ENERGY INNOVATION FELLOWS DATA CENTER SERIES PART 1 | SEPTEMBER 2026

Data Centers 101: Types, Loads, and Tradeoffs

Ian M. Hoffman, co-lead, DOE/LBNL Center of Expertise for Data Center Energy

Center of Expertise for Data Center Energy | Energy Technologies Area | Lawrence Berkeley National Laboratory

This work was funded by the U.S. Department of Energy's Office of Critical Minerals and Energy Innovation, under Contract No. DE-AC02-05CH11231.

BERKELEY LAB

CENTER OF EXPERTISE FOR DATA CENTER ENERGY

What we'll cover

1

Why Are Data Centers Proliferating?

What is driving the buildout? Why now? When's the bubble coming?

2

Data Center Types and Missions

What are the types and workloads of the more common data centers?

3

Data Center Energy and Water Use and Load Characteristics

How does data center power demand behave over time? How are energy and water use related?

4

Impacts: Pros and Cons

What are the advantages and disadvantages of the buildout? What other policy issues does the buildout surface (an incomplete list)? How is industry responding?

Why data centers? Why now? How are we responding?

Drivers

Cloud Computing and Web 2.0: the Third Industrial Revolution

- Digital economy – eCommerce runs on data centers
- Ubiquitous computing – your pocket, laptop bag and EV hold processing power that scientists 20 years ago could only dream of.
- Calls, texts, DM – every cell tower has a small data center

Big Data/Machine Learning/Artificial Intelligence: the Fourth Industrial Revolution

- LLMs growing in size & capability - more GPUs, larger neural networks, longer training runs.
- Inflection point for inference already here – agentic AI
- AGI and Superintelligence drawing closer – The "singularity" is close if not already here.

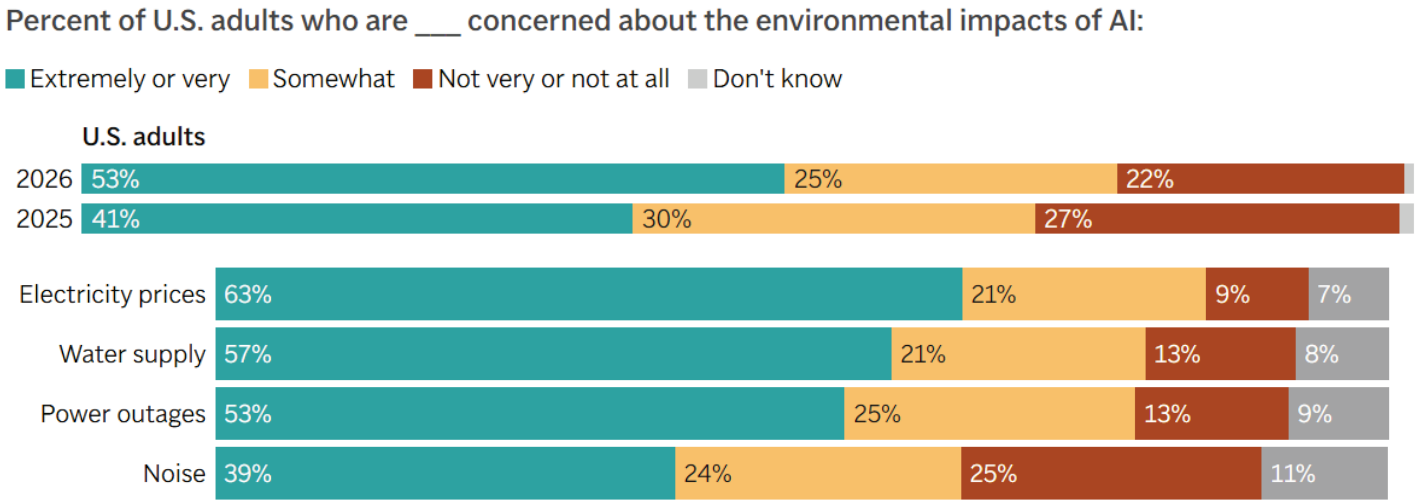
Bubble?

- Probably. But the industry does not expect a long-term correction (see, e.g., railroads, fiber networks). Data centers always will be needed.

Do you support or oppose the construction of an artificial intelligence data center in your local community?

Environmental Harm and Depletion of Natural Resources	28%	Oppose
Higher Utility Bills and Economic Burden on Residents	10%	Oppose
Local Disruption, Land Use, and Poor Community Fit	15%	Oppose
Rejection of Artificial Intelligence and Distrust of the Industry	13%	Oppose
Economic Benefits, Technological Necessity, and Competitiveness	12%	Support
Unsure, Indifferent, Ambivalent, or Uncodeable	22%	Neither/None

Source: Umass-Amherst poll released Sept. 14, based on polling in August. MoE: +/- 3.5%



Source: The Associated Press-NORC Center for Public Affairs Research and the Energy Policy Institute at the University of Chicago. Released Sept. 16, based on polling in August. MoE: +/- 2.2%

Many types of data centers, many missions

Enterprise / on-premise (~10s of kW-5 MW)

- Server rooms & small facilities, often under 10,000 sf and 1 MW, most often embedded within an office building
- Found wherever meaningful amounts of data are processed e.g., government, academia, hospitals, large offices.
- Generally low utilization (10-25%), low efficiency, diurnal load

Colocation (~5-100 MW)

- Leased facility (wholesale colo, single tenant) or facility with leased space for racks (retail colo, multiple tenants
- Colo provides space, power, cooling under a contract (SLA)
- Moderate variability in demand. Mostly diurnal.

Hyperscale (>100 MW)

- Large loads, dedicated substations, .
- Can be specialized (cloud, AI training, AI inference), but many are mixed use under a single, large tech firm.
- Not present in Maine.

Edge (~10s of kW-2 MW)

- Emerging: high-density telecom, EV support or low-latency inference/controls optimization.
- Critical for VR/AR, surgical telehealth, real-time industrial optimization (Internet of Things 2.0)

Neocloud (15-200+ MW)

- Large loads with dedicated substations or leased colo space
- Very specialized: GPU clusters as a service - quick way to train/fine-tune models and host inference. Can be low cost: ~65-80% of charges by major cloud provider for AI computing
- Lessees include…hyperscalers

Others:

- Crypto mining rigs (~100s of kW to 100 MW)
- High-performance computing (supercomputers, 1-100 MW)
- Telco - hubs, cell towers, satellite ground stations

