

STIFEL | IRIS

INTELLIGENCE • RESEARCH • INSIGHTS • SERVICE

ENERGY-AWARE AI

CONFRONTING THE POWER-HUNGRY REALITY OF SCALING AI

FEBRUARY 2025

INDUSTRY BRIEF – DATA CENTRES



Executive summary

In this Industry Brief, we explore the energy dilemma caused by the unprecedented data centre buildout. The emergence of Generative AI has ignited an arms race, as scaling laws fuel relentless demand for more compute infrastructure... and the risk of a power crunch. AI's future demands a redefinition of scalability, not just in terms of bigger models or faster chips, but also the energy systems powering them. If AI is to scale sustainably, the industry must become as creative with energy as it has been with algorithms.

AI's race for scale

The AI era is triggering a seismic shift in global capex priorities, with data centres poised to attract nearly USD500bn in 2025 alone. Despite challenges such as ROI pressure, the AI investment cycle shows no signs of slowing. This move to compute infrastructure is fuelled by hyperscalers and AI servers pushing the boundaries of data centre energy density and efficiency. From Nvidia's cutting-edge designs to liquid cooling breakthroughs, the race for computational dominance will shape the backbone of tomorrow's AI-driven world.

The energy reckoning of data centres

Data centres—once the silent workhorses of IT— are being transformed by AI into power-hungry behemoths. By 2030, their electricity use will double past 1000TWh, straining energy systems. With renewables already under pressure to scale up to decarbonise the grid, energy-intensive AI adds to the challenge. By embracing renewable and nuclear options—and integrating new tech like small modular reactors (SMRs) and battery energy storage systems (BESS)—the AI industry can chart a sustainable path forward.



Glossary

- **Artificial General Intelligence (AGI):** form of AI that can understand, learn, and apply intelligence across a wide variety of tasks, similar to human cognitive abilities.
- **AI clusters:** group of interconnected servers or machines used to run artificial intelligence workloads in parallel, increasing processing power.
- **AI pod:** in data centres, refer to modular, scalable units designed to support AI workloads and integrating AI accelerators, high-speed networking, and optimized cooling and power systems.
- **AI workload:** tasks or processes that require computational power to train, run, or analyse AI models.
- **Air cooling:** method of cooling hardware by circulating air to dissipate heat. In a data centre, it typically involves computer room air conditioners (CRACs) or computer room air handlers (CRAHs) to manage airflow efficiently.
- **Agentic AI:** type of AI system that act autonomously – as its own agent, making decisions and performing tasks without human intervention.
- **Automatic Transfer Switches (ATS):** devices that can automatically transfer power between primary and backup sources when an outage is detected.
- **Battery Energy Storage Systems (BESS):** set of technologies that store electricity in batteries for later use, enabling the management of energy supply and demand by discharging stored power when needed.
- **Busbars:** electrical conductors used to distribute power to multiple circuits within a system, often found in data centres.
- **“Chinchilla” scaling:** neural scaling law introduced by DeepMind, which optimizes the trade-off between model size and training data for large language models. It suggests that, given a fixed compute budget, models should be smaller but trained on significantly more data compared to previous approaches like OpenAI's GPT-3.
- **Colocation:** practice of housing servers and data centre equipment in third-party facilities for better resource management and reliability.
- **Chip-on-Wafer-on-Substrate (CoWoS):** advanced 2.5D packaging technology developed by TSMC that integrates multiple chips onto a single wafer. This approach has proven key to enhance bandwidth, reduces power consumption, and improves performance for high-performance computing (HPC) and AI applications.
- **Computer Room Air Handler (CRAH):** units used for cooling in data centres by circulating air and removing heat.
- **Central Processing Unit (CPU):** core processing unit of a computer that carries out instructions from programs by performing basic arithmetic, logic, control, and input/output operations, typically using multiple cores for parallel processing.
- **Data-lake:** centralised repository in a data centre that stores vast amounts of data in its native format. Unlike traditional databases, it allows flexible data ingestion and supports advanced analytics, machine learning, and big data processing.
- **Day-ahead electricity prices:** prices at which electricity is bought and sold for delivery in the next day, determined through a market auction based on supply and demand for each hour.
- **Dynamic Random Access Memory (DRAM):** type of memory used in computers and other devices to store data temporarily while the system is powered on.
- **Data Processing Unit (DPU):** specialized processor designed to offload data-intensive tasks from the Central Processing Unit (CPU), particularly in networking and storage systems.
- **Edge data centre:** small-scale data centre located closer to the end-user to reduce latency. These centres process data locally before sending it to larger cloud or core data centres.
- **Foundation models:** large pre-trained AI models that can be fine-tuned for specific tasks or applications.
- **Foundries:** companies or facilities that manufacture semiconductors and microchips for various applications.

Glossary

- **Generative AI (GenAI):** AI systems that can create new content, such as text, images, or videos, based on learned patterns from existing data.
- **Graphic Processing Units (GPU):** processors designed to accelerate graphics rendering and parallel computations, particularly important for AI and machine learning.
- **Grey space:** the areas that house the infrastructure necessary for the operation of the data centre but do not directly contain IT equipment. This notably includes electrical rooms, cooling systems, backup generators, and UPS systems that ensure the reliability of the "white space" (where IT racks and servers are located).
- **High Bandwidth Memory (HBM):** type of memory technology designed for faster data transfer rates, often used in AI and high-performance computing applications.
- **High-Performance Computing (HPC):** use of powerful computing systems to perform complex calculations and simulations, typically used in AI and large-scale data analysis.
- **Hyperscale data centre:** large, efficient facilities designed to support massive computing and storage needs, optimised for scalability, high availability, and energy efficiency, typically used by cloud providers and tech giants.
- **Inference:** stage in AI where a trained model applies learned patterns to new data in order to generate predictions, classifications, or decisions.
- **Levelised Cost of Energy (LCOE):** average cost of producing energy over the lifetime of a power generation asset, accounting for all costs and output.
- **Liquid cooling:** cooling technique using liquids to efficiently dissipate heat from electronic components, typically more effective than air cooling.
- **Load variation:** fluctuations in the demand for power or computing resources, which can affect system performance and efficiency.
- **Logic die:** portion of a semiconductor chip that handles the logic and processing operations, typically in CPUs and GPUs.
- **Large Language Model (LLM):** AI model trained on massive amounts of text data to understand and generate human-like language.
- **Megacampuses:** large data centres or tech facilities designed to accommodate significant amounts of computational and storage capacity.
- **Micro-modular designs:** small, self-contained data centre units that can be easily scaled and customised, offering flexibility in capacity and rapid deployment.
- **Moore's Law:** observation that the number of transistors on a microchip doubles approximately every two years, leading to increased computing power.
- **MV Transformers:** medium-voltage transformers used in electrical systems to convert voltage levels for optimal distribution.
- **Power Distribution Unit (PDU):** device often used in data centres and server rooms to distribute electrical power to multiple pieces of equipment.
- **Power Purchase Agreement (PPA):** long-term contract between a power producer and consumer to buy electricity at an agreed price.
- **Primary energy:** refers to natural energy sources that have not yet been converted or processed into another form, such as coal, oil, natural gas, sunlight, wind, or geothermal energy.
- **Power Usage Effectiveness (PUE):** metric used for measuring the energy efficiency of a data centre, calculated by dividing total facility energy consumption by the energy used by IT equipment.
- **Service Level Agreement (SLA):** formal contract between a service provider and a customer that defines the expected level of service, along with penalties or remedies if these standards are not met.
- **Small Modular Reactors (SMR):** compact nuclear reactors that are smaller and more flexible than traditional large reactors.
- **Switchgear:** electrical equipment used to control, protect, and isolate electrical circuits in power systems, ensuring safety and reliability.

Glossary

- **Synthetic data generation:** creation of artificial data that simulates real-world conditions, used to train AI models or test systems.
- **Test—time compute scaling:** adjustment of computational resources during the deployment of an AI model to optimise performance and efficiency.
- **Total Cost of Ownership (TCO):** complete cost of owning and operating a system, including purchase, maintenance, energy, and operational costs over its lifecycle.
- **Tensor Processing Unit (TPU):** specialised processor developed by Google for accelerating machine learning workloads, particularly for deep learning.
- **Training:** process of teaching an AI model by feeding it large volumes of data and adjusting its parameters to improve its performance.
- **Transformers:** electrical device used to step up or step-down voltage levels while maintaining the same frequency.
- **Uninterruptible Power Supply (UPS):** backup power system that provides temporary electricity during power outages to prevent downtime or data loss.
- **Wholesale data centre:** large-scale facility that leases out large portions or entire data centre spaces to clients requiring extensive computing resources.
- **White space:** the area where IT equipment (such as servers, storage, and networking devices) is installed, typically within server rooms or data halls. It contrasts with grey space, which houses infrastructure like cooling, power, and backup systems.



01

MEASURING THE AI DATA CENTRE CAPEX SPLURGE

HARDWARE IS EATING THE WORLD

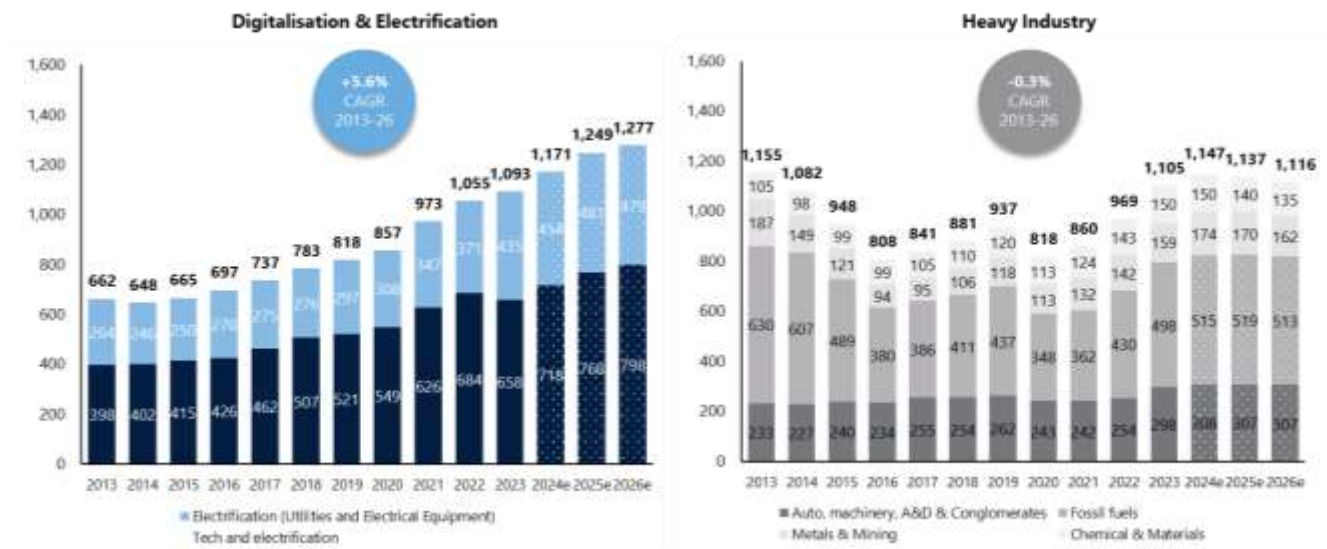
The AI era is redefining global capex priorities, shifting the spotlight to data centres, set to draw near USD500bn in capex in 2025. This shift reflects a broader capex pivot from traditional industrial infrastructure to the realm of digital assets, alongside a reallocation of IT spend from telecom networks to compute infrastructure. The advent of GenAI has sparked a trillion-dollar capex race for computational dominance among big tech companies, by unlocking economies of scale in model training and promising predictable performance progress with scaling laws. Yet, beneath the surface of this transformation lie significant challenges: mounting ROI pressures, the energy constraints of hyperscale infrastructure and the inherent fragility of scaling laws. Despite these hurdles, we believe the investment cycle will endure. The question is not if the AI capex splurge will continue, but at what scale and pace it will continue.

AI is accelerating the shift to a capex cycle focused on compute infrastructure

The AI boom has sped up the global shift in capex from heavy industries and fossil fuels to digital and electrical infrastructure. The IT capex cycle itself is continuing to transition from telco networks to computing infrastructure like data centres and advanced fabs. In 2024, data centre capex alone is set to have reached ~USD375bn, a 44% increase from 2023. A select group of hyperscalers and semiconductor manufacturers emerge as the new infrastructure giants, leading this capex cycle.

Global capex is shifting decisively from heavy industries towards IT and electrification. Our analysis of capex at the 7,500 largest listed companies shows heavy-industry spending has stalled since 2013, with a -0.3% CAGR to 2026e, trailing the commodity supercycle peak. Meanwhile, IT and electrification capex shows a +5.6% CAGR over the same period, exceeding heavy industry levels in 2020. The allocation of investment is undergoing a profound transformation: by 2026e, IT and electrification should reach 118% of heavy-industry capex levels, up sharply from 59% in 2013. Historically dominant sectors—fossil fuels, capital goods, and mining—have seen a marked erosion in their capex share. While IT has expanded steadily, electrification’s upswing is more recent, marking the early stages of an electricity supercycle and of the convergence of digitalisation and electrification trends. Power-grid capex alone rose from USD300bn in 2020 to nearly USD400bn in 2024 and is forecast at USD600bn by 2030, according to the IEA.

Fig. 1 – Tracking global capex: sector trends from 2013 to 2026e (USDbn)

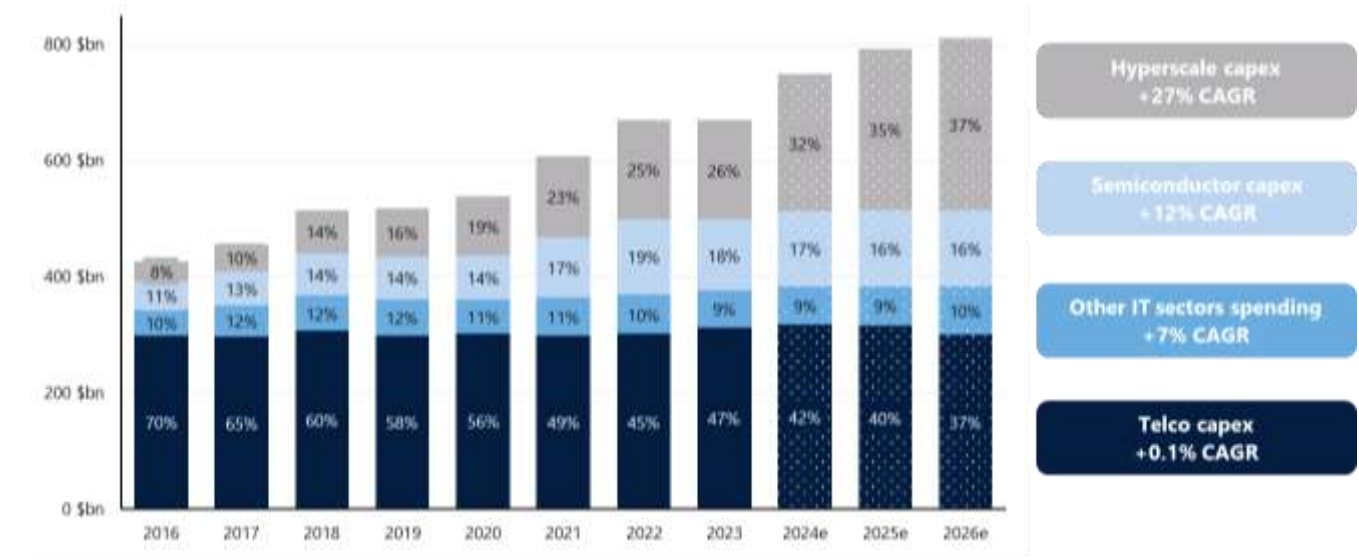


Source(s): Stifel* analysis, January 2025 Bloomberg consensus on the 7,500 largest listed firms with available data

Our analysis of IT capex underscores a structural shift from telecoms to data centres and semiconductor manufacturing, effectively reallocating investment from networks to computing power. Between 2016 and 2026e, telecom capex—70% of IT sector capex in 2016—is expected to stagnate at ~USD300bn, reflecting a subdued +0.1% CAGR. This lacklustre growth mirrors sustained pressure on telco revenues, which continues to constrain reinvestment capacity. In contrast, overall IT infrastructure capex is projected to expand at a robust +7% CAGR (2016–26e), driven by semiconductors (+12% CAGR) and Big Tech-led hyperscale capex (+27% CAGR). By 2025, these two segments are forecast to account for over 50% of total IT sector capex, upending previous allocation patterns.

Hyperscale capex growth has emerged as the defining trend in IT investment over the past decade, projected to account for 37% of total IT capex by 2026e. In our analysis, we isolated the seven largest hyperscalers—Amazon, Microsoft, Meta, Alphabet, Oracle, Alibaba, and Tencent—and found they are consistently expanding their capex at a remarkable CAGR of +27% from 2016 to 2026e. In 2016, hyperscale capex accounted for just 8% of global IT capex; by 2023, this had tripled to 26% and is forecast to reach 37% by 2026e, driven by data centre expansion amid rising cloud adoption and compute-heavy AI technologies.

Fig. 2 – IT infrastructure spending over 2016-2026e: from networks to computing



Source(s): Stifel* analysis, January 2025 Bloomberg consensus on the 7,500 largest listed firms with available data

Recent years have seen Big Tech entrench its position as a global capex heavyweight. In 2025, Microsoft, Amazon, Alphabet, and Meta will lead listed firms in capex, with IT accounting for half of the top 20 spenders. The concentration is striking: the top 10 IT firms account for 56% of sector capex versus 22% in heavy industries. The four hyperscalers alone (Alphabet, AWS, Meta, Microsoft) are expected to invest USD261bn in 2025, mainly in data centres—a 108% jump in two years. Capex growth will be largely offset by rising revenue, from USD402bn in CY19 (excl. non-AWS Amazon sales) to an estimated USD1,284bn in CY27e (+16% CAGR). However, the capex-to-sales ratio is expected to climb, averaging 24% in 2024–27 vs. 16% in 2019–23, with Microsoft’s rising the most, from 15% to 30%. Clustering is apparent in the semiconductor sector as well, with the five key logic foundries and memory players—TSMC, Samsung, Intel, SK Hynix, and Micron—projected to spend USD125bn in 2025, dominating sector investments.

Fig. 3 – Top 10 highest-capex companies by sector (USDbn)

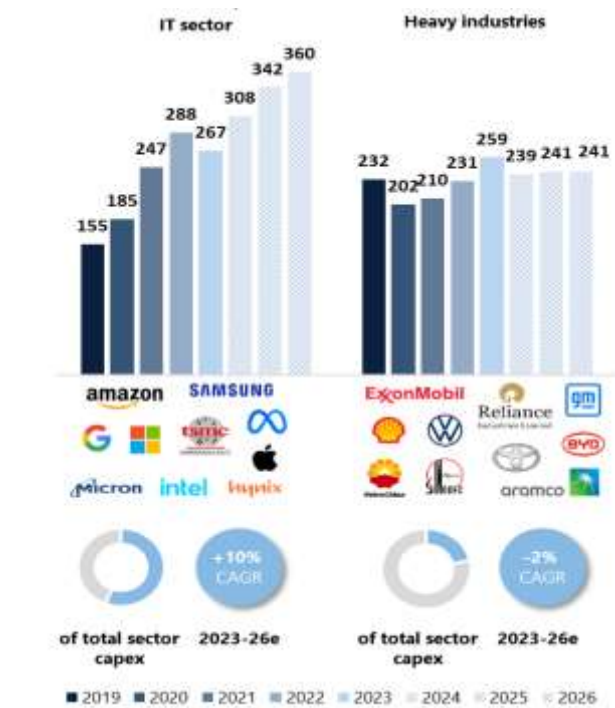
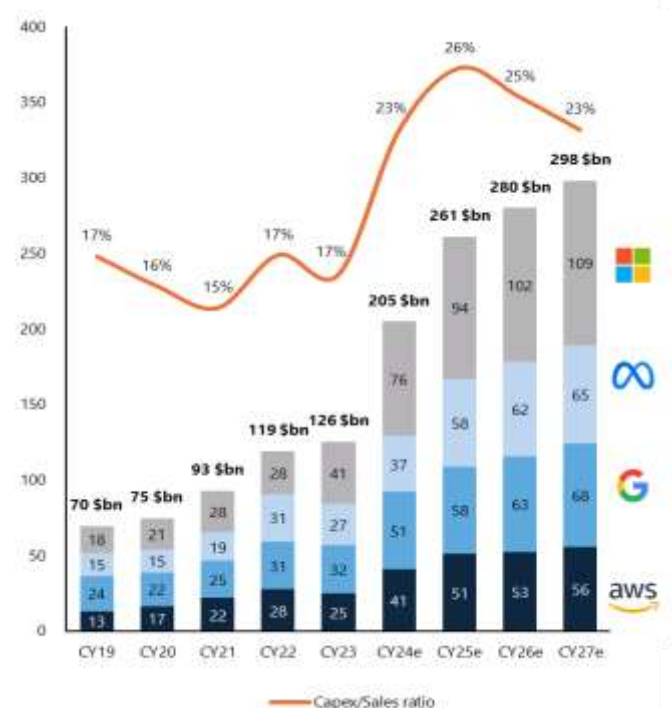


Fig. 4 – Hyperscale¹ capex and capex-to-sales ratio 2019-2027e



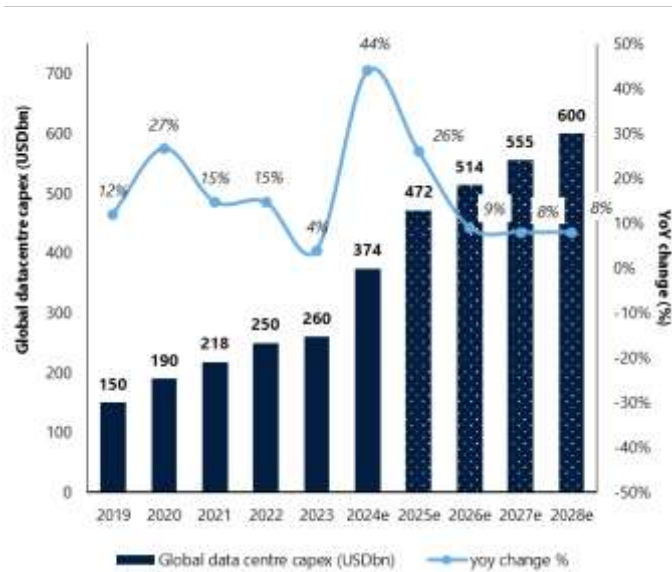
Source(s): Stifel*, Bloomberg consensus

(1) For Amazon, only AWS capex are included.

Ongoing investments in AI infrastructure are driving a sharp increase in data centre capex, which we estimate have reached USD374bn in 2024. We estimate that ~55% of this figure will come from US hyperscalers, ~20% from colocation and telcos and 10% from Chinese cloud services providers. Precise estimates remain challenging due to limited disclosures and inconsistencies in defining data centre-related spending. We believe data centre capex is generally underestimated, as most trackers rely on publicly available data from the largest cloud service providers. However, the capex growth trajectory is clearer, as it is currently driven predominantly by hyperscale data centres. As the backbone of the digital economy, data centre capex recorded a mid-teen CAGR over 2018–2023, fuelled by surging demand for cloud services and internal investments by large tech firms. According to Dell'Oro Group, global data centre capex grew 44% in 2024, supported by hyperscalers' aggressive AI-related investment. This followed subdued 4% growth in 2023, constrained by supply chain bottlenecks for accelerated computing and reduced general-purpose server investment.

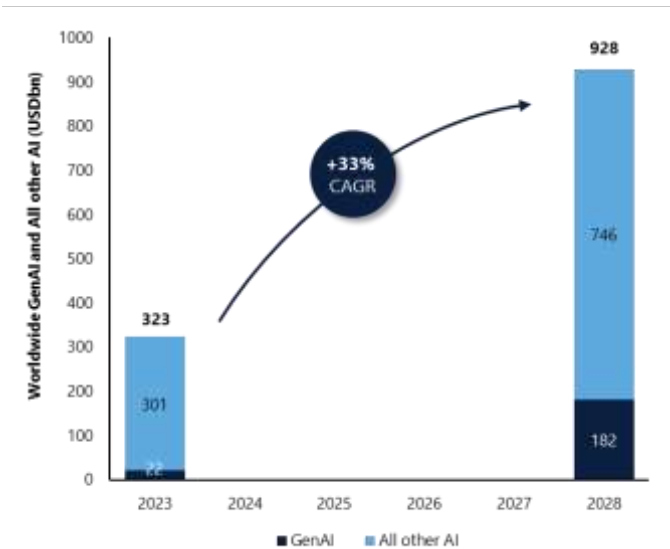
We anticipate another strong year in 2025, with data centre capex rising 26% yoy to nearly USD500bn. US hyperscalers will sustain the momentum, with an estimated 28% growth yoy, following an exceptional 63% increase in 2024. Microsoft and Meta will drive the largest absolute increase, having confirmed FY25 capex of over USD80bn and USD60–65bn, respectively, versus USD55.7bn and USD37.3bn in 2024. We expect capex growth to moderate in CY26e–CY28e to high-single digit growth to absorb the sharp rise from CY24–CY25. However, recent developments hint at a potential reacceleration, notably with the Stargate project, announced by Donald Trump in January 2025. The initiative targets up to 20 large-scale AI data centres in the U.S., with an initial USD100bn commitment and a total investment ambition of up to USD500bn by 2029. That said, details of the project remain opaque, and only a fraction of the funding required for the first phase (USD100bn) appears to be secured at this stage.

