CoreWeave

SUNK: A Unified System for Production-Grade AI Training

The industry's first unified training system for the most demanding AI workloads

Large-scale runs, made easy

Explore how SUNK sustains performance and reliability for AI research workloads across thousands of GPUs over days-long runs.

Learn more

SUNK (Slurm on Kubernetes) represents the industry's first unified training system for your most demanding AI workloads. It applies cloud-native scale and agility to deliver production-grade reliability and deep operational visibility for large, long-running training jobs.

Beyond stitched-together stacks

By collapsing the silo between research supercomputing (Slurm) and cloud-native infrastructure (Kubernetes), SUNK delivers a production-ready, researcher-first system with high performance, reliability, and deep operational visibility.

A dependable, repeatable approach

Intelligent scheduling and automated health management reduce disruption and keep large, synchronized GPU jobs running. Because distributed training is only as fast as the slowest GPU, SUNK groups GPUs with the best connectivity and avoids noisy or degraded hardware, preventing bottlenecks that can stall an entire run.

96%
Training goodput

97-98%
Effective training time (ETTR)

10x longer
Mean time to failure (MTTF)

The modern AI training cluster, redefined

Modern AI training demands sustained synchronization across large GPU fleets, where fragmented Slurm and Kubernetes environments create operational drag and limit visibility. SUNK unifies the training system to keep long-running, distributed jobs moving forward with production-grade reliability, scalable orchestration, and the operational insight teams need at thousand-GPU scale.

> “
>
> CoreWeave’s SUNK was first to market, is proprietary, and continues to be the only viable solution for running both Slurm and Kubernetes jobs on the same underlying cluster.
>
> **ClusterMAX™ 2.0:**
> *The Industry Standard GPU Cloud Rating System, November 6, 2025*

SUNK does more than streamline the management of Slurm clusters within a scalable Kubernetes environment—it unifies the full AI training lifecycle. The result is increased workload fungibility that improves your resource efficiency and streamlines workload deployment

01 Unify the full AI training lifecycle

Run the most demanding AI training workloads on a unified training system that delivers production-grade reliability, predictable performance, and deep operational visibility. Ensure large, long-running training jobs succeed at scale without forcing your teams to assemble or operate complex infrastructure.

02 Maximize productive AI training time

Maximize productive training time by running AI training workloads on a unified training system designed to deliver high goodput. Dedicate more GPU time to making forward progress rather than struggling with restarts, retries, or stalled jobs.

03 Train with confidence at scale

Run large-scale AI training with confidence, backed by built-in reliability, deep observability, and CoreWeave-managed lifecycle controls. Reduce operational burden for both your researchers and platform teams, allowing them to focus their energy on innovation.

Predictable performance for large-scale training jobs

As clusters grow larger and training runs extend across days or weeks, reliability, goodput, and operational consistency prove to be as critical as raw performance. SUNK is designed to maximize productive training time. It combines intelligent scheduling with automated reliability so large-scale jobs consistently make forward progress.

Bonus: Customers running mixed training and production workloads on SUNK benefit from higher utilization without maintaining separate reserved capacity pools, improving return on invested capital as workloads grow.

Greater utilization

up to 96% goodput

By combining topology-aware scheduling with automated health management, SUNK ensures that the vast majority of requested GPU time translates into useful training work—reducing wasted spend and accelerating time-to-results.

Minimized disruption

97–98% ETTR

Through automated recovery and job continuity behaviors, SUNK sustains high ETTR—allowing long-running training jobs to continue progressing despite inevitable hardware events and minimizing disruption across multi-day runs.

More model progress

>50% MFU

Topology-aware scheduling and tuned infrastructure on SUNK reduce communication bottlenecks and straggler effects, supporting >50% MFU on thousand-GPU clusters and converting the same GPU-hours into more model progress.

The first NVIDIA GB200 Exemplar Cloud for training

CoreWeave exceeded NVIDIA's own training performance targets—validating the reliability, predictability, and operational maturity of CoreWeave's training environments under real-world conditions.

"With CoreWeave SUNK, we seamlessly manage clusters with thousands of GPUs. On rack-scale GB200 systems, SUNK's topology-aware scheduling and custom dashboards enabled faster, more efficient training runs and higher cluster utilization.

Brian Belgodere
Senior Technical Staff Member, IBM

Reliability engineered into every training run

SUNK leverages health management with end-to-end observability from CoreWeave Mission Control to ensure training jobs run predictably at scale, while simplifying how teams deploy clusters, onboard users, and access training environments.

