HPC Applications Performance on V100

Authors: Frank Han, Rengan Xu, Nishanth Dandapanthula.

HPC Innovation Lab. August 2017

Overview

This is one of two articles in our Tesla V100 blog series. In this blog, we present the initial benchmark results of NVIDIA® Tesla® Volta-based V100™ GPUs on 4 different HPC benchmarks, as well as a comparative analysis against previous generation Tesla P100 GPUs. We are releasing another V100 series blog, which discusses our V100 and deep learning applications. If you haven't read it yet, it is highly recommend to take a look here.

PowerEdge C4130 with V100 GPU support

The NVIDIA® Tesla® V100 accelerator is one of the most advanced accelerators available in the market right now and was launched within one year of the P100 release. In fact, Dell EMC is the first in the industry to integrate Tesla V100 and bring it to market. As was the case with the P100, V100 supports two form factors: V100-PCIe and the mezzanine version V100-SXM2. The Dell EMC PowerEdge C4130 server supports both types of V100 and P100 GPU cards. Table 1 below notes the major enhancements in V100 over P100:

Table 1: The comparison between V100 and P100

	PCIe			SXM2		
	P100	**V100**	**Improvement**	**P100**	**V100**	**Improvement**
Architecture	Pascal	Volta		Pascal	Volta	
CUDA Cores	3584	5120		3584	5120	
GPU Max Clock rate (MHz)	1329	1380		1481	1530	
Memory Clock rate (MHz)	715	877	23%	715	877	23%
Tensor Cores	N/A	640		N/A	640	
Tensor Cores/SM	N/A	8		N/A	8	
Memory Bandwidth (GB/s)	732	900	23%	732	900	23%
Interconnect Bandwidth Bi-Directional (GB/s)	32	32		160	300	
Deep Learning (TFlops)	18.6	112	6x	21.2	125	6x
Single Precision (TFlops)	9.3	14	1.5x	10.6	15.7	1.5x
Double Precision (TFlops)	4.7	7	1.5x	5.3	7.8	1.5x
TDP (Watt)	250			300		

V100 not only significantly improves performance and scalability as will be shown below, but also comes with new features. Below are some highlighted features important for HPC Applications:

- Second-Generation NVIDIA NVLink™

All four V100-SXM2 GPUs in the C4130 are connected by NVLink™ and each GPU has six links. The bi-directional bandwidth of each link is 50 GB/s, so the bi-directional bandwidth between different GPUs is 300 GB/s. This is useful for applications requiring a lot of peer-to-peer data transfers between GPUs.

- New Streaming Multiprocessor (SM)

Single precision and double precision capability of the new SM is 50% more than the previous P100 for both PCIe and SXM2 form factors. The TDP (Thermal Design Power) of both cards are the same, which means V100 is ~1.5 times more energy efficient than the previous P100.

- HBM2 Memory: Faster, Higher Efficiency

The 900 GB/sec peak memory bandwidth delivered by V100, is 23% higher than P100. Also the DRAM utilization has been improved from 76% to 95%, which allows for a 1.5x improvement in delivered memory bandwidth.

More in-depth details of all new features of V100 GPU card can be found at this Nvidia website.

Hardware and software specification update

All the performance results in this blog were measured on a PowerEdge Server C4130 using Configuration G (4x PCIe V100) and Configuration K (4x V100-SXM2). Both these configurations have been used previously in P100 testing. Also except for the GPU, the hardware components remain identical to those used in the P100 tests as well: dual Intel Xeon E5-2690 v4 processors, 256GB (16GB*16 2400 MHz) Memory and an NFS file system mounted via IPoIB on InfiniBand EDR were used. Complete specs details are included in our previous blog. Moreover, if you are interested in other C4130 configurations besides G and K, you can find them in our K80 blog.

There are some changes on the software front. In order to unleash the power of the V100, it was necessary to use the latest version of all software components. Table 2 lists the versions used for this set of performance tests. To keep the comparison fair, we re-ran the P100 tests using the new software stack to normalize for the upgraded software.

