New NVIDIA V100 32GB GPUs, Initial performance results

Deepthi Cherlopalle, HPC and AI Innovation Lab. June 2018

GPUs are useful for accelerating large matrix operations, analytics, deep learning workloads and several other use cases. NVIDIA introduced the Pascal line of their Tesla GPUs in 2016, the Volta line of GPUs in 2017, and recently announced their latest Tesla GPU based on the Volta architecture with 32GB of GPU memory. The V100 GPU is available with both PCIe and NVLink version, allowing GPU-to-GPU communication over PCIe or over NVLink. The NVLink version of the GPU is also called an SXM2 module.

This blog will give an introduction to the new Volta V100-32GB GPUs and compare the HPL performance between different V100 models. Tests were performed using a Dell EMC PowerEdge C4140 with both PCIe and SXM2 configurations. There are several other platforms which support GPUs: PowerEdge R740, PowerEdge R740XD, PowerEdge R840, and PowerEdge R940xa. A similar study was conducted in the past comparing the performance of the P100 and V100 GPUs with the HPL, HPCG, AMBER, and LAMMPS applications.

Table 1 below provides an overview of Volta device specifications.

Table 1: *GPU Specifications*

	Tesla V100-PCIe	Tesla V100-SXM2
GPU Architecture	Volta	
NVIDIA Tensor cores	640	
NVIDIA CUDA Cores	5140	
GPU Max Clock Rate	1380MHz	1530MHz
Double precision performance	7TFlops	7.8TFlops
Single precision performance	14TFlops	15.7TFlops
GPU memory	16/32GB	16/32GB
Interconnect Bandwidth	32GB/s	300GB/s
System Interface	PCIe Gen3	NVIDIA NVLink
Max Power Consumption	250 watts	300 watts

The *PowerEdge C4140* Server is an accelerator optimized server with support for two Intel Xeon Scalable processors and four NVIDIA Tesla GPUs (PCIe or NVLink) in a 1U form factor. The PCIe version of the GPUs is supported with standard PCIe Gen3 connections between GPU to CPU. The NVLink configuration allows GPU-to-GPU communication over the NVLink interconnect. Applications that can take advantage of the higher NVLink bandwidth and the higher clock rate of the V100-SXM2 module can benefit from this option. The PowerEdge C4140 platform is available in four different Configurations: B, C, K, and G. The configurations are distinct in their PCIe lane layout and NVLink capability and are shown in **Figure 1** through **Figure 3**4.

In Configuration B, the GPU to GPU communication is through a PCIe switch, and the PCIe switch is connected to a single CPU. In Configuration C and G two GPUs are connected to each CPU, however in Configuration C the two GPUs are directly connected to each CPU, where as in Configuration G the GPUs are connected to the CPU via a PCIe switch. The PCIe Switch in Configuration G is logically divided into two virtual switches mapping 2GPUs to each CPU. In Configuration K, GPU-to-GPU communication is over NVLink, with all GPUs connected to a single CPU. As seen in the figures below all the configurations have additional x16 slots available apart from the GPU slots.

