대용량 스토리지 기반의 데이터 전송 노드 클러스터 설계 및 구축

홍원택*, 안도식**, 이재국**
*한국과학기술정보연구원 과학기술연구망센터
**한국과학기술정보연구원 슈퍼컴퓨팅인프라센터
{wthong, dsan, jklee}@kisti.re.kr

Designing and building a DTN cluster based on massively scalable storage

Wontaek Hong*, Dosik An**, Jaekook Lee**
*Advanced Kreonet Center, KISTI
**Supercomputing Infrastructure Center, KISTI

요 약

과학응용분야의 원활한 협업 지원을 위해서는 원거리간 대용량 연구데이터의 고속 전송이 반드시 요구된다. 이와 관련하여, 본 논문은 기 구축된 대용량 파일 시스템을 다수의 데이터 전송 노드(DTN)에 연동하기 위해 필요한 요구사항들을 정리하고, 이에 기반하여 DTN 클러스터를 설계하고 구축한 사례를 제시한다. 추가적으로, 종단간 왕복지연 시간이 약 130ms에 달하는 원거리 종단 포인트와 대용량 실험데이터를 송수신함으로써 구축된 결과물의 전송 성능을 측정하고 확인한다.

1. 서론

대용량 데이터의 고속 전송이 요구되는 과학응용분야를 위한 특화된 네트워크를 제공하기 위해 제안된 미국 에너지과학연구망(ESnet)의 Science DMZ 개념은 최근 천문학, 기상기후 분야 등의 대용량 데이터 전송이 필요한 국내외 연구망 커뮤니티에서 활발히 적용되어 오고 있다. [1][2] 이러한 Science DMZ 개념을 실현화하기 위해서는 일반 망과는 차별화된 고품질의 고성능 네트워크 경로 확보, 전용의 데이터 전송 노드(DTN) 및 전송 상태를 모니터링할 수 있는 시스템 등이 요구된다. 특히, 핵심 요소인 DTN을 적용하기 위해서는 RAID 컨트롤러를 기반의 내부 스토리지 연계 또는 확장성이 우수한 외부 스토리지 연계 모델의 구현이 필요하다. [3] 이와 관련하여 본 논문에서는 고성능 컴퓨팅(HPC) 환경과 같이 대용량 데이터 전송이 요구되는 분야에서 Science DMZ 구축 모델에 기반하여 기 구축된 병렬 파일시스템을 연계할 수 있는 DTN 클러스터를 설계하고 구축한다.

2. 본론

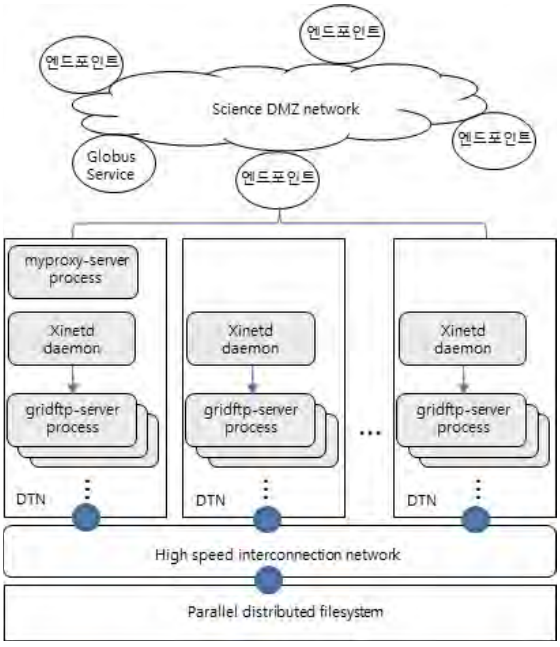
대용량 스토리지들을 Science DMZ 환경에 적용하기 위해 고속 마운트 기반의 외부 스토리지 연계 모델에 기초하여 DTN 클러스터를 구축한다. 이러한 접근 방법은 기 구축된 병렬 파일시스템을 다수의 DTN 서버에 마운트하여 확장성을 높일 수 있다. HPC 환경에서 계산 노드들과 스토리지 시스템들을 연결해 주는 인터커넥션 네트워크 기술이 DTN과 외부 파일시스템을 연동하기 위해 동일하게 적용될 수 있다. 특히, DTN 기반의 고속 마운트를 위해 병렬 파일시스템의 서버/클라이언트 모듈의 튜닝이 필요하고, 이러한 과정은 외부 스토리지에 적용된 파일시스템 프로토콜에 의존하여 서버/클라이언트 간의 성능 최적화 과정이 요구된다.

