PART 7 · SECOND EDITION · SEPTEMBER 2026 (FIRST PUBLISHED JULY 2026)

Compute, Silicon & System Integration

You never buy a chip. You marry an architectural family, queue for its memory, inherit its software stack, and take delivery of a rack — and only realized goodput ever pays it back.

END OF 2025 · THE MEMORY MARKET

Micron CY2026 HBM supply contracted

Micron had contracted its own calendar-2026 HBM supply by December 2025.

A supplier disclosure does not reserve your accelerator. Contract the qualified memory, package and system-release milestones.

MICRON · AS-OF 2025-12-17

THE REFRAME

An accelerator is a memory system, a software contract and a power envelope you pour into concrete.

The model must fit memory, its kernels must run, and the selected rack's power and cooling become building work.

ACT I

Choosing the silicon family

Classify accelerators on independent axes: availability, programmability, specialization, memory/interconnect, software and procurement. Familiar labels overlap; they are not mutually exclusive families.

THE TERRITORY

Four common labels — compare them on independent axes

MARKET LABEL	WHAT YOU BUY	WHAT IT LOCKS	PRIMARY CONSEQUENCE
Merchant GPU (NVIDIA, AMD)	Named chip/system and supported CUDA / ROCm release	Vendor roadmap and allocation queue	CUDA/ROCm skills and a hardware-resale option; qualify releases and date resale assumptions
Systolic TPU (Google)	Mostly rented pods; since Q2 2026 also in select customers' own data centers	XLA / JAX, Google's software stack	Renting TPU capacity leaves no owned chip to resell; exiting may require a compiler port
Hyperscaler XPU (Trainium, Maia, MTIA)	Trainium rental; Maia/MTIA internal capacity	Captive SDK, thin third-party support	Avoids a merchant purchase price but bears design, software, yield, packaging, capacity and lifecycle cost
Inference ASIC (SambaNova, d-Matrix)	A specialized, supported execution path	A fixed workload shape	Wins while the model architecture holds; obsolesces when it shifts

THE MASTER FORK

Rent a portable stack, or own a captive one?

Merchant buyer

— Pay the quoted complete-system price

— Hire CUDA/ROCm engineers; qualify the release and date the resale assumption

— Order against an allocation queue you do not control

— The default below frontier-self-build scale

Captive silicon program

— Avoid a merchant purchase price; bear NRE, yield, packaging, software and program risk instead

