Question – 3

Assignment - 5

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.

Answer:

Ampere A100 architecture

- **Third-Generation Tensor Cores**: The throughput and number of data types that ampere's Tensor Cores can handle are both higher than in earlier generations. They added capabilities for sparsity, which skip zero-value calculations in sparse matrices and essentially double the math throughput when working with AI tasks that use sparse neural networks. This new idea is especially helpful for deep learning because it can cut down on extra work and memory needs. This speeds up the training and inference stages without affecting the accuracy.
- **Multi-instance GPU**: With the MIG feature, a single A100 GPU can be split into multiple separate GPU instances. This keeps tasks from competing for GPU resources and keeps them safe from each other. This is very important for businesses and cloud service providers that want to get the most out of their hardware and make sure that performance stays the same across different tenants or workloads.
- **Enhanced Memory Architecture**: The 40 MB L2 cache and 40 GB HBM2 memory that came with the A100 are bigger. These changes make the bandwidth and memory bigger, which is needed to work with bigger information and models. As the size of the cache and memory grows, it becomes less necessary to reach slower external memory sources. This speeds up computation and makes it possible to work with bigger, more complicated models.
- **NVLink and PCIe Gen 4**: These technologies offer faster interconnects than older generations, which is very important for programs that need to quickly send data between multiple GPUs. Better connectivity helps AI and HPC apps run more efficiently by letting data be sent more quickly, which is important in multi-GPU setups for large-scale AI training sessions and complex simulations.

Hopper H100 architecture:

- **Fourth-Generation Tensor Cores**: Tensor Cores in the H100 handle new data types like FP8 and have better features for existing types, which speeds up computations by a large amount. These improvements make it possible for AI models to be trained and inferred

faster. This is especially helpful for next-generation AI apps, like those that need to work in real time.

- **Transformer Engine**: This new part is made to speed up the training and prediction of transformer-based models, like those used in natural language processing, better than GPU cores that are used for other tasks. It lets the H100 handle AI tasks up to 30 times faster, especially when it comes to handling big language models quickly and correctly.
- **Enhanced Memory and Connectivity**: When 80 GB HBM3 memory and upgrades to NVLink and PCIe Gen 5 are added to the H100 design, memory bandwidth and interconnect speeds go up by a lot. This speeds up data transfers, lowers latency, and makes it easier for connected GPUs to handle more users. This is very important for pushing the limits of how fast high-performance computing groups can work.
- **Advanced Security Features**: NVIDIA's Confidential Computing features are built into the H100. These features improve data protection during computation by blocking unauthorized access. With these security improvements, H100 can handle sensitive or private workloads, keeping data safe even in settings with multiple tenants.

Source: https://www.advancedclustering.com/wp-content/uploads/2022/03/gtc22-whitepaper-hopper.pdf?utm_source=chatgpt.com