노드간 통신을 위한 PCIe 어댑터 카드 성능 분석

차광호
한국과학기술정보연구원 슈퍼컴퓨터개발센터
e-mail: khocha@kisti.re.kr

Performance Evaluation of PCIe Adapter Card for Inter-node Communications

Kwangho CHA
Div. of Supercomputing, Center for Developement of Supercomputing System
Korea Institute of Science and Technology Information

요 약

PCIe 스위칭 기술을 이용하면 시스템 내부의 디바이스들의 연결을 넘어 노드간 통신에도 PCIe 버스를 활용할 수 있다. 이때 기존의 확장 케이블이 갖는 물리적인 한계를 극복하기 위하여 광통신을 활용하는 경우가 있다. 본 연구에서는 초소형 온 보드형 광모듈을 활용하여 자체 제작한 PCIe 어댑터카드의 NTB(Non-Transparent Bridging) 포트를 활용하여 노드간 통신을 수행한 결과를 소개한다.

1. 서론

PCIe(PCI Express) 스위칭 기술은 단순히 시스템 버스의 확장에 그치지 않고 타 시스템과의 연결도 가능케 하는 특징을 가지고 있다. 즉, 기존의 디바이스간 연결을 위한 시스템 버스 기능과 더불어 시스템 인터커넥트로의 활용도 모색할 수 있는 상황이라 할 수 있다[1]. 그러나 이러한 연결성을 축종시키기 위해서는 기존 확장 케이블의 제약을 해결하여야 하는 문제가 있다. 이를 위해 우리는 기존 연구[2,3,4]처럼 광통신을 통해 PCIe 버스 신호를 보낼 수 있는 방법을 모색하였고 특히 초소형의 온 보드형 광모듈을 활용한 PCIe 어댑터 카드를 개발한 바 있다. 이 어댑터 카드를 활용하는 이전 연구에서는 PCIe 버스를 확장하는 경우, 90Gbps 후반의 성능을 확인한 바 있다[5].

본 연구에서는 이 어댑터 카드의 NTB(Non-Transparent Bridging) 포트를 활용하여 노드간 통신이 수행됨을 검증하였고 그 성능과 향후 개선 방안을 소개하고자 한다.

2. 온 보드형 광모듈 기반 PCIe 어댑터 카드

본 연구에 사용된 그림 1의 PCIe 어댑터카드는 다음과 같은 구성 요소들을 활용하여 자체 제작되었다.

2-1. PCIe 스위칭

PCIe 어댑터카드의 주요 구성요소인 PCIe 스위칭 칩은 기존 PCIe 브릿지 칩의 발전된 형태로 PCIe 버스 신호의 확장과 더불어 PCIe 버스레인의 분기, 주소체계 변환 방법 등을 제공하고 있으며 환경 설정을 변경하여 스위칭 칩의 기능을 변경할 수 있는 방법을 제공하고 있다.

2-2. NTB 포트

본 연구에 사용된 어댑터 카드는 NTB(Non-Transparent Bridging) 포트를 가지고 있는 PCIe 스위칭 칩을 탑재하고 있다. 이 NTB 포트는 두 개의 시스템에 의해서 동일 디바이스가 검색되는 경우, 발생할 수 있는 문제를 해결하기 위하여 특정 시스템에는 해당 디바이스가 검색되지 않도록 격리시키는 목적으로 사용되었다. 특히 BAR(Base Address Register)를 이용한 주소 변환 기능은 트랜잭션 전송을 가능하게 하였고 메시지 확인용 레지스터를 사용하여 시스템들을 격리시키면서 통신은 가능하도록 하였다. 이러한 NTB 포트는 이후 패브릭 기능을 지원하기 위하여 각 칩 제조사별로 발전하게 되는데 Broadcom(구 PLX)의 경우에는 TWC(Tunneled Window Connection)를 그 예로 들 수 있다.

2-3. 온 보드형 광모듈

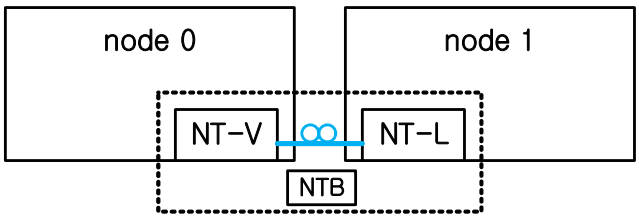
PCIe 버스를 확장하는 경우에는 주로 구리선 방식의 케이블을 활용하게 되는데 케이블의 굵기나 전송거리등에서 제약을 갖게 된다. 이러한 이유로 PCIe 버스 신호를 광통

(그림 1) 16배속 PCIe 어댑터 카드

신을 통해 전송하는 방법들이 제안되었고[2,3,4] 본 연구에 활용된 PCIe 어댑터 카드는 일반적은 광모듈과는 달리 물리적인 크기를 소형화한 온 보드형 광모듈을 사용하였다. 이때 온 보드형 광모듈은 소형화된 특징으로 인하여 하드웨어 메인보드 또는 어댑터 카드의 가장자리가 아닌 다양한 자리에 위치시킬 수 있다는 장점이 있다.

