

AI GPU Chip Market Size & Share Analysis - Growth Trends and Forecast (2026 - 2031)

Industry: Artificial Intelligence & ML

URL: <https://www.mordorintelligence.com/industry-reports/ai-gpu-chip-market>

AI GPU Chip Market Size and ShareMarket OverviewStudy Period | 2020 - 2031 |

Market Size (2026) | USD 288.40 Billion |

Market Size (2031) | USD 621.70 Billion |

Growth Rate (2026 - 2031) | 16.60 % |

Fastest Growing Market | Middle East and Africa |

Largest Market | North America |

Market Concentration | High |

Major Players*Disclaimer: Major Players sorted in no particular order

Image © Mordor Intelligence. Reuse requires attribution under CC BY 4.0.

Image © Mordor Intelligence. Reuse requires attribution under CC BY 4.0.

AI GPU Chip Market Analysis by Mordor IntelligenceThe AI GPU chip market size is expected to grow from USD 167.5 billion in 2025 to USD 288.4 billion in 2026 and is forecast to reach USD 621.7 billion by 2031 at 16.60% CAGR over 2026-2031. Demand is rising because AI model training and inference now require larger and denser compute fleets across hyperscale data centers, enterprise sites, and public compute programs. Government-backed procurement is adding a durable layer of spending because national compute capacity is now being treated as a strategic asset in several regions. The top of the vendor base remains concentrated because software compatibility, interconnect standards, and access to advanced packaging still influence buying decisions more than price alone. At the same time, enterprise fine-tuning, robotics workloads, and edge inference are widening the customer base for the AI GPU chip market beyond a small group of hyperscalers. The main opportunity through 2031 lies in systems that improve memory bandwidth, cooling efficiency, and deployment flexibility, even as packaging constraints and custom accelerators keep competition active.

Key Report Takeaways* By product type, data center AI GPUs held 93.11% of revenue in 2025, while edge AI GPUs are projected to expand at a 17.44% CAGR through 2031.

* By compute function, training GPUs held 52.33% of demand in 2025, while inference GPUs are projected to expand at a 17.62% CAGR through 2031.

* By deployment environment, hyperscale and cloud accounted for 72.42% of the AI GPU chip market size in 2025, while government and research institutions are projected to expand at a 17.73% CAGR through 2031.

* By workload, generative AI and large language models accounted for 48.12% of demand in 2025, while computer vision and robotics are projected to expand at a 17.32% CAGR through 2031.

* By geography, North America held 38.44% of the AI GPU chip market share in 2025, while the Middle East and Africa is projected to expand at a 17.42% CAGR through 2031.

Note: Market size and forecast figures in this report are generated using Mordor Intelligence's proprietary estimation framework, updated with the latest available data and insights as of January 2026.

Market Trends and InsightsDrivers Impact Analysis of AI GPU Chip Market*Driver | (~) % Impact on CAGR Forecast | Geographic Relevance | Impact Timeline | Expanding Enterprise Fine-Tuning of Proprietary Models | +3.1% | Global, strongest in North America and Europe | Medium term (2-4 years) | Rapid Scale-Up of Hyperscale AI Training Clusters | +2.9% | North America, Asia-

Pacific, and Europe | Short term (≤ 2 years) |
HBM4 Readiness and Advanced Packaging Upgrade Cycle | +2.6% | Global, centered on Asia-Pacific manufacturing, Taiwan and South Korea | Short term (≤ 2 years) |
Sovereign AI Procurement and Domestic Compute Security | +2.3% | Middle East, Europe, Asia-Pacific, expanding to South America and Africa | Medium term (2-4 years) |

NVLink-CXL and UALink Pooling of Accelerator Capacity | +1.9% | North America and Europe, early Asia-Pacific hyperscaler adoption | Long term (≥ 4 years) |
Liquid Cooling Standardization for High-TDP GPU Racks | +1.6% | Global, with early adoption in North America, Europe, and Japan | Medium term (2-4 years) |

