AI Infrastructure Market Size & Share Analysis - Growth Trends and Forecast (2026 - 2031)
Industry: Technology, Media & Telecom (TMT)
URL: https://www.mordorintelligence.com/industry-reports/ai-infrastructure-market

AI Infrastructure Market Size and ShareMarket OverviewStudy Period | 2020 - 2031 |
Market Size (2026) | USD 101.17 Billion |
Market Size (2031) | USD 202.48 Billion |
Growth Rate (2026 - 2031) | 14.89 % |
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.

AI Infrastructure Market Analysis by Mordor IntelligenceMarket Analysis

The AI infrastructure market size reached USD 101.17 billion in 2026 and is projected to reach USD 202.48 billion by 2031, reflecting a 14.89% CAGR over the forecast period. The expansion aligns with sustained allocations toward compute-intensive workloads, continued subsidy inflows for advanced semiconductor fabs, and a persistent premium on high-bandwidth memory that lengthens lead times for top-tier GPUs. Intensifying liquid-cooling adoption mitigates racks that now surpass 100 kilowatts, while export controls enacted by the United States in 2023 accelerate sovereign AI projects across the Middle East and Asia Pacific. Semiconductor policy has become a growth catalyst, as CHIPS-type incentives underpin fab expansions in the United States, Europe, and Japan. Hyperscalers, facing multi-year backlogs for NVIDIA H100 and H200 accelerators, have responded by pre-ordering next-generation devices and designing custom ASICs to secure capacity.

Key Report Takeaways* By offering, hardware led with 68.42% of revenue in 2025; the software segment is forecast to expand at a 16.02% CAGR to 2031.
* By deployment, on-premise architectures held 57.46% of the AI infrastructure market share in 2025, while cloud deployments are projected to advance at a 15.76% CAGR through 2031.
* By end user, enterprises commanded 42.22% share of the AI infrastructure market size in 2025; cloud service providers represent the fastest-growing cohort at 15.24% CAGR to 2031.
* By processor architecture, GPUs retained 88.82% of revenue in 2025, whereas FPGA and ASIC alternatives are poised to grow at 16.89% CAGR through 2031.
* By geography, North America accounted for 39.56% of 2025 revenue; Asia Pacific is expected to expand at a 16.44% CAGR between 2026-2031.
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 Infrastructure Market*Driver | (~) % Impact on CAGR Forecast | Geographic Relevance | Impact Timeline |
Soaring H100 and H200 GPU backlogs | +3.2% | Global, concentrated in North America and Asia Pacific | Medium term (2-4 years) |
Rapid AI-specific network fabrics | +2.8% | Global, early adoption in North America and Europe | Medium term (2-4 years) |

Energy-efficient liquid cooling adoption | +2.1% | Global, led by North America and Europe | Long term (≥ 4 years) |
Government CHIPS-type subsidies for AI fabs | +2.5% | North America, Europe, Asia Pacific | Long term (≥ 4 years) |
Cloud-native AI accelerator instances | +2.4% | Global, strongest in North America and Asia Pacific | Short term (≤ 2 years) |
Open-source AI framework optimization | +1.9% | Global | Medium term (2-4 years) |
Source: Mordor Intelligence |
Soaring H100 and H200 GPU Backlogs Among HyperscalersNVIDIA reported that 2025 pre-orders for H100 and H200 devices tripled available supply, prompting Microsoft to earmark USD 80 billion for multi-year allocations and AWS to expand its infrastructure budget by USD 100 billion through 2028.[1]Microsoft Communications, “Microsoft Announces USD 80 Billion AI Datacenter Investment,” Microsoft, microsoft.com High-bandwidth memory bottlenecks intensified the imbalance, as SK Hynix and Samsung controlled 95% of HBM3E output. Hyperscalers now co-design memory packaging directly with fabs, weakening the negotiating leverage of traditional GPU vendors. TSMC’s 3-nanometer capacity remained oversubscribed, extending device lead times past 12 months and accelerating a pivot toward custom ASICs such as Google TPU v6e. Enterprises, facing unpredictable delivery schedules, increasingly rent guaranteed instances from cloud providers even when on-demand prices exceed USD 30 per hour for eight-GPU bundles.

