EBOOK

BUYER'S GUIDE

CoreWeave

The Technical Buyer's Guide to Scaling AI

Find the right infrastructure partner to go from model development to deployment-ready

Contents

00 Executive Summary

The challenges of scaling AI workloads

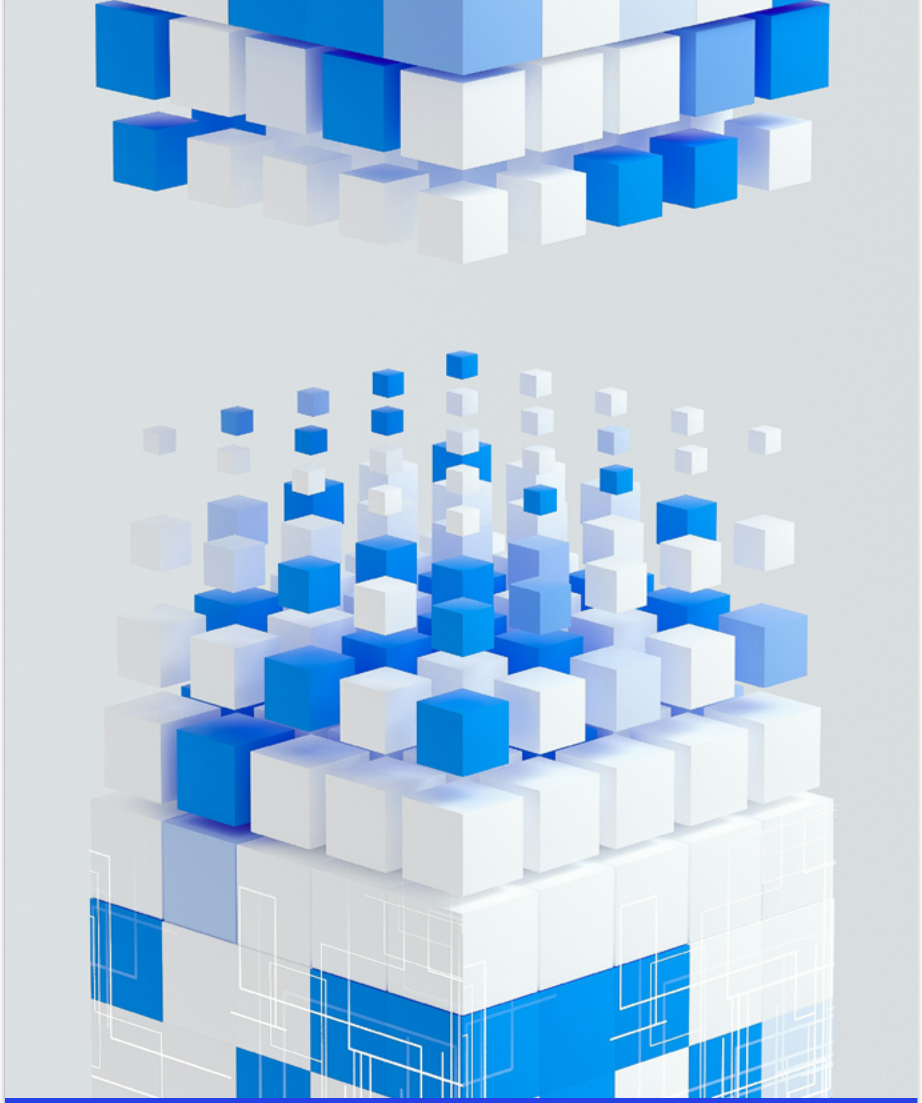
AI workloads are built different. That's why you need an AI cloud purpose-built to support them.

Whether you're expanding your AI startup into new verticals or working in an enterprise to get your AI innovations to market faster, scaling up to be deployment-ready isn't easy. You need the right cloud partner.

AI workloads push infrastructure harder than any traditional application: multi-week training runs, petabyte-scale datasets, and distributed systems that must stay online without degradation.

