Thermal Feasibility of Die-Stacked Processing in Memory

Yasuko Eckert Nuwan Jayasena Gabriel H. Loh
AMD Research
Advanced Micro Devices, Inc.
{yasuko.eckert, nuwan.jayasena, gabriel.loh}@amd.com

Abstract

Processing in memory (PIM) implemented via 3D die stacking has been recently proposed to reduce the widening gap between processor and memory performance. By moving computation that demands high memory bandwidth to the base logic die of a 3D memory stack, PIM promises significant improvements in energy efficiency. However, the vision of PIM implemented via 3D die stacking could potentially be derailed if the processor(s) raise the stack's temperature to unacceptable levels. In this paper, we study the thermal constraints for PIM across different processor organizations and cooling solutions and show the range of designs that are viable under different conditions. We also demonstrate that PIM is feasible even with low-end, fanless cooling solutions. We believe these results help alleviate PIM thermal feasibility concerns and identify viable design points, thereby encouraging further exploration and research in novel PIM architectures, technologies, and use cases.

1. Introduction

Processors have been increasing in computation performance and energy efficiency at a much faster pace than have improvements in bandwidth, latency, and energy of off-chip memory accesses. As a result, the memory system is often a performance bottleneck and accounts for an increasingly significant fraction of system-level energy consumption [11, 25]. Emerging workloads that exhibit memory-intensive behaviors with irregular access patterns, limited data reuse, and/or large working set sizes exacerbate this problem.

Moving computation closer to data has the potential to improve both the performance and the energy efficiency of memory accesses. One approach to achieve this is to integrate processing-in-memory (PIM) capabilities with memory dies using 3D die stacking. Memory-bound computations can then be offloaded from the main "host" processor to these auxiliary, in-memory processors. Such an organization is shown in Figure 1. Recent evaluations of 3D-stacked PIM have shown tremendous promise with an order of magnitude or more improvements in energy efficiency and/or performance for memory-intensive workloads [19, 27].

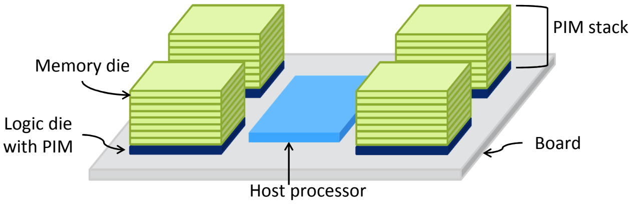

Figure 1: Example compute-node design with PIM

Key technologies necessary to realize 3D-stacked PIM are already being adopted in the industry. Recently released High Bandwidth Memory [7] and Wide I/O [8] JEDEC standards aim to commoditize DRAM that can be stacked on top of logic. Another recent memory technology, Hybrid Memory Cube (HMC), consists of DRAM stacked atop a "base" logic die [18]. While the logic die of HMC contains memory controllers and other miscellaneous logic today, one can easily envision adding more complex processing in the near future.

A potentially significant challenge for PIM is thermal management. Because the memory lies between the heat sink and the logic die, heat generated from the logic die raises the temperature of the memory.[1] The typical operating temperature range for DRAM is under 85°C. After the temperature exceeds that threshold, the refresh rate must be doubled for every $\sim 10^{\circ}$C increase [13]. Higher refresh rates not only consume more power but also reduce the availability of the DRAM, resulting in lower memory performance. PIM-type designs that use the stacked memory as main memory make thermal management even more challenging, as main memory does not traditionally have aggressive cooling solutions commonly used for processors. Hence, it is not immediately obvious whether it is thermally feasible to put a reasonable amount of computing inside a memory package.

The focus of this paper is to investigate whether the thermal constraints of 3D integration make the concept of die-stacked PIMs pointless for any further investigation, rather than to propose novel techniques. We consider a range of cooling solutions, from passive cooling (i.e., heat sink alone) to high-end-server active cooling (i.e., heat sink plus fan). Our evaluation shows that low-cost passive cooling is sufficient to cool a PIM stack with an 8.5W processor integrated with memory. This power budget is in the same range as that of modern, low-power, laptop processors [22]. With more expensive active cooling solutions, the allowable power budget increases to 55W, approaching that of high-performance processors. We therefore conclude that, although the exact cooling solution is a critical factor in determining how much compute can be integrated with memory, thermal constraints do not represent an insurmountable hurdle for PIM. We also find that the type of PIM processor, such as many simple single-instruction-multiple-data (SIMD) units or a few out-of-order cores, becomes more significant at greater power budgets as the variations in thermal hot spots become more pronounced. Hence, the contribution of this paper is to alleviate thermal concerns about PIM so that the community can focus on novel PIM-architecture research.