Fig. 5 - Data centres capex 2020-2028e



Source(s): Stifel*, Dell'Oro Group, Omdia

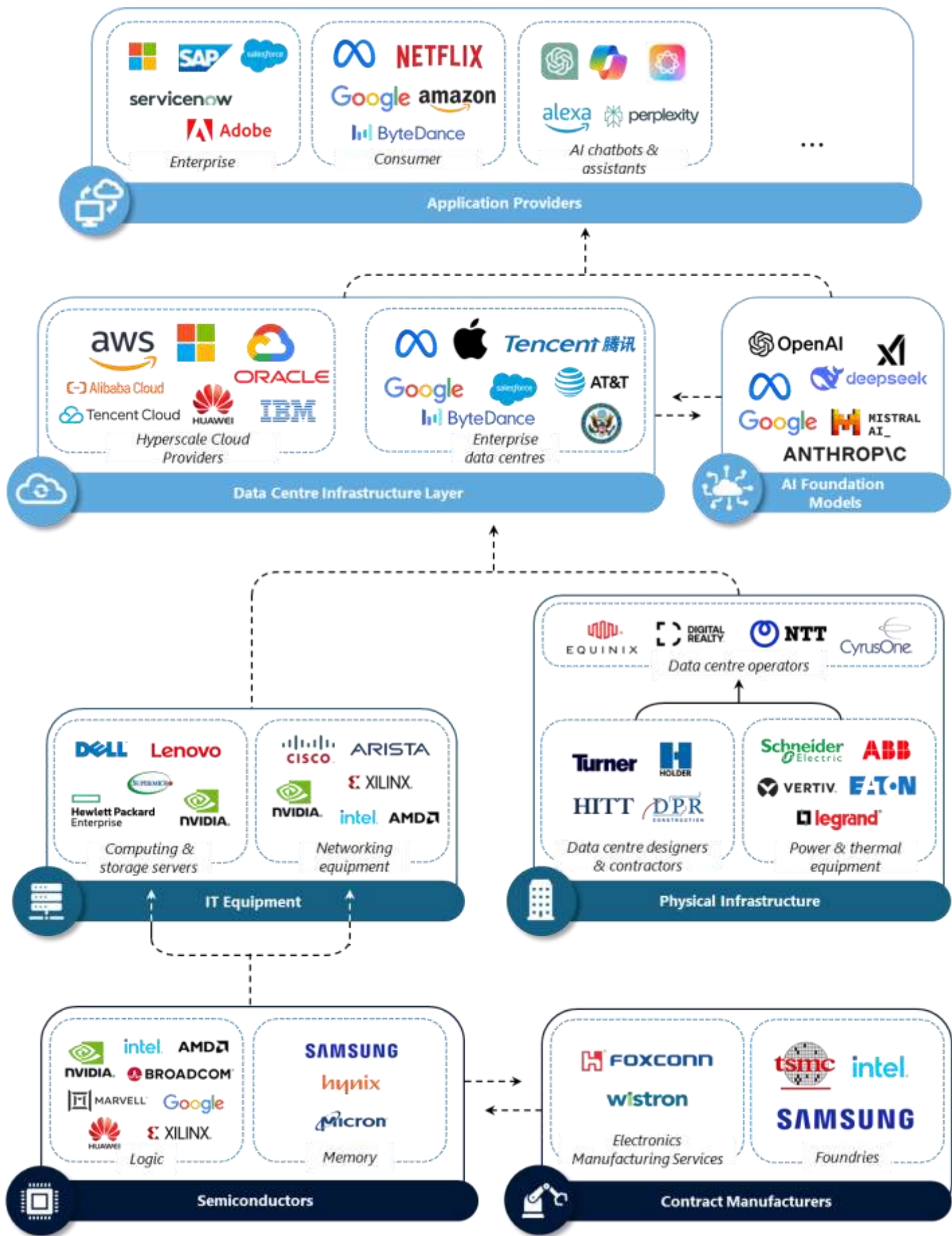
Fig. 6 – Worldwide AI spending (USDbn)



Source(s): IDC, Worldwide AI IT Spending Forecast, Oct 2024

The proliferation of AI workloads is compressing IT hardware renewal cycles, sustaining a prolonged capex cycle. Data centres are rapidly depreciating assets, as IT equipment needs to be replaced relatively often. Servers with accelerators and associated networking equipment constitute the largest cost driver for AI data centres, with replacement or upgrades typically occurring every 3–6 years. Despite Nvidia's dominance, the AI processing landscape remains highly competitive, which should continue to spur innovation, reinforcing the need for rapid hardware refresh cycles. As Moore's Law is fundamentally slowing down, the future of AI infrastructure is increasingly shifting from raw compute power to system-level optimisation—not solely at the transistor or chip level, but across entire data centres—, unlocking significant opportunities for continued performance gains in AI computing hardware. Hyperscalers may also challenge Nvidia's incumbency by increasingly adopting custom silicon, exemplified by Google's TPU and AWS's Trainium/Inferentia. Large-scale AI training workloads necessitate retrofitting existing facilities with high-density racks, enhanced power distribution and advanced cooling systems, further intensifying capex demands.

Fig. 7 – Snapshot of the data centre value chain



Source(s): Stifel* IRIS

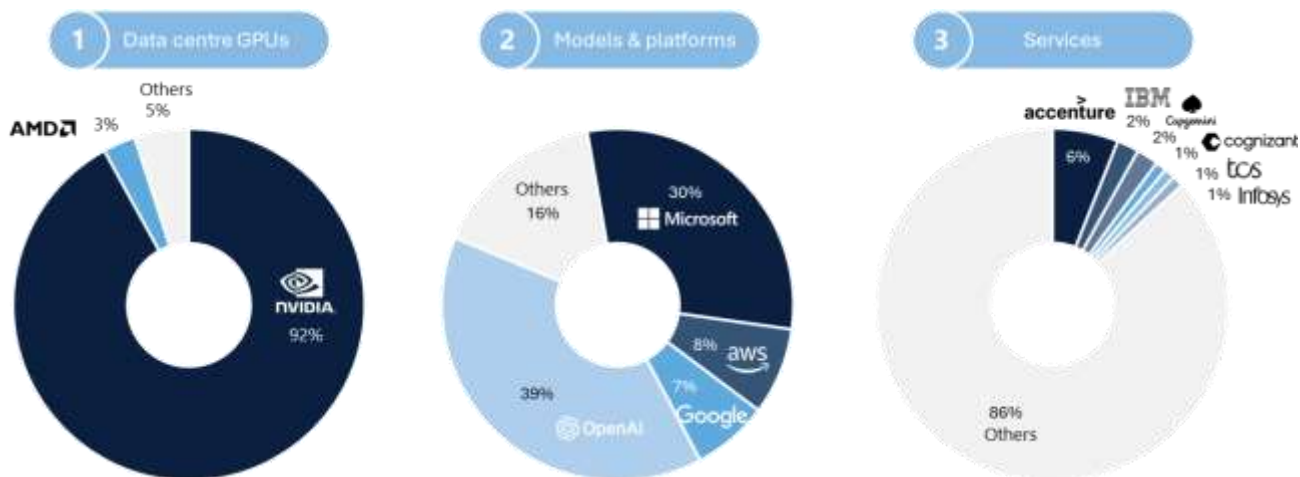
GenAI's race for capital and scale

The ongoing AI boom has fundamentally reshaped AI economics by unlocking economies of scale in model training, driving concentration in AI infrastructure and model development. Scaling laws have introduced a "Moore's law"¹ analogue for AI, promising greater predictability in performance progress and initiating a capital-intensive arms race in computational supremacy among a few Big Tech firms. This contest for compute power has elevated data centres to critical strategic assets.

GenAI and large language models (LLMs) have not only radically disrupted AI technology but also fundamentally changed the economics of the AI industry. Although AI is not entirely new—its transformative potential was evident well before the November 2022 launch of ChatGPT—the introduction of the Transformer architecture marked a pivotal turning point. Widely attributed to the 2017 paper "Attention Is All You Need" published by Google research scientists, this architecture superseded previous state-of-the-art neural network designs and paved the way for the rise of LLMs. However, beyond marking a trend breaker in AI performance progress, the transition to GenAI has prompted a major shift in AI economics with the emergence of foundation models: large-scale, general-purpose architectures that underpin a multitude of downstream applications. They reshaped cost structures across the AI industry due to the economies of scale in model training:

- **High upfront training costs, lower marginal costs per application:** The initial capital outlay for training state-of-the-art LLMs is considerable. However, once a state-of-the-art LLM is trained, it transforms into a universal backbone that can be fine-tuned for a wide array of specialized tasks across disparate domains. This approach shifts the expenditure paradigm from incurring frequent, moderate-scale training costs per individual project to a model where one massive, upfront training investment is amortized across numerous applications, resulting in significantly reduced marginal costs per deployment. This paradigm, however, has been disrupted by reasoning models, notably OpenAI's o1, released in September 2024, which shift a greater share of computational demand to inference, driving up marginal costs per use.
- **Shifts in AI market structure with consolidation of model providers:** the significant resources required to develop competitive foundation models have confined deployment to a few players—mainly US and Chinese tech giants and well-capitalized start-ups—that can absorb the high fixed costs of GenAI. Economies of scale create a "winner-takes-all" effect at the foundation model level, forming platforms for smaller 'model adapters' on the other side of the market. As a result, the GenAI market is expected to stay concentrated at lower levels (computing infrastructure, models) but fragmented in higher tiers (software applications). However, early debates have questioned whether Big Tech's proprietary models risk commoditisation by smaller, better-designed models or open-source alternatives. This concern resurfaced with DeepSeek R1's January 20, 2025, release, as its seemingly low-cost architecture challenges the idea of high barriers to entry in model training.

Fig. 8 – Market share across key layers of the GenAI industry, 2023



Source(s): IoT Analytics, Stifel*

Scaling laws are central to AI infrastructure development. Large language models (LLMs) have demonstrated a strong empirical correlation between model performance and the training compute, dataset size, and parameter count: model performance scales log-linearly as these inputs increase. This phenomenon underpins the industry's belief in "scaling laws," which posit a consistent relationship between expanding data and compute power and enhanced AI capabilities. However, defining and quantifying "model capability" remains contentious. What, for example, constitutes a doubling of capability? While specialists have broadly validated scaling laws in recent years, the belief in their ability to sustain over time and enable breakthroughs like AGI (Artificial general intelligence, i.e. an AI surpassing human performance across diverse tasks) remains a leap of faith.

The debate surrounding scaling laws persists. In the latter half of 2024, concerns about diminishing returns on scaling ignited widespread scrutiny: larger models demand exponentially more resources, yet gains in performance have begun to show diminishing returns when it comes to pre-training (training a model on a large, general dataset). Compounding this issue, growth in available training data lags advances in compute power, resulting in relative data scarcity and exposing inefficiencies such as suboptimal "Chinchilla scaling", the balance between the model's size (parameters) and the training dataset size (tokens) originally described by researchers at Google's DeepMind.

Despite these limitations, AI progress continues by scaling in new ways. AI scaling laws typically plateau over time (performance gains on scaling one input show diminishing returns), so new scaling laws must be identified to sustain AI progress. One approach is test-time compute scaling, where increased computational resources are allocated during inference. OpenAI's September 2024 o1 reasoning model exemplifies this approach, taking more time to 'think' and refine its answers. Other techniques, such as synthetic data generation, also demonstrate potential for leveraging computational resources to improve model performance beyond the pre-training phase.

Fig. 9 – Training computation of notable AI models (FLOP)

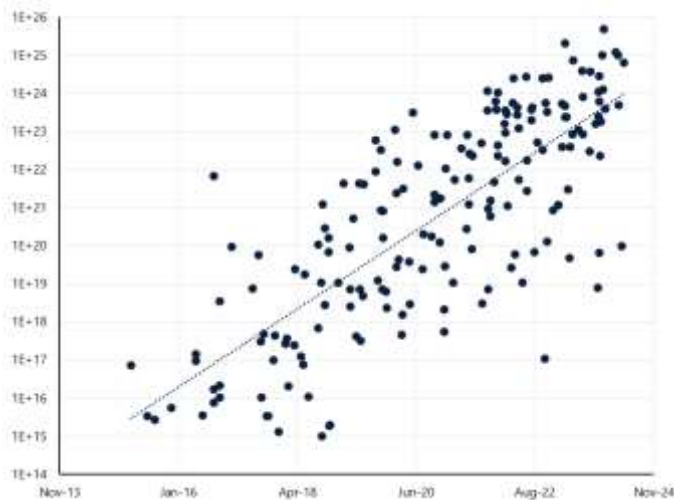
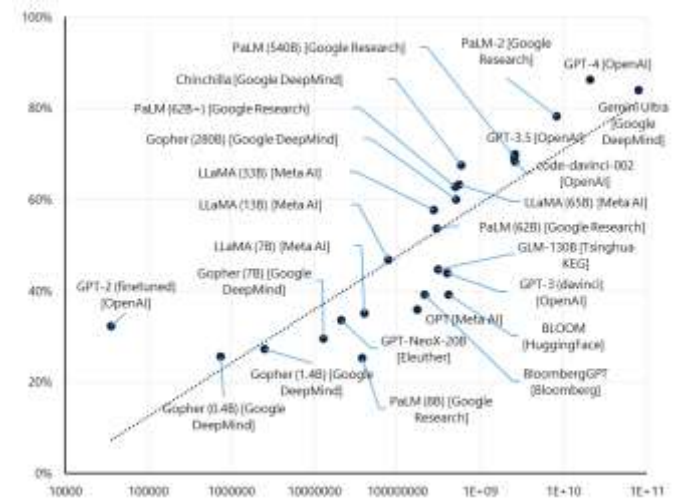


Fig. 10 – GenAI performance on knowledge tests vs. training computation (petaFLOP)

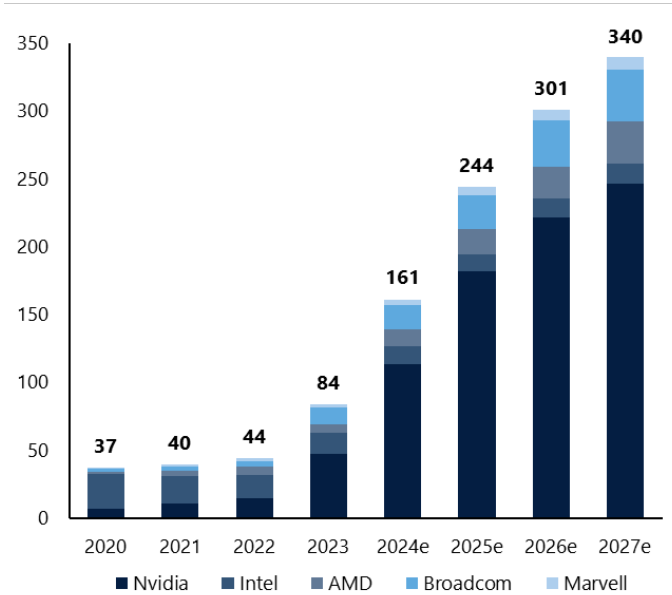


Source(s): Stifel* analysis of Epoch AI data. Performance on knowledge tests is measured with the MMLU benchmark.

The race to scale AI has driven an unprecedented surge in demand for computing power, reflected in extraordinary revenue growth among AI data centre chip providers, most notably Nvidia. Annual revenues from the data centre divisions of leading GPU/CPU vendors (Nvidia, AMD, Intel) and DPU/custom silicon players (Broadcom, Marvell) soared from USD37bn in 2021 to USD161bn in 2024, with projections pointing to USD340bn by 2027e—an 8x increase from pre-GenAI levels. Nvidia alone captures >70% of data centre processor revenues; however, its client base is similarly concentrated, with the four hyperscalers (Microsoft, Amazon, Meta, and Google) estimated to account for over 50% of Nvidia's 2023 data centre sales.

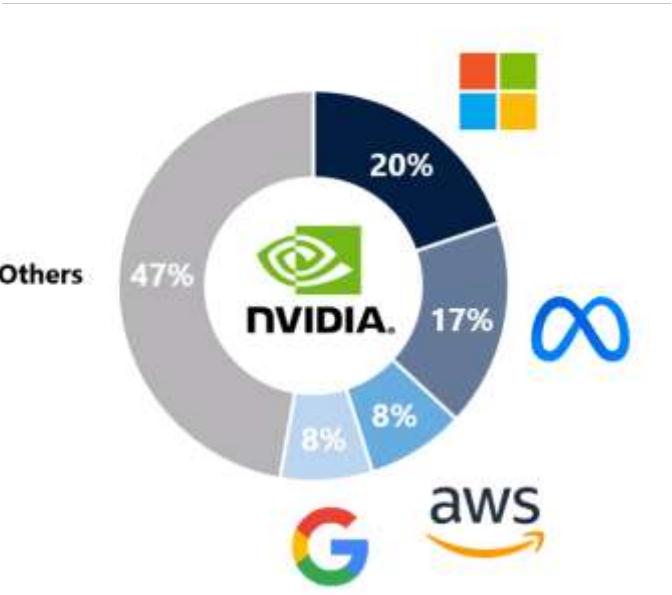
The slowdown in CPU sales further evidences a capital shift towards accelerated computing, with AI accelerators progressively displacing traditional CPU workloads. During Nvidia's Q2 2025 earnings call, CEO Jensen Huang highlighted the rapid depreciation of the USD1tn general-purpose data centres an asset base, with current CPU investments falling short of sustaining this infrastructure.

Fig. 11 – Data centre processor revenues (2020-2027e, USDbn)



Source(s): Bloomberg consensus. Note: Nvidia's fiscal year ends in January, Broadcom's in November, and Marvell's in February.

Fig. 12 – Nvidia's top customers (excluding gaming revenues, 2023)

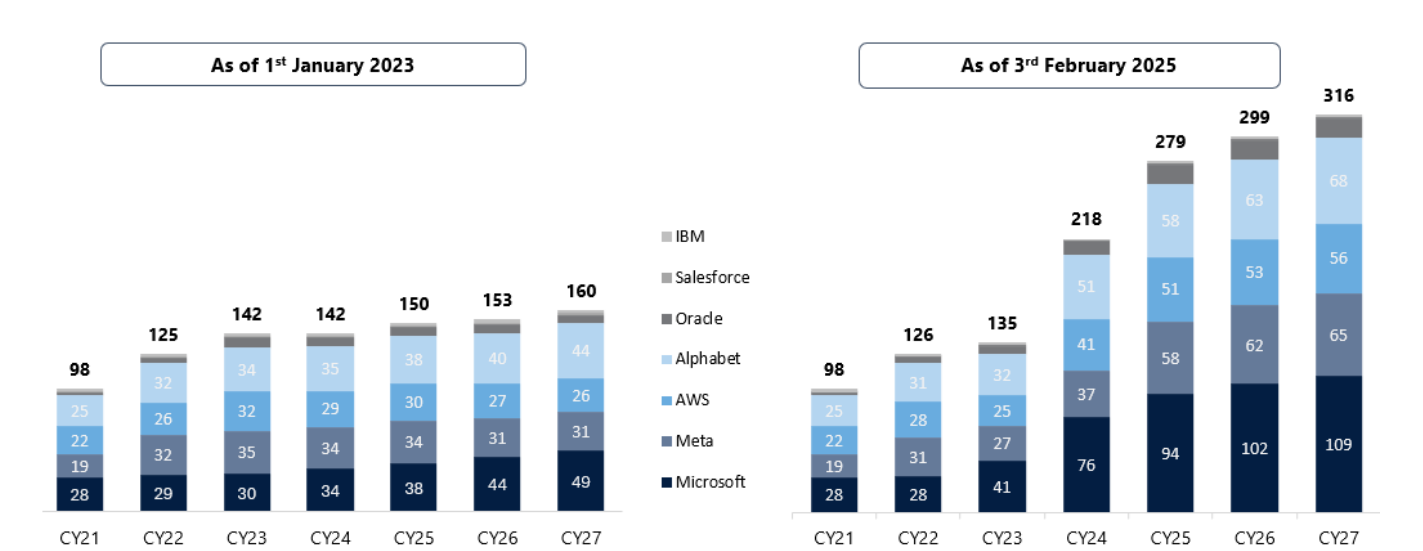


Source(s): Bloomberg Intelligence, Stifel*

Hyperscalers' aggregate capex surged 61% yoy in 2024 driven by their strategic pivot to GenAI, rising from USD135bn in 2023 to USD218bn. This marks a stark deviation from prior forecasts, which anticipated near-stagnant spending over 2023–27, following the 2022 tech downturn and aggressive cost containment by major tech firms. We estimate an incremental USD505bn in capex over 2024–27, which was unanticipated two years ago. The NASDAQ 100's summer 2024 weakness was largely attributed to investor concerns regarding the value creation of these AI capex, demanding transparency on capex-to-revenue dynamics.

Despite scrutiny over ROI, hyperscalers' AI-driven investments are expected to accelerate further into 2025. Bloomberg consensus projects a 28% capex increase in 2025, reaching USD279bn, before moderating to USD299bn in CY26 (+7% yoy) and USD316bn CY27 (+6% yoy), implying a 24% CAGR over 2023–27. 'The risk of under-investing is dramatically greater than the risk of over-investing', noted Alphabet's CEO Sundar Pichai during an earnings call. This suggests the AI data centre capex cycle could persist despite value-destructive ROIs, ultimately exerting downward pressure on margins.

Fig. 13 – Revisions to hyperscale capex consensus estimates

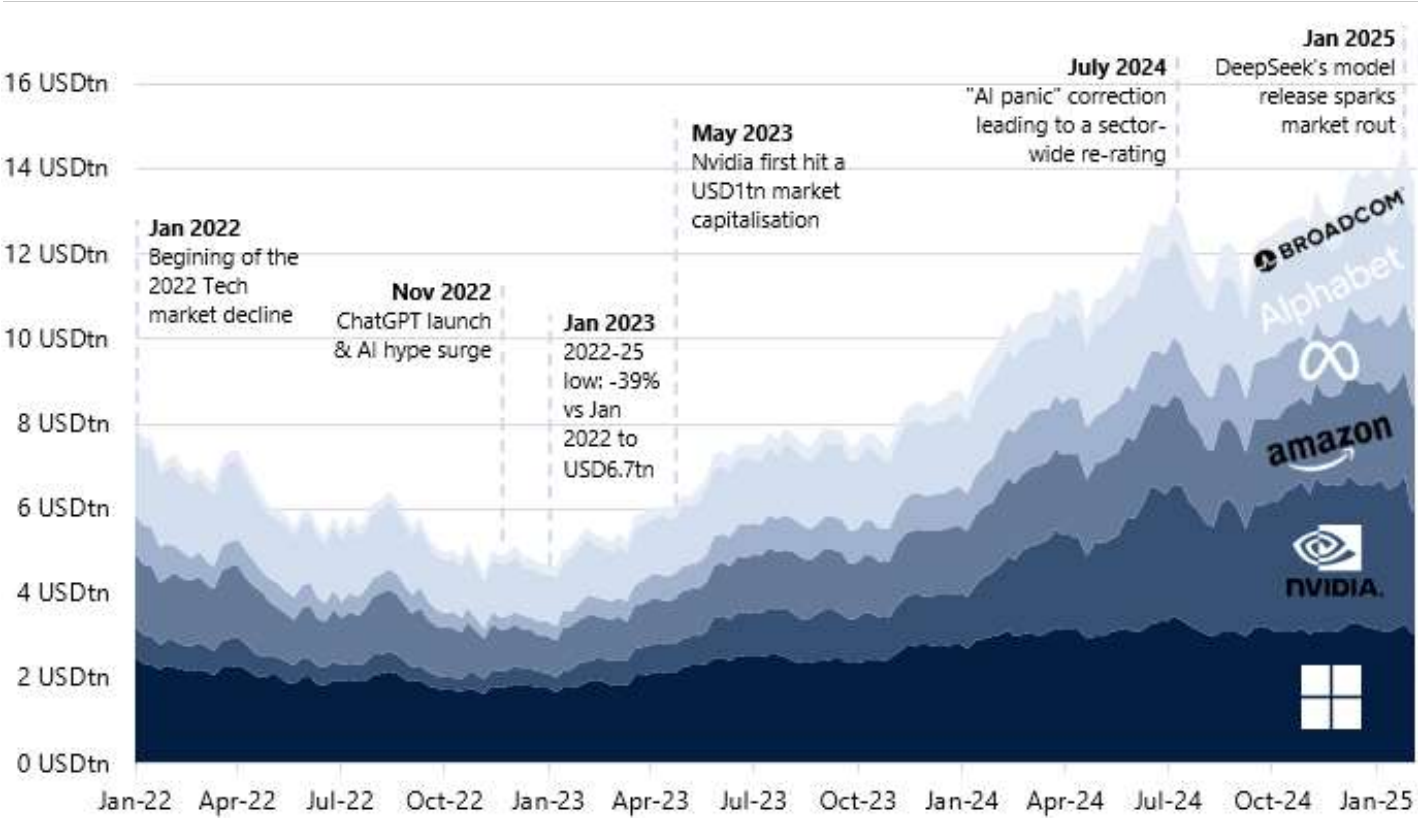


Source(s): Bloomberg, Stifel*

US tech giants defended rising capex plans in January earnings, despite DeepSeek’s concerns. DeepSeek’s R1 model reignited debate over AI infrastructure spending, challenging scaling laws by delivering high performance with comparatively limited resources. The prospect of open-weight alternatives and efficiency gains in model design has heightened the risk of hyperscalers reassessing capex. However, earnings calls indicate that DeepSeek is viewed as an accelerant rather than a deterrent to AI investment;

- **Microsoft:** The top data centre spender reaffirmed its USD80bn FY2025 capex plan (up 60% from USD50bn in FY2024) as its AI business now exceeds a USD13bn annual run rate, up 175% yoy. Microsoft CEO reiterated strong confidence in AI scaling and its integration into enterprise workflows, as adoption will unlock full AI’s ROI. Moreover, the CFO indicated Q3 and Q4 capex will match Q2, implying FY25 capex could most of the USD90bn.
- **Meta’s** AI strategy has prioritised open-source AI with Llama models, while allowing monetisation by embedding AI into its content ranking and ad optimisation engines. During its January 2025 results, Meta announced FY2025 capex plans to range between USD 60–65bn (vs USD39bn in FY24), with investments in AI capabilities driving capex growth, though most of the capex remains dedicated to its core business operations.
- **Alphabet** announced rising AI spending during 4Q24 results, guiding FY25 capex to USD75bn (vs USD59bn consensus). AI is both an asset and a challenge for Google: it can monetise it through Search and Cloud while safeguarding its edge with an integrated AI stack using in-house TPU chips and advanced AI models research capabilities (Gemini, DeepMind). But AI also threatens its search dominance by enabling new discovery methods.
- **Apple** has set itself apart from US Big Tech by prioritising on-device AI over large-scale data centre investments. The June 2024 launch of Apple Intelligence—its AI suite integrating both on-device and server-side processing—reinforced this strategic shift towards minimal cloud dependence. The January 2025 earnings call confirmed the initiative is designed to drive hardware sales rather than operate as a standalone AI platform.
- **Amazon** is ramping up AI cluster investments by deploying Nvidia GPUs and custom silicon (Trainium, Inferentia), but is seen as lagging behind major tech peers in model development and trailing Google in AI hardware. In 4Q24 results, it is expected to provide guidance for ~USD10bn in additional AWS capex for FY25 (FY24e: USD41bn). driven by AI.

Fig. 14 – USD10tn surge following ChatGPT's launch: the combined market capitalisation of major US hyperscalers and AI accelerator providers since January 2022



Source(s): Refinitiv, Stifel*

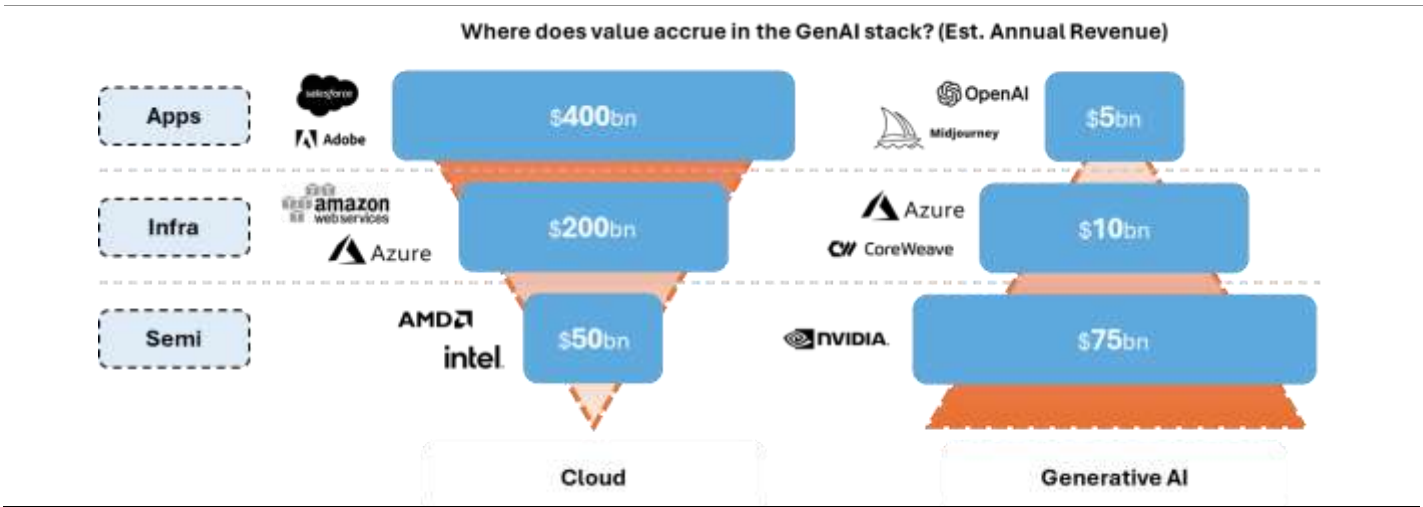
Can the GenAI investment cycle sustain its boom?

