benefits from a very strong educational environment and a highly competitive undergraduate and graduate educational system. The ongoing bachelor-master transformation and the fact that an increasing number of degree programs are taught in English, will further strengthen the European educational system.



4.2. WEAKNESSES

EUROPE IS FULL OF HORIZONTAL SPECIALISATION

Very few European companies offer a completely integrated vertical stack that they control from top to bottom (hardware, software, services). This sometimes makes it hard to compete with vertically-integrated US and Asian giants that can easily lock-in customers from the point that they buy a particular device, or at least entice customers into using the manufacturers' own solutions.

LOSS OF COMPETITIVENESS IN SOME DOMAINS

European players own fewer patents on emerging standards. While patents on standards are a controversial topic, a constant factor is that whoever owns patents which are part of widely accepted standards can make a lot of money, and others will have to pay. Looking at previous MPEG video and audio compression standards and comparing them to upcoming ones such as H.265, there is a clear shift in patent ownership towards Asian companies. Another weakness is the decreasing number of state-of-the-art foundries in Europe, which results in a decrease of know-how about making advanced complex circuits and a shift to fabrication outside of Europe. In the future, Europe might depend upon strategic decisions made outside of Europe.

BORDERS AND DIFFERENT LANGUAGES

Language and cultural diversity in Europe are handicaps to attracting bright international students to graduate programs outside of their home country. The lack of command of English by graduates in some countries also greatly hampers international networking, collaboration and publication. The fact that multiple scripts are being used throughout Europe (Latin, Greek, and Cyrillic) also makes integration more difficult. It would help if all students leaving high school were able to speak English at the C1 level (C1 is the level just below that of native speaker, which is the highest level (C2)).

WEAK ACADEMIA-INDUSTRY LINK

European computing research is characterised by a weak link between academia and industry, especially at the graduate level. Companies in the United States value PhD degrees much more than European companies, which often favour newly graduated engineers over PhD graduates. This leads to a brain drain of excellent computing systems researchers and PhDs trained in Europe to other countries where their skills are more valued. As a consequence, some of the successful research conducted in Europe ends up in non-EU products, or does not make it into a product at all.

EUROPE IS WEAK ON COMMERCIALISATION

Many technologies that are now commercialised by companies in the US and Asia were originally developed in Europe, but did not lead to commercial exploitation in Europe. The lack of a venture capitalist culture is definitely one of the causes. Banks are also reluctant to invest in new companies, especially after the 2008 crisis. It is much harder for a university or PhD graduate to start a company in Europe than in the United States. Yet, even with venture capital, the personal investment and identification of startup employees with the fate of the company attitude found in Silicon Valley startups is largely absent from the European work ethos and organised labor agreements. On top of this, bureaucracy and administrative procedures in some countries are preventing or killing several new initiatives. It is clear that Europe will have to become more entrepreneurial in the coming decades.

EUROPE LACKS A SILICON VALLEY

Silicon Valley is the place to be for would-be technology entrepreneurs. All resources to start a new business (venture capital, work force, network, expertise ...) are concentrated in the valley, leading to a unique ecosystem that generates new businesses. In the US, it is only rivalled by the Boston Area, and to a lesser extent by Austin, Texas. In Europe, there is no such hub, but every country/city seems to invest in its own small hub and ecosystem for technology entrepreneurship.

Both models have their pros and cons. In the US-model, there is a technological desert between the two coastal areas, while in Europe the technological companies are more spread out over the continent, and we have more (smaller) hubs (Munich, London, Paris, Barcelona ...). It is however unrealistic to assume that these smaller hubs will eventually be able to compete with Silicon Valley. They are just too local and too small. If we really want to compete, we will have to start building a major European technology hub, but that seems to be next to impossible, given the current organisation of Europe. There are, however, cases where this model was possible, and paid off. CERN is one of these. The Toulouse area, with Airbus, is another.

4.3. OPPORTUNITIES

COST EFFECTIVE CUSTOMISATION

Building a state-of-the-art fab is very expensive and requires a 95-100% load to make it profitable. In the previous years, most of the European Silicon makers have become fab-less or "fab light" (NXP, Infineon ...), and don't plan to keep up with the most advanced technology processes. Building modern monolithic SoCs requires a $0.5-1B investment or more, which can only be amortised with millions of pieces per year. This is very difficult for Europe, but we have previously seen that 2.5D technology (Silicon Interposers) could leverage current European fabs while offering a way for European companies to differentiate and integrate. This technology can also leverage the European know-how in embedded systems, MEMS and other integrated sensors and devices for building advanced composite circuits with an affordable design start cost ($10M vs. $1B).

Some roadblocks are still present: for example, building interoperable European components for assembly on a silicon interposer,

etc. Creating a European ecosystem, where all participants could complement each other with their own technology and solutions, is certainly a solution but it will require a strong willingness to break the European ivory towers and to have true collaboration and complementarity between stakeholders. Europe could act as the "man in the middle" between the few highly advanced processor manufacturers and the memory providers.

Perhaps, Europe can invest on programmable/reconfigurable industry (e.g. encouraging an European FPGA company to be leader like Altera and Xilinx).

Europe should not try to catch up on existing mainstream technologies, but to use the large ecosystem of research in Europe to explore alternative solutions to silicon. As the fab train has passed, Europe should not try to catch that train. It should focus on the emerging technologies. This could enable Europe to become the worldwide leader in such future technologies. Therefore, there should be a swarm of European SMEs working innovative technologies of one type or another.

LEVERAGING FREE/CHEAP/OPEN INFRASTRUCTURE

Open Source and Free Software has spread to virtually all corners of the computing market. This evolution can be, and in fact already is, leveraged by companies in order to lower the cost of implementing and maintaining a basic infrastructure that is not part of their core business. They may need the resulting tools themselves, or may want to build ecosystems around their products. Such ecosystems themselves offer many opportunities for starting new businesses, both in terms of directly using the software (hosting providers, social media, wireless routers, etc.) and in terms of offering consulting services to adapt such software for particular purposes.

