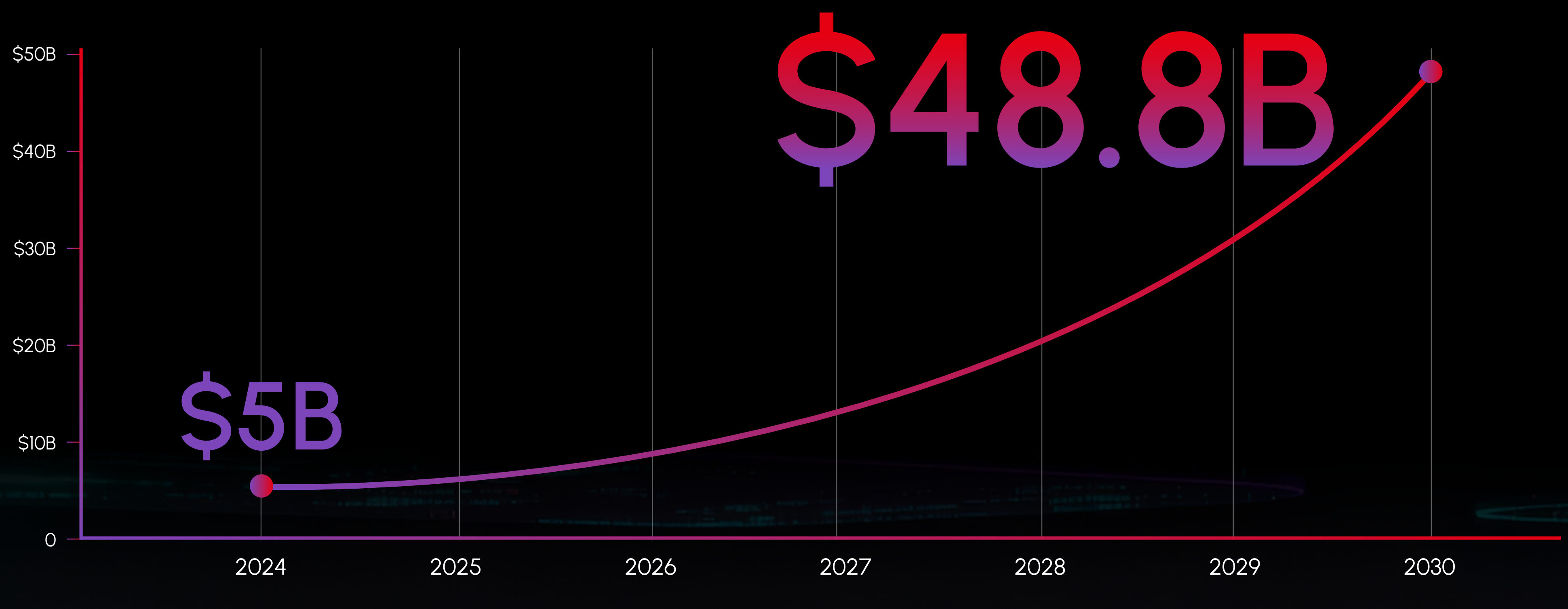
Futurum®

Lenovo

# AI Inferencing: Infrastructure Decisions That Will Shape Enterprise AI Outcomes

**Why inference—not training—is becoming the primary operational focus for enterprise AI**

## The Inference Market Is Expanding Rapidly



46.3% CAGR

Bull case: $137B by 2030

Bear case $26B by 2030

Inference is moving from pilot deployments toward broader production adoption, driving sustained infrastructure investment.

## Hybrid & Edge Are Influencing AI Deployment Architectures

65% Hybrid & Edge inference growing

vs.

46% Public Cloud

By 2030, hybrid and edge inference are projected to approach public cloud inference in overall market significance.

## Inference ≠ Training: Why Infrastructure Requirements Differ

| Training | Inference |
|---|---|
| Batch | Continuous |
| Centralized | Real-Time |
| Throughput-Optimized | Latency- and Memory-Sensitive |

Infrastructure optimized for training is often inefficient for large-scale inference workloads.

## Cost Is Emerging as a Key Competitive Metric

**Cost per Million Tokens** is an increasingly important unit of comparison

| | |
|---|---|
| Org A | $50/Million tokens |
| Org B | $25/Million tokens |

## Why Inference-Optimized Infrastructure Matters at Scale

Primary constraints affecting inference performance and efficiency:

- Memory bandwidth limits
- Latency sensitivity
- Power density & cooling considerations
- Accelerator utilization efficiency
- Operational tuning and workload placement

At scale, inference efficiency depends on both hardware choices and specialized AI infrastructure expertise; without it, utilization drops, costs rise, and time to value slows.

Lenovo services provide support across the design, deployment, and optimization of inference environments, which can help organizations manage complexity and improve production inference economics.

## Inference Infrastructure and AI Services Shape Business Outcomes

Right-sized decisions can improve efficiency and ROI; misalignment can drive higher costs and operational friction.

Read the Full Report + Research Click here!

Futurum®

Lenovo

2H 2025 AI Platforms Market Sizing & Five-Year Forecast, Futurum Research, December 2025.