DATA CENTER Frontier Special Report

The Rise of the Sustainable Data Center

By Paul Gillin

Photo: metamorworks / shutterstock.com

brought to you by

VERTIV

Contents

The Rise of the Sustainable Data Center

Sustainability is now a top-of-mind issue in data center construction and management, as cloud computing and digital transformation fuel the world's growing thirst for processing power. Energy Innovation Policy & Technology reports that global IP traffic increased ten-fold between 2010 and 2018, the number of compute instances running on the world's servers increased more than six-fold and data center storage capacity skyrocketed by a factor of 25. Some of the world's largest data centers now consume enough electricity to power 80,000 homes.

And workloads are only going to grow more. A profusion of GPUs and CPUs are being deployed to support a burgeoning number of artificial intelligence applications. Smart devices equipped with 5G wireless networks will give rise to smaller but numerous edge data centers.

The growth of connected devices is driving not only demand for power to operate them but also data centers that connect to them. Vertiv estimates that the move to 5G is likely to increase total network energy consumption by between 150% and 170% by 2026, with the largest increases coming in macro, node and network data center areas. One 2017 study forecast that the communications industry could use 20% of the world's electricity by 2025 and generate 14% of global emissions by 2040, or about as much as the entire U.S. does today.

All of this is occurring against the backdrop of growing global concerns about keeping water supplies safe and limiting the use of fossil fuels. To some degree there is a trade-off between the two. About 40% of the power consumed by data centers goes to the air-conditioning used to cool equipment. That figure can be significantly reduced through the use of evaporative cooling technologies, but the gains come at the expense of increased water usage and additional costs and steps needed to keep water supplies clean. Achieving the most sustainable solution requires balancing both variables.

10X INCREASE
Global IP traffic increased ten-fold between 2010 and 2018

6X INCREASE
The number of compute instances running on the world's servers increased more than six-fold

25X CAPACITY
Data center storage capacity skyrocketed by a factor of 25

80,000 HOMES
Some of the world's largest data centers now consume enough electricity to power 80,000 homes

New Ideas About Sustainability

Hyperscale data center operators have broadly embraced the goal of sustainable operations and have contributed some important innovations while significantly reducing their own power use. For example, Google's data centers today support an average of seven times more computing power than they did five years ago but use no more electricity. Data center electricity consumption increased just 4% between 2010 and 2014, far below the 24% increase of the previous five years, according to the U.S. Data Center Energy Usage Report. Energy use is expected to increase at an even slower rate in the future.

Part of this slowing growth is due to the widespread use of server virtualization, which has reduced the number of servers needed in data centers while driving up utilization rates. Hyperscalers have also been leaders in the use of power purchase agreements with renewable energy providers to create a cost-effective alternative to building out their own renewable infrastructure while funding investments in sustainable sources. Innovations like Google's "carbon-intelligent" data center design shifts workloads automatically to maximize the use of carbon-neutral power sources.

Voltage fluctuations and dropouts of as little as 30 milliseconds can damage IT equipment, triggering expensive outages.

But despite the industry's voracious appetite for power, data center operators for the most part have moved cautiously to adopt alternative energy sources on a large scale. This does not reflect reticence their part but rather the need to balance sustainability with resiliency.

Data centers require power that is both high-quality and predictable. Voltage fluctuations and dropouts of as little as 30 milliseconds can damage IT equipment, triggering expensive outages. Statista reported that the median cost of enterprise server downtime in 2019 was between $400,000 and $500,000 with 15% of enterprises reporting costs of greater than $5 million. For operators of multi-tenant data centers the potential costs are much higher, including lost business, penalties for failure to meet service level agreements and legal exposure.

UPSes not only protect against outages but eliminate the fluctuations of power taken directly from the grid.

Sustainable energy sources such as solar and wind are inherently unpredictable because of their dependence upon weather conditions. Fuel cells show promise, but they are currently expensive and immature. Solar and wind can be useful as an ancillary power supply but the cost of storing enough power to insure against disruptions—combined with the cost of siting, installing and maintaining equipment—is so high as to be impractical for many operators. Power purchase agreements can provide many of the same benefits without the high capital costs.

