An Approach for Adaptive DRAM Temperature and Power Management

Song Liu, Yu Zhang, Seda Ogrenci Memik, and Gokhan Memik

***Abstract*—High-performance DRAMs are providing increasing memory access bandwidth to processors, which is leading to high power consumption and operating temperature in DRAM chips. In this paper, we propose a customized low-power technique for high-performance DRAM systems to improve DRAM page hit rate by buffering write operations that may incur page misses. This approach reduces DRAM system power consumption and temperature without any performance penalty. We combine the throughput-aware page-hit-aware write buffer (TAP) with low-power-state-based techniques for further power and temperature reduction, namely, TAP-low. Our experiments show that a system with TAP-low could reduce the total DRAM power consumption by up to 68.6% (19.9% on average). The steady-state temperature can be reduced by as much as 7.84 °C and 2.55°C on average across eight representative workloads.**

***Index Terms*—DRAM, power, temperature.**

I. Introduction

Technological advances in microprocessor architectures enable high performance with an underlying assumption on increasing utilization of memory systems. On the other hand, increasing memory densities and data rates lead to higher operating temperatures in DRAM systems. Moreover, several techniques have been proposed to place DRAM closer to processor cores, such as 3-D ICs [10], and embedded DRAM [5]. With increasing power consumption and closer physical proximity to hot processor cores, modern DRAMs are operated under increasing temperatures. Prior studies have shown that DRAM temperature control has become a practical and pressing issue [4].

In this paper, we propose a DRAM architecture enhancement, which could harvest the largest peak temperature reduction without incurring any performance overhead. Specifically, we propose a customized method to reduce DRAM power consumption by improving DRAM page hit rate. Moreover, higher page hit rate also leads to less average DRAM access latency and thus improves system performance.

In our previous works, we have designed and analyzed the page-hit-aware write buffer (PHA-WB) [12] and the throughput-aware PHA-WB (TAP) [11]. PHA-WB provides a buffering mechanism to hold write operations that may cause a page miss. The TAP scheme was designed to dynamically adjust the tradeoff between the aggressiveness of the power optimization mechanism at the expense of more storage for buffering the data and orchestrating the buffer–DRAM coordination.

In this paper, we extend our work in two main aspects.

We take DRAM refresh operations into consideration. Experiments show that refresh operations have a strong impact on DRAM page hit rate. However, this impact decreases as DRAM traffic increases.

Manuscript received August 01, 2008; revised January 23, 2009. This work was supported in part by the US Department of Energy (DOE) under Award DEFG02-05ER25691 and in part by the National Science Foundation (NSF) under Awards CNS-0546305, CCF-0747201, CNS-0720691, IIS-0613568, CCF-0541337, CNS-0551639, IIS-0536994, CCF-0444405. The work of G. Memik was supported in part by Wissner-Slivka Chair funds.

The authors are with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: sli646@eecs.northwestern.edu; yzh702@eecs.northwestern.edu; seda@eecs.northwestern.edu; memik@eecs.northwestern.edu).

Digital Object Identifier 10.1109/TVLSI.2009.2014842

We also extend TAP with low-power-state-based techniques into TAP-low. We demonstrate that PHA-WB can actually increase the utilization of low-power states. PHA-WB also reduces average read delay by improving page hit rate, which reduces the performance penalty of low-power-state-based techniques.

Our experiments show that the TAP-low approach reduces the total DRAM system power consumption by as much as 68.6% (19.9% on average) and DRAM steady-state temperature by as much as 7.84 °C (2.55°C on average) for eight different workloads based on 20 SPEC CPU 2000 benchmarks running on a four-core CMP [18].

The remainder of this paper is organized as follows. Section II discusses related work. Our proposed technique is described in Section III. Section IV describes TAP-low, which is a combination of TAP and a low-power-state-based technique.Section V presents the experimental methodology and results. Finally, we conclude this paper in Section VI.

