EBOOK

Beyond PUE: Data center efficiency in the AI era

A multidimensional efficiency framework for AI infrastructure

Executive summary

AI holds enormous potential. Adoption across industries is accelerating, with more than 75 percent of organizations regularly using AI in at least one business function.[1] As data center operators race to increase capacity, tokens per watt, and tokens per dollar, their use of energy and natural resources requires all of us within the data center ecosystem to think critically about the products and services we provide that make it possible — and about the ways in which we measure efficiency. Frameworks must evolve in step with the transformation driven by AI workloads.

It's worth noting that "AI data centers" were not a distinct category until just a few years ago. Most facilities were general-purpose data centers, though some also handled AI workloads. Consequently, traditional efficiency models were built for CPU-centric environments with predictable workloads and modest power densities. Power usage effectiveness (PUE), introduced by The Green Grid in 2007, was developed as a simple, standardized metric to help organizations measure how efficiently a data center used energy. It has been the industry standard ever since.

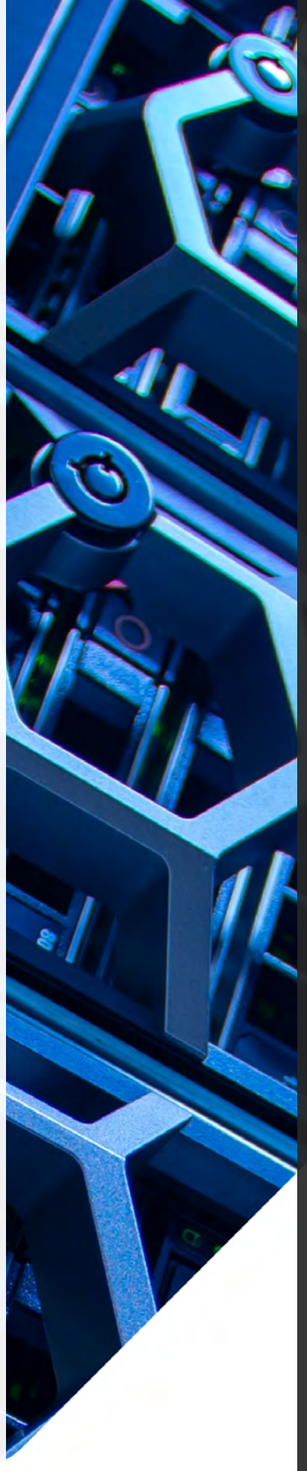
AI, however, operates on an entirely different scale. GPU clusters in megawatt-scale racks consume exponentially more power in bursts characteristic of training and inferencing. This has sparked debate around PUE's limitations and potential for misinterpretation as the nature of computing has evolved. While PUE provides a ratio of total facility energy against that consumed by IT equipment, advanced technologies that power AI and high-performance computing (HPC) are challenging its efficacy as hardware and infrastructure advance and thermal solutions such as liquid cooling and district heat reuse come to the fore.

Evaluating efficiency becomes a multidimensional equation, with key considerations at its core that have significant implications for the infrastructure, community, and bottom line. Among them:

- **How power is delivered** – What's the net energy loss from grid to rack to chip?
- **How compute is used** – Are servers sitting idle? Why?
- **How much fresh water is used** – Are cooling choices sustainable?
- **How much carbon is emitted** – Which mitigation factors are being employed?
- **How much energy can be reused** – What happens to waste heat?
- **How the facility interacts with the grid** – Are protection and reciprocation considered?

PUE is still a valuable part of the operational equation, but there are other factors that come into play when assessing the entirety of a data center's energy and resource usage in the AI era. The most suitable ways to measure efficiency are still being calibrated as operators seek to strike the right balance between capacity, demand, and utilization.

Partnership and co-innovation can drive better efficiency as technologies and standards evolve. In this eBook, we will **explore the benefits and limitations of PUE and the metrics that complement it**. Together, they form a framework that gives data center operators a more refined lens through which to evaluate efficiency and drive improvement across their facilities.



What's inside

4 **Which is better, higher or lower PUE?**
It's complicated.

6 **Has PUE improved?**
Legacy vs. new AI data centers

8 **A more nuanced view:**
Metrics that complement PUE

10 • Water usage effectiveness (WUE)

14 • Energy reuse effectiveness (ERE)

18 • Compute power efficiency (CPE)

22 • Carbon usage effectiveness (CUE)

26 • Grid-aware efficiency (GAE)

32 **Building a holistic framework for data center efficiency**

33 **Translate AI innovation into efficient infrastructure with Flex**

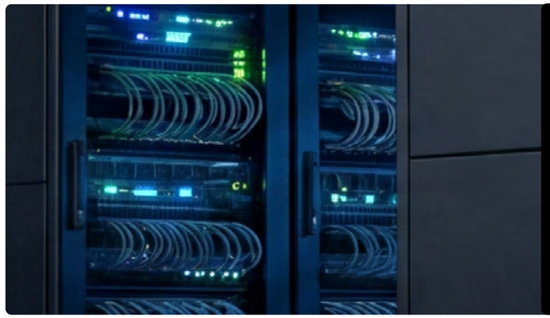
Which is better, higher or lower PUE? It's complicated.

According to the Uptime Institute, 67 percent of data center operators are very or somewhat concerned about improving energy performance for facility equipment.[2] And well they should be — energy consumption is a sizable portion of a data center's day-to-day operating costs. Data center electricity consumption is growing about 15 percent per year globally, but it is by no means spread evenly across countries.[3] By 2030, it is projected to increase 130 percent in the U.S. and 170 percent in China, the most significant regions. To put that in perspective, in 2024 data centers accounted for about 4 percent of electricity use in the U.S.[4] That will **more than double by 2030.**

Tracking PUE remains a standard industry practice, illuminating energy efficiency and operational performance over time. In general, if more energy is consumed by IT equipment rather than cooling, lighting, or infrastructure systems, a lower PUE will indicate better energy efficiency. Rising PUE suggests that more energy is going to cooling or power conversion rather than computing, and it can indicate infrastructure design flaws. But the situation can be nuanced, and **PUE should never be viewed in isolation.**

Among the metric's shortcomings:

- PUE only measures how efficiently power is delivered to IT equipment, not how effectively the IT equipment uses it to perform computing functions.
- Rapid data center expansion increases PUE temporarily; it settles as IT equipment comes online and balances the energy used by the infrastructure itself.
- Underutilized IT equipment can increase PUE because the space required to accommodate it still has to be lit and thermally managed, even if servers lie dormant or are only used intermittently.
- Rated power, which is the maximum amount of power an IT server could draw (though they rarely do), is often used as the basis for power and cooling infrastructure decisions, which puts power efficiency on the back foot from the outset.
- Shifting energy load classifications between categories (facility vs. IT equipment) can affect PUE in unintended ways.
- PUE does not take into consideration water usage or the data center's carbon footprint, both of which can have a meaningful impact on overall data center sustainability.



- Reusing the heat produced by increasingly dense racks to warm facilities on site and in the community is not reflected in PUE, nor are energy storage systems that bank or return excess energy to the grid.

Scenario: Efficient IT equipment

Is lower PUE better? Most of the time, yes. But look at what happens if greater IT equipment efficiency isn't matched by total facility energy improvements.

