

Overview

NVIDIA Run:ai is a GPU orchestration and optimization platform that helps organizations maximize compute utilization for AI workloads. By optimizing the use of expensive compute resources, NVIDIA Run:ai accelerates AI development cycles, and drives faster time-to-market for AI-powered innovations.

Built on Kubernetes, NVIDIA Run:ai supports dynamic GPU allocation, workload submission, workload scheduling, and resource sharing, ensuring that AI teams get the compute power they need while IT teams maintain control over infrastructure efficiency.

How NVIDIA Run:ai Helps Your Organization

For Infrastructure Administrators

NVIDIA Run:ai centralizes cluster management and optimizes infrastructure control by offering:

- **Centralized cluster management** – Manage all clusters from a single platform, ensuring consistency and control across environments.
- **Usage monitoring and capacity planning** – Gain real-time and historical insights into GPU consumption across clusters to optimize resource allocation and plan future capacity needs efficiently.

- **Policy enforcement** – Define and enforce security and usage policies to align GPU consumption with business and compliance requirements.
- **Enterprise-grade authentication** – Integrate with your organization's identity provider for streamlined authentication (Single Sign On) and role-based access control (RBAC).
- **Kubernetes-native application** – Install as a Kubernetes-native application, seamlessly extending Kubernetes for native cloud experience and operational standards (install, upgrade, configure).

For Platform Administrators

NVIDIA Run:ai simplifies AI infrastructure management by providing a structured approach to managing AI initiatives, resources, and user access. It enables platform administrators maintain control, efficiency, and scalability across their infrastructure:


- **AI Initiative structuring and management** – Map and set up AI initiatives according to your organization's structure, ensuring clear resource allocation.
- **Centralized GPU resource management** – Enable seamless sharing and pooling of GPUs across multiple users, reducing idle time and optimizing utilization.
- **User and access control** – Assign users (AI practitioners, ML engineers) to specific projects and departments to manage access and enforce security policies, utilizing role-based access control (RBAC) to ensure permissions align with user roles.
- **Workload scheduling** – Use scheduling to prioritize and allocate GPUs based on workload needs.
- **Monitoring and insights** – Track real-time and historical data on GPU usage to help track resource consumption and optimize costs.

For AI Practitioners

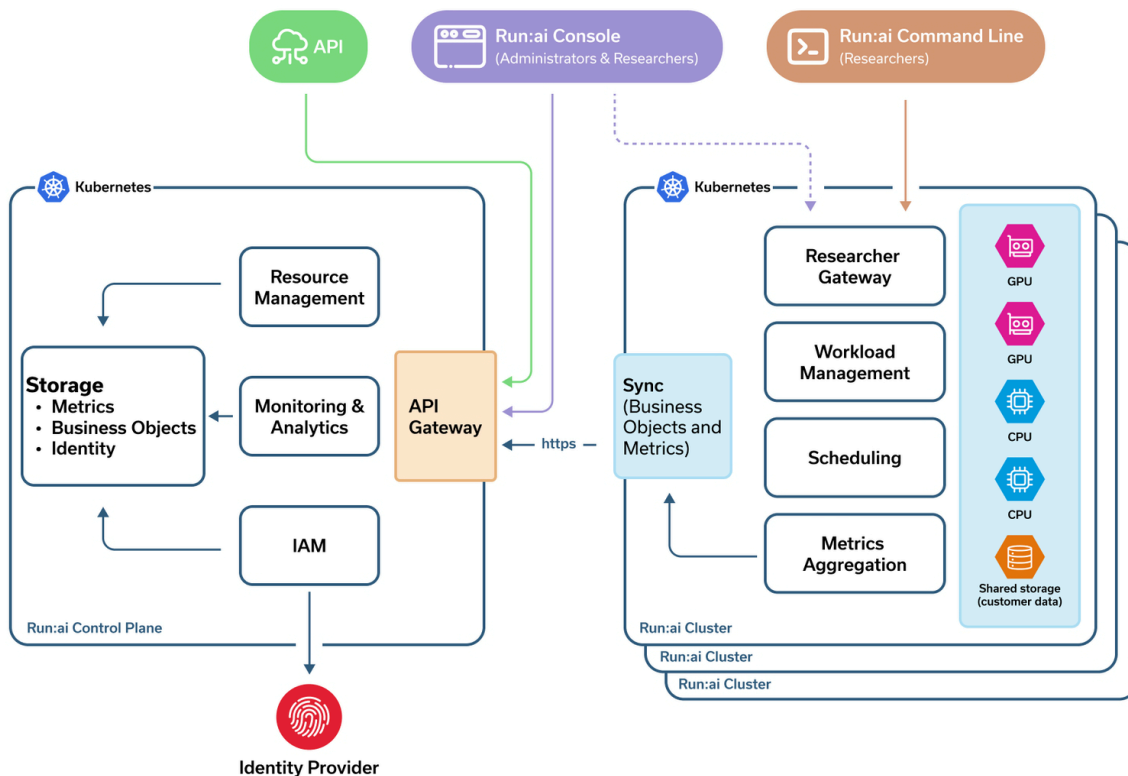
NVIDIA Run:ai empowers data scientists and ML engineers by providing:

