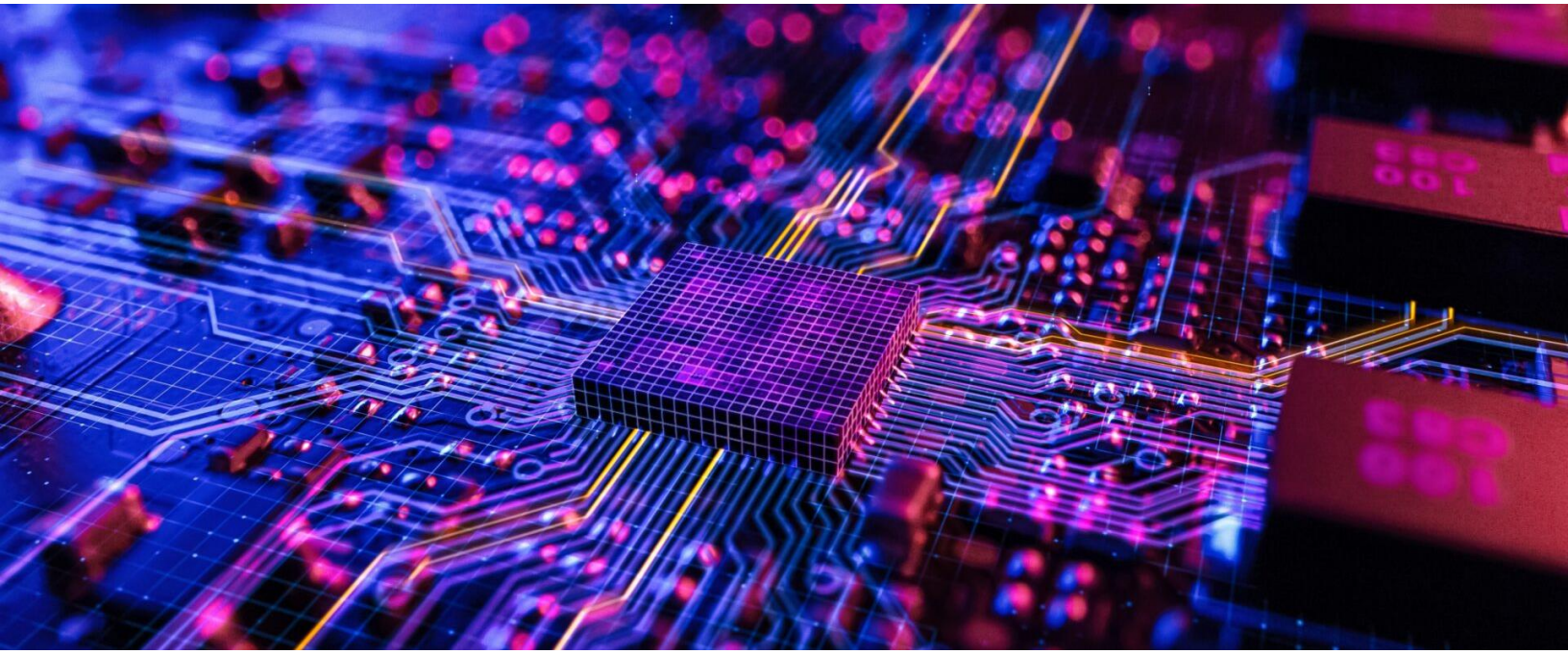
CONNECTED VISION ADVISORS

DEEP TECH INVESTMENT BANKING

AI COMPUTE IS GROWING FASTER THAN MOORE'S LAW

AI ACCELERATION MARKET LANDSCAPE

AUGUST 2026



SEMICONDUCTORS

AI & COMPUTER VISION

EDGE INTELLIGENCE

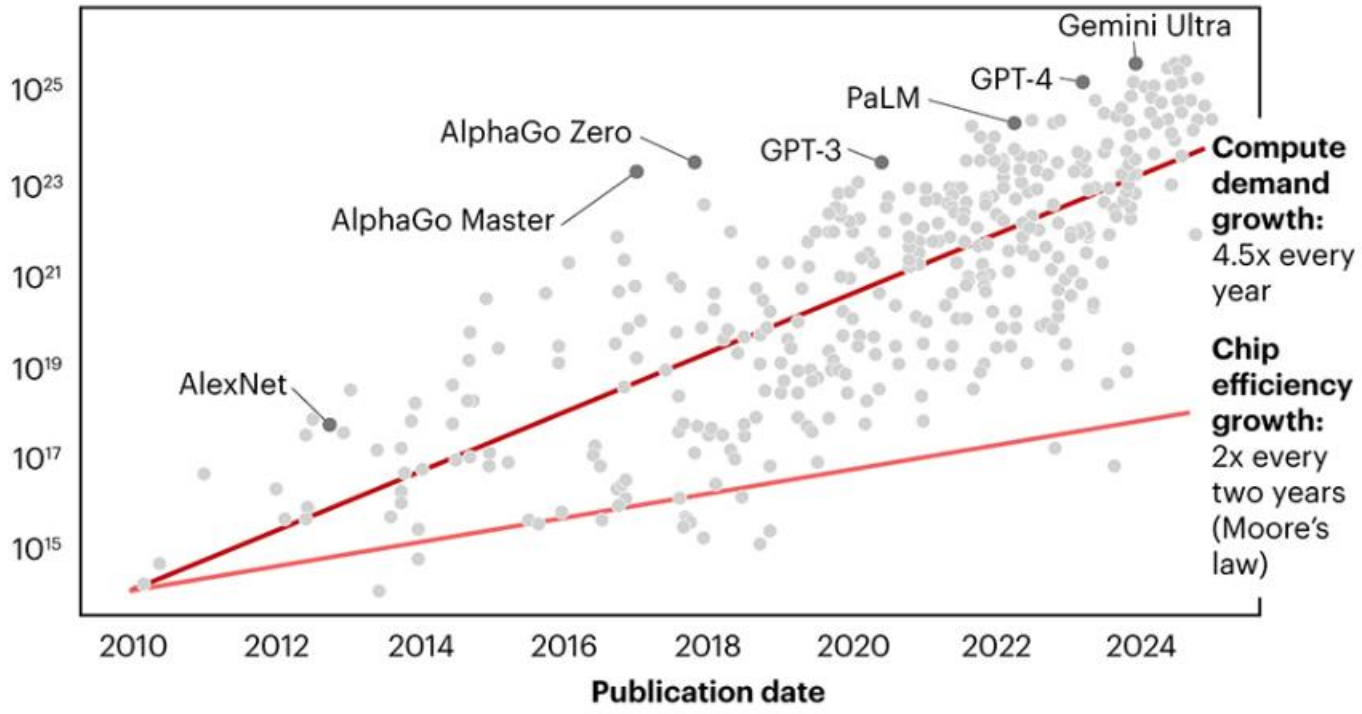
AI COMPUTE IS GROWING FASTER THAN MOORE'S LAW

SEMICONDUCTORS ARE CHANGING RAPIDLY

Artificial intelligence is no longer just a software trend, it is becoming one of the defining infrastructure challenges of the modern technology industry. Public attention often focuses on deepfakes, chatbots, and low-quality AI-generated content, but the larger impact is happening behind the scenes as businesses increasingly rely on AI for software development, medical imaging, fraud detection, logistics, customer support, cybersecurity, financial modeling, manufacturing and scientific research. These applications are not experimental side projects anymore; many companies are beginning to treat AI as a core productivity tool and competitive advantage. As a result, data centers are being asked to support increasingly large models, higher user demand, and constant inference workloads every time an AI system generates a response, analyzes an image, summarizes a document, or recommends an action.

