An Energy-Efficient 3D CMP Design with Fine-Grained Voltage Scaling

Jishen Zhao, Xiangyu Dong, Yuan Xie
Computer Science and Engineering Department, Pennsylvania State University
{juz138,xydong,yuanxie}@cse.psu.edu

***Abstract*—In this paper, we propose an energy-efficient 3D-stacked CMP design by both temporally and spatially fine-grained tuning of processor cores and caches. In particular, temporally fine-grained DVFS is employed by each core and L2 cache to reduce the dynamic energy consumption, while spatially fine-grained DVS is applied to the cache hierarchy for the leakage energy reduction. Our tuning technique is implemented by integrating an array of on-chip voltage regulators into the original processor. Experimental results show that the proposed design can provide an energy-efficient, direct, and adaptive control to the system, leading to 20% dynamic and 89% leakage energy reductions, and an average of 34% total energy saving compared to the baseline design.**

I. Introduction

Most research on chip multiprocessor (CMP) focused on finding sophisticated methods and strategies to improve system performance. Recent work has indicated that three-dimensional (3D) integration can improve the communication latency and the integration density for CMPs [1]. However, the continuous increase of power and energy budgets for CMP designs potentially brings in critical system design problems, such as power supply rail design, system reliability, and thermal issue.

A variety of energy-efficient designs have been proposed to decrease the system power consumption while preserving performance as much as possible. Dynamic voltage and frequency scaling (DVFS) is a widely-used power management technique to reduce the supply voltage according to the processor clock frequency [2]. Most existing DVFS mechanisms conform to a temporal granularity at tens of microseconds. Fine details of a workload's instantaneous activities may be hidden due to coarse-grained time intervals. Consequently, leveraging the workload activity details is a possible solution for further energy savings. Other research work focuses on reducing the power supply of cache hierarchies. In a CMP, some private caches are frequently accessed while others are not. Dynamic voltage scaling (DVS) for leakage control [3] addresses this issue by putting caches into drowsy state when they are inactive. Moreover, leveraging the low leakage power and non-volatility of Magnetic RAM (MRAM), SRAM-MRAM hybrid cache [4] uses heterogeneous 3D integration to improve the energy efficiency of L2 cache without much performance overhead, since most performance-critical write accesses are accommodated by the low-latency SRAM. However, even small fraction of SRAM-based L2 cache consumes over twice the leakage power as MRAM-based L2 cache with $32\times$ capacity, because it must always be online. In addition, read and write to the same cache can be significantly unbalanced for many workloads. Therefore, potential leakage power saving can be obtained by partitioning the cache hierarchy into many voltage regions and dynamically control the voltage supply of each one.

978-3-9810801-7-9/DATE11/

In this paper, we propose an energy-efficient 3D CMP design with fine-grained tuning of voltage and frequency control to processor cores and caches with only small performance overhead by integrating multiple on-chip voltage regulators (VRs). We further examine the area, performance, and power overhead brought in by on-chip VRs, and show considerable energy-efficiency that can be achieved by our fine tuning techniques with the presence of such overhead.

II. Background and Related Work

3D Integration: Modern CMPs can benefit from 3D integration with respect to footprint, speed, and heterogeneous integration. The critical paths can be significantly shortened and the bandwidth between processor cores and memories can be greatly increased [5]. 3D integration technology can be categorized into two major approaches, monolithic and stacking [6]. In this paper, we assume that the stacking approach is used for 3D CMP design, in which logic dies and memory dies are vertically stacked and connected by through-silicon vias (TSVs).

DVFS: Most existing DVFS mechanisms conform to a temporal granularity at tens of microseconds, due to both the software control delay and the limitation of voltage transition speed of traditional off-chip VRs. Recent work on on-chip VRs explored the possibility to overcome this limitation [7]. Several approaches with finer temporal granularities have been proposed. A thread motion-based power management scheme based on on-chip VRs was presented by Rangan ***et al.*** [8]. However, this technique is not scalable, since the number of available frequency and voltage levels is restricted by the performance variations of the cores. Kim ***et al.*** analyzed the effectiveness of DVFS on different temporal granularities [9]. However, they only employed per-core DVFS in their mechanism. Nevertheless, L2 caches can also benefit from fine-grained DVFS for more dynamic power reduction. In this paper, we present a hardware-based DVFS mechanism applied to both processor cores and L2 caches.

