# 1 Tensor Cores in Modern-Day GPU Architectures

Tensor cores are specialized processing units found in modern GPUs, which are designed to accelerate machine learning tasks, specifically deep learning computations [3]. Tensor cores have transformed the area of AI computation since their introduction with NVIDIA's Volta architecture. Their specialized design for matrix operations and mixed precision computing has made them very important for deep learning tasks. As the demand for AI continues to grow, tensor cores remains at the forefront of technological advancements, driving innovation across various industries and applications.
General Matrix-Matrix Multiplications (GEMMs) are the most important building blocks for AI and Deep Learning application. NVIDIA introduced Tensor Cores in their GPUs firstly to accelerate Matrix-Matrix Multiplications. Tensor Core provides significant performance boost and energy efficiency when performing GEMM operations, using mixed precision computation [5]. When Tensor cores were first introduced with the NVIDIA Volta architecture, they provided up to $12\times$ higher peak TFLOPs (Tera FLoating-point Operations Per second) for training of deep learning models. The Volta Stream Multiprocessor (SM) was composed of $64$ FP32 Cores, $64$ INT32 Cores, $32$ FP64 Cores, and $8$ Tensor Cores.

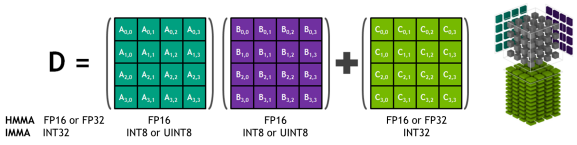
## 1.1 Key Features of Tensor Cores

### 1.1.1 Matrix Operations Optimization

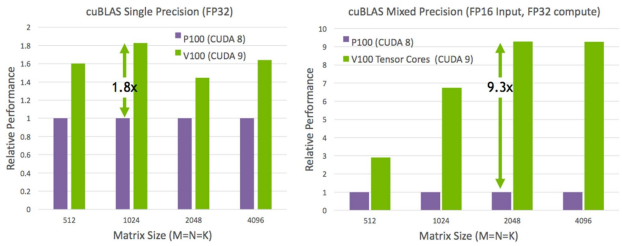
Deep learning relies on matrix multiplications for forward and backward propagation in neural networks. Tensor cores optimize these operations more efficiently than traditional CPU and GPU cores, performing multiple floating-point operations in parallel [6] and handling $4 \times 4$ matrices or larger. This significantly accelerates complex model computations and allows for simultaneous processing of numerous data points [2].

### 1.1.2 Mixed Precision Computing

Tensor cores support mixed precision training, where different parts of the computation can use different precisions (e.g., using FP16 for computation and FP32 for weight storage). This not only speeds up processing but also reduces memory usage, allowing larger models to fit into GPU memory [13, 15]. Many deep learning models can achieve similar accuracy with reduced precision due to techniques like loss scaling, which ensures that gradients do not become too small during training. This balance between speed and accuracy is crucial in practical applications.




(a) Tensor Cores provide fast matrix multiply-add with FP16 and FP32 compute capabilities. [8]




(b) CUDA 9 with Tensor Core deliver up to $9\times$ higher performance for GEMM operations.

### 1.1.3 Increased Throughput

While traditional CPUs are optimized for low-latency operations suited for general-purpose computing, tensor cores prioritize maximizing throughput, which is especially beneficial for training large neural networks that require rapid processing of vast data amounts. Additionally, tensor cores can perform several operations simultaneously, such as executing multiple matrix multiplications at once, significantly enhancing the overall speed of deep learning training and inference processes.

### 1.1.4 Enhanced Memory Bandwidth Utilization

AI models often need frequent access to large datasets, and tensor cores are designed to maximize memory bandwidth utilization, minimizing the limitations seen in traditional processing systems. They also effectively manage data locality, reducing the need for frequent data transfers between the GPU and memory, which is crucial for maintaining high throughput during AI computations.