This demand has made AI infrastructure a hot topic because the physical requirements are difficult to ignore. Large-scale AI systems require massive numbers of accelerators, high-bandwidth memory, advanced networking, cooling systems, and reliable power delivery. The semiconductor industry can no longer rely on Moore's Law in the same way it once did. For decades, performance improved as transistor density increased, allowing chips to become faster, cheaper, and more efficient over time. However, physical scaling limits, power constraints, memory bottlenecks, and the cost of advanced manufacturing have slowed that trend. AI compute demand is now growing much faster than traditional chip scaling can support, creating a major need for new acceleration methods, specialized hardware, and full-stack infrastructure designed specifically for AI workloads.

Training Compute Growth (FLOP)



Source: Bain & Company, https://www.consultancy-me.com/news/11792/ais-insatiable-demand-for-compute-power-is-smashing-moores-law

WHY TRADITIONAL SCALING IS NOT ENOUGH

PERFORMANCE GAINS NO LONGER COME FROM SMALLER TRANSISTORS

How Are Chip Companies Meeting Demand?

Chip companies must adapt to a new era of semiconductor trends. These developments affect everyone from established blue-chips to startups, and innovative approaches are rewarded.

Moore's Law Era

- Smaller transistors
- Faster CPUs
- Chip-level scaling

AI Acceleration Era

- ✓ Specialized accelerators
- ✓ Parallel GPU/NPU systems
- ✓ System-level scaling

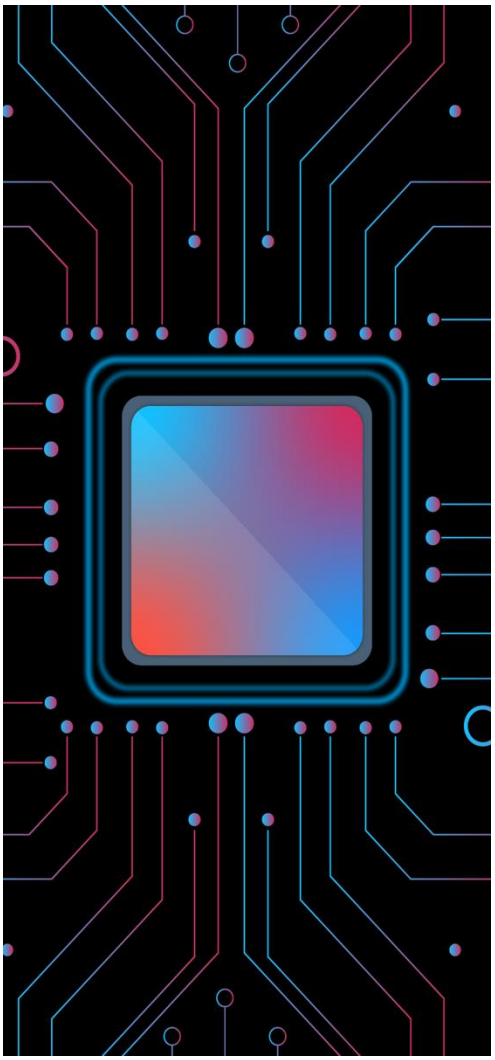
AI Acceleration is Needed to Solve the Moore's Law Bottleneck

Compute demand is growing fast, faster than Moore's law can support. AI models require enormous amounts of computation for training and inference, but traditional improvements in transistor density and CPU performance are no longer increasing fast enough to keep up. Acceleration is needed to provide larger performance gains through parallelism and specialized hardware rather than relying only on smaller transistors.

AI workloads are naturally parallel. Neural networks rely heavily on matrix multiplication and tensor operations, which can be split across many processing units at once. This makes GPUs, NPUs, and other accelerators much better suited for AI than general-purpose CPUs, because they are designed to process many similar operations simultaneously.

Memory bandwidth and data movement limit performance. AI systems must constantly move model weights, activations, and training data between memory and processors. Accelerated systems use technologies like high-bandwidth memory, advanced packaging, and faster interconnects to reduce these bottlenecks and keep compute units supplied with data.

AI performance now depends on full-system optimization. Traditional chip scaling focused mainly on improving individual processors, but AI requires coordination between accelerators, memory, networking, software frameworks, and data center infrastructure. Acceleration is needed because the performance problem has shifted from a single-chip issue to a full-stack system challenge.



AI ACCELERATOR TYPES

OPTIMIING MODEL TRAINING TO INFERENCE

From GPUs to TPUs, multiple approaches are available to train models and provide lightning-fast inference.

GPUs – Graphics Processing Units

GPUs are massively parallel processors originally designed for graphics but now widely used for AI because they can perform thousands of similar mathematical operations at once. They are especially useful for AI training, large-scale inference, simulation, and high-throughput computing because neural networks rely heavily on matrix multiplication and tensor operations. NVIDIA is the clear leader through its CUDA ecosystem, Tensor Cores, NVLink, and Hopper/Blackwell platforms, while AMD competes with Instinct accelerators and ROCm, and Intel participates through Xe GPUs and other data-center accelerator products.

NPUs – Neural Processing Units

NPUs are designed to run AI inference efficiently with lower power consumption than CPUs or GPUs. They are commonly used in phones, laptops, edge devices, automotive systems, and other environments where AI must run locally and efficiently rather than in a large data center. Major companies using NPUs include Apple with its Neural Engine, Qualcomm with Snapdragon and Hexagon NPU technology, Intel with Core Ultra NPUs, AMD with Ryzen AI, and Microsoft through the Copilot+ PC ecosystem.

TPUS – Tensor Processing Units

TPUs are custom tensor accelerators designed specifically for machine learning workloads rather than general-purpose computing. They are optimized for neural network training and inference, especially large matrix and tensor operations used in transformer models and generative AI. Google is the main company associated with TPUs, using them heavily in Google Cloud and its internal AI infrastructure as an alternative to GPU-based acceleration.

AI ACCELERATOR TYPES

OPTIMIING MODEL TRAINING TO INFERENCE (CONTINUED)

ASICs – Application-Specific Integrated Circuit

ASICs are custom chips designed for a specific workload or narrow set of workloads, making them less flexible than GPUs but often more efficient for the tasks they are built to perform. In AI, they are used for large-scale training, inference, recommendation systems, and cloud AI services where hyperscalers can justify custom silicon to improve performance per watt and reduce long-term cost. Major examples include AWS Trainium and Inferentia, Google TPU, Meta MTIA, Microsoft Maia, and custom AI chips developed with semiconductor partners.

