

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

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

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

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

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.

Contents

1 INTRODUCTION

1.1 STUDY OBJECTIVES

1.2 MARKET DEFINITION

1.3 STUDY SCOPE

1.3.1 MARKETS COVERED AND REGIONAL SCOPE

1.3.2 INCLUSIONS AND EXCLUSIONS

1.3.3 YEARS CONSIDERED

1.4 CURRENCY CONSIDERED

1.5 UNIT CONSIDERED

1.6 LIMITATIONS

1.7 STAKEHOLDERS

2 RESEARCH METHODOLOGY

2.1 RESEARCH DATA

2.1.1 SECONDARY AND PRIMARY RESEARCH

2.1.2 SECONDARY DATA

2.1.2.1 List of key secondary sources

2.1.2.2 Key data from secondary sources

2.1.3 PRIMARY DATA

2.1.3.1 List of primary interview participants

2.1.3.2 Breakdown of primaries

2.1.3.3 Key data from primary sources

2.1.3.4 Key industry insights

2.2 MARKET SIZE ESTIMATION METHODOLOGY

2.2.1 BOTTOM-UP APPROACH

2.2.1.1 Approach to arrive at market size using bottom-up analysis
(demand side)

2.2.2 TOP-DOWN APPROACH

2.2.2.1 Approach to arrive at market size using top-down analysis
(supply side)

2.3 DATA TRIANGULATION

2.4 RESEARCH ASSUMPTIONS

2.5 RISK ANALYSIS

2.6 RESEARCH LIMITATIONS

3 EXECUTIVE SUMMARY

4 PREMIUM INSIGHTS

4.1 ATTRACTIVE OPPORTUNITIES FOR PLAYERS IN AI INFERENCE MARKET

4.2 AI INFERENCE MARKET, BY COMPUTE

4.3 AI INFERENCE MARKET, BY MEMORY

4.4 AI INFERENCE MARKET, BY NETWORK

4.5 AI INFERENCE MARKET, BY APPLICATION

4.6 AI INFERENCE MARKET, BY END USER

4.7 AI INFRASTRUCTURE MARKET, BY REGION

4.8 AI INFERENCE MARKET, BY COUNTRY

5 MARKET OVERVIEW

5.1 INTRODUCTION

5.2 MARKET DYNAMICS

5.2.1 DRIVERS

5.2.1.1 Growing demand for real-time processing on edge devices

5.2.1.2 Growth of advanced cloud platforms offering specialized AI inference services

5.2.1.3 Enhanced GPU capabilities for inference tasks

5.2.2 RESTRAINTS

5.2.2.1 Computational workload and high power consumption

5.2.2.2 Shortage of skilled workforce

5.2.3 OPPORTUNITIES

5.2.3.1 Growth of AI-enabled healthcare and diagnostics

5.2.3.2 Advancements in natural language processing for improved customer experience

5.2.3.3 Increasing demand for real-time data processing and analytics

5.2.4 CHALLENGES

5.2.4.1 Data privacy concerns

5.2.4.2 Supply chain disruptions

5.3 TRENDS/DISRUPTIONS IMPACTING CUSTOMER BUSINESS

5.4 PRICING ANALYSIS

5.4.1 INDICATIVE PRICING OF KEY PLAYERS, BY COMPUTE

5.4.2 AVERAGE SELLING PRICE TREND, BY REGION

5.5 VALUE CHAIN ANALYSIS

5.6 ECOSYSTEM ANALYSIS

5.7 INVESTMENT AND FUNDING SCENARIO

5.8 TECHNOLOGY ANALYSIS

5.8.1 KEY TECHNOLOGIES

- 5.8.1.1 GenAI workload
- 5.8.1.2 High bandwidth memory (HBM)
- 5.8.1.3 High-performance computing (HPC)

