PART 12 · SECOND EDITION · SEPTEMBER 2026 (FIRST PUBLISHED JULY 2026)

Reliability, Resilience & Standards

The industry inherited a sixty-year discipline for keeping buildings up. The AI cluster earns its return on something no Tier certifies: the fraction of bought GPU-hours that become useful work.

THE DEFINITIVE GUIDE TO
AI Data *Centers*

LLAMA 3 405B · UP TO 16,384 H100S · META, 2024

1 every ~3 hr (419 in 54 days)

About one unexpected interruption every three hours, mostly hardware — facility availability was not reported.

The run still lost compute despite over 90% effective training time. The paper measures job interruptions, not building availability; this Part keeps those metrics separate.

META · AS-OF 2024-07

THE REFRAME

The reliability discipline we inherited rates the building. The AI cluster earns its return on the job.

Tiers, Rated levels, and Classes certify power and cooling topology. None of them has a clause for a synchronous 16k-GPU run, a CDU's blast radius, or a multi-megawatt load swing.

ACT I

The standards stop at the meter

Uptime Tiers, TIA-942, EN 50600: what each level actually guarantees, what each rung of redundancy costs — and why the two properties they bundle should be bought separately.

DIFFERENT STANDARDS, DIFFERENT SCOPES

Compare required outcomes — do not equate the labels

STANDARD FAMILY	CLASSIFICATION	PRIMARY SCOPE	CONTRACTING RULE
Uptime Institute	Tier I–IV	Topology and operational outcomes	Specify required capacity, maintenance and fault states; no availability percentage is certified
ANSI/TIA-942-C	Rated-1–4	Telecom, architectural, electrical and mechanical facility requirements	Verify the selected Rated requirements; do not infer equivalence from the numeral
EN 50600 / ISO/IEC 22237	Availability Class 1–4	Facility infrastructure, availability and management framework	Verify the applicable part, class and national adoption
Project service objective	Named maintenance and fault cases	End-to-end service path, transfer, controls, loading and workload recovery	Test the actual design state by state; a topology name alone proves none of these

THE DECISION INSIDE THE TIER

Buy the property, not the Tier

Concurrent maintainability

— Service any component or path without dropping load

— Define which component and path maintenance states must keep the contracted load

— Verify transfer interruption, post-maintenance loading and control dependencies

— Select it where the service objective and maintenance model require live work

Fault tolerance

— Survives the explicitly defined unplanned fault case under the standard or contract conditions

— Verify post-event loading, dual-cord paths, controls, common modes and recovery

— Select it when the service objective values that fault outcome more than alternatives

SO WHAT

Specify and test the maintenance and fault states per subsystem. Workload recovery economics are one input; a workload name does not output a topology.

THE PRICE OF BUNDLING

~25–40% facility

Whole-facility planning premium for Tier IV fault tolerance over Tier III concurrent maintainability

A dated, single-source planning heuristic — not a quote. Re-estimate the actual design at equal usable load and verify which maintenance and fault states the added topology supports.

