Special Session Paper
3D Nanosystems Enable *Embedded* Abundant-Data Computing

William Hwang[1], Mohamed M. Sabry Aly[1], Yash H. Malviya[1], Mingyu Gao[1], Tony F. Wu[1],
Christos Kozyrakis[1,2], H.-S. Philip Wong[1], Subhasish Mitra[1,2]
Department of Electrical Engineering[1] and Department of Computer Science[2], Stanford University

ABSTRACT
The world's appetite for abundant-data computing, where a massive amount of structured and unstructured data is analyzed, has increased dramatically. The computational demands of these applications, such as deep learning, far exceed the capabilities of today's systems, especially for energy-constrained embedded systems (e.g., mobile systems with limited battery capacity). These demands are unlikely to be met by isolated improvements in transistor or memory technologies, or integrated circuit (*IC*) architectures alone. Transformative nanosystems, which leverage the unique properties of emerging nanotechnologies to create new IC architectures, are required to deliver unprecedented functionality, performance, and energy efficiency. We show that the projected energy efficiency benefits of domain-specific 3D nanosystems is in the range of 1,000x (quantified using the product of system-level energy consumption and execution time) over today's domain-specific 2D systems with off-chip DRAM. Such a drastic improvement is key to enabling new capabilities such as deep learning in embedded systems.

3D NANOSYTEMS
Our three-dimensional (*3D*) nanosystems approach (Fig. 1a) is based on the Nano-Engineered Computing Systems Technology (N3XT) approach [1]. The specific technologies we focus on in this paper include: (a) energy-efficient digital logic using carbon nanotube field-effect transistors (*CNFETs*) [2], (b) high-density, nonvolatile memory using 3D vertical resistive RAM (*3D RRAM*) [3], and (c) ultra-dense (e.g., monolithic) 3D integration with fine-grained vertical connectivity. At the architecture level, we use: (d) domain-specific accelerators for deep learning inferencing and (e) reduced-precision arithmetic operations that trade off inference accuracy for improved execution time and energy consumption.

Digital systems implemented using CNFETs can enable an order of magnitude improved Energy Delay Product (*EDP*, simultaneously ~3x faster and ~3x lower energy) versus FinFETs [4], when analyzed using detailed models calibrated with experimental measurements [5]. The imperfection-immune paradigm [6, 7] overcomes imperfections and variations inherent to CNFETs. This approach enabled the demonstration of CNFET-based processors [8].

RRAM is denser than DRAM ($4F^2$ RRAM vs. $6F^2$ DRAM bit cell area), with <0.1 pJ/bit read and <1 pJ/bit write cell-level energy, in addition to latencies on the order of 1 and 10 ns for read and write, respectively [9]. 16 Gbit RRAM prototypes have been demonstrated [10]. 3D vertical RRAM [3], where multiple RRAM cells are placed on top of each other (and share the same access transistor), can provide up to 16x DRAM density for the same area footprint. 3D RRAM has similar access energy and latency as the aforementioned (traditional) RRAM.

Dense (e.g., monolithic) 3D integration vertically connects adjacent layers (logic or memory) using interlayer vias (*ILVs*). In today's 2D ICs, ILVs are used to connect the back-end-of-line metal layers. These ILVs can be >1,000x denser compared to through-silicon vias (*TSVs*) used in chip stacking today [1]. CNFETs and RRAM naturally enable monolithic 3D integration due to their low-temperature fabrication (<200°C). A recent paper [11] demonstrated a 4-layer monolithic 3D nanosystem that includes millions of CNFETs and 1 Mbit of RRAM on top of a silicon CMOS bottom layer.

CPU-based 3D nanosystems can provide significant energy efficiency benefits for servers and embedded systems [1, 12]. Here, we analyze domain-specific accelerator-based 3D nanosystems. Accelerators are optimized for domain-specific operations. There are many examples of accelerators, including those targeting deep learning (e.g., Google TPU [13], TETRIS [14], Eyeriss [15]). The massively concurrent memory accesses enabled by 3D nanosystems are key to fully harness the benefits of accelerators for compute- and memory-bound applications.

To further improve energy efficiency, reduced-precision operations (e.g., 8 bits vs. 32 or 64 bits) have long been used in application-specific designs. An in-depth study shows that deep learning applications with 8-bit representations achieve a 4x speedup compared to 32-bit representations, with <1% impact on classification accuracy [16].

