

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

Industry: Semiconductors & Hardware

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

GPU Semiconductor Market Size and ShareMarket Overview

Study Period | 2020 - 2031 |

Market Size (2026) | USD 228.61 Billion |

Market Size (2031) | USD 647.34 Billion |

Growth Rate (2026 - 2031) | 23.14 % |

Fastest Growing Market | Asia-Pacific |

Largest Market | North America |

Market Concentration | Medium |

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.

GPU Semiconductor Market Analysis by Mordor IntelligenceThe GPU semiconductor market size is expected to increase from USD 189.43 billion in 2025 to USD 228.61 billion in 2026 and reach USD 647.34 billion by 2031, growing at a CAGR of 23.14 % over 2026-2031. The GPU semiconductor market is expanding because AI training and inference workloads now require accelerated computing at a scale that conventional CPU-led systems cannot support efficiently. The spending cycle is also becoming more durable because hyperscalers, enterprises, and public sector buyers increasingly treat GPU capacity as core digital infrastructure rather than discretionary hardware. Supply discipline remains just as important as demand, since advanced packaging and high-bandwidth memory availability continue to shape shipment timing, pricing, and vendor positioning across the GPU semiconductor market. Competitive intensity is rising as NVIDIA and AMD defend the high-performance tier while Intel, automotive platform vendors, and regional suppliers target selective openings in inference, edge, and embedded workloads. The GPU semiconductor market also has room to widen beyond hyperscaler demand as sovereign AI build-outs, enterprise AI factories, automotive compute adoption, and on-device AI use cases create additional procurement channels.

Key Report Takeaways

- * By integration type, discrete GPUs accounted for 84.31% of revenue in 2025 and are projected to grow at a CAGR of 24.37% through 2031.
- * By device application, data center and server accelerators accounted for 73.52 % of the GPU semiconductor market size in 2025 and are projected to expand at a 24.35% CAGR through 2031.
- * By end user, cloud, data center, and enterprise IT accounted for 76.39% of the GPU semiconductor market share in 2025, and are expected to grow at a CAGR of 24.64% through 2031.
- * By memory type, HBM-based GPUs led with 66.54% share in 2025, and are projected to expand at a 24.16% CAGR through 2031.
- * By geography, North America led with 49.51% share in 2025, while Asia-Pacific is projected to post the fastest CAGR at 24.57% 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.

Global GPU Semiconductor Market Trends and Insights Drivers Impact Analysis*

Driver | (~) % Impact on CAGR Forecast | Geographic Relevance | Impact Timeline

Hyperscale AI Training and Inference Cluster Expansion | +7.5% | Global, North America, and the Asia-Pacific core | Short term (≤ 2 years) |
Enterprise AI Factory and Sovereign Compute Procurement | +3.8% | Global, Middle East, and South Asia are emerging | Medium term (2-4 years) |
Edge AI Upgrade Cycle in PCs and Mobile Devices | +2.3% | North America, Europe, Asia-Pacific | Short term (≤ 2 years) |
Rising ADAS and In-Cabin Compute Content per Vehicle | +1.8% | China, Europe, North America | Medium term (2-4 years) |
Chiplet-Based GPU Roadmaps Improving Yield and Product Scaling | +1.2% | Global, Asia-Pacific with TSMC-led ecosystem | Medium term (2-4 years) |
GPU-as-a-Service Broadening Access Beyond Hyperscalers | +1.4% | Global, North America leads, Asia-Pacific is growing | Short term (≤ 2 years) |

Source: Mordor Intelligence |

Hyperscale AI Training and Inference Cluster Expansion

The GPU semiconductor market is being pushed higher by the decision of the largest cloud operators to treat GPU capacity as foundational infrastructure. The combined capital expenditure of Amazon, Alphabet, Meta, Microsoft, and Oracle approached USD 600 billion in 2026, with a large share directed toward AI infrastructure build-outs. NVIDIA reported Q1 FY2027 data center revenue of USD 75.2 billion in May 2026, up 92% year over year, which shows how quickly this procurement cycle translated into supplier revenue. The Semiconductor Industry Association also showed that AI data center semiconductor revenue reached USD 670 billion in 2026, confirming that AI infrastructure has become the largest demand engine in the broader chip stack. The same source indicated that inference workloads are expected to grow faster than training workloads, suggesting the GPU semiconductor market should remain robust even after the first wave of large-model training clusters matures. Each leading-edge AI rack carries a very high system cost, so even a modest expansion in hyperscale capacity has an outsized effect on GPU volumes, mix, and pricing across the GPU semiconductor market.

