High-Efficiency Server Design

Eitan Frachtenberg, Ali Heydari, Harry Li, Amir Michael, Jacob Na, Avery Nisbet, Pierluigi Sarti
Facebook
{etc, alih, harrylihu, amir, mhnaJN, avery.nisbet, pierluigi.sarti}@fb.com

ABSTRACT

Large-scale datacenters consume megawatts in power and cost hundreds of millions of dollars to equip. Reducing the energy and cost footprint of servers can therefore have substantial impact. Web, Grid, and cloud servers in particular can be hard to optimize, since they are expected to operate under a wide range of workloads. For our upcoming datacenter, we set out to significantly improve its power efficiency, cost, reliability, serviceability, and environmental footprint. To this end, we redesigned many dimensions of the datacenter and servers in conjunction. This paper focuses on our new server design, combining aspects of power, motherboard, thermal, and mechanical design. We calculate and confirm experimentally that our custom-designed servers can reduce power consumption across the entire load spectrum while at the same time lower acquisition and maintenance costs. Importantly, our design does not decrease the servers' performance or portability, which would otherwise limit its applicability.

1. INTRODUCTION AND BACKGROUND

In the past decade, we have witnessed a fundamental change in personal computing. Many of the modern computer uses—networking and communicating; searching; creating and consuming media; shopping; and gaming—increasingly rely on remote servers for their execution. The computation and storage burdens of these applications has largely shifted from personal computers to datacenters of service providers, including Amazon, Facebook, Google, and Microsoft. These providers can thus offer higher-quality and larger-scale services, such as the ability to search virtually the entire Internet in a fraction of a second. It also lets providers benefit from the economies of scale and increase the efficiency of their services.

As one of these service providers, we leased datacenters and filled them with commodity servers. This choice makes sense at small to medium scale, while the relative energy cost is still small and the relative cost of customization outweighs the potential benefits. But as our site grew to become one of the world's largest, with a corresponding growth in computational requirements, we investigated alternative, more efficient designs for servers and datacenters, which we will start deploying this year. This paper primarily focuses on server design, with an eye to holistic combination with the datacenter for maximum gains [22]. We believe that this server design is general enough to appeal not only to Web and cloud services, but also to high-performance computing (HPC) centers and Grids, since they similarly use most of the commodity high-performance components that our servers use. Because of the large potential benefit of this design, we have decided to openly share the detailed specifications of our system architecture [7]. Within the scope of this paper, however, we cannot cover the lowest level of design detail published in these documents, so instead we focus on the high-level design principles that guided our high-efficiency servers.

How is server efficiency defined then? For the purposes of this paper, we look at the total cost of ownership (TCO) of the servers in a datacenter: the cost to equip and run the servers. Realistically, TCO is also affected by application performance and is often defined by metrics such as work per dollar or work per joule. But discussions of performance are often muddled by ill-defined or incompatible metrics [10]. The interactive nature of Web server applications can further complicate the metrics discussion. For example, we may care more about response time until we meet a certain user-perceptible threshold, and then prioritize throughput. Even with agreed-upon performance metrics, server efficiency is difficult to compare when all design choices are considered, because best-performance parts often cost a premium to acquire or operate. For this study, we try to avoid such parts since their cost premium outweighs their performance benefit in the TCO equation. On the other hand, we also exclude the low-end design choices that negatively affect performance, such as some lower-voltage processors, "wimpy" and high-density servers, or under-clocked components. Such designs typically increase cost by requiring more work aggregation over additional nodes ("scale-out") or higher component count. Moreover, they often imply slower sequential processing, a trade-off that does not make sense in many applications [16, 21]. For the interested reader, these topics have been studied extensively in related works [1, 2, 13, 22, 29].

Consequently, to simplify our efficiency analysis, we keep the performance constant, removing it altogether from the equation. Our discussion of efficiency will therefore focus

SC11, November 12-18, Seattle, Washington, USA

on reduced TCO for the same performance, which in turn implies fewer resources to equip and run for the same level of service. We also keep constant all aspects of networking and storage, although the next Open Compute specifications will include high-efficiency storage nodes as well.

What then comprises TCO? One obvious element is equipment cost (*capex*), which service providers can typically negotiate down through large-scale acquisitions of commodity hardware. Still, this is often the largest cost element in TCO [15]. Customized design choices can reduce acquisition cost further, so we explore this dimension at length in this paper.

Another large element of TCO is the energy cost to run these servers (part of operating costs, or *opex*), as well as the power and equipment to cool them, which grows with the servers' power. Some estimates put the fully burdened power cost of servers (including infrastructure and cooling) close to their purchase cost [3, 12, 22, 26, 28]. A recent study placed the total power consumption of US datacenters this year at 2% of total electricity use [19, 20]. Because of the magnitude of energy consumption, power efficiency is paramount in overall server efficiency, with each percentage point improvement having large impact over installations such as ours. For Web servers in particular, power efficiency matters not only at the highest continuous loads, but also at low and idle loads, since the workload can vary significantly based on diurnal cycles, application weight, external events, etc. Even HPC and cloud servers can have varying power load, because of synchronization delays, internal fragmentation, etc [9]. We will therefore place a special emphasis on power efficiency across all loads in this paper, and discuss some ways to reduce the infrastructure cost and overhead to deliver this power.

Reliability of components also affects TCO, since repairs can be costly in labor and reserve-capacity provisioning. Another similar but even less tangible element of TCO is the serviceability of servers, affecting the time (and therefore, cost) of repairs and upgrades. We address both elements in our design by simplifying component access and composition and reducing their overall count.

Most elements of a server's design affect more than one aspect of TCO. Furthermore, imposing or removing constraints from one component, such as the server's height, affect the constraints on other components. So server design requires the consideration of all components and tradeoffs together. For organizational purposes, however, we divided our contributions into the next four distinct sections: power supply design, motherboard design, thermal design and mechanical design. We then tie these design spaces back together in the evaluation and discussion section (Sec. 6), which includes experimental power and thermal efficiency results. Finally, Sec. 7 summarizes the main findings of the paper and explores avenues for further research.

2. POWER SUPPLY DESIGN

At first glance, the role of the power supply unit (PSU) appears limited to the conversion of the wall-power alternating current (AC) to the motherboard's lower direct current (DC) voltages, and therefore any efficiency gains are bounded by the losses of this conversion. In reality, however, PSU design choices can cascade all the way down to the motherboard and up to the datacenter's power distribution design, bearing substantial aggregate impact.

Electricity in the US is distributed from a local site transformer (or generator, in case of an outage) to the typical datacenter at $480Vac$ three-phase (or $277Vac$, Lines-to-Neutral) [14]. Inside the datacenter, it is fed to large uninterruptible power supply (UPS) units, and then further transformed to $208Vac$ by numerous power distribution units (PDUs), eventually reaching the servers [8]. These intermediate steps lose non-negligible amounts of power to inefficiency, so we designed our PSU to eliminate them both.