Table 2: The changes in software versions

Software	**Current Version**	**Previous version in P100 blog**
OS	RHEL 7.3	RHEL 7.2
GPU Driver	384.59	361.77/375.20
CUDA Toolkit	9.0.103RC	8.0.44
OpenMPI	1.10.7 & 2.1.2	1.10.1 & 2.0.1
HPL	Compiled with sm7.0	Compiled with sm6.0
HPCG	Compiled with sm7.0	-
AMBER	16, AmberTools17 update 20	16 AmberTools16 update3
LAMMPS	patch_17Aug2017	30Sep16

p2pBandwidthLatencyTest

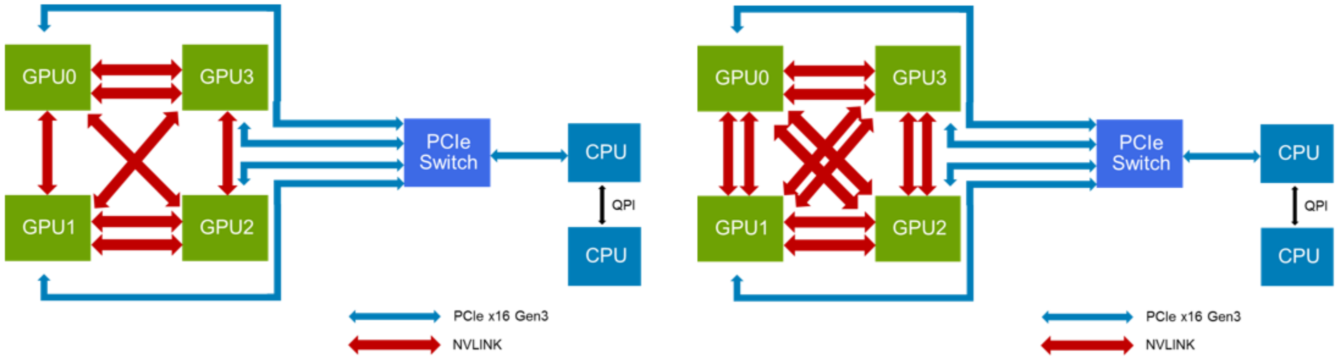
p2pBandwidthLatencyTest is a micro-benchmark included in the CUDA SDK. It tests the card to card bandwidth and latency with and without GPUDirect™ Peer-to-Peer enabled. Since the full output matrix is fairly long, we use the unidirectional P2P result as an example here to demonstrate the way to verify the NVLINKs speed on both V100 and P100.

In theory, V100 has 6x 25GB/s uni-directional links, giving 150GB/s throughput. The previous P100-SXM2 only has 4x 20GB/s links, delivering 80GB/s. The results of p2pBandwitdhtLatencyTest on both cards are in Table 3. “D/D” represents “device-to-device”, that is the bandwidth available between two devices (GPUs). The achievable bandwidth of GPU0 was calculated by aggregating the second, third and fourth value in the first line, which represent the throughput from GPU0 to GPU1, GPU2 and GPU3 respectively.

Table 3: Unidirectional peer-to-peer bandwidth

Unidirectional P2P=Enabled Bandwidth Matrix (GB/s). Four GPUs cards in the server.									
P100					V100				
D\D	0	1	2	3	D\D	0	1	2	3
0	231.53	18.36	18.31	36.39	0	727.38	47.88	47.9	47.93
1	18.31	296.74	36.54	18.33	1	47.92	725.61	47.88	47.89
2	18.35	36.08	351.51	18.36	2	47.91	47.91	726.41	47.95
3	36.59	18.42	18.42	354.79	3	47.96	47.89	47.9	725.02

It is clearly seen that V100-SXM2 on C4130 configuration K is significant faster than P100-SXM2, on:



C4130 ConfigK with P100

C4130 ConfigK with V100

1. **Achievable throughput**. V100-SXM2 has 47.88+47.9+47.93= 143.71 GB/s aggregated achievable throughput, which is 95.8% of the theoretical value 150GB/s and significant higher than 73.06GB/s and 91.3% on P100-SXM2. The bandwidth for bidirectional traffic is twice that of unidirectional traffic and is also very close to the theoretically 300 GB/s throughput.

