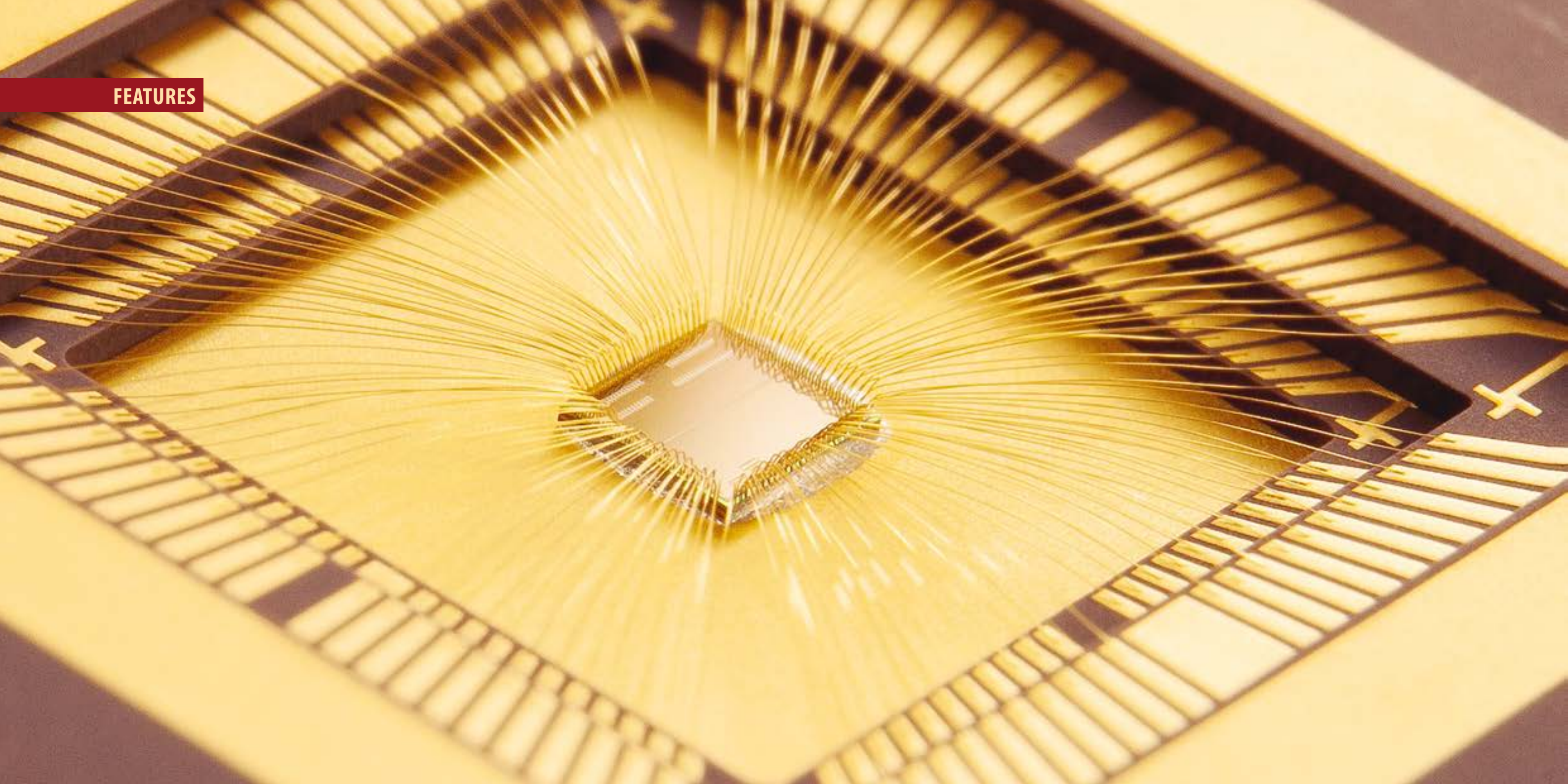


SCALING WITHOUT SCALING

A PARADIGM SHIFT IN THE SEMICONDUCTOR INDUSTRY

Cezar B. Zota, Lukas Czornomaz, Jean Fompeyrine and Stefan Abel – DOI: https://doi.org/10.1051/epn/2018504
IBM Research-Zurich, Säumerstrasse 4, 8803 Rüschlikon, Switzerland

Microelectronics is as pervasive as it is invisible. It is however impacting every moment and aspect of our daily lives and has radically transformed all industries. Yet, this adventure began only about 60 years ago, with the first integrated circuits by J. Kilby and R. Noyce.

