Power to the People

REDUCING DATACENTER CARBON FOOTPRINTS

JESSIE FRAZELLE

When you upload photos to Instagram, back up your phone to the cloud, send an email through Gmail, or save a document in a storage application like Dropbox or Google Drive, your data is being saved in a datacenter. These datacenters are airplane hangar-sized warehouses, packed to the brim with racks of servers and cooling mechanisms. Depending on the application you are using, you are likely hitting one of the datacenters operated by Facebook, Google, Amazon, or Microsoft. Aside from those major players, which I refer to as *hyperscalers*, many other companies run their own datacenters or rent space from a colocation center to house their server racks.

CARBON FOOTPRINTS

Most of the hyperscalers have made massive strides toward achieving carbon-neutral footprints for their datacenters. Google, Amazon, and Microsoft have pledged to decarbonize completely; however, none has yet succeeded in that quest.

If a company claims to be carbon neutral, this usually means it is offsetting its use of fossil fuels with RECs (renewable energy credits). A REC represents one MWh (megawatt-hour) of electricity that is generated and delivered to the electrical grid from a renewable energy resource such as solar or wind power. By purchasing RECs, carbon-neutral companies are essentially giving back clean energy to prevent someone else from emitting carbon. Most companies become carbon neutral by investing in *offsets* that primarily avoid emissions, such as paying people not to cut down trees or buying RECs. These offsets do not actually remove the carbon that the companies are emitting.

A *net zero* company actually has to remove as much carbon as it emits. Though the company is still creating carbon emissions, those emissions are equal to the amount of carbon the company removes.

If a company calls itself *carbon negative,* it is removing more carbon than it emits each year. This should be the gold standard for how companies operate. None of the FAANG (Facebook, Apple, Amazon, Netflix and Google) today claim to be carbon negative, but Microsoft issued a press release stating it would be by 2030.

POWER USAGE EFFICIENCY

PUE, or power usage efficiency, is defined as the total energy required to power a datacenter (including lights and cooling) divided by the energy used for servers. A perfect PUE would be 1.0, since 100 percent of electricity consumption would be used on computation. Conventional datacenters have a PUE of about 2.0, while hyperscalers have gotten theirs down to about 1.2. According to a 2019 study from the Uptime Institute, which surveyed 1,600 datacenters, the average PUE was 1.67.

PUE as a method of measurement is a point of contention. PUE does not account for location, which means a datacenter that is located in a part of the world that can benefit from free cooling from outside air will have a lower PUE than one in a very hot climate. PUE should be measured as an annual average since seasons change and affect the cooling needs of a datacenter over the course of a year. According to a study from the University of Leeds, "comparing a PUE value of datacenters is somewhat meaningless unless it is known whether it is operating at full capacity or not."

Google claims an average yearly PUE of 1.1 for all its datacenters, while individually some are as low as 1.08. One of the actions Google has taken for lowering its PUE is using machine learning to cool datacenters with inputs from local weather and other factors—for example, if the weather outside is cool enough the datacenter can use it without modification as free cold air. It can also predict wind-farm output up to 36 hours in advance. Google took all the data it had from sensors in its facilities monitoring temperature, power, pressure, and other resources to create neural networks to predict future PUE, temperature, and pressure in its datacenters. This way Google can automate and recommend actions for keeping its datacenters operating efficiently from the predictions. Google also sets the temperature of its datacenters to 80°F, rather than the usual 68-70°F, saving a lot of power for cooling. Weather local to the datacenter is a huge factor. For example, Google's Singapore datacenter has the highest PUE and is the least efficient of its sites because Singapore is hot and humid year-round.

According to a study from the University of Leeds, "comparing a PUE value of datacenters is somewhat meaningless unless it is known whether it is operating at full capacity or not."

Wired conducted an analysis of how Google, Microsoft, and Amazon stack up when comparing the carbon footprints of their datacenters. Google claims to be net zero for carbon emissions and publishes a transparency report of its PUE every year. While Microsoft claims it will be carbon negative by 2030, it is still carbon neutral today. It also claims to be pursuing 100 percent renewable energy by 2025.