Unlike typical 'experiment tracking' add-ons, SUNK's Weights & Biases integration is part of our end-to-end observability. By linking W&B runs to the infrastructure signals Mission Control sees (health, anomalies, stragglers), SUNK makes root-cause analysis actionable for both researchers and platform teams.

01 10x longer MTTF

CoreWeave clusters experience fewer interruptions in thousand-GPU clusters (~3.66 days vs. ~0.33 days / ~8 hours benchmark), reducing large-job disruption frequency and enabling predictable, multi-day training runs.

02 SUNK Grafana dashboards

Deep observability provides visibility from infrastructure health through job-level behavior, allowing teams to understand performance, diagnose issues, and trust training outcomes without manual intervention.

03 90 seconds to job restart

When failures do occur, automated self-healing and re-queueing (completed in ~90 seconds) minimizes interruption time and preserves productive training time across distributed jobs.

04 Transparent, simple operations

Simplified operations—including guided, opinionated cluster setup, streamlined user onboarding through SUNK User Provisioning (SUP), and isolated login environments—reduce day-to-day friction while preserving researcher control.

Monitor all SUNK jobs within your clusters

CoreWeave Mission Control continuously monitors cluster health, detects hardware anomalies and GPU stragglers, and automatically mitigates failures—reducing job interruptions and preserving productive training time.

Powered by CoreWeave Mission Control, this dashboard enables customers to:

- Monitor the resource usage of SUNK jobs across all clusters
- Track the rate of filesystem operations of SUNK jobs across all clusters
- View information about SUNK jobs on a per-user or per-partition basis

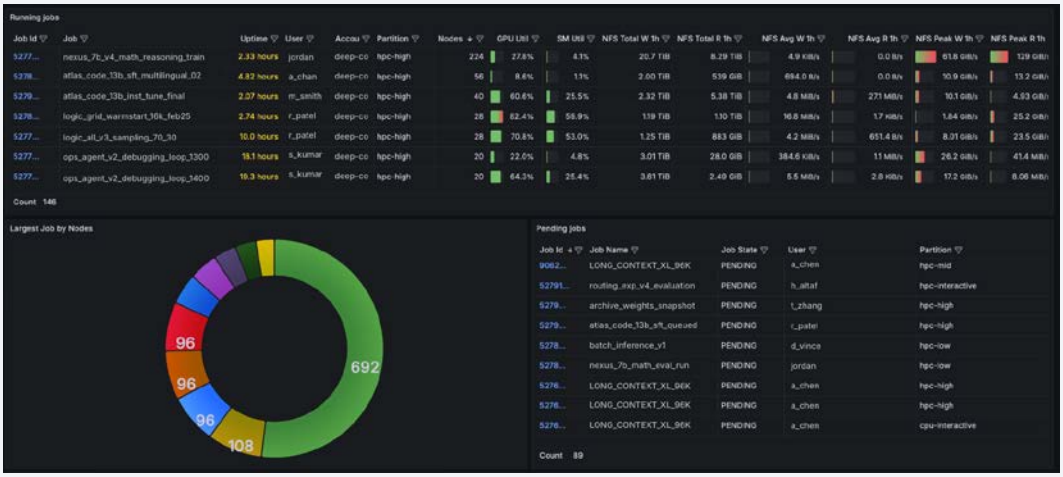


FIGURE 1

View of running and pending jobs in a Slurm cluster dashboard

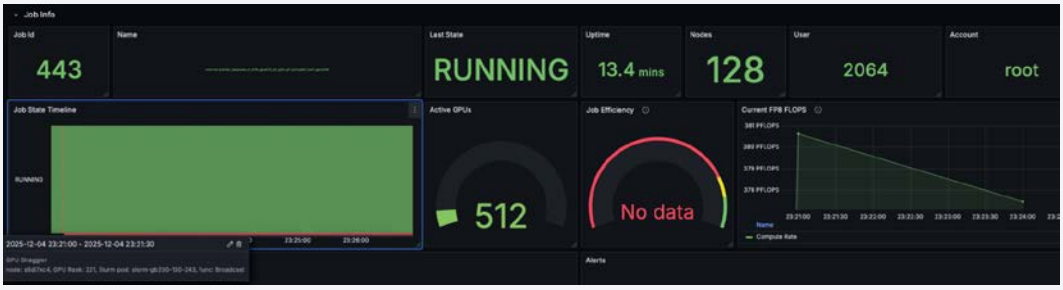


FIGURE 2

GPU Straggler Detection view for hung job in a Slurm job metrics dashboard

How SUNK works

CoreWeave's operational experience running research clusters at scale is built directly into SUNK's architecture. Proven patterns are standardized into a unified system that simplifies operations without sacrificing reliability or performance.