5.8.2 COMPLEMENTARY TECHNOLOGIES

- 5.8.2.1 High-speed interconnects
- 5.8.2.2 Edge computing infrastructure
- 5.8.2.3 Data center power management and cooling system

5.8.3 ADJACENT TECHNOLOGIES

- 5.8.3.1 Cloud AI services
- 5.8.3.2 AI development frameworks

5.9 PATENT ANALYSIS

5.10 TRADE ANALYSIS

- 5.10.1 IMPORT SCENARIO (HS CODE 854231)
- 5.10.2 EXPORT SCENARIO (HS CODE 854231)

5.11 KEY CONFERENCES AND EVENTS, 2025–2026

5.12 CASE STUDY ANALYSIS

- 5.12.1 AI-POWERED RADIATION THERAPY OPTIMIZATION WITH INTEL CORPORATION AND SIEMENS HEALTHINEERS
- 5.12.2 ARTIFICIAL INTELLIGENCE ACCELERATES DARK MATTER SEARCH WITH ADVANCED MICRO DEVICES, INC. FPGAS
- 5.12.3 SERVING INFERENCE FOR LLMS: A CASE STUDY WITH NVIDIA TRITON INFERENCE SERVER AND ELEUTHER AI
- 5.12.4 FINCH COMPUTING REDUCES INFERENCE COSTS USING AWS INFERENCE FOR LANGUAGE TRANSLATION

5.13 REGULATORY LANDSCAPE

- 5.13.1 REGULATORY BODIES, GOVERNMENT AGENCIES, AND OTHER ORGANIZATIONS
- 5.13.2 STANDARDS

5.14 PORTER'S FIVE FORCES ANALYSIS

- 5.14.1 THREAT OF NEW ENTRANTS
- 5.14.2 THREAT OF SUBSTITUTES
- 5.14.3 BARGAINING POWER OF SUPPLIERS
- 5.14.4 BARGAINING POWER OF BUYERS
- 5.14.5 INTENSITY OF COMPETITIVE RIVALRY

