Solution Overview

NVIDIA Run:ai

AI Workload and GPU Orchestration

The Hidden Challenges in AI Infrastructure

AI is rapidly transforming industries, but enterprises face critical challenges in managing AI infrastructure efficiently. Organizations adopting AI often struggle with:

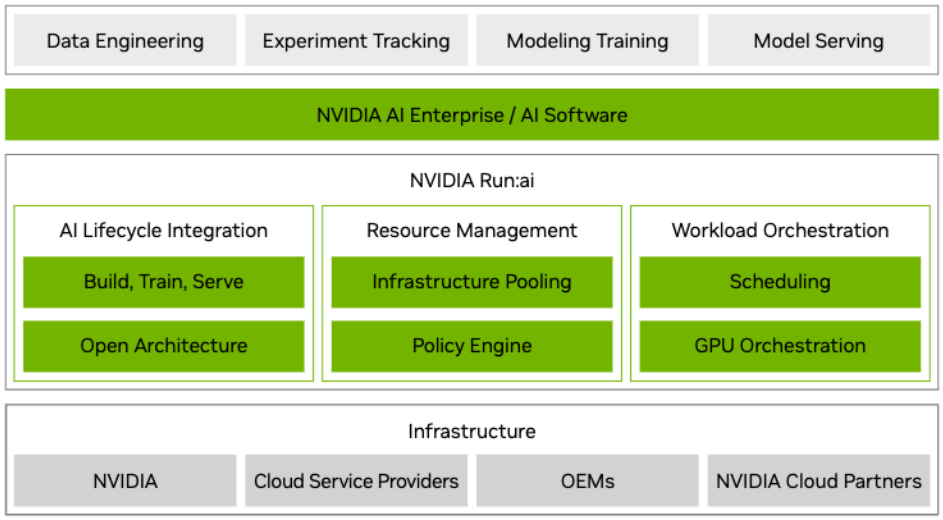
- Underutilized GPU resources, resulting in higher operational costs and inefficiencies, making AI investments less effective. Many AI workloads require dynamic scaling, but fixed configurations lead to idle GPUs, reducing overall return on investment and limiting scalability.
- Fragmented AI infrastructure makes it difficult to scale AI workloads across hybrid and multi-cloud environments, leading to management complexity. Organizations often deploy AI across multiple platforms, each with its own orchestration and governance requirements, creating inefficiencies and security risks.
- Lengthy AI development cycles delay business impact, making it harder for organizations to innovate and compete effectively. The time required for provisioning resources, configuring workloads, and managing dependencies slows AI projects, delaying real-world deployment and return on AI investments.
- Legacy infrastructure forces organizations into predefined allocations, preventing seamless experimentation, rapid iteration, and resource-sharing across teams.
- Ineffective workload orchestration leads to bottlenecks, preventing organizations from fully utilizing compute resources and optimizing performance. Default scheduling mechanisms struggle with AI-specific needs, failing to balance multiple priorities, allocate resources dynamically, and ensure optimal job execution.

Without a modern workload and GPU orchestration solution, these challenges slow down AI adoption, increase costs, and create inefficiencies that limit an organization's ability to realize AI-driven business value.

Key Challenges

- **Low GPU Utilization –** Inefficient allocation leads to wasted compute power and high costs.
- **Infrastructure Fragmentation –** Managing AI across multiple environments creates complexity and inefficiencies.
- **Slow AI Deployment –** Lengthy provisioning and dependency management slow innovation and time to market.
- **Static Resource Allocation –** Limits flexibility and responsiveness.
- **Inefficient Orchestration –** Default schedulers fail to optimize AI workloads, leading to bottlenecks.

Accelerating AI Workloads at Scale



NVIDIA Run:ai enables dynamic AI workload and GPU orchestration to build, train, and deploy AI workloads at scale.

