Analysis and Runtime Management of 3D Systems with Stacked DRAM for Boosting Energy Efficiency

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

***Abstract*—3D stacked systems with on-chip DRAM provide high speed and wide bandwidth for accessing main memory, overcoming the limitations of slow off-chip buses. Power densities and temperatures on the chip, however, increase following the performance improvement. The complex interplay between performance, energy, and temperature on 3D systems with on-chip DRAM can only be addressed using a comprehensive evaluation framework. This paper first presents such a framework for 3D multicore systems capable of running architecture-level performance simulations along with energy and thermal evaluations, including a detailed analysis of the DRAM layers. Experimental results on 16-core 3D systems running parallel applications demonstrate up to 88.5% improvement in energy delay product compared to equivalent 2D systems. We also present a memory management policy that targets applications with spatial variations in DRAM accesses and performs temperature-aware mapping of memory accesses to DRAM banks.**

I. Introduction

3D stacking has emerged as an attractive design technique that improves manufacturing yield, transistor density per chip footprint, and performance. One of the prominent advantages of 3D stacking is the ability to integrate heterogeneous technologies within the same chip, such as stacking memory layers with the processors. Designing 3D systems with on-chip DRAM is a promising solution to improve memory bandwidth and reduce memory access latency [1, 2]. Reducing the memory access overhead is especially beneficial for multicore systems, where long off-chip memory access latency has been a gating performance bottleneck.

3D systems with on-chip DRAM have the potential to increase the performance significantly; however, power densities and temperatures also increase following the performance improvement. In fact, high temperatures already bring major challenges because of their adverse effects on cooling costs and reliability. Existing temperature management methods for 3D systems include thermally-aware floorplanning, temperature-aware job allocation, and dynamic voltage and frequency scaling [3, 4, 5, 6]. So far, thermal management for 3D systems has been mostly disjoint from detailed performance evaluation. For example, recently proposed management policies for 3D systems use worst-case performance estimates without providing an architecture-level evaluation [7]. Performance evaluation for 3D systems, on the other hand, has mainly focused on a small number of cores (e.g., single-core, quad-core) running single-threaded workloads [2, 8, 9].

978-3-9810801-8-6/DATE12/©2012 EDAA

Another challenge in 3D systems with on-chip DRAM is that power and temperature of the DRAM layers can substantially increase because of the high memory access rate and the heat transfer from the logic layer. High DRAM temperature severely affects memory reliability and performance [10, 11]. Several research groups have examined 3D DRAM organization or access patterns [2, 10]. However, these techniques do not evaluate DRAM power and temperature connected with a detailed performance simulation of the multicore logic layer.

This paper's focus is analyzing the performance, energy, and temperature tradeoffs in 3D multicore systems with on-chip DRAM. We believe this is an essential step for optimizing the energy efficiency and reliability of future multicore 3D systems, and for understanding the benefits and limitations of 3D memory stacking. To the best of our knowledge, our work is the first to quantify the performance and energy benefits of 3D systems with stacked DRAM through a joint architecture-level performance, power, and temperature evaluation. We analyze several 3D multicore systems in comparison to equivalent 2D systems. Our specific contributions are as follows:

- We provide a simulation framework with joint performance, power, and thermal models for 3D systems with on-chip DRAM. Using the framework, we evaluate two 16-core 3D systems: a `high-performance` system and a `low-power` system. We run the parallel benchmarks in the PARSEC suite [12], and show that instructions per second (IPS) is improved by 109.7% in the `high-performance` system and 52.6% in the `low-power` system on average across all the benchmarks in comparison to the 2D baselines. The performance improvement causes an increase in core power by 29.98% on average in the `high-performance` system.
- We construct a detailed performance and thermal model for the on-chip DRAM in our simulation framework, and demonstrate the impact of high vertical bus widths and parallel access mechanisms enabled by the memory stacking in 3D multicore systems.
- We propose a memory management policy targeting memory-intensive applications that have spatial variations in memory access rates across different on-chip DRAM banks in 3D multicore systems. Our policy performs temperature-aware mapping of the accesses to DRAM banks to reduce the peak temperatures and thermal variations.

