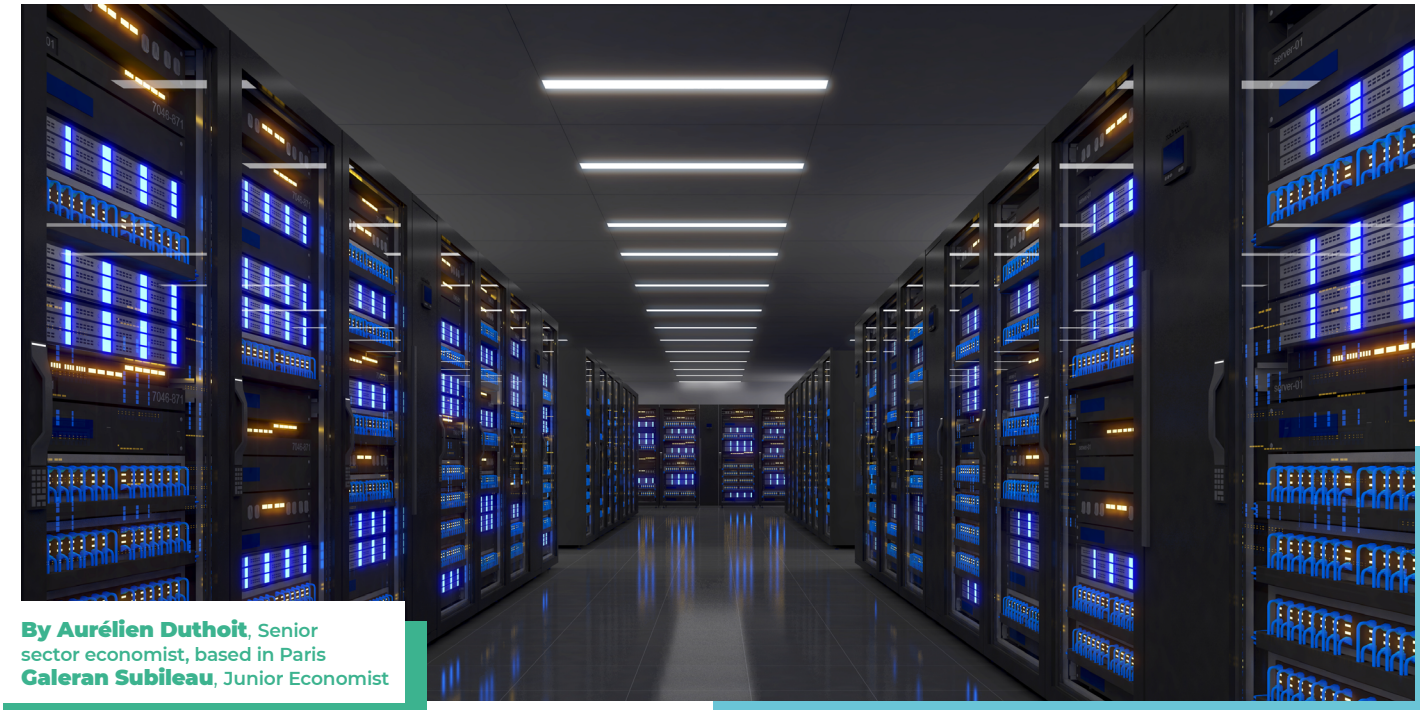
COFACE ECONOMIC PUBLICATIONS

NOVEMBER 2025

FOCUS



By Aurélien Duthoit, Senior sector economist, based in Paris
Galeran Subileau, Junior Economist

Data centers in the AI Age: Stakes, Limits and Risks of a Trillion-Dollar Gamble

EXECUTIVE SUMMARY

Can the trillion-dollar AI infrastructure boom keep outrunning the physical and economic systems that must sustain it, or is it growing too much, too fast? Unlike past digital waves that rode exponential advances in semiconductor performance and compounding network effects, data centers are constrained by land and water, specialized labor and equipment as well as power and grid capacity. Global capacities are expected to reach about 230 GW by 2030 - a 2.3x increase from 2024, the equivalent of 5% of today's global power generation capacity and enough to use as much electricity as the world's fourth-largest economy (Japan, 945 TWh).

With the US on track to deliver over half of global additions by doubling down on fewer than ten hubs, friction is inevitable and delays will materialize. If grid constraints could delay 20% of planned capacity by 2030 (IEA), projects worth more than USD 750 bn could slip, with the sharpest pain for the many construction, equipment and service companies building and outfitting this physical infrastructure. International diversification could ease pressure, but only at the margin: few locations can absorb US-scale builds, and we estimate a total-cost premium of about 20% for Europe versus the US considering the current data center economics.

