CoreWeave

FROM BOTTLENECKS TO BREAKTHROUGHS:

# Scale AI with Confidence

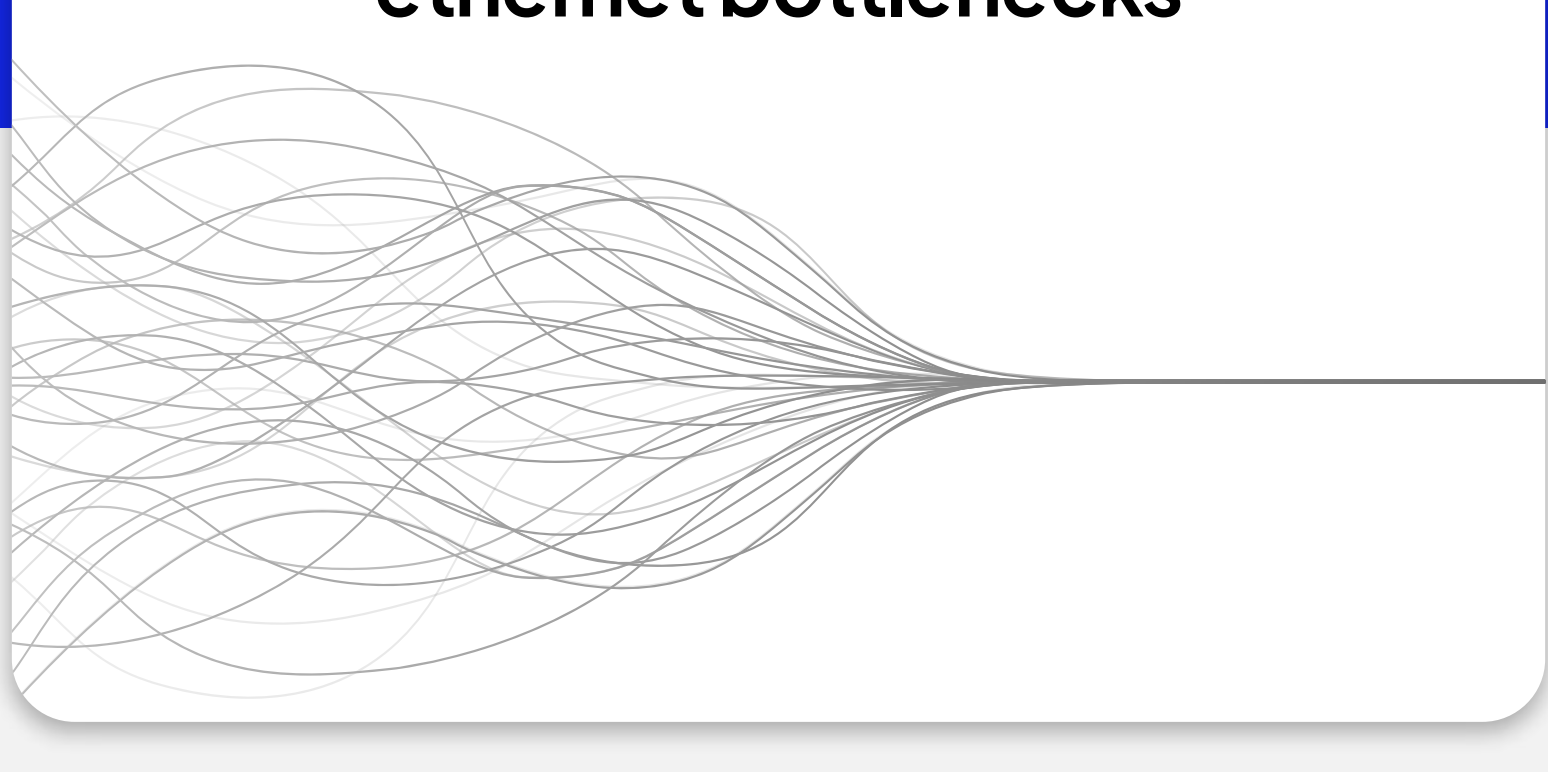
Scaling AI with predictable ROI isn't about adding more GPUs. It's about AI cloud infrastructure purpose-built to handle the pressure—not just patched together to survive it.

Here are five critical layers your infrastructure needs, and the big questions you should be asking every step of the way.

