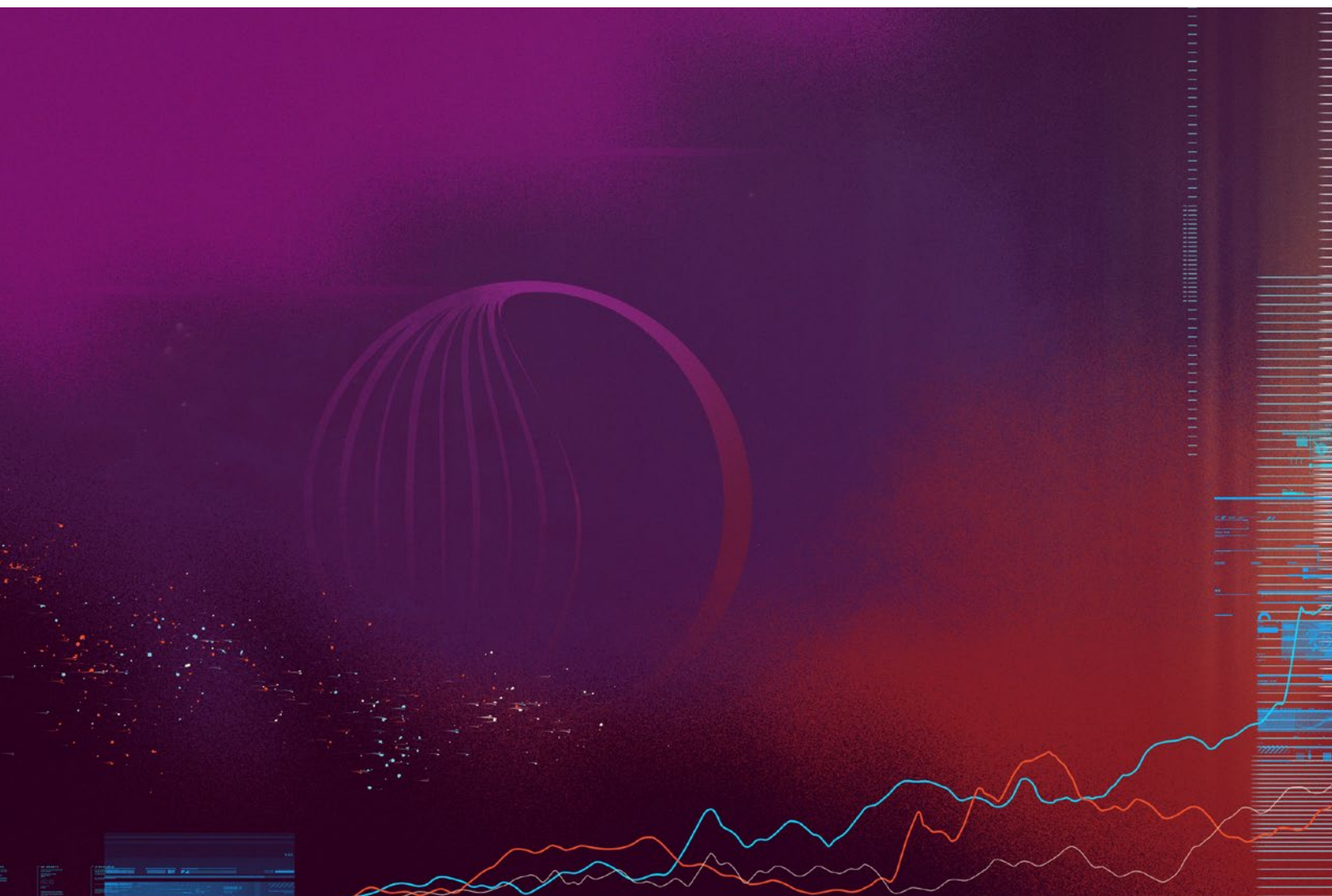




From Cost Per Token to Value Per Token: Validating the Economics of the Lenovo Hybrid AI Factory and Lenovo AI Library

