3D Systems with On-Chip DRAM for Enabling Low-Power High-Performance Computing

Jie Meng, Daniel Rossell, and Ayse K. Coskun
Electrical and Computer Engineering Department, Boston University, Boston, MA, USA
{jiemeng, drossell, acoskun}@bu.edu

1. INTRODUCTION

Communication between the last-level caches and main memory is one of the main performance bottlenecks of today's computer systems. 3D stacking enables integrating DRAM layers and processor cores on the same chip—improving memory bandwidth and reducing memory access time. We believe the performance increase achieved by 3D systems with on-chip DRAM can lead to high-performance system design without the need for adding more cores, using power-hungry architectures, or increasing the clock frequency. Higher performance, however, increases power densities and may create thermal challenges. Therefore, performance and temperature tradeoffs should be examined thoroughly to enable high-performance and reliable system operation.

Prior work has modeled performance in 3D systems with a small number of cores and single-threaded workloads [6]. Temperature management methods for 3D systems include static methods such as thermally-aware floorplanning [4], as well as run-time approaches such as task migration and dynamic voltage-frequency scaling [10]. Detailed performance analysis and thermal optimization for 3D systems have been mostly disjoint so far. This research provides a comprehensive approach to jointly evaluate performance, power, and temperature for 3D systems with on-chip DRAM. We use the evaluation framework to analyze 3D systems running parallel applications that represent future workloads. Our specific contributions are as follows.

- We present a model to compute the memory access latency of 3D systems with on-chip DRAM and use the model in architecture-level performance simulation. We integrate the performance simulation with power and thermal models to enable a thorough evaluation.
- We analyze the performance, power, and temperature of a 16-core 3D system with on-chip DRAM. We run the PARSEC parallel benchmarks [1], and show that instructions-per-cycle (IPC) for the applications is on average 72.6% higher in comparison to 2D systems. The performance improvement increases core power by up to 32.7%.
- We observe that the peak temperature of the 3D system increases by at most 1.5^oC with respect to the 2D peak temperature for a high-end system. For most benchmarks, there is a slight decrease in peak temperature, as the DRAM layers have low power density. However, for smaller or lower cost thermal packages as in embedded systems, we demonstrate the need for efficient thermal management strategies.

2. METHODOLOGY

Our target system is a 16-core processor. We model both the 2D baseline system and the 3D stack with on-chip DRAM, using the same architectural configurations for the cores and caches. We assume the system is manufactured at 45nm. The core architecture is based on the cores in the Intel 48-core single-chip cloud (SCC) [3]. The cores operate at 1GHz, 1.14V. Each core has a private 512KB L2 cache. The layout of the processor with on-chip DRAM is shown in Fig. 1.

To analyze the performance improvement of 3D architectures, we need an accurate model for memory latency. The two main components for main memory latency are the data request time spent at the memory controller and the data retrieving time spent at the DRAM-layers. Memory controller latency includes the time needed to translate physical addresses

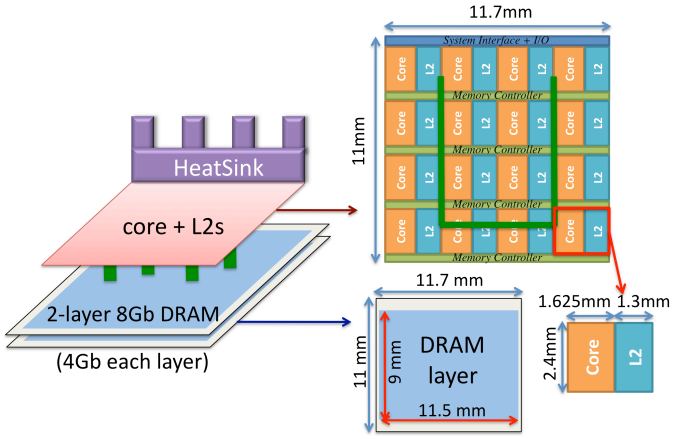

Figure 1: Layout of the 3D system with on-chip DRAM.

to memory addresses and the time for scheduling memory requests. The request scheduling time consists of time spent for converting memory transactions to command sequences and queuing time. We assume the memory controller address translation time is equal to the sum of memory controller processing time and controller-to-core delay.

DRAM access latency consists of address decoding time, column and row active time, and data transfer time. We consider a 1GB DRAM with two 4Gb layers, and set the row active time $t_{RAS} = 30ns$ and row precharge time $t_{RP} = 15ns$ [7]. We assume 1024 through-silicon-vias (TSVs), which provide a 128-Byte bus width with only 0.3% chip-area overhead. Table 1 summarizes the memory access times.