Rapid AI-Specific Network Fabrics (InfiniBand NDR, Ethernet 800G)InfiniBand NDR operated at 400 Gbps and connected about 70% of 2025 AI training clusters, delivering latency that was 40% lower than traditional Ethernet.[2]NVIDIA Networking Team, “InfiniBand Solutions,” nvidia.com Hyperscalers, however, began evaluating 800 Gbps Ethernet as Broadcom’s Tomahawk 5 and Spectrum-X switched traffic at competitive latencies with a 25% reduction in capital cost. Meta validated Ethernet performance by scaling its 10,000-GPU AI Research SuperCluster on 800 Gbps links, widening vendor choice and eroding InfiniBand lock-in. IEEE 802.3df work on 1.6 Tbps Ethernet continues, signaling more convergence between AI and standard data-center workloads.

Energy-Efficient Liquid Cooling AdoptionLiquid cooling penetration rose to 18% of AI racks in 2025 as power densities crossed 100 kilowatts, a threshold at which air systems struggle to remove heat. Direct-to-chip solutions lowered facility energy use by up to 40% and freed 60% of floor space compared with air-cooled equivalents. Microsoft piloted single-phase immersion baths that trimmed cooling infrastructure costs by 45% and expects to roll the approach across hyperscale campuses from 2026 onward. Regulatory incentives in the European Union, including looming carbon price mechanisms, reinforce adoption, whereas deployments in Asia Pacific lag due to lower electricity tariffs.

Government CHIPS-Type Subsidies for AI FabsThe United States allocated USD 52.7 billion for domestic semiconductor manufacturing, disbursing USD 8.5 billion to Intel, USD 6.6 billion to TSMC, and USD 6.4 billion to Samsung. Europe passed its EUR 43 billion (USD 47 billion) Chips Act to double regional wafer output by 2030. Japan set aside JPY 2 trillion (USD 13.5 billion) to back TSMC’s Kumamoto plant and a 2-nanometer roadmap led by Rapidus. Subsidies accelerate advanced-packaging capacity and de-risk geopolitical concentration in Taiwan, yet the Semiconductor Industry Association projects a 67,000-person talent gap that may delay full utilization.

Restraints Impact Analysis of AI Infrastructure Market*Restraint | (~) % Impact on CAGR Forecast | Geographic Relevance | Impact Timeline |
AI-class GPUs in chronic short supply through 2026 | -2.8% | Global, acute in Asia Pacific and Europe | Short term (≤ 2 years) |
400 V and 48 V power-conversion limits in legacy sites | -1.9% | Global, concentrated in North America and Europe | Medium term (2-4 years) |
Sovereign-AI export controls | -2.3% | Global, most severe in Asia Pacific | Long term (≥ 4 years) |

Rising Scope-2 emissions compliance costs | -1.6% | Europe, emerging in North America | Medium term (2-4 years) |
Source: Mordor Intelligence |
AI-Class GPUs in Chronic Short Supply Through 2026Lead times for H200 cards lengthened past 52 weeks in 2025, while AMD's MI300X backlog mirrored the constraint.[3]Reuters Staff, "NVIDIA AI Chip Orders Exceed Supply, Lead Times Stretch," Reuters, reuters.com CoWoS packaging capacity at TSMC hit 35,000 wafer starts per month, far below demand estimates above 100,000 equivalents. High-bandwidth memory remains scarce because each H100 device needs 80 GB of HBM3 stacked across five layers. Enterprises consequently delayed large-scale deployments and reprioritized model architectures that require fewer parameters. Cloud platforms countered by over-provisioning inventory, dropping utilization rates, and charging elevated spot prices, a tactic that distorts supply signals and suppresses near-term market adoption.

