Click Bait or Cause for Concern

Robert E. Young
COO Digital Iron Network,
Managing Director
Economists.com

May 21, 2025

Agenda

- Top 10 List
- US Data Center Landscape in Electric Load Growth
- Data Center Energy Use
- Data Center Energy Efficiency
- Server Rack Density and AI Compute
- Nvidia Chips
- Policy Issues
- Floating Point Precision
- AI Compute Forecasts

Top Ten Data Center Companies

Rank	Company	Est. Capacity (MW)
1	Amazon Web Services	10,000
2	Microsoft Azure	8,000
3	Google Cloud	7,000
4	Meta Platforms	6,000
5	Equinix	2,000
6	Digital Realty	2,000
7	NTT Global D.C.	1,500
8	CyrusOne	1,000
9	CoreWeave	1,000
10	Flexential	500

National Landscape

- US Electricity growth has returned
  - 2025 electricity sales expected to increase by 2%, or 83 TWH
  - 83 TWH = 83,000,000,000 kWh
- Data center are largest driver of growth
  - Share of US electricity will rise from about 4% to 8-12% of total
  - Expected to add another 43 TWH by 2030

The Long Pause is Over: Electricity Sales Are Growing

U.S. Electricity Consumption (1990-2026)

TWH (Source: EIA STEO May 2025)

Sources of Energy Use

- Data center energy use is driven by three main hardware categories and varies by:
  - Age
  - Configuration
  - Type and function
- Breakdown of Energy Consumption
  - IT Equipment (40%–50%)
  - Servers – Perform computation and processing
  - Storage – HDDs and SSDs for data retention
  - Network – Switches, routers, and connectivity hardware

Sources of Energy Use (cont'd)

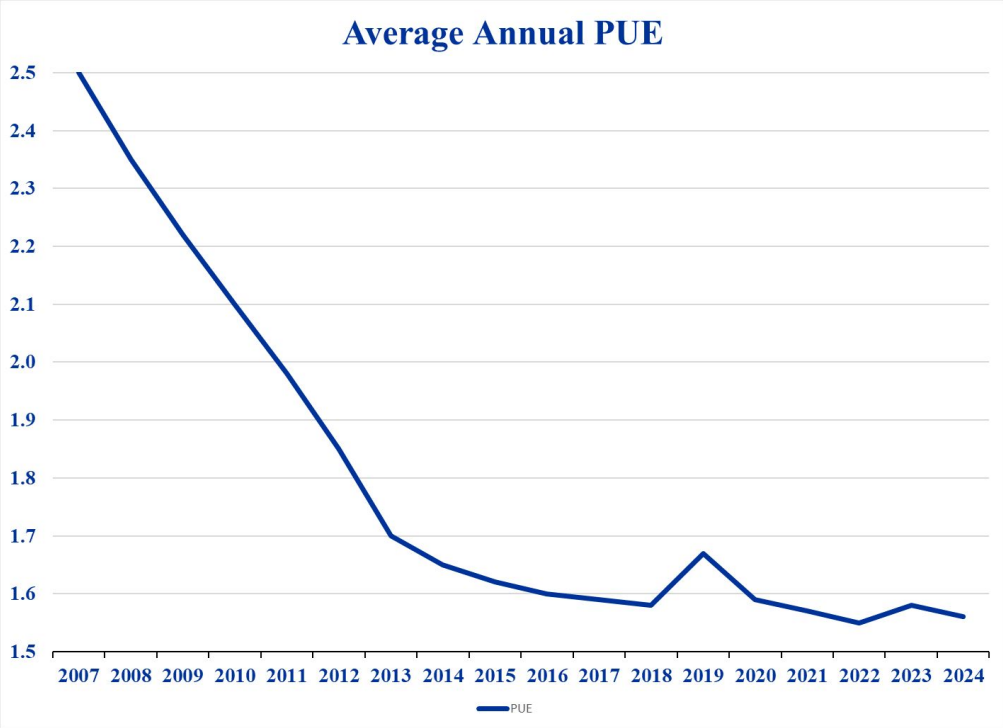
- Cooling Systems (30%–40%)
  - Maintain optimal temperatures
  - Shift from HVAC to specialized cooling technologies
- Auxiliary Components (10%–30%)
  - UPS (uninterruptible power supplies)
  - Security systems
  - Lighting and other infrastructure
  - Electric losses

Energy Efficiency

- Energy Efficiency Matters
  - Gauges how effectively electricity is used
  - Identifies trends and performance gaps
  - Drives improvement and optimization
  - Supports long-term operational strategy
- Key Metric is Power Use Effectiveness (PUE)
  - PUE = $\frac{\text{Total Facility Energy (kWh)}}{\text{IT Energy (kWh)}}$
  - Provides a standardized way to track data center efficiency over time
  - A lower PUE value indicates greater efficiency, with the theoretical minimum PUE being 1.0.

