KIPS Trans. Comp. and Comm. Sys.
Vol.9, No.9 pp.197~206
ISSN: 2287-5891 (Print), ISSN 2734-049X (Online)

https://doi.org/10.3745/KTCCS.2020.9.9.197

Deployment and Performance Analysis of Data Transfer Node Cluster for HPC Environment

Wontaek Hong† · Dosik An†† · Jaekook Lee††† · Jeonghoon Moon†††† · Woojin Seok††††

ABSTRACT

Collaborative research in science applications based on HPC service needs rapid transfers of massive data between research colleagues over wide area network. With regard to this requirement, researches on enhancing data transfer performance between major superfacilities in the U.S. have been conducted recently. In this paper, we deploy multiple data transfer nodes(DTNs) over high-speed science networks in order to move rapidly large amounts of data in the parallel filesystem of KISTI's Nurion supercomputer, and perform transfer experiments between endpoints with approximately 130ms round trip time. We have shown the results of transfer throughput in different size file sets and compared them. In addition, it has been confirmed that the DTN cluster with three nodes can provide about 1.8 and 2.7 times higher transfer throughput than a single node in two types of concurrency and parallelism settings.

Keywords : DTN Cluster, Wide Area Network, Data Transfer, Science DMZ, Parallel Filesystem

HPC 환경을 위한 데이터 전송 노드 클러스터 구축 및 성능분석

홍 원 택† · 안 도 식†† · 이 재 국††† · 문 정 훈†††† · 석 우 진††††

요 약

HPC(High Performance Computing) 서비스를 기반으로 한 거대과학 응용분야의 협업연구는 원거리에 떨어져 있는 연구자들 사이에서 대용량 데이터의 빠른 전송을 필요로 한다. 이와 관련하여 최근 미국 내의 주요 슈퍼컴퓨터들을 연계하여 고속 전송하기 위한 연구들이 수행되고 있다. 본 논문에서는 기 구축되어 운영 중인 한국과학기술정보연구원의 누리온 슈퍼컴퓨터 병렬 파일시스템 내의 대용량 데이터를 고속 전송하기 위해서 고성능 과학기술연구망 기반의 데이터 전송 노드(DTN) 클러스터를 구축하고 종단간 왕복지연 시간이 약 130ms에 달하는 원거리 전송 실험을 수행한다. 실험을 통해 다른 크기의 파일들로 구성된 실험 군들에 대해 DTN 클러스터링에 따른 전송 성능을 비교하였고, 3대의 멀티 노드로 구성된 DTN 클러스터는 두 종류의 병행성, 병렬성 설정에서 단일 노드 대비 각각 약 1.8, 2.7배의 전송 성능 향상을 가져올 수 있음을 확인하였다.

키워드 : DTN 클러스터, 원거리 네트워크, 데이터 전송, Science DMZ, 병렬 파일시스템

1. 서 론

최근 세계적으로 천문학, 기상기후, 고에너지물리 등과 같은 거대과학 응용 분야에서는 실험 장비, 관측 데이터, 시뮬레이션 결과 등을 포함한 다양한 데이터 출처로부터 전례 없는 대용량 데이터를 생산해 내고 있고, 유관 연구기관에서는 이러한 과학 빅 데이터를 안정적으로 수용하기 위해 대규모의 컴퓨팅, 스토리지, 네트워크 자원들을 구축하여 운영하고 있다. 특히, 이러한 대용량 데이터를 기반으로 수행되는 과학 응용 연구는 개별 연구자 단독으로 수행되지 않고, 주로 연구 그룹별로 협업을 통해 수행되므로, 원활한 협업 연구를 지원하기 위해서는 공동 연구에 참여하는 연구자들 간에 고성능 망을 기반으로 한 대용량 데이터의 고속 전송이 전제되어야 한다. 실 예로, 미국 에너지성 산하의 연구 기관인 OLCF(Oak Ridge Leadership Computing Facility)에서의 핵 분야 관련 페타스케일 시뮬레이션은 NERSC(National Energy Research Scientific Computing Center)에서 생산되는 핵 관련 상호작용 데이터 세트를 필요로 한다. 유사하게, ALCF(Argonne Leadership Computing Facility)의 기상/기후 분야 연구

※ 본 연구는 2020년도 한국과학기술정보연구원(KISTI) 주요사업 과제로 수행한 것입니다.
※ 이 논문은 2020년 한국정보처리학회 춘계학술발표대회에서 "대용량 스토리지 기반의 데이터 전송 노드 클러스터 설계 및 구축"의 제목으로 발표된 논문을 확장한 것임.
† 정 회 원 : 한국과학기술정보연구원 책임연구원
†† 비 회 원 : 한국과학기술정보연구원 슈퍼컴퓨팅인프라센터 선임기술원
††† 정 회 원 : 한국과학기술정보연구원 슈퍼컴퓨팅인프라센터 연구원
†††† 비 회 원 : 한국과학기술정보연구원 책임연구원
Manuscript Received : July 17, 2020
Accepted : July 28, 2020
* Corresponding Author : Wontaek Hong(wthong@kisti.re.kr)

자들은 관련 시뮬레이션 결과를 OLCF에 있는 기후 관측 데이터들과 비교함으로써 실험의 유효성을 검증한다. 즉, 이러한 대규모 데이터 센터들 간의 협업 연구는 필연적으로 연구 그룹들 간의 대용량 데이터 이동을 기반으로 이뤄진다[1].

한편 거대과학 응용 분야에서의 연구데이터 트래픽은 불특정 다수가 아닌 연구 그룹에 속한 상대적으로 적은 수의 사용자들이 데이터의 크기가 큰 플로우들을 활용한다는 점에서 일반 사용자들이 작은 크기의 플로우들을 빈번하게 이용하는 상용 망에서의 트래픽 패턴과는 구별된다[2]. 따라서 이러한 연구 그룹에 대해 효과적인 망 지원이 요구되므로 세계 각국에서는 전용화된 연구망을 구축 및 운영해 오고 있다. 특히, 미국의 에너지성 산하의 국가과학연구망인 ESnet(Energy Sciences Network)에서는 Science DMZ라는 일반 상용망과는 차별화된 전용의 연구망을 위한 네트워크 설계 패턴을 제안하였고, 이는 세계의 여러 연구망 커뮤니티에서 활발히 적용되어 오고 있다. 이러한 Science DMZ는 고품질의 고성능 네트워크, 대용량 데이터 전송 툴을 탑재한 전용의 데이터 전송 노드(DTN), 네트워크 성능 측정 서버 등으로 구성된다[3, 4]. 또한, 최근에는 HPC 환경을 제공하는 미국의 주요 슈퍼컴퓨팅 센터들(NERSC, OLCF, ALCF, National Center for Supercomputing Applications) 간에 ESnet을 중심으로 과학 빅데이터에 대한 전송 효율을 극대화시키기 위한 Petascale DTN 프로젝트가 수행되고 있다. 이 프로젝트의 궁극적인 목표는 Science DMZ 개념을 기반으로 미국 내 주요 슈퍼컴퓨터 사이트들이 1주에 1페타바이트의 대용량 데이터를 송수신할 수 있는 환경을 구축하는 함으로써 참여 기관의 HPC 환경을 이용하는 응용 연구자들에게 양질의 협업 연구 환경을 제공해 주는 것이다[5].

