데이터 그리드를 위한 네트웍 퍼포먼스 측정

권기환[0,1] 한대희[1] 조기현[1] 오영도[1] 서준석[1] 손동철[1] 이지수[2]
1. 경북대학교 고에너지물리연구소(CHEP) 2.한국과학기술정보연구원 수퍼컴퓨팅센터
{kihwan[0], hanbi, cho, suh, son}@knu.ac.kr, ydoh@hep.knu.ac.kr, jysoo@hpcnet.ne.kr

Performance Measurements on High Speed Network for Data Grid

Kihwan Kwon[0,1] Daehee Han[1] Kihyeon Cho[1] Youngdo Oh[1] JunSuhk Suh[1] Dongchul Son[1] Jysoo Lee[2]
1. Center for High Energy Physics, Kyungpook National University, Daegu 702-701, Korea
2. KISTI, Supercomputing Center, P.O. Box 122, Daejeon 305-806, Korea

요 약

고에너지 물리연구분야에서는 입자가속기에서 배출되는 많은 양의 데이터 (테라 혹은 페타바이트급)를 분산처리하기 위해 데이터 그리드를 구축하려고 시도하고 있다. 이러한 많은 양의 데이터를 다루는 데이터 그리드는 국제 간에 데이터를 신속하게 이동하는 것이 필수적이다. 현재 네트웍 인프라의 발전으로 초고속 대용량 네트웍 대역폭을 가지는 물리적인 경로가 생겨나고 있지만, 실제 퍼포먼스를 어떻게 테스트하고 어떤 요소들을 고려하여야 초고속 네트웍 상에서 좋은 퍼포먼스를 얻을 수 있는지 이해하지 못하는 경우가 많다. 이 논문에서는 한국과 외국의 초고속 연구용 네트웍을 중심으로 네트웍 퍼포먼스를 측정함으로써 초고속 네트웍 상에서의 네트웍 퍼포먼스 테스트의 이해를 돕고자 한다.

1. 서 론

고에너지물리연구는 가속기를 이용하여 우주 생성 및 진화에 관한 학문을 연구하는 기초과학분야이다. 고에너지물리실험은 검출기 설계, 제작, 신호처리 및 자료 수집, 분석에 이르기까지의 일련의 작업을 가속기가 있는 외국에서 국제 공동연구로 수행한다. 고에너지 물리연구에서는 입자가속기에서 배출되는 막대한 양의 데이터를 저장, 처리하기 위한 계산능력과 대용량의 저장장치를 확보하는데 어려움을 겪어왔다. 또한, 실제 실험에 참여, 연구, 분석하는 인력 또한 전세계에 흩어져 있어 실험의 효율성을 떨어뜨리고 있다. 이러한 어려움을 해결하고자 차세대 인터넷이라 불리는 그리드에 기반을 둔 데이터 그리드를 활용하고자 노력하고 있다.

이러한 데이터 그리드를 구성하기 위해서는 초고속 네트웍 기반이 필요하다. 초고속 네트웍 상에서만 무리없이 많은 양의 데이터를 운반할 수 있기 때문이다. 한국에도 연구망, 선도망을 중심으로 기가급 대의 네트웍이 구성되어 있고, 일본 미국 등과도 대용량 대역폭으로 연결되어 있다. 이러한 대용량 대역폭을 가지고 긴 RTT(Round Trip Time)를 가지는 LFN(Long Fat Network)상에서 어떤 요소들이 퍼포먼스를 높이기 위해서 중요하며, 실제로 어떤 결과를 얻을 수 있는지 실제 테스트를 통하여 보여줌으로써 초고속 네트웍 상에서의 퍼포먼스 테스트의 이해를 돕고, 주변의 네트웍의 성능을 바르게 측정함으로써 네트웍 활용에 도움을 주고자 한다. 본론의 2.1에서는 네트웍 퍼포먼스와 관련된 기본 개념을 살펴보고 2.2에서는 경북대학교와 한국 선도망(KOREN, Korea advanced Research Network) NOC(Network Operation Center)와의 네트웍 테스트를 TCP 윈도우 사이즈를 바꾸면서 테스트한 결과를 소개하였고, 2.3에서는 경북대와 미국 칼텍(California Institute of Technology)간의 네트웍 테스트를 통한 결과를 분석한다.