Sovereign-AI Export ControlsOctober 2023 regulations barred unlicensed shipment of NVIDIA's A100, H100, and H800 devices to China. China responded with a USD 50 billion chip program, and Huawei's Ascend 910C achieved parity with A100 on some inference tests in 2025. The European Union's AI Act adds EUR 5-15 million (USD 5.5-16.5 million) in compliance costs per cross-border deployment. Divergent standards risk a bifurcated ecosystem and elevate switching costs for international firms.

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

AI Infrastructure Market Segment AnalysisBy Offering:Software Gains as Inference Optimization Trumps Raw ComputeHardware commanded 68.42% of 2025 spending, reflecting capital-intensive GPU clusters, high-bandwidth memory, and NVMe fabrics that push rack densities beyond 100 kilowatts. Software is projected to rise at a 16.02% CAGR to 2031 as enterprises emphasize inferencing efficiency, model observability, and MLOps automation. Tools like Triton Inference Server compress latency by as much as 50% through quantization and kernel fusion. System vendors now bundle orchestration frameworks with observability dashboards, converting one-off licenses into subscriptions. The AI infrastructure market size attributed to software is therefore expanding faster than GPU capital investment, even though absolute spending on accelerators remains larger. Training workloads will stay GPU-centric, but inference is already moving toward purpose-built ASICs that lower total cost of ownership for production pipelines. Enterprises gaining cost relief redeploy freed budgets into data-quality initiatives and retrieval-augmented generation pipelines, pushing middleware adoption higher.

A second catalyst is the rise of large-language-model-as-a-service offerings that embed guardrails for content safety and bias mitigation. Vendors who package middleware with pre-trained models secure recurring revenue and deepen customer lock-in. Independent software providers respond by hardening open-source deployment stacks, ensuring that proprietary licensing does not impede model portability. The emergent dynamic elevates software gross margins toward 75%, well above hardware reselling levels, underscoring why investors favor code over silicon in later-stage funding rounds. The AI infrastructure market therefore shifts from a capital-expenditure cycle to a blended model where subscription revenue stabilizes earnings and mitigates hardware refresh volatility.

 By Deployment:Cloud Instances Erode On-Premise Moats Despite Sovereignty ConcernsOn-premise infrastructure held 57.46% of spending in 2025, driven by data-residency mandates and sectoral frameworks like HIPAA. Cloud deployments are forecast to grow at 15.76% CAGR as AWS Trainium2 and Google TPU v6e instances deliver multi-petaflop performance at favorable economics. The AI infrastructure market size asso

ciated with cloud offerings is thus expanding faster than enterprise capex, espe cially as hyperscalers standardize pay-per-inference pricing. Financial institut ions that once insisted on sovereign hosting now pilot confidential-compute encl aves that keep encryption keys under customer control, reducing regulatory frict ion.

Hybrid patterns proliferate as enterprises train sensitive models on-premise the n shift inference to geographic edge nodes that lower latency for end users. Sov ereign AI initiatives in Saudi Arabia and the United Arab Emirates inject more t han USD 140 billion to build domestic hyperscale campuses, sustaining a counterv ailing demand for local deployments. Cloud providers accommodate sovereignty by offering dedicated regions with jurisdictionally ring-fenced networking, certifi cations, and auditing. Over the long term, however, hardware obsolescence cycles of 18-24 months tilt the cost curve toward shared infrastructure, compelling on -premise defenders to adopt modular designs that swap node boards without re-cab ling entire halls.

By End User:Cloud Service Providers Outspend Enterprises to Lock In Competitive MoatsEnterprises accounted for 42.22% of the 2025 AI infrastructure market share , reflecting diversified use cases in manufacturing, retail, and professional se rvices. Cloud service providers are projected to post 15.24% CAGR as hyperscaler s pre-commit multi-billion-dollar blocks of HBM3E-attached accelerators. Bulk pr ocurement secures lower unit prices, enabling hyperscalers to offer burstable tr aining clusters at hourly rates still below the amortized cost of enterprise-own ed equivalents. Government and defense agencies adopt air-gapped, top-secret enc losures that insulate classified workloads yet benefit from cloud-like managemen t software.