2.1 Outside the (Server) Box

To improve power distribution, we designed the PSU to accept $277Vac$ directly, a first for motherboard-attached PSUs to the best of our knowledge. (For usability worldwide, our PSU actually accepts a voltage range of $180Vac \sim 305Vac$.) Aside from the large capex and opex savings associated with eliminating the PDUs, distributing the datacenter power at the higher voltage also wastes less energy on transmission inefficiency and requires less current, potentially reducing the copper cost (capex) in the wires. Using standard wires, on the other hand, could be simpler to deploy and offer lower resistance (translating to 0.5% less energy loss) which reduces opex instead.

Uninterruptible power is critical for business continuity, so eliminating the UPS altogether without a robust alternative solution is infeasible. The alternative we designed replaces the centralized UPS with rack-level DC energy storage, in the form of a custom battery cabinet (Figs. 1 and 2). This cabinet takes the same $277Vac$ input and outputs an industry-standard $48Vdc$ nominal of offline backup power directly to the racks, as well as online and backup power to the network switches. It uses high-efficiency rectifiers to convert AC to DC, but in normal operation, with the battery fully charged, they operate on a small amount of fixed current with an equivalent UPS efficiency of $\approx 99.5\%$. The cabinet also includes an impedance measurement scheme for real-time monitoring of the batteries' health. The cost savings of this customized solution are substantial: capex (per watt of backup power) is reduced by $\approx 8X$ (from $\approx \$2/W$ to $\approx \$0.25/W$) while the conversion energy loss is reduced by $\approx 20X$ (from 10% to 0.5%), compared to a standard UPS design.[1] On the output end of the PSU, we eliminated some material cost and power losses by attaching the PSU directly to the motherboard, providing a single output voltage ($12Vdc$), which is later converted as needed by the motherboard (Sec. 3).

2.2 Inside the PSU

The power factor of a PSU, defined as the ratio between the real and apparent power [18], is a key component of the performance of the PSU. The primary concerns with a low power factor are that it triggers higher energy rates from the utility company and degrades the power quality for the datacenter and external users. We specified a power factor correction scheme (PFC) with an active, single-phase interleaved topology, typically associated with lower-power applications [30]. This PFC is unique among PFCs of comparable power in that it can take up to $300Vac$ as input, and is optimized for efficiency and linear operation, and yet it

[1]Recent high-end UPS systems, such as those employing ESS, have twice the efficiency of the standard models, lowering the relative opex advantage of our design by half, but significantly increasing the capex difference.

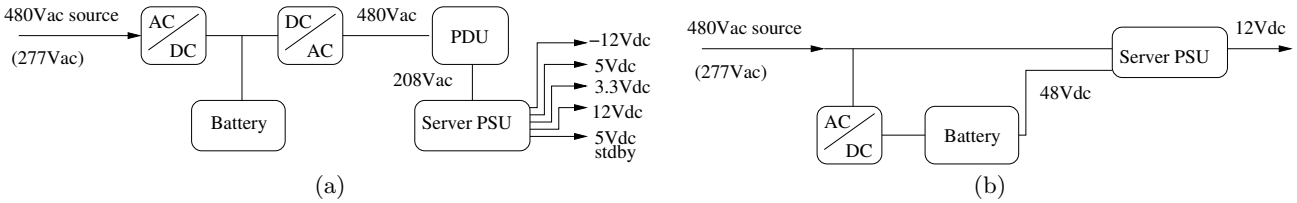

Figure 1: Schematic view of power distribution to the server. Inside the typical datacenter (a), an external three-phase $480Vac$ source provides power to large UPSs, consisting of rectifiers to convert AC to DC, a battery to store the energy, and an inverter to convert back to AC. Voltage is then transformed to $208Vac$ through PDUs and distributed to the servers. Each conversion loses some energy, typically around 5%, 5%, and 3%, respectively. Contrast this to our offline design (b), where in normal operation, $277Vac$ power from the AC source flows directly to the PSU with no conversion losses. The battery takes only a fixed amount of charge in normal offline operation, representing only $\approx 0.5\%$ of equivalent system loss. In backup operation, the battery feeds the server PSUs directly with $48Vdc$.

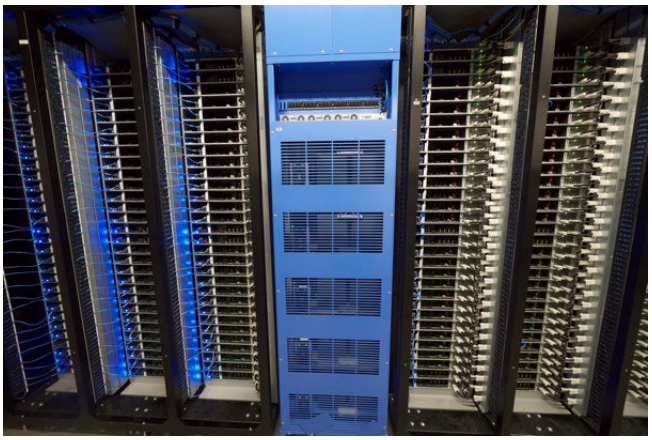

Figure 2: Picture of actual populated racks and battery cabinet. Racks are organized into triplets, each holding 90 servers and two 48-port switches. One battery cabinet provides backup power to the two triples pictured.

DC load (%)	Power factor	iTHD	Efficiency
$90W$(20%)	0.88	7.07%	90.50%
$112.5W$ (25%)	0.91	5.47%	91.95%
$225W$(50%)	0.97	3.56%	94.26%
$450W$(100%)	0.99	3.63%	94.61%

Table 1: Power supply performance at $277Vac$ and different loads, excluding PSU fan power and cable losses. Data was averaged over a random sample of PSUs, with the best samples exceeding 95% efficiency.

still costs less than comparable commodity PFCs. Its high power factor guarantees high performance even at low loads, with the PSU surpassing 90% efficiency from 20% load and peaking above 95% (Table 1). This efficiency profile exceeds even the Climate Savers Computing Initiative (CSCI) Platinum Standard [32]. (The standard is for $230Vac$, whereas our PSU meets Platinum requirements at an even stricter $200Vac$. Moreover, for a PSU of only $450W$ nominal power, the standard is particularly challenging to meet).

Although we prioritized sustained efficiency with the choice of topology, we also paid attention to the current total harmonic distortion (iTHD), which is often substantial with comparable PSUs [6]. iTHD measures the deviation of the input AC current from a perfect sinusoidal, and is defined as the ratio between the aggregate power of all harmonic components to the power of the fundamental frequency. It is typically much worse at lighter loads, which are common. Our PSU keeps iTHD at less than 5.5% starting from 25% load, and less than 4% distortion for most of the load range (Table 1), which helps to improve overall system efficiency, as well as power quality. For comparison, the iTHD values we measured in commodity PSUs were almost always higher than 15%.

A low iTHD also lowers the capacity and reliability margin requirements on the backup (diesel) emergency generators, since they may not be able to deliver the full active AC power in the presence of high current distortion, possibly stalling altogether. In addition, the PSU's improved power factor and iTHD have positive environmental effects outside the datacenter: they translate to higher power distribution quality (through reduced harmonic content) for other consumers, higher utilization of contracted power, and possibly lower over-provisioning need for the utility company, which is tied to indirect losses and power distribution losses. Finally, low iTHD values greatly reduce the current in the neutral conductors and associated losses and obviates the need for K-rated transformers [5].

