Proceedings of the ASME 2015 International Technical Conference and Exhibition on Packaging and Intergration of Electronic and Photonic Microsystems and ASME 2015 13th International Conference on Nanochannels, Microchannels and MiniChannels InterPACKICNMM2015 July 6-9, 2015, San Francisco, USA

InterPACKICNMM2015-48533

MULTI-PHYSICS DRIVEN CO-DESIGN OF 3D MULTICORE ARCHITECTURES

He Xiao*
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA, 30332

Wen Yueh
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA, 30332

Saibal Mukhopadhyay
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA, 30332

Sudhakar Yalamanchili
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA, 30332

ABSTRACT

The high heat flux and strong thermal coupling in the 3D ICs has limited the performance gains that would otherwise be feasible in 3D structures. The common practice of adopting worst-case design margins is in part responsible for this limitation since average-case performance would be limited by worst-case thermal design margins. The coupling between temperature and leakage power exacerbates this effect. However, worst-case thermal conditions are not the common state across the package at runtime. We argue for the co-design of the package, architecture, and power management based on the multi-physics interactions between temperature, power consumption and system performance. This approach suggests an adaptive architecture that accommodates the thermal coupling between layers and leads to increased energy efficiency over a wider operating voltage range and therefore higher performance.

In this paper, we target at a 3D multicore architecture where the cores reside on one die and the last level cache (LLC) resides on the other. The DRAM stack may be stacked on top of the package (e.g., 3D) or in the same package (e.g., 2.5D). We propose a novel adaptive cache structure - the constant performance model (CPM) cache - based on voltage adaptations to temperature variations. We construct a HSPICE model for the SRAM to explore the relationship between temperature, supply voltage, and the circuit delay in the context of the LLC. This model is used to investigate, characterize, and analyze the effect of the temperature-delay dependence of the SRAM LLC configuration on the system-level performance and energy efficiency. This analysis gives rise to an intelligent scheme for dynamic voltage regulation in the LLC cache that is sensitive to the temperature of the individual cache banks. Each cache bank is thermally coupled to the associated cores and thus is sensitive to the local core-level power management. We show that this local adaptation to the temperature-delay dependence leads to a significant power reduction in the LLC cache, and improvement of system energy efficiency computed as energy per instruction (EPI). We evaluate our approach using a cycle-level, full system simulation model of a 16-core x86 homogenous microarchitecture in 16nm technology that boots a full Linux operating system and executes application binaries. The advantages of the proposed adaptive LLC structure illustrate the potential of the co-design of the package, architecture, and power management in future 3D multicore architectures.

*Corresponding author; hxiao@gatech.edu

INTRODUCTION

The emerging technology of 3D stacked ICs brings about many advantages. Compared to the traditional 2D packaging technology, the 3D stacked IC realizes much higher integration density by stacking multiple dies. It provides a larger communication bandwidth between tiers [1] and reduces the average wire length and hence latency [2], enabling higher speed operation and therefore higher performance. However, as the integration level continues to increase, we encounter challenges in heat dissipation [3] within the stack. The operation of electrical circuits has a strong dependency on the temperature and the thermal coupling between layers exacerbates this dependence [4]. The common practice in the design of digital systems is to design for worst case conditions. The thermal challenges of 3D ICs expose significant performance penalties for such an approach since worst case conditions may not occur often on practice. Thermal behaviors are in fact driven by applications whose behaviors are time-varying.

In this paper we advocate an approach to converting the thermal headroom made available due to worst case design, to improve energy efficient operation. This requires understanding and controlling coupled interactions between workload behaviors, microarchitecture power management, circuit adaptation techniques, and choices of packaging. We demonstrate this approach on the operation of a 2-tier 3D structure comprised of a 16-core homogenous x86 processor die coupled with a shared LLC cache die. We first characterize the temperature-delay dependency of the SRAM in 16nm technology to understand the voltage margin available at each temperature relative to worst case design. Employing worst case design margins will fix the SRAM access delay corresponding to the worst case temperature. While maintaining this worst case delay, at lower temperatures we can lower voltage to maintain the performance (delay) but reduce energy consumption and thereby improve energy efficiency. This adaptation is driven by time-varying workload behaviors.

The state of the practice is to maximize performance for a given thermal budget under the context of the 3D IC structure compared to convention approaches [5] [6], which emphases more on the system performance over the temperature considerations. As the thermal issues become critical in the 3D IC chips, it may not be feasible to boost up the system performance without a sophisticatd cooling system. In this paper we address the problem of maximizing energy efficiency for a given thermal budget under the convection air cooling. Our co-design approach is based on i) picking a system optimization objective (system level energy efficiency), ii) characterization of interdependencies (temperature-delay behavior), iii) understanding the consequential impact on applications (performance vs. energy efficiency), iv) devising online solutions for optimizing combinations of applications, architecture and circuits (temperature-driven dynamic adaptation of voltage margins), and v) assessing the gains for alternative packaging options (2.5D and 3D).

