CHAPTER 5

Energy and Power Efficiency

Energy efficiency has been a major technology driver in the mobile and embedded areas for a long time. Work in this area originally emphasized extending battery life, but then expanded to include reducing peak power because thermal constraints began to limit further CPU performance improvements or packaging density in small devices. However, energy management is also a key issue for servers and data center operations, one that focuses on reducing all energy-related costs, including capital and operating expenses as well as environmental impacts. Many energy-saving techniques developed for mobile devices are natural candidates for tackling this new problem space, but ultimately a WSC is quite different from a mobile device. In this chapter, we describe some of the most relevant aspects of energy and power efficiency for WSCs, starting at the data center level and continuing to component-level issues.

5.1 DATA CENTER ENERGY EFFICIENCY

The broadest definition of WSC energy efficiency would measure the energy used to run a particular workload (say, to sort a petabyte of data). Unfortunately, no two companies run the same workloads and, as discussed in Chapter 2, real-world application mixes change all the time, so it is hard to benchmark WSCs this way. Thus, even though such benchmarks have been contemplated [Riv+07], they haven't yet been widely used [TGGb]. However, it is useful to view energy efficiency as the product of three factors we can independently measure and optimize:

$$\text{Efficiency} = \frac{\text{Computation}}{\text{Total Energy}} = \underbrace{\left(\frac{1}{\text{PUE}}\right)}_{\text{(a)}} \times \underbrace{\left(\frac{1}{\text{SPUE}}\right)}_{\text{(b)}} \times \underbrace{\left(\frac{\text{Computation}}{\text{Total Energy to Electronic Components}}\right)}_{\text{(c)}}.$$

In this equation, the first term (a) measures facility efficiency, the second (b) measures server power conversion efficiency, and the third (c) measures the server's architectural efficiency. We discuss these factors in the following sections.

5.1.1 THE PUE METRIC

Power usage effectiveness (PUE) reflects the quality of the data center building infrastructure itself [TGGc], and captures the ratio of total building power to IT power (the power consumed by the computing, networking, and other IT equipment). IT power is sometimes referred to as "critical power."

L. A. Barroso et al., *The Datacenter as a Computer*

PUE = (Facility power) / (IT Equipment power).

PUE has gained a lot of traction as a data center efficiency metric since widespread reporting started on it around 2009. We can easily measure PUE by adding electrical meters to the lines powering the various parts of a data center, thus determining how much power is used by chillers and UPSs.

Historically, the PUE for the average data center has been embarrassingly poor. According to a 2006 study [MB06], 85% of data centers were estimated to have a PUE greater than 3.0. In other words, the building's mechanical and electrical systems consumed twice as much power as the actual computing load. Only 5% had a PUE of 2.0 or better.

A subsequent EPA survey of over 100 data centers reported an average PUE of 1.91 [PUE10]. A few years back, an Uptime Institute survey of over 1,100 data centers covering a range of geographies and sizes reported an average PUE value between 1.8 and 1.89 [UpI12, Hes14]. More recently, a 2016 report from LBNL noted PUEs of 1.13 for hyperscale data centers (warehouse-scale computers) and 1.6–2.35 for traditional data centers [She+16]. Figure 5.1 shows the distribution of results from one of these studies [UpI12]. Cold and hot aisle containment and increased cold aisle temperature are the most common improvements implemented. Large facilities reported the biggest improvements, and about half of small data centers (with less than 500 servers) still were not measuring PUE.

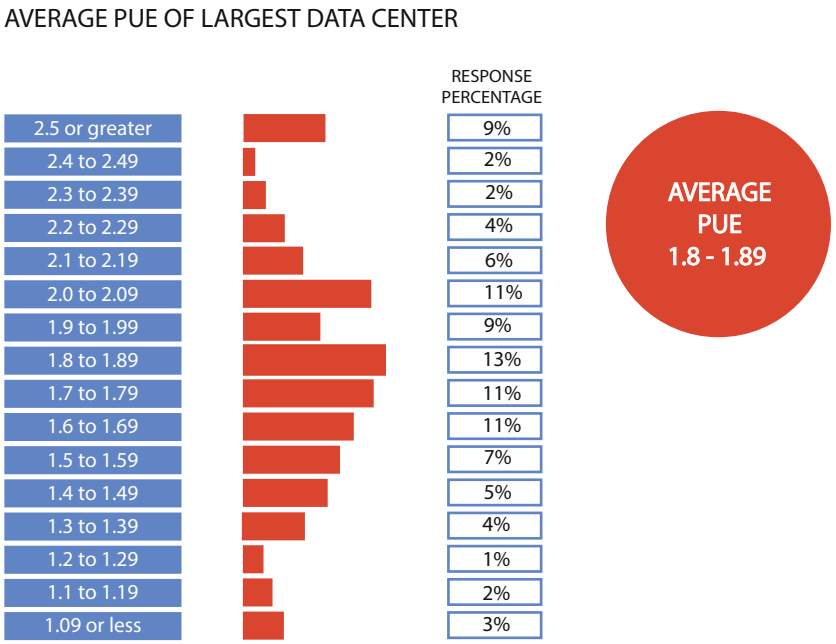

Figure 5.1: Uptime Institute survey of PUE for 1100+ data centers. This detailed data is based on a 2012 study [UpI12] but the trends are qualitatively similar to more recent studies (e.g., 2016 LBNL study [She+16]).

Very large operators (usually consumer internet companies like Google, Microsoft, Yahoo!, Facebook, Amazon, Alibaba, and eBay) have reported excellent PUE results over the past few years, typically below 1.2, although only Google has provided regular updates of its entire fleet based on a clearly defined metric (Figure 5.2) [GDCa]. At scale, it is easy to justify the importance of efficiency; for example, Google reported having saved over one billion dollars to date from energy efficiency measures [GGr].

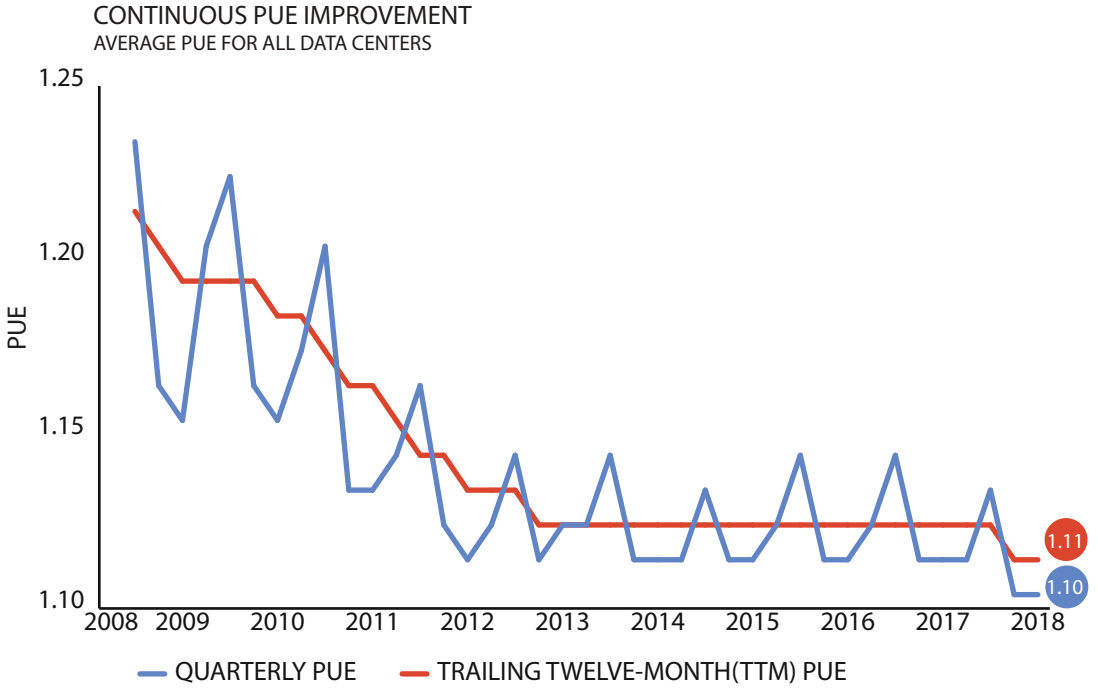

Figure 5.2: PUE data for all large-scale Google data centers over time [GDCa].

5.1.2 ISSUES WITH THE PUE METRIC

Although The Green Grid (TGG) publishes detailed guidelines on how to measure and report PUE [TGGd], many published values aren't directly comparable, and sometimes PUE values are used in marketing documents to show best-case values that aren't real. The biggest factors that can skew PUE values are as follows.

- Not all PUE measurements include the same overheads. For example, some may include losses in the primary substation transformers or in wires feeding racks from PDUs, whereas others may not. Google reported a fleet-wide PUE of 1.12 using a comprehensive definition of overhead that includes all known sources, but could have

reported a PUE of 1.06 with a more "optimistic" definition of overhead [GDCb]. For PUE to be a useful metric, data center owners and operators should adhere to Green Grid guidelines [TGGd] in measurements and reporting, and be transparent about the methods used in arriving at their results.

- Instantaneous PUEs differ from average PUEs. Over the course of a day or a year, a facility's PUE can vary considerably. For example, on a cold day it might be low, but during the summer it might be considerably higher. Generally speaking, annual averages are more useful for comparisons.

- Some PUEs aren't real-world measurements. Often vendors publish "design" PUEs that are computed using optimal operating conditions and nominal performance values, or they publish a value measured during a short load test under optimal conditions. Typically, PUE values provided without details fall into this category.

- Some PUE values have higher error bars because they're based on infrequent manual readings, or on coarsely placed meters that force some PUE terms to be estimated instead of measured. For example, if the facility has a single meter measuring the critical load downstream of the UPS, PDU, and low-voltage distribution losses will need to be estimated.

