2021 NSF Workshop on CMOS+X Technologies

Cointegration of CMOS with Emerging Technologies (X)

Organizers: Sankar Basu[1], Pinaki Mazumder[2], Sayeef Salahuddin[3]

Workshop Report

August 17-18, 2021

[1]*NSF,sbasu@nsf.gov*

[2]*NSF, pmazumde@nsf.gov*

[3]*UC Berkeley, sayeef@berkeley.edu*

Executive Summary

The semiconductor industry was created in the U.S. and over the last many decades the U.S. has enjoyed unfettered leadership in this space. However, its leadership in manufacturing chips has consistently eroded and it is now facing steep competition from around the world. With the erosion of manufacturing capability, facilities for advanced research on semiconductor technology have also seen substantial lack of investment. In fact, currently there are no U.S. facilities where exploratory research can be performed on advanced CMOS devices at the relevant dimension and at the relevant scale. This means that many U.S. innovations are tested at offshore facilities, leading to the loss of intellectual property. In some cases, the U.S. researchers start from prefabricated wafers obtained from foreign Foundries and integrate their own technology on top. While this method may be effective in a small number of cases, it is extremely limiting for most because it inherently precludes frontier research on advanced CMOS. In this context, it is important to note that the roughly 452 billion dollars of estimated revenue of the semiconductor industry in 2021 is predominantly derived from advanced CMOS.

On the other hand, recent trends in the semiconductor industry and research show that the computing hardware of the future will be built upon an intimate combination of CMOS and other emerging devices (we will call them the 'X' devices) that exploit phenomena such as ferroelectricity, ferromagnetism, spin, phase transformation etc. and materials such as oxide, nitride, carbon, chalcogenides etc. as transistor channel – a possibility that would not be considered seriously for computing just a few years ago.

Not having access to advanced CMOS research facilities disadvantages US researchers in 'X' research as well as the inability to integrate X with CMOS often takes away the very impact that the X technologies would otherwise have. Put together the challenges are so daunting that it discourages U.S. students to pursue a career in the semiconductor industry. Losing access to the smartest students will make it impossible for the U.S. to compete in this space in the future.

To revitalize the U.S. Semiconductor Industry it is, therefore, imperative to develop infrastructure where new ideas for future computing hardware can be explored, fabricated and tested at a relevant scale. This will need these facilities to be able to fabricate both advanced CMOS and new technologies (hitherto referred as the 'X') and integrate them in a seamless manner. Indeed, such facilities will greatly enhance U.S. competitiveness in the semiconductor research.

This Workshop on CMOS+X: Cointegration of CMOS with Emerging Technologies (X) brought together an international group of thought leaders and researchers on semiconductors with expertise covering Si CMOS, X devices and their integration, to discuss opportunities, known pitfalls and potential pathways for establishing U.S. capabilities that will meet this critical national need.

Organization of the Workshop:

The workshop took place on August 17 and 18 and consisted of three panels, two on the first day and one on the second day. It was organized by Dr. Sankar Basu and Dr. Pinaki Mazumder from the National Science Foundation, Dr. Sayeef Salahuddin from the University of California Berkeley and Dr. Damian Dudek from the Deutsche Forschungsgemeinschaft (DFG), German Research Foundation.

Each panel had 6 presentations from invited panelists covering roughly one hour, followed by another hour of open discussion. Panel 1 was titled Si technologies where panelists talked about the state of the art and the future roadmap of Si based technologies such as CMOS logic, RF etc. Panel 2 was titled X technologies where panelists presented various emerging devices such as quantum devices, magnetic memory, ferroelectric materials etc. Finally, Panel 3 was titled CMOS+X and Novel Functionality where panelists discussed integration of CMOS and X technologies that provide substantial improvement in terms of augmented functionality, energy efficiency and performance. The makeup of the panels is shown below:

August 17, 2021

Panel 1: Si Technologies ***(9:15 – 11:15 AM) : Moderator: Sayeef Salahuddin***

— Daniel Armbrust (Silicon Catalyst)
— Tahir Ghani (Intel)
— David Thompson (Applied Materials)
— Adrian Ionescu (EPFL)
— Gerd Fettweis (TU Dresden)
— Ullrich Pfeiffer (University of Wuppertal)

Panel 2: X Technologies ***(11:15 – 1:15 PM):Moderator: Muhammad Hussein***

— Heike Riel (IBM)
— Daniel Worledge (IBM)
— Max Lemme (RWTH Aachen University)
— Thomas Mikolajick (Namlab, TU Dresden)
— Rainer Waser (RWTH Aachen University)
— Dmitri Nikonov (Intel)

August 18, 2021

Panel 3: CMOS+X and Novel Functionality ***(9:00 – 11:00 AM) : Moderator: Suman Datta***

— Tsu-Jae King Liu (UC Berkeley)
— Naresh Shanbhag (UIUC)
— H.-S. Philip Wong (Stanford)
— Takashi Ando (IBM)
— Robert Weigel (FAU Erlangen Nürnberg)
— Arijit Raychowdhury (Georgia Tech)

Panel Summaries:

The following provides brief description of the three panels:

Panel 1: Si Technologies:

Si technologies form the core of computing hardware. However, as mentioned before, the U.S. facilities for academic research are severely outdated. As we think about how to develop next generation facilities, one effective way could be to create large centers where people from a broader spectrum are working together in, perhaps, a center or consortium. Today's research is often material inspired. However, experience in semiconductor industry shows that focusing on one material is not good enough. Without the knowledge of the entire ecosystem, research often veers into avenues that are considered irrelevant.

If an open platform of innovation can be developed, where researchers can explore innovation in physics, chemistry, materials, device design etc. within the context of a real device and at relevant scale, that could be a game changer. For example, is it possible to develop a facility that has an open-access platform for 20 nm channel length, gate-last devices? If it were, research could be done, e.g., for better contact resistance, better gate resistance, increasing mobility of the channel, better interconnect etc. – all within the relevant device scale and in the context of integrated process recipes. It is understood that, building such an open platform is going to be highly challenging; due to tremendous impact of these innovations, industries keep their process recipes a secret. However, even if the open platform technologies cannot have all the performance boosters of an industrial transistor, it could still be highly impactful in both advancing CMOS research and in attracting students by providing a pathway for immediate impact of their innovations.

One example is IMEC in Belgium that has established itself as one of the world-wide leaders in semiconductor research over the last two decades. It follows a hybrid model where it receives a core funding from the government, but most of the funding comes from membership fees from industry affiliate programs. It employs both full time engineers and graduate students and post-doctoral researchers. While the IMEC model may not be directly applicable for all cases, it serves as a model example of a large scale center for semiconductor research.

In Germany, creating a hierarchy of funding from European to province level has proven to be effective. Furthermore, focused centers on basic research such as the Max Planck Institutes and on applied research such as the Fraunhofer Institutes facilitate proper distribution of resources. Experience in Europe also shows that putting resources where students can freely innovate will be critically important. Local centers of excellence with people from different points of interest could prove to be very effective.

Today's students are also highly motivated by broad societal impact. 'Open source' protocol is therefore of high importance. As new facilities are established, it is necessary to explore and develop protocols for open innovation so that new research can stand on all the accumulated knowledge of previous work; thereby fostering impact and accelerated innovation.

Panel 2: X Technologies:

Within the CMOS+X paradigm the X covers a wide set of technologies. The computing workload in the recent years, dominated by AI, has become highly diverse. This is very different from the bygone era when, e.g, transactional load will constitute the main computational challenge. Deep learning and neuromorphic computing are starting to become prominent. For each, there are many different technologies that show promise. Quantum computing is one example where the computing itself is completely different. Right now, it may not be clear if Quantum Computing could do something that classical computers cannot, but it is only early days and continued research could potentially lead do substantial performance boost in the future.

It is clear that integration and realization of X technologies will require substantial research on materials. It is important, however, that X should not be defined in terms of materials. Rather X should be defined in terms of technologies and material, integration, tooling etc. should be explored within the context of that specific technology. In this context, it is also important to remember the cost effectiveness. Not every technology will warrant a pilot line. Exorbitant cost in a 300 mm line could easily become the bottleneck for exploratory research.

CMOS+X also involves an end-to-end framework that goes all the way from materials, devices to circuits and architecture design. In the same vein, it is important to note that device-to-system/application level research might only be disruptive if it can be shown to serve a broad set of applications and thus driving new platform-capable technologies.

In terms of X technologies and their integration concepts, that would be regarded too hard a couple of decades ago, are the ones that are being pursued today. This poses the challenge as to how the universities can remain engaged in relevant research. Creating centers around a specific technology and tooling the centers with the appropriate equipment (often too expensive today to acquire and maintain for an individual faculty) is a plausible way to move forward. Such centers can bring together people from different but related background such as materials, devices, measurement, circuits and system design and at the same time enable access to appropriate equipment. On the academic side, faculty will also have to come out of their traditional lone-wolf style of work and learn to reach out and work with each other.

Panel 3: CMOS+X and Novel Functionality:

CMOS+X systems are examples of what can be accomplished with an innovative integration of CMOS with an emerging technology. The panel discussed CMOS+NEMS, CMOS+MTJ, CMOS+Phase change memory and CMOS+quantum systems that offer new functionality and potentially orders of magnitude improvement in energy efficiency

CMOS+X offers exciting opportunities but it is very important to note that, to realize this potential, it will be necessary to have platforms where it is possible to work, optimize and advance both CMOS and X. In Europe, there are many activities focused on X but not so much on CMOS. As a result, most technologies do not really make an eventual impact. The important of knowing the fundamentals on CMOS cannot be overstated. All X technologies need some version of CMOS for their enablement. It is also very important to know CMOS fundamentals before students can think about technologies that can exceed CMOS capabilities. This will need infrastructure that enable frontier research on both CMOS and X. NSF has experience in establishing and operating large infrastructure in many different fields such as space, oceanography etc, and therefore may be the best poised among the U.S. federal agencies to build such infrastructure and maintain and operate them.

X technologies can help attracting students with broader diversity to the field. Students today are concerned about the impact of their work on life in general. If proper connections can be made between semiconductor technology to more sustainable, more equitable, and a healthier future, that will help substantially to attract the smartest and brightest students to this area. Semiconductor technology, of course, is the main driver for a sustainable and equitable future, but it needs to be packaged appropriately so that that connection becomes clear. Students want to have an impact on compelling societal problems. At the device and circuit level, the gap between what the students actually study in the classroom and how it impacts an important societal problem is often too large.

One way to approach this problem will be to focus on the fundamentals in our teaching rather than the latest fashion of the day. In fact, the current practice in academy is just the opposite where faculty mostly teach topics that are directly related to their own research –topics that are almost by definition highly specialized. One of the difficult challenges for students in the semiconductor technology is the long-winded nature of getting to the end results. After designing a chip, it takes more than 3 months to get that chip built. If the students are working in a lab to fabricate devices, it takes even longer to reach a meaningful result. The community needs to come together with NSF to think about how this process can be shortened.

Recommendations based on presentations and open discussions:

Presentations and discussions in the three panels indicate that functional augmentation of CMOS with emerging technologies (X), or the CMOS+X platform, is poised to become the building block of computing hardware in the future. Two decades ago NSF helped set up the NNIN centers across the nation. These centers have played a very important role in providing critical infrastructure for nanoscience research. However, the goal of the NNIN centers was to enable research in diverse set of scientific principles including physics, chemistry, materials and engineering. Research on CMOS+X will require focused infrastructure that enable research on advanced and exploratory semiconductor technology. In the following some summary recommendations based on the workshop are presented:

1. Computing hardware has evolved in a way that many activities that would be regarded too complex and therefore be discarded a few decades ago are becoming increasingly mainstream. This indicates that exploratory research in universities and other places now need access to complex equipment which are expensive to both acquire and maintain. There was almost a unanimous consensus among the panelists that NSF should play a major role in the development and continued operation of such infrastructure. NSF has prior experience in establishing similar infrastructure in other fields, including the NNIN centers.

