PCIe 기반 다중 NPU 데이터 전송 최적화

PCIe-Based Multi-NPU Data Transmission Optimization

오 세 현[†‡], 권용인[†‡], 이 제 민[†‡*]
[†]과학기술연합대학교대학원, [‡]한국전자통신연구원
(Sehyeon Oh, Yongin Kwon, Jemin Lee)
([†]University of Science and Technology, [‡]Electronics and Telecommunications Research Institute)

Abstract : 고성능 연산을 요구하는 인공지능 응용 프로그램의 증가로 분산 처리와 데이터 전송의 효율성이 중요해지고 있다. 본 논문에서는 Neubla Antara NPU를 활용하여 PCIe를 통해 다중 NPU 간의 데이터 전송 성능을 분석하고, 두 가지 전송 방안을 제안한다. 첫 번째는 BAR 기반 전송 방식으로, CPU가 NPU들간의 데이터를 복사하여 전송한다. 두 번째는 DMA와 CPU 전송을 결합한 하이브리드 전송 방식으로, CPU 개입을 최소화하여 더 빠른 데이터를 전송을 가능하게 한다. 성능 분석 결과, BAR 기반 전송은 24.33 MB/s, 하이브리드 전송은 1.94 GB/s의 속도를 기록하였다. 두 방식 모두 Peer-to-Peer 통신이 지원되지 않는 환경에서 데이터 전송을 위한 대안으로 활용 가능함을 확인할 수 있다.

Keywords : Nerual Processing Unit, Peer-to-Peer, Peripheral Component Interconnnect Express, Base Address Register, Direct Memory Access

Ⅰ. 서 론

최근 고성능 연산을 요구하는 응용 프로그램의 증가로, GPU와 NPU 같은 하드웨어 가속기를 활용한 분산 처리의 중요성이 커지고 있다[1]. 이 환경에서 가속기 간 데이터 전송 효율성은 시스템 성능에 큰 영향을 미친다. 이를 최적화 하기 위해 집합 통신 라이브러리(Collective Communication Library)[2]가 사용되며, NVIDIA는 NCCL, Intel은 OneCCL을 통해 GPU와 CPU 간의 데이터 전송을 최적화하고 있다. 이들 라이브러리는 P2P(Peer-to-Peer) 통신을 활용한다. 그러나 P2P 통신을 활용하지 않는 경우에도 효율적인 대안이 필요하다.

본 논문에서는 Neubla Antara NPU를 활용하여 PCIe 기반으로 다중 NPU 간의 데이터 전송 성능을 분석하고 P2P 통신이 지원되지 않는 환경에서 데이터 전송을 위한 두 가지 방안을 제안한다. 첫 번째는 BAR(Base Address Register) 기반 전송 방식으로, CPU가 NPU간 데이터를 복사하는 방식이다. BAR는 NPU의 메모리나 자원을 CPU의 주소 공간에 매핑하는 역할을 하며, CPU는 이를 통해 NPU의 메모리에 접근하여 데이터를 전송할 수 있다. 두 번째는 DMA(Direct Memory Access)와 CPU 전송을 결합한 하이브리드 전송 방식이다. 이 방식은 DMA를 활용하여 CPU 개입을 최소화하고, 보다 효율적인 데이터 전송을 가능하게 한다. 하이브리드 방식은 NPU와 CPU 간 데이터 전송을 DMA로 처리하고, CPU 간 메모리 복사를 결합하여 NPU 간 데이터를 간접적으로 전송하는 구조이다. 두 가지 방식 모두 P2P 통신이 불가능한 환경에서의 데이터 전송을 목표로 하며, 각 방식의 성능을 실험적으로 분석하고 비교한다.