— Multi-year design-to-volume lead time, every generation

— Captive compiler; effectively no resale market

— Pays when committed useful service covers program cost

SO WHAT

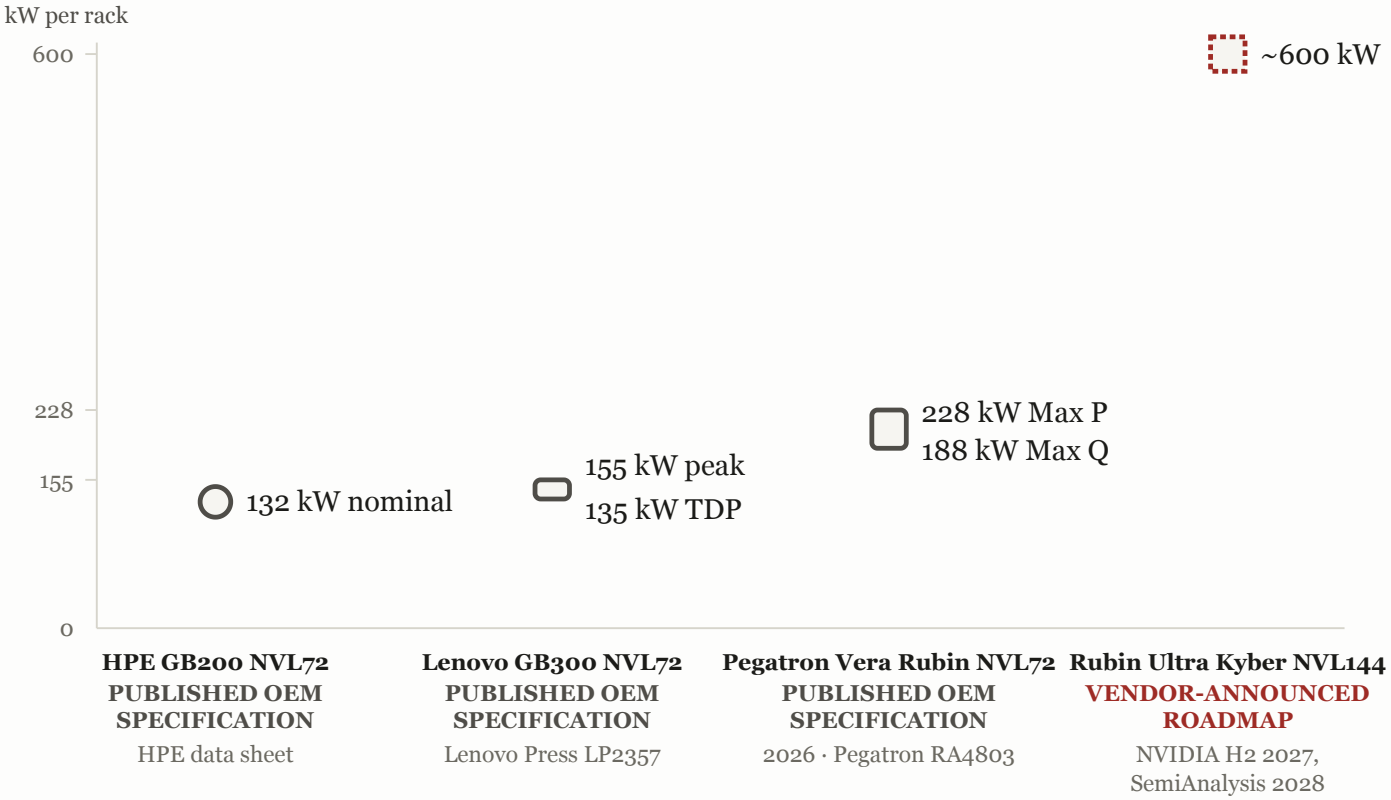
A TSMC/CoWoS design using HBM can share wafers, package slots and memory suppliers with NVIDIA/AMD; a new logo alone does not diversify that queue.

NAMED PROFILES

Keep OEM operating ratings separate from facility allowances

Rack power by named profile: published specs vs the ~600 kW roadmap

Vendor operating profiles, not one measured time series.



Rack kilowatts alone do not choose air or liquid cooling; heat split, component heat flux, inlet air and facility water do.

Schematic · Source: HPE + Lenovo + Pegatron + NVIDIA · as of October 2026

HPE, Lenovo and Pegatron profiles stay distinct. Kyber is a roadmap option; neither it nor a facility allowance is an operating measurement.

FOUR NAMED PROFILES

Keep the lifecycle attached to every rack-power point

132 kW nominal TDP: 115 kW liquid + 17 kW air

NVIDIA GB200 NVL72 by HPE — 132 kW nominal TDP with separate liquid and room-air duties

HPE, NVIDIA GB200 NVL72 BY HPE DATA SHEET, OCTOBER 2026 EDITION. · AS-OF 2026-10

NVIDIA RA up to 142 kW; Lenovo 135 kW TDP, up to 155 kW peak, ~90% liquid / ~10% air; HPE 132 kW nominal, ~155 kW EDPp

Lenovo GB300: 135 kW TDP / 155 kW peak; separate from NVIDIA's facility allowance

NVIDIA GB300 NVL72 ENTERPRISE REFERENCE ARCHITECTURE · AS-OF 2026-09

188 kW (Max Q) / 228 kW (Max P)

Pegatron Vera Rubin: 188 kW Max Q / 228 kW Max P; distinct from the DSX facility basis

PEGATRON RA4803-72N3 DATASHEET · AS-OF 2026-08

~600 kW

Kyber NVL144 (NVIDIA target H2 2027, contested): ~600 kW per-rack plan on 800 VDC

NVIDIA GTC · AS-OF H2 2027 ROADMAP (GTC 2025); 2028 REPORTED JULY 2026

THE DEPRECIATION FIGHT

5–6 yr book / 4–6 yr estimate / 2–3 yr bear case

Book policy vs published useful-life estimate vs the contested 2–3-year bear case

A contested figure — it hangs on residual demand for trailing silicon. Ride every generation, skip one, or hold: choose by substrate readiness and token economics, not habit.