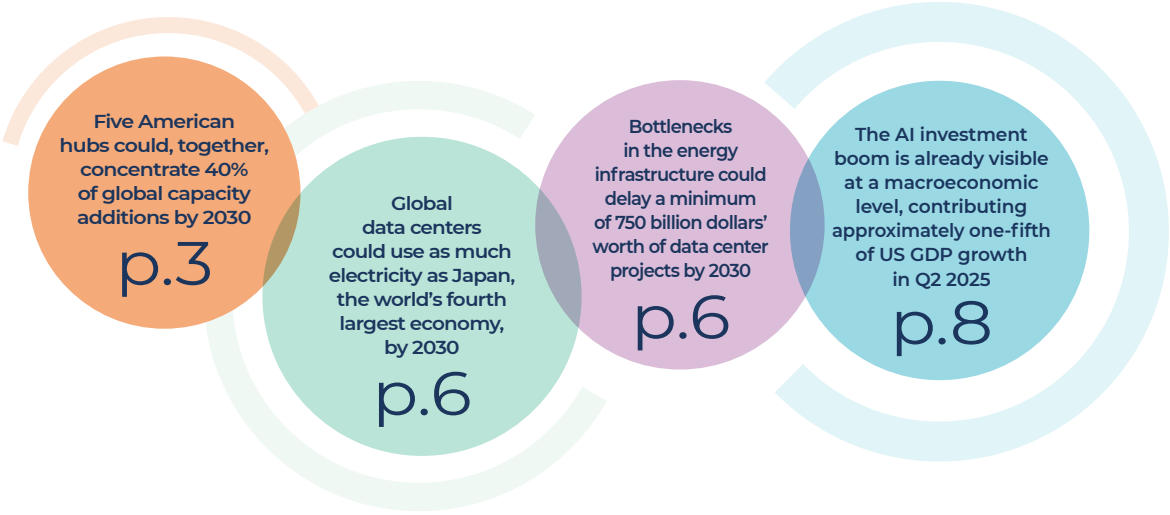
The magnitude of the current investment wave also raises legitimate questions about potential overcapacities. Forecasts for AI-driven compute demand diverge widely: the gap between low and high cases averages 80%, and even reputable sources' bullish scenarios differ by as much as 80%. An overcapacity shock would come from upstream and be felt not only by hyperscale, cash-rich platforms such as Microsoft, Meta, Alphabet and Amazon, but also and more severely by the lesser known and yet crucial colocation companies, which control more than two thirds of marketable capacities and are expected to contribute about half of additions by 2030. An investment freeze would have ripple effects up and down the supply chain, with companies feeling the scissor effect of rising committed capex against weaker realized demand.

•••

coface FOR TRADE

For now, the boom is already macro-relevant but not transformative. Using a bottom-up attribution focused on data-center construction and IT hardware, we estimate AI-related spending accounted for 17–23% of US GDP growth in Q2-2025. For the boom to have broader and more durable impact, it must evolve from an investment cycle into a productivity cycle, one that not only raises efficiency but also generates new goods and services at scale.

SOME OF THE THINGS YOU'LL LEARN...

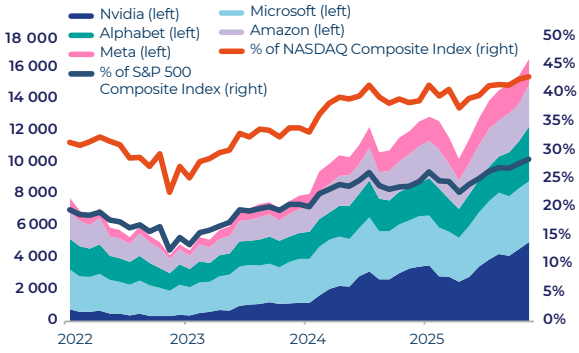


The global boom in artificial intelligence investment brings data centers to center stage

The public release of ChatGPT 3.5 in November 2022 marked a turning point in the global technology landscape, moving artificial intelligence (AI) from research labs to a mainstream service adopted by hundreds of millions of users within weeks. Growing business and consumer interest in AI has since triggered a surge in corporate spending and investor enthusiasm unmatched since the early days of the internet. In less than three years, the combined market capitalization of the world's largest AI-exposed firms (Nvidia, Microsoft, Alphabet, Amazon and Meta) has increased by more than 12 trillion dollars, accounting for much of the rise in major US equity indices **(Chart 1)**.

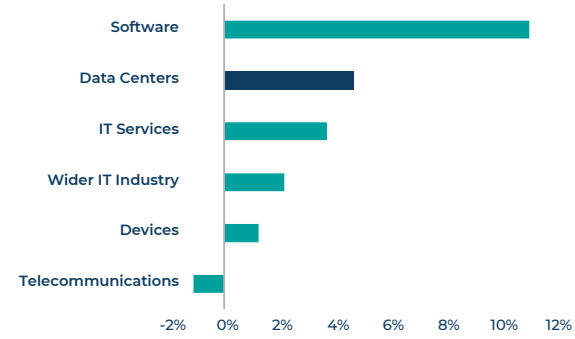
Behind this spectacular surge in equity valuations lies a massive and already tangible investment cycle. Hundreds of billions of dollars are being deployed to build the physical and digital infrastructure needed to train and operate large-scale AI models. While Nvidia, whose dominance in advanced graphic processing units (GPUs) has captured most of the headlines, sits at the visible end of this transformation, its success depends on investment further down the value chain in the lesser-known but highly strategic infrastructure of data centers, the facilities where thousands of servers equipped with these chips are installed, powered, cooled and interconnected. Since 2022, global annual spending on data centers did more than double, rising from 221 billion USD in 2022 to an estimated 475 billion USD in 2025. Booming investment in data centers did not start from a low point: over the past decade already, the migration from locally owned servers to remote, shared infrastructure managed by specialized providers known as cloud computing made data centers one of the fastest-growing components of global IT spending, second only to software **(Chart 2)**.

Chart 1 - Market capitalization of leading AI firms *(billion USD)* **and their share in US stock indices market value** *(%)*



Sources: LSEG, Workspace Coface

Chart 2 - Compound annual growth rate in global IT spending components, 2012-2022 *(%)*



Sources: Gartner, Coface

While dollar figures highlight the magnitude of the investment opportunity created by the hardware side of the AI boom, the data center industry tends to track its progress in terms of power demand capacities rather than currency. Capacities can however mean very different things whether data center development is seen through the lens of the power, information technology or real estate industries. **Chart 3** provides a snapshot of current global capacities from those three perspectives.

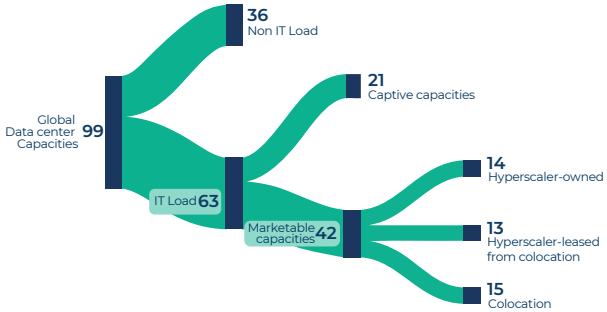
As of 2024, global data center capacity stood at roughly 99 gigawatts (GW), of which 63 GW correspond to computing equipment known as IT load and 36 GW to cooling, power management and building systems (lighting, security and other non-IT load). This distinction is key from an energy perspective, as it determines power consumption profiles and efficiency metrics.