SAVILLS RESEARCH, EUROPEAN DATA CENTRES, MAY 2024 · AS-OF 2024-05

PRACTITIONER CAPEX RANGES — ILLUSTRATIVE, UNREFERENCED

The redundancy ladder: what each rung costs and strands

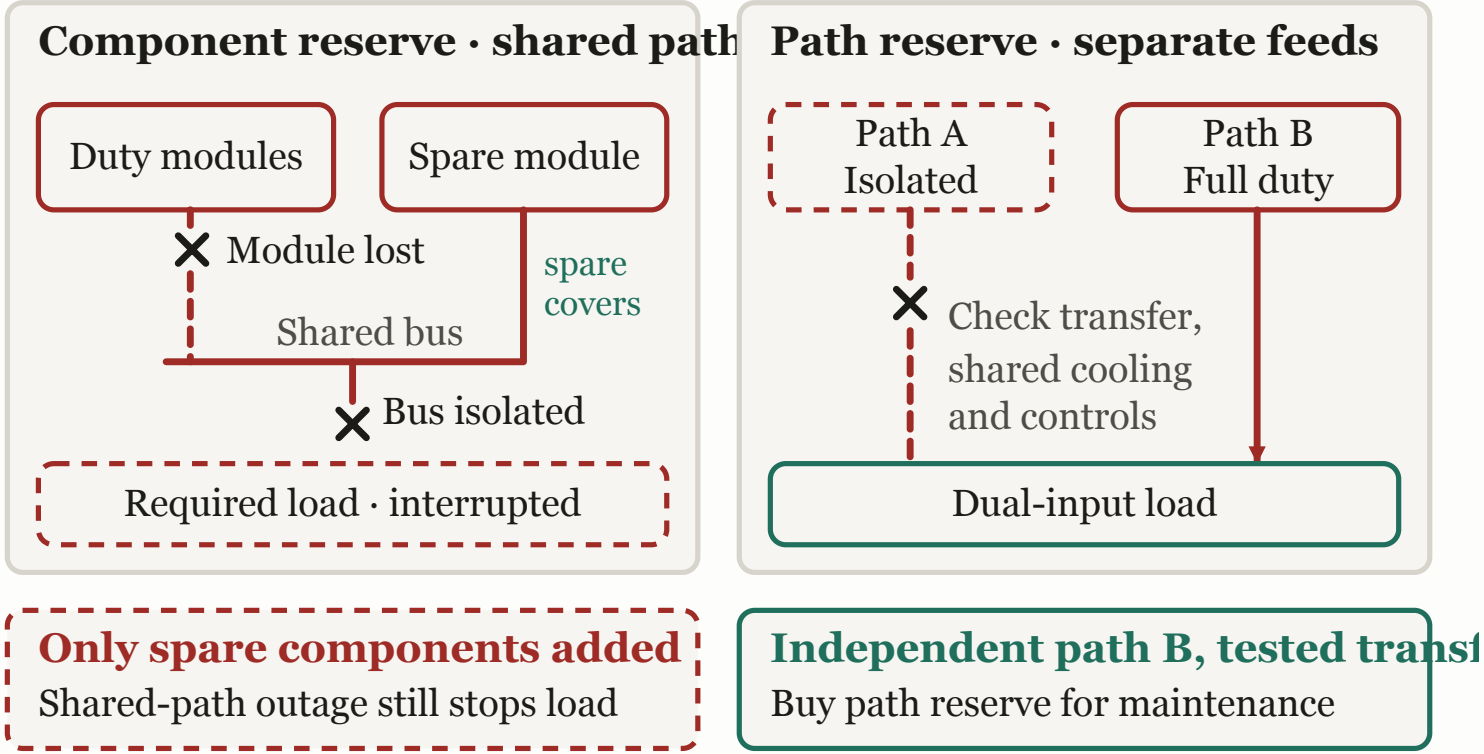
TOPOLOGY	RELATIVE MEP CAPEX	STRANDED CAPACITY	SURVIVES
N	Baseline (1.0x)	~0%	Nothing — any failure or maintenance event drops capacity
N+1	~1.1–1.2x	Small (~1/N)	One component failure OR one maintenance event
2N full mirror	~1.5x	~50%	Any single system fault plus concurrent maintenance
Distributed 3N/2, 4N/3	~1.2–1.35x	~25–33%	One block fault — survivors absorb its share; the 2026 hyperscale default
2(N+1)	~1.6x+	>50%	Multiple faults across both halves — highest-criticality only

THE LADDER, DRAWN

Draw one failure into each topology: what survives, and what you paid to strand

A spare component is not a second path

Illustrative power package · same required load, isolated maintenance



Schematic · Source: Guide derivation · as of Sep 2026

2N mirrors paths so one can carry the load. 3N/2 shares capacity across three half-size systems; verify transfer, post-event loading, shared controls and recovery. Labels do not prove outcomes.