If a data center uses 1,000 kWh of total energy per day, and 500 kWh of that is consumed by IT equipment, the PUE is:

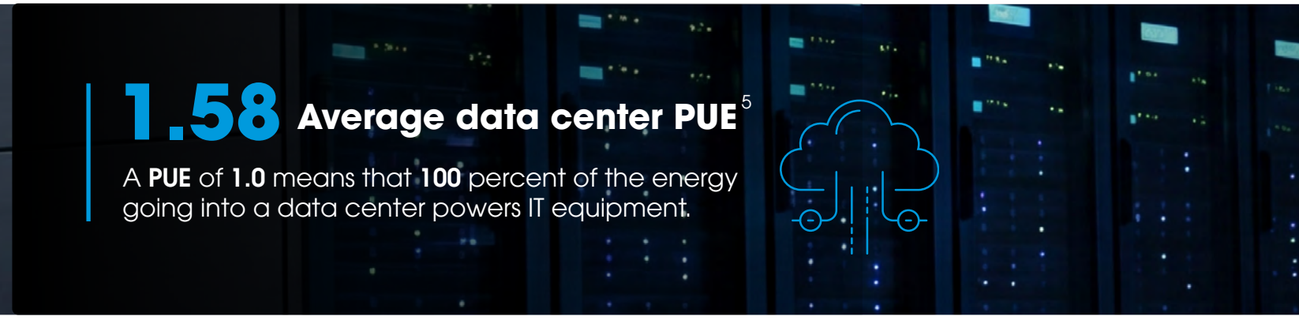
$$\textbf{PUE} = \frac{1000}{500} = 2.0$$

But if the IT equipment is more efficient, using just 300 kWh of energy per day while facility power use remains the same, the PUE becomes:

$$\textbf{PUE} = \frac{1000}{500 - 200} \approx 2.67$$

$$\text{PUE} = \frac{\textbf{Total Facility Energy}}{\textbf{IT Equipment Energy}}$$

- **Total facility energy** = all energy consumed by the data center, including cooling, lighting, power conversion, and IT equipment
- **IT equipment energy** = energy used specifically by computing, storage, and networking equipment



Has PUE improved?
Legacy vs. new AI data centers

In a survey that looked at PUE across a range of facility sizes, the Uptime Institute found that PUE varied widely and that new, larger facilities using technically advanced equipment and control systems fared better than older, less efficient ones.[6] Their PUE was 1.47, which is below the industry average of 1.58. Among hyperscalers, which build some of the largest data centers in the world, PUE tends to vary between 1.04 and 2.0.[7] In short, newer AI data centers are more efficient because they're built to be that way.

Nonetheless, **industry average PUE hasn't budged in at least seven years, remaining at roughly 1.55 to 1.58 despite advancements in data center infrastructure technologies.** This is due in part to the continued use of legacy infrastructure. According to the Uptime Institute, about half of all facilities have been in operation for more than a decade.[8] The organization theorizes that after The Green Grid introduced PUE in 2007, rapid improvements ensued as data center operators upgraded electrical systems, improved air flow management, and tackled other pragmatic changes until they reached a threshold that made further improvements too disruptive or cost prohibitive. Legacy facilities continue to influence global averages even as hyperscale sites achieve much lower PUE values.

Reality check: AI only accounts for 14% of power usage

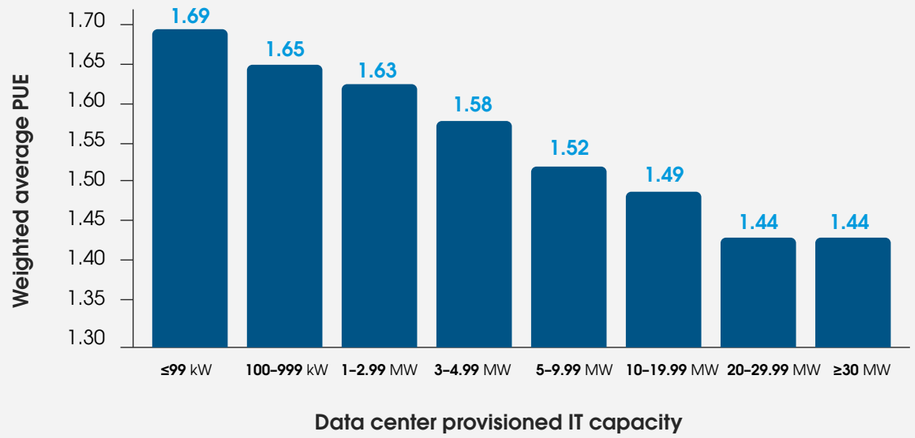
While AI gets all the hype, most workloads still run on traditional CPUs. Goldman Sachs estimates that AI currently accounts for just 14 percent of global data center power usage, well behind cloud computing and traditional workloads.[9] But hyperscalers and colocation providers are betting that AI workloads will increase exponentially and are building data center capacity accordingly.

It's important to note, however, that capacity and demand are not one in the same, and this has a direct impact on PUE.

Capacity is about potential — the total available infrastructure that can support data center operations, including power, physical space, cooling systems, and network connectivity. Demand is about actual usage. This further complicates PUE, in that racks within data centers may be drawing power in vastly different amounts depending on their purpose and workloads, yet the data center itself still requires the same infrastructure systems regardless of compute demand.

Weighted average PUE by data center IT capacity [10]

Efficiency improves as data centers increase in capacity size



A more nuanced view:
Metrics that complement PUE

PUE gives data center operators the big picture, but it's a bit like hammering a nail with a cannon — the precision is just not there. The same goes for its doppelganger, data center infrastructure efficiency (DCiE), an inverse calculation that simply presents the figure as a percentage (e.g., a PUE of 1.25 is a DCiE of 80 percent). Neither takes into account water usage, space utilization, carbon emissions, IT equipment performance, or renewable vs. nonrenewable energy sources.

For a more complete picture of operational efficiency, assessing several metrics can support better decision-making by helping data center operators understand resource usage, environmental impact, and compute productivity. Together, they present a multidimensional efficiency model for AI data centers.

METRICS COMPLEMENTING PUE	WHY IT MATTERS	EXAMPLE
WUE Water usage effectiveness	Highlights tradeoffs between energy and water efficiency	A data center in the Middle East might have a great PUE but a poor WUE due to the water required for cooling.
ERE Energy reuse effectiveness	Rewards energy recycling and encourages innovation	Capturing and reusing heat from racks elsewhere is energy-efficient, but PUE would likely stay the same or even go up.
CPE Compute power efficiency	Reflects compute productivity, not just energy ratios	For a given amount of compute, PUE would increase if CPE improved.
CUE Carbon usage effectiveness	Helps assess full environmental impact	Two data centers with the same PUE could have vastly different CUE scores if one uses renewable energy and the other uses fossil fuels.
GAE Grid-aware efficiency	Improves how the data center interacts with the grid, reducing peak grid stress and enhancing system reliability	Shifting AI training to times with available renewables avoids local overloads and helps keep system voltage stable. PUE does not capture this grid benefit and may remain unchanged.

