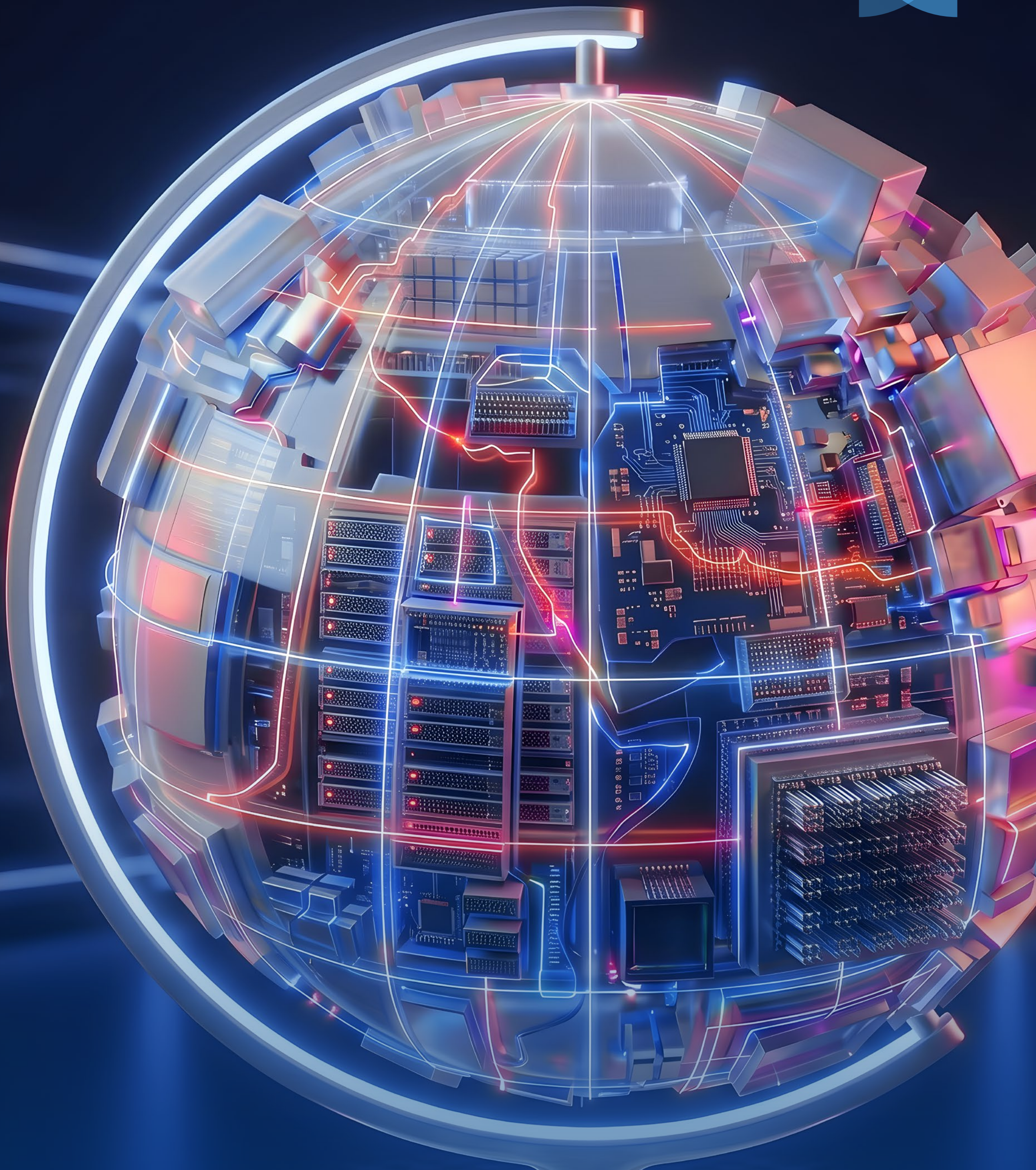
STUDY



The global build-out race

AI data center study | Part 1

Roland Berger

The global build-out race

AI data center study | Part 1

Data centers used to be a niche topic in enterprise IT. Today, they sit at the heart of a global race to build, control, and leverage the computational power that will define the artificial intelligence (AI) era – with consequences for GDP growth, economic influence, national sovereignty, and geopolitics. These facilities are the very foundation of the AI revolution – ignited by the emergence of generative AI in 2022 – with far-reaching implications for individuals, industries, and businesses alike. The predicted value generation potential of AI has attracted capital at an unprecedented pace. By 2026, hyperscaler capital expenditure (CapEx) is set to approach USD 750 billion – fueling tremendous growth along every layer of the data center value chain.

Two structural forces are driving demand: the exponential compute requirements of successive large language model (LLM) generations, and an adoption curve without precedent. The Roland Berger Data center Model projects global installed capacity growing from roughly 90 gigawatts (GW) in 2025 to nearly 350 GW by 2035, with AI's share of total load rising from 20 % to nearly 60 %.

This growth story is strong and fundamentally sound – yet financial dynamics and other structural risks introduce vulnerabilities that merit scrutiny. Both hyperscalers and AI model developers are currently investing more than their AI-linked revenues support, with some hyperscalers' overall free cash flow tightening. The strong growth along the supply chain also creates the conditions for a potential bullwhip correction, and the risk of some short-term, temporary slowdown – triggered by a tightening of external funding, regulatory headwinds, or supply chain bottlenecks – is non-negligible. Ecosystem participants that assume a resilient position across the stack will be those that capture sustainable long-term value through eventual cycles.

Geographically, the US is at least maintaining its lead in global capacity, reflecting the location of hyperscalers and frontier labs and a more enabling regulatory environment. Europe's share is anticipated to decline, constrained by grid connection lead times, high energy costs, and regulatory complexity. Easing these hurdles structurally – not just via financial support – is critical, and the cost of delaying action compounds every year that the gap widens.

The long-term growth outlook for AI and its supporting infrastructure is unambiguously strong, and the countries and companies that position themselves decisively will shape the economics of the AI era. With that framing in mind, this study revisits four questions that stakeholders across the data center ecosystem consistently raise – on build-out momentum, ecosystem risks, the public policy imperative, and diverging regional trajectories.

Contents

Cover AI GENERATED

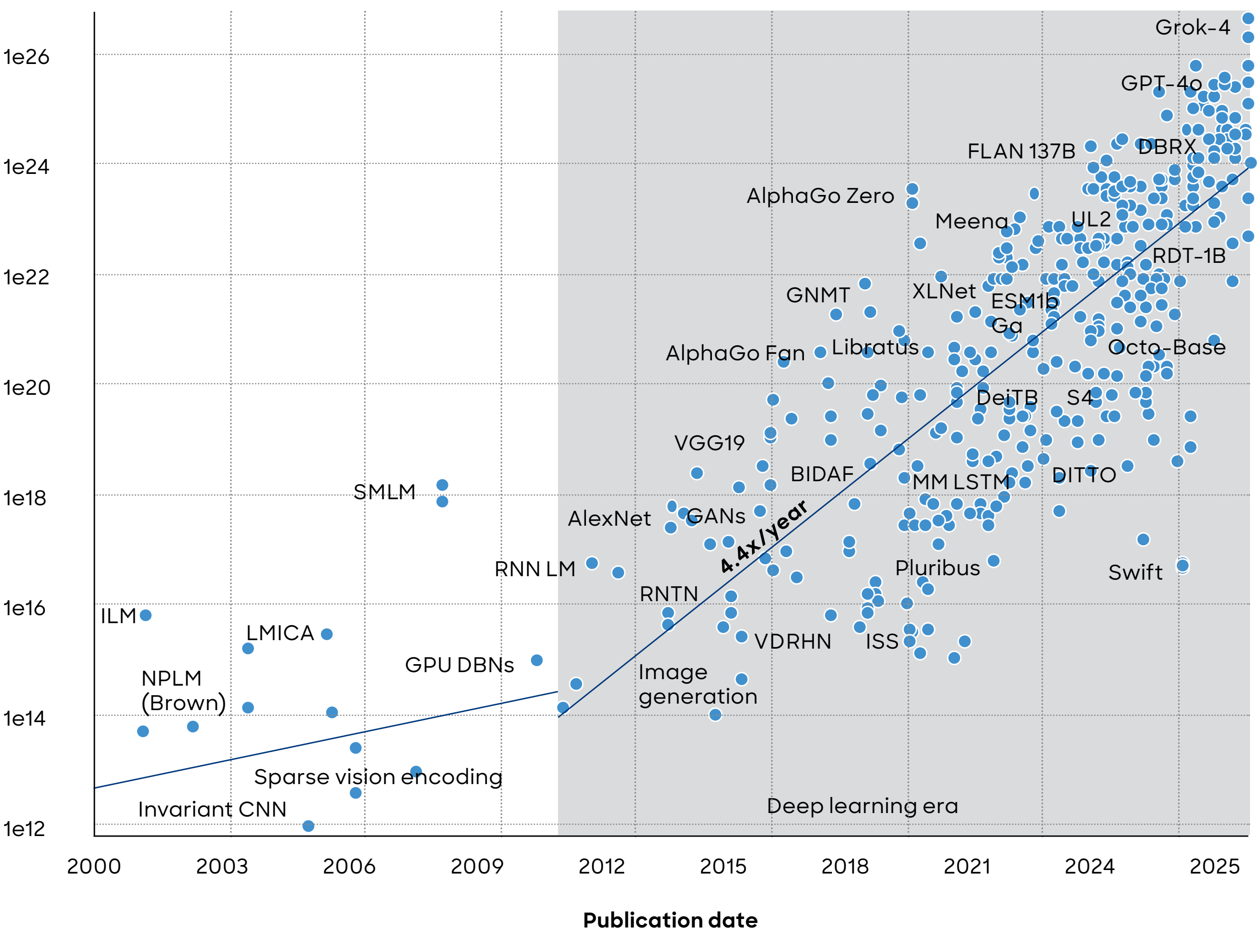
The AI infrastructure boom

1/

1.1/ Two forces driving data center demand

Few sectors have moved from specialist infrastructure to mainstream conversation as rapidly as the data center industry. Whether it's the scramble for power capacity, the extraordinary financial performance of enabling companies such as NVIDIA, growing community opposition to new developments, or even proposals for orbital data centers in space, the conversation starts with the accelerating adoption of AI technologies.