To become pervasive, being small and power efficient turned into the DNA of microelectronics. Scaling transistors has been the ultimate goal for many years, and allowed engineers to make circuits more complex, working at lower power, and cheaper to produce. This unique alignment of planets did last for 50 years.

Today, scaling alone cannot feed the growing needs for mobile and power electronics, automotive and healthcare applications to name a few. As an example, mobile electronics perfectly illustrates the burning challenges ahead of us: Being power efficient for better mobility, integrating functions to improve our mobile capacities and enabling cognitive technology for smart services. To support this evolution, the semiconductor industry needs disruptive concepts, from materials to devices and systems.

A winding road ahead for efficient logic

After four decades of pure geometrical scaling of Si CMOS technology, the semiconductor industry entered the 21st century with a new phase of scaling powered by material-driven innovations. Besides strained-Si and silicon-germanium, high-k gate dielectrics and metal-gates were introduced, as well as low-k dielectrics in the back-end-of-the-line or non-planar channels requiring conformal control of materials at the nanoscale. High-aspect ratio field effect transistors (FinFET) are today's CMOS standard technology envisioned until the 7 nm node [1]. They consist of a tall (≈ 50 nm) and thin (≈7 nm) channel of Si with a gate wrapped around it Fig. 1).They provide superior electrostatic control over the channel and superior channel perimeter per footprint compared to a planar transistor.

Beyond FinFETs, innovations on materials will keep being key drivers to continue improving the power vs performance vs density trade-off of CMOS technology. Near-term, fin-channels will evolve into stacked nanosheet channels [2]: sheets of semiconductors that are few nanometers thick and few tens of nanometers wide, stacked on top of each other to provide augmented channel perimeter per footprint and superior electrostatic control. These nanosheets are formed by epitaxy of crystalline superlattices, *e.g.*, Si/SiGe, that are patterned to form the width of the sheets. One of the two constituents

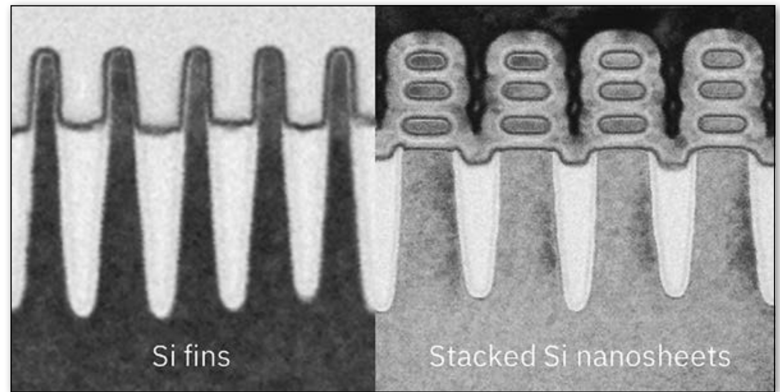
is selectively removed (*e.g.*, SiGe) to leave suspended stacked nanosheets of Si. The ability to control the deposition and patterning of gate dielectric, gate metal, sidewall spacers in the tiny horizontal gaps between the sheets is an essential area where materials and process innovation will ensure that this new technology becomes a reality.

Further innovation will follow two main parallel routes intrinsically linked to the transistor performance and energy efficiency metrics at a given footprint: the amount of current that can be driven, and the on-state to off-state current ratio that can be achieved, for a certain operating supply voltage. First, increasing the drive current of transistors implies boosting the transport properties of the channel material, *e.g.*, by replacing Si in the channel by other materials which are intrinsically better such as germanium or III-V compound semiconductors. They allow carriers to travel faster but require a whole new set of innovations to tackle their epitaxy on Si, the formation of high-quality reliable gate dielectrics, the realization of ultra-low resistance contacts, and the mitigation of off-state leakage owing to their lower bandgap than Si. Tremendous progress in the past ten years has led to the several demonstrations of CMOS circuits on Si with indium gallium arsenide (InGaAs) and silicon germanium as channel materials for n- and p- transistors [3].