This evolution is also starting to take place in areas like content/services (e.g. the UK-based OpenStreetmap) and hardware (e.g. the UK-developed Raspberry Pi, and the Dutch Fairphone project).

SOCIETAL CHALLENGES

Paradoxical as it may seem, several challenges faced by society are also huge opportunities for research and industry in computing systems. For example, the European ageing population will require the development of integrated health management and support systems that enable people to live at home for a longer time. Europe's national health systems are far better placed to take advantage of and coordinate such efforts than those of the United States or China. European expertise in low-power and embedded systems and its SME ecosystem is an asset for tackling other grand challenges such as the environment, energy and mobility. Experience in mission critical systems gives Europe a competitive advantage with the safety and security challenges that lie ahead in larger-scale consumer systems. Being a continent with few natural resources creates the need to develop technologies like urban mining, solar and wind power ...

CONVERGENCE

Disruptive technologies like cloud computing, and the convergence of HPC and embedded computing, represent opportunities for Europe as well. The trend towards more distributed environmentally integrated Cyber-Physical Systems could be beneficial for the European semiconductor industry, which has significant expertise in the wide range of required technologies. Europe has a recognised strength on embedded systems and real-time systems, and this knowledge is a key element for the Cyber-Physical Systems that interact with the world. Real-time systems and mixed criticality are domains that need further development and have a large market potential.

MICRO- AND NANO-ELECTRONICS

The recent move to classify micro- and nano-electronics as key enabling technologies [KET] for Europe creates significant opportunities for the computing systems industry in Europe. This move comes in response to Europe's ICT manufacturing share of the total ICT added value in Europe dropping from 12.2% to 9.4% between 2006 and 2010 *[PREDICT13]*. Such large and centralised research and development planning presents a great opportunity for computing systems research in Europe, bringing it on a par with other technologies such as energy, aerospace and automotive technology. The resources available for developing pan-European research capabilities could be used to address several of the weaknesses mentioned above, in particular the lack of tools and hardware development expertise. The emphasis on collaboration with industry, in particular SMEs, can provide mentoring and bridges for researchers wishing to industrialise their results, if used properly.

4.4. THREATS

COMPETING WITH FREE/CHEAP/OPEN INFRASTRUCTURE

SOFTWARE: FREELY AVAILABLE SOFTWARE IS "GOOD ENOUGH"

When software that is available for free has functionality similar to commercial offerings, the latter can become a hard sell. Even if a commercially supported version has more functionality, provides a better interface or comes with guaranteed support, it can be hard to convince current and prospective customers that these additional offerings are worth the financial premium at purchasing time. Moreover, in case the alternative is Open Source or Free Software, the ability to continue development should the original developers disappear is an advantage that is hard to beat.

HARD TO SELL DEVELOPMENT TOOLS

Since development tools are a fundamental infrastructure required by many companies, much of the commercially sponsored Open Source and Free Software development happens in this area.

Therefore the "freely available software" evolution particularly challenges companies whose core business consists of selling development tools. Moving towards adding value on top of free tools, and/or convincing customers that the added value of their own tool chain is worth the licensing cost, is far from trivial.

SERVICES: GOOGLE IS "GOOD ENOUGH"

The wide variety of services offered for free by Google and others can cannibalise business and opportunities for other companies. For European companies it is very hard to compete with super-efficiently organised US-based global companies. At the same time, companies that rely on these services to run their business are at risk, due to the fact that such services can be terminated at any time without notice and that their availability is not guaranteed.

HARDWARE: SUBSIDISED ABROAD

Developing hardware is considered strategic for some countries (e.g. China). Therefore, some governments directly or indirectly support development of microprocessors and other advanced digital circuits in order to gain strategic independence from external providers.

OPEN SOURCE/FREE SOFTWARE LICENSES AND BUSINESS MODELS

Not all Open Source and Free Software licenses are compatible with all business models. In particular the GNU General Public License (by design) can make it hard for proprietary software vendors to leverage code. Releasing the results from publicly funded projects under an additional license that makes such integration easier, such as the Modified BSD or MIT license, can help alleviate these problems. A parameter-designed solution may be useful, like the DESCA templates for FP7 consortium agreements.

