Q3:

3. (20) Read the Ampere whitepaper provided and then identify the key features that were introduced in the Ampere A100 architecture and compare those features against the Hopper-based H100 architecture (make sure to identify the source for the information you obtained on the H100). Please do not just repeat what you read in the Ampere whitepaper, go into more detail on each of the features you identify.

A100 Key Features:

The Ampere A100 GPU architecture was first released in May of 2020. Its release was a significant improvement to the previous industry standards due to several innovations.

For background, the GA100 silicon die is comprised of 8 GPU processing clusters, 8 texture processing clusters, 2 streaming multiprocessors per TPC, 16 streaming multiprocessors per GPC, 128 total SMs, and HBM2 memory controllers. The A100, a little scaled back, contains 108 SMs, with 64 FP32 CUDA cores per SM, and 4 tensor cores per SM. The importance of these features is described in more detail below.

It also possessed 192KB of shared memory and L1 data cache per streaming multiprocessor, which was 1.5 times larger than the V100. This integration, as well as the larger size, amplifies the performance with more efficient memory access.

The importance of these features is described in more detail below.

The third-generation tensor cores provided immense value to the A100. These enabled acceleration for every data type, as well as a sparsity feature for neural networks. This sparsity feature doubles compute throughput. It is enabled through a 2:4 sparse matrix allowing 2 non-zero values per 4 size vectors. This enables pruning, as it skips computation on matrix entries with a value of 0.

In addition to feature improvement, the product greatly reduced latency, improving bandwidth by 70% in HBM2 memory compared to the V100. It also possesses an enlarged L2 cache of 40 MB, reducing latency for memory-intensive workloads.

A100 vs H100:

The H100, a newer GPU version, advanced the improvements of the Ampere architecture. It included 4th-generation tensor cores, which added FP8 support and dynamic switching between FP8 and FP16. This is beneficial in supporting diverse AI workloads, and combined with a transform engine, greatly enhances the efficiency of computationally intensive processes. The sparsity feature in the H100 was enhanced to produce a 4 times speedup and FP8 operations. The performance of the H100 is much greater, 6.4 times the number of TLOPS for sparse operations at FP8 and FP16.

In terms of physical structure, the H100 introduced HBM3, improving memory bandwidth by double that of the A100. It also increased the size of the L2 cache by 25%, and the L1 cache/shared memory by 33%. These optimizations take place on a denser transistor architecture and improve performance for memory-heavy workloads in high performance computing.

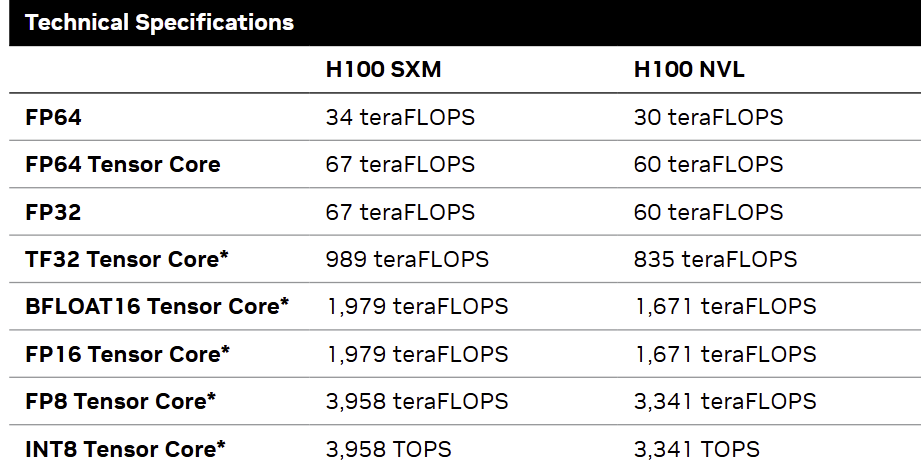
Overall, the performance of the H100 for FP64 workloads and FP16 AI training is greater than 6 times better, and AI inference with FP8 size indices was an innovation that was not present in A100.

These optimizations demonstrate the value of H100 for memory intensive workloads such as LLMs and other AI workloads.

Additional Notes:

H100 Features & Performance:

* <https://www.nvidia.com/en-us/data-center/h100/>



A100 Features & Performance:

* 40 GB HBM2 and 40 MB L2 Cache
* 1555 GB/sec of memory bandwidth -> 73% increase over V100
* Crossbar structure in L2 to cache read bandwidth
* Third gen NVLink

Full Implementation:

* 8 GPCs, 8 TPCs, 2 SMs, 16 SMs, 128 SMs per full GPU
* 64 FP32 CUDA cores per SM, 8192 FP32 CUDA cores per full GPU
* 4 third-generation tensor cores, 512 third-generation tensor cores per full GPU
* 4 third-gen tensor cores
* 6 HBM2 stacks, 12 512-bit memory controllers