2. While establishing such infrastructure:
    a. It is important to remember that platforms are needed that can enable frontier research in both CMOS and X. This is critical as co-optimization of both is necessary to unleash the full potential of CMOS+X platform.
    b. Cost effectiveness is an important consideration. Running in a 300 mm pilot line could be prohibitive due to the cost involved. In fact, at the initial stage, every technology does not need such a pilot line. Considering cost effectiveness and also the need to establish scalability, 200 mm facilities present a decent tradeoff.
    c. Hierarchical facilities in Germany have shown good effectiveness. Establishment of research infrastructure, e.g. in universities, in appropriate locations with proximity to industries such as in Dresden has proved to be a very effective model.

3. One of the major challenges is also that research funding is targeted almost completely towards new projects. However, enablement of research in semiconductor technology involves maintenance and operation of complex infrastructure that involve paying for engineers and often for technicians. In addition, maintaining tool chains for chip design is another such activity which is essential but almost impossible to fund. NSF should address this challenge,

perhaps by creating programs around infrastructure (point 1) that can fund maintenance of such infrastructure in addition to funding research projects.

4. The complexity of the technologies, need for compatibility with 100's of processes, and the diversity of the modern computing workloads indicate that center level effort will be needed for success. Recommendations include:
    a. Centers should be formed around technology and *not materials*. Material, synthesis, processing research should be done in relevance to the technology.
    b. Centers should involve researchers from device and materials technology all the way to the circuits and systems, but they should be rooted around one specific technology. It is important to bet on a few ideas and build separate, large efforts around each one of them.
    c. It is important to find models to involve industrial presence in such centers. This is specially important so that the research is guided by appropriate awareness of the ecosystem that will be needed to make that research successful.

5. Being able to attract the smartest students is imperative for continued success of semiconductor technology in the future. All panelists talked passionately about this issue:
    a. NSF should explore ways to shorten the loop from conception to result for students working in the semiconductor technology. This is related to establishing and maintaining research infrastructure. In addition, NSF could explore models to create and maintain an open innovation platform where researchers can focus on one specific aspect of the platform without having to worry about building every other part of it. Such a platform will have to be developed with close collaboration with industrial partners to ensure relevance.
    b. Students need a holistic education that go all the way from devices to end application so that they can appreciate the impact of their work on important societal problems. NSF could help creating such a curriculum by bringing the community together from all over the nation. This, in nature, will be very similar to curriculum often developed by NSF centers on specialized research topics.

CMOS + X: Current capabilities:

At the present, researchers take resort to a number of different approaches for CMOS+X research. A small number of university researchers has access to wafers from foundries through ad hoc personal connections. In this approach, typically, the CMOS driving circuitry is produced in the Foundry. Next, back-end-of-the-line (BEOL) compatible X devices are integrated to these CMOS. For example, NEMS switches have been integrated to TSMC 16 nm in this approach [1]. Similarly, RRAM circuits have been integrated with 130 nm CMOS [2]. It is to be noted that this approach requires alignment marks on the wafers so that university made devices on top of the CMOS circuits can be properly connected. This leads to substantial barriers – most often Foundries are not willing to provide alignment marks. Additionally, it is very difficult to align to contacts with small pitch due to registration differences between two completely different lithography tools. Finally, this approach also requires substantially complicated legal process between a Foundry and a University.

Another approach has been to use a 'two-chip' solution where CMOS circuitry fabricated in a Foundry has been used together with university-made devices by directly wiring up the contacts of the two different chips (such as in MEMS). This solution is the most inefficient as such wiring immediately dominates the performance of the overall system and any finegrain connection is impossible.

In a very small number of cases, where X devices are nearing commercialization (e.g. RRAM, MRAM, Si photonics), university researchers with special agreements may tape out CMOS+X circuits at foundries [3]. Nonetheless, such access, even when it is possible, is limited only to the macro level – a it is not possible to try new circuit topologies using individual transistors and memories, let alone to customize and improve the device characteristics. In addition, so far this has only been possible through large programs, mostly from DARPA, as these tapeouts are prohibitively expensive for an academic researcher.

Only a few universities have the capability to do a true CMOS+X integration but that is typically limited to single transistors. For example, direct integration of CMOS compatible ferroelectric material onto the gate of an advanced transistor has been demonstrated [4,5]. Similarly, integration of ferroelectric material directly onto the gate as well as a capacitor with BEOL compatible transistors have been demonstrated [6]. The difficulties involved in such integration comes from the fact that fabrication of a transistor with reasonable performance is very complex. For example, fabrication of one single transistor, with Lg=20 nm, following the gate last integration process takes larger than 250 processing steps. For this reason, despite success in CMOS+X integration with single transistor or a few transistors [7], currently with the available capabilities at the universities, it has not been

possible to integrate 100's-1000's devices with relevant gate length and pitch to build CMOS circuity that can demonstrate new functional prototypes.

As indicated above, it is possible to exploit Foundry capabilities for some X technologies nearing maturity. This will need special arrangements between the Foundries and researchers that the NSF can facilitate. However, the true potential of CMOS+X can only be unleashed when both CMOS and X can be integrated without restriction. Given the current capabilities in the US universities, small programs where certain X technologies are integrated with an underlying transistor can be considered. For example, a gain-cell memory can be constructed using a X based memory device (Ferroelectric, Ferromagnetic, resistive etc.) and a single transistor. The main goal of such a program will be to establish the feasibility of co-integration. In addition, such programs can demonstrate the possibility of CMOS processes up to M2 or M3 level. Models, calibrated to measured data from such small demonstrations, can be developed to project realistic system level performance.

Longer term programs should consider multiple centers of excellence where CMOS+X vehicles with more than 1000 devices can be prototyped. These programs should consider continued support so that the developed process recipe and the tool infrastructure can be maintained. Notably NSF has provided continued staff support for NNIN facilities. Multiple centers across the nation should be created in a networked fashion where complimentary capabilities can be leveraged and appropriate protocols are set in place to exchange wafers at any point of processing.

References:

U. Sikder et al., "Toward Monolithically Integrated Hybrid CMOS-NEM Circuits," in IEEE Transactions on Electron Devices, doi: 10.1109/TED.2021.3122404.

T. F. Wu *et al.*, "14.3 A 43pJ/Cycle Non-Volatile Microcontroller with 4.7µs Shutdown/Wake-up Integrating 2.3-bit/Cell Resistive RAM and Resilience Techniques," *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2019, pp. 226-228, doi: 10.1109/ISSCC.2019.8662402.

Naresh Shanbag et al, DARPA FRANC program

A. J. Tan *et al.*, "Ferroelectric HfO2 Memory Transistors With High-κ Interfacial Layer and Write Endurance Exceeding 1010 Cycles," in *IEEE Electron Device Letters*, vol. 42, no. 7, pp. 994-997, July 2021, doi: 10.1109/LED.2021.3083219.

D. Kwon *et al.*, "Near Threshold Capacitance Matching in a Negative Capacitance FET With 1 nm Effective Oxide Thickness Gate Stack," in *IEEE Electron Device Letters*, vol. 41, no. 1, pp. 179-182, Jan. 2020, doi: 10.1109/LED.2019.2951705.

S. Dutta *et al.*, "Monolithic 3D Integration of High Endurance Multi-Bit Ferroelectric FET for Accelerating Compute-In-Memory," *2020 IEEE International Electron Devices Meeting (IEDM)*, 2020, pp. 36.4.1-36.4.4, doi: 10.1109/IEDM13553.2020.9371974.

A. J. Tan *et al.*, "Experimental Demonstration of a Ferroelectric HfO2-Based Content Addressable Memory Cell," in *IEEE Electron Device Letters*, vol. 41, no. 2, pp. 240-243, Feb. 2020, doi: 10.1109/LED.2019.2963300.

APPENDIX

Panel 1: Si Technologies

2.1 U.S. Leadership in Advanced Microelectronics and Computing

Daniel Armbrust (Silicon Catalyst)

While Moore's law is ending, solid roadmaps remain for logic, memory, interconnects, and packaging [1, 2]. At present, there is no high probability candidate to replace the mainstream CMOS technology, and multiple paths need to be pursued for the next generation candidates, non-disruptive to the existing technology base. The challenge is to accelerate the pace of innovations and commercialization of these new candidates so that the U.S. can maintain leadership in semiconductor and advanced computing technologies.

To achieve advancements in computation, the entire system needs to be re-evaluated in the context of new materials and devices, their non-disruptive nature with conventional technologies, and innovations at the system and algorithm levels. For example, the deep neural networks-related innovations required us to go back down the stack and optimize both software and hardware. Similarly, promising logic and memory innovations driven by fundamentally new materials and physics need to be translated into new types of circuits and exploit them in higher-level algorithms, which are not within the capabilities of an individual company.

We need to build an eco-system by putting together the skills and interests across various government agencies, academic institutions, and technology industries. Currently, there is significant funding in fundamental research from various government and industry sources. However, there is also a significant gap between the early-stage research and the ability to translate them into something that the industries will pick up. The innovation gap includes lack of resources, prototyping an invention at scale, de-risking the technology elements, wrapping the idea around the suitable business model, and eventually transferring the technology to the industry. Over three years ago, I put together a proposal for creating such a public-private partnership among industry, academia, and national laboratories under the governance of various government agencies, that will make state-of-the-art facilities accessible for innovators and ensure entrepreneurial events of taking these early ideas and de-risk them.

Silicon Catalyst, established in 2015, worked as a startup accelerator by assembling a broad eco system, which includes over 200 semiconductor veterans serving as advisors, 55 stake-holders in industries as in-kind and strategic partners, and over 300 venture capital investors. It admitted >60 companies (new and serial entrepreneurs) into at least 24-month program after screening over 600 early-stage startups and significantly lowered the up-front cost and time of building the first prototype. Due to the unique eco system offered by Silicon Catalyst, these startups raised over $400 million from investors, in-kind, and grants. Now silicon catalyst has expanded globally to Israel, United Kingdom, and a power-electronics focused joint-venture (Si Power) in China and engaging a number of university researchers for potential entrepreneurs through speaking opportunities, judging competitions, offering them mentorships and internships.

2.2 Infrastructure and progress options for scaling silicon based transistor technology

Tahir Ghani (Intel)

Transistor scaling has enabled significant improvements in computational power, driven by materials and devices research. Figure 1 shows various transistor structures adopted by industries over the last ten years, mainly instigated at Intel. The first significant change in the transistor was the introduction of strain in the transistors in 2003, especially a uniaxial strain just along one direction rather than having the symmetric strain in a bidirectional fashion. Such a strained structure was developed for the 90 nm node. It was generally believed that the way to get strain into a MOS transistor is by growing the silicon germanium or silicon and doing pseudomorphic growth where the top ten layers will be under biaxial strain. There was a lot of effort by the industry and universities at that time to demonstrate the viability of that process. Intel introduced a structure of a completely different way of putting in strain, which is much more effective in terms of the amount of gain obtained through strain engineering. This was surprising for the rest of the device community at that time.