A Training compute (FLOP) over time



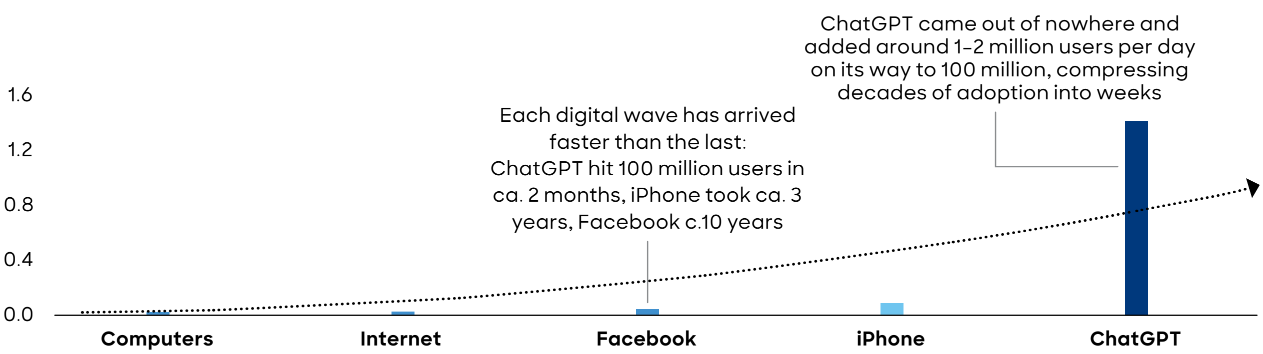
Source: Epoch AI, secondary research

The AI revolution has driven data center growth in two key ways. First is the compounding complexity of model training. With the entry into the deep learning era in 2010, the amount of compute deployed in model training grew by a factor of four year over year. This trend continued and accelerated with the arrival of transformer-based large language models (LLMs). So-called "scaling laws" describe how model performance improves as the number of parameters, the volume of training data, and the number of training iterations increases. The progression across GPT generations serves as an example: GPT-1 launched with 110 million parameters; GPT-4 reportedly operates at an estimated 1.7 trillion; GPT-5 projections range from three to five trillion. We expect such scaling dynamics to continue through at least 2030, as improving the intelligence and capabilities of LLMs requires exponentially more compute resources. Satya Nadella, CEO of Microsoft, described the link between model capability and exponential need for compute resources in an interview: "If intelligence is log of compute, whoever can do lots of compute is a big winner." ▶ A

The second driver is the pace and variety of AI adoption. The blast of AI adoption can, in hindsight, be traced to the years 2022 and 2023. Since then, demand for AI has not grown incrementally; it has outpaced virtually every previous technology shift. ChatGPT reached 100 million monthly active users in approximately two months – faster than TikTok (nine months) or Instagram (two and a half years) – adding between one and two million users per day in the process. That trajectory did not reflect pent-up demand for a familiar product; it was the creation of an entirely new category. By early 2026, weekly active users had reached 883 million, having surged roughly 70 % in just five months during 2025 before stabilizing. What would typically represent a decade of growth was compressed into a matter of quarters. ▶ B

The breadth of AI adoption may matter even more than its pace. Enterprise use cases remain largely in early deployment, compute-intensive workloads – reasoning, agentic AI, and video generation – are scaling rapidly, and the industry is shifting toward inference categories that are significantly more demanding than the text-based applications that drove the initial wave.

B Average new users added per day on the way to 100 million users [m]



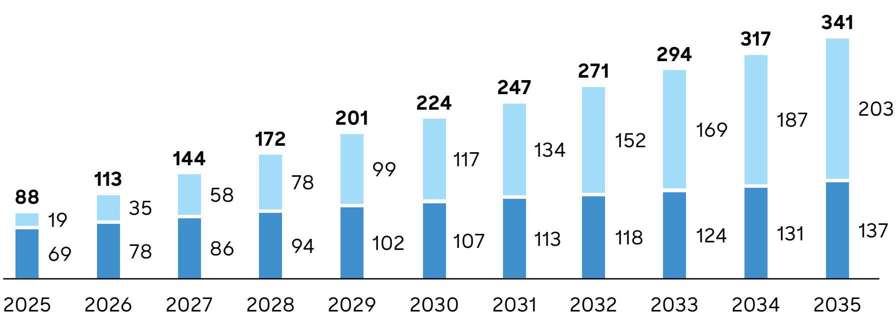
Source: The Conference Board

1.2/ Scale, trajectory, and the capital underwriting the build-out

The data center sector's current growth phase marks a dramatic step change – accelerating the rise of large IT campuses already reshaping the industry. Between 2015 and 2020, the industry underwent a significant expansion driven by the shift to cloud computing and the proliferation of virtualized, cloud-native use cases – the period that established the initial scale of hyperscalers such as Google and Amazon Web Services. That foundation matters, but AI-specific use cases are the dominant growth drivers today, and are expected to remain so for at least the next decade. Data center capacity is typically measured in gigawatts (GW) of power consumption, reflecting the energy and cooling needs of the data center infrastructure, rather than its compute capacity. This study uses the so-called "nameplate power consumption" and encompasses on-premise enterprise data centers, colocation facilities, and hyperscaler campuses. ▶ C

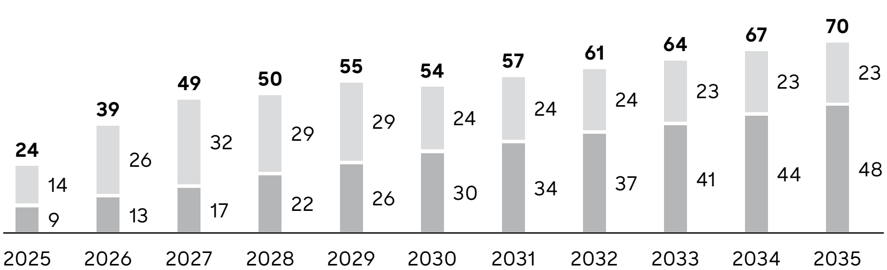
C Proprietary RB model foresees growth to ~220 GW in 2030 and ~340 GW in 2035

Evolution of nameplate facility data center capacity [in GW]



	CAGR '25 – '30	CAGR '30 – '35
Total	21%	9%
AI	44%	12%
Non-AI	10%	5%

Evolution of net-add capacity and replacement [in GW]



Total	18%	6%
Net-add	10%	~0%
Replacement	26%	10%

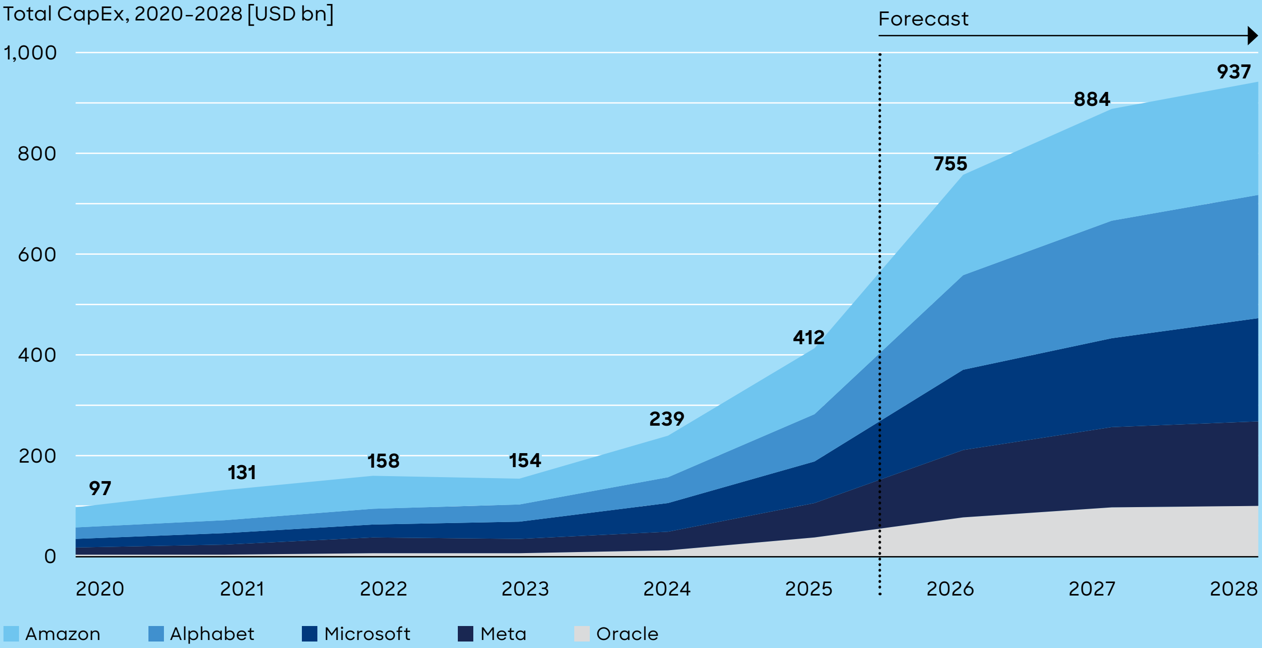
Source: Roland Berger analysis

The Roland Berger Data Center Model projects an expansion from 88 GW – with AI representing only approximately 20 % of that load – in 2025, to 224 GW with a 50 % AI share by 2030, and 341 GW with 60 % AI share by 2035. The model is based on chip production availability forecasts, revenue expectations for key chipmakers such as NVIDIA and Broadcom, as well as hyperscaler CapEx estimates. Across a range of published forecasts, consensus is broadly aligned through 2030 for AI-linked power demand growing at a 35-45 % compound annual growth rate (CAGR), driving the AI share of total data center load to roughly 50 %. The Roland Berger model extends this view to 2035, projecting an overall compound annual growth rate (CAGR) of 21 % from 2025 to 2030, and 9 % from 2030 to 2035. These figures are also consistent with the results of our survey of senior decision-makers spanning Europe's data center value chain, which indicated growth expectations of 16 % to 22 % between 2025 and 2030 and ~10 % for 2030-2035.