본 논문에서는 한국과학기술정보연구원(KISTI)에서 운영중인 슈퍼컴퓨터 5호기 누리온의 병렬 파일시스템인 Lustre 파일 시스템을 대상으로 원거리 대용량 데이터의 고속 전송이 가능하도록 Globus 소프트웨어 기반의 DTN 클러스터를 구축하고 전송 성능을 측정 및 분석한다. 특히, 전송 성능이 측정된 구간은 한미간 약 130ms의 왕복지연 시간이 존재하는 원거리 구간으로 종단간 원거리 전송 시 병렬성, 병행성 등을 반영한 DTN 클러스터링에 따른 송수신 전송 성능을 비교 분석한다. 세부적으로 크기가 다른 파일들로 구성된 데이터 세트에 대해 전송 성능의 차이를 비교하고, DTN 클러스터링에 따른 병행성, 병렬성의 이점을 잘 활용할 수 있는 데이터 세트를 분류한다. 또한, 이를 바탕으로 멀티 노드로 구성된 DTN 클러스터가 단일 노드와 비교하여 병행성, 병렬성 수준의 향상을 통해 전송 성능 향상을 가져올 수 있음을 정량적으로 제시한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 데이터 전송 노드 클러스터링 구성 시에 전송 성능 향상을 가져올 수 있는 전송 파라미터들을 간략히 기술하고, 이를 반영하여 구축된 병렬 파일시스템 연계 데이터 전송 노드 클러스터 시스템에 대해 설명한다. 3장에서는 구축된 시스템의 전송 성능 측정을 위한 실험 환경 및 한미간 국제망 구간에서 수행된 DTN 클러스터링의 전송 성능 평가 결과를 제공한다. 그리고 4장에서는 관련 연구들을 소개하고 마지막으로 5장에서 본 논문의 결론을 맺는다.

2. 병렬 파일시스템 연계 DTN 클러스터링

HPC 환경에서는 주로 Lustre, GPFS와 같은 병렬 파일시스템을 활용하여 대용량 데이터를 저장한다. 이러한 외부 스토리지에 저장된 대용량 데이터를 협업 연구자의 연구 환경으로 전송하기 위해서 일반적인 네트워크 환경을 이용하는 것은 전송 효율 저하로 이어진다. 이러한 상황을 극복하기 위해 제안된 Science DMZ 개념은 전용의 데이터 전송 노드(DTN)를 활용하여 데이터 전송 성능을 향상시킬 수 있다[3, 4]. 특히, Science DMZ 적용 모델 중 외부 스토리지를 DTN에 마운트하여 고속 전송하는 방법론은 병렬 파일시스템 내의 대용량 데이터를 효과적으로 전송 및 공유할 수 있게 한다.

2.1 고속 전송을 위한 전송 파라미터

HPC 환경에서 Science DMZ 기반의 DTN을 적용할 경우 대용량 데이터의 전송 효율 향상과 관련하여 Fig. 1과 같은 전송 파라미터들을 적용하여 전송 성능을 향상 시킬 수 있다[6, 7]. 첫째, 병렬성(Parallelism)은 하나의 파일들을 여러 블록으로 나누어 동시에 전송하여 전송 성능을 향상시킬 수 있는 방법으로 하나의 파일 전송 프로세스에서 지원 가능한 TCP 스트림들의 개수로 표현 가능하다. 둘째, 병행성(Concurrency)은 다수의 파일들을 각각의 파일 전송 프로세스에 매핑시켜 전송 성능을 향상 시킬 수 있는 방법으로, 다수의 전송 노드 및 노드 내의 다수의 CPU 코어 등을 최대한 활용할 수 있는 방법이다. 셋째, 파이프라이닝(Pipelining)은 하나의 파일 전송 프로세스에서 다수의 전송 커맨드를 수행함으로써 전송 성능을 향상시키고자 하는 방법으로, 응용 수준에서 전송 성능을 향상시킬 수 있다. 즉, 하나의 파일 전송 프로세스에서 단일 파일에 대해 전송을 개시한 후, 완료가 될 때까지 기다리지 않고, 다음 파일을 전송하고자 하는 경우 전송 성능의 향상을 가져올 수 있다.

위에서 언급된 전송 파라미터들은 다수의 대용량 파일들의 고속 전송을 위해 Gridftp, Globus 전송 서비스[8]와 같은 파일 전송 프로토콜 및 서비스에 적용되고 있다. 이러한 전송 파라미터들은 단독으로 또는 복합적으로 적용되어 활용될 수 있다. 경우에 따라 위와 같은 기법들이 전송 성능 향상을 항상 보장하는 것은 아니다. 일례로, 거대과학 응용분야의 대용량 데이터 전송에서 가장 활발히 이용되는 Globus 전송 서비스의 경우 병렬성, 병행성, 파이프라이닝 등을 지원하지만 실효성 측면에서 파이프라이닝의 수준을 1로 한정한다. 그러므로,

