AI Processor Competitive Landscape

A New Generation of Chips Challenging Nvidia's Dominance

Author: Gyre Research (team@gyreresearch.com)
Date: 2026-06-11 (v2 — corrected & updated; supersedes 14 May 2026 draft)
Classification: Institutional Circulation — Equity Research
Coverage: Apple · AMD · Amazon · Broadcom · Google · Meta · Microsoft · ARM · OpenAI · Nvidia

This report provides a comprehensive competitive analysis of next-generation AI processors competing with Nvidia across datacenter training, inference, and edge workloads. Chips covered include Apple M5, AMD MI350/MI400, Google TPU (Trillium / Ironwood / 8t·8i), Amazon Trainium3, Meta MTIA, Microsoft Maia 200, Broadcom's XPU platform, ARM, and OpenAI's Broadcom-built accelerator. Each is evaluated on processing power, memory architecture, software maturity, time to market, and competitive positioning versus Nvidia.

Analyst Revision Note — V2 (9 June 2026)

This version corrects factual errors in the 14 May draft and incorporates developments through 9 June 2026. Substantive changes:

- Broadcom Q2 FY2026 (reported 3 Jun 2026) supersedes prior Q1 figures: AI-semi revenue $10.8B (+143% YoY), >$30B AI bookings in-quarter, FY2026 AI guidance raised to ~$56B (prior draft said ">$30B").
- Apple removed as a "confirmed" Broadcom XPU customer — Broadcom confirms Google, Meta, OpenAI, Anthropic (+2 unnamed); Apple's Baltra uses Broadcom networking/IP, not a confirmed XPU.
- Google "TPU v7p" corrected to TPU7x (Ironwood); per-chip memory corrected to 192 GB (the 32 GB figure was Trillium's).
- TPU v8 (8t "Sunfish" / 8i "Zebrafish") announced 22 Apr 2026 — no longer roadmap-only.
- Apple M5 Ultra unified-memory ceiling corrected to ~512 GB (was ~192 GB); M5 Ultra did not appear at WWDC (9 Jun) and is reported slipped to ~Oct 2026.

- Microsoft Maia 200 is a Microsoft in-house design (codename Braga), not “designed by Marvell”; Marvell co-designs the next-gen Maia 300 / Braga-R.
- Process-node / cost corrections: AMD MI350 = N3P (3nm), not 4nm; B300 = 4NP, Vera Rubin = N3 / HBM4; Trainium3 ≈ half (not one-third) the manufacturing cost of a B200.
- OpenAI: “Titan” is a press/analyst codename, not official; the “$10B” figure is soft (deal is “multiple billions,” ~$18B financing snag reported May 2026); the $100B Nvidia figure is an investment into OpenAI (LOI), not a GPU purchase.
- Added: Arm’s first in-house CPU (24 Mar 2026, Meta lead customer); AMD–Meta ~6GW deal (24 Feb 2026) and CES 2026 MI455X lineup.
- Recency sweep (late May–9 Jun 2026): added Nvidia Q1 FY2027 results (20 May) and its Computex Arm-PC SoC; corrected the OpenAI–Nvidia “$100B” (never finalized — a ~$30B equity stake closed instead); corrected Anthropic’s Project Rainier to ~500,000 Trainium2; confirmed WWDC (8 Jun) was software-only (no M5 Ultra/Baltra); softened the UALink/Ultra Ethernet “eroding NVLink at rack scale” framing.

Executive Summary

Market Position

Nvidia still controls an estimated ~80% of the AI accelerator market (estimates range ~75–90% depending on methodology — revenue vs. deployed compute, datacenter-only vs. all accelerators), but the competitive picture in 2026 is materially different from 2024. A bifurcation is underway: merchant GPU vendors (Nvidia, AMD) compete on breadth and software ecosystem, while hyperscalers pursue custom ASIC strategies to reduce per-token cost and single-vendor dependency.

Scale of the Shift

Broadcom’s AI-semiconductor revenue reached $10.8B in Q2 FY2026 (+143% YoY), reported 3 June 2026, with more than $30B of new AI bookings in the quarter alone — roughly 3x the amount shipped (Q1 FY2026 was $8.4B, +106% YoY). This underscores how much custom-silicon spend is now flowing outside of Nvidia. The era of GPU-only AI infrastructure is ending.

