NVIDIA Run:ai on ThinkSystem Servers
Solution Brief

As AI workloads mature from pilot experimentation to enterprise-scale production, organizations face increased pressure to operationalize machine learning efficiently, maximize infrastructure ROI, and support ever-expanding AI teams. In partnership with NVIDIA, Lenovo introduces a unified solution that accelerates this journey: NVIDIA Run:ai on Lenovo AI Platforms.

This powerful combination addresses common friction points across the AI lifecycle — from experimentation to deployment — by unifying GPU resource management, improving workload orchestration, and supporting cross-functional collaboration across IT and data science teams.

By leveraging Lenovo's 285 and 289 AI infrastructure and NVIDIA Run:ai's intelligent GPU orchestration platform, enterprises can fully unlock the value of their AI investments, scale operationally with confidence, and reduce time-to-insight for data-driven outcomes.

Business and Technical Challenges

Despite substantial investment in AI hardware and software, many organizations struggle to efficiently scale their AI initiatives.

Key challenges include:

- For AI Practitioners:
  - Inconsistent access to GPU resources hampers experimentation and training cycles.
  - Fragmented environments delay progress from proof-of-concept to deployment.
  - Contention between teams results in idle time and lost productivity.
- For IT Leaders:
  - GPU infrastructure is often overprovisioned or underutilized due to lack of visibility.
  - Static resource allocation fails to align with dynamic AI workloads.
  - Difficulty enforcing usage policies across distributed teams and environments.
- For Executives:
  - AI investments yield diminishing returns without centralized orchestration.
  - Lack of observability across workloads delays AI roadmap execution.
  - Cloud overspend and infrastructure inefficiencies erode competitive advantage

Solution Overview: NVIDIA Run:ai on Lenovo Infrastructure

NVIDIA Run:ai is a Kubernetes-native AI workload orchestration platform designed to maximize the efficiency, agility, and governance of GPU resources in hybrid and on-prem environments. When deployed on Lenovo's purpose-built AI platforms, it delivers a scalable and flexible foundation for production-grade AI.

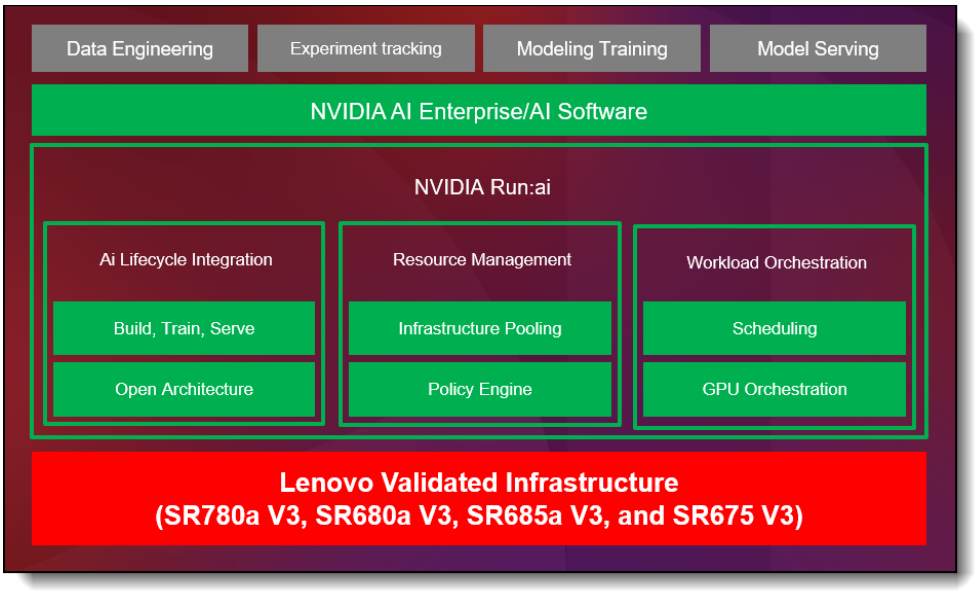

Figure 1. Solution Overview

Core capabilities of the solution:

- Fractional GPU allocation to optimize resource utilization.
- Priority-based workload scheduling to ensure mission-critical jobs are completed on time.
- Elastic scaling of training and inference jobs across distributed compute clusters.
- Lifecycle support for AI development, from Jupyter Notebooks to model serving.
- Policy-based governance for access control, security, and compliance.

NVIDIA Run:ai System Components

NVIDIA Run:ai is made up of two components both installed over a Kubernetes cluster. NVIDIA Run:ai control plane – Provides resource management, handles workload submission and provides cluster monitoring and analytics. NVIDIA Run:ai cluster – Provides scheduling and workload management, extending Kubernetes native capabilities

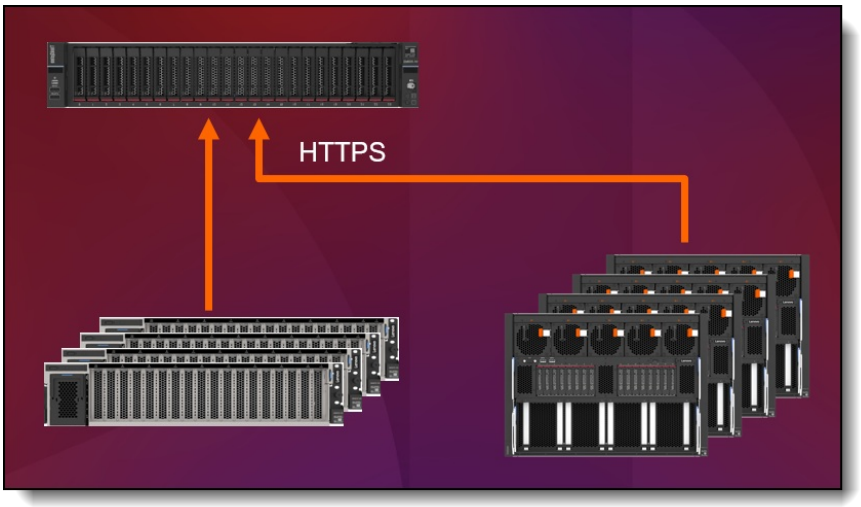

Figure 2. System components

The components are as follows:

