PART 10 · SECOND EDITION · SEPTEMBER 2026 (FIRST PUBLISHED JULY 2026)

Software, Orchestration & Service Delivery

Parts 1–9 build the machine. Part 10 decides whether it earns: the scheduler, the golden stack, the recovery loop, and the product wrapper that turn a fleet that fails every few hours into goodput you can sell.

THE DEFINITIVE GUIDE TO
AI Data *Centers*

META · LLAMA 3 405B · UP TO 16,384 H100S

1 every ~3 hr (419 in 54 days)

Llama 3 405B: 419 unplanned interruptions in a 54-day window; Table 5 attributes 148 to faulty GPUs and 72 to HBM3.

No facility tier prevents this. Whether it costs you the run is decided in software: detection, cordon, checkpoint, restart. That loop is Part 10.

META · AS-OF 2024-07

THE REFRAME

GPU utilization is the metric that lies. Goodput — the share of GPU-hours doing useful, committed work — is the metric that pays.

A GPU reads 100% busy while spinning on a hung collective or recomputing a corrupted step. Every chapter in this Part is a lever on the gap.

THE SCOREBOARD

90% vs 96% scenario

Illustrative 90%-versus-96% goodput sensitivity — not a population benchmark.

Two illustrative operating points, not measured results. Measure your own fleet's goodput before you price it (Chapters 1.3 and 1.8).

GUIDE ANALYSIS · AS-OF 2026-09-08

ACT I

Who runs where: the scheduling plane

The scheduler decides whether your GPUs run jobs or sit idle. Three decisions dominate: which control plane, whether placement honors topology, and how tenants share a GPU.

THE LAY OF THE LAND

Four control planes, stacked by timescale — bare metal to fleet, each its own loop

Four control loops, stacked by timescale — the scheduler fork sits in the middle

Read bottom-up. Each plane is its own control loop (ideally its own on-call rotation).

FLEET / CONTROL PLANE — continuous

DCGM/NVML + XID/SXID → detect → cordon → drain → hot spare → repair

WORKLOAD-SCHEDULING PLANE — seconds to minutes

Slurm or **K8s + Volcano / KAI** or **Slinky (both)** ← **the fork**

topology-aware: preserve an intact domain when the job requires it;
allow policy-safe backfill when it does not

↑ the scheduler trusts the runtime below: one mismatched NCCL = a collective hang

NODE RUNTIME PLANE — per rollout (hours to days)

GPU driver · CUDA/ROCm · NCCL/RCCL · firmware · container runtime · device plugin / DRA

hard rule: **pinned, attested, compatibility-tested tuple** within each scheduling pool

BARE METAL — provisioned as code — per build-out (weeks)

BMC/Redfish bring-up · PXE · golden images · join & register with the control plane

The architectural test of the whole topology:

a node fails → the fleet fences, cordons and drains it and swaps in a hot spare →
the job resumes with valid work and reconciled allocations — **no human on the critical path.**

Schematic · Source: Guide synthesis · as of October 2026

The test of the whole topology: a node fails and the job resumes with zero humans in the loop. Every fork in this Act is a choice about which plane owns that recovery.

THE MASTER FORK

Slurm, Kubernetes, or both?

Slurm (HPC-native)

— Gang scheduling and batch semantics built in

— Best fit: synchronous pre-training, tightly-coupled jobs

— Lightweight to run; familiar to HPC staff

— Thin at services, autoscaling, rollouts

Kubernetes + batch scheduler

— K8s 1.37 beta gang is opt-in; Volcano/KAI/Run:ai remain options

— Best fit: inference serving and cloud-native pipelines

— Vast ecosystem — GitOps, autoscaling, operators

— Heavier control plane; needs K8s + GPU expertise

Slinky hybrid

— Slurm semantics on a Kubernetes substrate

— Trains and serves side by side on one fleet

— Highest operational cost: you debug both stacks

SO WHAT

Migrating a production fleet later is a re-platforming project, not a config change. Choose for the fleet you will have in three years, not the one you have today.

