Data Center in the Generative AI Era

Contents

Executive summary

The emergence of generative artificial intelligence (GenAI) in late 2022 fundamentally transformed the technology industry. The size of the AI software market is expected to reach $218bn by 2029, thanks mainly to GenAI assistants, productivity tools, and GenAI-embedded enhancements to software-as-a-service (SaaS) solutions. The continual rollout of new foundation models has expanded GenAI capabilities beyond text to include sophisticated text-to-image generation, advanced speech recognition and synthesis, video generation, and comprehensive code generation.

Key focus areas include multimodal GenAI, reasoning models, and small foundation models designed for edge use cases. Meanwhile, the emergence of cost-efficient models—such as DeepSeek-V3, a 671 billion-parameter model trained at just under $6m—demonstrates that high-performance models can be developed without access to high-end GPUs. Moving forward, agentic AI is expected to be the main driving force behind enterprise GenAI usage. AI agents combine GenAI models with expert systems and machine learning to automate complex tasks with minimal supervision.

The increase in AI demand drives explosive growth in AI data centers. Global AI server shipments are projected to exceed 4 million units by 2029 with a 2024–29 cumulative average growth rate of approximately 14%. Capital expenditure on AI servers doubled from $22.9bn in 2022 to $56bn in 2023 and is projected to reach $138bn in 2024. Global cloud service providers are leading this investment surge. As a key player in the data center industry, GTI and Omdia believe telcos should seize the opportunity to become a key player in the AI data center market. Specifically, telcos must leverage their understanding of the network and compute demands of their local market and deploy specific technologies to capture this growth. Key GTI members such as China Mobile and Softbank are active players in the AI data center space with unique GenAI foundation model and service offerings. In order to accommodate the significant GenAI training and inference workloads, the industry has developed various data center solutions:

- For ultra-large-scale AI training and inference workloads, global cloud service providers and internet companies rely on hyperscale AI data centers, mega facilities with up to hundreds of thousands of AI training and inference chipsets designed for large-scale pretraining tasks.
- Banks, financial institutions, telcos, industrial conglomerates, and other large enterprises use large AI data centers optimized for training and inference in addition to support for general computing workloads.
- For business-critical functions, enterprises have deployed small AI data centers designed for environments with limited computing and energy resources or stringent data privacy requirements.

A diverse range of technological advances supports these data centers. Key technology drivers include exponential growth in AI chipsets, accelerated networking for model development and deployment, and service assurance through innovative power supply and thermal management solutions. Native liquid cooling, two-phase direct-to-chip cooling (DTC), and high-voltage direct current (HVDC) solutions are being implemented to address the increasing demands for power density and efficiency.

Nevertheless, adoption of AI data centers remains uneven around the world. Most activities have been focused on the US and China, and South-Eastern Asia and the Middle East are becoming the next frontiers. The US leads the data center server market at $140bn in 2024 and is dominated by hyperscalers and emerging neocloud vendors. China has established a strong domestic market for computing, storage, and networking data center technologies with an estimated server market size of more than $58bn in 2024. South-Eastern Asia's data center server market is valued at approximately $5.8bn, Singapore being the market leader. Meanwhile, the Middle East market is worth around $3.1bn and is led by the UAE and Saudi Arabia with strategic AI initiatives.

As AI data centers proliferate across all countries and territories, Omdia and GTI are in the view that telcos have are expected to play a key role. Any telco entering the AI data center space should adopt a strategic approach focused on four key principles:

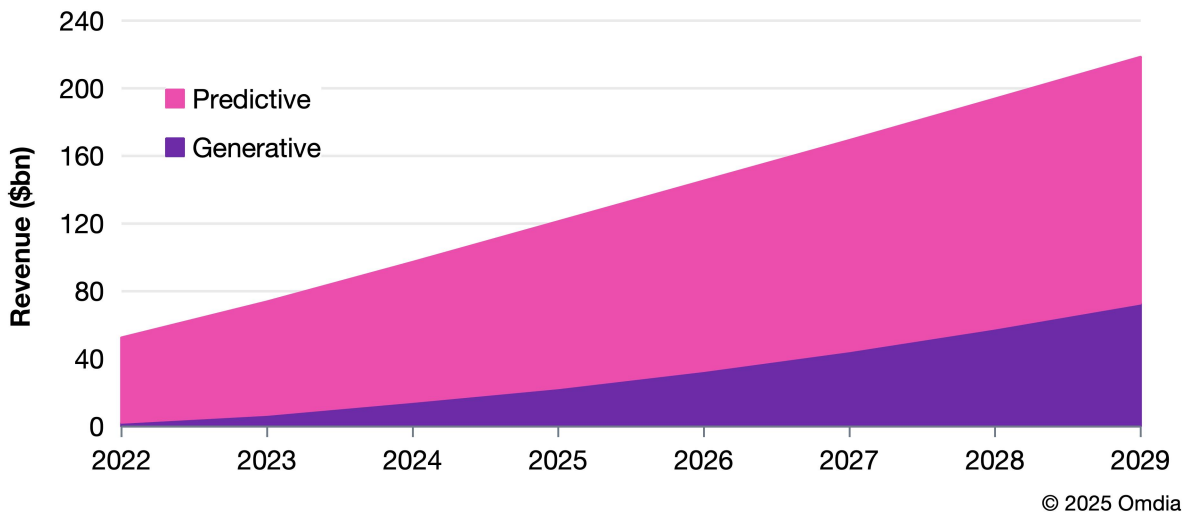
- Work backward from value creation by thoughtfully deploying GenAI infrastructure that delivers tangible client benefits rather than follow trends.
- Leverage existing fiber networks and connectivity infrastructure to provide edge computing capabilities and comprehensive service level agreements for latency-sensitive applications.
- Differentiate through strategic partnerships with AI hardware and software providers while developing value-added managed AI services, enhanced security measures, and expert consulting services.
- Embrace openness and collaboration by pursuing standardized solutions.