SIMULATION RESULTS
We analyze the inference phase of state-of-the-art deep neural networks (*DNNs*) [13]. Training phase analysis is part of our future work. In particular, we examine convolutional neural networks (*CNNs,* e.g., for embedded computer vision) and long short-term memory (*LSTMs,* e.g., for speech recognition and translation). We perform full physical design of the MAC unit (including the local and shared SRAM shown in Fig. 1). We use 28nm foundry PDK for the 2D baseline and 28nm CNFET PDK for our 3D nanosystem. The clock tree energy is scaled from published 65nm accelerator data [16]. We use ZSim [17] for architecture simulation, and the trace-based simulation framework to analyze accelerators [14].

The 2D baseline and the 3D nanosystem use the same architecture (Fig. 1). We use 8-bit precision (similar to [13]) MAC units. Our 3D nanosystems achieve major benefits (~1,000x range) in system-level EDP (total energy consumption × execution time) as shown in Fig. 2. In contrast, the benefits of

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
CODES/ISSS '17 Companion, October 15–20, 2017, Seoul, Republic of Korea
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5185-0/17/10…$15.00
https://doi.org/10.1145/3125502.3125531

chip stacking using TSVs (DRAM stacked on the accelerator chip with the same ratio between the number of MACs and TSVs as [14]) range from 2x to 9x for 8-bit accelerators. We use batch sizes commonly observed in embedded systems [18]. As LSTM networks are more memory-bound (also observed in [13]), they benefit the most from our 3D nanosystems approach.

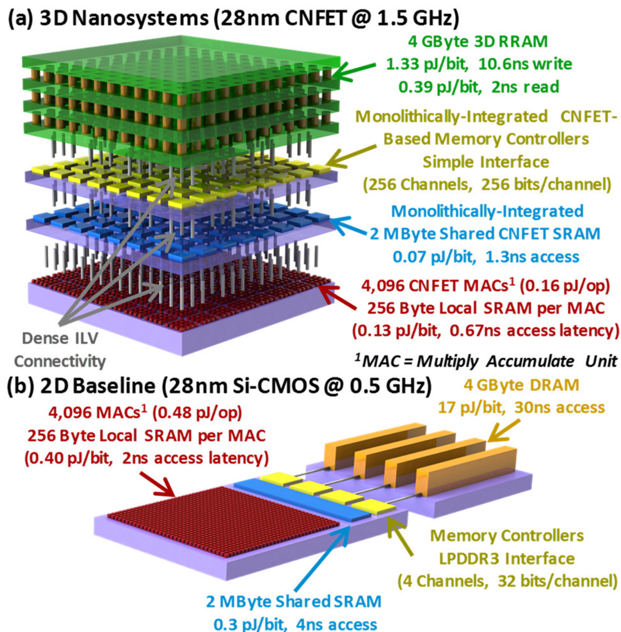

Figure 1: (a) 3D Nanosystem and (b) 2D Baseline

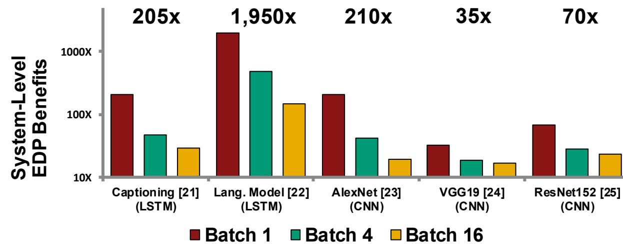

Figure 2: System-level EDP benefits of our 3D nanosystem vs. 2D baseline, with 8-bit MAC units, for DNNs. Maximum benefits for each network are highlighted above.

We performed thermal analysis of our 3D nanosystem and the 2D baseline using 3D-ICE [19]. The peak temperature and average power density of our 3D nanosystem is 35°C and 9.5 W/cm^2, respectively. For 2D baseline, the corresponding values are 36°C and 10.4 W/cm^2, respectively. These temperatures are comparable to typical workloads on today's mobile systems [20].

