제2장. AI 시스템 하드웨어 개요

이 작품은 AI를 사용하여 번역되었습니다. 여러분의 피드백과 의견을 환영합니다: translation-feedback@oreilly.com

---

슈퍼컴퓨터 수준의 AI 하드웨어를 단일 랙에 압축한다고 상상해 보십시오. NVIDIA의 최신 아키텍처가 바로 이를 실현합니다. 이 장에서는NVIDIA가 CPU와 GPU를 강력한 *슈퍼칩*으로 융합한 후 초고속 상호 연결기술로 수십 개를 결합해 '상자 속 AI 슈퍼컴퓨터'를 어떻게 구현했는지 살펴봅니다. 기본 하드웨어 구성 요소인 Grace CPU와 Blackwell GPU를 살펴보고, 이들의 긴밀한 통합과 방대한 메모리 풀이 AI 엔지니어의 작업을 어떻게 용이하게 하는지 알아보겠습니다.

이어서 72개의 GPU를 마치 하나의 기계처럼 연결하는 네트워킹 패브릭으로 시야를 넓힐 것입니다. 이 과정에서 컴퓨팅 성능, 메모리 용량, 효율성 측면에서 이 시스템에 초능력을 부여하는 도약들을 강조할 것입니다. 마지막에는 이 첨단 하드웨어가 이전에는 불가능해 보였던 수조 개 매개변수 모델의 훈련과 서비스를 어떻게 가능하게 하는지 이해하게 될 것입니다.

CPU와 GPU 슈퍼칩

AI 확장성을 위한 NVIDIA의 접근법은 단일 통합 CPU + GPU 슈퍼칩 모듈 수준에서 시작됩니다. Hopper 세대부터 NVIDIA는 ARM 기반 CPU와 하나 이상의 GPU를 동일한 유닛에 패키징하고 고속 인터페이스로 긴밀하게 연결하기 시작했습니다. 그 결과 통합 컴퓨팅 엔진처럼 작동하는 단일 모듈이 탄생했습니다.

슈퍼칩의 첫 구현체는 Grace CPU 하나와 Hopper GPU 하나를 결합한 Grace Hopper(GH200)였습니다. 다음으로 등장한 Grace Blackwell(GB200) 슈퍼칩은 동일한 패키지 내에 Grace CPU 하나와 Blackwell GPU 두 개를 결합했습니다. 그림 2-1에서 볼 수 있듯이, Grace CPU는 모듈 중앙에 위치하며 두 개의 Blackwell GPU 다이에 둘러싸여 있습니다.

그림 2-1. NVIDIA Grace Blackwell 슈퍼칩 모듈은 단일 모듈에 하나의 Grace CPU(중앙)와 두 개의 Blackwell B200 GPU(좌측 상단 및 우측 상단)를 포함하며, 공유된 통합 메모리 공간을 가지고 NVLink-C2C(칩 간)라는 맞춤형 고속 링크로 연결되어 있습니다.

기존 시스템에서는 CPU와 GPU가 별도의 메모리 풀을 가지며 상대적으로 느린 버스(PCIe 등)를 통해 통신하므로 데이터가 오가며 복사되어야 합니다. NVIDIA의 슈퍼칩인 는 NVLink-C2C(칩 간)라는 맞춤형 고속 링크로 CPU와 GPU를 연결하여 이러한 장벽을 제거합니다.

NVLink-C2C는 GB200 슈퍼칩 내 Grace CPU와 Blackwell GPU 간에 최대 약 900GB/s의 대역폭을 제공합니다. 비교를 위해, PCIe Gen5 x16(Blackwell B200)은 방향당 약 64GB/s, PCIe Gen6 x16(Blackwell Ultra B300)은 방향당 약 128GB/s입니다. NVLink-C2C의 상호 연결 속도는 일반적인 PCIe보다 한 차원 빠릅니다. 또한 중요한 점은 캐시 일관성을 유지한다는 것입니다.

캐시 일관성이란 CPU와 GPU( )가 일관된 통합 메모리 아키텍처를 공유함을 의미합니다. 따라서 양측은 항상 동일한 값을 인식합니다. 실제로 슈퍼칩 상의 Grace CPU와 Blackwell GPU는 하나의 거대한 메모리 풀처럼 서로의 메모리에

직접 접근할 수 있습니다. GPU는 CPU 메모리에 저장된 데이터를 읽거나 쓸 수 있으며, 그 반대의 경우도 명시적인 복사 없이 가능합니다. NVIDIA는 이 통합 메모리 아키텍처를 종종 '통합 CPU-GPU 메모리' 또는 '확장 GPU 메모리(EGM)'라고 부르며, 이는 CPU 메모리와 GPU 메모리 사이의 경계를 효과적으로 모호하게 만듭니다.

각 Grace Blackwell 슈퍼칩 은 엄청난 양의 메모리를 탑재합니다. Grace CPU에는 수백 기가바이트의 LPDDR5X DRAM이 연결되어 있으며, 각 Blackwell GPU는 자체 고속 대역폭 메모리(HBM) 스택을 보유합니다.

GB200 슈퍼칩에서 Grace CPU는 최대 ~480GB의 LPDDR5X 메모리를 최대 ~500TB/s로 제공하며, 두 개의 Blackwell GPU는 합쳐서 최대 ~384GB의 HBM3e 메모리(GPU당 총 192GB)를 제공합니다. 전체적으로 GB200 슈퍼칩은 GPU와 CPU가 통합 주소 공간에서 접근 가능한 약 900GB의 일관성 있는 통합 메모리를 제공합니다.

간단히 말해, 각 슈퍼칩은 거의 1테라바이트에 달하는 고속 통합 메모리를 자유롭게 활용할 수 있습니다. 이는 거대 AI 모델에 있어 판도를 바꾸는 요소입니다. 기존 시스템에서는 단일 GPU가 100GB 미만의 메모리로 제한될 수 있었으며, 이는 해당 용량을 초과하는 모델을 분할하거나 느린 저장 장치로 오프로드해야 함을 의미했습니다. 반면 여기서는 GPU가 CPU 메모리를 확장 메모리처럼 원활하게 활용할 수 있습니다.

신경망 레이어나 대형 임베딩 테이블이 GPU의 로컬 HBM에 들어가지 않을 경우 CPU 메모리에 상주할 수 있으며, GPU는 NVLink-C2C를 통해 여전히 이를 처리할 수 있습니다. 프로그래머 관점에서 통합 가상 주소 공간과 일관성은 정확성 확보를 단순화합니다. 그러나 성능을 위해 비동기 프리페치 및 단계적 파이프라인과 같은 기법을 사용해 배치와 메모리 이동을 명시적으로 관리해야 합니다. NVLink-C2C를 통해 LPDDR5X에 접근하는 것은 HBM에 직접 접근하는 것보다 더 높은 지연 시간과 약 10배 낮은 대역폭을 가집니다.

GPU 메모리는 여전히 CPU 메모리보다 훨씬 빠르고 GPU 코어에 더 가깝습니다. CPU 메모리는 크지만 다소 느린 확장 장치로 생각할 수 있습니다. LPDDR5X의 데이터 접근은 GPU의 HBM만큼 빠르지 않습니다. 대역폭은 약 10배 낮고 지연 시간은 더 깁니다. 지능형 런타임은 가장 자주 사용되는 데이터를 HBM에 보관하고, 오버플로우 또는 속도가 덜 중요한 데이터에는 CPU의 LPDDR5X를 사용합니다. 핵심은 오버플로우 시 더 이상 NVMe SSD나 네트워크를 통해 외부로 접근할 필요가 없다는 점입니다.

GPU는 CPU RAM에서 약 900GB/s(방향당 450GB/s) 속도로 데이터를 가져올 수 있습니다. 이는 HBM보다 느리지만 NVMe SSD 저장 장치에서 가져오는 것보다 훨씬 빠릅니다. 이러한 유연성은 매우 중요합니다. 예를 들어, 500GB 크기의 모델(단일 GPU의 HBM으로는 너무 큰 크기)도 HBM 192GB(사용 가능 180GB)와 CPU 메모리 약 500GB를 합친 용량에 접근할 수 있는 하나의 슈퍼칩 모듈 내에 완전히 배치할 수 있기 때문입니다. 이 모델은 여러 GPU에 걸쳐 모델을 분할하지 않고도 실행할 수 있습니다. GPU는 필요할 때 CPU 메모리에서 추가 데이터를 투명하게 불러옵니다.

본질적으로, 모델 전체가 슈퍼칩의 CPU + GPU 메모리 합계 내에 들어갈 수 있다면 메모리 크기는 초대형 모델을 수용하는 데 있어 더 이상 엄격한 한계가 되지 않습니다. 많은 연구자들이 모델이 GPU에 들어가지 않아 발생하는 두려운 "메모리 부족" 오류를 경험해 왔습니다. 이 아키텍처는 그 한계를 크게 확장하도록 설계되었습니다.

NVIDIA Grace CPU

이 슈퍼칩에서 그레이스 CPU의 역할은 범용 작업 처리, GPU에 공급할 데이터의 전처리 및 공급, 그리고 연결된 방대한 메모리 관리입니다. 클럭 속도는 다소 낮지만, LPDDR5X 메모리에 최대 약 500GB/s에 달하는 엄청난 메모리 대역폭과 100MB가 넘는 L3 캐시를 포함한 풍부한 캐시로 이를 보완합니다.

핵심 철학은 GPU로 데이터를 전달할 때 CPU가 절대 병목이 되어서는 안 된다는 점입니다. 스토리지에서 데이터를 스트리밍하거나 토큰화, 데이터 증강 같은 실시간 데이터 변환을 수행하며 NVLink-C2C를 통해 GPU에 매우 효율적으로 공급할 수 있습니다. 워크로드 일부가 CPU에서 더 효율적이라면 Grace 코어가 이를 처리하고 결과를 GPU가 즉시 활용할 수 있게 합니다.

이는 CPU가 GPU가 취약한 영역(예: 임의 메모리 접근이나 제어 중심 코드)에서 GPU의 역량을 확장하고, GPU는 CPU가 따라잡지 못하는 연산 집약적 작업을 가속화하는 조화로운 결합입니다.

CPU와 GPU 간 저지연 링크 덕분에 일반적인 오버헤드 없이 작업을 교환할 수 있습니다. 예를 들어, 명령이 느린 PCIe 버스를 통과할 필요가 없기 때문에 CPU에서 GPU 커널을 실행하는 것이 기존 시스템보다 훨씬 빠르게 이루어집니다. CPU와 GPU는 본질적으로 동일한 보드에 위치합니다. 이는 느린 원격 함수 호출과 빠른 로컬 함수 호출의 차이와 유사합니다. 다음으로, 슈퍼칩의 강력한 엔진인 블랙웰 GPU에 대해 살펴보겠습니다.

NVIDIA 블랙웰 "듀얼 다이" GPU

블랙웰은 이 GPU 세대에 대한 엔비디아의 코드명 으로, 컴퓨팅 성능과 메모리 측면에서 이전 호퍼(H100) GPU 대비 상당한 도약을 의미합니다. Blackwell B200 및 B300 "Ultra" GPU는 단일 칩이 아닙니다. 대신 단일 모듈에 두 개의 GPU 다이를 배치한 다중 칩 모듈(MCM) 설계를 사용합니다. 따라서 Blackwell은 *듀얼 다이 GPU*라고 불립니다(그림 2-2 참조).

이 섹션에서는 듀얼 다이 아키텍처의 세부 사항을 다루지만, 책의 나머지 부분에서는 Blackwell의 두 GPU 다이를 합쳐 단순히 "Blackwell GPU"로 지칭할 것입니다.

이 *칩릿* 접근 방식은 일반적으로 하나의 거대한 GPU로 구현될 부분을 더 작은 GPU 다이로 분할한 후, 초고속 온패키지 다이-투-다이 상호 연결로 결합합니다. 왜 이렇게 할까요? 단일 모놀리식 다이는 실리콘 칩의 크기 한계로 인해 제조 공정에 제약을 받기 때문입니다. 두 개의 물리적 GPU 다이를 단일 모듈로 결합함으로써 NVIDIA는 모듈의 총 트랜지스터 예산을 두 배로 늘릴 수 있습니다.

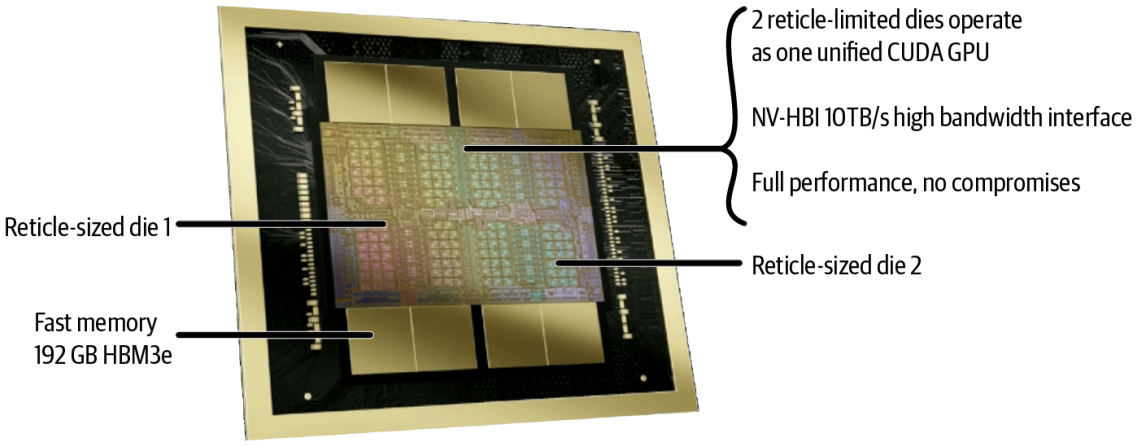


그림 2-2. 블랙웰 듀얼 다이 멀티칩 모듈(MCM) 설계

블랙웰 B200 MCM의 경우, 각 GPU 다이는 약 1040억 개의 트랜지스터와 96GB HBM3e 메모리를 갖습니다. 결합된 GPU 모듈은 B200 GPU당 약 2080억 개의 트랜지스터와 총 192GB(사용 가능 180GB) 메모리를 보유합니다. 비교하자면, Hopper H100 GPU는 약 800억 개의 트랜지스터와 80GB HBM3 메모리(Blackwell의 HBM3e 대비)를 탑재했습니다. 따라서 Blackwell의 B200은 트랜지스터 수를 두 배 이상 늘렸으며 메모리 용량은 약 2.4배 증가했습니다.

Blackwell의 두 GPU 다이 는 NV-HBI(High-Bandwidth Interface)라는 전용 고속 10TB/s 다이 간 인터커넥트를 통해 통신합니다. 이를 통해 모듈 내 두 GPU 다이는 단일 통합 GPU로 기능합니다. 상위 소프트웨어 계층에서는 단일 GPU로 인식됩니다.

시스템 관점에서 블랙웰 GPU는 하나의 단일 모듈, 즉 또는 장치로 인식됩니다. 이 장치에는 대용량 메모리 풀(192GB [실사용 가능 180GB] HBM3e)과 방대한 실행 유닛이 탑재되어 있지만, 내부적으로는 두 개의 칩이 협업하는 구조입니다. NVIDIA의 소프트웨어와 스케줄링은 두 GPU 다이 간 작업 부하를 균등하게 분배하고 메모리 접근이 일관되도록 보장합니다. 이로 인해 개발자는 NVIDIA가 의도한 대로 단일 GPU로 인식되므로 이러한 복잡성을 크게 무시할 수 있습니다.

