



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: -February 2026

Vol. - 07, Issue – 4

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 Engineering** is celebrating national science day by unveiling technical newsletter "**TeChronicle**" VOL-07,ISSUE-4 on 28th February 2026.*

*28 Feb is observed as National Science Day, which is celebrated in India on each year to mark the discovery of the Raman effect by Indian physicist **Sir C. V. Raman**, for his discovery, Sir C.V. Raman was awarded the Nobel Prize in Physics in 1930.*

GPU vs. AI-Specific Chips

[Name: Shraddha Ravindra Padol (TY-.B.Tech.)]

Abstract

Graphics processing units (GPUs) still power the majority of today's AI workloads, but they are no longer the only option. As AI moves from research prototypes to planet-scale services, hyper scalers and frontier model labs are increasingly

deploying AI-specific chips—custom accelerators

designed for their exact models and traffic patterns. The result is a two-track hardware landscape: general-purpose GPUs on one side, and purpose-built AI accelerators (TPUs, Trainium, MTIA/Iris, Maia, Jalapeño, etc.) on the other.

1. Where GPUs Stand Today

By mid-2026, GPUs account for roughly 70% of installed worldwide AI compute capacity, with custom silicon making up the remaining 30%, led by Google and Amazon. Nvidia alone held about 90% of the AI accelerator market in 2023; custom AI compute has since taken around 20 percentage points of share over three years as hyper scalers ramp their own designs.

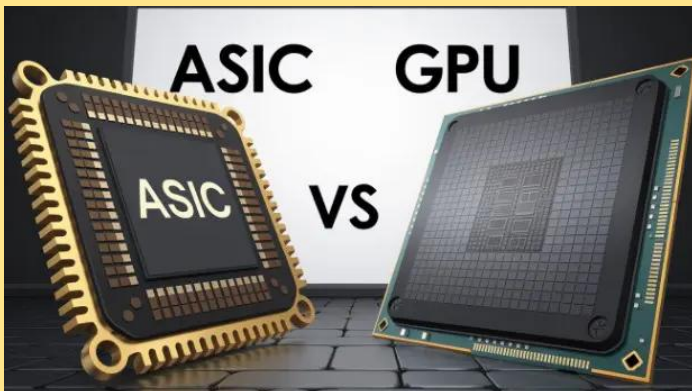
Nvidia's position remains formidable. Its latest datacenter platforms (e.g., the Vera Rubin family and NVL72 rack systems) combine GPUs with high-speed networking, specialized CPUs, and low-latency inference co-processors to push

both throughput per watt (TPS/MW) and tokens per user (TPS/user) higher. The company reports “superstrong” and accelerating demand, with guidance for roughly 70% year-over-year growth in the next fiscal year, constrained more by supply (TSMC wafers, HBM memory) than by demand. GPUs also benefit from a mature software stack (CUDA, libraries, frameworks) and broad compatibility across models and customers, which keeps them as the default for many training jobs and mixed workloads.

AMD has narrowed the gap with its MI300/MI455x-class GPUs and rack-scale Helios systems, growing its installed AI share from near zero to about 6%, though it still trails Nvidia in ecosystem depth and public workload benchmarks.

2. Why AI-Specific Chips Are Gaining Ground

AI-specific chips—often ASICs or domain-specific accelerators—are optimized for known, high-volume workloads rather than maximum flexibility. That specialization brings three advantages that matter at hyperscale:



- **Efficiency:** By removing logic not needed for AI (e.g., graphics pipelines) and tailoring data paths to tensor/math patterns, custom chips can deliver similar or better performance with far lower power per operation. In a world where power and wafers are the scarcest resources, even 10–30% gains in throughput/watt translate into massive cost and capacity advantages.
- **Cost at scale:** Designing separate training and inference chips used to be prohibitively expensive, but at today’s scales it is economical if the optimized silicon saves enough energy and area.

Workload alignment: Hyper scalers and model makers know their bottlenecks intimately (e.g., memory bandwidth for recommendation models, latency for

serving). They can design around those bottlenecks in ways a general-purpose GPU vendor cannot for every customer.

