BofA SECURITIES

Industrials/Multi-Industry

AI is driving a revolution in data center infrastructure

Industry Overview

Data center: tracking the money and megawatts

The wall of worry around data center spending continues to build from slow artificial intelligence (AI) monetization, to DeepSeek-related fears of improved algorithms, and, most recently, overbuilding risks. Despite concerns, AI firms have grown rapidly with OpenAI reaching a $10bn revenue run-rate in May, according to press reports. This report has (1) updated BofA data center infrastructure forecasts, (2) a deep dive on how AI will change data center architecture, and (3) an update on the training versus inference power debate.

BofA forecast 16% CAGR ('24-'28) for infrastructure

We forecast total data center spending (servers, infrastructure, & engineering) to rise at a 13% CAGR over 2024-28E. This compares to our prior 13% CAGR over 2023-26E, despite a 2024 starting point which is now 33% higher (e.g., $326bn versus $245bn forecast). We forecast a faster 16% CAGR over 2024-28E for infrastructure-related spending. We acknowledge that BofA's hyperscaler capex forecast is for a meaningful deceleration in 2026 versus 2025. Over the past 12 months, BofA's 2026 hyperscaler capex have been revised higher by 58%. We also see several reasons why infrastructure-related capex growth will prove to be more durable.

AI continues to drive changes to data center infrastructure

Nvidia's Blackwell chips are already driving adoption of direct-to-chip liquid cooling. However, the next generation of AI chips is likely to drive further changes, particularly in power distribution. Power and thermal engineering is becoming more important, with Nvidia, AMD, and Intel all introducing their own rack-level architecture. Operators, such as Microsoft and Meta, have also introduced reference designs to steer future development pathways. We see these changes as likely to benefit existing scaled incumbents, who are able to partner with semiconductor companies on new rack-level solutions. This should benefit Vertiv (VRT), Eaton (ETN), and Johnson Controls (JCI) in our coverage.

Inference is using more electricity…

Inference is turning out to be surprisingly power hungry. This is due, in part, to increased adoption of "chain of thought" or reasoning models, such as OpenAI's o1 model, DeepSeek's R1, and xAI's Grok-3. Even as the efficiency of AI algorithms improve, the models handling inference are growing in complexity and size. We also see AI use cases expanding from text to image and video, which require an order of magnitude more computing power.

…which means power remains the constraint

We think this bodes well for AI lateral stocks in the power generation space, such as Buy-rated GE Vernova (GEV). Behind-the-meter power generation, such as xAI's gas turbines at its Memphis location, offer a fast route to power. The Homer City redevelopment, which includes up to 4.5 gigawatts of gas power generation at a former coal power site, shows the potential for large-scale deployments.

17 June 2025

Equity
Americas
Industrials/Multi-Industry

Andrew Obin
Research Analyst
BofAS
+1 646 855 1817
andrew.obin@bofa.com

David Ridley-Lane, CFA
Research Analyst
BofAS
+1 646 855 2907
david.ridleylane@bofa.com

Sabrina Abrams
Research Analyst
BofAS
+1 646 556 3520
sabrina.abrams@bofa.com

Devin Leonard
Research Analyst
BofAS
+1 646 855 3698
devin.leonard@bofa.com

Refer to important disclosures on page 23 to 26. Analyst Certification on page 22. Price Objective Basis/Risk on page 21.
12842194

Timestamp: 17 June 2025 05:00AM EDT

BofA data center market forecasts

Exhibit 1: BofA Data Center Market forecast (2022-28E)

We forecast total data center spending to grow at a 13% CAGR over 2024-28E

($bn)	2022	2023	2024	2025E	2026E	2027E	2028E	'24-'28E CAGR
Servers	124	140	232	278	307	339	376	13%
Networking	18	20	21	22	24	25	27	7%
Storage	8	7	7	7	8	8	9	6%
Sub-total: IT equipment	**150**	**167**	**260**	**307**	**339**	**373**	**411**	**12%**
Construction & installation	23	26	30	35	40	46	51	14%
Electrical	13	15	18	21	25	28	32	15%
Air-cooling	*5*	*6*	*8*	*9*	*10*	*12*	*13*	13%
Liquid-cooling	*0*	*1*	*1*	*3*	*5*	*7*	*9*	65%
Thermal	6	7	9	12	15	19	22	24%
Generators	4	5	6	7	8	9	10	15%
Engineering	2	3	3	3	4	5	5	16%
Sub-total: Infrastructure	**48**	**55**	**67**	**79**	**92**	**106**	**121**	**16%**
Total	**198**	**222**	**326**	**386**	**431**	**479**	**532**	**13%**

Source: BofA Global Research

BofA GLOBAL RESEARCH

We see AI adoption driving a 13% market-wide CAGR over 2024-28E. Relative to our previous report (Who Makes the Data Center, September 2024), the 2024 market estimate is now 33% higher at $326bn. This significant upward revision was driven by greater spending on AI servers, which have higher pricing versus traditional servers. A few notes by category:

- **Electrical** market growth is closely tied to the electrical capacity of data centers, regardless of square footage or server costs.
- The **Thermal** market grows faster due to the additional requirements of higher rack densities. We view liquid-cooling offerings as largely additive to the existing air-cooling market.
- **Construction & installation** spending is tied both to the square footage of data centers and the amount of infrastructure equipment. Rising rack density puts downward pressure on the "white space" square footage for IT equipment, but increases the square footage of related "grey space" for infrastructure.
- **IT equipment** includes a greater portion of spending tied to refresh/replacement versus infrastructure-focused categories. For example, Amazon Web Services and Google Cloud both put the useful life of servers at six years. This results in a slower rate of growth for the market, as a larger portion reflects the installed base versus new data center build out.

Hyperscalers' capex: decelerating in '26

BofA analysts project capex by cloud service providers to rise 48% y/y to $386bn in 2025 and 5% in 2026 to $404bn. For more details, see their 5/9/25 note. Hyperscaler capex estimates continue to be revised higher. In the past 12 months, 2026 capex forecasts have been revised up by 58% (or $143bn).

We forecast data center infrastructure spending (e.g., excluding IT equipment) to rise 18% y/y in 2025 and 17% y/y in 2026. This is a far more modest deceleration. There are four reasons for this disconnect:

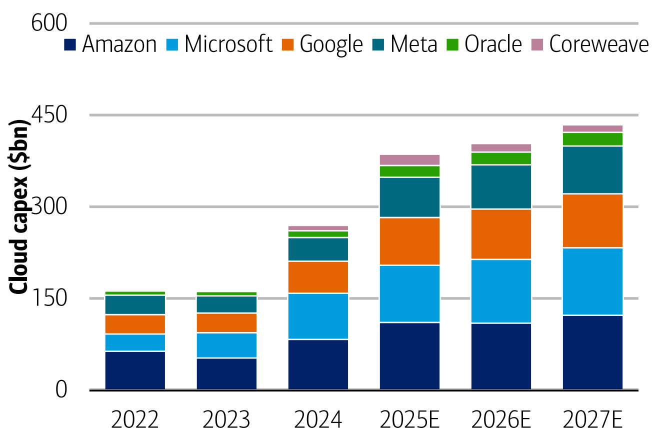
1. The majority of the incremental hyperscaler capex went for higher-priced AI accelerator GPU servers. Anecdotally, some in-process data center sites were repurposed for AI server racks.

2. Colocation companies needed time to raise equity and debt for AI-focused infrastructure investments. Leading colocation firms retrofit a portion of existing facility space for AI-dedicated leasing in 2024-25, while finalizing plans for new build infrastructure in 2026 and beyond.

3. Construction timelines are very difficult to accelerate. At industry conferences, we heard that bottlenecks around electricity availability, permitting, electrical equipment, and skilled labor all made it difficult to move more quickly.

4. It takes time for electrical and thermal equipment vendors to ramp capacity. Shipment volumes have grown, but so have backlogs.

Exhibit 2: Cloud capex (including leases)
All vendors projected to grow spending in 2025, 2026, and 2027



Source: BofA Global Research, company filings

BofA GLOBAL RESEARCH

Exhibit 3: Cloud capex spending (including leases)
BofA analysts forecast 44% y/y growth in 2025E

($bn)	Amazon	Microsoft	Google	Meta	Oracle	Coreweave	Total
2022	64	28	31	32	7	0	**162**
2023	53	41	32	28	7	0	**161**
2024	83	76	53	39	11	9	**261**
2025E	111	94	78	66	19	18	**386**
2026E	110	104	82	72	21	14	**404**
2027E	123	110	88	78	22	12	**434**

Source: BofA Global Research, company filings

BofA GLOBAL RESEARCH

Capex plans reaffirmed following Microsoft lease changes

- According to press reports, Microsoft has failed to move forward with pre-leasing agreements across a number of projects totaling ~2 gigawatts (GW). According to press reports, Microsoft has paused plans to build $1bn of data centers across three sites in Ohio (New Albany, Heath, and Hebron). The company continues to pre-construction activities (e.g., utility connections, roadway construction).

- Noelle Walsh, President of Microsoft Cloud, explained strong cloud- and AI-related growth is driving capex investment, but the company was "slowing or pausing some early-stage projects" to adjust the timing of capacity additions. The company has reiterated plans to spend over $80bn on data center capex in FY25.

