Data Center Power Oversubscription with a Medium Voltage Power Plane and Priority-Aware Capping

Varun Sakalkar*
Google LLC
sakalv@google.com

Vasileios Kontorinis*
Google LLC
vkontori@google.com

David Landhuis*
Google LLC
landhuis@google.com

Shaohong Li
Google LLC
shaohongli@google.com

Darren De Ronde
Google LLC
dderonde@google.com

Thomas Blooming
Google LLC
tblooming@google.com

Anand Ramesh
Google LLC
aramesh@google.com

James Kennedy
Google LLC
jimkennedy@google.com

Christopher Malone
Google LLC
cmalone@google.com

Jimmy Clidaras
Google LLC
canuckjc@gmail.com

Parthasarathy Ranganathan
Google LLC
partha.ranganathan@google.com

Abstract

As major web and cloud service providers continue to accelerate the demand for new data center capacity worldwide, the importance of power oversubscription as a lever to reduce provisioning costs has never been greater. Building on insights from Google-scale deployments, we design and deploy a new architecture across hardware and software to improve power oversubscription significantly. Our design includes (1) a new *medium voltage power plane* to enable larger power sharing domains (across tens of MW of equipment) and (2) a *scalable, fast, and robust power capping service* coordinating multiple priorities of workload on every node. Over several years of production deployment, our co-design has enabled *power oversubscription of 25% or higher*, saving hundreds of millions of dollars of data center costs, while preserving the desired availability and performance of all workloads.

• **Hardware → Enterprise level and data centers power issues**; • **Computer systems organization → Availability**.

Data center; Power; Electrical Design; Medium Voltage; Power Capping; Energy Efficiency; Availability; System Design

ACM Reference Format:
Varun Sakalkar, Vasileios Kontorinis, David Landhuis, Shaohong Li, Darren De Ronde, Thomas Blooming, Anand Ramesh, James Kennedy, Christopher Malone, Jimmy Clidaras, and Parthasarathy Ranganathan. 2020. Data Center Power Oversubscription with a Medium Voltage Power Plane and Priority-Aware Capping . In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '20), March 16–20, 2020, Lausanne, Switzerland.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3373376.3378533

1 Introduction

Worldwide spending on data center systems now exceeds $200B annually [1], with a sizable fraction of this investment going towards building out the physical infrastructure that provides power, cooling, and space for server machines. For example, considering only "hyperscale" data center operators (including Alibaba, Amazon, Apple, Baidu, Facebook, Google, IBM, JD.com, Microsoft, Oracle, and Tencent), total capital expenditures approached $120B for the year ending in September 2019 [2]. Based on typical ratios of data center to server equipment costs, this translates to hyperscale operators spending tens of billions of dollars each year on physical infrastructure [3]. These trends motivate increased emphasis on efforts to improve the cost efficiencies of building data centers, either by reducing costs of construction or by using built-out capacity more effectively.

One such effort, focused on increasing capacity utilization, is *power oversubscription.* Leveraging the observation that data center power consumption is typically far less than the theoretical maximum power draw of deployed equipment

*First three authors contributed equally to this work.

ASPLOS '20, March 16–20, 2020, Lausanne, Switzerland

ACM ISBN 978-1-4503-7102-5/20/03.
https://doi.org/10.1145/3373376.3378533

(and shows wide variability), oversubscription allows more IT equipment to be deployed on the same infrastructure. Essentially, this creates new data center capacity without additional construction, corresponding to significant cost savings for hyperscale operators. The power oversubscription potential (or *oversubscription ratio [OSR]*) is determined by the size of the power sharing domain. Additionally, in the infrequent cases when the load is at risk of exceeding the power capacity limits, a software technique, *power capping*, reduces power by shutting down workloads. Power capping for low-risk power oversubscription has developed over the last 15 years into an essential enabler of data center cost reduction [4–14].

In this paper, we discuss how we can improve power oversubscription significantly to achieve high data center cost savings. Leveraging insights from studying power consumption trends and workload requirements at Google-scale deployments, we propose a new power oversubscription architecture co-designed across facility electrical design, power control, and cluster management. Specifically, we make the following key contributions:

Medium voltage power plane. First, while it has been recognized from the early days of power capping that the potential oversubscription ratio (OSR) increases as the size of a data center power sharing domain increases, most architectures for power distribution and power capping have focused on sharing power with low voltage (~400 V or less) at the level of racks (tens of KW) and power distribution units (PDUs, 2-3 MW) [15]. In contrast, we propose power sharing across tens of MWs of equipment through a new *medium voltage power plane (MVPP)* (~15 kV).

In addition to a 10X larger pool for statistical multiplexing of workloads compared to the largest PDUs today, our design supports large variations in power density within and across data center rows, supporting regular server racks, but also higher-density accelerator racks (e.g., GPUs or TPUs) or lower-density storage racks. Not only is the oversubscription potential larger, complexity is also significantly reduced: we focus on preventing overloads of the entire power plane while other systems [11, 13, 16–18] must budget and control power at multiple levels of the power hierarchy.

To the best of our knowledge, our work is the first deployment of medium voltage power sharing in hyperscale data centers. In addition to build-out challenges (lack of design, construction, and operational experience in the industry, limited supply chain), our design also addresses key technical challenges around large-scale paralleling of backup generators, ensuring high availability, and the ability to dynamically track the power flow (generator awareness).

Priority-aware power capping. Second, not all workloads have the same availability requirements: *production workloads* are latency-sensitive and should go down rarely, but *non-production* workloads are much less sensitive to performance fluctuations and do not require stringent availability guarantees. In contrast to prior approaches, we design a new power capping service that incorporates generator awareness and information about job priorities, and works in conjunction with the MVPP for increased power oversubscription without compromising overall workload service-level objectives (SLOs). Specifically, the ability to affect only low-priority workloads on every node when power capping, means the achievable OSR is limited only by the aggregate power consumption of the high-priority workload.

Our cluster scheduler is able to assign tasks with different priorities (associated with different SLOs) to any node on the power plane. This maximizes scheduling flexibility and reduces resource stranding, ultimately leading to higher resource utilization. Using the node controllers of the cluster scheduler, our implementation suspends tasks to reduce power, allowing our power capping service to be effective at reducing power on a wide range of machine types (unlike power capping systems dependent on hardware-specific actuators like RAPL [19, 20] and DVFS [21]). In addition, our sharded implementation of the power capping service requires significantly fewer CPU resources than prior approaches that calculate and monitor power budgets at multiple levels of the distribution hierarchy [13].

Furthermore, the design of our power capping service can reliably inform tens of thousands of servers to reduce power consumption in 2-4 seconds. This is the total time required to read power meter measurements over a network, decide whether the value has exceeded the threshold for capping action, notify tens of thousands of machines, and have the majority of those machines reduce their power consumption. To our knowledge, this is the fastest reported software-based power controller at tens of MW scale. It allows data center power to safely peak near electrical design limits without triggering disruptive protections such as breaker trips.

Successful production deployment at scale. Our design is in full production at worldwide scale, and has been successfully deployed on multiple data center campuses, over multiple years. We present fleetwide data that show oversubscription ratios of 25% or more, significantly better than the single-digit oversubscription ratios in previous generations [3]. This translates to hundreds of MWs of data center capacity and a potential to reduce industry data center costs by billions of dollars annually. More importantly, we can achieve these savings without significantly impacting the availability SLOs of our production workloads.

Overall, power capping events become very rare if the power distribution is designed such that backup generator capacity is the only practical electrical limit: threshold violation events happen only in the infrequent cases when peak loads coincide with a utility power outage.

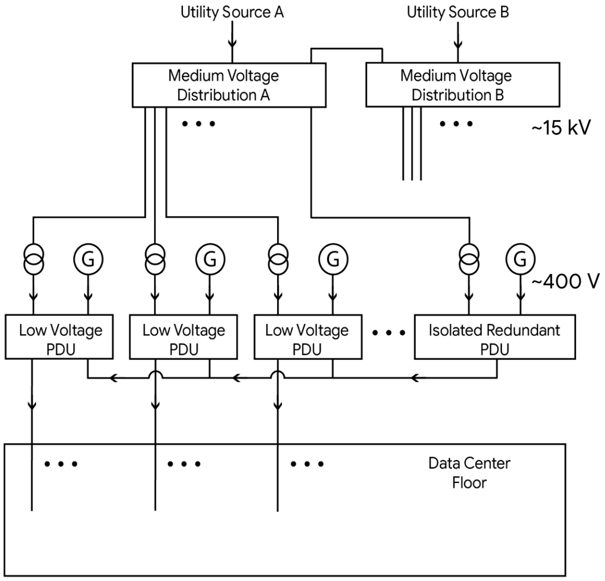


Figure 1. Schematic of a traditional power distribution architecture. *This baseline, traditional radial architecture provides N+1 redundancy at the PDU level. The PDU is the choke point for power distribution and limits the oversubscription possible in this architecture.*

Together, our power plane, power capping, and scheduling systems enable even more aggressive oversubscription that is limited only by the maximum fraction of power associated with the high-priority workloads that cannot be disrupted.