Telcos have a long-standing reputation for embracing and promoting open standards and open source technologies. In the GenAI era, this value will help to improve the overall GenAI system performance, level the competitive landscape, foster innovation, and accelerate GenAI adoption across the ecosystem.

Key GenAI trends

The emergence of GenAI in late 2022 has brought fundamental changes to the technology industry. Far from the early days of AI, when AI was winning chess competitions and computer games against human players, in the GenAI era AI is generating human-like responses with high fidelity. Conversational AI applications, such as ChatGPT and DeepSeek, are among the most popular and widely used consumer applications in human history. GenAI capabilities are integrated across many existing applications in social media, e-commerce, and productivity tools. It is not limited to consumer use cases: organizations worldwide have made use of GenAI for content creation, customer service, knowledge management, and sophisticated code generation and debugging processes. Across some specific industries, GenAI has made significant inroads, automating tasks and augmenting the capabilities of human employees.

Figure 1: Global predictive and generative AI software revenue, 2022–29



Source: Omdia

Continual rollout of new foundation models

Pretrained using publicly available data, GenAI models can understand human instructions and generate responses in text and speech. GenAI has expanded beyond text to encompass various forms of media, including sophisticated text-to-image generation, advanced speech recognition and synthesis, video generation and editing capabilities, and comprehensive code generation and analysis. As their parameter counts increase, these models can mimic human actions, demonstrate contextual understanding, perform and

execute tasks at a near-human level, and exhibit human-like behavior. To enhance the model readiness for the enterprise environment, GenAI model developers have introduced new techniques—including an efficient attention mechanism, a refined fine-tuning approach, inference acceleration, and retrieval augmented generation—to reduce training costs, improve inference speed, and minimize hallucination.

With the increasing adoption of GenAI, model developers are releasing new models with enhanced capabilities. The GenAI models launched in 2025 indicate the focus is on the following areas:

- Multimodal GenAI
  - These models are capable of understanding and generating various types of data using a collection of expert models. Key examples include o3 and o4-mini from OpenAI, Gemini-2.5 from Google, LlaMa 4 from Meta, and DeepSeek-V3-0324 from DeepSeek.
- Reasoning model
  - These models can perform in-depth analysis on the specific inquiry by drawing from multiple information sources and further improve the outputs through self-reflection and review. Key examples include DeepSeek-R1 from DeepSeek, QwQ-32B from Alibaba, X1 from Baidu, and T1 from Tencent.
- Small foundation model
  - These models have low parameter counts, usually between 0.5 billion and 7 billion, designed for edge use cases. Key examples include Qwen2.5 0.5B from Alibaba, Phi-4 from Microsoft, and Gemma-3 from Google.

At the same time, vendors continue to develop models with enhanced existing capabilities. Cohere X1, for example, can support 23 languages. Qwen2-72B demonstrates a strong command of regional languages, including Portuguese, Japanese, and Arabic. StepFun, a Chinese GenAI startup, launches foundation models focusing on enhanced image and video editing and generation.

Regardless of their capabilities, the training and inference workloads of these models are resource intensive. Test-time scaling, a computing approach that led to the emergence of the reasoning model, is already projected to be the next growth driver for AI computing resources. This approach enables models to think and refine their output during the testing or inference phase. The models can generate several responses and explore different reasoning paths, which requires more computing resources.

Agentic AI driving enterprise GenAI usage

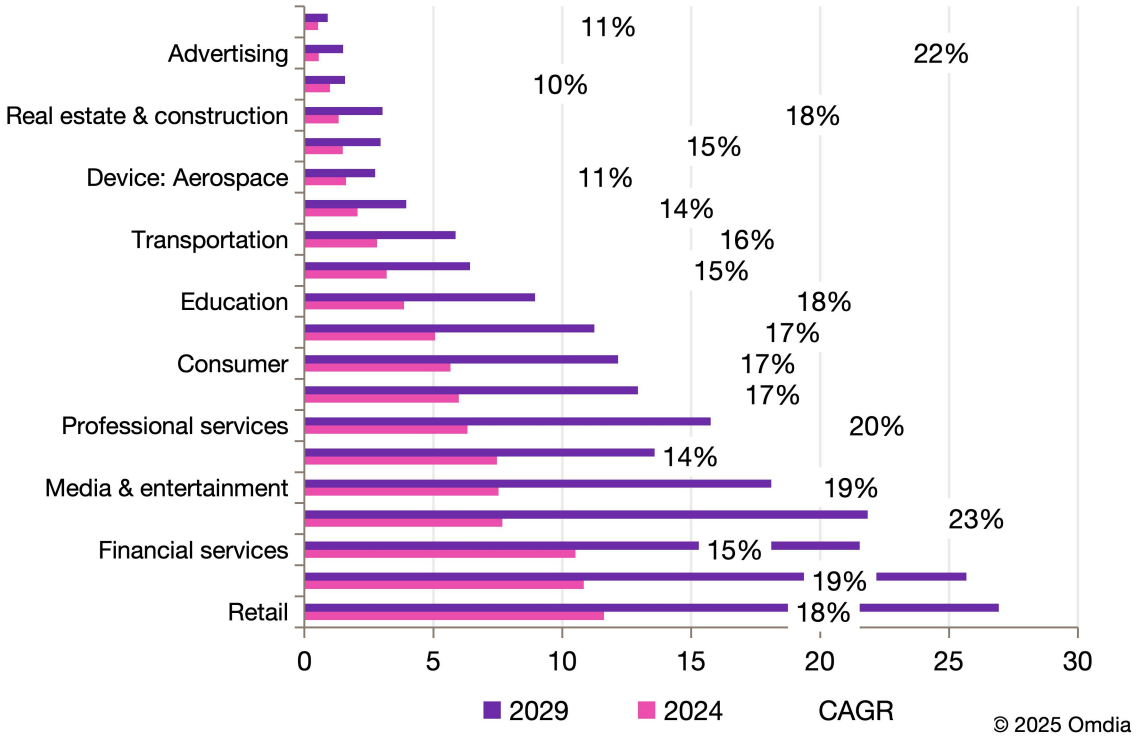
The business world has witnessed widespread adoption of GenAI, which has transformed various processes through the automation of routine tasks and the enhancement of

multimodal understanding. Aside from integrating GenAI models into daily workflow, enterprises are utilizing AI agents to augment the productivity of their workers. By combining GenAI models with expert systems, machine learning, and deep learning models, AI agents will be able to enhance all the use cases currently supported by AI assistants, thanks to the ability of agentic AI to automate even complex tasks while handling those processes with little or no supervision.