- In April, Alphabet CEO Sundar Pichai reaffirmed prior capex plans calling for $75bn to build out data center capacity to support AI.

- In the annual letter to shareholders in April, Amazon's CEO Andy Jassy repeated his belief in the ROIC and FCF generation of data center capex. "Our AI revenue is growing at triple digit YoY percentages and represents a multi-billion-dollar annual revenue run rate."

Colocation companies raising capital

However, collectively colocation companies are a larger part of the data center market. Colocation companies have been active in raising capital. While not comprehensive, colocation companies raised at least $41bn in equity and debt in 1H25 to date (e.g., through mid-June) This follows $30bn in equity and debt in 4Q24 and supports further spending for new development.

We also tracked over $28bn in US data center projects entering the planning stages in 1H25 to date. The largest of these is Red Wolf's Kansas facility, which has not yet been finalized but would be a six-building campus with each building valued at approximately $2bn each. Red Wolf has filed plans with the Unified Government of Wyandotte County/Kansas to build the data center campus, but plans have not been finalized to the planning department.

Exhibit 4: Selected capital raises by colocation firms
>$41bn in debt and equity raised in 1H25-to-date

Capital raises

Month	Name	Equity/Debt	Value ($mn)
Jan '25	Digital Edge	Equity	640
Jan '25	Digital Edge	Debt	1,000
Jan '25	DataBank	Equity	250
Jan '25	CyrusOne	Debt	7,900
Jan '25	CyrusOne	ABS	687
Jan '25	Data4	Debt	3,300
Feb '25	Sabey	ABS	410
Feb '25	T5 Data Centers	Equity	5,500
Feb '25	Compass	ABS	885
Feb '25	Goodman	Equity	2,500
Mar '25	Sabey	ABS	435
Mar '25	Coreweave	Equity	1,570
Apr '25	TierPoint	ABS	500
May '25	Crusoe	Equity & debt	11,600
May '25	Coreweave	Debt	2,000
May '25	Princeton Digital Group	Debt	1,200
May '25	Stack	Debt	1,400
May '25	iM Critical	Debt	33
Total			**41,810**

Source: Company press releases, news reports, BofA Global Research
Note: ABS = asset-backed security

BofA GLOBAL RESEARCH

Exhibit 5: Selected US data center projects entering planning stages
Over $28bn of projected started planning in 1H25 to date

Planning stages

Month	Name	State	Value ($mn)
Jan '25	Vantage Meridian	Mississippi	3,900
Jan '25	Amazon Jeffersonville	Ohio	500
Jan '25	Hunter's Ridge Phase 1	Virginia	250
Feb '25	Tract Park	Virginia	500
Feb '25	Microsoft (SAT93)	Texas	350
Feb '25	Microsoft (SAT94)	Texas	350
Mar '25	Stream - Genesee	New York	6,300
Apr '25	Apple - Maiden	North Carolina	175
Apr '25	Atlas - Phase 1	Georgia	857
Apr '25	ForgeLight Ventures	Arkansas	1,000
Apr '25	Bolingbroke - Phase 1	Georgia	1,933
Apr '25	Meta - Bowling Green	Ohio	800
Apr '25	Oppidan	Texas	31
Apr '25	Red Wolf DCD	Kansas	12,000
	Total		**28,946**

Source: Company press releases, news reports, BofA Global Research

BofA GLOBAL RESEARCH

We find nearly $2bn of US data center construction starts in 2Q25. This includes a hyperscaler project from Meta and Iron Mountain. For a more comprehensive view of the US market, we look at data from the US Census. They estimate construction spending on US data centers rose 32% y/y to reach an annualized pace of $35bn in April.

Exhibit 6: Selected US data center construction starts
>$25bn of US data centers started construction in 1H25

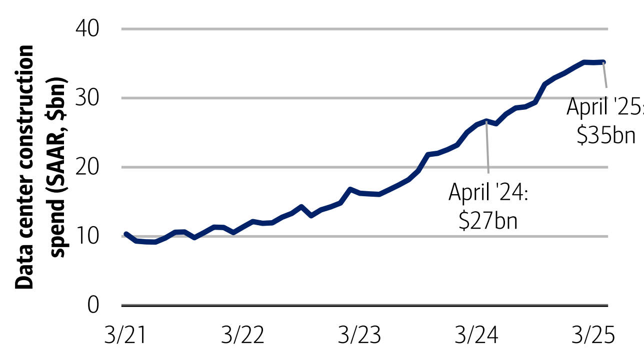
Construction starts

Month	Name	State	Value ($mn)
Jan '25	ATL11 Data Center	Georgia	1,500
Jan '25	AWS Ridgeland	Mississippi	1,400
Jan '25	Google Cedar Rapids	Iowa	600
Feb '25	Compass - Meridian	Mississippi	165
Feb '25	Corescale - Gainsville	Virginia	220
Feb '25	Aligned - Fredrick Co.	Maryland	300
Feb '25	Colovore	Nevada	82
Feb '25	Iron Mountain - Miami	Florida	66
Mar '25	Lancium Phase 2	Texas	4,100
Mar '25	Serverfarm HTX2, CTX2	Texas	274
Mar '25	CyrusOne - Fort Worth	Texas	200
Apr '25	Meta	Ohio	800
May '25	Iron Mountain	Virginia	1,000
May '25	Crusoe	Texas	15,000
	Total		**25,707**

Source: Company press releases, news reports, BofA Global Research

BofA GLOBAL RESEARCH

Exhibit 7: US data center construction spending ($bn, SAAR)
Data center spending rose 32% y/y to $35bn annualized pace in April



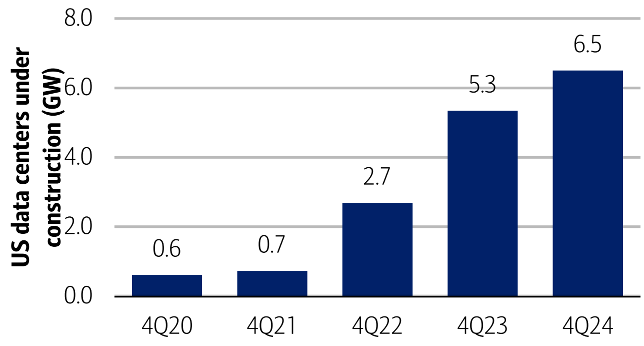
Source: US Census Bureau, BofA Global Research

BofA GLOBAL RESEARCH

Construction trends: heating up

US data centers under construction nearly doubled in 2023 to 5.3 GW, according to JLL. From there, it increased to nearly 6.5 GW in 2024. The pipeline (as measured in square footage) is also rising, but at a slower pace. However, this can be explained by rising power density.

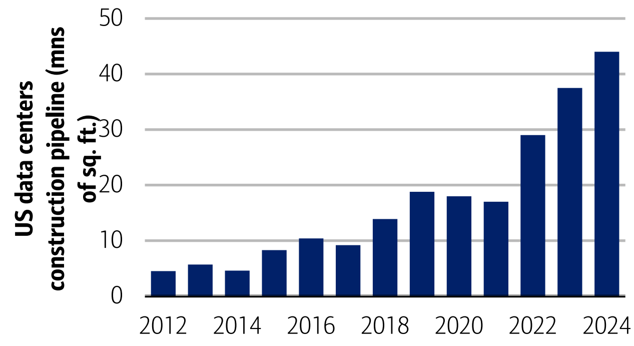
Exhibit 8: US data centers under construction (in gigawatts)
Data centers under construction increased 22% y/y in 2024



Source: JLL, BofA Global Research

BofA GLOBAL RESEARCH

Exhibit 9: US data center construction pipeline (millions of sq. ft.)
Square footage pipeline rising, but at a slower pace



Source: Newmark, BofA Global Research

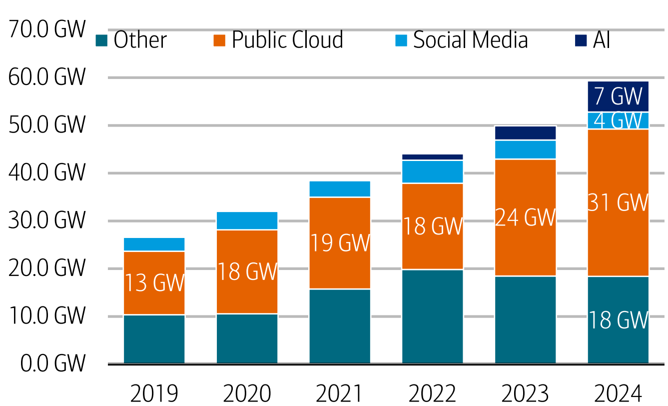
BofA GLOBAL RESEARCH

Public cloud is still most important component of demand

While AI uses of data center capacity is growing, public cloud remains the bulk of data center usage and has also been growing rapidly. As far as a share of total capacity goes, interestingly, social media has started to decline as AI has risen. Social media, public cloud, and AI are the three largest uses of global demand today. "Other" includes financial institutions, public sectors, software-as-a-service, and more.

An incremental 7GW of capacity in 2024 went to public cloud, as opposed to 3GW for artificial intelligence applications.