### 1.1.5 Efficient Resource Utilization

Tensor cores enable dynamic workload distribution, allowing them to adjust based on task complexity. Unlike traditional cores that may remain idle during certain operations, tensor cores continuously work on tensor operations. This optimization in operation scheduling and execution further reduces idle time, enhancing overall utilization of the GPU's resources.

### 1.1.6 Integration with Software Frameworks

Major AI frameworks like TensorFlow and PyTorch have built-in support for tensor cores, allowing developers to easily leverage their capabilities through simple API calls that automatically optimize computations. Additionally, the integration of tensor cores does not necessitate a complete review of existing codebases, as these frameworks can automatically detect and utilize tensor cores.

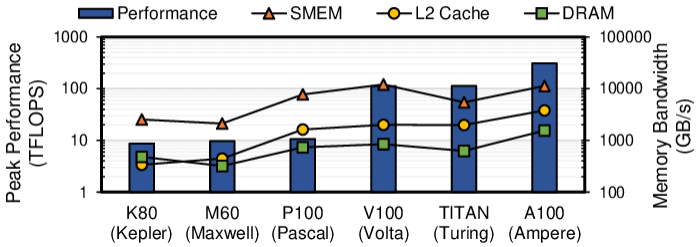
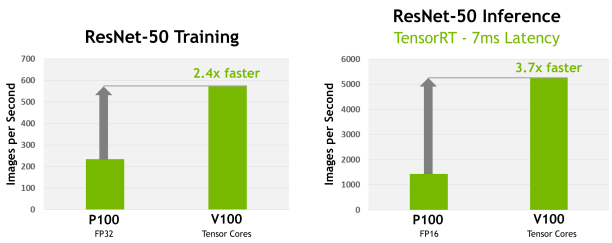
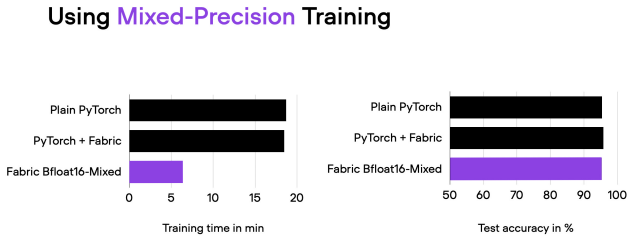



Figure 2: Performance of various NVIDIA GPU where V100, TITAN and A100 (latest) comprises of Tensor Cores shows the peak performance of Tensor Cores GPU over traditional GPUs (K80, M60 and P100). Sudden rise can be seen from P100 to V100 due to inclusion of Tensor Cores in GPUs. [1]




(a) Tesla V100 trains the ResNet-50 deep neural network 2.4x faster than Tesla P100.




(b) Comparison of Mixed Precision shows its supremacy over Single Precision Arithmetic. [4]

### 1.1.7 Reduced Latency for Inference

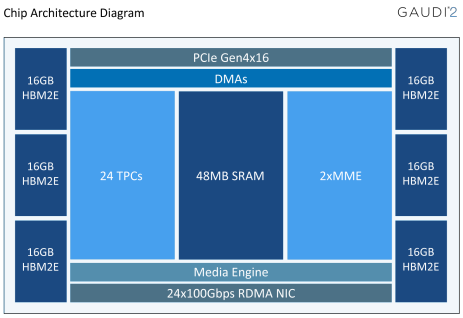
In applications like image recognition, natural language processing, and autonomous systems, low-latency inference is critical, and tensor cores are designed to perform these tasks rapidly, facilitating real-time responses. Additionally, tensor cores efficiently handle batched inputs, making them ideal for deployment scenarios where many inferences need to be processed simultaneously, such as in cloud services or edge devices.

