+357 96 030922
info@marketpublishers.com

The Global AI Inference Chip Market 2027-2037

https://marketpublishers.com/r/GA3C60562A8CEN.html
Date: October 2026
Pages: 677
Price: US$ 1,600.00 (Single User License)
ID: GA3C60562A8CEN

Abstracts

Investment in AI inference silicon has moved from a niche venture theme to one of the largest capital flows in semiconductors. Inference is expected to account for about three-quarters of AI compute demand by 2030, and buyers now judge hardware on cost per token rather than training speed. Investors have responded. Over US$5.2 billion of venture funding for inference AI chips was recorded across 31 deals in 2025, about 3.5 times the 2024 total. By 28 August 2026 the year had already reached US$4.2 billion across 22 deals. Rounds of US$25 million or more now account for almost all of the capital. Median pre-money valuations have risen from US$194 million in 2024 to more than US$1.7 billion in 2026.

The second half of 2026 brought a run of very large rounds:

Etched raised US$1 billion in two rounds, reaching a US$21 billion valuation.

SambaNova closed US$1 billion at US$11 billion, led by General Atlantic.

Positron AI raised up to US$875 million at US$5 billion.

Tenstorrent closed a US$1.5 billion Series E at US$6.2 billion, led by Trace Capital and Innovation X, with Fidelity, T. Rowe Price, Morgan Stanley and ARK Invest participating.

Others: Olix, MatX, Fractile, Rebellions, Axelera AI and SiMa.ai each raised nine-figure rounds.

Early stage: seed activity continues, for example Ubitium’s US$14 million round in Germany.

+357 96 030922
info@marketpublishers.com

Exits have arrived as well. Cerebras listed on Nasdaq in May 2026, raising US$5.55 billion. NVIDIA's licensing and hiring deal with Groq, reported at about US$20 billion, together with the Cerebras IPO, generated US$34.3 billion of exit value in 2026. Several other acquisitions followed:

Marvell acquired Celestial AI for US$3.25 billion upfront.

Microchip completed its acquisition of Hailo.

AMD agreed to acquire Taalas.

Analog Devices agreed to buy Alif Semiconductor for US$1.35 billion.

In China, Moore Threads, MetaX, Biren, Enflame and Iluvatar CoreX all listed.

Strategic and debt capital are increasingly important. Hyperscalers are committing gigawatts of capacity to custom accelerators: Anthropic has secured up to 1 GW of Google TPUs in 2026, rising to 5 GW in 2027. Sell-side analysts see Google's external TPU sales reaching US$100–250 billion by 2028. Investment-grade debt is now funding accelerator fleets directly, as with SpaceX's US$40 billion Apollo-led financing for NVIDIA chips.

The Global AI Inference Chip Market 2027–2037 provides a complete view of the market. It sizes every major chip architecture, from merchant GPUs and hyperscaler custom ASICs to dataflow, wafer-scale, compute-in-memory, photonic and neuromorphic designs, and edge NPUs. It analyses the forces reshaping competition: the memory wall, disaggregated prefill and decode, falling cost per token, power constraints, sovereign AI and export controls.

Report contents include:

Market size and forecasts for 2024–2037 by chip type, deployment location, workload, memory architecture, edge application, architecture and region, in US$ billions.

Unit shipments and ASPs for data center accelerators and edge devices.

Bear, base and bull scenarios, and a comparison with third-party estimates.

+357 96 030922
info@marketpublishers.com

Market shares by vendor and program for 2025, 2027, 2030 and 2037, including NVIDIA, AMD, Google TPU, AWS, Meta, Microsoft, OpenAI, Huawei and merchant challengers.

Technology analysis covering GPUs, TPUs and XPUs, LPUs, wafer-scale engines, transformer-specific silicon, in-memory compute, FP4 and logarithmic numerics, HBM4 and HBF memory, CoWoS and chiplets, NVLink, UALink and co-packaged optics, and rack-scale systems.

Data center strategies of hyperscalers, frontier labs, neoclouds and sovereign AI programs, with total cost of ownership analysis.

Edge inference across smartphones, AI PCs, automotive, robotics, industrial vision, smart cameras and wearables.

Regional analysis of North America, China, Europe, Asia-Pacific, the Middle East and Latin America.

Venture funding, valuations, capital efficiency, IPOs and M&A, plus a due-diligence framework for investors.

Supply chain analysis covering foundry, HBM, packaging, design services and ODMs.

125 company profiles, each with company facts, technology, key developments, funding and investment history, and a strategic assessment. Companies profiled include Advanced Micro Devices (AMD), Alchip Technologies, Alibaba Group (T-Head Semiconductor), Alif Semiconductor, Amazon Web Services (Annapurna Labs), Ambarella, Ambient Scientific, Anaflash, Apple, Arago, Arm Holdings, Axelera AI, Axera Semiconductor, Baidu (Kunlunxin), Biren Technology, Black Sesame Technologies, Blaize, BrainChip Holdings, Broadcom, ByteDance, Cadence Design Systems, Cambricon Technologies, Canaan Inc., Cerebras Systems, Ceva, CogniFiber, Corerain Technologies, d-Matrix, DEEPX, Denglin Technology, EdgeCortix, Efinix, Element Labs, EnCharge AI, Enflame Technology, Etched, Expedera, Extropic, Femtosense, Fractile, FuriosaAI, Global Unichip Corporation (GUC), Google (Alphabet), Graphcore, GreenWaves Technologies, Groq, Hailo, Horizon Robotics, Houmo.ai, Huawei (HiSilicon), Hygon Information Technology, IBM, Iluvatar CoreX, Imagination Technologies,

+357 96 030922
info@marketpublishers.com

Infineon Technologies, Innatera Nanosystems and more.....

