Accelerating enterprise AI:

Hardware advancements and compute architecture transformation

Agentic AI and multimodal technologies are driving rising enterprise AI demand, with inference growth expected to outpace training. As workloads extend from cloud to edge, AI compute architectures are undergoing restructuring.

DIGITIMES observes that six key applications—chatbots, software development, image and video generation, enterprise operations automation, and process automation—are gaining traction across enterprises. The resulting growth in inference demand, alongside increasingly diverse cloud and edge deployments, is driving a reconfiguration of AI infrastructure specifications. This report provides supply chain guidance on evolving specifications and hardware demand, serving as a reference for next-generation product development and emerging AI infrastructure opportunities.

No.8, 2026/4

SPECIAL REPORT

Executive summary

As generative AI moves from research and development into enterprise-scale deployment, the underlying AI infrastructure is undergoing a structural realignment driven by a new phase of compute demand. Whereas the initial wave of AI investment emphasized cloud-based training capacity, the proliferation of production-grade enterprise applications is shifting the balance toward inference compute, which is now growing faster than training compute. The reorientation is expanding enterprise infrastructure choices, as companies invest in hybrid and on-premises deployment alongside their core reliance on cloud platforms.

Large language models (LLMs) are advancing toward trillion-parameter scales while concurrently integrating capabilities such as chain-of-thought reasoning, multimodal outputs, and autonomous AI agents, and these four developments are driving adoption across six primary enterprise application areas—chatbots, software development, image generation, video generation, enterprise operations automation, and process automation. The expanding range of capabilities and diversified application demands are prompting organizations to reconsider infrastructure strategies, reducing their exclusive reliance on large-scale, centralized data centers for AI service provision. Instead, enterprises are evaluating and selecting infrastructure deployment models that reflect specific operational constraints and strategic priorities, including cost control, data sovereignty, latency, and reliability.

Cloud service providers (CSPs) delivering compute capacity are making substantial capital investments and expanding infrastructure- and platform-level AI services to maintain leadership in AI compute and to facilitate enterprise adoption. These providers are reassessing computing architectures to improve inference efficiency and to reduce customer compute costs. At the same time, several providers are broadening their software-as-a-service portfolios to capture demand for cloud compute and to raise the scale threshold required for independent compute investment, thereby creating disincentives for enterprises to develop and maintain on-premises compute infrastructure.

This analysis evaluates whether CSPs retain the ability to monopolize AI compute resources in the current market context, with particular attention to the durability of Nvidia’s platform leadership, which has historically supported large-scale training compute for LLMs. The report considers the implications of a market transition from training-centric workloads toward inference-centric deployment, assessing how this shift may alter competitive dynamics and resource concentration. It further examines the scalability of demand by estimating the potential volume of high-end AI servers that cloud compute vendors might ultimately introduce in future shipments in response to substantial enterprise AI adoption, to determine whether such deployment capacity could reinforce or erode concentrated control over AI compute.

In addition, the report examines enterprise AI adoption as the primary framework for evaluating shifts in compute architecture, providing a systematic analysis of how compute infrastructures are being reconfigured in response to evolving requirements. It evaluates the capacity of cloud compute to maintain strategic relevance amid rapidly expanding enterprise AI demand and delineates which market participants—major cloud providers, large language model providers, and compute platform providers—are likely to capture economic value as these dynamics unfold.

Finally, the report offers supply chain participants a clarified perspective on the prevailing architecture of future AI infrastructure, serving as a referential framework to inform the development of next-generation product

roadmaps, the specification of technical requirements, and the selection of strategic partners, thereby enabling precise alignment with the expanding infrastructure opportunities presented by the enterprise AI era.

Jim Hsiao

Jim Hsiao

DIGITIMES
Senior Analyst

Contents

Figure

LLM is moving beyond a singular focus on scale toward a phase defined by diversification and practical deployment, with trends consolidating into four principal directions by 2026. Parameterization is bifurcating into an M-shaped distribution: a segment of very large models continues to scale parameters to pursue peak performance, while a cohort of smaller models, typically under 10 billion parameters, emphasizes efficiency and deployability. Leading providers are intensifying efforts to decompose complex tasks and improve reasoning capabilities, enabling models to address multi-step problems with greater precision. Multimodal support has transitioned into the mainstream as models increasingly integrate and process text, image, audio, and video inputs and outputs. Concurrently, the emergence of Agentic AI is shifting LLMs from passive inference engines toward autonomous agents capable of automating processes, marking a broader movement from tool-like functionality to agentic application.

