The N3XT Technology for Brain-Inspired Computing

H.-S. Philip Wong
Department of Electrical Engineering and Stanford SystemX Alliance
Stanford University, Stanford, California 94305, USA

E-mail: hspwong@stanford.edu

21st century information technology (IT) must process, understand, classify, and organize vast amount of data in real-time. 21st century applications will be dominated by memory-centric computing operating on Tbytes of *active* data with little data locality. At the same time, massively redundant sensor arrays sampling the world around us will give humans the perception of additional "senses" blurring the boundary between biological, physical, and cyber worlds. Abundant-data processing, which comprises real-time big-data analytics and the processing of perceptual data in wearable devices, clearly demands computation efficiencies well beyond what can be achieved through business as usual.

Advances in brain-inspired computing are making rapid progress to meet the demands of abundant-data processing using a variety of techniques, including spiking neural networks, hyper-dimensional computing using sparse vectors, deep neural nets, deep belief nets, restricted Boltzmann machines, and their variants. It is therefore crucial to create a scalable and flexible brain-inspired technology platform that can support all the essential elements, and can be adapted for a wide variety of neural computational model.

The key elements of a scalable, fast, and energy-efficient computation platform that may provide another 1,000× in computing performance (energy-execution time product) for future computing workloads are [1]: massive on-chip memory co-located with highly energy-efficient computation, enabled by monolithic 3D integration using ultra-dense and fine-grained massive connectivity. There will be multiple layers of analog and digital memories interleaved with computing logic, sensors, and application-specific devices. We call this technology platform N3XT – Nanoengineered Computing Systems Technology. N3XT will support computing architectures that embrace sparsity, stochasticity, and device variability.

In this talk, I will give an overview of nanoscale memory and logic technologies for implementing N3XT. In particular, I give an overview of the use of nanoscale analog non-volatile memory devices for implementing brain-inspired computing [2]. Phase change memory (PCM) and resistive switching memory (RRAM) are used as examples to illustrate the need to co-design, co-optimize the device technology, circuit design, system architecture, and learning algorithms.

Acknowledgements: This work is supported in part by member companies of the Stanford Non-Volatile Memory Technology Research Initiative (NMTRI), member companies of the Stanford SystemX Alliance, the National Science Foundation (E2CDA, Expedition in Computing), and also in part by Systems on Nanoscale Information fabriCs (SONIC), one of the six SRC STARnet Centers, sponsored by MARCO and DARPA.

References:

[1] M.M. Sabry Aly, M. Gao, G. Hills, C.-S. Lee, G. Pitner, M.M. Shulaker, T.F. Wu, M. Asheghi, J. Bokor, F. Franchetti, K.E. Goodson, C. Kozyrakis, I. Markov, K. Olukotun, L. Pileggi, E. Pop, J. Rabaey, C. Re, H.-S. P. Wong, S. Mitra, "Energy-Efficient Abundant-Data Computing: The N3XT 1,000X," *IEEE Computer*, pp. 24 – 33, December 2015

[2] S. B. Eryilmaz, D. Kuzum, S. Yu, H.-S. P. Wong, "Device and System Level Design Considerations for Analog-Non-Volatile-Memory Based Neuromorphic Architectures," invited paper, *IEEE International Electron Devices Meeting (IEDM)*, paper 4.1, pp. 64 – 67, Washington, DC, December 7 – 9, 2015.