RETROSPECTIVE: Power Management for Online Data Intensive Services

David Meisner*, Christopher M. Sadler†, Luiz André Barroso‡,
Wolf-Dietrich Weber‡ and Thomas F. Wenisch‡

*Meta †Snowflake ‡Google

I. Summary

In ISCA 2011, the authors undertook the first comprehensive study of server power management techniques for a class of workload the authors coined "Online Data Intensive (OLDI) Services"—an application class that is central to the success of online Internet services, and yet presents formidable power management challenges. These workloads perform significant computing over massive data sets per user request but, unlike their offline counterparts (such as MapReduce computations), they require responsiveness in the sub-second time scale at high request rates. Large search products, online advertising, and machine translation are examples of workloads in this class. These workloads are highly latency sensitive, and their data set usage does not scale down with traffic, making it infeasible to simply turn off machines during off-peak periods.

Our study examined what, if anything, can be done to make OLDI systems more energy proportional. Specifically, we evaluated the applicability of active and idle low-power modes to reduce the power consumed by the primary server components (processor, memory, and disk), while maintaining tight response time constraints, particularly on 95th-percentile latency. At a high level we were asking if existing energy management knobs could allow us to save power by running a cluster a little slower during periods of lower traffic, while still meeting service responsiveness targets. Using Web search as a representative example of this workload class, we first characterized a production Web search workload at cluster-wide scale. We provided a fine-grain characterization and exposed the opportunity for power savings using low-power modes of each primary server component. Second, we developed and validated a performance model to evaluate the impact of processor- and memory-based low-power modes on the search latency distribution.

We were delighted to observe that some of the existing power savings modes in CPUs at the time could be used to obtain modest energy savings in OLDI workloads, even if the results did not approach ideal energy-proportionality. Our evaluations also identified new opportunities for useful power management features in both CPU and memory system designs.

II. 12 years later

Power management remains critical. Effective power management remains as critical today for OLDI applications, and more generally for cloud computing, as it was in 2011. Indeed, with the slowing of Moore's Law and the end of Dennard Scaling, the installed lifetime of server infrastructure has grown relative to server lifetime in 2011, which has the effect of shifting a larger fraction of total cost of ownership to operating (energy) expense—a greater fraction of the cost of OLDI services is driven by power demand than 12 years ago. Moreover, the total installed base of server infrastructure has grown drastically over the past decade.

The industry has embraced the need for energy-proportionality. Component vendors heeded the call to arms from our work to build more energy-proportional hardware. Indeed, power and energy management is a first-class aspect of CPU design, and modern server processors include dozens of power and voltage domains and employ extensive use of clock and power gating to reduce energy consumption of idle subsystems and cores. Power management has grown even more critical with the widespread use of turbo/over-clocking modes, where frequency and single-thread performance can improve substantially when only a subset of a server's cores are in use; CPUs automatically adapt frequency and voltage in response to time-varying workloads. Power management in the memory subsystem has grown much more sophisticated in recent generations of the DDR and LPDDR specifications. Memory modules routinely include low-power modes where DIMMs enter self-refresh, reducing energy requirements to maintain state. Even enterprise rotating disks now routinely offer low-power modes that enable better energy proportionality under time-varying load than was generally available in 2011.

Power state management has become a first-class aspect of Linux scheduling; the Linux scheduler is aware of the power state of all cores and includes optimizations to avoid the cost and delay of waking a core from a sleep state. Power management is also more tightly integrated with I/O, with some subsystems able to steer I/O notifications based on power state or trigger power state transitions in response to I/O.

Progress has also been made in managing and scheduling for large-scale clusters over the past 12 years. At the time, overall average utilization of OLDI infrastructure was often 40% at best [1]. Through improvement in our ability to co-schedule latency-sensitive OLDI services with other kinds of workloads (e.g., latency-tolerant batch processing) [3], and better mechanisms to automatically size CPU and memory resources for OLDI applications [2], typical infrastructure

utilization has improved to 50% or even better in large-scale clusters running OLDI workloads.

Hardware infrastructure has evolved. Despite the progress that has been made, new challenges have arisen as the design of OLDI infrastructure has changed drastically in the 12 years since our work has been published. The most notable change is that server CPUs have drastically more cores, and hence, drastically more granular power control, as each core typically has independent idle and active low power modes. In our 2011 study, we examined a state-of-the-art dual-socket server system with 8 cores per socket. Today, with chiplet-based integration of multiple discrete dies in a single CPU socket, server systems can easily have more than 100 cores per socket, and even larger systems are on the horizon.

Since our work was published, infrastructure processing units (also called Smart network interfaces) have become a critical part of OLDI system architecture. These components add a substantial additional compute sub-system within the I/O dataplane, which itself requires sophisticated power management to match those of the primary compute infrastructure.

In 2011, we reported that memory bandwidth tended to be underutilized for our representative web search application. This observation suggested that active low power modes for memory, that sacrifice memory bandwidth in exchange for energy savings, were a fruitful direction for server power management. However, with the drastic growth in the number of cores per socket, but constraints on the number of pins per socket limiting the total number of memory channels each socket can support, memory bandwidth now tends to be a much more scarce resource in OLDI systems. As such, it is unlikely that active low power modes for memory would be as fruitful today as they appeared to be in 2011.

New workloads create new challenges. Our OLDI power management study was completed before the present boom in machine learning and artificial intelligence. In particular, machine learning inference is now routinely on the critical serving path of OLDI applications, and TPU and GPU hardware accelerators can account for a significant fraction of the computation and energy consumption of OLDI applications. For ML serving infrastructure with unpredictable and time-varying load, energy-proportional design and effective power management therefore must also be applied to accelerators. Whereas GPUs and TPUs do provide some power management, such as clock and power gating of idle units, they do not provide the granularity of active low power modes (e.g., fine-grain voltage scaling) available in CPUs—an area ripe for further research and development.

REFERENCES

[1] L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," *Computer*, vol. 40, no. 12, pp. 33–37, 2007.

[2] K. Rzadca, P. Findeisen, J. Swiderski, P. Zych, P. Broniek, J. Kusmierek, P. Nowak, B. Strack, P. Witusowski, S. Hand *et al.*, "Autopilot: workload autoscaling at google," in *Proceedings of the Fifteenth European Conference on Computer Systems*, 2020, pp. 1–16.

[3] A. Verma, L. Pedrosa, M. Korupolu, D. Oppenheimer, E. Tune, and J. Wilkes, "Large-scale cluster management at google with borg," in *Proceedings of the Tenth European Conference on Computer Systems*, 2015, pp. 1–17.