Second, novel device concepts leveraging material innovations are explored to increase the steepness on the off-to-on transition in a given gate voltage range. III-V heterostructures have been used to demonstrate transistors relying on tunneling current to achieve a 10-fold increase of the current for a record-low gate voltage increase of only 48 mV, well beyond the 60 mV limit given by Boltzmann statistics in standard transistors [4]. An alternative approach is to introduce a new family of materials that possess a phase transition giving rise to an abrupt change of their conductivity (*e.g.*, with insulator-metal phase transition materials such as VO_2) or polarization (*e.g.*, with ferroelectric materials such as $Hf(Zr)O_2$). Such phase transitions can be exploited in novel device concepts to drastically increase the steepness of the off-to-on switching, by introducing these materials as part of the gate dielectric or source/drain contacts.

Merging functionalities in the third dimension

Future efficient scaling of hardware will demand not only device-level but also system-level innovations, referring to the organization of transistors or blocks of transistors on the semiconductor die. In the hyper-scaling era, 2021 and beyond, system-level improvements are predicted to be one of the main drivers for technology [5]. Such improvements can enable not only density scaling, but also new functionalities, so-called "more than Moore" approaches. In this era, two interrelated system-level technologies are envisioned: 3D sequential and heterogeneous integration.

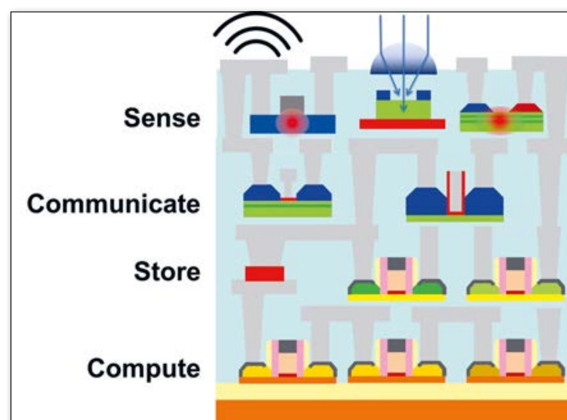


▲ **FIG. 1:** Transmission electron microscopy cross-sectional view along the gate (across the carrier transport direction) of FinFETs (left) and nanosheet-FETs. The fin channels are approximately 7 nm wide, while the 3-layer stacked nanosheets are approximately 5 nm thick.

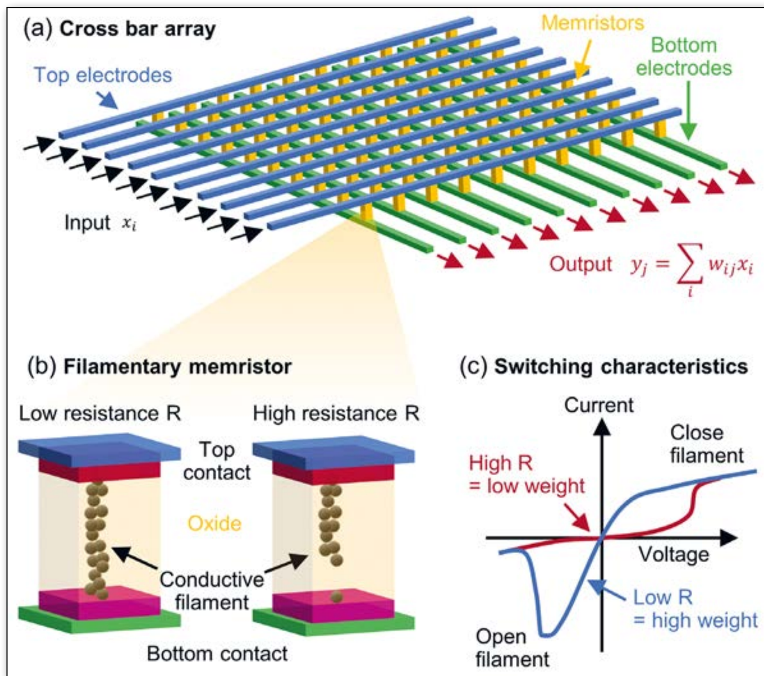
In 3D sequential integration multiple layers of transistors are fabricated *in situ* on top of each other (Fig. 2). This is different from 3D stacking packaging technologies utilizing through-silicon vias in that it offers transistor-level granularity of inter-level vias, resulting in two or three orders of magnitude higher via density, providing overall shorter wires, lower energy consumption and latency. 3D sequential integration will allow not only for subsequent stacking of Si CMOS layers, but also layers in different technologies – referred to as heterogeneous integration – such as III-V HEMTs, wide-band gap semiconductors and photonics, enabling high via densities between technologies which have hitherto required costly integration schemes in terms of both form-factor and energy consumption.