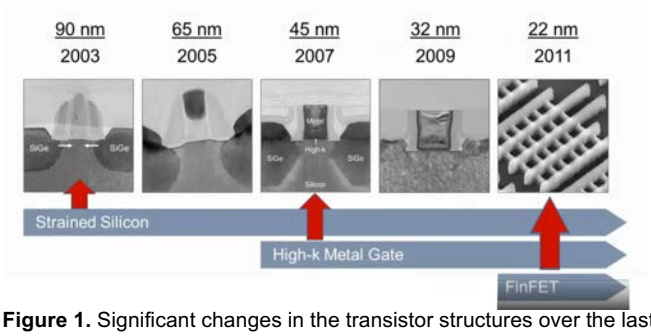

Figure 1. Significant changes in the transistor structures over the last ten years.

The most significant change was introduced around 2007 by introducing a high-k metal gate, which changed the fundamental ways we manufacture transistors. It required the gate at the last process, where typically the gate was done at first, so it inverted all of that flow. It also enabled the integration of many new materials into the transistors in a very well-controlled manner. Many issues needed to be overcome by understanding those materials and underlying physics, and eventually, it was implemented in the 45 nm technology node. The third change was in 2011, which was the introduction of FinFET into volume manufacturing. Researchers at that time were proposing various complicated structures to move away from a planar geometry. The most efficient structure, the FinFET, came from academia when UC Berkeley first presented results on FinFET in 1999 at IBM. This landmark invention shows the importance of industry-academia partnerships. The FinFET structure changed the whole envelope of transistor scaling. As we move forward, we need to think about a broad set of ideas to advance, more suitably pursued in academia.

Intel knew in the late 1990s that the age of extensive scaling is coming to an end, and to make substantial progress moving forward, we as a community need to come up with more innovative changes than just incremental scaling of the transistors. In response, Intel restructured its R&D model to expedite rapid innovations from fab prototypes to high volume manufacturing on schedule. The new organizational structure was critical to Intel's success in introducing revolutionary transistor innovations to the market. The new R&D structure had two components:

component research team and pathfinding team. A pathfinding team working on a technology node starts early in, focusing on selecting technology targets, process features, and completing design rule definition. The viability of each project was decided based on silicon learning. The success criteria to get selected to the development phase entails demonstrating agreed-upon success criteria, including a defect density number by the end of pathfinding. There were significant overlaps between component research and pathfinding and between pathfinding and development programs to ensure a smooth transition.

There is still significant room to scale transistor density beyond the current node in production. The semiconductor industries, including Intel, are trying to move to the next structure quickly, and promising candidates are gate all around and nanowire or nanoribbon FETs. Beyond ribbon FETs, 3D stacked transistors are also promising where we will have an n-type transistor on top of the p-type or vice versa. There is no bonding involved, but all are self-aligned monolithic stacked CMOS structures. These stacked structures will again bring a significant change, and continuous transistor innovations will enable 50% smaller cell and SRAM cell size and provides better performance per Watt for future computing systems. However, heat dissipation is a critical concern that needs to be carefully evaluated for 3D stacked nanowire or nanoribbon devices. The next step will be to combine these elements to 2D materials to get the ultimately scaled gate length device. For the interconnect side, we need to move away from copper for the very low levels of the metal and move towards subtractive metals like Mo or Ru.

2.3 New era in lithography scaling

David Thompson (Applied Materials)

The semiconductor industry brought us exponential improvements in computer performances over the last 70 years, mainly through innovations developed in the United States. For example, the first commercial digital computer (UNIVAC I) back in 1951 could perform 1900 floating-point operations per second (flops) for 125 kW. Supercomputers today can perform 147.6 petaflops (SUMMIT, IBM), which is about a factor of 100 trillion improvements in performance, for only 100x more power. The first 40 years of that innovation were largely built on two things. Firstly, the advent of the transistor from the home-grown work in the Bell Labs in the United States eventually replaced the vacuum tubes. Secondly, the sheer power of Moore's Law and lithographic scaling enabled unprecedented shrinking in device size and improvements in power, performance, and cost.

In the last 25 years, with the advent of compartmentalization, the confluence of lithographic scaling with materials intensity has been lifted. This is mainly because of the introduction of new materials with better conductivity for the interconnect network within the sheer array of switches present on a chip. Around the mid-2000s, the transistor gate dielectric was also replaced by a high-K dielectric material for better performance. The industry now has migrated into a new era where there will be continued incremental gains in lithography. The power performance cross trinity stems from innovations in new architectures, materials, and devices, entirely laid out in different fashions and going from a planar to 3D geometry.

As we are migrating from a purely lithographic scaling space where the delivery of a technology node without new materials introduction itself is challenging, the amount of effort required to develop and engineer new materials with the chip stack increases almost exponentially. About three to five hundred additional processes are needed that require different thermal budgets, considers different materials, interfaces and contacts, different solution phases, and different gas and plasma environments. The challenge in enabling these processes is understanding how those materials work, identifying a precursor to depositing material with a specific conductivity, characteristic, adhesion, and stress for a particular application.

It is a fascinating time because there is a beautiful renaissance of metallurgy and material science taking place on an entirely new scale and one of the first industrial revolutions we saw in the United States. To facilitate these types of innovations, we need to build a significant infrastructure with many skilled and talented scientists having diverse expertise and trained on these processes. The infrastructure required is no longer a simple type of facility that academic researchers independently set up inside their laboratory on their own. Instead, we need to make sure that the capabilities and infrastructure exist to have leading-edge technology, giving researchers access to total flows that will enable them to engineer a portion of the workflow in CMOS on legacy equipment and see the type of impact they could have.

2.4 The potential of agile innovation CMOS (+X) silicon platforms

Adrian M. Ionescu (Ecole Polytechnique Fédérale de Lausanne)

The scaling of technology nodes in silicon platforms has brought tremendous improvements in modern computation. However, as we are approaching the end of the scaling for CMOS field-effect transistors (FETs), we need various technology boosters with improved energy efficiency to keep driving the improvements. Gate-all-around FETs and nanosheets, which can be integrated into three-dimensional circuits, promise to meet the technological scaling goals beyond 2025 to reach near 1 nm node. However, the notion of the node itself does not make sense as a figure-of-merit [3], and we need new metrics to assess the progress technology, e.g., the densities of logic, main memory, and the interconnect, plus some other figure-of-merits like energy efficiency that quantifies device to system level performances for a given technology node.

Even with the continuing technological node scaling, we still expect fundamental limitations from energy efficiency and power density, so there is room for progress in this domain. One of the promising technology boosters in this domain is negative capacitance in CMOS that can benefit from the progress in doped high-k dielectric that can offer a steeper subthreshold slope, substantially reduced off current, and improved overdrive. We improved the subthreshold swing by a factor of three, from 60 mV/dec to 20 mV/dec (see Fig. 2a), and also improved the overdrive [4] with such doped high-k dielectric materials. Despite some controversies in the dynamic characterization, negative capacitance encompasses all performance metrics, resulting in excellent energy efficiencies and compatibility with any FETs. Another promising technology booster in this domain is junction-less FETs that can be combined with negative capacitance. Control of a

highly doped and extremely thin junction-less channel can be significantly enhanced by negative capacitance, leading to a steeper subthreshold slope and enhanced overdrive [5], see Fig. 2a, and making these future devices attractive.

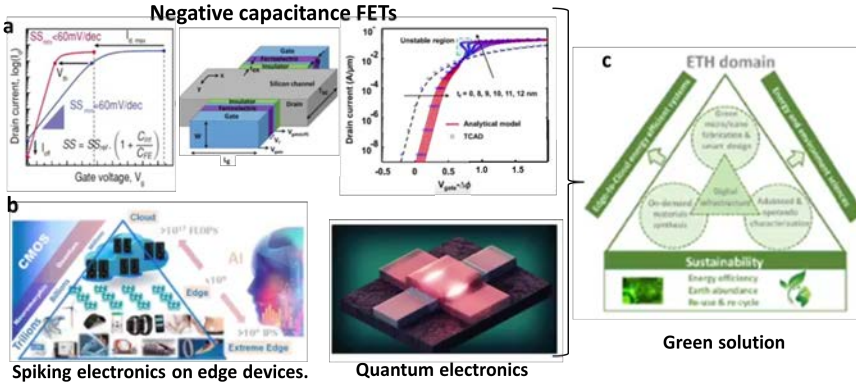

Figure 2. (a) Future of CMOS technology, negative capacitance as a booster technology. (b) New technology directions: spiking electronics for extreme edge devices and quantum electronics. (c) Proposal for a green sustainable CMOS technology platform.

On the other hand, we are processing more and more information from edge to the cloud, and we want edge devices to be autonomous and able to act in real-time. These devices cannot afford the latency to access the cloud back and forth. Moreover, these extreme edge devices are often battery-operated and need to be very energy efficient. Therefore, a possible paradigm change can be from traditional digital processing to bio-inspired information processing (see Fig. 2b), which is more like analog in nature. One possible route is to enhance CMOS with phase-change materials for energy-efficient spiking electronics (European Spike IoT Project), where sensing, processing, and communication will be unified in the spiking processes to realize architectures for end-to-end stochastic electronic functions on CMOS + phase change materials.

Quantum electronics is another domain of high interest where a functioning quantum processor needs to be scalable with millions of qubits. Even in this context, silicon-based technology seems to be the winner where we may now envision a solution to integrate both the qubit and cryo-CMOS in a scalable fashion and operate them together at a few mK to few K temperature range. One proof of concept from Hitachi consists of a silicon nanowire and two quantum dots to trap individual spins as qubits (Fig. 2b), controlled via electron spin resonance with microwave pulsing.

Finally, a digital infrastructure for synthesis, characterization, and fabrication of materials and devices is needed for a sustainable, energy-efficient, and environment-friendly green future for the metamorphosis of CMOS technologies. We must reduce the environmental impact on how we produce and consume materials and process these large-scale electronics. The green infrastructure proposal from the academic domain in Switzerland has the goal to envision an infrastructure (see Fig. 2c) that can enable a transition from automated to fully autonomous operation in materials processes and manufacturing and energy-efficient micro/nanofabrication with a target of ~30% reduction in materials waste by 2030, and ~30% reduction in hazardous chemicals and fabrication related energy consumptions by 2028.

2.5 Co-integration of CMOS with emerging technologies (X)

Gerd Fettweis (TU Dresden)

The German research landscape includes various institutes that bridge fundamental, strategic, and applied research. Long-term fundamental research is being run by Max Planck institute, shorter-term strategic research is being run by Helmholtz institute, and Leibniz Research Institute is bridging between the two. Fraunhofer institute is working on applied research and even advanced development that is looking at solving the problems in the development labs. Universities like TU Dresden, on the other hand, have the flexibilities to stretch among fundamental, strategic, all the way to applied research. Dresden have about 20 plus institutes from Max Planck, Helmholtz, Leibniz, and Fraunhofer, together they form the Dresden-concept. Four key branches of the Dresden-concept include biomedicine and bioengineering (SAC I), information technology and microelectronics (SAC II), materials and structures (SAC III), and culture and social change (SAC IV). The vision and topics of SAC II include the entire spectrum of electronic information processing from fundamental and applied research on data storage, processing, and transmission to the development and testing of new semiconductor materials, works on joint projects that make a difference, and often strengthens the local industries. SAC II's research topic includes all micro and nanoelectronics branches in the context of energy efficiency, beyond Moore's law scaling and integration, 5G-6G communication, IoT, big data, and artificial intelligence. Cool silicon was a 160-million-euro project focusing on advanced electronics with high energy efficiency, which was followed by the highly adaptive energy-efficient computing project. Based on this infrastructure, a new research institute, Barkhausen Institute, was formed to address the trustworthiness issues to make sure that democracy can happen in the future. Finally, all these were the foundation for the German equivalent of NSA. This is very different from the United States because in Europe, there is the possibility of getting the funding and building teams either on the European Union with clusters spread all over Europe or by using German funding within the country to have joint research projects between academia and industries. Finally, the German Science Foundation and others can even fund a local cluster like Dresden and work on big topics locally.

