A Global Perspective on Supercomputer Power Provisioning: Case Studies from United States and Europe

Tapasya Patki
Lawrence Livermore National Laboratory
Livermore, USA
patki1@llnl.gov

Barry Rountree
Lawrence Livermore National Laboratory
Livermore, USA
rountree4@llnl.gov

Torsten Wilde
Hewlett-Packard Enterprise
Munich, Germany
wilde@hpe.com

Andrea Bartolini
University of Bologna
Bologna, Italy
a.bartolini@unibo.it

Stephanie Brink
Lawrence Livermore National Laboratory
Livermore, USA
brink2@llnl.gov

Esa Heiskanen
CSC IT Center for Science Ltd.
Kajaani, Finland
esaheiskanen@gmail.com

Sachin Idgunji
NVIDIA Corporation
Santa Clara, USA
sidgunji@nvidia.com

Matthias Maiterth
Oak Ridge National Laboratory
Oak Ridge, USA
maiterthm@ornl.gov

James Rogers
Oak Ridge National Laboratory
Oak Ridge, USA
jrogers@ornl.gov

Ermal Rrapaj
Lawrence Berkeley National Laboratory
Berkeley, USA
ermalrrapaj@lbl.gov

Ralf Schneider
HLRS High Performance Computing Center Stuttgart
Stuttgart, Germany
ralf.schneider@hlrs.de

Woong Shin
Oak Ridge National Laboratory
Oak Ridge, USA
shinw@ornl.gov

Kathleen Shoga
Lawrence Livermore National Laboratory
Livermore, USA
shoga1@llnl.gov

Christian Simmendinger
Hewlett-Packard Enterprise
Munich, Germany
christian.simmendinger@hpe.com

Nicholas J. Wright
Lawrence Berkeley National Laboratory
Berkeley, USA
njwright@lbl.gov

Zhengji Zhao
Lawrence Berkeley National Laboratory
Berkeley, USA
zzhao@lbl.gov

ICS '25, Salt Lake City, UT, USA

Abstract

Electrical provisioning in high performance computing is transitioning from simple nameplate Thermal Design Power (TDP) models to more nuanced approaches based on expected electrical load. This paper captures current power

ACM ISBN 979-8-4007-1537-2/25/06
https://doi.org/10.1145/3721145.3734532

provisioning strategies across six international supercomputing centers and seven systems, three of which (Lumi, Summit, Sierra) were in the top 10 of the Top500 list at the time of data collection[1]. We present longitudinal and summary data of actual power consumption as well as a discussion of how each site approached the question of provisioning. We conclude with a discussion on future directions of hardware overprovisioning and its implications for machine and electrical utilization.

Keywords

High-performance computing, power provisioning

ACM Reference Format:
Tapasya Patki, Barry Rountree, Torsten Wilde, Andrea Bartolini, Stephanie Brink, Esa Heiskanen, Sachin Idgunji, Matthias Maiterth, James Rogers, Ermal Rrapaj, Ralf Schneider, Woong Shin, Kathleen Shoga, Christian Simmendinger, Nicholas J. Wright, and Zhengji Zhao. 2025. A Global Perspective on Supercomputer Power Provisioning: Case Studies from United States and Europe. In *2025 International Conference on Supercomputing (ICS '25), June 08–11, 2025, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3721145.3734532

1 Introduction

It has been just over fifteen years since power was expected to become a "first-order design constraint" [38] for high performance computing (HPC). Having now entered the exascale era[2], the reality is more nuanced than expected. The *Thermal Design Power* (TDP, the theoretical maximum power draw, measured in Watts) of new systems has increased as expected, with *nameplate TDP* (the sum of all components' TDP values) reaching tens of megawatts. However, much as sustained floating point operations per second (FLOPS) have diverged from theoretical or peak FLOPS, actual power consumption on recent systems often fails to reach 50% of nameplate TDP for scientific applications.

When nameplate TDP was still under a megawatt for new systems, provisioning power to that value (referred to as *worst-case provisioning*) was a reasonable choice. While applications may not have been able to reach theoretical maximum power, there was not that much additional cost in ensuring that any application that did could execute successfully. This paper captures a snapshot of the HPC community moving away from conservative nameplate TDP provisioning and taking up *hardware overprovisioning* [59]. We present a longitudinal study of power consumption and power provisioning from some of the world's fastest supercomputers. We discuss the impact of nameplate TDP and worst-case provisioning and provide diverse perspectives on power provisioning. The supercomputers considered at each site, identified by their name and associated Top500 rank at the time of data collection, are:

(1) Perlmutter (#12) and Cori (#60), at National Energy Research Scientific Computing Center (NERSC), United States
(2) Summit (#7), Oak Ridge National Laboratory (ORNL), United States
(3) Sierra (#10), Lawrence Livermore National Laboratory (LLNL), United States
(4) Marconi-100 (#35), CINECA, Italy
(5) Lumi (#5), CSC IT Center for Science Ltd., Finland
(6) Hawk (#42), High-Performance Computing Center Stuttgart (HLRS), Germany

Table 1 shows a high-level overview of the systems presented in this paper and associated datasets. Datasets comprise the computational power usage data collected at the HPC system-level, inclusive of the network and rack-level components. Every effort has been made to make these available publicly for reproducibility. However, two of the seven datasets from this paper will not be publicly available due to privacy policies of the collaborating supercomputing centers.

Figure 1 shows an overview of power usage across all seven systems using a box-and-whiskers plot, which captures the minimum, maximum and the quartile power consumed as a percentage of the nameplate TDP. The shaded region indicates the *provisioned power* (actual power obtained during system procurement).In four of the seven systems (Perlmutter, Cori, Sierra and Lumi), the provisioned power and the nameplate TDP are the same value (worst-case provisioning). For the remainder three systems (Summit, Marconi-100, and Hawk), the provisioned power was *deliberately less than* the nameplate TDP, resulting in a hardware overprovisioned system. As we observe from this overview graph, the maximum power consumed on all seven systems was well under the provisioned power for the timeframes presented in Table 1.

Adding tens of megawatts of new capacity is a significant effort for any site, particularly if much of that capacity remains stranded. As we show across several case studies, supercomputing centers are beginning to move away from provisioning against nameplate TDP. This new approach raises questions about how to handle the risk of applications that consume more power than is currently available; we detail how these risks are mitigated using existing power capping and scheduling capabilities.

To the best of our knowledge, this is the first paper to present a uniquely coordinated effort that discusses the state-of-the-practice in power provisioning across six international supercomputing centers and seven top-tier HPC systems. Given the diversity of the sites and their associated policies,

[1]Several months to years worth of data was collected across seven supercomputers through December 2023.
[2]Supercomputers that can perform up to 10^{18} floating point operations per second are referred to as exascale systems.

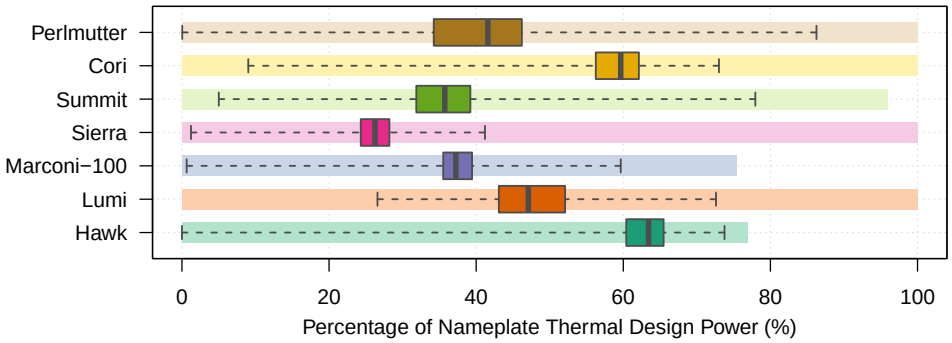

Figure 1: A summary of the quartile data for power consumed as a percentage of Nameplate Thermal Design Power (TDP) for seven HPC systems: Permultter and Cori at National Energy Research Scientific Computing Center (United States), Summit at Oak Ridge National Laboratory (United States), Sierra at Lawrence Livermore National Laboratory (United States), Marconi-100 at CINECA (Italy), Lumi at CSC IT Center for Science Ltd. (Finland) and Hawk at High-Performance Computing Center Stuttgart (Germany) The shaded area provides a visual guideline to represent the provisioned power. For four of the seven systems (Perlmutter, Cori, Sierra and Lumi), the provisioned power is the same as the TDP. Three of the seven systems (Summit, Marconi-100, and Hawk) are hardware-overprovisioned. The maximum power consumed for all systems is always significantly lesser than the provisioned power. Subsequent tables and figures in the paper provide granular detail.

System (Site)	Top500 Rank Nov. 2023 (Highest)	Year Installed	Architecture	Nameplate TDP (MW)	Provisioned Power (MW)	Dataset Begin Date	Dataset End Date	Sampling Rate	Dataset Publicly Available
Perlmutter (LBNL)	12 (5)	2021	HPE Cray EX 235n	6.9	6.9	1/1/23	12/31/23	60 sec.	Yes
Cori (LBNL)	60 (5)	2016	Cray XC40	5.7	5.7	6/1/22	5/31/23	30 sec.	Yes
Summit (ORNL)	7 (1)	2018	IBM AC922	15.022	14.4	1/1/21	12/31/21	15 sec.	No
Sierra (LLNL)	10 (2)	2018	IBM AC922	11	11	-	4.85 yrs.	3 hrs.	No
Marconi-100 (CINECA)	35 (9)	2019	IBM AC922	2.254	1.698	3/19/21	9/28/22	60 sec.	Yes
Lumi (CSC)	5 (3)	2023	HPE Cray EX235a	7.973	7.973	11/7/23	3/14/24	10 min.	Yes
Hawk (HLRS)	42 (16)	2020	HPE Apollo 9000	4.49	3.45	3/1/23	12/31/23	15 min.	Yes

Table 1: Description of the systems and the associated datasets. We show the Top500 rank of the systems from November 2023, which reflects their ranking at the time of data collection. Note that as of January 2025, three of these systems (Summit, Cori and Marconi-100) have been decommissioned. The remaining four systems are still in the Top500, with Lumi ranked at #8, Sierra ranked at #14, Perlmutter ranked at #19, and Hawk ranked at #66 (as of November 2024). Datasets are available at https://github.com/LLNL/LAST/tree/main/Power-Provisioning-Dataset

we focus on capturing the experiences and lessons learned from each site, as opposed to presenting a preferred or favored power provisioning or power management approach. We believe that such case studies can help supercomputing centers and the HPC community make sustainable future procurement decisions by reflecting on existing longitudinal data and learning from provisioning, management, and system software experiences across peer sites.

The remainder of this paper is organized as follows. Section 2 provides a background on hardware overprovisioning and dynamic power management. Section 3 presents the various challenges faced during telemetry and data collection. Sections 4–9 comprise case studies across six HPC sites located across United States and Europe. Section 10 presents a discussion on the current best practices with industry perspectives, as well as challenges faced in adoption of available power management solutions at scale. We summarize the paper in Section 11. Where possible, we provide total system longitudinal power consumption as well as a histogram of power measurements, along with power provisioning details.

