# Invited: A Systems Approach to Computing in Beyond CMOS Fabrics

Ameya Patil[1], Naresh Shanbhag[1], Lav Varshney[1], Eric Pop[2], H.-S. Philip Wong[2], Subhasish Mitra[2], Jan Rabaey[3], Jeffrey Weldon[4], Larry Pileggi[4], Sasikanth Manipatruni[5], Dmitri Nikonov[5], and Ian Young[5]
[1]Univ. of Illinois at Urbana-Champaign, IL, [2]Stanford Univ., Stanford, CA, [3]Univ. of California, Berkeley, CA, [4]Carnegie Mellon Univ., Pittsburgh, PA, [5]Intel Corp., Hillsboro, OR
{adpatil2, shanbhag, varshney}@illinois.edu, {epop, hspwong, subh}@stanford.edu, jan@eecs.berkeley.edu, {weldon, pileggi}@ece.cmu.edu, {sasikanth.manipatruni, dmitri.e.nikonov, ian.young}@intel.com

Emerging applications require computing platforms to extract task-relevant information from increasingly large amounts of data. These requirements place stringent constraints on energy efficiency, throughput, latency, and for certain data types, security and privacy of computing platforms. Traditionally, silicon CMOS scaling has been relied upon to meet these energy and delay constraints. However, the energy and delay benefits achievable via scaling are diminishing. Increased vulnerability to various sources of variations (e.g., process, voltage) further exacerbates these energy and speed challenges.

These trends have led to an intense exploration of beyond-CMOS devices. However, thus far, very few device options have been found to be more energy-efficient and faster than silicon CMOS. Some of the beyond-CMOS devices tend to be inherently stochastic in nature (e.g., spin). Special care is required to overcome variations in nano-scale devices (e.g., imperfection-immune paradigm to overcome imperfections and variations in CNFETs). There are also opportunities to leverage device-level stochasticity (by proper device design) for efficient implementations of new computation models.

We assert that for the semiconductor industry to continue to improve energy efficiency, it is critical that exploratory beyond-CMOS device research be conducted in concert with an end-to-end approach that includes an exploration of alternative *statistical models of computation* driven by the needs of emerging applications. These emerging workloads call for a migration from an *algorithmic* compute world dominated by Turing-inspired processes to a *learning-based* information processing paradigm. We further assert that this migration towards *information processing facilitates the exploration of beyond-CMOS devices* as the very same device properties (e.g., stochasticity), which are deemed as problematic for conventional computing, are in fact useful features to be leveraged by this renewed focus on data-driven learning systems.

We present *Shannon* and *brain-inspired statistical computing models* as a bridge between the requirements of emerging workloads and the unique properties of beyond-CMOS devices. For example, these models and workloads allow for systems to be designed using stochastic components such that their accuracy is equivalent to that of conventional von Neumann realizations. This focus on information extraction as opposed to data processing in emerging workloads motivates the use of *statistical information-based design metrics* such as mutual information, signal-to-noise ratio, and others. The use of such metrics opens the design space at both the component and system levels leading to new design methods and unique energy-delay-reliability trade-offs. For example, employing Shannon-inspired models of computation to realize spin-based inference kernels allows these devices to operate at an average switching error rate that is 1000X greater than that permitted by von Neumann architectures, thereby making spin-based devices competitive with CMOS. *Hyperdimensional computing* (HD) enables the realization of powerful cognitive systems on heterogeneous monolithic 3D semiconductor platforms using both beyond-silicon devices such as CNFET & RRAM, and CMOS. Variations in RRAM (monolithically 3D-integrated using CNFETs) in fact enable efficient realizations of some of the HD functions such as random mapping from the input to the HD space.

These statistical models of computation are best served when the components are application-aware. To realize this, we suggest the use of *nanofunctions*, rather than switches or a binary memory device, as the abstraction between device and systems design. Nanofunctions will have sufficiently complex functionality for a system designer to recognize and leverage for system design and yet be sufficiently simple for the device researcher to be able to exploit the inherent device properties, and fabricate beyond-CMOS prototypes. Thus, they provide a sandbox for exploring the synergies between systems/architectures and devices. For example, nanofunctions for emerging workloads include projection kernels such as vector *dot product*, distance kernels, squashing functions, nano-oscillators, and others. Of these, the dot product is the most complex and most commonly used. Dot product realizations in graphene, CNFET, RRAM, and nano-oscillators in RRAMs have been realized recently. As the Shannon/brain-inspired statistical computing models do not require these nanofunctions to exhibit deterministic behavior, it is possible to relax the specifications and benefit from the resultant energy and delay savings.

## ACKNOWLEDGMENTS

This work was supported in part by Systems on Nanoscale Information fabriCs (SONIC), one of the six SRC STARnet Centers, sponsored by MARCO and DARPA.