

The Global Market for Computing and AI for Data Centers 2026–2040

<https://marketpublishers.com/r/GDA26E51D214EN.html>

Date: April 2026

Pages: 490

Price: US\$ 1,500.00 (Single User License)

ID: GDA26E51D214EN

Abstracts

The global market for computing and artificial intelligence in data centers represents one of the most dynamic and capital-intensive segments of the semiconductor industry. Driven by the rapid proliferation of generative AI, large language models, and agentic AI systems, demand for specialised data center processors — encompassing GPUs, AI ASICs, CPUs, and FPGAs — has entered a period of extraordinary and sustained growth. From a market valued at approximately \$215 billion in 2025, the sector is projected to scale dramatically through 2040, as hyperscalers, cloud providers, and enterprises race to build the compute infrastructure required to train, fine-tune, and serve increasingly powerful AI models.

At the core of this expansion is the GPU, which remains the dominant processor architecture for AI workloads due to its unmatched parallel processing capability and mature software ecosystem. Nvidia continues to hold an overwhelming share of this segment, with successive generations — from Hopper to Blackwell to Rubin and beyond — each delivering step-change improvements in compute density, memory bandwidth, and energy efficiency. AMD provides meaningful competition with its MI-series accelerators, while the broader landscape is being reshaped by hyperscalers developing their own custom silicon to reduce dependency on merchant chip vendors and lower total cost of ownership.

AI ASICs represent the fastest-growing processor category, as companies including Google, Amazon Web Services, Microsoft, and Meta invest heavily in purpose-built chips optimised for specific workloads such as inference, recommendation, and training. These internally developed accelerators — including Google's TPU series, AWS Trainium and Inferentia, Microsoft MAIA, and Meta's MTIA — are increasingly displacing third-party GPUs for certain use cases, fundamentally altering the competitive dynamics

of the market and creating a parallel ecosystem of chip co-designers and advanced packaging specialists.

The server CPU market, though more mature, continues to evolve rapidly. Intel and AMD maintain leading positions with their x86 architectures, but face mounting pressure from Arm-based alternatives championed by hyperscalers such as AWS with Graviton, Google with Axion, Microsoft with Cobalt, and Nvidia with Grace and Vera. RISC-V is also emerging as a credible contender for specific workloads, particularly as open-source hardware ecosystems mature. Meanwhile, FPGAs continue to serve niche roles in low-latency and specialised inference applications.

Underpinning all of this is a complex and increasingly strained supply chain. Advanced semiconductor manufacturing is concentrated at TSMC, Samsung, and Intel Foundry, with leading-edge nodes below 5nm commanding the majority of AI chip demand. High Bandwidth Memory, supplied primarily by SK Hynix, Samsung, and Micron, has emerged as a critical bottleneck, while advanced packaging technologies such as CoWoS are operating at near-full capacity. Hyperscaler capital expenditure continues to flow into data centre construction, power infrastructure, and silicon procurement at a scale that is reshaping global semiconductor supply chains.

Geopolitics adds a further layer of complexity. US export controls on advanced AI chips have accelerated China's drive toward semiconductor self-sufficiency, with domestic players such as Huawei HiSilicon, Cambricon, Biren, and Hygon developing increasingly capable alternatives. The bifurcation of the global AI compute market into US-aligned and China-domestic supply chains is one of the defining structural trends of the decade, with profound implications for technology strategy, investment allocation, and national industrial policy.

The Global Market for Computing and AI for Data Centers 2026–2040 is a comprehensive strategic intelligence report covering the full landscape of data centre processor technology, market dynamics, competitive positioning, and long-range forecasting through to 2040. Produced for technology executives, semiconductor investors, strategic planners, and policy analysts, the report provides the depth of quantitative rigour and qualitative insight required to navigate one of the most rapidly evolving markets in the global economy.

The report opens with a set of preliminary materials including a detailed glossary of technical terms and abbreviations, a clear articulation of research objectives and scope, biographical profiles of the authoring team, and a candid retrospective on previous

forecast accuracy. This is followed by a three-page summary and a full executive summary designed for senior readers who require rapid orientation to the report's key findings without sacrificing analytical depth.

Chapter one establishes the macroeconomic and geopolitical context, examining global AI infrastructure investment trends, hyperscaler capital expenditure trajectories for both US and Chinese players, the evolving regulatory landscape including US export controls, and the widening technology divide between Western and Chinese semiconductor ecosystems.

