run:
ai

Virtualization Software for AI Infrastructure

Run:AI helps companies meet business objectives of their deep learning initiatives quickly and with more efficient use of resources.

Run:AI's software platform abstracts AI workloads from GPU compute power – creating 'virtual pools' where resources are automatically and dynamically allocated – so enterprises gain full GPU utilization. This helps IT and MLOps teams efficiently manage compute infrastructure at scale and within budget, while also enhancing the visibility of job scheduling and resource utilization for both IT and data science teams.

THE CHALLENGE

AI development is significantly different from traditional software development. It is based on experimentation and running an exceptionally large number of training models in parallel, which are typically highly compute-intensive, can run for hours, days, or even weeks, and require specialized and expensive processors such as GPUs. This makes established IT tools for tasks like capacity planning and quota management largely irrelevant. As a result, IT and MLOps teams often find themselves with few controls and limited visibility into compute resource allocation and utilization. In addition, data scientists may be limited to static GPU allocation or spend time waiting for specific GPU resources to become available for use, while other available resources across the organization stand idle.

OUR SOLUTION

Run:AI has built the world's first virtualization layer for deep learning workloads. By abstracting the underlying GPU infrastructure, Run:AI optimizes utilization of AI clusters by enabling flexible pooling and sharing of resources between users, teams, and projects. The software distributes workloads in an 'elastic' way – dynamically changing the number of resources allocated to a job – allowing data science teams to run more experiments on the same hardware. IT teams retain control by setting automated policies for resource allocation and gain real-time visibility - including run-time, queueing, and GPU utilization of each job. A centralized dashboard displays projects and jobs across multiple sites whether on premises or in the cloud. The Run:AI platform is built on top of Kubernetes, enabling simple integration with existing IT and data science workflows.

CONTROL AND VISIBILITY

Easy tools for control, policy setting, and tracking that provide a holistic view of GPU infrastructure utilization, usage patterns, workload wait times, and costs.

OPTIMIZED JOB SCHEDULING

Automated GPU cluster management, orchestration, and job queuing for efficient resource sharing and optimized utilization based on predefined policies.

BUILT ON KUBERNETES

Solution implemented as a Kubernetes plug-in, enabling simple integration with existing IT environments, open source tools, and data science platforms.

FLEXIBLE RESOURCE ALLOCATION

The software distributes workloads in an elastic way – dynamically changing the resources allocated to a job – to run more experiments on the same hardware.