DS 221 - Introduction to Scalable Systems: Architectures for AI

Shankaradithyaa Venkateswaran
Sr no: 22190

October 28, 2024

1 Tensor Cores

1.1 Introduction

Tensor cores have emerged as a key component in accelerating artificial intelligence computations, used in deep learning networks. They are specialized hardware units that are designed to perform matrix operations efficiently to speed up AI computations.
Tensor cores are found in the NVIDIA architectures where they have become an irreplaceable unit of computation for those architectures. Today I will be taking an indepth look into tensor cores and why they are better than regular GPUs and CPUs for AI computations.

1.2 What are Tensor Cores?

Tensor cores are specialized pieces of hardwares found in the NVIDIA Volta, Turing, Ampere, Hopper, and now Blackwell architectures. They are specially designed to perform matrix multiplications efficiently so that they can be used to accelerate deep learning computations.
Other companies do not have tensor cores in their GPUs, but they have similar units that perform the same function. For example, AMD has matrix cores in their GPUs that perform the same function as tensor cores. In a later section, I will look into such units in other GPU architectures.

1.3 Features of Tensor Cores

As I have mentioned above, Tensor cores are designed to perform matrix multiplications efficiently. They are highly specialized compute cores for matrix multiply and accumulate (MMA) math operations. These operations are the core aspect for deep learning computations and HPC. This operation is to use a lower precision such FP16 (floating point) or INT8 (integer) to perform matrix multiplications and accumulate the results in a higher precision format such as FP32 or FP64. The core multiplies two lower precision matrices and then adds the result to an accumulator matrix which is the same precision or a higher precision. This is then converted to a higher precision format to minimize the loss of precision. This process leads to a preservation of accuracy through the entire process of calculations. The tensor cores are designed to perform these operations at a very high speed and are capable of performing 64 FMA (fused multiply-add) operations per clock cycle.

Tensor cores have a high throughput and low latency, which makes them ideal for AI computations. They are capable of performing 125 teraflops of mixed-precision performance.
As we can infer from above, Tensor cores are made to support various datatypes, including FP16, FP32, BF16, INT8, and INT4. This makes them highly versatile and capable of performing a wide range of operations for AI computations.
These cores have structured sparsity present in newer architectures such as Ampere and onwards. This means that they can efficiently proccess sparse matrices and ignore the zero values in the matrix. This is a significant improvement over the older architectures where the tensor cores would have to process the zero values as well.
Overall, tensor cores are also highly energy efficient to perform matrix calculations on them. They are capable of performing a large number of operations in a short amount of time while consuming less power.

1.4 Tensor Cores over GPUs and CPUs

Now the topic comes to why chose tensor cores over GPU and CPU cores. While we can perform AI computations on such cores, it is highly inefficient as CPU and even GPU cores are made with a general purpose use in mind. In the field of a CPU core, it is made to perform a variety of calculations, not just matrix multiplications, hence reducing the efficiency of AI calculations. Now if we take the example of a GPU core, while it made to perform parallel computations along with matrix operations, it is still not as highly optimized as tensor cores. GPU cores have a variety of other functions as well, rendering pixels, paralle processing, image and video processing, scientific simulations, etc. This means, while they can be used for AI computations, they are not as efficient as tensor cores.
Now while CPU and GPU cores can perform matrix multiplications and both of them support MMA in some form or the other, it can safely be said from the statisitcs that tensor cores are much more efficient at performing these operations. As mentioned above, tensor cores can perform 64 FMA per clock cycle. This is a significant improvement over the regular GPU cores which can perform only 2 FMA operations per clock cycle.
Overall, it is just not efficient to run AI or matrix heavy computations on CPU or GPU cores. Tensor cores are highly optimized for these operations and are capable of performing them at a much higher speed and efficiency. Hence it is better to offload such operations to the tensor cores, and clear up space for different operations that are required by the CPU and GPU cores.

1.5 Conclusion

In conclusion, tensor cores are highly efficient and specialized hardware units that are designed to perform matrix multiplications efficiently which directly lead to a collosal speed up in AI computations over the use of regular CPU and GPU cores.

2 HPC and AI

Now we will explore the latest state of art HPC architectures tuned for AI. To do this, we will go through three architecture leaders, namely AMD, Intel, and NVIDIA. We will look into their latest architectures and how they are tuned for AI computations.

2.1 AMD

AMD has been a leader in the CPU market for a long time now. They have been making strides in the GPU market as well.

AMD INSTINCT™ MI325X ACCELERATOR

The figure is a AMD INSTINCT™ MI325X ACCELERATOR designed for increasing demands of AI computations and delievers a amazing HPC performance as well.

