BlueField-3 SmartNIC 기반 오프로드 연산의 성능 분석

박세영, 김현수, 임영빈
울산과학기술원

{seyeongpark, hyunsukim, ybim}@unist.ac.kr

Performance Analysis of Offloaded Operations on BlueField-3 SmartNICs

Seyeong Park, Hyunsu Kim, Youngbin Im
Ulsan National Institute of Science and Technology

요 약

본 논문은 NVIDIA BlueField-3 SmartNIC 을 대상으로 DMA, NVMe-over-Fabrics(NVMe-oF), 패킷 전송과 같은 대표적인 오프로드 연산의 성능 특성을 분석하였다. 각 연산의 단일 실행 및 복합 실행 환경에서 성능을 실험적으로 측정하고, 이를 제한하는 주요 병목 자원을 규명하였다.

Ⅰ. 서 론

최근 데이터센터에서는 AI 및 빅데이터 기술의 확산에 따라 100 Gbps 이상의 고성능 네트워크 인터페이스 카드(NIC)를 사용하고 있다. 그러나 이러한 고속 네트워크 환경에서는 패킷 처리 과정에서 과도한 CPU 사이클이 소모되어 데이터 연산에 활용 가능한 자원이 감소하는 문제가 발생한다. 이를 해결하기 위한 대안으로 프로세서, 메모리, 하드웨어 가속기 등을 내장하여 NIC 자체에서 복잡한 연산을 처리하는 SmartNIC 이 등장했다.

본 논문에서는 SmartNIC 의 대표적인 연산인 DMA, NVMe-oF 및 네트워크 전송을 대상으로, 단일 실행뿐 아니라 복수의 연산이 동시에 수행되는 상황에서의 성능을 실험적으로 분석한다. 특히 복합적인 SmartNIC 연산 간의 자원 경쟁 및 대역폭 제약이 전체 성능에 미치는 영향을 규명한다.

Ⅱ. 본론

1) SmartNIC 성능 병목 분석의 필요성

SmartNIC 은 일반적으로 FPGA, SoC, ASIC 기반 아키텍처로 구현되며 각각 비용, 유연성, 성능 측면에서 장단점을 가진다. 본 연구는 높은 유연성을 제공하는 SoC 기반의 NVIDIA BlueField-3 를 대상으로 한다. BlueField 는 Mellanox ConnectX NIC 과 SoC 를 결합하여 Arm 코어, DDR 메모리, 다양한 가속기와 DPA 를 통합한 구조를 가진다.

하지만 SmartNIC 이 제공하는 유연성에도 불구하고 일반적으로 호스트 프로세서 대비 연산 성능이 낮고 패킷이 SmartNIC 을 경유하면서 추가적인 복잡성이 발생해 성능의 이점이 상쇄될 수 있다는 한계가 있다. 따라서 SmartNIC 의 구조적 특성을 정확히 이해하고 병목 지점을 분석하는 것이 중요하다.

최근 SmartNIC 을 사용하여 호스트의 네트워크 및 I/O 스택을 오프로드하는 연구가 활발히 진행되고 있다. IO-TCP [1]는 NVMe-oF 를 사용하여 호스트의 I/O 연산을 SmartNIC 으로 오프로드하여 적은 호스트 코어만으로 콘텐츠 서버 성능을 향상시켰다. OS2G [2]는 GPU 와 SmartNIC 사이의 DMA 를 통해 데이터 경로를 최적화하였다. 이러한 오프로드 기술의 핵심이 되는 DMA, NVMe-oF, 패킷 전송은 모두 메모리와 PCIe bus 를 집중적으로 사용하는 연산이므로 동시에 실행될 경우 성능 병목이 생기기 쉽다. 하지만 기존 연구들은 주로 BlueField-2 와 같은 이전 세대 SmartNIC 을 사용하거나 400 Gbps 에 달하는 최대 대역폭을 완전히 활용하는 환경에서 테스트되지 못했기 때문에 앞으로의 연구에서 발생할 수 있는 성능 병목을 예측하기 어렵다. 따라서 본 논문은 최신 BlueField-3 의 대역폭을 최대로 사용하는 고부하 상태에서 성능 병목을 분석하여 오프로드 시스템을 설계하고 최적화하는데 중요한 자료를 제공하고자 한다.