각 Blackwell B200 GPU 모듈은 두 GPU 다이(각 96GB)에 걸쳐 결합된 192GB(사용 가능 180GB)의 HBM3e 메모리를 갖추고 있으며, 이는 8-Hi 스택으로 분할됩니다. 8-Hi HBM3e 스택은 각각 3GB 용량의 DRAM 다이 8개를 수직으로 적층하여 스택당 총 24GB를 구성합니다.

B200 GPU는 이러한 스택 8개(다이당 4개)를 사용하여 192GB(사용 가능 180GB, 192GB = 스택 8개 × 스택당 24GB)의 패키지 내 메모리를 제공합니다. 이는 이전 세대 Hopper GPU 대비 GPU당 스택 수와 용량을 증가시켜 모델 매개변수, 활성화, 기울기, 입력 데이터에 대한 여유 공간을 더 많이 확보합니다.

오류 정정 코드(ECC), 시스템 펌웨어 사용, 제조 제한 및 기타 문제로 인해 칩이 전체 192GB를 노출하지 못하기 때문에 B200당 192GB HBM3e 메모리 중 180GB만 사용 가능합니다. 따라서 Blackwell B200의 사용 가능 메모리는 전체 192GB가 아닌 180GB로 표기합니다.

메모리 속도도 향상되었습니다. Blackwell의 B200 HBM3e는 GPU당 최대 약 8TB/s의 통합 대역폭을 제공합니다. 비교를 위해, Hopper는 GPU당 약 3.35TB/s를 제공하는 이전 세대 HBM3를 사용합니다. 따라서 Blackwell의 메모리 대역폭 처리량은 Hopper보다 약 2.4배 높습니다.

초당 8테라바이트의 데이터를 공급받으며, 블랙웰 GPU 코어는 데이터 대기 시간으로 인한 빈번한 정지 없이 거대한 행렬 연산에 집중할 수 있습니다. 엔비디아는 온칩 캐싱도 강화했는데, 블랙웰은 총 126MB의 L2 캐시(다이당 63MB)를 탑재했습니다. 이 캐시는 GPU 내부에 위치한 소용량이지만 초고속 메모리로 최근 사용된 데이터를 보관합니다.

호퍼의 50MB L2 캐시 대비 2.5배 이상 증가한 L2 캐시 용량 덕분에 블랙웰은 더 많은 신경망 가중치나 중간 결과를 칩 내에 보관할 수 있어 HBM으로의 추가 데이터 이동을 피할 수 있습니다. 이는 GPU의 연산 유닛이 데이터 부족으로 가동 중단되는 상황을 최소화하는 데 기여합니다.

다음으로, 블랙웰 GPU가 전용 저정밀도 텐서 코어 세트와 어떻게 결합되는지 살펴보겠습니다. 또한 NVIDIA의 트랜스포머 엔진(Transformer Engine)이라 불리는 트랜스포머 최적화 하드웨어 및 소프트웨어 API도 함께 살펴보겠습니다. PyTorch와 같은 프레임워크와 vLLM과 같은 추론 엔진은 CUDA, CUTLASS, OpenAI의 Triton과 같은 라이브러리를 사용하여 이러한 최적화를 지원합니다. 이에 대해서는 후속 장에서 자세히 다룰 예정입니다.

이 책의 나머지 부분에서는 블랙웰의 듀얼 다이 GPU를 단순히 "블랙웰 GPU"로 지칭한다는 점을 기억하십시오.

NVIDIA GPU 텐서 코어와 트랜스포머 엔진

컴퓨트 유닛에 관해 말하자면, 블랙웰은 AI 워크로드를 특별히 겨냥한 개선 사항을 도입합니다. 그 핵심은 NVIDIA의 텐서 코어 기술과 트랜스포머 엔진(TE)입니다. 텐서 코어는 GPU의 각 스트리밍 멀티프로세서(SM) 내에 있는 특수 유닛으로, 매우 빠른 속도로 행렬 곱셈 연산을 수행할 수 있습니다.

텐서 코어는 이전 세대에도 존재했지만, 블랙웰의 텐서 코어는 8비트 및 4비트 부동 소수점과 같은 극히 낮은 정밀도를 포함한 더 많은 수치 형식을 지원합니다. 낮은 정밀도의 개념은 간단합니다. 숫자를 표현하는 데 더 적은 비트를 사용함으로써 동시에 더 많은 연산을 수행할 수 있습니다. 동일한 숫자를 표현하는 데 더 적은 비트가 사용되므로 메모리 효율성도 높아집니다. 이는 알고리즘이 수치 정밀도의 약간의 손실을 허용할 수 있다는 전제 하에 가능합니다. 요즘 많은 AI 알고리즘은 저정밀도 수치 형식을 염두에 두고 설계됩니다.

NVIDIA는 TE(Tensor Engine)를 통해 Deep Learning에서 혼합 정밀도 를 자동 조정 및 활용하는 방식을 선도했습니다. 핵심 레이어에는 높은 정밀도(FP16 또는 BF16)를, 덜 중요한 레이어에는 FP8을 사용하는 방식입니다. TE는 낮은 정밀도에서도 모델의 정확도를 유지하는 것을 목표로 정밀도의 균형을 자동으로 최적화합니다.

Hopper 세대에서 TE는 FP8 지원을 최초로 도입하여 FP16 대비 처리량을 두 배로 늘렸습니다. Blackwell은 FP8의 절반 비트 수를 사용하는 4비트 부동 소수점 형식인 NVIDIA FP4(NVFP4)를 도입하여 한 단계 더 발전시켰습니다. FP4는 매우 작아 FP8의 연산 처리량을 잠재적으로 두 배로 높일 수 있습니다. 그림 2-3은 FP16 대비 FP8 및 FP4의 상대적 속도 향상을 보여줍니다.

의 전체 NVL72 랙(72개 GPU)은 4비트 정밀도에서 이론상 1.4 엑사플롭스(1.4 $\times 10^{18}$) 이상의 텐서 코어 처리량을 가집니다. 이는 FP4 정밀도가 낮음에도 불구하고 이 단일 랙을 세계 최고 속도의 슈퍼컴퓨터 영역에 올려놓는 놀라운 수치입니다. 실제 워크로드가 항상 그 피크 성능에 도달하지는 않더라도, 그 잠재력이 존재한다는 사실 자체가 놀랍습니다.

현대 GPU는 향상된 스케일링 및 보정 기능과 함께 NVFP4 지원을 추가한 TE를 사용합니다. 실제로는 PyTorch와 같은 프레임워크에서 TE의 커널과 모듈을 활용하여 채택합니다. 이렇게 하면 정확도를 유지할 때 FP8 및 NVFP4가 적용됩니다. 이는 모든 프레임워크에서 완전히 자동화된 레이어별 결정은 아닙니다.

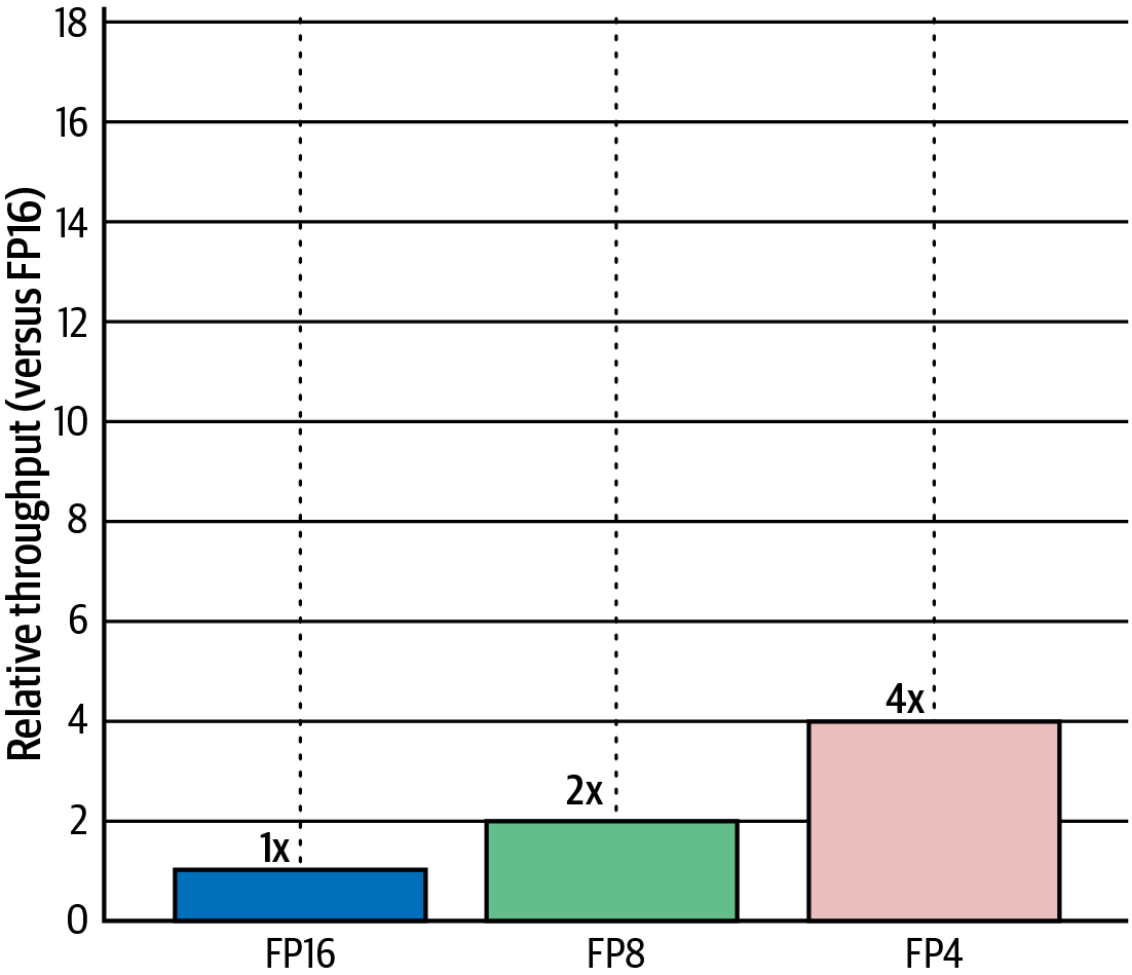


그림 2-3. FP16 대비 FP8 및 FP4의 상대적 속도 향상률

고급 기법에는 훈련 및 추론 과정에서 신경망의 각 레이어에 대해 정밀도를 동적으로 변경하는 것이 포함됩니다. 목표는 각 레이어에서 모델 정확도를 유지할 수 있는 최저 정밀도를 사용하는 것입니다. 예를 들어, 초기 레이어는 노이즈에 민감할 수 있으므로 신경망의 첫 번째 레이어는 FP16으로 유지할 수 있습니다. 그러나 경험적 규칙에 따라, 후속 레이어나 높은 정밀도가 중요하지 않은 거대한 임베딩 행렬에는 FP8이나 FP4를 사용할 수 있습니다.

이 모든 과정은 NVIDIA 라이브러리나 PyTorch 같은 AI 프레임워크 내부에서 자동으로 처리됩니다. 사용자는 혼합 정밀도만 활성화하면 되며, 그 결과는 사실상

'무료로' 얻는 엄청난 속도 향상입니다. 혼합 정밀도에 대해서는 9장에서 자세히 다루겠지만, 현재 많은 LLMs이 바로 이 이유로 혼합 정밀도를 사용한다는 점만 알아두세요. 이러한 정밀도 감소는 FP16 및 FP32 대비 훈련 속도를 향상시키면서도 정확도 손실을 줄입니다. Blackwell은 FP8과 FP4를 효율적으로 활용할 수 있도록 설계되었습니다.

이러한 정밀도 감소 형식은 메모리 사용량도 줄입니다. FP4 사용 시 FP8 대비 매개변수당 필요한 메모리가 절반으로 감소하며(FP8은 FP16 메모리 사용량을 절반으로 줄임), 이는 GPU 메모리에 더 큰 모델을 담을 수 있음을 의미합니다.

NVIDIA는 AI의 미래가 저정밀도 연산에 있다고 판단하고 Blackwell이 이를 탁월하게 수행할 수 있도록 설계했습니다. 이는 처리량(초당 토큰 수)과 지연 시간이 가장 중요한 대규모 모델의 추론 서비스에 특히 중요합니다.

호퍼에서 블랙웰로의 세대적 도약을 설명하기 위해, 엔비디아는 H100 기반 시스템이 1.8조 매개변수 규모의 대규모 MoE 모델에서 GPU당 초당 약 3.4 토큰만 생성할 수 있으며 첫 토큰에 5초 이상의 지연 시간이 발생한다고 보고했습니다. 이는 대화형 사용에는 너무 느립니다.

반면 Blackwell 기반 시스템(NVL72)은 동일한 모델을 GPU당 초당 약 150토큰으로 실행했으며, 첫 토큰 지연 시간은 약 50밀리초로 낮았습니다. 이는 Hopper 세대 대비 실시간 처리량에서 약 30배 개선된 수치입니다. NVL72는 이 대규모 모델이 실시간 응답을 제공할 수 있게 하여 훨씬 더 많은 저지연 사용 사례로 확장 가능하게 했습니다.

이러한 속도 향상은 순수한 FLOPS 성능, 더 빠른 GPU, 낮은 정밀도(FP4) 활용, 그리고 GPU에 데이터를 지속적으로 공급하는 NVLink 상호연결의 조합에서 비롯되었습니다. 이는 컴퓨팅과 통신을 아우르는 통합 설계가 실제 성능 향상으로 이어질 수 있음을 보여줍니다.

본질적으로 블랙웰 GPU는 이전 세대보다 더 강력하고, 더 스마트하며, 더 효율적으로 데이터를 공급받습니다. 텐서 코어, TE, 저정밀도 덕분에 수학적 연산을 더 빠르게 처리합니다. 또한 시스템 아키텍처는 거대한 메모리 대역폭, 대용량 캐시, NVLink 덕분에 데이터가 신속하게 제공되도록 보장합니다.

이어서 GPU 내부의 계층 구조를 간단히 살펴보겠습니다. 이는 후속 성능 튜닝 이해에 유용합니다. 스트리밍 멀티프로세서( )

스트리밍 멀티프로세서(SM), 스레드, 워프

는 이전 세대와 마찬가지로 각 GPU가 다수의 스트리밍 멀티프로세서(SM)로 구성됩니다. 이는 그림 2-4에 표시된 것처럼 GPU의 "코어"라고 생각하면 됩니다.