2.6 High Frequency and Communication Technology

Ullrich Pfeiffer (University of Wuppertal)

Terahertz (THz) frequencies are attractive for applications requiring ultra-high bandwidth. While the electronics community is scaling up in frequencies, the photonics community is scaling down in frequencies to enable technologies operating at the THz. Scaling up electronics systems in the THz regime offers several benefits, e.g., higher channel capacity, lower latency, new sensing phenomenology, higher spatial and temporal resolution, and increased dynamic range. However, the frequency scaling using existing technologies face challenges due to increased power consumption and thermal density. Also, we need to compromise on the system complexity, and systems are often bulky and costly. Thus, to leverage the benefits of THz electronics, we need to

develop the underlying transistor technologies to achieve higher-speed transistors, better-integrated circuit processes, better packaging technologies, and efficient system design tools.

There are several promising electron device technologies available for the THz regime, including III-V substrate (e.g., InP, GaN, and InP-GaAsSb), a silicon substrate (Si and SiGe), heterogeneous integration of III-V and silicon technologies (e.g., InP with SiGe), and electronic-photonic integration (e.g., Ge photodiode with silicon). III-V substrate technologies offer excellent performance, e.g., 25 nm InP HEMT technology have shown f_{max} of 1.5 THz, an amplifier gain of 9 dB at >1 THz, and can create around 3 mW of power at 850 GHz [6]. Pure silicon-based technologies offer much lower performance with f_{max} limited below 350 GHz. On the other hand, SiGe BiCMOS offers a high-speed heterojunction bipolar transistor (HBT), which has shown f_{max} of 700 GHz and can implement power amplifiers with 10 to 3 mW at 240 to 320 GHz. SiGe HBT started around 1995 with peak cut-off frequency < 100 GHz and operating voltage of 3.3 V. Since then, both the cut-off frequency and the operating voltage has been scaling (see Fig. 3), and we have achieved 500 GHz at 1.5 V around 2010 as an outcome of the 5dotfive project in Europe. Such a scaling enabled up to 160 GHz communication and radar application, and a follow-up project called 7dotseven pushed the performance further to enable THz imaging. Thus, the pathway is promising to scale up the silicon-based technologies to sub-mmWave frequencies.

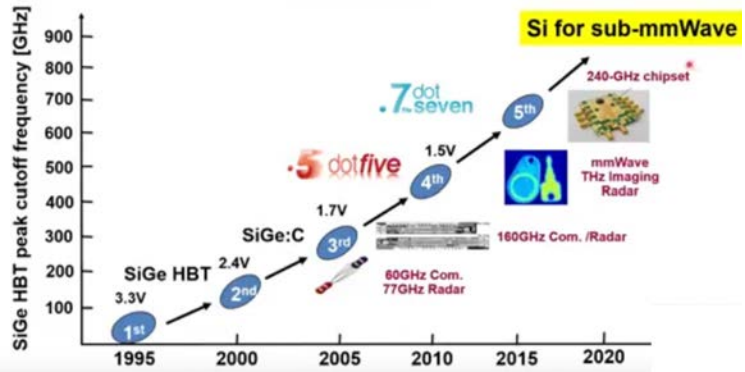

Figure 3. Roadmap for frequency scaling in the SiGe heterojunction bipolar transistor technology for THz applications.

Silicon technologies are already promising in various THz applications. A highly directive and fully integrated transmitter and receiver chipset has been demonstrated at 240 GHz using SiGe BiCMOS technology with a 110 Gbit/sec data rate in quadrature phase-shift keying modulation [7]. SiGe BiCMOS is also attractive for THz sensing and imaging for breast cancer detection applications [8]. THz imaging systems have also been developed using pure CMOS technologies to achieve THz light-field camera [9], and these chips can be integrated to achieve a light-field camera super array that has shown up to 9216 receiver channels at THz [9].

Panel 2: X Technologies

3.1 X = Quantum Computing

Heike Riel (IBM)

The future of computing comprises at least three technologies: the bits in conventional computers and supercomputers implemented with CMOS, the neurons in artificially intelligent systems, and the qubits in quantum systems. The power of quantum computing is based on using the laws of quantum physics like superposition, entanglement, and interference to achieve exponential performance. For example, in theory, n number of qubits can generate 2^n basis states for calculation, which enables a path to solve intractable problems. However, many problems in business and science are too complex for classical computers, including factorization, simulation of quantum systems, optimization, and graph problems, which a quantum computer can accelerate.

There are various qubit technologies to implement quantum computing, with different maturity and performance levels. Widely known technologies include superconducting wires, trapped ions, engineered defects (P atoms in ^{28}Si, NV centers in diamond, and dimers in SiC), and electron spins or quantum dots. However, working with quantum information is much more complicated and fragile than CMOS transistors and needs further research and optimization.

IBM has demonstrated a 27-qubit system (Falcon) in 2019, a 65 qubit system (Hummingbird) in 2020, and they plan to demonstrate a 127 qubit system (Eagle) by the end of 2021. IBM's quantum processors use Josephson junction-based superconducting qubits, which act as a nonlinear inductor and a two-level quantum system. The Josephson junction uses two superconductors with an insulator in between, which has the size of 100 nm x 100 nm with coherence time in the order of 100 µs, and a gate time of 10-500 ns. In addition, superconducting microwave resonators are also needed at the bottom to read out the qubit state and act as multi-qubit quantum bus and noise filters. Based on these achievements, IBM has projected a roadmap for scaling IBM quantum technology which promises up to 1121 qubit systems (Condor) by the end of 2023, see Fig. 4. Establishing such a roadmap provides an ecosystem for developing this new technology and helps identify the challenges.

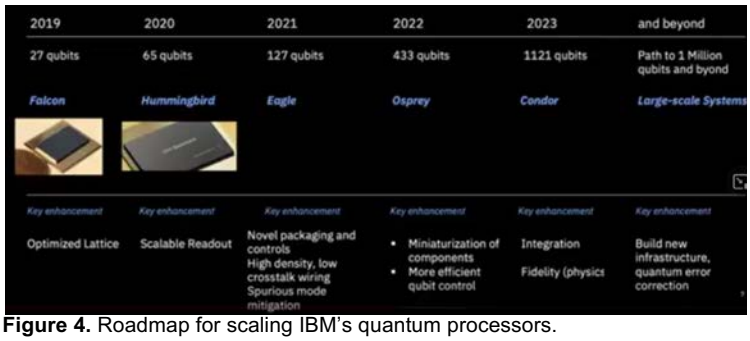

Figure 4. Roadmap for scaling IBM's quantum processors.

There are various challenges and opportunities in the development of quantum hardware. For example, the electronics in the refrigerator consist of cryo-CMOS circuitry, cryo-flex lines, amplifiers, attenuators, isolators, packaging technology, and qubit hardware, and there is a scope for potential innovations at each of these components. Quantum technology is at a very early stage, and there is much opportunity to explore new materials, devices, junctions, process controls, and

architectures to increase the coherence and fidelity and reduce the crosstalk. Moreover, we need efficient ways to measure and control the qubits, preferably by identifying new phenomena in nanoscale composition and structures.

3.2 X = Spin Transfer Torque MRAM

Daniel Worledge (IBM)

Spin transfer torque (STT) magnetoresistive random access memory (MRAM) is an emerging nonvolatile memory technology that was enabled by the invention of the magnetic tunnel junction (MTJ) device by John Slonczewski (IBM) in 1974. Slonczewski himself discovered the STT driven magnetization switching phenomena in MTJ in 1996, which enabled efficient write in MRAM with a much lower current than previously possible with a field-driven mechanism. Initial MTJ devices used amorphous aluminum oxide tunnel barriers, which led to low magnetoresistance, making them difficult to readout. In 2004, IBM and AIST demonstrated much higher magnetoresistance and solved the readout issue by replacing the oxide material with crystalline magnesium oxide, inspired by the theoretical prediction by Butler in 2000. Finally, the demonstration of the first perpendicular MTJ by IBM and Tohuku further lowered the switching current and enabled a new pathway for scaling. MTJ devices are now scaled down to < 10 nm.

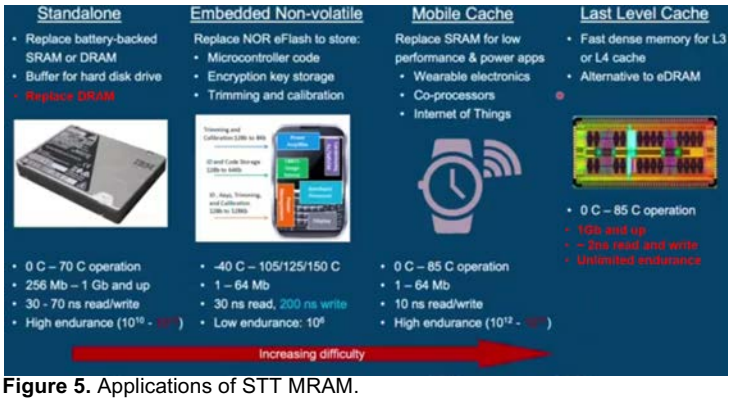

Figure 5. Applications of STT MRAM.

These four significant advances in developing MTJ devices lead to efficient nonvolatile STT MRAM technologies compatible with CMOS, see Fig. 5. The first general application category is standalone memory to replace battery-backed volatile SRAM and DRAM. MRAM for standalone application has high bit density (256Mb to 1Gb), fast read/write time of 30 to 70 ns, and very high endurance ~ 10^{10}. Currently, IBM's flash core module is using a 1 Gb MRAM chip as buffer memory. The second application category is embedded nonvolatile memory, which targets to replace embedded Flash that is not scalable below the 28 nm node. Major foundries, including Samsung, GlobalFoundries, and TSMC, have developed embedded MRAM to replace embedded Flash at the 28 nm node and below. This category of MRAM is usually operated over a wide temperature range and has a lower density (1~64 Mb), slower write time of 200 ns, and lower endurance of ~10^6. The third promising application is in the mobile cache to replace SRAM and embedded Flash in low-power applications (wearable electronics, internet-of-things, and coprocessors inside mobile devices). MRAM is very attractive for this category of application but has yet to enter the market. This category of MRAM does not need to be high performance, but its low power nature is useful at the edge devices where artificial intelligence can be implemented. The final category is to use MRAM as the last level of

cache (L3 and L4), which is an ongoing effort, and the critical requirement is to lower the switching current substantially, which is a material and device-intensive work. Ongoing works on MRAM have goals to achieve new perpendicular materials to enhance the magnetoresistance above 400% with a RA product <10 Ω-μm^2, new mechanisms to lower the switching current below 10 μA and to achieve on pitch etching so that we can have a 20 nm junction with 20 nm spacings. We need significant developments in new materials, devices, processes and tools, circuit design, and characterization tools to achieve these aggressive goals.

3.3 Two Dimension Materials for Silicon Integration

Max C. Lemme (RWTH Aachen University)

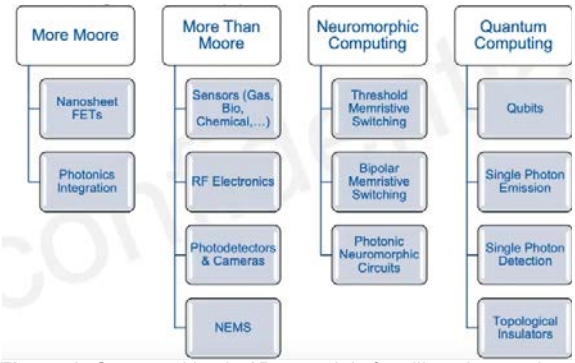

Figure 6. Opportunities in 2D materials for silicon integration.

