DTN 클러스터를 활용한 NFS 스토리지 기반 대용량 데이터 전송 성능 분석

홍원택*, 문정훈*, 곽재승*
*한국과학기술정보연구원 과학기술연구망센터
e-mail:wthong@kisti.re.kr

Analysis of large data transfer performance using DTN clustering based on NFS storage

Wontaek Hong*, Jeonghoon Moon*, Jaiseung Kwak*
*KREONET Center, Korea Institute of Science and Technology Information

요 약

최근 첨단 과학응용분야에서는 대용량의 데이터를 효과적으로 공유 및 전송하기 위해 전용의 연구망과 데이터 전송 노드(DTN)를 활용하여 전송 효율을 높이기 위한 연구들이 활발히 수행되고 있다. 특히, DTN 클러스터링은 복수 개의 전송 노드들을 이용하여 병행성을 높여 전송 효율을 극대화시키는 기법으로 고성능컴퓨팅 환경에서 적용되고 있다. 이와 관련하여, 본 논문에서는 초기 실험적 연구로 NFS 외부 스토리지를 연계한 DTN 클러스터링 환경을 구축하고, WAN 구간에서 종단간 대용량 데이터 전송 실험을 수행한다. 실험 결과, 파일의 크기 및 개수 등이 서로 다른 비교 군들을 대상으로 단일 DTN 대비 DTN 클러스터링에 대한 효과를 확인할 수 있었다.

1. 서론

최근 세계적으로 고에너지물리, 기상기후, 바이오/의료, 천문학 분야와 같은 과학 응용 분야에서는 년 간 수십~수백 PetaByte급 이상의 대용량 데이터를 생산하고 있고, 이러한 데이터를 원활히 저장하고 분석하기 위해 유관 기관에서는 대규모의 컴퓨팅, 스토리지 자원들을 구축 및 운영하고 있다 [1]. 또한, 이러한 대용량 데이터는 연구 그룹별로 공유되고, 원활한 협업 연구를 위해 필요한 시점에 협업 파트너들에게 고속으로 전송되어야 하므로 전용의 고성능 연구망의 역할은 매우 중요하다. 실 예로, 미국 DOE(Department of Energy) 산하의 연구 기관인 OLCF(Oak Ridge Leadership Computing Facility)에서의 핵 연구 관련 Petascale 시뮬레이션은 NERSC(National Energy Research Scientific Computing Center)에서 처리된 nuclear interaction 데이터 세트를 필요로 한다. 유사하게, ALCF(Argonne Leadership Computing Facility)의 기상 연구 과학자들은 시뮬레이션 결과를 ORNL(Oak Ridge National Laboratory)의 데이터센터에 있는 기후 관측 데이터 세트와 비교함으로써 실험의 유효성을 검증한다. 즉, 이러한 데이터 센터들 간의 협업은 유관 기관들 간의 대용량 데이터의 이동을 기반으로 이뤄진다 [2].

이러한 대용량 데이터를 원활히 전송하기 위해서 미국 DOE산하 ESnet(Energy Sciences Network)에서는 Science DMZ 개념을 제안하였고 [3][4], 세계의 유수의 연구망센터에서는 이러한 개념을 바탕으로 백본 망을 설계하고, Science DMZ 개념을 적용해 오고 있다. 최근에는 미국의 주요 슈퍼컴퓨팅 센터들 간의 과학 빅데이터 전송 효율을 높이기 위한 PetaScale DTN 프로젝트가 제안되어 NERSC/ESnet, ORNL, ALCF, NCSA(National Center for Supercomputing Applications) 등의 참여 기관들이 1주에 1 Petabyte의 데이터를 전송할 수 있는 환경을 구축하여 해당 기관의 연구자들에게 양질의 연구 환경을 제공해 주고 있다 [5].

이러한 동향과 관련하여, 본 논문에서는 Science DMZ 환경에서 데이터 전송 노드로 활용되는 DTN 서버들을 대상으로 Globus 소프트웨어 [6]에서 제공하는 복수 개의 I/O 노드들에 대한 클러스터링 기법을 적용하여 NFS 기반의 DTN 클러스터링 테스트베드를 구축하고, 전송 성능을 실험한다. 특히, 단일 DTN 서버 대비 복수개의 DTN 서버일 경우 망 자원의 증가에 따른 병행성(Concurrency)이 전송 성능에 미치는 영향을 한-미간 100G 네트워크 구간에서 측정 및 분석하고자 한다.