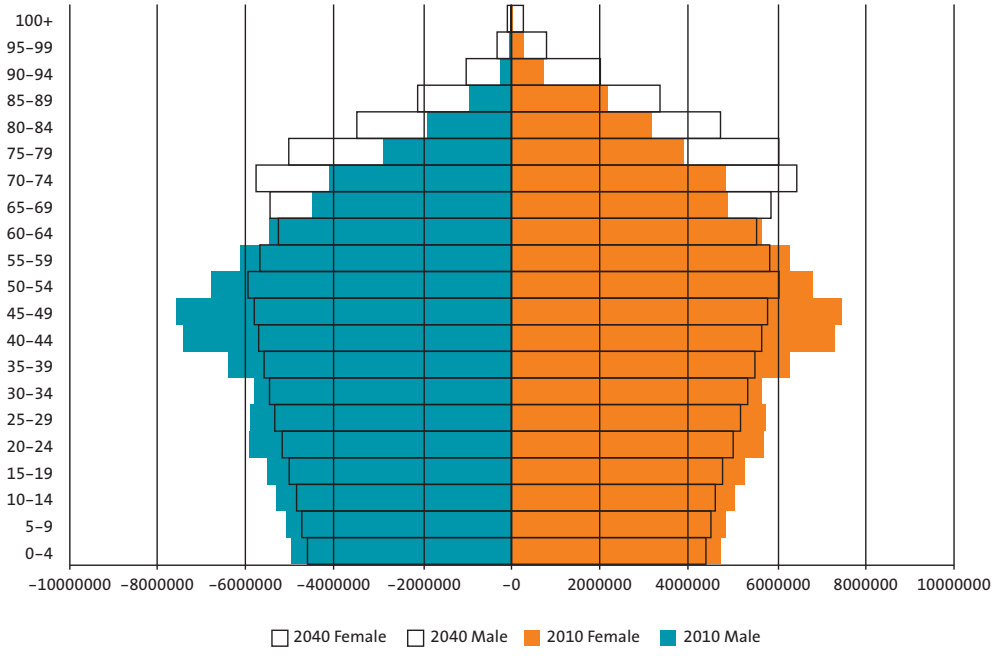
FINANCIAL CRISIS

The financial crisis has caused many companies to focus more on marketing and sales instead of R&D. Several governments are also reducing education and public research funds in their efforts to balance their budget, and to reduce the global debt to 60% of GDP. As it stands now, recovery is slow even in the economic strongholds of Europe. This situation, which has been ongoing since the financial crisis of 2008, is a serious threat to innovation in European companies.

AGEING POPULATION AND SHRINKING WORK FORCE

The population in Europe is growing older at a fast rate.
As it grows older, the work force is also shrinking, which means that there will be fewer hands and brains to create value in the economy. The effect is reinforced by the fact that young people tend to study longer, delaying their entrance into the labor market. Raising the retirement age can compensate for this, but not 100%. Hence, unless productivity increases, the economy may stagnate. The ageing population is definitely a threat to economic growth.

Population Structure: Western Europe, 2010 vs 2040 (projected)



Note: "Western Europe" here defined as EU-15 plus UK, CH, NO and smal adjacent islands and territories.
Source: US CENSUS IDB, available at http://www.census.gov/population/international/data/idb/informationGateway.php

GLOSSARY AND ABBREVIATIONS

Amorphous computing	A computing model similar to physics particles that have a state at a time, which is communicated to, influenced by, and influences the other particles.
API	Application Programming Interface
ASIC	Application-Specific Integrated Circuits are integrated circuits designed for a particular purpose, as opposed to being applicable for general use in many different situations.
Bayesian computing	Bayesian computing refers to computational methods that are based on Bayesian (probabilistic) statistics.
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.
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.
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
Declarative programming	Declarative programming is a programming paradigm that expresses the logic of a computation without describing its control flow. Many languages applying this style attempt to minimize or eliminate side effects by describing what the program should accomplish, rather than describing how to go about accomplishing it (the how is left up to the language's implementation). The opposite concept is imperative programming.
EUV	Extreme ultraviolet lithography is a next-generation lithography technology using an extreme ultraviolet (EUV) wavelength, currently expected to be 13.5 nm.
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).

FinFET	The term FinFET was coined by University of California, Berkeley researchers (Profs. Chenming Hu, Tsu-Jae King-Liu and Jeffrey Bokor) to describe 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).
FPGA	Field-Programmable Gate Array
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.
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.
Homomorphic encryption	Homomorphic systems send encrypted data to an application (generally executed on a remote server) and let application perform its operations without ever decrypting the data. As a result the application never knows the actual data, nor the results it computes.
ICT	Information & Communication Technology is a generic term used to refer to all areas of technology related to computing and telecommunications.
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.
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.
ISA	An Instruction Set Architecture is the definition of the machine instructions that can be executed by a particular family of processors.
JIT	Just-In-Time compilation is the method of compiling code from source or an intermediate representation at the time when it will execute. This allows for improved portability by generating the correct binary at execution time, when the final target platform is known. JIT compilation has been heavily leveraged in Java, Microsoft's C#, and OpenCL.
MEMS	Microelectromechanical systems refer to the technology of very small electromechanical devices. MEMS are made up of components between 1 to 100 µm in size, and MEMS devices generally range in size from 20 µm to a mm.
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.
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.
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, 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 been proven to be actually faster than regular computers executing very optimized algorithms.
QoS	Quality of Service.
Neuromorphic	Analog, digital, or mixed-mode analog/digital VLSI and software systems that implement models of neural systems.

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.
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.
Sandboxing	Sandboxing means that software is run in an isolated and restricted environment, in order to prevent it from performing unauthorized operations.
Sigma	The maturity of a manufacturing process can be described by a sigma rating, indicating its yield or the percentage of defect-free products it creates. A six sigma process is one in which 99.99966% of the products manufactured are statistically expected to be free of defects (3.4 defective parts/ million)
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.
Soft Errors	A soft error is a temporary wrong result, often caused by cosmic rays of temperature effects, and not by a permanent failure of the circuit (which is called a hard error). With increasing integration the chances of soft errors are said to increase, too.
Spike computations	A programming model where large collections of devices, modeled after neurons, interact through the transmission of spike signals
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).
Stochastic behavior	Stochastic behavior is non-deterministic behavior, which can only be analyzed and modeled using probabilistic approaches.
STREP	Specific Targeted Research Project – a type of European collaborative research and technology development project.
Time-on-market	The total number of days a product is available on the market
TRL	Technology Readiness Level
TSV	Through Silicon Via, a (vertical) electrical interconnect that goes through a silicon die or wafer ("via" = vertical interconnect access)
VLSI	Very-large-scale integration is the process of creating integrated circuits by combining thousands of transistors into a single chip.

