

Lenovo Confluent

Deploy and Manage Clusters at Scale

Abstract

Modern HPC and AI infrastructures continue to grow in scale and complexity. Administrators need a management platform that can provision, monitor, secure, and maintain large-scale server fleets while reducing operational overhead and accelerating time-to-production.

Lenovo Confluent is an open-source cluster provisioning and management platform developed by Lenovo for large-scale HPC and AI environments. It provides a unified framework for hardware discovery, bare-metal provisioning, operating system deployment, lifecycle management, and automation. For GPU-accelerated environments, Confluent applies the same repeatable operating model to GPU nodes and enables deployment workflows to include GPU drivers and the networking software stack required by the target solution.

Accelerate deployment

Move from discovered hardware to a configured cluster through repeatable, automated workflows.

Operate consistently

Manage heterogeneous node roles, architectures, and operating systems from a common interface.

Scale with confidence

Use Collective Mode for active/active load distribution, high availability, and large fleets.

Lenovo

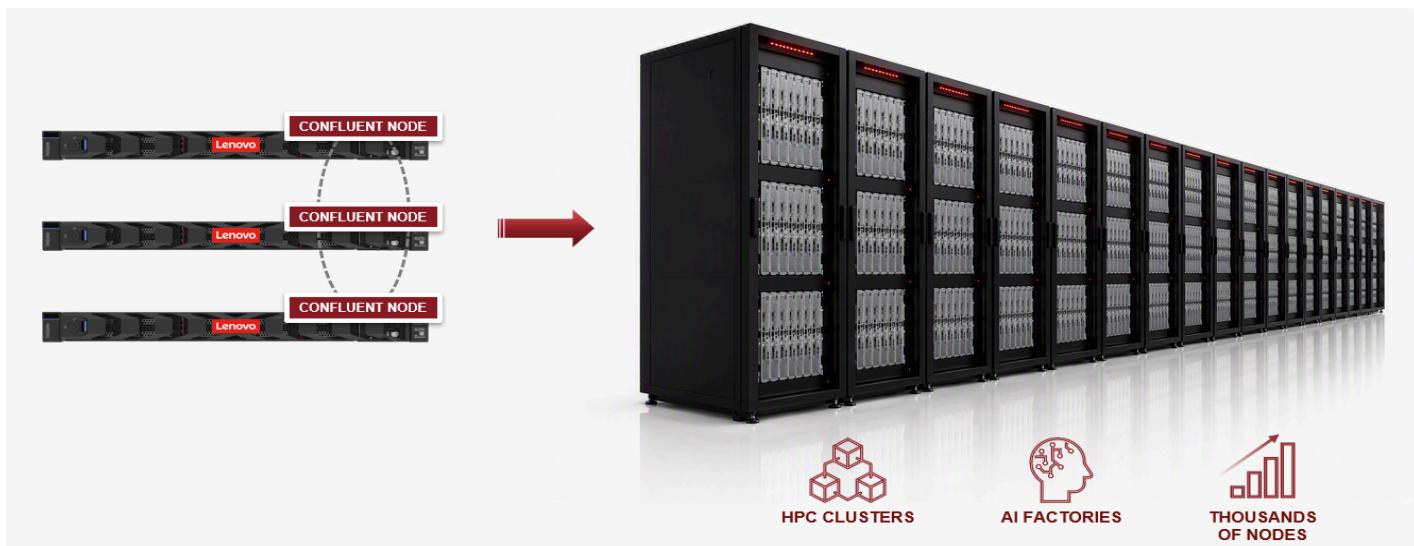


Figure 1. One management plane for cluster provisioning and managing ongoing operations.

Why Lenovo Confluent?

Automated cluster provisioning

Discover, configure, boot, and deploy servers at scale, reducing repetitive node-by-node tasks.

Scalability and availability

Extend a single management interface across multiple active Confluent servers for high availability with Collective Mode.

Hardware lifecycle management

Centralize power, firmware, BIOS/UEFI/BMC settings, health, telemetry, and support data.

Flexible OS deployment

Support disk-based and diskless images using PXE, HTTP(s) boot, or virtual media, with customization throughout the deployment process.

