Five Concerns About AI Data Centers, and What to Do About Them

By Hodan Omaar and Mitalee Pasricha | April 6, 2026

For decades, data centers were the quiet, reliable engines of the information economy, operating in the background of global commerce and daily life. But with the rise of artificial intelligence (AI), these facilities have been thrust into the public and political spotlight. Concerns continue to grow about what this expansion means for energy systems, water resources, and local infrastructure. But the root causes of these anxieties are poorly understood and frequently misattributed. Too often, policy responses target the scale of AI deployment rather than its systemic impact. This report examines five of the most consequential claims in that debate—electricity use, grid access, pricing, reliability, and water—and reaches a consistent conclusion: the core challenge is not AI infrastructure per se, but rather the frameworks used to measure, price, and manage its impact. Modernizing those frameworks could protect households and communities, strengthen grid performance, and reduce environmental impacts, while allowing AI infrastructure to scale in ways that support U.S. competitiveness and innovation.

INTRODUCTION

At the center of the debate shaping legislative proposals, regulatory action, and public opinion are five claims:

1. AI workloads use too much electricity.

2. AI workloads crowd out other uses of limited grid capacity.
3. AI workloads will raise household electricity bills.
4. AI workloads threaten grid reliability.
5. AI workloads strain local water resources.

These assertions arise for different reasons. In some cases, critics correctly identify real physical stresses—sharper power spikes, more volatile thermal loads—but reach for blunt responses such as bans or caps rather than technical and operational solutions that could address those stresses directly. In others, critics point to a legitimate concerns, such as higher household electricity bills, but misdiagnose the cause, blaming data center demand rather than the market design rules that govern how grid costs are recovered and passed on. In still others, the criticism reflects generalized opposition to large-scale AI deployment rather than a clearly defined, empirically grounded system risk.

This report examines each claim in turn.

First, regarding the concern that AI data centers use too much electricity, while AI workloads do increase power demand, data centers are not the primary, secondary, or even tertiary driver of rising global electricity demand. More importantly, electricity use on its own is not a policy problem unless it leads to a concrete failure, such as higher household costs, reduced grid reliability, or environmental harm. Treating consumption as the problem risks targeting scale rather than impact. To ensure that policymakers are evaluating energy use in a way that reflects real outcomes rather than headline numbers, Congress should direct the National Institute of Standards and Technology (NIST) and Department of Energy (DOE) to develop energy-per-unit-of-work metrics that measure power use relative to productive output, and support international alignment around these standards.

Second, regarding the concern that AI data centers crowd out other uses of limited grid capacity, the claim that data centers are “hogging” the grid implies that their demand is displacing more socially valuable uses of electricity and that these workloads are inherently less beneficial. That is not a fair characterization. Data centers support a wide range of economic and public benefits, and widely cited figures on their share of new demand often rely on interconnection queue data that overstates real capacity needs due to speculative and duplicative filings. That is not to say that there is no pressure on the grid. Policymakers should focus on making it easier for all projects, from clean energy and hospitals to housing and data centers alike, to connect to the grid, rather than restricting one category of demand. Congress should require utilities to publicly report queue management best practices that incorporate AI and automation, the Federal Energy Regulatory Commission (FERC) should tie grid operator cost

recovery to measurable reductions in study timelines, and Congress should use federal tax credits and loan programs to incentivize automated interconnection filings.

Third, regarding the concern that AI data centers will raise household electricity bills, the claim that data center growth will inevitably increase household bills misidentifies the source of the problem. If data center demand were inherently pushing up household electricity costs, similar growth would produce similar price increases across regions. It does not. In some regions, utilities pay generators a reservation fee based on forecasts of future demand, meaning projected AI load alone can trigger immediate cost increases for households—even before a single data center is built. In others, generators are paid only for electricity delivered, so similar growth does not produce the same price shock. The difference is not how much demand data centers add, but whether market rules decide when and how those costs are passed on to households. Current pricing structures also assume demand is largely unresponsive, but many AI workloads can be shifted across time or location in response to price signals. Policymakers should support grid-aware flexibility that allows large loads to adjust consumption during periods of peak stress, reducing the need to rely on the most expensive power sources and limiting price spikes that would otherwise be passed through to households.

Fourth, regarding the concern that AI data centers threaten grid reliability, AI workloads do introduce new operational challenges, particularly due to their highly variable and fast-changing power demand. The risk arises from how these load patterns interact with grid infrastructure that was not designed to handle rapid fluctuations. Addressing this requires ensuring that large loads manage their power draw in ways that support system stability. Congress should support the development of an industry code of practice for load smoothing, on-site buffering, and ramp-rate control, It should direct FERC to link favorable interconnection terms to adherence to these standards and encourage insurers to offer better premiums to operators that demonstrate compliance.

Fifth, regarding the concern that AI data centers strain local water resources, the debate often focuses on how much water data centers use, but water use is not the same as water harm. What matters is where water is withdrawn, how it is used, and whether it is returned in ways that sustain local ecosystems. Many data center operators are already taking steps to manage and replenish their water use, but differences in how water consumption and replenishment are measured make it difficult to assess their actual impact or compare performance across operators. Congress should direct the U.S. Environmental Protection Agency (EPA), in coordination with NIST, to establish a standardized water accounting framework that defines what must be measured and how replenishment is verified to ensure consistent, comparable reporting across operators. It should also direct EPA to identify water-stressed regions where withdrawal

risks from data centers is highest and incentivize lower-withdrawal cooling technologies and water productivity improvements in those areas. Policymakers should also support the integration of data centers into district heating systems to reuse waste heat.

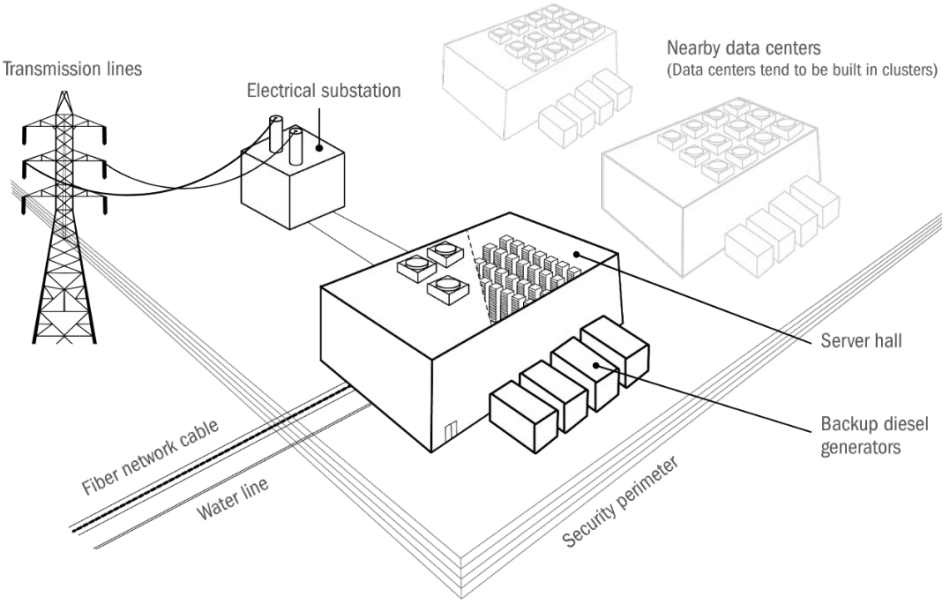
WHAT IS A DATA CENTER?

Data centers are facilities built to manage, store, process, and move large volumes of digital information. They support everything from everyday services such as email, video streaming, and cloud storage to enterprise systems that handle banking transactions, government records, and large-scale business operations. Increasingly, they also support more intensive computing tasks, including AI.

These workloads vary in complexity and intensity, but they all depend on the same core infrastructure: racks of servers to process data, high-speed fiber to move it, cooling systems to keep temperatures stable, and steady, high-volume power to keep everything running around the clock. Before AI entered the picture, companies designed most data centers to handle mixed workloads with relatively predictable patterns and moderate power requirements.

Figure 1 shows what a typical data center setup looks like—from how electricity and data enter the building to the equipment inside to the systems that support and secure it. Understanding this baseline is important to see how AI is beginning to push against it.

Figure 1: Common infrastructure and equipment at a typical data center[1]



As the diagram shows, a data center is more than a building full of servers. It's a tightly integrated system that brings together power, connectivity,

cooling, and personnel to ensure continuous operation. Electricity enters through high-voltage transmission lines, is converted at an on-site substation, and then distributed throughout the facility. To guard against outages, backup generators and battery systems are installed on site.

Fiber network cables connect the data center to the Internet, other data centers in a cluster, or both, allowing it to exchange information quickly and securely. Inside the building, the data hall houses the computing and networking equipment that processes and stores data. These machines generate significant heat, so the facility also includes cooling systems and environmental controls to keep temperatures stable.

The site is secured by perimeter fencing and staffed by teams of information technology (IT) professionals, engineers, and facility managers who monitor and maintain operations. Some space is set aside for offices and support functions, but the core of the facility is built around delivering stable, uninterrupted storage and compute.

Not all data centers are built the same. Most fall into one of three categories: enterprise, colocation, or hyperscale. Enterprise data centers are privately owned facilities built to serve a single organization. These are typically nontech companies, such as banks, insurers, and government agencies, that need to store and process data securely and in-house. Colocation data centers, or "colos," are shared facilities where multiple companies rent space to house their computing equipment. Each tenant brings its own servers and manages its own IT systems, but relies on the colocation provider for everything else, such as power, cooling, physical security, and network access. Colos serve a wide range of customers, from small start-ups to major retailers to government agencies. Finally, hyperscale data centers are designed to serve the needs of global technology companies such as Amazon, Google, Meta, and Microsoft, as well as their customers. Some are owned and operated directly by the tech companies that use them; others are built and leased by third-party providers. In either case, hyperscale data centers are optimized for efficiency, modularity, and performance, and are now the fastest-growing segment of the market.[2]

WHAT AI WORKLOADS MEAN FOR DATA CENTERS

AI workloads place different demands on data centers because the type of computation they require is fundamentally different from traditional tasks. Meeting those demands forces changes not just in processors, but also in memory, interconnects, and servers. And because servers are the building blocks of data centers, these shifts cascade upward, reshaping everything from rack layouts to power and cooling systems.

The computational work of traditional data center workloads, such as web servers and corporate databases, spans a wide range of operations. These tasks combine scalar arithmetic, conditional logic, and data movement. A

database query, for example, may require retrieving records from storage, comparing values, applying filters, and sorting results. Each step follows a deterministic sequence, and while the workload can be complex, it is generally predictable and well-suited to general-purpose processors that handle many different kinds of instructions.

AI workloads, by contrast, rely on a narrower set of operations—chiefly matrix multiplications—but these operations are more computationally intensive. During training, these matrices are multiplied repeatedly as the system adjusts billions of parameters to capture statistical patterns across vast datasets. Fine-tuning applies the same operations on a smaller scale, refining an already trained model with new data so it can specialize in a particular task or domain. Inference is the stage in which the trained model is put to use on fresh inputs, applying its parameters to generate an output—whether that means classifying an image, recommending a product, generating text, or translating speech. All three stages involve performing vast numbers of multiplications and additions in parallel.

AI relies far more heavily on Graphics Processing Units (GPUs) than Central Processing Units (CPUs) to perform the massive parallel matrix operations required for training, fine-tuning, and inference. But the shift is not limited to the processor itself; AI workloads reshape the entire server. AI-optimized systems cluster multiple accelerators tightly together, pair them with high-bandwidth memory (HBM) located physically close to the compute units, and connect them through ultra-fast links that move data between chips at enormous speeds. As models scale and techniques such as distributed training and reinforcement learning expand, designers must rethink how compute, memory, and networking are arranged within each machine. These changes increase the amount of power concentrated in a single server and the heat it generates, and because servers are the building blocks of data centers, the effects cascade outward into rack layouts, power delivery systems, and cooling infrastructure.