Chapter two forms the quantitative heart of the report, delivering granular market forecasts from 2021 to 2040 across all major processor categories. Revenue, average selling price, unit volume, wafer consumption, and server tray forecasts are provided at the vendor, product, and technology node level, enabling readers to build detailed bottom-up views of market opportunity and competitive exposure. Separate analytical lenses are provided for CPU, GPU, and AI ASIC dynamics, including HBM-driven revenue disaggregation and compute die forecasting.

Chapter three addresses the market forces shaping demand, including the falling cost of generative AI inference and training, the emergence of agentic and physical AI, the compute demands of recommendation engines and coding assistants, the competition between LLMs and traditional search, and broader questions around the CapEx and OpEx economics of AI infrastructure. An exploratory section examines the longer-term possibility of space-based data center architectures.

Chapter four maps the competitive landscape in detail, providing ecosystem maps for both the data center processor supply chain and the foundation model developer community. It includes financial benchmarking of leading chip designers, a deep-dive case study on OpenAI's revenue and compute trajectory, comprehensive market share analysis, and a dedicated section on Mainland China covering domestic market sizing, hyperscaler demand, manufacturer profiles, and supply chain structure.

Chapter five delivers an authoritative review of technology trends across all processor categories, covering process node roadmaps, chiplet architectures, rack-scale system designs, memory and packaging technology, and emerging computing paradigms including photonics, neuromorphic, and quantum computing. Unique assets include a full AI ASIC technology specification database and a start-up landscape analysis.

The report concludes with a forward-looking outlook chapter presenting bull, base, and

bear case scenarios for the market through 2031 and beyond to 2040, a comprehensive risk register, and strategic recommendations. An extensive company profiles section — covering 81 organisations with one dedicated page per company — rounds out the report, providing standardised strategic and financial snapshots of every major player in the ecosystem.

Report Contents include:

Global AI infrastructure and investment landscape

US and Chinese hyperscaler CapEx trends and projections

AI regulatory landscape and export controls

The US–China technology divide

Market Forecasts (2021–2040)

Total data centre processor revenue forecast

GPU, AI ASIC, CPU and FPGA revenue forecasts

Average selling price (ASP) forecasts by vendor and product tier

Processor unit shipment forecasts

Wafer starts by technology node and foundry (TSMC, Samsung, Intel Foundry)

GPU and AI ASIC compute die forecasts

HBM-driven revenue separation

Server tray volume forecasts

Dedicated CPU focus and GPU/AI ASIC focus sections

Market Trends

Cost of generative AI inference and training

From agentic AI to physical AI

Recommendation models for social networks

Coding assistants

Search engines vs. LLMs

OpenClaw

CapEx vs. OpEx in the generative AI era

The future of space-based AI data centres

Market Share & Supply Chain

Data center ecosystem map

Foundation models ecosystem map

US vs. China tech war timeline

Financial metrics of data center chip designers

Case study: OpenAI revenue and gigawatt forecast

Market share analysis — CPU, GPU, AI ASIC, XPU co-designers

Mainland China focus: market size, hyperscaler demand, manufacturer profiles, supply chain

Technology Trends

CPU: x86, Arm, RISC-V, workload specialisation

GPU: process nodes, chiplets, rack-scale architecture, HBM integration, interconnects

AI ASIC: hyperscaler roadmaps, start-up landscape, specification database, disaggregated inference

GPU vs. AI ASIC comparative analysis

Advanced packaging and HBM (HBM2E through HBM4), CoWoS, AI rack bill of materials

Emerging computing: photonics, neuromorphic, quantum

Outlook

Market outlook 2026–2040 with bull/base/bear scenarios

Technology outlook 2026–2040

Key risks and opportunities

Strategic recommendations

Company Profiles

81 individual company profiles, one page per company, covering strategy, products, financials, and roadmap. Companies profiled include 01.AI, Achronix Semiconductor, Advanced Micro Devices (AMD), AI21 Labs, Alchip Technologies, Aleph Alpha, Alibaba Group / T-Head Semiconductor, Amazon Web Services (AWS), Ampere Computing, Anthropic, Arm Holdings, Axelera AI, Baidu, Biren Technology, Broadcom, ByteDance, Cambricon Technologies, Cerebras Systems, China Mobile, Cisco Systems, Cohere, CoreWeave, d-Matrix, DeepSeek, Dell Technologies, Enflame Technology, Esperanto Technologies, Etched, Fujitsu, Furiosa AI, GlobalFoundries (GF), Google (DeepMind / TPU Programme), GrAI Matter Labs, Graphcore, Groq, GUC (Global Unichip Corp.), Hewlett Packard Enterprise (HPE), HiSilicon Technologies, Huawei Technologies, Hygon Information Technology, IBM, Iluvatar CoreX, Intel Corporation, Kalray, Lattice Semiconductor, Lightmatter and more.....