Two-dimensional (2D) materials provide exciting opportunities for integration with silicon, see Fig. 6. 2D materials are promising in areas where miniaturizing the devices is desired, often called "more Moore". There is a growing interest in transforming from 3D FinFETs to 2D nanosheet FETs as they provide the ultimate electrostatic control [10]. There is no loss of mobility in 2D nanosheet FETs where the mobility in silicon significantly reduces as the dimensions shrink. Moreover, 2D materials can be integrated at the back-end-of-the-line (BEOL), promising to enable 3D integrated circuits. 2D materials also function as active optoelectronic material and can provide high-speed photodetection and modulation. Thus, there is an opportunity for integration with silicon or silicon nitride photonics [11] and solving many of the data bottlenecks that we have today.

There is a "more than Moore" domain where added functionalities are desired in the integration, which is difficult in silicon CMOS as they require different processes, take up silicon footprint, and are costly. In this domain, 2D materials can offer the added functionality when integrated at the BEOL of a CMOS chip. For example, metal-insulator-graphene is attractive for analog RF flexible electronics with >167 GHz cutoff frequency, WiFi six-port receivers at 2.1-2.7 GHz, and IR photodetection with 1000x responsitivity compared to existing silicon technologies [12-14]. Graphene-Si diode can outperform existing commercial silicon-based counterparts by a factor of three in photodetection application [15]. Graphene accelerometers can offer 1000x mass reduction and 100x area reduction compared to the state-of-the-art [16]. Highly sensitive pressure sensors can be built using $PtSe_2$ to reduce 100x area compared to the state-of-the-art [17].

2D materials can serve as both nonvolatile memristor [18] and threshold memristor [19] and play a crucial role in enabling neuromorphic computing and compute-in-memory hardware. Resistive switching can come from various mechanisms in 2D materials, including phase change, ion

transport, soft breakdown, and filament-forming, and there is an enormous opportunity to explore these mechanisms within new materials. 2D materials can also host qubits and promising for hardware in quantum computing. There have been significant advancements in spin-valley qubits in bilayer graphene [20], WSe_2 based single-photon emitters [21], black phosphorous avalanche diode for single-photon detection [22], and topological insulators in Moiré heterostructure [23]. There are many engineering challenges in materials, growth, transfer process, etching, encapsulation, and electrical contacts, which we need to overcome to integrate 2D materials into BEOL CMOS processes and translate into real products.

3.4 Ferroelectric Materials and Devices

Thomas Mikolajick (NaMLab, TU Dresden)

Historically, ferroelectric materials were characterized by having a noncentrosymmetric crystal where certain ions can be switched between two stable positions. The switching is nonvolatile and purely electric-field driven. Therefore, it is very interesting for memory application, and ferroelectric memories are on the market since 1993. However, classical ferroelectric materials are not CMOS compatible due to the complex crystal structure and weakly bound oxygens, which are inferior to various CMOS processes. Hafnium oxide (HfO_2) is very promising to address this issue as they are compatible with CMOS processes. In the early days, it was believed that HfO_2 could not show ferroelectricity due to the instability arising from the orthorhombic phase. HfO_2 exhibits stable ferroelectricity [24] and similar polarization as the classical materials despite the early doubts. The coercive field of HfO_2 is high, and the k-value is low, which is helpful for transistor application but reduces the endurance in memory device applications. The integrated HfO_2 based ferroelectric devices also have enormous opportunities in non-memory applications: including analog circuits, neuromorphic computing, and supercapacitors.

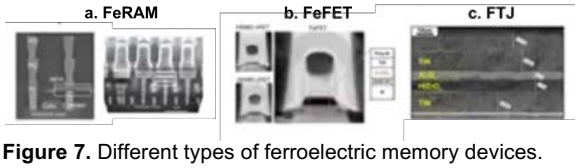

Figure 7. Different types of ferroelectric memory devices.

There are three types of memory devices using ferroelectric materials (see Fig. 7): ferroelectric random-access memory (FeRAM), ferroelectric FET (FeFET), and ferroelectric tunnel junction (FTJ). FeRAM measures the charge transfer at switching to read out the memory state [25] and is conceptually somewhat similar to DRAM. FeFET measures the transistor channel properties induced by the ferroelectric state for readout [26], which is conceptually similar to Flash memory but easier to integrate and operates at lower voltages. FTJ consists of two electrodes sandwiching a thin ferroelectric material [27]. Much research is needed to translate materials into these devices and integrate these devices into CMOS. Research on ferroelectric materials has two pillars: materials and integrated devices. Materials research is needed to understand the underlying physics that stabilizes the ferroelectricity in a particular structural phase of a material. Moreover, the role of dopants, oxygen, and stress on stability needs to be studied in the context of materials reliability, which will allow us to implement scaled test structures of new device demos. For integrated devices, NaMLab designs the devices and memory cell circuits and partners with companies like

FMC for chip design. For fabrication, NaMLab has partnered with production fabs in the industries and research organizations like Leti, IMEC, and Fraunhofer institute. The characterization is done at NaMLab or over the partner sites. NaMLab has demonstrated FeFET integration (see Fig. 7a) by partnering with GlobalFoundries, and an improve FeRAM integration which will be presented in IEDM 2021.

3.5 Energy Efficient Computing by Redox-Based Memristive Oxide Elements

Rainer Waser (RWTH Aachen University)

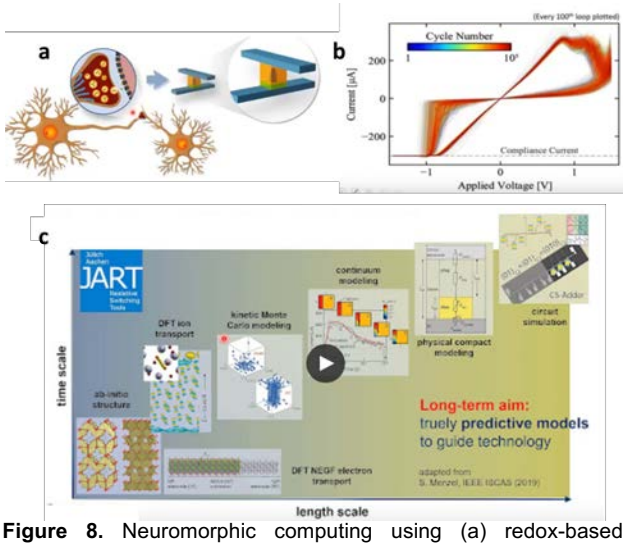

Figure 8. Neuromorphic computing using (a) redox-based memristive oxide element. (b) Current-voltage characteristics of the artificial synapse measured with a novel technique. (c) A comprehensive modeling and simulation framework from materials to system, which is available to the international community.

The alarming increase in energy consumption by CMOS-based information technology is of great concern in the present world, mainly because most of the electricity is still generated from fossil fuels. There should be a solution to this problem since the brain, also made from physical matter, is orders of magnitude energy-efficient than modern CMOS technology. Brain contains numerous neurons connected by synapses, using which brain can make decisions depending on data. To emulate the functionality of the brain in an integrated circuit, memristor devices can be used as artificial synapses (see Fig. 8a) and enable brain-inspired neuromorphic computing.

Various types of materials exhibits memristive switching via various mechanisms, which includes phase-change, valence change, Mott transition, thermochemical, electrochemical, electrostatic, and ferroelectric. These mechanisms show different characteristics which can be adapted for application-specific neuromorphic computing. Classical redox-based materials exploit the valence change mechanism (VCM), which can be engineered into a nanoscale structure where the thin oxide is sandwiched between an ohmic electrode (Ti, Ta, etc.) and an electronically active electrode (e.g., Pt, TiN, etc.). The resistance across the oxide barrier can be modulated with a voltage by tuning the mobile oxygen vacancies. An excellent memristive element must solve the so-called voltage-time dilemma, i.e.; it must show ultra non-linear switching kinetics with high-speed switching of nonvolatile memory and memory retention time up to 10 years [29]. Typical figure-of-merit for nonlinearity is given in terms of read/write times and voltages as

$$\mathrm{NL} = \frac{\log\left(t_{read}/t_{write}\right)}{\log\left(V_{write}/V_{read}\right)},$$

and the required value of NL for efficient practical application is >15. TaO_x based VCM cell already exhibits switching kinetics measured over > 14 decades on the same device [29].

The integration of energy-efficient neuromorphic circuits using these materials requires the development of fabrication technologies of these elements, dedicated measurement techniques, an understanding of microscopic physics of the processes, and proper modeling and simulation tools. For example, ultrafast current compliance-based measuring techniques have been developed in FZ-Jülich to measure over 10^6 current-voltage loops [28], see Fig. 8b, where a commercial source meter cannot measure even a single loop without damaging the device. On the simulation and modeling side, a multi-scale modeling framework has been developed in FZ-Jülich (see Fig. 8c) that starts with ab initio calculations at the material level to understand the electronic structure and density functional theory to understand electronic and ionic transports within the material. Furthermore, a powerful kinetic Monte Carlo and finite element methods are be used to generate compact physical models, allowing us to do the circuit and system-level simulations. The modeling package JART is available to the international community.

Last year, FZ-Jülich has set up a project called Neurotec II with many neighboring institutes, which covers all basic technologies for neuromorphic computing and memristor elements. The project's goal is to start from basic materials physics and proceed to demonstrate test chips. Some targeted applications of this project are low-power neuromorphic hardware with online learning to make artificial intelligence ubiquitous, high-performance scientific computing like neuroscience simulations, optimizations, and signal processing, and beyond deep learning, including brain-like functionality, memory augmented networks, and probabilistic/Bayesian networks.

3.6 X = Magneto-Electric Spin-Orbit (MESO) Logic

Dmitri Nikonov (Intel)

The demand for computing is growing exponentially, mainly due to the workloads, the data centers, and specifically artificial intelligence. There is also an exponential growth in the energy associated with the growth in computing such that within a decade, this required energy will approach the few percent of the total energy production in the world. Such an energy crisis has been identified as one of the seismic shifts by the decadal plan for the semiconductors from the Semiconductor Research Corporation [2], and efficient devices will be required to continue the sustainable development and lower the carbon emissions.

There is an ongoing quest to lower the energy for switching beyond CMOS devices, and one of the promising candidates is the magneto-electric spin-orbit (MESO) logic [30] (see Fig. 9a). MESO uses a magneto-electric (ME) effect for magnetization switching, typically observed in multiferroic materials like BiFeO3 at room temperature, which is much more energy-efficient than

existing mechanisms. The switching is accomplished by charging a capacitor; thus, the switching energy is determined by the device capacitance (C) and applied voltage (V) as $\sim CV^2$. Therefore, the main direction for lowering the energy is lowering the switching voltage.

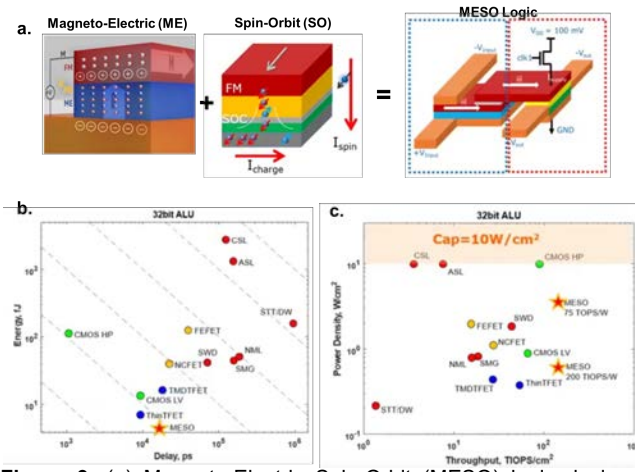

Figure 9. (a) Magneto-Electric Spin-Orbit (MESO) logic device. Comparison of (b) energy vs. delay and (c) density vs. throughput with various technologies.