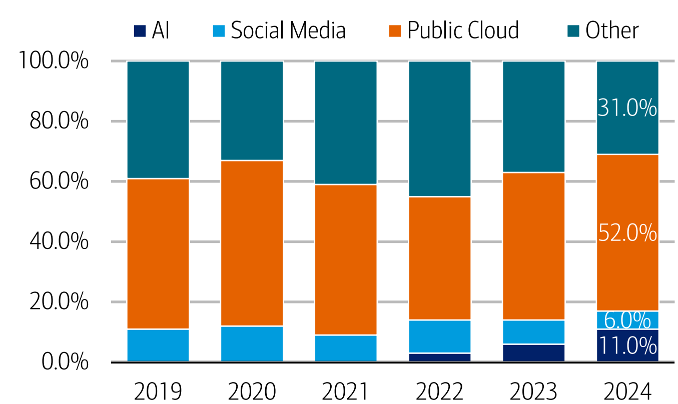
Exhibit 10: Data center capacity by most popular usage, total GW
Public cloud was the biggest source of growth in 2024



Source: DC Byte, BofA Global Research

BofA GLOBAL RESEARCH

Exhibit 11: Data center capacity by most popular usage, % of capacity
Public Cloud is still the largest share of the data center capacity



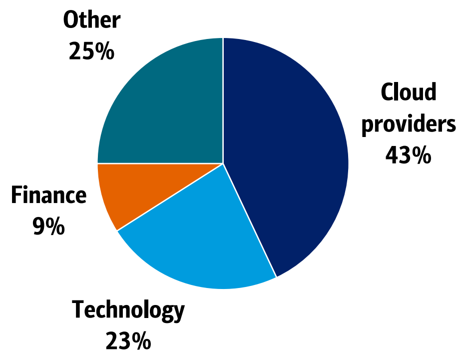
Source: DC Byte, BofA Global Research, JLL,

BofA GLOBAL RESEARCH

Colos had a great year in 2024

Investor perception has broadly been that colocations have been had a relatively difficult time getting capacity allocated from equipment and infrastructure providers. While the industry is clearly constrained, colocation capacity still grew 19% y/y in 2024. According to JLL, the largest customer for colocation is hyperscalers, with 43% of sales coming from hyperscalers. In 2024, AI represented 15% of data center workloads in colocation firms.

Exhibit 12: Colocation GW use by end customer, 2024
Cloud providers are the largest end customer of colocation firms

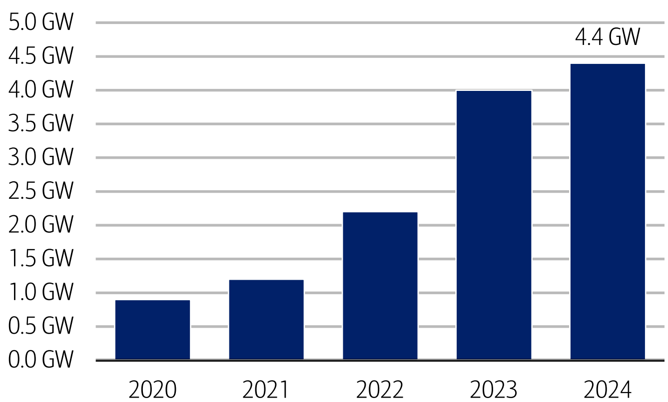


Source: BofA Global Research, JLL

BofA GLOBAL RESEARCH

Below, we show colocation absorption (a measure of incremental capacity filled) and vacancy rates. Colocation vacancy has dropped from 9.8% in 2020 to 2.6% at the end of 2024. Given vacancy is near 0%, absorption was almost entirely a reflection of pre-leasing activity.

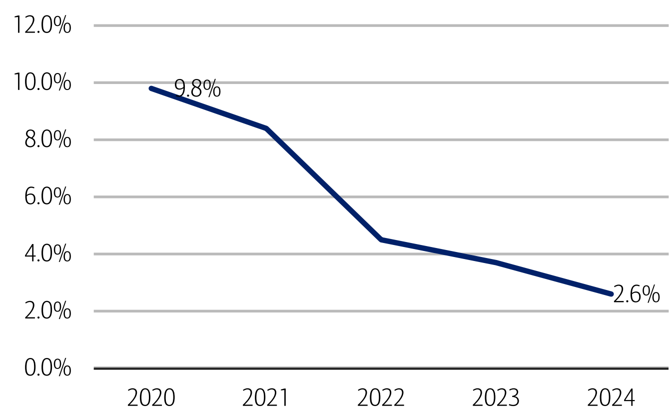
Exhibit 13: Colocation absorption, 2020-2024
Colocation absorption rose to 4.4GW in 2024...



Source: BofA Global Research, JLL

BofA GLOBAL RESEARCH

Exhibit 14: Colocation vacancy rates, 2002-2024
...as colocation vacancy dropped to 2.6%



Source: BofA Global Research, JLL

BofA GLOBAL RESEARCH

Data center infrastructure set to change

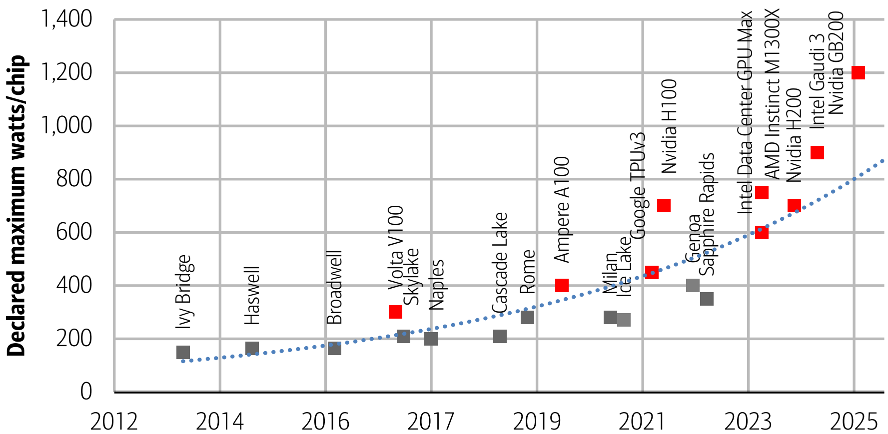
Data center infrastructure had remained relatively unchanged for nearly 20 years. However, the current wave of AI adoption accelerated two trends that require fundamental changes in design:

1. Watts per chip have been increasing with each generation of chip architecture. AI accelerator chips were a step-function increase.
2. AI's parallel processing requires more interconnections between chips. Putting more chips into a single rack becomes more effective.

1) Rising watts per chip

The power consumption per chip has increased 4x from Nvidia's first-generation Volta architecture to the current Blackwell. Many of the ways to increase computing performance require additional power. Put simply, supply voltage cannot scale down proportionally with node sizes. Putting more transistors on each chip requires more power. In addition, power consumption rises linearly with faster clock speeds.

Exhibit 15: Watts per chip is growing with GB200 setting a new high
Nvidia's Blackwell chips draw 4x the watts versus first-generation Volta chips



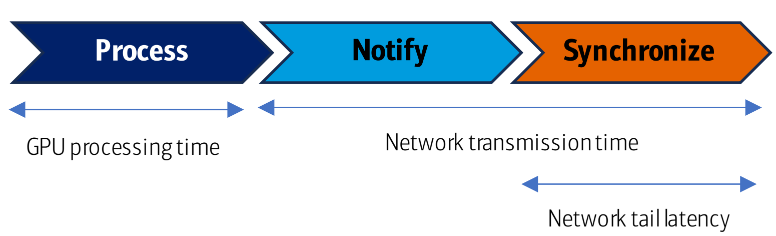
Source: Dell Technologies, Uptime Institute
Note: Red dots are GPUs, Grey dots are CPUs

BofA GLOBAL RESEARCH

2) Massively parallel processing

GPUs perform calculations in parallel. When thousands of GPUs work together in an AI cluster, if even one GPU lacks the data it needs, all other GPUs stall. This means that network latency delays can reduce overall performance significantly. Putting more GPUs within a single rack reduces the need for networking and high-speed interconnects.

Exhibit 15: Parallel processing time for GPUs in AI model training
Network transmission time can eat up 30% of job completion time



Source: BofA Global Research

BofA GLOBAL RESEARCH

These trends drive increase rack density…

In 2021, the average rack density was below 10 kilowatts (kW) per rack. A reference Hopper rack (H200 chips) would draw 35kW. A reference Blackwell rack (B200 chips) would draw 120kW. Based on released statistics from Nvidia, we estimate a reference Rubin rack would reach 600kW in a single rack.

Exhibit 16: Nvidia chip release schedule and reference rack architecture
Rack-level power density is rising exponentially

Release	GPU	TDP per chip (watts)	Rack	GPUs/rack	kW per rack
		Hopper architecture			
1H22	H100	700	DGX	8	35
2H24	H200	700	DGX		35
		Blackwell architecture			
2H24	B100	700	DGX	72	84
2H24	B200	1,000	DGX	72	120
2H25	GB200	1,350	Oberon	72	162
		Rubin architecture			
2026	Rubin	1,800	Oberon	144	600
2H27	Rubin Ultra	???	Kyber	144	???
		Feynman architecture			
2028	Feynman	???	Kyber	???	???

