EMPower: The Case for a Cloud Power Control Plane

Jonggyu Park Theano Stavrinos Simon Peter Thomas Anderson

UW FOCI (https://foci.uw.edu/)

Abstract. The combination of escalating application demand and the end of Dennard scaling has put energy management at the center of cloud operations, both in the core and at the edge. Because of the huge cost and long lead time of provisioning new data centers, many operators want to squeeze as much use out of existing data centers as possible, often limited by fixed power provisioning set at the time of construction. Workload demand spikes and the inherent variability of renewable energy, as well as customer desire for carbon free computing, make the data center power management problem even more challenging.

We believe it is time to build a power control plane to provide fine-grained observability and control over data center energy to operators. Our goal is to help make data centers substantially more elastic with respect to dynamic changes in energy sources and application needs, while still providing good performance to applications. There are many use cases for cloud power control, including increased power oversubscription and use of green energy, resilience to power failures, large-scale power demand response, and improved energy efficiency.

1 Introduction

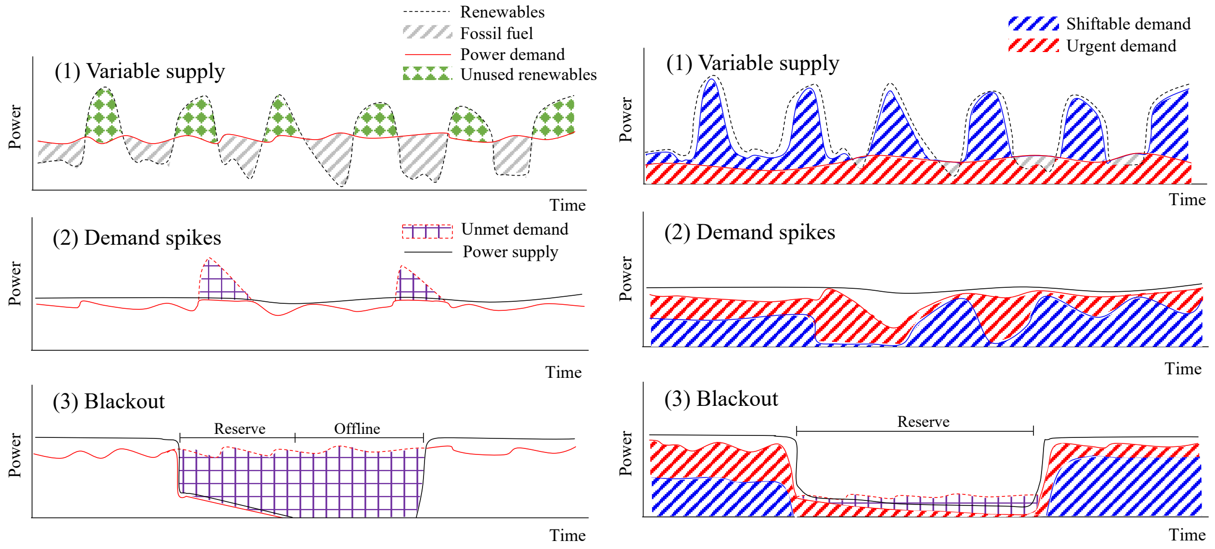
For a sustainable computing future, the age of abundant power in the cloud is nearing its end. A rapid increase in cloud application energy demand due to artificial intelligence and machine learning, as well as the tailing off of energy efficiency gains from Dennard scaling [5] has put power management at the center of cloud operations. In some regions, such as Northern Virginia and Ireland [8], data centers already consume more than 10% of grid power. That portion is projected to continue to grow, even relative to the increased supply needed to support decarbonization of transportation and building heating and cooling.

As a growing and already major power consumer, cloud data centers will increasingly have to balance fluctuating power demand and supply. In this scenario, there are three primary challenges for data center power management (cf. Figure 1a): (1) Solar and wind power—needed to support increased energy use by data centers, vehicles, and homes [22]—has volatile swings in power supply [60]. There can be undersupply and oversupply. (2) Power infrastructure is a significant capital expenditure; operators increasingly oversubscribe power to lower costs [38, 57, 65]. However, oversubscription reduces data center resilience to demand spikes. When demand spikes above supply, the demand cannot be met. (3) Extreme weather events, triggered in part due to climate change, are causing more volatility in demand, leading to blackouts and brownouts [14]. Data centers have limited power reserves, such as batteries and diesel generators, to be resilient to such situations, but they deplete quickly if power demand is not managed well.

To address these challenges, cloud data centers need to become more elastic with respect to dynamic changes in energy sources and application needs. To this end, we propose to build a *power control plane* (EMPower[1]) for cloud data centers. EMPower can observe and control power demand at a fine granularity and over short timescales (on the order of seconds) by making it *software-defined*. The key is to *gracefully trade off* power, performance, and application quality of service (QoS) over time. Our approach leverages the fact that application QoS requirements often allow for slack. This slack allows EMPower to conserve power during a power event by shedding and consolidating load, power-switching hardware components, and migrating critical workloads to less power-intensive processors, within QoS parameters, while less critical load is shifted to times with ample power supply (cf. Figure 1b).

Existing methods for addressing power-related challenges have been conservative, offering a narrow power control range. For instance, Google introduced a hardware-agnostic power capping system named Thunderbolt [40] that aims to reduce QoS violations while safely allowing power oversubscription. Thunderbolt regulates CPU power consumption by either limiting bandwidth or deactivating cores, balancing QoS with available

[1] EMPower = *E*lastic *M*anagement of *Power*.



(a) Without EMPower.

(b) With EMPower.