Section 2 describes the system architecture and the thermal cooling structure. Section 3 reports the characterization of the temperature-dependency of the SRAM at the 16nm technology node. Section 4 presents our voltage adaptation algorithm for the SRAM LLC cache. Sections 5 and 6 describe the simulation methodology and present the evaluation results. Section 7 concludes with some final thoughts and recommendations.

SYSTEM MODEL

We demonstrate this co-design approach using a system model of a 16-core homogenous x86 multicore architecture with IC stacking in 16nm technology. The system package consists of two IC tiers as illustrated in Figure 1.a. The dimensions of each die are 8.4mm X 8.4mm. The bottom tier contains the multicore processor with cores interconnected by a 2D torus - one core and its L1 cach are connected to a single router. The processor microarchitecture derives from a typical X86 out-of-order core including the front-end with and L1 instruction cache, the out-of-order scheduler, the integer ALU, the floating-point unit and a private L1 data cache of 32KB. The top tier is a shared L2 LLC SRAM cache with 16 banks of 2MB each - one bank is associated with each core. The LLC cache is a banked shared coherent cache implementing the MESI protocol. Each SRAM bank is connected to a core on the processor die via the through-silicon vias (TSV).

Both the processor and cache tiers contain 3 physical layers. The silicon layer represents the substrate in both tiers; the active layer contains the circuits; the back-end-of-line (BEOL) layer is used for wiring and routing. The face-to-back interconnection between the processor logic and cache structure resides in the BEOL layer of the processor tier and the silicon layer of the cache tier. The processor and cache tiers are placed on a silicon interposer.

The main DRAM memory can either be placed above the LLC cache tier as a stacked DRAM structure (3D packaging), or share the interposer with the (processor+LLC) package (2.5D packaging), as shown in Figure 1.b and Figure 1.c. The interconnection between the LLC cache and the DRAM memory is configured in a 2D torus topology. The memory bandwidth of the 3D package is 4X of that of the 2.5D package, and the latency of each memory request in the 3D package is improved by 40% compared to a 2D package [1]. The microarchitecture parameters are listed in Table 1.

We construct a conventional air-forced heat sink on top of both packages, shown in Figure 2. The package is attached directly to the copper heat sink with a dimension of 50mm X 50mm X 20mm. The spacing between the copper plates is set to 6mm. The thermal related parameters are listed in Table 2.

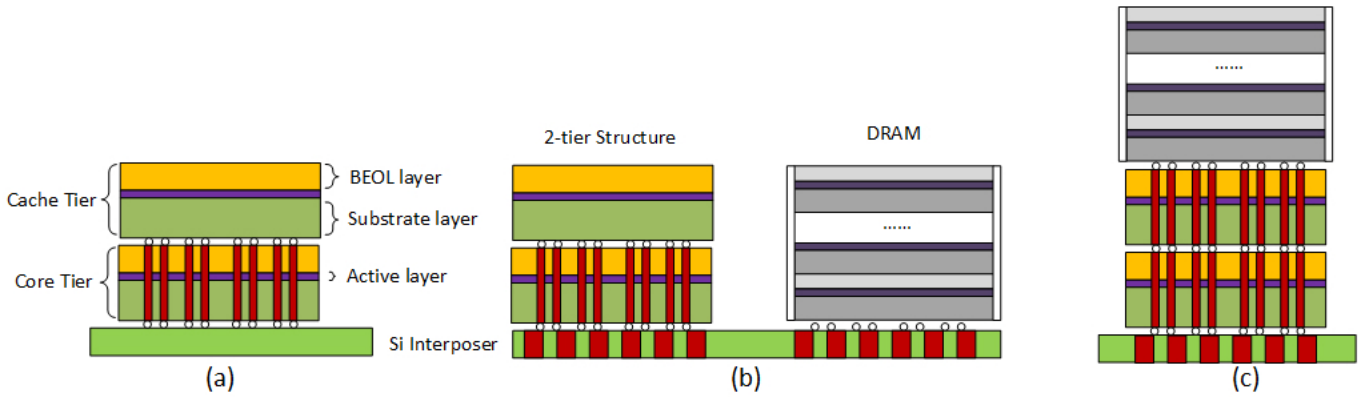

FIGURE 1. THE PHYSICAL MODEL OF: (a) THE 2-TIER STRUCTURE. (b) THE 2.5D PACKAGE WITH DRAM. (c) TRUE 3D PACKAGE WITH STACKED DRAM.

TABLE 1. THE PARAMETERS OF THE MICROARCHITECTURE CONFIGURATION.

