Adaptive Thermal Management for 3D ICs with Stacked DRAM Caches

Dawei Li, Kaicheng Zhang, Akhil Guliani, Seda Ogrenci-Memik

Department of EECS, Northwestern University

{daweili2016, kaichengzhang2016, akhulguliani2016}@u.northwestern.edu,
seda@eecs.northwestern.edu

ABSTRACT

We describe an adaptive thermal management system for 3D-ICs with stacked DRAM cache memories. We present a detailed analysis of the impact of 3D-IC hotspot aggregation on the refresh behavior of the stacked DRAM-based L3 cache. We also present the consequence of the refresh variation on the overall system performance and cache energy consumption. Our analysis demonstrates that memory intensive applications are influenced more strongly by the DRAM refresh variation. We show that there is an optimal operating point where, with a reduced clock frequency, processor cores would actually recover any performance loss induced by DRAM refresh and at the same time the cache energy consumption could be optimized. We propose a low overhead run-time method that can identify the best CPU frequency modulation factor to cool the system to minimize accelerated refresh rates in the DRAM caches. Our system can provide a customizable trade-off between performance of the processor and energy savings of the memory.

Keywords

3D-IC, thermal management, DRAM, cache, low power design

1. INTRODUCTION

Stacking of memory layers with processors has emerged as the primary driver for 3D integration allowing high bandwidth and low latency access to memory. Products are already being offered and some are under development by many institutions and companies such as Micron, Tezzaron, and Fermilab [1-3]. DRAM is used prominently in 3D memory stacking due to its low cost and high capacity. A variety of DRAM-based stacked L2 and L3 cache architectures have also been proposed [4-8].

An important challenge for 3D Integrated Circuit (3D IC) design is thermal management [9, 10]. The two prominent factors that have contributed to the thermal issues of 3D-ICs are: (i) The high effective power density resulting from multiple overlapping high power density blocks; (i) Multiple chip layers in the vertical direction increasing the total thermal resistance in the path of heat transfer from the chip to the heatsink. Various attempts are made, such as repositioning processor dies to avoid overlapping hotspots, using microfluidic channels, thermal vias, and thermoelectric devices for better heat transfer [11-15]. Still,

thermal management of 3D ICs has many problems to tackle.

Specifically, stacking 3D DRAM as caches in close proximity to the processor die will result in the formation of volumetric hotspots throughout DRAM layers. In effect, the processor will imprint its own thermal hotspots onto the stacked caches and the DRAM layer's own thermal resistance will further aggregate these 2D hotspots into 3D volumetric hotspots. This can be particularly troublesome since DRAM data retention is sensitive to the temperature. In response to aggravated thermal conditions experienced by DRAMs, new designs deploy variable refresh rates [16]. On the other hand, when DRAM refresh rate changes due to temperature fluctuation, the effective latency of the DRAM varies as well and it can degrade the performance of the processor cores. The refresh rate of the DRAM directly affects the performance of the system as a refresh command can block a full rank and decrease memory parallelism within the system. For memory intensive applications, the performance degradation is significant. Our experiments showed that for a memory intensive benchmark, if the DRAM operates at a higher refresh rate, the Instructions Per Second (IPS) metric of the application drops by up to 19%, even though the processor cores are executing at their full speed throughout. Furthermore, increased refresh rate results in higher energy consumed by the memory system. Refresh energy is becoming a dominant fraction of the overall DRAM energy as the DRAM density and capacity scale. While other leakage and dynamic activity driven fractions scale with process nodes, duration of the refresh operation and the resulting energy consumption actually increases with increasing DRAM capacity.

In this paper, propose an adaptive thermal management scheme for a multi-core processor chip with 3D stacked DRAM L3 cache. We investigate the potential of proactively managing the frequency of cores in an attempt to control the thermal footprint they will impose on the DRAM caches. Our analysis reveals that a judicious scale-down of core frequencies ultimately benefits the overall system. It prevents the DRAM caches from heating up to the point of operating at their highest refresh rate and the resulting performance boost of the cache reflects back onto the cores, largely making up (up to 100% for memory intensive applications and even beyond in special cases) for the slowdown of frequency throttled cores. While performance degradation is avoided, energy savings in the DRAM cache due to less frequent refresh and improvement of its reliability is achieved "for free".

