CoreWeave

Security at Cloud-Scale and the Future of Quant Research

How modern infrastructure delivers reliability, isolation, and observability for AI workloads

01 Executive summary

Security is no longer just about control. For quant teams, it is also about proving reliable execution at scale.

We break down the forces reshaping quant research, why legacy infrastructure fails, and what it takes to scale securely.

Mission-critical has expanded

For decades, quant research defaulted to on-prem infrastructure for good reasons. Control, isolation, deterministic performance, and trust in results mattered more than convenience. That foundation did not disappear, but the workloads running on top of it now operate at very different scale, duration, and levels of complexity.

Why legacy models break down

Training and backtesting can scale to hundreds of GPUs, with some teams operating even larger fleets. Hardware innovation outpaces traditional on-prem refresh cycles. As scale increases, reliability becomes harder to sustain, visibility degrades, and reproducibility suffers.

On-prem environments reach operational limits precisely as research intensity peaks. General-purpose clouds offer elasticity, but introduce performance variability and opaque execution that quant teams cannot accept.

The new requirement for quant security

Modern quant security requires more than isolation. It requires reliable execution, provable behavior, and continuous visibility across every workload. At scale, security and reliability are inseparable. Observability is what makes both verifiable.

02 Redefining mission-critical

The AI era leaves on-prem-only teams at risk

For decades, quant research teams relied on on-prem clusters to deliver security by design. Elite engineering teams could observe the full system, tune performance precisely, and trust that execution would remain consistent over time.

That model was sustainable as long as the workloads were predictable and manageable.

However, AI-driven quant workloads changed those boundaries for good, exceeding the physical and operational limits of on-prem environments and challenging traditional security assumptions. Quant teams pushing the limits of scale, duration, and iteration speed felt the shift first.

The shifts that redefined mission-critical

01 Larger, longer-running jobs

Training workloads increasingly span thousands of GPUs and run for days or weeks at a time. As job duration and scale increase, even small performance inconsistencies or hardware issues can materially impact outcomes.

02 Highly distributed training and backtesting

Backtesting has evolved into large-scale, continuous simulation. Distributed execution is now the norm, not the exception, increasing sensitivity to jitter, failures, and uneven performance across nodes.

03 Bursty demand and frequent iteration

Research demand is increasingly bursty and unpredictable. Teams need to scale experimentation rapidly as hypotheses change, market conditions shift, or new models emerge, without waiting on fixed capacity.

04 Hardware innovation outpacing on-prem cycles

GPU innovation now moves faster than traditional procurement and refresh cycles. For teams operating at the leading edge, staying competitive requires access to new architectures that are difficult to deploy and sustain on-prem.

02 Redefining mission-critical

When scale redefines what security means

The shifts reshaping quant workloads do more than strain infrastructure capacity. They challenge the assumptions that once defined security.

Scale introduces more operational fragility

Long-running, distributed jobs introduce more opportunities for small failures to accumulate. Performance degrades unevenly. Jobs stall or slow in ways that are difficult to detect without deep visibility. What once felt deterministic becomes fragile.

When critical workloads fail or degrade, the impact goes beyond wasted compute. Results become harder to reproduce. Confidence in outcomes erodes. And without a clear record of how workloads behaved, governance and validation break down.

Mission-critical infrastructure must now do more than protect data and models. It must ensure that execution itself can be trusted.

The new security imperative: reliability

Mission-critical did not disappear. It simply expanded.

For modern quant teams, mission-critical infrastructure is defined by reliable execution, provable behavior, and continuous visibility at cloud scale. Reliability is no longer implicit. It must be engineered, observed, and continuously validated across every workload, across the full stack.

This is the new security imperative for quant research.

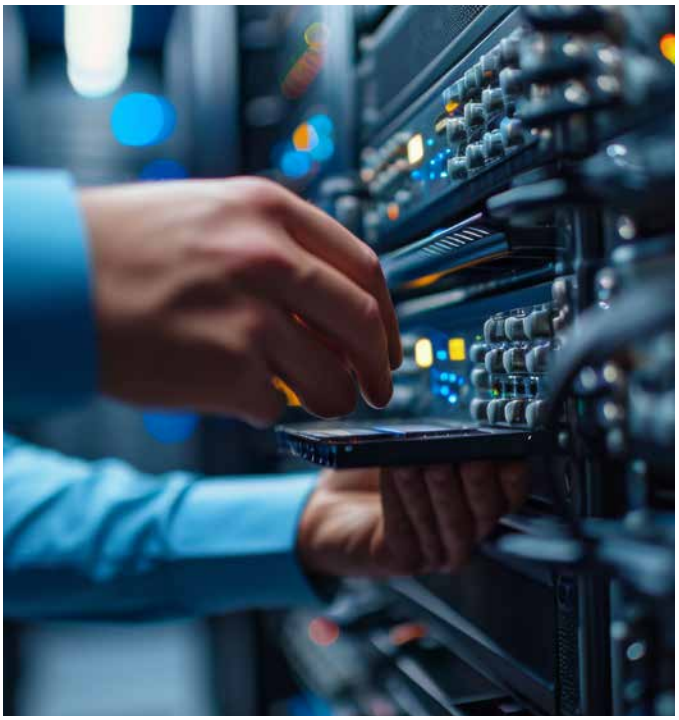
The question is: How to modernize infrastructure without sacrificing the control and reliability that made on-prem the default in the first place?

