소프트웨어 정의 네트워크, 우선순위 기반 흐름제어, RoCE기술을 결합한 효율적인 초고성능 컴퓨팅 네트워크 구조 제안

최대진, 권태경

서울대학교 컴퓨터 공학부

djchoi@mmlab.snu.ac.kr, tkkwon@snu.ac.kr

A Proposal of High-Performance Computing Network Architecture based on Software-Defined Network, Priority-based Flow Control, and RoCE

Daejin Choi, Ted "Taekyong" Kwon

School of Computer Science and Engineering, Seoul National University

요 약

초고성능 컴퓨팅은 여러 컴퓨터를 하나의 네트워크로 연결하여 대용량의 데이터 처리나 대규모의 연산을 처리하기 위한 기술이다. 최근 대규모 연산 응용에 대한 요구가 높아짐에 따라, PFC, RoCE 등 초고성능 컴퓨팅 네트워크를 효율적으로 구축하기 위한 기술에 대한 연구가 활발하게 진행되고 있다. 그럼에도 불구하고, 데이터를 여러 개의 호스트에게 전달하기 위한 브로드캐스트 기법이나, FPC 데드락, 정지 프레임 스톰 등 효율적인 네트워크 구축에 대한 연구는 미비하다. 본 논문에서는 현재 HPC 네트워크에 사용되는 핵심 기법들에 대한 문제점을 파악하고, 이러한 문제점들을 소프트웨어 정의 네트워크 기법을 활용하여 해결할 수 있는 HPC 네트워크 구조를 제안한다.

Ⅰ. 서 론

초고성능 컴퓨팅 (High-Performance Computing, HPC)은 대규모의 연산이나 데이터 처리를 많은 수의 컴퓨터를 네트워크 형태로 구성하여 수행하고자 하는 컴퓨터 구조이며, 유전체 분석 등 복잡한 연산 뿐 아니라 데이터센터에 적합한 네트워크로 최근 주목을 받고 있다. [1]

이러한 초고성능 컴퓨팅 네트워크는 저지연 (Low-latency), 혼잡 제어 및 회피 (Congestion Control/Avoidance), 신뢰성 있는 통신 (Reliable Communication)등의 요구사항을 가지고 있으며 [1], 이를 효율적으로 만족시키기 위해, Torus 등 다양한 네트워크 토폴로지에 대한 연구, CPU를 사용하지 않고 원격 컴퓨터의 메모리에 데이터를 송/수신 할 수 있는 원격 직접 메모리 접근 (Remote Direct Memory Access, RDMA) 기술[2] 및 우선순위 기반 흐름 제어 (Priority-based Flow Control, PFC) 기술[3, 4] 연구 등 다양한 기술들을 접목하고 발전시키기 위한 연구가 진행되고 있다. 비록 이러한 연구들이 효율적인 HPC네트워크 구축에 많은 기여를 했음에도 불구하고, 브로드캐스팅 (Broadcast) 시 발생하는 네트워크 내 중복된 트래픽 생성, PFC 데드락 (Deadlock)[4], 정지 프레임 스톰[2] 등 다양한 문제가 여전히 해결되지 못하고 있다.

위의 문제를 해결하기 위해, 본 논문에서는 소프트웨어 정의 네트워크 (Software-Defined Network, SDN) 기술을 최신 PFC, RDMA기술과 결합한 형태의 HPC 네트워크 구조를 제안한다. 제안된 네트워크 구조는 네트워크 내 중복된 트래픽을 제거하고, HPC 응용 (Application)에 기반한 네트워크 경로 설정 및 흐름 제어 등의 기능들을 수행한다. 이러한 연구는 초고성능 컴퓨팅 네트워크 구축의 실제에 있어 초석연구가 될 것으로 기대한다.

Ⅱ. 본론

본 장에서는 효율적인 HPC 네트워크 구축을 위해 현재 이용되고 있는 세가지 핵심 요소 기술과 해당 기술들이 가진 한계점을 설명하고, 제안된 SDN을 활용한 HPC 네트워크 구조로써 해결책을 설명한다.

2.1. 우선순위 기반 흐름 제어 (Priority-based Flow Control, PFC) 기법

PFC는 무손실 이더넷 (Lossless Ethernet)을 만들기 위한 기술로써[3] Hop-by-Hop으로 네트워크 혼잡을 관리하는 방법이다. 네트워크 스위치는 8개의 서로 다른 우선순위 (Priority)와 각 우선순위에 대응되는 버퍼 (Buffer)를 개별적으로 관리하며, 하나의 버퍼에 혼잡이 발생하면, 해당버퍼에 진입하는 모든 네트워크 흐름에 대해 정지 프레임 (Pause Fame)을 전송함으로써 버퍼의 오버플로우 (Overflow)가 발생하지 않도록 하고, 이를 통해 네트워크 혼잡을 제어한다. PFC는 현재 대부분의 스위치에서 이 기능을 지원하고 있고, 구현이 간단하기 때문에 데이터센터나 HPC 네트워크 등 무손실 이더넷이 필요한 곳에 널리 이용되고 있다.