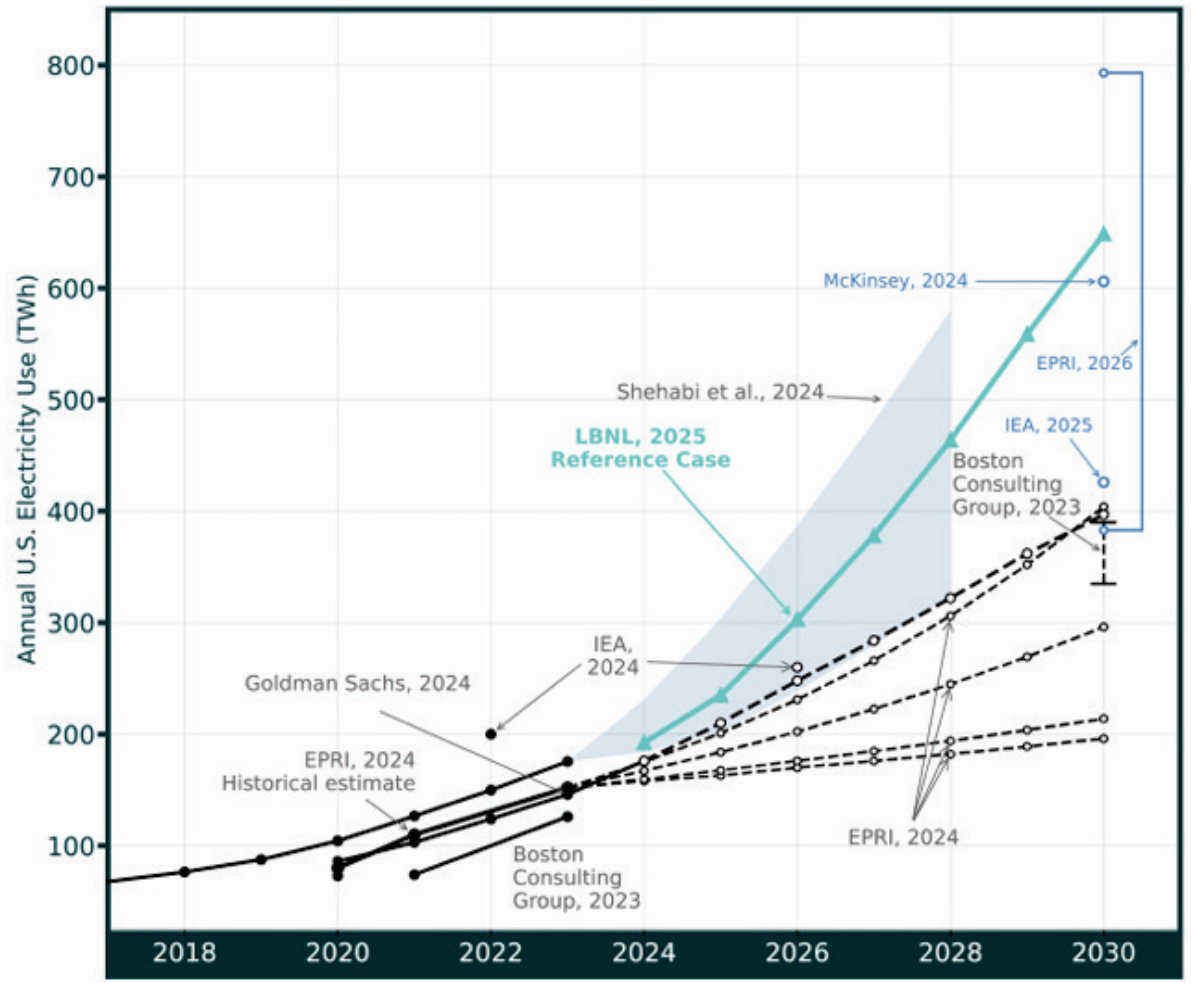
26).

The national picture: more and larger data centers, uncertainty in total magnitude

4.7% of U.S. electricity in 2024 (192 TWh)

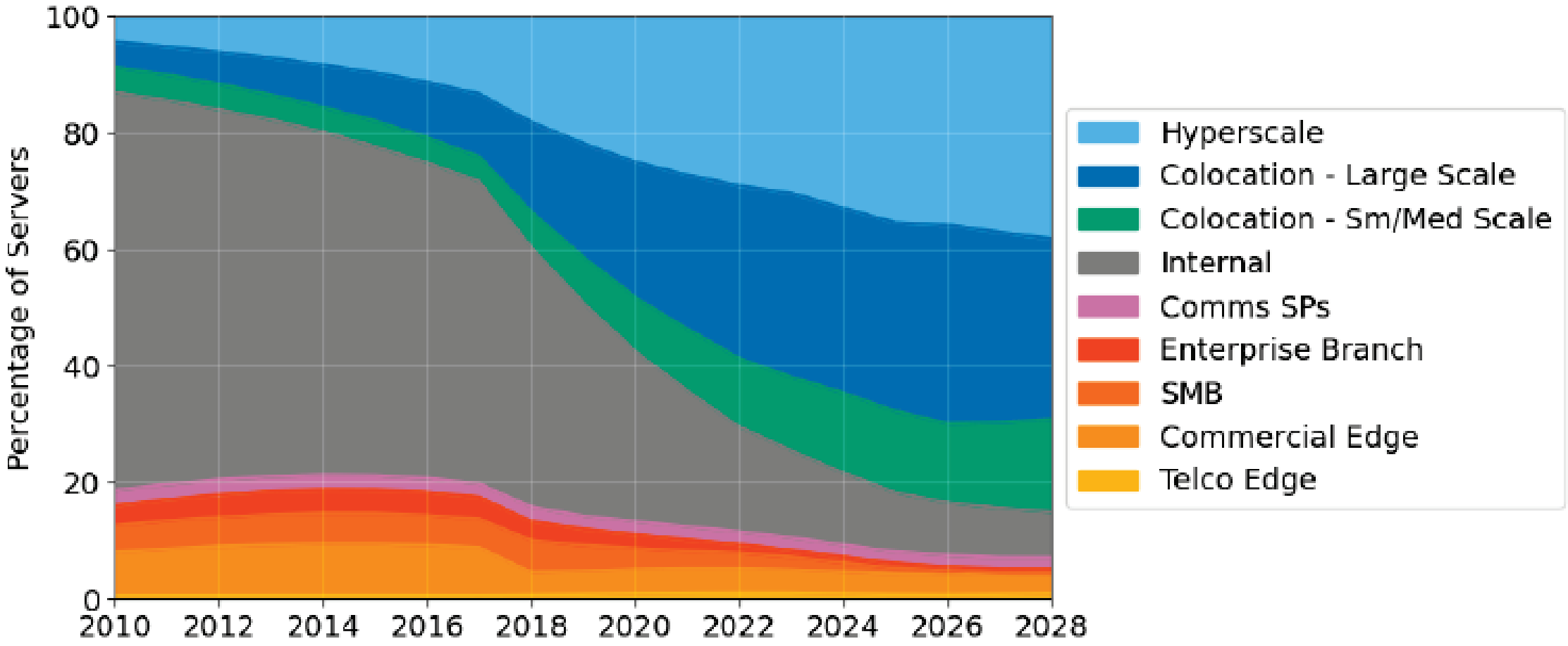
11.8% projected by 2030 (Reference Case, 649 TWh)

~1/3 of U.S. load growth, 2024-2030



Source: LBNL 2025 Update (LBNL-2001758).

