PART 14 · SECOND EDITION · SEPTEMBER 2026 (FIRST PUBLISHED JULY 2026)

Day-2 Operations, Upgrades & Lifecycle

An energized facility can still lose useful work. Meta's 54-day run shows why day 2 funds repeated-fault recovery as well as service continuity.

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

1 every ~3 hr (419 in 54 days)

One unplanned interruption every few hours, for 54 days — and the run still finished.

Most interruptions were attributed to confirmed or suspected hardware issues. The job recovered repeatedly; the paper does not measure building-power availability.

META · AS-OF 2024-07

THE REFRAME

An AI factory needs an available service and accelerators doing useful work on the critical path.

For a fleet with recurring interruptions, restore useful progress before the next fault compounds the loss; retain the service-continuity obligation.

ACT I

The metric that runs the factory

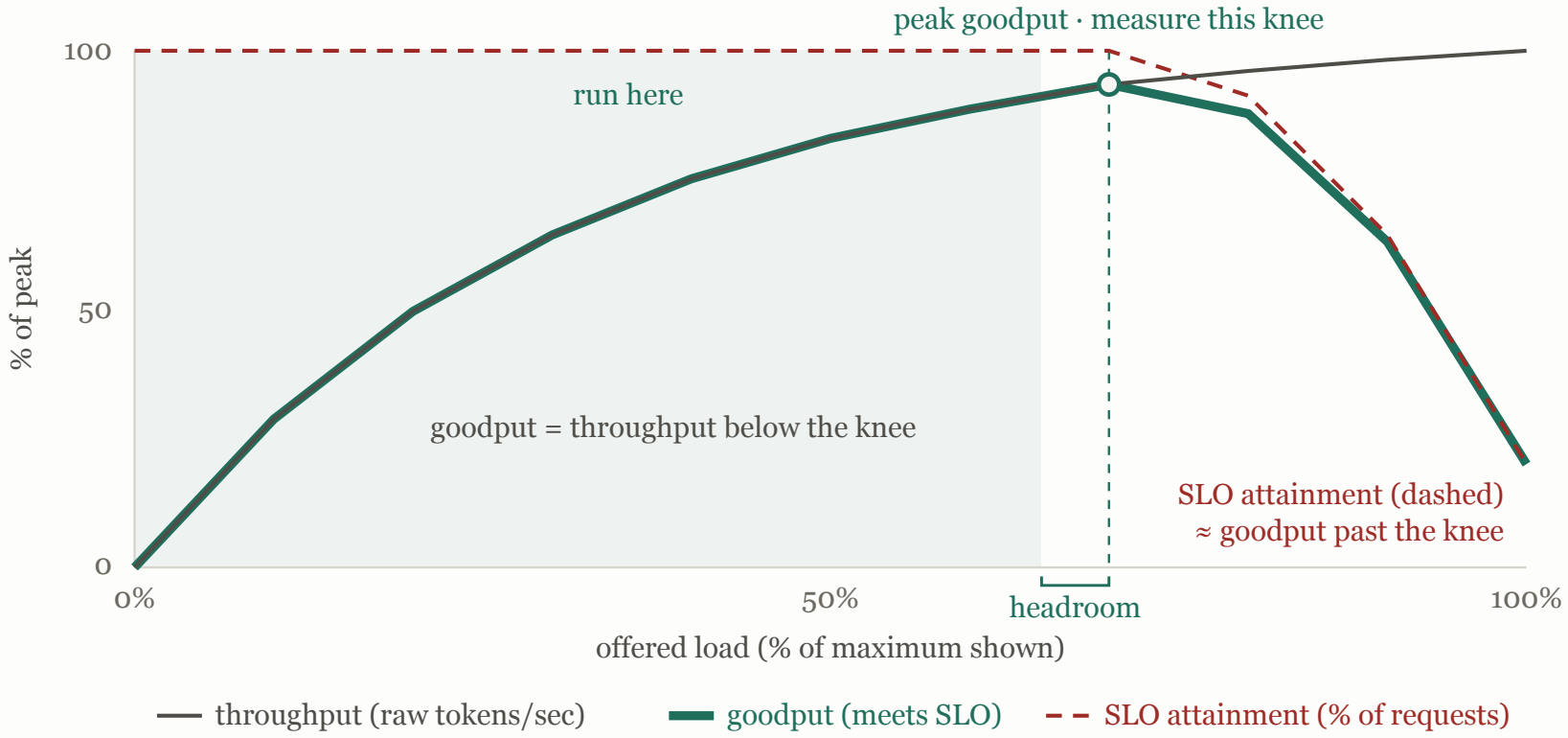
Availability reports the named service boundary. Separate service and progress ledgers expose silent corruption and show where reliability spending recovers usable capacity.

