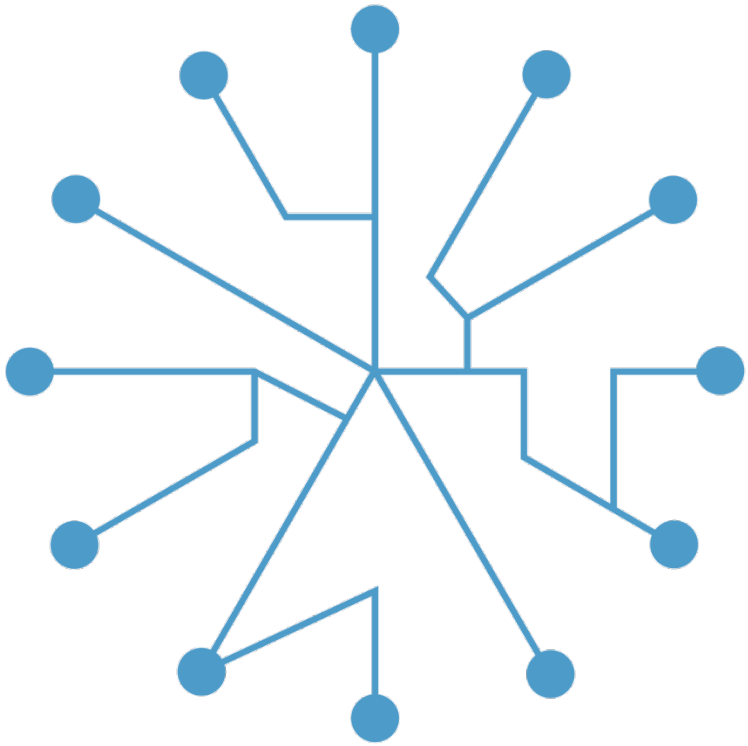
D2.1 Preliminary EuroLab-4-HPC Roadmap

V6-2016-08-4



EUROLAB-4-HPC

Document identifier: EUROLAB4HPC-DEL-DX.X	
Deliverable lead	UAU and BSC
Related Work package	WP2
Author(s)	Theo Ungerer (TU), Paul Carpenter (PC) and Mike Knebel (MK)
Main Contributor(s)	Theo Ungerer, Paul Carpenter, Avi Mendelson, Axel Tenschert, Babak Falsafi, Dietmar Fey, and Mike Knebel
Due date of deliverable	2016-08-31
Actual submission date	2016-08-30
Reviewed by	
Approved by	
Dissemination level	PU
Website	http://eurolab4hpc.eu/
Call	H2020-FETHPC-2014
Grant agreement no.	671610
Funding scheme	CSA - Coordination and Support Action
Project start date	01/09/2015
Duration	24 months

Review record

Rev N	Description	Author	Reviewers	Date
4.2	**Preliminary**	**TU, MK et al.**	**-**	**2016-06-24**
4.3	**Preliminary**	**TU, MK, PC, et al**	**-**	**2016-06-27**
4.4	**Preliminary**	**BF, AT, PC**	**-**	**2016-06-29**
5.0	**Internal and Collaborative Reviewers' Version**	**TU, PC, MK**	**Koen De Bosschere, Luca Benini, Marc Duranton (HiPEAC Vision), François Bodin (EXDCI/SRA)**	**2016-06-30**
6.0	**Final**	**TU, PC, MK**		**2016-08-04**

Executive Summary

This deliverable presents the "Preliminary Eurolab-4-HPC Roadmap" which summarizes the current state, as of August 2016, of the on-going road mapping effort within the EC CSA Eurolab-4-HPC. The "Preliminary Eurolab-4-HPC Roadmap" is a publicly available document, which will be used to foster discussions and further inputs towards the final "Eurolab-4-HPC Roadmap" due in August 2017.

The Eurolab-4-HPC Roadmap targets a long-term roadmap from 2022 to 2030 for High-Performance Computing (HPC). Because of the long-term perspective and its speculative nature, we started with an assessment of future computing technologies that could influence HPC hardware and software. This "Report on Disruptive Technologies for Years 2022-2030" is available in the Appendix. The proposal on research topics is derived from the report and discussions within the road mapping working groups.

The big picture: There is an ever-growing need of current and new applications for high performance in supercomputers, but also for mid-level and embedded computing.

High-performance computing (HPC) typically targets engineering simulations with numerical programs mostly based on floating-point computations. We expect the continued scaling of such engineering applications to continue beyond Exascale computers.

However, two trends are changing the landscape for high-performance computing and supercomputers. The first trend is the emergence of data analytics complementing simulation in scientific discovery. While simulation still remains as a major pillar for science, there are massive volumes of scientific data that are now gathered by sensors augmenting data from simulation available for analysis.

The second trend is the emergence of cloud computing and warehouse-scale computers (also known as data centers). Data centers consist of low-cost volume processing, networking and storage servers, aiming at cost-effective data manipulation at unprecedented scales. The scale at which they host and manipulate (e.g., personal, business) data has led to fundamental breakthroughs in data analytics.

There are a myriad of challenges facing massive data analytics including management of highly distributed data sources, and tracking of data provenance, data validation, mitigating sampling bias and heterogeneity, data format diversity and integrity, integration, security, sharing, visualization, and massively parallel and distributed algorithms for incremental and/or real-time analysis.

Large datacentres are fundamentally different from traditional supercomputers in their design, operation and software structures. Particularly, big data applications in data centres and cloud computing

centres require different algorithms and differ significantly from traditional HPC applications such that they may not require the same computer structures.

With modern HPC platforms being increasingly built using volume servers, there are a number of features that are shared among warehouse-scale computers and modern HPC platforms including dynamic resource allocation and management, high utilization, parallelization and acceleration, robustness and infrastructure costs. These shared concerns will serve as incentive for the convergence of the platforms.

There are also a number of ways traditional HPC ecosystems differ from modern warehouse-scale computers: efficient virtualization, adverse network topologies and fabrics in cloud platforms, low memory and storage bandwidth in volume servers. HPC customers must adapt to co-exist with cloud services; warehouse-scale computer operators must innovate technologies to support the workload and platform at the intersection of commercial and scientific computing.

It is unclear if a convergence of HPC with big data applications will arise. Investigating hardware and software structures targeting such a convergence is of high research and commercial interest.

However, further applications will emerge that may be unknown today. Recently, Deep Neural Networks (DNN) for back propagation learning of complex patterns emerged as new technique penetrating different application areas. DNN learning requires high performance and is often run on high-performance supercomputers. GPU accelerators show as very effective but also special purpose neuromorphic chips. It is widely assumed that it will be applied in future autonomous cars thus opening a very large market segment for embedded HPC.

Embedded high-performance computing demands are upcoming needs. It may concern smart phones but also applications like autonomous driving, requiring on-board high-performance computers. In particular the trend from current advanced ADAS (automatic driving assistant systems) to piloted driving (2018-2020) and to fully autonomous cars in next decade will increase on-board performance requirements and may even be coupled with high-performance supercomputers in the Cloud. The target is to develop systems that adapt more quickly to changing environments, opening the door to highly automated and autonomous transport, capable of eliminating human error in control, guidance and navigation and so leading to more safety. High-performance computing devices in cyber-physical systems will have to fulfil further non-functional requirements such as timeliness, (very) low energy consumption, security and safety.

Power and thermal management is considered as highly important and will continue its preference in future. Post-Exascale computers will target more than 1 Exaflops with less than 30 MW power consumption requiring processors with a much better performance per watt rate as available today. On the other side embedded computing needs high performance

with low energy consumption. The hardware target is widely the same, a high performance per watt.

Apart from mastering the technical challenges, reducing the environmental impact of the upcoming computing infrastructures is an important matter as well. Reducing CO_2 emissions and overall power consumption should be pursued. A combination of hardware like new processor cores, accelerators, memory and interconnect technology, and software techniques for energy and power management will need to be cooperatively deployed in order to deliver energy efficient solutions.

Because of the foreseeable end of CMOS scaling, new technologies are under development, such as, for example, Die Stacking and 3D Chip Technologies, Non-volatile Memory (NVM) Technologies, Photonics, Resistive Computing, Neuromorphic Computing, Quantum Computing, Nanotubes, Graphene, and diamond based transistors. Since it is uncertain when some of the technologies will mature, it is hard to predict which ones will prevail. The technologies will strongly impact the hardware and software of future HPC systems, in particular the processor logic itself, the (deeper) memory hierarchy, and new heterogeneous accelerators. This will significantly increase software complexity, demanding more and more intelligence across the programming environment like compiler, run-time and tool intelligence driven by appropriate programming models. Manual optimization of the data layout, placement, and caching will become uneconomic and time consuming, and will, in any case, soon exceed the abilities of the best human programmers.

But it is also possible, that new materials like graphene, nanotubes and diamonds could be used to run processors at much higher frequencies and with that may even enable to significantly increase the performance of single threaded programs. If accurate results are not necessarily needed, another speed up could emerge from more efficient special execution units, based on analog, or even on a mix between analog and digital technologies. An effective way to reason at run time on the amount of inaccuracy, which will be introduced to a system, is needed.

New memory technologies like memristors may allow on-chip integration, enabling a very tightly coupled communication between the memory and the processing unit. With the help of memory computing algorithms, data could be pre-processed “in-memory”.

Optical networks on die and Terahertz based connections may eliminate the need for preserving locality since the access time to local storage may not be as significant in future as it is today. Such advancements will lead to storage-class memory, which features similar speed, addressability and cost as DRAM combined with the non-volatility of storage. In the context of HPC, such memory can reduce the cost of checkpointing or eliminate it entirely.

The adoption of neuromorphic, resistive and/or quantum computing as new accelerators may have a dramatic effect on the system software and

programming models. It is currently unclear whether it will be sufficient to offload tasks, as on GPUs, or whether more dramatic changes will be needed. By 2030, disruptive technologies may have forced the introduction of new and currently unknown abstractions that are very different from today. Such new programming abstractions may include domain specific languages that provide greater opportunities for automatic optimization. Automatic optimization requires advanced techniques in the compiler and runtime system. We also need ways to express non-functional properties of software in order to trade various metrics: performance vs. energy, or accuracy vs. cost, both of which may become more relevant with near threshold, approximate computing or accelerators.

Nevertheless, today's abstractions will continue to evolve incrementally and will continue to be used well beyond 2030, since scientific codebases have very long lifetimes, on the order of decades.

Execution environments will increase in complexity requiring more intelligence, e.g., to manage, analyze and debug millions of parallel threads running on heterogeneous hardware with a diversity of accelerators, while dynamically adapting to failures and performance variability. This requires an evolution of the incumbent standards such as OpenMP to provide higher-level abstractions. An important question is whether and to what degree these fundamental abstractions may be impacted by disruptive technologies. Spotting anomalous behavior may be viewed as a big data problem, requiring techniques from data mining, clustering and structure detection.

The work needed: As new technologies require major changes across the stack, a vertical funding approach is needed, from applications and software systems through to new hardware architectures and potentially down to the enabling technologies. We see HP Lab's memory-driven computing architecture "The Machine" as an exemplary project that proposes a low-latency NVM (Non-Volatile Memory) based memory connected by photonics to processor cores. Projects could be based on multiple new technologies and similarly explore hardware and software structures and potential applications. Required research will be interdisciplinary. Stakeholders will come from academic and industrial research.

The opportunity: The opportunity may be development of competitive new hardware/software technologies based on upcoming new technologies to advantageous position European industry for the future. Target areas could be High-Performance Computing and Embedded High-Performance devices. The drawback could be that the chosen base technology may not be prevailing but be replaced by a different technology. For this reason, efforts should be made to ensure that aspects of the developed hardware architectures, system architectures and software systems could also be applied to alternative prevailing technologies. For instance, several NVM technologies will bring up new memory devices that are several magnitudes faster than current Flash technology and the developed system structures

may easily be adapted to the specific prevailing technologies, even if the project has chosen a different NVM technology as basis.

Contents

1. Introduction

The EC CSA Eurolab-4-HPC (Sept. 2015 – August 2017) will establish a long-term roadmap for excellence in European High-Performance Computing research, with a timescale beyond Exascale computers, i.e. a timespan of approximately 2022-2030.

The Eurolab-4-HPC roadmap will complement existing efforts such as the ETP4HPC SRA, an industry-led initiative to build a globally competitive HPC system value chain, and the HiPEAC Vision of HiPEAC CSA.

Development of the EuroLab-4-HPC roadmap will be aligned with ETP4HPC, ensuring that the multidisciplinary research done by the EuroLab-4-HPC consortium contributes with key technologies and research trends needed by the ETP4HPC SRA and its future revisions.

The EuroLab-4-HPC roadmap will also be developed in close collaboration with the upcoming HiPEAC Vision 2017 of HiPEAC CSA.

The current state of available roadmaps that are adjacent to the Eurolab-4-HPC roadmap is shown in the table below:

	Goal	Timespan	SWOT/ Political	Scope
HiPEAC Vision	Steer European academic research (driven by industry)	Short: 3 years, Mid: 6 years, Long: > 2020	Yes	HPC + embedded
ETP4HPC SRA/ EXDCI	Strengthening European [industrial] HPC ecosystem	6 years (2014 to 2020)	Yes	HPC except applications
PRACE Scientific Case	[Academic] need for European HPC infrastructure	8 years (2012 to 2020)	Yes	HPC applications
EESI (European Exascale Software Initiative)	Development of efficient Exascale applications	5 to 10 years	No	Exascale applications
BDVA (Big Data Value Association)	Big Data technologies roadmap	2020	-	Big data
Rethink Big	Roadmap for European Technologies in Hardware and Networking for Big Data		-	Big data
ECSEL MASRIA	European leadership in enabling and industrial technologies. Competitive EU ECS industry.	2015 roadmap to about 2025	Yes	Electronic components and systems (ECS)
Next Generation Computing Roadmap	Strengthening European industry	2014: 10 to 15 years		HPC extensively covered
Eurolab-4-HPC	Academic excellence in HPC	2022-2030	No	Whole HPC stack

The Eurolab-4-HPC roadmap is developed as a research roadmap with a substantially longer window than most of the roadmaps shown above. It is our target to stick to technical matters and provide an academic research perspective. Because targeting the post-Exascale era with a horizon of approximately 2022-2030 will be highly speculative, we proceed as follows:

1. Select disruptive technologies that may be technologically feasible in the next decade.
2. Assess the potential hardware architectures and their characteristics.
3. Assess what that could mean for the different working groups (WG) topics (concerns all WGs).

The roadmap roughly follow the structure:
"IF technology ready THEN foreseeable impact on WG topic could be"

The first task performed was to select potentially disruptive technologies and summarize its potential for the next decade with the help of experts in a "Report on Disruptive Technologies". The report has reached a stage of maturity and its impact on hardware and software is provided in working group zero, which is the basis for all other working groups. The full report is provided in the appendix.

0. Impact of disruptive technologies
   (Theo Ungerer, University of Augsburg, Germany)

Aside from a working group zero on disruptive technologies, we defined five more and assigned working group leaders:

1. New technologies and hardware architectures
   (Avi Mendelson, Technion, Haifa)
2. System software and programming environment
   (Paul Carpenter, BSC, Barcelona)
3. HPC application requirements
   (Paul Carpenter, BSC, Barcelona)
4. Vertical challenges: Green ICT, energy and resiliency
   (Bastian Koller and Axel Tenschert, HLRS, Stuttgart)
5. Convergence of HPC, with IoT and the Cloud
   (Babak Falsafi, EPFL, Lausanne)

Altogether about 46 contributors signed up to work on the roadmap.

The timescale concerns:

2016, April 30: We delivered an input to the EC consultation process regarding "game changing technology" (https://ec.europa.eu/futurium/en/content/fet-proactive).

2016, August 31: preliminary roadmap

2017, August 31: final roadmap

The rest of this document is structured following the topics of the working groups. The next section will provide the findings from the disruptive technologies and summarize the Potential Long-Term Impacts of Disruptive Technologies for HPC Hardware and Software in separate subsections. Section 3 targets New Technologies and Hardware Architectures, Section 4 focuses on System Software and Programming Environment, Section 5 on HPC Application Requirements, Section 6 on Green ICT, Energy and

Resiliency as Vertical Challenges, and Section 7 on the Convergence of HPC, with IoT and the Cloud.

2. Impact of Disruptive Technologies

Summary of Potential Long-Term Impacts of Disruptive Technologies for HPC Hardware

Potential long-term impacts of disruptive technologies could concern the processor logic, memory hierarchy, and future hardware accelerators.

Processor logic could be totally different if materials like graphene, nanotube or diamond would replace classical integrated circuits based on silicon transistors, or could integrate effectively with traditional CMOS technology to overcome its current major limitations like limited clock rates and heat dissipation.

