AI DATA CENTERS: Turning Compute Demand into a Strategic Power Constraint

Article by **Lucian Gavriliuc**, P.E., Executive Director, Energy Storage, **E3 Consulting, LLC** May 7, 2026

EXECUTIVE SUMMARY

AI data centers are rapidly becoming one of the largest new sources of electricity demand in modern history. Single campuses now require up to one gigawatt or more of continuous power, on par with a mid-sized city. At this scale, data centers are no longer "IT infrastructure;" they are heavy industrial loads with material implications for grid planning, capital allocation, and regulatory oversight.

The industry has largely exhausted efficiency gains at the facility level. Power Usage Effectiveness (PUE) has converged near theoretical limits, and further improvements will not offset the step-change in demand driven by AI workloads. The constraint has shifted decisively from facility efficiency to absolute power availability. AI and high performance computing (HPC) workloads require 10-20 times the rack power of traditional enterprise IT, compressing decades of grid load growth into just a few years. Grid interconnection timelines in key markets have stretched to 5-10 years and are fundamentally misaligned with AI deployment cycles of 2-3 years, creating structural bottlenecks for both IT companies and power providers.

For executive leadership, the implication is straightforward: compute (the processing capability of a system to perform calculations and execute workloads, typically delivered by CPUs, GPUs, and specialized accelerators) is now a power-constrained asset class. AI growth strategies are inseparable from grid conditions, regulatory frameworks, and access to firm, 24/7 power. Strategic responses are emerging around co-located generation, long-term power contracts (nuclear, gas, renewables plus storage), and deeper coordination with utilities. Those developing and investing in AI infrastructure should treat AI infrastructure as a core capital allocation and risk management decision. The ability to secure power, where, when, and at what cost, will define the ceiling on AI growth over the next decade.

Lucian Gavriliuc, P.E., serves as E3's Executive Director of Energy Storage, where he offers technical advisory services to developers, lenders, investors, and other stakeholders in the energy storage industry. In this article, he discusses the rapid adoption of AI and its impact on the power grid, which is presenting significant challenges for utilities, developers, and investors.

Lucian emphasizes that AI growth strategies must now be aligned with grid realities. Recognizing and addressing these constraints is crucial for those looking to lead or invest in the next wave of AI-driven innovation.

