Futurum
RESEARCH

Lenovo

From Cost Per Token to Value Per Token:

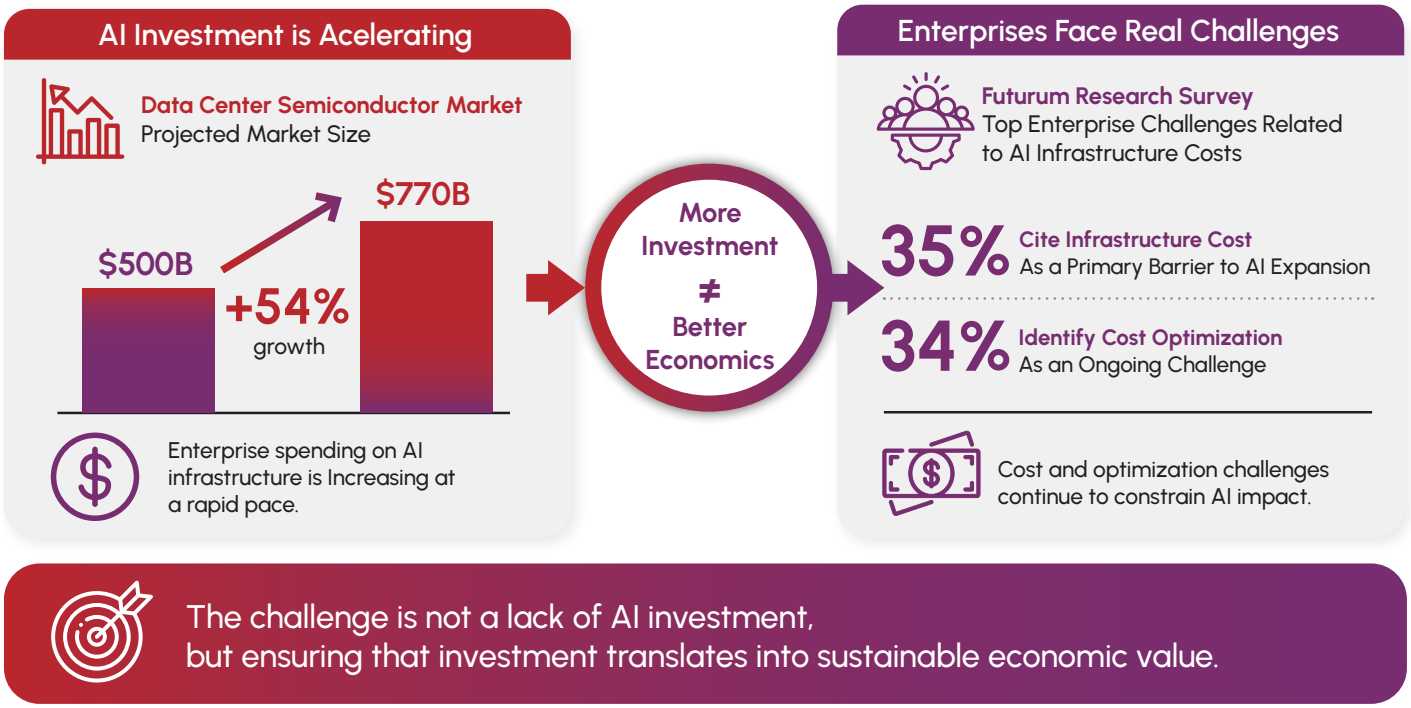
The AI Factory Imperative

SEPTEMBER 2026

The Real Problem: AI Infrastructure Without Economics

Enterprises are pouring capital into AI infrastructure, but capacity alone isn't producing the returns they expected. Nearly four in five organizations (78%) expect to increase AI budgets over the next 12 months. Yet in Futurum Research's AI Platforms 2026 Decision Maker Survey, 35% cite infrastructure cost as a primary barrier to AI expansion and 33% identify cost optimization as an ongoing challenge.[1] Organizations know how to acquire AI compute, but not how to reliably convert that capacity into productive, measurable output (see Figure 1).

Figure 1. AI Investment Is Accelerating, but Economic Challenges Remain
Enterprises are investing heavily in AI infrastructure,
yet cost and optimization challenges continue to limit impact.