Source: Mordor Intelligence |

Expanding Enterprise Fine-Tuning of Proprietary Models Enterprise fine-tuning has moved beyond one-off pilot work and is becoming a recurring operating practice for companies that want models trained on proprietary data. That shift matters for the AI GPU chip market because repeated retraining, evaluation, and deployment create ongoing hardware demand rather than a single purchase cycle. Many enterprises also want mixed clusters that support both fine-tuning and inference, which increases the value of versatile GPU configurations instead of narrow, fixed-purpose systems. On-premise economics are becoming easier to justify for high-utilization AI work, and Lenovo reported that on-premise generative AI deployments can reach breakeven against cloud in under 4 months for sustained workloads. Data control requirements in regulated sectors are also keeping some model customization closer to internal infrastructure, which broadens the buyer base of the AI GPU chip market beyond hyperscalers.

Rapid Scale-Up of Hyperscale AI Training Clusters The AI GPU chip market is still being shaped by larger hyperscale training clusters that need tightly integrated racks, dense networking, and more advanced cooling. These purchases are no longer limited to one training wave because serving fleets also need to grow after model deployment, which keeps procurement cycles active across both training and inference estates. The newest rack-scale systems are being ordered for large frontier workloads, and that raises demand for high-end accelerators, switching, power delivery, and memory in the same build cycle. AMD reinforced this pattern in February 2026 when it announced a multi-year 6-gigawatt partnership with Meta to deploy AMD Instinct GPUs across Meta's AI data centers.[1]AMD, "AMD and Meta Announce Expanded Strategic Partnership to Deploy 6 Gigawatts of AMD GPUs," AMD Newsroom, amd.com As long as hyperscalers continue to separate frontier training fleets from large inference fleets, the AI GPU chip market is likely to see more continuous buying than in earlier compute upgrade cycles.

HBM4 Readiness and Advanced Packaging Upgrade Cycle HBM4 is becoming an important demand trigger for the AI GPU chip market because bandwidth and memory capacity now influence system value as much as raw compute throughput. Siemens noted that JEDEC's HBM4 standard uses a 2,048-bit interface and can deliver up to 2.0 TB/s per stack, which raises the performance ceiling for next-generation AI processors.[2]Siemens, "HBM3e and HBM4, IC Design Guide for Next-Generation High Bandwidth Memory," Siemens Semiconductor Packaging Blog, siemens.com As suppliers ramp HBM4 in 2026, buyers are already aligning future platform decisions with the memory road map rather than treating memory as a secondary component. That change shortens the perceived life of current HBM3e-based fleets, especially for buyers that want to stay close to the leading performance tier. The result is that the AI GPU chip market is being supported not only by new deployments, but also by faster refresh decisions tied to memory architecture and package design.

Sovereign AI Procurement and Domestic Compute Security Sovereign procurement is adding a distinct policy-led demand layer to the AI GPU chip market, and this spending is less sensitive to short-term commercial return thresholds. South Korea's government selected operators for a KRW 2 trillion project, equivalent to USD 1.4 billion, with plans to secure 9,704 advanced GPUs including NVIDIA B300 and Vera Rubin systems. The UAE expanded its public AI infrastructure agenda through partnerships that include NVIDIA, while Abu Dhabi's Technology Innovation Institute and NVIDIA launched the Middle East's first joint AI and robotics research

lab in 2025. Africa is also building academic compute capacity, and the University of Cape Town launched the African Compute Initiative in 2026 to establish the continent's largest higher education AI compute cluster.[3]University of Cape Town, "UCT to Lead Africa's First Higher Education Dedicated AI Compute Initiative," UCT AI Initiative, ai.uct.ac.za Because these programs are framed around capability, resilience, and local control, they give the AI GPU chip market support even when commercial budgets tighten.