In practice, PUE values should be measured in real time. Not only does this provide a better picture of diurnal and seasonal variations, it also allows the operator to react to unusual readings during day-to-day operations. For example, someone may have left on a set of backup pumps after a periodic test. With real-time metrics the operations team can quickly correct such problems after comparing expected vs. actual PUE values.

The PUE metric has been criticized as not always indicating better energy performance, because PUEs typically worsen with decreasing load. For example, assume a data center's PUE is 2.0 at a 500 kW load vs. 1.5 at a 1 MW load. If it's possible to run the given workload with a 500 kW load (for example, with newer servers), that clearly is more energy efficient despite the inferior PUE. However, this criticism merely points out that PUE is just one of three factors in the efficiency equation shown earlier in this chapter, and overall the widespread adoption of PUE measurements has arguably been the driver of the biggest improvements in data center efficiency in the past 50 years.

5.1.3 SOURCES OF EFFICIENCY LOSSES IN DATA CENTERS

The section on data center power systems in Chapter 4 describes the efficient transformation of power as it approaches the data center floor. The first two transformation steps bring the incoming high-voltage power (110 kV and above) to medium-voltage distribution levels (typically less than

50 kV) and, closer to the server floor to low voltage (typically 480 V in North America). Both steps should be very efficient, with losses typically below half a percent for each step. Inside the building, conventional double-conversion UPSs cause the most electrical loss. In the first edition we listed efficiencies of 88–94% under optimal load, significantly less if partially loaded (which is the common case). Rotary UPSs (flywheels) and high-efficiency UPSs can reach efficiencies of about 97%. The final transformation step in the PDUs accounts for an additional half-percent loss. Finally, 1–3% of power can be lost in the cables feeding low-voltage power (110 or 220 V) to the racks (recall that a large facility can have a raised floor area that is over 100 m long or wide, so power cables can be quite long).

A significant portion of data center inefficiencies stems from cooling overhead, with chillers being the largest culprit. Cooling losses are three times greater than power losses, presenting the most promising target for efficiency improvements: if all cooling losses were eliminated, the PUE would drop to 1.26, whereas a zero-loss UPS system would yield a PUE of only 1.8. Typically, the worse a facility's PUE is, the higher the percentage of the total loss comes from the cooling system [BM06]. Intuitively, there are only so many ways to mishandle a power distribution system, but many more ways to mishandle cooling.

Conversely, there are many non-intuitive ways to improve the operation of the data center's cooling infrastructure. The energy for running the cooling infrastructure has a nonlinear relationship with many system parameters and environmental factors, such as the total system load, the total number of chillers operating, and the outside wind speed. Most people find it difficult to intuit the relationship between these variables and total cooling power. At the same time, a large amount of data is being collected regularly from a network of sensors used to operate the control loop for data center cooling. The existence of this large data set suggests that machine learning and artificial intelligence could be used to find additional PUE efficiencies [EG16].

Figure 5.3 shows the typical distribution of energy losses in a WSC data center. Much of this inefficiency is caused by a historical lack of attention to power loss, not by inherent limitations imposed by physics. Less than 10 years ago, PUEs weren't formally used and a total overhead of 20% was considered unthinkably low, yet as of 2018 Google reported a fleet-wide annual average overhead of 11% [GDCb] and many others are claiming similar values for their newest facilities. However, such excellent efficiency is still confined to a small set of data centers, and many small data centers probably haven't improved much.

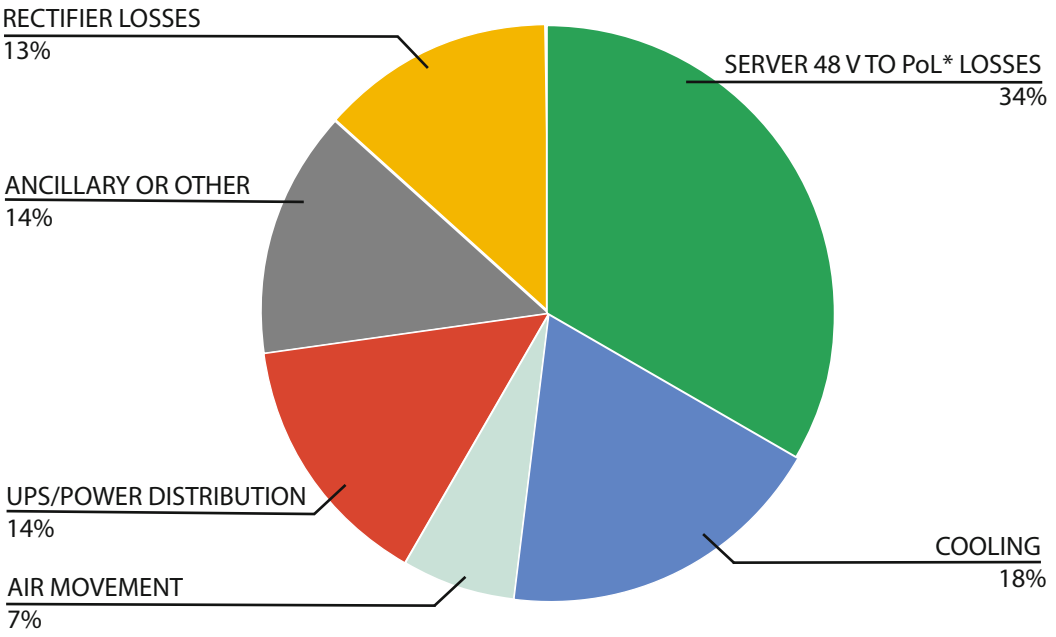

Figure 5.3: A representative end-to-end breakdown of energy losses in a typical datacenter. Note that this breakdown does not include losses of up to a few percent due to server fans or electrical resistance on server boards.

5.1.4 IMPROVING THE ENERGY EFFICIENCY OF DATA CENTERS

As discussed in the previous chapter, careful design for efficiency can substantially improve PUE [Nel+, PGE, GMT06]. To summarize, the key steps are as follows.

- *Careful air flow handling*: Isolate hot air exhausted by servers from cold air, and keep the path to the cooling coil short so that little energy is spent moving cold or hot air long distances.
- *Elevated temperatures*: Keep the cold aisle at 25–30°C rather than 18–20°C. Higher temperatures make it much easier to cool data centers efficiently. Virtually no server or network equipment actually needs intake temperatures of 20°C, and there is no evidence that higher temperatures cause more component failures [PWB07, SPW09, ES+].
- *Free cooling*: In most moderate climates, free cooling can eliminate the majority of chiller runtime or eliminate chillers altogether.

- *Better power system architecture*: UPS and power distribution losses can often be greatly reduced by selecting higher-efficiency gear, as discussed in the previous chapter.
- *Machine learning*: Apply novel machine learning techniques to discover non-intuitive techniques for controlling data center infrastructure to further reduce cooling requirements. Large amounts of data are being collected by many sensors in the data center, making this problem a natural fit for machine learning.

In April 2009, Google first published details on its data center architecture, including a video tour of a container-based data center built in 2005 [GInc09]. In 2008, this data center achieved a state-of-the-art annual PUE of 1.24, yet differed from conventional data centers only in the application of the principles listed above. Today, large-scale data centers commonly feature PUEs below 1.2, especially those belonging to cloud operators. Even in unfavorable climates, today's PUEs are lower than the state-of-the-art PUEs in 2008 For example, Google's data center in Singapore, where the average monthly temperature rarely falls below 25°C, the annual PUE is 1.18.

5.1.5 BEYOND THE FACILITY

Recall the energy efficiency formula from the beginning of this chapter:

$$\text{Efficiency} = \frac{\text{Computation}}{\text{Total Energy}} = \underbrace{\left(\frac{1}{\text{PUE}}\right)}_{\text{(a)}} \times \underbrace{\left(\frac{1}{\text{SPUE}}\right)}_{\text{(b)}} \times \underbrace{\left(\frac{\text{Computation}}{\text{Total Energy to Electronic Components}}\right)}_{\text{(c)}}.$$

So far we've discussed the first term, facility overhead. The second term (b) accounts for overheads inside servers or other IT equipment using a metric analogous to PUE: server PUE (SPUE). SPUE consists of the ratio of total server input power to its useful power, where useful power includes only the power consumed by the electronic components directly involved in the computation: motherboard, disks, CPUs, DRAM, I/O cards, and so on. Substantial amounts of power may be lost in the server's power supply, voltage regulator modules (VRMs), and cooling fans. As discussed in Chapter 4, the losses inside the server can exceed those of the entire upstream data center power train.

SPUE measurements aren't standardized like PUE but are fairly straightforward to define. Almost all equipment contains two transformation steps: the first step transforms input voltage (typically 110–220 VAC) to local DC current (typically 12 V), and in the second step VRMs transform that down to much lower voltages used by a CPU or DRAM. (The first step requires an additional internal conversion within the power supply, typically to 380 VDC.) SPUE ratios of 1.6–1.8 were common a decade ago; many server power supplies were less than 80% efficient, and many motherboards used VRMs that were similarly inefficient, losing more than 25% of input power in electrical conversion losses. In contrast, commercially available AC-input power supplies

today achieve 94% efficiency, and VRMs achieve 96% efficiency (see Chapter 4). Thus, a state-of-the-art SPUE is 1.11 or less [Cli]. For example, instead of the typical 12 VDC voltage, Google uses 48 VDC voltage rack distribution system, which reduces energy losses by over 30%.

The product of PUE and SPUE constitutes an accurate assessment of the end-to-end electromechanical efficiency of a WSC. A decade ago the true (or total) PUE metric (TPUE), defined as PUE * SPUE, stood at more than 3.2 for the average data center; that is, for every productive watt, at least another 2.2 W were consumed. By contrast, a modern facility with an average PUE of 1.11 as well as an average SPUE of 1.11 achieves a TPUE of 1.23. Close attention to cooling and power system design in combination with new technology has provided an order of magnitude reduction in overhead power consumption.