For RRAM, a key challenge is the limited write endurance (e.g., 10^7 RRAM vs. 10^{15} DRAM) [9]. We overcome this challenge using an architecture-level mechanism that shapes write distribution to RRAM by: (a) balancing the distribution of writes for each word, and (b) buffering a very small number of words (that are written to frequently) using SRAM. This approach enables a 10-year lifetime of our 3D nanosystem despite 10^7 RRAM endurance, while incurring <0.01% performance and energy impact, and 10 KByte SRAM per 1 GByte RRAM. Full details of this approach are not included here due to space constraints.

CONCLUSION

Our 3D nanosystems approach enables massive energy efficiency benefits over today's system architectures through advances across the computing stack, from device technology to architecture design. Such benefits pave a path towards new capabilities such as deep learning in energy-constrained embedded systems.

ACKNOWLEDGEMENT

We acknowledge the support of DARPA, STARnet SONIC, member companies of the Stanford SystemX Alliance, and the NSF Graduate Research Fellowship for W. Hwang.

REFERENCES

[1] M.M.S. Aly *et al.*, "Energy-Efficient Abundant-Data Computing: The N3XT 1,000X," *IEEE Computer,* 2015.

[2] J. Zhang *et al.*, "Carbon Nanotube Robust Digital VLSI," *IEEE Trans. CAD*, 2012.

[3] H.Y. Chen *et al.*, "HfOx based vertical resistive random-access memory for cost-effective 3D cross-point architecture without cell selector," *IEDM,* 2012.

[4] D.J. Frank and L. Chang, "Technology Optimization for High Energy-Efficiency Computation," *IEDM Short Course*, 2012.

[5] G. Hills, "Variation-Aware Nanosystem Design Kit", *https://nanohub.org/resources/22582*

[7] G. Hills *et al.*, "Rapid Co-optimization of Processing and Circuit Design to Overcome Carbon Nanotube Variations," *IEEE Trans. CAD*, 2015.

[8] M.M. Shulaker *et al*., "Carbon nanotube computer," *Nature,* 2013.

[9] H.-S.P. Wong and S. Salahuddin, "Memory Leads the way to better computing," *Nature*, 2015.

[10] R. Fackenthal *et al*., "A 16Gb ReRAM with 200MB/s Write and 1GB/s Read in 27nm Technology," *ISSCC*, 2014.

[11] M.M. Shulaker *et al*., "Three-dimensional integration of nanotechnologies for computing and data storage on a single chip," *Nature*, 2017.

[12] R. Braojos *et al*., "Nano-engineered architectures for ultra-low power wireless body sensor nodes," *CODES+ISSS*, 2016.

[13] N. Jouppi *et al*. "In-Datacenter Performance Analysis of a Tensor Processing Unit," *ISCA*, 2017.

[14] M. Gao *et al*., "TETRIS: Scalable and Efficient Neural Network Acceleration with 3D Memory," *ASPLOS*, 2017.

[15] Y.-H. Chen *et al*., "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," *IEEE JSSCC*, 2017.

[16] C. De Sa *et al.*, "Understanding and Optimizing Asynchronous Low-precision Stochastic Gradient Descent," *ISCA*, 2017.

[17] D. Sanchez *et al*., "ZSim: Fast and Accurate Microarchitectural Simulation of Thousand-Core Systems," *ISCA*, 2013.

[18] V. Sze *et al.*, "Efficient Processing of Deep Neural Networks:A Tutorial and Survey," arXiv preprint, 2017.

[19] A. Sridhar *et al*., "3D-ICE: A Compact Thermal Model for Early-Stage Design of Liquid-Cooled ICs," *IEEE Trans. Computers*, 2014.

[20] V. Chiriac *et al.*, "A figure of merit for mobile device thermal management," *IEEE ITherm*, 2016.

[21] O. Vinyals *et al.*, "Show and Tell: A Neural Image Caption Generator," *IEEE CVPR*, 2015.

[22] R. Jozefowicz *et al.*, "Exploring the Limits of Language Modeling," arXiv preprint, 2016.

[23] A. Krizhevsky *et al.*, "ImageNet Classification with Deep Convolution Neural Networks," *NIPS*, 2012.

[24] K. Simoyan *et al.*, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *ICLR*, 2015.

[25] K. He *et al.*, "Deep Residual Learning for Image Recognition," *IEEE CVPR*, 2016.