PART II · CHAPTER 3B

GPU Generation Roadmap

H100 to GB300, MI300X to MI455X — Three Generations, Two Vendors, $400 Billion in Silicon: How We Got Here and Where It Goes

Book: AI/ML Infrastructure from Silicon to Scale — V2 Edition | **Author:** Venkat Vinjam | **Cross-ref:** Ch 3A: GPU Architecture | Ch 4: Memory Hierarchy | Ch 5: Thermal/TCO

> *"The AI accelerator market is not a race between two companies. It is a race between two paradigms: NVIDIA's closed, fully-optimized stack versus an emerging open ecosystem. The winner shapes how AI is computed for the next decade."*
>
> — AI Infrastructure Strategy Perspective, 2026

3B.1 How to Read a GPU Roadmap

Before diving into specific chips, establish the right mental model for evaluating any GPU generation. Four lenses determine whether a generation will dominate workloads in production:

The Four Evaluation Lenses

① **Compute Density:**

FLOPs per watt, per dollar, per rack unit. Key metrics: Peak TFLOPS (BF16, FP8, FP4), TDP, TFLOPS/W.

② **Memory Architecture:**

Capacity, speed, scaling. Key: HBM GB, bandwidth TB/s, $/GB.

③ **Interconnect:**

Scale beyond a single GPU. Key: NVLink/Infinity Fabric GB/s, IB integration, topology.

④ **Software Stack:**

Programming model maturity. Key: CUDA/ROCm ecosystem, library coverage, framework support, debuggability.

A GPU that wins on compute but loses on interconnect is useless for training 100B+ parameter models. A GPU with great specs but poor software support costs months of engineering time to deploy productively.

Confidence Labels — Required Reading

Every specification in this chapter carries one of three confidence labels reflecting data reliability as of May 2026:

- **SHIPPED** Verified shipping hardware with measured benchmarks from MLPerf or vendor data sheets
- **ANNOUNCED** Vendor-disclosed specifications, not yet shipping at publication date

- ROADMAP From roadmap presentations — treat as directional, subject to change

3B.2 NVIDIA Hopper Generation — H100 and H200

3B.2.1 H100 — The Defining AI GPU SHIPPED

The H100 (2022–2023) is the chip that defined the modern AI infrastructure era. When historians write about the GPU moment in AI, they will cite the H100 as the inflection point where AI compute became a national strategic resource.

```
H100 SXM5 KEY SPECIFICATIONS:
  Architecture:    Hopper (GH100 die, 80B transistors, TSMC 4nm)
  Tensor Cores:    4th generation, FP8 native
  Peak BF16:       989 TFLOPS
  Peak FP8:        1,979 TFLOPS  ← Transformer Engine enabled
  HBM3:            80 GB, 3.35 TB/s bandwidth
  NVLink 4.0:      900 GB/s bidirectional per GPU
  NVSwitch 3.0:    Full bisection bandwidth in 8-GPU DGX
  TDP:             700W (SXM5), 350W (PCIe)
  Process:         TSMC 4nm
```

Critical Hopper Innovations

① **Transformer Engine** — Monitors activation magnitudes, computes per-tensor scaling factors, casts to FP8 for GEMM (near-2× throughput), accumulates in FP32. Result: ~1.9× GEMM throughput over pure BF16 with automatic quality preservation.

② **FP8 Format Pair (E4M3/E5M2)** — E4M3 for weights/activations (precision, limited range ±448); E5M2 for gradients (range, handles magnitude spikes). Together they enable FP8 mixed-precision training without manual per-layer configuration.

③ **Thread Block Clusters** — Groups of thread blocks sharing distributed SMEM across SMs, enabling super-large GEMM tiles spanning multiple SMs. Required for H100 to achieve near-peak TFLOPS on large matrix ops.

④ **NVLink 4.0 / NVSwitch 3.0** — Previous NVLink 3.0: 600 GB/s; H100 NVLink 4.0: 900 GB/s (+50%). NVSwitch 3.0 delivers 3× bandwidth per switch chip, making Tensor Parallelism within 8-GPU DGX dramatically more efficient.

Metric	A100 SXM SHIPPED	H100 SXM SHIPPED	Improvement
Peak BF16	312 TFLOPS	989 TFLOPS	**+3.2×**
Peak FP8	N/A	1,979 TFLOPS	**New format**
HBM Capacity	80 GB	80 GB	—
HBM Bandwidth	2.0 TB/s	3.35 TB/s	**+67%**
NVLink BW	600 GB/s	900 GB/s	**+50%**
TDP	400W	700W	**+75%**
LLaMA-2 70B training	Baseline	~3×	**3×**
LLaMA-2 70B inference	Baseline	~2.5×	**2.5×**

The Software Flywheel Effect

H100's performance improved ~2.9× over its hardware-only baseline through software updates alone:
Launch: ~1,100 tok/s → TRT-LLM (2023): ~1,900 tok/s → FP8 inference: ~2,800 tok/s → FlashAttention v3 + CUDA Graphs: ~3,200 tok/s final peak.

Lesson:

Hardware specs at launch are not the final performance number. NVIDIA's software flywheel is its true competitive moat.

3B.2.2 H200 — Memory Refresh, Not Architecture Refresh SHIPPED

⚠️ **Interview Trap**

"H200 has more TFLOPS than H100." —
FALSE.

"H200 is faster at everything." —
FALSE.