Enterprises weighing build versus rent must navigate capital risk, personnel sho rtages, and warranty uncertainty. A 1,000-GPU private cluster runs USD 15-30 mil lion upfront and becomes partially obsolete inside two years, while subscription models convert that outlay into predictable operating expense. Hyperscalers com pound the advantage with integrated data labeling, MLOps, and fine-tuning servic es. Governments, however, view AI sovereignty as strategic. Japan’s Ministry of Defense budgeted JPY 500 billion (USD 3.4 billion) for indigenous systems, refle cting geopolitical urgency rather than mere cost considerations.

 By Proc essor Architecture:FPGA and ASIC Alternatives Challenge GPU Hegemony in Inferenc eGPUs controlled 88.82% of 2025 revenue due to the entrenched CUDA ecosystem and parallelism requirements of transformer training. FPGA and ASIC devices are for ecast to expand at 16.89% CAGR as inference workloads prioritize energy efficien cy and predictable latency. Intel Gaudi 3 delivers 50% better performance-per-wa tt than H100 for transformer inference, while Cerebras WSE-3 packs 900,000 cores on a wafer-scale die suited to physics simulations. Google’s TPU v6e already ru ns production inference at 2.5 times GPU energy efficiency.

The AI infrastructure market therefore splinters between general-purpose GPUs an d domain-specific ASICs. Custom silicon carries high non-recurring engineering c osts, limiting feasibility to hyperscalers with trillions of inference queries p er quarter. FPGAs serve a middle niche in telecommunications and automotive, whe re algorithms evolve rapidly and field-upgrade flexibility is crucial. Vendors n ow develop chiplet-based SoCs that interconnect via die-to-die links like UCIe, cutting time-to-market and allowing incremental memory upgrades. NVIDIA’s acquis ition of interposer IP and AMD’s investment in chiplet packaging signal a future where modular substrates dilute single-vendor dominance.

Geography AnalysisNorth America AI Infrastructure MarketNorth America commanded 39.56% of 2025 spending, supported by USD 52.7 billion in CHIPS Act grants and b y hyperscalers that operate roughly 60% of global AI capacity. The Semiconductor

Industry Association warns of a 67,000-worker talent shortage by 2030, which could slow fab ramp-ups even as capital is plentiful. Canada positions Toronto and Montreal as research hubs backed by supportive immigration policy, whereas Mexico's grid reliability questions dampen large-scale build-outs. The United States Department of Defense awarded Amazon a USD 50 billion cloud contract, underscoring that sovereign security concerns coexist with a broader shift toward centrally managed compute.

APAC AI Infrastructure MarketAsia Pacific is expected to grow at a 16.44% CAGR through 2031, propelled by China's USD 50 billion semiconductor fund and India's USD 15 billion hyperscaler commitments. Alibaba deployed 100,000 Huawei Ascend 910C accelerators in 2025, illustrating rapid indigenous progress despite export curbs. Japan allocated JPY 2 trillion (USD 13.5 billion) for TSMC's Kumamoto site and 2-nanometer R&D to hedge geopolitical exposure. South Korea enjoys 95% share of HBM3E supply, an essential choke point in the AI supply chain. Australia's high power tariffs limit hyperscale, but Sydney and Melbourne still attract colocation players looking for resilient connectivity to submarine cables.

Europe and Middle East AI Infrastructure MarketEurope's growth moderates as AI Act compliance layers EUR 5-15 million (USD 5.5-16.5 million) in incremental cost per multi-nation deployment. Germany and France lead semiconductor subsidies, while Sweden leverages cold climate and hydroelectric power to tempt hyperscalers; Microsoft confirmed a USD 3.2 billion Stockholm campus for 2026. The United Kingdom confronts post-Brexit data transfer frictions that add latency and legal overhead to continent-wide services. Middle East sovereign wealth funds pledge USD 140 billion to converge energy advantage with AI ambitions, supporting Riyadh and Abu Dhabi data center corridors that operate largely outside Western export control regimes.