If your cloud was built for web apps, your models sit waiting for capacity instead of training. If your cloud treats AI like any other workload, you'll have a resiliency problem that stops jobs in their tracks. And if your cloud keeps you in the dark about performance and utilization, you can't optimize AI workloads at scale.

Choosing the right infrastructure partner is critical. This guide will equip you with five key questions to ask and the metrics to measure to ensure your AI cloud accelerates your breakthroughs, not your bottlenecks.



Learn how CoreWeave helped Mistral scale without compromise

Read the case study →

01 Scaling resiliency

Can it handle my workloads resiliently at scale?

A cloud built for AI provides lower TCO and competitive edge as you scale from experimentation to production.

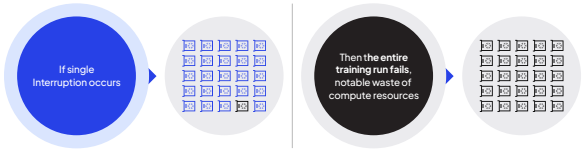
Traditional clouds are built for web apps, not the massive parallelism required by AI workloads. This changes the conversation around resiliency.

Web scale applications

CPUs are designed for graceful degradation. In scaled-out cloud environments, if a processor fails, workloads running on it are reallocated to other nodes using techniques such as live migration. Performance takes a hit, but it's small.

But when you run a training job on a large GPU cluster, that job is split up across many thousands of GPUs that work simultaneously. When a single GPU fails, the training job fails.

Training AI



This means a single failure mid-run can erase days of valuable compute and burn through your budget fast.

Any cloud solution that will handle AI workloads must be designed with unique requirements around managing for hardware events, as they have a tangible impact on success. Training or fine-tuning on traditional clouds built for web apps and retrofitted for AI workloads leads to delays, unexpected costs, and GPUs sitting idle.

The CoreWeave effect

CoreWeave Cloud was purpose-built for AI workloads. Our infrastructure delivers enterprise-grade resiliency through failure-tolerant hardware so you can scale your AI workloads with confidence.

With 96% goodput and 20% higher model FLOPS utilization (MFU) than industry peers, workloads run efficiently without wasted cycles or unexpected interruptions. CoreWeave's platform is engineered for elasticity across thousands of GPUs with automatic fault isolation, redundancy, and proactive monitoring.