Figure 1: Data center power management challenges.

power. However, by focusing on CPUs, such systems support only a small power control dynamic range. Moreover, power attribution is too coarse-grained to accurately determine how applications consume power. Similarly, application QoS is often specified by virtual machine, rather than at a finer granularity. As a result, it is challenging to determine which application loads to control and by how much.

To push data center power control well beyond existing capabilities, EMPower will incorporate several novel power-saving mechanisms and policies by leveraging the capabilities offered by emerging microservice development models and modern hardware devices. For example, disaggregated memory presents a unique opportunity to decouple application state from heterogeneous compute cores with minimal overhead. EMPower can leverage disaggregated memory for aggressive consolidation of compute across servers and accelerators, while shutting down unused components to expand the power control dynamic range. These selections will be made in real time, guided by our policies.

To realize EMPower, we require a hardware/software co-design of next-generation data centers, workloads, and their run-time systems. We identify five challenges to realizing EMPower: (1) An effective power control plane must scale to include most of a data center's hardware and applications. (2) There is currently no mechanism for applications to convey fine-grained service-level agreements (SLAs) to operators, forcing operators to be conservative when deciding how to respond to power demand. (3) Power control policies must be automatic and robust over both long and short timescales (seconds or less). (4) The range of power-controllable hardware devices must be expanded to unlock the full power control dynamic range available in data centers. (5) Power instrumentation and control mechanisms today provide information and actuation at the wrong granularities, making it difficult to identify opportunities for efficiency and to respond to fluctuations in power availability.

We lay out a research agenda to address these challenges: (1) designing scalable mechanisms for power demand and supply instrumentation and control, (2) designing an interface for expressing applications' SLAs to allow operators to trade off performance and power consumption, (3) designing automatic power control policies that are robust over varying timescales, (4) shutting down servers and incorporating low-power heterogeneous compute cores to maximize the power control dynamic range, and (5) designing power instrumentation and control mechanisms that can measure and actuate at a fine-grain level of remote procedure calls and microservices, over short timescales.

We expect that EMPower will dramatically improve the energy efficiency of data centers, enable more renewable energy use, reduce the time to recover from power outages, and allow data centers to outlive power disruption events by leveraging software-defined power control. For example, EMPower allows individual cloud data centers to handle more load by enabling further oversubscription of available power beyond what can be safely achieved today. EMPower's power instrumentation insights allow developers to focus on code debloating to improve software energy efficiency. By quickly shedding load and power-switching associated hardware resources, EMPower makes data centers resilient to power supply variability, including power disruption and green power availability. Finally, EMPower can keep critical applications in operation in a power crisis and gracefully reduce the power demand of a data center when power is in short supply.

2 Cloud Power Control—Why Now?

With the commercial success of internet-scale applications and cloud computing, cloud infrastructure has grown rapidly. Estimates place data centers as responsible for 1–2% of aggregate worldwide electricity consumption [34, 54] and project that data center power consumption will grow to 10% of global electricity use by 2030 [34, 42, 46]. In many power grids, data centers are already major load contributors. For example, in Northern Virginia, data centers account for 12% of power consumption (2022), and are predicted to reach 22% in 2032 [18, 19]. In Ireland, data centers account for 14% of national electricity use (2022) [8] and may be 30% by 2029 [21]. In response, the Ireland national grid manager recently canceled more than 30 planned data center projects to preserve the stability of the grid [35]. These are just the leading edge. With continued cloud and artificial intelligence growth, the power draw of data centers is expected to be a large factor for many regional grids [29, 34, 46]. A consequence of this rapid growth is that data centers will need to operate under tight and variable power envelopes, if they are to be allowed access to grid power.

Due to the high cost of provisioning peak power, some hyperscale cloud data centers already oversubscribe their power infrastructure [38, 57, 65]. Under oversubscription, more servers are placed on a circuit than can be fully powered at peak load simultaneously. To make this work, providers have deployed power capping systems [40] to automatically shed non-critical load in overload situations. These generally are designed to make adjustments within a small dynamic range.

However, power demand and supply variability can occur suddenly and with large swings. For example, Google observed a 30× increase in compute demand for some applications during the first quarter of 2020 due to the the pandemic-induced spike in home-office use [7] and a major problem was provisioning enough power to fulfill the demand of newly deployed servers to handle the spike. On the supply side, renewable energy is becoming a primary power source [22]. Wind and solar power plants have large swings in production around their nominal generating capacity [60]. Even without renewables, an increase in natural disasters has led to more blackouts and brownouts—observed grid failures worldwide are 4× above IEEE expectations for commercial power systems [23], and failure frequency is trending upward [14]. The problem may also become self-made: as the largest data centers become increasingly power proportional, with large load swings [41], there is an emerging possibility of grid-destabilizing power demand changes. Grid failures are especially a problem for edge data centers, as they often cannot afford the luxury of partnering with multiple power providers for redundancy and have limited power storage facilities.

3 Challenges

Power variability has traditionally not been an area of focus for system designers. Existing systems have a small dynamic range of power control, as well as coarse-grained instrumentation and load control. Challenges include high idle power consumption for servers, power instrumentation only at the chassis and CPU socket levels, coarse-grained load control at the virtual machine level, and missing integration with accelerators, such as GPUs and SmartNICs. We describe these challenges in this section and explain how they make it difficult to support efficient power control at scale. Building a power control plane requires us to overcome them.

Limited power control dynamic range. Cloud hardware traditionally has a small dynamic range for power consumption. Server power control features, such as dynamic voltage and frequency scaling (DVFS) and running average power limit (RAPL) allow only limited control over CPU, GPU, and memory power [9, 39, 53]. Cloud servers consume a large amount of idle power that cannot be controlled via either mechanism. We have tested a variety of servers, including on CloudLab [17]. We found that the total server idle power, where no application is running, ranges from 58 to 220 watts, depending on the machine in use. For GPU-optimized servers, idle power can be as high as 600 watts. For EMPower to be effective, we believe it is necessary to make servers more power-proportional—*i.e.*, more efficient at any utilization—to increase the data center's power control dynamic range.