+357 96 030922
info@marketpublishers.com

Contents

1 EXECUTIVE SUMMARY

1.1 The Inference Decade: Why the Market Is Changing Now
1.2 Market Definition and Scope at a Glance
1.3 Global Market Size and Growth, 2024–2037
1.4 Data Center Versus Edge
1.5 Chip Architecture Shift: GPUs, Custom ASICs and Challengers
1.6 Competitive Landscape in Brief
1.7 Fastest-Growing Segments
1.8 Capital Formation and Exits
1.9 Ten Predictions for 2027–2037
1.10 Strategic Implications by Stakeholder

2 INTRODUCTION, DEFINITIONS AND SEGMENTATION

2.1 Training Versus Inference
2.2 Anatomy of LLM Inference: Prefill, Decode and the KV Cache
  2.2.1 The Two Phases of a Request
  2.2.2 The KV Cache
  2.2.3 Mixture-of-Experts and the Memory Problem
  2.2.4 Batching
  2.2.5 Disaggregated Prefill and Decode
  2.2.6 Speculative Decoding
  2.2.7 The Throughput–Interactivity Trade-off
2.3 Inference Performance and Economic Metrics
  2.3.1 Latency Metrics
  2.3.2 Throughput and Interactivity Metrics
  2.3.3 Economic Metrics
2.4 Market Definition and What Is Counted
  2.4.1 The Unit of Value
  2.4.2 Attributing Revenue to Inference
  2.4.3 Edge and On-Device Value
  2.4.4 Exclusions
  2.4.5 Forecast Conventions
2.5 Segmentation Framework
  2.5.1 Deployment
  2.5.2 Chip Type

+357 96 030922
info@marketpublishers.com

2.5.3 Architecture
2.5.4 Workload
2.5.5 Memory Architecture
2.5.6 Region
2.5.7 Edge Application
2.6 Taxonomy of Inference Silicon
2.6.1 Programmable Parallel Processors
2.6.2 Domain-Specific Accelerators
2.6.3 Memory-Centric Architectures
2.6.4 Edge NPUs and SoCs
2.6.5 Emerging Physics and Number Systems
2.6.6 Reconfigurable Devices
2.6.7 How the Taxonomy Maps to the Market
2.7 A Short History of Inference Chips, 2015–2026
2.7.1 The First Inference Accelerators (2015–2019)
2.7.2 Scaling and the Transformer (2020–2022)
2.7.3 The Generative AI Surge (2022–2024)
2.7.4 Reasoning, Agents and Inference-First Silicon (2025–2026)

3 MARKET DRIVERS, TRENDS AND RESTRAINTS

3.1 Token Demand Explosion: Agentic AI, Reasoning Models and Multimodal
3.1.1 Measured token volumes
3.1.2 Agentic AI: more tokens per task, under latency constraints
3.1.3 Reasoning models and inference-time compute
3.1.4 Multimodal generation and physical AI
3.1.5 Power demand as a cross-check
3.2 The Economics of Inference: Falling Cost per Token
3.2.1 The Jevons dynamic
3.2.2 Where the cost goes
3.2.3 Sources of cost reduction
3.2.4 From capex to opex: what buyers optimize
3.3 The Memory Wall and the Rise of Memory-Centric Design
3.3.1 Why inference is memory-bound
3.3.2 KV cache growth
3.3.3 How suppliers are responding
3.4 Power, Grid and Siting Constraints
3.4.1 The binding constraint has moved to electricity
3.4.2 Community opposition and siting

+357 96 030922
info@marketpublishers.com

3.4.3 Rack density, cooling and power delivery
3.5 Disaggregated and Heterogeneous Inference
3.5.1 Splitting prefill and decode
3.5.2 Heterogeneous fleets and the KV cache tier
3.6 Software Ecosystems and Switching Costs
3.6.1 CUDA and NVIDIA's widening software perimeter
3.6.2 Why inference lowers switching costs
3.6.3 Where software still protects incumbents
3.7 Sovereign AI and National Compute Programs
3.7.1 Compute as strategic infrastructure
3.7.2 Regional programs
3.8 On-Device and Edge AI Demand Drivers
3.8.1 Why inference moves to the edge
3.8.2 Smaller models and lower precision
3.8.3 Device categories
3.8.4 The memory shock as a counter-force
3.9 Geopolitics, Export Controls and the US–China Silicon Curtain
3.9.1 Three generations of control
3.9.2 China's response: guaranteed domestic demand
3.9.3 Third markets
3.10 Restraints, Risks and Challenges
3.10.1 Power and grid
3.10.2 Memory and packaging supply
3.10.3 Monetization and capital sustainability
3.10.4 Efficiency shocks
3.10.5 Concentration and dependence
3.10.6 Software, execution and vendor claims
3.10.7 Policy, regulation and geopolitics
3.10.8 Edge-specific restraints