NVIDIA Run:ai accelerates AI and machine learning operations by addressing these key infrastructure challenges through dynamic resource allocation, comprehensive AI lifecycle support, and strategic resource management. As a unified AI workload and GPU orchestration platform, NVIDIA Run:ai represents a transformative approach to managing and optimizing AI resources and operations within an enterprise, designed to overcome the inherent challenges in traditional AI infrastructure by being dynamic, strategic, and integrally aligned with business objectives.

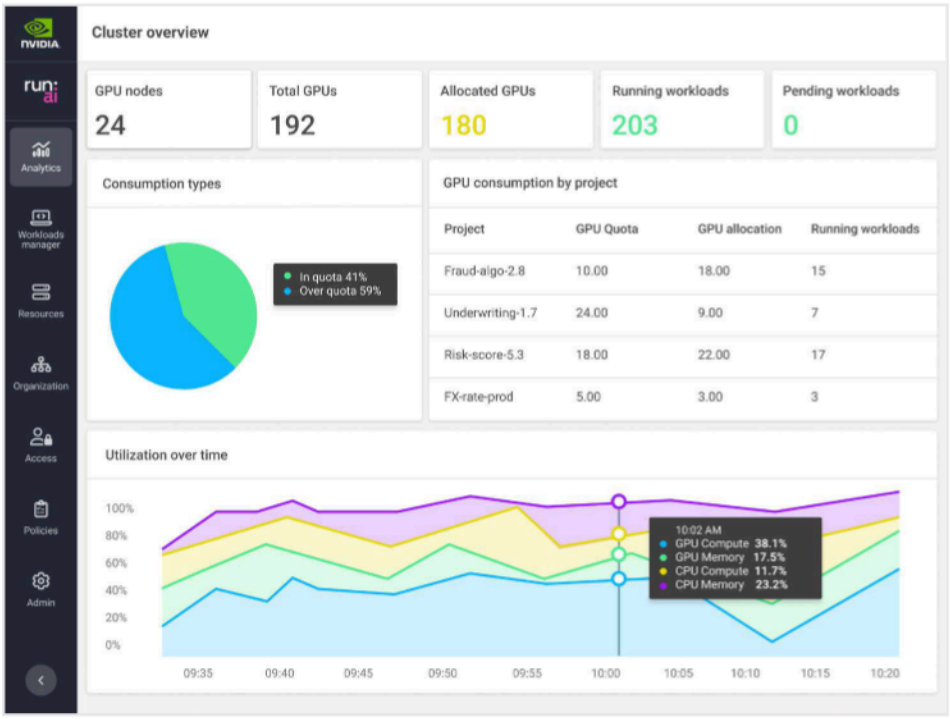
AI Lifecycle Integration

AI success depends on smooth transitions between development, training, and inference. NVIDIA Run:ai provides a flexible, open architecture that integrates seamlessly with leading ML tools, frameworks, and infrastructure, including:

- **Self-serve access to ML tools and compute –** NVIDIA Run:ai provides data scientists and engineers with on-demand access to ML environments, reducing setup time and increasing productivity.
- **Optimized AI framework support –** The NVIDIA Run:ai platform is optimized for the NVIDIA AI Enterprise software suite, which offers preconfigured, secure containers that simplify deployment of popular frameworks like TensorFlow, PyTorch, and Jupyter Notebooks, ensuring compatibility with enterprise AI workloads.
- **API-driven architecture for third-party integrations –** NVIDIA Run:ai enables seamless integration with preferred tools and frameworks, providing organizations with flexibility in managing and scaling AI workloads.

NVIDIA Run:ai Key Benefits

- **Maximized GPU Efficiency –** Intelligent resource pooling and dynamic allocation reduce costs and improve ROI.
- **Unified AI Infrastructure –** Centralized orchestration simplifies management across hybrid and multi-cloud environments.
- **Accelerated AI Lifecycle –** Seamless workflow integration speeds up development, training, and deployment.
- **Flexible and Scalable Compute –** Adaptive infrastructure supports evolving AI workloads and business needs.
- **Optimized Workload Performance –** Advanced scheduling and orchestration eliminate bottlenecks for high-performance AI workloads.