Enterprise AI Factory and Sovereign Compute Procurement

The GPU semiconductor market is also gaining support from enterprise AI factories and sovereign compute programs that follow policy goals rather than short-cycle return thresholds. The source text shows that this demand is tied to data sovereignty, national industrial planning, and the need to keep sensitive model development under domestic control. NVIDIA stated, in the context of the Sovereign AI Index cited in the draft, that its platforms account for a large share of tracked sovereign infrastructure projects, suggesting long procurement cycles once a national platform is selected. That pattern matters because governments and regulated institutions often buy premium systems even as commercial buyers slow down decision-making, helping keep a pricing floor in the GPU semiconductor market. The result is a broader demand base that is less dependent on the timing of hyperscaler ordering waves. It also gives leading vendors a chance to build durable relationships around hardware, software stacks, and ongoing support, which can lock in future replacement cycles across the GPU semiconductor market.

Edge AI Upgrade Cycle in PCs and Mobile Devices

The GPU semiconductor market is receiving another lift from the upgrade cycle in PCs and mobile devices as AI features move from premium systems into mainstream products. Microsoft extended Windows App SDK 2.2 AI support to NVIDIA GeForce RTX 3000-series and newer discrete GPUs in 2026, widening the installed base that can run local AI workloads without relying solely on new NPU-certified systems. The draft also shows why this matters, since local large language model inference depends heavily on memory bandwidth, and discrete GPUs still provide a large practical advantage over many NPUs for that workload. That advantage gives enter

prises and prosumer users a reason to refresh their systems with GPUs rather than wait for fully standardized AI PC platforms. On mobile devices, integrated GPU performance per watt remains an important differentiator for on-device AI use cases, so the same AI shift is reinforcing new SoC design cycles. The combined effect is that the GPU semiconductor market benefits from AI spending at both centralized data-center and distributed client-device scales.

Rising ADAS and In-Cabin Compute Content per Vehicle

The GPU semiconductor market is also being supported by the steady rise in compute content per vehicle, especially as ADAS and in-cabin functions move onto unified platforms. NVIDIA reported that DRIVE Orin had exceeded 1 million vehicle deployments at BYD by mid-2025, confirming that GPU-class automotive compute had already moved into mass-volume production. NVIDIA also stated that its automotive revenue reached USD 1.7 billion in FY2025, up 55% from the prior year, with Thor-based design wins across BYD, XPENG, Li Auto, Zeekr, Mercedes-Benz, and Volvo. The importance of the GPU semiconductor market is that automotive programs typically run on multi-year product cycles, which gives vendors better revenue visibility once a platform is designed into a vehicle lineup. Compliance barriers also matter, since safety certification makes it costly for automakers to switch suppliers mid-cycle. That raises the value of proven platforms and supports long-duration demand for certified GPU compute in automotive applications.

Restraints Impact Analysis*

Restraint	(~) % Impact on CAGR Forecast	Geographic Relevance	Impact Timeline
Export Controls and Tariff Volatility	-3.2%	US-China bilateral, global spill over	Short term (≤ 2 years)
Elevated GPU and Memory ASPs Slowing Mainstream Adoption	-2.0%	Global	Medium term (2-4 years)
HBM and CoWoS Allocation Bias Toward AI Racks	-1.3%	Global, bottleneck centered in Taiwan	Short term (≤ 2 years)
Grid Interconnection Delays for High-Density GPU Campuses	-0.9%	North America and Europe	Medium term (2-4 years)

Source: Mordor Intelligence | Export Controls and Tariff Volatility

The GPU semiconductor market faces a clear restraint from export control changes and tariff actions that affect pricing, compliance, and deal timing. The US Bureau of Industry and Security revised its license review policy for advanced semiconductors exported to China in January 2026, moving certain reviews to a case-by-case process subject to end-user certification and verification.[1]Bureau of Industry and Security, U.S. Department of Commerce, "Department of Commerce Revises License Review Policy for Semiconductors Exported to China," Bureau of Industry and Security, bis.gov BIS then clarified in May 2026 that license requirements for advanced computing items still apply to entities headquartered in Country Group D:5, regardless of their physical location. That keeps a structural limit on how freely high-end suppliers can address some overseas demand pools. The draft also notes that tariffs added a direct cost to cross-border AI chip transactions, complicating procurement planning for distributors and cloud operators. Together, these measures slow decisions, raise documentation burdens, and make revenue conversion less predictable across the GPU semiconductor market.