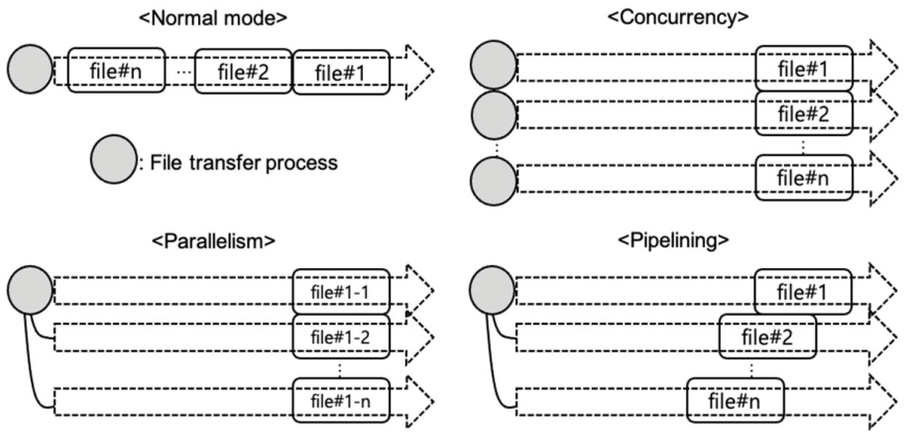

Fig. 1. Comparison of Transfer Parameters

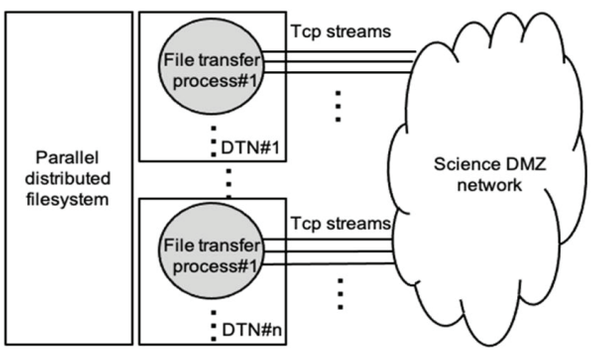

Fig. 2. Concurrency and Parallelism in DTN

전송규모, 전송 툴의 특성, 네트워크 성능 및 경로 등을 복합적으로 고려한 실질적인 전송 환경의 구축은 매우 중요하다.

특히, 위 전송 파라미터들 중 병렬성과 병행성은 원거리 대용량 데이터 전송의 효율의 향상을 위해 Globus 전송 서비스에서 활발히 활용되고 있는 개념으로 종단 포인트의 시스템 성능, 망 성능에 기반하여 병렬성과 병행성의 정도를 조절할 수 있다. 이러한 조절 가능한 전송 파라미터들은 전송 데이터의 규모 및 특성, 전송 속도, 스토리지 연계 등 다양한 전송 환경에 따라 전송 성능에 영향을 줄 수 있다. Fig. 2는 최근 거대과학 응용분야에서 활발히 적용 되고 있는 Science DMZ 모델에 기초하여 병렬 파일시스템상의 다수의 대용량 파일들을 전송하고자 할 때, 전용의 데이터 전송 노드들에 병렬성과 병행성을 지원하는 전송 프로토콜들이 어떻게 적용될 수 있는지를 개념적으로 보여준다.

2.2 데이터 전송 노드 클러스터 설계 및 구축

대용량 스토리지들을 Science DMZ 환경에 적용하기 위해 고속 마운트 기반의 외부 스토리지 연계 모델에 기초하여 DTN 클러스터를 설계하고 구축한다. 이러한 접근 방법은 기 구축된 병렬 파일시스템을 다수의 DTN 서버에 마운트함으로써 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 최적화를 위해 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 외부 망 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(Maximum Transmission Unit) 크기 세팅 등을 포함한 DTN 운영체제 및 전송 프로토콜에 대한 소프트웨어 최적화가 선행되어야 한다.

병렬 파일시스템 내의 다수의 대용량 파일들을 Science DMZ 환경에서 고속으로 전송하기 위해서는 DTN 서버들 또한 확장 가능한 병행성과 병렬성을 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 프로비져닝 등을 통해 병목이 없는 전용 경로를 확보한다. 또한, DTN 서버에 설정된 점보프레임 지원을 위한 MTU 설정

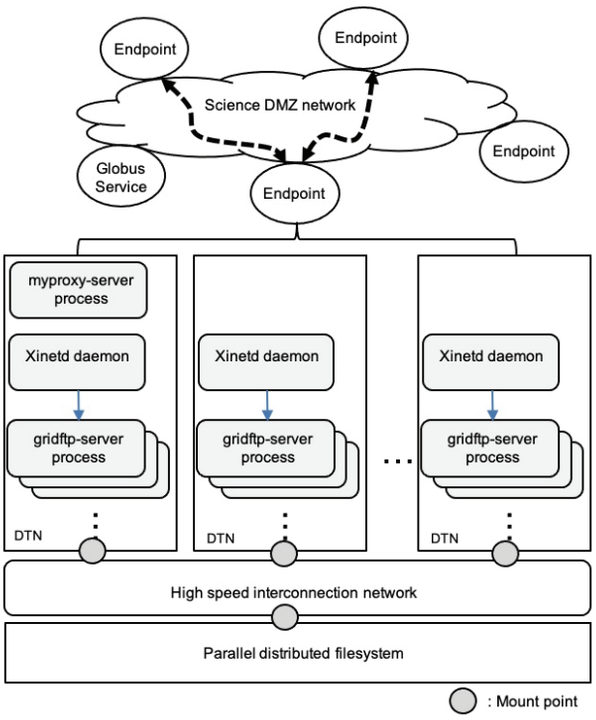

Fig. 3. Structure of DTN Cluster

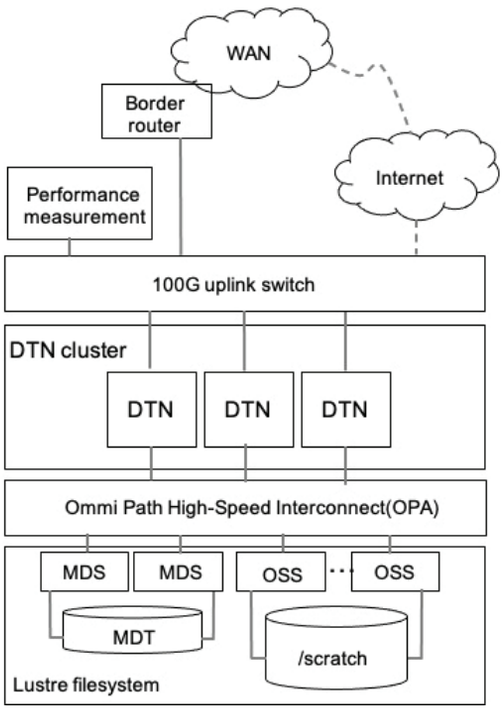

Fig. 4. DTN Cluster Deployment

과 더불어 상대방 DTN 서버들 또한 동일한 MTU 크기의 설정이 필요하다. 이러한 MTU 크기의 세팅은 DTN 네트워크 인터페이스뿐만 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.

Fig. 3은 위에서 언급된 요구사항들을 반영하여 병렬 파일시스템 내의 파일들을 고속 전송하기 위해 Globus Connect Server(GCS) 소프트웨어[8]를 활용하여 다수의 DTN 서버들을 클러스터링한 구조도를 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 각각의 DTN 내의 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 프로세스 당 N개의 TCP 스트림들을 생성한다면 산술적으로 최대 LxMxN개의 TCP 스트림들을 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.