2 Motivation

Energy efficiency and power-aware supercomputing research has been underway since the early 2000s. Several disruptive

approaches, such as hardware overprovisioning [59], variable electricity provisioning [83], advanced cooling [6], dynamic power capping with runtime systems and schedulers [20, 25], and co-scheduling [12, 13, 31, 67] have been proposed. While the benefits from these research approaches are significant and have been demonstrated repeatedly, their adoption in Top500 supercomputing systems has been limited.

Concerns from facilities and system administrator teams include: (1) lack of confidence in the electrical safety of techniques such as hardware overprovisioning, (2) limited understanding of system reliability with newer cooling technologies, and (3) lack of techniques to address security for dynamic power management. For users, the impact on the performance of their production workloads is not adequately understood. These issues, when combined with diverse energy efficiency priorities across geographies, have led to a large body of fragmented energy efficiency research in the community that does not get deployed at larger scales.

A key goal of this paper is to facilitate sharing of power provisioning and power management approaches with case studies from some of the world's fastest production supercomputers. We expect these case studies to enable meaningful reflections on past procurement decisions, encourage open communication between peer supercomputing centers, and bring the community together to learn from each others' experiences for future procurements. We believe that such case studies can help mitigate some of the barriers to adoption of existing (and upcoming) power provisioning research at production scales in the future.

2.1 Hardware Overprovisioning

To the best of our knowledge, Kondo's 2007 work [38] contains the first declaration that "power is a first-order design constraint" in parallel computing, and that there was a substantial gap (and opportunity) between nameplate TDP and the power consumed by production workloads. The introduction of Running Average Power Limit technology [63] allowed for hardware-enforced power capping at the processor level, which made cluster-level power control more reliable. Hardware overprovisioning was then reintroduced into the HPC community [59], with subsequent work covering SLURM plugins for power control [68], explorations of control algorithms and power capping techniques [69], economic analysis of hardware overprovisioning [60], and the intersection of overprovisioning and job malleability [12].

Hardware overprovisioning was independently discovered in the datacenter and cloud community, where it goes by the name of "power oversubscription." Early work likewise recognized the gap between nameplate TDP and the power required for production [26, 82]. Examples of more recent work focus on power oversubscription of large language models [54] and opaque virtual machines[41] as well as infrastructural work [43].

A note on definitions. *Hardware overprovisioning* has generally been used to describe adding more hardware in an environment with a fixed amount of power. *Power oversubscription* tends to describe reducing power for a fixed amount of hardware. The definitions have not been used precisely in the literature. The results are the same, of course, and the design and provisioning of new systems will likely adjust both parameters simultaneously.

2.2 Noteworthy Community Efforts

The HPC PowerStack Initiative started in 2017 with the goal of bringing together experts from academia, research laboratories and industry in order to design a holistic and extensible dynamic power management framework [20]. PowerStack explores hierarchical interfaces for dynamic power management at three levels: batch job schedulers, job-level runtime systems, and node-level managers. Two open-source implementations of the PowerStack have been developed. The first was sponsored by the Argo project within the DOE's Exascale Computing Project (ECP), and the latter came out of the European REGALE project funded by the EuroHPC Joint Undertaking. A parallel industry effort came from Intel, with the development of the open-source runtime system, GEOPM (Intel Global Extensible Open Power Manager) [25]. As part of the ECP PowerStack, the 2023 R&D100-winner Variorum was developed, which interfaces with the resource manager Flux, GEOPM, and LDMS [10]. As part of the REGALE project, the resource manager OAR and a runtime system EAR were developed, along with ExaMon and DCDB [8, 14, 49].

A complementary effort is PowerAPI, a portable vendor-neutral API for power measurement and control. It provides multiple levels of abstractions to satisfy the requirements of multiple types of users. The latest specification document is available at the PowerAPI Community website [30]. Vendor-specific implementations of PowerAPI have been deployed across some large-scale supercomputing sites.

The ExaDigiT Group [19, 75] is developing a community-driven digital twin framework of data centers and supercomputers. The primary goal of this effort is to gather experts in modeling, simulation, operational data analytics, telemetry, and AI to advance the development of digital twins of data centers. This includes modeling of cooling systems, electrical supply, and the compute systems hosted in the data centers themselves, along with power-aware job schedulers.

Several other researchers have also studied critical-path optimization under power constraints and power-aware scheduling. Notable research work includes application-level runtime optimization techniques such as Adagio [64] and Conductor [46], dynamic power scheduling policies such as PShifter [28], PERQ [55], DPS [24], Market-based Power Reduction (MPR) [34], window-based data-driven power scheduling [80], variable capacity scheduling [85] and leveraging wasted green power [83]. Several dynamic power-aware scheduling extensions for production-grade resource managers such as SLURM [68, 69, 71, 78] and Flux [40] have also been proposed, along with coordinated power management across components [27]. A detailed survey of power-aware scheduling techniques is presented in [45].

3 Dataset collection challenges

Measuring power and other system-level data at scale on some of the world's fastest supercomputers can pose several engineering challenges [50, 52]. First, power telemetry occurs at different levels in a supercomputer's machine room, and often includes measurements taken at different granularities from multiple sources such as wall meters, rack-level sensors, PDUs, as well as vendor-specific low-level dials such as Intel or AMD's Running Average Power Limit (RAPL) registers, IBM's Open Power Abstraction Layer (OPAL), or NVIDIA's Management Library (NVML). Collating coarse-grained and fine-grained measurements from different sources to obtain a meaningful picture of the entire system's power consumption can take significant post-processing effort [29]. Furthermore, different sites account for different components of the system during provisioning and for telemetry. For example, some sites account for cooling power as part of provisioned power, while others only consider compute power. There are also differences among the components considered when reporting provisioned power and the components considered for telemetry. Many sites do not provide explicit information on whether network and storage power was included in the telemetry data. Table 2 shows the components included in the provisioned and telemetry data for the systems in this paper.

Sites often rely on a combination of vendor-provided operational data analytics frameworks and site-level solutions for post-processing. Some examples of such frameworks include Lightweight Distributed Metric System (LDMS) [11], ExaMon [14], Data Center Database (DCDB) [49], OSISoft PI [51], IBM's Cluster System Management (CSM) [36, 57], Splunk [77] and Redfish [47].

While vendor-provided frameworks come with many advantages, these solutions are not portable, making it challenging to compare data from different HPC systems hosted by the same site. Additionally, maintaining the data collection framework reliably when physical sensors wear out or when meters fail can involve significant delays, resulting in some time windows with inconsistent data. Accurate data collection on high priority systems can be challenging when misconfigurations of specialized hardware require full system downtimes coordinated with outside vendor support to fix issues. Power is often considered a lower-priority metric for monitoring, so a long downtime that needs extra coordination can get deprioritized. Another challenge lies in the area of long-term data storage services. Standardized and portable vendor solutions in this domain are lacking.

System (Site)	Provisioned Power Components	Telemetry Components
Perlmutter (LBNL)	CPUs, GPUs, CDUs, Service Cabinets, & Storage	CPUs, GPUs, & CDUs
Cori (LBNL)	CPUs, CDUs, Service Nodes, & Storage	CPUs, CDUs, & Service Nodes
Summit (ORNL)	All (switchboard-level)	All (switchboard-level)
Sierra (LLNL)	CPUs, GPUs, Storage, Service Cabinets	CPUs, GPUs, Storage, Service Cabinets
Marconi100 (CINECA)	CPUs & GPUs only	CPUs & GPUs only
Lumi (CSC)	Lumi-G GPU partition & network	Lumi-G GPU partition, & network
Hawk (HLRS)	CPUs, GPUs, CDUs, Service Cabinets, & Storage	CPUs, GPUs, CDUs, Service Cabs., & Storage

Table 2: Summary table of components reported in provisioned power and telemetry data.

Power readings are also influenced to some extent by the ambient machine-room temperature and available cooling solutions at the site. While sites make every effort to maintain a standard temperature and cooling setup in their machine rooms, minor seasonal or load-based variations may occur during the course of a system's operation. Also, different sites may choose different cooling and ambient temperature strategies, making it challenging to compare power readings from different sites. Currently, these thermal and cooling effects cannot be easily captured or distinguished in the datasets that measure power independently.

Policies around releasing collected data publicly at a fine granularity vary, and often depend on the workload characteristics at the site and the constraints on the site-level administrators. As an example, in this paper, we observe data collection granularities ranging from 30 seconds to three weeks, and were able to make only five of the seven datasets available for reproducibility.

Community efforts to share insights on expected workloads, electrical provisioning decisions, and power management software are lacking. This paper presents a significant collaboration effort across six top-tier global supercomputing sites and is a first step toward encouraging community-level discussions on electrical power provisioning.

Equipment	Quantity	Total (kW)
GPU Cabinets	14	3869.04
CPU Cabinets	12	2528.40
Cooling Units (CDUs)	8	141.12
Services Cabinets	5	79.32
Storage Cabinets	16	290.61
System TDP	-	6908.49

Table 3: Perlmutter TDP breakdown by components.

4 Perlmutter and Cori Supercomputers

4.1 System Overview

The National Energy Research Scientific Computing Center (NERSC) at the Lawrence Berkeley National Laboratory is the home for Perlmutter [7], an HPE Cray EX system. Perlmutter consists of 1,792 GPU accelerated nodes each with one AMD EPYC 7763 processor (codename: Milan) and four NVIDIA A100 GPUs as well as 3,072 CPU-only nodes with two AMD EPYC 7763 processors each, interconnected with the HPE Slingshot network. Perlmutter has 90 non-compute nodes (service and I/O nodes) and 35 petabytes (PB) of an all flash file system.

NERSC also housed Cori, a Cray XC40 system, which was retired on May 31, 2023, after nearly seven years of service. Cori had 9,688 single-socket 68-core Intel Xeon Phi Processor 7250 ("Knights Landing") nodes and 2,388 dual-socket 16-core Intel Xeon Processor E5-2698 v3 ("Haswell") nodes. Cori nodes were interconnected with Cray's Aries network with Dragonfly topology.

4.2 Workload Characteristics

NERSC supports a broad range of science disciplines for the US Department of Energy (DOE) Office of Science. Workload characteristics are diverse, and include applications from physics, chemistry, biosciences, materials science, fusion energy and many others. A detailed analysis of NERSC workloads, with their parallelism, compute and I/O characteristics is available at [48].

4.3 Power Provisioning and Telemetry

Perlmutter's nameplate TDP is 6.9 MW. This number was provisioned by summing up the component TDPs. See Table 3 for the TDP breakdown by components (last column).

Figure 3.I (upper panel) shows the power usage of Perlmutter in 2023 (from January 1, 2023 to December 31, 2023) and Figure 4.I (left top panel) shows its power distribution histogram. The power data was read from the revenue-grade meters every 60 seconds. During this period, there were no hardware upgrades. The measured power includes GPU cabinets, CPU cabinets and cooling units. The power usage of the storage, and service racks is not included. As shown in the figure, power usage of Perlmutter is at about 50% of TDP most of the time.

Similarly, Figure 3.II (middle panel) shows the power usage of Cori, a retired Cray XC40 system, during its last production year (June 1, 2022 to May 31, 2023). The power data was read from the revenue-grade meters every 30 seconds. The measured power includes compute cabinets, including I/O and service nodes, as well as cooling units. Cori's TDP was 5.7 MW. Figure 4.II (top right panel) shows its power distribution histogram. Cori's power usage was at about 68% of TDP most of the time. Detailed analyses on power usage of Perlmutter and Cori are available in [65], [86], and [16].