In summary, we constructed our management scheme based on three fundamental observations of DRAM behavior:

- i. The refresh control mechanism depends on the peak temperature inside the DRAM stack.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

DAC '17, June 18-22, 2017, Austin, TX, USA

© 2017 ACM. ISBN 978-1-4503-4927-7/17/06...\$15.00

DOI: <http://dx.doi.org/10.1145/3061639.3062197>

¹ This work was partially supported by the DOE Grant DE-SC0012531 and NSF Grant CCF-1422489.

- ii. Memory bound applications are particularly sensitive to the changing refresh rate. (Reflected in their IPS values)
- iii. Since the processor thermal behavior contributes more to the DRAM hotspot temperature than the DRAM itself, core throttling can effectively change the refresh rate of the DRAM.

In this paper, we demonstrate through detailed simulation and analysis the impact of thermal hotspots on different DRAM refresh rates, energy of DRAM, and overall performance of the system. If we can correlate an application's sensitivity to the L3 DRAM cache's refresh rates, then, we can control the processor dies' influence on the temperature to prevent the chip from reaching the threshold for triggering that corresponding refresh rate. Our experiments show that with a low overhead policy we are able to obtain up to 72.5% DRAM energy saving with opportunities to actually increase performance (as measured in terms of progress in the workload) in refresh rate sensitive applications.

Remainder of the paper is organized as follows: Section 2 presents related work. Section 3 presents our analysis of a 3D IC system and the potential effect of an adaptive management system. Section 4 presents our adaptive model and the decision mechanisms involved in choosing the frequency to achieve the best operating point. Section 5 presents our conclusion and discussion on how our system can be implemented.

2. RELATED WORKS

A plethora of planar DRAM thermal management designs exit in literature [21-23]. In order to alleviate thermal issues in 3D-IC memory systems, Meng et al. [17] proposed a runtime optimization policy for DVFS under given power and thermal constraints. This work does not consider the impact of varying refresh rates on the DRAM cache access latency. The system is operated at selected (V, f) points so as to not exceed a given power cap and/or temperature threshold. In our work, we provide a policy that is aware of the consequences of faster refresh and directly influences DRAM cache refresh energy while not sacrificing performance as compared to the peak frequency operation.

Zhang et al. [18] proposes a PRAM (Phase Change RAM) and DRAM hybrid system in 3D-IC. PRAM is capable of raising the heat capacity of the memory and capturing the thermal energy before a threshold temperature is crossed. While PRAM might be a potential candidate for a thermally challenging environment such as 3D-IC, we aim to address the immediate challenges of existing systems which are purely based on stacked DRAM structures. Tran et al. [4] proposed the use of 3D-DRAM along with external DRAM to alleviate bandwidth and traffic issues. In this work thermal issues were not addressed.

Our novel approach does not require the use of a new technology, nor does it make any changes in the current refresh rate policy/management. Moreover, our approach can be applied to conventional DRAM systems with trivial overhead. The unique nature of 3D-IC thermal issues can be alleviated through refresh-correlation aware core frequency modulation while any conventional refresh rate based performance and energy saving techniques for planar DRAM can be easily co-implemented with our policy to further benefit the DRAM layers.

3. SETUP FOR OFFLINE ANALYSIS

In order to characterize the solution space and the nature of the trade-off between performance and DRAM thermal state, we first

	System Specifications
Base CPU Frequency	2GHz
L1 DCache	16kB
L1 ICache	16kB
L2 Cache	2MB
L3 DRAM Cache	512MB
L3 # of Channels	2
L3 # of Ranks	1
L3 # of Banks	16
64ms Refresh Temperature	40°C-85°C
32ms Refresh Temperature	85°C-95°C
8ms Refresh Temperature	95°C-115°C

Table 1. System Specifications for CPU and DRAM.