[1] While thermally attractive, placing the processor next to the heat sink is impractical because all of the power and IO signals must then be routed through all of the DRAM layers.

2. Background

2.1. PIM Designs

PIM attracted significant attention in the research community around the beginning of this century. Many of those efforts focused on embedded DRAM in logic processes (e.g., IRAM [5]) or logic implemented in DRAM processes (e.g., ActivePages [17], DIVA [3], FlexRAM [9]). However, these suffered from the reduced density of embedded DRAM or the reduced performance and high cost (in terms of lost DRAM density) of logic implemented in DRAM processes.

Recent work revisited the PIM concept by leveraging 3D die stacking with through-silicon vias (TSVs) [15, 19, 21, 27]. By implementing the in-memory processor in a logic process and the memory in a memory process and connecting them with high-bandwidth TSVs, they maintained memory density and high logic performance.

2.2. Prior Thermal Analyses of Memory-Logic Stacking

Milojevic et al. investigated three different cooling solutions for a two-die memory stacked on a larger, low-power 16-core ARM Cortex-A9 die and concluded that a passive heat sink is sufficient to keep the stacked memory under 90 °C [16]. Loh considered stacking memory on a high-performance, 92W, quad-core processor, demonstrating that an eight-die stacked memory plus a logic die can be kept below 95 °C with an active heat sink [14]. Our paper differs from these prior studies in that we evaluate a range of processor organizations and cooling solutions to understand the degree and type of PIM capabilities that are feasible under different design choices.

Other thermal-related prior work includes Thermal Herding [20] and work by Li et al. [12]. The former proposes microarchitectural techniques to lower the junction temperature of die-stacked CPU cores with a fixed cooling solution while the latter investigates the thermal constraints of various chip-multiprocessor configurations in the context of a 2D planar die. Neither of them evaluate the thermal interference between DRAM and logic packaged into the same 3D stack.

3. Modeling PIM Stack

3.1. PIM-Stack Configuration

Given PIM's forward-looking vision to add in-memory processing on the logic die of a memory stack, we assume process technology nodes that are likely to be mainstream in the future. Based on ITRS projections [24], we use 25nm and 10nm nodes for the memory and logic die, respectively, for our analysis. A recent Wide I/O DRAM implementation stacked two 64mm^2 die of 1Gbit DRAM in 50nm technology [10]. Vogelsang projects that a 25nm node quadruples the DRAM density of a 50nm node [26]. We further assume that DRAM die area remains constant (as has historically been the case) and that eight-die stacking of such DRAM is feasible in the assumed timeframe, providing a total of 4GB DRAM capacity per PIM stack.

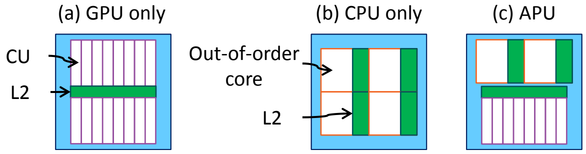

Figure 2: PIM logic-die design options evaluated

Table 1: Thermal-modeling parameters

Parameter	Value
Silicon thermal resistivity	0.0083 $\frac{m-K}{W}$ [14]
Metal-layer thermal resistivity	0.083 $\frac{m-K}{W}$ [14]
Die-to-die-layer thermal resistivity	0.0166 $\frac{m-K}{W}$ [14]
Die size	8mm x 8mm [16]
Die count	8 (memory) / 1 (logic)
Ambient temperature	25 °C [4]

We assume the base logic die area matches the stacked memory die area because PIM is intended to supplement the host processor rather than implement full processing capability. We use our in-house analysis for a 10nm node based on past industry trends and scaling projections to determine the PIM resources that fit within that area. We then consider three architecture design options with various power-density profiles under this area constraint. The first one is a general-purpose graphical processing unit (GPGPU). A GPGPU consists of many lightweight SIMD units for data-parallel execution and is capable of generating many concurrent memory accesses. Such an architecture, therefore, is a good match for exploiting the high bandwidth available to the integrated stacked memory in an energy-efficient manner. We base the GPGPU design on AMD's Graphics Core Next architecture [1]. The PIM logic die contains 16 Compute Units (CUs) and a 2MB shared L2 and is shown in Figure 2(a).