Coarse power instrumentation granularity. The advent of power instrumentation is a pivotal development in understanding energy consumption in data centers. Nevertheless, the granularity at which power instrumentation is available remains a challenge. Currently, these measurements exist primarily at coarse granularities, such as the full chassis through IPMI [31] or at the CPU socket level using RAPL [52]. However, in modern cloud environments, resource multiplexing is an essential mechanism for improving resource utilization. Existing methods for measuring power cannot attribute power consumption to individual applications or processes multiplexed on the same hardware. Per-application or per-process power measurements thus remain elusive, making it difficult to identify inefficiencies in software and to fully realize elastic power control. Finally, fine-grained power instrumentation is not a panacea in a cloud environment. Power consumption is a common attack vector in side-channel attacks and we have to make sure our collected data is safeguarded properly.

Coarse power control granularity. Cloud applications are increasingly built using a microservice development model, where applications are partitioned into fine-grained modules that encapsulate service and state. Microservices are naturally resilient to service failure and they are elastic—they can quickly add or remove service replicas. The microservices model thus simplifies migrating and shedding load in response to power supply or demand changes. These developments significantly reduce the barrier to effective load control at scale. Unfortunately, systems software does not yet control power at the same fine granularity, utilizing primarily core and socket-level power control. To realize fine-grained power control, we have to control CPU, IO, and memory utilization at a per-process level.

Further, many cloud applications are still designed as monoliths or leverage heavy-weight virtual machine (VM) technology to implement microservices. Shedding load with VMs involves shutting down an entire VM, and migration involves moving an entire VM's state among servers, which can take minutes. To support these applications we have to provide lighter-weight and fine-grained load control for legacy VMs.

Limited integration with accelerators and IO devices. Accelerators like GPUs, which are useful in handling specific parallelizable tasks, offer limited software control over power instrumentation and control mechanisms [53]. IO devices such as NICs and storage drives may have no established mechanisms at all. Understanding and controlling power in accelerators and IO devices is important for two reasons. First, such components contribute significantly to the power draw of servers [20, 47]. Second, accelerators, especially GPUs, contribute an increasing amount to the overall energy consumption of a data center. Integrating these devices into power control decisions is thus important for increasing the power control dynamic range.

Scalability. A perennial challenge of data center infrastructure design is scale. A data center power control plane must, in a timely manner, process information from power instrumentation and actuate power control over millions of heterogeneous processors and accelerators, applications, and hardware devices. It also must do it in a timely manner, without violating SLAs and under bursty power budgets. For EMPower to be successful, we have to design it from the ground up to be scalable, as well as with low-overhead and low-latency measurement and actuation mechanisms, down to OS- and VM-level CPU, accelerator, memory, and IO scheduling, socket allocation, and process/VM assignment.

4 EMPower: A Cloud Power Control Plane

We propose to build EMPower, a scalable cloud power control plane. EMPower will feature operating system mechanisms for power instrumentation and control that realize fine-grained and scalable power control policies. EMPower will address the challenges outlined in Section 3 by integrating (1) server shutdown to widen the available power control dynamic range, (2) fine-grain power instrumentation via performance counters and power models down to the process and procedure call level, (3) novel OS mechanisms to provide fine-grained power control for microservices via modern hardware interfaces, (4) new cloud infrastructure stacks that support low-power processing, and (5) hierarchical power instrumentation and control that can operate at scale.

The overview of EMPower is illustrated in Figure 2. Each physical server maintains records of SLA and power information, while a EMPower master controller collects this data through the network hierarchy. With this aggregated information, the EMPower master controller establishes the power budget and disseminates it via the network hierarchy. Switches may subdivide their budgets hierarchically, taking into account the power budget and workloads. Finally, the servers control process and VM load and implement power-saving mechanisms, such as node shutdown. We provide a detailed explanation of our approach in this section.

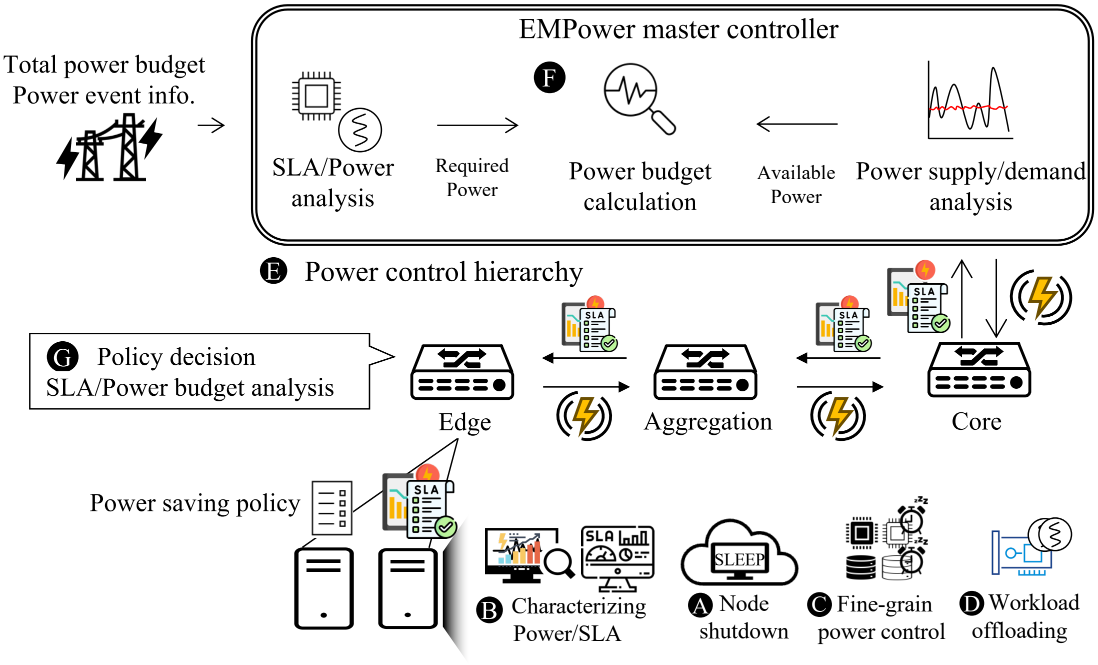


Figure 2: EMPower overview.