Components	Area	Dimension	Thermal Conductivity (W/(m*K))
Processor Die (Silicon)	14.7mm x 10.08mm	704um Thickness	117.5 @ 100°C
DRAM Stack (Silicon)	14.7mm x 10.08mm	9um (per die) Thickness	117.5 @ 100°C
Heat Sink (Copper)	125mm x 78.5mm	20mm Height	385
Heat Spreader (Copper)	125mm x 78.5mm	1mm Height	385
Micro-bumps	14.7mm x 10.08mm	25um Thickness	0.63

Table 2. Geometrical and thermal properties of 3D-IC stack.

performed thorough thermal and architectural performance analysis of a 3D multicore IC with stacked L3 DRAM cache. To construct our 3D-IC model for transient thermal simulation, we used industry standard thermal simulator FloTHERM V11.0. We used three applications from NAS benchmark [19] to represent different possible memory behaviors. We also used cycle accurate simulator GEM5 and power simulator McPAT [20] to obtain the transient power traces for each benchmark. Our 3D-IC processor core is based on the Alpha-like (21364) core and our DRAM is based on Tezzaron's Octopus 8-Port DRAM design [2]. Our transient thermal 3D-IC model consists of one 8-core processor layer and five stacked DRAM cache layers. Table 1 shows the system specifications of our processor and L3 DRAM cache. The DRAM refresh specifications of Octopus spans from 40°C to 115°C (115°C is the maximum rated temperature and 40°C is ambient temperature). Beyond 115°C, the system will throttle (or shutdown) the processor until the system temperature is below

Figure 1. Cross sectional view of 3D-IC with DRAM stack.

Figure 2. 3D view of the 3D-IC showing 8-cores with a shared L2 cache, and a TSV connected 5-layer DRAM stack.

115°C. Until 85°C the DRAM refreshes every 64ms, between 85°C and 95°C it refreshes every 32ms, and beyond 95°C it refreshes every 8ms.

3.1 Benchmarks and Transient Power Traces

To quantify the effect of hotspot temperature on DRAM performances and refresh rates, we choose CG, IS, and MG as our benchmarks. We chose CG to represent applications that are relatively insensitive to rise in temperature and the L3 DRAM cache refreshing at the highest rate. The IPS of CG varies by 6.4% between the cases when, while the cores are operating at the highest frequency, the cache DRAM refreshes every 64ms versus every 8ms. However, note that examining such an application still helps us identify opportunities for DRAM energy optimization, if that parameter is of higher priority in a low power system. We chose MG and IS to represent applications that experienced a pronounced performance hit by 10% and 19%, respectively, due to temperature and refresh rate changes. Finally, the transient power traces of these benchmarks also consider temperature-dependent leakage power as part of total power for the temperature range between 40°C to 115°C.

Figure 3. Thermal map of the 3D-IC showing the formation of volumetric hotspot in the stacked DRAM layers.

3.2 3D-IC Stacked DRAM Model

The first layer of our 3D-IC is an 8-core processor die. Five additional DRAM dies are stacked on the surface of the processor die. Fig. 1 shows a cross sectional view of our 3D-IC model which is representative of a typical DRAM stack. The processor die was placed closer to the heat sink for heat to be transferred towards the heatsink first. Table 2 shows the geometrical and thermal properties used in our model. Note that each DRAM layer is 9um in height and in total the DRAM stack makes up 45um of the entire chip height. Fig. 2 shows the 3D view of the entire structure. Our 3D-IC model consists of eight cores and a shared L2 cache placed in the center to avoid concentrated power density. A TSV array connects the top processor die to all the bottom DRAM layers. The first DRAM layer is placed face to face with the processor layer while each additional DRAM layer is placed face to back with the first DRAM layer. Since the power consumption of each component depends on the current temperature (due to its own dynamic activity state and also leakage), our thermal model contains a temperature probe for every functional block in each. Our transient simulation takes the current temperature of the probe at each step and assigns the corresponding power values (including temperature dependent leakage) for each functional unit. This ensure that we capture an accurate and complete transient behavior of our system. The memory layer has the same area as our processor layer. There are 16 banks per layer, corresponding to the eight processors.

3.3 Transient Thermal Analysis

Hotspots from the processor die influence the adjacent DRAM cache layers to form volumetric hotspots. This phenomenon can be seen in our 3D-IC thermal simulation presented in Fig. 3. Each of the hotspots from the processor die also appears on the DRAM cache layer, indicating a close correlation between the processor power density and the DRAM temperature. The next step is to investigate the correlation between accelerated DRAM cache refresh rate and the ultimate performance of the processor cores.