The capital fueling today's growth is mainly provided by a small number of hyperscalers. Combined CapEx across

D Selected hyperscaler[1] CapEx evolution

Total CapEx, 2020-2028 [USD bn]



1 Selected hyperscalers shown include Amazon, Google, Microsoft, Meta, and Oracle; broader data center & AI infrastructure estimates may differ depending on inclusion of colocation, neocloud, and power infrastructure spend

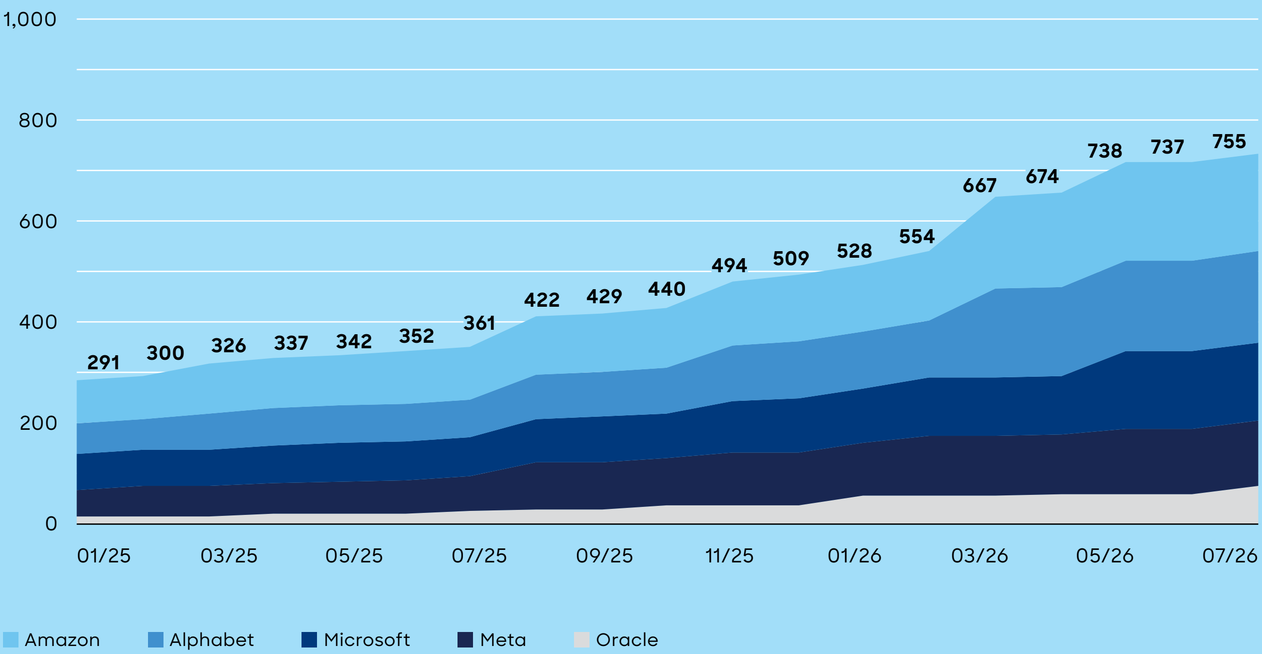
Source: S&P Capital IQ Pro

Amazon, Alphabet, Microsoft, Meta, and Oracle – relatively flat through 2022 – jumped sharply as AI became the dominant investment thesis. The individual commitments speak for themselves. Amazon expects to invest around USD 200 billion in CapEx, weighted toward AWS for AI and core workloads. Google has announced USD 180 – 190 billion in 2026 alone. Microsoft planned to increase its total AI capacity by over 80 % in 2025, and double its data center footprint over the following two years. However, new players, including AI labs, neocloud operators, and colocation providers, are now entering the CapEx cycle, broadening the base of capital flowing into the sector. ▶ D, E

The growth is cascading through semiconductor companies such as NVIDIA and TSMC, which have recorded substantial revenue growth from 2024 to 2025 – as well as data center infrastructure suppliers. Eaton's management described a backlog equivalent to twelve years of what was built in 2025, with the pipeline continuing to grow faster than the industry can build. Siemens reported data center revenue increases in Q1 of fiscal year 2025/26 of around 35 %.

E Evolution of 2026 CapEx estimates throughout 2025 and 2026

Evolution of 2026 forecast, 01/2025 – 07/2026 [USD bn]



Source: S&P Capital IQ Pro

1.3/ The physical reality of AI-scale infrastructure

Modern AI facilities are defined, above all else, by compute and power density. Over the past decade, the industry has undergone a structural shift in how much compute power and energy consumption is packed into a given physical footprint – and that shift has cascading consequences for every layer of the infrastructure. Rack-level power has moved from 10 kilowatts (kW) in legacy installations to over 100 kW today, with future projections reaching 1 megawatt (MW), while facility-level demand has scaled from a few megawatts to campuses now planned in the gigawatt range.

This densification is not just a function of the demand for more compute power and the economies of scale within a data center, it is a key technical requirement. Efficient large model training and high-speed inference both depend on tightly colocated processing, with minimal latency and communication overhead between units. In fact, today's racks containing close to 100 graphics processing units (GPUs) are designed to function as a single compute unit for the algorithms they run.

The density of modern AI compute is demanding on its supporting infrastructure. The industry has shifted from air to more effective but more complex liquid cooling as the primary method of managing rack-level heat – a change that affects facility design, floor loading, and plumbing infrastructure from the ground up. Further, innovations in direct-current power distribution within data centers, which reduce conversion losses and copper needs and simplify delivery to high-density racks, are moving from development into practical deployment.

The implications of this architecture on energy infrastructure are substantial. A one gigawatt data center draws power equivalent to the average consumption of a city of one million households or, arguing from the generation side, absorbs the power output of a conventional Western-style nuclear power plant.

Securing that power involves three distinct but related challenges:

- The mere availability of power from the electricity network or on-site power generation facilities
- The "right" grid access (increasingly to the high-voltage transmission grid) and on-site power distribution infrastructure required to deliver it reliably to the compute and non-compute (e.g., cooling and ventilation) equipment
- The capability of the grid and on-site facilities to deal with potentially hugely oscillating power demand within the data center.

Training runs frequently extend beyond 100 consecutive days, imposing extremely high availability and reliability requirements on the power infrastructure. Compounding this, data centers do not draw power at a constant rate. As a facility cycles between compute-intensive and communication-intensive phases during a training run, its aggregate power demand can oscillate materially within seconds – placing extreme stress on both the grid connection and on-site inline or standby backup systems designed to absorb or bridge those fluctuations. As electricity costs account for ~40 % of data center operating expenditure (OpEx), depending on the location, the combined challenge of costs and grid access has prompted several data center operators to invest in on-site generation capacity.

The supply chain serving the growing AI demand is under constant stress from multiple fronts. GPU shortages remain a live constraint, with Meta securing tens of billions in NVIDIA commitments precisely because production capacity continues to strain against demand. Grid access is increasingly constrained – for example, in the UK there is a significant shortfall of power available to support planned data centers, and in the United States, gas turbine manufacturing lead times have extended to five years, as hyperscaler-led power procurement reshapes global equipment supply priorities. These are not temporary hurdles; they are structural features of an industry building at a pace that physical infrastructure cycles were not designed to accommodate.

Hyperscaler CapEx – some USD 750 billion in 2026, mostly for data centers – is straining the global supply chain, from transformers, turbines, and cooling to chips.

Data center 101

Questions most people want to ask but never do

If AI lives in software, why does everybody talk so much about physical data centers?

Today's generative AI is built on large language models with hundreds of billions – sometimes trillions – of parameters. These parameters are the AI. Producing them requires a process called training: running vast datasets through the model repeatedly over more than 100 consecutive days, in a single, integrated facility. Once trained, models must also handle inference – responding to user queries at scale – which exceeds the processing power, memory, and energy capacity of individual devices and benefits enormously from economies of scale. Many AI labs also choose to keep their models centralized rather than distribute them to end users, reinforcing the need for large, purpose-built infrastructure.

Structural boom, cyclical fragility – Potential correction risks ahead

2/

2.1/ The financing gap beneath the growth story

The capital and rental commitments underwriting AI's infrastructure build-out are running well ahead of the revenues and cash flows available to support them. Despite revenue of USD 13 billion in 2025, OpenAI reportedly generated a cash burn of USD 8 billion. The shortfall was covered by further external funding. Similarly, Anthropic committed USD 50 billion in CapEx at the end of 2025 against USD 9 billion in revenue run rate at this time, requiring external funding like the USD 30 billion series G in February 2026. Both companies are, in effect, financing their infrastructure ambitions and expenses on the expectation of future revenue growth at a rate and scale that has no direct precedent in private markets.

This is not, in itself, a sign of an industry in distress. Land-grab phases in high-growth technology sectors routinely involve spending ahead of revenue, and both companies are operating in an environment where technological leadership compounds quickly. The risk is not the gap – it is the sensitivity of the model to its own assumptions. As Dario Amodei, CEO of Anthropic, noted in a February 2026 interview: "If I'm just off by a year in that rate of growth, or if the growth rate is 5x a year instead of 10x a year, then you go bankrupt."

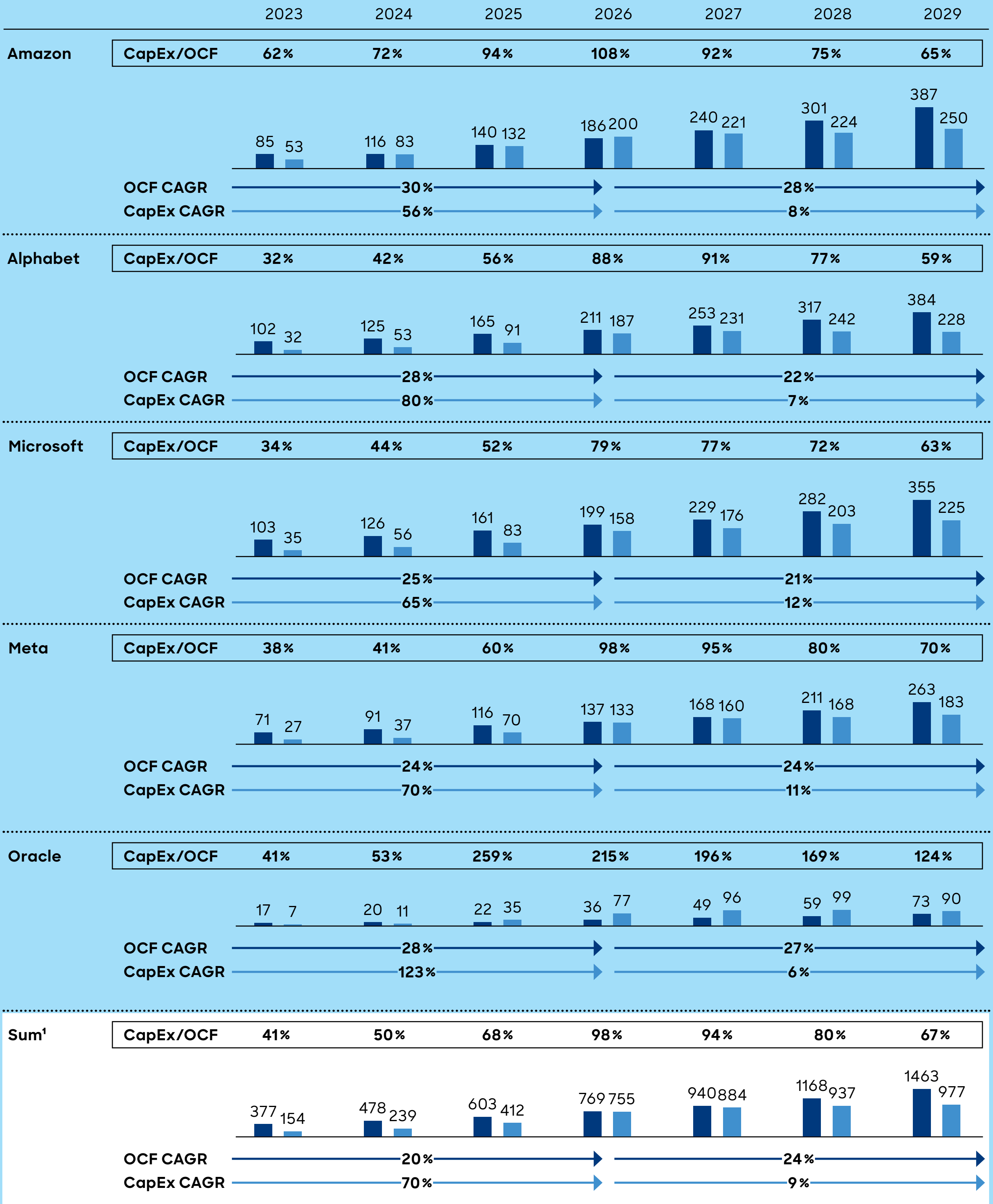
Hyperscalers are in a structurally stronger position, funding AI CapEx predominantly from the cash flows generated by their legacy cloud businesses. But even here, the math is tightening. Free cash flow for several of the major operators is expected to turn negative in 2026, and successive rounds of expanding CapEx guidance have been met with growing caution from investors. The market has not withdrawn its conviction in AI, but it is beginning to ask when the investment phase converts to returns, and on what timeline. ▶ F

