BHIKSHAPATHI KANIKI -24489(CDS) **ISS- ASSIGNMENT-2**

Question: 1. Find about tensor cores in modern-day GPUs/accelerators and describe the features of tensor cores that make them more suitable for AI computations than traditional CPU and GPU cores.

Introduction to Tensor Cores and Their Role in AI

Introduced by NVIDIA in its Volta architecture, **Tensor Cores** are specialized processing units specifically designed to accelerate matrix multiplications, a fundamental operation in deep learning. These cores have since evolved and are now integral to NVIDIA's Turing, Ampere, and Hopper architectures. Similar units have been developed by other companies, including AMD's **Matrix Cores**, Google's **TPU Matrix Multiply Units (MXUs)**, and Intel's **GPGPU tensor processors**.

A Tensor Core is engineered to handle large numbers of parallel computations efficiently, making it exceptionally effective for the complex mathematical operations essential to deep learning, such as:

- Matrix multiplications
- Convolutions
- Tensor transformations

These features enable Tensor Cores to significantly enhance the performance of deep learning and AI workloads, particularly in matrix-based operations central to these tasks.

Architectural Features of Tensor Cores

The architectural design of Tensor Cores includes unique features that make them particularly suited for deep learning and AI tasks:

- **Specialized Matrix Multiplication Units**: Tensor Cores perform highspeed matrix-matrix multiplications, using mixed-precision (e.g., FP16 or FP8 for inputs, FP32 for accumulation) to balance accuracy with speed.
- **Mixed Precision and Lower Precision Support**: Tensor Cores support formats like FP16, BF16, and FP8, which reduce memory requirements and processing time without sacrificing the model's effectiveness. For example, NVIDIA's Hopper Tensor Cores add support for FP8, enabling higher throughput while preserving accuracy in training and inference.

Key Features that Make Tensor Cores Suitable for AI Computations

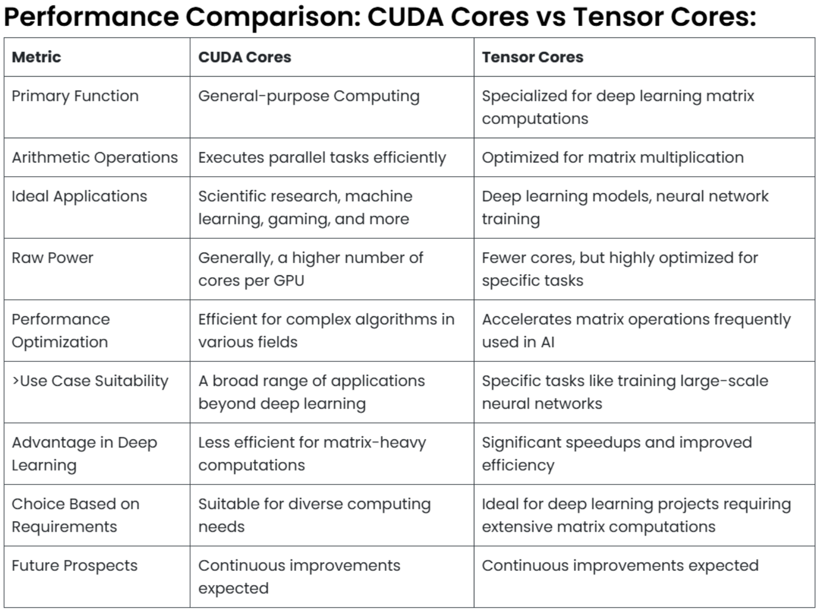
NVIDIA's Tensor Cores and similar architectures provide a range of features optimized for AI computations:

- **High Throughput with Matrix Multiplications**: Tensor Cores can execute high-throughput matrix operations essential for training neural networks. For example, an NVIDIA A100 GPU with third-generation Tensor Cores can achieve up to 312 teraflops in mixed precision, significantly improving computational efficiency.
- **Parallelism and Efficient Workload Distribution**: Tensor Cores process multiple operations in parallel, distributing workloads across thousands of cores within a GPU. This high degree of parallelism reduces latency, enabling faster model training and inference.
- **Sparsity and Zero-Skipping Techniques**: Tensor Cores in the Ampere and Hopper architectures support structured sparsity, which skips zero values in matrices, doubling performance for models with sparse data.

