CASE STUDY

Efficient compute orchestration yields more productive AI

How one defense organization created a private, managed GPU cloud for more than 100 researchers, and increased throughput for inference workloads by nearly 5X.

Run:ai's Kubernetes-based software platform for orchestration of containerized AI workloads enables GPU clusters to be utilized for different Deep Learning workloads dynamically from building AI models, to training, to inference. With Run:ai, jobs at any stage get access to the compute power they need, automatically.

Simplify management of shared GPU clusters with Run:ai

Users are typically allocated static, fixed numbers of GPU – 'two for me, two for you'. This one size fits all approach to scheduling and allocating GPU compute creates hassles for Ops and IT admins, and delays for researchers. In addition, a significant number of GPUs remain idle at any given time. The Run:ai Platform simplifies concepts from the world of High-Performance Computing and brings them to the world of cloud-native AI workloads.

Batch scheduling, advanced queuing mechanisms, and concepts like gang scheduling and topology awareness mean that each researcher is automatically allocated the amount of compute needed for his or her experiments. In addition, rules, policies, and requirements for each job are based on business priorities, and not first-come-first-served.

Case study: Defense industry

The customer use-case involved complex models being trained on huge data sets and real-time inferencing for a large defense organization. Inference workloads required maximum throughput and extremely low latency. With NVIDIA Triton Inference Server running multiple models on one GPU and with Run:ai coordinating the job scheduling on the inference server, maximum throughput and low latency were maintained while optimizing GPU utilization to nearly 100%. In addition, IT was able to manage pooled resources more efficiently, essentially creating a 'private GPU cloud' accessible on-demand to over 200 researchers and multiple teams.

Industry

Defense, Aerospace, Government

Challenges

- Complex to manage both training and inference on hundreds of on-premises NVIDIA DGX SystemsTM and NVIDIA GPUS (>200) for many researchers (>100) and AI teams.
- Resource management was non-existent.
- Inference workloads require low latency and maximum throughput.
- Fully air-gapped environment.

Solution

- Divide GPUs into logical pools, for build, train, and inference workloads.
- Apply advanced scheduling to manage how GPUs are allocated so every job gets the right amount of compute and memory
- Optimize orchestration of GPUs in the cluster for more than 80% utilization.

NVIDIA Products Used

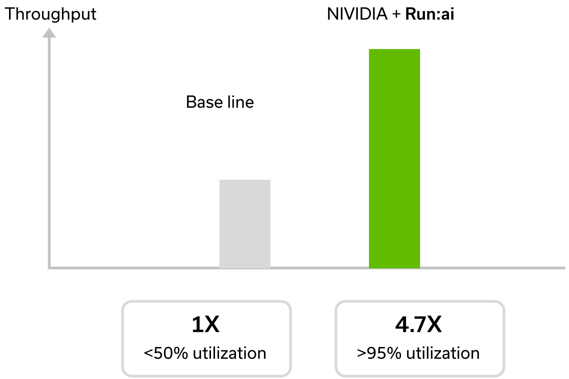
- NVIDIA GPU Operator, NVIDIA CUDA , and NVIDIA Triton Inference Server™
- NVIDIA T4 GPU servers, and NVIDIA DGX systems.

Results

By pooling all GPUs and applying pre-set priorities and policies, the customer was able to dynamically allocate resources to hundreds of users and multiple teams without management overhead. This resulted in greater availability of GPUs for training, speeding up research time. In addition, the example below shows nearly full utilization of the GPU cluster when NVIDIA Triton and Run:ai were used together with NVIDIA T4 GPUs running inference workloads. One can see an increase in throughput of 4.7x when using Run:ai and NVIDIA Triton together.

- Together with NVIDIA Triton, utilization of inference GPUs increased more than 4x to ~95%.
- Speed of model training increased while utilization of overall cluster was maximized.
- Customer successfully managed resource allocation of a shared pool of GPUs for the entire research team (>100 researchers) enabling on-demand access to all users and teams – essentially creating a private managed GPU cloud.
- Customer was able to see when additional server investments were necessary based on actual usage patterns, for better planning and ROI.

T4 Inference Test Results



About Run:ai

Run:ai has built the first cloud-native compute management platform for the AI era. The company's platform gives data scientists access to all the pooled compute power they need to accelerate AI development, whether on premises or in the cloud, while giving IT and MLOps visibility and control over scheduling and dynamic provisioning of GPUs. This delivers more than 2X gains in utilization of existing infrastructure.