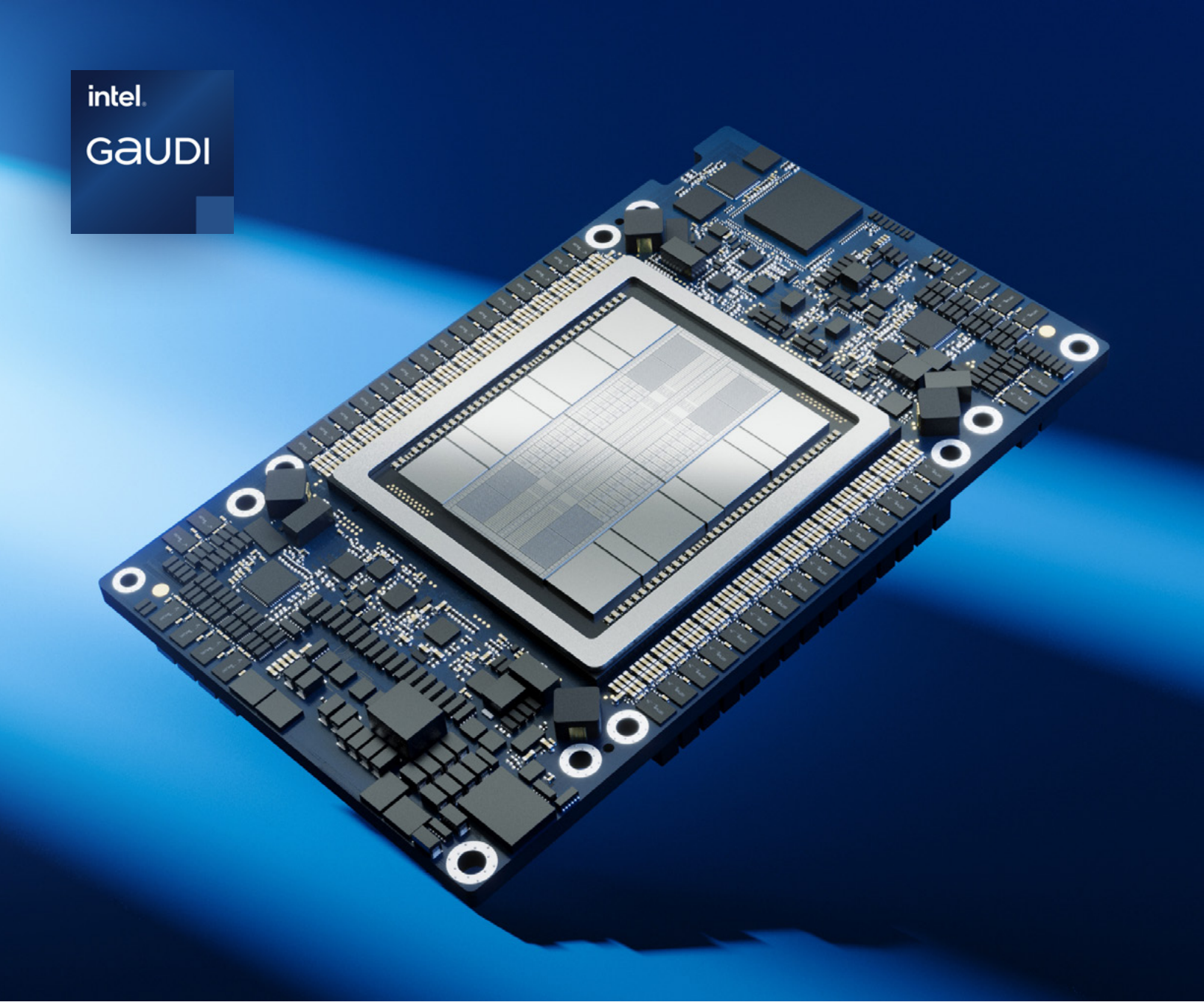


Technical Paper

Generative AI Compute
AI for Enterprise

Intel® Gaudi® 3 AI Accelerator

This technical paper introduces the next-generation AI accelerator from Intel®: The Intel® Gaudi® 3 AI Accelerator.

SEPTEMBER 2024 — Rev. 1

소개

딥 러닝 및 인공 지능 워크로드는 계속해서 더 높은 성능과 더 낮은 전력 소비를 요구합니다. 이 기술 문서에서는 인텔의 차세대 AI 가속기인 인텔® 가우디® 3 AI 가속기를 소개합니다. 새로운 가속기는 5세대 이기종 AI 가속 아키텍처를 특징으로 합니다.

인텔® 가우디® 3 AI 가속기는 대형 언어 모델 (LLM)과 스테이블 디퓨전 (Stable Diffusion 등 이미지 생성)과 같은 생성 응용 프로그램부터 표준 객체 인식, 분류 및 음성 더빙에 이르기까지 모든 AI 워크로드에 대해 최첨단 데이터 센터 성능을 제공하도록 설계되었습니다.

2022년에 도입된 인텔® 가우디® 2 AI 가속기는 PyTorch 프레임워크를 통합한 인텔® 가우디® 소프트웨어를 통해 지원됩니다. 인텔® 가우디® 3 AI 가속기는 한 차원 높은 AI 성능과 전력 효율성을 제공합니다. 인텔® 가우디® 2 가속기 7nm 프로세스에서 발전한 인텔® 가우디® 3 AI 가속기는 TSMC 5nm 프로세스로 제작되어 면적밀도와 전력 효율이 향상되었습니다.

인텔® 가우디® 3 AI 가속기는 성능과 전력 효율면에서 가능한 경계를 계속해서 허물고 있습니다. 인텔® 가우디® 2 가속기 아키텍처를 기반으로 구축된 인텔® 가우디® 3 AI 가속기는 컴퓨팅, 메모리 대역폭 및 아키텍처 효율성이 크게 향상되었습니다.

인텔® 가우디® 3 AI 가속기는 8개의 MME 엔진, 64개의 TPC 엔진 및 24개의 200Gbps RDMA NIC 포트를 포함하는 2개의 컴퓨팅 다이를 갖추고 있습니다. 또한 총 8개의 HBM2e 칩은 128GB의 통합 고대역폭 메모리 (HBM)로 구성됩니다.

인텔® 가우디® 3 AI 가속기는 훈련 및 추론에 있어서 FP8 및 BF16 연산에서 1.8 PFlops를 지원하고, 128 GB의 HBM2e 메모리 용량과 3.7 TB/s의 HBM 대역폭을 지원하는 탁월한 성능을 제공합니다.

Table of Contents

2x
FP8 GEMM FLOPs

4x
BF16 GEMM FLOPs

1.5x
Faster HBM Bandwidth

1.33x
Larger HBM Capacity

가우디® 3 AI 가속기 개요

AI 애플리케이션은 점점 더 빠르고 에너지 효율적인 하드웨어 솔루션을 요구하고 있으며 인텔® 가우디® 3 가속기는 이러한 요구에 부응하도록 설계되었습니다. 인텔® 가우디® 2 가속기에 비해 2배 이상의 FP8 GEMM FLOPs와 4배 이상의 BF16 GEMM FLOPs의 성능을 보유한 인텔® 가우디® 3 AI 가속기는 최첨단 AI 학습 성능을 제공합니다. 1.5배 더 빠른 HBM 대역폭과 1.33배 더 큰 HBM 용량을 갖춘 인텔® 가우디® 3 AI 가속기는 인텔® 가우디® 2 가속기에 비해 대규모 언어 모델 추론 성능이 크게 향상되었습니다.

인텔® 가우디® 3 AI 가속기 (그림 1)는 고대역폭과 낮은 지연시간을 제공하는 인터포저 브리지로 연결되는 2개의 동일한 컴퓨팅 다이를 가지고 있습니다. 다이와 다이간의 연결은 소프트웨어에 투명하여 통합된 하나의 대형 다이와 같은 성능과 동작을 제공합니다.

인텔® 가우디® 3 AI 가속기 컴퓨팅 아키텍처는 이기종이며 다음 두 가지 주요 컴퓨팅 엔진을 포함합니다: MME (Matrix Multiplication Engine)와 완전히 프래그래머블한 TPC (Tensor Processor Core) 클러스터. MME는 완전 연결된 레이어, 컨볼루션 및 일괄 처리 GEMM과 같이 매트릭스 곱셈으로 할 수 있는 모든 연산을 담당합니다. 딥러닝 애플리케이션을 위해 맞춤 제작된 VLIW (Very Long Instruction Word) 단일 명령 다중 데이터 (SIMD) 프로세서인 TPC는 모든 GEMM 이외의 작업을 가속화하는 데 사용됩니다.

인텔® 가우디® 가속기 제품 라인업

인텔® 가우디® 2와 인텔® 가우디® 3 AI 가속기의 기능 비교

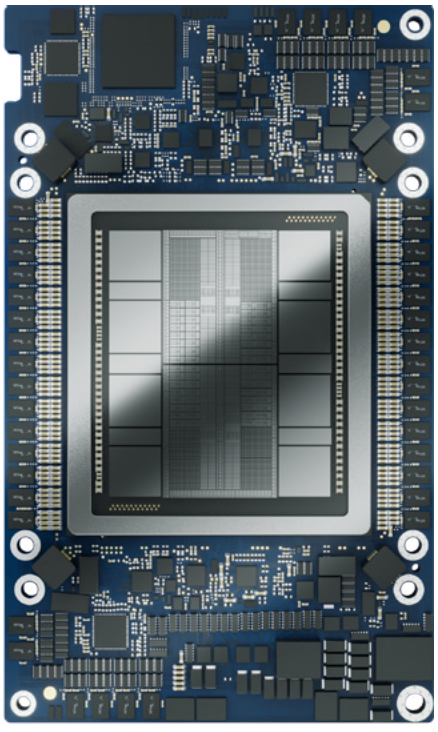
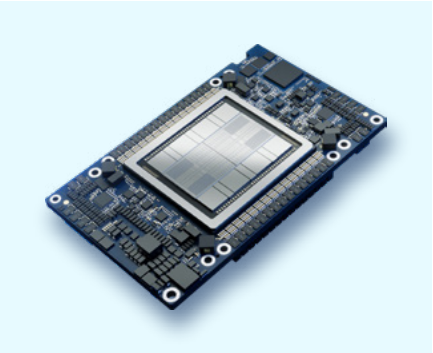


그림 1. 인텔® 가우디® 3 OAM 모듈

Feature/Product	Intel® Gaudi® 2 AI Accelerator	Intel® Gaudi® 3 AI Accelerator
BF16 MME TFLOPS	432	1835
FP8 MME TFLOPS	865	1835
BF16 Vector TFLOPS	11	28,7
MME Units	2	8
TPC Units	24	64
HBM Capacity	96 GB	128 GB
HBM Bandwidth	2.46 TB/s	3.7 TB/s
On-die SRAM Capacity	48 MB	96 MB
On-die SRAM Bandwidth	6.4 TB/s	12.8 TB/s
Networking (bidirectional)	600 GB/s	1200 GB/s
Host Interface	PCIe Gen4 X16	PCIe Gen5 X16
Host Interface Peak BW	64 GB/s (32 GB/s per direction)	128 GB/s (64 GB/s per direction)
Media Decoders	8	14

표 1. 인텔® 가우디® 2와 인텔® 가우디® 3 AI 가속기

하드웨어 시스템



HL-325L OCP 가속기 모듈

HL-325L OCP 가속기 모듈

인텔® 가우디® 3 AI 가속기 OAM (OCP Accelerator Module) 카드는 표준 OCP OAM 2.0 메자닌 카드 형태로 시스템 설계자에게 제공되며 패시브 냉각 시 최대 900W TDP (Total Device Power), 액체 냉각 시 최대 1.2KW TDP를 지원합니다.

표 2는 주요 인터페이스입니다:

Interface	Description
Host Link	x16 PCIe Gen5
Networking:Card-to-Card & Scale-out	48 x 112 Gb/s PAM4 SerDes Links
JTAG	In-field CPLD programming and low-level ASIC debug
UART	Low level debug & BMC access
I2C Master	On/Off-board Peripherals
I2C Slave / SMBUS	BMC control and monitoring interface

표 2. HL-325L OCP 가속기 모듈의 주요 인터페이스

HLB-325L 유니버설 베이스보드

HLB-325L 유니버설 베이스보드

HLB-325 유니버설 베이스보드는 OCP (Open Compute Project)에서 영감을 받은 또 다른 제품으로, 인텔® 가우디® 3 AI 가속기가 시스템 설계를 단순화할 수 있도록 합니다.

HLB-325 보드는 각 카드의 21개 NIC을 사용하여 non-blocking (비차단), all-to-all (올투올) 구성으로 PCB에 패시브 하게 상호 연결된 8개의 인텔® 가우디® 3 AI 가속기 카드를 지원할 뿐만 아니라 모든 인텔® 가우디® 3 AI 가속기 카드 (3x8=24)에서 나머지 3개의 200GbE NIC를 6개의 온보드 OSFP800 커넥터로 라우팅 하여 확장할 수 있습니다.

베이스보드에는 HIB (호스트 인터페이스 보드)에 대한 표준 인터페이스/커넥터가 있어 시스템 설계자가 특정 요구 사항에 맞게 맞춤 설계할 수 있으며, 다양한 토폴로지 및 애플리케이션을 위해 CPU와 가속기의 비율을 달리하여 원하는 시스템을 유연하게 구축할 수 있습니다.