2. 본 론

2.1 관련 용어

네트웍 성능을 얘기할 때 다음과 같은 기본적인 개념을 이해해야 할 것이다. Path Capacity 는 양 끝단의 호스트를 연결하는 네트웍 경로의 물리적인 능력을 의미한다. 어떤 네트웍 경로를 연결하는 링크는 여러개가 있을 수 있는데, 바틀넥(bottleneck)링크가 이 Path Capacity를 결정한다. Network Utilization은 어떤 네트웍 경로의 모든 네트웍 트래픽의 총 합을 의미한다. 그리고 Path Capacity에서 Utilization을 제외한 대역폭이 Available Bandwidth인 것이다. Achievable Bandwidth란 모든 주어진 상황을 고려하였을 때 실제 어떤 어플리케이션이 얻을 수 있는 퍼포먼스이다. 여기서 주어진 상황이란, 즉 전송 프로토콜(TCP 혹은 UDP), 하드웨어 (프로세서 속도, 버스 속도, 네트웍 카드), 운영체제, TCP 버퍼 사이즈 등을 의미한다. 이 논문에서 관심있게 테스트한 것은 Path Capacity와 Achievable Bandwidth이다.

최근 그리드의 열기에 대한 고조로 초고속 네트웍에서의 퍼포먼스에 대한 많은 연구가 진행되고 있다. TCP RFC 1323[1] 에서 언급되어 있듯이, TCP의 퍼포먼스는 전송속도 자체보다도, 전송속도와 왕복 딜레이(RTT)의 곱에 달려있다고 한다. 이 BDP (Bandwidth Delay

Product)는 양 끝단의 호스트들 간에 가상적으로 존재하는 파이프를 가득 채울 수 있는 데이터의 양을 의미한다. 그러나 이 BDP가 굉장히 클 때(Mega Byte 이상일 때), 즉, LFN에서 퍼포먼스에 문제가 생길 수 있다. 왜냐하면, TCP 프로토콜은 기본적으로 헤더에 16 비트를 씀으로서, TCP 윈도우 사이즈는 65킬로 바이트로 한정되기 때문이다. 이것을 해결하기 위해서 TCP RFC 1323은 여러 가지 옵션을 규정하고 있다.
그리고 패킷로스가 일어날 때 마다 그 파이프가 비워지며, 다시 슬로우 스타트를 시행한다. 이러한 한계를 극복하기 위해 여러 가지 옵션을 두고 있다.
가. tcp_window_scaling: 65킬로 바이트보다 더 큰 윈도우 사이즈를 지원하기 위해서 쓰인다.
나. tcp_sack: 데이터를 받는 호스트가 비연속적인 데이터를 acknowledge 할 수 있도록 한다. BDP가 클 적에 특히 도움이 된다.
다. ip_no_pmtu_disc: 가능하면 가장 큰 MTU (Max Transfer Unit)을 써서 패킷을 보내야 좋은 성능을 얻을 수 있다. 하지만 너무 큰 MTU로 패킷을 보내면 중간에서 이것을 감당할 수 없는 라우터는 이것을 나누기 때문에 오버헤드가 생긴다. 그러므로 어떤 네트웍 경로에서 이 분해(fragmentation)가 일어나지 않고 보낼 수 있는 가장 큰 MTU를 찾게 해주는 옵션이다.
라. rmem_max: 최대 receive 윈도우 사이즈
마. wmem_max: 최대 send 윈도우 사이즈
바. tcp_rmem: TCP receive 버퍼에 할당된 메모리이다.
사. tcp_wmem: TCP send 버퍼에 할당된 메모리