Regulatory LandscapeAI infrastructure investment and procurement increasingly track AI governance and cross-border technology controls. In the European Union, the AI Act (Regulation (EU) 2024/1689) reaches key applicability in August 2026, raising compliance and documentation requirements that affect how AI systems are developed and deployed across member states and, in turn, where training and inference workloads are hosted. In the United States, Executive Order 14365 (December 2025) established a National Policy Framework for Artificial Intelligence, and NIST continues to publish AI standards and guidance that vendors use to structure assurance, risk management, and testing practices for enterprise and public-sector deployments.

Trade and national-security policy also shapes accelerator availability and sovereign build strategies. U.S. export controls, including the October 2023 restrictions on high-end AI chips to China referenced in the report context, continue to reshape procurement paths, while 2026 policy discussions also include tighter licensing postures and additional conditions tied to security guarantees and domestic investment. In parallel, the U.S. Department of Commerce advanced the American AI Exports Program in 2026, positioning full-stack AI exports (hardware, models, and cybersecurity) as a coordinated initiative that can influence where AI capacity is built and which ecosystems receive preferential access to compliant supply.

Value Chain AnalysisThe AI infrastructure value chain runs from semiconductor materials and wafer fabrication through advanced packaging, HBM supply, server/OEM integration, high-speed networking, storage, power and cooling systems, data center build-out, and finally cloud and enterprise deployment software (orchestration, optimization, and MLOps). In 2025-2026, the most binding upstream constrain

ts are advanced-node foundry capacity and packaging, with CoWoS-type advanced pa ckaging and HBM3E allocations shaping which GPU and accelerator platforms can be delivered at scale and on what timelines. These constraints reinforce hyperscal er behavior described in the report context, including long-term pre-orders, dee per participation in packaging and memory planning, and a continued premium on h igh-bandwidth memory that extends lead times for top-tier accelerators.

Downstream, data center delivery is increasingly limited by site power readiness rather than only server availability. Grid interconnection and upgrades (transf ormers, substations, and transmission and distribution work) introduce long lead times that can delay a meaningful portion of planned capacity, and they interac t with the report context trend toward racks exceeding 100 kilowatts and higher liquid-cooling penetration. As a result, ecosystem coordination expands beyond c hip and server vendors to include utilities, engineering and construction partne rs, colocation providers, and liquid-cooling specialists, with procurement decis ions increasingly bundling compute, networking fabric (InfiniBand NDR and 800G E thernet), and facility-level power and cooling as an integrated bill of material s.

Competitive Landscape

Oligopolistic structure remains evident at the silicon layer, where NVIDIA captu red roughly 80% of 2025 accelerator revenue and maintains a 4-million-developer CUDA moat. Hyperscalers responded by designing ASICs such as Google TPU v6e, AWS Trainium2, and Microsoft Maia 100 that should reach 20% of training hours by 20 26, pressuring NVIDIA’s list prices by up to 30% for bulk orders. AMD’s MI325X l everages 288 GB of HBM3E to undercut H200 on price-per-gigabyte, finding early t raction in Oracle Cloud deployments. Intel’s Gaudi 3 emphasizes Ethernet connect ivity, appealing to enterprises wary of single-vendor ecosystems.

The interconnect layer witnesses consolidation around optical roadmaps, with Bro adcom’s Tomahawk 6 offering 1.6 Tbps switching that aligns with IEEE 802.3df mil estones. Patent filings trend toward chiplets and die-to-die protocols like UCIe , indicating that modular integration could dilute incumbent advantages by short ening time-to-market for challengers. Triton Inference Server and Apache TVM hol d growing mindshare, letting customers switch hardware without wholesale code re writes, thereby eroding proprietary middleware margins. Edge inference, defined by sub-75-watt power budgets, attracts startups such as Tenstorrent and Graphcor e, though deployments remain pilot-scale today.

