CoreWeave

Drive AI Innovation at Scale

An enterprise's guide to unlocking platinum AI cloud performance, based on the SemiAnalysis ClusterMAX™ Report

May 2025

Harness the full power of your AI Infrastructure.

Launching your AI products to market relies on so much more than access to GPUs in the cloud. Finding the right partner for your AI infrastructure needs is critical to accelerating your AI roadmap.

However, enterprises today have over 100 AI clouds to consider as partners in bringing their innovations to market. To cut through the noise, SemiAnalysis launched the ClusterMAX report™—the world's first independent GPU-cloud benchmark based on real GPU renters and AI-cloud experts—setting transparent, data-driven standards for the market. This inaugural report crowned CoreWeave as the first and only Platinum-Tier GPU Cloud Provider, the highest tier.[1]

By the end of this report, you'll have a clear understanding of how CoreWeave achieved this ranking as the top AI cloud provider and what it means for your AI ambitions. We've broken down the many topics outlined in the ClusterMAX report into three categories: performance, innovation, and enterprise readiness.

Cloud	Performance	Innovation	Enterprise Readiness	ClusterMAX™ Tier
CoreWeave	★★★★★	★★★★★	★★★★	Platinum
Azure	★★★★	★★★	★★★★	Gold
Oracle	★★★★	★★	★★★★	Gold
AWS	★★★	★★	★★★★	Silver
GCP	★★★	★★	★★★★	Bronze

Figure 1: This table offers a snapshot of five industry-leading AI clouds, based on rankings from the SemiAnalysis ClusterMAX™ Report and analysis from that research.

[1] The GPU Cloud ClusterMAX™ Rating System | How to Rent GPUs, SemiAnalysis, March 26, 2025.

Section 01:

Performance

Maximize goodput and cluster efficiency at scale for industry-leading AI performance.

In AI workloads, true performance isn't just raw TFLOP/s—it's goodput, the useful work your models actually complete. Platinum AI clouds marry best-in-class hardware with smart orchestration so jobs run faster, more reliably, and at scale. Look for:

- Reliable operation of 100K+ NVIDIA Blackwell GPU clusters at hyperscaler scale
- Higher Model FLOPs Utilization (MFU) for training workloads via topology-aware scheduling
- High-bandwidth NVIDIA Quantum InfiniBand with SHARP in-network reductions

Job dashboards expose utilization-adjusted FLOPS metrics in real time, so teams can pinpoint inefficiencies and respond faster. Proactive topology-aware scheduling further increases MFU by aligning job placement to the physical rack layout, driving up resource efficiency as you scale.

Validation of NCCL/RCCL networking performance is vital for maximizing training and inference performance. Only three cloud providers in the ClusterMAX™ rating that have correctly set up SHARP on InfiniBand for in-network reductions, offloading collective operations from GPUs into the network.

In fact, many companies and AI clouds struggle to scale seamlessly from proofs of concept to massive GPU fleets.

The CoreWeave Advantage

CoreWeave optimizes for every layer—from network fabrics to schedulers to hardware—to maximize throughput, minimize interruptions, and deliver consistent, predictable performance.

We consistently observe 50% fewer job interruptions when running GPU clusters of over 1K+ nodes.[2] By minimizing interruptions and recovering faster, we can help clients achieve a goodput rate as high as 96% versus industry average of 90% on clusters of 4K and 15K.[3]

Lower interruptions, faster recovery and higher goodput ultimately result in faster training times and several millions of dollars in saved costs.

higher GPU-cluster performance than alternative solutions [2] — **20%**

system goodput versus 90% industry average [3] — **96%**

fewer interruptions on 1,000+ GPU clusters [2] — **50%**

[2] CoreWeave Leads the Charge in AI Infrastructure Efficiency, with up to 20% Higher GPU Cluster Performance than Alternative Solutions, March 19, 2025. These results are for clusters of 4K and 15K and should not be extrapolated to other cluster sizes.
[3] Achieve AI Infrastructure Goodput of up to 96% with 3 Key Strategies, March 27, 2025. These results are for clusters of 4K and 15K and should not be extrapolated to other cluster sizes.