3.3.1 DRAM Sensitivity to Temperature

In the next stage of our analysis, we focused on the impact of temperature-induced variation of refresh rate on the energy of the DRAM cache memory and the performance of the processor. We define performance as the rate at which a workload executing on a processor progresses. More specifically, we measure it in terms of Instructions Per Second (IPS) for a given workload.

Fig. 4 shows the relative percentage of the total execution time that the DRAM cache spent in each of the three refresh states for different frequency settings of the processor, in a range from peak frequency 105% to 85% (100% is the default frequency). We observe that for all applications when cores are running at peak frequency, the heat dissipation of the processor cores progressively forces the DRAM cache into a temperature range above 85°C first and then ultimately beyond 95°C. As the core frequency is reduced the relative fraction of time spent at a temperature beyond 95°C (hence time spent refreshing every 8ms) shrinks and beyond a certain setting, the chip never exceeds that threshold, thereby, mostly remaining in the 32ms refresh regime. While this general trend is observed for all applications. The relative sensitivity of individual applications and the optimal frequency setting for completely avoiding the highest refresh rate varies.

In the following, we will present the consequences of spending more or less of the total execution time at peak refresh rate on the cache energy and core performance. Fig. 5 shows the DRAM

Figure 4. Percentage of time DRAM cache spent in each refresh interval setting with varying frequency modulation.

energy saving from applying CPU frequency scaling in the same range from 105% to 85% (100% is the default frequency). We define the *energy saving optimum point* of the benchmark as the frequency modulation that can obtain the highest energy saving percentage with respect to the baseline, which is operating the cores at peak frequency (100%). MG reaches its energy optimum point of 73.5% energy saving at frequency modulation of 92%. CG reaches its energy optimum point of 86% energy saving at frequency modulation of 86%. IS reaches its energy optimum point of 83.8% energy saving at frequency modulation of 88%.

Finally, we present the impact of DRAM cache refresh state on performance in Fig. 6. Similarly, we define the *performance optimum point* of the benchmark as the frequency modulation that can obtain the highest performance relative to the baseline (cores always running at default 100% peak frequency setting). The figure shows that the peaks for maximum performance are changing depending on how aggressively the heat output would drive the DRAM into higher refresh rates and on how sensitive the application is to the DRAM cache's varying latency.

In all benchmarks, properly scaling the frequency can in fact achieve gains in performance compare to the baseline 100%. In order to understand the benchmark behaviors under frequency modulation, we examine two regions: from 100% to 105% and from 100% to 85%. In the first region of interest, examining Fig. 5 it is clear that memory insensitive benchmark CG does not

Figure 5. DRAM performance under frequency modulation.

Figure 6. System energy saving under frequency modulation.

experience significant degradation in performance. At 105% frequency modulation, CG performance is rated at 104.5%. On the other hand, both memory sensitive benchmarks MG and IS show degradation in performance. At 105% frequency modulation, IS can only reach 101.2% in performance, and MG can only reach 99.8% in performance. This trend is also observed in Fig. 6 where CG's DRAM energy saving at 105% frequency modulation is -7% while IS and MG shows -18% and -88% respectively.

In the second region of interest, Fig. 5 indicates that CG, IS, MG can reach peak performance of 103.1%, 112.9%, 101.7%, respectively. We also observe the sharpest slope changes in the energy saving corresponding to when the DRAM refresh state changes. Fig. 4 shows clearly that the frequency modulation that changes the DRAM time spent from 8 ms to 32 ms refresh state corresponds to the highest performance increase in Fig. 5 and the largest DRAM energy saving in Fig. 6. The same behavior is observed for these cases for refresh state 32 ms to 64 ms although not as significant. In general, we define *our recommended optimum region of operation* as the frequency modulation range with performance at least as good as the baseline performance and positive DRAM energy saving percentage. Hence, this region is bounded by the frequency modulation with the highest performance and the highest DRAM energy saving (with performance at least 100%). For CG, this frequency modulation range is from 98% to 95%. For IS, it is from 99% to 86%. For MG, it is 99% to 97%.