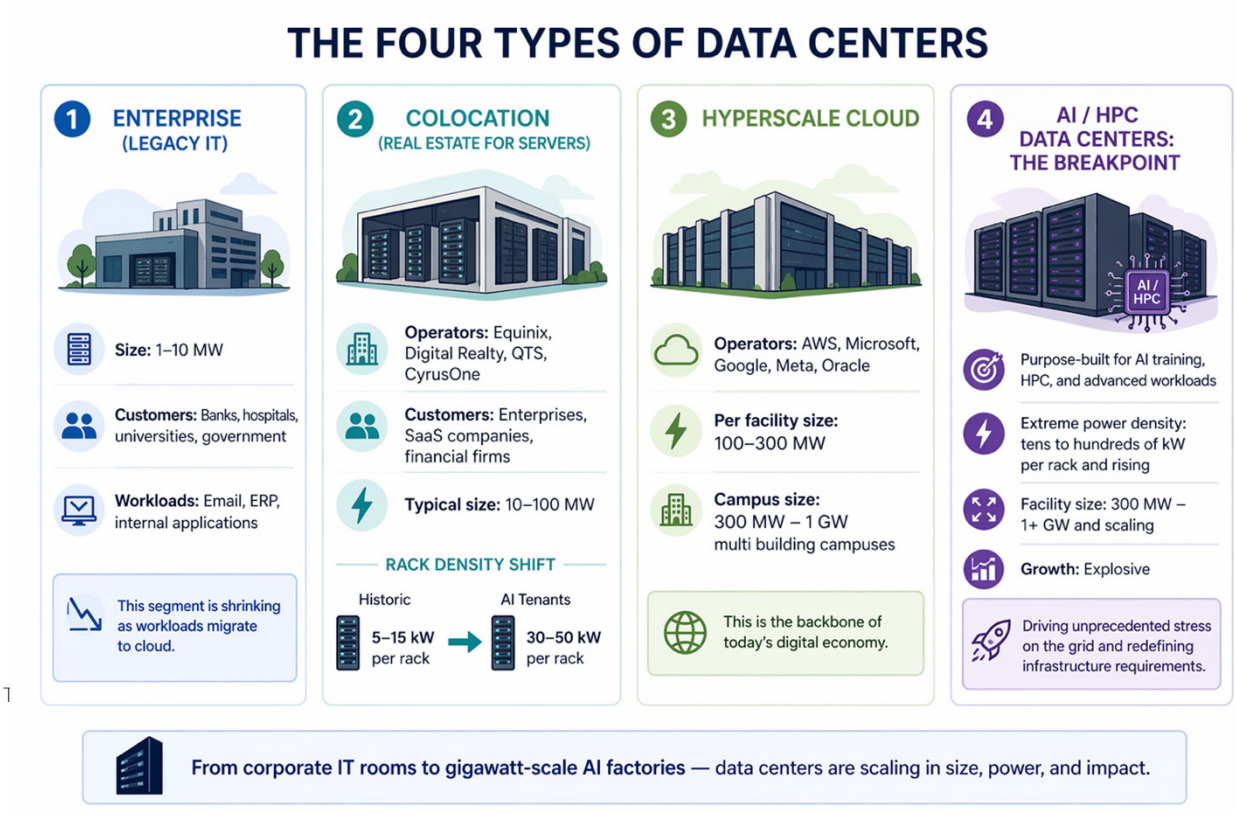
KEY TAKEAWAYS

- Power, not the availability of servers or land, is now the primary constraint on AI growth, with AI campuses scaling to 500 MW-1 GW+ of continuous demand.
- AI data centers behave like heavy industrial loads, reshaping utility planning, transmission build out, and regional economic development priorities.
- Efficiency gains are largely tapped out; incremental PUE improvements will not offset the step change in compute demand driven by graphics processing units (GPUs) and massive training clusters.
- Electrical interconnection and infrastructure permitting timelines are now a strategic risk, often longer than the business plans they support, requiring earlier, deeper engagement with utilities and regulators.
- Storage and firm generation are moving from optional to mandatory, becoming core components of AI campus economics, reliability, and social license to operate.

CONTEXT

Data centers have evolved from supporting internal IT to powering cloud, AI, and digital services at a global scale. Traditional enterprise and colocation facilities typically drew from 1-100 MW; hyperscale cloud campuses expanded to 100-300 MW. AI/HPC data centers now represent a fourth, distinct category: purpose-built "compute factories" operating at hundreds of megawatts to gigawatt-scale continuous load.

This new class of digital infrastructure is colliding with an aging grid, lengthy interconnection processes, and rising pressure to decarbonize, making AI driven load growth a central issue for utilities, regulators, and capital markets.