UPTIME INSTITUTE, 2025

What actually takes facilities down

45%

Impactful outages caused by power — most often the UPS bought to prevent them

UPTIME INSTITUTE, ANNUAL OUTAGE ANALYSIS 2026, FIGURE 4 · AS-OF 2025

58%

Human-error outages from staff not following procedures — and rising

UPTIME INSTITUTE ANNUAL OUTAGE ANALYSIS 2025 · AS-OF 2025

ACT II

Goodput: the number the money follows

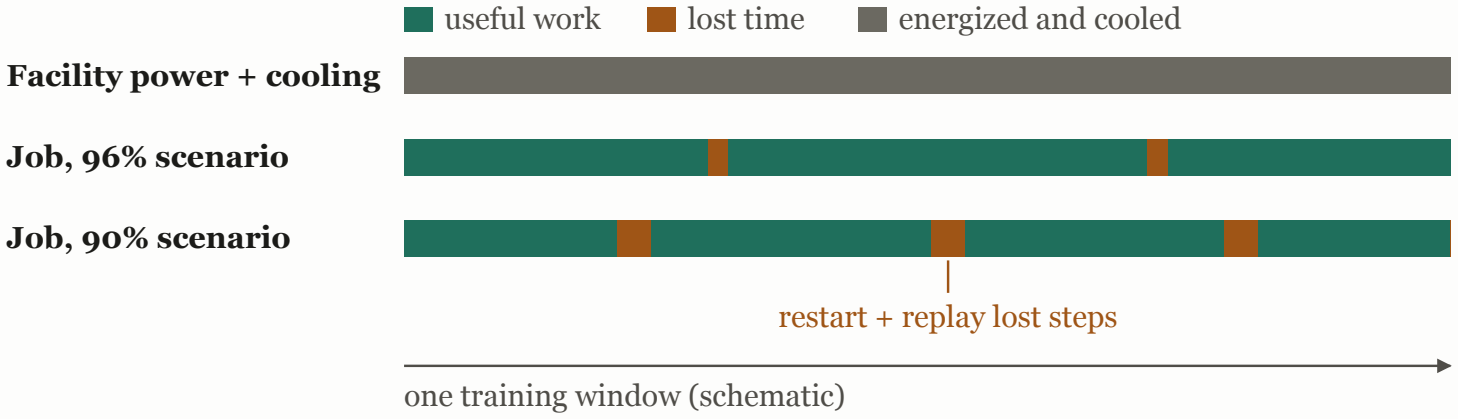
Availability asks whether the building is up. Goodput asks whether the money is working. This act follows the metric — where redundancy actually lives, why the grid now cares, and how to model every nine you claim.

TWO METRICS

You can lose a tenth of your compute in a Tier IV hall that never blinks

The facility never blinked; the job still lost time

Schematic timeline · 90% and 96% are guide scenarios, not measurements



Why gaps happen

Meta: 419 unplanned interruptions in 54 days, on up to 16,384 H100s

What shrinks them

Google, one large TPU v5p job:
ML Goodput +6.59%
Checkpoint restore <1 min across thousands of nodes (not full job recovery)

Measure the job's loss before buying added plant redundancy.

Schematic + measured · Source: Meta (2024) + Google Cloud (2025) + Guide derivation · as of September 2026

Availability is binary up/down at the meter. Goodput is continuous leakage below a perfect ceiling — and no facility certification measures it.

THE FAILURE ENVIRONMENT

The building was never the problem

90% vs 96% scenario

Illustrative 90%-versus-96% goodput sensitivity — replace with measured fleet results

GUIDE ANALYSIS · AS-OF 2026-09-08

~7 days (512-H100 cluster, typical)

Typical MTBF SemiAnalysis reported for a 512-H100 cluster — an availability risk

SEMIANALYSIS, AI NEOCLOUD PLAYBOOK AND ANATOMY · AS-OF 2024-10

~43.4%

Failure rate of the most resource-intensive large-LLM jobs (Alibaba)

UNICRON · AS-OF 2023-12

THE RECURRING DECISION IN AN AI FACTORY

Where the next redundancy dollar actually lives

REDUNDANCY SPEND	LAYER	RESTART PERMITTED WITHIN THE RTO	NO LOAD LOSS THROUGH THE DEFINED EVENT
N+1 → 2N facility power	Facility	Value depends on interruption cost and recovery target; restart tolerance does not select topology	A candidate only if path, control, common-mode, transfer, and post-event loading evidence close
Multi-tier async checkpointing	Storage + software	Directly reduces rollback and recovery badput	Reduces state-loss consequence where the service has recoverable state; it does not prevent interruption
Hot-spare GPUs + fast drain	Silicon + orchestration	Shortens repair-to-resume when orchestration and data placement are proven	Supports capacity continuity, but cannot replace independent power, cooling, network, and control paths
Per-GPU capacitance + rack BBU (roadmap figures)	Silicon + facility	Can avoid a restart for the qualified transient envelope	Can preserve service only for the qualified duration and complete downstream transfer chain

RIDE-THROUGH CUTS BOTH WAYS

~1,500 MW

Data-center load dropped during 82 seconds of faults — later cited in NERC's May 2026 Level 3 alert

Trip-to-protect turns a two-cycle sag into a full cluster restart and a grid event. Riding through is a goodput decision and a grid-citizenship decision at once.