Chips and servers drive demand for power, water. AI adds power intensity.



What's in that windowless box?

1

Pods of servers packed with 72 GPUs plus CPUs, Networking

High power density. Today, 2D packaging. Tomorrow, 3D.

2

Power intensifying from busbar to chip

Today, 275 kW/rack. Infra going from 480V AC to 800V DC.By 2030, MW racks.

3

Cooling: Fan Walls and Move From In-Row to D2C

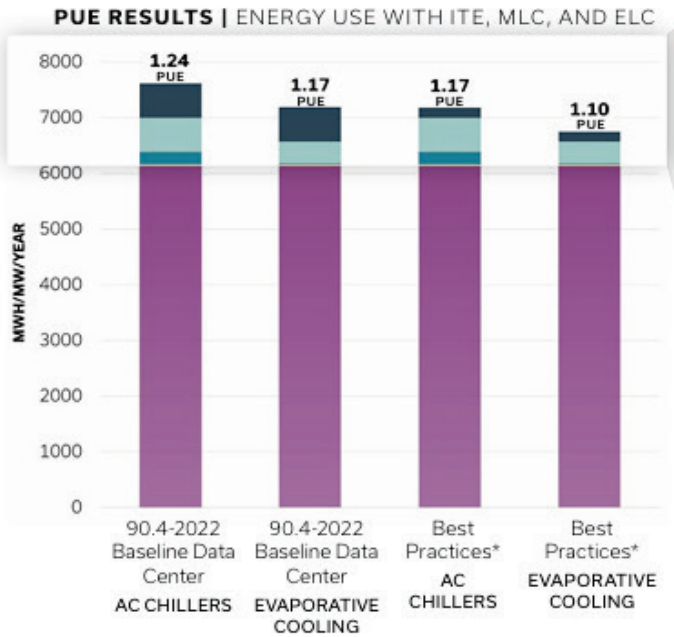
Coolant moving closer to the heat. Next step: direct-die cooling (see 3D packaging.)

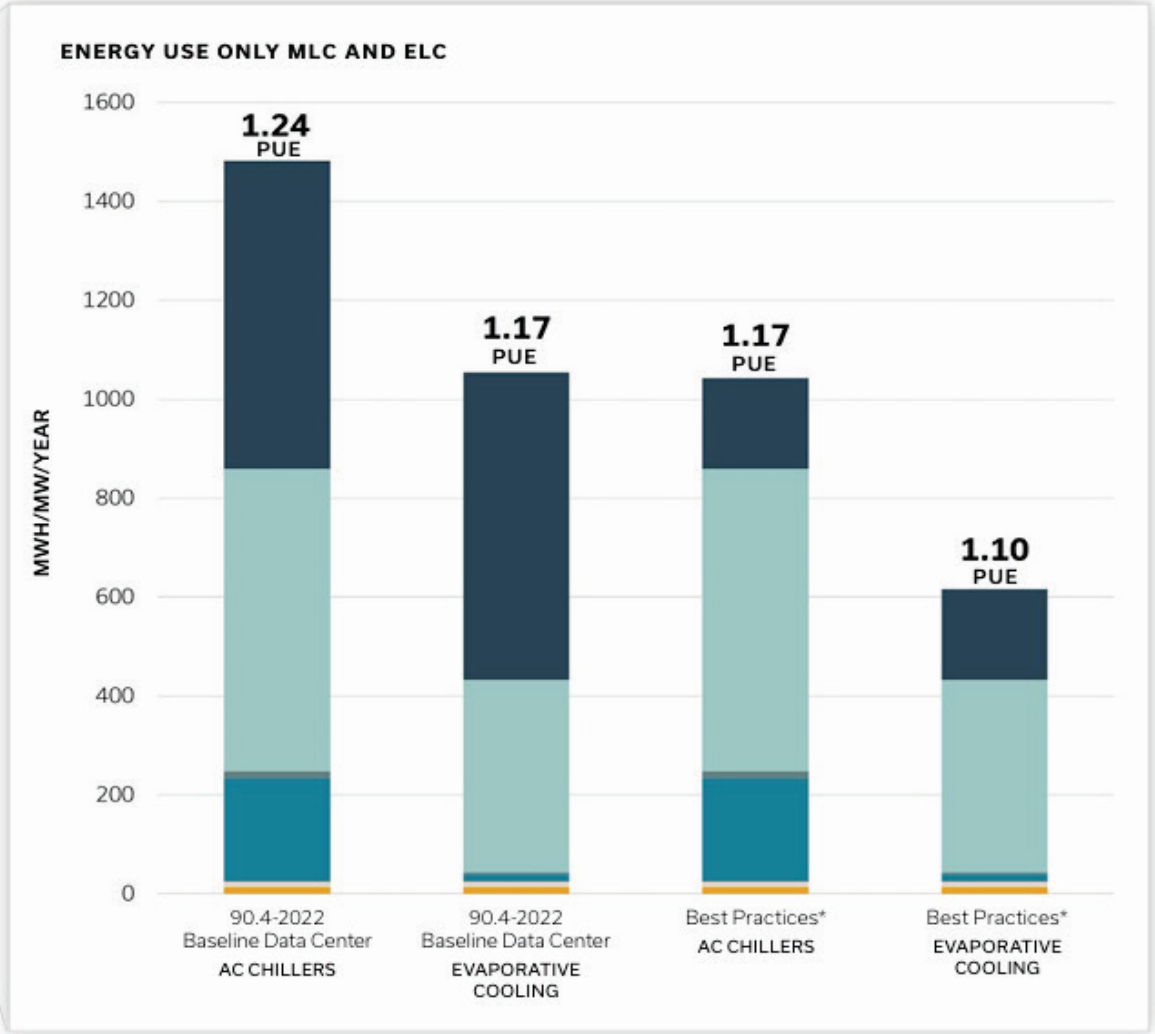
4

Computing virtualized in 2010s. Now digital twinning of the data centers and predictive controls

Most data centers still controlled by sophisticated heuristic controls with some machine learning. Coming fast: full-facility virtual models with high physics resolution and AI-driven collation of workloads, weather, power cost and availability, etc., and virtual validation on a digital twin before implementing new optimization.