FPGAs – Field-Programmable Gate Array

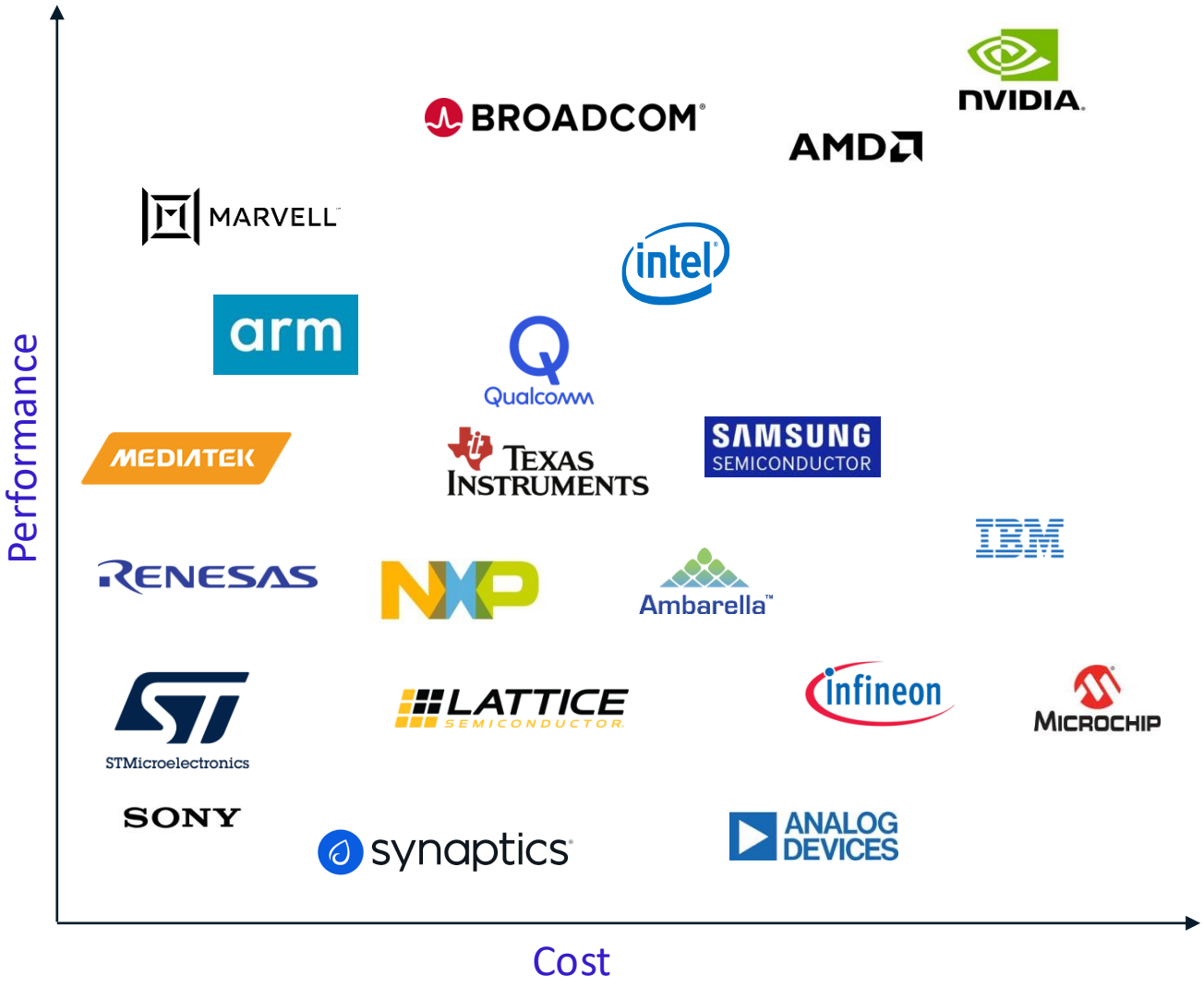
FPGAs are reconfigurable accelerators that can be programmed after manufacturing to create custom hardware pipelines. They are useful for low-latency inference, networking, signal processing, embedded systems, and specialized workloads where flexibility matters more than maximum training throughput. Major FPGA companies include AMD through Xilinx, Intel through Altera, and Microsoft, which has used FPGAs in Azure infrastructure for acceleration and networking.

Wafer-scale and Dataflow Accelerators

These accelerators use alternative architectures that move beyond the traditional GPU model by focusing on reducing data movement, improving throughput, or scaling compute in unusual ways. These systems are often aimed at large-model inference, AI training, or low-latency AI services where standard GPU clusters may face communication and efficiency bottlenecks. Major companies include Cerebras with wafer-scale processors, Groq with deterministic low-latency inference, SambaNova with reconfigurable dataflow architecture, and Tenstorrent with scalable AI processors.

ACCELERATOR MARKET ECOSYSTEM

TOP 20 SEMICONDUCTOR VENDORS



Insights and Products

NVIDIA is the clear market leader in GPU with the strongest full-stack position across GPU hardware and rack-scale AI systems.

- Blackwell B200 / GB200: B200 is a data-center GPU used for LLM training, large-scale inference, generative AI, and HPC. GB200 connects 36 CPUs and 72 GPUs in a rack-scale system, acting as one large AI compute system.
- GB300 NVL72 / Blackwell Ultra: GB300 integrates 72 Blackwell GPUs and 36 CPUs in a rack-scale AI system with 1.5x denser Tensor Core FLOPS and 2x higher attention performance compared to earlier models.
- Vera Rubin NVL72: Full-stack AI supercomputer that can train with a quarter of the GPUs and at one-tenth the cost per million tokens compared to Blackwell
- CUDA / TensorRT: AI software ecosystem made to allow for easier programming and optimization in hardware

Insights and Products (continued)

AMD provides alternative GPU competition, still catching up with software ecosystem maturity.

- Instinct MI300X / MI300A: MI300 is the major NVIDIA competitor for high-memory AI workloads, combining CPU and GPU elements for HPC-style accelerated computing.
- MI3125X: Began availability in Q4 2024, created as an intermediate step to MI3150. Earlier than expected release shows AMD's attempts at a move to faster AI product cadence.
- MI350: CDNA 4 data-center GPU, delivers 35x increase in interference performance compared to MI300.
- MI400 (future): Part of 2026 product family, shows movement towards full platform competition beyond just GPUs.
- ROCm: Open GPU software stack, provides drivers, APIs, compilers, and libraries to allow for parallel processing.

Intel is working on becoming a major player by using a heterogenous approach (CPUs for inference and edge, GPUs for large workloads and training).

- Gaudi 3: Intel's clearest dedicated AI accelerator, used for LLMs, multimodal models, training and inference.
- OpenVINO: Inference optimization toolkit deployed across CPUs, GPUs, NPUs, and edge/server environments.
- Jaguar Shores (future): AI strategy roadmap that shifts direction towards larger integrated platforms, not just standalone accelerator chips.

Qualcomm is strong in mobile/edge NPUs, entering data-center inference with cost-per-watt strategy.