Key Structural Themes

Three structural shifts define 2026: (1) memory bandwidth has replaced raw FLOPS as the critical bottleneck for LLM inference, driving AMD’s HBM4 strategy and Amazon’s UltraServer architecture; (2) open interconnect standards are emerging to challenge Nvidia’s NVLink moat —

Ultra Ethernet (UEC 1.0, Jun 2025) is shipping for scale-out today, while UALink (spec 1.0 Apr 2025, 2.0 Apr 2026) targets rack-scale scale-up but lacks volume switch silicon until 2027 (AMD's first Helios systems ship "UALink-over-Ethernet" in the interim); (3) Broadcom has captured an estimated 70%+ of the custom-silicon design-services market, making it the enabler of the hyperscaler ASIC wave (some analysts see this normalizing toward ~60% by 2027 as competition scales).

Bottom Line

Nvidia's CUDA ecosystem remains the dominant moat — it is a software moat, not a silicon moat. No competitor has closed this gap in 2026. AMD is the most credible merchant challenger.

Broadcom is the most asymmetric investment thesis. Apple's M5 redefines edge inference but is structurally outside the datacenter market. The real threat to Nvidia is not a single competitor — it is the cumulative effect of ten well-funded alternatives operating in parallel.

Chip-by-Chip Analysis

1. Nvidia — The Incumbent Standard

Chips: Blackwell Ultra (B300) · Vera Rubin (roadmap H2 2026)

Strengths

- CUDA ecosystem is the single most important moat in semiconductors. A decade of developer lock-in; most AI research and production code is written for CUDA first.
- Blackwell delivers up to 4x training and up to 30x inference performance versus the prior Hopper (H100) generation, at up to 25x better energy efficiency on large-model inference.
- At GTC (16 Mar 2026) Nvidia cited roughly $1 trillion of cumulative Blackwell + Vera Rubin revenue visibility across 2026–2027 — extraordinary forward visibility (up from the ~$500B figure a year earlier).
- Vera Rubin NVL144 system rated for up to 3.6 NVFP4 exaflops of inference (and ~1.2 FP8 exaflops training) — a ~3.3x gain over GB300 NVL72 and a target competitors are still building toward on paper.
- Most recent results (Q1 FY2027, reported 20 May 2026) underscore the scale: revenue ~$81.6B (+85% YoY) with data-center revenue ~$75B (+92%); management calls Blackwell demand "off the charts." At Computex (Jun 2026) Nvidia also unveiled an Arm-based PC SoC (N1/N1X), extending beyond the datacenter into client AI silicon.

Weaknesses

- Pricing power creates a ceiling. Hyperscalers are economically motivated to fund alternatives.
- Export controls on China continue to truncate the addressable market.
- Proprietary NVLink ecosystem creates vendor lock-in resentment at scale buyers, accelerating open-standard alternatives (UALink, Ultra Ethernet).

Time to Market / Maturity: Blackwell Ultra (B300) shipping and ramping since ~January 2026. Vera Rubin slated for 2H 2026 — engineering samples already at Microsoft, Dell, and CoreWeave; mass NVL72 shipments guided to fall 2026 (full-scale racks into 2027). Nvidia has halted China-bound H200 production, reallocating that TSMC capacity to Vera Rubin. Most mature stack in market by a wide margin.

---

2. AMD — The Credible Merchant Challenger

Chips: Instinct MI350/MI355X (shipping) · MI400/MI450 (2H 2026)

Strengths

- MI400 (MI455X) features 432 GB of HBM4 per GPU — ~50% more than the MI350's 288 GB HBM3e — with 19.6 TB/s bandwidth, roughly 2.4x the MI350's ~8 TB/s. A structural edge for LLM inference, where memory bandwidth is the primary bottleneck.
- Helios rack-scale platform built on the Open Rack Wide form factor Meta contributed to OCP in 2025. Oracle has committed to 50,000 MI450 GPUs (initial deployment Q3 2026); OpenAI is a named early MI400-series design partner.
- ROCm 7 claims up to 3.5x faster inference (and ~3x faster training) than ROCm 6, with 20–30% better performance than Nvidia on DeepSeek and Llama workloads and up to 40% more tokens per dollar.
- OpenAI is taking a warrant for up to 160 million AMD shares (~10% stake) to secure 6 GW of GPU supply (announced 6 Oct 2025) — the most significant customer-alignment signal AMD has received. AMD and Meta announced a further ~6 GW partnership (24 Feb 2026), and at CES 2026 AMD unveiled the flagship MI455X (~320B transistors, 40 PFLOPS FP4).