Source: Company reports, BofA Global Research

BofA GLOBAL RESEARCH

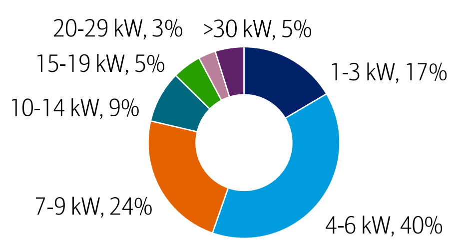
While we highlight Nvidia's roadmap, we note that other chip firms are following a similar trajectory. On 6/12, AMD announced a reference rack infrastructure for its Instinct MI350 GPUs. These racks feature up to 128 GPUs/rack, with each GPU drawing up to 1,400 watts, suggesting a 180+ kW rack density. AMD also announced its next generation Helios rack infrastructure, planned for release in 2026. This will feature 72 MI400 GPUs. In January 2025, Intel announced it was also developing a "system-level solution at rack scale" for its Gaudi data center accelerator chips.

All these firms are optimizing AI model performance across dimensions – chip-level, chip-to-chip bandwidth, and network throughput. This results in increased rack density, not as a goal, but as an outcome.

…which existing data centers are unprepared for

According to a 2024 Uptime Institute survey, only 5% of data centers have rack average densities above 30 kW. In other words, only 5% of data centers are designed to house even Hopper (H200) chips.

Exhibit 18: Global survey of over 700 data centers shows less than 5% are AI-ready
Only 5% of data centers have average rack density above 30 kW

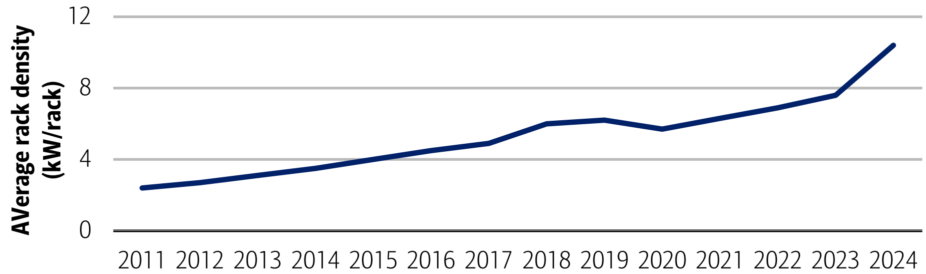


Source: Uptime Institute

BofA GLOBAL RESEARCH

Looking at average rack density over time, there was a clear inflection upward in 2024 as AI data centers began to go live. With rack densities 5-10x higher, even a small number of AI data centers drives the overall survey average up.

Exhibit 19: Average rack density rising, but remains low
Average rack density remains below 15kW per rack



Source: Uptime Institute

BofA GLOBAL RESEARCH

The existing power distribution architecture

Typically, data centers receive three-phase alternating current (AC) electricity at 13,800 volts or 34,500 volts. Over a series of power distribution units, transformers, busways, and uninterruptible power supplies (UPS), this electricity is stepped down and converted to 48-volt direct current (DC) which powers IT servers and other equipment.

Raising the voltage to save on wiring

As GPUs for AI applications have increasingly higher computational needs and demand more electricity, existing power distribution systems will struggle to cope. Large data center operators are proposing changes to existing power distribution systems.

In October 2024, Microsoft and Meta, as part of the Open Compute Project, announced a reference rack architecture called Mt. Diablo. This uses 400-volt direct current (DC), which is significantly higher than the current 48-volt.

On 5/28, Nvidia announced that it would develop a new power infrastructure with a 800-volt direct current (DC) architecture to deliver power requirements of 1+ megawatt server racks. The company has plans to deploy it by 2027.

Electrical power (measured in watts) can be broken down into voltage (measured in volts) and current (measured in amps). The carrying capacity for wiring is determined by the current (amps). Thus, higher voltage can carry more electrical power using the same diameter wire. This reduces the amount of copper needed within the rack. Compared to a 208-volt system, a 400-volt system would reduce the copper wire weight by 52% per Schneider Electric.

Industry participants noted that the Mt. Diablo/Open Compute Project proposal would allow equipment to be installed by electricians with low voltage certifications (e.g., below 600 volts). Nvidia's proposal would require installation by electricians with medium voltage certification, which would limit the potential workforce of installers.

Moving power equipment outside the rack to save space

Power supply units (PSUs), which convert AC to DC, take up valuable space within the rack. At higher voltage, these will take up even more space. This is why Nvidia and the Open Compute Project are proposing moving electrical equipment to a "side car" next to the rack containing servers.

Implications for incumbent equipment vendors

Given Eaton, Vertiv, nVent, and other IT infrastructure players specifically manufacture on/around the rack, this has raised investor questions on whether this will disrupt market share. We argue the data center industry prioritizes uptime/reliability, which historically has benefited incumbents. Service capabilities are another barrier to entry, particularly for operators adopting new equipment.

However, the proposed direct current architecture would result in simplified uninterruptible power supplies (UPS). The hypothetical cost of a direct current UPS is

10-20% lower than an alternating current UPS. However, we see this being largely offset by additional costs for higher-voltage switchgear and rectifiers (AC to DC convertors). Net-net, industry participants expect content per megawatt for electrical equipment to remain relatively similar.

More specifically, Nvidia's announcement regarding its new power infrastructure highlights Eaton, Schneider Electric, and Vertiv as partners on power systems, suggesting the incumbent power distribution players will retain an advantage.

Exhibit 19: Oberon rack
Servers placed horizontally; power equipment located within the rack

Source: Company presentations

BofA GLOBAL RESEARCH

Exhibit 20: Kyber rack
Servers placed vertically, power equipment located in a side car next to the rack

Source: Company presentations

BofA GLOBAL RESEARCH

Exhibit 21: Mt. Diablo rack
Power equipment located in a side car next the rack

Source: Open Compute Project

BofA GLOBAL RESEARCH

Details on the increase in Rubin Ultra infrastructure

Nvidia's Rubin Ultra GPU and its NVL576 Kyber racks were initially unveiled as a mockup in March 2025. Rubin Ultra will follow Rubin and Blackwell chips. The Rubin Ultra is intended to ship in 2H27. The current Blackwell B200 server racks can use up to ~120kW per rack. The first Vera Rubin rack, to launch in the 2H26, (the name for the combination of the Vera CPU and Rubin GPU) will use the same infrastructure as Grace Blackwell (the Grace CPU and Blackwell GPU combination).

However, the next iteration of Rubin – Rubin Ultra – will 2x the number of GPUs per rack. The single rack solution, dubbed Kyber, will be able to handle 600kW. Each rack will consist of four "pods" with 18 blades in each pod.

Thermal architecture: more change ahead

What is the COOLERCHIPS program?

The US Department of Energy's Advanced Research Projects Agency-Energy (ARPA-E) launched the COOLERCHIPS program in September 2022. COOLERCHIPS, which is an acronym for Cooling Operations Optimized for Leaps in Energy, Reliability, and Carbon Hyper-Efficiency for Information Processing Systems, provides federal funding to research with the explicit goal being to reduce the total cooling energy expenditure to less than 5% of a typical data center's IT load at any time and any location in the US for high-density compute systems (e.g., artificial intelligence workloads).

ARPA-E has granted over $83mn across 19 projects since its launch in 2022. We focus on an Nvidia-led consortium's project, which is likely to be used in conjunction with the Feynman chip architecture in 2028-29.

Vertiv, Nvidia partnership exploring combo of D2C and immersion cooling

Vertiv is part of a consortium led by Nvidia that received $5mn grant from the US in 2023 for COOLERCHIPS initiative. The consortium is exploring hybrid designs including partial immersion cooling. This program is working to develop a modular data center with a cooling system combining direct-to-chip, pumped two-phase and single-phase immersion in a rack manifold with built-in pumps and a liquid-vapor separator. According to program files, the design would cool chips with a two-phase cold plate with the rest of the server components are submerged in an immersion sled. The cooling design would incorporate green refrigerants for two-phase cooling, and dielectric fluid for immersion.

Immersion cooling is not the preferred technology by industry participants

Immersion cooling poses several challenges to data center operators. Before changing components in an immersion cooling tank, all IT components must be powered down. The tanks add considerable weight, which makes retrofit more difficult. The most used dielectric fluids for immersion cooling are fluorocarbons, such as perfluoroalkyl substances (PFAS). These introduce potential environmental and human health concerns. For these reasons, industry participants we spoke with would prefer cooling architectures that avoid immersion sleds.

Modest success on cooling, but still in early stages

As of March 2025, some of the projects that COOLERCHIPS has invested in are showing successful first prototypes on real servers. This includes teams achieving less than 5% energy used for cooling for a rack >120kW, meaning 95% or more of the energy is available for power and compute and a 90% reduction in cooling energy compared to today's energy usage (which is ~40% energy usage for cooling). Industry participants are still debating potential winners and losers for next-generation chip cooling between single-phase cooling, two-phase cooling, immersion cooling, and direct-to-chip cooling.