Contents

PRELIMINARY SECTIONS

Glossary of Terms and Abbreviations
Objective of the Report
Scope of this Report
About the Authors
What We Got Right, What We Got Wrong
3-Page Summary
Executive Summary

CHAPTER 1 — CONTEXT

1.1 Global AI Infrastructure and Investment Landscape
1.2 US and Chinese Hyperscaler CapEx Trends and Projections
1.3 AI Regulatory Landscape and Export Controls
1.4 The US–China Technology Divide

CHAPTER 2 — MARKET FORECASTS

2.1 Processor Revenue Forecast
2.1.1 Total Data Center Processor Market, 2021–2040 (\$B)
2.1.2 GPU Revenue Forecast, 2021–2040 (\$B)
2.1.3 AI ASIC Revenue Forecast, 2021–2040 (\$B)
2.1.4 Server CPU Revenue Forecast, 2021–2040 (\$B)
2.1.5 FPGA Data Center Revenue Forecast, 2021–2040 (\$M)
2.2 Average Selling Price (ASP) Forecast
2.2.1 GPU ASP Trends by Product Tier, 2021–2040 (\$K)
2.2.2 AI ASIC ASP Trends by Hyperscaler, 2021–2040 (\$K)
2.2.3 CPU ASP Trends — Intel Xeon vs. AMD EPYC, 2021–2040
2.3 Processor Volume Forecast
2.3.1 GPU Unit Shipments by Vendor, 2021–2040 (K units)
2.3.2 AI ASIC Unit Shipments by Hyperscaler, 2021–2040 (K units)
2.3.3 CPU Unit Shipments by Vendor, 2021–2040 (M units)
2.4 Wafer Forecast
2.4.1 GPU & AI ASIC Wafer Starts by Technology Node, 2021–2040
2.4.2 Wafer Starts by Foundry (TSMC, Samsung, Intel Foundry)
2.4.3 GPU & AI ASIC Compute Die Forecast, 2021–2040

- 2.4.4 HBM-Driven Revenue Separation from GPU & AI ASIC
- 2.5 Server Tray Volume Forecast
- 2.6 CPU Focus
- 2.7 GPU & AI ASIC Focus

CHAPTER 3 — MARKET TRENDS

- 3.1 Cost of Generative AI Inference and Training
- 3.2 From Agentic AI to Physical AI
- 3.3 Recommendation Models for Social Networks
- 3.4 Coding Assistants
- 3.5 Search Engine vs. LLM
- 3.6 OpenClaw
- 3.7 CapEx vs. OpEx in the Era of Generative AI
- 3.8 Is the Future of AI Data Centers in Space?

CHAPTER 4 — MARKET SHARE & SUPPLY CHAIN

- 4.1 Data Center Ecosystem Map
- 4.2 Foundation Models Ecosystem Map
- 4.3 U.S. vs. China Tech War — Timeline
- 4.4 Financial Metrics of Data Center Chip Designers
- 4.5 Case Study: OpenAI Revenue and Gigawatt
- 4.6 Market Share: CPU, GPU, AI ASIC & XPU Co-Designers
 - 4.6.1 GPU Market Share by Revenue and Units
 - 4.6.2 AI ASIC Market Share by Hyperscaler
 - 4.6.3 CPU Market Share by Vendor
 - 4.6.4 XPU Co-Designer Revenue Market Share
- 4.7 Focus on Mainland China
 - 4.7.1 Chinese DC Processor Market Size & Forecast
 - 4.7.2 Chinese Hyperscaler Processor Demand
 - 4.7.3 Chinese Processor Manufacturer Profiles & Roadmaps
 - 4.7.4 China DC Processor Supply Chain

CHAPTER 5 — TECHNOLOGY TRENDS

- 5.1 CPU Technology Trends
 - 5.1.1 x86 Architecture Evolution
 - 5.1.2 Arm-Based CPU Momentum in the Data Center