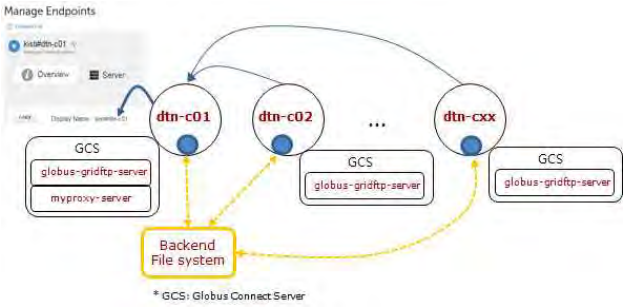
2. DTN 기반의 대용량 데이터 고속 전송

ESnet에서 제안된 Science DMZ 개념을 적용하기 위해서는 고품질의 고성능 네트워크 구성, 전용의 데이터 전송 노드(DTN), 네트워크 경로상의 건강도를 측정하는 모니터링 시스템이 필요하다. 특히, DTN 노드는 HW, SW, 고속 마운트 측면에서 성능 최적화가 선행되어야 한다. 첫째, HW 최적화와 관련해서는 CPU, 저장장치, 네트워크

인터페이스 카드, 메인보드에 대한 하드웨어 규격 등이 성능 최적화를 고려하여 선택되어야 한다. 둘째, SW 최적화와 관련해서는 DTN간의 종단간 대용량 데이터 전송을 위해 TCP 혼잡제어 알고리즘, 소켓버퍼 크기, 점보프레임 세팅 등을 고려한 OS 및 전송 프로토콜에 대한 튜닝이 이루어져야 한다.

이러한 DTN 노드를 활용하여 대용량 데이터를 전송하기 위해서는 DTN 내부의 다수 스토리지들을 RAID 방법으로 묶어서 하나의 볼륨으로 제공하는 방법과 외부 스토리지들을 NFS, Lustre, GPFS와 같은 네트워크 프로토콜들을 이용하여 고속 마운트 시키는 방법 등을 고려할 수 있다. 추가적으로 DTN과 외부 스토리지를 연계해 주는 네트워크 프로토콜에 대해서도 네트워크 프로토콜의 서버/클라이언트 파트 각각에 대해 성능 최적화가 필요하다.

DTN 노드를 활용하는 두 가지 방법 중에서 DTN 기반의 외부 스토리지 고속 마운트 기법은 Globus 소프트웨어를 활용하여 그림 1과 같은 방법으로 다수의 외부 스토리지 노드들을 하나의 엔드포인트로 등록하여 전송에 참여시킬 수 있다. 실제로 이러한 기법은 고속의 대용량 데이터 처리가 요구되는 HPC(High Performance Computing) 환경의 WAN 구간에서 대용량 데이터의 외부 전송 성능을 높이기 위해 활용되고 있다 [5]. 이와 관련하여 본 논문에서는 외부 스토리지를 연계하는 NFS 스토리지 기반의 DTN 클러스터링 환경을 구축하고, 실제 LFN(Long Fat Network) 환경에서 종단간 대용량 데이터 전송 실험을 수행한다.



(그림 1) Globus 기반의 DTN 클러스터링

3. DTN 클러스터링 전송 성능 실험

2절에서 언급된 DTN의 HW, SW 측면에서 성능 최적화된 각각의 DTN들을 대상으로 클러스터링을 구성한다. 본 실험에서는 2대의 DTN 노드들을 클러스터링하고, 같은 네트워크에 있는 고성능 서버에 NFS 서버 프로세스를 구동하여 DTN 노드들에 마운트시키는 환경을 구축한다. NFS 서버/클라이언트 최적화와 관련해서 sync/async 모드 설정, read/write 버퍼 크기 파라미터 등을 조정할 수 있으나 본 테스트베드에서는 NFS 클라이언트/서버에 대해 기본 값을 설정하여 구성한다. 데이터 전송을 위한 WAN 대상 구간은 한-미간 100Gbps 백본망을 대상으로 하고, 미국 쪽의 엔드포인트는 ESnet에서 운영하는 DTN@Starlight을 이용한다. 자세한 실험 환경은 표 1에서 제시되고, DTN 클러스터링에 포함되는 각각의 DTN 서버들에 대해 iperf를 이용한 Memory-to-Memory 전송 테스트는 약 83Gbps의 성능을 달성하였다.

