AI, HPC 와 대규모 언어 모델을 지원하기 위한 미래 데이터센터 네트워크 구조

김대업*, 이준기
한국전자통신연구원

*artkdu@etri.re.kr

Architecture of future data center network supporting AI, HPC, and large-scale language models

Dae-Ub Kim* and Joon Ki Lee
Electronics and Telecommunications Research Institute (ETRI).

요 약

본 논문은 AI/HPC 데이터 센터에서 고성능 AI 컴퓨팅을 실현하기 위해서 CPU, 가속기, 비부착 메모리의 모듈화, 새로운 연결 인터페이스와 네트워크 구조에 대하서 논의한다. 또한 관련 인터페이스와 네트워크 구조에 적용될 개방형 프로토콜에 대하여 분석한다.

Ⅰ. 서 론

고성능 인공지능(AI)과 고성능 컴퓨팅(HPC) 시스템에 대한 수요가 전 세계적으로 증가함에 따라, AI/HPC 데이터 센터 네트워크는 AI 의 복잡한 워크로드를 지원하도록 최적의 네트워크와 더욱 강력하고 효율적인 하드웨어 집합을 구성할 수 있도록 CPU, 가속기, 메모리 같은 컴퓨팅 자원들을 모듈 형태로 구현하고, 모듈 사이에 새로운 연결 구조를 도출하여야 한다. 그래서 AI/HPC 데이터 센터의 효율적인 확장 전략이 필요하고 동시에 AI/HPC 하드웨어와 네트워크의 혁신 및 모듈형 인프라 등 전반에 걸친 업계 협력이 중요하다. AI 반도체 기술에 있어서 새로운 소재와 제조 방법을 통해 반도체 성능의 경계를 넓혀 더 빠르고 효율적인 AI 계산이 가능하도록 하고, 칩렛같은 칩 설계에 대한 고급 모듈식 접근 방식을 제공하여 더 작고 특수한 칩들을 하나의 완전한 시스템으로 통합하도록 하여 기술 통합을 확대하고 유연성을 높여 전반적인 성능을 향상시킨다. 여러 부품을 하나의 응집된 단위로 조립하고 연결하는 패키징 기술은 칩을 수직으로 적층하고 고속 상호연결을 사용하는 등의 전략을 활용하여 부품 밀도를 높여 성능을 향상시킨다. 또한 고성능 AI 작업에 필요한 엄청난 컴퓨팅 수요를 충족하기 위해 고성능 AI/HPC 데이터 센터의 각 컴퓨팅 자원에서 분산 병렬 처리를 하고 병렬 처리를 위한 워크로드의 배분과 렉(Rack)이상 규모로 확장된 도메인에서도 모듈형 컴퓨팅 자원의 상호 인터페이스가 서로 연결되기 위해서 네트워크의 고도화가 중요하고 저지연 상호 연결이 가능해야 성능 확장이 가능하다.

Ⅱ. 본론

최근 인공지능과 다양한 과학 분야의 급속한 발전으로 인해 AI/HPC 데이터 센터에 더 높은 컴퓨팅 성능에 대한 필요성이 커지고 되었고, 시뮬레이션 및 대용량 워크플로우 실행을 목표로 하는 AI 등 다양한 연구에 대한 고성능 컴퓨팅(HPC)은 성능 측면에서 엑사플롭 경계를 넘어섰습니다. 거대한 워크로드를 분산처리하기 위해서 CPU, GPU 같은 컴퓨팅 장치 사이에 데이터를 전달하는 것이 매우 중요하다. 기존 네트워크 방식으로 병렬 노드에 워크로드 데이터를 배분할 때, 성능 개선이 가능하지만 분산 데이터 이동을 위한 대기시간이 커질 수 있다 [1].

