RETROSPECTIVE: Ensemble-level Power Management for Dense Blade Servers

Parthasarathy Ranganathan∗, Phil Leech†, David Irwin‡, Jeff Chase§
∗Google, †Hewlett Packard Enterprise, ‡University of Massachussets, Amherst, §Duke University

I. Motivation and Context

Seventeen years ago, in ISCA2006, we published a paper titled "Ensemble-level Power Management for Dense Blade Servers." At that time, there was broad research on mobile power management, but much less emphasis on server power management. We argued that this area warranted equal attention due to multiple trends: the cost/environmental implications of electricity, but more so, potential limits to power delivery (amperage per rack of servers) and thermal density (cooling capacity per rack). In even more of a departure from contemporary research, we focused on *peak* power provisioning. In contrast, most prior studies focused on *average* power.

Our key insight was the following: Trends towards blade servers and (then) emerging deployments at scale[1] meant that we could go beyond server power at the single-system level, and instead consider groups of servers (or *ensembles*).[2] Doing so enabled us to leverage higher-level statistical multiplexing trends in system-wide utilization and power consumption, and provision power for the *peak of the sum* of power usage across many servers rather than the *sum of the peak* power usage for each server. We likened our approach to the airline industry selling more tickets than seats, and then offering coupons in the uncommon case when everyone shows up.

Our proposal was, to put it mildly, very unconventional and challenged conventional wisdom (one internal reviewer called it "heresy"!). There were many concerns: Could we measure power at the right fidelity and granularity? What would we do if we ran out of power? How often would that happen? Would customers notice? How would we even describe this to them? This paper answered all of these questions. We presented a detailed analysis of power usage from real-world production deployments to back up our idea. We also proposed an architecture; designed an implementation that included hardware support for power monitoring and policies for power management; and demonstrated both prototype and simulation results that showed significant power reductions (up to 20% in overall power) with negligible workload slowdown.

II. Retrospective: Impact and Contributions

Widespread Industry Adoption. What started off as a research project, with some academic collaborations, eventually became a full-fledged product from HP's server group, significant enough to be called out by the CEO in HP's quarterly earnings statements. A year later, Fan, Weber, and Barroso published the other important paper in this area [4] leveraging an elegant distributed systems implementation for web-search-scale serving [4]. Ensemble-level power management has since been adopted by most large-scale datacenter deployments, including follow-on innovations documented by IBM [7], Facebook [15], Google [6], [11], Microsoft [5], [16], etc. The basic architecture (and mechanisms/policies) we outlined are now the approach *de jour* on most deployments.

Systems-Facilities Nexus and "No Power Struggles." Most of the work on enterprise power/thermal density at that time was focused on the facilities level. This paper introduced a novel *systems* approach to thinking about peak power provisioning that led to several follow-on papers from the broader research community. Notably, in discussing future work in the original paper, we identified "*...a promising avenue of research [...] when our ensemble-level control loop is interfaced with local per-server control and broader data center level control of power and cooling.*" Three years later, this led to a highly-cited study at ASPLOS [9] titled "No Power Struggles: A Unified Multi-level Power Management Architecture for the Data Center," which identified an elegant control-theoretic formulation to extend the systems thinking around facilities management both across hierarchical ensembles, and across average and peak power.

Ensembles: Beyond Enclosures to Planet-scale Computing. While our paper focused on blade server enclosures for prototyping, we argued that *"When these systems operate in the context of a larger collection of systems, such as a data center, the inefficiencies* [and benefits] *are compounded."* Later work (e.g., [4], [7]) demonstrated that indeed, the benefits from ensemble management were significantly more pronounced for larger statistical contexts. Additionally, though we focused on power, our work was the first to identify the significant differences between the sum-of-peaks and the peak-of-sums in broader *resource usage* within the datacenter. Later referred to as "resource stranding", this concept led to several follow-on studies targeting non-power use-cases (e.g., CPU oversubscription [3], [13], memory sharing [8]). Interestingly, the two broad classes of policies we identified in the original paper – reactive (*"use unless told you cannot"*) or proactive (*"don't assume, ask"*) – continue to be very relevant in all of these subsequent use-cases.

Modeling and Simulation for Enterprise Architectures. As

[1] A web company (the now commonly-used term "hyperscaler" was coined many years later) called Google was starting to alter many common assumptions about distributed enterprise servers.

[2] Reinforcing this shift from managing a single *box* to an ensemble of servers, one of our early title ideas was "*Out-of-the-box* power management!"

one of the first papers researching enterprise power optimizations, and at the peak of research community's transition to more quantitative and simulation-based research reporting, we had to invent new approaches to modeling and simulating enterprise architectures. Our work was the first to simulate power management in datacenters through resource usage proxies (an approach that was later codified in BladeSim [10]). Our hybrid approach of combining small prototypes and real-world system experiments that fed into more detailed simulations using fleet-scale traces and models was later adopted in several other follow-on studies. Recently, Google released power utilization traces from its fleet for broader research use [6].

III. Retrospective: Lessons and Observations

A Problem that is Still Relevant. Nearly two decades later, this area continues to be important. Interestingly, some of our secondary motivations in the original paper – around datacenter operating costs and sustainability – are the driving motivations now. Since large hyperscalers and cloud providers spend billions of dollars annually on new datacenters, our approach of capping power at the ensemble level has directly led to the current trend of oversubscription and "*the most cost-efficient data center is one that is not built*" movement.