Large power control dynamic range via server shutdown. To expand the power control dynamic range, EMPower remotely shuts down and starts up entire servers via built-in board management controllers to further reduce unnecessary idle power, for example using the intelligent platform management interface (IPMI) [31] or Redfish [16].

To allow us to do this without disrupting services, we plan to leverage disaggregated memory, such as via Compute eXpress Link (CXL) [13], to store virtual machine, operating system, and process snapshots, facilitating rapid power demand adjustments. Disaggregated memory allows its contents to stay accessible with low latency even after a server is powered down. This shift in capability enables power control over short timescales. EMPower can shut down entire servers and bring them back in a fraction of the time required for traditional servers that need to be restored entirely from block storage systems. Moreover, disaggregated memory remains accessible without requiring a host to be online. Hence, EMPower can enhance the power control dynamic range by allowing servers to be shut down while still allowing access to application data from other servers. Finally, disaggregated memory can facilitate coalescing of microservice state, thereby enabling efficient migration of compute loads. This increases the flexibility in scheduling workloads across machines while managing their power consumption [12, 63].

By leveraging these emerging technologies, we propose to develop a new hibernation technique (Ⓐ) that progressively saves system images to disaggregated memory through write-through mechanisms. This approach can markedly reduce the hibernation duration. Additionally, since disaggregated memory is byte-addressable, system states can be promptly recovered. Direct access to pooled memory allows for the immediate restoration of hot data to local memory. Consequently, EMPower can significantly reduce server idle power.

Fine-grained power instrumentation. To be able to effectively control power, we have to have a precise understanding of how application software is consuming power. Conventional hardware-based power management tools, such as RAPL and IPMI, are too coarse-grained to fulfill such a need. RAPL can measure the system-wide CPU and memory energy consumption while IPMI can measure the machine-level power consumption. However, EMPower needs to be able to measure the power in terms of applications, processes, and even remote procedure calls. To do so, we propose to develop power consumption models, APIs, and instrumentation tools (Ⓑ) that enable us to characterize power consumption by applications across all involved data center hardware components, as well as across the entire software stack. This accounting will be similar to the *perf* tool [55], which profiles applications' CPU usage for performance debugging. Our instrumentation will report power consumption at a similar level of detail, by instrumenting performance counters and leveraging power models to translate performance to power consumption information. To prevent abuse, this power information is retained within EMPower and accessible only by cloud operators. There is a potential benefit of making power instrumentation available also to cloud tenants. To minimize the potential for attacks, EMPower has to isolate power readings or report them with reduced fidelity or frequency to limit the exploitable bit-rate of this side-channel.

Coordinated and fine-grained power control via modern hardware interfaces. To realize fine-grained power control, we have to control CPU, IO, and memory load at a per-process level. Hardware support for fine-grained load control is increasingly available. Techniques, such as CPU jailing and bandwidth control [40], allow us to control CPU load at a fine-granularity. Interfaces, such as Intel's memory bandwidth allocation architecture [24] and cache allocation technology [49] allow us to control per-process memory bandwidth and cache utilization, while modern NICs [30] and SSDs [50] have the ability to limit IO bandwidth utilization to control IO load at a fine granularity. As hardware becomes more power proportional, it is important to utilize these knobs. Unfortunately, current operating systems do not exploit such hardware-provided load control mechanisms in concert. Our goal is to develop the necessary OS policies and mechanisms that collectively leverage these load control techniques to maintain power draw within its budget.

A practical illustration of this is the strategic confinement of applications to server sockets, while limiting access to memory and storage bandwidth (🅒) to reduce power consumption. For example, EMPower may decide to reduce the power consumption of a number of servers by shutting down one socket, but also reducing the number of active memory and SSD channels in tandem with the reduced compute load, to balance system resources and reduce idle power consumed by these resources. When a workload is memory intensive, EMPower may decide to leave memory bandwidth at capacity to allow the workload to finish within its SLA.

Cloud infrastructure software stack design for low-power processing. We plan to redesign the cloud software infrastructure stack with support for low power modes. For example, to be able to shutdown servers (🅐), cluster managers need to treat server shutdown as a new operational mode, distinct from a failure. To be able to use low-power processing, microservice runtimes need to support transparent migration of applications to low-power processors, such as SmartNICs (🅓). Similarly, disaggregated cloud services, such as storage, need to support these processors. We previously developed prototype runtimes and storage services leveraging low power options (e.g., iPipe [43] and E3 [44] for microservices, and LineFS [37] for storage) and we plan to extend them to support low-latency application and service migration when power budgets change. We also plan to redesign many further cloud services, such as network communication, locking, and load balancing, to support low-power operation.

Leveraging hierarchy for power instrumentation and control scalability. To efficiently utilize power and safely run applications, we need EMPower to react quickly to changes in power demand and supply. At the same time, EMPower has to be robust in the sense that it does not violate application SLAs, does not lead to tripped power breakers, and interacts well with power grids.