We use M5 [2] to build the performance simulation infrastructure, and run parallel applications from the PARSEC suite with sim-large input sets [1]. We model the 3D system with DRAM stacking in M5 by configuring the main memory access latency and the bus width between L2 caches and main memory to mimic the high data transfer bandwidth provided by the TSVs.

To compute run-time dynamic and leakage power of the cores, we utilize McPAT 0.7 [5]. L2 cache power is calculated using Cacti 5.3 [9]. For higher accuracy, we calibrate McPAT results to match reported average power values of Intel SCC cores. The DRAM power in the 3D system is calculated using MICRON's DRAM power calculator.

We use HotSpot 5.0 [8] for thermal simulations. Thicknesses of the logic and DRAM layers are $100\mu m$ and $20\mu m$, respectively. In addition to the default high-end thermal package, we simulate two lower-cost embedded system packages: Package A has the same convection resistance as the default (0.1 K/W) but there is no heat sink; Package B has a convection resistance of 1 K/W.

3. PERFORMANCE, POWER, AND TEMPERATURE ANALYSIS

Fig. 2 presents IPC improvements for the 3D system with DRAM stacking in comparison to the 2D-baseline. The average IPC increase across the benchmarks is 72.6%. `streamcluster`, `vips`, and `canneal` have higher IPC improvement, as these applications are highly memory-bound and benefit more from the reduction in memory access latency.

The core power increase for the 3D system with on-chip DRAM with respect to the 2D-baseline is presented in Fig.

Table 1: DRAM access latency

	2D-baseline design	3D with DRAM-stacking design
memory controller	4 cycles controller-to-core delay,116 cycles queuing delay 5 cycles memory controller processing time	4 cycles controller-to-core delay, 50 cycles queuing delay 5 cycles memory controller processing time
main memory	off-chip 1GB SDRAM $t_{RAS} = 40ns$, $t_{RP} = 15ns$, 10ns chipset request/return	on-chip 1GB SDRAM, 800MHz operating frequency $t_{RAS} = 30ns$, $t_{RP} = 15ns$ [7], no chipset delay
memory bus	off-chip memory bus, 200MHz, 8Byte bus width	on-chip memory bus, 2GHz, 128Byte bus width

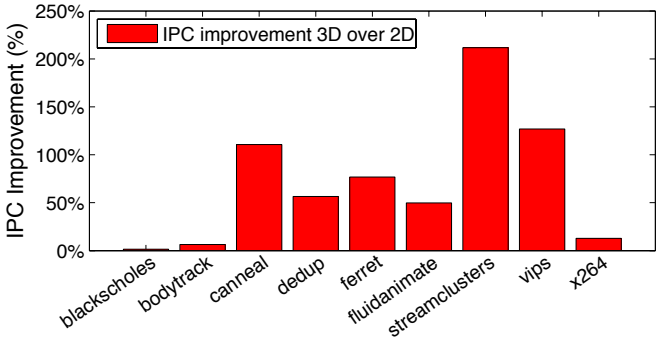

Figure 2: IPC improvement for the 3D system over the 2D baseline.

3. Power consumption increases by 16.6% on average for the 3D system across the benchmark set. `ferret` has the largest increase in core power, as it is already at a higher power range and has an additional 76.8% increase in IPC.

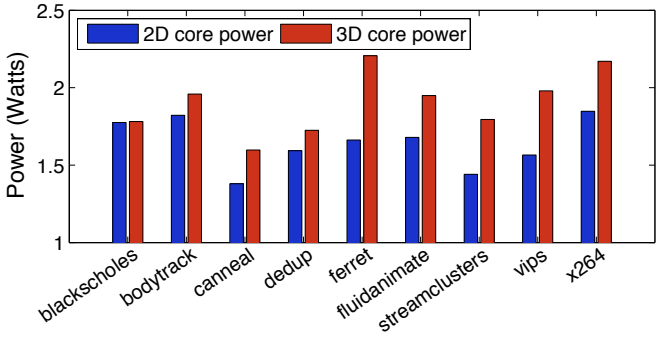

Figure 3: Average core power for the 2D and 3D systems.

For analyzing the power and temperature behavior of the DRAM layer, we present `dedup`'s per-layer DRAM power and temperature traces in Fig. 4. DRAM power varies following the changes in memory access rate. DRAM temperature changes due to the power variations as well as the core power variations on the adjacent layer.

