Review of Energy Saving Strategies for DRAM

Sai Kiran Talamudupula
BIOS Engineer, Intel Corporation, Chandler, Arizona

ABSTRACT

The ever increasing power consumption of the components within a computing system have resulted in tremendous costs and substantial failure rates which are a roadblock in achieving optimal performance at reasonable costs. To mitigate these issues, strategies that reduce the dynamic power consumption of these components are needed. In this paper, we review a survey a subset of those strategies with their salient features and their efficacy in providing energy savings. The paper reviews energy saving strategies proposed and verified in both simulators and real-time systems.

Keywords : DRAM Power Management, Power States.

I. INTRODUCTION

The ever increasing costs and constraints on power consumption are limiting the leap to the next generation exascale systems [34] as the power limit determined for these systems as per DoE guidelines is 20 MW. This means that nearly 60x improvement is watts per operation efficiency is needed. Considering the general trend of change in IPC moving from one Intel processor generation to another, the average increase in instructions per cycle (IPC) has been modest so achieving such a high ratio of power to performance in a relatively short duration could be nearly impossible. Therefore, there is an urgent need to limit the power consumption of modern computing systems to mitigate these issues.

To minimize the impact of the increasing power consumption on the computing performance, novel schemes/strategies are the need of the hour which could achieve it without hurting the performance by much. Traditionally, the focus of the energy saving strategies has been on the processor since it has been found to have the largest ratio in the power consumption of a full compute node. But off-late, DRAM (dynamic random access memory) power consumption has been found increasing in server platforms and its power consumption can even go beyond the level of processor power, given the right amount of memory intensity and DRAM modules. Similarly, the data intensive applications in the HPC (high performance computing) domain also demand an increase in memory capacity and subsequently, an increase in a share of its power consumption. For example, in an IBM server, it was determined that the average memory power consumption was determined to be 1200 watts compared to average processor power consumption of 840 watts. This shows that there is an urgent need to limit the power consumption of the memory as well.

Many strategies have been proposed in the past to reduce processor power consumption, primarily through DVFS. Profiling based strategies making use of performance counters have been used in [1], [3], [8], [9],[10],[12],[14],[16],[18],[21],[25],[26],[27],[59], [60],[61],[62][63],[69],[70],[71],[73].[74],[75], [76],[77], [78]. Communication interval determination in parallel applications making use of MPI and other communication mechanisms and subsequent application of DVFS was done in [15], [19], [23], [32], [36], [58], [59], [60], [61][62],[63]. The Intel Running

Average Power Limit (RAPL) has also been used to limit power consumption in modern computing systems [6],[9],[18],[26],[78].

In this paper, we survey several DRAM power saving techniques. We only review a subset of those strategies since it is impossible to survey all the available strategies. Therefore, we pick only the most representative ones to give an overview regarding current state of the work in DRAM power saving. Additionally, we only focus on strategies which were proposed, explicitly to address power saving instead of achieving it as a by product of their research work. Moreover, we focus on strategies proposed and verified in both simulation and real systems.

The rest of the paper is organized as follows. Section 2 provides the background about basic terminology for DRAM internals. Section 3 discusses the previously proposed power saving strategies for DRAM. Section 4 provides the conclusions for the paper.

II. DRAM Basics

DRAM consists of multiple arrays of one bit storages which are arranged in a 2-D form. These 2-D matrices are formed by the intersection of rows (word lines) and columns (bit lines). These grid matrix like structure is called a bank and when these banks are associated in the group of two, four and eight, these banks form DDR, DDR2 and DDR3 and they form the next higher logical unit in DRAM hierarchy (Fig.1) known as rank.

A DIMM (Dual in-line memory module) is a circuit board which consists of one of more DRAM ranks on it. This provides provides an interface to the memory bus. A DRAM channel formed by a one or a group of DIMMs that process the requests from the DRAM controller which is usually built on-chip so as to decrease the latency of the memory access operations..
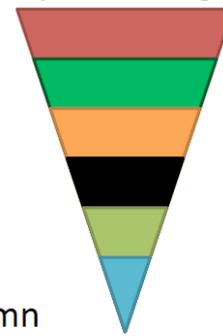

Figure 1 : Different levels of memory hierarchy.

DRAM is classified according to the maximum speed at which it, as well as the timings. DRAM Timings are numbers such as 3-6-7-10 which describe the performance of the DIMM modules using four timing parameters namely Column access strobe (CAS) latency, row to column address delay (tRCD), row precharge time (tRP) and row active time (tRAS), respectively. These timings are specified in BIOS and are managed through memory reference code (MRC).