- Hexagon NPU / Snapdragon AI Engine: On-device AI built for efficient inference under tight power constraints.
- Cloud AI 100: Data-center inference accelerator, Qualcomm's earlier attempt to expand past edge AI
- AI200: Inference targeted rack-scale accelerator focusing on high performance per $ per watt, large memory, and scalability
- AI250 (future): Introduces new memory architecture intended to improve bandwidth and efficiency, still focused on inference.

Broadcom is focusing on minimizing power consumption through custom AI accelerators.

- Custom AI XPUs: Custom AI accelerators built with large cloud customers for workload-specific training and inference.
- 3.5D XDSiP: Packing platform for custom accelerators, combining multiple techniques for higher performance and efficiency
- Tomahawk Ethernet Switching: Massive bandwidth requirements resulted in an AI data-center Ethernet switch silicon with 102.4 Tbps.
- Jericho Fabric Chips: Complements Tomahawk for large-scale networking, interconnects over one million XPUs across data centers.

Insights and Products (continued)

Marvell is infrastructure-focused, supports hyperscaler custom silicon, advanced packaging, AI networking.

- Custom AI ASICs: Customer-specific AI chips designed for cloud providers building proprietary AI training and inference systems.
- PAM4 Optical DSPs: Supports high-bandwidth links to carry data between racks and data centers.
- NVLink Fusion (Future NVIDIA partnership): Connects third-party XPUs into NVIDIA-style infrastructure, focuses on compatibility and interconnection.

ARM is a foundational IP provider leading in CPU/NPU architecture.

- Ethos-U55 / Ethos-U65: Embedded microNPU IP used for tinyML, IoT, MCU-class AI, and low-power local inference.
- Ethos-U85: Higher-performance edge NPU IP designed to support transformer models, smart cameras, and factory automation.
- Ethos-N: Higher-end NPU IP used for application-processor-class edge AI and more demanding inference workloads.
- Corstone Reference Platforms: Pre-integrated CPU, NPU, memory, and system IP platforms that help companies build AI chips faster.

MediaTek is a mobile SoC vendor, uses scalable NPUs on smartphones, IoT, and edge AI devices.

- Dimensity 9500: Flagship mobile SoC used for premium Android phones, mobile generative AI, camera AI, and real-time translation.
- NPU 990: Mobile NPU inside Dimensity 9500 designed for faster on-device token generation and lower AI power consumption.
- CIM-Based NPU Direction: Compute-in-memory NPU design intended to reduce data movement and improve mobile AI efficiency.

Texas Instruments is strong in embedded and industrial edge-AI, focused on low-power real-time inference.

- TDA4VM / Jacinto: Edge AI vision processor used for ADAS, robotics, industrial vision, radar, and perception analytics.
- C7x DSP + Matrix Multiply Accelerator: Embedded AI compute architecture used for neural-network inference, signal processing, vision, and sensor fusion.
- AM68A / AM69A: Industrial edge AI processors used for smart cameras, robotics, factory inspection, and machine vision.
- TIDL: TI Deep Learning software stack used to compile, optimize, and deploy neural networks on TI processors.

Insights and Products (continued)

Samsung Semiconductor is important in mobile AI processors and AI infrastructure memory, especially HBM.

- Exynos NPU: Mobile NPU inside Samsung SoCs used for smartphone AI, camera enhancement, voice AI, and on-device inference.
- HBM3E: High-bandwidth memory used in AI GPUs and accelerators to support large models and high memory bandwidth workloads.
- HBM4: Next-generation high-bandwidth memory designed for future AI accelerators with greater bandwidth and efficiency.
- HBM4E: Future AI memory product aimed at improving speed, energy efficiency, and thermal performance for next-generation AI infrastructure.

Renesas is strong in industrial and vision edge AI, focused on high performance-per-watt for embedded inference.

- RZ/V Series: Embedded AI MPU family used for industrial vision, robotics, smart factories, cameras, and endpoint inference.
- DRP-AI: Dynamically reconfigurable AI accelerator designed for low-power neural inference, especially in embedded vision applications.
- RZ/V2H: Higher-end embedded AI MPU used for robotics, multi-camera vision, industrial automation, and higher-performance endpoint AI.
- RZ/V2N: Mid-class vision AI MPU with DRP-AI3 used for smart factories, intelligent cities, and efficient edge vision.

Infineon is an edge-AI participant focused on secure, low-power sensor/HMI-oriented microcontrollers.

- PSOC Edge E84: AI-capable microcontroller used for wearables, smart home, HMI, voice AI, sensing, and IoT products.
- Ethos-U55 Integration: Embedded NPU inside PSOC Edge used for local neural-network inference on low-power devices.
- NNLite Accelerator: Ultra-low-power hardware accelerator used for always-on sensing and lightweight ML tasks.

Ambarella is a specialized leader in low-power computer vision AI and edge video systems.

- CVflow: Ambarella's core AI vision accelerator architecture used for computer vision, video analytics, and low-power edge inference.
- CVflow 3.0: Next-generation AI engine designed for edge GenAI, vision-language models, and reasoning workloads.
- CV7: 4nm edge AI 8K vision SoC used for multi-stream video, security cameras, drones, robotics, and automotive systems.
- CV-Series Vision SoCs: Broader family of vision processors combining video processing, image processing, and neural inference.

Insights and Products (continued)

Lattice utilizes FPGAs for small-footprint, low-power edge inference.

- sensAI / sensAI Studio: AI software stack used to train, optimize, and deploy machine-learning models onto Lattice FPGAs.
- CrossLink-NX: Low-power FPGA family used for vision pipelines, sensor bridging, robotics, and near-sensor preprocessing.
- Nexus / Certus / iCE40 FPGA Families: Low-power reconfigurable silicon used for always-on sensing, compact AI inference, and custom low-latency logic.
- NVIDIA / TI Collaboration Direction: Lattice is positioning its FPGAs as companion chips for sensor fusion and larger edge AI systems.

Microchip Technology uses low-power edge inference in industrial, medical, and defense fields.

- PolarFire FPGA: Low-power FPGA platform used for embedded vision, industrial AI, medical systems, aerospace, and deterministic edge inference.
- PolarFire SoC FPGA: FPGA with a RISC-V processor subsystem used for embedded Linux systems with hardware AI acceleration.

NXP is an established edge-AI player in automotive, industrial, and IoT with scalable NPUs.

- eIQ Neutron NPU: Scalable embedded NPU architecture used across NXP processors for low-power neural-network acceleration.
- i.MX RT700: Crossover MCU with integrated eIQ Neutron NPU used for wearables, smart home, medical devices, and edge AI.
- MCX N Series: Low-power MCU family that brings AI acceleration into lower-cost industrial, IoT, and sensing devices.
- eIQ Toolkit / Neutron SDK: Software tools used to convert and deploy trained AI models onto NXP processors.

Synaptics is a smaller AI silicon player targeting intelligent IoT, multimodal edge AI, and low-power connected devices.

- Astra Platform: AI-native IoT compute platform used for smart home, wearables, industrial IoT, robotics, healthcare, and connected edge products.
- SL2600 / SL2610 Product Line: Edge AI processor family used for multimodal AI, smart appliances, smart retail, robotics, and industrial devices.
- Torq Edge AI Platform: NPU and software platform for Astra products designed to support future edge AI and transformer-capable workloads.

