Design Principles in the Open Compute Project

Eitan Frachtenberg

Facebook, 1601 South California Avenue, Palo Alto, California, 94304, USA

etc@fb.com

Abstract: The Open Compute Project aims to capture the best principles in datacenter design and open them for third-party implementation and discussion. This paper summarizes them in the areas of electrical, thermal, building, and server design.

1. Introduction

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 have 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 and deployed alternative, more efficient designs for servers and datacenters. We have previously presented in detail the design specifications for our servers [1], as well as the rationale and performance of this design [2]. In this paper, we overview the principles and considerations that led to our choices in both server and datacenter design.

2. Electrical Design Principles

One of the most important decisions in the Open Compute Project (OCP) was to forgo standard datacenter power distribution at $208Vac$, and instead use the original $277Vac$ (lines-to-neutral) from the distribution station (see Fig. 1). The primary motivation for this choice was to reduce inefficient power transformations and their associated equipment. But it also required and enabled important changes in other electrical design choices. For example, for the backup power scheme we preferred a distributed offline scheme to the traditional inline centralized UPS (Fig. 1). Offline batteries reduced transformation waste, because they require little current while fully charged. However, they also necessitated a redesigned power supply unit (PSU) in the servers, to take both $277Vac$ as normal input and $48Vdc$ in backup operation. Such PSUs were only available for lower-power applications at the time, so we had to design our own specifications[1]. But having designed our own, we were able to obtain a high efficiency of over 94% throughout most load inputs while still reducing component cost [2]. Another aspect of the power distribution that we customized was the remote power panels (RPPs). Most rows in our datacenter include our own $277Vac$ servers and battery cabinets, so the RPP at each of these rows distributes power without transformation and with minimal loss. But to allow the flexibility to use the occasional legacy or commodity equipment, we also have a few rows where the RPP converts power down to $208Vac$.

3. Thermal Design Principles

The guiding principle in designing the thermal management of the datacenter building was to use 100% outside-air economization (See Fig. 2). Outside-air cooling offers significant power advantages over chillers but imposes

[1] Available at: http://opencompute.org/projects/power-supply/

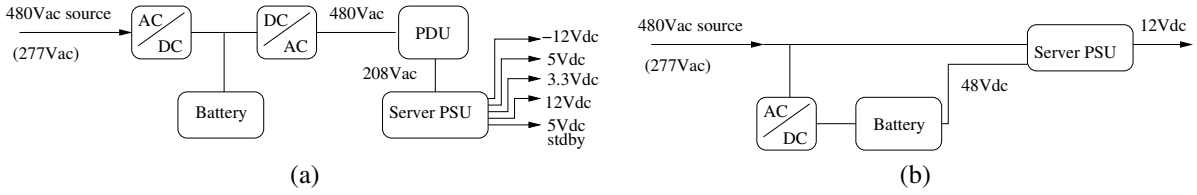

Figure 1. Schematic view of power distribution to the server. Inside the typical datacenter (a), an external three-phase 480*Vac* source provides power to large UPSs, consisting of rectifiers to convert AC to DC, a battery to store the energy, and an inverter to convert back to AC. Voltage is then transformed to 208*Vac* 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, 277*Vac* power from the AC source flows through RPPs 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 $48Vdc$.

restrictions on the site selection for the datacenter, since it works best with cool, dry air. (This constraint is one of the reasons we chose Prineville, Oregon, as the location for our first datacenter.) Two other important factors in our cooling scheme were letting the temperature and moisture limits in the datacenter to exceed ASHRAE's most liberal recommendations, and the high power efficiency of our servers [2], which consequently produce less heat. We achieved additional efficiency by containing the hot aisles and employing a drop-ceiling plenum for air return, which prevents the hot and cold air mixing that is typical in traditional datacenters without containment.