THE LANDSCAPE IN THREE NUMBERS

Who actually schedules AI fleets

~70% / ~20% / ~10%

2026 trade-press scheduler estimate; no sampled population or selection rule

HPCWIRE, SLURM VS KUBERNETES IN THE AGE OF AI, 15 MAY 2026 · AS-OF 2026-05-15

~90% K8s / ~50% Slurm

CoreWeave-relayed customer observation; no global share or overlap estimate

COREWEAVE, “COREWEAVE RANKS AS #1 AI CLOUD,” 10 APRIL 2025 · AS-OF 2025-04-10

GA in 1.34

Kubernetes DRA graduated to stable — the substrate change for GPU-native scheduling

KUBERNETES V1.34 RELEASE ANNOUNCEMENT, 27 AUGUST 2025: CORE DRA GRADUATES TO STABLE · AS-OF 2025-08-27

VENDOR GRAVITY

Apache-2.0, Apr 2025

NVIDIA now touches every rung: KAI open-sourced, Run:ai absorbed — and Slurm's developer acquired in Dec 2025.

NVIDIA has committed to keeping Slurm open and vendor-neutral. Weigh the concentration anyway — it is now part of your build-vs-buy call on the platform.

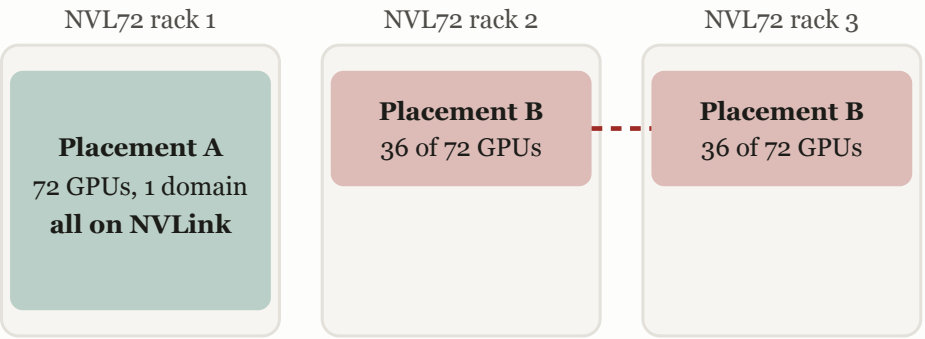
KAI-SCHEDULER PROJECT REPOSITORY · AS-OF 2025-04

THE CONTRACT

One job, two placements: full NVLink bandwidth, or off the cliff

The placement cliff — same job, two placements

An endpoint-bandwidth ratio, not a job slowdown: the cost depends on which collectives cross.



Per-GPU bandwidth, per direction
GB300 NVL72 + 800G ConnectX-8

A · ~0.9 TB/s

B · ~0.1 TB/s

~9× GB300 cliff
~18× GB200 · nominal endpoint ratio

Crossing the domain boundary drops traffic to the 800G scale-out NIC

How schedulers encode the boundary

- **Slurm**: topology/block — one NVL72 = one **18-node block**; a job that fits stays inside one block
- **Kubernetes**: ComputeDomains via DRA on the gpu.clique label; **the deployed driver may allow only one per node**
- **Fragmentation**: two 40-GPU holes cannot serve a TP/EP group that needs all 72 GPUs in one domain

Keep each tightly coupled group (TP/EP) inside one domain; let PP/DP cross.
Price the split against GPUs left idle waiting for a whole domain.

Schematic placements · Source: NVIDIA + DRA Driver for NVIDIA GPUs docs · as of October 2026

Keep TP/EP traffic local when the measured job requires it. Pack compatible smaller jobs only within the tested sharing, isolation and SLO boundary.

THE CLIFF, QUANTIFIED

~18× GB200 / ~9× GB300

Directional endpoint ratios differ by GB200/GB300 profile; they are not job-speed ratios.