2) 기존 SmartNIC 성능 분석 연구

많은 선행 연구에서 다양한 SmartNIC 아키텍처와 워크로드의 성능을 분석해왔다. [3]은 BlueField-2 의 하드웨어 구조를 분석하여 SmartNIC 의 통신 병목 현상을 분석했고, [4]는 NVMe-oF Target offloading 을 수행할 때 입출력 성능을 검증했으며, [5]는 BlueField-2,3 를 비교하여 연산 및 오프로딩 기능을 분석하였다.

그러나 이러한 연구들은 대부분 단일 기능 실행에 한정되어 있으며, 다양한 워크로드가 동시에 실행될 때 발생하는 간섭으로 인한 성능 하락의 원인을 알기 어렵다. 특히 Memory 또는 PCIe 대역폭 제한, CPU core 스케줄링 등의 문제로 이론적 최대 성능보다 낮은 결과가 도출될 수 있으며 이를 체계적으로 측정하는 연구가 필요하다.

3) 실험 방법 및 환경

실험 환경은 두 개의 서버로 구성된다. 각 서버는 Intel Xeon w5-3433 CPU 와 256 GB DDR5 메모리를 탑재하고 Ubuntu 20.04.6 을 실행한다. SmartNIC 은 BlueField-3 400 Gbps 이며 16 개의 코어로 구성된 Arm Cortex-A78 와 32 GB DDR5 메모리를 장착하고 있다.

DMA 는 DOCA DMA 라이브러리를 사용하여 호스트 메모리의 데이터를 SmartNIC 메모리로 읽는 처리량을 측정하였고, NVMe-oF 는 fio 를 사용하여 호스트에 장착된 9 개의 Seagate FireCuda-530 SSD 를 대상으로 무작위 읽기 워크로드를 실행하였다. 패킷 전송은 DPDK Pktgen 을 사용하여 외부 서버로 패킷을 전송하는 상황을 가정하였다.

4) 실험 결과 및 분석

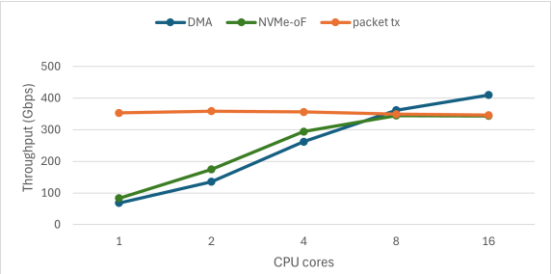

그림 1. 단독 사용시 최대 성능

그림 1 은 DMA, NVMe-oF, DPDK 패킷 전송이 각각 단독으로 실행될 때의 최대 성능을 보인다. DPDK 를 제외한 두 연산은 SmartNIC 코어 개수에 따라 성능도 증가함을 볼 수 있으며 각 연산의 성능을 제한하는 주요 병목 지점은 DMA 와 NVMe-oF 의 경우 PCIe 대역폭, 패킷 전송은 NIC line bandwidth 로 확인된다.

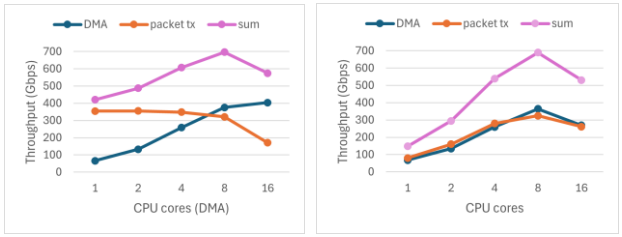

그림 2, 3. DMA 및 DPDK 패킷 전송 동시 사용 성능

그림 2, 3 은 DMA, Pktgen 을 동시에 실행한 결과이다. 그림 2 에서는 DPDK 코어 4 개를 사용하여 rate 를 고정한 상태에서 DMA 코어만 변화시킨 성능이고 그림 3 은 실제 오프로드 환경에서 DMA 연산에 따라 패킷을 전송하는 상황을 가정하여 DMA 성능과 동일하게 DPDK tx rate 를 조절한 성능이다. 두 연산의 처리량을 합한 최대 성능은 약 696 Gbps 로 측정되었다. 이는 두 연산이 SmartNIC 메모리 대역폭을 공유한다는 점에서 메모리 대역폭 병목으로 예상된다. 측정된 성능은 BlueField-3 의 이론적 최대 메모리 대역폭인 720 Gbps 에 근접하며 코어가 16 개일 때는 두 연산의 경합으로 최대 성능에 도달하지 못했다.