Restraints Impact Analysis of AI GPU Chip Market*
Restraint | (~) % Impact on CAGR Forecast | Geographic Relevance | Impact Timeline |
Advanced Packaging Capacity Bottlenecks | -2.1% | Global, centered in Taiwan, TSMC CoWoS, and South Korea | Short term (≤ 2 years) |
Rising Total Cost of Ownership for Cluster-Scale Deployments | -1.8% | Global, most acute in North America and Europe where power costs are high | Medium term (2-4 years) |
Export Controls and Geopolitical Supply Friction | -1.5% | China, Macau, spill-over risk to Southeast Asia and Middle East | Medium term (2-4 years) |
Competition From Custom ASICs and Proprietary Accelerators | -1.2% | North America, hyperscaler-internal, emerging in Asia-Pacific, Huawei Ascend | Long term (≥ 4 years) |

Source: Mordor Intelligence |

Advanced Packaging Capacity Bottlenecks Advanced packaging remains a practical limit on how fast the AI GPU chip market can convert design demand into shipped systems. Modern AI accelerators depend on complex integration of logic dies and stacked high-bandwidth memory, and that makes packaging yield and throughput as important as wafer supply. Siemens highlighted the growing complexity of HBM4 integration, and the move to higher bandwidth and denser stack configurations increases the burden on packaging lines. Even when vendors have strong product demand, delivery schedules can still stretch because memory, packaging, and backend assembly must all scale together. This restraint slows volume growth, favors vendors with stronger supply relationships, and keeps the AI GPU chip market dependent on a narrow manufacturing base in the near term.

Rising Total Cost of Ownership for Cluster-Scale Deployments Hardware cost is only one part of the spending burden in the AI GPU chip market because power, cooling, networking, staffing, and maintenance all rise with cluster size. Lenovo reported in 2026 that on-premise deployments can outperform cloud economics at high utilization, but that same analysis also shows how much planning and sustained usage are required to justify the capital outlay. This creates a split market where hyperscalers and very large enterprises can absorb the system-level cost, while smaller buyers face a higher barrier even when they want local infrastructure. Liquid cooling requirements are also becoming standard for top-tier racks, which adds facility work and extends project timelines. As a result, the AI GPU chip market still has strong demand, but not every interested buyer can move from evaluation to full deployment at the same speed.

*Our forecasts treat driver/restraint impacts as directional, not additive. The impact forecasts reflect baseline growth, mix effects, and variable interactions.

AI GPU Chip Market Segment Analysis
By Product Type: Data Centers Command Share as of 2025
Edge AI GPUs Accelerate
Data center AI GPUs held 93.11% of the AI GPU chip market share in 2025, and that concentration reflected where the newest hardware could be deployed at scale. The leading products are designed around dense racks, high-speed interconnects, and specialized cooling, which makes large data center environments the natural fit for current flagship platforms. This also keeps vendor competition centered on full system design rather than on the chip alone, because deployment success depends on memory, networking, and thermal management working together. The AI GPU chip market therefore still leans heavily toward centralized compute environments even as new demand pockets begin to appear.

Edge AI GPUs are projected to expand at a 17.44% CAGR through 2031, and that growth is tied to robotics, industrial automation, and localized inference needs. NVIDIA's robotics platform design, which links DGX systems for training with RTX PRO servers for simulation and Jetson hardware for on-device inference, shows how the edge stack is becoming part of a broader AI deployment model. Client AI GPUs remain a smaller part of the AI GPU chip industry, but they are gaining relevance as device makers add AI-native features to workstations and laptops. NVIDIA's RTX Spark announcement in 2026 showed that client devices are becoming another entry point for AI GPU adoption, especially where local model execution, design workflows, and compact inferencing are important.

Image © Mordor Intelligence. Reuse requires attribution under CC BY 4.0. By Compute Function: Training and Inference Platforms Diverge Architecturally Training GPUs accounted for 52.33% of the AI GPU chip market size in 2025, and that lead came from frontier model development and large public compute programs. Training platforms still need the highest interconnect density and the most aggressive scaling behavior, which supports continued demand for premium rack architectures. Inference GPUs, however, are projected to expand at a 17.62% CAGR through 2031, and that difference shows how model serving is becoming the larger recurring compute task after training is complete. The AI GPU chip market is therefore shifting from a training-first narrative to a more balanced model where deployment intensity matters as much as model creation.

