Electronic Design

By BEA KARRON, Contributing Editor

Thirsty Silicon: The Hidden Costs of Big Tech's AI Boom (Part 2)

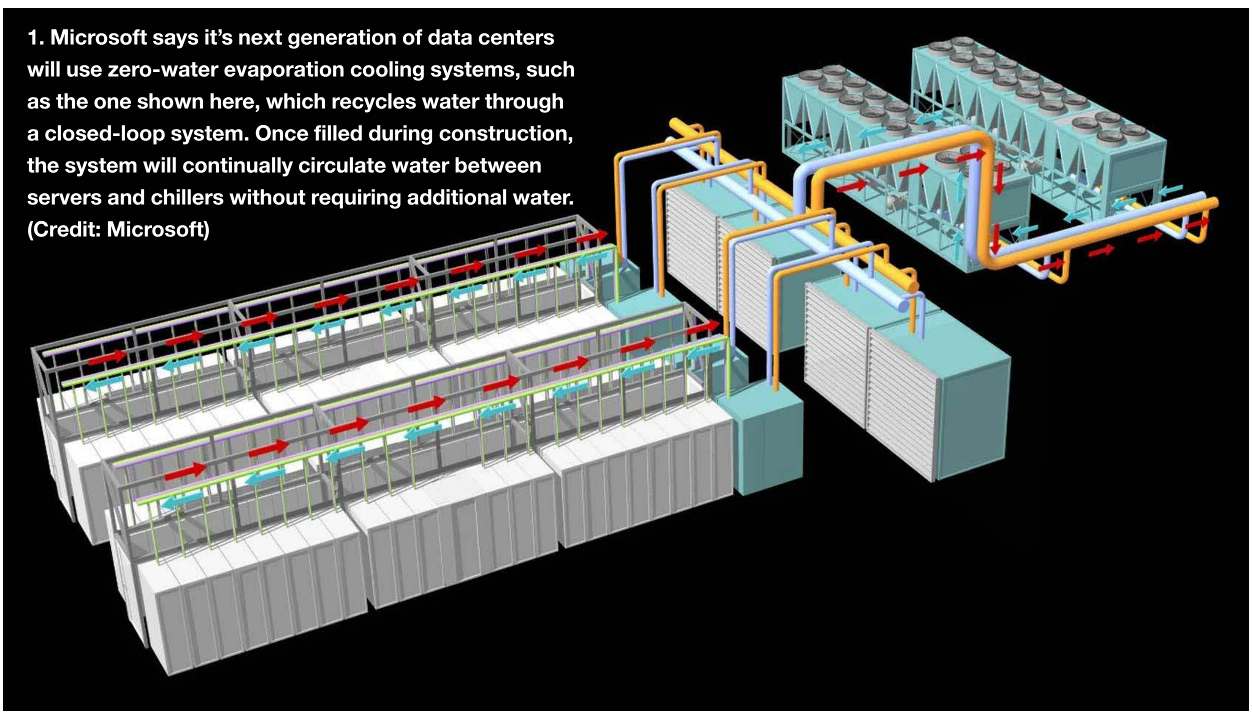
What's being done to reduce the extremely high water consumption needed to cool ever-more-powerful AI data centers? Part 2 delves into the innovative technology and other efforts to combat this escalating problem.

Part 1 of this two-part series focused on the increasing power demands of AI data centers, and the massive amounts of water being used to cool such systems, which is significantly depleting these resources. Part 2 looks at solutions to overcome these challenges and the commitments going toward making it happen.

Closed-Loop Cooling

While data centers are expected to rely on water cooling for the foreseeable future, a growing number of solutions can reduce how much water they consume. Most notably, closed-loop systems extract the heat from their water by passing it through a heat-exchanging cooling tower before it makes its next pass through the facility.

Although closed-loop systems tend to be more complex and costly, a European study estimated circular water solutions can reduce water usage by up to 70%. In addition, some operators are supplementing their coolant requirements with rainwater, reducing the amount needed to be drawn from wells or municipal supplies.



1. Microsoft says it's next generation of data centers will use zero-water evaporation cooling systems, such as the one shown here, which recycles water through a closed-loop system. Once filled during construction, the system will continually circulate water between servers and chillers without requiring additional water. (Credit: Microsoft)

In 2024, Microsoft announced that two of its new data center projects in Phoenix and Wisconsin would pilot new cooling systems with liquid-cooled zero-water evaporated designs in 2026. Though the company's currently operating data centers still use a mix of air- and water-cooled systems, Microsoft says all of its data center designs since August 2024 have implemented zero-water evaporation methods that mitigate energy impacts by using high-efficiency economizing chillers with higher water temperatures *(Fig. 1)*.

Meanwhile, Healixa IQ is tackling AI water consumption with its recently developed FrostPulse system, a compact cooling system with zero water loss. Rather than using water evaporation, FrostPulse captures and reuses thermal energy through matrix reservoirs, saving both energy and water, which the company claims can reduce cooling costs by up to 70%.