블록 다이어그램 및 주요 구성 요소

- HLB-325 has the following main components:
- 8 X dual B2B connectors for the HL-325 Mezzanine boards
- High speed connectors for x16 PCIe interconnect to HIB
- 2 Complex Programmable Logic Devices
- Power and reset control
- JTAG distribution to the mezzanines
- LED indications
- 6x OSFP connectors (6x800G using 112G PAM 8 SerDes)
- 3x PHY retimers
- 8x PCIe retimers
- USB connectors for Debug

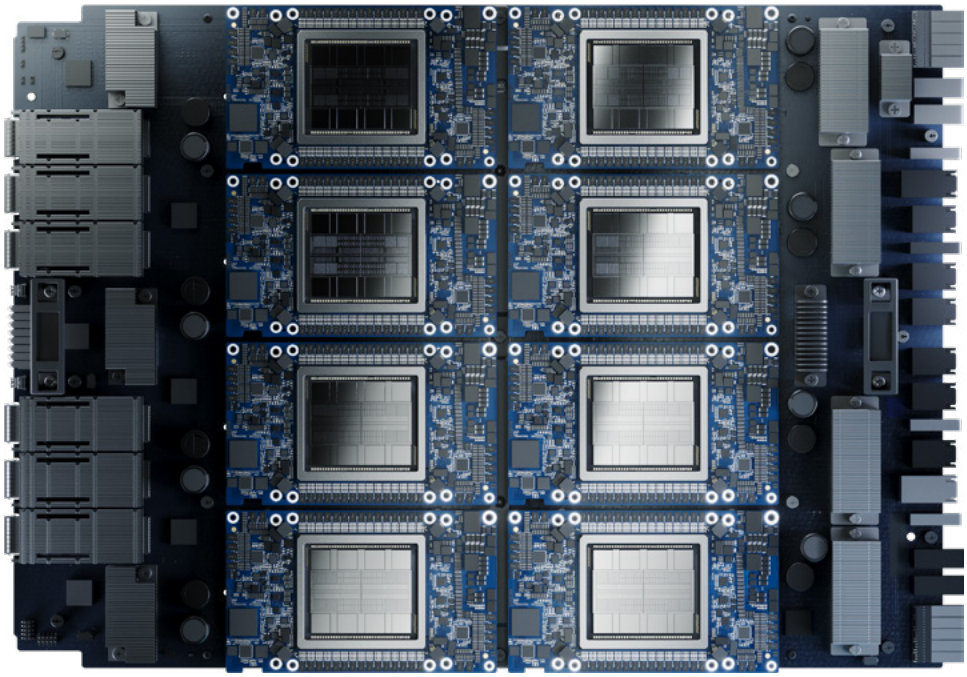


그림 2. HLB-325의 주요 구성 요소

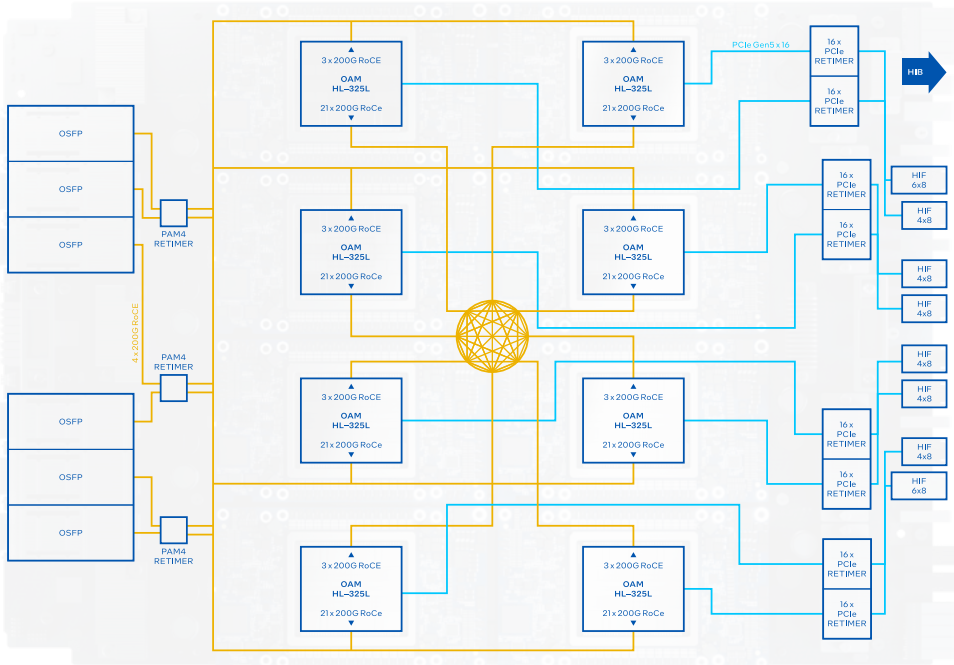
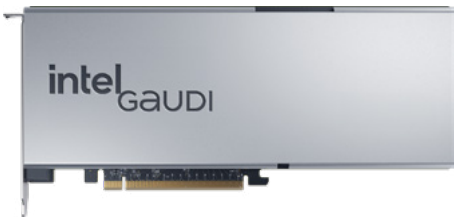


그림 3. HLB-325의 고속 블록 다이어그램

Feature	Description
OAM support	▪ OAM powered by 54V,12V and 3.3V ▪ Dual B2B connectors ▪ x16 PCIe Gen5 host interface per OAM ▪ 24x 200 GbE RoCE for scaleup and scale-out, via 48 112G PAM4 Serdes
Baseboard to HIB (Host Interface Board) Interface	▪ 8 X 16 PCIe Gen 5 connectors ▪ Power: 12V_Standby, 54V ▪ Side band signals: I2C, Reset, reference clocks, JTAG, UART, SGMII, USB ▪ Eight Amphenol connectors: 2x 160P (10131762-301LF) + 6 x 112P (10137002-101LF)
Networking: Card to Card & Scale-out	▪ Per OAM: 24x 200 GbE (48 112 Ghz PAM4 SerDes Links) split into: ▪ 21 x 200 GbE for OAM-to-OAM connections ▪ 3 x 200 GbE for scale-out ▪ Total Baseboard Scale-out: ▪ 8 x 3 x 200 GbE = 4.8 TbE connected to 6 OSFP800 ports
PCB dimension	▪ 585 mm x 417 mm x 4.6 mm

표 3. HLB-325의 특징



HL-338 PCIe 애드인 카드

HL-338 PCIe 애드인 카드

인텔® 가우디® 3 AI 가속기 PCIe 애드인 카드는 PCIe CEM 사양에 따라 시스템 설계자에게 제공됩니다. 리비전 5.1은 최대 600W TDP 패시브 냉각을 통한 전력 공급을 지원합니다.

표 4는 HL-338의 주요 인터페이스입니다:

Interface	Description
Host Link	x16 PCIe Gen5
Networking: ▪ Card-to-Card ▪ Scale-out	▪ 48 x 112 Gb/s PAM4 SerDes Links ▪ 2 x 400G QSFP112 ports
JTAG	In-field CPLD programming and low-level ASIC debug
I2C Slave / SMBUS	BMC control and monitoring interface

표 4. HLB-338 PCIe의 주요 인터페이스

4장의 HLTB-304 탑보드

HLTB-304 보드는 4장의 HL-338 카드와 6개의 200GbE 링크를 각각 연결할 수 있습니다. HL-338 카드와 나머지 3개의 HL-338 카드는 카드당 총 18개의 200GbE 링크로 연결됩니다.

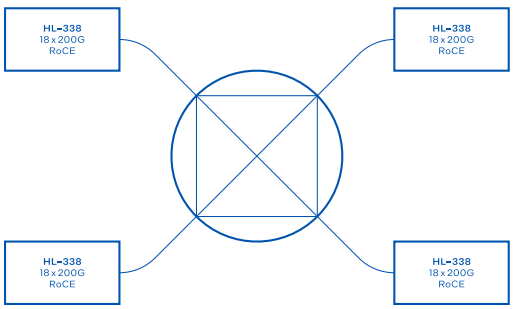


그림 4. HLTB-304 블록 다이어그램

인텔® 가우디® 3 AI 가속기 아키텍처

이기종 엔진의 병렬 실행

인텔® 가우디® 아키텍처는 모든 엔진을 병렬로 활성화할 수 있도록 설계되었습니다. 이것은 MME, TPC 및 NIC가 모두 동시에 동작할 수 있다는 것을 의미합니다.

서로 다른 엔진을 병렬로 실행하는 두 가지 주요 사용 사례는 다음과 같습니다 :

1. 엔진 유형의 입력과 출력 간에 종속성이 없는 경우.
   이 경우 특별한 소프트웨어 개입이 필요하지 않습니다. 그래프 컴파일러 (Graph Compiler)는 간단히 각 엔진이 실행되도록 트리거하여 전체 입력 및 출력 텐서의 크기를 제공할 수 있습니다.

2. 서로 다른 엔진에서 실행되는 연산 간에 종속성이 있는 경우.
   한 엔진의 출력이 다른 엔진의 입력으로 사용됩니다.

첫 번째 경우는 간단하며 MME, TPC 및 NIC를 병렬로 실행하도록 예약할 수 있습니다. 한 엔진이 실행 중인 작업을 완료하면, 입력이 준비되는 즉시, 다음 작업을 시작하도록 예약할 수 있습니다.

두 번째 경우는 인텔® 가우디® 소프트웨어에서 수행하는 작업 규모 관리 외에도 세분화된 스케줄링이 필요하기 때문에 더 복잡합니다. 이 경우, 종속 엔진은 생산자-소비자 관계의 파이프라인 방식으로 실행되도록 예약됩니다. 엔진 스케줄링 및 전체 오케스트레이션은 그래프 컴파일러가 수행합니다. 효율적인 엔진 스케줄링 및 실행을 위해 여러 소프트웨어 계층을 결합하여 함께 작동하는 방법에 대한 자세한 설명은 다음 섹션에 나와 있습니다.

그림 5는 인텔® 가우디® 3 AI 가속기의 전체 블록 다이어그램을 보여줍니다.

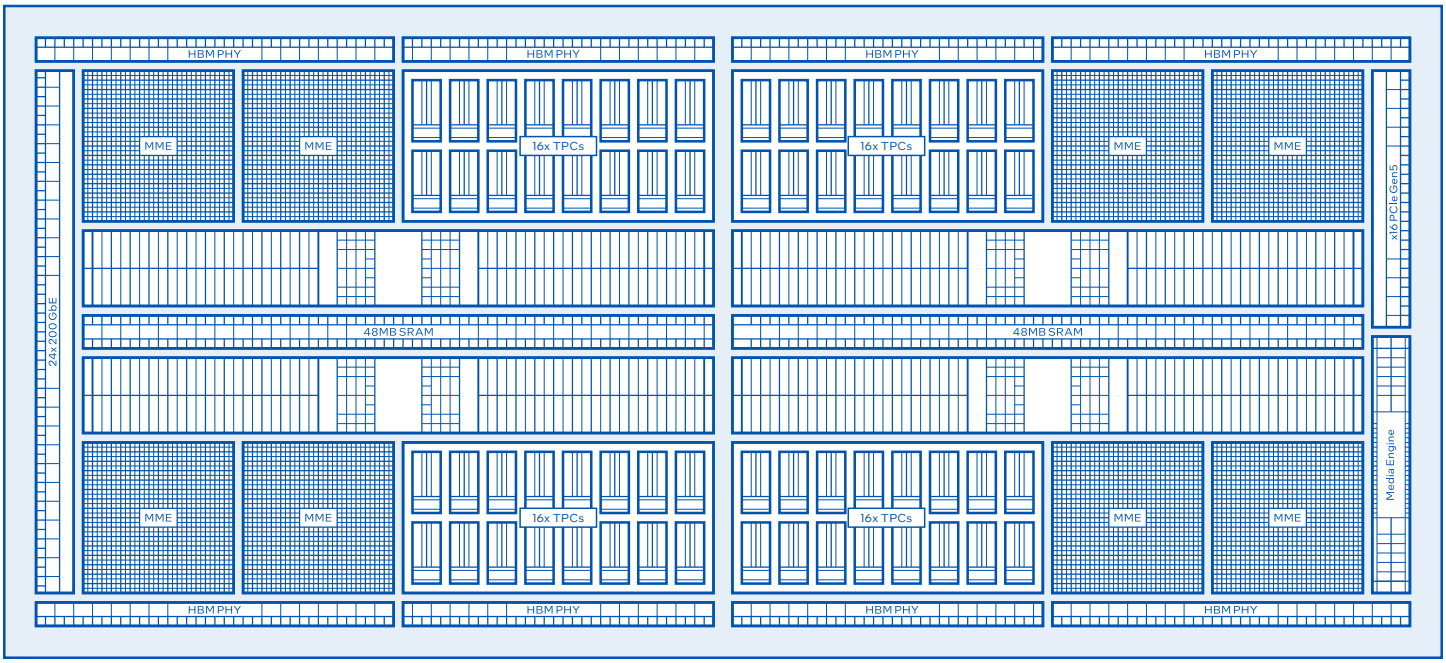
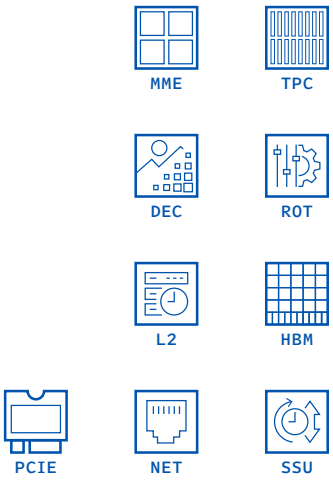


그림 5. 2 MME 및 16 TPC 유닛마다 L2 캐시가 있는 인텔® 가우디® 3 AI 가속기. L2의 일부는 그래프 컴파일러에서 공유 L3로 사용하도록 구성할 수 있습니다.

칩 각각의 구성 요소는 다음 장에서 자세히 설명합니다.

인텔® 가우디® 3 AI 가속기의 전체 구현에는 다음 장치들이 포함됩니다:



컴퓨트 엔진

- 8 Matrix Multiplication Engines (MMEs)
- 64 Tensor Processor Cores (TPCs)

미디어 엔진

- 14 Media Decoder Engines (DECs)
- 4 Rotator Engines (ROT)

메모리

- 96 MB of L2 Cache
- 128 GB of 8 HBM2e stacks

네트워킹

- PCIe Gen5 X16 port for communicating with host
- 24 Network ports and the accompanied RDMA Engine
- Scheduling and Synchronization Unit

물리적 파티션

인텔® 가우디® 3 AI 가속기 컴퓨팅 엔진은 4개의 클러스터로 나뉩니다.
각 클러스터는 DCORE (딥 러닝 코어)라고 하며 다음을 포함합니다:

- 2 Matrix Multiplication Engines (MMEs)
- 16 Tensor Processor Cores (TPCs)
- 24 MB of L2 Cache

그림 6은 DCORE 파티션, 미디어 서브 시스템, 네트워크 서브 시스템 및 호스트와의 연결을 통해 인텔® 가우디® 3 AI 가속기 아키텍처의 요소들을 보여줍니다.

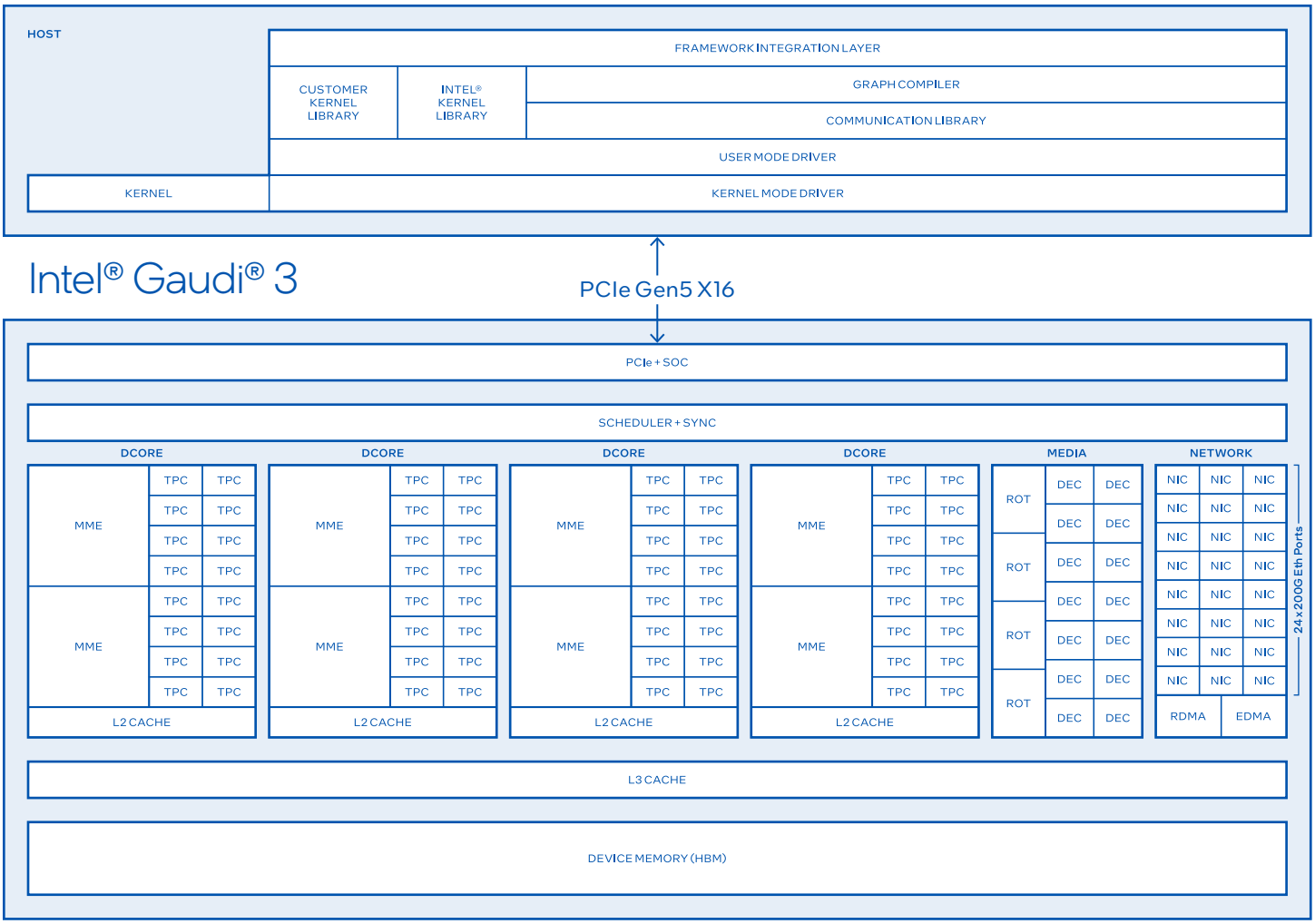


그림 6. DCORE 관점의 인텔® 가우디® 3 아키텍쳐와 지원 소프트웨어 레이어

호스트 인터페이스

인텔® 가우디® 3 PCIe 카드

인텔® 가우디® 3 AI 가속기에는 이전 세대 가속기에 탑재된 PCIe Gen 4에서 크게 업그레이드된 최첨단 PCI Express Gen 5 x 16 레인 인터페이스가 장착되어 있습니다. 이 고급 인터페이스는 총 128GB/sec의 대역폭을 제공하며, 각 방향에서 64GB/sec를 사용할 수 있습니다. 이는 PCIe Gen 4가 제공하는 총 대역폭 64GB/sec (각 방향 32GB/sec)에 비해 크게 개선된 것입니다.