4 TECHNOLOGY ANALYSIS

4.1 GPUs for Inference
4.1.1 NVIDIA Blackwell and Blackwell Ultra
4.1.2 Vera Rubin and Rubin CPX
4.1.3 AMD Instinct MI355X and MI450
4.1.4 Intel Gaudi 3 and Crescent Island
4.1.5 GPU Specification Comparison
4.2 Hyperscaler Custom ASICs

+357 96 030922
info@marketpublishers.com

4.2.1 Google TPU v6e (Trillium) and TPU v7 (Ironwood)
4.2.2 AWS Trainium2, Trainium3 and Inferentia2
4.2.3 Microsoft Maia 100 and Maia 200
4.2.4 Meta MTIA v2 and Successors
4.2.5 OpenAI, Anthropic and the Next Wave
4.3 Merchant Data-Center Inference ASICs: Dataflow, Deterministic SRAM (LPU), Wafer-Scale, Transformer-Specific and Hard-Wired
4.3.1 Cerebras WSE-3
4.3.2 The Groq LPU and the NVIDIA Licensing Agreement
4.3.3 Etched Sohu and MatX
4.3.4 SambaNova SN40L and SN50
4.3.5 Tenstorrent Wormhole and Blackhole
4.3.6 d-Matrix Corsair
4.3.7 FuriosaAI RNGD and Rebellions ATOM and REBEL
4.3.8 Qualcomm Cloud AI 100 Ultra, AI200 and AI250
4.3.9 Positron Atlas and Asimov
4.3.10 Taalas Hard-Wired Model Silicon
4.3.11 Comparing the Merchant Approaches
4.4 Memory-Centric Architectures: SRAM-Centric, Digital and Analog In-Memory Compute, PIM
4.4.1 SRAM-Centric Designs
4.4.2 Digital In-Memory Compute
4.4.3 Analog In-Memory Compute
4.4.4 Processing-in-Memory and Near-Memory Compute
4.5 Numerics: FP8, MXFP4/NVFP4, INT4, Ternary, Logarithmic
4.5.1 FP8
4.5.2 Microscaling: MXFP4 and NVFP4
4.5.3 INT4 and Integer Quantization
4.5.4 Ternary and Sub-2-Bit Models
4.5.5 Logarithmic Number Systems: Tensordyne
4.6 Photonic and Optical Inference Compute
4.6.1 Lightmatter
4.6.2 Lumai
4.6.3 Q.ANT
4.6.4 Arago
4.6.5 Other Developers
4.7 Neuromorphic Inference
4.7.1 Research and Large-Scale Systems
4.7.2 Commercial Edge Neuromorphic Processors

+357 96 030922
info@marketpublishers.com

4.8 Edge NPUs and SoC Integration
4.8.1 From TOPS to Bandwidth
4.8.2 Smartphone SoCs
4.8.3 PC Processors
4.8.4 Design Directions to 2030
4.9 Discrete Edge AI Accelerators
4.9.1 Why discrete accelerators persist
4.9.2 Vendor analysis
4.9.3 Competitive dynamics and outlook
4.10 Memory Technologies for Inference: HBM3E, HBM4/HBM4E, Custom HBM, GDDR7, LPDDR5X/6, SOCAMM, HBF and CXL
4.10.1 HBM3E: the 2025–2026 workhorse
4.10.2 HBM4 and HBM4E: wider interfaces and logic base dies
4.10.3 GDDR7: Rubin CPX and the disaggregated prefill tier
4.10.4 LPDDR5X, LPDDR6 and SOCAMM: capacity-first inference
4.10.5 High-bandwidth flash (HBF)
4.10.6 CXL memory pooling and KV-cache tiers
4.10.7 Processing-in-memory
4.11 Advanced Packaging and Chiplets: CoWoS, SoIC, Panel-Level and UCIe
4.11.1 TSMC CoWoS-S, CoWoS-L and CoWoS-R
4.11.2 SoIC and hybrid bonding
4.11.3 Panel-level packaging, system-on-wafer and glass substrates
4.11.4 Chiplets and UCIe
4.12 Scale-Up and Scale-Out Interconnect: NVLink, UALink, Ethernet (UEC and ESUN) and Co-Packaged Optics
4.12.1 NVLink 5, NVLink 6 and NVLink Fusion
4.12.2 UALink 1.0
4.12.3 Ethernet for scale-up and scale-out: UEC and ESUN
4.12.4 Hyperscaler fabrics
4.12.5 Co-packaged optics
4.13 Process Technology: N3, N2, A16, Intel 18A/14A, Samsung SF2 and Rapidus
4.13.1 TSMC: N3P, N2, A16 and A14
4.13.2 Intel 18A and 14A
4.13.3 Samsung SF2
4.13.4 Rapidus
4.13.5 China
4.14 Rack-Scale Systems, Power Delivery and Cooling
4.14.1 NVIDIA: GB300 NVL72, Vera Rubin NVL72 and the CPX and LPX extensions
4.14.2 AMD Helios

+357 96 030922
info@marketpublishers.com

4.14.3 Hyperscaler pods: Ironwood, TPU 8, Trainium and Maia
4.14.4 Huawei CloudMatrix 384 and Atlas SuperPoDs
4.14.5 Challenger systems: Cerebras CS-3, Groq and others
4.14.6 800 VDC power distribution
4.14.7 Liquid cooling
4.15 Inference Software, Compilers and Serving Frameworks
4.15.1 Serving engines: vLLM, SGLang and TensorRT-LLM
4.15.2 Distributed inference: Dynamo and llm-d
4.15.3 Compilers and kernel languages: XLA, Triton and vendor stacks
4.15.4 Challenger SDKs
4.16 Technology Readiness and Roadmap 2027–2037