5.2 THE ENERGY EFFICIENCY OF COMPUTING

So far we have discussed efficiency in electromechanical terms, the (a) and (b) terms of the efficiency equation, and largely ignored term (c), which accounts for how the electricity delivered to electronic components is actually translated into useful work. In a state-of-the-art facility, the electromechanical components have a limited potential for improvement: Google's TPUE of approximately 1.23 means that even if we eliminated all electromechanical overheads, the total energy efficiency would improve by only 19%. In contrast, the energy efficiency of computing has doubled approximately every 1.5 years in the last half century [Koo+11]. Although such rates have declined due to CMOS scaling challenges [FM11], they are still able to outpace any electromechanical efficiency improvements. In the remainder of this chapter we focus on the energy and power efficiency of computing.

5.2.1 MEASURING ENERGY EFFICIENCY

Ultimately, we want to measure the energy consumed to produce a certain result. A number of industry benchmarks try to do exactly that. In high-performance computing (HPC), the Green 500 [TG500] benchmark ranks the energy efficiency of the world's top supercomputers using LINPACK. Similarly, server-level benchmarks such as Joulesort [Riv+07] and SPECpower [SPEC] characterize other aspects of computing efficiency. Joulesort measures the total system energy to perform an out-of-core sort and derives a metric that enables the comparison of systems ranging from embedded devices to supercomputers. SPECpower focuses on server-class systems and computes the performance-to-power ratio of a system running a typical business application on an enterprise Java platform. Two separate benchmarking efforts aim to characterize the efficiency of storage systems: the Emerald Program [SNI11] by the Storage Networking Industry Association (SNIA) and the SPC-2/E [SPC12] by the Storage Performance Council. Both benchmarks measure storage servers under different kinds of request activity and report ratios of transaction throughput per watt.

5.2.2 SERVER ENERGY EFFICIENCY

Clearly, the same application binary can consume different amounts of power depending on the server's architecture and, similarly, an application can consume more or less of a server's capacity depending on software performance tuning. Furthemore, systems efficiency can vary with utilization: under low levels of utilization, computing systems tend to be significantly more inefficient than when they are exercised at maximum utilization.

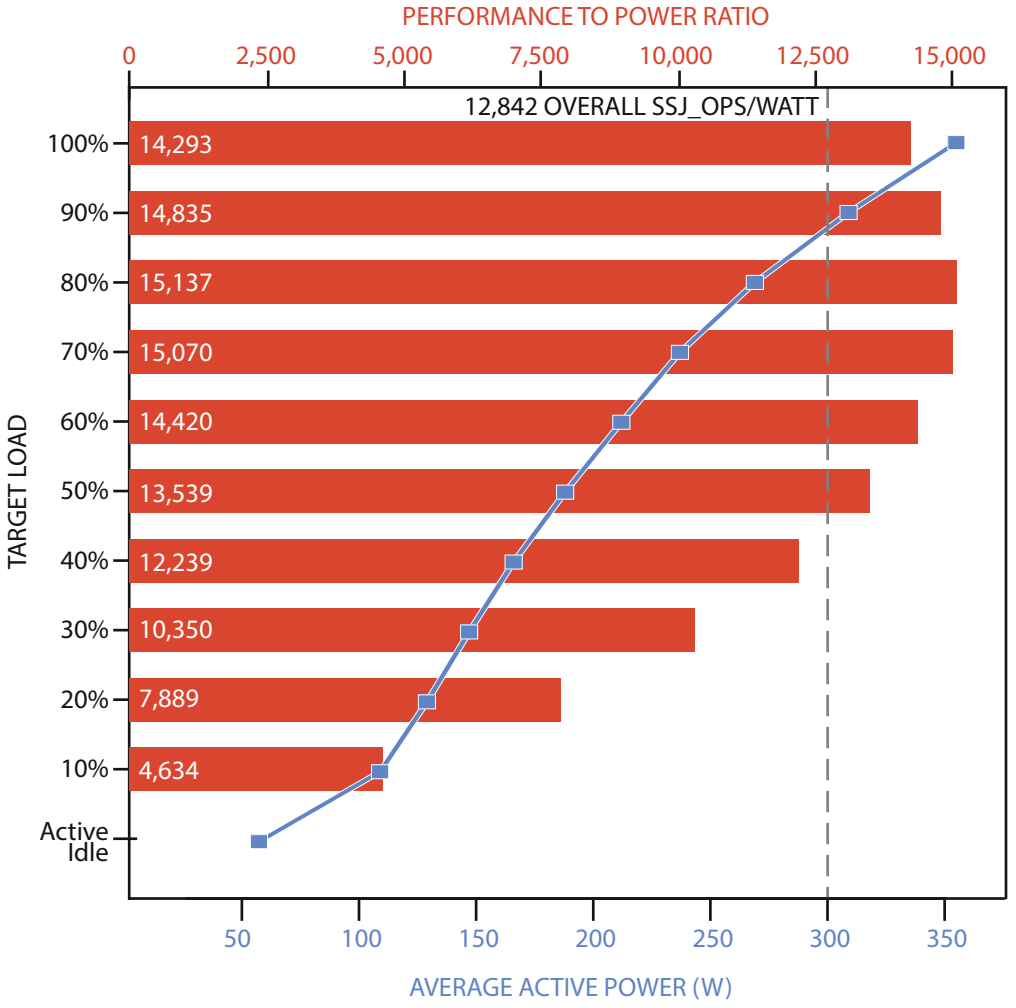

Figure 5.4: Example benchmark result for SPECpower_ssj2008; bars indicate energy efficiency and the line indicates power consumption. Both are plotted for a range of utilization levels, with the average energy efficiency metric corresponding to the vertical dark line. The system has two 2.1 GHz 28-core Intel Xeon processors, 192 GB of DRAM, and one M.2 SATA SSD.

Figure 5.4 shows the SPECpower benchmark results for the top performing entry as of January 2018 under varying utilization. The results show two metrics: performance- (transactions per second) to-power ratio and the average system power, plotted over 11 load levels. One feature in the figure is noteworthy and common to all other SPECpower benchmark results: the performance-to-power ratio drops appreciably as the target load decreases because the system power decreases much more slowly than does performance. Note, for example, that the energy efficiency at 30% load has 30% lower efficiency than at 100%. Moreover, when the system is idle, it is still consuming just under 60 W, which is 16% of the peak power consumption of the server.

5.2.3 USAGE PROFILE OF WAREHOUSE-SCALE COMPUTERS

Figure 5.5 shows the average CPU utilization of two Google clusters during a representative three-month period (measured between January and March 2013); each cluster has over 20,000 servers. The cluster on the right (b) represents one of Google's most highly utilized WSCs, where large continuous batch workloads run. WSCs of this class can be scheduled very efficiently and reach very high utilizations on average. The cluster on the left (a) is more representative of a typical shared WSC, which mixes several types of workloads and includes online services. Such WSCs tend to have relatively low average utilization, spending most of their time in the 10–50% CPU utilization range. This activity profile turns out to be a perfect mismatch with the energy efficiency profile of modern servers in that they spend most of their time in the load region where they are most inefficient.

Another feature of the energy usage profile of WSCs is not shown in Figure 5.5: individual servers in these systems also spend little time idle. Consider, for example, a large web search workload, such as the one described in Chapter 2, where queries are sent to a large number of servers, each of which searches within its local slice of the entire index. When search traffic is high, all servers are being heavily used, but during periods of low traffic, a server might still see hundreds of queries per second, meaning that idle periods are likely to be no longer than a few milliseconds.

The absence of significant idle intervals in general-purpose WSCs, despite the existence of low activity periods, is largely a result of applying sound design principles to high-performance, robust distributed systems software. Large-scale internet services rely on efficient load distribution to a large number of servers, creating a situation such that when load is lighter, we tend to have a lower load in multiple servers instead of concentrating the load in fewer servers and idling the remaining ones. Idleness can be manufactured by the application (or an underlying cluster management system) by migrating workloads and their corresponding state to fewer machines during periods of low activity. This can be relatively easy to accomplish when using simple replication models, when servers are mostly stateless (that is, serving data that resides on a shared NAS or SAN storage system). However, it comes at a cost in terms of software complexity and

energy for more complex data distribution models or those with significant state and aggressive exploitation of data locality.

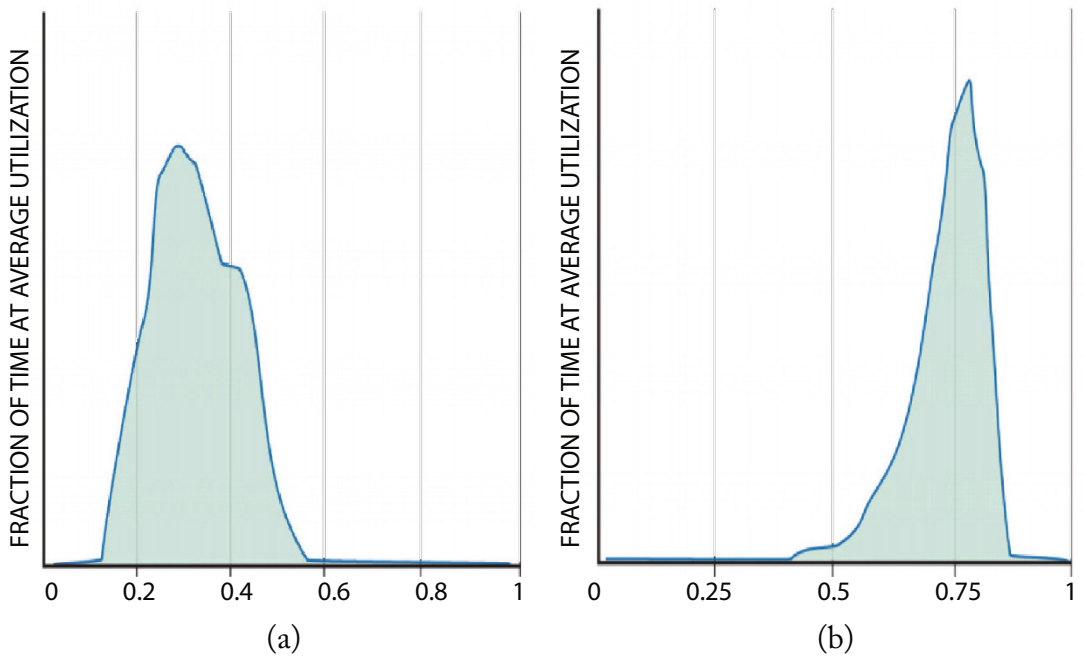

Figure 5.5: Average activity distribution of a sample of 2 Google clusters, each containing over 20,000 servers, over a period of 3 months.