Hotter Chips Create New Challenges

In addition to cooling the buildings that house the data center, challenges have emerged about how to extract the growing amounts of heat generated by the servers. Blowing chilled air through the server racks worked until recently, but as GPU power consumption approaches 1 kW per chip, air cooling becomes insufficient. Thus, alternative cooling methods are required to prevent meltdowns.

To address this, processor manufacturers developed packaging that improves heat transfer between the chip's case and whatever cooling medium it's attached to. This has usually been accomplished by embedding copper slugs within the CPU's/GPU's packaging material, which efficiently couples it to a heatsink. For lower power processors, a finned heatsink exposed to a constant flow of air can be sufficient.

However, more powerful CPUs/GPUs use servers and even "gamer-grade" PCs generate much more heat than can be handled by simple air cooling. These applications require a higher-efficiency solution, such as direct-to-chip cooling, a heatsink with coolant channels attached to the case of the processor. The channels allow the server's cooling system to feed a steady flow of water or dielectric fluid very close to the CPUs/GPUs without putting it in direct contact with them.

Meanwhile, the higher levels of heat generated by advanced processors now demands advanced packaging technologies, which were originally developed for devices that drive electric-vehicle (EV) motors, solar inverters, and other high-power applications. This would enable more efficient heat transfer between the chip and its coolant.

For example, Element Six offers materials like Cu-Diamond, a synthetic diamond composite whose thermal resistance is half that of copper. According to Bruce Bolliger and Dr. Ian Friel, commercial development managers at Element Six, Cu-Diamond cold plates and heat spreaders can significantly improve GPU performance, reduce GPU temperatures, and could allow for the use of higher temperature coolant fluids, potentially reducing cooling costs.

Breakthroughs in Ultra-Efficient Chip Technology

The growing impact of data centers is also encouraging development of new processor technologies that offer improved efficiencies, helping to diminish strain on power grids and groundwater levels. For instance, EnCharge AI recently introduced EN100, the first analog in-memory AI chip that consumes significantly less energy than an equivalent digital device. In addition, the memory is designed for use in AI edge devices that can deliver as much as 200 TOPS of processing power.

According to Naveen Verma, EnCharge AI CEO, deploying AI applications locally requires significantly less energy to move data from the server to devices, while also decreasing cost and latency. Verma added that this strategy distributes the thermal and electrical loads of AI applications across a larger area while reducing dependence on potentially vulnerable cloud infrastructures.

These more efficient CPU/GPU architectures and the advanced IC fab process that make them possible are beginning to "bend the curve" on the growing demands for energy and water they consume.

Bitcoin, for example, was once projected to consume more power than that of the world by 2020. Thanks to less energy-intensive crypto algorithms and slower than anticipated growth in demand for digital currencies, crypto mining accounts for just 0.5% of the world's energy. It's hoped that data centers could experience at least a modest decrease in their anticipated impact as breakthroughs in efficiency continue in cooling, IT, and power consumption.

Thinking Outside the Data Center

But what if data centers could benefit the environment rather than hurt it? This possibility is spurring the development of more wholistic solutions that, when combined with efficiency, might greatly minimize the overall impact of data centers.

Tierra Buena Regenerative, a Texas-based company, is developing a promising approach to mitigate the environmental impacts of large concrete campuses by restoring soil health to the land they occupy. To accomplish this, the company offers a proven set of scalable methods that build healthy soils, capture rain, recharge aquifers, reduce evaporation, and decrease flood risk.

The regenerative practices in Tierra Buena's arsenal include applying mineral bases, subsoiling, planting cover crops, and dragging specially designed eco-plows to break up lower layers of hardened soil without disturbing the fragile structure of its upper layer.

CEO Erik Fritz said these practices allow his team to re-build living soil at the molecular level, which restores its natural ability to function as a giant filter and a bio-matrix for countless species of soil-enhancing microbes. “We think about soil more like infrastructure. And we take an engineering approach to it; not with concrete, but manipulating the natural factors that allow it to function properly.”

Though Tierra Buena hasn’t worked with any data center clients yet, Fritz said he thinks data centers don’t need to use any less water. If data centers are creating water vapor, it should come back down as rain. “At the end of the day, if you could restore the water cycles, through repairing damage to soil and restoring ecosystem function, I think the data centers could actually become a tool to accelerate rainfall and make it more regular.”

Transparency, Oversight and Regulation

Despite ongoing advances, data centers remain thirsty beasts that continue to expand in the U.S. with little oversight.

At the federal level, there’s no binding legislation for data centers in the U.S. For example, the absence of a federal data registration program makes estimating the number of data centers, and their impact, very difficult — especially since some data center operators deliberately obscure their facilities’ locations.