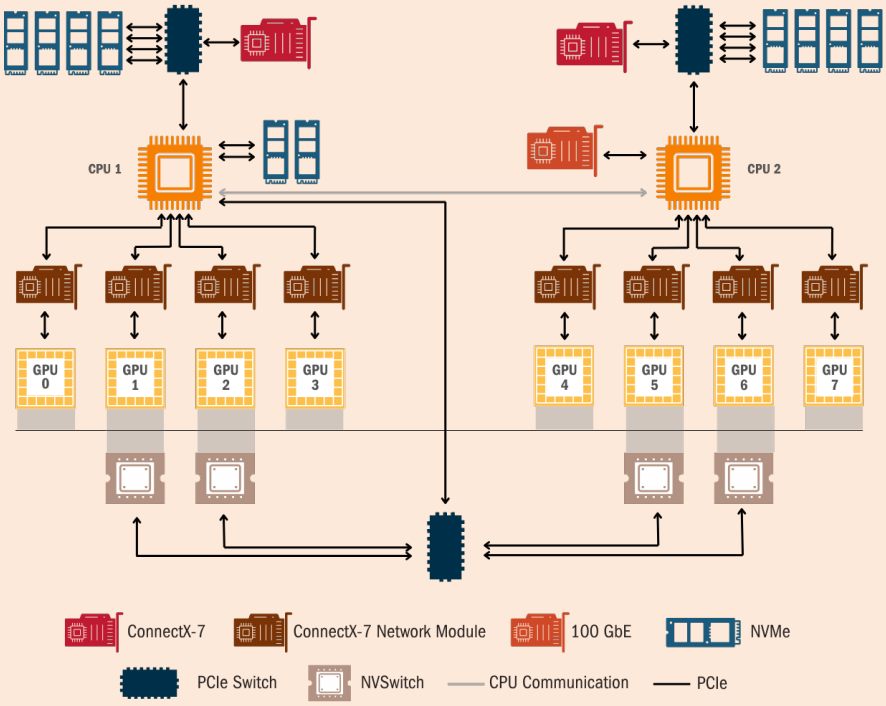
Before turning to the five concerns, it is useful to look closely at one example of what cutting-edge AI hardware looks like. NVIDIA's DGX H100 is a specialized server built around the company's H100 chips—among the most advanced AI processors on the market for several years, one of the most widely deployed AI processors in data centers today, and powerful enough to be subject to U.S. export controls. The DGX H100 is a full turnkey system designed and integrated entirely by NVIDIA.

Examining the DGX in the case study ahead, with all of NVIDIA's design choices on display, shows how a server built expressly for AI workloads differs from conventional machines and sets up the next sections on how those differences cascade outward to shape the design of entire data centers.

Inside NVIDIA's DGX H100: A Server Built for AI

At the heart of a single DGX server are eight NVIDIA H100 GPUs. The figure below illustrates the system's topology, with these GPUs at its center. Each GPU is built with two main types of cores: Tensor Cores and CUDA Cores. Tensor Cores are specialized for the matrix multiplications that dominate deep learning, and the H100 accelerates this work through automatic mixed-precision, meaning it can dynamically switch between different floating-point data precisions during calculation to balance speed and accuracy. CUDA Cores, meanwhile, are general-purpose processors that handle tasks less suited to matrix math, such as data preprocessing, activation functions, and managing the overall flow of AI programs.

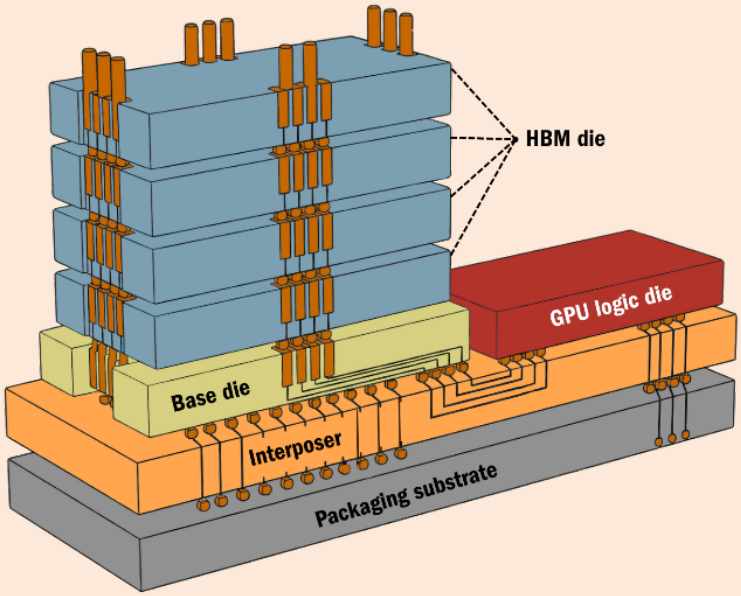
Figure 2: DGX H100 System topology[3]



Unlike conventional servers, wherein memory sits apart from the processor on a separate module, the H100 places its memory directly on the chip. This may sound like the memory is stacked directly above the processing cores, but that is not the case.

As figure 3 shows, the GPU logic die (in red) is the engine in which the processing cores sit. Next to it are HBM dies layered vertically, with a control layer known as a base die at the bottom that manages communication with the processor through a thin slice of silicon called an interposer. The base die is part of the HBM. The structure is far thinner and more compact than this diagram suggests, but the principle holds: HBM for GPUs is often built upward rather than spread out flat.

Figure 3: Vertical stacking of HBM next to a GPU[4]



GPU chips also come with built-in ports to link them with other devices in the server. On the H100, some of these ports are dedicated to NVLink, a specialized high-speed connection that links one GPU directly to another and provides a faster pathway for GPU-to-GPU communication than the standard interfaces do. Other ports on the H100 chip connect the GPUs to an NVSwitch, which acts like a hub and lets every GPU in a server talk to every other one at once—forming a tightly connected network for massive AI workloads. Finally, the GPU chip also includes ports to communicate with the server's CPUs.

The rest of figure 2 shows how these processors connect outward to networking and storage. On each side, the CPUs link up to high-speed network adapters that provide external connectivity, using technologies such as Ethernet or InfiniBand to move data between servers, and also to storage systems to access massive datasets. The system also includes a significant amount of internal, high-speed storage for data caching and local work. While the diagram shows two separate halves, they work together like two sides of a brain, acting as a single, unified system. Taken together, the DGX shows how changes within a GPU ripple far beyond the chip itself. The way these components are arranged—stacked vertically, packed more tightly, and linked through specialized bridges—creates more heat that must be pulled out through advanced cooling systems. The added weight and density, in turn, require new approaches to how data halls are built. In short, the innovations that make GPUs such as the H100 so effective for AI workloads also drive a cascade of new requirements for the design and operation of servers and entire data centers.

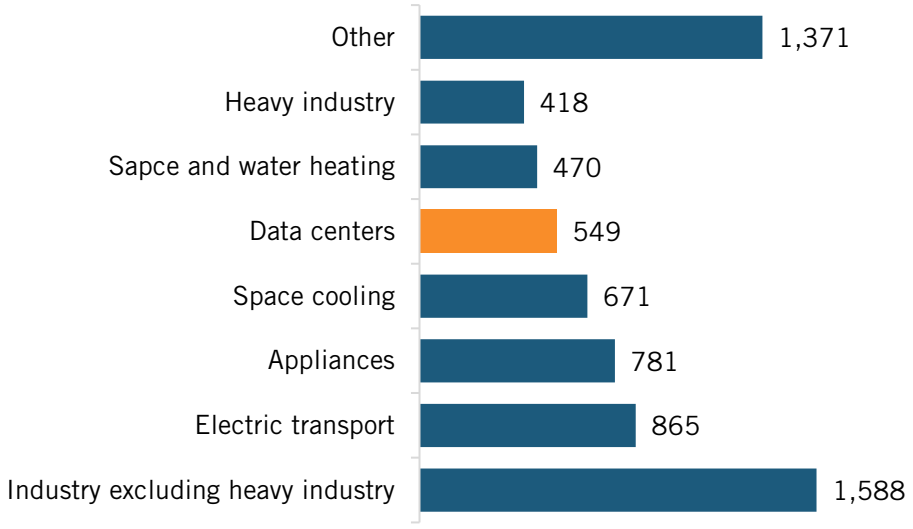
CONCERN 1: AI WORKLOADS USE TOO MUCH ELECTRICITY

Unlike the four concerns that follow, which tie higher power use for AI to specific downstream effects such as crowding out other grid capacity—raising household electricity bills, threatening grid reliability, and straining local water resources—the assertion that AI data centers "use too much electricity" frequently stands on its own.[5]

It is true that AI workloads increase the magnitude of power demand on electrical systems from data centers because the hardware needed to run and operate them consumes significantly more power than does the hardware used for traditional computing. According to research from SemiAnalysis, the average instantaneous power draw of a typical CPU and storage servers is about 1 kilowatt (kw), while a single AI server is in the order of 10 kw. Specifically, SemiAnalysis has found that the average power draw for a DGX H100 sever during normal operation is about 10,200 watts (W).[6]

But data centers are not the primary, secondary, or even tertiary driver of global electricity demand. Figure 4 illustrates data from the International Energy Agency showing projected increases in electricity demand by sector from 2024 to 2030, measured in terawatt-hours (TWh)—a standard unit for measuring electricity consumption at national or global scale, equivalent to one trillion watts of power sustained for an hour. According to the data, electricity use from data centers will account for less than 10 percent of the total increase in global electricity demand between 2024 and 2030.[7] Other factors, such as industrial output, the electrification of transport and buildings, rising air conditioning use, and the deployment of electric vehicles, are expected to contribute far more to overall demand growth.

Figure 4: Increase in global electricity demand by sector, 2024–2030 (TWh)[8]



Data centers are not uniquely straining the electric grid relative to other large sources of demand. So if the concern isn't tied to a specific downstream harm—higher consumer cost, environmental damage, reduced grid reliability, or displacement of other users—then it is not really about a measurable system failure. In that case, electricity use becomes a proxy for broader skepticism about AI's scale or pace. Treating absolute electricity consumption as inherently problematic substitutes reflexive resistance to AI deployment for a serious policy debate—without ever specifying what concrete problem needs to be fixed.

Policymakers Should Establish Energy-per-Unit-of-Work Metrics for AI to Ensure That Electricity-Use Comparisons Are Relative to Productivity

Even if policymakers accept that electricity use should be evaluated in terms of downstream effects, absolute consumption often becomes the shorthand proxy for those concerns. But aggregate electricity figures alone cannot distinguish between systems that draw more power because they deliver orders of magnitude more computational output and systems that draw more power while producing roughly the same level of throughput. A headline noting that data centers use twice as much electricity this year as last reveals nothing about whether AI systems are becoming less efficient or dramatically more productive. Without a way to relate energy use to output, such comparisons risk distorting rather than clarifying the policy debate.

Energy-per-unit-of-work metrics make that distinction explicit. By tying power use directly to output, they reveal whether additional electricity is translating into proportionally greater computational capacity, speed, or throughput—or whether power use is rising without meaningful gains in productivity. This reframes the energy debate away from raw consumption and toward how effectively electricity is being converted into useful work.

Industry has already begun moving in this direction by developing measures such as performance-per-watt or intelligence-per-watt. MLPerf Power has emerged as a leading benchmark, measuring the energy required to complete defined computational workloads, such as training a model or processing a fixed number of inference tasks.[9] For large language models, researchers are increasingly using tokens-per-joule to capture how much language output is generated per unit of energy.[10] At the hardware level, FLOPs-per-watt (FLOPS = Floating-point Operation) measures how efficiently chips convert power into raw computation, and newer AI accelerators are explicitly designed to maximize this ratio.

In the United States, NIST should work with DOE to develop recommended best practices for measuring energy per unit of productive work in AI systems. Rather than focusing on total power draw, these best practices should emphasize workload-level productivity for both training and inference, drawing on existing approaches such as task-based benchmarks, performance-per-watt measurements, and system-level

efficiency metrics. For example, the standards could specify a set of representative AI tasks, measurement methodologies, and reference hardware configurations that allow energy-per-unit-of-work to be compared across models and systems without requiring disclosure of proprietary model details or training data.