A physical property that these materials share is the high thermal conductivity: Diamonds for instance can be used as a replacement for silicon, allowing diamond based transistors with excellent electrical characteristics. Graphene and nanotubes are highly electrically conductive and could allow a reduced amount of heat generated because of the lower dissipation power, which makes them more energy efficient. With the help of those good properties, less heat in the critical spots would be expected which allows much higher clock rates and highly integrated packages. Whether such new technologies will be suitable for computing in the next decade is very speculative.

Furthermore, Photonics, a technology that uses photons for communication, can be used to replace communication busses to enable a new form of inter- and intra-chip communication.

Current CMOS technology may presumably scale continuously in the next decade, down to 6 or 5 nm. However, scaling CMOS technology leads to steadily increasing costs per transistor, power consumption, and to less reliability. Die stacking could result in 3D many-core microprocessors with reduced intra core wire length, enabling high transfer bandwidths, lower latencies and reduced communication power consumption.

3D stacking will also be used to scale flash memories, because 2D NAND flash technology does not further scale. In the long run even 3D flash memories will probably be replaced by memristor or other non-volatile memory (NVM) technologies. These, depending on the actual type, allow higher structural density, less leakage power, faster read- and write access, more endurance and can nevertheless be more cost efficient.

However, the whole memory hierarchy may change in the upcoming decade. DRAM scaling will only continue with new technologies, in fact NVMs, which will deliver non-volatile memory potentially replacing or being used in addition to DRAM. Some new non-volatile memory technologies could even be integrated on-chip with the microprocessor cores and offer orders of magnitude faster read/write accesses and also much higher

endurances than flash. Intel demonstrated the possible fast memory accesses of the 3D-XPoint NVM Technology used in their Optane Technology. HP's computer architecture proposal called "The Machine" targets a machine based on new NVM memory and photonic busses. The Machine sees the memory instead of processors in the centre. This so called Memory-Driven Computing unifies the memory and storage into one vast pool of memory. HP proposes advanced photonic fabric to connect the memory and processors. Using light instead of electricity is the key to rapidly accessing any part of the massive memory pool while using much less energy.

The Machine is a first example of the new Storage-class Memory (SCM), i.e., a non-volatile memory technology in between memory and storage, which may enable new data access modes and protocols that are neither 'memory' nor 'storage'. It would particularly increase efficiency of fault tolerance check pointing, which is potentially needed for shrinking CMOS processor logic that leads to less reliable chips. There is a major impact from this technology on software and computing. SCM provides orders of magnitude increase in capacity with near-DRAM latency which would push software towards in-memory computing.

Resistive Computing, Neuromorphic Computing and Quantum Computing are promising technologies that may be suitable for new hardware accelerators but less for new processor logic. Resistive computing promises a reduction in power consumption and massive parallelism. It could enforce datacentric and reconfigurable computing, leading away from the Von-Neumann architecture. Humans can easily outperform currently available high-performance computers in tasks like vision, auditory perception and sensory motor-control. As Neuromorphic Computing would be efficient in energy and space for artificial neural network applications, it would be a good match for these tasks. More lack of abilities of current computers can be found in the area of unsolved problems in computer science. Quantum Computing might solve some of these problems, with important implications for public-key cryptography, searching, and a number of specialized computing applications.

Summary of Potential Long-Term Impacts of Disruptive Technologies for HPC Software and Applications

New technologies will lead to new hardware structures with demands on system software and programming environment and also opportunities for new applications.

CMOS scaling will require system software to deal with higher fault rate and less reliability. Also programming environment and algorithms may be affected, e.g., leading to specifically adapted approximate computing algorithms.

The most obvious change will result from changes in memory technology. NVM will prevail independent of the specific memristor technology that will

win. The envisioned Storage-Class Memory (SCM) will influence system software and programming environments in several ways:

- Memory and storage will be accessed in a uniform way.
- Computing will be memory-centric.
- Faster memory accesses by the combination of NVM and photonics will lead to a shallower memory hierarchy envisioning a flat memory where latency does not matter anymore.
- Read accesses will be faster than write accesses, though, software needs to deal with the read/write disparity, e.g., by database algorithms that favour more reads over writes.
- NVM will allow in-memory checkpointing, i.e. checkpoint replication with memory to memory operations.
- Software and hardware needs to deal with limited endurance of NVM memory.

A lot of open research questions arise from these changes for software.

Full 3D stacking may pose further requirements to system software and programming environments:

- The higher throughput and lower memory latency when stacking memory on top of processing may require changes in programming environments and application algorithms.
- Stacking specialized (e.g. analog) hardware on top of processing and memory elements lead to new (embedded) high-performance applications.
- Stacking hardware accelerators together with processing and memory elements require programming environment and algorithmic changes.
- 3D multicores require software optimizations able to efficiently utilize the characteristics of 3rd dimension, .i.e. e.g., different latencies and throughput for vertical versus horizontal interconnects.
- 3D stacking may to new form factors that allow for new (embedded) high-performance applications.

Photonics will be used to speed up all kind of interconnects – layer to layer, chip to chip, board to board, and compartment to compartment with impacts on system software, programming environments and applications such that:

- A flatter memory hierarchy will be reached (combined with 3D stacking and NVM) requiring software changes for efficiency redefining what is local in future.
- It is mentioned that energy-efficient Fourier-based computation is possible as proposed in the Optalysys project.
- The intrinsic end-to-end nature of an efficient optical channel will favour broadcast/multicast based communication and algorithms.
- A full photonic chip will totally change software in a currently rarely investigated manner.

A number of new technologies will lead to new accelerators. We envision programming environments that allow defining accelerator parts of an

algorithm independent of the accelerator itself. OpenCL is such a language distinguishing "general purpose" computing parts and accelerator parts of an algorithm, where the accelerator part can be compiled to GPUs, FPGAs, or many-cores like the Xeon Phi. Such programming environment techniques and compilers have to be enhanced to improve performance portability and to deal with potentially new accelerators as, e.g., neuromorphic chips, quantum computers, in-memory resistive computing devices etc. System software has to deal with these new possibilities and map computing parts to the right accelerator.

Neuromorphic Computing is particularly attractive for applying artificial neural network and deep learning algorithms in those domains where, at present, humans outperform any currently available high-performance computer, e.g., in areas like vision, auditory perception, or sensory motor-control. Neural information processing is expected to have a wide applicability in areas that require a high degree of flexibility and the ability to operate in uncertain environments where information usually is partial, fuzzy, or even contradictory. The success of the IBM Watson computer is an example for such new application possibilities.
It is envisioned that neuromorphic computing could help understanding the multi-level structure and function of the brain and even reach an electronic replication of the human brain at least in some areas such as perception and vision.

Quantum Computing potentially solves problems impossible by classical computing, but posts challenges to compiler and runtime support. Moreover, quantum error correction is needed due to high error rates (10^{-3}). Applications of quantum computers could be new encryptions, quantum search, quantum random walk, etc.

Resistive Computing may lead to massive parallel computing based on data-centric and reconfigurable computing paradigms. In memory computing algorithms may be executed on specialised resistive computing accelerators.

Quantum Computing, Resistive Computing as well as Graphene and Nanotube-based computing are still highly speculative hardware technologies.

3. New Technologies and Hardware Architectures

Applying Disruptive Technologies More Aggressively

A valuable way to evaluate potential disruptive technologies is to examine their impact on the fundamental assumptions that are made when building a system using current technology, in order to determine whether future technologies have the potential to change these assumptions, and if yes what the impact of that change is.

Power is a First Class Citizen when Committing to New Technology

For the last decade, power and thermal management has been of high importance. The entire market focus has moved from achieving better performance through single thread optimizations, e.g., speculative execution, towards simpler architectures that achieve better performance per watt, provided that vast parallelism exists. The problem with this approach is that it is not always easy to develop parallel programs and moreover, those parallel programs are not always performance portable, meaning that each time the architecture changes, the code may have to be rewritten.

Research on new materials, such as graphene, nanotubes and diamonds as (partial) replacements for silicon can turn the tables and help to produce chips that could run at much higher frequencies and with that may even use massive speculative techniques to significantly increase the performance of single threaded programs. A change in power density vs. cost per area will have an effect on the likelihood of dark silicon.

The reasons why such technologies are not state of the art yet are their premature state of research, which is still far from fabrication, and the unknown production costs of such high performing chips. But we may assume that in 10 to 20 years the technologies may be mature enough or other such technologies will be discovered.

Going back to improved single thread performance may be very useful for many segments of the market. Reinvestment in this field is essential since it may change the way we are developing and optimizing algorithms and code.

Locality of References

Locality of references is a central assumption of the way we design systems. The consequence of this assumption is the need of hierarchically arranged memories, 3D stacking and more.

But new technologies, including optical networks on die and Terahertz based connections, may reduce the need for preserving locality, since the differences in access time and energy costs to local memory vs. remote storage or memory may not be as significant in future as it is today.

When such new technologies find their practical use, we can expect a massive change in the way we are building hardware and software systems and are organizing software structures.

The restriction here is purely the technology, but with all the companies and universities that work on this problem, we may consider it as lifted in the future.

Digital and Analog Computation

The way how today's computers are built is based on the digital world. This allows the user to get accurate results, but with the drawbacks of cost of time, energy consumption and loss of performance. But accurate results are not always needed. Due to this limitation the production of more efficient execution units, based on analog or even a mix between analog and digital technologies could be possible. Such an approach can revolutionize the way of the programming and usage of future systems.

Currently the main problem is, that we have no effective way to reason at run time on the amount of inaccuracy we introduces to a system.

End of Von Neumann Architecture

The Von Neumann architecture assumes the use of central execution units that interface with different layers of memory hierarchies. This model, serves as *the* execution model for more than three decades. But this model is not effective in terms of performance for a given power.

New technologies like memristors may allow an on-chip integration of memory which in turn grants a very tightly coupled communication between memory and processing unit.

Assuming that these technologies will be mature, we could change algorithms and data structures to fit the new design and thus allow memory-heavy "in-memory" computing algorithms to achieve significantly better performance.

Conclusions

Thermal dissipation and power consumption currently are important limitations. Dark Silicon (i.e. large parts of the chip have to stay idle due to thermal reasons) may not happen when specific new technologies ripen. New software and hardware interfaces will be the key for successfully applying future disruptive technologies.

We may need to replace the notion of general purpose computing with clusters of specialized compute solution. Accelerators will be "application class" based.

It is important to understand the usage model in order to understand future architectures/systems.

We propose to focus on:

- Cloud based (private and public) solutions
- IoT, mobile and end-devices
- Architectures for machine learning
- Architectures for streaming data manipulation
- Data-flow compute models
- New I/O devices

Open Questions und Research Challenges

The Discussion above leads to the following principal questions und research challenges for future HPC hardware architectures and implicitly for software and applications as well:

- Impact, if power and thermal will not be limiter anymore (frequency increase vs. many-cores)?
- Impact, if Dark Silicon can be avoided?
- Impact, if communication becomes so fast so locality will not matter?
- Impact, if data movement could be eliminated (and so data locality)?
- Impact, if memory and I/O could be unified and efficiently be managed?
- Evolution of system complexity: will systems become more complex or less complex in future?

4. System Software and Programming Environment

Scope

The system software is the part of the HPC software stack that is optimized by the HPC vendor and managed by the system's operator, and it includes the Operating System (OS), cluster management tools, distributed file systems, and resource management software (job scheduler). It is essential for an operational HPC system to have an efficient system software stack below the end user's application. The programming environment comprises the development tools used to build the end user's application (compilers, IDEs, debuggers, and performance analysis tools) along with the associated abstractions (e.g. programming models), as well as the runtime components: libraries and runtime systems. Workflow management tools and commonly pre-installed application libraries such as BLAS and LAPACK are also in the scope of this section.

Current Research Trends

Sustained Increases in System Complexity, Specialization and Heterogeneity

An important role of the system software and programming environment is to provide the application developers with common standardized abstractions. Such abstractions greatly improve programmer productivity and portability across systems. Today's dominant abstractions include Fortran, C, MPI, POSIX-style file systems, threads and locking, which are all relatively low-level. By 2030, disruptive technologies may have forced the introduction of new and currently unknown low-level abstractions that are very different from these, and this topic is addressed below. Nevertheless, today's abstractions will continue to evolve incrementally and probably increase in their level of abstraction, and will continue to be used well beyond 2030, since scientific codebases have very long lifetimes, on the order of decades. Developers are unwilling to adopt a new programming language or API until they are convinced that it will be supported for a long time.

Continuous CMOS scaling and 3D stacking are pointing towards increasingly complex hardware. High-bandwidth (3D integrated) and non-volatile memories (memristors, etc.) will lead to different memory hierarchies. Increasing performance per watt demands accelerators (many-core, GPU, vector, dataflow, and their successors), heterogeneous processors (big and small cores) and potentially reconfigurable logic (FPGA). The choice of processor cores will likely become increasingly heterogeneous (within a system) and varied (across systems). Certain techniques for energy efficiency (near threshold, DVFS, energy-efficient interconnects) increase timing variability among the processes in an HPC application. Virtualization, if adopted, will also increase timing variability. In addition to hardware

complexity, execution environments will also increase in complexity, through interactive use (which will require workloads to adjust to dynamically variable numbers of nodes, cores, memory capacities, and so on).

Hiding or mitigating this increasingly complex and varied hardware requires more and more intelligence across the programming environment. Manual optimization of the data layout, placement, and caching will become uneconomic and time consuming, and will, in any case, soon exceed the abilities of the best human programmers. There needs to be a change in mentality from programming "heroism" towards trusting the compiler and runtime system (as in the move from assembler to C/Fortran). Automatic optimization requires advanced techniques in the compiler and runtime system. In the compiler, there is opportunity for both fully automated transformations and the replacement of manual refactoring by automated program transformations under the direction of human programmers (e.g. Halide [14]). Advanced runtime and system software techniques, e.g., task scheduling, load balancing, malleability, caching, energy proportionality are needed.

Increasing complexity also requires an evolution of the incumbent standards such as OpenMP, in order to provide the right programming abstractions. There is as yet no standard language for GPU-style accelerators (CUDA is controlled and only well supported by a single vendor and OpenCL provides portability). Domain-specific languages (e.g. for partial differential equations, linear algebra or stencil computations) allow programmers to describe the problem in terms much closer to the original scientific problem, and they provide greater opportunities for automatic optimization. In general there is a need to raise the level of abstraction. In some domains (e.g. embedded) prototyping is already done in a high-level environment similar to a DSL (Matlab), but the implementation still needs to be ported to a more efficient language. A different opinion expressed the need to continue to provide a (simple) cost model, in similar terms to the correspondence of the programming language C to a von Neumann CPU, so that programmers could have an intuition about the effect on performance. There is scope for ways to express non-functional properties of software, as commonly done in embedded systems, in order to trade various metrics, e.g., performance vs. energy or accuracy vs. cost, both of which may become more relevant with near threshold, approximate computing or accelerators (quantum/neuromorphic).

There is a need for global optimization across all levels of the software stack, including OS, runtime system, application libraries, and application. Examples of global problems that span multiple levels of the software stack include a) support for resiliency (system/application-level checkpointing), b) data management transformations, such as data placement in the memory hierarchy, c) minimising energy (sleeping and controlling DVFS), d) constraining peak power consumption or thermal dissipation, and e) load balancing. Different software levels have different levels of information, and

must cooperate to achieve a common objective subject to common constraints, rather than competing or becoming unstable.

Complex Application Performance Analysis and Debugging

Performance analysis and debugging are particularly difficult problems beyond Exascale. The problems are two-fold. The first problem is the enormous number of concurrent threads of execution (millions), which provides a scalability challenge (particularly in performance tools, which must not unduly affect the original performance) and in any case there will be too many threads to analyse by hand. Secondly, there is an increasing gap between (anomalous) runtime behaviour and the user's changes in the source code needed to fix it, due to libraries, runtime systems and system software that the programmer may know little or nothing about.

Spotting anomalous behaviour, such as the root cause of a performance problem or bug, will be a "big data" problem, requiring techniques from data mining, clustering and structure detection, as well as high scalability through summarized data, sampling and filtering and special techniques like spectral analysis. As implied above, the tools need to be interoperable with programming abstractions, so that problems in a loop in a library or dynamic scheduling of tasks can be translated into terms that the programmer can understand.

Potential Implications of Disruptive Technologies

Disruptive Hardware Models of Computation

Many of the fundamental abstractions used in computing in general, and high-performance computing in particular, have evolved steadily since their introduction decades ago:

- Fortran programming language (introduced in the 1950s)
- C programming language (1973)
- Sockets communications (1983)
- File system in terms of files, directories, POSIX API (1988)
- POSIX threads, locks, condition variables, etc. (1988)
- MPI message passing API (1994)
- OpenMP (1997)

An important question is whether and to what degree these fundamental abstractions may be broken by new technologies, especially disruptive technologies. The above abstractions have stood the test of time and will endure in HPC, given the long lifetimes of scientific codebases. Nevertheless, certain disruptive technologies on the horizon have the potential to challenge certain basic assumptions.