Weaknesses

- Software ecosystem remains the primary liability. CUDA entrenchment means most ML engineers reach for Nvidia first.

- Cannot yet point to large-scale clusters matching Nvidia's deployment proof points (the major MI450 build-outs are still ahead of the company).
- AMD positions itself as a "critical high-performance alternative" — accurate, but strategically defensive framing.

Time to Market / Maturity: MI350/MI355X in production now (TSMC N3P). MI400/MI450 (TSMC N2) on track for 2H 2026; "Advancing AI" event scheduled July 2026. Maturing rapidly; credible at scale for inference workloads.

---

3. Apple Silicon M5 — Edge AI Dominance, Server Wildcard

Chips: M5 · M5 Pro · M5 Max (shipping) · M5 Ultra (expected ~Oct 2026)

Strengths

- Next-generation GPU architecture with a dedicated Neural Accelerator in each GPU core, delivering over 4x peak GPU compute for AI workloads vs. M4 and over 6x vs. M1.
- M5 offers unified memory bandwidth of 153 GB/s (~28% over M4); M5 Pro scales to 307 GB/s and M5 Max to 614 GB/s. Unified memory eliminates the CPU–GPU transfer bottleneck inherent in discrete-GPU setups.
- M5 Max delivers up to 8x faster AI image generation vs. M1 Max (M5 Pro: up to 7.8x vs. M1 Pro) — among the largest generational deltas Apple has cited for on-device generative work.
- M5 Ultra is projected by third-party analysts at ~600–800 TOPS (not an Apple-published figure); an effective local LLM inference platform for quantized models via llama.cpp and MLX.
- Apple and Broadcom are jointly developing "Baltra," a server-class AI inference chip on TSMC N3E, targeting ~2027 deployment — Apple's first move toward datacenter silicon.

Weaknesses

- macOS/Apple ecosystem lock-in. No CUDA, no ROCm. Not viable for enterprise PyTorch training clusters.
- Memory: M5 Ultra is expected to support up to ~512 GB unified memory (fusing two M5 Max dies) — large for a workstation, but still below HBM-equipped datacenter accelerators.
- Not sold as merchant silicon. Captive to Apple hardware. Baltra is internal-only.

Time to Market / Maturity: M5 / M5 Pro / M5 Max shipping now (Pro/Max announced 3 Mar 2026, TSMC N3P). M5 Ultra was not announced at WWDC (8 Jun 2026 — a software-only

team@gyreresearch.com ▪ gyreresearch.com ▪ (800) 933-4731

keynote; no M5 Ultra, Mac Studio, or “Baltra” shown) and is reported slipped to ~October 2026 on DRAM supply constraints. Mature for edge/workstation; nascent for datacenter.

4. Google TPU — The Most Mature Custom-Silicon Program

Chips: TPU v6e “Trillium” (GA) · TPU7x “Ironwood” (GA) · TPU 8t “Sunfish” / 8i “Zebrafish” (announced Apr 2026)

Strengths

- Trillium (TPU v6e) delivers 4.7x peak compute over TPU v5e, with 32 GB HBM per chip, GA since late 2024 with 100,000+ chip single-fabric deployments.
- Ironwood — officially TPU7x (the 7th generation) — delivers ~7.37 TB/s HBM bandwidth and 192 GB HBM per chip, scales pods to 9,216 chips for 42.5 FP8 exaflops per pod (~2x perf/watt vs. Trillium), and reached GA ~6 Nov 2025 — one of the most powerful AI systems built to date.
- Anthropic signed a deal (23 Oct 2025) to use up to 1 million TPUs, worth tens of billions, bringing 1 GW+ of capacity online in 2026. Meta also signed a multibillion-dollar deal to rent Google’s TPUs (with potential on-prem purchases from 2027).
- At Google Cloud Next (22 Apr 2026), Google announced its eighth generation as two chips — TPU 8t “Sunfish” (training) and TPU 8i “Zebrafish” (inference) — preview in 2H 2026, broad GA targeted ~late 2027. Software maturity (JAX/XLA) is significantly ahead of any competitor outside CUDA.