(WUE)

Water usage effectiveness

As data centers proliferate, concern about the amount of water they consume is growing. Data centers compete directly with humans for fresh water, and they use an extraordinary amount of it to cool chips, servers, and other IT equipment — up to 5 million gallons per day, per facility.[11] The U.S. alone has more than 4,300 data centers using billions of gallons of water annually.[12] This puts WUE at a premium. **To protect this finite resource — and to preserve their ability to scale — data center operators must invest in innovation.**

$$\text{WUE} = \frac{\textbf{Data center water consumption} \text{ (liters)}}{\textbf{IT equipment energy consumption} \text{ (kWh)}}$$

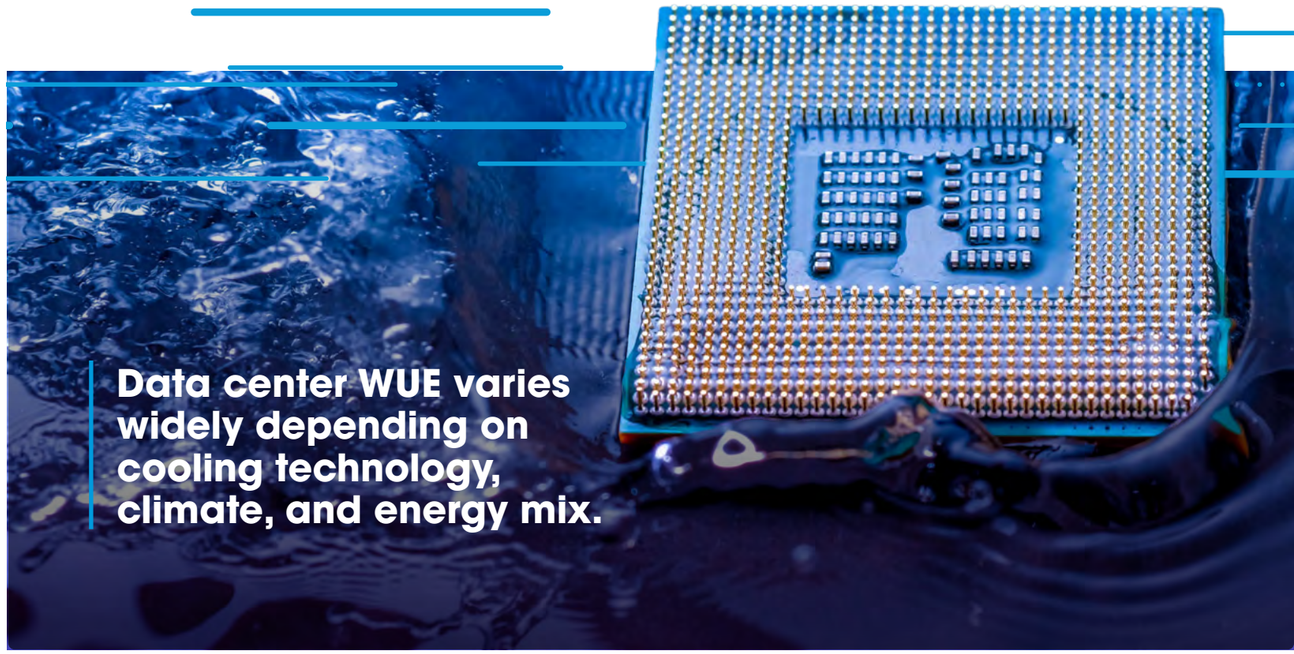
$$\text{EXAMPLE} \quad \frac{\textbf{100,000} \text{ liters}}{\textbf{50,000} \text{ kWh}} = \textbf{2.0 L/kWh}$$

For a detailed technical explanation of WUE, read White Paper #35 by The Green Grid.

AI data centers generate a lot of heat, which makes cooling an opportune place to start. There are two things to consider — cooling the equipment specifically and cooling the space altogether. Each cooling method has its benefits and drawbacks.

Traditional computer room air conditioning (CRAC) systems rely on mechanical refrigeration to lower the ambient temperature throughout the space, cooling the IT equipment as well as the room itself. They're not ideal from an energy efficiency standpoint, but they are water-friendly. Evaporative cooling systems have the opposite effect — impressive energy efficiency, but high water use. Neither is sufficient for the excessive heat produced by modern AI clusters.

Thermal management in AI data centers relies on liquid cooling solutions that circulate fluid (water or dielectric fluid) directly to the chip, cooling critical components within the server via closed-loop systems. This is more effective at the chip level, and it's also a boon for the rest of the space — ambient temperature in the data hall can be kept at a higher setpoint, reducing the overall amount of water required to cool the entire environment.



But here, too, there are tradeoffs. Even though this fluid is generally reused, the system still needs a way to get rid of the heat it absorbs — **this is called heat rejection, and it is the key design choice**. If heat is rejected using cooling towers, the system depends on water (through evaporation), which can drive significant facility water usage. Alternatively, heat can be rejected using refrigerant or air-based systems, which reduce or eliminate water use but typically come with higher energy consumption.

Bottom line: all liquid cooling systems reuse fluid, but the way heat is removed from the system determines whether more water or more energy is used.

From a practical standpoint, **higher rack densities will drive adoption of liquid cooling solutions** as thermal thresholds exceed what air-cooled systems can handle. Immersion cooling, in which servers are fully submerged in a bath of dielectric fluid, is largely in the investigative stage with hyperscalers.

As data center operators weigh their options, WUE serves as a touchstone for measuring improvements in water use against energy consumption.

WATER USAGE EFFECTIVENESS (WUE)

Reframing water conservation as a core design standard

FLEX
PERSPECTIVE

SCALABLE, WATER-WISE COOLING FOR THE AI ERA

With modern processors and AI accelerators generating more heat than ever, cooling systems are evolving. Liquid cooling is a case in point. Its superior heat absorption and transfer capabilities are paving the way for AI-era hardware — chips, network switches, etc. — that demands more than air-cooled systems can provide.

Advanced systems enable warm-water cooling, reducing dependence on mechanical chillers and enabling year-round "free cooling" in most environments. This is increasingly important as high-density compute and the quest for more tokens per watt — and more tokens per dollar — ups the ante for effective thermal management.

Broadening the scope of innovation to accommodate facility upgrades and purpose-built AI factories is essential, especially given that most data centers were built years ago and still use traditional air-cooling systems.

JetCool, a Flex company, collaborates with chipmakers and hyperscalers to develop innovative cooling methods that are increasingly energy efficient, scalable, and cost-effective.

For example:

- Microconvective cooling technology uses arrays of small fluid jets that precisely target hot spots on silicon devices such as ACISs, GPUs, and CPUs, transforming high-power electronic cooling performance at the chip or device level.
- High-performance liquid-to-liquid modular CDUs can cool larger in-rack densities to effectively manage heat in power-intensive environments; for instance, a 1MW CDU can cool a 1x MW IT rack or 2x 500 kW racks.
- Direct-to-chip liquid cooling cold plates are engineered to cool modern processors as well as accelerator chips such as GPUs, TPUs and NPUs exceeding 3,000W TDP and 500 W/cm^2 thermal loads.

As data centers proliferate and rack densities intensify, **water conservation should drive infrastructure and equipment design decisions** for operators and suppliers. Engineering scalable cooling solutions with WUE front and center is beneficial for all involved.

(ERE)

Energy reuse effectiveness

With AI chipsets and high-density racks requiring more power than ever, data center operators are seeking new ways to **capture and reuse excess energy** in their own facilities and surrounding communities. That's where ERE comes in.

"Waste energy" typically leaves the facility as warm water or warm air, both of which are byproducts of systems used to cool IT equipment. Absent a focus on reuse, heat dissipation can take a high toll on local electrical and water resources. With the right infrastructure in place, however, it can be distributed to other buildings on the data center campus through pipes that deliver hot water, steam, and chilled water. It can also be exported farther afield through thermal energy networks that benefit surrounding communities. **The key is intentional, coordinated design.**