Table 2 summarizes the overall power efficiency gains of our PSU design, from the point of entry to the datacenter and up to the server's motherboard.

2.3 Backup Operation

To handle power outages without interruptions, our PSU also accepts an industry-standard $48Vdc$ nominal input (specified input range is actually quite wider: $38Vdc \sim 59Vdc$). Note that this power can come from a common source to provide some economies of scale. In fact, each of our battery cabinets provides backup power to 90 servers for approximately $45sec$ at full load (Fig. 2).

The PSU converts the input voltage down to $12Vdc$ with a unique converter to ensure smooth transition to backup power operation [30], specified to a minimum of 90% efficiency from 40% load (intentionally not best-in-class in order to contain cost). The converter is isolated to guarantee

Customization description	Efficiency gain
DC-based UPS	$10 \sim 12\%$
High-efficiency PSU	$2 \sim 7\%$
Obviating PDUs	$\approx 3\%$
Low input AC iTHD	$\leq 0.5\%$
$277Vac$ distribution	$\approx 0.5\%$
Direct attachment	$\leq 0.25\%$
Overall	$13 \sim 25\%$

Table 2: Breakdown of potential power efficiency gains (or inefficiency avoided) from custom-designed power solution. Gains are represented as percent of power saved, so the overall gain is multiplicative. The first number, UPS efficiency gain, is derived as follows. A single UPS, running at full load, is typically around 94% efficient. [14]. These are rarely run in isolation however, but rather as part of an $N+1$ or $2N$ redundancy scheme, which lowers their effective load and consequently, their efficiency, to approximately $88\% \sim 90\%$. The other five gain numbers are derived from the specifications and manuals of commodity components in common use today.

very low noise and avoid recirculating currents through the server rack chassis ground, effectively equivalent to having a single battery per server with floating DC backup voltage. Another important converter feature of the DC-DC backup is its ability to share the current with the main AC-DC converter for a few milliseconds during the transition at AC loss, and even longer during the transition at recovery. This feature enables the lowering of start-up DC current, thus minimizing the potentially deleterious effect of impedance in the DC distribution hardware (battery cabinet and connections). This effect could catastrophically disrupt the backup operation at the very moment it is needed the most. This feature also means that the backup power can originate from a remote DC source, because it reduces the in-rush current, allowing the use of high-inductance DC distribution.

The PSU enters backup mode upon loss of AC power, providing emergency power for the next 10 to 15 seconds while the generators kick in. If at least $6sec$ went by before the PSU senses AC power again, it assumes the next source of the power is the generators, and uses a modified scheme to return to AC: each PSU waits for a random delay of up to $5sec$, letting the generators smooth in to full load operation in a linear fashion. The temporary AC-DC load-sharing ability makes this transition even smoother, as is the case for initial startup. Once more, this continuous transition permits cost and capacity savings on the generator side, as well as improving the reliability of backup operations. Finally, if the AC grid power is restored before the generators start, all the PSUs revert to normal operation without delay, transitioning smoothly—again by load-sharing—from DC to AC power.

3. MOTHERBOARD DESIGN

Motherboards are typically designed by original design manufacturers (ODMs) to fit in a large number of different stock-keeping units (SKUs), integrated to servers by original equipment manufacturers (OEMs). Consequently, flexibility of application and a wealth of features take high priority in motherboard design. HPC servers, cloud server farms, and datacenters, on the other hand, tend to be more homogeneous in their requirements, using only a handful of different SKUs. The wealth of oft-unused motherboard features then becomes a burden, since these features typically add to the purchase cost, power consumption, and service liability. (Generally speaking, the more features and components a motherboard has, the more opportunity there is for faults and reliability issues.)

One example is the baseboard management controller (BMC), ubiquitous in server motherboards. The BMC has numerous responsibilities, including monitoring hardware sensors, logging hardware errors, and providing a serial console. We found that nowadays we can eliminate the BMC altogether without adversely affecting the overall manageability of the server—by giving up some noncritical BMC functions and finding workable replacements for others. In practice, we found that the main network interface and BIOS combined can implement much of the BMC's functionality: SMBIOS system event logging, console access, hardware error reporting, and reboot-on-LAN. And while functioning normally, the operating system can monitor hardware and software health. The resulting TCO benefit is substantial enough to justify these workarounds, since a BMC can represent up to 2.5% of the cost of the server, while requiring up to 4% of its power. Other removed components are the on-board serial connector and boot progress (POST) LED display, which play an important role when someone has to physically troubleshoot the server, but otherwise remain unused. To satisfy this operational need, the motherboard has a special connector in the front for attaching a separate hot-pluggable debug card. The debug card includes the serial interface, POST codes with enhanced diagnostics, and other utilities. This solution is more cost effective (we expect only a few debug cards will be needed), more useful, and more convenient, since the server does not have to be pulled out of the rack to observe the POST codes.

Also expendable was the SAS controller on the motherboard, since we primarily use SATA-type drives in our servers—resulting in a significant reduction of another $\approx$ 2.25% in server cost and up to 3% of its power. In its stead, we added a customized external PCIe link (using a mini-SAS connector), which affords the system more extensibility, such as the ability to connect to a storage tray, video display, coprocessor, network, etc. This solution is more flexible and cost-effective than one based on on-board SAS, since it only adds back a negligible amount to the cost and power draw of the server. We also removed other redundant or obsolete functionality, such as extra PCI slots, extra USB connectors, PS/2 connectors, and VGA output (normally associated with the BMC). And to obviate the front-panel connector and its associated cables and harness, the motherboard has power/reset switches soldered directly to its front, as well as light indicators for power, hard-disk activity, and beep, where a visual signal is more useful than an auditory one in the datacenter environment.

As shown in Fig. 1, our PSU only outputs $12Vdc$. Some motherboard DC-DC converters (voltage regulators, or VRs) are therefore required to power different components, such as the hard drives' $5Vdc$. Again, we specified the principal DC-DC converters for high efficiency—minimum of 91%. This requirement saves up to 3% of the server's power compared to commodity parts and is not much more expensive

Component	Capex	Opex
Removing BMC	2.5%	4%
Lean BOM option	1.5%	negligible
Replacing SAS	2.25%	3%
Debug card	0.1%	negligible
On-board VRs	(0.3%)	3%
Overall	6%	10%

Table 3: Maximum capex and opex gains with the customized design for our most widely used SKU, compared to commodity motherboards with the same compute capacity. Values are normalized *per server* and are derived from ODM specifications. They vary based on vendor, choice of processor, negotiated price, etc. Opex gains depend on component load and include only power gains (as percentage of entire server power), since serviceability is hard to quantify before years of service.