III. Discussion of Power Saving Strategies

Authors in [30] use a mechanism which estimates the timing of memory access of to a bank bank and then switch it to low power mode appropriately. It also uses makes use of compiler analysis to modify the executable as per the memory accesses and based on that analysis different power saving modes are applied.

Lebeck et al. [47] describes a strategy for putting DRAM modules in low power modes by modifying virtual to physical address mapping so that physical pages are arranged into minimal number of chips and chips which have not been utilized are being put into low power mode.

An estimation of the idle time of memory chips was proposed in [38] which helped in deciding the time after which the DRAM can be put into low-power state. Asimilarstrategywas proposedin [49] for reducing memory energy consumption by adaptively transitioning memory module to low- power modes

with a added performance loss constraint which cannot be violated and provides an overall envelope within which, energy savings are to be maximized.

Authors in [31] propose an operating system scheduler based power mode strategy for DRAM energy efficiency which dynamically keeps a track of banks which are being utilized and accordingly switches on/off these banks at a context switch detected by the scheduler.

[47] proposes a technique for turning off DRAM chips in low-power mode. Their technique works by controlling virtual address to physical address mapping such that the physical pages of an application are clustered into a minimum number of DRAM chips and the unused chips are transitioned to low power modes. In addition, their technique also monitors the time period between accesses to a chip as a metric for measuring the frequency of reuse of a chip. When this time is greater than a certain threshold, the chip is transitioned to the low-power mode.

A strategy in Delaluz and Kandemir et al. [28] reduces DRAM power and energy consumption placing the arrays at run time with temporal affinity into the same set of banks, which helps in taking advantage of deeper sleep modes for longer durations of time. A virtual memory management based technique is proposed in [40] to reduce the DRAM power consumption which works by remapping to reduce the memory footprint of each application and turning off the under utilized DRAM modules. Zhou et al. [64] proposes a memory allocation scheme which depends on component utilization such that applications are allocated the memory depending on their utility. The unallocated memory is put down
into the low power mode to save energy and power consumption. [41] discusses method for saving memory energy by focusing the memory accesses to few memory ranks so that rest of the ranks can
be put into low- power modes. This strategy works

by dividing the pages as per their utilization into hot and cold ranks and increases the idle periods of cold ranks.

Memory voltage and frequency scaling was used in [33, 35] reduce DRAM power consumption which reduces the frequency of devices, channels etc. when the memory utilization is relatively lower while minimizing the performance degradation at the same time. The "rank-subsetting" strategy was proposed in Zheng et al. [65] to reduce memory power consumption by decreasing the amount of memory actively involved in providing memory accesses. This is done by putting a buffer between the memory modules and the memory bus which can provide variegated arrangement of DRAM ranks so that only one mini-rank is activated for a memory access with rest of them put into low power modes.

The idle time of memory banks was increased under a scheme [46] to reduce their power consumption such that if requested data can be recomputed using the active banks, the banks which are in low-power mode are not activated. A DMA (direct memory access) based scheme is discussed in [57] for saving energy since DMA transfers tend to be larger than the transfers done through the CPU, they get sliced into smaller transfer. As the gaps between these small transfers does not warrant turning off the chips to low power modes, the authors discuss aligning DMA requests coming from different I/O buses to the same memory device. Yoon et al. [66] discuss a strategy which exploits low- power mobile DRAM components to reduce memory power consumption by using a buffering mechanism which combines the data outputs from multiple ranks of low frequency mobile DRAM devices such that it expands and provide equivalent bandwidth similar to higher powered DRAM services. A study was done in [67] to explore the latency and reliability characteristics of modern DRAM when the supply voltage is modified. They determined that reducing the supply voltage

below a certain threshold introduced errors in the data and these errors can be avoided by increasing the latency of major memory operations. Finally, they proposed a DRAM energy reduction strategy termed "Voltron" which used a performance model to determine the extent to which the supply voltage of DRAM can be reduced without introducing errors and without exceeding a performance loss constraint.

A strategy was proposed in [17] to reduce the delay of synchronization of ranks to low power modes using the system I/O calls by using the fact that most I/O requests go through system calls and the OS can determine the completion time of these requests and this can be used to put the idle memory ranks into low-power modes. [68] proposed a main memory design which optimizes three different memory modules power consumption etc. at the expense of other parameters by profiling an application based on its last level cache misses and memory level parallelism.