GOLDMAN SACHS GLOBAL INSTITUTE, TRACKING TRILLIONS · AS-OF 2026-04

THE SECOND SOURCE

AMD: qualify the node port and the training fabric

Instinct nodes for inference

— More usable HBM can reduce model/KV sharding

— Test ROCm at the same model, precision, quality and latency

— Fund the port and recurring release support

— Keep a supported rollback path within the node

Helios rack-scale for training

— Qualify the delivered UALoE implementation

— Test collectives, congestion and fault recovery at scale

— Changing the fabric costs retuning and requalification

— Contract switch, firmware and delivery milestones

SO WHAT

The ROCm port and the Helios fabric are different commitments. Compare their measured useful-output benefit with the software and migration work each requires.

CAPACITY AND SOFTWARE

Memory is a specification; useful output belongs to a tested release

432 GB

AMD MI455X: 432 GB HBM4 capacity; specification checked September 8, 2026

AMD MI455X PRODUCT SPECIFICATIONS · AS-OF 2026-10

~29% lower (B200 1.41x)

MI355X / ROCm / SGLang: the dated FP8 workload example, not a vendor-wide win

SEMIANALYSIS INFERENCEX · AS-OF 2026-05

WORKED EXAMPLE

Custom silicon: does service volume repay the program?

Illustrative assumptions; equal useful service per device; undiscounted screen.

Design + software + integration	$400M
Three-year sustaining + delay reserve	$100M
Saving per matched device	$15,000
Committed volume	40,000 devices
Base surplus / volume threshold	**About $100M / 33,334 devices**

Exact threshold: $500M/$15k; 33,334 is the first positive integer. At 30,000: –$50M; one-year slip: –$100M. Appraisal: 1.8.

ACT II

The gate: memory and the package

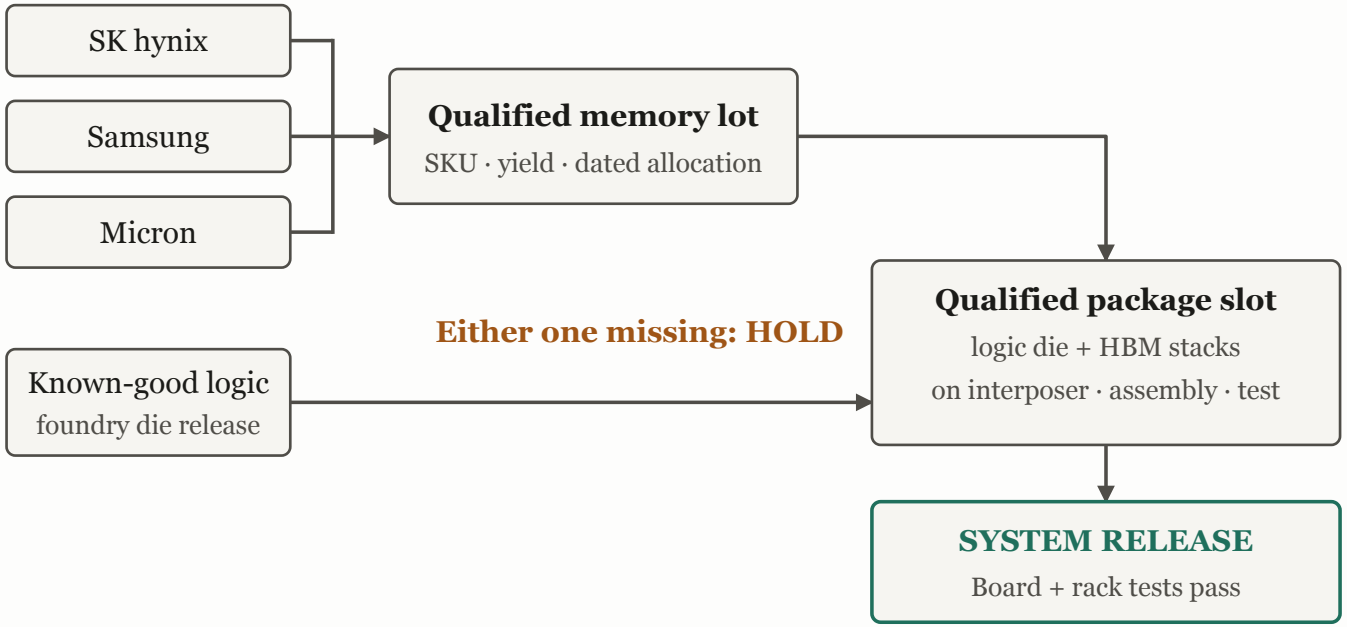
Whatever silicon you chose, it ships only when three DRAM houses and one packaging queue say it ships. The binding constraint on AI compute sits here, upstream of every purchase order.