II. Related Work

DRAM power consumption can comprise a large portion of the total system power. Modern DRAMs provide various power-saving states. Various low-power DRAM techniques focus on utilizing these idle states efficiently to achieve the best energy-delay product [2], [3], [6]. Our goal, however, is to tie the active periods of DRAM operation to power consumption. These low-power-state-based techniques can be used as complementary to our approach.

Memory controller reordering is a widely used technique in stream processors [7], [15]. In these systems, the memory controller reorders memory accesses, so that there are more chances to use efficient page and burst modes. On the other hand, our technique, which targets general-propose processors with write-back cache and fully buffered dual inline memory module (FB-DIMM), is a further enhancement for burst-accessed DRAM.

Dynamic thermal management (DTM) of DRAM has become a pressing issue in mobile systems [4]. In order to cool down the DRAM while keeping the performance penalty small, Lin *et al.* [8] proposed adaptive core gating and dynamic voltage and frequency scaling (DVFS) to CMP systems. However, DTM and DVFS are known to introduce system performance penalties. We refer to these techniques as memory-traffic-control-based temperature-aware techniques since they handle DRAM thermal emergencies by reducing DRAM access density.

Existing power- and temperature-aware techniques focus on two special cases. Power-state-based techniques are designed for applications with fewer DRAM accesses. On the other hand, memory-traffic-control-based techniques are applicable when DRAM traffic is heavy. Our proposed technique is designed for the latter case. However, instead of applying thermal management after the DRAM is overheated, our technique prevents potential thermal emergencies by improving page hit rate and thus reducing power consumption.

Shim *et al.* analyzed the impact of different memory mapping strategies on DRAM page hit rate and, thus, DRAM power consumption and performance [16]. Their experiments showed that different memory mapping strategies could lead to significant change in page hit rate. While this technique focuses on a smart memory mapping strategy, our hit-aware buffering mechanism aims to fully utilize locality of memory accesses.

To the best of our knowledge, the power optimization technique by Lin *et al.* [9] could be considered as the closest related work. They introduce a prefetching cache in the advanced memory buffer (AMB) to improve performance and reduce power consumption. However, they use close-page mode as the baseline system. In high-performance DRAM systems, open-page mode is a more likely choice, in which

the power and temperature benefit of prefetching would be minimal. Our technique is devised upon and evaluated with a baseline system in high-performance open-page mode.

III. Adaptive Management for Power and Temperature

This section reviews our previous work for adaptive DRAM temperature and power management. We first explore potential savings in DRAM power consumption. Then, we outline PHA-WB[12] and an adaptive enhancement, namely, the throughput-aware PHA-WB (TAP) [11].

A. DRAM Power Modeling

DDR2 DRAM power consumption has three main components: background power, activate power, and read/write power [13]. Background power is the power consumption without any accesses. Different power states have different background powers. Activate power is the power consumption of loading one row to the I/O buffer. Read/write power is the power consumption of reading or writing data stored in the I/O buffer. Reduction in read/write power occurs at the expense of performance. In order to reduce power with no performance penalty, we focus on reducing the activate power.

In open-page mode, activate power is a function of page hit rate. In other words, higher page hit rate means that the same number of read and write operations could be completed with fewer activates. Therefore, improving page hit rate could achieve power benefit without hurting performance.

B. PHA-WB

PHA-WB [12] is a buffer placed between the memory controller and the DRAM chips. Read operations pass through the buffer without delay. On the other hand, write operations that are not targeting an activated row will be buffered. PHA-WB checks the target address of each operation and maintains a table of the activated rows in each bank. When an operation accesses the DRAM, the target address is broadcast in PHA-WB. Buffered write operations with matching rows will access the DRAM after this operation with a page hit.

Note that, although PHA-WB only buffers write operations, it also improves the hit rate of read operations and thus reduces the average read delay.

C. TAP

The number of entries in PHA-WB is a tradeoff between buffer power overhead and DRAM power savings. In order to achieve better power savings and more aggressive control over the power density of the DRAM, we have introduced TAP[11], which dynamically chooses the size of the write buffer according to different DRAM access patterns.