Elevated GPU And Memory ASPs Slowing Mainstream Adoption

The GPU semiconductor market also faces pressure from elevated average selling prices that are no longer driven mainly by silicon economics. The source text shows that HBM memory and CoWoS advanced packaging account for a large share of the bill of materials for a leading-edge AI GPU, thereby creating a high-cost floor. Advanced packaging has therefore become a value center in its own right rather than a pass-through manufacturing step. That matters because many industrial, r

obotics, and telecom buyers want AI acceleration but still operate within tighter capital budgets than hyperscalers. In that setting, high system costs can delay replacement cycles and keep narrower FPGA or ASIC options in use for longer. The GPU semiconductor market remains strong at the top end, but broader adoption in budget-constrained verticals can still slow when platform costs stay elevated for multiple procurement cycles.

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

Segment Analysis By Integration Type: Discrete GPUs Anchor AI Compute Architecture

Discrete GPUs accounted for 84.31% of revenue in the GPU semiconductor market in 2025, underscoring how AI computing needs now shape the overall mix. That share reflects the simple fact that integrated graphics still cannot provide the memory bandwidth, dedicated VRAM capacity, and floating-point throughput needed for large-scale model training and inference. The Semiconductor Industry Association reported that AI accelerators account for most of the semiconductor value in AI server racks, which helps explain why discrete designs hold such a large share of the GPU semiconductor market. Product launches across NVIDIA and AMD have reinforced that position because buyers can adopt new hardware while keeping much of the same software environment and deployment logic. That continuity matters in the GPU semiconductor industry because replacement decisions now depend as much on ecosystem stability as on raw performance gains.

Intel's Crescent Island launch at Computex 2026 showed that the discrete data center segment remains open enough to attract major entrants with server compute relationships and packaging capabilities. At the same time, integrated GPUs remain strategically relevant in mobile devices, notebooks, display-centric automotive systems, and lighter edge inference tasks where power efficiency matters more than peak throughput. The draft also highlights the role of chiplet-based architecture, especially in AMD's Instinct roadmap, because it improves yield economics for large compute designs and allows suppliers to scale products without relying on a single monolithic die. That architecture supports broader product planning because one design base can be paired with different memory approaches for AI, gaming, and professional use cases. The result is a GPU semiconductor market where discrete GPUs keep control of high-value AI workloads, while integrated designs retain an important role in high-volume client and embedded deployments.

Image © Mordor Intelligence. Reuse requires attribution under CC BY 4.0. By Device Application: Data Center Accelerators Dominate Silicon Spending

Data center and server accelerators held a 73.52% share of the GPU semiconductor market in 2025 and are projected to expand at a 24.35% CAGR through 2031. The Semiconductor Industry Association reported that AI data center semiconductor revenue reached USD 670 billion in 2026, underscoring the centrality of this application to the broader GPU semiconductor market. The same report also shows that AI accelerators capture the majority of semiconductor content value in AI server racks, so application mix now follows infrastructure investment more than consumer unit shipments. Another important shift is that inference demand is expected to grow faster than training demand, which should keep installed platforms productive for longer while still supporting new deployments. That balance is useful because it supports continuing volume demand even if per-rack pricing becomes less aggressive over time.

Other device applications still matter because they spread GPU demand across several usage models and replacement cycles. PCs and workstations are benefiting from local AI inference needs, and Microsoft's 2026 software support expansion for

discrete RTX systems widened the refresh path for installed users. Automotive and ADAS remains the fastest-moving non-data-center application in the draft because higher compute content is moving into larger vehicle segments. Embedded and edge devices also provide a meaningful layer of demand because GPU IP licensing and modular platform design let chipmakers integrate graphics and AI capability without building full custom stacks from scratch. This leaves the GPU semiconductor market with a clear center of gravity in data centers, but not a single-point dependence on one application alone.

