State of AI Silicon 2026: The Definitive Technical Atlas

A Comprehensive Analysis of GPU Architectures, Engineering Constraints, and Market Dynamics

Anjan Goswami

General Manager, SmartInfer.com

December 2025

Abstract: The AI silicon landscape has entered a phase of architectural divergence, supply constraint, and value migration—training monopolies are giving way to fragmented inference ecosystems, and the memory wall has become the defining engineering constraint. NVIDIA maintains 86–94% data center GPU market share [1] but faces erosion as inference overtakes training (projected 55% of AI compute spending by 2026 [2]), hyperscalers deploy custom ASICs, and AMD achieves TCO parity for memory-bound workloads. The most significant developments: NVIDIA Blackwell's dual-die architecture doubles effective memory bandwidth to 8 TB/s, AMD MI325X delivers 256 GB HBM3e (the largest GPU memory pool available), and wafer-scale computing (Cerebras WSE-3 with 21 PB/s on-chip bandwidth [3, 4]) challenges fundamental assumptions about GPU architectures. Chinese domestic stacks reach ∼60% of H100 performance per chip but compensate through system-scale deployment [5], while India's Tata-PSMC fab targets commercial production in late 2026 at 28nm mature nodes [6, 7].

1. The Global Silicon Atlas

1.1 NVIDIA Blackwell and Hopper Architectures Dominate but Face New Constraints

NVIDIA's Blackwell architecture (B200/B100/GB200) represents the most significant architectural shift since Volta, moving from monolithic dies to dual reticle-limited dies connected via a 10 TB/s NV-HBI interconnect [8, 9]. The B200 packs 208 billion transistors across two dies on TSMC 4NP, presenting as a single unified GPU to CUDA applications—a $10 billion R&D investment in die-to-die coherency [10, 11, 12].

Table 1: NVIDIA GPU Architecture Comparison

Specification	H100 SXM	H200	B200	GB200 NVL72
Process	TSMC 4N	TSMC 4N	TSMC 4NP	TSMC 4NP
Transistors	80B	80B	208B (dual-die)	72 × 208B
HBM Capacity	80 GB HBM3	141 GB HBM3e	192 GB HBM3e	13.5 TB unified
Memory Bandwidth	3.35 TB/s	4.8 TB/s	8 TB/s	576 TB/s aggregate
FP8 TFLOPS	1,979	1,979	4,500	720 PFLOPS
FP4 TFLOPS	N/A	N/A	9,000	1.44 EFLOPS
TDP	700W	700W	1,000W	120 kW/rack
Pricing	$25–40K	$30–45K	$40–50K	∼$3M/rack

The 2nd-generation Transformer Engine enables NVFP4 (4-bit) precision with micro-tensor scaling—grouping 16 elements with separate FP8 scale factors yields <1% accuracy degradation versus FP8 on DeepSeek-R1 [13, 14, 15]. This delivers 3× throughput versus FP16 while reducing VRAM consumption by 60% [16]. The GB200 NVL72 configuration—72 Blackwell GPUs and 36 Grace CPUs in a single liquid-cooled rack—achieves 869,200 tokens/second on Llama2 70B, representing a 30× improvement over H100 for trillion-parameter inference [17, 18, 19].

1.2 AMD MI300X/MI325X Achieves TCO Parity Through Memory Density

AMD's chiplet-first strategy pays dividends where memory bandwidth matters most. The MI300X uses a 3.5D packaging approach—8 GPU XCDs (5nm) stacked on 4 IODs (6nm) via TSMC SoIC, mounted on the largest CoWoS interposer ever produced [20, 21].

Table 2: AMD Instinct Roadmap

Specification	MI300X	MI325X	MI355X (H2 2025)	MI400 (2026)
Architecture	CDNA 3	CDNA 3	CDNA 4 (3nm)	CDNA 5
Compute Units	304	304	256	TBD
HBM Capacity	192 GB HBM3	256 GB HBM3e	288 GB HBM3e	432 GB HBM4
Bandwidth	5.3 TB/s	6.0 TB/s	8 TB/s	19.6 TB/s
FP8 TFLOPS	2,615	2,615	9,200 (8-GPU)	20,000
TDP	750W	1,000W	1,400W	TBD
Price	$10–15K	$12–17K	TBD	TBD

SemiAnalysis TCO analysis reveals nuanced economics: MI300X delivers 33% cost advantage (tokens/dollar) versus H100 for memory-bound LLM inference (Llama 3 405B, DeepSeek v3 670B) [22]. For training, however, NVIDIA maintains performance-per-TCO advantage due to software ecosystem maturity—AMD's out-of-box experience requires "considerable patience" versus NVIDIA's "amazing" integration [23]. The MI300X's 256 MB Infinity Cache and 192 GB HBM3 enable single-GPU inference for 70B models that require 2× H100s, halving server count for memory-constrained deployments.

ROCm has closed the performance gap from 40–50% to 10–30% behind CUDA, with FlashAttention-2 now supported via the Composable Kernel backend [24]. The critical constraint remains ecosystem: 6 million CUDA developers versus AMD's nascent community.

1.3 Intel Gaudi Struggles While AI Startups Pioneer Specialized Architectures

Intel's Gaudi 3 delivers compelling specifications—128 GB HBM2e, 3.7 TB/s bandwidth, 1.8 PFLOPS FP8—at significant price discounts (∼$15,625 versus $30,000+ for H100) [25]. However, market share remains below 1% for discrete AI accelerators. Intel abandoned its $500M revenue forecast for Gaudi in October 2024, signaling strategic retreat. The Falcon Shores architecture (2025) attempts to merge Gaudi and Xe GPU DNA but faces the same ecosystem challenges.

