

3. Read the Pascal whitepaper provided, and then identify the key features that were introduced in the Pascal P100 architecture, comparing those features against the Ampere-based A100 architecture (make sure to identify the source for the information you obtained on the A100). Please do not just repeat what you read in the Pascal whitepaper, go into more detail on each of the features you identify.

Compute Capability:

The compute capability of a GPU can define its general specifications and available features. P100 supports 6.0 and A100 supports 8.0. The streaming multiprocessor of the 6.0 architecture consists of 64 or 128 CUDA cores to operate arithmetic tasks, 16 or 32 special function units to deal with single-precision transcendental tasks and 2 or 4 warp schedulers. The streaming multiprocessor of the 8.0 architecture offers more powerful and versatile cores, which include 64 FP32 cores for single-precision computation, 64 FP64 cores and 128 FP32 cores for double-precision computation, 64 INT32 cores for integer math, 4 mixed-precision Third Generation Tensor cores, 4 warp schedulers. The 6.0's shared memory behaves the same way as 5.x, but the 8.0's shared memory behaves the same way as Volta architecture, whose amount of the unified data cache is configurable on a per kernel basis. 8.0 allows a single thread block to address up to 163 KB of shared memory, but 6.0 only can address up to 64 KB. Also 6.0 don't have L2 cache residency management.

Process Technology:

A100 is designed to power computationally-intensive AI, HPC and data analytics tasks and it makes greater contribution to cloud and data center platform compared to P100. A100 includes 54.2 billion transistors with a die size of 826 mm², fabricated on TSMC's 7nm N7 manufacturing process, while P100 only supports 16 nm. The P100 supports 1190 MHz and the A100 supports 1065 MHz in base frequency, while the P100 supports 1329 MHz and the A100 supports 1410 Mhz in accelerated frequency. The A100 also has higher memory frequency and larger memory bit width than the P100, which also makes the A100 lower power consumption.