NERC, INCIDENT REVIEW: LARGE LOAD LOSS · AS-OF 2024-07

PICK BY THE QUESTION, ESCALATE AS ASSUMPTIONS BREAK

Four models, one escalation ladder

METHOD	ANSWERS	BREAKS WHEN
RBD / k-of-n algebra	Steady-state availability of a fixed topology	Repair is shared or failures correlate
Markov state-space	Degraded states and finite repair crews	Transitions aren't memoryless
Fault tree / minimal cut sets	The dominant failure path and its top contributors	Failures are sequence-dependent
Monte-Carlo simulation	Availability AND goodput under a realistic environment	The input rates are themselves wrong

INDEPENDENCE IS THE ASSUMPTION THAT LIES

0.5–5% logic; 1–10% sensors/final elements

The IEC 61508 beta-factor range — the share of faults your 'independent' paths actually share

A shared CDU, bus, or firmware image erases the parallel product. An idealized six-nines block collapses toward four once a realistic beta is applied.

IEC 61508-6:2010 ANNEX D · AS-OF 2010-04-30

META, 2024 — SENSITIVITY IS THE DELIVERABLE

What the model ranks first, proven in production

1.8 hr → 14 min

Projected job MTTF, 16k → 131k GPUs — it falls roughly inversely with scale

META A100 RESEARCH-CLUSTER MODEL. THIS IS A HYPOTHETICAL SCALE RESULT, NOT OBSERVED H100 BEHAVIOR OR A DATED DEPLOYMENT FORECAST. · AS-OF 2024-10

≈10 s checkpoint writes, or failure rate ≈1 (vs 6.50) per 1,000 node-days

ETTR 0.9 on a ~12,000-GPU run needs ~10 s checkpoint writes or far fewer failures

META, REVISITING RELIABILITY IN LARGE-SCALE ML RESEARCH CLUSTERS, ARXIV 2410.21680V2 · AS-OF 2025-02

14% → 4%

Large-job failure rate after ejecting lemon nodes from the pool

META PAPER. · AS-OF 2024-10

ACT III

Continuity, contracts, and proof

The rethink becomes money on two documents: the disaster-recovery plan that pre-pays for spare capacity, and the SLA that promises a number. Both fail when written against the wrong metric.

2026 CONVENTIONS — COST COLUMN ILLUSTRATIVE, UNREFERENCED

Five failover postures — and the power-bound question

POSTURE	REALISTIC RTO	STEADY COST VS ONE REGION	ENERGIZED IN ADVANCE?
Active-active	Seconds — load-balancer drain	~2x (a second live fleet)	Yes — both regions full
Hot / warm standby	Minutes	~1.4x	Yes — racked and powered
Pilot light	~1–4 hr scale-up	~0.2x of full redundancy	Partially — depends on slack that may not exist
Cold standby	Hours to days	Storage + code only	No — exposed to capacity-acquisition risk
Backup / restore only	Days+	Replication storage only	No — not viable for inference SLAs

TWO CONTINUITY FACTS THAT SET THE PLAN

Why the spare region is the hardest asset to fake

~14.5 hr (14 h 32 min)

AWS US-EAST-1 outage, Oct 2025 — one region's fault reached every service anchored there

AWS POST-EVENT SUMMARY · AS-OF 2025-10

about 220 Gbit/s payload

One-hour RPO for ~100 TB needs ~222 Gbps of replication — before failover capacity

AWS SUPPLIES RECOVERY-PATTERN BOUNDARIES ONLY. GUIDE UNIT CONVERSION FROM AN ASSUMED 100 TB DECIMAL COPY AND ONE-HOUR WINDOW · AS-OF 2026-09

THE MASTER SLA FORK

Availability commitment, or goodput contract?

Availability commitment

— Measured at the meter or the endpoint — trivially auditable

— Right for online inference, hosting, and the control plane

— Breached by power loss, network partition, node-down

— Defended with redundancy: dual power paths, spare cooling

Goodput contract

— Measured inside the job: telemetry, checkpoint logs, step counters

— Right for pre-training, large-scale RL, long batch

— Breached by SDC, stragglers, slow recovery, fabric errors

— Defended with health checks, hot spares, fast restore

SO WHAT

Most 2026 GPU contracts layer both. Promising facility availability to a customer who is actually paying for goodput is unenforceable for them and unprofitable for you.

BEFORE THE HEADLINE NUMBER, READ THE EXCLUSIONS

What the 2026 market actually signs

99% / 95%

Example GPU-cloud SLA: node / rack uptime with penalties (ClusterMAX 2.0, Nov 2025)

SEMIANALYSIS CLUSTERMAX · AS-OF 2025-11

10% / 30% / 100%

Reference service-credit ladder — % of monthly bill as uptime falls

AMAZON COMPUTE SLA, LAST UPDATED MAY 25, 2022. · AS-OF 2022-05-25

6–21%

Modeled goodput expense: one pretraining scenario, three provider tiers

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

THE THROUGH-LINE

An availability number you cannot reconstruct from failure rates and repair times is a marketing claim. The SLA is downstream of the model — never upstream of it.

