run:ai

How One Company Went from 28% GPU Utilization to 73% With Run:AI

CASE STUDY

A multinational company, focused on innovations in computer vision technology, managed to go from 28% GPU utilization to optimization of over 70%, and achieve a 2X increase the speed of their training models with Run:AI. After expecting to make an additional GPU investment in 2020, with a planned hardware purchase cost of over $1 million dollars, the company can now maintain their current infrastructure, meet the needs of their data science teams, and even improve training times.

CUSTOMER

A world leader in facial recognition technologies, the company provides AI services to many large enterprises, often in real-time. Accuracy, measured in terms of maximizing performance of camera resolution and FPS, density of faces, and field of view are critically important to the company and their customers.

CUSTOMER AI INFRASTRUCTURE AND TEAM

- On-Premises environment with 24 Nvidia DGX servers and additional GPU workstations
- 30 Researchers spread on two continents

CHALLENGES

- **Sharing resources across teams and projects was unsuccessful.** GPU resources were statically allocated, creating times with bottlenecks and other times with inaccessible, but available infrastructure.
- **Prioritizing and scheduling deep learning training tasks was ineffective.** Researchers lacked the ability to see and manage available resources which was slowing down their jobs.
- **Expensive investment in scaling hardware led to increased costs.** Although the utilization of existing hardware was extremely low, visibility issues and bottlenecks made it seem like additional hardware was necessary.

AFTER IMPLEMENTING RUN:AI'S PLATFORM

70% Average GPU utilization → HIGHER ROI

2x Experiments /GPUs → BETTER DATA SCIENCE

Multi-GPU Training by default → FASTER TIME-TO-VALUE

Simplified Workflows → REDUCED DS HASSLES

- Data science teams simplified GPU utilization workflows and increased productivity by 2X, allowing them to more quickly deliver value with deep learning models
- Removed bottlenecks, resulting in faster training times - shortened by 75% on average
- Gained control and visibility into GPU clusters and saw utilization go from 28% to over 70% for better budgeting and planning of new hardware needs
- Achieved ability to scale deep learning so new researchers and jobs easily gain access to infrastructure.

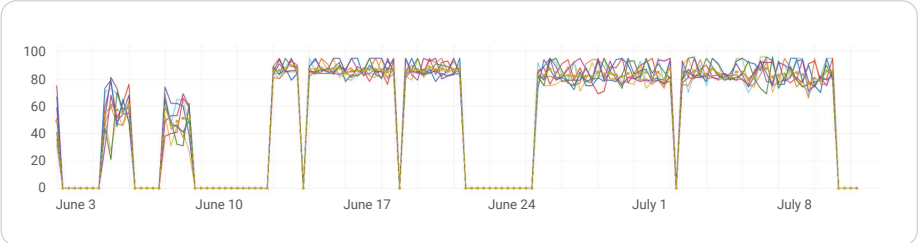
CHALLENGES

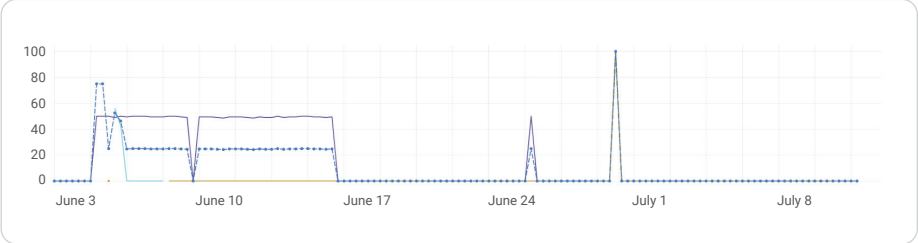
LOW GPU UTILIZATION

Some peaks, but mostly inefficient and underutilized resources

DIFFERENT USAGE PROFILES

'Build' 'Train' 'Retrain' with very different needs





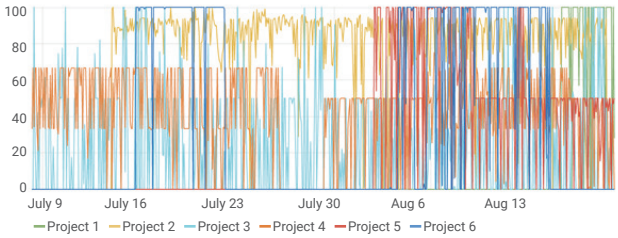
SOLUTION

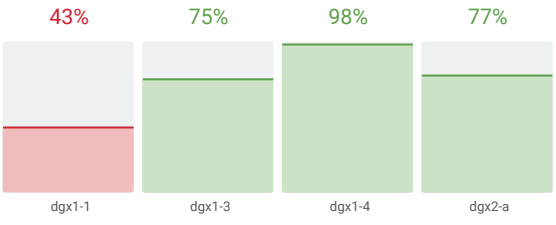
Run:AI's platform capabilities enabled the Company to achieve:

- **Increased GPU utilization** by moving teams from static, manual GPU allocations to pooled, dynamic resource sharing across the organization.
- **Increased productivity** for the data science teams using hardware abstraction, simplified workflows, and automated GPU resource allocations.
- **Visibility** into the GPU cluster, its utilization, usage patterns, wait times, etc., allowed the Company to better plan hardware spending.
- **Accelerated training times** using automated, dynamic allocation of resources which enabled the data science teams to complete training processes significantly faster.

INCREASED GPU UTILIZATION

Project and Node utilization is now visible to all teams.





ABOUT RUN:AI

Run:AI has built the world's first virtualization layer for deep learning training models. By abstracting workloads from underlying infrastructure, Run:AI creates a shared pool of resources that can be dynamically provisioned, enabling full utilization of GPU compute.

Data science and IT teams gain control and real-time visibility – including seeing and provisioning run-time, queueing, and GPU utilization of each job. A virtual pool of resources enables IT leaders and data scientists to view and allocate compute resources across multiple sites – whether on-premises or in the cloud. The Run:AI platform is built on top of Kubernetes, enabling simple integration with leading open source frameworks. Contact Run:AI at info@run.ai to learn more and see the platform in action.