REFERENCES

[Note: all web references were available on September 2014]

[Agui13]	Carlos Aguilar-Melchor, Simon Fau, Caroline Fontaine, Guy Gogniat, and Renaud Sirdey, "Recent Advances in Homomorphic Encryption", IEEE Signal Processing Magazine, pp. 108-117, March 2013
[Ailamaki14]	Quote from Anastasia Ailamaki, EPFL during her keynote "Running with scissors: Fast queries on just-in-time databases", at ACACES2014, July 2014
[Aldebaran]	http://www.aldebaran.com/en/a-robots/who-is-pepper
[Alibart10]	F. Alibart at al. "An Organic Nanoparticle Transistor Behaving as a Biological Spiking Synapse," Adv Functional Materials 20.2, 2010
[Android]	http://www.theregister.co.uk/2014/09/19/smartphones_its_the_economics_stupid/
[Androiddata]	http://www.washingtonpost.com/blogs/the-switch/wp/2014/09/18/newest-androids-will-join-iphones-in-offering-default-encryption-blocking-police/
[Approximate]	http://www.purdue.edu/newsroom/releases/2013/Q4/approximate-computing-improves-efficiency,-saves-energy.html
[Aron12]	http://www.newscientist.com/blogs/onepercent/2012/04/bloated-website-code-drains-yo.html
[ARWU2014]	Academic Ranking of World Universities, http://www.shanghairanking.com/ARWU2014.html
[Backdoors]	http://www.zdziarski.com/blog/wp-content/uploads/2014/07/iOS_Backdoors_Attack_Points_Surveillance_Mechanisms_Moved.pdf
[BadUSB]	https://srlabs.de/badusb
[Baikal]	http://en.itar-tass.com/economy/736804
[Bess14]	http://www.washingtonpost.com/blogs/the-switch/wp/2014/02/18/some-predict-computers-will-produce-a-jobless-future-heres-why-theyre-wrong/
[Blainey]	Bob Blainey, "Domain-specific models for innovation in analytics", PACT 2014 keynote.
[Borgh10]	J. Borghetti, G. Snider, P. Kuekes, J. Yang, D. Stewart, and R. Williams, "`Memristive' switches enable `stateful' logic operations via material implication," Nature, vol. 464, no. 7290, pp. 873–876, Apr. 2010
[Cadence]	http://www.cadence.com/Community/blogs/ii/archive/2013/08/06/wide-i-o-2-hybrid-memory-cube-hmc-memory-models-advance-3d-ic-standards.aspx

[Chafi]	Hassan Chafi, Arvind K. Sujeeth, Kevin J. Brown, HyoukJoong Lee, Anand R. Atreya, and Kunle Olukotun. "A domain-specific approach to heterogeneous parallelism." In Proceedings of the 16th ACM symposium on Principles and practice of parallel programming, PPoPP, pp. 35-46, New York, NY, USA, 2011. ACM.
[Cheem05]	S. Cheemalavagu, P. Korkmaz, K. Palem, "Ultra low-energy computing via probabilistic algorithms and devices: CMOS device primitives and the energy-probability relationship," International Conference on Solid State Devices. Tokyo, pp. 2-4, 2004
[Chua71]	L. Chua, "Memristor-The missing circuit element," IEEE Transactions on Circuit Theory, vol. 18, no. 5, pp. 507–519, 1971
[Cohe10]	Albert Cohen, Erven Rohou "Processor virtualization and split compilation for heterogeneous multicore embedded systems". DAC 2010: 102-107
[Cole14]	http://www.embedded.com/electronics-blogs/cole-bin/4433743/The-design-challenges-of-a-trillion-sensor-world
[Crocus]	http://www.eetasia.com/ART_8800690679_499486_NT_5e5c73a0.HTM?jumpto=view_welcomead_1409821305640&jumpto=view_welcomead_1409870487889
[Dailytech]	http://www.dailytech.com/Report+Windows+XP+Still+Running+on+Over+25+Percent+of+PCs+/article34627.htm
[Darksilicon1]	http://en.wikipedia.org/wiki/Dark_silicon
[Darksilicon2]	Nikos Hardavellas, Michael Ferdman, Babak Falsafi, and Anastasia Ailamaki. 2011. "Toward Dark Silicon in Servers." IEEE Micro 31 (4): 6-15.
[DARPA]	http://en.wikipedia.org/wiki/DARPA_Grand_Challenge
[DDHR]	http://www.ddhn.org/index-en.php
[Delfi-c3]	http://www.delfispace.nl/delfi-c3/autonomous-wireless-sun-sensor-payload
[Dennard]	http://en.wikipedia.org/wiki/Dennard_scaling
[Dijkstra72]	Edsger W. Dijkstra. The humble programmer. Commun. ACM 15(10):859-866, 1972.
[Diou10]	Diouf HiPEAC 2010
[Drones]	http://singularityhub.com/2014/08/11/top-10-reasons-drones-are-disruptive/
[DTC13]	http://www.digitaltrends.com/mobile/mobile-phone-world-population-2014/
[Dwave]	http://www.dwavesys.com/
[Eclipse]	https://www.eclipse.org/
[EMDIB]	http://www.eetimes.com/document.asp?doc_id=1323865&print=yes
[EndofWork1]	http://en.wikipedia.org/wiki/The_End_of_Work
[EndofWork2]	http://www.theatlantic.com/technology/archive/2012/10/the-consequences-of-machine-intelligence/264066/?single_page=true
[Esma12ACM]	Hadi Esmaeilzadeh, Adrian Sampson, Luis Ceze, and Doug Burger. 2012. Architecture support for disciplined approximate programming. In Proceedings of the seventeenth international conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS XVII). ACM, New York, NY, USA, 301-312.
[ETP4HPC]	http://www.etp4hpc.eu/wp-content/uploads/2013/06/Joint-ETP-Vision-FV.pdf
[EUV]	http://www.eetimes.com/document.asp?doc_id=1323865
[Exascale]	http://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/talk4-Harrod.pdf
[Fairphone]	http://www.fairphone.com
[Farab14Conf]	Quote from Faraboschi, HP Labs during his keynote "The Perfect Storm" at HiPEAC2014 Conference, January 2014