The rest of the paper starts with an overview of the related work. Section III introduces our simulation framework for

3D systems with on-chip DRAM. Section IV analyzes two 3D multicore systems using our framework and discusses our memory management policy. Section V concludes the paper.

II. Related Work

Most of the prior work on 3D systems with memory stacking considers performance, power, and thermal evaluations separately, focusing on the systems with a small number of cores or single-threaded workloads. For example, Liu et al. report that a single-core processor with 3D memory stacking increases system performance by 126%; however, their work does not consider the power or thermal impact [8]. Loh explores 3D-stacked memory architectures for 4-core processors [2] and performs thermal analysis using HotSpot [13]. Their thermal simulations use estimated power values that are not tied with detailed architecture-level performance evaluations.

Performance and power modeling is critical in 3D systems with stacked DRAM. Sun et al. study the architecture-level design of 3D stacked L2 caches [14]. Wu et al. use power density analysis and power delivery consideration in their 3D cost model [15]. However, they do not evaluate the power consumption of the memory components on 3D chips.

Several static thermal management techniques have been proposed for controlling temperature on 3D systems. Hung et al. present a thermally-aware floorplanner for 3D architectures [3]. Cong et al. propose transformation techniques for 3D IC placement [4]. These static optimization methods are implemented at design time without considering detailed runtime workload profiles. Dynamic thermal management methods for 3D systems include workload scheduling and dynamic voltage-frequency scaling (DVFS) methods (e.g., [16]). Zhu et al. propose runtime task migration and DVFS policies that utilize offline workload profiling [5]. However, these dynamic management methods do not target the on-chip DRAM layers.

Our research differentiates from prior work as we provide a detailed architecture-level performance, power, and thermal evaluation for the 3D systems with on-chip DRAM. We focus on future multicore architectures running parallel applications. Our simulation framework includes on-chip DRAM performance and power models based on memory access patterns collected at runtime. We quantify performance and energy efficiency improvements achieved by several 3D multicore systems. In addition, we provide a memory address management policy for reducing and balancing the DRAM temperature.

III. Methodology

In this section, we describe the target systems and introduce our simulation infrastructure for quantifying the performance, power, and temperature of 3D systems with on-chip DRAM.

A. Target Systems

Target systems in this work are multicore processors with stacked on-chip DRAM. Figure 1 provides an illustration of a 16-core 3D system with on-chip DRAM, where the processing cores and caches are on one layer and a 2-layer 3D DRAM is stacked below the logic layer. Through-silicon vias (TSVs) are used for vertically connecting the layers.

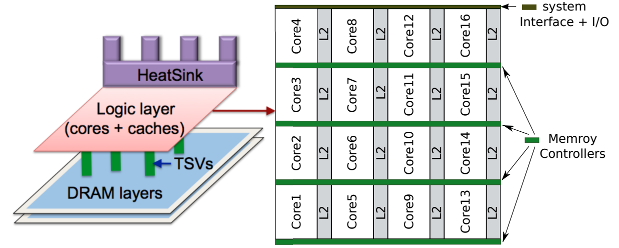

Fig. 1: The illustration of a generic 3D 16-core processor with on-chip DRAM stacking.

We experiment with two 16-core processors: a `high-performance` system and a `low-power` system. The core architecture for the `low-power` system is similar to the cores in the Intel 48-core Single-chip Cloud Computer (SCC) [17]. The `high-performance` system includes more aggressive core architectures, which are modeled based on the AMD Family 10h microprocessors in AMD Magny-Cours chips. We simulate both the 2D baselines (single-layer, off-chip memory) and 3D systems with on-chip DRAM for the two target architectures.

The architectural parameters for the cores and caches are listed in Table I. For each processor, we use the same architectural configuration for the 2D baseline and the 3D system. Each core has multiple-issue and out-of-order execution. We assume that both processors are manufactured at 45nm and have a supply voltage of 1.1V. The `low-power` system has a total die area of 128.7mm^2 and operates at 1 GHz, while the `high-performance` system has a total die area of 376mm^2 and operates at 2.1 GHz.