PCIe Gen 5 인터페이스를 통해 인텔® 가우디® 3 AI 가속기를 시중에 출시된 가장 강력한 CPU, 외장 NIC 및 SSD와 원활하게 연결할 수 있습니다. 이를 통해 최적의 성능과 효율성을 보장하므로 고성능 컴퓨팅 솔루션을 위한 최고의 선택입니다.

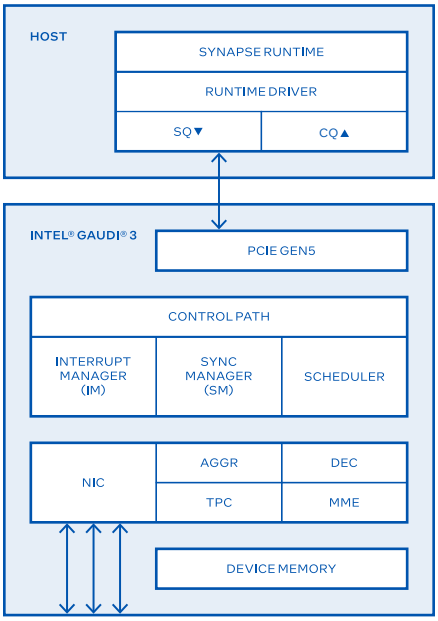


그림 7. 컨트롤 패스 블록 다이어그램

인텔® 가우디® 3 컨트롤 패스 (제어 경로)

다양한 엔진의 병렬 및 효율적인 실행을 관리하기 위해 인텔® 가우디® 3 AI 가속기에는 프로그래밍이 가능한 컨트롤 패스 엔티티 (Control Path Entity)가 통합되어 있습니다. 이 엔티티는 높은 처리량과 짧은 지연 시간을 위해 설계되었습니다. 그림 7은 이 기능의 주요 구성 요소를 제공합니다.

가우디® 3의 제어 경로는 다음과 같은 요소로 구성됩니다

- **제출 대기열** (Submission Queues - SQ) : 이 대기열은 런타임 시스템에서 발급합니다.
- **완료 대기열** (Completion Queues - CQ) : 작업 완료 보고에 사용됩니다.
- **프로그래밍 가능한 스케줄링 메커니즘** : 이 메커니즘은 작업 스케줄링에 활용됩니다.
- **프로그래밍 가능한 하드웨어 동기화 메커니즘** : 이 메커니즘은 다이어그램에서 '싱크 관리자 (Sync Manager - SM)' 라고 불리며 하드웨어 동기화에 사용됩니다.
- **프로그래밍 가능 인터럽트 서비스 메커니즘** : 다이어그램에서 '인터럽트 매니저 (INTR)' 라고 불리는 이 메커니즘을 통해 하바나 드라이버에 비동기 이벤트를 전달할 수 있습니다.

이러한 각 구성 요소는 인텔® 가우디® 3 AI 가속기 엔진의 원활하고 효율적인 작동을 보장하는 데 중요한 역할을 합니다.

다양한 엔진의 병렬 및 효율적인 실행을 제어하기 위해, 인텔® 가우디® 3 AI 가속기에는 프로그래밍이 가능한 저지연, 고 처리량 (Low Latency, High Throughput)의 제어 경로 엔티티 (Control Path Entity)가 포함되어 있습니다. 그림 7은 이 기능의 주요 구성 요소를 보여줍니다.

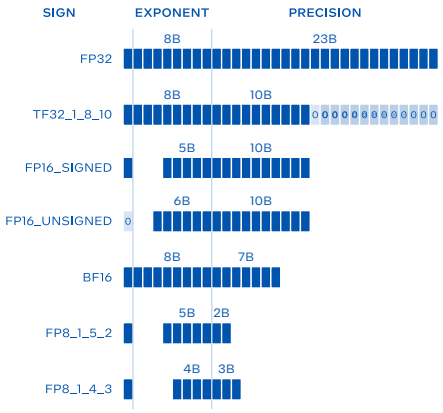


그림 8. 지원되는 부동 소수점 데이터 유형

컴퓨팅

그림 8은 인텔® 가우디® 3 AI 가속기 엔진이 지원하는 다양한 부동 소수점 데이터 유형을 보여주며, 그 다양성과 적응성을 보여줍니다.

표 5는 행렬 곱셈 (MME에서 수행)과 벡터 처리 (TPC에서 수행)의 초당 피크 연산에 대한 자세한 분석을 제공합니다.

이 정보는 인텔® 가우디® 3 AI 가속기의 인상적인 연산 능력을 보여줍니다.

		Intel® Gaudi® 3 AI Accelerator	
Computation Type	Datatype	OAM Peak TFLOP/sec	PCIe Peak TFLOP/sec
MME (Matrix)	FP8	1835	1835
	BF16	1835	1835
	FP16 (signed)	459	459
	TF32	459	459
	FP32	229	229
TPC (Vector)	FP8	57.3	57.3
	BF16	28.7	28.7
	FP16	28.7	28.7
	FP32	14.3	14.3

표 5. 인텔® 가우디® 3 OAM 및 PCIe 매트릭스 및 벡터 컴퓨팅 기능

인텔® 가우디® 3 MME

MME 소개

인텔® 가우디® 3 AI 가속기 매트릭스 곱셈 엔진 (MME)은 인텔® 가우디® 가속기 제품군의 제5세대 MME 엔진입니다. 이 MME는 딥 러닝 알고리즘의 기본이 되는 연산 유형인 행렬 연산을 위해 특별히 설계된 특수 고성능 컴퓨팅 코어입니다. 인텔® 가우디® 3 AI 가속기에는 8개의 MME가 탑재되어 있으며, 각각 64K 병렬 연산을 수행할 수 있습니다.

이러한 대규모 병렬 처리를 통해 높은 수준의 계산 효율을 구현할 수 있으므로, 딥 러닝 워크로드에 널리 사용되는 복잡한 행렬 연산을 처리하는 데 특히 효과적입니다.

인텔® 가우디® 3 AI 가속기의 MME는 현재 딥러닝 모델에서 수행되는 곱셈 연산의 효율성을 위해 맞춤화되었습니다. 다양한 MME 간에 작업을 유연하게 배분하고 MAC의 활용률을 극대화할 수 있는 메모리 지침을 제공하는 풍부한 프로그래머 모델을 갖추고 있습니다.

딥러닝 모델의 크기와 복잡성이 계속 증가함에 따라, 효율적인 고성능 행렬 곱셈 엔진에 대한 수요도 증가할 것으로 예상됩니다. 따라서 인텔® 가우디® 3 AI 가속기의 MME와 같은 기능은 딥러닝 기술의 지속적인 발전에 매우 중요합니다.

MME 아키텍처

인텔® 가우디® 3 AI 가속기는 8개의 MME (행렬 곱셈 엔진)를 갖춘 강력한 연산 능력의 집합체입니다. 각 엔진에는 64K MAC (Multiply-Accumulate Units - MACs) 이 장착되어 있으며, MME 당 200 테라플롭스 이상의 피크 처리량을 구현할 수 있습니다.

이러한 높은 처리량은 인텔® 가우디® 3 AI 가속기의 인상적인 성능 잠재력을 보여줍니다.

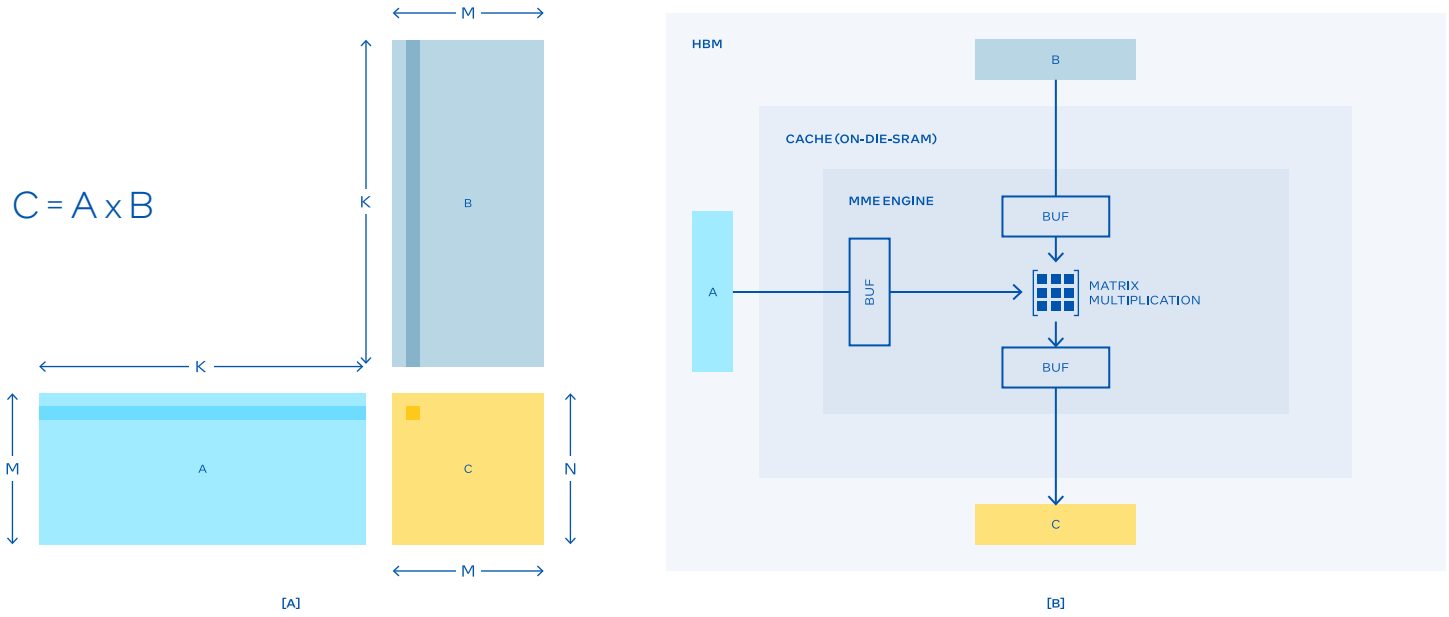


그림 9. 일반적인 행렬 곱셈과 MME 엔진 블록 다이어그램에 대한 매핑

그림 9는 단일 엔진에 대한 기능을 보여줍니다.
이것을 통해 사용자는 MME의 복잡한 작동 방식과 인텔® 가우디® 3 AI 가속기의 전반적인 성능에서 MME의 역할을 더 잘 이해할 수 있습니다. 이러한 지식을 바탕으로 사용자는 컴퓨팅 요구 사항에 맞게 가속기의 기능을 최대한 활용할 수 있습니다.

그림 9A는 일반 행렬 곱셈 (General Matrix Multiplication - GEMM) 연산, 특히 A x B 곱셈의 알고리즘을 보여줍니다. 이 연산은 두 개의 입력 텐서인 A [N x K]와 B [K x N]에서 텐서 C [N x M]를 생성합니다. 행렬 곱셈에서 계산된 각 요소는 세 텐서의 어두운 부분에서 알 수 있듯이 A의 행과 B의 열의 도트 곱입니다.

그림 9B는 데이터 흐름을 자세히 설명하는 블록 다이어그램입니다.
MME는 필요한 차원, 위치, 데이터 유형, 다양한 실행 피연산자로 프로그래밍됩니다.
그런 다음 메모리에서 텐서 A와 B를 검색하여 행렬 곱셈을 하기 위해 스트리밍 버퍼로 가져옵니다.
행렬 곱셈은 최대 64K의 곱하기와 누적 연산을 병렬로 실행할 수 있습니다.
연산이 완료되면 텐서 C를 다시 메모리로 푸시합니다.

메모리 시스템은 캐시와 실제 HBM 메모리로 구성됩니다. 이러한 각 텐서는 MME 동작에 관계없이 온다이 (on-die) SRAM에서 독립적으로 가지고 오거나 (Pull) 또는 SRAM에 집어넣을 수 (Push) 있습니다.

자세한 내용은 메모리 섹션을 참조하세요.
8개의 MME 엔진을 함께 프로그래밍하여 더 큰 작업을 수행할 수 있습니다.
다음의 그림은 8개의 MME를 나타냅니다.

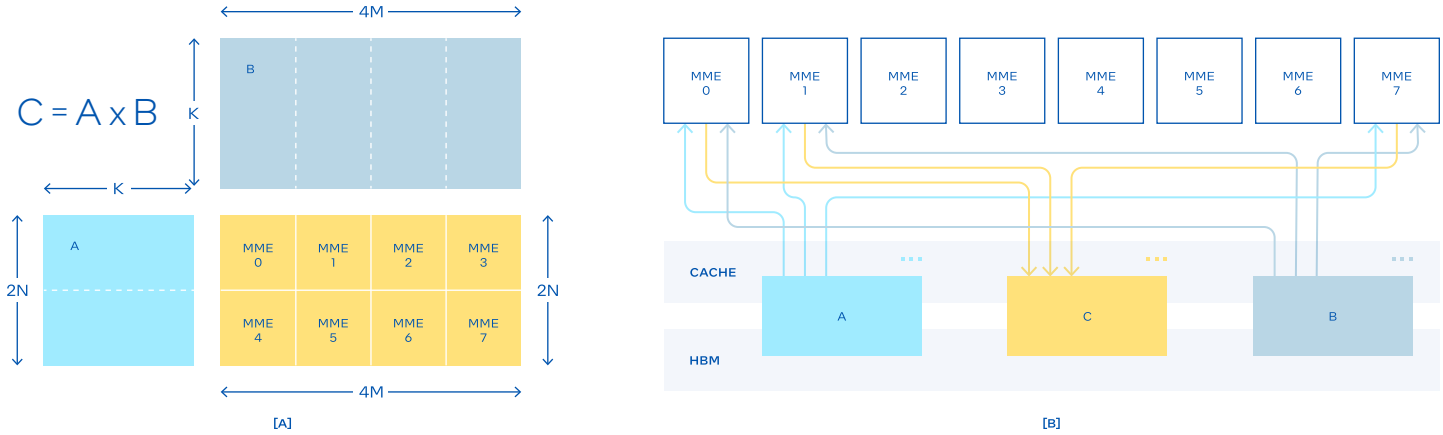


그림 10. 일반적인 행렬 곱셈과 MME 엔진 블록 다이어그램에 대한 매핑

그림 10A는 8개의 MME를 나타내며, A x B 행렬 곱셈이 8개의 MME로 나눠지는 알고리즘 프로세스를 보여줍니다. 각 MME는 공통 차원인 K를 공유하면서 작업의 N x M 슬라이스를 수행합니다.