Amazon, on the other hand, has the worst carbon footprint of the large tech companies. As noted previously, the location of the datacenter matters, so some Amazon regions might be greener than others because of the weather conditions in those areas or having more access to solar or wind energy. Amazon founder and CEO Jeff Bezos has pledged to get to net zero by 2040. Greenpeace seems to believe otherwise, claiming in a 2019 report that Amazon is not dedicated to that pledge since its Virginia datacenters were at only 12 percent renewable energy.

In 2018, Apple claimed 100 percent of its energy was from renewable sources. Facebook claims it will be at 100 percent renewable energy by the end of 2020. While U.S. companies have followed suit on pledging to lower their carbon footprints, Chinese Internet giants such as Baidu, Tencent, and Alibaba have not.

WHAT IS USING POWER IN A DATACENTER?

According to a study from Procedia Environmental Sciences, 48 percent of the power in a datacenter goes to equipment such as servers and racks, 33 percent to HVAC (heating, ventilation, and air conditioning), 8 percent to UPS (uninterrupted power supply) losses, 3 percent to lighting, and 10 percent to everything else.

HVAC requires a delicate process of making sure hot air from server exhaust doesn't mix with cool air and raise the temperature of the entire datacenter. This is why most datacenters have hot and cold aisles. The goal is to have the cold air flow into one side of the racks, while the hot air exhaust comes out the other side. Optimizing air flow throughout the racks and servers is essential for HVAC efficiency.

Power comes off the grid as AC power. This can be single-phase, which has two wires (a power wire and a neutral wire); or three-phase, which has three wires, each 120 electrical degrees out of phase with each other. The key difference between the two is that three-phase power can handle higher loads than single-phase. The frequency of the power off the grid can be either 50 or 60Hz. Voltage is any of the following: 208, 240, 277, 400, 415, 480, or 600V.

Since most equipment in a datacenter uses DC power, the AC power needs to be converted. This results in power losses and wasted energy adding up to around 21-27 percent. To break this down, there is a 2 percent loss when utility medium voltage, defined as greater than 1000V and less than 100 kV, is transformed to 480VAC; a 6-12 percent loss within a centralized UPS because of conversions from AC to DC and DC back to AC; and a 3 percent power loss at the PDU (power distribution unit) level resulting from the transformation from 480VAC to 208VAC. Standard power supplies for servers convert 208VAC to the required DC voltage, resulting in a 10 percent loss, assuming the power supply is 90 percent efficient. This is all to say that power is wasted throughout traditional datacenters in transformations and conversions.

In an attempt to lessen the amount of wasted power from conversions, some people rely on high-voltage DC power distribution. The Lawrence Berkeley National Lab conducted a study in 2008 in which the use of 380VDC power distribution for a facility was compared with a traditional 480VAC power-distribution system. The results showed that the facility using DC power eliminated multiple conversion stages, resulting in a 7 percent decrease in energy consumption compared with a typical facility with AC power distribution. This is rarely done at hyperscale, however. Hyperscalers tend to have three-phase AC going to the rack, then convert to DC at the rack or server level.

In an attempt to lessen the amount of wasted power from conversions, some people rely on high-voltage DC power distribution.

MORE POWER-EFFICIENT COMPUTE

In addition to RECs and using 100 percent renewable energy, there are other ways hyperscalers have made their datacenters more power efficient. In 2011, the Open Compute Project started out of a basement lab in Facebook's Palo Alto headquarters. Its mission was to design from a clean slate the most efficient and economical way to run compute at scale. This led to using a 480VAC electrical distribution system to reduce energy loss, removing anything in the servers that didn't contribute to efficiency, reusing hot aisle air in winter to heat the offices and the outside air flowing into the datacenter, and removing the need for a central power supply. The Facebook team installed the newly designed servers in the Prineville datacenter, which resulted in 38 percent less energy to do the same work as the existing datacenters. It also cost 24 percent less.