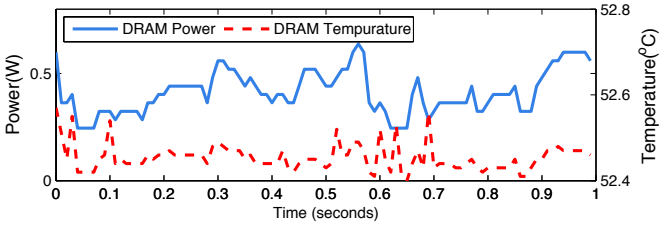

Figure 4: DRAM power and temperature traces (for one layer) for dedup running on the 3D system.

The peak core temperatures in the 2D and 3D systems with the high-performance package are shown in Fig. 5. DRAM stacking causes limited temperature rise in these cases. The maximum peak temperature increase is 1.52^oC for streamcluster. In fact, most of the benchmarks running on the 3D system have lower peak temperatures as the lower power DRAM layer shares the heat of the hotter cores.

The thermal behavior for 2D and 3D systems with the embedded packages are shown in Fig. 6. The peak temperature

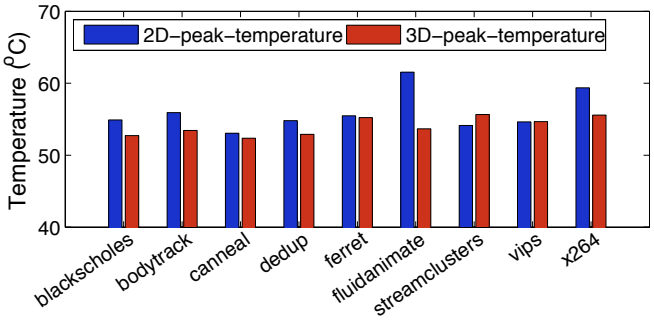

Figure 5: Peak core temperatures for 2D and 3D systems with the default HotSpot package.

increase is more visible. For example, `ferret`'s peak temperature increases by 3.31^oC and 8.04^oC for the two embedded system packages in comparison to the peak on the 2D system. Efficient thermal management or low-power design is needed to ensure reliable operation for 3D systems with lower cost or smaller packages.

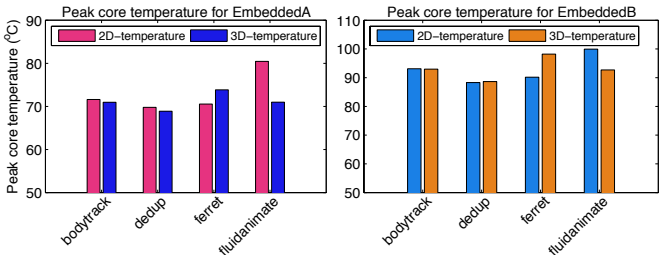

Figure 6: Peak core temperatures for 2D and 3D systems with the embedded packages.

4. CONCLUSION

3D systems with on-chip DRAM offer a promising solution for designing high-performance systems without introducing power-hungry architectures or additional cores. We have shown significant IPC increases for parallel workloads running on 3D systems. The thermal impact of the performance increase has been minimal for high-end packages. For embedded systems with smaller or lower cost packages, efficient thermal management strategies are needed to ensure reliable operation.

5. REFERENCES

[1] Bienia, C. *Benchmarking Modern Multiprocessors*. PhD thesis, Princeton University, January 2011.

[2] Binkert, N., et al. The M5 simulator: Modeling networked systems. *IEEE Micro 26*, 4 (july-aug. 2006), 52 –60.

[3] Howard, J., et al. A 48-core IA-32 message-passing processor with DVFS in 45nm CMOS. In *ISSCC* (2010), pp. 108 –109.

[4] Hung, W.-L., et al. Interconnect and thermal-aware floorplanning for 3D microprocessors. In *ISQED* (2006).

[5] Li, S., et al. McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures. In *MICRO-42* (2009), pp. 469 –480.

[6] Loh, G. 3D-stacked memory architectures for multi-core processors. In *ISCA* (2008), pp. 453 – 464.

[7] Micron. 1GB DDR SDRAM datasheet.

[8] Skadron, K., et al. Temperature-aware microarchitecture. In *ISCA* (2003).

[9] Thoziyoor, S., Muralimanohar, N., Ahn, J. H., and Jouppi, N. P. CACTI 5.1. Tech. rep., HP Laboratories, Palo Alto, April 2008.

[10] Zhu, C., et al. Three-dimensional chip-multiprocessor run-time thermal management. *IEEE Trans. on CAD 27*, 8 (August 2008), 1479 – 1492.