Mixed training and inference platforms are gaining a practical role in enterprise environments that cannot justify separate fleets for each workload. These buyers often need a shared cluster that can fine-tune models, run evaluation cycles, and serve applications from the same installed base. That operating pattern broadens the middle of the market and keeps demand from concentrating only in the most expensive training hardware. It also explains why the AI GPU chip industry is seeing more interest in memory-rich and flexible configurations that trade some peak specialization for higher overall utilization.

By Deployment Environment: Hyperscale Clouds Lead as Governments Surge Hyperscale and cloud accounted for 72.42% share of the AI GPU chip market size in 2025, and that dominance came from the build-out by the largest cloud providers. These operators still set the pace for high-end volume because they can buy complete fleets, fill them quickly, and absorb the supporting cost of power, networking, and cooling. Enterprise data centers formed the next layer of demand as companies invested in local clusters for proprietary inference, fine-tuning, and internal AI services. The AI GPU chip market remains highly exposed to hyperscale build cycles, but the base of enterprise buyers is widening as local deployment cases become clearer.

Government and research institutions are projected to expand at a 17.73% CAGR through 2031, which makes them the fastest-growing deployment environment. OECD analysis from 2025 showed that public compute planning is increasingly assessed through sovereignty and national capability criteria rather than only through commercial cost logic. Enterprise buyers are also using utilization thresholds to compare cloud and on-premise economics, and Lenovo reported that breakeven can arrive in under 4 months for sustained high-use deployments. Edge and endpoint deployments remain smaller, but they are adding incremental demand where local response time, data handling, or physical system control matters.

Image © Mordor Intelligence. Reuse requires attribution under CC BY 4.0. By Workload: Generative AI Dominates While Robotics and Vision Rise Generative AI and large language models held 48.12% of workload demand in 2025, and that lead reflected both training intensity and the serving infrastructure needed after deployment. This category continues to account for the largest installed demand base because large language models require substantial memory bandwidth, cluster coordination, and constant inference support once they are in production. Recommendation and search, speech and natural language processing, and scientific computing all

so remain part of the AI GPU chip market, especially where workloads run continuously against large data sets. The current mix shows that generative AI remains the main revenue anchor, even as adjacent workloads begin to scale faster.

Computer vision and robotics are projected to expand at a 17.32% CAGR through 2031, and that growth is being supported by physical AI use cases across industrial, logistics, and service settings. NVIDIA's 2026 robotics system framework shows how training, simulation, and edge inference are being linked into one deployment path for machine vision and autonomous action. AMD and OpenCV also announced collaboration in 2026 to accelerate computer vision and vision AI workloads on AMD hardware, which shows that software support is widening beyond a single ecosystem. As these use cases scale, the AI GPU chip market is likely to see more demand for localized inference, lower latency, and systems that can operate reliably outside traditional hyperscale environments.

Geography Analysis
North America AI GPU Chip MarketNorth America held 38.44% of the global AI GPU chip market in 2025, and the region remained the largest buyer because it combines hyperscale spending with the deepest developer ecosystem. The United States still anchors most of that demand through cloud platform investment, software compatibility, and system-level integration around CUDA and NVLink. Canada is becoming more active in sovereign compute, and Bell and Cohere signed a USD 220 million agreement in June 2026 to deploy 2,304 NVIDIA Grace Blackwell GB200 NVL72 systems in British Columbia. Mexico benefits more through manufacturing and assembly ties to the United States than through large domestic AI GPU deployments at this stage. This keeps North America at the center of near-term volume for the AI GPU chip market even as more regions build local compute agendas.