03 Infrastructure shortfalls

Traditional models fail modern quant workloads

Most quant teams assume they have two viable options for running mission-critical workloads: on-prem infrastructure or a general-purpose cloud. However, neither was designed for the scale, execution characteristics, and reliability requirements of modern, AI-driven quant research.

As workloads grow larger, longer-running, and more distributed, the limitations of both models become harder to ignore.



On-prem shortfalls

For teams operating on-prem environments, the infrastructure that once protected research velocity now becomes the constraint slowing it down. As scale and complexity exponentially increase, reliability degrades precisely when research intensity is highest.

Common challenges include:

01 Fixed capacity ceilings

On-prem clusters cannot easily absorb sudden spikes in training or backtesting demand, forcing teams to queue work or limit experimentation.

02 Hardware procurement

Long procurement and deployment cycles can make it difficult to keep up with the latest hardware generations and innovations.

03 Operational fragility at scale

As clusters grow larger and jobs run longer, small failures become more frequent and harder to isolate, increasing operational fragility.

04 Limited operational visibility

Many on-prem environments rely on manual diagnostics and fragmented tooling, increasing overhead and slowing root-cause analysis.

03 Infrastructure shortfalls

General-purpose cloud weaknesses

General-purpose clouds promise elasticity, but they were not built for the execution guarantees quant workloads require. For quant research, trading capacity for consistency and control introduces execution risk that teams cannot tolerate.

Key limitations include:

01 Multi-tenant performance variability

Virtualization layers introduce jitter, noisy-neighbor effects, and unpredictable performance that undermine deterministic execution.

02 Opaque infrastructure behavior

Limited visibility into underlying infrastructure restricts auditability and makes root-cause analysis difficult.

03 Inconsistent execution outcomes

Performance variability directly affects backtesting fidelity and model reproducibility.

04 Lack of operational proof

Teams often lack verifiable records of how workloads actually ran, complicating governance and validation.

Neither model provides the combination quants require: isolation + elasticity + real-time operational proof.

How much of an impact does a new GPU generation have?

10x

overall performance improvement (at 25 tokens/sec/user) for NVIDIA GB200 NVL72 vs. previous generation

12x

better performance per dollar vs. NVIDIA H200 (factoring pricing difference)

Access to the latest hardware generations can have a profound impact on quant researchers' ability to scale, iterate, and stay cost-competitive.

"From Dense to Mixture of Experts: The New Economics of AI Inference," Signal65, Ryan Shrout, December 29, 2025.

04 The new security standard

What's required for quant security at cloud-scale?

As workloads grow larger, longer-running, and more distributed, neither traditional on-prem nor general-purpose clouds were built for this new AI reality.

What quant teams need is infrastructure that delivers reliable, deterministic execution at cloud scale, paired with the visibility required to prove it.

This evolution demands a new architectural approach

Meeting this mission-critical baseline requires infrastructure purpose-built for AI workloads from the ground up.

In this model, reliability is not separate from security—it is a prerequisite for it. Isolation must persist at scale, execution must be deterministic, and observability must provide operational proof, not just logs after the fact.

This is the gap modern quant teams are left to navigate.

This is the gap purpose-built AI clouds are designed to close.

The new baseline for mission-critical execution

Maintain isolation and performance consistency	Execution must remain predictable as workloads scale rapidly across hundreds or thousands of GPUs.
Tolerate failure without cascading disruption	Small failures are inevitable at scale. Infrastructure must contain them without compromising running jobs or results.
Detect and explain issues in real time	Problems must be surfaced, diagnosed, and addressed while workloads are still running, not after results are invalidated.
Provide continuous operational visibility	Teams need auditable insight into how workloads run, where they run, and how resources are used across every execution.