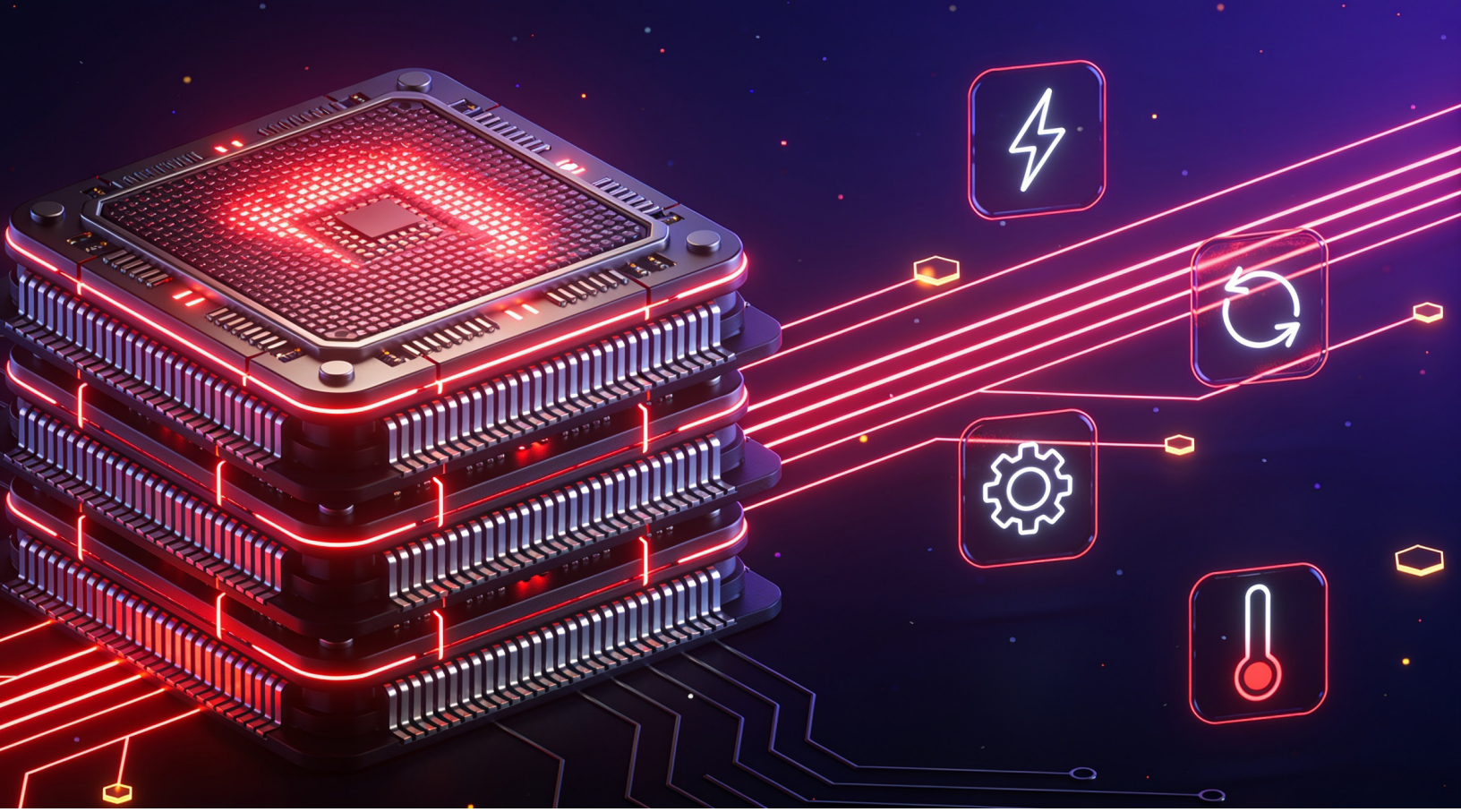


Source: Futurum Research

The gap between capital deployed and value returned comes from operating, optimizing, and adapting that compute as workloads evolve. Closing that gap is the differentiator: the ability to manage compute as an economic system across its full lifecycle. That challenge is intensifying as the industry's center of gravity shifts from training to inference. As inference scales, the economics of AI infrastructure become token-level economics. Cost per token, token throughput, GPU utilization, and token output per watt are infrastructure concerns, shaped by architectural decisions made at the time of procurement and refined continuously through active management.

Understanding this is essential for infrastructure decision-makers. Spending on AI infrastructure without a framework for tracking and improving these metrics is a path to stranded investment.