1.1

Open AI, Generated April 2026

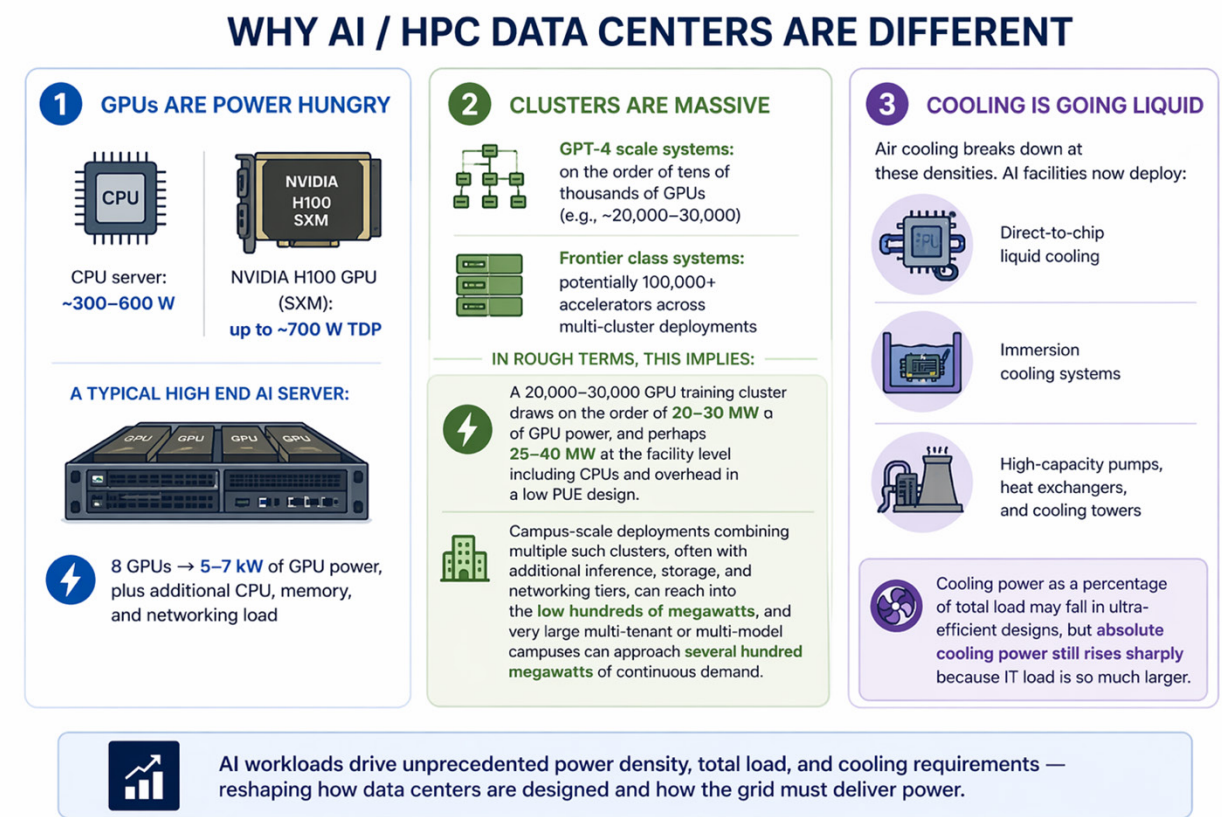
ANALYSIS

1. From Incremental IT Load to Mega Load

AI and HPC workloads have transformed data centers from incremental, distributed IT loads into concentrated "mega loads" comparable to steel mills, LNG terminals, or semiconductor fabrication plants. Rack power has risen from 5-10 kW for enterprise IT to 100-150+ kW for AI training, driven by GPU based servers and massive, tightly coupled clusters.

At the same time, single-site power requirements have escalated from 5-10 MW (circa 2005) to 300 MW-1 GW by 2025, with campus footprints rivaling those of major generation assets. Utilities now face multi gigawatt development pipelines concentrated in a few regions, compressing what used to be decades of load growth into a single planning cycle.

The limiting factor is no longer capital or technology. It is the ability to deliver reliable power at scale, within the required timeframe, under regulatory constraints.



Open AI, Generated April 2026

2. Efficiency has Peaked; Demand tracks Compute

Over the past two decades, data center PUE improved from ~2.0+ to 1.1-1.2 for leading hyperscalers, meaning 85-90% of facility power now reaches IT equipment. Cooling's share of load has fallen as a percentage, but in absolute terms, both compute and cooling power continue to climb as workloads become increasingly power dense.

Further gains at the facility level are marginal relative to the increase in compute intensity. Cooling remains significant in absolute terms, but it is no longer the primary lever. Cost and risk are shifting from "how efficiently can we cool servers?" to "how much firm power can we secure, at what price and carbon profile, over **what horizon?"** This reframes AI infrastructure as a long duration energy commitment rather than a short cycle IT spend.

3. System-Wide Impact

For utilities and grid operators, AI campuses join electrification (EVs, heating, industrial) and flexible digital loads (e.g., crypto mining) as one of three major new load classes, driving changes in generation planning, transmission expansion, and storage deployment. Total grid side investment for a single AI campus can reach $100M-$500M+ before a single server is installed.