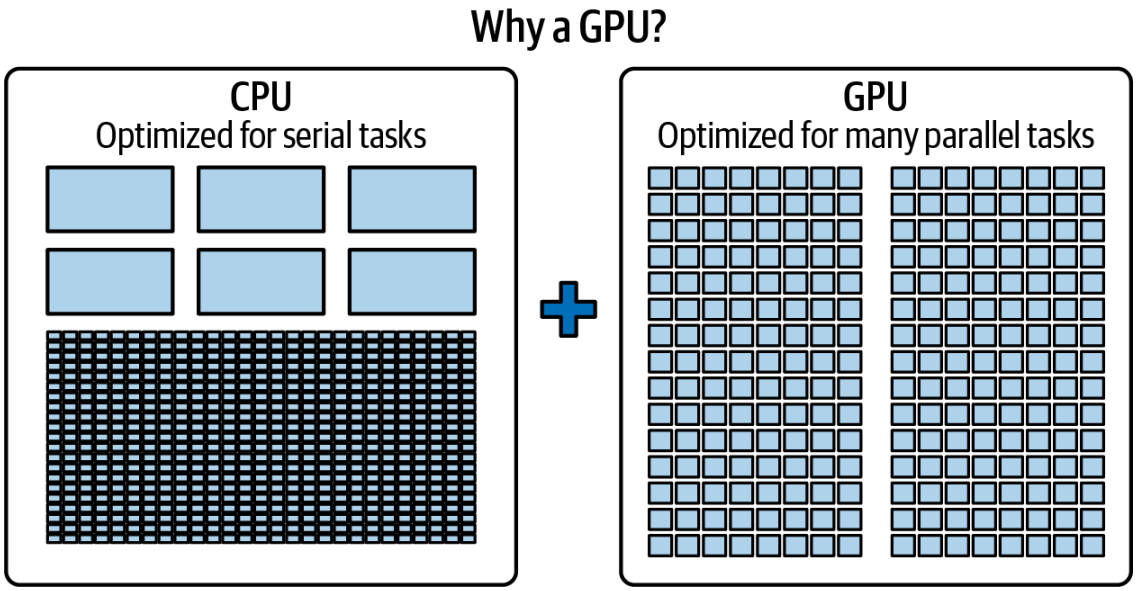


그림 2-4. CPU 코어와 GPU 코어비교 (출처: *https://oreil.ly/003EH*, *https://oreil.ly/Z25Tf*)

각 SM은 FP32, INT32 등을 위한 산술 연산 장치, 행렬 연산을 위한 텐서 코어, 메모리 작업을 위한 로드/스토어 장치, 초월 함수 연산 등을 위한 특수 기능 장치 등을 다수 포함합니다. GPU는 또한 레지스터, 공유 메모리, L1 캐시 등을 포함한 자체 초고속 메모리 풀을 보유하고 있습니다.

SM은 스레드 를 고정 크기 그룹인 *워프*(warp) 단위로 실행합니다. 각 워프는 정확히 32개의 스레드로 구성되며, 이들 스레드는 동일한 명령어를 동기화 상태로 동시에 실행합니다 . 이를 단일 명령어 다중 스레드(SIMT) 실행 모델이라 합니다.

SM은 글로벌 메모리에서 액세스된 데이터를 기다리는 스레드의 지연 시간을 상쇄하기 위해 많은 활성 워프 를 병렬로 실행합니다. 수십 개의 워프(수백 개의 스레드)가 동시에 실행 중인 SM을 생각해 보십시오. 한 워프가 메모리 페치 대기 중일 때 다른 워프가 실행될 수 있습니다. 이를 *지연 시간 숨김(latency hiding)*이라고 합니다. 이 책에서 지연 시간 숨김을 여러 번 다시 살펴볼 것입니다. 이는 튜닝 도구 상자에 반드시 포함해야 할 매우 중요한 성능 최적화 도구입니다.

Blackwell과 같은 하이엔드 GPU에는 수백 개의 SM이 탑재됩니다. 각 SM은 수천 개의 스레드를 동시에 실행할 수 있습니다. 이렇게 해서 단일 GPU에 수만 개의 활성 스레드를 구현하는 것입니다. 앞서 언급한 바와 같이, 이 모든 SM은 126MB L2 캐시를 공유하며 HBM에 연결되는 메모리 컨트롤러도 공유합니다. 메모리 계층 구조는 그림 2-5와 같이 레지스터(스레드별) → 공유 메모리(스레드

블록별, 각 SM 내) → L1 캐시(SM별) → L2 캐시(GPU 내 모든 SM이 공유) → HBM 메모리(오프칩)로 구성됩니다.

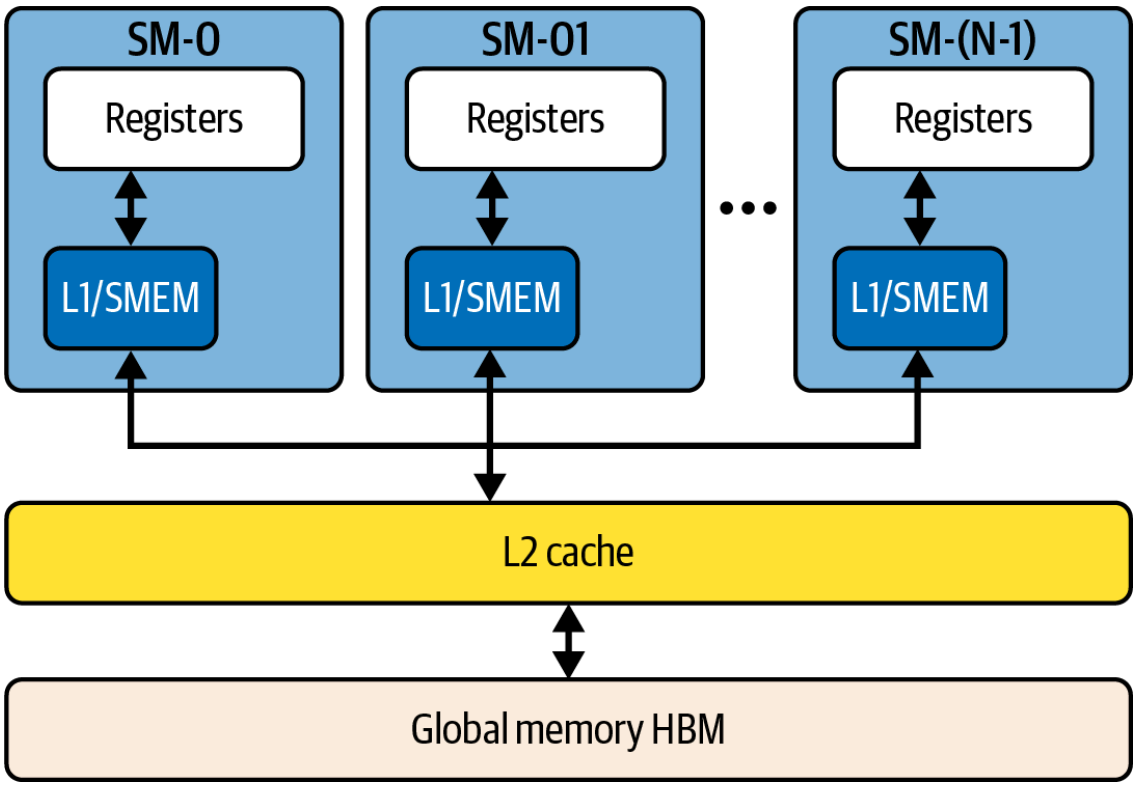


그림 2-5. GPU 메모리 계층 구조

최상의 성능을 위해 데이터는 이 계층 구조에서 가능한 한 높은 위치에 머물러야 합니다. 모든 연산이 8TB/s 속도의 HBM으로 전송된다면, 오프칩 메모리 접근 지연 증가로 인해 GPU가 너무 자주 정지할 것입니다. 재사용 가능한 데이터를 SM 로컬 메모리나 L2 캐시에 보관함으로써 GPU는 엄청난 처리량을 달성할 수 있습니다. Blackwell 아키텍처의 캐시 및 대역폭 두 배 증가는 바로 이 GPU 괴물을 계속 먹여 살리고 만족시키기 위한 것입니다.

성능 엔지니어로서 우리는 커널 성능이 컴퓨팅뿐만 아니라 메모리 트래픽과 처리량에 의해 바운디드되는 수많은 사례를 목격하게 될 것입니다. NVIDIA는 블랙웰을 설계하면서 많은 AI 워크로드에서 FLOPS와 메모리 대역폭 간의 균형이 잘 맞도록 분명히 고려했습니다.

블랙웰의 설계는 연산과 메모리를 균형 있게 조정하여 많은 AI 커널에서 GPU가 최소한의 정지 상태로 연산을 지속할 수 있도록 합니다. 실제로 잘 최적화된 고밀도 수학 연산은 온칩 메모리의 데이터를 재사용하여 심각한 메모리 제약 없이 최대 FLOPS에 근접할 수 있습니다.

이 모든 것은 최적화된 코드가 주어지면 GPU가 데이터 대기보다 계산에 바쁘게 움직인다는 것을 의미합니다. 대규모 축소 연산이나 무작위 메모리 접근과 같은 특정 작업은 여전히 메모리 바운디드가 발생할 수 있으나, 업데이트된 GPU, 메모리 및 상호 연결 하드웨어로 인해 이 문제가 다소 완화됩니다.

초대규모 네트워킹: 다수의 GPU를 하나의 장치로 통합

두 개의 GPU와 CPU 를 슈퍼칩에 통합함으로써 우리는 놀라울 정도로 강력한 노드를 확보했습니다. 다음 과제는 이러한 슈퍼칩들을 다수 연결하여 더 큰 규모의 모델 훈련으로 확장하는 것입니다.

NVIDIA는 GB200/GB300 슈퍼칩을 활용한 대규모 랙 구성 인 NVL72 시스템을 제공합니다. NVL72는 72개의 Blackwell GPU와 36개의 Grace CPU로 구성된 시스템을 의미하며, 모두 NVLink로 상호 연결됩니다. 이는 본질적으로 단일 랙에 담긴 AI 슈퍼컴퓨터입니다.

GB200/GB300 NVL72는 18개의 컴퓨팅 노드로 구성된 로, 각 노드에는 두 개의 GB200/GB300 슈퍼칩이 포함되어 있어 컴퓨팅 노드당 총 4개의 Blackwell GPU와 2개의 Grace CPU를 갖추고 있습니다(그림 2-6 참조).

그림 2-6. GB200/GB300 NVL72 랙 내부의 1U 컴퓨팅트레이와 두 개의 Grace Blackwell Superchip (출처: developer.nvidia.com)

여기서 각 슈퍼칩 모듈에는 하나의 Grace CPU와 두 개의 Blackwell GPU(각 B200은 듀얼 다이 MCM)가 있습니다. NVL72는 이러한 트레이 18개를 서로 연

결합니다. 18개의 컴퓨팅 노드를 연결함으로써 GB200/GB300 NVL72는 72개의 Blackwell GPU(18개 노드 × 4개 GPU)와 36개의 Grace CPU(18개 노드 × 2개 CPU)를 하나로 묶어 강력한 통합 CPU-GPU 클러스터를 형성합니다.

NVL72의 흥미로운 점은 모든 GPU가 단일 NVLink 도메인 내에서 NVLink 스위치 패브릭을 통해 다른 모든 GPU와 초고속으로 통신할 수 있다는 것입니다. NVIDIA는 GPU에 탑재된 NVLink 5 연결과 NVSwitch라는 전용 스위치 실리콘을 조합하여 이를 구현했습니다.

NVLink 및 NVSwitch

각 Blackwell GPU는 18개의 NVLink 5 포트를 노출합니다( ). NVL72가 모든 포트를 NVLink 스위치 시스템에 연결함으로써, 집계 양방향 NVLink 대역폭은 GPU당 1.8TB/s(18개 NVLink 링크 × 양방향 100GB/s)를 제공합니다. 각 NVLink 스위치 트레이는 100GB/s 속도로 144개의 NVLink 포트를 제공합니다. 9개의 트레이 전체에서 각 GPU의 18개 NVLink 5 링크는 NVSwitch 칩당 하나씩 배선되어 72개의 GPU가 전체 분할 대역폭으로 완전히 연결됩니다. 집계 양방향 NVLink 5 대역폭은 GPU당 1.8TB/s(18개 NVLink 링크 × 양방향 100GB/s)입니다.

이는 Hopper GPU가 사용한 이전 세대 대비 GPU당 NVLink 대역폭이 두 배입니다. Hopper H100은 18개의 NVLink 4 포트를 사용하지만 NVLink 5의 절반 속도로 작동합니다. NVLink를 통한 GPU 간 지연 시간은 한 자릿수 마이크로초 범위입니다.

GPU들은 NVSwitch 칩을 통해 네트워크로 연결됩니다. NVSwitch는 네트워크 스위치와 유사한 스위칭 칩이지만, NVLink 전용으로 설계되었습니다. 이는 NVLink 스위치 시스템 내에서 모든 GPU가 단일 스위치 단계를 통해 다른 모든 GPU에 완전한 분할 대역폭으로 접근할 수 있음을 의미합니다. 이 단일 단계 특성은 단일 NVL72 랙 내에서 유효합니다. 각 GPU가 18개의 NVLink 링크를 사용하여 18개의 NVSwitch 칩에 연결되므로 단일 스위치를 통한 경로가 가능하기 때문입니다. 그림 2-7은 NVL72에 사용되는 NVLink 스위치 트레이를 보여줍니다.