Reducing thermal resistance for a 1kW chip

Currently, in high-density data centers, the room is 40 degrees lower than the chip's heat (e.g., 70 degrees Celsius for the chip; 30 degrees Celsius for the room). This is in order to effectively transfer the heat from the chip. If COOLERCHIPS program designs are successful, for a 1KW chip (in line with Blackwell Ultra density), the temperature difference could be reduced to 10 degrees Celsius. An operator would then be able to run chips at 40 degrees Celsius, which would yield more efficient chips. Running the building at 60 degrees Celsius would clearly not be possible in our view unless the infrastructure was redesigned to not need human presence in the data center.

Potential shifts to cooling infrastructure

According to De Bock, shifts to cooling technology for servers with >3kW density could include better flow manipulation of cold plates to transfer heat more efficiently, shifting the materials in the cold plate (e.g., silicon to replace copper/aluminum), and increasing the length of replacement cycle for immersion cooling fluids (which currently need to be replaced every ~6 months).

Rise of modular data centers

Modular data centers are prefabricated-portable data centers that are built in a factory and then shipped to a specific location. These can be built in 2-3 months, as opposed to a regular data center, which can take a little over a year to construct (per Delta, 400 days). They are also built on a factory floor, which leads to stronger labor pool availability and leverage on production processes.

Pat Lynch of CBRE, a data center operating real estate company, stated in an industry publication that "Modular prefab data centers are a good fit for clients needing a small amount of additional IT capacity," meaning less than 2MW or for temporary capacity needs.

Due to their prefabricated nature, modular data centers are faster to market but standardized. As a result, we believe they are likely to see more adoption amongst colocation and enterprise utilization, rather than hyperscalers who prefer to customize their infrastructure.

Eaton's EMEA data center segment leader, Juan Colina, has written: "Modularity data centers… decrease total cost of ownership and increase quality and consistency across sites." According to industry websites, Huawei is the leader in the global prefabricated modular data center market, followed by Schneider and Vertiv.

DeepSunk Part III: Training power still rising; inference is more energy intensive

In December, Chinese startup DeepSeek released an artificial intelligence (AI) model, DeepSeek-V3. V3's performance across several benchmark tests is comparable to OpenAI's GPT-4. This sparked investor concern that algorithmic/software advancements would result in less spending on AI-related infrastructure & power generation.

Six months after DeepSeek's release, we find that:

1. Training energy usage is still rising. DeepSeek's V3 training power usage was still above the power usage curve, per research estimates.

2. Industry participants think that data center designs for training and inference are tending to more similar versus different

3. Inference estimates are for acceleration – which displays Jevons paradox in action. While DeepSeek may have accelerated cost declines, the shape of the cost curve remains the same. As far as inference power consumption (not cost), DeepSeek's use of "chain of thought" methodology means more power consumption per inference task.

We look at Epoch AI's estimates for total power required to train AI models. Total power draw is the power draw of the hardware used to train the model, measured in watts. The cost has been rising exponentially with the recent Llama 3.1 requiring 22,600kW to train. Despite efficiency gains, the cost is still rising exponentially.

Per Epoch AI, power to train models is doubling every year. Training compute needs have grown even faster (around 4x a year), which is offset partially by hardware efficiency gains, adoption of lower precision formats, and longer training runs. The increase to train models has been driven by GPU count. While power per GPU is growing, according to Epoch AI this is only at the rate of a few percentage points a year.

Training power draw continues to accelerate

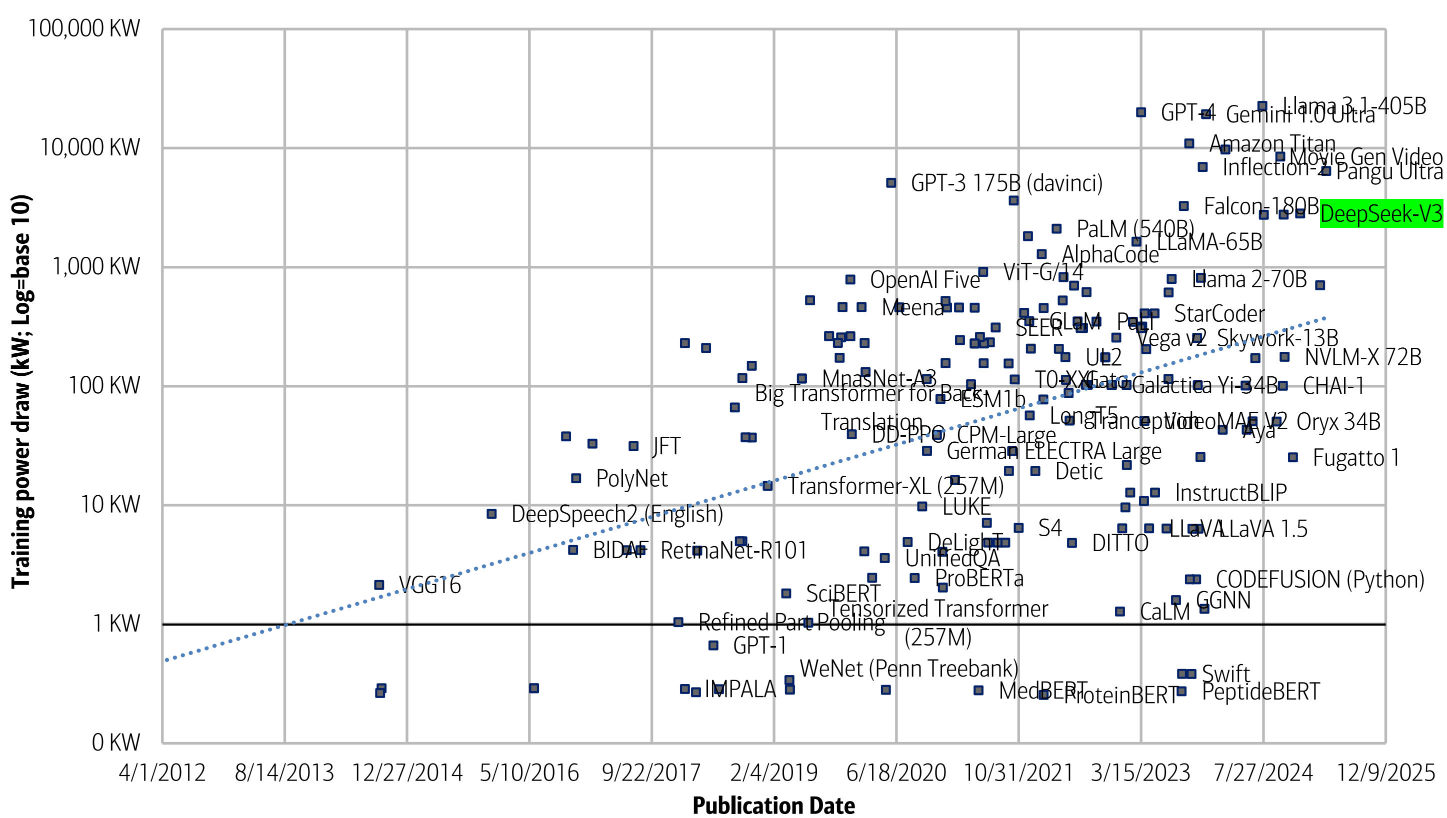
Training power draw is not going down, despite changes to the cost curve, and continues to rise. Power draw is defined as the electricity used by a data center for servers simultaneous working on an individual model. A few key points:

1) DeepSeek V-3 (highlighted in green) is estimated by Epoch AI, an AI research firm, to have used 2,818 kW to train. This is above the curve and likely reflects the large amount of datasets and parameters used to develop the computational capabilities of DeepSeek, even at a lower cost.

2) Open-source models released after DeepSeek, including LG's EXAONE Deep 32B (released 3/16/25) and Huawei's Pangu Ultra (released 4/10/25) remain above the exponential curve of training power draw. Pangu Ultra is estimated to have used 6,426 kW at once, more than 2x DeepSeek. LG's model is estimated to have taken 703 kW.

In July 2024, Anthropic CEO Dario Amodei stated that while current models like ChatGPT-4o only cost about $100mn, that the cost of training these models will increase to $10bn and beyond quickly. Amodei stated, "Right now, $100mn. There are models in training today that are more like a billion... I think if we go to ten or a hundred billion, and I think that will happen in 2025, 2026, maybe 2027, and the algorithmic improvements continue a pace, and the chip improvements continue a pace, then I think there is in my mind a good chance that by that time we'll be able to get models that are better than most humans at most things."

Exhibit 23: Estimated training power draw for AI models, 1/2012-4/2025
Power needed to train AI has increased exponentially



Source: Epoch AI (Creative Commons License); BofA Global Research

BofA GLOBAL RESEARCH

Training has gotten longer and more costly. Training state-of-the-art models, such as Gemini's 1.0 Ultra (released in 2023) took around 100 days to train. In contrast, AlexNet in 2012 (which utilized GPUs) was trained in 5-6 days. This reflects increasingly large datasets despite increasingly advanced hardware. The training compute of notable AI models has been increasing exponentially, doubling roughly every 5 months and particularly in the last five years. Per Stanford AI, cutting edge AI models require colossal amounts of data, compute power, and financial resources.