Europe and APAC AI GPU Chip MarketEurope is building a larger sovereign compute role in the AI GPU chip market, with policy, public funding, and compliance all pushing demand toward domestic infrastructure. OECD work on public cloud compute availability supports the view that public-sector AI capacity is increasingly being evaluated through resilience and sovereignty criteria. Asia-Pacific presents a broader mix, from South Korea's USD 1.4 billion national GPU program to growing demand in India and Southeast Asia as domestic model development expands. France also signaled willingness to diversify vendors for sovereign systems, which suggests the region may support more than one software and hardware stack as procurement matures.

MEA and South America AI GPU Chip MarketThe Middle East and Africa is projected to expand at a 17.42% CAGR through 2031, which gives it the fastest regional growth rate in the AI GPU chip market. The UAE continues to build institutional AI capacity, and the Technology Innovation Institute's partnership with NVIDIA gives the region a formal research base in robotics and advanced AI systems. Africa is also adding academic compute infrastructure, and the University of Cape Town launched the African Compute Initiative in 2026 to expand research access to high-end AI systems. South America remains smaller in current scale, but Brazil's plan for a USD 360 million AI supercomputer due in 2027 shows that the region is entering the procurement cycle with more visible public ambition.

Image © Mordor Intelligence. Reuse requires attribution under CC BY 4.0.

Competitive Landscape

The AI GPU chip market remains highly concentrated at the top, and NVIDIA held approximately 80-85% of data center AI accelerator revenue in 2026. That position reflects a durable advantage in software compatibility, proprietary interconnects, and the ability to sell tightly integrated systems rather than stand-alone c

hips. AMD remained the main merchant alternative with approximately 5-7% share, which means the competitive field is still narrow despite rising buyer interest in second-source options. The result is a market where most large customers still buy into a system ecosystem first and a chip vendor second.

Strategic moves in 2026 showed that the AI GPU chip market is now being contested through platform partnerships as much as through silicon launches. AMD strengthened its position in February 2026 through a multi-year, multi-generation 6-gigawatt deployment partnership with Meta, which gives AMD scale, validation, and a direct place in one of the largest AI infrastructure build-outs. OpenAI and Broadcom then introduced the Jalapeño Intelligence Processor in April 2026, which signaled that some of the highest-volume inference workloads are moving toward custom silicon where efficiency matters at fleet scale. NVIDIA also continued to widen its reach across deployment layers with new client AI hardware in 2026, which helps protect its software base even outside large data center racks. These moves show that competition is broadening, but it is broadening around ecosystems, supply access, and workload specialization rather than around price alone.

Another important shift in the AI GPU chip market is that custom accelerators are taking a larger share of internal hyperscaler workloads, especially in inference. That does not remove demand for merchant GPUs, but it changes the addressable opportunity because not all AI spending now flows through the same procurement channel. China is also developing a more independent accelerator base, and domestic suppliers led by Huawei's Ascend line increased their presence in 2025 as local procurement diversified away from a single foreign vendor. Even with that shift, sovereign and enterprise buyers still value flexibility for multi-workload clusters, and that continues to favor GPU platforms with mature software support over narrower fixed-purpose alternatives.

AI GPU Chip Industry Leaders* NVIDIA Corporation

- * Advanced Micro Devices, Inc.
- * Intel Corporation
- * Moore Threads Technology Co., Ltd.
- * Biren Technology Co., Ltd.

* *Disclaimer: Major Players sorted in no particular order

Image © Mordor Intelligence. Reuse requires attribution under CC BY 4.0. AI GPU Chip Market Companies Covered in this Report* NVIDIA Corporation

- * Advanced Micro Devices, Inc.
- * Intel Corporation
- * Moore Threads Technology Co., Ltd.
- * Biren Technology Co., Ltd.
- * Huawei Technologies Co., Ltd.
- * Google LLC
- * Amazon.com, Inc.
- * Microsoft Corporation
- * Alibaba Group Holding Limited
- * Meta Platforms, Inc.
- * Cerebras Systems, Inc.
- * Graphcore Limited
- * SambaNova Systems, Inc.
- * Qualcomm Incorporated
- * Tenstorrent Inc.
- * Shanghai Iluvatar CoreX Semiconductor Co., Ltd.
- * Shanghai Denglin Technology Co., Ltd.
- * Hygon Information Technology Co., Ltd.
- * Glenfly Tech Co., Ltd.