Hyperscalers' reliance on debt markets is growing. In 2025, hyperscalers issued more than USD 120 billion in debt – against an annual average below USD 30 billion per year between 2020 and 2024. Amazon alone placed USD 54 billion in debt in March 2026. This is not inherently problematic for companies of this scale and credit quality, but it marks a meaningful structural shift in how the build-out is being financed. Equally notable, the aggregate CapEx expectation for 2026 more than doubled – from approximately USD 350 billion to USD 750 billion – between July 2025 and July 2026. If that pattern of upward revision continues, the assumptions underpinning current return-on-investment timelines will require revisiting.

AI CapEx is racing ahead of revenue. Cash flow from other core business earnings and external financing bridge the gap for now, but momentum will hold only if breakeven of revenues and costs comes within reach.

F Hyperscaler CapEx and operating cash flow (OCF), 2020-2029

Actuals and consensus estimates [USD bn, %]



Source: Company filings, S&P Capital IQ Pro

1 Sum not drawn to scale to enhance layout

OCF CapEx

2.2/ Mixed market signals, supply chain risk, and the case for structural long-term growth

Potential margin pressure compounds the financing risk. On-demand and spot GPU rental rates have been fluctuating, making the economics of the pure GPU-as-a-Service model difficult to predict for infrastructure players ("neoclouds") that depend on this segment. At these spot pricing levels, the model is potentially a thin-margin business, requiring tight control over energy costs, CapEx planning, and hardware lifecycle management. This applies to individual GPU-hour spot and short-term on-demand rates. For larger installations – particularly training clusters and increasingly inference fleets – the picture looks different: scarcity and long-term commitments locking in capacity years in advance were evident in spring 2026, as demonstrated by the Anthropic, xAI, and AWS agreements. The rationale: growing demand for ever-larger training installations and Anthropic's surge following its Claude Code success.

The supply chain structure itself introduces a further layer of instability. Due to the length of procurement cycles and the time required to build data centers, modest shifts in end demand can produce disproportionately large swings in upstream orders for data center equipment such as GPUs, memory, and cooling equipment. Each stage in the value chain has to adjust its operations and inventory levels in response – amplifying, rather than absorbing, potential demand volatility. Operators hedging against shortage risk by over-ordering or locking in long-term supply commitments reinforce this dynamic by delaying the feedback that would otherwise correct it. This bullwhip dynamic produced multi-billion-dollar misallocations during technology boom cycles. The current AI build-out, with its extreme capital intensity and long lead times, is structurally well suited to reproduce it.

The correction risk is non-negligible. A sudden drop in the current level of external funding, a significant regulatory change, a supply bottleneck, or technological displacement, of the kind briefly demonstrated by the DeepSeek disruption in 2025, could result in a sudden

Data center 101

Questions most people want to ask but never do

Who captures the most upside from the AI data center expansion?

AI chip companies – led by NVIDIA and Broadcom – capture the largest share of near-term value, as GPUs and AI accelerators sit at the core of every data center and command premium pricing and margins. Memory manufacturers are close behind: AI chips require high-bandwidth memory (HBM) to move data rapidly during training and inference, and supply remains tight enough to constrain the entire build-out. The second part of this study examines where else in the supply chain significant value is being created. Hyperscalers such as Amazon, Microsoft, and Google buy this hardware, deploy it at scale, and rent AI capacity to downstream clients – including model developers like OpenAI and Anthropic – who in turn use that capacity to build products and services, though many have yet to reach sustainable economics.

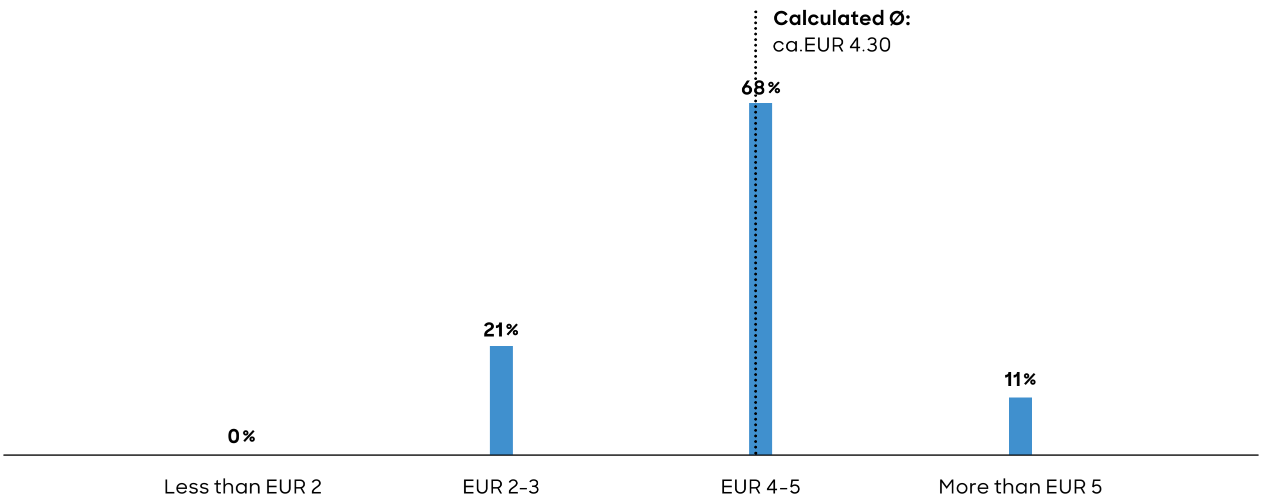
slowdown of the build-out. However, the underlying long-term demand for compute is growing. Enterprise AI adoption is still in early deployment phases and physical AI, autonomous systems, and large-scale agentic workloads are scaling. Therefore, any moderation will likely be short-lived.

The challenge for operators, investors, and suppliers is how to support the data center expansion at its current pace and scale in a way that is resilient to a period of moderation. That means focusing on demand transparency and predictable contract structures with customers downstream, careful selection of financing instruments, and maintaining balance sheet health. Scenario-based operations planning is potentially needed in addition to mobilizing resources to capture the data center growth and value. ▶ G

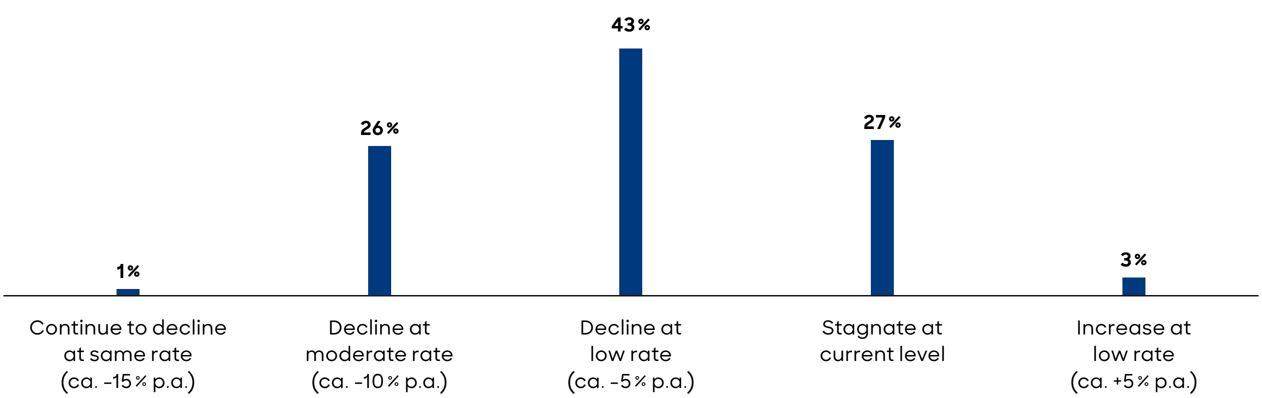
The correction risk is non-negligible. The current AI build-out, with its extreme capital intensity and long lead times, is structurally well suited to reproduce the bullwhip dynamic that produced multi-billion-dollar misallocations during technology boom cycles – yet the underlying long-term demand for compute continues to grow.

G Industry survey: GPU/hr pricing trends & outlook[1]

Assuming that on-demand pricing was the dominant pricing model in 2025, what was the average hourly price per GPU in 2025 for the H100 GPU class?



Looking at only one cycle of H100 class GPU and assuming no refreshment with new GPU models, how do you expect the average price per GPU/hr to develop in 2026/2027?