Weaknesses

- Captive infrastructure — TPUs are rented via Google Cloud, not sold as merchant silicon in the traditional sense.
- DA Davidson analyst Gil Luria estimated Alphabet could capture ~20% of the AI-chip market (a ~$900B opportunity) if it sold TPUs to third parties — a move Google has not yet made at scale.
- Customer concentration in Google’s own cloud limits market influence relative to the silicon’s specs.

Time to Market / Maturity: Trillium and Ironwood in production/GA. TPU 8t/8i preview 2H 2026. Most mature custom-silicon program; significant scale advantage.

team@gyreresearch.com ▪ gyreresearch.com ▪ (800) 933-4731

5. Amazon Trainium3 — Cost-Competitive With Real Customers

Chips: Trainium3 (launched re:Invent Dec 2025) · Trainium4 (timing not yet AWS-stated)

Strengths

- Trainium3 delivers 2.52 PFLOPS FP8 per chip with 144 GB HBM3e and 4.9 TB/s bandwidth (TSMC 3nm). A full Trn3 UltraServer (144 chips) delivers 362 PFLOPS with up to 4.4x higher performance and 4x better energy efficiency vs. the Trainium2 UltraServer.
- AWS claims customers achieve up to 50% lower training and inference costs vs. GPU alternatives (Decart cites ~4x faster inference at half the GPU cost). Amazon Bedrock serves production workloads on Trainium3.
- Independent estimates put a Trainium3 chip at roughly half the manufacturing cost of an Nvidia B200 (~$3,200 vs. ~$6,400 per Epoch AI) — a structural TCO advantage for captive workloads.
- Trainium4 will support Nvidia's NVLink Fusion (NVLink 6 / MGX, up to 72 chips at 3.6 TB/s/chip), enabling heterogeneous Trainium+GPU clusters — announced at re:Invent (Dec 2025).

Weaknesses

- Neuron SDK adoption is narrow. Trainium3 has Day-0 support only for LNC=1 or LNC=2; LNC=8 — preferred by much of the ML research community — is not planned until mid-2026.
- Anchor customer Anthropic runs production today largely on Trainium2 (Project Rainier, ~500,000 Trainium2 chips activated, scaling toward 1M+) and is moving toward Trainium3 — so Trn3 production at Anthropic is forward-looking rather than fully in place.
- Effectively captive to AWS; ecosystem depth far behind CUDA and even Google's JAX/XLA. (AWS has not stated a Trainium4 ship date; "late 2026/early 2027" is an external estimate.)

Time to Market / Maturity: Trainium3 launched at re:Invent (Dec 2025), ramping in 2026. Trainium4 timing not yet disclosed by AWS. Strong for AWS-committed workloads; limited external appeal.

6. Meta MTIA — Aggressive Roadmap, Internally Focused

Chips: MTIA 300 · 400 · 450 · 500 (all disclosed March 2026)

Strengths

- Meta disclosed four MTIA generations in March 2026, built on the open-source RISC-V ISA and co-developed with Broadcom on TSMC. The lineup reports a 4.5x increase in HBM bandwidth and a 25x increase in compute FLOPs from MTIA 300 to MTIA 500.
- Meta's pace — roughly a new chip generation every six months ("four chips in two years") — is faster than any competitor in the market.
- RISC-V adoption eliminates ARM licensing costs and architectural dependencies — a strategic long-term bet on an open ISA.
- Handles Meta's highest-volume workloads: ad ranking, recommendation, and image/video generation at billions-of-users scale.

Weaknesses

- Entirely internal deployment. Not sold commercially; no third-party ecosystem.
- RISC-V for AI at this scale is relatively unproven; mature RISC-V inference toolchains lag ARM and x86.
- Betting on an open ISA is directionally sound, but near-term software support is thinner than established competitors.