More compute remains our base case

Our base case is that AI-related compute, and hence power consumption, will continue to increase. Companies will continue to reassess the composition of their capex. Jon Gray, COO of Blackstone (owner of $80bn of data center assets) noted on January 30th: "Maybe there's a little less training that's done as a result of the less intensity. But at the same time, there's more inference, maybe there's more cloud, maybe there's more to do with enterprise. As usage goes up significantly, there's still a vital need for data centers, the form of that use may change… power usage… will continue because our lives are migrating online…there'll be even more questions coming even if the amount of power used on an individual question goes down."

More efficient, but still on the cost curve

Dario Amodei, the co-founder and CEO of Anthropic (builder of the Claude large language model), noted that while DeepSeek "produced a model close to the performance of US models 7-10 months older, for a good deal less cost (but not anywhere near the ratios people have suggested)" fundamentally DeepSeek is "not a unique breakthrough or something that fundamentally changes the economics of LLMs; it's an expected point on an ongoing cost reduction curve". For example, the cost of GPT-4 output tokens has fallen ~80% in less than a year since release.

Inference power forecast to accelerate

Chain of thought requires more power

DeepSeek is a reasoning model, using "chain of thought" methodology that breaks the task into parts. This allows the model to do better on tasks related to logic, math, and pattern finding. This requires greater computational intensity during inference. According to an MIT Technology Review article published on Jan 31st, DeepSeek used 87% more energy than Meta's Llama 3, a model with the same number of parameters, for a similar inquiry as it generates much longer responses. This is particularly relevant as compute is expected to shift more towards inference versus training over time. (O'Donnell, James. "DeepSeek might not be such good news for energy after all.")

Videos are even more energy intensive

Different inferencing tasks require different levels of computing power. An MIT Technology Review study suggests that video generation is by far the most intensive by a factor of ~1000x (versus a large text model).

MIT Technology Review measured energy used for inference for text generation using Meta's Llama models which are open source. The smallest model in the cohort, Llama 3.1 8B, had 8 billion parameters (adjustable inputs in the model). The largest model, Llama 3.1 405B, had 405 billion parameters. Notably, DeepSeek has over 600 billion parameters. More parameters drive more energy per response. Prompts themselves also can drive higher energy needs, with more complicated requests driving more energy usage. According to Microsoft researchers in 2024, doubling the energy required by the GPUs running the model is roughly equivalent to the total operation's energy needs, including the cooling and the CPU's as well as other infrastructure.

Image generation models require a different architecture called diffusion, which transformers noise/words into images. The energy requirement is not dependent on the prompt, but rather by the size of the model, image resolution, and the number of steps the process takes which can generate higher quality. Per MIT, image generation typically has fewer parameters than a large text model, which drives the lower energy usage.

MIT used a Chinese AI startup video model called CogVideoX to estimate video generation energy. This went from 109,000 joules (in the August 2024 model) for eight frames per second to 3.4mn joules in November 2024 for 16 frames per second for a five-second video. MIT notes that these are lower-quality videos with 16 frames per second similar to a silent-era (1920's) film quality, whereas closed-loop models such as OpenAI's Sora have much higher quality and therefore likely require more energy.

Exhibit 23: Inference tasks ranked by energy usage, 2025

Video Generation requires 1000x more energy than large scale text models

Inference	Joules per GPU	Total joules	W per hour
Text Model - Small	57	114	0.016
Image Generation - Small	1,141	2,282	0.317
Image Generation - Large	2,282	4,564	0.634
Text Model - Large	3,353	6,706	0.931
Video Generation	3,400,000	6,800,000	944.445

Source: MIT Technology Review, BofA Global Research

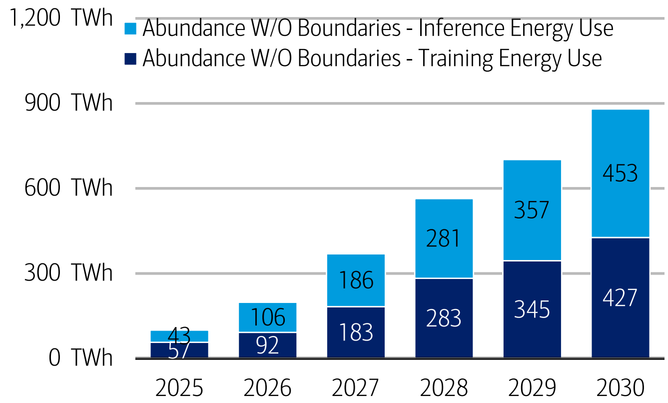
BofA GLOBAL RESEARCH

According to the Stanford 2025 AI Index report, inference energy efficiency has been improving at a 40% per annual rate and hardware costs have been declining (30)% annually. Depending on the usage, inference costs in 2024 fell from 9 to 900 times relative to 2022. Despite this, Schneider forecasts that Generative AI inference will become the primary driver of AI electricity consumption within the AI sector by 2027-20228. This reflects increasing deployment of LLMs, largely being integrated into industrial applications. As usage scales, this more than offsets any hardware efficiency gains.

In Schneider Electric's Sustainable AI whitepaper from December 2024, the company outlines a series of scenarios, including Sustainable AI and "Abundance Without Boundaries." Abundance Without Boundaries assumes a "Jevons Paradox" scenario, which assumes rapid expansion of AI workloads and increased energy consumption despite some improvements in hardware efficiency and algorithmic optimization. The "Abundance Without Boundaries" scenario assumes energy demands for data centers are met and adoption increases. It assumes modest hardware efficiencies, with joules per GenAI token decreasing at a rate of 1% annually from 2025-2030, but offset by an acceleration in adoption.

The Sustainable AI case study assumes slightly more modest adoptions (but still at a 20% CAGR for industry use cases and a 43% CAGR for consumer users) but with a more rapid hardware efficiencies. Sustainable AI assumes that the joules per GenAI token decrease at a rate of 5% annually every year until 2030, but that an acceleration in adoption more than offsets the improvement in inference. In this scenario, inference outgrows training even more given training is more likely to lend itself to efficiencies and adoption has a larger impact on the inference power demand. This blends to a total AI energy usage CAGR of 44%.

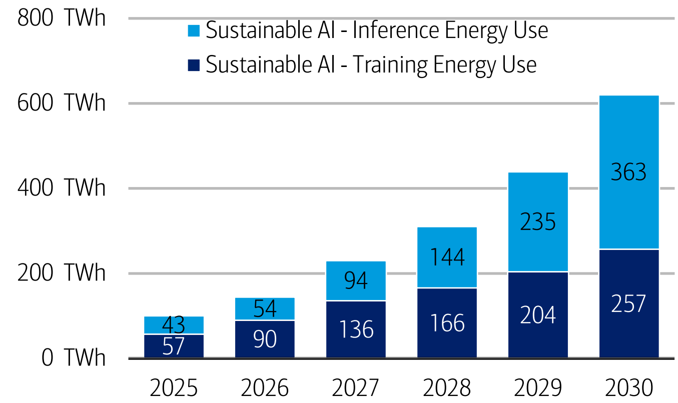
Exhibit 25: Abundance training vs. inference, TWh, 2025-2030
Schneider's abundance without boundaries scenario forecasts 54% AI CAGR



Source: Schneider Energy Whitepaper, BofA Global Research

BofA GLOBAL RESEARCH

Exhibit 26: Sustainable AI training vs. inference, TWh, 2025-2030
Schneider's sustainable scenario forecasts 44% AI CAGR



Source: Schneider Energy Whitepaper, BofA Global Research

BofA GLOBAL RESEARCH

Below we showcase key assumptions in each of the models made by Schneider for abundance without boundaries and sustainable AI.

Exhibit 26: Key input assumptions in "Abundance without borders" and Sustainable AI scenarios
Schneider assumes materially increased efficiency in the sustainable AI scenario

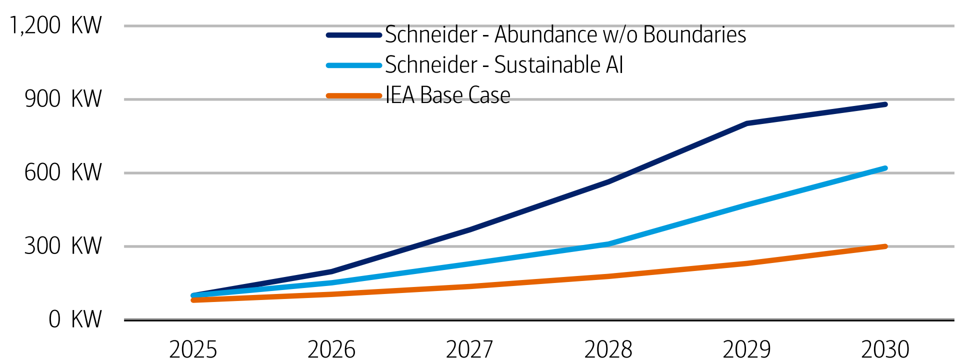
	2025-2030 CAGR	
	Sustainable AI	**Abundance without boundaries**
AI Growth	44.0%	54.5%
Training Growth	37.1%	54.0%
Inference Growth	72.1%	78.4%
Industry Users	20.1%	29.7%
Consumer Users	43.1%	49.6%
Joule per GenAI token	-4.7%	-1.0%
Tokens per GenAI Output	2.5%	4.1%
Joule per Output	-2.3%	3.0%

Source: Schneider Energy Whitepaper, BofA Global Research

BofA GLOBAL RESEARCH

Below, we show the AI scenarios we discuss in this report. We note that Schneider's assumption for AI adoption, even at its Sustainable AI scenario, is more aggressive than the IEA assumption.