The startup landscape reveals architectural divergence for specialized workloads:

Table 3: AI Chip Startup Landscape

Company	Architecture	Key Innovation	Performance	Valuation	Status
Groq	TSP	230 MB SRAM-only	1,660 TPS Llama 70B	$6.9B→$20B	Acquired
Cerebras	WSE	46,225 mm² single die	3,000 TPS on 120B	$8.1B (IPO)	Production
SambaNova	Dataflow	1.5 TiB off-package	3.7× faster vs H100	∼$1.6B	Production
Tenstorrent	Tensix+RISC-V	Open-source stack	328 TFLOPS FP8	∼$1–3B	Scaling
Graphcore	IPU (BSP)	Bulk Sync. Parallel	350 TFLOPS	∼$600M	Acquired

Groq's LPU architecture merits special attention: by eliminating HBM entirely and using 230 MB on-chip SRAM, Groq achieves 80+ TB/s effective memory bandwidth versus 3.35 TB/s for H100's HBM3 [26, 27]. This enables sub-10ms first-token latency and 10× energy efficiency. The tradeoff: large models require hundreds of chips (576 for Llama2 70B) due to limited on-chip capacity. NVIDIA's December 2025 acquisition validates inference-specialized architectures as strategic.

Cerebras WSE-3 represents the most radical departure from conventional design: the entire 300mm wafer functions as a single 4-trillion-transistor chip with 44 GB on-wafer SRAM delivering 21 PB/s bandwidth—7,000× more than H100 [3, 4]. Yield management uses built-in redundancy with scribe-line stitching across reticle boundaries.

1.4 China's Domestic Ecosystem Achieves System-Level Competitiveness Despite Chip-Level Gaps

Huawei's Ascend 910C employs a dual-chiplet architecture fusing two 910B SoCs via silicon interposers on SMIC's N+2 (7nm-class) process [28, 29]. Individual chip performance reaches ∼800 TFLOPS FP16—roughly 60% of H100 [30]—but the CloudMatrix 384 system (384 × 910C chips in optical mesh) delivers 300 PFLOPS BF16, approximately 2× GB200 NVL72's aggregate compute, albeit at 5× higher power consumption (559 kW versus ∼120 kW) [31, 32].

Table 4: Chinese AI Chip Landscape

Chinese Chip	Process	Memory	Compute	Status
Huawei Ascend 910C	SMIC N+2 (7nm)	128 GB HBM3	∼800 TFLOPS FP16	Production, ∼450K shipped
Huawei Ascend 910B	SMIC N+2	64 GB HBM2E	256 TFLOPS FP16	Deployed at scale
Cambricon Siyuan 590	SMIC N+2	32 GB HBM2	345 TFLOPS FP16	Production
Biren BR100	TSMC 7nm (blocked)	64 GB HBM2E	2,048 TOPS INT8	Halted by sanctions
Moore Threads S4000	Domestic	48 GB GDDR6	200 TOPS INT8	Production

CANN software stack maturity remains the critical weakness. Developers report the ecosystem as "a road full of pitfalls" with "disorganized documentation" [33, 34]. Huawei has committed to open-sourcing CANN, MindSpore, and openPangu by December 31, 2025, attempting to accelerate community development [35, 36]. The `torch_npu` adapter enables PyTorch code migration with minimal changes, and DeepSeek now provides native Ascend support.

December 2024 export controls target HBM directly—new ECCN 3A090.c restricts memory exceeding 2 GB/s/mm² bandwidth density, effectively banning HBM2E and beyond [37, 38]. Chinese companies stockpiled HBM aggressively; TSMC's "die bank" extends Huawei's production runway approximately 9 months from mid-2025 [39]. Domestic HBM2 production via CXMT (ChangXin Memory Technologies) remains years behind cutting-edge density requirements.

1.5 European and Indian Initiatives Focus on Strategic Nodes Rather Than Leading Edge

SiPearl's Rhea processor—80 Arm Neoverse V1 cores on TSMC N6 with HBM2E—will power JUPITER, Europe's first exascale supercomputer, but arrives 2–3 years behind schedule with first samples in early 2026 [40, 41]. The European Processor Initiative combines Arm-based general-purpose compute (Rhea) with RISC-V accelerators (EPAC) for heterogeneous

HPC [42]. The €43 billion EU Chips Act targets 20% global market share by 2030—ambitious given current ∼10% and execution delays [43].

Tata Electronics' Dholera fab represents India's most significant semiconductor initiative: $11 billion investment with PSMC Taiwan providing technology transfer for 28nm–110nm mature nodes [44, 45]. Timeline: sample chips by late 2025, commercial production by late 2026, 50,000 wafers/month at full capacity [6, 7]. Target markets include power management ICs, display drivers, and automotive—not leading-edge AI chips.

IIT Madras's SHAKTI processor project has produced the first entirely India-designed RISC-V chips, including the IRIS chip for ISRO space applications (180nm at SCL Chandigarh) [46]. India's design ecosystem—home to 20% of global semiconductor design engineers—remains stronger than its manufacturing capability.

2. Engineering Challenges and the Walls That Define 2026

2.1 The Memory Wall Dominates All Architectural Decisions

LLM inference is fundamentally memory-bound, not compute-bound. During autoregressive generation, loading entire model weights per token generated results in arithmetic intensity far below GPU operational efficiency. Llama 2 7B achieves ∼62 ops/byte versus A10's 208 ops/byte ratio—meaning tensor cores sit idle waiting for memory [47].