NERSC procures a new HPC system approximately every 5 years. NERSC upgrades the facility to accommodate the increasing power requirements for new systems. Usually, NERSC hosts two supercomputers simultaneously, e.g., Cori and Perlmutter had about two years of overlap before Cori was retired. NERSC was at 12.5 MW power before Perlmutter. An additional 12.5 MW power was added for Perlmutter. The provisioned power for NERSC is 25 MW as of this writing, with 20 MW dedicated to HPC loads.

4.4 Power Management

No power management techniques, static or dynamic, were utilized.

5 Summit Supercomputer

5.1 System Overview

The Summit supercomputer at ORNL was a leadership-class system that debuted in 2018 as the fastest system in the world and has recently been decommissioned. It had a hybrid architecture with 4,608 compute nodes. Each node contained two IBM POWER9 CPUs and six NVIDIA Volta GPUs. Each node had over half a terabyte of coherent memory addressable by all CPUs and GPUs, plus 1.6TB of non-volatile RAM that could be used as a burst buffer or as extended memory. The nodes were connected in a non-blocking fat-tree using a dual-rail Mellanox EDR InfiniBand interconnect [9].

5.2 Workload Characteristics

Summit was designed to support open science and artificial intelligence applications. It has been used for research

and development for advanced genomics, earthquake simulations, extreme weather and climate simulations, materials science and physics studies [33].

5.3 Power Provisioning and Telemetry

The TDP for Summit was 15.022 MW, and provisioned power was 14.4 MW, making it a hardware overprovisioned system.

5.3.1 Infrastructure and Power Provisioning. ORNL stipulates strict requirements for their unit substations. These substations, typically consisting of a 3000/4000 Kilovolt-amperes (kVA) [3] liquid-insulated 3-phase 60Hz substation distribution transformer and an integrated 480 Volt (V) 5000 Ampere (A) secondary, must meet specific requirements that include the use of FR3 as the immersion fluid, external forced-air-cooling (KNAF), and a K-4 rating to handle any additional stress (and winding loss) imposed by harmonics [2]. These considerations, along with generally favorable ambient temperature conditions, allows ORNL to comfortably budget power distribution to 90% of the transformer nameplate. Brief excursions above 90% that may be experienced based on load from the HPC systems are well tolerated without impact to the longevity of the transformer.

For Summit, each 3000/4000 kVA unit substation was budgeted to provide 4,000 kVA*90% or 3,600 kVA. Four unit substations were provisioned specifically for the HPC load, providing an aggregate of 4*3,600 kVA or 14,400 kVA (equivalent to 14.4 MW). Line losses from the four unit substations to the main switchboards were small.

5.3.2 Nameplate TDP and Expected Load. From Summit's perspective, the rating of the nodes installed is at 3.36 kVA, with a maximum power consumption at 3,260 W as provided by the Machine Unit Specification by IBM [3]. The original design installed 4608 nodes, fully subscribing the four switchboards. This design gave a total specified capacity rating of 15.483 MVA or a TDP of 15.022 MW.

From experience with previous systems, normal high load operation at 60% of TDP was expected, while High-Performance Linpack (HPL) could briefly reach up to 80%. The compute load at 80% TDP was matched with the 90% of transformer capacity. The combination of 80% HPL and 90% transformer rating provided a comfortable margin, where short periods of high-demand could be absorbed by the surge capacity above 90% on the transformer. In practice, ORNL has never experienced conditions that would cause the unit substation breaker to open due to excessive demand.

[3] kVA is a measure for *apparent* power, which is a combination of true (or *working*) power and reactive power in a circuit. On the other hand, kW is a measure of true power.*Power factor* is the ratio between the true power and apparent power, and is a measure of electrical efficiency. In a system with 100% electrical efficiency, kVA and kW are the same [2].

Applying these expectations to the TDP of 15.022 MW resulted in an expected load of 12.018 MW. With an added safety margin of 5% this was increased to 12.8 MW. The expected maximum load for the system was advertised at 13.0 MW, which was observed by one highly tuned application at acceptance that utilized both CPUs and GPUs with mixed precision capabilities. [37]. The HPL power consumption had a peak of 10.8 MW (including the filesystem), averaging to 10.1 MW. High-Performance Conjugate Gradient (HPCG) had an average power consumption of 6.1 MW. Additional data for system behavior can be found in [73] and [74]. For budgetary considerations, ORNL uses 75% of the measured HPL average. This generated an initial estimate for 7.58 MW. Historically, this has tracked well, with actual average consumption through five years of service at 6.5 MW.

5.3.3 One year of Summit power data: 2021. Figure 3.III presents power data for the Summit system of the year 2021 in normal operation. The measurements are taken at the main switch boards, and sampled on a 15-second interval.

The figure shows that the expected maximum load of 12.8 MW (green dashed line) was not exceeded during the year, while normal operation reached into 70% to 80% of TDP. A notable artifact of the plot is when observing high loads close to 12 MW (blue dots), in general, the running average goes down (black line). This is due to the fact that before a large scale run on the full system, the scheduler has to make room for such job resulting in a lower running averaged power, even if observed loads during the run are very high. The histogram in Figure 4.III shows that while the majority of samples are in the 5000-6000 kW range, the full range of the system was used, as measurements are present in the 11500 - 12000 kW range. The system is a capability system (prioritized for jobs that use no less that 20% of the system), and codes capable of running at full scale were able to use the full potential of the machine without compromises.

5.4 Power Management

The system is hardware overprovisioned. Frequency selection, throttling and other software power management features were discussed, but not put into production as the system should be able to run at full performance if needed by applications, while guaranteeing reliable operation.

6 Sierra Supercomputer

6.1 System Overview

The Sierra supercomputer at Lawrence Livermore National Laboratory is a 125-petaflop system and is the 14th fastest supercomputer in the world as of November 2024. The Sierra supercomputer was installed in 2018 as part of the the CORAL partnership [79]. It is built by IBM in partnership with NVIDIA

Corporation and Mellanox technologies and is a heterogeneous supercomputer that uses IBM Power9 CPUs and NVIDIA Tesla V100 Tensor Core GPUs.

Sierra has a total of 4,320 compute nodes with 190,080 total cores and 17,280 GPUs. Each node has two 22-core, 3.45 GHz IBM POWER9 processors and four NVIDIA V100 GPUs. Two of the cores on each socket are reserved for system use, leaving 40 usable compute cores per node. Each node also has 256 GB of system memory and 64 GB of GPU memory. The nodes are connected by Mellanox EDR InfiniBand at 100 gigabits per second [4].

6.2 Workload Characteristics

Scientists and engineers use Sierra to assess the performance of nuclear weapon systems. These calculations are necessary to understand key issues of physics. This work on Sierra has important implications for other global and national challenges such as nonproliferation and counterterrorism [5].

6.3 Power Provisioning and Telemetry

Figure 2 (left panel) shows a timeline of the power consumption data from the Sierra supercomputer, collected at three week intervals from over 1,765 days (~4.85 years). In the dataset, the median power consumption was 3.186 MW and the maximum power consumption was 4.091 MW. Figure 2 (right panel) shows a histogram of the power consumption data from the Sierra supercomputer, collected at three hour intervals over 1,779 days (~4.85 years). The data in the histogram (right panel) is finer grained than the timeline data presented in the left panel. In the histogram, the median power consumption was 2.888 MW and the maximum power consumption was 4.301 MW. Note that the finer grained dataset was used for the overview figure and table (Figure 1 and Table 1). The provisioned power for Sierra, which is the same as the nameplate TDP in this case, was 11 MW. As can be observed from the data in Figure 2, the system always consumed less than 50% (less than 5.5 MW) of the provisioned power during its operation over multiple years [66].

6.4 Power Management

No power management techniques, static or dynamic, were utilized.

7 Marconi-100 Supercomputer

7.1 System Overview

The Marconi-100 (M100) system was installed at the CINECA datacenter located in Casalecchio di Reno, Italy in early 2020. It was made available on April 20, 2020, with its production use starting on May 4, 2020. The system was decommissioned in late 2023.

Marconi-100 featured 980 IBM Power 9 AC922-GTH computes nodes (with NVIDIA Volta V100 GPUs) assembled in 55 racks. The network was Mellanox Infiniband EDR DragonFly+ at 100 Gb/s. In addition to air cooling, the racks also had cold water-cooled rear-door heat exchangers (RDHx). Roughly half of the heat being removed was from the RDHx and the other half with air-cooling. The room was cooled with six air-conditioning units (CDUs), four of which could operate in free cooling due to their spatial location. The cold water for the RDHx was provided by an external chiller.

7.2 Workload Characteristics

The workloads on Marconi-100 were mostly classical HPC simulations and artificial intelligence applications.

7.3 Power Provisioning and Telemetry

The nameplate power consumption, according to IBM documentation, was 2,300 W for a single node, which sums up to 2,254 kW for the system. The room's provisioned power was configured with an estimated load of 1698 kW for compute and 404 kW for the cooling.

Marconi-100 featured the ExaMon holistic monitoring system[14], which, in the case of the power measurements of the compute cluster, includes two sources: Intelligent Platform Management Interface (IMPI) with out-of-band node-level power measurements and Logics switchboard power measurements. The former reports power data at the compute node level and includes per-component instantaneous power sampled every 20 seconds. Logics, on the other hand reports energy measurements from switchboards including IT and cooling equipment every minute[18]. Figure 3.IV and Figure 4.IV show the computation power usage data with roughly 60-second samples from March 19, 2021 to September 28, 2022. As can be observed, the median power consumption was significantly lower than the nameplate TDP as well as the provisioned power.

7.4 Power Management

The system is hardware overprovisioned. However, no power management techniques, static or dynamic, were utilized.

8 LUMI Supercomputer

8.1 System Overview

The LUMI system is based on the Hewlett Packard Enterprise (HPE) Cray EX architecture. It is the fastest supercomputer in Europe and has a sustained computing power of 380 petaflops. The largest partition of the system is LUMI-G, which consists of 2,978 GPU nodes. Each node has one 64-core AMD Trento CPU and four AMD MI250X GPUs.

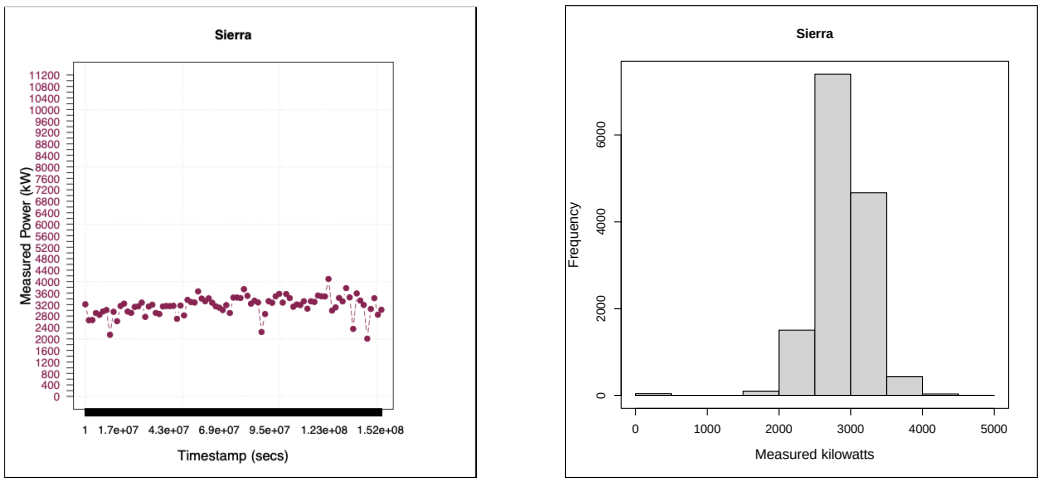

Figure 2: Timeline and histogram of the LLNL Sierra supercomputer. Provisioned power is 11 MW.