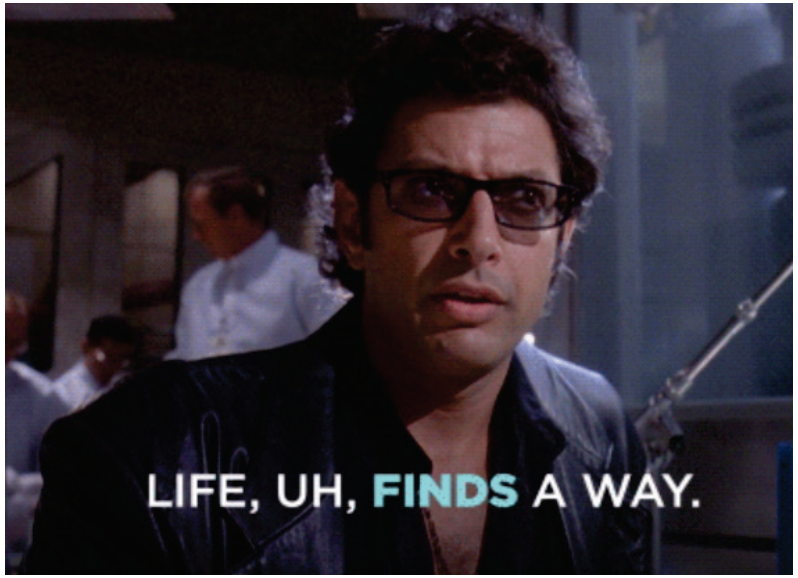
Power Usage Effectiveness = Total Power / IT Power, the breakdown





Chemistry of direct-to-chip coolants for hyperscale AI/Supercomputing

- Deionized water with generally 25% polyethylene glycol, known as PG25. Small energy penalty. vs. water alone.
- Glycol is not for freeze protection but a stable medium for additives:
  - Bio-inhibitors (tolyltriazole, benzotriazole, or both)
  - Corrosion inhibitors (nitrite, molybdate, silicates)
- pH @ 7.0 to 8.5 or aluminum and copper corrosion rates increase nonlinearly.



Higher thermal loads → higher flow volumes and velocities → less corrosion/erosion protection

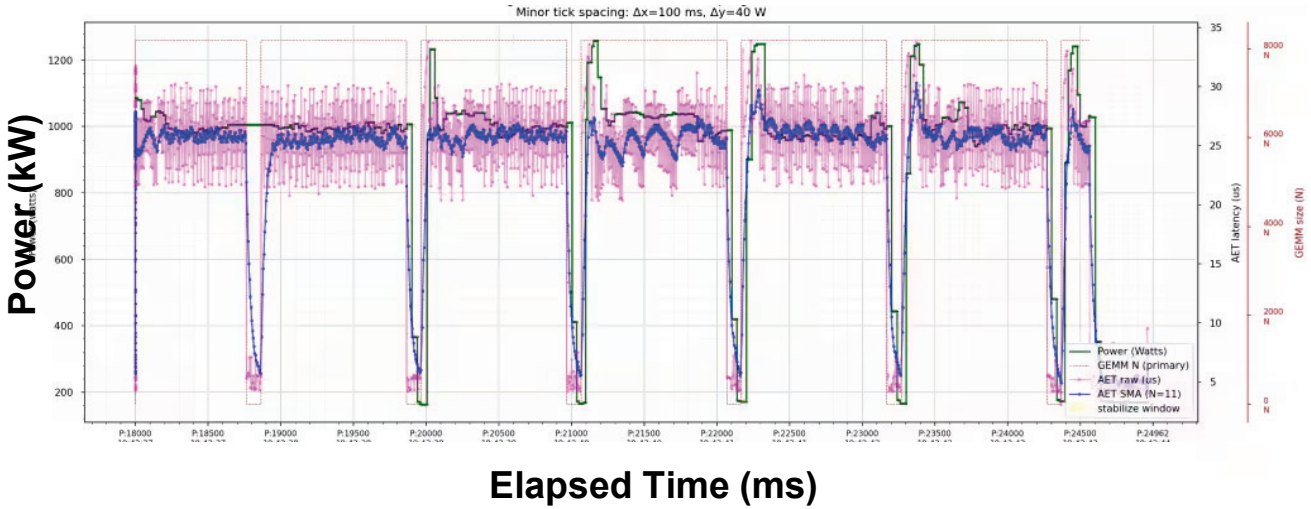
Potential solution: two-phase cold plates that flow refrigerants over chips. Raises “new” issues: PFAS vs. toxicity, flammability, high GWP

Load shapes: two timescales

Type	Sub-second (grid stability)	Diurnal (hours to seasons)
Hyperscale - AI training	Synchronized, rapid oscillations — the outlier	High utilization, high load factor
AI inference / edge	Bursty, not synchronized	Diurnal; business-hour peaks; may be in flux
Hyperscale (non-AI cloud)	Virtualized; managed for steadiness	High load factor
Colocation / mid-size	Relatively virtualized	Moderate; mixing workloads smooths the aggregate
Enterprise / on-premise	Increasingly virtualized	Flat; chronically underutilized (10-25% @ server; often 2-6 kW power/rack)

Sources: NERC (2026); Choukse et al. (2025, preprint); measurement studies (2026); E3 (2025).