AUTHORS

Mitch Lewis

Performance Analyst | Signal65

Cameron Moccari

Operations Director | Signal65

IN PARTNERSHIP WITH

Lenovo

OCTOBER 2026

Executive Summary

Enterprises are no longer asking whether they can acquire AI compute; they are asking whether that compute reliably converts into business value. As inference workloads scale, the economics of AI infrastructure become token-level economics, and two metrics come to the fore: the cost to produce each token, and the value each token delivers. Building an economic foundation that delivers more predictable AI infrastructure costs is essential for sustained enterprise AI operations; however, the challenge is converting that cost efficiency into valuable, predictable business outcomes. Lenovo's answer to both is the Lenovo Hybrid AI Factory — integrated infrastructure, data, models, and platforms — paired with the Lenovo AI Library of validated use cases and lifecycle services including Lenovo AI Fast Start and the Lenovo AI Express program.

Signal65 measured real token throughput on a Lenovo hybrid AI factory system and combined those measurements with Lenovo-provided representative pricing to evaluate the economics of self-hosting AI on Lenovo infrastructure against public cloud alternatives. Key findings:



Value per token in as little as 90 days from deployment – Organizations can start driving value per token from deployment of Lenovo AI Library use cases on right-sized hybrid AI factory configurations within 90 days. The Lenovo AI Express program speeds order-to-ship to 15 business days for eligible hybrid AI factory configurations, while Lenovo AI Fast Start drives value realization in as little as 90 days from deployment¹.



More than two-thirds lower cost per token – At full utilization the Lenovo Hybrid AI Factory delivers tokens at roughly 70% below a hyperscaler's per-token API list price, and about two-thirds (66%) below renting an 8-GPU cloud instance around the clock, assuming a 30% discount on the instance.



On-premises AI breaks even at just over 30% utilization – Compared with API fees, cost per token runs 39–62% lower as utilization reaches 50–80%. Long-running, always-on agentic workloads keep hardware busy around the clock, and that sustained demand is where on-premises pulls ahead.



An integrated AI stack, not just raw tokens – An API sells tokens and a GPU instance rents compute. Enabled by Lenovo Services, Lenovo-validated Hybrid AI Factories integrate infrastructure, enterprise data and knowledge, and agent models and platforms to simplify deployment and management of use cases in the Lenovo AI Library — while processing tokens at up to 70% lower cost than cloud alternatives.

The Shift from Cost Per Token to Value Per Token

Enterprise AI investment continues to accelerate, but capacity alone is not producing the returns organizations expected. Futurum Research's AI Platforms 2026 Decision Maker Survey finds that nearly four in five organizations expect to increase AI budgets over the next twelve months, yet 35% cite infrastructure cost as a primary barrier to AI expansion and roughly a third identify cost optimization as an ongoing challenge.² Organizations know how to acquire AI compute; converting that capacity into productive, measurable output is the harder problem.

¹ Lenovo AI Fast Start program documentation: production-ready AI pilot with value realization in as little as 90 days from deployment.

² Futurum Research, AI Platforms Decision Maker Survey, 1H 2026, as reported in "From Cost Per Token to Value Per Token: The AI Factory Imperative," Futurum Research, September 2026.

That challenge intensifies as the industry's center of gravity shifts from training to inference. At inference scale, infrastructure economics become token economics: cost per token, token throughput, GPU utilization, and token output per watt are determined by architectural decisions made at procurement and refined through active management. Cost per token is the efficiency metric. Value per token — what each token of AI output actually enables in faster decisions, reduced manual processing, or new service capability — is the outcome metric. A deployment can excel at the first and still fail at the second if the organization has no use case ready to consume the capacity.

This report evaluates Lenovo's approach to both sides of that equation: measured cost per token on Lenovo infrastructure relative to public cloud alternatives, and the mechanisms — the Lenovo AI Library, Lenovo AI Fast Start, and Lenovo AI Express — by which Lenovo compresses the time between infrastructure investment and value-producing tokens.

The Lenovo Approach: Hybrid AI Factory, AI Library, and Lifecycle Services

The Lenovo Hybrid AI Factory provides an integrated technology foundation across infrastructure and devices, data, models, and platforms, supported by lifecycle services that adapt deployments as workload demand and economics change. Rather than selling compute capacity alone, the factory model pairs validated hardware configurations with the software and services required to operate AI as an ongoing economic system.

