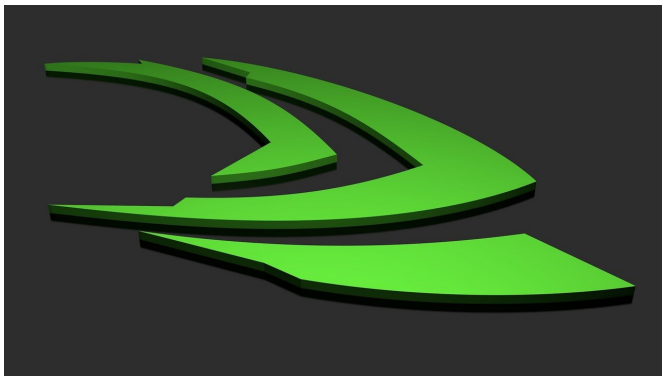
Nvidia brings Ampere A100 GPUs to Google Cloud

9 July 2020, by Peter Grad



Credit: Pixabay/CC0 Public Domain

Just over a month after announcing its latest generation Ampere A100 GPU, Nvidia said this week that the powerhouse processor system is now available on Google Cloud.

The A100 Accelerator Optimized VM A2 instance family is designed for enormous artificial intelligence workloads and data analytics. Nvidia says users can expect substantive improvements over previous processing models, in this instance up to a 20-fold performance boost. The system maxes out at 19.5 TFLOPS for single-precision performance and 156 TFLOPS for AI and high performance computing applications demanding TensorFloat 32 operations.

The Nvidia Ampere is the largest 7 nanometer chip ever constructed. It sports 54 billion transistors and offers innovative features such as multi-instance GPU, automatic mixed precision, an NVLink that doubles GPU-to-GPU direct bandwidth and faster memory reaching 1.6 terabytes per second.

The accelerator features 6,912 CUDA cores and has 40GB of HBM2 memory.

In describing the Ampere architecture, Invidia stated its improvements "provide unmatched acceleration at every scale."

"Scientists, researchers and engineers—the da Vincis and Einsteins of our time—are working to solve the world's most important scientific, industrial and big data challenges with AI and high-performance computing," Nvidia said. "Meanwhile businesses and even entire industries seek to harness the power of AI to extract new insights from massive data sets… NVIDIA Tensor Core technology has brought dramatic speedups to AI, bringing down training times from weeks to hours and providing massive acceleration to inference."

The new cloud service is currently in alpha mode. Service will be available in five configurations depending on business needs. The configurations range from one to 16 GPUs and 85GB Ram to 1,360GB RAM.

Google said businesses can tap into the Ampere A100 GPUs without much difficulty.

"We want to make it easy for you to start using the A2 VM shapes with A100 GPUs," Google said in an online statement. "You can get started quickly on Compute Engine with our Deep Learning VM images, which come preconfigured with everything you need to run high-performance workloads. In addition, A100 support will be coming shortly to Google Kubernetes Engine, Cloud AI Platform and other Google Cloud services.

Pricing has not yet been announced. Google announced that Cloud services will be available to the public after this year.

The quick availability of the new service to cloud operations is an indication of increasing demands of AI innovators exploring areas from facial detection to robotics to research into a vaccine for COVID-19.

NVIDIA says the A100 "has come to the cloud faster than any NVIDIA GPU in history." After the release of Nvidia's K80 GPU in 2014, it took two years before cloud access was made available. Following the 2016 release of its successor, the Pascal P4 GPU accelerator, the wait time for cloud availability was cut to one year. And it took only five months for the Volta V100 GPU accelerator to become available on the cloud.

This week's announcement of Ampere A100 cloud service comes about seven weeks after the latest GPU was unveiled.

More information:
techcrunch.com/2020/07/07/nvid … ome-to-google-cloud/

APA citation: Nvidia brings Ampere A100 GPUs to Google Cloud (2020, July 9) retrieved 1 December 2020 from https://techxplore.com/news/2020-07-nvidia-ampere-a100-gpus-google.html