5 GLOBAL MARKET FORECASTS 2024–2037

5.1 Forecast Methodology and Key Assumptions
5.1.1 What the model measures
5.1.2 Data center GPUs and hyperscaler ASICs: top-down anchors
5.1.3 The inference-share allocation
5.1.4 Bottom-up lines: merchant accelerators, China and novel architectures
5.1.5 Edge anchor and segmentation
5.1.6 Derived segmentations and scenarios
5.2 Total Market
5.3 Data Center Inference by Chip Type
5.3.1 Merchant GPUs
5.3.2 Hyperscaler Custom ASICs
5.3.3 Merchant, Chinese Domestic and Novel-Architecture Accelerators
5.4 Data Center by Deployment Location
5.5 Data Center by Workload
5.6 Data Center by Memory Architecture
5.7 Data Center Unit Shipments and ASP
5.8 Edge and On-Device by Application
5.9 Edge Unit Shipments and Silicon Content
5.10 Market by Architecture
5.11 Segment Growth Summary
5.12 Scenario Analysis
5.13 Reconciliation with Other Future Markets Reports and Third-Party Estimates
5.13.1 Future Markets' Computing and AI for Data Centers 2027–2040
5.13.2 Future Markets' Generative AI Hardware Materials 2027–2037
5.13.3 Future Markets' Edge AI Chips 2026–2036

+357 96 030922
info@marketpublishers.com

5.13.4 Third-party estimates

6 DATA CENTER INFERENCE MARKET ANALYSIS

6.1 Hyperscaler Inference Strategies (Google, AWS, Microsoft, Meta, Oracle)
6.1.1 Google: TPU as an Inference Business
6.1.2 AWS: Trainium Becomes the Bedrock Engine
6.1.3 Microsoft: Maia 200 and a Multi-Vendor Fleet
6.1.4 Meta: From Ranking ASIC to Generative Inference
6.1.5 Oracle: The Neutral Hyperscaler
6.1.6 Comparing Hyperscaler Strategies
6.2 Frontier Labs as Silicon Buyers (OpenAI, Anthropic, xAI)
6.2.1 OpenAI
6.2.2 Anthropic
6.2.3 xAI
6.2.4 Frontier Lab Agreements Compared
6.3 Neoclouds and Inference-as-a-Service
6.3.1 Capacity Neoclouds
6.3.2 Inference-API Providers
6.4 Sovereign and National AI Facilities
6.4.1 Gulf States
6.4.2 Europe
6.4.3 Asia-Pacific
6.5 Enterprise On-Premise Inference
6.6 Total Cost of Ownership: GPU vs ASIC vs Challenger
6.6.1 From Accelerator-Hours to Tokens
6.6.2 Utilization Sensitivity
6.6.3 TCO Over the Forecast Period
6.7 Procurement Behavior and Contract Structures
6.7.1 Multi-Year Capacity Commitments
6.7.2 Tokens-as-a-Service
6.7.3 Leasing, Off-Balance-Sheet Structures and Asset-Backed Debt
6.7.4 Circular Financing and Vendor Equity
6.7.5 Pricing Leverage, Allocation and Multi-Sourcing
6.7.6 Outlook for Procurement

7 EDGE AND ON-DEVICE INFERENCE MARKET ANALYSIS

7.1 Smartphones

+357 96 030922
info@marketpublishers.com

7.1.1 Flagship platforms in 2025–2026
7.1.2 GenAI phone penetration and the 2026 memory shock
7.1.3 Outlook
7.2 AI PCs and Workstations
7.2.1 The Copilot+ floor and 2026 platforms
7.2.2 The 2026 PC shipment decline
7.2.3 Outlook: from NPU checkbox to local inference workstation
7.3 Automotive
7.3.1 Compute tiers and the value ladder
7.3.2 NVIDIA DRIVE Thor
7.3.3 Tesla AI5 and AI6
7.3.4 Merchant challengers: Qualcomm, Horizon Robotics, Mobileye, Black Sesame, Ambarella
7.3.5 In-house chips from Chinese OEMs
7.3.6 The cockpit and the convergence of compute
7.4 Robotics and Physical AI
7.4.1 NVIDIA Jetson Thor as the reference platform
7.4.2 Humanoid compute requirements
7.4.3 Market sizing and reconciliation
7.5 Industrial and Machine Vision
7.5.1 Applications and compute requirements
7.5.2 Silicon supply
7.6 Smart Cameras, Video Analytics and Smart Cities
7.6.1 From CNN detection to on-camera VLMs
7.6.2 Vendors: Ambarella, Hailo, Axera and the Chinese ecosystem
7.6.3 Smart cities
7.7 Wearables, XR, Smart Home and Consumer IoT
7.7.1 Smart glasses: the new AI form factor
7.7.2 XR headsets
7.7.3 Watches, hearables, smart home and consumer IoT
7.8 Edge Infrastructure: Telco MEC and On-Site Servers
7.8.1 Telco MEC and AI-RAN
7.8.2 On-site enterprise and industrial servers
7.9 Edge versus Cloud: Where Inference Will Run
7.9.1 The placement decision
7.9.2 Hybrid execution as the default architecture
7.9.3 Indicative placement by workload
7.9.4 Implications for silicon vendors