The discrete AMD Instinct MI325X GPU delievers a superior performance on a broad set of data types needed for AI software, including FP16, BF16, FP8, and INT8. It has an industry leading 256 GB of HBM3E memory and 6 TB/s of memory bandwidth which allows a single acclerator to contain and process a one-trillion parameter model while reducing the cost of ownership for selec LLMs. It has the following performance:

AI Peak Performance	without sparsity	with sparsity
TF32 (TFLOPs)	653.7	1307.4
FP16 (TFLOPs)	1307.4	2614.9
BFLOAT16 (TFLOPs)	1307.4	2614.9
INT8 (TOPs)	2614.9	5229.8
FP8 (TFLOPS)	2614.9	5229.8
HPC Peak Performance		
FP64 vector (TFLOPs)	81.7	NA
FP32 vector (TFLOPs)	163.4	NA
FP64 matrix (TFLOPs)	163.4	NA
FP32 matrix (TFLOPs)	163.4	NA

The AMD Instinct MI325X accelerator is an AMD CDNA 3 architcture-based accelerator with high throughput based on improved AMD Matrix cores. Matrix cores are the equivalent of tensor cores in NVIDIA GPUs. They are designed to perform matrix multiplications efficiently. Each discrete MI325X GPU offers a 16-lane host interface with PCIe Gen 5 support, enabling high-speed data transfer between the host and the accelerator.

The features of the MI325X accelerator which accelerate AI computations can be broadly described as the increased memory bandwidth and the memory capacity of the accelerator. The architecture itself is purpose built to drive the most demanding AI computations. The specifications include:

Feature	Capacity
Form factor	OAM (Open Accelerator Module)
GPU compute units	304
Matrix cores	1216
Stream processors	19,456
Peak Engine Clock	2100 MHz
Memory Capacity	Up to 356 GB
Memory Bandwidth	6 TB/s (peak theoretical)
Memory Interface	8192-bit HBM3E
PCIe Gen 5 x16	128 GB/s
Infinity Fabric Link	7x 128 GB/s

the infinity fabric link is a high-speed interconnect that allows for high-speed data transfer between the GPUs. Each OAM module includes: Eight accelerated compute dies (XCDs) with 38 compute units (CUs), 32 KB of L1 cache per CU, 4 MB shared L2 cache shared across CUs, and 256 MB of AMD Infinity Cache™ shared across 8 XCDs.

These compute units support a broad range of precisions for both AI/ML and HPC acceleration, native hardware support for sparsity, and enhanced computations throughput. The OAM also includes: Four supported decoders for HEVC/H.265, AVC/H.264, VP9, or AV1, each with an additional 8-core JPEG/MPEG CODEC, 256 GB of HBM3E memory with 6 TB/s on-package peak throughput, and SR-IOV for up to 8 partitions.

(Note: I did not go too indepth into the above terms)

AMD has also introduced the ROCm (Radeon Open Compute) platform which is an open-

source software platform that allows for the development of high-performance computing and AI applications. It is designed to work with AMD GPUs and accelerators. All these combine to form an architecture that is highly optimized for AI computations and HPC. Now these AMD MI325X accelerators can be stacked on each other on a AMD Universal Base Board (UBB 2.0) with HGX host connecters to form the AMD Instinct MI300X MI325X Platform.

2.2 Intel

Moving on let us now talk about Intel and their latest architecture for HPC which are tuned for AI computations. Intel, just like AMD has made its own language to support HPC which is called the oneAPI. This supports HPC standards including C/C++, Fortran, Python, OpenNP, and MPI for integration.

Intel® Data Center GPU Max Series

Like the Intel® Xeon® CPU Max Series we have the Intel® Data Center GPU Max Series. The previous code name of this series was the Ponte Vecchio.

This GPU product is based on the Intel X^e HPC micro-architecture. The GPu takes advantage of highly parallelized computing models associated with AI and HPC. Just like the CPU max series, the GPU max series is supported by the oneAPI open exosystem with the flexibility of Single Instruction Multiple Data (SIMD) and Single Instruction Multiple Thread (SIMT) programming models.

The heart of the product is the Intel X^e HPC stack. The components of the stack are as shown in the above figure. All these components reside within a Multi Tile Package.

This GPU will have OAM modules, just like the AMD MI325X accelerators. The Max Series 1350 GPU has a 450W TDP and 96GB of HBM2e memory. It will have 112 X^e cores.

The larger OAM module will have full 600 W 128 GB HBM2e memory and a full 128 X^e cores. All this will be called the Max Series 1550 GPU.

To link these OAM modules, Intel has its own connection system called the Intel X^e Link. These links will provide high speed interconnects between the OAMs to provide high performance for AI and HPC workloads.

Both tehe Max Series 1350 and 1550 GPUs are connected with the PCI express 5.0 interface and with both of them having up to 4x HBM2e 3.2GT/s

The performances of each are as follows:

	Max Series 1350	Max Series 1550
FP64 PEAK FLOPS	44 TFLOPS	52 TFLOPS
FP32 PEAK FLOPS	44 TFLOPS	52 TFLOPS
BF16 PEAK FLOPS	832 FLOPS	704 TFLOPS
INT8 PEAK FLOPS	1664 TFLOPS	1408 TFLOPS

Overall the performance of the Intel Data Center GPU Max Series is comparable to other GPUs in the market.