## 1.2 Comparison between CPU, GPU and Tensor Cores

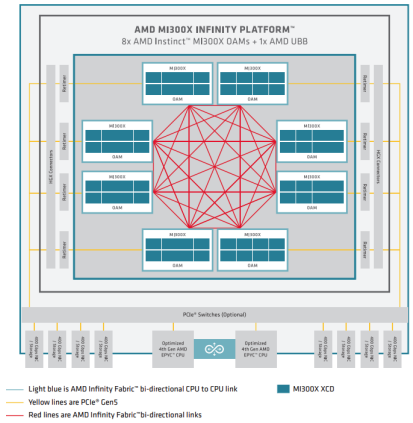
| Feature | CPU Cores | GPU Cores | Tensor Cores |
|---|---|---|---|
| Purpose | General computing | Parallel processing | AI and deep learning |
| Core Count | Few (4-64) | Thousands (up to several thousand) | Specialized units integrated with GPUs |
| Processing Type | Low-latency, sequential | High-throughput, parallel | Optimized for matrix operations |
| Mixed Precision Support | Limited | Basic mixed precision support | Advanced mixed precision support |
| Optimal Use Cases | General sequential tasks, low-latency tasks | Deep learning, graphics rendering, parallelizable tasks | Deep learning training and inference |
| Architecture | Complex control logic | SIMD (Single Instruction, Multiple Data) | Specialized for tensor operations |
| Data Handling | Cache-focused, optimized for small data sets | High bandwidth for large data sets | Efficient memory access for tensors |
| Throughput | Moderate | High | Extremely high for specific tasks |
| Power Efficiency | Optimized for diverse workloads | High efficiency for parallel tasks | Very efficient for AI-specific workloads |
| Latency | Very low | Higher than CPU but optimized for throughput | Low, especially for inference tasks |
| Instruction Set | Rich and complex | Simplified for parallel tasks | Specialized instructions for tensor operations |
| Development Frameworks | Supported in most programming languages | CUDA, OpenCL, and many AI libraries | Optimized support in frameworks like TensorFlow and PyTorch |
| Cost | Generally higher per core for high-end CPUs | Cost-effective for parallel workloads | Typically found in high-end GPUs, adding to overall GPU cost |
| Scalability | Limited by core count | Highly scalable with multiple GPUs | Scales well with GPU architectures |
| Thermal Design | Focused on thermal efficiency | High power consumption, managed with cooling solutions | High performance but requires efficient thermal management |
| Error Handling | Advanced error detection and handling | Basic error handling, focused on throughput | Optimized for reliability in AI workloads |

# 2 High Performance Computing (HPC) Architectures


(a) SM Architecture for NVIDIA Hopper H100. [12]




(b) Intel Gaudi 2 Processor High-level Architecture. [19]




(c) AMD MI300 Series Node-level Architecture. [23]

Figure 4: High Performance Computing (HPC) Architectures

## 2.1 NVIDIA Hopper Architecture (H100 GPU) [Launch Year $-2022$]

Hopper is NVIDIA's latest generation of GPU architecture, specifically designed to accelerate AI and HPC workloads. The flagship product in this lineup is the NVIDIA H100 Tensor Core GPU, which offers significant advancements over the A100 architectures.

### 2.1.1 Architecture Implementation

The **NVIDIA GH100 GPU** is composed of multiple GPU processing clusters (GPCs), texture processing clusters (TPCs), streaming multiprocessors (SMs), L2 cache, and HBM3 memory controllers. The full implementation of the GH100 GPU includes [12],

1. 8 GPCs, 72 TPCs (9 TPCs/GPC), 2 SMs/TPC, 144 SMs per full GPU
2. 128 FP32 CUDA Cores per SM, 18432 FP32 CUDA Cores per full GPU
3. 4 $4^{th}$ Gen Tensor Cores per SM, 576 per full GPU
4. 6 HBM3 or HBM2e stacks, 12 512-bit memory controllers
5. 60 MB L2 cache
6. Fourth-generation NVLink and PCIe Gen 5

### 2.1.2 Key Features

1. **Process Technology:** Built on TSMC's $4N$ process technology for improved performance and efficiency.
2. **Transistor Count:** Over $80$ billion transistors, significantly increasing computational power.