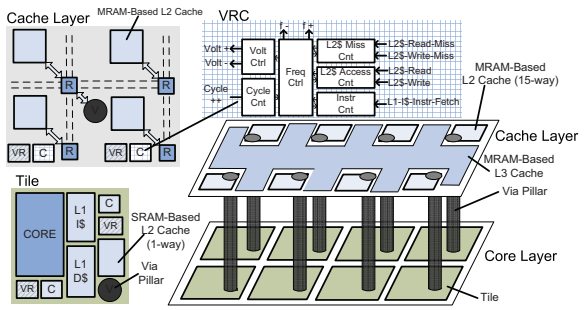

Fig. 1. Overview of the energy-efficient 3D CMP configuration.

Hybrid Cache: Due to its non-volatility, MRAM can be used in cache system to save leakage power consumption. Nevertheless, MRAM has a long write latency and high write energy consumption. To address this problem, a hybrid L2 cache design was proposed by Sun ***et al.*** [4] to integrate both SRAM and MRAM in cache hierarchy for performance and power benefits. According to their study, SRAM cache with a quarter of the capacity could consume over $8\times$ more leakage power than MRAM cache. Consequently, considerable power reduction can be achieved if we discard the SRAM cache as much as possible in the hybrid cache hierarchy. In this work, we propose to apply spatially fine-grained DVS to both SRAM- and MRAM-based caches, with on-chip VRs to provide the power supply.

III. Configuration and Parameters

The baseline 3D CMP is an 8-core in-order processor based on UltraSPARC III architecture. SRAM-MRAM hybrid L2 caches similar to the design presented by Sun ***et al.*** [4] are adopted, each consisting of 15-way MRAM cache and 1-way SRAM cache to accommodate L2 cache writes. As shown in Fig. 1, the baseline processor is partitioned into two layers, namely the core layer and the cache layer. The core layer consists of 8 tiles, each of which contains a processor core, its private L1 instruction and data caches, and SRAM L2 cache. The cache layer consists of MRAM private L2 caches and a shared MRAM L3 cache. Communications between the two layers are conducted by TSV pillars. Table I lists the detailed baseline parameters. We estimate the area and dynamic power consumption of our baseline under $65nm$ technology using processor area and power estimation tool McPAT [10]. The SRAM and MRAM cache areas and leakage power are estimated with our modified CACTI [11]. The results are listed in Table II.

In order to leverage the fine-grained tuning of the voltage supply for each core and cache, we integrate 25 on-chip VRs into the baseline. A four-phase on-chip VR from the work of Hazucha ***et al.*** [7] is adopted. Table III lists the related parameters of such an on-chip VR, and an off-chip switching regulator from the work of Namgoong ***et al.*** [12] for comparison. On the core layer, each tile contains a VR. Another 8 VRs are integrated to supply voltage to the SRAM-based L2 cache ways. All the 16 VRs increase the area of the core layer by $20.32mm^2$ (16.6%), to $142.32mm^2$. We stack the other 9 VRs for both MRAM L2 caches and MRAM L3 cache on the cache layer. The area of the cache layer is increased to $133.74mm^2$. VR controllers (VRCs) are used to configure the output of each VR.

TABLE I
Baseline 3D CMP configuration.

Cores	8-core, 1 GHz, in-order, 14-stage pipeline
L1 I/D-caches	8 x 64KB, 64B line, 2-way, write-through, private, SRAM, read/write latency: 1 cycle
L2 cache	8 x 256KB, 64B line, private, 16-way (1-way SRAM, 15-way MRAM), write-back, read latency: 1 cycle(SRAM)/5 cycles(MRAM), write latency: 1 cycle(SRAM)/16 cycles(MRAM)
L3 cache	32MB, 64B line, 32-way, write-back, shared, MRAM read latency: 15 cycles, write latency: 37 cycles
Main memory	4GB, read/write latency: 100 cycles
Communications	inter-layer: TSV pillars, 1-cycle latency intra-layer: routers, 2-cycle latency

TABLE II
Area and power consumption of the baseline.