[Farab14CSW]	Quote from Paolo Faraboschi, HP Labs during his keynote "The Memory is the Computer", At the HiPEAC 10th anniversary, May 2014
[FlashCrash]	http://en.wikipedia.org/wiki/2010_Flash_Crash
[FoJ14]	http://www.pewinternet.org/2014/08/06/future-of-jobs/
[Footprint]	http://www.footprintnetwork.org/en/index.php/GFN/page/earth_overshoot_day/
[FPGA]	http://www.wired.com/2014/06/microsoft-fpga/
[FTH12]	"The challenge of mobile backhaul", page 11, in http://www.ftthcouncil.eu/documents/Publications/DandO_White_Paper_2013_Final.pdf
[FTTH]	http://www.ftthcouncil.eu/documents/Publications/DandO_White_Paper_2013_Final.pdf
[Gaisler]	http://www.gaisler.com
[Gang11]	Amlan Ganguly et al, "Scalable Hybrid Wireless Network-on-Chip Architectures for Multicore Systems", IEEE Transactions on Computers, October 2011, vol. 60 no. 10
[Gans14]	http://www.embedded.com/electronics-blogs/break-points/4433990/A-trillion-sensors
[Gunt14]	http://www.theguardian.com/sustainable-business/intel-conflict-minerals-ces-congo-electronics
[Halide]	Jonathan Ragan-Kelley, Connelly Barnes, Andrew Adams, Sylvain Paris, Frédo Durand, and Saman Amarasinghe. "Halide: A language and compiler for optimizing parallelism, locality, and recomputation in image processing pipelines." In ACM SIGPLAN Conference on Programming Language Design and Implementation, Seattle, WA, June 2013.
[Hey03]	A. Hey and A. Trefethen. "The data Deluge: An e-Science Perspective", in F. Berman, G. Fox and A. Hey, Eds. Grid Computing – Making the Global Infrastructure a Reality, pp. 809-824. Wiley, 2003
[Hill & Kozyrakis '12]	Mark D. Hill and Christos Kozyrakis, "Advancing Computer Systems without Technology Progress", ISAT Outbrief, DARPA/ISAT Workshop, March 26-27, 2012.
[Highfreqtrading]	http://en.wikipedia.org/wiki/High-frequency_trading#May_6.2C_2010_Flash_Crash
[HPkeynote]	https://www.youtube.com/watch?v=Gxn5ru7klUQ
[IBMpr]	http://www-03.ibm.com/press/us/en/pressrelease/44357.wss
[IBMWatson]	http://www.ibm.com/smarterplanet/us/en/ibmwatson/
[IndIntern]	http://www.ge.com/docs/chapters/Industrial_Internet.pdf
[IOS-8data]	http://appleinsider.com/articles/14/09/17/apple-says-incapable-of-decrypting-user-data-with-ios-8-even-for-government-agencies
[ITRS2013]	http://www.itrs.net/Links/2013ITRS/2013Chapters/2013ERD.pdf
[Kuzum11]	D. Kuzum, R. Jeyasingh, B. Lee, P. Wong, "Nanoelectronic Programmable Synapses Based on Phase Change Materials for Brain-Inspired Computing," in Nano Letters, June 2011
[Lee12]	H. D. Lee et al. "Integration of 4F2 selector-less crossbar array 2Mb ReRAM based on transition metal oxides for high density memory applications," in 2012 Symposium on VLSI Technology (VLSIT), 2012, pp. 151 –152
[Lustre]	Lustre, Esterel Technologies Scade 6
[Machine]	http://www8.hp.com/hpnext/posts/discover-day-two-future-now-machine-hp - .U_MYlEh3Ras
[Manchester]	http://www.graphene.manchester.ac.uk/future/
[Mann14]	http://www.eetimes.com/document.asp?doc_id=13245877
[Mercedes]	http://www.cnet.com/news/mercedes-autonomous-car-re-creates-worlds-first-long-distance-drive/
[Merr12]	http://www.eetimes.com/document.asp?doc_id=1262861
[Merr14]	http://www.eetimes.com/document.asp?doc_id=1323476&elq=4f0843b28caa4d22add34b751d6adf54&elqCampaignId=18725