Convergence Between Storage and Memory

All existing computing systems make a strong distinction between memory and storage. Random-access memory is fast (in both bandwidth and

latency), it is byte addressable and randomly accessible by the processor, it has high cost-per-bit, and its contents are volatile. Storage is slow, in both bandwidth and latency, data is accessed through at I/O device in 512-byte (or larger) blocks, it has lower cost-per-bit, and the data is persistent.

This (hardware) correspondence between persistence on the one hand and speed, addressability and granularity on the other is the basis for the different roles of memory and storage. Temporary data structures are held in memory, and manipulated using random accesses. Data that must be persistent and/or passed among programs is serialized to a file as a byte stream.

Storage-class memory, including HPE's Persistent Memory, has similar speed, addressability and cost as DRAM with the non-volatility of storage. In the context of HPC, such memory can reduce the cost of checkpointing or eliminate it entirely. There is also work on persistent objects, e.g., NV-Heaps [12], and further work is needed.

Neuromorphic, Resistive and Quantum Computing

The adoption of neuromorphic, resistive computing and/or quantum computing may have a dramatic effect on the system software and programming model. It is currently unclear whether it will be sufficient to offload tasks, as on GPUs, or whether more dramatic changes will be needed.

5. HPC Application Requirements

Scope

Industrial and scientific applications are the *raison d'etre* of high-performance computing. HPC systems must therefore be designed to meet the needs of the users, and they must anticipate future evolutionary and disruptive changes in these requirements. This must be done while mitigating the negative impacts of the end of Moore's law and vertical challenges such as energy efficiency, programmer productivity and resiliency. This section collects the main requirements of HPC users, including applications, numerical libraries, and algorithms. The focus is on the impact of HPC requirements on HPC computing systems, rather than the applications themselves.

This section will be extended during the second year of EuroLab-4-HPC, in cooperation with the EXDCI Project, as part of the update to the 2012 PRACE Scientific Case [2].

Current Research Trends

Need for More Performance

There is no doubt that all user communities see a continued demand for ever-more computational performance well beyond Exascale. In addition, many users highlight increasing challenges related to data storage and processing. More quantitative details on future computational requirements are in the U.S Advanced Scientific Computing Advisory Committee (ASCAC) report [1] and the 2012 PRACE Scientific Case [2].

Adapting Applications for Scalability and Heterogeneity

HPC applications need to be adapted for Exascale systems and beyond. It will be some time after the introduction of the first Exaflops machine before more than a handful of applications are able to take full advantage of such a machine. The biggest issues relate to scalability (identifying and managing sufficient levels of parallelism), heterogeneity (including accelerators), and parallel I/O. Scientific codebases have very long lifetimes, on the order of decades, over which they have earned their users' trust [7]. For this reason, HPC application developers are reluctant to rewrite their software, and are keen to follow an incremental path [8].

There is strong interest in higher-level programming abstractions to provide independence and portability from the details of particular hardware implementations and execution environments, including varying degrees of parallelism, application-specific designs, heterogeneity, accelerators, and complex (deeper) memory hierarchies [9]. Compilers and runtime systems should perform complex transformations such as overlapping computations and communications [8], auto-tuning [9], scheduling and load balancing

(especially difficult with multi-scale multiphysics codes). New abstractions are needed to improve parallel I/O. Domain-Specific Languages (DSLs) help by separating domain science from the programming challenges [9]. Much more research is needed in these areas, but from the application point-of-view, the main barriers to their adoption are lack of standardization or long-term support in compilers and libraries [4], as well as difficulties in the interoperability of multiple programming models in large codebases. Regarding accelerators, there are currently too many incompatible programming interfaces, e.g. CUDA, OpenCL, OpenACC, and OpenMP 4.0, and consolidation on an open, vendor-neutral and widely used standard is needed [9].

There are serious difficulties with performance analysis and debugging, and existing techniques based on printf, logging and trace visualization will soon be intractable. Existing debuggers are good for small problems, but more work is needed to (graphically) track variables to find out where the output first became incorrect, especially for bugs that are difficult to reproduce. Performance analysis tools require lightweight data collection using sampling, folding and other techniques, so as not to increase execution time or disturb application performance (leading to non-representative analysis). There is a need for both superficial on-the-fly analysis and in-depth AI and deep learning analytics. As compilers and runtime systems become more complex, there will be a growing gap between runtime behaviour and the changes in the application's source code required to improve performance—although this does not yet seem to be a significant problem.

There is a concern that future systems will have worse performance stability and predictability, due to complex code transformations, dynamic adapting for energy and faults, dynamically changing clock speeds, and migrating work [7]. This is problematic when predictability is required, e.g., for applications such as weather forecasting and for making proposals for access to HPC resources (since proposals need an accurate prediction of application performance scalability).

Need for Co-Design

Application communities are keen to be involved in co-design activities, in order to ensure appropriate future system designs, with appropriate memory capacities, memory hierarchies, networks and topologies and storage systems well suited to a class of applications [9]. Users need early access to prototypes, in order to test algorithm performance and provide feedback to system designers [7]. As machines fail to follow Moore's law and Dennard's scaling, raw LINPACK performance is seen as non-representative of real world performance (e.g. Tianhe-2 was reportedly switched off much of the time because it was not useful for real applications, despite having at the time the world's highest LINPACK performance). Long-term partnerships are needed between vendors, HPC centres, research institutes and universities, as is being done in the U.S. ExMatEx (extreme materials), CESAR (advanced reactors) and ExaCT (combustion in turbulence) co-design centres.

Extreme Data

A new paradigm for scientific discovery is emerging due to the exponentially increasing volumes of data generated by HPC simulations and collected from telescopes, colliders, and other scientific instruments or sensors [6]. From the application point of view, the major problem is how to extract new knowledge or insights from the data [5][6]. Specific problems related to computing systems are *managing data* (streaming data processing, archiving, curation, metadata, provenance, distribution and access), *data analytics* (statistical streaming data analysis, machine learning on high-dimensional data), *data-intensive simulation* (large-scale multi-physics and multi-scale simulations), *data-driven inversion and assimilation* (high-dimensional Bayesian inference, e.g., Full Waveform Inversion for oil and gas), and *statistics and stochastic methods* (direct-inverse uncertainties and extreme event statistics) [3]. Users may wish to continue using a trusted (but inefficient) algorithm that has worked well on smaller data volumes [10].

Data movement is a major problem, including distributing data among scientists worldwide at acceptable cost and movement across infrastructure from the point of generation or collection. There is a need for *in situ* analytics and data reduction, with pre-processing, simulation, post-processing and visualization executed on the same HPC cluster. This requires batch and interactive workloads to coexist and it needs interoperable file formats [8].

More details on the convergence of HPC and big data are given in Section 6.

Interactivity and Usage Models

There are two broad categories of HPC usage. Capability computing refers to very large jobs that use (almost) the entire machine, e.g., brain simulation, or high-resolution turbulence model, and such a job must complete in the minimum time. Capacity (or throughput) computing refers to a large number of concurrent jobs, with a trade-off between minimising individual job execution time and maximising overall throughput. Capacity computing currently uses perhaps a few thousand cores per job, and it is commonly used for large ensembles of moderate-scale computations, e.g., for weather or climate simulations (in order to understand the distribution of possible outcomes) and for design space exploration.

There is increasing interest in "real time" and interactive supercomputing. High priority simulations are needed for extreme weather and mission-critical jobs (e.g. at NASA). Interactive jobs are also needed, as described above, for *in situ* visualization, as well as for computational steering: changing parameters in a simulation model as it runs, and changing resolutions in certain places of importance. Interactive and batch jobs should adapt to dynamic resource availability [9], which is an opportunity for new algorithms and programming models.

Finally, there is an opportunity to execute HPC workloads in the cloud, especially for SMEs and to support real time or high priority jobs. There have been some pilots, that show problems with the cost model, data security [2] and privacy (e.g. for medical data), licencing problems and data transfer costs.

Other Application Issues

Bit Reproducibility: Many users are not prepared for the loss of bit reproducibility across similar runs on the same system. This is a significantly increasing problem due to heterogeneity and elastic numbers of cores, and it will increase the difficulty in validating algorithms. The situation is, however, similar to physical experiments, for which random experimental error is an old and known problem.

Resiliency: is a vertical problem, and Application-Based Fault Tolerance (ABFT) techniques handle detectable, correctable and silent errors inside the application. Some algorithms have better fault tolerance than others, for example iterative solvers, which are widely used in Computational Fluid Dynamics (CFD) and other areas tolerate errors (or approximations like analog computing) by executing more iterations.

Energy Minimization: Since energy consumption is a major concern, users require better tools to measure the energy consumption. More importantly, they also need to be incentivized to minimise their energy use.

Other Application Issues: outside the scope of this roadmap (because they can be dealt with inside the application communities themselves) include: development of ultra-scalable solvers based on hierarchical algorithms [4], mesh generation [4], verification and validation and uncertainty quantification (VVUQ) [4], difficulty of coupling models at different scales, etc. [16], parallelization in time [4], methods to extract information/understanding from large quantities of scientific data [5], parallelization in time [4].

References

[1] The Opportunities and Challenges of Exascale Computing. Summary Report of the Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee. Fall 2010.

[2] PRACE Scientific Case for HPC in Europe 2012–2020. October 2012.

[3] Jean-Pierre Vilotte. Data and Data-intensive computing challenges in Earth and Universe Sciences. BDEC 2016. http://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/Vilotte-BDEC16Jun16.pdf

[4] EESI2 D3.3 Final Report on Applications Working Groups. EESI2 312478. September 2015.

[5] Andrew Connolly. The Challenge of Data in an Era of Petabyte Surveys. https://www.nrao.edu/meetings/bigdata/presentations/May3/5-Connolly/Greenbank.pdf

[6] Synergistic Challenges in Data-Intensive Science and Exascale Computing. DOE ASCAC Data Subcommittee Report, March 2013.

[7] Department of Energy Advanced Scientific Computing Advisory Committee Exascale Computing Initiative Review, August 2015

[8] EESI2 D5.3 WP5 Final Report on Cross Cutting Issues Working Groups. EESI2 312478. July 2015.

[9] EESI2 D4.3 Final Report on Enabling Technologies. EESI 312478.

[10] IDC Update on How Big Data Is Redefining High Performance Computing. 2014. https://www.tacc.utexas.edu/documents/1084364/1136739/IDC+HPDA+Briefing+slides+10.21.2014_2.pdf

6. Vertical Challenges: Green ICT, Energy and Resiliency

GreenICT

GreenICT approaches include novel emergent business models such as where the heat produced by HPC computation can be used to provide heat for business/residential buildings. Currently AoTerra is offering a small all-in-one data centre that feeds the produced heat into a buildings water circuit. Green ICT is the synergy of technological advancements and business models for application use-cases delivering cost effective low-power solutions for real problems.

Energy efficient virtual machines, software stacks and runtime libraries for native applications need to be researched and constructed. A combination of hardware (new cores, accelerators, memory/interconnect technology) and software techniques for energy and power management will need to be cooperatively deployed in order to deliver energy efficient solutions. It will be important to ensure that hardware and software techniques act as a single unified control system. Otherwise it is possible that there will be short time periods where software and hardware mechanisms will be in conflict, leading to increased energy consumption. Techniques need to be developed that enable software to automatically and optimally exploit heterogeneity in compute resources.

Further, environmental impact of IT infrastructures is another import issue. A typical benchmark for making assumptions regarding the environmental impact is the eco-efficiency calculated by the economic value of a product compared to the external effects to the environment. In addition, GreenICT considers the life cycle assessment analysing the environmental impact of a product during its complete life cycle including the input resources (e.g. kWh and resources in kg) and the output as well as changes in the inventory (e.g. CO_2 emissions). Generally it needs to be distinguished between two terms, the eco-efficiency (see above) and the eco-effectiveness. While eco-efficiency describes an approach for reducing resources and pollutants as well as increasing the outcome, eco-effectiveness is a strategy for producing zero environmental pollution and 100 % recycling.

For the HPC sector typical aspects in the scope of GreenICT are the power consumption of the IT infrastructure and additionally the power consumption for non-IT infrastructure such as cooling. Further, CO_2 emissions are influenced by HPC related infrastructure. In addition, when thinking of the complete life cycle of an IT infrastructure, the power consumption and CO_2 emissions are not only relevant for maintenance but also for production and taking out of operation.

Approaches for reducing the power consumption including aspects of CO_2 emissions were analysed in the European funded ECO_2Clouds project [1]. In

addition, a resource-efficient cooling of IT infrastructure was evaluated in the European funded CoolEmAll project [2]. As part of the mechanism to access HPC resources, users need to be incentivised to reduce their energy consumption and/or environmental impact.

Energy

Non-volatile memories offer new potential to deliver energy savings. Combined hardware and software techniques are expected to be required. Further away on the horizon are the potential impacts of new technologies concerning how chips are fabricated and new circuit structures built from materials such as graphene.

When thinking of power consumption, PUE (power usage effectiveness) and DCIE (data centre infrastructure efficiency) are of high importance.

The PUE describes the complete amount of power delivered to a datacentre divided by the amount of power used by IT devices. In this scope it needs to be distinguished between IT devices such as server, hard drives, network devices and so forth and non-IT devices such as light, cooling and so forth. Ideally the optimal PUE is 1, which would mean that the total amount of power was used by the IT devices.

The DCIE is the inverse PUE. For instance, a DCIE of 0,3 means that 30 % of the total amount of power delivered to a data centre is used by the IT devices.

Both terms, PUE and DCIE, enable a benchmark of the delivered power. An optimal usage of delivered power is necessary in the HPC sector for reducing power consumption costs and taking care of the environment.

Resiliency

Preserving data consistency in case of faults is an important topic in the HPC area. Traditional methods for handling resiliency issues are reaching limits due to growing data amounts. For instance redundant devices and backups alone are not sufficient anymore. There is a strong need for exploring novel strategies and methods going beyond existing resiliency methodologies.

We are now approaching the era of unreliable computing where it cannot be guaranteed that all computing resources within a full HPC system, a single node, or even in a single chip containing processing cores and/or memory can be relied upon to be fault free. This requires changes to the programming model, the runtime systems and the hardware used to solve HPC application problems. NVM offers a new technology that has the potential to support efficient low-cost techniques for performing checkpointing with a view to supporting transactional techniques/models for reliability. Computer systems wear out and reliability concerns mean that the localised on-chip temperature must be controlled and managed in order to maximise lifespan and to address the related problems of Dark Silicon.

The on-going shrinking of semiconductor feature sizes towards the physical limits of CMOS technology will further increase the fault rates of silicon devices. Effects like voltage fluctuation, cosmic radiation, wear out, thermal cycling, or variability may cause transient, intermittent, or permanent hardware faults leading to silent data corruption (SDC) or broken and unreachable subsystems. However, it is unclear, if and when future technologies, like photons, graphene, or nanotubes will replace silicon transistors and whether these technologies are more reliable.

In massively-parallel HPC-Systems, which may consist of millions of heterogeneous computing devices, the increasing fault rates of silicon-based chips together with the growing number of CPUs, GPUs, caches, interconnects, or off-chip memories will reduce the system's Mean-Time-Between-Failures (MTBF) to a few minutes or seconds.

In order to cope with potential hardware faults, current server processors support basic hardware fault-tolerance mechanisms, like error correcting (ECC) or error detecting codes (EDC) for the memory and the cache hierarchy, the register files and the interconnection network. NVidia also offers ECC-memory for some of their GPUs. However, current hardware fault-tolerance mechanisms shipped with commodity server systems do not cover the complete hardware of the processor. Fault-tolerance mechanisms, which cover the complete device (e.g. high-availability or safety-critical lockstep execution) are expensive, because they require a duplication or triplication of the hardware. Therefore, hardware failures cannot be detected and masked in all cases by current server hardware, which requires that long-running HPC-applications must be able to cope with the effects of frequent hardware faults on the software level. However, not all layers of the HPC software stack (operating system, middleware, application, and programming model) already support fault-tolerant execution.

For HPC-Systems software-based global checkpoint/restart mechanisms represent the state-of-the-art recovery technique. These checkpoint/restart mechanisms usually assume a fail-stop behaviour in case of a fault, which also means that faults leading to silent data corruptions, cannot be corrected. Furthermore, global software-based checkpoint mechanisms come with a significant runtime and storage overhead and provide reduced scalability when the number of computing devices is growing.

Future disruptive hardware and software technologies may solve some of the reliability problems of current HPC-systems:

- Non-volatile memory may be used for efficient hardware-supported checkpoint creation.
- Magnetic-field-based resistive memories are immune to radiation-induced soft-errors.
- Neuromorphic Computing on GPUs and custom designs can tolerate transient and permanent faults.
- Approximate computing may apply new algorithms to HPC applications that do not need the full precision.

