이기종 자원을 위한 버스 확장 시스템 구현

차광호, 구경모
한국과학기술정보연구원 슈퍼컴퓨팅기술개발센터
khocha@kisti.re.kr, kookm@kisti.re.kr

Implementation of Bus Expansion System for Heterogeneous Computing Resources

Kwangho CHA, Kyungmo Koo
Center for Supercomputing Technology Development,
Korea Institute of Science and Technology Information

요 약

여러 인공지능 서비스의 보급은 초고성능 컴퓨팅 시스템 아키텍처의 변화를 야기하였고 다양한 계산 자원들의 활용이 모색되고 있다. 본 연구에서는 이러한 계산 자원들의 수용을 위해 범용적으로 사용되는 PCIe 버스를 기반으로 시스템 버스 확장 장치를 설계하고 구현하였다. PCIe 4.0 스위치를 기반으로 하는 확장 보드와 어댑터 카드를 개발하였고 GPU를 활용하여 실제 시스템으로의 활용 가능성을 검증하였다.

1. 서론

최근 다양한 AI 서비스가 활발히 보급되면서 초고성능 컴퓨팅 시스템 아키텍처에도 이러한 변화가 반영되고 있다. GPU, NPU, FPGA 등 이기종 자원의 활용을 그 예로 들 수 있는데 해당 계산 자원을 위한 전용 시스템 버스 또는 인터커넥션 네트워크를 사용할 수도 있지만 오랜 기간 범용 IO 버스 역할을 수행한 PCIe 버스를 활용하는 방법도 고려할 수 있다.

본 연구에서는 이러한 계산 자원을 수용하기 위해 자체 개발한 PCIe 4.0 기반 확장 시스템을 소개하고자 한다.

2. 시스템 버스 확장 기술

GPU나 FPGA와 같은 이기종 계산 자원을 Scale up 방식으로 집적하기 위해서는 이 자원들을 연동하기 위한 시스템 버스 기술이 요구된다. NVLink[1], Infinity Fabric[2]등 같은 전용의 인터커넥션 네트워크 기술을 활용할 수도 있지만 기존 시스템과의 호환성과 비용 등을 감안하면 범용 IO 버스인 PCIe 버스의 사용도 대안이 될 수 있다.

PCIe 버스는 오랜 시간 동안 IO 버스의 역할을 수행하면서 많은 사용자층을 확보하고 있으며 표 1과 같이 꾸준히 그 성능을 개선하고 있어서 새로 등장하는 시스템 버스 기술들의 참조 모델이 되어 왔다. 특히 최근 많은 관심을 받고 있는 CXL 버스 역시 물리 계층은 PCIe 버스 기술을 사용하고 있다는 점도 PCIe 버스의 영향력을 보여준다고 할 수 있다.

본 연구에서 사용된 PCIe 스위치는 PCIe 버스를 확장하고 지능화된 서비스를 제공하기 위한 일종의 SoC 장치이다. 단순히 연결 가능한 디바이스 수를 증가시키는 기능 이외에도 NTB(Non-Transparent Bridge)를 통한 호스트 PC간 통신 및 고속 데이터 전송을 위한 DMA 기능 등을 제공하는 제품들이 존재한다.

<표 1> PCIe 버스의 주요 제원[3]

PCIe 버전	데이터 전송률 (GT/s)	인코딩 방식	대역폭 (단방향 x16 기준, GB/s)
1.0	2.5	8b/10b	4
2.0	5.0	8b/10b	8
3.0	8.0	128b/130b	15.75
4.0	16.0	128b/130b	31.51
5.0	32.0	128b/130b	63.01
6.0	64.0	FLIT (PAM-4)	~128

3. PCIe 4.0 기반 확장 시스템

본 연구에서 목표로 하는 시스템 버스 확장 시스템은 다양한 이기종 계산자원의 수용과 경우에 따라서는 호스트 PC간 통신에도 활용할 수 있도록 계획되었다.

이를 감안해서 PCIe 스위치 프로세서로는 PCIe 4.0 지원을 기본으로 NTB와 DMA 기능을 제공하는 제품군을 검토하여 Microchip사의 Switchtec PM40100과 PM40036을 사용하였다[4]. 각각 PCIe 4.0 100레인과 36레인을 제공하며 NTB 포트와 DMA엔진을 포함하고 있다. PM40100은 그림 1과 같은 형상으로 구성된 백플레인 보드의 개발에 사용되었으며 이 백플레인 보드는 다양한 디바이스들을 위한 인터페이스를 준비하여 다음의 기능을 제공할 수 있다.

① 다양한 계산 자원 수용

② 호스트 PC간의 인터커넥션 네트워크 제공

③ 백플레인 보드간 연결 기능 제공

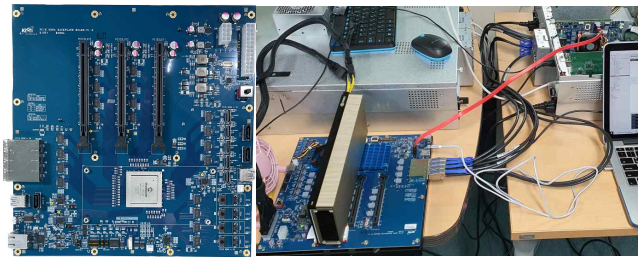
또한 PM40036은 호스트 PC와 백플레인 보드에 장착되는 어댑터 카드를 위한 SoC로 사용되었으며 이 어댑터 카드는

① 호스트와 백플레인 보드와의 연결, 또는

② 호스트 PC간의 직접 연결에 사용할 계획이다. 이러한 확장 시스템들의 연결용 미디어로는 mini SAS HD케이블과 광 케이블을 사용할 수 있도록 하였다.