The need for stable and clean power is one of the reasons most data centers maintain multiple uninterruptible power supplies. UPSes not only protect against outages but eliminate the fluctuations of power taken directly from the grid. Unfortunately, most data center UPSes are under-utilized by design. The units are most efficient when operating at about 80% capacity, but data center operators typically run them at no more than 20% capacity to ensure availability in the event of an outage. That adds up to a substantial amount of battery capacity that is effectively wasted.

As Kevin Hagen, Iron Mountain's vice president of environment, social, and governance strategy, told Data Center Knowledge, UPSes batteries are "useless capital. We're spending millions of dollars of insurance policy just sitting there in the hope we don't use it. That's a terrible use of money."

This unused capacity could potentially be a resource to stabilize the electrical grid, boost usage of sustainable energy and create financial benefits for both data center operators and the utilities from which they buy their power. Unused UPSes battery capacity could be used to make sustainable energy sources more practical and even improve the overall health of the energy grid.

A Smarter Grid

The grid that provides most of the electrical power in the U.S. was designed in the 1890s and has been incrementally improved upon, but it is limited by its aging design. Power needs in the home were modest in the days the grid was designed. No one conceived of 500-watt PCs, electric ovens and battery-powered cars, much less 100-MEGAWATT data centers.

The grid was designed to be a one-way path from the utility to the customer, who was billed monthly. The scope of power outages has been steadily reduced through sub-zoning, but failures today still have the potential to affect a lot of people. Perhaps the greatest shortcoming of the current power grid is its "one-wayness". It was designed as a conduit between a utility and its customers. While the network has been upgraded over time to permit two-way transmission, there is still no easy way for customers to exchange electricity with each other. This has increasingly become a liability as many homes and businesses have become miniature power stations, often generating more electricity than they need.

The US Department of Energy's Smart Grid initiative reimagines the power grid in a manner similar to the Internet. The 9,200 electric generating units on the current grid would be complemented by thousands of switching stations that intelligently route power between nodes as needed. Customers could form collectives to share the power they generate to lower costs or improve reliability. Blackouts would last minutes instead of hours. The overall resilience of the grid would improve and sustainable sources become more practical to use.

Researchers at leading data center technology companies are developing technologies that can make UPSes smart nodes on the energy grid. The 80% of the average data center UPSes that go unused could become a node on the network, offering temporary power storage and provisioning. UPS batteries could not only serve as backup power sources in case of an outage, but because UPSes inherently deliver a more stable flow of power, they could help to compensate for fluctuations in power levels and quality across an entire grid.

For example, the UPS battery capacity needed to power a 15-MEGAWATT data center for five minutes could supply electricity to approximately 15,000 homes for the same duration. While five minutes doesn't seem like a long time, it is more than enough for utilities to rebalance capacity during periods of high demand or disruption. Power could be rerouted through the smart grid so quickly that most consumers wouldn't even see their lights flicker. UPSes can also mitigate the impact of power surges that damage delicate electronic equipment, enhancing the overall quality of the grid.

Data Center UPSes

A Smart Grid that includes data center UPSes has the potential to be a win-win-win solution for all parties.

- **Utilities** would gain greater reliability across their service area, suffer fewer outages and lower the cost of emergency repairs. Expanding the scope of power sources to include businesses and even homes would reduce the need for the construction of new generating facilities.
- **Data center operators** could charge for the use of their UPSes and benefit from tax breaks and favorable tariffs. They would have greater visibility into costs and the potential to reduce power expenses through supplier arbitrage. There would be more incentive for operators to invest in their own sustainable energy sources thanks to improved power predictability and the ability to sell the electricity they generate to a wider range of customers. Multiple data center operators could even pool resources to reduce their overall UPS needs and share power more cheaply with each other.
- **Everyone on the grid** would enjoy a more stable and reliable supply of power, a wider range of buying options and less risk of damage to electronic equipment from power fluctuations.

There are currently only a small number of data center operators engaged in smart grid experiments, most of them in Europe. However, improvements in technology, combined with lower costs will make this option widely available within a few years. The data center technology companies that are leading the charge will bring these benefits to their customers first. Some are even developing technologies that will enable existing UPSes to be retrofitted as smart hubs on the smart grid.