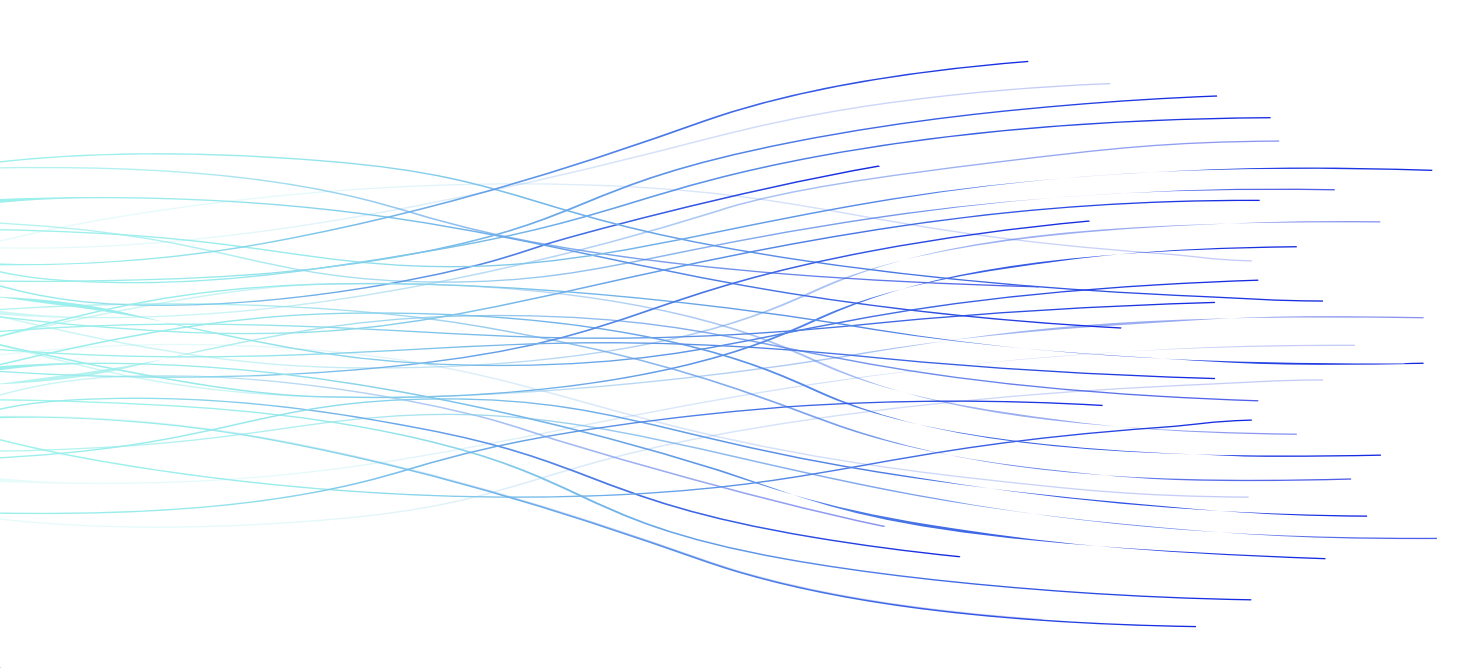
## 01 The foundation: purpose-built parallelized infrastructure

Does your network treat GPU clusters as a single supercomputer—or just individual machines?

**General-purpose cloud: ethernet bottlenecks**



**CoreWeave Cloud: free-flowing Infiniband fabric**



**What it takes:**

Non-blocking network fabric enabling any-to-any GPU communication across massive GPU clusters without contention