Layer	Area	Dynamic Power	Leakage Power
Core layer	$122.6mm^2$	$2.35W$(peak $24.75W$)	$2.04W$
Cache layer	$122.3mm^2$	$4.1W$(peak $62W$)	$6.7W$

IV. Energy-Efficient Design

A. Temporally Fine-Grained DVFS

The granularity of conventional DVFS methodologies ranges from 10 to $100\mu s$ due to the limited voltage transform speed of off-chip VRs. By employing on-chip VRs, we are able to apply DVFS to both the cores and L2 caches at a much finer temporal granularity of 100 to $200ns$ [9]. To exploit such a fine-grained DVFS, we evaluate three variables in given time intervals, which are instructions per cycle (IPC), number of L2 cache accesses (NA), and L2 cache read and write misses (M_r and M_w). In order to deal with the unbalanced L2 cache read and write miss penalties, i.e., different read and write latencies of the MRAM-based L3 cache, we compute the weighted miss rate (WMR) at i-th time interval by

$$\text{WMR}(C_i) = \frac{M_{r_i} + \lambda M_{w_i}}{\text{NA}}, \quad \lambda = \frac{l_w}{l_r} \tag{1}$$

, where l_r and l_w are the L3 cache read and write latencies respectively.

Fig. 2 is a plot of IPC and WMR of the multi-threaded benchmark *apsi* running on a 4-core processor, examined with different time intervals. When evaluated with a large time scale, the activities appear to be stable for all the 4 cores during the time interval 1500-1700μs. However, if we look into a much smaller time scale, e.g., $200ns$ time interval, a different picture is shown in the figures on the right. Even within $2\mu s$, the IPC and the L2 cache miss rate of each core still vary. Some cores are idle for a great proportion of time. Consequently, dynamically adjusting the VF level of each core and cache in small time intervals can potentially save more power. Another observation is that the same activity pattern is likely to repeat for several time intervals in each core, when evaluated by the fine temporal granularity. Therefore, the core's activity of the next time interval can potentially

TABLE III
PARAMETERS OF OFF-CHIP [12] AND ON-CHIP [7] VR.

Type	Physical Area	Transition delay	Power efficiency
Off-chip	$13.7mm^2$	$1\mu s$	93%
On-chip	core: $0.14mm^2$ Decap+circuitry: $1.13mm^2$	$20ns$	87%

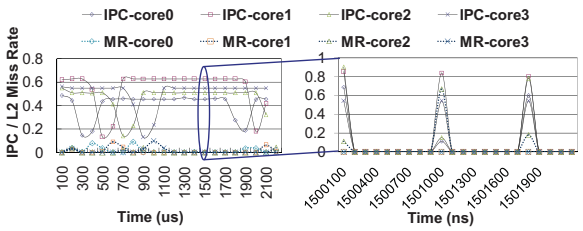

Fig. 2. System activity in terms of each core's IPC and the L2 cache miss rate examined at different granularities of time interval.

be predicted using the core's history records. Based on these observations, we propose a temporally fine-grained DVFS technique to improve the dynamic power consumption. Two steps are involved with the technique: workload profiling and online fine-grained DVFS.

Before applying DVFS to the system, we obtain the power and performance characteristic profiles of the workloads by simulations with the highest voltage and frequency level. This profile will be used as the performance constraint to the online DVFS mechanism. During run-time, we employ online DVFS to each core and cache. VF level of each core and cache is controlled by the corresponding VRC (shown in Fig. 1) according to IPC, NA, and WMR. When the IPC of a core is low and the miss rate of its L2 cache is high, much of the execution time is spent on servicing the L2 cache misses. Therefore, we reduce the VF level to save system power consumption. When L2 cache miss rate is low, we increase VF level to improve performance. The detailed VF control policy is shown in Table IV.