1 Question type: Single select; sample size: 70

Source: Roland Berger industry survey (Jan 2026)

US dominance – Why Europe is falling behind in the race

3/

3.1/ US dominance and the scale of the gap

The US is the undisputed center for global data center capacity. North America had 4,322 data centers in 2025, with the US alone having 4,032 – more than Western Europe (2,715) and Asia (1,409) combined. The NA share of global installed capacity remained above 55 % from 2023 to 2026. According to the Roland Berger Data center Model, that dominance is set to persist: the US global share is projected to climb slightly to around 58 % by 2030, while Europe declines from roughly 13 % of installed capacity today to around 10 % – a continuation of a downward trend already visible since 2023. In the scenario discussed here, China is expected to stay at roughly 15 % due to constrained GPU access. NA data center capacity is projected to expand from 49 GW in 2025 to 130 GW by 2030 – a CAGR of approximately 22 % – with hyperscalers expected to represent the majority of that base by the end of the decade and the fastest-growing segment concentrated in facilities of 250 MW and above. ▶ H

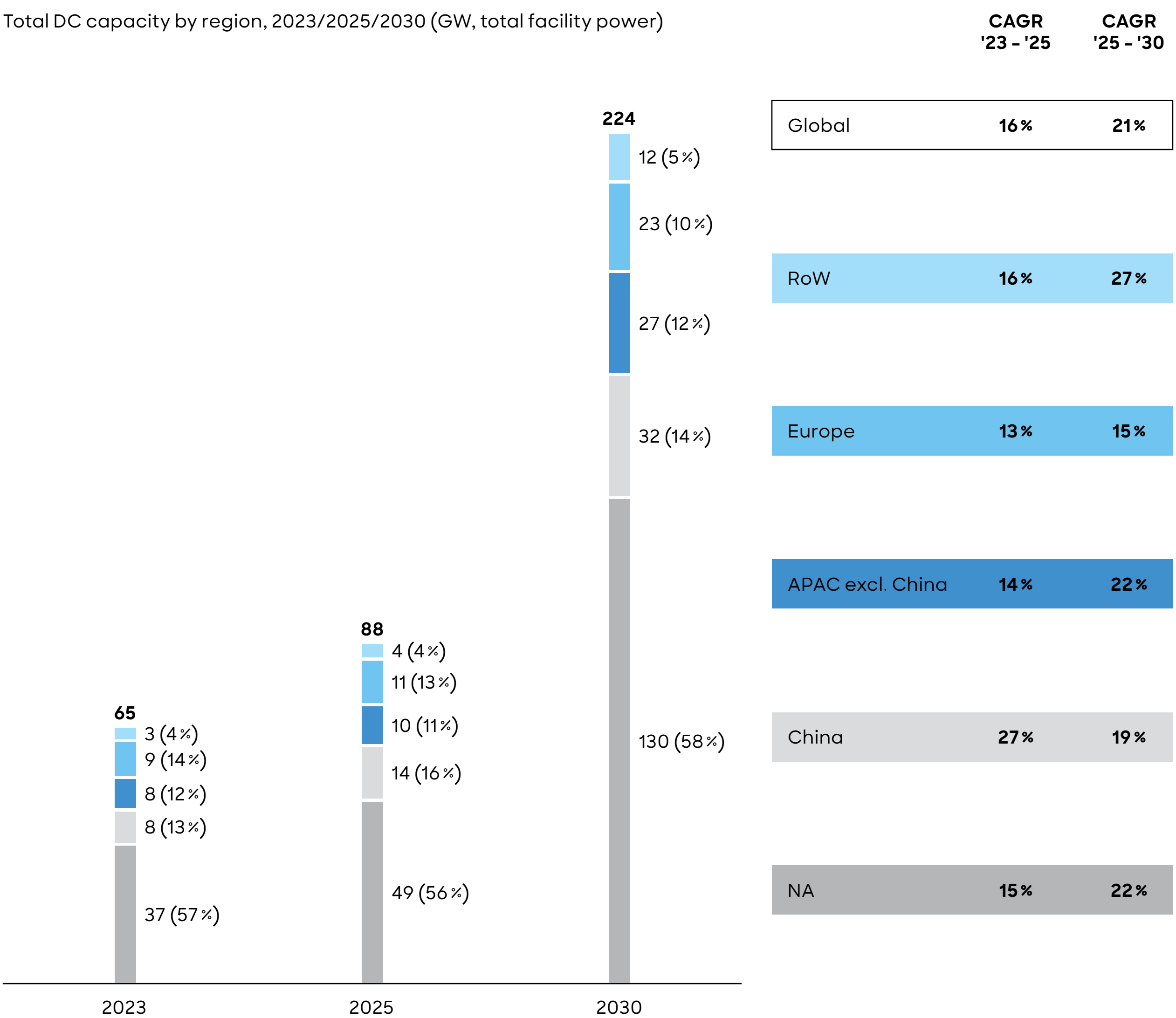
Of the largest data center developments announced to date – all above one gigawatt in planned capacity – every project but one (Portugal's SINES Data Campus, targeting approximately 1.2 GW by 2030) is US-based: Stargate at 7–10 GW, Project Rainier at 5+ GW in Indiana, xAI's Colossus Campus in the Memphis area at roughly 2 GW, Meta Hyperion in Louisiana targeting an initial 2 GW, and Microsoft's Fairwater Campus at approximately 2 GW. Europe's most prominent projects are the SINES Data Campus in Portugal, which targets approximately 1.2 GW by 2030, and the Cambois site in the UK, planned at above 700 MW – both meaningful by European standards but well below the leading US developments.

The concentration of data center investment in the US has not gone uncontested. Between May 2024 and March 2025, more than USD 60 billion in US data center projects were blocked or delayed as a result of local opposition, with concerns including grid overload, water consumption, rising power bills, and the value of tax incentives extended to developers. Several states introduced legislation in 2026 to pause construction of large new energy consumers while those issues are studied – a dynamic that adds a layer of political uncertainty to a build-out that, on paper, still shows no signs of slowing.

In terms of the nature of the AI data center load, US-based developers including Google, Meta, Microsoft, and Amazon/AWS maintain the large-scale research and training infrastructure that drives frontier model development, while European AI activity is predominantly limited to inference-facing: smaller facilities, optimized for reaction speed and low latency, serving end users rather than producing foundational models. Among major AI developers, only Mistral has a meaningful European operational base. The consequence is that Europe's data center investment is weighted toward the less compute-intensive workloads – a position that limits the scale of infrastructure investment the region can organically attract.

> Power availability has become one of the tightest constraints on AI data centers. In the US, developers are increasingly turning to on-site gas-based power generation, trusting a pipeline network that has held up better than the electricity grid; in Europe, decommissioned power plant sites are drawing interest for their existing grid connections.

H Proprietary RB Data Center Model foresees continued dominance of North America



Source: Roland Berger analysis

3.2/ Structural constraints limiting European build-out

Europe's data center power draw is expected to grow at approximately 15 % per year between 2023 and 2030, compared to 22 % for the US and 19 % for China. The difference reflects the physical and regulatory conditions that govern how quickly new capacity can be brought online – European enterprises are adopting cloud and AI at a meaningful pace, but the infrastructure to support that demand is harder and slower to build. ▶|

Three structural constraints are limiting Europe's build-out. The first is grid access. In Europe, connecting a new data center to the electricity network can take up to seven years, compared to one to three years in the US. European permitting processes are longer, regulatory frameworks are more prescriptive, and the workarounds available to US operators – specifically on-site gas turbine generation – are largely blocked by European environmental rules. In Europe, obtaining permits for large gas turbine plants to bridge grid constraints is difficult, and fuel costs are higher than in the US or China. The UK illustrates the tension in concrete terms: the government's Compute Roadmap targets 6 GW of AI-capable data center capacity by 2030, but approximately 80 % of existing UK facilities are clustered around London, straining a regional grid that was never designed to absorb demand at this scale. The Clean Power 2030 Action Plan commits to reducing gas generation by approximately 85 % while adding offshore wind and battery storage as primary replacements – a transition that creates a real supply risk during precisely the years when data center demand is projected to be strongest.

The second constraint is energy cost. Europe carries among the highest industrial power prices of any major data center market globally, materially above the US, China, and the Middle East. High dependence on imported gas and oil exposes operators to supply disruption in ways that markets with more diversified or domestic energy sources do not face. As electricity accounts for approximately 40 % of data center operating expenditure, the energy cost compounds directly into competitive disadvantage – particularly as AI workload intensity and, therefore, power consumption per rack increases. Respondents from our survey conducted for the purpose of this report echo this concern: 69 % of data center professionals expect energy availability to become more challenging by 2035. This makes energy availability and affordability not just an economic issue, but a decisive factor in Europe's digital sovereignty.

The third constraint is regulatory complexity. Operating a data center in Europe, particularly in a major market such as Germany, means navigating a compliance landscape that spans a variety of regulatory dimensions: data protection under GDPR and national implementing legislation; cybersecurity obligations under NIS2; telecommunications law; energy and grid connection regulation; building, zoning, and environmental approvals; workplace health, safety, and fire protection; commercial and contract law; sector-specific rules for regulated industries; and certification expectations under frameworks including EN 50600 and ISO 27001.

For example, Germany's 2023 Energy Efficiency Act defined specific requirements for data centers, including maximum PUE (power usage effectiveness) levels that became stricter depending on the year of commissioning, and mandatory obligations around an energy reuse factor (ERF) – measuring the proportion of waste heat that is reused, for example in district heating networks. While improving PUE is ultimately in an operator's own economic interest, mandatory heat reuse was adding a further burden: it required proximity to suitable heat sinks, coordination with municipal heat planners, and investment in heat recovery infrastructure. Data centers are thus generally also viewed as hubs for various forms of energy – a fundamentally correct perspective, but one that doesn't make life any easier for anyone involved. In June 2026, the German government's cabinet approved a draft bill – still subject to parliamentary approval – that scales back several of these requirements beyond what the EU Energy Efficiency Directive itself mandates, including relaxed PUE thresholds and longer compliance timelines, alongside a heat reuse obligation that is only in place if it passes a cost-benefit analysis in terms of technical and economic feasibility. While welcomed as a step toward reducing bureaucratic burden, industry associations such as Bitkom argue the revisions still fall short of making Germany genuinely competitive as a data center location. The episode is itself illustrative of the compliance burden: operators must navigate national rules that exceeded EU requirements, only to see them walked back roughly three years later. The cumulative effect of the European regulatory expectations creates a strong compliance burden. Survey respondents identified regulatory complexity as one of the three primary bottlenecks for European data center development, alongside grid access and energy cost. ▶ J

That said, the constraints shaping Europe's build-out could also serve as potential points of differentiation. Asked to rank the three most important requirements or differentiation factors for a European data center in 2030, industry survey respondents placed digital sovereignty and