Keep jobs whose communication requires scale-up inside that domain. Backfill only when sharing, isolation, fragmentation and SLO rules permit.

GUIDE, ENDPOINT-RATIO ARITHMETIC FROM NVIDIA NCX REFERENCE ARCHITECTURE: NVLINK 5 AND NAMED GB200/CX-7 OR GB300/CX-8 PROFILES. · AS-OF 2026

EXPRESSING TOPOLOGY

Two ways to teach the scheduler the rack

DIMENSION	SLURM BLOCK SCHEDULING	KUBERNETES COMPUTEDOMAINS (DRA)
Topology unit	A block in topology.yaml = one NVLink domain	ComputeDomain over nodes sharing a clique label
Gang scheduling	Built in — all-or-nothing allocation	KAI/Volcano/Kueue, or opt-in 1.37 beta; readiness remains separate
Partial-node sharing	Supported within a node	One ComputeDomain per node in the documented IMEX workflow
Best fit	Large synchronous training, HPC-heritage fleets	Mixed train-and-serve, cloud-native platforms

THE DEFAULT THAT SETS EVERYTHING

Which placement meets the job contract?

Pack compatible jobs

— Keep each required TP/EP group inside its scale-up domain

— Share only within the qualified isolation and SLO boundary

— Reduce reserved GPU-hours when the measured job still passes

Reserve a complete domain

— Use when the rank layout or service guarantee requires it

— Price stranded capacity and tested reclaim time

— Verify the application; link rate alone cannot prove completion

SO WHAT

Compare reserved GPU-hours and completion time for the same job. Packing fails when its communication crosses the measured limit; reserving everything wastes paid capacity.

SHARING IS TWO DECISIONS

CVSS 9.0

NVIDIAScape: a container-to-host escape on shared GPU nodes, in the toolkit nearly every fleet runs.

MIG supplies scoped device isolation. Qualify the host, driver, VM/TEE, network and reassignment boundary for distrustful tenants; a container is not the whole control.

WIZ RESEARCH · AS-OF 2025-07

ACT II

One versioned organism: node, fleet, loop

Driver, CUDA, NCCL, and firmware move together or the job hangs. Provisioning, telemetry, and recovery are a single continuous loop that renews the fleet against failure.

THE HARD RULE

One NCCL version skew between two nodes does not slow a synchronous job. It can hang a collective with no useful error, just a collective waiting forever.

Ship a pinned, attested and compatibility-tested tuple inside each scheduling pool. A heterogeneous fleet may carry multiple approved pools; control cross-pool placement and rollouts explicitly.

THE SECOND-SOURCE QUESTION

AMD/ROCm adds software pools, firmware matrices and support ownership.

Compare matched hardware quotes plus qualification and operating cost; test the intended workload before paying for a second stack.

ILLUSTRATIVE — STATED ASSUMPTIONS

ATP-10A: admit the tested tuple, not a boot flag

Assume Q10, two DGX B200 nodes and a two-hour loaded soak. Synthetic results; OEM temperature limits and the complete receipt remain mandatory.

09:10 — n02 RDMA path fails; preserve the failed test	HOLD
10:00 — repair, correctness and recovery checks pass	Start soak
12:00 — loaded soak completes inside recorded limits	Admit Q10
Missing manifest, OEM limit or ownership evidence	**Keep fenced**