By End User: Enterprise Cloud Spending Steers GPU Procurement

Cloud, data center, and enterprise IT held 76.39% of the GPU semiconductor market share in 2025, and this segment is projected to expand at a 24.64% CAGR through 2031. That mix confirms that institutional AI deployment has replaced gaming as the main revenue and procurement anchor in the GPU semiconductor market. NVIDIA's FY2026 revenue reached USD 215.9 billion, with USD 193.7 billion from data centers, underscoring how enterprise and hyperscaler spending now strongly shapes supplier outcomes.^[2] NVIDIA Corporation, "NVIDIA Announces Financial Results for Fourth Quarter and Fiscal Year 2026," U.S. Securities and Exchange Commission archive, sec.gov AMD also reported USD 5.8 billion in Q1 2026 data center revenue, showing that large buyers are actively funding a second high-performance supply path rather than relying on a single vendor. This end-user profile matters because infrastructure buyers typically purchase not only chips, but also complete system platforms, software compatibility, and long-term service commitments.

The remaining end-user groups still shape the depth and resilience of the GPU semiconductor market. Consumer electronics and gaming continue to support high-volume GDDR-oriented design cycles, even though their revenue weight has fallen relative to data center GPUs. Automotive buyers work from a smaller base but often remain locked into a platform for the life of a vehicle program after certification. Telecom, industrial, robotics, healthcare imaging, scientific simulation, and other specialized users add a layer of steady procurement that is less exposed to consumer demand swings. This means the GPU semiconductor market remains led by enterprise-scale compute buyers, while several smaller end-user channels help maintain breadth across the cycle.

Image © Mordor Intelligence. Reuse requires attribution under CC BY 4.0. By Memory Type: HBM Architecture Redefines AI GPU Performance Boundaries

HBM-based GPUs held 66.54% share of the GPU semiconductor market size in 2025 and are projected to expand at a 24.16% CAGR through 2031. The draft makes clear that this position comes from the memory bandwidth demands of large-model AI workloads, which have pushed new premium GPU generations toward HBM architectures. That shift is important because memory choice is no longer a secondary design decision in the GPU semiconductor market. It now determines which platforms can support frontier training and high-throughput inference at commercial scale. In practice, HBM has become part of the performance boundary for premium AI platforms rather than a component that can be swapped without altering the system profile.

GDDR-based GPUs still play a significant role in gaming, professional visualization, PCs, and automotive display applications, where cost efficiency and power limits matter more than absolute bandwidth. Shared DDR and LPDDR memory configurations remain important in smartphones, tablets, and entry-level notebooks, where unit volumes are high even though revenue per unit is lower. The input also notes that standards compliance for HBM4 and future generations acts as an entry barrier for memory suppliers seeking access to the next wave of premium AI GPU programs. That creates another layer of concentration in the supply chain because only qualified vendors can participate in the most advanced platforms. As a result, the GPU semiconductor market continues to widen by unit volume across several memory types, but the highest-value growth remains tightly linked to HBM availability and qualification.

Geography Analysis

North America held a 49.51% share of the GPU semiconductor market in 2025, keeping the region in the lead entering 2026. The main reason is that the largest hyperscaler buyers are headquartered in the United States, and their AI infrastructure budgets continue to drive demand for GPUs in the semiconductor market. The draft states that Amazon, Microsoft, Alphabet, Meta, and Oracle were on track to spend more than USD 600 billion on AI infrastructure in 2026, which helps explain why regional demand remained concentrated even when data center build-outs were geographically distributed. NVIDIA's FY2026 revenue and AMD's Q1 2026 data center performance both reflect that spending pattern, since North American buyers remain central to high-end procurement and product qualification. Canada is also drawing more AI compute investment through policy support, while Mexico is gaining relevance as a secondary data center location tied to nearshoring and regional diversification. Together, these factors keep North America at the core of demand, platform testing, and early deployment activity in the GPU semiconductor market.