The United States should also work through international forums such as the G7 and the Organization for Economic Co-operation and Development (OECD) to encourage convergence around these productivity-based metrics. International alignment would help ensure that energy-per-unit-of-work metrics become a common reference point for evaluating AI systems, rather than giving rise to fragmented or inconsistent measurement regimes across jurisdictions—particularly as some countries consider mandatory reporting requirements related to AI energy use.

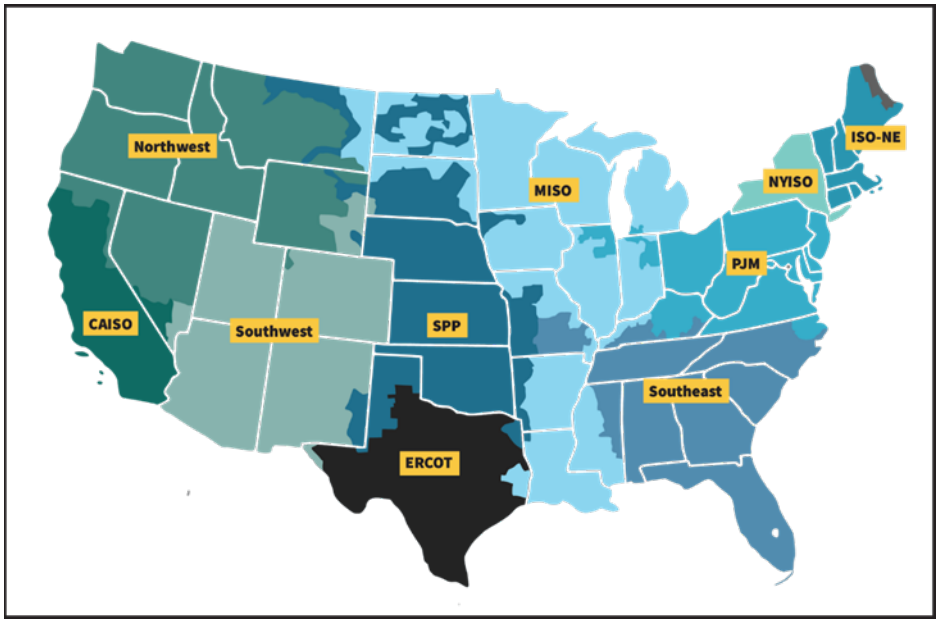
CONCERN 2: AI WORKLOADS CROWD OUT OTHER USES OF LIMITED GRID CAPACITY

Critics contend that data centers for AI crowd out other socially valuable uses of electricity. As electricity systems become increasingly constrained, scarce power that could otherwise support the electrification of homes, transportation, or industrial decarbonization is instead diverted to private data centers running commercial AI workloads. From this perspective, AI does not merely add demand to the grid; it also competes directly with other priorities for limited capacity.[11] The critique is especially acute in regions where grid expansion has lagged behind demand growth.

To understand this concern as it manifests in the United States, it helps to start with how the U.S. power system is organized. The United States is not served by a single, unified electric grid. Instead, it is divided into regional power systems, each with its own operating rules, planning processes, and reliability standards. In most parts of the country, day-to-day grid operations are managed by nonprofit entities known as Regional Transmission Organizations (RTOs) or Independent System Operators (ISOs).

RTOs and ISOs, such as PJM in the Mid-Atlantic, MISO in the Midwest, and CAISO in California, do not own power plants or transmission lines. Their role is to operate the grid.[12] They coordinate the real-time flow of electricity across multistate regions, ensure that supply and demand remain balanced, and manage the technical rules that govern who can connect to the system and under what conditions.

Figure 5: RTOs and ISOs in the United States[13]



A central responsibility of these grid operators is determining which resources are allowed to inject electricity into the grid and which large customers are permitted to draw significant amounts of power. Because electricity must be delivered instantaneously and reliably, any new connection—whether a supply-side power plant or a demand-side large industrial load—must undergo a formal technical review known as interconnection. This process is designed to ensure that new projects do not overload transmission lines, destabilize voltage or frequency, or increase the risk of outages.

Interconnection therefore functions as a gatekeeping mechanism. Grid operators study the impact of proposed projects on substations, transmission lines, and other shared infrastructure, and may require upgrades before approving a connection. These studies take time and are conducted sequentially, which means projects are placed in queues and evaluated in order.

Because access to electricity is governed by a slow, sequential interconnection process, and because grid expansion itself proceeds incrementally, critics argue that the system has limited ability to absorb large, sudden additions of demand. In that context, they contend that the arrival of gigascale AI data centers tilts scarce electricity toward private data center deployments and away from other forms of electrification and clean energy use that must compete within the same constrained system.[14] In response, critics have called for measures to slow or pause new data center approvals, such as a bill introduced in Virginia that would temporarily halt additional projects until existing interconnection requests

are processed, explicitly citing the risk of already strained queues becoming further congested.[15]

Speculative Queue Requests Distort Perceptions of Grid Use

Many of the headline statistics used to argue that data centers are hogging the grid, such as claims that they account for more than 90 percent of new power requests, are drawn from interconnection queue data that substantially overstates real demand.[16] As documented by Lawrence Berkeley National Laboratory (LBNL) in its 2025 "Queued Up" report, U.S. interconnection queues are dominated by speculative and ultimately nonviable projects. Developers frequently submit multiple, overlapping interconnection requests for the same project across different locations to preserve optionality while siting, permitting, and cost negotiations are underway. Once a viable site is identified, the remaining requests are withdrawn, often removing hundreds of megawatts (MW) of "capacity" from the queue at once.

Now, even if a withdrawn request never draws a single watt of power, its presence in the interconnection queue makes grid operators perform complex, time-consuming reliability studies as if the project were real. This creates a logjam wherein legitimate projects such as new housing developments, hospitals, or renewable energy sites get stuck behind these speculative data center placeholders, delaying their access to the grid.

That is why regulators such as FERC introduced stricter queue reforms in 2023, including higher withdrawal penalties and milestone-based study requirements designed to ensure that developers had meaningful financial commitments before reserving capacity. Many RTOs and ISOs also increased deposits and tightened site-control standards. After these reforms took effect, interconnection queues shrank in 2024 for the first time in years even though underlying demand for new projects did not materially decline.[17] Treating inflated queue positions as equivalent to actual electricity use confuses administrative congestion with physical scarcity and risks responding to a paperwork backlog as though it reflected a genuine shortage of local power supply.

The Hogging Critique Overlooks Self-Supplied Power

The claim that data centers are "hogging the grid" also overlooks the extent to which operators are trying to step out of the line altogether. Rather than competing for scarce grid capacity, many developers are accelerating a shift toward behind-the-meter solutions, generating or contracting power directly at or near their sites to reduce reliance on the public utility grid. McKinsey estimated that self-generation could meet up to 30 percent of new data center demand by 2030, up from less than 5 percent in 2023, as companies seek to bypass multi-year interconnection delays.[18] In practice, this means building localized microgrids that can operate independently, combining large-scale battery storage, on-site

generation such as high-efficiency natural gas turbines, and emerging technologies such as fuel cells.[19]

Even so, shifting demand off the grid does not eliminate interconnection challenges—it reshapes them. Behind-the-meter generation, dedicated renewables, and new firm power still require grid approval in order to connect, synchronize, or export power. When the timelines for approving new load and the timelines for connecting new supply diverge, and when new load and new supply cannot be safely coordinated within the grid's operational limits, projects can stall.

Ireland is experiencing this acutely. Despite a strong pipeline of offshore wind, new data center development has slowed because the national grid operator EirGrid cannot safely accommodate additional high-density loads without increasing the risk of system instability, including transient faults and frequency imbalances.[20] In 2024, EirGrid warned of a potential "mass exodus" of data centers if connection agreements remained stalled.[21] Amazon has already paused further investment in the country, citing uncertainty around how and when new offshore wind projects will be connected to the grid, as well as a broader lack of clarity on the requirements for securing energy access for data centers.[22] Interconnection delays aren't just a technical problem; they can derail investment and weaken national digital competitiveness. In the United States, a similar issue exists, but how quickly and reliably supply and demand align varies by regional electricity market.

Policymakers Should Reduce Interconnection Backlogs by Scaling AI-Enabled Interconnection Processes

Policymakers should focus on scaling and institutionalizing AI-enabled interconnection processes to reduce the time it takes for large energy users and new generation to connect to the electric grid. While pilot programs have demonstrated that AI can accelerate interconnection studies, those gains remain uneven and highly localized, with no clear pathway to routine adoption across the interconnection process.

DOE's AI for Interconnection (AI4IX) initiative is a useful starting point, but its impact remains constrained by both scale and scope. As currently structured, AI4IX functions primarily as a pilot-funding program, supporting partnerships between grid operators, project developers, and software providers to automate discrete components of the interconnection process.[23] For example, it might support a project focused on grid capacity and impact simulation, using AI to rapidly assess how proposed projects would affect existing grid conditions and thereby speed up the technical studies required for interconnection approval.

These efforts are valuable proof points, showing what is technically possible and helping to de-risk new approaches for utilities and regulators. But unless their outputs are adopted systematically across regions and

embedded into standard utility practice, such pilots will not materially reduce interconnection backlogs. Driving diffusion will require a combination of interventions.

First, Congress should require public utility transmission providers to adopt and share interconnection queue-management best practices that incorporate advanced computing tools, including AI, machine learning, and automation. A version of this approach appeared in the proposed Department of Energy AI Act 2024, which would have required transmission providers to "share and employ, as appropriate, queue management best practices with respect to the use of computing technologies … in evaluating and processing interconnection requests, in order to expedite study results."[24] These reports should be made publicly available, both enabling researchers, regulators, and ratepayer advocates to benchmark performance across utilities and creating natural accountability pressure on those that lag behind. Reviving this concept would help move AI-enabled interconnection from isolated pilots to a durable, system-wide capability

Second, FERC should introduce interconnection productivity credits by changing how grid operators are paid for running the interconnection process. RTOs and ISOs are responsible for reviewing interconnection requests and conducting the technical studies needed to connect new generation and large energy users to the grid. Because they are generally not-for-profit entities that do not sell electricity, they recover the costs of doing this work through fees and tariffs approved by FERC.[25] Today, those costs are generally recovered regardless of how long interconnection studies take—a model that relies on traditional cost-of-service regulation which often has poor incentive properties.[26] FERC could instead tie a portion of that cost recovery to performance, such as measurable reductions in study timelines. This would mirror existing performance-based regulation frameworks that reward utilities for meeting specific reliability or efficiency targets.[27] Under this approach, grid operators that meaningfully reduce interconnection timelines would be rewarded, whether through AI and machine learning tools validated by DOE, improved workflows, or other innovations. AI-enabled tools are particularly promising because they offer a scalable way to accelerate study processes across the system.

Third, Congress should build on DOE's Speed to Power Initiative—launched in late 2025 to accelerate the development of multi-gigawatt energy projects needed to support AI-driven load growth and reindustrialization—by using federal financial support to incentivize the adoption of federal financial support for the use of modern, automated interconnection processes.[28] Specifically, Congress should require that energy generation and grid infrastructure projects seeking federal tax credits under the Inflation Reduction Act or debt financing through the DOE Loan Programs Office use standardized, automated interconnection filings wherever the

relevant regional transmission organization or utility has such tools available. By linking federal capital to the use of automated interconnection interfaces, Congress would accelerate adoption at scale and move AI-enabled interconnection from isolated pilots into routine practice for the energy projects most critical to near-term grid expansion.

CONCERN 3: AI WORKLOADS WILL RAISE HOUSEHOLD ELECTRICITY BILLS

Critics contend that the rapid growth of power-intensive AI facilities will inevitably push up monthly electricity bills for households. According to this view, the costs associated with AI-driven data center growth are ultimately passed on to households through two main channels: regulated infrastructure costs and wholesale electricity prices.