+357 96 030922
info@marketpublishers.com

8 REGIONAL MARKET ANALYSIS

8.1 Regional Overview
8.1.1 Data Center and Edge Mix by Region
8.1.2 Regional Growth Rates
8.1.3 Where Value Is Designed, Made and Consumed
8.1.4 Sovereign Demand as a Regional Swing Factor
8.1.5 Regional Summary
8.2 North America
8.2.1 Hyperscaler and Frontier-Lab Demand
8.2.2 Power, Siting and Local Opposition
8.2.3 Supplier Base and Capital Formation
8.2.4 Manufacturing Onshoring
8.2.5 Edge and On-Device Inference
8.2.6 Outlook
8.3 China
8.3.1 Export Controls and the Domestic Substitution Mandate
8.3.2 Huawei Ascend
8.3.3 Cambricon, Hygon, Kunlunxin and Alibaba T-Head
8.3.4 The GPU Listings and Capital Formation
8.3.5 Supply-Side Constraints: Lithography and Memory
8.3.6 Edge and On-Device Inference
8.3.7 Interpreting the Forecast
8.4 Europe
8.4.1 Policy Framework: AI Act, Chips Act and AI Factories
8.4.2 Demand Characteristics
8.4.3 European Inference Chip Developers
8.4.4 The European Capital Gap
8.4.5 Edge and On-Device Inference
8.4.6 Outlook
8.5 Asia-Pacific (ex-China)
8.5.1 Taiwan
8.5.2 South Korea
8.5.3 Japan
8.5.4 India
8.5.5 Singapore, Southeast Asia and Australia
8.5.6 Outlook
8.6 Middle East and Africa
8.6.1 Saudi Arabia and the UAE

+357 96 030922
info@marketpublishers.com

8.6.2 Capital Exporter, Israeli Design and Africa
8.6.3 Risks and Outlook
8.7 Latin America

9 COMPETITIVE LANDSCAPE

9.1 Market Structure and Concentration
9.1.1 Barriers to Entry
9.1.2 Concentration at the Edge
9.2 Market Shares by Vendor
9.3 NVIDIA's Inference Position and Response
9.3.1 Scale and Financial Position
9.3.2 Rubin CPX and the Disaggregated Hardware Portfolio
9.3.3 Dynamo: The Orchestration Layer
9.3.4 The Groq Licence and the LPX Rack
9.3.5 NVLink Fusion: Opening the Fabric
9.3.6 Capital as a Competitive Instrument
9.3.7 Assessment
9.4 AMD, Intel, Qualcomm and Other Diversified Vendors
9.4.1 AMD
9.4.2 Intel
9.4.3 Qualcomm
9.4.4 Broadcom
9.4.5 Marvell, MediaTek and Other Design Partners
9.4.6 Chinese and Edge-Centric Incumbents
9.5 Challenger Landscape by Segment
9.5.1 Data Center Inference ASICs
9.5.2 Compute-in-Memory, Analog and Novel Architectures
9.5.3 Photonic and Optical Compute and Interconnect
9.5.4 Edge Inference Silicon
9.5.5 Semiconductor IP Licensing
9.5.6 Chinese Domestic Challengers
9.6 Strategic Partnerships and Ecosystems
9.7 Competitive Positioning Matrix

10 INVESTMENT, FUNDING, M&A AND EXITS

10.1 Venture Capital Deal Activity
10.1.1 Quarterly Activity

+357 96 030922
info@marketpublishers.com

10.1.2 Stage, Size and Geography
10.1.3 Reconciling Funding Trackers
10.2 Deal Sizes and Valuations
10.2.1 Valuation Step-Ups
10.3 Capital Efficiency
10.3.1 Data Discrepancies
10.4 Major Funding Rounds 2025–2026
10.4.1 Who Is Investing
10.5 Exits: IPOs and Acquisitions
10.5.1 The Cerebras IPO
10.5.2 The Licence-and-Hire Model
10.5.3 Chinese GPU IPOs
10.5.4 The IPO Pipeline
10.6 Hyperscaler and Corporate Strategic Investment
10.6.1 Compute-for-Equity and Warrant Structures
10.6.2 NVIDIA and Other Incumbents as Investors
10.7 Due Diligence Framework for Investors
10.7.1 Valuations Against the Forecast Market

11 SUPPLY CHAIN

11.1 Value Chain Overview
11.2 Foundry
11.3 Memory Suppliers
11.3.1 Memory content per accelerator and demand outlook
11.4 Advanced Packaging and OSAT
11.5 IP, EDA and Design Services
11.5.1 IP and EDA
11.5.2 Design services and the hyperscaler ASIC programs
11.6 Systems and ODMs
11.7 Supply Chain Risks
11.7.1 Resilience scenarios and indicators

12 MARKET OUTLOOK

12.1 Outlook 2027–2037
12.1.1 What will decide the outcome
12.2 Key Risks
12.2.1 The risks that matter most

+357 96 030922
info@marketpublishers.com

12.3 Opportunities
12.3.1 Reading the map
12.4 Recommendations
12.4.1 Chip vendors
12.4.2 Hyperscalers and AI labs
12.4.3 OEMs and system builders
12.4.4 Investors
12.4.5 Policymakers