Each core has a private L2 cache. All the L2 caches are located on the same layer as the cores and connected by a shared bus. MESI protocol is used for cache coherence. Both the 2D and 3D systems have on-chip memory controllers. The dimensions for the components of the 16-core processors are listed in Table II. We assume face-to-back, wafer-to-wafer bonding for building the 3D systems. Wafer-to-wafer bonding allows for reliably manufacturing larger 3D systems approaching $20mm \times 20mm$ with the current technology.

B. Modeling DRAM Accesses

To analyze the performance characteristics of 3D architectures, we need an accurate DRAM access latency model. For the `low-power` system, we assume a two-layer DRAM, where each layer has two ranks. A single-layer DRAM is

TABLE I: Core architecture parameters.

Parameter	`High-performance`	`Low-power`
CPU Clock	2.1GHz	1.0 GHz
Issue	out-of-order	out-of-order
Decode Width	3-way	2-way
Reorder Buffer	84 entries	40 entries
BTB size	2048 entries	512 entries
RAS size	24 entries	16 entries
Integer/FP ALU	3/3	2/1
Load/Store Queue	32/32 entries	16/12 entries
L1 ICache	64KB@2ns	16KB@2ns
L1 DCache	2-way/64B-block	2-way/64B-block
L2 Cache	512KB@6ns	512KB@5ns
L2 Cache	16-way/64B-block	4-way/64B-block

TABLE III: DRAM access latency

	2D-baseline design	3D system with on-chip DRAM
memory controller	4ns controller-to-core delay, 48ns queuing delay	4ns controller-to-core delay, 24ns queuing delay
DRAM access	$t_{RAS} = 36ns$, $t_{RP} = 15ns$	$t_{RAS} = 36ns$, $t_{RP} = 15ns$
total access time	*103ns* for off-chip 1GB SDRAM	*79ns* for on-chip 1GB SDRAM
memory bus	off-chip memory bus, 200MHz, 8-Byte bus width	on-chip memory bus, 2GHz, 64/128-Byte bus width

stacked with the logic layer in the `high-performance` system, which consists of four identical ranks. The total DRAM capacity in each system is 1GB. Each rank has four internal banks allowing low-order interleaved accesses.

As shown in Table III, the main memory access latency consists of memory controller processing time, DRAM access latency, and time spent communicating through the bus. We assume a 4ns round-trip memory controller to core delay by using 183ps/mm wire propagation delay for 45nm technology [18] and estimating average core to memory controller distance as 10mm. Memory controller processing time is dominated by the memory request queuing delay, which is estimated as 100 cycles at 2.1GHz for 2D systems [19]. In 3D systems with stacked DRAM, faster data transfer time enables reduction in request queuing delay. We assume 50% lower delay in comparison to the 2D baseline based on prior work's analysis [8]. We assume that both `high-performance` and `low-power` systems have the same delay (in ns) as memory access time is not strongly dependent on core frequency.

DRAM access latency is mainly composed of precharge time (t_{RP}), row active time (t_{RAS}), and data transfer time. We obtain t_{RP} and t_{RAS} from Micron's datasheet [20] and assume they are the same for both 2D and 3D systems as modeled in prior work [2]. To simulate data transfer between the memory controller and on-chip DRAM, we assume 512 TSVs that provide a 64-Byte bus at 2GHz, while data transfer in 2D design is limited by an 8-Byte off-chip bus operated at 200MHz. TSVs have $10\mu m \times 10\mu m$ dimensions and a $10\mu m$ pitch. The total area devoted to TSVs is less than 0.2% of the chip area for both the `high-performance` and `low-power` systems. For the two-layer DRAM in the `low-power` system, we do not distinguish between the access times to different layers, as the additional vertical distance to research the second layer is very short (i.e., around $70\mu m$). Figure 2 demonstrates the layout for the single-layer DRAM.

We also model a parallel memory access scenario using a 128-Byte wide memory bus enabling two different cores to access the on-chip DRAM at the same time. We model the system to fully utilize the 128-Byte memory bus by reducing the memory latency by half in our performance simulation.

TABLE II: Dimensions of the components in the 3D `high-performance` and `low-power` systems.