*Corresponding Author (leejaymin@etri.re.kr)
이제민: 한국전자통신연구원
※ 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통기획평가원의 지원을 받아 수행된 연구 결과임 (No.RS-2024-00459797, 온디바이스 AI를 위한 ML컴파일러 프레임워크 기술 개발)

Ⅱ. PCIe 데이터 전송 구조

그림1은 NPU 및 PCIe BAR 기반 데이터 전송 토폴로지를 보여준다. CPU는 루트 컴플렉스를 통해 NPU0와 NPU1에 연결되며, BAR를 사용해 각 NPU 메모리를 CPU 주소 공간에 매핑한다. 그림2는 NPU 카드 내부의 PCIe BAR 매핑 구조를 보여준다. NPU BAR는 NPU의 DDR, SFR, PCIe 컨트롤러 등 다양한 자원에 매핑된다. PCIe 설정 영역을 통해 PCIe 컨트롤러를 제어하고 DMA를 설정할 수 있으며, NPU DDR 영역을 통해 호스트 CPU가

NPU DDR에 접근할 수 있다. NPU BAR는 CPU 주소 공간에 할당된 주소와 연결되어, CPU는 이 주소를 통해 NPU 자원에 접근한다.

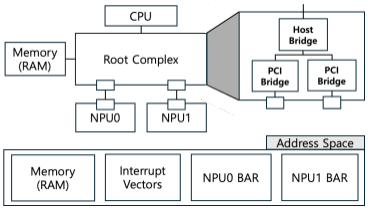

그림 1. PCIe 데이터 전송 토폴로지
Fig. 1. Topology of PCIe Data Transmission

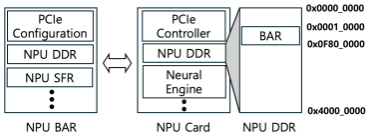

그림 2. NPU와 BAR 간의 PCIe 데이터 전송 구조
Fig. 2. PCIe Data Transmission Structure Between NPU and BAR

Ⅲ. 제안하는 시스템

본 논문에서는 BAR 기반 전송 방식과 DMA와 CPU 전송을 결합한 하이브리드 방식을 비교하여 제안한다. 그림 3은 BAR 기반 전송 방식을 보여준다. 이 방식에서는 NPU0과 NPU1의 DDR 메모리가 각각 BAR를 통해 호스트 PC의 주소 공간에 매핑되며, CPU는 NPU0에서 NPU1으로 데이터를 복사한다. 이 방식은 CPU가 데이터 전송을 중개하는 방식으로, CPU의 처리 능력과 메모리 대역폭에 의해 전송 속도가 제한된다. 그림 4는 DMA와 CPU 전송을 결합한 하이브리드 전송 방식을 나타낸다. NPU0와 NPU1의 DDR 메모리는 각각 DMA 채널을 통해 DMA 버퍼0과 DMA 버퍼1로 데이터를 전송하고, CPU는 이 두 버퍼간의 데이터를 처리한다. 이 방식은 DMA를 사용하여 CPU 개입을 최소화하면서, 더 효율적인 데이터 전송을 가능하게 한다.

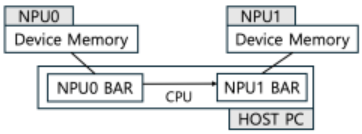

그림 3. BAR 기반 전송 방식
Fig. 3. BAR-based Transmission Method

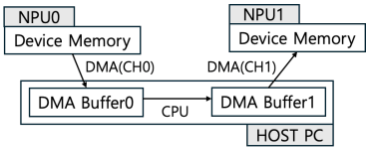

그림 4. DMA와 CPU의 전송을 결합한 하이브리드 전송 방식
Fig. 4. Hybrid Transmission Method combining DMA and CPU Transfer

Ⅳ. 실험 결과

1. 실험 환경

실험은 Intel I7-12700 CPU와 두 개의 Neubla Antara NPU 보드를 장착한 호스트 컴퓨터에서 진행되었다. NPU0는 PCIe Gen5 x16 슬롯, NPU1은 PCIe Gen3 x4 슬롯에 장착되었으며, 두 보드는 PCIe Gen4 x8을 지원한다.

2. 결과 및 평가

그림 5는 BAR PCIe Gen5 x 16 슬롯에 NPU를 장착하여 BAR 전송 속도와 DMA 전송 속도를 측정한 결과를 보여준다. NPU는 PCIe Gen4 x8을 지원하므로, 실제 전송 속도는 PCIe Gen4 x8 대역폭에 의존한다. BAR 전송 속도는 CPU에서 NPU로 6.76 MB/s, NPU에서 CPU로 28.11 MB/s로 측정되었다. DMA 전송 속도는 CPU에서 NPU로 8014.24 MB/s, NPU에서 CPU로 9558.77 MB/s로 측정되었으며, 이는 PCIe Gen4 x8의 이론 대역폭인 15.754 GB/s 대비 49.68%와 59.25%에 해당된다.

