CMP SCI 635: Modern Computer Architecture

NVIDIA Tesla P100 and NVIDIA TURING GPU ARCHITECTURE

Name: Kunjal Panchal Date: 19th Nov, 2019 Student ID: 32126469

Paper: 20 – Data Parallel [NVidia16+NVidia18]

Strengths:

1. [Pascal] NVIDIA's Tesla P100 GPU Accelerator is a well-rounded high-end processor for both high performance computing and data learning applications. A new native data type, half precision or FP16, is introduced to essentially double the TFLOP performance for DL applications.
2. [Pascal] With support for PCIe Gen 3, which offers a bandwidth of 32GB per second, each P100 processor provides the throughput of more than 32 CPU-based nodes. Its 18.7 teraflops of half-precision performance, however, means that the PCIe version will take a slight performance hit compared to the original P100, which offers 21 teraflops. That difference won't be much of an issue for massive server banks.
3. [Turing] Nvidia made a gamble by devoting a significant portion of a huge die to tensor and RT cores to make this leap, as it is very expensive to produce on 12nm. The hardware had to be launched before the game devs showed interest in creating games that require raytracing, and Nvidia picked a time when they are far ahead of their competition in the performance high end.
4. [Turing] Overclocking headroom is good – GPU Boost 4.0 works with the NVIDIA/PrecisionX1 scanner utility very well even though it is in beta.

Weaknesses:

1. [Pascal] The big cloud players end up with GPUs that are out of date. Google announcement about adding P100 to GCP mentions the K80 finally reaching general availability yet the K80 was launched by NVIDIA in 2014. That's ancient, especially in the world of GPUs. GPUs typically have a four-year lifecycle so the K80 will actually be end-of-life very shortly. With the new generation (Volta) coming out after that, it would be hard to make the case that these GPUs are worth anything with the exception of maybe certain use-cases where performance really doesn't matter.
2. [Pascal] Despite being 3x expensive than K80, P100 is claimed to gove 2-4x performance improvements. But it is known that, DL benchmarks as cuDNN has Pascal specific optimizations so the improvements compounds, a lot. Also, contrary to GCP claims, when we attach 1 K80 to our instance we only

get one half of the board, we need to attach 2 to get the full K80. Google can't do that with the P100.

3. [Pascal] The Tesla P100 comes in 2 form factors. NVIDIA's P100 PCIe form factor features the traditional PCIe 3.0 x16 interface for card-to-card and card-to-CPU communications. For some applications the PCIe interface can be a bottleneck limiting the overall system performance.
4. [Turing] It feels like the cards have all basically the same performance with the exception of the 2080ti. Nvidia just flip-flops the names to make consumer think they are actually getting a performance, but in reality, they were just paying for a pascal card with tensor cores for raytracing.

Questions/Assertions:

1. [Pascal] Nvidia Tesla P100 gives bad performance on Google Cloud. The main issue in gaming is the driver optimization, GeForce GPUs are optimised for gaming but the GRID drivers have the primary purpose of computing. If we can port desktop GPU drivers for it, we can get the level of performance we are expecting from the hardware power.

2. [Pascal] Is there any reason for why we would want to get anything pre-Volta; as Volta is a paradigm shift architecturally [Independent thread scheduling].

3. [Pascal] Half precision is implemented on the software layer, but not on the hardware layer for these cards. This means you can use 16-bit computation but software libraries will instead upcast it to 24-bit to do computation (which is equivalent to 32-bit computational speed. This means that you can benefit from the reduced memory size, but not yet from the increased computation speed of 16-bit computation.

4. [Turing] Is it possible to use two of these cards for extreme SLI performance with new higher-bandwidth bridges?

5. I keep wondering, both TU102 (RTX2080Ti) & TU104 (2080) has a very large die. Despite 12nm is a mature process. Such large die will have enough defective chips that couldn't make it to 2080Ti/2080 SKU. Where those chips go? Do they actually use those chips in Quadro cards? How does that work?