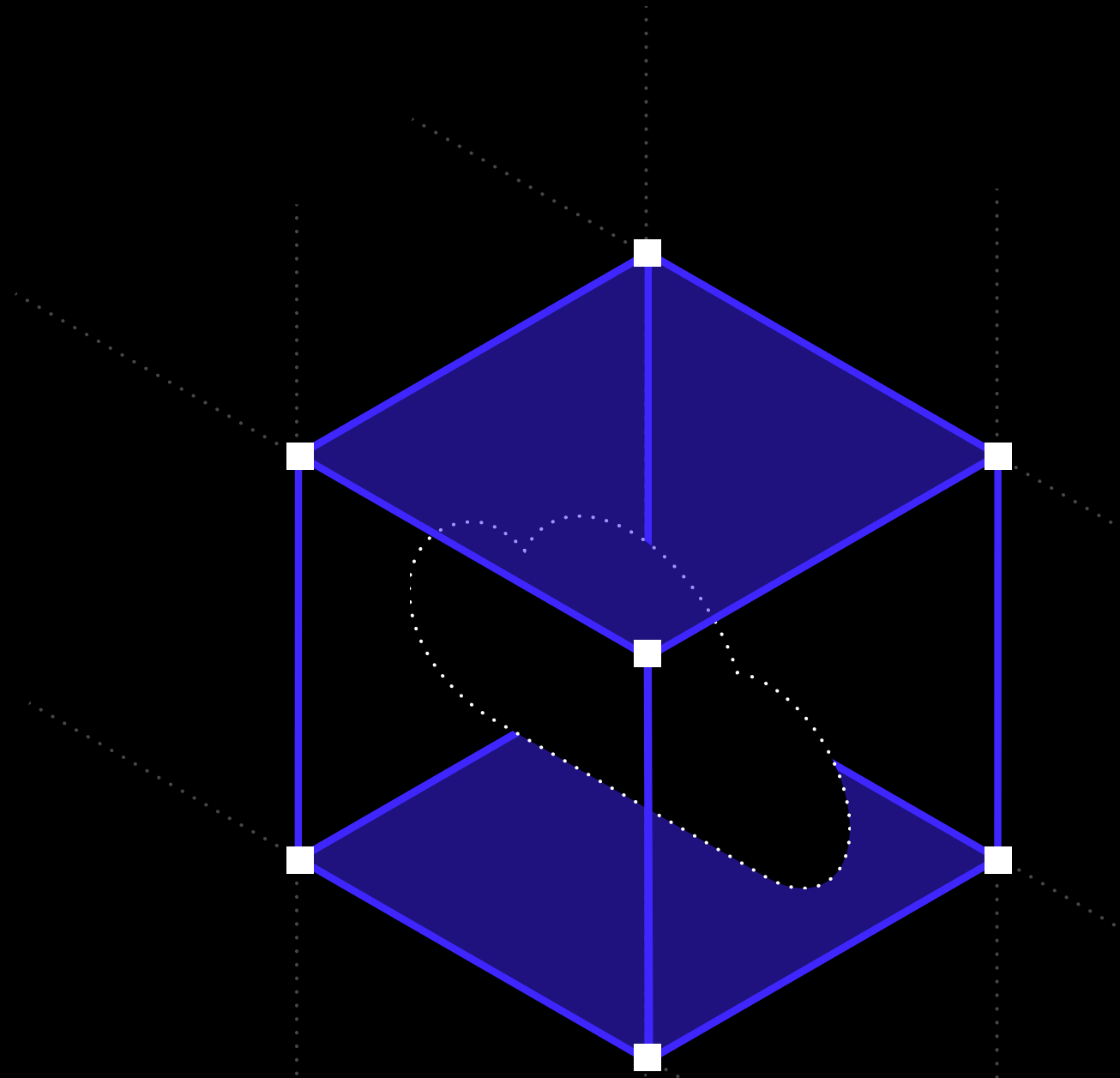
CASE STUDY

How Lambda built a hyperscaler cluster in 90 days



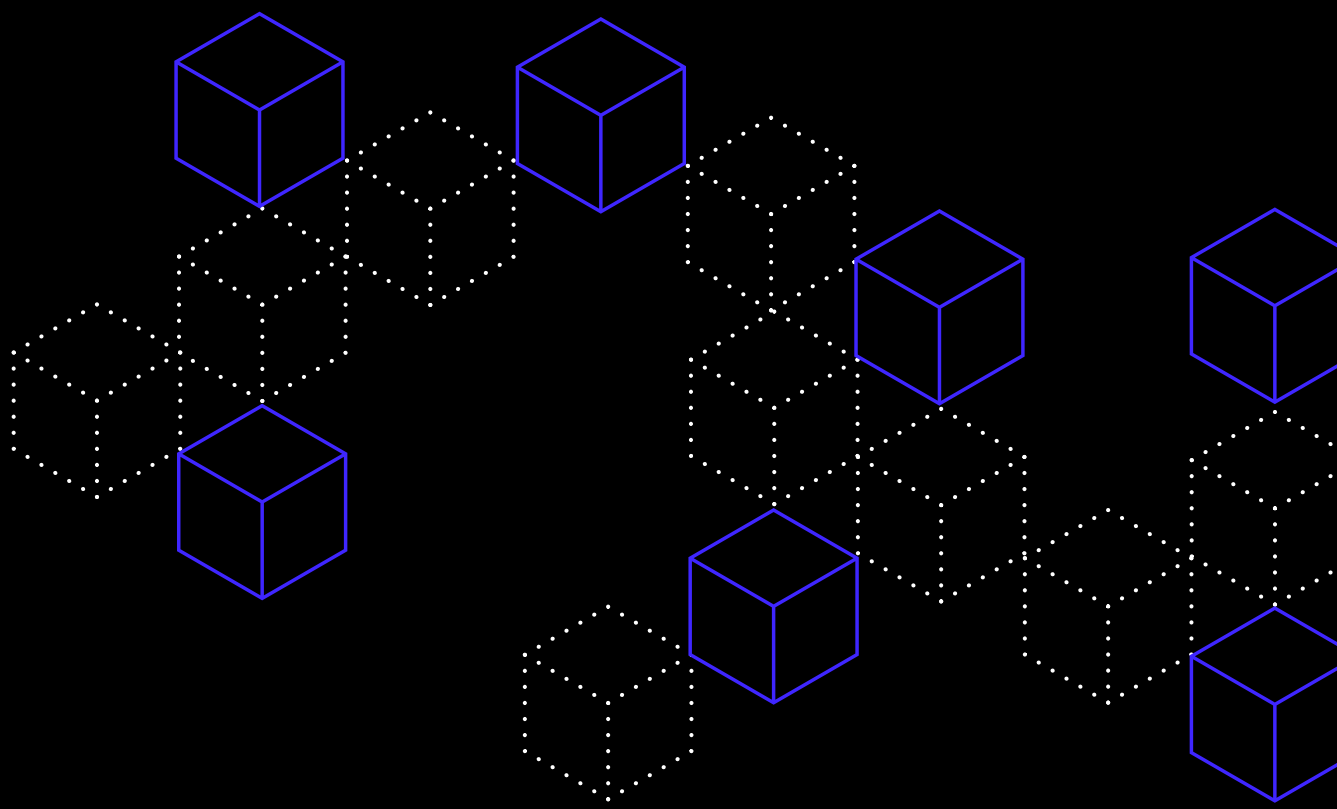
Lambda

Overview

Building private GPU clouds for AI workloads meant to run at hyperscale requires getting countless details right: access to the latest NVIDIA GPUs, ultra-low-latency networking, cabling perfection, scalable orchestration, and strict security postures. Getting thousands of GPUs operational in a private cloud, especially on a tight schedule, comes down to careful planning, reliable processes, and a team that knows what they're doing.

When Lambda was asked to deliver thousands of GPUs as a hyperscale private cloud within just 90 days, the team knew exactly what they were getting into. Tight deadlines didn't leave room for surprises, every possible risk had to be anticipated, mitigated, and managed without compromising quality or security. To achieve this, Lambda relied on meticulous pre-deployment planning, diligent procurement, proactive contingency management, and automated validation.

The result: a fully operational, enterprise-grade GPU infrastructure, delivered exactly as promised, ready to handle critical AI workloads from day one.



Challenge

Building for Hyperscalers on a tight timeline

Delivering thousands of GPUs as a private cloud cluster in 90 days required Lambda to overcome several challenges:

No room for delays

Equipment lead times were longer than the project timeline. Lambda had to pre-secure key inventory before formal specs were locked-in, and build a deployment strategy around it.