### 2.1.3 Performance Enhancements

1. **Tensor Cores:** Enhanced $4^{th}$ Gen Tensor Cores support new data types (e.g., **FP8**) for improved AI model training and inference.
2. **Streaming Multiprocessor (SM):** H100 SM quadruples the previous A100 peak per SM floating point computational power due to the introduction of FP8 precision computations, and doubles the A100 raw SM computational power on all previous Tensor Core, FP32, and FP64 data types, clock-for-clock.
3. **Transformer Engine:** A new Transformer Engine is designed specifically to accelerate transformer-based models (like LLMs). This engine utilizes a combination of floating-point and integer matrix operations to provide up to $6\times$ higher performance in transformer models compared to A100.
4. **Thread Block Clusters:** H100 introduces a new thread block cluster architecture that exposes control of locality at a granularity larger than a single thread block on a single SM.
5. **Multi-Instance GPU (MIG):** Allows multiple users to share a single GPU, enhancing resource utilization and flexibility. Hopper GPUs partitioned into multiple instances, each acting as a separate GPU.
6. **Dynamic Scheduling:** Supports dynamic task scheduling to optimize GPU workloads and improve efficiency.

### 2.1.4 Memory Architecture

1. **High Bandwidth Memory (HBM3):** Up to $80$ **GB of HBM3 memory** with a bandwidth of over $3$ **TB/s**, facilitating large-scale data processing.
2. **Unified Memory:** Enhanced unified memory architecture for seamless memory access across different GPU components.

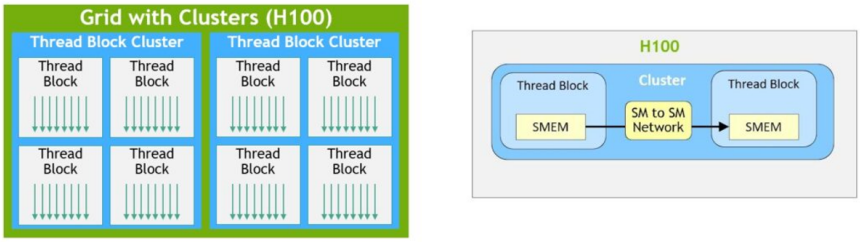



Figure 5: Block Cluster Architecture and Distributed Shared Memory Architecture for H100 Architecture.

#### 2.1.5 Software Ecosystem

1. **NVIDIA AI Software Stack:** Compatibility with frameworks like TensorFlow, PyTorch, and NVIDIA's own CUDA for optimized performance.
2. **DPX Instructions:** The H100 introduces DPX instructions to accelerate the performance of Dynamic Programming (DP) algorithms by up to $7\times$ compared to NVIDIA Ampere GPUs. These new instructions provide support for advanced fused operands for the inner loop of many DP algorithms.
3. **Support for NVIDIA Omniverse:** Enables collaborative 3D content creation and simulation.

#### 2.1.6 Connectivity

1. **NVLink:** Support for NVLink interconnect technology to enable high-speed communication (upto 900 GB/s GPU-to-GPU bandwidth) between multiple GPUs.
2. **PCIe Gen** 5**:** Faster data transfer rates with support for PCIe Gen 5 interfaces.

### 2.2 Intel Gaudi2 (AI Accelerator) [Launch Year $-2023$]

Intel Gaudi2, developed by Habana Labs owned by Intel, is an AI training processor designed to accelerate deep learning training workloads. It's part of Intel's growing portfolio of AI-focused hardware aimed at providing alternatives to GPUs for AI training. It offers AI training at lower costs due to its efficient use of power and silicon area. Its design focuses on high throughput and scalability, targeting data centers running large-scale AI models.
The Intel Gaudi accelerator architecture includes three main subsystems – compute, memory, and networking – and is designed from the ground up for accelerating Deep Learning training workloads. The compute architecture is heterogeneous and includes two compute engines – a Matrix Multiplication Engine (MME) and a fully programmable Tensor Processor Core (TPC) cluster.

#### 2.2.1 Architecture Implementation

The Intel Gaudi2 AI Accelerator comprises multiple processing clusters and memory units designed to optimize deep learning tasks and model training. The architecture includes the following features [16],

1. 32 Tensor Processor Cores (TPCs), each containing 2 Matrix Multiply Units (MMUs).
2. 8 HBM2e stacks with 2.45 TB/s total memory bandwidth.
3. 24 MB L2 cache for efficient data access and reduced latency.
4. PCIe Gen 4 for high-speed data transfer between devices.
5. Integrated Ethernet Networking with 24x 100GbE RDMA for fast communication and reduced bottlenecks.
6. Support for BF16 (Brain Floating Point), FP32, FP16, and INT8 operations for versatile data handling.