DVFS brings in performance overhead, in terms of additional cycles for VRCs to adjust the VF levels at each time interval and the voltage transition time of VRs. In each time interval, we compare the current execution time with the one obtained from the workload profile with the same instruction count. If the execution time increases over $\mu\%$, we will increase n_u and decrease n_d in the next time interval so that the frequency tends to be increased to improve the performance. Although the VF adjustment strategy cannot guarantee to achieve the optimal result, it is an easy and effective method at the system design stage. We show in Section V that our DVFS scheme can save energy in a considerable manner.

B. Spatially Fine-Grained DVS

Leakage has become an increasingly dominant component of cache power consumption as feature sizes shrink. Although inactive MRAM-based caches can be turned off to save leakage power, the SRAM-based L2 cache ways, which consume over $2.5\times$ more leakage power than the MRAM-based L2 cache ways in our baseline, cannot be frequently

TABLE IV
ONLINE DVFS POLICY. σ_1, σ_2, AND σ_3 ARE AVERAGE IPC, NA, AND WMR CALCULATED USING THE WORKLOAD PROFILE. n_d IS THE VF LEVELS TO BE DECREASED. n_u IS THE VF LEVELS TO BE INCREASED.

IPC	WMR	Core VF	NA	WMR	L2$ VF
$< \sigma_1$	$> \sigma_2$	$-n_d$ levels	$< \sigma_3$	$> \sigma_2$	$-2n_d$ levels
$> \sigma_1$	$> \sigma_2$	N/A	$< \sigma_3$	$< \sigma_2$	$-n_d$ levels
$< \sigma_1$	$< \sigma_2$	$+n_u$ levels	$> \sigma_3$	$> \sigma_2$	N/A
$> \sigma_1$	$< \sigma_2$	$+2n_u$ levels	$> \sigma_1$	$< \sigma_2$	$+n_u$ levels

powered off because of SRAM's volatility. Consequently, we propose spatially fine-grained DVS to reduce the leakage power consumptions of caches by lowering the supply voltage when they are inactive.

Although SRAM is volatile and we cannot completely shut it down, it is possible to put them into drowsy state to just keep the present data when they are not accessed. In this case, the supply voltage is reduced to $1.5V_t$, where V_t is the threshold voltage. Since both voltage and leakage current can be reduced, the total leakage power consumption is significantly saved. Only 1-cycle is needed to wake up the SRAM caches when they are re-accessed. The corresponding VRC receives signals from the cache controllers, indicating accesses to the cache banks within the voltage region. When the cache banks are inactive, the VRC keeps the voltage supply low. Otherwise, the VRC turns on the corresponding VR. Note that the performance penalty to configure a VR is approximately 1 to 2 cycles in our design, since the voltage transition of on-chip VRs is at nanosecond time scale.

V. EXPERIMENTS

A. Benchmarks and Metrics

We use Simics [13] to run our simulations. Simulation workloads are multi-threaded benchmarks selected from SPEC OMP2001 [14] and PARSEC [15]. Each simulation runs for 1 billion instructions. Since our energy-efficient design techniques are closely related to memory access behavior, we intentionally select several benchmarks with various memory read and write intensities.

Taking into account of both energy saving and performance overhead, we evaluate both power-delay product (PDP) and energy-delay product (EDP) in our experiments by

$$\text{PDP} = (P_d + P_l)t, \quad \text{EDP} = \text{PDP} \cdot t \tag{2}$$

, where dynamic (P_d) and leakage (P_l) powers are calculated as

$$P_d = P_{d0} \frac{\sum_{j=1}^{m} \sum_{i=1}^{n} V_{ij}^2 f_{ij}}{mnV^2 f} \tag{3}$$

$$P_l = \sum_{j=1}^{m} (\sum_{i=1}^{n} (L_1 R_{2ij} + L_2 W_{2ij}) + L_3 RW_{3j}) \tag{4}$$

Here we assume that the effective capacitance does not change in the experiment. m is the number of DVFS time intervals, n is the number of cores in the CMP, V_{ij} and f_{ij} are supply voltage and frequency level applied to the core or cache during the j-th time interval. P_{d0} is the total dynamic power of the cores or L2 caches of the baseline without DVFS.