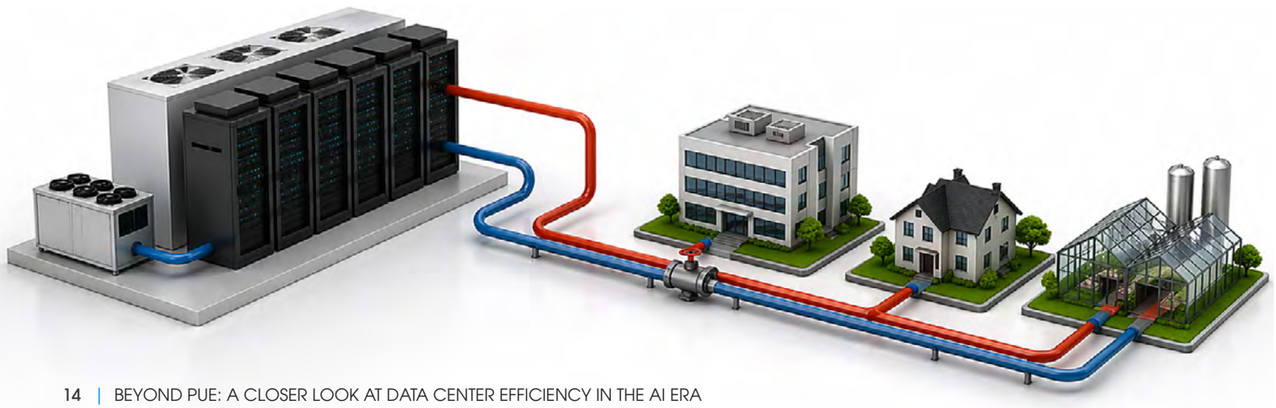
When waste heat is reused within the data center, it is part of the PUE calculation — but PUE does not account for waste heat reused elsewhere. Outside the data center, it becomes a measure of ERE. An ERE of 0.0 means that 100 percent of the energy brought into the data center is reused somewhere else.

$$ERE = \frac{\text{Total facility energy} - \text{energy reused}}{\text{IT equipment energy}}$$

For a detailed technical explanation of ERE, read White Paper #29 by The Green Grid.

Data center infrastructure can be designed for efficiency (a low PUE) with or without reusing energy — and it can reuse its own waste energy without impacting PUE whatsoever. PUE and ERE are complementary metrics that give data center operators a more robust picture of energy use within and outside of their facility.

Data centers consume ~415 TWh of energy annually.[17] As much as 90 percent can be recovered as heat, which could provide thousands of megawatts of thermal power globally — but only a small fraction is ever captured and reused.[18]



Olympic Aquatic Centre

Waste heat from a colocation facility was used to warm swimming pools at the 2024 Paris Olympics — along with 1,000+ homes and a greenhouse on the data center's roof. [13]

Trout farm

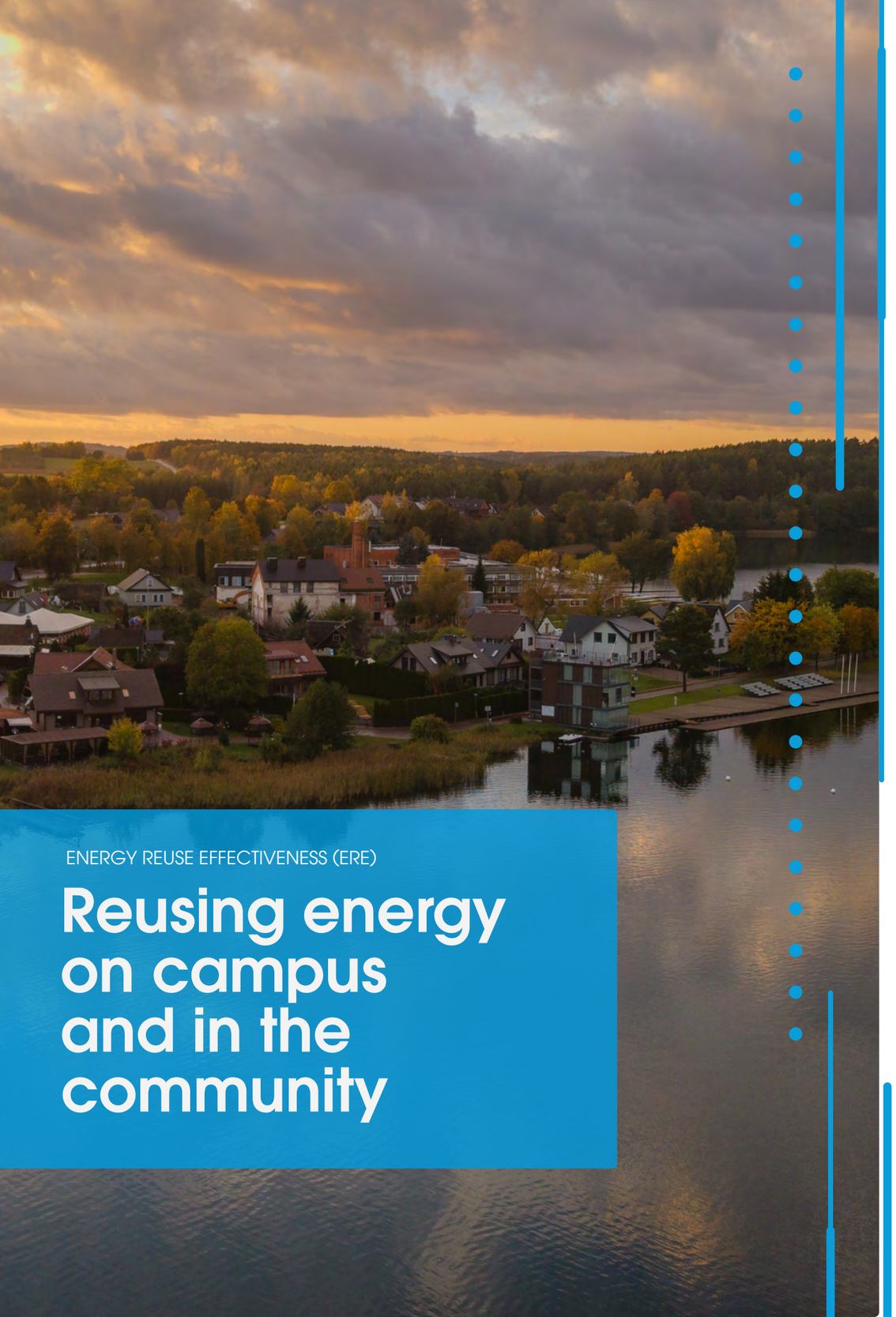
In Norway, excess heat from data center operations is supporting the world's largest land-based aquaculture facility for trout, with plans to scale the system to 8 MW. [14]

Wood pellets

A Swedish data center exports surplus energy to a local thermal power plant for the production of wood pellets, reducing CO_2 emissions in and around the city. [15]

Community heating

Heat from a data center in Dublin warms 500,000+ sq. ft. of public, commercial, and residential buildings via a district heating system, reducing carbon emissions in the area by more than 1,500 tons per year. [16]



ENERGY REUSE EFFECTIVENESS (ERE)

Reusing energy on campus and in the community

FLEX
PERSPECTIVE

WILL AI-ERA PHYSICS MAKE HEAT REUSE MORE VIABLE?

When it comes to ERE, "heat quality" is a central consideration. High-quality heat, typically 122°-140° F, can be used with little or no additional input from heat pumps. With AI factories poised to generate extraordinary amounts of it, the energy reuse conversation has met its moment. **What do we need to do collectively as an industry to put what seems like the most sensible of ideas into action?**

The barriers are well understood:

- Air-cooled data centers reject heat at low temperatures compared with the requirements of district heating systems and industrial processes, which means heat pumps that add cost and complexity are needed to make it usable.
- Data centers are often located in industrial zones or rural campuses where land is cheap and grid interconnects are plentiful, but demand for waste heat is nonexistent. Expensive piping would be required to transport it to urban areas.
- Thermal storage, heat sinks, or cooling systems are still required to deal with excess heat during warm months, because data centers produce it all day, every day.
- Traditional data centers can rely on relatively inexpensive cooling towers that are well understood, simple to operate, and do not require entering into long-term agreements with utilities or municipalities.