Data center 101

Questions most people want to ask but never do

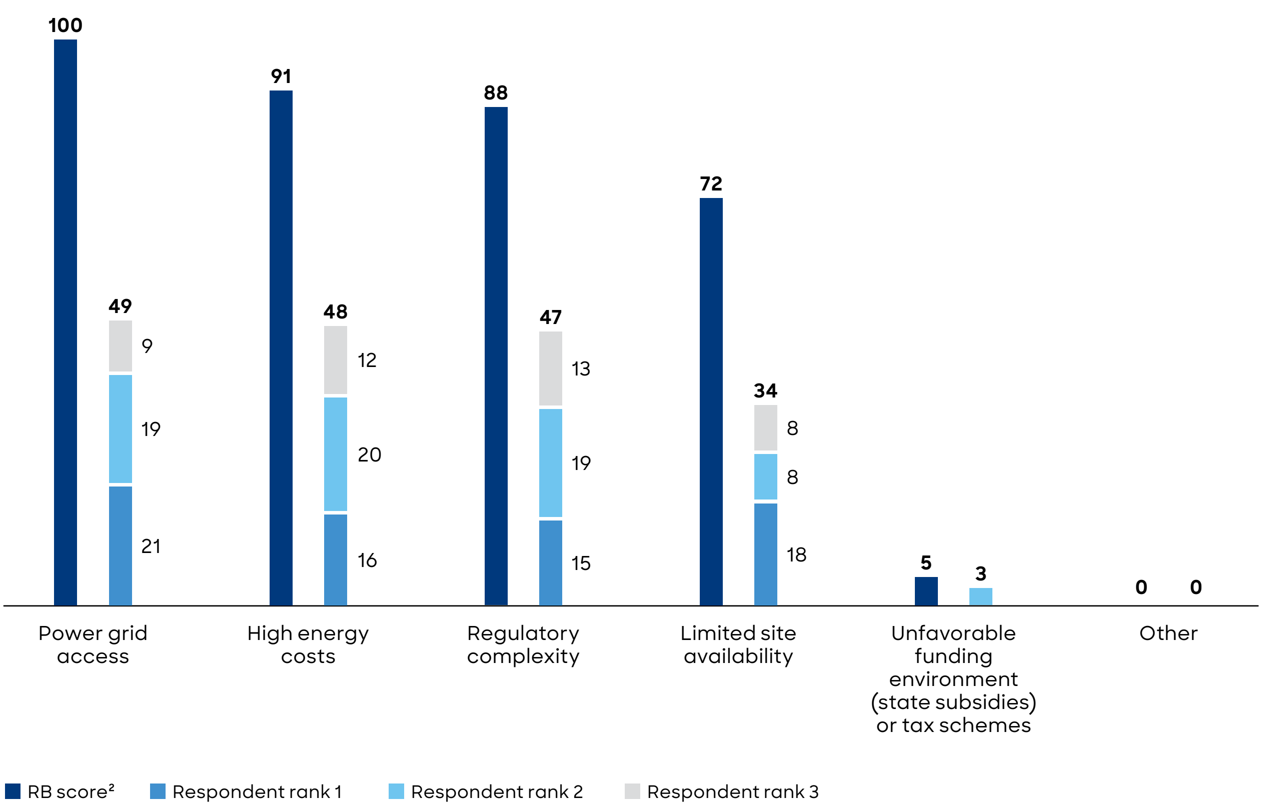
Why have megawatts and terawatt-hours become the language used to describe AI data center scale?

AI data centers are dense concentrations of high-performance chips, each drawing up to one kilowatt of power on its own. A facility housing hundreds of thousands of chips can continuously consume 100 megawatts or more – before accounting for networking, other IT functions, and cooling, which add substantially to that figure. Since power supply and thermal management are the central engineering and operational challenges of any data center, megawatts have become the natural unit to describe both a facility's power draw and its cooling capacity. That level of demand is comparable to a mid-sized industrial plant or a small city. When total energy consumption over time is needed, the unit shifts to terawatt-hours per year, factoring in average utilization across the installed base.

regulatory compliance first, followed by sustainability and energy integration, and then availability and resilience. The implication is that the region's long-term position rests on trust, compliance, and green credentials – a differentiated proposition, but one that will require further regulatory and policy support, without adding wasteful bureaucracy, to translate into meaningful growth and economic wealth.

Industry survey: Bottlenecks for data center development in Europe[1]

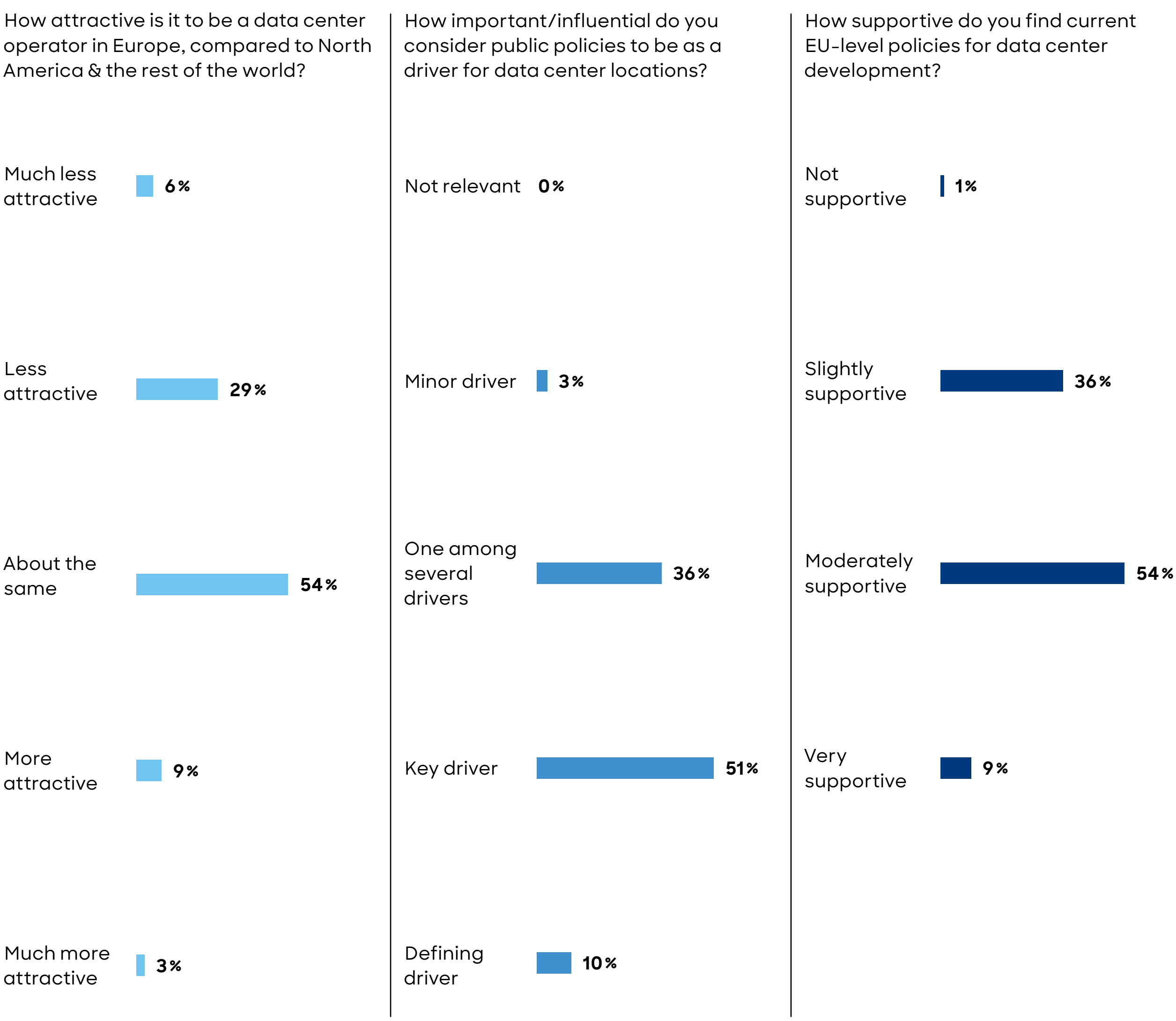
Which of the following do you see as the three main bottlenecks for data center development in Europe?



1 Question type: Select and rank up to three choices; Sample size: 70
2 Data normalized by following method: Scores assigned Fibonacci numbers (#1:8, #2:5, #3:3); highest overall score set at 100 (indexed)

Source: Roland Berger industry survey (2026)

J Industry survey: European location & public policy perception[1]



1 Question type: Single select; sample size: 70

Source: Roland Berger industry survey (2026)

The public policy imperative for Europe

4/

4.1/ The policy rationale for domestic AI infrastructure

European public policy debates around support for (AI) data centers are often framed in terms of jobs, sovereignty, and growth. All three motives matter, but they differ materially in economic weight and strategic relevance. Direct local job creation is likely the weakest argument for subsidizing AI data centers. Construction of a large facility does create temporary jobs in civil works, MEP, trades and local services, and upstream demand spills over into the value chain for power, cooling, and IT equipment. However, once operational, hyperscale AI data centers are highly automated and require only a small permanent workforce for operations, maintenance, and security. Compared to traditional industries such as automotive manufacturing, the number of sustained jobs per million euros of invested capital is two orders of magnitude lower. A EUR 5 billion car plant might support tens of thousands of direct and indirect jobs, whereas a EUR 5 billion AI campus will create only a few hundred permanent roles.