THE CURVE

Useful output peaks below raw throughput

Goodput peaks before throughput

Serving, schematic: useful output (goodput) tracks throughput only until the latency SLO starts to break.



Schematic · Source: Guide synthesis · as of August 2026

Contention, restarts or SLO violations can put a knee in useful output. Measure the workload and service boundary; the schematic is not a universal utilization target.

DECLARED SENSITIVITY

90% vs 96% scenario

Illustrative 90%-versus-96% goodput sensitivity — not an industry baseline

Use these two points only to test operational economics.
Replace them with the named fleet's measured goodput.

GUIDE ANALYSIS · AS-OF 2026-09-08

WHAT THE GAP COSTS

6–21%

Modeled goodput expense: one pretraining scenario, three provider tiers

SemiAnalysis's April 2026 model prices extra rental time or reduced output, including fault-tolerance overhead. Price your own recovery, spare capacity and setup costs.

SEMIANALYSIS, HOW MUCH DO GPU CLUSTERS REALLY COST?, APRIL 20, 2026, SCENARIO 1. · AS-OF 2026-04-20 MODEL; AUGUST 2025 PRICING BASIS

THE MASTER FORK

How do you reconcile service availability and workload goodput?

Service availability

— Measures whether the contracted service is reachable and usable

— Defines maintenance and fault outcomes, transfer interruption, and recovery

— Must include the whole service path and common-mode dependencies

— Does not by itself select N, N+1, 2N, or a Tier topology

Workload goodput / ETTR

— Measures useful work after stalls, rollback, corruption, and recovery

— Values checkpoints, hot spares, health checks, and fast restore

— Quantifies interruption consequence for the named workload

— Does not erase the facility continuity requirement

SO WHAT

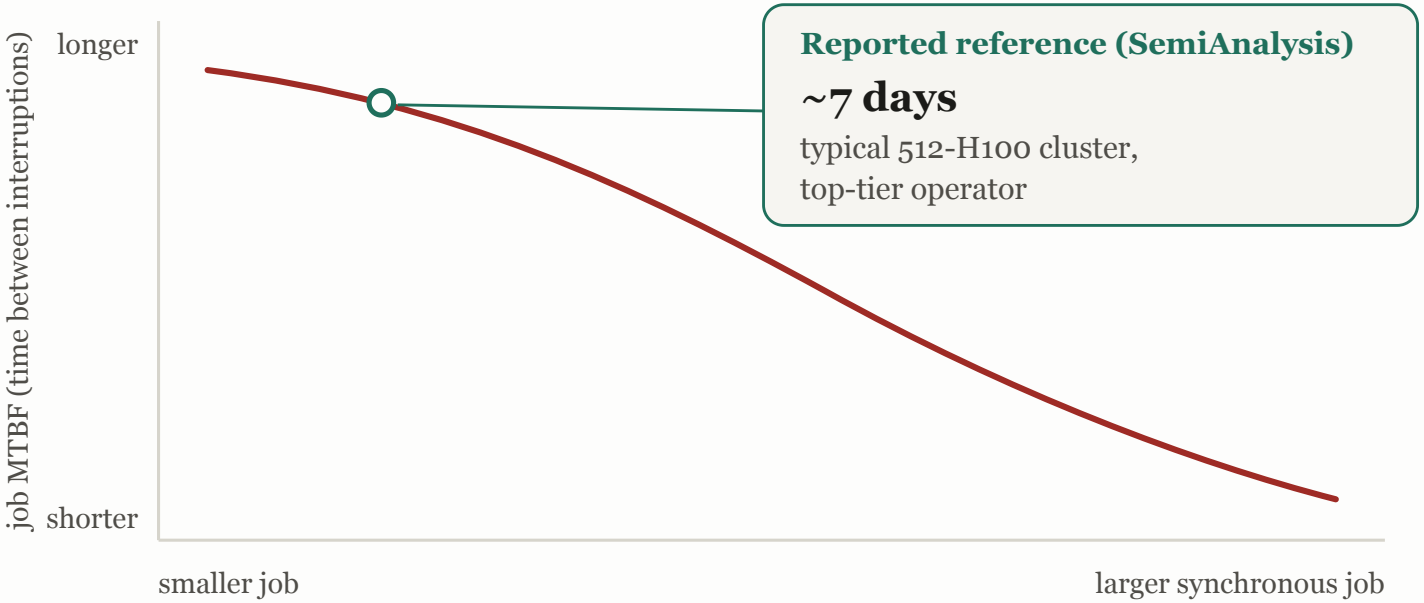
Every project reports both. Their weighting—and the marginal reliability investment—follows the declared maintenance, fault, recovery, service, workload, and contract model.