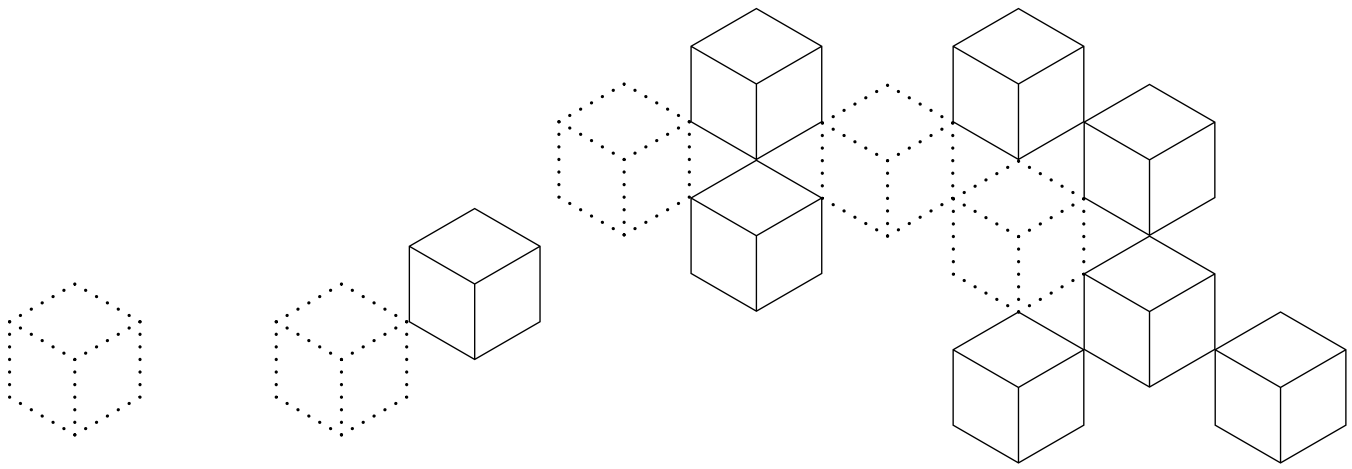
Tight tolerances at scale

Tens of thousands of cable terminations, a complex multi-tier network topology, and a massive host fleet meant even small issues could cascade into major setbacks. Lambda's automation and validation workflows had to work from the first rack to the last.

Hyperscale-ready at handoff

Every GPU, every switch, every layer of the stack was tested and verified before delivery.

Every detail mattered. The entire system had to come online cleanly and ready for production at hand-off.



Solution

Coordinated Execution

To deliver on time, Lambda orchestrated a tightly integrated plan across real estate, supply chain, networking, and platform engineering builds in parallel, and relied on automation to compress timelines without sacrificing reliability.

Power & Facilities Readiness

+ Secured several megawatts of infrastructure capacity ahead of real estate closure
+ Proactively engaged with contractors to align construction, power, and inspections to the deployment timeline
+ Strategically pre-ordered long-lead power gear to avoid delays

Supply Chain & Logistics

+ Sequenced hardware delivery with build milestones to avoid floor congestion
+ Maintained buffer inventory to absorb failure rates without slowing down progress

Data Center Deployment

+ Compressed rack deployment timelines by automating provisioning and parallelizing workflows
+ On-site technical team scaled dynamically to meet daily milestones

Networking & Fabric Architecture

+ Deployed a multi-tier high-performance interconnect designed for distributed AI workloads across thousands of GPUs
+ Designed for NVIDIA NCCL and SHARP compatibility to support tightly-coupled distributed training workloads
+ Automated network configuration through Zero Touch Provisioning (ZTP) tools
+ Built for high-throughput, low-latency traffic patterns with full redundancy

Platform Automation & Provisioning

+ Rolled out a new stress-tested provisioning workflow
+ Enabled concurrent provisioning, significantly reducing time-to-operational readiness

Security & Compliance

+ Delivered a fully isolated, single-tenant environment with no Lambda access to customer systems
+ Aligned cluster operations with SOC 2 controls and zero-touch infrastructure requirements
+ Integrated with the customer's kubernetes orchestration stack to support hybrid workload requirements

Results

Lambda delivered the full hyperscale private cloud on schedule, with all systems validated, secured, and production-ready at handoff. Despite the scale and complexity, the deployment met every operational and infrastructure milestone without compromising performance or reliability.

Key outcomes:

01 Thousands of NVIDIA GPUs deployed, configured, and handed off in a fully operational, production-ready state in 90 days.

02 Full-stack validation and burn-in completed in under 10 days.

03 Network, power, and system deployment executed with minimal rework and consistent throughput.

04 Delivered within a tight security envelope, aligned with SOC 2 and customer-specific compliance expectations.

The success of this build reinforces Lambda's position as the partner of choice for AI infrastructure, capable of executing complex deployments with precision and speed, without compromising quality, control, or deadlines.