Core configuration	
Fetch width	4
Execution width	5 (4 INT ports, 1 FP port)
InstQ size	32
ROB size	128
LSQ size	48 (32 loads, 16 stores)
Cache configraution	
IL1	4-way 16KB, 1 cycle
DL1	8-way 32KB, 1 cycle
LLC	32-way 32MB, 16 banks, 30 cycles
Memory configraution	
2.5D Memory	4 mem controllers, 50ns per access
3D Memory	16 mem controllers, 30ns per access

SRAM Cache Characterization

In order to exploit the thermal dependency of the LLC cache in the 2-tier 3D structure, we implement a SRAM bank model with thermal interaction. In this work, the sub-array is synthesized with a schematic level memory compiler for a given memory array configuration, as shown in Figure 3. To evaluate the temperature dependencies, the sub-array's temperature-delay and temperature-leakage interactions are simulated through Hspice model in 16nm technology. The transistor

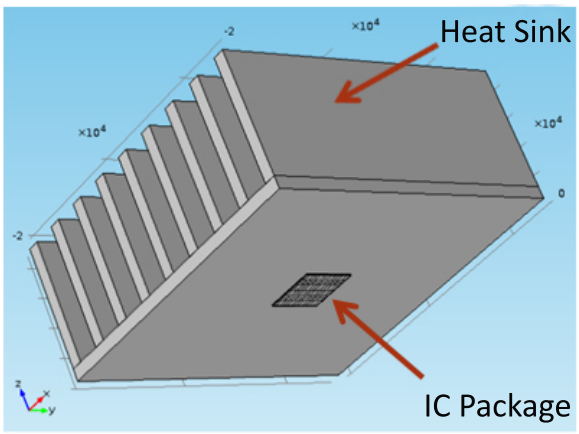

FIGURE 2. THE THERMAL STRUCTURE OF THE TARGETED 16-CORE SYSTEM MODEL.

sizing and cell configurations are optimized for predictive model from Sinha *et al.* [7]. The DC full array simulation across temperature provides leakage information to calibrate leakage trend for the temperature-leakage model. For delay-temperature analysis, the critical path is simulated for switching activities. The critical path of the sub-array delay across room temperature to thermal throttling threshold is tabulated for look-up, as shown in Figure 4.

Because of the regularity of the SRAM array, the extracted critical path of the sub-array is deterministic. The critical path of a conventional SRAM bank is determined by the word-line driver, cell drive bit-line, sensamp sensing, and bit-line precharge/sensamp reset. This model assumes the wordline-reset is masked during sensamp evaluation with a divided-bitline multiplexing architecture where the critical path is defined as:

TABLE 2. THE THERMAL PARAMETERS OF THE 3D STACKED STRUCTURE

Air convection heat sink	
Heat transfer coefficient	1e-10 W/μm^{2}K
BEOL layer	
Thickness	25 μm
Thermal conductivity	2.25e-6 W/μmK
Volumetric heat capacity	2.175e-12 J/μm^{3}K
Silicon Substrate layer	
Thickness	50 μm
Thermal conductivity	1.30e-4 W/μmK
Volumetric heat capacity	1.628e-12 J/μm^{3}K
Active layer	
Thickness	10 μm

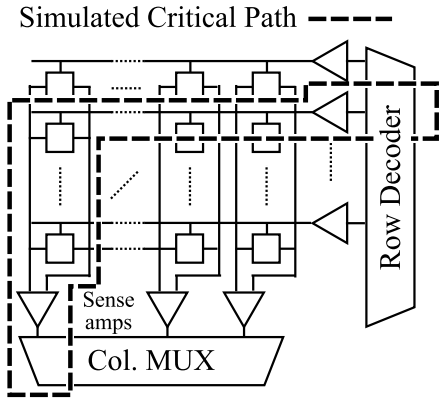

FIGURE 3. THE DELAY MODELING OF THE SRAM SUBARRAY CRITICAL PATH.

$$T_{random-cycle} = T_{wordline-driver} + T_{cell-drive-bitline} + T_{sensamp} + T_{sensamp-precharge} \quad (1)$$

When the LLC cache operates at 0.8V, the SRAM access time at 300°K is reduced to 42% compared to the latency at 400°K, which indicates potential benefits in energy efficiency in the compared to the worst-case design paradigm. If the supply voltage of the cache can be scaled down during runtime according to its temperature, we can achieve a better system energy efficiency without sacrificing the system performance.

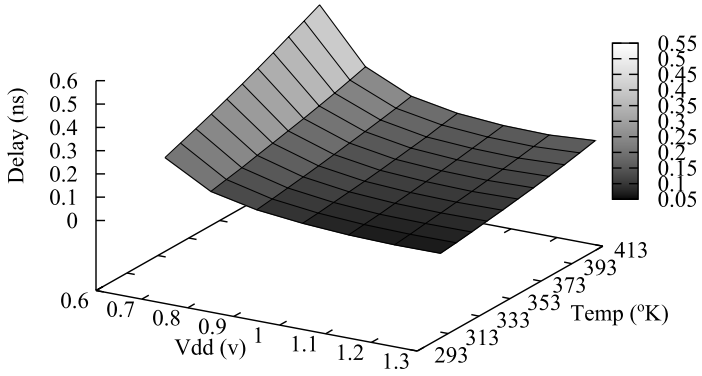

FIGURE 4. THE DELAY OF THE STATIC TIMING MODEL IN TERMS OF TEMPERATURE WITH VOLTAGE VARIASION.

