

AI performance depends on the whole system

Agentic AI puts the network at the center of infrastructure.

Benchmark real workloads to expose bottlenecks before they constrain performance, cost, or scale.

Trace the work. Find the wait.

Follow data, computation, and communication through a complete run.

DATA ACCESS

Can storage keep the workload supplied?

COMPUTATION

Are accelerators working—or waiting?

COMMUNICATION

How much time is spent exchanging and synchronizing?

Pressure reveals the limits

Change conditions. Measure workload response.

SCALE

Add accelerators. Measure useful work gained.

CONTENTION

Share resources. Measure each workload's slowdown.

DISRUPTION

Interrupt the path. Measure recovery.

Training and inference need different scorecards.

TRAINING

Goal: time to target quality.

Measure:
End-to-end training time, scaling efficiency, and accelerator utilization.

Stress test:
Collective communication, checkpointing, and recovery.

INFERENCE

Goal: throughput within latency targets.

Measure:
Tokens per second, response latency, and throughput against the service-level performance target.

Stress test:
Concurrency, latency under load, and service recovery.

Make results decision-ready

- Compare performance and cost
- Capture model, workload, and configuration
- Repeat tests. Report variability.

Take the next step

Get in touch with Cisco

Turn measured results into smarter investment, clear optimization priorities, and confidence to scale.