The Lenovo AI Library is a curated catalogue of prebuilt agents and validated use cases built on proven Hybrid AI solutions, spanning functions from contact-center automation to computer vision for manufacturing and retail. Agents built with the Lenovo xIQ Agent Platform and the AI Library use cases run on the Lenovo Hybrid AI Factory. In prior work, Signal65 validated the Lenovo **Super Agent for Retail** and the **Lenovo Knowledge Super Agent** AI Library solutions.

The Lenovo AI Express program speeds order-to-ship to 15 business days for eligible hybrid AI factory configurations.³ Lenovo Hybrid AI Services then help enterprises rapidly achieve AI business value. Lenovo AI Fast Start is a structured services engagement that takes a customer's chosen use case from hardware delivery to a production-ready AI pilot, driving value realization in as little as 90 days from deployment; for organizations still identifying their opportunity, Lenovo AI Discover provides the on-ramp.

³ Lenovo, "Lenovo AI Express with NVIDIA Accelerates Path to Hybrid AI Factory in Weeks," news.lenovo.com, 2026.

Validation Overview

As a baseline for this analysis, Signal65 completed an economic evaluation of Lenovo on-premises hardware and cloud solutions. Signal65 benchmarked token throughput on a Lenovo ThinkSystem SR675 V3 with eight NVIDIA H100 PCIe GPUs serving the open-weight gpt-oss-120b model. The measured throughput and three-year pricing for a representative Lenovo configuration were combined to calculate an on-premises cost per token. Pricing data includes hardware costs, three-year 24x7 onsite support, management licensing, NVIDIA AI Enterprise, and electricity.

The on-premises cost per token value was then utilized to create two economic comparisons to cloud-hosted AI. The first compares against a cloud API endpoint, a hosted service that runs the model and charges per token, serving the same model, gpt-oss-120b. The second evaluates the cost of an equivalent cloud instance with 8 NVIDIA H100 GPUs. Cloud instance costs were modeled with publicly available on-demand hourly costs, with an assumed discount of 30% applied.

Appendix A details the test configuration, measured performance, cost inputs, and assumptions.

Tokenomics Findings

Roughly 70% lower cost per token at full utilization

At full utilization of measured throughput, the Lenovo system delivers tokens at roughly 70% below the cloud API list price for the same model. Although organizations are unlikely to achieve full utilization, on-premises token cost advantages begin much sooner. The cost per token of on-premises hardware breaks even at approximately 30% utilization. As utilization grows beyond 30%, Lenovo on-premises infrastructure achieves further cost savings, with 39% lower cost per token at 50% utilization, 57% lower cost per token at 70% utilization, and 62% lower cost per token at 80% utilization.

