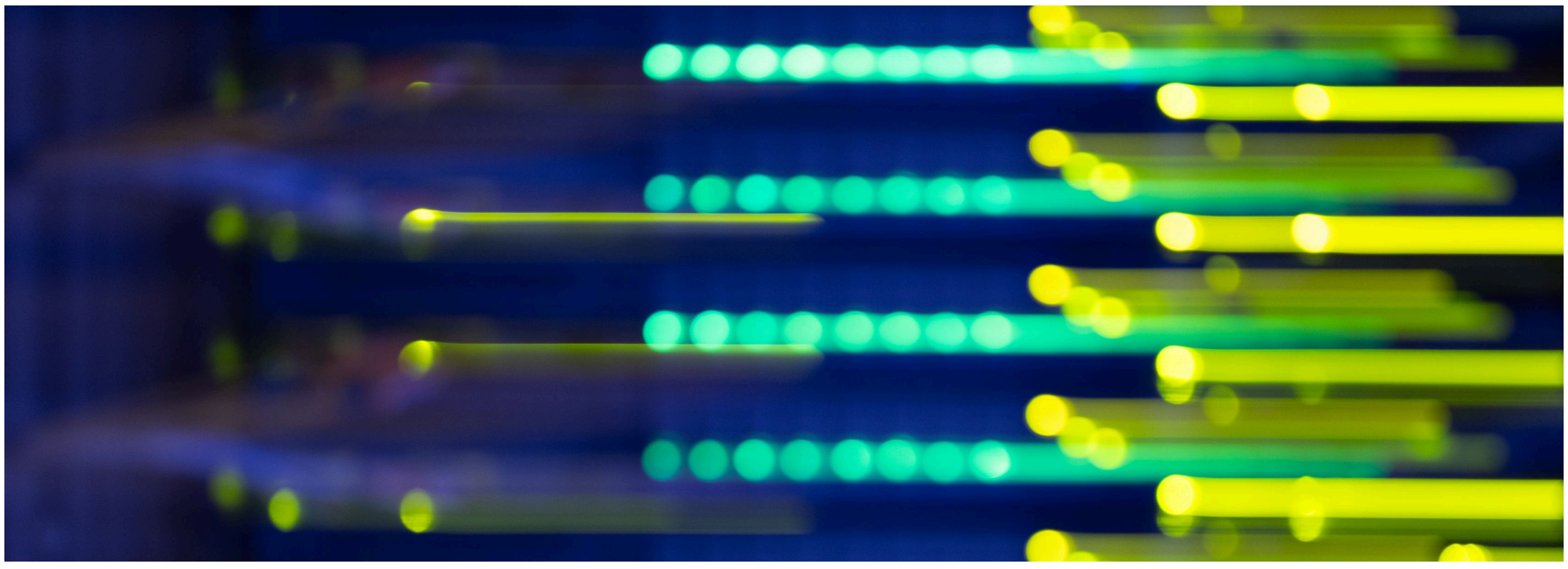
CoreWeave

Inference Is Your Product's Reliability Layer

Where user experience, economics, and operational risk now converge



Production inference breaks in predictable ways

Inference infrastructure is now a product decision

Production AI doesn't pause. Every response, every agent action, every workflow step depends on inference behaving predictably at scale, under load, as models change and traffic shifts. The infrastructure underneath determines whether it does.

As inference scales, failure modes compound before they surface. Latency instability, cost opacity, and limited control develop gradually. The infrastructure decision underneath determines whether teams can diagnose and correct them before users feel the gap.

When inference hits production, three things must hold

Reliability: Latency instability develops before it's diagnosable. Models update, concurrency grows, and by the time symptoms are user-facing, the window to intervene without impact has closed.

Cost: Spend disconnects from infrastructure behavior at sustained volume. Token pricing can work for exploration; as traffic patterns shift and context lengths grow, the bill becomes hard to attribute or forecast.

Control: Abstraction limits what teams can see and fix in production. The longer they operate on a limited foundation, the more expensive the correction becomes.

Inference behavior is product behavior

The right inference foundation turns production pressure points into operational variables—pricing, control, and visibility matched to how each workload actually behaves.

Purpose-built inference that holds at scale

CoreWeave runs inference directly on bare-metal GPU infrastructure. Three execution paths sit on the same cloud. Teams choose the control level a workload requires today and move between paths as requirements evolve.

Combined with W&B Weave for model-level observability and Mission Control for infrastructure telemetry, CoreWeave delivers end-to-end visibility from metal to model in a single stack. Teams start anywhere and scale without replatforming.

Serverless Inference

Fully managed

Deploy leading open-source models immediately with autoscaling included and no infrastructure to manage. Access a curated catalog including Llama, Qwen, DeepSeek, and Kimi with built-in observability through W&B Weave.

For speed to integration over fine-grained runtime control.

Dedicated Inference

Managed runtimes

Custom model weights on chosen GPU classes. CoreWeave manages production operations and SLAs. Full OpenAI API compatibility and infrastructure-aligned pricing keeps spend tied to actual GPU resources, not token abstraction.

For teams moving to production with strict SLA, isolation, or governance requirements.

Inference on CKS

Self-managed

Full Kubernetes-native ownership. Custom runtimes and infrastructure-level tuning on bare-metal GPUs. Supports vLLM, SGLang, NVIDIA Dynamo, and other open runtimes with explicit GPU selection for latency-critical or regulated workloads.