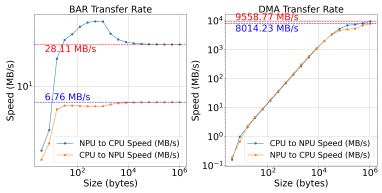

그림 5. BAR 및 DMA 전송 방식의 CPU/NPU 간 데이터 전송 속도 비교 (PCIe Gen5 x 16)

Fig. 5. Comparison of CPU/NPU Data Transfer Speeds for BAR and DMA Transmission Methods(PCIe Gen5 x16)

그림 6은 PCIe Gen3 x4 슬롯에 NPU를 연결한 경우의 BAR 전송 속도와 DMA 전송 속도를 나타낸다. BAR 전송 속도는 CPU에서 NPU로 3.22 MB/s, NPU에서 CPU로 28.12 MB/s로 측정되었다. DMA 전송 속도는 CPU에서 NPU로 2452.07 MB/s, NPU에서 CPU로 3023.30 MB/s로 측정되었으며, 이는 PCIe Gen3 x4의 이론 대역폭인 3.94 GB/s 대비 각각 약 60.80%와 74.96%에 해당한다.

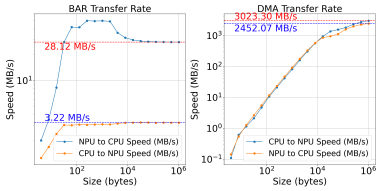

그림 6. BAR 및 DMA 전송 방식의 CPU/NPU 간 데이터 전송 속도 비교 (PCIe Gen3 x4)

Fig. 6. Comparison of CPU/NPU Data Transfer Speeds for BAR and DMA Transmission Methods(PCIe Gen3 x4)

그림 7은 NPU 간 데이터 전송을 비교한 결과로, BAR 기반 전송과 하이브리드 전송을 보여준다. BAR 전송 속도는 24.33 MB/s, 하이브리드 전송 속도는 1988.28 MB/s로 측정되었다. 하이브리드 방식은 DMA를 활용해 CPU 개입을 최소화하여 더 빠르고 효율적인 전송을 가능하게 한다. 이 방식은 BAR 방식보다 PCIe 대역폭을 더 효율적으로 활용하며, 대역폭이 높아질수록 전송 속도도 더 빠르게 향상된다.

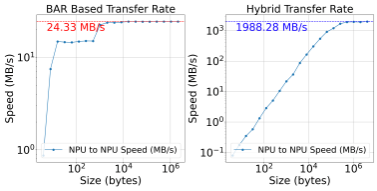

그림 7. BAR 기반 전송과 하이브리드 전송 방식의 NPU 간 데이터 전송 속도 비교

Fig. 7. Comparison of Inter-NPU Data Transfer Speeds for BAR-based and Hybrid Transmission Methods

V. 결 론

본 논문에서는 PCIe 기반 다중 NPU 간의 데이터 전송 성능을 분석하고, 두 가지 전송 방안을 제안하였다. 첫 번째는 BAR 기반 전송 방식으로, CPU가 NPU 간 데이터를 복사하여 전송하는 방식이다. 두 번째는 DMA와 CPU 전송을 결합한 하이브리드 전송 방식으로, CPU 개입을 최소화하여 더 빠른 데이터 전송을 가능하게 한다. 성능 분석 결과, BAR 기반 전송은 24.33 MB/s, 하이브리드 전송은 1.94 GB/s의 전송 속도를 기록하였다. 이를 통해 두 방식 모두 P2P 통신이 지원되지 않는 환경에서 데이터 전송을 위한 대안으로 활용 가능함을 확인할 수 있다.

Reference

[1] C. Silvano, D. Ielmini, F. Ferrandi, L. Fiorin, S. Curzel, L. Benini, and R. Birke, "A survey on deep learning hardware accelerators for heterogeneous HPC platforms," arXiv preprint, arXiv:2306.15552, 2023.

[2] A. Weingram, Y. Li, H. Qi, D. Ng, L. Dai, and X. Lu, "xCCL: A survey of industry-led collective communication libraries for deep learning," Journal of Computer Science and Technology, vol. 38, no. 1, pp. 166-195, 2023.