05 The AI cloud

Security at scale. Delivered by the AI cloud.

Secure quant research at cloud scale is not achieved by layering controls onto legacy infrastructure. It requires an AI cloud designed to deliver reliable execution, enforce isolation, and provide continuous operational proof as workloads scale. These three pillars are inseparable. Reliability enables trust in results. Isolation preserves control. Observability turns guarantees into proof. Together, they define what security means for quant research at scale.

Reliability

Deterministic, predictable, jitter-free execution makes results reproducible and verifiable across long-running, distributed workloads. At scale, infrastructure must tolerate failure without cascading disruption. When reliability breaks down, reproducibility suffers, governance erodes, and trust in results disappears.

Isolation

Secure quant research requires clear, enforceable boundaries at both the physical and logical layers. Those boundaries must persist under load and at scale. Multi-tenant ambiguity introduces security and governance risk by making execution environments harder to reason about, control, and protect.

Observability

Quant teams need continuous, real-time visibility into every job, GPU, node, and performance pattern. They must be able to explain what ran, where it ran, and how it behaved. Without workload-level telemetry, execution cannot be validated and security becomes an assumption rather than a fact.

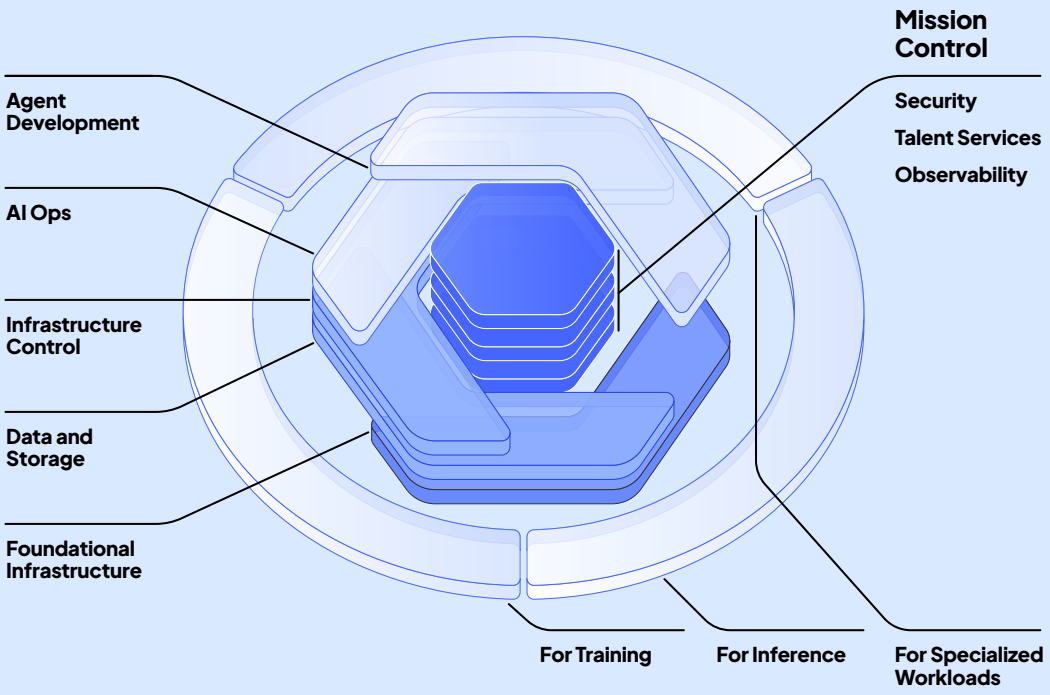
06 CoreWeave architecture

Inside the AI cloud built for secure quant research

Delivering security at scale requires more than individual controls. It requires an AI cloud designed as a single, integrated system—where reliability, isolation, and observability are built into every layer of the stack.

Where general-purpose clouds abstract infrastructure behind shared layers, CoreWeave integrates compute, networking, storage, orchestration, and operations into a unified AI cloud stack. The result: reliability, isolation, and observability built directly into the architecture.

Together, these layers deliver secure execution that remains reliable, isolated, and observable as demand scales.



Designed as a system, not a collection of services.
The CoreWeave AI cloud integrates infrastructure, orchestration, and operations into a system designed for secure executive at scale.

06 CoreWeave architecture

How CoreWeave delivers security at scale

CoreWeave's purpose-built AI cloud is designed to deliver secure, elastic quant research without sacrificing performance, control, or governance.

This architecture gives quant teams on-prem-level sovereignty with cloud-scale flexibility. Execution remains deterministic and reproducible, failures are contained rather than amplified, and every workload is observable and auditable as scale increases.