Table 5: HBM Generation Roadmap

Generation	Bandwidth	Capacity	Timeline
HBM2E	3.35 TB/s	80 GB	2022 (H100)
HBM3	5.2 TB/s	141 GB	2023 (H200)
HBM3E	8 TB/s	192 GB	2024–25 (B200)
HBM4	12+ TB/s	256+ GB	2026 (projected)

The "Memory-Parkinson" dynamic ensures models grow to fill available HBM, perpetually hitting the memory wall. Solutions emerge across multiple dimensions:

PagedAttention (vLLM) applies OS-style virtual memory paging to KV cache, achieving <4% memory waste versus 80%+ without—enabling 2–5× more concurrent users per GPU. FlashAttention-3 on Hopper achieves 1.2 PFLOPS FP8 through tiled computation in SRAM, leveraging the Tensor Memory Accelerator (TMA) for 30–40% throughput improvement [47, 48]. Speculative decoding with EAGLE-3 on Blackwell breaks 1,000 tokens/sec/user on Llama 4 Maverick via draft-and-verify parallelism.

In-memory compute represents the most radical solution: Samsung's HBM-PIM integrates compute within HBM stacks, demonstrating 2.5× performance and 62% energy reduction [49, 50]. SK Hynix's AiMX achieves 3× LLM speed improvement versus mobile DRAM at equivalent power. Academic research (Stanford CHIMERA, IBM's 64-core analog chip) reports 10–20× efficiency gains but faces precision and noise challenges for production deployment [51].

2.2 Energy Efficiency Becomes a Data Center Constraint

Individual GPU TDPs have reached 1,000–1,400W (B200, MI355X), pushing rack densities to 120–140 kW—10× traditional deployments [52]. Traditional air cooling maxes out at ∼50 kW per rack; liquid cooling becomes mandatory [53, 54].

Table 6: Data Center Cooling Technologies

Cooling Technology	Max Density	Energy Savings	Complexity	Leaders
Direct-to-Chip (D2C)	100kW+	70–80% heat capture	Moderate	CoolIT, Asetek, JetCool
Single-Phase Immersion	100kW+	PUE <1.03	High	GRC, Submer
Two-Phase Immersion	100kW+	Highest	Very High	LiquidStack, Iceotope

Silicon photonics addresses interconnect energy: electrical signaling consumes 30 pJ/bit while optical approaches achieve <5 pJ/bit—a 6× reduction [55, 56]. Lightmatter's Passage technology and Broadcom's Tomahawk 6 CPO switch (102.4 Tbps) represent production-ready solutions. NVIDIA's Quantum-X Photonics switches (commercial 2026) target 3.5× power reduction for rack-scale fabrics [57].

2.3 Packaging Innovations Bypass Reticle Limits

Maximum die size is constrained to ∼800mm² by lithography reticle limits (26mm × 33mm). Three approaches have emerged:

Chiplet architecture (AMD MI300X): 13 chiplets on 3.5× reticle interposer—smaller dies yield better individually, but inter-chiplet communication adds latency [58, 59]. Infinity Fabric provides 6 TB/s bisection bandwidth but inter-die hops add nanoseconds.

Dual-die coherent (NVIDIA Blackwell): Two reticle-limited dies connected via 10 TB/s NV-HBI, presenting as unified GPU [60, 61]. CUDA programming model preserved at cost of $10B R&D investment in coherency protocols.

Wafer-scale (Cerebras WSE-3): Entire 300mm wafer as single chip with 900,000 cores [62, 63]. Built-in redundancy handles defects; thermal solution allows wafer expansion/contraction. Tradeoff: fixed architecture versus mix-and-match chiplets.

TSMC CoWoS capacity remains structurally constrained: ∼70,000 wafers/month by end 2025, with NVIDIA and AMD reserving 100% through 2025 [64]. Global demand reaches 1 million wafers by 2026—NVIDIA alone requires 595,000 (60% of projected global capacity). Price increases of 10–20% expected in 2025 [65].

2.4 Programmability Remains NVIDIA's Deepest Moat

Despite Triton achieving 76–82% of CUDA performance for LLM inference on H100/A100, and torch.compile delivering 43% average training speedup [66, 67], the CUDA ecosystem—6 million developers, 300+ acceleration libraries, 18 years of optimization—remains unmatched [68].

ROCm has closed the gap from 40–50% to 10–30% behind CUDA in real-world performance [69, 24]. Flash Attention-2 support on Composable Kernel backend, official PyTorch support, and major customer wins (Meta, Microsoft, Oracle) signal maturation. However, SemiAnalysis reports AMD's out-of-box experience remains "riddled with bugs rendering training impossible" without significant engineering investment [70, 71].

Triton's promise—write portable GPU kernels in Python—succeeds for standard operations but fails for specialized workloads requiring thread-level control. MLIR provides infrastructure for cross-platform compilation but hasn't "democratized AI compute" due to corporate fragmentation and NVIDIA's vertical integration advantage.

2.5 Neuromorphic Computing Remains 3–5 Years from Production Relevance

Intel Loihi 2 (1 million neurons, Intel 4 process) and SpiNNaker2 (ARM cores + accelerators) demonstrate research viability for spiking neural networks [72, 73]. BrainChip Akida represents the only commercially deployed neuromorphic processor, achieving 3–5× power reduction for audio recognition versus CPU/GPU—but limited to edge classification tasks.