NVIDIA Run:ai provides MLOps managers and admins with a unified dashboard where they can see their entire AI/ML compute usage by team, jobs, and more.

NVIDIA Run:ai Key Results[1]

- **10X GPU Availability –** More access to GPUs empowers data scientists to accelerate AI innovation.
- **20X Workloads Running –** Increased pace of development enables organizations to scale AI initiatives faster.
- **5X GPU Utilization –** Optimized resource efficiency eliminates idle GPUs and maximizes performance.
- **Zero Manual Resource Intervention –** Intelligent automation removes bottlenecks and streamlines operations.

Resource Management

Efficiently managing GPU and compute resources is critical to AI scalability. NVIDIA Run:ai features policy-driven resource pooling, which ensures maximum utilization and cost efficiency by dynamically adapting to fluctuating AI demands.

- **Dynamic GPU allocation –** NVIDIA Run:ai adapts compute resources to match the real-time needs of AI workloads. By dynamically allocating GPU resources, organizations can avoid idle hardware and reduce costs, ensuring high performance even during peak demand.
- **Policy engine –** NVIDIA Run:ai enables organizations to easily define and adjust rules for resource consumption in alignment with business priorities. It provides a simple interface to express policies that prioritize critical workloads, optimize resource usage, and adapt quickly to changing business requirements.
- **Centralized orchestration –** NVIDIA Run:ai offers a unified interface for managing GPU resources across on-premises, cloud, and hybrid environments. This centralized control reduces complexity and provides organizations with the flexibility to scale their AI initiatives seamlessly.

Workload Orchestration

NVIDIA Run:ai ensures that AI workloads run efficiently and at scale, whether for development, training, or inference. The platform's advanced orchestration capabilities maximize GPU efficiency and support seamless scaling across diverse environments.

- **Advanced AI scheduling –** NVIDIA Run:ai employs intelligent scheduling algorithms that dynamically allocate compute resources based on workload priority, resource availability, and real-time demand. This dynamic approach minimizes idle resources and ensures critical AI tasks receive priority, enhancing overall system performance.
- **GPU fractioning –** The NVIDIA Run:ai platform enables GPU fractioning, allowing GPUs to be divided into logical units. This capability supports mixed workloads by allocating just the right amount of GPU power to each task, optimizing hardware efficiency and allowing more workloads to run simultaneously.
- **Inference optimization –** NVIDIA Run:ai is designed to reduce cold start times, enabling faster model deployment and more responsive AI services. The platform's metrics-driven scaling, GPU memory swapping, and model streaming capabilities ensure inference workloads can scale rapidly to handle spikes in demand without compromising performance.

Run Anywhere. Scale Everywhere. Maximize AI Efficiency.

NVIDIA Run:ai delivers unparalleled flexibility and efficiency for AI workloads, enabling organizations to run seamlessly across on-premises, cloud, and hybrid environments. The platform's intelligent orchestration dynamically adapts to evolving business needs, ensuring efficient resource utilization and scalability.

By unifying GPU and compute management under a single platform, NVIDIA Run:ai optimizes performance, reduces idle resources, and accelerates AI initiatives. Automated policies make it simple to align resource consumption with strategic priorities, helping enterprises achieve faster time to market and maximize the return on their AI investments.

Through seamless integration with leading ML tools and platforms, NVIDIA Run:ai supports the entire AI lifecycle—from development and training to large-scale inference—providing organizations with a consistent and efficient path to AI-driven innovation.

Ready to Get Started?

To learn more, visit nvidia.com/run-ai

1. A leading US bank confirmed achieving 1,000% greater GPU availability, 2,000% increase in volume of workloads running, 500% increase in GPU utilization, and 0% manual intervention while using the Run:ai Platform in February 2024.

NVIDIA