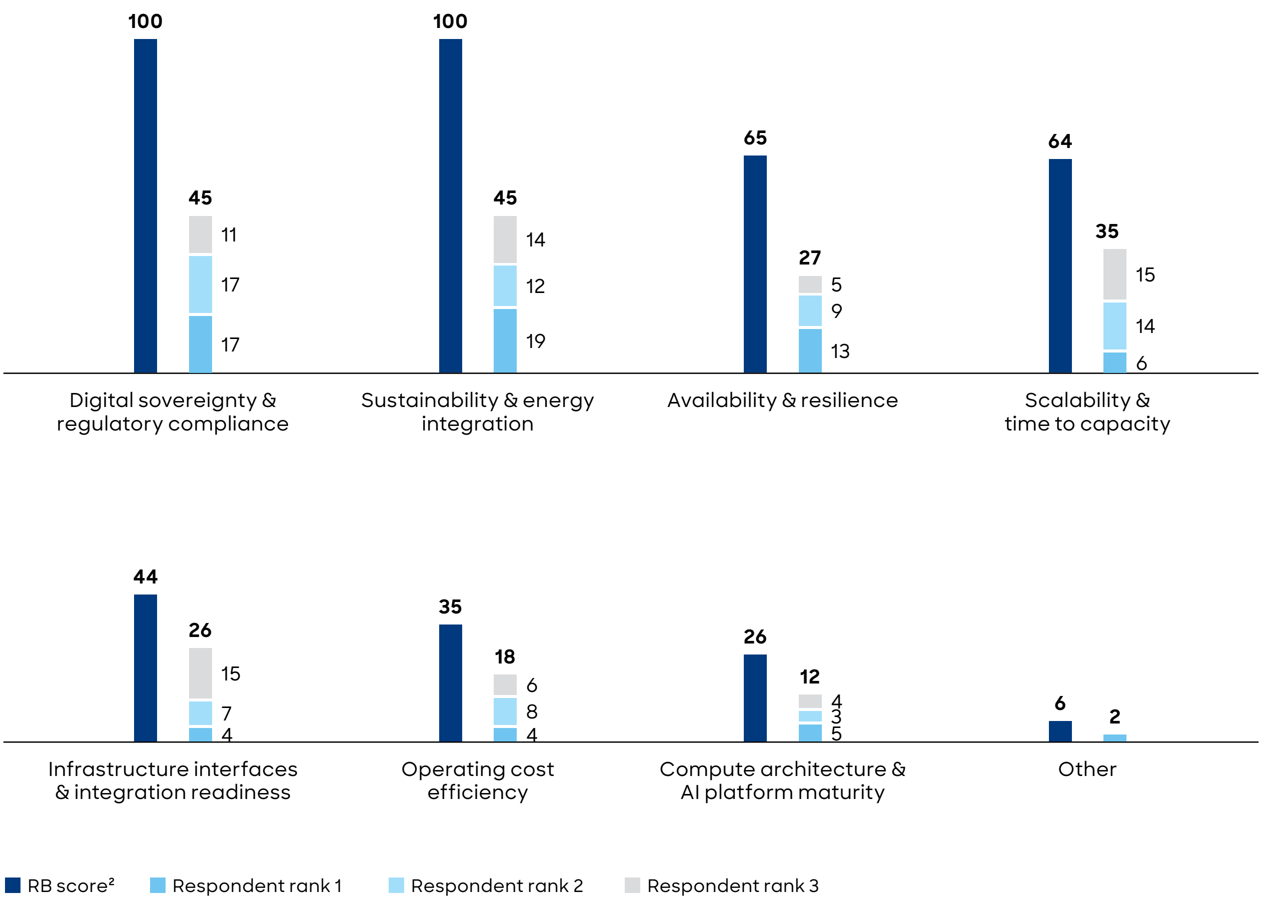
Arguments around digital sovereignty, data privacy, and resilience are stronger. Compliance with the General Data Protection Regulation (GDPR) and related national laws is not optional for European enterprises. Data-intensive sectors, such as healthcare, finance, public administration, and critical infrastructure, need cloud and AI infrastructure that can guarantee lawful processing, enforce European data subject rights, and avoid exposure to extraterritorial laws such as the US CLOUD Act, which can compel US-based providers to hand over data stored abroad. The French concept of "trusted cloud," operationalized through SecNumCloud certification – a national framework that restricts foreign access to certified infrastructure – and the broader idea of sovereign cloud show how public policy can shape infrastructure models that align with European privacy norms and limit foreign jurisdictional reach.

Beyond data privacy, local infrastructure matters for resilience and sovereignty. In the context of rising geopolitical tension, depending almost entirely on hyperscalers hosting data and models abroad creates vulnerabilities: governments. Critical sectors may face access risks if relations sour, if sanctions are imposed, or if foreign regulators take divergent stances. Domestic AI data centers and sovereign cloud hosting increase the ability of European states and firms to maintain operations, enforce their own rules, and respond to crises on their own terms. They also give regulators leverage to insist on practices around bias mitigation, training data provenance, and safety that reflect European values. ▶K

The most significant policy motive, however, is securing domestic access to A- driven GDP growth. AI is expected to reshape virtually every sector – from manufacturing and logistics to healthcare, public administration, and finance – and thus a significant share of future productivity gains. In August 2025, The Economist wrote that roughly 20–40 % of the United States' then current 2 % annual GDP growth was already linked to AI, with a substantial part attributable to investment in AI infrastructure. A 2025 IMF Working Paper, "The Global Impact of AI: Mind the Gap' projected that AI could lift global GDP by 1.3–4 % cumulatively over the next decade, but not evenly: the paper argues that national gains will depend on each country's "AI exposure, preparedness, and access" and anticipates that the US could capture up to 5.4 % of additional GDP over ten years. The World Bank's "Strengthening AI Foundations" report, in turn, identifies investment in AI data centers as one of the essential "4Cs" underpinning AI readiness: connectivity, compute, context, and competency.

K Industry survey: Key differentiating factors for data centers in 2030[1]

What would be the three most important requirements or differentiation factors for a data center in Europe in 2030?



1 Question type: Select and rank up to three choices; sample size: 70
2 Data normalized by following method: Scores assigned Fibonacci numbers (#1:8, #2:5, #3:3); highest overall score set at 100 (indexed)

Source: Roland Berger industry survey (2026)

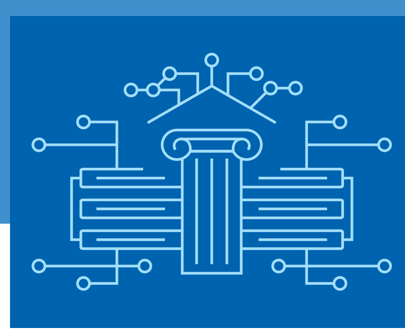
The link between data centers and AI-enabled growth is not automatic, but it is increasingly tight. Domestic AI data centers are, in several respects, a prerequisite for a competitive domestic AI ecosystem. They make it possible to train and deploy models in full compliance with local regulations, including for sensitive public sector and critical infrastructure use cases. They lower latency and improve reliability for local firms and institutions building AI-enabled products and services. They also provide a platform around which a broader ecosystem of model developers, cloud and edge providers, integrators, and application startups can cluster. Without competitive local compute, Europe risks missing out on future GDP growth as AI value creation – and associated high-productivity jobs in software, data, and R&D – moves toward other regions. A development in June 2026 brought these arguments into sharp relief. Citing national security, the US Commerce Department barred any foreign national – including Anthropic's own non-US employees – from accessing Fable 5 and Mythos 5, the company's most advanced models. Unable to separate foreign from domestic users, Anthropic disabled both models for all customers. Even though access was restored weeks later, the episode illustrates the second and third arguments above: a jurisdiction cannot claim digital sovereignty if access to frontier technology can be curtailed overnight by a foreign government, and future AI-driven growth will diverge if some economies keep access to the most capable systems while others lose it. For Europe, the incident read as a wake-up call – a further reason to build domestic data centers and a home-grown AI capability that cannot be switched off from outside.

Data center 101

Questions most people want to ask but never do

Do data centers create jobs – and why do policy-makers care?

Data centers generate sizeable employment during construction, drawing on engineering, electrical, and mechanical trades – but once operational, hyperscale facilities are highly automated and require only a small permanent workforce. Compared to traditional industries, sustained job creation per euro invested is negligible: a EUR 5 billion AI campus might support a few hundred permanent roles, against tens of thousands for an equivalent automotive plant. Policy-makers are nonetheless deeply invested in the sector because data centers are the physical backbone of the AI revolution – and because participating in that revolution on sovereign, regulation-compliant terms requires local infrastructure. Without it, the economic gains from AI-driven productivity growth risk flowing elsewhere.



4.2/ From policy ambition to structural reform

From a European perspective, the urgency of this challenge is increasingly recognized. Mario Draghi's 2024 report on European competitiveness warns that "competition for computing power and lack of investment in connectivity could soon translate into a digital bottleneck." In March 2026, Philip R. Lane, Member of the Executive Board of the European Central Bank (ECB), highlighted "a striking, and widening, gap between the Euro Area and the United States" when comparing AI-related investment levels. Our own survey of industry experts shows that data center location decisions in Europe are perceived as strongly influenced by public policy frameworks, yet most current EU-level measures are seen as only "slightly to moderately" supportive. ▶L

In response, the European Union (EU) and its member states have begun to craft policies to foster domestic AI infrastructure. The InvestAI initiative, announced in February 2025, proposes a EUR 20 billion package to support the establishment of five AI "gigafactories" for large model training, complementing the European High Performance Computing Joint Undertaking (EuroHPC) and other programs. Initial interest was substantial – 76 expressions of interest were submitted – but the formal call for proposals, originally planned for December 2025, had not been launched at the time of writing. The delay illustrates the tension between political ambition and administrative execution. At the national level, subsidy and regulatory initiatives vary widely across member states and have often added complexity rather than accelerating the build-out.

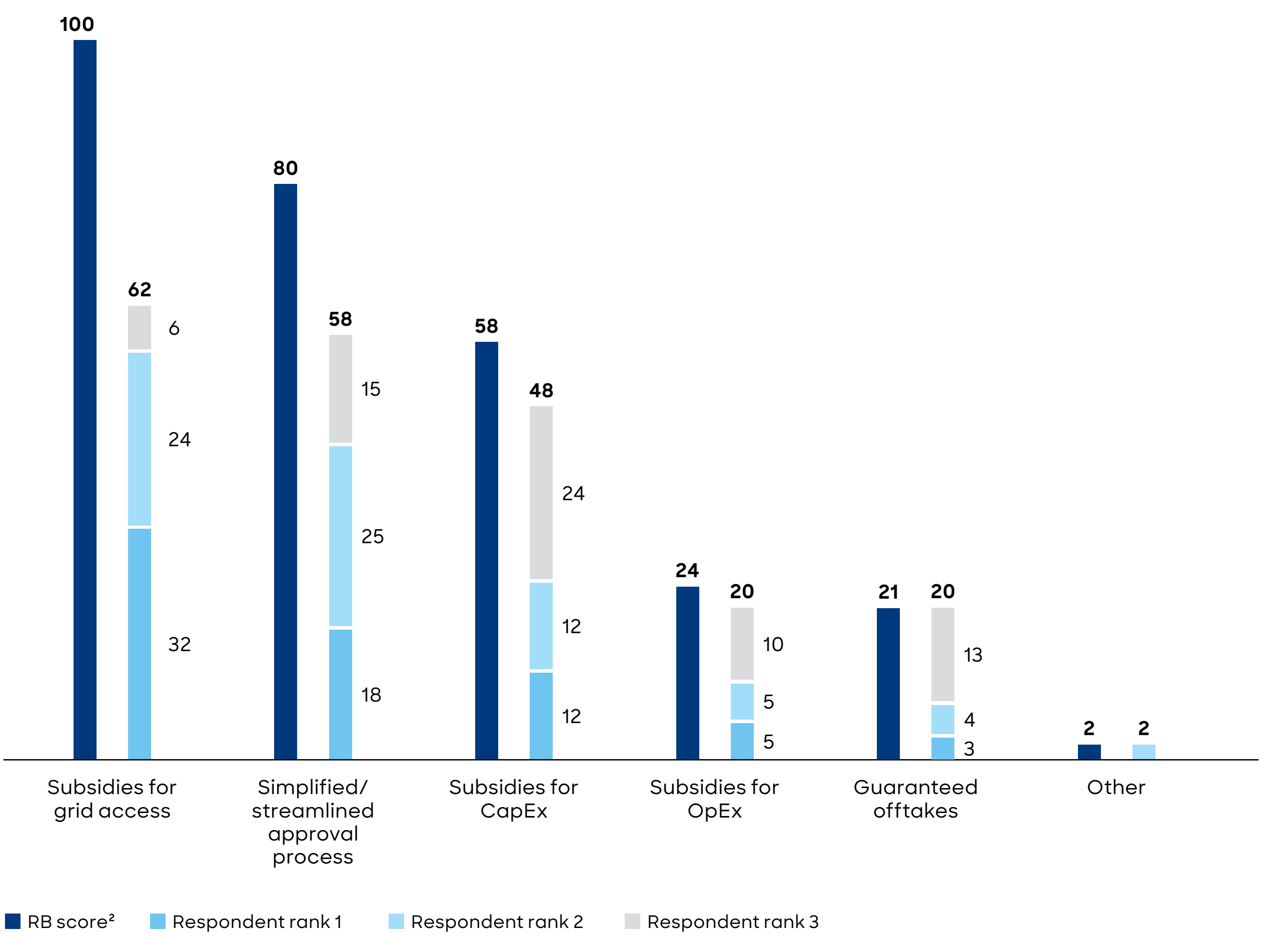
As our analysis and survey results underscore, financial incentives alone cannot resolve the structural obstacles that are holding back Europe's build-out – namely, regulatory complexity, grid and energy access, and the absence of home-grown hyperscalers. Addressing these obstacles requires changes to energy and planning frameworks: anticipatory grid planning, dedicated high-capacity nodes for digital infrastructure, streamlined permitting, and technology-neutral efficiency rules – if any. Financial tools such as InvestAI and EU-level champions funds can supplement the weaker private investor base, but only if deployed quickly and with clear, predictable criteria that invite industrial innovation rather than impose procedural delays.