Open and automation-ready

Use the web UI, CLI, REST API, and Ansible-based automation workflows.

Secure by design

Use TLS-secured communication, TPM2, Secure Boot, and SSH certificate-based management patterns.

Core capabilities and customer value

Capability	Customer value
Device onboarding	Automates discovery and inventory collection, helping administrators identify and onboard systems easily and consistently.
Hardware lifecycle management	Controls power and boot behavior; upgrades firmware; manages BIOS/UEFI/BMC settings; and reports health, telemetry, events, and firmware levels.
Collective Mode	Distributes management workloads across peer servers and provides active/active scale-out and high availability.
Operating system lifecycle	Builds and customizes images for deployment of disk-based or diskless operating systems across x86 and Arm systems, including cross-architecture Arm image builds from x86 Confluent nodes.
Automation interfaces	Combines Linux administration tools, a REST API, Python client library, and integration points for scripts and Ansible playbooks.
Unified web interface	Provides fleet monitoring, real-time telemetry, rack views, and multi-console access through a web-based experience.
Powerful admin toolbox	Provides a comprehensive set of command-line tools designed to simplify the daily administration of large-scale clusters.

Deployment architecture and sizing considerations

Confluent uses a client-server architecture. Management nodes run the "confluentd" service, which administrators interact with using the CLI, web UI, or REST API. Nodes are provisioned over the management network and controlled out-of-band through the BMC network.

Profile	When it fits	Design consideration
Standalone 1 Confluent node	Smaller environments or deployments where management HA is not required.	A physical or virtual server can be used. Size networking and storage for the selected deployment method.
Highly available 3 Collective members Shared storage	Environments that require quorum-based HA and active/active management.	Collective configurations require at least three quorum members. Shared access to /var/lib/confluent is strongly recommended.
Large-scale 5 or more members Shared storage	Very large clusters that need additional load distribution and fault tolerance.	Add active members according to deployment traffic, node count, topology, and failure-tolerance objectives.

Sizing note: The profiles above are positioning examples, not final sizing rules. Validate the number of management nodes, shared storage design, and network bandwidth against the target node count, deployment model, boot concurrency, topology, and availability requirements.

HPC and AI ecosystem integration

Confluent provides the provisioning and infrastructure-management foundation for GPU-accelerated HPC and AI clusters. GPU servers are managed as defined node roles, enabling administrators to standardize software, operating system deployments, and infrastructure configurations across the environment. Through operating system images, automated deployment workflows, configuration management, and integration with external tools such as Ansible, Confluent can be used to deploy and maintain the software components commonly found in modern HPC and AI platforms.

- **GPU driver and software:** Deploy NVIDIA CUDA and AMD ROCm software stacks, including data center GPU drivers and supporting packages, as part of disk-based or diskless operating system definitions. Groups and reusable node definitions help maintain a consistent GPU software environment and configuration across the cluster.
- **Network driver:** Deploy supported networking drivers and software, including NVIDIA DOCA-OFED for InfiniBand and RoCE environments, as well as the Cornelis Omni-Path software stack.
- **Operating systems and hypervisors:** Confluent supports the deployment of Red Hat Enterprise Linux, Rocky Linux, Ubuntu, AlmaLinux, Oracle Linux, SUSE Linux Enterprise Server, and Debian, as well as virtualization platforms such as VMware ESXi and Proxmox. Operating systems can be deployed in disk-based or diskless configurations across supported x86 and Arm platforms.
- **High-Performance storage:** Confluent can be used to deploy and configure environments based on leading parallel file systems, including IBM Storage Scale, Lustre, and BeeGFS, helping ensure consistent configuration of storage clients, servers, and supporting services.
- **Observability and monitoring:** Confluent can bootstrap commonly used monitoring and observability platforms, including Icinga2, Grafana, Prometheus, Alertmanager, and Loki. These tools complement Confluent's native hardware-health and telemetry capabilities with broader infrastructure, metrics, alerting, and log visibility.
- **Scheduling and resource orchestration:** Confluent can prepare the infrastructure and system foundation for environments using NVIDIA Slurm, NVIDIA Slinky, Siemens PBS Professional, Siemens Grid Engine, IBM Spectrum LSF, and Kubernetes distributions.
- **Energy monitoring and optimization:** Confluent can support the deployment of EAR, the Energy Aware Runtime, alongside the selected scheduler or orchestration platform. EAR provides capabilities for monitoring energy consumption and optimizing cluster power efficiency.

