



Maratha Vidya Prasarak Samaj's
Karmaveer Adv. Baburao Ganapatrao Thakare
College Of Engineering
Nashik-13.



(NAAC ACCREDITED INSTITUTE WITH 'A' GRADE)

DEPARTMENT OF ELECTRONICS & TELECOMMUNICATION ENGG.

Departmental **TeChronicle**

Month: - April 2026

Vol. - 08, Issue – 1

Department Vision:-

To be recognized as an excellent department offering competent technical education to create competent electronics & telecommunication engineers for the benefit of the common masses.

Department Mission:-

Committed to serve the needs of society through innovative teaching learning processes, promoting industry-institute interaction to provide competent and cultured electronics and telecommunication engineers.

Program Educational Objectives:-

- 1. To impart state of art technical education in the Electronics & Telecommunication Engineering.*
- 2. To promote society beneficial projects and activities.*
- 3. To develop soft skill, team work, professional ethics and multidisciplinary approach for the career enhancement.*
- 4. To bridge the gap between Industry-Institute through collaboration with Industries, Institutions and Universities.*
- 5. To provide suitable infrastructure and facilities in tuned with advancing technological evaluation.*

Greeting,

Department of Electronics and Telecommunication
is unveiling technical newsletter "TeChronicle"
VOL-08, ISSUE-4 on 4th April 2026.

The day is observed every year on April 4 on occasion of the birth anniversary of Ex Sarchitnis of Maratha Vidya Prasarak Samaj's (MVPS's, Nashik) late Dr. V. N. Pawar, recipient of the prestigious Dr. B. C. Roy award.

Introduction to the AI Chip Wars

[Name: Nikita Rajendra Gaikwad]

The AI chip market is no longer just about raw computing power—it's about efficiency, scalability, and accessibility. Companies across industries are racing to deploy AI models that can process vast datasets, train large language models, and deliver real-time insights. The processor at the core of these

operations determines not only the speed of innovation but also the cost and environmental impact of AI deployment.

NVIDIA has dominated this space since the early 2010s, largely due to its CUDA platform, which provides developers with a robust ecosystem for parallel computing. But as AI models grow exponentially in size, the limitations of traditional GPUs—particularly in power efficiency and cost—have become apparent. This has opened the door for competitors to innovate.

Key insight: The AI chip market is undergoing a paradigm shift. While NVIDIA remains the gold standard, the rise of alternative architectures—such as custom ASICs, FPGAs, and wafer-scale engines—is forcing the industry to rethink how AI is processed. This shift is driven by

the need for lower latency, reduced power consumption, and more sustainable computing.

In this landscape, four major players stand out: NVIDIA, AMD, Intel, and a new wave of AI startups. Each brings a unique approach to AI acceleration, whether through specialized GPUs, novel chip designs, or revolutionary architectures. The question is no longer if these alternatives will gain traction, but when and how they will reshape the industry.

Why the AI Chip Market Matters

The stakes in the AI chip market extend far beyond corporate profits. The processor you choose can determine the feasibility of your AI project, its long-term sustainability, and even its ethical implications.

For example:

Cost efficiency: High-performance AI chips are expensive. A single NVIDIA H100 GPU can cost tens of thousands of dollars, making it inaccessible for many startups and researchers. Alternatives like AMD's MI300 or Intel's Gaudi 3 aim to provide comparable performance at a lower price point.

Power consumption: Data centers consume 1% of the world's electricity, with AI workloads contributing significantly to this demand. More efficient chips can reduce operational costs and environmental impact.

Performance flexibility: Some AI workloads require specialized hardware. For instance, large language models (LLMs) benefit from tensor cores, while computer vision tasks may perform better on GPUs with higher memory bandwidth.

Future-proofing: As AI models grow larger and more complex, the hardware required to train and run them must evolve. Companies investing in the right chip today may avoid costly upgrades tomorrow.

The Major Players in the AI Chip Market

NVIDIA: The incumbent leader with a comprehensive ecosystem.

AMD: A strong challenger with its Instinct MI-series GPUs.

Intel: Leveraging its Xe architecture and Gaudi accelerators.

Startups: Cerebras, Samba Nova, and others pushing

boundaries with novel designs.

Each of these players brings a unique value proposition to the table.

The Evolution of AI Chips: A Brief History

Early 2010s: GPUs like NVIDIA's Tesla K20 become popular for AI workloads due to their parallel processing capabilities.