Fig. 4에서와 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. 슈퍼컴퓨터의 login 노드 등이 이용하는 기존 망(국내 망)을 경유하여 DTN 전송을 수행할 경우 전용 경로 확보 측면에서 Science DMZ의 망 설계 개념에 부합되지 않으므로 외부 망(국제 망)을 직접 통하여 전송하는 망 경로를 선택한다. 이럴 경우, 기존 망에서 Firewall 이슈 등을 국제 망의 Border 라우터/스위치의 ACL(Access Control List)에서 세분화된 수준으로 반영해야 하므로 세심한 작업이 요구된다. 부수적으로 Science DMZ 기반의 DTN 클러스터링의 경우 일반적인 쉘을 통한 접근을 엄격히 통제하므로, 보안적인 측면에서 안전성을 높일 수 있다. 추가적으로 DTN 클러스터들이 경유하는 업링크 스위치에서도 많은 사용자들의 트래픽이 집중될 경우를 대비하여 면밀한 전송성능 분석을 위해 망 성능 측정 서버의 적용이 필요하다.

3. 실험 및 성능 분석

3.1 실험환경 구성

Fig. 5는 HPC 환경에서 Lustre 파일시스템을 기반으로

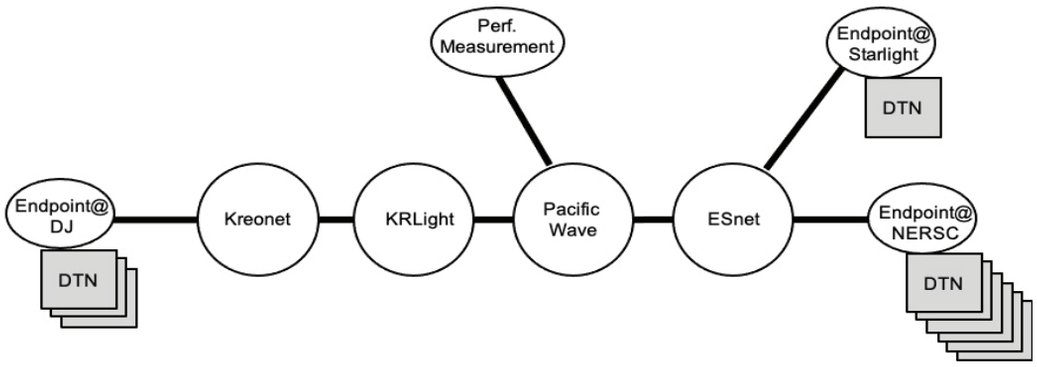

Fig. 5. Experiment topology

구축된 DTN 클러스터의 원거리 전송 성능을 측정하기 위한 실험 환경을 보여준다. 실험 환경은 다음의 3가지 전송 실험들을 수행하기 위해 구성된다. 첫째, 궁극적인 Disk-to-Disk (D2D) 전송 성능을 측정하기에 앞서, 개별 DTN 노드들의 Memory-to-Memory(M2M) 전송 성능을 측정한다. 둘째, DTN 클러스터들 간의 전송 성능 측정에 앞서 DTN 클러스터와 ESnet Starlight에 위치하는 단일 DTN 간의 D2D 전송 실험을 수행하여 제한적인 DTN 클러스터링의 전송 성능을 측정한다. 마지막으로, 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 상대편 종단 포인트도 DTN 클러스터로 구성되어야 자원들을 충분히 활용할 수 있다. 이러한 맥락에서 미국 NERSC의 Cori 파일시스템과 연동되어 있는 DTN 클러스터 종단 포인트와의 전송 테스트를 수행한다.

Table 1은 Fig. 5의 전송 실험 환경을 좀 더 자세히 표 형태로 정리한다. 각 DTN에는 CentOS v7.5.1804이 설치되고, DTN 자체 성능 최적화와 관련하여 원거리 전송에 적합한 TCP 혼잡제어 알고리즘, TCP 소켓버퍼 크기 설정, MTU 크기 설정 등이 기본적으로 이뤄진다. DTN 클러스터들이 연결되는 외부 스토리지 시스템으로는 Lustre 병렬 파일시스템이 100Gbps 인터커넥션 네트워크로 연결되고, DTN 클러스터 연계를 위해 Lustre client v2.10.7이 설치되어 적용된다. 전송 툴과 관련하여 M2M 및 D2D 전송 테스트를 위해 각각 iperf와 GCS(Globus Connect Server)가 적용된다. 특히, GCS는 DTN 클러스터링의 효과를 극대화하기 위한 방법으로 병행성, 병렬성의 수준을 조절할 수 있도록 지원한다. 이러한 DTN 클러스터링 효과에 대한 병렬성, 병행성 모니터링을 위해서는 DTN 노드별로 ps, top, netstat 등을 설치하여 구동중인 전송 프로세스의 수, CPU 이용률, TCP stream의 수 등을 각각 모니터링 한다. 구축된 종단 포인트와 M2M 전송 성능을 측정하기 위해 약 112ms의 RTT(Round Trip Time)를 갖는 미국 시애틀에 위치한 성능 측정 노드를 이용하고, D2D 전송 성능을 측정하기 위해서 각각 약 157ms와 130ms의 RTT를 갖는 미국 시카고에 위치한 ESnet DTN과 미국 버클리에 위치한 NERSC Cori DTNs을 이용한다. 특히, NERSC Cori DTN 종단 포인트는 6대의 DTN 서버들로 구성되어 있고, 본 논문에서 구축된 종단 포인트와의 네트워크 경로는 한국의 과학기술연구망(KREONET), 미국의 ESnet 등을 통해 제공되고, 종단간 대역폭은 100Gbps이다.

Table 1. Experiment Environment in Detail

	Description
Operating system	CentOS v7.5.1804
TCP algorithm	HTCP
Socket buffer size	approx. 2.5Gbytes (100Gbps*0.2sec*1/8byte) Tuned for Large BDP
MTU	9000 bytes (for Jumbo frame)
Mounted FS client	Lustre client v2.10.7
Transfer tool	iperf v2.0.13 Globus connect server v4
Monitoring tool	ps, top, netstat
M2M endpoint	Perf. node (RTT: approx. 112ms)
D2D endpoint	ESnet DTN (RTT: approx. 157ms) NERSC DTNs (RTT: approx. 130ms)
Bandwidth	100Gbps

3.2 Memory-to-Memory 전송 실험

DTN의 종단 포인트들 간의 D2D 전송을 수행하기 전에 (Fig. 5)에서의 성능 측정 노드를 활용하여 M2M 전송 성능을 먼저 확인한다. DTN 클러스터 종단 포인트들 간의 전송은 무수한 TCP stream들을 기반으로 수행되므로 iperf 등의 전송 툴을 활용한 M2M 전송 성능의 측정 과정을 거친다. 가장 이상적인 방법은 DTN 클러스터에 해당하는 노드들을 대상으로 M2M 전송 성능 테스트를 수행하는 것이지만, Globus를 기반으로 종단 포인트화 된 노드들은 보안관리 측면에서 쉘 접근을 원칙적으로 허용하지 않으므로, 본 테스트는 한미간 원거리 구성에서 100ms 이상의 RTT를 갖는 성능 측정 노드를 이용하여 수행된다.