Environmental scrutiny grows. European carbon mechanisms may add 5-8% to operati ng expenses by 2028, nudging providers toward renewable energy and liquid coolin g. Hyperscalers lead renewable power purchase agreements exceeding 25 GW cumulat ive, putting sustainability on par with latency as a competitive factor. Talent scarcity also shapes rivalry; NVIDIA and AMD opened combined training academies for 30,000 engineers per year to defend ecosystem loyalty. Overall, competitive intensity rises, but architectural lock-in has begun to erode as open standards mature.

AI Infrastructure Industry Leaders* NVIDIA Corporation

* Intel Corporation

* Advanced Micro Devices (AMD)

* Microsoft Corporation

* Amazon Web Services, Inc.

* *Disclaimer: Major Players sorted in no particular order

AI Infrastructure Market Companies Covered in this Report* NVIDIA Corporation
* Intel Corporation
* Advanced Micro Devices (AMD)
* Amazon Web Services, Inc.
* Microsoft Corporation
* Google LLC
* IBM Corporation
* Cisco Systems, Inc.
* Hewlett Packard Enterprise
* Dell Technologies, Inc.
* Samsung Electronics Co., Ltd.
* Micron Technology, Inc.
* Arm Holdings plc
* Synopsys, Inc.
* Baidu, Inc.
* Alibaba Cloud
* Tencent Cloud
* Cerebras Systems
* Graphcore
* Huawei Technologies Co., Ltd.
Read Analysis of AI Infrastructure Companies

Market Opportunities and Future Outlook

Power-ready capacity and energy-integrated site development are emerging as key whitespace for AI infrastructure, as high-density training clusters shift from being typical data center loads to regionally significant grid anchors. Meta's announced expansion of the Hyperion AI campus in Richland Parish, Louisiana, to a 5 GW supercluster (July 2026) illustrates how hyperscale AI projects increasingly tie compute roadmaps to generation and transmission planning, favoring locations with scalable interconnection and permitting pathways. This shift also expands opportunity for liquid cooling and power distribution modernization in facilities that must accommodate racks above 100 kilowatts, aligning with the report context on liquid-cooling adoption and 400 V and 48 V limitations in legacy sites.

A second opportunity is supply chain and architecture diversification across accelerators, packaging, and networking to reduce exposure to constrained GPU allocations and export-control frictions. The report context highlights chronic AI-class GPU short supply through 2026, extended lead times, and the growing role of custom ASICs (for example, Google TPU v6e and AWS Trainium2) alongside Ethernet-based fabrics competing with InfiniBand for large clusters. That environment supports alternative accelerator platforms (FPGA/ASIC options for inference), open optimization stacks such as Triton Inference Server and Apache TVM that improve portability, and OSAT and advanced packaging expansion that relieves CoWoS and HBM-related bottlenecks, enabling more predictable delivery for enterprises and cloud service providers.

Recent Industry Developments in AI Infrastructure Market* July 2026: NVIDIA introduced a partner-oriented AI compute buildout approach that includes capital support and revenue-sharing structures for AI cloud providers building AI factories. The approach broadens NVIDIA's route-to-market beyond direct hardware sales by shaping how third-party clouds finance and monetize GPU capacity, strengthening its ecosystem influence while supply remains constrained.
* June 2026: Microsoft confirmed its Fairwater AI campus in Mount Pleasant, Wisconsin, is operational, linking large-scale NVIDIA Blackwell-class GPU deployments over 800G Ethernet. The deployment underscores the shift toward campus-scale systems designed as unified supercomputers and highlights Ethernet's role as a co

st and vendor-flexible alternative in AI cluster interconnects.
* March 2026: AWS and NVIDIA announced an expanded strategic collaboration to deploy more than 1 million NVIDIA GPUs across AWS regions starting in 2026 through 2027. Multi-year volume commitments of this scale change procurement dynamics for accelerators and associated networking, memory, and cooling, while tightening competitive pressure on alternative silicon and smaller buyers seeking guaranteed capacity.