- **Optimized workload scheduling** – Ensure high-priority jobs get GPU resources. Workloads dynamically receive resources based on demand.
- **Fractional GPU usage** – Request and utilize only a fraction of a GPU's memory, ensuring efficient resource allocation and leaving room for other workloads.
- **AI initiatives lifecycle support** – Run your entire AI initiatives lifecycle – Jupyter Notebooks, training jobs, and inference workloads efficiently.
- **Interactive session** – Ensure an uninterrupted experience when working on Jupyter Notebooks without taking away GPUs.
- **Scalability for training and inference** – Support for distributed training across multiple GPUs and auto-scales inference workloads.
- **Integrations** – Integrate with popular ML frameworks - PyTorch, TensorFlow, XGBoost, Knative, Spark, Kubeflow Pipelines, Apache Airflow, Argo workloads, Ray and more.
- **Flexible workload submission** – Submit workloads using the NVIDIA Run:ai UI, API, CLI or run third-party workloads.

NVIDIA Run:ai System Components

NVIDIA Run:ai is made up of two components both installed over a Kubernetes  cluster:

- **NVIDIA Run:ai cluster** – Provides scheduling and workload management, extending Kubernetes native capabilities.
- **NVIDIA Run:ai control plane** – Provides resource management, handles workload submission and provides cluster monitoring and analytics.



NVIDIA Run:ai Cluster

The NVIDIA Run:ai cluster is responsible for scheduling AI workloads and efficiently allocating GPU resources across users and projects:

- **NVIDIA Run:ai Scheduler** – Applies AI-aware rules to efficiently schedule workloads submitted by AI practitioners.
- **Workload management** – Handles workload management which includes the researcher code running as a Kubernetes container and the system resources required to run the code, such as storage, credentials, network endpoints to access the container and so on.
- **Kubernetes operator-based deployment** ↗ – Installed as a Kubernetes Operator to automate deployment, upgrades and configuration of NVIDIA Run:ai cluster services.
- **Storage** – Supports Kubernetes-native storage using Storage Classes ↗, allowing organizations to bring their own storage solutions. Additionally, it also integrates with external storage solutions such as Git, S3, and NFS to support various data requirements.

- **Secured communication** – Uses an outbound-only, secured (SSL) connection to synchronize with the NVIDIA Run:ai control plane.
- **Private** – NVIDIA Run:ai only synchronizes metadata and operational metrics (e.g., workloads, nodes) with the control plane. No proprietary data, model artifacts, or user data sets are ever transmitted, ensuring full data privacy and security.