One of the key challenges facing these integration schemes is the combination of traditional Si CMOS fabrication processes with the additional thermal budget requirements of a subsequent layer. Due to the sensitivity of the gate oxide/channel interface in most MOSFET technologies, fabrication of layers beyond the first is typically limited to 450 °C, significantly reducing available process options. Moreover, when combining Si CMOS technology with non-standard materials, *e.g.* III-V compounds or photoactive materials such as $BaTiO_3$., process and material incompatibilities must be resolved.

Nevertheless, significant progress in system-level integration has been made in recent years, with approaches following two main routes. In the first route, subsequent active layers are grown *ex situ* and integrated on the 3D stack by direct wafer bonding, a transfer method whereby



◂ **FIG. 2:** Schematic figure showing how sequential 3D heterogeneous integration could merge different functionalities on the same chip. From the bottom up: Si CMOS, novel memories, III-V RF transistors and photoactive materials, all linked by electrical paths or vias (grey color).



▲FIG. 3: **(a)** Illustration of array of memristors that can be used to perform a power-efficient, analog vector matrix multiplication based on Ohm's and Kirchhoff's laws. **(b)** Illustration of a filamentary memristor. Oxygen vacancies form a conductive path that can partially be switched. **(c)** Schematics of the IV characteristics of a memristor with two different resistance values.

two oxide-coated surfaces are placed in physical contact, adhering after thermalization. The strength of this approach is the flexibility of material choices – due to the ex situ growth, there is no concern of crystal lattice mismatch between stacked layers – manifested in a host of experimental demonstrators, ranging from GaN power switches to III-V high-frequency transistors and $BaTiO_3$ optical modulators, all integrated on silicon substrates [6,7].

In the alternative route, subsequent active layers are grown *in situ* using various selective epitaxial schemes, typically entailing the patterning of a growth mask as well as a method of maintaining low crystal defect density, such as aspect ratio trapping or confined epitaxial overgrowth. This approach promises reduced fabrication costs compared to direct wafer bonding, since it minimizes the quantity of grown material and omits the need for a costly non-Si growth substrate. It has been explored for instance using vertical nanowires, or growing templates in cavities, both to integrate III-V's on silicon [8,9].

Neuromorphic cores for the data-driven economy

While device and system-level innovations will continue to optimize computing systems, they will still fundamentally support the von Neumann architecture, where the processing unit, the control unit, and the external memory are three separate entities. In spite of being excellent at addressing a vast amount of computational problems, the von Neumann architecture has limitations to handle workloads related to big data processing and to train and execute neural networks. For algorithms such as deep learning, the memory-processor bottleneck strongly limits the computational task, because the system is busy exchanging data between the memory and the processor during two distinct phases. First, the training phase of the network utilizes a large set of examples to iteratively tune the network. Second, the pre-trained networks are executed in the inference phase to analyze new, unknown input data.

Novel types of non-von Neumann computing architectures are therefore in the focus of intense research, leading to new microprocessors operating either in the digital or in the analog domain. Memory and processing units are close or co-located, similar to the synaptic and neuronal functions in brains. They are often referred to as "brain-inspired" or "neuromorphic" computing.