The second design consists of four two-way multithreaded, out-of-order (OoO) cores with a two-level cache hierarchy per core as shown in Figure 2(b). This was chosen to stress the thermal feasibility of PIM as OoO cores have areas of high power density and high peak temperatures.

The last design combines half the resources from each of the first two designs to form an Accelerated Processing Unit (APU). This design includes two OoO cores, eight CUs, and a 2MB GPU L2 as shown in Figure 2(c). To simplify the thermal modeling, we did not include resources other than compute and caches (e.g., memory controllers) into the logic die. This is because we do not expect other resources to become thermal hot spots under normal operating conditions.

3.2. Evaluation Methodology

We use the HotSpot thermal simulation tool [23] and calibrated it against thermal models from our product teams. We decompose each DRAM and logic die into multiple sub-layers consisting of bulk silicon, active device layer, and metal layers, and we also model the die-to-die via layer. The die furthest away from the heat sink is the logic die. Table 1 lists key parameters for the thermal modeling.

We obtained other inputs to HotSpot, detailed floorplans and power-density distributions of GPU and CPU, from our product teams. The power-density model assumes a compute-intensive behavior for the CPU and many concurrent vector operations and memory accesses for the GPU. At a high level,

Table 2: Evaluated cooling-solution types

Cooling	**Convection Thermal Resistance (°C/W)**
Passive heat sink	4.0 [2]
Low-end active heat sink	2.0
Commodity-server active heat sink	0.5 [16]
High-end-server active heat sink	0.2

the GPU has a fairly uniform power distribution. The OoO CPU tends to consume much more power in the execution units than in the rest of the core, resulting in nearly 2x higher worst-case power density relative to the GPU. For the stacked-memory power, we used power from 1Gb Wide I/O DRAM in 50nm [10] as a baseline and scaled it up for eight 4Gb DRAM chips. We then scaled down the power to a 25nm node using the projected DRAM voltage scaling trend by Vogelsang [26]. The result is 0.7W for an eight-die DRAM stack, which is assumed to be uniformly distributed in the stack. We also evaluate cases in which the memory consumes more power in Section 4.

We assess a wide range of low-cost to high-cost cooling solutions to understand the required level of cooling for a given desirable PIM capability without exceeding DRAM's 85 °C threshold for the nominal refresh rate. As listed in Table 2, we evaluate one passive heat sink (i.e., heat sink only) and three different active heat sinks (i.e., heat sink plus fan) with different cost-performance trade-offs. Note the low-end active heat sink refers to inexpensive consumer-level cooling. We also experimented with a no-heat-sink case, but we found that it severely limits the PIM logic power budget to sub-1W.

4. Thermal Profiles and Analysis

We varied the amount of power allocated to the PIM logic die while holding the total stacked-memory power constant at 0.7W. Distribution of the allocated logic-die power is based on the in-memory processor's power-density characteristics described in the previous section. The APU-die configuration assumes a 40:60 power split between the CPU and GPU in this experiment due to the higher energy efficiency of GPUs. Our goal is to find the maximum logic-die power that can sustain the peak memory temperature below 85 °C for each logic-design and cooling-solution combination.

Figure 3 plots the results when including the thermal effects of all eight layers of memory. Even the low-cost passive heat sink can sustain up to 8.5W of logic power while keeping the nominal DRAM refresh rate and therefore maintaining the DRAM bank availability. The PIM stack can afford the large logic power because of the very low projected total memory power in 25nm. 8.5W of power is in the same power-budget range as that of low-power laptop processors [22]. This suggests that PIM capability can be made comparable to those mobile processors', opening up the possibility of integrating a fairly intelligent in-memory processor into a memory stack to combat the increasing off-chip-memory-access inefficiency.

The addition of even an inexpensive fan doubles the logic-die power budget, to 17W, compared to the passive heat sink. Employing a server-grade active heat sink further increases the power budget. With the commodity-server active heat sink, the power budget goes up to 37-42W, and it is pushed out to the 47-55W range with the high-end-server active heat sink. Although the large power budget can accommodate more complex, higher-performance PIM, it comes with more expensive cooling equipment costs, higher fan-generated heat, and noise [6].