#### 2.2.2 Key Features

1. **Optimized for AI Workloads:** Intel Gaudi2 is specifically designed to handle the demands of AI and machine learning tasks.
2. **High-Performance Compute:** Intel Gaudi2 have enhanced computational capabilities tailored for training large neural networks. [19]

#### 2.2.3 Performance Enhancements

1. **Increased Core Count:** Features a higher number of compute cores compared to its predecessor, improving parallel processing.
2. **Matrix Multiplication Acceleration:** Optimized for matrix operations, critical for deep learning tasks.

#### 2.2.4 Memory Architecture

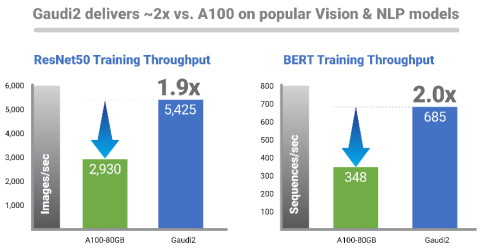
1. **High Bandwidth Memory (HBM):** Integration of HBM2e technology for improved memory bandwidth and reduced latency.
2. **Unified Memory Architecture:** Allows seamless access to memory resources for both CPU and GPU, enhancing efficiency.

### 2.2.5 Software Ecosystem

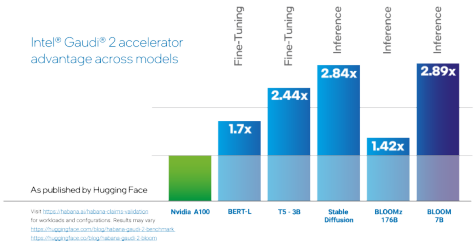
1. **Intel OneAPI Support:** Compatibility with Intel's OneAPI, providing a unified programming model for diverse workloads.
2. **Framework Support:** Works with popular AI frameworks such as TensorFlow, PyTorch, and MXNet for optimized performance.

### 2.2.6 Connectivity

1. **Integrated Networking:** Gaudi2 has integrated RDMA over Converged Ethernet (RoCE v2) engines on-chip. It offers 2.4 Terabits of networking bandwidth with the native integration on-chip of $24 \times 100$ Gbps RoCE V2 RDMA NICs, which enable inter-Gaudi communication via direct routing or via standard Ethernet switching [17].
2. **PCIe Gen 5:** Intel Gaudi2 features PCIe Gen 5 for high-speed data transfer, enhancing overall system performance.
3. **Multi-GPU Scalability:** Designed for scaling across multiple Gaudi2 accelerators in a system to tackle larger AI models.




(a) Throughput comparison of NVIDIA A100 and Intel Gaudi2 based on Deep Learning models.




(b) Intel Gaudi2 (AI Accelerator) advantages across various models. [18] [Experiment conducted by HuggingFace]

Figure 6: Intel Gaudi2 Advantages and Comparisons.

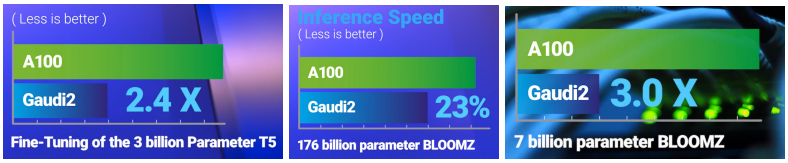



Figure 7: Comparison of NVIDIA A100 model and Intel Gaudi2 model with various parameters. [18]

## 2.3 AMD Instinct MI300 (CDNA 3 Architecture) [Launch Year $-2023$]

AMD's Instinct MI300 is the latest GPU from AMD, built on the CDNA 3 architecture and designed to accelerate AI and HPC workloads. This architecture represents a shift toward more AI-centric hardware design from AMD, competing directly with NVIDIA's Hopper and Intel's AI accelerators.