THE SCALE LAW

Job MTBF falls with exposure — calibrate the curve, do not extrapolate it

Bigger synchronous jobs get interrupted more often

Schematic: more failure domains, more exposure. The ~7-day anchor is not a per-GPU rate.



Measure your own fleet; do not extrapolate the anchor
(correlation and software change the shape).
Budget for fast detection, checkpointing and recovery.

Schematic · Source: SemiAnalysis (Oct 2024) · as of October 2026

Larger jobs expose more failure domains, but topology, software, correlation, maturity, and event definitions move the curve. Fit fleet telemetry, then budget detection, checkpointing, and recovery.

THE FAILURE WITH NO SIGNATURE

~1 fault per 1,000 devices; exposure unstated

Meta CPU-fleet SDC evidence: about one fault per thousand devices; exposure unstated

CPU study rates do not establish GPU/HBM incidence. Escaped corruption can surface as a loss spike days into a multi-week run, forcing progress to be replayed.

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

META'S LEMON-NODE INTERVENTION

14% → 4%

Large-job failure rate, before and after ejecting the fleet's 'lemon' nodes

A sliver of the footprint touched an outsized share of jobs. Meta ejected repeat offenders proactively — before they had proven themselves bad — and job completion jumped.

META PAPER. · AS-OF 2024-10

ACT II

Recover faster than you fail

Every minute of detect-to-recover comes straight out of goodput. Act II is the machinery that shrinks it: checkpoint tiers, local spares, predictive maintenance, and firmware discipline.

THE RECOVERY STACK

Checkpoint tiers trade recovery coverage against capacity and reload cost

TIER	RECOVERY LATENCY	WHAT IT SURVIVES
In-memory (host RAM)	Seconds — no network reload	Transient process or GPU error; same-node restart
Peer (in-cluster replica)	Tens of seconds — intra-fabric copy	Single-node loss without a durable reload
Durable (parallel FS / object store)	Minutes — full reload across the fabric	Whole-job, rack, or facility event
Async drain (overlay)	N/A — hides the write cost in background	Makes the durable-tier cadence affordable

A COMPATIBLE SPARE CAN SHORTEN RESTORATION; REPLENISHMENT SETS STOCK DEPTH

What recovery speed is worth

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

Checkpoint restore, one clock inside MTTR: durable storage vs multi-tier (Google-reported)

GOOGLE CLOUD · AS-OF 2025-06

148 GPU + 72 HBM interruptions

GPU/HBM interruptions in one Llama 3 run on up to 16,384 H100s — not equipment AFR

META AI, THE LLAMA 3 HERD OF MODELS, §3.3.4/TABLE 5 · AS-OF 2024-07

5% → 2%

Delta model: faster recovery cuts synchronous-job overprovisioning

CUI ET AL., STORY OF TWO GPUS: CHARACTERIZING THE RESILIENCE OF HOPPER H100 AND AMPERE A100 GPUS, SC '25 · AS-OF 2025-12

THE PLANT UNDERNEATH

Four maintenance rungs — select per asset from consequence and detectability

Pick the maintenance rung per asset

The top two rungs pay only if the failure gives warning.