1 Futurum Research. AI Platforms Decision Maker Survey, 1H 2026.



Why AI Factories Break Traditional IT Models

An AI factory is not a conventional IT deployment. Traditional infrastructure is designed around stable, predictable workloads: a transaction system runs the same processes at predictable volumes, and capacity planning is relatively straightforward. AI factories operate on different terms. Training and inference workloads are dynamic and concurrent. Resource demands are non-linear. Power and thermal requirements fluctuate in ways that static provisioning cannot absorb. Real-time orchestration and dynamic resource balancing are not optional features; they are operational prerequisites.

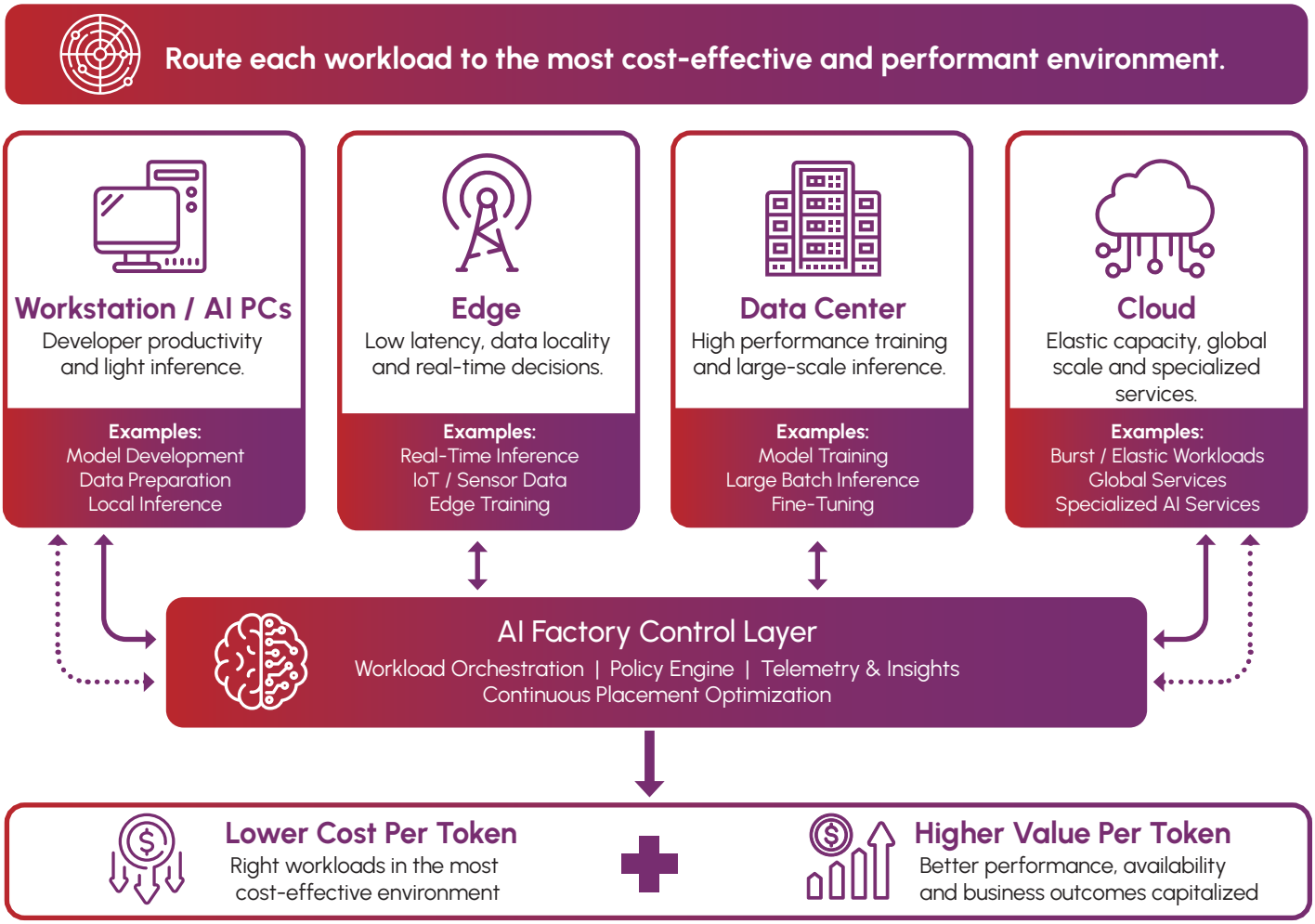
The deployment environment adds further complexity. Futurum Research data shows that enterprise AI compute is distributed across public cloud (40.7%), on-premises data centers (35.6%), colocation (13.5%), bare metal/HPC (5.6%) and edge (4.7%).[2] Consumption models are equally fragmented, split across CapEx purchases, reserved capacity, on-demand, and leasing arrangements. There is no single architectural pattern that fits all workloads, and no stable end state to plan toward. The implication is that organizations need the ability to make and continuously revisit placement decisions across the AI continuum, based on the economics and requirements of each workload, not that every workload should be consolidated in one environment.