AI training loads can pose challenges on- and offsite

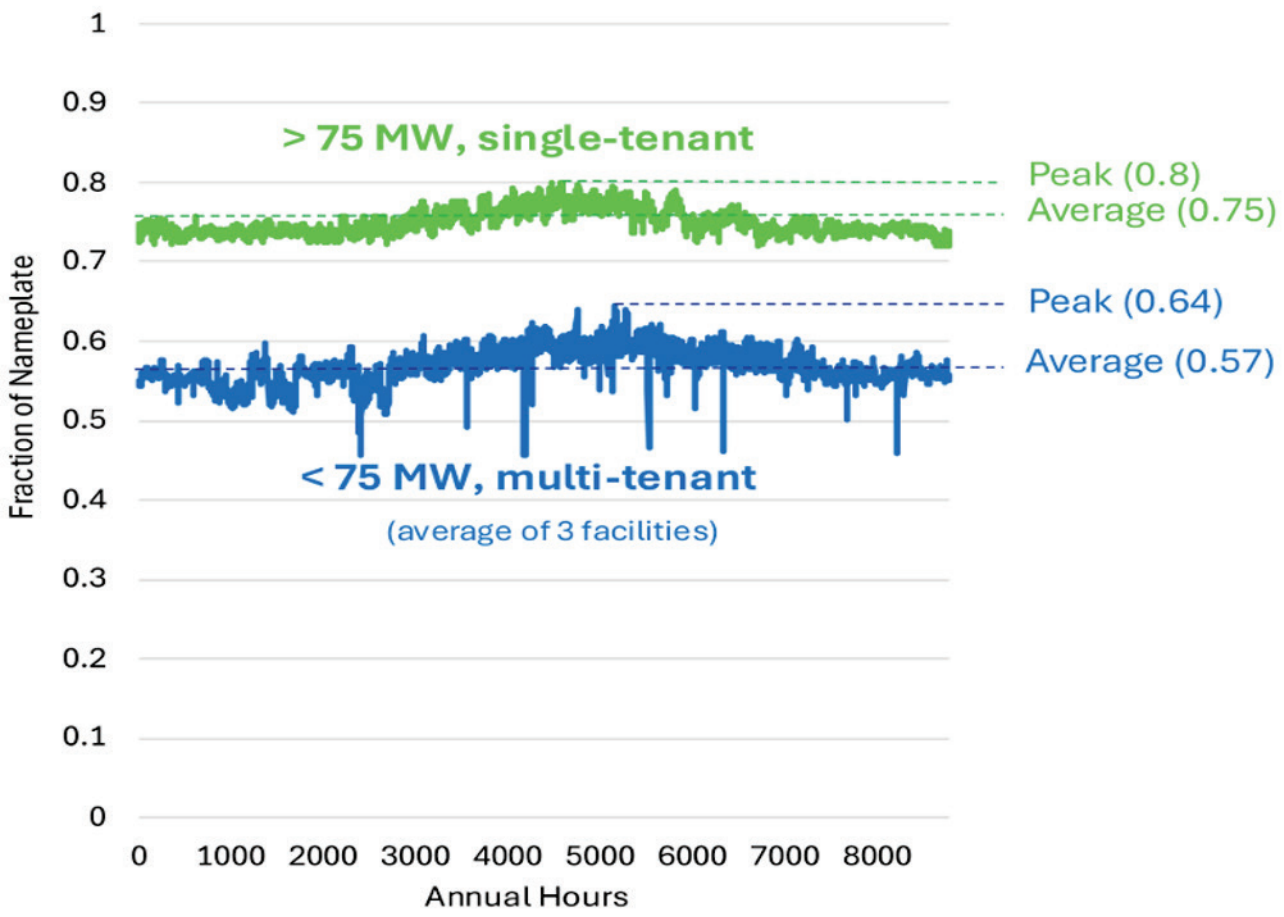


Solution	Reliability	Performance	Energy	Cost	Ability to meet tightest spec	Dependency on the developer	Lifetime
Software-only mitigation	Medium	Medium	High	Medium	High	High	High
GPU power smoothing	High	Medium	High	Low	Medium	Medium	Medium
Rack-level energy storage	High	High	Low	High	High	Low	High

Takeaway: Synchronized AI training clusters produce sub-second power transients with implications for grid-stability. Very large AI inference clusters can have a similar pattern but generally smaller in magnitude and less of a concern. Mid-sized facilities showing interest in Maine are more likely to be mixed-use loads and unlikely to cause these issues.

Source: Oracle Cloud Infrastructure Blog, March 2026, Behind the Scenes: GPU Power Smoothing for Large-Scale AI Training: :

Colo loads vary by size and lessee workload



- Colos are agile and growing fast in more markets than hyperscale facilities
- Loads are bifurcated between small retail facilities and large wholesale facilities.
- Wholesale: typically hyperscalers with high utilization rates
- Retail colos: Mash of tenant loads that vary in magnitude and timing but generally lower in utilization

What drives siting of large data centers

Siting Determinants

- Power availability and cost – Timing, cost and conditions for interconnection
- Potential for onsite generation
- Land – Cost and room for expansion
- Fiber network – Data flows are critical
- Climate and water – Cooler locations means less energy and water use.
- Community Approvals/Permitting - A month of delay can be tens of $M in lost revenue. Rising in importance. Some would rank as #1.
- Proximity to users (major metros, heavy industry, etc.)

Counterweights

- Community acceptance - A month of delay can be tens of $M in lost revenue. Rising in importance. Some would rank #1 among siting criteria.
- Close doesn't mean inside a beltway – With speed of light communications, a data center serving most needs can be hundreds of miles away, barring stringent latency requirements (VR/AR, IoT 2.0).
- Onsite generation means speed to power and freedom from utilities and grid regulation – but also air permits, new community objections (e.g., noise, water use), and critically access to natural gas. Available prime movers (e.g., ICEs) are small and short-lived.
- Water - Dry cooling can make water availability less stringent but comes with added energy cost and offsite water use (more on this tradeoff shortly)

Data center buildout economics

Rate impacts

Where data center load is highly concentrated, increased system demand may raise costs for other ratepayers

Where data centers are not showing up at scale or where operators are bringing their own generators or other capacity (DR), rate pressure is still there but principally is coming from other drivers: grid modernization, new manufacturing, electrification.

Indirect effects of the surge, regardless of siting:

- Significant demand growth could increase (or decrease) costs, depending on when and where it occurs.
- Higher demand for capital, equipment and skilled labor can increase costs and timelines.
  - Transformers: 24-month lead times, longer for certain voltage ranges.
  - Switchgear: 18 to 24-month lead times.
  - Onsite generators: longer lead times. Driving demand for smaller prime movers (reciprocating engines) or quirky ones
  - Experienced electrical engineers, project managers, master electricians, industrial plumbers.
  - Some competition for equipment and talent with the electric power sector, ramping up in part to serve data centers

.