Event-based vision via Dynamic Vision Sensors (Prophesee Metavision, Sony neuromorphic sensors) gains traction in autonomous vehicles (30%+ of prototypes) and robotics (45+ companies), offering microsecond temporal resolution versus milliseconds for frame cameras. The programming model—fundamentally different from CNNs—limits adoption to specialized applications.

3. Academic Research Frontiers Promising 10–100× Improvements

University research labs are pursuing architectural approaches claiming order-of-magnitude improvements:

Table 7: Academic Research Frontiers

Research Area	Claimed Improvement	Key Source	Production Timeline
MEMS Analog Computer	300× power, 100× speed	GE Aerospace/Stanford 2024	2027–2028
Photonic Tensor Cores	880 TOPS/mm², 5.1 TOPS/W	Frontiers Physics 2024	2026–2027
IBM Analog AI (PCM)	14× energy efficiency	Nature 2023	2026
Silicon Photonics CPO	80% "optics tax" reduction	Industry 2025	2026 (NVIDIA)
Processing-in-Memory (Sangam)	10× energy vs H100	arXiv 2024	2027

MIT-GlobalFoundries partnership (February 2025) focuses on silicon photonics integration combining RF SOI, CMOS, and optical features on single chips [74]. Stanford SystemX produced the CHIMERA chip (0.92 TOPS, 2.2 TOPS/W with 2 MB on-chip RRAM) and NeuRRAM (2× energy efficiency with software-equivalent accuracy) [75]. Georgia Tech EIC Lab won Best Paper at MICRO 2024 for Fusion-3D 3D-stacked memory integration [76].

IBM Research's 64-core analog AI chip using phase-change memory achieved 12.4 TOPS/W sustained—14× more energy efficient than digital counterparts for speech recognition [77]. Their 3D analog architecture for Mixture-of-Experts outperforms GPUs in throughput, area, and energy efficiency per Nature Computational Science (2025).

HBM4 enters mass production Q3–Q4 2025 with 2048-bit interface (doubled from HBM3's 1024-bit), >40% power efficiency improvement, and 2+ TB/s data transfer speeds [78]. SK Hynix, Samsung, and Micron have all sold out 2026 supply.

4. Market Dynamics and Value Migration

4.1 Inference Spending Overtakes Training in 2026

The AI accelerator market reached $123 billion in 2024 and is projected to reach $207 billion in 2025 and $286–445 billion by 2030 [1, 2, 79]. The critical shift: inference spending overtakes training in 2026, with >65% of AI IaaS spending supporting inference by 2029.

Table 8: AI Accelerator Market Segmentation (Billions USD)

Market Segment	2024	2025	2030
Total AI Accelerator	$123B	$207B	$286–445B
AI Inference	$91B	$103–106B	$254–255B
Edge AI Chip	$20B	$25–26B	$59–66B
Data Center GPU	$18B	$10.5B (DC only)	$114B

NVIDIA maintains 86–94% data center GPU share but faces pressure from custom ASICs (Google TPUs at 13.1% share), AMD's TCO advantage (33% cost reduction for memory-bound inference), and hyperscaler vertical integration (Amazon Trainium powers 35% of new AWS AI workloads) [1, 24].

NVIDIA gross margins of 70–75% represent peak profitability, with Blackwell transition creating short-term pressure. AMD's aggressive pricing (MI300X at 2–4× lower acquisition cost than H100) and roadmap (MI400 targeting 40% better tokens/dollar than B200) creates sustained competitive pressure.

4.2 Supply Chain Constraints Persist Through 2027

HBM supply is fully committed: SK Hynix CFO confirmed "entire 2026 HBM supply sold out"; Micron's 2025–2026 capacity "fully booked"; Samsung acknowledges demand exceeds supply through 2026 [78]. HBM prices doubled in some cases through 2024–25.

TSMC CoWoS packaging represents the binding constraint: 70,000 wafers/month by end 2025 versus 1 million wafer demand by 2026 [64, 65]. NVIDIA alone requires 595,000 wafers—60% of projected global capacity. Intel's EMIB/Foveros gains interest from capacity-blocked customers.

CHIPS Act impact: $52.7B in US manufacturing incentives has catalyzed 140+ projects across 28 states with >$630B announced investments. TSMC Arizona, Intel Ohio, and Samsung Texas fabs represent strategic supply chain diversification from Taiwan concentration risk.

4.3 Sovereign AI Stacks Reshape Global Competition

China's full-stack approach (Ascend + CANN + MindSpore) reaches functional competitiveness at system level despite 3–5 year per-chip disadvantage [80, 31]. Nearly 50% of Chinese LLMs now train on Ascend chips [28]. Huawei's CloudMatrix 384 matches GB200 NVL72 aggregate compute at 5× power consumption—brute-force scaling compensates for efficiency gaps [32].

Middle East emergence: Saudi Arabia's HUMAIN ($100B mandate) and UAE's MGX ($100B fund) and Stargate UAE ($500B data center project) represent largest AI infrastructure investments outside US/China. Both remain dependent on US chip access, creating geopolitical leverage.

India's mature-node strategy (Tata Dholera at 28nm, 2026 production) addresses global chip shortage for power management, automotive, and IoT rather than competing on leading-edge AI chips [7, 6, 46]. Design ecosystem strength (20% of global semiconductor engineers) enables long-term capability building.

