London Medical Imaging & AI Centre Speeds Up Research with Run:AI

CASE STUDY

The London Medical Imaging & AI Centre for Value Based Healthcare, which works to develop, test and deploy pioneering AI systems across the NHS, has nearly doubled the number of experiments it runs on its AI hardware with Run:AI. The center has also doubled its average GPU utilization as it makes efficient use of its valuable hardware.

CUSTOMER

The London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare is a consortium of academic, healthcare and industry partners, led by King's College London and based at St Thomas' Hospital. It uses medical images and electronic healthcare data held by the UK National Health Service to train sophisticated deep learning algorithms for computer vision and natural-language processing. These algorithms are used to create new tools for effective screening, faster diagnosis and personalized therapies, to improve patients' health.

AI INFRASTRUCTURE AND TEAM

- Heterogeneous environment including DGX-1s and DGX-2s on premises
- Diverse research team with backgrounds in clinical research, data science and AI with varying usage patterns

AFTER IMPLEMENTING RUNAI'S PLATFORM

2.1x Higher GPU utilization → BETTER VALUE FOR MONEY

31x Faster experiments → INCREASED PACE OF DEVELOPMENT

Elastic workloads → LESS DELAYS AND BOTTLENECKS

1.85x More experiments → FASTER TIME TO INNOVATION

CHALLENGES

- **Low overall utilization of expensive AI hardware.** Total GPU utilization was below 30%, with significant idle periods for some GPUs despite demand from researchers.
- **Overloaded system with jobs requiring more resources.** The system was overloaded on multiple occasions where more GPUs were needed for running jobs than were available.
- **Poor visibility and scheduling led to delays and waste.** Bigger experiments requiring a set large number of GPUs were sometimes unable to begin, because smaller jobs using only a few GPUs were blocking them out of their resource requirements.

SOLUTION

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

Increased GPU utilization: GPU utilization rose *by 110%*, with resultant increases in experiment speed. Researchers ran more than 300 experiments in a 40-day period, compared to just 162 experiments that were run in a simulation of the same environment without Run:AI. By dynamically allocating pooled GPU to workloads, hardware resources were shared more efficiently.

Improved visibility: with advanced monitoring and cluster management tools, data scientists are able to see which GPU resources are not being used and dynamically adjust the size of their job to run on available capacity.

Fairer scheduling and guaranteed resource quotas, allowing large ongoing workloads to use the optimal amount of GPU during low-demand times, and automatically allowing shorter, higher-priority workloads to run alongside. In one instance, a single data scientist was able to submit more than 50 concurrent jobs, which were smoothly run as resources became available.

More completed experiments, allowing the AI Centre to iterate faster and develop critical diagnostic tools and therapeutic pathways that will save lives.

> ***"Our experiments can take days or minutes, using a trickle of computing power or a whole cluster. 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."***
>
> *Dr. M. Jorge Cardoso, Associate Professor & Senior Lecturer in AI at King's College London and CTO of the AI Centre*

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.