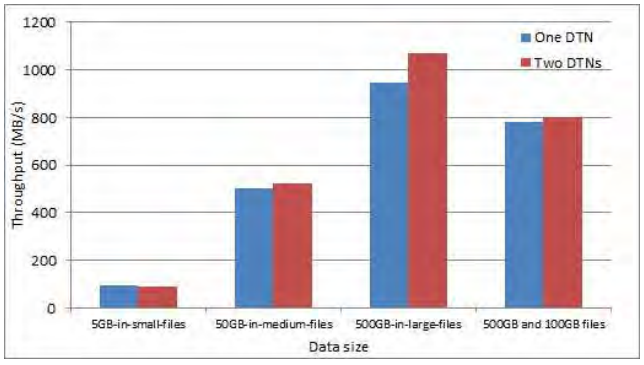
<표 1> 한-미간 전송 실험 환경

구성요소	Description
HW Spec	AMD EPYC 7451 CPU 2.3 GHz 128GB RAM, 2*480GB SSD 100Gps Mellanox ConnectX-5 EN PCIe 3.0 x16 slot
OS	CentOS 7.5
TCP algorithm	HTCP
Socket buffer	Tuned for Large BDP
Backend file system	100G server running NFS
Transfer Tool	Globus (with parallel transfer)
Source Endpoint	ESnet DTN@Starlight
Bandwidth	100Gbps
종단간 Hop count 및 RTT	Three hops, approximately 157 ms
Test file set	Case I: 5GB-in-small-files Case II: 50GB-in-medium-files Case III: 500GB-in-large-files Case IV: 500GB and 100GB files

Globus 소프트웨어에서는 복수 개의 I/O 노드 참여에 따른 Network use 옵션의 변화를 반영할 수 있다. 이러한 파라미터들로는 물리적인 노드의 증가에 따른 변화 파라미터인 병행성(Concurrency)과 하나의 태스크를 여러 개로 나누어 처리하는 병렬성(Parallelism)을 포함한다. 본 실험에서는 DTN 클러스터링을 구성할 때 증가하는 I/O 노드 수 변화에 따른 WAN에서의 전송 성능 변화를 비교하고자 한다. 표 1에서 제시된 각각의 테스트 파일 세트에 대해서 단일 DTN 노드, 두 개로 구성된 DTN 클러스터링마다 종단간 데이터 전송을 수행하여 전송 성능을 측정한다. 참고로, 500GB-in-large-files은 100MB, 200MB, 500MB 파일들이 {a/a, a/b, ..., a/z, b/a, b/b, ..., b/z, ..., z/a, z/b, ..., z/z}와 같은 디렉토리에 각각 저장된다. 유사하게, 50GB-in-medium-files은 10MB, 20MB, 50MB 파일들이 각각의 디렉토리에 저장된다.

그림 2는 위에서 수행한 실험에 대해 각 파일 세트 별로 전송 성능을 비교하여 보여준다. 실험 결과에서 보듯이 상대적으로 작은 크기의 파일 대비 큰 크기의 파일 전송의 경우 DTN 클러스터링의 효과는 더 큰 것으로 나타난다. 실제로, 실험에서 500GB-in-large-files 경우 단일 DTN의 경우 대비 두 개의 DTN으로 구성된 DTN 클러

스터링은 약 13% 정도의 성능 향상을 보인다. 하지만, 나머지 실험의 경우, DTN 클러스터링의 효과는 크지 않았다. 특히, 500GB-in-large-files 경우에 비해 전송되는 파일의 개수가 적고, 전체 파일 크기가 큰 Case IV 경우 노드 수 증가에 따른 병행성(Concurrency)의 이점을 활용하지 못하는 경우로 볼 수 있다.



(그림 2) 전송성능 비교 그래프

4. 결론

본 논문에서는 NFS 기반의 DTN 클러스터링 테스트베드를 구축하고, 한-미간 100G 네트워크 구간에서 전송성능을 실험하였다. 실험 결과 상대적으로 작은 크기의 파일 대비 큰 크기의 파일 전송의 경우와 파일의 개수가 많은 경우 DTN 클러스터링의 효과는 더 크게 나타났다. 향후, 제한된 실험 환경의 개선을 위해서 내포된 성능 저하 요인을 해결할 필요가 있다. 우선, 본 테스트베드 환경에서는 DTN 서버들의 업링크 경로와 NFS 서버/클라이언트 간의 경로가 같은 경로를 공유함으로써 종단간 전송 성능에 영향을 미친다. 추가적으로, NFS는 “Parallelism”에 특화된 네트워크 프로토콜이 아니므로 최적의 클러스터링 효과를 얻기 위해서는 Lustre, GPFS와 같은 “Parallel” distributed protocol을 대상으로 고속 마운트 기법을 적용하여 유사한 실험을 수행할 필요가 있다.

ACKNOWLEDGMENT

본 연구는 2019년도 한국과학기술정보연구원(KISTI) 주요사업 과제로 수행한 것입니다.

참고문헌

[1] B. HARROD, “US Department of Energy Big Data and Scientific Discovery”, http://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/talk4-Harrod.pdf.

[2] A. Khan et al., “SCISPACE: A Scientific Collaboration Workspace for File Systems in Geo-Distributed HPC Data Centers”, arXiv:1803.08228v1, 2018.

[3] Science DMZ, http://fasterdata.es.net/science-dmz/.

[4] 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.

[5] Petascale DTN project, https://cs.lbl.gov/news-media/news/2017/esnets-petascale-dtn-project-speeds-up-data-transfers-between-leading-hpc-centers/.

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