5.15 KEY STAKEHOLDERS AND BUYING CRITERIA

- 5.15.1 KEY STAKEHOLDERS IN BUYING PROCESS
- 5.15.2 BUYING CRITERIA

6 AI INFERENCE MARKET, BY COMPUTE

6.1 INTRODUCTION

6.2 GPU

6.2.1 ABILITY TO HANDLE AI WORKLOADS AND PROCESS VAST DATA VOLUMES TO BOOST ADOPTION

6.3 CPU

6.3.1 RISING DEMAND FOR VERSATILE AND GENERAL-PURPOSE AI PROCESSING TO BOOST MARKET GROWTH

6.4 FPGA

6.4.1 INCREASING NEED FOR FLEXIBILITY AND CUSTOMIZATION FOR AI WORKLOADS TO SPUR DEMAND

6.5 NPU

6.5.1 RISING DEMAND FOR HIGH-END SMARTPHONES TO DRIVE SEGMENTAL GROWTH

6.6 TPU

6.6.1 NEED FOR FASTER PROCESSING IN AI RESEARCH AND APPLICATION DEVELOPMENT TO BOOST DEMAND

6.7 FSD

6.7.1 DEMAND FOR HIGH-PERFORMANCE, ENERGY-EFFICIENT AI PROCESSING IN AUTONOMOUS VEHICLES TO FUEL ADOPTION

6.8 INFERENTIA

6.8.1 ABILITY TO TRAIN COMPLEX AI AND DEEP LEARNING MODELS TO DRIVE ADOPTION

6.9 T-HEAD

6.9.1 RISING DEMAND FOR CUSTOMIZED, HIGH-PERFORMANCE AI CHIPS ACROSS CHINESE DATA CENTERS TO STIMULATE MARKET GROWTH

6.10 MTIA

6.10.1 META'S EXPANSION INTO AR, VR, AND METAVERSE TO FUEL MARKET GROWTH

6.11 LPU

6.11.1 INCREASING NEED TO HANDLE COMPLEX NLP AND LANGUAGE-BASED AI TASKS TO ACCELERATE DEMAND

6.12 OTHER ASICS

7 AI INFERENCE MARKET, BY MEMORY

7.1 INTRODUCTION

7.2 DDR

7.2.1 RISING ADOPTION OF AI-ENABLED CPUS IN DATA CENTERS TO SUPPORT MARKET GROWTH

7.3 HBM

7.3.1 ELEVATING NEED FOR HIGH THROUGHPUT IN DATA-INTENSIVE AI TASKS TO FUEL MARKET GROWTH

8 AI INFERENCE MARKET, BY NETWORK

8.1 INTRODUCTION

8.2 NIC/NETWORK ADAPTERS

8.2.1 INFINIBAND

8.2.1.1 Growing utilization of HPC and AI models to minimize latency and maximize throughput to boost segmental growth

8.2.2 ETHERNET

8.2.2.1 Rising demand for scalable and cost-effective networking solutions to propel growth

8.3 INTERCONNECTS

8.3.1 GROWING COMPLEXITY OF AI MODELS REQUIRING HIGH-BANDWIDTH DATA PATHS TO FUEL DEMAND

9 AI INFERENCE MARKET, BY DEPLOYMENT

9.1 INTRODUCTION

9.2 ON-PREMISES

9.2.1 GROWING DATA PRIVACY CONCERNS TO DRIVE MARKET

9.3 CLOUD

9.3.1 ABILITY TO SCALE RESOURCES ON DEMAND TO BOOST GROWTH

9.4 EDGE

9.4.1 INCREASING APPLICATION IN HEALTHCARE, AUTOMOTIVE, AND INDUSTRIAL AUTOMATION TO FOSTER MARKET GROWTH

10 AI INFERENCE MARKET, BY APPLICATION

10.1 INTRODUCTION

10.2 GENERATIVE AI

10.2.1 RULE-BASED MODELS

10.2.1.1 Integration with ML and deep learning to offer lucrative growth opportunities

10.2.2 STATISTICAL MODELS

10.2.2.1 Growing application in finance, economics, and healthcare sectors to fuel market growth

10.2.3 DEEP LEARNING

10.2.3.1 Ability to advance AI technologies to boost demand

10.2.4 GENERATIVE ADVERSARIAL NETWORKS (GANS)

10.2.4.1 Need to handle large-scale data to fuel market growth

10.2.5 AUTOENCODERS

10.2.5.1 Increasing use in data processing, anomaly detection, and feature extraction to accelerate demand

10.2.6 CONVOLUTIONAL NEURAL NETWORKS (CNNs)

10.2.6.1 Rising number of autonomous vehicles and smart cities to drive market

10.2.7 TRANSFORMER MODELS

10.2.7.1 Growing popularity of GPT models and BERT to offer lucrative growth opportunities

10.3 MACHINE LEARNING

10.3.1 RISING APPLICATION FOR REAL-TIME DECISION-MAKING AND DATA ANALYSIS TO FOSTER MARKET GROWTH

10.4 NATURAL LANGUAGE PROCESSING

10.4.1 GROWING DEMAND FOR SENTIMENT ANALYSIS, LANGUAGE TRANSLATION, AND SPEECH RECOGNITION TO DRIVE MARKET

10.5 COMPUTER VISION

10.5.1 ESCALATING NEED FOR ADVANCED PROCESSING CAPABILITIES TO BOOST DEMAND

11 AI INFERENCE MARKET, BY END USER

11.1 INTRODUCTION