각 전송별로 100초간 M2M 전송 테스트의 수행결과, DTN 클러스터를 구성하는 각각의 단일 노드들과 성능 측정 노드 간의 TCP 단일 스트림에 대한 전송 성능은 DTN 클러스터의 단일 노드 기준으로 송신 시 약 17.8Gbps, 수신 시 약 26.2Gbps를 기록하였다. 추가적으로 수행된 5개의 TCP 복수 스트림들에 대한 전송 성능은 DTN 클러스터의 단일 노드 기준으로 송신 시 약 71.9Gbps, 수신 시 약 73.3Gbps를 기록하였다.

3.3 Disk-to-Disk 전송 실험

1) 멀티노드와 단일노드 간 전송

구축된 DTN 클러스터와 ESnet Starlight에 위치하는 단일 DTN 사이의 D2D 전송 실험에서 이용되는 전송 파일 set은 ESnet DTN의 Climate-Large 디렉토리에 있는 약 244GB 크기의 파일들로서 10개의 21.5GB의 파일과 1개의 28.8GB 파일로 구성된다. 참고로, 이 테스트 데이터 set은 기상기후 응용분야 ICNWG(International Climate Network Working Group) 멤버 사이트에서의 활용을 위해 제공되고 있고, 기상기후 모델 데이터를 포함한다[9].

본 실험은 멀티노드 기반의 병행성을 지원하는 DTN 클러스터 노드와 단일 노드만으로 구성된 DTN 간의 전송 성능 테스트이므로, 제한된 전송 성능이 예상된다. 실제로 ESnet DTN은 단일 노드의 제한된 구성 외에 100G NIC이 아닌 10G NIC으로 구성되어 본 논문에서 구축된 DTN 클러스터의 전송 성능을 확인하기에는 부족한 환경이다. 이러한 제한된 환경에서 DTN 클러스터에 속한 노드 수를 조절하면서 전체 전송 성능을 확인한 결과 최대 9.28Gbps의 D2D 전송 성능을 확인하였고, Table 2에서와 같이 ESnet DTN의 단일 노드에 따른 제한된 병행성 및 10G NIC의 병목현상을 감안할 때 노드 증가에 따른 제한된 병렬전송 성능을 확인할 수 있었다.

Table 2. Transfer throughput from ESnet single DTN

	One DTN	Two DTNs	Three DTNs
Throughput (Gbps)	8.48	9.2	9.28

2) 멀티노드와 멀티노드 간 전송

위 전송 실험에서 확인하였듯이, 구축된 DTN 클러스터의 전송 성능을 확인하기 위해서는 전송에 참여하는 종단 포인트도 비슷한 스펙의 DTN 클러스터로 구성되어 양 종단간 대용량 데이터의 전송 시 병행성, 병렬성을 높일 수 있어야 한다. 이러한 관점에서 실험에 활용되는 상대방 종단 포인트로서 원거리에 존재하는 미국 NERSC Cori DTN 클러스터와의 전송 실험을 진행한다.

본 논문에서 구축된 DTN 클러스터와 NERSC Cori DTN 사이의 D2D 전송 성능에서 이용되는 전송 파일 set은 ESnet DTN에서 제공되는 약 250GB~500GB의 데이터 set와 이 데이터 set을 기반으로 재생산된 약 500GB~1.5TB의 데이터 set이다. 재생산된 데이터 set은 Climate-Small과 Climate-Huge을 제외한 나머지 데이터 set들을 바탕으로 구성된다. Table 3은 실험에 이용되는 데이터 set에 대한 세부 정보를 제공한다.

Table 3. Transfer Data Sets

	Size	Description
Climate-Small	246GB	1,496 files, ranging in size from 29MB to 425MB
Climate-Medium	245GB	117 files, ranging in size from 1.2GB to 6GB
Climate-Large	244GB	10 files, each 21.5GB plus one 28.8GB file
Climate-Huge	245GB	two files, each ~120GB
500GB-in-large-files	500GB	100MB, 200MB, and 500MB files in each leaf directory
AddedSet#1	490GB	2*Climate-Medium
AddedSet#2	978GB	AddedSet#1 plus 2*Climate-Large
AddedSet#3	1.48TB	AddedSet#2 plus 500GB-in-large-files

DTN 클러스터링은 외부 스토리지 상의 다수의 대용량 파일들을 병행성, 병렬성의 증가를 통해 전체적인 전송 성능을 향상시키는 방법이다. 이러한 외부 파일시스템 대상 마운트 기반의 파일 전송 시, 내부 파일시스템에 직접 전송하는 경우 대비 성능 감소 정도를 파악하는 것이 필요하다. 이와 관련하여, 단일 파일을 마운트 기반의 Lustre 외부 파일시스템에 저장하는 경우와 엔드포인트 접근 계정의 홈 디렉토리에 직접 저장하는 경우를 비교한다. Climate-Huge 디렉토리에 있는 약 120GB의 단일 파일을 대상으로 실험을 한 결과, 내부 파일시스템 및 외부 파일시스템에 전송하는 경우 각각 약 7.2Gbps, 5.4Gbps의 전송 성능을 나타내 약 25% 성능 차이를 보였다. 이러한 전송 성능의 차이는 복수 개의 전송 파일 set을 대상으로 DTN 클러스터링에 따른 병렬성, 병행성의 증가 이점을 활용하면 충분한 성능 보상을 기대할 수 있을 것이다.

