GreenDM: A Versatile Hybrid Drive for Energy and Performance

Zhichao Li, Ming Chen, and Erez Zadok
{*zhicli, mchen, ezk*}*@cs.stonybrook.edu*
Department of Computer Science, Stony Brook University

Abstract

Studies show that power consumption in the IT infrastructure is critical [5, 13, 14] with up to 40% of that power consumption comes from storage [21]. Therefore, power consumption has become an important factor influencing storage systems design [1, 17, 22, 23, 25, 26, 28]. Modern computer components such as CPU, RAM, and disk drives tend to have multiple power states with different operational modes [6, 9, 15]. Among them, magnetic HDDs achieve the worst *power-proportionality* [2], which states that systems should consume power proportional to the amount of work performed.

With the advent of Flash-based Solid State Drives (SSDs) that are more power- and performance efficient than HDDs, many considered SSDs as the front tier storage for caching or data migration [3, 7, 10, 12, 16, 19, 20, 24]. However, those projects mainly focused on boosting performance. Only some studies [4, 8, 10, 11, 18, 26] considered both performance and energy consumption. Even for these studies, their designs are often based on fixed and inflexible policies that make it difficult for the system to adapt well to different workloads. Moreover, caching-based systems work well when workloads exhibit strong data locality but can perform poorly otherwise. Lastly, many projects usually rely on simulations and refer to manufacturer's energy and performance specifications for benchmarks, rather than empirical, real-world results.

We designed and implemented a Linux Device Mapper [27] (DM) target named *GreenDM*; it maps one virtual block device onto several devices (e.g., SSD and SATA). GreenDM receives data requests from the hybrid virtual device, and then transparently redirects the resultant requests to the underlying block devices. The DM framework offers additional benefits that can be used with any target device (e.g., replication, multi-path, encryption, redundancy, and snapshots). The framework is also highly scalable: one can easily configure the virtual device to use multiple physical devices transparently.

In our GreenDM, we separate hot data from cold data based on their access patterns (recency, frequency, etc.): hot data is stored directly on the SSD and colder data is stored on the HDD. When cold data becomes hot, we migrate it from the HDD to the SSD; conversely, when hot data is not accessed enough or we need to free up space for hotter data, we migrate some colder data from the SSD to the HDD. GreenDM includes several versatile configuration parameters to determine the threshold for those migrations from the SSD to the HDD and vise verse. By using the SSD for hot data before using the HDD, we improve performance and reduce energy use—because SSDs are typically faster and consume less energy than HDDs. To further optimize the system, we decouple the migrations between the SSD and the HDD, and serve data requests directly from RAM once it is buffered. By keeping mostly cold data on the HDD, we can spin it down at times and further reduce whole-system energy consumption. To better work with workloads exhibiting poor data locality, GreenDM maps accesses from a Virtual LBA (VLBA) space to a Physical LBA (PLBA) space—starting from the lowest numbered SSD PLBAs to create hot regions on the SSD and cold regions on the HDD. Lastly, SSDs have a limited number of erasures to each block; to increase the SSD's lifetime, we drop migration attempts when there are concurrent accesses on the same data block. We have evaluated our prototype extensively using a variety of micro-benchmarks and general-purpose benchmarks. We experimented with several configurable GreenDM parameters, analyzed the results, and demonstrated their impact on performance and energy. We showed performance improvements of up to 160% and 330%, and concurrent energy savings of up to 60% and 76%, for a Video-server workload and a Web-search workload, respectively.

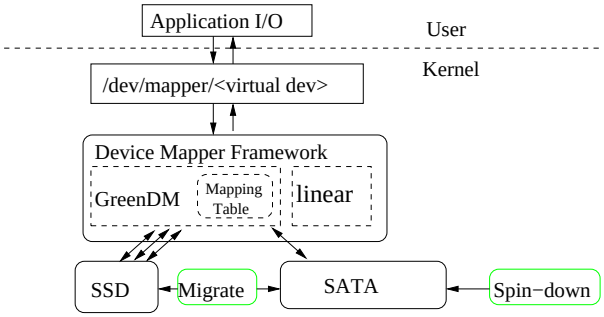

Figure 1: ***Architecture of the Green Virtual Device***

References

[1] D. G. Andersen, J. Franklin, M. Kaminsky, A. Phanishayee, L. Tan, and V. Vasudevan. FAWN: A Fast Array of Wimpy Nodes. In *Proceedings of the 22nd ACM Symposium on Operating Systems Principles (SOSP '2009)*. ACM SIGOPS, October 2009.

[2] L. A. Barroso and U. Hölzle. The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines. *Synthesis Lectures on Computer Architecture*, 4(1):1–108, 2009.

[3] Bcache. *http://bcache.evilpiepirate.org/*.

[4] T. Bisson, S. A. Brandt, and D. D.E. Long. A Hybrid Disk-Aware Spin-Down Algorithm with I/O Subsystem Support. In *In Proceedings of the 26th IEEE International Performance, Computing and Communications Conference*, 2007.

[5] J. Chang, J. Meza, P. Ranganathan, C. Bash, and A. Shah. Green Server Design: Beyond Operational Energy to Sustainability. In *Proceedings of the 2010 International Conference on Power Aware Computing and Systems*, HotPower'10, 2010.

[6] V. Delaluz, A. Sivasubramaniam, M. Kandemir, N. Vijaykrishnan, and M. J. Irwin. Scheduler-Based DRAM Energy Management. In *Proceedings of the 39th annual Design Automation Conference*, pages 697–702, New York, USA, 2002.