Tackling Water Usage

The seven-year-long California drought that ended in early 2019 and the wildfires that ensued are just two recent events that have cast a spotlight on the far-reaching consequences of worsening water shortages. Those concerns have been compounded by lack of certainty about what the impact of climate change will be on water supplies. Even without environmental change, the world is consuming water at ever-faster rates. Global population has doubled over the past 40 years but use of water has quadrupled. The Water Resources Group forecasts that global water demand may outstrip sustainable use by 40 percent as soon as 2030.

Photo: David Greitzer / shutterstock.com

Lake Shasta, CA, August 20, 2014 – California's lingering drought exposes the 180-200 foot drop in water levels.

Data centers are under particular scrutiny. In the U.S. alone they are expected to consume an estimated 174 billion gallons of water in 2020. A 15-megawatt data center can use up to 360,000 gallons of water a day. Regulatory pressure is being turned up by state and local bodies that are concerned about scarcity of public resources while activist groups like Greenpeace have put a special emphasis on drawing attention to data centers' use of both power and water. All this is happening as cooling options are narrowing due to restrictions on the use of refrigerants.

In addition to overall data center growth, demand for cooling is being driven by new applications like machine learning, cryptocurrency mining and internet of things devices. Artificial intelligence workloads make extensive use of power-hungry graphics processing units and AI training algorithms can require days of heavy processing.

Data center water use has broad impacts. It affects the quality and availability of local water supplies, particularly when groundwater is involved. It invites increased regulatory scrutiny and operators may have to make expensive infrastructure investments to clean recycled water.

Hyperscale operators are also leading the charge toward innovative cooling strategies, pioneering the use of free-air cooling, raising operating temperatures for server inlet air and experimenting with new technologies like indirect air cooling, refrigerant economization, water to the chip, immersion cooling and rear-door chilling units. However, there is no clear solution to reducing water usage while meeting the growing cooling requirements of ever-denser compute infrastructure.

And the equation is more complicated than just the tradeoff of water versus air cooling. In fact, air cooling can actually increase overall water consumption. Air conditioning is an enormous consumer of power. A typical data center uses 35% of its power for HVAC, according to The Engineering Mindset.

The role of water in power consumption is often overlooked. Few people are aware that power generation consumes more than half the water in the U.S. and Europe. Generating a single kilowatt of electricity with coal or nuclear fuel requires about 15 gallons of water. That means that the most effective way for some data centers to reduce water usage is to cut back on power consumption. The balance comes in seeking efficient technologies that reduce or even eliminate the need for water consumption.

Seeking a Balance

The best strategy is to seek a balance. Since 2006, Green Grid's Power Usage Efficiency (PUE) rating has been used as the gold standard for data center efficiency. PUE is calculated by dividing the amount of power entering a facility by the amount used to support the IT load, with the goal being to get as close to a 1:1 ratio as possible. Although the Green Grid has stated that PUE was strictly meant to be a metric for data center operators to track their own power efficiency, it was adopted in the ANSI/ASHRAE/IES Standard 90.1-2013, thereby providing an incentive for the use of water cooling. By shifting the burden from electricity to water, facilities can improve their PUE and thereby better comply with regulations because water consumption is invisible to the calculation.

By shifting the burden from electricity to water, facilities can improve their PUE and thereby better comply with regulations because water consumption is invisible to the calculation.

In 2011 the Green Grid introduced the Water Usage Effectiveness (WUE) metric as a corollary to PUE. WUE is calculated by dividing the amount of water a site uses for cooling, regulating humidity and producing electricity by IT equipment energy usage. It's meant to enable data center managers to understand the effect water consumption has on the local electric grid.

Data center operators need to consider both metrics in achieving a balance that is good for both their bottom lines and sustainability objectives. Investments in alternative energy sources like wind and solar, both of which use far less water than coal-or nuclear-powered plants, can help reduce both scores.

The best course of action is to map the choice of thermal technology to the use case. Low- or zero-water solutions that also minimize power use are the most desirable. There are several types of cooling alternatives that are both water- and power-efficient.

