

Data Center GPU Market Forecasts To 2034 – Global Analysis By GPU Type (Discrete GPUs, Integrated GPUs and Hybrid GPUs), GPU Architecture, Form Factor, Cooling Type, Function, Deployment, Data Center Type, Memory Technology, Application, End User and By Geography

<https://marketpublishers.com/r/D69920ACD72CEN.html>

Date: October 2026

Pages: 200

Price: US\$ 4,150.00 (Single User License)

ID: D69920ACD72CEN

Abstracts

According to Statistics MRC, the Global Data Center GPU Market is accounted for \$150.5 billion in 2026 and is expected to reach \$1118.6 billion by 2034 growing at a CAGR of 28.5% during the forecast period. The Data Center GPU Market covers graphics processing technologies developed for intensive data center applications, including artificial intelligence, machine learning, high-performance computing, cloud services, scientific workloads, and analytics. These GPUs enable extensive parallel computation for processing large datasets and complex computational models. The market incorporates GPU hardware, software ecosystems, memory solutions, networking technologies, and related infrastructure. Major applications include AI training and inference, accelerated computing, simulation, virtualization, and enterprise data processing. The competitive landscape includes GPU manufacturers, cloud platforms, and technology companies offering specialized architectures and solutions for advanced data center computing requirements.

Market Dynamics:

Driver:

Increasing Data Processing Requirements

The expanding quantity of data produced by businesses, connected devices, digital

services, and software applications is increasing the need for accelerated processing within data centers. Companies must analyze extensive datasets to support business intelligence, cybersecurity, customer insights, recommendation engines, and real-time operational decisions. GPUs can perform large numbers of data-processing operations concurrently, complementing CPUs in computationally intensive analytics and processing workloads. As data-centric applications become increasingly widespread, enterprises and infrastructure providers are adopting accelerated computing systems to address processing requirements. The growing necessity for efficient analysis and handling of large datasets is consequently supporting greater integration of GPU technologies into data center infrastructure.

Restraint:

High Cost of Data Center GPUs

The substantial investment required for sophisticated data center GPUs can limit their adoption across organizations. GPUs designed for intensive AI, machine learning, and high-performance computing workloads generally involve higher procurement costs than standard computing components. Deployment expenses can also extend beyond the GPU itself, encompassing specialized servers, advanced memory, high-speed networking, thermal management, electricity infrastructure, and supporting software. These requirements can significantly increase the overall expenditure associated with GPU-based computing environments. Small and medium-sized businesses or organizations operating with constrained budgets may find large GPU deployments challenging, especially when their workloads do not provide sufficient utilization to justify the considerable infrastructure and equipment investment required.

Opportunity:

Integration of GPUs with Advanced Networking and Memory Technologies

Combining GPUs with advanced memory and networking technologies creates opportunities for developing more capable data center computing platforms. AI and high-performance applications frequently transfer substantial quantities of information among processors, memory, storage, and networking components. GPU manufacturers can work with memory suppliers, networking companies, semiconductor developers, and infrastructure providers to create integrated systems incorporating high-bandwidth memory, fast interconnects, advanced networking, and optimized software. Improved coordination among these components can support communication across large accelerator clusters and demanding workloads. This trend can encourage strategic

collaboration, ecosystem development, and integrated platform offerings, allowing technology providers to deliver comprehensive accelerated computing infrastructure for data-intensive applications.

Threat:

Rapid Technological Competition and Product Substitution

Fast innovation across computing and semiconductor technologies can increase substitution risks for data center GPUs. New processor architectures, dedicated AI chips, custom accelerators, neuromorphic technologies, and other emerging solutions are being designed for specialized computational workloads. Advances in these alternatives could enable some applications to meet performance requirements without extensive use of traditional GPUs. Data center operators may consequently compare multiple processor architectures based on workload suitability, power requirements, software support, and total infrastructure costs. Rapid technology cycles can also reduce the effective lifespan of existing solutions, requiring GPU companies to maintain substantial investments in research, architecture development, software platforms, and continuous product innovation.