Let's dive into some of the details of the Open Compute designs that allow for power efficiency. The Open Rack design includes a power-bus bar with either 12VDC or 48VDC of distributed power to the nodes. The bus bar runs along the back of the rack vertically. It transmits power from the rack-level PSUs (power supply units) to the servers in the rack. The bus bar allows the servers to plug in directly to the rack for power, so when you service an Open Rack you do not need to unplug power cords; you can just pull the server out from the front of the rack. With the Open Compute designs, network connections to servers are at the front of the rack so the technician never has to go to the back of the rack (i.e., the hot aisle).

REDUNDANCY

Conventional designs have PSUs in every server. The Open Rack design has centralized PSUs for the rack, which allow for N+M redundancy, the most common deployment being N+1. This means there is an extra PSU per rack of servers. In a conventional system this would be 1+1 since there is one extra PSU in every individual server. Keeping the PSUs centralized to the rack reduces the number of power-converting components; this increases the efficiency of the system.

RIGHT-SIZED PSUS

Server designers tend to choose PSUs that have enough headroom to deliver power for the maximum configuration. Server vendors would rather carry a small number of oversized power-supply SKUs than a large number that are right-sized to purpose, since economies of scale prefer the former. This leads to an oversizing factor of at least two to three times the required capacity for conventional power supplies. In comparison, a rack-level PSU will be less oversized since it is right-sized for purpose. The hyperscalers also have the advantage of economies of scale for their hardware. The typical Open Rack-compliant power supply is oversized at only 1.2 times the required capacity, if that.

OPTIMAL EFFICIENCY

Every power supply has a sweet spot for load versus efficiency. The 80 Plus certification program measures PSU efficiency using these different grades: bronze, silver, gold, platinum, and titanium. The most power-efficient grade is titanium. The most common grade of PSU used in datacenters is silver, which has a maximum efficiency of 88 percent, meaning it wastes 12 percent electric energy as heat at the various load levels. In comparison, the 12V and 48VDC PSUs have data showing maximum efficiencies at 95 percent and 98 percent, respectively. This means the rack-level PSUs waste only between 5 and 2 percent of energy.

While the efficiency of the rack-level PSU is important, you still need to weigh the cost of the number of conversions being made to get the power to each server. For every unnecessary power conversion, you are paying an efficiency cost. For example, with a 48VDC rack-level power supply, the server might need to convert the rack provided from 48VDC to 12VDC, then that 12VDC to V_{CORE}. V_{CORE} is the voltage supplied to the CPU, GPU, or other processing core. With its 48VDC power supply, Google advocates for using 48V to PoL (point of load) to deliver power to the servers. This means placing a DC-to-DC or linear power-supply regulator going from the rack-level PSU to the server, which would reduce the number of conversions needed to get the power to the processing cores. The 48VDC-to-DC regulators required for Google's implementation, however, are not common and come at a premium cost. It is likely that Google's motivation for opening the specs for the 48VDC rack is to drive more volume to those parts and thus drive down costs. In contrast, 12VDC-to-DC regulators are quite common and low cost.