Each GPU node features four 200 Gbit/s network interconnect cards and has an 800 Gbit/s injection bandwidth. The MI250X GPU comes with a total of 128 GB of HBM2e memory, offering over 3.2 TB/s of memory bandwidth.

8.2 Workload Characteristics

LUMI has wide user base from different European countries, resulting in a diverse workload. It is one of the world's leading platforms for artificial intelligence applications.

8.3 Power Provisioning and Telemetry

The nameplate TDP of the LUMI-G system is 7,973 kW. Procurement design basis power for provisioning was slightly over the nameplate TDP at 8,000 kW. For the purposes of this paper, we assume the TDP and provisioned power to be 7,973 kW. The maximum power for direct liquid cooled IT-load was 9,300 kW. The latter number also includes the CPU partition, LUMI-C. The TDP of the MI250X GPU (also known as Total Graphics Power, or TGP) is 560 W.

In November 2023, the performance of High-Performance Linpack (HPL) on LUMI-G was 379.7 petaflops and the average power was 7,106.82 kW. For the High-Performance Conjugate Gradient (HPCG) benchmark, the maximum power draw was 7,405.56 kW.

Figures 3.V and 4.V show six months of LUMI-G power usage. The data depicted starts from November 2023, and captures the time after the HPL runs for the Top500 list took place and after the last maintenance break for LUMI ended. Data is collected every 10 minutes. All liquid cooled cabinet components are captured in these measurements, including the high speed slingshot network. The median power consumption was observed to be 3,767 kW, and the maximum power consumption was 5,808 kW. With the exception of HPL and HPCG, the user-level applications and workloads did not consume beyond 73% of provisioned power.

8.4 Power Management

In order to avoid feed breakers from tripping when the GPUs are fully utilized and to fully populate the cabinet to meet the 400 V per distribution line requirement in Europe, the GPUs have to be capped at 500 W each (TGP was 560 W).

9 Hawk Supercomputer

9.1 System Overview

The Hawk supercomputer is the flagship system at HLRS and is one of Europe's fastest computing systems. The main computing partition of the Hawk system is composed of 44 HPE Apollo 9000 racks, each hosting four chassis with eight compute blades. Each blade has four two-socket compute nodes equipped with 256 GB of DDR4 memory and two 64-core AMD Epyc Rome 7742 processors. Additional four racks with 24 Apollo 6500 compute nodes with eight NVIDIA A100 GPUs are deployed for the evaluation and development of hybrid HPC-AI workflows. The system's components and nameplate TDP are specified in Table 4.

9.2 Workload Characteristics

HLRS delivers over 94% of its computing time to academic users for research and development in the engineering domain. The rest is consumed by commercial customers, nearly all of which are deploying simulation applications in Computational Fluid Dynamics (CFD) and structural mechanics. Over 50% of the computing cycles are consumed by CFD applications [32]. Over 75% of these applications can be considered memory-bound, and only a small fraction of these have been ported to a GPU. Based on experience from the

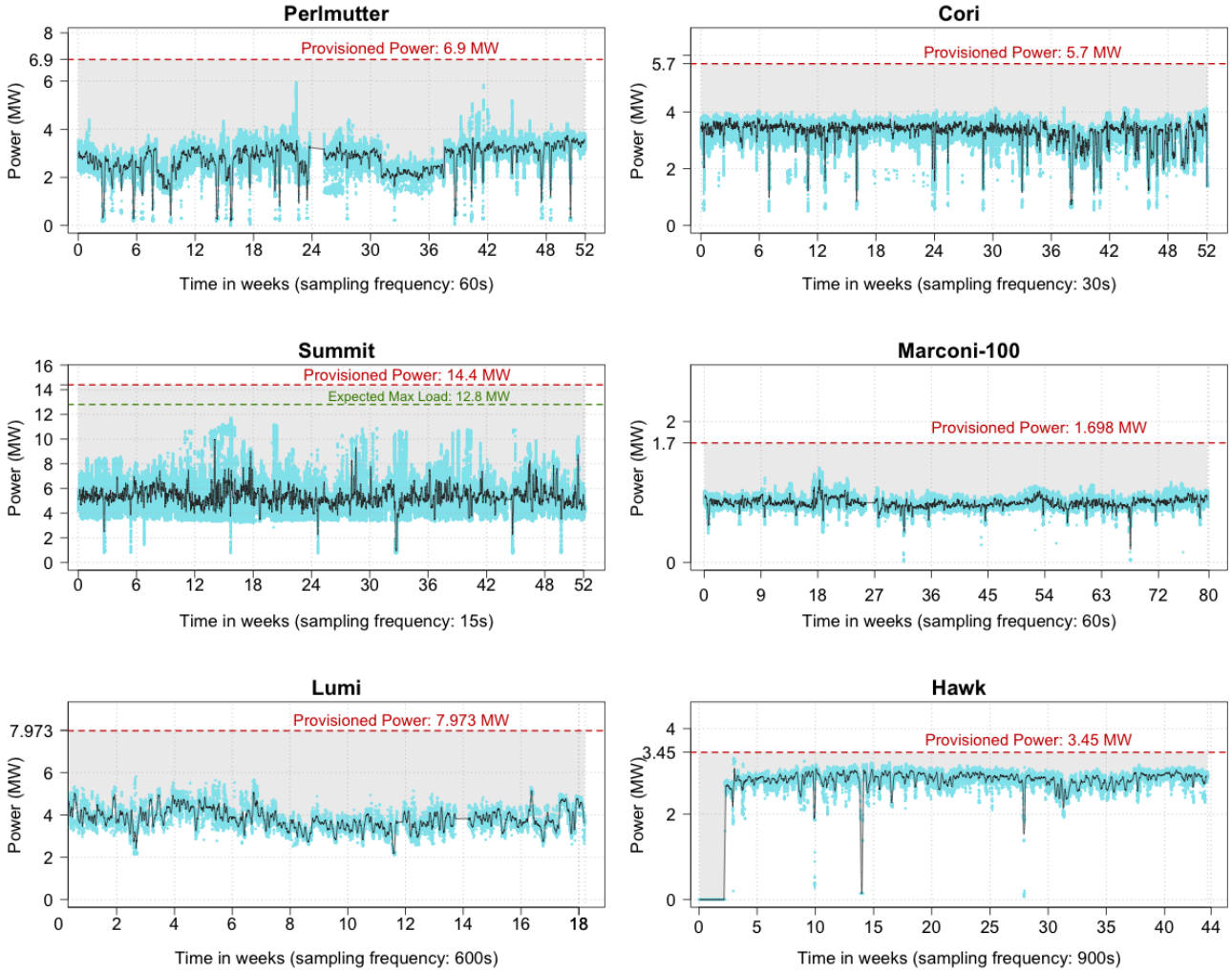

Figure 3: Timeline (blue dots) and the moving average (black line) data from six supercomputers: Perlmutter, Cori, Summit, Marconi-100, Lumi and Hawk. The gray shaded area represents unused power and the red line indicates provisioned power.

operation of Hawk's predecessors, the performance of these applications is not significantly impacted by changing clock frequencies or setting power caps. As a result, a hardware overprovisioned system can successfully deliver the required sustained performance for Hawk's users.

9.3 Power Provisioning and Telemetry

HLRS decided to provision power for Hawk based on the performance and energy requirements of typical user workloads, resulting in a hardware overprovisioned system. The main computing partition's nameplate TDP (4,491 kW) exceeds the provisioned 3,450 kW power of HLRS-II by 21.2%.

Power at HLRS is supplied by two different sources: HLRS-I (1 MW) and HLRS-II (3.45 MW). The main computing partition of Hawk is connected to the HLRS-II power supply. Other components such as the administration and system-infrastructure storage servers and cooling components are connected to HLRS-I. This main partition is organized in three rack-sets with three fly-wheel uninterrupted power supply (UPS) systems. Each rack-set is power capped and the capping factor is shown in Table 5.

Hawk's power data is collected at various levels with different granularities, as follows:

- Per UPS, at the outlets of the three fly-wheel UPS;
- At the sub-distributor level for the administration, storage and Apollo 6500 parts;

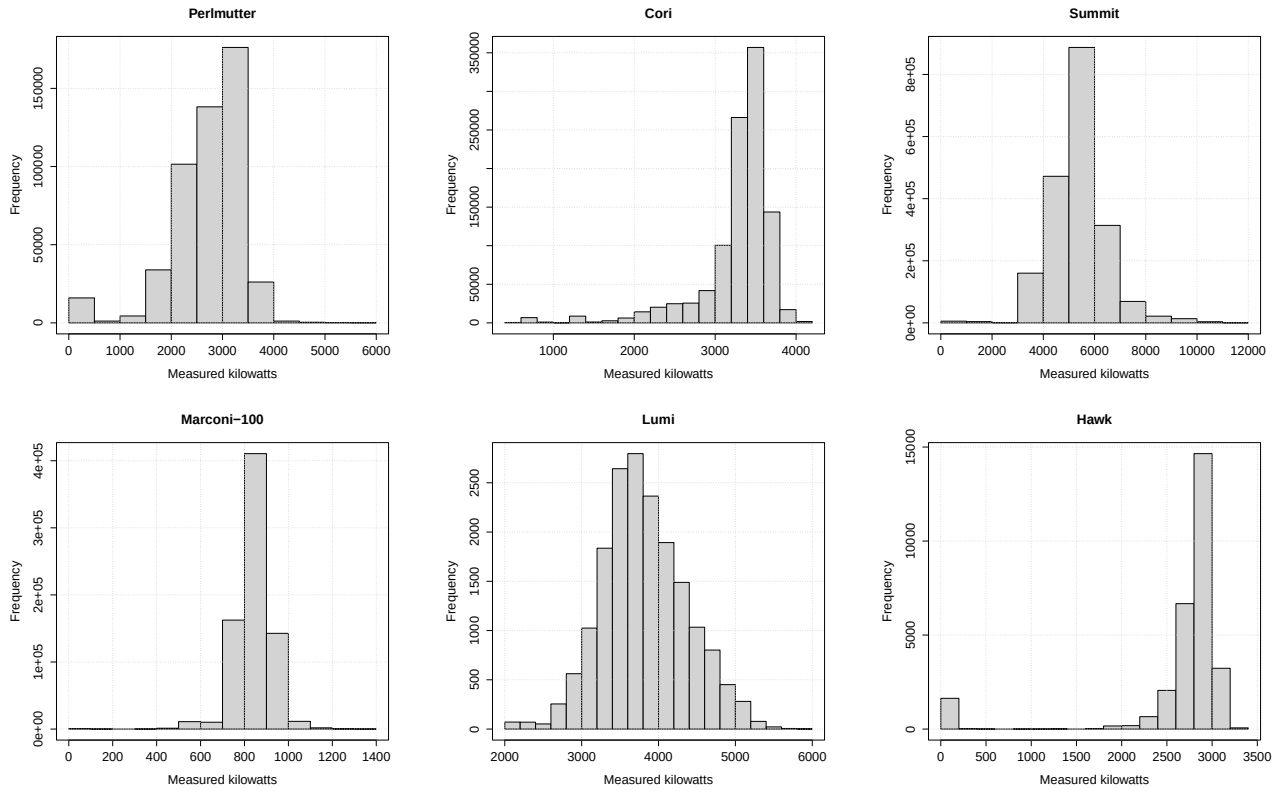

Figure 4: Histograms showing the distribution of power usage from six HPC systems (Sierra shown in Figure 2).