H200 is the same Hopper GH100 die with upgraded HBM. Identical compute-bound performance. Dramatically better memory-bound performance.

```
H200 vs H100 — THE CRITICAL DIFFERENCE:

  H100 → H200 changes:
    HBM3    → HBM3e       (faster interface standard)
    80 GB   → 141 GB      (+76% capacity)
    3.35 TB/s → 4.8 TB/s  (+43% bandwidth)
    700W    → 700W        (IDENTICAL TDP)
    TFLOPS  → same        (IDENTICAL compute units)

  LLaMA-3 70B in BF16:
    H100: 140 GB weights → MUST use TP=2 (2 GPUs to fit)
    H200: 140 GB weights → fits in 141 GB → SINGLE-GPU serving!

  Result: H200 achieves ~40% more throughput on 70B inference vs H100
          despite identical compute — purely from memory capacity/bandwidth.
```

The H200 shipped in early 2024 as a "soft upgrade" path — same SXM5 form factor, same power envelope, same cooling, allowing customers to swap modules without rack redesign. This is the "server platform continuity" play.

3B.3 NVIDIA Blackwell Generation — B200, GB200, GB200 NVL72

3B.3.1 Blackwell Architecture — The Dual-Die Breakthrough SHIPPED

Blackwell (2024–2025) represents NVIDIA's most dramatic architectural leap since the original GPU. The fundamental design breakthrough is the dual-die MCM (Multi-Chip Module), driven by a hard physical constraint: the EUV reticle limit.

```
BLACKWELL DUAL-DIE DESIGN:

  Problem: Modern EUV has a maximum exposure area (~858 mm²)
  GH100 (H100): 814 mm² — already at the reticle limit
  Solution: Two dies connected at high bandwidth (NV-HBI)

  B200 construction:
    2× B100 compute dies (each ~450 mm², TSMC 4NP)
    NV-HBI bandwidth:  10 TB/s (3× H100's HBM bandwidth!)
    Total transistors:  208 billion (vs 80B on H100 — 2.6×)
    HBM3e:             192 GB, 8.0 TB/s (2.4× H100)
    TDP:               1,000W

  KEY PERFORMANCE:
    Peak FP8:     4.5 PFLOPS     (2.3× H100)
    Peak NVFP4:   9.0 PFLOPS     (new format, B200-native)
    NVLink 5.0:   1.8 TB/s/GPU   (2× H100 NVLink 4.0)
```

3B.3.2 NVFP4 — The Format That Changes Training Economics

NVFP4 is a 4-bit float: 1 sign bit, 2 exponent bits, 1 mantissa bit. Block scaling lets every 16 weights share one FP8 scale factor, giving better-than-naive-INT4 precision due to floating-point representation.

> **Why FP4 Matters More Than FP8 Did**
>
> FP8 vs BF16: 2× throughput gain (H100 era). FP4 vs FP8: another 2× gain. Cumulative: 4× throughput gain from H100 BF16 → GB300 FP4.
>
> More critically: GB300 demonstrated
> **FP4 training**
> in MLPerf v5.1 — the first vendor to achieve FP4 training at scale with accuracy targets met. NVIDIA is the only vendor to submit FP4 training results.

3B.3.3 GB200 NVL72 — The Rack as a Single Computer SHIPPED

```
GB200 NVL72 ARCHITECTURE:

  The GB200 "superchip":
    1× Grace CPU (ARM Neoverse V2, 72 cores)
    2× B200 GPUs
    NVLink-C2C: CPU↔GPU = 900 GB/s (7× PCIe Gen5)
    CPU memory:  480 GB LPDDR5X at 960 GB/s (coherent with GPU HBM!)

  GB200 NVL72 rack — the new unit of compute:
    72 B200 GPUs + 36 Grace CPUs
    NVLink fabric:  130 TB/s total rack bandwidth
    Compute:        1.4 exaFLOPS FP8 / 2.8 exaFLOPS FP4
    HBM memory:     72 × 192 GB = 13.8 TB GPU HBM
    CPU memory:     36 × 480 GB = 17.3 TB LPDDR5X (coherent)
    Total:          ~31 TB unified (GPU + CPU)
    Power:          ~120 kW per rack
    Cooling:        Direct liquid cooling MANDATORY

  TENSOR PARALLELISM ADVANTAGE:
    Standard cluster (H100):  TP=8 within node, PP spans nodes via IB
    GB200 NVL72:              TP=72 — ENTIRE RACK as one TP group!
    Practical result: Models needing 8-way TP (H100) can use
    72-way TP at NVLink speeds. Sub-millisecond intra-GPU latency
    for trillion-parameter inference throughout.
```

3B.4 NVIDIA Blackwell Ultra — GB300 / B300

Spec	B200 SXM SHIPPED	GB300 SXM ANNOUNCED	Delta
HBM type	HBM3e	HBM3e (12-high stacks)	New stacking
HBM Capacity	192 GB	279 GB	**+45%**
HBM Bandwidth	8.0 TB/s	~12 TB/s	**+50%**
NVFP4 Compute	~9 PFLOPS	~20 PFLOPS	**+122%**
TDP	1,000W	~1,400W	**+40%**
Cooling	Liquid required	DLC mandatory (98%)	**Stricter**

MLPerf Training v5.1 — GB300 NVL72 Results

Llama 3.1 405B, 512 GPUs:

64.6 minutes to target accuracy

— industry record.

vs H100 equivalent: ~6.2× faster. vs B200: ~2× faster.