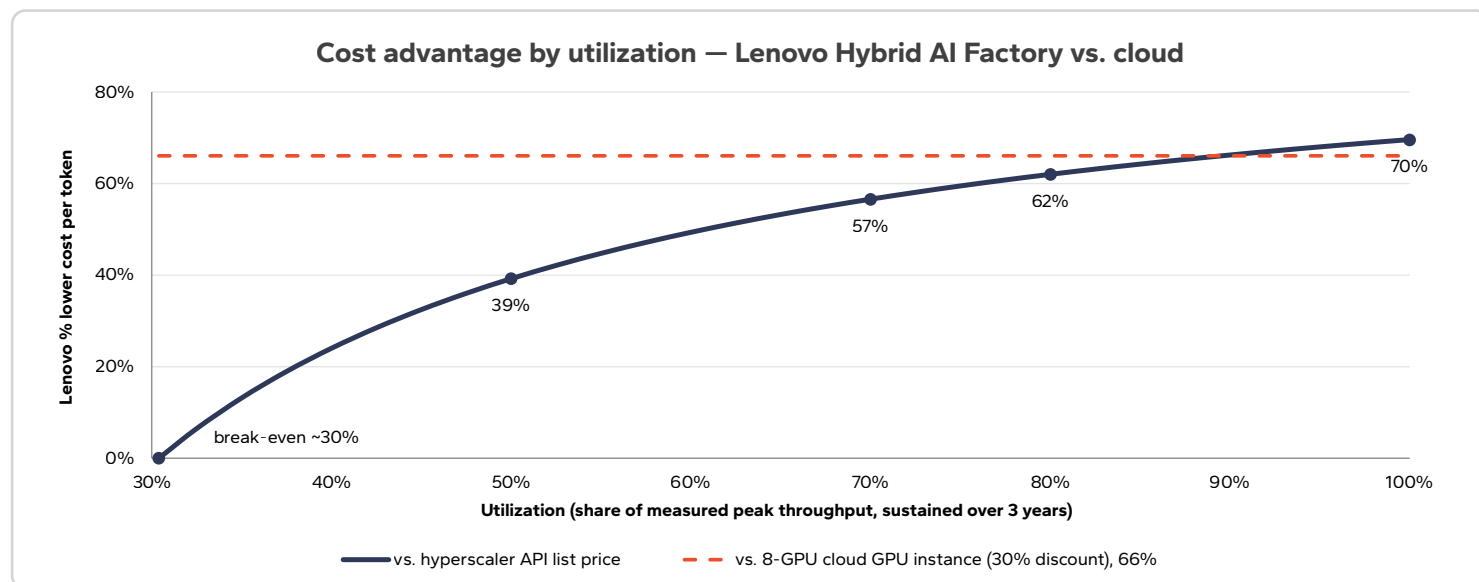


Figure 1: Signal65 analysis · gpt-oss-120b balanced workload, measured on an 8-GPU NVIDIA H100 system · Lenovo and cloud pricing as of October 7, 2026

While the cloud API is more cost-effective at low utilization, a break-even point near 30% is likely attainable for many agentic workloads. Long-running, always-on AI agents can complete tasks around the clock, helping sustain consistent GPU utilization. As enterprises increasingly adopt agentic AI, that sustained GPU demand can create a strong economic advantage for on-premises infrastructure.

Sustained utilization	Cost per token vs. Cloud API
~30%	Break-even
50%	Lenovo ~39% lower
70%	Lenovo ~57% lower
80%	Lenovo ~62% lower
100%	Lenovo ~70% lower

About two-thirds lower cost than comparable cloud GPUs

Instead of utilizing hosted models via a cloud API endpoint, organizations can also rent full GPU instances in the cloud; however, this analysis again shows a large cost advantage for on-premises hardware. Compared to a cloud instance with eight comparable NVIDIA H100 GPUs, the Lenovo system costs about two-thirds (66%) less over the three-year period. In this case, the advantage holds at any utilization level, since both systems are assumed to run continuously. Assuming equivalent GPUs achieve equivalent throughput, this price advantage also applies to cost per token, with the cloud instance costing approximately 3x as much per token as Lenovo on-premises hardware. This analysis assumes a 30% discount to cloud pricing; at full on-demand list price, the on-premises cost advantage grows to approximately 76%.



Beyond Tokenomics: Achieving Value per Token

Cost per token presents a strong economic argument for deploying enterprise AI on-premises; however, it only tells part of the story. While AI workloads are priced in tokens, the actual value is found in what those tokens achieve. Enterprise organizations don't want to simply buy tokens; they want to deploy meaningful AI applications and agentic workloads to improve customer experiences, accelerate internal operations, or automate complex processes.

The economic comparison alone understates a structural difference in what is actually purchased. An API simply sells tokens and a GPU instance rents compute by the hour. In both cases, the customer is required to assemble everything else to achieve a valuable AI workload. This involves months of custom development, followed by ongoing management of data integration, security, and drift monitoring. This complexity results in significant developer time and cost before achieving a valuable AI solution.

Alternatively, Lenovo-validated Hybrid AI Factories integrate infrastructure, enterprise data and knowledge, and agent models and platforms to simplify deployment and management of use cases in the Lenovo AI Library. This streamlines the process, enabling organizations to quickly achieve real business value. An example of this value can be seen in the Lenovo Knowledge Super Agent solution, **previously validated by Signal65**. In this solution, which provides an AI-powered enterprise knowledge platform, Signal65 found that

employee time spent on knowledge retrieval tasks could be reduced by 30%, saving organizations up to 120 hours per employee annually. By leveraging the Lenovo AI Library, Lenovo Hybrid AI Factory, and Lenovo Hybrid AI Services, organizations not only achieve a lower cost per token from the underlying infrastructure but also achieve significant value per token.

