RETROSPECTIVE: DRPM: Dynamic Speed Control for Power Management in Server Class Disks

Sudhanva Gurumurthi[1] Anand Sivasubramaniam[2] Mahmut T. Kandemir[2] Hubertus Franke[3]

[1]Advanced Micro Devices, Inc.
Austin, TX, USA
sudhanva.gurumurthi@amd.com

[2]The Pennsylvania State University
University Park, PA, USA
{axs53,mtk2}@psu.edu

[3]IBM Research
Yorktown Heights, NY, USA
frankeh@us.ibm.com

Our ISCA 2003 DRPM paper was unique compared to computer architecture papers of that time in two respects. First, this paper focused on a layer of the memory hierarchy that was not a common topic in most computer architecture conferences then: storage. Second, the paper delved into the *microarchitecture* of the then predominant storage device in servers: the hard disk drive (HDD), whose internal composition is primarily electro-mechanical and details of which were rarely the focus of study in computer architecture. As maverick as this paper may have been, it was grounded in certain key trends and shifts in computing at the turn of the 21st century.

This research was carried out in an era when there was a major expansion in data centers, which would eventually lead to paradigms such as cloud computing. Mobile computing devices, such as smart phones, were still at their infancy and research was ongoing on how the future may hold prospects for such devices running applications that leverage remote computing resources. As an example, our own research group at Penn State was exploring ideas along these lines in the context of mobile spatial databases, where query processing is partitioned between a resource-constrained mobile device communicating with servers in data centers [1]. These emerging applications and use cases would require data center infrastructures composed of compute, memory, and storage deployed at scale. HDDs were the bulwark of the storage systems in this era. It was common to organize HDDs in the form of RAID arrays [6] for performance and resilience reasons. This often required provisioning more HDDs than necessary solely to meet storage capacity goals. As the trend towards data centers and "warehouse-scale computing" [2] continued, there was a growing concern about the energy usage of these large infrastructures. The issue of energy efficiency and sustainability is as important at the time of writing this retrospective as it was when we carried out this research.

There was vigorous research activity across academia and industry on reducing the energy usage of CPUs and main memory. The general approach taken was to design hardware to have a set of operating states or modes, where each state consumes a certain amount of power, provides a certain level of performance, and then define a state-machine that governs the transition between these modes and the transition costs, typically the latency to exit one state to go to another. For example, in CPUs, these states would be defined in terms of points along the voltage-frequency curve and some additional states that may suspend clocking or gate power to specific parts of the processor. The key attribute that made this state-machine style power management attractive is that the processor or the memory system could continue to do useful work when residing in these power states and the transition latencies between most of these states are small enough for effective power management to be practical. Hard disk drives, on the other hand, presented a big challenge in the context of servers.

There were essentially three power states for a HDD: (1) An "active" state where the disk drive is fully powered up, the spindles are spinning at their rated speed (denoted in Rotations Per Minute - RPM), and the device is doing I/O; (2) An "idle" state where the device is not doing any work but the spindles keep rotating; (3) A "standby" state where the spindles are spun down. Since the spindles need to spin for an HDD to serve read and write requests, the device cannot serve any I/O requests when it is in the standby mode. Most of the energy used by an HDD was due to the spindle motor that kept the platter assembly spinning, which meant that any meaningful energy gains in an HDD required putting the device into the standby state. For example, our ISCA paper shows that server-grade HDDs at that time consumed approximately 40 Watts in the active state vs. 4 Watts in standby. However, it took many *seconds* to exit from standby and transition to one of the other states. A second is several orders of magnitude longer than the exit latency of any of the power states in a CPU or memory, which meant that disks could never go into this state if the server(s) were being actively used.

To understand the scope of the problem and identify opportunities to manage storage energy usage, we carried out a trace-driven simulation study of two transaction processing workloads on various RAID array configurations [5]. There were two key insights from this study that drove the direction of our research. First, we found that most of the energy used by HDDs is when they are in the *idle* state, when suggested that there may be opportunities to reduce energy usage by

transitioning the HDDs to standby. However, we found that the duration of these idle periods was much shorter than the time it takes to transition to and from the standby state. The second key insight was that the predictability of these idle periods was low. This implied that trying to craft policies that would predict idleness and proactively transition the HDD to standby or back to active is challenging. We explored various ways to either increase the inter-arrival time between I/O requests at the HDD or by exploring various parameters of the RAID array itself. The benefits were found to be small. We needed to take a step back and think differently about this problem.