Direct evaporative cooling

Direct evaporative cooling uses direct evaporation of water to produce significant cooling and humidification with low energy consumption. A wetted media like CELdek or GLASdek is used as a substrate to allow a large volume of air contact evaporating water. This is widely regarded as the simplest, most cost-effective method of cooling and humidification for air. This technology is deployed by those operators that are willing to let the air in the data center fluctuate with the outdoor air and are willing to take the risk of bringing in air from outside. To minimize these risks, direct evaporative systems must include full compressorized backup systems in the event outdoor air is not suitable for the data centers as is the case with fires that have occurred in California and other parts of the U.S.

Indirect evaporative cooling

Indirect evaporative cooling uses two opposing airstreams that contact a different side of a polymer heat exchanger. The outer wall of the exchanger contacts air that needs to be conditioned before it is delivered to the occupied space. The inner wall is in contact with air that comes from the ambient environment or building exhaust. Cooling occurs when water sprayed on the interior wall of a heat exchanger evaporates, imparting a cooling effect to the outer wall of the heat exchanger. This allows the airstream that contacts the outer wall to be cooled. This technology, while efficient, can consume more water than direct evaporative cooling due to the heat exchanger being between the indoor and outdoor air. The benefit of this technology over direct evaporative cooling is that it eliminates the risks associated with bringing outdoor air into the data center and the need to size backup or compressorized cooling to the full load of the data center.

The benefit of indirect evaporative cooling over direct evaporative cooling is that it eliminates the risks associated with bringing outdoor air into the data center and the need to size backup or compressorized cooling to the full load of the data center.

Evaporative freecooling chillers

Evaporative freecooling chillers use ambient air that is passed through wet pads or atomizing spray systems to extend the hours the chiller can operate in free cooling mode. During warmer days evaporative cooled air is injected into the heat rejection (condenser) coils, thereby reducing energy costs. The combination of operating modes between freecooling with the adiabatic or evaporative system and lowered condensing air temperature not only provides for a more efficient chiller operation, but also reduces the peak power requirements for the chiller to operate, reducing generator and power distribution needed to support the cooling load.

Refrigerant economizer

Refrigerant economizer is an efficient technology that leverages the phase change of refrigerants to economize or provide freecooling during times when the temperatures outside will support it. These systems use common compressor-based technology and pumps to circulate the refrigerant through the system to provide cooling. They can provide very efficient cooling without the use of water and have the best 10-year total cost of ownership because of their simple design.

Liquid cooling – direct-to-chip

Liquid cooling – direct-to-chip takes water or refrigerant directly to the chip to take the heat off the hottest components within the server. Direct-to-chip cooling extracts about 80% of the heat into the liquid while the remaining heat is cooled through air cooling systems in the data center. This technique is not only an efficient means of cooling equipment but it also reduces the fan energy within the equipment, thus improving operational efficiency.

Liquid cooling – immersion cooling

Liquid cooling – immersion cooling is a relatively new cooling technique that submerges computer equipment in a non-conductive fluid. Heat is removed by circulating liquid into direct contact with hot components, then through cool heat exchangers. The benefit of this approach over direct-to-chip liquid cooling is that all of the server heat is captured by the liquid, so fans can be eliminated from the servers.

Bracing for Change

Balancing power efficiency and water usage is a process that is unique to each data center and is affected by continual technology change. Operators should look for solutions that provide the best combination of PUE and WUE given local operating costs, sustainability objectives, regulations and community considerations. Facility size is an important factor in making the right choice as there is no one size that fits all. Look to hyperscale operators for ideas, as they are motivated to do some of the most innovative work in this area.

Also keep an eye on emerging trends in hardware design and AI. Component makers are constantly looking for ways to reduce power consumption, such as scaling back processor cycles or shutting down components during periods of low activity. AI-based controls will complement human operators by monitoring infrastructure at a much larger scale and adjusting workloads to optimize energy use.

Facility size is an important factor in making the right choice as there is no one size that fits all. Look to hyperscale operators for ideas, as they are motivated to do some of the most innovative work in this area.

The growing trend toward edge computing may enable further efficiencies by moving some processing to air-cooled servers in the field and relieving the load on networks and data centers.

Operators should also monitor the development of the Smart Grid and the opportunities it will present to monetize their UPS infrastructure while simplifying the path toward sustainable power generation. The field of power and water efficiency is the scene of constant innovation, making it one of the most exciting dimensions of data center operations.