2. **Real world application**. Symmetric access is the key for real world applications, on each chipset, P100 has 4 links, out of which three are connected to each of the other three GPUS. The remaining fourth link is connected to one of the other three GPUs. So, there are two links between GPU0 and GPU3, but only 1 link between GPU0 and GPU1 as well as GPU0 and GPU2. This is not symmetrical. The above numbers of p2pBandwidthLatencyTest in blue show this imbalance, as the value between GPU0 to GPU3 reaches 36.39 GB/s, which is double the bandwidth between GPU0 and GPU1 or GPU0 and GPU2. In most real world applications, it is common for the developer to treat all cards equally and not take such architectural differences into account. Therefore it will be likely that the faster pair of GPUs will need to wait for the slowest transfers, which means that 18.31 GB/s is the actual speed between all pairs of GPUs.

   On the other hand, V100 has a symmetrical design with 6 links as seen in Figure 1. GPU0 to GPU1, GPU2, or GPU3 all have 2 links between each pair. So 47.88 GB/s is the achievable link bandwidth for each, which is 2.6 times faster than the P100.

Figure1: V100 and P100 Topologies on C4130 configuration K

High Performance Linpack (HPL)

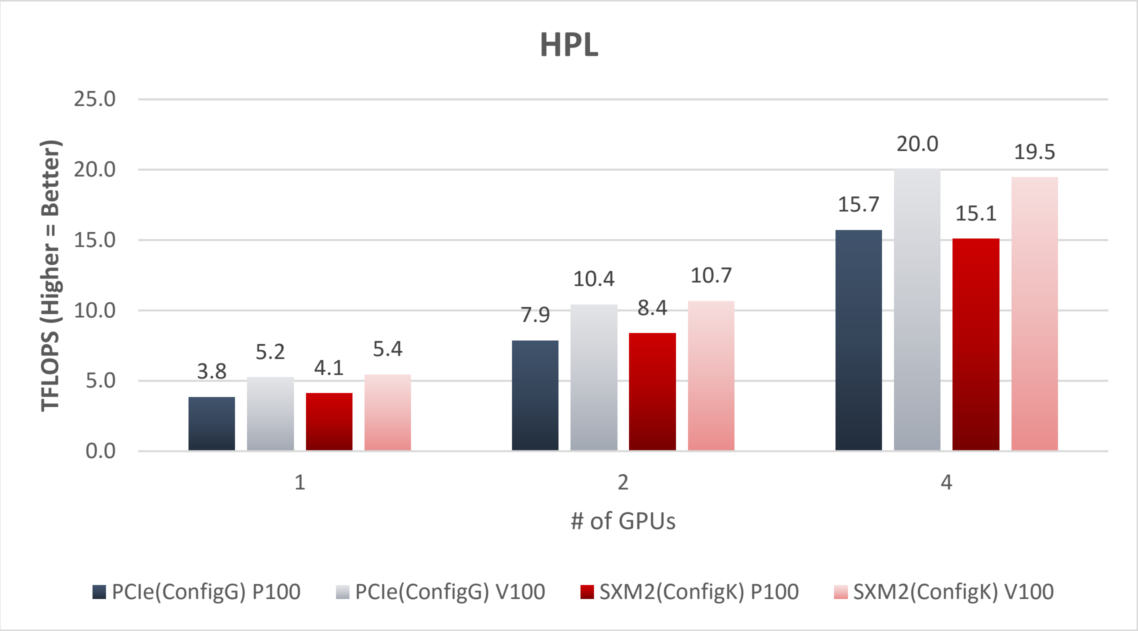


Figure2: HPL Multi-GPU results with V100 and P100 on C4130 configuration G and K

Figure 2 shows the HPL performance on the C4130 platform with 1, 2 and 4 V100-PCIe and V100-SXM2 installed. P100's performance number is also listed for comparison. It can be observed:

1) Both P100 and V100 scaling well, performance increases as more GPUs are added.

2) V100 is ~30% faster than P100 on both PCIe (Config G) and SMX2 (Config K).

3) A single C4130 server with 4x V100 reaches over 20TFlops on PCIe (Config G).

HPL is a system level benchmark and its performance is limited by other components like CPU, memory and PCIe bandwidth. Configuration G is a balanced design, which has 2 PCIe links between CPU and GPU and this is why it outperforms configuration K with 4x GPUs in the HPL benchmark. We do see some other applications perform better in Configuration K, since SXM2 (Config K) supports NVLink, higher core clock speed and peer-to-peer data transfer, these are described below.

