Sergi Aliaga

Homework 5

Question 3

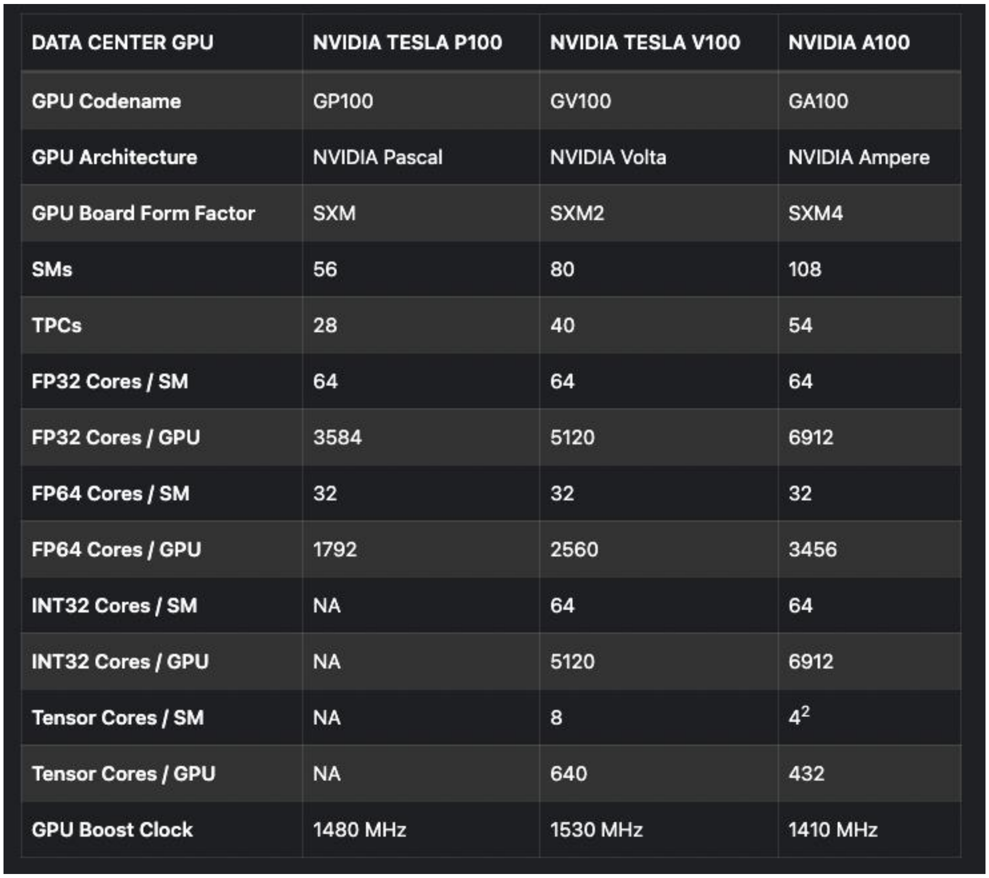
Read the Pascal whitepaper provided, and then identify the key features that were introduced in the Pascal P100 architecture, comparing those features against the Ampere-based A100 architecture (make sure to identify the source for the information you obtained on the A100). Please do not just repeat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

NVIDIA's Pascal P100 architecture was introduced in 2016 as the 'most advanced datacenter accelerator' architecture. As indicated, such architecture was destined to improve the performance for computing hardware in data centers.

On the other hand, the Ampere A100 architecture, recently introduced (May 2020) is NVIDIA's first 7nm architecture focused on AI and neural network related computations. Among other advancement, the A100 architecture leverages third-generation Tensor Cores, i.e. processing units that accelerate the process of matrix multiplication, one of the most recurrent operations in the field of AI and Deep Learning.

Table 1 presents a summary of the technical aspects of the most recent architectures from NVIDIA: the Pascal, Volta and Ampere architectures.

We observe that the Ampere architecture offers higher specs in almost every category, in addition to the inclusion of tensor cores, which are not present in the P100 architecture. It is worth mentioning that the Ampere architecture has a reduced number of tensor core per GPU, which might lead to think that it would result into a worse performance. However, the A100 architecture achieves a 20x overall mixed-precision compute capabilities compared to P100 thanks to its main differential characteristic: the use of fine-grained structured sparsity. As studied in previous homework assignments, the use of sparse matrix representations can greatly increase the throughput of the general matrix multiplication, and it's the main secret of the A100 performance improvements.



DATA CENTER GPU	NVIDIA TESLA P100	NVIDIA TESLA V100	NVIDIA A100
GPU Codename	GP100	GV100	GA100
GPU Architecture	NVIDIA Pascal	NVIDIA Volta	NVIDIA Ampere
GPU Board Form Factor	SXM	SXM2	SXM4
SMs	56	80	108
TPCs	28	40	54
FP32 Cores / SM	64	64	64
FP32 Cores / GPU	3584	5120	6912
FP64 Cores / SM	32	32	32
FP64 Cores / GPU	1792	2560	3456
INT32 Cores / SM	NA	64	64
INT32 Cores / GPU	NA	5120	6912
Tensor Cores / SM	NA	8	4[2]
Tensor Cores / GPU	NA	640	432
GPU Boost Clock	1480 MHz	1530 MHz	1410 MHz

Table 1: Specifications of the three most recent architectures from NVIDIA

Sources:

https://www.hardwaretimes.com/nvidia-ampere-architectural-analysis-a-look-at-the-a100-tensor-core-gpu/

https://technical.city/en/video/Tesla-P100-PCIe-16-GB-vs-Tesla-A100

https://en.wikipedia.org/wiki/Ampere_(microarchitecture)