Materials Challenges for AI Hardware Accelerators

James B. Hannon

IBM Research Division, T.J. Watson Research Center, Yorktown Heights, NY, USA

Email: jbhannon@us.ibm.com

The growth of artificial intelligence (AI) workloads presents a serious challenge to high-performance computing. These computationally-intensive applications include image classification, speech recognition, and anomaly/fraud detection. In typical applications, neural networks with tens of millions of free parameters are "trained" using huge labeled datasets. Training of sophisticated networks can take weeks, even in state-of-the-art data centers. Analysis of the algorithms shows that a large fraction of the training time is spent performing matrix operations. These tasks are amenable to hardware acceleration, and the use of graphical processing units (GPUs) to accelerate training is now common. In this talk, I will describe a new class of exotic hardware accelerators based on *analog computing*. These accelerators employ cross point arrays in which the conductance of each cross-point element can be individually tuned. By applying voltage to the rows, and summing the currents in the columns, matrix multiplications can be efficiently performed. In principle, this scheme can be used to accelerate AI workloads by many thousands of times compared to conventional GPUs. The requirements of the training algorithms place serious constraints on the materials properties of the cross-point elements [1]. While many materials have been proposed, and implemented, none completely satisfy the needed requirements. I will review the materials requirements in detail and describe results from recent implementations using phase change materials.

References:

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