Specifically, the agent understands the goals set by humans, breaks them down into specific tasks, and executes those tasks with no human intervention. Furthermore, AI agents can interact with and utilize enterprise datasets and applications through agentic AI frameworks such as AutoGen, MCP, and A2A. Several AI agents can be deployed in a multi-agent framework through robust orchestration to further enhance the capabilities of AI agents. Orchestration can abstract complexity and ensure agents collaborate effectively and in alignment with the overall goals that have been set.

Omdia expects the AI software market to reach $218bn in 2029. Growth will be driven primarily by the burgeoning supply of and demand for GenAI assistants, productivity tools, and GenAI-embedded enhancements to widely used SaaS solutions. According to Omdia's AI maturity survey, 37% of companies with revenue exceeding $1bn were scaling AI deployment across more than one business function or unit.

Figure 2: Global AI software revenue by end-user vertical, 2024 & 2029



Source: Omdia

Retail takes the top spot with $11.6bn in revenue in 2024 and an 18% CAGR. Data from Omdia's AI Enterprise Contract Tracker shows AI contracts for retailers jumped in 2H23 and 1H24, and retail vied with financial services as the industry with the most contracts. Large e-commerce players drive this vertical as they increasingly personalize product targeting. The top retail AI use cases are product recommendations, virtual assistants, and image and video content analytics.

Healthcare is the second-largest vertical with $10.8bn in revenue in 2024 and a 19% CAGR. The top use cases are medical image analysis, computational drug discovery, and virtual and voice assistants. The industry is also leveraging AI and other technologies within AI for clinical trials, drug development, bioinformatic analysis, and medical diagnosis.

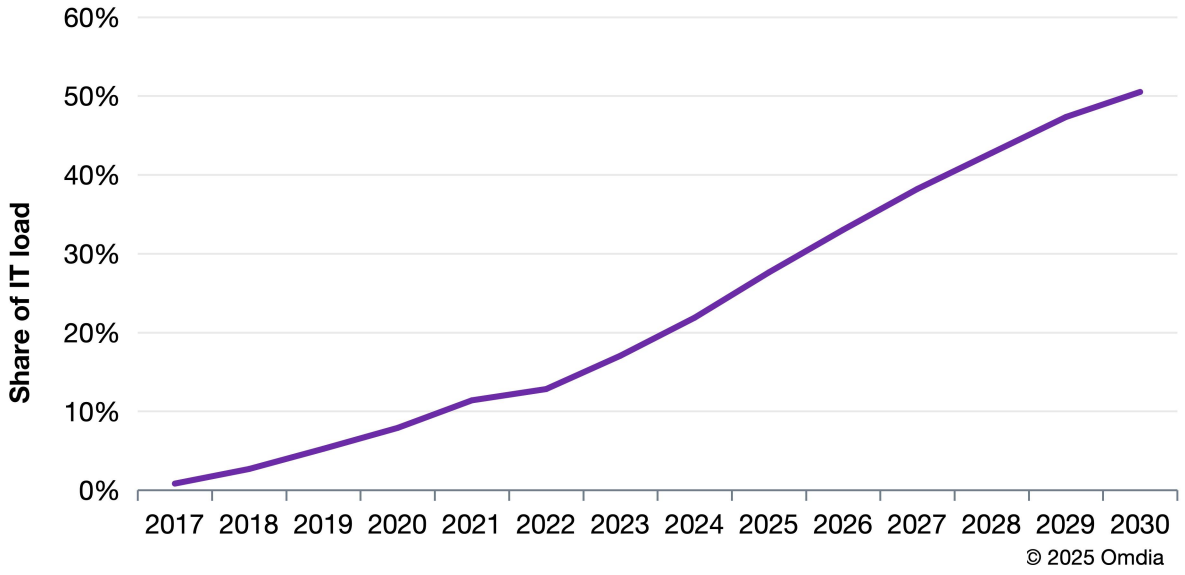
Financial services is closely tied with healthcare in 2024 at $10.5bn but, as a mature AI adopter, will have a slower CAGR of 15% to 2029. Key use cases include virtual assistants, fraud detection, intelligent document processing, algorithmic trading strategy analytics, risk assessment, and compliance.

GenAI is also making inroads into other verticals. In education, it enables personalized learning experiences; in legal sectors, it facilitates document analysis and contract review; and in manufacturing, it optimizes processes and quality control.

The proliferation of AI data centers

The explosive growth of GenAI has driven high market demand for data centers, particularly those optimized for GenAI training and inference workloads. As more and more enterprise solutions rely on GenAI capability, modern AI data centers have become the backbone of technological advance.