Because hardware faults will manifest themselves much more frequently, hardware-based mechanisms such as ECC or Chipkill will not suffice to mitigate these errors. Therefore software-based reliability schemes will take an ever-increasing role in the effort to keep the MTTF rates reasonable. Another problem is that currently there is a separation between hardware-based resilience solutions and software-based ones. In the future software-based resilience techniques need to be co-designed together with hardware-based resilience to provide a powerful, seamless, integrated solution.

Furthermore new software-based fault-tolerance solutions should ideally leverage the future hardware developments such as 3D stacking, non-volatile memory (NVM) and accelerators.

Compiler-level Resilience

Future Exascale systems are expected to feature more heterogeneous computing substrates, incorporating CPUs, GPUs, vector processors, FPGAs and application specific neural or quantum accelerators. Each of these substrates have different resilience properties, for example vector instructions may be more vulnerable since they reside in the processor pipeline for longer periods; while FPGAs can be reconfigured to "heal" local permanent failures. During code generation, the compiler has to consider these resilience properties so as to maximize the overall system reliability.

Reliability at Runtime and Programming Models

With increased fault rates, it is important that the error handling/recovery is transparent to the user. However, in order for that to happen, those errors should be intercepted by the runtime and channelled to transparent error handler recovery units at the runtime and application layers. Current programming models such as MPI or OpenMP are not designed to be fault-tolerant so upon receipt of an error message, the whole application could be terminated. Furthermore, it is not clear when and at what scale those standards will adopt fault tolerance extensions; MPI fault-tolerance proposals have been discussed for decades without much impact on the standard. Instead, we argue that the runtime should provide efficient wrappers to message-passing or task-based programming models so that an error detection signal could be delivered and handled by the message or task that experienced the error.

Once the error message is delivered to the application, it must be handled efficiently as well. For this, the programming models should be elastic enough to provide the efficiency; if we adopt a stop-the-world type of synchronous error recovery, this will not be a viable solution for Exascale systems. Instead asynchronous recovery mechanisms that overlap computation with recovery are needed. For example, for task based programming models, we need the task schemes to be elastic enough so that only the tasks that were affected by the error are involved in the

recovery process while the rest of the tasks that do not depend on the error can make forward progress overlapping recovery with computation.

Checkpointing schemes also need to be adapted to the higher fault rates expected in the future. If system-wide synchronous single-level checkpointing schemes are not scalable, according to predictions, they would dominate the execution time if deployed in future Exascale systems. In comparison, user-level asynchronous multi-level checkpointing schemes will be more apt for the future. Multi-level checkpointing schemes such as FTI already exist and they are very efficient thanks to splitting up what to checkpoint intelligently between the fast local memory and slower disks. Likewise, recent HPC systems have started to deploy asynchronous checkpointing leveraging burst buffers composed fast solid-state memory; designed to drain checkpoints from local memory and thus overlapping computation with check pointing. In the future where NVM main memory will co-exist with volatile DRAM, multi-level asynchronous checkpointing schemes will become even more relevant.

In the future timeframe with increased error rates predicting failures before they manifest themselves will be increasingly more important. The runtime will need to monitor system state and based on fault-related symptoms, and predict that a possible error is imminent. The prediction mechanism should be efficient minimizing false positives, because predicting a failure where there is none reduces the performance. The same goes for false negatives where the system will not predict a fault, but an error will then happen. This will lead to costly fault recovery actions. If a fault is predicted, an correction action can be triggered employing, for example application-based fault tolerance proactively, or the device that may potentially become faulty can be disabled.

Application/Algorithm Based Fault Tolerance (ABFT)

Compared to checkpoint-restart the advantage of ABFT mechanisms is the relatively lower overhead of fault-free execution. Checkpoint restart error-recovery mechanisms are “backward” in the sense that the execution is rolled “back” upon detecting an error to the last checkpoint. In comparison, ABFT error-recovery mechanisms are “forward” in the sense that they do not need to go back to a former safe state in computation but rather they try to restore the lost data or correct it by using application/algorithm knowledge. An example is iterative linear algebra operations where the faulty segment of the matrix could be recomputed using existing partial results. The problem of ABFT is that it requires heavy modification of the application by an expert to incorporate the error recovery code. This high cost was in one sense limiting the adoption of ABFT. However, it is foreseen that relatively fewer applications/algorithms will scale to Exascale and therefore this might justify allocating more resources for these applications that do scale. In this scenario, the perceived high development costs of ABFT could be tolerable; and indeed ABFT might become a more attractive alternative.

Certain future NVM memories exhibit slower write speeds compared to reads; additionally they will likely suffer from write endurance: i.e. the memory cell will start to fail if the writes exceed a certain number. A likely research topic at the algorithm and application level is then to minimize the number of writes to degrade the cell as little as possible; here ideas similar to communication-avoiding algorithms could be utilized to maximize lifetime.

References

[1] The ECO_2Clouds project website: http://eco2clouds.eu/, last visited: 16.03.2016.
[2] The CoolEmAll project website: https://www.hlrs.de/about-us/research/past-projects/coolemall/, last visited: 16.03.2016.

7. Convergence of HPC, with IoT and the Cloud

Convergence of HPC and Cloud Computing

High-performance computing refers to technologies that enable achieving a high-level computational capacity as compared to a general-purpose computer [1]. High-performance computing in recent decades has been widely adopted for both commercial and research applications including but not limited to high-frequency trading, genomics, weather prediction, oil exploration. Since inception of high-performance computing, these applications primarily relied on simulation as a third paradigm for scientific discovery together with empirical and theoretical science.

The technological backbone for simulation has been high-performance computing platforms (also known as supercomputers) which are specialized computing instruments to run simulation at maximum speed with lesser regards to cost. Historically these platforms were designed with specialized circuitry and architecture ground up with maximum performance being the only goal. While in the extreme such platforms can be domain-specific [2], supercomputers have been historically programmable to enable their use for a broad spectrum of numerically-intensive computation. To benefit from the economies of scale, supercomputers have been increasingly relying on commodity components starting from microprocessors in the eighties and nineties, to entire volume servers with only specialized interconnects [3] taking the place of fully custom-designed platforms [4].

In the past decade, there have been two trends that are changing the landscape for high-performance computing and supercomputers. The first trend is the emergence of data analytics as the fourth paradigm [5] complementing simulation in scientific discovery. While simulation still remains as a major pillar for science, there are massive volumes of scientific data that are now gathered by instruments, sensors augmenting data from simulation available for analysis. The Large Hadron Collider and the Square Kilometre Array are just two examples of scientific experiments that generate in the order of petabytes of data a day. This recent trend has led to the emergence of data science and data-centric analytics as a significant enabler not just for science but also for humanities.

The second trend is the emergence of cloud computing and warehouse-scale computers (also known as data centres) [8]. Today, the backbone of IT and the "clouds" are data centres that are utility-scale infrastructure. Datacentres consist of low-cost volume processing, networking, and storage servers aiming at cost-effective data manipulation at unprecedented scales. Datacentre owners prioritize capital and operating costs (often measured in performance per watt) over ultimate performance. Typical high-end datacentres draw around 20 MW, occupy an area equivalent to 17 times a football field and incur a 3 billion Euros in investment. While datacentres are primarily designed for commercial use, the scale at which they host and

manipulate (e.g., personal, business) data has led to fundamental breakthroughs in data analytics.

Massive Data Analytics

We are witnessing a second revolution in IT, at the centre of which is data. The emergence of e-commerce and massive data analytics for commercial use in search engines, social networks and online shopping and advertisement has led to wide-spread use of massive data analytics (in the order of Exabytes) for consumers. Data now also lies at the core of the supply-chain for both products and services in modern economies. Collecting user input (e.g., text search) and documents online not only has led to ground-breaking advances in language translation but is also in use by investment banks mining blogs to identify financial trends. The IBM Watson experiment is a major milestone in both natural language processing and decision making to showcase a question answering system based on advanced data analytics that won a quiz show against human players.

The scientific community has long relied on generating (through simulation) or recording massive amounts of data to be analysed through high-performance computing tools on supercomputers. Examples include meteorology, genomics, connectomics (connectomes: comprehensive maps of connections within an organism's nervous system), complex physics simulations, and biological and environmental research. The proliferation of data analytics for commercial use on the internet, however, is paving the way for technologies to collect, manage and mine data in a distributed manner at an unprecedented scale even beyond conventional supercomputing applications.

Sophisticated analytic tools beyond indexing and rudimentary statistics (e.g., relational and semantic interpretation of underlying phenomena) over this vast repository of data will not only serve as future frontiers for knowledge discovery in the commercial world but also form a pillar for scientific discovery [7]. The latter is an area where commercial and scientific applications naturally overlap, and high-performance computing for scientific discovery will highly benefit from the momentum in e-commerce.

There are a myriad of challenges facing massive data analytics including management of highly distributed data sources, and tracking of data provenance, data validation, mitigating sampling bias and heterogeneity, data format diversity and integrity, integration, security, sharing, visualization, and massively parallel and distributed algorithms for incremental and/or real-time analysis.

With respect to algorithmic requirements and diversity, there are a number of basic operations that serve as the foundation for computational tasks in massive data analytics (often referred to as “dwarfs” [6] or “giants” [7]). They include but are not limited to: basic statistics, generalized n-body

problems, graph analytics, linear algebra, generalized optimization, computing integrals and data alignment. Besides classical algorithmic complexity, these basic operations all face a number of key challenges when applied to massive data related to streaming data models, approximation and sampling, high-dimensionality in data, skew in data partitioning, and sparseness in data structures. These challenges not only must be handled at the algorithmic level, but should also be put in perspective given projections for the advancement in processing, communication and storage technologies in platforms.

Many important emerging classes of massive data analytics also have real-time requirements. In the banking/financial markets, systems process large amounts of real-time stock information in order to detect time-dependent patterns, automatically triggering operations in a very specific and tight timeframe when some pre-defined patterns occur. Automated algorithmic trading programs now buy and sell millions of dollars of shares time-sliced into orders separated by 1ms. Reducing the latency by 1ms can be worth up to $100 million a year to a leading trading house. The aim is to cut microseconds off the latency in which these systems can reach to momentary variations in share prices [21].

Warehouse-Scale Computers

Large-scale internet services and cloud computing are now fuelled by large datacentres which are a warehouse full of computers. These facilities are fundamentally different from traditional supercomputers and server farms in their design, operation and software structures and primarily target delivering a negotiated level of internet service performance at minimal cost. Their design is also holistic because large portions of their software and hardware resources must work in tandem to support these services [8].

High-performance computing platforms are also converging with warehouse scale computers primarily due to the growth rate in cloud computing and server volume in the past decade. James Hamilton, VP and Distinguished Engineer at Amazon and the architect of their datacentres commented on the growth of Amazon Web Services (AWS) stating in 2014 that “every day AWS adds enough new server capacity to support Amazon’s global infrastructure when it was a $7B annual revenue enterprise (in 2004).”

Silicon technology trends such as the end of Dennard Scaling [10] and the slowdown and the projected end of density scaling [11] are pushing computing towards a new era of platform design tokened ISA: (1) technologies for tighter Integration of components (from algorithms to infrastructure), (2) technologies for Specialization (to accelerate critical services), and (3) technologies to enable novel computation paradigms for approximation. These trends apply to all market segments for digital platforms and reinforce the emergence and convergence of volume servers in warehouse-scale computers as the building block for high-performance computing platforms.

With modern high-performance computing platforms being increasingly built using volume servers, there are a number of salient features that are shared among warehouse-scale computers and modern high-performance computing platforms including dynamic resource allocation and management, high utilization, parallelization and acceleration, robustness and infrastructure costs. These shared concerns will serve as incentive for the convergence of the platforms.

There are also a number of ways traditional high-performance computing ecosystems differ from modern warehouse-scale computers [12]. With performance being a key criterion, there are a number of challenges facing high-performance computing on warehouse-scale computers. These include but are not limited to efficient virtualization, adverse network topologies and fabrics in cloud platforms, low memory and storage bandwidth in volume servers, multi-tenancy in cloud environments, and open-source deep software stacks as compared to traditional supercomputer custom stacks. As such, high-performance computing customers must adapt to co-exist with cloud services given these challenges, while warehouse-scale computer operators must innovate technologies to support the workload and platform at the intersection of commercial and scientific computing.

Embedded Systems and IoT Impacts

Internet of Things (IoT) is also having an impact on traditional high-performance computing because of a number of industrial applications that have historically adopted embedded technologies but can benefit from higher performance. Sensors and cyber-physical systems are prominent examples of embedded technologies that require managing and analysing massive amounts of data. In these applications, the embedded systems must collaborate hand-in-hand to filter and analyse data locally due to the massive scale of the data generated prior to consulting with a cloud service for high quality decisions.

In the Large Hadron Collider (LHC) in CERN, beam collisions occur every 25ns, which produces up to 40 million events per second. All these events are pipelined with the objective of distinguishing between interesting and non-interesting events to reduce the number of events to be processed to a few hundreds [18].

Bridges are monitored in real-time [19] with information collected from more than 10,000 sensors processed every 8ms, managing responses to natural disasters, maintaining bridge structure, and estimating the extent of structural fatigue.

In intelligent transportation systems, complex event processing systems are being developed to allow for fuel consumption reduction of railway systems, managing throttle positions, elaborating big amounts of data and sensor information, such as train horsepower, weight, prevailing wind, weather, traffic, etc. [20].

The automotive and avionics domains are continually demanding increasing levels of intelligence, efficiency and environmental performance, whilst fulfilling safety requirements. High performance brings the opportunity to fulfil these fundamental drivers, to develop systems that adapt more quickly to changing environments, opening the door to highly automated and autonomous transport, capable of eliminating human error in control, guidance and navigation and so leading to more safety [16, 17].

References

[1] en.wikipedia.org/wiki/Supercomputer

[2] en.wikipedia.org/wiki/Anton_(computer)

[3] www.cray.com/products/computing/xc-series

[4] Supercomputing Strategy Shifts in a World without BlueGene, nextplatform.com, April 14th, 2015.

[5] The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research.

[6] The Landscape of Parallel Computing Research: A View from Berkeley, UC Berkeley Tech. Report, Asanovic, et. al., EECS-2006-183.

[7] Frontiers in Massive Data Analysis, The National Academies Press, 2013.

[8] The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines.

[9] AWS Innovation at Scale, www.youtube.com/watch?v=JIQETrFC_SQ.

[10] Towards Dark Silicon in Servers, Hardavellas et. al., IEEE Micro 2011.

[11] After Moore's Law, The Economist, March 2016.

[12] HPC Cloud Bad; HPC in the Cloud Good, John Simmons, IPDPS Keynote, 2015.

[13] P-SOCRATES FP7 project, www.p-socrates.eu.

[14] Next Generation Computing Roadmap, a study prepared for the European Commission DG Communications Networks, Content & Technology under the contract: 30-CE-0528423/00-42 SMART 2012/0052.

[15] The HiPEAC vision for advanced computing in Horizon 2020, http://www.hipeac.net/system/files/hipeac_roadmap1_0.pdf.

[16] STRATEGIC RESEARCH AGENDA OF EPoSS - THE EUROPEAN TECHNOLOGY PLATFORM ON SMART SYSTEMS INTEGRATION, September 2013

[17] 2016 Multi Annual Strategic Research and Innovation Agenda for ECSEL Joint Undertaking.

[18] M. Shapiro, "Supersymmetry, Extra Dimensions and the Origin of Mass: Exploring the Nature of the Universe Using PetaScale Data Analysis", Google TechTalk, June 18, 2007.

[19] NTT DATA Technology Foresights 2012, "Big Data Demonstration – Bridge Monitoring System", http://www.nttdata.com/global/en/insights/foresight/pdf/2012/foresight2012_vol1_02.pdf

[20] "SAP enters complex-event processing market", http://www.cio.com.au/article/377688/sap_enters_complex-event_processing_market/.

[21] R. Tieman, "Algo trading: the dog that bit its master", Financial Times, March 2008.