GenAI investments are on track to exceed USD 1tn by 2030, but limited near-term revenue generation could create significant ROI headwinds. We believe the trajectory of AI data centre capex will ultimately depend on: (i) the emergence of transformative applications, (ii) the continued viability of scaling laws, and (iii) growing energy constraints. We present three scenarios shaped by these dynamics.

The debate over AI capex ROI has intensified, scrutinising whether AI can transcend the current "picks and shovels" phase to deliver sustained returns. The emergence of breakthrough AI-powered applications that rapidly drive widespread adoption is widely regarded as the ultimate arbiter of the AI cycle's trajectory. At the heart of this issue is the stark gap between revenue forecasts driving AI infrastructure investment and realised ecosystem growth. In June 2024, Sequoia Capital framed this as 'AI's USD600bn Question,' underscoring that sustaining AI-related capex would require c. USD600bn in annual AI software revenue, against identified near-term revenues of just USD75bn. Indeed, The GenAI ecosystem remains disproportionately weighted, with semiconductor firms, notably Nvidia, capturing the lion's share of value, while downstream applications remain under-monetised. This imbalance has created an unsustainable inverted pyramid market structure (see Fig. 15).

Whether AI spending requires the advent of such applications to validate its current trajectory remains an open question. The emphasis on monetising new applications or upselling existing services may underestimate the structural dynamics of this cycle—dominated by entrenched incumbents rather than disruptive entrants. A substantial portion of AI capex appears defensive, a front-loaded "arms race" aimed at safeguarding core revenue streams (e.g., Google Search or Meta's social networks) in an AI-powered future. These incumbents leverage access to deep capital pools and exceptionally low costs of capital, ensuring sustained investment capabilities irrespective of short-term returns. Agentic AI (autonomous decision-making AI systems) and Physical AI (systems embodied in hardware, such as robots interfacing with the physical world) have recently emerged as widely discussed potential sources of AI revenue.

Fig. 15 – The “pyramid inversion” of GenAI value today



Source(s): Estimations from Apoorv Agrawal in April 2024, Stifel*

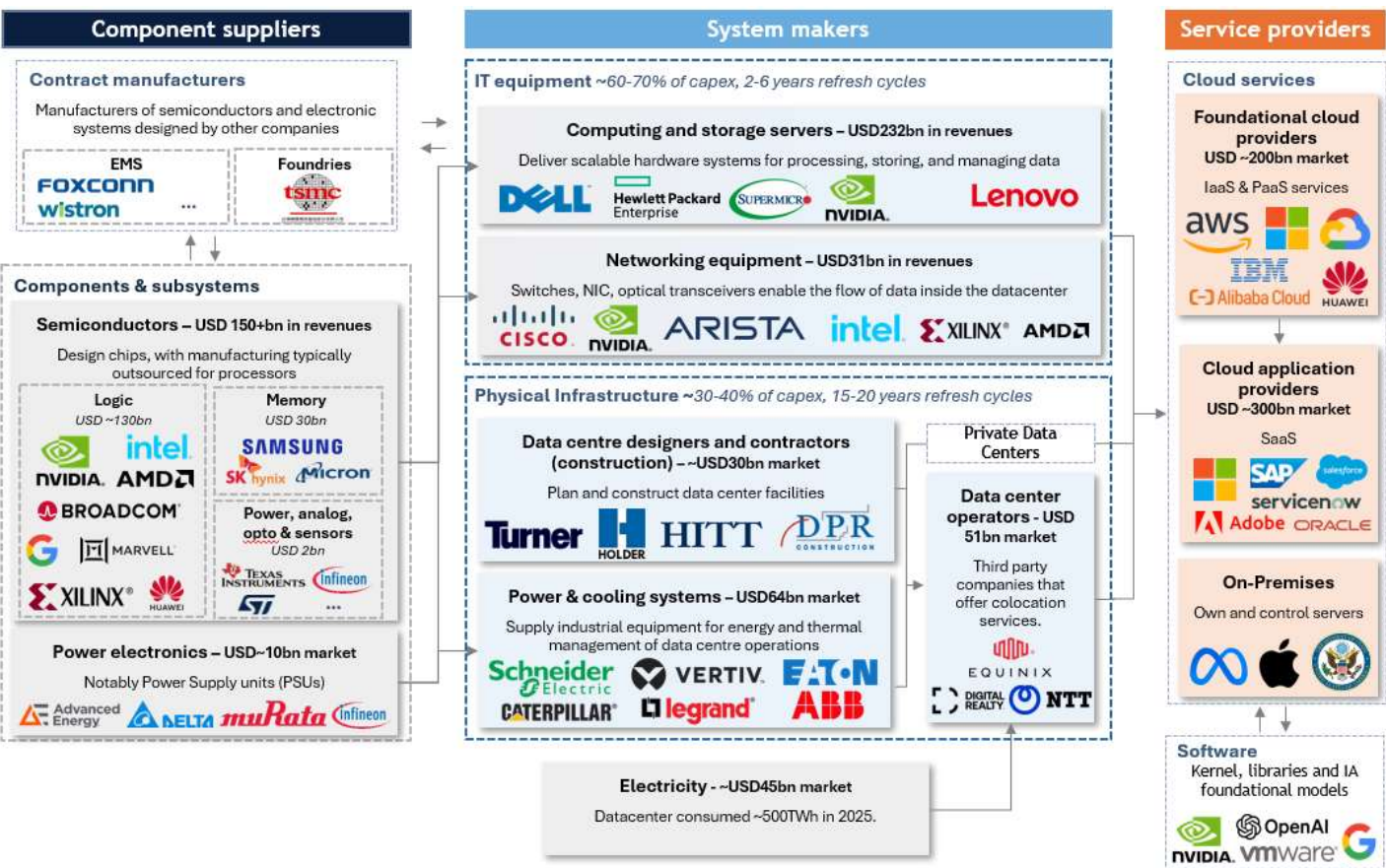
Continuation of scaling laws remains central to AI investment trends. AI performance continues to scale with compute—though benchmark-dependent—and has yet to reach its limits. However, scaling now extends beyond pre-training (larger models trained on more data improve performance), the original paradigm. Two more laws have emerged: post-training scaling (fine-tuning and reinforcement learning boost performance) and test-time scaling (more compute at inference improves accuracy). A key risk is continued technical scalability while user-level gains stall.

The accelerated cadence of AI hardware cycles, exemplified by Nvidia, remains pivotal in sustaining the exponential compute growth underpinning AI model scaling. AI processor architectures now evolve on markedly compressed timelines relative to historical computing trends, with Nvidia shifting from a biennial to an annual GPU release cycle.

Multi-layered supply chain constraints could also impede AI data centre expansion. We believe the most severe effects are likely to manifest within a three- to five-year horizon, primarily due to power infrastructure limitations. We identify three key bottlenecks: (i) semiconductors, (ii) electrical equipment and (iii) energy constraints.

- **The initial AI scale-up phase saw shortages of GPUs and AI accelerators, largely driven by constrained HBM (High-Bandwidth Memory)¹ and CoWoS (Chip-on-Wafer-on-Substrate)¹ capacity.** While supply constraints have eased, they remain structural bottlenecks. HBM is critical to GPUs and AI accelerators, enabling to process vast datasets in parallel by stacking DRAM dies on a logic die. DRAM majors (Samsung, SK Hynix, Micron) are shifting production towards AI-centric memory, but supply remains tight. Similarly, CoWoS, an advanced packaging technology for HPC interconnects, is another bottleneck, hindering AI chip production since ChatGPT's late 2022 launch. TSMC plans to more than double CoWoS capacity in 2024, with another near-doubling by 2025, yet supply remains constrained. Nvidia, AMD, and ASIC makers cite CoWoS limits as a major challenge. For instance, Nvidia's Blackwell series has faced delays, attributed in part to related packaging issues.
- **Supply chain issues extend to industrial equipment, notably within the electrical systems domain.** The rapid expansion of hyperscale and edge facilities has created acute shortages of critical electrical components, particularly MV transformers, with lead times now extending to two to three years. These shortages reflect structural supply-demand imbalances exacerbated by pandemic-induced disruptions, coupled with limited global manufacturing capacity for specialized components.
- **Energy constraints are an escalating concern, particularly in high-density data centre regions.** In the US, ageing grid infrastructure and regulatory hurdles are delaying or cancelling planned expansions due to restricted high-capacity connections. European markets face rising energy costs and stringent sustainability mandates, complicating site selection and operational strategies. Emerging markets, while appealing for lower costs, often contend with grid reliability issues and limited redundancy, heightening risks for hyperscalers and colocation providers. This is explored in greater detail in Part 3 of this report.

Fig. 16 – The data centre submarkets



Source(s): Stifel*

(1) Refer to the Glossary (p. 3-5) for definition

We have drawn up three medium-term scenarios for AI investment trajectories, each anchored in distinct assumptions about AI adoption, model performance progress, and supply chain constraints :

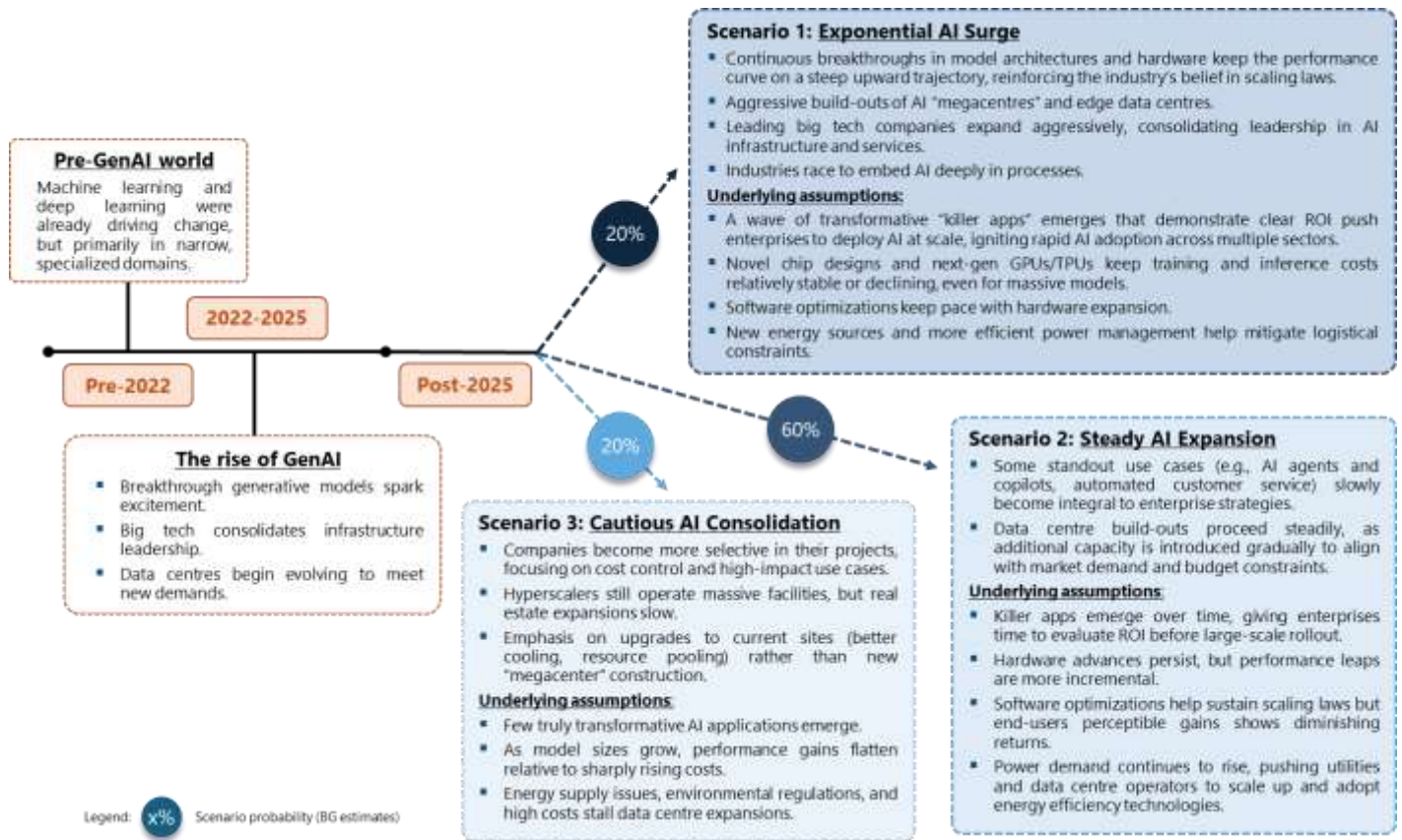
- **Adoption and “Killer Applications”:** The pace of enterprise and consumer adoption hinges on the emergence of disruptive AI-driven solutions capable of delivering demonstrable value across industries. Without such applications, the economic justification for sustained infrastructure investment weakens. However, early AI monetisation may rely more on integration into existing products and revenue streams than on new applications.
- **Scaling Laws and Computational Efficiency:** The materialisation of scaling laws remains a linchpin for continued AI infrastructure investment. A plateau in these trends could significantly diminish the rationale for expanding compute capabilities.
- **Power and Supply Chain Constraints:** While medium-term investments in chip production capacity and power infrastructure may alleviate some bottlenecks, systemic challenges—notably energy grid limitations—could limit the scalability of AI infrastructure to meet exponential demand.

The “**Steady AI Expansion**” base case projects gradual AI growth after the 2023-25 acceleration, with hardware and model performance improving under scaling laws. Data centre growth is primarily limited by ROI pressures, as AI monetisation takes time. AI capex jumps 26% in 2025 before stabilising at high-single-digit growth through 2026–30.

In the bullish “**Exponential AI Surge**” scenario, algorithmic and hardware breakthroughs trigger an inflection point in scaling laws, driving exponential compute demand. This accelerates hyperscale and sovereign data centre investments, potentially catalysing a race towards Artificial General Intelligence (AGI)¹ development. Growth in this scenario is primarily capped by energy constraints.

The bearish “**Cautious AI Consolidation**” scenario assumes slower progress in scaling laws, hardware, and AI adoption prompting a shift to strict cost controls. Investment focuses narrowly on transformative use cases, favouring existing infrastructure over expansion. Strong data centre capex growth in 2024–25 is followed by a digestion phase, with an MSD decline in 2026–27 before stabilizing at a +7% CAGR in 2027–30.

Fig. 17 – Our three long-term scenarios



Source(s): Stifel*

(1) Refer to the Glossary (p. 3-5) for definition

02

THE NEW BLUEPRINT FOR DATA CENTRES

HOW ARE AI WORKLOADS PUSHING THE LIMITS OF DATA CENTRES' ELECTRICAL AND COOLING SYSTEMS?

While data centre cost structures remain predominantly IT-focused, the facility infrastructure is also critical and alone accounts for ~30% of new-build data centre capex. Global power capacity is projected to double by 2030, up from the current ~60GW—implying nearly USD1tn in facility infrastructure spend alone. This expansion is primarily driven by AI server deployments, which are redefining data centre design with a focus on energy efficiency and thermal innovation to support this transition. Vendors are gearing up for a one-to-two-decade AI data centre capex cycle, adapting rapidly to AI-driven business needs. The shift towards AI-centric workloads has driven rack power densities to unprecedented levels. Nvidia's Blackwell architecture has redefined benchmarks, surpassing 130kW per rack—three times the density of its H100 predecessor. As even higher-density configurations are under consideration, liquid cooling is emerging as a key enabler for high-performance implementations.

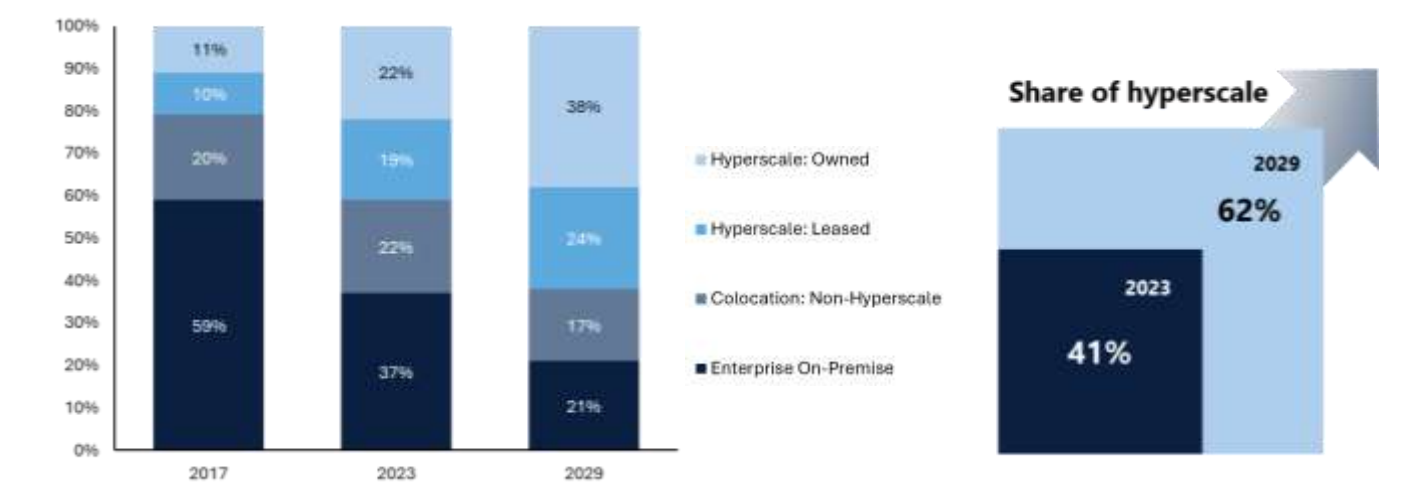
Deconstructing data centre infrastructure

Data centre cost structures remain IT-centric, but facility infrastructure accounts for a material ~30% of capex. The AI-driven investment cycle is reshaping dynamics for electrical and thermal system suppliers, as the demands of accelerated computing drive convergence between IT equipment and infrastructure (power and cooling). Ever larger data centres and stricter energy efficiency needs add to the pressures.

Power, measured in megawatts (MW), is the key metric for data centre scale because it defines how much IT equipment can be supported. Utility power capacity represents the maximum draw from the grid, but industry convention focuses on IT power (or critical power)—the portion allocated to IT equipment, net of non-IT consumption such as cooling and electrical losses. Data centre rental pricing is typically quoted in power terms (USD/kW/month). As HPC data centres' power densities rise, site selection increasingly depends on utility power availability over land area.

The most pivotal development in the data centre market over the past decade has been the rise of hyperscale data centres—the pinnacle of scale. They typically provide over 40MW of critical power and require more than USD1bn in capex, inclusive of IT equipment. These facilities may form part of expansive campuses where multiple data centres interconnect, cumulatively delivering several hundred MW. Current trends indicate a shift towards ever larger footprints and the emergence of megacampuses ('megacentres'), potentially reaching multiple GW, although capacity expansion is generally executed progressively. Major tech firms own and operate them for internal cloud services; alternatively, colocation providers develop and lease them to hyperscalers. Synergy Research reports their share of global capacity rose from 22% in 2017 to 41% in 2023, with projections reaching 62% in 2029.

Fig. 18 – Data centre capacity trends: share of critical IT load (Worldwide, MW)



Source(s): Synergy Research group – August 2024, Stifel*

The colocation model has become another key trend as enterprises seek to optimise IT capex deployment. Colocation data centres, operated by providers such as Equinix and Digital Realty, host multiple organisations under flexible capacity commitments via multi-year contracts, enabling rapid scalability to meet evolving requirements.

- **Wholesale data centres** typically range from 10–40MW, typically these are located in regional hubs to serve large enterprises, government agencies, and cloud service providers. Leases tend to be long-term, generally 5–10 years.
- **Retail data centres** represent the smallest tier, generally under a few MW, often in urban centres to minimise latency and network congestion. These sites host multiple small tenants leasing a limited number of racks, catering to customers with lower power capacity requirements within shared data halls.

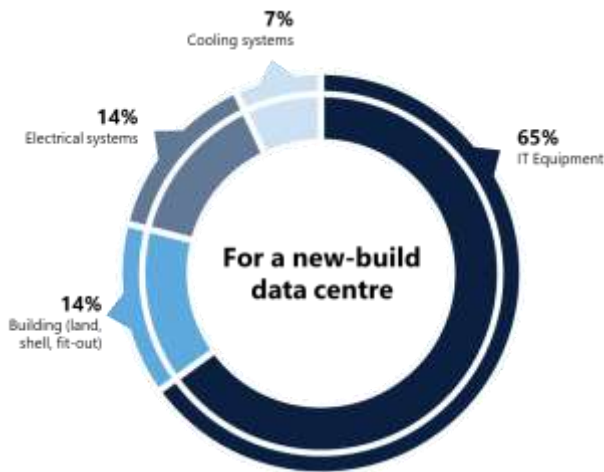
There are, however, many more types and classifications of data centres. For example, edge data centres, typically smaller, are decentralised facilities designed for local processing near data sources. However, the shift to cloud computing has constrained this segment, concentrating workloads in larger, centralised facilities. In 2018, Gartner estimated that only 10% of enterprise data was processed at the edge. To cut costs and expedite deployment, operators have explored modular data centres: standardised, prefabricated structures such as containerised units and prefabricated data halls.

IT equipment dominates data centre cost structures, accounting for ~75% of the ~USD375bn global data centre capex in 2024. Global total capex is skewed towards compute, networking and storage infrastructure due to their shorter asset lifecycle (typically 3 to 6 years) compared with infrastructure (typically 20 years). The proportion allocated to IT equipment is poised to grow further as the adoption of HPC accelerates, given that IT hardware for AI workloads typically command an even larger share of data centre cost structures.

The focus on IT investments often overshadows the significant costs associated with constructing the facilities themselves. For new builds, we estimate physical infrastructure—including electrical and thermal systems—accounts for on average ~30-40% of data centre capex (~20% for HPC sites):

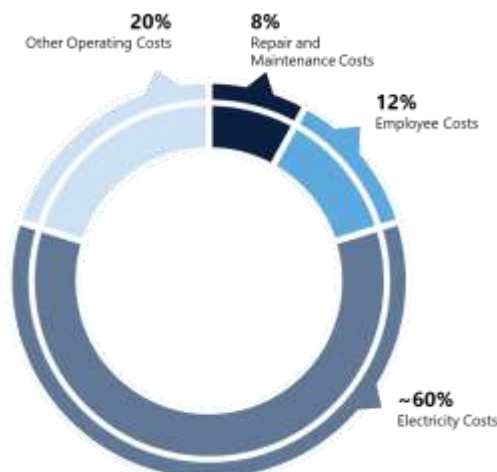
- **For large, greenfield data centres, facility infrastructure costs (core shell, power distribution, cooling systems) typically range from USD7–10m per MW.** Costs vary widely depending on power density, resiliency requirements, scale, and location. Economies of scale drive down per-MW costs, while smaller-scale sites may cost toward USD15m per MW. High redundancy requirements and high-density data centres (e.g., HPC) push costs higher. Location is another critical factor, with key variables including land acquisition, labour, and electricity costs. Retrofits or expansions can be several million dollars cheaper per MW but are often necessary rather than opportunistic, as infrastructure obsolescence accelerates with the AI shift.
- **The core building structure accounts for a smaller cost share (~40%) than electrical (~40%) and cooling (~20%) systems supporting servers.** Moreover, electrical-thermal costs typically scale with power, meaning that an increase in data centre size or power density yields only minor savings.
- **Electricity is the dominant operating expense for data centre operators,** typically representing >50% of total opex, though subject to significant variability based on local power tariffs, energy efficiency, and other factors.

Fig. 19 - Data centre construction cost breakdown



Source(s): Stifel*

Fig. 20 – Data centre operating costs breakdown



Source(s): Stifel*

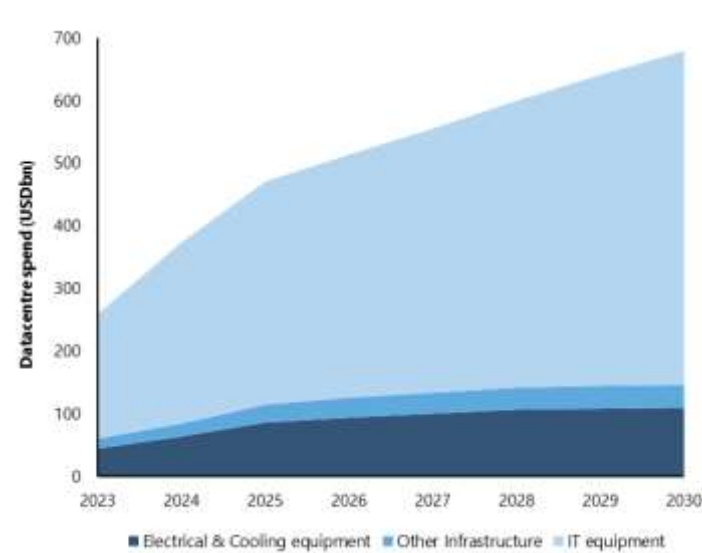
A typical timeline for a large-scale hyperscale facility spans approximately 30 months, varying slightly based on modularity or extending for highly advanced, cutting-edge designs.

- The initial six months focus on site selection, prioritising power access, connectivity, latency considerations, low natural disaster risk and other location-specific factors. These decisions shape the data centre’s design and operational capabilities, influencing subsequent phases.
- The following six months are allocated to design and budgeting, succeeded by regulatory approval (‘permitting’). Construction commences 12 months into the timeline.
- Over the next 18-24 months, the facility’s physical structure is completed, followed by systems integration, testing, and commissioning. This phase includes rigorous evaluation of power, cooling and redundancy systems to ensure compliance with operational and regulatory standards. By month 30, the facility is handed over to its owners, ready for deployment.

We estimate the data centre electrical and thermal equipment market, valued at ~USD50bn in 2024, will expand at a double-digit rate over the next five years. Dell'Oro Group estimates the data centre physical infrastructure market was USD28bn in 2023. However, factoring in the substantial 2024 growth and a broader product scope, we add ~USD20bn to the estimate. We expect the electrical and cooling segment to expand slightly below overall data centre capex (+15% CAGR, 2023–30) as AI IT hardware gains share. Nonetheless, this still presents a compelling double-digit growth opportunity, particularly in hyperscale and colocation, where we anticipate mid- to high-teens expansion. As hyperscalers and enterprises accelerate AI adoption, components optimised for higher power densities and advanced cooling are poised for above-market growth. We understand the sector is preparing for a one- to two-decade growth cycle and is scaling production capacity accordingly, particularly in the US. Vendor FY25 backlogs are nearly full, with clients securing long-term orders, offering strong visibility for equipment players.

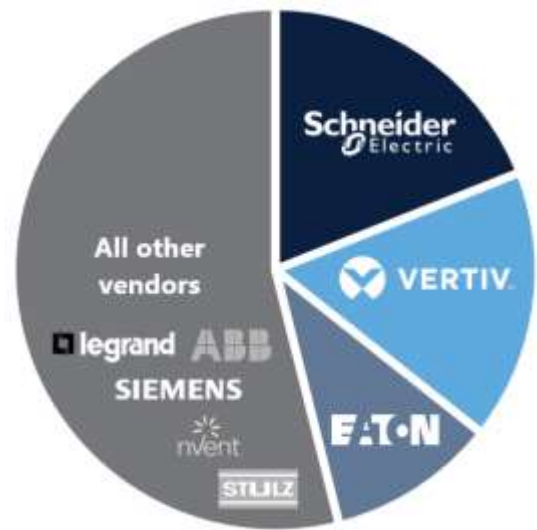
This data centre electrical and cooling market is dominated by Schneider Electric, Vertiv and Eaton. Collectively, the Big Three command ~45% of core electrical and cooling capex in data centres, with offerings that span most product categories. However, the market is more fragmented when including broader electrical and cooling equipment (e.g., generators, MV systems). Moreover, the competitive landscape is more fragmented at the component level, due to product specialisation. Some firms dominate specific subsegments, like Legrand in white space products. Hyperscalers typically prefer multi-vendor, best-of-breed solutions over integrated end-to-end systems, heightening competition among manufacturers. While end-to-end solutions offer integration and simplicity, they introduce risks such as supply chain bottlenecks and vendor lock-in, which hyperscalers try to avoid.