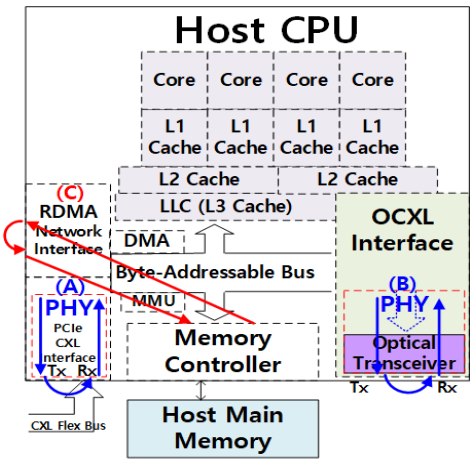

그림 1. AI/HPC 확장 인터페이스 검증 플랫폼

그림 1 은 AI/HPC 확장을 위해 현재 새롭게 논의되는 인터페이스의 검증을 위한 플랫폼으로 컴퓨팅 자원 장치인 가속기나 분리 메모리에 상호 연결이 가능한 호스트 CPU 에 세 가지 인터페이스가 구현되었으며, 성능 비교가 가능하다. 각 인터페이스는 CXL 에 활용 가능한 PCIe 물리 계층을 지원하고[4], 광 연결을 위한 이더넷 물리 계층도 지원한다. 플랫폼은 Gen1 ~ Gen5 PCIe 모드 또는 다중 레인을 갖춘 10Gbps 및 25Gbps 광 모드를 지원하는 AMD UltraScale+ 장치에서 CXL 의

전기 PCIe PHY, 광 연결 CXL 의 광 PHY, RDMA 를 지원하는 광 이더넷 등 세 가지 인터페이스를 통해 성능 비교가 가능하다.

그림 2 은 그림 1 의 각 인터페이스의 적용을 통해 AI/HPC 컴퓨팅 성능 요구에 따라 기존 데이터 센터와는 다른 AI/HPC 데이터 센터 내부 네트워크 구조를 도시한 것이다. AI/HPC 데이터 센터는 기존 데이터 센터 네트워크 구조와는 다르게 세가지 네트워크 구조를 가져야 한다. 네트워크 구조가 세분화되는 이유는 바로 컴퓨팅 자원의 모듈화와 LLM(Large Language Model) 워크로드의 AI 분산 고성능 처리를 위해서 모듈 형태의 자원 간에 서로 다른 속성의 네트워크가 필요하기 때문이다.