This is precisely the operating context in which Lenovo's Hybrid AI Advantage™ becomes relevant. The challenge is not simply to provide compute in multiple locations. It is to help customers make better decisions across the AI continuum – matching the right workload, model, data, infrastructure, and operating environment to the outcome required. Lenovo is positioned across the continuum represented in Figure 2: from AI PCs and workstations to edge, data center, and cloud environments. That breadth matters because workload placement is itself an economic decision. Running inference closer to the user or data may improve latency, privacy, resilience, and data-movement economics; centralized infrastructure may provide the scale, control, and predictable performance required for other workloads. Hybrid workload placement can therefore directly affect cost per token and value delivered per token.

2 Futurum Research. AI Chipsets Decision Maker Survey, 1H 2026

Lenovo's relevance is not limited to its technology footprint. AI factories also require the ability to connect architecture, deployment, operations, and ongoing optimization as one lifecycle. The Lenovo Hybrid AI Factory provides the integrated technology foundation across infrastructure and devices, data, models, and platforms; lifecycle expertise helps customers adapt those decisions as workload demand, technology, and economics change.

Figure 2. Hybrid Workload Placement Optimizes AI Factory Economics
Routing workloads to the most cost-effective and performant environment directly improves cost per token and value delivered per token.



Dynamic workload placement across workstations, edge, data centers and cloud drives better economics and better outcomes.

Source: Futurum Research

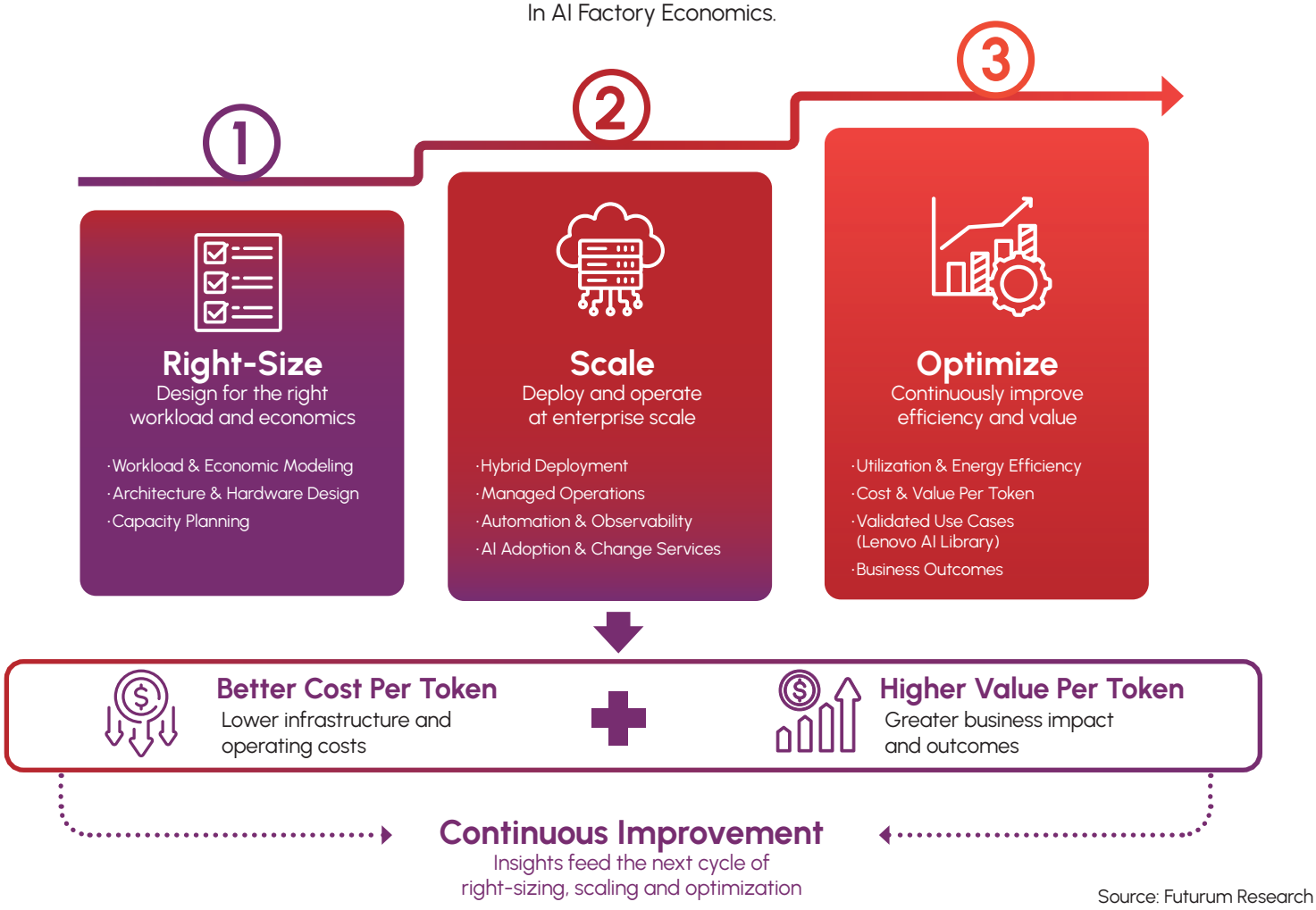
Full-Stack Economics: Where Token Value Is Created