인텔® 가우디® 3 AI 가속기 메모리 서브시스템은 런타임 최적화와 함께, 가능한 경우, 재사용된 데이터를 HBM에서 한 번만 가져오도록 보장합니다. 예를 들어, mme0, mme2, mme4, mme6는 모두 텐서 A의 상단에서 가져오는 반면, mme0과 mme1은 텐서 B의 4분의 1을 공유합니다.

HL GC 런타임은 필요할 때 가져온 데이터가 캐시에 저장되도록 합니다.

다른 차원 분할도 가능하며 그래프 컴파일러는 다양한 옵션을 분석하여 가장 효율적인 설정을 선택한다는 점에 주목할 필요가 있습니다.

그림 10B는 데이터 흐름을 자세히 설명하는 블록 다이어그램입니다. MME는 병렬로 작동할 수 있으며, 각각 필요한 A와 B의 하위 집합을 가져와서 C 내에서 N x M 하위 집합을 생성합니다. 인텔® 가우디® 3 AI 가속기의 8개의 MME는 병렬로 작동하여 0.5M의 연산을 수행하며, 최대 1.8 TB/s 를 달성합니다.

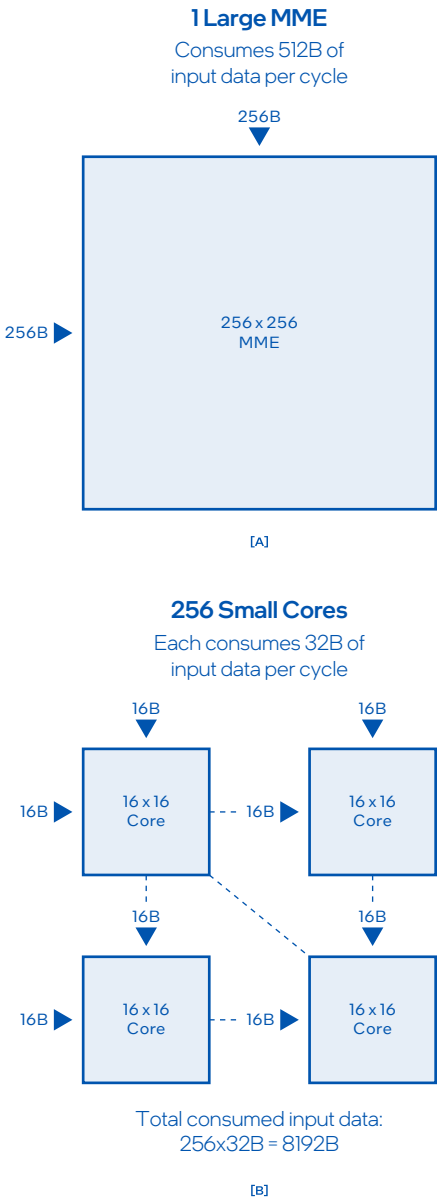


그림 11. 하나의 대형 MME와 256개의 소형 코어 비교. 동일한 연산 능력을 가지고 있음에도 불구하고 MME는 소형 코어보다 16배 더 적은 입력 데이터를 소비합니다.

여러 개의 소형 유닛보다 하나의 대형 매트릭스 곱셈 유닛의 이점

위에서 언급한 바와 같이, 인텔® 가우디® 3 AI 가속기는 8개의 대형 MME를 갖추고 있으며, 각 MME는 사이클당 64k의 MAC을 수행합니다. MME를 AI 워크로드를 위해 수정된 최신 GPU와 비교하면, 인텔® 가우디® 3 AI 가속기는 소수의 대형 행렬 곱셈 유닛을 갖추고 있는 반면, GPU에는 다수의 소형 행렬 곱셈 유닛이 포함되어 있습니다. 그림 11은 GEMM 가속기의 두 가지 옵션, 즉 하나의 큰 유닛과 여러 개의 작은 유닛을 비교한 것입니다.

그림 11는 사이클당 64k MAC (Multiply-Accumulates)을 수행하는 인텔® 가우디® 3 AI 가속기의 MME 한 개와 256개의 소형 GEMM 코어들이 각 사이클당 256 MAC을 수행하여 총 64k MAC을 연산하는 경우를 비교한 것입니다. 이 다이어그램은 MME와 코어가 제곱된 2D 행렬로 구성되어 있다고 가정합니다. MME는 256개의 열에 256개의 행을 포함하고, 소형 코어는 각각 16개의 열에 16개의 행을 포함합니다. 이 비교에서는 입력 요소당 1바이트가 필요한 FP8의 입력 데이터 유형이라고 가정합니다.

그림 11의 두 옵션의 컴퓨팅 성능은 모두 64k MAC/cycle을 수행할 수 있다는 점에서 동일합니다. 그러나 대역폭 관점에서 보면 두 옵션은 크게 다릅니다.

그림 11A에서 대형 MME는 사이클당 두 세트의 256B 입력이 필요하며, 사이클당 총 512B까지 합산됩니다.

반면에, 그림 11B는 각 소형 코어에 사이클당 두 세트의 16B 입력이 필요하며, 이는 사이클과 코어당 최대 32B를 합산합니다. 256개의 소형 코어를 모두 공급하는 데 필요한 입력 데이터의 총 양은 32B의 256배인 8192B에 달합니다. 이는 하나의 대형 MME가 요구하는 것보다 16배나 더 많은 양입니다. MME에 필요한 입력 데이터의 양이 작아진다는 것은 여러 가지 이점이 있다는 것을 의미합니다.

첫째, 입력 대역폭이 16배 감소하면 데이터 전송이 줄어들고 에너지 효율이 높아집니다.
둘째, 입력 대역폭에 대한 요구 사항이 커지면, 시스템이 높은 컴퓨팅 활용도에 도달할 수 있도록 하는 최소 GEMM 차원에 제약이 생깁니다. 예를 들어, 작은 행렬 곱셈 코어가 많은 최신 GPU에서 80%의 컴퓨팅 활용률에 도달하려면 m=n=k=~3K의 GEMM 차원이 필요합니다.
인텔® 가우디® 3 AI 가속기에서 MAC을 100% 활용하려면 m=n=k=1K면 충분합니다.

96MB L2 캐시를 통해 활성화가 파이프라인 되는 경우 (일반적인 경우), MME를 100% 활용하기 위해서는 m=n=k=512이면 충분합니다. 즉, 인텔® 가우디® 3 AI 가속기는 컴퓨팅 활용률이 80%에 불과한 최신 GPU에 비해 컴퓨팅을 100% 활용하기 위해서 GEMM 작업에서 25배 ~ 200배 적은 MAC이 필요합니다. 역설적이게도 상대적으로 큰 행렬 곱셈 가속기를 만들면 대안에 비해 더 작은 GEMM 크기에서 하드웨어를 더 효율적으로 활용할 수 있다는 것을 알 수 있습니다.

MME 데이터 유형

인텔® 가우디® 3 AI 가속기 MME는 모든 주요 AI 컴퓨팅 데이터 유형을 지원합니다 : FP8 (E4M3 및 E5M2 모두), BF16, FP16, TF32 및 FP32. 모든 데이터 유형은 FP32 어큐뮬레이터에 축적됩니다.

FP8이 학습과 추론에 가장 선호되는 컴퓨팅 데이터 유형이 됨에 따라, 인텔® 가우디® 3 AI 가속기의 5세대 MME는 빠르게 FP8 입력 스케일링을 통합하여 FP8로의 스케일링 또는 FP8 로부터의 스케일링을 위한 TPC의 컴퓨팅 부하를 줄입니다.

인텔® 가우디® 3 TPC

텐서 프로세서 코어 소개

인텔® 가우디® 3 AI 가속기는 5세대 텐서 프로세서 코어를 내장하고 있습니다.

TPC는 범용 단일 명령어, 다중 데이터 (SIMD) VLIW 프로세서입니다. 폭은 256B이며 FP32, BF16, FP16 및 FP8 (E4M3 및 E5M2 모두) 데이터 유형을 지원합니다.
또한, 다음과 같은 정수 데이터 타입이 지원됩니다 : UINT32, INT32, UINT16, INT16, UINT8 및 INT8.

로컬 SRAM으로 피연산자를 가져오고 내보내는 데 DMA가 필요한 일반적인 DSP와 달리, TPC는 고급 마이크로 아키텍처 기술로 구현된 DMA 없는 프로그래밍 모델을 통해 소프트웨어 개발을 크게 간소화할 수 있습니다. 또한 동일한 고급 마이크로 아키텍처를 통해 커널 간에 유휴 시간 없이 연속 실행이 가능합니다. 이를 통해 입력 및 출력(캐시 또는 DRAM)의 위치에 관계없이 마이크로초 규모의 커널에서도 TPC를 100% 활용할 수 있습니다.

MME와 마찬가지로 TPC는 작은 크기의 입력을 처리할 때에도 높은 컴퓨팅 활용도를 달성합니다.

TPC 아키텍처

그림 12는 TPC 블록 다이어그램이며 그 기능을 설명합니다.

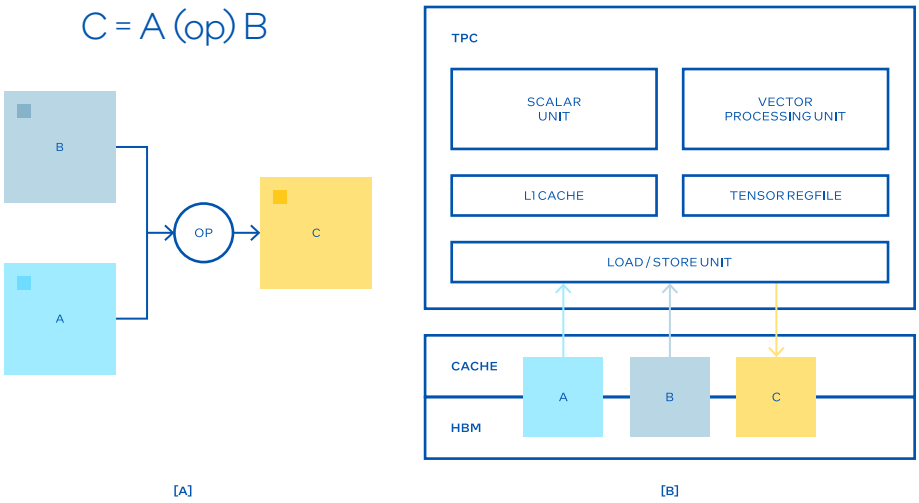


그림 12. 일반적인 행렬 곱셈과 MME 엔진 블록 다이어그램에 대한 매핑

인텔® 가우디® 3 AI 가속기 미디어 엔진

디코더 엔진

인텔® 가우디® 3 AI 가속기에는 14개의 미디어 디코딩 유닛이 있습니다. 다음 형식이 지원됩니다.

Level 4 Video Formats

- 최대 10개의 HEVC 프로파일, 최대 8192 x 8192 해상도
- 프로그레시브 H.264 및 SVC 베이스 레이어 및 최대 4096 x 4096 해상도의 MVC
- 최대 2개의 VP9 프로파일 (10비트) 최대 8192 x 8192 해상도

Image Formats

- 최대 8192 x 8192 해상도의 JPEG
- 최대 8192 x 8192 해상도의 프로그레시브 JPEG

이 블록은 디코딩 외에도 스트림의 후처리도 지원합니다.

Post Processing Features

- 이미지 다운스케일링 (이미지 크기 조정):
- 수직 및 수평 스케일링은 서로 다른 스케일링 비율을 사용할 수 있습니다
- 최대 출력 사진 크기 4096 x 4096
- 이미지 업스케일링 (최대 3배):
- 수직 및 수평 스케일링은 서로 다른 스케일링 비율을 사용할 수 있습니다
- 최대 출력 사진 크기 4096 x 2160
- 이미지 자르기:
  - 시작 위치, 너비 및 높이가 정의 가능한 4 픽셀 정확도의 자르기 매개 변수 설정 사용
  - 디지털 줌
  - 자르기와 업스케일링을 결합하여 지원
- PP는 바이리니어 스케일링, 랜코스 스케일링 지원

인텔® 가우디® 3 AI 가속기는 디코더 블록당 두 개의 포스트 프로세싱 채널을 구현합니다. 하나는 스칼라 (업 / 다운), 다른 하나는 원본 이미지 출력용입니다.

성능

모든 하드웨어 인스턴스의 전반적인 성능은 표 7에 나와 있습니다.

지원하는 형식

동영상 디코더는 다음 기능을 지원합니다:

Feature	Support
Input stream format	▪ YCbCr444, 422, 420 ▪ YCbCr440, 411, 400
Output stream format	YCbCr420 or RGB/BGR packet or per planer

표 7. 디코더 형식

Video Format*	1080p30 Streams
HEVC	250
VP9	300
H.264	200
Image Format*	**1080 img/sec**
Jpeg 420	12000

* 디코더의 실제 성능은 이미지 해상도, 이미지 품질 및 형식 등 다양한 요인에 따라 달라질 수 있습니다.

표 6. 포맷 성능

로테이터 엔진

인텔® 가우디® 3 AI 가속기는 하드웨어 로테이터 엔진을 통합하여 입력 이미지의 다음과 같은 변환을 수행할 수 있습니다 :

- 2D rotation
- 3D rotation
- Projection
- Mesh : 이미지 왜곡 및 왜곡 해제
- 리샘플러 : 사용자 정의 좌표에서 입력 데이터 리샘플링 (Re-sampling)
- 다상 (Polyphase) 필터를 사용한 리스케일링

인텔® 가우디® 3 AI 가속기의 메모리 특성

인텔® 가우디® 3 AI 가속기 온다이 (On-die) 메모리는 이전 가속기 대비 두 가지 주요 개선 사항이 있습니다.

첫 번째는 크기가 48MB에서 총 96MB의 온다이 SRAM으로 2배 증가한 것입니다.

두 번째는 2-레벨 캐시의 통합입니다. 96MB의 온다이 SRAM은 균일하게 액세스 할 수 있는 최종 레벨 캐시 (L3)로 사용하거나 각각 24MB L2 캐시 4조각으로 분할할 수 있으며, 각 슬라이스는 2개의 MME와 16개의 TPC에 액세스 할 수 있습니다.

L2는 L3에 비해 2배 더 높은 캐시 I/O 처리량을 제공합니다. 온다이 메모리를 L2 또는 L3 캐시로 사용하는 것은 인텔® 가우디® 소프트웨어 스택을 통해 구성할 수 있으며, 이 소프트웨어 스택은 I/O 텐서당 최적의 캐시 할당을 동적으로 결정합니다.