TAP employs a performance counter to track the number of clock cycles for every k consecutive DRAM accesses. More cycles for k consecutive DRAM accesses mean lower DRAM throughput. Moreover, when the processor switches between two different applications or two distinct phases of an application, the size of the buffer could also be switched. TAP utilizes clock gating to modify the number of activated entries in the write buffer.

We have presented the structure of TAP in our previous work[11].Fig. 1 further shows the structure of the write buffer in TAP. Compared with a regular cache, the write buffer is modified as follows. First, the tags are divided into row and column addresses to represent row matches and address matches. Second, the buffer entry lines are put into three groups with 16, 16, and 32 entries (2, 2, and 4 entries are drawn in the figure for simplicity), respectively. Each group has an extra "in use" bit denoting whether this group is currently used by the system. Finally, latches are added at the input of address comparators (shown as gray blocks). When the corresponding entries are neither in use nor valid, the inputs of these comparators are latched to save energy. When the size of the buffer is reduced, some buffer entries are valid but not "in use." The system takes all these entries as read match.

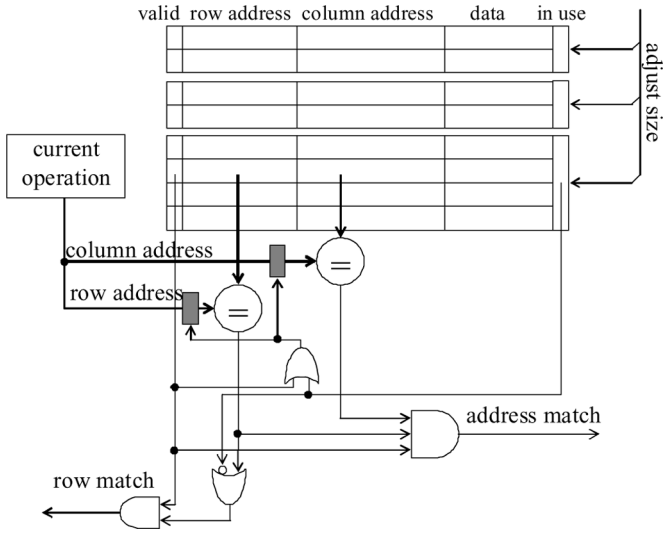

Fig. 1. Circuit-level diagram of the write buffer.

IV. Incorporating Adaptive Active State Optimizations With Low-Power-State-Based Techniques

Low-power-state-based optimization is widely used in battery-powered systems to tradeoff system performance for longer battery life. In this section, we will illustrate that, with a minor modification, our proposed TAP could utilize low-power state to further reduce DRAM power consumption at the cost of a slightly longer average DRAM read delay. Compared with a low-power-state-based technique, our proposed TAP combined with low-power-state extension (TAP-low) could achieve more power reduction for similar performance degradation.

In a system with a low-power-state-based technique, when the memory system is idle for a given period, the DRAM chips switch to low-power state. The DRAM chips have to switch to active state before later accesses, regardless of whether the subsequent access is a read or a write operation. However, as we have discussed earlier, write operations could be delayed without hurting system performance. It is not necessary to switch to active state for each write operation.

Fig. 2 shows a simple example of incorporating the write buffer with a low-power-state-based technique. In Fig. 2, "R" and "W" indicate the execution times of read and write operations, respectively. The shadowed blocks indicate that the DRAM is in low-power state. Fig. 2(a) shows a system with a low-power-state-based technique. The DRAM operations inFig. 2(a) are executed as soon as possible. On the other hand, Fig. 2(b) shows a system employing TAP-low. For the same memory requests in Fig. 2(a), the system in Fig. 2(b) could finish these requests with less switching between different power states. Therefore, the system would enjoy more power savings in low-power state. Note that the second read operation in Fig. 2(b) suffers more delay than that in Fig. 2(a). This is because it takes extra time for the DRAM chips to switch to active state. However, since PHA-WB could reduce the average read delay by improving the page hit rate of read operations, the final average read delay is similar between these two systems.