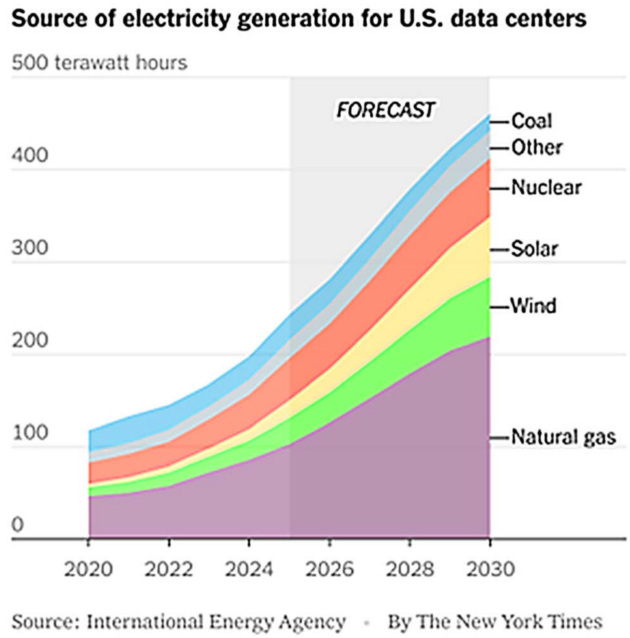
Some states, like Virginia, California, and Connecticut, have tried to pass legislation for mandatory reporting or energy use, water consumption and emissions. However, intensive lobbying caused most of these measures to be vetoed or indefinitely delayed.

According to the *Bay Journal*, Maryland Gov. Wes Moore (D) told state legislators that he intended to veto a bill that would study the economic, environmental, and energy consequences of data center development due to financial constraints and concerns about federal funding.

In New Jersey, Gov. Phil Murphy conditionally vetoed a bill that would require data centers to report water and electricity usage, citing a need for further research into whether data centers’ power and cooling demands are unreasonably burdening N.J. ratepayers.

In Europe, some regulatory framework is beginning to emerge to support critical AI infrastructure with the target of achieving carbon-neutral data centers by 2030. The European Commission’s regulations, to be proposed at the beginning of 2026, aim to triple EU data center processing capacity in the next five to seven years and will set new energy and water efficiency standards in the Data Centre Energy Efficiency Package.

It’s also possible that the answer to the balancing act between AI demand, water, and power could lie in AI itself. Climate Change AI, a U.S. non-profit, is exploring the relationship between machine learning and climate change. When used responsibly, the organization states that machine learning can help tackle climate change by lowering the cost of research, monitoring global changes, and accelerating the transition to low-carbon and renewable energy sources.



2. As AI data centers continue to proliferate, most of the power they consume will continue to be produced by fossil-fueled power plants. (Credit: NY Times)

Conclusions

As the appetite for AI services continues to surge, most forecasts anticipate their energy demand to triple by 2028, eventually consuming nearly 12% of the nation’s existing generating capacity. AI’s water consumption is also expected to grow at a similar rate. A recent study projects global demand to reach between 4X to 6X the annual water usage of a country like Denmark.

To combat intensive water use, Google, Microsoft, and Meta have all pledged to reach at least net-zero by 2030 (Amazon set its goal for 2040). They’re aiming to be water positive in the next five years, though corporations often fail to fulfill loft sustainability goals.

Compounding the problem, some experts are concerned that renewable energy sources won’t be able to meet data centers’ growing energy needs alone. This is causing many U.S.-based AI enterprises to turn to fossil-fuel plants to power their facilities, at least for now. This is driven in part by the changing realities surrounding the transition to wind, solar, and other renewable energy sources.

While renewables continue to provide a growing percentage of America’s energy needs, President Donald Trump’s green-energy policies and endorsement of natural gas, oil, and coal are slowing the nation’s transition to a low-carbon economy. While nuclear power shows some promise, there aren’t enough active plants to meet demand. And new atomic power plants based on commercially proven designs could take a decade or more to build and commission, and cost much more than conventional plants.

As a result, the country will probably continue to lean heavily on fossil fuels for the near future, a trend that extend the lifetime of many aging coal mines *(Fig. 2)*. In addition, U.S. Energy Secretary Chris Wright told Reuters that plans are being made to maximize the existing grid by ending the previous administration's planned phase-out of coal plants and extend their lifetimes by operating diesel-powered backup generators.

This means that even if the projections for the American AI industry's anticipated water and energy demands in the next three to five years won't be as large as predicted, they won't be met by renewable sources. That's due to changes in regulation at the federal level as well as elimination of nearly all of the government's initiatives to stimulate their funding.

Further reading on this topic:

If you're interested in calculating your own AI water "footprint," the news site The Conversation published a formula by Leo S. Lo, Dean of Libraries and Advisor to the Provost for AI Literacy at the University of Virginia. It estimates data center water use from on-site cooling and water required to generate power to the data center. The formula is:

Energy per prompt (watt hours) × Water factor (millimeters per watt-hour) = Water per prompt (in milliliters)

Click here for details on how the formula was derived and how to use it.

Bea Karron is a recent graduate of Marist University, where she earned a B.A. in Journalism and Public Relations. While at school in the Hudson Valley, she developed a passion for storytelling focused on science and sustainability, and how climate issues connect to our everyday lives.

Although she doesn't have an engineering background, she's picked up a fair amount of technical knowledge via osmosis from her father, an electrical engineer. Bea is excited to contribute to Electronic Design by exploring the complex intersection between technology and the environment.