Component	#Racks	Power per rack [kW]	Total power [kW]
Administration & access servers	6	8.3	49.8
In-row chillers	2	1.0	2.0
Storage	7	9.3	65.0
Cooling (CDU)	6	12.0	72.0
Compute Apollo 6500	4	30.6	122.4
Subtotal HLRS I			311.2
Compute Apollo 9000 (HLRS II)	44	95.0	4180.0
Total			4491.2

Table 4: Nameplate TDP of Hawk's components.

Component	#Racks	Total power [kW]	Available power [kW]	Capping factor [-]
rack-set 01-16	16	1520.0	1200.0	0.789
rack-set 17-30	14	1330.0	1125.0	0.846
rack-set 31-44	14	1330.0	1125.0	0.846

Table 5: Power capping of compute racks.

- At the rack level, per power distribution unit (PDU) for all racks and cooling devices; and
- At the node level, by the chassis management controller (CMC) in the Apollo 9000 compute part.

Figures 3.VI and 4.VI show Hawk's power usage between March 1, 2023 and December 31, 2023, reported every 15 minutes. The zero timeframe at the beginning of the time-series is explained by the UPS systems being bypassed and powered down in winter time outside of the thunderstorm season to save energy. The four drops in power consumption visible in the time series also result from bypass mode of the UPS systems due to infrastructure maintenance tasks. As we observe from this dataset, the workloads never exceed the provisioned power, and can be managed well through dynamic power management.

9.4 Power Management

Of the supercomputers discussed in this paper, Hawk is the only system using software provided by HPE for dynamic power management in production.

PowerSched is a proof-of-concept prototype from HPE deployed at HLRS [76]. It implements a reliable, robust, transparent and extensible framework for in-band, application-aware power and energy management. It can manage hardware overprovisioned systems while simultaneously steering

App.	Capping Mode	Power limit per node [W]	Measured power consumption [kW]	Perf. [GFlop/s]
HPL	UC	742	44.5	2.05E+05
HPL	SPC	585	36.8	1.27E+05
HPL	DPC	521-649	38.4	1.56E+05
HPCG	UC	742	40.6	2515
HPCG	SPC	585	35.1	2401
HPCG	DPC	521-649	33.6	2402

Table 6: Impact of power capping on two benchmarks, HPL and HPCG. UC refers to uncapped mode, SPC refers to static power capping, and DPC refers to dynamic power capping.

HPC workloads toward a chosen optimization goal, such as an energy efficiency or a total cost of ownership (TCO) goal.

PowerSched records CPU profiling data while dynamically changing system runtime parameters, such as the available power per CPU package. One key element of working toward an optimization goal is the idea of a *steady state*. A steady state denotes a near time-constant footprint in profiling counters, similar to the state being used in performance projections, albeit on smaller timescales. PowerSched can measure these steady-state footprints transparently with minimal impact on application performance and without user intervention. From this steady state, PowerSched classifies workloads with unsupervised machine learning algorithms or uses a direct optimization strategy for a given energy-runtime metric. The latter approach has been shown to deliver high application performance with minimal energy without requiring extensive training data. It is also extensible for hybrid architectures and is able to handle complex and load-imbalanced applications.

We now demonstrate the effects of the power capping applied to the Hawk system, including the usage of *dynamic* power capping from HPE's PowerSched. We present in Table 6 the average power consumption along with the performance in FLOPs for the HPL (compute-bound) and HPCG (memory-bound) benchmarks, collected across 64 nodes. For each benchmark, we show three operating modes: uncapped execution mode (UC), statically power capped mode at 585 W per node (SPC), and dynamically power capped (DPC mode. We consider these two applications as representatives of compute-bound and memory-bound applications respectively. As can be seen, static power capping has a major impact on the performance of HPL, where static power capping by 21.2% per node reduces performance by 38.0%. On the other hand, minimal impact is observed on HPCG, where the performance reduces only by 4.5% for the same static power cap. The results illustrate that naive static power capping can result in significant slowdowns in compute-bound applications. In order to remedy this situation, HPE's PowerSched framework has been deployed on Hawk since January 2024. This framework distributes the available power budget per power domain (rack-set) in an optimal manner, allowing for maximum performance under minimal power for both frequency bound as well as memory bound applications, as can be observed from DPC rows in Table 6. This is also reflected in Figure 5, which shows the HPCG timeline data.

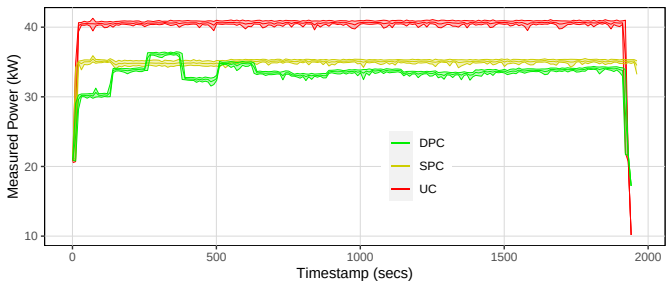

Figure 5: HPCG with Uncapped Execution (UC), Static (SPC) and Dynamic Power Capping (DPC) modes.

10 Discussion

In the previous sections, we presented case studies from seven supercomputers across six sites. In this section, we reflect on the common observations and lessons learned.

10.1 Supercomputers are unique

Supercomputers are procured with different purposes and at different scales across sites as well as within a single site. Some systems are designed for capacity, where the goal is to support a large number of diverse workloads with various job geometries (small-scale to large-scale jobs with a broad distribution of durations). Other systems, such as leadership-class capability systems, are designed for full-scale mission-critical use cases, supporting a more selective set of high-priority applications. On such systems, the goal is to run as fast as possible without necessarily optimizing for system efficiency metrics such as energy consumption or resource utilization. In this paper, we presented examples of both these system types. Summit and Sierra supercomputers were more capability-focused, whereas the other five supercomputers were more capacity-focused.

The purpose of a system drives the type of workloads and jobs that execute on it—ranging from mission-critical applications on capability systems, to open-science applications on capacity systems. The power draw is dependent on the workload, and as a result, it is not particularly useful to compare absolute power values or ranges across different

supercomputers and sites. Workloads can also vary across time and across the machine lifetime, making it challenging to categorize the overall usage characteristics. Often, a mix of jobs are executed on a system, making it difficult to draw universal conclusions about workload behavior. In short, comparing measurements across system should be done with care.

10.2 Most HPC applications are not power-hungry

A key observation from this paper is that regardless of the site-specific workloads differences and the type (capability or capacity) and scale of the system, none of the supercomputers ran at TDP. This is primarily because the most HPC applications are not as power-hungry. It is often incorrectly assumed that giving more power to an application will always improve its performance, and that enforcing a power cap will always slow an application down. While this is true for frequency-bound and computationally intense applications such as HPL, it does not apply to most scientific workflows and applications, *e.g.*, the HPCG example in Section 9. Similar results have been widely presented, where setting power caps or changing frequencies have had little to no effect on the performance of the application [59, 64, 66, 71].

Even as workloads shift toward using AI/ML and multi-binary science workflows [23, 44], similar trends have been observed [39]. A study done to understand performance, power and energy scaling of Large-Language Models (LLMs) on NVIDIA Volta (V100) and Ampere (A100) GPUs shows that the TGP was not reached for either GPU [70]. A bioinformatics study on the Hopper H100 GPU demonstrated that while performance improved by 20x compared to an FPGA implementation, only 550 W of the 700 W TGP was utilized [72]. A longitudinal study on LLMs at the Acme datacenter of Shanghai AI Laboratory with over 4,000 GPUs shows a similar result [35]. Another example of an AI-based workflow is the Multi-scale Machine-Learned Modeling Infrastructure (MuMMI). It is a large-scale, multi-binary and award-winning workflow for cancer research [17]. Capping the per-GPU power from 300 W (peak) to 175 W had no impact on the GPU component (ddCMD molecular dynamics) performance within MuMMI [58]. With some LLM applications, periodic power swings are observed, where the applications reach TDP for a short amount of time and then operate significantly below TDP for majority of the time [53]. Strategies to mitigate such swings with dynamic power management and overprovisioning are being actively researched [53].

Compute-bound applications are likely to consume more power. However, many applications exhibit specific dynamic phase behaviors and tend to be bound by memory, I/O (input/output), and network usage instead. Data transfers across nodes (and between CPUs and GPUs) as well as data staging contribute to significant non-compute time during an application's execution. Similar considerations apply to GPU-based applications. In many cases, only a small portion of the entire workflow can be delegated to the GPU. Also, many GPU-based applications that simulate real-world use cases have branch instructions due to the nature of the underlying problem. This greatly limits the way in which they can utilize a GPU, making the application less compute bound [35, 84].

10.3 Dynamic power management can be successful at production-scale

Under-utilizing servers due to the overestimation of power needs has several consequences. The upfront costs associated with building a high-capacity data center or supercomputer are often significant. Excess infrastructure, including cooling systems and backup generators, represent a substantial financial investment that may not be fully utilized for years, if ever. That excess infrastructure also increases the carbon footprint of such data centers. Maintaining such a center can be more expensive than necessary. Large, underutilized floor spaces in the machine room can present challenges in cooling, potentially leading to wasted energy and higher operational costs. From a user perspective, this leads to limited scalability, where applications cannot run a larger simulation even when enough power is available.

As a result, moving toward a dynamically managed hardware-overprovisioned system may be beneficial, especially for systems that are designed for capacity. In this paper, the Hawk supercomputer demonstrates how a dynamic power management solution can be successful in production. Similarly, the Summit and the Marconi-100 supercomputers demonstrate that hardware overprovisioning can be accomplished while ensuring electrical safety and without compromising on performance, even for leadership-class capability systems. We hope these demonstrations encourage more sites in the Top500 list to consider adopting hardware overprovisioning and dynamic power management techniques.

10.4 Geographic location matters

Electricity pricing as well as local environmental policies can influence power provisioning decisions. As a result, the geographic location of a site can play a role in adoption of hardware overprovisioning and dynamic power management strategies. The relationships and contracts between the electricity service providers and the site determine which power management solutions can be considered.

The Energy-Efficient High Performance Computing Working Group (EEHPC WG) was established to encourage the implementation of energy conservation measures, energy efficient design, and share related ideas in HPC [1]. It has over

900 members worldwide, 50% of which are supercomputing sites, 30% are vendors, and 20% are academic partners. Teams within EEHPC WG have studied the impact of contractual relationships between electricity service providers and supercomputing centers and have surveyed sites across US and Europe to better understand their electricity pricing models. These papers have indicated that the price of electricity is often significantly higher in Europe than it is within United States, and also determined that the overall awareness to establish an efficient energy contract is higher in the United States [15, 22, 56, 62].