이러한 옵션을 가동시키고, 조정하므로써, 더 나은 퍼포먼스를 얻을 수 있다. 현재 경북대 네트웍 테스트 머신에 셋업된 것은 다음과 같다.

tcp_rmem: 4096 87380 128388607
tcp_wmem: 4096 87380 128388607
rmem_max/wmem_max: 128388607

2.2 경북대와 대전 KOREN과의 테스트

경북대와 대전에 있는 초고속 선도망 NOC는 기가빗으로 연결되어 있다. 그림 1은 Iperf[2]를 이용하여 하나의 TCP 스트림을 사용하여 약 10초간 데이터를 전송하였을 때의 결과이다. traceroute 툴을 이용하여 경북대와 NOC간에 RTT를 측정하면, 약 2ms 이다. 그러므로 BDP (Bandwidth Delay Product)는 다음과 같이 계산될 수 있다. 2ms * 1Gbps = 256KB/s 이다. 이 테스트에서는 윈도우 사이즈를 256KB에서 시작하여 4MB까지 변화시키면서 여러번 테스트를 실행하였다. 결과 그래프에 나타나듯이 윈도우 사이즈가 실제 사용 가능한 대역폭에 미치는 영향을 잘 보여주고 있다. 하지만 실제 테스트에서 보여주듯이 적어도 TCP 윈도우 사이즈가 1MB는 되어야 기가빗 라인을 충분히 사용할 수 있다는 것을 보여준다. BDP는 이론적으로 최적의 TCP 윈도우 사이즈를 찾아 준다고 할 수 있지만, 실제로는 그것보다 크거나 작은 경우를 시도해 보면서 경험적으로 최적의 TCP 윈도우 사이즈를 찾아야 할 것이다.

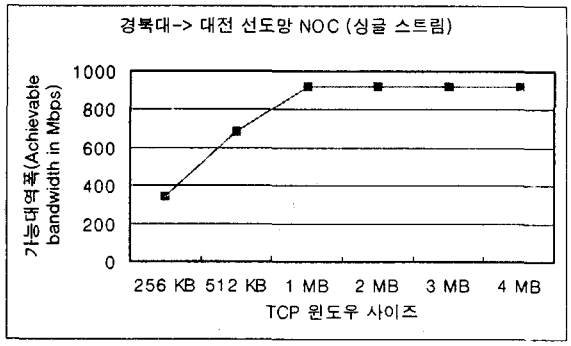

그림 1. 경북대와 KOREN NOC와의 퍼포먼스 테스트

.이 테스트에 퍼포먼스 측정툴로 Iperf1.7.0을 사용하였다. 메모리에서 바로 비트 스트림을 생성하여 보여줌으로서 하드 디스크의 속도의 제약을 받지 않고 네트웍을 테스트 할 수 있다. 네트웍에 결함이 없고 TCP 윈도우 사이즈가 충분히 클 때, 패킷 손실 없이 믿을 수 있는 네트웍에서는 하나의 스트림을 사용하던지 여러 개의 병렬 스트림을 사용하건 비슷한 퍼포먼스를 낼 것이다. 5개의 TCP 스트림을 사용하여 동시에 데이터를 전송할 경우 선도망에서 각 스트림이 대역폭을 나누어 가지고 이 스트림의 총 대역폭의 합은 Single Stream의 대역폭과 비슷하다(그림 2).