Figure 3: AI share of global IT workload, 2017–30

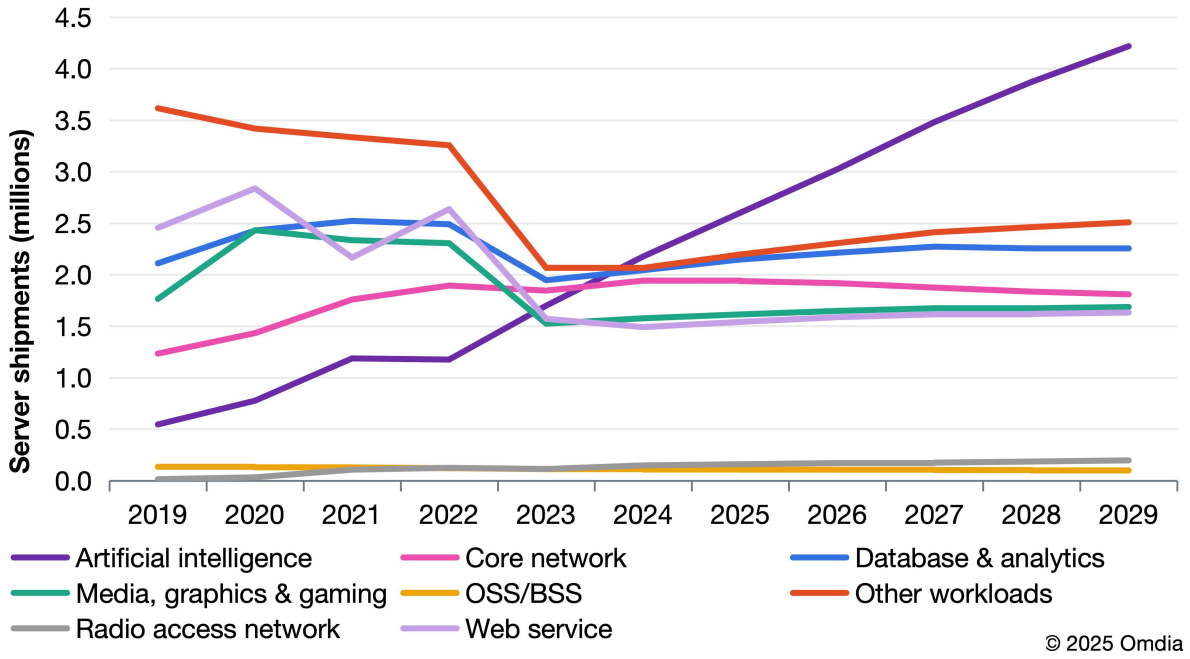


Source: Omdia

Most importantly, GenAI has completely changed the configuration of servers needed to run AI. First, the demand for computing has grown significantly. At the same time, the growing model demands higher memory capacity and networking capability, often spread across several AI processors in a single server or across several servers. The exorbitant price of multi-GPU servers capable of running large language models will consume a significant portion of the annual capex of cloud service providers and enterprises. Cloud services and on-premises infrastructure-as-a-service (IaaS) services will be the most efficient way for enterprise customers to consume compute capacity when the cost of capital remains high and macroeconomic uncertainties persist. **Figure 4** shows Omdia's

forecast for AI server shipments, which is the best indicator of the growth of the AI data center.

Figure 4: Global server shipments by application category, 2019–29

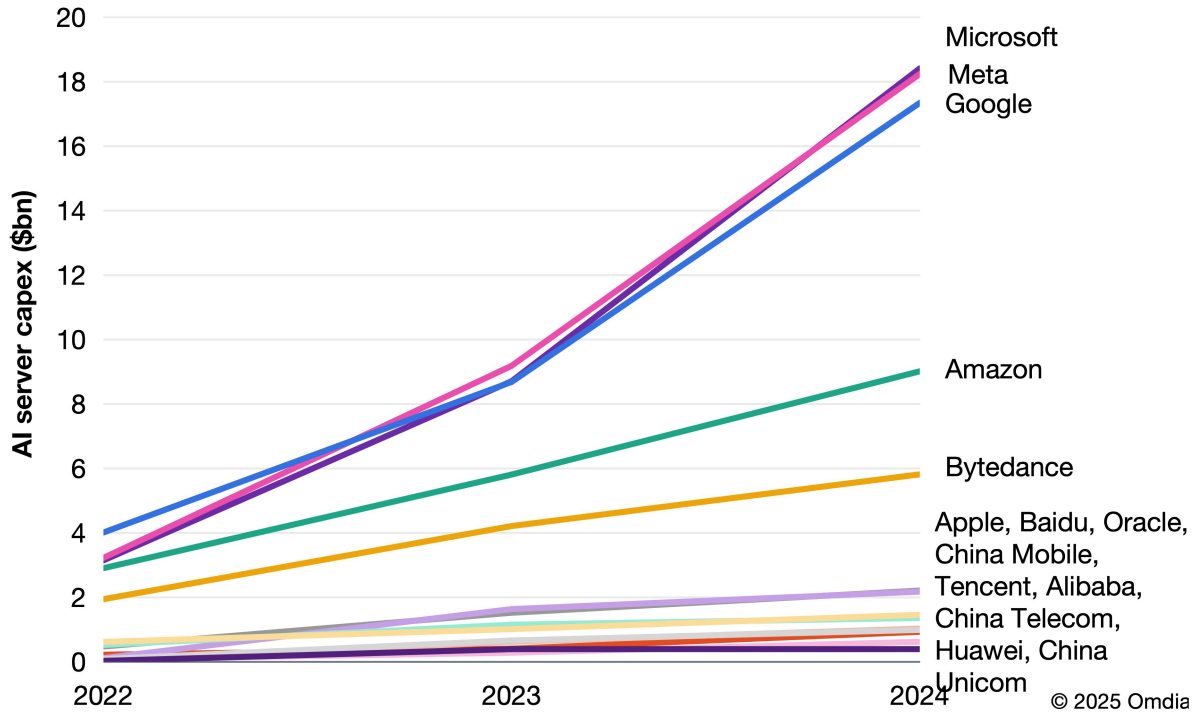


Source: Omdia

Omdia expects global AI server shipments to exceed 4 million units by 2029, driven by the acceleration initiated in 2023 by the emergence of GenAI. The CAGR is estimated to be around 14% between 2024 and 2029.