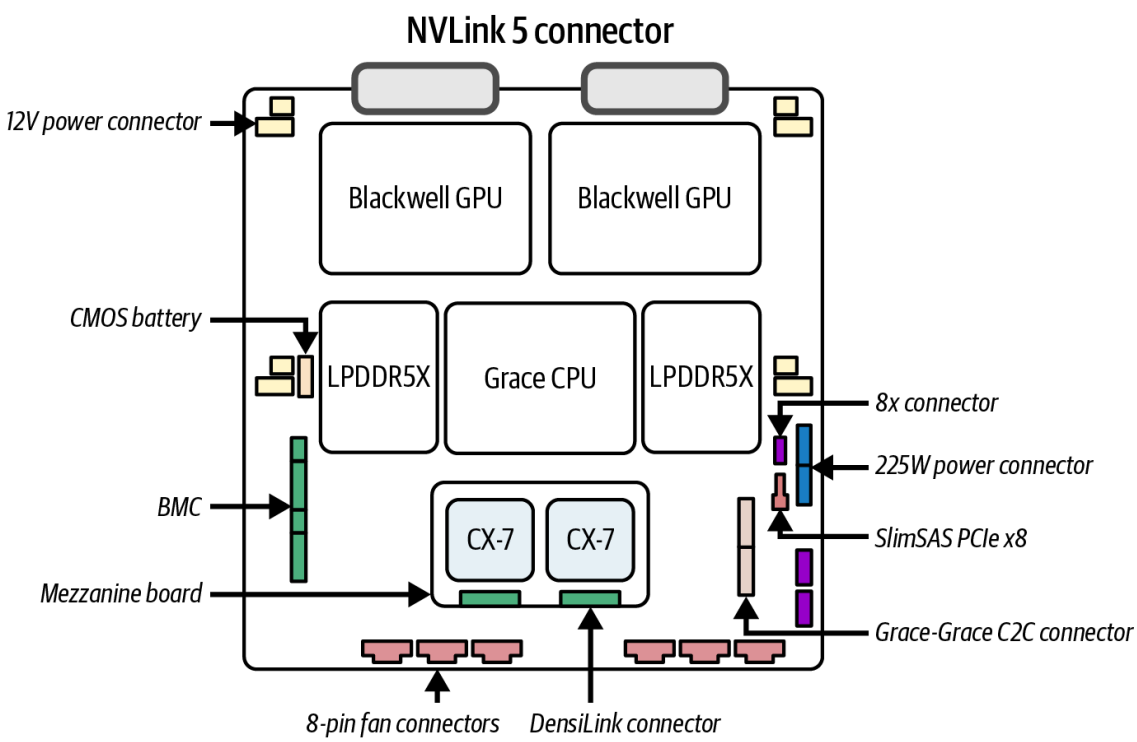
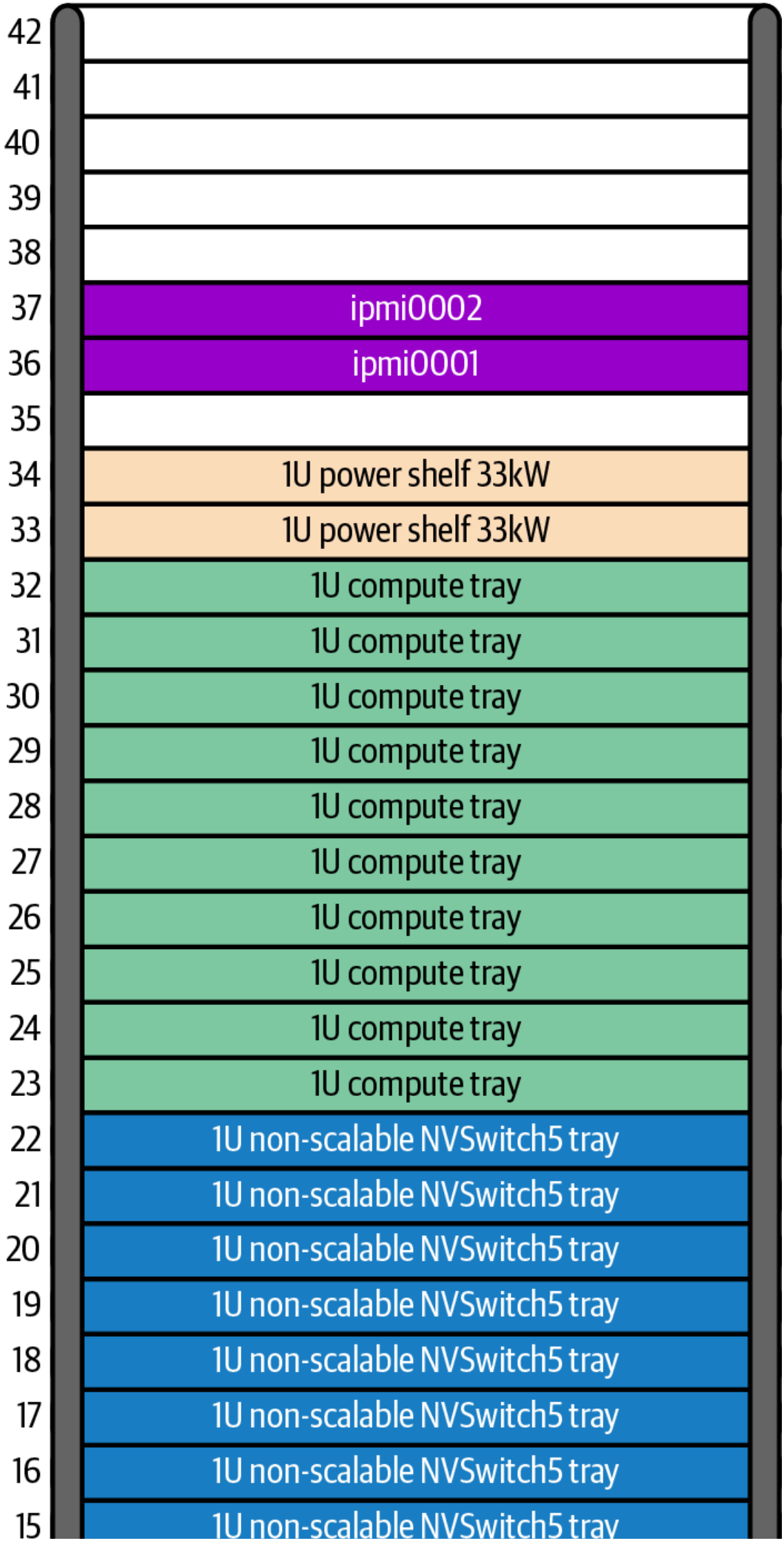


그림 2-7. NVL72 내부의 NVLink 스위치 트레이1개 (출처: https://oreil.ly/h7seG)

각 스위치 트레이에는 두 개의 NVSwitch 칩과 여러 개의 고속 포트가 포함됩니다. NVL72 랙은 그림 2-8과 같이 9개의 이러한 스위치 트레이와 18개의 컴퓨트 트레이로 구성됩니다.


42
41
40
39
38
37
ipmi0002
36
ipmi0001
35
34
1U power shelf 33kW
33
1U power shelf 33kW
32
1U compute tray
31
1U compute tray
30
1U compute tray
29
1U compute tray
28
1U compute tray
27
1U compute tray
26
1U compute tray
25
1U compute tray
24
1U compute tray
23
1U compute tray
22
1U non-scalable NVSwitch5 tray
21
1U non-scalable NVSwitch5 tray
20
1U non-scalable NVSwitch5 tray
19
1U non-scalable NVSwitch5 tray
18
1U non-scalable NVSwitch5 tray
17
1U non-scalable NVSwitch5 tray
16
1U non-scalable NVSwitch5 tray
15
1U non-scalable NVSwitch5 tray

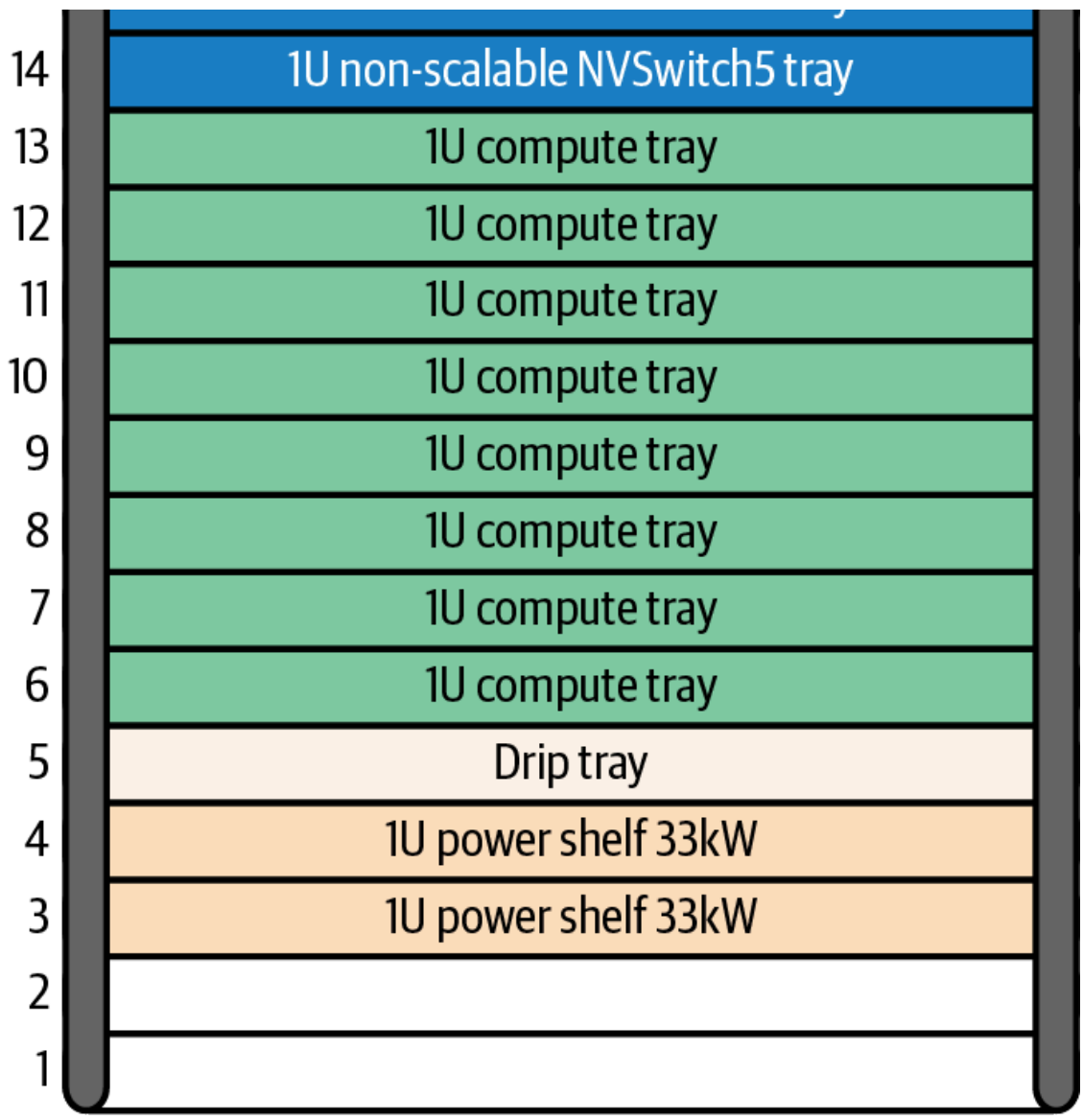


그림 2-8. NVL72 랙 내부의 9개 트레이로 구성된NVSwitch 시스템 (출처: *https://oreil.ly/h7seG*)

9개의 스위치 트레이 각각에 두 개의 NVSwitch 칩이 포함되므로, NVL72 시스템에는 총 18개의 NVSwitch 칩이 있습니다. 네트워크는 풀 크로스바(full cross-bar) 형태로 구성되어 모든 GPU가 모든 NVSwitch에 연결되고, 모든 NVSwitch가 모든 GPU에 연결됩니다. 이는 어떤 두 GPU 사이에도 고대역폭 경로를 제공합니다.

각 스위치 트레이는 144개의 NVLink 포트를 노출하여 각 GPU의 18개 NVLink 링크를 완전히 연결합니다. 구체적으로, 각 GPU는 자체 18개 NVLink 링크를 사용하여 18개 NVSwitch 칩(각 스위치당 1개 링크)에 연결됩니다. 이는 모든 GPU가 다른 GPU에 1홉(GPU → NVSwitch → GPU)으로 도달할 수 있으며, 그 과정에서 엄청난 대역폭을 확보할 수 있음을 의미합니다. 그림 2-9는 72개의 완전히 연결된 GPU(36개의 GB200 슈퍼칩)와 18개의 NVSwitch로 구성된 전체 NVL72 아키텍처를 보여줍니다.

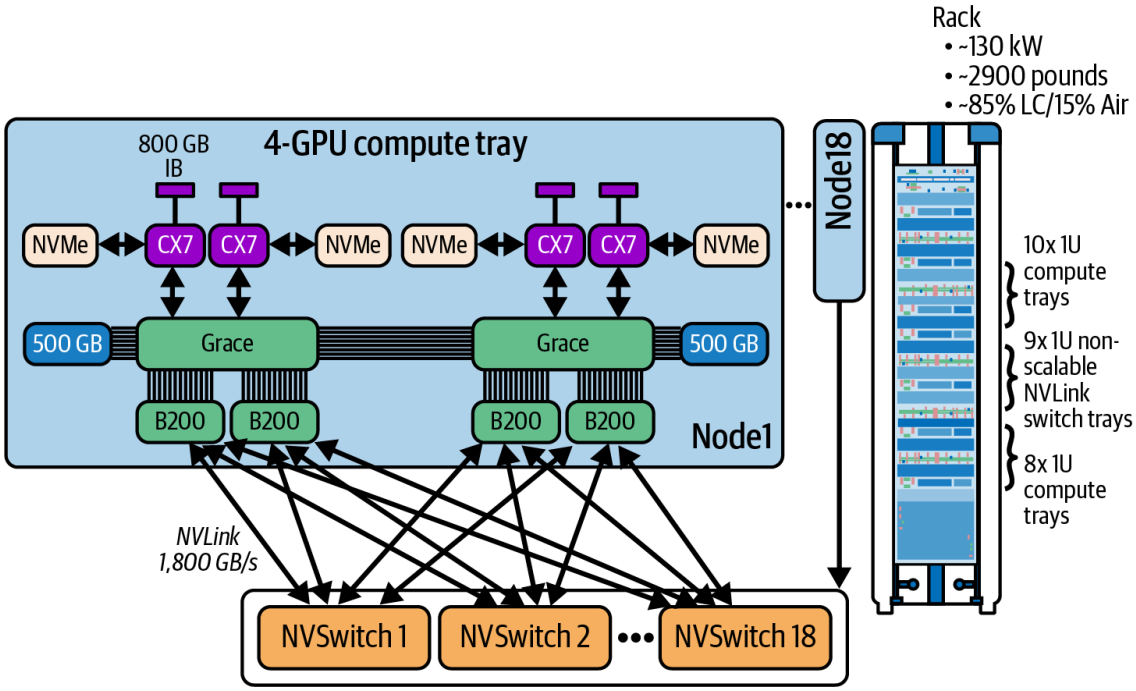


그림 2-9. 각 GPU는 각 NVSwitch에 연결됨 (스위치당 하나의 링크)

전체 72개 GPU 네트워크를 가로지르는 총 분할 대역폭은 NVL72 랙 내에서 약 130TB/s입니다. 비교를 위해 말하자면, 이는 유사한 규모의 최상위 인피니밴드 클러스터보다도 몇 배나 높은 수치입니다. 이 설계는 GPU 전반에 걸친 글로벌 주소 공간을 갖춘 완전 연결형 고대역폭 패브릭을 제공합니다. 이를 통해 동기화와 일관성에 대한 명시적인 소프트웨어 제어를 유지하면서 효율적인 집단 연산 및 일방적 연산을 수행할 수 있습니다.

다중 GPU 프로그래밍

프로그래밍 모델 관점에서, 각 GPU는 피어 투 피어(peer-to-peer) 및 NVIDIA SHMEM(NVSHMEM)과 같은 분할 글로벌 주소 공간(PGAS) 모델을 사용하여 NVLink를 통해 다른 GPU의 메모리에 직접 접근할 수 있습니다. 글로벌 주소 공간은 존재하지만, GPU 캐시는 GPU 간에 전역적으로 일관성을 유지하지 않습니다. NVLink-C2C를 통한 CPU-GPU 경로만 캐시 일관성을 유지합니다. NCCL 및 NVSHMEM과 같은 소프트웨어 스택은 올바른 다중 GPU 액세스에 필요한 동기화 및 순서를 제공합니다. 하드웨어 캐시 일관성과 소프트웨어 동기화 기술이 결합되어 NVL72는 본질적으로 하나의 대형 GPU로 인식될 수 있습니다.

원격 직접 메모리 액세스(RDMA)는 InfiniBand 및 RDMA over Converged Ethernet(RoCE) 전송 계층을 통해 호스트 간 직접적인 제로-카피 메모리 전송을 가능하게 하는 네트워크 기술입니다. 선택적 원격 원자적 작업은 InfiniBand 및 RoCE를 위해 InfiniBand Trade Association(IBTA) 에 의해 정의됩니다.