Reliable performance

Infrastructure designed for predictable execution

- ☐ Kubernetes on bare metal with Slurm integration for consistent scheduling and execution
- ☐ Greater than 99.9 percent uptime across production environments
- ☐ Proximity to global trading hubs to support latency-sensitive workloads
- ☐ Predictable, jitter-free execution environments for trusted backtesting and distributed training
- ☐ Deep observability integrated into the execution layer

True isolation

Security built at the architectural layer

- ☐ Physically and logically isolated single-tenant clusters
- ☐ DPU-enforced network segmentation with EVPN/VXLAN per-tenant virtual networks
- ☐ No hypervisors, no noisy neighbors, no shared L2 or L3 network surfaces

Elasticity without exposure

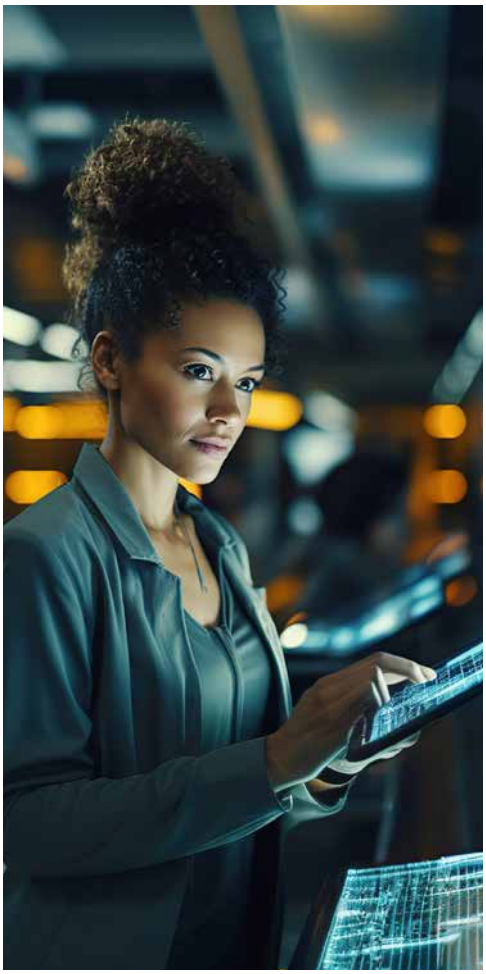
Scale securely without changing your security posture

- ☐ Securely scale to accommodate tenfold workload spikes
- ☐ Access to next-generation GPUs without procurement or deployment delays
- ☐ Maintain consistent isolation, performance, and governance while scaling
- ☐ Integration with experiment tracking and governance platforms used by quant teams

07 CoreWeave Mission Control

Mission Control: making security actionable

CoreWeave Mission Control is the operational layer of the purpose-built AI cloud. It turns architectural guarantees into observable, governable, and auditable reality.



GPU Straggler Detection

Surface performance issues before results are compromised. By identifying root causes in distributed jobs, teams can intervene early and prevent silent degradation that undermines outcomes.

Direct-to-expert support

Resolve mission-critical issues without delay. Execution-level signals route incidents directly to CoreWeave engineers, reducing time to resolution and limiting research disruption.

Telemetry Relay

Stream real-time operational and audit signals out of the platform. Workload telemetry flows directly into customer systems and SIEMs without post-hoc reconstruction.

Execution-level observability

Understand how workloads actually behaved. Mission Control provides continuous visibility into GPU, node, job, and network behavior across every run.

Governance built into workflows

Make auditability a byproduct of execution. Mission Control integrates with tools such as Weights & Biases to support experiment tracking, governance, and compliance across distributed workflows.

Reproducibility and drift detection

Validate consistency across runs and over time. Teams can detect drift, compare executions, and maintain reproducible results as models and scale evolve.

08 Quant use principles

The force multiplier for quants

Across high-frequency trading and systematic strategies, CoreWeave Mission Control powers production research workflows where execution behavior must be trusted, validated, and repeatable.

PRINCIPLE 1

Large-scale signal discovery

GPUs accelerate feature generation, factor evaluation, and model training across massive historical datasets. Teams run thousands of experiments in parallel to surface weak signals that only emerge at scale.

Mission Control impact:

Provides execution-level visibility across parallel experiments, allowing teams to detect drift, validate performance consistency, and c ompare results with confidence.

PRINCIPLE 2

High-frequency model training

Latency-sensitive models are trained and retrained continuously as market conditions evolve. Deterministic execution is essential so training outcomes translate reliably into live environments.