3. 실험 환경

앞서 설명한 그림 1의 16배속용 PCIe 어댑터 카드 2개를 그림 2처럼 서로 다른 서버에 장착하고 전용 광케이블로 연결하였다. 이 중 하나의 카드는 NTB 포트를 활성화하는 DIP 스위치를 설정하고 통신을 위한 레지스터 설정 값들을 EEPROM에 저장하여 NT-Virtual용 엔드 포인트로 활용할 수 있게 하였다. 또한 나머지 카드는 NTB 포트에 대한 설정 없이 연결하여 NT-Link용 엔드 포인트로 사용하였다.



(그림 2) 광통신 기반 연결 개념도

시스템간 통신을 위해 PCIe 스위칭 칩은 메시지 확인용 레지스터 이외에 DMA 엔진을 제공하며 단편화되지 않은 데이터를 고속으로 전송하는데 유용하게 사용된다. 통신 성능을 측정하는데 이 DMA 기능을 활용하였다.

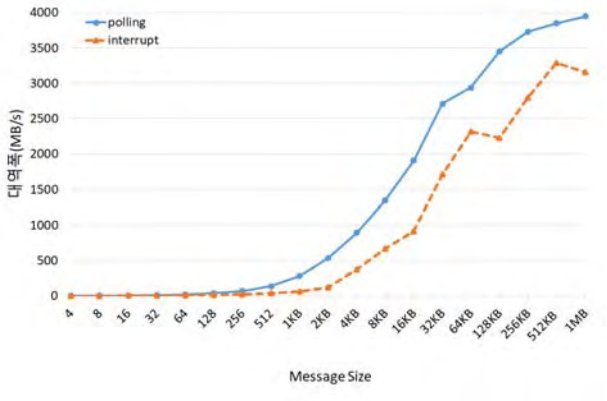
4. 성능 측정 결과

그림 3은 전송 데이터의 크기를 변경하면서 측정한 시스템간 통신 대역폭이다. 폴링 방식에 의한 DMA 상태 확인의 경우 약 30Gbps이상의 대역폭을 보이고 있다.

DMA 엔진의 데이터 전송 완료 여부를 확인하는 방식에 따라 성능차가 존재한다. DMA 엔진이 전송 완료시 인터럽트를 발생시키는 경우에는 시스템용 프로세서의 부하는 경감시킬 수 있으나 데이터 전송 성능은 그만큼 저하되는 것을 볼 수 있다. 전송하는 데이터의 크기가 작을 때는 DMA 인터럽트 오버헤드가 크게 나타남에 따라 폴링 방식 때의 대역폭 대비 21% 수준의 성능을 보였다. 데이터 크기가 증가함에 따라 어느 정도 인터럽트 오버헤드를 상쇄하는 모습을 보였지만 폴링 방식 대비 85% 수준의 대역폭을 보이고 있다.

5. 결론 및 향후 계획

본 연구에서는 PCIe 스위칭 칩과 온 보드형 광모듈을 사용하는 PCIe 어댑터 카드를 사용하여 노드 간 통신 성능을 확인하였다. NTB와 DMA를 활용하는 제약으로 인해 PCIe



(그림 3) 노드간 통신 대역폭

버스를 확장하는 경우에 비해서는 성능 저하가 확인되었다. 그러나 전문 인터커넥션 네트워크를 활용하지 않고 얻을 수 있는 성능임을 감안하면 그 활용도를 무시할 수만은 없는 상황이다. 향후 자체적으로 개발하고 있는 PCIe 기반 통신 라이브러리[6]와 본 PCIe 어댑터 카드를 연계하여 발전시킬 계획을 가지고 있다.

참고문헌

[1] 차광호, 유정록, 김성호, "노드간 통신을 위한 PCIe 스위칭 기술 연구," 한국정보처리학회 추계학술발표회 논문집, 56~58, 2015년 10월.

[2] Derek Percival, "PCIe over Fiber Optics: Challenges and Pitfalls," https://pcisig.com/sites/default/files/files/02_05_PCIe_Over_Fibre_Optics_Challenges_and_Pitfalls_FROZEN.pdf

[3] PLX Technology and Avago Technologies, "A Demonstration of PCI Express Generation 3 over a Fiber Optical Link" white paper, https://docs.broadcom.com/docs/AV02-3245EN

[4] A Triossi, D Barrientos, M Bellato, D Bortolato, R Isocrate, G Rampazzo and S Ventura, "A PCI Express optical link based on low-cost transceivers qualified for radiation hardness," Journal of Instrumentation, vol. 8, Feb. 2013

[5] Kyungmo Koo, Junglok Yu, Sangwan Kim, Min Choi, and Kwangho Cha, "Implementation of Multipurpose PCI Express Adapter Cards with On-Board Optical Module," The Journal of Information Processing Systems, vol. 14, no. 1, pp. 270~279, Feb. 2018.

[6] Cheol Shim, Kwangho Cha, Min Choi, "Design and implementation of initial OpenSHMEM on PCIe NTB based cloud computing," Cluster Computing, First Online: Feb. 2018.