HPCG

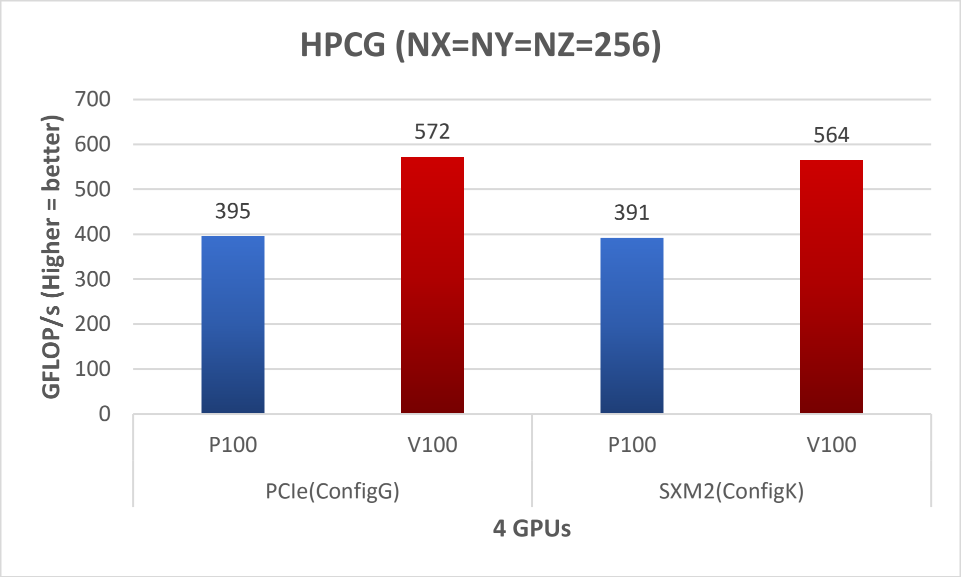


Figure 3: HPCG Performance results with 4x V100 and P100 on C4130 configuration G and K

HPCG, the High Performance Conjugate Gradients benchmark, is another well-known metric for HPC system ranking. Unlike HPL, its performance is strongly influenced by memory bandwidth. Credit to the faster and higher efficient HBM2 memory of V100, the performance improvement we observed is 44% over P100 on both Configuration G and K.