MEMORY GATES

Memory fit, bandwidth and qualified delivery are separate gates

A system ships only when the memory lot and the package slot both clear

Supplier qualification and package assembly are separate gates.



Separate test: does the workload fit and feed?

Capacity: weights + KV + reserve must fit usable HBM

Speed: bytes moved ÷ sustained bandwidth is only one part of step time

Schematic · Source: Micron + NVIDIA + guide synthesis · as of September 2026

The model must fit and run fast enough; a qualified HBM lot and a package-release slot must both arrive.

WORKED EXAMPLE

HBM: one replica can fit and still miss its latency target

Illustrative BF16/GQA decode; decimal GB; full trace in 7.6.

Weights + KV + reserve	140 + about 21 + 20 GB
Usable GB200 capacity	186 GB
Memory lower bound at 4 TB/s	40 ms
Capacity ceiling	**9 sequences**

Exact tensor counts permit nine sequences; ten need about 187 GB and fail 186 GB. About 3.2 TB/s is the rounded latency boundary; test the complete service.

THE RETICLE WALL

858 mm² (26 × 33 mm)

The largest area one scanner exposure can print — the limit all advanced packaging exists to defeat

CoWoS stitches multiple reticles of silicon and a dozen HBM stacks into one package. How many reticles the interposer spans is the real gate on capacity per package.