Fig. 21 – Projected data centre capex breakdown, 2023-30e



Source(s): Stifel*

Fig. 22 – Vendor market share in data centres electrical and thermal infrastructure



Source(s): Stifel*. Back-up generators not included

Vendors are forced to compete against one another to win hyperscaler projects. Selection criteria predominantly hinge on technical capabilities, lead times, and total cost of ownership. Technical differentiation is increasingly centred on delivering high-density equipment, which remains a critical priority for hyperscalers. However, in the large-scale data centre segment, hyperscalers often leverage significant in-house expertise in design, favouring the development of proprietary solutions. Vendors must demonstrate agility, accommodating frequent design revisions imposed by hyperscalers. During Covid, supply chains became severely disrupted, which had a knock-on effect on data centre operators, benefiting vendors who navigated them best.

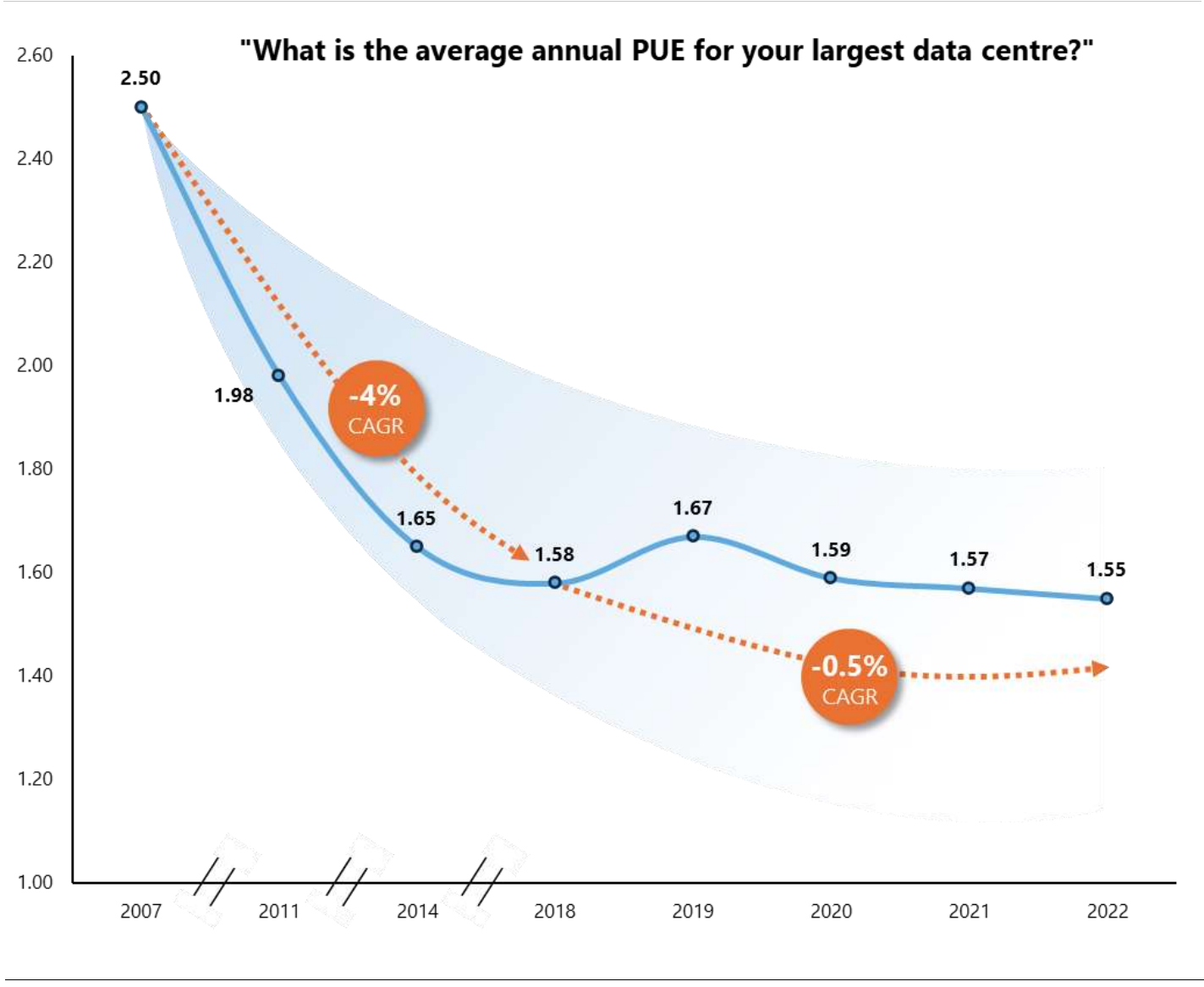
Electrical and cooling system providers are integral to maintenance services. A significant portion of what hyperscaler customers pay for is the long-term warranties (10–15 years). These warranties ensure the rapid dispatch of engineers—typically within hours, regardless of location—in the event of system failure. This critical, on-demand support necessitates 24/7 global networks of highly trained engineers, a resource currently in short supply. Furthermore, third-party service providers offer limited utility, as certain repairs require manufacturer-trained engineers with in-house equipment expertise.

The industry takes Power Usage Effectiveness (PUE) as a key metric to evaluate the energy efficiency of data centres. PUE is calculated by dividing the total power entering a data centre by the power consumed solely by IT equipment. Enterprise colocation facilities typically operate at PUE levels of 1.5–1.6, while hyperscale data centres often achieve sub-1.4 PUE metrics, with some purpose-built facilities—such as those by Google—reporting figures as low as 1.10. AI-specific data centre designs commonly target PUEs below 1.3. According to the Uptime Institute, the industry-wide average PUE has seen a marked decline, from 2.5 in 2007 to an estimated 1.55 by 2022, representing one of the most significant drivers of energy savings and a critical factor in and curbing the exponential growth of data centre power consumption.

The data centre industry's pursuit of decreasing PUE is facing a deceleration in efficiency advancements. Despite initial substantial improvements, recent analyses indicate that average PUE values have plateaued, with minimal reductions observed over the past decade. This stagnation is exacerbated by the proliferation of HPC applications, notably those associated with AI workloads, which necessitate increased power and cooling resources. This substantial power density challenges traditional data centre cooling and power distribution infrastructures, which were originally designed for lower-density deployments. Consequently, achieving further reductions in PUE becomes increasingly complex, as the marginal gains from traditional efficiency measures diminish.

To mitigate these challenges, the industry is advancing liquid cooling solutions capable of reducing PUE by 0.2–0.3 by decreasing reliance on less energy-efficient air cooling. However, for modern hyperscale data centres, where PUE is already below 1.30, this likely represents the final phase of material PUE improvements.

Fig. 23 – Data centre average annual PUE: progress is stalling



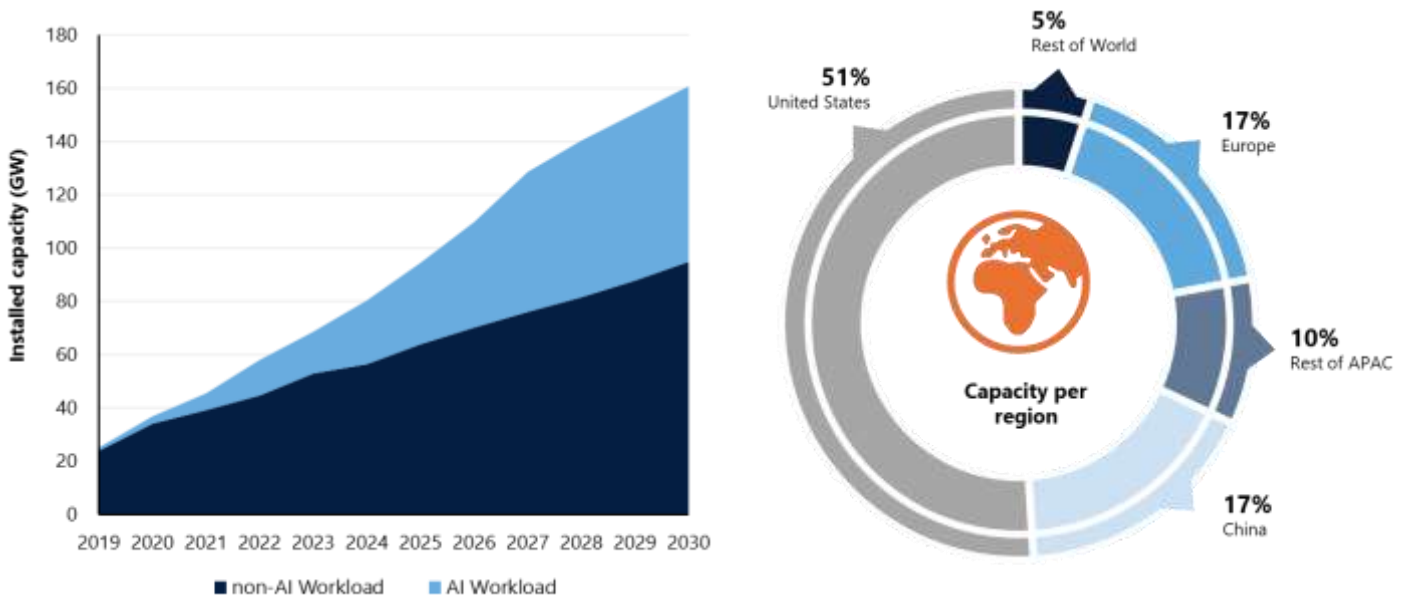
Source(s): Uptime Institute Global Survey of IT and data centre managers 2007-2022

AI transition is driving data centre densification

AI's growing power needs are fundamentally altering data centre architectures to enable deployment of high-tier processors. HPC applications, in particular, are pushing rack densities to new heights. Nvidia's Blackwell architecture suggests the standardisation of >100kW racks, speeding the shift to liquid cooling, to accommodate GPU advancements.

Global data centre power capacity is projected to more than double between 2023 and 2030. As of 2023, installed capacity stands at 57GW, according to Schneider Electric, with AI accounting for no more than 4.5GW (8% of the total). Estimates of global capacity vary by ± 10 GW, reflecting both methodological differences and the inherent opacity of data centre markets. However, multiple forecasts converge on a doubling—or even tripling—of capacity by 2030. Omdia projects an additional ~90GW of installed capacity between 2023 and 2030, exceeding 160GW. In its November 2024 CMD, Vertiv raised its forecast to ~100GW of new capacity from 2023 to 2029, with annual additions of 13GW to 20GW. This sharp acceleration in data centre expansion is clearly AI-driven, with the AI momentum confirmed by the substantial capex increases announced by US hyperscalers in early 2025. However, our base case remains more conservative at ~12GW of annual capacity additions from 2023 to 2030. Beyond 2027, data centre capex trends remain difficult to forecast, in our view, contingent on AI scaling laws and monetisation dynamics.

Fig. 24 – Estimated total data centre installed capacity, in GW



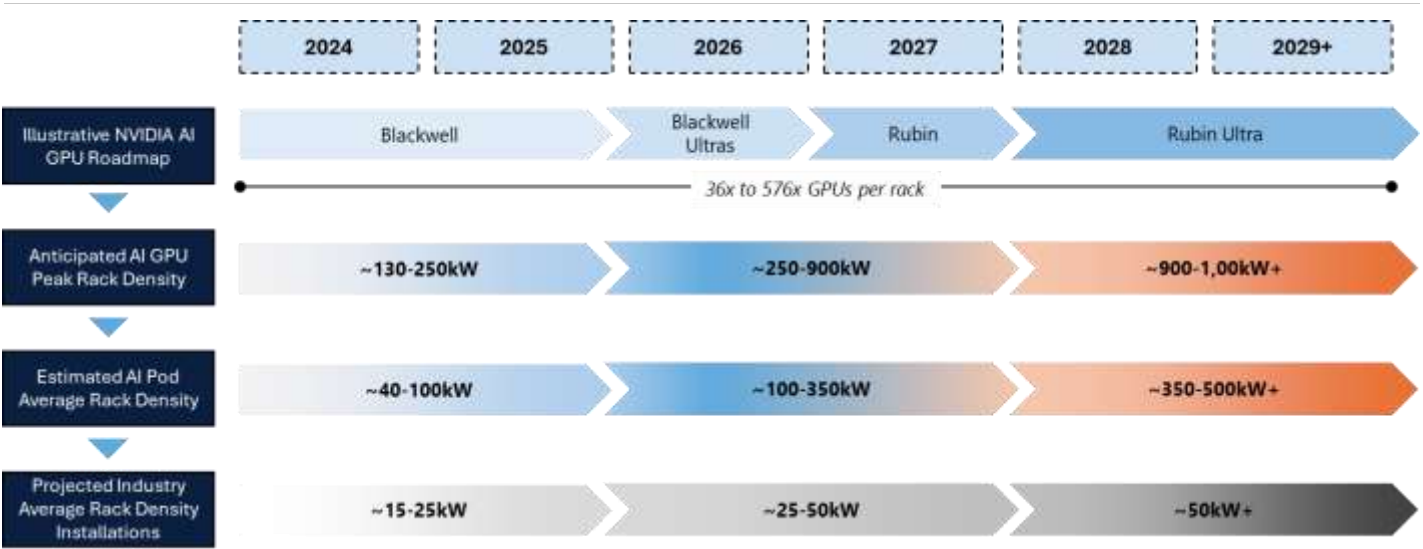
Source(s): Omdia 2024, Stifel*

The evolution towards AI-centric workloads is redefining data centre architectures, notably through increasing rack power densities. In 2022, before AI adoption accelerated, average rack densities were at ~10kW, according to Digital Realty, though hyperscalers typically utilised 2–3x higher-density racks. The industry has incrementally moved towards key systems enabling higher power densities—critical for AI workloads—but establishing common standards has been challenging. Google remains an exception, having pioneered higher-density racks and the early adoption of liquid cooling. However, Nvidia's Blackwell architecture marks a turning point, heralding ultra-dense configurations exceeding 100kW. For example, Nvidia's GB200 NVL72 rack-scale server, unveiled in March 2024 as part of the Blackwell series, integrates 72 GPUs, each consuming 1.2kW, thereby pushing total rack density beyond 130kW—a new benchmark that triples the density of an H100 rack. The H100, released just two years ago, was itself a breakthrough in performance. Therefore, ultra-dense configurations such as 300kW racks are already being actively discussed by industry players. Densities in the 120–140kW range present significant operational challenges, requiring even hyperscale facilities to implement bespoke designs for large-scale deployment.

This evolution reflects a broader trend in data centre design, where compute resources are increasingly concentrated at the rack level. Effectively, this represents an extension of Moore’s Law beyond the chip, with performance gains transitioning from silicon-level scaling to system-level optimisations. Rack densification mirrors the efficiency gains historically driven by transistor miniaturisation, reinforcing the industry’s push towards higher performance per watt. These include enhanced interconnect and networking technologies, such as Nvidia’s NVLink, which reduce bottlenecks and allow multiple GPUs within a rack to operate as a unified compute fabric.

As power densities rise, traditional air-cooling methods struggle to dissipate heat effectively, making liquid cooling an essential technology for next-generation data centres. The latest Nvidia chips are being designed for direct-to-chip cooling configurations with rack densities ranging between 60kW and 120kW. While active rear door cooling should be serviceable at some of the higher rack densities (<80kW) as chips and rack densities continue to increase towards densities greater than 150kW, direct-to-chip cooling and immersion will be the only viable offerings available, although the latter is far more niche at the moment.

Fig. 25 – Hardware is (h)eating the world: Vertiv’s AI rack density roadmap



Source(s): Vertiv (CMD 2024), Stifel*

>300kW rack densities would push low voltage (LV) systems to their limits, if not create bottlenecks, hence industry discussions about medium voltage (MV) power distributions. Compared to LV, MV significantly cuts resistive losses, thus improving energy efficiency. We estimate that this change would involve using MV equipment (UPS, switchgear, busbars, transformers) up until the rack PDU, where it would be stepped down for the IT equipment. In doing so, the length of LV cables would be drastically reduced, minimising resistive losses.

Data centres are increasingly built using standardised, scalable designs to enable rapid expansions to meet the surging demand for AI. Large facilities are assembled from small, replicable units, forming extensive structures. The smallest IT unit, known as a "pod", typically spans up to a few MW and is designed around GPUs, TPUs, or other HPC components. These pods feature high-bandwidth interconnects (e.g., InfiniBand, NVLink, RoCE) and specialised storage systems, with each pod powered by dedicated electrical infrastructure. Multiple pods combine to form a data hall, while larger centres aggregate data halls into buildings, which can be interconnected to create sprawling campuses exceeding several hundred MW. This enables progressive scaling to handle increases in demand while standardising infrastructure equipment for streamlined procurement.

Pre-fabricated modular data centres are also gaining traction as a scalable and efficient alternative to traditional brick-and-mortar facilities. These "data centres in a box" are factory-assembled, shipped to deployment sites in sections, and assembled like LEGO bricks. Each module can integrate power distribution, cooling infrastructure, and structural components such as corridors and lift shafts. While logistics remain a challenge, the key advantage lies in accelerated deployment. Factory-built and rigorously tested before shipment, these modules reduce on-site construction, lowering labour costs and mitigating installation risks.

AI workloads can be broadly categorised into training and inference, each with distinct data centre requirements. Training workloads are significantly more power-intensive, necessitating deployment in high-density data centres. However, as they are latency-insensitive, such workloads can be situated virtually anywhere. In contrast, inference workloads, while typically lower in density, require proximity to population centres to minimise latency.

- **Training workloads are used to train AI models** such as large language models (LLMs). Beyond servers, training necessitates substantial data storage and robust networking to interconnect these components. These elements form AI clusters—arrays of racks functioning as a single computer. Clusters vary widely in size, ranging from a few racks to hundreds, and are often characterised by the number of accelerators deployed. The requirements of AI training workloads diverge sharply from conventional data centre hardware. Training is latency-agnostic and don't need to be near major population hubs. Moreover, AI clusters operate near full utilisation throughout their training cycles, which may span hours to months.
- **Inference, by contrast, entails deploying trained models to generate outputs for new inputs (queries).** For users, inference involves a trade-off between output accuracy and latency. Inference workloads frequently utilise accelerators for large models and, depending on the application, may rely heavily on CPUs. Applications such as autonomous vehicles, recommendation engines, and conversational AI (e.g., ChatGPT) each demand bespoke IT stacks tailored to their operational needs. Hardware requirements for inference can range from edge devices (e.g., smartphones) to multi-rack server configurations, with rack densities spanning from a few hundred watts to over 10 kW. Unlike training, inference server requirements scale with the number of users and query volumes.

A key takeaway is that inference workloads will rise over time as newly trained models transition to production. A greater share of inference-optimised data centres could temper data centre densification, given that inference typically operates on less dense configurations than training.

Fig. 26 – Classification of AI data centres

	Data-Lake (existing workloads)	Inference	Training
Functioning	Huge <u>data Lakes</u> to be utilized by AI systems	User requests answered, or <u>inferred</u> by trained model	Mass amounts of data used to <u>train</u> an AI model
Rack Density	Low to Medium (5-15kW)	Medium to High (15-25kW)	Very High (35-100kW)
Proximity	Not sensitive, green power preference	Closer to end users	Not sensitive, green power preference
Capacity Block	5-15MW	1-5MW	5-100MW
Interconnection	Collect data and feed training	Training updates models and service user requests	Access to public & private data sets

Source(s): Stifel*

Infrastructure requirements for next-generation AI are forcing to explore new cooling architectures

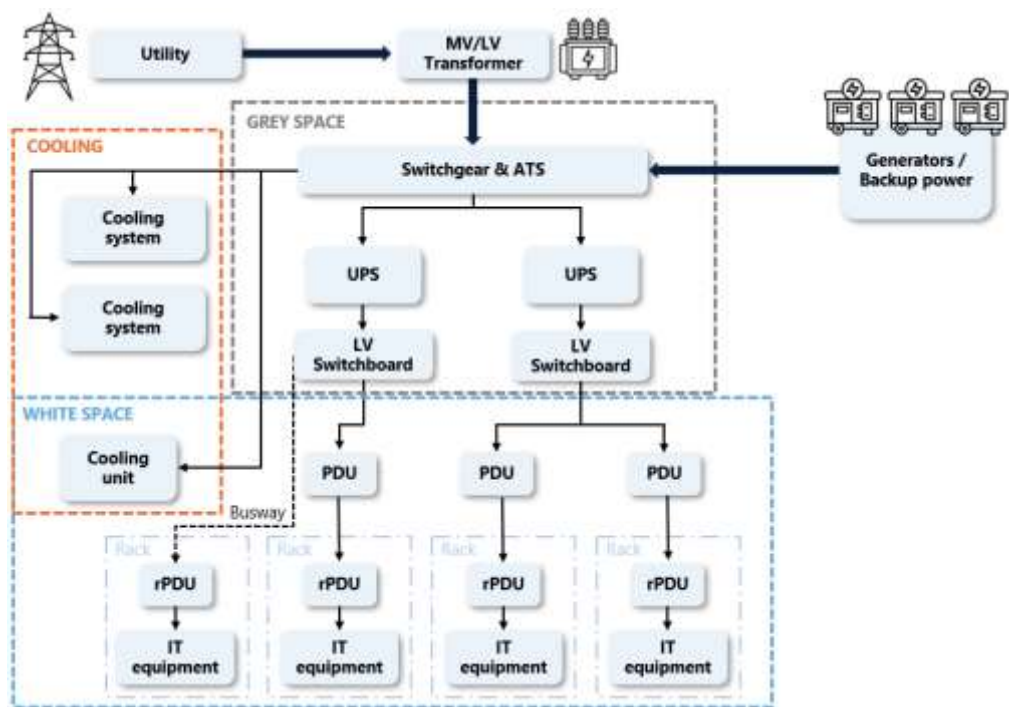
As electrical and cooling systems become increasingly important in data centre design, we examine their architecture. The key shift is the standardisation of liquid cooling for higher-density builds, though air cooling remains in hybrid use with rear door heat exchangers (RDHx) or direct-to-chip (DTC) technologies. Immersion cooling remains a niche solution, with widespread adoption still years away.

Data centres hinge on specialised energy and thermal management systems to ensure uninterrupted operations.

While the components themselves are not exclusive to these facilities, their design and specifications are uniquely tailored to address the immense power demands and heat dissipation challenges posed by modern data centre operations. These systems, essential to sustaining uptime and efficiency, fall into four primary categories:

- **Grey space:** Encompasses back-end electrical infrastructure, including, notably, switchgear and uninterruptible power supplies (UPS). These systems maintain stability across IT and non-IT loads, representing 30–40% of total electrical and thermal infrastructure costs. Switchgear acts as a sophisticated circuit breaker, protecting downstream components from overheating or damage during power surges. The UPS delivers instantaneous backup power during mains disruptions – usually via battery packs – ensuring IT equipment uptime. Most UPS configurations sustain operations for 5–15 minutes, bridging the gap until backup generators engage.
- **White space:** IT-room electrical equipment such as racks housing IT hardware, power distribution units (PDUs), and cable management. White space accounts for 10–15% of costs. PDUs distribute power within the IT room. While basic models function as power strips, advanced units integrate monitoring and rack-level power conversion.
- **Cooling system:** mitigate the heat generated by IT operations to protect the equipment. Air-cooling dominates deployments, though liquid cooling is gaining traction in high-density setups. Cooling systems represent 30–35% of electrical and thermal infrastructure costs, with this share increasing due to the shift towards liquid cooling.
- **Back-up generators:** Predominantly diesel-powered, these systems activate within two minutes of a mains failure to sustain critical loads. Automatic transfer switches (ATS) facilitate seamless power transitions. Generators are a major expense, representing ~15% of electrical and thermal equipment costs.

Fig. 27 - From grid to rack: illustrative electrical distribution system in data centres

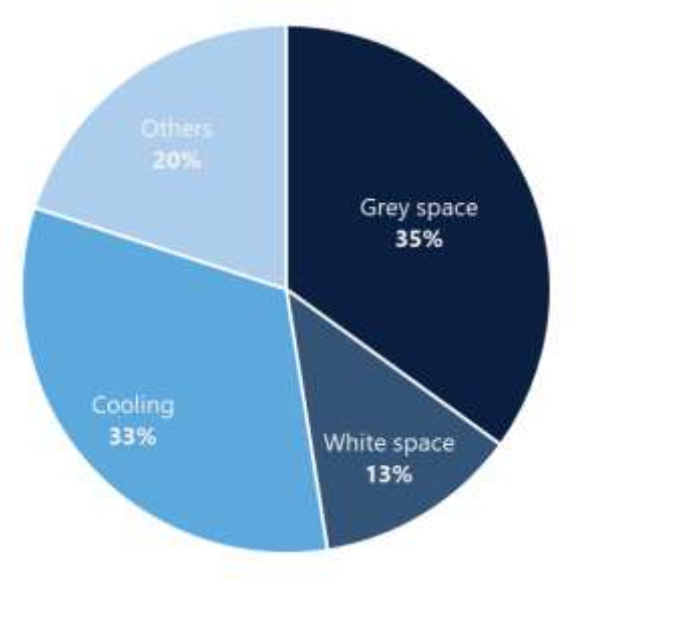


Source(s): Schneider Electric, Stifel*

Most of the value of the equipment is concentrated in a few components. We estimate five pieces of equipment account for ~70% of infrastructure costs: chillers, CRAH units, UPS systems, switchgear, and backup generators. Within electrical systems, grey space amounts to ~3x the value of white space items.

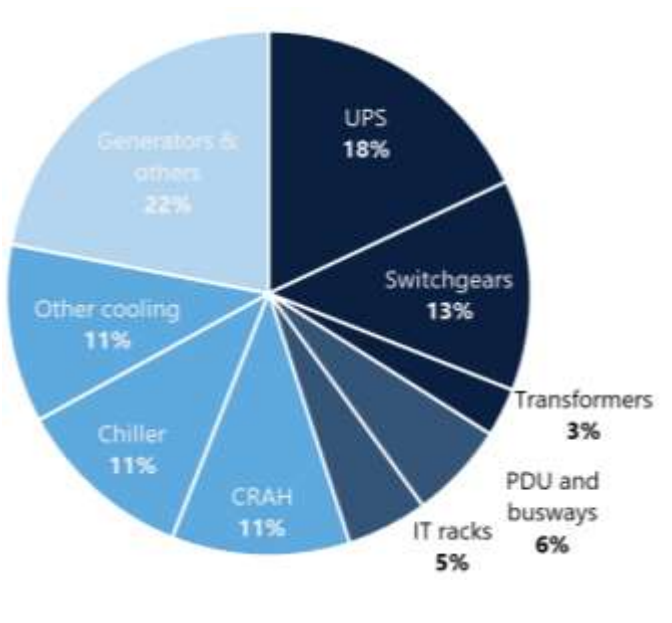
The exact cost breakdown by component varies significantly, depending on data centre design choices and redundancy requirements. The diversity of possible architectures affects the cost allocation across product types. A key determinant is the targeted downtime and redundancy, typically measured by the “Tier” classification (data centres are rated from 1 to 4, with Tier 4 offering the highest resilience). Shifting from an N+1 configuration (N operational components with one redundant) to 2N (N operational components with N redundant) can almost double the cost of the affected components. Tier 3 facilities are the most prevalent, generally requiring N+1 redundancy for upstream components (transformers, generators) while downstream components such as UPS and PDU typically adhere to a 2N standard.