The 63 GW of IT load, which comprise computing equipment including servers, storage systems and networking hardware, represent the most relevant scope from a computing-capacity standpoint, as they directly reflect the scale of active processing power deployed globally. This measure also forms the basis for assessing trends in cloud expansion, AI workloads and enterprise digitalization. About 21 GW of this capacities are captive, that is owned and managed directly by corporations (big technology companies themselves, financial services companies) or large institutions (universities).

The remaining 42 GW of marketable capacities are available for lease and represent the commercially accessible portion of the global data center inventory. This scope is particularly relevant for real estate investors, as it reflects the active supply side of the market where capital deployment, leasing activity and valuation dynamics are most visible. Hyperscalers, a term referring to large cloud service providers like Alphabet, Amazon, or Microsoft that operate massive-scale data centers not only own about 14GW of marketable capacities, but also an unknown but significant share of captive capacities and lease another 13GW from real-estate operators known as colocation companies. Colocation companies are specialists in the leasing of buildings providing shared infrastructure (space, power, cooling), but which leave the investment and management of the IT part of the infrastructure to one or multiple tenants. Colocation companies also own 15GW of capacities that are leased to non-hyperscaler companies. While mostly unknown to the public, it is still colocation companies, and not hyperscalers, that own the largest share of marketable global data center capacities.

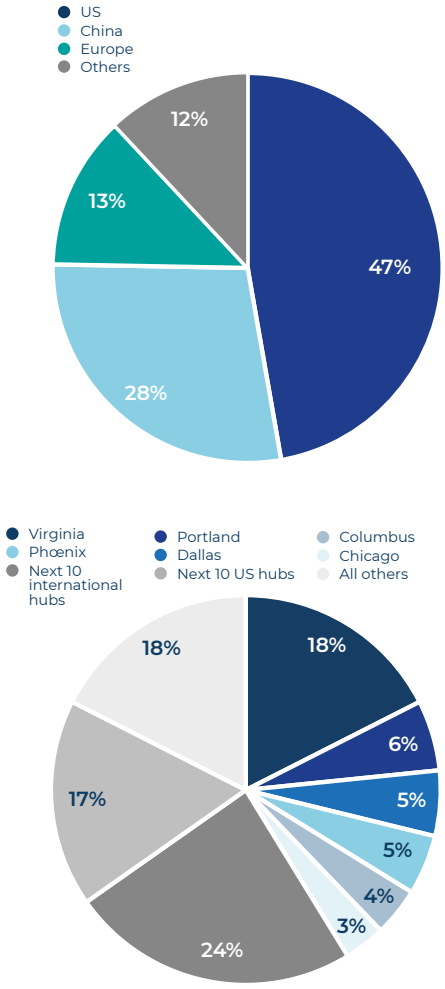
From a geographic point of view, global data center capacities in their broadest definition (IT load and non IT load, captive and marketable) are largely dominated by the US and China whose combined share stands at 75% of global capacities **(Chart 4)**. Because the most widely used digital platforms, from social networks to streaming services and enterprise software, are dominated and hosted by US companies, the United States has established a disproportionate influence over global data center capacities. China remains the only significant exception to the reach of US technology firms for having promoted its homegrown digital ecosystem. The data center industry itself rarely breaks down capacities by country, but rather by hub or cluster, that is loosely defined geographic areas sharing common supply-side (energy, water, land, regulation) and demand-side features (proximity to major cloud providers, enterprises, latency-sensitive users) which are more relevant to analyze market trends. Those data, typically provided by real estate investment services companies and focusing only on marketable IT load, reflect the same US dominance and highlight the predominant role of just five US hubs which, combined, account for over 40% of global capacities.

Chart 3 - Global data center capacities by load type, commercial availability and operational model, 2024 *(GW)*



Sources: Coface calculations based on estimates from IEA, CBRE, Cushman & Wakefield, Turner & Townsend and Synergy Research

Chart 4 -Global data center capacities by region *(2025 estimates) and hub (2024 estimates)*



Sources: IEA, Cushman & Wakefield, Coface

Bigger, hotter, faster: AI is a game changer for the data center industry

The differences between traditional data centers and the facilities purposely built to train and operate advanced AI models can hardly be understated.

Traditional data centers are designed to support general-purpose computing workloads such as web hosting, enterprise applications, databases and cloud services with flexibility, reliability and cost-efficiency as priorities. Compute is delivered primarily by central processing units (CPUs, the family of all-around, versatile processors found in every personal computer and smartphone) deployed inside servers that are mounted in racks (standardized metal frames), powering moderate, distributed workloads.

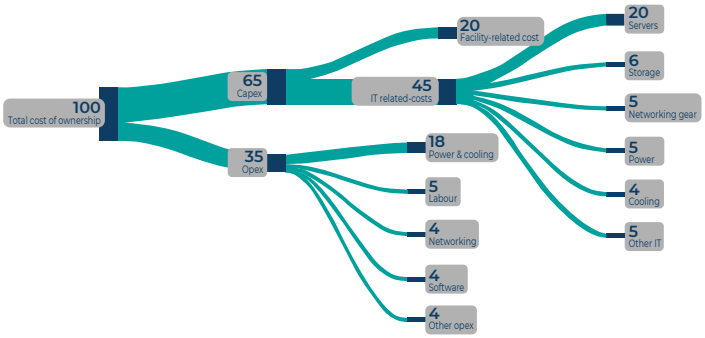
AI-optimized data centers, by contrast, are purpose-built for massive parallel computation (many calculations are performed at the same time, instead of one after another). They cluster thousands of accelerators such as graphic processing units (GPUs), each with thousands of cores and high-bandwidth memory connected through ultra-fast interlinks. This architecture delivers the parallelism required to process billions of parameters across huge datasets, making it indispensable for training and serving large AI models.

