

보안 중심 네트워크에서 빅데이터 전송방안 연구

석우진, 문정훈, 이상권

한국과학기술정보연구원

wjseok@kisti.re.kr, jhmoon@kisti.re.kr, sglee@kisti.re.kr

A Study on Big Data Transfer over Security-centric Network

Woojin Seok, Jeonghoon Moon, Sanggeun Lee

Korea Institute of Science and Technology Information

요약

본 논문은 보안중심의 공공기관에서 빅데이터 전송을 위한 성능개선 방안을 소개한다. 공공기관 특성 및 국가보안정책 강화 지침에 의거하여 구성된 보안중심 네트워크에서는 빅데이터 전송에 큰 장애가 발생하고 있다. 여러 계층의 보안장치 배치로 인하여, 데이터 크기가 페타규모로 증가하는 과학분야 빅데이터 전송이 어렵게 되었고, 이것을 해소하는 방안으로써 네트워크 설계 구조안을 제안한다. 이는 네트워크 전송 경로상의 보안장비를 우회하는 방안으로 경로 우회를 위한 매개서버를 두고 전송성능을 최적화를 적용하는 방안이다.

I. 서론

최근 과학기술 연구의 패러다임 변화는 융합 및 AI 연구 중심으로 변화하고 있어, 이를 위한 빅데이터 중심의 데이터 집약형 연구 활동 중심으로 진화하고 있다. 예로써, 천문 연구분야에서는 해외 빅데이터를 전송이 필요하며, SKA 데이터 총 4.9ZB/year(157TB/s), LSST 데이터 총 0.5EB/10year, SDSS 데이터 200GB/day 등의 빅데이터 전송을 요구하고 있다. 바이오/생명과학분야는 정부 정책적으로 빅데이터를 통합 관리를 진행하고 있으며, 부처별로 30PB, 524TB, 2PB, 3PB 등 페타 혹은 그 이상급의 전송을 요구하고 있다. 또한, 이러한 데이터를 활용하여 분석하는 방법에 있어서도, 계산 중심 분석방법에서 대용량 데이터 기반의 슈퍼컴퓨팅 및 인공지능 활용기법으로 진화하고 있다.

하지만, 상용망 인터넷 수준의 네트워크 구조와 복잡한 보안정책/장비로 인해 거대 과학데이터 전송에 큰 장애가 발생하고 있다. 과학분야 공공기관에서는 연구 데이터 보안을 위해 엄격한 보안체제와 보안장비를 거쳐야 하는 문제로 빅데이터 전송에 현실적인 어려움이 있다. A기관의 경우, 보안 정책에 의하면 8단계¹⁾의 엄격한 보안장치 구조를 통해 연구원 보안이 구현되어 있어, 안전성은 보장되지만 융합 및 AI 연구를 위한 빅데이터 전송은 거의 불가능한 수준이다.

II. 빅데이터 전송을 위한 ScienceDMZ 적용 제안

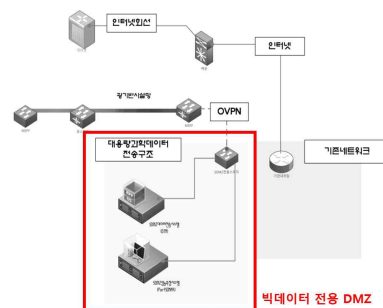
기존 네트워크 프로토콜 및 구조방안으로는 보안중심 체계에서의 빅데이터 전송의 한계가 발생하고 있으며, 이를 극복하기 위한 방안으로 ScienceDMZ 방안을 제안한다.

2010년 초에, 국제 연구망 커뮤니티에서 과학기술 전용의 글로벌 고속 및 안전한 전송을 위한 목적으로 '네트워크 서비스 인터페이스(NSI)' 기술이 있었으나, NSI는 국제 연구망 상호간 표준 인터페이스 구현의 복잡성을 이유로 실제 적용에 이르지 못하는 등 실제 글로벌 적용 단계로 진화

되지 못한 한계점이 발생하였다.

반면, ScienceDMZ 기술은 미에너지성 국가연구망(ESnet)에서 제안된 기술로서 글로벌 과학기술 분야에 적용이 용이하여 국제 연구망 커뮤니티에서 사실표준(defacto standard)으로서 역할이 가능한 고신뢰·고성능 빅데이터 전송 네트워크 기술이다. 전파망원경, 바이오게놈, 전자현미경, 입자가속기 등 과학기술 분야에서의 거대 실험연구 목적의 대형 실험장비에서 발생하는 과학빅데이터²⁾를 고속으로 전송하는 기술로써, 전송 구간 전용화, 전송 장비 최적화, 네트워크 프로토콜 최적화 등 많은 기술적인 요소를 포함하고 있음

핵융합연구소(NFRI)의 경우, KSTAR 대용량 연구데이터의 원활한 전송을 위해 과학빅데이터 전용 ScienceDMZ³⁾ 기반의 100Gbps 전용망을 구축하여, 10Gbps 대역폭상에서 120Mbps의 저속 전송 성능이었으나, ScienceDMZ 적용 후 9,000Mbps로 약 75배의 전송 성능이 개선되었다. 또한, 천문연구원(KASI)에서는 전파망원경(KVN)에서 생성된 과학빅데이터를 중앙상관센터로 전송, 분석하기 위하여 ScienceDMZ 기반의 전용망 기술을 적용하여 고속의 빅데이터 전송을 구현하였다.



