FEDERAL | NATIONAL

1818 LIBRARY STREET, SUITE 500 | RESTON, VA 20190
CONTACT@CTGFEDERAL.COM | CONTACT@CTGNATIONAL.COM
703.278.3885
CTGFEDERAL.COM | CTGNATIONAL.COM
Cohesive Technology Group

AT A GLANCE

4 Days
Unboxing to deployment ready

24 Nodes
GPU compute servers fully provisioned

192 GPUs
Configured with drivers, CUDA, and AI stack

85%+
Reduction in deployment time

100%
Node configuration consistency

0
Manual configuration errors

AI-Driven Automation for Rapid Deployment of a NeoCloud GPU Cluster

A Field Implementation by CTG Federal | National

Executive Summary

Deploying large-scale AI infrastructure has traditionally been a weeks-long effort spanning hardware provisioning, operating system installation, networking, and cluster configuration. This paper describes how CTG Federal used AI-assisted software development and purpose-built automation to deploy a 24-node, 192-GPU AI cluster in just four days, from unboxing to production readiness.

Using AI tools to accelerate development, the CTG team built a lightweight orchestration application that integrated PXE boot, DHCP, TFTP, HTTP, Redfish APIs, and Ansible to automate the entire cluster deployment lifecycle. The system provisioned bare-metal GPU servers, installed and configured the operating system, applied GPU drivers and cluster software, and validated the environment across all nodes simultaneously.

The result is a repeatable NeoCloud-style deployment framework that turns raw hardware into a production-ready AI compute environment in a fraction of the time required by traditional methods. For organizations planning GPU infrastructure at any scale, this approach eliminates the bottleneck between procurement and production.

The Challenge

Organizations building NeoCloud GPU environments face a common bottleneck: the gap between receiving hardware and running production AI workloads. Traditional deployment methods require sequential manual work that stretches timelines into weeks, introduces configuration drift, and creates risk at every step.

A typical GPU cluster deployment involves:

- BIOS and firmware configuration on every node individually
- Physical node provisioning and operating system installation
- Network and storage configuration across the fabric
- GPU driver and CUDA stack installation per machine
- Cluster orchestration setup and end-to-end validation

Each step is performed manually and sequentially. In a 24-node, 192-GPU environment, this means engineers repeat the same procedures dozens of times, trusting that each repetition is identical. The result is a four-to-six week timeline, a high risk of human error, and significant cost in skilled labor.

Why Traditional Methods Fall Short

Beyond the time cost, conventional deployment practices create several downstream risks that are particularly acute for organizations building NeoCloud infrastructure at scale.

Risk Area	Impact on Your Organization
Configuration drift	Nodes end up with different driver versions, BIOS settings, or OS patches — causing unpredictable behavior in production
Vendor lock-in	Deployment scripts written for one OEM break when supply constraints force a switch to another vendor
No audit trail	Manual processes produce inconsistent documentation, making compliance and troubleshooting difficult
Non-repeatable	Each deployment is a one-off project — the next cluster starts from scratch
Sequential bottleneck	Nodes provisioned one at a time, with no ability to parallelize across the cluster
Talent dependency	Specialized engineers required on-site for the full duration of every deployment

For NeoCloud builders navigating supply chain volatility — where one quarter may bring Dell hardware and the next requires HPE, or a mixed environment based on availability — these problems compound. A deployment practice locked to a single vendor becomes a liability the moment supply conditions shift.

The CTG Solution: AI-Driven, Vendor-Agnostic Deployment

CTG Federal built a deployment practice designed from the ground up around open standards and AI-assisted engineering. Rather than hard-coding automation for a single OEM, the platform uses industry-standard protocols — PXE, Redfish, DHCP, TFTP, HTTP, and Ansible — that work across hardware vendors.

This matters because the automation is rapidly tailored to each customer's specific infrastructure and requirements. When hardware changes — whether due to supply availability, a vendor switch, or a mixed-vendor strategy — the CTG team uses AI-assisted development to adapt the automation quickly rather than starting over. The result is a deployment practice that is bespoke to your environment but built on a portable, reusable foundation.

> **Vendor flexibility in practice:**
> Open protocols like Redfish and PXE are supported across Dell, HPE, Lenovo, Supermicro, and other major server vendors. When your hardware mix changes, the CTG framework adapts — the deployment process does not restart. AI-assisted development accelerates this adaptation, turning what would be weeks of re-engineering into days of targeted refinement.