These different purposes have profound implications for energy use, cooling and networking. AI-ready data centers host far denser racks, each integrating many more processors than traditional ones. Average power density has jumped from about 7–10 kW per rack to 30–100 kW today, with some future designs already exceeding 600 kW - enough to power more than a thousand French homes. Stargate, a 10 GW mega-project jointly developed by OpenAI, Oracle, Softbank and MGX announced in January 2025, would represent nearly 10 % of total US nuclear output.

Higher density inevitably means more heat. Conventional air-cooling systems are not suitable for demanding AI computations, prompting a rapid shift toward liquid and immersion cooling. Meanwhile, AI workloads require thousands of GPUs to exchange data simultaneously at extremely high speed, driving the adoption of specialized high-bandwidth interconnects far beyond the capabilities of traditional Ethernet networks.

Changes in the purpose and technology of data centers primarily affect their cost structure. Over a data center's lifecycle, capital expenditures generally account for around 65% of total costs, while operating expenditures represent about 35% **(Chart 5).**

Chart 5 - Indicative breakdown of total cost of ownership of a data center *(% of total)*



Sources: Coface calculations based on estimates from IEA, CBRE, Cushman & Wakefield, Turner & Townsend and Synergy Research

Investment in servers, power and cooling equipment (about 30% of total data center costs) as well as power and cooling operating expenditures (17%) are most probably the items seeing their cost increase the most, even though the secretive nature of AI technologies does not allow for precise cost measurements and comparisons over time or across companies and locations. Using public data for the Stargate project as proxy, we can derive from its 10 GW nameplate capacity for a price tag of 500 billion USD a cost of about 50 million USD per MW, that is a minimum of twice the cost of conventional data centers.

Another consequence of the greater power, cooling and connectivity demands of AI data centers is that the list for suitable destinations is much narrower – those resources are, at scale, rare and the most obvious choices for data center location are already used by legacy facilities, whose retrofitting is not always economically and technically possible.

Speed-to-market pushes hyperscalers all-in on US investment

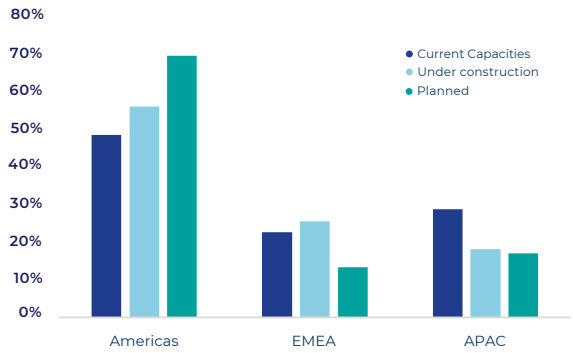
The transition toward AI-optimized workloads is translating into an exceptional wave of data center construction driven mostly by a handful of hyperscalers. This urgency stems partly from fear of missing out, but even more from the belief that greater computing capacities translate directly into superior AI performance. This belief is grounded in the so-called scaling laws, that is empirical observations showing that model capability improves predictably as computing resources increase. Another widely held conviction among AI investors fueling this race is the belief that the first company to achieve artificial general intelligence (AGI) could gain a lasting and potentially insurmountable competitive edge. AGI refers to a hypothetical form of artificial intelligence that can perform any intellectual task a human can - often imagined as surpassing human cognitive capabilities across a broad range of domains. Although the concept remains loosely defined and its feasibility debated[1], many investors treat it as a strategic milestone - OpenAI's founder and CEO Sam Altman reiterated, in a September 2025 interview[2], that AGI could be reached before 2030. The perceived advantage lies in the notion that an AGI system could recursively improve itself, accelerating its own development and capabilities. This self-enhancing loop, sometimes referred to as an intelligence explosion, could allow the pioneering company to outpace competitors dramatically, both technologically and economically.

To move fast, hyperscalers are doubling down on their actual data center infrastructure located in the US. US hubs offer a unique combination of advantages that, arguably, no other international market can match to bring new data center capacities to the market, including a) a familiar regulatory environment which reduces compliance uncertainty and accelerates decision-making, b) generally supportive state-level legislation (36 US states provide tax incentives specifically designed to attract new data center development[3]), c) strong backing from the federal government (the current US administration has publicly endorsed and, as a matter of fact, announced the Stargate project) and d) abundant land parcels with robust fiber connectivity and access to scalable energy solutions, mostly gas-powered power plants (about 40% of US data center electricity consumption, and a major advantage of the US in terms of costs).

1 - *AAAI 2025 Presidential Panel on the Future of AI Research*, March 2025. 76% of the 475 AI specialists surveyed found that "scaling up current AI approaches to yield AGI is "unlikely" or "very unlikely" to succeed, suggesting doubts about whether current machine learning paradigms are sufficient for achieving general intelligence".
2 - *Ich glaube nicht, dass die KI Menschen wie Ameisen behandeln wird*, Die Welt, September 2025
3 - *An Overview of State Data Center-related Tax Incentives*, NAIOP, Winter 24/25 issue

The future domination of the US in the global data center landscape shows in the various estimates of the capacities currently being added to the market. Looking at one such dataset provided by commercial real estate services company Cushman & Wakefield, we find that the weight of the Americas region (of which, 96% alone for the US) is only set to increase for the foreseeable future, with 49% of actual capacities, 56% of new capacities being built and 70% of new capacities planned[4] **(Chart 6)**.