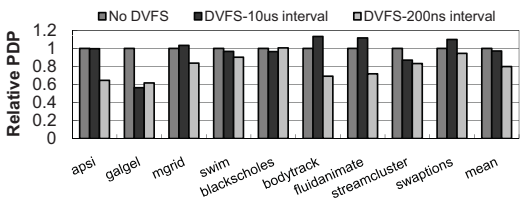

Fig. 3. Relative PDP (energy consumption) with different DVFS mechanisms.

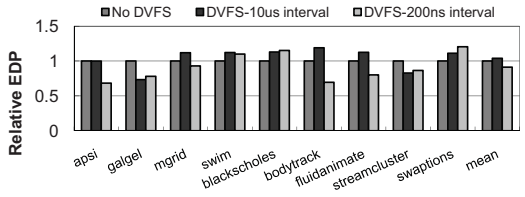

Fig. 4. Relative EDP with different DVFS mechanisms.

R_{2ij}, W_{2ij}, and RW_{3j} are defined as whether the L2 cache read banks, L2 cache write banks, and the L3 cache are turned on (1) or not (ϵ). Here ϵ is the ratio of leakage when the cache bank is shut down. The total execution time is computed as the sum of each time intervals by $t = \sum_{j=1}^{m} t_j$.

B. Results

Dynamic energy reduction We run simulations to evaluate our temporally fine-grained DVFS method with granularities of both $100\mu s$ (coarse-grained) and $200ns$ (fine-grained), which respectively represent off-chip VRs and on-chip VRs being applied to the baseline system. Note that it is impossible to include such a large number of off-chip VRs to implement DVFS to both cores and L2 caches, therefore we only apply coarse-grained DVFS to processor cores for comparison. Although on-chip VRs bring in more overhead than off-chip VRs, there is still 29.6% power reduction on average by applying our fine-grained DVFS method. When applying DVFS in each time interval, we force the increased execution time to be within 30%, so that the system performance is not degraded too much. Relative PDP (energy) reductions are shown in Fig. 3. On average, our fine-grained DVFS strategy achieves 17% more energy reduction than coarse-grained DVFS. Fig. 4 shows the results of performance and energy correlated together. On average, 9% EDP reduction can be observed in fine-grained DVFS, while only 3% for coarse-grained DVFS.

Leakage energy reduction Energy savings in terms of cache leakage reductions are illustrated in Fig. 5. On average, a 88.9% leakage energy reduction is achieved with cache DVS at $200ns$ time interval.

Total energy reduction Finally, we consider both dynamic and leakage energy reductions achieved by our design, and illustrate the relative total energy consumptions of different configurations in Fig. 6. On average, a 3% energy saving compared to baseline can be achieved with off-chip VRs, while a 34% total energy reduction can be achieved with on-chip VRs.

VI. Conclusion and Future Work

We present an energy-efficient 3D CMP design by applying fine-grained control to both processor cores and cache hierarchy with on-chip VRs. Temporally fine-grained DVFS is used to reduce the system dynamic energy, and spatially fine-grained cache DVS is used to save cache leakage energy. Simulations of a set of multi-threaded benchmarks show that our proposed design results in 20% dynamic energy reduction, 88.9% cache leakage energy reduction, and 34% total energy reduction on average. The average performance overhead induced by our design is only 14%.

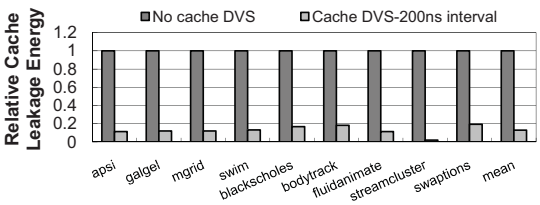

Fig. 5. Comparison of cache leakage energy with different configurations.