Thus, for instance, for a real-world scenario where high-priority workloads never consume more than 60% of "peak power" on a power plane, an OSR of $1/0.60 - 1 \simeq 66\%$ is possible. In comparison, Google's first-generation power capping architecture achieves single-digit OSRs.

Paper organization. The rest of the paper is organized as follows. Section 2 first provides an overview of Google's prior power oversubscription architecture as an example baseline design, and discusses key terminology for power oversubscription. Section 3 presents our new proposed architecture with detailed discussion of the implementation of the medium voltage power plane and the workload-aware power capping service. Section 4 discusses the deployment of our architecture at scale, including results showing the increased oversubscription benefits and fast response times. Section 5 discusses related work, and Section 6 summarizes our conclusions.

2 Background

Below, we provide an overview of Google's first-generation power oversubscription architecture as an illustration of the traditional baseline for power oversubscription. Section 2.1 describes the power distribution and cluster scheduling in the data center, and Section 2.2 discusses the basics of power oversubscription and its implementation.

2.1 Data center design

Data center power distribution. Figure 1 illustrates the common "radial" power distribution architecture used in large-scale data centers over the last decade. Power is delivered from a utility substation to a data center building at a medium voltage (25-35 kV) and is stepped down at building entry to another medium voltage less than 15 kV. Power fans out radially from each medium voltage bus to multiple power distribution units (PDUs), each supporting 2-3 MW of IT equipment at low voltage across multiple rows of machine racks on the data center floor.

Each PDU has a backup generator associated with it. $N+1$ redundancy is provided at the level of PDUs. This significantly increases power availability without the cost of duplicating every generator and PDU lineup; so long as only one PDU at a time becomes unavailable due to failure or planned maintenance, the single redundant PDU can take its place. The power feeds entering the building have $2N$ redundancy. In the parlance of data center tier classifications [22, 23], this architecture maps roughly onto *Tier III*: while concurrent maintenance is possible, some medium voltage component failures could lead to either an outage or prolonged operation of many PDUs on generators. In practice, a leading cause of downtime in this architecture is generator failure when utility power is unavailable. When a generator fails, automatic switching of loads from the failed PDU to the redundant PDU—powered by its own generator—mitigates the risk of downtime.

Data center cluster scheduler. A typical cluster scheduling system (such as Borg [24]) manages and runs hundreds of thousands of jobs, from many thousands of different applications, across a number of clusters, each with up to thousands of machines. As illustrated in Figure 2, the cluster generally has a centralized master controller, which orchestrates node control agents that run on each machine in a cluster.

It supports high availability with run-time features that minimize fault-recovery time, and scheduling policies that reduce the probability of correlated failures.

The cluster runs a heterogeneous workload with two basic application tiers:

- *Production tier:* long-running, production services that should rarely go down and are assumed to have a 100% availability target in this paper. They also handle short-lived latency-sensitive requests (a few μs to a few *ms* for critical internal software infrastructure).
- *Non-production tier:* jobs that take from a few seconds to a few days to complete; these are much less sensitive

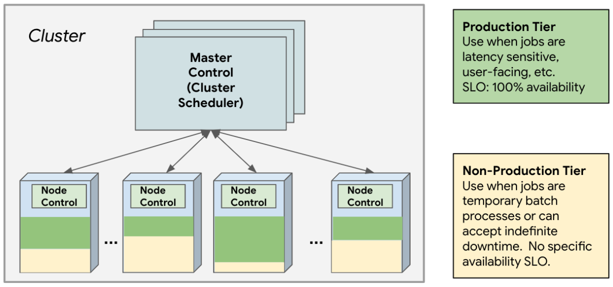

Cluster
Master Control (Cluster Scheduler)
Node Control
Node Control
Node Control
Node Control
...
...
Production Tier
Use when jobs are latency sensitive, user-facing, etc.
SLO: 100% availability
Non-Production Tier
Use when jobs are temporary batch processes or can accept indefinite downtime. No specific availability SLO.

Figure 2. A simplified view of the data center cluster management system. *At a cluster level, scheduled tasks are relayed from a distributed master controller to node controllers running on individual machines. At a high level, tasks can be considered to fall into two different prioritization tiers: production and non- production.*

to short-term performance fluctuations and typically do not require specific availability guarantees.

The mix of these two workload classes varies across clusters, and also varies over time: batch jobs come and go, and many end-user-facing service jobs exhibit diurnal usage patterns.

2.2 Power oversubscription

Power oversubscription is a technique proposed to improve data center efficiency [4, 5] that is widely deployed commercially (e.g., [4, 5, 11, 13]). The key intuition behind power oversubscription is that the data center capacity is often underutilized: actual power draw is much lower than rated power. Spiky and variable workloads on servers leads to median (or even 95th percentile) power consumption that is significantly lower than the peak power. Across ensembles of servers (rows or clusters), spikes are usually not correlated, and this statistical multiplexing leads to even greater underutilization or "stranding" of power. Similar to the airline industry overselling flight capacity, power oversubscription allows more IT equipment to be provisioned on the same power bus, essentially providing new data center capacity without costly construction, therefore reducing costs (dollar-per-watt). This fraction of additional power is called the *oversubscription ratio (OSR)*.

Power oversubscription "choke points". The power oversubscription potential in a data center is determined by the maximum power expected at a *choke point*, i.e., the point where an electrical limit will be reached first as load increases (e.g., a circuit breaker trips). Typically, the larger the power capacity of the choke point, the more statistical multiplexing will occur downstream for a given set of workloads and equipment. More statistical multiplexing means a reduction in the peak power at the choke point and a larger oversubscription potential.

In the traditional baseline discussed in Section 2.1, the capacity of the commodity backup generators associated with the PDUs (measured in megawatts [MW]) represent the choke point of the architecture. The other components of the PDU – the switch, transformer, breaker, and bus – all typically have larger capacity than the generator. Similarly, the capacity of the upstream medium voltage distribution is typically much larger than the combined capacity of the PDUs, and the total capacity of the low voltage buses bringing power from each PDU to the IT equipment is much larger than the PDU capacity. Consequently, in the baseline architecture, power oversubscription means deploying IT equipment per PDU that exceeds its generator capacity. This is a relatively small level of statistical multiplexing compared to the broader datacenter.

Power oversubscription "safety valves". While the oversubscription ratio is typically set to a value to avoid the risk of exceeding infrastructure limits, the statistical nature of power variation still requires a "safety valve" to deal with overloaded circuits. The default hardware response, *load preservation* (discussed further in Section 3.3), where the generator controls disconnect some electrical loads by opening bus breakers, will work but can be disruptive. *Power capping* is an alternate software technique that rapidly reduces power when it approaches the electrical limit of a choke point. At a high level, power capping monitors the total power within the domain associated with a choke point. When a specified power threshold is exceeded, power capping notifies the node controller on every server (Figure 2) in the domain. The node controllers react by suspending or killing jobs to reduce power (response time in order of seconds).

This reactive suspension of workloads is both effective and tolerable, so long as the events are rare enough to preserve the applicable SLO for availability. From the point of view of the broader scheduler, power capping looks just like another failure mode. Figure 3 illustrates the holistic treatment of power capping unavailability in conjunction with other failure scenarios in a typical cluster. The data center is managed as a hierarchy of different failure domains such that a failure in one domain can impact some or all of the domains below it. The cluster scheduler captures the key constituents of this hierarchy of failure domains and their dependencies.