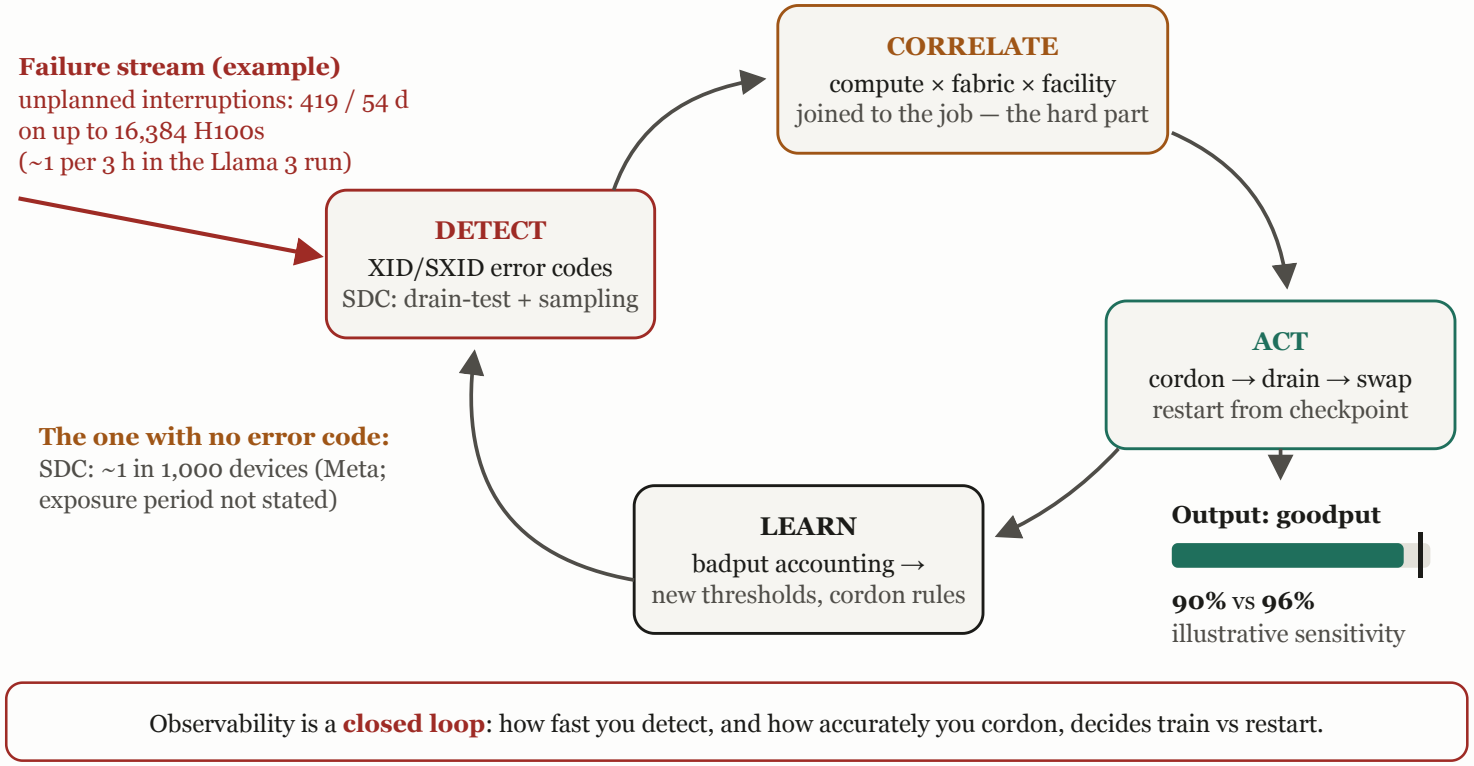
Assumed procedure, not an OEM soak prescription. Recheck active allocation before destructive retry. Chapter 10.5 contains the full trace and flip.

THE CLOSED LOOP

Observability is a loop, not a dashboard: failures in, goodput out

The telemetry loop — failure stream in, goodput out

Detection speed and cordon accuracy decide whether the fleet trains or restarts.



Observability is a **closed loop**: how fast you detect, and how accurately you cordon, decides train vs restart.

Schematic · Reported + illustrative · Source: Meta + guide · as of September 2026

Detection speed and cordon accuracy set the split. Miss a fatal error and a synchronous job hangs on a dead GPU; over-cordon and healthy nodes flap out of the fleet.

THE FAILURE WITH NO ERROR CODE

~1 fault per 1,000 devices; exposure unstated

Meta: about one SDC fault per thousand devices; the exposure period was not stated.

