Centralizing AI Training on GPUs

As AI workloads continue to grow, training large models across multiple nodes and GPUs becomes a daunting task. Run:ai's AI Compute Orchestration Platform provides a comprehensive solution for centralizing AI training on GPUs, optimizing AI workloads through distributed training, hyperparameter optimization, and more.

Distributed Training and HPO

Enable large-scale distributed training across multiple nodes and GPUs. With unattended training sessions, data scientists can easily create a training and/or HPO session and let the platform's scheduler manage the workload from start to finish.. This streamlines the training process, enabling data scientists to focus on developing better models and improving accuracy.

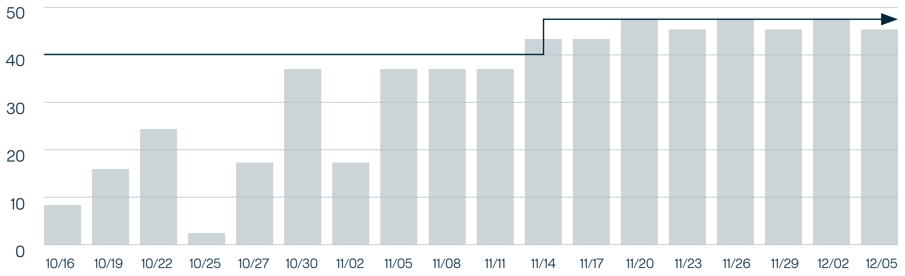
Utilize Idle Resources

Run:ai's platform pools the resources and uses dynamic quota-management that supports over-quota, enabling the platform to utilize idle or unused resources. The platform automates and optimizes the placement of workloads minimizing fragmentation and ensuring large-scale training can run efficiently. Unused resources and resources used by lower priority workloads are automatically released.

Ecosystem Integration

Easily integrate with MLOps tools like Kubeflow, W&B, Tensorboard and different distributed computing and training frameworks like Ray, Horovod and others. This enables practitioners to use their preferred tools while benefiting from Run:ai's workload scheduling and resource management capabilities.

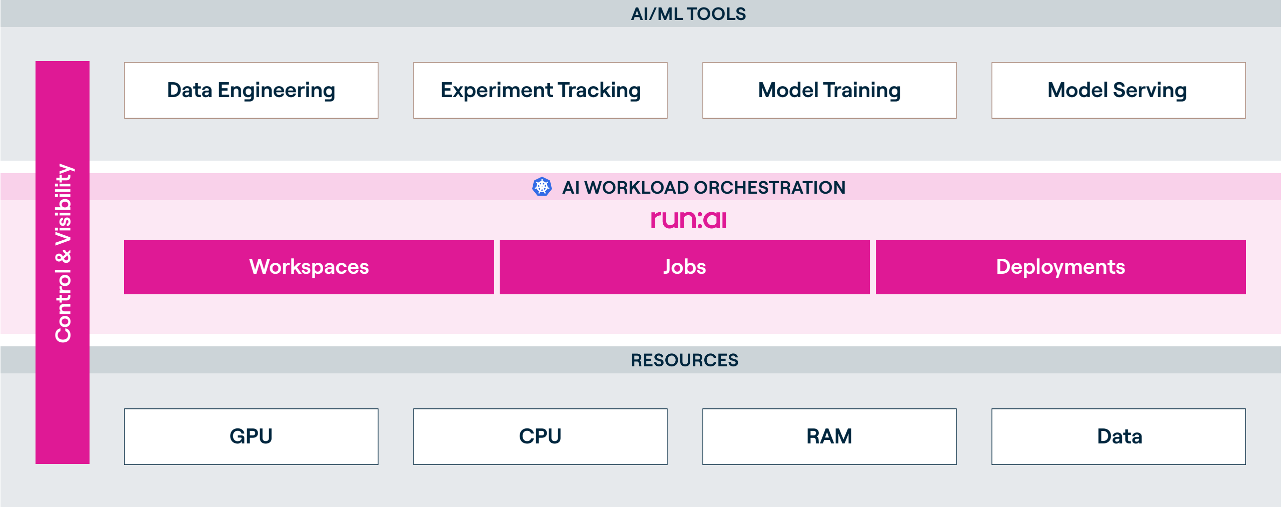
5X UTILIZATION



■ Total ▨ Total Allocated

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.



AI/ML TOOLS

Data Engineering | Experiment Tracking | Model Training | Model Serving

AI WORKLOAD ORCHESTRATION

run:ai

Workspaces | Jobs | Deployments

RESOURCES

GPU | CPU | RAM | Data

Control & Visibility

Feature Highlights

Gang Scheduling

Allows multi-node workloads to be launched together, start together, recover from failures together, and end together.

Batch Scheduling

Multiple queues and a sophisticated fairness algorithm automatically queue, preempt, restart and run workloads based on predefined policies, priorities and resource availability.

Dynamic Quotas

Guaranteed GPU quotas ensure resources are available to data scientists, projects or departments whenever needed. Combined with the ability to go over-quota and use unused or idle resources ensures optimal utilization and removes infrastructure complexity.

Customers Accelerating AI with Run:ai Atlas

SONY | BNY MELLON | ZEBRA | xiaomi