3 New power oversubscription architecture

Below, we discuss our new power oversubscription architecture. Section 3.1 discusses the two key insights that underpin our design. Building on these, Section 3.2 discusses our new power delivery architecture based on an MVPP that enables larger power capping domains, and Section 3.3 discusses our co-designed power capping service that incorporates generator awareness and information about job priorities to

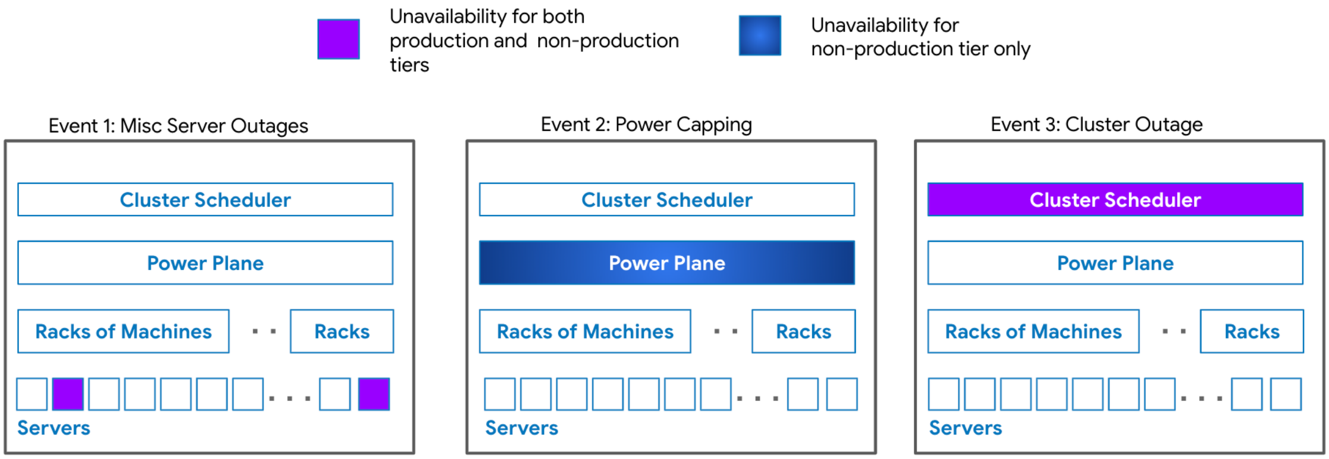

Unavailability for both production and non-production tiers
Unavailability for non-production tier only
Event 1: Misc Server Outages
Event 2: Power Capping
Event 3: Cluster Outage
Cluster Scheduler
Power Plane
Racks of Machines
Racks
Servers

Figure 3. Holistic treatment of power capping along with other failure scenarios. *When a small fraction of the server machines stop responding (Event 1), and all the resources (e.g. CPU, memory, disk) associated with these machines become unavailable. When a power capping event (Event 2) occurs, some tasks across that power domain are suspended; other jobs and servers are not impacted. Finally, an outage of the cluster scheduler prevents scheduling of new jobs (Event 3), effectively taking away resources from workloads.*

increase OSR without compromising workload service-level objectives (SLOs).

3.1 Two key opportunities

In the most conservative implementation of oversubscription, the largest OSR possible without a risk of exceeding infrastructure limits is given by:

$$OSR = \frac{1}{U_{Total}} - 1 \tag{1}$$

where U_{Total} is the maximum total power utilization across both production and non-production workloads. We define power utilization as the ratio of actual power consumption to the theoretical peak power due to all deployed equipment (see Section 4.1). Depending on the deployment's tolerance to power budget violation events, other less conservative implementations can replace U_{Total} with the 95th or 99th percentile of total power utilization. However, there are two other insights that offer opportunities to significantly increase oversubscription.

Larger power domains for increased oversubscription. First, as discussed in Section 2.1, the choke point in the baseline architecture is the generator capacity at the PDU. This is a relatively small level of statistical multiplexing compared to the broader data center, and the power oversubscription potential is limited by the diversity of workloads running within each PDU domain. Increasing the power capacity of the choke point, for example, to a broader cluster level, will correspondingly increase the statistical multiplexing and the OSR opportunity. Thus, OSR can potentially be increased to:

$$OSR_{new} = \frac{1}{U_{Cluster}} - 1 \tag{2}$$

where $U_{Cluster}$ is the power utilization associated with a broader scheduling domain, and typically smaller than U_{PDU}.

Workload awareness for increased oversubscription. Second, as discussed in Section 2.1, not all workloads have the same availability requirements. It is therefore possible to increase OSR aggressively by taking advantage of the weak availability requirements for *non-production* workloads. High total power utilization is tolerable as long as the residual power associated with production workloads never exceeds electrical limits. Thus, the greatest OSR possible with power capping can potentially be set to:

$$OSR_{new} = \frac{1}{U_{Prod}} - 1 \tag{3}$$

where U_{Prod} is the power utilization associated with production workloads. If non-production workloads are substantial, U_{Prod} will be significantly smaller than U_{Total}.

3.2 Medium voltage power plane

Larger power domains and higher availability. Figure 4 illustrates our new power distribution architecture. Two significant differences from the previous generation in Figure 1 are the medium voltage distribution and the generator pooling. The capacity of utility feeds is significantly larger than the size of the generator farm, essentially moving the choke point to the generator farm. Correspondingly, in combination with the power capping service (discussed in Section 3.3), very large power oversubscription is possible, significantly

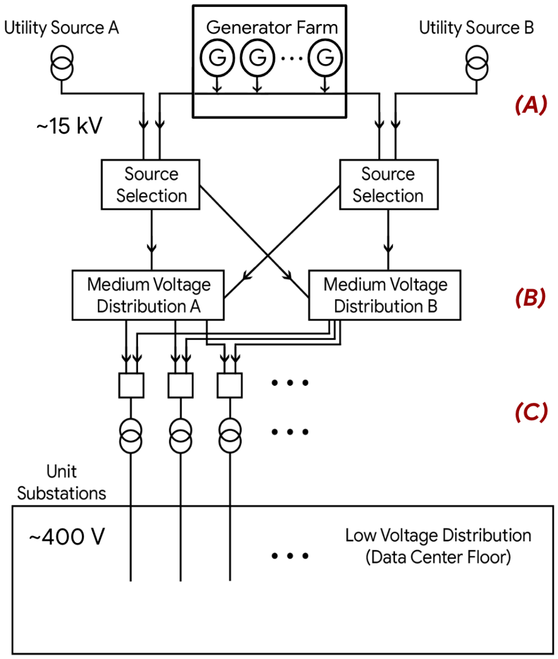


Figure 4. Schematic of the medium voltage power plane (MVPP) architecture. *The design features 2N redundancy upstream of the unit substations, apart from the generator farm. The generator farm has $N + M$ redundancy. When at least one utility source is available, the choke point is the utility source transformer (whose limit will typically not be reached with oversubscription). When both utility sources are unavailable, the choke point becomes the generator farm.*

improving cost per watt — even after considering the relatively minor cost of oversizing other electrical components to make the generator farm the choke point.

The new topology of this design also achieves overall higher power availability than the baseline design for the same cost. Single failures and maintenance operations related to the power distribution will not cause any significant downtime. All upstream points (see labels *A* and *B* in Figure 4) now have alternate pathways for every distribution component, essentially a 2N scheme. For example, if the bus for Medium Voltage Distribution A goes out of service, every unit substation can be powered via Medium Voltage Distribution B. Similarly, during outages, this design also provides higher power availability while minimizing the costs of generators. This is achieved by paralleling the output of $N + M$ generators on a single bus where M is much smaller than N and can be tuned to meet any desired availability target. (This paralleling needs additional hardware support as discussed below.) Note that close to the data center floor, unit substations power rows of IT equipment.

Given the high reliability of these components, an $N + 1$ redundancy scheme is used to switch from these to an alternative source (not shown in the Figure) in case of failure or maintenance.

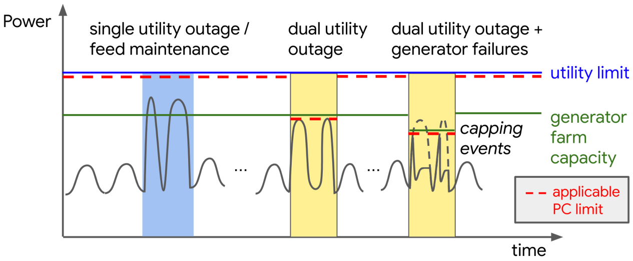


Figure 5. Examples of generator awareness by the MVPP power capping (PC) system during three episodes. *(left) Single utility loss where the capping threshold remains near the utility limit: no capping occurs; (middle) Dual utility outage where the capping threshold is near the generator farm capacity: no capping occurs; (right) Dual utility outage with some generator failures: the capping threshold is near the capacity of the remaining generators, and capping events occur.*

Support for generator awareness. Our architecture supports the ability to dynamically track the power flow and enforce the appropriate power limit every time the active path changes. Although generator awareness is also possible in the baseline design described in Section 2, the significant relative difference (~33% of the generator farm capacity) between the utility feed capacity and the generator farm capacity of the MVPP makes generator awareness more profound. In practice, capping events are expected only when loads are powered by generators due to failure of both utility feeds. Given dual utility feed outages are rare, the expected unavailability due to power capping is correspondingly small.