Fig. 28 - Illustrative electrical and cooling cost breakdown



Source(s): Stifel*, Schneider Electric

Fig. 29 – Illustrative electrical and cooling cost breakdown (by components)



Source(s): Stifel*, Schneider Electric

The stringent thermal requirements of high-end AI processors are already driving the shift towards liquid cooling. Despite its advantages, market penetration remains low: Dell’Oro Group estimates the market (direct liquid and immersion) was worth just USD700m in 2023. However, is set to grow at a +44% CAGR to USD4.3bn by 2028, marking a gradual yet inevitable transition. There are three primary liquid cooling technologies:

- **Rear-door heat exchangers (RDHx)** represent the approach most akin to conventional cooling, with liquid-cooled heat exchangers installed at the back of the rack to manage hot air exhaust. While often marketed as a liquid cooling solution, RDHx is best described as a hybrid system in which liquid is used to cool air rather than directly cooling server components. Common in space-limited data centres with 40–60 kW rack densities, RDHx is less efficient than other liquid cooling methods, as it still depends on air-based heat transfer.
- **Direct-to-chip (DTC):** In this method, a liquid circulates through a cold plate positioned near the highest power-density components (usually processors), directly absorbing heat. DTC currently leads deployments as it can support power densities of 60–120kW (i.e., most of Nvidia’s Blackwell products) and its relative ease of integration into existing data centre infrastructure.
- **Immersion cooling** involves immersing servers in a tank filled with dielectric fluid. This is the most energy-efficient liquid cooling method achieving a PUE of ~1.05, and it has been used in racks exceeding 150 kW. However, it poses structural challenges, requiring reinforced floors and more space due to the weight of loaded cooling baths. Capex is also significantly higher than that of other cooling methods. Near-term adoption will likely remain limited to AI facilities due to reliability and maintenance challenges.

Fig. 30 – Overview of cooling technologies

Cooling System				
PUE	1.3	1.2-1.3	1.1	1.05
Description	CRAC or CRAH units supply cold air through raised floors into server racks, absorbing heat from the servers and expelling it through hot aisles.	Radiator-like doors attached to the back of server racks, where hot air flows through liquid-filled coils, efficiently absorbing heat and reducing server temperatures.	Cold plates directly attached to heat-generating components like CPUs or GPUs, where liquid coolant (water or dielectric fluid) absorbs heat and cools the system at the source	A method where servers are fully submerged in thermally conductive dielectric liquid. The liquid absorbs heat directly from the components.
Capital Investment	—	Low	High	Higher
Ease of Retrofit	—	Easy	Harder	Difficult
Cost per watt	—	~\$5.00-\$6.00	\$3.50-\$5.00	\$6.00-\$7.00
Heat removal efficiency	—	~50%	~70%-75%	~95%

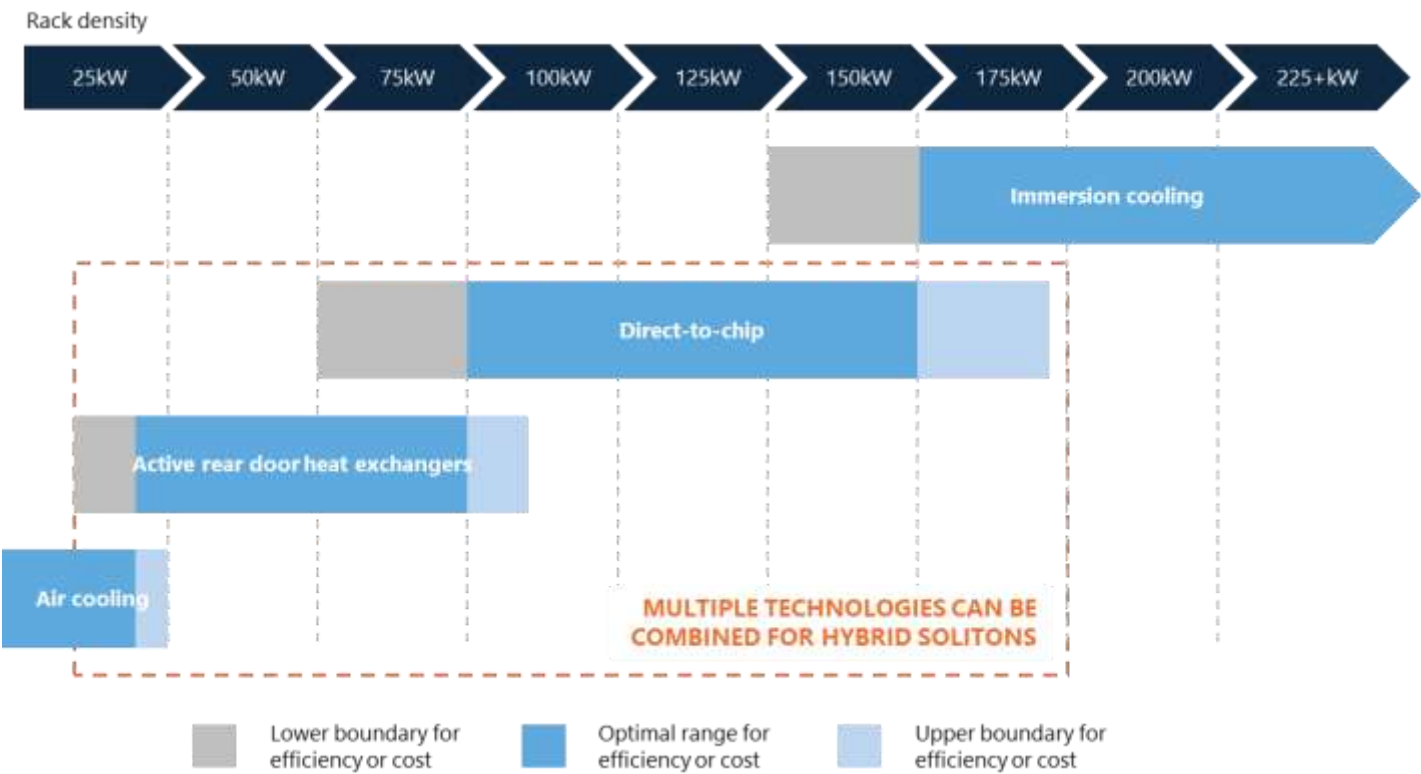
Source(s): Stifel*

Air cooling will remain the primary solution in the near term (1–3 years) and likely in the medium term (4–7 years), albeit with less certainty. The key constraint on liquid cooling adoption is the need for facility redesigns and retrofits. Active rear-door cooling provides a pragmatic alternative as rack densities increase, requiring no structural modifications. Given its simplicity and broad applicability across server density ranges, rear-door cooling should be the preferred approach. However, its total cost of ownership rises with density, which is why new high-performance facilities are almost exclusively designed for liquid cooling.

Despite the shift towards liquid cooling-centric architectures, air cooling will remain relevant. Legacy air conditioning remains viable for datacentres operating below 25kW rack densities. At higher densities, a hybrid model is already common, integrating liquid cooling with residual air cooling. While direct-to-chip cooling removes 70–80% of the heat from the IT equipment, air cooling will be relied on for the last part of heat removal. Therefore, even in the very power-dense AI data centres being built, air-cooling is partly relied on for thermal management designs.

Immersion cooling is expected to remain a niche solution due to high implementation costs and operational complexities. While it delivers the lowest PUEs (sub-1.05), practical limitations persist: IT equipment must dry before reintegration, and concerns exist regarding dielectric fluid fumes. Immersion cooling should remain confined to extreme-density compute environments, although wider adoption should emerge if rack densities make another quantum leap.

Fig. 31 – Applicable cooling technologies by rack density



Source(s): JLL Research, Vertiv, Stifel*

03

SOLVING THE DATA CENTRE ENERGY DILEMMA

BALANCING RENEWABLE POWER & STABILITY

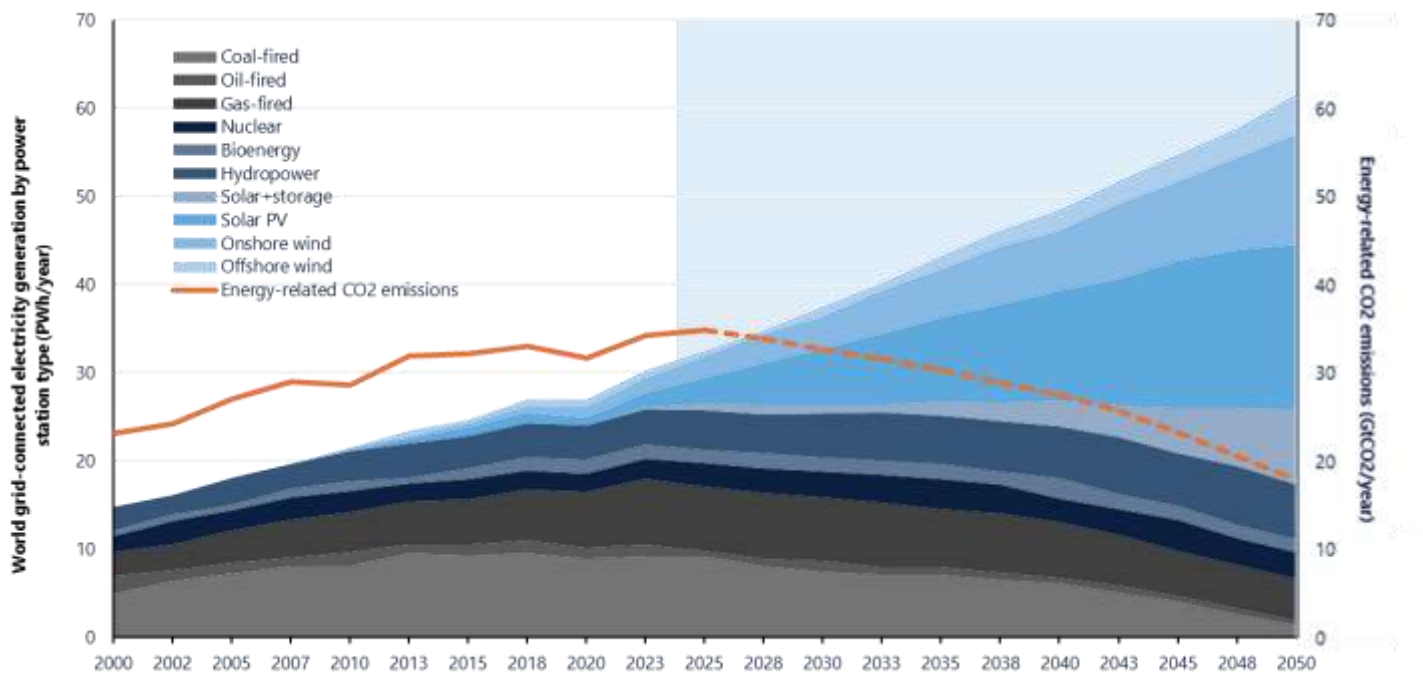
Data centres are a rapidly growing driver of global electricity demand, accounting for about 2% of electricity consumption in 2024. With the rise of energy-intensive technologies like AI, this demand is expected to double by 2030, surpassing 1,000 TWh annually, such that their escalating energy needs pose a significant challenge for sustainable energy systems. Meanwhile, electricity generation, historically reliant on fossil fuels, is shifting towards renewables like wind and solar to meet rising demand while reducing CO2 emissions. This transition is crucial, as global electricity use is set to double by 2050, driven by population growth and economic expansion. In this context, data centres embody the tension between increasing power requirements and the need for decarbonisation. With their energy intensity rising, integrating stable, low-carbon energy sources is vital to support their growth without undermining climate goals. By embracing renewable and nuclear energy solutions, and integrating emerging technologies, the AI industry has an opportunity to chart a sustainable path forward.

Data centre deployment is one catalyst for the emergence of a new electricity supercycle

As fossil fuels continue to dominate global electricity production, the combination of rising demand and increasing CO2 emissions stresses the need for a shift in paradigm. By 2050, global electricity demand should double from 30PWh in 2023 to 62PWh, driven by the accelerating shift to electrification. Amidst this surge, data centres power use (~500TWh in 2025, i.e. 2% of global demand) should exceed 1,000TWh by 2030, largely fueled by the energy-intensive demands of AI-driven workloads.

Historically, coal has been the largest source of global power supply followed closely by gas. On the low-carbon side, hydropower and nuclear power are the largest power contributors, with nuclear playing a particularly significant role in some countries. While wind and solar are growing rapidly, fossil fuels — coal, oil and gas — still dominate global power production. Combined, they are also the largest source of global carbon dioxide emissions, highlighting the need for a sustainable shift in power generation.

Fig. 32 - World grid-connected electricity generation by power station type (PWh/year)



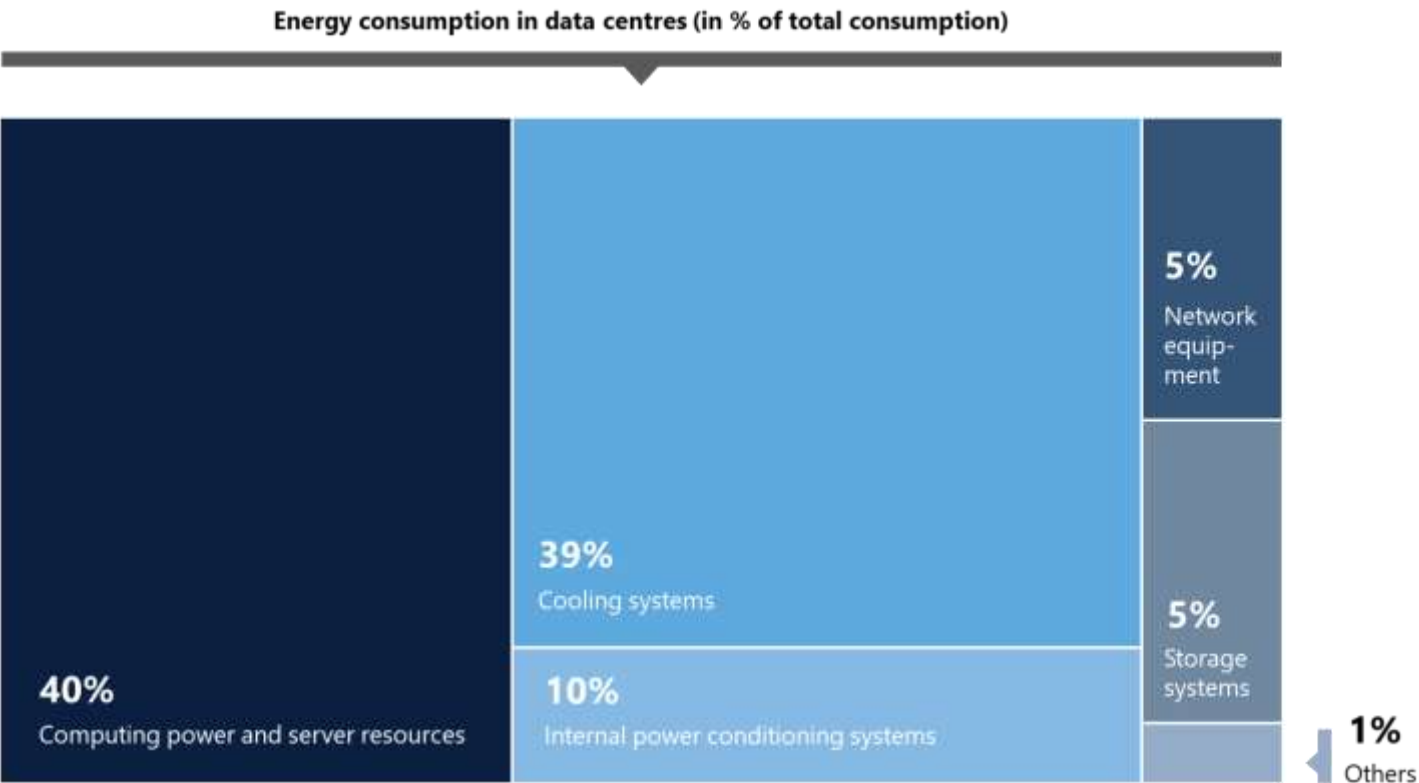
Source(s): IEA Web (2024), Global Data (2024), Stifel*

By 2050, the global population is expected to grow by 1.7bn while global GDP is set to increase by 89%. As a result, global energy demand is projected to rise significantly over the next 20 to 30 years. Whereas in 2023, electricity accounted for about 20% of the world's final energy use, by mid-century, this share should increase to close to 40%, with electricity demand doubling from 91EJ/yr in 2023 to approximately 180EJ/yr in 2050, reflecting average annual growth of 2.6%. Electricity is set to gradually replace other forms of energy, driving global electricity demand from 30PWh in 2023 to 62PWh by 2050 according to estimates from Det Norske Veritas (DNV).

Data centres will become a major source of growth in global electricity demand. In 2023, data centres contributed +90TWh (+0.3%) to growth in global electricity demand, with an average annual growth rate of nearly 17% since 2019. For instance, it is estimated that the hardware, specifically graphic processing units (GPUs), used to train GPT-3, the forerunner to ChatGPT, used 1,300MWh of electricity. That is equivalent to the energy used by 1,450 average US households each month. Once an AI model is trained, the process of running live data through it to produce a result is called 'inference' — in other words, inference is effectively the use of AI. A single inference uses a tiny amount of energy, but billions of inferences add up significantly.

Two broad areas drive most of the electricity consumption in a data centre: computing power and server resources (roughly 40% of data centre power consumption) and cooling systems (these consume 38-40% of power). Hyperscalers and large-scale data centre operators supporting generative AI and high-performance computing require particularly high-density infrastructure. While traditional data centres relied on CPUs consuming 150-200W per chip, modern GPUs now demand significantly more power. Nvidia’s latest AI GPUs ran at 400W until 2022, increased to 700W in 2023, and reached 1,200W in 2024. A large rack can house up to 72 GPUs and 36 CPUs, requiring power demands of 130kW. By 2027, average power density is expected to rise from 36kW per rack in 2023 to 50kW per rack, reflecting the growing energy intensity of AI-driven data centres. Recent announcements related to the Chinese company DeepSeek suggest, however, that the development of alternative models that are drastically less energy-intensive than the initial ones created could help reduce AI-related energy consumption.

Fig. 33 – Roughly 80% of data centre energy consumption comes from two sources



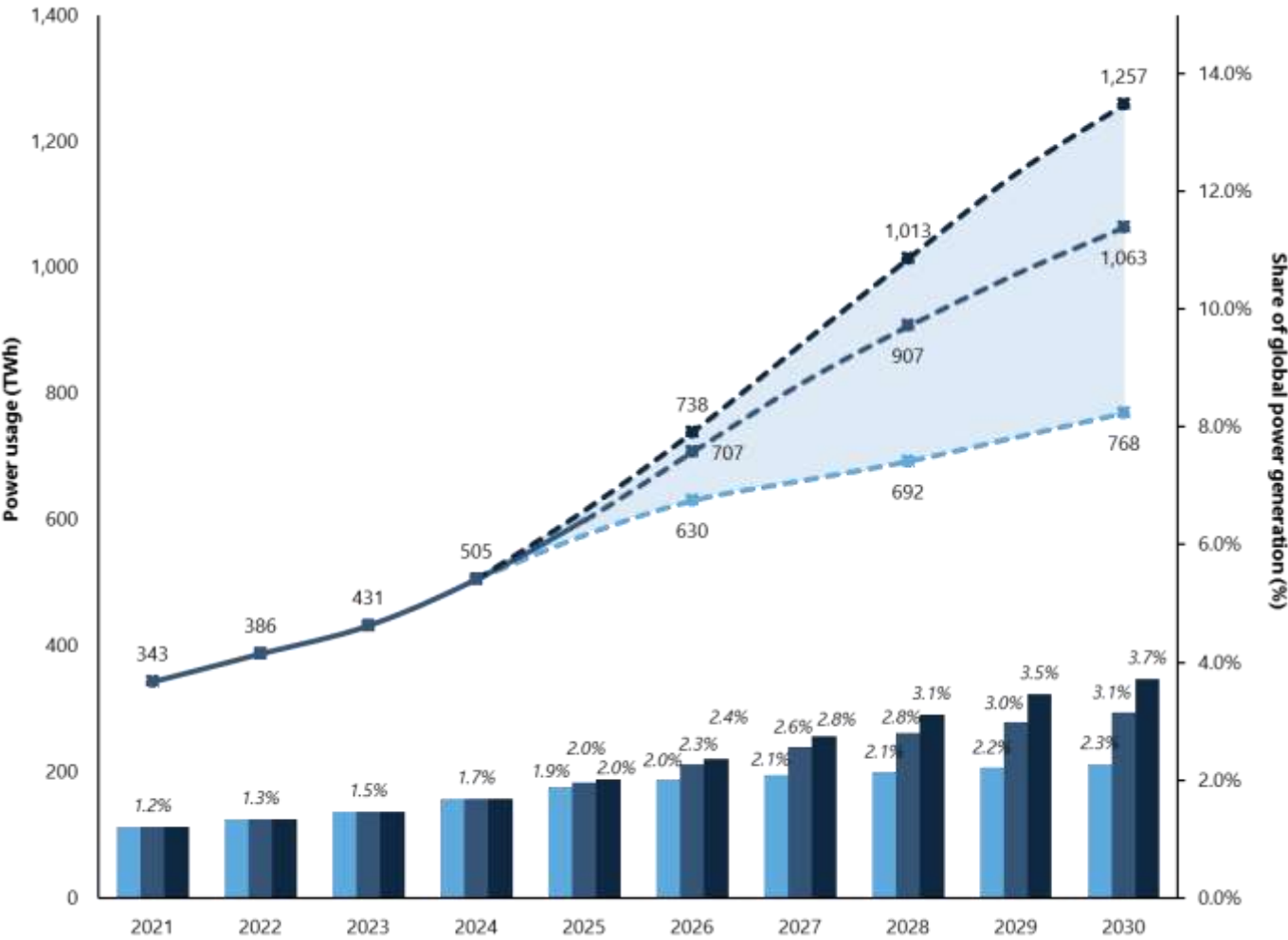
Source(s): Deloitte, Stifel*

We model global data centre power demand through 2030 across the three scenarios outlined in Part 1. Key assumptions include data centre capex growth, critical IT power additions, utilisation rates, and energy efficiency improvements (PUE). The model’s primary sensitivity lies in critical IT power additions: (i) capacity additions exert a near-linear impact on power demand as efficiency gains plateau; and (ii) infrastructure investment is largely a derivative of total data centre capex, given IT equipment’s relatively recurring capex share (3–6-year renewal cycles), while infrastructure remains a long-lived asset.

Our base case projects ~150% growth in data centre power demand from 2023 to 2030, reaching 1,063TWh, with its share of global demand doubling from 1.5% to 3.1%. It assumes an additional 82 GW of critical IT power over this period (i.e. +12 GW per year), reaching 139 GW, underpinned by strong investment in 2025 and a normalisation of expansion in 2027–2030 without a digestion phase.

- All scenarios indicate a material increase in data centres’ share of power demand, driven by current build-out investment and slowing efficiency gains. Even in our bear case—where investment contracts in 2026–2027—power demand grows at a CAGR of +9% from 2023 to 2030.
- Our bull case projects 2030 power demand to be just 18% above the base case, as energy and supply chain constraints would still limit infrastructure scaling significantly beyond our base case projections.

Fig. 34 – AI & data centre power demand set to exceed 1,000TWh/yr by 2030



1

Scenario
Cautious AI Consolidation

- **Data centre capex:** +3% CAGR 2024–30 to USD447bn, with an MSD contraction in 2026–27 before normalising at +7% CAGR in 2027–30
- **Capacity additions (2023–30):** 43GW, a 1.8x increase, averaging 6GW per year
- **Data centre power demand:** +9% CAGR (2023–30) to 768TWh, equating to 2.3% of global electricity consumption

2

Scenario
Steady AI Expansion

- **Data centre capex:** +10.5% CAGR 2024–30 to USD680bn, reflecting +26% growth in 2025 before moderating to sustained HSD growth through 2026–30
- **Capacity additions (2023–30):** 82GW, a 2.4x increase, averaging 12GW per year
- **Data centre power demand:** +14% CAGR (2023–30) to 1,063TWh, equating to 3.1% of global electricity consumption

3

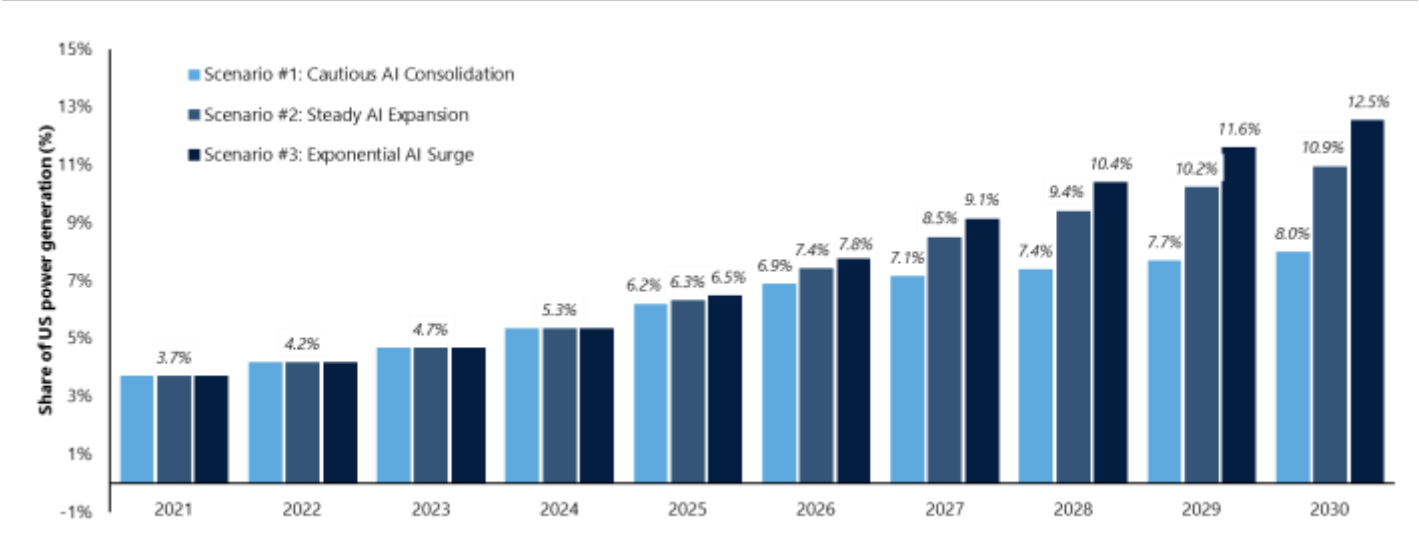
Scenario
Exponential AI Surge

- **Data centre capex:** +14% CAGR 2024–30 to USD822bn, driven by +35% in 2025 and DD growth in 2026–27, before normalising to HSD growth through 2027–30
- **Capacity additions (2023–30):** >100GW, a 2.8x increase, averaging 15GW per year
- **Data centre power demand:** +17% CAGR (2023–30) to 1,257TWh, equating to 3.7% of global electricity consumption

The rapid expansion of data centres in the US and Europe will intensify grid strain at a pace unmatched by other regions. While these two markets collectively represented roughly one-third of global electricity consumption in 2023, they accounted for nearly two-thirds of total data centre power demand—a share we estimate will rise to 70% by 2030. With data centre electricity usage projected to grow at a CAGR of +14% through to the end of the decade, the strain will be disproportionately concentrated in these regions, where the sector already commands a significant share of power consumption.