Time to Market / Maturity: Multiple generations disclosed simultaneously; production status varies by variant (MTIA 450 mass production ~early 2027, MTIA 500 ~6 months later). Strategically significant; not a merchant competitor.

---

7. Microsoft Maia 200 — Inference-First, Quietly Formidable

Chip: Maia 200 (codename "Braga," in production) — Microsoft in-house design

Strengths

- Microsoft claims Maia 200 delivers 3x the FP4 performance of Amazon's Trainium3 and FP8 performance above Google's seventh-generation TPU, with 30% better performance per dollar vs. prior-generation hardware in its fleet (vendor figures, not independently benchmarked).

team@gyreresearch.com ▪ gyreresearch.com ▪ (800) 933-4731

- Specs: native FP8 and FP4 tensor cores, 216 GB HBM3e at 7 TB/s, 272 MB on-chip SRAM, 750W SoC envelope, ~140B transistors, TSMC 3nm. Inference-optimized design.
- Powers GPT-5.2 models from OpenAI, Microsoft Foundry, and Microsoft 365 Copilot — real production-scale deployment (US Central / Iowa live, US West 3 / Phoenix next), not a benchmark exercise.

Weaknesses

- Not externally available. Azure customers cannot directly provision Maia 200; it operates behind Azure AI services.
- Inference-specific orientation limits training-workload applicability.
- Software stack is proprietary and narrow; limited developer community.

Time to Market / Maturity: Announced 26 Jan 2026; in production at Microsoft datacenters. Maia 200 is a Microsoft in-house design — Marvell co-designs the next-generation Maia 300 / "Braga-R," not Maia 200. Real deployment, but opaque to the market.

---

8. Broadcom — The Infrastructure Backbone of Custom Silicon

Role: Custom ASIC design partner (XPU) · Not a branded chip vendor

Strengths

- Q2 FY2026 AI-semiconductor revenue: $10.8B, up 143% YoY (reported 3 Jun 2026); >$30B of AI bookings in the quarter; total revenue $22.19B (+48%). CEO Hock Tan reiterated a $100B+ AI-chip revenue target for FY2027, with the quarter putting Broadcom on track for roughly $56B of AI revenue in FY2026 (Q3 guided to ~$16B, +~200% YoY).
- Confirmed XPU customers include Google, Meta, OpenAI, and Anthropic (Broadcom cites six core custom-chip customers; two remain unnamed). Broadcom commands an estimated 70%+ of the custom AI-accelerator design-services market.
- The 3.5D XDSiP face-to-face (F2F) packaging platform integrates compute dies, I/O tiles, and HBM in a single package (production since ~Feb 2026), lowering time-to-market for partners using Broadcom IP.
- Tomahawk 6 — the first 102.4 Tbps Ethernet switch (shipping ~Mar 2026) — provides the fabric connecting custom chips at 1M+ XPU scale. This vertical slice (custom silicon plus networking) is the structural advantage.

team@gyreresearch.com ▪ gyreresearch.com ▪ (800) 933-4731

Weaknesses

- Customer concentration: HSBC estimates Google TPUs at ~58% of Broadcom's ASIC unit shipments and ~78% of ASIC revenue (TPUs price ~$13K/chip vs. ~$5K for other programs).
- Dependent on TSMC advanced-node capacity, which is constrained through at least 2027.
- Revenue is backend-weighted to customer deployment timelines, creating lumpiness (the stock fell ~14% post-Q2 — despite the +143% AI beat, the Q3 guide (~$16B) and softer software-segment results fell short of the most bullish expectations).
- If hyperscalers develop full in-house design capability, Broadcom's design-partner moat erodes — Marvell has explicitly flagged this dynamic.

Time to Market / Maturity: Operationally mature; AI revenue accelerating at 140%+ YoY. The most asymmetric pure-play on hyperscaler custom-silicon capex without binary chip-architecture risk; the $69B VMware acquisition (closed Nov 2023) adds high-margin software revenue (~$7.2B in Q2 FY2026).