Section 02:

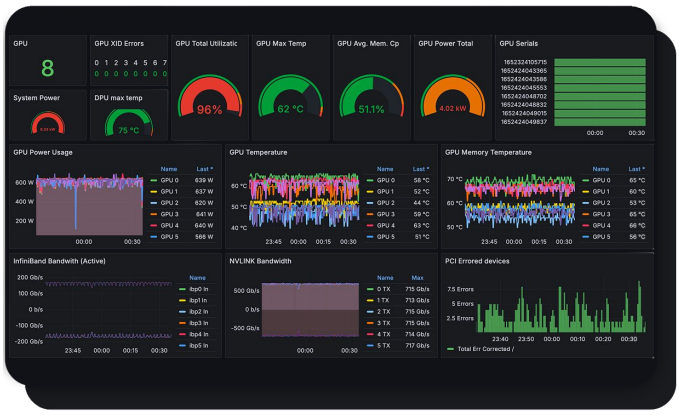
Innovation

Accelerate your AI roadmap with infrastructure purpose-built for next-gen AI workloads.

Cutting-edge AI demands bleeding-edge infrastructure. Platinum AI clouds turn advanced research into production reality by rolling out features that unlock new use cases and slash development cycles.

- Early access to the latest GPUs, including NVIDIA Blackwell now generally available
- In-network SHARP and topology-aware scheduling
- Container-native Slurm & Kubernetes integration
- Automated NCCL tuning and health-check integrations

Staying ahead of hardware launches means you can prototype on the latest GPU architectures sooner and get to market faster. This also demands that cloud providers hand over a Day-1 deployment-ready cluster that's been rigorously tested and is ready to operate at peak efficiency.



CoreWeave dashboards offer detailed metrics and data available for your clusters out-of-the box.

Innovators in the AI cloud space continuously push the industry forward with hardware and software solutions that help improve customers overall performance and experience. Offloading collective operations with SHARP in the switches and transport tasks to RDMA-aware NICs enables predictable multi-GPU scaling so large-scale training jobs behave more predictably. NVIDIA BlueField DPUs enable greater security and compute performance by offloading networking, security, and storage management tasks from GPUs and CPUs.[1]

Software innovations like turnkey orchestration plugins—whether you prefer Slurm, Kubernetes, or both—streamline DevOps handoffs and let you onboard new teams quickly. Built-in performance tests and alerts proactively catch misconfigurations before they derail production workloads.

The CoreWeave Advantage

Today, few clouds have mastered all these optimizations, and even fewer have done so with operable, large-scale clusters. CoreWeave stands out here and has continually been first to market for many of the latest GPU architectures.

In addition, CoreWeave Mission Control provides advanced cluster validation, health monitoring, proactive node replacement, and deep observability. These capabilities help ensure your workloads run on healthy infrastructure, significantly reducing the likelihood of disruptions and helping improve goodput.

[1] The GPU Cloud ClusterMAX™ Rating System | How to Rent GPUs, SemiAnalysis, March 26, 2025.

Section 03:

Enterprise Readiness

Combine hyperscaler-grade security with the agility and technical expertise of a specialized provider.

Being enterprise-ready means a cloud platform not only delivers top-end performance, but also meets the strictest corporate and regulatory demands: protecting your data, guaranteeing uptime, and enforcing governance at every layer. This looks like:

- SOC 2, ISO 27001, FedRAMP, HIPAA, and CSA STAR Level 1 certifications
- DPU, VLAN, and InfiniBand partition-key tenant isolation
- Automated burn-in tests, continuous passive telemetry, and weekly diagnostics
- 99.9%+ server level agreements (SLAs) backed by 24/7 global support

For regulated industries and mission-critical applications, only Platinum-tier clouds combine hyperscaler-grade controls and security with the agility of a specialist provider.