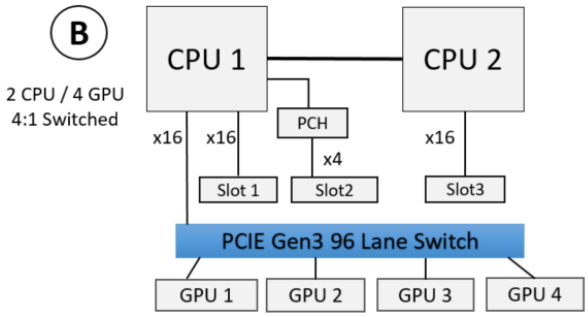

Figure 1: *PowerEdge C4140 Configuration B*

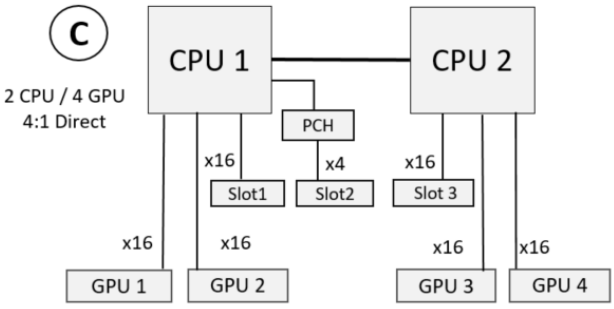

Figure 2: *PowerEdge C4140 Configuration C*

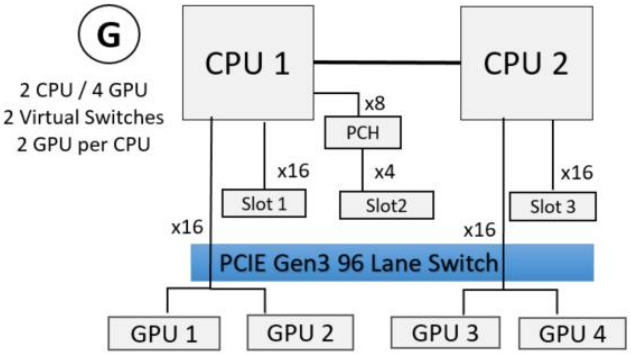

Figure 3: *PowerEdge C4140 Configuration G*

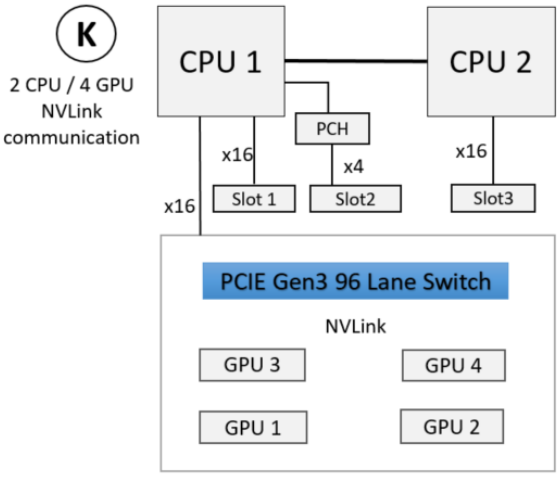

Figure 4*: PowerEdge C4140 Configuration K*

The PowerEdge C4140 platform can support a variety of Intel Xeon CPU models, up to 1.5 TB of memory with 24 DIMM slots, multiple network adapters and provides several local storage options. For more information on this server click here.

To evaluate the performance difference between the V100-16GB and the V100-32GB GPUs, a series of tests were conducted. These tests were run on a single PowerEdge C4140 server with the configurations detailed below in **Table 2**, **Table 3** and **Table 4**.

Table 2: *Tested Configurations Details*

Tests	Configuration	GPU
Test 1	B	4*V100-16GB-PCIe
Test 2	B	4*V100-32GB-PCIe
Test 3	K	4*V100-16GB-SXM2
Test 4	K	4*V100-32GB-SXM2

Table 3: *Hardware Configuration*

Server	Dell EMC Power Edge C4140
Processor	Intel Xeon Gold 6148. 20 cores, 2.40 GHz
Memory	384 GB @ 2667 MTps
GPU	NVIDIA V100-16GB PCIe,V100-32GB, V100-16GB-SXM2, V100-32GB-SXM2
Storage	M.2 120GB SSD
Power Supplies	Dual 2000W

Table 4: *Software/Firmware Configuration:*

Component	Version
BIOS	1.1.7
OS	Red Hat Enterprise Linux 7.4
Kernel	3.10.0-693.17.1.el7.x86_64
System Profile	Performance optimized (Turbo enabled, C-States disabled, Power management set to Max Performance)
CUDA driver	390.46
CUDA toolkit	9.1
Compilers	gcc- 4.8.5 , OpenMPI – 1.10.2
Intel MKL	From Intel Parallel Studio 2017

HPL performance

High Performance Linpack (HPL) is a standard HPC benchmark used to measure computing power. It is also used as a reference benchmark by the Top500 list to rank supercomputers worldwide. This benchmark provides a measurement of the peak computational performance of the entire system. There are few parameters that are significant in this benchmark:

- *N* is the problem size
- *NB* is the block size

- *Rpeak* is the theoretical peak of the system.
- *Rmax* is the maximum measured performance achieved on the system.
- The *efficiency* is defined as the ratio of Rmax to Rpeak.

The resultant performance of HPL is reported in GFLOPS.

N is the problem size provided as input to the benchmark and determines the size of the dense linear matrix that is solved by HPL. HPL performance tends to increase with increasing N value (problem size) until limits of system memory, CPU or data communication bandwidth begins to limit the performance. For GPU system, the highest HPL performance will commonly occur when the problem size is close to the size of the GPUs memory and the performance will be higher when a larger problem size will fit in that memory.