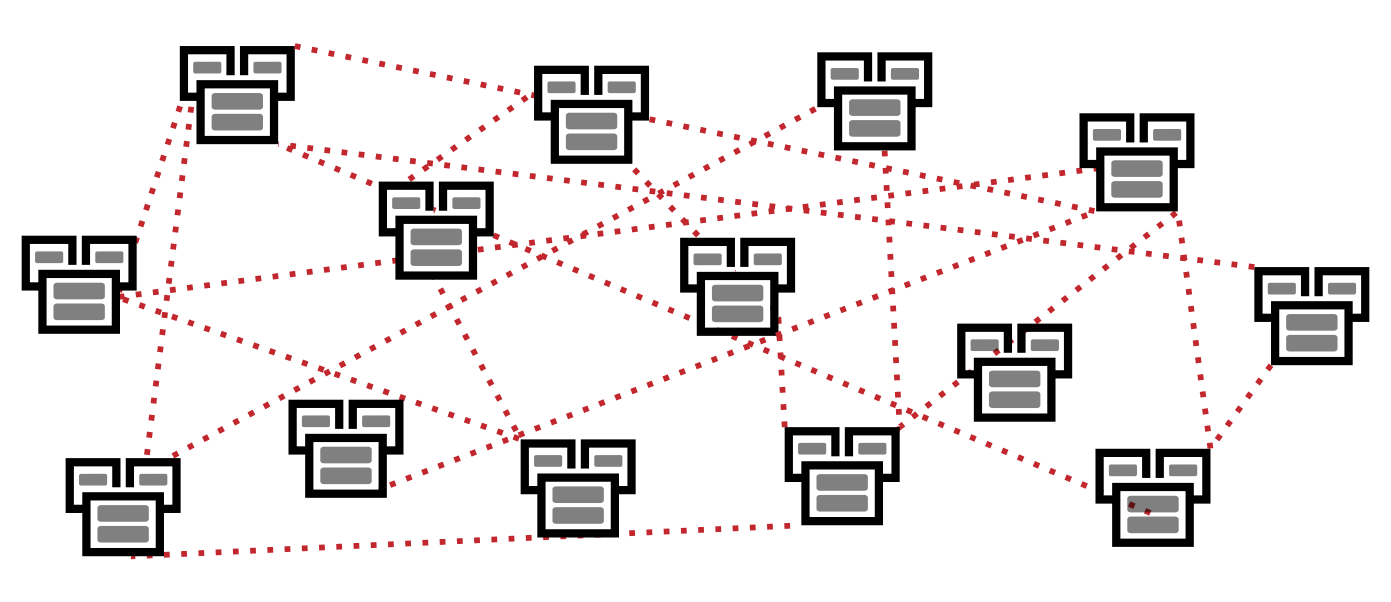
**By the numbers:**

- Up to **6,400 Gbps** per node
- **Sub-microsecond latency**
- **2× network performance** with SHARP

## 02 Intelligent orchestration and resource management

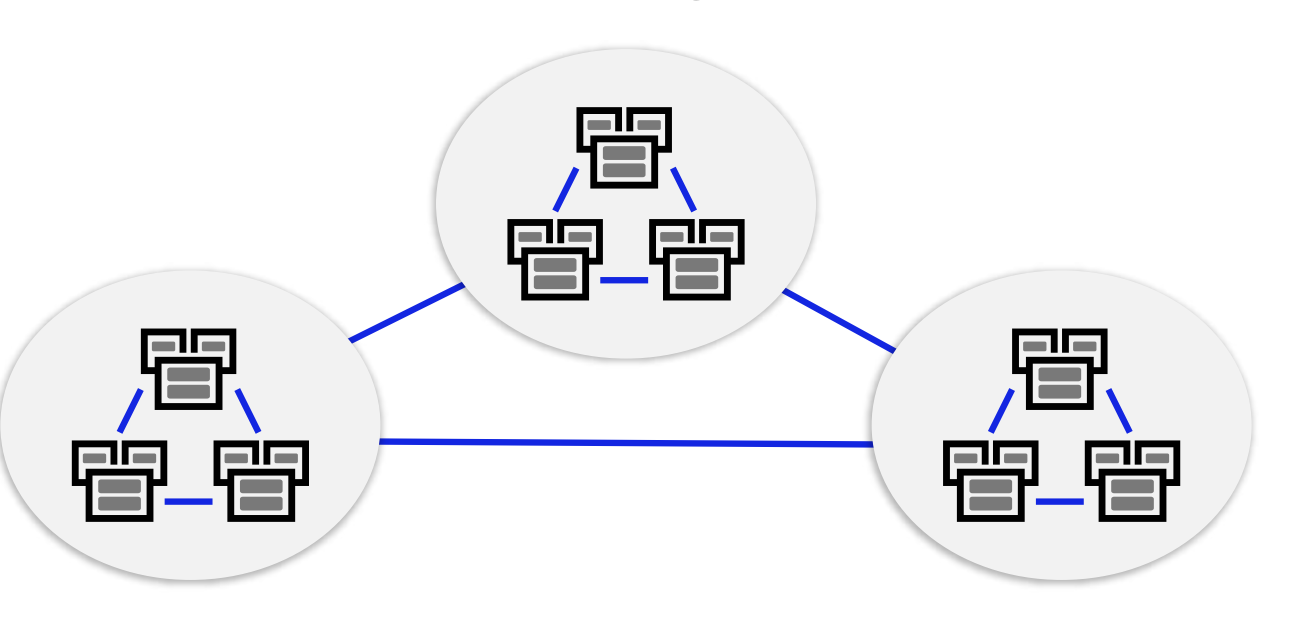
Can your orchestration pin jobs to the right GPU topology—or does it scatter workloads randomly?