인텔® 가우디® 3 AI 가속기는 3.6GHz 주파수에서 실행되는 8개의 HBM2e 디바이스를 통합하여 인텔® 가우디® 2 AI 가속기보다 50% 높은 3.7TB/s의 피크 HBM 대역폭을 제공합니다. 각 HBM2e 디바이스 용량은 16GB로, 총 128GB에 달하며, 이는 2세대 가속기보다 33% 높고, 80GB의 HBM 메모리를 탑재한 경쟁 GPU 솔루션보다 1.6배 더 높습니다.

더 큰 메모리 용량의 장점은 두 가지입니다.

첫 번째 장점은 더 작은 HBM 용량으로 더 많은 디바이스를 필요로 하는 구성을 실행할 수 있다는 점입니다. 또 다른 장점은 배치 크기를 늘리거나 사전 계산을 피하는 등 컴퓨팅 효율이 더 높은 구성을 사용할 수 있다는 점입니다. 빠르게 진화하는 딥 뉴럴 네트워크 (DNN) 가속 환경에서 인텔® 가우디® 3 AI 가속기의 혁신적인 메모리 서브시스템은 뛰어난 성능을 발휘합니다. 이 서브시스템은 제품의 핵심 구성 요소로, 매트릭스 곱셈 엔진 (MME) 및 텐서 프로세서 코어 (TPC)와 조화롭게 작동하여 탁월한 성능을 제공하도록 설계되었습니다.

	Gaudi® 2 OAM	Gaudi® 3 OAM
PCIe	Gen4 x 16	Gen5 x 16
PCIe Peak BW	64 GB/s bidirectional	128 GB/s bidirectional
HBM	6 x HBM2E	8 x HBM2E
HBM Capacity	96 GB	128 GB
HBM Peak BW	2.46 TB/s	3.7 TB/s
On-die-SRAM	48 MB	96 MB
On-die-SRAM BW	6.4 TB/s	19.2 TB/s
TDP	600 W	900 W

표 8. 가우디® 2 OAM과 가우디® 3 OAM의 메모리 특성

가상 공간 접근성

인텔® 가우디® 3 AI 가속기 메모리 서브시스템의 핵심은 사용자가 VRAM에 액세스 할 때 가상공간에서 작동할 수 있도록 하는 메모리 관리 유닛 (MMU)입니다. 이 기능은 메모리 관리의 복잡성을 추상화하여 원활한 사용자 경험을 제공합니다.

고급 캐싱 시스템

인텔® 가우디® 3 AI 가속기 메모리 서브시스템은 L2 및 L3 캐시가 장착되어 있으며, 각각의 DCORE 및 HBM 메모리 채널에 각각 연결됩니다. 캐시 시스템은 몇 가지 주요 기능을 통해 데이터 액세스를 최적화하도록 설계되었습니다 :

- **높은 처리량** : 이 시스템은 L2 액세스의 경우 최대 19.2 TB/s, L3 액세스의 경우 6.4 TB/s의 총처리량을 제공합니다.
- **대용량 및 세트 연결성** : 96MB의 용량과 12방향 세트 연결성을 갖춘 캐시 시스템은 대용량 데이터를 효과적으로 처리할 수 있습니다.
- **할당 힌트** : 사용자는 L2에 캐싱할 것인지 L3 또는 둘 다에 캐싱할지 지정할 수 있어 데이터 관리를 더욱 효과적으로 제어할 수 있습니다.
- **연령 (Age) 대체 알고리즘** : 시스템은 사용자 정의 클래스를 고려한 연령 교체 알고리즘을 사용하여 사용자가 정의한 클래스 및 우선순위를 고려하여 캐시 리소스를 효율적으로 사용할 수 있도록 합니다.
- **유지 관리 명령** : 이 명령은 캐시 활용도를 높이고 불필요한 데이터가 HBM 리소스를 소모하는 것을 방지합니다.

고대역폭 메모리 인스턴스

인텔® 가우디® 3 AI 가속기 메모리 서브시스템에는 8개의 고대역폭 메모리 (HBM) 인스턴스를 포함하며, 최대 128GB의 용량과 3.7TB/s의 총 대역폭을 제공합니다. 이러한 막대한 용량과 처리량 덕분에 시스템은 대용량의 데이터를 효과적으로 처리할 수 있습니다.

결론적으로, 인텔® 가우디® 3 AI 가속기 메모리 서브시스템은 DNN 가속화의 한계를 뛰어넘기 위한 노력의 증거입니다. 진보된 기능과 고성능은 우리 제품의 필수적인 부분입니다. 오늘날 DNN 애플리케이션의 까다로운 요구 사항을 충족하는 솔루션을 제공할 수 있게 되었습니다.

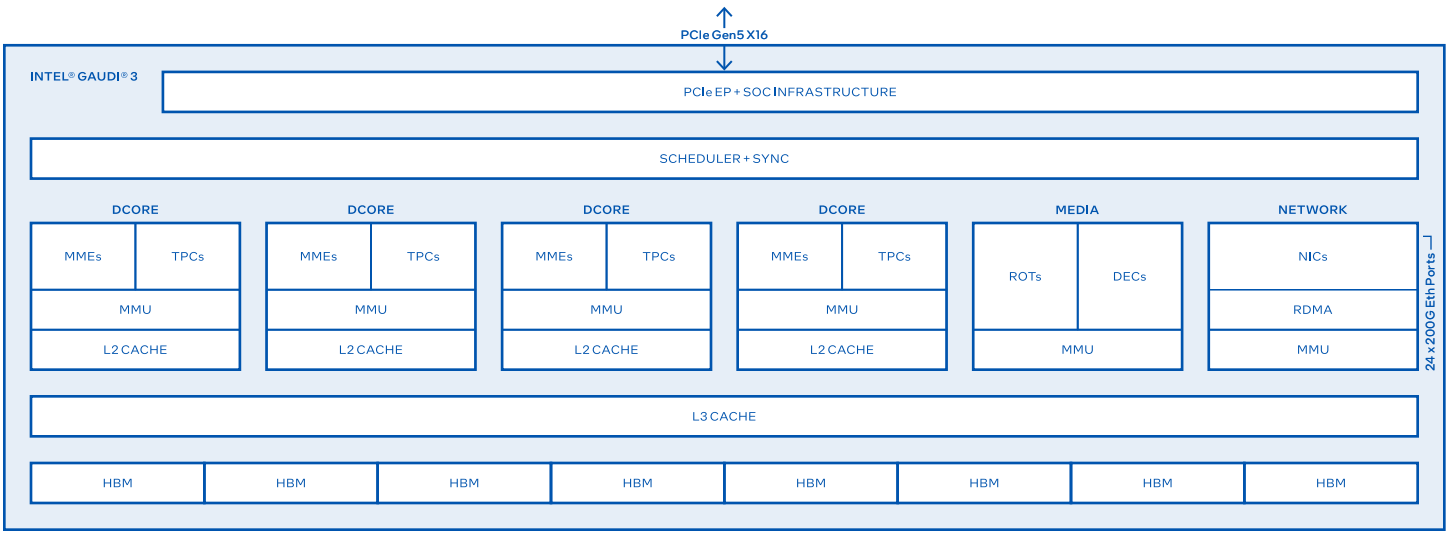


그림 13. DCORE 관점의 인텔® 가우디® 3 아키텍처와 및 지원 소프트웨어 계층

네트워킹

인텔® 가우디® 3 AI 가속기에 컨버지드 이더넷을 통한 RDMA를 통합하면 단일 노드에서 수천 개의 노드까지 대규모로 유연하게 확장할 수 있는 뚜렷한 이점을 제공합니다.
솔루션의 확장 기능의 장점을 설명하려면 네트워크의 기초인 인텔® 가우디® 3 AI 가속기에 포함된 네트워킹 아키텍처에서 시작해야 합니다.

인텔® 가우디® 3 AI 가속기의 혁신적인 NW 서브시스템은 원활한 데이터 이동과 효율적인 작업 관리의 원동력입니다. 이 시스템의 핵심은 데이터 이동을 조율 (오케스트레이션) 하는 마스터 컨덕터인 인텔® 가우디® 커뮤니케이션 라이브러리 (IGCL)입니다. 이 시스템에는 프로그래밍이 가능한 스케줄링 메커니즘이 탑재되어 있어 작업 종속성을 유지하면서 엔진을 원활하게 활성화할 수 있습니다.

인텔® 가우디® 3 AI 가속기 네트워킹 서브 시스템은 24개의 200기가 비트 이더넷 NIC 포트, 레이어 2 MAC, 및 RDMA 엔진을 갖추고 있습니다. 이 강력한 설정은 고속 데이터 전송과 뛰어난 성능을 지원합니다.

무엇보다도 인텔® 가우디® 3 AI 가속기에는 4개의 전용 어그리게이션 엔진이 있습니다. 이 엔진은 통신 라이브러리 대신 작동하여 합산 작업을 수행합니다. 이는 더 빠른 계산과 더 효율적인 데이터 처리를 의미합니다.

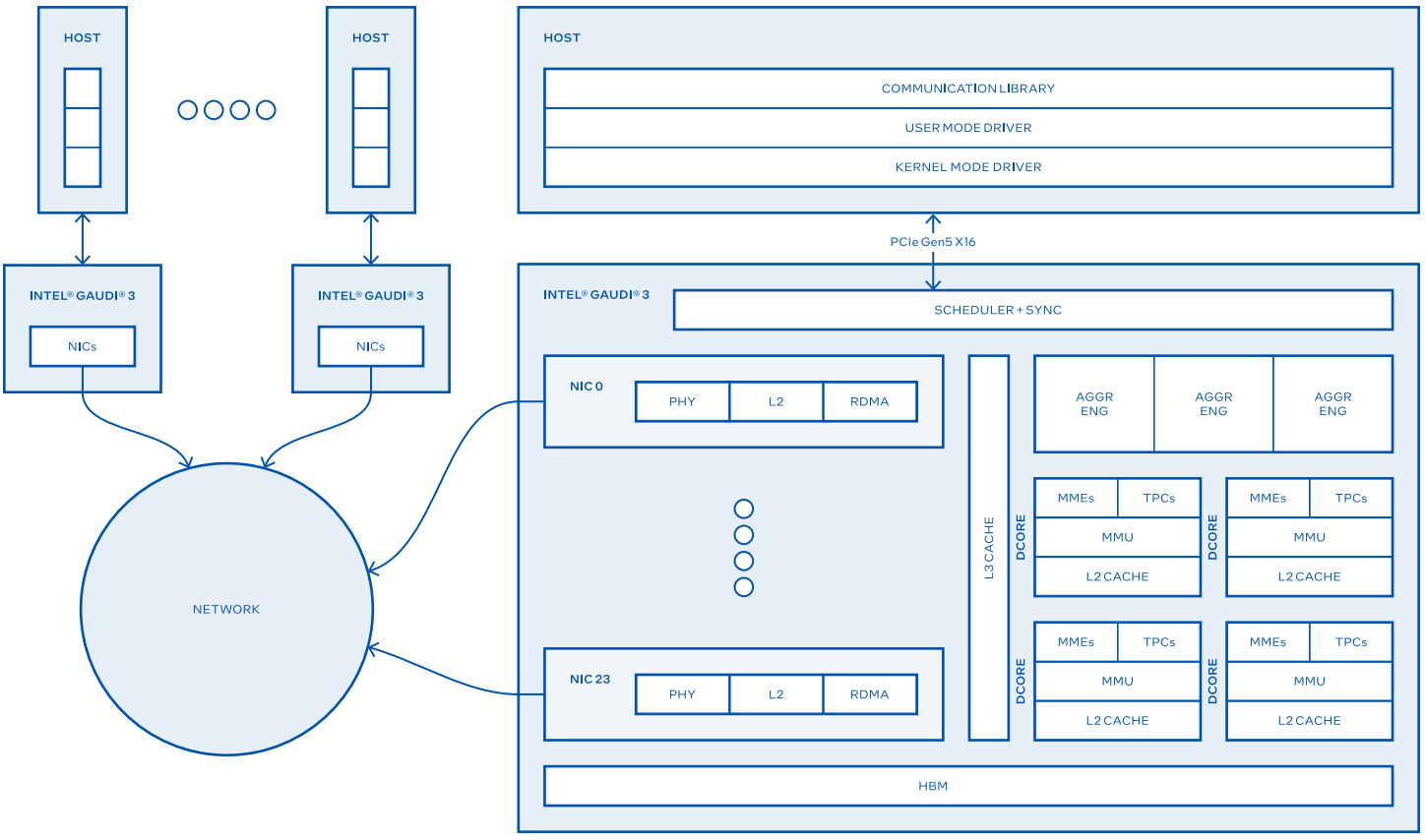
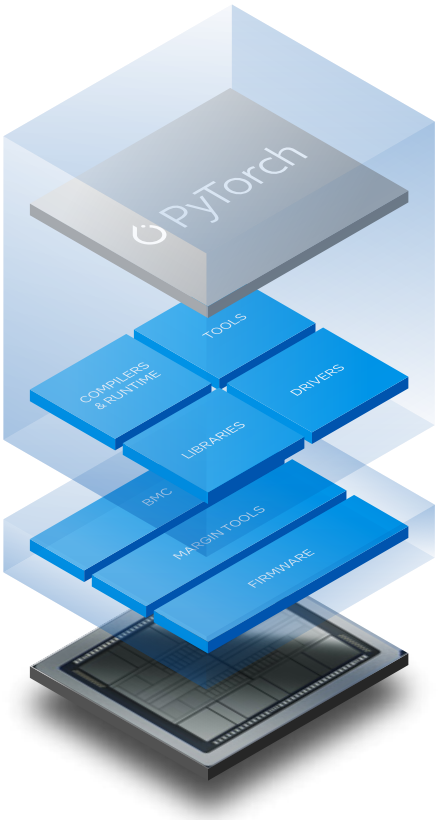


그림 14. 일반적인 행렬 곱셈과 MME 엔진 블록 다이어그램에 대한 매핑



가우디 소프트웨어 제품군

인텔® 가우디® 소프트웨어 제품군

인텔® 의 AI 가속기에서 고성능 딥 러닝 (DL) 트레이닝 및 추론을 용이하게 하도록 설계된 인텔® 가우디® 소프트웨어 제품군은 신경망 토폴로지를 인텔® 가우디® 하드웨어 제품군에 효율적으로 매핑합니다. 이 소프트웨어 제품군에는 그래프 컴파일러, 자동 커널 퓨저, 사전 컴파일된 커널 라이브러리와 같은 low-level 구성 요소와 AI에코시스템 (PyTorch, DeepSpeed, Hugging Face, vLLM, Ray 등) 이 통합되어 있습니다. 인텔® 가우디® 소프트웨어에는 Paged Attention, Flash Attention등과 같은 인기 알고리즘의 맞춤형 구현도 포함되어 있습니다.