Authors in [82] and Li et al. [83] discuss strategies for minimizing memory power consumption exploiting spatial and temporal image data correlations in video processing applications by mapping image data in DRAM such that the row activations are minimized, subsequently reducing memory power consumption. Authors in [84] proposes refresh method to reduce refresh power consumption of memory devices, such memory row refreshes are bypassed using counters for each row in the memory module. An Analytical model is used for saving memory power in Lyuh et al. [51] by choosing a low-power mode for memory banks using scheduling of memory access operations etc.

A strategy for tuning the garbage collector is used in [24] to reduce DRAM power by switching off the banks that are not storing live data. An application level strategy is proposed in S. Liu et al. [50] which reduces refresh level power in memory by identifying critical and non-critical data in the application and allocated them in different modules of the memory

and then use lower refresh rates for non-critical data and higher refresh rates for critical data.

A strategy is proposed in [85] to improve upon DRAM temperature by working on thread scheduling and page allocation such that threads are grouped together and mapped to certain DIMMs and at any time only single group and the DIMM associated with it remain active and the rest are deactivated to reduce their temperature. A buffering based technique is studied in Trajkovic et al. [86] to reduce DRAM power consumption by examining the activate-precharge operation and avoiding the cost of activation. Their strategy works by prefetching cache blocks and combining multiple blocks for write accesses to the same DRAM row.

[37] proposes a strategy to reduce the peak power consumption of DRAM to a certain power budget by using basic algorithms such as knapsack etc. to determine the schedule which decides when the memory should be put to low power modes. A granularity modification based strategy to reduce memory power is proposed in Yoon et al. [87] which manages the virtual memory to specify access granularity for each page based on the spatial locality of an application.

David et. al [6] propose a memory-frequency scaling based strategy which essentially utilizes memory bandwidth usage metrics in a manner such that frequency scaling is applied whenever memory-bandwidth usage goes below a certain threshold. In addition to providing a memory frequency scaling strategy, the work also provides detailed power and performance models which discusses effect of memory frequency scaling on performance and power consumption.

A coordinated mechanism for processor and memory frequency scaling is proposed in [88] with detailed performance model which relies multiple performance counters to gauge the effect of memory and processor frequency scaling in conjunction on the

overall application performance and they depict their results on a simulation framework by using a timeslice based approach. A relatively modest approach of jointly applying processor and memory frequency scaling is proposed in [27] where the authors use a basic timeslice based strategy coupled with power and performance models based on feedback to the performance misprediction. They show their results in real-time by emulating memory frequency modification from the BIOS.

IV. CONCLUSIONS

The desire to extract maximum performance from the modern computing systems has greatly increased their power and energy consumption. To mitigate this issue, several strategies have been proposed in the past to reduce power consumption of the memory by various ways. In this work, we have reviewed many power saving strategies for DRAM which make use of frequency scaling, memory mapping and many other ways to implement the strategy. We hope that this survey work will inform researchers, processor architects and software engineers to regarding the prominent strategies employed to reduce memory power consumption.

V. REFERENCES

[1]. A. Intel Power Governor. https://software.intel.com/en- us/articles/intel-power-governor.

[2]. Intel Software Developer's Manual. https://software.intel.com/en-us/articles/intel-sdm

[3]. N. Almoosa, W. Song, Y. Wardi and S. Yalamanchili. A Power Capping Controller for Multicore Processors. In 2012 American Control Conference (ACC), pages 4709-4714. June 2012.

[4]. D.H. Bailey et. Al. The NAS Parallel Benchmarks Summary and Preliminary Results. In Proceedings of the 1991 ACM/IEEE conference on Supercomputing, pages 158-165, 1991.

[5]. T. Vogelsang. Understanding the Energy Consumption of Dynamic Random Access Memories. In MICRO 2010.

[6]. H. David, E. Gorbatov, U.R. Hanebutte, R. Khannal, and C. Le RAPL: Memory Power Estimation and Capping. In Proceedings of the 16th ACM/IEEE international symposium on Low power electronics and design, ISLPED'10, pages 189-194, New York, NY, USA, 2010. ACM.

[7]. V. Anagnostopoulou et. al. Barely alive memory servers: Keeping data active in a low-power state. in ACM Journal on Emerging Technologies in Computing Systems, Special issue on Sustainable and Green Computing Systems, April 2012.