Overall, many supercomputing centers are moving toward more sustainable choices. In this paper, all three systems located in Europe (Marconi-100, Lumi, and Hawk) were open to power management solutions. Marconi-100 was hardware overprovisioned, Lumi had GPU power capping enabled during full utilization, and Hawk was both hardware overprovisioned and deploying a dynamic power-aware scheduler. In the United States, only the Summit supercomputer was hardware overprovisioned. None of the systems in the United States used dynamic power management.

10.5 Meaningful energy-efficiency metrics are needed

Traditional metrics for datacenter energy efficiency include floating-point-operations-per-second (FLOPS) per watt and Power-Usage Effectiveness (PUE). Metrics such as Total Cost of Ownership (TCO) are also used to determine the capital and operational expenditures, of which power procurement and energy costs are a significant component. These metrics, while necessary, are not sufficient to evaluate energy efficiency [21]. For example, FLOPS per watt often does not capture the inherent costs of data transfers, storage, networking and I/O. Similarly, average PUE data across an entire supercomputer's lifetime or a large time window is not effective—dynamic and instantaneous PUE metrics are needed.

A correlated and often overlooked metric for energy efficiency is *utilization*: if all the procured components (CPUs, GPUs, memory, network, I/O subsystem) are highly utilized, the system is expected to maximize throughput per watt. Current systems do not measure per-component utilization; and many user workloads either leave GPUs idle or several compute cores idle [57]. Metrics such as Total Usage Effectiveness, TUE, which considers entire IT center [61], and Data center Workload Power Efficiency, DWPE [81], which considers the site-specific workload, have been proposed but are not utilized as often. None of the existing metrics capture carbon footprint or electricity pricing [42], which are also crucial from the perspective of energy efficiency and power provisioning. Designing and utilizing better metrics will allow us to compare and evaluate energy efficiency during provisioning of future systems.

10.6 Testbeds for disruptive research approaches are needed

Testing hardware overprovisioning, power-aware scheduling, and modern cooling infrastructure at scale is necessary for adoption in production systems. Better understanding of electrical safety, system reliability, security, and performance impact can be gained from simulation and longitudinal analysis. Digital twins for supercomputers, such as the one proposed by ExaDigIT [19] can provide a safe and necessary testbed for disruptive energy efficiency research, allowing the community to easily adopt solutions such as hardware overprovisioning in production.

11 Conclusions: Lessons for Future Systems

System design is the art of picking where the bottlenecks will be in a new machine. Infrastructural bottlenecks are the most expensive to remediate. What we have demonstrated in this paper, first and foremost, is that nameplate TDP is a poor estimate of power requirements, and limiting a new system to what can fit within nameplate TDP results in a machine that is likely half the size it could have been. More heterogeneity within a system will make nameplate TDP even less relevant: HPC applications are not designed to be maximally taxing every aspect of the system simultaneously, and those that do are unable to sustain that for very long.

Hardware-enforced power caps that slow execution to stay within a power bound were first introduced into HPC with Intel's Sandy Bridge architecture around 2010, and this capability is now available in all server-class CPUs and GPUs. System software that allows safe execution up to a predefined power limit has been mostly limited to research prototypes, with the notable exceptions of Intel's GEOPM (and, perhaps, the HPE system used in Hawk). This kind of software will be most trusted if it originates with the system integrator, but the assumption that machines must be designed to nameplate TDP prevented any demand for this feature. Based on our survey here, those assumptions are changing.

Sustainability is also becoming a significant concern in future machine design [42], and at first glance that goal appears to conflict with the "Use *all* the power!" approach we advocate here. In the case where better estimates of power consumption lead to provisioning more compute resources, the result is indeed a larger supercomputer that uses more resources. We suspect, though, that the more common case will be users who find they can fit new machines within the existing limits of their power infrastructure without needing to bring in additional power to the site.

Finally, as a community, we have a generations-deep understanding of how to make code run fast. We are now, however, entering a regime where minimizing execution time (and thus maximizing peak power) forces us to accept machines that would be smaller than we would like. Dynamic power management allows us to (mostly) avoid this tradeoff: power is routed to where it does the most good in the moment (be that optimizing for speedup, problem size, accuracy, or throughput) while the total system power remains under the provisioned power bound. Designing machines to run under static power caps is the simple, easy win. The largest gains will require rethinking power as just another schedulable and dynamic resource.

Acknowledgments

Authors wish to thank Pekka Lehtovuori from CSC IT Center for Science Ltd., Finland for help with collection, review and release of the Lumi dataset. Authors thank Norm Bourassa at NERSC for assisting with Cori power data collection and providing the provisioned power for Perlmutter and the NERSC facility. Part of this work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-CONF-862432). Part of this material is based upon work supported by the Advanced Scientific Computing Research Program in the U.S. Department of Energy, Office of Science, under Award Number DE-AC02-05CH11231 and used resources of the National Energy Research Scientific Computing Center (NERSC), which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Part of this research was sponsored by and used resources of the Oak Ridge Leadership Computing Facility (OLCF), which is a DOE Office of Science User Facility at the Oak Ridge National Laboratory (ORNL) supported by the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This work was partially funded by the German Federal Ministry of Education and Research (BMBF) and the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) as part of the SiVeGCS project.

References

[1] [n. d.]. Energy Efficient High Performance Computing Working Group (EE HPC WG). https://eehpcwg.lbl.gov

[2] 2010. *IEEE Standard Terminology for Power and Distribution Transformers, IEEE Std C57.12.80-2010 (Revision of IEEE Std C57.12.80-2002).* IEEE Standards Organization. 1–60 pages.

[3] 2018. Model 8335-GTW and 8335-GTX server specifications. https://www.ibm.com/docs/en/power9/8335-GTX?topic=specifications-model-8335-gtw-8335-gtx-server

[4] 2018. Sierra: NNSA's largest and most advanced supercomputer will help solve the nation's most demanding computational challenges in support of nuclear security. https://asc.llnl.gov/sites/asc/files/sierra-fact-sheet.pdf

[5] 2018. Sierra Supercomputer. https://hpc.llnl.gov/hardware/compute-platforms/sierra

[6] 2023. NREL Joins $40 Million Effort To Advance Data Center Cooling Efficiency. https://www.nrel.gov/news/program/2023/nrel-joins-effort-to-advance-data-center-cooling-efficiency.html

[7] 2023. Perlmutter Architecture. Retrieved August 1, 2023 from https://docs.nersc.gov/systems/perlmutter/architecture/

[8] 2023. REGALE: Open Architecture for Exascale Supercomputers, Software Suite. https://regale-project.eu/?page_id=400

[9] 2023. Summit: Oak Ridge National Laboratory's 200 petaflop supercomputer. https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/

[10] 2023. Variorum: Vendor-Agnostic Computing Power Management. https://ipo.llnl.gov/sites/default/files/2023-08/Final_variorum-rnd-100-award.pdf

[11] Anthony Agelastos, Benjamin Allan, Jim Brandt, Paul Cassella, Jeremy Enos, Joshi Fullop, Ann Gentile, Steve Monk, Nichamon Naksinehaboon, Jeff Ogden, Mahesh Rajan, Michael Showerman, Joel Stevenson, Narate Taerat, and Tom Tucker. 2014. The Lightweight Distributed Metric Service: A Scalable Infrastructure for Continuous Monitoring of Large Scale Computing Systems and Applications. In *SC '14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 154–165. https://doi.org/10.1109/SC.2014.18

[12] Eishi Arima, A Isaías Comprés, and Martin Schulz. 2022. On the Convergence of Malleability and the HPC PowerStack: Exploiting Dynamism in Over-provisioned and Power-constrained HPC Systems. In *ISC-HPC*.

[13] Eishi Arima, Minjoon Kang, Issa Saba, Josef Weidendorfer, Carsten Trinitis, and Martin Schulz. 2022. Optimizing Hardware Resource Partitioning and Job Allocations on Modern GPUs under Power Caps. In *Workshop Proceedings of the 51st International Conference on Parallel Processing*. https://doi.org/10.1145/3547276.3548630

[14] Andrea Bartolini, Francesco Beneventi, Andrea Borghesi, Daniele Cesarini, Antonio Libri, Luca Benini, and Carlo Cavazzoni. 2019. Paving the way toward energy-aware and automated datacentre. In *Workshop Proceedings of the 48th International Conference on Parallel Processing (ICPP Workshops 19)*. 1–8. https://doi.org/10.1145/3339186.3339215

[15] Natalie Bates, Girish Ghatikar, Ghaleb Abdulla, Gregory A Koenig, Sridutt Bhalachandra, Mehdi Sheikhalishahi, Tapasya Patki, Barry Rountree, and Stephen Poole. 2015. Electrical grid and supercomputing centers: An investigative analysis of emerging opportunities and challenges. *Informatik-Spektrum* 38 (2015), 111–127.

[16] Sridutt Bhalachandra, Brian Austin, and Nicholas J. Wright. 2021. Understanding power variation and its implications on performance optimization on the Cori supercomputer. In *2021 International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS)*. 51–62. https://doi.org/10.1109/PMBS54543.2021.00011

[17] Harsh Bhatia, Francesco Di Natale, Joseph Y. Moon, Xiaohua Zhang, Joseph R. Chavez, Fikret Aydin, Chris Stanley, Tomas Oppelstrup, Chris Neale, Sara Kokkila Schumacher, Dong H. Ahn, Stephen Herbein, Timothy S. Carpenter, Sandrasegaram Gnanakaran, Peer-Timo Bremer, James N. Glosli, Felice C. Lightstone, and Helgi I. Ingolfsson. 2021. Generalizable Coordination of Large Multiscale Workflows: Challenges and Learnings at Scale. In *SC21: International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–16. https://doi.org/10.1145/3458817.3476210

[18] Andrea Borghesi, Carmine Di Santi, Martin Molan, Mohsen Seyedkazemi Ardebili, Alessio Mauri, Massimiliano Guarrasi, Daniela Galetti, Mirko Cestari, Francesco Barchi, Luca Benini, Francesco Beneventi, and Andrea Bartolini. 2023. M100 ExaData: a data collection campaign on the CINECA Marconi100 Tier-0 supercomputer. *Scientific Data* 10, 1 (2023), 288. https://doi.org/10.1038/s41597-023-02174-3

[19] Wesley Brewer, Matthias Maiterth, Vineet Kumar, Rafal Wojda, Sedrick Bouknight, Jesse Hines, Woong Shin, Scott Greenwood, David Grant, Wesley Williams, and Feiyi Wang. 2024. A Digital Twin Framework for Liquid-cooled Supercomputers as Demonstrated at Exascale. In *SC24: International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–18. https://doi.org/10.1109/SC41406.2024.00029

[20] Christopher Cantalupo, Jonathan Eastep, Siddhartha Jana, Masaaki Kondo, Matthias Maiterth, Aniruddha Marathe, Tapasya Patki, Barry Rountree, Ryuichi Sakamoto, Martin Schulz, et al. 2018. *A Strawman for an HPC powerstack*. Technical Report. https://www.osti.gov/biblio/1466153

[21] Andrew A. Chien, Chaojie Zhang, Liuzixuan Lin, and V. Trivikram Rao. 2022. Beyond PUE: Flexible Datacenters Empowering the Cloud to Decarbonize. https://par.nsf.gov/servlets/purl/10400420

[22] Anders Clausen, Gregory Koenig, Sonja Klingert, Girish Ghatikar, Peter M. Schwartz, and Natalie Bates. 2019. An Analysis of Contracts and Relationships between Supercomputing Centers and Electricity Service Providers. In *Workshop Proceedings of the 48th International Conference on Parallel Processing* (Kyoto, Japan) *(ICPP Workshops '19)*. Association for Computing Machinery, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3339186.3339209

[23] Alex de Vries. 2023. The growing energy footprint of artificial intelligence. *Joule* 7, 10 (2023), 2191–2194. https://doi.org/10.1016/j.joule.2023.09.004

[24] Jianru Ding and Henry Hoffmann. 2023. DPS: Adaptive Power Management for Overprovisioned Systems. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '23)*. Association for Computing Machinery, New York, NY, USA, Article 27, 14 pages. https://doi.org/10.1145/3581784.3607091

[25] Jonathan Eastep, Steve Sylvester, Christopher Cantalupo, Brad Geltz, Federico Ardanaz, Asma Al-Rawi, Kelly Livingston, Fuat Keceli, Matthias Maiterth, and Siddhartha Jana. 2017. Global Extensible Open Power Manager: A Vehicle for HPC Community Collaboration on Co-Designed Energy Management Solutions. In *High Performance Computing*, Julian M. Kunkel, Rio Yokota, Pavan Balaji, and David Keyes (Eds.). Springer International Publishing, Cham, 394–412.