[7] Flashcache. *https://github.com/facebook/flashcache/*.

[8] J. Guerra, H. Pucha, J. Glider, W. Belluomini, and R. Rangaswami. Cost Effective Storage Using Extent Based Dynamic Tiering. In *USENIX FAST*, 2011.

[9] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke. DRPM: Dynamic Speed Control for Power Management in Server Class Disks. In *Proceedings of the 30th International Symposium on Computer Architecture*, pages 169–179, San Diego, Californai, USA, 2003.

[10] N. Joukov and J. Sipek. GreenFS: Making Enterprise Computers Greener by Protecting Them Better. In *Proceedings of the 3rd ACM SIGOPS/EuroSys European Conference on Computer Systems 2008 (EuroSys 2008)*, Glasgow, Scotland, April 2008. ACM.

[11] R. T. Kaushik and M. Bhandarkar. GreenHDFS: Towards An Energy-Conserving, Storage-Efficient, Hybrid Hadoop Compute Cluster. In *Proceedings of the 2010 International Conference on Power Aware Computing and Systems*, HotPower'10, 2010.

[12] I. Koltsidas and S. D. Viglas. Flashing up the Storage Layer. *Proceedings of the VLDB Endowment*, 1:514–525, 2008.

[13] J. G. Koomey. Growth in Data Center Electricity Use 2005 to 2010. Technical report, Standord University, 2011. *www.koomey.com*.

[14] Z. Li, K. M. Greenan, A. W. Leung, and E. Zadok. Power Consumption in Enterprise-Scale Backup Storage Systems. In *Proceedings of the Tenth USENIX Conference on File and Storage Technologies (FAST '12)*, San Jose, CA, February 2012. USENIX Association.

[15] Z. Li, R. Grosu, P. Sehgal, S. A. Smolka, S. D. Stoller, and E. Zadok. On the Energy Consumption and Performance of Systems Software. In *Proceedings of the 4th Israeli Experimental Systems Conference (ACM SYSTOR '11)*, Haifa, Israel, May/June 2011. ACM.

[16] T. Luo, R. Lee, M. Mesnier, F. Chen, and X. Zhang. hStorage-DB: Heterogeneity-Aware Data Management to Exploit the Full Capability of Hybrid Storage Systems. *Proceedings of the VLDB Endowment*, pages 1076–1087, 2012.

[17] D. Narayanan, A. Donnelly, and A. Rowstron. Write off-loading: Practical Power Management for Enterprise Storage. In *Proceedings of the 6th USENIX Conference on File and Storage Technologies (FAST 2008)*, 2008.

[18] E. Pinheiro and R. Bianchini. Energy Conservation Techniques for Disk Array-Based Servers. In *Proceedings of the 18th International Conference on Supercomputing (ICS 2004)*, pages 68–78, 2004.

[19] T. Pritchett and M. Thottethodi. SieveStore: A Highly-Selective, Ensemble-level Disk Cache for Cost-Performance. In *Proceedings of the 37th Annual International Symposium on Computer Architecture*, ISCA '10, 2010.

[20] M. Saxena, M. M. Swift, and Y. Zhang. FlashTier: a Lightweight, Consistent and Durable Storage Cache. In *Proceedings of the 7th ACM European Conference on Computer Systems*, EuroSys '12, pages 267–280, 2012.

[21] G. Schulz. Storage Industry Trends and IT Infrastructure Resource Management (IRM), 2007. *www.storageio.com/DownloadItems/CMG/MSP_CMG_May03_2007.pdf*.

[22] P. Sehgal, V. Tarasov, and E. Zadok. Evaluating Performance and Energy in File System Server Workloads Extensions. In *Proceedings of the Eighth USENIX Conference on File and Storage Technologies (FAST '10)*, pages 253–266, San Jose, CA, February 2010. USENIX Association.

[23] M. W. Storer, K. M. Greenan, E. L. Miller, and K. Voruganti. Pergamum: Replacing Tape with Energy Efficient, Reliable, Disk-based Archival Storage. In *Proceedings of the Sixth USENIX Conference on File and Storage Technologies (FAST '08)*, San Jose, CA, February 2008. USENIX Association.

[24] A Tiered Block Device. *http://sourceforge.net/projects/tier/*.

[25] A. Verma, R. Koller, L. Useche, and R. Rangaswami. SRCMap: Energy Proportional Storage Using Dynamic Consolidation. In *Proceedings of the 8th USENIX Conference on File and Storage Technologies*, FAST'10, 2010.

[26] C. Weddle, M. Oldham, J. Qian, A. A. Wang, P. Reiher, and G. Kuenning. PARAID: a Gear-Shifting Power-Aware RAID. In *Proceedings of the Fifth USENIX Conference on File and Storage Technologies (FAST '07)*, pages 245–260, San Jose, CA, February 2007. USENIX Association.

[27] Device Mapper. *http://en.wikipedia.org/wiki/Device_mapper*.

[28] Q. Zhu, Z. Chen, L. Tan, Y. Zhou, K. Keeton, and J. Wilkes. Hibernator: Helping Disk Arrays Sleep Through the Winter. In *Proceedings of the 20th ACM Symposium on Operating Systems Principles (SOSP '05)*, pages 177–190, Brighton, UK, October 2005. ACM Press.