(all values in mm except TSVs)	`High-perf.`		`Low-power`	
	Length	Width	Length	Width
Chip	20	18.8	11.7	11
Core	4.5	3.5	2.4	1.625
L2 Cache	4.5	1.2	2.4	1.3
DRAM	20	18.8	11.5*	9*
TSVs	diameter $10\mu m$, pitch $10\mu m$			

* *This system includes 2 DRAM layers*

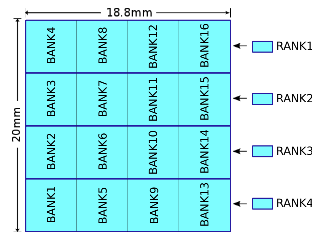

Fig. 2: The layout of the single-layer on-chip DRAM in the `high-performance` 3D system with stacked DRAM.

C. Performance Simulation

We use the M5 simulator [21] to build the performance simulation infrastructure for our target systems. We use the Alpha instruction set architecture (ISA) as it is the most stable ISA currently supported in M5. The full-system mode in M5 models a DEC Tsunami system to boot an unmodified Linux 2.6 operating system. We run the PARSEC parallel benchmark suite [12], representing future multicore workloads.

M5 models a split-transaction bus that is configurable in both latency and bandwidth. We model the 3D systems with on-chip DRAM in M5 by configuring the main memory access latency and the bus width/speed between L2 caches and main memory. The simulator, thus, mimics the high data transfer bandwidth provided by the TSVs. The core and cache architectures are outlined in Table I. The bus and memory access delay configurations are based on our analysis summarized in Table III.

We run the PARSEC benchmarks in M5 with sim-large input sets and collect the performance statistics at regular intervals. We implement thread-binding in M5 for the PARSEC benchmarks to control thread allocation. Each thread is bound on a specific core during the interval. We fast-forward the M5 simulation to the region of interest (ROI) and execute each PARSEC benchmark with the detailed out-of-order CPUs for 1 second (100 time steps, collecting statistics at 10ms intervals). In addition to the performance statistics provided by M5, we track the number of memory accesses to each DRAM bank at every interval by observing the least significant bits for the physical memory addresses.

D. Power Model

We use McPAT 0.7 [22] for 45nm technology to obtain the run-time dynamic power of the cores. McPAT computes the power consumption by taking the system configuration parameters and M5 performance statistics as inputs.

To improve the accuracy of run-time power computations, we calibrate McPAT's run-time dynamic power values for the cores to match the published or measured power data for the target core architectures. We derive the average dynamic core power values from McPAT across the benchmark suite, and

compute the calibration factor, R, to translate the McPAT raw data to the target power scale. Then, we use R to scale each benchmark's dynamic core power consumption. A similar calibration approach has been introduced in prior work [23]. At nominal temperature, we assume the leakage power for the cores is 35% of the total core power.

L2 cache power is calculated using CACTI 5.3 [24]. The dynamic power obtained from CACTI is scaled using the L2 cache access rates collected from M5. For the on-chip memory controllers in both `high-performance` and `low-power` systems, we estimate the power consumption as 5.9W based on the memory controller power reported for Intel SCC [17]. The system interface and I/O power as well as the on-chip bus power are negligible with respect to the total chip power.

The DRAM power in the 3D system is calculated using MICRON's DRAM power calculator [25], which takes the memory read and write access rates as inputs. We obtain detailed DRAM power traces for each of the DRAM banks sampled every 10ms interval.

E. Thermal Model

For thermal simulations, we use HotSpot 5.0 [13], which includes basic 3D modeling features. We use a sampling interval of 10ms. We configure the on-chip DRAM's thickness as 0.05mm and thermal conductivity as 100W/mK. We set the other chip and package parameters using the default configuration in HotSpot to represent efficient packages in high-end systems. Calibrated power traces are the inputs for the thermal model. All simulations use the HotSpot grid model for higher accuracy and are initialized with the steady-state temperatures. The parameters in HotSpot simulations for 2D and 3D architectures are listed in Table IV. We do not explicitly model the thermal impact of the TSVs, considering TSVs occupy less than 0.2% of the chip area. As previously demonstrated, low TSV densities have limited impact on temperature [16].

IV. Analysis of Performance, Power, and Temperature of 3D Systems with Stacked DRAM

In this section, we evaluate the performance, power, and thermal results for the 16-core `high-performance` system and `low-power` system running the PARSEC parallel benchmarks. We also introduce a memory address management policy for reducing and balancing the DRAM temperature.