Regulated Infrastructure Costs

In most of the United States, electricity is supplied by regulated utilities—private companies or public entities that operate as monopolies within a defined service territory. Because these utilities do not face direct competition, their investment decisions and customer rates are overseen by state-level Public Utility Commissions (PUCs).

When a utility determines that new infrastructure is needed to serve growing demand, such as transmission lines, substations, or grid reinforcements, it must seek approval through a formal process known as a rate case. In a rate case, the utility asks the PUC for permission to recover its costs and earn a regulated return, and it proposes how those costs should be allocated across customer classes—residential, commercial, and industrial—using the principle of cost causation, which holds that customers should pay in proportion to the costs they impose on the system.[29]

Critics contend that multibillion-dollar grid upgrades are increasingly being spread across the entire rate base, caused by a small number of large data centers driving those investments.[30] In this scenario, the cost of new infrastructure is embedded in the monthly delivery charged to households. Critics further warn that if data center demand were to slow or shift, these long-lived assets could become stranded, leaving residential customers responsible for decades of debt on infrastructure they did not request.[31]

Wholesale Electricity Prices

Utilities do not generate all the electricity they sell. In much of the United States, they purchase power from regional wholesale markets that operate through continuous auctions.[32] These markets are run by regional grid operators such as PJM and MISO, which coordinate electricity supply and demand across multistate areas.

In these markets, power plants submit bids indicating how much electricity they can supply and the minimum price at which they are willing to operate. Grid operators then stack these bids from cheapest to most expensive and dispatch enough generation to meet total demand at that moment. The key feature of this system is that all electricity is paid at the price charged by the last power plant needed to meet demand.[33] This is typically a higher-cost plant, and can be a gas-fired "peaker" plant that is expensive to run but can ramp up quickly. Even though most electricity may come from cheaper sources, such as nuclear, coal, or renewables, the price paid to all generators reflects the cost of that marginal, last-needed unit.

Critics argue that large AI data centers affect this pricing mechanism because they add large amounts of steady, round-the-clock demand. When overall demand rises closer to the system's available supply, grid operators must call on these higher-cost plants more often to meet peak or near-peak conditions. In practical terms, this means the market clears at a higher price more frequently because expensive plants are needed more often to keep the lights on.

When wholesale prices rise in this way, utilities pass those higher costs through to customers as increased supply charges on monthly bills. From the critics' perspective, households are therefore exposed to higher electricity prices because data center demand pushes the system into relying on more costly generation, even though residential customers are not the source of that additional load.

The Real Driver of Higher Household Bills Is Market Design, Not Data Center Load

Rising household electricity bills linked to data center growth are primarily a market design failure, not a demand problem. Electricity markets do not distinguish between different types of demand. An electron used by a data center is no different from one used by a household appliance, a factory, or an electric vehicle. The grid responds only to total demand, not who is consuming it. If the increase in demand from data centers were inherently driving higher prices, similar price increases would appear wherever that demand is expanding.

The evidence shows that that is not the case. A recent analysis from SemiAnalysis comparing PJM and ERCOT shows that while both regions are experiencing rapid data center expansion, their electricity price trajectories have diverged. SemiAnalysis estimated that in the PJM area, the spike in reservation fees will translate to a $25 to $30 surcharge on the average monthly household bill in the 2025–2026 period compared with 2024.[34] In ERCOT, no comparable price shock is projected to occur. If two systems experience similar demand growth but produce very different price outcomes, demand alone cannot explain the result. The difference lies in how each system translates forecasted demand into prices

PJM relies on what is called a capacity market that sets prices based on forecasts of future demand.[35] It uses a mathematical model to estimate how much electricity will be needed years in advance and then determines how much to pay power plant owners to keep their facilities on standby. In effect, this functions as a reservation fee paid to generators to ensure that capacity is available when needed. That cost is then passed on to households and businesses through electricity bills. The price is tied not to electricity actually consumed today, but rather to a central planner's estimate of what demand might look like in the future.

Because this reservation fee is built on a simulation rather than actual usage, even modest adjustments in the forecast can produce extreme jumps in cost. If the model predicts a surge in AI load, the formula automatically triggers a price hike to ensure that enough plants are waiting in reserve—even if those data centers are currently just empty lots. In the 2025–2026 cycle, this forecasting mechanism has caused the total cost of these standby payments in PJM to increase 9.3x over the prior year.[36] This has resulted in $16 billion in total charges that are being passed directly to households, forcing them to pay to reserve power for demand that does not yet exist.[37]

ERCOT, by contrast, relies on what is called an energy-only market, wherein generators are paid only for the electricity they produce and deliver to the grid.[38] There is no forward payment to keep plants on standby. Instead, the system uses real-time pricing to balance supply and demand. In this model, prices only rise when electricity is physically scarce. If demand spikes, the price of power increases at that moment, providing a natural incentive for power plants to turn on and for investors to build new capacity to capture those higher revenues.

This design limits how much speculative demand can affect prices. Delays or overestimates in projected data center growth do not translate into immediate costs for households because prices respond to actual conditions, not modeled expectations. The tradeoff is greater price volatility. During periods of extreme demand or supply constraints, prices can spike sharply. But those spikes are tied to real conditions on the grid, not projections of what might happen years in the future, and can be managed through a range of tools (discussed elsewhere in this report).

The comparison makes clear that higher household electricity bills are a function of market design, not data center electricity use itself.

Wholesale Price Fears Incorrectly Assume AI Demand Is Inelastic

To see why concerns about higher wholesale electricity prices resulting from AI demand do not necessarily translate into unavoidable cost spillovers for households, it is useful to borrow the classic "fish market" analogy from economics.

Imagine a local fish market early in the morning. The fishermen have already returned to port; the day's catch is fixed and cannot be increased in the short run. Supply is perfectly inelastic. If a sudden influx of new buyers arrives, prices rise sharply because everyone is competing for the same fixed quantity of fish.

Electricity markets resemble this scenario. In the short term, generation capacity is largely fixed because the fleet of available power plants is constrained by physical and regulatory lead times; many baseload units (e.g., nuclear or large coal) cannot ramp their output up or down quickly to meet sudden shifts.[39] When demand rises against that fixed supply, prices adjust upward to balance the system.

In a well-functioning market, rising prices would cause some buyers to reduce consumption or exit the market altogether. For instance, some fish buyers will simply leave the market and come back tomorrow or buy chicken instead. That response is what limits price spikes and prevents the most expensive suppliers from setting prices for everyone.

In today's electricity markets, however, demand is unresponsive because retail rates are decoupled from wholesale reality. Most households and data centers pay a regulated flat rate that shields them from seeing the actual cost of power in real time. Because they don't see power becoming expensive, they don't stop buying. The utility has to pay the spike price and then "true up" the difference later by adding surcharges or raising the base rate for everyone.[40] This turns a market spike into a long-term price hike for the entire rate base.

However, AI demand is not inherently inelastic. Unlike households or critical services, data centers running AI workloads can behave as flexible consumers of electricity in ways that most households and critical services cannot. Many AI workloads—particularly training and other non-latency-sensitive tasks—can be paused, slowed, deferred, or relocated without losing progress. This flexibility is the focus of cutting-edge research at the National Renewable Energy Laboratory (NREL), where work on workload-aware grid management shows how data centers can automatically throttle or pause training during periods of grid stress, shift computation to off-peak hours when prices are lower, or move workloads geographically to regions with excess renewable generation.[41]

Not only do AI data centers not have to be responsible for inevitable raises in household electricity prices, but they can also reduce the price pressure that would otherwise be passed through to consumers by stabilizing the net load on the grid. According to a 2025 MIT Sloan study, data centers that implement flexible, grid-aware workloads can reduce total system costs by 2 to 5 percent.[42] When these large-scale users shift their intensive AI training to off-peak hours, they flatten the demand curve, which reduces the grid's reliance on expensive, high-emission peaker plants that typically set the high wholesale prices for everyone else. In this scenario, the data

center doesn't just pay its way, it also acts as a stabilizing force that lowers the average cost of electricity for the entire rate base

Policymakers Should Support Grid-Aware Flexibility to Suppress Price Spikes

Blocking data center development and socializing their costs are both policy failures. The task is to ensure that AI infrastructure is integrated into electricity markets in ways that reflect actual system conditions rather than force households to absorb avoidable risk.

Policymakers should restore a clearer link between grid conditions and large-scale electricity consumption. When prices fail to reflect real-time system constraints, even flexible demand behaves as if they are fixed. DOE highlighted the role of price-responsive demand in its 2024 *Pathways to Commercial Liftoff: Virtual Power Plants* report, noting how exposing large loads to the true cost of power can unlock flexibility the grid increasingly requires. [43] Enabling data centers to function as "virtual power plants" that self-regulate and adjust consumption can allow AI growth to support grid stability rather than translate short-run stress into permanent household cost increases.[44]

The 2025 ITIF report, "The United States Needs Data Centers, and Data Centers Need Energy—But That's Not Necessarily a Problem," offers several recommendations to DOE, FERC, and data centers themselves on how to support data centers as partners in peak demand management. [45]

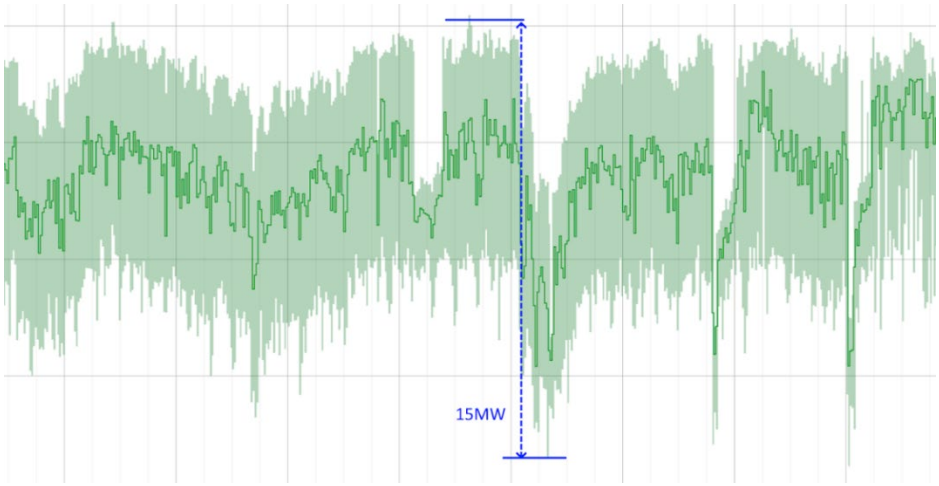
One particularly important reform would require RTOs to explore how they can broadcast transparent, machine-readable congestion and price signals in real time. If a data center's workload scheduler can see an impending price spike, it can automatically defer discretionary computing tasks or shift load across regions. When large loads reduce consumption during peak stress, they lessen the need to dispatch the most expensive marginal generators that set the clearing price for the entire market.

CONCERN 4: AI WORKLOADS THREATEN GRID RELIABILITY

Concerns about grid reliability are often highly localized rather than system-wide. Power interruptions and equipment failures typically manifest at specific pinch points such as substations, feeders, or transformers, where legacy infrastructure was not designed to accommodate rapid, concentrated shifts in load. While the magnitude of AI demand creates a baseline pressure on these assets, data centers are only one contributor to a much broader surge in electrification across transport, buildings, and industry (as covered in concern 1).

The more distinctive reliability challenge from AI workloads comes from how the consumption of this workload behaves. Figure 6 shows time-series data from Google illustrating these dynamics.

Figure 6: Time-series plot of power draw from an AI workload[46]



Data center power is on the y-axis and time on the x-axis. The data shows that AI workloads create an almost 15 MW swing in load, causing power to spike from approximately 1 MW to 15 MW.[47] The fluctuations are due to the synchronous operations of the GPUs.[48] During a job, all GPUs work together in a tightly coordinated manner. This leads to periods when power consumption is extremely high during active computation, followed by brief, sharp drops when the GPUs are waiting for data or synchronizing with one another. This means power delivery needs to be sized for those much higher peaks.

Concern about the behavioral dimension of AI workloads affecting grid reliability, specifically because of their extreme volatility, is therefore a valid one policymakers should address.