Mid-2010s: NVIDIA introduces CUDA cores and tensor cores, accelerating AI training and inference.

Late 2010s: FPGAs and custom ASICs (e.g., Google's TPU) emerge as alternatives for specialized AI tasks.

Early 2020s: Wafer-scale engines (e.g., Cerebras WSE) and heterogeneous architectures gain traction.

2024–2026: The race intensifies with AMD, Intel, and startups introducing chips designed for edge AI, real-time inference, and energy efficiency.

NVIDIA's Key AI Chips

NVIDIA's AI portfolio is built around several key product lines:

Tesla Series: Designed for data center AI training.

GeForce RTX: Consumer GPUs optimized for AI inference and gaming.

Jetson: Embedded AI platforms for edge devices.

Grace Hopper: A combination of CPU and GPU for high-performance computing.

However, the crown jewel of NVIDIA's lineup is the H100 and H200 GPUs, which are built on the Hopper architecture. These chips are designed specifically for AI workloads and offer:

Fourth-Generation Tensor Cores: Deliver up to 9x faster training compared to previous generations.

Transformer Engine: Optimized for training large language models like L.Lama and Mistral.

Multi-Instance GPU (MIG): Allows a single GPU to be partitioned into multiple isolated instances, improving resource utilization.

How NVIDIA's AI Chips Work

NVIDIA's GPUs are optimized for parallel processing, which is essential for AI workloads.

Unlike CPUs, which excel at sequential tasks, GPUs can handle thousands of threads simultaneously, making them ideal for matrix operations that dominate AI training.

The **Hopper architecture** introduces several innovations:

FP8 Precision: Allows mixed-precision training with lower memory usage and faster computation.

Confidential Computing: Ensures secure execution of AI workloads in untrusted environments.

NVLink and NVSwitch: High-speed interconnects that enable multi-GPU scaling for large models.

For example, training a large language model like GPT-4 requires thousands of GPUs working in tandem. NVIDIA's DGX systems are specifically designed for such workloads, offering pre-configured racks of H100 GPUs with optimized cooling and power delivery.

How AMD's AI Chips Work

AMD's Instinct GPUs are built on the CDNA 3 architecture, which introduces several innovations for AI acceleration:

Matrix Cores: Dedicated hardware units for matrix multiplication, the backbone of deep learning.

High-Bandwidth Memory (HBM): AMD's MI300X features HBM3, which provides 5.3 TB/s of memory bandwidth, significantly reducing bottlenecks in training large models.

Unified Memory: The MI300X combines CPU and GPU memory, allowing for seamless data sharing between the two. This is particularly useful for workloads that require frequent CPU-GPU communication, such as inference with large models.

AMD's approach is to provide a balanced solution that doesn't sacrifice performance for cost. For example, the MI300X is designed to handle both training and inference, whereas NVIDIA's H100 is primarily optimized for training.

How Intel's AI Chips Work

Intel's Gaudi accelerators are built on a heterogeneous architecture that combines:

Tensor Processor Cores: Dedicated hardware for matrix multiplication.

Matrix Multiplication Engines: Optimized for AI

workloads.

High-Bandwidth Memory (HBM): Gaudi 3 features HBM2e, providing high memory bandwidth for training large models.

Intel's approach is to provide a flexible solution that can adapt to different AI workloads. For example, Gaudi 3 is designed to handle both training and inference, making it a versatile option for data centers.

Additionally, Intel's Xe architecture is being used in its high-performance GPUs, such as the Arc Alchemist line, which targets gaming and AI workloads alike.

What Distinguishes Each Player?

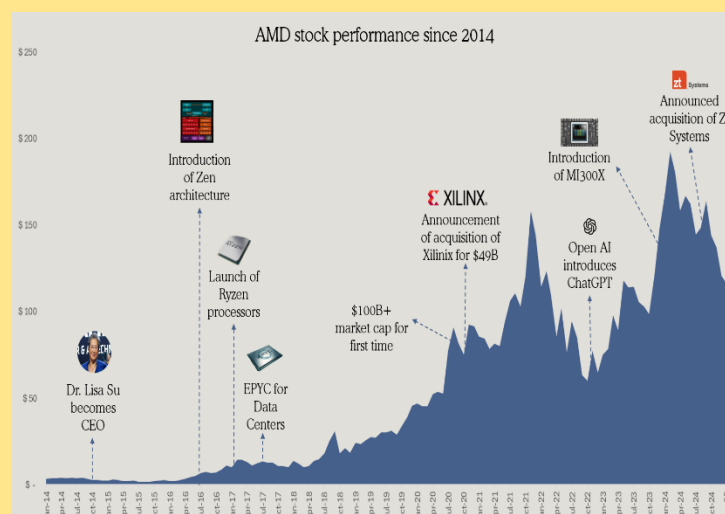
Each company in the AI chip market brings a unique value proposition to the table:

