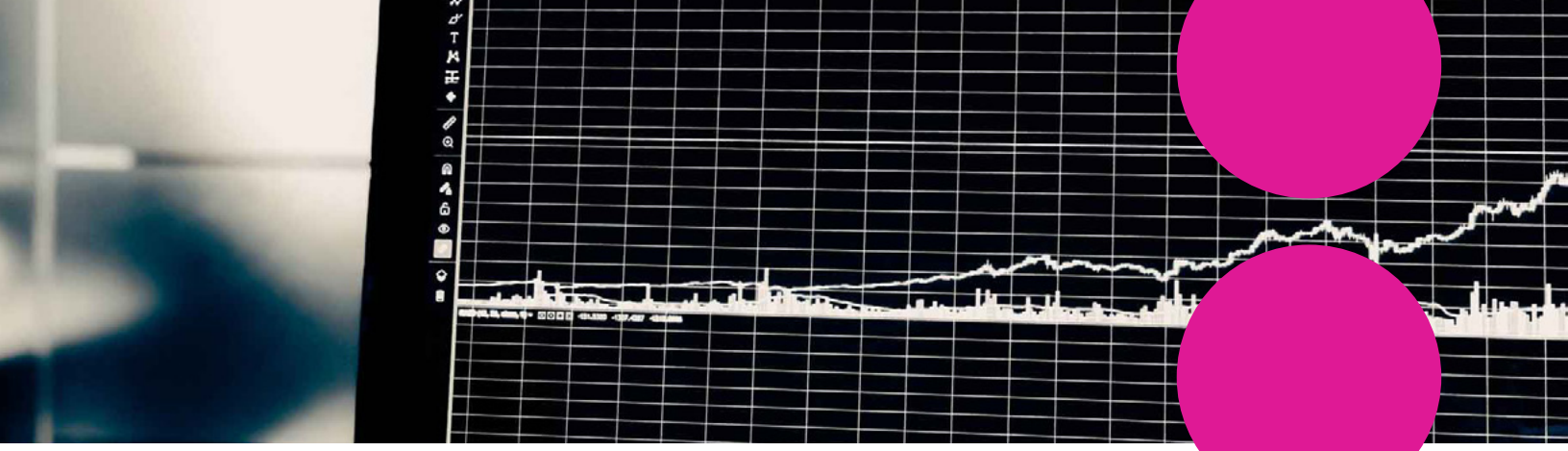

Run:ai MLOps Compute Platform powered by NVIDIA DGX Systems

Speeding the ROI of AI with Intelligent Infrastructure Utilization and Access

Artificial Intelligence (AI) is top of mind for many organizations as a way to reduce operational costs, increase efficiency, grow revenue and improve customer experience. To deliver on these AI-driven initiatives enterprises invest heavily in data science and AI development. Organizations need reliable AI infrastructure to support their AI practitioners, ensuring access to the resources they need to be successful. However, IT departments often encounter difficulty juggling individual systems, clusters, and multiple teams with differing priorities and needs.When differing needs are present, teams compete for GPU computing time, leading to delays and misalignment in AI workflows.

Perhaps the most common risk is the rise of shadow AI. It occurs when individual teams buy their own infrastructure or use cloud compute resources to quickly enable their AI initiatives. While no organization wants to stifle the ambition of its data science teams, this decentralized approach results in enormous waste, as well as overspending on compute resources. As AI infrastructure expands, administrators must simplify the developer experience, efficiently allocate resources, and avoid the dangers of shadow AI.

Solution

To maximize efficient utilization and ROI of AI infrastructure, Run:ai and NVIDIA offer the Run:ai MLOps Compute Platform (MCP) integrated joint solution. It includes world-class AI infrastructure with NVIDIA DGX™ systems, along with complete control and visibility of all compute resources with Run:ai Atlas in an easy-to-use solution. DGX systems deliver unmatched flexibility and AI performance, and combined with Run:ai Atlas provide an AI infrastructure solution that seamlessly orchestrates the complexities of AI development.

With the Run:ai MCP solution, compute resources are gathered into a centralized pool that can be managed and provisioned by one team, but delivered to many users with self-service access. An innovative, cloud-native operating system helps IT manage everything from fractions of GPUs to large-scale distributed training. Run:ai's workload-aware orchestration ensures that every type of AI workload gets the right amount of compute resources when needed. The solution preserves freedom for developers to use their preferred tools via integrations with Kubeflow, Airflow, MLflow and more. With Run:ai MCP, organizations benefit from a simplified deployment, and direct access to NVIDIA experts with world-class enterprise support.

