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

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 Run:ai's Platform

2.1X	Higher GPU Utilization	Better Value for Money
31X	Faster Experiments	Increased Pace of Developments
1.85X	More Experiments	Faster Time to innovation
Elastic Workloads		Fewer Delays and Bottlenecks

Challenges

○ **Low Overall Utilization of 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 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 AI Center 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.

○ **Fair Scheduling and Guaranteed Resources:** 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.

Make data science more productive with the compute power to run more experiments and deploy AI initiatives faster. Run:ai helps organizations accelerate their AI journey - from building initial models to scaling AI in production. Using Run:ai's Atlas software platform, companies streamline the development, management and scaling of AI applications across any infrastructure (on-premises, edge, cloud). Researchers gain on-demand access to pooled resources for any AI workload. An innovative, cloud-native operating system helps IT manage everything from fractions of GPUs to large-scale distributed training. Learn more at www.run.ai.