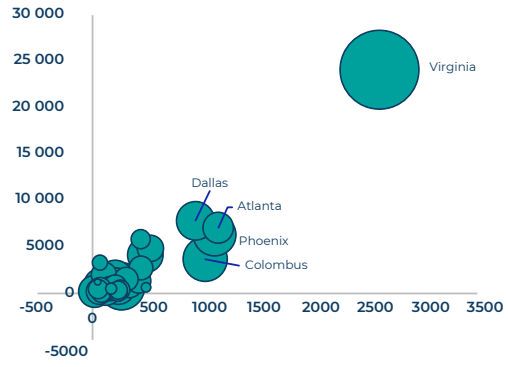
Chart 6 -Share of regions in global data center current capacities, capacities under construction and planned capacities *(% of total IT load, marketable capacities)*



Sources: Cushman & Wakefield, Coface

Looking this time not at regions, but hubs (the wider economic area hosting data centers), we find the same polarizing trend by which the largest current hubs will be the main contributors to new capacities coming online in the next few years **(Chart 7)**, with the US hubs of Virginia, Dallas, Atlanta, Phoenix and Columbus, accounting for about 40% of new capacities being built or planned globally.

Chart 7 -Global data center capacities by hub: currentcapacities (size of the bubble), capacities being built*(horizontal scale, MW)*, **capacities planned** *(vertical scale,MW)*



Sources: Cushman & Wakefield, Coface

Too much, too fast: the US data center pipeline is too large to go online without friction

The possible doubling of US data center capacities by 2030 suggested by data center pipeline data has prompted growing skepticism among industry participants and analysts alike regarding the feasibility of bringing all planned capacities online as scheduled. Investment concentration driven by the same local economies of scale allowing fast speed-to-market could lead to situations where costs and delays actually boom - a phenomenon known as diseconomies of scale, driven here by resource limitations - as local supply growth outpaces available resources **(Table 1)**. Anecdotal evidence is already pointing to substantial delays for the construction of new data centers since late 2022 and even project cancellations in top US hubs including Virginia, Atlanta and Phoenix owing to a mix of rising construction costs, land and zoning conflicts, power grid interconnection delays and environmental (including water) impact.

Table 1 - Key resources: scale advantages and scale challenges

Resources	Economies of scale – efficiency gains at scale	Diseconomies of scale – friction and constraints at scale		
		Scarcity	Delays	Costs
Land	Lower acquisition and planning costs. Ideal for scalable expansion.	Competition for well-connected parcels (power, telecoms).	Long zoning reviews and higher likelihood of community pushbacks for large campuses.	Booming land prices near hubs with high potential.
Water	Reduced water consumption per kW of IT load with larger advanced cooling systems.	AI datacenters require potable water. Water consumption among hyperscalers growing at 10-20% per annum since 2015. 80-90% of data center water consumption is lost to evaporation.	Longer environmental impact studies in hubs already under strain.	Regulators likely to ask data center managers to invest in higher efficiency water management systems.
Energy	Larger facilities enable dedicated substations, direct PPAs, renewable integration, and more efficient transmission build-out.	Grid competition, constrained availability.	Interconnection delays, grid bottlenecks.	Rising wholesale prices in high-growth regions[5].
Labor	Lower hiring and training costs. Faster project completion.	Construction unemployment at a decade low and specialized trades in short supply. Adverse impact of the US immigration policy.	Construction backlogs are growing, with lead times now averaging nine months for data center projects	Construction wages growing between 4–6% since 2023.

Source: Coface

Of all the possible limiting factors hurting future US AI data center capacity growth, energy has arguably been the most debated[6]. The challenge spans the entire power value chain, from generation to delivery, with each link introducing possible bottlenecks that could derail hyperscale ambitions. Using the base case developed by the International Energy Agency (IEA) for the US, we find that every billion dollars invested in AI data centers demands about 125 million dollars invested in the energy sector, with the power grid (transmission and distribution) alone accounting for two thirds of this amount and power generation for the remaining third. Yet these figures are modest compared with the trillions to be deployed in data center infrastructure itself, and capital availability is unlikely to be the binding constraint. The real obstacles are physical rather than financial: above all, time but also labor, equipment and permitting capacities.

4 - *Global Data Center Market Comparison*, Cushman & Wakefield, 2025. Please note that, unlike industry observers from the energy industry, real estate companies tend to focus only on "marketable" capacities, which exclude in-house data centers, and use the IT load as a proxy for the capacity, which excludes non-IT sources of power consumption.
5 - *How AI Data Centers Are Sending Your Power Bill Soaring*, Bloomberg, October 2025. The analysis for the US points to a tripling of electricity costs in areas located near large data centers.
6 - For a comprehensive review of the possible impact of AI on energy markets, see Energy and AI, International Energy Agency, 2025.

Upstream, the US faces a mismatch between fast AI load growth, slow power plant build cycles and the decommissioning of obsolete coal and nuclear power plants. Dispatchable resources such as gas and nuclear power are constrained by permitting, supply-chain fragility and long lead times, while emerging alternatives such as small modular reactors (SMR) remain years away from scale. Renewables, though abundant, require massive storage and transmission upgrades to provide reliability.

Down the power chain, delays are systemic in grid infrastructure where new transmission projects take a minimum of four years to be completed in advanced economies. Beyond permitting and construction, the IEA points to widespread shortages of critical grid equipment such as transformers and estimates that grid constraints alone could delay about 20% of global data center capacities due to come online by 2030. Based on the IEA's baseline projection of a 130 GW increase in global data center capacity by 2030, and our estimates of cost structure and current construction costs, this would translate into delayed projects worth at least 750 billion USD – a conservative calculation that does not factor in increases in data center construction costs (+8% per annum since 2019) or the more expensive equipment mix of AI-grade data centers. America's largest power grid manager, PJM, has also stated that timelines for the interconnection of large data centers can stretch to five years or more owing to technical studies, necessary transmission upgrades, lack of available grid equipment and the complexity to manage hubs which sometimes span across different states.