Figure 5 describes three example outage scenarios and how the capping limits change accordingly. In the first scenario, one of the two utility feeds to the MVPP becomes unavailable due to either a failure or planned maintenance. Since the remaining feed can support the MVPP on its own, the power capping system continues to protect the limit applicable for utility power. In the second scenario, both utility sources become unavailable, e.g. due to a complete failure of local utility power. The power capping system is aware that the MVPP is receiving power from the generator farm, and it enforces a limit associated with generator farm capacity until utility power is restored. Finally, in the third scenario, there is not only a complete loss of utility power but also several generators have failed to start. In this case, the power capping system is aware of how many generators are available and enforces an appropriately lower power limit.

Section 3.3 discusses how generator awareness is implemented in our power capping service, and Section 4 presents data showing the effectiveness of this approach for handling rare situations where utility power is unavailable and power consumption exceeds generator farm capacity.

Power fungibility and reduced stranding. In addition to enabling higher power oversubscription, the MVPP architecture features progressively larger power headrooms going from the building level to the level of individual rows. (Headroom, in this context, is the difference between the capacity of the power distribution infrastructure and maximum expected power.) This architecture makes *power fungible across a data center floor*, with total capacity up to several tens of MW. Consequently, the MVPP enables deployment flexibility for different kinds of IT equipment (servers, accelerators, storage, etc.) while avoiding power stranding. With a more conventional baseline power distribution architecture, it can be difficult to avoid stranding power capacity, especially when there are large regions of low-power-density (e.g. storage) equipment on the data center floor. The ability to support a wide range of power densities at any location on the data center floor has increased in importance with the advent of high-power-density hardware accelerators (e.g. TPUs or GPUs for machine learning workloads). Another aspect of co-design for a high availability cluster is the use of high-bandwidth network fabrics which can also scale to several tens of MW. This implies that both intra-cluster and inter-cluster bandwidth can be uniformly large across a power plane. Coupled with the additional flexibility to consume power anywhere on the data center floor, there is more freedom to spatially organize IT equipment. In particular, data storage can be physically separated from workload-intensive machines, and this enables programming models and cluster architectures which take advantage of close physical proximity between compute nodes.

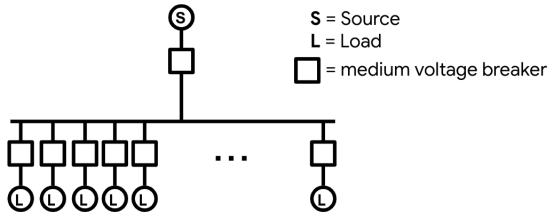


Figure 6. Reliable, fast-acting fault protection on the medium voltage distribution bus. *Medium voltage class relays support high-speed communication between devices in case of a fault, resulting in fast protection without complex wiring, and enabling relatively large distances (> 100 m) between breakers.*

Generator and power bus features. We developed several hardware features to support the operation of the medium voltage power plane. The generator paralleling discussed earlier requires carefully managing constraints on waveform, phase sequence, frequency, phase angle difference, and voltage differences across multiple sources [25]. For a medium voltage plane at the scale of tens of MW, the time needed to resolve these constraints and synchronize generators is longer than the practical battery backup time for IT equipment. We worked with generator manufacturers to optimize the startup time of generator farms and confirm reliable synchronization behavior with large-scale tests. Additionally, to implement the generator awareness discussed earlier, we worked with generator vendors to create an API that reports how many generators are available (used by the power capping service to adjust its thresholds accordingly).

Finally, because the MVPP involves distribution buses that are large both in capacity and physical extent, there is a challenge to provide reliable, fast-acting hardware protection in case of an electrical fault. In a system that distributes tens of MW, fast protection is necessary to minimize physical damage and maximize safety for people in the vicinity of a fault. While low voltage electrical equipment often lacks reliability and advanced protection features, both are readily available for medium voltage equipment. In particular, Figure 6 shows medium voltage devices that can communicate with each other via a peer-to-peer protocol. This not only eliminates wiring complexity but also enables near-instantaneous circuit breaker responses to isolate a fault. Each individual load in the figure represents a unit substation powering a row of IT equipment on the data center floor.

Software architecture. Figure 7 describes the high-level architecture of the power capping service that reduces power on the MVPP whenever power approaches an applicable power limit. At a frequency of 1 Hz, a *meter watcher* module continuously polls several fast-response power meters located at the utility feeds and generator farm (corresponding to the choke points in the power distribution from Section 3.2). These measurements are fed to the *power notifier* module that is tasked with aggregating readings and comparing measured power values with PC thresholds derived from MVPP electrical design data.

Whenever the power capping threshold is exceeded, the notifier contacts the *machine manager* module to send remote procedure calls (RPCs) to individual node-level schedulers on the power plane. These in turn suspend low-priority tasks using a SIGSTOP signal. Each node scheduler will prevent low priority tasks from running as long as power remains above threshold. Once power drops sufficiently, the node scheduler will issue a SIGCONT signal, allowing tasks to resume. Typically a significant portion of tasks (about a third) die and reschedule once they receive the SIGCONT. These tasks commonly fail to respond to application level watchdogs. SIGSTOP was chosen over a SIGKILL signal, to allow a portion of tasks to survive power capping which reduces wasted work. In addition, SIGSTOP minimizes the amount of work required at the node level at the very moment of load shedding (no process cleanup happens with SIGSTOP). However, there are two special cases where we prefer SIGKILL

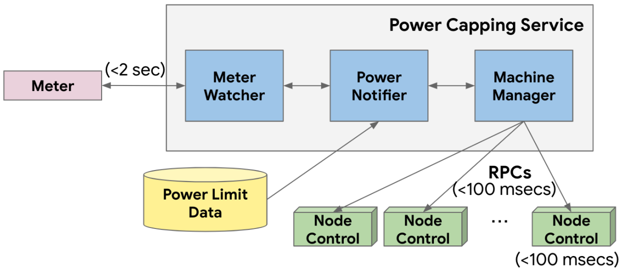


Figure 7. Simplified block diagram for the power capping service. *A watcher module polls fast-response power meters at choke points. The notifier module checks aggregated power with thresholds, and the machine manager sends RPCs to individual nodes suspending non-production tasks.*

over SIGSTOP signal. We have found that accelerators may continue crunching data and consume power several seconds after we suspend the host task. Sending a SIGKILL signal instead flushes the data queues between host and accelerator, idles the accelerator, and reduces power faster. We also choose to use a SIGKILL signal for cloud virtual machines. SIGSTOP and SIGCONT signals are used for virtual machine management and there were concerns about being able to distinguish the different use cases.

For the traditional baseline design, we have historically deployed one master-elected power capping job instance per PDU. For the much larger MVPP domain, we shard the power capping service such that each instance is responsible for notifying roughly 2 MW worth of node schedulers. The limited fan-out of this design helps achieve fast response times and also bounds the resources needed per job. We provision roughly 1 CPU and 1 GB of memory per power capping instance which is orders of magnitude less overhead than other power capping implementations in the literature [13].

3.3 Power Capping Service

Fast response times. The end-to-end response time for power reduction is bounded to < 4 s, with an average case of < 2 s. Limiting the notification fan-out with the sharded design bounds the notification time to be a small portion of the overall response time budget. Given that reaching all the machines in the domain takes less than 100 ms and suspending tasks in the node also happens in less than 100 ms, most of the time is spent waiting on meter telemetry.

Although our industry-standard power meters are capable of measuring power at sub-second granularity, they report power values over the network reliably only once per second. As a result, we expect in the worst case a one second delay in reporting from the meter. Given that the power capping service also polls the meters once per second, we conservatively budget for a total of 2 s round trip latency for reading power meters.

The notification from the power capping jobs to the individual machine level controllers takes milliseconds. To obtain this optimized delay, we pre-allocate the memory for RPC stubs and maintain warm TCP connections from the power capping controller to all the node-level controllers in the domain. The latency of the node-level controller from the arrival of the remote procedure call until the power drops is only tens of milliseconds.

On average, the expected response time is well below 2 s. However, we budget overall 4 s to account for network tail latencies. The fast response times allow the capping threshold to be just slightly lower than the power limit we protect (see "Robustness" below), with little risk of power exceeding the limit before power capping reacts. Note that we can tolerate machines that are slow to, or do not, respond. Since we suspend all low-priority jobs and typically shed much more power than we need, these stragglers do not matter.

Generator awareness of power limit. As discussed in Section 3.2, the MVPP architecture supports generator awareness. The power capping service reads power meters deployed at the MVPP level (*A* in Figure 4) and determines the active power source. The protected power limit is the utility feed capacity whenever utility power is the active source, and it is the available generator farm capacity if the generator farm is the active source. If some generators fail to start or synchronize to their common bus, the power limit will be dynamically adjusted to match the capacity of the successfully synchronized generators. The power capping service uses the hardware API discussed earlier to determine generator availability and get the capacity of the responsive generators.