READING A POWER-EFFICIENCY GRAPH

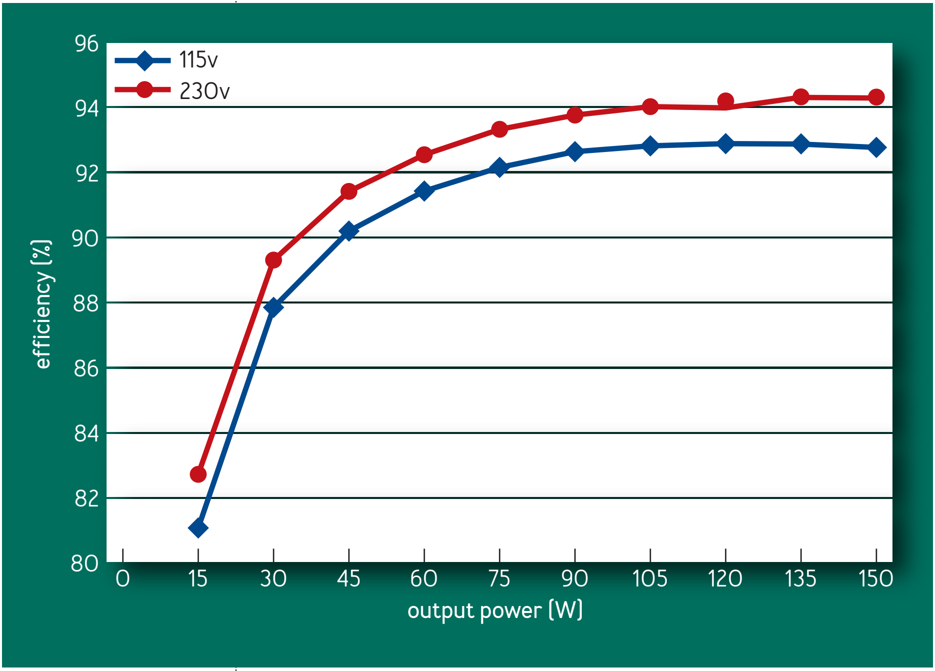
Figure 1 is an example of a power-efficiency graph for a power supply. You can see that the peak of the graph is where the PSU is the most efficient. Divide the output power by the input power to calculate efficiency. The x-axis of the graph measures the load of the power supply in watts, while the y-axis measures efficiency.

If you know the peak load is 120W and idle is 60W, as shown in figure 1, then this power supply would be more than is needed since it can handle up to 150W. At a peak load of 120W with 230VAC, this power supply would have a maximum efficiency of around 94 percent and a minimum efficiency at idle of around 92 percent with 230VAC. You now know the losses of this specific power supply and can compare it with other supplies to see if they are more efficient. This allows you to choose the right power supply for the load.

OPEN COMPUTE SERVERS WITHOUT A BUS BAR

Not all Open Compute servers include a power-bus bar. Microsoft's Olympus servers require AC power. The Olympus power supply has three 340W power-supply modules, one for each phase, with a total maximum output of 1,000W. Therefore, these power supplies assume

FIGURE 1: **POWER EFFICIENCY GRAPH EXAMPLE**



all deployments are three-phase power. The minimum efficiency of the PSU is 89-94 percent, depending on the load. This places the grade of the Olympus power supply around an 80 Plus platinum.

Like all technical decisions, using per-server AC power supplies versus rack-level DC is a tradeoff. By having separate power supplies, different workloads can balance the power they are consuming individually rather than at a rack level. In turn, though, Microsoft needs to build and manufacture multiple power supplies to ensure they are right-sized to run at maximum efficiency for each server configuration. Serviceability also requires technicians to unplug power cables and go to the back of the rack.

At the time Microsoft made the decision to use individual AC power supplies per server, the Open Rack design was at v1 (not v2 like it is today), the cost of the copper for the power-bus bar was higher, and the loss of efficiency to resistance was a factor. The Open Rack v1 design had an efficiency concern with power loss resulting from heating the copper in the bus bar. If a rack holds 24 kW of equipment, a 12VDC power-bus bar must deliver 2 kA of current. This requires a very thick piece of copper, which has a significant power loss because of resistance in the bus bar.

Let's break down how to measure the relationship of power to resistance. Ohm's law declares electric current (I) is proportional to voltage (V) and inversely proportional to resistance (R), so $V=IR$. To see the relationship of power to resistance, combine Ohm's law ($V=IR$) with $P=IV$, which translates to power (P) being the product of current (I) and voltage (V). Substituting $I=V/R$ gives $\mathrm{P}=(V/R)V=V^2/R$. Then, substituting $V=IR$ gives $P=I(IR)=I^2R$. So, $P=I^2R$ is how you can calculate the power loss resulting from resistance in the bus bar.

In making its decision, Microsoft balanced the conversion efficiency against the material cost of the bus bar and the resistive loss. Open Rack v2, however, changes the tradeoffs of the original decision. With a 48VDC bus bar, a rack that holds 24 kW of equipment requires only 500A, as opposed to the 2kA required by the 12VDC power-bus bar from the v1 spec. This translates into a much cheaper bus bar and lower losses from resistance. The bus bar still has more loss than 208VAC cables, but there is an improved efficiency from the power-supply unit at the rack level, which makes it compelling. As stated earlier, however, you need to be mindful of the number of conversions needed to get the power to the components on the motherboard. If your existing equipment is 12VDC, you would want to avoid any extra conversions using that with a 48VDC bus bar. Save the 48VDC bus bar for new equipment that has 48V to PoL to avoid extra conversions.