Insights and Products (continued)

Analog Devices is relevant for tinyML and battery-powered inference, not established in mainstream acceleration.

- MAX78000: Ultra-low-power AI microcontroller with a CNN accelerator used for tinyML, sensor classification, face detection, and audio/vision inference.
- MAX78002: Higher-capability AI microcontroller used for larger tinyML tasks such as face identification, object detection, and audio classification.
- MAX7800x Development Tools: Software tools used to train, quantize, and deploy CNNs onto MAX78000 and MAX78002 devices.

STMicroelectronics is a strong low-cost embedded AI vendor focused on real-time neural inference in microcontrollers.

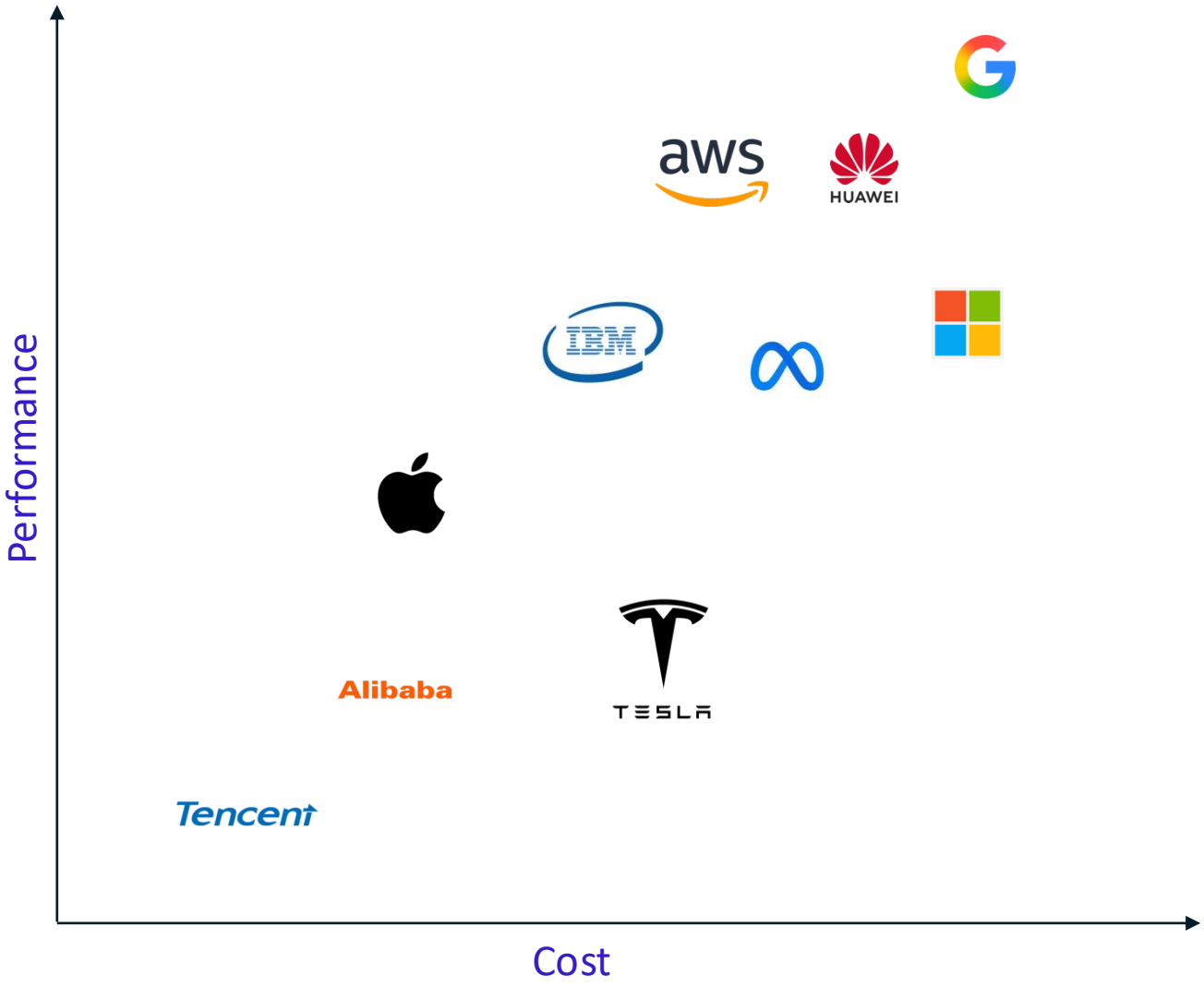
- STM32N6: AI-focused microcontroller used for computer vision, audio AI, smart sensors, wearables, and embedded edge AI.
- Neural-ART Accelerator: ST's embedded NPU inside STM32N6 that accelerates neural-network inference directly on the microcontrollers
- TM32Cube.AI / ST Edge AI Developer Cloud: Software tools used to optimize and deploy trained neural-network models onto STM32 devices.

Sony focuses exclusively on computer vision, moved inference directly into image sensor for low-latency edge vision.

- IMX500: Intelligent vision sensor with AI processing built directly into the image sensor for smart cameras, retail analytics, industrial vision, and privacy-sensitive edge AI.
- AITRIOS: Edge AI sensing and management platform used to build, deploy, and manage vision AI applications using Sony intelligent image sensors.
- Raspberry Pi AI Camera: Developer-accessible camera module using Sony's IMX500 sensor for compact local vision AI prototyping.
- In-Sensor AI Research Direction: Future direction focused on running more advanced vision models directly inside the sensor with lower latency and bandwidth use.

ACCELERATOR MARKET ECOSYSTEM

TOP 10 HYBRID VENDORS



Insights and Products

Google is the strongest hybrid AI accelerator company, with a mature TPU ecosystem used for large-scale AI training, inference, and Google Cloud infrastructure.

- TPU v5e / v5p: Cloud TPU generations used for large-scale AI training and inference, with v5e focused more on cost-efficient workloads and v5p focused more on high-performance training.
- Trillium TPU: Sixth-generation TPU used for improved performance-per-dollar, energy efficiency, and large AI model training/inference in Google Cloud.
- Ironwood TPU: Seventh-generation TPU designed for large-scale inference, reasoning models, and high-throughput AI serving.
- AI Hypercomputer: Google's full-stack AI infrastructure platform combining TPUs, GPUs, CPUs, storage, networking, and software orchestration for large AI workloads.

Insights and Products (continued)

AWS is a major hybrid AI accelerator provider focused on lowering the cost of cloud training and inference through custom silicon.

- Trainium: AWS custom AI training chip used for large-scale deep learning, generative AI, and cost-efficient model training.
- Trainium2: Newer AWS training accelerator used for larger generative AI models, higher performance, and improved scaling across AWS infrastructure.
- Inferentia: AWS custom inference chip used to reduce the cost of deep learning and generative AI inference on Amazon EC2.
- Inferentia2: Newer AWS inference accelerator used for higher-performance, lower-cost inference for large language models, recommendation systems, and generative AI services.
- Neuron SDK: AWS software stack used to compile, optimize, and deploy models onto Trainium and Inferentia.