CONSTANT PERFORMANCE MODEL

In this section, we propose a constant performance model (CPM), to fully utilize the thermal headroom during operation. Since the I_{on} current of a CMOS device is a quadratic function of the supply voltage and I_{off} current is an exponential function of V_{dd}, the supply voltage scaling is an effective way to reduce the power consumption. The CPM model, derived from dynamic voltage scaling, regulates the supply voltage of the SRAM cache banks individually to reduce the runtime power consumption. As the delay of the SRAM crtical path is fixed, the voltage drop will not introduce extra bit errors at runtime.

Initially, the voltage of each bank corresponds to the maximum SRAM access delay which corresponds to that for maximum temperature, i.e., worst case conditions. The goal of the CPM is to enable bank-level voltage regulation in SRAM LLC cache, to dynamically reduce the unnecessary voltage margin at lower temperature mitigating the effects of using worst-case design voltage margins. The voltage can be reduced without compromising timing integrity since the critical path delay also reduces. Thus the CPM model decreases the supply voltage of the cache bank according to the temperature level of that bank while maintaining a constant cache access time, as depicted in Figure 5, based on the SRAM temperature-dependency curve. Since the cache latency remains constant throughout execution, system performance will not be degraded.

Meanwhile, as the voltage drop of the cache banks reduces power consumption, it provides a positive feedback to reduce the temperature of the whole system. Figure 6 demonstrates the improvement of the thermal behavior of the SRAM cache tier when running the typical memory bounded application *lu-nc* - one of the benchmark applications used in this analysis. The maximum temperature (hotspot) is reduced by around 8°K.

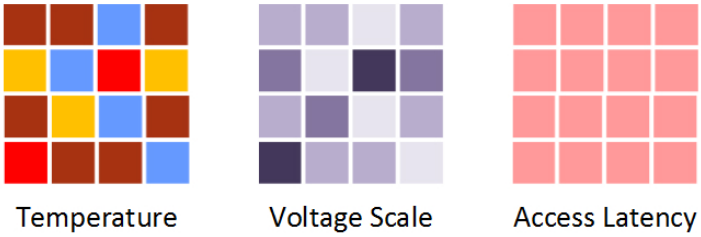

FIGURE 5. THE RUNTIME DIAGRAM OF THE CPM MODEL WITH TEMPERATURE VARIATION AMONG BANKS.

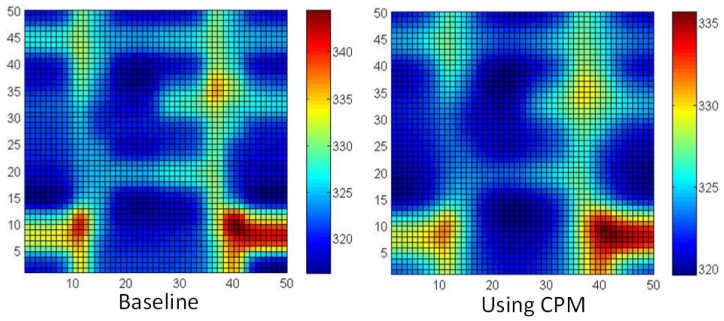

FIGURE 6. THE TEMPERATURE COMPARATION OF THE SRAM TIER BETWEEN THE BASELINE AND THE CPM RUNNING LU-NC APPLICATION.

Constant Delay SRAM Access

Our baseline LLC cache operates at 0.8V in 3GHz, and the hit time is 30 cycles. The worst case thermal conditions lead to bank temperatures of approximately 400°K. To characterize the temperature dependency of the cache, we run simulations across wide range of voltages - 0.6V to 1.1V and corresponding temperatures. Figure 7 illustrates the results of this analysis showing the voltages required to maintain this baseline SRAM latency (corresponding to 0.8V under 400°K) at different temperatures. The supply voltage can be reduced to 0.66V without any performance degradation when a cache bank temperature drops to 300°K.

When the LLC cache initializes at the start-up, the supply voltage of the entire 16 banks are set to 0.8V. The voltage is then scaled down when its temperature is below the scaling threshold. By assuming an ideal integrated voltage regulator (IVR) , the voltage transition completes instantly.

Voltage Adaptation Algorithm

The basic idea of thermal adaptation is to trade off the circuit timing headroom with supply voltage reduction in the SRAM cache. The algorithm is depicted in Algorithm 1. The new voltage of the cache bank is determined by both the current temperature of the cache bank and the power of its associated core. The timing margin of the SRAM access can be directly calculated using the temperature of the local cache bank, and then be converted to the correct voltage drop. The power of the associated core gives hints as to the pipeline execution performance, and sets up the minimal voltage constraints for the SRAM cache bank to guarantee correct functionality. The new voltage is updated by striking a balance between the two parameters.