구축된 DTN 클러스터와 NERSC DTN 클러스터 사이에서 송수신 시의 전송 성능을 비교하기 위해 약 250GB 크기의 Climate-{Small, Medium, Large, Huge} 데이터 set을 대상으로 전송 실험을 수행한다. 수행결과 Fig. 6에서와 같이 최소 약 11.4Gbps에서 최대 약 37.0Gbps의 전송 성능을 보였고, 구축된 DTN 클러스터를 기준으로 모든 데이터 set에 대해 수신 시의 성능이 송신 시 보다 우수하였다. 이는 NERSC DTN 클러스터를 구성하는 서버의 스펙과 누리온 DTN 클러스터를 구성하는 서버 스펙의 차이 및 NERSC DTN의 TCP 소켓버퍼 크기가 미국 내 주요 HPC 센터들과의 BDP(Bandwidth Delay Product)를 기준으로 설정되어

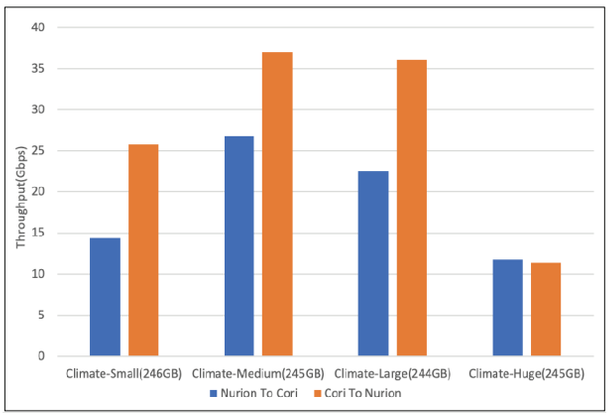

Fig. 6. Comparison of Transfer Throughput (I)

100ms 이상의 한미간 원거리에 대한 BDP를 충분히 반영하여 설정되지 못한 것이 주요 원인이 될 수 있다. 추가적으로 데이터 크기에 따른 전송 성능의 영향과 관련하여 Climate-{Medium, Large} 데이터 set의 경우가 Climate-{Small, Huge} 데이터 set에 비해 성능이 우수하였다. Climate- Huge 데이터 set의 경우 약 120GB 크기의 파일 2개로 구성되어 있어 구축된 DTN 클러스터의 병행성을 충분히 활용하지 못한 경우로 볼 수 있고, Climate-Small 데이터 set의 경우는 Climate-{Medium, Large} 데이터 set에 비해 다수의 작은 파일들로 구성되어 있어서 발생하는 상대적인 전송 성능 저하로 판단할 수 있다.

Fig. 7은 위의 실험에서 DTN 클러스터링의 효과가 우수한 데이터 set을 기반으로 새로운 데이터 set을 재생산하여 전송 실험을 수행한 결과를 보여준다. 본 실험에서는 Climate-{Medium, Large} 데이터 set과 500GB-in-large -files 데이터 set을 기반으로 약 500GB, 1TB, 1.5TB 크기로 재생산된 데이터 set을 활용하여 구축된 DTN 클러스터를 기준으로 송수신 시의 전송 성능을 측정하여 비교한다. 이전 실험에서처럼 모든 데이터 set에 대해 수신시의 전송 성능이 송신시의 전송 성능 보다 우수함이 유지되었고, 동일한 Climate-{Medium,

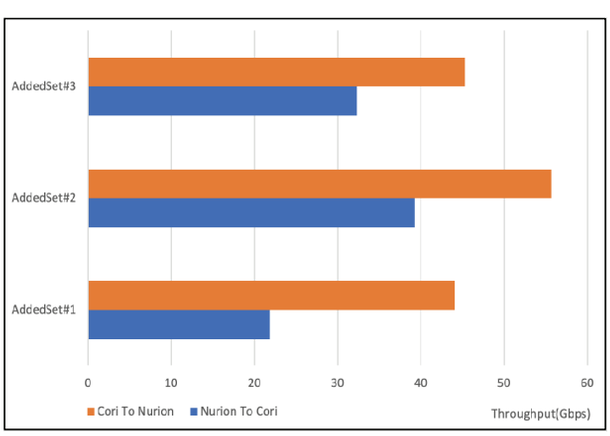

Fig. 7. Comparison of Transfer Throughput (II)

Large} 데이터 set이지만 전송 규모가 증가함에 따라 Fig. 6의 결과에 비해 DTN 클러스터링에 의한 전송 성능의 이득을 더 가져오는 것을 확인할 수 있다. 특히, AddedSet#2의 수신시 전송 성능은 약 55.7Gbps를 기록하였고, 이는 DTN 클러스터링에 따른 병행성, 병렬성 증가의 효과를 상대적으로 잘 활용한 결과에 해당된다.

Fig. 8은 DTN 클러스터링의 전송 성능 이득이 가장 컸던 AddedSet#2을 이용하여 DTN 노드 수의 증감 및 TCP 스트림 수에 영향을 주는 병행성(C), 병렬성(P)에 변화를 주면서 DTN 클러스터링의 효과를 비교한 결과를 제공한다. 실험결과, 노드 수(N)가 1에서 3으로 증가하면서 Mode I(C:8, P:16)에서는 약 30.3Gbps에서 54.9Gbps로 전송 성능이 증가하였고, Mode II(C:4, P:8)에서는 약 18.2Gbps에서 49.8Gbps로 전송 성능이 증가하는 것을 볼 수 있다. 이러한 전송 성능의 향상은 전체적인 병행성 및 병렬성의 증가에 기반하여 증가된 전체 TCP 스트림 수에 기인한다. 실제로 Mode I에서 노드 수가 3인 경우, 최대 384개(N*C*P)의 전체 TCP 스트림들이 생성되어 전송에 참여할 수 있고, Fig. 9는 이렇게 생성되

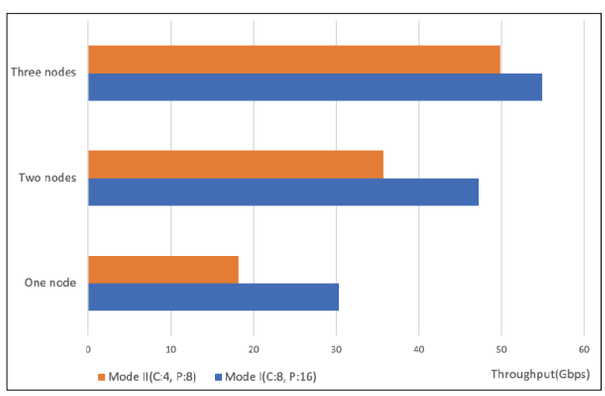

Fig. 8. Comparison of Transfer Throughput (III)