In this section of the blog, the HPL performance of the NVIDIA V100-16GB and the V100-32GB GPUs is compared using PowerEdge C4140 configuration B and K (*refer to Table 2*). Recall that configuration B uses PCIe V100s with 250W power limit and configuration K uses SXM2 V100s with higher clocks and 300W power limit. Figure 5 shows the maximum performance that can be achieved on different configurations. We measured a 14% improvement when running HPL on V100-32GB with PCIe versus V100-16GB with PCIe, and there was a 16% improvement between V100-16GB SXM2 and V100-32GB SXM2. The size of the GPU memory made a big difference in terms of performance as the larger memory GPU can accommodate a larger problem size, a larger N.

As seen in ***Table 1*** the V100-16GB, V100-32GB PCIe and V100-16GB, V100-32GB SXM2 have the same number of cores, double precision performance and GPU Bandwidth except for the GPU memory. We also measured ~6% HPL performance improvement from PCIe to SXM2 GPUs which is a small delta in HPL performance but Deep learning frameworks like Tensor Flow and Caffe show much more performance improvement.

Running HPL using only CPUs yields ~2.3TFLOPS with the Xeon Gold 6148; therefore, one PowerEdge C4140 system with four GPUs provides floating point capabilities equal to about nine two socket Intel Xeon 6148 servers.

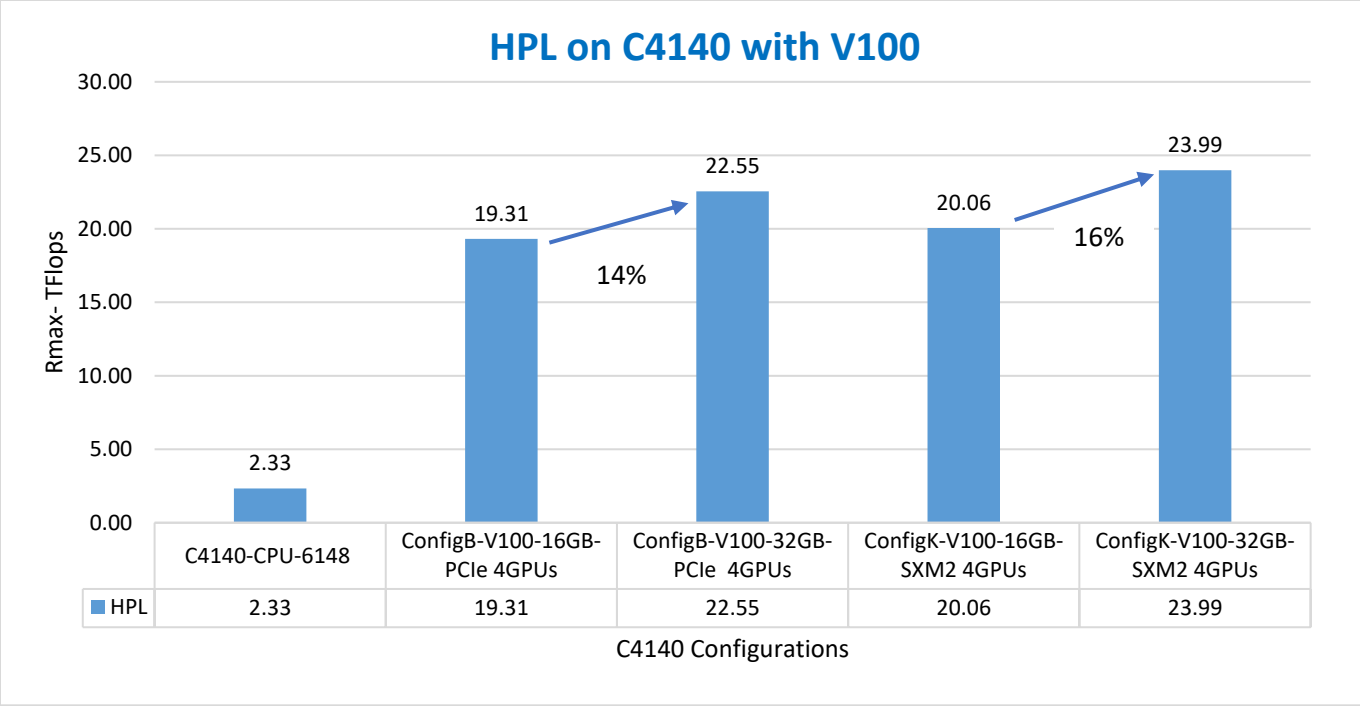

***Figure 5**: HPL Performance on different C4140 configurations.*

Figure 6 and Figure 7 shows the performance of V100 16GB vs 32GB GPU for different values of N. **Table 2** shows the configurations used for this test. These graphs helps us visualize how the GPU cards perform with different problem sizes. As explained above, the problem size is calculated based on the size of the GPU memory, the 32GB GPU can accommodate a larger problem size than the 16GB GPU. When a problem size that is larger than what will fit in GPU memory is executed on a GPU system, the system memory attached to the CPU is used, and this leads to a drop in performance as the data must move from system memory to GPU memory. For ease of understanding the test data is split into two different graphs.