11.2 CONSUMER

11.2.1 GROWING ADOPTION OF AI-ENABLED PERSONAL DEVICES TO PROPEL MARKET

11.3 CLOUD SERVICE PROVIDERS

11.3.1 SURGING AI WORKLOADS AND CLOUD ADOPTION TO STIMULATE MARKET GROWTH

11.4 ENTERPRISES

11.4.1 HEALTHCARE

11.4.1.1 Growing demand for personalized treatment to fuel market growth

11.4.2 BFSI

11.4.2.1 Rising focus on enhancing security and improving customer services to foster market growth

11.4.3 AUTOMOTIVE

11.4.3.1 Growing focus on safe and enhanced driving experiences to fuel demand

11.4.4 RETAIL & E-COMMERCE

11.4.4.1 Rapid shift toward data-centric models to enhance customer engagement to accelerate demand

11.4.5 MEDIA & ENTERTAINMENT

11.4.5.1 Rising demand for content recommendation engines and interactive media experiences to foster market growth

11.4.6 OTHERS

11.5 GOVERNMENT ORGANIZATIONS

11.5.1 GROWING NEED TO ENHANCE PUBLIC SAFETY AND SECURITY TO OFFER LUCRATIVE GROWTH OPPORTUNITIES

12 AI INFERENCE MARKET, BY REGION

12.1 INTRODUCTION

12.2 NORTH AMERICA

12.2.1 MACROECONOMIC OUTLOOK FOR NORTH AMERICA

12.2.2 US

12.2.2.1 Presence of established AI inference manufacturers to drive market

12.2.3 CANADA

12.2.3.1 Growing emphasis on commercializing AI to offer lucrative growth opportunities

12.2.4 MEXICO

12.2.4.1 Rapid digital transformation and surging adoption of cloud computing to fuel market growth

12.3 EUROPE

12.3.1 MACROECONOMIC OUTLOOK FOR EUROPE

12.3.2 UK

12.3.2.1 Growing investments in data center infrastructure to boost demand

12.3.3 GERMANY

12.3.3.1 Increasing adoption of smart technologies to boost manufacturing to drive market

12.3.4 FRANCE

12.3.4.1 Rising government-led initiatives to strengthen AI technology to fuel market growth

12.3.5 ITALY

12.3.5.1 Rising emphasis on developing digital infrastructure to offer lucrative growth opportunities

12.3.6 SPAIN

12.3.6.1 Rapid adoption of cloud computing to accelerate demand

12.3.7 REST OF EUROPE

12.4 ASIA PACIFIC

12.4.1 MACROECONOMIC OUTLOOK FOR ASIA PACIFIC

12.4.2 CHINA

12.4.2.1 Proliferation of IoT devices to drive market

12.4.3 JAPAN

12.4.3.1 Rising investments to boost cloud infrastructure to foster market growth

12.4.4 INDIA

12.4.4.1 Government-led initiatives to boost AI infrastructure to offer lucrative growth opportunities

12.4.5 SOUTH KOREA

12.4.5.1 Thriving semiconductor industry to drive market

12.4.6 REST OF ASIA PACIFIC

12.5 ROW

12.5.1 MACROECONOMIC OUTLOOK FOR ROW

12.5.2 MIDDLE EAST

12.5.2.1 Growing emphasis on digital transformation and technological innovation to drive market

12.5.2.2 GCC

12.5.2.3 Rest of Middle East

12.5.3 AFRICA

12.5.3.1 Growing need for managing advanced data processing requirements to fuel market growth

12.5.4 SOUTH AMERICA

12.5.4.1 Growing need for flexible and secure cloud storage solutions to accelerate demand

13 COMPETITIVE LANDSCAPE

13.1 INTRODUCTION

13.2 KEY PLAYER STRATEGIES/RIGHT TO WIN, 2020–2024

13.3 REVENUE ANALYSIS, 2022–2024

13.4 MARKET SHARE ANALYSIS, 2024

13.5 COMPANY VALUATION AND FINANCIAL METRICS

13.6 BRAND/PRODUCT COMPARISON

13.7 COMPANY EVALUATION MATRIX: KEY PLAYERS, 2024

13.7.1 STARS

- 13.7.2 EMERGING LEADERS
- 13.7.3 PERVASIVE PLAYERS
- 13.7.4 PARTICIPANTS
- 13.7.5 COMPANY FOOTPRINT: KEY PLAYERS, 2024

- 13.7.5.1 Company footprint
- 13.7.5.2 Compute footprint
- 13.7.5.3 Memory footprint
- 13.7.5.4 Network footprint
- 13.7.5.5 Deployment footprint
- 13.7.5.6 Application footprint
- 13.7.5.7 End user footprint
- 13.7.5.8 Region footprint