The majority of compute for generative AI applications is currently provisioned in cloud environments, and the placement of compute resources across cloud-only deployments, hybrid configurations, or on-device execution is governed by an interrelated set of variables. Latency requirements dictate proximity of inference and training workloads to end users, while data openness and sensitivity inform where data can be processed and stored in cloud. Task complexity and compute demand determines the scale and specialized hardware necessary for model development and inference, and data scale influences both storage and throughput considerations. The degree of standardization across application domains affects portability and the feasibility of distributed deployments, and prevailing levels of cloud adoption shape organizational readiness to offload workloads. These factors interact to create trade-offs between performance and cost, leading organizations to select deployment topologies that optimize responsiveness, compliance, resource utilization, and economic efficiency for their specific generative AI use cases.

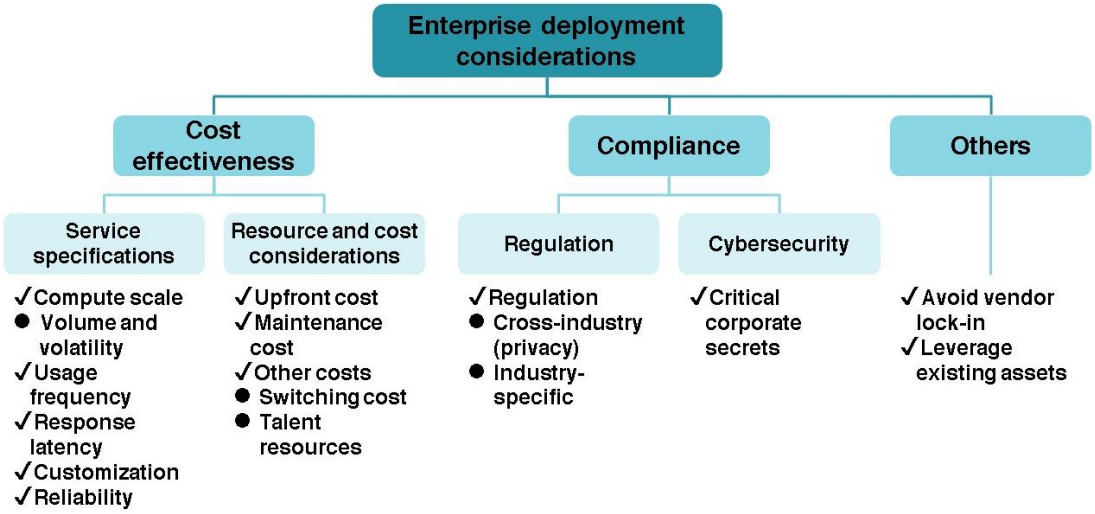
Figure: Compute location for generative AI applications

Application	Deployment model factors							Deployment
	Latency	Data openness	Task Complexity	Data scale	Compute demand	Standardization	Current cloud adoption	
Chatbots	Low	Medium	Medium	High	High	High	High	Cloud
Enterprise IT automation	Medium	Low	High	High	High	Medium	Low	Hybrid
Process automation	Low	Low	Medium	Medium	Medium	Low	Low	Edge
Software development	Medium	Medium	Low	Low	Low	High	High	Hybrid
Image generation and processing	Medium	High	Low	Low	Medium	Medium	High	Hybrid
Video generation and processing	High	High	Medium	Medium	High	Low	High	Cloud

Source: DIGITIMES, 2026/3

The global AI industry is shifting from R&D to service deployment. Enterprises planning AI investments must choose between cloud and on-premises deployments, weighing high potential value against substantial implementation costs. Key deciding factors are cost-effectiveness, compliance, and issues like corporate autonomy—factors that shape deployment choices and commercial opportunities across the AI software and hardware supply chain.