Recent Industry Developments in AI GPU Chip Market* June 2026: NVIDIA confirmed that the Vera Rubin platform has entered full production, with first systems scheduled to ship to customers in the second half of 2026, as announced at the COMPUTEX keynote on June 1. Simultaneously, NVIDIA unveiled RTX Spark, an ARM-based, Windows-on-chip system-on-chip featuring a 6,144-CUDA-core Blackwell GPU and 20-core MediaTek Grace CPU, entering the client AI GPU market with Microsoft Surface Laptop Ultra and partners including Dell, HP, ASUS, Lenovo, and MSI.

* April 2026: OpenAI and Broadcom unveiled the Jalapeño Intelligence Processor, OpenAI's first custom AI accelerator architected for LLM inference, co-developed from initial design to manufacturing tape-out in nine months. Jalapeño is targeted for initial deployment by the end of 2026 and represents the first step in a multi-generation compute platform built with Broadcom silicon and Celestica rack manufacturing expertise.

* April 2026: The UALink Consortium ratified its next-generation specification, adding In-Network Compute, Chiplet Definition, and Manageability capabilities to the 200G UALink standard. UALink 1.0 hardware from AMD, Intel, and Astera Labs is targeted for late 2026 deployment, supporting up to 1,024 accelerators in a single fabric, scaling beyond NVLink's 576-GPU ceiling.

* February 2026: AMD and Meta announced a multi-year, multi-generation 6-gigawatt partnership to deploy AMD Instinct GPUs across Meta's AI data centers, with the first gigawatt of custom MI450-based systems scheduled for shipment in the second half of 2026 under the AMD Helios rack-scale architecture. Meta received a warrant to acquire up to 160 million AMD shares tied to deployment milestones.

Table of Contents for AI GPU Chip Industry Report1. INTRODUCTION

- * 1.1 Study Assumptions and Market Definition
- * 1.2 Scope of the Study
- 2. RESEARCH METHODOLOGY
- 3. EXECUTIVE SUMMARY
- 4. MARKET LANDSCAPE
 - * 4.1 Market Overview
 - * 4.2 Market Drivers* 4.2.1 Expanding Enterprise Fine-Tuning of Proprietary Models
 - * 4.2.2 Rapid Scale-Up of Hyperscale AI Training Clusters
 - * 4.2.3 HBM4 Readiness and Advanced Packaging Upgrade Cycle
 - * 4.2.4 Sovereign AI Procurement and Domestic Compute Security
 - * 4.2.5 NVLink-CXL and UALink Pooling of Accelerator Capacity
 - * 4.2.6 Liquid Cooling Standardization for High-TDP GPU Racks
 - * 4.3 Market Restraints* 4.3.1 Advanced Packaging Capacity Bottlenecks
 - * 4.3.2 Rising Total Cost of Ownership for Cluster-Scale Deployments
 - * 4.3.3 Export Controls and Geopolitical Supply Friction
 - * 4.3.4 Competition From Custom ASICs and Proprietary Accelerators
 - * 4.4 Industry Value Chain Analysis
 - * 4.5 Regulatory Landscape
 - * 4.6 Technological Outlook
 - * 4.7 Impact of Macroeconomic Factors on the Market
 - * 4.8 Porter's Five Forces Analysis* 4.8.1 Bargaining Power of Suppliers
 - * 4.8.2 Bargaining Power of Buyers
 - * 4.8.3 Threat of New Entrants
 - * 4.8.4 Threat of Substitutes
 - * 4.8.5 Industry Rivalry