DTN을 하드웨어적으로 구성하는 측면에서 개별 DTN 시스템은 제한된 기능만을 수용할 수 있도록 최대한 단순하게 구성하는 것이 중요하다. 특히, 고속 마운트 기반의 파일시스템 연계를 위해서는 외부망 트래픽을 위한 단일 네트워크 인터페이스와 파일시스템으로 향하는 단일 네트워크 인터페이스로 분리하여 DTN을 구성해야 하고, 이러한 인터페이스들은 Ethernet, InfiniBand, Intel Omni-Path와 같은 인터커넥션 네트워크에 연결되어야 한다. 또한, DTN 하드웨어 성능 최적화와 더불어 원거리 전송에 적합한 TCP 알고리즘의 선택, 소켓버퍼 크기, 점보 프레임 지원을 위한 MTU 크기 세팅 등을 포함한 DTN 운영체제 및 전송 프로토콜에 대한 소프트

웨어 최적화가 선행되어야 한다.

병렬 파일시스템 내의 다수의 대용량 파일들을 Science DMZ 환경에서 고속으로 전송하기 위해서는 DTN 서버들 또한 확장 가능한 병행성(Concurrency)와 병렬성(Parallelism)을 지원해야 한다. 병행성은 전송하고자 하는 DTN 서버의 수 및 CPU 코어 수의 증가를 의미하고, 이러한 증가된 자원들은 각각의 전송 프로토콜 프로세스에 매핑되어 전송 성능을 향상시킬 수 있다. 또한, 하나의 파일을 전송하는데 있어서 다수의 TCP 스트림들로 나누어 전송할 수 있는 병렬성을 지원함으로써 전송 효율을 향상시킬 수 있다. 이러한 병행성과 병렬성은 전송하고자 하는 파일의 크기, 파일의 수 등의 전송 환경에 따라 최적 값이 변할 수 있음을 추가적으로 고려해야 한다.

DTN 클러스터가 연결된 외부 전송망은 네트워크 패스 프로비저닝 등을 통해 병목이 없는 전용 경로를 확보한다. 또한, DTN 서버에 설정된 점보프레임 지원을 위한 MTU 설정과 더불어 상대방 DTN 서버들 또한 동일한 MTU 크기의 설정이 필요하다. 이러한 MTU 크기의 세팅은 DTN 네트워크 인터페이스 뿐만 아니라, 종단 목적지까지의 모든 경로 상에 있는 네트워크 장비 상의 In/Outbound에 대해 동일하게 설정해야 한다. 추가적으로 더 엄격한 망 분리를 고려하는 경우에는 데이터 전송 프로토콜의 제어 채널을 위한 제어 망과 데이터 채널을 위한 전송망을 분리한다.



(그림 1) DTN 클러스터 구축

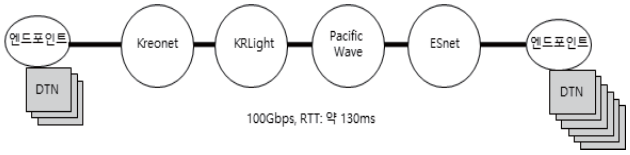
그림 1은 위에서 언급된 요구사항들을 반영하여 병렬 파일시스템 내의 파일들을 고속 전송하기 위해 Globus Connect Server(GCS) 소프트웨어 [4]를 활용하여 다수의 DTN 서버들을 클러스터링한 구축 환경을 보여준다. GCS에서는 DTN과 같은 다수의 I/O 노드들을 하나의 종단 포인트로 표현할 수 있는 메커니즘을 제공하고, 이러한 메커니즘에 기반하여 원거리 전송에 적합하게 하드웨어, 소프트웨어들이 성능 튜닝된 다수의 DTN 서버들을 대상으로 클러스터링화 한다.

다수의 파일들을 동시에 전송하기 위해서는 Xinetd 데몬에서 fork된 gridftp-server 프로세스들이 각각의 파일들을 전송하기 위해 생성되어 전송에 참여한다. 이렇게 생성된 각각의 gridftp-server 프로세스들은 DTN에서 제공하는 다수의 CPU 코어들로 매핑되어 병행성을 증가시킨다. 추가적으로 각각의 gridftp-server 프로세스들은 하나의 파일을 전송하는데 있어서, 다수의 TCP 스트림들을 생성하여 병렬성을 높이게 된다. 예를 들어, L개의 DTN 서버에서 M개의 CPU 코어를 갖고, 각 전송 프로세스당 N개의 TCP 스트림들을 생성한다면 산술적으로 최대 LxMxN개의 TCP 스트림을 생성할 수 있게 된다. 이렇게 병렬 파일시스템을 고속 마운트하는 DTN 클러스터를 바탕으로 설정된 단일의 종단 포인트는 Globus 전송 기반의 협업에 참여하고자 하는 다른 종단 포인트들과 상호 검색 및 고속 전송이 가능하다.