[22] By Patricia Derler, Edward A. Lee, Alberto Sangiovanni Vincentelli, Modeling Cyber–Physical Systems, Invited paper on Proceedings of IEEE 2011

Terms and abbreviations

AWS	Amazon Web Services
CPS	Cyber Physical System
DCIE	Data Center Infrastructure Efficiency
EC	European Commission
EC	Embedded Computing
HPC	High Performance Computing
ISA	Instruction Set Architecture
PUE	Power Usage Effectiveness

8. Appendix:

EUROLAB-4-HPC

Report on Disruptive Technologies for years 2020-2030

Authors:

Prof. Dr. Theo Ungerer, University of Augsburg
Prof. Dr.-Ing. Dietmar Fey, University of Erlangen-Nuremberg

Compiled by:

Mike Knebel, University of Augsburg

With contribution of:

Nader Bagherzadeh, University of California, Irvine – *3D stacking*
Sandro Bartoli, University of Siena- *photonics*
Koen Bertels, Delft University of Technology - *quantum computing*
Christian Bradatsch, University of Augsburg - *graphene*
Jose Manuel García Carrasco, University of Murcia - *photonics*
Koen De Bosschere, Ghent University - *overall comments*
Marc Duranton, CEA LIST DACLE - *various technologies*
Babak Falsafi, Ecole Polytechnique Federale de Lausanne - *various technologies*
Martin Frieb, University of Augsburg – *CMOS scaling*
Florian Haas, University of Augsburg - *photonics*
Said Hamdioui, Delft University of Technology - *quantum and resistive computing*
Florian Kluge, University of Augsburg - *3D stacking*
Mike Knebel, University of Augsburg - *diamond computing, overall compilation*
Avi Mendelson, Technion - *diamond computing*
Jörg Mische, University of Augsburg - *memristors, resistive computing*
Nizar Msadek, University of Augsburg - *quantum computing*
Benjamin Pfundt, University of Erlangen-Nuremberg - *3D stacking, memristors, resistive computing*

Ulrich Rückert, University of Bielefeld - *neuromorphic computing*
Alexander Stegmeier, University of Augsburg - *nanotubes*
Sebastian Weis, University of Augsburg - *neuromorphic computing*

Abstract

This report is part of the roadmapping effort within the EC CSA Eurolab-4-HPC. The roadmap itself targets a long-term roadmap (2022-2030) for High-Performance Computing (HPC) and it was decided, because of the speculative nature, to start with an assessment of future computing technologies that could influence HPC hardware and software.

The report covers the following technologies: CMOS scaling, die stacking and 3D chip technologies, Non-volatile Memory (NVM) technologies, Photonics, Resistive Computing, Neuromorphic Computing, Quantum Computing, Nanotubes, Graphene and Diamond Transistors.

From the assessment of these technologies we derive some potential long-term impacts of Disruptive Technologies for HPC hardware.

The report is the draft (August 2016) of an on-going assessment process and will be extended in the future.

Table of contents of Appendix Report on Disruptive Technologies

1. Introduction

Roadmapping beyond the upcoming Exascale machines (2022-2030) is extremely speculative. The basic idea of Eurolab-4-HPC roadmap is therefore to assess potentially disruptive technologies and summarize its impacts on HPC hardware as IF .. THEN .. statements, i.e. IF disruptive technology will be available THEN potential impact on hardware could be.

To sort the different technologies we define Types of Innovation adapted to HPC as:

Sustaining: An innovation that does not principally affect existing HPC. An innovation that improves HPC hardware in ways that were generally expected.

Discontinuous: An innovation that is unexpected, but nevertheless does not affect existing HPC.

Disruptive: An innovation that creates a new line of HPC hardware by applying a different set of values, which ultimately (and unexpectedly) overtakes existing HPC techniques.

We survey the currents state of research and development and its potential for the future of the following hardware technologies:

- CMOS scaling
- Die stacking and 3D chip technologies
- Non-volatile Memory (NVM) technologies
- Photonics
- Resistive Computing
- Neuromorphic Computing
- Quantum Computing
- Nanotubes
- Graphene and
- Diamond Transistors

We categorize the technologies as:

- Sustaining technologies: CMOS scaling and Die stacking, see section 2
- Disruptive technologies that potentially create a new line of HPC hardware: NVM and Photonics, see section 3
- Disruptive technologies that potentially create alternative ways of computing: Resistive, Neuromorphic, and Quantum Computing, see section 4
- Disruptive technologies that potentially replace CMOS for processor logic: Nanotube, Graphene, and Diamond technologies, see section 5.

We summarize potential long-term impacts of Disruptive Technologies on HPC hardware in section 2 of the preliminary roadmap. Such impacts could concern the processor logic, the memory hierarchy, and potential hardware accelerators.

2. Sustaining Technology (improving HPC HW in ways that are generally expected)

Continuous CMOS scaling

Current (2016) high-performance multiprocessors feature 14 to 16nm technology. In April 2015, TSMC announced that the 10nm production would begin at the end of 2016. On 23 May 2015, Samsung Electronics showed off a 300mm wafer based on 10nm FinFET chips. Intel delayed their 10nm manufactured Cannonlake processor until the second half of 2017 [5] due to problems with the manufacturing process with 10nm technology. Intel's difficulties and changed plans show the continuing challenges with keeping pace with Moore's law.

Continuing Moore's Law and managing power and performance tradeoffs remain as the key drivers of the International Technology Roadmap For Semiconductors 2015 Edition (ITRS 2015) [1] grand challenges. Silicon scales according to the ITRS 2013 Roadmap until around 7 to 8nm in 2025 and 6 to 5nm in 2028 for MPUs or ASICs. DRAM half pitch (i.e., half the distance between identical features in an array) is projected to scale down to 10nm in 2025 and 7.7nm in 2028 allowing up to 32 GBits per chip. However, DRAM scaling below 20 nm is very challenging [1]. This results in an increasing cost of transistors at nodes below 10nm: the cost per transistor may increase from one technology node to the next [2].

The ITRS roadmap does not guarantee that silicon-based CMOS will extend that far because transistors with a gate length of 6 nm or smaller are significantly affected by quantum tunneling [3]. As a result of the limited further CMOS scaling the ITRS redirected their focus [4].

One trend to improve the density on chips will be 3D integration. A revolutionary DRAM/SRAM replacement will be needed [1]. As a result, non-silicon extensions of CMOS, using III-V materials or Carbon nanotube/nanowires, as well as non-CMOS platforms, including molecular electronics, spin-based computing, and single-electron devices, have been proposed [3].

Impact on hardware: "Scaling von Neumann systems leads to steadily increasing power consumption, high voltage density and high clock frequency leading away from the operating points of a biological brain" [3].

For a higher integration density, new materials and processes will be necessary. Since there is a lack of knowledge of the fabrication process of such new materials, the reliability might be lower, which may result in the need of integrated fault-tolerance mechanisms [1].

References

[1] Semiconductor Industry Association. "International technology roadmap for semiconductors (ITRS), 2015 edition." Hsinchu, Taiwan, 2015.

[2] HiPEAC Vision, 2015

[3] en.wikipedia.org/wiki/10_nanometer

[4] http://www.nature.com/nnano/journal/v11/n2/full/nnano.2016.8.html

[5] http://arstechnica.com/gadgets/2015/07/intel-confirms-tick-tock-shattering-kaby-lake-processor-as-moores-law-falters/

Die Stacking and 3D-Chip

Die Stacking and 3D chip integration denote the concept of stacking integrated circuits (e.g. processors and memories) vertically in multiple layers. 3D packaging assembles vertically stacked dies in a package, e.g., system-in-package (SIP) and package-on-package (POP).

Die stacking can be achieved in a stacking approach by connecting separately manufactured wafers or dies vertically either via wafer-to-wafer, die-to-wafer, or even die-to-die integration. The mechanical and electrical contacts are realized either by wire bonding as in SIP and POP devices or microbumps. SIP is sometimes listed as a 3D stacking technology, although it should be better denoted as 2.5 D technology.

Another approach is arranging dies (called chiplets) horizontally connected with Interposers onto silicon substrate. The advantages of 3D technology based on Interposer are numerous: Firstly, short communication distance between dies, thus reducing communication load and then reducing communication power consumption. Secondly, the possibility of stacking dies from various heterogeneous technologies, like stacking memory on top of logic like flash, nonvolatile memories, or even photonic devices, in order to benefit of the best technology where it best fits. And thirdly, an improved system yield and cost by partitioning the system in a divide & conquer approach: multiple similar dies are fabricated, tested and sorted before the final 3D assembly, instead of fabricating ultra large dies with much reduced yield.

Die stacking can also be achieved by stacking active layers vertically on a single wafer in a monolithic approach. Such kind of 3D chip integration does not use off-chip signaling for communication but it applies direct signaling between layers. Contacts are implemented in true 3D technology without mechanical contacts using inductive or capacitive effects or by vertical conductive channels through the chip substrate, so-called through-silicon-vias (TSV).

Since the TSV technology offers the densest connectivity, it is currently the most promising and favored 3D stacking technology for future high-performance microprocessors. Besides, there is also monolithic 3D technology, where layers are grown on top of the other. This is also more compact, and allows smaller grain integration between layers.

Current state: The monolithic approach of die stacking is already used in 3D flash memories from Samsung and also for smart sensors. Commercial

prototypes of 3D technology date back until 2004 when Tezzaron released a 3D IC microcontroller [1]. Intel evaluated chip stacking for a Pentium4 already in 2006 [2]. Recent multicore designs using Tezzaron's technology include the 64 core 3D-MAPS (3D MAssively Parallel processor with Stacked memory) research prototype from 2012 [3] [4] and the Centip3De with 64 ARM Cortex-M3 Cores also from 2012 [5]. Fabs are able to handle 3D packages (e.g. [6]). In 2011 IBM announced 3D chip production process [7]. Intel announced "3D XPoint" memory in 2015 (assumed to be 10x the capacity of DRAM and 1000x faster than NAND flash [8]). Both NVIDIA and AMD already exploit the high-bandwidth and low latencies given by 3D stacked memories for a high-dense memory processor, called high-bandwidth memory (HBM). AMD's GPUs based on the Fiji architecture with HBM are available since 2015, and NVIDIA released Pascal-based GPUs in 2016 [17]. A direction towards future 3D stacking of memory dies on processor dies is the Hybrid Memory Cube from Micron. It stacks multiple DRAM dies and a separate layer for a controller which is vertically linked with the DRAM dies. This interposer approach is used in high end FPGAs to reduce cost.

Perspective: 3D NAND Flash may be prevailing. 3D flash memories may enable SSDs with up to 10 TB of capacity in the short term [9]. In 2007, earliest potential was seen in memory stacks for mobile applications [10]. It is to expect that 3D chip technology will widely enter the market for mainstream architectures within the next 5 years. Representative for this current development are, e.g., Intel's Xeon Phi Knights Landing processors which will be equipped with package-integrated DRAMs in 2016 as a result of their cooperation with Micron.

It is also to be expected that in a long-term perspective the technology will be expanded progressively from 3D packaging technologies towards real 3D chip stacking and possibly towards 3D ICs in 3D packages in order to profit from all the benefits such technology will offer in particular for HPC architectures.

The main challenge in establishing this 3D chip stacking technology is gaining control of the thermal problems that have to be overcome to realize reliably very dense 3D interconnections. This requires the availability of appropriate design tools, which are explicitly supporting 3D layouts. Both topics represent an important issue for research in the next 10 to 15 years.

Impact on hardware: 3D stacking has a series of beneficial impacts on the hardware in general and on the possibilities how to design future processor-memory-architectures in particular. Wafers can be partitioned into smaller dies because comparatively long horizontally running links are relocated to the third dimension and thus enable smaller form factors. 3D stacking also enables heterogeneity, by integrating layers, manufactured in different processes, e.g., different memory technologies, like SRAM, DRAM, Spin-transfer-torque RAM (STT-RAM) and also memristor technologies, which would be incompatible among each other in monolithic circuits. Due to short connection wires, reduction of power consumption is to be

expected. Simultaneously, a high communication bandwidth between layers connected with TSVs can be expected leading to particularly high processor-to-memory bandwidth.

The last-level caches will probably be the first to be affected by 3D stacking technologies, which will increase bandwidth and reduce latencies by a large cache memory stacked on top of logic circuitry. In a further step it is consequent to expand 3D chip integration also to main memory in order to make a strong contribution in reducing decisively the current memory wall which is one of the strongest obstructions in getting more performance in HPC systems. Furthermore, possibly between 2026 and 2030, 3D arithmetic units will undergo the same changes ending up in complete 3D many-core microprocessors, which are optimized in power consumption due to reduced wire lengths.

3D stacking will also be used to scale flash memories, because 2D NAND flash technology does not scale beyond 16 nm [9, 12]. 3D stacking can also be used for image sensors. A technology was presented by Olympus in which more than 4 million microbumps have been used for stacking a 16 megapixel array sensor directly on top of a circuit implementing a global shutter control logic. Sony used TSV technology to combine image sensors directly with column-parallel analogue-digital-converters and logic circuits [13,14].

References

[1] Tezzaron 3D-IC Microcontroller Prototype [Online] February 11, 2016. http://www.tachyonsemi.com/OtherICs/3D-IC_8051_prototype.htm

[2] Black, B.; Annavaram, M.; Brekelbaum, N.; DeVale, J.; Lei Jiang; Loh, G.H.; McCauley, D.; Morrow, P.; Nelson, D.W.; Pantuso, D.; Reed, P.; Rupley, J.; Sadasivan Shankar; Shen, J.; Webb, C., "Die Stacking (3D) Microarchitecture," in *International Symposium on Microarchitecture (MICRO)*., pp.469-479, 2006

[3] http://arch.ece.gatech.edu/research/3dmaps/3dmaps.html

[4] Dae Hyun Kim; Athikulwongse, K.; Healy, M.; Hossain, M.; Moongon Jung; Khorosh, I.; Kumar, G.; Young-Joon Lee; Lewis, D.; Tzu-Wei Lin; Chang Liu; Panth, S.; Pathak, M.; Minzhen Ren; Guanhao Shen; Taigon Song; Dong Hyuk Woo; Xin Zhao; Joungho Kim; Ho Choi; Loh, G.; Hsien-Hsin Lee; Sung Kyu Lim, "3D-MAPS: 3D Massively parallel processor with stacked memory," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pp.188-190, 2012

[5] Fick, D.; Dreslinski, R.G.; Giridhar, B.; Gyouho Kim; Sangwon Seo; Fojtik, M.; Satpathy, S.; Yoonmyung Lee; Daeyeon Kim; Liu, N.; Wieckowski, M.; Chen, G.; Mudge, T.; Sylvester, D.; Blaauw, D., "Centip3De: A 3930DMIPS/W configurable near-threshold 3D stacked system with 64 ARM Cortex-M3 cores," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC*, pp.190-192, 19-23, 2012

[6] 3D *& Stacked-Die Packaging Technology Solutions [Online] February 11, 2016.* http://www.amkor.com/go/3D-Stacked-Die-Packaging

[7] IBM Press Release [Online], in German, February 11, 2016. http://www-03.ibm.com/press/de/de/pressrelease/36129.wss

[8] Intel® Optane™: Supersonic memory revolution to take-off in 2016 [Online] February 11, 2016. http://www.intel.eu/content/www/eu/en/it-managers/non-volatile-memory-idf.html

[9] Intel offers ingenious piece of 10TB 3D NAND chipper [Online] February 11, 2016. http://www.theregister.co.uk/2014/11/21/intel_offering_an_ingenious_piece_of_10tb_3d_nand_chippery/

[10] Lu , Jian-Qiang; Rose, Ken; Vitkavage, Susan, "3D Integration: Why, What, Who, When?" in *Future Fab International , Issue 23, 2007,* http://homepages.rpiscrews.us/~luj/FutureFab23_Luj_Reprint.pdf.

[11] AMD's high-bandwidth memory explained [Online] February 11, 2016. http://techreport.com/review/28294/amd-high-bandwidth-memory-explained

[12] Eun-Seok Choi; Hyun-Seung Yoo; Han-Soo Joo; Gyu-Seog Cho; Sung-Kye Park; Seok-Kiu Lee, "A Novel 3D Cell Array Architecture for Terra-Bit NAND Flash Memory," in *3rd IEEE International Memory Workshop (IMW)*, pp.1-4,2011

[13] Kondo, T.; Takemoto, Y.; Kobayashi, K.; Tsukimura, M.; Takazawa, N.; Kato, H.; Suzuki, S.; Aoki, J.; Saito, H.; Gomi, Y.; Matsuda, S.; Tadaki, Y., "A 3D stacked CMOS image sensor with 16Mpixel global-shutter mode and 2Mpixel 10000fps mode using 4 million interconnections," in *Symposium on VLSI Circuits (VLSI Circuits), 2015*, pp.C90-C91, 2015

[14] ISSCC 2013: Sony Stacked Sensor Presentation (ISSCC 2013) [Online] February 11, 2016 http://image-sensors-world-blog.blogspot.de/2013/02/isscc-2013-sony-stacked-sensor.html

[15] Y. Xie, J. Zhao, Die-stacking Architecture, Morgan & Claypool Publishers series, Synthesis Lectures on Computer Architecture, 2016.