2.3 NVIDIA

Finally, we come to NVIDIA. NVIDIA has been a leader in the GPU market for a long time now. We will dive into their latest architectures and see what makes them the leaders of the AI and HPC market.

NVIDIA Hopper Architecture

NVIDIA's Hopper architecture is the latest architecture from NVIDIA on the market. It is stated to securely deliver the highesh performance computing with lowe latency.

The H100 is NVIDIA's 9th generation data center GPU designed for AI and HPC. It is magnitudes faster thatn the prior generation NVIDIA A100. The H100 carries over the major design

focus of A100 and improved the strong scaling for AI and HPC workloads. The following are the results of benchmarking the H100 against the A100. The H100 has multiple new features and improvements over the A100. These features are:

- New Streaming Multiprocessor:
  - New fourth generation Tensor Cores which are upt o 6x faster chip to chip compared to A100, including per SM speedup, additional SM count, adn higher clocks of H100. Overall, this a giant upgrade to the A100 as we have seen that tensor cores are vital for AI computations, and the H100 is just better at it than the A100.
  - 3x faster IEEE FP64 and FP32 processing raters chip to chip compared to A100.
  - New thread block cluster feature which allows programmiatically control of a localily at a granularity larger than a single Thread block on a single SM.
  - A new Asynchronous Execution to enable a new Tensor Memory Accelerator (TMA) to transfer larger blocks of data efficiently between global memory and shared memeory. I need not say how important this is for offloading large datasets for AI computations.
- New Transformer Engine:
  The H100 has a new transformer engine which is a combination of software and custom Hopper Tensor COre technology which directly accelerates Transformer model training and inferences. This lets it deliver up to 9x faster AI training and up to 30x faster AI inference speedups on LLMs compared to A100.
- HBM3 Memory:
  The H100 provides a 2x bandwidth increase compared to A100. As stated by NVIDIA, teh H100 SXM5 GPU is the world's first GPU with HBM3 memory delivering a class-leading 3 TB/sec of memory bandwidth.
- The H100 has a 50MB L2 cache which caches large portions of models and datasets for repeated access.

The performances of the H100 are given below:

	H100 SXM5	H100 SXM5 Sparse
FP8 tensor core	1978.9 TFLOPS	3957.8 TFLOPS
FP16	133.8 TFLOPS	NA
FP16 tensor core	989.4 TFLOPS	1978.9 TFLOPS
BF16 tensor core	989.4 TFLOPS	1978.9 TFLOPS
FP32	66.9 TFLOPS	NA
TF32 tensor core	494.7 TFLOPS	989.4 TFLOPS
FP64	33.5 TFLOPS	NA
FP64 tensor core	66.9 TFLOPS	NA
INT8 tensor core	1978.9 TFLOPS	3957.8 TFLOPS

From this table we can clearly see that tensor cores accelerate the computations, which makes this architecture perfect for both HPC And AI.
The H100 is an engineering marvel which promote HPC and AI, but to truly boost our HPC and dig deep for AI we need to go even further beyond.

NVIDIA DGX

This is NVIDIA's foundational building block for Data centers for AI. The DGX system uses

chips such as the A100 to build a foundational block that can handle large scale computations and, well, HPC. With each generation the DGX can get upgraded with a new chip. Here I will talk about the NVIDIA DGX H100. This is the version of NVIDIA's DGX system with the H100 chip.
The NVIDIA DGX H100 has the following specifications to boost HPC for AI specific tasks:

Feature	Capacity
GPU	8x H100 Tensor Core GPUs
Tensor Cores	4th gen Tensor Cores with sparsity handling
NVLink	4th gen NVLink (connections inside the chips)
NVSwitch	4x - 3rd gen NVSwitch
ConnectX	8x ConnectX - 7 (400 GB/s InfiniBand / Ethernet)
DPU	2x Bluefield 3- DPUs
PCIe	PLCe Gen5

This type of architecture allows for insane data center scalability which is why it is one of the top options in the market today.

Conclusion

This is the end of my report on Tensors cores and why they are helpful for AI and the different architecture leaders and their leading architecutres for HPC tuned for AI.

3 References

Sparsity Docs from NVIDIA
Medium Article on Tensor Cores
Tensor Core Slides by NVIDIA
Tensor Cores Article
Tensor Core docs from NVIDIA
Hopper Whitepaper
DGX A100 Datasheet from NVIDIA
DGX Docs from NVIDIA
AMD Instinct MI325X Datasheet
Article on Intel Data Center GPU Max Series
Intel Data Center GPU Max Series Docs from Intel
Intel Data Center GPU Max Series Overview from Intel
Article Introducing Intel Xeon CPU Max Series and Intel Data Center GPU Max Series
HPC docs from Intel

Disclaimer:

While completing this report, the teams message requiring the document to be at max 6 pages was sent. I have had to cut out a tremendous portion of whatever I wanted to report in response to make the document fit the requirements. Thank you for understanding.