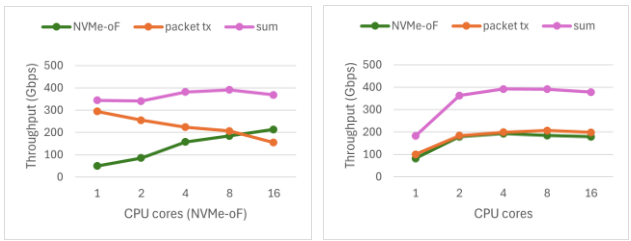

그림 4, 5. NVMe-oF, DPDK 패킷 전송 동시 사용 성능

그림 4, 5 는 NVMe-oF 와 Pktgen 을 동시에 실행한 결과이다. 그림 2, 3 과 마찬가지로 그림 4 에서는 DPDK 코어 4 개와 rate 를 고정하였고 그림 5 에서는 NVMe-oF 처리량만큼 패킷을 전송하도록 tx rate 를 조절하였다. 두 연산 모두 NIC 과 PCIe switch 사이의 PCIe 를 통해 외부로 나가는 데이터가 발생하기 때문에 PCIe bandwidth 에 제한이 생긴다. 이 경우, 두 연산의 데이터가 경유하는 PCIe Gen 5.0 최대 대역폭인 512 Gbps 를 성능 상한으로 예상했지만 실제 측정값은 400 Gbps 에 근접하였다. 이는 PCIe 버스상의 데이터 경합뿐만 아니라 프로토콜 처리 오버헤드 등 복합적인 요인이 성능 저하의 원인으로 작용하고 있음을 보인다.

Ⅲ. 결론

본 논문에서는 BlueField-3 의 DMA, NVMe-oF, DPDK 패킷 전송 성능을 단일 및 복합 실행 환경에서 측정하고, 연산별 병목 요인을 규명하였다. DMA 는 PCIe 대역폭, NVMe-oF 는 CPU 활용도, 패킷 전송은 NIC 라인 대역폭이 주요 병목 지점으로 작용함을 확인하였다. 복합 실행시에는 이러한 병목 요인이 중첩되면서 메모리 및 PCIe 대역폭에서 추가적인 자원 경쟁이 발생하여 이론적 최대 성능에 미치지 못하는 결과가 나타났다.

결론적으로, SmartNIC 을 활용한 오프로드 시스템에서 단일 연산의 성능 평가만으로는 실제 성능 특성을 충분히 설명할 수 없음을 보여준다. 따라서 향후 연구에서는 연산별 자원 소모 특성을 기반으로 한 효율적인 자원 관리 및 스케줄링 기법을 모색하여 SmartNIC 의 성능을 극대화하는 방향이 필요하다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학 ICT 연구센터(ITRC)의 지원(IITP-2025-II211817, 25%)과 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(RS-2025-00349594, 25%)과 과학기술정보통신부 및 정보통신기획평가원(IITP)의 재원으로 차세대 클라우드 네이티브 셀룰러 네트워크 리더십 프로그램의 지원(RS-2025-00418784, 25%)과 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(RS-2025-00405128, 25%)을 받아 수행된 연구임.

참 고 문 헌

[1] Kim, T. et al., “Rearchitecting the TCP Stack for I/O-Offloaded Content Delivery”, in Proc. 20th USENIC Symposium on Networked Systems Design and Implementation (NSDI), Boston, MA, USA, 2023

[2] Jin, Z. et al., “OS2G: A High-Performance DPU Offloading Architecture for GPU-based Deep Learning with Object Storage”, in Proc. 30th ACM International Conference on Architectural Support for Programming Language and Operating Systems (ASPLOS), Rotterdam, Netherlands, 2025

[3] Wei, X. et al., “Characterizing Off-path SmartNIC for Accelerating Distributed Systems,” in Proc. 17th USENIX Symposium on Operating Systems Design and Implementation (OSDI), Boston, MA, USA, 2023.

[4] Xu, J. et al., “Performance Characterization of SmartNIC NVMe-over-Fabrics Target Offloading,” in Proc. 17th ACM International System and Storage Conference (SYSTOR), Virtual, Israel, 2024

[5] Michalowicz, B. et al., “Battle of the BlueFields: An In-Depth Comparison of the BlueField-2 and BlueField-3 SmartNICs,” 2023 IEEE Symposium on High-Performance Interconnects (HOTI), CA, USA, 2023, pp. 41-48, doi: 10.1109/HOTI59126.2023.00020.