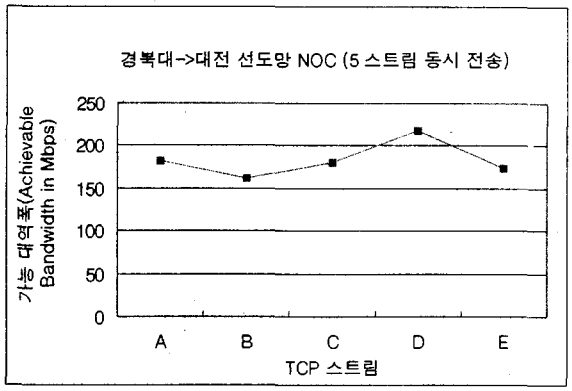

그림 2. 다섯 개의 TCP 스트림 A, B, C, D, E가 동시에 데이터 전송

이 다섯 스트림의 사용 대역폭의 합은 906Mbps로서 하나의 스트림이 사용한 920Mbps대와 비슷함을 알 수 있다.

2.3 경북대와 미국 칼텍(Caltech)과의 테스트

경북대에서 미국 칼텍으로의 경로는 그림 3과 같다.

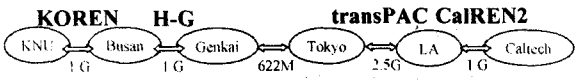

그림 3. 경북대에서 칼텍으로의 네트웍 경로

한국의 KOREN 선도망(2.5Gbps 백본, 경북대는 1Gbps로 연동 되어 있음)을 타고 부산으로 간 다음 현해/겐카이(1Gbps) 선을 타고 일본으로 간다. 겐카이 XP와 동경 XP 사이에 622Mbps로 연결되어 있다. 그 다음 transPAC (2.5Gbps)을 타고 미국 LA의 exchange point를 거쳐 CalREN2(1Gbps)을 타고 칼텍으로 간다. 일본에서 현해/겐까이와 transPAC을 연결하는 부분이 622Mbps인 OC-12라인을 사용하고 있어 바틀넥(bottleneck)으로 작용하고 있다. 실제 Path Capacity 측정 툴인 pathrate[3]을 이용하여 측정하여 보았을 때, 450~457Mbps를 얻을 수 있었다.

또한, UDP 스트림을 흘려 보냈을 때, 400Mbps에서는 거의 0%에 가까운 datagram loss를 보였지만, 450Mbps에서는 30%의 손실을 보였다. 그리고 622Mbps의 UDP 스트림을 보냈을 땐, 거의 80%에 가까운 손실을 보였다. UDP는 TCP 측정과는 달리, packet loss와 jitter에 대해 측정할 수 있게 해준다. 그리고 UDP 테스트에서는 대역폭을 지정한 만큼 Iperf에서 생성하여 전송한다. UDP 테스트 결과로 부터 Path Capacity는 약 400Mbps 정도인 것을 알 수 있다.

BDP를 계산해 보았을 때, 135.5ms * 457Mbps= 7.74MByte 정도이다. 굉장히 큰 TCP 윈도우 사이즈를 요구하며, 이것은 LFN(Long Fat Network)의 특징이다. 그러므로 적어도 약 8MByte의 윈도우 사이즈를 이론적으로 요구하는 것이다. 디폴트 TCP 옵션으로는 가능하지 않으며, RFC 1323에서 제시한 window_scaling 옵션을 비롯한 다른 옵션에서 적당한 사이즈를 지정해 주어야 한다. 그림 4는 병렬 전송 TCP 스트림의 수를 10개씩 증가시키면서 퍼포먼스를 측정한 것이다. 처음에는 선형적으로 어느 정도 대역폭이 증가하다가 50개 스트림 이후로는 거의 변화가 없다. 너무 많은 TCP 스트림은 서로 자원을 빼앗기 위해 경쟁하기 때문에 퍼포먼스는 증가하지 않는다[4]. 이것은 그림 5에서 더욱 잘 나타난다. 그러므로 50개 정도의 병렬 전송 TCP 스트림이 바람직하며 거의 사용가능한 대역폭을 다 소비하고 있다고 볼 수 있다.