Hierarchical aggregation and budgeting (🅔) will scale power measurement and actuation. In the limit, we will leverage the hierarchy of data center network topologies [2] via programmable switches by extending existing network control planes to aggregate power measurements and SLA information at the server, rack, and pod (set of racks) level. The aggregates will be forwarded to a central power controller that takes power supply measurements, executes a global power control policy, and then distributes power budgets back to switches and servers for effective power actuation, local to each server.

In detail, EMPower calculates the available power budget considering both power supply and demand. Meanwhile, EMPower also estimates the required power to guarantee application SLAs using the SLA and power information. Using the available power budget and required power, EMPower determines per-pod power budgets (🅕) and delivers them to the corresponding switches by leveraging the data center network hierarchy. Pod-level switches then subdivide the budget to the rack-level. Rack-level switches finally subdivide to per-server budgets and determine necessary power-saving policies depending on the power budget and available power-saving mechanisms (🅖). OSes and hypervisors may further subdivide the budget to per-socket, per-core, and per-application budgets.

Note that power distribution does not always correspond to network hierarchy in data centers. As such, each rack and pod might not represent a power failure domain from a physical perspective. However, simpler power control hierarchies are possible. For example, servers can aggregate and actuate per-core and per-socket power measurements in a virtual hierarchy to and from the central controller.

5 Use Cases

There are many use cases for cloud power control, including increased power oversubscription, resilience to power failures, large-scale power demand response, improved energy efficiency, and use of green energy. We detail these examples here.

Power oversubscription. The capacity demand of data centers has been increasing, drastically raising cost. Power distribution infrastructure, in particular, is a growing cost center. In the meantime, the actual power consumption of data centers is often less than the maximum power draw of the deployed equipment [57]. For this reason, hyperscalers often oversubscribe power—the provisioning of more machines than the power supply can fully support at 100% utilization. However, oversubscription can threaten to push a data center's power draw beyond its supply unexpectedly.

Continuously controlling the load and power demand to prevent overloading power equipment is the job of power capping systems [38, 40, 65, 66]. In comparison to existing systems, EMPower enables increased power oversubscription by providing a larger power control dynamic range by shutting down servers and migrating workloads to low-power processors, as well as finer grained control over power demand and supply, allowing power demand to be controlled much closer to the given power supply envelope.

Power resilience. Data center power supply is increasingly threatened by blackouts and brownouts from natural disasters—in particular climate events, which are becoming more frequent due to global warming—and failures of aging grid infrastructure [14, 15, 23]. This trend is especially salient for edge data centers which often receive power from only one utility [48].

EMPower can handle these unexpected events by keeping track of available reserved energy, such as batteries and generators, and shedding an adequate amount of non-critical load to fit the power budget.

Power demand response. Demand response refers to adjusting power draw in response to changes in the power supply. Effective demand response has financial benefits and offers power resiliency. For instance, grid operators may increase the cost of power to incentivize lower demand from grid customers, e.g., when renewable energy makes up a small share of the energy mix or when the grid is particularly strained.

Supporting power demand response is poised to become table-stakes for new data center deployments. For instance, the Irish government has expressed a preference for energy-efficient and carbon-conscious data center developments [33]. Included in the preference are techniques to adapt to variable grid demands on power consumption. By gracefully degrading non-critical load to reduce a data center's power draw at times of extreme grid-wide power demand, EMPower can support demand response to enable the deployment of new data centers that not only preserve grid health, but can also reduce operational costs.

Energy efficiency. Many existing cloud APIs layer inefficient implementations, wasting energy. EMPower can help identify and debug software inefficiencies by providing a granular accounting of the power consumption of individual components of the application stack. A power control plane would profile applications' power usage to help developers identify code and application architectures that may be streamlined. The power profiling information supplied by EMPower could also be used to quantitatively compare system and application designs for energy efficiency.

Low-carbon computing. The large-scale computing industry is growing faster than green energy sources can be brought online. To date, the price of emitting carbon is too low to drive the data center's carbon efficiency. This will change over time. In the longer term, the tools we build for managing power consumption can be used for minimizing data center carbon and will make that transition much easier.

Many of the techniques employed in EMPower can form a foundation of low-carbon computing. For example, common features of low-carbon computing platforms are to handle power supply swings caused by renewable energy and to estimate the carbon emissions of servers [1, 61]. EMPower natively addresses power volatility (§2), and its power consumption measurements can support carbon emissions estimation. We anticipate cloud operators will be forced to adopt power control planes to address urgent power control problems. We believe a side benefit is that these power control planes can also facilitate the longer-term social goal of low-carbon computing.

6 Related Work

Power capping. Data center operators are deploying power capping systems [40] that enable increasing oversubscription [38, 57, 65] of data center power infrastructure. The key criterion for such systems is to uphold quality-of-service guarantees, while shedding load that would exceed a predefined power envelope. Server overload control that preserves latency targets has also been investigated [11] for non-power-capped scenarios. EMPower builds on this work to enhance the power control dynamic range and provide elastic power control beyond oversubscription, including supply variability, resilience, and demand control.

Intermittent computing. Many system-level hardware and software techniques address continued operation under variable power supply, for instance for Internet-of-Things devices, edge computing, and energy-harvesting computing environments. The most extreme scenario is intermittent computing, where only a fraction of peak power is available for extended periods [47, 59, 64]. Some techniques include fast snapshotting and restoration during low-power events and the use of heterogeneous hardware with different power profiles. We adapt some of this work to operate beyond the server scale. In particular, EMPower combines OS and networking techniques to enable fast reaction and control at rack scale and beyond.

Resource disaggregation. Resource disaggregation is a recent hot topic in the systems and architecture communities [26, 27, 58], supported by emerging hardware and protocols [13, 51]. Its primary purpose is to pool hardware resources, including memory and storage, to enable more efficient sharing and to raise utilization. We plan to build on recent resource disaggregation support for fine-grained power control, such as shutting down a server chassis, while keeping pooled memory online.

