RETROSPECTIVE: Power Provisioning for a Warehouse-Sized Computer

Xiaobo Fan, Wolf-Dietrich Weber and Luiz André Barroso
{xiaobo, wolf, luiz}@google.com

Abstract - **This paper appeared at ISCA 2007, just a few years after the appearance of the first generation of datacenters that were custom-built for massive internet services. Our study was concerned with understanding the dynamic energy usage behavior of thousands of servers provisioned in such facilities to run popular internet-based services, and with identifying opportunities for optimizing the way power was provisioned in the datacenter. Our measurements also accidentally unveiled new opportunities for reducing the energy consumption of server-class hardware. As the first ISCA paper concerned primarily with the subject of datacenter design, this paper invited the architecture community to consider datacenters a worthy research topic.**

I. HISTORICAL CONTEXT

In the first years of this millennium, companies like Google used to deploy their computing equipment in 3rd party co-location facilities. As cloud-based services began to rapidly grow in popularity, it became clear that Google would need to build our own facilities, in addition to our own computing equipment (servers, networking switches and storage systems). That gave us the opportunity to design the building with some knowledge of the kind of computing equipment and workloads that would occupy it, as well as to optimize the computing equipment in order to make the facility itself more efficient. Such an all inclusive and integrated design problem is rarely available in an engineer's career, so our team jumped at the opportunity. Our guiding insight was that we were not designing a building, power delivery systems, cooling systems, computing systems, and software systems, but we were designing one massive building-sized machine. This zoomed out view of the design problem opened new areas of innovation in several engineering disciplines, including power provisioning and energy efficiency, which were the areas we explored in this paper.

Undertaking such a new design challenge gave us deep appreciation for the capital expenses (bricks, transformers, bus ducts, cooling towers, etc.) and operational expenses (electricity usage, facility operations) involved in providing access to massive internet services. One opportunity for cost optimization was to determine how much computing equipment could be safely deployed in a datacenter building that had a given maximum power delivery capacity by design. For example, if we could add 20% more servers to a building, the building construction costs per server deployed would be reduced by about 17%. These kinds of opportunities incentivized us to scrutinize the dynamic behavior of servers in further detail, and gave rise to the measurements that resulted in the findings of this paper.

II. KEY IDEAS

Whenever one is out of ideas it is useful to begin measuring some aspect of a system that hadn't been quantified before. In this case this started with an observation. We had built a datacenter with the capacity to deliver X Megawatts of power to computing equipment, filled that datacenter with such equipment, and observed that at the facility level we would never get anywhere near the full utilization of that power delivery capacity. The reason was that provisioning decisions were made assuming the peak power usage of servers, but servers rarely were exercised to peak compute (and therefore power) capacity and never at the same time, so power went unused. At the compute cluster level, such underutilization could cause some clusters to never reach beyond 60% of its provisioned capacity. The work that resulted in this paper sought to understand the joint probability of groups of servers being above a given fraction of their peak power demand. That understanding might allow us to establish a statistical guarantee that it would be safe to oversubscribe the provisioned facility power (adding more servers than the peak provisioned math would allow). To that end we had to develop an accurate

model of machine power usage based on instantaneous activity levels of components in a server, validate those models, and gain permission from our production teams to run those measurements in live datacenters.

The results were indeed promising, identifying sizable opportunities for safe power oversubscription, which in turn could translate into lowering datacenter costs substantially. However, our paper recognized that even with strong statistical guarantees of safe oversubscription, we needed mechanisms to react to possible overload situations in a graceful manner, as going above the provisioned capacity could result in unacceptable large scale service outages. If a statistically abnormal power peak were to emerge, the facility management software could prevent an overload by rapidly throttling down computing activity in subsets of the cluster. Google engineers subsequently implemented and launched such a power throttling system [3], first deployed over a decade ago and still an important part of our infrastructure efficiency solutions.

III. IMPACT

We identify three areas in which this paper has left a legacy in our community. The first area has less to do with the results of our paper, but more with the fact that the paper was published at all. The field of datacenter design was rather young then, and very few members of our community were aware of this rich new area for research and experimentation. A reasonable objection to accepting this paper back then would have been to declare it outside of the bounds of computer architecture. We are grateful to the ISCA 2007 PC for taking a chance on this paper. The vibrant area of Warehouse-Scale Computing exists today because of their willingness to expand our field to new topics. Within the sub-area of power management alone, interesting new research from companies like Facebook [4], Microsoft [5] and Google [3] is frequently published in architecture conferences

The second area our paper contributed to our community was in the description of the design of datacenters and the dynamic behavior of massive internet service workloads. We believe it made architects aware of the many opportunities to improve the design of large datacenters. In some ways, our work builds on Ranganathan at al's [2] ISCA-2006 paper that studies power provisioning at the blade server (or ensemble) level.

Looking back, however, the biggest impact derived from our work in this paper may have been an observation somewhat buried in subsection 5.2: *Improving Non-Peak Power Efficiency*. Our detailed measurements of machine activity against their corresponding power draw showed that underutilized servers were still consuming a lot of power. In fact, much of the inefficiency of server equipment back then could be addressed by challenging the hardware community to design systems that consumed power in a way that was proportional to their level of activity, and a subsequent paper from our group tried to make that call-to-action clear [1]. Energy proportionality was embraced as an industry goal and our servers and datacenters today are much more efficient because of the industry-wide effort to pursue it.

REFERENCES

[1] L.A. Barroso and U. Hölzle, The Case for Energy-Proportional Computing, IEEE Computer, Vol. 40, No. 12, December 2007.

[2] P. Ranganathan, P. Leech, D. Irwin, and J. Chase. Ensemble-level power management for dense blade servers. In ISCA ’06: Proceedings of the 33rd annual international symposium on Computer Architecture, pages 66–77, 2006.

[3] V. Sakalkar, V. Kontorinis, D. Landhuis, S. Li, D. D. Ronde, T. Bloom- ing, A. Ramesh, J. Kennedy, C. Malone, J. Clidaras, and P. Ranganathan, “Data Center Power Oversubscription with a Medium Voltage Power Plane and Priority-Aware Capping,” in ASPLOS, March 2020.

[4] Q. Wu, Q. Deng, L. Ganesh, C.-H. Hsu, Y. Jin, S. Kumar, B. Li, J. Meza, and Y. Song, “Dynamo: Facebook’s Data Center-Wide Power Management System,” in ISCA, June 2016.

[5] C. Zhang, A. Kumbhare, I. Manousakis, D. Zhang, P. Misra, R. Assis, K. Woolcock, N. Mahalingam, B. Warrier, D. Gauthier, L. Kunnath, S. Solomon, O. Morales, M. Fontoura, and R. Bianchini, “Flex: High- Availability Datacenters with Zero Reserved Power,” in ISCA, June 2021.