13 COMPANY PROFILES (125 COMPANY PROFILES)

14 RESEARCH METHODOLOGY

14.1 Objectives and Scope
14.2 Market Definition and Allocation Rules
14.3 Consistency with the Future Markets Catalogue
14.4 Primary and Secondary Research
14.5 Forecast Approach
14.6 Conventions and Limitations

15 GLOSSARY OF TERMS AND ABBREVIATIONS

16 REFERENCES

+357 96 030922
info@marketpublishers.com

List Of Tables

LIST OF TABLES

Table 1. Market Scope at a Glance
Table 2. Headline Forecast Summary, AI Inference Chip Market (US$ Billions)
Table 3. Data Center vs Edge Share of AI Inference Chip Revenue, 2024–2037 (%)
Table 4. Key 2025–2026 Funding, IPO and M&A Events in Inference Silicon
Table 5. Ten Predictions for the AI Inference Chip Market, 2027–2037
Table 6. Strategic Implications for Chip Vendors, Hyperscalers, OEMs and Investors
Table 7. Training vs. Inference Workload Characteristics
Table 8. Illustrative KV Cache Footprint by Model Architecture and Context Length
Table 9. Prefill vs. Decode Hardware Requirements
Table 10. Key Inference Metrics and Definitions
Table 11. Market Scope – Included and Excluded Items
Table 12. Segmentation Framework Used in This Report
Table 13. Milestones in AI Inference Silicon, 2015–2026
Table 14. Global Data Center Power Demand by Workload, 2025–2030 (GW)
Table 15. Token Volume and Agentic AI Demand Indicators, 2024–2030
Table 16. Inference Cost per Million Output Tokens by Model Class and Hardware, 2024–2040 (US$)
Table 17. Inference Cost Breakdown by Infrastructure Component, 2025 (% of Delivered Cost per Token)
Table 18. Selected Token-Economics Claims by Inference Silicon Suppliers, 2025–2026
Table 19. Memory Bandwidth and Capacity Requirements by Model Class
Table 20. Power-Related Constraints and Their Impact on Inference Chip Design
Table 21. Disaggregated Serving Approaches
Table 22. Inference Software Stacks by Vendor
Table 23. Selected Sovereign AI Programs and Inference Silicon Implications
Table 24. Edge AI Drivers and Their Impact on Silicon
Table 25. Export Control Timeline Affecting Inference Chips, 2022–2026
Table 26. Market Restraints – Impact Assessment
Table 27. Data-Center GPU Specifications for Inference (NVIDIA, AMD, Intel), 2024–2027
Table 28. Hyperscaler Inference ASIC Specifications
Table 29. Merchant Inference Accelerator Benchmarking
Table 30. In-Memory and Near-Memory Computing Approaches Compared
Table 31. Numeric Formats for Inference
Table 32. Photonic Inference Compute Developers and Status

+357 96 030922
info@marketpublishers.com

Table 33. Neuromorphic Processors for Inference
Table 34. Smartphone and PC NPU Performance, 2024–2026
Table 35. Discrete Edge AI Accelerator Comparison
Table 36. Memory Technology Roadmap for Inference
Table 37. Packaging Technologies for Inference Accelerators
Table 38. Scale-Up Interconnect Comparison
Table 39. Leading-Edge Process Nodes for Inference Silicon
Table 40. Rack-Scale Inference Systems Compared
Table 41. Compiler and Runtime Approaches of Challenger Vendors
Table 42. Technology Readiness Levels of Inference Architectures
Table 43. Inference Silicon Technology Roadmap, 2027–2037
Table 44. Inference Share of Accelerator Revenue by Chip Class, Selected Years
Table 45. Key Forecast Assumptions
Table 46. Global AI Inference Chip Market, 2024–2037 (US$ Billions)
Table 47. Data Center AI Inference Chip Revenue by Chip Type, 2024–2037 (US$ Billions)
Table 48. Merchant GPU Revenue Attributable to Inference, 2024–2037 (US$ Billions)
Table 49. Hyperscaler Custom ASIC Inference Revenue by Program, 2024–2037 (US$ Billions)
Table 50. Merchant, Chinese Domestic and Novel-Architecture Inference Accelerator Revenue, 2024–2037 (US$ Billions)
Table 51. Data Center & Infrastructure Inference Chip Revenue by Deployment Location, 2024–2037 (US$ Billions)
Table 52. Data Center Inference Chip Revenue by Workload, 2024–2037 (US$ Billions)
Table 53. Data Center Inference Chip Revenue by Memory Architecture, 2024–2037 (US$ Billions)
Table 54. Data Center Inference Accelerator Shipments and Blended ASP, 2024–2037
Table 55. Edge & On-Device AI Inference Chip Revenue by Application, 2024–2037 (US$ Billions)
Table 56. Edge AI Inference Chip Unit Shipments by Application, 2024–2037 (Million Units)
Table 57. Average AI Silicon Value per Edge Device by Application, 2025–2037 (US$ per Unit)
Table 58. Global AI Inference Chip Revenue by Architecture, 2024–2037 (US$ Billions)
Table 59. Segment Growth Summary, 2027–2037 (US$ Billions, CAGR)
Table 60. AI Inference Chip Market Scenarios: Bear, Base and Bull, 2026–2037 (US$ Billions)
Table 61. Scenario Drivers and Assumptions
Table 62. Comparison of AI Inference Chip Market Estimates