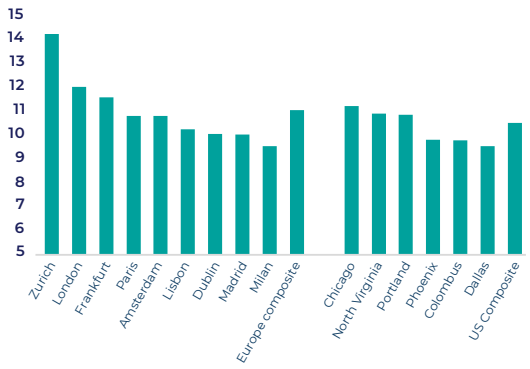
Beyond power generation, transmission and distribution, the rapid and massive deployment of AI data centers also raises issues regarding system operations and overall grid security. Reserve margins (the extra generating capacity above expected peak demand that grid operators maintain to ensure reliability), are tightening in the US, forcing utilities to rely on reserve gas capacities to plug the gap and curtailment protocols to keep an appropriate supply / demand balance. Greater data center flexibility, by which data centers would agree to reduce their power consumption and/or rely on on-site storage and power generation during regional peaks in demand, will be critical to maintain local grid stability… but could also increase costs and encourage hyperscalers to reconsider their siting decisions.

Diversification beyond the US would be challenging and certainly not equivalent

The surge of investment in secondary US hubs such as Atlanta (900 MW of capacity, 8,200 MW under construction or planned), Austin (680 MW, 5,400 MW), and Reno (350 MW, 6,300 MW) contrasts with the growing, but comparatively limited international pipeline: Tokyo, the largest non-US pipeline, plans for "just" 1,900 MW of additional capacity. While the business case for international diversification is not always compelling from a cost perspective, our calculations indicate that the cost impact may be less prohibitive than commonly thought.

To illustrate this point, we build composite data-center cost indices for leading US and European hubs, weighted by their share in current data-center capacity. Our comparison shows that top European hubs report average construction costs of 11.05 million USD per MW of capacity, only around 5 % higher than the 10.52 million USD per MW observed in top US hubs **(Chart 8).**

Chart 8 - Data center construction costs by hub and region *(million USD per MW)*

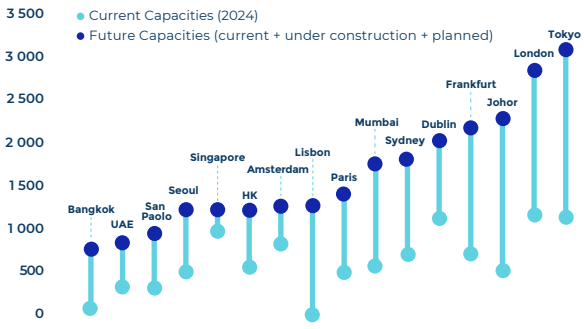


Sources: Coface estimates based on Turner & Townsend and Cushman & Wakefield

Assuming that construction accounts for about 20 % of total cost of ownership, this difference represents only a 1 % premium for top European locations. Even including energy (18% of total cost of ownership) and despite the significant gap between the two regions (European wholesale prices for energy intensive industries are currently twice as high in Europe vs the US[7]), the cost gap would "only" reach 19% - a premium we consider manageable for hyperscalers seeking to accelerate capacity deployment and which could be substantially lower in countries or regions offering more competitive energy costs. In our view, the main barrier to international diversification is therefore not cost but context, and a uniquely favorable US environment still anchors most investment decisions. International destinations remain scarce, with few locations offering the scale, resources, and infrastructure needed to absorb large-scale data center projects.

To identify hubs that could nonetheless benefit from a partial reallocation of capacity away from the United States, we compare current vs future data center capacities – assuming all projects under construction and planned will materialize, and narrowing down the scope down to locations with at least 1% of current or future capacities. We find that large scale diversification options outside of the United States remain limited with only 16 hubs, of which 8 in Asia Pacific, 6 in Europe, 1 in the Middle East and 1 in Latin America **(Chart 9)**.

Chart 9 - Current and future data center capacities by major hub outside of the United States *(MW)*



Sources: Cushman & Wakefield, Coface

7 - *Electricity Mid-Year Update 2025*, IEA, 2025

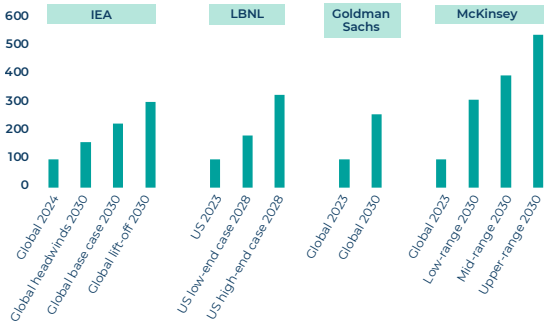
The picture could however shift at the margin with a few countries including the UAE, Saudi Arabia and Argentina signalling active bids to attract US-led, hyperscale data center bids. For host countries and regions, attracting even a small share of the US data center investment spill-over could translate into large numbers. At a time when Europe is seeking to ramp-up its AI infrastructure and boost its digital autonomy, securing just 1 % of US planned capacities would expand Europe's existing capacities by 5% and inject around 19 billion USD in additional data center construction investment.

The height of the current investment wave raises fear over future overcapacities

As the data center industry races to expand its capacities, the challenge could in later years shift to whether capacities are not being built faster than needed. While historical data center growth has followed relatively predictable patterns tied to digitalization and cloud adoption, accurately assessing future supply and demand for AI-specific infrastructure will most certainly prove challenging, as the underlying drivers are both fast-evolving and deeply interdependent.

To illustrate our point, we compare data center consumption forecasts from four reputable sources spanning the energy, academic, investment banking and consulting universes used for the preparation of this report **(Chart 10)**. While they all share a central scenario of strong global demand growth by 2030, which would be broadly consistent with projects in the pipeline coming to the market and finding enough demand, we also observe large gaps in their low case and high case scenarios (80% difference on average), as well as important dispersion between sources themselves (80% growth difference between the IEA's and McKinsey's most bullish scenarios, for instance).