그림 1에서와 같이 슈퍼컴퓨터 5호기 누리온의 약 20PB 용량의 Lustre 파일시스템을 연계하기 위해 3 대의 고성능 DTN 서버를 활용하여 DTN 클러스터를 구축하였다. 각 DTN 서버는 2*Intel Xeon Gold 3.5GHz 8 코어 CPU, 12*32GB 메인메모리, 1.92TB SSD 디스크, 100Gbps Ethernet NIC, 100Gbps 인터커넥션 NIC 등으로 구성된다. 추가적으로 각 DTN에는 CentOS v7.5.1804, Globus connect server v4, Lustre client v2.10.7이 설치되어 적용된다.

그림 2는 Lustre 파일시스템을 기반으로 구축된 DTN 클러스터의 전송 성능을 측정하기 위한 실험 환경을 보여준다. DTN 클러스터는 다수의 대용량 파일들에 대해 병행성과 병렬성을 극대화하기 위한 접근 방법이므로, 실험 환경의 구성 시에도 이러한 점을 충분히 고려해야 한다. 즉, 구축된 DTN 클러스터의 종단 포인트와 더불어 전송에 참여하는 상대

편 종단 포인트도 DTN 클러스터로 구성되어야 자원들을 충분히 활용할 수 있다. 이러한 맥락에서 미국 에너지성 슈퍼컴센터(NERSC)의 Cori 파일시스템과 연동되어 있는 DTN 클러스터 종단 포인트와의 전송 테스트를 수행한다.



(그림 2) 전송 실험 환경

NERSC DTN 종단 포인트는 6대의 DTN 서버들로 구성되어 있고, 그림 2에서처럼 두 종단 포인트들 간의 네트워크 경로는 한국의 국가과학기술연구망(KREONET), ESnet 등을 통해 제공되고, RTT가 약 130ms, 종단간 대역폭은 100Gbps으로 제공된다. 이러한 환경에서 종단간 Disk-to-Disk 전송 테스트를 위해 ESnet Read-Only DTN에서 제공하는 약 250GB의 데이터 set들을 대상으로 송수신 실험을 진행한다. 표 1은 실험에 이용되는 데이터 set에 대한 세부 정보를 보여준다.

<표 1> 전송 파일 정보

종류	크기	구성
Climate-Small	246GB	29MB에서 425MB의 크기로 분포된 1,496개의 파일들로 구성
Climate-Large	244GB	10개의 21.5GB의 파일과 1개의 28.8GB 파일로 구성

약 250GB 크기의 Climate-{Small, Large} 데이터 set을 대상으로 송수신 실험을 한 결과, Climate-Small 데이터 set의 경우 송수신시 각각 14.4Gbps, 25.76Gbps의 전송 성능을 기록하였고, Climate-Large 데이터 set의 경우 각각 22.48Gbps, 36.08Gbps의 전송 성능을 기록하였다.

3. 결론

본 논문에서는 병렬 파일시스템의 대용량 데이터를 효율적으로 전송하기 위해 Globus 기반의 DTN 클러스터를 설계하고 구축하였다. 또한, 한미간 원거리 전송 환경에서 실험 데이터를 송수신함으로써 DTN 클러스터링에 따른 전송 성능을 측정하였다. 향후, DTN 클러스터링의 효과를 면밀히 확인하기 위해 전송 규모, 파일의 크기 등 다양한 전송 환경을 고려한 성능 분석에 대한 연구가 필요하다.

ACKNOWLEDGMENT

본 연구는 2020년도 한국과학기술정보연구원(KISTI) 주요사업 과제로 수행한 것입니다.

참고문헌

[1] J. Crichigno, E. Bou-Harb, and N. Ghani, “A Comprehensive Tutorial on Science DMZ”, IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 2041-2078, 2019.

[2] Z. Liu et al., “A Comprehensive Study of Wide Area Data Movement at a Scientific Computing Facility”, IEEE International Conference on Distributed Computing Systems, Vienna, Austria, Jul. 2018.

[3] E. Dart et al., "The Science DMZ: A Network Design Pattern for Data-Intensive Science", SC'13 Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, Denver, USA, Nov. 2013.

[4] Globus Connect Server, https://docs.globus.org/globus-connect-server-installation-guide/.