5. Architecture Predictions for 2026–2028

5.1 Winners and Losers Crystallize

Winners:

- **NVIDIA**: Maintains dominance through ecosystem lock-in, Vera Rubin roadmap (2026–2027), and strategic acquisitions (Groq partnership)
- **Google TPUs**: 60% of funded GenAI startups now on Google Cloud; Trillium achieves 67% better perf/watt than GPUs
- **CXL memory pooling**: 200–500ns latency versus 100μs for NVMe; essential for LLM serving infrastructure by 2027
- **Optical interconnects**: CPO becomes mandatory for AI data centers; NVIDIA commercial deployment 2026 [81]

Losers:

- **Intel AI accelerators**: $<$1% discrete share; Gaudi struggles despite spec competitiveness; Falcon Shores faces uphill battle
- **Pure-play analog compute startups**: Most consolidate or pivot; only IBM and well-funded players survive
- **Copper interconnects at rack scale**: Physical limits reached for high-bandwidth AI fabrics

5.2 The Next Architecture Converges on Heterogeneity

Specialized architectures don't converge—they coexist within heterogeneous systems:

Table 9: Architecture Role Specialization by 2028

Architecture	Optimal Use Case	2028 Role
GPU (NVIDIA/AMD)	Training, flexible inference	Primary training, high-throughput batch
TPU (Google)	Large-scale cloud training	Hyperscaler-captive
LPU/SRAM-based	Real-time low-latency inference	Voice AI, agents, robotics
Custom ASIC	Cost-optimized inference at scale	Hyperscaler internal deployment
NPU (edge SoC)	On-device inference	Mobile, automotive, IoT

CXL 4.0 (November 2025) doubles bandwidth to 128 GT/s, enabling true memory pooling across heterogeneous compute. By 2027, rack-scale coherent memory fabrics—KV cache distributed across disaggregated pools—become standard for LLM serving.

Optical interconnects transition from research to requirement: NVIDIA's Quantum-X Photonics (2026) and Marvell's Celestial AI acquisition ($3.25B) signal industry commitment [82]. Silicon photonics market grows from $2.86B (2025) to $28.75B (2034) at 29% CAGR.

6. Actionable Engineering Recommendations

For AI engineers and systems architects making infrastructure decisions in 2026:

Training infrastructure:

- NVIDIA Blackwell (GB200 NVL72) remains optimal for large-scale training where ecosystem maturity and maximum performance matter
- AMD MI350X/MI400 becomes viable for organizations with engineering capacity to optimize ROCm stack
- Reserve CoWoS/HBM capacity 18–24 months ahead—supply constraints persist through 2027

Inference infrastructure:

- Evaluate AMD MI300X/MI325X for memory-bound LLM inference: 33% TCO advantage for 70B+ models
- Consider Groq/Cerebras for latency-critical applications (sub-10ms TTFT requirements)
- Deploy liquid cooling for any new $>$50 kW rack installations—air cooling obsolete for AI density

Architectural choices:

- Implement PagedAttention (vLLM) for KV cache management: 2–5$\times$ batch capacity improvement
- Adopt FP8/FP4 quantization with micro-scaling: 2–3$\times$ throughput at $<$1% accuracy loss
- Plan for CXL memory pooling (2027+) for disaggregated inference serving

Supply chain strategy:

- Diversify beyond TSMC for non-leading-edge packaging (Intel EMIB gaining customers)
- Monitor HBM4 availability (H2 2025)—current suppliers sold out through 2026
- Track China domestic ecosystem—Ascend 910C functional for inference at 60% H100 performance

The defining constraint of 2026 is not compute—it is memory bandwidth, packaging capacity, and software ecosystem maturity. Architectural decisions must optimize across all three dimensions simultaneously.

Author's Note

This paper reflects applied analysis informed by work across multiple enterprise and hyperscale AI systems. The views expressed are intended to support architectural reasoning and strategic decision-making, rather than prescribe specific vendors, products, or implementations.

About the Author

Anjan Goswami is a Senior ML & AI Strategist with 20+ years building 0-to-1 systems across Microsoft, Adobe, Salesforce, Walmart, eBay, and Amazon. His experience spans search and recommendation systems, multimodal LLMs, and large-scale distributed ML platforms, with current work focusing on Copilot systems and document intelligence at Microsoft PowerPoint.

Executive Contact

Senior technical leaders or executives with questions related to AI silicon strategy, GPU infrastructure planning, or inference optimization architecture may contact the author for further discussion.

Report compiled December 2025. Data represents best available estimates from public sources, company announcements, and industry analysis. Projections involve inherent uncertainty.

References

[1] SQ Magazine. AI Chip Statistics 2025: Funding, Startups & Industry Giants, 2025. URL https://sqmagazine.co.uk/ai-chip-statistics/.

[2] Gartner. Gartner Says AI-Optimized IaaS Is Poised to Become the Next Growth Engine for AI Infrastructure, October 2025. URL https://www.gartner.com/en/newsroom/press-releases/2025-10-15-gartner-says-artificial-intelligence-optimized-iaas-is-poised-to-become-the-

[3] Cerebras Systems. Cerebras Wafer-Scale Cluster Datasheet. Technical report, Cerebras Systems, 2024. URL https://8968533.fs1.hubspotusercontent-na1.net/hubfs/8968533/Cerebras%20Wafer%20Scale%20Cluster%20datasheet%20-%20final.pdf.

[4] Various Authors. A Comparison of the Cerebras Wafer-Scale Integration Technology with Nvidia GPU-based Systems for Artificial Intelligence. *arXiv preprint*, 2025. URL https://arxiv.org/html/2503.11698v1.