Chart 10 - 2030 data center demand forecasts
(selected sources[8] , 2023/2024 = 100)



Sources: IEA, LBNL, Goldman Sachs, McKinsey, Coface

Those gaps reflect not just possible differences in methods, which generally comprise a mix of bottom-up (extrapolating general trends from granular data) and top-down approaches (anchoring forecasts in macroeconomic indicators or historical consumption patterns), but also different assumptions across a large range of variables. **Table 2** summarizes these assumptions, distinguishing between supply-side and demand-side factors and showing how each could act as a catalyst or a drag on global data center capacity growth. Within each quadrant, these factors span technological (hardware progress and model efficiency), economic (AI adoption and financing conditions) and structural dimensions (energy access, regulation and geopolitics) that together will shape both the pace and the geography of future capacity expansion. The uncertainty around those assumptions is all the greater as the underlying factors are interdependent (higher financing costs could weigh on AI adoption, for instance), non-linear (a single breakthrough in model efficiency could sharply reduce computing needs) and fragile (one missing critical link such as a sufficient supply of dedicated semiconductors could derail investment expectations).

Table 2 - Tailwinds and headwinds for future data center capacities

	TAILWINDS	HEADWINDS
Demand-side factors	Clear monetization and productivity gains sustain investment and rapid deployment across sectors.	Return on investment of AI use cases comes below expectations; clients cut budgets.
	Model size, context length and multimodal features continue to expand, keeping compute intensity on a steep upward path.	Breakthroughs in efficiency and hardware utilization lower compute needs per task and moderate total demand.
	Greater use of video, real-time agents and applications such as augmented and virtual reality drive higher bandwidth and compute intensity	A predominance of text and code workloads keeps bandwidth and latency requirements moderate
	Edge computing[9] adds new use cases but little substitution. Centralized hyperscale demand still dominates.	Wider adoption of edge computing and on-device computing substitutes part of hyperscale demand.
Supply side factors	Rapid grid expansion, renewable integration and flexible PPAs enable faster deployment and stabilize long-term prices.	Slow grid upgrades and limited interconnection capacity delay new data centers and increase power costs.
	Improvements in chip performance and abundant semiconductor supply ease capacity pressure.	Persistent bottlenecks in GPU and memory supply concentrate access among a few firms, reinforcing the race toward larger and more compute-intensive frontier models.
	Lower funding costs and supportive regulation accelerate financing, grid build-out and project delivery.	Higher rates and tighter credit delay projects, reduce utility and colocation investment.
	Streamlined permitting, stable incentives and clear siting rules accelerate new capacities.	Stricter permitting, environmental rules and interconnection procedures delay projects and raise costs.
	Geopolitical alignment, freer data flows and cooperative standards improve utilization and supply stability.	Export controls, data-localization mandates and trade frictions fragment supply chains and duplicate workloads.

Source: Coface

Capacity bottlenecks and oversupply would expose credit vulnerabilities across the AI infrastructure chain

The same assumptions that underpin diverging projections of future data center demand also reveal how risks could propagate across the AI infrastructure chain. Annex 1 provides a simplified view of the main goods and services traded within the ecosystem - from software and IT services to computing capacities, infrastructure, power, construction, and equipment - and the key stakeholders involved: AI application developers, hyperscalers, independent and colocation cloud providers, engineering and construction firms with their subcontractors, suppliers of IT and power systems, and utilities. It highlights the central position of hyperscalers and colocation companies, whose investment decisions shape activity, pricing, and cash flows across the entire value chain.

This interdependence also determines how financial stress would travel through the system under adverse scenarios.

8 - See United States data center energy usage, LBNL, 2024; AI to drive 165% increase in data center power demand by 2030, Goldman Sachs, 2024; AI data center growth: Meeting the demand, Mc Kinsey, 2024
9 - Edge computing refers to the processing of data closer to its source (on devices or local servers, typically) rather than in centralized hyperscale data centers. This reduces latency and bandwidth needs for applications requiring real-time responses.

In a bottleneck environment, where the rush to build runs into limits of land, energy, labour or equipment, stress would first emerge downstream, among construction and equipment suppliers struggling with higher costs and delayed payments. Conversely, an oversupply situation would start upstream, as hyperscalers and colocation operators face lower utilization rates and respond by postponing or cancelling projects, pushing the shock down the chain through reduced orders and tighter financing conditions. **Table 3** summarizes how each scenario would propagate through the ecosystem and which segments would be most vulnerable to credit deterioration.

Table 3 - Propagation of risks along the AI infrastructure value chain

	CAPACITY BOTTLENECKS	OVERSUPPLY
Origin of the shock	Rapid build-out constrained by limited land, energy, labour, or critical equipment.	AI computing capacities exceed profitable demand.
Propagation mechanism	Cost escalation, delays, and supply shortages ripple through construction and equipment layers.	Capex cuts, leasing pauses, and project cancellations reduce order books and utilization downstream.
Primary transmission channel	Margin compression, delayed cash inflows, and working-capital buildup.	Revenue contraction, asset write-downs, and tighter credit conditions.
Most exposed segments	EPC and MEP contractors, specialized subcontractors, component vendors, utilities facing grid congestion.	Colocation operators, independent AI cloud providers, highly leveraged developers, upstream component suppliers.
Typical financial effects	Pressure on margins, liquidity stress, growing receivables, and higher default risk.	Falling utilization, valuation declines, liquidity pressure, covenant breaches.
Relative intensity of credit stress	Highest downstream (execution layer).	Highest upstream (investment layer).