In summary, we observe that if we can prevent the DRAM from reaching the highest refresh rate, we will be able to achieve energy savings. The performance is more complex as it depends also on the sensitivity of the benchmark to refresh rate changes. In general, the optimum point for energy and performance is the same when the application is sensitive to DRAM refresh rate. This complex behavior needs to be incorporated into an adaptive thermal management scheme in order to harness the potential for optimization of performance and energy saving.

4. ADAPTIVE THERMAL MANAGEMENT

Memory refresh rate directly affects the availability of DRAM caches to regular requests from upper-level caches. Increasing memory refresh rate will increase the average latency of the memory requests. Such increase will become more severe when the memory is already under heavy load by a memory intensive application. In the following section, we describe our proposed policy, which draws its roots from the observations made in the previous section.

4.1 Design and Implementation

We propose a policy that picks a frequency setting to control the heat output of the cores and its influence on the cache layer’s temperature so that the memory would refresh less frequently. However, we aim to configure our policy to reflect respective sensitivity for an application to cache latency.

The essence of our management policy is to identify an operating condition where the effective slowdown of cores due to pure frequency throttling would be balanced out by the performance recovered thanks to the DRAM cache’s improved access latency. Our scheme consists of two main components: (i) system profiling to assess the sensitivity of an application to the L3 cache’s latency at different refresh rates; (ii) a scale-and-check strategy, which monitors proximity of the chip temperature to the threshold for different refresh rates. We describe these components in more detail in the following subsections.

4.1.1 System Profiling

Before we can scale the CPU frequency, we need to assess the performance metrics under different memory refresh settings. As the refresh is being executed at 64ms intervals by the memory controller (through the REF commands directed at the DRAM cache), the CPU can order the memory controller to send new REF commands at a higher frequency, emulating the higher refresh rates even though the temperature threshold has not been exceeded. The high refresh rate is maintained for a single segment of 64ms duration to make sure that every row in the memory is refreshed. We count the number of instructions executed during this 64ms period and use the IPS (instruction per second) to represent performance. Thus, we have a natural way to profile the

Temperature	Frequency Modulation Scaling
< 80 C	100%
80 – 85 C	IPS32 / IPS64
85 – 95 C	IPS8 / IPS32
> 95C	100%

Table 3. Adaptive Thermal Management.

performance at 64ms, 32ms, and 8ms refresh intervals. We denote IPS64, IPS32, and IPS8 for profiled the IPS metrics at 64ms, 32ms, and 8ms refresh intervals. The time spent in profiling is negligible and therefore there would be no performance loss due to the profiling. Thermal response is slow compared to frequency response. Thus we invoke the profiling process when the temperature reaches 80°C, i.e., 5°C below the critical temperature to change into 32ms refresh rate (85°C). It can be safely assumed that the processor would not reach 85°C before the completion of our extremely short profiling segment.

4.1.2 CPU Frequency Modulation

Our management policy makes use of a run-time prediction on the amount of performance that can be “recovered” by improving the DRAM cache latency (i.e., preventing it from entering high refresh rates). This prediction is derived from the run-time profiling of an application as described in Section 4.1.1. The measured variation in IPS between different “emulated” refresh states while operating at the peak processor frequency is used as the prediction vector to drive the specific decisions of our policy. Our analysis presented in Section 3 led to the conclusion that pursuing aggressive frequency scaling beyond the relative drop of performance directly induced by DRAM refresh will inevitably result in performance degradation. Hence, our policy utilizes these predicted rates of performance recovery as a guideline. It starts frequency scaling when the DRAM temperature reaches 80°C, right after we finish the system profiling stage. We develop a simple yet effective frequency-scaling scheme based on a temperature-based look up table shown in Table 3. This table is created at run-time according to the IPS measurements collected during profiling. When the temperature reaches 80°C, our algorithm will scale down the CPU frequency by IPS32/IPS64. In this stage, if the cores and the DRAM cache remain under 85°C, the refresh rate will continue to remain at 64 ms interval. In the

