최적성능노드 경유 고속전송 방안 연구

석우진[1],
[1]한국과학기술정보연구원 과학기술디지털융합본부
wjseok@kisti.re.kr

Faster Data Transfer using Optimized Intermediate Node

Woojin seok[1]
[1]Div. of Science Digital Convergence, KISTI

요 약

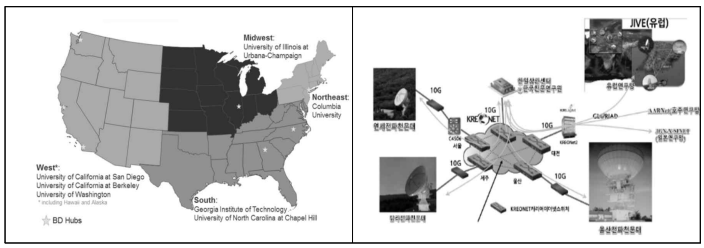
본 논문에서는 과학 빅데이터를 위한 고속 데이터 전송 방식을 제안한다. 최근의 과학연구는 이전보다 훨씬 더 많은 양의 데이터를 요구하지만, 잘 알려진 네트워킹 문제인 라스트마일 문제로 인해 여전히 데이터를 수신하는 데 시간이 오래 걸린다. 과학 빅데이터 전송시 라스트마일 문제로 인한 패킷 손실에 대해 더 나은 방법을 제안한다. 제안하는 방법은 원격 전송에 최적화된 중간 서버를 사용하고 종단간 네트워크 경로에서 라스트마일을 분리한다. 전송 측정을 통해 향상된 성능을 확인한다.

1. 서론

과학분야는 실험장비에서 생성된 데이터를 중심으로 데이터 분석 방식으로 발전하고 있다. 다양한 과학분야에서 실험장비 데이터를 위한 관측 장비, 분석 장비들의 기술발전이 비약적으로 이루어지고 있으며, 이렇게 생성된 과학데이터는 별도의 데이터센터를 구축하여 관리되고 있다. 미국은 NSF(국가연구재단)의 재원으로 미 전역에 분야별 지역별 빅데이터혁신허브를 구축하여 (그림1 왼쪽) 과학분석을 촉진하고 있으며, 이를 위하여 고속전송인프라, 대용량 저장데이터센터 기술을 적용하고 있다. 또한, 저장된 데이터는 분석을 위하여 고성능 컴퓨팅 자원들과 연계하여 데이터의 저장, 전송, 처리가 SW 적으로 연계될 수 있도록 진화하고 있다. 예를 들어, 천문분야에서는 분산된 전파망원경을 통하여 생성된 과학빅데이터 전송을 위하여 전용의 고속 네트워크를 구축하여(그림1 오른쪽) 대용량의 관측데이터를 분석 장비에 전송하고 있다.

이러한 대용량의 데이터를 전송하기 위해서 데이터 전송은 중요한 기술적 요소이지만, 여전히 기존의 전송 프로토콜인 TCP 프로토콜 기반의 FTP(File Transfer Protocol) 프로토콜이 사용되고 있다. 또한 성능개선을 위해서 bbcp, bbFTP, GridFTP, BDSS와 같은 병렬 전송 방법이 제안되어 더 나은 성능을 보여주고 있다 [1][2].

본 논문에서는 End-to-End 경로를 에지 네트워크와 백본 네트워크로 분할하고 경계지역에 전송성능에 최적화된 서버를 배치하는 전송 방법을 제안한다. 기존의 TCP 전송방식과 추가되어 성능을 더욱 개선할 수 있는 방법으로써, 라스트마일 문제로 인해 엣지 네트워크에서 많이 발생하는 패킷 손실로 인한 성능 저하가 전체 성능에 영향을 미치지 않는 것을 추구한다. 제안된 방법은 병렬 FTP와 함께 작동할 수 있으므로 데이터 전송 성능이 훨씬 향상된다.