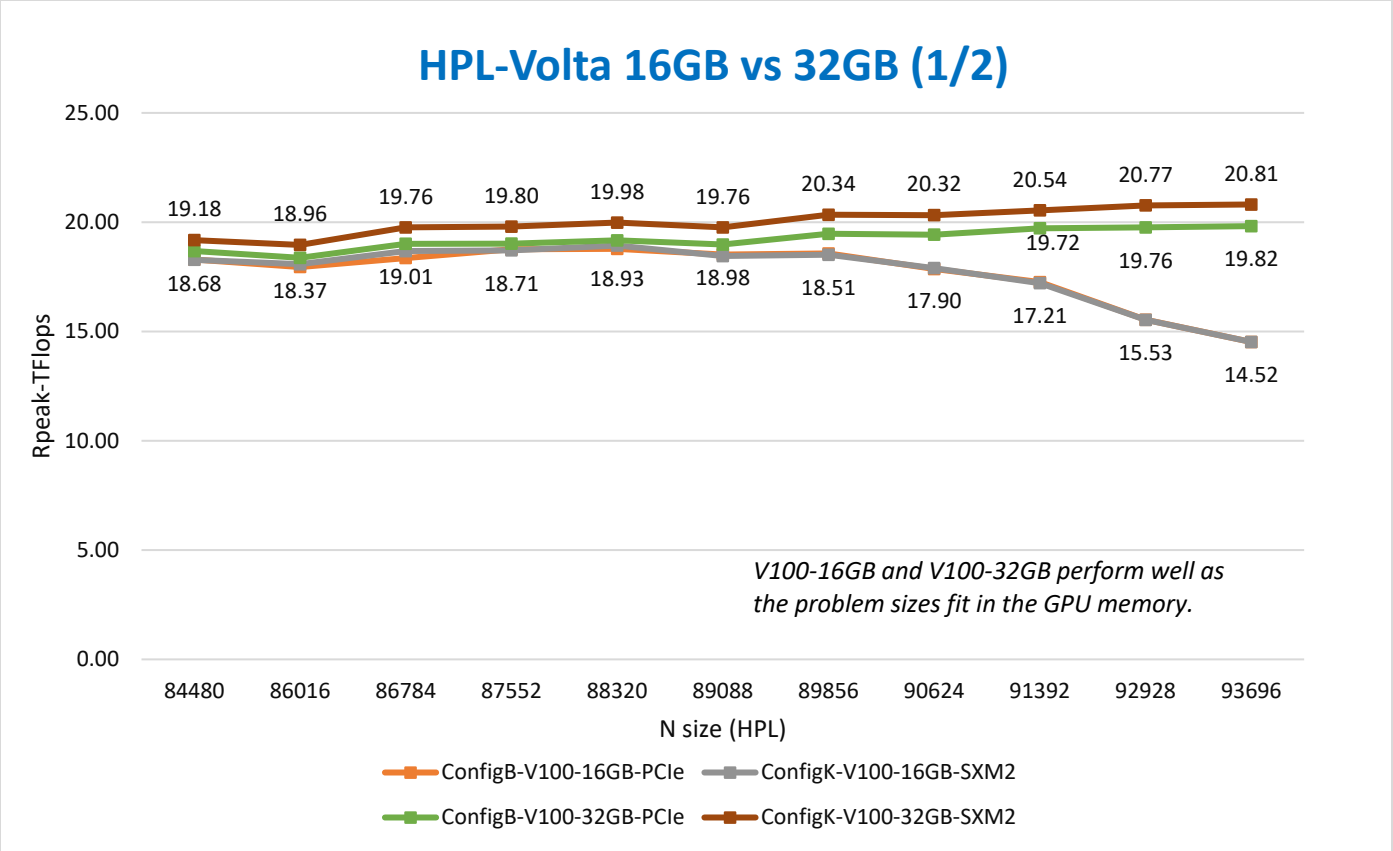

Figure 6: HPL performance with different problem sizes (N)

In Figure 6 we notice that the HPL performance for both the cards is similar until the problem size (N) approximately fills up V100-16GB memory, the same problem size (N) would approximately fill up half the memory for V100-32GB GPUs. In the second graph in Figure 7 we notice that the performance of the V100 16GB GPU drops as it cannot fit larger problem sizes in the GPU memory and must start to use system host memory. The 32GB GPU continues to perform better with larger and larger N until the problem size reaches the maximum capacity of the V100 32GB memory.