[16] Y. Xie, J. Cong, and S. Sapatnekar, Three-Dimensional Integrated Circuit Design: EDA, Design and Microarchitectures. Springer, 2009

[17] https://images.nvidia.com/content/pdf/tesla/whitepaper/pascal-architecture-whitepaper.pdf

3. Disruptive Technology in Hardware/VLSI (innovation that creates a new line of HPC hardware superseding existing HPC techniques)

Non-volatile Memory (NVM) Technologies

Currently NAND Flash is the most common NVM technology, which finds its usages on SSDs, memory cards and memory sticks. NAND flash uses floating-gate transistors for storing single bits. This technology is facing a big challenge, because scaling down decreases the endurance and performance significantly [25]. Hence the importance of other NVM technologies increases.

Resistive memories, i.e. memristors, are an emerging class of non-volatile memory technology. The memristors electrical resistance is not constant but depends on the history of current that had previously flowed through the device. The device remembers its history—the so-called non-volatility property: when the electric power supply is turned off, the memristor remembers its most recent resistance until it is turned on again [1].

Among the most prominent memristor candidates and close to commercialization are phase change memory (PCM) [2, 3, 4, 5, 6], metal oxide resistive random access memory (RRAM or ReRAM) [7, 8], and conductive bridge random access memory (CBRAM) [9].

PCM can be integrated in the CMOS process and the read/write latency is only by tens of nanoseconds slower than DRAM whose latency is roughly around 100ns. The write endurance is hundreds of millions of writes per cell at current processes. This is why PCM is currently positioned only as a Flash replacement. [21]. RRAM offers a simple cell structure which enables reduced processing costs. The endurance can be more than 50 million cycles and the switching energy is very low [22]. RRAM can deliver 100x lower read latency and 20x faster write performance compared to NAND Flash [23]. CBRAM can also write with relatively low and with high speed. The read/write latencies are close to DRAM.

Spintronics is the technology of manipulating the spin state of electrons. Instead of using the electrons charge, spin states can be utilized as a substitute in logical circuits or in traditional memory technologies like SRAM. An STT-RAM [10] memory cell stores data in a magnetic tunnel junction (MTJ). Each MTJ is composed of two ferromagnetic layers (free and reference layers) and one tunnel barrier layer (MgO). If the magnetization direction of the magnetic fix reference layer and the switchable free layer is anti-parallel, resp. parallel, a high, resp. a low, resistance is adjusted, representing a digital "0" or "1". Recently it was reported that by adjusting intermediate magnetization angles in the free layer 16 different states can be stored in one physical cell, enabling to realize multi-cell storages in MTJ technology [11].

The read latency and read energy of STT-RAM is expected to be comparable to that of SRAM. The expected 3x higher density and 7x less leakage power consumption in the STT-RAM makes it suitable for replacing SRAMs to build large NVMs. However, a write operation in an STT-RAM memory consumes 8x more energy and exhibits a 6x longer latency than a SRAM. Therefore, minimizing the impact of inefficient writes is critical for successful applications of STT-RAM [12].

NRAM, short for Nano Ram is a proprietary technology of Nantero. The ram uses a fabric of carbon nanotubes (CNT) for saving bits. The resistive state of the CNT fabric determines, whether a one or a zero is saved in a memory cell. The resistance depends on if the CNTs are in contact with each other. With the help of a small voltage, the CNTs can be brought into contact or be separated. Reading out a bit means to measure the resistance. Nantero claims that their technology features the same read- and write latencies as DRAM, has a high endurance and reliability even in high temperature environments and is low power with essentially zero power consumption in standby mode. Furthermore NRAM is compatible with existing CMOS fabs without needing any new tools or processes, and it is scalable even to below 5nm [26].

Current state: IBM announced MLC-PCM technology replacing flash. Intel and Micron announced the new Breakthrough Memory 3D XPoint Technology [14] as revolutionary flash replacement. It is expected that the X-Point technology could become the dominating technology as an alternative to RAM devices offering in addition NVM property in the next ten years.

IBM also developed a neuromorphic core with a 64-K-PCM-cell as Synaptic-Array (256 Axone x 256 Dendrite) to implement SNNs (Spiking Neural Networks) [19].

Adesto is currently offering CBRAM technology in their serial memory chips [24].

The circuit-level performance, energy, and area model of the emerging non-volatile memory simulator NVSim [20] allows the investigation of architectural structures for future NVM based high-performance computers.

Perspective: It is foreseeable, that other NVM technologies will supersede current flash memory. PCM for instance might be 1000 times faster and 1000 times more resilient. Some NVM technologies have been considered as a feasible replacement for SRAM [15, 16, 17]. Studies suggest that replacing SRAM with STT-RAM could save 60% of LLC energy with less than 2% performance degradation [15].

It is unclear when most of the new technologies may be mature enough and which of them will prevail. But this is not of importance, because all have the same goal, namely to revolutionize the current storage technology.

Impact on hardware: Memristors will deliver non-volatile memory which can be used potentially in addition to DRAM, or as a complete replacement.

The latter will lead to a new Storage-Class Memory (SCM), i.e., a technology that blurs the distinction between memory and storage by enabling new data access modes and protocols that serve both 'memory' and 'storage'. These new SCM types of non-volatile memory could be integrated on-chip with the microprocessor cores as they use CMOS-compatible sets of materials and require different device fabrication techniques than flash. In a VLSI post-processing step they can be integrated on top of the last metal layer (see the note on Back-end of line service in section Resistive Computing). One of the challenges for the next decade is the provision of appropriate interfacing circuits between the SCMs and the microprocessor cores. The benefits of memristor devices in integration density, energy consumption and access times may not get lost by costly interface circuitry. This holds in particular for exploiting the multi-level cell storage capability of NVMs for future systems, e.g., for big data applications. Moreover, memristors offer orders of magnitude faster read/write accesses and also much higher endurance. They are resistive switching memory technologies, and thus rely on different physics than that of storing charge on a capacitor as is the case for SRAM, DRAM and Flash [18].

Spin-transfer torque magnetic random access memory (STT-RAM) devices are also an important class of non-volatile memory that primarily targets the replacement of DRAM, e.g., in Last-Level Caches (LLC). However, the asymmetric read/write energy and latency of NVM technologies introduces new challenges in designing memory hierarchies. Spintronic allows integration of logic and storage at lower power consumption.

Also new hybrid PCM / Flash SSD chips could emerge with a processor-internal last-level cache (STT-RAM), main processor memory (PCRAM), and storage class memory (ReRAM) [18].

References

[1] en.wikipedia.org/wiki/Memristor

[2] B. C. Lee, P. Zhou, J. Yang, Y. Zhang, B. Zhao, E. Ipek, O. Mutlu, and D. Burger, "Phase-change technology and the future of main memory," IEEE Micro, vol. 30, pp. 143–143, Jan. 2010.

[3] C. Lam, "Cell design considerations for phase change memory as a universal memory," in VLSI Technology, Systems and Applications, pp. 132–133, 2008.

[4] B. C. Lee, E. Ipek, O. Mutlu, and D. Burger, "Architecting phase change memory as a scalable dram alternative," in 36th Annual International Symposium on Computer Architecture (ISCA-2009), pp. 2–13, 2009.

[5] M. K. Qureshi, V. Srinivasan, and J. A. Rivers, "Scalable high performance main memory system using phase-change memory technology," in 36th Annual International Symposium on Computer Architecture (ISCA-2009), pp. 24–33, 2009.

[6] P. Zhou, B. Zhao, J. Yang, and Y. Zhang, "A Durable and Energy Efficient Main Memory Using Phase Change Memory Technology" ," in 36th Annual International Symposium on Computer Architecture (ISCA-2009), pp. 14-23, 2009.

[7] C. Xu, D. Niu, N. Muralimanohar, R. Balasubramonian, T. Zhang, S. Yu, and Y. Xie, "Overcoming the challenges of crossbar resistive memory architectures," in IEEE 21st International Symposium on High Performance Computer Architecture (HPCA 2015), pp. 476–488, Feb 2015.

[8] C. Xu, X. Dong, N. Jouppi, and Y. Xie, "Design implications of memristor-based rram cross-point structures," in Design, Automation Test in Europe Conference Exhibition (DATE), 2011, pp. 1–6, March 2011.

[9] William Wong: Conductive Bridging RAM. electronic design, 2014, http://electronicdesign.com/memory/conductive-bridging-ram

[10] D. Apalkov, A. Khvalkovskiy, S. Watts, V. Nikitin, X. Tang, D. Lottis, K. Moon, X. Luo, E. Chen, A. Ong, A. Driskill-Smith, and M. Krounbi, "Spin-transfer Torque Magnetic Random Access Memory (STT-MRAM)," Journal Emerg. Technol. Comput. Syst., vol. 9, pp. 13:1–13:35, May 2013.

[11] D. Bernard, "Spintronic devices for memristor applications, Talk at Meeting of EU COST ACTION MemoCis IC1401, "Memristors: at the crossroad of Devices and Applications", Milano, 28th March 2016.

[12] H. Noguchi, K. Kushida, K. Ikegami, K. Abe, E. Kitagawa, S. Kashiwada, C. Kamata, A. Kawasumi, H. Hara, and S. Fujita, "A 250-MHz 256b-I/O 1-Mb STT-MRAM with advanced perpendicular MTJ based dual cell for nonvolatile magnetic caches to reduce active power of processors," in 2013 Symposium on VLSI Technology (VLSIT), pp. 108–109, 2013.

[13] Gary Hilson: IBM Tackles Phase-Change Memory Drift, Resistance. EETimes 5/1/2015. http://www.eetimes.com/document.asp?doc_id=1326477

[14] http://www.intel.com/content/www/us/en/architecture-and-technology/non-volatile-memory.html

[15] H. Noguchi, K. Ikegami, N. Shimomura, T. Tetsufumi, J. Ito, and S. Fujita, "Highly reliable and low-power nonvolatile cache memory with advanced perpendicular STT-MRAM for high-performance CPU," in Symposium on VLSI Circuits Digest of Technical Papers, pp. 1–2, June 2014.

[16] H. Noguchi, K. Ikegami, K. Kushida, K. Abe, S. Itai, S. Takaya, N. Shimomura, J. Ito, A. Kawasumi, H. Hara, and S. Fujita, "A 3.3ns-access-time 71.2 uW/MHz 1Mb embedded STT-MRAM using physically eliminated read-disturb scheme and normally-off memory architecture," in IEEE International Solid-State Circuits Conference (ISSCC), pp. 1–3, Feb 2015.

[17] J. Ahn, S. Yoo, and K. Choi, "DASCA: Dead Write Prediction Assisted STT-RAM Cache Architecture," in Proceedings of the 2014 IEEE 20th International Symposium on High Performance Computer Architecture, HPCA '14, 2014.

[18] Evangelos Eleftheriou, Future Non-Volatile Memories: Technology Trends and Applications, Keynote, CSW Milan, 2015

[19] Neuromorphes System für SNNs, 09.12.2015; http://www.elektroniknet.de/halbleiter/prozessoren/artikel/126062/

[20] X. Dong, C. Xu, Y. Xie, and N. Jouppi, "NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 31, pp. 994–1007, July 2012.

[21] Lee, B. C., Ipek, E., Mutlu, O., & Burger, D. (2009). Architecting phase change memory as a scalable dram alternative. ACM SIGARCH Computer Architecture News, 37(3), 2-13.

[22] Govoreanu, B., Kar, G. S., Chen, Y. Y., Paraschiv, V., Kubicek, S., Fantini, A., ... & Jossart, N. (2011, December). 10× 10nm 2 Hf/HfO x crossbar resistive RAM with excellent performance, reliability and low-energy operation. In Electron Devices Meeting (IEDM), 2011 IEEE International (pp. 31-6). IEEE.

[23] http://www.crossbar-inc.com/technology/rram-advantages/

[24] http://www.adestotech.com/products/mavriq/

[25] http://www.anandtech.com/show/7237/samsungs-vnand-hitting-the-reset-button-on-nand-scaling

[26] http://nantero.com/technology/

Photonics

The general idea is to replace electrons with photons in intra-chip connections, inter-chip, memory and logic. First, interconnects between devices and chips are built from optical fiber, which allows higher throughput and possibly lower latency (head-flit latency). Such connections already exist (e. g. Thunderbolt or optical PCI Express by Intel). Optical inter-chip signals are then expected to be conveyed also on different mediums to facilitate integrability with CMOS process, e.g., polycarbonate as in some IBM research prototypes and commercial solutions.

The next step is to use optical interconnects to connect chips on the same circuit board. For future devices, also intra-chip connections may be optical. In general, the trend is going towards photonics integrated with electronics and closer to the chips and cores. However, conversion from photons to electrons is costly and for this reason there are currently strong efforts in improving the crucial physical modules of an integrated optical channel (e.g. modulators, photodetectors and thermally stable and efficiently integrated laser sources). Therefore, overall improvement in the effective adoption of photonics links closer and closer to the cores is to be expected.

On another direction, to eliminate the electrical-optical conversion overhead, research tries to build chips which completely are based on photonics. Silicon photonics describes the attempt to integrate photonics in CMOS-like production [1].

Optical or photonic computing uses photons produced by lasers or diodes for computation. Most research projects focus on replacing current computer components with optical equivalents, resulting in an optical digital computer system processing binary data. This approach appears to offer the best short-term prospects for commercial optical computing, since optical components could be integrated into traditional computers to produce an optical-electronic hybrid [2]. This approach is currently far from general but it is however suitable for some specific application domains, e.g., extremely energy-efficient Fourier-based computation proposed in the Optalysys project (http://optalysys.com) [5].

Current State: Optical fiber connections between devices exist. Currently, optical PCI Express connections allow high bandwidth and high speed over long distances (up to around 80m). Dedicated units for conversion between optical and electronic signals are required. Some integrated photonics solutions exist and are mainly aimed at replacing point-to-point electric wires (e.g. IBM HPC systems).

Perspective: Inter-chip connections on a single circuit board will become available soon. This requires silicon photonics, which allow the integration of photonics in a CMOS-like manufacturing process. Silicon photonics implement lasers, detectors, and waveguides on-chip with silicon only [3].

Research is working on optical intra-chip and inter-chip connections, with prototypes being already available. In this direction researchers have already identified the importance of a vertical design space exploration and design of a computer system endowed with integrated photonics. This can be combined with 3D technologies, replacing an active silicon interposer by a photonic interposer. Further challenges will arise from the evidence in current research proposals and prototypes that lower-layer design choices (e.g. physical layer, topologies, access strategies, sharing of resources), can have a significant impact in higher layers of the design (e.g. NoC-wise and up to memory coherence and programming model implications) and vice versa. This is mainly due to the scarce experience in using photonics technology for serving computing needs (close to processing cores requirements) and, most of all, due to the intrinsic end-to-end nature of an efficient optical channel, which is completely opposed to the well-established and mature knowledge of "store-and-forward" electronic communication paradigm. Then, intrinsic low-latency properties of optical interconnection (on-chip and inter-chip) could imply a re-definition of what is local in a future computing system, especially at high-scale like in a perspective HPC, together with the programming paradigms able to take advantage of the induced new optimal organization of the overall machine.

Further research targets photonic non-volatile memory [4]. This could reduce latencies of memory accesses by eliminating costly optoelectronic conversions. A revolution of micro architecture design is possible, since latencies and differences in speed between CPU and main memory in fully optical chips will not exist anymore.

There are disagreements between researchers about the future capabilities of optical computers: Will they be able to compete with semiconductor-based electronic computers on speed, power consumption, cost, and size? For optical logic to be competitive beyond a few niche applications, major breakthroughs in non-linear optical device technology would be required, or perhaps a change in the nature of computing itself [2].

The English company Optalysys, however, is of different opinion and announces that "Optalysys's initial products will launch in 2017 and are expected to enable existing computers to achieve HPC-levels of performance up to an equivalent processing rate of 9 Petaflops –

comparable to the 5th fastest computer in the world today. Following that we plan to pursue the design of larger systems capable of achieving multiple Exaflops by 2020" [5].

Impact on hardware: With both memory and connections becoming faster, 3rd level caches in Von Neumann architectures may become obsolete. Lots of main memory could be accessed with small latencies. Possibly, if the whole microarchitecture is implemented on silicon photonics, computational units and memory could work with the same speed. The elimination of the von-Neumann bottleneck promises completely new and different architectures.

References

[1] http://researcher.watson.ibm.com/researcher/view_group.php?id=2757

[2] http://en.wikipedia.org/wiki/Optical_computing

[3] http://adsabs.harvard.edu/abs/2005Natur.433..292R

[4] http://www.nature.com/nphoton/journal/vaop/ncurrent/full/nphoton.2015.182.html

[5] http://optalysys.com/optalysys-prototype-proves-optical-processing-technology-will-revolutionise-big-data-analysis-computational-fluid-dynamics-cfd/

4. Disruptive Technology (alternative ways of computing)

Resistive Computing

Apart from using memristors as non-volatile memory, there are several other ways to use memristors in computing [1, 2]. Using memristors as memristive synapses in neuromorphic computing [2, 3, 4] and using memristors in quantum computing [2] are discussed in separate sections. In this section, resistive computing is discussed.