MESO utilizes materials with spin-orbit (SO) coupling, and more specifically, the inverse Rashba-Edelstein effect (IREE), which converts the magnetization information into an electrical current in the SO element by applying a current into it through the magnet. Thus, the direction of the magnetization determines the direction of the current.

These write and read elements can be cascaded to get the MESO device, where the ME element acts as input, and the SO element acts as output. Such an element is expected to switch at a substantially low voltage ~0.1V, lowering the energy dissipated in the interconnects. The downside of this device is that it does not have a shut-off mechanism intrinsic to this device; therefore, one needs to regulate the operation of this device using CMOS, i.e., integration with CMOS is required for clocking or gating of the MESO devices. MESO is expected to be >10 times more energy-efficient than CMOS when benchmarked against a 32-bit arithmetic logic unit, see Fig. 9b. However, this energy efficiency comes with a trade-off from slower speed, about ten times slower than high-performance CMOS. Note that such trade-off was also observed between CMOS transistors and bipolar transistors, and industry moved towards the CMOS due to energy efficiency. CMOS is limited by the dissipated power density in this new comparison, while MESO is not limited by power. Thus, MESO-based architectures can achieve better computing throughput (see Fig. 9c) in energy-constrained applications.

Panel 3: CMOS+X and Novel Functionality

4.1 X-tending CMOS Technology to Sustain Kurzweil's Law

Tsu-Jae King Liu (University of California, Berkeley)

The semiconductor industry has relied on innovations in transistor materials and structures in the front end of line process and wiring innovations at the back end of the line process to increase the density of transistors on a chip with each new generation. Today, Intel's most advanced technology uses air gaps between wires to reduce coupling capacity. This back end of the line process with two metals separated by an air gap can be adapted to form nanometer-scale electromechanical switches (NEMS) over CMOS circuitry and enable three-dimensional integration.

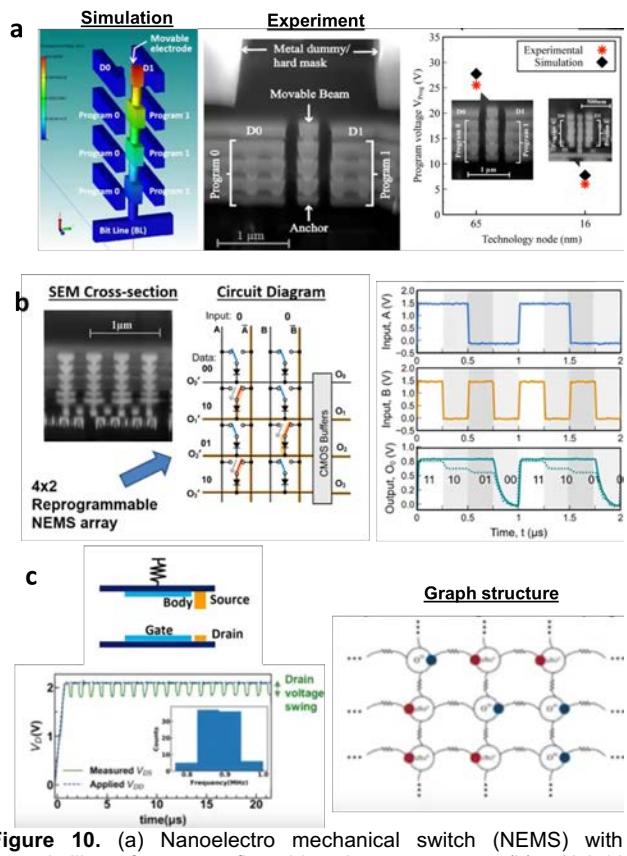

Figure 10. (a) Nanoelectro mechanical switch (NEMS) with nonvolatility for reconfigurable interconnects. (b) Hybrid NEMS+CMOS circuit for energy-efficient edge computing. (c) Four terminal microelectron mechanical relay as self-oscillator, a key component for implementing Ising networks.

These NEMS can be designed to achieve reconfigurable interconnects by enabling nonvolatility. Fig. 10a shows a simulated NEMS where the movable electrode comprises multiple layers of metals interconnected by vias, forming a compliant structure for lower program voltage. A voltage pulse is applied to the program 1 electrode on the right to induce an electrostatic force to actuate the beam into contact with the data-1 (D1) electrode on the right. The movable electrode will remain in contact with the data electrode when the program voltage is removed if the contact adhesive force is greater than the spring restoring force of the beam. Similarly, to actuate the beam into contact with the data-0 (D0) electrode on the left side, a voltage pulse is applied to the program 0 electrode on the left side. Note that the vertically oriented beam provides for a compact footprint conducive to implementation in a crosspoint array. The actuation or program electrodes and conducting electrodes run parallel across the array and the bit lines run along the orthogonal direction. Also, it should be noted that a non-linear device such as a diode can be incorporated in a series with each beam to prevent undesirable leakage current in a crosspoint array.

These NEM switches and associated hybrid CMOS circuitry has been designed at UC Berkeley using the process design kit for the 65 nm CMOS process from Texas Instruments (see Fig. 10a) and the 16 nm CMOS process from TSMC. Since the electrostatic force increases with decreasing electrode spacing, tighter metal pitch in the 16 nm CMOS process should provide a lower program voltage. The measurements on the 65 nm node and 16 nm node switches (see Fig. 10a) show that the program voltage scales down as expected. The program voltage is projected to be less than 3 volts at the 7 nm node, which is compatible with CMOS transistors.

These NEMS switches can be used to design hybrid CMOS+NEMS-based energy-efficient hardware for edge computing leveraging the back end of line processes. Fig. 10b is showing such a circuit implemented using TSMC 16 nm to process. Each switch in the 2x4 array of switches is programmed such that each output line will be driven to low only if the input string matches the stored data for that line. Fig. 10b is also showing the successful operation of this 2x4 line decoder. Note that the diodes were used to ensure that the current flows in only one direction through the beam, resulting in degraded bit line high voltage.

Our group at UC Berkeley has also developed a low thermal budget process using polycrystalline silicon germanium as a structural material to enable the fabrication of micro-electro-mechanical systems (MEMS) such as relays which can be used as volatile switches for logic applications. This contrasts with the reconfigurable interconnect discussed before. A simple three-terminal MEMS switch relies on the electrostatic force between the gate electrode and the source electrode to turn on, which is not optimal since the source voltage is not necessarily fixed. A four-terminal MEMS relay employs a body electrode as the reference electrode (see Fig. 10c) so that the state of the switch is determined primarily by the gate to body voltage difference, which is independent of the source voltage. These four-terminal relays can be made to oscillate between on and off states with only DC voltages applied [33]. A voltage slightly less than the turn-on voltage is applied to the gate, followed by a voltage at the drain through a resistive load. The extra electrostatic force induced between the drain and the source electrodes causes the relay to turn on and simultaneously the current flows between the drain and the source to discharge the drain voltage, resulting in a reduction in the electrostatic force between the source and the drain. Eventually the relay turns off and the drain will charge back up, inducing electrostatic force once again to turn on the relay and discharge drain voltage. Thus, the drain oscillates as the relay turns on and off repeatedly.

The oscillation can be locked into phase with a small oscillatory perturbation applied at the body electrode. The perturbation signal has a frequency roughly double that of the relay oscillation frequency. This harmonic injection-locked signal could have one of two possible phases differing by 180 degrees and the two oscillatory phases can be used to encode binary information. These non-linear oscillators can be used to implement an Ising machine to solve combinatorial optimization problems efficiently. Such problems often have a graph structure (see Fig. 10c) that can be mapped to a network of coupled oscillators. The coupling strengths correspond to the graph edge values, depending on the values of the coupling elements in order to minimize energy consumption and thereby reach the optimal combinatorial configuration.

4.2 Facilitating Devices-to-Systems Research at Scale via CMOS+X

Naresh Shanbhag (UIUC)

CMOS+X technology will be a critical component in facilitating devices to systems research at scale, i.e., across the nation. When we try to optimize across the computer stack, we have unmatched functionalities and efficiencies across different semiconductor systems. Once we know how to build these systems, it becomes challenging for competitors to copy them. This is where the devices to systems research are essential.

I directed the SONIC center between 2013 and 2017, which individually worked on CMOS and X technologies. The SONIC center was a vertically integrated research program with extensive expertise: information theorists, neuroscientists, and communication theorists at one end and device researchers at the other extreme. The goal there was to build systems and system theories that were cognizant of the stochasticity of the underlying device fabrics, and this thread from systems to devices was created in the center. To encourage exciting collaborations and ensure that they end up with real-life prototypes, we had a CMOS fab initiative in SONIC, where a separate bucket of funds was set up for faculty members to bid for it. So, every year we called for proposals and faculty research that connected devices on the systems could put together a short one-pager. As a result of this initiative, we observed fascinating CMOS prototypes from year three and onwards. One significant outcome of this initiative was to launch the area of in-memory computing, which is of great current interest to the community. We were looking at this technology in 2014 well before the rest of the world thought about it. Faculty members also explored various types of non-CMOS technologies, striving to build complex systems as they could. However, only systems with limited functionalities were demonstrations as it is challenging to scale up those results.

2017 onwards, there was the electronics resurgence initiative (ERI), joint university microelectronics program (JUMP), and Energy-Efficient Computing: from Devices to Architectures (E2CDA) focusing on integrated CMOS+X. ERI is also a vertically integrated initiative, and most of the program goes from systems to fabrication and testing. SONIC had a strong influence in the formulating of this program. We are now members of the Foundations of Novel Compute (FRANC) program within ERI, and in this program, we are working on MRAM-based deep in-memory architectures for RF communication which goes all the way from foundries to systems. This program focuses explicitly on CMOS+X, where X is now MRAM, and the CMOS part is a 22 nm FDX technology from GlobalFoundries. MRAM is at the core of the processor, where the data is being stored and computed. MRAM is highly dense and nonvolatile, but on its own, it is insufficient to demonstrate system scalability. CMOS circuitry is at the peripherals to drive the array and read the signals. MRAM is also highly stochastic, and we need to compensate for the stochasticity using error compensation methods, which we call Shannon-inspired computing models. We have already demonstrated MRAM-DIMA ICs and shown that these digital methods effectively compensate for the signal-to-noise ratio.

Now, how do we facilitate the devices to systems research at scale? There are two initiatives that I would like to recommend. The first one is purely CMOS, especially a chip design and test center.

The model for this center is adopted from the chip implementation center of Taiwan that had a tremendous impact on government federally funded centers and the semiconductor industry and promoting CMOS. That center acts as an interface between CAD tool companies and foundries. The other center is a CMOS+X fabrication and test center where the facilities should be available to make individual devices and a collection of devices to demonstrate meaningful functionalities. Also, this center will focus on the integration of advanced CMOS with the beyond CMOS devices, which will add a lot of value similar to the MRAM+DIMA project. Finally, there is an opportunity to create a shared space with industry, where industry researchers can work with academic researchers.