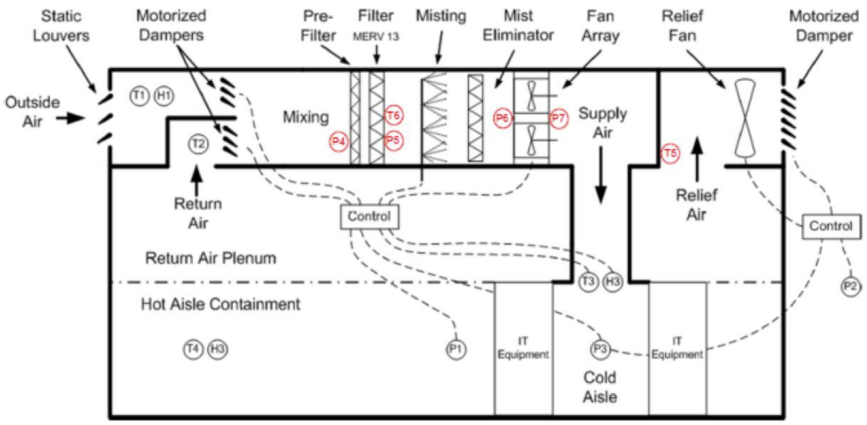

Figure 2. Schematic view of air flow throughout the datacenter. Outside air is pulled in at the top level of the datacenter and cooled by misting, as necessary. Excess moisture is screened out before the cold air is allowed to drop into the cold aisles of the compute floor. The air is pressurized through the server racks to the back, hot aisles (that can heated above human comfort levels because they require no servicing). The hot air is dumped outside or mixed with incoming air when outside temperatures are too cold. Excess heat can also be used for the office space in the winter. Distributed sensors monitor the environment and adaptively control temperature, moisture, and air pressure in real time.

4. Building Design Principles

As a consequence of the thermal design described in the previous section, the datacenter building requires no ductwork or cooling pipelines for the data hall, and subsequently eliminates the need for access flooring. Thus, we were able to tile the datacenter floors with concrete slab-on-grade. This choice not only reduces cost but also has higher bearing

capacity and enables increased building height. Because we control the building's humidity and can keep it high enough, we were also able to forgo static dissipation flooring.

The ceiling uses lightweight composite deck concrete instead of the traditional corrugated metal, increasing the diaphragm load and environmental protection. The stiffer ceiling permitted structural changes that yielded more open space, with strategically located diagonal braces. The lighting system uses power-over-ethernet for reduced power consumption and cabling. Each fixture also includes diagnostics: temperature for cooling efficiency, as well as luminosity, lamp life, and occupancy sensors for automatic adjustment of its output.

Since we use custom-sized racks with battery cabinets in between at fixed distances, the weight-bearing columns in the building are aligned behind the battery cabinets (in the hot aisles), minimizing impedance to airflow behind the servers that produce the heat. We also separated the office area as an attachment to the compute area (as opposed to being carved out of it) for optimized airflow and to provide it its own, more lenient environmental requirements.

5. Server Design Principles

One area where we made significant departures from commodity offerings is in server design. Our servers employ not only customized dimensions optimized for cooling, but also customized motherboard, power supply, and mechanical design. The three main design principles were: (1) prefer efficiency over aesthetics; (2) optimize for high-impact use cases instead of generality of application; and (3) maximize serviceability and limit it to the front of the server. The following examples show how we applied these principles, with more details provided in previous work [2].

- We designed high-efficiency PSUs (as described in Sec. 2) with both $277Vac$ and $48Vdc$ inputs and a single $12Vdc$ output. The PSU has special provisions for smooth startup, as well as transitioning from normal to backup power and back. This PSU results in a power efficiency gain in the range of $13 \sim 25\%$ while retaining cost competitiveness with commodity PSUs.

- The motherboards were designed for increased efficiency, reduced cost, and improved airflow. We removed any extraneous components that are not used in our applications but are provided by commodity vendors to address the flexibility requirements of a large market, such as BMC, SAS controllers, serial and graphic outputs, etc. We provided low-cost and low-power alternatives to manage the servers instead.

- The side-by-side motherboard layout for the dual processors, combined with a custom-sized chassis ($1.5U$ height), high efficiency fans and HDDs located in the back, provide optimal airflow and cooling, and resulted in substantially improved cooling efficiency at less power than commodity servers.

- "Vanity-free" functional design: no resources were spent on appearance—no stickers, paint, plastic bezels, or face plates. Virtually all components were designed for quick-release servicing without screws.

- Each rack contains three columns of 30 servers each, together with two 48-port switches. We thus amortize some of the rack's cost and improve the server-to-switch ratio. The switches fit in a customized quick-release tray that can accommodate any standard $19"/1U$ equipment.

- All cables, including network and power, are accessed from the front and 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.

6. Conclusion

Going the custom route is not always a feasible choice. But when the scale justifies it, there are many choices in datacenter design that can evince significant capex and opex savings. The principles summarized here resulted in a datacenter that is 24% less expensive to equip and uses 38% less power to run than a traditional datacenter with the same compute capacity. In addition, the average power usage effectiveness (PUE) measured over the first year of operation resulted in an average of 1.07, arguably the lowest sustained PUE recorded to date.

References

1. Facebook. The Open Compute server architecture specifications. www.opencompute.org, April 2011.
2. Eitan Frachtenberg, Ali Heydari, Harry Li, Amir Michael, Jacob Na, Avery Nisbet, and Pierluigi Sarti. High-efficiency server design. In *Proceedings of the 24th IEEE/ACM International Conference on High Performance Computing, Networking, Storage and Analysis (SC)*, Seattle, WA, November 2011.