For technology companies and large compute buyers, interconnection, permitting, and access to firm power have become board level constraints on product roadmaps and AI expansion plans. For industrial and large commercial customers, AI driven demand can change regional power prices, availability, and reliability profiles, altering the economics of long lived manufacturing and infrastructure investments.

Across all industrial sectors, expansion planning must now consider power availability and grid strategy as part of location decisions, M&A, and long term capital planning, not just as an operational afterthought.

4. Capital Allocation, Risk, and Timing

Three structural timing gaps are emerging:

- **Generation:** A one GW AI campus consumes ~8.7 TWh/year, requiring nuclear scale output or several GW of renewables plus storage, yet new firm generation assets typically have multi year development and regulatory approval timelines.
- **Transmission and distribution:** High voltage lines, bulk substations, and 230-500 kV interconnections can take 5-10 years to plan and build, while AI deployment targets 2-3 years to revenue.
- **Regulation and markets**: Interconnection queues and approval processes are being overwhelmed by the volume and size of AI driven requests, with regulators and ISOs still adapting their frameworks.

This creates a real risk that compute capacity is deployed ahead of power availability, resulting in stranded assets, delayed revenues, and cost escalation. There is also increasing reputational and political risk where projects are perceived to compete with other loads or conflict with decarbonization goals.

IMPLICATIONS FOR DECISION-MAKERS

*For CEOs and corporate board*s, AI infrastructure is now a strategic energy position as much as a technology decision. Where AI campuses are located, how firm power is secured 24/7, and how utilities are engaged will determine the ability to scale AI products and services at an acceptable cost and risk.

For CFOs, long-term PPAs, equity stakes in generation or storage, and co-investment in grid upgrades are moving from edge cases to mainstream tools for securing AI capacity, with balance-sheet and rating implications.

F*or CIOs and CTOs*, architecture decisions (how much to centralize vs. distribute, how much flexibility to design into workloads, how to use storage and flexible loads) now have a direct impact on grid interconnection feasibility, regulatory posture, and resilience strategies.

STRATEGIC OPTIONS

1. Make Power a Core Criterion in AI Site Selection

Embed regional grid capacity, interconnection timelines, and regulatory posture as hard gates in AI data center and cloud region decisions, alongside tax, labor, and real estate costs.

2. Pursue Co Planned Models with Utilities

Shift from "serve load" to "co plan load" by engaging early with utilities on dedicated generation, renewables plus storage portfolios, and multi GW regional energy hubs tied to data center campuses.

3. Secure Firm, 24/7 Supply through Structured Power Deals

Evaluate long term contracts with nuclear (including SMR where credible), combined-cycle gas turbines (CCGT), and large scale storage as part of a diversified, resilient supply stack, rather than relying solely on spot markets or generic "green" claims.

4. Design for Flexibility Where Possible

Where workloads allow, build controllable or curtailable AI capacity that can participate in demand response, similar to how some crypto miners operate as dispatchable load, creating revenue streams and easing interconnection constraints.

5. Integrate Storage as Core Infrastructure, Not an Add On

Treat storage as mandatory for interconnection deferral, reliability, and renewable firming, especially where AI load is highly continuous but supply is increasingly variable.

FORWARD-LOOKING PERSPECTIVE

Over the next 3-7 years, expect:

- Persistent tension between AI deployment and grid readiness, with interconnection queues, permitting, and public scrutiny acting as rate limiters on AI expansion in certain regions.
- Acceleration of new market structures that explicitly value flexible load, storage, and co located generation around large AI campuses.
- Growing regulatory and political scrutiny of AI data center footprints, especially where they intersect with decarbonization targets, local reliability, and water or land use.

Organizations that secure interconnection early, lock in firm power, and integrate energy strategy into AI planning will capture a disproportionate advantage. Those that treat power as an afterthought will find themselves constrained, not by capital or technology, but by the grid itself. The question is no longer whether AI infrastructure can be built. It is whether it can be powered.

At E3 Consulting, we are working alongside developers, utilities, and investors to navigate these challenges, from interconnection strategy and power supply planning to the integration of storage and co-located generation.

If you are interested in learning more about our Data Center Services, or would like to discuss your project, please reach out to Lucian Gavriliuc, PE, at Lucian.Gavriliuc@e3co.com or (425) 393-3598.