TAP-low has a similar system structure as TAP. The only difference is in the control logic. First, in TAP-low, the write operation will be

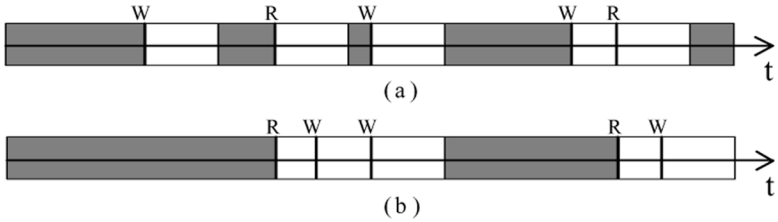

Fig. 2. Example of incorporating PHA-WB with a low-power-state-based technique.

TABLE I
WORKLOAD MIXES

Workload	Benchmarks
W1	swim, swim, swim, swim
W2	swim, lucas, applu, mgrid
W3	wupwise, apsi, fma3d, facerec
W4	swim, lucas, wupwise, apsi
W5	swim, lucas, sixtrack, galgel
W6	equake, gap, gcc, wupwise
W7	ammp, apsi, vpr, parser
W8	mcf, vortex, mesa, gzip

sent to the write buffer when the DRAM chips are in low-power state. Second, before the DRAM chips switch from active state to low-power state, the system releases some buffer entries by issuing some buffer operations. These entries are used to buffered write operations when the DRAM is in low power state.

V. EXPERIMENTAL RESULTS

In this section, we first introduce the power and thermal models used in our evaluation. Then, we outline our experimental methodology. Finally, we present the results demonstrating the effectiveness of our technique.

A. Power and Thermal Modeling

We use a DRAM power model based on the data published by Micron Technology, Inc.[14]. We utilize an RC-analogy-based thermal model for the DIMM, which is similar to the model presented by Lin *et al.* [8]. The DRAM system is modeled at the chip level as the DRAM chips and the AMB. We present static temperature values for each workload. We also adopted the AMB power model used by Lin *et al.* [8]. We have estimated the power consumption of PHA-WB by modeling the broadcast structure and the buffer array using CACTI 3.2 [17]. The power overhead of PHA-WB has been considered in our evaluation.

B. Experimental System Configuration

We used M5 simulator [1] as our architectural simulator to extract memory read and write traces for SPEC CPU 2000 applications. We assume a CMP with four processors and 1-GB DDR2 FB-DIMM. Each processor executes one of the SPEC applications (the workloads are described in Table I). The major parameters for the processor and memory are listed in Table II. Then, we analyze the DRAM page hit rate under different configurations with and without PHA-WB. A System Power Calculator [14] is used to calculate DRAM power consumption. Finally, we calculate AMB power and static DRAM system temperature for each trace based on these calculations.

Different DRAM refresh periods would result in different page hit rates. For the first set of our experiments, we assume that the refresh period is 64 ms. We analyze the impact of refresh periods on page hit rate in Section 5.C.

We evaluate PHA-WB with 16, 32, and 64 entries, as well as TAP and TAP-low. In TAP, the adaptive adjustor monitors the total clock cycles for ten consecutive DRAM accesses. If these accesses take less than 500 cycles, 64 buffer entries are enabled; if between 500 and 1000 cycles, 32 buffer entries are enabled; if between 1000 and 4000 cycles, 16 buffer entries are enabled; and if ten accesses take more than 4000 cycles, all the buffer entries are disabled, and the system works as if there is no write buffer. In TAP-low, a low-power-state-based technique is enabled when the number of active buffer entries is smaller than 64.

C. Results

We evaluate the impact of PHA-WB on page hit rate, power consumption, and the static temperature of DRAM. We performed a comparison of PHA-WB structures with different number of entries for all workloads. Then, we compare the performance of a system with only a low-power-state-based technique and that with TAP-low.

Fig. 3 shows the page hit rates of different workloads on different system configurations. We observe that page hit rate increases by up to 18.75% (9.63% on average) based on comparison of an architecture without PHA-WB and one with a 16-entry PHA-WB. The maximum and average page hit rate improvements for the 64-entry PHA-WB are 25.43% and 13.29%, respectively. The introduction of even a small-sized PHA-WB makes a significant impact compared to the base case. As we increase the PHA-WB size further, we observe a gain; however, the benefits diminish because most of the writes that access a later read row are captured, even with a smaller PHA-WB.