[8]. R. Ge, X. Feng, W. Feng, and K.W. Cameron. CPU MISER: A Performance Directed, Run-time System for Power-aware Clusters. In Parallel Processing, 2007. ICPP 2007. International Conference on, page 18, Sep. 2007.

[9]. R. Ge, X. Feng, Y. He, and P. Zou. The Case for Cross- component Power Coordination on Power Bounded Systems. In 2016 45th International Conference on Parallel Processing (ICPP), pages 516-525, Aug 2016.

[10]. R. Ge, X. Feng, S. Song, H.C. Chang, D. Li, and K.W. Cameron. PowerPack: Energy Profiling and Analysis of High- performance systems and Applications. Parallel and Distributed Systems, IEEE Transactions on, 21:658-671, 2010.

[11]. A. Amin et. al. Rank-aware cache replacement and write buffering to improve DRAM energy efficiency. In Proceedings of the 16th ACM/IEEE international symposium on Low power electronics and design, pp. 383-388, ACM, 2010.

[12]. S. Huang and W. Feng. Energy-efficient Cluster Computing via Accurate Workload Characterization. In Cluster Computing and the Grid, 2009. CCGRID'09. 9th IEEE/ACM International Symposium on, pages 68- 75, May 2009.

[13]. R. Ayoub et. al. Energy efficient proactive thermal management in memory subsystem. In International Symposium on Low-Power Electronics and Design (ISLPED), pp. 195-200, IEEE, 2010.

[14]. C. Isci and M. Martonosi. Runtime Power Monitoring in High- end Processors: Methodology and empirical data. In Proceedings of the 36th annual IEEE/ACM International Symposium on Microarchitecture, MICRO 36, pages 93-, Washington, DC, USA, 2003. IEEE Computer Society.

[15]. K. Kandalla, E.P. Mancini, S. Sur, and D.K. Panda. Designing Power-aware Collective Communication Algorithms for InfiniBand Clusters. In Parallel Processing (ICPP), 2010 39th International Conference on, pages 218-227, 2010.

[16]. Kristopher Keipert, Gaurav Mitra, Vaibhav Sundriyal, Sarom S. Leang, Masha Sosonkina, Alistair P. Rendell, and Mark S. Gordon. Energy Efficient Computational Chemistry: Comparison of x86 and ARM Systems. Journal of Chemical Theory and Computation, 11(11):5055-5061, 2015. PMID:26574303.

[17]. M. Bi et. al. Delay-Hiding Energy Management Schemes for DRAM. In IEEE HPCA 2010.

[18]. Gary Lawson, Vaibhav Sundriyal, Masha Sosonkina, and Yuzhong Shen. Runtime Power Limiting of Parallel applications on Intel Xeon Phi Processors. In Proceedings of the 4th International Workshop on Energy Efficient Supercomputing, E2SC '16, pages 39- 45, Piscataway, NJ, USA, 2016. IEEE Press.

[19]. M.Y. Lim, V.W. Freeh, and D.K. Lowenthal. Adaptive, Transparent Frequency and Voltage scaling of Communication Phases in MPI Programs. In Proceedings of the 2006 ACM/IEEE conference on Supercomputing, 2006.

[20]. K. Chandrasekhar, B. Akesson and K. Goossens. Run- time power-down strategies for real-time SDRAM memory controllers. In Proceedings of the 49th Annual DAC, pp. 988-993, ACM, 2012.

[21]. K. Ma, X. Li, W. Chen, C. Zhang, and X. Wang. GreenGPU: A Holistic Approach to Energy Efficiency in GPU-CPU Heterogeneous Architectures. In 2012 41st International Conference on Parallel Processing, pages 48-57, Sept 2012.

[22]. Jayanth Srinivasan. An Overview of Static Power Dissipation. Technical report.

[23]. V. Sundriyal and M. Sosonkina. Per-call Energy Saving Strategies in All-to-all Communications In Proceedings of the 18th European MPI Users'

Group conference on Recent advances in the message passing interface, EuroMPI'11, pages 188-197, Berlin, Heidelberg, 2011. SpringerVerlag.

[24]. G. Chen, R. Shetty, M. Kandemir, N. Vijaykrishnan, M. Irwin and M. Wolczko. Tuning garbage collection for reducing memory system energy in an embedded java environment.In ACM Transactions on Embedded Computing Systems. 2002.