Covid-19 Impact:

The COVID-19 pandemic affected the Data Center GPU Market through accelerated digitalization, increased remote working, expanded cloud usage, and greater dependence on online platforms. Businesses and institutions adopted cloud-based infrastructure, virtual communication, digital services, and data-intensive applications, increasing requirements for computing resources and data center capacity. However, pandemic-related restrictions disrupted semiconductor production, transportation networks, component sourcing, and hardware deliveries, creating obstacles for GPU availability and infrastructure expansion. Data center operators also encountered operational challenges resulting from workforce limitations and facility restrictions. Consequently, COVID-19 highlighted the importance of advanced computing infrastructure while creating simultaneous supply chain and deployment difficulties for GPU manufacturers and data center operators.

The Discrete GPUs segment is expected to be the largest during the forecast period

The Discrete GPUs segment is expected to account for the largest market share during the forecast period, driven by its dedicated processing resources, high performance,

and suitability for computationally intensive data center applications. Unlike integrated solutions, discrete GPUs function separately from CPUs and provide extensive parallel processing capabilities for artificial intelligence, machine learning, analytics, high-performance computing, and scientific workloads. Their ability to scale across multi-GPU configurations allows data centers to address demanding computational requirements efficiently. Growing needs for accelerated processing, sophisticated AI models, extensive datasets, and specialized computing infrastructure are contributing to the increasing deployment of discrete GPUs within advanced data center environments.

The Generative AI segment is expected to have the highest CAGR during the forecast period

Over the forecast period, the Generative AI segment is predicted to witness the highest growth rate, supported by the expanding use of large language models, multimodal systems, AI-generated content, and advanced artificial intelligence applications. These workloads require significant computational capabilities for training, model optimization, fine-tuning, and inference, increasing the need for powerful GPUs within data center infrastructure. Cloud providers, technology companies, enterprises, and research institutions are developing accelerated computing environments to accommodate increasingly complex AI models. The broader integration of generative AI into software development, customer support, content production, healthcare, financial services, and other applications is encouraging investment in specialized GPU infrastructure capable of supporting demanding and computationally intensive AI workloads.

Region with largest share:

During the forecast period, the North America region is expected to hold the largest market share, supported by its extensive data center ecosystem, concentration of leading technology companies, and broad adoption of cloud and artificial intelligence technologies. The region benefits from established GPU suppliers, semiconductor manufacturers, cloud platforms, and large data center operators. Increasing deployment of generative AI, machine learning, high-performance computing, analytics, and other demanding workloads is supporting the use of GPU-accelerated infrastructure. The presence of hyperscale facilities and continued investment in advanced computing systems further strengthens regional demand. In addition, mature digital infrastructure, technical expertise, and established technology ecosystems support North America's prominent position in the market.

Region with highest CAGR:

Over the forecast period, the Asia Pacific region is anticipated to exhibit the highest CAGR, supported by increasing data center capacity, expanding cloud services, and greater adoption of artificial intelligence technologies. Countries throughout the region are developing hyperscale data centers, AI-focused infrastructure, and advanced digital platforms, increasing demand for GPU-accelerated computing resources. Growing applications of generative AI, machine learning, analytics, and high-performance computing are encouraging businesses to adopt sophisticated accelerator solutions. Accelerating digital transformation among enterprises and public-sector organizations is further supporting infrastructure requirements. In addition, developing technology ecosystems and investments in semiconductor manufacturing and data center facilities are creating favorable conditions for expanding GPU deployment across the region.

Key players in the market

Some of the key players in Data Center GPU Market include Intel Corporation, Advanced Micro Devices, Inc., NVIDIA Corporation, Huawei Technologies Co., Ltd., Biren Technology, Moore Threads Technology Co., Ltd., Iluvatar CoreX, Cambricon Technologies Corporation Limited, Jingjia Microelectronics Co., Ltd., Enflame Technology Co., Ltd., MetaX Integrated Circuits Co., Ltd., Alibaba Group Holding Limited, Baidu, Inc., Fujitsu Limited, IBM Corporation, Lenovo Group Limited, Inspur Group and Foxconn Industrial Internet Co., Ltd.

Key Developments:

In July 2026, AMD and Microsoft expanded their long-term strategic partnership, with Microsoft planning to deploy AMD Helios rack-scale systems and next-generation AMD Instinct GPUs on Azure for frontier-model inference and AI services.