Oversubscription at Multiple Levels and Safety Valves. One of the important elements we missed in our original paper was the power of this approach at multiple dimensions. While we focused on overall power consumption, our approach of oversubscription has been demonstrated at multiple levels – for thermal oversubscription, for cluster-level oversubscription, for busduct-level oversubscription, for fillrate-oversubscription, for storage-oversubscription, etc. Similarly, beyond the (now) relatively simple throttling safety valve we considered, a much richer array of safety valves span all the way from workload quality-of-service tiers to planetary task migration to even physical shutdown and movement of machines and power generators. We had certainly not anticipated many of these!

Considering Machine Learning. A retrospective in the 2020s will likely not be complete without a discussion of machine learning (ML)[3]! While ML was far from our minds when we wrote the original paper, it is clear that machine learning can supplement our control-theoretic approaches to manage power budgets, and also anticipate and react to power oversubscription events. Oversubscription and design of safety valves for future datacenter fleets comprised entirely of machine learning infrastructure, is an interesting area of future research.

IV. Closing remarks

We conclude with a few interesting non-technical anecdotes and observations.

- This paper was originally called "*Enclosure* power management." Luiz Barroso suggested changing the name to *Ensemble* power management. We think this captures the idea much better. Thanks, Luiz!
- David Irwin, the only student author of the original paper, was an intern at HP Labs at the time. Inspired (partly) by this work, David went on to continue broader research on energy and sustainability [1], [2], [12], [14] to this day as a professor at the University of Massachusetts Amherst.
- When this paper was presented at ISCA, a senior researcher in the community puzzled over why "server architecture" papers were being presented at ISCA since: "Server architecture is not computer architecture!" We are happy to report that this researcher is now also working on server architecture research!

We are humbled and gratified at being featured in this ISCA50 retrospective. We continue to be excited by opportunities in this area and hope the insights and contributions from this work continue to pave the way for future advances.

[3]No large-language models (ala ChatGPT or Bard) were used in writing this retrospective!

References

[1] A. S. Bansal, T. Bansal, and D. Irwin, "A Moment in the Sun: Solar Nowcasting from Multispectral Satellite Data using Self-Supervised Learning," in *e-Energy*, June 2022.

[2] N. Bashir, Y. Chandio, D. Irwin, F. Anwar, J. Gummeson, , and P. Shenoy, "Jointly Managing Electrical and Thermal Energy in Solar- and Battery-powered Computer Systems," in *e-Energy*, June 2023.

[3] N. Bashir, N. Deng, K. Rzadca, D. E. Irwin, S. Kodak, and R. Jnagal, "Take it to the Limit: Peak Prediction-driven Resource Overcommitment in Datacenters," in *EuroSys*, April 2021.

[4] X. Fan, W.-D. Weber, and L. A. Barroso, "Power Provisioning for a Warehouse-sized Computer," in *ISCA*, June 2007.

[5] A. G. Kumbhare, R. Azimi, I. Manousakis, A. Bonde, F. V. Frujeri, N. Mahalingam, P. A. Misra, S. A. Javadi, B. Schroeder, M. Fontoura, and R. Bianchini, "Prediction-based power oversubscription in cloud platforms," in *USENIX Annual Technical Conference*, July 2021.

[6] S. Li, X. Wang, X. Zhang, V. Kontorinis, S. Kodakara, D. Lo, and P. Ranganathan, "Thunderbolt: Throughput-Optimized, Quality-of-Service-Aware Power Capping at Scale," in *OSDI*, November 2020.

[7] Y. Li, C. R. Lefurgy, K. Rajamani, M. S. Allen-Ware, G. J. Silva, D. D. Heimsoth, S. Ghose, and O. Mutlu, "A Scalable Priority-Aware Approach to Managing Data Center Server Power," in *HPCA*, February 2019.

[8] K. Lim, J. Chang, T. Mudge, P. Ranganathan, S. K. Reinhardt, and T. F. Wenisch, "Disaggregated memory for expansion and sharing in blade servers," in *ISCA*, June 2009.

[9] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, and X. Zhu, "No "Power" Struggles: Coordinated Multi-level Power Management for the Data Center," in *ASPLOS*, March 2008.

[10] P. Ranganathan and P. Leech, "Simulating Complex Enterprise Workloads using Utilization Traces," in *Workshop on Computer Architecture Evaluation using Commercial Workloads (CAECW)*, February 2007.

[11] V. Sakalkar, V. Kontorinis, D. Landhuis, S. Li, D. D. Ronde, T. Blooming, 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.

[12] A. Souza, N. Bashir, J. Murillo, W. Hanafy, Q. Liang, D. Irwin, and P. Shenoy, "Ecovisor: A Virtual Energy System for Carbon-Efficient Applications," in *ASPLOS*, March 2023.

[13] M. Tirmazi, A. Barker, N. Deng, M. Haque, Z. Qin, S. Hand, M. Harchol-Balter, and J. Wilkes, "Borg: The Next Generation," in *EuroSys*, April 2020.

[14] J. Wamburu, N. Bashir, D. Irwin, and P. Shenoy, "Data-driven Decarbonization of Residential Heating Systems," in *BuildSys*, November 2022.

[15] 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.

[16] 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.