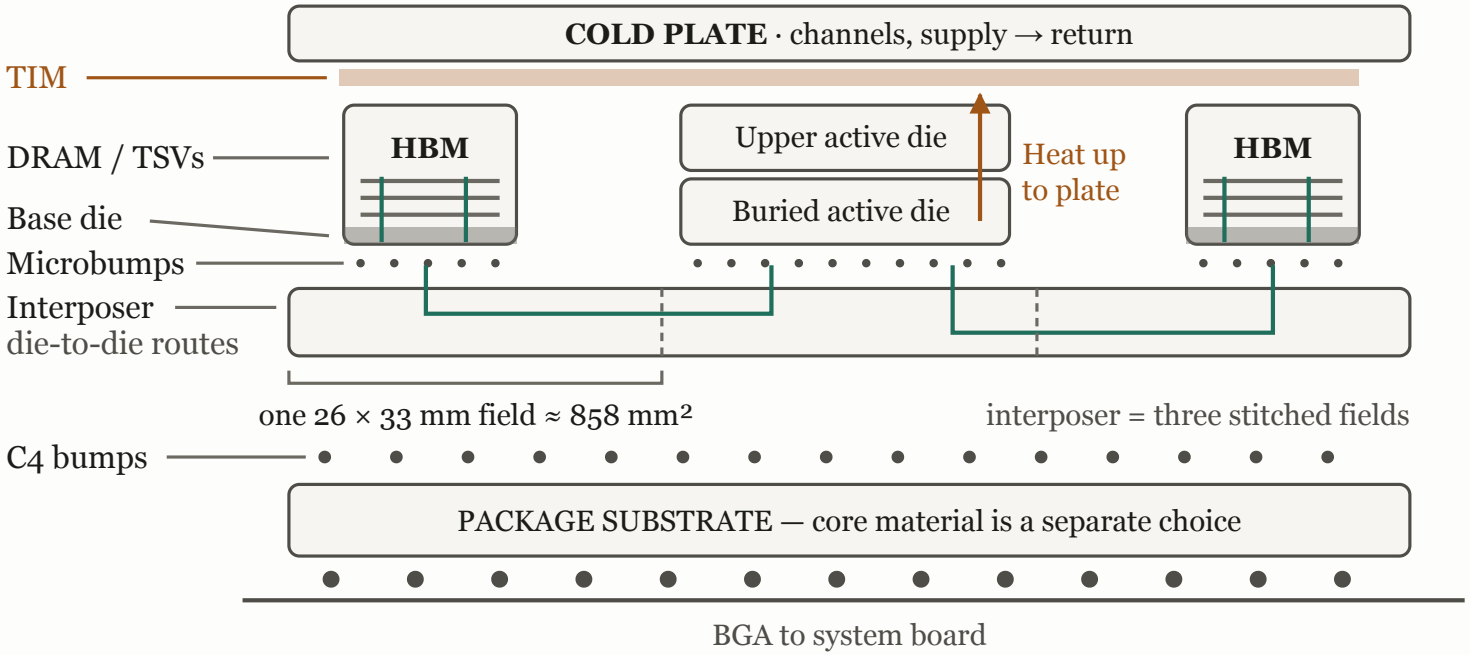
ASML · AS-OF 2026-09

PACKAGE PATHS

Stitch the interposer, then prove its heat path

A package must carry bits and remove heat

Conceptual cross-section, not to scale; compute drawn as a 3D-stacked option under a top cold plate.



Stitching adds alignment and yield risk.

Test package yield, test access and heat removal together.

Schematic · Source: ASML + TSMC + UCIe + guide synthesis · as of September 2026

A reticle field is about 858 mm²; stitching adds alignment and yield constraints. HBM layers, base die and TSVs remain distinct from the buried-die heat route.

WORKED EXAMPLE

Package choice must close link power and heat

Illustrative one-way payload and combined PHY energy.

1 TB/s × 8 × 1.10 framing	8.8 Tb/s
PHY A at 1 pJ/bit	8.8 W
PHY B at 3 pJ/bit	26 W
600 W × 0.040 K/W	24 K rise
20 W link / 30 K thermal gates	**A passes; B fails**

A flips above about 2.3 TB/s or 0.050 K/W. UCIe version alone does not qualify interchangeability or assembly yield.

THE LAST METER

up to ~2,000 A (typical peak, per Vicor)

Peak core current Vicor calls typical at under a volt, where microohms burn hundreds of watts

The same synchronized load step that droops the die is the seed of the facility transient three layers up. Power delivery is moving onto the package — and next, under it.

VICOR · AS-OF 2026-10

ACT III

The system you actually run

The chip is a third of the machine. Host composition, the software regime, and the precision ladder decide whether purchased FLOPS ever become delivered tokens.

WORKED EXAMPLE

Buy CPU service capacity before adding GPUs

Illustrative host pool; CPU busy time excludes tool/network wait.

100 req/s × 0.20 CPU-s/req	20 busy cores
At 70% occupancy ceiling	29 cores required
Request state + runtime	100 GB
32 cores / 128 GB	**Passes screen**

The thin 16-core host clears at most 56 req/s; the 32-core pool clears 112 req/s. Profile queues and p99 before release.

WORKED EXAMPLE

A second stack needs a qualified release and a funded owner

Illustrative ROCm/SGLang port, not a vendor tax.

Port work × loaded rate	200 h × $150/h
Initial port / annual sustaining	$30,000 / $36,000
Gross saving × 12 months	$8,000 × 12
First-year surplus	**$30,000**

Flip below $5,500/month gross saving. Pin the full hardware/software/model tuple and pass quality, scale, recovery and rollback.

WORKED EXAMPLE

A faster precision recipe can still fail the service

Synthetic fixed 10,000-item evaluation; floor 89%; p99 ≤50 ms.

FP8 correct / rate / p99	8,940 / 1,200 tok/s / 45 ms
NVFP4 correct / rate / p99	8,880 / 1,600 tok/s / 42 ms
Selection	**FP8**

NVFP4 needs at least 8,900 correct items while preserving latency and useful output. A percentage-point drop is not a relative-percent drop.

ACT IV

From purchase order to producing cluster

Selection is a constrained optimization; delivery is a logistics and acceptance discipline. The contest is scored in cost per useful token — and decided on the install floor.

WORKED EXAMPLE

Technical eligibility precedes the finance model

Illustrative 48,000 tok/s service and 1,000 kW IT-input budget.

A: 48 nodes, 960 kW	Fails 180 GB memory need
B: 60 nodes, 960 kW	Passes memory and power
C: 69 nodes, 1,035 kW	Fails power budget
B nominal arrival	**Week 20**

A one-week system-release slip leaves no eligible bid. Send B's 60-node, 960 kW service, cost and delivery figures to Chapter 1.8.

