+357 96 030922
info@marketpublishers.com

AI Inference Market by Compute (GPU, CPU, FPGA), Memory (DDR, HBM), Network (NIC/Network Adapters, Interconnect), Deployment (On-premises, Cloud, Edge), Application (Generative AI, Machine Learning, NLP, Computer Vision) - Global Forecast to 2030

https://marketpublishers.com/r/A5047029838CEN.html
Date: February 2025
Pages: 366
Price: US$ 4,950.00 (Single User License)
ID: A5047029838CEN

Abstracts

The AI Inference market is expected to be worth USD 106.15 billion in 2025 and is estimated to reach USD 254.98 billion by 2030, growing at a CAGR of 19.2% between 2025 and 2030. The AI inference market is being driven by the exponential increase in data generation, fueled by the widespread use of connected devices, social media platforms, and digital transformation initiatives. This massive influx of data necessitates efficient inference systems to extract meaningful insights in real time, enabling businesses to stay competitive and responsive. Additionally, the growing emphasis on personalized user experiences, such as recommendation systems in e-commerce and content platforms, has heightened the demand for AI inference to deliver tailored outcomes swiftly and accurately. Furthermore, regulatory and compliance requirements in sectors like healthcare and finance are pushing organizations to adopt AI inference for tasks such as fraud detection, risk assessment, and diagnostics, ensuring both accuracy and scalability.

“Machine Learning segment holds highest market share in 2024.”

Machine Learning holds high market share in the AI inference market, which is driven by the expanding use of ML applications across various industries. Machine learning models, especially deep learning and reinforcement learning algorithms, require extensive computational resources to train and deploy effectively. This requirement of robust infrastructure, such as high performance GPUs, TPUs and dedicated AI accelerators, becomes essential as organizations continue to bring in machine learning

+357 96 030922
info@marketpublishers.com

for prediction analytics, recommendation engines, autonomous systems, etc. Technology companies such as Google Cloud (USA), Amazon Web Services (USA), and Microsoft Azure (USA) are enhancing their AI products to accommodate more complex ML models and providing solutions such as TPU V4 and NVIDIA'S A100 GPUs. Recent advancements such as Gcore's introduction of 'Inference at the Edge' in June 2024 accelerate this trend even further through provision of nanosecond-order low-latency AI processing utilizing high-performance, strategically located nodes equipped with NVIDIA L40S GPUs. These platforms support both fundamental and custom machine learning models, including popular open-source foundation models like LLAMA Pro 8B, Mistral 7B, and Stable-Diffusion XL, paving the way towards versatility and flexibility for various scenarios. This alliance of scalability, accessibility, and state-of-the-art infrastructure reinforces machine learning's dominance in the AI inference market.

”Enterprises is projected to grow at a high CAGR of AI Inference market during the forecasted timeline”

The enterprise segment will have the highest growth rate in the AI Inference market. Enterprises have widely adopted AI solutions for better operational efficiency, offer personalized customer experience and to drive innovation. Enterprises have resources and infrastructure to deploy large-scale AI models in domains such as customer service, supply chain optimization, and predictive analytics. Healthcare enterprise use AI for medical imaging and diagnostics, financial organizations for fraud and risk detection, and retailer for AI-based recommendation system and inventory management. This growth is further propelled by rise in advancements in enterprise-focused AI platforms that simplify the deployment and scale AI applications. For instance, In May 2024, Nutanix (US) collaborated with NVIDIA Corporation (US) in order to boost adoption for generative AI . This integration of Nutanix's GPT-in-a-Box 2.0 with NVIDIA'S NIM inference microservices will enable enterprises to deploy scalable, secure, and high-performance GenAI applications both centrally and at the edge. With its platform, Nutanix simplifies the deployment of AI models and reduces the need for specialized AI expertise and empowers businesses to implement AI strategies. These innovations highlight the increasing rate at which enterprises are investing in AI inference for competitive advantages and operational improvement.

“Asia Pacific is expected to hold high CAGR in during the forecast period.”

The AI inference market in Asia Pacific will grow at a high CAGR in the forecast period. Asia Pacific has seen remarkable progress in AI research, development, and deployment. Countries like China, Japan, South Korea, and Singapore are making

+357 96 030922
info@marketpublishers.com

substantial investments in AI research and infrastructure. Strong collaborations among academia, industry and government in these countries have resulted in innovations in machine learning, natural language processing, computer vision, and robotics. For instance, In October 2024, Nvidia Corporation (US) made strategic plans and collaborations in India, such as partnerships with Yotta, E2E Networks, and Netweb, to promote the use of AI technologies and create AI 'factories' specific to the Indian market. These collaborations are aimed at accelerating AI inference with Nvidia's high-end GPUs, software, and networking features, including Yotta's Shakti Cloud providing Nvidia Inference Microservices (NIM) and E2E for access to Nvidia's H200 GPUs. Netweb's manufacturing of Tyrone servers based on Nvidia's MGX reference design also complements these efforts. These developments will substantially increase demand for AI inference solutions in India by allowing companies to handle sophisticated workloads, drive AI adoption in Asia Pacific, and assist startups with innovative accelerator programs.