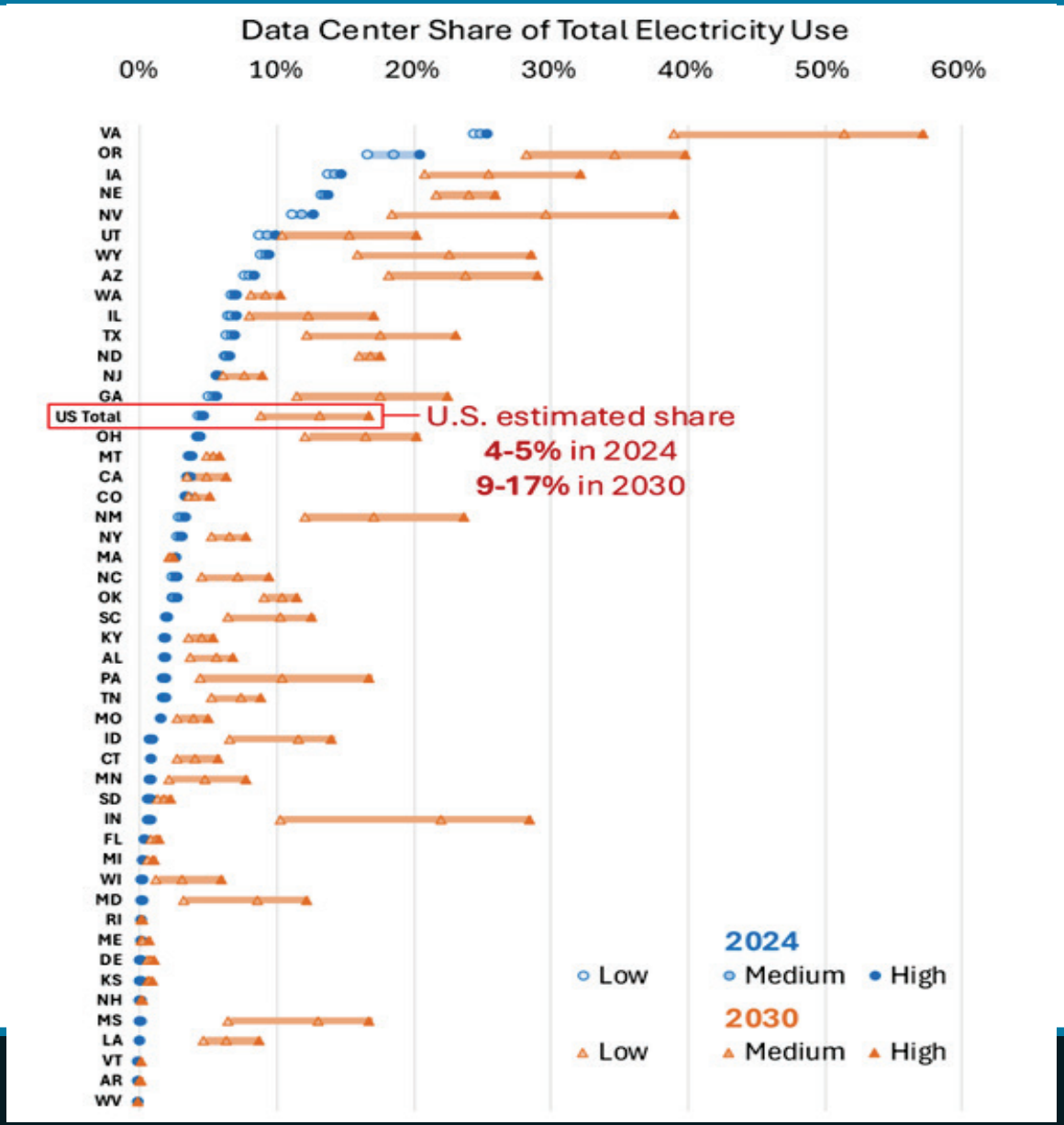
Off-grid and islanding feasibility

- Virtually all data centers of any size have backup generators, mostly diesel, some natural gas-fired, some fuel cells
- Now many data centers are securing their own primary power and other capacity
  - Onsite – often via a microgrid
    - Combined cycle gas plants (CCGTs)
    - Fuel cells
    - Single-cycle turbines, jet engines, reciprocating engines, cruise ship engines
    - Solar PV with battery storage
    - Small modular reactors (online ~2030)
  - Offsite
    - Solar and wind via power purchase agreements
    - CCGTs shared with local utilities
    - Demand response via CSPs
    - Nuclear plants (Three Mile Island)

Where is the growth concentrated? Will it stay that way?

- EPRI projected data center loads as a percent of total state load under low, medium, and high scenarios.
- Top states: VA, TX, OR, NV, UT, WY, NM, IN, MS, LA, IA, NE, AZ, GA, OH, ND, IL
- Strong executive or legislative action: TX, VA, ME
- **Takehome**: There may be a saturation point beyond which communities will not allow more.

[EPRI Home | Powering Intelligence 2026]



What utilities and grid operators can do: large-load policies and tariffs

Speed to Power: Solutions for Accelerating Large Load Connections, June 2026

Solution Area	Solutions
Load Forecasting	
S1. Load forecasting approaches	S1.1 Separate load forecasts for large loads S1.2 Regular customer engagement and market research S1.3 New methods for large load forecasting S1.4 Milestone requirements for inclusion in forecasts S1.5 Processes for monitoring project duplication in forecasts S1.6 Forecast reporting requirements
S2. Load forecast coordination	S2.1 RTO* guidelines on large load forecasting S2.2 State involvement in RTO load forecast adjustments S2.3 Data sharing and exchange on large load forecasting
Interconnection	
S3. Load interconnection processes	S3.1 Clear thresholds for large load interconnection S3.2 Disclosure rules for similar interconnection requests S3.3 Coordination between load interconnection and planning processes S3.4 Coordinated interconnection for load and paired generation S3.5 Clear disclosure rules for behind-the-meter generation and storage S3.6 Standard interconnection process elements** S3.7 Cluster studies S3.8 Interconnection data and queue transparency
S4. Fast-track transmission interconnection service	S4.1 Provisional interconnection service S4.2 Surplus interconnection service S4.3 BYOG*** and pay-for-transmission incentives
S5. ERIS**** for loads (transmission)	S5.1 ERIS-like interconnection service for loads
S6. Flexible interconnection (distribution)	S6.1 Staggered connections S6.2 Static flexible interconnection S6.3 Dynamic flexible interconnection
S7. Large load construction of transmission and distribution facilities	S7.1 Option to build

41 potential solutions: Protecting the grid, other customers & potentially easing rate, tax pressures

- Load forecasting
- Interconnection
- Resource planning and procurement
- Markets and operations
- Cost allocation and ratemaking

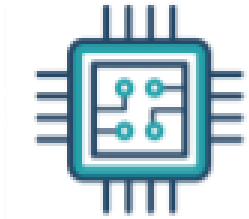
https://emp.lbl.gov/publications/speed-power-solutions-accelerating

Utility Procurement	
S8. Utility procurement approaches	S8.1 Advance procurement of equipment S8.2 Flexible procurement for new resources
Markets and Operations	
S9. Energy markets and transmission service	S9.1 Load participation in real-time markets S9.2 Non-firm transmission service for loads
S10. Operating reserves	S10.1 Changes in regulation reserve procurement S10.2 Downward contingency reserves S10.3 Changes in reserve cost allocation
Cost Allocation and Ratemaking	
S11. Cost allocation methods	S11.1 Review and update of allocation factors S11.2 Direct assignment of costs
S12. Large load tariffs	S12.1 Clear eligibility requirements S12.2 Standard rates S12.3 Risk mitigation measures S12.4 Interruptible service requirements S12.5 Customer choice programs for large loads
S13. Wholesale rates and access for large loads	S13.1 Wholesale rates under utility service S13.2 Direct access programs

Key potential policy advance: Data centers can support the grid, not only stress it

Increasing operational impacts

Computational Load Flexibility



- Job scheduling
- Computational intensity adjustments
- Model/service selection

Core Facility Infrastructure

- Cooling system temperature setpoint adjustments
- Uninterruptible Power Supplies (UPS)

Energy Storage

- Thermal energy storage
- Electrical energy storage (batteries)

On-Site Electricity Generation

- Diesel (low-emissions)
- Natural Gas
- Hydrogen

Increasing planning and investment

Source: Smith framework (LBNL).