Microsoft is a major cloud hybrid focused on custom AI silicon for Azure, OpenAI workloads, and Microsoft Copilot services.

- Azure Maia 100: Microsoft's custom AI accelerator designed for large-scale AI training and inference inside Azure data centers.
- Maia 200: Microsoft's newer inference-focused accelerator designed to improve token generation economics, memory bandwidth, and low-precision AI serving.
- Azure Cobalt CPU: Microsoft's custom Arm-based CPU used alongside AI accelerators to improve cloud infrastructure efficiency.
- Azure AI Infrastructure: Full-stack Microsoft system combining custom silicon, liquid cooling, networking, servers, and software for Azure AI and Copilot workloads.

Meta is an internal-scale hybrid accelerator company focused on inference, recommendation systems, ranking, and social media AI workloads.

- MTIA v1: First-generation Meta Training and Inference Accelerator designed to improve efficiency for Meta's internal recommendation and ranking workloads.
  Next-generation MTIA: Updated Meta accelerator with improved compute, memory capacity, and bandwidth for production AI inference.
- MTIA 300: Meta inference chip used for ranking and recommendation systems across Meta's large-scale platforms.
- MTIA 400 / 450 / 500: Future Meta accelerator roadmap aimed at expanding workload coverage, especially for AI inference and user-facing AI services.
- PyTorch / Triton Integration: Meta's software path for integrating internal accelerator hardware with AI model development and deployment.

Insights and Products (continued)

Huawei is a vertically integrated AI infrastructure hybrid focused on Ascend processors, Atlas systems, and large-scale AI compute.

- Ascend 910 / 910B / 910C: Huawei AI processors used for model training, inference, and large-scale AI infrastructure.
- Atlas 800 Training Server: Huawei AI training server using Ascend processors for deep learning, smart city, healthcare, scientific, and industrial AI workloads.
- Atlas 900 A3 SuperPoD: Large-scale Huawei AI system using Ascend processors and designed to operate as a large AI compute cluster.
- CloudMatrix: Huawei cloud AI architecture built around Ascend processors and high-bandwidth interconnect for large AI model training and serving.
- CANN Software Stack: Huawei's software ecosystem used to develop, optimize, and deploy AI models on Ascend hardware.

IBM is an established enterprise technology company relevant for AI acceleration through research-driven and enterprise-focused chip programs, but it is not a leading hyperscale accelerator vendor.

- Artificial Intelligence Unit: IBM's AI accelerator designed to support enterprise AI workloads with efficient neural-network inference and lower-precision arithmetic.
- NorthPole: IBM's experimental brain-inspired AI chip architecture focused on reducing data movement by placing memory closer to compute for efficient inference.
- Telum Processor / Telum II: IBM mainframe processor family with integrated AI acceleration used for real-time enterprise inference, fraud detection, financial workloads, and transaction processing.
- Spyre Accelerator: IBM AI accelerator direction used to bring additional generative AI and enterprise inference capability to IBM Z and LinuxONE systems.

Apple is the strongest consumer-device hybrid AI accelerator company, focused on efficient on-device inference rather than mainstream data-center acceleration.

- Apple Neural Engine: Dedicated neural processing hardware built into Apple A-series and M-series chips for on-device machine learning and AI inference.
- M-series GPU Neural Accelerators: Apple Silicon direction that adds neural acceleration into GPU cores to improve local AI workloads such as diffusion models and LLM inference.
- Core ML: Apple's supported software framework for deploying optimized machine learning models onto Apple hardware.
- MLX: Apple machine learning framework used for efficient model development and local AI experimentation on Apple Silicon.
- Apple Intelligence Hardware Platform: Apple's device-level AI approach combining CPU, GPU, Neural Engine, unified memory, and software frameworks for private local inference.

Insights and Products (continued)

Tesla is a vertical AI hardware company focused on autonomous driving and robotics rather than general cloud AI acceleration.

- FSD Computer / Hardware 3: Tesla's in-vehicle AI computer used for real-time autonomous driving inference from vehicle camera data.
- Hardware 4: Newer Tesla in-vehicle inference platform with improved camera and compute capability for Full Self-Driving and driver-assistance workloads.
- AI5 / AI6: Tesla's future AI chip direction focused on real-time inference for vehicles and robotics.
- Dojo D1: Tesla's custom training chip originally designed for large-scale video-based autonomous driving training.

Alibaba is a cloud and e-commerce hybrid with custom AI inference silicon through T-Head, but it is less established in current mainstream acceleration than Google, AWS, Microsoft, or Huawei.

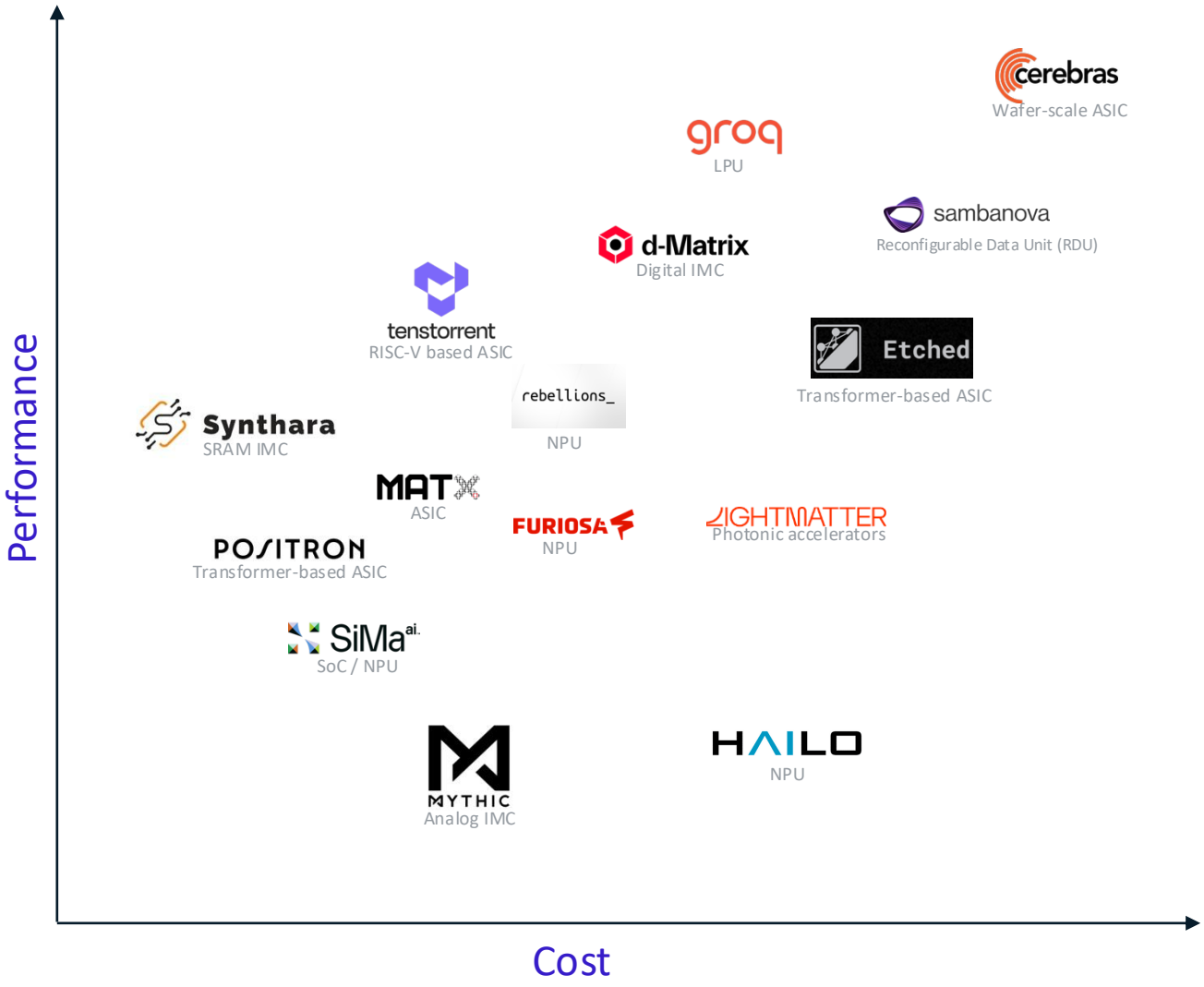
- Hanguang 800: Alibaba's custom AI inference chip designed for high-throughput image, video, search, recommendation, and e-commerce AI workloads.
- Alibaba Cloud AI Inference Instances: Cloud infrastructure using Alibaba's AI acceleration resources for enterprise inference and internal Alibaba workloads.
- T-Head Semiconductor Unit: Alibaba's chip design group responsible for Hanguang and other custom silicon efforts.