- 5.1.3 RISC-V in the Data Center
- 5.1.4 CPU Specialisation for AI Workloads
- 5.2 GPU Technology Trends
 - 5.2.1 Process Node Roadmap and Transition
 - 5.2.2 Chiplet and Multi-Die Architectures
 - 5.2.3 Rack-Scale GPU Architectures (NVL72 and Beyond)
 - 5.2.4 Memory Bandwidth and HBM Integration
 - 5.2.5 Networking and Interconnect Evolution
- 5.3 AI ASIC Technology Trends
 - 5.3.1 Hyperscaler ASIC Product Roadmaps
 - 5.3.2 AI ASIC Start-Up Landscape
 - 5.3.3 AI ASIC Technology Specification Database
 - 5.3.4 Compute Disaggregation for AI Inference
- 5.4 GPU vs. AI ASIC: Comparative Analysis
- 5.5 Advanced Packaging and HBM Memory
 - 5.5.1 HBM Technology Roadmap (HBM2E to HBM4)
 - 5.5.2 CoWoS and Advanced Packaging Capacity
 - 5.5.3 Custom HBM and Co-Design Trends
 - 5.5.4 AI Rack Bill of Materials
- 5.6 Emerging Computing Architectures
 - 5.6.1 Photonic Computing
 - 5.6.2 Neuromorphic Computing
 - 5.6.3 Quantum Computing Outlook

CHAPTER 6 — OUTLOOK

- 6.1 Market Outlook 2026–2040
- 6.2 Technology Outlook 2026–2040
- 6.3 Key Risks and Opportunities
- 6.4 Strategic Recommendations

CHAPTER 7 — COMPANY PROFILES (81 COMPANY PROFILES)

List Of Figures

LIST OF FIGURES

- Fig. 1.1 Global AI Infrastructure Investment Forecast, 2021–2040 (\$B)
- Fig. 1.2 US vs. Chinese Hyperscaler CapEx, 2021–2040 (\$B)
- Fig. 1.3 Data Center Power Consumption Forecast, 2024–2040 (GW)
- Fig. 1.4 AI-Related Data Center Construction Starts by Region, 2022–2028
- Fig. 1.5 US Export Controls on AI Chips — Key Milestones, 2019–2026
- Fig. 1.6 US–China Technology Decoupling Timeline, 2018–2026
- Fig. 2.1 Total Data Center Processor Market Revenue Forecast, 2021–2040 (\$B)
- Fig. 2.2 Revenue Breakdown by Processor Type (CPU, GPU, AI ASIC, FPGA), 2021–2040
- Fig. 2.3 Data Center Processor CAGR by Category, 2025–2040 (%)
- Fig. 2.4 GPU Market Revenue Forecast, 2021–2040 (\$B)
- Fig. 2.5 GPU Revenue Split by Vendor (Nvidia, AMD, Others), 2021–2040
- Fig. 2.6 Nvidia GPU Revenue by Product Generation, 2021–2028 (\$B)
- Fig. 2.7 AMD GPU Revenue by Product Generation, 2021–2028 (\$B)
- Fig. 2.8 AI ASIC Market Revenue Forecast, 2021–2040 (\$B)
- Fig. 2.9 AI ASIC Revenue Split by Hyperscaler, 2021–2040
- Fig. 2.10 Server CPU Market Revenue Forecast, 2021–2040 (\$B)
- Fig. 2.11 Server CPU Revenue Split by Architecture (x86 vs. Arm), 2021–2040
- Fig. 2.12 FPGA Data Center Revenue Forecast, 2021–2040 (\$M)
- Fig. 2.13 GPU ASP Evolution by Product Tier, 2021–2040 (\$K)
- Fig. 2.14 AI ASIC ASP Trends by Hyperscaler, 2021–2040 (\$K)
- Fig. 2.15 Server CPU ASP Trends — Intel Xeon vs. AMD EPYC, 2021–2040 (\$)
- Fig. 2.16 GPU Unit Shipments by Vendor, 2021–2040 (K units)
- Fig. 2.17 Nvidia GPU Unit Shipments by Product Generation, 2021–2028
- Fig. 2.18 AMD GPU Unit Shipments by Product Generation, 2021–2028
- Fig. 2.19 AI ASIC Unit Shipments by Hyperscaler, 2021–2040 (K units)
- Fig. 2.20 Google TPU Unit Deployment Forecast, 2021–2040
- Fig. 2.21 AWS Trainium & Inferentia Unit Forecast, 2021–2040
- Fig. 2.22 Microsoft MAIA Unit Forecast, 2021–2040
- Fig. 2.23 CPU Unit Shipments — Data Center, 2021–2040 (M units)
- Fig. 2.24 Intel vs. AMD CPU Market Share in Unit Terms, 2021–2040 (%)
- Fig. 2.25 Hyperscaler Custom CPU Unit Adoption, 2022–2040 (M units)
- Fig. 2.26 GPU & AI ASIC Wafer Starts by Technology Node, 2021–2040 (KW/month)
- Fig. 2.27 Wafer Consumption Split: Advanced Nodes (

