

Voices

Making sustainability central to the data centers of the future

As data center expansion accelerates in response to surging demand for AI, how can we ensure that data centers are built and operated sustainably across environmental, technological, and social dimensions? Drawing on a panel of experts worldwide, this Voices piece examines challenges in developing and operating data centers sustainably.



Wei Sun
School of Engineering, The University of Edinburgh

Aligning data center demand with renewable supply across time and location

The data center construction cycle is moving faster than grid infrastructure development. Powering this growth sustainably requires more than building more renewable capacity. The central challenge is to match renewable supply with rapidly growing data center demand across time and location.

Unlike traditional large loads, data centers can also serve as potential grid assets by providing various demand-side flexibility. The biggest part of a data center's power demand is driven by computational workloads. Workloads involving AI training and batch processing can be delayed by deadlines or transferred between facilities, constituting the principal sources of flexibility in data centers. Coordinating temporal and spatial flexibility with renewable availability and network conditions would better align flexible data center demand with renewable energy supply. Meanwhile, uninterruptible power supply, battery energy storage systems, and on-site generation can further reshape the power demand of data centers.

Temporal flexibility allows delay-tolerant workloads to follow renewable output over hours or days. Scheduling these workloads during periods of abundant wind or solar generation can reduce reliance on fossil fuel generation and renewable energy curtailment. Such shifting must follow service-quality requirements and avoid rebound peaks when deferred workloads resume.

Spatial flexibility extends this coordination across locations. In the context of Great Britain, substantial wind generation is concentrated in Scotland, while major electricity demand is concentrated further south. The transmission constraints can lead to wind curtailment even when demand remains high elsewhere. Shifting flexible workloads between facilities could reduce this renewable curtailment and relieve regional grid congestion. However, spatial shifting must account for latency, data sovereignty, and local environmental impacts. Sustainable growth therefore requires data center demand to follow renewable supply in both time and location.



Vinod Narayanan
Department of Mechanical and Aerospace Engineering, UC Davis; UC Davis Western Cooling Efficiency Center

Cooling for AI data centers: Sustainability opportunities and challenges

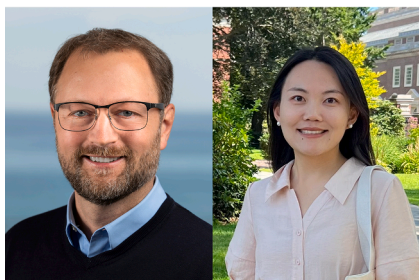
Efficient cooling has become a foundational enabler of modern AI data centers. Because nearly all electrical power consumed by servers is ultimately converted to heat, cooling is central to data center sustainability. Cooling systems influence energy use, water consumption, noise, and land requirements. The conventional air-cooling approach relies on conditioning and moving large volumes of air through servers. In contrast, closed-loop direct-to-chip liquid cooling systems allow current and future graphics processing units (GPUs), many approaching or exceeding 2,000 W with heat fluxes above 200 W/cm², to operate reliably without performance throttling. Liquid cooling supports higher-power GPUs and dense rack-scale architectures that improve compute performance, memory bandwidth, and interconnect efficiency, reducing the energy required per AI token generated.

Liquid cooling also enables higher coolant inlet temperatures, which reduces reliance on chillers and enables dry cooling rather than evaporative cooling, thereby substantially reducing water use in most climates. Coolant outlet temperatures in the



55°C–60°C range make heat reuse more attractive for applications such as membrane desalination, district heating, indoor agriculture, and hot water production. However, spatial and temporal mismatches between heat generation and demand limit widespread heat reuse.

When heat reuse is unavailable and evaporative cooling is minimized, rejection of large thermal loads to the ambient environment becomes a key challenge. Localized heat accumulation may reduce cooling efficiency, increase reliance on mechanical chillers, and accelerate component degradation and corrosion. Key research questions include how to improve cooling-system reliability and whether distributed data center architectures can mitigate heat island and noise impacts while enabling beneficial heat reuse.



Eric Masanet and Yingbo Zhang
University of California, Santa Barbara

Turning AI data center waste heat into a resource

Data center waste heat recovery is the next frontier of energy system integration, and one that holds substantial, if yet largely untapped, potential for accelerating sustainable energy transitions. The concept is not new. Some small data centers already repurpose low-grade waste heat from IT racks, generally for heating nearby spaces or for injection into district heating networks.

However, the emergence of large, liquid-cooled AI data centers can be a game changer. These data centers can concentrate much higher quantities of waste heat, and at higher temperatures than conventional air-cooled data centers can. This “high-value” heat greatly expands the opportunity space, both externally and internally.

Externally, high-value heat can be a viable source of process heating in industrial plants, space heating in greenhouses, and water heating in aquaculture—and of driving thermal processes in water treatment plants—to name a few large-scale examples. Sustainability benefits accrue when exported waste heat replaces fossil fuel use in these applications, especially when the data center runs on renewable power.

Internally, high-value heat can drive sorption-based chillers for data hall cooling or organic Rankine cycles for onsite power generation, both of which can be used for power load reduction and/or flexibility to benefit renewable grids. Soon, high-value heat might even enable on-site atmospheric water harvesting in water-scarce regions to support local water utilities or direct air capture to accelerate climate change mitigation.