In May 2026, NVIDIA and IREN announced a strategic partnership supporting deployment of up to 5 GW of NVIDIA DSX-aligned AI infrastructure across IREN's data-center pipeline, combining NVIDIA accelerated computing with IREN's data-center and infrastructure capabilities.

In April 2026, Intel and Google announced a multiyear collaboration to advance AI and cloud infrastructure, including continued use of Intel Xeon processors and expanded co-development of custom ASIC-based infrastructure processing units (IPUs) for large-scale AI infrastructure.

GPU Types Covered:

Discrete GPUs

Integrated GPUs

Hybrid GPUs

GPU Architectures Covered:

CUDA-Based Architectures

CDNA-Based Architectures

Xe-Based Architectures

Other GPU Architectures

Form Factors Covered:

Single-Slot GPUs

Dual-Slot GPUs

Multi-Slot GPUs

SXM Modules

Rack-Scale GPU Systems

Cooling Types Covered:

Air-Cooled

Liquid-Cooled

Immersion-Cooled

Functions Covered:

AI Training

AI Inference

High-Performance Computing

Graphics and Visualization

Media and Video Processing

Deployments Covered:

Cloud

On-Premises

Data Center Types Covered:

Hyperscale Data Centers

Enterprise Data Centers

Colocation Data Centers

Edge Data Centers

Memory Technologies Covered:

HBM

GDDR

Other Memory Technologies

Applications Covered:

Generative AI

Machine Learning

Data Analytics

High-Performance Computing

Graphics Rendering

Media Processing

Virtualization and VDI

End Users Covered:

Cloud Service Providers

Enterprises

Government and Research Institutions

Regions Covered:

North America

United States

Canada

Mexico

Europe

United Kingdom

Germany

France

Italy

Spain

Netherlands

Belgium

Sweden

Switzerland

Poland

Rest of Europe

Asia Pacific

China

Japan

India

South Korea

Australia

Indonesia

Thailand

Malaysia

Singapore

Vietnam

Rest of Asia Pacific

South America

Brazil

Argentina

Colombia

Chile

Peru

Rest of South America

Rest of the World (RoW)

Middle East

Saudi Arabia

United Arab Emirates

Qatar

Israel

Rest of Middle East

Africa
South Africa

Egypt

Morocco

Rest of Africa

What our report offers:

Market share assessments for the regional and country-level segments

Strategic recommendations for the new entrants

Covers Market data for the years 2023, 2024, 2025, 2026, 2027, 2028, 2030, 2032 and 2034

Market Trends (Drivers, Constraints, Opportunities, Threats, Challenges, Investment Opportunities, and recommendations)

Strategic recommendations in key business segments based on the market estimations

Competitive landscaping mapping the key common trends

Company profiling with detailed strategies, financials, and recent developments

Supply chain trends mapping the latest technological advancements

Free Customization Offerings:

All the customers of this report will be entitled to receive one of the following free customization options:

Company Profiling

Comprehensive profiling of additional market players (up to 3)

SWOT Analysis of key players (up to 3)

Regional Segmentation

Market estimations, Forecasts and CAGR of any prominent country as per the client's interest (Note: Depends on feasibility check)

Competitive Benchmarking

Benchmarking of key players based on product portfolio, geographical presence, and strategic alliances

Contents

1 EXECUTIVE SUMMARY

- 1.1 Market Snapshot and Key Highlights
- 1.2 Growth Drivers, Challenges, and Opportunities
- 1.3 Competitive Landscape Overview
- 1.4 Strategic Insights and Recommendations

2 RESEARCH FRAMEWORK

- 2.1 Study Objectives and Scope
- 2.2 Stakeholder Analysis
- 2.3 Research Assumptions and Limitations
- 2.4 Research Methodology
 - 2.4.1 Data Collection (Primary and Secondary)
 - 2.4.2 Data Modeling and Estimation Techniques
 - 2.4.3 Data Validation and Triangulation
 - 2.4.4 Analytical and Forecasting Approach