Asia-Pacific is projected to expand at a 24.57% CAGR through 2031, making it the fastest-growing regional layer in the GPU semiconductor market. The region is growing through a different mix of drivers, since it combines end demand, memory supply, packaging capacity, equipment investment, and sovereign compute programs in one broad ecosystem. South Korea remains especially important because Samsung Electronics and SK Hynix anchor HBM production, which makes the region central to the premium AI supply chain even when final system demand is located elsewhere. Japan also supports the GPU semiconductor market upstream, and the Semiconductor Equipment Association of Japan forecast that the country's semiconductor and FPD manufacturing equipment market would reach JPY 5,500.4 billion (USD 36.7 billion) in FY2026.[3] Semiconductor Equipment Association of Japan, "Semiconductor and FPD Manufacturing Equipment Forecast, January 2026," Semiconductor Equipment Association of Japan, seaj.or.jp That rise reflects AI-linked investment in advanced logic and memory capacity, which means GPU demand is spreading back through tools, materials, and production infrastructure.

Europe's role in the GPU semiconductor market is being supported by compliance, data residency, and the preference for locally operated AI compute in regulated use cases. Germany, the United Kingdom, and France remain the largest demand pools in the region, and domestic cloud operators are scaling local GPU capacity in response to enterprise requirements. The region's demand profile is less tied to consumer replacement cycles and more tied to trusted deployment conditions for finance, healthcare, and public sector workloads. Outside Europe, South America and the Middle East and Africa remain smaller in share but important in strategic terms because energy availability, digital infrastructure expansion, and sovereign technology priorities are creating new pockets of premium AI compute demand. This leaves the GPU semiconductor market with a regional structure where North America leads in current demand, Asia-Pacific grows fastest through supply and deployment depth, and other regions expand where regulation, infrastructure, or state-backed programs create clear buying triggers.

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

Competitive Landscape

The GPU semiconductor market remains moderately concentrated at the premium AI data center tier. NVIDIA still holds the clearest leadership position because its hardware is reinforced by CUDA, TensorRT, and NeMo, which raise switching costs for developers and enterprise users. NVIDIA's FY2026 revenue reached USD 215.9 billion, with data center contributing USD 193.7 billion, demonstrating the comp

any's strong position at the highest-value end of the GPU semiconductor market. The company also used 2026 to extend that lead through the Blackwell ramp, the Vera Rubin platform, and RTX Spark systems for Windows AI PCs. Those moves show a strategy that links data centers, scientific computing, networking, and client-side AI under a single platform roadmap.

AMD has become the primary challenger in the GPU semiconductor market by building a steady annual cadence around the Instinct family rather than responding only at long intervals. AMD reported USD 5.8 billion in Q1 2026 data center revenue, up 57% year over year, and tied that result to deployment momentum for Instinct accelerators in AI training and inference. That gives hyperscalers and enterprise buyers a real second source at the high-performance tier, which matters in a market shaped by allocation risk and long lead times. Intel is pursuing a different opening by targeting inference workloads and standard air-cooled deployments with Crescent Island, while also using EMIB packaging to challenge the concentration around CoWoS-based system approaches.[4]Intel Corporation, "Intel Announces New AI Innovations at Computex, Chip to Rackscale AI Solutions Delivered to Customers," Intel Corporation, [intc.com](https://www.intel.com) These strategies do not erase NVIDIA's lead, but they do widen the field enough to make the GPU semiconductor market more competitive than a single-vendor reading would suggest.

The GPU semiconductor market also remains broad because not every opportunity sits in hyperscaler-scale AI clusters. Mobile, embedded, industrial vision, medical imaging, and automotive edge inference all require different balances of cost, power, certification, and latency. That creates room for suppliers with narrower architectures, licensing models, or platform specialization even if they do not challenge the top tier of AI training systems. It also means that compliance standards, software compatibility, and packaging access can matter as much as peak benchmark performance when buyers choose a vendor. The result is a competitive structure where premium AI data center revenue is concentrated, but the full GPU semiconductor market still contains several layers of differentiated competition.

GPU Semiconductor Industry Leaders* NVIDIA Corporation

* Advanced Micro Devices, Inc.

* Intel Corporation

* Qualcomm Incorporated

* Samsung Electronics Co., Ltd.

* *Disclaimer: Major Players sorted in no particular order

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

Recent Industry Developments* June 2026: NVIDIA unveiled RTX Spark at GTC Taipei in partnership with Microsoft, a superchip integrating a Blackwell RTX GPU with a 20-core NVIDIA Grace CPU via NVLink chip-to-chip interconnect, delivering 1 petaflop of AI performance and up to 128 GB of unified memory for Windows AI PC applications. Systems from ASUS, Dell, HP, Lenovo, and Microsoft Surface are planned for fall 2026 availability.