13.8 COMPANY EVALUATION MATRIX: STARTUPS/SMES, 2024

- 13.8.1 PROGRESSIVE COMPANIES
- 13.8.2 RESPONSIVE COMPANIES
- 13.8.3 DYNAMIC COMPANIES
- 13.8.4 STARTING BLOCKS
- 13.8.5 COMPETITIVE BENCHMARKING: STARTUPS/SMES, 2024
 - 13.8.5.1 Detailed list of key startups/SMEs
 - 13.8.5.2 Competitive benchmarking of key startups/SMEs

13.9 COMPETITIVE SCENARIO

- 13.9.1 PRODUCT LAUNCHES
- 13.9.2 DEALS

14 COMPANY PROFILES

14.1 KEY PLAYERS

- 14.1.1 NVIDIA CORPORATION
 - 14.1.1.1 Business overview
 - 14.1.1.2 Products/Solutions/Services offered
 - 14.1.1.3 Recent developments
 - 14.1.1.3.1 Product launches
 - 14.1.1.3.2 Deals
 - 14.1.1.4 MnM view
 - 14.1.1.4.1 Key strengths
 - 14.1.1.4.2 Strategic choices
 - 14.1.1.4.3 Weaknesses and competitive threats
- 14.1.2 ADVANCED MICRO DEVICES, INC.
 - 14.1.2.1 Business overview

- 14.1.2.2 Products/Solutions/Services offered
- 14.1.2.3 Recent developments
 - 14.1.2.3.1 Product launches
 - 14.1.2.3.2 Deals
- 14.1.2.4 MnM view
 - 14.1.2.4.1 Key strengths
 - 14.1.2.4.2 Strategic choices
 - 14.1.2.4.3 Weaknesses and competitive threats
- 14.1.3 INTEL CORPORATION
 - 14.1.3.1 Business overview
 - 14.1.3.2 Products/Solutions/Services offered
 - 14.1.3.3 Recent developments
 - 14.1.3.3.1 Product launches
 - 14.1.3.3.2 Deals
 - 14.1.3.4 MnM view
 - 14.1.3.4.1 Key strengths
 - 14.1.3.4.2 Strategic choices
 - 14.1.3.4.3 Weaknesses and competitive threats
- 14.1.4 SK HYNIX INC.
 - 14.1.4.1 Business overview
 - 14.1.4.2 Products/Solutions/Services offered
 - 14.1.4.3 Recent developments
 - 14.1.4.3.1 Product launches
 - 14.1.4.3.2 Deals
 - 14.1.4.4 MnM view
 - 14.1.4.4.1 Key strengths
 - 14.1.4.4.2 Strategic choices
 - 14.1.4.4.3 Weaknesses and competitive threats
- 14.1.5 SAMSUNG
 - 14.1.5.1 Business overview
 - 14.1.5.2 Products/Solutions/Services offered
 - 14.1.5.3 Recent developments
 - 14.1.5.3.1 Product launches
 - 14.1.5.3.2 Deals
 - 14.1.5.4 MnM view
 - 14.1.5.4.1 Key strengths
 - 14.1.5.4.2 Strategic choices
 - 14.1.5.4.3 Weaknesses and competitive threats
- 14.1.6 MICRON TECHNOLOGY, INC.