WORKED EXAMPLE

Rail loss and shelf response have different boundaries

Illustrative path, shelf and ideal storage screen.

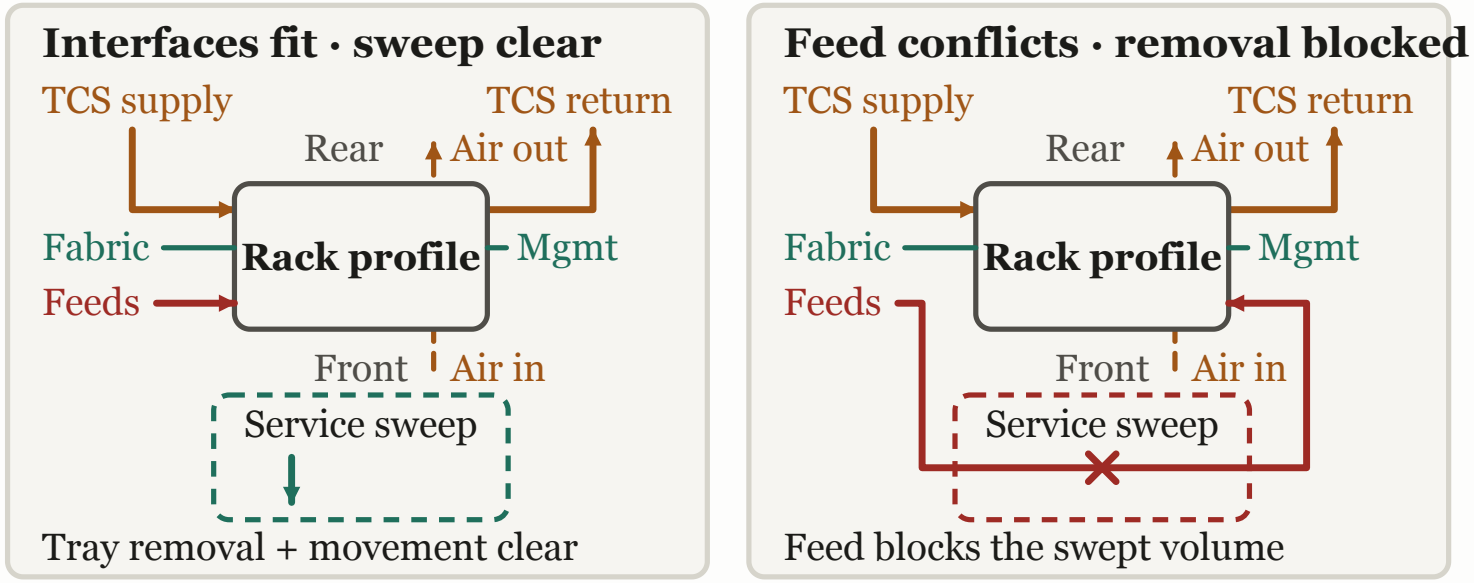
2,000 A squared × 20 μΩ	80 W rail-path loss
R/L step versus 40 mV budget	30 mV
120 kW / 96% shelf efficiency	125 kW input
20 kW × 10 ms	200 J
50→46 V ideal capacitance	**1.0 F fails; 1.1 F screens**

1.0 F supplies 192 J, below 200 J. Ideal 1.1 F supplies about 210 J; validate ESR, aging and dynamics. At 95% efficiency the input limit fails.

RACK INTERFACES **Busbar, manifolds, ToR and NVLink share the rack**

Keep feeds out of the tray-removal space

OCP MGX Rack & Trays Rev. 1.1 · illustrative plan view · TCS: technology cooling system



All interfaces + sweep pass
Release the rack position

Any interface or sweep fails
Reroute the feed or change the layout

Guide scenario · Source: Guide derivation · as of Sep 2026

Compute trays and power shelves need power, liquid and fiber routes that leave repair access. Set the mating geometry before long-lead fabrication orders.

WORKED EXAMPLE

Spare electrical capacity does not clear the rack

Wholly hypothetical Rack A, not a mixed OEM profile.

Electrical / liquid margins	12 kW / 20 kW
Room-air deficit	2.0 kW
Rear service deficit	0.10 m
Nominal release	**HOLD**

Both failed interfaces must be ready by week 12 for the two-week test to finish in week 14; readiness in week 15 moves release to week 17.