Extensive primary interviews were conducted with key industry experts in the AI Inference market space to determine and verify the market size for various segments and subsegments gathered through secondary research. The break-up of primary participants for the report has been shown below: The study contains insights from various industry experts, from component suppliers to Tier 1 companies and OEMs. The break-up of the primaries is as follows:

By Company Type: Tier 1 – 40%, Tier 2 – 25%, and Tier 3 – 35%

By Designation: C-level Executives – 50%, Directors – 20%, and Others – 30%

By Region: North America – 40%, Europe – 20%, Asia Pacific – 30%, and RoW – 10%

The report profiles key players in the AI Inference market with their respective market ranking analysis. Prominent players profiled in this report are NVIDIA Corporation (US), Advanced Micro Devices, Inc. (US), Intel Corporation (US), SK HYNIX INC. (South Korea), SAMSUNG (South Korea), Micron Technology, Inc. (US), Apple Inc. (US), Qualcomm Technologies, Inc. (US), Huawei Technologies Co., Ltd. (China), Google (US), Amazon Web Services, Inc. (US), Tesla (US), Microsoft (US), Meta (US), T-Head (China), Graphcore (UK), and Cerebras (US), among others.

Apart from this, Mythic (US), Blaize (US), Groq, Inc. (US), HAILO TECHNOLOGIES

+357 96 030922
info@marketpublishers.com

LTD (Israel), SiMa Technologies, Inc. (US), Kneron, Inc. (US), Tenstorrent (Canada), SambaNova Systems, Inc. (US), SAPEON Inc. (US), Rebellions Inc. (South Korea), Shanghai BiRen Technology Co., Ltd. (China) are among a few emerging companies in the AI Inference market.

Research Coverage: This research report categorizes the AI Inference market based on compute, memory, network, deployment, application, end user, and region. The report describes the major drivers, restraints, challenges, and opportunities pertaining to the AI Inference market and forecasts the same till 2030. Apart from these, the report also consists of leadership mapping and analysis of all the companies included in the AI Inference ecosystem.

Key Benefits of Buying the Report The report will help the market leaders/new entrants in this market with information on the closest approximations of the revenue numbers for the overall AI Inference market and the subsegments. This report will help stakeholders understand the competitive landscape and gain more insights to position their businesses better and plan suitable go-to-market strategies. The report also helps stakeholders understand the pulse of the market and provides them with information on key market drivers, restraints, challenges, and opportunities.

The report provides insights on the following pointers:

Analysis of key drivers (Growing need for real-time processing at edge devices, Advanced cloud platforms offering specialized AI inference services, and Enhanced GPU capabilities for inference tasks) influencing the growth of the AI inference market.

Product Development/Innovation: Detailed insights on upcoming technologies, research & development activities, and new product & service launches in the AI inference market.

Market Development: Comprehensive information about lucrative markets – the report analysis the AI inference market across varied regions

Market Diversification: Exhaustive information about new products & services, untapped geographies, recent developments, and investments in the AI inference market

Competitive Assessment: In-depth assessment of market shares, growth

+357 96 030922
info@marketpublishers.com

strategies, and service offerings of leading players like NVIDIA Corporation (US), Advanced Micro Devices, Inc. (US), Intel Corporation (US), SK HYNIX INC. (South Korea), SAMSUNG (South Korea), among others in the AI inference market.

+357 96 030922
info@marketpublishers.com

Contents

+357 96 030922
info@marketpublishers.com

+357 96 030922
info@marketpublishers.com

+357 96 030922
info@marketpublishers.com

6 AI INFERENCE MARKET, BY COMPUTE

7 AI INFERENCE MARKET, BY MEMORY

+357 96 030922
info@marketpublishers.com

+357 96 030922
info@marketpublishers.com

+357 96 030922
info@marketpublishers.com

+357 96 030922
info@marketpublishers.com

13 COMPETITIVE LANDSCAPE

+357 96 030922
info@marketpublishers.com

+357 96 030922
info@marketpublishers.com

+357 96 030922
info@marketpublishers.com

+357 96 030922
info@marketpublishers.com

+357 96 030922
info@marketpublishers.com

15 APPENDIX

+357 96 030922
info@marketpublishers.com

I would like to order

Product name: AI Inference Market by Compute (GPU, CPU, FPGA), Memory (DDR, HBM), Network (NIC/Network Adapters, Interconnect), Deployment (On-premises, Cloud, Edge), Application (Generative AI, Machine Learning, NLP, Computer Vision) - Global Forecast to 2030

Product link: https://marketpublishers.com/r/A5047029838CEN.html

Price: US$ 4,950.00 (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/A5047029838CEN.html