The Economic Recovery Case

Four validated recovery vectors. Combined potential of $15 to $40 billion annually.

Every number in this brief comes from production deployments or peer-reviewed research.

“The question is not whether recovery is possible.
The question is who builds the platform that captures it.”

Series	Synestra Research Series | Brief No. 3 of 3
Date	June 2026
Status	Public — For Distribution to Industry Analysts and Investors

EXECUTIVE SUMMARY

The losses are real. So is the recovery.

The Hidden Cost of Operational Hyperscale documented four structural loss categories that drain economic value from AI factories while conventional monitoring systems show green across every dashboard. This brief documents what happens when you fix them.

Each recovery vector in this brief is grounded in production data from operators who have already deployed the underlying coordination approach. These are not projected outcomes or theoretical estimates. They are measured results from facilities operating at the scale that Synestra targets.

The four vectors together represent a $15 to $40 billion annual recovery opportunity across the current hyperscale AI campus market. That number will grow with the market.

40%	25%+	15-30%	$250B
Cooling energy reduction	Power oversubscription	Server fleet recovery	Design cost recovery
(Google DeepMind)	*(Google ASPLOS 2020)*	*(LBNL 2024)*	*(McKinsey 2025)*

All four numbers are drawn from production deployments or peer-reviewed studies. Each represents a recovery vector that Synestra's coordination architecture is designed to capture.

RECOVERY VECTORS

Four mechanisms. Four production data points.

01 Cooling Coordination: 40 percent reduction

In 2016, Google DeepMind deployed machine learning-based cooling control across Google's data center fleet. The result was a 40 percent reduction in cooling energy and a 15 percent improvement in facility-level PUE. This was not achieved on underinvested infrastructure. Google's data center engineering team is among the most capable in the world and had already spent years optimizing the same systems.

The result is significant for two reasons. First, the magnitude: 40 percent cooling energy reduction is not a marginal improvement. Second, the source: this was an additive gain on top of infrastructure that had already been extensively optimized by skilled human engineers. The coordination layer captured something human monitoring and domain-level optimization could not.

Cooling coordination works by treating the chiller plant, cooling towers, and computer room air handlers as a single optimization surface. When one system changes, the others respond as a coordinated unit. Without coordination, each system responds independently, producing suboptimal outcomes for the campus as a whole.

Source: Evans, R. and Gao, J. (2016). DeepMind AI Reduces Google Data Centre Cooling Bill by 40%. Google DeepMind Research Blog.

02 Power Oversubscription: 25 percent and above

Data centers are engineered to provision power conservatively. Every rack slot is allocated peak theoretical draw. The aggregate provisioned load is almost never reached. The gap between provisioned and actual load is not random variance. It is structural.

Sakalkar et al. at Google demonstrated that software-coordinated power oversubscription of 25 percent or higher is achievable in production without compromising workload availability or performance. Over several years of deployment across tens of megawatts of infrastructure, the approach saved hundreds of millions of dollars in data center costs. The approach uses a medium-voltage power plane with priority-aware software capping coordinated across all nodes simultaneously.

At current AI factory revenue densities, 25 percent more deployable capacity from existing infrastructure translates directly to 25 percent more productive compute output without a dollar of

additional capex or power infrastructure.

Source: Sakalkar et al. (2020). Data Center Power Oversubscription with a Medium Voltage Power Plane and Priority-Aware Capping. ASPLOS 2020, ACM.

03 Server Fleet Recovery: 15 to 30 percent

The Lawrence Berkeley National Laboratory's 2024 U.S. Data Center Energy Usage Report estimates that 15 to 30 percent of the server fleet energy can be recovered through identification and elimination of zombie and idle servers. Gartner research estimates that more than 30 percent of enterprise and colocation servers are comatose: powered on, drawing energy, delivering no productive compute output.

In AI factory environments, this problem is concentrated and expensive. GPU servers cost $200,000 to $400,000 per unit and consume 10 to 15 kilowatts each. A server at 10 percent utilization is consuming 6 to 9 kilowatts and producing almost nothing economically. A coordination layer that identifies low-utilization compute, consolidates workloads, and actively manages power state recovers both the energy cost and the productive compute capacity for redeployment.

This is not a hardware problem. The servers are already there. The utilization data is already being collected. What is missing is a coordination layer that reads across systems and acts on what it sees.