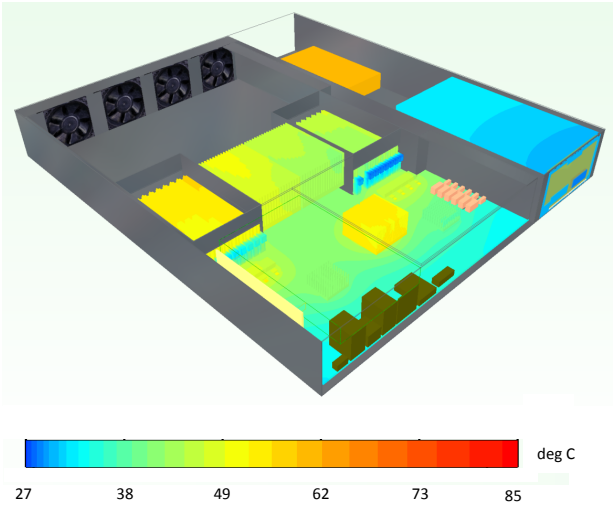

Figure 3: Flotherm isometric view of thermal design shows chassis, motherboard (with dual processors and memory slots side-by-side), fans, and the hard-disk drive (HDD) behind the PSU. The temperature range here assumes an inlet temperature of 27^oC. The air duct on top is elided for visualization purposes.

than stock VRs. Table 3 summarizes the cost gains of these features.

Cost is not only a function of component choice, but also one of research, design, development, and verification. To curb these costs, we designed a single board (PCB) with different stuffing options to match multiple applications. For example, the board has the space for up to 18 memory slots, but our most widely used application requires only a third of the slots. For these servers, the bill-of-materials (BOM) can simply include components for a memory slot on every third position, eliminating the waste of the unused slots and possibly benefiting from higher memory speeds. Another example of a BOM option that shares the same board design is the choice of a higher- or lower-end network interface, based on the application.

One more interesting feature that can have a large impact on the reliability and efficiency of the server and datacenter as a whole is HDD startup. Hard drives can draw a large current when they spin up, which on aggregate can cause large in-rush current that exceeds the PSU's maximum rated current and causes it to fail. To avoid this spike, when multiple hard drives are connected to the motherboard, it staggers each disk' spin up by a few seconds. Another benefit of this feature is that it allows for a cheaper and lower-power PSU, since it does not have to over-provision for HDD startup.

Crucially important is also the motherboard layout, which directly affects airflow through the chassis and consequently its thermal efficiency. Cables impede airflow, which is another reason why we connect the PSU directly to the motherboard (Sec. 2.1). We designed our motherboard for minimal cables and high airflow efficiency, as depicted in Fig. 5.

4. THERMAL DESIGN

The goal of server thermal design is to cool down the hot components to their operating temperatures with a minimal expenditure of energy and component cost. The typical mechanism used to cool servers at the datacenter level is to cool air at large scale and push it through the servers using their internal fans. The cool air picks up heat from the server components, exits from the server outlet, and is then pushed back to the atmosphere or chilled and recirculated. More efficient cooling is achieved with air containment in aisles, with the front, (or inlet), side of the server facing the "cold aisle" and the back facing the "hot aisle." Yet another technique to improve cooling efficiency is to create an air-pressure differential between the aisles using large datacenter fans.

Our specific design goal was to be able to cool our upcoming datacenter with unchilled outside air almost year round by allowing effective server cooling even with relatively high inlet air temperature and humidity. To achieve this goal, we needed a more effective design for heat transfer than our commodity servers'. Improving airflow through the server is a key element here: when internal server components impede airflow, more cooling energy is expended (for example, by faster fans, cooler inlet air, or higher air pressure). One technique by which we improved airflow in the chassis was to widen the motherboard and spread the hot components side by side, not behind each other. We also moved the hottest components—processors and memory—to receive the coldest air first. (Note that they are also located closer to the air inlet than in the typical back-mounted motherboard.) Another modified dimension was the server height: given a relatively constant rack height (for servicing purposes), a taller server reduces cooling energy but also the rack's computational density. Our Flotherm simulation found that the optimal server height to maximize the compute-capacity per cooling-energy ratio to be the uncommon $1.5U$ height with large-surface-area heat sinks (confirmed experimentally in Sec. 6.3). This height also allows for an air duct that sits on top of the motherboard and "surgically" directs airflow to the thermal components in parallel heat tracks, reducing leaks and air recirculation inside the chassis. Obstructions to airflow are kept to a minimum, decreasing the number of fans required to push the air out (Fig. 3).

Although a recent study found that newer HDDs do not necessarily fail more with higher temperatures [27], we designed the chassis to keep all six possible HDDs well within their specified operating temperatures. And since the high-efficiency PSU gen-

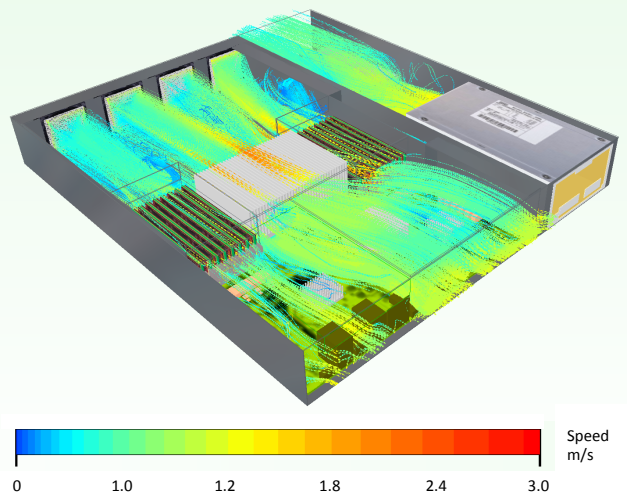

Figure 4: Flotherm simulation of airflow speed at minimum continuous fan speed.

erates less than 20W of waste heat under load and has its own fan to dissipate this heat, even the HDDs behind it operate at less than 40^oC (verified in simulation and empirically). Contrast this with typical server designs that locate the HDD in the front of the chassis to meet its cooling requirements.

Also reduced is the amount of airflow required through the system to keep it cool—up to half the volume per unit time compared to standard $1U$ servers, for the same inlet-to-outlet temperature difference (Fig. 4). This low requirement, combined with smart fan-speed controllers, results in fans that spin at their minimum continuous speed nearly year-round, depending on ambient temperature and workload. An additional advantage of this low speed, continuous operation is a longer expected fan lifetime compared to the typical fan's start-stop cycles, leading to overall improved server reliability. It also naturally translates to lower power and operating costs for server cooling—approximately 1% of the total server power—compared to the more typical 10% in commodity servers. Somewhat surprisingly, even the capex of the server's cooling components alone is about $40 \sim 60\%$ lower than a typical server's, depending on OEM component pricing. The two main reasons for this improvement are that we can use thinner fans (owing to the reduced airflow) and simpler heatsinks without a heat pipe (owing to the larger surface area).

Closing the cycle, these efficiency gains carry forward to the datacenter level as well. Our server is capable of working reliably at air inlet temperatures of 35^oC and a relative humidity of 90%, exceeding the most liberal ASHRAE recommendations for datacenter equipment [17]. In practice, this allows us to cool our upcoming datacenter almost exclusively on free (outside) air, relying on infrequent evaporative cooling instead of chillers for particularly hot days.

5. MECHANICAL DESIGN

Three principles guided our mechanical server design:

1. Prefer efficiency over aesthetics.
2. Optimize for high-impact use cases instead of generality of application.
3. Maximize serviceability and limit it to the front of the server.

Let us elaborate on these principles.