Dedicated digital neuromorphic chips, such as TrueNorth (IBM) and Loihi (Intel Corp.), are based on standard CMOS processes and devices. In TrueNorth, packets are digitally routed between one million neurons, which are small, programmable logic units [10]. The synaptic weights are stored in SRAM (static random-access memory) cells distributed across the chip. This co-location of memory and processing units allows low-power inference tasks *e.g.* on image classification. To allow learning in such distributed networks, other approaches (*e.g.* SpiNNaker) connect many processors in large clusters and apply local, brain-inspired learning rules such as spike-time dependent plasticity (STDP).

Analog or mixed-analog approach to neuromorphic microprocessors are also proposed. Bio-inspired processors emulate the biological functions of neural networks such as synapses and neurons with analog electrical circuits designed in CMOS technology [11]. While the spiking neurons are analog circuits, synapses and intra- and inter-chip communication is typically done in the digital domain. In this domain, Europe is at the forefront of research (*e.g.* www.neuram3.eu, to which the cover image is related).

A second type of analog neuromorphic circuits rely on memristive materials as an analog local memory, motivated by the increase in the local density of synaptic weights. A simple exploitation of Ohm's and Kirchhoff's laws allow to multiply and accumulate values in the analog electrical domain. Assembled as cross-bar arrays (Fig. 3a), networks of memristors can then be applied to strongly accelerate both the inference and the training phase of neural networks by orders of magnitude compared to von Neumann-based hardware [12][13]. Having such analog accelerators depend on the discovery of novel materials with dedicated non-volatile resistive switching properties, such as low drift, a large number of analog states, and low-power and voltage operation. Great progress has been made using different mechanism such as phase-change in GeSbTe, filamentary conduction in dielectrics (Fig. 3b) such as HfO2, and domain switching in STT-MRAM (Spin-transfer torque magnetoresistive random-access memory) or ferroelectrics. However, no satisfying solution has been found yet. Novel circuit concepts might have to be co-developed to compensate for the imperfection of such analog devices [14]. Nevertheless, since its first description in the late 1980's, neuromorphic computing has made great progress in the last few years, One

example of its potential use is the detection of correlations in weather forecast data as shown in Fig. 4 a&b.

Conclusions

Through these selected examples, we have shown that the end of geometrical scaling does not imply that microelectronics technologies are frozen and can only be incrementally optimized. The potential for innovation is remarkable, be it to improve the power efficiency of digital devices, to integrate functionalities in complex, yet compact systems, or to enable new type of computational tasks. At the core of these innovations, novel materials, devices and integration concepts will be the key elements that will make such a disruptive evolution happen. It will require different skills from scientists and engineers. First, a broadening of the technical horizon will be needed to include unusual materials and revolutionary devices as possible solutions. Moreover, there will be a clear demand for a holistic view on the technologies, leading to the co-development of materials, devices and system towards specialized chips. ■

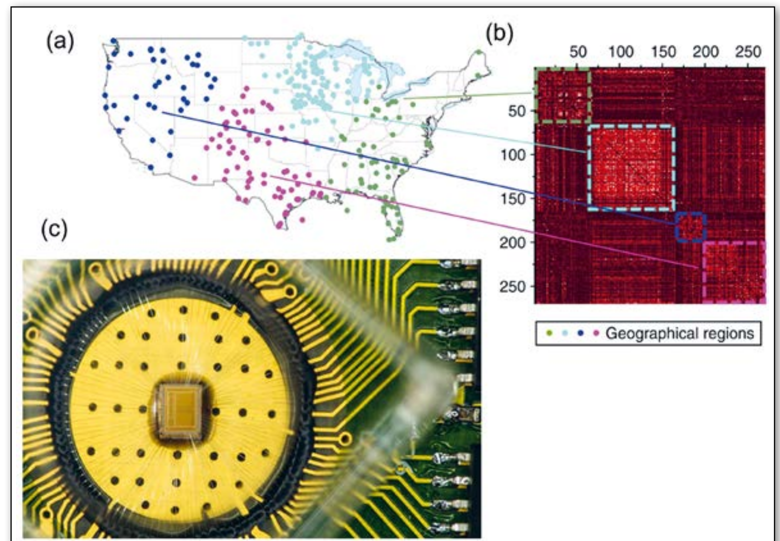
About the Authors

Cezar B. Zota received the Ph.D. degree in electrical engineering from Lund University, Sweden, in 2017. Presently, he is a researcher at the IBM Research Zurich laboratory, Switzerland, where his research interests include high-frequency electronics and quantum computers.