3 MARKET DYNAMICS AND TREND ANALYSIS

- 3.1 Market Definition and Structure
- 3.2 Key Market Drivers
- 3.3 Market Restraints and Challenges
- 3.4 Growth Opportunities and Investment Hotspots
- 3.5 Industry Threats and Risk Assessment
- 3.6 Technology and Innovation Landscape
- 3.7 Emerging and High-Growth Markets
- 3.8 Regulatory and Policy Environment
- 3.9 Impact of COVID-19 and Recovery Outlook

4 COMPETITIVE AND STRATEGIC ASSESSMENT

- 4.1 Porter's Five Forces Analysis
 - 4.1.1 Supplier Bargaining Power
 - 4.1.2 Buyer Bargaining Power
 - 4.1.3 Threat of Substitutes
 - 4.1.4 Threat of New Entrants

- 4.1.5 Competitive Rivalry
- 4.2 Market Share Analysis of Key Players
- 4.3 Product Benchmarking and Performance Comparison

5 GLOBAL DATA CENTER GPU MARKET, BY GPU TYPE

- 5.1 Discrete GPUs
- 5.2 Integrated GPUs
- 5.3 Hybrid GPUs

6 GLOBAL DATA CENTER GPU MARKET, BY GPU ARCHITECTURE

- 6.1 CUDA-Based Architectures
- 6.2 CDNA-Based Architectures
- 6.3 Xe-Based Architectures
- 6.4 Other GPU Architectures

7 GLOBAL DATA CENTER GPU MARKET, BY FORM FACTOR

- 7.1 Single-Slot GPUs
- 7.2 Dual-Slot GPUs
- 7.3 Multi-Slot GPUs
- 7.4 SXM Modules
- 7.5 Rack-Scale GPU Systems

8 GLOBAL DATA CENTER GPU MARKET, BY COOLING TYPE

- 8.1 Air-Cooled
- 8.2 Liquid-Cooled
- 8.3 Immersion-Cooled

9 GLOBAL DATA CENTER GPU MARKET, BY FUNCTION

- 9.1 AI Training
- 9.2 AI Inference
- 9.3 High-Performance Computing
- 9.4 Graphics and Visualization
- 9.5 Media and Video Processing

10 GLOBAL DATA CENTER GPU MARKET, BY DEPLOYMENT

- 10.1 Cloud
- 10.2 On-Premises

11 GLOBAL DATA CENTER GPU MARKET, BY DATA CENTER TYPE

- 11.1 Hyperscale Data Centers
- 11.2 Enterprise Data Centers
- 11.3 Colocation Data Centers
- 11.4 Edge Data Centers

12 GLOBAL DATA CENTER GPU MARKET, BY MEMORY TECHNOLOGY

- 12.1 HBM
- 12.2 GDDR
- 12.3 Other Memory Technologies

13 GLOBAL DATA CENTER GPU MARKET, BY APPLICATION

- 13.1 Generative AI
- 13.2 Machine Learning
- 13.3 Data Analytics
- 13.4 High-Performance Computing
- 13.5 Graphics Rendering
- 13.6 Media Processing
- 13.7 Virtualization and VDI

14 GLOBAL DATA CENTER GPU MARKET, BY END USER

- 14.1 Cloud Service Providers
- 14.2 Enterprises
- 14.3 Government and Research Institutions

15 GLOBAL DATA CENTER GPU MARKET, BY GEOGRAPHY

- 15.1 North America
 - 15.1.1 United States
 - 15.1.2 Canada

- 15.1.3 Mexico
- 15.2 Europe
 - 15.2.1 United Kingdom
 - 15.2.2 Germany
 - 15.2.3 France
 - 15.2.4 Italy
 - 15.2.5 Spain
 - 15.2.6 Netherlands
 - 15.2.7 Belgium
 - 15.2.8 Sweden
 - 15.2.9 Switzerland
 - 15.2.10 Poland
 - 15.2.11 Rest of Europe
- 15.3 Asia Pacific
 - 15.3.1 China
 - 15.3.2 Japan
 - 15.3.3 India
 - 15.3.4 South Korea
 - 15.3.5 Australia
 - 15.3.6 Indonesia
 - 15.3.7 Thailand
 - 15.3.8 Malaysia
 - 15.3.9 Singapore
 - 15.3.10 Vietnam
 - 15.3.11 Rest of Asia Pacific
- 15.4 South America
 - 15.4.1 Brazil
 - 15.4.2 Argentina
 - 15.4.3 Colombia
 - 15.4.4 Chile
 - 15.4.5 Peru
 - 15.4.6 Rest of South America
- 15.5 Rest of the World (RoW)
 - 15.5.1 Middle East
 - 15.5.1.1 Saudi Arabia
 - 15.5.1.2 United Arab Emirates
 - 15.5.1.3 Qatar
 - 15.5.1.4 Israel
 - 15.5.1.5 Rest of Middle East