Energy Efficiency

- Trends in PUE
  - Declined rapidly from 2007 through 2018, then plateaued as the ratio approached 1.5



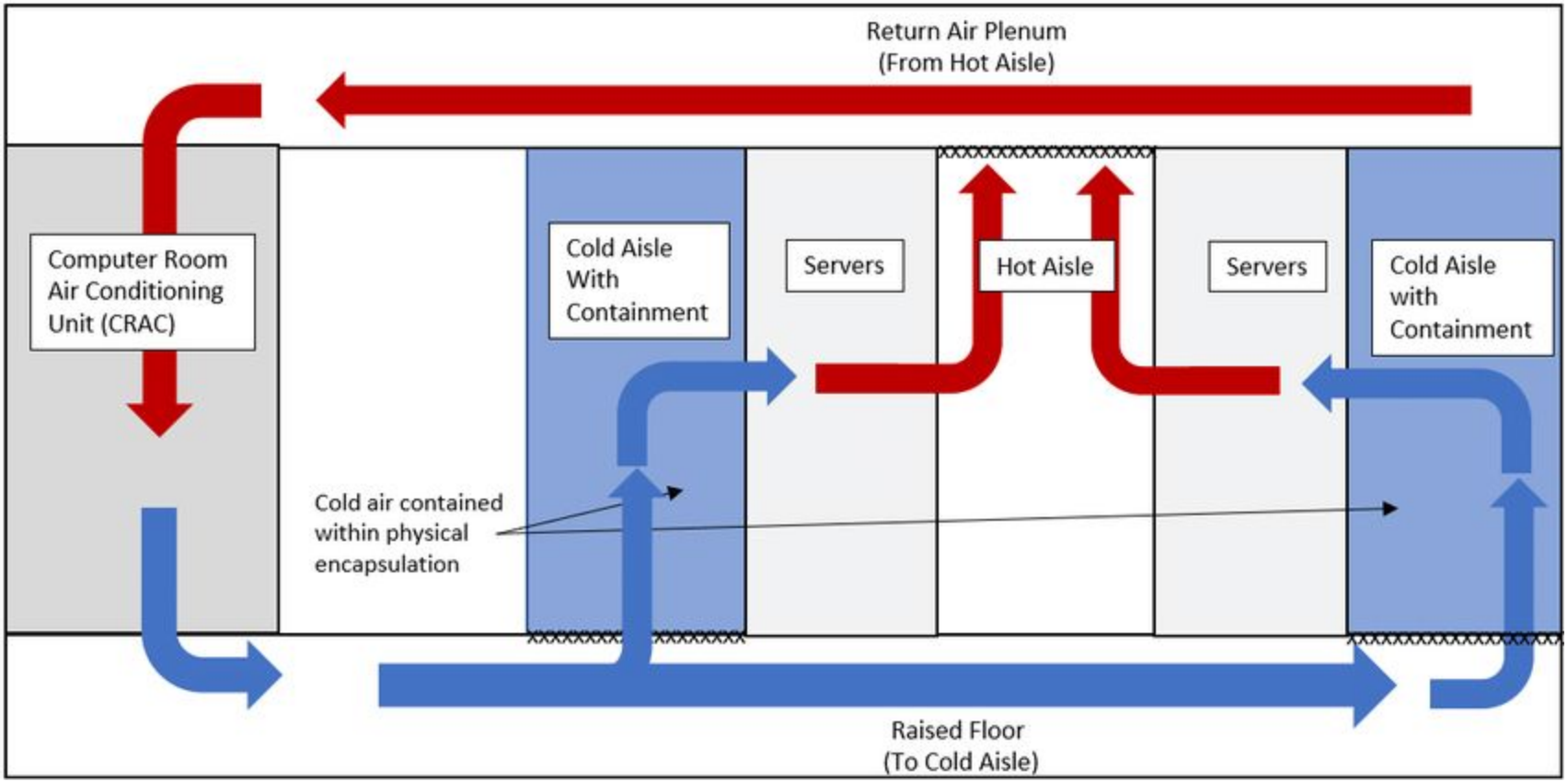
Energy Efficiency

- Improvements in PUE
  - While similar from the outside, internal data center designs are not standardized
  - Legacy data centers are difficult to modify and cost prohibitive
- Cooling is the key area to improve PUE
  - Hot/cold aisle containment
    - Prevents air mixing and improves cooling efficiency
  - Leverage outside air during suitable conditions
  - Liquid Cooling for AI Compute
- Utility Conservation 101