more sensing, data, setup

Predictive	Schedule from a life forecast if warning ≥ lead time + error
Condition-based	Act on a measured threshold if a sensor sees degradation
Preventive	Service on a calendar if code or warranty says so
Run-to-failure	Fix it when it breaks if the loss is accepted

Detectable degradation

Abrupt or undetectable failure: prove redundancy, spares and access

> Condition-based example: correlate coolant inlet and flow with lost GPU output.

Bad sensors or short warning turn a forecast into lost service.
Commissioned baselines and maintainable paths make intervention safe.

After a failure, four clocks run separately and can overlap:

- Detect / isolate
- Restore service
- Repair unit
- Replenish stock

Schematic · Source: Guide synthesis · as of September 2026

Keep detection/isolation, service restoration, repair and replenishment as separate clocks. A transformer lead time above 160 weeks is not outage duration or a predictive-maintenance rule.

FILLING THE POWER CEILING

~3% training / ~21% inference

POLCA headroom depends on workload overlap and arrivals

POLCA observed ~3% training and ~21% inference headroom. Correlated arrivals and phase overlap can consume it and trip an overloaded path; qualify caps and independent protection.

MICROSOFT RESEARCH · AS-OF 2024-04

FIRMWARE AT FLEET SCALE

A web fleet tolerates a drifted node. A synchronous training job does not — every step blocks on the slowest rank.

A firmware regression on one rank can slow the collective. Qualify compatibility tuples; gate canaries and later waves on measured behavior and full recovery time.

WHY THE WINDOW MATTERS

~$12–13B/GW-yr

A revenue-per-GW scenario; price maintenance from lost contribution

SemiAnalysis supplies one annual revenue scenario. Price the full reset, retraining and readmission window from lost output, demand and marginal contribution.

SEMIANALYSIS · AS-OF 2026-06

ACT III

The human layer

Operators, procedures and handoffs can amplify a fault. Staff the qualified roles, then decide command authority before a cross-domain incident.

UPTIME INSTITUTE, 2022 AND 2025

~85%

Human-error outages involving failure to follow or flaws in procedures

Roughly 85% applies to the human-error-outage population, not to all data-center outages. Uptime pairs better procedures with staff training and real-time operational support.

UPTIME INSTITUTE ANNUAL OUTAGE ANALYSIS 2025 · AS-OF 2025

AFTER THE INCIDENT

Redesign the error trap and train the changed procedure

Blame and reminders only

— Write-up or a reminder without changing the work

— The wrong label or ambiguous line-up remains

— The next operator meets the same trap

— No evidence that the error path is closed

Redesign the trap

— Fix the ambiguous line-up, wrong label or MOP gap

— Qualified independent checks on critical steps

— Targeted training on the changed procedure

— Drill the expected state, HOLD and restoration gates

SO WHAT

Match SOP, MOP and switching authority to the action. CAB review is one part of engineering MOC; neither authorizes an ambiguous isolation.

HANDING OVER THE KEYS

The autonomy ladder: promote loops, not facilities

RUNG	WHAT THE AGENT DOES	HUMAN POSTURE
0 — Observe	Ingests telemetry, flags anomalies	Agent has no authority; established local controls remain active
1 — Recommend	Proposes action; human executes	In the loop — approves every action
2 — Act within envelope	Executes inside a hard-bounded action space	On the loop — supervises, can intervene
3 — Fully autonomous	Owns the loop, escalates on its own judgment	Out of the loop for routine action

THE RESULT THAT STARTED THE CLIMB

~40%

Cooling-energy reduction in DeepMind's 2016 live test of ML-recommended settings

Google's historical cooling result does not confer actuator authority. Qualify each loop's constraints and fallback; automatic protection and transfer retain their separate role.