그래프 컴파일러 및 런타임

인텔® 가우디® 그래프 컴파일러는 인텔® 가우디® AI 가속기에서 주어진 모델 토폴로지를 구현하는 최적화된 바이너리 코드를 생성합니다. 연산자 융합, 데이터 레이아웃 관리, 병렬화, 파이프라이닝 및 메모리 관리, 그래프 레벨 최적화를 수행합니다. 그래프 컴파일러는 다양한 성능 최적화된 연산 (예 : 요소별, 비선형, non-GEMM 연산자 등)이 포함된 풍부한 TPC 커널 라이브러리를 사용합니다. 인텔® 가우디® 3 AI 가속기 하드웨어 (MME, TPC 및 DMA)의 이기종 특성을 고려할 때, 인텔® 가우디® 그래프 컴파일러는 프레임워크 그래프의 병렬 및 파이프라인 실행을 통해 효과적인 활용을 가능하게 합니다. 인텔® 가우디® 소프트웨어는 스트림 아키텍처를 사용하여 비동기 작업의 동시 실행을 관리하고, 인텔® 가우디만의 컴퓨팅과 네트워킹 조합을 지원하여 프레임워크에 멀티스트림 아키텍처를 제공합니다. 컴퓨팅, 네트워킹, DMA 등 다양한 유형의 스트림이 호스트의 개입 없이 최소한의 지연 시간으로 서로 동기화됩니다.

TPC 프로그래밍

인텔® 가우디® 소프트웨어 TPC SDK에는 LLVM 기반의 TPC-C 컴파일러, 시뮬레이터 및 디버거가 포함되어 있습니다. 이러한 도구는 맞춤형 TPC 커널을 쉽게 개발할 수 있도록 지원합니다. SDK는 고성능 커널을 빌드하는 데 사용됩니다. 사용자는 인텔® 가우디® AI 가속기에서 맞춤형 딥 러닝 모델과 알고리즘을 개발하여 고유한 요구 사항에 맞게 혁신하고 최적화할 수 있습니다.

TPC 프로그래밍 언어인 TPC-C는 프로세서 고유의 SIMD 기능을 쉽게 활용할 수 있도록 언어 데이터 유형이 추가된 C99의 파생 언어입니다. 기본적으로 넓은 벡터 데이터 유형을 지원하여 SIMD 엔진의 프로그래밍을 지원합니다 (예 : float64, uchar256 등). 텐서 기반 메모리 액세스, 특수 함수를 위한 가속, 난수 생성 및 여러 데이터 유형을 포함하여 딥 러닝을 위한 많은 명령어가 내장되어 있습니다.

에코시스템 통합

인텔® 가우디® 소프트웨어는 1.x와 2.x 모두 PyTorch에 기본적으로 통합되어 있습니다.
또한 많은 인기 있는 소프트웨어 패키지들에도 통합되어 있습니다 : 분산 트레이닝 및 추론을 위한 DeepSpeed, 트랜스포머 및 디퓨저 모델 사용을 위한 Hugging Face, 최첨단 LLM 서빙 처리량을 위한 vLLM 등이 있습니다. 인텔® 가우디® 소프트웨어 PyTorch Python 패키지는 Flash Attention과 같은 가우디에 최적화된 여러 연산을 제공하여 LLM 학습 및 추론과 관련된 기존 생태계에 혁신을 이끕니다.

양자화 (Quantization)

인텔® 가우디® 3 AI 가속기는 이전 제품보다 FP8 데이터 유형을 훨씬 더 많이 지원합니다. 인텔® 가우디® 소프트웨어는 사용자에게 높은 정확도와 향상된 처리량으로 기존 모델을 변환하는 자동화된 정량화 도구의 형태로 제공하며, 기존 모델과의 호환성을 위해 트랜스포머 엔진과 유사한 API를 지원합니다.

또한 인텔® 가우디® 소프트웨어는 AWQ 및 GPTQ와 같은 int4 가중치 전용 양자화 체계를 지원하며, 인텔® 뉴럴 컴프레서의 산하에 있는 양자화 도구를 오픈 소스화하여 해당 영역에서 사용자의 혁신을 가능하게 합니다.

자동 커널 융합 (Automatic Kernel Fusion)

커널 융합은 트레이닝 및 추론, 메모리 대역폭 개선, 오버헤드 상각, 추론의 경우 전체 메모리 용량을 줄이고 배치 크기를 늘려 효율성을 높일 수 있는 여러 이점이 있습니다. 인텔® 가우디® 소프트웨어에는 사용자의 개입 없이도 사용자 그래프의 원시 커널 시퀀스에서 융합 커널을 자동으로 생성할 수 있는 최첨단 MLIR 기반 커널 퓨저가 포함되어 있습니다.

그런 다음 이러한 커널은 그래프 컴파일러에 인터페이스 되어 인텔® 가우디® 가속기의 이기종 아키텍처를 활용합니다.

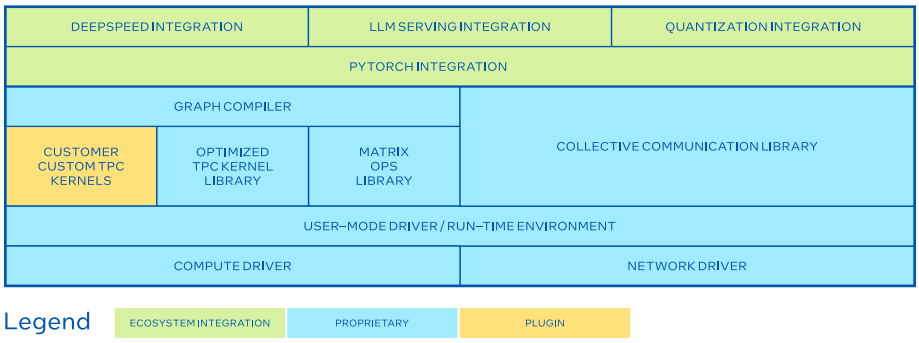


그림 15. 인텔® 가우디® 소프트웨어 스택

네트워킹

딥 러닝 트레이닝은 일반적으로 여러 다양한 장치에서 수행되므로 인텔® 가우디® 3 AI 가속기의 NIC (네트워크 인터페이스 컨트롤러) 은 전체 인텔® 가우디® 3세대 트레이닝 솔루션에서 필수적인 구성 요소입니다. 인텔® 가우디® 3 AI 가속기 NIC는 네트워크 디바이스들 간의 DNN 그래프 분포에 맞게 맞춤화되었습니다. (스케일 아웃)

NIC는 소프트웨어 개입 없이 안정적인 연결을 통해 높은 대역폭과 짧은 지연 시간을 특징으로 하는 RDMA (원격 직접 메모리 액세스)를 컴퓨팅 엔진에 제공합니다. 일반적인 클라우드 인프라에 적합하도록 NIC 포트는 각 방향에서 총 대역폭이 4.8Tb/s인 이더넷 (Eth) 연결을 사용하며, 다양한 포트 구성을 지원합니다. NIC는 일반적으로 널리 사용되는 이더넷 인프라와 InfiniBand (IB) 프로토콜의 안정적이고 지연시간이 짧은 RDMA의 이점을 활용하여 RoCE v2 사양을 구현합니다.

인텔® 가우디® 가속기 구현은 DNN 애플리케이션 및 대규모 배포에 더 적합하도록 RoCE v2 사양을 확장하여 수천 개의 인텔® 가우디® 가속기에 대한 선형 확장성을 지원합니다.

다음 섹션에서는 인텔® 가우디® 3 AI 가속기가 지원하는 주요 RoCE 확장에 대해 중점적으로 설명합니다.

MPI 집합 연산을 RDMA에 매핑하기

RDMA 프로토콜은 자연 읽기 및 쓰기 연산을 사용하여 원격 메모리 액세스를 지원합니다. RDMA 읽기 및 쓰기 작업은 개시자가 로컬 및 원격 메모리에 대한 포인터를 모두 가지고 있다고 가정합니다. 그러나 DNN 애플리케이션은 일반적으로 송수신 접근 방식을 기반으로 하는 MPI 스타일의 집합 연산을 사용합니다.

이 접근 방식은 세 가지 주요 요소를 정의합니다.

첫째, 송신 버퍼에 대한 포인터를 가진 발신자 측

둘째, 수신 버퍼에 대한 포인터를 가진 수신자 측

셋째, 양쪽 간에 데이터를 이동하는 랑데뷰 플로우입니다.

따라서 MPI 집합 연산은 RDMA 읽기 및 쓰기 연산에 자연스럽게 매핑되지 않습니다.
이 매핑을 수행하는 방법에는 여러 가지가 있으며, 각 방법마다 장단점이 있습니다. MPI 작업을 RDMA 송수신 작업에 매핑하는 것도 한 가지 옵션입니다. 이 옵션은 랑데뷰 플로우를 해결하지 못합니다.

수신자가 수신 작업을 게시하기 전에 발신자가 수신자에게 데이터를 보내면 재전송을 하기 위해 RNR NACK이 전송되어 성능이 크게 저하됩니다. 또 다른 옵션은 임시 버퍼를 사용하여 수신자 측에서 랑데뷰를 구현하고 나중에 수신자 버퍼를 사용할 수 있게 되면 mem-copy를 시작하는 것입니다. 이 옵션에는 높은 지연 시간 및 높은 메모리 용량과 같은 많은 단점이 있습니다.

인텔® 가우디® 가속기는 발신자 측에서 랑데뷰 플로우를 해결하는 하드웨어 기반을 구현하여 데이터가 mem-copy 없이 수신자에게 한 번만 전송되도록 합니다. 이 접근 방식을 사용하면 간단한 집합 API를 사용자에게 노출하여 CPU에서 NIC의 하드웨어로 복잡성을 오프로드하여 지연 시간을 최소화하고 메시지 전송 속도를 높일 수 있습니다.

콜렉티브 커널을 하드웨어로 오프로드하기

실제로는, 집합 연산의 실행은 순위 간에 여러 개의 송수신 연산으로 분할하여 수행됩니다. HLS-3 레퍼런스 서버에는 8개의 서로 다른 인텔® 가우디® 가속기가 포함되어 있습니다. 각 디바이스에는 고유한 랭크 ID가 할당됩니다. 장치 간 연결은 여러 포트를 사용하여 이루어집니다.

따라서 집합 연산은 랭크 간 및 각 랭크를 연결하는 여러 포트 간에 분할되어야 합니다. 분할 프로세스는 CPU 리소스를 소모하여 잠재적으로 포트 사용률을 낮추고 전송 대역폭을 낮출 수 있습니다. 인텔® 가우디® 3 AI 가속기 NIC는 하드웨어에 집합 연산을 오프로드하여 하드웨어가 300KB의 작은 버퍼 크기로 전체 대역폭을 달성할 수 있도록 합니다.

혼잡 제어-적시 기반 (Congestion Control – Timely Based)

DNN 클러스터가 점점 더 커짐에 따라 네트워크의 혼잡이 더 큰 문제로 대두되고 있습니다. 손실이 발생하는 네트워크의 혼잡은 패킷 드롭으로 인해 성능이 크게 저하될 수 있습니다. 무손실 네트워크를 달성하기 위해 우선순위 기반 흐름 제어 (PFC)를 활성화하면 패킷 드롭을 방지할 수 있지만 스위칭 레이어 간에 혼잡이 확산될 수 있습니다. RoCE v2는 명시적인 혼잡 알림 (ECN)을 기반으로 RoCE 혼잡 관리 (RCM)를 구현합니다. 그러나 RCM은 정제되지 않은 방법이기 때문에 처리량 변동성이 큽니다.

인텔® 가우디® 3 AI 가속기에서 혼잡 제어는 ECN을 지원할 뿐만 아니라 SWIFT® 와 같은 시간 기반 혼잡 체계도 지원하도록 확장되었습니다. 이러한 알고리즘은 지연 (RTT 계산)을 혼잡 신호로 사용하므로 ECN을 훨씬 더 세밀하게 제어할 수 있습니다.

다중 경로 로드 밸런싱 (패킷 스프레이)

대규모 노드 클러스터를 연결하려면 멀티 레이어 토폴로지가 필요합니다. 이러한 경우, 노드 간의 네트워크 연결에는 여러 경로가 포함될 수 있습니다. 두 노드 간의 네트워크 대역폭을 최대한 활용하고 혼잡을 줄이려면 가능한 모든 동일한 비용의 경로 간에 트래픽이 균형을 이루어야 합니다. 동일 비용 다중 경로 (ECMP : Equal-cost multipath)를 활용하면 해결할 수 있습니다. 그러나 위에서 설명한 것처럼 대규모 클러스터는 다른 경로에 영향을 미치는 정체가 발생하여 처리량이 감소하고 워크플로우의 완료 시간이 늘어날 수 있습니다.

이러한 혼잡을 완화하기 위해 로드 밸런싱 시스템을 도입했습니다. 이 시스템은 경로의 부하를 고려하고 비용 함수를 조정하여 대역폭 사용률을 높게 유지하고 지연 시간을 낮춥니다. 로드 밸런싱 시스템은 다른 경로로 전송되는 패킷의 순서를 변경하는 방법을 제공합니다.

RDMA 안정적 연결 (Reliable Connection) 메모리 사용량

RDMA 안정적 연결을 사용하여 대규모의 올투올 연결 클러스터를 배포하면 확장할 수 없는 메모리 풋프린트로 인해 문제가 발생할 수 있습니다. 각각 P개의 프로세스가 있는 N개의 노드로 구성된 클러스터를 생각해 보세요. 모든 P 프로세스가 모든 노드의 모든 프로세스와 통신하려는 경우, RDMA RC (안정적 연결) 서비스는 각 노드에 P^2× (N-1) QP를 필요로 합니다.

각 QP에는 100 바이트 크기의 컨텍스트와 10KB바이트 크기의 작업 대기열이 포함됩니다. 우리의 구현에서 QP는 인텔® 가우디® 가속기의 집합 통신 라이브러리 (CCL)에 의해 처리됩니다. CCL은 가장 일반적으로 클러스터의 각 피어 노드에 대해 4개의 QP를 열므로 각 노드의 총 QP 수는 4× (N-1) 입니다. 따라서 메모리 풋프린트는 노드 수에 따라 확장 가능합니다.

네트워크 내 축소 (In-network Reduction)

축소 연산에 대한 컴퓨팅 요구 사항을 줄이고 통신 단계와 우수한 중첩을 제공하기 위해 인텔® 가우디® 3 AI 가속기는 네트워크 경로에서 축소 연산을 수행할 수 있는 기능을 지원합니다. 지원되는 연산은 sum, min 및 max입니다. 또한 축소 연산은 FP32, FP16, BF16 및 FP8을 포함한 다양한 데이터 유형을 지원합니다. 또한 정확도를 높이기 위해 BF16 및 FP16 축소를 FP32 누적과 함께 수행할 수 있습니다.

네트워크 및 컴퓨팅 동기화

일부 DNN 가속기 시스템은 개별 NIC를 사용하여 클러스터의 다른 노드와 통신합니다.
이러한 시스템에서, 데이터를 전송하는 네트워킹과 데이터를 소비하는 컴퓨팅 엔진 간의 동기화는 높은 호스트 CPU 사용률로 인한 높은 레이턴시 (지연 시간)의 어려움을 겪습니다.
인텔® 가우디® 3 AI 가속기는 NIC와 컴퓨팅 엔진을 모두 통합하며, 이들 간의 동기화는 칩 내에서 호스트의 개입 없이 최소한의 지연 시간으로 이루어집니다.

텐서 시맨틱스 (Tensor Semantics)

표준 RDMA 연산은 연속 버퍼에서 작동하도록 설계되었지만, DNN 애플리케이션은 텐서 및 서브-텐서 시맨틱에서 작동하도록 설계되었습니다. RDMA 연산을 위해 서브 텐서를 연속 버퍼에 매핑하는 것은 매우 복잡하거나 확장성이 없을 수 있습니다. 따라서 인텔® 가우디® 3 AI 가속기는 칩의 다른 모든 엔진과 마찬가지로 텐서 시맨틱스에서 로컬 및 원격 메모리에 모두 액세스 할 수 있는 텐서 엔진을 NIC에 도입했습니다.