In resistive computing, logic circuits are built by memristors [5]. Memristive gates have a lower leakage power, but switching is slower than in CMOS gates [2]. However, the integration of memory into logic allows to reprogram the logic, providing low power reconfigurable components [12] and can reduce energy and area constraints in principle due to the possibility of computing and storing in the same device (computing in memory). Memristors can also be arranged in parallel networks to enable massively parallel computing [13].

Resistive computing is one of the emerging and promising computing paradigms [5,6,7]. It takes the data-centric computing concept much further by interweaving the processing units and the memory in the same physical location using non-volatile technology, therefore significantly reducing not only the power consumption but also the memory bottleneck. Resistive devices such as memristors have been shown to be able to perform both storage and logic functions [5,8,9,10,11].

Resistive computing provides a huge potential as compared with the current state-of the art:

- It significantly reduces the memory bottleneck as it interweaves the storage, computing units and the communication [5,6,7].
- It features low power leakage [2].
- It enables maximum parallelism [7,13].
- It allows full configurability and flexibility [12].
- It provides order of magnitude improvements for the energy-delay product per operations, the computation efficiency, and performance per area [7].

Serial and parallel connections of memristors were proposed for the realization of Boolean logic gates with memristors by the so-called memristor ratio logic. In such circuits the ratio of the stored resistances in memristor devices is exploited for the set-up of Boolean logic. Memristive circuits realizing AND, OR gates and the implication function were presented in [14,15,19]. Hybrid memristive computing circuits consist of memristors and CMOS gates. The research of Singh [17], Xia et.al. [18], and Rothenbuhler et.al.[19] are representative for numerous proposals of hybrid memristive circuits, in which most of the Boolean logic operators are handled in the memristors and the CMOS transistors are mainly used for level restoration to retain defined digital signals.

Perspective: Resistive computing, if successful, will be able to significantly reduce the power consumption and enable massive parallelism; hence, increase computing energy and area efficiency by orders of magnitudes. This will transform computer systems into new highly parallel architectures and associated technologies, and enables the computation of currently infeasible big data and data-intensive applications, fueling important societal changes.

Research on resistive computing is still in its infancy stage, and the challenges are substantial at all levels, including material and technology, circuit and architecture, tools and compilers, and algorithms. As of today most of the work is based on simulations and small circuit designs. It is still unclear when the technology will be mature and available. Nevertheless, some start-ups on memristor technologies are emerging such as KNOWM.

A couple of start-up companies appeared in 2015 on the market who offer memristor technology as BEOL (Back-end of line) service in which memristive elements are post-processed in CMOS chips directly on top of the last metal layers. Also some European institutes reported just recently at a workshop meeting "Memristors: at the crossroad of Devices and Applications" of the EU cost action 1401 MemoCis the possibility BEOL integration of their memristive technology to allow experiments with such technologies. This offers new perspectives in form of hybrid CMOS/memristor logic which use memristor networks for high-dense resistive logic circuits and CMOS inverters for signal restoration to compensate the loss of full voltage levels in memristive networks. Multi-level cell capability of memristive elements can be used to face the challenge to handle the expected huge amount of Zettabytes produced annually in a couple of years. Besides, proposals exist to exploit the multi-level cell storing property for ternary carry-free arithmetic [20], [21] or both compact storing of keys and matching operations in future associative memories realized with memristors [22].

Impact on hardware: Due to its nature which is Non-Von Neumann, resistive computing will significantly change the way we used to design our computers, both from software as well as from hardware perspective. It will enforce datacentric and reconfigurable computing. Hybrid memristive networks can reduce energy and area requirement of logic circuits compared to pure CMOS. Massively parallel networks of memristors could form specialized accelerators to solve NP-hard problems [7].

Currently the interfacing and the peripheral to access memristive elements costs a lot of energy. Appropriate low energy driver circuits and new design flows are necessary to face the challenges for memristive circuits to use them. At the end extreme low-energy consumption devices can be realized both for big data applications with computing-in-memory and high-performance embedded computing devices that can be operated completely by energy harvesting mechanisms thanks to the use of memristors which offer low energy consumption circuits, high storage densities and low

number of latency states in arithmetic circuits due to carry-free additions with ternary number representations.

References

[1] Massimiliano Di Ventra, Yuriy V. Pershin: The parallel approach, Nature Physics 9, 200–202 (2013)

[2] Pershin, Y.V., Di Ventra, M.: Neuromorphic, Digital, and Quantum Computation With Memory Circuit Elements, Proc. IEEE , 100(6) 2071-2080 (2011)

[3] Matthew D. Pickett, Gilberto Medeiros-Ribeiro, Stanley Williams: A scalable neuristor built with Mott memristors, Nature Materials 12, 114–117 (2013

[4] Sung Hyun Jo, Ting Chang, Idongesit Ebong, Bhavitavya B. Bhadviya, Pinaki Mazumder and Wei Lu: Nanoscale Memristor Device as Synapse in Neuromorphic Systems, Nano Lett., 10 (4), pp 1297–1301 (2010)

[5] Julien Borghetti, Gregory S. Snider, Philip J. Kuekes, J. Joshua Yang, Duncan R. Stewart, R. Stanley Williams: 'Memristive' switches enable 'stateful' logic operations via material implication, Nature 464, 873–876 (2010)

[6] M. Di Ventra et al., "Memcomputing: a computing paradigm to store and process information on the same physical platform," arXiv preprint arXiv:1211.4487, 2012.

[7] Said Hamdioui, Lei Xie, Hoang Anh Du Nguyen, Mottaqiallah Taouil, Koen Bertels, Henk Corporaal, Hailong Jiao, Francky Catthoor, Dirk Wouters, Linn Eike, Jan van Lunteren: "Memristor Based Computation-in-Memory Architecture for Data-Intensive Applications', Proceedings of the 2015 Design, Automation & Test in Europe & Exhibition, pp. 1718-1725, 2015.

[8] G. Snider, "Computing with hysteretic resistor crossbars," Applied Physics A, vol. 80, pp. 1165–1172, 2005.

[9] J. Borghetti et al., "Memristive switches enable stateful logic operations via material implication," Nature, vol. 464, pp. 873–876, 2010.

[10] L. Gao et al., "Programmable cmos/memristor threshold logic," IEEE Transactions on Nanotechnology, vol. 12, pp. 115–119, 2013.

[11] Lei Xie, Hoang Anh Du Nguyen, Mottaqiallah Taouil, Koen Bertels, Said Hamdioui: Fast boolean logic mapped on memristor crossbar. The IEEE International Conference on Computer Design, pp. 335-342, 2015.

[12] Julien Borghetti, et al.: A hybrid nanomemristor/transistor logic circuit capable of self-programming, Proc. Natl. Acad. Sci. USA 106(6) 1699–1703 (2009)

[13] Yuriy V. Pershin, Massimiliano Di Ventra: Solving mazes with memristors: A massively parallel approach, Phys. Rev. E 84, 046703 (2011)

[14] J. J. Yang, D. B. Strukov, and D. R. Stewart. Memristive devices for computing. Nature nanotechnology, 8(1):13-24, 2013.

[15] S. Kvatinsky, A. Kolodny, U. C. Weiser, and E. G. Friedman. Memristor-based imply logic design procedure. In Proceedings of the 2011 IEEE 29th International Conference on Computer Design, ICCD'11, pages 142-147, Washington, DC, USA, 2011. IEEE Computer Society.

[16] S. Kvatinsky, N. Wald, G. Satat, A. Kolodny, U. Weiser, and E. Friedman. Mrl - memristor ratioed logic. In Cellular Nanoscale Networks and Their Applications (CNNA), 2012 13th International Workshop on, pages 1-6, Aug 2012.

[17] T. Singh. Hybrid memristor-cmos (memos) based logic gates and adder circuits. CoRR,abs/1506.06735, 2015.

[18] Q. Xia, W. Robinett, M. W. Cumbie, N. Banerjee, T. J. Cardinali, J. J. Yang, W. Wu, X. Li, W. M. Tong, D. B. Strukov, and Others. Memristor- CMOS Hybrid Integrated Circuits for Reconfigurable Logic. Nano letters, 9(10):3640{3645, 2009.

[19] A. Rothenbuhler, T. Tran, E. H. B. Smith, V. Saxena, and K. A. Campbell. Recon_gurable threshold logic gates using memristive devices. Journal of Low Power Electronics and Applications, 3(2):174-193, 2013.

[20] A. El-Slehdar, A. Fouad, and A. G. Radwan. Memristor based n-bits redundant binary adder. Microelectronics Journal, 46(3):207-213, 2015.

[21] D. Fey. Using the multi-bit feature of memristors for register files in signed-digit arithmetic units. Semiconductor Science and Technology, 29(10):104008, 2014.

[22] P. Junsangsri, F. Lombardi, and J. Han. A memristor-based tcam (ternary content addressable

memory) cell. In Nanoscale Architectures (NANOARCH), 2014 IEEE/ACM International Symposium on, pages 1-6, July 2014.

Neuromorphic Computing

Neuromorphic Computing, as developed by Carver Mead in the late 1980s, describes the use of very-large-scale integration (VLSI) systems containing electronic analog circuits to mimic neuro-biological architectures present in the nervous system.

The basic idea of Neuromorphic Computing is to exploit the massive parallelism of such circuits and to create low-power and fault-tolerant information-processing systems. Aiming at overcoming the big challenges of deep-submicron CMOS technology (power wall, reliability, and design complexity), bio-inspiration offers alternative ways to (embedded) artificial intelligence. The challenge is to understand, design, build, and use new architectures for nanoelectronic systems, which unify the best of brain-inspired information processing concepts and of nanotechnology hardware, including both algorithms and architectures [9]. A key focus area in further scaling and improving of cognitive systems is decreasing the power density and power consumption, and overcoming the CPU/memory bottleneck of conventional computational architectures [14].

In recent times, the term neuromorphic has also been used to describe analog, digital, and mixed-mode analog/digital VLSI and software systems that implement models of neural systems (for perception, motor control, or multisensory integration). The implementation of neuromorphic computing on the hardware level can be realized by oxide-based memristors, threshold switches and transistors [1, 2, 3, 4]. Such kind of research is still in its infancy.

Current state: Large scale neuromorphic chips exist based on CMOS technology, replacing processor cores by artificial neural networks. Research projects on neuromorphic computing are the following.

Mapping brain-like structures and processes into electronic substrates has recently seen a revival with the availability of deep-submicron CMOS technology. Large programs on brain-like electronic systems have been launched worldwide. At present, the largest programs are the SyNAPSE program (Systems of Neuromorphic Adaptive Plastic Scalable Electronics) in the US (launched in 2009, [10]) and the EC flagship Human Brain Project (launched in 2013, [11]).

SyNAPSE is a DARPA-funded program to develop electronic neuromorphic machine technology that scales to biological levels. More simply stated it is an attempt to build a new kind of computer with similar form and function to the mammalian brain. Such artificial brains would be used to build robots whose intelligence matches that of mice and cats. The ultimate aim is to build an electronic microprocessor system that matches a mammalian brain in function, size, and power consumption. It should recreate 10 billion neurons, 100 trillion synapses, consume one kilowatt (same as a small electric heater), and occupy less than two litres of space ([10]).

The "Cognitive Computing via Synaptronics and Supercomputing" (C2S2) project is a funded project from DARPA's SyNAPSE initiative. Headed by IBM the group will turn to digital special-purpose hardware for brain emulation. The True North chip is an impressive outcome of this project integrating a two-dimensional on-chip network of 4096 digital application-specific cores (64 x 64) and over 400 Mio. bits of local on-chip memory (~100 Kb SRAM per core) to store synapses and neuron parameters as well as 256 Mio. individually programmable synapses on-chip. One million individually programmable neurons can be simulated time-multiplexed per chip, sixteen-times more than the current largest neuromorphic chip. The chip with about 5.4 billion transistors is fabricated in a 28nm CMOS process (4.3 cm^2 die size, 240µm x 390 µm per core). By device count, True North is the largest IBM chip ever fabricated and the second largest (CMOS) chip in the world. The total power, while running a typical recurrent network at biological real-time, is about 70mW resulting in a power density of about 20mW/cm^2 (about 26pJ) which is in turn comparable to the cortex but three to four orders-of magnitude lower compared to 50-100W/cm^2 for a conventional CPU [12].

Another US initiative is the Brain Corporation (qualcomm venture). It is a pioneer in developing novel algorithms based on the functioning of the nervous system, with applications to vision, motor control, and autonomous navigation. It is working with partners to design specialized hardware that will bring to market the next generation of smart consumer products with artificial nervous systems [5].

The Human Brain Project (HBP) is a European Commission Future and Emerging Technologies Flagship. The HBP aims to put in place a cutting-

edge, ICT-based scientific research infrastructure that will allow scientific and industrial researchers to advance our knowledge in the fields of neuroscience, computing and brain-related medicine. The Project promotes collaboration across the globe, and is committed to driving forward European industry. Within the HBP the subproject SP9 designs, implements and operate a Neuromorphic Computing Platform with configurable Neuromorphic Computing Systems (NCS). The platform provides NCS based on physical (analogue or mixed-signal) emulations of brain models, running in accelerated mode (NM-PM1, wafer-scale implementation with about 200.000 analogue neurons on a wafer in 180nm CMOS), numerical models running in real time on digital multicore architectures (NM-MC1 with 18 ARM cores per chip in 130nm CMOS), and the software tools necessary to design, configure and measure the performance of these systems. The platform will be tightly integrated with the High Performance Analytics and Computing Platform, which will provide essential services for mapping and routing circuits to neuromorphic substrates, benchmarking and simulation-based verification of hardware specifications [12].

Closely related to HBP are the Blue Brain Project and the BrainScales project. The goal of the Blue Brain Project (EPFL and IBM, launched 2005): "… is to build biologically detailed digital reconstructions and simulations of the rodent, and ultimately the human brain. The supercomputer-based reconstructions and simulations built by the project offer a radically new approach for understanding the multilevel structure and function of the brain." The project uses an IBM Blue Gene supercomputer (100 TFLOPS, 10TB) with currently 8,000 CPUs to simulate ANNs (at ion-channel level) in software [15].

The European funded research project BrainScaleS (Brain-inspired multiscale computation in neuromorphic hybrid systems) aimed at understanding and emulating functions and interactions of multiple spatial and temporal scales in brain-information processing. Both, numerical simulations on Petaflop supercomputers and fundamentally different non Von Neumann hardware architectures were employed for this purpose. Within its broad scope of advancing neuromorphic computing, the hardware part is a very-large-scale, mixed-signal implementation of a highly connected, adaptive network of analogue neurons. The basic element is the HICANN (High Input Count Analog Neural Network) chip hosting one analogue network core and necessary support circuitry for communication as well as controlling. The HICANN was implemented in a 180 nm CMOS technology, has a total of 112K synapses and 512 neuron circuits and is the basic component of the HBP NM-PM1 platform [13].

According to Olivier Temam's website [6], the following things have been achieved for hardware neural networks:

- ASIC-like energy efficiency on a digital CMOS and an analog design.
- Tolerance to permanent faults on both GPUs and a custom design.
- Tolerance to transient faults.

- That about half of the PARSEC benchmarks could benefit from an NN accelerator.
- A small-footprint high-throughput accelerator for enabling state-of-the-art machine-learning in data centers or embedded systems [7].
- Tape out of a 3D stacked NN to outline that 3D stacking might be a particularly suitable scalability path for neuromorphic architectures [8].

Perspective: Software implemented artificial neural networks on HPC-clusters, multi-cores (OpenCV), and GPGPUs (NVidia cuDNN) are already commercially used. FPGA acceleration of neural networks is available as well. From a short term perspective these software implemented neural networks may be accelerated by commercial transistor-based neuromorphic chips or accelerators. Future emerging hardware technologies, like memcomputing and 3D stacking [8] may bring neuromorphic computing to a new level and overcome some of the restriction of Von Neumann based VLSI systems in terms of scalability, power consumption or performance.

The building blocks for ICs and for the Brain are the same at nanoscale level: electrons, atoms, and molecules, but their evolutions have been radically different. The fact that reliability, low-power, reconfigurability, as well as asynchronicity have been brought up so many times in recent conferences and articles, makes it compelling that the Brain should be an inspiration at many different levels, suggesting that future nano-architectures could be neural-inspired. The fascination associated with an electronic replication of the human brain has grown with the persistent exponential progress of chip technology. The present decade 2010–2020 has also made the electronic implementation more feasible, because electronic circuits now perform synaptic operations such as multiplication and signal communication at energy levels of 10 fJ, comparable to biological synapses. Nevertheless, an all-out assembly of 1014 synapses will remain a matter of a few exploratory systems for the next two decades because of several challenges [9].

Impact on hardware: Neuromorphic computing would be efficient in energy and space and applicable as hardware accelerator.

