Khaula Molapo

20240001

Computer Architecture – Week 11

Research and Compare: NVIDIA Hopper Architecture:

The NVIDIA Hopper architecture, launched in 2022, is one of the most advanced GPU architectures ever built. It is designed mainly for AI training, high-performance computing (HPC), and data center workloads. The flagship chip, H100, is based on this architecture and built using TSMC's 4nm process technology.

Hopper introduces a large number of Streaming Multiprocessors (SMs) — up to 132 SMs in total. Each SM contains multiple CUDA cores that handle thousands of threads in parallel. This design allows Hopper to perform trillions of calculations per second, making it ideal for deep learning and large-scale simulations.

In terms of memory, the Hopper GPU uses HBM3 (High Bandwidth Memory), providing extremely high memory speed and bandwidth. The H100 has up to 80 GB of HBM3 memory with a bandwidth of over 3 TB/s, allowing massive data to be accessed quickly. It also includes a memory hierarchy that consists of registers, L1 cache per SM, shared L2 cache, and global memory. This structure reduces latency and improves data reuse between threads.

One of Hopper's biggest advancements is the introduction of the Transformer Engine, which is specially designed to accelerate AI models like GPT and BERT. It uses FP8 precision to balance speed and accuracy during deep learning computations. The architecture also supports Multi-Instance GPU (MIG) technology, allowing one GPU to be split into smaller independent GPUs for different workloads.

The target domain for NVIDIA Hopper is AI computing, large data centers, and scientific research. While it can also be used for high-end graphics, its real strength is in parallel data processing, neural network training, and cloud-scale AI workloads.

In short, Hopper is not just faster; it's smarter — designed to handle the massive and complex data demands of modern computing. It shows how GPU architecture continues to evolve from gaming hardware into the backbone of global AI and HPC infrastructure.

2. Performance Analysis
Given:
- Task is 95% parallelizable (p = 0.95)
- CPU: 32 cores

- GPU: 4096 cores

Amdahl's Law:
Speedup = 1 / [(1 – p) + (p / n)]
= 1 / [(1 – 0.95) + (0.95 / 4096)]
= 1 / [0.05 + 0.000232]
= 1 / 0.050232
≈ 19.9× speedup

So, in theory, the GPU could make the task almost 20 times faster than before.

Reflection

In real life, we rarely reach this theoretical speedup. Many practical factors limit performance. One big issue is memory bandwidth — if the GPU cannot feed data fast enough, the cores will sit idle. Communication overhead between the CPU and GPU also slows things down, especially when large amounts of data must be transferred repeatedly.

Another problem is warp divergence, which happens when threads in a GPU take different execution paths, reducing parallel efficiency. Some parts of a program may still be sequential and cannot use all cores effectively. Thermal limits, power constraints, and software optimization also affect actual performance.

In summary, while the GPU has thousands of cores, real-world results depend on how efficiently the code and memory system are designed to use them. Theoretical performance is the ceiling — practical performance is usually much lower.