Mission Control impact:

Detects performance anomalies and GPU stragglers during distributed training runs, keeping long-running jobs stable and results reproducible.

PRINCIPLE 3

Massively parallel backtesting

Distributed GPU clusters enable continuous simulation across decades of tick-level data. Results directly inform high-stakes capital allocation decisions and must be auditable.

Mission Control impact:

Delivers end-to-end traceability into how simulations ran, where they ran, and how resources were used—without manual log reconstruction.

PRINCIPLE 4

Intraday strategy optimization

GPUs power rapid intraday retraining and parameter tuning as markets shift. Teams need elastic capacity without introducing execution variability.

Mission Control impact:

Surfaces real-time performance signals across rapidly scaling workloads, helping teams validate results as experimentation velocity increases.

PRINCIPLE 5

Risk modeling and stress testing

Firms use GPUs to run large-scale scenario analysis and stress testing across portfolios. These workloads demand predictable execution and clear governance trails.

Mission Control impact:

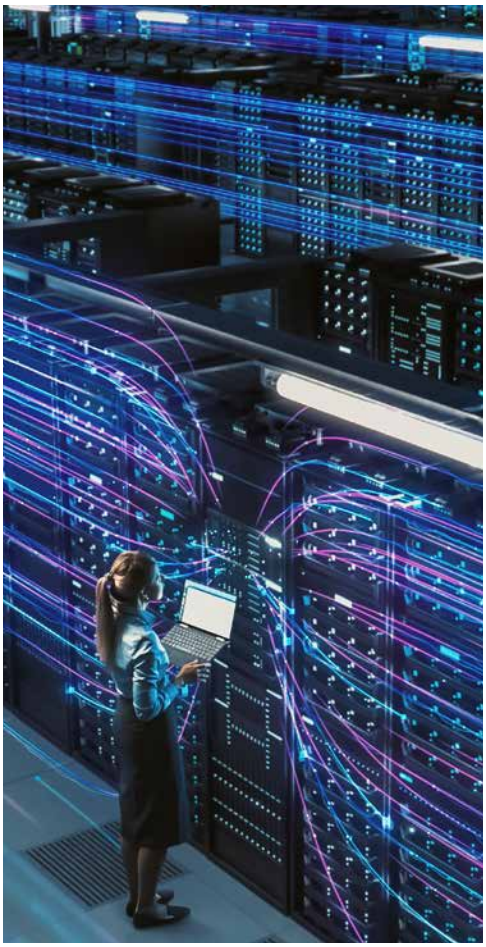
Streams audit and operational telemetry into customer systems, supporting governance, review, and reproducibility across risk workflows.

09 Action guide

Modernize security without compromise

Quant teams no longer have to choose between on-prem sovereignty and cloud-scale flexibility. The right infrastructure makes security, reliability, and velocity mutually reinforcing rather than mutually exclusive.

Use the following principles to guide secure modernization of quant research infrastructure.



01 Treat reliability as a security requirement

If reliability degrades, reproducibility and governance degrade with it. Long-running, distributed workloads must behave predictably under sustained load, with failures detected and contained before results are compromised.

02 Preserve control through isolation

Elasticity only matters if isolation holds at scale. Clear, enforceable boundaries at the physical and logical layers are essential to protect proprietary models, data, and research workflows.

03 Demand operational proof, not assumptions

Mission-critical research requires more than logs and post-hoc analysis. Observability is what turns infrastructure guarantees into auditable reality, providing quant teams with continuous, real-time visibility into how workloads run, where they run, and how resources are used.

04 Scale without sacrificing governance

Modern quant research depends on rapid iteration across the latest hardware. Infrastructure should make it possible to scale securely while maintaining reproducibility, auditability, and integration with existing research and governance tools.

05 Choose a platform built for quant security at scale

Infrastructure must be designed for today's AI-accelerated quant strategies. CoreWeave combines isolated, high-performance infrastructure with Mission Control's transparency and operational intelligence to deliver a secure, governable, and elastic research environment.

10 Next steps

Put secure execution into practice

Modern quant research now operates at a scale where infrastructure behavior directly shapes outcomes. The next step is to move from principle to practice by evaluating platforms built to deliver reliable, observable execution under sustained load.

Explore CoreWeave Mission Control

Discover what makes CoreWeave Mission Control the operating standard for large-scale AI workloads.

Download →

Jane Street Scales, Securely

Learn how Jane Street gained a flexible, full-stack solution through CoreWeave.

Read more →

Redefining Mission Critical Infra

See how quant teams are reevaluating their infrastructure for the AI era.

Read more →