+357 96 030922
info@marketpublishers.com

Table 63. Data Center Inference Chip Revenue Share by Deployment Location, 2025–2037
Table 64. Hyperscaler Inference Silicon Strategies
Table 65. Frontier Lab Compute Agreements, 2025–2026
Table 66. Neocloud and Inference-API Providers and Their Silicon
Table 67. Sovereign AI Facilities and Their Inference Silicon Choices
Table 68. Enterprise On-Premise Inference Platforms
Table 69. Illustrative TCO Comparison per Million Tokens (Mid-Tier Model, 2026 Basis)
Table 70. Illustrative Sensitivity of Cost per Million Tokens to Fleet Utilization
Table 71. Inference Compute Procurement Structures and Their Implications
Table 72. Edge and On-Device Inference Segments at a Glance
Table 73. Flagship Smartphone AI Processors, 2025–2026
Table 74. AI PC Platforms, 2025–2026
Table 75. AI PC and Total PC Shipment Forecasts Compared
Table 76. Mobileye ADAS and Autonomy ASP Ladder, 2026
Table 77. Automotive AI SoCs
Table 78. Robotics Compute Platforms
Table 79. Industrial Edge AI Silicon, Selected Products
Table 80. Smart Camera and Video Analytics AI Silicon
Table 81. Edge Infrastructure Inference Platforms
Table 82. Workload Placement Decision Factors
Table 83. Indicative Primary Placement of Inference Workloads, 2026 and 2032
Table 84. Global AI Inference Chip Revenue by Region, 2024–2037 (US$ Billions)
Table 85. AI Inference Chip Revenue by Region and Deployment, 2027 and 2037 (US$ Billions)
Table 86. Regional Roles in the AI Inference Chip Value Chain
Table 87. Sovereign and Government AI Processor Demand by Region, 2025–2040 (US$ B)
Table 88. Regional AI Inference Chip Market Summary, 2027 and 2037
Table 89. North America: AI Inference Chip Revenue, 2024–2037 (US$ Billions)
Table 90. North American Inference Silicon Suppliers by Segment
Table 91. China: AI Inference Chip Revenue, 2024–2037 (US$ Billions)
Table 92. Chinese AI Accelerator Capital Markets Events, 2025–2026
Table 93. Chinese Domestic Inference Accelerators
Table 94. Composition of China's Data Center Inference Revenue, 2025–2037 (US$ B)
Table 95. Europe: AI Inference Chip Revenue, 2024–2037 (US$ Billions)
Table 96. European Inference Chip Developers
Table 97. Asia-Pacific (ex-China): AI Inference Chip Revenue, 2024–2037 (US$ Billions)

+357 96 030922
info@marketpublishers.com

Table 98. Asia-Pacific (ex-China) Inference Silicon Ecosystem by Country
Table 99. Middle East & Africa: AI Inference Chip Revenue, 2024–2037 (US$ Billions)
Table 100. Gulf Inference Compute Programs and Silicon Suppliers
Table 101. Latin America: AI Inference Chip Revenue, 2024–2037 (US$ Billions)
Table 102. Data Center Inference Concentration Indicators, 2025–2037
Table 103. Data Center AI Inference Chip Revenue and Share by Vendor/Program, 2025–2037 (US$ Billions, % share)
Table 104. Change in Data Center Inference Share and Growth by Vendor/Program, 2025–2037
Table 105. Inference Silicon Programs of Diversified Vendors, 2026
Table 106. Challenger Segments: Representative Companies, Architectural Theses and Status, Q4 2026
Table 107. Major Inference Silicon Partnerships, 2024–2026
Table 108. Competitive Positioning of Leading Inference Chip Suppliers
Table 109. Inference AI Chip Venture Capital Deal Activity, 2016–2026 YTD
Table 110. Inference AI Chip VC Deal Activity by Quarter, Q1 2025–Q3 2026
Table 111. Comparison of AI Chip and Inference Chip Funding Trackers, 2025–2026
Table 112. Inference AI Chip VC Deal Sizes and Valuations by Stage (US$ Millions)
Table 113. Valuation Step-Ups at Selected Inference Chip Companies, 2024–2026
Table 114. Top Inference AI VC-Backed Companies by Capital Efficiency (as of 30 September 2026)
Table 115. PitchBook Versus Publicly Announced Funding: Selected Discrepancies
Table 116. Largest Inference Chip Funding Rounds, 2025–2026
Table 117. Inference Chip M&A and IPO Transactions, 2024–2026
Table 118. Hyperscaler Strategic Investments and Acquisitions in AI Silicon and Infrastructure
Table 119. Due Diligence Checklist for Inference Chip Investments
Table 120. Inference Chip Value Chain and Key Suppliers
Table 121. Foundry Capacity Relevant to Inference Silicon
Table 122. HBM Supplier Positions
Table 123. Design-Services Partners and the Hyperscaler Programs They Serve
Table 124. Supply Chain Risk Register
Table 125. Base-Case Milestones in the AI Inference Chip Market, 2026–2037
Table 126. Risk Register (Probability/Impact)
Table 127. Opportunity Map
Table 128. Summary of Strategic Recommendations by Stakeholder
Table 129. Companies Profiled in This Report
Table 130. Glossary of Terms and Abbreviations

