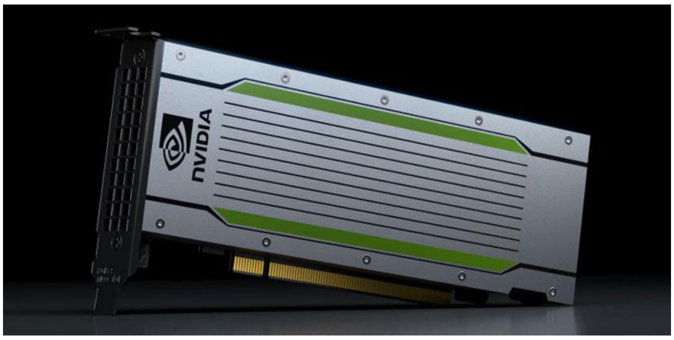
Nvidia Intros TensorRT Hyperscale Platform

Written by Marco Attard
14 September 2018

Nvidia uses GTC Japan 2018 to announce the TensorRT Hyperscale Platform-- a piece of datacentre hardware the company describes as the "most advanced inference accelerator for voice, video, image, recommendation services."



“Our customers are racing toward a future where every product and service will be touched and improved by AI,” the company says. “The Nvidia TensorRT Hyperscale Platform has been built to bring this to reality — faster and more efficiently than had been previously thought possible.”

The TensorRT Hyperscale Platform is built on Turing-based Tesla T4 GPUs and a comprehensive set of inference software, namely the TensorRT 5 inference optimiser and the TensorRT inference server. The Tesla T4 GPU carries 320 Turing tensor cores and 2560 CUDA cores, allowing for flexible multi-precision capabilities ranging from FP32 to FP16 to INT8, as well as INT4, all inside a 75W PCIe form factor. Performance clocks at 65 teraflops for FP64, 130 TOPS for INT8 and 260 TOPS for INT4.

The TensorRT 5 inference optimise and runtime engine supports the tensor cores and expands the neural network optimisations for multi-precise workloads, while the inference server enables applications to use AI models in datacentre production. The TensorRT inference server is available from the Nvidia GPU Cloud container registry, and maximises datacentre throughput and GPU utilisation, supports all popular AI models and frameworks, and integrates with Kubernetes and Docker.

Also seen at GTC announced is AGX, a high-performance computer series based on Xavier, the Nvidia processor built for autonomous machines. The company already offers the Jetson AGX Xavier dev kit for companies wanting to work on such vehicles. Another announcement is the Clara Platform, a combination of hardware and software bringing AI to next-gen medical imaging systems.

Go New Nvidia Datacentre Inference Platform to Fuel Next Wave of AI-Powered Services