Won ALL 7 training benchmarks in v5.1. First platform to achieve Llama 405B training in under 2 hours at 512-GPU scale.

3B.5 AMD CDNA 3 — MI300X

3B.5.1 Architecture: The Memory Moat Strategy SHIPPED

⚠️ **Common Misconception — MI300X Architecture**

WRONG:

"MI300X is a 3-chiplet design."

CORRECT:

MI300X has

10 dies total

— 6 XCDs (compute) + 4 MCDs (memory controllers).

```
MI300X CORRECT ARCHITECTURE:

  6 × XCDs (eXtended Compute Dies):
    CDNA 3 compute architecture, TSMC 5nm
    ~50 Compute Units per XCD
    Total CUs: 304 across all 6 XCDs

  4 × MCDs (Memory Controller Dies):
    HBM controllers, TSMC 6nm
    3 HBM3 stacks per MCD = 12 HBM3 stacks total

  Package: Silicon interposer connecting all 10 dies

  KEY SPECIFICATIONS:
    Peak BF16:       ~1.3 PFLOPS (1,307 TFLOPS)
    HBM3:            192 GB, 5.3 TB/s bandwidth
    Infinity Fabric: ~896 GB/s GPU-GPU (intra-node)
    TDP:             750W
```

3B.5.2 The Memory Moat — Why 192 GB Matters

Model	BF16 Size	H100 GPUs Needed	MI300X GPUs Needed	Winner
LLaMA-3 8B	16 GB	1×	1×	Tie
LLaMA-3 70B	140 GB	2× (TP=2 required)	**1× (single GPU!)**	**MI300X**
LLaMA-3 405B	810 GB	12× (TP=8, PP≥2)	**5×**	**MI300X**
Mixtral 8×22B	281 GB	4×	**2×**	**MI300X**

Single-GPU serving of 70B eliminates: AllReduce per transformer layer, 1 GPU × $25K cost, networking overhead, and 2-process deployment complexity. **This is AMD's core TCO argument — and it is mathematically valid.**

3B.5.3 MI300X Performance Reality — Honest Assessment

```
WHERE MI300X WINS:
  ✓ Memory capacity: 192 GB vs 80 GB (2.4× advantage)
  ✓ LLM inference (large models): 1×MI300X vs 2×H100 → system-level cost win
  ✓ Memory bandwidth: 5.3 TB/s vs 3.35 TB/s (+58%)
  ✓ Long-context (128K+ tokens): full KV cache resident in single GPU
  ✓ MoE models (Mixtral, DeepSeek): larger expert cache fits in HBM

WHERE H100 WINS:
  ✗ FLOPS/watt: H100 more efficient compute per watt
  ✗ Tensor Core efficiency: NVIDIA kernel library maturity
  ✗ Software ecosystem: CUDA, TRT-LLM vs ROCm
  ✗ Tokens/sec/GPU (small models): H100 wins on latency at batch=1

MLPerf Inference v4.0 — Llama 2 70B, 8-GPU server:
  MI300X:    ~11,700 tokens/sec/GPU
  H100:      ~23,500 tokens/sec/GPU

  MI300X at ~50% of H100 tok/s per GPU...
  BUT for 70B: MI300X serves on 4 GPUs vs H100 requires 8 GPUs!

  System-level (8 GPUs each):
  H100 (8-GPU, 2× TP groups):   ~23,500 tok/s
  MI300X (8-GPU, 8 single-GPU): ~93,600 tok/s (8 independent models!)

  AMD's counter: "Compare SYSTEMS not chips." For 70B inference: valid.
```

3B.6 AMD CDNA 3+ — MI325X

```
MI325X — THE BRIDGE CHIP (2024):

  vs MI300X:
    Same XCD compute dies (identical CDNA 3 compute)
    HBM3 → HBM3e: 192 GB → 256 GB (+33% capacity)
    Bandwidth: 5.3 → 6.0 TB/s (+13%)
    TDP: 750W → ~1,000W

  KEY TIMING ISSUE:
    MI325X mass shipments: Q2 2025
    B200 launch:          Also Q2 2025
    Result: Most hyperscalers SKIPPED MI325X

  ONE NOTABLE WIN:
    MLPerf Training v5.0 — Llama 2 70B LoRA fine-tuning:
      MI325X: 21.1 min  ← AMD's FIRST-EVER MLPerf Training win!
      H200:   23.1 min
    Why AMD won: LoRA keeps base model frozen; larger HBM = more
    adapter candidates resident simultaneously.
```

> **Lesson: Interim Chip Timing Risk**
>
> Interim chip refreshes only succeed if they ship well *ahead* of the competitor's next generation. AMD learned this with MI325X — shipping simultaneously with B200 meant most buyers skipped straight to MI355X.