4.3 CMOS+X, X = Memory Integration in 3D Integrated Circuits

H.-S. Philip Wong (Stanford)

In today's computing systems, there is a bandwidth deficit, as illustrated in Fig. 11a. The blue data points in Fig. 11a are the throughput of GPU, and the orange data points are the bandwidth between the main memory and the computing unit. The throughput is growing at about 1.81 times every two years. On the other hand, the bandwidth is growing at a much slower rate, at 1.56 times every two years. Thus, today, there is a large bandwidth deficit, making it a bottleneck for computing systems, both in terms of throughput and energy consumption.

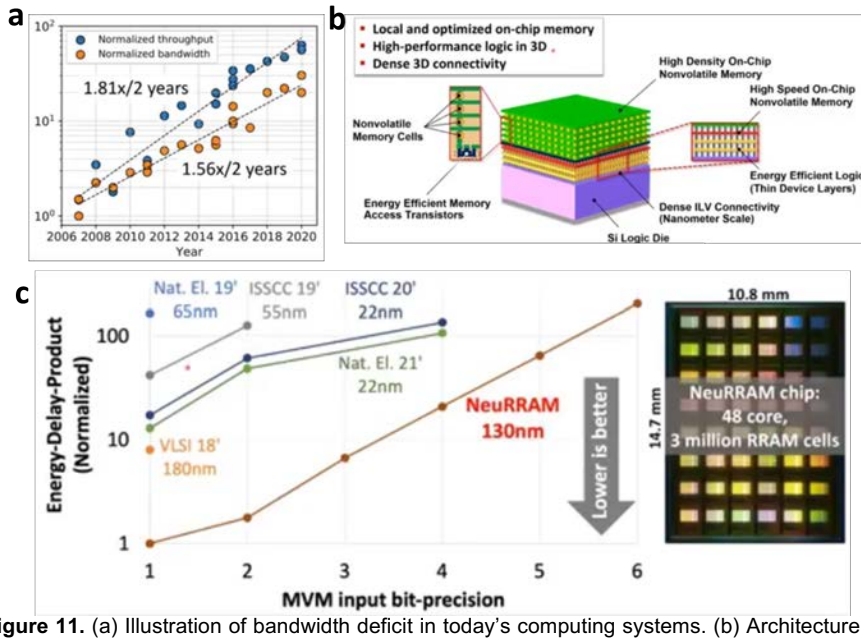

Figure 11. (a) Illustration of bandwidth deficit in today's computing systems. (b) Architecture of a N3XT nano system. (c) NeuRRAM neuromorphic computation system based on the N3XT nanosystem, which is energy-efficient compared to existing results in the literature.

The future is in systems integration logic with a memory and logic fusion. The N3XT nanosystem (see Fig. 11b) offers such a fusion in three-dimension (3D). N3XT system would have local and optimized memory, high-performance logic in 3D, and dense 3D connectivity. It will be built upon a silicon logic-based layer plus a high-performance logic layer in the upper layer of the 3D integrated systems. These transistors perform energy-efficient logic that access both the high-speed memory and high-density non-volatile memory. A broad system

design space is available within which memory and logic can be partitioned in a chip. The partitioning of memory and logic on the system occurs not only at the chip level but also at the package level. Understanding how to partition the system is itself an important research topic.

An N3XT system has been implemented using a 256x256 RRAM array integrated with 130 nm CMOS [34] which was used to demonstrate a restricted Boltzmann machine for image recovery. Starting from this chip, a much bigger chip, NeuRRAM, has been designed consisting of 48 core and 3 million RRAM cells. NeuRRAM has an energy-delay product that is substantially better than any results in the literature (paper in submission, see Fig. 11c). This energy efficiency is achieved by innovative circuit design and integrating the memory right on the CMOS chip.

Another example of an N3XT system is the Stanford associate RRAM chip for life-long edge learning and search (SAPIENS chip) [35], built with TSMC 40 nm RRAM technology. It performs an approximate search using an associative of memory. SAPIENS chip is highly energy-efficient as the memory is integrated with CMOS circuitry, achieved very accurate one-shot learning compared to the software models, and offers very robust inferencing for millions of cycles.

Logic technology is equally important, and it has to be integrated into 3D. Several years ago, a four-layer 3D integrated chip was demonstrated [36] where more than two million carbon nanotube transistors have been integrated along with one million bit RRAM cells and silicon logic as the bottom layer, illustrating the N3XT nanosystem concept. Today, this concept has been translated into SkyWater technology foundry in a 200 mm wafer with a CMOS baseline.

The semiconductor industry has been moving along for the past 50 years in only one direction ahead, scaling down the devices and miniaturizing in two dimensions. Nevertheless, we had a good run, and we are not looking back because the future now has many more possibilities in CMOS+X.

4.4 The Future of Computing: Bits, Neurons, and Qubits

Takashi Ando (IBM)

The mission of IBM Research is to explore the future of computing in the context of bits, neurons, and qubits. Bits is where mathematics and information intersect and covers today's computers and High-Performance Computing. IBM is heavily investing in this area and working with many partners to develop the next generation technology. Neurons are where biology and information intersect, and this covers applications like Deep Learning or other neuromorphic computing. IBM recently launched its new center called AI hardware center to address this area. Finally, qubits are where physics and information intersect and discover quantum systems.

More than 15 years ago, IBM launched a 300 mm semiconductor line in the facility of SUNY Polytechnic Institute and had been using this forum to develop the next generation CMOS Technologies, working with many tool vendors and device manufacturing companies. AI hardware center was established in 2019 in the same facility to leverage the foundation built over the last 15 years. Fig. 12a shows the roadmap of IBM's semiconductor technology development. CMOS scaling used to be limited by the lithographic capability. The new challenge is to continue the density scaling by introducing new materials and device architecture. IBM no longer manufactured semiconductor chips and adopted an R&D-based model working with partners such the process is transferred to the partners once it is developed. The model is working very well, and recently IBM announced the first demonstration of a 2 nm chip. IBM also adopted this model for neurons and qubits. Although IBM's AI hardware center is primarily located in Albany, to leverage various expertise on new materials and architecture, the center has been expanded to Zurich, Tokyo, London, and Yorktown Heights and partnered with external entities. The center's goal is to create a unique ecosystem to co-develop across the stack: from materials to application.

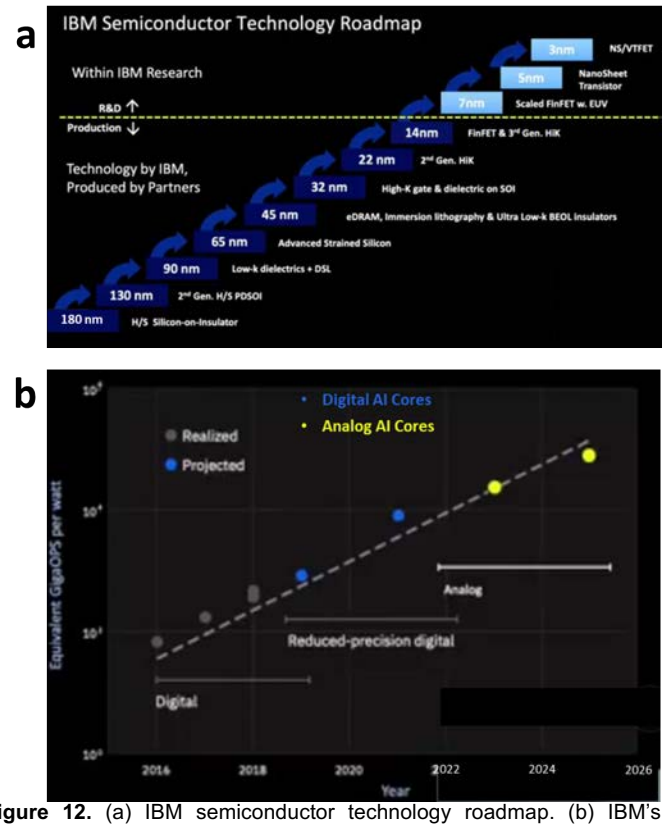

Figure 12. (a) IBM semiconductor technology roadmap. (b) IBM's projection on deep learning performance/watt.

The compute demand for AI is creating significant challenges as the compute requirements are doubling every 3.5 months. Significant innovation both in hardware and software in order to continue this trajectory. IBM's approach to push this trend up to 2022 is to use the digital AI cores by utilizing approximate computing, see Fig. 12b. Beyond this, a paradigm shift in the hardware is required to continue this trend, and analog AI cores can provide a solution up to 2025. The key idea of analog AI core is to map the deep learning weights to the analog crosspoint arrays such that the computation is performed at the location of the data. Phase-change materials are promising candidates to serve as nonvolatile analog storage. This approach will address the von Neumann bottleneck.

4.5 Microelectronics is on the move: from the viewpoint of a European

Robert Weigel (FAU Erlangen Nürnberg)

Micro-electronics is not only an enabler of technological products, but micro-electronics is also playing an essential role in the fight for pre-eminence in the world between China and the United States. Although Europe has put much money into supporting its semiconductor industries, Europe's role has been declining in the last decade, with only a 7% share of the worldwide business volume. European governments have done a lot in this area, but the Asian players and the United States have done much more. Europe has two chip manufacturers and currently plays no role in the CMOS-dominated internet, computer, notebook, and smartphone industries. There is no big design house, no role in chip-packaging and testing.

However, Europe holds strong positions in power, electronics, analog chips, optoelectronics, sensors and MEMS, embedded systems, and crypto and security chips. Overall, Europe still has a good ecosystem for microelectronics. A strong European user industry is expected to be seen in the automotive area when edge computing gains importance in the near future. That is part of the reason why Intel and TSMC are interested in building new advanced CMOS fabs in Europe. In parallel, the European Commission is about to invest in establishing a leading-edge fab in Europe. So Europe might catch up a bit in CMOS business and its ecosystem.

In the CMOS+X, X represents emerging technologies at which Europe is very strong. There are significant activities in Europe on CMOS+X research among various industries, research institutes, and academia. However, there is no strategic pooling of interest, a typical problem in Europe and many countries. The first step towards strategic pooling has recently been done in Germany with many research facilities such as Max Planck, Helmholtz, Leibniz, and Fraunhofer. 11 Fraunhofer institutes and 2 Leibniz institutes; those previously engaged with microelectronics research have been brought under the same roof called Forschungsfabrik Mikroelektronik Deutschland (FMD) to work on the CMOS+X.

One example of the advancement of CMOS+X from FMD is that they have combined a BiCMOS heterojunction-bipolar-transistor (HBT) with CMOS in a single die and showed that frequencies up to 1 THz are feasible [37]. Recently, silicon photonics has become a game-changer for photonic communication, and compared to CMOS devices, SiGe HBTs within an electronic-photonic integrated circuit (EPIC) offer advantages for f_T, f_{max}, and breakdown voltages. The bandwidth of Ge-photodiode has been pushed towards 110 GHz, enabling the use of SiGe technology for silicon photonics data communication. Monolithic EPIC platforms can combine optical modulators, broadband photodetectors, and optical waveguides with SiGe:C HBT devices. Heterogeneous integration of different technologies will further boost the functionality and performance, especially at higher frequencies. The combination of SiGe BiCMOS with InP HBTs is a promising pathway. Moreover, RF MEMS switches can also be realized in BiCMOS and BiCMOS integrated microfluidic packaging in wafer bonding.

Finally, the revitalization of Europe’s microelectronics industry will require a long-term strategic masterplan, improvements in the boundary conditions for the industries, significant sources of

fundings like in the U.S., South Korea, and China, settlement of leading-edge manufacturers, and support programs for all relevant parts of the value chain.