AI Reliability Risks Vary by Workload Profile

In practice, AI workloads fall into two dominant categories—training and inference—and each interacts with grid assets in fundamentally different ways. Treating these workloads separately makes clear how their load profiles translate into distinct stresses on grid infrastructure and where reliability risks actually arise.

Training workloads is a finite process that ends once a model reaches a target level of accuracy. During training, GPUs cycle through several distinct phases. In forward propagation, the GPU processes data at a sustained, high power draw. This is followed by backward propagation, wherein the model updates its parameters and power demand spikes in short bursts. Periodically, the system enters checkpoint phases, brief pauses when the model saves its progress, during which power consumption drops sharply, similar to a pause screen in a video game. Even then, GPUs do not fully idle. They typically maintain a high baseline load, often around 60 to 70 percent of peak, so they can immediately resume computation.[49]

The repeated transitions between sustained high load, rapid spikes, and partial idling create a characteristic sawtooth power profile for AI training. A 2025 paper, *AI Load Dynamics: A Power Electronics Perspective* by Yuzhuo Li and Yunwei Li at the University of Alberta illustrates this.[50] The researchers measured GPU power consumption during checkpoint events while fine-tuning OpenAI's GPT-2 and created time-series recordings of how power use changes over time. The data shows GPU power consumption sitting idle, then jumping abruptly to a high, jagged plateau of activity (like the teeth on a saw blade) before dropping suddenly back down—a pattern that repeats throughout the training run. For grid operators, this pattern matters because it introduces frequent and steep load transitions rather than a smooth or predictable demand curve.

The primary reliability risk associated with a sawtooth load profile is thermal cycling. A transformer is essentially a large, oil-filled vessel containing copper windings. When a data center draws huge megawatts, those copper windings heat up and physically expand. When the workload hits a checkpoint and demand drops sharply, the windings cool and contract. In an AI training environment, this cycle of heating and cooling can occur dozens of times per day. The repeated expansion and contraction place mechanical stress on the copper windings and the surrounding paper insulation, gradually causing the materials to become brittle and crack.[51] This process, known as thermal fatigue, follows the same principle as bending a paperclip back and forth until it breaks—it isn't the total weight that breaks it, but the repeated motion. These transitions can make the transformer undergo years' worth of mechanical fatigue in a matter of months.[52]

Inference workloads, by contrast, are not finite jobs but rather ongoing services. GPUs used for inference sit at a low-power idle state while waiting for user requests. When a request arrives, the GPU rapidly ramps up to near-peak power to process it as quickly as possible, then drops back down once the task is complete. This cycle repeats continuously, producing short bursts of high power rather than sustained load.

Li and Li also examined power consumption patterns during inference operations run on Meta's Llama 3.1 model, finding that power draw cycles from idle to peak and back again in fractions of a second.[53] The distinguishing feature of inference is not its average power demand, but rather the speed and frequency of these load changes. The pattern is a rapid flickering between high and low that never fully settles, which differs markedly from both training workloads and traditional data center computing.

For inference workloads, the dominant stress falls on the grid's electronic control systems—the equipment that keeps voltage steady and determines when to shut circuits off for safety. A rapid surge of AI inference requests can cause momentary drops in electrical pressure called voltage sags.

Think of the grid like a water pipe: if one were to suddenly throw open a massive valve and then slam it shut every half-second, the water pressure in the whole plumbing system would fluctuate. In the grid, this erratic pressure can cause two specific types of damage. First, it can create electronic fatigue as substation sensors and capacitors are forced to fight these millisecond-speed pressure waves thousands of times a day, causing their internal components to degrade prematurely. Second, because the speed and intensity of these AI spikes look like a short circuit, the grid's automated safety breakers can be tricked into tripping and causing a false-positive blackout.[54]

Policymakers Should Incentivize Data Centers to Internalize Volatility Costs

To protect the reliability of the American power system, policymakers should establish incentives that reward data centers for internalizing the costs of their volatility. These incentives should specifically target the two key physical risks to grid infrastructure: thermal fatigue and electrical pressure instability.

First, regulators should reward the use of advanced software that orchestrates compute to reduce peak electrical stress. For example, Google's DeepMind has applied AI to optimize the massive amount of electricity required for its data centers. By continuously analyzing vast amounts of electrical data related to server loads, power distribution, and the energy consumed by cooling equipment (e.g., pumps, chillers, and cooling towers), their AI system learned to make precise, real-time adjustments to operate these systems more efficiently. This capability led to a 40 percent reduction in the energy used for cooling alone, which translated to a 15 percent reduction in the data center's overall power usage.[55] Similarly, start-ups such as Emerald AI have demonstrated that software "mediators" can reduce an AI cluster's power draw by 25 percent during peak grid stress without violating performance guarantees.[56]

Second, regulators should incentivize the use of on-site shock absorbers to maintain voltage stability (electrical pressure). Microsoft, for one, has begun scaling such technology across its newer campuses. Traditionally, data center batteries sit idle most of the time, waiting for a blackout. Microsoft's system repurposes these massive lithium-ion banks to respond to the grid in real time. With a reaction time of approximately 80 milliseconds, the facility can "pull" power from the batteries to stabilize grid frequency or "absorb" the micro-spikes caused by AI workload bursts.[57]

Unfortunately, while such examples demonstrate what is technically feasible, existing policy frameworks do little to encourage this kind of behavior at scale. The AI Action Plan calls for data centers to "optimize existing grid resources as much as possible" through better management technologies, transmission upgrades, and new ways for large consumers to manage their usage during critical periods.[58] But the plan stops short of

creating concrete incentives or performance standards that would reward data centers for investing in internal load smoothing, ramp-rate control, or reliability-enhancing power management.

Policymakers should not only treat verified load profile management as a grid-supporting service and design incentives accordingly, but also support an industry code of practice, developed collaboratively by data center operators, utilities, and RTOs, that defines clear standards for on-site buffering, load smoothing, and ramp-rate control. Congress should direct FERC to make adherence to the code a condition of favorable interconnection terms, while property and business interruption insurers should offer more favorable premiums to operators that demonstrate compliance—creating parallel regulatory and market-based pressure that reinforces adoption at scale.

CONCERN 5: AI WORKLOADS STRAIN LOCAL WATER RESOURCES

A fundamental fact of modern computing, rooted in thermodynamics, is that nearly all electrical energy consumed by a processor is ultimately converted into heat.[59] As AI workloads grow more power intensive, the challenge of removing that heat has become a central design constraint for data centers. For decades, most facilities relied on air-based cooling systems, using large volumes of conditioned air to dissipate heat. That approach was viable when server power densities were lower. It is increasingly inadequate for modern AI hardware, whose heat is generated at much higher intensity and in much more concentrated spaces.

To manage these thermal loads, data center operators are rapidly shifting toward liquid-based cooling systems. Water and other specialized coolants are far more effective than air at absorbing and transporting heat away from high-density AI servers. This transition improves energy efficiency and enables higher-performance computing. But it has also raised environmental concerns, particularly around water use.

These concerns fall into two related categories. The first is local and ecological. Liquid cooling systems can increase water withdrawals at specific sites, raising fears about competition with municipal water supplies, agriculture, or fragile local ecosystems, especially in water-stressed regions. Communities worry that large data centers could exacerbate scarcity, strain infrastructure, or redirect water away from essential uses.

The second concern is systemic and perceptual, centered on how water use is framed in relation to AI activity. Media coverage has increasingly translated data center water consumption into vivid, per-task analogies that portray everyday AI use as environmentally wasteful. A September 2024 *Washington Post* article titled “A bottle of water per email: the hidden environmental costs of using AI chatbots” cites estimates that training GPT-3 is comparable to the water needed to produce 100 pounds of beef and

Meta's training of its LLaMA-3 model has been estimated to use 22 million liters of water, which the article equates to producing more than 4,000 pounds of rice.[60] For AI use, the article cites estimates that generating a 100-word email with GPT-4 could require the equivalent energy of roughly half a liter of water.

Much like the "hogging the grid" critique discussed in Concern 2, these comparisons frame AI resource use as inherently low value or socially wasteful, implicitly questioning whether the benefits of AI justify the consumption of scarce natural resources.

Misleading Water Metrics and Mismatched Comparisons

Many of the claims about data center water use rely on comparisons and calculations that do not withstand closer scrutiny. Widely cited figures, from water-per-email estimates to analogies equating AI systems with beef production or staple crops, are often built on inconsistent assumptions, mismatched units, or selective accounting choices that exaggerate the apparent impact of AI workloads.

Consider the comparison that training a model such as GPT uses as much water as producing 100 pounds of beef. The issue is not necessarily the number itself, but the unit of comparison. It measures a one-off computational process, whose output is used billions of times afterward, against 100 pounds of beef, which must be produced anew each time people want 100 pounds of beef. These are not comparable units.

A fairer comparison would look at entire facilities. Take Colossus 2, xAI's Memphis facility, one of the largest AI data centers in the world. A bottom-up estimate puts its full annual water footprint at 346 million gallons. While that sounds immense, a single high-volume In-N-Out Burger location carries a total footprint of roughly 147 million gallons when the water required to raise the cattle is accounted for.[61] In other words, one of the most powerful data centers consumes the water equivalent of just two and a half fast-food restaurants. Yet, no one is calling for a moratorium on burger joints to save the local watershed.

The per-task framing that dominates media coverage makes things worse. The method critics use is simple: take a facility's total water use, divide it by the number of queries processed, and present that as the water cost of generating an email or an image. But a data center's cooling systems do not switch on for each individual request. They run continuously, at roughly the same intensity, whether the facility is processing 10 queries or 10 billion. The water consumed maintaining operating temperature would be consumed regardless. Attributing a share of that fixed overhead to each individual query implies that sending one more email causes the cooling towers to draw more water, which it does not. The per-task figure is not a measure of what a query costs. It is the facility's total water bill divided by its output, presented in a way that makes routine AI use look

environmentally consequential when the underlying math does not support that conclusion.

The result is a distorted debate driven in part by anti-data-center rhetoric rather than sound evidence. AI data centers do use water, and in some locations, that use may raise legitimate concerns. A serious discussion requires moving past spurious narratives and focusing on the real determinants of water risk.

Water Use Is Not the Same as Water Harm

It is true that AI workloads generate far more heat than traditional computing does. One way to quantify the increased cooling demands of AI workloads is to look at a metric called Thermal Design Power (TDP). TDP is primarily a chip-level specification a manufacturer can provide for thermal engineers. It is a power number, not a temperature, and represents the amount of heat (in watts) the cooling system must be able to remove when the chip runs under typical load. For example, a CPU with a 125 W TDP means the cooler should be able to carry away 125 W of heat.

TDP of modern AI chips is high and typically escalates with each new generation. NVIDIA's flagship data center GPUs have jumped from the A100's 400 W TDP in 2020 (SXM4 model) to the H100's 700 W in 2022, and its new Blackwell B200 is reported to have a TDP of 1,000 W. This trend is also evident in other major players. AMD's Instinct accelerators have increased from 500 W for the MI250X to 750 W for the MI300X, and Google's seventh-generation Ironwood TPU and Intel's Gaudi 3 are both reported to have a TDP of 600 W.[62]

However, all the claims focus on how much water is used, rather than how much water is used without being replenished. This is a critical distinction because many companies have active water replenishment initiatives to offset their consumption. Google, Microsoft, Meta, and Amazon have all pledged to be water positive by 2030, meaning they plan to put more water back into the environment than they use.[63]

The Post article notes that Google's 2024 environmental report shows that it had "replenished only 18 percent of the water it consumed—a far cry from the 120 percent it has set as a goal by 2030."[64] But its 2025 report released in June shows that it has raised that number to 64 percent already—and if it keeps pace at that rate of change, it will reach 120 percent by the year 2027.[65] This has been achieved by funding more than 100 local stewardship projects, such as restoring wet meadows in California's Central Valley to act as natural sponges for groundwater recharge and investing in irrigation efficiency in the Colorado River Basin. These projects are strategically placed in the same water-stressed regions where their data centers operate, helping ensure that the replenishment happens in the actual communities being impacted.