Rigorous third-party audits and certifications prove that your data and IP are protected at every layer from hardware to hypervisor. SOC 2 and ISO 27001 are non-negotiables; enterprises should expect event greater security measures throughout the stack.

Isolation is enforced in hardware, not just in software. Hardware-enforced isolation (DPUs and VLANs) ensures tenant traffic never crosses boundaries—eliminating "noisy neighbor" risks and data-leakage concerns. Meanwhile, continuous health-checks catch drifting network or hardware issues before they impact jobs.

When outages do occur, robust SLAs and a dedicated, around-the-clock response team get you back online fast. Cross-region snapshots and automated failover paths guard against data loss, giving your compliance and risk teams peace of mind.

The CoreWeave Advantage

CoreWeave sets a high standards of security and reliability to our physical data centers, hardware, and software. We stay one step ahead of threats and failures through a robust lifecycle management and monitoring software, advanced cluster validation, proactive health checking capabilities, and deep observability capabilities.

This deep and broad visibility extends to our customers, as well. CoreWeave includes an array of detailed metrics and data for your clusters out-of-the box—from performance to temperature to health checks—with no setup or extra charges.

> "CoreWeave is clearly leading in providing the best GPU cloud experience and has very high goodput and are entrusted to manage the large-scale GPU infrastructure for AGI labs like OpenAI, high frequency trading firms like Jane Street.

SemiAnalysis ClusterMAX™ Report

Page 54[1]

[1] The GPU Cloud ClusterMAX™ Rating System | How to Rent GPUs, SemiAnalysis, March 26, 2025.

Summary

Fast-track your AI roadmap with the Platinum Cloud provider.

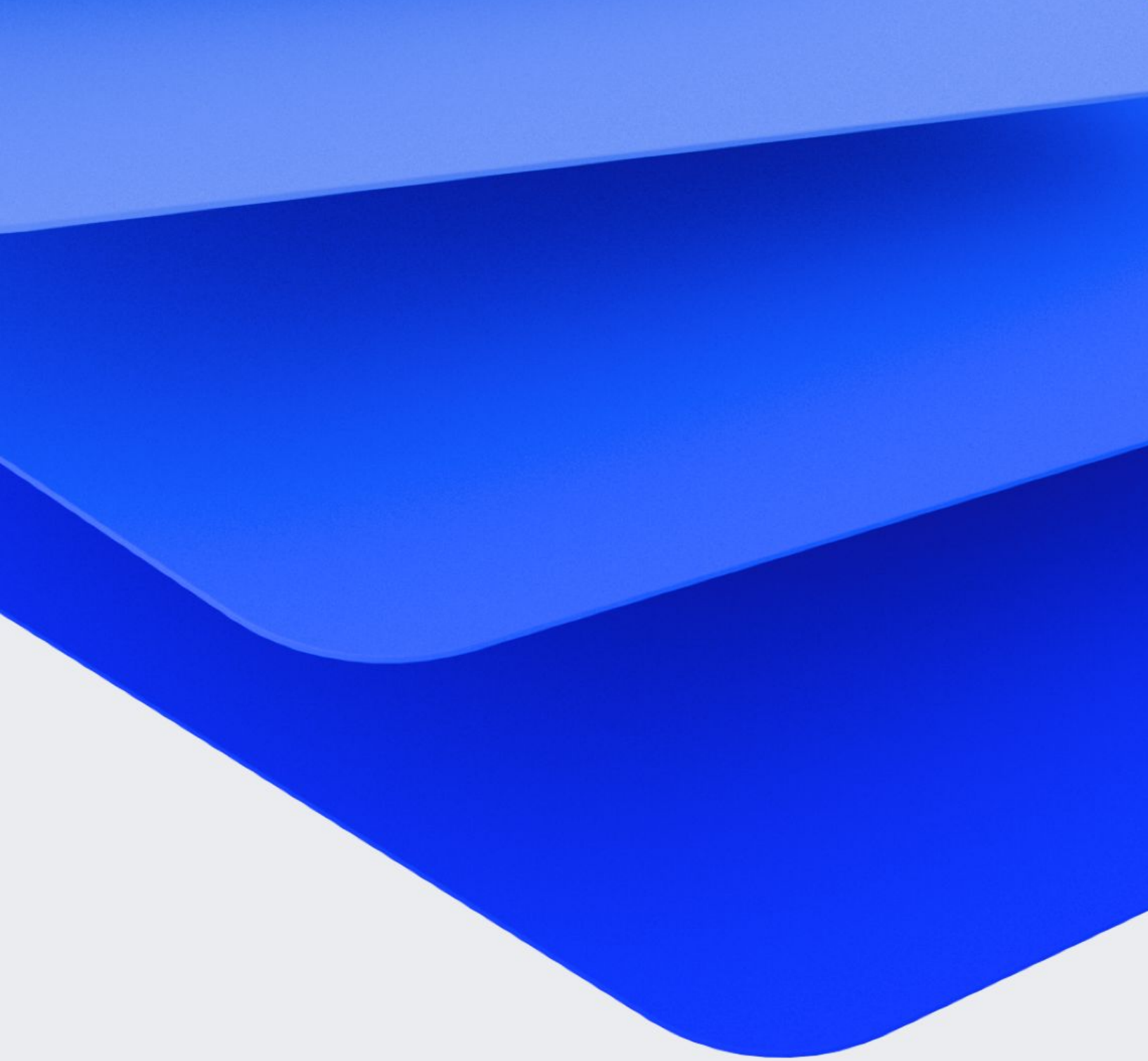
By evaluating AI clouds against these three pillars, you can safeguard your AI initiatives, accelerate time-to-value, and drive measurable ROI. As the first and only AI cloud to achieve Platinum status in SemiAnalysis's ClusterMAX™ Report, CoreWeave embodies the essential capabilities that a Platinum-tier cloud must deliver.