	Cloud API endpoint	Cloud GPU instance	Lenovo Hybrid AI Factory with AI Library
What you buy	Tokens	Compute by the hour	A validated stack with agents and services, on hardware you own
Ready-to-customize agents and use cases	No; you build them	No; you build them	Yes, from the Lenovo AI Library
Time to value	Typically months of custom development	Typically months of custom development	Production-ready AI pilot in 90 days with Lenovo AI Fast Start; eligible configurations ship in 15 business days with Lenovo AI Express
Where data and models run	Provider's environment	Your cloud tenancy	On your premises or hybrid, governed by design
Cost per token	Cost advantage only at low utilization (below ~30%)	About 3x Lenovo's cost per token, even with a 30% discount	Break-even at ~30% utilization; more than two-thirds lower than the API at full utilization
Services	Provider support	Provider support	Lenovo AI Services, from discovery through ongoing optimization

Valuable Tokens in 90 Days

Along with economics and business value, another critical component for enterprises deploying AI solutions is time. For enterprises, AI workloads can generate massive value such as cost savings, accelerated development and innovation, creation of new business opportunities, and more. This value creates an imperative for organizations to begin deploying AI workloads as quickly as possible. However, as previously discussed, token APIs and GPU instances do not provide a full AI solution, and custom AI engineering introduces additional time, complexity, and cost.

Lenovo's deployment motion compresses this process, enabling organizations to achieve valuable AI applications more quickly. Eligible Lenovo Hybrid AI Factory configurations ship in 15 business days through Lenovo AI Express. The Lenovo AI Library use cases provide a validated starting point rather than a blank page. Signal65 **previously validated** a single-week deployment of the Retail Super Agent use case.

Lenovo AI Fast Start drives value realization in as little as 90 days. The 90 days start at deployment, once hardware is delivered and installed. Lenovo AI Express covers the order-to-ship stage before that, at 15 business days. The figure assumes a chosen use case delivered through Lenovo AI Fast Start; organizations still deciding begin with Lenovo AI Discover. "In as little as" is Lenovo's design target, and actual timelines depend on use-case complexity, data readiness, and integration work. In earlier testing, Signal65 saw an AI Library use case deployed and running within a week.

Final Thoughts: Lower Cost, High Value Use Cases, Accelerated Time to Value

Deploying AI workloads in the enterprise weaves together several key considerations – what does the solution cost, how much business value does it bring, and how quickly can it be deployed. This evaluation shows that Lenovo Hybrid AI Factories with the Lenovo AI Library present compelling answers to all of these questions.

Signal65's measured results show that organizations running sustained inference on Lenovo on-premises hardware can achieve per-token cost advantages of roughly 70% compared to model APIs and 66% compared to discounted on-demand cloud GPU instances. Meanwhile, the Lenovo Hybrid AI Factory and Lenovo AI Library elevate the solution beyond economics, providing a validated stack — with infrastructure, data, models, validated designs, agents and services — on hardware you own. Additionally, Lenovo AI Fast Start and Lenovo AI Express compress the time between infrastructure investment and efficient, valuable token production — enabling organizations to achieve real AI value in as little as 90 days.

Appendix A: Methodology, Results, and Assumptions

This appendix documents the configuration, measured performance, cost inputs, and calculations behind the results in this report.

Test configuration

Item	Detail
System	Lenovo ThinkSystem SR675 V3
GPUs	8× NVIDIA H100 PCIe (80 GB)
Serving software	NVIDIA NIM 2.0.5 (vLLM backend)
Model	gpt-oss-120b, served as four replicas of two GPUs each
Load generator	NVIDIA GenAI-Perf
Workload	Balanced: about 1,000 input and 1,000 output tokens per request
Operating point	512 concurrent requests

Measured performance

All cost calculations use the results below, measured at the operating point above.