[MPM13]	http://www.tech-pundit.com/wp-content/uploads/2013/07/Cloud_Begins_With_Coal.pdf?c761ac
[NELP12]	http://www.nelp.org/page/-/Job_Creation/LowWageRecovery2012.pdf?nocdn=1
[Nielsen]	http://www.nngroup.com/articles/law-of-bandwidth/
[NISE13]	http://www.businessinsider.com/50-percent-unemployment-robot-economy-2013-1
[Nuzm11]	Nuzman et al. "Vapor SIMD: Auto-Vectorize Once, Run Everywhere" at CGO 2011, p. 151-160
[NWiener]	http://en.wikipedia.org/wiki/Norbert_Wiener
[Obsolescence]	http://en.wikipedia.org/wiki/Planned_obsolescence
[OpenPower]	http://openpowerfoundation.org/
[OpenStream]	Antoniu Pop and Albert Cohen. OpenStream: Expressiveness and data-flow compilation of OpenMP streaming programs. ACM Transactions on Architecture and Code Optimization (TACO), January 2013.
[Orba1403]	http://electroiq.com/blog/2014/03/moores-law-has-stopped-at-28nm/
[Orba1412]	http://www.eetimes.com/author.asp?doc_id=1323497
[Overseasdata]	http://www.theregister.co.uk/2014/07/31/microsoft_overseas_data_ruling/
[Partha12]	http://h30507.www3.hp.com/t5/Innovation-HP-Labs/Electrons-for-compute-photons-for-communication-ions-forstorage/ba-p/115067 retrieved on November 2012.
[Pershin11]	Y. Pershin and M. Di Ventra, "Memory effects in complex materials and nanoscale systems," Advances in Physics, vol. 60, no. 2, p. 145, 2011
[PLUM11]	http://arxiv.org/ftp/arxiv/papers/1201/1201.5543.pdf
[PLUM13]	http://www.washingtonpost.com/blogs/wonkblog/wp/2013/02/28/how-the-recession-turned-middle-class-jobs-into-low-wage-jobs/
[PREDICT13]	http://is.jrc.ec.europa.eu/pages/ISG/documentPREDICT2013.pdf
[PRISM]	http://en.wikipedia.org/wiki/PRISM_%28surveillance_program%29
[ProjectAra]	http://www.projectara.com/
[Qualcomm]	http://www.eetimes.com/document.asp?doc_id=1323993&print=yes
[Ragh11]	http://www1.icsi.berkeley.edu/~barath/papers/emergy-hotnets11.pdf
[Razor]	http://infocenter.arm.com/help/index.jsp?topic=/com.arm.doc.arp0015a/index.html
[Riscv]	http://riscv.org/
[Roadm13]	http://www.hipeac.net/roadmap
[Robobrain]	http://robobrain.me/#/about
[Robocars]	http://spectrum.ieee.org/cars-that-think/transportation/self-driving/nisssans-ghosn-now-says-robocar-sales-could-start-in-2018
[Robolaw]	http://www.robolaw.eu/RoboLaw_files/documents/robolaw_d6.2_guidelinesregulatingrobotics_20140922.pdf
[Samp11]	A. Sampson, W. Dietl, E. Fortuna, D. Gnanapragasam, L. Ceze, and D. Grossman. "EnerJ: Approximate Data Types for Safe and General Low-Power Computation," PLDI 2011, pp. 164-174, 2011
[Sampa]	http://sampa.cs.washington.edu/research/approximation/enerj.html
[Schirber]	http://physics.aps.org/articles/v5/24
[Sciencedaily14]	http://www.sciencedaily.com/releases/2014/06/140605113709.htm
[Sege14]	http://www.eetimes.com/author.asp?section_id=36&doc_id=1323248&_mc=NL_EET_EDT_EET_daily_20140728&cid=NL_EET_EDT_EET_daily_20140728&elq=bf95fd30a7e04cdf89266d8820464357&elqCampaignId=18259