Figure: Enterprise AI deployment decision factors



Source: DIGITIMES, 2026/3

Market projections indicate a marked acceleration in AI application deployment beginning in 2026. This transition will present both strategic challenges and material opportunities for

the compute architectures and critical hardware that underpin modern data centers. As scaling laws for frontier LLMs approach diminishing returns, a paradigm shift is underway. With the integration of Chain of Thought (CoT), inference scaling is now outpacing training-side gains. However, since effective CoT reasoning remains tethered to top-tier models with trillions of parameters, these workloads continue to be centralized within large-scale cloud clusters. Major cloud providers' initial investments and competitive initiatives in cloud AI computing are projected to generate abundant, cost-efficient compute and development resources for the AI applications market in the coming years. Consequently, enterprises engaged in AI application development are likely to preferentially adopt AI cloud computing solutions from these large providers, incentivized by lower total cost of ownership and diminished technical and operational barriers to development inherent in those offerings.

Recent advances in multimodal inference models have increased their suitability for cloud-based deployment, driven by the divergent computational and data-type requirements for generating images, audio, and video compared with those for text. Contemporary mainstream multimodal architectures, such as Sora and Google Veo 3, adopt a diffusion-transformer approach that integrates diffusion models with the transformer core architecture typical of LLMs to support multimodal visual generation. This hybrid design introduces greater architectural complexity compared to conventional language models, and producing high-quality visual content imposes substantially higher computational demands on accelerators. As a result, inference for these models is more efficiently executed on cloud cluster compute environments, where the requisite scale and heterogeneous resources are more readily available.

Inference-related demand is developing rapidly, prompting a reassessment of prevailing AI server architectures that were historically optimized for training workloads and driving targeted hardware and software enhancements. Nvidia has responded with a portfolio of initiatives aimed at optimizing inference, including the Dynamo software solution and the Rubin CPX GPU, along with its associated hardware architecture, which the company has indicated will enter the market in early 2027. These offerings represent vendor-level attempts to close the gap between existing server designs and the requirements of inference workloads; however, the rapid proliferation of AI applications has generated substantially larger inference requirements than current deployments were designed to support, leaving

material scope for further refinement of AI server designs. Market attention has increasingly focused on the potential of new memory architectures as a pathway to address the evolving performance and efficiency demands of inference, reflecting a broader industry recognition that changes beyond incremental tuning of existing components may be necessary to meet future workload characteristics.