Table of Contents for AI Infrastructure Industry Report1. INTRODUCTION

* 6.4.20 Huawei Technologies Co., Ltd.

7. MARKET OPPORTUNITIES AND FUTURE OUTLOOK

* 7.1 White-Space and Unmet-Need Assessment

AI Infrastructure Market Report Scope and Research MethodologyMarket Definition and CoverageThis market covers the spending and revenues linked to the infrastructure needed to train and run AI workloads at scale, including specialized compute, supporting storage and memory, and the system software layers that help deploy and optimize these workloads across data centers and cloud environments.

Scope exclusions: We exclude consumer-grade edge devices and general IT services that are not directly tied to accelerating AI training or inference workloads.

Segments Covered in This Report* By Offering* Hardware* Processor
* Storage
* Memory

* Software* System Optimisation
* AI Middleware and MLOps

* By Deployment* On-Premise
* Cloud

* By End User* Enterprises
* Government and Defence
* Cloud Service Providers

* By Processor Architecture* CPU
* GPU
* FPGA/ASIC (TPU, Inferentia, Gaudi, Cerebras)
* Other Processor Architectures

* By Geography* North America* United States
* Canada
* Mexico

* South America* Brazil
* Argentina
* Rest of South America

* Europe* United Kingdom
* Germany
* France
* Sweden
* Rest of Europe

* Asia Pacific* China
* Japan
* India
* Australia
* South Korea
* Rest of Asia Pacific

* Middle East* Saudi Arabia
* United Arab Emirates
* Turkey
* Rest of Middle East

* Africa* South Africa
* Nigeria
* Rest of Africa

Data Sources, Market Sizing, and ValidationDesk Research

Desk research was used to build the factual base for the model and to keep the s cope consistent across regions and buying channels. We relied on public sources such as the US Census Bureau and US International Trade Commission trade data, B ureau of Economic Analysis ICT investment series, OECD digital economy indicator s, and International Telecommunication Union connectivity and data traffic stati stics.

To keep assumptions grounded, we also reviewed company filings and investor pres entations, standards and technical notes from bodies such as IEEE and NIST, and open academic literature on AI training compute trends and data center efficienc y. In several steps, we also used paid subscriptions focused on company financia ls and intelligence, patents, and shipment-level import and export records to fl ag capacity signals and validate supplier and buyer coverage. The sources listed here are illustrative, and additional public references were also used during d ata collection, clarification, and cross-checking.

Primary Interviews and Surveys

Primary interviews and structured surveys were used to pressure-test the demand pool and to confirm what buyers count as AI infrastructure in real budgets, incl uding cloud, on-premises, and hybrid deployments. We spoke with infrastructure s uppliers, cloud and colocation stakeholders, and enterprise and public-sector bu yers across Americas, EMEA, and APAC so that regional deployment patterns and pr icing shifts could be reflected in the final model.

Distribution of primary research fieldwork respondents

Company type | Respondent position | Region |
Top tier: 33% | CXOs: 14% | APAC: 47% |
Mid tier: 53% | Functional/Unit leaders: 41% | EMEA: 30% |
Smaller Players: 14% | Managers: 45% | Americas: 23% |

Market-Sizing & ForecastingThe core sizing starts from a top-down build where pu blic cloud and data-center investment signals are translated into an AI-ready in frastructure demand pool, then filtered by AI workload intensity and typical har dware and software attach rates. After that shape is built, we corroborate the t otals with selective bottom-up approximations, such as sampled ASP times unit vo lumes for accelerator-rich servers and related system software, plus channel che cks on deployment mix.

Inputs treated as sizing fingerprints included accelerator server adoption rates , GPU and AI accelerator availability cycles, high-bandwidth memory penetration, data-center buildouts, and power and cooling constraints, along with cloud vers us on-premises workload placement. When the bottom-up checks exposed gaps, the m odel used conservative proxy logic, for example by applying regional deployment shares from interviews and then using a consistent price progression for compara ble configurations.