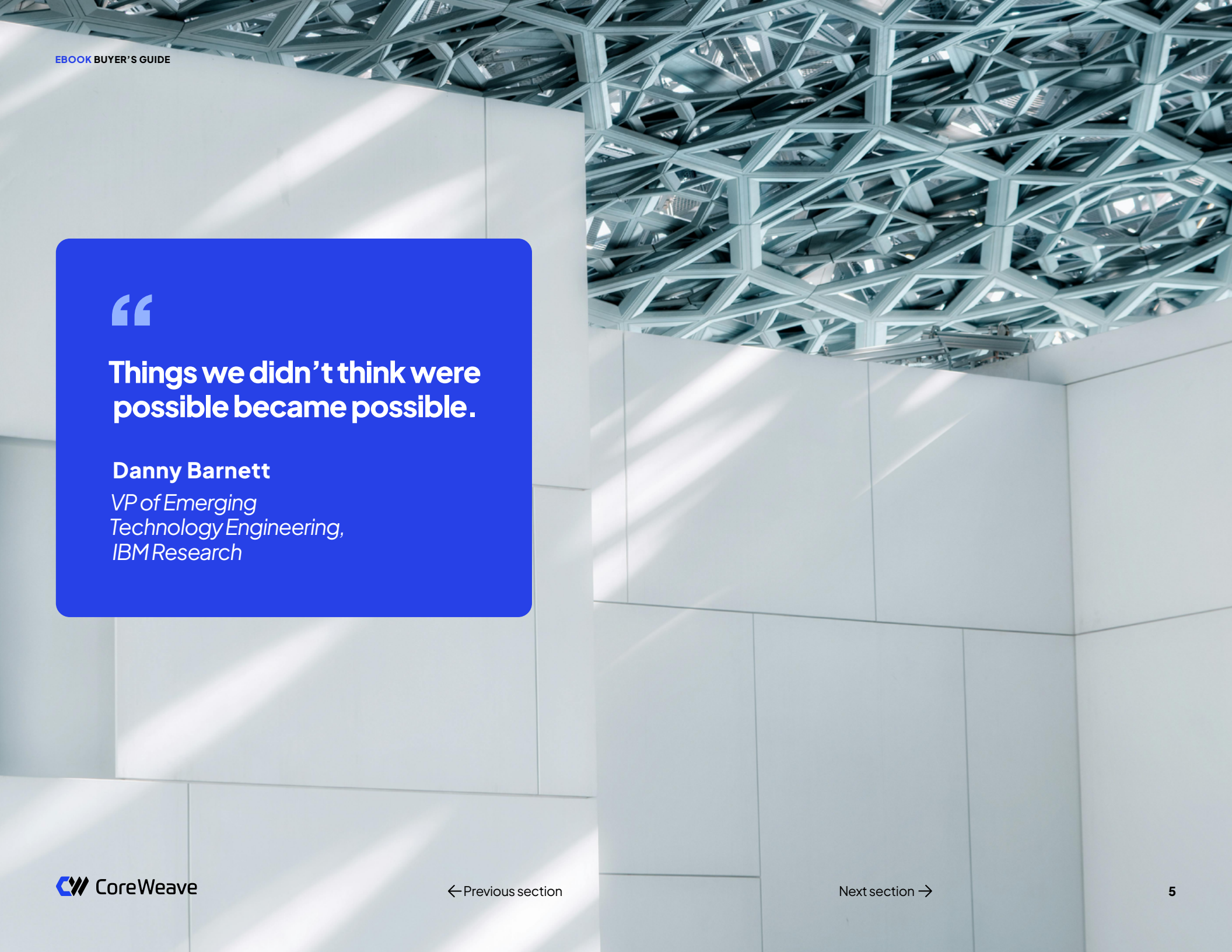
96% MFU
goodput

20% higher
model FLOPs utilization

50% fewer
Interruptions per day

Learn how IBM boosted training speed by 80% with CoreWeave

Read more →



> **Things we didn't think were possible became possible.**
>
> **Danny Barnett**
> *VP of Emerging Technology Engineering, IBM Research*

02 Operational efficiency

How big is the operational burden for my team?

Make sure your team spends minimal time running the infrastructure and maximum time building AI.

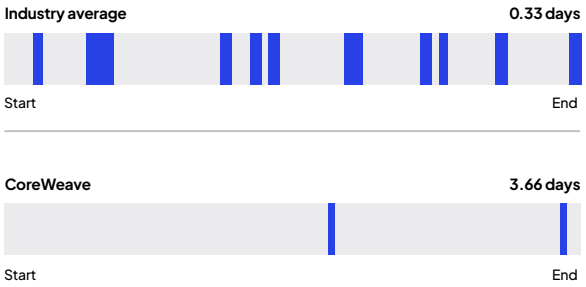
Every hour engineers spend managing infrastructure is an hour they're not training, fine-tuning, or deploying models. The wrong cloud shifts operational complexity onto your team, forcing you to script around provisioning delays, debug opaque performance issues, or scale processes manually. This drains velocity and limits how fast new ideas can move from experiment to production.

Architects need to ask whether a cloud provider simplifies or complicates operations at scale. Purpose-built AI clouds integrate orchestration, observability, and proactive cluster health checks into a unified environment, so your team can focus on innovation, not firefighting.

Choosing a provider that minimizes operational drag doesn't just improve efficiency—it compounds ROI. Less time managing infrastructure means faster iteration, more reliable pipelines, and a team free to build the innovations that actually drive value for the business.