NVIDIA의 RDMA 프로토콜 구현인GPUDirect RDMA는 네트워크 인터페이스 컨트롤러(NIC)가 GPU 메모리를 등록하고 nvidia-peermem 드라이버를 사용하

여 GPU 메모리와 직접 RDMA를 수행할 수 있도록 합니다. 이를 통해 GPU는 CPU를 거치지 않고 노드 간에 데이터를 교환하고 원자적 작업을 실행할 수 있습니다. 또한 NIC는 호스트 RAM을 경유하지 않고 GPU 메모리와 직접 DMA를 수행할 수 있습니다.

원격 원자 연산 및 노드 간 단방향 연산은 NVSHMEM과 같은 상위 계층 라이브러리를 통해 제공되며, 이 라이브러리들은 RDMA 전송 계층 위에 해당 의미론을 구현합니다. GPUDirect RDMA는 원자 API 자체보다는 직접적인 데이터 경로를 제공한다는 점에 유의하십시오. 분산 훈련 및 추론 워크로드는 다수의 GPU 간에 정보를 자주 동기화하고 교환해야 합니다.

기존에는 GPU가 서로 다른 컴퓨팅 노드와 랙에 배치되었습니다. 따라서 동기화는 InfiniBand나 이더넷과 같은 상대적으로 느린 네트워크 링크를 통해 이루어졌습니다. 이는 대규모 AI 모델을 지원하기 위해 다수의 GPU로 확장할 때 흔히 병목 현상이 발생합니다.

NVL72 시스템에서는 이러한 교환이 NVLink와 NVSwitch를 통해 초고속으로 이루어집니다. 이는 최소한의 통신 오버헤드로 훈련 작업이나 추론 클러스터를 최대 72개의 GPU까지 확장할 수 있음을 의미합니다. 또한 GPU들이 서로의 데이터를 기다리는 시간이 훨씬 줄어들기 때문에, 전체 처리량은 72개의 GPU까지 거의 선형적으로 확장됩니다.

반면, 동일한 작업을 8개의 Hopper H100 GPU를 탑재한 9개의 별도 컴퓨팅 서버로 구성된 유사한 규모의 72-GPU H100 클러스터에 확장하는 경우를 고려해 보십시오. 이 구성은 InfiniBand를 필요로 하며, 이는 네트워크 병목 현상을 발생시켜 클러스터의 확장 효율성을 크게 저하시킵니다.

구체적인 수치를 통해 NVL72와 72-GPU H100 클러스터를 분석하고 비교해 보겠습니다. 단일 NVL72 랙 내에서 GPU 간 대역폭은 GPU당 최대 1.8TB/s(양방향 합산)이며, 킬로바이트 단위의 소규모 메시지 전송 시 지연 시간은 1~2마이크로초 수준입니다. 대용량 메시지는 더 오래 걸리며 일반적으로 대역폭에 의해 제한됩니다. 기존 인피니밴드 네트워크에서는 GPU당 대역폭이 NIC 수와 속도에 따라 20~80GB/s 수준이며, 지연 시간은 5~10마이크로초 이상일 가능성이 높습니다.

NVL72 네트워크는 호스트 NIC 패브릭 대비 GPU당 대역폭이 현저히 높고 지연 시간이 낮습니다. 구체적으로 NVLink 5는 GPU당 약 1.8TB/s의 집계 대역폭을 제공하는 반면, 최신 호스트 NIC는 400~800Gb/s 회선 속도에서 포트당 약 50~100GB/s를 제공합니다. 이로 인해 집단 연산 오버헤드가 수십 퍼센트에서 불과 몇 퍼센트로 감소합니다.

실질적으로 NVLink로 연결된 NVL72 시스템 내에서는 기존 노드 간 패브릭 대비 집단 오버헤드가 현저히 낮지만, 정확한 반복 시간 비율은 워크로드에 따라 달라집니다. 예를 들어, NVIDIA는 1조 8천억 매개변수 MoE 모델이 H100에서 GPU당 초당 약 3.4토큰, 첫 토큰까지 약 5초가 소요되던 것이 GB200 NVL72에서는 GPU당 초당 약 150토큰, 첫 토큰까지 약 50ms가 소요되는 것으로 개선되었다고 보고했습니다. 이러한 속도 향상은 Blackwell의 높은 컴퓨팅 처리량 외에도 NVL72 랙 내부에서 GPU 간 통신 병목 현상을 제거한 데 기인합니다.

단일 NVL72 랙 내에서 통신은 매우 빠르기 때문에 통신 병목 현상은 거의 완전히 제거되어 낮은 우선순위가 됩니다. 반면 기존 InfiniBand 및 이더넷 클러스터에서는 통신이 종종 주요 병목 현상이 되어 소프트웨어 수준에서 신중한 최적화와 튜닝이 필요합니다.

요약하면, 고속 NVLink 및 NVSwitch 하드웨어의 이점을 활용하기 위해 워크로드의 통신을 가능한 한 랙 내부("랙 내부")에 유지함으로써 NVL72 구성을 최대한 활용하는 소프트웨어를 설계하고 구현해야 합니다. NVL72의 컴퓨팅 및 메모리 리소스를 넘어 확장해야 할 경우에만 랙 간("랙 간") 느린 InfiniBand 또는 이더넷 기반 통신을 사용하십시오.

NVIDIA SHARP를 활용한 네트워크 내 집계

하드웨어 기반 최적화 기술 의 또 다른 예는 NVIDIA 확장형 계층적 집계 및 축소 프로토콜(SHARP)입니다. NVLink 스위치 시스템 랙의 경우, 네트워크 내 축소 작업은 NVSwitch ASIC에 통합된 SHARP 엔진을 활용하여 축소 및 기타 집합 연산을 네트워크 내에서 오프로드합니다(그림 2-10 참조).

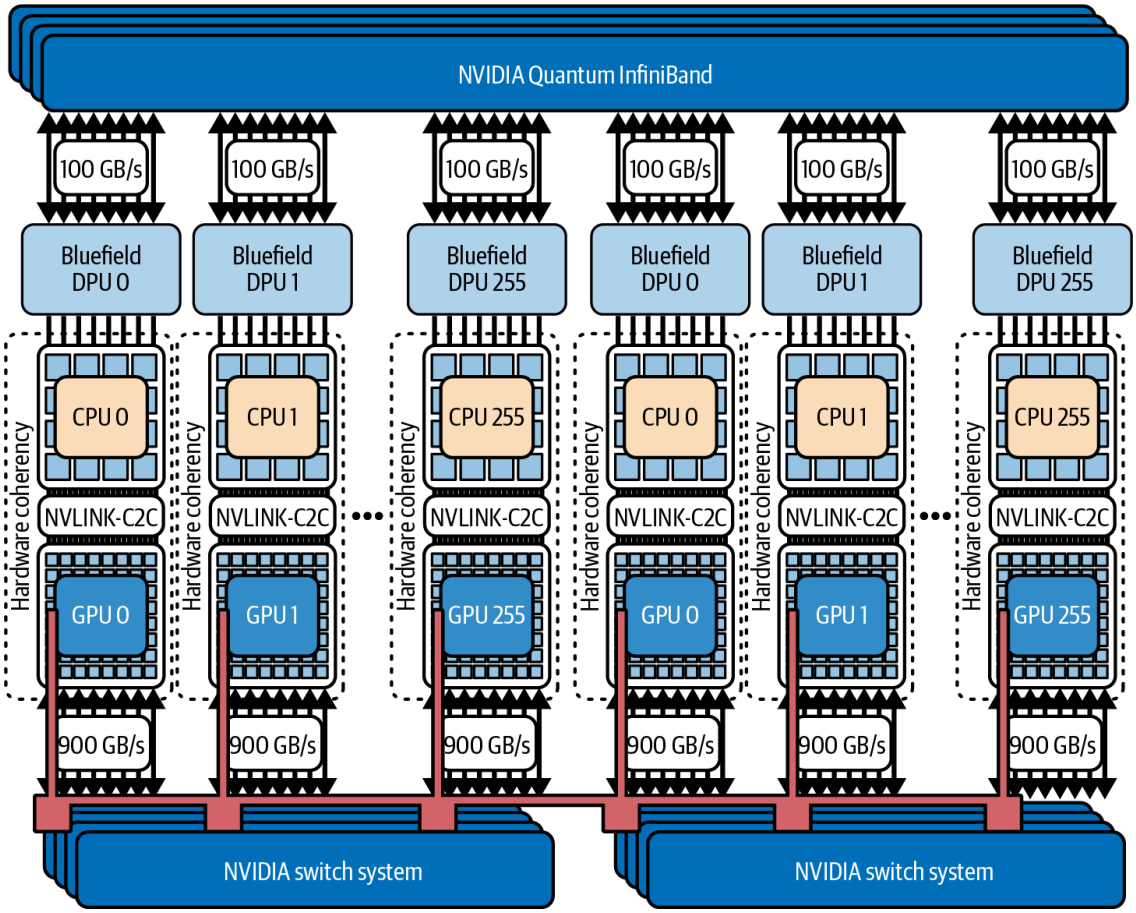


그림 2-10. NVSwitch 내 SHARP 축소 엔진을 사용한 NVIDIA 네트워크 하드웨어로의 연산오프로딩

NVSwitch 패브릭은 데이터가 GPU를 거치지 않고도 부분 결과를 결합합니다. GPU에서 스위치 하드웨어 자체로 집단 연산을 오프로드함으로써 SHARP는 GPU가 더 복잡한 연산에 집중할 수 있게 하고, 집단 연산 지연 시간을 낮추며, 네트워크를 통과하는 전체 데이터 양을 줄이고, 시스템 효율성을 높입니다.

SHARP의 향상된 효율성은 분산 훈련 중 기울기 집계나 매개변수 동기화와 같은 중량 작업을 NVSwitch의 전용 SHARP 엔진이 처리함을 의미합니다. 그 결과 랙 내부 및 랙 간 구성 모두에서 훨씬 더 효율적인 확장성을 제공합니다.

SHARP를 사용하면 GPU 수가 증가해도 거의 선형적인 성능 향상을 확인할 수 있습니다. 이러한 네트워크 내 컴퓨팅 기능은 초거대 모델 훈련 시 특히 중요합니다. 집단 연산에서 절약되는 모든 마이크로초가 상당한 전체 속도 향상으로 이어질 수 있기 때문입니다.

SHARP( )는 NVIDIA가 2019~2020년 Mellanox 인수 과정에서 획득한 가장 영향력 있는 혁신 기술 중 하나입니다. 현재 SHARP를 사용하지 않는다면 반드시 살펴보시기 바랍니다. SHARP는 콜렉티브 작업의 지연 시간과 트래픽을 크게 줄일 수 있으며, 통신에 제약받는 훈련의 확장 효율성을 종종 향상시킵니다.

멀티랙 및 스토리지 통신

다음으로 NVL72 랙 이 다른 NVL72 랙 또는 공유 파일 시스템과 같은 외부 스토리지 시스템과 어떻게 통신하는지 살펴보겠습니다. 앞서 설명한 바와 같이 NVL72 랙 내부에서는 NVLink가 모든 GPU 간 트래픽을 처리합니다. 그러나 랙 외부에서는 보다 전통적인 네트워킹 하드웨어에 의존합니다.

NVL72의 각 컴퓨팅 노드에는 고속 네트워크 인터페이스 카드(NIC)와 데이터 처리 장치(DPU)가 장착됩니다. DPU는 네트워킹, 스토리지, 보안 및 관리 작업을 호스트 CPU에서 오프로드하고 가속화하며 분리합니다. 이러한 작업을 NIC에서 직접 실행함으로써 DPU는 CPU 오버헤드와 지연 시간을 줄입니다.

NVL72 설계에서 BlueField-3 DPU는 라인 속도 패킷 처리, RDMA 및 NVMe over Fabrics(oF) 작업을 처리합니다. NVMe-oF는 네트워크 패브릭 전반에 스토리지를 확장하는 NVMe의 프로토콜 변형입니다. 따라서 DPU는 CPU 개입 없이 네트워크, 스토리지 및 GPU 메모리 간에 데이터를 직접 이동시킵니다. 이는 전체 시스템 처리량과 효율성을 극대화합니다.

GB200/GB300 NVL72 랙은 Quantum-X800 InfiniBand 또는 Spectrum-X800 이더넷 패브릭과 통합됩니다. 컴퓨트 트레이는 일반적으로 높은 외부 대역폭을 위해 노드당 ConnectX-8 800Gb/s NIC 4개를 사용합니다. BlueField-3 DPU는 스토리지, 보안 및 제어 평면 작업에 네트워크 내 가속화 또는 오프로드가 필요한 곳에 사용됩니다.

800Gb/s NIC 4개를 사용하면 컴퓨팅 노드당 3.2Tbit/s, 랙당 약 57.6Tbit/s(57.6Tbit/s = 노드당 3.2Tbit/s × 18노드)의 성능을 제공합니다. 이 처리량은 놀라울 정도로 높지만, 랙을 벗어나면 여전히 초고속 네트워크가 필요하다는 점을 기억해야 합니다. 이를 통해 멀티랙 확장 시 랙 경계에서 병목 현상이 발생하지 않습니다. NVIDIA는 이러한 멀티랙 배포 *환경을 'AI 팩토리'라* 명명했습니다. 또한 NVL72가 노드당 4개의 NIC를 활용해 더 큰 네트워크 패브릭에 연결될 수 있도록 설계했습니다.

각 노드의 BlueField-3 DPU는 RDMA, TCP/IP, NVMe SSD 스토리지 액세스와 같은 네트워킹 작업을 오프로드하는 데 도움을 줍니다. 이를 통해 Grace CPU가 네트워크 인터럽트 관리에 매달리지 않도록 합니다. DPU는 본질적으로 스마트 네트워크 컨트롤러 역할을 하며, NVIDIA의 GPUDirect RDMA 소프트웨어를 사용하여 NIC와 GPU 메모리 간에 데이터를 직접 이동시킵니다. 이는 호스트 메모리를 통한 데이터 스테이징이나 CPU 사이클 사용이 필요하지 않습니다.

BlueField DPU는 CPU 개입을 피하므로, 대규모 훈련 작업을 위해 스토리지 서버에서 대용량 데이터셋을 스트리밍할 때 특히 유용합니다. 구체적으로 DPU는 데이터 전송을 처리하고 GPU 메모리에 직접 저장하는 동안 CPU는 데이터 전처리 같은 다른 작업에 집중할 수 있습니다.

성능 오프로드 기능 외에도 DPU는 안전한 멀티테넌시를 지원합니다. 서로 다른 작업과 사용자의 네트워크 트래픽을 분리하여 노드 상에서 스마트 방화벽/스위치 역할을 수행합니다.

여러 NVL72 랙으로 확장할 때 NVIDIA는 Quantum 시리즈 InfiniBand 스위치를 사용합니다. 이러한 InfiniBand 스위치를 통해 여러 NVL72 랙을 상호 연결하여 대규모 NVL72 랙 클러스터를 구성할 수 있습니다.