Token economics are not determined by any single layer of the AI stack. Infrastructure architecture, networking fabric, storage configuration, model selection, orchestration, and lifecycle services all shape it. Within the broader set of considerations that matter for infrastructure decision-makers – cost, value, speed, risk – cost per token and value per token are the two primary economic metrics.

The objective is to manage AI for cost, value, speed, and risk – improving efficiency, accelerating time to useful outcomes, and maintaining the security, resilience, and governance required for production AI, not simply lowering cost per token.

Focusing in isolation produces a partial picture. A well-optimized inference pipeline that runs on poorly matched hardware will still underperform. Training and inference carry different infrastructure profiles: training typically demands a large footprint, often rented from neoclouds rather than built on-prem, while inference requires meaningfully less capacity once deployed. Infrastructure sized for peak training but left in place – or left idle – as workload shift to inference carries unnecessary cost. Lifecycle services are the mechanism through which these layers are continuously aligned. Three operational capabilities are central to this lifecycle model (see Figure 3):

Figure 3. Lifecycle Services Drive AI Factory Economics
Three Operational Capabilities Deliver Continuous Improvement In AI Factory Economics.



Source: Futurum Research

Right-Sizing: Advisory and Design

The starting point for AI factory economics is architectural fit. Overbuilt infrastructure wastes capital; underbuilt infrastructure constrains performance and limits the throughput metrics that determine cost efficiency. With 78% of organizations expecting to increase AI budgets over the next 12 months,[2] the risk of misallocation is material.

Effective right-sizing requires workload characterization even before procurement decisions are made. This means understanding the ratio of training to inference workloads, the expected token throughput requirements, the sensitivity of target use cases to latency, and the likely evolution of workload mix over time. Architecture design and phased deployment planning should follow from that analysis, not precede it. Infrastructure decisions made at this stage – GPU selection, memory configuration, networking topology, storage tiering – set the baseline token economics that subsequent optimization can improve upon but cannot easily overcome.

Scaling: Implementation and Managed Services

Most AI infrastructure programs encounter a predictable obstacle between proof of concept and production: the operational demands of running AI at scale exceed what the original deployment model was designed to handle. The enterprise and edge AI infrastructure market is projected to grow from $45.8 billion in 2025 to $152 billion by 2030,[3] with tier 2 cloud AI infrastructure growing at 107% year-on-year.[3] Much of that investment will be challenged by the gap between initial deployment and sustained productive operation.