Mean time to failure

Mean time to failure, industry average vs. CoreWeave[1 2]



Data for a training run of 1,024 NVIDIA H100 GPUs

The CoreWeave effect

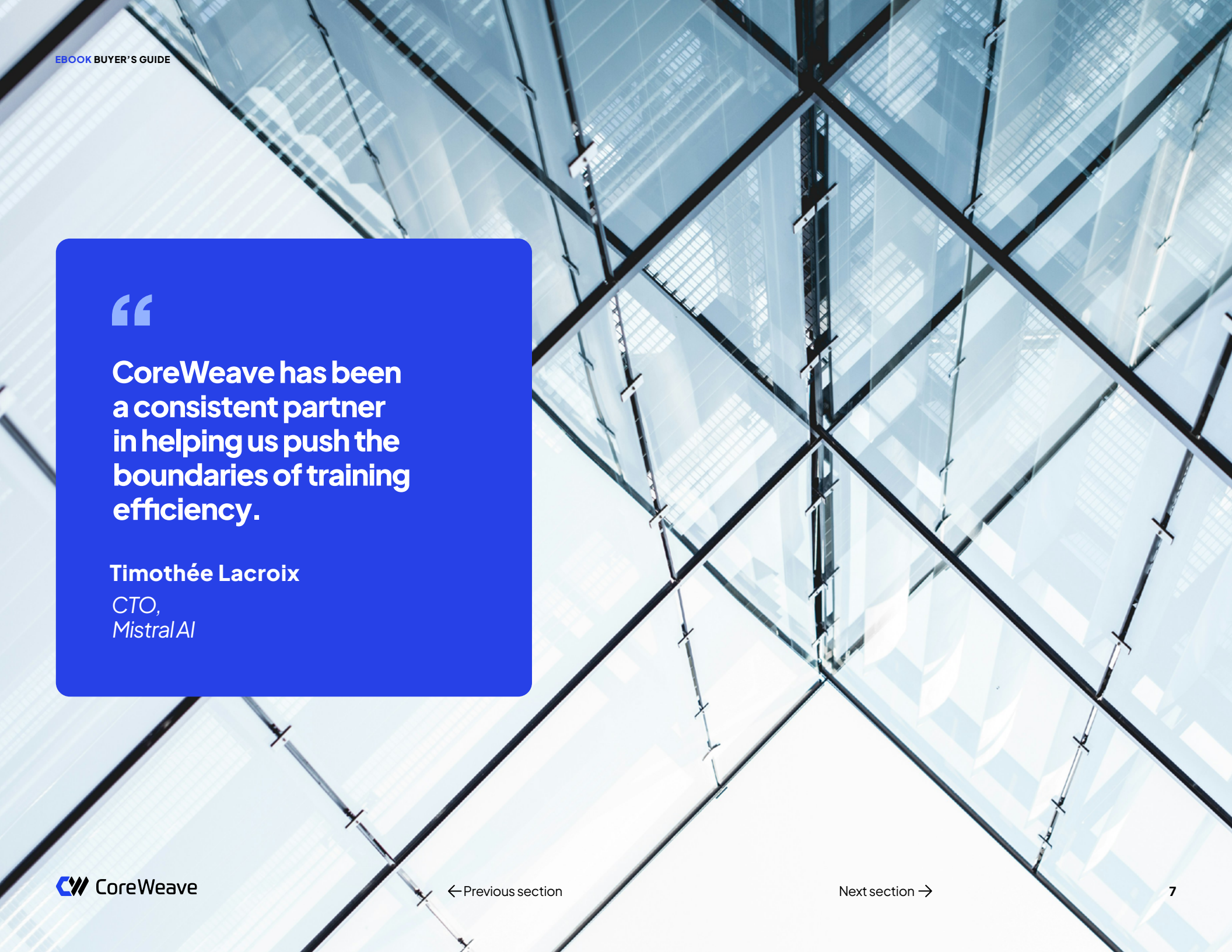
Mission Control is CoreWeave's orchestration and observability layer, purpose-built for AI-scale infrastructure. Instead of forcing teams to manage separate systems for scheduling, monitoring, and scaling, Mission Control unifies everything into a single operational pane of glass.

For cloud architects and DevOps engineers, that means less time coordinating resources and more time building. Mission Control automates cluster management, optimizes workload placement, and surfaces performance telemetry in real time—from GPU utilization to cost efficiency—without additional tooling or integration work.