3B.7 AMD CDNA 4 — MI355X

3B.7.1 Specifications and the Competitive Leap SHIPPED

```
MI355X — CDNA 4 (2025):

  Architecture: NEW XCD design (not just memory refresh)
    2 XCDs (vs 6 smaller ones in MI300X) — fewer cross-die crossings
    185B transistors per XCD (high density)
    FP4/FP6 support: New Matrix Cores (matching B200)
    HBM3e:     288 GB (vs 256 GB MI325X)
    Bandwidth: ~8 TB/s (matching B200's 8 TB/s)
    TDP:       ~1,400W
    Peak FP8:  ~5 PFLOPS (vs MI300X ~1.3 PFLOPS — 3.85×)

MLPerf Inference v5.0 vs B200 (Llama 2 70B):
  Offline throughput:    MI355X at 97% of B200  ← nearly competitive!
  Server (interactive):  MI355X at 119% of B200 ← AMD WINS latency!

  AMD winning the interactive scenario is strategically significant:
  for latency-sensitive apps (chatbots, code completion),
  MI355X is actually FASTER per-GPU than B200.
```

> **Software vs Hardware Contribution to MI355X Gains**
>
> ~30% of MI355X's MLPerf improvement over MI325X came from **software**
>
> (ROCm 7.1): better FlashAttention ROCm implementation, improved GEMM kernel selection, reduced CPU overhead in ROCm runtime.
>
> Lesson: Hardware matters, but closing the software gap mattered equally. This is AMD's most important signal that ROCm is maturing.

3B.8 AMD MI400 Series — The Next Generation

The MI400 series (announced CES 2026, shipping H2 2026) represents AMD's first architecture with HBM4 and is the basis for the landmark Meta and OpenAI customer commitments. ANNOUNCED

Product	Target Market	Memory	Interconnect	Availability
MI430X	Sovereign AI / HPC / Hybrid	HBM4	UALink	H2 2026
MI440X	Mainstream AI Training/Inference	HBM4	UALink	H2 2026
MI450 / MI450X	Hyperscale (OpenAI, Meta custom)	432 GB HBM4	IF128 Ethernet	H2 2026 (1 GW)
MI455X / Helios Rack	Rack-Scale / Yotta-Scale	31 TB HBM4/ rack	UALink fabric	Q3 2026 est.

```
HBM4 TECHNICAL LEAP (MI450 vs MI355X):
  MI355X HBM3e: 288 GB at ~8 TB/s
  MI450  HBM4:  432 GB at 19.6 TB/s

  Capacity: +50%
  Bandwidth: +145% — the largest single-generation HBM BW jump ever

  HBM4 BW gains come from:
    1. Wider buses per stack (2048-bit vs 1024-bit HBM3e)
    2. Higher per-pin data rate
    3. More stacks per GPU (8 stacks vs ~6 on MI355X)

  Performance implication for decode (LLaMA-3 70B):
    MI355X (~8 TB/s):  baseline decode tok/s
    MI450 (~19.6 TB/s): ~2.5× MI355X decode throughput
```

3B.8.1 The Meta and OpenAI Deals — Strategic Significance

Dimension	OpenAI Deal	Meta Deal	Combined
Announced	Oct 6, 2025	~Feb 2026	—
Total committed	6 GW of GPUs	6 GW of GPUs	12 GW
First deployment	1 GW, H2 2026	1 GW, H2 2026	2 GW by 2026
Est. GPU count	~550–600K	~550–600K	~1.1–1.2M
Est. revenue	~$20B+	~$20B+	~$40B+ GPU alone
Warrants issued	160M AMD shares	160M AMD shares	320M (~20% dilution)
Vesting condition	GPU shipment	GPU shipment	Performance-based

AMD currently earns ~$16.6B/year in data center revenue total. These two deals alone could represent ~$40B+ in GPU revenue over ~5 years — potentially 2.4× AMD's current annual data center revenue. The equity-for-demand structure is unprecedented in semiconductors: AMD offers performance-based warrants that only vest when GPUs ship.

> **Why Meta and OpenAI Agreed**
>
> Both are NVIDIA's largest customers facing: supply constraints (NVIDIA pricing power), price concentration risk ($25–55K/GPU, single vendor), and technology dependency (CUDA lock-in).
>
> Diversifying to AMD provides: negotiating leverage with NVIDIA, supply chain redundancy, technical co-engineering influence on MI450 architecture, and potential cost savings at hyperscale.

3B.9 Power and TCO Across Generations

Generation	GPU	TDP	Cooling Required	Rack Density
NVIDIA Hopper	H100 SXM	700W	Air or liquid	~82 kW/rack
NVIDIA Hopper+	H200 SXM	700W	Air or liquid	~82 kW/rack
NVIDIA Blackwell	B200 SXM	1,000W	**Liquid MANDATORY**	~115 kW/rack
NVIDIA Blk Ultra	GB300 SXM	1,400W	**DLC required (98%)**	~120 kW/rack
AMD CDNA 3	MI300X	750W	Air or liquid	~88 kW/rack
AMD CDNA 3+	MI325X	~1,000W	**Liquid recommended**	~115 kW/rack
AMD CDNA 4	MI355X	~1,400W	**Liquid mandatory**	~120 kW/rack

> **Key Infrastructure Trend**
>
> Every new generation requires liquid cooling. Air-cooled datacenters are becoming AI-legacy infrastructure. Operators with only air cooling CANNOT run B200/GB300/MI355X without significant facility upgrades (~$5–15M per 100 kW of new liquid capacity).

Tokens-per-Watt Efficiency (LLaMA-2 70B, 8-GPU server)

GPU	Tokens/sec	Power Draw	Tokens/kW	vs H100
H100	23,500	5.6 kW	4,196	Baseline
H200	33,000	5.6 kW	5,893	**+40%**
B200	75,000	8.0 kW	9,375	**+123%**
GB300	1M+ tok/s/rack	~100 kW	~10,000	**~2.4×**
MI325X	32,350	8.0 kW	4,044	~H100 parity
MI355X	100K+ tok/s	11.2 kW	~9,000	**~3×**

Despite rising TDP, efficiency (performance-per-watt) consistently improves. The extra watts buy more tokens-per-watt than the baseline generation — this is the architectural efficiency gain from process node improvements and architectural optimization.