**General-purpose cloud: scattered pods**



Non-topology-aware placement = higher latency

**CoreWeave Cloud: topology-aware placement for reduced fragmentation**



Topology-aware = minimal overhead

**What it takes:**

Bare-metal Kubernetes with topology-aware scheduling and gang scheduling for distributed AI workloads

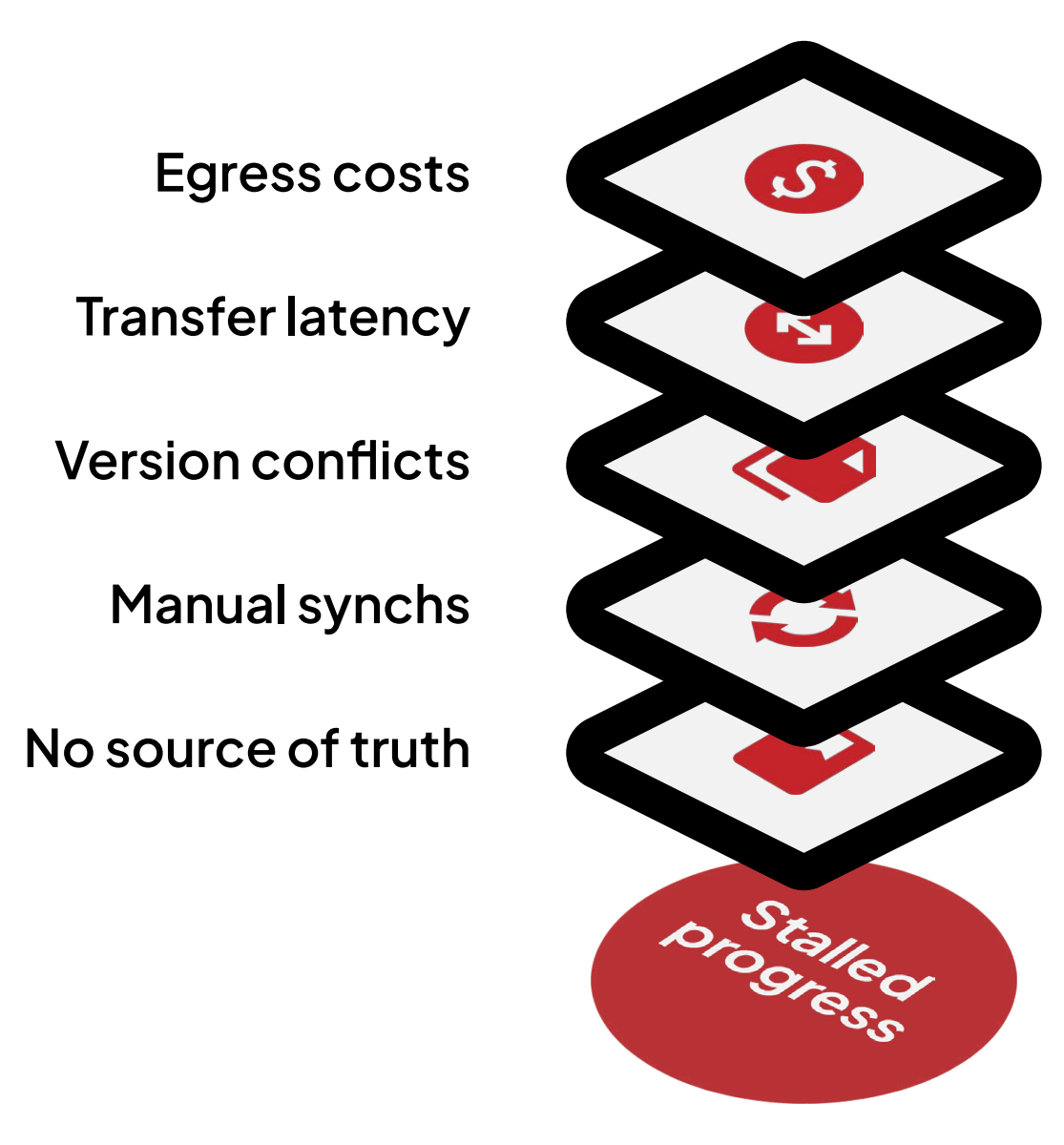
**By the numbers:**

- **50%+ MFU** (vs. 35-45% typical)
- **20% higher** cluster performance
- **96%** goodput rate (vs. 90% typical)