Pillar	CoreWeave Differentiators
Performance	• 96% system goodput vs. 90% industry average[3] • 20% higher cluster throughput[2] • 50% fewer interruptions[2] • Improved NCCL throughput on NVIDIA Quantum InfiniBand and SHARP enablement
Innovation	• First to market for NVIDIA GB200 and H100 GPU clusters • Container-native Slurm & Kubernetes integration with SUNK • In-switch SHARP compute offload for low-latency collectives • Automated NCCL tuning, validation, and health-check integrations
Enterprise Readiness	• SOC 2 & ISO 27001 compliant • Hardware-enforced tenant isolation with NVIDIA BlueField DPUs and VLANs • 99.9%+ SLAs backed by 24/7 global support • Proactive health-check, monitoring, and human-in-the-loop automation • Out-of-the-box dashboards for GPUs, networking, cluster operations, temperature, and more

CoreWeave combines hyperscaler-grade scale and security with the agility of a specialist. Our automated health checks, deep observability, and transparent economics empower your teams to focus on model innovation—not infrastructure headaches.

Ready to see how a Platinum-rated AI cloud accelerates your AI roadmap and slashes TCO? Contact our team to schedule a technical deep dive at info@coreweave.com or visit www.coreweave.com/contact-us.

[2] CoreWeave Leads the Charge in AI Infrastructure Efficiency, with up to 20% Higher GPU Cluster Performance than Alternative Solutions, March 19, 2025. These results are for clusters of 4K and 15K and should not be extrapolated to other cluster sizes.
[3] Achieve AI Infrastructure Goodput of up to 96% with 3 Key Strategies, March 27, 2025. These results are for clusters of 4K and 15K and should not be extrapolated to other cluster sizes.



Unleash AI's full potential. Our cloud platform is purpose-built for AI, operating at the bleeding-edge of technology with the speed and efficiency necessarily to perform at scale.

CoreWeave