TABLE IV: Thermal simulation configuration in HotSpot.

Parameter	Value
Chip thickness	0.1mm
Silicon thermal conductivity	100 W/mK
Silicon specific heat	1750 kJ/m^{3}K
Sampling interval	0.01s
DRAM thickness	0.05mm
DRAM thermal conductivity	100 W/mK
Interface material thickness	0.02mm
Interface material conductivity	4 W/mK
Spreader thickness	1mm
Spreader thermal conductivity	400 W/mK
Heat sink thickness	6.9mm
Heat sink resistance	0.1K/W

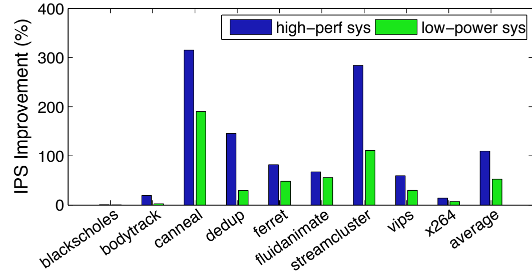

Fig. 3: Percentage of IPS improvements for the 3D systems with stacked DRAM compared to the 2D baselines.

A. Performance Evaluation

Figure 3 compares the performance of the 3D systems with on-chip DRAM against the 2D baselines. We use instructions retired per second (IPS) as our performance metric. Using 3D DRAM stacking, we achieve an average IPS improvement of 109.7% for the `high-performance` system and 52.6% for the `low-power` system across the 9 PARSEC benchmarks compared to the equivalent 2D baselines with off-chip memory. The `high-performance` system has larger IPS improvements compared to the `low-power` system because of its more advanced core architecture, which provides better instruction-level parallelism. Among all the benchmarks, `streamcluster` and `canneal` achieve higher IPS improvements (over 100%) in both 3D systems, as they are highly memory-bound and therefore benefit more significantly from the reduction in memory access latency. On the other hand, CPU-bound benchmarks, such as `blackscholes` and `x264` have very limited performance improvement.

We select two PARSEC benchmarks, `streamcluster` and `fluidanimate`, to show the temporal performance trends. In Figure 4, we observe that for both the 2D and 3D systems the IPS of `streamcluster` is stable during the simulation, while the IPS of `fluidanimate` changes periodically. These trends are similar for both the

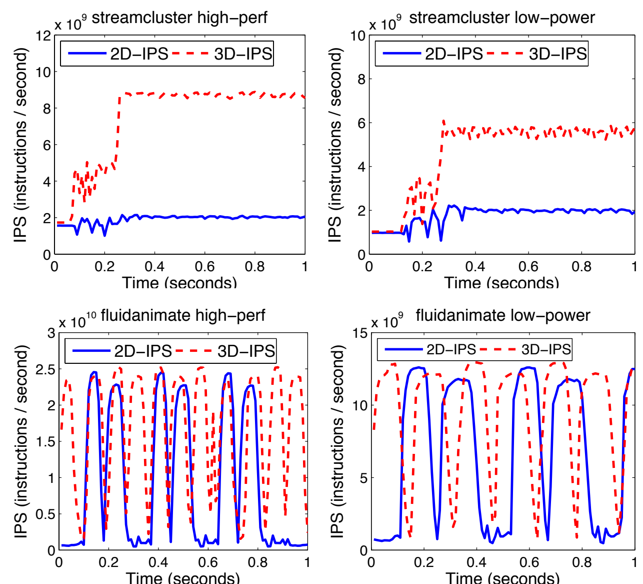

Fig. 4: Temporal IPS changes for 2D and 3D systems for `streamcluster` and `fluidanimate`.