- 14.1.6.1 Business overview
- 14.1.6.2 Products/Solutions/Services offered
- 14.1.6.3 Recent developments
 - 14.1.6.3.1 Product launches
 - 14.1.6.3.2 Deals
- 14.1.7 APPLE INC.
 - 14.1.7.1 Business overview
 - 14.1.7.2 Products/Solutions/Services offered
 - 14.1.7.3 Recent developments
 - 14.1.7.3.1 Product launches
 - 14.1.7.3.2 Deals
- 14.1.8 QUALCOMM TECHNOLOGIES, INC.
 - 14.1.8.1 Business overview
 - 14.1.8.2 Products/Solutions/Services offered
 - 14.1.8.3 Recent developments
 - 14.1.8.3.1 Product launches
 - 14.1.8.3.2 Deals
- 14.1.9 HUAWEI TECHNOLOGIES CO., LTD.
 - 14.1.9.1 Business overview
 - 14.1.9.2 Products/Solutions/Services offered
 - 14.1.9.3 Recent developments
 - 14.1.9.3.1 Product launches
 - 14.1.9.3.2 Deals
- 14.1.10 GOOGLE
 - 14.1.10.1 Business overview
 - 14.1.10.2 Products/Solutions/Services offered
 - 14.1.10.3 Recent developments
 - 14.1.10.3.1 Product launches
 - 14.1.10.3.2 Deals
- 14.1.11 AMAZON WEB SERVICES, INC.
 - 14.1.11.1 Business overview
 - 14.1.11.2 Products/Solutions/Services offered
 - 14.1.11.3 Recent developments
 - 14.1.11.3.1 Product launches
 - 14.1.11.3.2 Deals
- 14.1.12 TESLA
 - 14.1.12.1 Business overview
 - 14.1.12.2 Products/Solutions/Services offered
- 14.1.13 MICROSOFT

- 14.1.13.1 Business overview
- 14.1.13.2 Products/Solutions/Services offered
- 14.1.13.3 Recent developments
 - 14.1.13.3.1 Product launches
 - 14.1.13.3.2 Deals
- 14.1.14 META
 - 14.1.14.1 Business overview
 - 14.1.14.2 Products/Solutions/Services offered
 - 14.1.14.3 Recent developments
 - 14.1.14.3.1 Product launches
 - 14.1.14.3.2 Deals
- 14.1.15 T-HEAD
 - 14.1.15.1 Business overview
 - 14.1.15.2 Products/Solutions/Services offered
- 14.1.16 GRAPHCORE
 - 14.1.16.1 Business overview
 - 14.1.16.2 Products/Solutions/Services offered
 - 14.1.16.3 Recent developments
 - 14.1.16.3.1 Product launches
 - 14.1.16.3.2 Deals
- 14.1.17 CEREBRAS
 - 14.1.17.1 Business overview
 - 14.1.17.2 Products/Solutions/Services offered
 - 14.1.17.3 Recent developments
 - 14.1.17.3.1 Product launches
 - 14.1.17.3.2 Deals
- 14.2 OTHER PLAYERS
 - 14.2.1 MYTHIC
 - 14.2.2 BLAIZE
 - 14.2.3 GROQ, INC.
 - 14.2.4 HAILO TECHNOLOGIES LTD.
 - 14.2.5 SIMA TECHNOLOGIES, INC.
 - 14.2.6 KNERON, INC.
 - 14.2.7 TENSTORRENT
 - 14.2.8 SAMBANOVA SYSTEMS, INC.
 - 14.2.9 SAPEON INC.
 - 14.2.10 REBELLIONS INC.
 - 14.2.11 SHANGHAI BIREN TECHNOLOGY CO., LTD.

15 APPENDIX

15.1 DISCUSSION GUIDE

15.2 KNOWLEDGESTORE: MARKETSANDMARKETS' SUBSCRIPTION PORTAL

15.3 CUSTOMIZATION OPTIONS

15.4 RELATED REPORTS

15.5 AUTHOR DETAILS

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>