AMBER

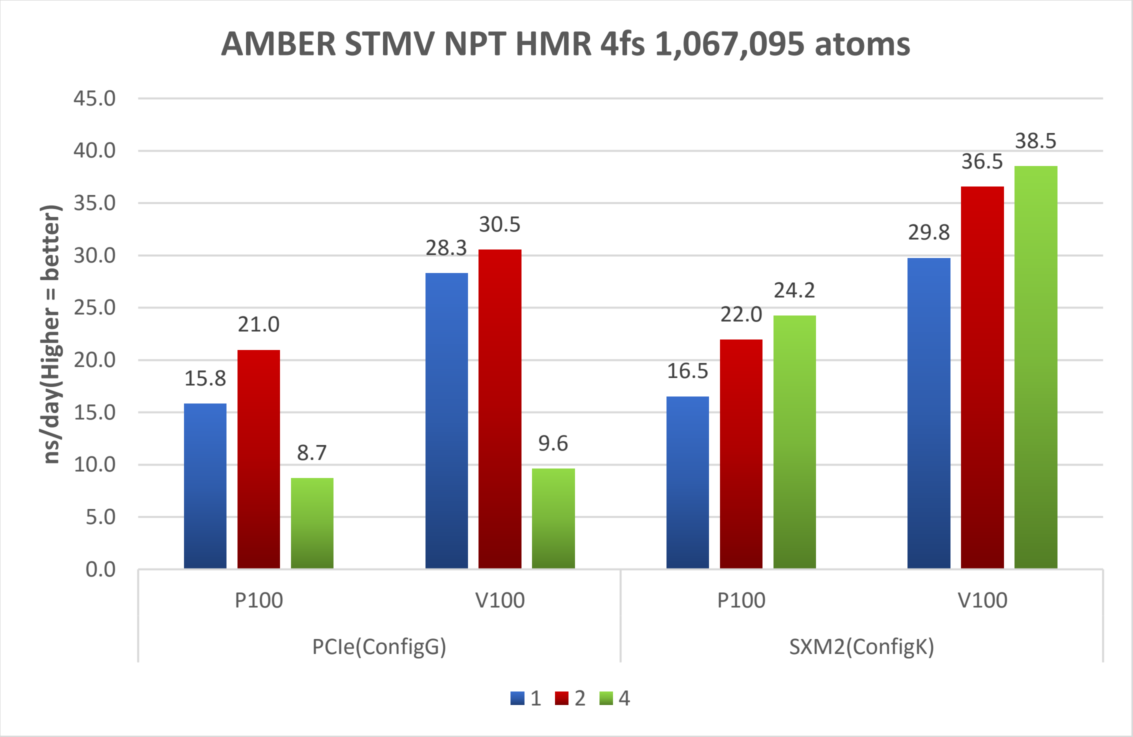


Figure 4: AMBER Multi-GPU results with V100 and P100 on C4130 configuration G and K

Figure 4 illustrates AMBER's results with Satellite Tobacco Mosaic Virus (STMV) dataset. On SXM2 system (Config K), AMBER scales weakly with 2 and 4 GPUs. Even though the scaling is not strong, V100 has noticeable improvement than P100, giving ~78% increase in single card runs, and 1x V100 is actually 23% faster than 4x P100. On the PCIe (Config G) side, 1 and 2 cards perform similar to SXM2, but 4 cards' results dropped sharply. This is because PCIe (Config G) only supports Peer-to-Peer access between GPU0/1 and GPU2/3 and not among all four GPUs. Since AMBER has redesigned the way data transfers among GPUs to address the PCIe bottleneck, it relies heavily on Peer-to-Peer access for performance with multiple GPU cards. Hence a fast, direct interconnect like NVLink between all GPUs in SXM2 (Config K) is vital for AMBER multiple GPU performance.

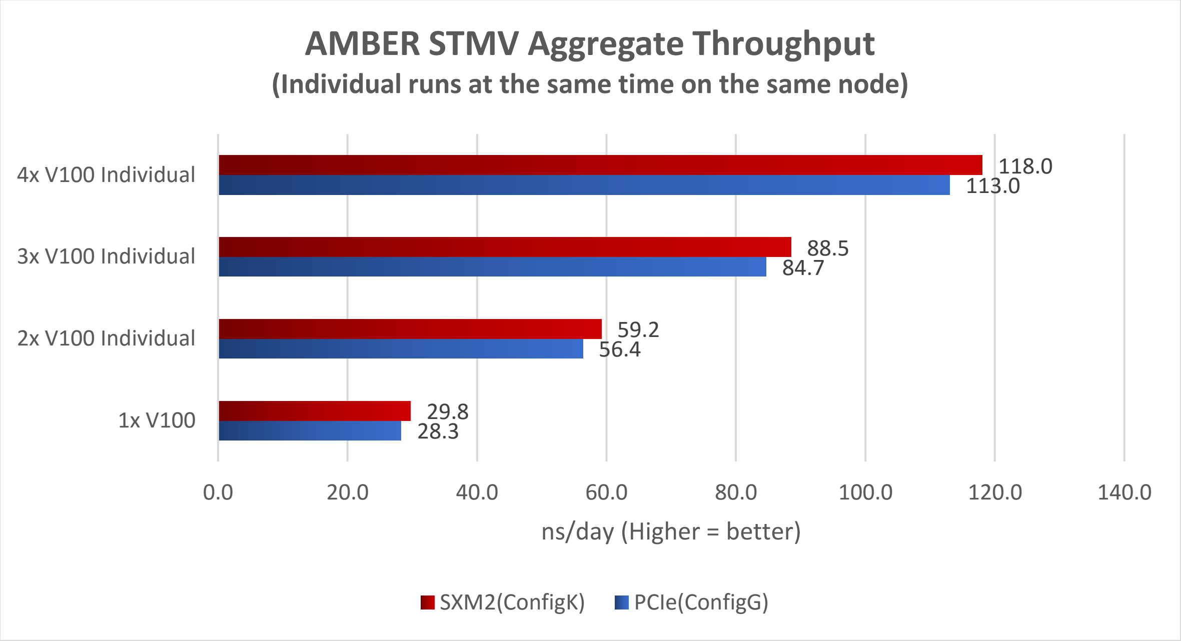


Figure 5: AMBER Multi-GPU Aggregate results with V100 and P100 on C4130 configuration G and K