[25]. V. Sundriyal and M. Sosonkina. Initial Investigation of a Scheme to use Instantaneous CPU Power Consumption for Energy Savings format. In Proceedings of the 1st International Workshop on Energy Efficient Supercomputing, E2SC '13, pages 1:1-1:6, New York, NY, USA, 2013. ACM.

[26]. M. Chen et. al. Coordinating Processor and Main Memory for Efficient Server Power Control. In ICS 2011.

[27]. V. Sundriyal and M. Sosonkina. Joint Frequency Scaling of Processor and DRAM. The Journal of Supercomputing, 72(4):1549-1569, 2016.

[28]. V DeLaLuz et. al. Automatic data migration for reducing energy consumption in multi-bank memory systems," in DAC, 2002, pp. 213- 218, IEEE, 2002

[29]. V DeLaLuz, M. Kandemir, N. Vijaykrishnan, M. Irwin, A. Sivasumbramaniam, and I. Kolcu. Compiler-directed array interleaving for reducing energy in multi-bank memories. In ASPDAC 2002.

[30]. V DeLaLuz, M. Kandemir, N. Vijaykrishnan, M. Irwin and A. Sivasumbramaniam. DRAM energy management using software and hardware directed power mode control. In HPCA 2001.

[31]. V DeLaLuz, M. Kandemir, N. Vijaykrishnan, M. Irwin and A. Sivasumbramaniam. Scheduler Based DRAM Energy Management. DAC 2002.

[32]. V. Sundriyal, M. Sosonkina, and A. Gaenko. Runtime Procedure for Energy Savings in Applications with Point-to- point Communications. In Computer Architecture and High- Performance Computing (SBAC-PAD), 2012 IEEE 24th International Symposium on, pages 155-162, 2012.

[33]. Q. Deng et. al. MultiScale: Memory System DVFS with Multiple Memory Controllers. ISPLED 2012.

[34]. DRAM Basics. Chapter 12. https://www.ece.cmu.edu/~ece740/f11/lib/exe/fetch.php?media =wiki:lectures:onur-740-fall11-lecture25-mainmemory.pdf

[35]. Q. Deng et. al. Memscale: active low-power modes for main memory. ACM SIGPLAN Notices, vol. 46, no. 3, pp. 225-238, 2011.

[36]. V. Sundriyal, M. Sosonkina, and Z. Zhang. Achieving Energy Efficiency During Collective Communications. Concurrency and Computation: Practice and Experience, 25(15):2140-2156, 2013.

[37]. B. Diniz, D. Guedes, W. Meira Jr, and R. Bianchini, "Limiting the power consumption of main memory," in ACM SIGARCH Computer Architecture News, vol. 35, pp. 290-301, ACM, 2007.

[38]. X. Fan, C. Ellis, and A. Lebeck, "Memory controller policies for DRAM power management," in Proceedings of the 2001 international symposium on Low power electronics and design, pp. 129-134, ACM, 2001.

[39]. M. Floyd, S. Ghiasi, T. Keller, K. Rajamani, F. Rawson, J. Rubio, and M. Ware, "System power management support in the IBM POWER6 microprocessor," IBM Journal of Research and Development, vol. 51, no. 6, pp. 733-746, 2007.

[40]. H. Huang, P. Pillai, and K. Shin, "Design and implementation of power-aware virtual memory," USENIX Annual Technical Conference, pp. 57-70, 2003.

[41]. H. Huang, K. Shin, C. Lefurgy, and T. Keller, "Improving energy efficiency by making DRAM less randomly accessed," in Proceedings of the 2005 international symposium on Low power electronics and design, pp. 393-398, ACM, 2005.

[42]. I. Hur and C. Lin, "A comprehensive approach to DRAM power management," in 14th International Symposium on High Performance Computer Architecture, 2008. HPCA., pp. 305-316, IEEE, 2008.

[43]. S. Irani, S. Shukla, and R. Gupta, "Online strategies for dynamic power management in systems with multiple power-saving states," ACM Transactions on Embedded Computing Systems (TECS), vol. 2, no. 3, pp. 325-346, 2003.

[44]. M. Kandemir, U. Sezer, and V. Delaluz, "Improving memory energy using access pattern classification," in International Conference on Computer-Aided Design, pp. 201-206, IEEE Press, 2001.

[45]. B. Khargharia, S. Hariri, and M. S. Yousif, "Self-optimization of performance-per-watt for interleaved memory systems," in 14th international conference on High performance computing, HiPC'07, (Berlin, Heidelberg), pp. 368-380, Springer-Verlag, 2007.