Automation Framework Architecture

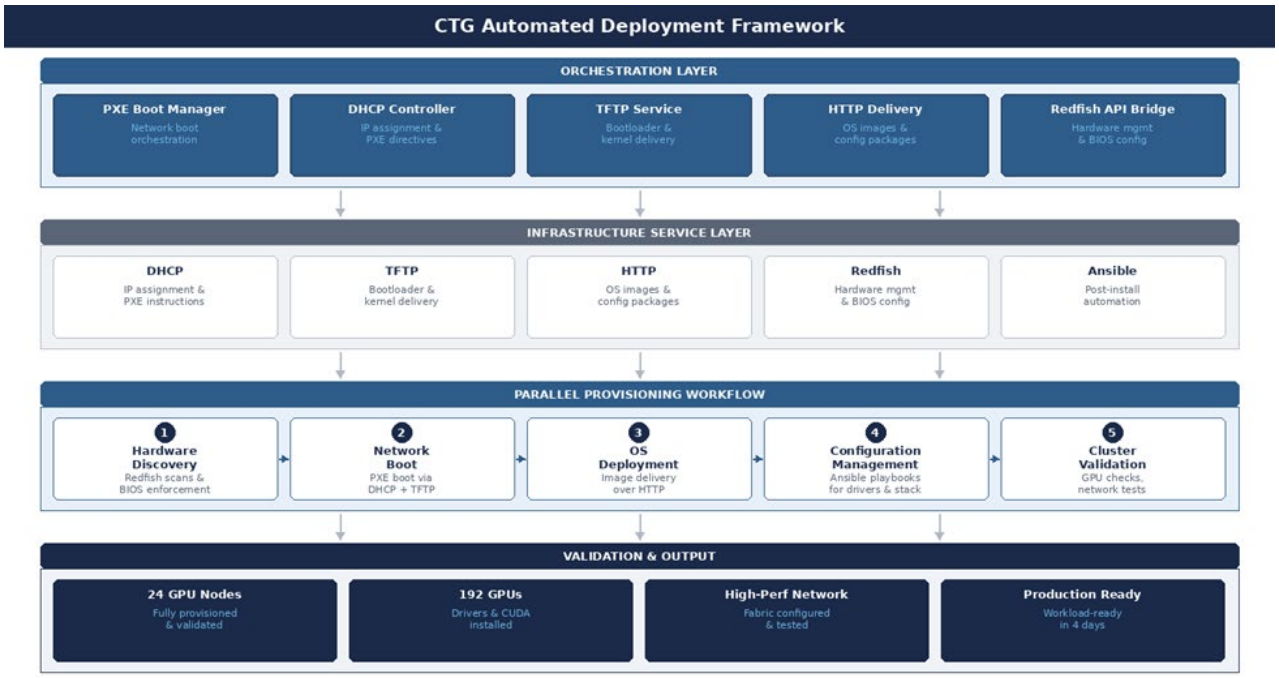

Figure 1: CTG Automated Deployment Framework

The framework operates through four integrated layers:

- Orchestration Layer coordinates all provisioning stages from a single control plane
- Infrastructure Services handle DHCP, TFTP, HTTP, Redfish, and Ansible independently
- Parallel Workflow provisions all nodes simultaneously, not sequentially
- Validation Output confirms production readiness across 24 nodes and 192 GPUs

Benefits: What This Delivers for Your Organization

Speed: From Weeks to Days

The CTG approach compresses a four-to-six week deployment into four days:

Traditional Approach	CTG Automated Approach
4–6 weeks end-to-end	4 days end-to-end
Nodes configured one at a time	All 24 nodes provisioned in parallel
Manual BIOS setup per server	Redfish API configures all nodes at once
USB/media-based OS installs	PXE network boot with HTTP delivery
Hand-applied driver packages	Ansible automates GPU/CUDA stack
Ad-hoc validation and testing	Automated validation across all nodes

What this means for you:

An 85%+ reduction in deployment time that translates directly into faster time-to-value. Your AI teams start running workloads weeks sooner, and your infrastructure investment begins producing returns from day one.

Repeatability: Deploy Once, Use Everywhere

The CTG framework is a reusable platform, not a one-time script. Once the orchestration and Ansible playbooks are tested for a given configuration, deploying the next cluster requires no additional development.

This delivers:

- Zero configuration drift — Node 1 and node 24 receive identical OS images, drivers, CUDA stacks, and cluster configs
- No forgotten steps — Every action is encoded in automation, not dependent on engineer memory
- Faster subsequent deployments — The second cluster deploys just as fast as the first
- Lower lifetime cost — The framework becomes a standing asset, not a one-time project expense

Auditability: Know Exactly What Was Deployed

In regulated environments, demonstrating exactly what was deployed, when, and how is not optional. The CTG platform generates a complete audit trail automatically. The deployment record captures:

- BIOS settings applied to each node via Redfish API
- Operating system image version installed on each machine
- GPU driver and CUDA versions deployed
- Network and cluster configuration parameters applied
- Validation test results with timestamps and pass/fail outcomes

This traceability is difficult to achieve through manual processes and nearly impossible to reconstruct after the fact. With the CTG framework, it is automatic.

Scalability: Built to Grow With You

The framework is not limited to a 24-node deployment. Scaling to 48, 96, or more nodes requires no fundamental changes to the automation.