하지만, PFC기술은 특정 경우에 대해 부가적인 문재를 초래한다. 예를 들어, 정지 프레임을 전달하는 경로에 사이클(Cycle)이 발생할 경우 PFC 데드락 현상이 발생하며[4], 네트워크 트래픽이 높은 양으로 지속적으로 발생할 경우 발생하는 정지 프레임 스톰[1] 등이 발생한다. 그럼에도 불구하고, PFC 기술이 기본적으로 무손실 이더넷을 구현하는 핵심 요소기술로 인지되고 있기 때문에, 해당 문제들을 해결할 수 있는 다양한 연구들이 진행되고 있으며 PFC는 여전히 많이 이용되고 있다.

2.2. 원격 직접 메모리 접근 (Remote Direct Memory Access, RDMA) 및 RoCE

일반적인 네트워킹에서 하나의 호스트가 수신한 패킷은 네트워크 인터페이스 카드 (Network Interface Card, NIC)에 있는 버퍼에 복사되고, CPU연산에 의해 다시 호스트의 메모리로 옮겨지게 된다. 따라서, 최대한 많은 CPU연산을 요구하는 초고성능 컴퓨팅 응용에는 일반적인 네트워킹 방식을 사용할 경우 최대치의 CPU작업을 수행할 수 없기 때문에 비효율이 발생한다.

RDMA는 이러한 문제점을 해결하고자 별도의 CPU 연산없이 원격 호스트의 메모리에 바로 패킷이 전달되게끔 하는 기술로써, NIC에서 바로 DMA장치를 통해 CPU제어 없이 NIC의 버퍼에 있는 패킷을 직접 메모리

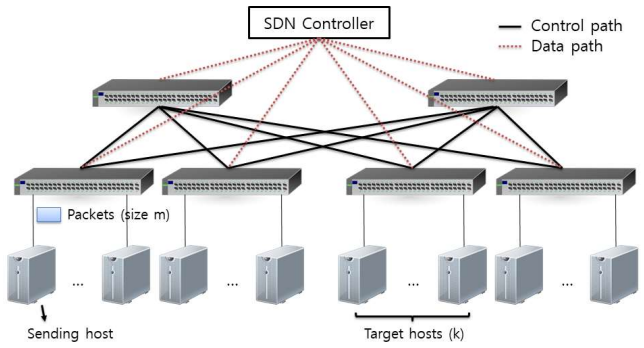
에 쓰는 작업을 수행한다[1]. 기존의 RDMA기술은 인피니밴드 (InfiniBand, IB)의 일대일 (One-to-One) 통신을 위해 개발되었으나, 최근에는 이더넷 상에서도 RDMA를 사용하는 RoCE (RDMA over Converged Ethernet) 기술이 개발되어 이용되고 있다. RoCE는 신뢰성 있는 통신을 지원하지 않기 때문에, PFC 등의 기술들이 이용된 무손실 이더넷 상에서 기능이 효율적으로 동작하는 것으로 알려져 있다. [1]

PFC, RoCE 두 기술만으로도 이더넷 기반 HPC나 데이터센터 네트워크에서 효율적으로 동작하는 것으로 알려져 있음에도 불구하고 [1, 2], RoCE기술이 일대일 통신을 위해서 개발되었기 때문에, 멀티캐스트나 브로드캐스트등 다수의 호스트로 데이터를 전송할 때 각 호스트와의 연결을 모두 맺고 중복된 트래픽을 전송하는 비효율성이 발생한다. 한편, 이러한 일대다수(one-to-many) 통신은 대용량 데이터 처리 등에서는 빈번히 발생하기 때문에, HPC나 데이터센터 네트워크에서는 필수적으로 해결해야 할 문제라고 볼 수 있다.

2.3. 소프트웨어 정의 네트워크 (Software-Defined Network, SDN)

SDN은 네트워크 내에 데이터를 전달하는 부분과 스위치들의 네트워크 경로를 제어하는 부분을 분리하여, 프로그래밍 가능한 SDN 콘트롤러 (Controller)를 통해 네트워크 상태에 따라 경로 및 혼잡 제어 등의 기능을 수행할 수 있도록 하는 기술이다. 기존 네트워크 시스템과 달리, SDN 기술은 프로그래밍 측면에서 유연성 (Flexibility)을 가지고 있음은 물론, 스위치에서 패킷 그룹화, 복사 등 일대다 통신을 위한 다양한 연산을 지원하기 때문에, 데이터센터 네트워크, HPC 네트워크에 그 활용 가능성이 높다고 볼 수 있다.

2.4. PFC, RoCE, SDN을 결합한 HPC 네트워크 구조



〔그림 1〕 팻-트리(Fat-Tree) 토폴로지에 기반한 SDN, RoCE, PFC가 결합된 HPC 네트워크 구조.