Another good indicator is the level of capex. Since 2022, capital expenditure on AI servers has doubled annually, increasing from $22.9bn in 2022 to $56bn in 2023 and $138bn in 2024. Hyperscalers in the US have been the top spenders. In comparison, Chinese cloud service providers are lagging in capex; however, their spending is also trending upward because of increased cloud usage, particularly from 2025 onward.

Figure 5: Global server capex by top cloud service providers, 2022–24



Source: Omdia

Categories of modern AI data centers

Since modern AI data centers are designed to support GenAI workloads, their designs and configurations are influenced by the type of workloads they support. In the view of GIT and Omdia, the diversity of GenAI workloads has led to the emergence of three major categories of AI data centers.

Hyperscale AI data center

A hyperscale AI data center is a mega data center with tens of thousands or hundreds of thousands of AI training and inference chipsets. Though a hyperscale data center excels at both tasks, this type of data center is typically designed for large-scale pretraining tasks that require extremely high operating conditions and complex maintenance. Training is the most cost-consuming part of the GenAI model development process. More efficient training results in shorter training times, improved market competitiveness, energy savings, and cost reduction. At the same time, the rise of reasoning models means more computing is required to support test-time scaling. As the use of reasoning models increases, more efficient computing is necessary for faster inference speeds. Currently, the industry primarily employs ultra-high-performance AI computing clusters that comprise 10,000 or even 100,000 AI training and inference chipsets. Such a cluster is equipped with a scale-

out networking architecture to increase the computing capability. These data centers have the best cooling technologies to ensure energy efficiency.

Large AI data center

Large AI data centers are optimized for training and inference in addition to supporting general computing tasks and ensuring high energy efficiency. Hyperscale AI data centers are designed for global and national workloads, but large AI data centers are typically designed to support regional workloads. Like hyperscale AI data centers, they must meet stringent energy-saving and cost-efficiency requirements while safeguarding user experience, application security, and reliability. The hybrid deployment of multiple models is especially critical in large AI data centers, enabling GenAI applications to be more flexible and scalable.

Small AI data center

Small AI data centers are designed for an operating environment with limited computing and energy resources or stringent data privacy and security requirements. Generally, a small AI data center can be deployed quickly and offers an easy-to-use toolchain that supports remote operations and maintenance when necessary. Because of their operating environments, adequate security measures must be taken to ensure the entire system is secure and highly resilient when operating in a distributed environment.

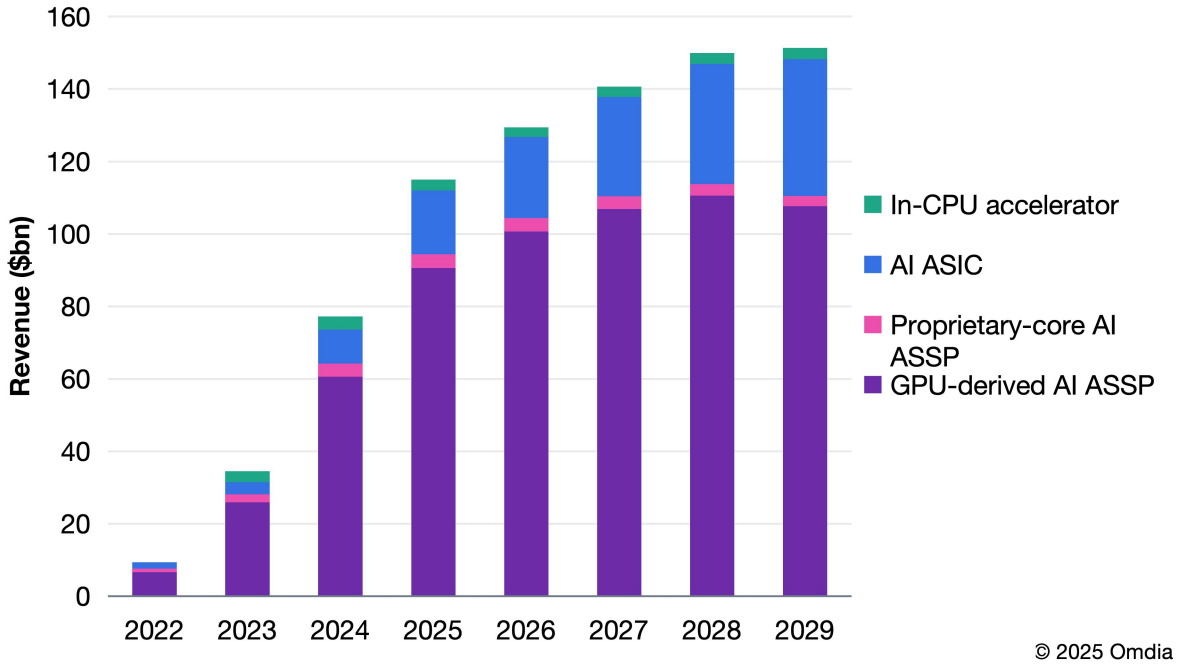
Key technology drivers

Though GenAI is the single most significant market driver, the technology drivers for AI data centers can be divided into three major areas: computing, networking, and infrastructure. Each area has a significant impact on the industry's future direction, and in combination they enable more efficient GenAI training and inference.

Exponential growth in AI chipsets

The surge of 2022–24 has reshaped the industry. NVIDIA's data center business overtook Intel's DC&AI and grew to 10× its size. However, 2024 was the year when competitors took share from NVIDIA.

Figure 6: Global, AI processors for data centers, forecast by architecture, 2022–29



Source: Omdia

On one front, AMD's MI-series GPUs have begun to sell in scale, specifically to the hyperscale and high-performance computing (HPC) markets. AMD is still behind NVIDIA in software, but a significant effort has brought its ROCm SDK closer to a good-enough status, and the GPUs are now deployed with all the major hyperscalers.

AMD's design approach is based on chipsets and TSMC's 3D packaging technology. This has enabled it to set out a roadmap of incremental improvements on the GPUs. Infinity Fabric has also begun to address the backend network, aligning with a broader industry trend toward Ethernet in the backend; however, this also requires significant additional work.