하지만 이 테스트에서 하나의 TCP 스트림으로는 10Mbps 이하의 대역폭 밖에 얻을 수 없었다. 아마도 QoS나 혹은 어떤 종류의 traffic shaping 걸려 있는 것 같다.

3. 결 론

데이터 그리드에서는 국제 간에 많은 양의 데이터 전송을 요구하며 기본적으로 초고속 네트웍을 필요로 한다. 이 논문에서는 경북대와 선도망 NOC, 경북대와 칼텍 간의 네트웍 퍼포먼스를 측정함으로써 초고속 네트웍 상에서 좋은 퍼포먼스를 얻기 위하여 고려해야할 요소들을 점검하였다. 유휴한 대역폭을 제대로 사용하기 위

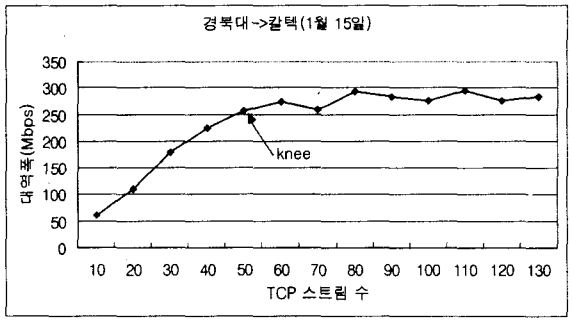

그림 4 . 경북대에서 칼텍으로의 네트웍 퍼포먼스 측정

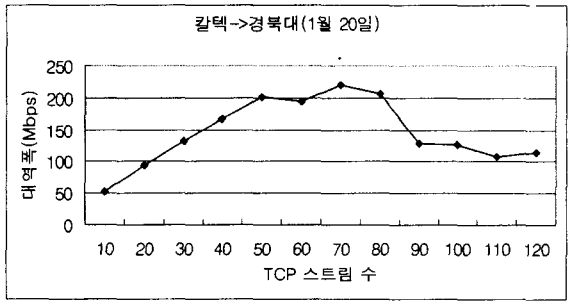

그림 5. 칼텍에서 경북대로의 네트웍 퍼포먼스 측정

해서는 TCP 윈도우 사이즈의 조정이 필요하며 다수의 TCP 스트림을 이용한 병렬 전송이 중요하다. 경북대와 대전 사이의 테스트를 통해서 BDP가 이론적으로는 최적의 TCP 윈도우 사이즈를 계산할 수 있게 해 주지만, 실제에 있어서는 그 보다 높거나 낮은 윈도우 사이즈를 시도해 봄으로써 최적의 윈도우 사이즈를 경험적으로 찾는다는 것이 바람직하다는 것을 보여준다. 경북대와 칼텍과의 테스트를 통해서 여러 개의 병렬 전송시 TCP 스트림의 수에는 최적의 수가 있으며, 이 보다 더 많은 스트림을 사용할 경우에 다른 트래픽에 영향을 주며, 서로 자원을 차지하려고 경쟁하기 때문에 가용대역폭을 증가시켜주지는 않는다는 사실을 확인시켜 주었다. 앞으로 경북대는 칼텍에서 개발한 TCP인 FAST[5]를 이용하여 현재의 대부분의 OS에 설치되어 있는 TCP Reno를 대치하여 테스트할 계획이다. 이런 테스트 결과를 활용하여 유럽의 입자가속기에서 생성되는 테라바이트~페타바이트급 자료를 네트웍으로 전송 받을 예정이다.

4. 참고문헌

[1]http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc1323.html
[2]http://dast.nlanr.net/Projects/Iperf/
[3]http://www.cc.gatech.edu/fac/Constantinos.Dovrolis/pathrate.html
[4]Netest: A Tool to Measure Maximum Burst Size, Available Bandwidth and Achievable Throughput", G. Jin, B. Tierney, Proceedings of the 2003 International Conference on Information Technology Research and Education, 2003
[5]http://netlab.caltech.edu/FAST/