+357 96 030922
info@marketpublishers.com

Global AI Inference Market Size Study, By Compute (GPU, CPU, FPGA), By Memory (DDR, HBM), By Network (NIC/Network Adapters, Interconnect), By Deployment (On-premises, Cloud, Edge), By Application (Generative AI, Machine Learning, NLP, Computer Vision), and Regional Forecasts 2022-2032

https://marketpublishers.com/r/GE954415DD53EN.html
Date: March 2025
Pages: 285
Price: US$ 3,750.00 (Single User License)
ID: GE954415DD53EN

Abstracts

The global AI inference market was valued at USD 74.71 billion by 2023 and is projected to reach USD 362.97 billion by 2032, growing at a compound annual growth rate (CAGR) of 19.2% from 2024 to 2032. The increasing adoption of AI inference solutions across industries is driven by advancements in specialized AI inference chips and hardware that improve real-time processing, efficiency, and scalability.

With industries increasingly deploying AI models for applications such as autonomous driving, healthcare diagnostics, smart assistants, and data center optimizations, the demand for high-performance AI inference processors has surged. These chips enable faster, more energy-efficient AI inference processes, which is particularly critical in edge computing and cloud-based AI systems.

Major market players such as NVIDIA Corporation (US), Advanced Micro Devices, Inc. (US), Intel Corporation (US), SK HYNIX INC. (South Korea), and SAMSUNG (South Korea) are leading innovation in AI inference technology. Companies are expanding their global footprint through product launches, strategic alliances, acquisitions, and research collaborations to enhance their AI inference portfolios.

Advancements in AI Inference Hardware Driving Market Growth

+357 96 030922
info@marketpublishers.com

The rapid evolution of AI inference chips has enabled businesses to optimize machine learning and AI model execution, particularly in real-time applications. Key developments include dedicated AI inference processors, tensor processing units (TPUs), field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs). These hardware solutions are designed to accelerate AI inference performance, enabling enterprises to deploy scalable and power-efficient AI systems.

For instance, in October 2024, Advanced Micro Devices, Inc. (US) launched the 5th Gen AMD EPYC processors, optimized for AI inference, cloud computing, and high-performance workloads. The EPYC 9005 series provides enhanced GPU acceleration and maximized per-server performance, making it ideal for data center AI workloads.

Similarly, in October 2024, Intel Corporation (US) and Inflection AI (US) announced a strategic collaboration to accelerate AI inference adoption in enterprises through the launch of Inflection for Enterprise. Powered by Intel Gaudi processors and Intel Tiber AI Cloud, this system provides customizable AI solutions for enterprise AI workloads.

Market Expansion Fueled by Edge AI and On-Premises AI Inference Solutions

As AI applications continue to evolve, the focus is shifting toward low-latency, high-speed AI inference processing at the edge. Edge AI inference solutions are critical in autonomous systems, IoT devices, real-time analytics, and smart surveillance. The growing adoption of edge AI inference hardware enables businesses to reduce reliance on cloud-based inference models, providing faster decision-making capabilities while maintaining data privacy and security.

Furthermore, on-premises AI inference solutions are gaining traction as enterprises seek cost-effective, high-performance AI models for mission-critical applications. Cloud-based inference solutions continue to dominate, driven by the scalability and processing power offered by hyperscale cloud providers such as Google, Amazon Web Services (AWS), and Microsoft Azure.

Major Market Players Included in This Report:

NVIDIA Corporation (US)

Advanced Micro Devices, Inc. (US)

Intel Corporation (US)

+357 96 030922
info@marketpublishers.com

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)

Cerebras (US)

The Detailed Segments and Sub-Segments of the Market Are Explained Below:

By Compute:

GPU

CPU

FPGA

+357 96 030922
info@marketpublishers.com

By Memory:

- DDR
- HBM

By Network:

- NIC/Network Adapters
- Interconnect

By Deployment:

- On-Premises
- Cloud
- Edge

By Application:

- Generative AI
- Machine Learning
- Natural Language Processing (NLP)
- Computer Vision

By Region:

North America

+357 96 030922
info@marketpublishers.com

- U.S.
- Canada
- Mexico

Europe

- UK
- Germany
- France
- Italy
- Spain

Asia-Pacific

- China
- Japan
- India
- South Korea
- Australia

Latin America

- Brazil
- Argentina

+357 96 030922
info@marketpublishers.com

Middle East & Africa

- Saudi Arabia
- UAE
- South Africa

Years Considered for the Study:

- Historical Year – 2022
- Base Year – 2023
- Forecast Period – 2024 to 2032

Key Takeaways:

- Market Estimates & Forecast for 10 years (2022-2032)
- Annualized revenue and segment-wise breakdowns
- Regional-level market insights
- Competitive landscape analysis of major players
- Emerging trends in AI inference hardware and software
- Investment opportunities in AI inference processors and AI-optimized memory solutions

+357 96 030922
info@marketpublishers.com

Contents

+357 96 030922
info@marketpublishers.com

CHAPTER 3. GLOBAL AI INFERENCE MARKET DYNAMICS

CHAPTER 4. GLOBAL AI INFERENCE MARKET INDUSTRY ANALYSIS

CHAPTER 5. GLOBAL AI INFERENCE MARKET SIZE & FORECASTS BY COMPUTE 2022-2032

+357 96 030922
info@marketpublishers.com

+357 96 030922
info@marketpublishers.com

+357 96 030922
info@marketpublishers.com

CHAPTER 12. RESEARCH PROCESS

+357 96 030922
info@marketpublishers.com

List Of Tables

LIST OF TABLES

(This list is not complete; the final report contains more than 100 tables. The list may be updated in the final deliverable.)

+357 96 030922
info@marketpublishers.com

List Of Figures

LIST OF FIGURES

(This list is not complete; the final report contains more than 50 figures. The list may be updated in the final deliverable.)

+357 96 030922
info@marketpublishers.com

I would like to order

Product name: Global AI Inference Market Size Study, By Compute (GPU, CPU, FPGA), By Memory (DDR, HBM), By Network (NIC/Network Adapters, Interconnect), By Deployment (On-premises, Cloud, Edge), By Application (Generative AI, Machine Learning, NLP, Computer Vision), and Regional Forecasts 2022-2032

Product link: https://marketpublishers.com/r/GE954415DD53EN.html

Price: US$ 3,750.00 (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/GE954415DD53EN.html