Exhibit 28: Schneider versus IEA assumptions for energy usage, 2025-2030
Schneider vs. IEA demand, 2025-2030



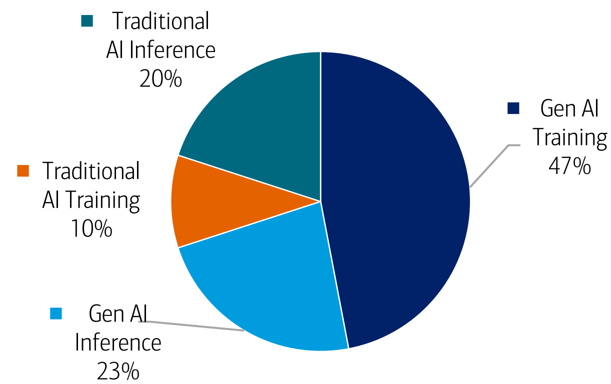
Source: Schneider Energy Whitepaper, BofA Global Research

BofA GLOBAL RESEARCH

Generative AI inference is the key driver of growth

Below we show drivers of electricity usage for artificial intelligence applications in 2025 and 2030. We use Schneider's "Sustainable AI" scenario.

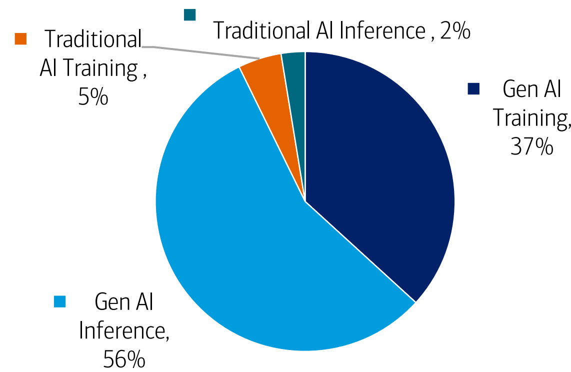
Exhibit 29: Split of AI energy usage, % of total TWh, 2025
Generative AI training is the largest AI data center use of energy in 2025



Source: Schneider Electric Whitepaper, BofA Global Research

BofA GLOBAL RESEARCH

Exhibit 30: Split of AI energy usage, % of total TWh, 2030
Generative AI inference is the largest AI data center use of energy by 2030



Source: Schneider Electric Whitepaper, BofA Global Research

BofA GLOBAL RESEARCH

What about efficiency gains in inference?

Per industry sources (e.g., April 2023 article Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning - ScienceDirect), efficiency for AI inference is improving. However, this largely reflects efficiency in scaled models rather than cutting edge models, which continue to increase in power usage. For cutting edge models, efficiency is not canceling out performance increases.

In Schneider's Sustainable AI scenario, which assumes rapid hardware efficiencies (e.g., a 5% CAGR for reduction in energy per GenAI output over 2025-2030), inference energy usage still grows at a 53% CAGR over this timeframe (vs. 60% in the abundance without boundaries scenario). This is a 7% reduction in the CAGR. The largest reduction in the CAGR is in training, which is forecast to go from a 50% CAGR to 35%. This reflects more hardware efficiencies reflected in training, with adoption driving more of an acceleration. Per Schneider: it's likely that a popular model (e.g., ChatGPT) requires many more times the quantity racks for inference as it did for training since their queries are now in the millions per day.

Schneider also notes that as data becomes scarce, a trade-off between data efficiency and energy efficiency may emerge, with more data-efficient models potentially requiring greater computational resource.

Data centers use 1-2% of total electricity today…

Data centers consume approximately 1-2% of global electricity production. Estimates range from 240-340 terawatt hours (TWh) in 2022 according to the International Energy Agency, 409 TWh in 2023 according to a meta-analysis[1] of 46 forecasts, and 499 TWh in 2023 per Schneider.[2]

In the US, Berkeley Lab estimated data center electricity usage at 176 TWh in 2023.[3] Using a separate methodology, McKinsey estimated it at 147 TWh in 2023Load growth forecast for data center TWh.

[1] Mytton, D., & Ashtine, M. (2022). Sources of data center energy estimates: A comprehensive review. Joule.
[2] Avelar, V., Donovan, P., Lin, P., Torell, W., Torres Arango, M. (2023). The AI Disruption: Challenges and Guidance for Data Center Design. White Paper 110.
[3] Shehabi, A., Smith, S.J., Hubbard, A., Newkirk, A., Lei, N., Siddik, M.A.B., Holecek, B., Koomey, J., Masanet, E., Sartor, D. 2024. 2024 United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, Berkeley, California.

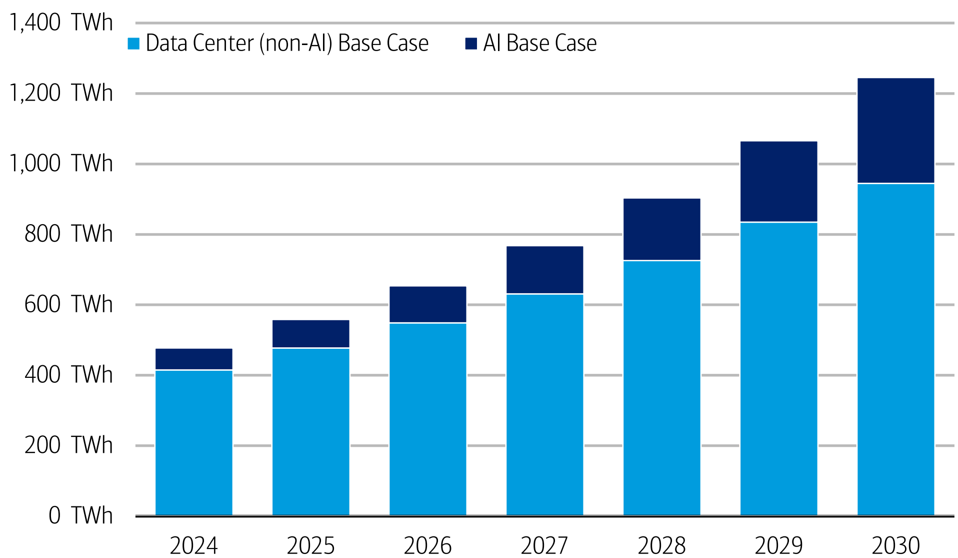
...and likely to grow at mid-teens CAGR globally

Forecasting data center electricity consumption has been a prevalent topic among academics and industry research. Forecast methodologies range from bottom-up (e.g., equipment shipments), top-down (e.g., data consumption), and extrapolation (e.g., energy intensity per chip).

We highlight the International Energy Agency (IEA) forecast, which is global and was published in April 2025. The IEA forecast is for data centers to grow at an overall 15% CAGR from 2043-2030, with growth accelerating in 2025-2030 from the past two years. The growth is broad-based, with artificial intelligence forecast to grow at a 30% CAGR from 2025-2030. Conventional servers are forecast to grow at 9% a year. All other data center-related demand, including conventional servers, other IT equipment, and cooling/other infrastructure, are forecast to grow at a 11% CAGR.

By 2030, the IEA forecasts data centers will consume 945 TWh of electricity, roughly equivalent to 3% of total electricity globally. Roughly 32% of this usage will be driven by AI applications.

Exhibit 31: Data center global energy consumption forecasts, 2024A-2030E
The IEA forecasts 15% data center energy consumption CAGR in TWh



Source: International Energy Agency, BofA Global Research

BofA GLOBAL RESEARCH

Power remains a limitation on growth

Power constraints to data centers have been a focus for industry participants and investors alike. This has led some data center operators and developers to bring their own power. In practice, this looks like agreements for nuclear, natural gas, and coal power plants.

...but short-term bridges are possible

xAI: bridging the gap with gas-powered turbines

xAI sourced 100MW of power in 100 days for its Memphis AI data center. The Memphis site powers xAI's Chatbot, Grok. The site went live in July 2024. It reported an initial capacity of 150MW and an 8MW grid connection with local power operator Tennessee Valley Authority.

xAI ramped quickly as they obtained 35 gas-powered turbines. We estimate 0.5 gigawatt of total capacity spread across the 35 turbines, or roughly 14 megawatt per turbine. xAI relied on legislation that allowed temporary turbines in one location for less than a year to operate without air permits. In January, xAI applied to the county health department for permits to run 15 turbines. These will be removed when a second electric substation is completed and connected to the grid. Afterwords, these gas-power turbines will be used for backup power.

Hyperscalers: increasing renewable capacity

Hyperscalers seem in very early stages of investing in renewable capacity for data centers. We take a look at Google, Amazon, Microsoft and Meta's most recent sustainability reports.

- **Google (July 2024):** Google signed contracts to purchase 4 GW of clean energy generation capacity in 2023. The company has maintained a global average of approximately 64% carbon-free-energy across all of its data center sites, inclusive of those operated by third parties. The company was able to maintain this 64% carbon-free energy target despite increasing its electricity load increasing 17% y/y.