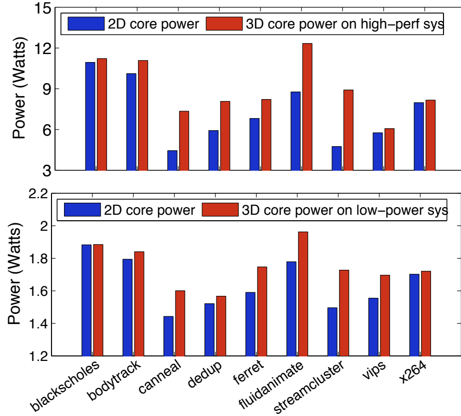

Fig. 5: Average core power for the 3D systems with on-chip DRAM and the 2D systems for the `high-performance` system (*top*) and the `low-power` system (*bottom*).

`high-performance` system and the `low-power` system. Also, `streamcluster` has 284% higher IPS on average in the `high-performance` system while `fluidanimate` has 67.3% IPS improvement compared to the 2D baseline. This is because `streamcluster` has a significantly higher number of DRAM accesses compared to `fluidanimate`.

The significant performance improvement for benchmarks, such as for `streamcluster`, suggests corresponding increases in core power. In addition, temporal changes of IPS for some benchmarks, such as `fluidanimate`, demonstrate that using average power/temperature or coarse-grained performance estimates in analysis of 3D systems cannot capture the runtime trends accurately. Dynamically changing workload characteristics can only be observed by detailed architecture-level performance, energy, and thermal evaluations and periodic sampling of runtime events, which are integrated in our simulation approach.

B. Power Evaluation

Figure 5 demonstrates the average core power increases for the 3D systems with stacked DRAM compared to the 2D baselines. Power consumption per core increases by 29.98% and 6.9% on average for the `high-performance` and the `low-power` systems, respectively. `canneal` and `streamcluster` have the largest increases in core power, as they have the highest performance improvements. The core power of `fluidanimate` also increases considerably, as it is already at a high power range and the IPS of `fluidanimate` has additional 67.3% increase in 3D `high-performance` system. Our results demonstrate an average energy delay product (EDP) improvement of 51.3% for the `high-performance` system and 37.9% for the `low-power` system compared to their equivalent 2D baselines. `canneal` running on `high-performance` system has 88.5% EDP reduction, which is the largest energy efficiency improvement across the benchmarks.

C. Temperature Analysis

We illustrate the thermal behavior for 3D systems in Figure 6. We select four benchmarks from PARSEC benchmark

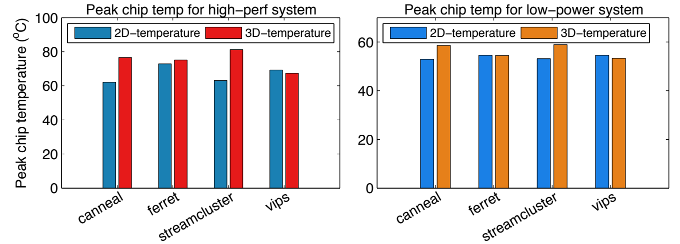

Fig. 6: Peak chip temperatures for the 2D baselines and the 3D systems with stacked DRAM.

sets (`canneal`, `ferret`, `streamcluster`, and `vips`). Their peak chip temperatures on the 2D and 3D systems for both the `high-performance` system and `low-power` system are shown in the figure. The maximum peak temperature increase is 18.1^oC for running `streamcluster` in `high-performance` system and 5.8^oC in `low-power` system. This is because `streamcluster` has the highest DRAM access rate across all the benchmarks. We notice that `vips` running on our target 3D systems obtain a peak temperature decrease. This is a result of the relatively low memory access rates of `vips`. Low frequency of memory accesses results in low DRAM power, which already has lower power density compared to the logic layer. The lower power DRAM layer shares the heat of the hotter cores, decreasing the adjacent logic layer temperature for benchmarks with low frequency of memory accesses. These results suggest that more aggressive architectures or wider/faster memory access links can be leveraged to boost energy efficiency further within safe thermal operating points.

D. Evaluating the Impact of DRAM Access Bandwidth

In addition to simulating 3D systems with a 64-Byte DRAM bus, we evaluate a higher bandwidth scenario where a 128-Byte link connects the logic and DRAM layers. The 3D system utilizes the increased bandwidth to allow parallel memory accesses to two DRAM banks simultaneously. The DRAM temperature profile for the `high-performance` 3D system with 128-Byte and 64-Byte bandwidths are shown in Figure 7. We notice that with higher bus-width, the temperature of the 3D systems reaches 89^oC for memory-bound benchmarks such as `streamcluster`. The thermal variation across the DRAM layer increases to 6.5^oC.