[5] Information Technology and Innovation Foundation. How Innovative Is China in Semiconductors?, August 2024. URL https://itif.org/publications/2024/08/19/how-innovative-is-china-in-semiconductors/.

[6] The Volt Post. Tata Group, PSMC 28nm Chip Plant is First Mega Fab Plant, 2024. URL https://thevoltpost.com/tata-group-psmc-28nm-chip-fab-plant-in-dholera/.

[7] Dholera Times. India's First Semiconductor Fab in Dholera (2024–2028): Tata Electronics Timeline and Progress, 2024. URL https://www.dholeratimes.com/dholera-updates/blogs/tata-semiconductor-dholera-project-timeline-2024-2028.

[8] Wikipedia. Blackwell (microarchitecture), 2024. URL https://en.wikipedia.org/wiki/Blackwell_(microarchitecture).

[9] Hyperstack. NVIDIA Blackwell GB200 NVL72: Price and Specs Included, 2024. URL https://www.hyperstack.cloud/nvidia-blackwell-gb200.

[10] Server Simply. Blackwell B200 vs. Hopper H200 vs. H100, 2024. URL https://www.serversimply.com/blog/blackwell-b200-and-hopper-h200.

[11] AceCloud. Blackwell Architecture And Its Impact On Scalable Generative AI, 2024. URL https://acecloud.ai/blog/nvidia-blackwell-architecture-for-generative-ai/.

[12] BaCloud. Nvidia B200 vs AMD Instinct MI355X: Next-Gen AI Data Center GPU Showdown, 2024. URL https://www.bacloud.com/en/blog/203/nvidia-b200-vs-amd-instinct-mi355x-next-gen-ai-data-center-gpu-showdown.html.

[13] NVIDIA. The Engine Behind AI Factories — NVIDIA Blackwell Architecture, 2024. URL https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/.

[14] NVIDIA Developer. Introducing NVFP4 for Efficient and Accurate Low-Precision Inference, 2024. URL https://developer.nvidia.com/blog/introducing-nvfp4-for-efficient-and-accurate-low-precision-inference/.

[15] Emergent Mind. NVFP4 Quantization Algorithm Overview, 2024. URL https://www.emergentmind.com/topics/nvfp4-quantization-algorithm.

[16] Lambda. Accelerate Your AI Workflow with FP4 Quantization on Lambda, 2024. URL https://lambda.ai/blog/lambda-1cc-fp4-nvidia-hgx-b200.

[17] Fibermall. Introduction to NVIDIA GB200 Superchip and Liquid-Cooled Servers and Cabinets, 2024. URL https://www.fibermall.com/blog/nvidia-gb200-superchip.htm.

[18] Wccftech. NVIDIA Deep-Dives Into Blackwell Infrastructure: NV-HBI Used To Fuse Two AI GPUs Together, 5th Gen Tensor Cores, 5th Gen NVLINK & Spectrum-X Detailed, 2024. URL https://wccftech.com/nvidia-blackwell-ai-deep-dive-nv-hbi-fuse-two-ai-gpus-together-5th-gen-tensor-cores-5th

[19] NVIDIA. GB200 NVL72, 2024. URL https://www.nvidia.com/en-us/data-center/gb200-nvl72/.

[20] Tom's Hardware. AMD unveils Instinct MI300X GPU and MI300A APU, claims up to 1.6X lead over Nvidia's competing GPUs, 2023. URL https://www.tomshardware.com/pc-components/cpus/amd-unveils-instinct-mi300x-gpu-and-mi300a-apu-claims-up-to-16x-lead-over-nvidias-compet

[21] 36Kr. AMD's Aggressive Pricing Stabs Intel but Fails to Outperform NVIDIA, 2024. URL https://eu.36kr.com/en/p/3541331537719433.

[22] SemiAnalysis. MI300X vs H100 vs H200 Benchmark Part 1: Training - CUDA Moat Still Alive. December 2024. URL https://semianalysis.com/2024/12/22/mi300x-vs-h100-vs-h200-benchmark-part-1-training/.

[23] The Register. AMD slaps together a silicon sandwich with MI300 APUs, GPUs, 2023. URL https://www.theregister.com/2023/12/06/amd_mi300_gpu/.

[24] Thundercompute. ROCm vs CUDA: Which GPU Computing System Wins in December 2025?, 2025. URL https://www.thundercompute.com/blog/rocm-vs-cuda-gpu-computing.

[25] Intel. Gaudi 3 AI Accelerator White Paper. Technical report, Intel, 2024. URL https://cdrdv2-public.intel.com/817486/gaudi-3-ai-accelerator-white-paper.pdf.

[26] Various Authors. LPU: A Latency-Optimized and Highly Scalable Processor for Large Language Model Inference. *arXiv preprint*, 2024. URL https://arxiv.org/html/2408.07326v1.

[27] Groq. Inside the LPU: Deconstructing Groq's Speed, 2024. URL https://groq.com/blog/inside-the-lpu-deconstructing-groq-speed.

[28] IEEE Spectrum. China's AI Chip Race: Tech Giants Challenge Nvidia, 2024. URL https://spectrum.ieee.org/china-ai-chip.

[29] Tech Startups. Huawei begins mass shipment of Ascend 910C AI chip to fill Nvidia void in China, April 2025. URL https://techstartups.com/2025/04/21/huawei-begins-mass-shipment-of-ascend-910c-ai-chips-to-fill-nvidia-void-in-china/.