[46]. H. Koc, O. Ozturk, M. Kandemir, and E. Ercanli, "Minimizing energy consumption of banked memories using data recomputation," in International Symposium on Low Power Electronics and Design, 2006., pp. 358- 361, 2006.

[47]. A. Lebeck, X. Fan, H. Zeng, and C. Ellis, "Power aware page allocation," ACM SIGPLAN Notices, vol. 35, no. 11, pp. 105-116, 2000.

[48]. X. Li, R. Gupta, S. Adve, and Y. Zhou, "Cross-component energy management: Joint adaptation of processor and memory," ACM Transactions on Architecture and Code Optimization (TACO), vol. 4, no. 3, p. 14, 2007.

[49]. X. Li, Z. Li, F. David, P. Zhou, Y. Zhou, S. Adve, and S. Kumar, "Performance directed energy management for main memory and disks," in ACM SIGARCH Computer Architecture News, vol. 32, pp. 271-283, ACM, 2004.

[50]. S. Liu, K. Pattabiraman, T. Moscibroda, and B.Zorn, "Flikker: Saving DRAM refresh-power through critical data partitioning," ACM SIGPLAN Notices, vol. 46, no. 3, pp. 213-224, 2011.

[51]. C. Lyuh and T. Kim, "Memory access scheduling and binding considering energy minimization in multi-bank memory systems," in Proceedings of the 41st annual DAC, pp. 81-86, ACM, 2004.

[52]. K. T. Malladi, F. A. Nothaft, K. Periyathambi, B. C. Lee, C. Kozyrakis, and M. Horowitz, "Towards energy- proportional datacenter memory with mobile DRAM," in ISCA, pp. 37 - 48, june 2012.

[53]. D. Meisner, B. Gold, and T. Wenisch, "PowerNap: eliminating server idle power," ACM Sigplan Notices, vol. 44, no. 3, pp. 205-216, 2009.

[54]. J. Mukundan and J. F. Martinez, "MORSE: Multi- objective reconfigurable self-optimizing memory scheduler," HPCA, vol. 0, pp. 1-12, 2012.

[55]. O. Ozturk, G. Chen, M. Kandemir, and M. Karakoy, "Cache miss clustering for banked memory systems," in IEEE/ACM international conference on Computer- aided design, ICCAD '06, pp. 244-250, ACM, 2006.

[56]. O. Ozturk and M. Kandemir, "ILP-Based energy minimization techniques for banked memories," ACM Trans. Des. Autom. Electron. Syst., vol. 13, pp. 50:1-50:40, July 2008.

[57]. V. Pandey, W. Jiang, Y. Zhou, and R. Bianchini, "DMA- aware memory energy management," in HPCA, pp. 133- 144, feb. 2006.

[58]. V. Sundriyal, M. Sosonkina, and Z. Zhang. Automatic Runtime Frequency Scaling System for Energy Savings in Parallel Applications. The Journal of Supercomputing, 68(2):777-797,2014.

[59]. Vaibhav Sundriyal, Ellie Fought, Masha Sosonkina, and Theresa L. Windus. Power Profiling and Evaluating the Effect of Frequency Scaling on NWChem. In Proceedings of the 24th High Performance Computing Symposium, HPC '16, pages 19:1-19:8, San Diego, CA, USA, 2016. Society for Computer Simulation International.

[60]. Vaibhav Sundriyal, Ellie Fought, Masha Sosonkina, and Theresa L. Windus. Evaluating effects of application based and automatic energy saving strategies on NWChem. In Proceedings of the 25th High Performance Computing Symposium, HPC '17, pages 16:1-16:12, San Diego, CA, USA, 2017. Society for Computer Simulation International.

[61]. Vaibhav Sundriyal and Masha Sosonkina. Runtime power- aware energy-saving scheme for parallel applications. 2015.

[62]. Vaibhav Sundriyal, Masha Sosonkina, Alexander Gaenko, and Zhao Zhang. Energy saving strategies for parallel applications with point-to-point communication phases. Journal of Parallel and Distributed Computing, 73(8):1157 - 1169, 2013.

[63]. A. Vishnu, S. Song, A. Marquez, K. Barker, D. Kerbyson, K. Cameron, and P. Balaji. Designing Energy Efficient Communication Runtime Systems for Data Centric Programming Models. In Proceedings of the 2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing, GREENCOM-

CPSCOM '10, pages 229-236, Washington, DC, USA, 2010. IEEE Computer Society.