Another reason why it may be difficult to manufacture useful idle periods in large-scale distributed systems is the need for resilient distributed storage. GFS [GGL03] achieves higher resilience by distributing data chunk replicas for a given file across an entire cluster instead of concentrating them within only a small number of machines. This benefits file system performance by achieving fine granularity load balancing, as well as resiliency, because when a storage server crashes (or a disk fails), the replicas in that system can be reconstructed by thousands of machines, making recovery extremely efficient. The consequence of otherwise sound designs is that low traffic levels translate into lower activity for all machines instead of full idleness for a significant subset of them. Several practical considerations may also work against full idleness, as networked servers frequently perform many small background tasks on periodic intervals. The reports on the Tickless kernel project [SPV07] provide other examples of how difficult it is to create and maintain idleness.

5.3 ENERGY-PROPORTIONAL COMPUTING

In an earlier article [BH07], we argued that the mismatch between server workload profile and server energy efficiency behavior must be addressed largely at the hardware level; software alone cannot efficiently exploit hardware systems that are efficient only when they are in inactive idle modes (sleep or standby) or when running at full speed. We believe that systems are inefficient when lightly used largely because of lack of awareness by engineers and researchers about the importance of that region to energy efficiency.

We suggest that *energy proportionality* should be added as a design goal for computing components. Ideally, energy-proportional systems will consume almost no power when idle (particularly in active idle states where they are still available to do work) and gradually consume more power as the activity level increases. A simple way to reason about this ideal curve is to assume linearity between activity and power usage, with no constant factors. Such a linear relationship would make energy efficiency uniform across the activity range, instead of decaying with decreases in activity levels. Note, however, that linearity is not necessarily the optimal relationship for energy savings. As shown in Figure 5.5(a), since servers spend relatively little time at high activity levels, it might be fine to decrease efficiency at high utilizations, particularly when approaching maximum utilization. However, doing so would increase the maximum power draw of the equipment, thus increasing facility costs.

Figure 5.6 illustrates the possible energy efficiency of two hypothetical systems that are more energy-proportional than typical servers. The curves in red correspond to a typical server, circa 2009. The green curves show the normalized power usage and energy efficiency of a more energy-proportional system, which idles at only 10% of peak power and with linear power vs. load behavior. Note how its efficiency curve is far superior to the one for the typical server; although its efficiency still decreases with the load level, it does so much less abruptly and remains at relatively high efficiency levels at 30% of peak load. The curves in blue show a system that also idles at 10% of peak but with a sublinear power versus load relationship in the region of load levels between 0% and 50% of peak load. This system has an efficiency curve that peaks not at 100% load, but around the 30–40% region. From an energy usage standpoint, such behavior would be a good match to the kind of activity spectrum for WSCs depicted in Figure 5.5(a).

The potential gains from energy proportionality in WSCs were evaluated by Fan et al. [FWB07] in their power provisioning study. They used traces of activity levels of thousands of machines over six months to simulate the energy savings gained from using more energy-proportional servers—servers with idle consumption at 10% of peak (similar to the green curves in Figure 5.6) instead of at 50% (such as the corresponding red curve). Their models suggest that energy usage would be halved through increased energy proportionality alone because the two servers compared had the same peak energy efficiency.

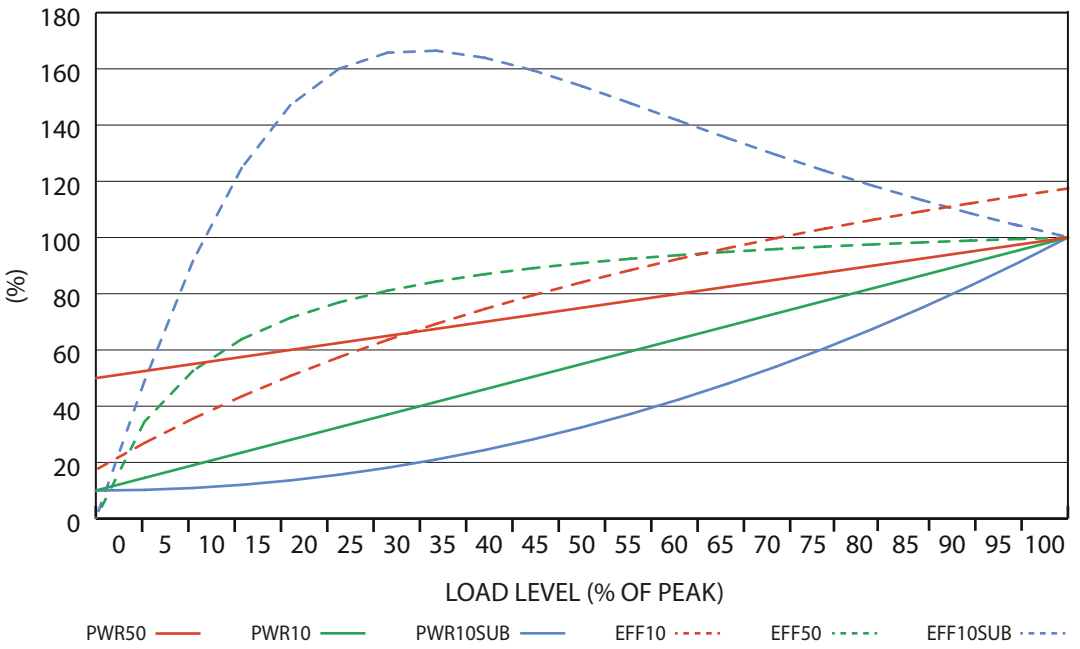

Figure 5.6: Power and corresponding power efficiency of three hypothetical systems: a typical server with idle power at 50% of peak (`Pwr50` and `Eff50`), a more energy-proportional server with idle power at 10% of peak (`Pwr10` and `Eff10`), and a sublinearly energy-proportional server with idle power at 10% of peak (`Pwr10sub` and `Eff10sub`). The solid lines represent power % (normalized to peak power). The dashed lines represent efficiency as a percentage of power efficiency at peak.

5.3.1 CAUSES OF POOR ENERGY PROPORTIONALITY

Although CPUs historically have a bad reputation regarding energy usage, they are not necessarily the only culprit for poor energy proportionality. Over the last few years, CPU designers have paid more attention to energy efficiency than their counterparts for other subsystems. The switch to multicore architectures instead of continuing to push for higher clock frequencies and larger levels of speculative execution is one of the reasons for this more power-efficient trend.

The relative contribution of the memory system to overall energy use has decreased over the last five years with respect to CPU energy use, reversing a trend of higher DRAM energy profile throughout the previous decade. The decrease in the fraction of energy used in memory systems is due to a combination of factors: newer DDR3 technology is substantially more efficient than previous technology (FBDIMMs). DRAM chip voltage levels have dropped from 1.8 V to below 1.5 V, new CPU chips use more energy as more aggressive binning processes and temperature-controlled "turbo" modes allow CPUs to run closer to their thermal envelope, and today's systems have

a higher ratio of CPU performance per DRAM space (a result of DRAM technology scaling falling behind that of CPUs).

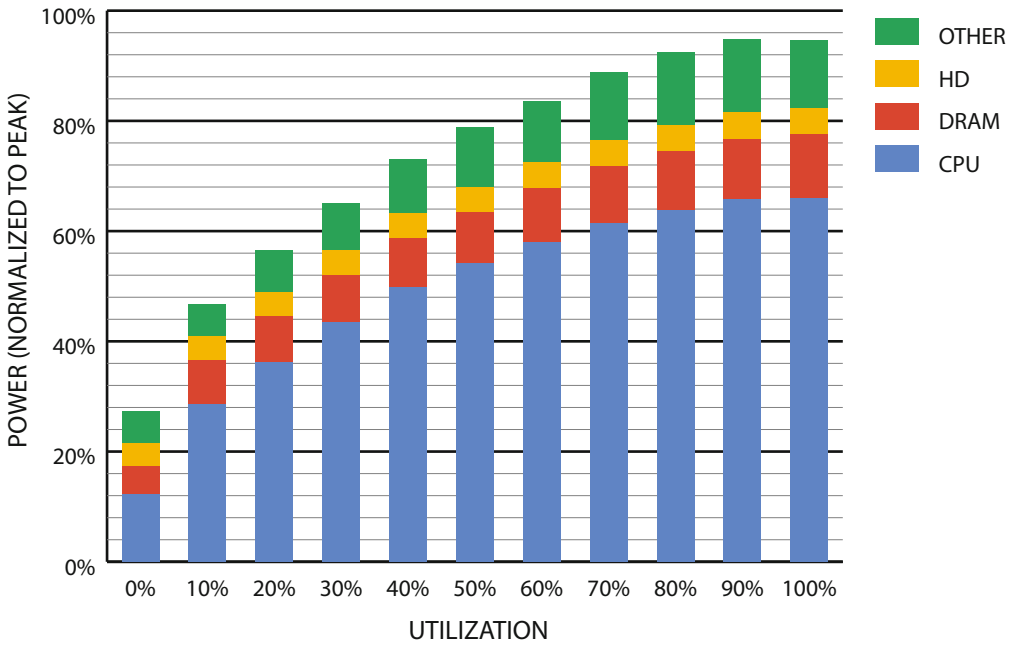

Figure 5.7: Subsystem power usage in an x86 server as the compute load varies from idle to full usage.