OpenAI is an emerging hybrid AI infrastructure company developing custom accelerators to reduce dependence on third-party GPUs and improve large-scale inference economics.

- Jalapeño Intelligence Processor: OpenAI's first custom AI accelerator, designed with Broadcom for LLM inference, token generation efficiency, memory movement, networking, and large-scale AI serving.
- OpenAI-Broadcom Custom Accelerator Program: Multi-generation custom AI accelerator effort planned around 10 gigawatts of deployment, with OpenAI designing accelerators and systems and Broadcom supporting development and deployment.
- AMD Instinct Deployment Partnership: OpenAI's large-scale deployment agreement for 6 gigawatts of AMD Instinct GPUs, showing that its infrastructure strategy combines custom silicon with external accelerators.

ACCELERATOR MARKET ECOSYSTEM

TOP 15 STARTUPS



Insights and Products

Cerebras is the leading wafer-scale AI accelerator company, focused on replacing multi-GPU clusters with extremely large chips for high-performance training and inference.

- WSE-3: Wafer-Scale Engine 3 is Cerebras' third-generation wafer-scale AI chip, built with 4 trillion transistors, 900,000 AI-optimized cores, and extremely high on-chip memory bandwidth for large AI workloads.
- CS-3 System: Third-generation Cerebras AI system built around WSE-3, used for large-scale model training, inference, scientific computing, and high-performance AI clusters.
- Cerebras Wafer-Scale Cluster: Scaled-out Cerebras systems used to train and serve large models while reducing the complexity of distributing workloads across many separate GPUs.
- Cerebras Inference Cloud: Cloud inference service used to run open models and serve LLM workloads with very high token throughput.

Insights and Products

SambaNova is a mature AI accelerator startup focused on dataflow computing for enterprise-scale inference, fine-tuning, and large-model deployment.

- SN40L RDU: Reconfigurable Dataflow Unit used for enterprise AI inference and training, especially large language models, mixture-of-experts, and workloads limited by the memory wall.
- SN50 RDU: Newer SambaNova accelerator designed for agentic inference, with higher compute and network bandwidth than the SN40 generation.
- DataScale / SambaNova Suite: Full-stack AI system combining RDU hardware, software, models, and enterprise deployment tools for private AI infrastructure.
- SambaNova Composition of Experts: Model deployment approach that uses many smaller expert models to reduce inference cost and complexity compared with massive monolithic models.

Groq is a low-latency AI inference accelerator company focused on extremely fast token generation for LLMs and real-time AI applications.

- LPU: Language Processing Unit designed specifically for deterministic, low-latency inference rather than general-purpose GPU-style acceleration.
- GroqChip: Custom inference processor used inside Groq systems to accelerate large language models, speech, vision, and generative AI workloads.
- GroqRack: Rack-scale inference system combining many Groq accelerators for high-throughput, low-latency AI serving.
- GroqCloud: Developer-facing cloud platform used to access Groq inference hardware through APIs for text, audio, and vision models.

d-Matrix is a memory-centric AI inference startup focused on reducing latency, power, and cost for generative AI in datacenters.

- Corsair: d-Matrix AI inference platform designed for datacenter generative AI, especially low-latency LLM inference, reasoning, agents, and video generation.
- 3DIMC Architecture: In-memory computing architecture that brings compute closer to memory to reduce data movement and improve inference energy efficiency.
- JetStream: Networking acceleration technology used to scale d-Matrix inference systems across servers and racks.
- Aviator Software: d-Matrix software stack used to compile, optimize, and deploy models onto Corsair inference hardware.
- SquadRack: Rack-scale d-Matrix system designed for high-throughput, sustainable generative AI inference at datacenter scale.

Insights and Products

Tenstorrent is a RISC-V-based AI accelerator startup focused on scalable, programmable AI processors and developer-accessible hardware.

- Wormhole: Tenstorrent AI accelerator architecture using Tensix cores and a mesh network-on-chip, available in PCIe card form for AI and research workloads.
- Blackhole: Newer Tenstorrent AI processor family designed for higher-performance AI acceleration, developer systems, and scalable server deployments.
- Tenstorrent Galaxy: AI server platform using Tenstorrent accelerators for scalable, lower-cost AI compute.
- TT-Metalium / Software Stack: Tenstorrent's low-level programming and software environment used to develop, optimize, and deploy workloads on its AI processors.

Synthara is a compute-in-memory AI startup focused on improving AI efficiency by performing computation directly inside SRAM memory, reducing data movement for ultra-low-power edge inference.

- ComputeRAM: SRAM-based compute-in-memory architecture that performs matrix operations directly within memory arrays, significantly reducing memory access and power consumption for AI workloads.
- Synthara AI IP: Licensable compute-in-memory accelerator IP that integrates into custom SoCs to accelerate neural network inference, DSP, and embedded AI applications.
- Edge AI Platform: Ultra-low-power AI solution designed for always-on sensing, IoT devices, wearables, industrial automation
- Synthara Software Development Kit: Software tools and compiler that optimize and deploy AI models onto Synthara's compute-in-memory architecture for efficient execution.

FuriosaAI is a South Korean AI accelerator startup focused on power-efficient datacenter inference for LLMs and multimodal models.