Power proportionality. Power proportionality is a requirement of EMPower, since it increases the power control dynamic range of the data center [45]. Existing work targets power proportionality on single servers, for instance using dynamic voltage and frequency scaling (DVFS) and hardware sleep states [3, 6, 32, 53, 56]. However, DVFS provides only a small power dynamic range [39], and these mechanisms are typically only scoped to a single server. Other work aims for system-wide power proportionality, i.e., across multiple nodes, often by leveraging the different power profiles and capabilities of heterogeneous hardware [4, 10, 45, 47]. We plan to enhance data center power proportionality by integrating low-power processors (e.g., SmartNICs [37, 44]), server hibernation, and fine-grained instrumentation and control.

Energy attribution. There have been several research efforts to measure power consumption in cloud servers and applications [25, 28, 36, 62]. For example, EnergAt presents a thread-level, NUMA-aware energy attribution for CPU and DRAM in multi-tenant environments [28]. However, EnergAt uses up to 10% of an application's energy to determine its energy consumption, which is too high for continuous use, such as in EMPower. Other systems use performance counters, accessed through hardware interfaces or *perf* [55], to estimate the power consumption at a container level [25, 36, 62]. However, such event monitoring from VMs is unavailable in cloud settings since it is a privileged task. To support monitoring in a cloud setting, systems often rely on CPU occupation time, which is inaccurate, or a customized hypervisor to estimate container-level CPU power consumption. Moreover, the related work focuses on CPU and memory power consumption without considering other components, including accelerators and peripherals. EMPower collects power consumption information beyond CPU occupation time, allowing for precise attribution of energy to applications in a cloud computing environment. EMPower avoids the security risks of existing approaches by supplying applications with appropriate power data aggregates, limiting the attack surface of side-channels.

7 Conclusion

We underscore the critical need for a power control plane in cloud data centers, driven by the end of Dennard scaling, rising power costs, increased use of renewables, increased extreme weather events, and sudden power demand surges. We propose EMPower to provide fine-grained, scalable control over data center power use, aiming to enhance data center elasticity in response to dynamic changes in energy demand and supply. New technologies, including disaggregated memories, low-power compute devices, programmable switches, and fine-grained development models, open the opportunity for EMPower. We envision several use cases for cloud power control, including increased power oversubscription, use of green energy, resilience to power failures, and improved energy efficiency.

Acknowledgments

This work is supported by the National Science Foundation grants CNS-2104548 and CNS-2148209, as well as funding from Alibaba, Amazon, Cisco Systems, Google, Microsoft, and VMware.

References

[1] Bilge Acun, Benjamin Lee, Fiodar Kazhamiaka, Kiwan Maeng, Udit Gupta, Manoj Chakkaravarthy, David Brooks, and Carole-Jean Wu. Carbon Explorer: A Holistic Framework for Designing Carbon Aware Datacenters. In *Proceedings of ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, page 118–132, 2023.

[2] Mohammad Al-Fares, Alexander Loukissas, and Amin Vahdat. A Scalable, Commodity Data Center Network Architecture. In *Proceedings of the ACM SIGCOMM Conference on Data Communication*, page 63–74, 2008.

[3] AMD. AMD PowerNow! Technology. https://www.amd.com/content/dam/amd/en/documents/archived-tech-docs/white-papers/24404a.pdf.

[4] David G Andersen, Jason Franklin, Michael Kaminsky, Amar Phanishayee, Lawrence Tan, and Vijay Vasudevan. FAWN: A Fast Array of Wimpy Nodes. In *Proceedings of ACM SIGOPS Symposium on Operating Systems Principles*, pages 1–14, 2009.

[5] Thomas Anderson, Adam Belay, Mosharaf Chowdhury, Asaf Cidon, and Irene Zhang. Treehouse: A Case For Carbon-Aware Datacenter Software. In *HotCarbon: Workshop on Sustainable Computer Systems Design and Implementation*, 2022.

[6] Esmail Asyabi, Azer Bestavros, Erfan Sharafzadeh, and Timothy Zhu. Peafowl: In-application CPU Scheduling to Reduce Power Consumption of In-memory Key-Value Stores. In *Proceedings of ACM Symposium on Cloud Computing*, pages 150–164, 2020.

[7] Brian Barrett. How Google Meet Weathered the Work-From-Home Explosion. https://www.wired.com/story/how-google-meet-weathered-work-from-home-explosion/.

[8] John Campbell. Data Centres Used 14% of Republic of Ireland's Electricity Use. https://www.bbc.com/news/world-europe-61308747.

[9] Aaron Carroll and Gernot Heiser. An Analysis of Power Consumption in a Smartphone. In *Proceedings of USENIX Annual Technical Conference*, pages 1–14, 2010.

[10] Geoffrey Challen and Mark Hempstead. The Case for Power-agile Computing. In *Proceedings of USENIX Conference on Hot Topics in Operating Systems*, pages 1–5, 2011.

[11] Inho Cho, Ahmed Saeed, Joshua Fried, Seo Jin Park, Mohammad Alizadeh, and Adam Belay. Overload Control for us-scale RPCs with Breakwater. In *Proceedings of USENIX Symposium on Operating Systems Design and Implementation*, pages 299–314, 2020.

[12] Jonathan Corbet. Live Migration of Virtual Machines over CXL. https://lwn.net/Articles/931528/.

[13] CXL™ Consortium. Compute Express Link. https://www.computeexpresslink.org/about-cxl.

[14] DARPA/ISAT Workshop on Energy-Resilient Systems, December 2020.