At the same time, the landscape of AI data centers has been revolutionized by the increasing deployment of specialized hardware. The quest for efficiency and sustainability will lead to further optimizations in server hardware, resulting in numerous custom silicon deployments. Google's deployment of its custom Cloud Tensor Processing Units (TPUs) is accelerating rapidly, and Google Cloud Platform has achieved profitability. Aside from Google, ByteDance, Tencent, and others are developing custom CPUs based on the Arm architecture, following the footsteps of other hyperscalers such as Amazon, Alibaba, and Microsoft.

Finally, computing capacity is a national priority for many countries and regions because of its growing importance globally and the impact of geopolitical situations. There will be a push from governments to have domestic AI computing technology, particularly in areas

such as cloud computing and AI. Hyperscale cloud service providers are well positioned to lead these efforts, which will serve as a tailwind for custom silicon design efforts.

Countries developing their semiconductor industries aim to create more resilient supply chains that are less susceptible to international disruptions such as trade disputes. China already has a strong domestic market (Xin Chuang) for computing, storage, and networking data center technologies, driven by local vendors, government agencies, and enterprise customers. Chinese server vendors ship servers with domestic CPUs. These CPUs employ diverse architectures, including x86, Arm, MIPS, and RISC-V. However, RISC-V in particular is gaining significant popularity among government-funded projects and large cloud service providers thanks to its open instruction set architecture.

Accelerated networking for model development and deployment

To further improve the AI workload performance, data center operators adopt two major strategies: scale up and scale out. Scaling up involves increasing the computing, networking, and storage within a single server, and scaling out involves adding more servers to distribute the workload. Each strategy demands a different type of networking approach.

Because scale up increases the resources in a single server, data movement between various computing units, including CPU, GPU, AI ASIC, and storage, must accelerate. Chipset vendors traditionally rely on a network interface card (NIC) to distribute the workload. The emergence of SmartNIC offers additional processing capabilities and programmability to offload network-related tasks from the host CPU.

In recent years, Intel, NVIDIA, and cloud service providers have focused on data processing units (DPUs) to improve the overall efficiency of GenAI model training and inference operations. DPUs excel at optimizing data movement between storage systems and AI ASICs while effectively reducing latency and enabling seamless scaling of AI workloads across multiple nodes. Additionally, DPUs can help reduce security concerns by providing robust hardware-level security for AI model protection. They ensure secure data movement during training and inference operations and enable the isolation of AI workloads from other applications.

In scale-out scenarios, GenAI workloads demand massive data transfers between compute nodes, where traditional TCP/IP networking often creates significant bottlenecks for AI training. Data centers are responding to these demands by making substantial infrastructure adaptations, including upgrades to higher-bandwidth networks supporting 400GbE and beyond, implementing advanced congestion control mechanisms, enhancing network fabric designs for AI clusters, adopting new protocols such as remote direct memory access (RDMA), and improving load-balancing and traffic management systems.

Specific RDMA enhancements have focused on three key areas. In latency optimization, the technology has achieved sub-microsecond latency, reduced CPU overhead, and enabled direct GPU-to-GPU communication. Scalability improvements have delivered better support for large-scale AI clusters.

Service assurance through power supply and thermal management

The rapid expansion of AI data centers has brought sustainability concerns to industry discussions. Organizations are now actively seeking innovative solutions to balance the growing computational demands of AI with environmental responsibility. As data centers face increasing power density and efficiency demands, they are not efficient enough to continue with traditional server design guidelines. The industry seeks a new approach to address power supply and thermal management.

One of the leading innovations is the design of the liquid-cooling system. Native liquid cooling integrates liquid-cooling technology directly into server and data center hardware designs rather than retrofitting existing air-cooled designs. This liquid-cooling approach optimizes in-server thermal flow, component layout, IT performance, and compact space. Consequently, this trend will accelerate the adoption of liquid cooling and mean a permanent update in the guidelines for server design. For example, NVIDIA's Blackwell server directs cooling liquids straight to heat sources, achieving higher performance in a smaller footprint while significantly reducing power usage. The modular design simplifies maintenance and upgrades, enhancing system reliability and sustainability by minimizing noise and waste heat.

In addition, native liquid cooling has inspired new design directions. Full-server liquid cooling features a cold-plate cover that protects components such as the power supply and memory in addition to the main chips. The heat from these components is conducted via parallel cooling loops, eliminating the need for additional air cooling. Ideally, all server-generated heat can be dissipated entirely via the coolant through pipelines. This solution improves thermal management efficiency, streamlines system architecture, and cuts operational costs.

The increasing demand for power density from AI and high-performance computing drives the rapid growth of two-phase DTC. Traditional cooling methods struggle to meet these needs, but two-phase DTC supports a lower power usage effectiveness (PUE) score and reduced energy and carbon emissions. This system absorbs significant latent heat during evaporation within the cold plate and condenses back to liquid in the condenser, forming a closed-loop system. It utilizes regular, mature refrigerants and ensures stable operating temperatures, making it ideal for higher-density computing environments. Supported by chipmakers such as NVIDIA, AMD, and Intel as well as server companies such as Dell, two-phase cooling is expected to boom and become essential for ultra-high-density computing.

Traditional alternating current (AC) power systems suffer from energy losses during transmission and require complex rectification and inversion equipment, resulting in additional inefficiencies. Some cloud providers are experimenting with HVDC solutions in their data centers to address these issues. HVDC can boost energy efficiency, simplify infrastructure, and cut operational costs. When paired with onsite energy storage and photovoltaic microgrids, HVDC architectures enhance overall system efficiency.