- RNGD / Renegade: FuriosaAI's flagship datacenter AI accelerator using Tensor Contraction Processor architecture for efficient LLM and multimodal inference.
- RNGD PCIe Card: Server accelerator card using RNGD for lower-power AI inference deployment in enterprise and datacenter systems.
- NXT RNGD Server: Multi-card FuriosaAI server platform designed to scale RNGD accelerators for higher-throughput inference.

Lightmatter is a photonic AI infrastructure startup focused on solving data movement and interconnect bottlenecks in AI systems.

- Passage: Photonic interconnect platform designed to connect chips and chiplets with very high bandwidth and lower power for large AI systems.
- Passage M1000: Photonic superchip/interposer direction designed to support next-generation AI infrastructure with dense chip-to-chip communication.
- Guide: Light engine platform used to support optical connectivity for AI infrastructure.

Insights and Products

Rebellions is a South Korean AI accelerator startup focused on datacenter inference and energy-efficient large-model deployment.

- ATOM: AI inference accelerator built for datacenter workloads, using Samsung 5 nm process technology and designed for efficient neural-network inference.
- REBEL: Newer Rebellions accelerator direction using chiplet architecture and HBM for large-scale AI inference.
- REBEL-Quad: Multi-chiplet / multi-accelerator system direction aimed at higher-performance LLM inference and energy-efficient datacenter deployment.
- Rebellions Software Stack: Deployment tools used to integrate Rebellions accelerators into datacenter AI workflows.

MatX is a newer AI accelerator startup focused on high-throughput chips for frontier-scale large language models.

- MatX One: AI accelerator chip designed for high-throughput LLM training, prefill, decode, reinforcement learning, and long-context workloads.
- MatX Server Direction: System-level deployment approach focused on serving the large-model needs of frontier AI labs.
- LLM-Specific Architecture: MatX's design direction focuses on optimizing for transformer-heavy large model workloads rather than broad general-purpose acceleration.
- MatX Software Direction: Software and compiler path intended to make large AI models run efficiently on MatX hardware.

Positron AI is an inference-focused startup building transformer accelerators for lower-power LLM serving.

- Atlas: Positron's first inference accelerator system, using multiple Archer accelerators for transformer model inference with lower power draw than GPU-based systems.
- Archer Accelerator: Positron's accelerator chip used inside Atlas systems for LLM inference and high-efficiency token generation.
- Asimov: Positron's next-generation accelerator direction aimed at larger models, higher memory capacity, and more scalable transformer inference.

Hailo is an edge AI accelerator startup focused on efficient local inference rather than datacenter-scale AI training.

- Hailo-8: Edge AI accelerator delivering up to 26 TOPS for real-time computer vision, object detection, smart cameras, robotics, industrial, and automotive edge workloads.
- Hailo-8L: Lower-power edge accelerator used for embedded AI applications such as Raspberry Pi AI kits, smart sensors, and compact vision systems.
- Hailo-10H: Edge generative AI accelerator designed to run LLMs, VLMs, and other generative AI workloads locally on PCs, automotive systems, and edge devices.
- Hailo AI Software Suite: Tools used to compile, optimize, and deploy neural networks onto Hailo edge AI processors.

Insights and Products

SiMa.ai is an edge AI startup focused on machine learning system-on-chip platforms for industrial, robotics, automotive, and embedded.

- AIMLSoC: Machine learning system-on-chip designed to run computer vision and AI inference workloads at the edge with low power and deterministic performance.
- Palette Software: SiMa.ai software platform used to deploy, optimize, and manage AI models on MLSoC hardware.
- Modalix Platform: Newer SiMa.ai edge AI platform direction designed for multimodal edge AI and more flexible deployment.
- Edge AI Developer Tools: Software and model optimization tools used to move trained AI models onto SiMa.ai embedded hardware.

Etched is a specialized AI chip startup focused on transformer inference rather than general-purpose acceleration.

- Sohu: Transformer-specific ASIC designed to accelerate transformer models such as LLMs and diffusion-style models by hardwiring transformer operations into silicon.
- Sohu Server: Multi-chip inference server designed to deliver extremely high token throughput for large language model serving.
- Transformer ASIC Architecture: Etched's core approach, trading general-purpose flexibility for much higher performance on transformer-based AI workloads.

Mythic is an analog/in-memory AI accelerator startup focused on low-power edge inference.

- M1076 Analog Matrix Processor: Mythic AI processor using analog compute-in-memory to accelerate neural-network inference with low power consumption.
- M.2 Accelerator Card: Mythic deployment card used for edge AI inference, computer vision, drones, robotics, and smart camera workloads.
- Analog Compute-in-Memory Architecture: Mythic's core approach, storing model weights close to computation to reduce data movement and energy use.
- Mythic Software Tools: Development tools used to compile and deploy neural networks onto Mythic analog AI processors.

CONCLUSION

WHAT DOES THIS MEAN FOR THE INDUSTRY?

Compute demand will strain supply chains. By outpacing semiconductor efficiency, current infrastructure is not equipped to meet this demand. As a result, bottlenecks will emerge not only in chip production, but also in memory, advanced packaging, networking, power delivery, and data center deployment.

Space is limited. As we approach the physical limit for transistor density on a chip, progress will have to be made in other areas to meet the compute demand. This will be in the form of parallel computing, specialized accelerators, high bandwidth memory, and full-stack infrastructure.

Power will become a limiting factor. As compute demand grows, data centers will require significantly more electricity and cooling, making energy efficiency just as important as raw performance.

Memory and data movement are becoming bottlenecks. Even if processors get faster, moving data between memory, chips, and servers consumes time and power, forcing innovation in HBM, interconnects, and chip packaging.

Compute will shift from chip-level scaling to system-level scaling. Instead of relying only on smaller transistors, companies will scale performance through chiplets, multi-GPU clusters, rack-scale systems, and optimized software ecosystems.

The future of AI acceleration will likely be defined by companies that can solve the entire compute problem, not just produce the fastest individual chip. As Moore's Law slows and AI demand continues to grow, success will depend on balancing raw performance with power efficiency, memory bandwidth, software support, scalability, and supply chain execution. GPUs will likely remain the dominant platform in the near term because of their flexibility, mature software ecosystems, and strong deployment base, especially through NVIDIA and its full-stack infrastructure. However, custom accelerators from hyperscalers such as Google, AWS, Microsoft, Meta, and OpenAI will become increasingly important as companies look to reduce the cost of training and inference at massive scale. At the same time, edge AI accelerators, NPUs, and embedded inference chips will grow as more AI workloads move onto phones, laptops, vehicles, cameras, and industrial devices.

Over time, the most successful architectures will not necessarily be the most powerful in isolation, but the ones that fit their workload best. Training large frontier models, serving billions of LLM requests, running private on-device assistants, and processing real-time sensor data all require different tradeoffs. This means the market will likely become more specialized rather than converging on a single universal accelerator. The winners will be companies that combine efficient hardware with high-bandwidth memory, advanced packaging, fast interconnects, reliable manufacturing, and developer-friendly software ecosystems. AI acceleration is therefore not just a response to growing compute demand; it represents a broader shift in the semiconductor industry from transistor-level scaling to full-system optimization.