Source: Coface

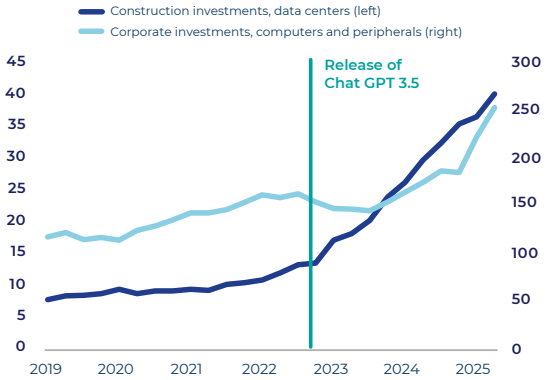
The oversupply scenario would outline the infrastructure nature of data centers. Above break even, variable costs increase only marginally, and very profitable. Below however, low utilization rates weigh on margins, encourage price competition to recover fixed costs.

The AI infrastructure investment is powering, but not transforming the economic cycle yet

Beyond forward-looking uncertainties, the AI investment wave is already having significant macroeconomic impact, raising expectations of a short-term acceleration in US growth with some observers already identifying AI-related investment as the main engine of the current expansion[10]. We qualify that view for two reasons. First, most estimates start from a very wide broad base, sometimes including the whole information-processing equipment category (which includes medical equipment), software investment (which has been dynamic but not accelerating) or intellectual property investment (which includes any type of the latter – including film production). Second, most headline attributions do not remove the generally strong and pre-existing uptrend, implicitly treating the entire post-2022 increase in spending as somehow related to AI, rather than the increment over the counterfactual path absent the AI shock.

To provide a more conservative but accurate estimate, we take a different, bottom-up approach by first looking at investment items directly relevant to data centers, that is investment in data center construction and corporate IT equipment, whose growth has very significantly accelerated since November 2022 **(Chart 11)**, albeit with a lag corresponding to the lead time of OpenAI's rivals to plan and build their own facilities.

Chart 11 - Corporate investment in data center construction and IT equipment purchases *(billion USD)*



Sources: BEA, Coface

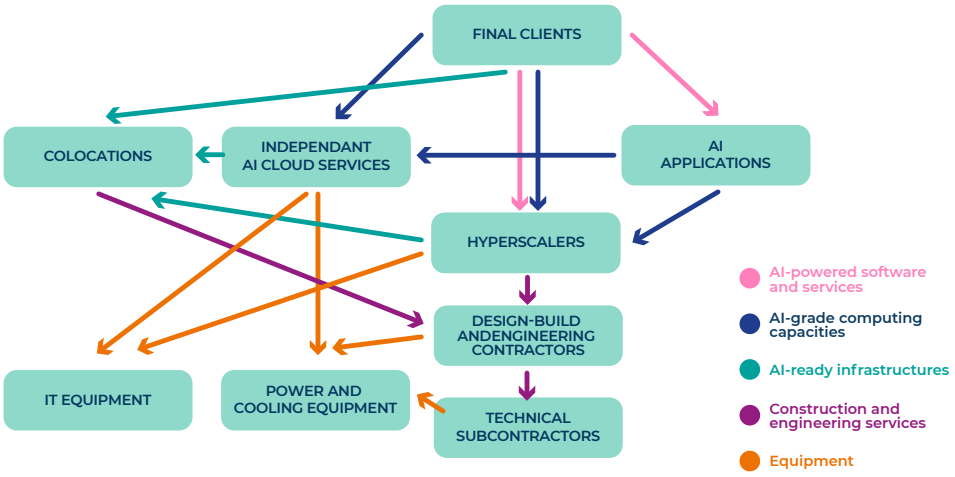
We then net actuals against the pre-Q4-2022 counterfactual to isolate the post-ChatGPT uplift. On this basis, Q2-2025 shows an "excess" of 83 billion USD versus trend - about 17% of the year-on-year increase in GDP, or roughly 0.36 pp of the 2.1% year-on-year growth rate. Using our earlier working assumption that IT hardware accounts for about two-thirds of total data center capex, this maps to an all-in AI data center capex uplift of 112 billion USD, equivalent to about 23% of year-on-year GDP growth (0.48 pp of the 2.1% increase).

Beyond this initial, investment-driven contribution to GDP growth, the macroeconomic significance of AI will hinge on whether massive investment in infrastructure will translate into measurable productivity gains. In principle, broader AI adoption should enhance efficiency across sectors by automating routine tasks and optimizing resource allocation. The short-term effects could, however, be more ambivalent: the first beneficiaries are likely to be large firms with the scale and data access needed to integrate AI tools rapidly, translating into cost reductions and efficiency gains rather than a surge in output. In the short run, AI adoption is therefore more likely to raise profit margins than to expand overall production, as companies produce the same goods and services with fewer inputs. In aggregate, this means a greater share of value going to capital vs labour, reinforcing the capital bias that has characterized much of the digital economy over the past thirty years a dynamic theorized by 2024 Nobel Prize–winning professor Daron Acemoglu and Pascual Restrepo, who describe AI as an automation-biased technological change that raises efficiency mainly by replacing tasks rather than augmenting workers. A contrasting view, supported by Philippe Aghion, 2025 Nobel laureate, sees AI as a general-purpose technology whose benefits broaden as complementary investments in skills, data and organisational change diffuse through the economy. In this perspective, AI ultimately creates new tasks, firms and sources of growth, raising aggregate productivity and tempering distributional effects as adoption spreads. While it has only been three years since the launch of ChatGPT and it is certainly too early to tell which view will prevail, the evidence so far suggests that AI is powering the economic cycle rather than transforming it.

A follow-up report examining the potential transformative impact of AI on the labor market and its broader implications is scheduled for release in Q1 2026.

10 - See, for a more median view, *Economic shifts in the age of AI*, Bank of America Institute, October 2025

Annex 1 - Simplified view of the main goods and services traded within the data center ecosystem



COFACE SA
1, place Coste et Bellonte
92270 Bois-Colombes
France

www.coface.com

November 2025 - Layout: Cécile Bélonie - Photo © iStock