* May 2026: NVIDIA reported record Q1 FY2027 data center revenue of USD 75.2 billion, up 92% year over year, driven by the Blackwell and Vera Rubin platform ramps and expansion of AI networking solutions. NVIDIA's board authorized an additional USD 80 billion share repurchase and raised the quarterly cash dividend from USD 0.01 to USD 0.25 per share.

* May 2026: AMD reported Q1 2026 data center segment revenue of USD 5.8 billion, up 57% year over year, with OEMs and ODMs deploying Instinct MI350X GPUs at scale for AI training and inference. AMD guided Q2 2026 total revenue to USD 11.2 b

illion, which represented 46% year over year growth, and outlined the Instinct M I400 Helios rack architecture for large-scale AI deployments.
* January 2026: The US Bureau of Industry and Security issued a final rule effective January 15, 2026, shifting the export license review policy for NVIDIA H200 and AMD MI325X-equivalent chips from presumption of denial to case-by-case review for exports to China and Macau, subject to end-user certification and performance verification requirements. The White House simultaneously imposed a 25% tariff on advanced AI chips under Section 232, adding a cost layer to advanced GPU cross-border transactions.

Table of Contents for GPU Semiconductor 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 Hyperscale AI Training and Inference Cluster Expansion
 - * 4.2.2 Enterprise AI Factory and Sovereign Compute Procurement
 - * 4.2.3 Edge AI Upgrade Cycle in PCs and Mobile Devices
 - * 4.2.4 Rising ADAS and In-Cabin Compute Content per Vehicle
 - * 4.2.5 Chiplet-Based GPU Roadmaps Improving Yield and Product Scaling
 - * 4.2.6 GPU-as-a-Service Broadening Access Beyond Hyperscalers
 - * 4.3 Market Restraints* 4.3.1 Export Controls and Tariff Volatility
 - * 4.3.2 Elevated GPU and Memory ASPs Slowing Mainstream Adoption
 - * 4.3.3 HBM and CoWoS Allocation Bias Toward AI Racks
 - * 4.3.4 Grid Interconnection Delays for High-Density GPU Campuses
 - * 4.4 Industry Value Chain Analysis
 - * 4.5 Regulatory Landscape
 - * 4.6 Technological Outlook
 - * 4.7 Porter's Five Forces Analysis* 4.7.1 Bargaining Power of Suppliers
 - * 4.7.2 Bargaining Power of Buyers
 - * 4.7.3 Threat of New Entrants
 - * 4.7.4 Threat of Substitutes
 - * 4.7.5 Competitive Rivalry
- 5. MARKET SIZE AND GROWTH FORECASTS (VALUE)
 - * 5.1 By Integration Type* 5.1.1 Integrated GPUs
 - * 5.1.2 Discrete GPUs
 - * 5.2 By Device Application* 5.2.1 Mobile Devices and Tablets
 - * 5.2.2 PCs and Workstations
 - * 5.2.3 Data Center and Server Accelerators
 - * 5.2.4 Gaming Consoles and Handheld Devices
 - * 5.2.5 Automotive and ADAS
 - * 5.2.6 Embedded and Edge Devices
 - * 5.3 By End User* 5.3.1 Consumer Electronics and Gaming
 - * 5.3.2 Cloud, Data Center, and Enterprise IT
 - * 5.3.3 Telecom
 - * 5.3.4 Automotive
 - * 5.3.5 Industrial and Robotics

- * 5.3.6 Other End Users
- * 5.4 By Memory Type* 5.4.1 GDDR-Based GPUs
- * 5.4.2 HBM-Based GPUs
- * 5.4.3 Shared DDR/LPDDR Memory GPUs
- * 5.4.4 Other Memory Types

- * 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 Vendor 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 Qualcomm Incorporated
- * 6.4.5 Samsung Electronics Co., Ltd.
- * 6.4.6 MediaTek Inc.
- * 6.4.7 Imagination Technologies Group Ltd.
- * 6.4.8 Huawei Technologies Co., Ltd.
- * 6.4.9 Moore Threads Intelligent Technology (Beijing) Co., Ltd.
- * 6.4.10 Biren Technology Co., Ltd.
- * 6.4.11 VeriSilicon Microelectronics (Shanghai) Co., Ltd.
- * 6.4.12 Zhaoxin Semiconductor Co., Ltd.
- * 6.4.13 UNISOC Technologies Co., Ltd.
- * 6.4.14 Rockchip Electronics Co., Ltd.
- * 6.4.15 Bolt Graphics, Inc.