Drain-and-test spends GPU-hours up front; production sampling is probabilistic. Measure coverage and capacity cost for both; doing nothing risks corrupt gradients and days of rollback.

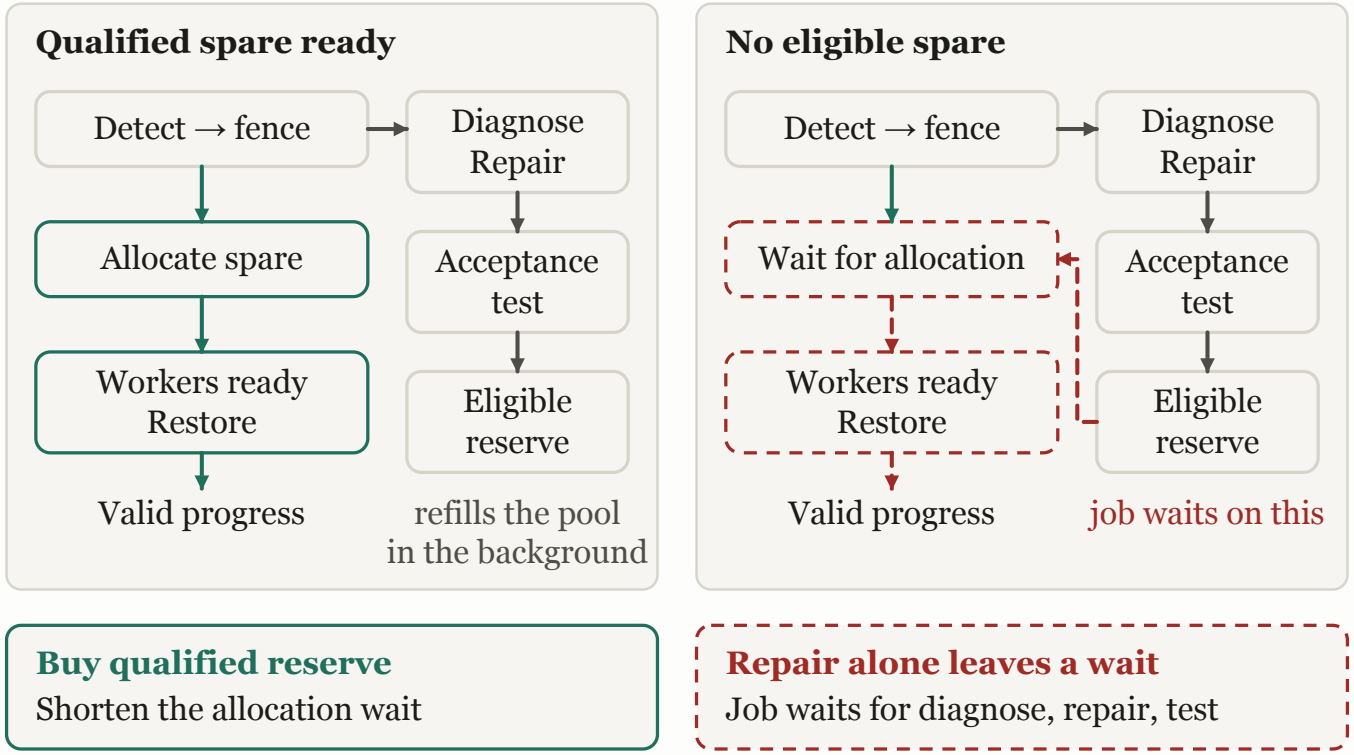
META ENGINEERING, HOW META KEEPS ITS AI HARDWARE RELIABLE, 22 JULY 2025. · AS-OF 2025-07-22

THE MINUTES BUDGET

Detect, eject, swap, restore, resume — every stage runs against a minutes budget

A qualified spare takes repair off the critical path

Illustrative · same failure, two states: arrows show dependencies, not durations



In both states, replaying lost progress follows restore; checkpoint age sets how much.