Water use is a siting and design variable, not a fixed quantity

Water ranks among the top community concerns nationally, second only to energy. (Quinnipiac poll, March 2026).

Evaporative/open or closed loop

- Lowest energy use
- Highest onsite water use
- Discharge chemistry to manage

Closed-loop / dry

- Near-zero onsite water
- Higher energy
- Can shift water use offsite (Scope 2)

Hybrid air / liquid

- Where AI thermal density has driven the industry
- Latest GPUs: ~100% liquid cooling
- Next: Direct-die microfluidics, two-phase cold plate?

Efficiency and the energy-water paradox: Evaporative cooling systems use less energy but more water. For new builds, hyperscalers are moving heavily toward closed-loop systems with dry coolers and air-cooled chillers that use virtually no water but more energy. It's a choice that shifts water use from cooling the data center to cooling/evaporative losses in power generation, from inside the fence line to the power plant.

Sources: PAE/CTUIR (2025); Microsoft (Dec. 2024; Build 2026).

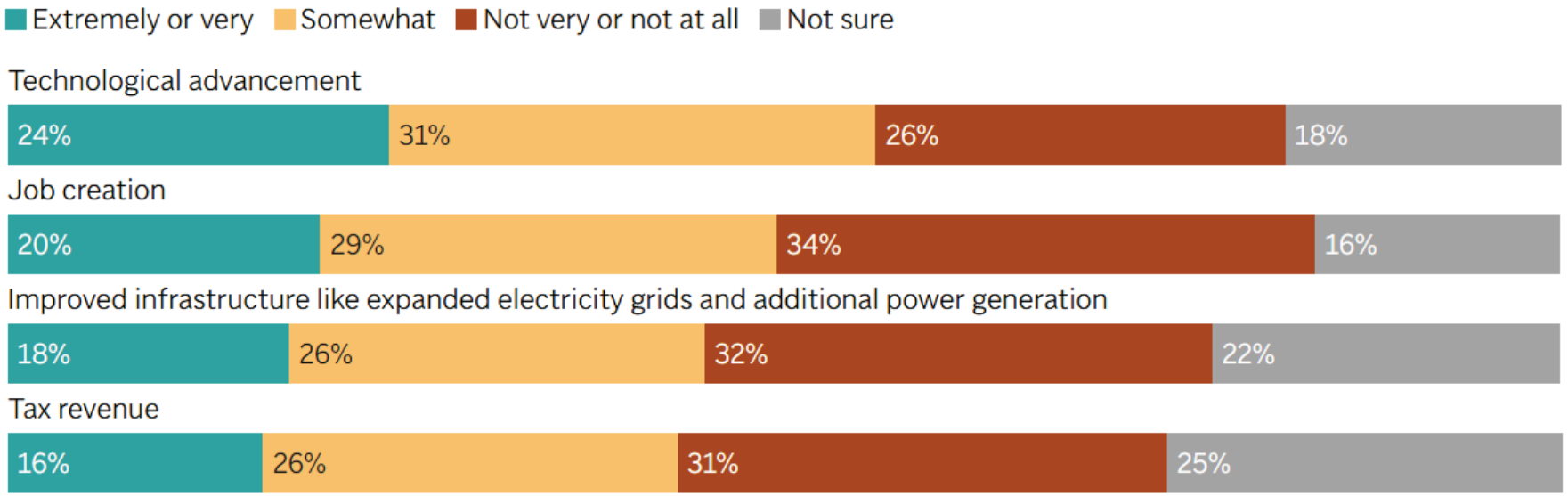
Positive impacts are meaningful

Technological advances / geopolitical competitiveness, construction jobs and tax benefits are among the top perceived benefits.

Tax Revenues and Improved Local Services

Direct Payments/Incentives to Communities

Net positive water use



Source: The Associated Press-NORC Center for Public Affairs Research and the Energy Policy Institute at the University of Chicago. Released Sept. 16, based on polling in August. MoE: +/- 2.2%

For More Information

Contact: Ian Hoffman, ihoffman@lbl.gov

Acknowledgment

This work was funded by the U.S. Department of Energy's Office of Critical Minerals and Energy Innovation, under Contract No. DE-AC02-05CH11231.

Disclaimer

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, or The Regents of the University of California.

Ernest Orlando Lawrence Berkeley National Laboratory is an equal opportunity employer.