OEMs compete to sell servers, and as such care about presentation: expensive hardware needs to look expensive. For example, servers are adorned with plastic front plates, logos, emblems, and paint that serve no function; increase cost, transportation weight and complexity; and end up in a landfill. Sometimes their effect is downright detrimental, such as front plates that obstruct airflow and slow down servicing. Additionally, OEMs have to design for modularity, expandability, and conformity with standard racks, since they aim for as many unknown applications as possible. This necessarily leads to a plethora of brackets, plastic clips, ports, hard-disk bays, and other options that are not always used but always add parts and cost. We have the luxury of a small set of well-understood applications, allowing us to design for a small number of SKUs with the minimal required functionality and expandability, and no "vanity features" whatsoever. And yet we believe that even applications outside our own can benefit from the efficient design choices outlined below, because they share the same computational components: CPU, RAM, etc.

Finally, the third principle may not affect the server's efficiency directly, but has a crucial role for datacenter efficiency and serviceability. Eliminating the requirement to service the back of servers can lead to better thermal management in hot/cold aisles in the datacenter and avoids the unnecessary exposure of operators to the hot (back) side of the rack. It also reduces the time to fix servers and consequently the required on-site staff size.

With these principles in mind, we customized our server mechanical design as follows.

- No resources were spent on appearances: no stickers, paint, plastic bezels, or face plates.
- Use of low-cost pre-plated metal: cost effective in bulk and more efficient than plating and painting the completed chassis. Although it is possible for corrosion to build up around the edges, the effect is only cosmetic.
- No screws required anywhere except for the attachment of heatsinks, enabling efficient assembly and servicing. Motherboard is attached to chassis with press-fit stand-offs.
- Components designed for quick-release for servicing: Drive-cage and rails; direct-attached PSU, and fans.
- Wherever possible, recycled materials substitute for new plastic parts.
- Servers use no cover lids, requiring fewer parts and less work when servicing. Instead, the server above acts as a lid.
- In addition, the server chassis dimensions are specifically designed to accommodate server components, not the other way around. For example, the motherboard itself is of nonstandard size to accommodate the PSU direct-attach connector and to spread out the hot components. Similarly, we designed our racks to accommodate the servers. The height of the servers ($1.5U$) was calculated to be thermally optimal, compared to the more typical $1U$ height (Sec. 4).

- Since all servers have the same height, our racks require no rails and are built much more economically by simply cutting and bending metal tabs to hold the uniformly sized servers. The bent metal tabs can develop corrosion, but again with no functional repercussions. A plunger on the front right side of the server locks the server in place (Fig. 5).

- To efficiently accommodate standard 48-port switches while maintaining a serviceable rack height, a single rack actually holds three columns of servers, with two switches (Fig. 2). We thus amortize some of the rack's cost and optimize the server-to-switch ratio. The switches fit in a customized quick-release tray that can accommodate any standard 19"/1U gear.

- Motherboard is front-mounted in chassis, as opposed to the typical back-mounted server. This improves cooling, allows easy servicing from the front (including debug card attachment), and eliminates the need for a front panel with its associated connectors and cables.

- All cables, including network and power, are accessed from the front. The uniform server height again allows for optimal cable design, since both power and network cables can have the minimal length required to have the same pitch as the server they attach to. This design reduces cost and clutter, as well as simplifies service. The short power cables connect to a bare-bones, pre-plated-metal power strip designed for the known server pitches.

Moving away from standard parts often incurs higher costs, because of economies of scale. However, we designed most of these customizations for reduced cost, and for large installations the acquisition scale is such that the added cost of customization is minimal. We therefore estimate a modest net capex gain of roughly 1% per server with this chassis and rack design, lacking a precise chassis cost of commodity servers. A more significant reduction in TCO stems from the hard-to-quantify gains in reliability and serviceability improvements. We will try to assess and report these gains as we gather statistically meaningful data over the next year.

6. EVALUATION AND DISCUSSION

6.1 Methodology

We have evaluated the power, thermal and performance properties of a prototype of our design against two commodity servers. Some component and specification details in these servers are proprietary to our business partners and we cannot disclose them. However, both commodity servers are a common off-the-shelf product from two major OEMs, with dual Xeon X5650 processors, 12GB DDR3 ECC memory, on-board Gigabit Ethernet, and a single 250G SATA HDD in a 1U standard configuration. The first server, "Commodity A," is widely deployed in our leased datacenters for our main Web application. The second server, "Commodity B," is a three-year-old model that was updated to accept the latest generation processors and obtains similar application performance to server A. All three servers had the same chipset, and motherboard technology hasn't changed significantly during this time.

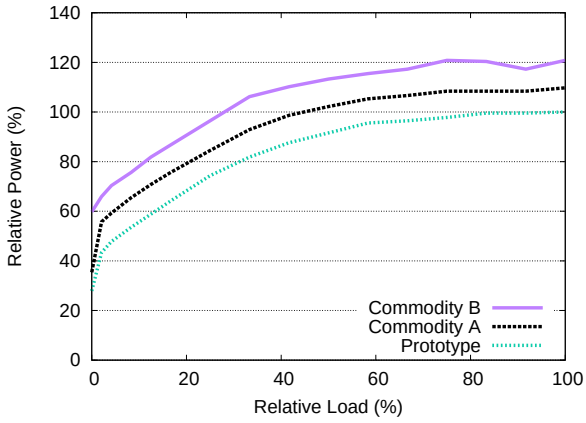

Figure 6: Power comparison of prototype and commodity servers. All values are normalized to the load and power the prototype server at its throughput saturation point (rightmost prototype data point is always 100%/100%).

To ensure a fair comparison, we used the exact same CPUs, DIMMs, and HDD unit in turn, moving them from server to server. The only differing components between the three servers were therefore the chassis, motherboard, fans, power supply, and power source (208Vac/277Vac). The HDD contained a fresh operating system and application image to match our production Web server configuration. We also updated the BIOS versions and attempted to keep their settings identical between servers; but different vendors expose different settings and implement different algorithms, so some BIOS differences remain, accounting for some performance differences. Note, however, that any option that improves performance, such as QPI rate or turbo mode, also increases power consumption.

We used the Apache *ab* tool [31] to generate Web load with an increasing number of concurrent clients, from zero ("active-idle") up to the 100% (saturation) point, where any increase in load did not increase throughput. This homogeneous request stream may not represent a realistic workload mix for absolute performance metrics, which is immaterial here because we are only looking for comparative metrics. What does matter is that experiments are reproducible, the offered load is easy to control, and performance and power scale with the offered load. For each offered load point, we measured the performance metrics reported by *ab* and the power consumption using a Yokogawa WT3000 power meter connected to the AC input, averaged over 256 samples.