[Self-driving]	http://spectrum.ieee.org/transportation/self-driving/

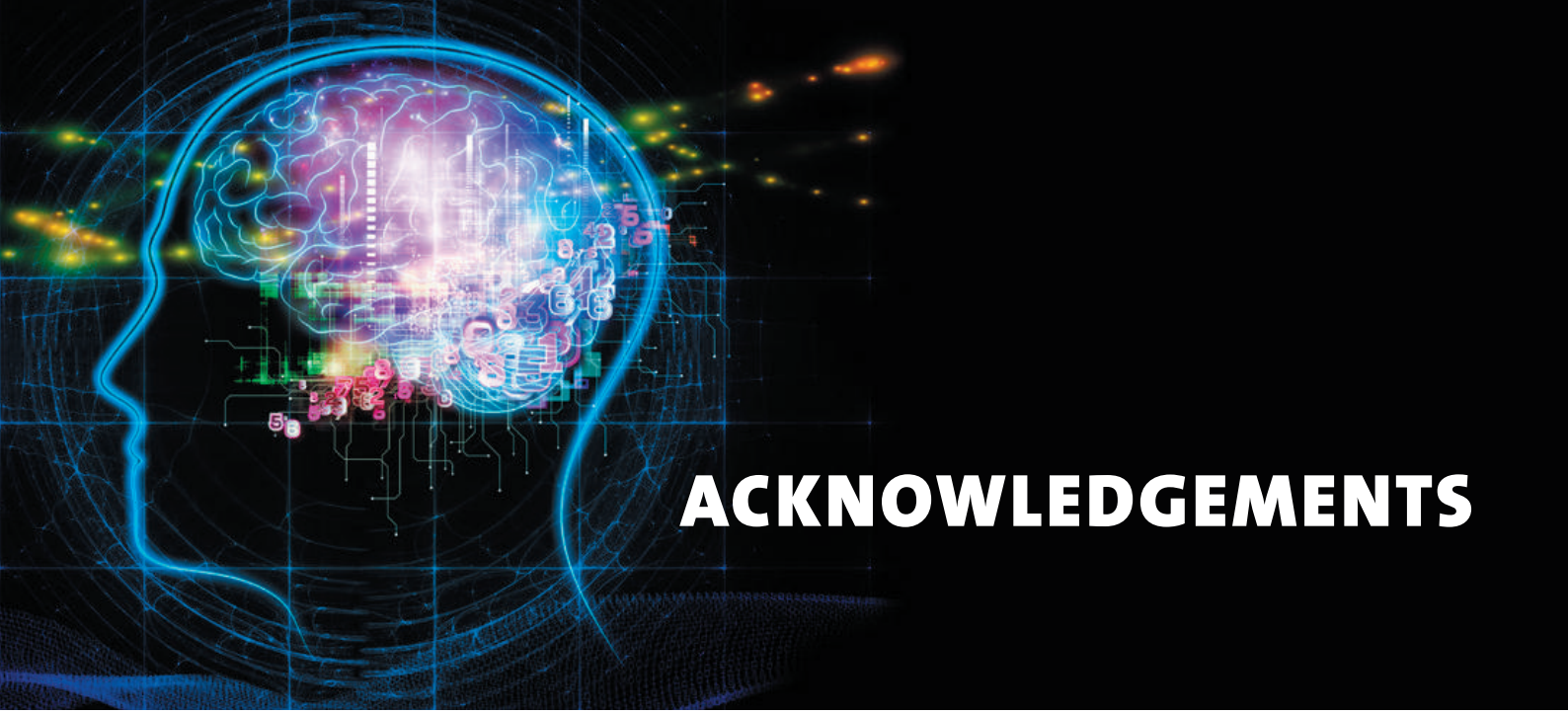
[Shei14]	http://www.theguardian.com/sustainable-business/rare-earth-metals-upgrade-recycle-ethical-china
[SKinvestm]	http://www.smartcompany.com.au/technology/42997-south-korean-government-responds-to-japan-s-robot-revolution-with-2-69-billion-investment-in-robotics.html
[SMR]	http://www.seagate.com/tech-insights/breaking-areal-density-barriers-with-seagate-smr-master-ti/
[Snider08]	G. Snider, "Spike-timing-dependent learning in memristive nanodevices," in Nanoscale Architectures, 2008. NANOARCH 2008. IEEE International Symposium on, 2008, pp. 85–92
[Softwarecrisis]	http://en.wikipedia.org/wiki/Software_crisis
[SPIRAL]	M. Pueschel, B. Singer, J. Xiong, J. Moura and J. Johnson, D. Padua, M. Veloso and R. W. Johnson. "SPIRAL: A Generator for Platform-Adapted Libraries of Signal Processing Algorithms." Journal of High Performance Computing and Applications, special issue on Automatic Performance Tuning. 2004, 18(1), pp. 21-45.
[STARNet]	https://www.src.org/program/starnet/
[StarSs]	Judit Planas, Rosa M. Badia, Eduard Ayguadé, Jesús Labarta, "Hierarchical Task-Based Programming With StarSs." IJHPCA 23(3): 284-299, 2009.
[Strukov08]	D. Strukov, G. Snider, D. Stewart, and R. Williams, "The missing memristor found," Nature, vol. 453, no. 7191, pp. 80–83, May 2008
[SyNAPSE]	https://www.ibm.com/smarterplanet/us/en/business_analytics/article/cognitive_computing.html
[TECS07]	Aman Kansal, Jason Hsu, Sadaf Zahedi, and Mani B. Srivastava. "Power management in energy harvesting sensor networks". ACM Trans. Embed. Comput. Syst. 6, 4, Article 32, 2007
[Tick-Tock]	http://en.wikipedia.org/wiki/Intel_Tick-Tock
[TopLanguages]	http://spectrum.ieee.org/static/interactive-the-top-programming-languages
[Toshiba]	http://techon.nikkeibp.co.jp/english/NEWS_EN/20141001/379923/?ST=msbe&P=2
[Ts014]	Quote by Theodore Ts'o during his lecture "File Systems and Storage Technologies" during the ACACES2014 Summer School, July 2014
[Uber]	https://www.uber.com/
[UrbanMining]	http://www.mining.com/urban-mining-the-electronic-waste-gold-mine-34760/
[VanH14]	http://www.ibcn.intec.ugent.be/sites/default/files/docs/ICT%20electricity%20consumption%202007-2012.pdf
[Watsonmedic]	http://www.wired.co.uk/news/archive/2013-02/11/ibm-watson-medical-doctor
[whitepaper]	https://ec.europa.eu/digital-agenda/en/news/white-paper-broadband-access-technologies
[Wong14]	http://www.businessweek.com/articles/2014-01-28/all-the-food-thats-fit-to-3d-print-from-chocolates-to-pizza
[Xeonchip]	http://www.extremetech.com/extreme/184828-intel-unveils-new-xeon-chip-with-integrated-fpga-touts-20x-performance-boost
[Xia09]	Q. Xia et al., "Memristor–CMOS Hybrid Integrated Circuits for Reconfigurable Logic," Nano Letters, vol. 9, no. 10, pp. 3640-3645, Sep. 2009