This progress also shows why replenishment volume alone is not the right metric. Stories focused solely on gross water consumption figures, while attention-grabbing, are a distraction from the questions policymakers should be asking, such as if the replenished water is of good quality and supports the health of the surrounding ecosystem. If water is returned to its source at a much higher temperature, it can cause thermal pollution. This heated water can disrupt fragile aquatic ecosystems by reducing oxygen levels and harming local wildlife.

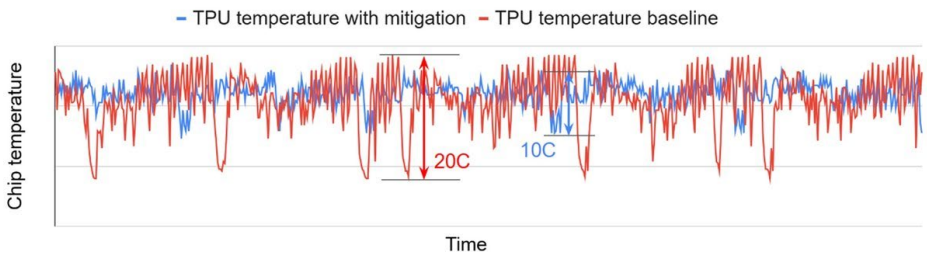
Moreover, if a large volume of water is withdrawn too quickly, or from a region that is already water stressed, it can disrupt the natural flow of rivers and strain the local supply for other essential uses such as agriculture and drinking water. What matters is the balance between AI infrastructure and the health of the watersheds that sustain it.

Changing Power Demands Create New Cooling and Water Constraints

What is largely missing from debates about data center water use is that AI workloads introduce fundamentally new thermal challenges. AI infrastructure does not merely generate more heat; it also produces heat in volatile bursts and in new, hard-to-reach hotspots created by advanced chip packaging, pushing cooling systems beyond what they were originally designed to manage.

Figure 7 is a time-series plot that shows the temperature fluctuations of a Google TPU chip over time, comparing a baseline scenario to one where a mitigation technique is used. The red line, representing the TPU temperature baseline, shows a wide and volatile range of temperatures, with sharp and rapid swings of up to 20°C. The blue line represents the mitigated scenario and shows how software controls to actively smooth power draw across the chip can reduce the intensity of temperature swings by roughly 50 percent.

Figure 7: Chip temperature fluctuations of Google's TPU chips[66]



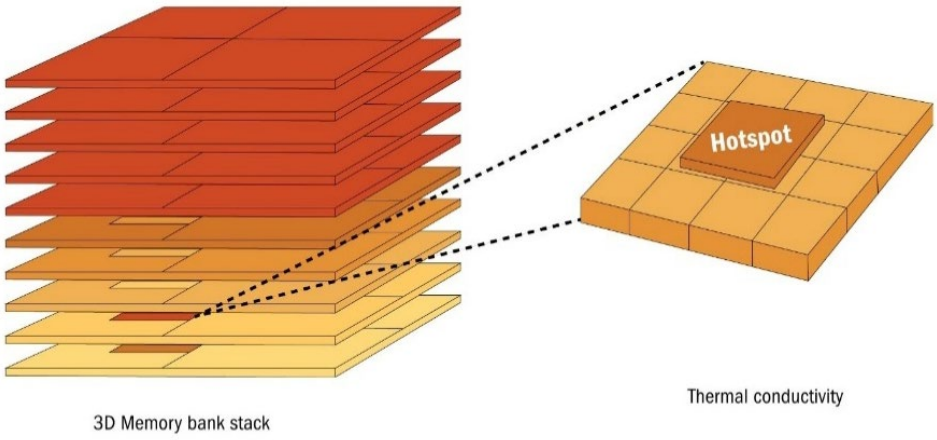
As the power draw of the chip rapidly spikes and drops, its temperature follows suit with equally fast and significant swings. The volatility of these thermal fluctuations can vary significantly depending on the type of workload (training versus inference) and task.

Importantly, this is not the whole chip's average temperature; it's the hottest point on one die, what engineers usually call the hotspot temperature. That's the spot most likely to fail first, and it's where rapid fluctuations matter most. Even if the overall chip temperature looks tame, the hotspot can be swinging significantly every few seconds, which can lead the chip to degrade and even fail. This occurs because a chip is made of multiple materials, such as silicon, copper, and solder, that each expand and contract at different rates when they heat up and cool down. Wildly fluctuating temperatures cause these materials to continuously expand and contract. Over time, this constant stress from the temperature fluctuations can cause microscopic cracks to form and grow in the solder joints and interconnects, eventually leading to a complete failure, much like how repeatedly bending a metal wire may cause it to break.

AI workloads are also changing where these hotspots appear on a chip, presenting a new challenge for thermal engineers. On traditional server CPUs, the main hotspot may be where the main processing cores are. However, modern AI chips use advanced methods to place components much closer together, either by putting them side-by-side on a small, shared platform or by stacking them directly on top of one another.[67] This new level of integration is leading to the creation of new hotspots.

The location of memory on the AI chips is a prime example of a new hotspot. As AI models grow more complex, the systems that support them demand memory with greater capacity, faster throughput, lower latency, and better energy efficiency.[68] To meet this demand, a key innovation is HBM, which involves stacking multiple layers of memory chips vertically on top of each other, as shown in figure 8. This is a critical development that has helped solve the "memory wall" bottleneck between a processor's speed and the speed of the memory it communicates with.[69]

Figure 8: Heat buildup in 3D-stacked memory[70]



The challenge is that heat can only escape in two main directions: upward toward a heat sink or sideways through the chip's edges. Memory layers in the middle of a stack are effectively "trapped."[71] They cannot easily release heat upward because layers above them block the path or downward because layers below them do the same. Heat from chips higher or lower in the stack also spreads into these middle layers, moving both vertically and sideways like water in a puddle. The center of the stack ends up furthest from any escape route, turning it into the hottest point.

The complex challenges of escalating TDP, rapid thermal fluctuations, and new hotspot locations on AI chips illustrate why traditional cooling methods are no longer sufficient. When dozens of these chips are put into a single server and thousands of servers into a data center, the problem isn't just cumulative, it's multiplicative. The heat generated at the chip level cascades, overwhelming traditional air-cooling systems and forcing new innovations at every level of the infrastructure stack, from direct-to-chip liquid cooling to the design of the entire data center's thermal management.

Data Centers Are Actively Reducing Cooling and Water Intensity

Cooling is one of the largest expenses in running a data center, second only to electricity.[72] Operators are therefore strongly motivated to reduce that cost. From chip packaging to rack design, and from AI-driven thermal optimization to city-scale district cooling, innovations are reshaping how data centers manage heat while cutting energy and water use.

Cooling Innovations in Data Centers

Starting with chip-level innovations, companies are exploring cooling methods built directly into the chip package itself. SK hynix, which is a leader in HBM, is using a proprietary technique called Molded Underfill (MR-MUF), which fills the tiny air gaps between stacked chips with a heat-conducting material, allowing heat to escape more efficiently and keeping the chips cooler.[73] TMSC, meanwhile, is developing a technology called the Integrated Micro Cooler (IMC-Si) that etches tiny, microscopic fluid channels directly into the body of the silicon interposer.[74] The coolant flows through these channels, which are located just a few hundred microns away from the heat-generating transistors on the chips above them, allowing for extremely efficient heat removal before it can even spread across the rest of the chip.

At the server level, the most significant advance is Direct-to-Chip Liquid Cooling (DLC). It is designed to take heat away from CPUs and GPUs by placing thin metal cold plates made of copper or aluminum directly on top of these chips, Inside each cold plate are tiny channels that carry coolant through, allowing the metal to absorb heat from the chip and transfer it into the fluid. The coolant never touches the electronics; it only flows through the sealed cold plate.

The type of liquid used varies. While water is an excellent conductor of heat, it is also electrically conductive and can corrode components. To prevent damage, data centers use specially treated and deionized water. Once the coolant has absorbed heat from the chips, the heated liquid is pumped back to a Coolant Distribution Unit (CDU). There, a liquid-to-liquid heat exchanger removes the heat from the server coolant loop without the fluids ever mixing. The now-cooled liquid is sent back to the chips to repeat the cycle. Finally, the extracted heat is released outside the system. NVIDIA's most cutting-edge rack architecture uses DLC as a standard feature. Its GB200 NVL72 rack-scale system, which houses 72 GPUs and 36 Grace CPUs, has cold plates mounted directly on the chips and a built-in CDU.[75]

Another server-level innovation gaining traction is liquid immersion cooling, wherein entire servers are fully submerged in a tank of a nonconductive, dielectric fluid.[76] Because the fluid is not electrically conductive, the servers can run while fully immersed, and the liquid absorbs heat directly from all the components, not just the CPUs and GPUs. Immersion cooling comes in two forms. In single-phase systems, the fluid behaves much like water in a radiator: it absorbs heat, warms up, and is then pumped out to a cooling unit before being cycled back in. In two-phase systems, the fluid is designed to boil at relatively low temperatures. When components get hot, the liquid touching them instantly turns to vapor, carrying the heat upward. The vapor then condenses back into liquid on cooler surfaces of the tank and drips back down to repeat the cycle. This boiling-and-condensing loop makes two-phase cooling extremely efficient at pulling heat away right where it is generated, but it also requires more specialized fluids and equipment.

A data hall–level strategy is adiabatic cooling, which uses evaporation to lower air temperature. In this approach, outside air is drawn through damp pads or sprayed with a fine mist before it enters the facility. As the water evaporates, the air cools down, and this cooler air is then circulated through the data hall to absorb heat from the servers. Because the process avoids energy-hungry chillers, it makes it an efficient option in dry climates where evaporation works best. The trade-off, however, is that adiabatic systems require a reliable supply of water, and their sustainability depends on where and how that water is sourced.

These innovations help improve data center energy and water responsibility. By capturing heat directly from the chips, servers, and racks, they drastically reduce the need for energy-intensive air conditioning and fans, allowing more power to be used for computing.

Thermal Optimization

Thermal optimization is a strategy data centers are using to optimize cooling operations and reduce costs by dynamically matching cooling to a facility's IT load in real time, preventing overcooling. A network of rack-level

sensors continuously collects temperature and power data, which is fed to an AI system that automatically calculates and makes precise adjustments to cooling units, fans, and pumps. This ensures that only the necessary amount of cooling is delivered, eliminating manual oversight and minimizing wasted energy.

Siemens is a leading provider of such systems and has deployed its White Space Cooling Optimization in financial data centers using hundreds of wireless rack-level sensors. In one facility, the system cut the number of cooling units in operation from 72 to 35 and reduced cooling energy consumption by approximately 70 percent.[77]

District Cooling

District cooling is a strategy that moves beyond managing individual data centers to coordinating cooling at the urban scale. Instead of each facility relying on its own equipment, a centralized plant produces chilled water and distributes it through underground pipes to a network of connected buildings. This allows data centers to forgo the installation and maintenance of their own energy-intensive chillers and cooling towers, offloading their needs to a more efficient, large-scale system.

One prominent example is the Deep Lake Water Cooling (DLWC) system in Toronto, Canada, operated by Enwave Energy Corporation. Equinix's Toronto data center is a major customer. The DLWC system draws frigid water from deep in Lake Ontario, pumps it to a central transfer station, and runs it through heat exchangers that pass the coolness to a closed-loop system serving downtown buildings. By leveraging this steady natural cold source, the system eliminates much of the need for mechanical chillers and cooling towers. Today, it connects more than 180 buildings in Toronto, and Enwave has reported that it saves 220 million gallons of water annually.[78]

Policymakers Should Establish a Common Standard for Data Center Water Accounting

Many of the largest AI data center operators have pledged to be water positive by 2030, an ambitious and welcome commitment. But without an agreed standard for how water use is measured and replenishment verified, these pledges are difficult to assess. A company can credibly report progress toward water positivity while measuring something entirely different from its peers.