The main difference between Microsoft's design with individual power supplies and the 24VDC and 48VDC Open Rack designs is the way the initial power is delivered to the servers. Microsoft's design distributes three-phase power to the servers individually through power supplies, while the 24VDC and 48VDC power-bus bars distribute the power delivery to the servers. Once power is delivered to the server, it is typically sent through a DC-to-DC power-supply regulator, which in turn powers the components on the motherboard. This step is shared whether the power is coming from a single power-bus bar or individual power supplies.

Another interesting bit comes into play with UPSes. As noted earlier, there are losses in efficiency because of UPSes. What does this mean in terms of a DC bus bar or individual AC PSUs? When AC power is going into each individual server, you have two choices: a UPS on the AC before it gets distributed to the individual servers, or a UPS per server integrated into each server's PSU. Deploying and servicing individual batteries per server is a nightmare for maintenance. Because of this, most facilities that use AC power to the servers wind up using rack-wide or building-wide UPSes. Since the batteries in a UPS are

DC, an AC UPS has an AC-to-DC converter for charging the batteries and a DC-to-AC inverter to provide AC power from the battery. For online UPSes, meaning the battery is always connected, this requires two extra conversions from AC to DC, and DC back to AC, with power-efficiency losses for both.

With a DC rack-level design, battery packs can be attached directly to the bus bar. The rack-level PSUs are the first AC-to-DC conversion state so there is no need for another conversion since everything from there runs on DC. The downside is that the rack-level PSU needs to adjust the voltage level to act as a battery charger. This means the servers need to accept a fairly wide tolerance on the 48V target, around +/-10V, so 40-56V isn't unreasonable. Because DC-to-DC converters are fairly tolerant about input voltage ranges, this is fairly straightforward, without any significant loss in power efficiency. It's important to note that for hyperscalers UPSes are present only to allow for a generator to kick in—a few seconds rather than 10-15 minutes for a traditional datacenter.

With commodity servers, such as Dell or Supermicro, the cost of individual power supplies is much higher in terms of power efficiency since those PSUs do not have as high an 80 Plus grade and do have much more oversizing. They also tend to lack power-supply regulators that minimize power-conversion losses in supplying power to the components on the board. This would lead to around an 8-12 percent gain in power efficiency by moving from a bunch of commodity servers in a rack to an Open Compute Project design—not to mention that the serviceability ease of the bus bar would benefit technicians as well.

Related articles

Cooling the Data Center
What can be done to make cooling systems in data centers more energy efficient?
Andy Woods
https://queue.acm.org/detail.cfm?id=1737963

Virtualization: Blessing or Curse?
Managing virtualization at a large scale is fraught with hidden challenges.
Evangelos Kotsovinos
https://queue.acm.org/detail.cfm?id=1889916

Words Fail Them
Dedesignating and other linguistic hazards
Stan Kelly-Bootle
https://queue.acm.org/detail.cfm?id=1569209

By designing rack-level architectures, huge improvements can be made for power efficiency over conventional servers, since PSUs will be less oversized, more consolidated, and redundant for the rack versus per server. While the hyperscalers have benefited from these gains in power efficiency, most of the industry is still waiting. The Open Compute Project was started as an effort to allow other companies running datacenters to benefit from the power efficiencies as well. If more organizations run rack-scale architectures in their datacenters, the wasted carbon emissions caused by conventional servers can be lessened.

Acknowledgments

Huge thanks to Rick Altherr, Amir Michael, Kenneth Finnegan, Arjen Roodselaar, and Scott Andreas for their help with the nuances in this article.

Jessie Frazelle *is the cofounder and chief product officer of the Oxide Computer Company. Before that, she worked on various parts of Linux, including containers, and the Go programming language.*