NVIDIA Run:ai Control Plane

The NVIDIA Run:ai control plane provides a centralized management interface for organizations to oversee their GPU infrastructure across multiple locations/subnets, accessible via Web UI, [API](#) and [CLI](#). The control plane can be deployed on the cloud or on-premise for organizations that require local control over their infrastructure (self-hosted).

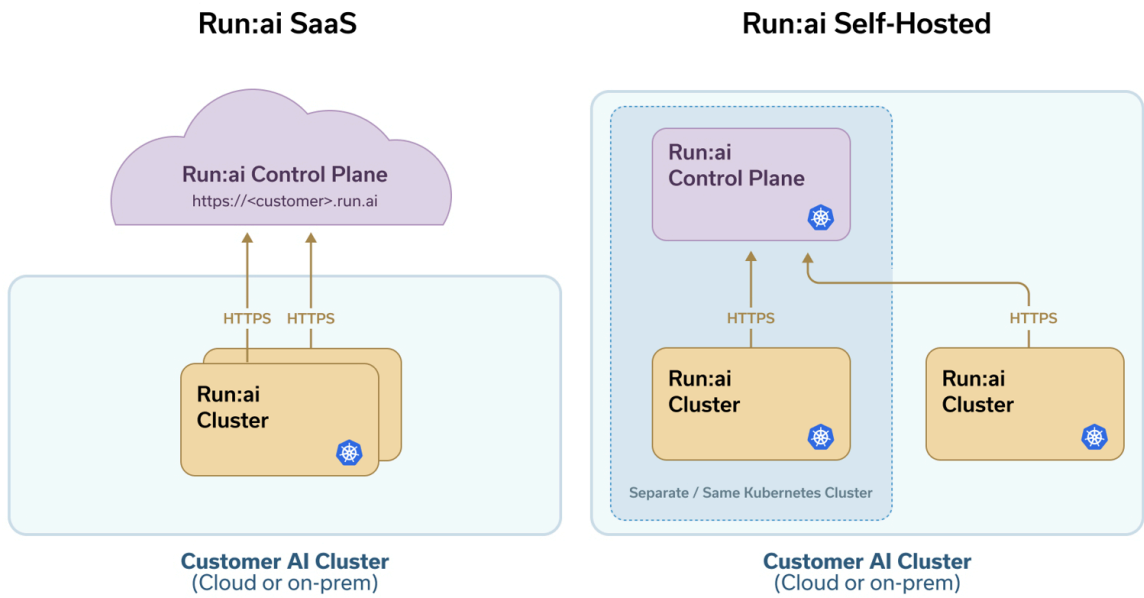
- **[Multi-cluster management](#)** – Manages multiple NVIDIA Run:ai clusters for a single tenant across different locations and subnets from a single unified interface.
- **[Resource and access management](#)** – Allows administrators to define Projects, Departments and user roles, enforcing policies for fair resource distribution.
- **[Workload submission and monitoring](#)** – Allows teams to submit workloads, track usage, and monitor GPU performance in real time.

Installation types

There are two main installation options:

Installation Type	Description
SaaS	NVIDIA Run:ai is installed on the customer's data science GPU clusters. The cluster connects to the NVIDIA Run:ai control plane on the cloud (<a href="https://<tenant-name>.run.ai">https://<tenant-name>.run.ai). With this installation, the cluster

	requires an outbound connection to the NVIDIA Run:ai cloud.
Self-hosted	The NVIDIA Run:ai control plane is also installed in the customer's data center



Run:ai Deployment Options

Next
What's New

Last updated 1 month ago



Corporate Info

[NVIDIA.com Home](#)

[About NVIDIA](#)

[Privacy Policy](#)

[Manage My Privacy](#)

[Terms of Service](#)

NVIDIA Developer

[Developer Home](#)

[Blog](#)

Resources

[Contact Us](#)

[Developer Program](#)

Copyright © 2025, NVIDIA Corporation.