At the rack level, OCP-driven direct current (DC) rack solutions centralize power distribution through power shelves, delivering 48V DC directly to servers and IT equipment. This is especially beneficial for high-density computing environments with power requirements of approximately 100kW, where 48V DC outperforms the older 12V DC standard. As densities increase to several hundred kilowatts or even 1MW, voltage levels may need to rise to 800V

DC. This technology, already mature in electric vehicles, could soon be adopted in data center racks such as NVIDIA's NVL72.

Cost-efficient foundation models to further drive GenAI adoption

Launched in December 2024, DeepSeek-V3 is a 671 billion-parameter model that achieves comparable and, in some cases, stronger performance than rival US open source and closed source models. It utilizes 2,048 NVIDIA H800 GPUs and required 2.788 million H800 GPU hours for its training, which the company says cost just shy of $6m. DeepSeek's achievements suggest that high-performance models can be developed and trained without access to high-end GPUs, at a significantly lower cost than leading foundation models that currently dominate the market.

Since the emergence of DeepSeek, increasingly resource-efficient foundation models have been developed. Though it may be logical to assume that the emergence of these models will reduce the reliance on advanced GenAI infrastructure, the direct opposite has been true. This is known as the Jevons paradox: increased efficiency in resource use (in this case, the usage of AI compute resources) can paradoxically lead to an increase in overall consumption. The emergence of DeepSeek and other similar foundation models has encouraged more enterprises to adopt and deploy GenAI-based applications.

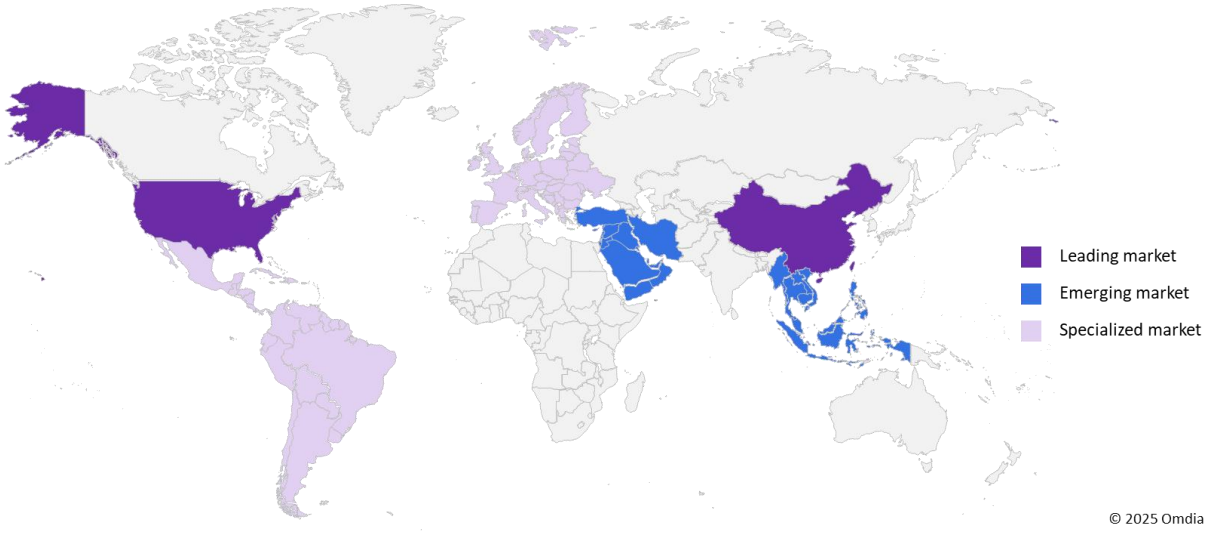
A critical reflection is that innovation is not dependent on brute force, large-scale models or easy access to powerful and expensive compute infrastructure. Instead, it depends on how these resources are leveraged and to what effect. DeepSeek has shown that agility, creativity, and the ability to think outside the box can drive innovation and represent a valuable source of competitive advantage.

Country-specific AI data center trends

Regional characteristics in AI data center deployment

The global race to build AI data centers has gained unprecedented momentum. North America, particularly the US, stands out as a leading hub because of its robust technological infrastructure, substantial capital investment, and concentration of major tech giants such as Google, Microsoft, and Amazon. China follows closely behind thanks to its aggressive push for AI sovereignty, which is supported by state policies and massive investments from companies including Alibaba and Tencent. Outside the US and China, Omdia and GTI observe that emerging markets such as South-Eastern Asia and the Middle East are quickly becoming key players in the global AI data center landscape thanks to their strategic geographic positioning, abundant power resources and capital, and supportive government policies. Latin America and Europe are also carving out their niches in the AI data center boom. Their momentum is bolstered by a strong emphasis on data privacy and compliance, which makes them attractive destinations for companies seeking AI infrastructure. Together, these regions represent the forefront of AI data center development, leveraging unique strengths to drive global AI innovation.

Figure 7: AI data center markets



Source: Omdia and GTI

The US

Omdia estimates the US data center server market was valued at $140bn in 2024, with over 172 billion square feet of space. As discussed earlier, the US AI data center market is dominated by hyperscalers. In 2024, there were unprecedented increases in capex from AWS, Google, Meta, and Microsoft with an overall increase of 94%. These facilities are primarily concentrated in northern Virginia, northern California, and Phoenix, Arizona, serving hyperscale cloud providers and social media companies requiring extensive computing capabilities.

Other key players in the US market are the neocloud vendors. These vendors are AI-specialized startup cloud service providers that focus on providing GPU as a service. They are growing quickly thanks to a continuous and increasing influx of funding. This is an ecosystem that has already matured. Collectively, market leaders such as CoreWeave, Lambda Labs, Crusoe, and Groq have raised over $1bn. CoreWeave is currently publicly traded and is aggressively planning global expansion. At the same time, many smaller neocloud vendors are struggling. Paperspace was acquired by DigitalOcean, a Tier 2 cloud service provider; others, including Nyriad, are looking for buyers.