WORKED EXAMPLE

A fast cluster that cannot recover stays on HOLD

Synthetic two-node run with a predeclared recovery gate.

Useful output / run time	3.78M tokens / 3,600 s
Throughput; quality; p99	1,050 tok/s; 89.2%; 48 ms
Fault recovery / allowed	150 s / 120 s
Release	**HOLD**

Fix the 30-second recovery excess; retest performance, quality, manifest, telemetry, update and rollback. Zero unresolved critical defects.

THE VELOCITY EXEMPLAR

Colossus: a dated build history, a separate crew calculation

NVIDIA reported October 28, 2024: 122 days for the build; 19 days from first rack to training.

Reported 100k-GPU build	122 days
First rack → training	19 days
Illustrative station/crew case in 7.15	5 pre-floor + 10 floor days

The 72-rack case needs six crews and ≤12.5% rework. Colossus's 100,000/64 ≈1,563 rack-equivalents is arithmetic, not observed racks or a reusable crew rate.

THE THROUGH-LINE

Choose a family, queue for its memory, inherit its software and sign for the rack. Useful output is what repays those commitments.

TAKE IT HOME

What to do Monday

01 Pick the accelerator family by software and fabric lock-in, not the datasheet — the stack outlives the silicon.

02 Force every comparison to dense, one precision, sustained MFU on your own model — never the headline FLOPS.

03 Secure HBM and CoWoS allocation as a matched pair — a stack without a packaging slot is stranded BOM.

04 Compose the host for the workload: agentic inference at a training-era GPU:CPU ratio strands the fleet.

05 Name your denominator — $/GPU-hr, $/M-tokens, or tokens/MW — before you shortlist a single chip.

06 Contract the L12 gate on goodput with a multi-day burn-in; pre-terminate the fiber and screen optics off the floor.

THE DEFINITIVE GUIDE TO

AI Data *Centers*

This deck is the map. The guide is the territory — every number here is dated, sourced, and kept current in the full chapters.

Ch 7.1 The four accelerator families and the merchant-vs-captive fork

Ch 7.6 HBM: the binding constraint on AI compute

Ch 7.9 Software ecosystems, lock-in, and the realized-MFU gap

Ch 7.14 Integration levels, burn-in, and the goodput acceptance gate

READ THE FULL PART, WITH EVERY SOURCE AND DERIVATION, AT AIDATACENTERGUIDE.COM

SOURCES

Every figure in this deck traces to the guide's dated register (1/4)

Micron — Micron FQ1 2026 earnings presentation, December 17, 2025. · as-of 2025-12-17 · https://micron.gcs-web.com/static-files/088991c5-a249-4f66-a0a6-258d9b66f3f9

HPE, NVIDIA GB200 NVL72 by HPE data sheet, October 2026 edition. · as-of 2026-10 · https://www.hpe.com/psnow/generateDDS/NVIDIA%20GB200%20NVL72%20by%20HPE%20data%20sheet-PSN1014890104MYEN.pdf

NVIDIA GB300 NVL72 Enterprise Reference Architecture; Lenovo Press LP2357; HPE GB300 NVL72 QuickSpecs v3 · as-of 2026-09 · https://docs.nvidia.com/enterprise-reference-architectures/nvl72-ai-factory-with-gb300-nvl72-dual-plane-networking-architecture.pdf

Pegatron RA4803-72N3 datasheet · as-of 2026-08 · https://svr.pegatroncorp.com/SpecFile/PEGATRON_RA4803-72N3_DS_2026v1.pdf

NVIDIA GTC (Jensen Huang); DCD, Tom's Hardware · as-of H2 2027 roadmap (GTC 2025); 2028 reported July 2026 · https://www.tomshardware.com/pc-components/gpus/nvidia-shows-off-rubin-ultra-with-600-000-watt-kyber-racks-and-infrastructure-coming-in-2027

SOURCES

Every figure in this deck traces to the guide’s dated register (2/4)

Goldman Sachs Global Institute, Tracking Trillions (April 2026), p.5, for the 4–6 yr useful-life estimate; Amazon 2025 Form 10-K, Meta and CoreWeave 2025 Form 10-K for book policies; the 2–3 yr case is the guide's stress assumption · as-of 2026-04 · https://www.goldmansachs.com/pdfs/insights/articles/tracking-trillions-the-assumptions-shaping-scale-of-the-ai-build-out/GS_Apr_GSGI_Tracking_Trillions.pdf