- Adding nodes to an existing cluster follows the same automated path
- Standing up an entirely new cluster reuses the same playbooks and orchestration
- Your investment in the first deployment pays dividends on every subsequent one

Reduced Operational Risk

Manual configuration is the leading source of deployment errors in data center environments. Automation removes this risk:

- Same configuration applied to every node through the same code path, every time
- Changes made once and applied uniformly — no node-by-node updates
- No missed BIOS settings, mismatched drivers, or inconsistent network configs
- Fundamentally more reliable than manual processes at any scale

Deployment Lifecycle

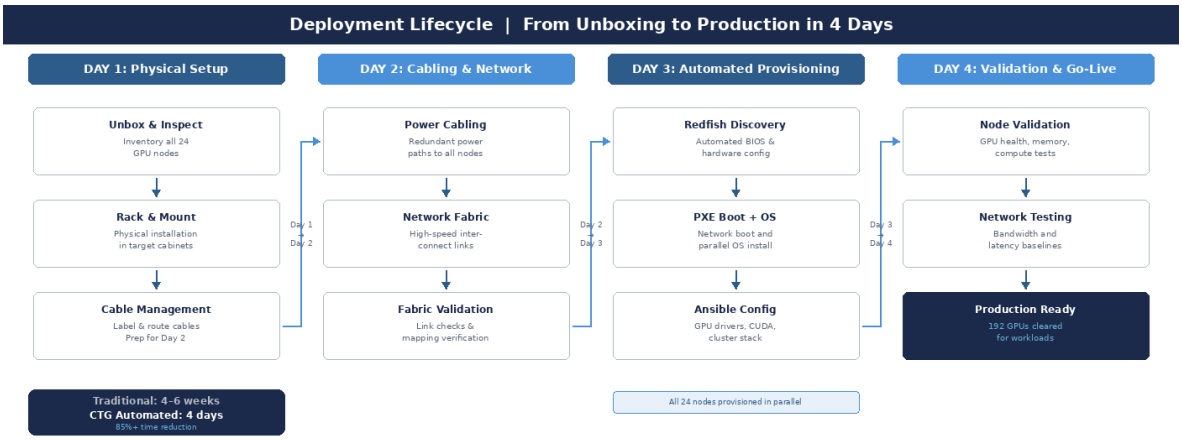

Figure 2: Deployment Lifecycle — From Unboxing to Production in 4 Days

Core Infrastructure Services

Component	What It Does for Your Deployment
DHCP	Assigns IP addresses and delivers PXE boot instructions so servers begin provisioning the moment they power on.
PXE Boot	Enables network-based bare-metal provisioning — no USB drives, optical media, or manual OS installation required.
TFTP	Delivers bootloaders and kernel images during PXE boot, keeping the boot chain lightweight and fast.
HTTP	High-speed delivery of OS images and config packages, supporting parallel downloads across all nodes simultaneously.
Redfish APIs	Automated hardware management and BIOS configuration from a central control point — vendor-agnostic across Dell, HPE, Lenovo, and others.
Ansible	Post-install automation including OS hardening, GPU driver installation, CUDA deployment, network setup, and validation.

Deployment Results

Day 1	Day 2	Day 3	Day 4
Physical Setup Unbox, inspect, and rack all 24 GPU nodes	**Cabling & Network** Redundant power, high-speed fabric, management connectivity	**Automated Provisioning** All 24 nodes provisioned in parallel: BIOS, OS, drivers, CUDA, cluster stack	**Validation & Go-Live** GPU health checks, network testing, 192 GPUs cleared for production

Infrastructure deployed:

- 24 GPU compute nodes with 192 total GPUs
- High-performance networking fabric connecting all nodes
- Fully automated OS, GPU drivers, CUDA, and cluster configuration
- Every node validated and cleared for production workloads

Within four days, the cluster progressed from unopened hardware to a fully operational AI compute environment.

Conclusion

CTG Federal built a deployment practice that turns raw GPU hardware into production-ready NeoCloud infrastructure in four days. By combining open infrastructure protocols, AI-assisted development, and infrastructure-as-code automation, the framework eliminates the traditional bottleneck between procurement and production.

For organizations navigating hardware supply variability, vendor diversity, or rapid scaling requirements, the CTG approach delivers:

- Speed — 85%+ reduction in deployment time, from weeks to days
- Repeatability — Same automation, same result, every time
- Auditability — Complete deployment records generated automatically
- Vendor flexibility — Open standards that work across Dell, HPE, Lenovo, and others
- Scalability — Framework that grows with your infrastructure needs
- Bespoke fit — Rapidly tailored to your specific hardware, network, and operational requirements

This approach enables organizations to focus on running AI workloads rather than building the infrastructure to support them.

About the Author

Brandon Moore is a senior infrastructure architect at CTG Federal with more than 20 years of experience designing and engineering enterprise and federal data center environments. His work has focused on building high-performance computing, AI infrastructure, and large-scale data center platforms for Fortune 500 organizations and U.S. federal agencies.