List Of Tables

LIST OF TABLES

- Table 2.1 Data Center Processor Market Revenue Summary, 2021–2040 (\$B)
- Table 2.2 GPU Revenue by Vendor, 2021–2040 (\$B)
- Table 2.3 AI ASIC Revenue by Hyperscaler, 2021–2040 (\$B)
- Table 2.4 Server CPU Revenue by Vendor, 2021–2040 (\$B)
- Table 2.5 GPU ASP by Product Tier, 2021–2040 (\$K)
- Table 2.6 AI ASIC ASP by Hyperscaler, 2021–2040 (\$K)
- Table 2.7 GPU Unit Shipments by Vendor, 2021–2040 (K units)
- Table 2.8 AI ASIC Unit Shipments by Hyperscaler, 2021–2040 (K units)
- Table 2.9 CPU Unit Shipments by Vendor, 2021–2040 (M units)
- Table 2.10 GPU & AI ASIC Wafer Starts by Node and Foundry, 2021–2040
- Table 2.11 AI Server vs. General-Purpose Server Tray Volume, 2021–2040 (M units)
- Table 2.12 CPU Processor Roadmap Summary — Major Vendors, 2024–2030
- Table 2.13 GPU & AI ASIC Product Roadmap Summary, 2024–2030
- Table 3.1 Cost per Token by Model Size and Hardware Configuration, 2024–2040
- Table 3.2 Agentic AI Use Cases by Industry and Hardware Requirements
- Table 3.3 Coding Assistant Market Share and Underlying Infrastructure, 2024
- Table 4.1 Financial Metrics — Top 10 Data Center Chip Designers, 2021–2025
- Table 4.2 US and Chinese Hyperscaler CapEx Summary, 2021–2026 (\$B)
- Table 4.3 AI Semiconductor Start-Up Fundraising Database, 2019–Q1 2026
- Table 4.4 GPU Market Share Summary by Revenue and Units, 2021–2025
- Table 4.5 AI ASIC Specifications — Google, AWS, Microsoft, Meta, 2024–2026
- Table 4.6 Chinese Data Center Processor Manufacturer Overview
- Table 4.7 China DC Processor Supply Chain — Key Component Suppliers
- Table 5.1 CPU Specifications — Intel, AMD, AWS, Google, Microsoft, Huawei, Nvidia, 2024–2026
- Table 5.2 GPU Specifications — Nvidia Blackwell, Rubin; AMD MI350X, MI450, 2024–2026
- Table 5.3 AI ASIC Technology Specification Database (Full, All Major Vendors)
- Table 5.4 HBM Specification Comparison — HBM2E, HBM3, HBM3E, HBM4
- Table 5.5 AI Server Rack BoM — Itemised Cost Breakdown, 2025 (\$K)
- Table 5.6 Emerging Computing Technology Readiness Assessment
- Table 6.1 Market Forecast Summary — Bull / Base / Bear Scenarios, 2026–2040 (\$B)
- Table 6.2 Key Risk Register — Probability and Impact Assessment

I would like to order

Product name: The Global Market for Computing and AI for Data Centers 2026–2040

Product link: <https://marketpublishers.com/r/GDA26E51D214EN.html>

Price: US\$ 1,500.00 (Single User License / Electronic Delivery)

If you want to order Corporate License or Hard Copy, please, contact our Customer Service:

info@marketpublishers.com

Payment

To pay by Credit Card (Visa, MasterCard, American Express, PayPal), please, click button on product page <https://marketpublishers.com/r/GDA26E51D214EN.html>