[15] Jacqueline Davis. Data Center Operators Will Face More Grid Disturbances. https://journal.uptimeinstitute.com/data-center-operators-will-face-more-grid-disturbances/.

[16] DMTF. Redfish Developer Hub. https://redfish.dmtf.org/.

[17] Dmitry Duplyakin, Robert Ricci, Aleksander Maricq, Gary Wong, Jonathon Duerig, Eric Eide, Leigh Stoller, Mike Hibler, David Johnson, Kirk Webb, Aditya Akella, Kuangching Wang, Glenn Ricart, Larry Landweber, Chip Elliott, Michael Zink, Emmanuel Cecchet, Snigdhaswin Kar, and Prabodh Mishra. The Design and Operation of CloudLab. In *Proceedings of USENIX Annual Technical Conference*, pages 1–14, 2019.

[18] Dominion Energy. 2020 Virginia Integrated Resource Plan. https://www.dominionenergy.com/-/media/pdfs/global/2020-va-integrated-resource-plan.pdf.

[19] Dominion Energy. 2021 Update to the 2020 Integrated Resource Plan. https://www.dominionenergy.com/-/media/pdfs/global/company/2021-de-integrated-resource-plan.pdf.

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

[21] Robbie Galvin. Data Centers Are Pushing Ireland's Electric Grid to the Brink. https://gizmodo.com/data-centers-are-pushing-ireland-s-electric-grid-to-the-1848282390.

[22] Greenpeace. Clicking Clean: Who is Winning the Race to Build a Green Internet? https://www.greenpeace.de/publikationen/20170110_greenpeace_clicking_clean.pdf.

[23] C. Heising. IEEE Recommended Practice for the Design of Reliable Industrial and Commercial Power Systems, 2007. IEEE 493-2007.

[24] Andrew J Herdrich, Marcel David Cornu, and Khawar Munir Abbasi. Introduction to Memory Bandwidth Allocation. https://www.intel.com/content/www/us/en/developer/articles/technical/introduction-to-memory-bandwidth-allocation.html.

[25] Hubblo. Scaphandre. https://github.com/hubblo-org/scaphandre.

[26] Jaehyun Hwang, Qizhe Cai, Ao Tang, and Rachit Agarwal. TCP ≈ RDMA: CPU-efficient Remote Storage Access with i10 . In *Proceedings of USENIX Symposium on Networked Systems Design and Implementation*, pages 127–140, 2020.

[27] Jaehyun Hwang, Midhul Vuppalapati, Simon Peter, and Rachit Agarwal. Rearchitecting Linux Storage Stack for μs Latency and High Throughput. In *Proceedings of USENIX Symposium on Operating Systems Design and Implementation*, pages 113–128, 2021.

[28] Hongyu Hè, Michal Friedman, and Theodoros Rekatsinas. EnergAt: Fine-Grained Energy Attribution for Multi-Tenancy. In *Proceedings of Workshop on Sustainable Computer Systems*, pages 1–8, 2023.

[29] Kevin Imboden. 2022 Global Data Center Market Comparison. https://cushwake.cld.bz/2022-Global-Data-Center-Market-Comparison.

[30] Intel Corporation. Intel 82599 10 GbE Controller Datasheet. Revision 2.6.

[31] Intel Corporation. Intelligent Platform Management Interface Specification Second Generation v2.0. https://www.intel.com/content/www/us/en/products/docs/servers/ipmi/ipmi-second-gen-interface-spec-v2-rev1-1.html.

[32] Intel Corporation. Overview of Enhanced Intel SpeedStep® Technology for Intel® Processors. https://www.intel.com/content/www/us/en/support/articles/000007073/processors.html.

[33] Ireland Government Statement on the Role of Data Centres in Ireland's Enterprise Strategy. https://assets.gov.ie/231142/e108d6fa-c769-4286-8fb4-0e2ff07548fe.pdf.

[34] Nicola Jones. How to Stop Data Centres From Gobbling Up the World's Electricity. *Nature*, 561(7722):163–167, 2018.

[35] Peter Judge. EirGrid Pulls Plug on 30 Irish Data Center Projects. https://www.datacenterdynamics.com/en/news/eirgrid-pulls-plug-on-30-irish-data-center-projects/.

[36] Kubernetes Efficient Power Level Exporter (Kepler). https://github.com/sustainable-computing-io/kepler.

[37] Jongyul Kim, Insu Jang, Waleed Reda, Jaeseong Im, Marco Canini, Dejan Kostić, Youngjin Kwon, Simon Peter, and Emmett Witchel. LineFS: Efficient SmartNIC Offload of a Distributed File System with Pipeline Parallelism. In *Proceedings of the ACM SIGOPS Symposium on Operating Systems Principles*, page 756–771, 2021.

[38] 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. Prediction-Based Power Oversubscription in Cloud Platforms. In *Proceedings of USENIX Annual Technical Conference*, pages 473–487, 2021.

[39] Etienne Le Sueur and Gernot Heiser. Dynamic Voltage and Frequency Scaling: The Laws of Diminishing Returns. In *Proceedings of International Conference on Power Aware Computing and Systems*, page 1–8, 2010.

[40] Shaohong Li, Xi Wang, Xiao Zhang, Vasileios Kontorinis, Sreekumar Kodakara, David Lo, and Parthasarathy Ranganathan. Thunderbolt:

Throughput-Optimized, Quality-of-Service-Aware Power Capping at Scale. In *Proceedings of USENIX Symposium on Operating Systems Design and Implementation*, pages 1241–1255, 2020.

[41] Liuzixuan Lin and Andrew A Chien. Adapting Datacenter Capacity for Greener Datacenters and Grid. In *Proceedings of ACM International Conference on Future Energy Systems*, page 200–213, 2023.