---

9. ARM — The Architecture Layer Beneath Everything

Role: ISA licensor and chip-design platform · Now also a first-party chip vendor

Strengths

- ARM's Neoverse architecture underpins AWS Graviton, Ampere Altra, Google Axion, and Microsoft Cobalt — virtually every hyperscaler CPU. Every non-x86, non-RISC-V custom AI chip runs on ARM IP (>1B Neoverse cores deployed).
- Edge AI processors including Qualcomm Snapdragon and Apple Silicon are ARM-based — dominant in mobile and on-device inference; Nvidia's new Arm-based PC SoC (N1/N1X, Computex Jun 2026) extends the ISA further into AI PCs.
- The royalty model captures value from the entire custom-silicon wave regardless of which architecture wins; ARM CSS (Compute Subsystem, via "Arm Total Design") is increasingly licensed to hyperscalers, deepening ARM's role in custom builds.
- On 24 Mar 2026, ARM launched its first in-house CPU with Meta as lead customer (production 2H 2026; launch partners include OpenAI, Cloudflare, Cerebras). A first-party AI ASIC is expected ~early 2027 — a strategic shift from pure IP licensor toward chip vendor.

Weaknesses

- As an IP licensor, revenue leverage is indirect and capped by royalty rates (its new first-party chip ambitions also risk channel conflict with licensees).
- Meta's RISC-V bet — and growing RISC-V interest at other hyperscalers — is a long-term structural threat to ARM's licensing position in custom inference silicon.
- ARM has no hardware manufacturing capability; entirely dependent on ecosystem and foundry execution.

Time to Market / Maturity: Architecture is ubiquitous and mature across all segments. Defensive position with royalty leverage on the whole market; the move into first-party silicon (2026–2027) is the new variable to watch.

---

10. OpenAI Custom Accelerator — Emerging Wild Card

Chip: OpenAI-designed XPU (press codename "Titan") · Broadcom-built · TSMC N3 · production from 2H 2026

Strengths

- OpenAI's first custom AI accelerator, developed with Broadcom (collaboration announced 13 Oct 2025): OpenAI designs, Broadcom builds, TSMC manufactures on 3nm, targeting 10 GW of inference-focused deployment.
- Mizuho estimates the full 10 GW OpenAI–Broadcom program could represent $150–200B over multiple years — among the largest custom-silicon commitments in history if realized.
- Vertical integration at OpenAI's inference scale would generate structural cost advantages over renting Nvidia capacity at current margins.

Weaknesses

- "Titan" is a press/analyst codename, not an official product name; the widely-cited "$10B" is soft (Broadcom characterized it as "multiple billions," and a ~$18B financing snag for phase one (~1.3 GW) was reported in May 2026 — reportedly because Broadcom sought a Microsoft commitment to buy ~40% of output before funding). No performance specifications have been disclosed.
- Official guidance is production "starting 2H 2026," but volume deployment now looks more likely in 2027 (the earlier Q2→Q3 2026 slip has extended further).
- The parallel Nvidia arrangement began as a Sept 2025 letter of intent for Nvidia to invest up to $100B into OpenAI (an investment, not a GPU purchase) — but the $100B was never

finalized; Jensen Huang called it “probably not in the cards” (Mar 2026), and what actually closed was a ~$30B Nvidia equity stake (Feb 2026, ~$730B valuation) alongside Vera Rubin commitments. OpenAI’s custom chip is supplementary, not a Nvidia replacement.

Time to Market / Maturity: Pre-production; first deployment targeted 2H 2026, volume more likely 2027. High optionality, high execution risk — first-generation custom chips from software-first organizations have a poor historical track record.