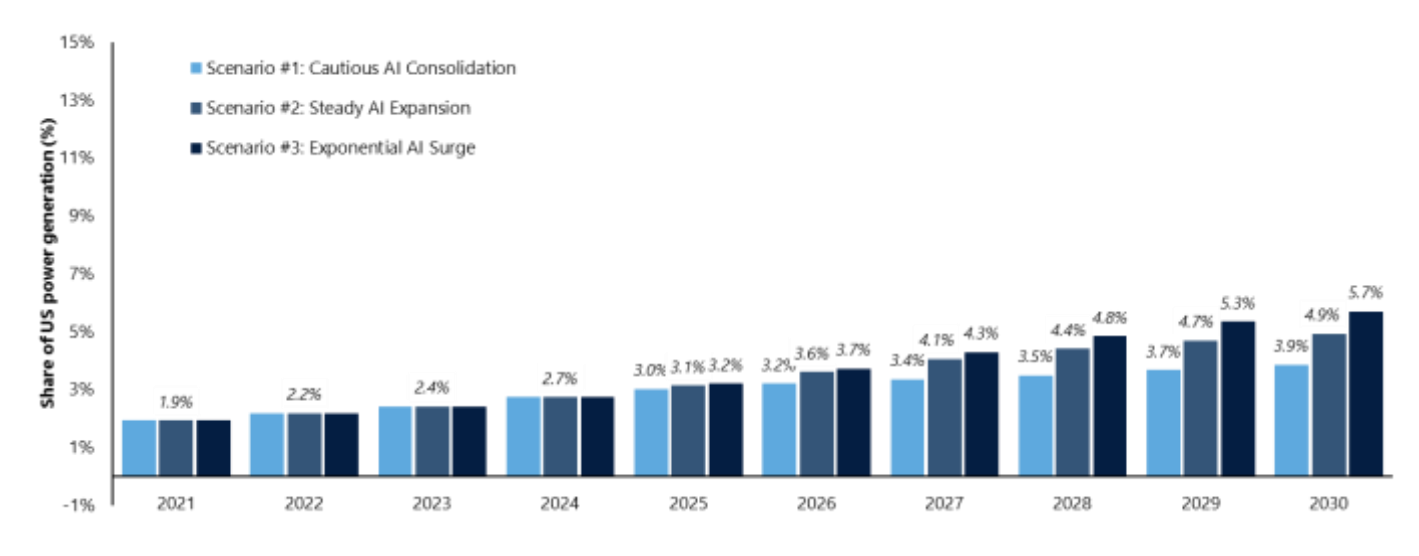
- **The impact will be strongest in the US, which has emerged as the global epicentre for AI-driven infrastructure expansion.** The country hosts most of both operational and planned AI clusters, making it the primary driver of surging data centre power needs. As a result, we project that data centres will account for nearly 11% of total US electricity consumption by 2030 at 556TWh—more than doubling from less than 5% in 2023.
- **In Europe, power constraints and high electricity costs may slow capacity growth, but data centre expansion will still reshape the energy landscape.** The data centre share of electricity consumption is expected to double, rising from an estimated 2.4% today to 4.9% by 2030. Currently, Europe holds close to 20% of global data centre capacity. Our base-case projections indicate a decline of 3pp by 2030, as the region’s data centre power demand is expected to grow at a +12% CAGR by 2030 (versus +16% in the US and +14% in RoW).

Fig. 35 – Data centres share in US power generation (%)



Source(s): Stifel*

Fig. 36 – Data centres share in Europe (EU30¹) power generation (%)



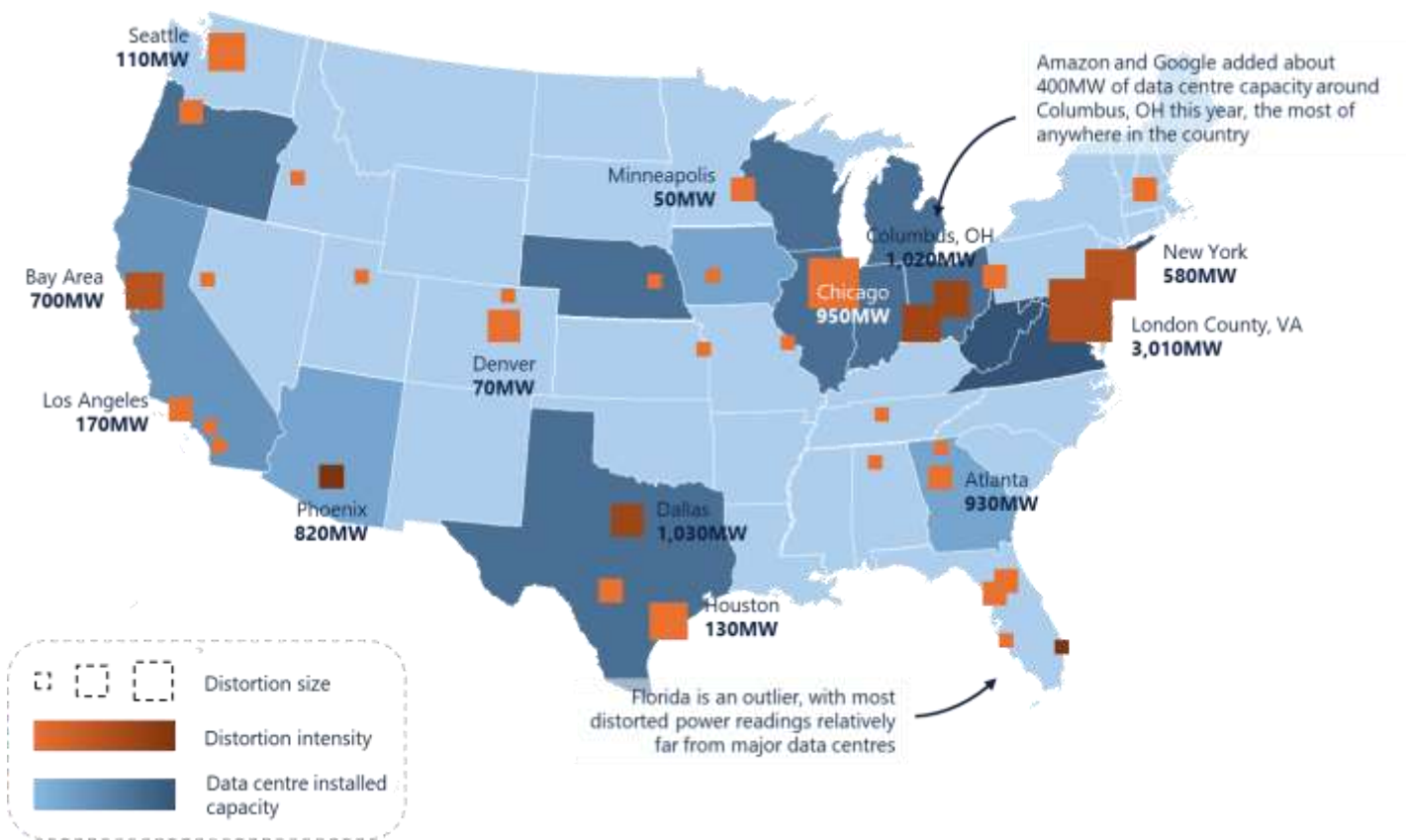
Source(s): Stifel*

Despite increased power density in data centres, efficiency gains have plateaued, meaning further improvements in energy usage per computational output are not strictly keeping pace. As a result, as power-intensive generative AI training and inference continues to grow faster than other uses and applications, global data centre electricity consumption could roughly double to more than 1,000TWh by 2030.

On top of this, growing demand for electricity from AI data centres is straining local power grids and leading to power quality issues in nearby residential areas. These centres consume vast amounts of energy, with some facilities requiring more than 100MW — equivalent to powering tens of thousands of homes. This demand sometimes exceeds the capacity of local grids, resulting in voltage distortions and irregular power supplies that can damage household appliances or reduce their lifespan.

A recent Bloomberg analysis found that 75% of distorted power readings in over 770,000 US homes occurred within a 50-mile radius of AI data centres. These distortions can manifest as voltage sags, surges or harmonic disruptions, creating significant challenges for residential power systems. These findings highlight the broader issue of integrating energy-intensive technologies into existing infrastructure, which often lacks the robustness to manage such increased demands efficiently.

Fig. 37 – AI data centres are responsible for power distortions across the US



Source(s): Bloomberg analysis of Whisker Labs and DC Byte Data, S&P Global Market Intelligence, Stifel*

To mitigate these challenges, grid upgrades, such as implementing advanced energy storage solutions and real-time monitoring systems, as well as the development of more energy-efficient AI systems will be required. Without these measures, the continued expansion of AI data centres risks worsening power reliability issues in residential areas and amplifying the environmental impact of heightened energy consumption. This becomes especially critical in the context of an increasing strain on the energy grid from the growing integration of renewable energy sources into energy mixes.

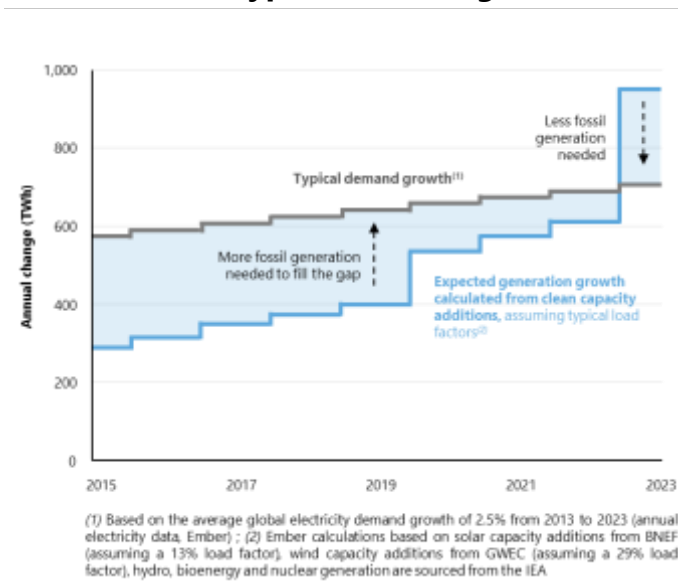
This has raised a challenge: securing enough stable, low-carbon power for data centres

Data centres face increasing challenges linked to the energy trilemma – availability, stability, sustainability – as demand surges, driven by AI and other energy-intensive technologies. While renewables play a growing role in the electricity mix, their intermittency - seen in over 100 hours of negative pricing in Europe in 2023 - creates grid instability, sometimes incompatible with the 24/7 energy supply requirement of data centres.

Historically, traditional electricity generation from coal, gas and oil was highly inefficient and lost a lot of primary energy as heat. It is estimated that in 1990, the global power system efficiency was a mere 39%. However solar PV and wind directly convert power to electricity without heat loss, making them close to 100% efficient. This shift to renewables has significantly improved the conversion rate of primary energy to useful energy, from 50% to around 75% today.

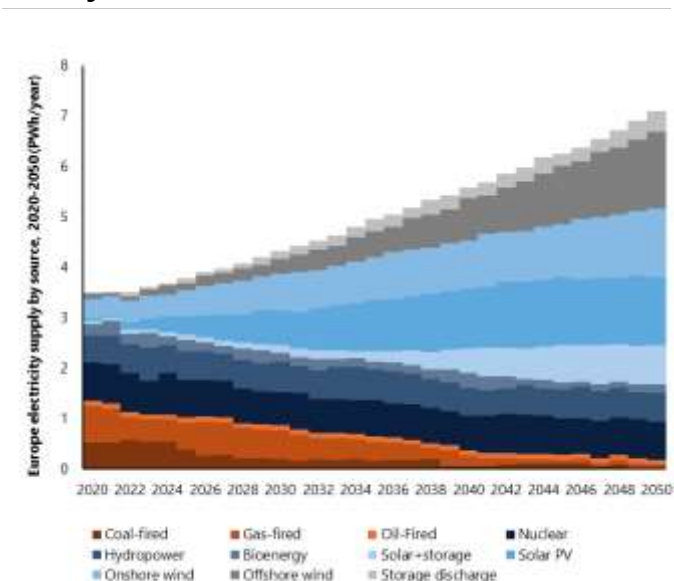
Today, the growing and greening of electricity is the standout feature of the energy transition. Global electricity demand is projected to double by 2050, with the power sector anticipated to achieve 90% decarbonisation by then. Electricity generation is expected to rise significantly from 9.2PWh/yr in 2023 to 24.4PWh/yr by 2035, driven largely by the rapid deployment of renewable energy technologies such as solar and wind.

Figure 38 - Clean power capacity growth in 2023 exceeded typical demand growth



Source(s): Ember (Global Electricity Review, 2024) Stifel*

Figure 39 – Europe’s electricity supply will mostly come from renewables

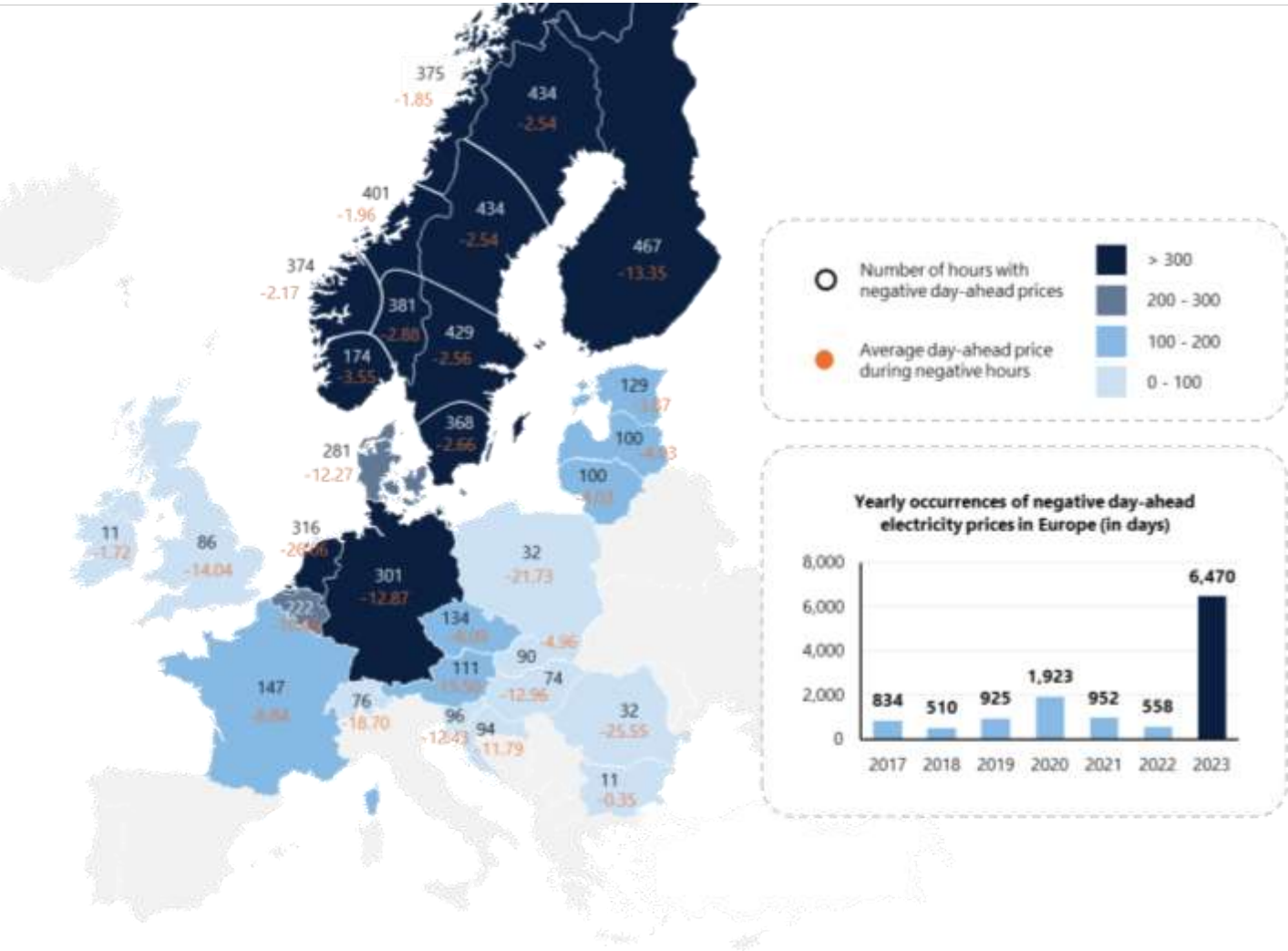


Source(s): Det Norske Veritas (DNV), Stifel*

However, the surge in electricity demand—forecast to increase by 12.1PWh/yr by 2035—will absorb much of this additional capacity in the short to medium term. This growth is fuelled by electrification in various sectors, including transport, industry and heating, as well as expanding access to electricity in developing regions. Despite these challenges, the transition away from fossil fuels is expected to accelerate in the late 2030s. By then, the sustained growth in renewable energy capacity should outpace the increase in demand, enabling a more significant decline in fossil fuel use.

But, the more renewables, the more intermittency. Unlike fossil fuel or nuclear power plants, which can produce electricity consistently, renewables depend on natural conditions that fluctuate over time. Solar power generation is affected by weather conditions such as cloud cover, and the day-night cycle, while wind power relies on wind speed, which can be unpredictable and varies by location, time of day, and season.

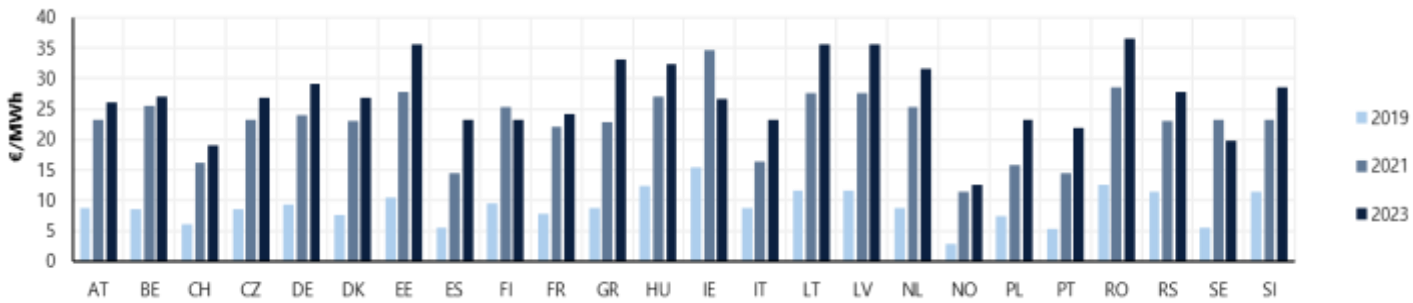
Fig. 40 - Negative hours and average day-ahead price during negative hours in EU in 2023



Source(s): ACER (2023), Stifel*

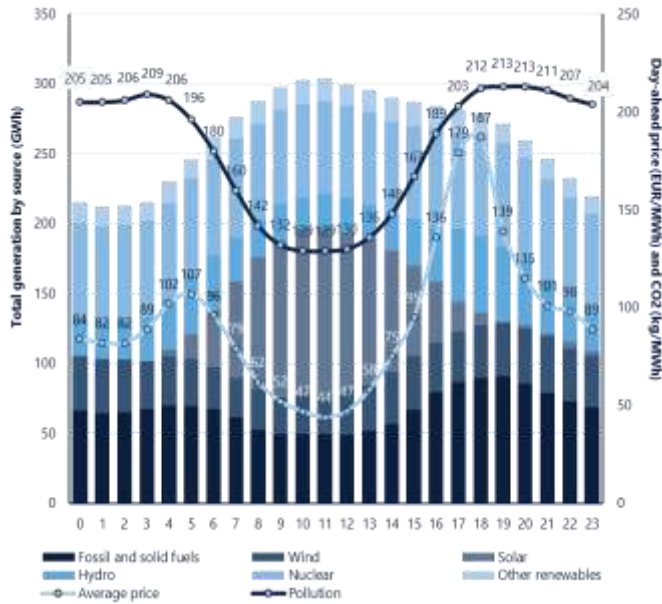
Consequently, the renewable energy mix is a key determinant of power supply stability. Indeed, relying too heavily on a single renewable source increases the risk of power shortages during periods of low generation and excess supply during periods of high generation. Case in point, the rise in negative price events in the last few years highlights the growing challenge of balancing electricity markets amid the intermittency of renewable energy production. In 2023, the average daily standard deviation of day-ahead electricity prices in Europe was three times higher than in 2020. Additionally, the number of prices exceeding EUR100/MWh in 2023 was 67% greater than in 2021 and the number of negative wholesale electricity prices surged significantly.

Fig. 40 - Negative hours and average day-ahead price during negative hours in EU in 2023



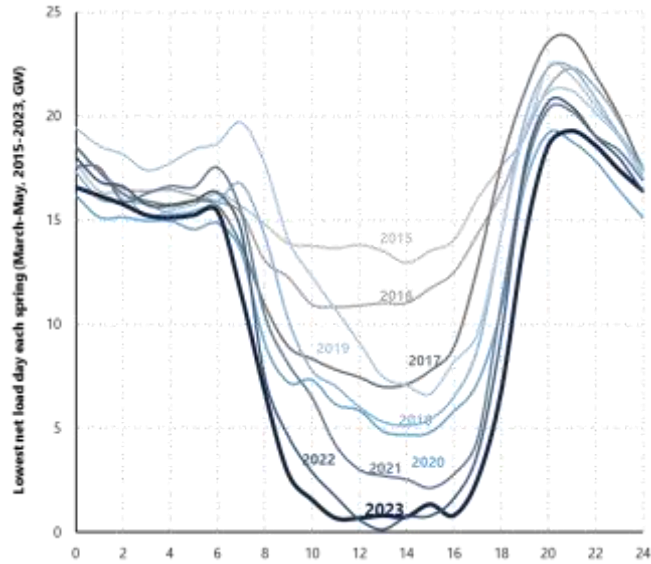
Source(s): Forschungsstelle für Energiewirtschaft (FfE), Stifel*

Figure 42 – Average hourly generation in the EU-27 in August 2024



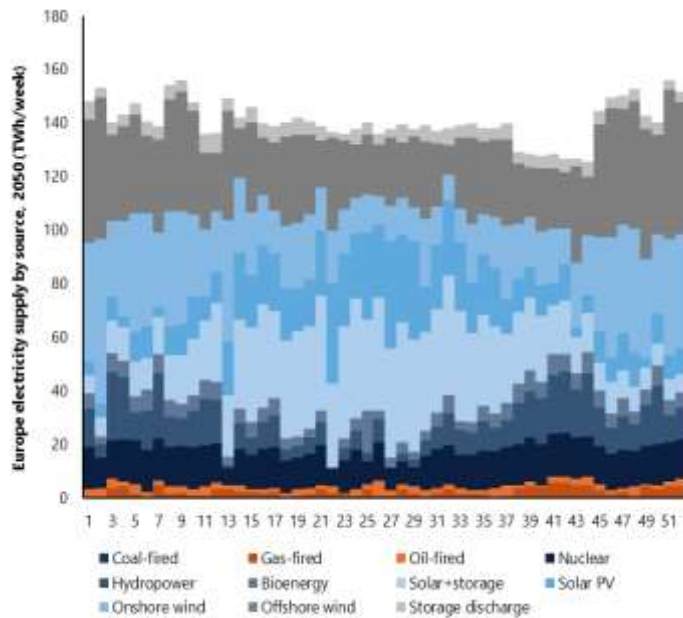
Source(s): ACER (2024), Stifel*

Figure 43 – California duck-curve, lowest net load day each spring (March-May 2015-23)



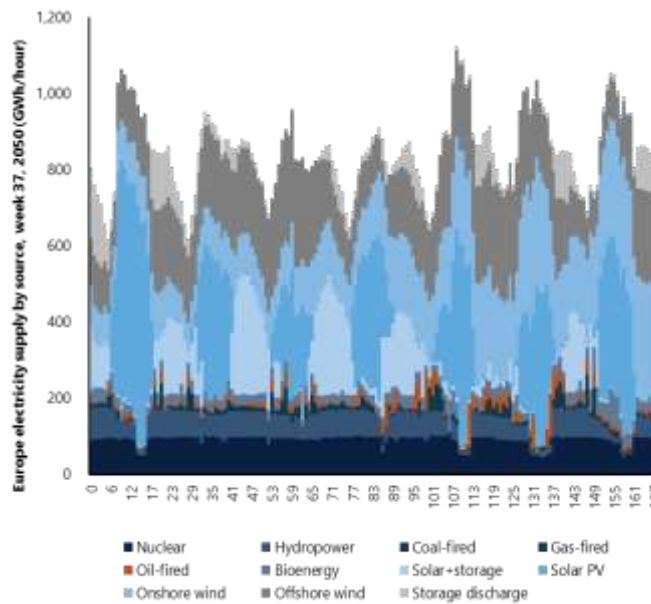
Source(s): CAISO, Stifel*

Figure 44 - Europe electricity supply by source, 2050



Source(s): Det Norske Veritas (DNV), Stifel*

Figure 45 – Europe electricity supply by source, week 37, 2050



Source(s): Det Norske Veritas (DNV), Stifel*

The frequency of negative day-ahead electricity prices in Europe increased 12x from 2022 to 2023. In Finland, Norway and Sweden, the occurrence rose 20x, driven primarily by high hydropower feed-in combined with strong wind and solar generation in the Northern Europe. In countries with substantial solar capacity, such as Germany, the Netherlands and Spain, negative prices are most common at midday when solar irradiance peaks. This is also the case in the US where some states, such as California, are subject to ample daily load variations. This general trend underlines how rising renewable energy generation affects grid stability, which is essential for the 24/7 energy supply data centres require.

The sector needs alternatives, nuclear innovations can help fill the gap

Data centre operators have objectives and regulatory obligations that require them to green their activities, particularly by adopting cleaner electricity. However, as discussed earlier, renewable energy in its current form has limitations. In light of these constraints, it is becoming essential for hyperscalers to explore alternatives. Among these options, nuclear energy is increasingly emerging as part of the solution.

Regulations for data centres vary significantly across the globe. Some countries and regions have established clear standards, for instance, Germany mandates that both new and existing data centres meet specific Power Usage Effectiveness (PUE) targets within the next four years. In contrast, regulation in the US remains relatively lax, with the exception of California, which enforces certain efficiency requirements through building regulations. Nevertheless, many US-based companies may still be impacted by regulatory developments in other countries.

Figure 46 – Selected regional and data centre operators’ climate-related targets

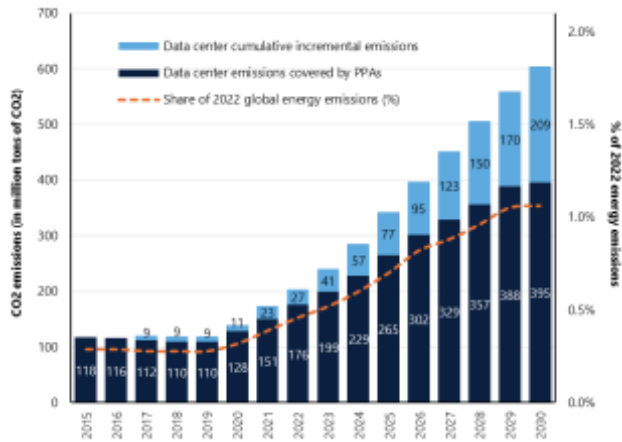


Source(s): Company disclosures, S&P Global, Stifel*

However, many companies have been proactive in establishing their own targets. By 2030, many companies aim to reduce direct (Scope 1) and indirect (Scope 2) carbon emissions, with some even pursuing net-zero or carbon-neutral goals across their entire value chain (Scope 3) - using carbon credits, carbon removal strategies and increased reliance on renewable or low-carbon energy sources. Many major tech firms have set specific goals to cut power-related emissions by 25% to 50%, despite growing energy demand and competition for renewable capacity.