SUNK implements Slurm natively on Kubernetes, combining Slurm's proven scheduling with Kubernetes' declarative infrastructure and lifecycle management. This unified design enables teams to run training and inference on the same cluster, share compute resources efficiently, and increase workload fungibility across environments.

> "Using SUNK User Provisioning has made adding new users to our Slurm cluster trivial and given us the ability to easily set up the right set of permissions for our different use cases. If only all my activities were this straightforward.
>
> **Tim McNerney**
> *Member of Technical Staff, Inflection.AI*

Syncer

SUNK's Syncer maintains consistency between Slurm and Kubernetes resources, ensuring scheduling decisions and cluster state remain aligned. By reducing coordination gaps between control planes, Syncer preserves predictable behavior as clusters scale.

Topology-aware scheduler

At thousand-GPU scale, placement is hard: small differences in GPU-to-GPU connectivity can create stragglers that slow an entire distributed job. SUNK accounts for real GPU topology and communication patterns to place workloads where synchronization stays fast, sustaining high goodput and Model FLOPs Utilization as clusters grow.

Kubernetes-native lifecycle management

Slurm workloads run as first-class Kubernetes resources, inheriting declarative infrastructure, automated rescheduling, and predictable lifecycle behavior. Clusters scale dynamically while maintaining consistency across long-running jobs.

SUNK User Provisioning (SUP)

SUP automates secure onboarding and user isolation for production training environments. Programmatic access controls reduce manual configuration while preserving researcher autonomy and enterprise governance.

Get started with SUNK

SUNK offers two engagement models so teams can align operational ownership to their needs. Customers can start with a CoreWeave-supported rollout using standard SUNK primitives and hands-on Solution Architecture support, or choose guided self-service (in preview) for a more direct adoption path with structured guidance to help teams get training runs live quickly while maintaining clear operational responsibility.

What is guided self-service on SUNK?

SUNK provides a streamlined, self-service path to a production-ready SUNK cluster, built on proven operational patterns. CoreWeave manages deployment, upgrades, and lifecycle controls, enabling teams to focus on training workloads rather than infrastructure management.

Designed for fast onboarding and predictable operations, this guided self-service accelerates cluster creation by leveraging proven operational patterns without changing what customers fundamentally buy.

Key benefits

Fast onboarding	Stand up a production-ready SUNK cluster in minutes using guided setup with safe, sensible defaults
Predictable upgrades	Clear release channels and coordinated change windows reduce operational risk.
Reduced operational burden	CoreWeave manages lifecycle operations and common failure modes.
Researcher-friendly workflows	Built-in support for researcher pods, shared environments, and self-service access.
Enterprise alignment	Programmatic user access, permissions, and isolation aligned with security requirements.

Choose the solution purpose-built for AI training at scale

AI innovation no longer hinges on isolated performance gains. It depends on the ability to sustain model progress across large, synchronized training runs without disruption, drift, or operational bottlenecks.

By delivering a unified training system designed for production scale, SUNK enables teams to execute multi-day, thousand-GPU workloads with more reliability, less operational burden, and higher utilization.

More sustained model progress. More consistent outcomes at AI scale.

> “
>
> **We needed infrastructure that scales without dragging operations along with it. SUNK delivered that out of the box:** shared file systems, automated user provisioning, and customizable environments that let our researchers focus on research instead of fighting their tooling.
>
> **Sam Kottler**
> *ML Infra, Cursor AI*

Monitor jobs. Recover faster.

With continuous health management and deep observability, SUNK gives teams real-time visibility into cluster and job behavior—automatically mitigating failures and preserving forward progress when issues occur.

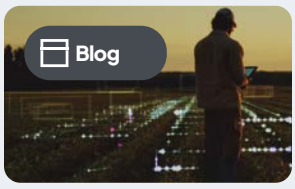
Streamline training. Drive efficiency

By replacing fragmented research infrastructure with a production-grade training platform, SUNK allows teams to execute at scale with confidence—improving workload efficiency and accelerating iteration cycles.

Run AI training on SUNK

See the difference of a unified training system that’s built for speed, scale, and stability.

Contact us



AI Training Cluster, Redefined

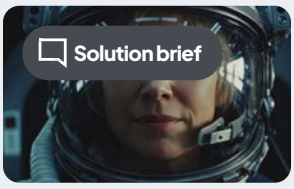
Discover how SUNK enables greater performance, reliability, and elasticity of training workloads.

Read more →

See SUNK in Action

In this demo video, you get an inside look into how SUNK works in CoreWeave Cloud.

Watch now →



Full-Stack AI Observability

See what makes CoreWeave Mission Control the operating standard for the AI cloud.

Learn more →