[그림1] 데이터 기반의 과학분석 방식의 진화 사례: (좌) 과학빅데이터 지역적 관리 (우)전파망원경의 과학빅데이터전송 사례

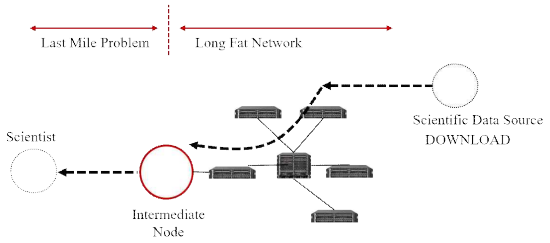
2. 최적화된 중간노드 활용 고속전송방안

데이터 전송 성능은 TCP 성능에 따라 달라진다. TCP의 본질적인 특성상 패킷 손실로 인해 후속 패킷 재전송 및 TCP 창 복구가 발생하여 성능이 저하된다. 패킷 손실은 라스트 마일 문제로 인해 백본 네트워크보다 에지 네트워크에서 더 자주 발생한다.

처리량은 아래 [3]과 같이 패킷 손실률의 제곱에 반비례한다.

$$Throughput \leq \frac{MSS}{RTT\sqrt{P_{loss}}} \quad \text{(수식 1)}$$

RTT가 길수록 재전송 및 창 복구에 더 많은 시간이 소요된다. 즉, 패킷 손실로 인해 장거리 데이터 전송의 경우 성능이 많이 저하된다. 원격 과학 응용의 경우 대부분의 과학자가 전 세계에 위치한 과학 데이터가 필요하기 때문에 전송 거리가 훨씬 더 길다. 긴 경로는 사무실이나 캠퍼스 네트워크의 패킷 손실에 대해 훨씬 높은 성능 저하를 초래할 수 있다 [4]. 본 논문에서는 에지 네트워크에서의 패킷 손실로 인한 피해를 전체 성능과 분리하기 위해 데이터 전송 세션 분할을 제안한다.



[그림2] 전송경로 분할하여 전송성능향상 방안

제안하는 구조는 에지에서의 패킷 손실로 인한 성능 저하에 영향을 미치지 않도록 에지 네트워크를 백본 네트워크와 분리하여, 데이터 저장 및 전송에 최적화된 중간 노드에 데이터를 저장한다. 중간 노드는 데이터를 저장하고, 데이터 저장이 끝나면 즉시 다음 노드로 전달한다. 이러한 목적으로 중간 노드에는 데이터 전송에 최적화된 소켓 버퍼을 구성한다.

중간 노드는 데이터 전송을 위한 전송에 최적화된 서버이며, 하드웨어로써, 고성능 저장 및 전송을 위한 SSD(Solid State Drive)와 10G 네트워크 인터페이스 카드로 구성한다. 전송을 최적화하려면 BDP(Bandwidth Delay Product) 계산에서 소켓 버퍼를 적절하게 설정해야 한다.

256Mbyte = 10Gbps * 0.2초 * 1/8Byte (수식 2)

대부분의 과학 응용 분야에는 국제 거리 데이터 전송이 필요하고 국제 거리에 대한 대기 시간은 약 100ms이므로 RTT의 경우 약 200ms로 설정된다. 결과적으로 BDP를 기준으로 소켓 버퍼 크기는 256Mbyte가 된다. 최적화된 중간노드는 큰 BDP에 대해 뛰어난 성능을 보여주기 때문에 해밀턴 TCP를 혼잡 제어로 사용한다. 중간 노드는 성능 저하 설정을 제거하기 위해 “path MTU”를 설정하지 않으며, 또한 방화벽은 전송 성능을 크게 저하시키기 때문에 방화벽을 우회한다. 보안 목적으로 시스템 내부에서 자체적인 방화벽을 설정하여 사용한다.

대부분의 과학 데이터는 Luster 파일 시스템을 주로 많이 사용하므로, 본 실험에서도 Luster 파일 시스템을 기반으로 파일 시스템을 구성하였고, 중간경유 서버는 이러한 파일시스템에 통합하여 사용한다. 데이터를 더 빠르게 처리하기 위해서 중간 노드와 파일 시스템의 연계를 최적화 하여야 한다. 즉, 네트워크 버퍼 크기 및 동기/비동기 쓰기 옵션과 같은 일부 마운트 옵션을 사용하며, Linux 커널을 기반으로 읽기/쓰기 버퍼 크기 옵션(rsize, wsize)을 조정하면 파일 시스템과 중간 노드 간의 전송 성능을 향상시킬 수 있다. 또한 중간 노드의 wring 정책은 신뢰성보다는 처리량에 초점을 맞추기 위해 비동기식으로 설정한다.