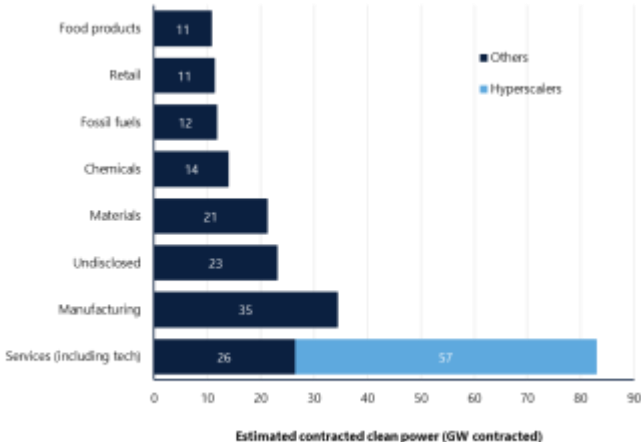
Nonetheless, in the short term, data centre operators are expected to prioritise power stability over the adoption of clean energy sources. Tech majors have traditionally been the leading buyers of Power Purchase Agreements (PPAs) – i.e. agreements whereby a buyer commits to purchasing electricity producer at a predetermined price – with renewable energy producers. In 2022, Amazon led globally as the largest buyer with 10.9GW of corporate PPA deals. In comparison, other major tech companies with significant clean power purchases that year included Meta (2.6GW), Google (1.6GW), and Microsoft (1.3GW). Hyperscalers now dominate in securing PPAs and successfully competing with other industries for new capacity, largely aided by their size and financial strength.

Figure 47 – Most data centres’ CO2 emissions are now covered by PPAs



Source(s): IEA (2023), Stifel*

Figure 48 – Hyperscalers dominate clean energy procurement

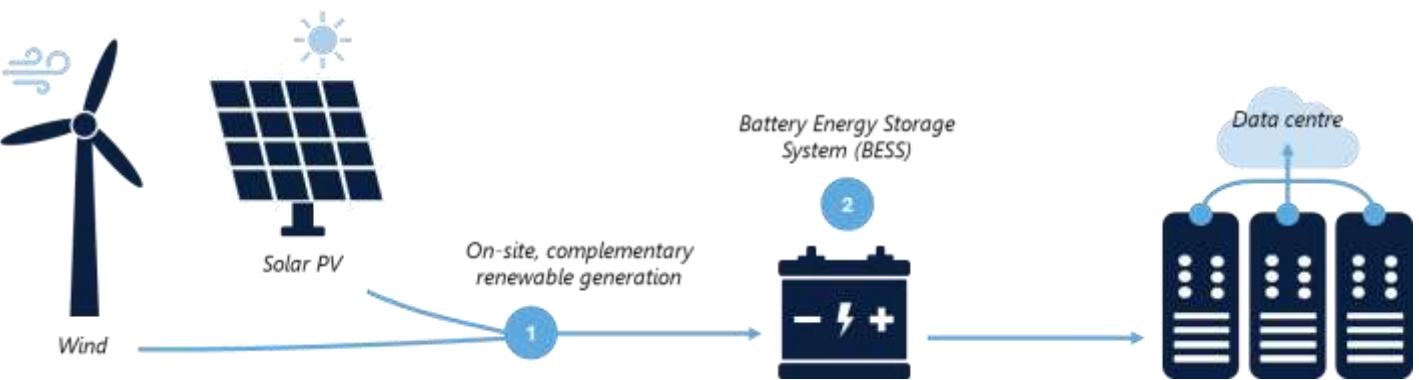


Source(s): S&P Global, Stifel*

However, PPAs, which were originally developed under vastly different energy conditions, are struggling to meet their obligations at the necessary levels. While these programmes were intended to promote the adoption of renewable energy, they have proven insufficient in bridging the gap between demand and availability.

On-site generation and Battery Energy Storage Systems (BESS), especially when combined, can present an alternative and enable data centre operators to stabilise their procurement of clean energy. On-site generation, such as solar panels or wind turbines, allows data centres to produce their own clean energy directly at the facility, reducing dependence on external energy sources. When combined with Battery Energy Storage Systems (BESS), these centres can store excess energy generated during periods of high production and use it during times of low generation or high demand. This combination ensures a more stable and reliable supply of clean power, minimising fluctuations and enhancing sustainability.

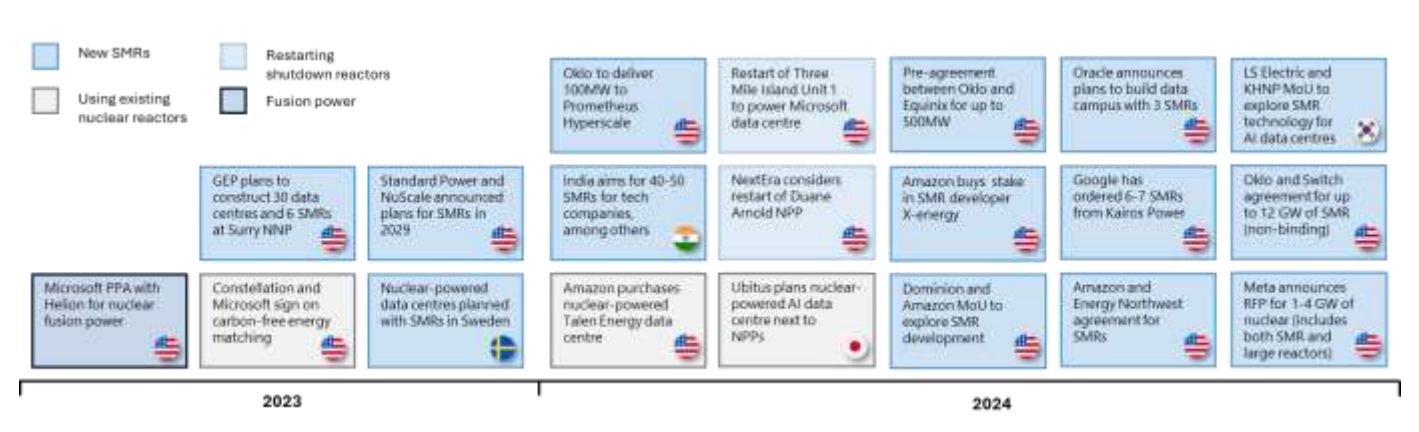
Fig. 49 – On-site generation and energy storage are two alternatives to help provide clean power to data centres



Source(s): IEA (2023), Stifel*

However, recent announcements from leading tech companies about agreements to source nuclear power indicate that they view the technology as crucial to their efforts in securing large-scale, carbon-free, and reliable energy. Although these solutions may not meet short- to medium-term energy needs, longer lead times and development timelines (seven years on average since the 2000s) for data centre projects make innovative nuclear technology developments increasingly justifiable.

Fig. 50 – Nuclear project announcements by data centre operators have multiplied since 2023



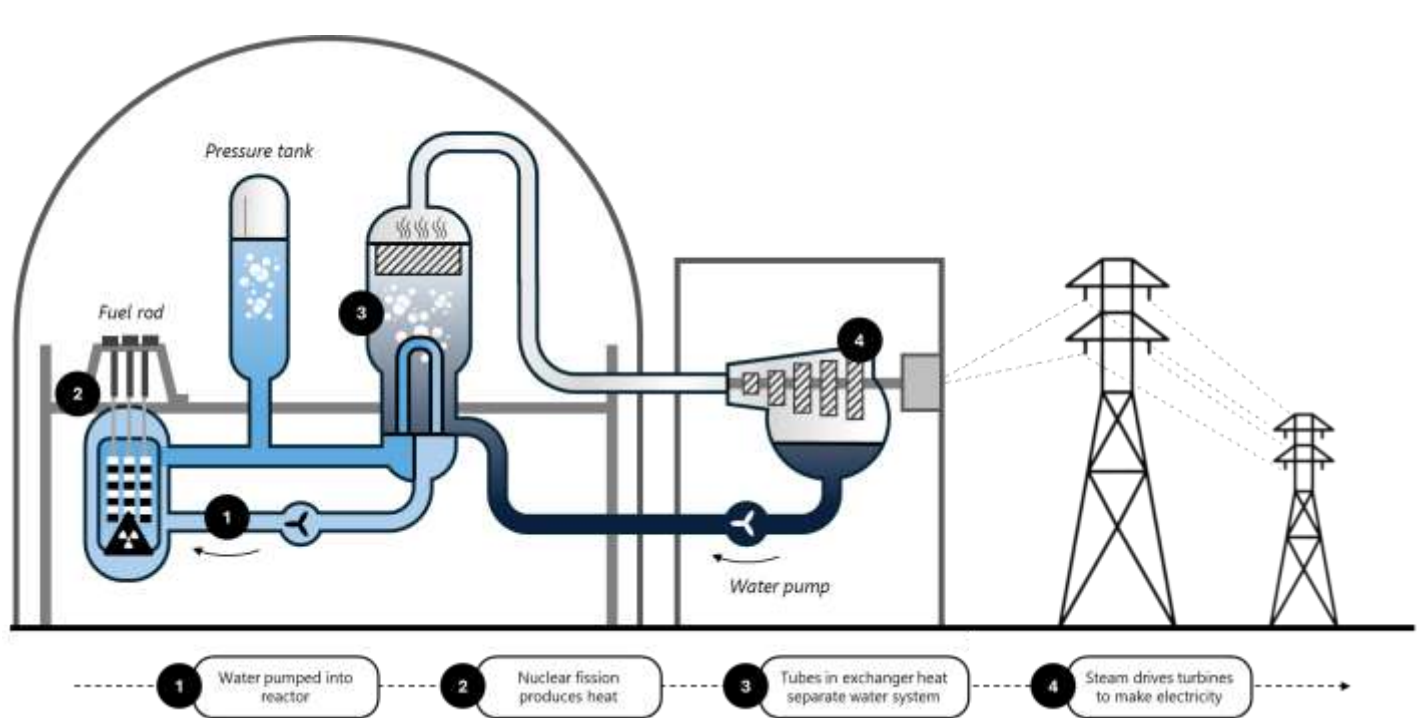
Source(s): IEA, Company disclosures, Stifel*

Footnote(s): PPA: Power Purchase Agreement ; RFP: Request For Proposals ; NPP: Nuclear Power Plant

Nuclear power has traditionally been a reliable, carbon-free source of continuous electricity. Nuclear energy is now the largest source of non-emitting electricity generation in OECD countries and the second largest source world-wide after hydro-power. Although access to enriched uranium fuel can be a challenge, it provides a stable energy supply with reduced reliance on large-scale fuel imports.

A variety of technologies have been developed but a nuclear power plant typically generates electricity through nuclear fission, where atoms (such as uranium-235 or plutonium-239) are split to release heat, which is absorbed by a coolant and transferred to produce steam that spins a turbine connected to a generator, converting mechanical energy into electricity, with the steam then cooled, condensed and recycled for reuse.

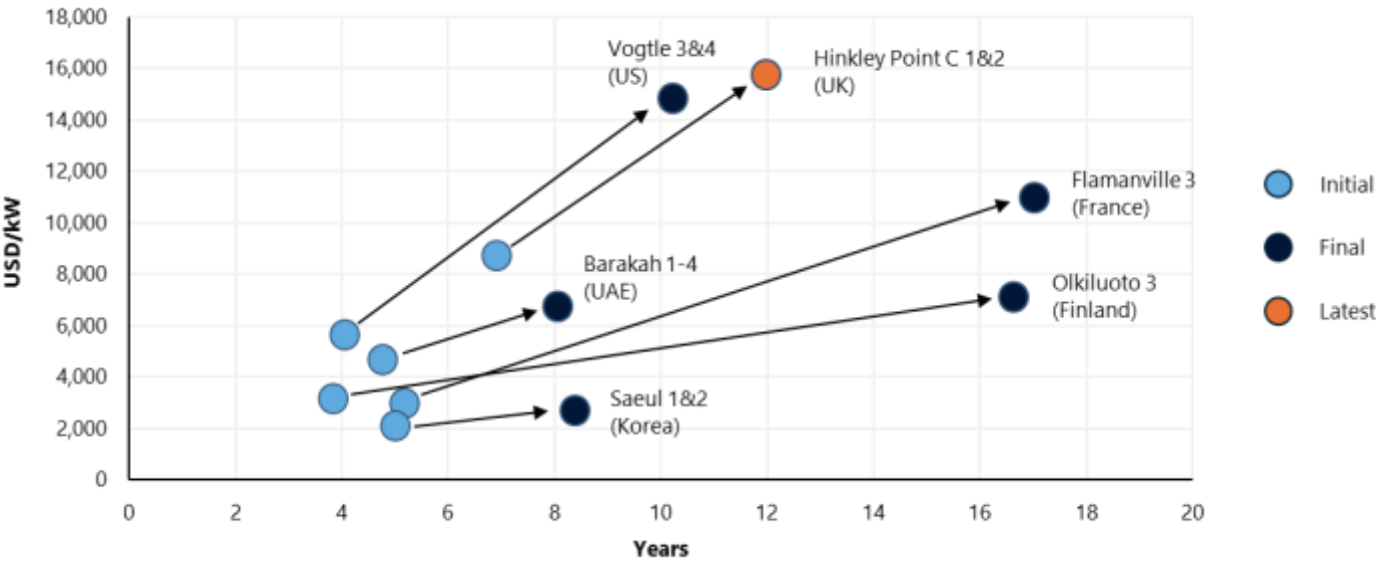
Fig. 51 – Schematic layout of a water-cooled nuclear power plant



Source(s): Stifel*

But construction times for traditional nuclear power plants typically play a crucial role in total costs, as delays result in significant cost overruns. Nuclear plants generally take longer to build than fossil fuel or renewable power stations due to their larger scale, technological complexity and stricter regulations. Globally, since 2000, the average construction time for a nuclear reactor has been about seven years, although advanced economies have often experienced much longer delays, sometimes exceeding a decade.

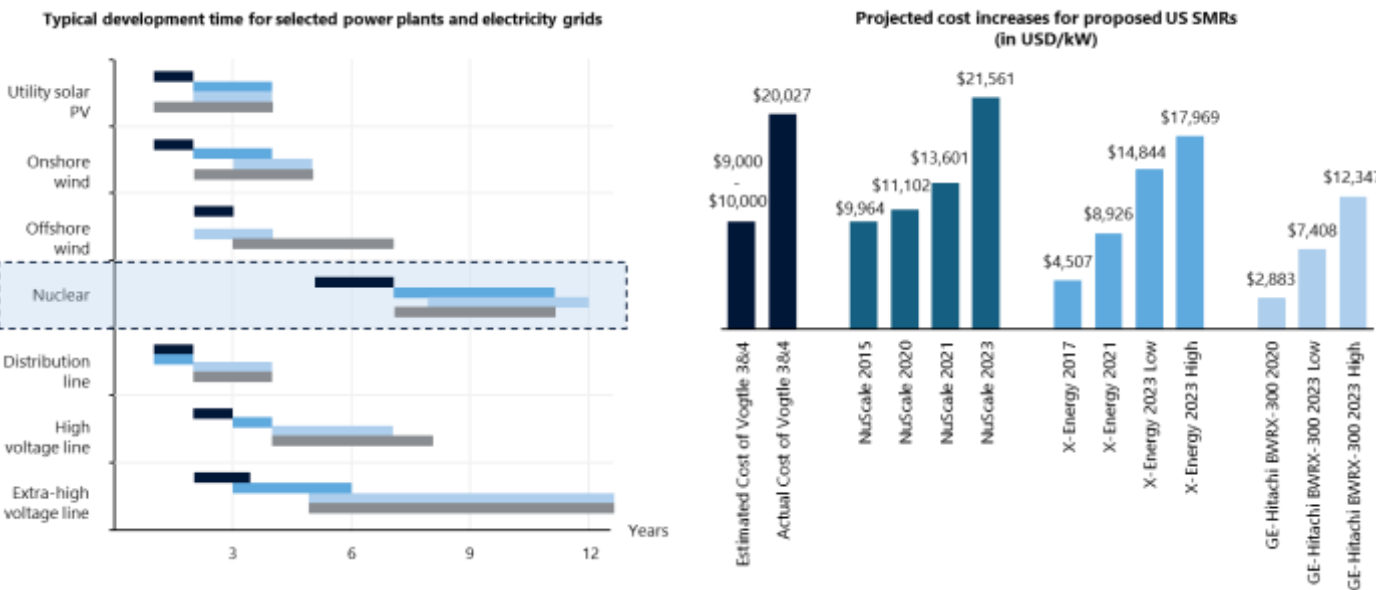
Fig. 52 – Initial and latest capital cost estimates and construction time for selected nuclear projects



Source(s): IEA, Company disclosures, Stifel*

The time and costs required to build traditional nuclear power plants varies considerably across different countries and regions. Countries like China have managed to build nuclear reactors more quickly, averaging seven years between 2017 and 2023, while Korea has completed recent projects within four to six years. Projects in the United States, such as Vogtle Units 3 and 4, and in Europe, like Olkiluoto 3, Hinkley Point C, and Flamanville 3, have faced extensive delays and cost escalations due to factors such as new reactor designs, regulatory challenges and the need to rebuild industrial skills after a period of inactivity.

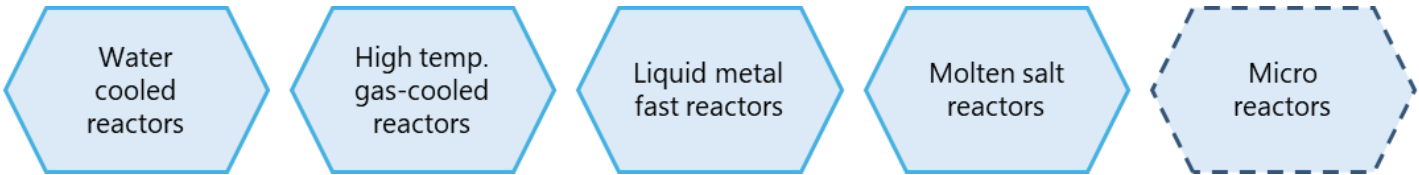
Fig. 53 – Nuclear projects often face developing time and cost overruns



Source(s): IEA, IEEFA, Company disclosures, Stifel*

Given conventional nuclear constraints, the Small Modular Reactors (SMR) technology is increasingly praised as the next-generation technology that can provide a solution to power new data centre needs. SMRs are a design concept referring to the size, capacity and standardisation for in-factory mass production of nuclear reactors. As a class of reactors, SMRs are defined by their smaller size, but a considerable variety within this class of reactors exist: they vary by power output, temperature output, technology and fuel cycle.

Fig. 54 – Major technology lines of Small Modular Reactor (SMR) technologies



Source(s): Internation Atomic Energy Agency (IAEA), Stifel*

A number of SMRs are based on existing commercially deployed light water technologies, while others are based on advanced design concepts. Typically, they offer a range of sizes, from as small as 1MWe to over 300MWe, and a range of temperatures, from 285°C to more than 850°C, to meet the specific energy needs of different sectors. Some SMR technologies are already demonstrated (at lab and commercial scales), while others are still in development. Timelines for deployment vary based on technology and regulatory readiness levels. Although SMR developers aim to roll out commercially available Small Modular Reactors (SMRs) between the late 2020s and early 2030s, they face significant challenges, including refining the technology, securing regulatory approval and establishing sustainable business models.

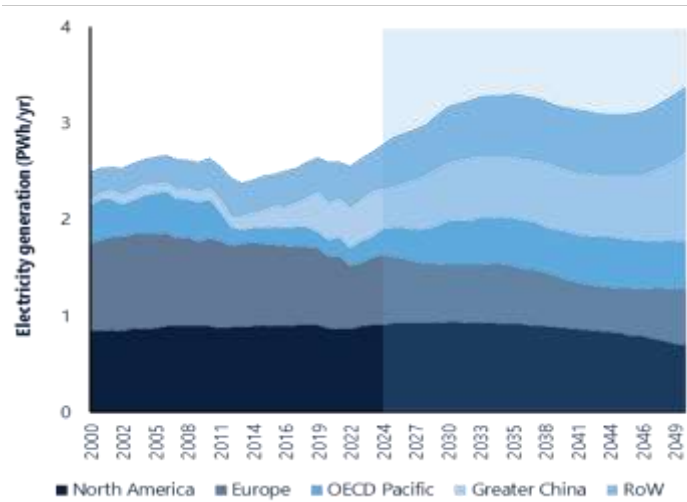
Fig. 55 – Momentum for SMRs is growing around the world



Source(s): Nuclear Energy Agency (NEA), Stifel*

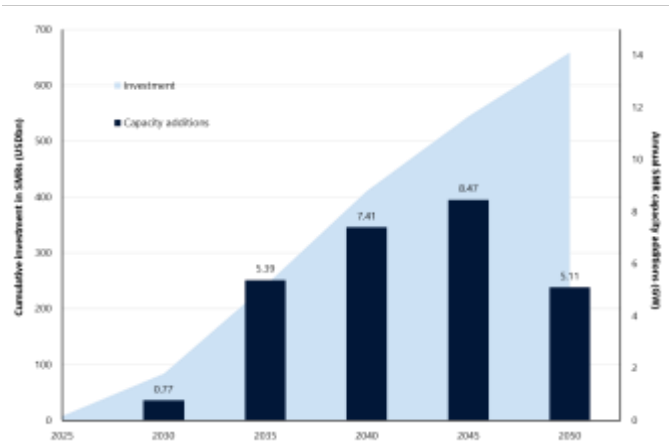
Substantial market growth is expected by the private sector nuclear industry, with nuclear energy output projected to steadily grow by 2% p.a. from current levels over the next decade. Output is expected to stabilise in the mid-2030s before slightly declining in the mid-2040s, not due to a reduction in new capacity, but because many older nuclear plants will be decommissioned. From now until 2030, most new capacity will come from traditional site-built, large-scale reactors already in development. After 2030, the additional capacity **will likely include a mix of site-built and factory-manufactured Small Modular Reactors (SMRs).** Nuclear energy output is expected to peak at 3,400TWh/year by 2050, a 30% increase from current levels.

Figure 56 – Change in nuclear power output by region (PWh/yr)



Source(s): IEA Web (2024), Det Norske Veritas (DNV), Stifel*

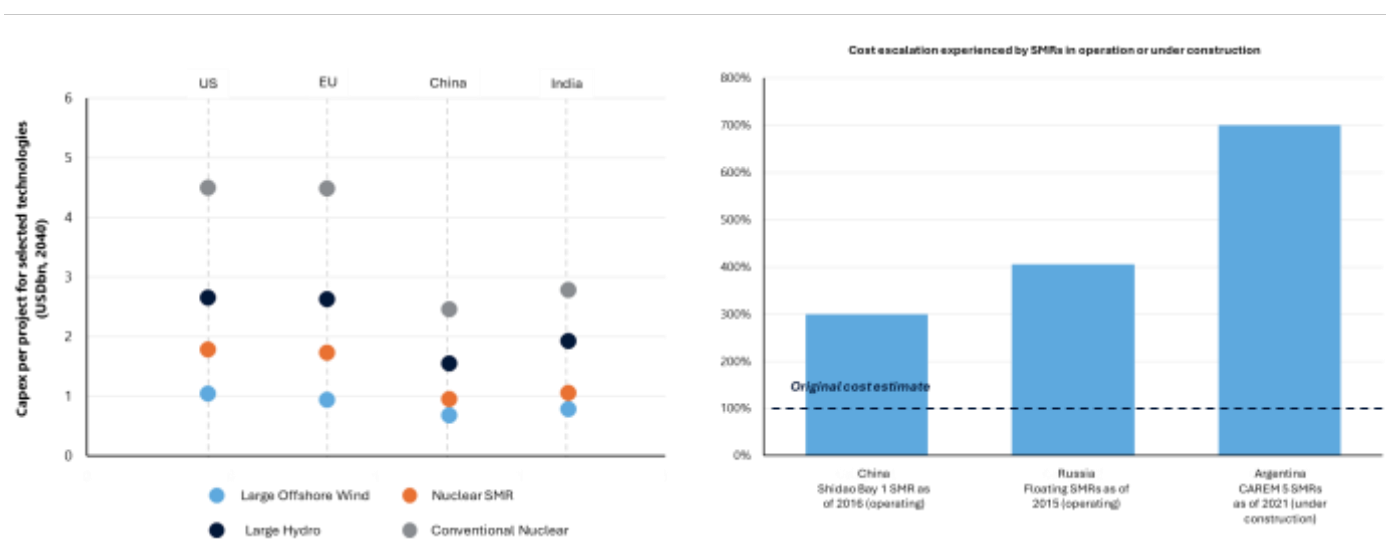
Figure 57 – Global SMR cumulative investment and capacity additions



Source(s): IEA Web (2024), Det Norske Veritas (DNV), Stifel*

One of the key advantages of deploying Small Modular Reactors (SMRs) is that they have the potential to foster increased participation from the private sector. SMRs, with their smaller scale, offer a more attractive option for investors, requiring less capital compared to the over USD10bn typically needed for conventional nuclear plants. This lower investment makes SMRs more accessible to private finance institutions and reduces the need for multiple investors to share risks. If SMRs achieve cost parity per megawatt through standardised designs, their shorter pre-project and construction timelines, combined with lower financing costs, could significantly reduce the payback period, potentially by up to 10 years. This faster return on investment then allowing for earlier net cash inflows, enabling the reinvestment of capital into new projects and stimulating further market growth.

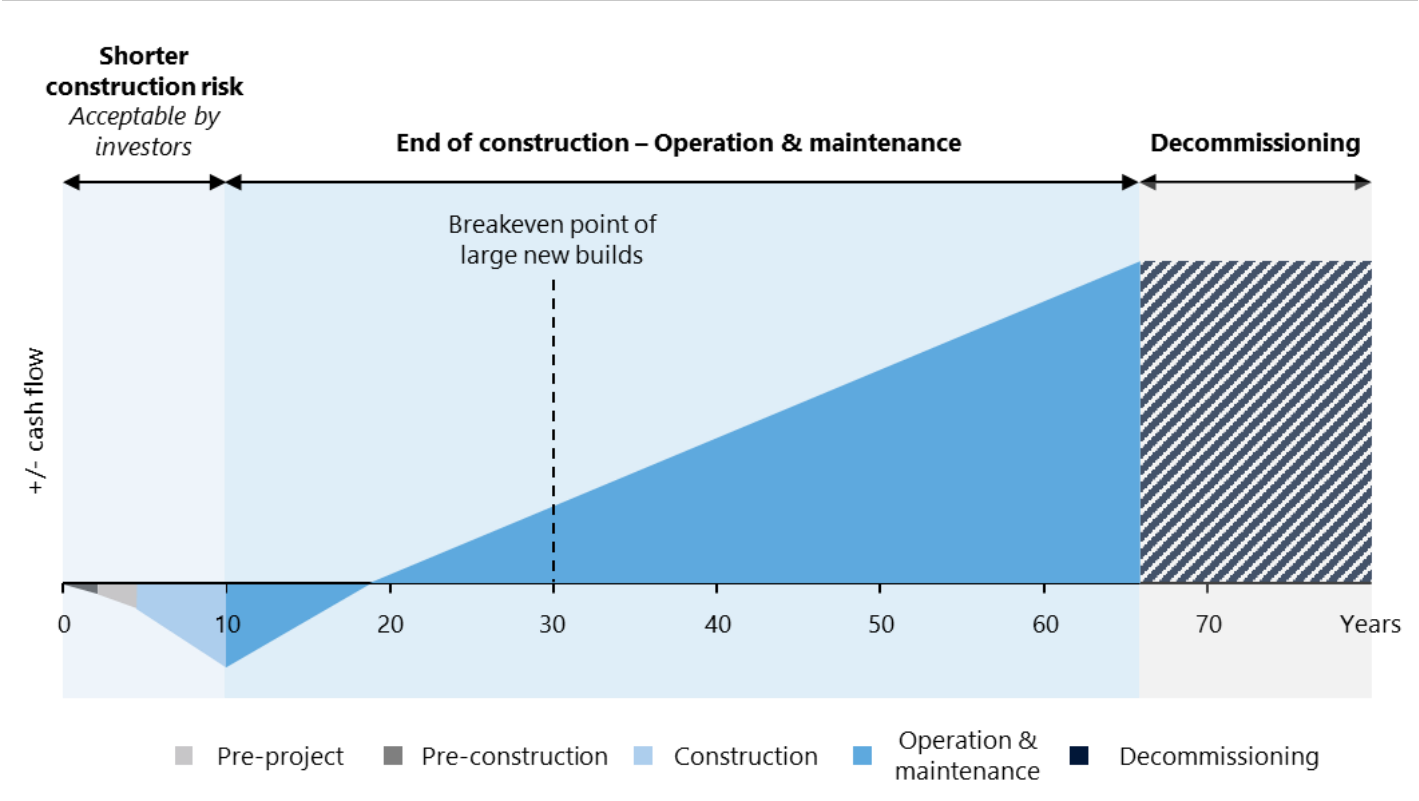
Fig. 58 – Capex for SMR projects tend to be lower than some alternative renewables but are subject to significant overruns



Source(s): IEA, IEEFA, Stifel*

However, most proposed SMRs still require further design work, regulatory licensing, scaling to commercial size, and pre-operational testing. Experience with past reactor projects suggests potential cost increases and delays. This means potential investors may want to insist upon fixed-price contracts, as a developer's willingness to agree to one could indicate confidence in their costs estimates and help manage associated risks.