Generator load demand response. For greater efficiency and reliability, it is desirable to run a smaller number of generators near full capacity than all generators at low capacity. *Load demand response* dynamically adjusts the number of running generators to better match the actual load. The challenge is this: a power capping event could cause generators to idle, which in turn reduces the threshold to activate power capping; however, after we have already shed non-production workload, subsequent power capping may be ineffective, and we could potentially lose the entire power plane due to generator overload. To avoid this situation, we implemented a load demand algorithm which maintains a conservative buffer of generator farm capacity above demand.

Robustness. Our MVPP power capping implementation emphasizes robustness. First, we use accurate and dual - redundant meters at the electrical distribution choke points. This is a simpler and safer strategy than trying to infer power

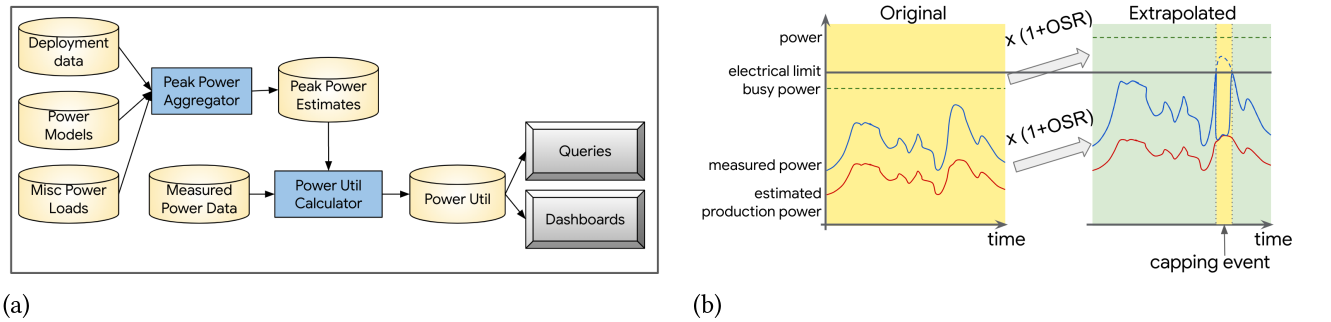


Figure 8. (a) Automated analysis pipeline to capture historical power utilization values and estimate oversubscription potential. *Measured historical power meter data is divided by modeled total peak power to generate power utilization series.* **(b) Example simulation of power capping events at a given OSR.** *Historical power data can be scaled to simulate power capping events at hypothetical OSRs, and impact of production workloads.*

from many machine-level measurements that are not aligned in time.

This strategy also minimizes the compute and network resources required to monitor choke points. In the event of a dual utility outage, no meter readings are available for a period of tens of seconds while generators are starting and preparing to accept load. During this time, the IT equipment draws power from local batteries. Since dual utility outages are rare, our power capping service can afford to proactively shed low-priority load while generators are starting to ensure that generators are not overloaded.

We trigger capping at 98% of the power limit that we seek to protect. The 2% buffer accounts for possible increases in power consumption within our typical response time (see "Fast response times" above). In contrast to the approach followed in [11], we do not implement any hysteresis to avoid power oscillations. Instead, we allow tasks to resume gradually over time. Given that a significant portion of low priority tasks are aborted before they can resume, rapid power oscillations are unlikely.

Fail-safe load preservation. *Load preservation* is a generator control mechanism used as a fail-safe backup for non-production load shedding. When power capping fails to reduce power sufficiently and the generator farm becomes overloaded, generator controls rapidly disconnect some electrical loads in order to preserve the remaining loads. This can be done by opening row-level bus breakers, for example. In our implementation, individual buses (data center rows) are assigned a priority ranking, and the rows are disconnected in an overload emergency according to their ranking. It is highly desirable to activate power capping before load preservation kicks in, because the former affects only non-production workload while the latter disrupts all workloads. We achieve this in practice by activating power capping based on a power threshold and load preservation based on generator frequency falling below an “underfrequency” limit. The thresholds are chosen such that power capping will always act before load preservation.

4 Deployment at scale in Google

In this section, we discuss the deployment of our new power oversubscription architecture at Google scale. We first describe our software pipeline to monitor power and determine OSR thresholds (Section 4.1). We then present results demonstrating the benefits of the larger power domain and awareness of SLOs fleetwide as well as in individual clusters (Sections 4.2 and 4.3).

4.1 Continuous Power Estimation

We have built a software pipeline to facilitate continuous power utilization analysis in order to determine safe levels of oversubscription for different workload mixes. We focus on deriving two primitives through this pipeline: (a) total power utilization and (b) production power utilization at the choke point of the power architecture, which is the PDU or MVPP. Total power utilization can be translated to a rate of power capping events at a given OSR (see Equation 1), while production power utilization tells us whether sufficient load can be shed at a specific OSR (see Equation 3).

Estimation of deployed power. At one end of the pipeline (Figure 8), estimates of *peak power*, the worst-case consumed power of connected equipment, are calculated for each choke point. Peak power is the denominator for both total power utilization and production power utilization. We derive peak power from:

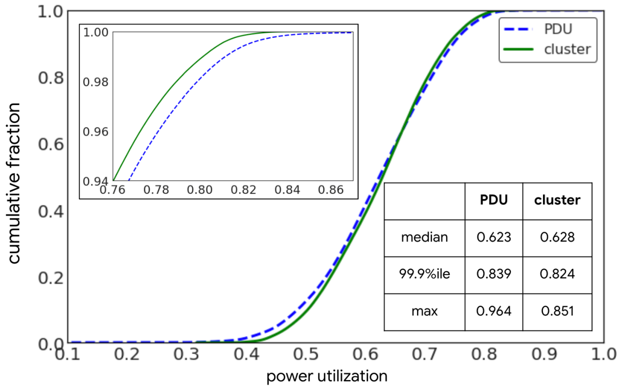
- Records of deployed hardware configurations
- Machine peak power values. Worst case machine power is determined experimentally during the qualification process for the new hardware platforms by running Google benchmarks and stress tests.
- Peak power values for non-machine loads on the data center floor (e.g. cooling and networking equipment)

Measurement of consumed power. On another branch of the pipeline, a state-of-the-art monitoring infrastructure records *actual consumed power* every ~2 seconds at different levels of the power distribution hierarchy, including levels labeled *A* and *C* in Figure 4. Both the raw power measurements and the summarized time series data are held in a distributed data store to accelerate data retrieval.

To get total power utilization, we normalize measured consumed power with the estimated deployed peak power (see Figure 8a). Using Equation 1, the power utilization time series can be analyzed to determine the maximum possible OSR for an MVPP or PDU before power capping events are expected. We can also use the time series to estimate the total duration of capping events as a function of OSR, which yields an approximate probability for power capping without "generator awareness" (see Section 3.2).

Modeling of generator failures for refined probability estimates. We assess the impact of generator awareness through a Monte Carlo simulation which estimates the capping rates and the resultant impact to non-production tier availability based on utility downtime and frequency. The System Average Interruption Frequency Index (SAIFI) and System Average Interruption Duration Index (SAIDI) are good approximations for average utility outage duration and frequency in the US. This data can be obtained from the US Energy Information Administration [26]. Because of generator awareness, the probability of capping is dramatically reduced.

Estimation of production power. We define production power as the power consumption attributable to the production workload tier. This is the power consumption we expect to remain after power capping eliminates the non-production workload. Production power cannot be directly measured. Instead we reduce machine-level CPU utilization by the known utilization of non-production tasks and estimate each machine's production power from machine-level power models that interpolate idle and peak machine power according to CPU utilization. This approach builds on the methodology described in [4, 5]. To get production power utilization we normalize production power again with estimated deployed power. Using Equation 2, the production power utilization time series can be analyzed to determine the maximum possible OSR for an MVPP or PDU before the production power would exceed the critical limit that the capping mechanism protects. For the PDU-based designs, the limit is the generator capacity, while for the MVPP-based designs, the limit and baseline for oversubscription is the non-redundant generator farm capacity.



	PDU	cluster
median	0.623	0.628
99.9%ile	0.839	0.824
max	0.964	0.851

Figure 9. A comparison of the cumulative distribution functions (CDFs) of power utilization for clusters and PDUs. *Each CDF represents an average over many single-cluster workload mixes. The max power utilization observed at the cluster level is 85%, corresponding to an OSR potential of ~17%.*

4.2 Benefits at scale

To characterize the benefits from our approach at scale, we present data collected over a period of 1.5 years, from January 2018 to June 2019, for a statistically significant set of clusters in PDU-based data centers as well as clusters in two MVPP data centers. The empirical power utilization distributions reported below are based on 1-hour maximum values from the estimated time series. We also filter out PDUs and MVPPs with peak power less than 80% of electrical limits, since we have observed that clusters at the beginning of their life cycle demonstrate lower total power utilization and higher production power utilization, and they are not representative of fully-deployed clusters where power capping matters. (This happens because it takes more time for batch workloads to be migrated by users into new clusters as compared to production workloads.) We are making a sample of these power utilization traces available to the research community [27].