- **Run:ai Control Plane**: Centralized resource management, user access policies, workload prioritization, built on Lenovo ThinksSystems. Refer to Control Plane System Requirements for specs and recommendations
- **Run:ai Cluster**: GPU scheduling, workload orchestration, Kubernetes-native scalability. Built on Lenovo AI server. Refer to the Lenovo Hybrid AI 285 Platform Guide for specs and recommendations

Role-Based Value Proposition

NVIDIA Run:ai software delivers distinct value to each stakeholder. Our co-tailored solution aligns with the priorities of AI practitioners, IT managers, and platform admins—driving technical efficiency, operational control, and strategic impact.

For IT Managers:

- **Centralized Control**: Manage multiple GPU clusters from a single console.
- **Usage Analytics**: Gain insights into GPU allocation, job performance, and bottlenecks.
- **Policy Enforcement**: Set consumption thresholds, scheduling rules, and user permissions.
- **Authentication & RBAC**: Integrate with enterprise identity platforms (e.g., LDAP, SSO).
- **Kubernetes-Native Design**: Install and manage using familiar cloud-native operations.

For AI Practitioners:

- **Self-Service GPU Access**: Launch training, fine-tuning, or inference jobs on-demand.
- **Interactive Development**: Run uninterrupted Jupyter Notebook sessions using fractional GPUs.
- **Model Lifecycle Integration**: From data prep to deployment — with support for key tools (PyTorch, TensorFlow, Ray, Kubeflow).
- **Scalable Training & Serving**: Leverage multiple GPUs with support for auto-scaling.

For Platform Admins:

- **Team Structuring**: Map projects, teams, and departments for intelligent resource allocation.
- **User and Access Control**: Assign permissions aligned to org structure and security policies.
- **Scheduling and Monitoring**: Allocate resources based on workload priority and urgency.
- **Cost Optimization**: Reduce idle GPU time and increase infrastructure ROI.

Subscription model and Part number information

Run:ai is licensed per GPU with options for education, enterprise, and public sector usage. The following table lists the ordering part numbers from Lenovo.

Table 1. NVIDIA Run:ai

Part number	Feature 7S02CTO1WW	NVIDIA part number	Description
Software subscription			
7S02004UWW	SDYT	744-RA7001+P3CMI12	NVIDIA Run:ai Subscription per GPU 1 Year
7S02004XWW	SDYW	744-RA7001+P3CMI36	NVIDIA Run:ai Subscription per GPU 3 Years
7S020050WW	SDYZ	744-RA7001+P3CMI60	NVIDIA Run:ai Subscription per GPU 5 Years
7S02004VWW	SDYU	744-RA7001+P3EDI12	NVIDIA Run:ai Subscription per GPU EDU 1 Year
7S02004YWW	SDYX	744-RA7001+P3EDI36	NVIDIA Run:ai Subscription per GPU EDU 3 Years
7S020051WW	SDZ0	744-RA7001+P3EDI60	NVIDIA Run:ai Subscription per GPU EDU 5 Years
7S02004WWW	SDYV	744-RA7001+P3INI12	NVIDIA Run:ai Subscription per GPU INC 1 Year
7S02004ZWW	SDYY	744-RA7001+P3INI36	NVIDIA Run:ai Subscription per GPU INC 3 Years
7S020052WW	SDZ1	744-RA7001+P3INI60	NVIDIA Run:ai Subscription per GPU INC 5 Years
Support Services subscription			
7S020053WW	SDZ2	744-RA7002+P3CMI12	24x7 Support Services for NVIDIA Run:ai Subscription per GPU 1 Year
7S020056WW	SDZ5	744-RA7002+P3CMI36	24x7 Support Services for NVIDIA Run:ai Subscription per GPU 3 Years
7S020059WW	SDZ8	744-RA7002+P3CMI60	24x7 Support Services for NVIDIA Run:ai Subscription per GPU 5 Years
7S020054WW	SDZ3	744-RA7002+P3EDI12	24x7 Support Services for NVIDIA Run:ai Subscription per GPU EDU 1 Year
7S02005AWW	SDZ9	744-RA7002+P3EDI60	24x7 Support Services for NVIDIA Run:ai Subscription per GPU EDU 5 Years
7S020057WW	SDZ6	744-RA7002+P3EDI36	24x7 Support Services for NVIDIA Run:ai Subscription per GPU EDU 3 Years
7S020055WW	SDZ4	744-RA7002+P3INI12	24x7 Support Services for NVIDIA Run:ai Subscription per GPU INC 1 Year
7S020058WW	SDZ7	744-RA7002+P3INI36	24x7 Support Services for NVIDIA Run:ai Subscription per GPU INC 3 Years
7S02005BWW	SDZA	744-RA7002+P3INI60	24x7 Support Services for NVIDIA Run:ai Subscription per GPU INC 5 Years

Author

Carlos Huescas is the Worldwide Product Manager for NVIDIA software at Lenovo. He specializes in High Performance Computing and AI solutions. He has more than 15 years of experience as an IT architect and in product management positions across several high-tech companies.

Related product families

Product families related to this document are the following:

- AI Servers
- Artificial Intelligence

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION ”AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP2254, was created or updated on July 10, 2025.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2254
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2254.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®
ThinkSystem®

Other company, product, or service names may be trademarks or service marks of others.