What's Next in Computing – In Bits, AI & Qubits?

H. Riel[1]
[1]*IBM Research Europe – Zurich, 8803 Rüschlikon, Switzerland*
hei@zurich.ibm.com

Advances in semiconductor science and technology have influenced and powered nearly every aspect of our life for more than 50 years and continue to push the limits of computing performance. The transistor has been demonstrated for the first time in 1947 and is still the basic building block of microprocessors like CPUs and GPUs – the core of our modern digital computers. Today, more than 100 billions of transistors with critical dimensions of nanometers are built in a chip fabricated on 300 mm substrates, all this enabled by heroic engineering and scaling efforts. Sustained innovation in materials, new device structures and architectures as well as in fabrication techniques including lithography have been driving the industry over these many years. Recently the gate-all-around nanosheet architecture has been introduced enabling 45% higher performance or 75% lower power consumption compared to the 7-nm-FinFET technology thus leading to robust, higher performant devices for future technology nodes [1]. Stacking of devices and new architectures like the Vertical-Transport Nanosheet Field Effect Transistor (VTFET) are currently explored as future possible options [2]. Reducing the power consumption and increasing the performance as well as density will remain the driving force for future advancements [3].

However, these important innovations in classical computing are not sufficient to cope with the changing workloads we encounter in artificial intelligence (AI). The compute requirements needed to train large AI models is doubling every six months which is unsustainable without significant hardware and software innovation. To fuel this trend specialized technologies are developed to accelerate AI workloads. One approach to accelerate AI computing is based on reduced precision named approximate computing. Hereby, the precision of training systems can be aggressively scaled down to even 4-bits without significant loss in accuracy across application domains while enabling significant hardware acceleration [4]. In another approach, we tackle the problem of the Von-Neumann architecture of split memory and processing unit requiring costly data transfer. A solution to this problem is analog in-memory computing where synaptic weights are stored locally in conductance values of non-volatile memristive devices. A crossbar layout allows to perform multiply-accumulate operations, the dominant compute operation in deep neural networks, in an analog manner and promises a significant speed up and decrease of power consumption [5, 6].

Despite the advances described above there are still many significant and relevant problems that are intractable to classical computers and AI accelerators but could be addressed by quantum computers. The last few years have witnessed a strong evolution in quantum computing technologies developing the entire stack from the bottom up. These quantum systems continue to scale in size, quality and speed; utilizing modularity and quantum communication will enable scaling and will boost computational capacity. The implementation of recent error mitigation approaches starts to enable interesting computational regimes in which quantum computers run circuits beyond the reach of brute-force classical simulations called the era of quantum utility [7]. The IBM Quantum roadmap lays out how the technological improvements therein open up new opportunities not only for large-scale applications utilizing error-mitigation, but also pave the way toward future error-corrected systems within the next decade.

References

[1] R. Bao et al., IEDM, DOI:10.1109/IEDM45741.2023.10413745 (2023).
[2] W. Cao, et al., Nature **620**, 501 (2023).
[3] S. Datta, et al., Science **378**, 733 (2022).
[4] X. Sun et al., Advances in Neural Information Processing Systems **33**, 1796 (2020).
[5] S. Ambrogio, et al. Nature **620**, 768 (2023).
[6] M. Le Gallo, et al. Nature Electronics **6**, 680 (2023).
[7] Y. Kim et al. Nature **618**, 500 (2023).