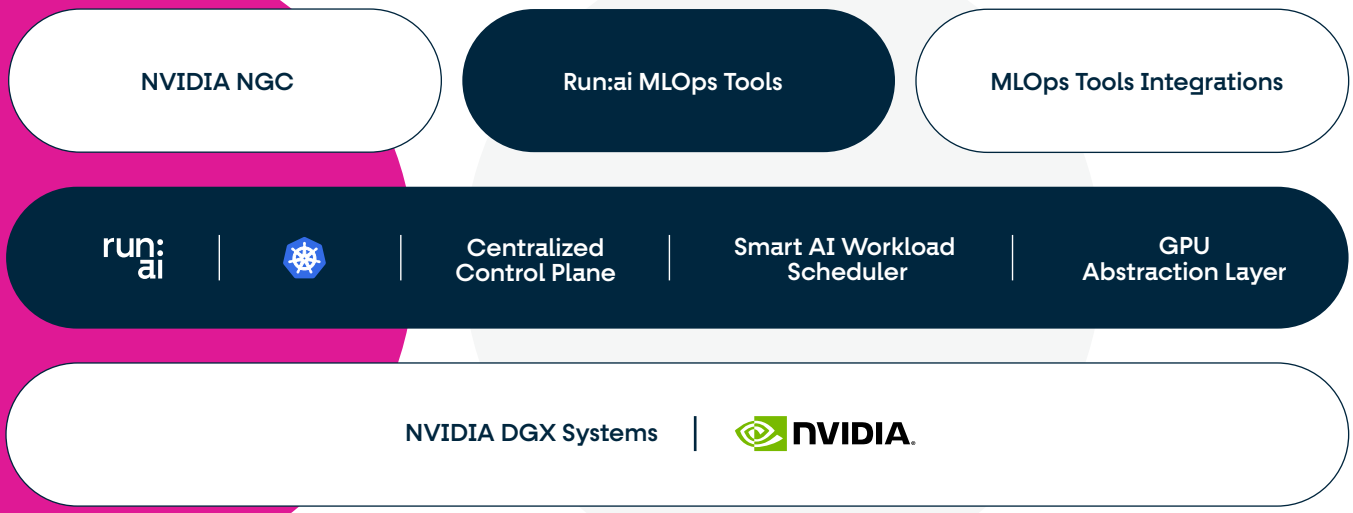

Figure 1.Solution Architecture

Key Benefits of MCP as a Full Stack Solution

Seamless Orchestration

MCP allows enterprises to speed time-to-insight and achieve faster ROI for AI initiatives with optimized MLOps workflows and world-class AI infrastructure. MCP provides seamless orchestration of MLOps workloads combined with the unmatched power of NVIDIA DGX systems. For data science teams large or small, MCP simplifies and accelerates the AI development process and helps organizations achieve their business goals

Centralized Control & Visibility

Run:ai Atlas offers dashboards and analytics giving IT insight across all resources and workloads. Align resource allocation to business goals by setting policies and priorities across departments, projects or users.

Optimal GPU Utilization

Automated resource management and efficient sharing of GPU resources enables organizations to achieve a higher utilization and therefore more value per GPU. Additionally, MCP takes full advantage of the advanced multi-GPU topology of DGX systems to ensure maximum performance.

Truly Open & Extensible

Use the built-in workflows in Run:ai Atlas which are optimized for the full AI development lifecycle or extend the platform by easily integrating 3rd party MLOps tools.

Accelerate Hybrid Cloud

Run:ai has the unique capability of delivering centralized control and visibility across resources that are located on-premises or in the cloud, enabling organizations to make hybrid cloud AI infrastructure a reality.

World-class AI Infrastructure

NVIDIA DGX systems are the universal AI system for every AI workload, offering unprecedented compute density, performance and flexibility. With integrated access to unmatched AI expertise, AI practitioners can get up and running quickly and stay running smoothly, dramatically improving time to insights. NVIDIA DGX systems underpin flexible AI infrastructure that scales throughout your AI journey.

Customer Story: Investment Banking Services Company

A leading US-based investment banking services company has established an AI center of excellence, to centralize AI infrastructure and increase governance without constraining data scientists' freedom and flexibility to use their preferred tools. Boasting the world's largest financial data set, this organization is using AI in a multitude of initiatives across the business, including client inquiry classification and processing, anomaly detection, risk management, forecasting trade fails and work order volume for resource planning.

AI Infrastructure and Team

- **Large, Growing Environment –** expanding the pool of accelerated computing resources for data scientists
- **Diverse Research Team –** with backgrounds in risk assessment and predictive AI with varying usage patterns

Challenges

- **Low Overall Utilization of AI Hardware –** Total GPU utilization was below 30%, with significant idle periods for some GPUs despite demand from researchers
- **Overloaded system with jobs requiring more resources –** The system was overloaded on multiple occasions where more GPUs were needed than were available for running jobs
- **Lack of IT Visibility and Governance –** Siloed and disparate technologies sprawled across the enterprise resulted in the desire to bring data science initiatives from the organization's peripheries to its core. There was also a need to centralize visibility and control of data to reduce regulatory and data compliance risks

Solution

The customer selected Run:ai MCP to address these challenges, consolidating stranded or siloed infrastructure and eliminating static allocation of resources. Pools of shared DGX systems were created allowing teams to access more compute resources, to run more workloads, and scale productivity. Many jobs are now submitted to the system by researchers every day, regardless of team, and jobs are queued and launched automatically by Run:ai Atlas when resources become available.

Run:ai Atlas' Smart AI workload scheduler running on Kubernetes enables crucial features for the management of DL workloads, like advanced queuing and quotas, managing priorities and policies, automatic preemption, multi-node training, and more. It provides an elegant solution to simplify complex scheduling processes, leading to more utilization of accelerated computing resources.

Outcomes

10X	20X	5X	0
GPU Availability	Workloads Running	GPU Utilization	Manual Resource Intervention
Enabled AI Innovation	Increased Pace of Development	Eliminated Idle Resources	Less Delays & Bottlenecks

Visit run.ai/DGX for more information or

Request a Demo