+357 96 030922
info@marketpublishers.com

List Of Figures

LIST OF FIGURES

Figure 1. Global AI Inference Chip Market, 2024–2037 (US$ Billions)
Figure 2. Global AI Inference Chip Revenue by Architecture, 2024–2037 (%)
Figure 3. Data Center Inference Chip Market Share by Vendor/Program, 2030 (%)
Figure 4. Segment CAGR, 2027–2037 (%)
Figure 5. Taxonomy of AI Inference Silicon
Figure 6. Global Data Center Power Demand by Workload, 2025–2030 (GW)
Figure 7. Inference Cost per Million Output Tokens by Model Class and Hardware, 2024–2040
Figure 8. Inference Cost Breakdown by Component, 2025 (% of Delivered Cost per Token)
Figure 9. Data Center AI Inference Chip Revenue by Chip Type, 2024–2037 (US$ Billions)
Figure 10. Data Center Inference Revenue Share by Chip Type, 2024–2037 (%)
Figure 11. Share of Accelerator Revenue Deployed for Inference, 2024–2037 (%)
Figure 12. Hyperscaler Custom ASIC Inference Revenue by Program, 2024–2037 (US$ Billions)
Figure 13. Challenger Segments: Inference Accelerator Revenue, 2024–2037 (US$ Billions)
Figure 14. Data Center Inference Revenue by Deployment Location, 2024–2037 (%)
Figure 15. Data Center Inference Chip Revenue by Workload, 2024–2037 (US$ Billions)
Figure 16. Data Center Inference Revenue by Memory Architecture, 2024–2037 (%)
Figure 17. Data Center Inference Accelerator Shipments, 2024–2037 (Million Units)
Figure 18. Edge & On-Device AI Inference Chip Revenue by Application, 2024–2037 (US$ Billions)
Figure 19. Edge Inference Chip Revenue Share by Application, 2024–2037 (%)
Figure 20. Edge AI Inference Chip Unit Shipments by Application, 2024–2037 (Million Units)
Figure 21. AI Inference Chip Market Scenarios, 2024–2037 (US$ Billions)
Figure 22. Global AI Inference Chip Revenue by Region, 2024–2037 (US$ Billions)
Figure 23. AI Inference Chip Market CAGR by Region, 2027–2037 (%)
Figure 24. Data Center Inference Chip Market Share by Vendor/Program, 2025 (%)
Figure 25. Data Center Inference Chip Market Share by Vendor/Program, 2037 (%)
Figure 26. AI Inference Chip Market Map, Q4 2026
Figure 27. Inference AI Chip VC Deal Value and Count, 2016–2026 YTD
Figure 28. Inference AI Chip VC Deal Value by Quarter, 2025–2026

+357 96 030922
info@marketpublishers.com

Figure 29. Advanced Micro Devices (AMD): AMD Instinct MI300-series accelerator
Figure 30. Ambarella: Ambarella CV7 vision SoC
Figure 31. Cerebras Systems: Cerebras Wafer-Scale Engine 3 (WSE-3)
Figure 32. d-Matrix: d-Matrix Pavehawk 3D-stacked in-memory compute silicon
Figure 33. DEEPX: DEEPX DX-series NPU
Figure 34. EnCharge AI: EnCharge AI EN100 M.2 accelerator card
Figure 35. Etched: Etched Sohu transformer inference board
Figure 36. FuriosaAI: FuriosaAI first-generation Warboy inference chip
Figure 37. Google (Alphabet): Google TPU v7 Ironwood accelerators on a server board
Figure 38. Graphcore (SoftBank): Graphcore Colossus MK2 GC200 IPU
Figure 39. GreenWaves Technologies: GreenWaves GAP8 and GAP9 processors
Figure 40. Groq: Groq LPU (originally the Tensor Streaming Processor)
Figure 41. Hailo: Hailo-10H edge AI accelerator
Figure 42. Horizon Robotics: Horizon Robotics Journey 5 automotive AI processor
Figure 43. Innatera Nanosystems: Innatera Pulsar spiking neural processor
Figure 44. Intel Corporation: Intel Loihi 2 neuromorphic research chip
Figure 45. Lightmatter: Lightmatter Envise photonic computing chip
Figure 46. MediaTek: MediaTek Pentonic 2000 smart-TV SoC with integrated AI processing
Figure 47. Meta Platforms (MTIA): Meta Training and Inference Accelerator (MTIA)
Figure 48. Microsoft (Maia): Microsoft Azure Maia 100 AI accelerator and Cobalt 100 CPU
Figure 49. Mythic: Mythic MP10304 quad-AMP PCIe card
Figure 50. NVIDIA: NVIDIA Rubin platform: Vera CPU, Rubin GPU, NVLink 6 Switch, ConnectX-9 SuperNIC, BlueField-4 DPU and Spectrum-6 Ethernet switch
Figure 51. Qualcomm: Qualcomm Cloud AI 100 inference accelerator
Figure 52. SambaNova Systems: SambaNova Cardinal SN10 RDU (first generation)
Figure 53. SiMa.ai: SiMa.ai MLSoC
Figure 54. Synaptics: Synaptics Astra SL2610 edge AI processor
Figure 55. Tenstorrent: Tenstorrent AI accelerator card
Figure 56. Untether AI (AMD): Untether AI speedAI240 inference chip (product line discontinued after AMD hired the team in 2025)

+357 96 030922
info@marketpublishers.com

I would like to order

Product name: The Global AI Inference Chip Market 2027-2037

Product link: https://marketpublishers.com/r/GA3C60562A8CEN.html

Price: US$ 1,600.00 (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/GA3C60562A8CEN.html