4. 성능 평가

그림 2와 같이 개발된 시스템의 Throughput 성능을 검증하였다. Nvidia A100 GPU와 호스트 메모리간의 Throughput을 측정하였는데 GPU를 호스트 PC에 직접 연결하는 경우와 개발된 백플레인 보드에 연결한 경우로 나누어서 성능을 측정하였다. 벤치마크 프로그램으로는 CUDA-Tool인 bandwidthTest를 사용하였다.

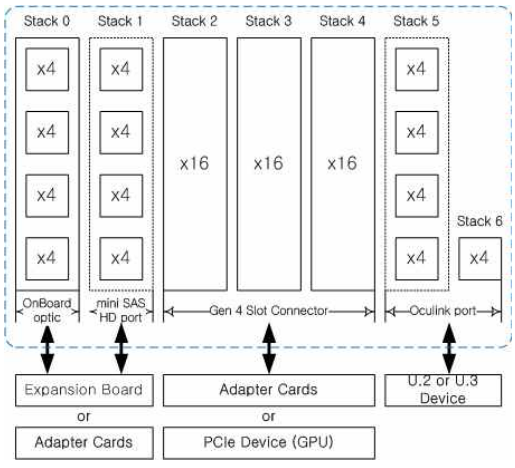


(그림 2) 백플레인 보드 외형(좌) 및 서버 시스템과 연동 모습(우)

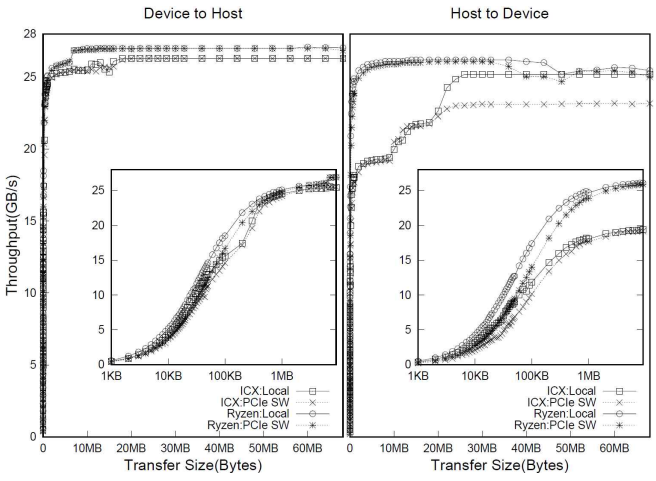
GPU가 백플레인 보드에 연결된 경우에는 백플레인 보드, miniSAS HD 케이블 그리고 어댑터카드를 거쳐서 호스트 PC에 도달하게 되며 2개의 PCIe 스위치를 통과하게 된다. 호스트 PC로는 인텔 Xeon Ice Lake를 사용하는 서버와 AMD Ryzen 5를 사용하는 PC를 사용하였다.

그림 3에 실험 결과를 정리하였다. 전송 데이터의 크기가 작은 경우에는 확장 시스템을 거치는 경우의 성능이 로컬 PCIe 버스에 직접 연결된 경우 보다 하회하는 성능을 보였으나 전송 데이터의 크기가 증가하여 일정 수준에 도달한 경우에는 성능 저하가 아주 미비하였고 이러한 현상은 GPU에서 호스트 메모리로 데이터를 전송하는 경우에 더욱 확실하게 나타났다.

다만 호스트 메모리에서 GPU로 데이터를 전송하는 경우에는 성능 저하가 좀 더 눈에 띄게 나타났다. 특히 인텔 Ice Lake를 사용하는 시스템에서는 약 8.2% 정도의 성능 저하를 보였다. 이에 대해서는



(그림 1) PCIe 4.0 스위치 기반 백플레인 보드 형상



(그림 3) 네트워크 처리량 측정 결과

개발된 확장 시스템뿐만 아니라 실험에 사용된 호스트 서버의 특징을 함께 검토하면서 원인을 분석 중이다.

5. 결론

본 연구에서는 이기종 계산 자원 수용을 위한 PCIe 버스 확장 시스템을 제작하고 그 성능 평가 과정을 소개하였다. PCIe 버스의 확장을 위해 PCIe 스위치들이 데이터 경로에 추가되는 만큼 어느 정도의 성능 저하를 피할 수는 없으나 전송 데이터의 크기가 증가한 경우에는 성능 저하도가 일정 부분 상쇄됨을 확인할 수 있었다.

현재는 확장 시스템의 안정성 개선을 위한 리비젼을 진행 중이며 디바이스들의 확장 연결 기능뿐만 아니라 호스트 PC간 통신에 활용하기 위한 기능 및 성능 검증을 계획하고 있다.

ACKNOWLEDGMENTS

본 연구는 2023년도 한국과학기술정보연구원(KISTI)의 기본사업으로 수행된 연구입니다. (과제번호: K-23-L02-C06)

참고문헌

[1] Nvidia, "NVLink and NVSwitch,", Retrieved Sep. 26, 2023 from https://www.nvidia.com/ko-kr/data-center/nvlink/

[2] Ian Cutress, "AMD Moves From Infinity Fabric to Infinity Architecture: Connecting Everything to Everything," AnandTech, Retrieved Sep. 26, 2023 from https://www.anandtech.com/show/15596/amd-moves-from-infinity-fabric-to-infinity-architecture-connecting-everything-to-everything

[3] Casey Morrison and Jonathan Bender, “Seamless Transition to PCIe 5.0 Technology in System Implementations,” PCI-SIG Webinar, Dec. 9, 2020.

[4] Microchip, "Switchtec™ PCIe® Switches," Retrieved Sep. 26, 2023 from https://www.microchip.com/en-us/products/interface-and-connectivity/pcie-switches