3-Year TCO Comparison — Production Decision Making

Metric	H100	H200	B200	MI300X	MI355X
Hardware (capex)	$350K	$450K	$550K	$300K	$500K
Power 3yr ($0.07/kWh)	$43K	$43K	$62K	$46K	$86K
Cooling infra	Low (air OK)	Low	**HIGH (liquid)**	Medium	**HIGH (liquid)**
Software risk	Low (CUDA)	Low	Low	Medium (ROCm)	Lower (ROCm 7)
Total 3yr est.	$420K	$520K	$650K+	$370K	$620K
Performance vs H100	Baseline	**1.4×**	**~3×**	0.7× (chip-level)	**~2.8×**
Tokens/$K (normalized)	100	108	**184**	95*	**181**

*MI300X serves 70B on 1 GPU vs 2 H100s → system-level TCO is actually competitive for large-model inference workloads.

3B.10 Competitive Landscape and Forward Roadmap

NVIDIA Vera Rubin — After Blackwell Ultra ROADMAP

```
NVIDIA VERA RUBIN (2026–2027, estimated):
  Named after: Vera Rubin (astronomer who discovered dark matter evidence)
  Process: TSMC N3 (3nm) — first NVIDIA chip on 3nm
  Architecture: New "Rubin" GPU (post-Blackwell)
  Memory: HBM4 (matching/exceeding AMD MI450)
  NVLink 6: Expected ~3.6 TB/s per GPU
  NVSwitch 5: Full mesh for NVL rack-scale systems

  Expected performance:
  GB300 FP4: ~20 PFLOPS → Rubin FP4: ~35–40 PFLOPS (estimated)
  VR NVL72 equivalent: Target 2–3 exaFLOPS FP4 per rack
  HBM4 capacity: 400+ GB per GPU (estimated)
```

The Strategic Battle — 2026 Perspective

Dimension	NVIDIA (Sustainable)	AMD (Achievable, Uncertain)
Software	**CUDA: 18+ year head start, millions of devs**	**ROCm 7.x: 3–5 year gap remains**
Inference	**TRT-LLM: best-in-class optimization**	**vLLM ROCm: improving but behind**
Ecosystem	**AWS, Azure, GCP all deeply committed**	**Fewer cloud providers, fewer integrations**
Memory	Matched by B200/GB300	**Historically 2–3 gen ahead on HBM**
Open ecosystem	**CUDA lock-in**	**ROCm + PyTorch + Triton = no lock-in**
Strategic customers	All major hyperscalers	**12 GW committed: Meta + OpenAI**
Compute gap	**GB300 leads MI355X by ~3–4×**	Closed: MI355X at 97–119% of B200
Market share (2027 est.)	~75% (down from ~85%)	**~12–15% (up from ~5–8%)**

> **The $12B Question**
>
> Can AMD ship 12 GW of MI450 by 2031? This is the execution risk — AMD has never shipped at this volume. The total market is growing fast enough that AMD can gain share without NVIDIA losing absolute revenue. Not a zero-sum game.

Architecture Diagrams

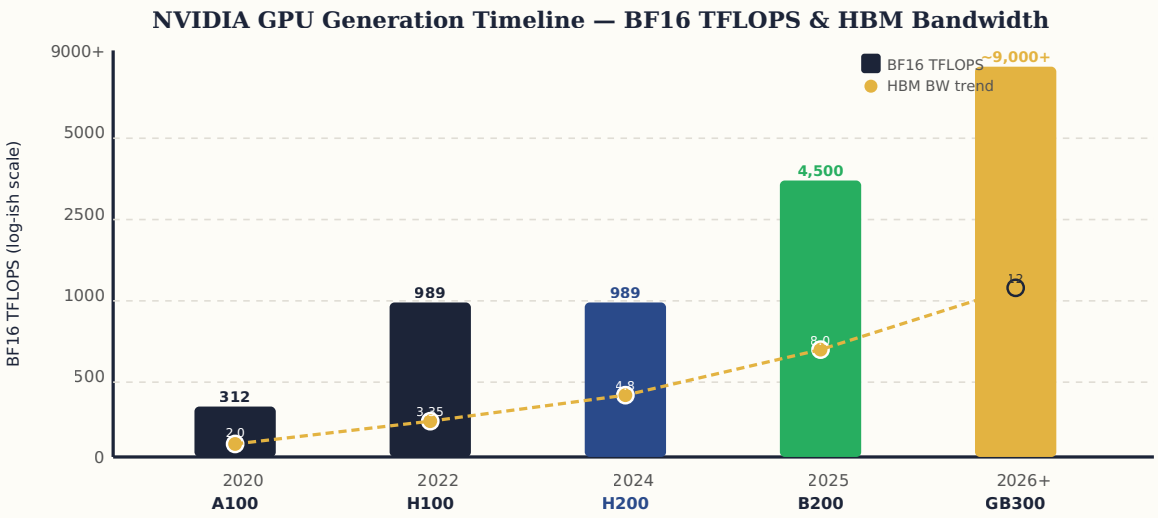


Figure 3B.1 — NVIDIA GPU generation timeline. Navy bars = BF16 TFLOPS (left scale, log-compressed). Gold circles/dashed line = HBM bandwidth (TB/s). Note the H200 has identical TFLOPS to H100 — its gain is purely from HBM3e capacity and bandwidth. GB300 TFLOPS refers to NVFP4 (FP4 equivalent throughput).