Fig. 4 shows the percentage power savings under different PHA-WB sizes, relative to the baseline system. Power saving is a function of page hit rate and memory throughput. Some applications with high improvements in page hit rate fail to yield significant power savings because the DRAM throughput is very low. The average and peak page power savings for TAP are 22.94% and 6.23%, respectively. TAP is entirely disabled when running workload W8. Hence, the power consumption for our architecture equals that of the base case. Similarly, the steady-state temperature and read delay (shown in Figs. 5 and 6) are identical to that of the base case. Fixed-size PHA-WBs are not disabled when running workload W8, thus yielding negative power and temperature reduction (i.e., degradation in power and temperature). TAP-low yields significant power savings for the last three workloads. The average power reduction achieved by TAP-low is 19.87%. Fig. 5 shows the reduction in DRAM temperature using different PHA-WB sizes compared to the base case. TAP-low reduces the DRAM steady-state temperature by 2.55°C on average.

Our proposed PHA-WB also improves the performance by reducing the average delay of read operations. We analyze how our technique influences the average read delay (duration of time between when the AMB assigns a read operation and when the data are available in the DRAM). Fig. 6 shows the reduction in average read operation delay for each design. The average delay is reduced by up to 14.34% (8.29% on

TABLE II
EXPERIMENTAL CONFIGURATION OF PROCESSOR, MEMORY SYSTEM, AND PHA-WB

Parameters	Values
Processor	4-core, 2.5GHz
Instruction Cache (per core)	64KB, 2-way, 64B line, 1 cycle hit latency, 16 MSHRs
Data Cache (per core)	64KB, 2-way, 64B line, 2 cycle hit latency, write-back, 16 MSHRs
L2 Cache (shared)	16MB, 8-way, 128B line, 13 cycle hit latency, write-through, 40 MSHRs
FB-DIMM	1GB, 512Mb per chip, 8 DQs per chip, 667MHz, 250 cycle baseline delay
DDR2 DRAM Chip	4 banks per chip, 16384 rows per bank, 256 columns per row, 4 Bytes per column
Burst Mode	Burst length of 8
Major DRAM Timing Parameter	Active to Read tRCD=15ns, Read to Read Data tCL=15ns, Precharge to Active tRP=15ns
No. of 64 Byte Entries in PHA-WB	16/32/64/adaptive

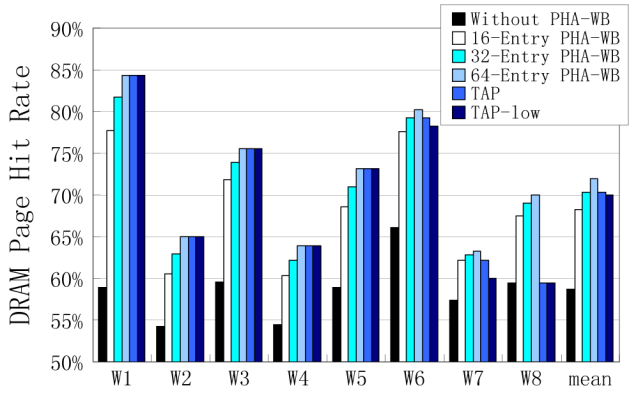

Fig. 3. DRAM page hit rates for different workloads with different PHA-WBs.

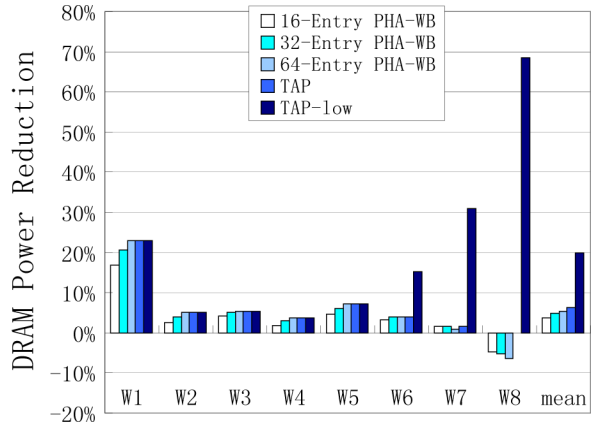

Fig. 4. DRAM power reduction for different workloads with different PHA-WBs.