This unified control plane turns infrastructure management into a high-signal, low-effort process. You can provision, tune, and scale massive GPU environments with precision, while the platform continuously optimizes for throughput, reliability, and cost. It's built to minimize the operational friction that slows down AI development, so you can move as fast as your ideas

Unlock higher performance and usage for faster time to market.

Discover Mission Control →



> **CoreWeave has been a consistent partner in helping us push the boundaries of training efficiency.**
>
> **Timothée Lacroix**
> *CTO,*
> *Mistral AI*

03 AI-native storage

How does storage help optimize for efficiency and performance?

Storage is an active performance layer that determines whether your AI environment scales smoothly or stalls under pressure.

Storage isn't just a backend component—it's foundational to AI performance, scalability, and reliability.

AI workloads depend on high-throughput, low-latency data pipelines that can sustain terabytes per second across distributed systems. Storage must be designed for parallel access, massive concurrency, and dynamic scaling. Legacy file systems built for transactional workloads severely limit your ability to experiment and scale.

Reliable storage also underpins fault tolerance and reproducibility—critical when training runs span thousands of GPUs and weeks of runtime. Checkpointing and snapshotting must be instantaneous and durable to prevent costly retraining after a single node failure.

The CoreWeave effect

CoreWeave Cloud provides high-throughput object and distributed file storage designed for reliability and scale that eliminates AI data bottlenecks. CoreWeave AI Object Storage, launched in October 2025, eliminates the friction of moving data between regions, clouds, and tiers. Powered by our Local Object Transport Accelerator (LOTA) technology, it combines simplicity, scalability, and transparency, offering throughput of up to 7 GB/s per GPU, zero fees (egress, ingress, or request), and automated usage based billing levels that cut existing customers' costs by more than 75%.

7 GB/s
throughput per GPU

Zero
egress, ingress, or request fees

75% lower
costs for existing customers

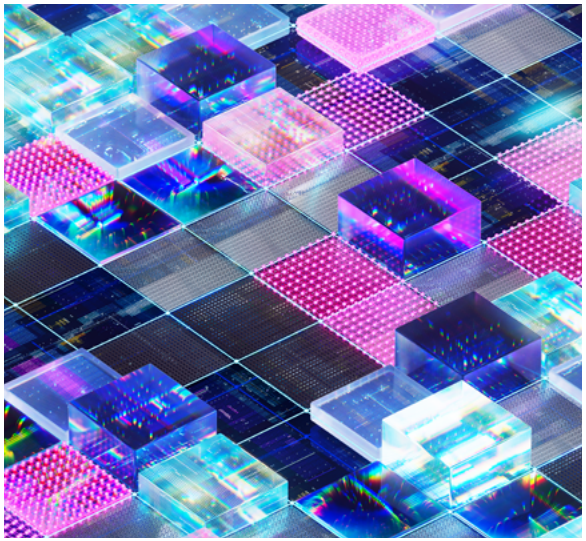
Take a deep dive into AI Object Storage.

Explore storage without limits →

04 Security at scale

Are my workloads secure, private, and compliant?

Why AI at scale requires an infrastructure that embeds isolation, visibility, and compliance directly into the platform.



As AI moves from experimentation to production, the attack surface expands dramatically. You'll need to protect not just applications, but the models, data pipelines, and orchestration layers that power them. Training data may contain sensitive information, models can be reverse-engineered, and inference endpoints become high-value targets for data exfiltration or prompt injection.

Scaling exacerbates these risks. Distributed systems multiply points of failure—across GPUs, storage clusters, and APIs—while complex toolchains introduce vulnerabilities through open-source dependencies, container images, or unmanaged access credentials. Traditional cloud security models, designed for stateless web apps, often lack the granular controls, isolation, and auditability AI workloads demand.

Finally, compliance and governance expectations rise sharply at production scale. Teams must enforce data lineage, encryption, and access control without sacrificing performance or velocity. Balancing security with scalability becomes a balancing act: lock too tightly, and you lose agility; move too fast, and you risk exposure.

