Boost AI Compute Power to Speed Data Science Innovation

Run:ai Atlas - The Foundation for AI Clouds

Challenge - AI has an Execution Problem

Most AI research initiatives never make it to production. Why? Researchers consume compute power, typically using Graphics Processing Units (GPU), in order to build and train Deep Learning (DL) algorithms and bring AI initiatives to production. These GPU resources are allocated to researchers in a static way. This means that often, expensive compute resources allocated to one researcher sit idle - even as another researcher is waiting for GPUs. In fact, most AI teams within enterprises are using only 25% of their GPU infrastructure on average. Bringing models to production is painfully slow as static compute allocation limits progress. The inability to efficiently utilize resources slows experimentation, and is one of the primary reasons most enterprises don't see ROI trom AI.

Solution - Building an AI Foundation

Run:ai customers move models to production 10x faster than those without Run:ai. Using Run:ai's Atlas software platform, companies streamline the development, management and scaling of AI applications across any infrastructure (on-premises, edge, cloud). Researchers gain on-demand access to pooled GPU and CPU resources for any AI workload. An innovative, cloud-native operating-system helps IT manage everything from fractions of GPUs to large-scale distributed training. Greater efficiency yields faster modeling; one Run:ai customer recently executed 6,700 parallel hyperparameter tuning jobs and completed modeling in record time.

Benefits of the solution include:

Faster Time to Innovation
By using Run:ai's resource pooling, queueing, and prioritization mechanisms researchers are removed from infrastructure management hassles and can focus exclusively on data science. Run as many workloads as needed without compute bottlenecks.

Doing More with Less
Run:ai's fairness algorithms guarantee that all users and teams get their fair-share of resources. Policies around priority projects can be pre-set, and the platform allows dynamic allocation of resources from one user / team to another, ensuring that all users get timely access to coveted GPU resources.

Deploy AI Faster
Run:ai Atlas allows users to easily make use of fractional GPUs, integer GPUs, and multiplenodes of GPUs, for distributed training on Kubernetes. In this way, AI workloads run based on needs, not capacity. Data science teams will be able to run more AI experiments on the same infrastructure.

The Atlas Platform Consists of Four Layers

○ **INFRASTRUCTURE RESOURCES -** Orchestrate AI workloads across compute resources whether they are on-premises or in the cloud.

○ **OPERATING SYSTEM -** The OS consists of a GPU abstraction layer and a Kubernetes-based AI workload scheduler. Schedule and manage any AI workloads - build, train, inference - via our cloud-native operating system. Dynamic scheduling features allow for automated usage of fractions of GPU to multi-node distributed training.

○ **CONTROL PLANE-** Gain centralized visibility and control across multiple clusters no matter where they are located. See real-time and historical analytics showing users, jobs clusters and projects. Admins can set priorities and polices for users and teams across the infrastructure. SSO and advanced enterprise features are included.

○ **APPLICATION LAYER -** Develop and run your AI applications on accelerated infrastructure using Run:ai integrated tools or any data science tools of your choice. Run:ai integrates with Pytorch, Tensorflow and many DS tools, as well as Kubeflow, MLflow, Seldon, Weights & Biases and a host of other MLOps tools.

Automate Scheduling of AI Workloads on Kubernetes-based Architecture

Common practice today is to build deep learning infrastructure around containers and Kubernetes. Run:ai has built a smart scheduler for AI workloads directly into Kubernetes in order to simplify the learning curve for IT and data science teams and to improve infrastructure efficiency.

Manage AI Infrastructure Aligned with Business Goals

Run:ai Atlas manages tasks as batch processes using multiple queues on top of Kubernetes. The Control Plane allows system admins to define different rules, policies, and requirements for each queue based on business priorities. Combined with a quota-based system and configurable fairness policies, the allocation of resources can be automated by admins and optimized to allow maximum utilization of cluster resources.

Faster Experimentation - from 46 Days to Just Two Days with Run:ai

With better orchestration and management of compute resources, companies using Run:ai Atlas are seeing massive reductions in the time it takes to get AI models into production. One customer slashed the time taken to complete its experiments from 46 days (their previous average) to the current average which is now just a day and a half - an improvement of 3000%.