Question 3

Comparing Pascal P100 against Volta V100 architecture

Following are the differences between Pascal P100 and Volta V100.

- Volta V100 has a total of 640 Tensor Cores which accounts for 125 TeraFlops per seconds, and it is 5 times faster than Pascal, whereas Tensor cores are not available in Pascal P100.
- Tesla V100 has 5120 GPU cores whereas Pascal P100 has 2016 GPU Cores.
- P100 supports only a RAM of 16GB but the newly developed V100 supports twice that of P100 that is upto 32 GB.
- P100 supports only a RAM of 16GB but the newly developed V100 supports twice that of P100 that is upto 32 GB.
- CUDA Version 7.0 is supported by V100 while Pascal only supports CUDA version 6.0.
- Newly developed versions of the frameworks such as MXNet, Caffe2, NTK and Tensorflow developed to improve performance in the training time and higher multi-node training performance.
- The P100 architecture supports a memory bandwidth of 720.9 GB/s whereas the advanced V100 supports a range of 900.1 GB/s.
- The memory clock speed of V100 is 1758 MHz while P100 supports a memory clock speed of almost 1430 MHz.
- Tesla P100 operates at a GPU frequency of 1480 MHz and comparatively V100 delivers a GPU Boost clock of 1530 MHz.
- V100 is equipped with a new streaming Multiprocessor optimized for deep learning functionalities.
- Both these GPUs are provided with HBM2 memory since it is in the same package as the GPU and hence it provides space savings.
- When accounted for the Floating-point performance V100 leads with 14029 GFlops whereas its competitive P100 provides only 10609 GFlops.
- V100 provides a streamlined instruction set for simpler decoding and reduced instruction latencies.
- L1 cache is enhanced in V100 than in P100 and hence it provides higher performance and lower latency.

References:

1. https://developer.nvidia.com/blog/inside-volta/
2. https://www.e2enetworks.com/tesla-v100-vs-p100-do-you-know-the-differences/
3. https://segmentnext.com/nvidia-volta-tesla-v100-p100/