Extended Abstracts of the 2016 International Conference on Solid State Devices and Materials, Tsukuba, 2016, pp113-114

B-7-01(Invited)

Memory Devices for Brain-Inspired Computing

S. Burc Eryilmaz*, Haitong Li, Weier Wan, and H.-S. Philip Wong*

Department of Electrical Engineering and Stanford SystemX Alliance, Stanford University, Stanford, CA 94305, USA
*E-mail: eryilmaz@stanford.edu, hspwong@stanford.edu

Abstract

This paper focuses on brain-inspired hardware for energy efficient cognitive computing using analog resistive memories as synaptic arrays. Resistive memories are superior to its counterparts in terms of area and system-level energy consumption. Through experimental measurements, we illustrate the need for algorithm-device co-design for energy efficient inter-array communication as well as accurate compact models that capture analog conductance change to meet the challenges of scaling up in system size.

1. Introduction

As the amount of data that is generated increases exponentially in the new era of smart consumer products (smart phones, etc) and internet-of-things (IoT), scalable solutions for energy efficient information extraction from huge amounts of data are required. In a wide range of mobile devices that have a power budget from high-end smartphones [1] to low end IoT front end sensor devices [2,3]; power consumption due to connectivity/data transmission dominates the overall power consumption. These applications require feature extraction, data compression, and data processing in the front end device, in order to reduce the amount of data transmitted to back end platform for further processing with little to no information loss; as well as for rapid real-time response. Since such algorithms can easily become computationally and energetically expensive, careful customization of hardware is needed in the front-end sensor device. For such hardware, memory systems (on-chip or off-chip memory) can consume from 25 to 60% of power of the application SoC (not including radio), emphasizing the importance of memory technology in front-end devices [1-3].

Further processing of data aggregated from sensor nodes requires significant amounts of computational power and is typically implemented on cloud-based platforms. Deep neural networks, which have shown promise for extracting information from big amounts of data, are currently trained on conventional hardware such as CPUs [4,5] or GPUs [6-8]. However, such large-scale systems can easily consume 100s of kWs of power during training, mainly due to moving big amounts of data between massive amounts of off-chip memories and thousands of processors [4,5]. For data-intensive tasks, energy consumption due to memory accesses are > 40% of the overall system energy [9]. Scaling up these systems for energy efficient (deep) learning and data mining requires careful hardware customization combined with low-energy memory technologies that are scalable in size and that can enable 3-dimensional monolithic integration for fast and energy efficient on-chip memory.

2. Brain-inspired hardware

Brain inspired hardware is a class of hardware that is intended to implement brain-inspired algorithms under 2 classes: 1) biology based network models and learning rules and 2) artificial neural networks that are inspired by the biological brain to some extent, but do not strictly mimic the brain. Hardware customization with conventional CMOS technologies for some brain-inspired algorithms has already proved to be low power for training or inference tasks [10-12]. Brain-inspired hardware aims to realize the real-time processing power and energy efficiency of biological brain, and gets inspirations from real brain in terms of connectivity, processing, and/or communication schemes on the device, circuit, and architecture level. Because the number of synapses in neural networks is way larger than the number of neurons and scales quadratically, implementation of synaptic weights (memories) deserves special attention. Resistive memory technologies such as resistive metal oxide memory (RRAM) [13], phase change memory (PCM) [14], and conductive bridge memory (CBRAM) [15] offer significant advantages compared to their conventional counterparts such as SRAM or DRAM: low switching energy, excellent size-scalability, monolithic 3-D integration and analog programmability. Brain-inspired (neuromorphic) architectures with analog resistive memory can benefit from all these characteristics of emerging memories. This type of hardware utilize analog programmability of resistive memories for implementing synaptic weights, both for realizing biologically realistic networks such as Hopfield network [16]; as well as artificial neural networks such as feedforward network [17], restricted Boltzmann machine [18], and convolutional kernel [19]. Here, we review the device requirements and design considerations for brain-inspired hardware with analog resistive synapses.

3. Design considerations for brain-inspired hardware with analog resistive synapses

Device level considerations

It is worth noting that to determine the requisite memory characteristics for a given application or learning algorithm, analysis should be performed with accurate device models that can capture cycle-to-cycle variations, device-to-device variations, nonlinearities in weight update; while also considering the noise in the available data, and the accuracy needed for the application. We study algorithm-device interaction for a restricted Boltzmann machine (RBM) with

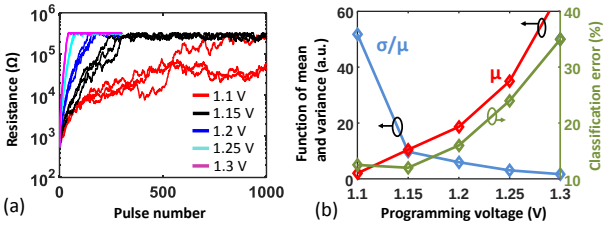

Fig. 1 (a) Gradual resistance change in RRAM for different programming voltages (b) effect of programming voltage on learning performance. μ refers to average log-conductance change for one pulse, σ refers to standard variation of fluctuations of log-conductance change around a smooth fit over a cycle.

one hidden layer, with MNIST digit recognition as the task in hand. RRAM model used is adapted from [20], and is calibrated with experimental data. Gradual resistance change (gradual RESET) of resistive device is shown in Fig. 1(a). Note that for this particular device, gradual resistance change occurs in one direction; whereas in SET direction, resistance changes abruptly. To overcome this problem, a differential weight encoding scheme is used [21]. Fig. 1(b) shows the relation between the programming voltage and test error for MNIST digit data. Programming voltage determines 1) conductance change for a single gradual RESET pulse (averaged over cycles), and 2) cycle-to-cycle variations; both of which affect the classification error for MNIST test set after training. We observe that device-to-device variations are tolerated during training to the extent of the accuracy of gradual weight increments.