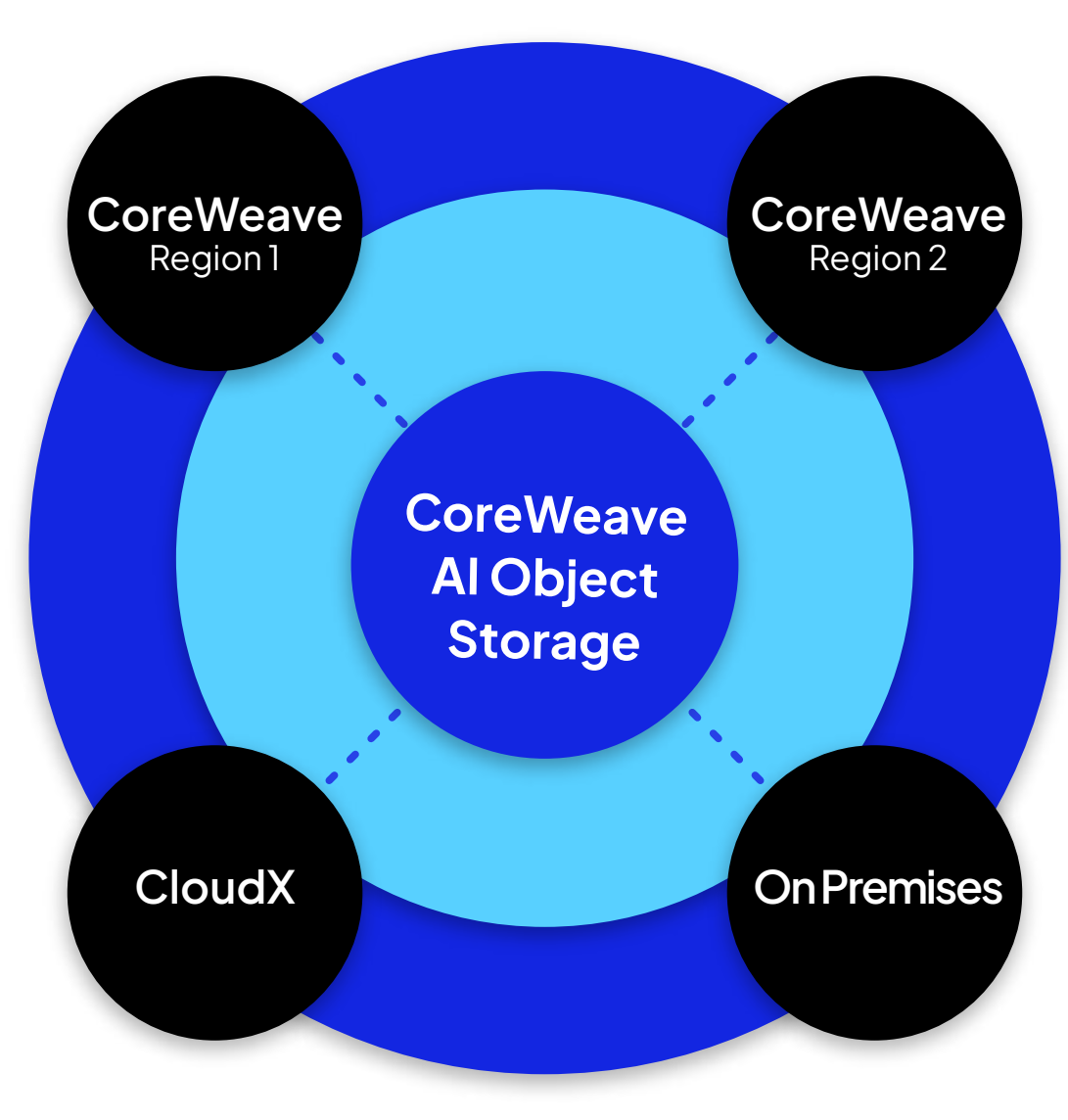
## 03 AI-native storage and data pipeline

Does your storage keep GPUs fed—or do they burn budget on idle GPUs waiting for data?

**General-purpose cloud: fragmented regional storage**



**CoreWeave Cloud: global unified storage with zero egress**



**What it takes:**

CoreWeave AI Object Storage with embedded local caching eliminates data-loading bottlenecks without requiring separate caching infrastructure—keeping GPUs continuously fed during training and fine-tuning.