6.2 Power/Performance Efficiency

Fig. 6 shows the power consumption of all three servers. The power consumption of the prototype is on average 16% better than system A's and 28% better than system B's. The relative gain in an actual deployment should be higher, since production loads are typically much less than 100%, to allow for load spikes and to curb response times. In absolute terms, the power difference is much higher at low load and at

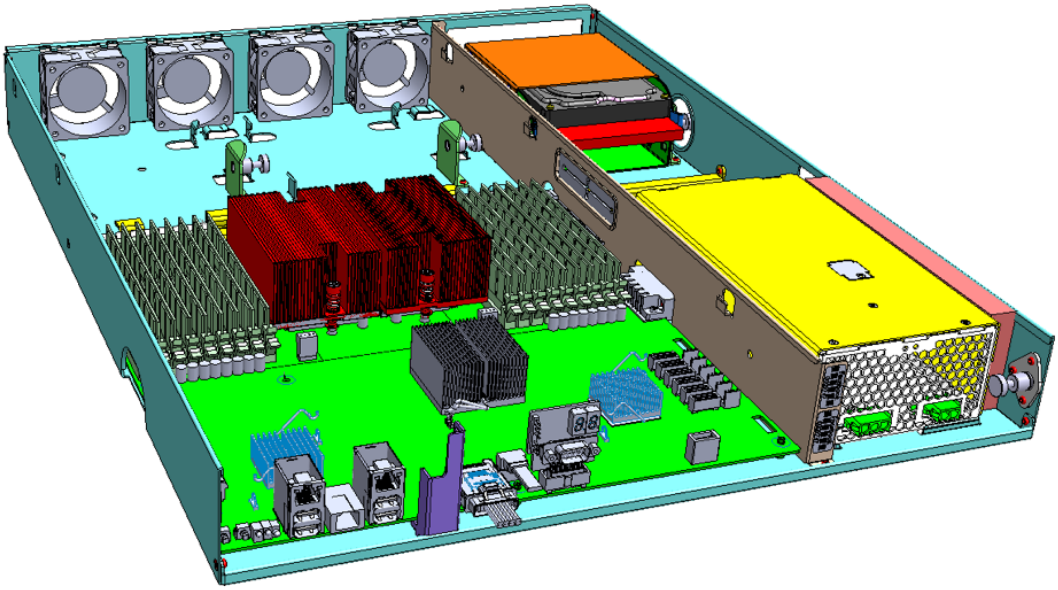

Figure 5: Motherboard and chassis layout. Two CPUs and Eighteen DIMMs are placed side by side, yielding the best thermal result based on airflow coming from front to rear. They are the primary power consumers and heat generators, taking up over 80% of the total system power, so spreading them apart ensures that they do not heat up each other. All the I/O ports, including the debug card (shown attached here, but normally unneeded), are placed on the front to facilitate easy access for rack service. The PCIe x16 slot placement (front left) can accommodate a full-height horizontal PCIe card on a riser board (obstructed by side panel). The PSU power mates directly with the motherboard, eliminating PSU power cable routing (only remaining cables to the fans and HDD are flush to the sides). Mounting holes' placement offers the best support for motherboard. Extra space for up to five more HDDs lies behind the motherboard and PSU.

the leftmost, active-idle measurement, the prototype system is more than 20% and 40% power-efficient than servers A and B respectively. Although no service provider wants to run machines at low load, some number of machines are always in this state—to account for reserve capacity to handle load spikes, for example [2], or because of external fragmentation in clusters and Grids [9].

In terms of relative performance, Figures 7 and 8 show a consistent picture across most load points. The prototype system exhibits $\approx$ 10% improvement in both throughput and mean latency over server A and $\approx$ 8% improvement over server B. Since most of the performance-related hardware components are shared between tests, we attribute this performance difference to BIOS differences and not to our design. At the very least, this verifies that our design does not hurt performance, which was an explicit constraint, as discussed in Sec. 1.

6.3 Thermal Efficiency

Thermal efficiency is another important element of the TCO, both in terms of cooling energy in the server (fan energy) and in the datacenter. Recall from Sec. 4 that our thermal design is based on a spread and unpopulated board placed in a $1.5U$ pitch open chassis, and employs four high-efficiency custom $60 \times 25mm$ axial fans. In contrast, the commodity servers use a thermally shadowed, densely populated $1U$ chassis with six off-the-shelf $40 \times 25mm$ fans. To evaluate the thermal efficiency, we put each server in a specially-built airflow chamber that simulates conditions in an actual rack deployment. This chamber also isolates

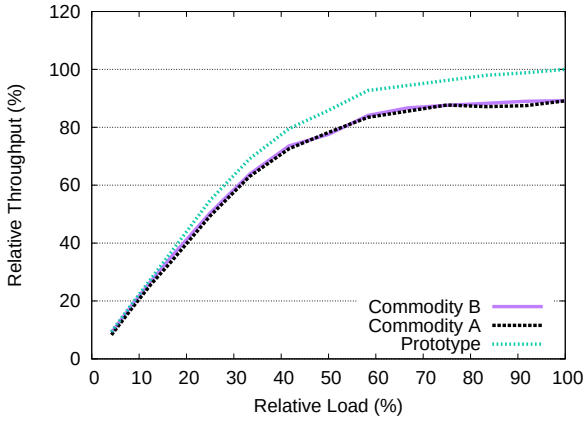

Figure 7: Throughput comparison of prototype and commodity servers. All values are normalized to the load and throughput (in requests per second) of the prototype server at its saturation point.

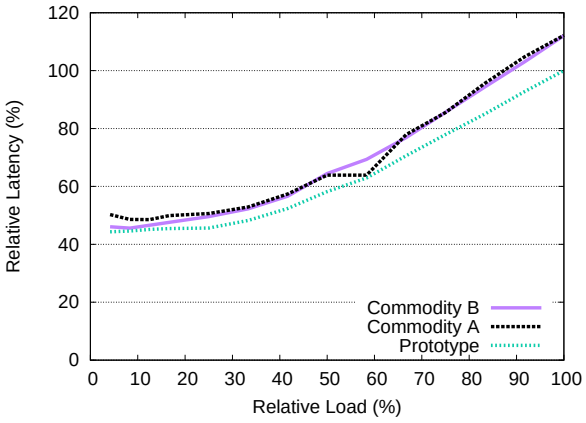

Figure 8: Mean response time (latency) comparison of prototype and commodity servers. All values are normalized to the load and latency of the prototype server at its saturation point.

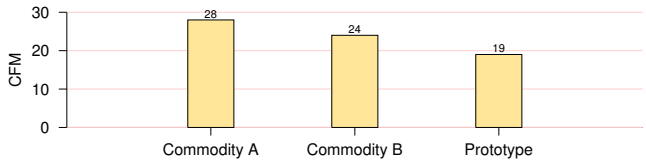

Figure 9: Airflow comparison (in CFM) at $200W$

Component	Capex	Opex
Power supply	$1 \sim 2\%$	$2 \sim 7\%$
Motherboard	6%	10%
Thermal	1%	$2 \sim 9\%$
Mechanical	1%	unavailable
Overall	$9 \sim 10\%$	$11 \sim 28\%$

Table 4: Efficiency gains by component *per server*, excluding datacenter-level gains. The opex values are derived from our measurements, Table 1, and [8].

and measures the airflow through the server, expressed in cubic-feet-per-minute (CFM). We also confirmed the measured CFM value analytically by measuring the server's AC power and temperature difference between inlet and outlet. We loaded the servers with an artificial load resembling our production power load (around $200W$, with leakage power at less than $10W$), while maintaining the constraint that all components remain within their operating thermal specifications. Our results (Fig. 9) show a significant improvement with our prototype. For a typical $7.5MW$ datacenter, this reduced airflow translates to a reduction of approximately $8 \sim 12\%$ of the cooling opex. More importantly, it enables free air cooling for our upcoming datacenter.