5. MARKET SIZE AND GROWTH FORECASTS (VALUE)

- * 5.1 By Product Type* 5.1.1 Data Center AI GPUs
- * 5.1.2 Edge AI GPUs
- * 5.1.3 Client AI GPUs
- * 5.2 By Compute Function* 5.2.1 Training GPUs
- * 5.2.2 Inference GPUs
- * 5.2.3 Mixed Training and Inference GPUs
- * 5.3 By Deployment Environment* 5.3.1 Hyperscale and Cloud
- * 5.3.2 Enterprise Data Centers
- * 5.3.3 Government and Research Institutions
- * 5.3.4 Edge and Endpoint Deployments
- * 5.4 By Workload* 5.4.1 Generative AI and Large Language Models
- * 5.4.2 Computer Vision and Robotics
- * 5.4.3 Speech and Natural Language Processing
- * 5.4.4 Recommendation, Search, and Graph Analytics
- * 5.4.5 Scientific Computing and Other AI Workloads
- * 5.5 By Geography* 5.5.1 North America
- * 5.5.1.1 United States
- * 5.5.1.2 Canada
- * 5.5.1.3 Mexico
- * 5.5.2 Europe
- * 5.5.2.1 Germany
- * 5.5.2.2 United Kingdom
- * 5.5.2.3 France
- * 5.5.2.4 Italy
- * 5.5.2.5 Rest of Europe
- * 5.5.3 Asia-Pacific
- * 5.5.3.1 China
- * 5.5.3.2 Japan
- * 5.5.3.3 South Korea
- * 5.5.3.4 India
- * 5.5.3.5 Southeast Asia
- * 5.5.3.6 Rest of Asia-Pacific
- * 5.5.4 South America
- * 5.5.5 Middle East and Africa

6. COMPETITIVE LANDSCAPE

- * 6.1 Market Concentration
- * 6.2 Strategic Moves
- * 6.3 Market Positioning Analysis
- * 6.4 Company Profiles (includes Global Level Overview, Market Level Overview, Core Segments, Financials as available, Strategic Information, Market Rank/Share, Products and Services, Recent Developments)* 6.4.1 NVIDIA Corporation
- * 6.4.2 Advanced Micro Devices, Inc.
- * 6.4.3 Intel Corporation
- * 6.4.4 Moore Threads Technology Co., Ltd.
- * 6.4.5 Biren Technology Co., Ltd.
- * 6.4.6 Huawei Technologies Co., Ltd.
- * 6.4.7 Google LLC
- * 6.4.8 Amazon.com, Inc.
- * 6.4.9 Microsoft Corporation
- * 6.4.10 Alibaba Group Holding Limited
- * 6.4.11 Meta Platforms, Inc.
- * 6.4.12 Cerebras Systems, Inc.
- * 6.4.13 Graphcore Limited
- * 6.4.14 SambaNova Systems, Inc.

- * 6.4.15 Qualcomm Incorporated
- * 6.4.16 Tenstorrent Inc.
- * 6.4.17 Shanghai Iluvatar CoreX Semiconductor Co., Ltd.
- * 6.4.18 Shanghai Denglin Technology Co., Ltd.
- * 6.4.19 Hygon Information Technology Co., Ltd.
- * 6.4.20 Glenfly Tech Co., Ltd.

7. MARKET OPPORTUNITIES AND FUTURE OUTLOOK

* 7.1 White Space and Unmet-Need Assessment

Global AI GPU Chip Market Report ScopeThe Global AI GPU Chip Market encompasses the worldwide industry dedicated to the design, manufacturing, and distribution of specialized graphics processing units optimized for artificial intelligence workloads, including deep learning, machine learning, natural language processing, and data analytics.

The AI GPU chip market Report is Segmented by Product Type (Data Center AI GPUs, Edge AI GPUs, and Client AI GPUs), Compute Function (Training GPUs, Inference GPUs, and Mixed Training and Inference GPUs), Deployment Environment (Hyperscale and Cloud, Enterprise Data Centers, Government and Research Institutions, and Edge and Endpoint Deployments), Workload (Generative AI and Large Language Models, Computer Vision and Robotics, Speech and Natural Language Processing, Recommendation, Search, and Graph Analytics, and Scientific Computing and Other AI Workloads), and Geography (North America, Europe, Asia-Pacific, South America, and Middle East and Africa). The Market Forecasts are Provided in Terms of Value (USD).