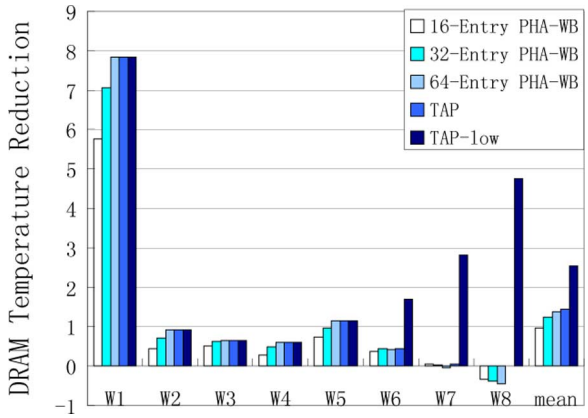

Fig. 5. DRAM temperature reduction for different workloads with different PHA-WBs.

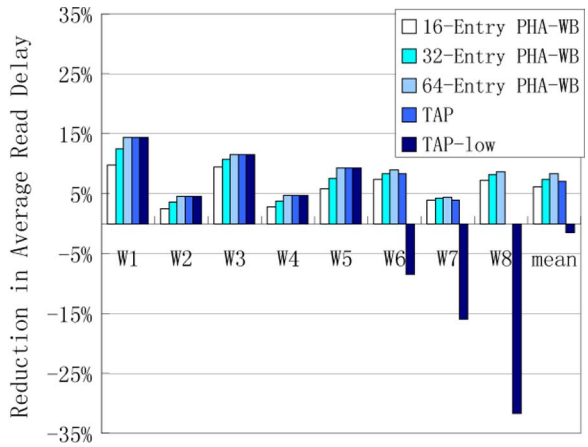

Fig. 6. Reduction of average read delay for different workloads with different PHA-WBs.

average across all benchmarks) for the 64-entry PHA-WB. TAP is designed for maximum power and temperature reduction. The 64-entry PHA-WB achieves more reduction in average read delay than TAP. Hence, when compared to the biggest fixed-size PHA-WB, TAP makes a tradeoff between performance improvement and reduction in power consumption. TAP-low is a further tradeoff between power and performance.

Fig. 7 shows page hit rates under different refresh periods. The page hit rate of W1 is similar in different refresh periods for both the baseline system and TAP-low. This is because the number of read and write operations is much higher than the number of refresh operations and the impact of refresh operations is negligible. The page hit rates of other workloads are slightly different under different refresh periods. However, refresh operations do not hurt the benefit of our proposed technique.

We also compare the efficiency of TAP-low and that of a simple low-power-state-based technique on W6. On the one hand, the presence of PHA-WB introduces more opportunities for using low power state. TAP-low achieves much higher power reduction (15.1%) than the low-power-state-based technique (7.5%). On the other hand, the average delay is similar among these two approaches. The average delay of TAP-low (7.7%) is even lower than the simple low-power-state-based technique (8.4%). Therefore, compared with the simple low-power-state-based technique, TAP-low achieves higher power reduction on relatively small or no performance degradation.

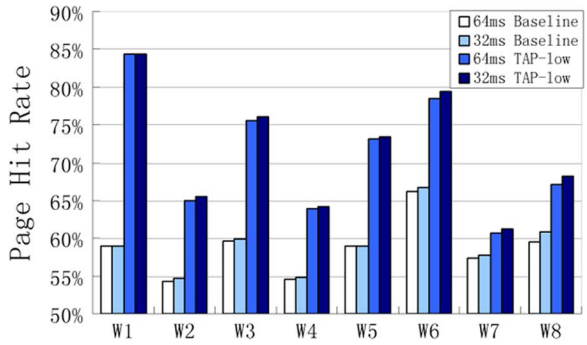

Fig. 7. Page hit rates for different workloads under different refresh periods.