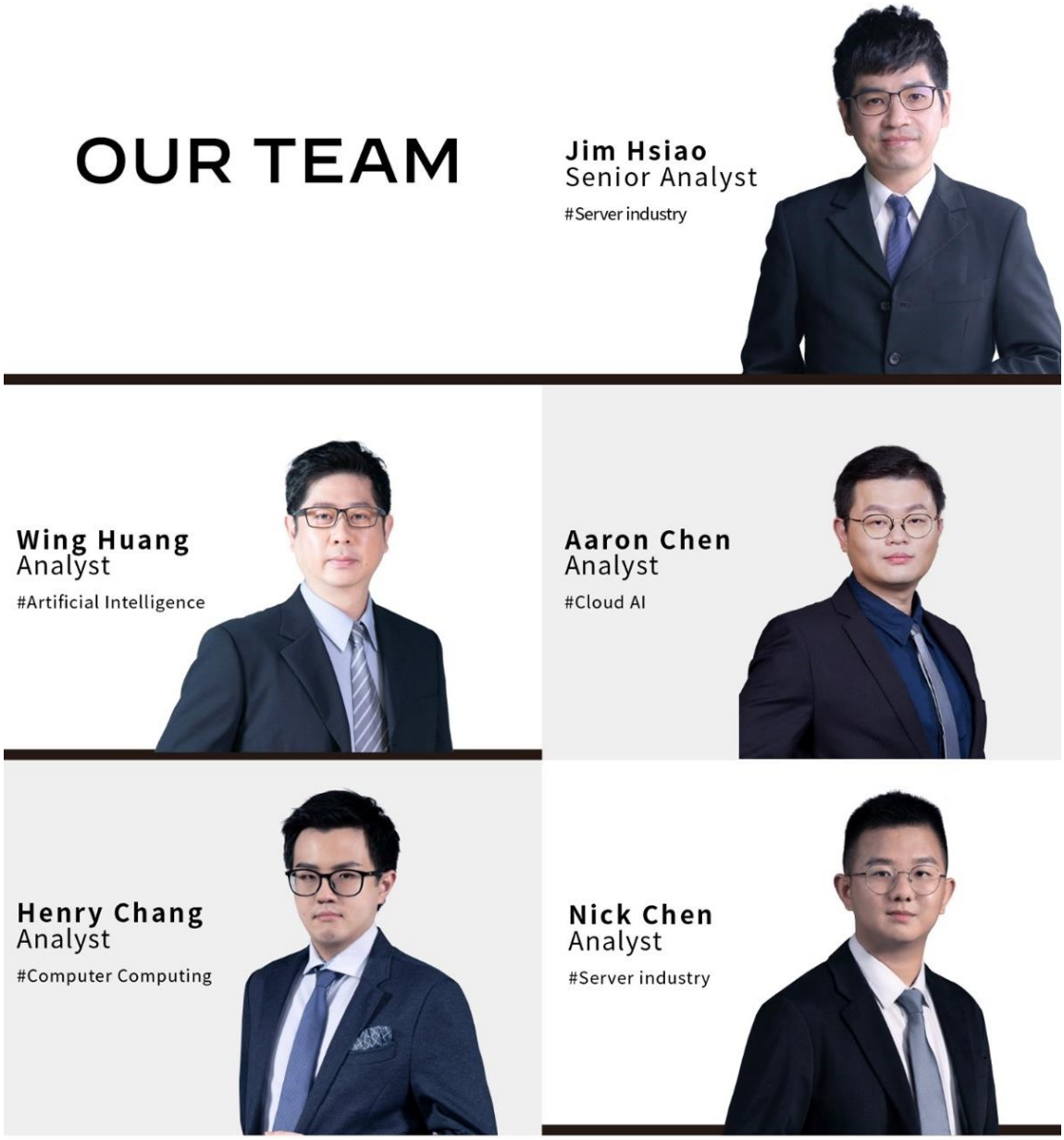
AWS delivers its IaaS-level compute through EC2, while leveraging Bedrock as the PaaS layer for model orchestration and agent development. Model development is supported through SageMaker, and AWS App Studio enables application creation via natural language. Enterprise-oriented software services include the conversational agent Amazon Q and the e-commerce shopping assistant Rufus. AWS complements its proprietary ASIC-based compute with substantial GPU capacity to address customer requirements for large-scale generative AI training and inference, incorporating hardware from Nvidia and AMD. Industry reporting from DIGITIMES indicates that Amazon has been a consistent leading purchaser in procurements of UBB-architecture AI servers, reflecting continued investment in diverse accelerator ecosystems to meet varied customer workloads.

Figure: Amazon AI IaaS evolution

		2016~2022	2023~1H25	2H25
AI IaaS	Key service	**Anthropic compute supply** • Partnered since 2021; early P4d user; used to trained Claude 1.	**Anthropic compute supply** • Claude 3 and 3.5 trained on AWS P5 and TPU; Trainium used for experiments and inference. • Claude 4 first model trained using Trainium 2. **Self-use** • Trained own Nova series models.	**Anthropic compute supply** • Claude 5 expected fully trained on Trainium 2 and Trainium 3. **Open AI compute supply** • Invested US$50B; will use Trainium 3 and Trainium 4 from 2027. **Self-use** • Trained own Nova series models.
AI compute deployment		**Nvidia** • 2017 P3 GPU provided V100 compute. • 2020 launched P4d large cluster; 4,000 A100 synchronous compute to meet early LLM training. **Amazon** • 2019 launched first Inferentia; 2022 first Trainium online.	**Amazon** • 2023 launched second-gen Inferentia; 2024 launched second-gen Trainium. **Nvidia** • 2023 began building P5 (H100); 2024 launched P5e (H200). • mid-2025 launched P6e-GB200 and P6-B200.	**Amazon** • Project Rainier to build >1M Trainium 2/3 units. **Nvidia** • P6e-GB300 service started Dec 2025. • Vera Rubin system planned 2H26; Rubin Ultra 2H27.