### 2.3.1 Architecture Implementation

The AMD Instinct MI300, based on the CDNA 3 architecture, combines GPU and CPU processing units in a single package, optimized for HPC and AI workloads. The key features include [21],

1. 8 Compute Die Stacks with 128 Compute Units (CUs) each, supporting FP64, FP32, BF16, and INT8 operations
2. 8192 Stream Processors across all compute units
3. Unified HBM3 memory with up to 128 GB capacity and 5.2 TB/s memory bandwidth
4. Infinity Fabric Link interconnect technology for seamless scaling across multiple GPUs
5. 2nd Generation Matrix Core Technology for high-performance AI acceleration
6. PCIe Gen 5 support for high-speed data transfers across devices
7. Chiplet-based 3D Packaging for integrated CPU and GPU die stack in a single module

### 2.3.2 Key Features

1. **Unified GPU and CPU Architecture:** Integrates CPU and GPU resources on a single chip for improved performance and efficiency.
2. **Advanced Chiplet Design:** Utilizes AMD's chiplet technology to enhance scalability and performance. There are 4 base IO die chiplets on the MI300. These four chiplets are linked together using a new AMD Infinity Fabric AP (Advanced Package) Interconnect. The MI300A is comprised of 13 chiplets ($3\times$ CCD, $6\times$ XCD, $4\times$ IO) while the MI300X uses 12 ($8\times$ XCD, $4\times$ IO). The MI300X have memory 192GB while HBM3 have memory 128GB.

### 2.3.3 Performance Enhancements

1. **Accelerator Complex Die (XCD):** XCD contains the GPU computational elements of the processor along with the lower levels of the cache hierarchy. The XCD has 40 CUs that include 38 active CUs at the aggregate level and 2 disabled CUs for yield management. The CUs all share a 4 MB L2 cache that serves to coalesce all memory traffic for the die. [20]

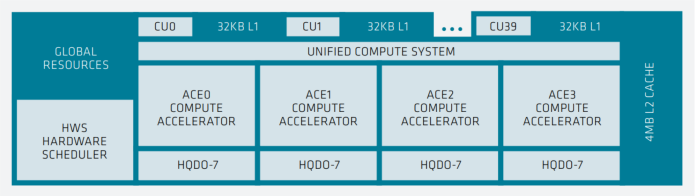



Figure 8: XCD-level system architecture of MI300. [23]

2. **Increased Compute Density:** Higher compute core count compared to previous generations, optimizing for parallel processing.
3. **Enhanced Tensor Performance:** Specialized for machine learning workloads with optimized tensor operations.

### 2.3.4 Memory Architecture

1. **High Bandwidth Memory (HBM3):** Supports HBM3 technology with increased memory bandwidth to accelerate data-intensive workloads [22]. The MI300 Series integrates up to 8 vertically stacked XCDs and 8 stacks of High-Bandwidth Memory 3 (HBM3).
2. **Large Memory Capacity:** Up to 128 GB of memory on a single GPU, enabling the handling of large datasets.

### 2.3.5 Software Ecosystem

1. **ROCm Support:** Compatible with AMD's ROCm (Radeon Open Compute) platform, providing a comprehensive programming model for HPC and AI.
2. **Framework Compatibility:** Works seamlessly with popular AI frameworks like TensorFlow, PyTorch, and others.

### 2.3.6 Connectivity

1. **PCIe Gen 5:** AMD CDNA 3 features support for PCIe Gen 5, enabling high-speed data transfer and reduced latency.
2. **Infinity Fabric:** AMD's proprietary interconnect technology for efficient communication between multiple GPUs and CPUs. The MI300 features 4 I/O dies (containing system infrastructure) with 256MB of AMD Infinity Cache. Infinity fabric has 128 total 32GB links [23].