[26] Xiaobo Fan, Wolf-Dietrich, and Luiz Andre Barroso. 2007. Power Provisioning for a Warehouse-sized Computer. *ACM SIGARCH Computer Architecture News* 35, 2 (Jun 2007), 13–23.

[27] Rong Ge, Xizhou Feng, Tyler Allen, and Pengfei Zou. 2021. The Case for Cross-Component Power Coordination on Power Bounded Systems. *IEEE Transactions on Parallel and Distributed Systems* 32, 10 (2021), 2464–2476. https://doi.org/10.1109/TPDS.2021.3068235

[28] Neha Gholkar, Frank Mueller, Barry Rountree, and Aniruddha Marathe. 2018. Pshifter: Feedback-based dynamic power shifting within hpc jobs for performance. In *Proceedings of the 27th International Symposium on High-Performance Parallel and Distributed Computing*. 106–117.

[29] Alfredo Giménez, Todd Gamblin, Abhinav Bhatele, Chad Wood, Kathleen Shoga, Aniruddha Marathe, Peer-Timo Bremer, Bernd Hamann, and Martin Schulz. 2017. ScrubJay: deriving knowledge from the disarray of HPC performance data. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (Denver, Colorado) *(SC '17)*. Association for Computing Machinery, New York, NY, USA, Article 35, 12 pages. https://doi.org/10.1145/3126908.3126935

[30] R. E. Grant, M. Levenhagen, S. L. Olivier, D. DeBonis, K. T. Pedretti, and J. H. Laros III. 2016. Standardizing Power Monitoring and Control at Exascale. *Computer* 49, 10 (Oct 2016), 38–46. https://doi.org/10.1109/MC.2016.308

[31] Jason Hall, Arjun Lathi, David K. Lowenthal, and Tapasya Patki. 2023. Evaluating the Potential of Coscheduling on High-Performance Computing Systems. In *Job Scheduling Strategies for Parallel Processing (JSSPP)*.

[32] High-Performance Computing Center Stuttgart (HLRS). 2023. Next-Generation Supercomputing, 2022 Annual Report. https://www.hlrs.de/fileadmin/about/Annual_Report/HLRS-Annual_Report_2022.pdf

[33] Jonathan Hines. 2018. Five Gordon Bell Finalists Credit Summit for Vanguard Computational Science. https://www.olcf.ornl.gov/2018/09/17/uncharted-territory/

[34] Md Rajib Hossen, Kishwar Ahmed, and Mohammad A. Islam. 2023. Market Mechanism-Based User-in-the-Loop Scalable Power Oversubscription for HPC Systems. In *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. 485–498. https://doi.org/10.1109/HPCA56546.2023.10071006

[35] Qinghao Hu, Zhisheng Ye, Zerui Wang, Guoteng Wang, Meng Zhang, Qiaoling Chen, Peng Sun, Dahua Lin, Xiaolin Wang, Yingwei Luo, Yonggang Wen, and Tianwei Zhang. 2024. Characterization of Large Language Model Development in the Datacenter. arXiv:2403.07648 [cs.DC]

[36] IBM. [n. d.]. IBM Cluster Systems Management. https://www.ibm.com/docs/en/power6?topic=overview-cluster-systems-management

[37] Wayne Joubert, Deborah Weighill, David Kainer, Sharlee Climer, Amy Justice, Kjiersten Fagnan, and Daniel Jacobson. 2018. Attacking the Opioid Epidemic: Determining the Epistatic and Pleiotropic Genetic Architectures for Chronic Pain and Opioid Addiction. In *SC18: International Conference for High Performance Computing, Networking, Storage and Analysis*. 717–730. https://doi.org/10.1109/SC.2018.00060

[38] Masaaki Kondo, Yoshimichi Ikeda, and Hiroshi Nakamura. 2007. A High Performance Cluster System Design by Adapti[v]e Power Control. In *IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 1–8. https://doi.org/10.1109/IPDPS.2007.370535

[39] Adam Krzywaniak, Pawel Czarnul, and Jerzy Proficz. 2022. GPU Power Capping for Energy-Performance Trade-Offs in Training of Deep Convolutional Neural Networks for Image Recognition. In *Computational Science – ICCS 2022*, Derek Groen, Clélia de Mulatier, Maciej Paszynski, Valeria V. Krzhizhanovskaya, Jack J. Dongarra, and Peter M. A. Sloot (Eds.). Springer International Publishing.

[40] Naman Kulshreshtha, Tapasya Patki, Jim Garlick, Mark Grondona, and Rong Ge. 2024. Vendor-neutral and Production-grade Job Power Management in High Performance Computing. In *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1845–1855. https://doi.org/10.1109/SCW63240.2024.00231

[41] Alok Gautam Kumbhare, Reza Azimi, Ioannis Manousakis, Anand Bonde, Felipe Frujeri, Nithish Mahalingam, Pulkit A. Misra, Seyyed Ahmad Javadi, Bianca Schroeder, Marcus Fontoura, and Ricardo Bianchini. 2021. Prediction-Based Power Oversubscription in Cloud Platforms. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*. 473–487.

[42] Baolin Li, Rohan Basu Roy, Daniel Wang, Siddharth Samsi, Vijay Gadepally, and Devesh Tiwari. 2023. Toward Sustainable HPC: Carbon Footprint Estimation and Environmental Implications of HPC Systems. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM. https://doi.org/10.1145/3581784.3607035

[43] Shaohong Li, Xi Wang, Xiao Zhang, Vasileios Kontorinis, Sreekumar Kodakara, David Lo, and Parthasarathy Ranganathan. 2020. Thunderbolt: Throughput-Optimized, Quality-of-Service-Aware Power Capping at Scale. In *Proceedings of the 14th USENIX Symposium on Operating Systems Design and Implementation*. 1241–1255.

[44] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2022. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. arXiv:2211.02001 [cs.LG]

[45] Matthias Maiterth, Gregory Koenig, Kevin Pedretti, Siddhartha Jana, Natalie Bates, Andrea Borghesi, Dave Montoya, Andrea Bartolini, and

Milos Puzovic. 2018. Energy and Power Aware Job Scheduling and Resource Management: Global Survey — Initial Analysis. In *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. 685–693. https://doi.org/10.1109/IPDPSW.2018.00111

[46] Aniruddha Marathe, Peter E Bailey, David K Lowenthal, Barry Rountree, Martin Schulz, and Bronis R de Supinski. 2015. A run-time system for power-constrained HPC applications. In *High Performance Computing: 30th International Conference, ISC High Performance 2015, Frankfurt, Germany, July 12-16, 2015, Proceedings 30*. Springer, 394–408.

[47] Steven J. Martin, David Rush, Kevin Hughes, and Matt Kelly. [n. d.]. Modernizing Cray Systems Management: Use of Redfish APIs on Next Generation Cray Hardware. In *Cray User Group Meeting*. ACM. https://cug.org/proceedings/cug2018_proceedings/includes/files/pap107s2-file1.pdf

[48] NERSC. 2025. *NERSC 10 Workload Analysis*. Technical Report. National Energy Research Scientific Computing Center (NERSC). https://portal.nersc.gov/project/m888/nersc10/workload/N10_Workload_Analysis.latest.pdf Accessed: 2025-04-29.

[49] Alessio Netti, Micha Müller, Axel Auweter, Carla Guillen, Michael Ott, Daniele Tafani, and Martin Schulz. 2019. From facility to application sensor data: modular, continuous and holistic monitoring with DCDB. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM. https://doi.org/10.1145/3295500.3356191

[50] Alessio Netti, Woong Shin, Michael Ott, Torsten Wilde, and Natalie Bates. 2021. A Conceptual Framework for HPC Operational Data Analytics. In *2021 IEEE International Conference on Cluster Computing (CLUSTER)*. 596–603. https://doi.org/10.1109/Cluster48925.2021.00086

[51] OSISoft. [n. d.]. OSISoft PI System. https://techsupport.osisoft.com/products/

[52] Michael Ott, Woong Shin, Norman Bourassa, Torsten Wilde, Stefan Ceballos, Melissa Romanus, and Natalie Bates. 2020. Global Experiences with HPC Operational Data Measurement, Collection and Analysis. In *2020 IEEE International Conference on Cluster Computing (CLUSTER)*. 499–508. https://doi.org/10.1109/CLUSTER49012.2020.00071

[53] Pratyush Patel, Esha Choukse, Chaojie Zhang, Íñigo Goiri, Brijesh Warrier, Nithish Mahalingam, and Ricardo Bianchini. 2024. Characterizing Power Management Opportunities for LLMs in the Cloud. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3* (La Jolla, CA, USA) *(ASPLOS '24)*. Association for Computing Machinery, New York, NY, USA, 207–222. https://doi.org/10.1145/3620666.3651329

[54] Pratyush Patel, Esha Choukse, Chaojie Zhang Íñigo Goiri, Brijesh Warrier, Nithish Mahalingam, and Ricardo Bianchini. 2023. POLCA: Power Oversubscription in LLM Cloud Providers. (Aug 2023). arXiv:2308.12908 [cs.DC].

[55] Tirthak Patel and Devesh Tiwari. 2019. PERQ: Fair and Efficient Power Management of Power-Constrained Large-Scale Computing Systems. In *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing* (Phoenix, AZ, USA) *(HPDC '19)*. Association for Computing Machinery, New York, NY, USA, 171–182. https://doi.org/10.1145/3307681.3326607

[56] Tapasya Patki, Natalie Bates, Girish Ghatikar, Anders Clausen, Sonja Klingert, Ghaleb Abdulla, and Mehdi Sheikhalishahi. 2016. Supercomputing centers and electricity service providers: A geographically distributed perspective on demand management in Europe and the United States. In *High Performance Computing: 31st International Conference, ISC High Performance 2016, Frankfurt, Germany, June 19-23, 2016, Proceedings*. Springer, 243–260.