The AI shift may hasten a transition. Not only do these data centers generate high-grade waste heat at industrial scale, they often use liquid cooling, which produces hotter return water temperatures. Levers that could influence adoption include:

- Creating economically viable, win-win scenarios between data centers, municipalities, and utilities that lower energy costs and reduce stress on the grid
- Building a district heating infrastructure; capital invested today can produce long-term advantages for waste heat producers and consumers alike
- Instituting carbon pricing, incentivizing product development, and implementing sustainability policies that make the economics feasible and the optics undeniable
- For new builds, designing in heat reuse systems early with the intent of using them for district heating, greenhouses, industrial processes, domestic hot water, or other uses

Companies in the data center sector have long collaborated on products that are increasingly heat-efficient. For instance, advanced electronic designs such as **high-efficiency AC/DC and DC/DC converter topologies** designed in consultation with hyperscalers maximize energy conversion and minimize heat production. We have also seen companies come together to define common standards across the data center infrastructure. Doing the same for energy reuse could be powerful. Putting ERE on par with PUE is a good place to start.

(CPE)

Compute power efficiency

Energy use has a direct impact on data center profitability. With AI workloads drawing power at an unprecedented rate, **CPE has moved beyond the technical and into the fiscal arena.** CPE measures how much useful work gets done for each unit of energy consumed, focusing on server utilization and hardware efficiency. When compute capacity outpaces power efficiency gains, expenses can skyrocket.

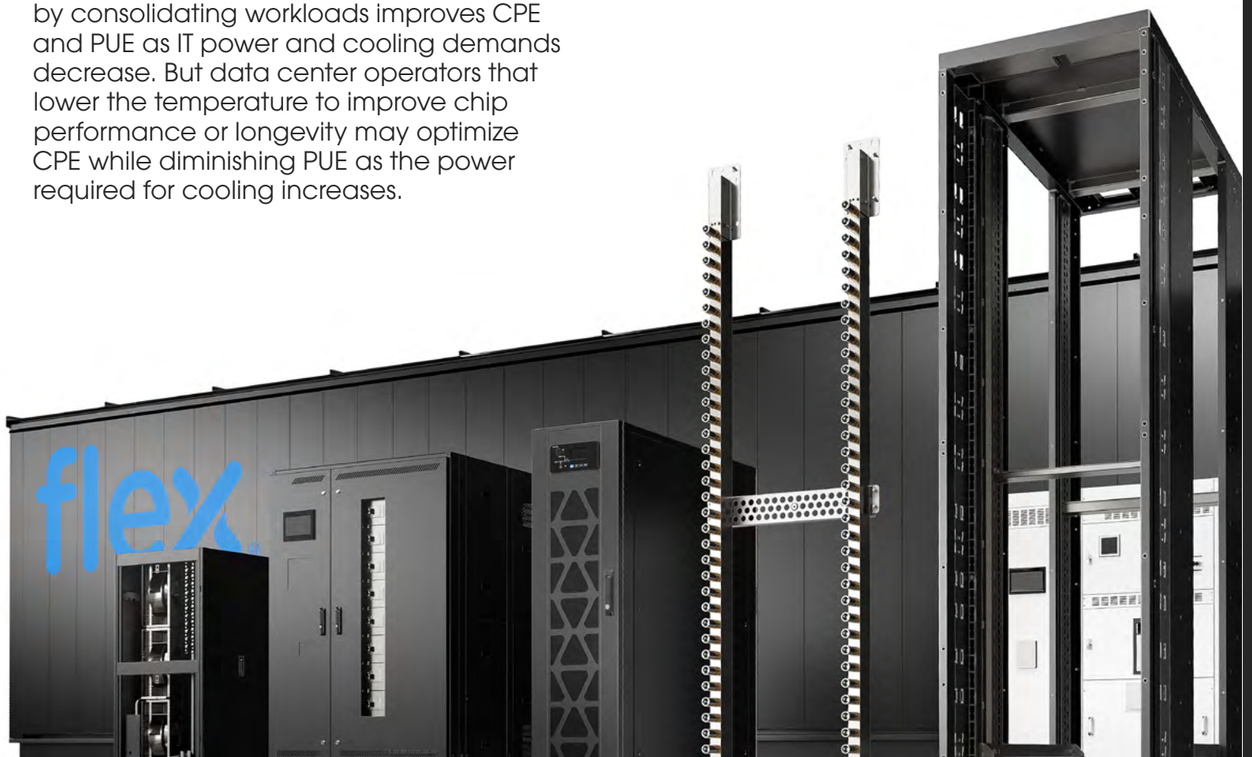
The quest for greater CPE rests on more efficient chips and data center architectures that maximize hardware utilization without negatively affecting computing output, availability, or reliability. **CPE and PUE are related, but they measure different layers of efficiency** (IT and facility, respectively). Whether they complement or conflict with each other depends on how improvements are implemented.

For instance, improving server utilization by consolidating workloads improves CPE and PUE as IT power and cooling demands decrease. But data center operators that lower the temperature to improve chip performance or longevity may optimize CPE while diminishing PUE as the power required for cooling increases.

There is also a circularity to it: as CPE improves, workloads scale, IT power consumption increases, and infrastructure expands, until the cycle begins anew.

The best overall efficiency comes from high CPE plus low PUE. Because PUE does not measure how much useful compute is produced, GPU utilization, or performance per watt, CPE can help data center operators refine their approach to total data center efficiency.

$$\text{CPE} = \frac{\text{TRADITIONAL: } \frac{\text{\# of compute operations}}{\text{Watt}}}{\text{AI: } \frac{\text{\# of tokens}}{\text{Watt}}}$$

