Lenovo | AMD

5 Myths of AI Inference – and what actually works

Don't let these myths stand between you and real-world AI success.

Adoption of AI inference is essential for turning AI models into real outcomes.

See what modern inference really requires for success, and how Lenovo and AMD deliver optimized platforms that make it possible – from edge to data center.

Why AI inference matters now

Inference is **where AI starts delivering repeatable business value** – quickly turning trained models into live insights, automation, and decisions across a range of use cases.

Real-time decisions
Act on data as events happen

Personalized experiences
Adapt content and interactions for each user in the moment

Automation
Reduce manual effort with AI-driven processes

Predictive maintenance
Detect issues early to prevent downtime and failures

Fraud and risk detection
Detect and handle suspicious activity as it occurs

Quality inspection
Use AI vision to identify defects and ensure consistent quality at scale

Time to rethink AI inference

Myth #1:

AI inference belongs in the cloud

- Edge and on-prem inference aren't practical
- Data must travel back to a central environment for processing
- Local servers can't handle AI workloads

Reality

- AI leaders are running inference on premises and at the edge
- Processing closer to the data can reduce latency and bandwidth use
- Local CPU- and GPU-based inferencing can deliver outstanding value

Myth #2:

AI inference is too slow for real-time insight

- Inference is best for offline analysis
- Performance can't meet real-time user or operational needs

Reality

- Optimized inference delivers low-latency, real-time results
- AI can power rapid decisions and responsive user experiences

Myth #3:

AI infrastructure is rigid and hard to scale

- Scaling requires disruptive hardware changes
- You must overprovision up front for future demand
- Infrastructure can't adapt to new AI workloads

Reality

- Inference environments can scale up or out as demand grows
- You can start with CPU inference and add GPUs as needed for more

Myth #4:

Inference and training must happen together

- The same systems must handle training and inference
- All AI projects need large, training-class infrastructure

Reality

- Inferencing runs trained models efficiently - optimized to power real-time use cases like fraud detection or vision analytics
- Training builds large foundation models - LLMs and multimodal models - using massive, scale-out GPU infrastructure

Myth #5:

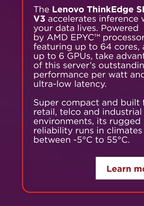
AI inference is too costly for my business

- Inference requires prohibitively expensive, specialized hardware
- Only big organizations can afford production AI

Reality

- Inference-optimized systems are designed for efficient performance and GPU, memory and network utilization
- Efficient CPUs and GPUs help minimize TCO
- On-prem inference can help eliminate consumption-based pricing and reduce cost per token

Deliver AI innovation where data lives, with Lenovo and AMD



Designed to support inference workloads from the edge to the data center, Lenovo's AI inferencing-optimized servers and infrastructure solutions powered by AMD EPYC™ processors combine speed, efficiency, and scalable AI performance.

Real-time AI insights at the edge

The **Lenovo ThinkEdge SE455i V3** accelerates inference where your data lives. Powered by AMD EPYC™ processors featuring up to 64 cores, and up to 6 GPUs, take advantage of this server's outstanding performance per watt and ultra-low latency.

Super compact and built for retail, telco and industrial environments, its rugged reliability runs in climates between -5°C to 55°C.

Learn more

Powerful and optimized for inference at scale

The **Lenovo ThinkSystem SR675I V3**, a GPU dense accelerated infrastructure that supports full AI lifecycle, scales flexibly to meet your AI inference demands – from a single node deployment to supercomputing.

With **2x AMD EPYC processors** and up to 8x full-height double-width GPUs, you can accelerate computer vision, NLP, and generative AI across industries.

Learn more

Move from AI myth to AI reality

Turn inference plans into action
with expert guidance to design and deploy infrastructure that fits your goals.

Accelerate what happens next

Lenovo | AMD