15.5.2 Africa

15.5.2.1 South Africa

15.5.2.2 Egypt

15.5.2.3 Morocco

15.5.2.4 Rest of Africa

16 STRATEGIC MARKET INTELLIGENCE

16.1 Industry Value Network and Supply Chain Assessment

16.2 White-Space and Opportunity Mapping

16.3 Product Evolution and Market Life Cycle Analysis

16.4 Channel, Distributor, and Go-to-Market Assessment

17 INDUSTRY DEVELOPMENTS AND STRATEGIC INITIATIVES

17.1 Mergers and Acquisitions

17.2 Partnerships, Alliances, and Joint Ventures

17.3 New Product Launches and Certifications

17.4 Capacity Expansion and Investments

17.5 Other Strategic Initiatives

18 COMPANY PROFILES

18.1 Intel Corporation

18.2 Advanced Micro Devices, Inc.

18.3 NVIDIA Corporation

18.4 Huawei Technologies Co., Ltd.

18.5 Biren Technology

18.6 Moore Threads Technology Co., Ltd.

18.7 Iluvatar CoreX

18.8 Cambricon Technologies Corporation Limited

18.9 Jingjia Microelectronics Co., Ltd.

18.10 Enflame Technology Co., Ltd.

18.11 MetaX Integrated Circuits Co., Ltd.

18.12 Alibaba Group Holding Limited

18.13 Baidu, Inc.

18.14 Fujitsu Limited

18.15 IBM Corporation

18.16 Lenovo Group Limited

18.17 Inspur Group

18.18 Foxconn Industrial Internet Co., Ltd.

List of Tables

Table 1 Global Data Center GPU Market Outlook, By Region (2023-2034) (\$MN)

Table 2 Global Data Center GPU Market Outlook, By GPU Type (2023-2034) (\$MN)

Table 3 Global Data Center GPU Market Outlook, By Discrete GPUs (2023-2034) (\$MN)

Table 4 Global Data Center GPU Market Outlook, By Integrated GPUs (2023-2034) (\$MN)

Table 5 Global Data Center GPU Market Outlook, By Hybrid GPUs (2023-2034) (\$MN)

Table 6 Global Data Center GPU Market Outlook, By GPU Architecture (2023-2034) (\$MN)

Table 7 Global Data Center GPU Market Outlook, By CUDA-Based Architectures (2023-2034) (\$MN)

Table 8 Global Data Center GPU Market Outlook, By CDNA-Based Architectures (2023-2034) (\$MN)

Table 9 Global Data Center GPU Market Outlook, By Xe-Based Architectures (2023-2034) (\$MN)

Table 10 Global Data Center GPU Market Outlook, By Other GPU Architectures (2023-2034) (\$MN)

Table 11 Global Data Center GPU Market Outlook, By Form Factor (2023-2034) (\$MN)

Table 12 Global Data Center GPU Market Outlook, By Single-Slot GPUs (2023-2034) (\$MN)

Table 13 Global Data Center GPU Market Outlook, By Dual-Slot GPUs (2023-2034) (\$MN)

Table 14 Global Data Center GPU Market Outlook, By Multi-Slot GPUs (2023-2034) (\$MN)

Table 15 Global Data Center GPU Market Outlook, By SXM Modules (2023-2034) (\$MN)

Table 16 Global Data Center GPU Market Outlook, By Rack-Scale GPU Systems (2023-2034) (\$MN)

Table 17 Global Data Center GPU Market Outlook, By Cooling Type (2023-2034) (\$MN)