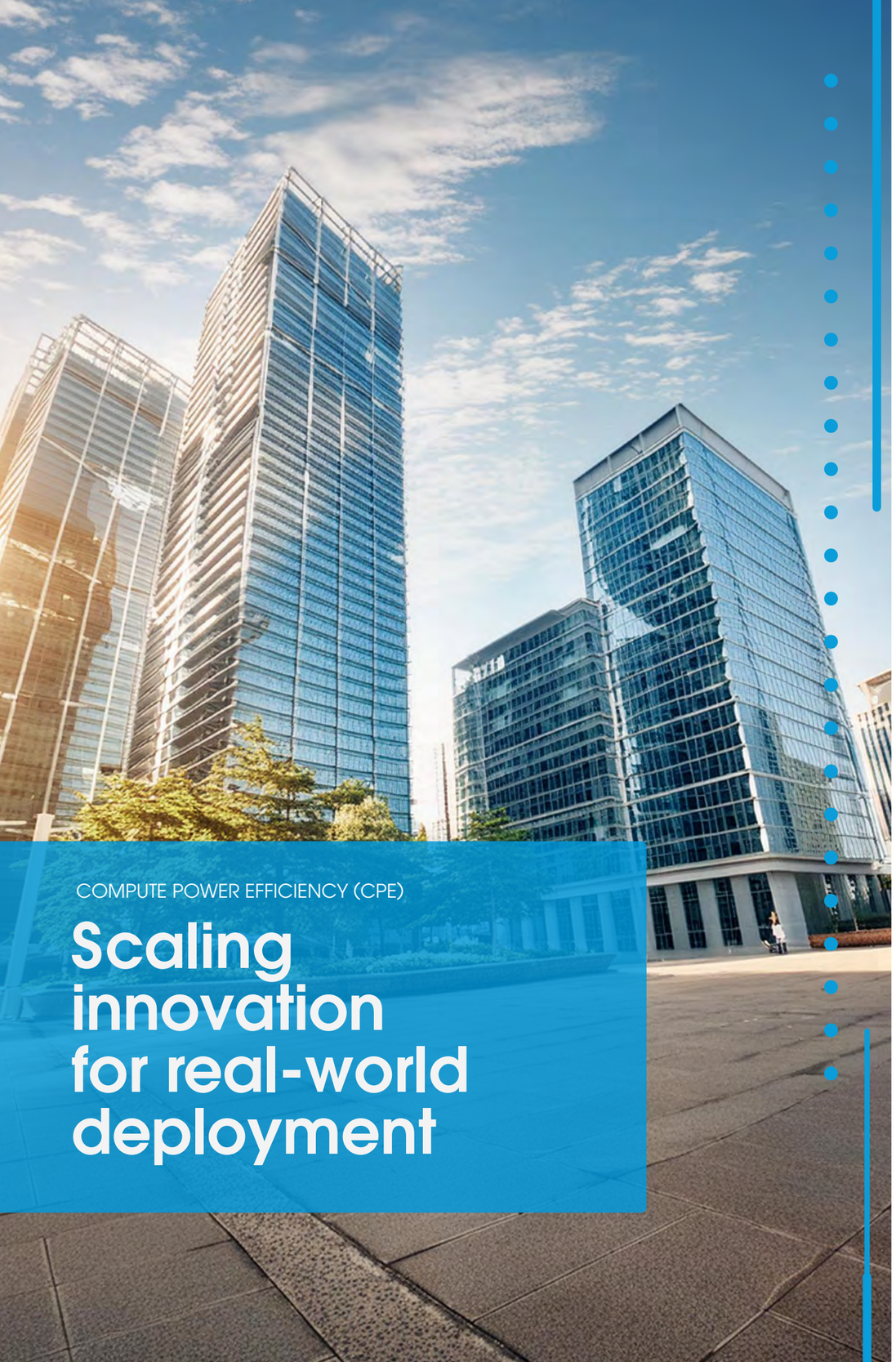


Energy efficiency in AI accelerators roughly doubles every two years, with TPUs sometimes improving faster than GPUs due to specialization.[19]

What is a token?

A token is the basic unit of data derived from breaking down larger blocks of information. The more tokens processed per watt, the more energy-efficient AI becomes.

Hardware, model, and system setup can affect tokens per watt, thus it is a useful metric for comparison and optimizing infrastructure. Hyperscalers also use tokens per watt to help determine customer cost structure and pricing competitiveness. Higher tokens per watt lowers costs per token.



COMPUTE POWER EFFICIENCY (CPE)

Scaling innovation for real-world deployment

FLEX
PERSPECTIVE

ACCELERATING TIME TO MARKET WITH FAST, SCALABLE DEPLOYMENT

CPE in the data center is fundamentally driven by silicon innovation. Chipmakers drive progress, packing more functionality into each generation of processors and dramatically increasing compute density at the rack. Greater energy consumption and thermal loads subsequently compel hyperscalers to redesign power distribution, cooling systems, and network topologies to accommodate advances in hardware and sustain efficiency at scale.

CPE improves as more compute capabilities are integrated directly onto the silicon and as data moves more efficiently between chips. For instance, **bringing optics closer to the processor** reduces repeated optical-to-electrical conversions, lowering latency and power consumption. Collapsing traditional spine-leaf network layers into flatter architectures reduces switching overhead and cuts energy losses. Liquid cooling dissipates higher heat loads directly at the chip, improving thermal performance while reducing overall energy use.

While Flex does not engineer the chips, we enable their rapid and scalable deployment by **co-developing production-ready solutions in sync with chipmaker and hyperscaler roadmaps.** Our solutions are grounded in deep knowledge of:

- Data center architectures
- Systems engineering
- Power distribution
- Liquid cooling
- Networking and optical integration
- Advanced manufacturing

Through prequalification of emerging technologies, accelerated new product introduction (NPI), and system-level integration capabilities, Flex helps enable advances in silicon and architecture to be deployed in real-world data center environments.

(CUE)

Carbon usage effectiveness

There are many direct and indirect sources of carbon dioxide (CO_2) emissions related to data center operations, from the electricity that powers IT hardware, cooling systems, and water sourcing — power that is largely dependent on electrical grids that rely on fossil fuels — to the manufacturing and shipping of data center equipment and construction materials. On-site backup generators also release CO_2 into the atmosphere.

With demand for compute capacity increasing exponentially, renewable energy sources and innovation in cooling technology and power infrastructures have taken on greater urgency. Intense competition among hyperscalers can shift the balance toward the economically expedient rather than the environmentally valiant, but these priorities need not be mutually exclusive. As data center operators vie for dominance and accelerate capacity buildout, there are many well-established levers at their disposal to help them do so in an ecologically responsible manner.

Among them:

Purchasing energy from renewable sources to **reduce carbon emissions**

Using direct-to-chip cooling technologies to **lower the consumption of fresh water**

Making infrastructure improvements that **reduce energy loss** as voltages are stepped down from the grid to the chip

Choosing low-carbon production methods for building materials such as concrete, steel, and aluminum

Building AI data centers in colder regions with abundant renewable energy resources to **reduce reliance on carbon-intensive cooling systems**

Engaging suppliers and manufacturers within proximity to data center locations to **limit emissions associated with shipping**

Designing **advanced power and cooling infrastructure**

$$\text{CUE} = \frac{\textbf{Total } CO_2 \textbf{ emissions} \text{ (kg)}}{\textbf{Total IT equipment energy} \text{ (kWh)}}$$

For a detailed technical explanation of CUE, read White Paper #32 by The Green Grid.

PUE captures how efficiently a data center uses power. CUE provides insight into how much carbon is used in doing so. In essence, they measure distinct aspects of sustainability. A lower CUE means that a data center is more carbon efficient, with 0.0 indicating that it relies solely on carbon-free energy sources. Calculators for tracking and reporting emissions abound. CUE is another tool in the toolbox, **enabling data center operators to quantify their facility's carbon footprint and make choices accordingly.**

CARBON USAGE EFFECTIVENESS (CUE)

Making smarter material and design decisions

FLEX
PERSPECTIVE

ADVANCING CUE THROUGH DESIGN AND LIFECYCLE STRATEGIES

Aware of the complexity of data center operations — and mindful that energy security risks and the rising costs of decarbonization mechanisms (such as RECs, GECs, and CCUs) are becoming increasingly important to operators — Flex collaborates with customers on design for sustainability (DfS) and lifecycle assessment (LCA) solutions that **reduce embodied and operational emissions at the source,** rather than relying on external compensatory measures.

Together, DfS and LCA help customers make smarter material and design decisions to reduce embodied carbon and improve the efficiency of power and cooling systems by evaluating impacts across the entire product lifecycle, from manufacturing through end of life. Applied to data center infrastructure, these methods decrease total carbon footprint and support measurable improvements in CUE.

Design for sustainability

DfS can integrate environmental considerations into system and component design from the outset. For data center power and cooling solutions, this includes designing for:

- Reduced operational energy consumption
- Lower embodied carbon through optimized material selection
- "Circularity by design" that enables reuse, repair, refurbishment, and end-of-life recovery
- Extended product lifespan and less waste
- Lower water and resource use throughout the product lifecycle

By improving materials, architecture, and manufacturing processes early in the design cycle, DfS directly decreases operational emissions and upstream carbon intensity, contributing to lower CUE.

Lifecycle assessment solutions

LCA quantifies environmental impact across all stages of the product lifecycle. When applied to data center power and cooling systems, LCA can help operators and original equipment manufacturers (OEMs):

- Identify carbon hotspots across the value chain
- Compare technology choices (e.g., air vs. liquid cooling) using total lifecycle impact
- Justify investments in higher-efficiency solutions with evidence-based modeling
- Quantify improvements in CUE, WUE, and other sustainability metrics
- Support regulatory reporting and compliance (CSRD, SBTi, Scope 3 accounting)

Flex also applies circular economy principles across its supply chain — from material recovery to component reuse — extending the impact of DfS and LCA beyond product design and into real operational practice.