6.4 Acquisition Cost

On the capex front, a fair and direct comparison of the servers is not entirely feasible, since OEM pricing varies by many factors, and typically includes elements such as support, delivery, taxes, spare parts, etc. That said, our pricing data indicates a $\approx 10\%$ advantage of the ODM cost of building our own design over the OEMs' server pricing at scale.[2] As Table 4 shows, this figure agrees with the expected capex benefit of each component, summarized from the previous sections.

6.5 Discussion

These experiments represent perhaps a single data point in the large design space of prototype implementation choices, commodity choices, and workload,[3] but they exemplify how the benefits of our design can be realized in practice for actual gains. To account for implementation variations, we review and sum up the potential gains from each of the four design dimensions, focusing our attention on server-level gains. Table 4 normalizes the acquisition cost and efficiency gains of each of the previous four sections to a single-server basis, based on comparable commodity servers that we buy. The numbers should be taken more as order-of-magnitude guidance than precise measurements, since they vary based on the myriad underlying assumptions, such as workload, specific vendor and components, ambient environment, etc. Nevertheless, our experimental data falls within this range and validates our design. Again, these numbers reflect the savings at the server level only. As elaborated in the previous sections, combining this design with the datacenter's, the overall gains of the equipment and fully burdened power are even higher, with capex reduction of 24% and power savings of 38%.

We excluded performance from this discussion for multiple reasons. Most workloads' performance is constrained by components that are outside our design control—we can only choose them from the market (primarily, the CPU, with contributions from RAM and network). So our design space cannot affect performance much. But we also wanted to create a very general design. By constraining ourselves and fixing the performance-critical components, we can make the generalized argument that any application that performs well on these commodity parts and servers will perform the same, but with better TCO, on our new design.[4] This argument cannot be made, for example, for designs incorporating "wimpy" nodes. Nevertheless, one area for future research is to explore many different choices for CPUs and evaluate their effect on aggregate performance, efficiency, and thermal density.

Although the evaluation reported here is based on a single server, we observed later on that the power, temperature, and performance data scales nearly linearly to the rack and cluster level. Using our initial evaluations, we calculated that a large datacenter's power usage efficiency (PUE), based mostly on this customized server and evaporative cooling, to be less than 1.1. The actual PUE mea-

[2]Smaller-scale operators that lease computing services can also take advantage of this benefit if servers are provided by the larger-scale company that offers the lease. This is particularly true for large cloud and HPC datacenters.

[3]For example, the commodity system we chose has a high-end efficient PSU.

[4]For example, cloud applications that cannot always make many specific assumptions on the underlying hardware.

sured in our Prineville datacenter so far averages around 1.07—among the industry's best [11, 13]. But PUE calculations do not reveal the entire picture, since they do not account for improved server efficiency [24]. If PUE is defined as the ratio between the power going into the datacenter and the power going into the servers, more efficient servers subtract the same amount from both terms, actually increasing PUE. Many datacenters today are bound by power more than space, so it is likely that any efficiency gain will translate to additional servers (and throughput). In our case, the 16% more power-efficient servers translate to 19% of additional capacity without even taking single-server performance improvements into account. Since PUE calculations do not take throughput per watt into account, the improved efficiency is not reflected in the PUE number.

Anecdotally, we also weighed the prototype and Commodity A servers, and found the prototype system to be almost 6 pounds lighter. At our scale, this represents significant savings in raw materials, manufacturing cost, reliability, landfill waste, and even transportation resources.

7. CONCLUSIONS AND FUTURE WORK

This paper describes design changes to datacenter and HPC servers in four areas: power supply, motherboard, thermal and mechanical. Our main innovations include $277Vac$ power distribution with no PDUs; dual-input voltage PSUs with integrated UPS functions; custom dimensions optimized for cooling; and a distributed backup power solution that is offline, efficient, economical and either local or remote at choice. These changes measurably reduce TCO without reducing performance. Going back to the elements of TCO described in the introduction, we have seen how a customized server design can:

1. Reduce operating and cooling power (e.g., efficient power conversions, higher-quality power characteristics, fewer components, thinner and slower fans, improved airflow).

2. Lower the acquisition cost and server weight (e.g., fewer and simpler components, lower density, fewer expansion options).

3. Cut costs on supporting infrastructure (e.g., no centralized UPS, no PDUs, no chillers).

4. Increase overall reliability (e.g., fewer and simpler components, distributed and redundant batteries, smooth normal / backup transitions, staggered HDD startup, slower fans).

5. Improve serviceability (e.g., all-front service access, simpler cable management, no extraneous plastics or covers).

At large scale, this design translates to substantial savings. This year, we are deploying our first customized datacenter in Prineville, Oregon, to be populated primarily with servers that follow the design guidelines above. We calculate that over the course of the next three years, our upcoming datacenter servers will have at least 19% more throughput, cost approximately 10% less, and use several tons less raw materials to build than a comparable datacenter of the same power budget, populated with commodity servers. When this server design is matched with a corresponding datacenter design (including all aspects of cooling, power distribution, backup power, and rack design), the power savings grow to 38% and the cost savings to 24%, with a corresponding measured PUE of ≈ 1.07.

This synergy between servers and datacenter is our primary focus for future study. We plan to conduct thermal, power, and performance analysis of each of the datacenter components and report these results in a future publication, which will also include more details on the datacenter deployment. We also plan to experiment with design choices that do affect performance in limited ways and analyze the trade-offs involved (such as our work on Mecached performance [4]). It would also be interesting to collect reliability and serviceability metrics over time and compare them to our leased datacenters.

Focusing our attention on HPC clusters, we find numerous similarities in the server requirements to our own. Many supercomputers today consist of a large heterogeneous set of compute servers, often employing x86-based CPUs as well. Variations from our servers, such as a PCI-based accelerator card or a higher-end NIC can easily be accommodated. We therefore think that an HPC cluster based on our design principles and open specifications is plausible, and will result in lower overall power consumption. We plan to investigate this opportunity further.

Another direction for future work is to help commoditize our improvements by the opening of our design [25, 23]. Although our design reduces cost, it does not benefit from the commodity economies of scale. But this design is general enough and capable enough that many types of servers and applications can take advantage of it. As demand for these TCO-efficient datacenters grows, we may see designs like ours trickling back to OEMs, eventually becoming commodity products and closing the loop that led us away from commodity designs in the first place.

Acknowledgments

We would like to thank all the people who helped with the experiments and reviews: Victor Li, Manish Modi, Giovanni Coglitore, Frank Frankovsky, Jonathan Heiliger, Jimmy Langston, Harry Li, and Pete Bratach.

8. REFERENCES

[1] David G. Andersen, Jaosn Franklin, Michael Kaminsky, Amar Phanishayee, Lawrence Tan, and Vijay Vasudevan. FAWN: a fast array of wimpy nodes. In *Proceedings of the 22nd ACM SIGOPS Symposium on Operating Systems Principles (SOSP)*, pages 1–14, New York, NY, USA, 2009. ACM. `portal.acm.org/citation.cfm?id=1629577`.

[2] Luiz André Barroso and Urs Hölzle. The case for energy-proportional computing. *IEEE Computer*, 40(12):33–37, 2007. `citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.128.5419`.