For platform teams with strict performance or compliance requirements.

Independently validated

#1 inference benchmark

Leading **DeepSeek R1** inference performance on NVIDIA GB200 NVL72, **MLPerf Inference v6.0**

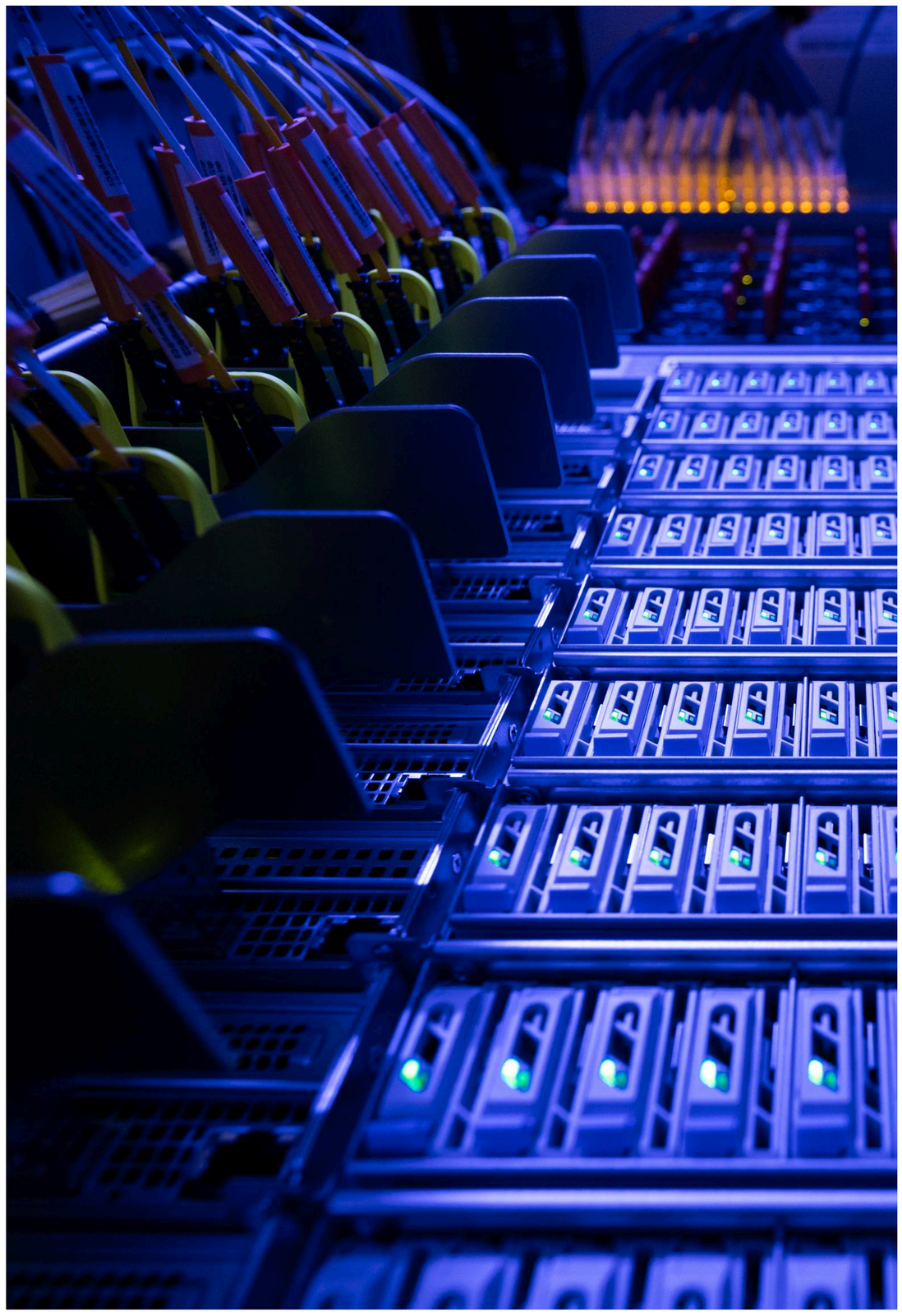
Exemplar Cloud Status

NVIDIA GB200 NVL72 Exemplar Cloud for inference—throughput and latency validated at production scale

#1 AI Cloud

The only AI cloud rated **SemiAnalysis ClusterMAX Platinum i**n both evaluations

Where infrastructure architecture becomes product advantage



Bare-metal performance

No virtualization overhead between workload and GPU. Latency and throughput stay consistent as concurrency grows—without the overhead abstracted environments absorb.

Infrastructure-aligned pricing

Cost tied to infrastructure choices, from pay-per-token to pay-per-GPU/hour. Reservations and Flex Reservations cover steady-state and burst, with no egress, ingress, or per-request fees so spend stays explainable to finance and actionable for engineering as workloads scale.

Metal-to-model observability

Infrastructure and runtime signals through CoreWeave Mission Control; model-level evaluation through W&B Weave. Teams diagnose root cause and intervene before the user experience degrades.

Take the next step with CoreWeave

Talk to an inference specialist about matching the right execution path to your workload.

Contact us

Explore CoreWeave Inference

See the full inference portfolio—Serverless, Dedicated, and CKS—and how each path maps to your workload.

Learn more