SYNESTRIA

RESEARCH SERIES · WHITE PAPER NO. 2

WHITE PAPER · JUNE 2026

The Hidden Cost of Operational Hyperscale

Why the facilities that never fail are losing the most.

ABSTRACT

AI factories do not fail. They forget. The most sophisticated data center infrastructure ever built generates millions of telemetry signals per second and acts on almost none of them. The cost of this inaction does not appear on any dashboard because the measurement framework does not exist. This paper documents four structural loss categories using published research and production data. At $20 to $38 million per megawatt per year in recoverable value intensity — the conservative to base scenario in the Synestria commercial model — every percentage point of EA recovered at a 1GW facility is worth $200 to $380 million annually. The infrastructure investment is already made. The question is how much of it is actually delivering.

PUBLICATION	Synestria Research Series | White Paper No. 2	**DATE**	June 2026
STATUS	Public — For Distribution	**CONTACT**	research@synestria.com

Most AI factories do not fail. They forget.

A facility that has never missed an SLA, never had a major outage, and never failed a compliance audit is still likely operating at 70 to 80% of its designed Economic Availability. The remaining 20 to 30% disappears quietly, a few percentage points at a time, across four structural loss categories that no existing monitoring system was designed to surface.

This is not a failure problem. It is a measurement problem. The industry built its measurement framework around availability. Whether the systems are running and accessible. Availability is the right metric for a general purpose data center. It is the wrong metric for the AI factory.

The AI factory is a compute production system. Its primary product is not uptime. Its primary product is economic output per unit of provisioned infrastructure. A facility that is always available but never operating at full economic output is not a well run facility. It is an expensive one.

$10M	$15M	25%
REVENUE PER MW PER YEAR — AI FACTORY COMPUTE DENSITY (LOW ESTIMATE)	REVENUE PER MW PER YEAR — AI FACTORY COMPUTE DENSITY (HIGH ESTIMATE)	ESTIMATED AVERAGE EA GAP — SHARE OF DESIGNED OUTPUT NOT BEING DELIVERED

At a 500MW AI factory with a 25% EA gap, every percentage point of recovery is worth $50 to $75 million per year. The tools to measure that gap do not exist yet. The data to close it is already being generated.

> *Current monitoring infrastructure is optimized for fault detection. Finding things that are broken. The four loss categories documented in this paper are not broken. They are operating within normal parameters. No alarm will fire. No ticket will be opened. The loss is invisible to every system in the facility.*

Four categories. Four different mechanisms. One shared root cause.

The four structural loss categories below are documented independently across peer reviewed research and production deployments. None of the source papers uses the term Economic Availability. Each measures a different loss category in isolation. Together, they document a facility operating well below its designed output.

01

Stranded Capacity

Lawrence Berkeley National Laboratory (2024) confirmed that more than 30% of servers in production facilities are in a comatose state. Powered on. Passing health checks. Producing nothing. Technical availability: 99.995%. Economic availability of that capital: zero. The facility is paying for full power, cooling, and floor space on hardware that is delivering no revenue. LBNL's methodology used direct server health audits across production facilities. The finding is not an outlier. It is a structural feature of facilities without operational memory.

02

Cooling Inefficiency

Google DeepMind's 2016 deployment documented a 40% reduction in cooling energy at Google hyperscale data centers by applying machine learning to historical sensor and operational data. The cooling infrastructure was not replaced. The chiller plants, cooling towers, and CRAHs were the same. The difference was a system capable of reading the operating history and acting on what it learned. Cooling inefficiency does not appear on utilization dashboards because power usage effectiveness (PUE) measures total power consumption, not cooling efficiency relative to workload. A facility with consistently high PUE and consistently average cooling efficiency is invisible to every standard metric.

03

Coordination Loss

Google's ASPLOS 2020 publication documented that software coordination sustained power oversubscription of 25% or more across tens of megawatts, saving hundreds of millions of dollars. No new hardware was deployed. The capacity existed. It was not being used because the workload scheduler had no visibility into the infrastructure consequence of its decisions. Brookhaven National Laboratory and Lawrence Berkeley National Laboratory (2025) documented a 4x swing in facility power draw from a single parameter change on one H100 node. The scheduler made an optimal compute decision. The facility paid a 300% energy penalty. Neither system knew about the other.

04

Knowledge Loss

McKinsey's 2025 analysis estimates $250 billion in design cost recovery from accumulated operational data applied to future facility planning. This is not a soft estimate. It is based on documented cases where facilities operating their second and third campuses made substantially better design decisions because they had data from the first. The inverse is also documented: greenfield facilities repeating the same commissioning errors, the same stranded capacity patterns, and the same thermal inefficiencies as prior builds. Infrastructure that does not remember cannot improve across generations.

The data exists. The framework does not.

Every data element required to measure, attribute, and recover these four loss categories is already present in a modern AI facility. BMS telemetry. NMS statistics. DCIM exports. Ticketing history. Power draw at the rack level. GPU utilization logs. The data is there. The system to connect it across domain boundaries has never been built.

> *The four loss categories documented above share a common root cause. Infrastructure that does not retain operational experience cannot learn from it. Facilities that capture every intervention, every consequence, and every outcome across all five domains create a compounding advantage. Facilities that do not repeat the same losses every year.*

Why existing tools cannot close this gap

BMS monitors the physical plant. NMS monitors the network. DCIM tracks power and cooling assets. Each system is world class at its own domain. None of them is designed to trace a consequence that runs from the physical layer to the network layer to the compute layer to the economic output layer. The accountability boundary between systems runs directly through the most economically consequential failure modes in the facility.

What the measurement framework requires

Economic Availability requires a continuous intelligence layer above existing domain systems. One that ingests data from all five domains simultaneously. One that builds consequence memory from every event. One that can trace a cooling inefficiency to its effect on compute throughput to its dollar value in lost revenue. That layer is Synestria. It does not replace existing systems. It reads what they already produce and connects what they cannot see.

4	0	$150M
LOSS CATEGORIES — EACH DOCUMENTED INDEPENDENTLY IN PEER REVIEWED RESEARCH	EXISTING TOOLS DESIGNED TO OBSERVE ALL FOUR SIMULTANEOUSLY	ANNUAL VALUE OF 1 POINT EA RECOVERY AT A 1GW FACILITY