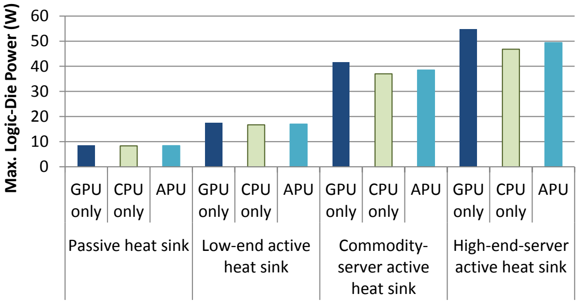

Figure 3: Maximum sustainable logic-die power

Figure 3 also shows a lack of differentiation in the maximum logic power among the design choices when using the passive heat sink, despite the CPU-only design's significantly higher worst-case power density. This is because the variations in hot spots are muted due to the limited thermal headroom afforded by the low-cost passive heat sink. The logic-design choices start to make more impact on the power budgets (up to 8W) with the higher cooling capacity of the active heat sinks. As expected, the CPU's high power density raises the CPU-only design's die temperature more than the GPU-only die's for the same total logic power. Figure 4 illustrates that the latter spreads power, and consequently heat, more evenly throughout the die. The higher logic temperature of the CPU-only die heats up the stacked memory to a greater degree, reducing the maximum sustainable logic power under the 85 °C memory-temperature cap. Hence, the power distribution in the PIM logic layer must be more carefully managed when using server-grade active heat sinks (at the higher power consumption levels).

Although we assumed that the DRAM cannot exceed the 85 °C limit to maintain the performance, we could potentially relax the constraint dynamically during non-memory-bound application phases. A 10 °C increase in the DRAM-temperature limit at the cost of a 2x higher refresh rate allows 10-16% more logic power, boosting the PIM compute performance when computation becomes a bottleneck (results not included due to space constraints). Such a dynamic scheme to trade DRAM performance for in-memory processing capability is especially desirable for stand-alone PIM without a host processor.

Sensitivity to Ambient Temperature: The ambient temperature of a PIM stack may vary substantially depending on the packaging, the distance from the fan (e.g., in dense servers), and the room temperature. We vary the ambient temperature to evaluate the extent of impact on the PIM logic-die power budget. The results are shown in Figure 5. The server-level cooling solutions show more dramatic variations with ambient temperature as their larger cooling capacity is over-provisioned in the baseline case of 25 °C while the low-cost

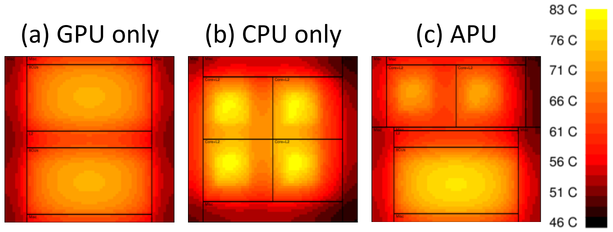

Figure 4: Logic-die thermal maps with high-end-server active heat sink. Logic-die power is 45W.

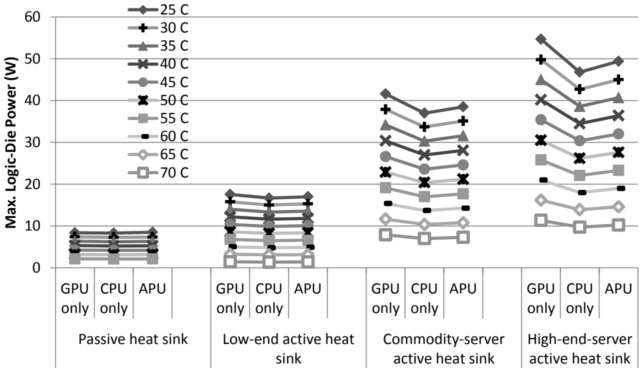

Figure 5: Sensitivity to ambient temperature. Note that a passive heat sink is not feasible at ambient temperatures above 55°C.

heat sink is already under-provisioned. Similar to the earlier observation, higher ambient temperatures reduce the effective cooling capacity, especially for the server-level heat sinks, resulting in less differentiation in the maximum logic power among the three logic designs. Overall, we conclude that for the PIM framework we examined, forced-air cooling becomes necessary to cool the stacked memory plus the logic die once the ambient temperature rises by 20 °C or more. For ambient temperatures higher than 60 °C, server-level heat sinks are necessary.

Sensitivity to Stacked-Memory Power: The previous two studies assumed that the stacked memory always consumes a fixed amount of power. However, the memory power depends on the access patterns, voltage-frequency settings, and other operating conditions. Therefore, we next scale the total stacked-memory power from our projection of 0.7W to 16W while holding the rest of the parameters constant. Figure 6 plots the impact on the maximum sustainable logic power for the two low-end cooling solutions: the passive heat sink and low-end active heat sink. Although only the APU-die results are shown due to space constraints, the GPU- and CPU-design results are within ±0.5W.