Seizing these waste heat integration opportunities will not be easy. All applications will require private and public investments in new data center designs; co-location partnerships; heat, water, and/or carbon export networks; business models; and the enabling of policies. However, the sustainability benefits may very well be worth it.



Ayse K. Coskun^{1,2} and Varun Sivaram²
¹Boston University; ²Emerald AI

AI is an energy technology: Building flexible data centers

AI is often framed as an energy challenge. Rapidly growing data centers are placing unprecedented new electricity demand on power grids that are already constrained. But there is another side to this story: AI can itself become an energy technology. Advances in computing and software can transform data centers from rigid electricity consumers into flexible, controllable grid resources.

Computing is uniquely suited to flexibility because computation is programmable. AI workloads can be shifted in time, moved across locations, or run at different power levels. Software can orchestrate these options intelligently, reducing power when the grid is stressed and increasing consumption when electricity is abundant, while protecting application performance.

Importantly, this is no longer hypothetical. In our work across five real-world demonstrations in operational AI data centers, we have shown that computing workloads can respond directly to grid needs, curtailing power consumption for several hours and

dynamically following changing power targets. Software can unlock this flexibility while maintaining the service levels users expect.

This capability has important implications for sustainability. Reducing data center demand during grid peaks can decrease reliance on expensive, carbon-intensive peaker plants and prevent the rush to build short-term generation solutions. Conversely, flexible data centers can consume more electricity when renewable generation is plentiful, helping accommodate the inherent variability of wind and solar. Beyond sustainability, flexible AI data centers can avoid or defer expensive grid upgrades, which helps protect electricity affordability for local communities, and they can bolster the reliability of power systems by providing support during periods of peak grid stress.

As AI infrastructure expands, efficiency alone will not be enough. We must design data centers around both efficiency and flexibility. Done well, growth in AI and growth in clean energy need not be opposing forces: more computing can help enable more renewables, and ultimately, more AI tokens can run on a cleaner grid.



Ana Valdivia
University of Oxford

Making data center growth transparent and accountable

The sustainability of data centers remains a contested question as their value as critical infrastructure increasingly collides with the finite water, energy, and land resources that data centers depend on. Given the large volumes of water that data centers use to cool their servers, alternatives are being developed. For instance, chip-level cooling systems are gaining attention. However, these alternatives remain at an early stage, and water continues to provide a cheap resource for the data center industry. Renewable energy has also been promoted to address the large, sometimes carbon-dependent, amounts of energy needed to power servers and cooling systems, but it is not yet possible to power compute-intensive AI workloads through the use of renewable sources alone. As a result, data centers continue to contribute significantly to greenhouse gas emissions worldwide. Beyond water and energy, data centers also generate toxic waste and acoustic pollution, and they affect soil, fauna, and flora through, for instance, the discharge of wastewater into rivers. Amid this expansion, many governments and regions lack specific legal frameworks to regulate the industry beyond urban planning permits and general environmental regulations.

Given these challenges, concrete regulatory measures are needed to ensure transparency and accountability within this expansion of data centers. First, public participatory debate should be fostered so that local residents can democratically decide whether they want this industry operating near their communities. Second, greater transparency and independent accountability are needed around the environmental metrics of data centers. In turn, this information should be made publicly available and independently monitored. Third, there is a need to debate whether this expansion and resource demand are compatible with a world of limited resources and accelerating climate change. The sustainability dilemma of data centers creates tension between continuing to develop large AI models and meeting climate goals, and, unfortunately, there is nothing on the horizon to suggest that these two objectives will soon become mutually compatible.



Dillon Mahmoudi
Department of Geography and Environmental
Systems, University of Maryland, Baltimore County

Data centers and digital equity in the Global North and Global South

Investments in tech and energy capital are prompting the creation of data centers across the Global South, but the North dwarfs their scale and growth. Frontier AI compute, where the largest models are trained, stays concentrated in regions with access to capital, chips, and cloud demand. The South is becoming a compute consumer, not a producer, on terms it does not set. Microsoft's pledge to invest \$50 billion across the South does not change this, despite regulatory tolerance and the freedom to externalize water, grid strain, and carbon. Where facilities appear, such as Singapore's spillover into Johor or India's pivot from fossil assets to compute capacity, they serve regional cloud and inference, not frontier training: bets on demand that may never arrive.

The real divide is access. The South trains no frontier models. It lacks the GPU clusters, the capital, and the chip supply that Northern firms monopolize. It pays computational and data rent to consume models built elsewhere, at prices set by the firms that own them. The infrastructure arrives as dependency, not capacity. The gap that concretizes is not environmental burden migrating south, though those burdens are real where facilities land. It is exclusion from the digital means of production dressed as market integration. Consigned to the demand side of a rent relation, the South consumes AI it cannot make, while the extractive frontiers underwriting Northern compute fall on people who will never benefit from them. In its current form, digital equity is a colonial relation that is not solved by building data centers in the South but reproduced by them.

DECLARATION OF INTERESTS

V.S. is the founder and CEO of Emerald AI, a startup developing software created to enable power flexibility in data centers. A.K.C. is chief scientist at Emerald AI. Both V.S. and A.K.C. are shareholders of Emerald AI. Emerald AI has filed patent applications related to data center power flexibility and the software infrastructure that enables it.