Figure 5.7 shows the power usage of the main subsystems for a Google server (circa 2012) as the compute load varies from idle to full activity levels. Unlike what we reported in the first edition, the CPU portion (everything inside a CPU socket) is once more the dominant energy consumer in servers, using two-thirds of the energy at peak utilization and about 40% when (active) idle. In our experience, server-class CPUs have a dynamic power range that is generally greater than 3.0x (more than 3.5x in this case), whereas CPUs targeted at the embedded or mobile markets can do even better. By comparison, the dynamic range of memory systems, disk drives, and networking equipment is much lower: approximately 2.0x for memory, 1.3x for disks, and less than 1.2x for networking switches. This suggests that energy proportionality at the system level cannot be achieved through CPU optimizations alone, but instead requires improvements across all components. Networking and memory are both notable here. Future higher bandwidth memory systems are likely to increase the power of the memory subsystems. Also, given switch radix scaling challenges, the ratio of switches to servers is likely to increase, making networking power more important. Nevertheless, as we'll see later, increased CPU energy proportionality over the last five years, and an increase in the fraction of overall energy use by the CPU, has resulted in more energy proportional servers today.

5.3.2 IMPROVING ENERGY PROPORTIONALITY

Added focus on energy proportionality as a figure of merit in the past five years has resulted in notable improvements for server-class platforms. A meaningful metric of the energy proportionality of a server for a WSC is the ratio between the energy efficiency at 30% and 100% utilizations. A perfectly proportional system will be as efficient at 30% as it is at 100%. In the first edition (in early 2009), that ratio for the top 10 SPECpower results was approximately 0.45, meaning that when used in WSCs, those servers exhibited less than half of their peak efficiency. As of June 2018, that figure has improved almost twofold, reaching 0.80.

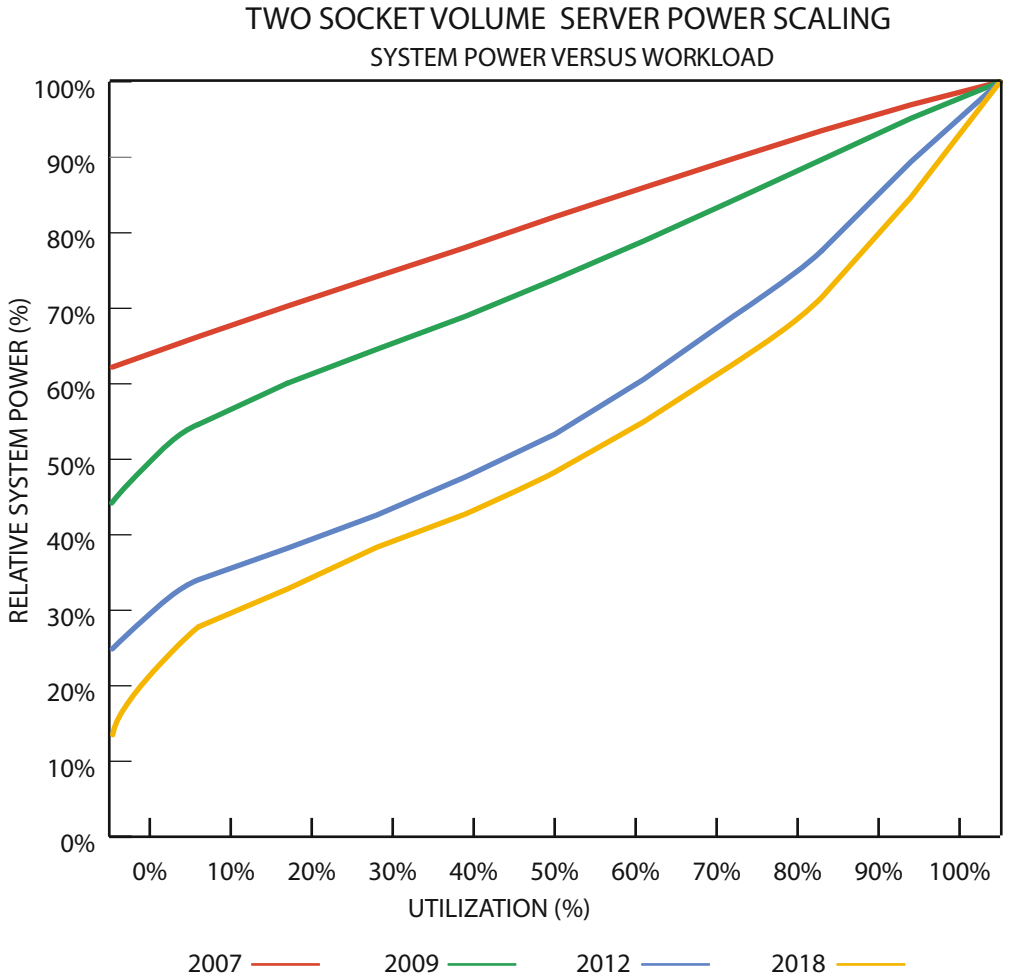

Figure 5.8: Normalized system power vs. utilization in Intel servers from 2007–2018 (courtesy of David Lo, Google). The chart indicates that Intel servers have become more energy proportional in the 12-year period.

Figure 5.8 shows increasing proportionality in Intel reference platforms between 2007 and 2018 [Sou12]. While not yet perfectly proportional, the more recent systems are dramatically more energy proportional than their predecessors.

5.3.3 ENERGY PROPORTIONALITY IN THE REST OF THE SYSTEM

While processor energy proportionality has improved, greater effort is still required for DRAM, storage, and networking. Disk drives, for example, spend a large fraction of their energy budget (as much as 70% of their total power for high RPM drives) simply keeping the platters spinning. Improving energy efficiency and proportionality may require lower rotational speeds, smaller platters, or designs that use multiple independent head assemblies. Carrera et al. [CPB03] considered the energy impact of multi-speed drives and combinations of server-class and laptop drives to achieve proportional energy behavior. Sankar et al. [SGS08] explored different architectures for disk drives, observing that because head movements are relatively energy-proportional, a disk with lower rotational speed and multiple heads might achieve similar performance and lower power when compared with a single-head, high RPM drive.

Traditionally, data center networking equipment has exhibited rather poor energy proportionality. At Google we have measured switches that show little variability in energy consumption between idle and full utilization modes. Historically, servers didn't need much network bandwidth, and switches were expensive, so their overall energy footprint was relatively small (in the single digit percentages of total IT power). However, as switches become more commoditized and bandwidth needs increase, networking equipment could become responsible for 10–20% of the facility energy budget. At that point, their lack of proportionality will be a severe problem. To illustrate this point, let's assume a system that exhibits a linear power usage profile as a function of utilization (u):

$$P(u) = Pi + u(1\text{-}Pi).$$

In the equation above, Pi represents the system's idle power, and peak power is normalized to 1.0. In such a system, energy efficiency can be estimated as u/P(u), which reduces to the familiar Amdahl Law formulation below:

$$E(u) = \frac{1}{1 - Pi + Pi/u}.$$

Unlike the original Amdahl formula, we are not interested in very high values of u, since it can only reach 1.0. Instead, we are interested in values of utilization between 0.1 and 0.5. In that case, high values for Pi (low energy proportionality) will result in low efficiency. If every subcomponent of a WSC is highly energy proportional except for one (say, networking or storage), that subcomponent will limit the whole system efficiency similarly to how the amount of serial work limits parallel speedup in Amdahl's formula.

Efficiency and proportionality of data center networks might improve in a few ways. Abts et al. [Abt+10] describe how modern plesiochronous links can be modulated to adapt to usage as well as how topology changes and dynamic routing can create more proportional fabrics. The IEEE's Energy-Efficient Ethernet standardization effort [Chr+10], (802.3az), is also trying to pursue interoperable mechanisms that allow link-level adaptation.

Finally, energy-proportional behavior is not only a target for electronic components but for the entire WSC system, including power distribution and cooling infrastructure.

5.3.4 RELATIVE EFFECTIVENESS OF LOW-POWER MODES

As discussed earlier, long idleness intervals would make it possible to achieve higher energy proportionality by using various kinds of sleep modes. We call these low-power modes *inactive* because the devices are not usable while in those modes, and typically a sizable latency and energy penalty is incurred when load is reapplied. Inactive low-power modes were originally developed for mobile and embedded devices, and they are very successful in that domain. However, most of those techniques are a poor fit for WSC systems, which would pay an inactive-to-active latency and energy penalty too frequently. The few techniques that can be successful in this domain have very low wake-up latencies, as is beginning to be the case with CPU low-power halt states (such as the ACPI C1E state).

Unfortunately, these tend to be the low-power modes with the smallest degrees of energy savings. Large energy savings are available from inactive low-power modes such as spun-down disk drives. A spun-down disk might use almost no energy, but a transition to active mode incurs a latency penalty 1,000 times higher than a regular access. Spinning up the disk platters adds an even larger energy penalty. Such a huge activation penalty restricts spin-down modes to situations in which the device will be idle for several minutes, which rarely occurs in servers.

Active low-power modes save energy at a performance cost while not requiring inactivity. CPU voltage-frequency scaling is an example of an active low-power mode because it remains able to execute instructions albeit at a slower rate. The (presently unavailable) ability to read and write to disk drives at lower rotational speeds is another example of this class of low-power modes. In contrast with inactive modes, active modes are useful even when the latency and energy penalties to transition to a high-performance mode are significant. Because active modes are operational, systems can remain in low-energy states for as long as they remain below certain load thresholds. Given that periods of low activity are more common and longer than periods of full idleness, the overheads of transitioning between active energy savings modes amortize more effectively.

The use of very low-power inactive modes with high-frequency transitions has been proposed by Meisner et al. [MGW09] and Gandhi et al. [Gan+] as a way to achieve energy proportionality. The systems proposed, PowerNap and IdleCap, assume that subcomponents have no useful low power modes other than full idleness and modulate active-to-idle transitions in all subcomponents

in order to reduce power at lower utilizations while limiting the impact on performance. The promise of such approaches hinges on system-wide availability of very low power idle modes with very short active-to-idle and idle-to-active transition times, a feature that seems within reach for processors but more difficult to accomplish for other system components. In fact, Meisner et al. [Mei+11] analyze the behavior of online data intensive workloads (such as web search) and conclude that existing low power modes are insufficient to yield both energy proportionality and low latency.