VI. Conclusion

We have proposed and evaluated an adaptive power and thermal management scheme for DRAM systems. Our proposed technique, namely, TAP-low, which incorporates TAP with a low-power-state-based technique, reduced the DRAM power consumption by as much as 68.6% (19.9% on average). The peak and average temperature reductions achieved by TAP-low are 7.84°C and 2.55 °C, respectively.

References

[1] N. L. Binkert, R. G. Dreslinski, L. R. Hsu, K. T. Lim, A. G. Saidi, and S. K. Reinhardt, "The M5 simulator: Modeling networked systems," *IEEE Micro*, vol. 26, no. 4, pp. 52–60, Jul./Aug. 2006.

[2] V. Delaluz, M. Kandemir, N. Vijaykrishnan, A. Sivasubramaniam, and M. J. Irwin, "Hardware and software techniques for controlling DRAM power modes," *IEEE Trans. Comput.*, vol. 50, no. 11, pp. 1154–1173, Nov. 2001.

[3] X. Fan, C. S. Ellis, and A. R. Lebeck, "Memory control policies for DRAM power management," in *Proc. ISLPED*, 2001, pp. 129–134.

[4] J. Iyer, C. L. Hall, J. Shi, and Y. Huang, "System memory power and thermal management in platforms built on Intel Centrino duo mobile technology," *Intel Technol. J.*, vol. 10, no. 2, pp. 123–132, May 2006.

[5] S. S. Iyer and H. L. Kalter, "Embedded DRAM technology: Opportunities and challenges," *IEEE Spectr.*, vol. 36, no. 4, pp. 56–64, Apr. 1999.

[6] A. R. Lebeck, X. Fan, H. Zeng, and C. Ellis, "Power aware page allocation," in *Proc. ASPLOS-IX*, 2000, pp. 105–116.

[7] K.-B. Lee, T.-C. Lin, and C.-W. Jen, "An efficient quality-aware memory controller for multimedia platform SoC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 5, pp. 620–633, May 2005.

[8] J. Lin, H. Zheng, Z. Zhu, H. David, and Z. Zhang, "Thermal modeling and management of DRAM memory systems," in *Proc. ISCAS*, 2007, pp. 312–322.

[9] J. Lin, H. Zheng, Z. Zhu, Z. Zhang, and H. David, "DRAM-level prefetching for fully-buffered DIMM: Design, performance and power saving," in *Proc. ISPASS*, 2007, pp. 94–104.

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

[11] S. Liu, S. Ogrenci Memik, Y. Zhang, and G. Memik, "An approach for adaptive DRAM temperature and power management," in *Proc. ICS*, 2008, pp. 63–72.

[12] S. Liu, S. Ogrenci Memik, Y. Zhang, and G. Memik, "A power and temperature aware DRAM architecture," in *Proc. DAC*, 2008, pp. 878–883.

[13] *"Calculating Memory System Power for DDR2,"* Micron, Boise, ID, 2005.

[14] *"System Power Calculator,"* Micron, Boise, ID, 2006 [Online]. Available: http://www.micron.com/support/designsupport/tools/powercalc/powercalc.aspx

[15] S. Rixner, "Memory controller optimizations for web servers," in *Proc. MICRO-37*, 2004, pp. 355–366.

[16] H. Shim, Y. Joo, Y. Choi, H. G. Lee, and N. Chang, "Low-energy off-chip SDRAM memory systems for embedded applications," *ACM Trans. Embed. Comput. Syst.*, vol. 2, no. 1, pp. 98–130, Feb. 2003.

[17] P. Shivakumar and N. P. Jouppi, "CACTI 3.0: An integrated cache timing, power, and area model," Western Res. Lab., Palo Alto, CA, WRL Research Report 2001/2, 2001.

[18] *"SPEC CPU2000,"* Standard Perform. Eval. Corp., Warrenton, VA, 2000. [Online]. Available: www.spec.org