선택적 재전송 및 고장 배달 (Selective Retransmission and Out of Order Delivery)

높은 처리량과 짧은 지연 시간을 제공하기 위해 현재의 RoCE 구현은 네트워크의 무손실성에 의존합니다. 이는 이더넷 기반 네트워크가 손실을 감수하는 반면, 인피니밴드 (IB) 네트워크는 크레딧에 의존하기 때문입니다.

인피니밴드 (IB)에서 패킷 손실에 대한 복구 구현은 Go-Back-N, 즉 NACK이 도착하면 ONA에서 다시 재전송하는 방식입니다. 이는 대역폭과 흐름 완료 시간 및 꼬리 지연 시간에 큰 영향을 미치며, 산발적으로 드롭하는 경우에도 마찬가지입니다. 이는 ONA에서 NTS로 전송되는 모든 패킷이 재전송되기 때문입니다.

데이터 센터가 주로 우선순위 기반 흐름 제어 (PFC)의 확장성 제한으로 인해 손실 아키텍처로 이동하고 있기 때문에, 인텔® 가우디® 3 AI 가속기 RoCE 구현은 인피니밴드 전송 계층 사양을 확장하고 응답자의 선택적 ACK 및 요청자의 선택적 재전송을 허용합니다. 다른 모든 용도로는 IB 사양이 여전히 유효합니다. 따라서 인텔® 가우디® 3 AI 가속기의 RoCE는 선택적 ACK 구현을 통해 TCP/IP보다 훨씬 더 확장성이 뛰어납니다.

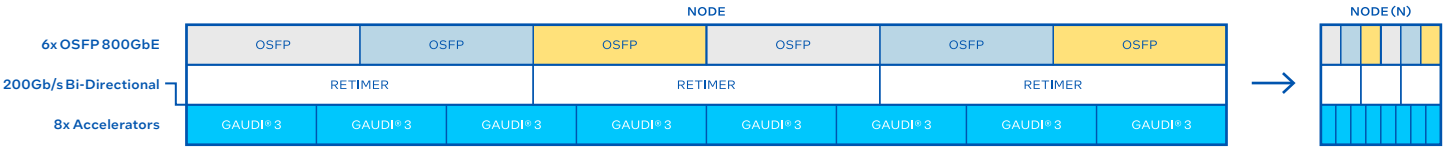
클러스터 아키텍처

표준 이더넷 스위치를 사용하여 모듈식 고성능 클러스터를 원하는 규모에 맞게 구축할 수 있습니다. 다음은 16노드 하위 클러스터 빌딩 블록을 사용하여 512노드 클러스터 (4096개의 인텔® 가우디® 3 AI 가속기)를 구축하는 예시입니다. 인텔® 가우디® 3 AI 가속기 기반 서버에서 각 OAM 카드에는 3개의 OSFP 스케일아웃 포트에 연결된 NIC 포트가 있습니다.

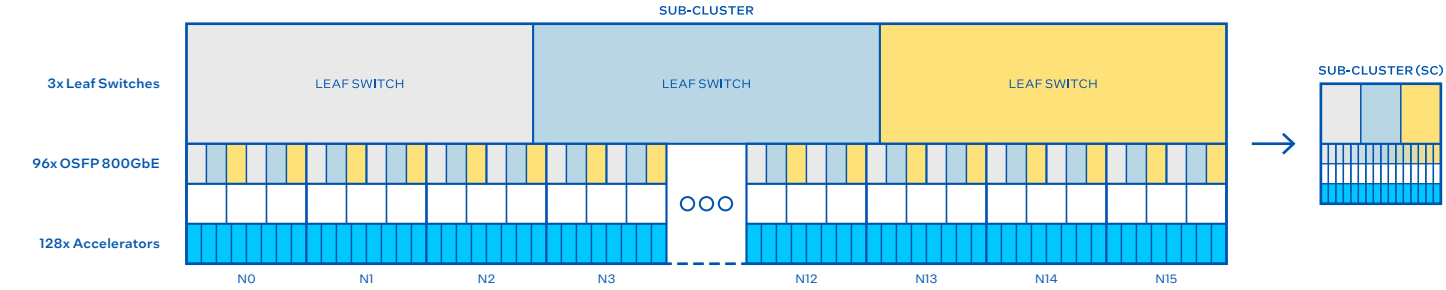
그런 다음 16대의 서버를 3개의 64포트 800Gbps 이더넷 리프 스위치에 연결하여 서브클러스터를 구축합니다. 서브 클러스터에서 시스템의 모든 카드는 3개의 리프 스위치를 통해 다른 시스템의 다른 카드와 통신할 수 있습니다. 마지막으로 32개의 서브 클러스터는 48개의 64포트 800Gbps 이더넷 스파인 스위치를 사용하여 함께 네트워크로 연결됩니다. 이 토폴로지는 각 리프 및 스파인 스위치의 64포트가 모두 활용되는 3중 네트워크를 형성합니다.

그림 16은 단일 노드, 16노드 하위 클러스터, 512노드 클러스터로 확장된 GenAI 시스템을 보여줍니다.

Node Level Architecture



Sub-Cluster Level Architecture (16 Nodes)



Cluster Level Architecture (32 Sub-Clusters, 512 Nodes)

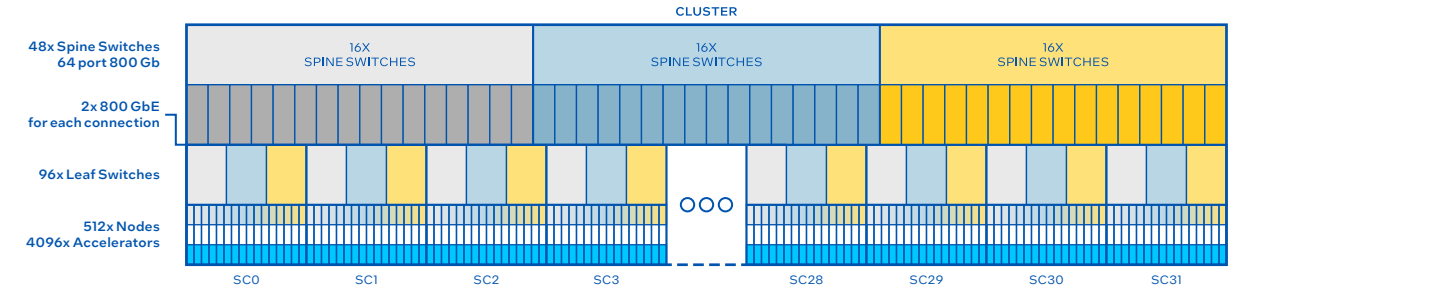


그림 16. 인텔® 가우디® 3 AI 가속기-스케일아웃 클러스터 아키텍처

모든 것을 요약하면 : 하드웨어와 소프트웨어를 결합한 통합 AI 가속 솔루션

인텔® 가우디® 소프트웨어는 인텔® 가우디® 3 AI 가속기 하드웨어 활용도를 크게 향상시키는 포괄적인 기능을 제공합니다. 이 장에서는 인텔® 가우디® 소프트웨어의 다양한 구성 요소를 원활하게 통합하여 워크로드 런타임을 단축하는 방법을 설명합니다. LLM (대규모 언어 모델)에서 가져온 실제 사례를 살펴봄으로써, 소프트웨어 레이어가 하드웨어 효율성에 미치는 영향을 주의 깊게 살펴봅니다.

트랜스포머 하위 시퀀스의 나이브한 실행

대규모 언어 모델 (LLM)은 일련의 반복되는 트랜스포머 레이어로 구성됩니다.
각 트랜스포머 레이어에는 다음과 같은 복잡한 작업 순서가 포함됩니다 :

1. 일반 행렬 곱셈 (GEMM) : 선형 트랜스포메이션을 위한 기본 연산.
2. 배치-GEMM : 여러 입력을 효율적으로 처리하는 최적화된 GEMM 변형.
3. 정규화 : 소프트맥스, 레이어 정규화 또는 RMSNorm과 같은 기술을 포함.
4. 잔여 추가 (Residual-Add) : 정보 흐름을 보존하기 위한 중요한 구성 요소.
5. 비선형 활성화 기능 : GELU 또는 SwiGLU를 선택할 수 있습니다.
6. 드롭아웃 (훈련 전용) : 오버피팅을 방지하기 위한 정규화 기법.

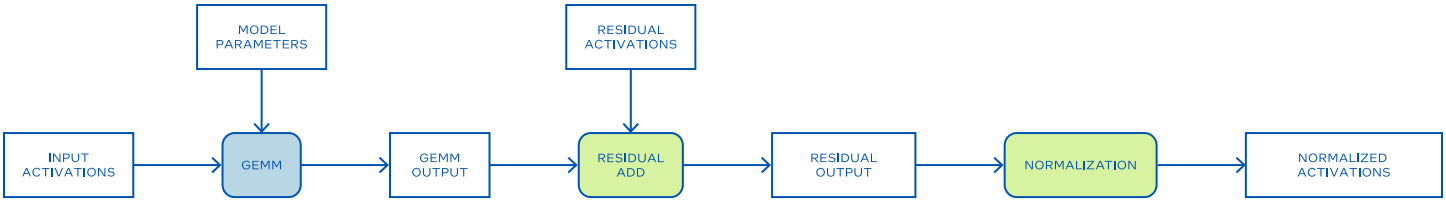


그림 17. 대형 언어 모델의 트랜스포머 레이어 하위 그래프 일러스트

그림 17에서는 트랜스포머 레이어 내에서 두 번 반복되는 작업의 하위 시퀀스를 시각화합니다.

파란색 그래프 노드 (모서리가 둥근 직사각형)는 MME 연산을 나타내고, 녹색 그래프 노드는 TPC 연산을 나타냅니다.
작업의 하위 시퀀스는 다음 단계로 구성됩니다 :

1. GEMM : MME에 의해 실행됨
2. 잔여-추가 : TPC에 의해 실행
3. 정규화 : TPC에 의해 실행됨

최적화 없이 일련의 작업을 실행하면 그림 18에 표시된 런타임이 발생합니다.
8개의 MME 유닛은 모두 논리적으로 하나의 유닛으로 작동하며 GEMM을 완료까지 실행합니다.
각 TPC 커널은 TPC의 ISA와 마이크로아키텍처를 효율적으로 사용하는 인텔® 가우디® 가속기 TPC SDK를 사용하여 작성되었습니다.

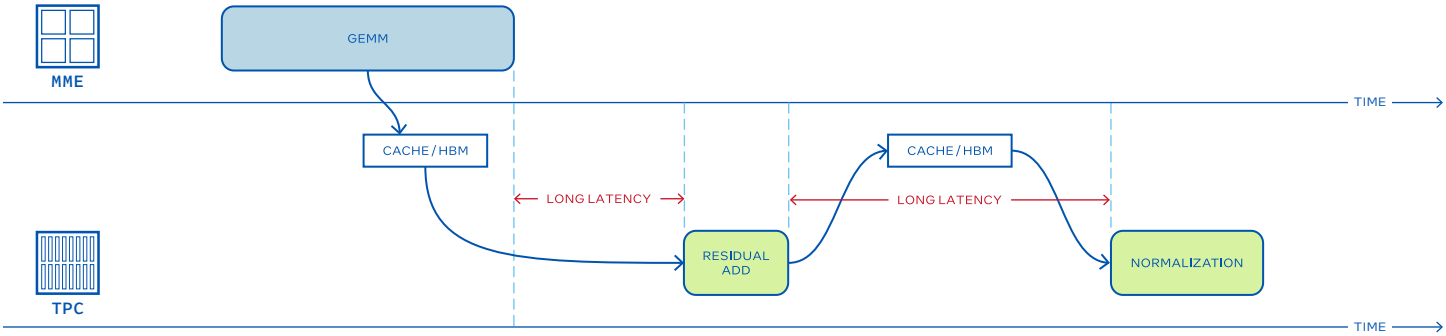


그림 18. 그림 17의 하위 그래프에 대한 나이브한 소프트웨어 구현의 타임라인 실행 일러스트

그림 18에 표시된 실행에서는 전체 GEMM 출력이 캐시에 기록됩니다. 그러나 일반적으로 대규모 워크로드에서와 같이 전체 GEMM 출력은 캐시 용량을 초과할 수 있으므로 캐시에 맞지 않는 모든 GEMM 출력은 HBM에 기록됩니다. TPC는 모든 GEMM 결과 쓰기가 완료된 후 첫 번째 커널 실행을 시작합니다.

TPC가 실행을 시작하면 입력 크기가 크기 때문에 일부 입력을 HBM에서 읽게 되므로 상대적으로 긴 지연 시간 (1~2초) 이 발생하며, 이는 HBM에 의해 결정됩니다. 두 번째 TPC 커널은 첫 번째 커널과 동일한 긴 지연 시간을 경험합니다.

자동 커널 퓨전

인텔® 가우디® 소프트웨어 제품군이 제공하는 즉각적인 개선 사항 중 하나는 자동 커널 융합입니다. 두 개의 TPC 커널이 자동으로 융합되어 개별 커널 내의 연산을 결합한 새로운 커널을 생성합니다. 융합된 커널의 입력과 출력은 융합된 커널의 외부 입력과 출력입니다. 커널을 융합하면 오리지날 커널 간의 중간 결과를 읽거나 쓰는 I/O를 절약할 수 있습니다.

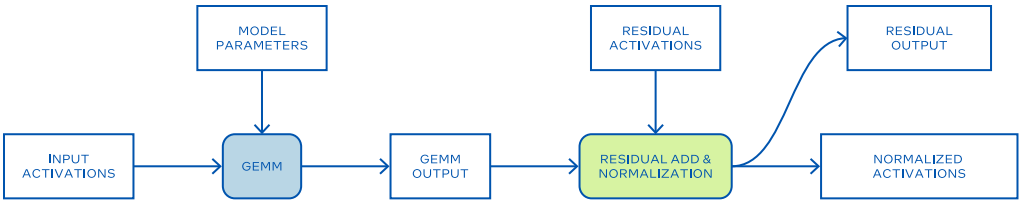


그림 19. 퓨저가 두 개의 연속된 TPC 커널을 하나로 융합한 후의 그림 17의 하위 그래프 일러스트

그림 17의 서브-그래프를 융합하면 그림 19의 서브-그래프가 됩니다.

1. 한 커널에서 결과를 쓰고 다음 커널에서 동일한 결과를 읽는 데 드는 입출력 시간을 절약할 수 있습니다.
2. 커널 간 지연 시간 절약.

그림 19의 서브-그래프를 실행하면 그림 20에 표시된 실행 결과가 됩니다. 런타임 이득은 위에서 설명한 두 가지 방식으로 나타납니다.

첫째, 실행 중인 엔진 사이에는 지연 시간이 하나만 존재합니다.

둘째, 융합된 커널의 정규화 부분이 외부 I/O가 아닌 TPC 내부에서 입력을 읽기 때문에 전체 TPC 런타임이 감소하여 I/O 시간을 절약할 수 있습니다.