예를 들어, 총 576개의 GPU를 갖춘 8개의 랙 NVL72는 NVLink 스위치 시스템을 사용하여 하나의 NVLink 5 도메인으로 연결됩니다. 그런 다음 InfiniBand 또는 이더넷을 사용하여 해당 NVLink 도메인을 다른 도메인(예: 다른 NVL72 랙)이나 외부 스토리지에 연결합니다(단, 랙 간 InfiniBand 또는 이더넷 통신 성능은 랙 내 NVLink/NVSwitch 통신보다 낮습니다).

요약하면, NVIDIA의 ConnectX 및 BlueField DPU와 같은 InfiniBand 및 이더넷 NIC는 일반적으로 NVLink와 함께 사용됩니다. 이들은 랙 간에 높은 대역폭 연결성을 제공하고 DPU의 네트워크 내 컴퓨팅을 사용하여 프로토콜을 오프로드합니다.

사전 통합 랙 어플라이언스

NVL72는 매우 복잡한 시스템이므로, NVIDIA는 이를 단일 캐비닛에 사전 통합된 랙 "어플라이언스" 형태로 제공합니다. 이 어플라이언스는 18개의 컴퓨팅 노드, 9개의 NVSwitch 유닛, 내부 NVLink 케이블링, 전원 분배 장치, 냉각 시스템이 모두 조립된 상태로 제공됩니다. 조직은 이를 단일 유닛으로 주문하면 도착 즉시 사용 가능한 상태로 활용할 수 있습니다. 랙을 시설 전원 공급 장치에 연결하고, 수냉 인터페이스를 연결한 후, InfiniBand 케이블을 네트워크에 연결하고 전원을 켜기만 하면 됩니다.

이 시스템은 기본적으로 개봉 후 바로 사용 가능하며, AI 워크로드 실행을 시작하기 위해 최소한의 설정만 필요합니다. NVIDIA가 랙 내부에서 이미 처리해 놓았기 때문에 72개의 GPU를 개별적으로 NVLink로 케이블링할 필요가 없습니다. 곧 설명할 액체 냉각 설정 역시 자체적으로 완비되어 있습니다.

이러한 어플라이언스 방식은 배포를 가속화하며 시스템이 NVIDIA에 의해 올바르게 구축되고 검증되었음을 보장합니다. 랙에는 NVIDIA Base Command Manager 클러스터 관리 소프트웨어와 함께, 클러스터 작업 스케줄링 및 오케스트레이션을 위한 SLURM(Simple Linux Utility for Resource Management) 및 쿠버네티스도 포함됩니다.

요약하자면, NVL72 랙은 사용자의 환경에 바로 설치하여 생산용 AI 워크로드를 즉시 실행할 수 있도록 설계되었습니다. 수동 설치나 복잡한 구성이 필요하지 않습니다.

공동 패키징 광학 장치: 네트워킹 하드웨어의 미래

네트워킹 데이터 처리 속도 가 800Gbit/s, 1.6Tbit/s 및 그 이상으로 상승함에 따라 NVIDIA는 실리콘 포토닉스와 공동 패키징 광학(CPO)을 자사 네트워킹 하드웨어에 통합하기 시작했습니다. 여기에는 Quantum-X800 InfiniBand 및 Spectrum-X800 이더넷 플랫폼이 포함됩니다. 이 플랫폼들은 800Gb/s 종단 간 연결성과 네트워크 내 컴퓨팅 기능(예: SHARP)을 탑재해 출시됩니다. CPO를 통해 광 송신기가 스위치 실리콘 바로 옆에 통합됩니다. 이는 전기 경로를 극적으로 단축시켜 랙 간 더 높은 대역폭 링크를 가능하게 하고, 전력 소모를 줄이며, 전반적인 통신 효율을 향상시킵니다.

실질적으로 CPO와 같은 기술은 수백, 수천 개의 랙(AI 공장)을 단일 통합 패브릭으로 연결하는 길을 열어주고 있으며, 이로 인해 랙 간 대역폭이 더 이상 병목 현상이 되지 않습니다. 이러한 광네트워킹 발전은 초고성능 GPU를 갖춘 네트워크가 초고성능을 유지할 수 있도록 보장하는 데 필요한 고성능 랙 간 대역폭에 매우 중요합니다.

요약하면, NVL72 랙 내부에서 NVIDIA는 NVLink와 NVSwitch를 활용해 72개의 GPU 간 초고속 전수 상호 연결 네트워크를 구축합니다. 이 상호 연결은 매우 빠르고 균일하여 GPU들이 다수의 집단 연산에서 사실상 하나의 장치처럼 작동합니다. 랙 외부에서는 고속 NIC(예: InfiniBand 또는 이더넷)가 랙을 다른 랙이나 스토리지에 연결하며, DPU가 데이터 이동을 효율적으로 관리합니다.

NVL72는 매우 강력한 엔드-투-엔드( ) 독립형 시스템이자 대규모 AI 슈퍼컴퓨터 또는 *AI 팩토리의* 기본 구성 요소입니다. 여러 개의 이러한 랙으로 구성된 대규모 AI 데이터 센터인 AI 팩토리 개념이 이제 현실화되고 있습니다. NVIDIA는 HPE 및 Supermicro와 같은 OEM 및 시스템 공급업체와 협력하여 GB200 NVL72 시스템을 공급합니다. NVIDIA의 하드웨어 및 네트워크 로드맵은 AI 팩토리 비전을 실현하는 데 정확히 초점을 맞추고 있습니다. 요약하자면, NVL72는 GPU, 네

트워킹, 물리적 랙 하드웨어가 수천, 수백만 개의 GPU로 확장될 수 있도록 최대한 원활하고 효율적으로 함께 설계된 코디자인의 극한을 보여줍니다.

연산 밀도와 전력 요구 사항

NVL72 랙은 컴퓨팅 측면에서 놀라울 정도로 높은 밀도 를 자랑하며, 이는 단일 랙 기준으로 매우 높은 전력을 소모함을 의미합니다. 완전히 로드된 NVL72는 최대 부하 시 약 130kW의 전력을 소비할 수 있습니다. 이는 약 50~60kW를 소비했던 NVIDIA의 이전 세대 AI 랙보다 2배 이상 높은 수치입니다. 72개의 최첨단 GPU와 모든 지원 하드웨어를 하나의 랙에 집적하는 것은 데이터 센터 인프라가 감당할 수 있는 한계를 뛰어넘는 것입니다.

NVL72 랙에 130kW를 공급하려면 표준 전원 공급 장치 하나만으로는 부족합니다. 데이터 센터는 일반적으로 이러한 전력을 공급하기 위해 여러 개의 고용량 회로를 준비합니다. 예를 들어, 데이터 센터는 완전히 독립적인 두 개의 전원 공급 장치를 배치할 수 있습니다. 이 경우 각 공급 장치는 한쪽 공급 장치에 장애가 발생할 경우 전체 랙 부하를 감당할 수 있도록 설계됩니다.

한 공급원이 오프라인 상태가 되면, 남은 회로가 전체 130kW 전력을 지원하여 회로 과부하를 방지할 수 있습니다. 이러한 중복성은 중요한 보호 장치입니다. 그렇지 않으면 전력 중단으로 인해 수개월에 걸친 훈련 작업이 중단될 수 있습니다.

랙 내부에서는 전력이 각 1U 컴퓨팅 노드의 전원 공급 장치로 분배됩니다. 전력은 로컬 전자 장치를 위해 교류(AC)에서 직류(DC)로 변환됩니다. NVL72의 각 컴퓨팅 노드에는 두 개의 Grace Blackwell Superchip이 탑재되어 있으며, 이들이 합쳐서 약 6kW를 소비합니다. 18개의 컴퓨팅 노드를 고려하면 총 전력 소비량은 약 110kW입니다. NVSwitch 트레이, 네트워크 스위치, 공기 냉각 시스템, 수냉 펌프 등이 약 20kW를 차지하여 전체 NVL72 랙의 총 소비 전력은 130kW입니다.

일반적인 팜스프링스 데이터센터( )에서 사용되는 전압(예: 415V 3상 교류)의 전류는 매우 크기 때문에 모든 장비가 고전류 설계로 제작됩니다. 운영자는 이러한 랙을 호스팅하기 위해 신중하게 계획해야 하며( ), 이는 종종 전용 전원 분배 장치(PDU)와 세심한 모니터링을 필요로 합니다. 전력 과도 현상도 고려 대상입니다. 72개의 GPU가 유휴 상태에서 최대 전력으로 가속될 때, 단 몇 밀리초 만에 수십 kW의 전력을 급격히 소비할 수 있기 때문입니다. 우수한 설계에는 커패시터나 시퀀싱이 포함되어 큰 전압 강하를 방지합니다.

시스템은 GPU 부스트 클럭을 미세한 간격으로 시차를 두어 활성화함으로써, 모든 GPU가 정확히 동일한 마이크로초에 전력을 급증시키는 것을 방지하고 서지를 완화할 수 있습니다. 이러한 전기 공학적 세부 사항들이 130kW 랙을 관리 가능한 수준으로 만드는 핵심 요소입니다.

고밀도 컴퓨팅의 최첨단을 달리는 이 NVL72 랙을 소형 변전소라 부르는 것도 무리가 아닙니다. 576개의 GPU를 탑재한 이 랙 8대를 합치면 거의 1MW(랙당 130kW × 8랙)의 전력을 소비하는데, 이는 소규모 데이터 센터의 전체 용량에 해당합니다! 긍정적인 측면은 130kW가 랙 하나에 과도한 전력량이지만, 와트당 처리량 또한 매우 높다는 점입니다.

기존 장비 여러 랙을 NVL72 한 대로 대체하면 전체 효율은 향상됩니다. 하지만 이처럼 집중된 전력 소비를 지원할 인프라가 반드시 필요합니다. 또한 NVL72 랙을 수용하는 모든 시설은 다음에서 논의할 것처럼 충분한 전력 용량과 냉각 시스템을 확보해야 합니다.

액체 냉각 대 공기 냉각

130kW를 단일 랙에서 냉각하는 것은 기존 공기 냉각 기술의 한계를 넘어섭니다( ). 각각 약 1,200와트의 열을 방출하는 72개의 GPU에 공기를 불어넣으려면 허리케인 수준의 기류가 필요하며, 이는 극도로 시끄럽고 비효율적일 뿐만 아니라 배출되는 뜨거운 공기도 치명적일 것입니다. 따라서 이 정도의 전력 밀도로 작동하는 NVL72 랙에는 액체 냉각이 유일한 실용적 해결책입니다.

NVL72는 완전 액체 냉각 시스템입니다. 각 Grace Blackwell Superchip 모듈과 NVSwitch 칩에는 콜드 플레이트가 부착됩니다. 콜드 플레이트는 내부 배관이 있는 금속판으로, 부품에 직접 접촉합니다. 수성 냉각수가 배관을 통해 흐르며 열을 제거합니다. 모든 콜드 플레이트는 호스, 매니폴드, 펌프로 연결되어 냉각수를 시스템 전체에 순환시킵니다.

일반적으로 랙에는 각 노드에 대한 퀵 디스커넥트 커플링이 있어 냉각수를 흘리지 않고 서버를 삽입하거나 제거할 수 있습니다. 랙은 외부 시설의 냉각수 시스템에 공급 및 반환 연결부를 갖습니다. 종종 랙 내부에 내장되거나 바로 옆에 위치한 냉각수 분배 장치(CDU)라는 열교환기가 있습니다. CDU는 랙 내부 냉각수 루프의 열을 데이터 센터의 냉각수 루프로 전달합니다.

시설은 20~30°C의 냉각수를 공급합니다. 물은 열교환기를 통해 열을 흡수합니다. 가열된 물은 다시 냉각기나 냉각탑으로 펌핑되어 재냉각됩니다. 현대식 설계에서는 온수 냉각 방식을 채택하기도 하는데, 이때 냉각수는 30°C로 유입되어

45°C로 배출됩니다. 이 물은 능동 냉각 없이 증발식 냉각탑으로 냉각될 수 있어 전체 효율을 향상시킵니다. 핵심은 물이나 액체 냉각제가 공기보다 유량 단위당 훨씬 더 많은 열을 운반할 수 있다는 점입니다. 따라서 좁은 공간에서 고전력으로 작동할 때 액체 냉각이 훨씬 더 효과적입니다.

액체 냉각은 GPU와 CPU 온도를 공기 냉각 대비 훨씬 낮게 유지함으로써 열에 의한 GPU 스로틀링을 줄입니다. GPU는 온도 한계에 도달하지 않고 최대 클럭을 유지할 수 있습니다. 또한 칩을 더 낮은 온도에서 작동시키면 전력 손실이 감소하므로 신뢰성과 효율성까지 향상됩니다.

NVL72는 부하 상태에서 GPU 온도를 50~70°C 범위로 유지하는데, 이는 전력 소모가 큰 장치에 매우 우수한 성능입니다. 콜드 플레이트와 냉각수 루프는 각 GPU가 1,000W, 각 CPU가 500W의 열을 시스템에 방출할 수 있도록 매우 정교하게 설계되었습니다. 또한 냉각수 유량은 해당 열을 신속히 제거할 수 있을 만큼 충분해야 합니다. 대략적인 추정치에 따르면 약 130kW의 열을 방출하려면 물 온도 상승 10~12°C 조건에서 분당 150~200리터의 냉각수가 필요합니다.

이 시스템에는 냉각수 온도, 압력, 누수 감지 센서와 제어 장치가 반드시 포함됩니다. 누수 감지 센서나 압력 손실 센서를 통해 누수가 감지되면 시스템은 해당 구역을 신속히 차단하거나 격리할 수 있습니다. 유체 누출 위험을 최소화하기 위해 자가 밀봉 연결부 사용과 보조 격리 트레이 설치를 권장합니다.

랙 내 이 수준의 액체 냉각은 한때 특수한 기술이었으나, 현재는 이러한 대규모 AI 클러스터의 표준이 되었다. 메타(Meta), xAI, 구글(Google)과 같은 기업들은 공기 냉각으로는 이러한 시스템에서 소비되는 막대한 전력을 감당할 수 없기 때문에 AI 클러스터에 액체 냉각을 도입하고 있다.

따라서 NVL72는 액체 냉각 루프를 포함한 더 복잡한 시설을 요구하지만, 많은 데이터 센터가 이제 액체 냉각을 고려하여 건설됩니다. 내장형 내부 액체 냉각 시스템을 갖춘 NVL72 랙은 냉각 루프에 직접 연결될 수 있습니다.

내부 액체 냉각의 부작용 중 하나는 랙의 무게입니다. 하드웨어와 냉각제가 채워진 NVL72 랙의 무게는 약 3,000파운드(1.3~1.4톤)에 달합니다. 이는 소형 자동차 무게에 해당하지만 바닥 몇 평방피트에 집중된 것으로, 랙으로서는 극히 무거운 편입니다. 고가 바닥을 갖춘 데이터 센터는 평방피트당 파운드 단위로 측정되는 이 하중을 바닥이 견딜 수 있는지 확인해야 합니다. 고밀도 랙은 종종 보강 슬래브 위에 배치되거나 추가 스트럿으로 지지됩니다. 이러한 랙을 이동하려면 지게차와 같은 특수 장비가 필요합니다. 이는 AI 슈퍼컴퓨터를 설치할 때 발생하는 독특한 물리적·물류적 과제를 고려해야 하는 배포 과정의 일부입니다.