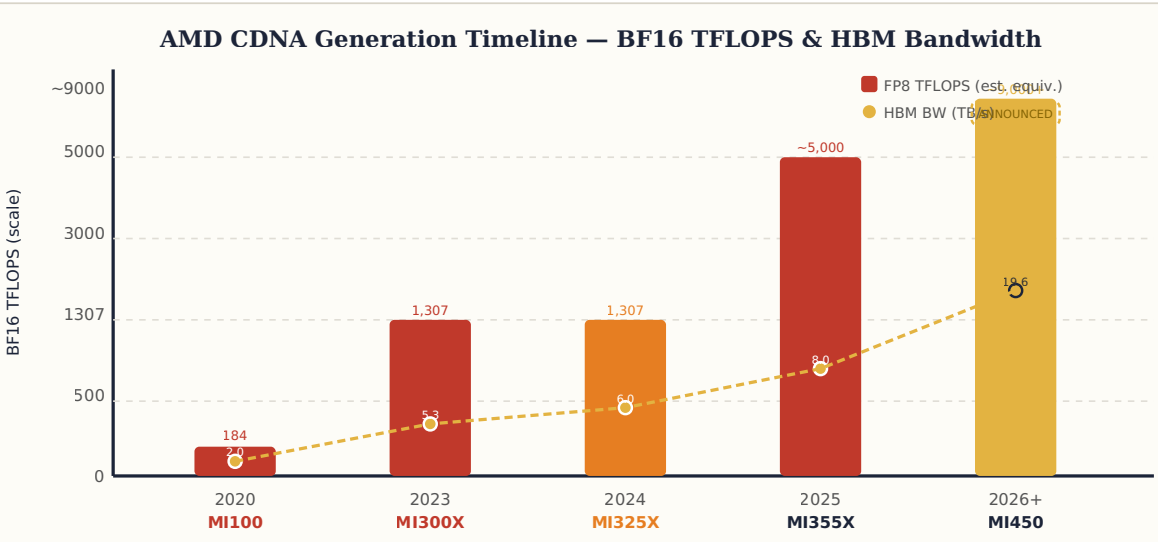


Figure 3B.2 — AMD CDNA generation timeline. Red bars = FP8/BF16 TFLOPS. Gold dashed line = HBM bandwidth (TB/s). Note MI325X has identical compute to MI300X — it is a memory upgrade only. MI450 specs are ANNOUNCED, subject to change. MI450 BW of 19.6 TB/s represents a 145% jump — the largest single-generation HBM bandwidth increase ever.

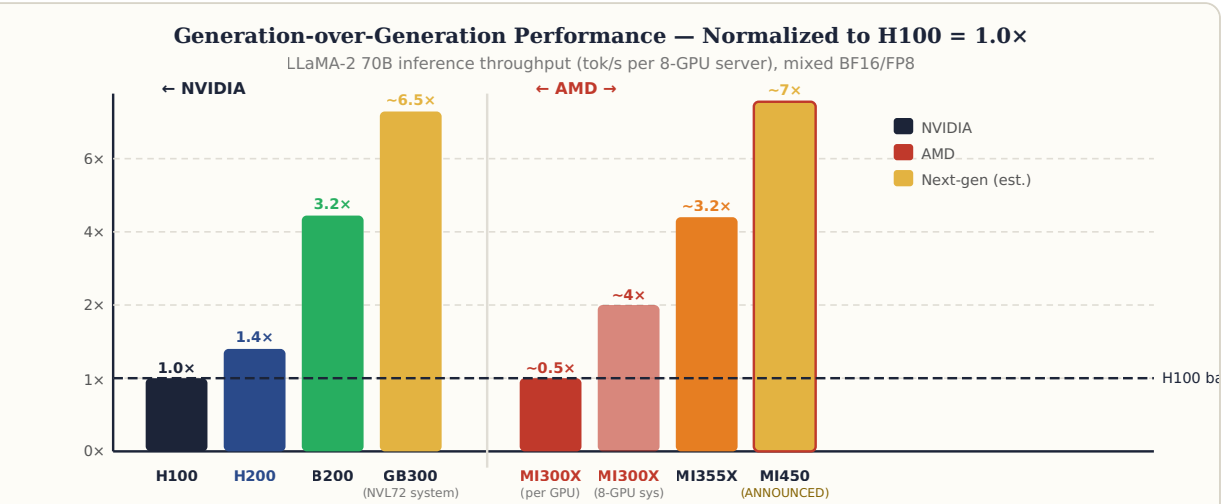


Figure 3B.3 — Generation-over-generation performance comparison, normalized to H100=1.0×. AMD MI300X is shown twice: per-GPU (0.5×) and 8-GPU system-level for 70B inference (4×, because MI300X can run 8 independent 70B models vs H100's 4 TP=2 groups). MI450 is ANNOUNCED; numbers are estimates based on HBM4 bandwidth scaling.