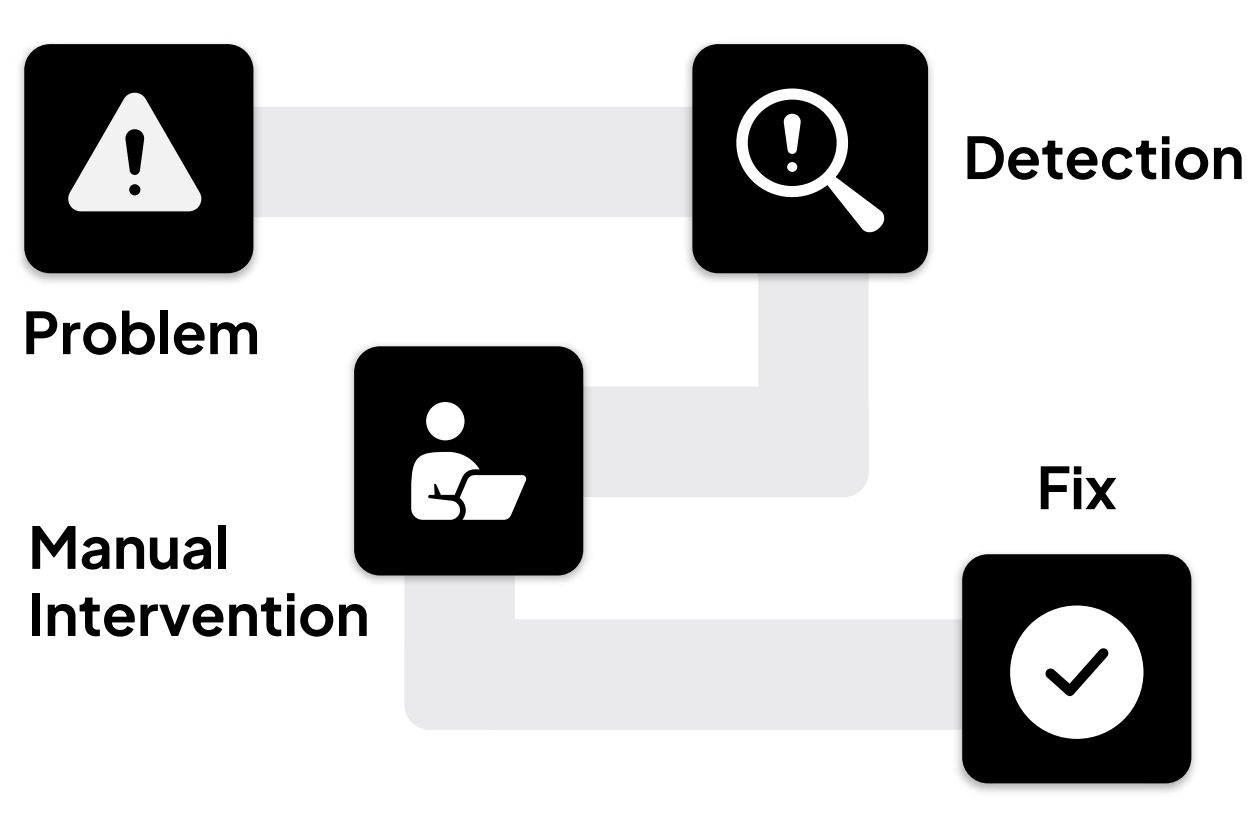
**By the numbers:**

- Up to **7 GB/s** per GPU
- **Zero egress fees**
- up to **75% lower** storage costs

## 04 Proactive observability and automated recovery

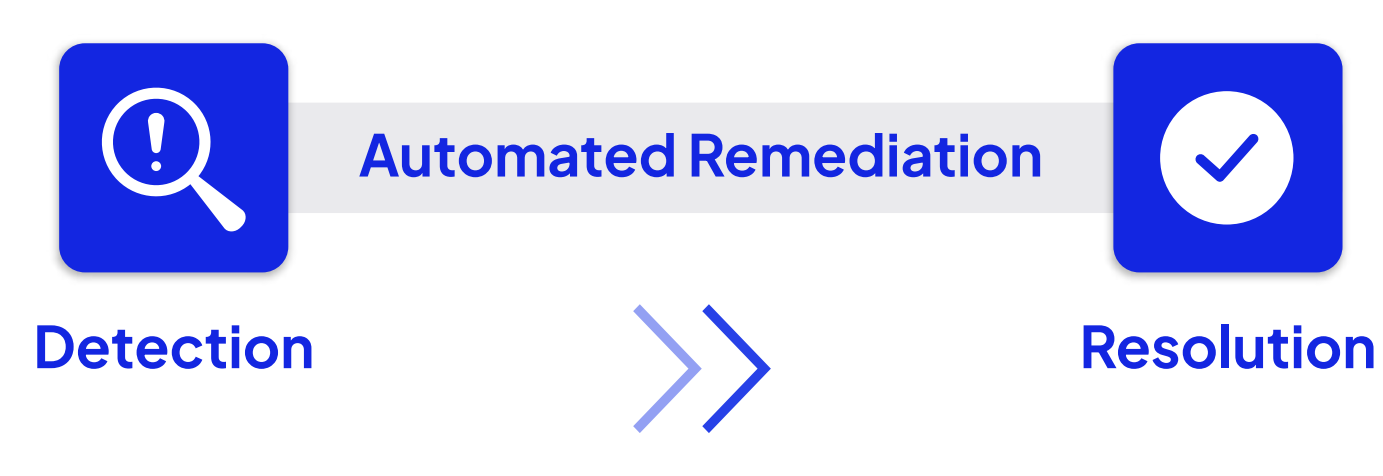
Can you see problems before they become failures—and get them fixed before the costs pile up?