5.3.5 THE ROLE OF SOFTWARE IN ENERGY PROPORTIONALITY

We have argued that hardware components must undergo significant improvements in energy proportionality to enable more energy-efficient WSC systems. However, more intelligent power management and scheduling software infrastructure plays an important role too. For some component types, achieving perfect energy-proportional behavior may not be a realizable goal. Designers will have to implement software strategies for intelligent use of power management features in existing hardware, using low-overhead inactive or active low-power modes, as well as implementing power-friendly scheduling of tasks to enhance energy proportionality of hardware systems. For example, if the activation penalties in inactive low-power modes can be made small enough, techniques like PowerNap (Meisner et al. [MGW09]) could be used to achieve energy-proportional behavior with components that support only inactive low-power modes.

This software layer must overcome two key challenges: encapsulation and performance robustness. Energy-aware mechanisms must be encapsulated in lower-level modules to minimize exposing additional infrastructure complexity to application developers; after all, WSC application developers already deal with unprecedented scale and platform-level complexity. In large-scale systems, completion of an end-user task also tends to depend on large numbers of systems performing at adequate levels. If individual servers begin to exhibit excessive response time variability as a result of mechanisms for power management, the potential for service-level impact is fairly high and can lead to the service requiring additional machine resources, resulting in minimal net improvements.

Incorporating end-to-end metrics and service level objective (SLO) targets from WSC applications into power-saving decisions can greatly help overcome performance variability challenges while moving the needle toward energy proportionality. During periods of low utilization, latency slack exists between the (higher latency) SLO targets and the currently achieved latency. This slack represents power saving opportunities, as the application is running faster than needed. Having end-to-end performance metrics is a critical piece needed to safely reduce the performance of the WSC in response to lower loads. Lo et al. [Lo+14] propose and study a system (PEGASUS) that combines hardware power actuation mechanisms (Intel RAPL [Intel18]) with software control policies. The system uses end-to-end latency metrics to drive decisions on when to adjust CPU power in response to load shifts. By combining application-level metrics with fine-grained

hardware actuation mechanisms, the system is able to make overall server power more energy proportional while respecting the latency SLOs of the WSC application.

Software plays an important role in improving cluster-level energy efficiency despite poor energy proportionality of underlying servers. By increasing the utilization of each individual server, cluster management software can avoid operating servers in the region of poor energy efficiency at low loads. Cluster scheduling software such as Borg [Ver+15] and Mesos [Hin+11] take advantage of resource sharing to significantly improve machine-level utilization through better bin-packing of disparate jobs (encapsulation). This is a net win for energy efficiency, where the poor energy proportionality of the servers that make up a WSC is mitigated by running the server at higher utilizations closer to its most energy efficient operating point. An even larger benefit of higher utilization is that the number of servers needed to serve a given capacity requirement is reduced, which lowers the TCO dramatically due to a significant portion of the cost of a WSC being in concentrated in the CapEx costs of the hardware.

However, as server utilization is pushed higher and higher, performance degradation from shared resource contention becomes a bigger and bigger issue. For example, if two workloads that would each completely saturate DRAM bandwidth are co-located on the same server, then both workloads will suffer significantly degraded performance compared to when each workload is run in isolation. With workload agnostic scheduling, the probability of this scenario occurring increases as server capacity increases with the scaling of CPU core counts. To counter the effects of interference, service owners tend to increase the resource requirements of sensitive workloads in order to ensure that their jobs will have sufficient compute capacity in the face of resource contention. This extra padding has an effect of lowering server utilization, thus also negatively impacting energy efficiency. To avoid this pitfall and to further raise utilization, contention aware scheduling needs to be utilized. Systems such as Bubble-Up [Mar+11], Heracles [Lo+15], and Quasar [DK14] achieve significantly higher server utilizations while maintaining strict application-level performance performance requirements. While the specific mechanisms differ for each system, they all share a common trait of using performance metrics in making scheduling and resource allocation decisions to provide both encapsulation and performance robustness for workloads running in the WSC. By overcoming these key challenges, such performance-aware systems can lead to significantly more resource sharing opportunities, increased machine utilization, and ultimately energy efficient WSCs that can sidestep poor energy proportionality.

Raghavendra et al. [Rag+08] studied a five-level coordinated power management scheme, considering per-server average power consumption, power capping at the server, enclosure, and group levels, as well as employing a virtual machine controller (VMC) to reduce the average power consumed across a collection of machines by consolidating workloads and turning off unused machines. Such intensive power management poses nontrivial control problems. For one, applications may become unstable if some servers unpredictably slow down due to power capping. On the

implementation side, power capping decisions may have to be implemented within milliseconds to avoid tripping a breaker. In contrast, overtaxing the cooling system may result in "only" a temporary thermal excursion, which may not interrupt the performance of the WSC. Nevertheless, as individual servers consume more power with a larger dynamic range due to improving energy proportionality in hardware, power capping becomes more attractive as a means of fully realizing the compute capabilities of a WSC.

Wu et al. [Wu+16a] proposed and studied the use of Dynamo, a dynamic power capping system in production at Facebook. Dynamo makes coordinated power decisions across the entire data center to safely oversubscribe power and improve power utilization. Using Intel RAPL as the node-level enforcement mechanism to cap machine power, the system is workload-aware to ensure that high priority latency-sensitive workloads are throttled only as a measure of last resort. As a result of deploying Dynamo, the authors note a significant boost in power capacity utilization at their data centers through increased use of dynamic core frequency boosting; namely, Intel Turbo Boost [IntTu], which can run CPU cores at higher frequencies given sufficient electrical and thermal headroom. Much like PEGASUS, Dynamo combines application-specific knowledge with fine-grained hardware knobs to improve the realizable compute capability of the WSC while respecting application performance boundaries.

Technologies such as Turbo Boost reflect a growing trend in CPU design of adding additional dynamic dimensions (CPU activity) to trade off power and performance. The behavior of Turbo Boost is highly dependent on the number of active cores and the compute intensity of the workload. For example, CPU core frequency can vary by as much as 85% on Intel Skylake server CPUs [IntXe]. Another manifestation of this phenomenon takes the form of wider vector instructions, such as AVX-512, which can cause large drops in CPU core frequency due to its usage. On the one hand, these techniques enable higher peak performance, but on the other hand, they increase performance variability across the WSC. Dynamic frequency scaling decisions made in hardware present a set of new challenges in achieving performance robustness, and software designers must be cognizant of such effects in hardware and to handle the resulting performance variation.

5.4 ENERGY EFFICIENCY THROUGH SPECIALIZATION

So far we have assumed traditional WSCs: collections of servers, each with CPUs, DRAM, networking, and disks; all computation handled by general purpose CPUs. However, recapping the discussion in Chapter 4, Dennard scaling has now ended (due to fundamental device limitations that prevent operating voltage from further being scaled due to static leakage concerns), and Moore's Law is well on its way to being sunset (as chip manufacturers struggle with maintaining high yield while further shrinking transistor sizes). Looking forward, general purpose CPUs are facing a

daunting task when it comes to further energy efficiency improvements. This issue is orthogonal to energy proportionality, as it is about improving energy efficiency at peak compute load.

While general-purpose CPUs improve marginally over time when it comes to energy efficiency improvements at peak load, the demand for compute is growing at a steady rate. Currently, this demand is being driven by technologies powered by artificial intelligence and machine learning, which require extraordinary amounts of compute commensurate with large model sizes and gargantuan amounts of data being fed into such workloads. While general-purpose CPUs are fully capable of performing the operations necessary for artificial intelligence, they are not optimized to run these kinds of workloads.

Specialized accelerators are designed for running one particular class of workloads well. The hardware for these accelerators can be general purpose graphics processing units (GPGPUs), field programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs), to name a few. Unlike general-purpose CPUs, specialized accelerators are incapable of running all kinds of workloads with reasonable efficiency. That is because these accelerators trade off general compute capabilities for the ability to run a subset of workloads with phenomenal performance and efficiency. High-performance server class CPUs are designed to extract the maximum performance out of challenging workloads with a wide variety of potential kinds of computations, unpredictable control flows, irregular to non-existent parallelism, and complicated data dependencies. On the other hand, specialized accelerators need to perform well only for a specific kind of computation that provides opportunities for domain-specific optimizations.

For example, consider Google's Tensor Processing Unit (TPU) [Jou+17]. This custom ASIC was designed to handle the inference portion of several machine learning workloads. The energy efficiency of the TPU benefited greatly from the specialization of compute. The TPU is powered by a systolic array, an energy efficient construct that excels at performing regular computations, such as matrix multiplication. The use of a systolic array allows the TPU to avoid a high access rate to large SRAM arrays that would otherwise consume significant amounts of power. In addition, compared to a modern superscalar out-of-order CPU, the control logic for a TPU is relatively simple and thus much more energy efficient. Since parallelism in machine learning applications is easier to extract, the TPU has no need for the complicated and energy hungry control hardware found in CPUs. These and other design decisions for the TPU unlocked a vast improvement in energy efficiency.

Figure 5.9 shows how the TPU is orders of magnitude more energy efficient for inference tasks compared to a contemporary server CPU of its time (Intel Haswell). However, while the energy efficiency is high, the energy proportionality of the TPU happens to be much worse than that of the CPU, as it consumes 88% of peak power at 10% load. The designers note that the poor energy proportionality is not due to fundamental reasons but to tradeoffs around design expediency. Nevertheless, there is an open opportunity to apply the same learnings from general compute, such as improved energy proportionality, to specialized accelerators as well.