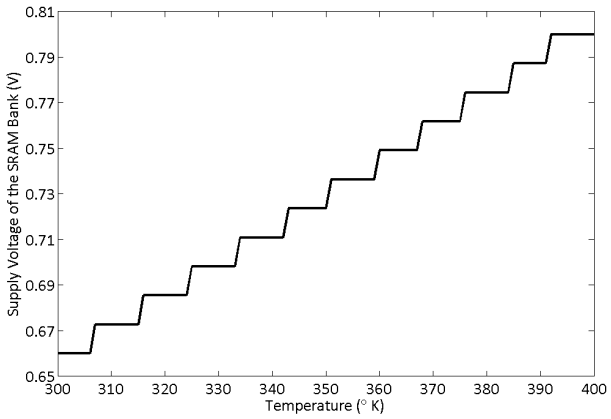

FIGURE 7. SUPPLY VOLTAGE SCALING OF THE SRAM LLC CACHE TO MAINTAIN CONSTANT ACCESS LATENCY WITH RESPECT TO TEMPERATURE.

Algorithm 1 Thermal Adaptaion of the SRAM Supply Voltage

```
1:  function AdptFrmwrk(void)
2:      updatePower(core[], cache[]);
3:      updateTemperatre();
4:      synchronizationBarrier();
5:      for i = 0 to cache.banknum-1 do
6:          cache[i].volt=updateVoltage(cache[i].temp,
7:                        core[i].power);
8:      end for
9:      synchronizationBarrier();
10: end function
11:
12: function updateVoltage(cacheT, coreP)
13:     index=genIndx(cacheT, coreP);
14:     newVolt=voltTbl[index];
15:     return newVolt;
16: end function
```

For compute bounded applications, the temperature of the LLC bank is largely affected by the activity of its associated core. As there are relatively fewer LLC access when running

these type of applications, we only need to maintain the minimal supply voltage for a cache access. In contrast, the power consumption of the cache banks is much larger when the system executes memory bounded applications, and thus the bank temperature is mainly determined by the activity of the LLC bank. For the other applications, the voltage drop is a result of a combination of memory and compute behaviors.

The Effect of the Stack Organization

The system performance and power profile depends on the DRAM memory configuration as well. When the DRAM memory is stacked on the LLC cache tier (3D package), the overall system has a much larger memory bandwidth and a lower memory request latency. As a result, the cores tend to consume more power and generate more heat, and the thermal coupling between cores and the LLC is significant. On the other hand, the stacked DRAM adds thermal resistance to the heat sink, and the cooling capacity is reduced. Therefore, the power/energy improvement from the CPM scheme is limited compared to the use of 2.5D packaging where the DRAM is instead placed on the silicon interposer.

SIMULATION FRAMEWORK

We construct the 16-core simulator based on the manifold cycle-based simulation framework [8] integrated with the Energy Introspector (EI) multi-physics architecture level modeling library [9]. The EI library includes common open source power, thermal, and reliability models integrated to capture multi-physics interactions, e.g., coupling between temperature and leakage power. The simulator provides a full functional, timing, power and thermal analysis infrastructure. The timing model of the microarchitecture is driven by the Linux OS and applications. Simulations are fast-forwarded to the regions of interest to initialize the processor and warm up the cache. The detailed pipeline execution and cache access information are collected every 10 ms and drive the power model (McPAT [10]) to generate power traces of each 3D tier. At the end of the sampling window, the power maps of the system components are provided to the 3D interlayer thermal library (3D-ICE [11]) to compute the thermal grids. Finally, the thermal information is collected by the voltage scaling controller to adjust the supply voltage of the SRAM banks.

RESULTS AND DISCUSSION

In this section, we evaluate the CPM model relative to the baseline system with the supply voltage of the LLC cache fixed to 0.8V. The test applications are picked up from the SPLASH-2 benchmark [12], a parallel application suite written for the multicore shared-memory architecture. The cache characterization of the applications are shown in Table 3 running on a baseline configuration. The LLC miss rate between 2.5D and 3D differs due to the reason that 3D chips have more memory controllers, and thus the route from L2 to DRAM is shorter.

TABLE 3. SPLASH-2 BENCHMARK CHARACTERIZATION. HR - HIT RATE. MR - MISS RATE.

App	L1 HR	LLC MR 2.5D	LLC MR 3D
barnes	96.9%	16.4%	17.0%
fmm	98.1%	40.0%	40.6%
lu-nc	93.6%	45.3%	43.1%
ocean-c	93.6%	44.2%	44.3%
radiosity	99.2%	17.8%	17.4%
radix	97.3%	43.8%	44.7%
raytrace	96.5%	25.0%	24.7%
water-ns	98.6%	25.1%	25.3%

The hit rate of L1 cache and the miss rate of the LLC cache are calculated as the geometric mean of the cache banks. The compute bound applications have a high L1 hit rate, and most of the memory request are can be served in the L1 cache (e.g. *radiosity*). The memory bound applications otherwise have a high interaction with the LLC and the main memory (e.g. *lu-nc* and *ocean-c*).