Larger power domains. We start by examining the relationship between workload domain sizes (PDUs and clusters) and power utilization distributions in aggregate. Note that a cluster domain in our fleet consists of at least a small group of PDUs, and an MVPP includes more than 10X the number machines as a single PDU. On a fleetwide average basis, the power utilization of a cluster has a smaller variance than that of a constituent PDU. Although the average power utilization does not change when the size of the power domain increases, the tail behavior does. At high percentiles, the difference between PDU- and cluster-level utilization increases, roughly 2% at 99.9%-ile and 10% at the maximum. (Figure 9). The additional statistical multiplexing at larger power domains leads to an increase in OSR potential without power

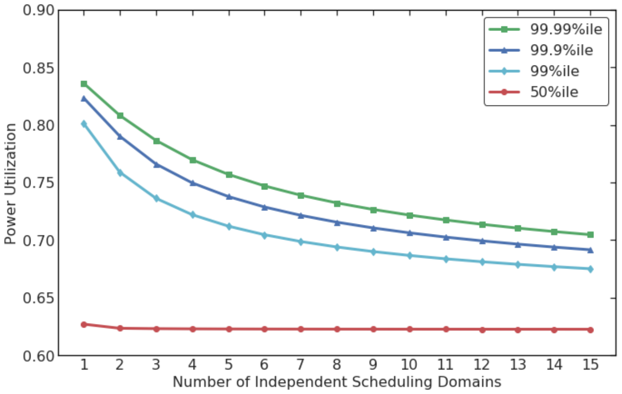


Figure 10. Total power utilization distributions as a function of k independent scheduling domains, $1 \leq k \leq 15$. *The distributions were calculated by Monte Carlo aggregation of many combinations of empirical, cluster-level power utilization distributions, taken k clusters at a time. This illustrates the significant narrowing of the power utilization distribution, corresponding to an increase in OSR potential, as more independent workloads are combined on a power plane.*

capping (Equation 1). Based on the maximum cluster power utilization observed in the fleet (85%), increasing the domain size (from PDUs to MVPP) should enable an OSR of at least 17%.

Multiple scheduling domains. An orthogonal optimization that further increases OSR is to place multiple scheduling domains on an MVPP. We have previously illustrated the 1:1 mapping of a single cluster scheduler to a power plane in Figure 3. However, we can also choose to deploy multiple independent cluster schedulers on the same power plane. The cluster scheduling domain in that case covers a unique set of machines against which jobs can be co-scheduled within that domain. Typically, the smaller the scale of a scheduling domain (where scale refers to the set of machines the scheduler governs), the higher the likelihood of stranded resources for the machines. Incidentally, the large size of MVPP allows us the ability to carve out multiple scheduling domains where each can scale to minimize stranded resources while also taking advantage of statistical multiplexing of power at the MVPP to achieve higher OSR. In Figure 10, we present the effect of increasing the number (k) of independent scheduling domains on the total power utilization of the power plane. A power utilization distribution is generated for each value of k by convolving empirical power utilization distributions from many random combinations of k clusters. The standard deviation of the power utilization distribution is proportional to $1/\sqrt{k}$, as expected from the central limit theorem. As a result, for a desired long-term average non-production tier availability, indicated by the power utilization percentile, we can estimate the expected increase in OSR potential. For example, going from one scheduling domain to two enables roughly 5% additional OSR when we cap at the 99%-ile power utilization.

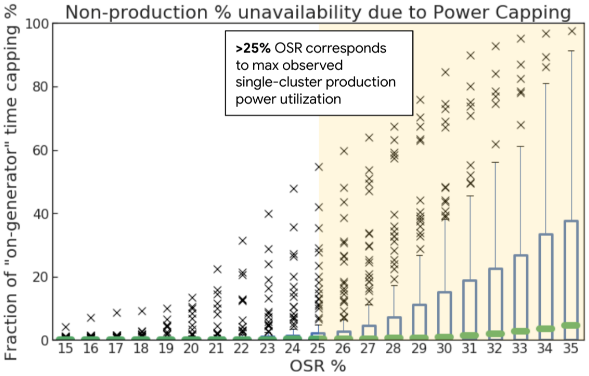


Figure 11. Expected fraction of time that the non-production tier will be capped while on generator. *Markers correspond to individual clusters fully deployed on a power plane to the indicated power oversubscription (OSR). The boxes indicate first and third quartiles of the distribution of clusters.*

Generator and SLO awareness. To further increase OSR, we combine the impact of the larger MVPP power domain with generator awareness. In accordance with our cluster scheduling availability requirements, only non-production workloads are prevented from running during a power capping event. In order to characterize the impact on non - production tier unavailability as a function of OSR, we ran a time-domain Monte Carlo simulation for a single-cluster MVPP at different OSR values using the historical power utilization time series of actual clusters. The simulation employed conservative assumptions about maintenance and failure rates as well as the utility power outage frequency and duration mentioned in Section 4.1. Over centuries of simulated time, we probabilistically generated all three types of events depicted in Figure 5. From the results, we estimated the fraction of generator time where non-production workloads cannot run due to power capping in a given cluster (see Figure 11). The plot shows the range of non-production workload unavailability across clusters as a function of OSR value *while running on generators*. Note that the time spent running on generators is small to begin with, since utility outages occur for at most a few hours per year in typical data center locations. Generator awareness ensures that the impact of power capping on even the non-production workloads on an MVPP is rare and limited in duration.

While most clusters would see impact to non-production workload for at most a few percent of the generator run time at OSR = 30%, we have conservatively limited the initial OSR of MVPPs to 25% based on the largest production power utilization (~ 80%) observed in any single cluster domain to date. This choice minimizes the risk of not being able to

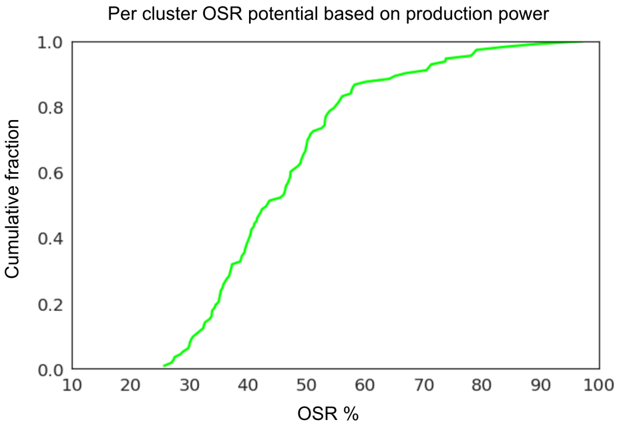


Figure 12. CDF of potential OSR values derived per cluster in our fleet according to the worst (99.99%-ile) production power utilization observed. *These results indicate that we can safely oversubscribe any existing workload mix by 25% while ensuring that power will be lower than the actual limit we are trying to protect: the generator farm capacity of the MVPP. OSR values are related to production power utilization by Equation 3. For example, the smallest 99.99%-ile production power utilization observed for a cluster is 50.7%, correponding to the largest potential OSR:* $1/50.7\% - 1 = 97\%$.

sufficiently reduce power in a utility outage. If in the future, we observe a production power utilization larger than 80% in any cluster domain, we can decide to reduce the OSR of any single-cluster MVPP based on a risk assessment that comprehends the production power utilization trend of that MVPP.

The distribution of worst production power observed per cluster in our fleet is depicted in Figure 12. Unless higher levels of production power utilization are achieved in the future, we can be confident of avoiding the situation where power capping cannot sufficiently reduce power in a utility outage. As more confidence is gained both in our power utilization predictions and in our power capping system, we can consider even more aggressive oversubscription, approaching the limit implied by the maximum observed production power utilization in the affected clusters.

4.3 Specific cluster measurements

We also present example data from production MVPP data centers. These illustrate the impact of aggregation across scheduling domains on power utilization. In addition, we measured the time response of the MVPP power capping service.

Power utilization of single and dual-scheduling domains. Figure 13 shows empirical results based on the partial deployment of MVPPs involving one and two scheduling domains, respectively. In this case, MVPP1 supported a single cluster and was deployed to a peak power of 20 MW during the analysis period. MVPP2 provided power for two independent clusters and was deployed to a total peak power of 22 MW. Standard deviations of the power utilization for MVPP1 and MVPP2 are 0.050 and 0.048, respectively.