Source: Shehabi et al. (2024). United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, LBNL-2001637.

04 Design and Construction Efficiency: $250 billion

McKinsey's August 2025 analysis projects global data center capital expenditures excluding IT hardware will exceed $1.7 trillion by 2030. Of that, McKinsey identifies up to $250 billion as recoverable through improved design and construction methods: standardized campus designs, optimized construction sequences, and economies of scale in modular build-out.

STL Partners adds a time dimension to the construction recovery story. Delays in data center construction cost developers an average of $14.2 million per month for a typical 60 megawatt facility. A three-month delay collapses project IRR from 17.1 percent to 12.6 percent. Scaled to a 2.5 GW campus, a one-month delay costs approximately $590 million in direct economic losses.

Synestra's Operations Management module addresses this vector directly: establishing an Economic Availability baseline during commissioning, before the first production workload, so operators know exactly where the facility stands and what needs to be addressed before revenue operations begin.

Source: McKinsey and Company (2025). Building Data Centers Bigger, Faster. McKinsey Capital Excellence Practice, August 2025. STL Partners (2025). Preventing Multimillion Dollar Data Centre Losses Through Reporting.

AGGREGATE RECOVERY OPPORTUNITY

What the four vectors add up to.

Each recovery vector has been documented independently. Coordinating all four simultaneously, through a unified campus model, produces compounding returns that cannot be achieved by applying any single vector in isolation.

Recovery Vector	Documented Rate	Annual Impact at Scale
01 Cooling Coordination	40% cooling energy reduction 15% PUE improvement	Hundreds of millions per campus
02 Power Oversubscription	25%+ additional deployable capacity	25% revenue uplift from existing infrastructure
03 Server Fleet Recovery	15-30% of fleet energy recovered	$10B+ industry-wide annually
04 Design and Construction	$250B of $1.7T capex recoverable	26% IRR delta from 3-month delay avoided
Combined (conservative)	3-5 percentage points EA improvement	$2.28B – $3.15B/year per 2.5 GW campus

Conservative estimate: 3 percentage points of Economic Availability improvement on a 2.5 GW campus generates $2.28 billion in annual revenue recovery. At a $35 million Year 1 investment (hardware plus license), the conservative payback period is 5.6 days.

WHY THE TIMING IS NOW

The scale problem has already arrived.

The four recovery vectors documented here are not new. The research is five to ten years old in some cases. What has changed is the economic consequence of not acting on them.

In the general-purpose colocation era, a 10 percent efficiency gap on a facility generating $1 million per megawatt per year cost $100,000 per megawatt. Annoying, but not strategic. At $12.5 million per megawatt per year in AI factory revenue density, the same 10 percent gap costs $1.25 million per megawatt per year. At 2.5 GW campus scale, it costs $3.1 billion per year. The infrastructure did not change. The economics did.

And the telemetry problem grows with every rack added. At 100 racks operating on 800VDC bus systems, per-rack power, cooling, and compute telemetry already generates thousands of data points per second across systems that do not talk to each other. A NOC team can monitor 100 racks. It cannot monitor 1,500. At 1,500 racks per building approaching one gigawatt of draw, the gap between what human monitoring can track and what the facility actually needs becomes structurally dangerous.

The industry is building at this scale right now. The campuses being constructed today at 2 to 3 GW will not have adequate coordination infrastructure when they come online in 2026 and 2027. That is the first-mover window.

INVESTOR IMPLICATIONS

The recovery opportunity is not speculative.

Each of the four recovery vectors documented in this brief was validated in production at real facilities, under real workloads, at real cost. The recovery rates are not projections. They are measured outcomes from operators who built the coordination infrastructure and measured what happened.

The aggregate recovery opportunity, conservatively estimated at $15 to $40 billion annually across the current hyperscale AI market, will grow with the market. U.S. data center capacity is projected to grow from 30 gigawatts in 2025 to 90 gigawatts by 2030. Every gigawatt built without a coordination layer is a gigawatt operating at 70 to 78 percent of its economic potential.

The question is not whether the recovery opportunity is real. The research record settled that. The question is who builds the coordination platform that captures it at the campus scale the market is now requiring.

Part of a three-brief series. See also: The Hidden Cost of Operational Hyperscale and the Research Gaps and Evidence Map. *| synestra.ai | john.chavner@synestra.ai*