To compensate for a single job's weak scaling on multiple GPUs, there is another use case promoted by AMBER developers, which is running multiple jobs in the same node concurrently but where each job uses only 1 or 2 GPUs. Figure 5 shows the results of 1-4 individual jobs on one C4130 with V100s and the numbers indicate that those individual jobs have little impact on each other. This is because AMBER is designed to run pretty much entirely on the GPUs and has very low dependency on the CPU. The aggregate throughput of multiple individual jobs scales linearly in this case. Without any card to card communication, the 5% better performance on SXM2 is contributed by its higher clock speed.

LAMMPS

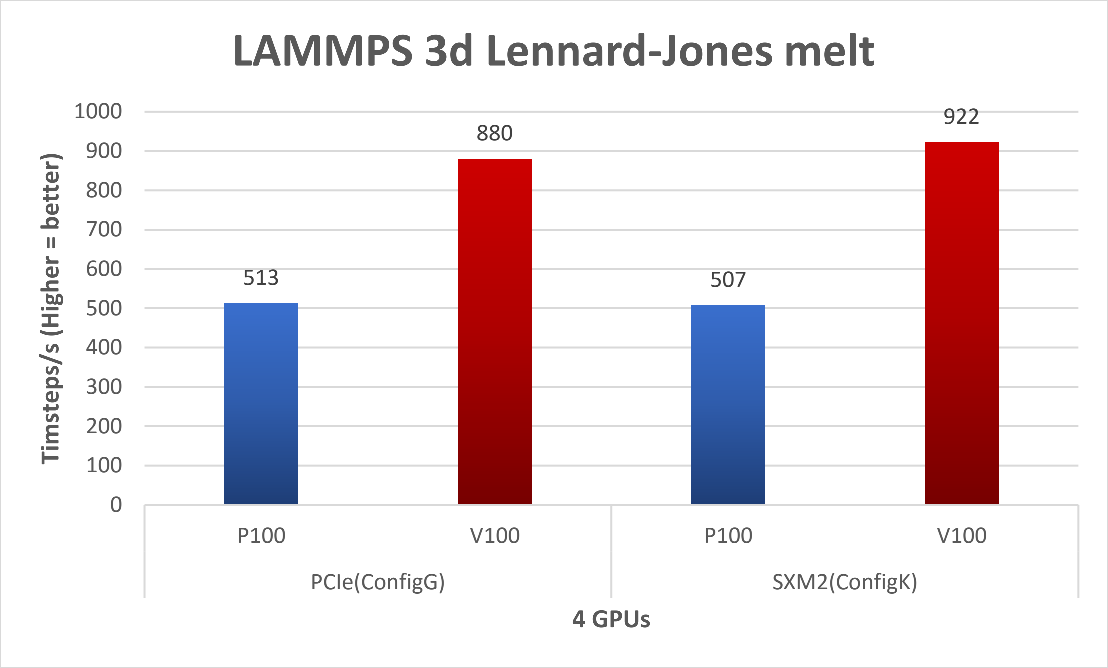


Figure 6: LAMMPS 4-GPU results with V100 and P100 on C4130 configuration G and K

Figure 6 shows LAMMPS performance on both configurations G and K. We use the Lennard-Jones liquid dataset, which contains 512000 atoms, and we compiled LAMMPS with the kokkos package installed. V100 is 71% and 81% faster on Config G and Config K respectively. Comparing V100-SXM2 (Config K) and V100-PCIe (Config G), the former is 5% faster due to NVLINK and higher CUDA core frequency.