Managed scaling across hybrid environments addresses this directly. Implementation and deployment orchestration that spans on-premises, colocation, and cloud, routing workloads based on cost, latency, and capacity, is what converts infrastructure investment into operational throughput. AIOps-driven automation reduces the manual overhead of managing distributed AI infrastructure, and creates the observability needed to track utilization and efficiency metrics at the token level. This requires a continuous operating discipline that connects workload telemetry, utilization, cost, energy consumption, performance, and business outcomes with the ability to adjust placement, capacity, and operating practices over time. Scaling, in this context, is an ongoing operational capability, not a one-time build event.

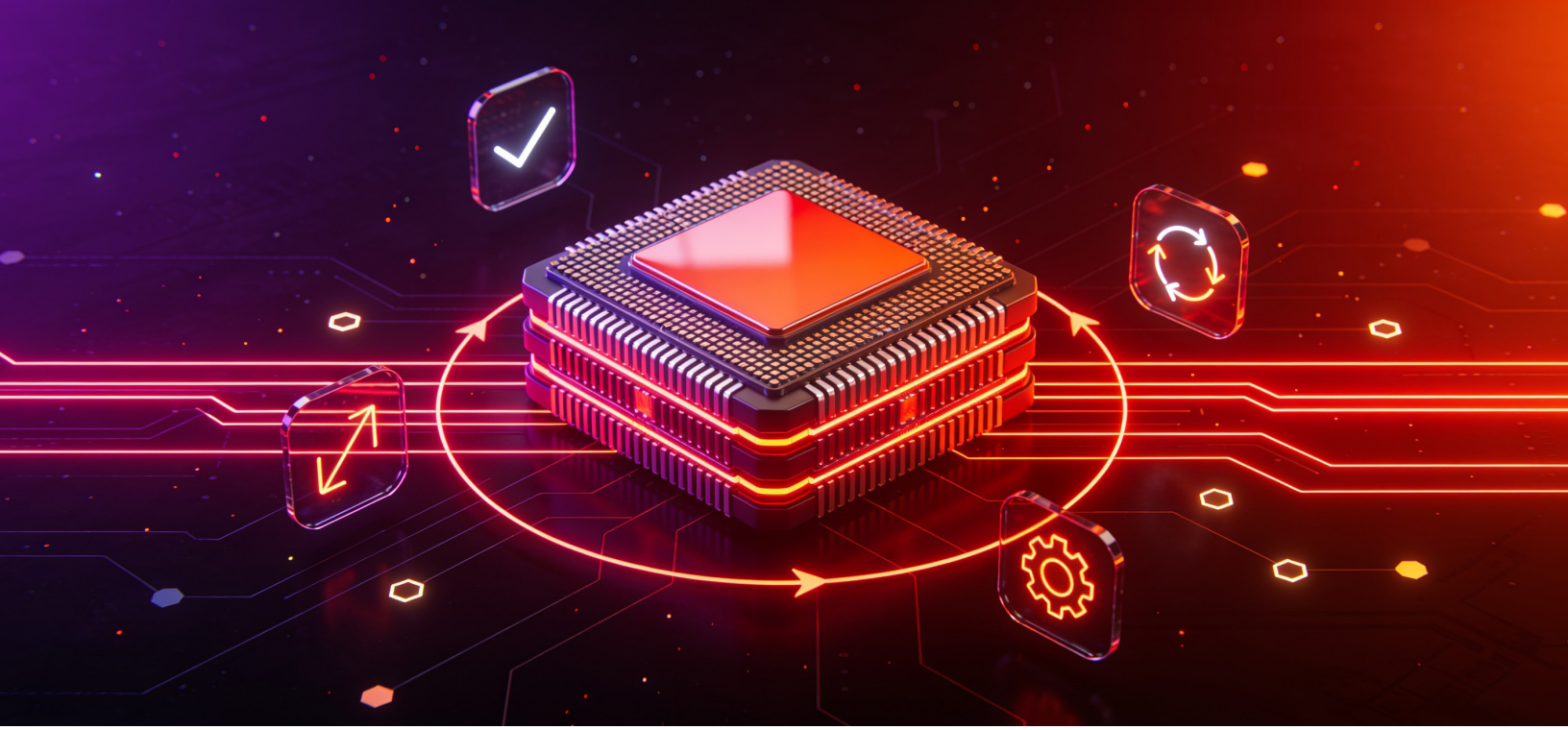
Monetization: Continuous Optimization

Underutilized infrastructure is a stranded investment. The structural challenge is that AI factory workloads are inherently dynamic: inference demand patterns shift, model versions change, and business priorities evolve. Infrastructure configured to handle one workload profile will drift out of alignment with the next unless actively managed.