When hyperscalers engage Flex on relevant data center projects, we can support customer objectives related to resource efficiency and responsible use of materials.

(GAE)

Grid-aware efficiency

Power quality and availability are intrinsic to compute-intensive, time-sensitive AI workloads that are particularly vulnerable to power anomalies such as rapid voltage changes, frequency deviations, harmonics, outages, and transient (one-off) events. Data centers rely on consistent, uninterrupted power to ensure uptime, protect equipment, and maintain operational efficiency. But when IT operations run 24/7 and depend on energy-intensive cooling systems, **drawing maximum power without insight into grid capacity or energy sources is not operationally economical.**

Utilities design the grid to ensure everyone using it has reliable access to power. They take into account peak energy usage, power draw trends, and grid stress abatement. But the **immense, "spiky" power draw characteristic of AI workloads is difficult for utilities to plan for and balance**, because fast changes in demand put sudden pressure on the local grid. Load spikes may also cause voltage and frequency instability in the grid that can disrupt other users' sensitive devices. In extreme cases, this can trigger protection relays and take transformers or generators offline. While that is rare, the probability increases when grid systems are weak or an AI factory is processing immense workloads.

In practice, this impact is assessed at the point of common coupling (PCC) — the shared grid connection where a data center's load changes may be experienced by other customers using the same network.

As a shared resource, assuring power quality and reliability is the responsibility of both the utility and the data center operator. PUE, while providing insight into usage within the data center, does not take into consideration grid conditions, the timing of electrical consumption, carbon intensity, or load shifting. Fortunately, **data center operators have many levers they can use** to mediate their impact on the grid, from energy-efficient IT equipment and cooling systems to load shedding, shifting, and modulation.

The goal is to become grid-interactive, not merely grid-dependent. GAE marries facility energy consumption data with real-time grid conditions such as carbon intensity, peak load times, the availability of renewable sources, and other performance indicators to help make that possible.

Managing GPU power spikes in AI-era data centers

Solutions to minimize the effect of load steps from AI data centers are already in use. Energy storage systems and uninterrupted power supply (UPS) systems can smooth fast changes when properly sized and controlled. Flex engineers and manufactures products that can improve GAE, including:

Capacitive energy storage system (CESS) – Deployed within racks to modulate power fluctuations through dynamic energy storage and power management technologies

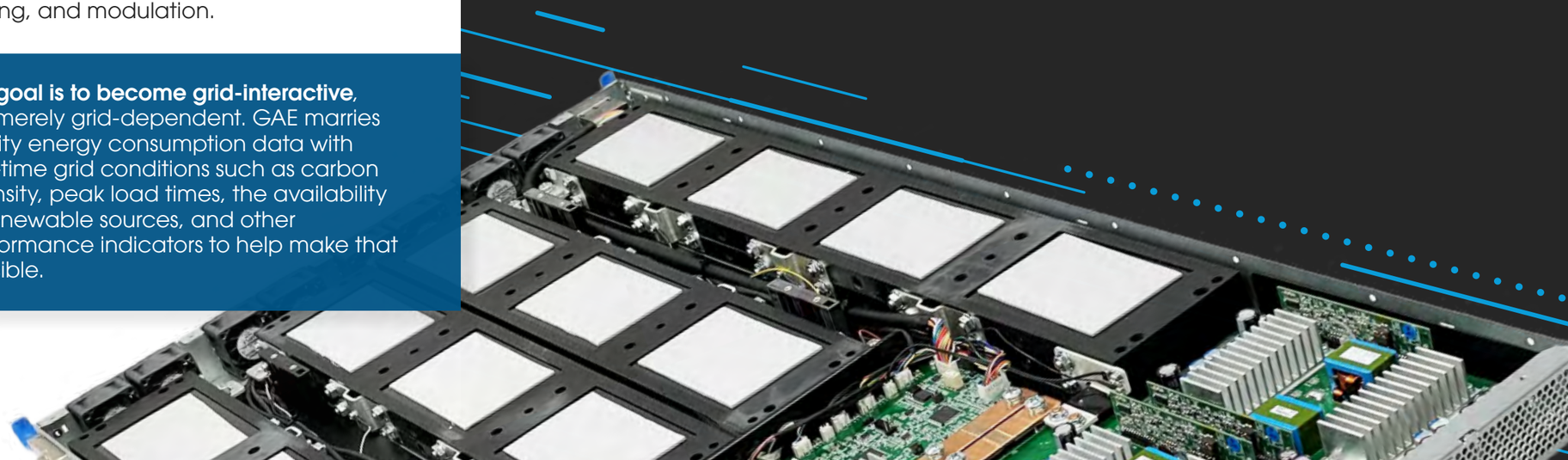
E-houses – Integrate switchgear distribution, protection, and control systems in pre-engineered, isolated environments that enable rapid deployment, fault containment, and stable, sure power delivery

Switchgear – Rapidly isolates faults, absorbs and coordinates spikes, maintains voltage stability, and prevents localized disturbances from damaging equipment or causing upstream grid instability

Metering and monitoring software – Provides real-time visibility into rack- and system-level power consumption to detect rapid load transients, mitigate spikes, and coordinate control with UPS, CESS, and switchgear systems to optimize stability and grid interaction

UPS systems – Stabilize voltage and frequency instantly while providing backup power to prevent power outages and sudden load swings from propagating upstream

Explore critical power products



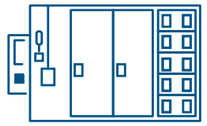
GRID-AWARE EFFICIENCY (GAE)

The catalyst for an adaptive, resilient power grid

FLEX
PERSPECTIVE

COLLABORATING TO REDUCE GRID STRESS AND INCREASE ENERGY RELIABILITY

Rapid growth in AI training and inferencing is reshaping how utilities think about grid capacity and utilization. Large data centers add hundreds of megawatts of demand in a single location, often on compressed timelines. That scale challenges traditional load forecasting, in which growth is assumed to be gradual and geographically diffused. It also introduces enormous, intermittent spikes in energy demand that are unlike those that utility operators have dealt with in the past. In response, utilities are shifting toward scenario-based planning that explicitly models a range of data center build-out trajectories and usage patterns.



This shift is creating an opportunity for constructive collaboration between utilities and data center operators. There are several options worth exploring, including:

- **Improving transparency around load profiles** – Actively monitoring grid signals and using load profile management strategies enable data center operators to fine-tune consumption with **dynamic controls** that reduce costs and support grid stability. By sharing detailed usage data and realistic ramp assumptions, operators can help utilities avoid overbuilding power generation or transmission infrastructure. Conversely, data center operators can mitigate the risk of extended build timelines and power disruptions that impact computing.
- **Mediating grid stress** – Data centers can mediate grid stress by switching to onsite backup power sources during a grid event or when the utility triggers a request to step down power consumption, and by sharing power reserves from **battery energy storage systems (BESS)**. Cooperative agreements such as demand response programs and interruptible tariffs can also bolster grid reliability while reducing required reserve margins, improving grid utilization, and avoiding expensive peak power generation. And if needed, penalties can be imposed when data center power use exceeds contractual agreements.

Bottom line: **AI-driven energy demand does not have to erode grid reliability.** Joint planning of transmission upgrades, integration of renewables, power sharing arrangements, and load interconnection timelines can reduce risk for both parties. With earlier engagement, shared data, flexible rate design, and co-investment in reliance solutions, utilities and data center operators can turn a potential strain into a catalyst for a more adaptive, well-utilized, and resilient power grid.