Conclusion

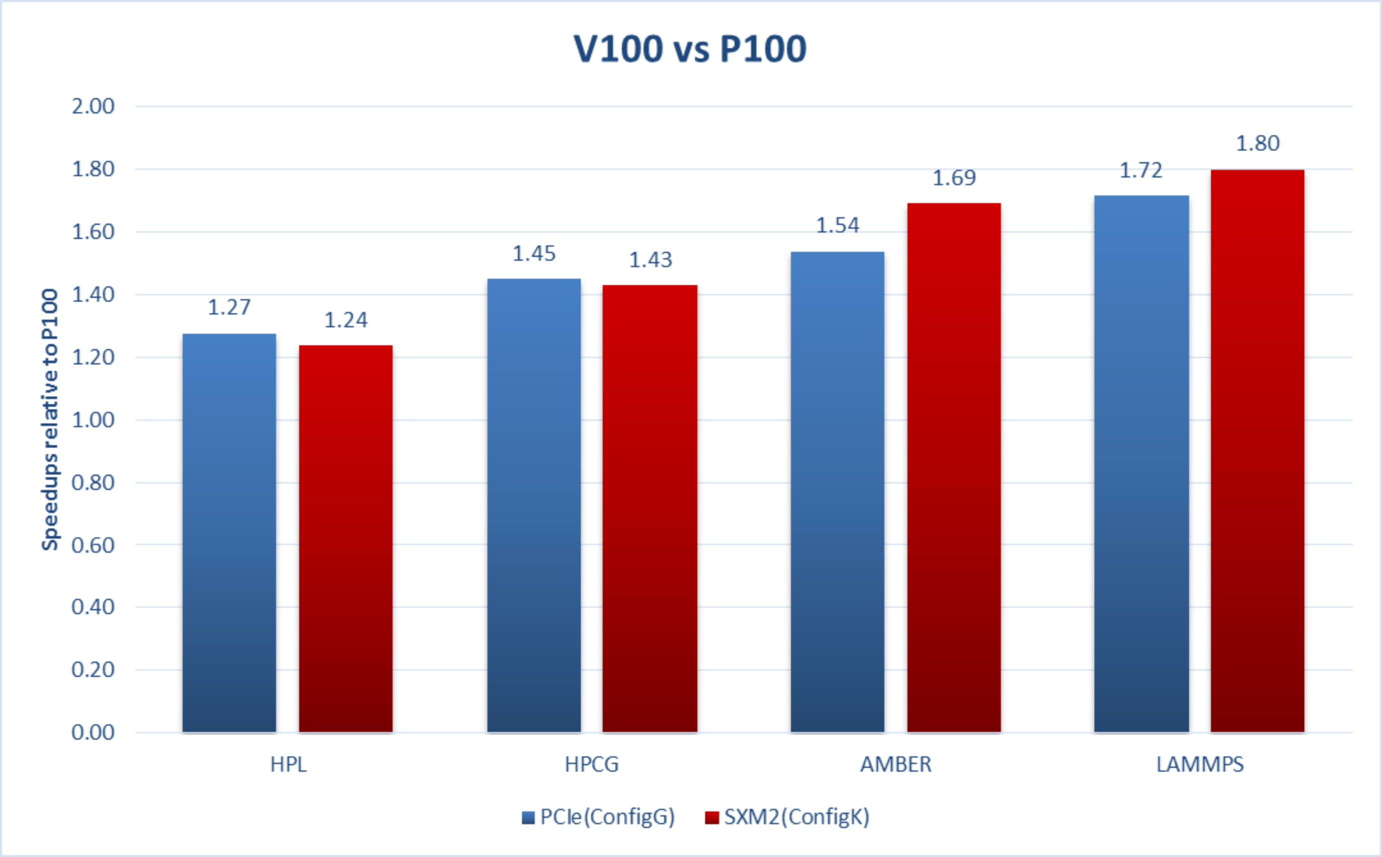


Figure 7: V100 Speedups on C4130 configuration G and K

The C4130 server with NVIDIA® Tesla® V100™ GPUs demonstrates exceptional performance for HPC applications that require faster computational speed and highest data throughput. Applications like HPL, HPCG benefit from the additional PCIe links between CPU and GPU that are offered by Dell PowerEdge C4130 configuration G. On the other hand, applications like AMBER and LAMMPS were boosted with C4130 configuration K, owing to P2P access, higher bandwidth of NVLink and higher CUDA core clock speed. Overall, a PowerEdge C4130 with Tesla V100 GPUs performs 1.24x to 1.8x faster than a C4130 with P100 for HPL, HPCG, AMBER and LAMMPS.