[30] Tom's Hardware. DeepSeek research suggests Huawei's Ascend 910C delivers 60% of Nvidia H100 inference performance, 2025. URL https://www.tomshardware.com/tech-industry/artificial-intelligence/deepseek-research-suggests-huaweis-ascend-910c-delivers-60-percent-nvidia-h100-inference

[31] Debuglies. China's AI Hardware Ecosystem in 2025: Huawei's Ascend Series, Indigenous Chip Development and the Trajectory of Global Technological Divergence, April 2025. URL https://debuglies.com/2025/04/29/chinas-ai-hardware-ecosystem-in-2025-huaweis-ascend-series-indigenous-chip-development-a

[32] Tech Startups. Huawei's Ascend 910C-powered system reportedly outperforms Nvidia's H100 on key metrics, April 2025. URL https://techstartups.com/2025/04/28/huaweis-ascend-910c-system-reportedly-outperforms-nvidias-h100-in-key-metrics/.

[33] Tom's Hardware. Huawei's Ascend AI chip ecosystem scales up as China pushes for semiconductor independence, 2024. URL https://www.tomshardware.com/tech-industry/semiconductors/huaweis-ascend-ai-chip-ecosystem-scales.

[34] ChinaTalk. Can Huawei Take On Nvidia's CUDA?, 2024. URL https://www.chinatalk.media/p/can-huawei-compete-with-cuda.

[35] Poniak Times. Huawei Open-Sources AI Stack: CANN, MindSpore & openPangu at Connect 2025, 2025. URL https://www.poniaktimes.com/huawei-open-source-ai-connect-2025/.

[36] Huawei. Groundbreaking SuperPoD Interconnect: Leading a New Paradigm for AI Infrastructure, September 2025. URL https://www.huawei.com/en/news/2025/9/hc-xu-keynote-speech.

[37] Center for Strategic and International Studies. Understanding the Biden Administration's Updated Export Controls, 2024. URL https://www.csis.org/analysis/understanding-biden-administrations-updated-export-controls.

[38] Baker McKenzie. US Department of Commerce Significantly Expands Controls Targeting Indigenous Production of Advanced Semiconductors in China, 2024. URL https://sanctionsnews.bakermckenzie.com/us-department-of-commerce-significantly-expands-controls-targeting-indigenous-production

[39] SemiAnalysis. Huawei Ascend Production Ramp: Die Banks, TSMC Continued Production, HBM is The Bottleneck, 2024. URL https://newsletter.semianalysis.com/p/huawei-ascend-production-ramp.

[40] The Register. SiPearl details Rhea1 chip for European exascale system, 2024. URL https://www.theregister.com/2024/05/14/sipearl_rhea1_specs/.

[41] AnandTech. SiPearl's Rhea-2 CPU Added to Roadmap: Second-Gen European CPU for HPC, 2024. URL https://www.anandtech.com/show/21295/sipearls-rhea2-cpu-added-to-roadmap-second-gen-european-cpu-for-hpc.

[42] EuroHPC JU. The European Processor Initiative (EPI), 2024. URL https://eurohpc-ju.europa.eu/research-innovation/our-projects/european-processor-initiative-epi_en.

[43] Wilson Center. The European Chips Act: A Vital Step In the Right Direction, 2024. URL https://www.wilsoncenter.org/article/european-chips-act-vital-step-right-direction.

[44] Tata Group. Tata Group to Build the Nation's First Fab in Dholera, 2024. URL https://www.tata.com/newsroom/business/first-indian-fab-semiconductor-dholera.

[45] Yahoo Finance. Construction starts on AI-enabled semiconductor fab in Gujarat, India, 2024. URL https://finance.yahoo.com/news/construction-starts-ai-enabled-semiconductor-115605262.html.

[46] Santhosh Gandhi. Decoding Semiconductor Market for India, 2024. URL https://medium.com/@isanthoshgandhi/decoding-semiconductor-market-for-india-9765a6102fe8.

[47] Deepak Kumar Sahoo. FlashAttention-3: The Engine Powering Next-Gen LLMs, 2024. URL https://medium.com/the-synaptic-stack/flashattention-3-the-engine-powering-next-gen-llms-30b2843bb182.

[48] Dao-AILab. flash-attention: Fast and memory-efficient exact attention, 2024. URL https://github.com/Dao-AILab/flash-attention.

[49] Various Authors. Memory Is All You Need: An Overview of Compute-in-Memory Architectures for Accelerating Large Language Model Inference. *arXiv preprint*, 2024. URL https://arxiv.org/html/2406.08413v1.

[50] IEEE Spectrum. Memory Chips That Compute Will Accelerate AI, 2024. URL https://spectrum.ieee.org/amp/processing-in-dram-accelerates-ai-2656045722.

[51] IBM Research. In-memory computing, 2024. URL https://research.ibm.com/projects/in-memory-computing.

[52] Clarifai. NVIDIA B200 Vs. H100: Choosing The Right GPU For Your AI Workloads, 2024. URL https://www.clarifai.com/blog/nvidia-b200-vs-h100.

[53] Introl. Structured Cabling vs Liquid Cooling for 100 kW Data Center Racks, 2024. URL https://introl.com/blog/structured-cabling-vs-liquid-cooled-conduits-designing-for-100-kw-plus-racks.

[54] Datacenters.com. Why Liquid Cooling Is Becoming the Data Center Standard, 2024. URL https://www.datacenters.com/news/why-liquid-cooling-is-becoming-the-data-center-standard.

[55] Microelectronics UK. Photonics powering AI data centres: The latest innovations, 2025. URL https://microelectronicsuk.com/blog1/photonics-powering-ai-data-centres-latest-innovations.

