Centralizing AI Development on GPUs

The demand for AI continues to grow, and IT and DevOps teams are faced with the challenge of building and managing complex AI infrastructures. One of the biggest challenges is providing data scientists with access to GPUs for model training and development. Run:ai's AI Compute Orchestration Platform offers a solution for centralizing AI development on GPUs, providing a simplified, efficient, and scalable way to manage AI workloads.

Self-service Development Environments

Simplify the process of creating, managing, and sharing development environments by enabling data scientists to self-provision workspaces in a secured way with their preferred model development tools (e.g. Jupyter Notebooks, Weights & Biases and many more), the required compute resources and the data they need.

Dynamically share GPUs

Easily run multiple workloads, such as notebooks and workspaces, on a single GPU using isolated GPU fractions or NVIDIA MIG partitions. Allowing more users to get access to GPUs, run more experiments and optimize the utilization of expensive GPUs.

Automated Resource Management

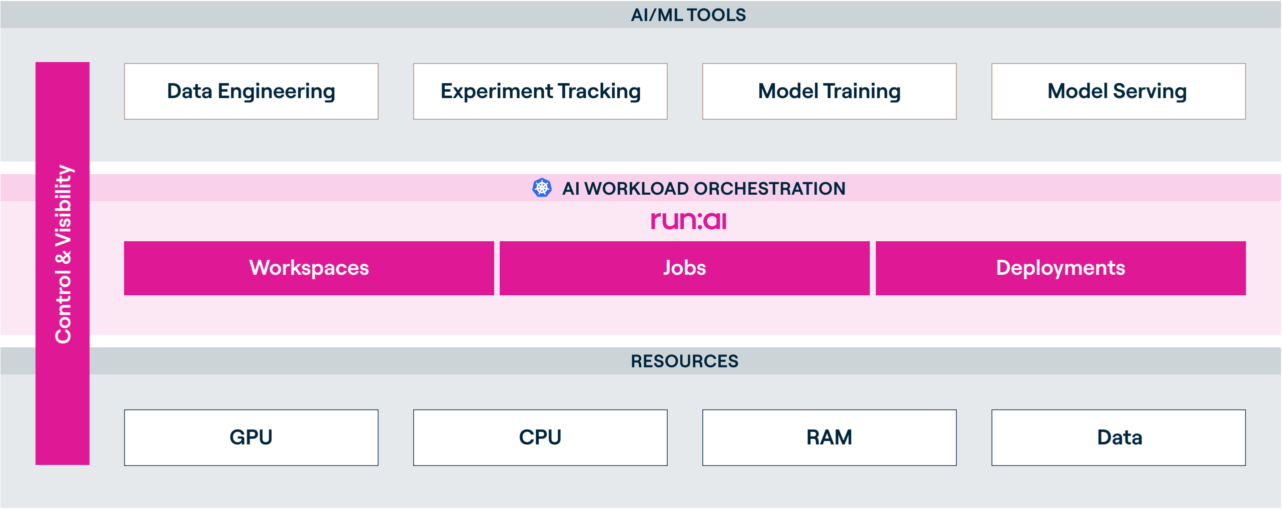
Ensure data scientists get access to the resources they need by using guaranteed quotas, removing users need to "hug" GPUs. Automated preemption of lower priority workloads ensures guaranteed quotas are met, and frees up resources when they are not being used.

25X FASTER

> *With Run:ai we've seen great improvements in speed of experimentation and GPU hardware utilization. Average experimentation time was reduced from 49 days to 2 days.*
>
> **Dr. M. Jorge Cardoso**
> Associate Professor & Senior Lecturer in AI at King's College London and CTO of the AI Centre

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

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.

AI Workload Scheduler

Run:ai's K8s Scheduler uses multiple queues to manage batch tasks, with customizable rules and policies for each queue based on business priorities. Combined with over-quota and fairness policies, resource allocation is automated and optimized for maximum cluster utilization.

Workspaces

Workspaces enable data scientists to self-provision the model development tools (like Jupyter Notebook, W&B, MLflow) together with the compute and data they need, in a simplified, streamlined and secured manner.

Customers Accelerating AI with Run:ai Atlas

SONY BNY MELLON ZEBRA xiaomi