PROCESS

The HiPEAC vision is a bi-annual document that presents the trends that have an impact on the community of High Performance and Embedded Architecture and Compilation. The document is based on information collected through different channels.

- Meetings with teachers and industrial partners during the ACACES 2013 and 2014 summer schools.
- Two workshops with HiPEAC members and external invitees:
  - "New computing approaches and approximate computing", 27 May 2014, Brussels.
  - A meeting on more general topics, 13 June 2014, Brussels.
- Two workshops organised in cooperation with the DG Connect: Complex Systems and Advanced Computing:
  - "*Next Generation Computing Systems: components and architectures for a scalable market*", 10 December 2013, Brussels. Report available on: http://ec.europa.eu/digital-agenda/en/news/report-next-generation-computing-workshop
  - *"Software tools for next generation computing"*, 24 June 2014, Brussels. More details available on: http://ec.europa.eu/digital-agenda/en/news/software-tools-next-generation-computing-0
- A survey sent to all HiPEAC members in the spring of 2014 with more than 50 responses collected and analysed.
- A dedicated open session during the HiPEAC CSW of May 2014 in Barcelona.
- Free participation and feedback from HiPEAC members and partners.
- Feedback session during the Autumn CSW in Athens.

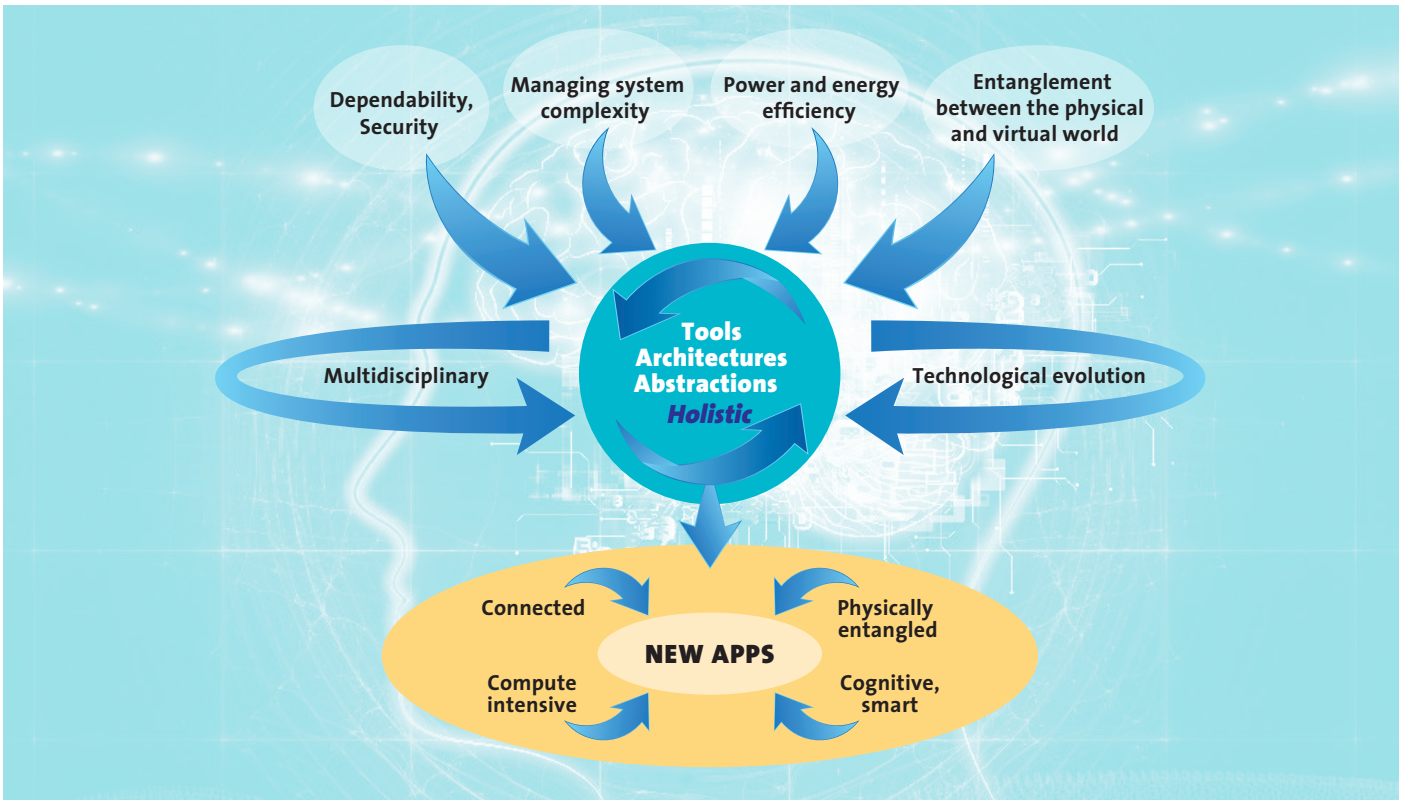
The document is called a "Vision" because it is the result of the interpretation of the trends and directions as seen by the HiPEAC community. As HiPEAC has no direct power to enforce the recommendations, the timeline associated with the potential implementation of the recommendations is uncertain; this is why the document is not a roadmap *per se*.



ACKNOWLEDGEMENTS

This document is based on the valuable inputs form the HiPEAC members. The editorial board, composed of Marc Duranton (CEA), Koen de Bosschere (Ghent University), Albert Cohen (Inria), Jonas Maebe (Ghent University), and Harm Munk (Astron), would like to thank particularly: Peter Maat (ASTRON), Vicky Wandels (UGent), Eneko Illarramendi (Ghent University), Jennifer Sartor (Ghent University), Emre Özer (ARM), Sabri Pllana (Linnaeus University), Lutz Schubert (University of Ulm), Dietmar Fey (Friedrich-Alexander-Universität Erlangen-Nürnberg), Praveen Raghavan (IMEC), Hamed Fatemi (NXP), Per Stenström (Chalmers), Pier Stanislao Paolucci (INFN Roma), Jose Daniel Garcia (University Carlos III of Madrid), Kerstin Eder (University of Bristol), Babak Falsafi (EPFL), Frank Oppenheimer (OFFIS), Sergey Tverdyshev (SYSGO), Paolo Faraboschi (HP Labs), Chris Fensch (University of Edinburgh).

HIGHLIGHTS OF THE HIPEAC VISION 2015



Information technology is one of the corner stones of modern society. It is also a key driving force of the western economy, contributing substantially to economic growth by productivity gains and the creation of new products and services. All this was made possible by the uninterrupted exponential growth in computation, communication and storage capacity of the previous decades. Unfortunately, this now seems to have come to an end. Growth is predicted to slow down due to the fundamental laws of physics, and an increasing number of people are getting worried about the adverse effects of the progressively omni-present use of information technology: loss of privacy, the impact on the job market, security and safety.

Devices will be more and more connected, entangled physically, cognitive and smart and requiring more and more computational, storage and communication resources. To ensure the further growth of information technology and to address these concerns, several key challenges need to be tackled.

- Future computing systems will have to be more reliable - preferably secure, safe and dependable by design.
- As the complexity of systems will further increase, we need better tools to manage this increase in overall complexity.
- Energy is currently the limiting factor of performance. We need solutions to dramatically increase the energy efficiency of information systems.
- The Internet of Things and cyber physical systems monitor and control the physical world and lead to an entanglement between the physical and virtual world. We need models and techniques for information systems to naturally interact, analyze and interpret the physical world.
- Computing is used in all domains of human activity – and not just by computer scientists. We must develop better tools and models for non-computer scientists to design their own information systems and push for multidisciplinary approaches.
- Systems of those systems which are only locally optimized are not globally optimal. We need a more holistic approach to system design, taking into account functional and non-functional requirements.
- In the next decade, several exponential growth laws will slow down. We must search for and industrially develop alternative technologies to continue the performance growth of information systems.

design: www.magelaan.be

HiPEAC COMPILATION ARCHITECTURE

roadmap@hipeac.net
http://www.HiPEAC.net/roadmap