NVIDIA는 또한 랙 관리 컨트롤러 형태로 관리 및 안전 기능을 통합합니다. 이 컨트롤러는 냉각수 펌프, 밸브 위치, 전력 사용량 등을 감독하고 모든 노드의 상태를 모니터링합니다. 관리자는 이를 통해 모든 노드의 펌웨어 업데이트나 시스템의 안전한 종료와 같은 작업을 수행할 수 있습니다.

이러한 모든 고려 사항은 NVL72가 데이터 센터 인프라를 염두에 두고 공동 설계되었음을 보여줍니다. NVIDIA는 전력과 냉각 방안을 마련한 시스템 엔지니어, 설치 및 운영 방식을 규정한 시설 엔지니어와 협력하여 컴퓨팅 아키텍처를 개발했습니다. 단순히 빠른 칩이 아닌 균형 잡히고 실용적인 시스템을 제공하는 것이 핵심입니다.

이러한 복잡성의 보상은 막대합니다. 전력 및 냉각의 한계를 뛰어넘음으로써, 엄청난 양의 컴퓨팅 성능이 단일 랙에 집중되어 와트당 컴퓨팅 성능이 크게 향상됩니다. 물론 130kW는 상당한 전력량이지만, GPU당 또는 1조 FLOP(TFLOP)당으로 계산하면, 동일한 GPU를 효율성이 낮은 냉각 시스템으로 여러 랙에 분산 배치하는 방식에 비해 오히려 효율적입니다.

실무에서의 성능 모니터링 및 활용

이처럼 강력한 시스템을 보유할 때는 그 성능을 최대한 활용하고 있는지 확인해야 합니다. NVL72를 효과적으로 운영하려면 성능, 활용도, 전력에 대한 세심한 모니터링이 필요합니다. NVIDIA는 데이터 센터 GPU 관리자(DCGM)와 같은 도구를 제공하여 각 GPU의 활용도 비율, 메모리 사용량, 온도, NVLink 처리량 등의 메트릭을 추적할 수 있습니다.

성능 엔지니어로서 훈련 실행 및 추론 워크로드 중 이러한 지표를 주시해야 합니다. 이상적으로는 훈련 작업 중 대부분의 시간 동안 GPU 활용률이 100%에 근접해야 합니다. GPU 활용률이 50%로 나타난다면, 절반의 시간 동안 GPU가 유휴 상태로 유지되고 있음을 의미합니다. 데이터 로딩 병목 현상이나 동기화 문제가 있을 수 있습니다.

마찬가지로 NVLink 사용량도 모니터링할 수 있습니다. NVLink 링크가 자주 포화 상태에 이르면 통신이 원인일 가능성이 높습니다. BlueField DPU와 NIC에는 자체 통계 기능이 있어 데이터 읽기 시 스토리지 링크 포화를 방지하도록 모니터링됩니다. NVL72 같은 최신 시스템은 이러한 텔레메트리 데이터를 노출합니다.

전력 모니터링 또한 중요합니다. 약 130kW 규모의 시스템에서는 사소한 비효율이나 잘못된 구성도 막대한 전력 및 비용 낭비로 이어질 수 있습니다. 시스템은 노드별 또는 GPU별 전력 소모량을 모니터링할 수 있도록 지원할 것입니다. 관리

자는 최대 성능이 필요하지 않은 경우 에너지 절약을 위해 GPU의 전력 또는 클럭을 제한할 수 있습니다.

NVIDIA GPU는 전력 제한 설정이 가능합니다. 예를 들어, 성능을 극한까지 끌어올릴 필요가 없는 소규모 작업을 실행할 경우 GPU 클럭을 낮춰 효율성(와트당 성능으로 측정)을 개선하면서도 처리량 요구사항을 충족할 수 있습니다. 이로 인해 수 킬로와트의 전력을 절약할 수 있습니다. 수주간의 훈련 기간 동안 이는 상당한 절감 효과와 비용 효율성으로 이어질 수 있습니다.

공유 및 스케줄링

또 다른 측면은 NVL72에서 워크로드 공유 및 스케줄링입니다. 모든 작업이 72개의 GPU를 모두 필요로 하는 경우는 거의 없습니다. 여러 팀이나 실험이 GPU의 일부 집합에서 동시에 실행될 수 있습니다. SLURM이나 쿠버네티스 같은 클러스터 스케줄러에 NVIDIA 플러그인을 사용하면, 동일한 랙 내에서 한 사용자에게 8개 GPU, 다른 사용자에게 16개 GPU, 또 다른 사용자에게 48개 GPU를 할당하는 식으로 분할할 수 있습니다.

또한 NVIDIA의 멀티 인스턴스 GPU(MIG) 기능 을 사용하면 단일 물리적 GPU를 하드웨어 수준에서 분할된 더 작은 GPU로 나눌 수 있습니다. 예를 들어, 180GB의 GPU 메모리를 가진 하나의 Blackwell GPU를 더 작은 조각으로 분할하여 여러 작은 추론 작업을 동시에 실행할 수 있습니다.

각 Blackwell GPU는 최대 7개의 완전히 분리된 MIG 인스턴스를 지원합니다. 이를 통해 하나의 물리적 GPU를 전용 메모리와 SM을 갖춘 최대 7개의 소형 GPU로 분할할 수 있습니다. MIG 크기는 제품 세대에 따라 고정되어 있습니다. MIG 파티션에 대한 자세한 내용은 다음 장에서 살펴보겠습니다.

실제 환경에서는 이렇게 큰 GPU에서 MIG를 활용해 하나의 GPU로 여러 모델을 동시에 서비스하는 추론 시나리오에 적용할 수 있습니다. BlueField DPU의 존재는 DPU가 방화벽 및 가상 스위치 역할을 수행할 수 있으므로 안전한 다중 테넌트 환경을 가능하게 합니다. 이는 서로 다른 작업과 사용자의 네트워크 트래픽을 격리합니다. 즉, 조직은 서로 간섭 없이 서로 다른 부서 또는 외부 고객이 시스템의 파티션을 안전하게 사용할 수 있도록 허용할 수 있습니다. 이는 클라우드 제공업체가 안전한 다중 테넌트 격리를 통해 대규모 서버를 여러 고객을 위해 분할하는 방식과 유사합니다.

비용 측면에서 NVL72와 같은 시스템은 수백만 달러 규모의 자산이며, 매월 수만 달러의 전력을 소모할 수 있습니다. 따라서 가능한 한 많은 유용한 작업, 즉 굿풋

(goodput)을 수행하는 것이 중요합니다. 시스템이 유휴 상태라면 막대한 자본 및 운영 비용이 낭비되는 셈입니다. 따라서 시간 경과에 따른 활용도 모니터링이 중요합니다. 사용 가능한 시간 대비 실제 사용된 GPU 시간을 추적할 수 있습니다.

시스템 활용도가 낮은 것으로 확인되면 워크로드를 통합하거나 추가 팀에 더 많은 프로젝트를 위해 제공할 수 있습니다. 일부 조직은 내부 팀이 자체 예산으로 GPU 사용 시간당 비용을 지불하는 차징백 모델을 구현합니다. 이는 효율적인 사용을 장려하고 전기 및 감가상각 비용을 반영합니다. 이러한 투명성은 사람들이 자원을 소중히 여기도록 합니다.

하드웨어 업그레이드의 투자 수익률(ROI)

와 같은 최첨단 하드웨어에 투자할 가치가 있는지 의문이 들 수 있습니다. 투자 수익(반환하다)을 분석할 때, 답은 종종 달러당 성능으로 귀결됩니다. 예를 들어 NVL72가 구형 랙 4대의 작업을 처리할 수 있다면, 장기적으로 하드웨어와 전력 비용 모두를 절감할 수 있습니다. 본 장 초반에 블랙웰 GPU 한 대가 처리량 측면에서 호퍼 GPU 2~3대를 대체할 수 있다고 논의한 바 있습니다. 이는 업그레이드 시 동일한 작업을 수행하는 데 필요한 총 GPU 수가 줄어들 수 있음을 의미합니다.

간단한 사례 연구를 분석해 보겠습니다. 현재 100개의 H100 GPU로 워크로드를 처리하고 있다고 가정해 보겠습니다. 각 GPU가 두 배 이상 빠르거나(FP8/FP4 사용 시 그 이상) 더 빠르기 때문에 50개의 Blackwell GPU로도 처리할 수 있습니다. 따라서 100개 대신 50개의 GPU를 구매하게 됩니다. 각 블랙웰의 가격이 H100보다 비싸더라도, 절반만 구매하면 비용이 비슷하거나 더 저렴해질 수 있습니다. 전력 측면에서, 동일한 작업을 처리할 때 100개의 H100은 70kW를 소비할 수 있지만, 50개의 블랙웰은 50kW만 소비할 수 있습니다. 이는 상당한 전력 절감 효과입니다.

1년 동안 이 전력 차이로 수만 달러를 절약할 수 있습니다. 또한 GPU 수가 줄면 유지 관리해야 할 서버도 줄어들어 해당 서버의 CPU, RAM, 네트워킹에 대한 오버헤드가 감소하여 추가 절감 효과를 제공합니다. 종합하면, 특히 24시간 가동할 만큼 충분한 작업량이 있다면, 새 하드웨어로의 업그레이드는 경우에 따라 1~2년 안에 투자 비용을 회수할 수 있습니다.

물론 정확한 가격과 사용 패턴에 따라 계산은 달라지지만, 핵심은 대규모 배포 시 최신 AI 하드웨어 도입의 투자 수익률(ROI)이 매우 높을 수 있다는 점입니다. 가시적인 ROI 외에도, 여러 개의 소형 시스템 대신 단일 고성능 시스템을 사용함으로써 시스템 아키텍처를 단순화할 수 있는 간접적 이점도 있습니다. 이러한 단순

화는 전력 소비를 줄이고 네트워크 복잡성을 감소시켜 운영 효율성을 향상시킵니다.

예를 들어 메모리 한계로 인해 여러 구형 GPU에 모델을 분할 배치할 필요가 없어지면 소프트웨어가 단순화되고 엔지니어링 복잡성이 감소합니다. 또한 최신 하드웨어를 보유하면 최신 소프트웨어 최적화를 활용할 수 있으며, 업그레이드하는 경쟁사들과의 격차를 유지할 수 있습니다. 누구도 경쟁사보다 절반 속도로 모델을 훈련하고 서비스하는 처지가 되고 싶어 하지 않습니다. 업그레이드는 성능을 향상시키는 동시에 더 큰 모델, 더 빠른 반복 작업, 신속한 응답을 가능하게 합니다.

NVL72를 효과적으로 운영하는 것은 하드웨어적 성과만큼이나 소프트웨어 및 관리 측면의 도전 과제입니다. 하드웨어는 놀라운 잠재력을 제공하지만, 성능 모니터링, 높은 활용도 유지, 작업의 스마트한 스케줄링을 통해 하드웨어의 모든 성능을 활용하는 것은 엔지니어의 몫입니다.

좋은 소식은 NVIDIA가 드라이버, 자질, 컨테이너 런타임, 클러스터 오케스트레이션 도구 등 성능 모니터링 및 개선을 위한 풍부한 소프트웨어 스택을 제공한다는 점입니다. 이 책의 나머지 부분에서는 GB200/GB300 NVL72와 같은 시스템을 완전히 활용하기 위해 소프트웨어를 최적화하는 방법을 살펴볼 것입니다. 지금 당장 중요한 점은, 박스 안에 엑사플롭스(exaFLOPS) 규모의 성능을 가진 AI 시스템을 제공받았을 때, 모든 플롭(flop)과 모든 바이트(byte)를 최대한 활용하기 위해서는 그만큼 진보된 전략이 필요하다는 것입니다.

미래를 엿보다: NVIDIA의 로드맵

본문 작성 시점( ) 기준, Grace Blackwell NVL72 플랫폼은 AI 하드웨어의 최첨단을 대표합니다. 그러나 NVIDIA는 이미 차세대 도약을 준비 중입니다. 향후 몇 년간의 NVIDIA 하드웨어 로드맵을 간략히 살펴볼 가치가 있는데, 이는 명확한 확장 패턴을 보여주기 때문입니다. NVIDIA는 성능, 메모리, 통합성 측면에서 계속해서 두 배로 투자할 계획입니다.

Blackwell Ultra 및 Grace Blackwell Ultra

NVIDIA의 Blackwell Ultra(B300)와 그에 대응하는 Grace Blackwell Ultra 슈퍼칩( )은 NVL72 아키텍처에 대한 드롭인 업그레이드입니다. 각 Blackwell Ultra B300 GPU는 B200(180GB)보다 약 50% 더 많은 메모리 용량(288GB)을 제공하며, AI 연산 성능이 1.5배 향상되고, 어텐션 연산 및 정밀도 감소(예: NVFP4)를 위해 특별히 설계된 더 큰 온다이 가속기를 탑재하고 있습니다. 이는

Blackwell B300이 B200보다 45~50% 더 높은 추론 처리량을 생산한다는 것을 의미합니다.

72개의 GB300 GPU로 구성된 랙은 36개의 Grace Blackwell Ultra 모듈(각 모듈당 GPU 2개 + CPU 1개), 약 20.7TB의 HBM(72 × 288GB), 약 18TB의 DDR(36 × 500GB)으로 구성됩니다. 합산하면 GB300 NVL72 랙당 약 38TB의 고속 메모리를 제공합니다. 또한 GB300 NVL72 Ultra의 랙 내 NVLink 및 NVSwitch 네트워크는 GB200 NVL72와 동일한 NVLink 5 세대를 사용합니다.

요약하면 GB300은 동일한 아키텍처를 사용하므로 GB200의 진화적 업그레이드입니다. 그러나 더 많은 SM, 더 높은 메모리 용량, 더 빠른 클럭 속도를 포함해 모든 측면에서 향상되었습니다.

베라 루빈 슈퍼칩(2026)

암흑 물질의 증거를 제시한 여성 천문학자 베라 루빈( )의 이름을 딴 코드명인 베라 루빈 슈퍼칩(VR200)은 차세대 주요 아키텍처 단계입니다. 베라는 그레이스 CPU의 후속 ARM 기반 CPU이며, 루빈은 블랙웰의 후속 GPU 아키텍처입니다. 엔비디아는 그레이스 블랙웰(GB200/GB300) 구성과 유사하게 단일 모듈(VR200)에 베라 CPU 1개와 루빈 GPU 2개를 결합하여 슈퍼칩 개념을 이어갑니다.

베라 CPU는 TSMC의 3nm 반도체 공정을 사용하며, 더 많은 CPU 코어와 약 1TB/s 속도의 더 빠른 LPDDR6 메모리를 탑재합니다. 루빈 GPU는 약 13~14TB/s 속도의 더 높은 GPU 고대역폭 메모리(HBM)를 지원합니다.

NVLink 역시 6세대인 NVLink 6으로 진화할 것으로 예상되며, 이는 CPU-GPU 및 GPU-GPU 링크 대역폭을 두 배로 증가시킬 것입니다. 또한 베라 루빈은 랙당 더 많은 노드(또는 NVLink 도메인당 더 많은 랙)를 허용하여 8랙 구성의 GB200/GB300 NVL72 클러스터가 가진 576 GPU 제한을 넘어 확장할 수 있을 것이라는 추측도 있습니다.