Table 18 Global Data Center GPU Market Outlook, By Air-Cooled (2023-2034) (\$MN)

Table 19 Global Data Center GPU Market Outlook, By Liquid-Cooled (2023-2034) (\$MN)

Table 20 Global Data Center GPU Market Outlook, By Immersion-Cooled (2023-2034) (\$MN)

Table 21 Global Data Center GPU Market Outlook, By Function (2023-2034) (\$MN)

Table 22 Global Data Center GPU Market Outlook, By AI Training (2023-2034) (\$MN)

Table 23 Global Data Center GPU Market Outlook, By AI Inference (2023-2034) (\$MN)

Table 24 Global Data Center GPU Market Outlook, By High-Performance Computing (2023-2034) (\$MN)

Table 25 Global Data Center GPU Market Outlook, By Graphics and Visualization (2023-2034) (\$MN)

Table 26 Global Data Center GPU Market Outlook, By Media and Video Processing (2023-2034) (\$MN)

Table 27 Global Data Center GPU Market Outlook, By Deployment (2023-2034) (\$MN)

Table 28 Global Data Center GPU Market Outlook, By Cloud (2023-2034) (\$MN)

Table 29 Global Data Center GPU Market Outlook, By On-Premises (2023-2034) (\$MN)

Table 30 Global Data Center GPU Market Outlook, By Data Center Type (2023-2034) (\$MN)

Table 31 Global Data Center GPU Market Outlook, By Hyperscale Data Centers (2023-2034) (\$MN)

Table 32 Global Data Center GPU Market Outlook, By Enterprise Data Centers (2023-2034) (\$MN)

Table 33 Global Data Center GPU Market Outlook, By Colocation Data Centers (2023-2034) (\$MN)

Table 34 Global Data Center GPU Market Outlook, By Edge Data Centers (2023-2034) (\$MN)

Table 35 Global Data Center GPU Market Outlook, By Memory Technology (2023-2034) (\$MN)

Table 36 Global Data Center GPU Market Outlook, By HBM (2023-2034) (\$MN)

Table 37 Global Data Center GPU Market Outlook, By GDDR (2023-2034) (\$MN)

Table 38 Global Data Center GPU Market Outlook, By Other Memory Technologies (2023-2034) (\$MN)

Table 39 Global Data Center GPU Market Outlook, By Application (2023-2034) (\$MN)

Table 40 Global Data Center GPU Market Outlook, By Generative AI (2023-2034) (\$MN)

Table 41 Global Data Center GPU Market Outlook, By Machine Learning (2023-2034) (\$MN)

Table 42 Global Data Center GPU Market Outlook, By Data Analytics (2023-2034) (\$MN)

Table 43 Global Data Center GPU Market Outlook, By High-Performance Computing (2023-2034) (\$MN)

Table 44 Global Data Center GPU Market Outlook, By Graphics Rendering (2023-2034) (\$MN)

Table 45 Global Data Center GPU Market Outlook, By Media Processing (2023-2034) (\$MN)

Table 46 Global Data Center GPU Market Outlook, By Virtualization and VDI
(2023-2034) (\$MN)

Table 47 Global Data Center GPU Market Outlook, By End User (2023-2034) (\$MN)

Table 48 Global Data Center GPU Market Outlook, By Cloud Service Providers
(2023-2034) (\$MN)

Table 49 Global Data Center GPU Market Outlook, By Enterprises (2023-2034) (\$MN)

Table 50 Global Data Center GPU Market Outlook, By Government and Research
Institutions (2023-2034) (\$MN)

Note: Tables for North America, Europe, APAC, South America, and Rest of the World
(RoW) Regions are also represented in the same manner as above.

I would like to order

Product name: Data Center GPU Market Forecasts To 2034 – Global Analysis By GPU Type (Discrete GPUs, Integrated GPUs and Hybrid GPUs), GPU Architecture, Form Factor, Cooling Type, Function, Deployment, Data Center Type, Memory Technology, Application, End User and By Geography

Product link: <https://marketpublishers.com/r/D69920ACD72CEN.html>

Price: US\$ 4,150.00 (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/D69920ACD72CEN.html>