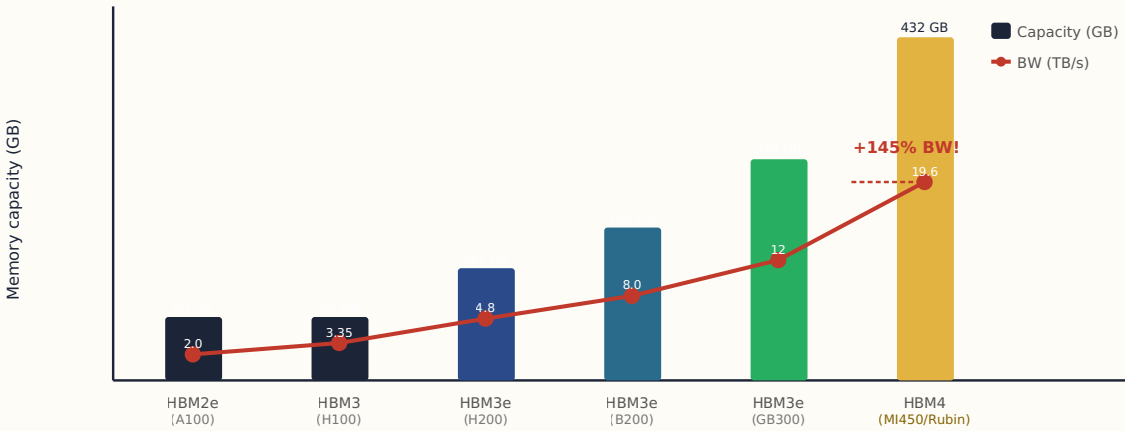


Figure 3B.4 — HBM technology evolution across GPU generations. Navy/green bars = capacity (GB). Red line = bandwidth (TB/s). The jump from HBM3e (GB300) to HBM4 (MI450/Rubin) represents a 145% bandwidth increase — driven by wider 2048-bit buses and higher per-pin data rates. Capacity nearly doubles to 432 GB per GPU.

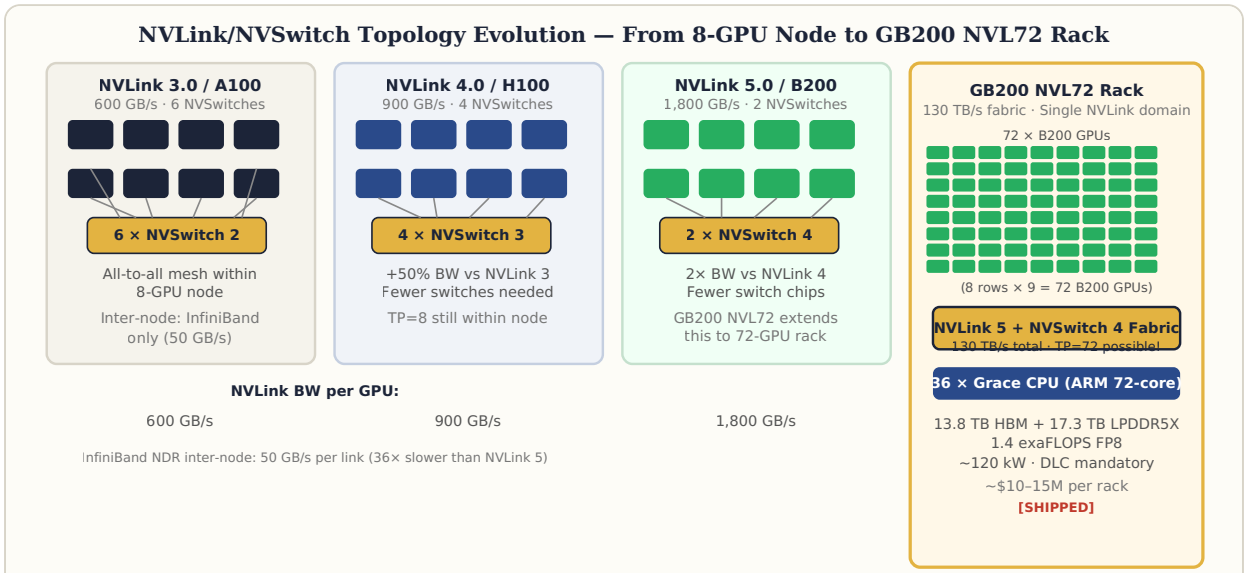


Figure 3B.5 — NVLink/NVSwitch topology evolution. Left three panels: 8-GPU DGX-style nodes with increasing NVLink bandwidth (600 → 900 → 1,800 GB/s per GPU). Right panel: GB200 NVL72 rack — 72 GPUs unified in a single NVLink domain at 130 TB/s total fabric bandwidth, enabling TP=72 at NVLink latency, eliminating InfiniBand overhead for intra-rack tensor parallelism.