As DRAM performance is severely affected from high temperatures due to the impact of temperature on refresh rates, we implement a memory address management policy

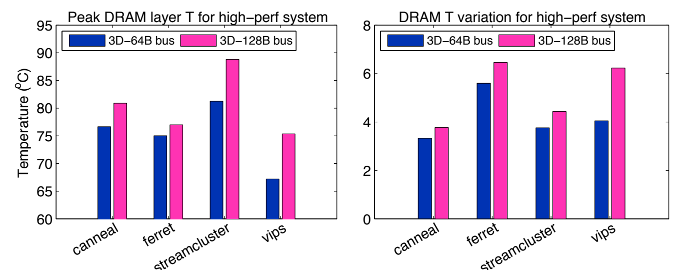

Fig. 7: DRAM peak temperatures and thermal variations in the 3D `high-performance` system.

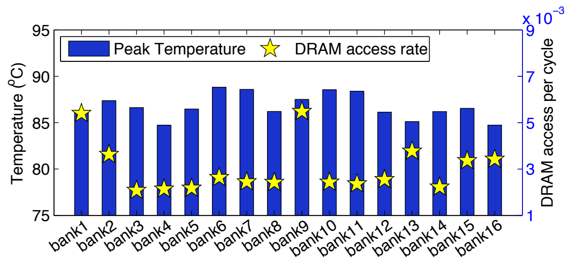

Fig. 8: DRAM bank temperature and access rate for `streamcluster` in 3D `high-performance` system with 128-Byte memory bus.

for temperature reduction on the DRAM layer. Our policy targets memory-intensive applications with high spatial variations in their access rates across different DRAM banks. Figure 8 illustrates the peak temperatures and the number of accesses per cycle across the 16 DRAM banks while running `streamcluster` on the 3D `high-performance` system with 128-Byte memory bus. The location of each bank is shown in Figure 2. Banks 6, 7, 10, 11, which are located on the center of the DRAM layer have higher temperatures than banks 1, 4, 13, 16, which are on the corners. The variations in DRAM bank access rates indicate differences in power consumption across the DRAM banks. In Figure 8, the most accessed DRAM bank 9 and least accessed bank 3 have average power consumption of 5.1W and 1.9W, respectively.

Based on this analysis, our memory management policy maps more frequently accessed memory address ranges, such as the address range for bank 9 in the default mapping, to physical banks with lower temperatures (e.g., bank 1). The memory address mapping is implemented by the OS when virtual memory addresses are translated into physical addresses. The specific memory mapping strategy matching the virtual memory address ranges to physical locations can be determined based on average case analysis statically. This approach has no additional cost compared to existing memory mapping mechanisms. The mapping policy can also be updated if average case workload dynamics change significantly.

Simulation results show that our policy reduces DRAM peak temperature by 1.42^oC and the thermal variations by 1.6^oC for `streamcluster` running on the 3D `high-performance` system with 128-Byte memory bus in comparison to the worst-case allocation, where the banks receiving higher number of accesses are located in the center of the DRAM layer. Note that our memory address mapping policy would reduce temperature further for 3D systems with larger variations in core power (e.g., when there are idle cores).

V. Conclusion

In this paper, we have proposed an integrated simulation framework for detailed analysis of performance, power, and temperature of 3D systems with on-chip DRAM. We have quantified the benefits and challenges for two 16-core 3D systems. Our results show remarkable improvements in energy efficiency: a `high performance` 3D system has up to 88.5% lower energy delay product in comparison to an equivalent 2D system with off-chip memory. We have also introduced a memory management policy for memory-intensive applications with variations in DRAM bank accesses.

Acknowledgements

We thank Dr. E. Kursun and T. Brunschwiler from IBM for their valuable feedback. This work has been in part funded by the Design Automation Conference A. Richard Newton Scholarship.