[3] Christian L. Belady. In the data center, power and cooling costs more than the IT equipment it supports. *Electronics Cooling*, 23, February 2007. `www.electronics-cooling.com/2007/02/`.

[4] Mateusz Berezecki, Eitan Frachtenberg, Mike Paleczny, and Kenneth Steele. Many-core key-value store. In *Proceedings of the Second International Green Computing Conference*, Orlando, FL, August 2011. IEEE. `frachtenberg.org/eitan/pubs/`.

[5] Controlled Power Company. What is a K-rated transformer? `www.centralyacht.com/library/`

electrics/kratedtransformer.pdf, January 2005.

[6] Ismail Daut, Rosnazri Ali, and Soib Taib. Design of a single-phase rectifier with improved power factor and low THD using boost converter technique. *American Journal of Applied Sciences*, 3(9):1902–1904, September 2006. www.scipub.org/fulltext/ajas/ajas392025-2028.pdf.

[7] Facebook. The Open Compute server architecture specifications. www.opencompute.org, April 2011.

[8] Xiaobo Fan, Wolf dietrich Weber, and Luiz André Barroso. Power provisioning for a warehouse-sized computer. In *Proceedings of the 34th International Symposium on Computer Architecture (ISCA)*, San Diego, CA, June 2007. citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.84.9052.

[9] Dror G. Feitelson. Metrics for parallel job scheduling and their convergence. In Dror G. Feitelson and Larry Rudolph, editors, *Seventh Workshop on Job Scheduling Strategies for Parallel Processing*, volume 2221 of *Lecture Notes in Computer Science*, pages 188–1205. Springer Verlag, 2001. www.cs.huji.ac.il/~feit/parsched/.

[10] Eitan Frachtenberg and Dror G. Feitelson. Pitfalls in parallel job scheduling evaluation. In Dror G. Feitelson, Eitan Frachtenberg, Larry Rudolph, and Uwe Schwiegelshon, editors, *11th Workshop on Job Scheduling Strategies for Parallel Processing*, volume 3834 of *Lecture Notes in Computer Science*, pages 257–282. Springer-Verlag, Boston, MA, June 2005. frachtenberg.org/eitan/pubs/.

[11] Google. Quarterly PUE benchmark data. www.google.com/corporate/datacenter/efficiency-measurements.html, July 2011.

[12] James Hamilton. Cost of power in large-scale data centers. perspectives.mvdirona.com/2008/11/28/CostOfPowerInLargeScaleDataCenters.aspx, November 2008.

[13] James Hamilton. Cooperative expendable micro-slice servers (CEMS): Low cost, low power servers for internet-scale services. In *Fourth Biennial Conference on Innovative Data Systems Research (CIDR)*, Asilomar, CA, January 2009. citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.145.1407.

[14] James Hamilton. Internet-scale service infrastructure efficiency. In *Proceedings of the 36th International Symposium on Computer Architecture (ISCA)*, page 232. ACM, 2009. www.mvdirona.com/jrh/TalksAndPapers/JamesHamilton_ISCA2009.pdf.

[15] James Hamilton. Internet scale storage. In *ACM Special Interest Group on the Management of Data Annual Conference (SIGMOD'11)*, Athens, Greece, June 2011. ACM. Slides available at mvdirona.com/jrh/TalksAndPapers/JamesHamilton_Sigmod2011Keynote.pdf.

[16] Urs Hölzle. Brawny cores still beat wimpy cores, most of the time. *IEEE Micro*, 2010. static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/pubs/archive/36448.pdf.

[17] Mark Hydeman. Implications of current thermal guidelines for data center energy use. *ASHRAE journal*, pages 30–41, August 2010.

[18] IEEE. Authoritative dictionary of standards terms, 7th edition, 2000.

[19] Jonathan Koomey. Growth in data center electricity use 2005 to 2010. Technical report, Analytics Press, Oakland, CA, August 2011. www.analyticspress.com/datacenters.html.

[20] Andrew Krioukov, Prashanth Mohan, Sara Alspaugh, Laura Keys, David Culler, and Randy Katz. NapSAC: design and implementation of a power-proportional web cluster. *ACM SIGCOMM Computer Communication Review*, 41:102–108, 2011. ccr.sigcomm.org/drupal/files/p102-green12t-krioukovA.pdf.

[21] Willis Lang, Jignesh M. Patel, and Srinath Shankar. Wimpy node clusters: What about non-wimpy workloads? In *Proceedings of the Sixth International Workshop on Data Management on New Hardware (DaMoN'10)*, Indianapolis, IN, January 2010. pages.cs.wisc.edu/~jignesh/publ/nonwimpy.pdf.

[22] Kevin Lim, Parthasarathy Ranganathan, Jichuan Chang, Chandrakant Patel, Trevor Mudge, and Steven Reinhardt. Understanding and designing new server architectures for emerging warehouse-computing environments. *ACM SIGARCH Computuer Architecture News*, 36(3):315–326, 2008. portal.acm.org/citation.cfm?id=1382148.

[23] Rich Miller. Facebook opens its server, data center designs. www.datacenterknowledge.com/archives/2011/04/07/facebook-opens-its-server-data-center-designs/, April 2011.

[24] Rich Miller. Microsoft eliminates server fans, despite PUE hit. www.datacenterknowledge.com/archives/2011/01/31/microsoft-eliminates-server-fans-despite-pue-hit/, January 2011.

[25] Rich Miller. Will Open Compute alter the data center market? www.datacenterknowledge.com/archives/2011/04/14/will-open-compute-alter-the-data-center-market/, April 2011.

[26] Jennifer Mitchell-Jackson, Jonathan G. Koomey, Bruce Nordman, and Michele Blazek. Data center power requirements: Measurements from silicon valley. *Energy*, 28(8):837–850, June 2003. citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.132.254&rep=rep1&type=pdf.

[27] Eduardo Pinheiro, Wolf-Dietrich Weber, and Luiz André Barroso. Failure trends in a large disk drive population. In *Proceedings of the Fifth USENIX Conference on File and Storage Technologies (FAST'07)*, pages 17–28. USENIX Association, 2007. www.usenix.org/events/fast07/tech/full_papers/pinheiro/pinheiro_html/.

[28] Asfandyar Qureshi. Plugging into energy market diversity. In *Proceedings of the Seventh ACM Workshop on Hot Topics in Networks (HotNets)*, October 2008. citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.161.8738.

[29] Vijay Janapa Reddi, Benjamin C. Lee, Trishul Chilimbi, and Kushagra Vaid. Web search using mobile cores: Quantifying and mitigating the price of efficiency. In *Proceedings of the 37th International Symposium on Computer Architecture (ISCA)*. ACM, June 2010. portal.acm.org/citation.cfm?id=1815961.1816002.

[30] Pierluigi Sarti. *450W Power Supply Hardware 1.0*, April 2011. opencompute.org/specs/Open_Compute_Project_Power_Supply_v1.0.pdf.

[31] ab - Apache HTTP server benchmarking tool. httpd.apache.org/docs/2.0/programs/ab.html.

[32] Climate savers platinum standard. www.climatesaverscomputing.org.