Google has run this playbook for a decade with TPUs, alternating generations optimized for training and inference. In 2026 it is launching TPU 8t (training) and TPU 8i (inference) together, with TPU 8i physically larger due to more HBM to serve inference’s memory demands. Amazon’s Trainium/Inferentia family follows a similar logic inside AWS.

3. Meta’s Iris and the MTIA Roadmap: A Case Study

Meta’s move into custom silicon illustrates the strategic shift. In September 2026, Meta began production of Iris, a datacenter AI accelerator under its MTIA (Meta Training and Inference Accelerator) program. Iris is not a GPU; it is tuned for Meta’s core workloads—ranking, recommendations, ad systems, and growing generative-AI features across Facebook, Instagram, and WhatsApp.

Key points from Meta’s program:

- **Four chips in two years:** Meta is compressing hardware cadence, targeting multiple MTIA generations within a short window to reduce dependence on external GPUs.
- **Measured gains:** The first-generation MTIA v1 (TSMC 7nm, 800 MHz) delivered 102.4 TOPS at INT8 and 51.2 TFLOPS at FP16, explicitly balanced for ranking/recommendation serving rather than raw peak throughput. Meta reports the next generation (v2) achieves roughly 3× performance over v1 on four key models, justifying continued investment.
- **14-gigawatt target:** Meta aims to roughly double its datacenter AI capacity to about 14 GW by 2027. Doing this entirely with merchant GPUs would intensify competition for constrained supply; custom silicon gives Meta a second lever to convert scarce power into usable AI compute more efficiently.
- **Long-term manufacturing tie-up:** Broadcom is extending its partnership to support MTIA chips through 2029, signaling a multi-generation commitment rather than a one-off project.

Iris and later MTIA chips (e.g., MTIA 500 targeted for mass deployment in 2027) are focused on inference. There is no indication yet that Meta is moving its largest training runs off Nvidia GPUs; instead, it is shifting repetitive, high-volume inference onto custom hardware while keeping GPUs for training and flexible workloads.

4. New Entrants and the Widening Field

Meta is not alone. The custom-chip field now includes:

- Google TPUs: Long-established, with major users including Anthropic, OpenAI, Meta, and Apple; Morgan Stanley has projected TPU-related sales could reach tens of billions annually by 2027–2028.
- Amazon Trainium/Inferentia: Mature programs inside AWS for both internal and customer workloads.
- Microsoft Maia: Reports describe a Maia 300 accelerator for Azure/Copilot workloads, with possible unveiling in late 2026 and large TSMC capacity orders for 2027.
- OpenAI Jalapeño: OpenAI's first inference accelerator, presented at Hot Chips 2026, benchmarks competitively against Nvidia's GB300 and is estimated to be roughly in the same class as Nvidia's Vera Rubin systems for certain workloads.
- Others: Cerebras (CS-4/CS-5 roadmaps for ultra-low-latency inference), Samba Nova, D-Matrix, and various startups add further pressure on the GPU incumbents.

Together, these programs form a four-way race (Meta, Google, Microsoft, Amazon) to reduce reliance on merchant GPUs for at least part of their AI stack, even as all remain major Nvidia customers.

5. Technical Trade-offs: Flexibility vs. Efficiency

From an architecture standpoint, the GPU vs. AI-specific debate boils down to flexibility versus efficiency:

- GPUs are general-purpose parallel processors originally designed for graphics, repurposed for AI via APIs like CUDA. They excel when models change rapidly, when workloads are diverse, or when broad software support is critical.

- AI accelerators strip away non-AI logic and optimize memory hierarchies, interconnects, and math units for specific patterns (e.g., matrix multiplies, sparse embeddings). This yields better performance per watt and often lower latency for stable workloads, at the cost of narrower applicability.

IBM's analogy is useful: a GPU is like a pickup truck—good at many jobs—while an AI-specific accelerator is like an 18-wheeler or a sports car, superior for particular tasks but less versatile.

6. Market Dynamics: Share, Supply, and the “Growth-Rate” Battle

Custom AI chips are growing faster than merchant GPUs for the first time. Industry analysis projects ~44.6% growth in custom AI chip shipments in 2026, compared with ~16.1% for merchant GPUs, as hyper scalers scale purpose-built silicon. This does not mean GPUs are losing in absolute terms—Nvidia still guides for strong growth—but it does mean custom silicon is capturing a larger slice of incremental AI compute.