- **Optimized for AI Data Types**: Tensor Cores support AI-specific data types, such as FP16, BF16, INT8, and FP8, which are critical for efficiently handling the vast amounts of data in AI models.
- **Advanced Instructions for Dynamic Workloads**: Some Tensor Cores, such as those in NVIDIA's Hopper architecture, support DPX (Dynamic Programming Acceleration) instructions for complex AI algorithms, extending Tensor Core utility beyond matrix multiplications to real-time environments.

Comparison with Traditional CPU and GPU Cores in AI Computations

- **CPU Limitations in AI**: CPUs handle tasks serially, limiting their scalability for deep learning models. Their smaller core counts and limited memory bandwidth make CPUs less suitable for matrix-heavy AI tasks.
- **GPU Cores vs. Tensor Cores**: While traditional GPU cores improve over CPUs with parallel processing, they lack the specialized architecture of Tensor Cores. Standard GPU cores often operate in FP32, reducing throughput compared to Tensor Cores' mixed-precision and sparsity handling capabilities, making Tensor Cores faster and more power-efficient



Performance Comparison: CUDA Cores vs Tensor Cores:

Metric	CUDA Cores	Tensor Cores
Primary Function	General-purpose Computing	Specialized for deep learning matrix computations
Arithmetic Operations	Executes parallel tasks efficiently	Optimized for matrix multiplication
Ideal Applications	Scientific research, machine learning, gaming, and more	Deep learning models, neural network training
Raw Power	Generally, a higher number of cores per GPU	Fewer cores, but highly optimized for specific tasks
Performance Optimization	Efficient for complex algorithms in various fields	Accelerates matrix operations frequently used in AI
>Use Case Suitability	A broad range of applications beyond deep learning	Specific tasks like training large-scale neural networks
Advantage in Deep Learning	Less efficient for matrix-heavy computations	Significant speedups and improved efficiency
Choice Based on Requirements	Suitable for diverse computing needs	Ideal for deep learning projects requiring extensive matrix computations
Future Prospects	Continuous improvements expected	Continuous improvements expected

---

Question: 2. Explore the latest state-of-art HPC-based architectures tuned for AI from three architecture leaders, namely, NVIDIA, Intel and AMD. Give detailed description of at least three such architectures (e.g., DGX-H100 from NVIDIA), particularly focusing on the architecture features that help in AI computations.

Introduction to NVIDIA DGX H100 and the Hopper Architecture

The NVIDIA DGX H100 system represents a cutting-edge AI platform powered by the NVIDIA Hopper architecture. Featuring the H100 GPU, the system targets high-performance applications in data centers, AI, and HPC (High-Performance Computing). Built on TSMC's 4N process with over 80 billion transistors, the H100 GPU advances beyond the A100, delivering new efficiencies in speed, compute power, and scalability.

The DGX H100 excels in tasks such as deep learning (DL), machine learning (ML), and AI applications, supporting diverse fields like natural language processing (NLP) and scientific computing.

Detailed Architecture of the H100 GPU

The Hopper architecture offers enhanced performance across multiple dimensions:

- **Streaming Multiprocessor (SM) Enhancements**: The new H100 SM architecture quadruples A100's FP8 throughput, integrates an enhanced shared memory and L1 cache, and introduces efficient asynchronous execution. The asynchronous execution feature allows data transfers to be conducted independently, boosting overall SM throughput.

- **FP8 Data Format**: A major innovation in the Hopper architecture is the 8-bit floating point (FP8) format, a compact data type that retains performance and precision while improving data throughput. By supporting FP8 alongside FP16, BF16, and INT8, the H100 achieves double the performance of FP16 while reducing data storage requirements—ideal for large-scale AI models such as LLMs.