[42] Liuzixuan Lin, Victor M. Zavala, and Andrew A. Chien. Evaluating Coupling Models for Cloud Datacenters and Power Grids. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*, page 171–184, 2021.

[43] Ming Liu, Tianyi Cui, Henry Schuh, Arvind Krishnamurthy, Simon Peter, and Karan Gupta. Offloading Distributed Applications onto SmartNICs using iPipe. In *Proceedings of ACM Special Interest Group on Data Communication*, pages 318–333, 2019.

[44] Ming Liu, Simon Peter, Arvind Krishnamurthy, and Phitchaya Mangpo Phothilimthana. E3: Energy-Efficient Microservices on SmartNIC-Accelerated Servers. In *Proceedings of USENIX Annual Technical Conference*, pages 363–378, 2019.

[45] David Lo, Liqun Cheng, Rama Govindaraju, Luiz André Barroso, and Christos Kozyrakis. Towards Energy Proportionality for Large-scale Latency-critical Workloads. In *Proceedings of ACM/IEEE International Symposium on Computer Architecture*, pages 301–312, 2014.

[46] Eric Masanet, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey. Recalibrating Global Data Center Energy-Use Estimates. *Science*, 367(6481):984–986, 2020.

[47] David Meisner, Brian T Gold, and Thomas F Wenisch. PowerNap: eliminating server idle power. *ACM SIGARCH Computer Architecture News*, 37(1):205–216, 2009.

[48] Bruce Myatt and Russell Carr. Advanced Microgrids as a Resiliency Strategy for Federal Data Centers. https://datacenters.lbl.gov/sites/default/files/Designing%20and%20Managing%20Data%20Centers%20for%20Resilience%20-%20Demand%20Response%20and%20Microgrids_3Dec2019_0.pdf.

[49] Khang T Nguyen. Introduction to Cache Allocation Technology in the Intel® Xeon® Processor E5 v4 Family. https://www.intel.com/content/www/us/en/developer/articles/technical/introduction-to-cache-allocation-technology.html.

[50] NVM Express Workgroup. NVM Express 1.2.1. http://www.nvmexpress.org/wp-content/uploads/NVM_Express_1_2_1_Gold_20160603.pdf.

[51] NVM Express Workgroup. NVM ExpressTM over Fabrics Revision 1.1. https://nvmexpress.org/wp-content/uploads/NVMe-over-Fabrics-1.1-2019.10.22-Ratified.pdf.

[52] Srinivas Pandruvada. Running Average Power Limit. https://01.org/blogs/2014/running-average-power-limit-%E2%80%93-rapl.

[53] Pratyush Patel, Zibo Gong, Syeda Rizvi, Esha Choukse, Pulkit Misra, Thomas Anderson, and Akshitha Sriraman. Towards Improved Power Management in Cloud GPUs. *IEEE Computer Architecture Letters*, pages 1–4, 2023.

[54] Fred Pearce. Energy Hogs: Can World's Huge Data Centers Be Made More Efficient? *Yale Environment 360*, 2018.

[55] Perf Wiki. perf: Linux Profiling With Performance Counters. https://perf.wiki.kernel.org.

[56] George Prekas, Mia Primorac, Adam Belay, Christos Kozyrakis, and Edouard Bugnion. Energy Proportionality and Workload Consolidation for Latency-critical Applications. In *Proceedings of ACM Symposium on Cloud Computing*, pages 342–355, 2015.

[57] Varun Sakalkar, Vasileios Kontorinis, David Landhuis, Shaohong Li, Darren De Ronde, Thomas Blooming, Anand Ramesh, James Kennedy, Christopher Malone, Jimmy Clidaras, and Parthasarathy Ranganathan. 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*, page 497–511, 2020.

[58] Yizhou Shan, Yutong Huang, Yilun Chen, and Yiying Zhang. LegoOS: A Disseminated, Distributed OS for Hardware Resource Disaggregation. In *Proceedings of USENIX Conference on Operating Systems Design and Implementation*, page 69–87, 2018.

[59] Navin Sharma, Sean Barker, David Irwin, and Prashant Shenoy. Blink: Managing Server Clusters on Intermittent Power. In *Proceedings of International Conference on Architectural Support for Programming Languages and Operating Systems*, page 185–198, 2011.

[60] Hans-Werner Sinn. Buffering Volatility: A Study on the Limits of Germany's Energy Revolution. *European Economic Review*, 99:130–150, 2017.

[61] Abel Souza, Noman Bashir, Jorge Murillo, Walid Hanafy, Qianlin Liang, David Irwin, and Prashant Shenoy. Ecovisor: A Virtual Energy System for Carbon-Efficient Applications. In *Proceedings of ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, page 252–265, 2023.

[62] Spirals Research Group. PowerAPI. https://github.com/powerapi-ng.

[63] Dragan Stancevic. nil-migration: Nearly Instantaneous Live Migration of Virtual Machines, Containers, and Processes. https://nil-migration.org/.

[64] Sumanth Umesh and Sparsh Mittal. A Survey of Techniques for Intermittent Computing. *Journal of Systems Architecture*, 112:101859–101859, 2021.

[65] 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*, page 469–480, 2016.

[66] Chaojie Zhang, Alok Gautam Kumbhare, Ioannis Manousakis, Deli Zhang, Pulkit A. Misra, Rod Assis, Kyle Woolcock, Nithish Mahalingam, Brijesh Warrier, David Gauthier, Lalu Kunnath, Steve Solomon, Osvaldo Morales, Marcus Fontoura, and Ricardo Bianchini. Flex: High-Availability Datacenters With Zero Reserved Power. In *Proceedings of ACM/IEEE Annual International Symposium on Computer Architecture*, pages 319–332, 2021.