About the author: Lucian Gavriliuc is the Executive Director of E3's Energy Storage Practice. An electrical engineer with more than two decades of power industry experience, he has designed electrical power systems and conducted power studies in accordance with IEEE standards and applicable codes, including the NEC, NFPA 70E, NESC, and ANSI, and remains active in IEEE. He holds a BS in Electrical Technology (Power) from the University of Houston, where he graduated cum laude, and an MBA from the University of Georgia.

References

ISO Large Load & Interconnection Materials

NYISO

Large Load Interconnections https://www.nyiso.com/documents/20142/33938587/LargeLoadForecast_ManualUpdates_LFTF_20221021_V1.pdf
Interconnection & Planning Portal https://www.nyiso.com/interconnections
NYISO Gold Book (Load Forecast) https://www.nyiso.com/documents/20142/51231901/2025-Gold-Book-Baseline-Forecast-Tables.xlsx/29e041cb-52b1-49a0-ac48-010b71a0eea7

ERCOT

Large Load Interconnection Process Q&A https://www.ercot.com/files/docs/2025/12/24/Large-Load-Interconnection-Process-Q-A.pdf
System Planning & Weatherization Update (Board Presentation) https://www.ercot.com/files/docs/2025/12/02/16.2-System-Planning-and-Weatherization-Update_Revised.pdf

CAISO

Large Load Integration Initiative https://www.caiso.com/generation-transmission/load/large-load
Interconnection Queue Reports https://www.caiso.com/library/interconnection-queue-reports
Large Load Considerations Issue Paper (2026) https://www.caiso.com/documents/issue-paper-large-load-consideration-jan-20-2026.pdf

PJM

Large Load Interconnection - Primary PJM documentation Interconnection Process Manuals and Large Load Guidance https://www.pjm.com/planning/rtep-development/stakeholder-process/developers
Large Load Interconnection - Secondary Commentary Zero Emission Grid - "PJM Large Load Interconnection Process (2026)" (blog summary of PJM processes) https://www.zeroemissiongrid.com/insights-press-zeg-blog/pjm-large-load-interconnection-process-2026/

Data Center Energy & Efficiency

Uptime Institute Global Data Center Survey https://datacenter.uptimeinstitute.com/rs/711-RIA-145/images/2024.GlobalDataCenterSurvey.Report.pdf
International Energy Agency (IEA) - Data Centres and Data Transmission Networks https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks
Lawrence Berkeley National Laboratory (LBNL) - Data Center Energy Use Reports https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report_1.pdf

AI Infrastructure & Compute

NVIDIA H100 Technical Overview https://www.nvidia.com/en-us/data-center/h100/
AMD Instinct MI300 Series Overview https://www.amd.com/en/products/accelerators/instinct/mi300.html
Google TPU Overview https://cloud.google.com/tpu
AWS Trainium https://aws.amazon.com/machine-learning/trainium/
Clarifai - "NVIDIA H100: Price, Specs, Benchmarks & Decision Guide" (2025) https://www.clarifai.com/blog/nvidia-h100

Grid Load Growth & Planning

Grid Strategies - National Load Growth Report (2025) https://gridstrategiesllc.com/wp-content/uploads/Grid-Strategies-National-Load-Growth-Report-2025.pdf
Utility Dive - ERCOT Large Load Queue Growth https://www.utilitydive.com/news/ercots-large-load-queue-jumped-almost-300-last-year-official/808820/
Uptime Institute - Global Data Center Survey / PUE Trends https://journal.uptimeinstitute.com/large-data-centers-are-mostly-more-efficient-analysis-confirms/

Crypto Mining & Flexible Load

ERCOT Demand Response & Large Flexible Load Materials https://www.ercot.com/gridinfo/resource
Texas Blockchain Council / ERCOT Market Participation Materials (various public filings and ERCOT stakeholder discussions)
Texas Bitcoin Miners Demand Response Press (e.g., Texas Blockchain Council) https://www.prnewswire.com/news-releases/texas-bitcoin-miners-turn-off-to-conserve-power-for-the-grid-301547420.html