- **Transformer Engine**: Another unique aspect is the Transformer Engine, which dynamically adjusts between FP8 and FP16 for efficient precision management. This helps reduce processing time significantly when running models such as GPT-3 and BERT, enabling up to 9x faster training and 30x faster inference compared to the A100.

Key Tensor Core Features for Accelerated AI

NVIDIA's DGX H100 integrates fourth-generation Tensor Cores with specialized features to accelerate AI computations:

- **High Throughput with Mixed Precision**: The H100 Tensor Cores support a wide range of data types, from FP8 to FP64, enabling flexible, high-speed mixedprecision operations. This design optimizes for AI-specific tasks, resulting in up to 6x the throughput of previous GPU generations.

- **Sparsity Support**: By leveraging sparsity—skipping computations for zero values—Tensor Cores can double processing efficiency for dense matrix operations. This technique enhances deep learning performance without compromising accuracy.

- **DPX (Dynamic Programming Acceleration) Instructions**: The H100 introduces DPX instructions, accelerating dynamic programming algorithms by up to 7x. This feature is particularly useful for recursive models and applications in genomics and real-time robotics, extending Tensor Core utility beyond traditional deep learning.

Software Ecosystem for the DGX H100

The DGX H100 system is integrated into NVIDIA's broader software ecosystem, which provides a suite of tools, libraries, and models optimized for AI and HPC:

- **CUDA and AI SDK**: The DGX H100 is fully compatible with the CUDA platform, offering libraries like cuDNN for deep neural networks and TensorRT for model optimization. CUDA's broad compatibility with major frameworks (TensorFlow, PyTorch, etc.) simplifies the development and deployment of AI solutions [**?**].

- **NVIDIA NGC Catalog**: NVIDIA's NGC platform provides a repository of pretrained models, model scripts, and containers. This catalog enables data centers to quickly adopt and scale AI applications across various industries. For example, NGC containers are pre-configured for NLP, computer vision, and reinforcement learning, reducing setup times and streamlining development workflows [**?**].

- **Distributed Training with NVLink and NVSwitch**: The DGX H100 supports distributed training across multiple GPUs with NVLink, an interconnect technology that enables high-bandwidth GPU-to-GPU communication. NVIDIA's NVLink Switch System connects up to 256 GPUs, allowing data centers to build massive AI supercomputers for handling complex model training [**?**][**?**].

Real-World Applications and Performance Benefits

The DGX H100 excels in various applications:

1. **Natural Language Processing (NLP)**: The Transformer Engine and FP8 tensor cores enable rapid training for large NLP models, improving real-time applications like language translation and text generation.

2. **Genomics and Scientific Research**: DPX instructions support accelerated genomics algorithms, such as the Smith-Waterman algorithm, enhancing sequence analysis for disease research.

3. **Autonomous Systems and Robotics**: Enhanced Tensor Cores support real-time decision-making in autonomous vehicles, drones, and industrial robots, providing reliable performance in safety-critical applications.

Introduction: A Detailed Study on Gaudi2

The Gaudi2 architecture, developed by Habana Labs (an Intel company), is one of the most advanced hardware accelerators built for deep learning. Gaudi2 is specifically designed to support AI workloads, offering high-performance and cost-effective alternatives to traditional GPU-based architectures. With features such as specialized compute engines, advanced memory systems, and an Ethernet-based networking design, Gaudi2 provides substantial improvements in both performance and scalability for large-scale AI tasks.

Gaudi2 Chip Architecture

The Gaudi2 chip architecture is purpose-built for deep learning and optimized across three primary subsystems: compute, memory, and networking. Each subsystem contributes to the chip's overall performance, efficiency, and scalability in training complex AI models.

- **Compute Engines**: Gaudi2 has a heterogeneous compute architecture with two types of engines: the Matrix Multiplication Engine (MME) and the Tensor Processor Core (TPC) cluster. The **MME** efficiently handles matrix multiplications like fully connected layers and convolutions with up to 2.4 Tbps bandwidth. The **TPC cluster** acts as a programmable SIMD processor, managing non-GEMM operations essential for deep learning training.