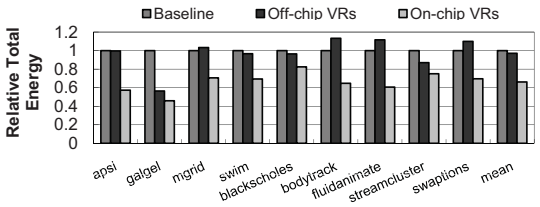

Fig. 6. Relative total energy consumptions of different configurations.

References

[1] B. Zhao, Y. Du, Y. Zhang, and J. Yang, "Variation-tolerant non-uniform 3D cache management in die stacked multicore processor," in *Proceedings of the International Symposium on Microarchitecture*, 2009, pp. 222–231.

[2] K. Choi, R. Soma, and M. Pedram, "Fine-grained dynamic voltage and frequency scaling for precise energy and performance trade-off based on the ratio of off-chip access to on-chip computation times," in *Proceedings of the conference on Design, automation and test in Europe - Volume 1*, 2004, pp. 18–28.

[3] K. Flautner, N. S. Kim, S. Martin, D. Blaauw, and T. Mudge, "Drowsy caches: simple techniques for reducing leakage power," in *Proceedings of the International Symposium on Computer Architecture*, 2002, pp. 148–157.

[4] G. Sun, X. Dong, Y. Xie, J. Li, and Y. Chen, "A novel architecture of the 3D stacked MRAM L2 cache for CMPs," in *Proceedings of the International Conference on High-Performance Computer Architecture*, 2009, pp. 239–249.

[5] C. C. Liu, I. Ganusov, M. Burtscher, and S. Tiwari, "Bridging the processor-memory performance gap with 3D IC technology," *IEEE Design and Test*, vol. 22, no. 6, pp. 556–564, 2005.

[6] Y. Xie, G. Loh, B. Black, and K. Bernstein, "Design space exploration for 3D architectures," *ACM Journal of Emerging Technologies in Computing Systems*, vol. 2, no. 2, pp. 65–103, 2006.

[7] P. Hazucha, G. Schrom, J. Hahn, B. A. Bloechel *et al.*, "A 233-MHz 80%-87% efficient four-phase DC-DC converter utilizing air-core inductors on package," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 4, pp. 838–845, 2005.

[8] K. K. Rangan, G.-Y. Wei, and D. Brooks, "Thread motion: fine-grained power management for multi-core systems," in *Proceedings of the International Symposium on Computer Architecture*, 2009, pp. 302–313.

[9] W. Kim, M. S. Gupta, G. yeon Wei, and D. Brooks, "System level analysis of fast, per-core DVFS using on-chip switching regulators," in *Proceedings of the International Conference on High-Performance Computer Architecture*, 2008, pp. 1–12.

[10] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman *et al.*, "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures," in *Proceedings of the International Symposium on Microarchitecture*, 2009, pp. 469–480.

[11] HP Labs, "Cacti, http://www.hpl.hp.com/research/cacti/," 2010.

[12] M. Namgoong, M. Yu, and T. Meng, "A high-efficiency variable-voltage CMOS dynamic DC-DC switching regulator," in *Proceedings of the International Solid-State Circuits Conference*, 1997, pp. 380–381.

[13] P. S. Magnusson, M. Christensson, J. Eskilson, D. Forsgren *et al.*, "Simics: a full system simulation platform," *IEEE Transactions on Computer*, vol. 35, no. 2, pp. 50–58, 2002.

[14] The Standard Performance Evaluation Corporation, "SPEC OMP2001, http://www.spec.org/omp/."

[15] C. Bienia, S. Kumar, J. P. Singh, and K. Li, "The PARSEC benchmark suite: characterization and architectural implications," in *Proceedings of the International Conference on Parallel Architectures and Compilation Techniques*, 2008, pp. 239–249.