Array and system level considerations

Brain-inspired algorithms mentioned earlier in this work typically require network architectures with 1000-10000 fan-out. Two major components of energy consumption within synaptic arrays are 1) wire energy (CV^2) and 2) programming energy ($V^2 t_{pulsewidth}/R$). Wire energy begin to dominate array level energy consumption in 1k×1k arrays (few pJs), and constitute a larger fraction as the array gets bigger. Energy consumption analysis on array level for the case study presented above is shown in Fig. 2(a). Higher average energy/epoch at the beginning of the training is due to the initial SET programming of all devices. As training progresses, devices are reprogrammed with gradual RESET pulses; and the effect of initial SET state diminishes after 50 epochs. Both device energy and wire energy scales quadratically with programming voltage, hence low voltage programming is desired for lower energy consumption. Array size should be determined with the following considerations in mind: 1) communication overhead (latency and energy) for using multiple smaller arrays, 2) wire energy within large, arrays 3) IR drop within large arrays, 4) connectivity/fan-out of network architecture [22].

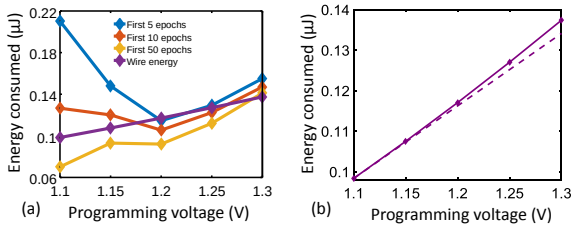

Fig. 2 (a) Energy consumption within RRAM and wires (b) Quadratic dependence of energy on V compared with linear curve.

4. Open research questions

Rapid progress is being made in this field. Some of the open research problems and needs for the field are listed below:

- Beyond demonstration of functionality, application level benchmarking is needed for fully integrated brain-inspired hardware with resistive synaptic arrays. Quantities of interest can be performance per power, performance per area, or how reliable the system is to the variability of a given memory technology.
- Scalable solutions for communicating between arrays are needed for scaling up in system size and achieving high fan-outs needed for brain-inspired computing. Hierarchical connectivity should be explored [23] for energy efficiency; since it is analogous to the connectivity of biological brain, where connections are locally dense and globally sparse [23].
- Low voltage memory devices should be explored, since both wire and programming energy scales quadratically with programming voltage during training
- Biologically realistic (spiking) networks allow continuous real time online learning [24]; however, more research is needed to identify problems that can efficiently be solved by STDP. So far, RBM and S2M (synaptic sampling machine) are 2 problems where STDP excel at [24,25], but it is still in its infancy compared to commonly used backpropagation.
- Algorithm-device co-design is needed to identify what device characteristics are important for different applications. Device models that can accurately model the statistics of gradual conductance change are needed.
- Monolithic 3D integration under low temperature is required for the future synaptic device technologies.

Acknowledgements

This work is supported in part by SONIC, one of six centers of STARnet, a Semiconductor Research Corporation program sponsored by MARCO and DARPA, the NSF Expedition in Computing (Visual Cortex on Silicon, award 1317470), and the member companies of the Stanford Non-Volatile Memory Technology Research Initiative (NMTRI) and the member companies of the Stanford SystemX Alliance.

References

[1] A. Carroll et al., USENIX, **14** (2010). [2] H. Kim et al., IEEE Trans. Biomed Circuits Syst., **8**, 2 (2013). [3] D. Bol et al., IEEE J. Solid-State Circuits, **48**, 1 (2012). [4] R. Ananthanarayanan et al., SC, pp. 1-12 (2009). [5] Q. V. Le et al., ICML, 2012. [6] A. Krizhevsky et al., NIPS, 2012. [7] R. Raina et al., ICML, pp. 873-880, 2009. [8] V. Vanhoucke et al., Deep Learning and Unsupervised Feature Learning Workshop, 2010. [9] M. M. S. Aly et al., Computer, **12**, 12 (2015). [10] Y. H. Chen et al., ISSCC (2016). [11] S. Han et al., arXiv preprint arXiv:1602.01528 (2016). [12] E. H. Lee et al., ISSCC (2016). [13] S. Yu et al., IEEE TED (2011). [14] D. Kuzum et al., Nano Letters, **12**, 5 (2011). [15] T. Ohno et al., Nature Mat., **10**, 8 (2011). [16] S. B. Eryilmaz et al., IEDM (2013). [17] G. Burr et al., IEDM (2014). [18] S. B. Eryilmaz et al., IEEE TED (in review). [19] L. Gao et al., IEEE EDL (2016). [20] Z. Jiang et al., SISPAD, 41-44 (2014). [21] O. Bichler et al., IEEE TED, 2012. [22] S. B. Eryilmaz et al., IEDM (2015). [23] S. Joshi et al., CNNA (2010). [24] E. Neftci et al., Front Neurosci, **7**, 272 (2014). [25] E. Neftci et al., Front Neurosci, **10**, 241 (2016).