Schematic · Source: Guide derivation · as of September 2026

Every stage a human touches costs minutes times the whole cluster's GPUs. Automate the full loop and keep escalation as the exception, only for faults the machine cannot classify.

WHERE THE RECOVERY DOLLAR GOES

+6.59% goodput; save <5 min; restore <1 min

Google's thousands-of-nodes restore time omits detection, fencing, allocation, validation.

Spend on the dominant measured recovery stage, including qualified spares. Quarantine lemon nodes until retest; automatic re-admission makes a repeat offender a flapping goodput sink.

GOOGLE CLOUD · AS-OF 2025-06

ACT III

From FLOPS to product

The stack exists to sell something — trained weights, served tokens, or GPU-hours. Frameworks, serving engines, contracts, and data provenance decide the margin on every one.

TRAINING FRAMEWORKS

34% → 54%

BF16 MFU on one 128-H100 cluster, Jan vs Dec 2024: the gain was software, not silicon.

Megatron-Core, FSDP2 and DeepSpeed expose different parallelism and checkpoint paths. Staff and qualify the selected workload; no framework guarantees a one-engineer team.

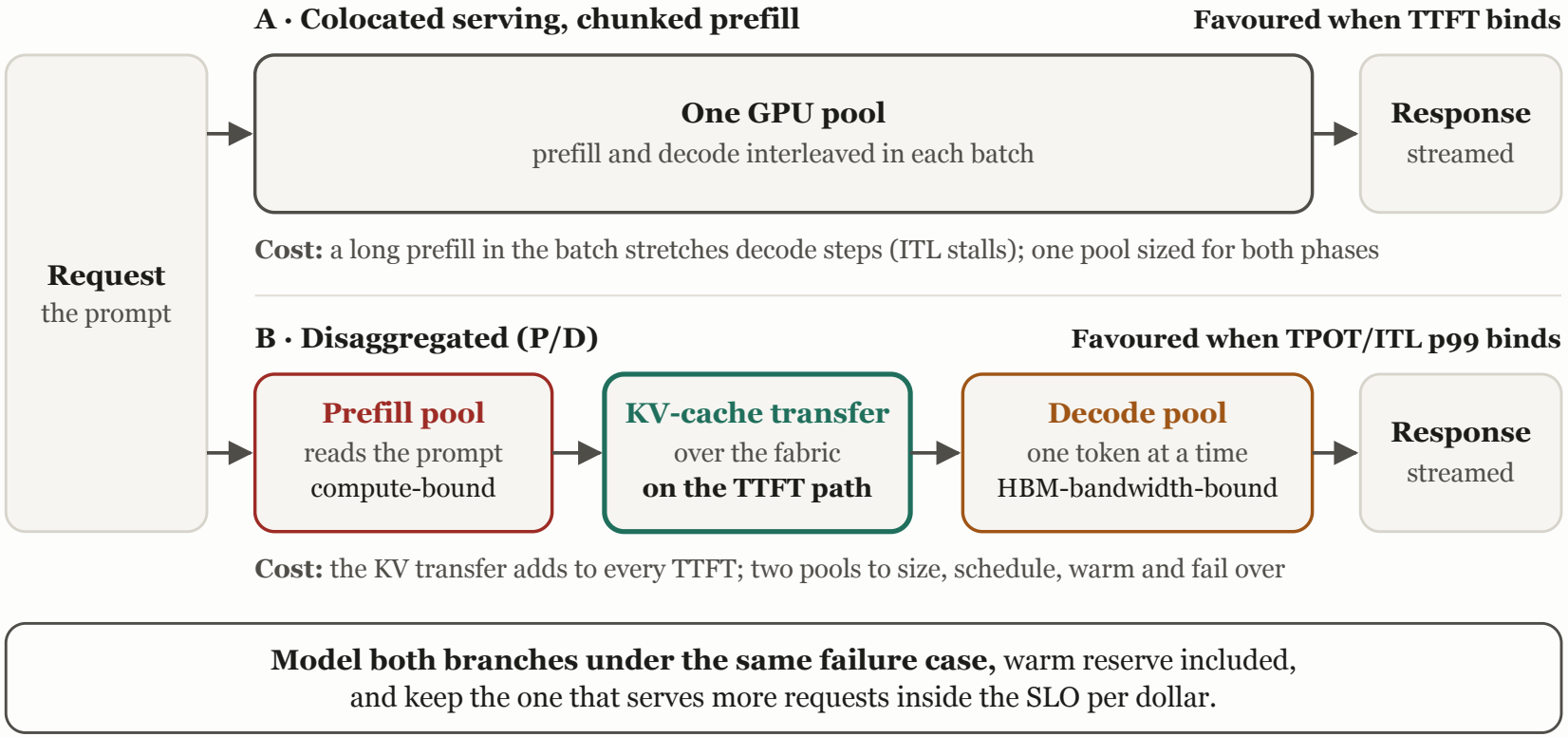
SEMIANALYSIS, H100 VS GB200 NVL72 TRAINING BENCHMARKS, 2025-08-20 · AS-OF 2024-12