There is precedent for rapid policy correction when strategic interests are at stake. The current German administration moved the EU toward "technology neutrality" in vehicle powertrains within approximately six months, reopening the door for internal combustion engines after the initially agreed 2035 ban. The same decisiveness is needed on AI infrastructure. Germany, and by extension Europe, would benefit from taking a more enabling regulatory stance on data center infrastructure. The German government's June 2026 draft law easing parts of the Energy Efficiency Act is a step in the right direction, but industry voices argue it still falls short of making Germany a genuinely competitive data center location. Revising the regulation further would be a concrete signal that Europe intends to build AI infrastructure, not just regulate AI.

The correction risk discussed in chapter two should not, in public policy debates, be read as a reason to slow down. A temporary adjustment in valuations or order books along the value chain does not change the structural risk of Europe falling further behind, nor does it diminish the long-term importance of domestic AI and AI infrastructure. Despite the dot-com bust of the early 2000s, the continent spent the following two decades trying to catch up on digitalization and the IT economy – a gap that shaped its technology sector for a generation. The cost of repeating that experience in AI compounds with every year of inaction.

L Industry survey: Ranking of subsidy importance[1]

Assuming you were the President of the European Commission and could spend EUR 50 billion to support data centers, how would you rank the importance of the following subsidies?



1 Question type: Select and rank up to three choices; sample size: 70
2 Data normalized by following method: Scores assigned Fibonacci numbers (#1:8, #2:5, #3:3); highest overall score set at 100 (indexed)

Source: Roland Berger industry survey (2026)

Conclusion

Data centers are the basis for the ongoing AI revolution, similar to the steam engines that powered the Industrial Revolution. The data center build-out itself can drive GDP growth, but more importantly, it enables the AI-driven productivity gains that will determine which economies grow fastest, which companies capture value, and which nations have genuine strategic autonomy. The positions in this build-out race can still be shaped – but not for long. The US is pulling ahead, and Europe risks being left behind by constraints that are structural and often self-imposed. For operators and investors, the priority is capturing growth combined with resilience. For policymakers, particularly in Europe, the window to act is narrowing. The countries that get this right will capture a disproportionate share of the AI era's economic gains.

Key ecosystem stakeholder questions

#1 AI infrastructure momentum

Is the AI infrastructure cycle at its peak – or is this still the right moment to enter?
The 2025–2026 period marks an unprecedented step change in the global data center build-out. Hyperscaler capital expenditure is set to roughly double from approximately USD 400 billion in 2025 to around USD 750 billion in 2026 – with the bulk representing genuine volume expansion, not merely price inflation. The Roland Berger model projects annual capacity net additions of around 25 to 30 GW in the coming decade. Moreover, the pace of GPU innovation and release cycles approaching an annual cadence have fueled an ongoing debate over the true useful life of a GPU. Our model anticipates replacement activity overtaking net new additions within roughly five years. Since each GPU generation brings its own power and cooling requirements, the replacement cycle likely extends beyond compute. Rapid technology innovation across every discipline is continuously reshuffling the value chain, creating entry points for players who can displace incumbents on performance or efficiency grounds. The AI infrastructure cycle is not at its peak – it is still accelerating.

#2 Ecosystem risks

What risks do data center ecosystem participants face – and how material are they?
Two distinct risk categories define the challenge for data center ecosystem participants. The first is competitive: AI data centers are tightly integrated systems orchestrated by hyperscalers and leading compute vendors against clear performance criteria – compute density, power efficiency, uptime, and time to market. Any supplier overtaken by a more capable alternative risks rapid displacement. The second risk is cyclical: despite structural demand strength, the build-out is partially financed ahead of revenues, with AI labs spending multiples of their income on infrastructure and hyperscaler free cash flow tightening. Grid access bottlenecks, regulatory shifts, or a change in investor sentiment could trigger a short-term correction. Participants who build resilience into their operating and financial structures now will be best placed to sustain momentum through any temporary moderation.

#3 Public policy imperative

Why does data center infrastructure demand urgent public policy attention - and what is the cost of inaction?

Data centers are the foundational compute infrastructure for the AI revolution – a transformation already reshaping GDP trajectories and competitive positioning across economies. Economies that participate actively in AI are projected to capture a disproportionate share of productivity gains over the coming decade. That participation requires local data center infrastructure: to ensure compliance with frameworks such as GDPR, protect data sovereignty, and keep high-value activity within reach of local regulation. The cost of inaction is already visible – mandatory heat reuse obligations, delayed grid connections, and regulatory complexity are deterring investment and slowing the European build-out. Every year of inaction widens a gap that, as the post dot-com era demonstrated, can take a generation to close.

#4 Implications of diverging regional trajectories

Given diverging regional build-out speeds and local requirements, what are the implications for ecosystem participants?

The data center market is global in its technology foundations but distinctly local in its execution. For technologies within the data hall, standards are set by hyperscalers and chip vendors, and product qualification travels across geographies. Beyond the data hall, local conditions take over: grid connection timelines, permitting frameworks, energy pricing, and regulatory obligations vary materially and directly influence site design and equipment choices. For ecosystem participants, this demands a dual strategy – maintaining a globally competitive technology roadmap while adapting to regional realities. Europe's slower build-out and more complex local requirements, while presenting headwinds, also open specific entry points for infrastructure specialists, EPC contractors, and suppliers with established local presence that purely US-centric players are structurally less equipped to address.

METHODOLOGY

The Roland Berger Data center Model draws on multiple external sources – including forecasts from the International Energy Agency (IEA) and Epoch AI – to establish a baseline and validate consensus estimates for future capital expenditure (CapEx), revenue trajectories, and manufacturing ramp-up timelines across relevant hyperscalers and chipmakers. Key assumptions, including IT equipment replacement rates, are stress-tested against input from industry experts and cross-referenced with the findings of our proprietary survey.

This study also draws on an independent survey of 70 senior decision-makers spanning the data center value chain, including data center operators, information technology, energy, power systems and utilities, and construction and engineering professionals. Half of respondents directly operate data center facilities – hyperscale, colocation, enterprise, or internal – while the remaining half supply products and services to data center owners and operators, encompassing power, cooling, IT, security, construction, and software and monitoring solutions.

All respondents represented organizations across at least ten countries in Europe, with ~80 % holding direct involvement in purchasing or strategic decisions related to data center infrastructure and services. This ensured participants brought both the relevant domain expertise and decision-making authority needed to provide informed perspectives on the topics examined. Participating organizations varied in scale, with respondents drawn from small (26 %), medium (39 %), and large enterprises (36 %).

Glossary

AC	Alternating current
AI	Artificial intelligence
AWS	Amazon Web Services
CAGR	Compound annual growth rate
CapEx	Capital expenditure
DC	Direct current
ERF	Energy reuse factor
ECB	European Central Bank
EuroHPC	European High Performance Computing Joint Undertaking
EU	European Union
EUR	Euro
FLOPS	Floating point operations per second
GPU	Graphics processing unit
GDP	Gross domestic product
GDPR	General Data Protection Regulation
GW	Gigawatt
kW	Kilowatt
LLM	Large language model
MEP	Mechanical, electrical, and plumbing
MW	Megawatt
NIS2	Network and Information Security Directive 2
OpEx	Operating expenditure
PUE	Power usage effectiveness
R&D	Research and development
UK	United Kingdom
US	United States
USD	United States dollar

Credits

AUTHORS

Edeltraud Leibrock
Senior Partner | Germany
edeltraud.leibrock@rolandberger.com

Nikolaus Lehmann
Partner | Germany
nikolaus.lehmann@rolandberger.com

Urs Neumair
Partner | Germany
urs.neumair@rolandberger.com

Wim D'Hondt
Director | Canada
wim.dhondt@rolandberger.com

Siyi Hao
Principal | UK
siyi.hao@rolandberger.com

07.2026
ROLANDBERGER.COM

FURTHER READING

Power, not GPUs, is the limiting factor for AI data centers

CONTACTS

Isaac Chan | United States
isaac.chan@rolandberger.com

Kyle Gordon | United States
kyle.gordon@rolandberger.com

Jonas Kick | Germany
jonas.kick@rolandberger.com

Michael Knott | United Kingdom
michael.knott@rolandberger.com

Marie Lê de Narp | France
marie.ledenarp@rolandberger.com

Yuzuru Ohashi | Japan
yuzuru.ohashi@rolandberger.com

Carsten Rossbach | Germany
carsten.rossbach@rolandberger.com

Daniel Tao | United States
daniel.tao@rolandberger.com

Nicolas Teisseyre | France
nicolas.teisseyre@rolandberger.com

Alex Xu | Greater China
alex.xu@rolandberger.com

Andrew Yi | South Korea
andrew.yi@rolandberger.com

Ron Zheng | Greater China
ron.zheng@rolandberger.com

ROLAND BERGER is the only leading global strategy consultancy of European origin. The firm combines deep industry expertise with broad experience across core management functions and transformation programs. Founded in 1967 and headquartered in Munich, Roland Berger supports companies worldwide in shaping and executing complex transformations – from strategic repositioning and performance improvement to the development and application of data-driven, AI-enabled solutions. The firm is committed to embedding sustainability across all its projects. In 2025, Roland Berger generated revenues of over EUR 1 billion.

Publisher

Roland Berger GmbH
Sederanger 1
80538 Munich
Germany
+49 89 9230-0

26_2011_STU