앞 절에서 언급하였듯이, PFC, RoCE, SDN은 각각 HPC나 데이터센터 네트워크의 핵심기술로써 널리 이용되고 있지만, 각 기술이 가지는 한계를 상호보완적으로 극복하려는 시도는 아직 부족하다. 본 논문에서 제시하는 HPC 네트워크는 PFC를 사용한 무손실 이더넷 상에서 RoCE를 통해 호스트 간 통신을 수행하도록 구성되어 있다. 이 때, HPC 응용이 서로 다른 네트워크에 있는 호스트 간에 브로드캐스트를 요구할 경우, 호스트가 직접 일대다 연결을 맺을 필요 없이, SDN 콘트롤러에서 패킷을 복사하여 전달하여 네트워크 내 중복된 트래픽을 최소화 하도록 한다.

그림 1은 HPC네트워크에서 많이 사용되는 팻-트리 (Fat-Tree)기반 토폴로지의 제안된 HPC 네트워크 구조를 나타낸다. 그림 1에서 하나의 호스트가 k 개의 호스트에 크기가 m인 데이터를 전송해야 할 경우, 기존 네트워크에서는 k개의 호스트에 대해 모두 연결을 맺고 각 호스트에 m만큼의 데이터를 전송하므로 송신 호스트로부터 수신 호스트까지의 경로에 m*k 만큼의 트래픽이 발생하게 되며, 평균 홉 수 (Hop Count)를 d라고 했을 때, 네트워크 전체에 m*k*d의 트래픽 부하가 발생하게 된다. 반면, 제안된 HPC 네트워크에서는 스위치 간 패킷 전송 시 m만큼의 트래픽만 이동하게 되고, 따라서 네트워크 전체에 m*(d-2) + 2*mk = m*(d-2+2k) 만큼의 트래픽이 발생하게 된다. 즉, 동일 크기의 패킷이 전달된다고 가정했을 때, 기존 네트워크에서는 홉 수와 전달하는 호스트의 수의 곱만큼 발생하던 트래픽이 제안된 네트워크에서는 선형함수만큼만 발생한다.

또한, 제안된 네트워크는 SDN 콘트롤러에 현재 네트워크 상황이 샘플링 (Sampling)되어 보고되므로, 기존 HPC 네트워크에 문제점으로 보고되었던 PFC 스톰, PFC 데드락 등에 대해서도 유연하게 대처가 가능하다. 예를 들어, 그림 1에서 특정 스위치에 PFC 스톰이 발생할 경우, 일시적으로 패킷들이 처리될 때까지 네트워크 흐름을 다른 스위치를 통해서 이동할 수 있도록 경로설정을 함으로써 정지 프레임이 발생하지 않도록 하는 등의 기능을 수행할 수 있다.

Ⅲ. 결론

본 논문에서는 초고성능 컴퓨팅 네트워크 구축을 위해서 현재까지 연구되는 핵심 기술들이 가지는 문제점을 분석하고, SDN 기술을 적용하여 효과적으로 네트워크 트래픽을 절감하고, 기존 시스템이 가지고 있던 문제를 해결하는 새로운 HPC 네트워크 구조를 제안하였다. 기존 시스템이 RoCE를 사용함에 있어 브로드캐스트를 지원하지 못하여 발생하는 네트워크 내 중복된 트래픽 발생 문제나, PFC 데드락, 정지 프레임 스톰 문제 등 개별적인 기술이 가지고 있는 한계점을 SDN 기술을 접목함으로써 트래픽의 양을 지수함수에서 선형함수 만큼 줄일 수 있었고, 네트워크 상황을 모니터링하고 유연하게 제어할 수 있는 SDN의 특성을 이용하여 기존 PFC기술이 가지고 있는 기능적 한계점도 극복이 가능하였다. 본 연구진은 본 논문에 제시한 구조를 바탕으로 실제 테스트베드를 구축하여, 다양한 응용에 적합한 SDN기반 HPC 계산 네트워크 구조를 만들 예정이다. 이러한 HPC 네트워크 연구는 추후 초고성능 컴퓨팅 연구의 중요한 기반 연구가 될 것으로 예상된다.

ACKNOWLEDGMENT

본 연구는 한국연구재단에서 지원하는 2017년도 차세대정보컴퓨팅기술 개발사업 (PF급 이종 초고성능 컴퓨터 개발, NRF-2016M3C4A7952587) 연구수행으로 인한 결과물임을 밝힙니다.

참 고 문 헌

[1] C. Guo, H. Wu, Z. Deng, G. Soni, J. Ye, J. Padhye, and M. Lipshteyn. RDMA over Commodity Ethernet at Scale. In Proceedings of the 2016 ACM SIGCOMM Conference. ACM, New York, NY, USA.

[2] R. Mittal, A. Shpiner, A. Panda, E. Zahavi, A. Krishnamurthy, S. Ratnasamy, S. Shenker, Revisiting Network Support for RDMA,

[3] IEEE DCB. 802. 1Qbb - Priority-based Flow Control. (http://www:ieee802:org/1/pages/802:1bb:html.)

[4] S. Hu, Y. Zhu, P. Cheng, C. Guo, K. Tan, J. Padhye, and K. Chen, Tagger: Practical PFC Deadlock Prevention in Data Center Networks, In Proceedings of the 2017 ACM CoNext Conference, ACM, Incheon, Republic of Korea.