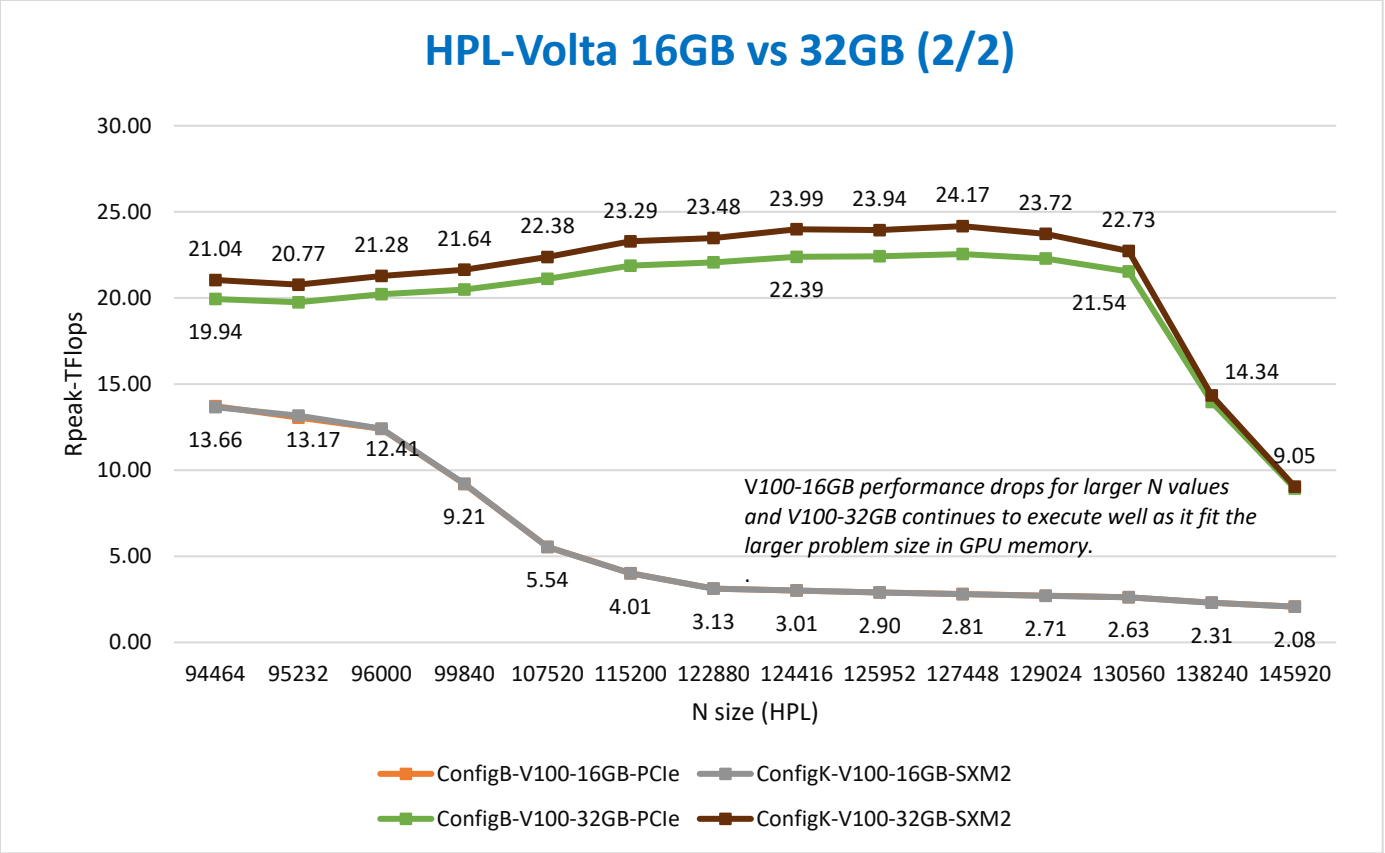

Figure 7: HPL performance with different problem sizes (N)

Conclusion and Future work:

PowerEdge C4140 is one of the most prominent GPU based server options for HPC related solutions. We measured a 14-17% improvement in HPL performance when moving from the smaller memory V100-16GB GPU to the larger memory V100-32GB GPU. For memory bound applications, the new Volta 32GB cards would be the preferred option.

For future work, we will run molecular dynamic applications, deep learning workloads and compare the performance between different Volta cards and C4140 configurations.

Please contact HPC innovation lab if you'd like to evaluate the performance of your application on PowerEdge Servers.