Cold Aisle Containment

- Traditional data centers used room-based cooling.
  - Inefficient airflow, with cold and hot air mixing - wasted energy
  - Increased rack densities made room cooling unsustainable
- CAC was developed to as a targeted solution to improve cooling efficiency and reduce operating costs to improve PUE
  - Hot/cold aisle containment
    - Prevents air mixing and improves cooling efficiency
  - Leverage outside air during suitable conditions

Energy Efficiency – Cold Aisle Containment



Energy Efficiency – Cold Aisle Containment

Digital Iron Network

Direct Current Supply

- Electric grids supply alternating current (AC) to customers
- Server racks and AI compute run on Direct Current (DC)
  - About 5-10% of data center energy use is AC/DC losses
- Infineon and Nvidia will develop direct current delivery systems

AC/DC Loss Percentages		
Source	**Efficiency**	**Loss**
Server PSUs	90–96%	4-10%
UPS systems	92–95%	5-8%
PDUs	98–99%	1-2%
Total		**10%**

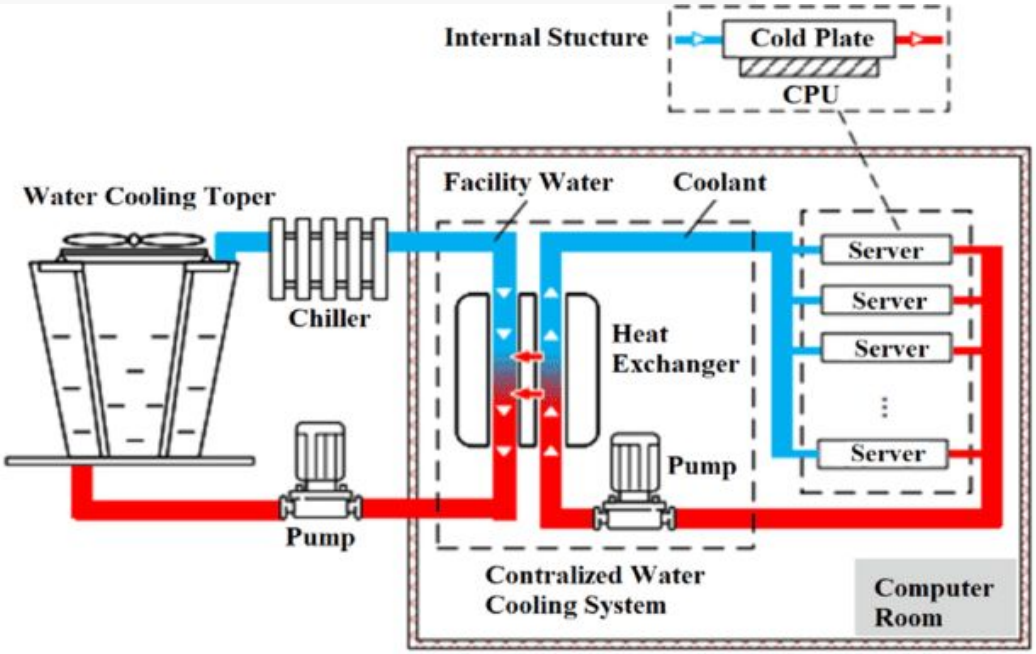
Server Rack Density and AI Compute

Rack Density and AI Compute

- Traditional rack densities (pre-2015)
  - Typically ranged from 3–6 kW per rack
  - Designed for general-purpose IT workloads (web, storage, email)
- Rise of High-Performance Computing (2016-2020)
  - Densities increased to 8–15 kW per rack
  - Driven by cloud, analytics, and virtualization
- AI/ML acceleration era (2020–present)
  - AI workloads require GPUs and custom accelerators (e.g., TPUs)
  - Rack densities now commonly 20–40 kW, and often exceed 50 kW

Rack Density and AI Compute

- Ultra-dense AI racks emerging
  - Leading AI systems (e.g., Nvidia DGX, Meta Grand Teton) reach 60–100 kW per rack
  - Require liquid or immersion cooling



Nvidia Chips – AI Compute Workhorse

- Nvidia dominant position in AI compute ~ 92%
- Early Investment in CUDA (2006) allowed use of sing units (GPU)
- Optimized for parallel processing – matrix heavy AI workloads like LLMs
- Strong developer ecosystem and software stack
- Continuous chip innovation