This is where the distinction between cost per token and value per token becomes practically important. Cost per token, driven by GPU utilization rates, token throughput per watt, and workload-to-hardware fit, alongside the broader cost of delivering AI such as storage, data movement, and labor, is the efficiency metric. Value per token is determined by what the AI output actually enables: faster decisions, reduced manual processing, new service capabilities, or accelerated time to value. Lenovo's AI Library provides repeatable, pre-validated AI use cases that can address the value side of this equation by reducing time to value, complexity, and risk, while increasing the productive output of each token generated by the factory.

Power and energy efficiency are increasingly central to both metrics. With 33% of data centers expected to operate on 100% on-site power by 2030, energy cost is becoming a direct input to token economics.[4] Workload and model optimization, energy efficiency tuning, and shared infrastructure models are the services levers that translate infrastructure investment into sustained economic returns.

3 Futurum Research. Data Center Semiconductors Market Sizing & Five-Year Forecast, 1H 2026.
4 Futurum Research. AI Grid Constraints Will Push Over 33% of Data Centers Off-Grid by 2030, March 2026.



From Infrastructure to Lifecycle Economics

AI factories are ongoing economic systems, not capital projects with a completion date. Their performance is determined by how well the full stack – business use case, data, model, infrastructure, workload placement, operations, and services – is continuously aligned to changing demand and business objectives. This is the central challenge of token economics. The right model, infrastructure configuration, or workload location today may not remain the right choice as utilization, power availability, data requirements, performance expectations, and business priorities evolve. Sustainable returns therefore require more than a one-time infrastructure decision. They require a continuous operating discipline: design, place, deploy, operate, optimize, scale, and refresh. Lenovo describes this discipline as TokenOps: an AI-native approach to making token economics and AI outcomes more visible, measurable, and manageable across the lifecycle.

Lenovo's Hybrid AI Advantage brings together the capabilities required to support that discipline. Lenovo's Hybrid AI Factory provides the integrated technology foundation across infrastructure and devices, data, models, and platforms. The Lenovo AI Library helps customers convert capacity into business value faster through repeatable, outcome-oriented use cases, reducing the time, complexity, and risk of moving from AI investment to productive consumption. Lenovo AI Services provide the expertise to design, place, deploy, operate, optimize, scale, and refresh AI environments as workloads and requirements change.

Together, these capabilities enable TokenOps in practice: making the economics and outcomes of AI more visible, measurable, and manageable across the lifecycle. The objective is to continuously improve how effectively AI investment is converted into useful business outcomes, not simply to generate the lowest-cost token.

For organizations procuring AI infrastructure, the question is therefore not only which hardware to buy, but how to make every AI token count. Lenovo's ambition is to help customers get more value from AI – faster, more efficiently, with greater confidence and less risk – while continuously improving token economics over time.

Important Information About This Report

AUTHORS

Nick Patience
Vice President & Practice Lead,
AI Platforms | The Futurum Group

PUBLISHER

Futurum Research

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

This document, including any supporting materials, is owned by The Futurum Group. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of The Futurum Group.

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

Lenovo

ABOUT LENOVO

Lenovo is a US$83 billion revenue global technology powerhouse, ranked #153 in the Fortune Global 500, and serving millions of customers every day in 180 markets. Guided by its vision of "Smarter Technology for All", Lenovo is executing a Hybrid AI strategy that spans Personal AI – one personal AI, multiple devices; and Enterprise AI – helping customers turn data into insights and value. This strategy is delivered through the Group's commitment to world-class innovation and a full-stack AI portfolio, including devices (PCs, workstations, smartphones, tablets, accessories), infrastructure solutions (server, storage, edge, high-performance computing and software defined infrastructure), as well as software, solutions, and services. With a global footprint spanning more than 20 research and development locations and a global supply chain that includes more than 30 manufacturing sites across 11 markets, Lenovo is widely recognized for its operational excellence. Lenovo is listed on the Hong Kong stock exchange under Lenovo Group Limited (HKSE: 992) (ADR: LNVGY). Learn more at www.lenovo.com and follow the latest news in our newsroom.

Futurum®

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

Futurum®
CONTACT INFORMATION: The Futurum Group LLC | futurumgroup.com | (833) 722-5337

© 2026 The Futurum Group. All rights reserved.