Competitive Matrix

CHIP / PLATFORM	USE CASE	MEMORY	PROCESS	STATUS	SW MATURITY	NVIDIA THREAT
Nvidia B300 (Blackwell Ultra)	Training + Inference	288 GB HBM3e	TSMC 4NP	Shipping	★★★★★	Baseline
Nvidia Vera Rubin (NVL144)	Training + Inference	288 GB HBM4	TSMC N3	H2 2026	★★★★★	Baseline
AMD MI350 / MI400	Training + Inference	288 GB HBM3e / 432 GB HBM4	N3P / N2	Shipping / 2H 2026	★★★□□	High
Apple M5 Ultra	Edge Inference / LLM	up to ~512 GB unified	TSMC N3P	~Oct 2026 (exp.)	★★★□□	Low (diff. market)
Google Ironwood (TPU7x)	Training + Inference	192 GB/chip (pod-scale)	adv. node (undisc.)	GA	★★★★□	High (internal)
Amazon Trainium3	Training + Inference	144 GB HBM3e	TSMC 3nm	Launched (ramping)	★★★□□	Medium
Meta MTIA 300–500	Inference (ranking / gen-AI)	Undisclosed	TSMC	Multi-gen disclosed	★★□□□	Low (internal)
Microsoft Maia 200	Inference	216 GB HBM3e	TSMC 3nm	In production	★★□□□	Medium
Broadcom XPU (TPU/MTIA/OpenAI)	Custom (customer-defined)	Customer-defined	3nm	Revenue ramping	N/A (foundry)	Very High (enabler)
OpenAI XPU (“Titan”)	Inference	TBD	TSMC N3	Pre-production	N/A	TBD

Note: *B300 and Vera Rubin are listed separately (the 14 May draft merged them into one “288 GB HBM3e / 3nm” row, conflating a 4NP/HBM3e part with an N3/HBM4 part). Ironwood memory corrected to 192 GB/chip; its process node is undisclosed by Google.*

Key Analyst Conclusions

1. **The moat is CUDA, not silicon.**

   AMD's MI400 will likely match or exceed Nvidia on raw memory specs. It will not match CUDA. Until ROCm reaches parity in developer mindshare — which has not happened and is not imminent — AMD is a cost-arbitrage option for inference, not a platform replacement for training.

2. **Broadcom is the most differentiated investment thesis in the space.**

   It captures value regardless of which hyperscaler architecture wins. Broadcom is on track for ~$56B of AI-semi revenue in FY2026 and has reiterated a $100B+ FY2027 target, with the VMware-driven software segment adding high-margin, insulating revenue. It is the arms dealer in a war where everyone is buying weapons — though customer concentration (Google ~78% of ASIC revenue) and software-segment softness are the watch-items.

3. **Custom silicon is structurally deflationary for Nvidia's cloud revenue.**

   Every token served on Trainium3, a TPU, or Maia is a token not served on an H100 or B200. Google's TPU program is the most mature and available at a scale few others can match — already diverting significant enterprise spend that would otherwise flow to Nvidia.

4. **Apple's M5 is a legitimate local inference platform, not a datacenter play.**

   A distributed Neural-Accelerator architecture and (for the forthcoming M5 Ultra) up to ~512 GB of unified memory make it a strong option for on-premises, privacy-sensitive, mid-scale inference. "Baltra" (Apple + Broadcom server chip, ~2027) is the one to watch if Apple decides to commercialize server silicon.

5. **Meta's 6-month chip cadence is the most aggressive in the industry.**

   It structurally lowers cost-per-token for Meta's ad business and compounds a competitive moat in recommendation and generative-AI workloads. Not a merchant chip; an internal competitive advantage for Meta's core revenue engine.

6. **The real risk to Nvidia is aggregate, not singular.**

   The AI-chip market has moved from one-horse to a fragmented, multi-architecture landscape across training, inference, and edge. Nvidia remains dominant. The question for equity analysts is the rate of share erosion and what it implies for margin trajectories through 2027 and beyond.

team@gyreresearch.com ▪ gyreresearch.com ▪ (800) 933-4731

Important Disclosures

This report is produced for informational purposes only and does not constitute investment advice, an offer, or a solicitation to buy or sell any security. All data sourced from company disclosures, earnings calls, and public reporting current as of 9 June 2026 (sources include company press releases and SEC filings, Tom's Hardware, The Next Platform, CNBC, Reuters/Bloomberg, and analyst notes from HSBC, Mizuho, and DA Davidson). Vendor-published performance figures (e.g., Microsoft, AMD, AWS marketing claims) are not independently benchmarked and are identified as such. Forward-looking statements involve risks and uncertainties; actual results may differ materially. Intended for institutional circulation only.

team@gyreresearch.com ▪ gyreresearch.com ▪ (800) 933-4731