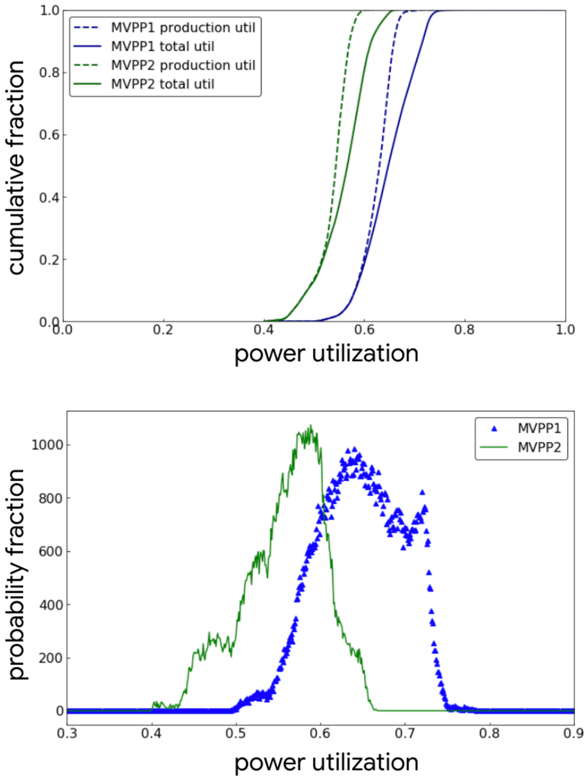


Figure 13. Power utilization of single and dual scheduling domains. *Top: CDFs of total power utilization ("total util") and the production power utilization ("production util") of two different MVPPs over one year. Bottom: Probability distribution functions (PDFs) corresponding to the total power utilization CDFs.*

We have observed that the medians and the form of the distributions are determined by the workload mixes on the respective power planes. Based on the empirical data the highest observed production power utilization for MVPP1 is 65% and MVPP2 is 59%, which is consistent with the expectation that power planes with more scheduling domains will generally have higher OSR potential (see Figure 13). Also, for a single scheduling domain like MVPP1, these results confirm the ability to safely oversubscribe by at least 25% and potentially by as much as 66% for this particular workload.

Response time measurements. A power versus time trace for a sample MVPP power capping event is shown in Figure 14. As described in Section 3.3, the power capping service

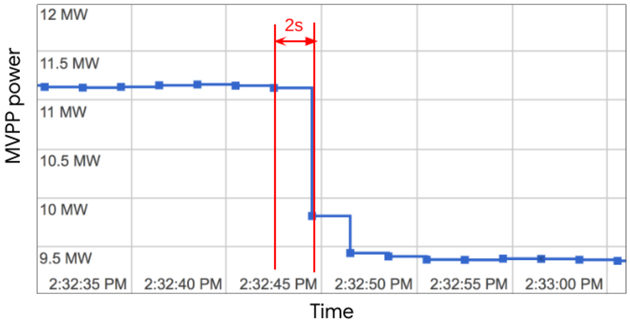


Figure 14. Measured power on a partially-deployed production MVPP during a power capping test.*The power capping event was induced by temporarily lowering the power capping threshold. In about 2s, the MVPP power drops by 15%, reflecting the relative proportion of non-production workload.*

shards the power plane into ∼ 2 MW chunks, and a service instance is responsible for notifying node schedulers within a single chunk. This notification fan-out, combined with fast meter telemetry and our design choice to suspend all non-production workload simultaneously, allows power capping to achieve the largest possible power reduction in a few seconds.

5 Related Work

To the best of our knowledge, our work is the first to discuss improved data center power oversubscription through a co-designed architecture using a medium voltage power plane and a power capping service that is aware of both generators and workload SLOs.

Like our power plane co-design, Facebook's Dynamo [11] is a production hyperscale power management system. More recently, IBM proposed and evaluated CapMaestro [13] as a solution for public clouds that comprehends workload priorities. Our work differs from these in several ways. Instead of measuring and budgeting power at multiple levels of the power hierarchy, we focus on capping a novel medium voltage power plane. Also, regardless of OSR, our power capping events do not affect the performance of high-priority workloads. Our system is able to schedule tasks of any priority on any machine; both Dynamo and CapMaestro depend on segregation of different priorities onto different machines. Together, these differences allow our architecture to achieve significantly higher levels of oversubscription (more than 25%) compared to previous generations. Also, our power capping service starts and stops tasks by sending platform-agnostic signals to the node controllers of our cluster scheduler. Dynamo uses RAPL, and CapMaestro uses the Intel Node Manager, both of which are available only for Intel platforms.

Other researchers have investigated a variety of node-level and multi-level control systems to throttle power and enable oversubscription while minimizing performance impact, e.g. [4, 6, 8, 16, 28–30]. Summarizing many of the available options, [31] discussed interactions between controls at different levels (server, rack, group of racks) with different objectives (energy efficiency, capping, performance) and actuators (p-states, turning machines off, admission control, power budget management). We also note [32] as an example of recent power management work for supercomputers.

Although some medium voltage components are used for data center power distribution [33, 34], we believe we are the first to tackle the fault protection, backup generation, and other operational challenges associated with using a medium voltage power plane in a hyperscale data center context. A separate body of work discusses tapping into stored energy to enable higher oversubscription [35–37]. We could add battery capacity to a medium voltage power plane to enhance its OSR potential, but it is unclear whether the cost savings achieved for other infrastructure would exceed the battery costs. Similarly Fu *et al.* [38] discuss how control algorithms can be used to leverage the current-time curves of circuit breakers for additional oversubscription. Since the variance in the time response of generators and breakers is difficult to know, we have taken the more conservative approach of capping power as quickly as possible whenever a threshold is reached.

Finally, we note there have been substantial contributions on data center resource management at scale (e.g. [39, 40]) that should be considered complementary to the hyperscale power management work summarized here.

6 Summary

With the rapid growth of cloud computing, demand for physical data center infrastructure continues to rise. Furthermore, new types of data center hardware (like machine learning accelerators) demand more power in the same space. To meet these demands at sustainable cost, large-scale service providers must find ways to improve infrastructure utilization while maintaining high availability for the most critical workloads. Power oversubscription will continue to be one of the most important ways. This paper describes a co-designed power delivery, power control, and cluster scheduling solution that enables OSR values of more than 25%, a significant increase relative to what has been previously reported for hyperscale data centers.

We introduced a novel medium voltage power plane architecture for power distribution and discussed how it achieves higher availability than the traditional baseline architecture at comparable cost. Our design pools tens of MW under a choke point, enhancing OSR potential via statistical multiplexing of loads and scheduling domains. We co-designed our medium-voltage power plane with a fast and robust power capping service that can coordinate multiple workload priorities. Because the applicable electrical limits are significantly larger with utility power than with generator power, the probability of power capping remains low even at high OSR. Since the events are rare, we can cap by a method that is both simple and reliable: we use our scheduler to reactively suspend low-priority workloads on the power plane. Thus, power is maintained at safe levels without violating the SLOs of either high- or low-priority workloads. We have constructed multiple data centers with our proposed architecture, serving large-scale production workloads over multiple years. Our measured power utilization distributions verify the expected oversubscription potential, and the fast power control response times. We are also releasing Google power usage data relevant to this paper with the goal of encouraging the broader community to pursue further research in this area.

The co-design in this paper is one example of optimizing across the data center technology stack to achieve significant cost efficiencies. We believe this approach can be extended further. By tailoring the OSR of a power plane to its workload and by further leveraging the stratification of workload according to multiple performance and availability SLOs, even larger infrastructure cost reductions will be possible.

7 Acknowledgements

This work is built on the infrastructure and contributions of a large number of Google colleagues across multiple teams. We are particularly grateful to Luiz Barroso, Dan Costello, Chung-Ching Chang, Mike Dahlin, Gaurav Dhiman, Mark Hennecke, Urs Hölzle, Rohit Jnagal, Joe Kava, Ed Lao, Frank McLoughlin, Peter Pellerzi, Jyoti, Sastry, Michal Tabor, Steve Webster, Tom Wenisch, Steven Zhao and John Zipfel for their feedback and support on this paper and their contributions to the system architecture development. We would also like to thank the anonymous reviewers for their feedback.

References

[1] Gartner, Inc. Gartner says global IT spending to grow 1.1 percent in 2019. http://bit.ly/gartner-2019-04-07. Accessed: 2020-01-23.

[2] Synergy Research Group. Hyperscale operator capex returns to growth mode in Q3. http://bit.ly/q3-2019-hyperscale-capex. Accessed: 2020-01-21.

[3] Luiz André Barroso, Urs Hölzle, and Parthasarathy Ranganathan. *The Datacenter as a Computer: Designing Warehouse-Scale Machines, Third Edition.* Morgan & Claypool. Available at http://bit.ly/dc-as-a-computer, 2018.