Source: DIGITIMES, 2026/3

Analysts


OUR TEAM
Jim Hsiao
Senior Analyst
#Server industry
Wing Huang
Analyst
#Artificial Intelligence
Aaron Chen
Analyst
#Cloud AI
Henry Chang
Analyst
#Computer Computing
Nick Chen
Analyst
#Server industry

About DIGITIMES

DIGITIMES boasts a network of over 1,200 members from Taiwan's leading tech companies, covering key players in the industry. This unique advantage allows DIGITIMES to access firsthand industry information and accurately track global technology supply chain trends. We are dedicated to advancing cutting-edge technology research, providing critical supply chain insights, and leading technological development. Our research areas include semiconductors, AI, IoT, information technology, consumer electronics, telecommunications, automotive technology, and display technologies.

Supply Chain Analysis

In today's challenging economic environment, companies need deep insights and precise data to formulate strategies. DIGITIMES provides comprehensive coverage of the entire process, from semiconductor design and manufacturing to servers and end products, including all stages of component production and distribution. We offer accurate, transparent, and systematic analysis to help businesses make informed decisions within complex supply chains.

Expert Team

DIGITIMES consists of experienced industry experts who bring deep expertise in their respective fields, delivering high-quality and thorough research reports. Our research is based on reliable and authoritative data sources, including collaborations with industry-leading companies. We employ rigorous scientific research methodologies to ensure the accuracy and credibility of our reports.

Research Report Content

DIGITIMES's reports cover global and Taiwanese production and sales data, industry development trends, technological advancements, strategies of leading companies, and competitive dynamics. We also focus on supply chain trends in regional and emerging markets and the development of key components.

Customized Research and Consulting Services

DIGITIMES provides customized research and consulting services tailored to businesses' unique needs. Our offerings include technology trend forecasting, competitor analysis, and supply chain insights. With deep industry expertise, we help businesses seize innovation opportunities, make informed decisions, and strengthen their competitive edge in a rapidly evolving tech landscape.

DIGITIMES Services: https://www.digitimes.com/reports/services/

Contact us

For any inquiries, feel free to contact us. We're here to help!

Service hours: Mon - Fri, 09:00-18:00 (UTC+8)

TEL: +886 2 8712 8866

Fax: +886 2 8712 3366

Email: subscription@digitimes.com

Disclaimer

The contents of the report provided by our company are based on information from sources recognized by us and judgments made as of a specific date. However, due to rapid industry changes, incomplete information, and other uncertain factors, we do not guarantee the accuracy and completeness of this research report in the future. Any opinions and estimates in the report are subject to change without notice.

The information in this research report is provided for general reference only and is not intended as specific advice for any particular individual or entity. Users should exercise their own judgment and take responsibility for the outcomes if they use or reference this information for decision-making purposes. Except in cases where liability is clearly attributable to DIGITIMES, users may not hold us responsible for any direct or indirect damages resulting from the use of this research report. The content of this investment report is based on information believed to be reliable but does not make any explicit or implicit representations or warranties regarding the accuracy, completeness, or correctness of the data. Opinions presented in this research report may be amended or withdrawn without notice to users. The contents of this research report are copyrighted by DIGITIMES Inc. (hereinafter referred to as DIGITIMES) and are strictly protected against copying and imitation. For specific details, please refer to the copyright statement included in the report.

Copyright statement

All content published on the DIGITIMES website, such as articles, photos, images, illustrations, audio, video, files, website layout and design, are protected by the laws of the Republic of China (Taiwan), international copyright laws, and relevant intellectual property laws. This intellectual property, including but not limited to trademarks, patents, copyrights, trade secrets, and proprietary technology, is owned by DIGITIMES or its collaborative content providers.

Violation of the copyright policy may result in legal consequences, including but not limited to fines and lawsuits. Inadvertent violation of copyright, such as not realizing a piece of content is copyrighted, will also be regarded as illegal.

Users may download or copy website and print content or services for personal, non-commercial use only. Users must comply with all relevant copyright laws. Without explicit authorization, users may not alter, publish, broadcast, resell, reproduce, modify, distribute, perform, display, or utilize any part or the entirety of the content and services on the DIGITIMES website for profit.