Metric	Result
Aggregate output throughput	~16,600 tokens per second
Requests completed	~16.8 per second
Generation speed per user	~35 tokens per second
Time to first token (median / 95th percentile)	~0.37 s / ~0.79 s

Three-year Lenovo cost

Component	Three-year cost	Notes
System, 3-year 24x7 onsite support, and XClarity management licensing	\$280,000	Lenovo ThinkSystem SR675 V3, 8× NVIDIA H100 PCIe (based on Lenovo-provided representative street pricing, derived under Lenovo's standard pricing methodology from a same-chassis current-generation configuration, as of October 7, 2026)
NVIDIA AI Enterprise	\$52,000	\$6,500 per GPU for three years × 8 GPUs
Power and cooling	\$29,021	6,135 W × \$0.12/kWh × PUE of 1.5, around the clock for three years
Total cost	~\$361,000	About \$13.74 per hour over 26,280 hours

Cloud comparison pricing

Item	Value
Pricing date	Published cloud pricing as of October 7, 2026
Cloud API (gpt-oss-120b)	\$0.15 per 1M input tokens and \$0.60 per 1M output tokens
Cloud GPU instance	8× NVIDIA H100 at \$55.04 per hour on demand, with an assumed 30% discount
NVIDIA AI Enterprise on the instance	Same price as the Lenovo system (\$6,500 per GPU for three years)
Term	Three years, running around the clock

How the results are calculated

Cost per token – Total three-year cost divided by the tokens processed over three years at a given utilization, with input and output tokens blended at the balanced workload shape.

Utilization – The share of measured peak throughput sustained over the term. The 50–80% levels are illustrative.

Break-even – The utilization at which Lenovo's cost per token equals the API price for the same requests: about 30%.

Cloud instance comparison – The ratio of three-year costs. Because equivalent throughput is assumed on both sides, the result holds at any utilization.

Sensitivities

Change to the assumptions	Effect on the result
Full on-demand list price for the cloud instance (no discount)	Advantage over the instance rises from ~66% to ~76%
NVIDIA AI Enterprise bought pay-as-you-go on the cloud marketplace (\$8.00 per hour)	Advantage over the instance rises to ~70% (~78% at list)
Cloud instance delivers more throughput than the tested system	Advantage narrows; Lenovo stays lower cost unless the instance delivers about 3× the throughput
Lower throughput on the priced system	At 80% utilization, Lenovo stays below API pricing down to about 38% of measured throughput
Facility overhead (PUE) of 1.3 instead of 1.5	Results change by less than 1 point
Expressing the results in absolute dollars	In absolute terms, the 66% advantage equals approximately \$700K per 8-GPU system over three years (about \$235K per year) versus the discounted instance. Savings scale linearly with system count only where sustained utilization holds. The break-even framing above is the guardrail for fleet-level extrapolation.

Scope and limitations

Workload – Results are based on one model (gpt-oss-120b) and a single request shape (balanced input and output). Other models and request shapes may produce different results.

Pricing – All pricing is point-in-time and indicative; cloud rates can change.

Not included – Hosting, networking, staff, services and taxes.

Disclosures

This report was commissioned by Lenovo. Signal65 is part of The Futurum Group; Futurum Research, cited in this report, is an affiliated research organization. Signal65 provides research, analysis, and evaluation services to many high-technology companies, including those mentioned in this report. The analysis and conclusions herein reflect Signal65's independent assessment based on the measurements and sources described. Performance and cost results are specific to the configurations, workloads, pricing, and dates documented in Appendix A; other configurations and market conditions will produce different results.

Important Information About this Report

CONTRIBUTORS

Mitch Lewis

Performance Analyst | Signal65

Cameron Moccari

Operations Director | Signal65

PUBLISHER

Ryan Shrout

President and GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 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 "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

This document, including any supporting materials, is owned by Signal65. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of Signal65.

DISCLOSURES

Signal65 provides research, analysis, advising, and lab services to many high-tech companies, including those mentioned in this paper.

ABOUT SIGNAL65

Signal65 exists to be a source of data in a world where technology markets and product landscapes create complex and distorted views of product truth. We strive to provide honest and comprehensive feedback and analysis for our clients in order for them to better understand their own competitive positioning and create optimal opportunities to market and message their devices and services.



IN PARTNERSHIP WITH



CONTACT INFORMATION

Signal65 | signal65.com