The figure demonstrates an inverse linear relationship between the stacked-memory power and the maximum logic power regardless of the cooling-solution types examined. For a 1W increase in the stacked-memory power, the logic-die power budget needs to decrease by approximately 1W to keep the memory temperature below 85 °C. This makes intuitive sense because the total power density of the PIM stack remains roughly the same even when power is shifted from the logic die to the eight memory die. Once the memory power reaches 5x (or beyond) of our initial projection, the logic-die power budget with the passive heat sink may become too constrained,

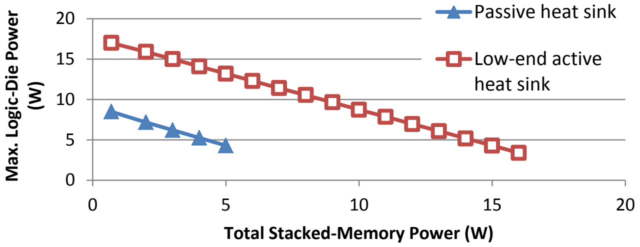

Figure 6: Sensitivity to stacked-memory power

such that an inexpensive fan may need to be attached to enable desirable in-memory processing capability.

5. Conclusions

While processing in memory is a very interesting approach that is enabled by recent advancements in die stacking, all of this excitement could be for naught if the thermal coupling of the logic forces the memory layers into inoperable temperature ranges. Our analyses showed that even with a low-cost passive heat sink, nominal DRAM temperatures can be achieved while still provisioning a useful amount of power for in-memory computation. We also showed that exactly how much power is available for compute is highly sensitive to the chosen cooling solution, but overall the potential thermal concerns of a die-stacked PIM do not impose a big enough challenge to discourage PIM explorations. We feel that these results remove an important roadblock for PIM and help to shed light on the range of feasible PIM design points, thereby encouraging the community to conduct more research on novel architectures and use cases for exploiting in-memory processing capabilities without the need to repeatedly justify PIM thermal feasibility in each independent study.

References

[1] Advanced Micro Devices, Inc. AMD Graphics Core Next (GCN) Architecture. *AMD White Paper*, June 2012.

[2] Chris Gonzales and Hwan Ming Wang. Thermal Design Considerations for Embedded Applications. *Intel White Paper*, December 2008.

[3] J. Draper, J. Chame, M. Hall, C. Steele, T. Barrett, J. LaCoss, J. Granacki, J. Shin, C. Chen, C. W. Kang, I. Kim, and G. Daglikoca. The Architecture of the DIVA Processing-in-Memory Chip. In *Proc. of the 16th Intl. Conf. on Supercomputing*, pages 14–25, New York, NY, June 2002.

[4] P. Emma, A. Buyuktosunoglu, M. Healy, K. Kailas, V. Puente, R. Yu, A. Hartstein, P. Bose, and J. Moreno. 3D Stacking of High-Performance Processors. In *Proc. of the 20th Intl. Symp. on High Performance Computer Architecture*, Orlando, FL, February 2014.

[5] B. R. Gaeke, P. Husbands, X. S. Li, L. Oliker, K. A. Yelick, and R. Biswas. Memory-Intensive Benchmarks: IRAM vs. Cache-Based Machines. In *Proc. of the 16th Intl. Parallel and Distributed Processing Symposium*, Fort Lauderdale, FL, April 2002.

[6] X. Han and Y. Joshi. Energy Reduction in Server Cooling Via Real Time Thermal Control. In *Proc. of the Annual IEEE Semiconductor Thermal Measurement and Management Symposium*, pages 20–27, San Jose, CA, March 2012.

[7] JEDEC. High Bandwidth Memory (HBM) DRAM. http://www.jedec.org/standards-documents/docs/jesd235.

[8] JEDEC. Wide I/O Single Data Rate (Wide I/O SDR). http://www.jedec.org/standards-documents/docs/jesd229.

[9] Y. Kang, W. Huang, S.-M. Yoo, D. Keen, Z. Ge, V. Lam, J. Torrellas, and P. Pattnaik. FlexRAM: Toward an Advanced Intelligent Memory System. In *Proc. of the Intl. Conf. on Computer Design*, Austin, TX, October 1999.