과학적 빅데이터 전송을 가속화하기 위해 데이터 전송 경로를 엣지 네트워크와 백본 네트워크로 분리한다. 검증은 3개의 노드로 구성된 토폴로지에서 이루어졌으며 중간 노드는 데이터를 저장하고 전달하는 단계가 전혀 다르다. node0과 node1 사이의 경로는 패킷 손실률이 0.05%, 대기 시간은 10ms, 대역폭은 10Mbps로 설정된 에지 네트워크를 설정하고 실험한다. node1과 node2 사이의 경로는 패킷 손실률이 0.005%, 대기 시간은 100ms, 대역폭은 1Gbps로 설정된 백본 네트워크를 설정하고 실험한다. 노드의 시스템은 모두 Dell R710이고 Cisco 4507로 연결되어 있으며, 지연 시간과 패킷 손실은 노드 사이에 숨겨진 추가 노드에서 실행되는 더미넷 소프트웨어에 의해 실험한다.

결과를 보면 레거시 FTP는 node0에서 node2로 1.35Bbyte의 데이터를 전송하고 3.6Mbps의 처리량을 측정하였다. 그러나 제안된 방식은 처리량이 5.5Mbps이다. 데이터 전송 속도를 높이기 위해 중간 노드를 사용하여 53%의 성능 향상을 달성하였다. 윈도우 복구 및 패킷 재전송 동안 TCP ACK 세그먼트에 대한 누적 대기 시간은 RTT의 제곱에 크게 비

레한다 [5]. 따라서 제안된 방법보다 레거시 FTP에서 성능 저하가 더 많이 발생하는 것으로 파악된다.

[표1] 성능 측정

Table Head	Throughput		
	Node0-Node2	Node0－Node1	Node1-Node2
Legacy FTP	2978.05 second	NA	NA
Proposed Transfer	1959.96 second	1144.37 second	815.59 second

3. 결론

제안된 방법은 과학 빅데이터를 저장하고 전달하기 위해 성능이 최적화된 스토리지 서버를 사용했다. 레거시 대비 53% 향상된 성능을 보여준다. 성능 측정을 위해 최적화된 스토리지 서버를 사용하여 실제 파일의 성능 측정에 대한 추가 분석이 필요하다. 서버는 빅데이터 다운로드 측면에서 길고 대역이 큰 네트워크에 맞게 조정하여 사용이 가능하다. 과학 분야의 CERN LHC 데이터와 같은 글로벌 데이터 소스와 세계적으로 유명한 슈퍼컴퓨터 센터나 국립 연구소의 과학 빅데이터로부터 데이터를 받는 것이며 국내 과학자들은 대개 국내 사이트가 아닌 글로벌 데이터 소스에서 데이터를 가져오는 경향이 있다. 제안된 방법은 한국 과학자들의 글로벌 과학 활동에 훨씬 더 나은 성능을 제공할 것이다.

Acknowledge

이 논문은 2024년도 한국과학기술정보연구원(KISTI)의 기본사업으로 수행된 연구입니다. (양자암호통신 기반 공동활용 네트워크 기반 구축, K-24-L04-M01-C02-S01)

참고문헌

[1] http://fasterdata.es.net/host-tunning/

[2] N.A. Watts and F.A. Feltus, “Big Data Smart Socket(BDSS): a system than abstracts data transfer habits from end users”, International Journal of Bioinfomatics, vol. 33, no. 4, pp.627-628, 2016.

[3] Jitendra Padhe, Victor Firoju, Don Towsley, and Jim Kurose, “Modeling TCP throughtpu: a simeple model and its empirical validation”, Proceedings of the ACM SIGCOMM, Vancouver Canada, August 31-September 3, 1998.

[4] Eli Dart, Lauren Rotman, Brian Tierry, Mary Hester, and Jason Zurawsk, “Science DMZ: A Network Design Pattern for Data-Intersive Science”, SC'13 Proceedings of the International Conerence on High Performance Comuting, Networking, Storage and Analysis, Denver USA., November 17-22, 2013.

[5] W. Seok, M. Lee, J. Kong, and J. Kwak, “Accelerating Data Transfer over Tier system for LHC experiment”, Journal of the Korean Physical Society, vol. 55, no. 2, pp. 2253-2257.