Meanwhile, a ideal system is used to capture an upper bound of the performance of the CPM model - this model maintains the delay of the SRAM corresponding to to 300°K and sets the supply voltage to 0.66V.

System Performance

We compare the system performance between the 2.5D and 3D packaging in terms of instruction per cycle (IPC), as shown in Figure 8. The memory bound applications are more sensitive to the packaging differences because of the intensive interaction between processors and DRAM memory. For example, the *lu-nc* application suffers from over 50% performance degradation from 3D to 2.5D, while the *water-ns* application has only 0.9% IPC reduction.

There are 3 main reasons for the impact of the 2.5D packaging on system performance: i) DRAM bandwidth is reduced, as the available memory controllers in 2.5D packaging is limited compared to 3D, ii) the DRAM access time is higher, and iii) the average routing distance between the LLC cache and the DRAM controller is longer.

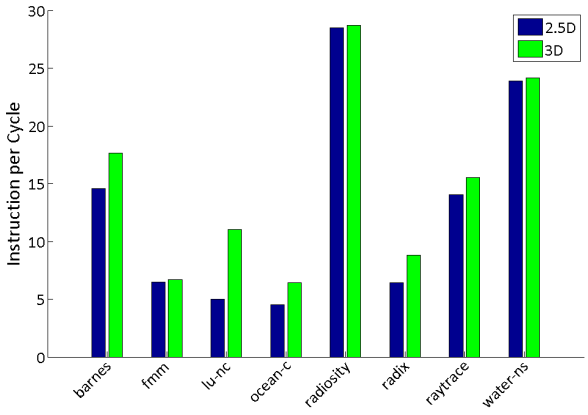

FIGURE 8. THE SYSTEM PERFORMANCE COMPARISON BETWEEN 2.5D AND 3D PACKAGING IN TERMS OF IPC.

Power Reduction

The power reduction of the LLC cache using CPM is shown in Figure 9. The CPM reduces an overall 15% of the maximum SRAM power among the 8 applications, and an average of 23% of the minimum power both in the 2.5D and 3D packages. The significant power saving of the comes from the reduction of the unnecessary voltage margin. Memory intensive applications such as *ocean-c* saves up to 30% of the power.

Energy Consumption

The CPM also reduces the total energy consumed in the system throughout the execution. Figure 10 illustrates the normalized energy reduction of the LLC cache. The energy saving of the SRAM system is over 20% in average when CPM is employed.

Cache Temperature

The voltage drop in the LLC cache that reduces the power consumption will also reduce the temperature of the cache bank, which in turn helps to further decrease the voltage of the LLC bank. As shown in Table 4, the CPM will reduce the hotspot of the SRAM cache by an average of 4.3°K. The 2.5D package has a little better temperature reduction, as the system runs slower than the system with 3D DRAM stacking. The temperature of the 2.5D system is lower, enabling greater voltage drop during execution. The *lu-nc* application has the largest temperature reduction of 8.8°K and 7.4°K respectively in the two package configurations for 2 reasons. First, it is a memory bound application, and the power reduction is significant when there is a voltage drop in the cache; second, it has the largest amount of LLC activity of all memory bound applications.

TABLE 4. THE MAX TEMPERATURE VARIATION BETWEEN THE BASELINE AND CPM OVER THE SPLASH-2 BENCHMARK.

Δ T (°K)	2.5D DRAM	3D DRAM
barnes	6.9	4.6
fmm	5.7	5.2
lu-nc	8.8	7.4
ocean-c	7.4	4.9
radiosity	2.6	2.5
radix	4.4	4.6
raytrace	5.1	3.2
water-ns	4.1	2.3

Energy Efficiency

The power consumed in the LLC cache can take up 10% 35% of the total power of the 2-tier structure, and the runtime SRAM power reduction will improve the system energy efficiency in terms of energy per instruction, as shown in Figure 11. The EPI records the average energy used to execute a single instruction. For both of the 2.5D and 3D configurations, the memory bound applications have better EPI improvement (11%) than compute bound applications (5%), as the proportion of power consumption in the cache system is higher.

CONCLUSIONS

The paper presents the CPM model for thermal adaptation of an SRAM LLC cache to improve the energy efficiency and power consumption in the 3D stacked ICs. The key idea of the CPM is to construct a bank-level supply voltage regulator that maintains a constant SRAM access time based on the temperature-delay dependence of the SRAM LLC cache. The novelty of the CPM lies in that the voltage scaling of the LLC cache is controlled by the temperature and the power of its associated cores, directly addressing the new thermal challenges of 3D ICs. We evaluate the system performance, power/energy consumption, and the energy efficiency of the proposed adaptation technique to both 2.5D and 3D packaging. The simulation results show up to 30% reduction of the peak power and 27% saving in energy consumption of the SRAM cache, compared to the conventional worst-case SRAM design. The memory bounded applications are most benefited from the CPM mechanism. The EPI of the 16-core processor is improved on average by 5% in the 2.5D packaging and 8% in the 3D packaging. The co-design approach to the adaptation SRAM structure indicates potential opportunities to build an high performance, power and energy efficient system using 3D stacked