AMD MI455X product specifications (page lists a 2026-07-23 launch date; read 2026-10-19). · as-of 2026-10 · https://www.amd.com/en/products/accelerators/instinct/mi400/mi455x.html

SemiAnalysis InferenceX · as-of 2026-05 · https://inferencex.semianalysis.com/blog/mi355x-glm5-fp8-sglang-40-cheaper-than-b200

Synopsys, AI chip silicon-success and verification guidance — method reference; guide-authored hypothetical scenario September 8, 2026. No supplier quote or test measurement. · as-of 2026-09 · https://www.synopsys.com/blogs/chip-design/ai-chip-silicon-success-tips.html

NVIDIA GB200 specifications and CUDA performance memory model — method reference; guide-authored hypothetical scenario September 8, 2026. No supplier quote or test measurement. · as-of 2026-09 · https://www.nvidia.com/en-us/data-center/gb200-nvl72/

SOURCES

Every figure in this deck traces to the guide's dated register (3/4)

ASML — TWINSCAN NXT:1965Ci specification, maximum exposure field 26 × 33 mm; guide area conversion. · as-of 2026-09 · https://www.asml.com/en/products/duv-lithography-systems/twinscan-nxt-1965ci

UCIe Consortium specification catalogue — method reference; guide-authored hypothetical scenario September 8, 2026. No supplier quote or test measurement. · as-of 2026-09 · https://www.uciexpress.org/specifications

Vicor · as-of 2026-10 · https://www.vicorpower.com/resource-library/articles/high-performance-computing/vertical-power-delivery-enables-cutting-edge-processing

NVIDIA CUDA best-practices profiling guidance — method reference; guide-authored hypothetical scenario September 8, 2026. No supplier quote or test measurement. · as-of 2026-09 · https://docs.nvidia.com/cuda/cuda-c-best-practices-guide/

Guide derivation from a worked example with hypothetical inputs, Chapter 7.9, 2026-09-08 · as-of 2026-09

Guide derivation from a worked example with hypothetical inputs, Chapter 7.10, 2026-09-08; no supplier quote or test measurement. · as-of 2026-09

Guide derivation from a worked example with hypothetical inputs, Chapter 7.11, 2026-09-08; no supplier quote or test measurement. · as-of 2026-09

Texas Instruments, power-management fundamentals — method reference; guide-authored hypothetical scenario September 8, 2026. No supplier quote or test measurement. · as-of 2026-09 · https://www.ti.com/document-viewer/lit/html/SSZTCQ4

SOURCES

Every figure in this deck traces to the guide's dated register (4/4)

Guide-authored hypothetical scenario, September 8, 2026; no supplier quote or test measurement. · as-of 2026-09 · https://www.opencompute.org/wiki/Open_Rack/SpecsAndDesigns

Guide derivation from a worked example with hypothetical inputs, Chapter 7.14, 2026-09-08 · as-of 2026-09

NVIDIA Newsroom (Spectrum-X / xAI Colossus), 2024-10-28, for the 122 days; xAI, Colossus page (x.ai/colossus), for the 92-day doubling (company claim) · as-of 2024-10 (122 days, NVIDIA); 2025-02 (92-day doubling, xAI) · https://nvidianews.nvidia.com/news/spectrum-x-ethernet-networking-xai-colossus

NVIDIA Newsroom, 2024-10-28 · as-of 2024-10 · https://nvidianews.nvidia.com/news/spectrum-x-ethernet-networking-xai-colossus

Supermicro / ServeTheHome, Colossus case study · as-of 2024-10 · https://www.supermicro.com/CaseStudies/Success_Story_xAI_Colossus_Cluster.pdf

Fluke Networks, fiber inspection and testing guidance — method reference; guide-authored hypothetical scenario September 8, 2026. No supplier quote or test measurement. · as-of 2026-09 · https://www.flukenetworks.com/edocs/wp-fiber-cleaning-and-inspection-white-paper

SOURCES

About these figures

Figures are resolved by reference from the guide's claim register at build time (Second edition · September 2026 (first published July 2026)) — a number appears here only as it appears, dated and sourced, at aidatacenterguide.com. Newer data may exist: check the live register.