NVIDIA announcements at GTC 2022

Silverton Consulting, Inc. StorInt™ Briefing

NVIDIA held their Spring GTC 2022 conference this month and announced many new products as well as updates to current solutions. The biggest news was a new GPU, the Hopper H100, a new engine for NVIDIA AI/HPC/advanced data analytics processing.

NVIDIA H100 GPU

NVIDIA's new GPU has anywhere from 2.5x to 30x the speed of the previous generation, A100 GPU, and it comes in 3 flavors

- **GH100 GPU**: 144 Streaming Multiprocessors (SMs), 18.4K CUDA cores, 576 4th gen TensorCores, 6 HBM3 or HBM2E stacks, 60MB L2 cache, 4th gen NVLINK and PCIe gen 5
- **H100 GPU SXM5**: 132 SMs, 16.9K CUDA cores, 528 4th gen TensorCores, 80GB with 5 HBM3 stacks, 50MB L2 cache, 4th gen NVLINK and PCIe gen 5
- **H100 PCIe**: 114 SMs, 14.6K CUDA cores, 456 4th gen TensorCores, 80GB with 5 HBM2E stacks, 50MB L2 cache, 4th gen NVLINK and PCIe gen 5

The GH100 and H100 SXM5 are built for datacenter workloads and don't support server/workstation graphics workloads. The H100 PCIe supports both datacenter and graphics workloads.

The H100 SXM5 has about **25% more** FP and integer arithmetic performance than the H100 PCIe GPU. For example, the H100 SXM5 can perform 30TFLOPS of Peak FP64 calculations while the H100 PCIe can only perform 24TFLOPS. This 25% performance advantage spans arithmetic processing from FP64 all the way down to FP8. There were no estimates supplied on GH100 performance in comparison to SXM5, but given the additional CUDA and TensorCores, we would expect at least 10% more FP performance for the GH100 over the H100 SXM5.

For comparison, the A100 had 7K CUDA cores. Moreover, with respect to the previous generation A100, each H100 GPU SM is:

- **Up to 6X faster** in chip-to-chip performance, this includes higher SM counts, faster SMs, and higher clock rate
- **Up to 2x faster** in Matrix Multiply Accumulate instruction performance,
- **Up to 4X faster** in Matrix Multiply Accumulate for FP8 on H100 vs. FP16 on the A100.

In addition, the H100 has **DPX instructions** for faster dynamic programing used in genomics, which is 7X faster than A100. It also has **3X faster IEEE FP64 and FP32** arithmetic over the A100, more (1.3X) shared memory, a new asynchronous execution engine, new **Tensor Memory Accelerator** functionality, and a new distributed shared memory with **direct SM to SM data transfers**.

The H100 is available in a new, server building block, the **HGX H100** with 4 or 8 GH100 GPUs in a single board configuration that includes NVLink and NVSwitch electronics for a build your own GPU server.

The H100 is also available in a **DGX H100 server** with 8 H100 GPUs in a rack mountable server that includes BlueField-3 (when available) DPU 400Gbps Ethernet/InfiniBand or OSFP 400Gbps Ethernet/InfiniBand networking. Up to 32 DGX H100 servers can be combined into a **DGX H100 Pod**. And NVIDIA is in the process of building out **EOS. a new SuperPod,** with 18 DGX H100 Pods in a single cluster.

Other news

BlueField-3 DPUs are scheduled to start shipping sometime later this year. Many storage solutions are starting to incorporate DPUs, including VAST Data (BlueField-1 DPU now, BlueField-3 when available), LANL new Advanced Box of Flash and others. **Spectrum-4**, a new InfiniBand switch was introduced included in DGX H100 SuperPods.

They also announced something like 60 SDKs have been updated, including DOCA3, CUDnn, and other NVIDIA APIs. Furthermore, many of their AI solutions have been updated including **RIVA 2.0** speech recognitions/text-to-speech, **Merlin 1.0** hyperscale recommendation engine, **Maxine** video conferencing, and the **Nem**o NLP model.

Omniverse was also discussed. Omniverse is NVIDIA's physics simulation/digital twin solution. NVIDIA also announced a new **OVX server** based on A40 GPUs and ethernet networking for Omniverse.

The other networking item discussed was the **ConnectX-7**, an intelligent NIC, that showed up in a converged **CNX H100 card** and in HGX H100 servers. The CNX H100 allows direct GPU to network data transfers without having to go through CPUs.

Significance

NVIDIA seems to be hitting on all cylinders. The only negative over the last year was the loss of the ARM acquisition. Barring that, NVIDIA is creating a whole new processing ecosystem based on their GPU technology for AI/HPC/advanced analytics.

The new Hopper H100 GPU is a nice evolution of their Ampere GPUs. It's almost like NVIDIA is introducing a new GPU each year and a new technology stack every 2 years. It's a big bet, but so far successful one. No doubt, crypto's GPU demand has helped underpin this investment.

AMD and Intel, the other GPU competition, aren't nearly as active pushing the technology. This may be acceptable for CPU vendors, as their main profit engine is CPUs, a competing processing technology. On the other hand, it does point to one advantage of having companies with a single-minded purpose and focus. For it seems these sorts of companies are the only ones that can re-invent whole industries. Stay tuned, it's going to be an interesting decade for IT.

Silverton Consulting, Inc., is a U.S.-based Storage, Strategy & Systems consulting firm offering products and services to the data storage community.