run:ai

SOLUTION BRIEF

Boost AI Compute Power to Speed Data Science Innovation

Run:AI – Orchestration and Acceleration for AI Workloads

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 from AI.*

SOLUTION - AI WORKLOADS GET A BOOST

Run:AI customers move models to production 10x faster than those without Run:AI. Run:AI has built an Orchestration Platform for AI Computing in order to help researchers consume GPUs more efficiently. We do this by automating the orchestration of AI workloads and the management and virtualization of hardware resources across teams and clusters. Run:AI pools compute and applies dynamic allocation mechanisms to boost resource availability at any given time. With pre-set scheduling and prioritization policies, researchers have access to as many GPUs as they need, and achieve model accuracy faster. *Greater efficiency yields faster modeling; one Run:AI customer recently executed 6,700 parallel hyperparameter tuning jobs and completed modeling in record time.*

Our AI Orchestration Platform for GPU-based computers running AI/ML workloads provides:

- **Fair scheduling** to allow users to easily and automatically share clusters of GPUs,
- **Distributed training** on multiple GPU nodes to accelerate model training times,
- **Fractional GPUs** to seamlessly run multiple workloads on a single GPU of any type,
- **Visibility** into workloads and resource utilization to improve user productivity.

Benefits of the solution include:

FASTER TIME TO INNOVATION

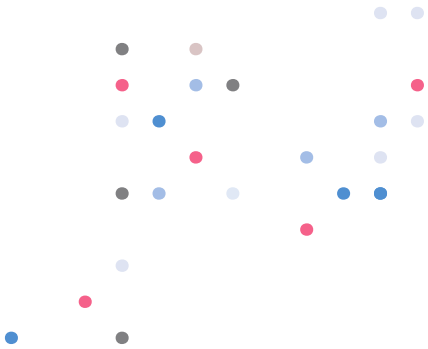
By using Run:AI's resource pooling, queueing, and prioritization mechanisms researchers are removed from infrastructure management hassles and can focus exclusively on data science. Run as many workloads as needed without compute bottlenecks.

INCREASED PRODUCTIVITY

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.

IMPROVED GPU UTILIZATION

The Run:AI Scheduler allows users to easily make use of fractional GPUs, integer GPUs, and multiple-nodes 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.



run:ai

ALL RIGHTS RESERVED RUNAI. NO PART OF THIS CONTENT MAY BE USED WITHOUT EXPRESS PERMISSION OF RUNAI

AUTOMATE SCHEDULING OF AI WORKLOADS KUBERNETES-BASED ARCHITECTURE

Common practice today is to build deep learning infrastructure around containers and Kubernetes. Run:AI has built a super-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 GPU INFRASTRUCTURE EFFICIENTLY BASED ON BUSINESS GOALS

The Run:AI Scheduler manages tasks as batch processes using multiple queues on top of Kubernetes, allowing 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 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%.

> *"With Run:AI we've seen great improvements in speed of experimentation and GPU hardware utilization. Reducing time to results ensures we can ask and answer more critical questions about people's health and lives."*
>
> M. Jorge Cardoso, Associate Professor & Senior Lecturer in AI London Medical Imaging & AI Centre for Value-Based Healthcare

Looking to implement Run:AI?
Please reach out to us for a demo

ABOUT RUN:AI

Run:AI has built the world's first Orchestration Platform for AI Computing. By abstracting workloads from underlying hardware, Run:AI creates a shared pool of GPU resources that can be dynamically provisioned, enabling efficient orchestration of AI workloads and optimized utilization of GPUs. Data scientists can seamlessly consume massive amounts of GPU power to improve and accelerate their research while IT teams retain centralized, cross-site control and real-time visibility over resource provisioning, queuing, and utilization. The Run:AI platform is built on top of Kubernetes, enabling simple integration with existing IT and data science workflows. Learn more at www.run.ai.

run:ai

ALL RIGHTS RESERVED RUNAI. NO PART OF THIS CONTENT MAY BE USED WITHOUT EXPRESS PERMISSION OF RUNAI