DEEPMIND · AS-OF 2016

ACT IV

The lifecycle bill comes due

Silicon turns over in a few years; the building lasts decades. Act IV is where the depreciation forecast gets settled in cash — and where the proof your commissioning bought quietly decays.

THE CONTESTED NUMBER

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

GPU accounting book life vs frontier economic life

Contested: hyperscalers moved depreciation lives in opposite directions. Refresh execution (cascade, resell, destroy only media that fail validation) turns the forecast into cash or a write-down.

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

END OF LIFE, POWER-BOUND ERA

An energized shell can preserve years of interconnection work, if its rights and physical envelope support the next use.

Price repower, conversion and demolition against retained grid rights, structure, cooling and demand. Resolve lease, permit and restoration obligations before selecting the end state.

WHY COMMISSIONING NEVER ENDS

132 kW → ~600 kW

HPE GB200 132 kW rack TDP to Kyber ~600 kW roadmap figure, boundary unpublished — a density step re-opens commissioning

Kyber-class density is a roadmap input. Retain unchanged evidence; retest isolated affected blocks on drift and before admitting a load outside the proved envelope.

NVIDIA · AS-OF HPE GB200 QUICKSPECS 2026-09-08; NVIDIA KYBER TARGET H2 2027 (CONTESTED)

TAKE IT HOME

What to do Monday

01 Report both service availability and workload goodput/ETTR; define their denominators before writing the SLA.

02 Spend the marginal reliability dollar where the named service, fault, recovery, and workload model returns the required outcome.

03 Keep CPU evidence scoped; fund SDC detection and checkpoint validation against the costly progress late corruption can destroy.

04 Derive each fabric tier from measured traffic, placement, failure headroom, and the governing SLO—not workload labels.

05 Set canary width and tested rollback/replacement scope; gate later waves and name the security/emergency exception authority.

06 Redesign the error trap, train and drill it; retain unchanged commissioning proof and retest changed behavior.

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 14.1 Goodput, ETTR, and the reliability economics of AI factories

Ch 14.4 Checkpoint tiering, lemon nodes, and the detect-to-recover loop

Ch 14.8 Firmware campaigns: canary width, rollback, and supply-chain integrity

Ch 14.9 Refresh triggers, depreciation, sanitization, and residual value

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

SemiAnalysis, How Much Do GPU Clusters Really Cost?, April 20, 2026, Scenario 1. · as-of 2026-04-20 model; August 2025 pricing basis · https://newsletter.semianalysis.com/p/how-much-do-gpu-clusters-really-cost

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/

Meta paper. · as-of 2024-10 · https://arxiv.org/abs/2410.21680

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

Meta AI, The Llama 3 Herd of Models, §3.3.4/Table 5 · as-of 2024-07 · https://arxiv.org/html/2407.21783

Cui et al., Story of Two GPUs: Characterizing the Resilience of Hopper H100 and Ampere A100 GPUs, SC '25 (arXiv 2503.11901 v4) · as-of 2025-12 · https://arxiv.org/abs/2503.11901

SOURCES

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

Microsoft Research / ACM, Patel et al., Characterizing Power Management Opportunities for LLMs in the Cloud, ASPLOS, April 2024; publisher page gives month precision. · as-of 2024-04 · https://www.microsoft.com/en-us/research/publication/characterizing-power-management-opportunities-for-llms-in-the-cloud/

SemiAnalysis ('To Boldly Go', Jun 2026) · as-of 2026-06 · https://newsletter.semianalysis.com/p/to-boldly-go-the-case-for-space-datacenters

Uptime Institute Annual Outage Analysis 2025 · as-of 2025 · https://uptimeinstitute.com/about-ui/press-releases/uptime-announces-annual-outage-analysis-report-2025

DeepMind / Google · as-of 2016 · https://deepmind.google/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40/

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

SOURCES

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

NVIDIA / Tom's Hardware / The Register (CES 2026, GTC) · as-of HPE GB200 QuickSpecs 2026-09-08; NVIDIA Kyber target H2 2027 (contested) · https://www.tomshardware.com/pc-components/gpus/nvidia-shows-off-rubin-ultra-with-600-000-watt-kyber-racks-and-infrastructure-coming-in-2027

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.