Centralizing AI Model Deployment on GPUs

Deploying machine learning models on GPUs can be a complex and resource-intensive task, especially when dealing with large datasets and multiple models. Run:ai's platform simplifies this process by centralizing AI model deployment on GPUs, giving you the control, visibility, and scalability you need to deploy your models with ease. Whether you're working with small or large models (like LLMs), Run:ai has the tools to help you right-size your resources and achieve your SLA goals.

Right-size your resources

Easily deploy your model servers on fractional GPUs (or MIG partitions), autoscale them as needed, and orchestrate their scheduling for maximum efficiency. Avoid over provisioning of resources and save cost by only using what you need when you need it.

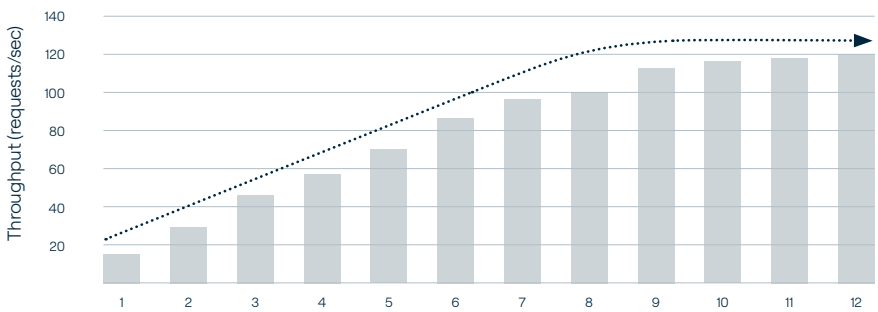
Control, Visibility and SLAs

SLA settings for your model deployment guarantees that your models will perform as required. Get insights into performance metrics, such as GPU utilization and model server response time, to ensure that your models are performing optimally. The platform provides centralized control and visibility across all your entire AI infrastructure so you can easily manage your resources and track performance.

Choose the Model Server You Want

Deploy single or multiple models per GPU and choose the model server that best suits your needs. The platform has built-in support for NVIDIA Triton and Seldon, two popular model serving frameworks that provide high-performance, scalable, and flexible solutions for ML model deployment.

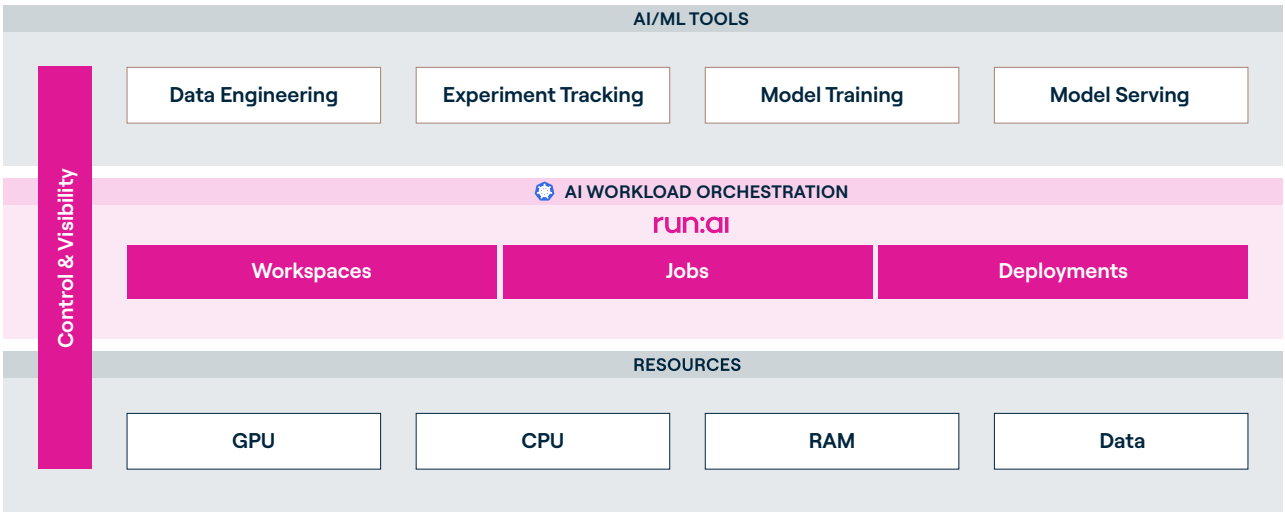
8X
THROUGHPUT



A100 | Replicas on a single GPU

Platform Overview

The Run:ai Atlas platform sits in between the infrastructure and the AI workloads that require access to these valuable resources. Platform teams gain centralized control and visibility across all AI infrastructure, whether on-premises or cloud. AI/ML teams get streamlined and self-service access to all the compute they need, when they need it, using the tools they prefer.



Feature Highlights

Autoscaling

Automatically scale model deployments up or down based on predefined thresholds using built-in or custom metrics, ensuring model SLAs are met and results in an optimal end-user experience.

Performance Monitoring

Get insights into model performance by drilling down into realtime and historical analytics regardless of where the models are deployed (on-premises or cloud).

Fractional GPU

Fractional GPU allows GPU resources to be shared without memory overflows or processing clashes. Using virtualized logical GPUs, with their own memory and computing space, containers can use and access GPU Fractions as if they were self-contained processors.

Customers Accelerating AI with Run:ai Atlas

SONY BNY MELLON ZEBRA xiaomi