By Product TypeData Center AI GPUs |
 Edge AI GPUs |
 Client AI GPUs |
 By Compute FunctionTraining GPUs |
 Inference GPUs |
 Mixed Training and Inference GPUs |
 By Deployment EnvironmentHyperscale and Cloud |
 Enterprise Data Centers |
 Government and Research Institutions |
 Edge and Endpoint Deployments |
 By WorkloadGenerative AI and Large Language Models |
 Computer Vision and Robotics |
 Speech and Natural Language Processing |
 Recommendation, Search, and Graph Analytics |
 Scientific Computing and Other AI Workloads |
 By GeographyNorth America | United States |
 | Canada |
 | Mexico |
 Europe | Germany |
 | United Kingdom |
 | France |
 | Italy |
 | Rest of Europe |
 Asia-Pacific | China |
 | Japan |
 | South Korea |
 | India |
 | Southeast Asia |
 | Rest of Asia-Pacific |
 South America |
 Middle East and Africa |
 By Product Type | Data Center AI GPUs |
 | Edge AI GPUs |

- | Client AI GPUs |
- By Compute Function | Training GPUs |
 - | Inference GPUs |
 - | Mixed Training and Inference GPUs |
- By Deployment Environment | Hyperscale and Cloud |
 - | Enterprise Data Centers |
 - | Government and Research Institutions |
 - | Edge and Endpoint Deployments |
- By Workload | Generative AI and Large Language Models |
 - | Computer Vision and Robotics |
 - | Speech and Natural Language Processing |
 - | Recommendation, Search, and Graph Analytics |
 - | Scientific Computing and Other AI Workloads |
- By Geography | North America | United States |
 - | | Canada |
 - | | Mexico |
 - | | Europe | Germany |
 - | | United Kingdom |
 - | | France |
 - | | Italy |
 - | | Rest of Europe |
 - | | Asia-Pacific | China |
 - | | Japan |
 - | | South Korea |
 - | | India |
 - | | Southeast Asia |
 - | | Rest of Asia-Pacific |
 - | | South America |
 - | | Middle East and Africa |

Key Questions Answered in the Report What is the current and forecast value of the AI GPU chip market? The AI GPU chip market size stood at USD 167.5 billion in 2025, reached USD 288.4 billion in 2026, and is forecast to reach USD 621.7 billion by 2031 at a 16.60% CAGR.

Which product type leads AI GPU demand today? Data center AI GPUs led the market with 93.11% share in 2025 because the most advanced systems still depend on centralized racks, dense networking, and specialized cooling.

What part of the workload mix is growing fastest? Computer vision and robotics is the fastest-growing workload category, with a projected 17.32% CAGR through 2031, as physical AI, automation, and localized inference expand.

Why are governments becoming major buyers of AI GPU systems? Public buyers are treating compute capacity as a strategic asset tied to sovereignty, resilience, and domestic AI capability, which is why government and research deployments are projected to grow at 17.73% through 2031.

Which region leads today and which region is growing fastest? North America held 38.44% share in 2025, while the Middle East and Africa is projected to record the fastest regional CAGR at 17.42% through 2031.

How are custom AI chips affecting GPU vendors? Custom accelerators are taking a larger share of internal hyperscaler inference work, which narrows the merchant GPU opportunity in some workloads even though GPUs remain preferred for flexible multi-workload clusters.

Related Reports Explore More ReportsAI Inference GPU Market ShareEdge AI GPU Market ShareGenerative AI GPU Market ShareMarket Report on AI ChipsetsIndustrial GPU Market GrowthGPU Software Market ForecastsIntegrated GPU Market TrendsScientific Computing GPU Market SizeMarket Data for GPU Programming PlatformGPU In Robotics and Smart Manufacturing Market Share