TAKE IT HOME

What to do Monday

01 Define maintenance/fault states, transfer interruption, post-event loading, common modes and workload consequence before topology.

02 Draw fault domains on purpose — the blast radius of the worst single failure is a design output, not a discovery.

03 Rank recovery, cooling, ride-through, and path redundancy against the named service/failure model; no order is universal.

04 Treat regions as correlated, not independent — decouple the single-region control plane before buying a second fleet.

05 Energize failover capacity years in advance; in a power-bound market you cannot rent your way out on the day.

06 Write the SLA from the model: design basis → failure environment → modeled number → promised number with margin.

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 12.1 Standards crosswalk, redundancy topologies, fault domains

Ch 12.2 The goodput rethink and where redundancy lives

Ch 12.3 Failover postures and the spare-region problem

Ch 12.4 Availability commitments vs goodput contracts

Ch 12.5 RBD, Markov, fault trees, Monte-Carlo — the model

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

Savills Research, European Data Centres, May 2024; publisher report. Quoted secondary estimate; not measured project evidence. · as-of 2024-05 · https://pdf.euro.savills.co.uk/european/european-commercial-markets/spotlight-eu-data-centre---may-2024---final.pdf

Uptime Institute, Annual Outage Analysis 2026, Figure 4 (n=96) — 45% is the 2025 outage year, restated there and down from 54% in 2024. · as-of 2025 · https://datacenter.uptimeinstitute.com/rs/711-RIA-145/images/2026.AnnualOutageAnalysis.pdf

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

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

SemiAnalysis, AI Neocloud Playbook and Anatomy · as-of 2024-10 · https://newsletter.semianalysis.com/p/ai-neocloud-playbook-and-anatomy

Unicron (He et al., arXiv 2401.00134) · as-of 2023-12 · https://arxiv.org/abs/2401.00134

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

SOURCES

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

NVIDIA technical announcement, 16 March 2026. Vendor design statement, not an independent field test. · as-of 2026-03 · https://developer.nvidia.com/blog/nvidia-vera-rubin-pod-seven-chips-five-rack-scale-systems-one-ai-supercomputer/

NERC, Incident Review: Large Load Loss (2025-01-08), pp.7–8; NERC comments to FERC, Docket EL25-49 (2025-04-23), Attachment A (April 17 presentation), PDF p.17 · as-of 2024-07 · https://www.nerc.com/globalassets/our-work/reports/event-reports/incident_review_large_load_loss.pdf

IEC 61508-6:2010 Annex D (Table D.4 score bands); exida guidance on the β ranges · as-of 2010-04-30 · https://www.exida.com/blog/the-importance-of-considering-common-cause-with-redundant-systems

Meta A100 research-cluster model. This is a hypothetical scale result, not observed H100 behavior or a dated deployment forecast. · as-of 2024-10 · https://arxiv.org/abs/2410.21680

Meta, Revisiting Reliability in Large-Scale ML Research Clusters, arXiv 2410.21680v2 (2025-02-06) · as-of 2025-02 · https://arxiv.org/abs/2410.21680v2

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

SOURCES

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

AWS Prescriptive Guidance states a typical multi-Region architecture can cost about twice a single-Region one and that hot standby doubles workload cost; the 2× here is a guide-derived capacity factor, so storage, transfer and licenses are priced separately. · as-of 2026-09 · https://docs.aws.amazon.com/wellarchitected/latest/reliability-pillar/rel_planning_for_recovery_disaster_recovery.html

AWS post-event summary; InfoQ; ThousandEyes · as-of 2025-10 · https://aws.amazon.com/message/101925/

AWS supplies recovery-pattern boundaries only. Guide unit conversion from an assumed 100 TB decimal copy and one-hour window; no tariff or measured path performance. · as-of 2026-09 · https://docs.aws.amazon.com/wellarchitected/latest/reliability-pillar/rel_planning_for_recovery_disaster_recovery.html

SemiAnalysis ClusterMAX · as-of 2025-11 · https://newsletter.semianalysis.com/p/clustermax-20-the-industry-standard

Amazon Compute SLA, last updated May 25, 2022. · as-of 2022-05-25 · https://aws.amazon.com/compute/sla/

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

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.