		Performance %	% of DRAM Energy Saving	Frequency Modulation	% of Time spent @64ms	% of Time spent @32ms	% of Time spent @8ms
CG	Default	100%	0%	100% @ 40°C	2.5%	10.8%	86.7%
	OPT	103.1%	72.5%	98% @ 40°C	2.8%	97.2%	0%
	Delay80_OPT	105.1%	72.5%	98% @ 80°C	2.8%	97.2%	0%
	Delay90_OPT	103.1%	72.5%	98% @ 90°C	2.4%	97.6%	0%
	Adaptive	105.3%	72.5%	99.3% @ 80°C 94.3% @ 85°C	2.5%	97.5%	0%
IS	Default	100%	0%	100% @ 40°C	5.6%	23.8%	70.6%
	OPT	112.9%	68.6%	99% @ 40°C	6.1%	93.9%	0%
	Delay80_OPT	114%	68.6%	99% @ 80°C	6%	93.4%	0%
	Delay90_OPT	113%	68.5%	99% @ 90°C	5.6%	94.4%	0%
	Adaptive	114%	68.6%	98.1% @ 80°C 82.5% @ 85°C	6.4%	93.6%	0%
MG	Default	100%	0%	100% @ 40°C	7.4%	63.4%	29.2%
	OPT	101.7%	49.2%	99% @ 40°C	8%	92%	0%
	Delay80_OPT	102.7%	49.2%	99% @ 90°C	8.5%	91.5%	0%
	Delay90_OPT	101.9%	49.1%	99% @ 90°C	7.8%	92.2%	0%
	Adaptive	103%	52.5%	99.2% @ 80°C 91.4% @ 85°C	21.3%	78.7%	0%

Table 4. Comparison of different frequency modulation policies and their effects on the system performance and DRAM energy.

case that the frequency scaling cannot prevent temperature from reaching 85°C, the refresh state will reach 32 ms interval. Upon reaching 85°C our policy scales the CPU frequency by IPS8/IPS32 to introduce more aggressive control on the core's heat dissipation, yet, *the scaling is not purely chosen to control peak temperature, but rather it is specifically based on the predicted performance recovery given the application-specific sensitivity on the DRAM refresh rate*. If the chip temperature increases up to 95°C under this new frequency setting, our policy determines that any further scaling will not be able to move the DRAM cache out of the 8ms refresh interval and any sustained frequency scaling of the cores will only serve slowdown with no DRAM energy savings. Then, the policy sets the processor frequency back to 100% to recover the baseline performance since it is apparent that we cannot prevent the DRAM cache from going into the 8ms refresh state, without sacrificing performance. We note that, alternative implementations of our policy could indeed pursue more aggressive scaling, if the system level priorities are strictly inclined towards higher memory energy efficiency allowing performance to be a secondary objective.

4.2 Results

We present several frequency modulation policies including our adaptive thermal management policy in Table 4 for each of the three benchmarks. The baseline setting is **Default** with 100% of the original CPU frequency and thus there is no degradation in performance and no energy saving. **OPT** is a frequency modulation setting that uses the optimum frequency setting derived from the results presented in Fig. 4, Fig. 5, and Fig. 6. This setting represents the left boundary of our recommended optimum region of operation and hence the frequency modulation with the highest performance. **Delay80_OPT** and **Delay90_OPT** uses the optimum frequency modulation from OPT, but the scaling is applied at the start of 80°C mark and 90°C mark at runtime, respectively. These two settings show the effect of different initiation points for frequency modulation. Note that for cases Delay80_OPT, Delay90_OPT, the performance gain is greater than the OPT setting. In addition, Delay80_OPT achieves higher performance than Delay90_OPT. Thus it is better to let the applications run until it reaches 80°C to start frequency modulation. Finally, the policy marked **Adaptive** is our proposed scheme that sets the frequency modulation to IPS32/IPS64 when the processor reaches 80°C. If the temperature continues to rise, the subsequent scaling of IPS8/IPS32 is applied at the start of 85°C mark. If the temperature rises above 95°C then the scaling did not prevent the DRAM from reaching 8ms refresh state. In this case, our system will set the CPU frequency to full speed again without scaling. For benchmarks MG and IS, our adaptive policy can obtain 52.5% and 68.6% DRAM cache energy saving while achieving performance gain of 103% and 114% respectively. For workload CG, we can obtain 72.5% energy saving while achieving performance gain of 105%.