In the long run, pure-play GenAI model developers and service providers are expected to be key disrupters. For example, OpenAI is planning a new initiative, Project Stargate, under which major tech companies led by Softbank would invest $500bn in AI in the US. Anthropic and Inflection aim to choose to build their server clusters in the long term.

Moving forward, the US AI data center market is expected to continue evolving rapidly with significant investments in infrastructure development, particularly in secondary and tertiary markets, as organizations adapt to increasing data demands and technological advances.

China

China's data center infrastructure is extensive and continuously expanding. In July 2017, the Chinese government issued the "Development Plan for a New Generation of Artificial Intelligence," which explicitly mentioned the need to cultivate a high-end and efficient innovative economy and called for accelerating the in-depth integration of AI with the economy and society and focusing on enhancing the scientific and technological innovation capacity of a new generation of AI.

Several influential players across different categories dominate the Chinese data center market. In the domestic technology sector, giants such as Alibaba Group (through Alibaba Cloud), Tencent Holdings, Baidu, Huawei, and Inspur lead the market with innovative solutions and extensive infrastructure. The three major telcos, China Mobile, China Telecom, and China Unicom, play crucial roles in data center operations and connectivity. Additionally, specialized data center operators such as GDS Holdings Ltd., VNET Group Inc., Chindata Group, GLP China, Space DC Pte Ltd., and Shanghai Athub have established strong market positions.

Overall, Omdia estimates the size of the Chinese data center server market to be over $58bn in 2024, with a total of over 77 billion square feet. According to the latest version of the Ministry of Industry and Information Technology's National Data Center Application Development Guidelines, the four first-tier cities—Beijing, Shanghai, Guangzhou, and Shenzhen—have the highest concentration of data centers. In recent years, Chinese hyperscalers have gradually begun to deploy data centers in areas surrounding the first-tier cities and the central and western regions. Guizhou is a typical case. The province has made rapid progress in recent years with its climate, ecology, power supply, resource endowments, and policy advantages.

South-Eastern Asia

In 2024, the South-Eastern Asia data center server market is estimated to be worth around $5.8bn with a total deployed space of 18 million square feet. Singapore's robust infrastructure, market-friendly policies, and power supply security make it the market leader, and it continues to be the preferred destination for regional data center infrastructure. The country launched its first National AI Strategy in 2019 and its second in 2024 with a strong focus on AI governance and GenAI.

Over the past year, Johor Bahru in Malaysia has emerged as South-Eastern Asia's fastest-growing data center market. In September 2024, Malaysia unveiled its National Guidelines on AI Governance and Ethics (AIGE). The guidelines aim to help all AI developers and practitioners in Malaysia deploy safe and ethical AI based on seven principles. This announcement came at the end of the country's National AI Roadmap. At the same time, the government has been actively strengthening its digital governance framework, having gazetted the Cyber Security Act 2024 in June 2024 and reviewed its Personal Data Protection Act 2010.

The market is particularly driven by the demand from hyperscale companies, because South-Eastern Asia is one of the regions where Western and Chinese hyperscalers are actively rolling out in-country data centers. Local telcos such as Singtel, Telkomsel, Indosat, Telekom Malaysia, and Viettel are also actively acquiring NVIDIA's GPUs and rolling out in-

country AI data centers. The growth is further supported by favorable regional government initiatives, including increased land availability for development, reduced electricity tariffs, tax incentives, and support for renewable energy procurement.

Middle East

Omdia estimates the Middle East data center server market to be around $3.1bn in size with a total deployed space of 2.5 million square feet. Key trends shaping the Middle East data center landscape include a strong emphasis on sustainability with increased use of renewable energy sources, the deployment of 5G networks and edge facilities, improvements in inland connectivity through submarine cables, and the integration of AI.

Countries such as the UAE and Saudi Arabia are leading this transformation through strategic programs including the UAE's National AI Strategy 2031 and Saudi Arabia's National Strategy for Data & AI. The most prominent project is the Stargate UAE, the world's largest data center outside the US. The project occupies a 26-square-kilometer site, featuring tens of thousands of GPUs with an eventual power consumption of 5GW. Organizations from the UAE, such as the Technology Innovation Institute (TII), G42, and e&, focus on the GenAI model and platform development. The market is also witnessing significant developments in digital infrastructure, exemplified by the recent memorandum of understanding between the UAE's Ministry of Investment and Egypt's Ministry of Communications, which aims to develop 1GW of network infrastructure center capacity.

Recommendations for telcos

GTI and Omdia believe telcos need to be proactive in the era of GenAI and AI data center. They must take the following actions to ensure they further expand their influence beyond traditional connectivity:

- Work backward from value creation.
  - When deploying AI data centers, telcos must be more forward-looking and should not mindlessly follow the current trends. They should think deeply about how AI technology can create value for their clients and deploy advanced GenAI infrastructure that enables enterprises to capture the expected revenue.
- Offer a combination of computing and connectivity.
  - Telcos entering the AI data center industry should begin by leveraging their existing extensive fiber networks and connectivity infrastructure for latency-sensitive and geographically distributed applications.
  - Telcos can also offer edge computing capabilities that further enhance the user experience.
  - Additionally, telcos can offer comprehensive end-to-end service-level agreements.
- Differentiate through partnership and managed services.
  - Telcos should forge strong partnerships with AI hardware and software providers and establish clear differentiation from traditional data center providers by developing value-added managed AI services, implementing advanced encryption and data protection measures, and offering expert consulting services for the implementation of GenAI.
- Embrace openness and collaboration.
  - Telcos should collaborate with other telcos and ecosystem players to seek open and standardized solutions that enhance the performance and efficiency of the overall AI data center system. Solutions based on open source and open standards level the playing field, allowing vendors to innovate on top of these solutions and encouraging GenAI adoption.
  - A good example would be GTI. Formed by China Mobile, SoftBank, Vodafone and other operators, GTI can help telcos to explore new opportunities in data center through industry collaboration and dialogues.

Commissioned