4.6 CMOS and the X-factor

Arijit Raychowdhury (Georgia Tech)

There is much ongoing work in integrating dense memory with CMOS, including spin-based nonvolatile memory, resistive RAMs, phase change memory, ferroelectric, and oxide memory, which are at various maturity levels. In terms of integration, monolithic front-end-of-the-line will produce the highest performance and density. However, there are interesting ongoing works on monolithic back-end-of-the-line integration where lower temperature processes are enabled, allowing different layers and technologies to be integrated but with a sacrifice in the density. In addition, a higher density of chiplets in the package integration is also a viable solution for memory and logic. High-capacity memory is required for applications like high-performance computing, artificial intelligence (AI), graphs, and optimization. Fig. 13a shows the energy efficiency vs. the power of different technologies for AI acceleration. The orange data points are examples of novel technologies integrated with CMOS, showing much higher performances than the CMOS counterparts.

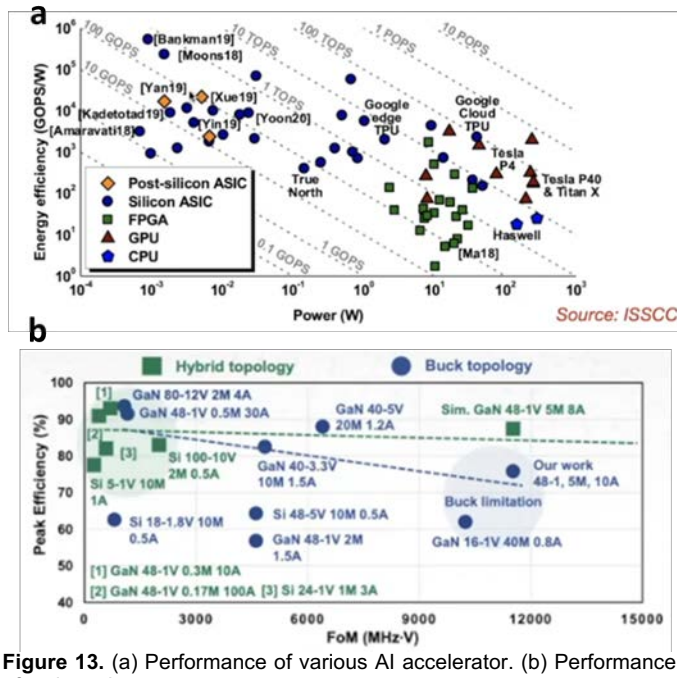

Figure 13. (a) Performance of various AI accelerator. (b) Performance of various down converters.

Apart from memory, integrating wide bandgap materials with CMOS can improve high-voltage power conversion and regulation technologies. The inputs are typically 48 V in servers, which is down-converted to 12 V to power the mainboard. Thus, the mainboard takes a lot of current, leading to I2R losses, and high voltage converters capable of converting 48V to 1V are gaining much attention in recent days. One of the booster technologies enabling such conversion is GaN devices, and GaN converters are showing very high figure-of-merits at high efficiencies, as shown in Fig. 13b. GaN devices are also promising in RF applications, and the GaN RF device market is expected to exceed $2 billion by 2025. Although the application started with military and defense, it is entering the commercial market with the advent of 5G.

There are also new physical computing paradigms in which the devices themselves are not necessarily the switches. One of the examples is the physical computing paradigm. In conventional Boolean logic, the transistor physics is essentially converted into the mathematical domain via logic gates. Logic gates act as the boundary between physics and mathematics, which allows us to use various algorithms in Turing machines. For this, the amount of resource requirement (e.g., time or energy) goes high when dealing with problems like optimization or weather prediction. However, the amount of resources can be reduced if we can increase the boundary between physics and mathematics by implementing the physics of the device as the nano functions. Moreover, a transition from the local interaction of nano functions to a global interaction can substantially reduce the amount of resources needed to solve a complex computational problem. Thus, integrating new device technologies with CMOS is a pathway to enable this new physical computing paradigm with significantly low resource requirements..

Acknowledgement

This report was written primarily by Dr. Shehrin Sayed , Dr. Micahel Bartl and Prof. Sayeef Salahuddin from UC Berkeley

List of the invited participants:

Invitee Name	Affiliation
Alan Seabough	Notre Dame
Alp Sipahigil	UC Berkeley
Anand Raghunathan	Purdue
Arun Ashok	FZ Juelich
Asif Khan	Georgia Tech
Atulsimha Jayasimha	VCU
Barbara De Salvo	Facebook
Birgit Schwenzer	NSF
Carlos Diaz	TSMC
Damian Dudek	DFG
Debdeep Jena	Cornell
Dongik Suh	SK Hynix
Erik Brunvand	NSF
Gary Miner	AMAT
George Janini	NSF
Giuseppe Iannaccone	University of Pisa
Huili Xing	Cornell
James Edgar	NSF
Jean Anne Incorvia	UT Austin
Joachim Knoch	RWTH Aachen
Joerg Appenzeller	Purdue
John Paul Strachan	FZ Juelich
Joydeep Brunvand	NSF
Kaushik Roy	Purdue
Luigi Colombo	UT Dallas
Margaret Martonosi	NSF
Matthias Passlack	TSMC
Michael Jank	Fraunhofer Institute
Michael Schiek	FZ Juelich

Mike Niemier	Notre Dame
Mircea Stan	Virginia
Muhammad Hussain	UC Berkeley
Nikhil Shukla	Virginia
Rainer Leupers	RWTH Aachen
Rance Cleaveland	NSF
Regina Dittmann	FZ Juelich
Robert Wallace	UT Dallas
Rosa Martonosi	NSF
Ryan Guo	NSF
Sharon Hu	Notre Dame
Shimeng Yu	Georgia Tech
Simka Hosono	Samsung
Stephan Menzel	FZ Juelich
Suman Datta	Notre Dame
Supriyo Bandyopadhyay	VCU
Supriyo Datta	Purdue
Susanne Hoffmann-Eifert	FZ Juelich
Thomas Ernst	Leti
Tobias Gemmeke	RWTH Aachen
Todd Youngkin	SRC
Tom Kuech	NSF
Tony Chiang	AMAT
Uwe Schroder	NamLab
Uygar Avci	Intel
Vikas Rana	FZ Juelich
Woobin Song	Samsung
Zhihong Chen	Purdue

References

[1] IMEC Technology Forum 2021.
[2] SRC, Decadal Report, 2021.
[3] S. K. Moore, “A better way to measure progress in semiconductors”, IEEE Spectrum, July 2020.
[4] A. Saeidi, F. Jazaeri, I. Stolichnov and A. M. Ionescu, "Double-Gate Negative-Capacitance MOSFET With PZT Gate-Stack on Ultra-Thin Body SOI: An Experimentally Calibrated Simulation Study of Device Performance," in IEEE Transactions on Electron Devices, vol. 63, no. 12, pp. 4678-4684, Dec. 2016.
[5] A. Rassekh, J. -M. Sallese, F. Jazaeri, M. Fathipour and A. M. Ionescu, "Negative Capacitance Double-Gate Junctionless FETs: A Charge-Based Modeling Investigation of Swing, Overdrive and Short Channel Effect," in *IEEE Journal of the Electron Devices Society*, vol. 8, pp. 939-947, 2020.
[6] X. Mei et al., "First Demonstration of Amplification at 1 THz Using 25-nm InP High Electron Mobility Transistor Process," in IEEE Electron Device Letters, vol. 36, no. 4, pp. 327-329, 2015.
[7] P. Rodriguez-Vazquez, J. Grzyb, B. Heinemann and U. R. Pfeiffer, "A QPSK 110-Gb/s Polarization-Diversity MIMO Wireless Link With a 220–255 GHz Tunable LO in a SiGe HBT Technology," in IEEE Transactions on Microwave Theory and Techniques, vol. 68, no. 9, pp. 3834-3851, 2020.
[8] P. Hillger et al., "A 128-pixel 0.56THz sensing array for real-time near-field imaging in 0.13µm SiGe BiCMOS," 2018 IEEE International Solid - State Circuits Conference - (ISSCC), 2018.
[9] R. Jain et al., "A 32×32 Pixel 0.46-to-0.75THz Light-Field Camera SoC in 0.13µ m CMOS," 2021 IEEE International Solid- State Circuits Conference (ISSCC), 2021.
[10] www.imec-int.com/en/articles/imec-introduces-2d-materials-logic-device-scaling-roadmap
[11] Daniel Schall, Daniel Neumaier, Muhammad Mohsin, Bartos Chmielak, Jens Bolten, Caroline Porschatis, Andreas Prinzen, Christopher Matheisen, Wolfgang Kuebart, Bernhard Junginger, Wolfgang Templ, Anna Lena Giesecke, and Heinrich Kurz, “50 GBit/s Photodetectors Based on Wafer-Scale Graphene for Integrated Silicon Photonic Communication Systems”, ACS Photonics 1 (9), 781-784, 2014.
[12] Wang et al. ACS Apl. El. Mat. 2019.
[13] Wang et al. Adv. El. Mat. 2021.
[14] De Fazio et al., ACS Nano, 2020.
[15] Riazimehr et al., ACS Photonics, 2019.
[16] Fan et al., Nat Electron. 2019.
[17] Lukas et al., Adv. Func. Mat. 2021.
[18] Belete et al., Adv. El. Mat. 2020.
[19] Völkel et al., IEEE SNW, 2021.
[20] Banszerus et al. Nano. Lett. 2020.
[21] He et al., Nat. Nano., 2015.
[22] Atalla and Koester, DRC, 2017.
[23] Kennes et al., Nat. Phys. 2021.
[24] T. Mikolajick and U. Schröder, Nat. Mat. 2021.
[25] T. Mikolajick, Rev. Mod. in Mat. Sc. and Mat. Eng. Elsevier, 2016; J. Okuno et al. VLSI 2020.
[26] H. Mulasomanovic et al., Nanotechnology, 2021.
[27] E. Y. Tsymbal and H. Kohlstedt, Science, 2006; B Max et al., JEDS, 2019.
[28] T. Hennen et al. Rev. Sci. Instrum. 92 (2021).
[29] S. Menzel et al., Adv. Func. Mater. 25, 2015.
[30] S. Manipatruni et al., Nature 565, 35-42, 2019.
[31] K. Kato et al., IEEE Electron Device Letters, 37(12), 1563-1565, 2016.
[32] U. Sikder et al., IEEE IEDM 2020.
[33] X. Hu et al., IEEE IEDM 2019; IEEE IEDM 2021.
[34] W. Wan, R. Kubendran, S. B. Eryilmaz, W. Zhang, Y. Liao, D. Wu, S. Deiss, B Gao, P. Raina, S. Joshi, H. Wu, G. Cauwenberghs, and H.-S. P. Wong, "A 74 TMACS/W CMOS-RRAM Neurosynaptic Core with Dynamically Reconfigurable Dataflow and In-situ Transposable Weights for Probabilistic Graphical Models," International Solid-State Circuits Conference (ISSCC), 2020.

[35] H. Li et al., "One-Shot Learning with Memory-Augmented Neural Networks Using a 64-kbit, 118 GOPS/W RRAM-Based Non-Volatile Associative Memory," 2021 Symposium on VLSI Technology, pp. 1-2, 2021.
[36] Shulaker, M., Hills, G., Park, R. *et al.* Three-dimensional integration of nanotechnologies for computing and data storage on a single chip. *Nature* **547,** 74–78 (2017).
[37] D. Kissinger, G. Kahmen and R. Weigel, "Millimeter-Wave and Terahertz Transceivers in SiGe BiCMOS Technologies," in IEEE Transactions on Microwave Theory and Techniques, doi: 10.1109/TMTT.2021.3095235.