Example HPC and AI platforms

Confluent provides the infrastructure foundation upon which higher-level HPC and AI platforms can be deployed. Examples include:

- A Slurm-based HPC cluster for research, engineering, simulation, and technical computing.
- A Kubernetes-based platform for AI training, inference, and containerized services.
- A hybrid HPC/AI environment combining Slurm and Kubernetes, including architectures that use NVIDIA Slinky.

Positioning note: These are solution examples rather than an exhaustive list of supported platforms. Confluent handles the hardware and operating-system foundation; scripts, configuration workflows, and Ansible can be used to bootstrap the selected platform layer.

Business and operational outcomes

Faster time-to-production

Standardize discovery, provisioning, and configuration from the first node to the full cluster.

Lower operational friction

Bring provisioning, hardware control, telemetry, consoles, and automation into one management plane.

Greater consistency

Apply reusable definitions across compute, GPU, login, storage, and other node roles.

Built for growth

Move from a standalone server to a resilient Collective architecture without changing the operational model.

Lenovo support and resources

While **Confluent remains fully open source**, Lenovo offers subscription-based support per managed node, providing customers with Lenovo-backed assistance for their Confluent deployment. Available terms include:

Table 1. Lenovo Confluent Support

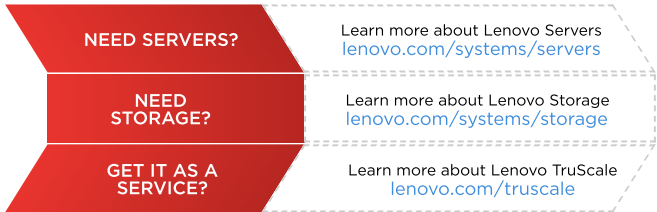
Part number	Feature code	Description
7S090039WW	S9VH	Lenovo Confluent 1 Year Support per managed node
7S09003AWW	S9VJ	Lenovo Confluent 3 Year Support per managed node
7S09003BWW	S9VK	Lenovo Confluent 5 Year Support per managed node
7S09003CWW	S9VL	Lenovo Confluent 1 Extension Year Support per managed node

For more information, consult these resources:

- [Implementing Lenovo Confluent Management Software](#)
- [Lenovo Confluent Online Documentation](#)
- [Lenovo EveryScale HPC & AI Software Stack](#)

About Lenovo

Lenovo is a US\$83 billion revenue global technology powerhouse, ranked #153 in the Fortune Global 500, and serving millions of customers every day in 180 markets. Guided by its **vision** of “Smarter Technology for All”, Lenovo is executing a **Hybrid AI strategy** that spans Personal AI – one personal AI, multiple devices; and Enterprise AI – helping customers turn data into insights and value. This strategy is delivered through the Group’s commitment to **world-class innovation** and a full-stack AI portfolio, including devices, infrastructure solutions, software, and services. Lenovo is widely recognized for its **operational excellence**, with a global footprint spanning more than 20 R&D locations and a supply chain that includes more than 30 manufacturing sites across 10 markets.



© 2026 Lenovo. All rights reserved.

Availability: Offers, prices, specifications and availability may change without notice. Lenovo is not responsible for photographic or typographic errors. **Warranty:** For a copy of applicable warranties, write to: Lenovo Warranty Information, 1009 Think Place, Morrisville, NC, 27560. Lenovo makes no representation or warranty regarding third-party products or services. **Trademarks:** Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo. AMD® and ROCm® are trademarks of Advanced Micro Devices, Inc. Linux® is the trademark of Linus Torvalds in the U.S. and other countries. IBM® and IBM Spectrum® are trademarks of IBM in the United States, other countries, or both. NVIDIA®, CUDA®, and NVIDIA DOCA® are trademarks of NVIDIA Corporation. Other company, product, or service names may be trademarks or service marks of others. Document number DS0217, published September 30, 2026. For the latest version, go to lenovopress.lenovo.com/ds0217.