Lukas Czornomaz received his engineering degree in physics and material sciences from the National Institute of Applied Sciences in Toulouse and PhD from the University of Grenoble Alpes. Since joining IBM Research – Zurich in 2010, he has been focusing his research on the material and device integration of III-Vs on silicon for CMOS, high-speed RF and optical communication applications.

Jean Fompeyrine holds an engineering degree from the national school of engineer in Caen (France) and a PhD from the University of Bordeaux (France). His expertise relates to functional oxide thin films, used as materials for CMOS and integrated photonics. He has also focused on new methods for the monolithic heterogeneous integration of advanced materials. He is currently focusing on the development of dedicated hardware for neuromorphic computing, specifically novel analog non-volatile resistance.

Stefan Abel studied nanoscale engineering at the University of Würzburg (Germany) and holds a PhD from the University of Grenoble (France). He has gathered strong knowledge in materials science related to oxide materials, integrated photonics, and neuromorphic computing. His work currently focuses on the research and development of new memristors, including their application for accelerating deep-learning algorithms and implementing non-von Neumann reservoir computing schemes.



▲ **FIG. 4: (a)** Rainfall data from different weather stations across the USA, **(b)** Covariance matrix mapped in an array of phase change memory (PCM) cells used for detecting correlations in the weather data. **(c)** Photograph of a neuromorphic PCM chip for cognitive data analysis [15].

References

[1] R. Xie *et al.*, *Proc. of IEEE Int. Electron Device Meeting (IEDM)*, San Francisco, California, December 2016, pp. 47-50.

[2] N. Loubet *et al.*, *Symp. on VLSI Technology*, Digest of Technical Papers, Kyoto, Japan, *June* 2017, p. T17-5.

[3] L. Czornomaz, V. Djara, V. Deshpande, E. O'Connor, M. Sousa, D. Caimi, K. Cheng, and J. Fompeyrine, *Symp. on VLSI Technology*, Digest of Technical Papers, Honolulu, Hawaii, USA, June 2016.

[4] C. Convertino, C. B. Zota, H. Schmid, A. M. Ionescu and K. E. Moselund, *J. Phys.: Condens. Matter* 30, p. 264005 (2018).

[5] S. Salahuddin, K. Ni and S. Datta, *Nat. Electron.* **1**, 442 (2018).

[6] C. Zota, C. Convertino, V. Deshpande, T. Merkle, M. Sousa, D. Caimi and L. Czornomaz, *Symp. on VLSI Technology*, Digest of Technical Papers, Honolulu, Hawaii, USA, June 2018, p. T15-5.

[7] F. Eltes, M. Kroh, D. Caimi, C. Mai, Y. Popoff, G. Winzer, D. Petousi, S. Lischke, J. E. Ortmann, L. Czornomaz, L. Zimmermann, J. Fompeyrine, S. Abel, *Proc. of IEEE Int. Electron Device Meeting (IEDM)*, San Francisco, California, *December* 2017, p. 24.5.1.

[8] O.-P. Kilpi, J. Svensson, J. Wu, A. R. Persson, R. Wallenberg, . Lind, and L.-E. Wernersson, *Nano Lett.* **17(10)**, 6006 (2017).

[9] L. Czornomaz, E. Uccelli, M. Sousa, V. Deshpande, V. Djara, D. Caimi, M. D. Rossell, R. Erni, and J. Fompeyrine, *Symp. on VLSI Technology*, Digest of Technical Papers, Kyoto, Japan, June 2015, p. T172.

[10] P. Merolla *et al.*, *Science* **345(08)**, 668 (2014).

[11] G. Indiveri *et al.*, *Front. Neurosci.* **5**, 1 (2011).

[12] T. Gokmen and Y. Vlasov, *Front. Neurosci.* **10**, 333 (2016).

[13] S. Ambrogio *et al.*, *Nature* **558**, 60 (2018).

[14] M. Le Gallo *et al.*, *Nat. Electron.* **1**, 246 (2018).

[15] A. Sebastian *et al.*, *Nat. Commun.* **1706.00511**, 1115 (2017).