[64]. P. Zhou et. al. Dynamic Tracking of Page Miss Ratio Curve for Memory Management. ACM SIGOPS 2004.

[65]. Honzong Zheng et. al. Mini-Rank: Adapting DRAM Architecture for Improving Memory Power Efficiency. MICRO 2008.

[66]. D.H Yoon, J. Chang, N. Muralimanohar and P. Ranganathan. BOOM: Enabling Mobile Memory Based Low-Power Server DIMMs. ISCA 2012.

[67]. Kevin K. Chang et. al. Understanding Reduced Voltage Operations in Modern DRAM Devices: Experimental Characterization, Analysis, and Mechanisms. Proceedings of the ACM on Measurement and Analysis of Computing Systems. Volume 1, Issue 1, June 2017.

[68]. S. Phadke and S. Narayanasamy. MLP Aware Heterogeneous Memory Systems. IEEE DATE 2011.

[69]. C. Lefurgy, K. Rajamani, F. Rawson, W. Felter, M. Kistler, and T.W. Keller. Energy management for commercial servers. Computer, 36(12):39- 48, December 2003.

[70]. S. Borkar. The exascale challenge, 2011. Keynote speech, the 12th International Conference on Parallel Architectures and Compilation Techniques.

[71]. R. Gonzalez, B. M. Gordon, and M. A. Horowitz. Supply and Threshold Voltage Scaling for Low Power CMOS. IEEE Journal of Solid-State Circuits, 32(8):1210-1216, Aug 1997.

[72]. G. Semeraro, G. Magklis, R. Balasubramonian, D. H. Albonesi, S. Dwarkadas, and M. L. Scott. Energy-Efficient Processor Design Using Multiple Clock Domains with Dynamic Voltage and Frequency Scaling. In Proceedings Eighth International Symposium on High Performance Computer Architecture, pages 29-40, Feb 2002.

[73]. Linux Kernel Governors. https://www.kernel.org/doc/Documentation/cpufreq/governors.t xt. Online].

[74]. V Pallipadi and A Starikovskiy. The Ondemand Governor: Past, Present and Future. 2:223-238, 01 2006

[75]. R. Efraim, R. Ginosar, C. Weiser, and A. Mendelson. Energy Aware Race to Halt: A Down to Earth Approach for Platform Energy

Management. IEEE Computer Architecture Letters, 13(1):25-28, Jan 2014.

[76]. J. P. Halimi, B. Pradelle, A. Guermouche, N. Triquenaux, A. Laurent, J. C. Beyler, and W. Jalby. Reactive DVFS Control for Multicore Processors. In 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, pages 102-109, Aug 2013.

[77]. Vaibhav Sundriyal, Masha Sosonkina, Fang Liu, and Michael W. Schmidt. Dynamic Frequency Scaling and Energy Saving in Quantum Chemistry Applications. In IEEE International Symposium on Parallel and Distributed Processing Workshops and PhD Forum, pages 837-845, 2011.

[78]. A. Marathe et. al. A Run-time System for Power Constrained HPC Applications. International Conference on High Performance Computing. ISC 2015.

[79]. Z. Wang and X. Hu, "Energy-aware variable partitioning and instruction scheduling for multibank memory architectures," ACM Transactions on Design Automation of Electronic Systems (TODAES), vol. 10, no. 2, pp. 369-388, 2005.

[80]. H. Zheng and Z. Zhu, "Power and Performance Trade- Offs in Contemporary DRAM System Designs for Multicore Processors," IEEE Transactions on Computers, vol. 59, no. 8, pp. 1033-1046, 2010.

[81]. D. Kaseridis, J. Stuecheli, and L. K. John, "Minimalist open-page: a DRAM page-mode scheduling policy for the many-core era," in MICRO, MICRO-44 '11, (New York, NY, USA), pp. 24-35, ACM, 2011.

[82]. H. Kim and I.-C. Park, "High-performance and low- power memory-interface architecture for video processing applications," IEEE Transactions on Circuits and Systems for Video Technology,, vol. 11, pp. 1160 - 1170, nov 2001.

[83]. Y. Li and T. Zhang, "Reducing DRAM Image Data Access Energy Consumption in Video Processing," IEEE Transactions on Multimedia, vol. 14, pp. 303 -313, april