NVIDIA: Dominates the market with its CUDA ecosystem, offering unparalleled software support and developer tools. Its H100 GPU remains the gold standard for AI training.

AMD: Focuses on cost efficiency and open standards with its ROCm platform. The MI300X is a strong contender for organizations looking to avoid vendor lock-in.

Intel: Offers a flexible solution with its Gaudi accelerators and oneAPI platform. Intel's chips are particularly well-suited for mixed AI workloads.

Cerebras: Disrupts the market with its wafer-scale design, offering unparalleled performance for large language models. However, its high cost and specialized nature limit its appeal.



SambaNova: Differentiates itself with a dataflow architecture that unifies memory and compute. Its DataScale system is ideal for enterprise AI workloads.

Increasing competition beyond NVIDIA

So far we have spoken about NVIDIA and AMD, but there is increasing risk from Amazon, Google, and Microsoft. As Sid Pardeshi founder of Blitzy an autonomous software development startup, and former NVIDIA engineer shared:

"I am more bullish on Amazon and Google than AMD for two key reasons. First, Amazon's exclusive partnership with Anthropic positions it uniquely, leveraging Trainium for training and Inferentia for inference. Second, companies like Amazon, Google, Microsoft are deeply incentivized to challenge NVIDIA's dominance. They have their own model services like Bedrock, Vertex, and Azure, enabling them to gather customer usage insights and refine their hardware."

To date, Google TPU is probably the second-best accelerator in the world, offering high-bandwidth interconnects and strong performance over long contexts. However, TPUs are exclusive to Google Cloud Platform (GCP), which limits their appeal and usability in the broader market, and offer AMD a big advantage if they manage to improve the overall developer experience. It will also be interesting to see in the coming years if AMD partnership with Anthropic proves to be more fruitful than AMD partnership with Meta.

Final Verdict: Who Will Win the AI Chip Wars?

The battle for AI chip dominance is far from over, but several trends are emerging:

NVIDIA remains the leader due to its mature ecosystem and developer tools. Its H100 GPU is the gold standard for AI training, and its CUDA platform ensures compatibility with most AI frameworks.

AMD and Intel are gaining ground by offering more cost-effective and open alternatives. The MI300X and Gaudi 3 are strong contenders for organizations looking to avoid vendor lock-in.

Startups like Cerebras and SambaNova are pushing the boundaries of what's possible with novel architectures. However, their high cost and specialized

nature limit their appeal to large enterprises and research institutions.

In the short term, NVIDIA is likely to maintain its dominance, thanks to its unmatched software ecosystem and market share. However, AMD and Intel are making significant strides, particularly in cost efficiency and open standards. The long-term winner will depend on several factors:

Developer adoption: The company that can attract the most developers to its platform will likely win the AI chip wars.

Performance scalability: As AI models grow larger, the ability to scale performance efficiently will be critical.

Cost and power efficiency: Organizations are increasingly prioritizing power efficiency and cost effectiveness in their AI deployments.

Ultimately, the AI chip market is evolving rapidly, and the landscape may look very different in a few years. For now, the battle is heating up, and the stakes have never been higher.

References:

<https://www.barrons.com/articles/ai-nvidia-chips-40bc6704>
<https://www.flybridge.com/ideas/the-bow/0myfoswfo75ry3bdfefbhc5lpe7uxy>
<https://tech.eslamgaming.com/the-battle-for-ai-chip-dominance-nvidia-vs-amd-vs-intel-vs-startups/>

Committee Members	
Dr. B. N. Rajole	Editor in Chief
Ms. T. S. Deshmukh	Co-Editor
Ms. D. V. Patil	Staff Coordinator
Mr. Soham Paradkar	Student Coordinator
Mr. Avinash Gholap	Student Coordinator
Ms. Nikita Gaikwad	Student Coordinator

Website: www.kbtcoe.org

Email Id: techronicle.etc@gmail.com