Companies make different choices about what to count. Some include only the water that evaporates directly from their cooling towers on site. Others also account for the water embedded in their supply chains, for instance, the ultra-pure water used to manufacture the chips inside their servers. Some count only freshwater drawn from rivers and aquifers, while others include rainwater absorbed by crops grown to feed cattle raised near their

facilities. The result is that two companies can both claim to be on track toward water positivity while measuring entirely different things.

Congress should direct EPA, working with NIST, to develop a standardized water accounting methodology for large data centers—one that specifies what categories of consumption must be reported, how replenishment claims are verified, and how net consumption should be indexed against the stress level of the local watershed where the water is actually being withdrawn—because the same volume of water used matters far more in, say, drought-prone Arizona than in water-abundant Oregan or Washington.

Policymakers Should Incentivize Lower-Withdrawal Cooling Technologies in Water-Stressed Regions

As explained earlier, water use is not the same as water harm. But in specific water-stressed regions, the act of withdrawal itself—even if the water is later replenished—may deplete local aquifers, reduce river flows, and strain supplies that communities depend on for drinking water and agriculture. A data center drawing on already-strained aquifers in Arizona or West Texas presents an entirely different risk profile from one operating in the rain-abundant Pacific Northwest.

Congress should direct EPA to identify and designate water-stressed regions, drawing on existing federal drought and water scarcity data, and within those regions create targeted incentives for data centers to adopt lower-withdrawal cooling technologies. The choice between wet evaporative, dry, and hybrid adiabatic cooling systems can produce dramatically different withdrawal profiles for facilities of identical size. Wherever data centers are sited in water-stressed areas, policymakers should encourage the use of lower-withdrawal alternatives, recycled municipal wastewater as a cooling source, and hybrid approaches that reserve wet cooling for only the hottest periods.

Policymakers Should Require Water Productivity Reporting for Data Centers in Water-Stressed Regions Unable to Switch Cooling Systems

Not every data center in a water-stressed region can immediately switch to lower-withdrawal cooling technologies. Retrofitting cooling infrastructure is expensive, technically complex, and, in some cases, physically constrained by a facility's existing design. For these operators, the most actionable near-term lever is maximizing the computational value extracted from every gallon withdrawn.

To illustrate why this matters, imagine two AI data centers handling the same workload in the same water-stressed region. Both use the same cooling architecture and withdraw the same amount of water per hour. But the first center runs older, less-optimized software and takes 100 hours to complete the job. The second runs a more efficient AI stack and finishes in 20 hours. Both withdraw the same amount of water per hour, but the

second withdraws five times less water in total to deliver exactly the same output. In a region where every gallon matters, that difference is significant.

Congress should direct NIST to develop a water productivity metric for data centers—one that measures computational output per unit of water withdrawn and require facilities operating in EPA-designated water-stressed regions to report against it. This would create accountability for operators who cannot yet transition to lower-withdrawal cooling, while incentivizing them to maximize the value of every gallon they do use.

Policymakers Should Incentivize Circular Heat and District Heating Integration

Policymakers should transition from viewing data center heat as a waste product to be mitigated to treating it as a community asset to be harvested. High-density AI clusters are particularly well suited for this because liquid cooling systems produce a more concentrated, steady stream of thermal energy than traditional air-cooled facilities do.

The Technological University of Dublin's Tallaght campus partnership with nearby Amazon Web Services (AWS) data centers to heat dormitories and lecture halls provides a blueprint for this synergy. Excess heat from the AWS data centers is collected through their water-based cooling system, which produces a steady stream of warm water.[79] That water is then pumped through insulated district-heating pipes to the Tallaght campus, where heat exchangers transfer the energy into the university's existing heating system for dorms and classrooms. The system now supplies most of the university's heating.

Arrangements such as this create a virtuous cycle. For data center operators, exporting heat lowers cooling demand and operating costs. For communities and institutions, heating local dormitories, hospitals, or greenhouses lowers their reliance on fossil-fuel boilers and reduces energy bills. At a system level, heat reuse turns AI infrastructure from a localized burden into a shared asset. This opportunity has shown particularly significant in regions such as Appalachia, where a 2026 report finds that data center waste heat could heat entire towns, supply fresh produce through community greenhouses, and support industrial co-location on a former coal plant.[80]

Policymakers should therefore support the integration of data centers into district heating networks by reducing the practical barriers that prevent heat reuse from happening at scale. In many cases, the challenge is not technical feasibility but rather coordination and upfront cost. Moving waste heat offsite requires basic infrastructure—such as heat exchangers to capture thermal energy, pumps to move heated water, and insulated pipes to deliver it to nearby buildings—that individual projects often struggle to finance or permit on their own.

Governments can play a catalytic role by helping cover these upfront infrastructure costs and by streamlining permitting for the pipes and equipment needed to move heat beyond the data center fence line. Just as importantly, policymakers can clarify the regulatory treatment of recovered heat so operators and local institutions can enter long-term agreements without uncertainty over whether heat transfers trigger utility-style regulation, resale restrictions, or new compliance obligations.

Policymakers should also require greater transparency and planning around heat reuse. As part of siting and environmental review processes, large new data centers could be asked to assess whether nearby campuses, hospitals, housing developments, or industrial facilities could realistically use recovered heat. Standardized assessment and contracting approaches would make it easier for communities to plan around this resource and for operators to integrate heat reuse into facility design from the start.

CONCLUSION

Data centers have become a lightning rod for anxieties about the AI economy. But the risks they pose—from rising household bills to grid instability to local water stress—are not inherent to AI technology. They are the predictable result of infrastructure systems still operating under analog-era rules. Today's friction stems from measurement frameworks that track raw electricity and water inputs while ignoring productive output, market designs that socialize peak costs rather than signal them, and planning processes that mistake administrative congestion for physical scarcity.

The choice for the United States is not between technological growth and environmental stewardship. If the policy framework is right, AI infrastructure can strengthen the grid rather than strain it, stabilize prices rather than inflate them, and transform heat and flexible demand into system assets. Getting data centers right is not about capping AI. It is about updating the institutions that govern how energy, water, and infrastructure performance are measured, priced, and managed in a digital economy.

REFERENCES

1. Adapted from Hamilton Lombard, "The 2024 JLARC Report on Data Centers in Virginia," Weldon Cooper Center for Public Service, University of Virginia, accessed March 19, 2026, https://www.coopercenter.org/research/jlarc-report-data-centers-virginia.
2. Synergy Research Group, "The World's Total Data Center Capacity is Shifting Rapidly to Hyperscale Operators," June 24, 2025, https://www.srgresearch.com/articles/the-worlds-total-data-center-capacity-is-shifting-rapidly-to-hyperscale-operators.
3. "Introduction to NVIDIA DGX H100/H200 Systems," NVIDIA, accessed March 19, 2026, https://docs.nvidia.com/dgx/dgxh100-user-guide/introduction-to-dgxh100.html.
4. Adapted from Chris Mellor, "Rambus HBM Subsystem More Than Doubles HBM2E Speed," *Blocks and Files*, August 16, 2021, https://blocksandfiles.com/2021/08/16/rambus-hbm-subsystem-more-than-doubles-hbm2e-speed.
5. Lauren Leffer, "The AI Boom Could Use a Shocking Amount of Electricity," *Scientific American*, October 13, 2023, https://www.scientificamerican.com/article/the-ai-boom-could-use-a-shocking-amount-of-electricity.
6. Dylan Patel et al., "AI Datacenter Energy Dilemma - Race for AI Datacenter Space," *SemiAnalysis*, March 13, 2024, https://semianalysis.com/2024/03/13/ai-datacenter-energy-dilemma-race.
7. International Energy Agency (IEA), "Energy and AI" (Paris: IEA, April 2025), https://www.iea.org/reports/energy-and-ai.
8. Adapted from "Increase in Electricity Demand by Sector, Base Case, 2024–2030," International Energy Agency (IEA), last updated April 10, 2025, https://www.iea.org/data-and-statistics/charts/increase-in-electricity-demand-by-sector-base-case-2024-2030.
9. Arya Tschand et al., "MLPerf Power: Benchmarking the Energy Efficiency of Machine Learning Systems from μWatts to MWatts for Sustainable AI" (working paper, arXiv, October 2024), https://arxiv.org/html/2410.12032v2.
10. "Best Intelligence per Joule," *SambaNova Systems* (blog), November 12, 2025, https://sambanova.ai/blog/best-intelligence-per-joule.
11. "The Power Crunch Threatening America's AI Ambitions," *Financial Times*, December 7, 2025, https://ig.ft.com/ai-power.
12. "RTOs and ISOs," Federal Energy Regulatory Commission (FERC), accessed March 19, 2026, https://www.ferc.gov/power-sales-and-markets/rtos-and-isos.
13. Ibid.
14. Oliver Milman, "More than 200 environmental groups demand halt to new US datacenters" *The Guardian*, December 8, 2025, https://www.theguardian.com/us-news/2025/dec/08/us-data-centers.
15. Jason Ma, "Virginia House of Delegates May Consider Temporary Data Center Moratorium," *Data Center Dynamics*, February 3, 2026,

https://www.datacenterdynamics.com/en/news/virginia-house-of-delegates-may-consider-temporary-data-center-moratorium/.

16. Laila Kearney and Tim McLaughlin, "Power Costs Soar in PJM Region as Data Center Demand Spikes," *Reuters*, August 7, 2025, https://www.reuters.com/business/energy/power-costs-soar-pjm-region-data-center-demand-spikes-2025-08-07/.
17. Steven Zhang and Chris Talley, "In 2024, Interconnection Queues Shrank for the First Time in Years," *Interconnection.fyi* (blog), June 1, 2025, https://www.interconnection.fyi/blog/2023-to-2024-queue-changes.
18. Adam Barth, Humayun Tai, and Ksenia Kaladiouk, "Powering a New Era of US Energy Demand," McKinsey & Company, April 29, 2025, https://www.mckinsey.com/industries/public-sector/our-insights/powering-a-new-era-of-us-energy-demand.
19. Lynne Kiesling, "The Future of Data Center Electricity Use and Microgrids," Reason Foundation (commentary), September 5, 2024, https://reason.org/commentary/the-future-of-data-center-electricity-use-and-microgrids.
20. EirGrid and SONI, *All-Island Generation Capacity Statement 2021–2030* (Dublin and Belfast: EirGrid and SONI, 2021), https://cms.eirgrid.ie/sites/default/files/publications/208281-All-Island-Generation-Capacity-Statement-LR13A.pdf.
21. "Data Centres in Ireland: A Strategy for Sustainable Growth," KPMG Ireland, November 2024, https://kpmg.com/ie/en/home/insights/2024/11/data-centres-in-ireland-strategy.html.
22. Charles Kennedy, "Amazon Scraps New Irish AI Facility Amid Power Grid Shortfall," *OilPrice.com*, July 25, 2025, https://oilprice.com/Latest-Energy-News/World-News/Amazon-Scraps-New-Irish-AI-Facility-Amid-Power-Grid-Shortfall.html.
23. "AI for Interconnection (AI4IX)," U.S. Department of Energy (DOE), Office of Electricity, accessed March 19, 2026, https://www.energy.gov/gdo/ai-interconnection-ai4ix.
24. Department of Energy AI Act, S. 4664, 118th Cong. (2024), https://www.congress.gov/bill/118th-congress/senate-bill/4664/text.
25. Richard A. Drom, "New Metrics for Measuring the Success of a Non-Profit RTO," *Energy Law Journal* 28, no. 2 (2007), https://www.eba-net.org/wp-content/uploads/2023/02/11-New_Metrics.pdf.
26. P.L. Joskow, "The Expansion of Incentive (Performance-Based) Regulation of Electricity Distribution and Transmission in the United States," *Review of Industrial Organization 65* (September 2024): 455–503, https://doi.org/10.1007/s11151-024-09973-x.
27. Ibid.
28. "Speed to Power: Accelerating the Path to a Clean Energy Future," DOE, accessed March 19, 2026, https://www.energy.gov/speed-to-power.
29. Joe Dammel, "What's Up With Rate Cases?" Fresh Energy, December 21, 2021, https://fresh-energy.org/whats-up-with-rate-cases.
30. Eliza Martin and Ari Peskoe, "Extracting Profits from the Public: How Utility Ratepayers Are Paying for Big Tech's Power," Harvard Law School, Environmental & Energy Law Program, March 2025,

https://eelp.law.harvard.edu/extracting-profits-from-the-public-how-utility-ratepayers-are-paying-for-big-techs-power/.