[10] J.-S. Kim et al. A 1.2V 12.8GB/s 2Gb Mobile Wide-I/O DRAM with 4x128 I/Os Using TSV-Based Stacking. In *ISSCC*, 2011.

[11] C. Lefurgy, K. Rajamani, F. Rawson, W. Felter, M. Kistler, and T. W. Keller. Energy Management for Commercial Servers. *Computer*, 36(12):39–48, 2003.

[12] Y. Li, B. Lee, D. Brooks, Z. Hu, and K. Skadron. CMP Design Space Exploration Subject to Physical Constraints. In *Proc. of the 12th Intl. Symp. on High Performance Computer Architecture*, Austin, TX, February 2006.

[13] J. Liu, B. Jaiyen, R. Veras, and O. Mutlu. RAIDR: Retention-Aware Intelligent DRAM Refresh. In *Proc. of the 39th Intl. Symp. on Computer Architecture*, Portland, OR, June 2012.

[14] G. H. Loh. 3D-Stacked Memory Architectures for Multi-Core Processors. In *ISCA-35*, 2008.

[15] J. Menon, L. D. Carli, V. Thiruvengadam, K. Sankaralingam, and C. Estan. Memory Processing Units. Poster at the 26th Hot Chips, Cupertino, CA Cupertino, CA.

[16] D. Milojevic, S. Idgunji, D. Jevdjic, E. Ozer, P. Lotfi-Kamran, A. Panteli, A. Prodromou, C. Nicopoulos, D. Hardy, B. Falsafi, and Y. Sazeides. Thermal Characterisation of Cloud Workloads on a Low-power Server-on-Chip. In *Proc. of the 30th Intl. Conf. on Computer Design*, pages 175–182, Montreal, Canada, September 2012.

[17] M. Oskin, F. T. Chong, and T. Sherwood. Active Pages: a Computation Model for Intelligent Memory. In *Proc. of the 25th Intl. Symp. on Computer Architecture*, pages 192–203, Barcelona, Spain, June 1998.

[18] J. T. Pawlowski. Hybrid Memory Cube: Breakthrough DRAM Performance with a Fundamentally Re-Architected DRAM Subsystem. In *Hot Chips 23*, 2011.

[19] S. Pugsley, J. Jestes, H. Zhang, R. Balasubramonian, V. Srinivasan, A. Buyuktosunoglu, A. Davis, and F. Li. NDC: Analyzing the Impact of 3D-Stacked Memory+Logic Devices on MapReduce Workloads. In *Proc. of the Intl. Symp. on Performance Analysis of Systems and Software*, Monterey, CA, March 2014.

[20] K. Puttaswamy and G. H. Loh. Thermal Herding: Microarchitecture Techniques for Controlling HotSpots in High-Performance 3D-Integrated Processors. In *Proc. of the 13th Intl. Symp. on High Performance Computer Architecture*, pages 193–204, Phoenix, AZ, February 2007.

[21] M. Scrbak, M. Islam, K. Kavi, M. Ignatowski, and N. Jayasena. Improving Node-level MapReduce Performance using Processing-in-Memory Technologies. The 7th Workshop on UnConventional High Performance Computing, August 2014.

[22] T. Singh, J. Bell, and S. Southard. Jaguar: A next-generation low-power x86-64 core. In *Proc. of the Intl. Solid-State Circuits Conference*, pages 52–53, San Francisco, CA, February 2013.

[23] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan. Temperature-Aware Microarchitecture. In *Proc. of the 30th Intl. Symp. on Computer Architecture*, pages 2–13, San Diego, CA, May 2003.

[24] The Intl. Technology Roadmap for Semiconductors. http://www.itrs.net. 2013.

[25] A. Udipi, N. Muralimanohar, N. Chatterjee, R. Balasubramonian, A. Davis, and N. P. Jouppi. Rethinking DRAM Design and and Organization for Energy-Constrained Multi-Cores. In *ISCA-37*, 2010.

[26] T. Vogelsang. Understanding the Energy Consumption of Dynamic Random Access Memories. In *Proc. of the 43rd Intl. Symp. on Microarchitecture*, pages 363–374, Atlanta, GA, December 2010.

[27] D. P. Zhang, N. Jayasena, A. Lyashevsky, J. Greathouse, L. Xu, and M. Ignatowski. TOP-PIM: Throughput-Oriented Programmable Processing in Memory. In *Proc. of the 23rd Intl. ACM Symp. on High Performance Parallel and Distributed Computing*, Vancouver, Canada, June 2014.