7. MARKET OPPORTUNITIES AND FUTURE OUTLOOK

- * 7.1 White-Space and Unmet-Need Assessment

Global GPU Semiconductor Market Report ScopeThe GPU semiconductor market refers to the market for semiconductor components and integrated circuits designed to perform graphics processing, parallel computing, and accelerated computing functions. The scope of the report covers GPU semiconductor products used across applications such as consumer electronics, gaming, data centers, artificial intelligence, automotive, and industrial computing.

The GPU Semiconductor Market Report is Segmented by Integration Type (Integrated GPUs and Discrete GPUs), Device Application (Mobile Devices and Tablets, PCs and Workstations, Data Center and Server Accelerators, Gaming Consoles and Handheld Devices, Automotive and ADAS, and Embedded and Edge Devices), End User (Consumer Electronics and Gaming, Cloud, Data Center, and Enterprise IT, Telecom, Automotive, Industrial and Robotics, and Other End Users), Memory Type (GDDR-Based GPUs, HBM-Based GPUs, Shared DDR/LPDDR Memory GPUs, and Other Memory Types), 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 Integration Type
Integrated GPUs |
Discrete GPUs |
By Device Application
Mobile Devices and Tablets |
PCs and Workstations |
Data Center and Server Accelerators |
Gaming Consoles and Handheld Devices |
Automotive and ADAS |
Embedded and Edge Devices |
By End User
Consumer Electronics and Gaming |
Cloud, Data Center, and Enterprise IT |
Telecom |
Automotive |
Industrial and Robotics |
Other End Users |
By Memory Type
GDDR-Based GPUs |
HBM-Based GPUs |
Shared DDR/LPDDR Memory GPUs |
Other Memory Types |
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 |
By Integration Type | Integrated GPUs |
| Discrete GPUs |
By Device Application | Mobile Devices and Tablets |
| PCs and Workstations |
| Data Center and Server Accelerators |
| Gaming Consoles and Handheld Devices |
| Automotive and ADAS |
| Embedded and Edge Devices |
By End User | Consumer Electronics and Gaming |
| Cloud, Data Center, and Enterprise IT |
| Telecom |
| Automotive |
| Industrial and Robotics |
| Other End Users |
By Memory Type | GDDR-Based GPUs |
| HBM-Based GPUs |
| Shared DDR/LPDDR Memory GPUs |

| Other Memory Types |
|
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 ReportWhat is the current and forecast value of the GPU semiconductor space? The GPU semiconductor market size stood at USD 189.43 billion in 2025, reached USD 228.61 billion in 2026, and is forecast to reach USD 647.34 billion by 2031 at a 23.14% CAGR.

Which application leads revenue generation for GPU chips? Data center and server accelerators led with 73.52% share in 2025, showing that AI infrastructure remains the main demand center.

Which end-user group is driving the strongest demand? Cloud, data center, and enterprise IT held 76.39% share in 2025 and are projected to post the fastest 24.64% CAGR through 2031.

Why are HBM-based GPUs gaining so much importance? HBM-based GPUs led with 66.54% share in 2025 and are projected to grow at 24.16% CAGR because large-model AI workloads require much higher memory bandwidth.

Which region is growing fastest for GPU deployment and supply expansion? Asia-Pacific is projected to record the fastest 24.57% CAGR through 2031, supported by memory production, equipment investment, and broad AI infrastructure activity.

What is the main risk to wider GPU adoption outside hyperscalers? High platform costs remain the biggest hurdle because advanced memory and packaging keep system prices elevated for telecom, industrial, and other budget-sensitive buyers.

Related Reports Explore More Reports Find Latest Data onWorkstation GPU Market ResearchMarket Data for North America GPUGPU Cloud Market ResearchMarket Outlook for Industrial GPUIndia Discrete GPU Market AnalysisScientific Computing GPU Market ReportMarket Trends in Desktop GPUIndustry Outlook of GPU SubstrateMarket Report on United States Discrete GPUGPU Foundry MarketGrowth Forecasts for ABF Substrate Market For GPUs