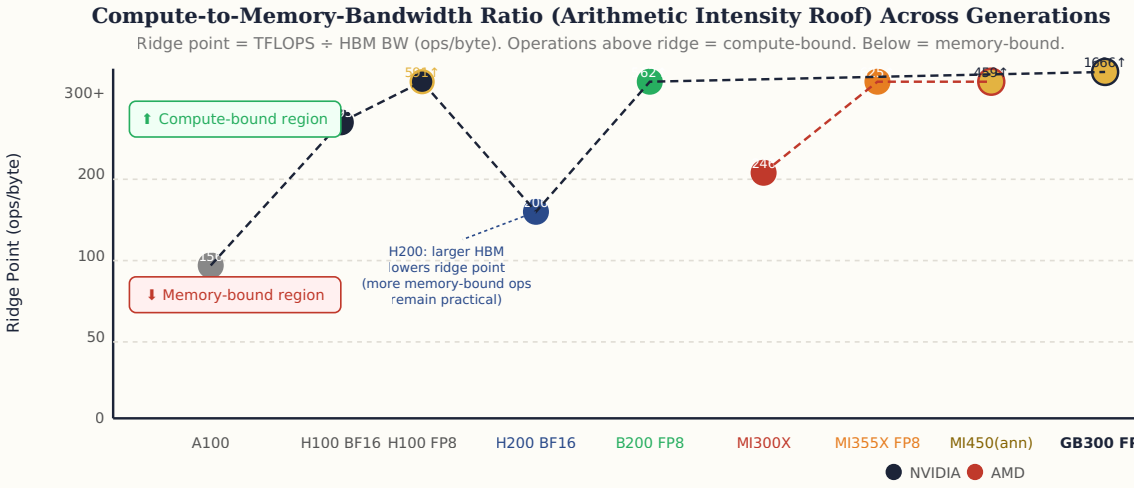


Figure 3B.6 — Arithmetic intensity ridge points across GPU generations. Higher ridge = GPU can execute more compute-bound workloads efficiently. Values above the chart cap (marked ↑) indicate extremely high ridge points where almost all practical LLM operations are memory-bound at batch=1 but compute-bound at batch≥16. Note: H200's lower ridge than H100 FP8 reflects its expanded HBM capacity making memory-bound operations cheaper — a feature, not a regression.

Quick Reference — GPU Generation Numbers

GPU	Vendor	Process	TDP	HBM GB	HBM TB/s	Peak BF16 TFLOPS	Status
A100 SXM	NVIDIA	TSMC 7nm	400W	80	2.0	312	SHIPPED
H100 SXM	NVIDIA	TSMC 4nm	700W	80	3.35	989	SHIPPED
H200 SXM	NVIDIA	TSMC 4nm	700W	141	4.8	989 (same!)	SHIPPED
B200 SXM	NVIDIA	TSMC 4NP	1,000W	192	8.0	4,500 (FP8)	SHIPPED
GB300 SXM	NVIDIA	TSMC 4NP	~1,400W	279	~12	~20,000 (FP4)	ANNOUNCED
GB200 NVL72	NVIDIA	—	~120 kW	13,824	130 TB/s fabric	1.4 exaFLOPS FP8	SHIPPED
MI300X	AMD	TSMC 5/6nm	750W	192	5.3	1,307 (6XCD+4MCD)	SHIPPED
MI325X	AMD	TSMC 5/6nm	~1,000W	256	6.0	1,307 (same compute)	SHIPPED
MI355X	AMD	TSMC 4nm	~1,400W	288	~8.0	~5,000 (FP8)	SHIPPED
MI450	AMD	TSMC 3nm	TBD	432	19.6	~9,000 (FP8 est.)	ANNOUNCED
Vera Rubin	NVIDIA	TSMC N3	TBD	400+	TBD	35–40K (FP4 est.)	ROADMAP

Chapter 3B Review Questions

3. **The H200 has the same Tensor Core count and clock speed as the H100, yet achieves 40% higher throughput on LLaMA-3 70B inference. Explain the exact mechanism.** What class of operation benefits and what does not? Why does fitting the model in a single GPU matter beyond just avoiding a second GPU cost?

3. **Explain the Blackwell dual-die design.** What physical constraint drove this architectural decision? What is the bandwidth of the NV-HBI inter-die connection, and why is that bandwidth critical — specifically, what failure mode occurs if inter-die bandwidth is lower than HBM bandwidth?

3. **A company is choosing between 8× H100 GPUs and 8× MI300X GPUs for LLaMA-3 70B inference serving (BF16, batch=32).** Compare the two setups including: required tensor parallelism degree, effective tokens/sec per dollar, and the single most important qualitative factor that could swing the decision toward AMD even if H100 wins on raw tok/s.

3. **AMD's MI325X shipped in Q2 2025 and most hyperscalers skipped it.** What specific timing dynamic explains this? What does this teach about the economics of GPU generation transitions? Under what conditions would a bridge chip refresh succeed?

3. **The Meta and OpenAI deals with AMD include performance-based warrants.** Explain why this structure benefits both parties. What is the maximum dilution AMD faces if all milestones are met, and what must have happened for that dilution to occur? Why would Meta want AMD's stock to appreciate?

3. **Compute-to-memory-bandwidth ratio (ridge point) has been increasing with each GPU generation for FP8/FP4 formats, yet decreasing for the memory-upgrade variants (H200 vs H100 BF16).** Is this a contradiction? Explain what practical implication each trajectory has for inference serving at batch=1 vs batch=128.

Chapter Cross-References

Ch 3A GPU architecture fundamentals — SM structure, CUDA programming model, Tensor Cores
Ch 4 Memory Hierarchy — HBM deep dive, bandwidth roofline, KV cache management
Ch 5 Thermal & TCO — cooling infrastructure, liquid cooling design, $/token analysis
Ch 6 Inference systems — batching strategies, continuous batching, throughput vs latency
Ch 14 Networking — InfiniBand vs NVLink, RDMA, collective communications
Ch 17 Telemetry — GPU performance monitoring, dcgm-exporter, NVML metrics