Specialized accelerators have an important role to play in improving the energy efficiency of WSCs of the future. Large emerging workloads such as machine learning are ripe targets for acceleration due to the sheer volume of compute they demand. The challenge is to identify workloads that benefit from being implemented on specialized accelerators and to progress from concept to product in a relatively short timespan. In addition, the same insights from improving energy efficiency of servers (such as energy proportionality) also apply to accelerators. Nevertheless, not all workloads can be put on specialized compute hardware. These can be due to the nature of the workload itself (general-purpose CPUs can be viewed as accelerators for complex, branchy, and irregular code) or due to it not having enough deployment volume to justify the investment in specialized hardware. Thus, it is still important to improve the overall energy efficiency of the entire data center, general-purpose servers included, in conjunction with improving energy efficiency of accelerators.

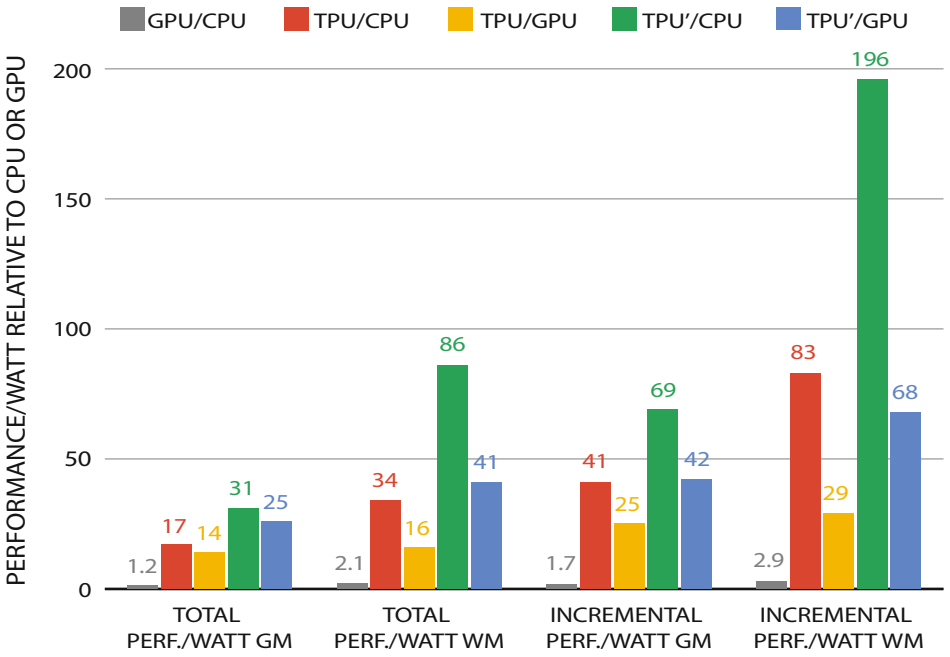

Figure 5.9: Relative performance/watt (TDP) of GPU server (blue bar) and TPU server (red bar) to CPU server, and TPU server to GPU server (orange bar). TPU' is an improved TPU. The green bar shows its ratio to the CPU server and the blue bar shows its relation to the GPU server. Total includes host server power, but incremental doesn't. GM and WM are the geometric and weighted means [Jou+17].

5.5 DATA CENTER POWER PROVISIONING

Energy efficiency optimizations reduce electricity costs. In addition, they reduce construction costs. For example, if free cooling eliminates the need for chillers, then we don't have to purchase and install those chillers, nor do we have to pay for generators or UPSs to back them up. Such construction cost savings can double the overall savings from efficiency improvements.

Actually *using* the provisioned power of a facility is equally important. For example, if a facility operates at 50% of its peak power capacity, the effective provisioning cost per watt used is doubled. This incentive to fully use the power budget of a data center is offset by the risk of exceeding its maximum capacity, which could result in outages.

5.5.1 DEPLOYING THE RIGHT AMOUNT OF EQUIPMENT

How many servers can we install in a 1 MW facility? This simple question is harder to answer than it seems. First, server specifications usually provide very conservative values for maximum power consumption. Some vendors, such as Dell and HP, offer online power calculators [DEC, HPPC] to provide better estimates, but it may be necessary to measure the actual power consumption of the dominant applications manually.

Second, actual power consumption varies significantly with load (thanks to energy proportionality), and it may be hard to predict the peak power consumption of a group of servers. While any particular server might temporarily run at 100% utilization, the maximum utilization of a group of servers probably isn't 100%. But to do better, we'd need to understand the correlation between the simultaneous power usage of large groups of servers. The larger the group of servers and the higher the application diversity, the less likely it is to find periods of simultaneous very high activity.

5.5.2 OVERSUBSCRIBING FACILITY POWER

As soon as we use anything but the most conservative estimate of equipment power consumption to deploy clusters, we incur a certain risk that we'll exceed the available amount of power; that is, we'll *oversubscribe* facility power. A successful implementation of power oversubscription increases the overall utilization of the data center's power budget while minimizing the risk of overload situations. We will expand on this issue because it has received much less attention in technical publications than the first two steps listed above, and it is a very real problem in practice [Man09].

Fan et al. [FWB07] studied the potential opportunity of oversubscribing facility power by analyzing power usage behavior of clusters with up to 5,000 servers running various workloads at Google during a period of six months. One of their key results is summarized in Figure 5.10, which shows the cumulative distribution of power usage over time for groups of 80 servers (Rack), 800 servers (PDU), and 5,000 servers (Cluster).

Power is normalized to the peak aggregate power of the corresponding group. For example, the figure shows that although rack units spend about 80% of their time using less than 65% of their peak power, they do reach 93% of their peak power at some point during the six month observation window. For power provisioning, this indicates a very low oversubscription opportunity at the rack level because only 7% of the power available to the rack was stranded. However, with larger machine groups, the situation changes. In particular, the whole cluster never ran above 72% of its aggregate peak power. Thus, if we had allocated a power capacity to the cluster that corresponded to the *sum of the peak* power consumption of all machines, 28% of that power would have been stranded. This means that within that power capacity, we could have hosted nearly 40% more machines.

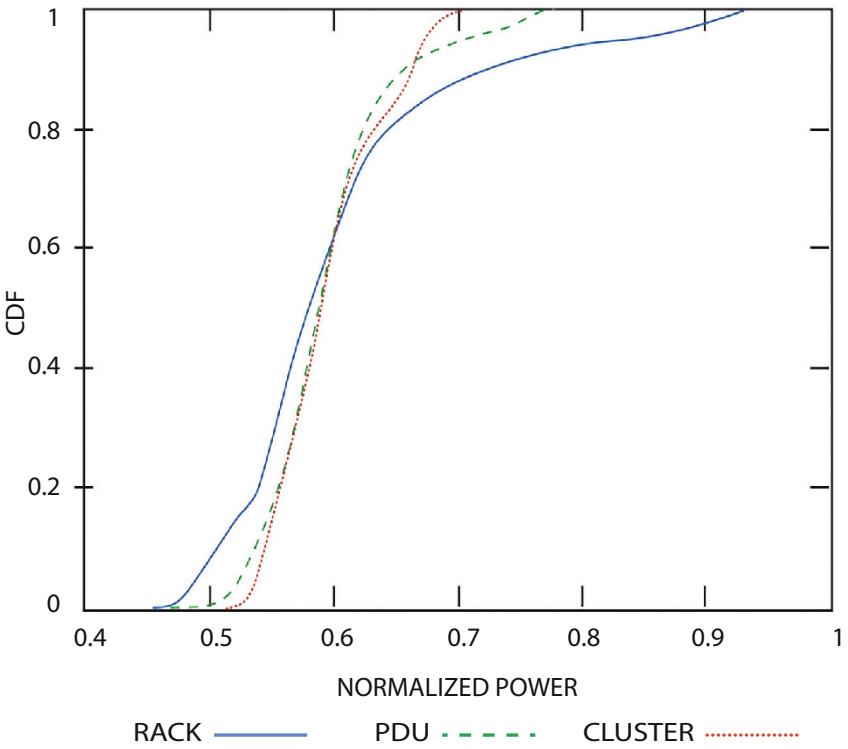

Figure 5.10: Cumulative distribution of time that groups of machines spend at or below a given power level (power level is normalized to the maximum peak aggregate power for the corresponding grouping) (Fan et al. [FWB07]).

This study also evaluates the potential of more energy-proportional machines to reduce peak power consumption at the facility level. It suggests that lowering idle power from 50% to 10% of peak (that is, going from the red to the green curve in Figure 5.6) can further reduce cluster peak

power usage by more than 30%. This would be equivalent to an additional 40%+ increase in facility hosting capacity.

The study further found that mixing different workloads within a cluster increased the opportunities for power oversubscription because this reduces the likelihood of synchronized power peaks across machines. Once oversubscription is applied, the system needs a safety mechanism to handle the possibility that workload changes may cause the power draw to exceed the data center capacity. This can be accomplished by always allocating some fraction of the computing resources to a workload that runs in a lower priority class or that otherwise does not have strict deadlines to meet (many batch workloads may fall into that category). Such workloads can be quickly paused or aborted to reduce facility load. Provisioning should not be so aggressive as to require this mechanism to be triggered often, which might be the case if oversubscription is applied at the rack level, for example.

In a real deployment, it's easy to end up with an underutilized facility even when you pay attention to correct power ratings. For example, a facility typically needs to accommodate future growth, but keeping space open for such growth reduces utilization and thus increases unit costs. Various forms of fragmentation can also prevent full utilization. Perhaps we run out of space in a rack because low-density equipment used it up, or we can't insert another server because we're out of network ports, or we're out of plugs or amps on the power strip. For example, a 2.5 kW circuit supports only four 520 W servers, limiting utilization to 83% on that circuit. Since the lifetimes of various WSC components differ (servers might be replaced every 3 years, cooling every 10 years, networking every 4 years, and so on) it's difficult to plan for 100% utilization, and most organizations don't.