**General-purpose cloud: Limited visibility, manual troubleshooting**



Limited visibility into failures
Issues discovered after job failure
Manual diagnosis and remediation

**CoreWeave Cloud: Deep visibility, rapid resolution**



Full-stack visibility from bare metal up
Proactive health checks detect issues early
Automated remediation for common failures
24/7 direct-to-expert support for complex issues

**What it takes:**

Integrated observability from bare metal to application combined with automated health monitoring and node lifecycle management (Mission Control), backed by 24/7 FleetOps engineers.

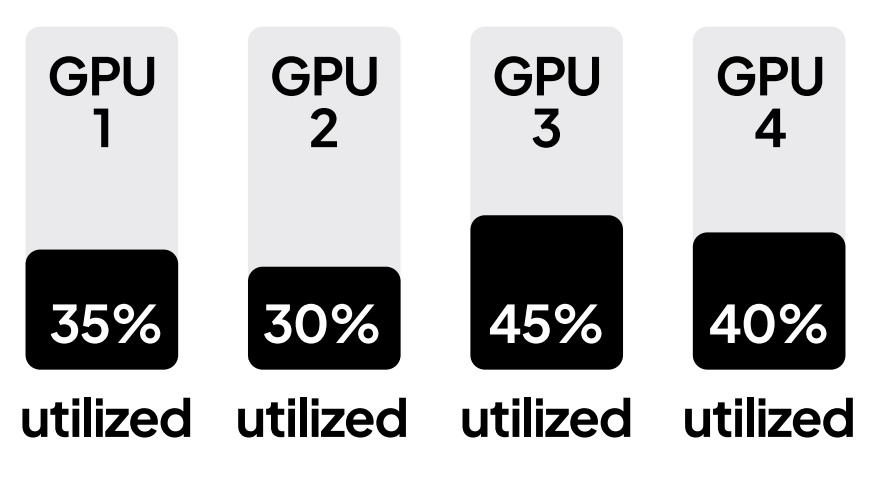
**By the numbers:**

- **50% fewer** interruptions
- **10× longer uptime** (MTTF 3.66 vs. 0.33 days)
- **3× faster** recovery

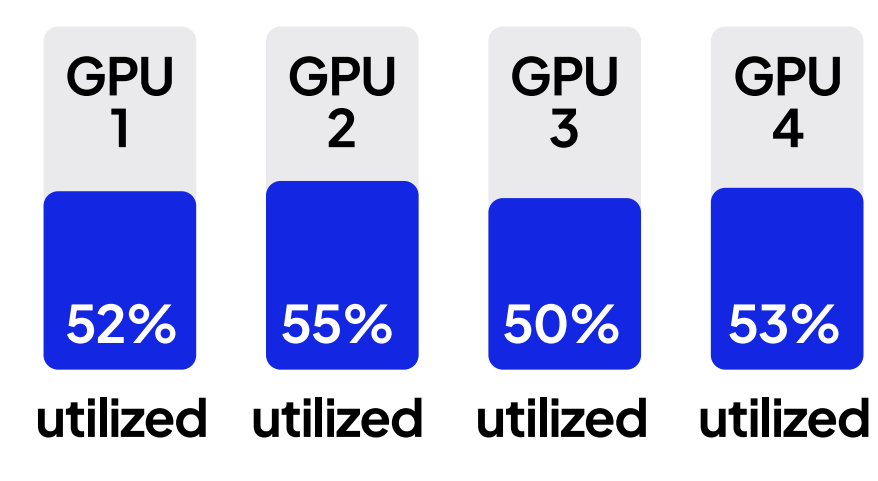
## 05 Continuous workload optimization and elastic scaling

Does your infrastructure extract maximum performance from every GPU—or leave efficiency on the table?