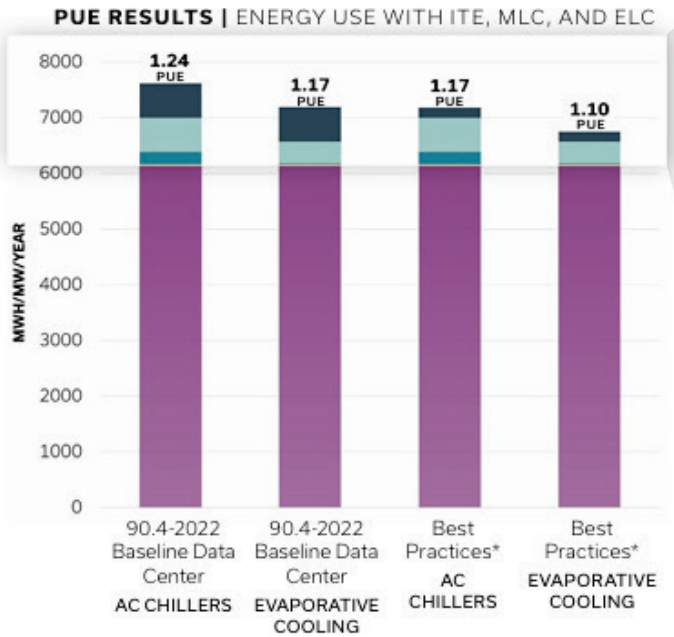
Copyright Notice

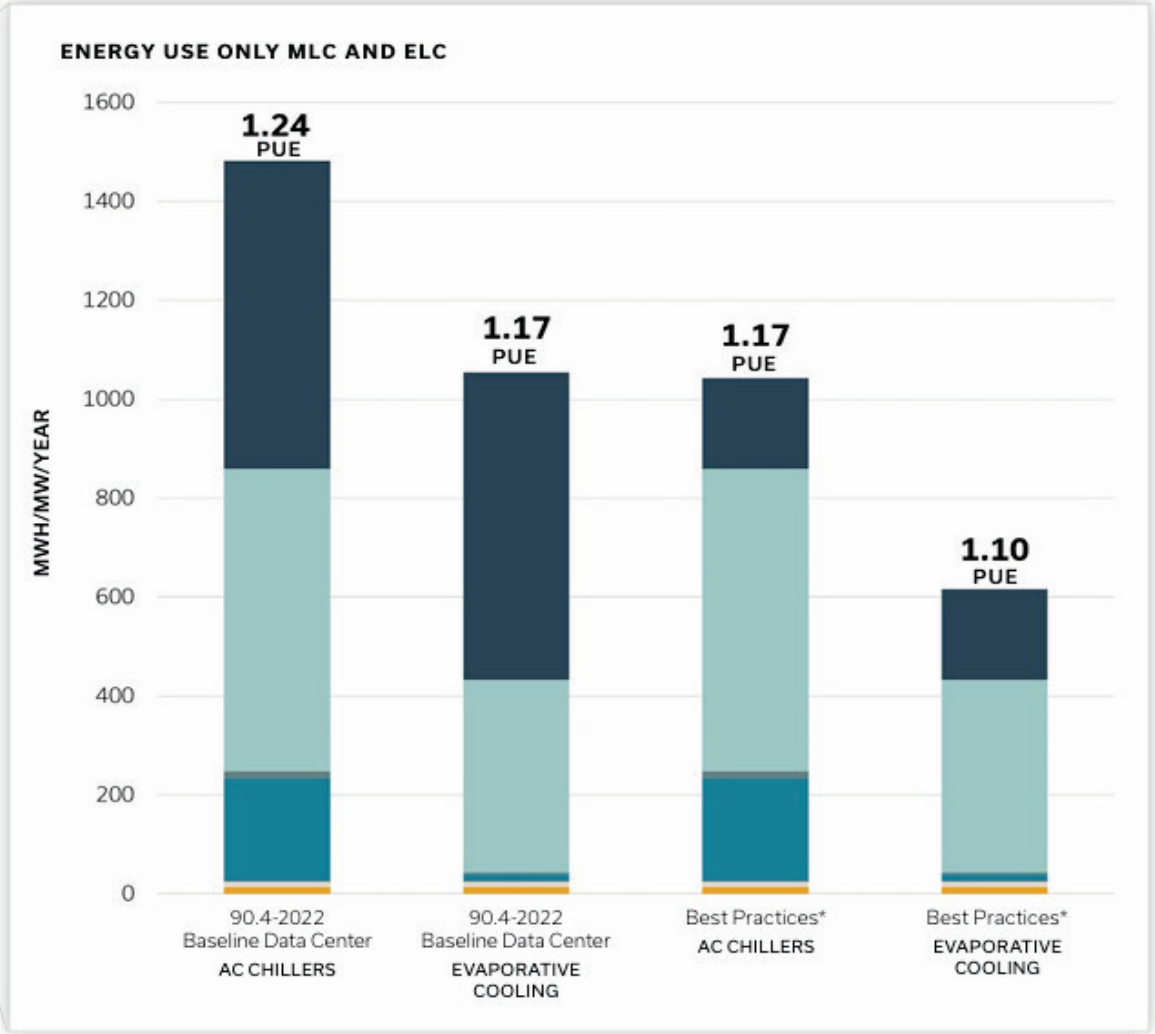
This manuscript has been authored by an author at Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231 with the U.S. Department of Energy. The U.S. Government retains, and the publisher, by accepting the article for publication, acknowledges, that the U.S. Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes

Backup

Reference material for discussion

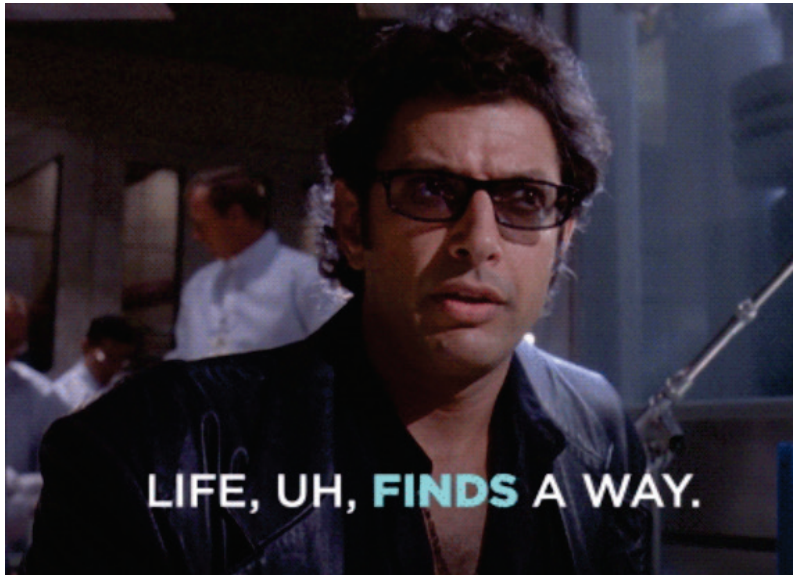
Power Usage Effectiveness: What It Says, What It Doesn't





Typical chemistry for direct-to-chip coolant (hyperscale AI/HPC)

- Deionized water with generally 25% polyethylene glycol, known as PG25. Small energy penalty. vs. water alone.
- Glycol is not for freeze protection but a stable medium for additives:
  - Bio-inhibitors (tolyltriazole, benzotriazole, or both)
  - Corrosion inhibitors (nitrite, molybdate, silicates)
- pH @ 7.0 to 8.5 or aluminum and copper corrosion rates increase nonlinearly.



Higher thermal loads → higher flow volumes and velocities → less corrosion/erosion protection

Cue two-phase cold plates that flow refrigerants over chips. Raises “new” issues: PFAS vs. toxicity, flammability, high GWP

Data centers and grid gain from demand flexibility

- **Streamlined interconnection**
- **Availability/uptime**
  - Power failures - internal and external - are the #1 cause of major data center outages
  - DR programs and capacity markets provide visibility into grid conditions
  - DR makes it less likely that the local node or circuit on the grid will go down
    - More reliability for the data center and neighboring community
- **Cost containment**
  - Applications for rate increases are up
  - Reductions in peak demand reduce facility exposure to higher rates (e.g., demand charges)
- **Demand flexibility is entwined with operational flexibility, enhanced resilience**
  - Workload management
    - System balancing
    - Prioritization and timing (e.g., of batch work)
  - Processor power capping
  - Cooling setpoints
  - All help ensure backup/onsite supplies can deliver on the most mission-critical workloads

Rate pressure tracks concentration

PJM (Mid-Atlantic to PA/NJ)

- Data center load is concentrated.
- Capacity prices rose sharply across recent auctions. [Get figures and driver quant from PJM independent monitor, Monitoring Analytics]
- Virginia created a separate large-load rate class.
- Other states considering large-load tariffs.

ISO-NE / Maine

- No comparable concentration.
- About 285 MW in study region-wide.
- Maine's medium and large data center load today is effectively de minimis.

Rate pressure tracks concentration, not the mere presence of a data center. Cost allocation and interconnection terms decide who pays. Big operators increasingly fund new generation, both on the grid and onsite.

Sources: ISO-NE 2026 CELT. PJM figures pending primary verification.