[57] Tapasya Patki, Adam Bertsch, Ian Karlin, Dong H. Ahn, Brian Van Essen, Barry Rountree, Bronis R. de Supinski, and Nathan Besaw. 2021. Monitoring Large Scale Supercomputers: A Case Study with the Lassen Supercomputer. In *2021 IEEE International Conference on Cluster Computing (CLUSTER)*. 468–480. https://doi.org/10.1109/Cluster48925.2021.00057

[58] Tapasya Patki, Zachary Frye, Harsh Bhatia, Francesco Di Natale, James Glosli, Helgi Ingolfsson, and Barry Rountree. 2019. Comparing GPU Power and Frequency Capping: A Case Study with the MuMMI Workflow. In *2019 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS)*. 31–39. https://doi.org/10.1109/WORKS49585.2019.00009

[59] Tapasya Patki, David K. Lowenthal, Barry Rountree, Martin Schulz, and Bronis R. de Supinski. 2013. Exploring hardware overprovisioning in power-constrained, high performance computing. In *Proceedings of the 27th international ACM conference on International conference on supercomputing (ICS13)*. 173–182. https://doi.org/10.1145/2464996.2465009

[60] Tapasya Patki, David K. Lowenthal, Barry Rountree, Martin Schulz, and Bronis R. de Supinski. 2016. Economic Viability of Hardware Overprovisioning in Power-Constrained High Performance Computing. In *Fourth International Workshop on Energy Efficient Supercomputing (E2SC)*. 8–15. https://doi.org/10.1109/E2SC.2016.007

[61] Michael K. Patterson, Stephen W. Poole, Chung-Hsing Hsu, Don Maxwell, William Tschudi, Henry Coles, David J. Martinez, and Natalie Bates. 2013. TUE, a New Energy-Efficiency Metric Applied at ORNL's Jaguar. In *Supercomputing*, Julian Martin Kunkel, Thomas Ludwig, and Hans Werner Meuer (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 372–382.

[62] Marcin Pospieszny, Jean-Philippe Nominé, Ladina Gilly, François Robin, and Radosław Januszewski Norbert Meyer. [n. d.]. Electricity in HPC Centres. *Partnership for Advanced Computing in Europe (PRACE)* ([n. d.]). https://prace-ri.eu/wp-content/uploads/hpc-centre-electricity-whitepaper-2.pdf

[63] Barry Rountree, Dong Ahn, Bronis R. de Supinski, David K. Lowenthal, and Martin Schulz. 2012. Beyond DVFS: A First Look at Performance Under a Hardware-Enforced Power Bound. In *8th Workshop on High-Performance, Power-Aware Computing (HPPAC)*.

[64] Barry Rountree, David K. Lownenthal, Bronis R. de Supinski, Martin Schulz, Vincent W. Freeh, and Tyler Bletsch. 2009. Adagio: making DVS practical for complex HPC applications. In *Proceedings of the 23rd International Conference on Supercomputing* (Yorktown Heights, NY, USA) *(ICS '09)*. Association for Computing Machinery, New York, NY, USA, 460–469. https://doi.org/10.1145/1542275.1542340

[65] Ermal Rrapaj, Sridutt Bhalachandra, Zhengji Zhao, Brian Austin, Hai Ah Nam, and Nicholas Wright. 2024. Power Consumption Trends in Supercomputers: A Study of NERSC's Cori and Perlmutter Machines. In *To appear in Proceedings of the ISC High Performance '24*.

[66] Brian S. Ryujin, Arturo Vargas, Ian Karlin, Shawn A. Dawson, Kenneth Weiss, Adam Bertsch, Michael Scott McKinley, Michael R. Collette, Simon D. Hammond, Kevin T. Pedretti, and Robert N. Rieben. 2022. Understanding Power and Energy Utilization in Large Scale Production Physics Simulation Codes. *CoRR* abs/2201.01278 (2022). arXiv:2201.01278 https://arxiv.org/abs/2201.01278

[67] Issa Saba, Eishi Arima, Dai Liu, and Martin Schulz. 2022. Orchestrated Co-scheduling, Resource Partitioning, and Power Capping on CPU-GPU Heterogeneous Systems via Machine Learning. In *Architecture of Computing Systems*.

[68] Ryuichi Sakamoto, Thang Cao, Masaaki Kondo, Koji Inoue, Masatsugu Ueda, Tapasya Patki, Danielle Ellsworth, Barry Rountree, and Martin Schulz. 2017. Production Hardware Overprovisioning: Real-World Performance Optimization Using an Extensible Power-Aware Resource Management Framework. In *IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 957–966. https://doi.org/10.1109/IPDPS.2017.107

[69] Ryuichi Sakamoto, Tapasya Patki, Thang Cao, Masaaki Kondo, Koji Inoue, Masatsugu Ueda, Danielle Ellsworth, Barry Rountree, and Martin Schulz. 2018. Analyzing Resource Trade-offs in Hardware Overprovisioned Supercomputers. In *IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 526–535. https://doi.org/10.1109/IPDPS.2018.00062

[70] Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadepally. 2023. From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. arXiv:2310.03003 [cs.CL]

[71] Osman Sarood, Akhil Langer, Abhishek Gupta, and Laxmikant Kale. 2014. Maximizing throughput of overprovisioned hpc data centers under a strict power budget. In *SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 807–818.

[72] Bertil Schmidt, Felix Kallenborn, Alejandro Chacon, and Christian Hundt. 2023. CUDASW++4.0: Ultra-fast GPU-based Smith-Waterman Protein Sequence Database Search. *bioRxiv* (2023). https://doi.org/10.1101/2023.10.09.561526

[73] Woong Shin, J. Austin Ellis, Ahmad Maroof Karimi, Vladyslav Oles, Sajal Dash, and Feiyi Wang. 2022. Long Term Per-Component Power and Thermal Measurements of the OLCF Summit System. (4 2022). https://doi.org/10.13139/OLCF/1861393

[74] Woong Shin, Vladyslav Oles, Ahmad Maroof Karimi, J. Austin Ellis, and Feiyi Wang. 2021. Revealing power, energy and thermal dynamics of a 200PF pre-exascale supercomputer. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC21)*. https://doi.org/10.1145/3458817.3476188

[75] Woong Shin, James B. White, Wael Elwasif, Rafael Ferreira Da Silva, Christopher Zimmer, Bronson Messer, Reuben Budiardja, Antigoni Georgiadou, Verónica Melesse Vergara, Jack Lange, Matthias Maiterth, Tim Osborne, Leah Huk, John Holmen, Nick Hagerty, Ahmad Maroof Karimi, Thomas Naughton, Ryan Adamson, Ryan Prout, Feiyi Wang, Scott Atchley, Kevin G. Thach, Thomas Beck, and Sarp Oral. 2024. Towards Sustainable Post-Exascale Leadership Computing. In *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1790–1794. https://doi.org/10.1109/SCW63240.2024.00225

[76] Christian Simmendinger, Marcel Marquardt, Jan Mäder, and Ralf Schneider. 2024. PowerSched - Managing Power Consumption in Overprovisioned Systems. 1–8. https://doi.org/10.1109/CLUSTERWorkshops61563.2024.00012

[77] Splunk. [n. d.]. Splunk OpenTelemetry and Performance Monitoring Frameworks. https://www.splunk.com/en_us/solutions/opentelemetry.html

[78] Tapan Srivastava, Huazhe Zhang, and Henry Hoffmann. 2023. Penelope: Peer-to-peer Power Management. In *Proceedings of the 51st International Conference on Parallel Processing* (Bordeaux, France) *(ICPP '22)*. Association for Computing Machinery, New York, NY, USA, Article 43, 11 pages. https://doi.org/10.1145/3545008.3545047

[79] Sudharshan S. Vazhkudai, Bronis R. de Supinski, Arthur S. Bland, Al Geist, James Sexton, Jim Kahle, Christopher J. Zimmer, Scott Atchley, Sarp Oral, Don E. Maxwell, Veronica G. Vergara Larrea, Adam Bertsch, Robin Goldstone, Wayne Joubert, Chris Chambreau, David Appelhans, Robert Blackmore, Ben Casses, George Chochia, Gene Davison, Matthew A. Ezell, Tom Gooding, Elsa Gonsiorowski, Leopold Grinberg, Bill Hanson, Bill Hartner, Ian Karlin, Matthew L. Leininger, Dustin Leverman, Chris Marroquin, Adam Moody, Martin Ohmacht, Ramesh Pankajakshan, Fernando Pizzano, James H. Rogers, Bryan Rosenburg, Drew Schmidt, Mallikarjun Shankar, Feiyi Wang, Py Watson, Bob Walkup, Lance D. Weems, and Junqi Yin. 2018. The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems. In *SC18: International Conference for High Performance Computing, Networking, Storage and Analysis*. https://doi.org/10.1109/SC.2018.00055

[80] Sean Wallace, Xu Yang, Venkatram Vishwanath, William E. Allcock, Susan Coghlan, Michael E. Papka, and Zhiling Lan. 2016. A Data Driven Scheduling Approach for Power Management on HPC Systems. In *SC '16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 656–666. https://doi.org/10.1109/SC.2016.55

[81] Torsten Wilde, Axel Auweter, Michael K. Patterson, Hayk Shoukourian, Herbert Huber, Arndt Bode, Detlef Labrenz, and Carlo Cavazzoni. 2014. DWPE, a new data center energy-efficiency metric bridging the gap between infrastructure and workload. In *2014 International Conference on High Performance Computing & Simulation (HPCS)*. 893–901. https://doi.org/10.1109/HPCSim.2014.6903784

[82] Qiang Wu, Qingyuan Deng, Lakshmi Ganesh, Chang-Hong Hsu, Yun Jin, Sanjeev Kumar, Bin Li, Justin Meza, and Yee Jiun Song. 2016. Dynamo: Facebook's data center-wide power management system. *ACM SIGARCH Computer Architecture News* 44, 3 (2016), 469–480. https://doi.org/10.1145/3007787.3001187

[83] Fan Yang and Andrew A. Chien. 2016. ZCCloud: Exploring Wasted Green Power for High-Performance Computing. In *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 1051–1060. https://doi.org/10.1109/IPDPS.2016.96

[84] Kohei Yoshida, Rio Sageyama, Shinobu Miwa, Hayato Yamaki, and Hiroki Honda. 2023. Analyzing Performance and Power-Efficiency Variations among NVIDIA GPUs. In *Proceedings of the 51st International Conference on Parallel Processing* (Bordeaux, France) *(ICPP '22)*. Association for Computing Machinery, New York, NY, USA, Article 65, 12 pages. https://doi.org/10.1145/3545008.3545084

[85] Chaojie Zhang and Andrew A. Chien. 2021. Scheduling Challenges for Variable Capacity Resources. In *Job Scheduling Strategies for Parallel Processing*, Dalibor Klusáček, Walfredo Cirne, and Gonzalo P. Rodrigo (Eds.). Springer International Publishing, Cham, 190–209.

[86] Zhengji Zhao, Ermal Rrapaj, Sridutt Bhalachandra, Brian Austin, Hai Ah Nam, and Nicholas Wright. 2023. Power Analysis of NERSC Production Workloads. In *Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis* (<conf-loc>, <city>Denver</city>, <state>CO</state>, <country>USA</country>, </conf-loc>) *(SC-W '23)*. 1279–1287. https://doi.org/10.1145/3624062.3624200