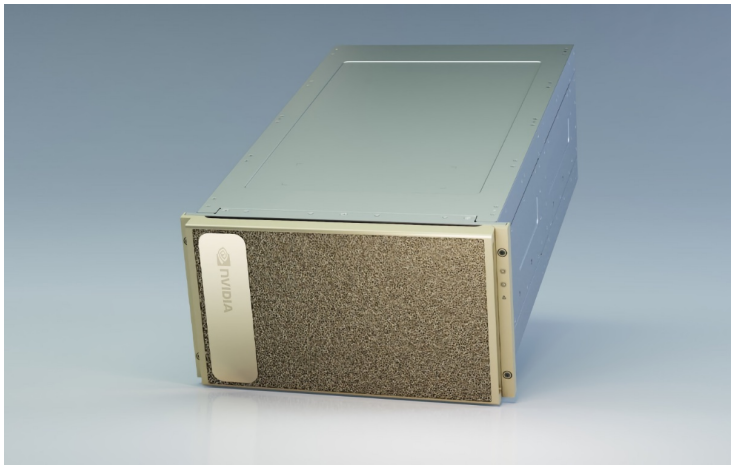
BioSpectrum
the business of Bio & Health Sciences
ASIA EDITION

NUHS Builds AI Production Platform Using NVIDIA DGX A100 For Better Healthcare Predictions

01 December 2021 | News

Singapore healthcare group uses AI to improve patient care and treatment



NVIDIA today announced that National University Health System (NUHS) has built an AI production platform based on an NVIDIA DGX A100 system to become the first healthcare group in Singapore to have real-time streaming capabilities to deliver better patient care and treatment, collaborate on biomedical research and transform how illnesses are managed and treated.

At the heart of NUHS' newly-launched Endeavour AI platform, NVIDIA DGX A100 will run AI tools that make real-time predictions on diagnosis, progression of diseases, readmissions, risk of falls, and others.

The new system will be integrated with NUHS' Discovery AI training platform to form a complete training and inference system as part of the group's digital transformation.

"There are many demands on healthcare these days and we are undertaking a digital transformation throughout the cluster. In the centre of our digital transformation is the use of AI. Advances in healthcare require great compute resources, and NVIDIA DGX A100 delivers easy access to performance needed to aid in a world class hospital," said Dr Ngiam Kee Yuan, group chief technology officer of NUHS and deputy chief medical informatics officer of National University Hospital (NUH).

NUHS is one of three public healthcare clusters in Singapore and an integrated academic health system and regional health system that delivers value-driven, innovative and sustainable healthcare. Its network covers 19 hospitals,

polyclinics, specialist centres, medical centre, and academic health science institutions.

CPU limitations

Discovery AI, which runs on NVIDIA GPUs, is used for training models using large data sets while CPUs are used for inferencing.

“As the number, volume and speed at which we are running inferencing increase, GPUs become necessary. Otherwise, we need to expend a lot more CPUs to run the same inferencing at that speed,” said Ngiam.

For example, one AI tool would run about 100 to 200 inferences per second. For every patient who turns up at its hospitals and polytechnics, every time a doctor clicks, saves or free texts, or when new lab test results are out, an AI tool will be running in the background. All the data gets processed by the AI tools. This is done hundreds of times per second throughout the whole cluster at a large volume. If only CPUs are used, NUHS will run out of processing speed very quickly.

“That's why we strategised, planned and built in NVIDIA DGX A100 from Day One when we deployed Endeavour AI. This is because we will be using it for high speed and large volume inference processed by our AI tools,” said Ngiam.

NVIDIA DGX A100 is the universal system for all AI infrastructure, from analytics to training to inference. It packs five petaFLOPS of AI performance into a 6U form factor, replacing legacy infrastructure silos with one platform for every AI workload.

Streaming data, real-time output

Endeavour AI is a software and hardware stack that features streaming data as well as AI tools running micro services off a Kubernetes backbone to process all the streaming data and produce outputs on a real-time basis.

With a capacity to handle up to 150 projects, Endeavour AI will start off with dozens of projects initially before scaling up.

Among the first projects are those that impact the whole cluster, ranging from predictions on how a patient with a certain condition will fare when admitted to a hospital to analysing magnetic resonance imaging (MRI) images. The projects will involve everything from structured medical data to text-based medical data that form the basis for generating chatbots that are conversational in nature.

NUHS produces between 20 and 30 GB of structured data and text daily, or between 1,800 and 2,500 messages per second for one hospital. This translates to about 10,000 to 15,000 messages per second at peak for the entire cluster. The AI tools need to run quickly in the background to absorb all the data on a day-to-day basis.

“We do not want to build one project at a time. We are building a platform that enables multiple projects to run at a time. We have multiple uses for GPUs, largely in training at this point, but certainly we are well underway in operationalising the production use cases. Without the GPU, we cannot do a lot of these things,” said Ngiam.

Operationalising AI, optimising healthcare

With Endeavour AI, an inferencing platform for streaming data powered by NVIDIA A100 GPUs on X86 server, NUHS has become the first healthcare group in Singapore to achieve stream capability by operationalising AI tools in real time for the entire cluster.

Patients interacting with AI-powered chatbots will experience improvement in appointment making, reduction in waiting time, and enjoy optimised care due to some of the work on patient trajectories.

Radiologists and clinicians benefit from improved accuracy and speed of processing of images, X rays and scans, thanks to the stream capability.

In day-to-day hospital operations, a number of predictions is done automatically without even needing to click a button. Data that streams out of the electronic medical record system are processed and the doctor is alerted if a patient meets a certain set of weights for high risk.

“These are tangible realities and outcomes that we expected when we Endeavour AI went live,” said Ngiam.

Looking towards for compute power

Even though the new NVIDIA DGX A100 has just gone live, NUHS is already looking forward to the next generation of the system to tackle the expected growth in datasets and speed needed to process those data in the next few years.

“We have invested in programmes that look at genomics. When this genomic data hits us, we are not talking about gigabytes but one terabyte of data per day. The amount of compute required to run genomic type processing going forward is going to be exponentially larger. Until then, our next step is to look at how to optimally use our GPUs for the next few years,” said Ngiam.

“NVIDIA DGX A100 lets NUHS consolidate training, inference and analytics into a unified AI infrastructure. It will provide the computing power to help the hospital group achieve operational and scientific breakthroughs in the healthcare sector, benefitting clinicians and patients in Singapore,” said Dennis Ang, director of enterprise business for the SEA and ANZ Region at NVIDIA.