The battleground is increasingly about fab and memory allocation, not just list prices. TSMC wafer capacity and HBM supply are finite; every wafer devoted to MTIA, TPU, Trainium, or Maia is one not available for Nvidia/AMD GPUs or smaller AI chip firms. Over time, successful custom programs can slow the growth rate of GPU orders from hyper scalers even as their total AI compute rises, because custom chips absorb workloads that would otherwise require more GPUs.

7. What This Means for the “AI Chip War”

For your broader narrative on the AI chip war, the GPU vs. AI-specific subtopic can be framed around three tensions:

1. Dominance vs. Displacement: GPUs remain dominant in installed base and training, but custom accelerators are carving out large, defensible niches in inference and repetitive workloads.^{ibm+1}
2. Merchant vs. In-House: Nvidia and AMD sell to everyone; hyper scalers design for themselves. As in-house programs mature (TPU, Trainium, MTIA/Iris, Maia, Jalapeño), the market shifts from a pure merchant model to a hybrid where

the biggest customers are also competitors in silicon.

References:

- 1.<https://semiengineering.com/can-gpus-continue-to-dominate-ai-compute>
- 2.<https://www.ibm.com/think/topics/ai-accelerator-vs-gpu>
- 3.<https://tech-insider.org/meta-iris-ai-chip-mtia-production-2026>
- 4.<https://contabo.com/blog/gpu-vs-ai-accelerator-differences>

NVIDIA Blackwell Ultra: Powering the Next Generation of Artificial Intelligence

[Name: Ms. T. S. Deshmukh]

Artificial Intelligence is rapidly moving from simple machine-learning applications toward Generative AI, AI agents and reasoning models. To support these increasingly demanding workloads, NVIDIA has developed its Blackwell Ultra architecture, with the GB300 NVL72 platform designed specifically for large-scale AI training and inference.

The GB300 NVL72 combines 72 NVIDIA Blackwell Ultra GPUs and 36 NVIDIA Grace CPUs in a liquid-cooled rack-scale system. Each Blackwell Ultra GPU provides 288 GB of HBM3e memory, while the complete platform provides up to 20 TB of GPU memory and up to 576 TB/s of aggregate GPU memory bandwidth.

A key technology is the Tensor Core, a specialized processing unit designed for matrix and tensor calculations used extensively in neural networks. Blackwell Ultra also provides enhanced acceleration for Transformer models, which form the foundation of many modern Large Language Models (LLMs). NVIDIA states that Blackwell Ultra provides 2× higher attention performance and 1.5× more AI compute than the previous Blackwell GPU generation. Another important technology is fifth-generation NVLink, which allows multiple GPUs to communicate at extremely high speed. In the GB300 NVL72, 72 GPUs are connected with 130 TB/s of NVLink bandwidth, allowing them to operate as a tightly integrated AI computing system.

The platform also supports FP4 low-precision AI computation, which reduces the amount of data that must be stored and moved while enabling high AI throughput. NVIDIA reports up to 1,440 PFLOPS of FP4 Tensor Core performance for the GB300 NVL72 configuration with sparsity.

The Blackwell Ultra GPU incorporates several

specialized hardware technologies to improve AI processing efficiency. Its Tensor Cores accelerate matrix multiplication, while the enhanced Transformer Engine dynamically supports different numerical precisions for Transformer workloads. The GPU uses HBM3e for high-speed access to model weights and activations and incorporates high-speed NVLink connectivity for GPU-to-GPU communication. It also includes hardware features for decompression, reliability, security and data movement. Because these processors operate at very high computational density, the GB300 NVL72 uses liquid cooling to efficiently remove heat from the GPUs and maintain stable operating temperatures. Together, these components allow the Blackwell Ultra architecture to handle the extremely large computational and memory requirements of modern AI models.

References:

1. NVIDIA, “NVIDIA GB300 NVL72: Blackwell Ultra AI Platform,” NVIDIA Data Center. [NVIDIA GB300 NVL72](#)
2. NVIDIA, “Blackwell Architecture,” NVIDIA. [NVIDIA Blackwell Architecture](#)

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. Shraddha Padol	Student Coordinator

Website: www.kbtcoe.org

Email Id: techronicle.etc@gmail.com