In our initial studies, we considered the internals of the HDD to be a black box and focused solely on workload or RAID array level tuning approaches. Given the results of our characterization study, we decided to examine and understand the internal mechanisms of the HDD to identify avenues for reducing energy usage. We learned that the time it takes for an already spinning spindle motor to go from one non-zero RPM to another is *linear* with the RPM, whereas the the power of the motor is a *quadratic* function of its angular momentum. In other words, a linear decrease in the RPM can yield a quadratic reduction in its energy usage and that this linear RPM change is possible at a relatively lower transition latency than it would be to transition between active/idle and standby. This was an important insight that drove our research.

Our inspiration on leveraging this property of the spindle motor arose from an unusual source: a ceiling fan. Ceiling fans often have a knob (a voltage regular) on the wall that allows one to control the speed of the fan, slowing it down or speeding it up based on how much air circulation was desired. The time (exit latency, if you will) of a ceiling fan to go from a slow to a fast state is far lower than from the off state to full speed due to the need to overcome inertia. So, we asked the following question: *Could an HDD support multiple operating modes between active and standby, where each intermediate point is at a different RPM and the device can serve I/O requests in each such mode?* This would make the mechanism to enable disk power management analogous to multiple voltage-frequency points in a CPU, where the RPM of the HDD is the power management knob. We called this concept "Dynamic RPM" or "DRPM".

We had to work through the details of these functions and their parameters for HDD designs, various practical electro-mechanical and electronic issues for multi-RPM operation, and how DRPM could be leveraged to manage storage energy usage while minimizing performance impact. This involved a combination of deep dives into HDD design and its mechanics and studying publicly available data about commercial HDDs to construct architectural simulation models to evaluate DRPM. The interested reader is referred to our ISCA paper for these details. Interestingly, a paper on storage energy management published around the same time as our ISCA paper also included the idea of a dual-RPM HDD to reduce energy usage, but only when in the idle state [3]. In addition to energy reduction, our subsequent research also showed that DRPM is an effective means to thermal management of a disk drive [4].

As we reflect upon this paper two decades later, a few highlights stand out. This was Sudhanva's first paper at ISCA and laid the foundation for his dissertation and many of his subsequent projects in academia and industry. The multi-RPM concept was commercialized in the Hitachi Deskstar™ 7K400 HDD. Looking at the citations to our paper, we believe the more lasting impact of our work is that it spurred a large body of research into storage energy efficiency from the computer architecture and systems communities. We believe this paper also served as an exemplar for future papers in these communities that explored innovations in the design of storage devices. With the confluence of memory and storage in recent years enabled by technologies such as non-volatile memory, novel data center fabrics to leverage them at scale, and the practical realization of in-memory and in-storage compute paradigms, the future appears bright for continued innovation in this space.

We thank our colleagues in the Computer Systems Laboratory (CSL) and the Microsystems Design Laboratory (MDL) at Penn State for their technical inputs and encouragement. We thank the National Science Foundation for funding this research.

REFERENCES

[1] N. An, A. Sivasubramaniam, N. Vijaykrishnan, M. Kandemir, M. Irwin, and S. Gurumurthi, "Analyzing energy behavior of spatial access methods for memory resident data," in *Proceedings of the International Conferences on Very Large Databases (VLDB)*, September 2001, pp. 411–420.

[2] L. Barroso, U. Hölzle, and P. Ranganathan, *The Datacenter as a Computer: Designing Warehouse-Scale Machines (Synthesis Lectures on Computer Architecture*, 3rd ed. Morgan and Claypool Publishers, 2019.

[3] E. Carrera, E. Pinheiro, and R. Bianchini, "Conserving disk energy in network servers," in *Proceedings of the International Conference on Supercomputing (ICS)*, June 2003, pp. 86–97.

[4] S. Gurumurthi, A. Sivasubramaniam, and V. Natarajan, "Disk drive roadmap from the thermal perspective: A case for dynamic thermal management," in *Proceedings of the International Symposium on Computer Architecture (ISCA)*, June 2005, pp. 38–49.

[5] S. Gurumurthi, J. Zhang, A. Sivasubramaniam, M. Kandemir, H. Franke, N. Vijaykrishnan, and M. Irwin, "Interplay of energy and performance for disk arrays running transaction processing workloads," in *Proceedings of the International Symposium on Performance Analysis of Systems and Sofware (ISPASS)*, March 2003, pp. 123–132.

[6] D. Patterson, G. Gibson, and R. Katz, "A case for redundant arrays of inexpensive disks (RAID)," in *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, June 1988, pp. 109–116.