Particularly attractive is the application of ANNs in those domains where, at present, humans outperform any currently available high-performance computer, e.g., in areas like vision, auditory perception, or sensory motor-control. Neural information processing is expected to have a wide applicability in areas that require a high degree of flexibility and the ability to operate in uncertain environments where information usually is partial, fuzzy, or even contradictory. Even more computational power may be obtained by emerging technologies like quantum computing, molecular electronics, or novel nano-scale devices (memristor, spintronics, nanotubes (CMOL)) [9].

References

[1] https://en.wikipedia.org/wiki/Neuromorphic_engineering

[2] Pershin, Y.V., Di Ventra, M.: Neuromorphic, Digital, and Quantum Computation With Memory Circuit Elements, Proc. IEEE , 100(6) 2071-2080 (2011)

[3] Matthew D. Pickett, Gilberto Medeiros-Ribeiro, Stanley Williams: A scalable neuristor built with Mott memristors, Nature Materials 12, 114–117 (2013

[4] Sung Hyun Jo, Ting Chang, Idongesit Ebong, Bhavitavya B. Bhadviya, Pinaki Mazumder and Wei Lu: Nanoscale Memristor Device as Synapse in Neuromorphic Systems, Nano Lett., 10 (4), pp 1297–1301 (2010)

[5] http://www.fundingpost.com/venturefund/venture-fund-profile.asp?fund=1012

[6] http://pages.saclay.inria.fr/olivier.temam/homepage/research.html

[7] Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam. 2014. DianNao: a small-footprint high-throughput accelerator for ubiquitous machine-learning. ASPLOS '14. ACM, New York, NY, USA, 269-284.

[8] Belhadj, B.; Valentian, A.; Vivet, P.; Duranton, M.; He, L.; Temam, O., "The improbable but highly appropriate marriage of 3D stacking and neuromorphic accelerators," CASES 2014, pp.1-9, 2014.

[9] Rueckert, U., "Brain-Inspired Architectures for Nanoelectronics",. In: Hoefflinger B. (Ed.) "CHIPS2020, Vol. 2", Chap. 18, Springer, 2016.

[10] http://www.artificialbrains.com/darpa-synapse-program

[11] http://www.humanbrainproject.eu

[12] Merolla, P.A. et al.: A million spiking-neuron integrated circuit with a scalable communication network and interface, Science 345, pp. 668-673, 2014.

[13] http://brainscales.kip.uni-heidelberg.de

[14] Evangelos Eleftheriou, Future Non-Volatile Memories: Technology Trends and Applications, Keynote, CSW Milan, 2015

[15] http://bluebrain.epfl.ch/page-56882-en.html

Quantum Computing

Today's computers, both in theory (Turing machines) and practice (personal computers) are based on classical bits which can be either 0 or 1 to perform operations. Modern Quantum Computing systems operate differently as they make use of quantum bits (qubits) which can be in a superposition state and entangled with other qubits [1]. Superposition and entanglement are thus the two main phenomena that one tries to exploit in quantum computing. Superposition implies that a qubit is both in the ground and the excited state. Entanglement means that two (or more) qubits can be combined with each other such that their states have become inseparable. This gives rise to very interesting properties that can be exploited algorithmically.

The computational power of a quantum computer is directly related to these phenomena and the number of qubits. Two qubits can hold four values at any given time, namely (00, 01, 10, and 11). With each qubit that is added, the compute capacity of the quantum computer is doubled and thus

increases exponentially. All these qubits states (in superposition and entangled with each other) can then be manipulated in parallel as, e.g., gates are applied on them which gives the exponential computing power. The problem is that building a qubit is an extremely difficult task as the quantum state that is needed is very fragile and decoheres (losing the state information due to dynamic coupling with the external environment) rapidly. In addition, it is impossible to read out the state of a qubit, which ultimately is necessary to get the answer of a computation, without destroying the superposition state, thus destroying information contained in the qubit state. Basically, it turns into a classical bit that houses only a single value [2].

Current state: A well-known but highly debated example of a quantum computer is the D-Wave machine built by the Canadian company with the same name [2]. It is not yet proven that D-Wave actually uses the above mentioned quantum phenomena nor has any exponential speedup been shown except in one isolated case but which was not considered conclusive by the independent researchers such as M. Schroyer from ETH Zurich [3]. In addition, D-Wave is based on quantum annealing and thus only usable for specific optimization problems.

An alternative direction is to build a universal quantum computer based on quantum gates, such as Hadamard, rotation gates and CNOT. Google, IBM and Intel have all initiated research projects in this domain and currently superconducting qubits seem to be the most promising direction [4] [5] [6] [8].

Currently, the European Commission is preparing the ground for the launch in 2018 of a €1 billion flagship initiative on quantum technologies [9].

Perspective: Making use of Quantum Computing has the benefit to improve the speed-up of certain computations enormously, and even allows solving problems that are impossible for classical computing. Even though the challenges are substantial, they can be separated in physics oriented and engineering oriented ones. The physics challenges primarily have to address the lifetime of qubits and the fidelity of qubit gate operations. The engineering challenges go from identifying relevant algorithms and provide compiler and runtime support. It is also clear that a quantum computer will require a supercomputer to provide the necessary quantum error correction mechanisms as error rates of around 10^{-3} are not uncommon. As the quantum phenomena require mK (milli Kelvin) conditions, the control logic should be brought as close as possible to reduce the transfer of data up to room temperature computers. Understanding how conventional CMOS behaves under cryo-conditions is another challenge.

Quantum Computing might have the advantage to solve some problems that couldn't be solved with classical computers - one example is Shor's algorithm for decryption which, at least assuming that a large scale quantum computer can be built consisting of millions of qubits, could

decrypt a 2000 bit word in around one day which is completely impossible for conventional supercomputers.

In the short term, the Quantum Key Distribution algorithm (QKD) [6] can be used as a new encryption technology that relies on the fact that, when a third party tries to eavesdrop, the entangled state is immediately destroyed.

Further quantum algorithms are [7]:

- Grover's Algorithm is the second most famous result in quantum computing. Often referred to as "quantum search," Grover's algorithm actually inverts an arbitrary function by searching n input combinations for an output value in √n time.
- Binary Welded Tree is the graph formed by joining two perfect binary trees at the leaves. Given an entry node and an exit node, The Binary Welded Tree Algorithm uses a quantum random walk to find a path between the two. The quantum random walk finds the exit node exponentially faster than a classical random walk.
- Boolean Formula Algorithm can determine a winner in a two player game by performing a quantum random walk on a NAND tree.
- Ground State Estimation algorithm determines the ground state energy of a molecule given a ground state wave function. This is accomplished using quantum phase estimation.
- Linear Systems algorithm makes use of the quantum Fourier Transform to solve systems of linear equations.
- Shortest Vector problem is an NP-Hard problem that lies that the heart of some lattice-based cryptosystems. The Shortest Vector Algorithm makes use of the quantum Fourier Transform to solve this problem.
- Class Number Computes the class number of a real quadratic number field in polynomial time. This problem is related to elliptic-curve cryptography, which is an important alternative to the product-of-two-primes approach currently used in public-key cryptography.
- It is expected that machine learning will be transformed into quantum learning - the prodigious power of qubits will narrow the gap between machine learning and biological learning [3].

In general, the focus is now on developing algorithms requiring a low number of qubits (a few hundred) as that seems to be the most likely reachable goal in the 10-15 year time frame.

Impact on hardware: An interesting point to investigate is a better hardware architecture supporting the power efficiency of quantum better. If this is too complex, it should be at least possible to provide a hybrid architecture of both systems enabling to run the simplest sequences of an application as usually on classical computers and the complexe ones on quantum co-processors. By doing this, the system performance can be improved during runtime [8].

As pointed out earlier, a quantum computer will always be a heterogeneous computing platform where conventional supercomputing facilities will be

combined with quantum processing units. How they interact and communicate is clearly a challenging line of research [7].

References

[1] Quantum Computing – University of Waterloo: https://uwaterloo.ca/institute-for-quantum-computing/quantum-computing-101

[2] Cade Metz: Google's Quantum Computer Just Got a Big Upgrade; http://www.wired.com/2015/09/googles-quantum-computer-just-got-a-big-upgrade-1000-qubits/

[3] Tom Simonite: Google's Quantum Dream Machine, MIT Technology Review, December 18, 2015, www.technologyreview.com/s/544421/googles-quantum-dream-machine/

[4] Tom Simonite: Microsoft's Quantum Mechanics. MIT Technology Review, October 10, 2014. https://www.technologyreview.com/s/531606/microsofts-quantum-mechanics/

[5] Tom Simonite: IBM Shows Off a Quantum Computing Chip. MIT Technology Review, April 29, 2015www.technologyreview.com/s/537041/ibm-shows-off-a-quantum-computing-chip/

[6] A. Odeh, K. Elleithy, M. Alshowkan, E. Abdelfattah: "Quantum Key Distribution by Using Public Key Algorithm(RSA)" London, United Kingdom: Third International Conference on Innovative Computing Technology (INTECH), August 2013.

[7] Daniel Kudrow, Kenneth Bier, Zhaoxia Deng, Diana Franklin, Yu Tomita, Kenneth R. Brown, and Frederic T. Chong: Quantum Rotations: A Case Study in Static and Dynamic Machine-Code Generation for Quantum Computers. 2013 International Symposium on Computer Architecture (ISCA-2013).

[8]http://www.wsj.com/articles/intel-to-invest-50-million-in-quantum-computers-1441307006

[9] https://ec.europa.eu/digital-single-market/en/news/european-commission-will-launch-eu1-billion-quantum-technologies-flagship

5. Beyond CMOS

Nanotubes

Carbon nanotubes (CNTs) are tubular structures of carbon atoms. These tubes can be single-walled (SWNT) or multi-walled nanotubes (MWNT). Their diameter is in the range of a few nanometers. Their electrical characteristics vary, depending on their molecular structure, between metallic and semiconducting [1].

A CNTFET consists of two metal contacts which are connected via a CNT. These contacts are the drain and source of the transistor. The gate is located next to or around the CNT and separated via a layer of silicon oxide [4].

Current state: In September 2013, Max Shulaker from Stanford University published a computer with digital circuits based on carbon nanotubes. It contains a 1 bit processor, consisting of 178 transistors and runs with a frequency of 1 kHz. [2]

Nanotube-based RAM is a proprietary memory technology for nonvolatile random access memory developed by Nantero (this company also refers to this memory as NRAM) and relies on the effect that nanotubes lying cross over can either be touching each other or are slightly separated, depending on their position. A NRAM "cell" consists of a non-woven fabric matrix of CNTs located between two electrodes. The resistance state of the fabric is high (representing "off" or "0" state) when (most of) the CNTs are not in contact and is low (representing "on" or "1" state) vice versa. To switch the NRAM between states, a small voltage greater than the read voltage is applied between top and bottom electrodes. In theory NRAM can reach the density of DRAM while providing performance similar to SRAM. [5]

Perspective: It will take a an unknown number of years before NRAM drives might be in production stage [3].

Impact on hardware: CNTs can be utilized for a lot of different applications in several areas of research. The most promising ones for HPC are the construction of carbon nanotube field-effect transistors (CNTFETs), nanotube-based RAM (or Nano-RAM) and the improvement of chip cooling. CNTs are very good thermal conductors. Thus, they could significantly improve conducting heat away from CPU chips [6].

References

[1] https://en.wikipedia.org/wiki/Carbon_nanotube

[2] Max M. Shulaker (Stanford University, Stanford) et al.: Nature, http://www.nature.com/nature/journal/v501/n7468/full/nature12502.html

[3] http://www.computerworld.com/article/2929471/emerging-technology/fab-plants-are-now-making-superfast-carbon-nanotube-memory.html

[4] Lorraine Rispal: Large Scale Fabrication of Field-Effect Devices based on In Situ Grown Carbon Nanotubes. Dissertation, Technische Universität Darmstadt, http://tuprints.ulb.tu-darmstadt.de/2021/

[5] https://en.wikipedia.org/wiki/Nano-RAM

[6] http://www.extremetech.com/extreme/175457-this-carbon-nanotube-heatsink-is-six-times-more-thermally-conductive-could-trigger-a-revolution-in-cpu-clock-speeds

Graphene

In 2010 two physicists at Manchester University in the U.K. shared a Nobel Prize in Physics for their work on a new wonder material: graphene, a flat sheet of carbon with the thickness of a single atom. Konstantin Novoselov and Andre Geim discovered the material by applying plain old sticky tape to simple graphite [1].

Graphene grows on semiconductor i.e. on the surphase of a germanium crystal, which is seen as big step towards manufacturability, see [5, 6].

Current state: In 2010, IBM researchers demonstrated a radio-frequency graphene transistor with a cut-off frequency of 100 Gigahertz. This is the highest achieved frequency so far for any graphene device. In 2014, engineers at IBM Research have built the world's most advanced graphene-based chip, with performance that's 10,000 times better than previous graphene ICs. The key to the breakthrough is a new manufacturing technique that allows the graphene to be deposited on the chip without it being damaged [4].

Graphene Project is an EC Flagship project with considerable research efforts in making graphene useful, however, still focused more on the material science perspective than on its potential usage for future computer technology. Graphene is among the strongest materials known and has attractive potential also outside of computer technology, e.g., as electrodes for solar cells, for use in sensors, as the anode electrode material in lithium batteries and as efficient zero-band-gap semiconductors [2].

Perspective: Graphene is a promising technology in laboratory. Due to the fact that the new graphene manufacturing method is actually compatible with standard silicon CMOS processes, it will probably be possible to realize commercial graphene computer chip in future [4].

Since in its current form graphene is not suitable for transistors, researchers have been working on a way to convert it for this use.

Impact on hardware: Graphene has an excellent capacity for conducting heat and electricity.

References

[1] Moskvitch, Katia. A Graphene Discoverer Speculates on the Future of Computing. [Online] Scientific American, January 23, 2015. http://www.scientificamerican.com/article/a-graphene-discoverer-speculates-on-the-future-of-computing/.

[2] Rodewald, Mike. Researchers discover method for mass production of nanomaterial graphene. [Online] November 10, 2008. http://newsroom.ucla.edu/releases/method-for-mass-production-of-70969.

[3] IBM. Made in IBM Labs: IBM Scientists Demonstrate World's Fastest Graphene Transistor. [Online] February 5, 2010. http://www-03.ibm.com/press/us/en/pressrelease/29343.wss.

[4] Anthony, Sebastian. IBM builds graphene chip that's 10,000 times faster, using standard CMOS processes. [Online] January 30, 2014. http://www.extremetech.com/extreme/175727-ibm-builds-graphene-chip-thats-10000-times-faster-using-standard-cmos-processes

[5] http://hackaday.com/2015/10/14/graphene-grown-on-semiconductors-big-step-toward-manufacturability/

[6] http://www.nature.com/ncomms/2015/150810/ncomms9006/full/ncomms9006.html

Diamond Transistors

Diamonds can be processed in a way that they act like a semiconductor. Diamond based transistors can be fabricated.

Current state: Researchers at the Tokyo Institute of Technology fabricated a diamond junction field-effect transistors (JFET) with lateral p-n junctions. The device shows excellent physical properties such as a wide band gap of 5.47 eV, a high breakdown field of 10 MV/cm (3–4 times higher than 4H-SiC and GaN), and a high thermal conductivity of 20 W/cm*K (4–10 times higher than 4H-SiC and GaN). It has been found that this diamond transistor works with excellent electrical characteristics, up to 723 K [1].

Perspective: Currently the gate length of the fabricated diamond transistors is in the single-digit micrometer range. Compared with the current 22nm technology with gate lengths of about 25nm [2], a reduction in size is absolutely necessary in order to allow fast working circuits (limitation of the propagation delays).

Producing reasonable diamond wafers for mass production could be possible with the method of [3]. The time for producing diamond wafers is another factor that has to be reduced drastically to compete with other technologies.

Impact on hardware: The high thermal conductivity of diamond, which is several magnitudes higher than that of conventional semiconductor material, allows faster heat dissipation. This could solve the temperature problem of stacked dies. Switching energy of a diamond based

semiconductor is expected to be much smaller than silicon and the maximum operating temperature can be much higher. It may "revive" the traditional Moore law.

References

[1] T. Iwasaki et al., "High-Temperature Operation of Diamond Junction Field-Effect Transistors With Lateral p-n Junctions," in IEEE Electron Device Letters, vol. 34, no. 9, pp. 1175-1177, Sept. 2013.

[2] https://en.wikipedia.org/wiki/22_nanometer

[3]Aida, H., Kim, S. W., Ikejiri, K., Kawamata, Y., Koyama, K., Kodama, H., & Sawabe, A. (2016). Fabrication of freestanding heteroepitaxial diamond substrate via micropatterns and microneedles. Applied Physics Express, 9(3), 035504.