[4] Parthasarathy Ranganathan, Phil Leech, David Irwin, and Jeffrey Chase. Ensemble-level power management for dense blade servers. *ISCA '06: International Symposium on Computer Architecture*, 2006.

[5] Xiaobo Fan, Wolf-Dietrich Weber, and Luiz André Barroso. Power provisioning for a warehouse-sized computer. In *ISCA '07: International Symposium on Computer Architecture*, 2007.

[6] Charles Lefurgy, Xiaorui Wang, and Malcolm Ware. Power capping: a prelude to power shifting. *Cluster Computing*, 2008.

[7] Anshul Gandhi, Mor Harchol-Balter, Rajarshi Das, Jeffrey O. Kephart, and Charles Lefurgy. Power capping via forced idleness. *Workshop on Energy Efficient Design*, 2009.

[8] Kai Ma and Xiaorui Wang. PGCapping: exploiting power gating for power capping and core lifetime balancing in CMPs. In *Proceedings of the 21st International Conference on Parallel Architectures and Compilation Techniques*, 2012.

[9] Arka A. Bhattacharya, David Culler, Aman Kansal, Sriram Govindan, and Sriram Sankar. The need for speed and stability in data center power capping. *Sustainable Computing: Informatics and Systems*, 2013.

[10] Guosai Wang, Shuhao Wang, Bing Luo, Weisong Shi, Yinghang Zhu, Wenjun Yang, Dianming Hu, Longbo Huang, Xin Jin, and Wei Xu. Increasing large-scale data center capacity by statistical power control. In *Proceedings of the Eleventh European Conference on Computer Systems*, 2016.

[11] Qiang Wu, Qingyuan Deng, Lakshmi Ganesh, Chang-Hong Hsu, Yun Jin, Sanjeev Kumar, Bin Li, Justin Meza, and Yee Jiun Song. Dynamo: Facebook's data center-wide power management system. In *Proceedings of the 43rd International Symposium on Computer Architecture*, ISCA '16, pages 469–480, Piscataway, NJ, USA, 2016.

[12] Huazhe Zhang and Henry Hoffman. Maximizing performance under a power cap: A comparison of hardware, software, and hybrid techniques. In *ASPLOS '16*, 2016.

[13] Young Li, Charles R. Lefurgy, Karthick Rajamani, Malcome S. Allen-Ware, Guillermo J. Silva, Daniel D. Heimsoth, Saugata Ghose, and Onur Mutlu. A scalable priority-aware approach to managing data center server power. In *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2019.

[14] Sulav Malla and Ken Christensen. A survey on power management techniques for oversubscription of multi-tenant data centers. *ACM Computing Surveys*, 2019.

[15] Steven Pelley, David Meisner, Pooya Zandevakili, Thomas F. Wenisch, and Jack Underwood. Power routing: dynamic power provisioning in the data center. In *ASPLOS*, 2010.

[16] Sherief Reda, Ryan Cochran, and Ayse K. Coskun. Adaptive power capping for servers with multithreaded workloads. *IEEE Micro*, 2012.

[17] Reza Azimi, Masoud Badiei, Xin Zhan, Na Li, and Sherief Reda. Fast decentralized power capping for server clusters. In *2017 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2017.

[18] Song Wu, Yang Chen, Xinhou Wang, Hai Jin, Fangming Liu, Haibao Chen, and Chuxiong Yan. Precise power capping for latency-sensitive applications in datacenter. *IEEE Transactions on Sustainable Computing*, 2018.

[19] Howard David, Eugene Gorbatov, Ulf R. Hanebutte, Rahul Khanna, and Christian Le. RAPL: Memory power estimation and capping. In *ISLPED '10*, 2010.

[20] Intel Corporation. Intel 64 and IA-32 Architectures Software Developer's Manual, Volume 3B: System Programming Guide, Part 2. http://bit.ly/intel-rapl, 2016.

[21] Wonyoung Kim, Meeta Sharma Gupta, Gu-Yeon Wei, and David Brooks. System level analysis of fast, per-core DVFS using on-chip switching regulators. In *HPCA '08: 14th International Conference on High-Performance Computer Architecture*, pages 123–134, 2008.

[22] Uptime Institute. Data center site infrastructure tier standard topology. https://uptimeinstitute.com/tiers. Accessed: 2020-01-23.

[23] Telecommunications Industry Association standards. Telecommunications infrastructure set. http://global.ihs.com/tia_telecom_infrastructure.cfm?RID=Z56&MID=5280. Accessed: 2020-01-23.

[24] Abhishek Verma, Luis Pedrosa, Madhukar R. Korupolu, David Oppenheimer, Eric Tune, and John Wilkes. Large-scale cluster management at Google with Borg. In *Proceedings of the European Conference on Computer Systems (EuroSys)*, Bordeaux, France, 2015.

[25] Cummins Inc. Basics of paralleling. http://bit.ly/paralleling-gens. Accessed: 2020-01-23.

[26] Energy Information Administration. Average U.S. electricity customer interruptions totaled nearly 8 hours in 2017. http://bit.ly/saidi-saifi. Accessed: 2020-01-23.

[27] Vasileios Kontorinis, Varun Sakalkar, David Landhuis, and Parthasarathy Ranganathan. Google power data. https://rebrand.ly/28npdrt, February 2020.

[28] Yiyu Chen, Amitayu Das, Wubi Qin, Anand Sivasubramaniam, Qian Wang, and Natarajan Gautam. Managing server energy and operational costs in hosting centers. In *Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '05, pages 303–314, New York, NY, USA, 2005.

[29] Xiaorui Wang, Ming Chen, Charles Lefurgy, and Tom W. Keller. SHIP: A scalable hierarchical power control architecture for large-scale data centers. *IEEE Trans. Parallel Distrib. Syst.*, 23(1):168–176, 2012.

[30] Arka Aloke Bhattacharya, David E. Culler, Aman Kansal, Sriram Govindan, and Sriram Sankar. The need for speed and stability in data center power capping. *Sustainable Computing: Informatics and Systems*, 3:183–193, 2013.

[31] Ramya Raghavendra, Parthasarathy Ranganathan, Vanish Talwar, Zhikui Wang, and Xiaoyun Zhu. No "power" struggles: Coordinated multi-level power management for the data center. In *Proceedings of the 13th International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS XIII, pages 48–59, New York, NY, USA, 2008.

[32] Daniel Ellsworth, Tapasya Patki, Swann Perarnau, Sangmin Seo, Abdelhalim Amer, Judicael Zounmevo, Rinku Gupta, Kazutomo Yoshii, Henry Hoffman, Allen Malony, Martin Schulz, and Pete Beckman. Systemwide power management with Argo. In *2016 IEEE International Parallel and Distributed Processing Symposium Workshops*, 2016.

[33] Marco Basili. Medium voltage products. http://bit.ly/abb-medium-voltage-products. Accessed: 2020-01-23.

[34] Schneider Electric. Medium-voltage product offer. http://bit.ly/schneider-medium-voltage-products. Accessed: 2019-08-15.

[35] Sriram Govindan, Anand Sivasubramaniam, and Bhuvan Urgaonkar. Benefits and limitations of tapping into stored energy for datacenters. In *Proceedings of the 38th Annual International Symposium on Computer Architecture*, ISCA '11, pages 341–352, New York, NY, USA, 2011.

[36] Vasileios Kontorinis, Liuyi Eric Zhang, Baris Aksanli, Jack Sampson, Houman Homayoun, Eddie Pettis, Dean M. Tullsen, and Tajana Simunic Rosing. Managing distributed UPS energy for effective power capping in data centers. In *Proceedings of the 39th Annual International Symposium on Computer Architecture*, ISCA '12, pages 488–499, Washington, DC, USA, 2012.

[37] Di Wang, Chuangang Ren, Anand Sivasubramaniam, Bhuvan Urgaonkar, and Hosam Fathy. Energy storage in datacenters: What, where, and how much? In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '12, pages 187–198, New York, NY, USA, 2012.

[38] Xing Fu, Xiaorui Wang, and Charles Lefurgy. How much power oversubscription is safe and allowed in data centers? In *Proceedings of the 8th ACM International Conference on Autonomic Computing*, ICAC '11, pages 21–30, New York, NY, USA, 2011.

[39] Eli Cortez, Anand Bonde, Alexandre Muzio, Mark Russinovich, Marcus Fontoura, and Ricardo Bianchini. Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms. In *SOSP '17*, 2017.

[40] Christina Delimitrou and Christos Kozyrakis. Quasar: Resource-efficient QoS-aware cluster management. In *ASPLOS '14*, 2014.