31. Steve Clemmer et al., "Data Center Power Play: How Clean Energy Can Meet Rising Electricity Demand While Delivering Climate and Health Benefits," Union of Concerned Scientists (UCS), January 21, 2026, https://www.ucs.org/resources/data-center-power-play.
32. "Electricity Markets," National Governors Association (NGA), accessed March 19, 2026, https://www.nga.org/electricity-markets/.
33. "An Introductory Guide to Electricity Markets," FERC, accessed March 19, 2026, https://www.ferc.gov/introductory-guide-electricity-markets-regulated-federal-energy-regulatory-commission.
34. Aishwarya Mahesh et al., "Are AI Datacenters Increasing Electric Prices?" *SemiAnalysis*, March 3, 2026, https://newsletter.semianalysis.com/p/are-ai-datacenters-increasing-electric.
35. "Understanding Wholesale Capacity Markets," FERC, accessed March 19, 2026, https://www.ferc.gov/understanding-wholesale-capacity-markets.
36. Mahesh et al., "Are AI Datacenters Increasing Electric Prices?"
37. Ibid
38. "Electric Reliability Council of Texas (ERCOT)," FERC, accessed March 19, 2026, https://www.ferc.gov/industries-data/electric/electric-power-markets/ercot.
39. "Energy Explained," U.S. Energy Information Administration (EIA), accessed March 19, 2026, https://www.eia.gov/energyexplained/
40. "Ratemaking Fundamentals and Principles," *Commissioners' Desk Reference Manual*, National Association of Regulatory Utility Commissioners (NARUC), accessed March 19, 2026, https://www.naruc.org/commissioners-desk-reference-manual/3-ratemaking-fundamentals-and-principles/.
41. Tyler Norris et al., "Rethinking Load Growth," Duke University, Nicholas Institute for Energy, Environment & Sustainability, 2025, https://nicholasinstitute.duke.edu/sites/default/files/publications/rethinking-load-growth.pdf; "U.S. electricity peak demand set new records twice in July," EIA, August 5 2025, https://www.eia.gov/todayinenergy/detail.php?id=65864.
42. Dylan Walsh, "Flexible Data Centers Can Reduce Costs—If Not Emissions," MIT Sloan School of Management, Ideas Made to Matter, October 20, 2025, https://mitsloan.mit.edu/ideas-made-to-matter/flexible-data-centers-can-reduce-costs-if-not-emissions.
43. *Pathways to Commercial Liftoff: Virtual Power Plants*, DOE, updated January 2025, https://climateprogramportal.org/wp-content/uploads/2025/06/LIFTOFF_DOE_VPP_2023.pdf.
44. Arthur van Benthem et al., "How Can We Improve the Efficiency of Electricity Pricing Systems?" University of Pennsylvania, Kleinman Center for Energy Policy, July 24, 2024, https://kleinmanenergy.upenn.edu/research/publications/how-can-we-improve-the-efficiency-of-electricity-pricing-systems/.
45. Robin Gaster, "The United States Needs Data Centers, and Data Centers Need Energy—But That Is Not Necessarily a Problem" (ITIF, November 24,

2025,) https://itif.org/publications/2025/11/24/united-states-needs-data-centers-data-centers-need-energy-but-that-is-not-necessarily-a-problem/.

46. Houle Gan and Parthasarathy Ranganathan, “Balance of Power: A Full-Stack Approach to Power and Thermal Fluctuations in ML Infrastructure,” *Google Cloud Blog*, February 11, 2025, https://cloud.google.com/blog/topics/systems/mitigating-power-and-thermal-fluctuations-in-ml-infrastructure.
47. Jeremie Eliahou Ontiveros, Dylan Patel, and Ajey Pandey, “AI Training Load Fluctuations at Gigawatt-scale—Risk of Power Grid Blackout?” *SemiAnalysis*, June 25, 2025, https://semianalysis.com/2025/06/25/ai-training-load-fluctuations-at-gigawatt-scale-risk-of-power-grid-blackout/#problem-1-managing-fast-power-fluctuations.
48. Gan and Ranganathan, “Balance of Power.”
49. Yuzhuo Li and Yunwei Li, “AI Load Dynamics—A Power Electronics Perspective,” University of Alberta (preprint), February 6, 2025, https://arxiv.org/pdf/2502.01647v2.
50. Ibid.
51. Ontiveros, Patel, and Pandey, “AI Training Load Fluctuations.”
52. Li and Li, “AI Load Dynamics.”
53. Ibid.
54. Ontiveros, Patel, and Pandey, “AI Training Load Fluctuations.”
55. Rich Evans and Jim Gao, “DeepMind AI Reduces Energy Used for Cooling Google Data Centers by 40%,” *The Keyword* (Google blog), July 20, 2016, https://blog.google/outreach-initiatives/environment/deepmind-ai-reduces-energy-used-for.
56. Marc Spieler, “How AI Factories Can Help Relieve Grid Stress,” *NVIDIA Blog*, July 1, 2025, https://blogs.nvidia.com/blog/ai-factories-flexible-power-use/.
57. “Microsoft Dublin: Grid-Interactive UPS for Frequency Regulation,” World Economic Forum, accessed March 19, 2026, https://initiatives.weforum.org/energy-and-industry-transition-intelligence/case-study-details/microsoft-dublin:-grid-interactive-ups-for-frequency-regulation/aJYTG0000000Y4v4AE.
58. The White House, “Winning the Race: America’s AI Action Plan,” July 23, 2025, https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf.
59. “Is This True That Computers Convert Electrical Energy Into Heat?” FlameIT, accessed March 19, 2026, https://flameit.io/kb/faq/is-this-true-that-computers-convert-electrical-energy-into-heat.
60. Evan Halper. "AI Is Already Wreaking Havoc on Global Energy Goals." *Washington Post*, September 18, 2024.https://www.washingtonpost.com/technology/2024/09/18/energy-ai-use-electricity-water-data-centers/.
61. Nicolas Bontigui and Dylan Patel, “From Tokens to Burgers: A Water Footprint Face-Off,” *SemiAnalysis*, January 15, 2026, https://newsletter.semianalysis.com/p/from-tokens-to-burgers-a-water-footprint.

62. Advanced Micro Devices, Inc., “AMD Instinct MI300X Accelerator Specifications,” accessed March 2026, https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html; Advanced Micro Devices, Inc., “AMD Instinct MI250X Specifications,” accessed March 2026, https://www.amd.com/en/products/accelerators/instinct/mi200/mi250x.html; Google Cloud, “TPU v5p Architecture and Performance,” accessed March 2026, https://docs.cloud.google.com/tpu/docs/v5p; Intel Corporation, “Intel Gaudi 3 AI Accelerator White Paper,” August 2025, https://www.intel.com/content/www/us/en/content-details/817486/intel-gaudi-3-ai-accelerator-white-paper.html.

63. Google, “2025 Environmental Report” (Mountain View, CA: Google, 2025), 41, 86, https://sustainability.google/reports/google-2025-environmental-report/; Microsoft, “Microsoft Will Be Water Positive by 2030,” September 21, 2020, https://blogs.microsoft.com/blog/2020/09/21/microsoft-will-be-water-positive-by-2030/; Meta, “Our Commitment to Water Stewardship,” August 19, 2021, https://tech.facebook.com/engineering/2021/8/water-stewardship/; Amazon, “Amazon to Be Water Positive by 2030,” November 28, 2022, https://www.aboutamazon.com/news/sustainability/amazon-to-be-water-positive-by-2030.

64. Halper, “AI Is Already Wreaking Havoc.”

65. Google, *2025 Environmental Report*.

66. Gan and Ranganathan, “Balance of Power.”

67. “2.5D and 3D IC Packaging,” ASE Global, accessed March 19, 2026, https://ase.aseglobal.com/3d-ic-packaging/.

68. Dylan Patel, Myron Xie, Tanj Bennett, et al., “Scaling the Memory Wall: The Rise and Roadmap of HBM,” *SemiAnalysis*, August 11, 2025, https://semianalysis.com/2025/08/12/scaling-the-memory-wall-the-rise-and-roadmap-of-hbm/.

69. Ibid

70. Adapted from Mehdi Elahi et al., “On the Thermal Vulnerability of 3D-Stacked High-Bandwidth Memory Architectures,” (preprint), August 30, 2025, https://arxiv.org/html/2509.00633v1.

71. Ibid

72. Jeremie Eliahou Ontiveros et al., “Datacenter Anatomy Part 2—Cooling Systems,” *SemiAnalysis*, February 13, 2025, https://semianalysis.com/2025/02/13/datacenter-anatomy-part-2-cooling-systems/.

73. “Rulebreakers’ Revolutions: MR-MUF Unlocks HBM Heat Control,” SK hynix Newsroom, July 30, 2024, https://news.skhynix.com/rulebreaker-revolutions-mr-muf-unlocks-hbm-heat-control/.

74. SemiVision Research, “Cooling is the New Architecture: TSMC’s IMC-Si and the Future of AI Packaging,” *TSPA Semiconductor*, July 3, 2025, https://tspasemiconductor.substack.com/p/cooling-is-the-new-architecture-tsmcs.

75. Marc Hamilton, “Chill Factor: NVIDIA Blackwell Platform Boosts Water Efficiency by Over 300x,” *NVIDIA Blog*, April 22, 2025, https://blogs.nvidia.com/blog/blackwell-platform-water-efficiency-liquid-cooling-data-centers-ai-factories/.

76. Kyle Chien, “A Guide to Data Center Cooling: Future Innovations for Sustainability,” *Digital Realty Blog*, March 7, 2025 https://www.digitalrealty.com/resources/articles/future-of-data-center-cooling.

77. Siemens, “Dynamically Match Cooling to IT Load in Real Time: How Artificial Intelligence Is Cooling Data Center Operations,” white paper, https://assets.new.siemens.com/siemens/assets/api/uuid:c815cadf-8221-479a-b396-2a2651bb92c7/Whitepaper-White-Space-Cooling-Optimization.pdf.

78. “Toronto’s Sustainable District Cooling: Enwave’s Deep Lake Water System,” Alfa Laval, January 2, 2025, https://www.alfalaval.com/media/stories/sustainability/toronto-sustainable-district-cooling-system/.

79. April Roach and Tasmin Lockwood, “This University Campus Is Heated by an AI Data Center: Your Home Could Be Next,” *CNBC*, January 27, 2026, https://www.cnbc.com/2026/01/27/data-centers-ai-district-heating-aws-amazon-ireland.html.

80. Deborah Stine, *Catching Heat: The Opportunities and Challenges of Using Waste Heat from Appalachian AI Data Centers* (ReImagine Appalachia, February 5, 2026), https://reimagineappalachia.org/wp-content/uploads/2026/02/Catching-Heat_Using-Waste-Heat-Generated-from-Data-Centers.pdf.

ABOUT THE AUTHORS

Hodan Omaar is a senior policy manager at ITIF's Center for Data Innovation. Previously, she worked as a senior consultant on technology and risk management in London and as an economist at a blockchain start-up in Berlin. She has an M.A. in Economics and Mathematics from the University of Edinburgh.

Mitalee Pasricha is a Google Public Policy Fellow with the Center. She is currently pursuing Bachelor of Science degrees in Environmental Sciences and Environmental Economics and Policy at the University of California, Berkeley.

ABOUT THE CENTER FOR DATA INNOVATION

The Center for Data Innovation studies the intersection of data, technology, and public policy. With staff in Washington, London, and Brussels, the Center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as open data, artificial intelligence, and the Internet of Things. The Center is part of the Information Technology and Innovation Foundation (ITIF), a nonprofit, nonpartisan think tank.

Contact: info@datainnovation.org

datainnovation.org