[56] STMicroelectronics. Light into data: How silicon photonics is powering the AI data center revolution, 2024. URL https://blog.st.com/data-silicon-photonics-ai/.

[57] NVIDIA Developer. Scaling AI Factories with Co-Packaged Optics for Better Power Efficiency, 2024. URL https://developer.nvidia.com/blog/scaling-ai-factories-with-co-packaged-optics-for-better-power-efficiency/.

[58] SemiAnalysis. AMD MI300 – Taming The Hype – AI Performance, Volume Ramp, Customers, Cost, IO, Networking, Software, 2024. URL https://newsletter.semianalysis.com/p/amd-mi300-taming-the-hype-ai-performance.

[59] Wikipedia. AMD Instinct, 2024. URL https://en.wikipedia.org/wiki/AMD_Instinct.

[60] AMAX. Comparing NVIDIA Blackwell Configurations, 2024. URL https://www.amax.com/comparing-nvidia-blackwell-configurations/.

[61] Exxact Corp. NVIDIA Blackwell Architecture, 2024. URL https://www.exxactcorp.com/blog/hpc/nvidia-blackwell-architecture.

[62] Cerebras Systems. Cerebras Systems Unveils World's Fastest AI Chip with Whopping 4 Trillion Transistors, 2024. URL https://www.cerebras.ai/press-release/cerebras-announces-third-generation-wafer-scale-engine.

[63] IEEE Spectrum. Cerebras WSE-3: Third Generation Superchip for AI, 2024. URL https://spectrum.ieee.org/cerebras-chip-cs3.

[64] 36Kr. Who Will Divide Up the CoWoS Production Capacity in 2026?, 2024. URL https://eu.36kr.com/en/p/3580962946874242.

[65] Tom's Hardware. TSMC's CoWoS packaging capacity reportedly stretched due to AI demand, 2024. URL https://www.tomshardware.com/tech-industry/semiconductors/intel-gains-ground-in-ai-packaging-as-cowos-capacity-remains-stretched.

[66] Nikulsinh Rajput. Triton vs CUDA for Mortals, 2024. URL https://medium.com/@hadiyolworld007/triton-vs-cuda-for-mortals-2bcf95399223.

[67] PyTorch. PyTorch 2.x, 2024. URL https://pytorch.org/get-started/pytorch-2-x/.

[68] Yicai Global. Who is the domestic computing power chip leader? Ten questions AI big model (II), 2023. URL https://www.yicaiglobal.com/star50news/2023_10_10661071923731496960 2.

[69] SCIMUS. ROCm vs CUDA: A Practical Comparison for AI Developers, 2024. URL https://thescimus.com/blog/rocm-vs-cuda-a-practical-comparison-for-ai-developers/.

[70] TechPowerUp. AMD's Pain Point is ROCm Software, NVIDIA's CUDA Software is Still Superior for AI Development: Report, 2024. URL https://www.techpowerup.com/330155/amds-pain-point-is-rocm-software-nvidias-cuda-software-is-still-superior-for-ai-developr

[71] Citizen Watch Report. In Depth Benchmarking tests of AMD MX300 vs Nvidia H100 and H200 by SemiAnalysis, 2024. URL https://citizenwatchreport.com/in-depth-benchmarking-tests-of-amd-mx300-vs-nvidia-h100-and-h200-by-semianalysis/.

[72] Technology Newsroom. Loihi 2 for Neuromorphic Computing, 2024. URL https://technologynewsroom.com/tech-news/loihi-2-for-neuromorphic-computing/.

[73] Various Authors. Neuromorphic hardware for sustainable AI data centers. *arXiv preprint*, 2024. URL https://arxiv.org/html/2402.02521v2.

[74] Stanford SystemX Alliance. SystemX Alliance, 2024. URL https://systemx.stanford.edu/.

[75] Stanford SystemX Alliance. Stanford engineers present new chip that ramps up AI computing efficiency, 2022. URL https://systemx.stanford.edu/news/2022-08-18-000000/stanford-engineers-present-new-chip-ramps-ai-computing-efficiency.

[76] Georgia Tech EIC Lab. EIC lab - Georgia Tech, 2024. URL https://eiclab.scs.gatech.edu/.

[77] IBM Research. Analog in-memory computing could power tomorrow's AI models, 2024. URL https://research.ibm.com/blog/how-can-analog-in-memory-computing-power-transformer-models.

[78] NPR. Memory loss: As AI gobbles up chips, prices for devices may rise, December 2025. URL https://www.npr.org/2025/12/28/nx-s1-5656190/ai-chips-memory-prices-ram.

[79] Business Research Insights. Artificial Intelligence (AI) Chips Market Size — CAGR of 36.6%, 2024. URL https://www.businessresearchinsights.com/market-reports/artificial-intelligence-ai-chips-market-105026.

[80] Recode China AI. Huawei Claims AI Chips Surpass Nvidia's A100, 2024. URL https://recodechinaai.substack.com/p/huawei-claims-ai-chips-surpass-nvidias.

[81] AI World Journal. 2026 AI Compute Predictions: The Shift Beyond Silicon Has Begun, 2026. URL https://aiworldjournal.com/2026-ai-compute-predictions-the-shift-beyond-silicon-has-begun/.

[82] Optics & Photonics News. Marvell Looks to Acquire Celestial AI, December 2025. URL https://www.optica-opn.org/home/industry/2025/december/marvell_looks_to_acquire_celestial_ai/.