- **Memory Architecture**: Gaudi2 includes 96GB of HBM2E memory with 2.45 TB/sec bandwidth and 48MB of SRAM for rapid data access. This setup allows the MME and TPC engines to avoid memory bottlenecks, supporting parallel processing for large AI workloads. A dedicated transpose engine efficiently manages data format alignment.

- **Networking**: Gaudi2 integrates RDMA NICs with 2.4 Tbps bandwidth across 24 x 100 Gbps Ethernet ports, enabling direct inter-Gaudi communication for scalability in multi-node configurations. This Ethernet-based approach is cost-effective and avoids vendor lock-in.

Parallel Computation and Overlapping Workloads

Gaudi2's defining feature is its ability to handle overlapping workloads between its MME and TPC compute engines, enabling simultaneous task execution. This capability is crucial for AI models, where GEMM and non-GEMM operations often run sequentially on GPUs.

- **Overlapping Computation**: Gaudi2 allows the MME and TPC to execute tasks in parallel. While the MME processes matrix-heavy tasks, the TPC can handle elementwise transformations simultaneously, reducing compute time and enhancing throughput.

- **Comparison with GPUs**: Traditional GPUs require large matrix dimensions for optimal performance. In contrast, Gaudi2 maintains high utilization with smaller matrices (as small as 1,000x1,000), allowing support for a wider range of models and smaller tensors without underutilization.

Scalability of Gaudi2

Gaudi2's scalability is a significant advantage for AI workloads, allowing expansion from single-node setups to large clusters to meet growing computational demands.

- **Rack-Scale Configurations**: The Gaudi2 architecture supports standard interfaces, enabling scalability from individual nodes to racks and larger clusters. Its Ethernetbased communication structure enhances scalability by allowing flexible interconnections across multiple racks without the need for proprietary network interfaces. This setup is ideal for data centers aiming to integrate AI hardware at scale without vendor lock-in.

- **DDN A3I Reference Architecture**: Gaudi2 integrates seamlessly with DDN AI storage solutions, like the AI400X2, to support data-intensive AI workloads. The DDN A3I (Accelerated, Any-Scale AI) architecture combines DDN's storage capacity with Supermicro's Gaudi2 AI servers, delivering predictable performance and capacity for AI training. This configuration enables Gaudi2 systems to achieve peak performance in both single-server and multi-server environments, simplifying deployment for enterprises with high data throughput needs.

Introduction to AMD CDNA 3 Architecture

The AMD CDNA 3 architecture, part of the AMD Instinct MI300 Series, represents a transformative step in computational acceleration, targeting workloads in AI, machine learning (ML), and HPC. CDNA 3 builds on the previous CDNA 2 architecture with revolutionary features, including a 3D chiplet-based design, the integration of Infinity Fabric technology, and enhanced matrix processing capabilities.

These advancements make the MI300 series highly efficient, allowing data centers to achieve exceptional computational power and scalability. For instance, the MI300X GPU focuses on high-throughput matrix computations, while the MI300A APU model provides a unified memory pool for CPU-GPU collaboration. This architecture is ideal for large-scale AI and deep learning models, which demand high data processing speeds and efficiency.

Detailed Chiplet-Based Architecture

The AMD CDNA 3 architecture uses a 3D chiplet-based design, shifting away from traditional monolithic architectures. Each MI300 processor integrates up to eight **Accelerator Complex Dies (XCDs)** and four **Input/Output Dies (IODs)**. XCDs serve as the computational cores, while IODs manage data routing and system I/O functions. These components are interconnected via **AMD Infinity Fabric**, a high-bandwidth, low-latency interconnect that ensures seamless data transfer across the processor.

Each XCD die contains 38 compute units (CUs), creating a dense configuration that maximizes performance per watt. The MI300 series scales effectively by allowing multiple XCDs to be combined within a processor, reaching up to 304 compute units per processor. By distributing compute, memory, and communication functions across specialized dies, AMD CDNA 3 optimizes workload distribution and scalability, making it suitable for standalone and multi-GPU configurations.