INFERENCE SERVING

Prefill and decode stress different hardware — modern serving splits them apart

Prefill and decode: share one pool, or split and pay the KV transfer

The winning branch depends on workload and fabric — above all, on which latency target binds.



Model both branches under the same failure case, warm reserve included, and keep the one that serves more requests inside the SLO per dollar.

Schematic · Source: Guide synthesis · as of October 2026

Separate prefill/decode only when measured latency and full cost beat the aggregated pool, including KV transfer, worker failure, warm reserve and rollback.

WHERE THE COMPUTE GOES

~2/3

Deloitte forecast roughly two-thirds of 2026 AI compute for inference; not a power share.

Count accepted output from requests meeting quality, errors and latency jointly. Keep rejects and partial streams in attainment; price all provisioned capacity, including reserve.

DELOITTE TMT PREDICTIONS 2026, “MORE COMPUTE FOR AI, NOT LESS,” PUBLISHED 2025-11-18. · AS-OF 2025-11-18

THE COMMERCIAL FORK

Capacity vendor or goodput vendor — the ladder decides

RUNG	WHO OWNS GOODPUT	MARGIN POSTURE
Bare-metal	Customer	Price capacity; customer operates the workload
GPUaaS (IaaS)	Shared	Meter the agreed capacity and service boundary
Managed	Operator	Price operating labor, guarantees and credits
Serverless	Operator	Price accepted output plus idle and cold-start risk

THE CONTRACT TRAP

Never sell a goodput SLA you cannot instrument. If your telemetry can't attribute lost time to a root cause, every dispute resolves against you.

Facility availability is the easy promise; effective training time is the one the customer values. Climbing the ladder means carrying that risk — knowingly and measurably.

DATA GOVERNANCE

$1.5B settlement fund

Bartz: $1.5B settlement fund approved on 20 July 2026; not proof of completed distribution.

Record rights, jurisdiction and intended use for each source. A signed data hash cannot grant permission; quarantine unresolved inputs before they reach weights and embeddings.

US DISTRICT COURT, FINAL APPROVAL AND JUDGMENT, DOCUMENT 680, 20 JULY 2026 · AS-OF 2026-07-20

TAKE IT HOME

What to do Monday

01 Put goodput on the wall, not GPU utilization — the metric you display is the one your operators optimize.

02 Choose for the future fleet; reserve intact scale-up domains only for jobs whose measured communication requires them.

03 Pin and attest a compatibility-tested driver/runtime/collective/firmware tuple per pool; admit nodes through measured gates.