**General-purpose cloud: 35-45% MFU Provisioning delays Inefficient scaling**



**CoreWeave Cloud: 50%+ MFU Instant provisioning**



**What it takes:**

Full-stack optimization—from bare-metal Kubernetes and topology-aware scheduling to rapid model loading and AI-optimized storage—engineered to maximize MFU and minimize wasted GPU time.

**By the numbers:**

- Up to **20% higher** GPU cluster performance
- **96%** goodput = maximum ROI from every GPU hour
- **Faster** time-to-train » **Faster** innovation

## Don't settle for infrastructure that wasn't purpose-built for AI

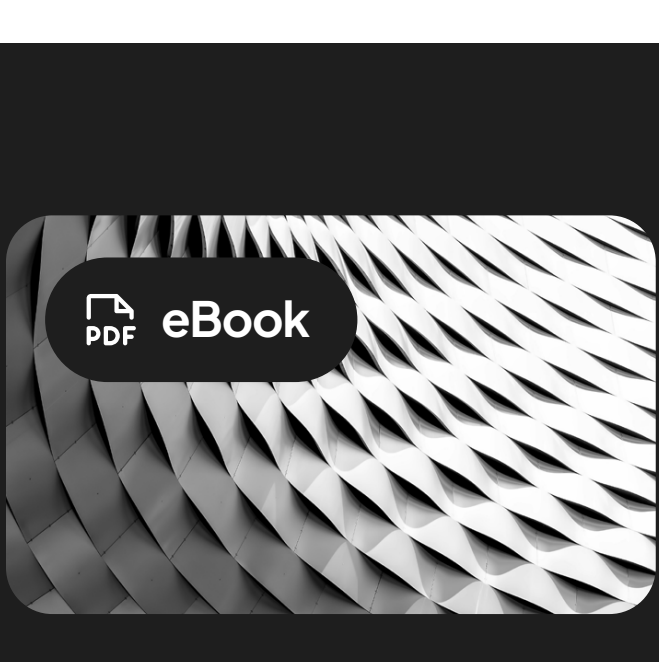
Answering yes to each of the five questions above isn't optional—they're requirements for scaling AI with predictable ROI. CoreWeave is the only cloud purpose-built to deliver on all of them with the industry's first operating standard for running AI at scale.

**Learn more about why the world's leading AI labs choose CoreWeave to train, fine-tune, and deploy at scale.**

## Take the next step

Scale your AI workloads with transparency, industry-leading performance, and confidence on CoreWeave Cloud

Contact Us

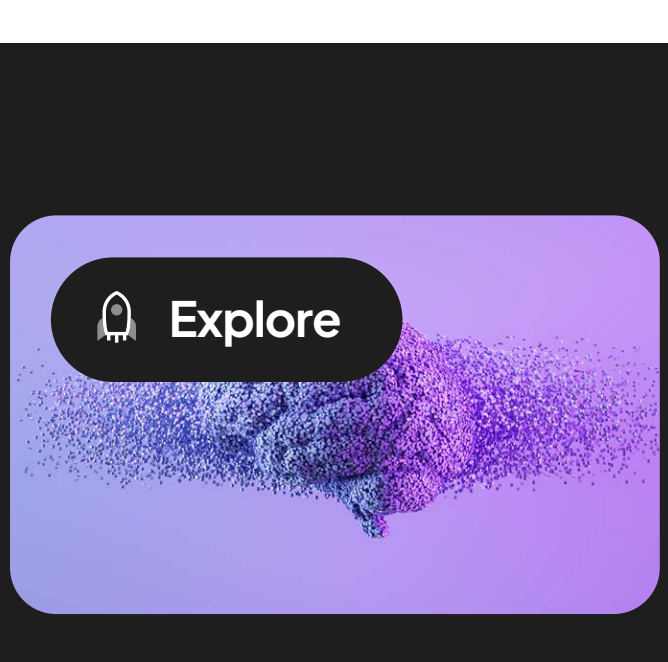


eBook

**Clarity, unlocked**

Five Surprises that Derail AI Scale-And How To Build with Confidence

**Read more →**



Explore

**How IBM boosted performance by 80% on CoreWeave**

Read the case study

**Learn more →**

Explore

**CoreWeave Mission Control: The Operating Standard for the AI Cloud**

Read the solution brief

**Learn more →**

CoreWeave WWW.COREWEAVE.COM

© Copyright CoreWeave 2025. All rights reserved. CoreWeave, its logo, and coreweave.com are trademarks of CoreWeave.
This informationis provided "as is" without any warranty. This document is current as of the initial date of publication and may be changed at any time.