NVIDIA A100 GPU Overview

PowerEdge Support and Benchmark Performance

The A100 will be most impactful on PCIe Gen4 compatible PowerEdge servers, such as the PowerEdge R7525, which currently supports 2 A100s and will support up to 3 A100s within the first half of 2021. PowerEdge support for the A100 GPU will roll out on different Dell EMC next-gen server platforms over the course of H1 CY21.

Benchmarking data comparing performance on various workloads for the A100 and V100 are shown below:

**Inference**

Figure 2 displays the performance improvement of the A100 over the V100 for two different inference benchmarks – BERT and ResNet-50. The A100 performed 2.5x faster than the V100 on the BERT inference benchmark, and 5x faster on the RN50 inference benchmark. This will translate to significant time reductions spent on inferring trained neural networks to classify and identify known patterns and objects.

**HPC**

Figure 3 displays the performance improvement of the A100 over the V100 for four different HPC benchmarks. The A100 performed between 1.4x – 1.9x faster than the V100 for these benchmarks. Users looking to process data and perform complex HPC calculations will benefit from reduced completion times when using the A100 GPU.
Training

Figure 4 displays the performance improvement of the A100 over the V100 for two different training benchmarks – BERT Training TF32 and BERT Training FP16. The A100 performed 5x faster than the V100 on the BERT TF32 benchmark, and 2.5x faster on the BERT FP16 benchmark. Users looking to swiftly train their neural networks will greatly benefit from the A100 GPUs improved specs, as well as new features (such as TF32), which are further discussed below.

A100 Specifications

At the heart of NVIDIA’s A100 GPU is the NVIDIA Ampere architecture, which introduces double-precision tensor cores allowing for more than 2x the throughput of the V100 – a significant reduction in simulation run times. The double-precision FP64 performance is 9.7 TFLOPS, and with tensor cores this doubles to 19.5 TFLOPS. The single-precision FP32 performance is 19.5 TFLOPS and with the new Tensor Float (TF) precision this number significantly increases to 156 TFLOPS; ~20x higher than the previous generation V100. TF32 works as a hybrid of FP16 and FP32 math models that uses the same 10-bit precision mantissa as FP16, and 8-bit exponent of FP32, allowing for speedup increases on specific benchmarks.

Furthermore, the A100 supports a massive 40GB of high-bandwidth memory (HBM2) that enables up to 1.6TB/s memory bandwidth. This is 1.7x higher memory bandwidth over the previous generation V100 (see Figure 5).

New Features

In addition to the product specification improvements noted above, the NVIDIA A100 introduces 3 key new features that will further accelerate High-Performance Computing (HPC), Training and Artificial Intelligence (AI) Inference workloads:

1. **3rd Generation NVIDIA NVLink™** – The new generation of NVLink™ has 2x GPU-to-GPU throughput over the previous generation V100.

2. **Multi-Instance GPU (MIG)** – This feature enables a single A100 GPU to be partitioned into as many as seven separate GPUs, which benefits cloud users looking to utilize their GPUs for AI inference and data analytics workloads.

3. **Structural Sparsity** – This feature supports sparse matrix operations in tensor cores and increases the throughput of tensor core operations by 2x (see Figure 6).
User Implementation

It is important to know how the A100 can accelerate varying HPC, Training and Inference workloads:

**HPC Workloads** – Scientific computing simulations are typically so complex that FP64 double-precision models are required to translate the mathematics into accurate numeric models. At nearly 20 TFLOPs of double-precision performance, simulation run times are reduced by half with A100 double-precision tensor cores, allowing for 2x the normal FP64 output.

**Training Workloads** – Learning applications, such as recognition and training, typically require FP32 single-precision models to extract high level features from raw input. This means that the Tensor Float (TF32) computational model is an excellent alternative to FP32 for these types of applications. Running TF32 will grant up to 20x greater performance than the V100, allowing for significant train time reductions. Applications that need higher performance offered by a single server can do so by leveraging efficient scale-out techniques using low latency and high-bandwidth networking supported on the R7525. Additionally, specific training applications will also benefit from an additional 2x in performance with the new sparsity feature enabled.

**Inference Workloads** – Inference workloads will greatly benefit from the full range of precision models available, including FP32, FP16, INT8 and INT4. The Multi-Instance GPU (MIG) feature allows multiple networks to operate simultaneously on a single GPU so server users can have optimal utilization of compute resources. Structural sparsity support is also ideal for inference and data analytics applications, delivering up to 2x more performance on top of A100’s other inference performance gains.

**Conclusion**

The NVIDIA A100 GPU offers performance improvements and new feature sets that were designed to accelerate HPC, Training and AI Inference workloads. A server configured with A100 GPUs will be enabled to utilize these capabilities working in concert with other system components to yield best performance.