04 Automate tested recovery paths; retain fencing, checkpoint validity, qualified spares and escalation.

05 Qualify the complete host, GPU, network and reassignment boundary for each tenant class.

06 Sell the instrumented contract; carry rights and erasure decisions through every derived artifact.

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 10.1 The scheduling plane and the Slurm-vs-Kubernetes fork

Ch 10.2 Topology-aware placement and the bandwidth cliff

Ch 10.7 Fault tolerance and the autonomous recovery loop

Ch 10.11 Inference serving: SLOs, batching, disaggregation

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 (1/3)

Meta (Llama 3 Herd of Models) · as-of 2024-07 · https://arxiv.org/abs/2407.21783

Guide analysis — stipulated sensitivity scenario; no claim of an industry measurement. · as-of 2026-09-08

HPCwire, Slurm vs Kubernetes in the Age of AI, 15 May 2026; denominator not specified. · as-of 2026-05-15 · https://www.hpcwire.com/2026/05/15/slurm-vs-kubernetes-in-the-age-of-ai/

CoreWeave, "CoreWeave Ranks as #1 AI Cloud," 10 April 2025; distinct inference and training customer observations. · as-of 2025-04-10 · https://www.coreweave.com/blog/coreweave-ranks-as-1-ai-cloud-backed-by-semianalysiss-platinum-clustermax-tm-rating

Kubernetes v1.34 release announcement, 27 August 2025: core DRA graduates to stable; subsequent features have separate maturity states. · as-of 2025-08-27 · https://v1-34.docs.kubernetes.io/blog/2025/08/27/kubernetes-v1-34-release/

KAI-Scheduler project repository · as-of 2025-04 · https://github.com/kai-scheduler/KAI-Scheduler

Guide, endpoint-ratio arithmetic from NVIDIA NCX reference architecture: NVLink 5 and named GB200/CX-7 or GB300/CX-8 profiles. · as-of 2026 · https://docs.nvidia.com/ncx/ncp-software-reference-guide/latest/data-center-architecture.html

NVIDIA Developer — Slurm block scheduling on GB200 NVL72 · as-of 2026-05 · https://developer.nvidia.com/blog/achieving-peak-system-and-workload-efficiency-on-nvidia-gb200-nvl72-with-slurm-block-scheduling/

SOURCES

Every figure in this deck traces to the guide's dated register (2/3)

NVIDIA Developer — MNNVL on Kubernetes · as-of 2025-11 · https://developer.nvidia.com/blog/enabling-multi-node-nvlink-on-kubernetes-for-gb200-and-beyond/

DRA Driver for NVIDIA GPUs docs (kubernetes-sigs), v0.5.x · as-of 2026-10 · https://dra-driver-nvidia-gpu.sigs.k8s.io/docs/

Wiz Research · as-of 2025-07 · https://www.wiz.io/blog/nvidia-ai-vulnerability-cve-2025-23266-nvidiascape

Meta Engineering, How Meta keeps its AI hardware reliable, 22 July 2025. · as-of 2025-07-22 · https://engineering.fb.com/2025/07/22/data-infrastructure/how-meta-keeps-its-ai-hardware-reliable/

Google Cloud (multi-tier checkpointing) · as-of 2025-06 · https://cloud.google.com/blog/products/ai-machine-learning/using-multi-tier-checkpointing-for-large-ai-training-jobs

SemiAnalysis, H100 vs GB200 NVL72 Training Benchmarks, 2025-08-20 · as-of 2024-12 · https://newsletter.semianalysis.com/p/h100-vs-gb200-nvl72-training-benchmarks

Deloitte TMT Predictions 2026, "More compute for AI, not less," published 2025-11-18. · as-of 2025-11-18 · https://www.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2026/compute-power-ai.html

SOURCES

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

US District Court, final approval and judgment, Document 680, 20 July 2026; official settlement-administrator court-document copy. · as-of 2026-07-20 · https://assets-us-01.kc-usercontent.com/1eeb16db-4934-006e-40a6-38fa91285ebb/36cce252-a32c-4cbb-a624-31dd8ac9574c/2026-07-20%20Order%20Granting%20Final%20_dckt%20680_0_.pdf

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.