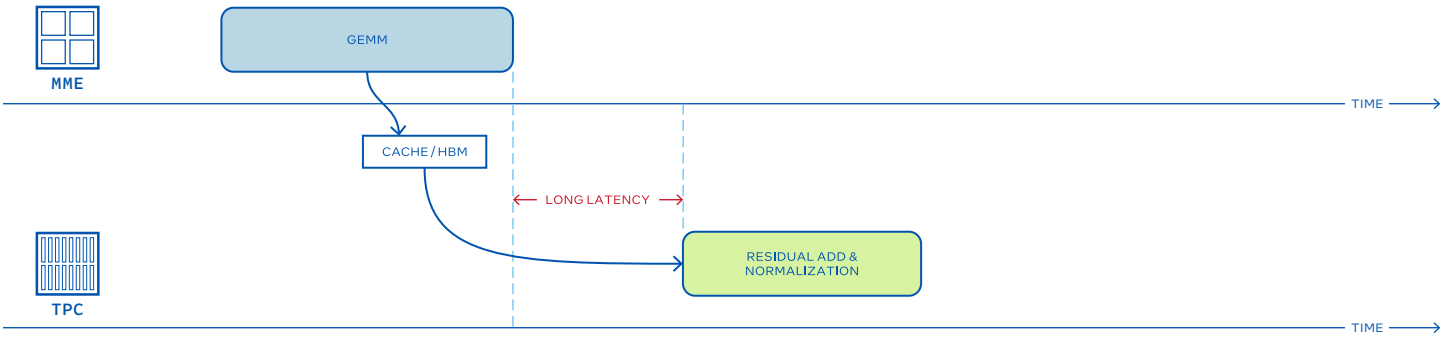


그림 20. 그림 19의 하위 그래프에서 디바이스 실행을 보여주는 일러스트

그래프 컴파일러를 통한 생산자-소비자 파이프라이닝

이전 섹션에서 설명한 그래프 컴파일러는 인텔® 가우디® 3 AI 가속기의 이기종 아키텍처에서 워크로드 실행을 최적화하도록 특별히 설계되었습니다. 논리적으로 실현 가능한 경우, 종속 엔진은 파이프라인 방식으로 작동하도록 예약되어 생산자-소비자 실행 종속성을 확립합니다. 생산자는 MME와 TPC가 공유하는 최상위 캐시 계층 구조인 L2 캐시에 출력을 씁니다. 출력이 완전히 쓰여지면 소비자는 캐시에서 이 출력을 입력으로 읽습니다. 인텔® 가우디® 소프트웨어는 생성된 데이터가 캐시 내에 잘 맞도록 보장하며, 세분화된 작업을 통해 효율적인 디바이스 활용을 가능하게 합니다.

캐시 기반 파이프라이닝을 활용하면 생산자와 소비자 간의 지연 시간을 최소화하여 모든 디바이스 엔진을 최적으로 활용할 수 있습니다.

그림 21은 이러한 생산자와 소비자의 관계, 즉 MME가 실행하는 GEMM 연산과 TPC가 실행하는 융합 잔여 덧셈 및 정규화 연산을 보여줍니다. GEMM 연산의 입력과 출력은 4개의 슬라이스로 분할되며, 각 출력 슬라이스는 MME에서 완전히 생성된 다음 TPC에서 입력으로 읽힙니다.

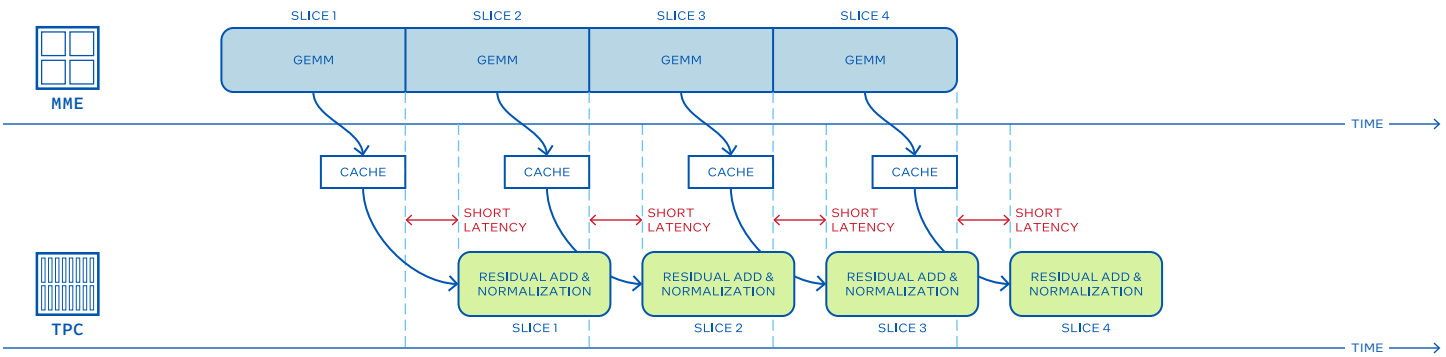


그림 21. MME가 생산자이고 TPC가 소비자인 경우 캐시를 통한 MME와 TPC 간의 파이프라이닝 일러스트

Mix에 네트워크 추가하기

위에서 언급한 바와 같이, 인텔® 가우디® 3 AI 가속기 아키텍처는 본질적으로 NIC를 포함한 모든 엔진의 병렬 실행을 지원합니다. 그림 19의 예시를 확장하여 GEMM 연산에 이어 잔여 추가 전에 All-reduce 집단 통신을 추가합니다. 그림 22는 All-reduce 집단 연산을 추가할 때 형성되는 서브-그래프를 보여줍니다. 훈련 및 추론 사용 사례를 위해 여러 장치 간에 LLM을 텐서 병렬식으로 분할하는 경우 All-reduce가 필요합니다.

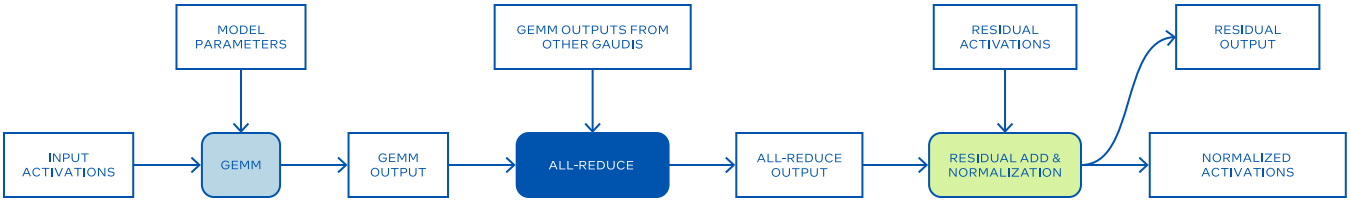


그림 22. 그림 19의 하위 그래프에 GEMM과 융합된 TPC 커널 간의 All-reduce 집단 통신 작업이 표시된 일러스트

나이브한 실행에서는 All-reduce가 파이프라인을 방해하게 됩니다. 그 결과 그림 23에 표시된 것처럼 GEMM, All-reduce 및 융합된 TPC 커널이 순차적으로 실행됩니다. 엔진 간 병렬 처리와 엔진 활성화 사이에 긴 대기 시간이 발생하지 않습니다.

LLM 코드를 병렬 실행이 가능한 방식으로 구조화함으로써 그래프 컴파일러와 HCL은 여러 엔진에 걸쳐 워크로드를 효율적으로 워크로드를 여러 엔진에 효율적으로 분산할 수 있습니다. 이 접근 방식은 기기 활용도를 극대화합니다.

그림 24에서는 그림 22에 표시된 하위 그래프의 인텔® 가우디® 3 AI 가속기에서의 효과적인 실행을 시각화합니다. 구체적으로, 이 다이어그램은 다음과 같은 데이터 흐름을 보여줍니다. NIC에 대한 생산자로서의 MME, NIC는 MME의 출력에 대한 소비자이자 TPC에 대한 생산자, 그리고 TPC는 NIC 출력의 소비자입니다.

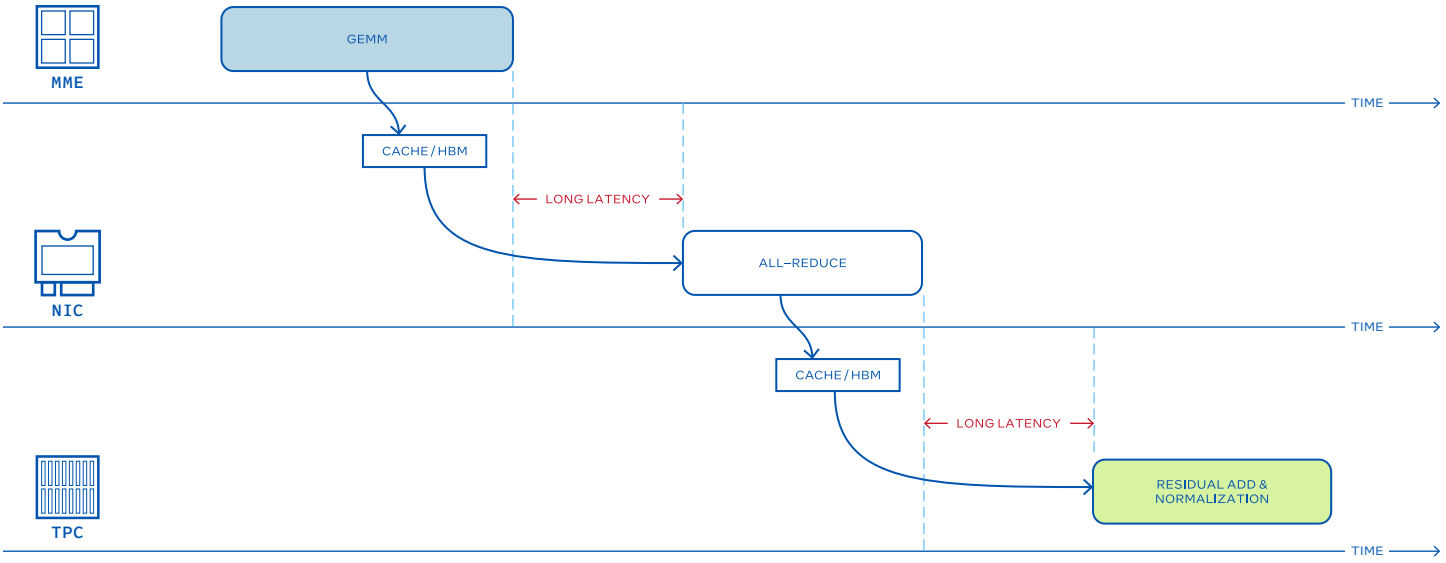


그림 23. 그림 22의 하위 그래프에서 연산을 최적화없이 실행한 경우

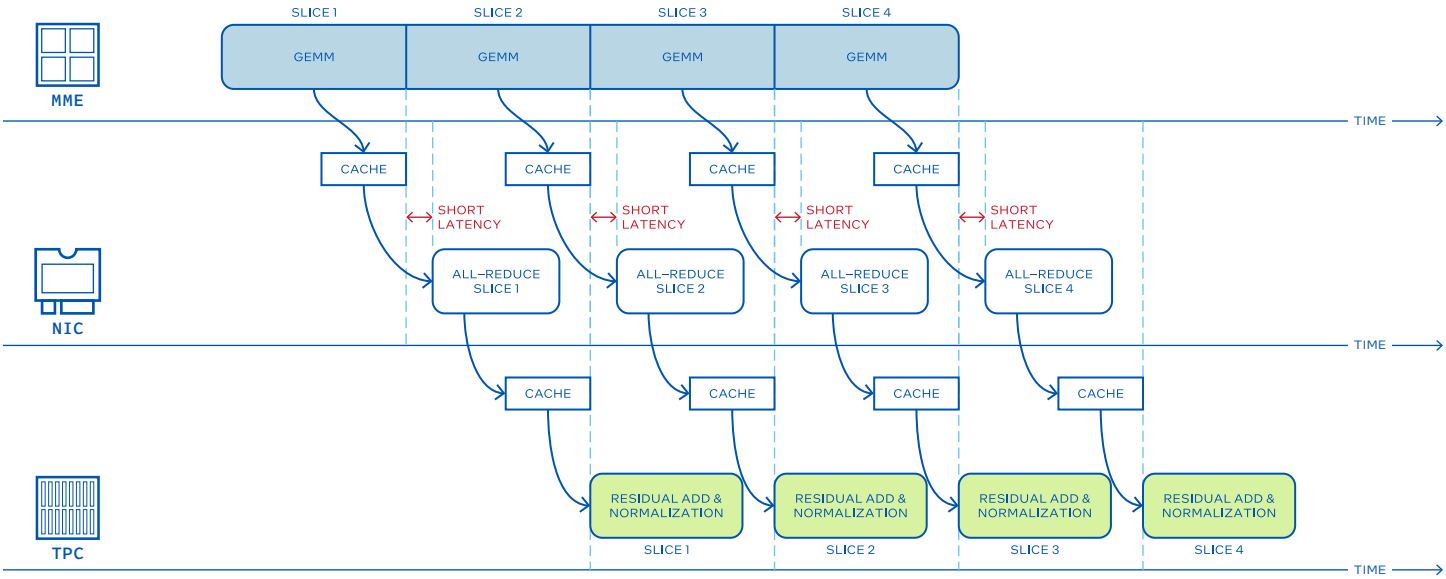


그림 24. 세 엔진 간의 생산자-소비자 관계를 통한 MME, NIC 및 TPC의 최적화된 스케줄링

그림 25. 인텔® 가우디® 3 AI 가속기 제품 라인

인텔® 가우디® 3의 성능 향상

인텔® 가우디® 3 AI 가속기는 인텔® 가우디® AI 가속기 제품군의 3세대입니다. 대용량 HBM 용량과 넓은 대역폭은 인텔® 가우디® 3 AI 가속기가 최첨단 성능의 GenAI 트레이닝 및 추론 성능을 가능하게 합니다.

학습 시나리오에서는 이전 세대에 비해 인텔® 가우디® 3 AI 가속기의 거의 모든 향상된 기능이 사용됩니다. 트레이닝은 컴퓨팅 집약적인 작업이기 때문에 향상된 컴퓨팅 성능은 즉각적인 성능 향상을 제공합니다. 증가된 HBM 대역폭은 더 큰 컴퓨팅 연산을 처리할 수 있습니다. 또한, 더 큰 HBM 용량은 성능 향상에도 기여합니다. HBM 용량이 커지면 배치 크기가 증가하여 컴퓨팅 활용도를 높일 수 있으며, 워크로드의 특정 부분을 다시 계산하거나 런타임 중에 네트워킹 작업을 추가하는 모델-병렬 분할을 피할 수 있습니다.

일반적으로, LLM의 추론 처리량은 모델의 매개변수와 컨텍스트 윈도우를 읽는 데 사용되는 사용 가능한 HBM 대역폭에 의해 결정됩니다.

인텔® 가우디® 3 AI 가속기와 인텔® 가우디® 2 AI 가속기를 비교했을 때, 소형 LLM (13B 크기 이하 모델)의 속도 향상은 두 세대의 가속기 간 HBM 대역폭 비율과 비슷한 약 1.5배임을 알 수 있습니다. 그러나, LLama-70B와 같은 더 큰 LLM 모델을 비교해보면, 개선 효과는 HBM 대역폭 비율보다 훨씬 더 크고, 2배 이상을 초과할 것으로 예상됩니다.

이러한 성능 향상은 인텔® 가우디® 3 AI 가속기에서 사용할 수 있는 메모리 용량이 더 커졌기 때문입니다. 메모리 용량이 커지면 배치 크기를 늘릴 수 있으므로 주어진 시간당 더 많은 샘플을 처리할 수 있습니다.

인텔® 가우디® 3 AI 가속기의 측정된 성능은 인텔® 가우디® 소프트웨어 릴리스와 동시에 인텔® 가우디® 3 AI 가속기의 모델 성능에 업데이트 및 게시될 예정입니다.