- **Amazon (July 2024):** Amazon had 513 renewable energy projects announced since January 2024 contributing to 28MW of capacity. Amazon's renewable energy procured globally plus utility energy supply means that 100% of electricity is matched with renewable energy sources. In Oregon, Amazon works with its local utility to allow AWS to choose the energy supply to power its data centers, including renewable energy.

- **Microsoft (July 2024):** Microsoft increased contracted portfolio of renewable energy assets to >19.8 GW in 2023. The company does not directly reference using renewables to power data centers in the report.

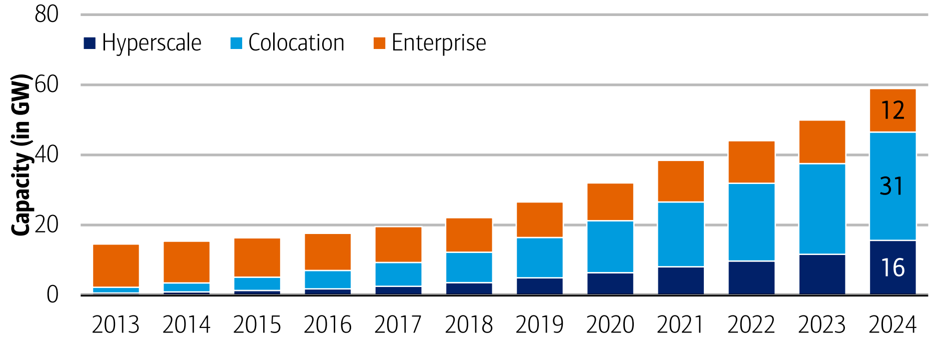
- **Meta (August 2024):** Meta highlights that it contracts 11,700 MW of renewable energy. The company matched 100% of its electricity use in 2023 with renewable energy by adding new wind and solar projects to local grids, including where its data centers are located. The company has more than 6,700 MW of renewable energy online. 100% of the company's data center buildings have earned at minimum the LEED Gold certification, for a total of 42 LEED Gold data center buildings totaling nearly 28mn square feet.
  - The company's data centers, on average, exhibited a power usage effectiveness (PUE) of 1.08 and water usage effectiveness of 0.18 in 2023.
  - META's newest AI-optimized data centers will feature dry-cooling technology, which uses air as the cooling medium.

Slicing and dicing the data center market

A look at the last eleven years of growth

In 2024, colocation firms drove 40% of the incremental load additions. This reflects a 19% y/y growth rate. Hyperscalers grew high-20% y/y. Enterprise was flat, consistent since 2021 levels. In total, an incremental 9GW of capacity was added in 2024. Industry sources estimate an incremental 10GW projected to break ground in 2025.

Exhibit 32: Global data center capacity, 2013-2024
Hyperscale and colocation have driven nearly all the growth since 2013



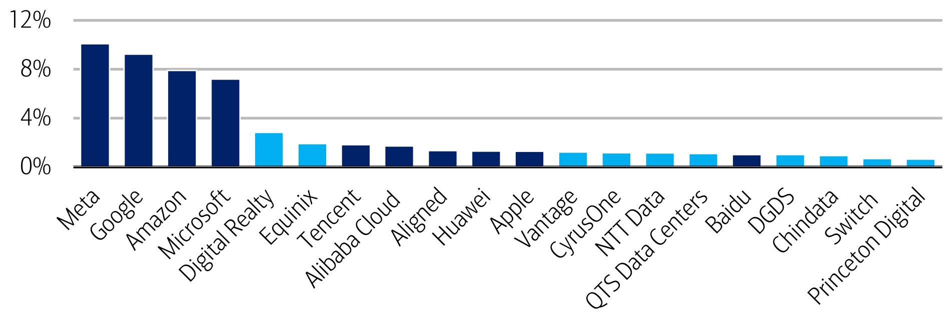
Source: BofA Global Research, Schneider Electric, IEA, McKinsey's Investing in the Rising Data Center Economy

BofA GLOBAL RESEARCH

We find that small colocation companies (e.g., <10 data centers) collectively comprise a meaningful portion of data center space (20-25%). These firms typically own smaller sites (e.g., <20 megawatts) outside major markets. This results in a downward bias from most bottom-up estimates of the data center market. For example, market estimates from commercial real estate brokers tend to capture only the top 15-25 US markets.

The International Energy Agency (IEA) estimates megawatt capacity for the leading hyperscaler and colocation providers as of 2024. Meta, Google, Amazon, and Microsoft have a combined share of 34% of global capacity.

Exhibit 33: Global data center capacity by top players, 2024
Meta, Google, Amazon, and Microsoft combined represent 34% of the global market



Source: BofA Global Research, International Energy Agency

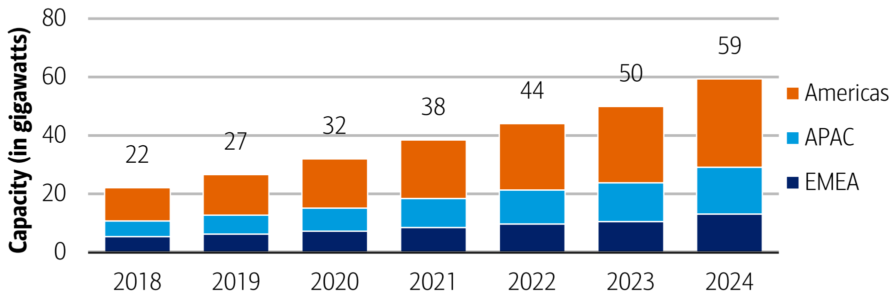
BofA GLOBAL RESEARCH

Data center growth is global phenomenon

We estimate data centers (measured by electrical capacity) grew at an 18% CAGR over 2018-24. We find only relatively modest differences in growth by region. EMEA has been a relative laggard (16% CAGR), Asia Pac a touch better (20%), and the Americas region right in line (18%).

In 2024, we estimate that EMEA had the fastest growth off the lowest base (+25%), with APAC growing +20% y/y and Americas growing +16% y/y.

Exhibit 34: Global data center capacity, 2018-24
Electrical capacity has grown at an 18% CAGR

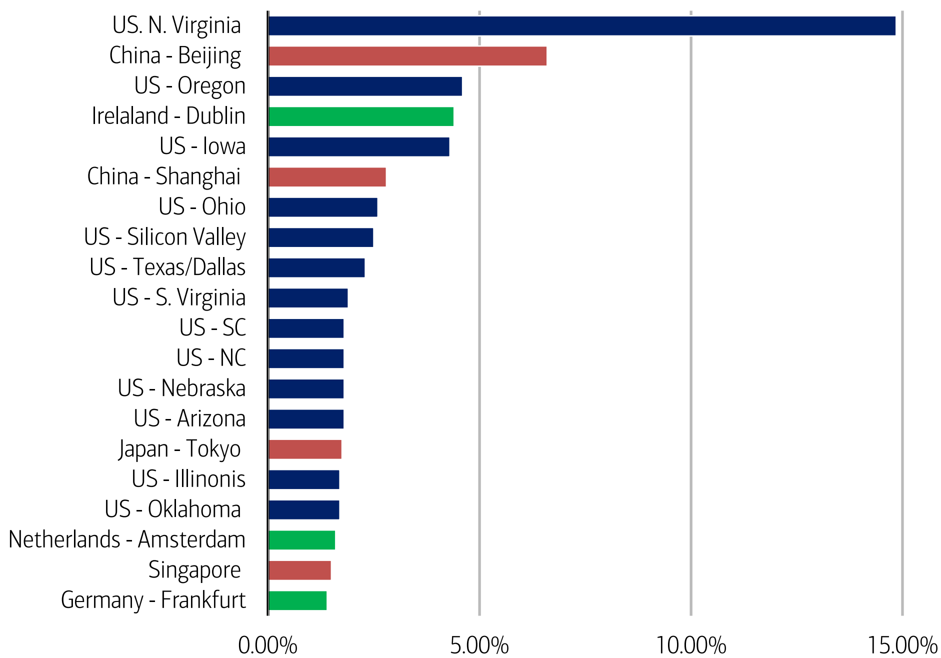


Source: BofA Global Research, Schneider Electric, JLL, IEA

BofA GLOBAL RESEARCH

Hyperscaler capacity is largely concentrated in the same key regions globally. Below, we look at the top 20 locations for hyperscaler data center capacity globally. Navy represents North America, Red represents APAC, and Green represents EMEA. The largest single hyperscaler location globally is US Northern Virginia, which represents almost 15% of global capacity for hyperscalers. The second largest capacity is in Beijing, with ~7% capacity.

Exhibit 35: Hyperscale data center capacity by country/region, ranked by top 20 largest regions
US Northern Virginia has the largest concentration of data center capacity



Source: Synergy Research, BofA Global Research

BofA GLOBAL RESEARCH

Price objective basis & risk

GE Vernova (GEV, C-1-7, $487.88)

We base our $550 price objective on a 26x EV/EBITDA multiple of our 2026E adjusted EBITDA estimate. Our target multiple is a premium to the 15x peer average on 2025 estimates. We argue a premium multiple is warranted given above-peer earnings growth and margin trajectory.

Risks to our price objective are 1) changes to government incentives for wind turbines, 2) execution risks on targeted $500mn cost-out by 2028, 3) secular decline in natural gas turbine demand, and 4) cost overruns on services & equipment contracts.