References

[1] B. Black *et al.*, "Die stacking (3D) microarchitecture," in *International Symposium on Microarchitecture (MICRO)*, 2006, pp. 469 – 479.
[2] G. Loh, "3D-stacked memory architectures for multi-core processors," in *International Symposium on Computer Architecture (ISCA)*, 2008, pp. 453 – 464.
[3] W.-L. Hung *et al.*, "Interconnect and thermal-aware floorplanning for 3D microprocessors," in *International Symposium on Quality Electronic Design (ISQED)*, March 2006.
[4] J. Cong *et al.*, "Thermal-aware 3D IC placement via transformation," in *Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2007, pp. 780 – 785.
[5] C. Zhu *et al.*, "Three-dimensional chip-multiprocessor run-time thermal management," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, pp. 1479 – 1492, August 2008.
[6] A. K. Coskun, T. S. Rosing, J. Ayala, and D. Atienza, "Modeling and dynamic management of 3D multicore systems with liquid cooling," in *International Conference on Very Large Scale Integration (VLSI-SoC)*, 2009, pp. 60 – 65.
[7] A. K. Coskun, D. Atienza, T. S. Rosing, T. Brunschwiler, and B. Michel, "Energy-efficient variable-flow liquid cooling in 3D stacked architectures," in *Design, Automation, and Test in Europe (DATE)*, 2010, pp. 111 – 116.
[8] C. Liu *et al.*, "Bridging the processor-memory performance gap with 3D IC technology," *IEEE Design Test of Computers*, pp. 556 – 564, 2005.
[9] G. Loi *et al.*, "A thermally-aware performance analysis of vertically integrated (3-D) processor-memory hierarchy," in *ACM/IEEE Design Automation Conference (DAC)*, July 2006, pp. 991 – 996.
[10] M. Ghosh and H.-H. Lee, "Smart refresh: An enhanced memory controller design for reducing energy in conventional and 3D die-stacked DRAMs," in *MICRO'07*, 2007, pp. 134–145.
[11] S. Liu *et al.*, "Hardware/software techniques for dram thermal management." in *HPCA*, 2011, pp. 515–525.
[12] C. Bienia, "Benchmarking modern multiprocessors," Ph.D. dissertation, Princeton University, January 2011.
[13] K. Skadron, M. R. Stan, W. Huang, V. Sivakumar, S. Karthik, and D. Tarjan, "Temperature-aware microarchitecture," in *ISCA'03*, 2003.
[14] G. Sun, X. Wu, and Y. Xie, "Exploration of 3D stacked L2 cache design for high performance and efficient thermal control," in *International symposium on Low power electronics and design*, 2009, pp. 295 – 298.
[15] X. Wu *et al.*, "Cost-driven 3D integration with interconnect layers," in *ACM/IEEE Design Automation Conference (DAC)*, 2010, pp. 150 – 155.
[16] A. K. Coskun, T. S. Rosing, J. Ayala, D. Atienza, and Y. Leblebici, "Dynamic thermal management in 3D multicore architectures," in *Design Automation and Test in Europe (DATE)*, 2009, pp. 1410 –1415.
[17] J. Howard *et al.*, "A 48-core IA-32 message-passing processor with DVFS in 45nm CMOS," in *International Solid-State Circuits Conference (ISSCC)*, 2010.
[18] Y. Jin *et al.*, "Adaptive data compression for high-performance low-power on-chip networks," in *MICRO'08*, pp. 354 – 363.
[19] M. Awasthi *et al.*, "Handling the problems and opportunities posed by multiple on-chip memory controllers," in *PACT'10*, pp. 319 – 330.
[20] Micron Technology, Inc. DRAM component datasheet. [Online]. Available: http://www.micron.com
[21] N. Binkert *et al.*, "The M5 simulator: Modeling networked systems," *IEEE Micro*, vol. 26, no. 4, pp. 52 –60, July 2006.
[22] S. Li *et al.*, "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures," in *MICRO'09*, 2009, pp. 469 –480.
[23] R. Kumar *et al.*, "Single-ISA heterogeneous multi-core architectures: the potential for processor power reduction," in *MICRO'03*, pp. 81 – 92.
[24] S. Thoziyoor *et al.*, "CACTI 5.1," HP Laboratories, Tech. Rep., 2008.
[25] Micron Technology, Inc. DRAM power calculations. [Online]. Available: http://www.micron.com