Become a grid partner, not just a grid user

GAE is an emerging framework

There is no single "GAE score." In practice, grid-aware behavior is judged at the PCC using parameters such as voltage stability, demand flexibility, ramp rate, rapid voltage change/flicker, and harmonics. Exact limits and acceptance criteria are set by the local utility or independent system operator (ISO). Data center operators and utilities may look at:

Grid-stress-weighted efficiency – This is a practical choice if a utility signal, ISO price, or internal grid stress index is known, enabling operators to weight consumption during periods of grid stress. It directly rewards shifting or curtailing energy use during constrained periods.

Rapid voltage change – Changes that occur when the load step from the data center is large relative to local grid strength should stay within ±3% (exact limits are set by the local utility or ISO). Beyond this range, nearby customers may experience flickering lights or minor equipment disturbances.

Coincident peak contribution – Scheduling flexible AI workloads for off-peak hours lowers stress on feeders and substations and reduces the need for reserves, protecting grid reliability and avoiding unnecessary costs.

Demand flexibility – During stressed system conditions, the utility or ISO may request a rapid reduction in power. An AI data center should be able to reduce a defined amount of load within a specified response time and hold that reduction briefly so the net change seen at the PCC is smaller and slower.

Building a holistic framework for data center efficiency

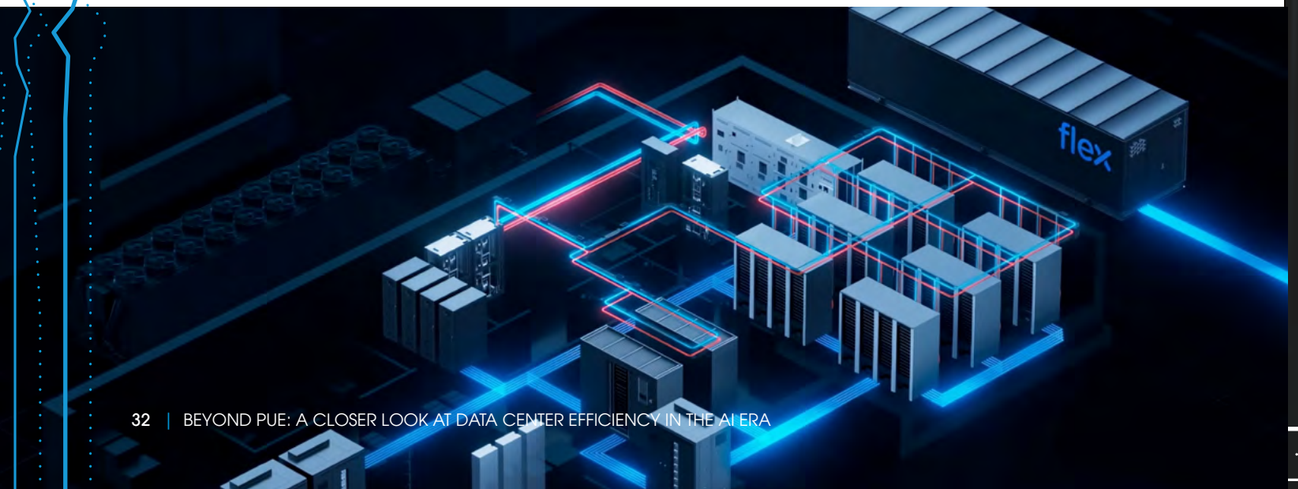
In the AI era, data center efficiency requires a more holistic framework. PUE remains an important operational benchmark, but on its own is too narrow for infrastructure shaped by GPU-intensive workloads, megawatt rack densities, and dynamic power profiles. The pressure to manage carbon, water, and grid impact more responsibly is growing. **A holistic approach means evaluating efficiency across power, compute, water use, carbon emissions, energy reuse, and grid interactions.** This broader view matters, because AI changes both the scale and character of demand.

Capacity and utilization do not always move in lockstep. Facilities may appear to be efficient by one measure while creating tradeoffs elsewhere, such as lower PUE at the expense of water use, or improving compute performance while increasing power and cooling complexity. The operators best positioned to lead will the ones that manage these interdependences deliberately rather than optimizing one metric in isolation.

Efficiency is now a systems-level discipline. Data center operators should:

- Track PUE alongside complementary metrics such as WUE, ERE, CPE, and emerging grid-aware frameworks.
- Design infrastructure around realistic workload behavior rather than theoretical maximums.
- Prioritize cooling and power architectures that support performance and resource efficiency at high density.
- Make decisions with sustainable design and lifecycle impact, from materials and manufacturing through operations and end of life.

The goal is to broaden one's lens beyond capacity and tokens per dollar to encompass intelligent, responsible, and resilient facilities and operations.



Translate AI innovation into efficient infrastructure with Flex

Flex helps hyperscalers translate rapid advances in AI — across silicon, systems, and architectures — into scalable infrastructure solutions that increase efficiency and tokens per dollar. With expertise spanning power, cooling, systems integration, networking, manufacturing, and lifecycle strategy, **Flex enables customers to deploy next-generation architectures while improving operational and environmental performance.**

The company's critical and embedded power solutions are designed to help address energy efficiency, reliability, and dynamic load challenges. Advanced thermal solutions for high-density AI environments are engineered with a focus on water usage considerations. In addition, Flex collaborates with data center operators through co-development, prequalification, and rapid NPI to facilitate the deployment and scaling of emerging chip and system technologies.

Beyond deployment, Flex works with customers to reduce operational and environmental impacts at the source. Through DfS and LCA services, we help hyperscalers make smarter choices about design materials, architectures, power, and cooling systems to reduce embodied and operational emissions, improve CUE and WUE outcomes, and support circularity across the product lifecycle.

Tackle AI-era data center efficiency with Flex.

Contact us

Resources

1. McKinsey & Company, The state of AI: How organizations are rewiring to capture value, March 12, 2025
2. Uptime Institute Globa Data Center Survey 2025, July 2025
3. IEA, Energy demand from AI, Accessed April 2026
4. Pew Research Center, What we know about energy use at U.S. data centers amid the AI boom, Rebecca Lepppert, October 24, 2025
5. Uptime Institute, Large data centers are mostly more efficient, analysis confirms, Jacqueline Davis, February 7, 2024
6. Uptime Institute, Large data centers are mostly more efficient, analysis confirms, Jacqueline Davis, February 7, 2024
7. The New Stack, Cloud PUE: Comparing AWS, Azure and GCP Global Regions, Adrian Cockcroft, January 10th, 2025
8. Uptime Institute, Large data centers are mostly more efficient, analysis confirms, Jacqueline Davis, February 7, 2024
9. Goldman Sachs, AI to drive 165% increase in data center power demand by 2030, Feb 4, 2025
10. Uptime Institute, Large data centers are mostly more efficient, analysis confirms, Jacqueline Davis, February 7, 2024
11. Environmental and Energy Study Institute, Data Centers and Water Consumption, Miguel Yanez-Barnuevo, June 25, 2025
12. Data Center Map, USA Data Centers, Accessed April 2026
13. Equinix, The internet is sustainably heating Paris and the Olympic training pool, accessed April 2026
14. Green Mountain, Turning waste heat into value, 4 February 2026
15. EcoDataCenter, Enabling the Green Transition, Accessed April 2026
16. The Irish Times, Excess heat from Dublin data centre to warm local buildings in first for Ireland, Kevin O'Sullivan, April 6, 2023
17. International Energy Agency, Energy and AI, 10 April 2025
18. Uptime Institute, Heat reuse: a management primer, Max Smolaks, October 2023
19. Epoch AI, Leading ML hardware becomes 40% more energy-efficient each year, Robi Rahman, Oct. 23, 2024
20. Nature Sustainability, Environmental impact and net-zero pathways for sustainable artificial intelligence servers in the USA, T. Xiao, F. Nerini, et. al., 10 November 2025

For more information, visit **flex.com/connect**