Nvidia Chips – AI Compute Workhorse

Nvidia DGX Power and Performance Metrics						
Nvidia Model	Year	Cost	Pflops (FP8)	$/Pflop	kW	Pflop/kW
DGX-1	2016	$129,000	1	$129,000	3.5	0.29
DGX-2	2018	$399,000	2	$199,500	10	0.20
DGX A100	2020	$399,000	5	$79,800	6.5	0.77
DGX H100	2022	$200,000	32	$6,250	10.20	3.14
DGX H200	2023	$200,000	32	$6,250	10.20	3.14
DGX B100	2023	$250,000	56	$4,464	14.3	3.92
DGX B200	2024	$350,000	72	$4,861	14.3	5.03
GB200 NVL72	2025	$3,000,000	720	$4,167	120	6.00
GB300 NVL72	2026	$3,700,000	756	$4,894	140	5.40
Ruben NVL576	2027	??	5,000	??	600	8.33

Nvidia Chips – What is FP Precision

- Floating point precision refers to how many bits are used to represent numbers in calculations
  - FP128 – 128 bits or 34 digits to FP4 – 4 bits or 2 digits
- More bits = higher precision, accurate calculation
- Fewer bits = faster compute, a lot of rounding
- Important because AI models tolerate lower precision
  - FP4 allows faster compute, less power, lower memory
  - FP128 for fluid dynamics, weather modeling, moon launch

Nvidia Chips – Move to FP4

- Massive Gains in Performance & Efficiency
- Model Quantization - algorithms allow LLMs and vision models to retain accuracy even when quantized from FP16 or INT8 to FP4.
- Reduced memory bandwidth
- Higher compute density 2x more on same chip area
- Ideal for inference workloads – After a model is trained
  - Chatbots, image recognition, recommendations, autopilot

Nvidia Chips – Move to FP4

- Robert's Rules for AI - Don't do math on Chatbots
- Question to ChatGPT: What is 6,265,108/665,223?
- Answer: ≈ 9.42
- FYI – Precision in Excel is FP64, or 15 significant digits

Policy Issues

- Generally, utilities can meet need for new generation capacity
- Concern over grid reliability & capacity
  - Hyperscale data centers demand 100–300+ MW per site
  - Stress on local and regional grids (e.g., VA, OR, IE)
  - Accelerates need for long-range grid planning reform
- Interconnection Delays
  - Lengthy interconnection queues delaying data center deployment
  - Utilities and ISOs overwhelmed by high-volume load requests
  - Calls for streamlined grid access policies

Policy Issues

- Grid Connection challenges
  - CenterPoint Energy, TX connection queue increased from 1GW to 8GW in less than a year
  - Virginia connection requests can take 7 years
  - Mirrors huge queues and long delays for solar and wind interconnections
- Larger data centers operators turning to onsite generation
  - Google, Microsoft, Amazon, & Meta interest small nuclear reactors
  - Amazon built near Susquehanna, PA nuclear plant for direct connect

Policy Issues

- Grid Connection challenges –AI to the Rescue
- PJM Interconnection queue was 2,600 GW in Dec 2023
  - Average 40-month delay
- Google AI worked with PJM to:
  - Automate Application Reviews
  - Integrate siloed data bases
  - Accelerate project approvals
- Outcome: Cleared 72 GW as of May 2025

Policy Issues

- Emissions & Clean Energy Goals
  - Risk of increasing fossil fuel generation if clean firm supply is lacking
  - Conflict with state climate laws and decarbonization mandates
  - Push for 24/7 carbon-free energy standards (e.g., Google, Microsoft)
- Planning & Siting Challenges
  - NIMBY resistance growing in residential or agricultural zones
  - Lack of coordination between land-use and energy planning
  - Need for electric transmission planning reform

Policy Issues

- Ratepayer Cost Allocation
  - Transmission and substation upgrades often socialized
  - Potential for cross-subsidization by residential ratepayers
  - Consider cost-of-service tariffs, impact fees, or reservation pricing
- Regulatory Oversight Gaps
- Many data centers operate outside PUC jurisdiction
- Lack of centralized oversight for non-utility large loads
- Emerging need for data center load forecasting

AI Compute Forecast

- Without AI compute, data center energy growth is manageable
- The key driver is AI compute electricity use
- What is known is energy use per Nvidia GPU
- Nvidia does not release sales of GPUs and servers
  - ☞ But slips happen😎
- Nvidia Nov 2024 Earnings Summary page 5 said H200 "NVIDIA H200 sales increased significantly to double-digit billions since launch, Aug 2024
- Divide $10 billion by $32k cost per H200, so min of 312,550 sold in 3 months

Thank you for your time and attention!

Robert E. Young
COO Digital Iron Network
Managing Director
Economists.com

May 21, 2025