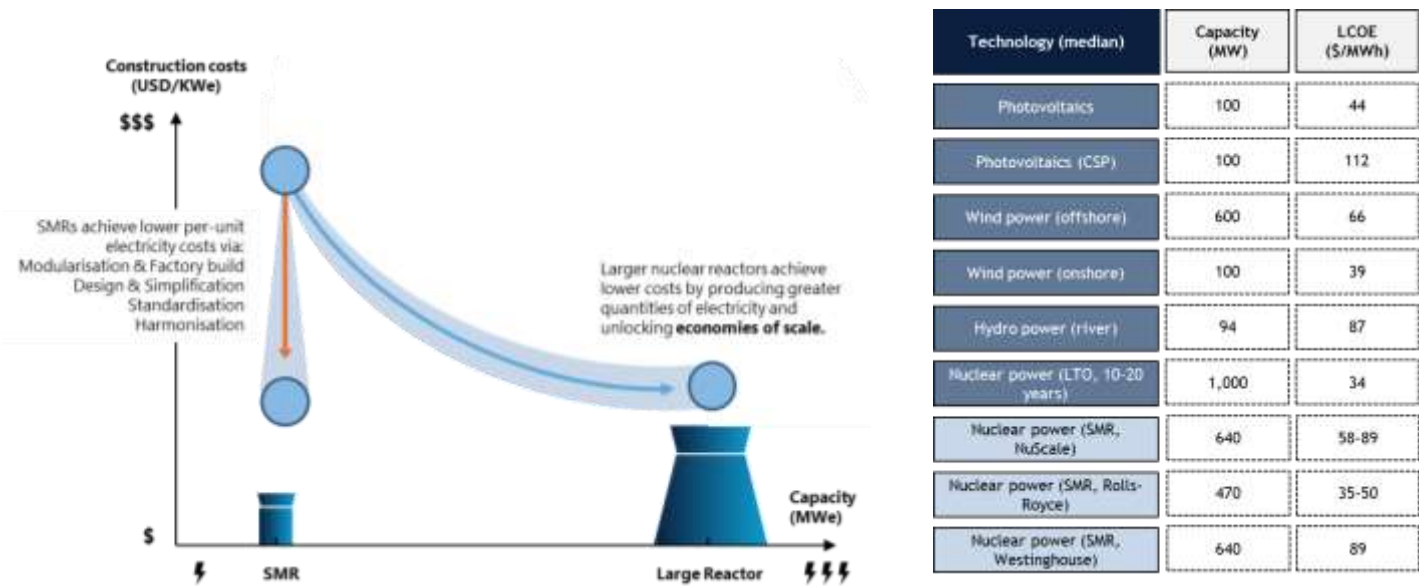
Fig. 59 – Indicative cumulative cash flow profile of an SMR power plant assuming cost parity with a conventional large-scale nuclear plant



Source(s): IEA, Stifel*

As a result, nuclear power presents a complex picture. As nuclear power production shifts increasingly to emerging countries, and with the long-term potential of Small Modular Reactors (SMRs), the average Levelized Cost of Electricity (LCOE) of nuclear energy is expected to decline to USD70-80/MWh. This reduction will be driven by shorter construction timelines and more favorable financing conditions, with some projects possibly reaching as low as USD50/MWh. Even so, nuclear predictions should be approached with caution, as sector data is often shaped by a few large projects, and cost overruns in developed countries have driven up the LCOE.

Fig. 60 – Key economic benefits of SMRs should partially offset diseconomies of scale

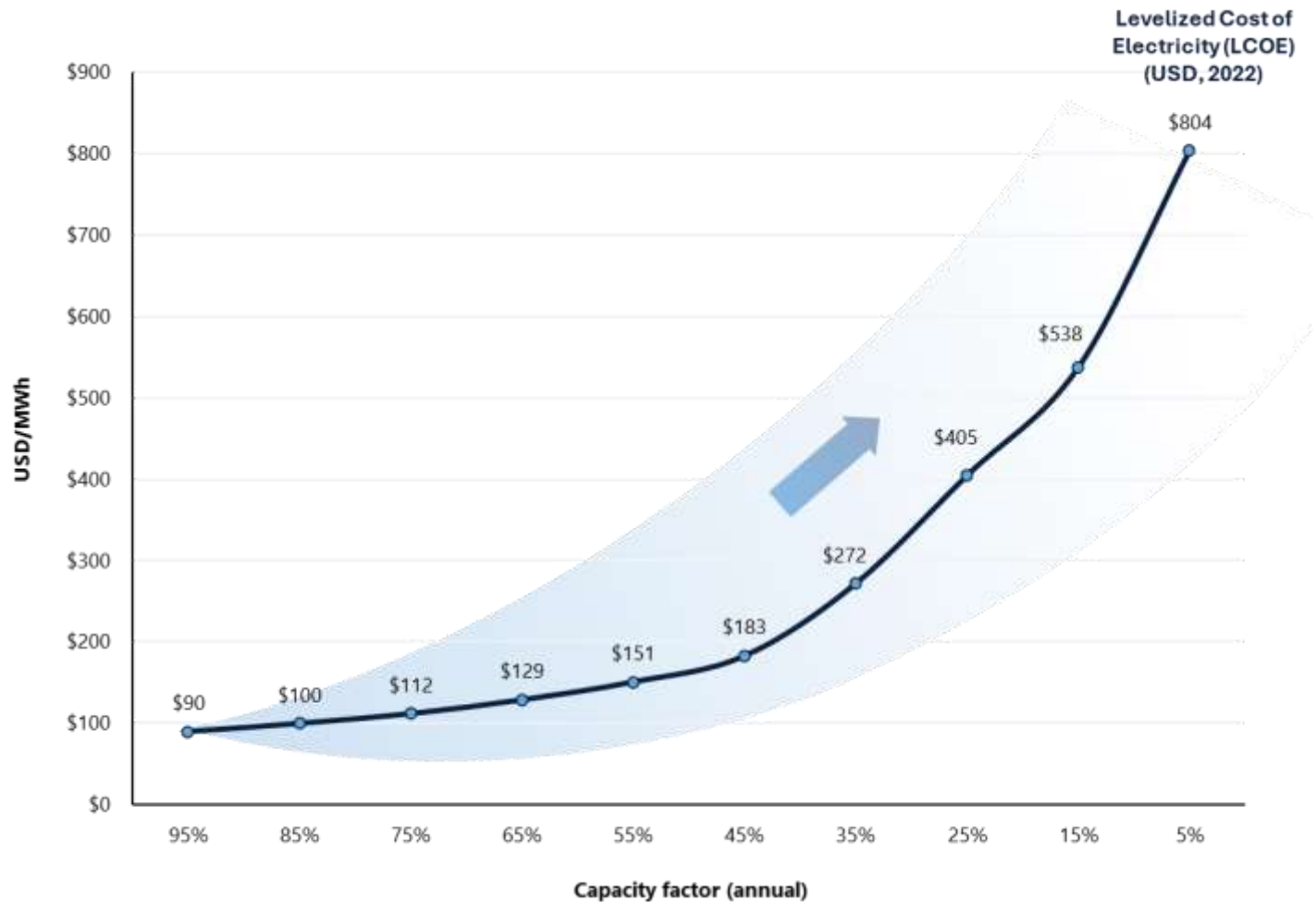


Source(s): Nuclear Energy Agency (NEA), Schneider Electric, Stifel*

Footnote(s): LCOE = Levelised Cost of Energy

SMR innovations may lower nuclear costs in the long-term but are expected to increase costs in the short term due to the need for investment in manufacturing facilities, higher material requirements and the time needed for economies of mass production to take effect. The achievement of a positive learning curve in SMR development will largely depend on the number of units built for each design. With around 80 SMR designs currently being proposed and marketed worldwide, according to the International Atomic Energy Agency (IAEA), it remains highly uncertain how many of each will be constructed. If too few are built, cost savings over time may not materialise, and modular factory-based construction may not be economically viable. Current cost estimates for SMRs exceed those of conventional renewables, but **their value lies in providing secure and continuous power rather than intermittent energy**.

Fig. 54 – Major technology lines of Small Modular Reactor (SMR) technologies

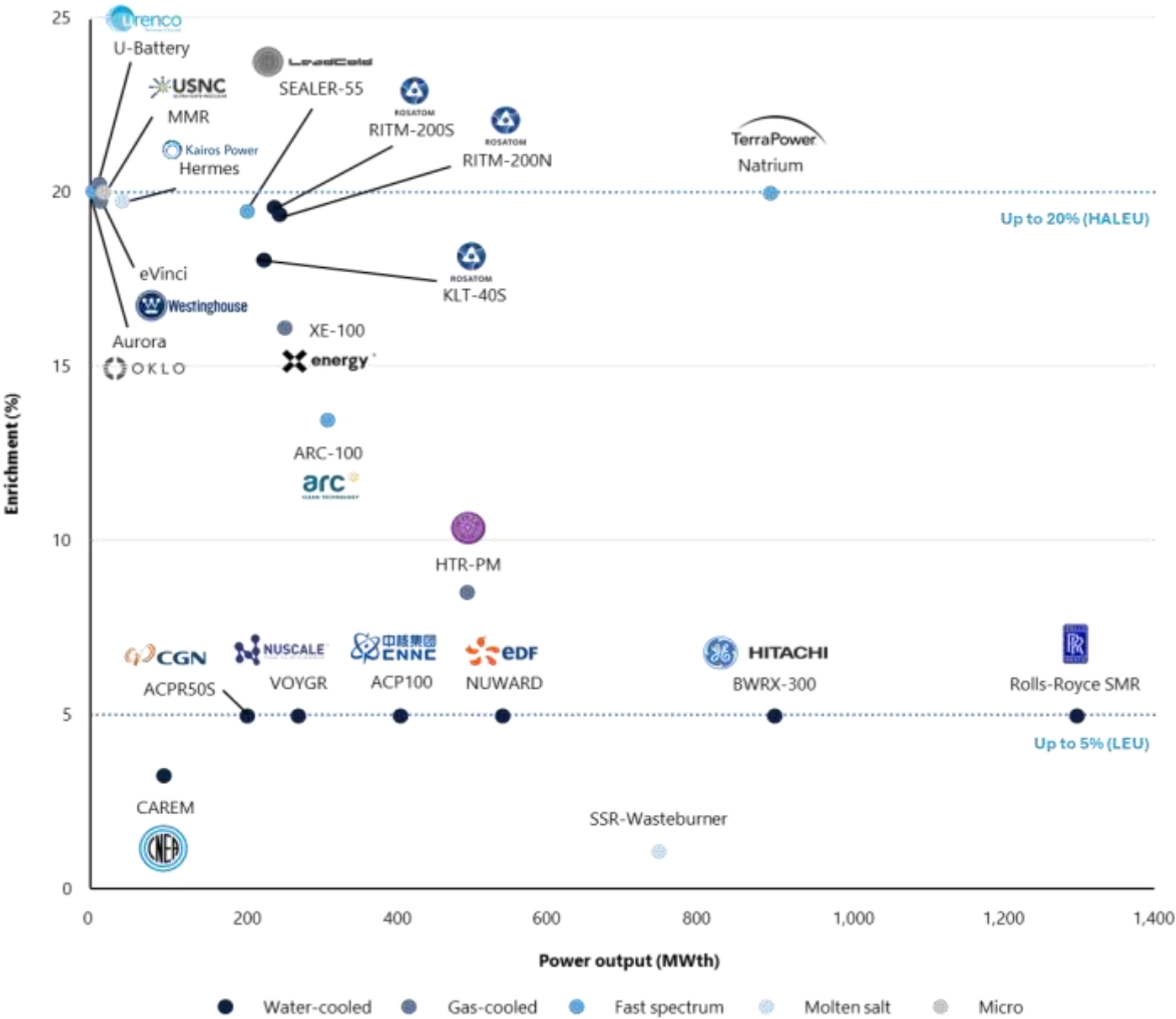


Source(s): IEEFA, Stifel*

The experience with the few SMRs that have been built or are under construction suggests that projects often face budget overruns and longer-than-expected timelines, reflecting broader challenges in the nuclear industry. Additionally, new data from proposed SMR projects in the U.S. indicates a rising trend in cost estimates, raising questions about whether these plants can be built as economically as often claimed.

Nonetheless, according to estimates from Det Norske Veritas (DNV), between 2030 and 2050, around 45% (230GW) of the 500GW of nuclear capacity under construction will be based on SMRs, resulting in approximately 600 SMRs starting construction by 2050. By 2040, regional average costs are likely to remain above the 2030 reference costs, but by 2050, costs should be up to 20% lower in several regions. While SMRs could reduce costs by up to 60% by 2050, their higher initial costs, combined with limited cost reductions in conventional nuclear, will mean that the overall cost of new capacity will only be slightly lower than the cost of large-scale nuclear by 2030.









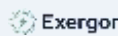
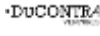







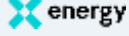

Fig. 62 – Main SMR concepts by type, thermal power and enrichment levels



Footnote(s): HALEU = High-Assay Low-Enriched Uranium ; LEU = Low-Enriched Uranium
Source(s): Nuclear Energy Agency (NEA), Stifel*

Ultimately, SMRs present an interesting solution for reliable, low-carbon energy in data centres, balancing sustainability goals with key challenges. While nuclear fuel is a plentiful source of firm, low-carbon baseload electricity, legacy designs have been criticised regarding their lack of adaptability, long development times and safety concerns. SMR technology, however, is maturing and has piqued the interest of data centre operators due to its safety and reliability claims. While there is enthusiasm for the technology, it must undergo thorough testing and regulatory approval before widespread deployment and the industry is working on creating an effective deployment model for integrating SMRs into data centres. SMRs are competitive with photovoltaic and wind technologies in terms of waste production and can complement renewable energy sources for green power generation. However, the data centre industry must also factor in the higher upfront costs and long development times of nuclear power, balancing this with the need to meet sustainability goals.

Fig. 63 – SMR-related transactions over the past 24-months

Deal Date	Companies	HQ Location	Description	Main investors	Deal Size	Deal Type
Jan-25		FRA	Developer of sodium-cooled fast neutron reactors for flexible electricity production and direct process heat.	    	25	Grant + VC
Jan-25		USA	Operator of Small Modular Reactors (SMRs) technology.		n.a.	VC
Dec-24		USA	Developer of advanced laser technology for uranium enrichment, providing energy-efficient lasers for small modular reactors and civil nuclear fuel production.	   	21	VC
Dec-24		USA	Leading developer of advanced small modular reactor (SMR) and fuel technology for clean energy generation.	   	476	VC
Sep-24		FRA	Developer of clean nuclear technology using particle accelerators to reduce radioactive waste and prevent nuclear accidents.	   	136	VC
Jul-24		USA	Developer of integral molten salt reactor technology to replace fossil fuels with safe, economical clean energy for industrial heat applications.	Undisclosed	27	VC
Jul-24		FRA	Developer of low -carbon small modular reactors providing high-temperature vapor solutions for industrial heat needs.	  	2	Seed
May-24		USA	Developer of small modular nuclear power projects designed to reduce construction time and costs while achieving sustainability goals.	   	37	VC
Apr-24		SWE	Developer of lead-cooled nuclear reactors for safe and reliable power production in various sites.	     	7	VC
Mar-24		CAN	Developer of molten salt reactors that recycle nuclear waste to produce clean electricity and save costs.		2	Grant
Dec-23		USA	Leading developer of advanced small modular reactor (SMR) and fuel technology for clean energy generation.	   	217	VC
Oct-23		USA	Operator of a clean energy technology company developing inherently safe, economical small modular reactors.		7	Grant
Oct-23		USA	Developer of integral molten salt reactor technology to replace fossil fuels with safe, economical clean energy for industrial heat applications.	Undisclosed	3	VC
Sep-23		USA	Developer of modular light water reactor nuclear power plants for various energy applications including district heating and hydrogen production.		n.a.	Secondary Transaction
Aug-23		USA	Developer of advanced laser technology for uranium enrichment, providing energy-efficient lasers for small modular reactors and civil nuclear fuel production.	Undisclosed	1	Seed
Jul-23		SGP	Developer of small modular reactor technology aimed at achieving carbon neutrality by replacing fossil fuel-based energy with nuclear energy.	Undisclosed	n.a.	Seed
Jun-23		SWE	Operator of a Nordic nuclear energy company offering flexible small modular reactors for clean energy production.		2	Seed
Mar-23		CAN	Developer of molten salt reactors that recycle nuclear waste to produce clean electricity and save costs.		27	Corporate
Mar-23		USA	Operator of a clean energy technology company developing inherently safe, economical small modular reactors.	Undisclosed	1	VC
Mar-23		FRA	Developer of sodium-cooled fast neutron reactors for flexible electricity production and direct process heat.	Undisclosed	n.a.	Spin-Off
Jan-23		USA	Developer of the Xe-100, an advanced small modular high-temperature gas-cooled reactor (HTGR) for safe, efficient power generation.	  	23	VC
Jan-23		SWE	Developer of small lead-cooled reactors for reliable and safe nuclear power in Stockholm, Sweden.		n.a.	VC

Stifel Financial Corp.: A Global Overview

We're dedicated to advising growth companies and their investors at every stage of their journey, leveraging our expertise and insights to guide them towards becoming global champions

Group

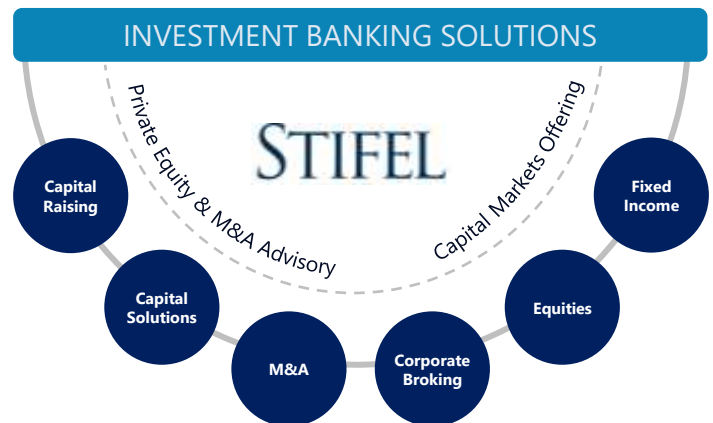
- Founded in 1890; publicly listed since 1983 (NYSE)
- \$11.8bn⁽¹⁾ market capitalisation (NYSE)
- \$4.97bn revenues in 2024 with a 14% revenue CAGR since 2006
- Over 9,000 professionals globally

Institutional

- Full-service investment bank with a global presence
- Leading advisor to middle market companies
- Deep sector and product competencies
- Over 600 investment banking professionals
- Largest Equity Research Platform globally
- Extensive and differentiated distribution capabilities

Global Wealth Management

- Private Client Group:
 - 2,300+ financial advisors managing USD 500bn+ in client assets⁽²⁾
 - Full suite of corporate and individual wealth management solutions
- Banking Services:
 - Bank and Trust with USD 31bn+ in assets⁽²⁾
 - Full suite of deposit and lending products and services



CORE SECTOR EXPERTISE



Diversified Industries



Healthcare



Natural Resources



FIG



Infrastructures



Real Estate



Technology

AUTHORS



**Antoine
Lebourgeois, CFA**

Analyst

antoine.lebourgeois@stifel.com



**Mahaut
Arnaud**

Analyst

mahaut.arnaud@stifel.com



LEGAL DISCLAIMER

This white paper is provided on a confidential basis for informational purposes only and is not intended to, and does not, constitute a recommendation with respect to any potential transaction or investment. Any opinions expressed are solely those of Stifel and applicable only as at the date of this white paper. This white paper is necessarily based upon economic, market, financial and other conditions as they exist on, and on the information made available to Stifel as of, the date of this white paper, and subsequent developments may affect the analyses or information set forth in this white paper. This white paper does not purport to give legal, tax or financial advice. Recipients should not rely on the information contained in this white paper and must make their own independent assessment and such investigations as they deem necessary. Stifel is not soliciting any action based upon this white paper. This white paper does not constitute or form part of any offer or invitation to sell, or issue, or any solicitation to any offer to purchase or subscribe for, any shares, financial instruments, or other securities, nor shall it (or any part of it), or the fact of its distribution, form the basis of, or be relied on in connection with or act as any inducement to enter into, any contract whatsoever relating to any securities, financial instruments or financial services of Stifel or of any other entity or constitute an invitation or inducement to any person to underwrite, subscribe for or otherwise acquire securities. The information in this white paper is not complete and is based upon information that Stifel considers reliable, but it has not been independently verified. Stifel does not represent, guarantee, or warrant, expressly or implicitly, that this white paper or any part of it is valid, accurate or complete (or that any assumptions, data or projections underlying any estimates or projections contained in the white paper are valid, accurate or complete), or suitable for any particular purpose, and it should not be relied upon as such. Stifel accepts no liability or responsibility to any person with respect to or arising directly or indirectly out of the contents of or any omissions from this white paper.

The distribution of this white paper may be restricted by law. Accordingly, this white paper may not be distributed in any jurisdiction except in accordance with the legal requirements applicable to such jurisdiction. Persons into whose possession this document comes are required to inform themselves about and to observe any such restrictions. This white paper is only be addressed to and directed at specific addressees who: (A) if in member states of the European Economic Area (the "EEA"), are persons who are "qualified investors" within the meaning of Article 2(e) of Regulation (EU) 2017/1129 (as amended) (the "Prospectus Regulation") ("Qualified Investors"); (B) if in the United Kingdom, are Qualified Investors within the meaning of Article 2(e) of the Prospectus Regulation as it forms part of domestic law by virtue of the EU (Withdrawal) Act 2018 (as amended from time to time) and who are: (i) persons having professional experience in matters relating to investments who fall within the definition of "investment professionals" in Article 19(5) of the Financial Services and Markets Act 2000 (Financial Promotion) Order 2005 (the "Order"); or (ii) high net worth entities falling within Article 49(2)(a) to (d) of the Order; or (C) are other persons to whom it may otherwise lawfully be communicated (all such persons referred to in (B) and (C) together being "Relevant Persons"). This white paper must not be acted or relied on in (i) the United Kingdom, by persons who are not Relevant Persons; (ii) in any member state of the EEA by persons who are not Qualified Investors; or (iii) in the United States ("U.S.") by persons who are not Qualified Institutional Buyers ("QIBs") as defined in and pursuant to Rule 144A under the U.S. Securities Act of 1933, as amended. Any investment activity to which this white paper relates (i) in the United Kingdom is available only to, and may be engaged in only with, Relevant Persons; (ii) in any member state of the EEA is available only to, and may be engaged in only with, Qualified Investors; and (iii) in the U.S. is available only to, and may be engaged in only with, QIBs. If you have received this white paper and you are (A) in the United Kingdom and are not a Relevant Person; (B) are in any member state of the EEA and are not a Qualified Investor; or (C) are in the U.S. and are not a QIB, you must not act or rely upon or review the white paper and must return it immediately to your Stifel representative (without copying, reproducing or otherwise disclosing it (in whole or in part)).

No person shall be treated as a client of Stifel, or be entitled to the protections afforded to clients of Stifel, solely by virtue of having received this document.

This paper was produced by Bryan, Garnier & Co Limited, prior to the acquisition by Stifel Financial Corp; some contributors may have since left the organisation.

Independence of Research

Stifel prohibits its employees from directly or indirectly offering a favourable research rating or specific price target, or offering to change a rating or price target, as consideration or inducement for the receipt of business or for compensation.

Basis of Presentation

References herein to "Stifel" collectively refer to Stifel, Nicolaus & Company, Incorporated, Stifel Nicolaus Europe Limited ("SNEL"), Stifel Europe AG ("STEA"), Stifel Europe Advisory GmbH, Stifel Nicolaus Canada Incorporated, Bryan Garnier & Co Limited, Bryan Garnier Securities SAS, Bryan Garnier & Co GmbH, Bryan Garnier & Co AS and other affiliated broker-dealer subsidiaries of Stifel Financial Corp. SNEL and STEA also trade as Keefe, Bruyette & Woods ("KBW"). For a list of Stifel affiliates and associated local regulatory authorisations please see here: www.stifel.com/disclosures/emaildisclaimers. References herein to "Stifel Financial" refer to Stifel Financial Corp. (NYSE: SF), the parent holding company of Stifel and such other affiliated broker-dealer subsidiaries. Unless otherwise indicated, information presented herein with respect to the experience of Stifel also includes transactions effected and matters conducted by companies acquired by Stifel (including pending acquisitions publicly announced by Stifel), or by Stifel personnel while at prior employers.

If you no longer wish to receive these marketing communications, please e-mail StifelEurope.GDPR@stifel.com and we will arrange to have you taken off the relevant mailing list(s).

Copyright 2025 Stifel. All rights reserved.

www.StifelEurope.com

STIFEL | IRIS

INTELLIGENCE • RESEARCH • INSIGHTS • SERVICE



London, United Kingdom

Stifel Nicolaus Europe Limited
150 Cheapside
London, EC2V 6ET
Tel: +44 20 7710 7600

Frankfurt, Germany

Stifel Europe AG
Kennedyallee 76
60596 Frankfurt am Main
Tel: +49 69 788080

Paris, France

Bryan Garnier Securities SAS
26 avenue des Champs-
Élysées
75008 Paris
Tel: +33 1 56 68 75 00

Paris, France

Stifel Europe AG – Paris Branch
80 Avenue de la Grande
Armée
75017 Paris
Tel: +33 1 7098 3940

Frankfurt, Germany

Stifel Europe Advisory GmbH
Bockenheimer Landstrasse 24
60323 Frankfurt am Main
Tel: +49 69 247 4140

Munich, Germany

Stifel Europe AG – Munich
Branch
Maffeistrasse 4
80333 Munich
Tel: +49 89 9992 9820
Tel: +49 89 2154 6000

Munich, Germany

Stifel Europe GmbH
Königinstraße 9
80539 Munich
Tel: +49 89 242 262 11

Milan, Italy

Stifel Europe AG – Milan
Branch
Via Privata Maria Teresa, 8
20123 Milan
Tel: +39 02 85465761

Oslo, Norway

Bryan Garnier & Co AS
Haakon VII's Gate 1, 2nd Floor
0161 Oslo
Postbox: 0117 Oslo
Tel: +47 908 45 025

Zurich, Switzerland

Stifel Schweiz AG
Tessinerplatz 7
8002 Zurich
Tel: +41 43 888 6100

Geneva, Switzerland

Stifel Schweiz AG – Geneva
Office
Place de la Fusterie 12
1204 Geneva
Tel: +41 22 994 0610