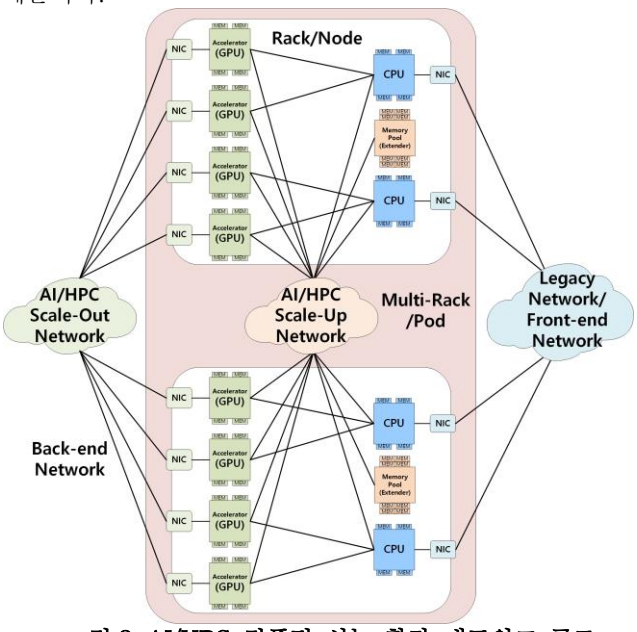

그림 2. AI/HPC 컴퓨팅 성능 확장 네트워크 구조

첫번째 기존 네트워크는 CPU 의 네트워크 포트 또는 네트워크 인터페이스 카드 (NIC)를 통해 연결되는 네트워크로 기존의 데이터 통신용 네트워크이다. 주로 기존 노드의 개수를 증가시키고, 노드 간의 네트워크 규모 확장 (Scaling-out)을 고려한 네트워크이며 이더넷 (Ethernet)과 IP 기반의 네트워크로 L2/L3 스위치 라우터를 활용해서 구성되는 네트워크로 광 전달망 장비등을 통해 외부망 또는 외부 데이터센터와 연결이 가능하고, 흔히 Front-end 네트워크 라고도 한다. 두번째 네트워크와 관련하여 AI/HPC 와 성능을 높이기 위해서 기존의 데이터 통신 개념의 네트워크는 컴퓨팅 성능을 향상시키기에는 한계가 있어, RDMA over Converged Ethernet (RoCE)와 InfiniBand 프로토콜을 활용하여 주로 AI 워크로드를 가속기에 분산하기 위한 목적이 크다. RDMA 를 통해 CPU 개입 없이 데이터를 전송하여 지연 시간 감소, 확장성이 증가하므로 GPU, 가속기에 직접 NIC 인터페이스 적용할 수 있게 되었다. AI/HPC 를 위한 데이터 센터의 Back-end 네트워크로 구분되어 가속기의 연결 네트워크 규모 확장 (Scaling-out)과 관련이 있다. 세번째는 AI/HPC 의 근본적인 컴퓨팅 성능 개선을 위해서 필요한 네트워크이다. 컴퓨팅 자원 간의 상호 연결을 네트워크 규모로 확장하기 위한 네트워크이며, CPU, 가속기, 메모리 모듈의 연결을 스위치를 이용하여 네트워크 형태로 확장하여 가속기를 활용한 AI 컴퓨팅 성능 향상에 있어 공유 메모리 등을 활용하여 대기 시간을 획기적으로 줄이고 CPU, 가속기, 메모리의 연결과 메모리 인식을 통해 직접 연결을 통한 성능을 확장 (Scaling-Up)을 하기 위한 네트워크이다.

Ⅲ. 결론

AI 성능 확장 구조는 CPU, GPU, 가속기, 메모리 등의 컴퓨팅 자원이 모듈식으로 구현되어서 하나의 장치로 인식되게 될 것이다. 이 자원 간의 연결 네트워크는 저지연, 광대역으로 각 자원 모듈을 연결하여야 한다. CPU, GPU, 가속기와 같은 처리장치 자원과 각 처리장치에 부착된 부착 메모리와 원격 메모리가 구분이 필요하고, 이 컴퓨팅 자원들이 네트워크 요소로 등장할 수 있다. 현재 관련 네트워크 시장은 NVIDIA 고유 연결 프로토콜(NVLink)과 InfiniBand 프로토콜을 바탕으로 시장을 지배하고 있다[3]. 하지만 현재 CXL 을 비롯하여 컴퓨팅 자원간 연결을 위한 추가적인 개방형 연결 프로토콜이 등장하면서 개방형 네트워크로 발전할 수 있는 초기 국면에 진입했다[4-6]. 그래서 CPU-GPU-Memory 연결은 여러 기업의 제품이 다양하기 때문에 함께 참여하는 공통 연결 프로토콜을 통해 연결하려고 하는 요구 높아 졌고, 대부분의 주요 기업이 개방형 연결 프로토콜 개발에 참여하여 앞으로 해당 기술이 적용된 여러 CPU, 가속기, 메모리 모듈 또는 ASIC 칩이 많아 질 것이다.

ACKNOWLEDGMENT

이 논문은 2024 년 정부 (과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임. (No. 2019-0-00002, 광 클라우드 네트워킹 핵심원천 기술 개발).

참 고 문 헌

[1] C.-C. Yang and G. Cong, "Accelerating Data Loading in Deep Neural Network Training," 2019 IEEE 26th International Conference on High Performance Computing, Data, and Analytics (HiPC), Hyderabad, India, 2019, pp. 235-245.

[2] D.-U. Kim, at al., "Optically Networked Heterogeneous Data-centric Computing System with Silicon Photonics Transceivers," 2024 Optical Fiber Communications Conference and Exhibition (OFC), San Diego, CA, USA, 2024, M1G.2, pp. 1-3.

[3] https://www.nvidia.com/ko-kr/ accessed on March 2025

[4] Compute Express Link (CXL) Specification, Rev.3.1, Aug. 2023.

[5] https://computeexpresslink.org/ accessed on March 2025

[6] https://ualinkconsortium.org/ accessed on March 2025.