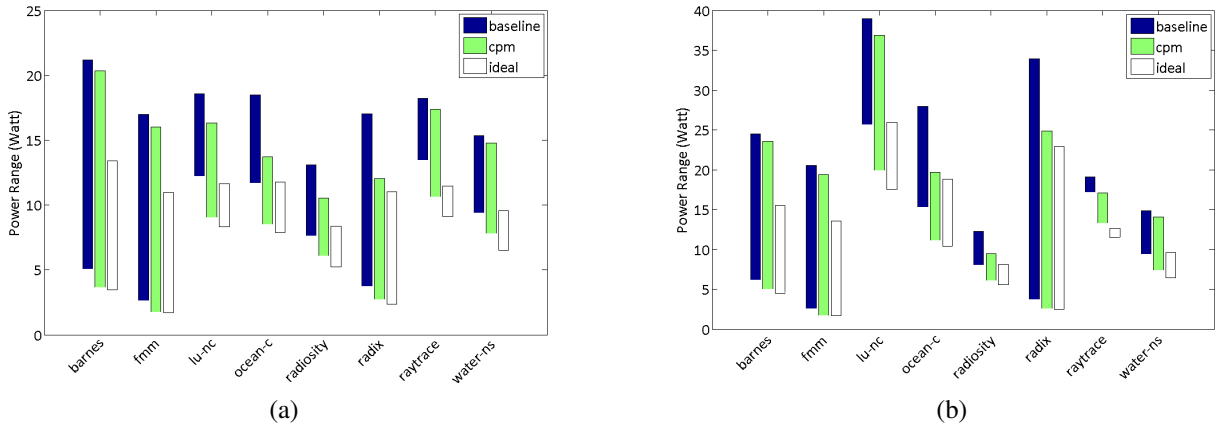

FIGURE 9. THE POWER RANGE PROFILE OF: (a) THE SYSTEM WITH 2.5D DRAM PACKAGING (b) THE SYSTEM WITH 3D DRAM STACKING.

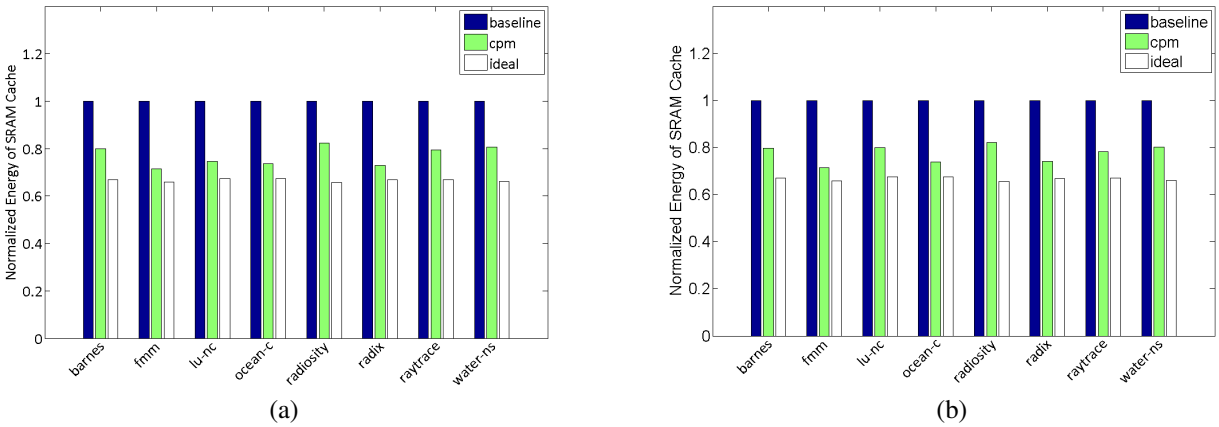

FIGURE 10. THE TOTAL NORMALIZED SRAM ENERGY SAVING OF: (a) THE SYSTEM WITH 2.5D DRAM PACKAGING (b) THE SYSTEM WITH 3D DRAM STACKING.

IC technology.

ACKNOWLEDGMENT

This work is supported and sponsored by the Semiconductor Research Corporation task 2318.001.

REFERENCES

[1] Loh, G., 2008. "3d-stacked memory architectures for multi-core processors". In Computer Architecture, 2008. ISCA '08. 35th International Symposium on, pp. 453–464.

[2] Khan, N., Alam, S., and Hassoun, S., 2009. "System-level comparison of power delivery design for 2d and 3d ics". In 3D System Integration, 2009. 3DIC 2009. IEEE International Conference on, pp. 1–7.

[3] Xiao, H., Wan, Z., Yalamanchili, S., and Joshi, Y., 2014. "Leakage power characterization and minimization in 3d stacked multi-core chips with microfluidic cooling". In Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), 2014 30th Annual, pp. 207–212.