The CoreWeave effect

CoreWeave's platform was designed for AI from the ground up, and that includes security. Every layer—from the data pipeline to the network fabric to user access—is engineered for protection and performance.

Data security. All data is encrypted in transit and at rest, and customers can even manage their own keys. Immutable logging pipelines give complete traceability. In addition, CoreWeave AI Object Storage features encryption in transit.

Network isolation. Each compute node has its own NVIDIA BlueField-3 DPU that enforces hardware-level isolation—real physical separation between tenants.

Identity and access. Fine-grained IAM and zero-trust patterns ensure users get only what they need. Automated User Provisioning (AUP) federates identity from Okta or Microsoft Entra into CoreWeave IAM, and SUNK User Provisioning (SUP) automates Slurm user provisioning for instant, secure access across clusters.

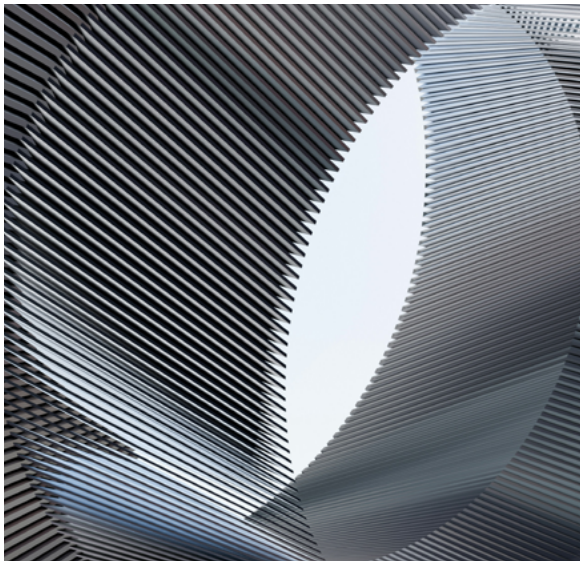
See why we lead the industry on security and privacy.

Discover CoreWeave security →

05 Partnership and support

Do I have the right partner to help me scale?

The right AI cloud is more than a provider—it's a consultative partner and a critical extension of your engineering team.



Infrastructure isn't static—it's a living system that evolves with every model, framework, and architecture update. Engineers and architects need a cloud partner that evolves just as fast. A strong partner anticipates bottlenecks, shares best practices, and co-engineers solutions that sustain performance as workloads grow.

True partnership means access to deep architectural collaboration: tuning workloads for optimal utilization, aligning network topology with training patterns, and refining orchestration strategies to match how your teams actually build and deploy. It's about solving hard, system-level challenges together, not filing tickets into a neverending queue.

No team scaling AI succeeds in a vacuum. A cloud provider that acts as a strategic extension of your engineering organization—with shared visibility, aligned incentives, and continuous optimization—becomes a force multiplier.

The CoreWeave effect

CoreWeave was uniquely designed for teams building at the frontier edge. Partnership here means deep collaboration between your architecture and ours. CoreWeave moves first on next-generation GPU architectures, interconnects, and liquid-cooled systems, so your models never wait on infrastructure maturity. You build faster because we build ahead.

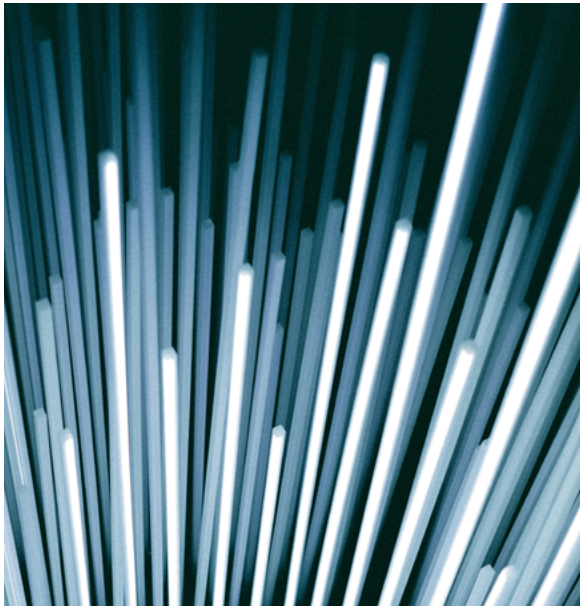
We provide deep consultative expertise and direct-to-expert technical support to troubleshoot issues, optimize performance, and fine-tune configurations for AI workloads.