Management of energy, peak power, and temperature of WSCs are becoming the targets of an increasing number of research studies. Chase et al. [Cha+01c], G. Chen et al. [Che+07], and Y. Chen et al. [Che+05] consider schemes for automatically provisioning resources in data centers, taking energy savings and application performance into account. Raghavendra et al. [Rag+08] describe a comprehensive framework for power management in data centers that coordinates hardware-level power capping with virtual machine dispatching mechanisms through the use of a control theory approach. Femal and Freeh [FF04, FF05] focus specifically on the issue of data center power oversubscription and describe dynamic voltage-frequency scaling as the mechanism to reduce peak power consumption. Managing temperature is the subject of the systems proposed by Heath et al. [Hea+06] and Moore et al. [Moo+05]. Finally, Pedram [Ped12] provides an introduction to resource provisioning and summarizes key techniques for dealing with management problems in the data center. Incorporating application-level knowledge to safely save power by re-shaping its latency distribution through DVFS is studied by Lo et al. [Lo+14], Kasture et al. [Kas+15], and Hsu et al. [Hsu+15].

5.6 TRENDS IN SERVER ENERGY USAGE

While in the past dynamic voltage and frequency scaling (DVFS) was the predominant mechanism for managing energy usage in servers, today we face a different and more complex scenario. Given lithography scaling challenges, the operating voltage range of server-class CPUs is very narrow, resulting in ever decreasing gains from DVFS.

Figure 5.11 shows the potential power savings of CPU dynamic voltage scaling (DVS) for the same server by plotting the power usage across a varying compute load for three frequency-voltage steps. Savings of approximately 10% are possible once the compute load is less than two thirds of peak by dropping to a frequency of 1.8 GHz (above that load level the application violates latency SLAs). An additional 10% savings is available when utilization drops to one third by going to a frequency of 1 GHz. However, as the load continues to decline, the gains of DVS once again return to a maximum of 10%. Instead, modern CPUs increasingly rely on multiple power planes within a die as their primary power management mechanism, allowing whole sections of the chip to be powered down and back up quickly as needed. As the number of CPUs in a die increases, such coarse-grained power gating techniques will gain greater potential to create energy proportional systems.

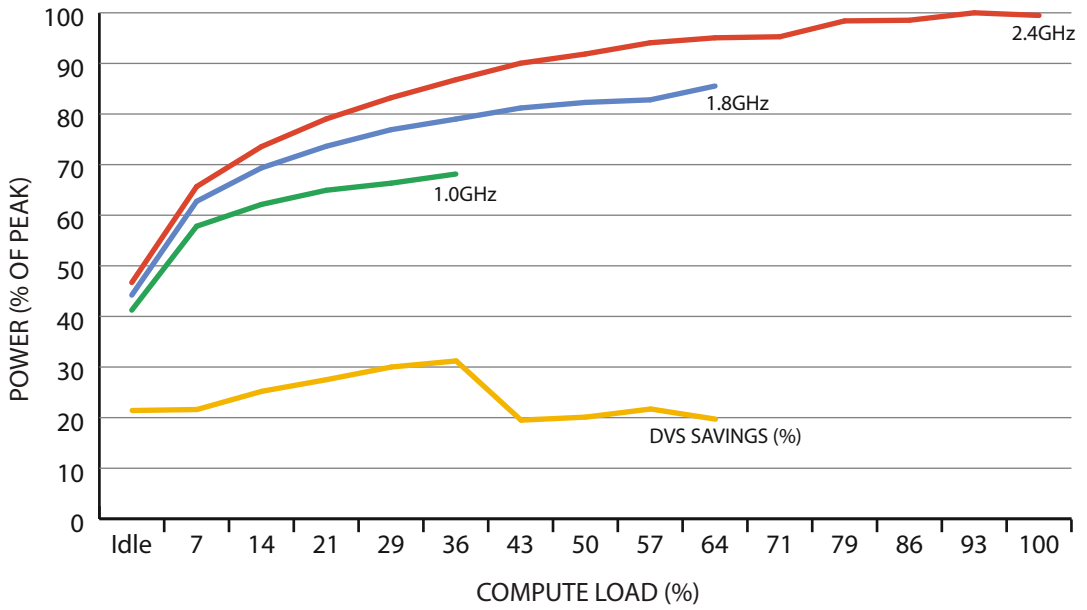

Figure 5.11: Power vs. compute load for a server at three voltage-frequency levels and corresponding energy savings.

A second trend is that CPUs continue to outpace other server components in energy proportionality improvements. The result is a power budget breakdown with larger energy fractions from non-CPU subsystems at *lower* utilizations.

5.6.1 USING ENERGY STORAGE FOR POWER MANAGEMENT

Several studies [Gov+, Wan+, Kon+12] propose using energy stored in the facility's backup systems (such as UPS batteries) to optimize facility performance or reduce energy costs. Stored energy could be used to flatten the facility's load profile (using less utility power when it's most expensive), mitigate supply variability in a wind-powered facility, or manage short demand peaks in oversubscribed facilities (using stored energy instead of capping the load).

In our opinion, the most promising use of energy storage in power management consists of managing short demand peaks or short-term supply reductions (say, when a data center is partly powered by renewable energy sources, such as wind). Power capping systems need some time to react intelligently to demand peak events, and may need to set peak provisioning levels well below the maximum breaker capacity in order to allow time for power capping to respond. A power capping system that can draw from energy storage sources for just a few seconds during an unexpected peak would allow the facility to safely operate closer to its maximum capacity while requiring a relatively modest amount of additional energy storage capacity.

To our knowledge no such power management systems have yet been used in production systems. Deploying such a system would be difficult and potentially costly. Besides the control complexity, the additional cost of batteries would be significant, since we couldn't just reuse the existing UPS capacity for power management, as doing so would make the facility more vulnerable in an outage. Furthermore, the types of batteries typically used in UPS systems (lead-acid) don't age well under frequent cycling, so that more expensive technologies might be required. While some have argued that expanded UPSs would be cost effective [Kon+12], we believe that the economic case has not yet been made in practice.

5.7 SUMMARY

Energy efficiency is a key cost driver for WSCs, and we expect energy usage to become an increasingly important factor in WSC design. The current state of the industry is poor: the average real-world data center and the average server are far too inefficient, mostly because efficiency has historically been neglected and has taken a backseat relative to reliability, performance, and capital expenditures. As a result, the average WSC wastes two thirds or more of its energy.

The upside of this history of neglect is that sizable improvements are almost trivial to obtain—an overall factor of two in efficiency improvements is possible, without much risk, by simply applying best practices to data center and server designs. Unfortunately, the path beyond this

low-hanging fruit is more difficult, posing substantial challenges to overcome inherently complex problems and often unfavorable technology trends. Once the average, data center achieves state-of-the-art PUE levels, and servers are deployed with high-efficiency power supplies that are available today, the opportunity for further efficiency improvements in those areas drops to below 40%. From a research and development standpoint, greater opportunities for gains in energy efficiency from now on will need to come from computer scientists and engineers, and less so from mechanical or power conversion specialists (though large opportunities remain for mechanical and power engineers in reducing facility costs in particular).

First, power and energy must be better managed to minimize operational cost. Power determines overall facility cost because much of the construction cost is directly related to the maximum power draw that must be supported. Overall energy usage determines the electricity bill as well as much of the environmental impact. Today's servers can have high maximum power draws that are rarely reached in practice, but that must be accommodated or limited to avoid overloading the facility's power delivery system. Power capping promises to manage the aggregate power of a pool of servers, but it is difficult to reconcile with availability; that is, the need to use peak processing power in an emergency caused by a sudden spike in traffic or by a failure in another data center. In addition, peak server power is increasing despite the continuing shrinking of silicon gate sizes, driven by a combination of increasing operating frequencies, larger cache and memory sizes, and faster off-chip communication (DRAM and I/O buses as well as networking speeds).

Second, today's hardware does not gracefully adapt its power usage to changing load conditions, and as a result, a server's efficiency degrades seriously under light load. Energy proportionality promises a way out of this dilemma but may be challenging to implement across all subsystems. For example, disks do not naturally lend themselves to lower-power active states. Systems for work consolidation that free up and power down entire servers present an avenue to create energy-proportional behavior in clusters built with non-energy-proportional components but are harder to implement and manage, requiring transparent process migration and degrading the WSC's ability to react to sudden upticks in load. Furthermore, high-performance and high-availability distributed systems software tends to spread data and computation in a way that reduces the availability of sufficiently large idle periods on any one system. Energy-management-aware software layers must then manufacture idleness in a way that minimizes the impact on performance and availability.

Third, energy optimization is a complex end-to-end problem, requiring intricate coordination across hardware, operating systems, VMs, middleware, applications, and operations organizations. Even small mistakes can ruin energy savings; for example, when a suboptimal device driver generates too many interrupts or when network chatter from neighboring machines keeps a machine from quiescing. There are too many components involved for perfect coordination to happen naturally, and we currently lack the right abstractions to manage this complexity. In contrast to

hardware improvements, such as energy-proportional components that can be developed in relative isolation, solving this end-to-end problem at scale will be much more difficult.

Fourth, the hardware performing the computation can be made more energy efficient. General purpose CPUs are generally efficient for any kind of computation, which is to say that they are not super efficient for any particular computation. ASICs and FPGAs trade off generalizability for better performance and energy efficiency. Special-purpose accelerators (such as Google's tensor processing units) are able to achieve orders of magnitude better energy efficiency compared to general purpose processors. With the sunset of Moore's Law and the breakdown of Dennard scaling, specializing compute will likely keep its place as one of the remaining tools in the shrinking toolbox of hardware changes that can further improve energy efficiency.

Finally, this discussion of energy optimization shouldn't distract us from focusing on improving server utilization, since that is the best way to improve cost efficiency. Underutilized machines aren't only inefficient per unit of work, they're also expensive. After all, you paid for all those servers, so you'd better keep them doing something useful. Better resource sharing through cluster-level scheduling and performance-aware scheduling have made very promising forward progress in increasing server utilization while maintaining workload encapsulation and performance robustness.