Memory Hierarchy and Infinity Cache

In CDNA 3, the memory hierarchy is re-engineered to support AI and HPC demands. The MI300 series introduces **256MB of AMD Infinity Cache** on the IOD, acting as the last-level cache (LLC) for all XCDs. This cache structure enhances memory bandwidth and reduces latency by storing frequently accessed data, which is essential for high-speed matrix computations in ML applications.

The MI300X model includes **eight HBM3 (High-Bandwidth Memory) stacks**, offering a capacity of up to 192GB and a peak memory bandwidth of 5.3 TB/s, suitable for handling large AI models. The MI300A model further integrates a unified memory pool accessible by both CPU and GPU, improving data transfer rates and power efficiency. This design minimizes latency and simplifies programming by reducing the need for CPU-GPU memory transfers.

Matrix Core Features and Sparse Data Optimization

A key feature of CDNA 3 is the advanced **Matrix Cores**, which offer specialized processing for AI data types. These cores support a range of formats essential for ML, including **FP16, BF16, FP8, TF32, and INT8**. The inclusion of FP8 and TF32 formats enables efficient data handling with reduced precision, allowing the Matrix Cores to process large neural networks while minimizing computational overhead and power consumption.

CDNA 3 also supports **4:2 sparsity**, which enables the Matrix Cores to ignore zerovalue elements in data matrices. This sparsity feature doubles throughput for models with high sparsity, such as transformer models and other neural networks with a high proportion of zero values. With these features, AMD CDNA™ 3 achieves a peak performance of 2.6 petaFLOPS for FP8 operations, making it highly effective for both AI training and inference.

Advanced Communication and Scaling Features

The AMD CDNA 3 architecture scales efficiently due to the **Infinity Fabric** interconnect, which connects components within a processor and across multiple GPUs/APUs in a system. Each IOD die in CDNA 3 has **two**

bi-directional 16-lane Infinity Fabric links operating at up to 32 Gbps, allowing fast data exchange between XCDs and IODs. In an 8-GPU MI300X configuration, the Infinity Fabric provides 896 GB/s aggregate bandwidth, enabling efficient distributed training for large-scale AI applications.

The MI300 series can be configured to meet diverse workload needs. For example, the MI300X supports up to eight GPU partitions, making it ideal for multi-stream AI inference, while the MI300A APU supports unified memory, reducing the need for CPU-GPU data transfers. These flexible configurations allow data centers to optimize their setup for either high-throughput, low-latency training or scalable inference, enhancing the utility of the MI300 series in AI and HPC environments.

COMPARSION OF ALL PROCESSORS

Feature	**Intel Habana Gaudi2**	**AMD M1 CND**	**NVIDIA H100**
Architecture	Custom HPU, optimized for AI	Data center optimized GPU	Ampere Next-gen (Hopper) GPU
Core Specialization	AI accelerators with HBM2e	General-purpose with AI focus	Tensor Cores optimized for AI
Memory Bandwidth	1 TB/s with HBM2e	~768 GB/s	1 TB/s with HBM3
Networking Bandwidth	800 GB/s	Limited details on bandwidth	900 GB/s NVLink
Precision Support	FP32, BFloat16	FP32, BFloat16	FP32, BFloat16, INT8, FP8
Power Efficiency	High, with AI optimization	Balanced	Efficient but power-demanding
Performance	~1.4x faster in vision-language models compared to H100 in some tests	AI-enhanced performance	Industry-leading in FP8 and INT8 performance
Ecosystem	Optimum Habana forPyTorch, TensorFlow	ROCm for PyTorch and other libraries	CUDA and cuDNN, widely adopted
Cost Efficiency	Competitive, optimized for TCO	TCO-focused	Premium tier

References:

[1] AMD. AMD CDNA 3 White Paper. Accessed from AMD’s official resources.

[2] Habana Labs. Gaudi2 White Paper. Intel Corporation, 2022. Available through Habana Labs documentation.

[3] NVIDIA. GTC 2022 White Paper: Hopper Architecture. Presented at NVIDIA GTC, 2022. Retrieved from NVIDIA’s official publications.