CoreWeave acts as an extension of your engineering team: a partner that shares your performance goals, accelerates your iteration loops, and scales with you as your ambitions grow.

06 Action guide

Five key questions, summarized

A quick hit of the five key questions you need answered before you choose an AI cloud provider



01

Can it handle my workloads resiliently at scale?

CoreWeave Cloud delivers enterprise-grade resiliency so you can scale your workloads with confidence.

02

How big is the operational burden for my team?

CoreWeave Mission Control minimizes the operational friction of AI development.

03

How does storage help optimize for efficiency and performance?

CoreWeave Cloud's high-throughput object and distributed file storage eliminates AI data bottlenecks.

04

Is it secure?

CoreWeave designs for protection and performance, from the data pipeline to the network fabric to user access.

05

Is it the right partner to help me scale?

CoreWeave acts as an extension of your engineering team and a partner that shares your performance goals.

07 Next steps

MLPerf V5.0 results

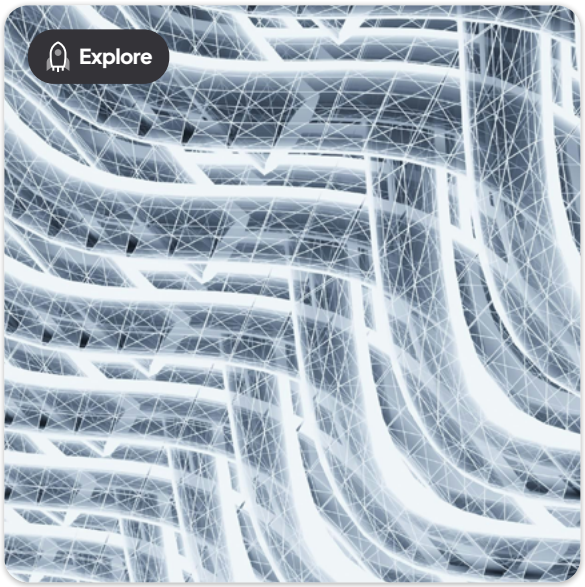
See why the CoreWeave Cloud leads MLPerf benchmarks across both training and inference performance.

Download →

Achieving 20% higher MFU

Read the technical report on how CoreWeave achieved 20% higher MFU and 10x reliability on NVIDIA Hopper GPUs.

Read →



Why CoreWeave Cloud

CoreWeave Cloud is the force multiplier for AI development, offering the pace, performance, and partnership you need to accelerate AI innovation.

Read →

Notes and references

01 Kokolis, Apostolos et al. Revisiting Reliability in Large-Scale Machine Learning Research Clusters. arXiv, February 2025. https://arxiv.org/abs/2410.21680v1

02 Brown, Wes et al. Purpose-Built Cloud for AI at Scale: Achieving 20% Higher MFU and 10x Reliability on Thousand-GPU Clusters. CoreWeave, August 2025. https://cdn.prod.website-files.com/62bc66d283fd9c34ffec780a/689fa33a0e99f19c11c8ecbe_CoreWeave%20Training%20Benchmarks%20Whitepaper%20August%202025.pdf com/62bc66d283fd9c34ffec780a/689fa33a0e99f19c11c8ecbe_CoreWeave%20Training%20Benchmarks%20Whitepaper%20August%202025.pdf

CoreWeave

CoreWeave

The Technical Buyer's Guide to Scaling AI

Find the right infrastructure partner to go from model development to deployment-ready

Contact us

WWW.COREWEAVE.COM

290 W Mt Pleasant Ave #4100,
Livingston, NJ 07039

December 2025