# References

[1] Lee, Sunjung, Seunghwan Hwang, Michael Jaemin Kim, Jaewan Choi, and Jung Ho Ahn. "Future scaling of memory hierarchy for tensor cores and eliminating redundant shared memory traffic using inter-warp multicasting." IEEE Transactions on Computers 71, no. 12 (2022): 3115-3126.
[2] Hanindhito, Bagus, and Lizy K. John. "Accelerating ml workloads using gpu tensor cores: The good, the bad, and the ugly." In Proceedings of the 15th ACM/SPEC International Conference on Performance Engineering, pp. 178-189. 2024.
[3] San Juan, Pau, Pedro Alonso-Jordá, and Enrique S. Quintana-Ortí. "High performance and energy efficient integer matrix multiplication for deep learning." In 2021 29th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP), pp. 122-125. IEEE, 2021.
[4] Raschka, Sebastian. "Accelerating PyTorch Model Training." Sebastian Raschka's Magazine. Accessed October 27, 2024. https://magazine.sebastianraschka.com/p/accelerating-pytorch-model-training.
[5] DigitalOcean. "Understanding Tensor Cores." 2023. Accessed October 27, 2024.
[6] Evanson, Nick. "Explainer: What Are Tensor Cores? Mixed-Precision Computing." TechSpot, 2020. Accessed October 27, 2024.
[7] NVIDIA. "NVIDIA H100 Tensor Core GPU." Accessed October 27, 2024. https://www.nvidia.com/en-us/data-center/h100/.
[8] NVIDIA Corporation. "NVIDIA Ampere Architecture Whitepaper." 2020. Accessed October 27, 2024.
[9] NVIDIA. "Tensor Cores: Mixed Precision for AI and Scientific Computing." 2021. https://developer.nvidia.com/blog/tensor-cores-mixed-precision-scientific-computing/. Accessed October 27, 2024.
[10] NVIDIA. "NVIDIA H100 Tensor Core GPU." Accessed October 27, 2024. https://www.nvidia.com/en-in/data-center/h100/.
[11] NVIDIA. "Inside Volta." Accessed October 27, 2024. https://developer.nvidia.com/blog/inside-volta/.
[12] NVIDIA. "NVIDIA Hopper Architecture: In Depth." Accessed October 27, 2024. https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/.
[13] NVIDIA. "Tensor Cores." Accessed October 27, 2024. https://www.nvidia.com/en-in/data-center/tensor-cores/.
[14] NVIDIA. "NVIDIA Volta GPU Architecture." 2024. https://www.nvidia.com/en-in/data-center/volta-gpu-architecture/. Accessed October 27, 2024.
[15] Wevolver. "Tensor Cores vs. CUDA Cores: The Powerhouses of GPU Computing from NVIDIA." 2023. Accessed October 27, 2024.
[16] Intel. "Habana Gaudi2 Processor for Deep Learning." Accessed October 27, 2024. https://www.intel.com/content/www/us/en/developer/articles/ technical/habana-gaudi2-processor-for-deep-learning.html.
[17] Intel. "Gaudi Overview." Accessed October 27, 2024. https://www.intel.com/content/www/us/en/ products/details/processors/ai-accelerators/gaudi-overview.html
[18] Intel. "Gaudi2." Accessed October 27, 2024. https://www.intel.com/content/www/us/en/products/ details/processors/ai-accelerators/gaudi2.html.
[19] Habana Labs. "Gaudi Architecture." 2024. https://docs.habana.ai/en/latest/Gaudi_Overview/ Gaudi_Architecture.html. Accessed October 27, 2024.
[20] AMD. "AMD Instinct MI300 Accelerator." Accessed October 27, 2024. https://www.amd.com/en/products/ accelerators/instinct/mi300.html.
[21] AMD. "AMD Instinct MI300 Series Accelerators." Accessed October 27, 2024. https://www.amd.com/en/products/accelerators/instinct/mi300.html
[22] AMD. "CDNA." 2024. https://www.amd.com/en/technologies/cdna.html. Accessed October 27, 2024.
[23] AMD. "MI300 Architecture." Accessed October 27, 2024. https://rocm.docs.amd.com/en/latest/conceptual/gpu-arch/mi300.html.