For forecasting, scenario analysis was used so the outlook could handle uncertai nties such as supply lead times, export controls, and faster changes in model tr aining intensity. Scenario weights were adjusted using expert input on expected capacity additions, typical procurement timing, and how quickly software layers standardize across deployments.

Data Validation & Update CycleEach step is reviewed through triangulation checks where the sized totals are compared with independent signals, including data-ce

nter capex trends, accelerator shipment momentum, and cloud infrastructure spend ing direction. Outliers are investigated through variance checks by region and d eployment type, and then reviewed again in an internal analyst pass before sign-off.

Reports are refreshed annually, with interim updates triggered when material eve nts change pricing, supply, or deployment patterns, for example major capacity e xpansions or policy shifts that affect availability. Before delivery, a final re view is performed so the outputs reflect the most recent information and the ass umptions remain traceable to the model inputs.

Mordor Intelligence's AI Infrastructure Market Size Compared Against Other Publi shed EstimatesPublished market size numbers for AI infrastructure do not always match because the boundaries can shift, especially around what gets counted as i nfrastructure versus adjacent software and services. Differences also come from the starting year, the treatment of cloud spend versus on-premises purchases, an d how pricing is updated during fast product cycles.

Data-center capex direction, accelerator-rich server adoption signals, and cloud infrastructure spending checks are used to keep Mordor Intelligence's 2026 tota l focused on AI training and inference infrastructure only, rather than broader AI platform spend. Some publishers also mix spending measures with revenue measu res, apply aggressive price escalation, or include a wider stack that counts man aged services, which can push the total beyond what most buyers consider infrast ructure.

Benchmark comparison

Source	Market Size	Gaps in Research Methodology
Mordor Intelligence	USD 101.17 B (2026)	
Global Consultancy A	USD 487.00 B (2026)	Uses a spending tracker scope that concentrates on worldwide server and storage outlays and can capture hyperscaler capex surges, which is not the same as a revenue-defined infrastructure market and may include broader procurement buckets.
Industry Research Group B	USD 135.81 B (2024)	Starts from a 2024 base year a nd includes a wider offering mix and deployment definitions, which can shift tot als when price progression and attach assumptions are applied during a rapid acc elerator cycle.

The table shows that the spread is mainly explained by scope and measurement cho ices, not just growth expectations. When the market is kept aligned to training and inference infrastructure, and the cross-checks are anchored in observable bu ildout and adoption signals, the resulting value stays easier to reconcile and r epeat across updates.

Key Questions Answered in the ReportWhat is the projected value of the AI infras tructure market in 2031? The market is forecast to reach USD 202.48 billion by 2 031, expanding at a 14.89% CAGR over the period.

Which region will grow the fastest in AI infrastructure spending? Asia Pacific i s expected to post a 16.44% CAGR through 2031, spurred by large semiconductor fu nds in China and hyperscaler expansions in India.

How dominant are GPUs in current AI accelerator revenue? GPUs held 88.82% of pro cessor architecture revenue in 2025, although ASIC and FPGA devices are now grow ing faster for inference workloads.

Why are liquid-cooled data centers gaining momentum? Rising rack densities above 100 kilowatts and stricter carbon pricing make liquid cooling essential, reduci ng facility energy use by up to 40%.

How do export controls influence AI infrastructure strategy? U.S. restrictions o
n high-end GPU shipments to China drive parallel investment in domestic chips, l
eading to divergent technology stacks and higher compliance costs for multinatio
nal deployments.

Related Reports Explore More ReportsIndustry Trends in Agentic AIAI In IoT Marke
t SizeAnalysis of AI Software Market In Legal IndustryCloud AI IndustryEnterpris
e AI Market ReportMarket Data for Explainable AIArtificial Intelligence Market F
orecastsIndustrial AI Software MarketIntelligent Virtual Assistant (IVA) Industr
yEdge AI Accelerators Market