결론적으로 베라 루빈 세대는 코어 수, 메모리 용량, 대역폭, TFLOPS 등 대부분의 지표에서 약 2배의 성능 향상을 가져올 것입니다. 루빈 GPU는 다이당 약 200개의 SM(Streaming Multiprocessor)을 탑재해 효율성을 더욱 높일 수 있습니다. 2세대 FP4나 실험적인 2비트 정밀도 같은 신규 기능도 통합될 수 있으나, 이는 현재로서는 추측에 불과합니다.

특히 흥미로운 가능성은 루빈의 288GB HBM RAM이 여전히 대규모 AI 모델의 병목 현상이라는 점입니다. 이에 따라 엔비디아는 GPU 모듈에 직접 2차 메모리

를 통합할 수 있습니다. 예를 들어, GPU 모듈 베이스에 LPDDR 메모리를 직접 배치하여 베라의 CPU DDR 메모리와 별개로 GPU용 더 크지만 느린 메모리 풀로 활용할 수 있습니다.

이 경우 단일 GPU 모듈은 총 약 550GB(288GB HBM + 256GB LPDDR)의 캐시 일관성 있는 통합 메모리를 보유할 수 있습니다. 이는 GPU가 자체적인 다단계 메모리 계층 구조를 갖게 됨에 따라 CPU와 GPU 메모리 간의 경계를 더욱 모호하게 만들 것입니다. 루빈 GPU 세대에서 이 기술이 구현되든 아니든, 주목해야 할 방향입니다.

전체적으로 베라 루빈 및 베라 루빈 울트라 랙은 GB200/GB300 NVL72 대비 5배의 성능을 제공합니다. 또한 랙당 약 600kW에 달하는 5배의 전력을 소모합니다. VR200/VR300 NVL 시스템은 모든 루빈 GPU에 걸쳐 랙당 총 GPU HBM 용량이 방대하며(GPU당 288GB HBM), 여기에 수십 TB의 CPU 메모리가 추가됩니다. 또한 랙 내 NVLink 6은 NVLink 5보다 통신 오버헤드가 적습니다.

루빈 울트라 및 베라 루빈 울트라 (2027)

기존 패턴에 따라, 루빈(R300)과 베라 루빈의 '울트라' 버전 이 출시 1년 후에 등장합니다. 한 보고서에 따르면 NVIDIA는 그때까지 4개의 다이로 구성된 GPU 모듈로 전환할 수 있다고 합니다. 이는 두 개의 듀얼 다이 루빈 패키지를 결합하여 쿼드 다이 루빈 GPU를 구현하는 방식입니다. 이 R300 루빈 울트라 GPU 모듈은 하나의 패키지에 4개의 GPU 다이를 탑재하며, 단일 R300 GPU 모듈에 총 1TB의 HBM 메모리를 제공하는 16개의 HBM 스택을 포함합니다. 4개의 다이는 듀얼 다이 B300 모듈의 코어 수를 두 배로 늘립니다.

특히 베라 루빈 NVL144 시스템은 랙 전체에 걸쳐 144개의 해당 다이를 탑재합니다. 이는 각 4개의 다이를 가진 36개의 슈퍼칩 모듈에 해당합니다. 또한 전체 시스템에 다중 다이 패키지를 적용해 GPU 수를 4배로 늘린 베라 루빈 NVL576 구성도 존재합니다.

2027년까지 각 랙은 3~4 엑사플롭스의 연산 성능과 총 165TB의 GPU HBM RAM(루빈 GPU당 288GB HBM × 576개 GPU)을 제공할 수 있을 것입니다. 이러한 수치는 아직 다소 추측에 불과하지만, 엑사플롭스 단위의 연산 성능과 테라바이트 단위의 GPU HBM RAM을 갖춘 초대형 AI 시스템으로의 발전 방향은 분명합니다.

파인만 GPU(2028년)와 매년 두 배 증가하는 성능

엔비디아는 루빈 이후 세대인 ''의 코드명을 파인만(Feynman)으로 정했으며, 2028년 출시를 목표로 하고 있습니다. 세부 사항은 아직 부족하지만, 파인만 GPU는 더 정밀한 2nm TSMC 공정 노드로 전환될 가능성이 높습니다. HBM5를 채택하고 모듈 내부에 더 많은 DDR 메모리를 포함할 것으로 예상됩니다. 또한 다이 수를 4개에서 8개로 두 배로 늘릴 수도 있습니다.

2028년까지는 추론 수요가 AI 워크로드를 확실히 주도할 것으로 예상됩니다. 특히 AI 모델에서 추론 기능이 계속 발전함에 따라 더욱 그러할 것입니다. 추론은 기존 비추론 모델에 비해 수백 배에서 수천 배 더 많은 추론 시간 계산을 요구합니다. 따라서 칩 설계는 대규모 추론 효율성을 최적화할 것으로 보이며, 여기에는 더 많은 새로운 정밀도, 더 많은 온칩 메모리, NVLink의 처리량을 더욱 향상시키기 위한 패키지 내 광학 링크 등이 포함될 수 있습니다.

엔비디아는 가능한 한 매년, 매 세대마다 무언가를 두 배로 늘리는 것 같습니다. 한 해는 메모리를 두 배로 늘리고, 다른 해에는 다이 수를 두 배로 늘리며, 또 다른 해에는 상호 연결 대역폭을 두 배로 늘리는 식입니다. 몇 년에 걸쳐 이 두 배 증가의 복합 효과는 엄청납니다. 엔비디아의 공격적인 성장 궤적은 매 세대마다 중요한 요소를 두 배로 늘리는 방식에서 확인할 수 있습니다. 예를 들어, 블랙웰은 듀얼 GPU 다이(모듈당 1개 대신 2개)를 도입했으며, NVLink 링크당 양방향 대역폭은 약 900GB/s에서 약 1.8TB/s로 두 배 증가했습니다. 또한 GPU당 메모리는 블랙웰의 180GB에서 블랙웰 울트라 세대의 약 288GB로 증가했습니다. 루빈과 파인만은 컴퓨팅 성능, 메모리, 대역폭을 더욱 향상시켰습니다.

엔비디아는 랙이 AI 모델 생산라인인 'AI 팩토리'를 반복적으로 언급합니다. 엔비디아는 파트너사를 통해 '랙 서비스(Rack as a Service)'를 제공해 기업이 자체 구축 대신 슈퍼컴퓨터의 일부를 임대할 수 있도록 구상 중입니다. 이 추세는 최첨단 하드웨어가 배포 가능한 통합형 포드 형태로 제공되면서 지속될 전망입니다. 또한 각 세대마다 새 포드로 교체해 용량을 두 배로 늘리고 성능을 향상시키며 비용을 절감할 수 있습니다.

성능 엔지니어인 우리에게 중요한 것은 하드웨어가 계속해서 새로운 수준의 확장을 가능하게 한다는 점입니다. 현재는 불가능한 모델도 몇 년 후에는 일상적으로 처리될 수 있습니다. 이는 또한 새로운 정밀도 형식, 더 큰 메모리 풀, 개선된 상호 연결 기술 등을 활용하기 위해 소프트웨어를 지속적으로 적응시켜야 함을 의미합니다. 최첨단 모델의 발전이 이러한 하드웨어 혁신과 밀접하게 연결되어 있기 때문에 지금은 매우 흥미로운 시기입니다.

핵심 요약

다음 혁신 기술들은 종합적으로 NVIDIA 하드웨어가 초거대 AI 모델을 전례 없는 속도, 효율성, 확장성으로 처리할 수 있도록 합니다:

통합 슈퍼칩 아키텍처

NVIDIA는 ARM 기반 CPU(Grace)와 GPU(Hopper/Blackwell)를 단일 슈퍼칩으로 융합하여 통합 메모리 공간을 창출합니다. 이 설계는 CPU와 GPU 간 수동 데이터 전송 필요성을 제거함으로써 데이터 관리를 단순화합니다.

통합 메모리 아키텍처

통합 메모리 아키텍처와 일관성 있는 상호 연결은 프로그래밍 복잡성을 줄여줍니다. 개발자는 명시적인 데이터 이동을 걱정하지 않고 코드를 작성할 수 있어 개발 속도가 빨라지고 AI 알고리즘 개선에 집중할 수 있습니다.

초고속 상호연결망

NVLink(NVLink-C2C 및 NVLink 5 포함)와 NVSwitch를 활용하여 시스템은 극히 높은 랙 내 대역폭과 낮은 지연 시간을 달성합니다. 이는 GPU들이 하나의 거대한 프로세서 일부인 것처럼 거의 동일하게 통신할 수 있음을 의미하며, 이는 AI 훈련 및 추론 확장에 매우 중요합니다.

고밀도, 초대형 시스템 (NVL72)

NVL72 랙은 72개의 GPU를 하나의 컴팩트한 시스템에 통합합니다. 이 통합 설계는 높은 컴퓨팅 성능과 방대한 통합 메모리 풀을 결합하여 대규모 모델을 지원하며, 기존 환경에서는 불가능했던 작업을 가능하게 합니다.

고급 냉각 및 전력 관리

NVL72는 정교한 액체 냉각 및 강력한 전력 분배 시스템을 기반으로 하며 랙당 약 130kW(130kW = 18노드 × 노드당 6kW + 약 20kW NVSwitch/냉각/오버헤드)로 작동합니다. 이러한 냉각 및 전력 용량은 고밀도 고성능 구성 요소를 관리하고 안정적인 작동을 보장하는 데 필수적입니다.

상당한 성능 및 효율성 향상

Hopper H100과 같은 이전 세대에 비해 Blackwell GPU는 연산 및 메모리 대역폭에서 약 2~2.5배의 성능 향상을 제공합니다. 이는 훈련 및 추론

속도의 상당한 개선으로 이어지며, Blackwell의 FP4 텐서 코어와 트랜스포머 엔진을 활용하는 경우 최대 30배 빠른 추론 속도를 달성할 수 있습니다. 또한 GPU 수를 줄여 잠재적인 비용 절감 효과도 기대할 수 있습니다.

최신 소프트웨어 스택 지원

NVIDIA의 소프트웨어 및 프레임워크는 최신 하드웨어를 완전히 활용하고 공동 설계된 시스템 최적화를 지원하기 위해 지속적으로 진화하고 있습니다. 여기에는 통합 메모리 관리 및 네이티브 FP8/FP4 정밀도 지원이 포함됩니다. 따라서 엔지니어는 최소한의 코드 변경으로 시스템의 전체 성능을 활용할 수 있습니다.

미래 대비 로드맵

NVIDIA의 개발 로드맵(Blackwell Ultra, Vera Rubin, Vera Rubin Ultra, Feynman 포함)은 컴퓨팅 처리량 및 메모리 대역폭과 같은 핵심 매개변수의 지속적인 배가 성장을 약속합니다. 이 발전 경로는 향후 더욱 거대해지는 AI 모델과 복잡한 워크로드를 지원하기 위해 설계되었습니다.

결론

NVIDIA NVL72 시스템은 그레이스 블랙웰 슈퍼칩, NVLink 패브릭, 고급 냉각 기술을 통해 AI 하드웨어 설계의 최첨단을 보여줍니다. 이 장에서는 모든 구성 요소가 AI 워크로드 가속이라는 단일 목표를 위해 공동 설계된 방식을 살펴보았습니다. CPU와 GPU는 하나의 유닛으로 융합되어 데이터 전송 병목 현상을 제거하고 거대한 통합 메모리를 제공합니다.

수십 개의 GPU는 초고속 네트워크로 연결되어 통신 지연이 최소화된 하나의 거대한 GPU처럼 작동합니다. 또한 메모리 서브시스템은 확장 및 가속화되어 GPU 코어의 막대한 요구를 충족시킵니다. 심지어 전력 공급 및 열 관리도 이 수준의 컴퓨팅 밀도를 가능하게 하기 위해 새로운 차원으로 발전했습니다.

그 결과 단일 랙으로 다중 랙 슈퍼컴퓨터에서만 볼 수 있었던 성능을 구현했습니다. 엔비디아는 칩, 보드, 네트워킹, 냉각 등 컴퓨팅 스택 전체를 종단 간 최적화하여 초대형 AI 모델의 훈련 및 서비스를 초고성능으로 가능하게 했습니다.

그러나 이러한 하드웨어 혁신에는 특수 시설, 전력 및 냉각에 대한 세심한 계획, 그리고 이를 완전히 활용하기 위한 정교한 소프트웨어가 필요하다는 도전 과제가 따릅니다. 하지만 그 보상은 엄청납니다. 연구자들은 이제 결과를 몇 주 또는 몇 달 동안 기다리지 않고도 전례 없는 규모와 복잡성을 지닌 모델을 실험할 수 있습니다. 기존 인프라에서는 한 달이 걸렸을 모델 훈련이 NVL72에서는 며칠 만

에 완료됩니다. 간신히 상호작용이 가능했던 추론 작업(쿼리당 수 초 소요)이 이제 실시간(밀리초 단위)으로 구현됩니다. 이는 수조 개의 매개변수를 가진 대화형 AI 어시스턴트 및 에이전트처럼 이전에는 실현 불가능했던 AI 애플리케이션의 문을 열어줍니다.

NVIDIA의 빠른 로드맵은 이것이 시작에 불과함을 시사합니다. Grace Blackwell 아키텍처는 Vera Rubin과 Feynman으로 진화할 것이며 그 이상으로 발전할 것입니다. NVIDIA CEO 젠슨 황이 설명하듯, "AI는 빛의 속도로 진화하고 있으며, 기업들은 추론 AI의 처리 수요와 추론 시간 확장에 대응할 수 있는 확장성을 갖춘 AI 공장을 구축하기 위해 경쟁하고 있습니다."

NVL72와 그 후속 모델들은 AI 공장의 핵심입니다. 이는 산더미 같은 데이터를 처리하여 놀라운 AI 역량을 생산해낼 중장비입니다. 성능 엔지니어로서 우리는 이 하드웨어 혁신의 어깨 위에 서 있습니다. 이는 우리에게 엄청난 원시적 역량을 제공하며, 우리의 역할은 하드웨어의 잠재력을 최대한 활용하는 소프트웨어와 알고리즘을 개발함으로써 이 혁신을 활용하는 것입니다.

다음 장에서는 하드웨어에서 소프트웨어로 전환할 것입니다. NVL72와 같은 시스템에서 운영체제, 드라이버, 라이브러리를 최적화하여 이 놀라운 하드웨어가 제대로 활용되도록 하는 방법을 탐구할 것입니다. 후속 장에서는 소프트웨어 아키텍처를 보완하는 메모리 관리 및 분산 훈련/추론 알고리즘을 살펴볼 것입니다.

이 책의 주제는 공동 설계(codesign)입니다. 하드웨어가 AI를 위해 공동 설계된 것처럼, 하드웨어의 성능을 극대화하기 위해서는 소프트웨어와 방법론 역시 공동 설계되어야 합니다. 이제 하드웨어의 기본 원리를 명확히 이해했으니, AI 시스템 성능 향상을 위한 소프트웨어 전략으로 깊이 들어가 볼 준비가 되었습니다. AI 슈퍼컴퓨팅 시대가 도래했으며, 이를 최대한 활용하는 여정은 정말 흥미진진할 것입니다.

시작해 보겠습니다!