[4] Wan, Z., Xiao, H., Joshi, Y., and Yalamanchili, S., 2013. "Co-design of multicore architectures and microfluidic cooling for 3d stacked ics". In Thermal Investigations of ICs and Systems (THERMINIC), 2013 19th International Workshop on, pp. 237–242.

[5] Semeraro, G., Albonesi, D., Dropsho, S., Magklis, G., Dwarkadas, S., and Scott, M., 2002. "Dynamic frequency and voltage control for a multiple clock domain microarchitecture". In Microarchitecture, 2002. (MICRO-35). Pro-

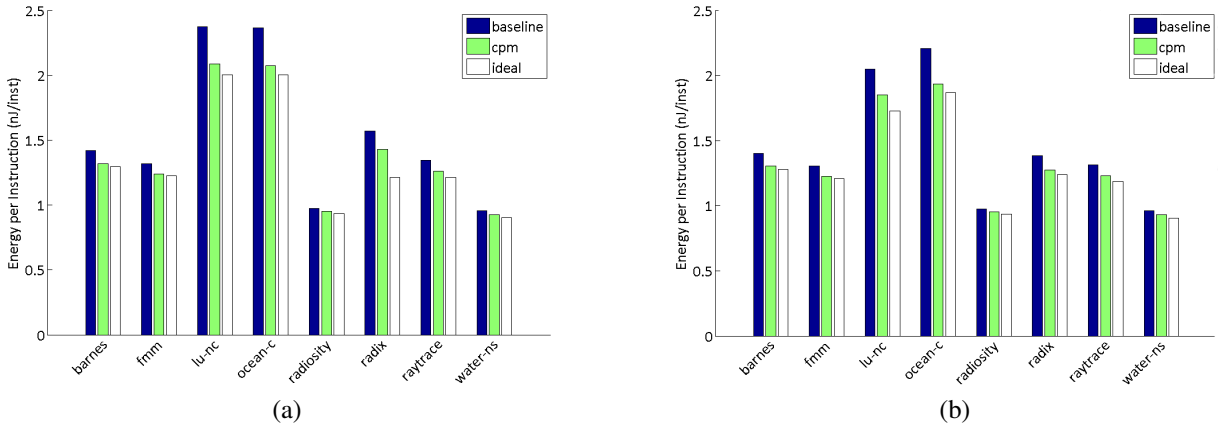

FIGURE 11. THE ENERGY EFFICIENCY COMPARISON OF: (a) THE SYSTEM WITH 2.5D DRAM PACKAGING (b) THE SYSTEM WITH 3D DRAM STACKING.

ceedings. 35th Annual IEEE/ACM International Symposium on, pp. 356–367.

[6] Raghavan, A., Luo, Y., Chandawalla, A., Papaefthymiou, M., Pipe, K. P., Wenisch, T., and Martin, M., 2012. "Computational sprinting". In High Performance Computer Architecture (HPCA), 2012 IEEE 18th International Symposium on, pp. 1–12.

[7] Sinha, S., Yeric, G., Chandra, V., Cline, B., and Cao, Y., 2012. "Exploring sub-20nm finfet design with predictive technology models". In Proceedings of the 49th Annual Design Automation Conference, ACM, pp. 283–288.

[8] Wang, J., Beu, J., Bheda, R., Conte, T., Dong, Z., Kersey, C., Rasquinha, M., Riley, G., Song, W., Xiao, H., Xu, P., and Yalamanchili, S., 2014. "Manifold: A parallel simulation framework for multicore systems". In Performance Analysis of Systems and Software (ISPASS), 2014 IEEE International Symposium on, pp. 106–115.

[9] Song, W., Mukhopadhyay, S., and Yalamanchili, S., 2014. "Energy introspector: A parallel, composable framework for integrated power-reliability-thermal modeling for multicore architectures". In Performance Analysis of Systems and Software (ISPASS), 2014 IEEE International Symposium on, pp. 143–144.

[10] Li, S., Ahn, J.-H., Strong, R., Brockman, J., Tullsen, D., and Jouppi, N., 2009. "Mcpat: An integrated power, area, and timing modeling framework for multicore and manycore architectures". In Microarchitecture, 2009. MICRO-42. 42nd Annual IEEE/ACM International Symposium on, pp. 469–480.

[11] Sridhar, A., Vincenzi, A., Ruggiero, M., Brunschwiler, T., and Atienza, D., 2010. "3d-ice: Fast compact transient thermal modeling for 3d ics with inter-tier liquid cooling". In Computer-Aided Design (ICCAD), 2010 IEEE/ACM International Conference on, pp. 463–470.

[12] Bienia, C., Kumar, S., and Li, K., 2008. "Parsec vs. splash-2: A quantitative comparison of two multithreaded benchmark suites on chip-multiprocessors". In Workload Characterization, 2008. IISWC 2008. IEEE International Symposium on, pp. 47–56.