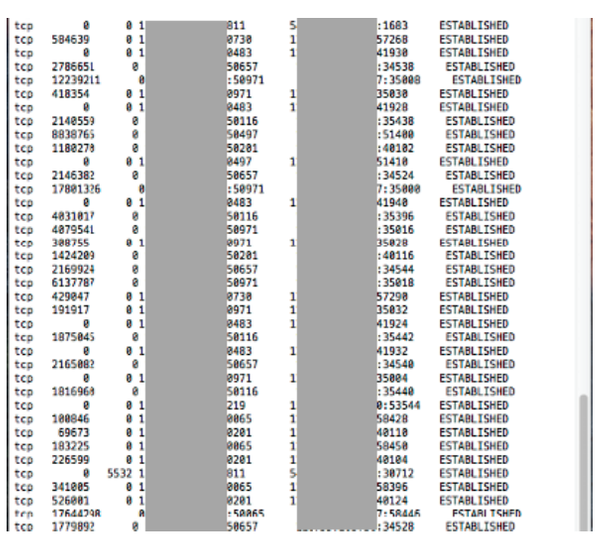

Fig. 9. TCP Stream Status Using Netstat

어 전송에 참여한 TCP 스트림의 상태 정보를 “netstat” 명령어를 이용하여 3대의 노드 중 1대에서 일부를 캡쳐하여 보여준다. Fig. 9에서 볼 수 있듯이 구축된 DTN 클러스터로의 수신이기 때문에 “Local Address”의 “Send-Q”가 아닌 “Recv-Q” 값이 카운트되는 것을 확인할 수 있다.

4. 관련 연구

기상기후, 천문학, 고에너지물리 분야와 같은 거대 과학응용 분야에서 생성되는 대용량 데이터의 전송 성능을 향상시키기 위한 연구들이 수행되고 있다. 이러한 연구들은 GridFTP, Globus 전송 서비스 등과 같이 현재 거대 과학응용 분야에서 활발히 활용되는 전송 툴에 대한 누적 로그들을 기반으로 데이터 전송에 대한 특성 및 패턴을 분석하는 연구[7,10,11], 종단간 대용량 파일 전송 시 잠재적인 병목 지점에 대한 분류 및 성능 측정[12], 전용 데이터 전송 노드의 고성능 하드웨어 스펙을 충분히 활용하기 위한 전송 프로토콜의 설계 및 구현[13] 등을 포함한다.

[7]에서는 GridFTP 및 Globus 전송 서비스와 관련하여 누적된 전송 로그들을 기반으로 원거리 전송에 대한 특성을 체계적으로 분석하여 제공한다. 이러한 정보들에는 데이터 전송 노드의 활용 정도, 원거리 전송에서의 데이터 손상 및 반복 전송, 전송 파일의 유형 및 전송 성능, 사용자 이용 패턴 등을 포함한다. 또한, 전송 파라미터의 설정과 관련하여 누적되어 활용되어 온 병행성과 병렬성 정도의 평균적인 수치를 제시함으로써 전송 파라미터 설정 시 참고할 수 있는 정보를 제시한다.

R. Kettimuthu et al.[10]는 미국 내의 ANL(Argonne National Laboratory)과 NCSA(National Center for Supercomputing Applications) 사이에서 복수 개의 DTN 노드들을 활용하여 하루에 총 1PB의 천문학 관련 대용량 데이터를 전송한 사례를 제시한다. 단, 이 실험 환경은 각각 NCSA 사이트에 12대의 DTN, ANL 사이트에 28 대의 DTN 노드들이 적용되었고, 두 사이트 간의 RTT는 약 6ms에 달한다. 추가적으로 Globus 전송 서비스의 부가적인 기능인 전송 데이터의 “무결성 검증” 기능을 활성화했을 때와 하지 않았을 때의 성능 차이를 측정하고, 주어진 실험 환경에서 최대의 전송 성능을 기록할 수 있는 최적의 파일 크기, 수 등을 실험하여 제시한다.

Z. Liu et al.[11]는 [7]과 유사하게 HPC facilities에서의 데이터 전송의 특징을 GridFTP 로그와 TSTAT 로그를 기반으로 분석한다. 특히, Science DMZ 기반의 데이터 전송 노드에 GridFTP 서버 프로세스의 파이프라이닝이 활성화 된 상태에서 일정 수준의 병행성을 적용 시 성능 저하 문제, 단일 GridFTP 서버 프로세스에서 병렬 TCP 스트림들이 적용되었을 때의 부하 불균형 문제를 지적한다. 이에 대한 대응으로 파이프라이닝의 수준을 낮게 설정하여 부하 불균형의 정도를 낮춘 결과를 제시하고, 병행성과 병렬성의 적용 등을 통해 부하 불균형의 수준을 개선시킬 수 있음을 언급한다.

Y. Liu et al.[12]는 거대과학 응용에서 발생하는 상대적으로 작은 크기의 파일들을 HPC facilities 간에 원거리 전송할 때 전송 성능이 낮은 원인을 분석한다. 면밀한 분석을 위해 스토리지 시스템, 메모리, 네트워크 구간 등을 고려하여 읽기, 쓰기 시의 오버헤드를 구간별로 측정함으로써 종단간 파일 전송 시의 파일 당 발생하는 오버헤드가 서브시스템들에서 발생하는 오버헤드보다 훨씬 크다는 것을 확인한다. 또한, 파일 당 발생하는 오버헤드를 줄이기 위한 방법으로 전송 시 병행성을 높여 복수 개의 파일들을 동시에 전송하는 방법과 송신지 스토리지 시스템 상에서 다음 전송 파일의 일부분을 프리패치하는 방식을 통해 성능 향상을 가져올 수 있음을 보인다.

[13]에서는 고성능 데이터 전송을 위한 전송 툴인 mdtmFTP를 설계하고 구현한다. mdtmFTP는 데이터 전송 노드의 멀티코어 하드웨어를 충분히 활용하지 못하는 단점을 지적하고, 최적화된 쓰레드 스케줄링이 가능한 MDTM(Multicore-Aware Data Transfer Middleware) 미들웨어 서비스 상에서 Pipelined I/O 중심 구조로 설계된다. 또한, 전송 성능 저하의 원인인 LOSF(lots of small files) 문제[14]를 해결하기 위해 크기가 작은 파일들의 일괄 처리를 가능하게 하는 “large virtual file” 전송 메커니즘을 제안한다. 성능 검증을 위해 ESnet 100G 네트워크 테스트베드에서 GridFTP, FDT, BBCP 등 기존의 전송 툴과의 전송 성능을 비교한다.