5. CONCLUSION/DISCUSSION

Cores will largely influence the thermal state of the stacked DRAM cache in a 3D-IC. This is especially important given the DRAM refresh rate depends on temperature. We propose an adaptive thermal management scheme that utilizes the CPU frequency modulation in order to maintain the DRAM cache in the lower refresh rate states. The optimum point obtained by our adaptive thermal management policy is very close to the optimum point reported from sweeping all frequency values. In addition, this policy will not interfere with DRAM refresh rate management

mechanism already in place, as our policy adjusts the CPU frequency-scaling factor only.

6. REFERENCES

- [1] Hybrid Memory Cube: <http://www.hybridmemorycube.org/technology.html>.
- [2] Tezzaron Semiconductor, "Octopus 8-port DRAM for die-stack applications", May 2010.
- [3] D. Li, et al., "A Methodology for Power Characterization of Associative Memories", IEEE ICCD, 2015.
- [4] L.-N. Tran, et al., "Heterogeneous Memory Management for 3D-DRAM external DRAM with QoS", IEEE ASP-DAC, 2013.
- [5] G.-H. Loh, "3D-Stacked Memory Architectures for Multi-Core Processors", IEEE ISCA, 2009.
- [6] H. Sun, et al., "3D DRAM Design and Application to 3D Multicore Systems", IEEE CD & CASS, 2009.
- [7] D. Jevdjic, S. Volos, B. Falsafi. "Die-stacked DRAM caches for servers: hit ratio, latency, or bandwidth? have it all with footprint cache." ACM SIGARCH Computer Architecture News 41.3 (2013)
- [8] N. M. Tshibangu, et al. "Design of controller for L2 cache mapped in Tezzaron stacked DRAM." IEEE 3D Systems Integration Conference (3DIC), 2013
- [9] P. Jain, et al., "Thermal and Power Delivery Challenges in 3D ICs", Three-Dimensional Integrated Circuit Design – Chapter 3, Y. Xie et al.(eds.), Integrated Circuits and Systems, Springer, 2010.
- [10] S. Ogrenici-Memik, Heat Management in Integrated Circuits: On-chip and System-level Monitoring and Cooling. IET, 2015.
- [11] B. Shi, et al. "Hybrid 3D-IC cooling system using micro-fluidic cooling and thermal TSVs." IEEE ISVLSI, 2012.
- [12] J. Cong, et al., "Thermal-Aware Cell and Through-Silicon-Via Co-Placement for 3D ICs", IEEE/ACM DAC, 2011.
- [13] D. Li, et al., "Integrating Thermocouple Sensors into 3D ICs", IEEE ICCD, 2013
- [14] Y. Pi, et al., "Microfluidic Cooling for distributed hot-spots", IEEE ECTC, 2016.
- [15] I. Chowdhury, et al., "On-Chip Cooling by Superlattice-Based Thin-Film Thermoelectrics", Nature Nanotechnology, p. 235-238, 2009.
- [16] I. Bhati, et al., "DRAM Refresh Mechanisms, Penalties, and Trade-Offs", IEEE Transactions on Computers, 2016.
- [17] J. Meng, et al., "Optimizing Energy Efficiency of 3-D Multicore Systems with Stacked DRAM under Power and Thermal Constraints", IEEE DAC, 2012.
- [18] W. Zhang, T. Li, "Exploring Phase Change Memory and 3D Die-Stacking for Power/Thermal Friendly, Fast and Durable Memory Architectures", IEEE PACT, 2009.
- [19] NAS Division. NAS parallel benchmarks. <http://www.nas.nasa.gov/Resources/Software/npb.html>.
- [20] S. Li, et. al., "McPAT: An integrated power, area, and timing modeling framework for multicore and many core architectures", IEEE MICRO, 2009
- [21] S. Liu, B. Leung, A. Neckar, S. Ogrenici-Memik, G. Memik, N. Hardavellas, "Hardware/software techniques for DRAM thermal management", IEEE HPCA, 2011
- [22] S. Liu, S. Ogrenici-Memik, Y. Zhang, G. Memik, "A power and temperature aware DRAM architecture", IEEE DAC, 2008
- [23] S. Liu, S. Ogrenici-Memik, Y. Zhang, G. Memik, "An approach for adaptive DRAM temperature and power management", IEEE TVLSI, 2010