5. 결 론

본 논문에서는 기 구축되어 운영 중인 KISTI 누리온 슈퍼컴퓨터의 병렬 파일시스템인 Lustre 파일시스템을 대상으로 원거리 전송성능을 향상시킬 수 있는 Science DMZ 기반의 데이터 전송 노드 클러스터를 구축하고 전송 성능을 측정 및 분석하였다. 실험 결과 왕복지연 시간이 100ms 이상인 실험 환경에서 멀티 노드들로 구성된 종단 포인트들 간의 전송 시에 병렬성, 병행성의 정도에 따른 전송 성능의 향상을 확인할 수 있었고, 추가적으로 DTN 클러스터링에 따른 병렬성, 병행성의 증가 시 생성된 실제 TCP 스트림들의 상태 정보를 제시하였다. 또한, 다른 크기의 파일들로 구성된 데이터 세트에 대해 전송 성능의 차이를 확인하고, DTN 클러스터링에 따른 병행성, 병렬성의 효과를 극대화할 수 있는 데이터 세트를 분류하였다. 앞으로 거대과학 응용 분야에서의 효과적인 협업 지원을 위해서 대용량 실험 데이터의 고속 전송은 꾸준히 요구될 것이고, 본 논문에서 제시된 결과들은 누리온 슈퍼컴퓨터 및 과학기술연구망을 활용하는 연구자들의 협업연구 환경 개선에 도움을 줄 것으로 기대된다.

References

[1] A. Khan, T. Kim, H. Byun, and Y. Kim, "SCISPACE: A scientific collaboration workspace for geo-distributed HPC data centers," *International Journal of Future Generation Computer Systems*, Vol.101, pp.398-409, 2019.

[2] C. Laat, E. Radius, and S. Wallace, "The rationale of the current optical networking initiatives," *International Journal of Future Generation Computer Systems*, Vol.19, No.6, pp.999-1008, 2003.

[3] E. Dart, L. Rotman, B. Tierney, M. Hester, and J. Zurawski, "The Science DMZ: A Network Design Pattern for Data-Intensive Science," *Proceedings of IEEE/ACM Annual SuperComputing Conference (SC13)*, Denver, USA, Nov. 2013.

[4] 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.

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

[6] E. Yildirim, E. Arslan, J. Kim, and T. Kosar, "Application-Level Optimization of Big Data Transfers through Pipelining, Parallelism and Concurrency," *IEEE Transactions on Cloud Computing*, Vol.4, No.1, pp.63-75, 2016.

[7] Z. Liu, R. Kettimuthu, I. Foster, and N. Rao, "Cross-Geography Scientific Data Transferring Trends and Behavior," *Proceedings of the 27th International Symposium on High-Performance Parallel and Distributed Computing*, New York, USA, Jun. 2018.

[8] Globus project [Internet], https://docs.globus.org/ globus-connect-server/.

[9] ESnet DTNs [Internet], https://fasterdata.es.net/ performance-testing/DTNs/.

[10] R. Kettimuthu, Z. Liu, D. Wheeler, I. Foster, K. Heitmann, and F. Cappello, "Transferring a Petabyte in a Day," *International Journal of Future Generation Computer Systems*, Vol.88, pp.191-198, 2018.

[11] Z. Liu, R. Kettimuthu, I. Foster, and Y. Liu, "A Comprehensive Study of Wide Area Data Movement at a Scientific Computing Facility," *Proceedings of IEEE International Conference on Distributed Computing Systems*, Vienna, Austria, Jul. 2018.

[12] Y. Liu, Z. Liu, R. Kettimuthu, N. Rao, Z. Chen, and I. Foster, "Data transfer between scientific facilities - bottleneck analysis, insights, and optimizations," *Proceedings of the 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, Larnaca, Cyprus, May 2019.

[13] L. Zhang, W. Wu, P. DeMar, and E. Pouyoul, "mdtmFTP and its evaluation on ESNET SDN testbed," *International Journal of Future Generation Computer Systems*, Vol.79, pp.199-204, 2018.

[14] J. Bresnahan, M. Link, R. Kettimuthu, D. Fraser, and I. Foster, "GridFTP Pipelining," *Proceedings of the TeraGrid2007 Conference*, Madison, USA, Jun. 2007.

홍 원 택

https://orcid.org/0000-0002-8057-5204
e-mail : wthong@kisti.re.kr
1998년 성균관대학교 정보공학과(학사)
2000년 성균관대학교 전기전자 및 컴퓨터공학과(석사)
2019년 성균관대학교 전기전자컴퓨터공학과(박사)
2000년 ~ 2002년 ㈜콤텍시스템 기술연구소 연구원
2002년 ~ 현 재 한국과학기술정보연구원 책임연구원
관심분야 : 망 성능 분석, 대용량 데이터 전송 프로토콜, 고성능 네트워크 구조, 클라우드 네트워킹

안 도 식

https://orcid.org/0000-0002-4075-7580
e-mail : dsan@kisti.re.kr
2008년 전북대학교 전자·정보공학부(학사)
2010년 전북대학교 컴퓨터공학(석사)
2017년 전북대학교 컴퓨터공학(박사)
2017년 ~ 현 재 한국과학기술정보연구원 슈퍼컴퓨팅인프라센터 선임기술원
관심분야 : High Performance Computing & Interconnect Network

이 재 국

https://orcid.org/0000-0002-6159-3124
e-mail : jklee@kisti.re.kr
2002년 충남대학교 컴퓨터과학과(학사)
2004년 충남대학교 컴퓨터과학과(석사)
2012년 충남대학교 컴퓨터공학과(박사)
2010년~2013년 한국인터넷진흥원 책임연구원
2013년 ~ 현 재 한국과학기술정보연구원 슈퍼컴퓨팅인프라센터 연구원
관심분야 : 시스템 및 네트워크 보안, HPC 시스템

문 정 훈

https://orcid.org/0000-0003-2870-7676
e-mail : jhmoon@kisti.re.kr
1997년 경일대학교 컴퓨터공학과(학사)
1999년 경북대학교 컴퓨터공학과(석사)
2000년 ~ 현 재 한국과학기술정보연구원 책임연구원
관심분야 : ScienceDMZ, 네트워크 QoS, 가상화, SDN, 클라우드 컴퓨팅, TCP 성능향상

석 우 진

https://orcid.org/0000-0002-7340-9961
e-mail : wjseok@kisti.re.kr
1996년 경북대학교 컴퓨터공학과(학사)
2002년 University of North Carolina at Chapel Hill, Computer Science(MS)
2008년 충남대학교 컴퓨터공학과(박사)
2018년 ~ 현 재 한국과학기술정보연구원 과학기술연구망센터 센터장
1998년 ~ 현 재 한국과학기술정보연구원 책임연구원
관심분야 : 미래인터넷, TCP 전송성능, 무선 IoT 네트워크, 양자암호통신 키관리시스템