[그림] 핵융합연구소 ScienceDMZ 설계안

2) 과학빅데이터: 일반적인 빅데이터는 작은 크기의 데이터들이 모여서 빅데이터를 형성.

과학분야 빅데이터는 하나의 데이터의 규모가 거대하여 코끼리데이터라고 명명하기도 함

3) ScienceDMZ(과학빅데이터 전용 전송체계): 과학 빅데이터 고속전송을 위한 네트워크 전용화/최적화 기술체계

1) 8단계 보안: 라우터, DDOS, 기본방화벽, 웹방화벽, 바이러스방화벽, 침입차단, 유출경로차단, 내부망연결 등

미국의 경우, 미국 DOE(에너지성)는 LBNL 산하 대형과학장비에서 발생하는 과학 빅데이터 전송에서 발생하는 액세스단의 저속화 문제(last one mile problem) 해결을 위하여 ScienceDMZ 기술을 적용하였으며, 해양대기국(NOAA)과 NERSC 전송의 경우, NOAA의 방화벽 뒤 FTP 서버 기반 전송 시 8~16Mbps의 저속 전송을 개선하여 ScienceDMZ 적용 후 약 200배의 처리량 증가 효과(3,160Mbps, 총 크기 239.5GB의 데이터를 10분 이내 전송 가능)가 발생하였다.

이러한 보안중심 네트워크에서의 빅데이터 전송체계를 완성하기 위하여 ScienceDMZ 기술 적용이외에 각 기관간의 데이터 접근을 위한 사용자인증 부분에서도 기술적 체계의 일원화가 추가로 필요하다. 즉 다수 연구자원에 파편화된 사용자인증 적용 시, 파편화된 사용자 인증체계를 적용하게 되며 이는 개별 데이터 접근에 필요한 계정 생성·관리로 인한 멀티도메인 간의 연구자원 통합활용 불가 등 다수의 문제점 발생한다. 이러한 문제점을 개선하기 위하여, 인증체계의 일원화 필요하여, 계정연합(Identity federation) 방식을 통하여 사용자 인증체계 표준화, 연합기반 통합인증 등 다양한 사용자 인증인가 기술의 출연(연) 적용으로 해결이 가능하다.

III. 결론

핵융합연구소(NFRI)는 KSTAR 장비를 통해 발생하는 핵융합에너지 관련 빅데이터를 미국의 ORNL(Oak ridge National Laboratory), PPPL(Princeton Plasma Physics Laboratory)과 카이스트, 포항공대, 서울대 등의 국내 파트너들과 공유를 위한 고속전송 요구되고 있다. 천문연구원의 SKA(호주), LSST(칠레), SDSS(미국) 데이터 및 화학연구원의 NoMaD(독일) 데이터 등 출연(연) 및 과기대와 공유를 위한 국내외 연계 고속 전송 요구되고 있다. 또한, SKA 데이터 총 4.9ZB/y(157TB/s), LSST 데이터 총 0.5EB/10y(12.6GB/1jpg), SDSS 데이터 200GB/day를 예상하고 있으며, 국내 바이오/생명과학분야의 각 부처별 발생 데이터 규모가 방대하여 전송에 소요되는 시간(수 개월 ~ 수 년)이 길어져 연구에 지장을 초래하고 있다.

이러한 국내 과학분야 공공기관들에서 지속적으로 요구되는 데이터는 크게 증가하고 있으며 더욱 요구되고 있는 상황이다. 또한 이러한 빅데이터들을 국내 각 공공기관들 간의 공유하기 위한 전송체계도 동시에 요구되고 있다. 보안중심의 각 공공기관에서 이러한 빅데이터를 전송하여 공유하는 체계로 지속적으로 변화되어야 한다. 이러한 빅데이터 전송체계를 바탕으로 슈퍼컴퓨터, 클라우드 등의 거대 분석도구와의 연계도 추가로 필요하게 될 것이다.

또한 빅데이터 및 이러한 거대 분석도구들의 사용자 접근체계 역시 각 공공기관들에 대한 일원화된 접근체계가 필요하며, 연합인증방식으로 가능할 것으로 판단된다. 빅데이터 전송체계인 ScienceDMZ와 연합인증의 접근체계를 통하여 빅데이터 중심의 연구 환경이 개선이 시급하며 또한 이를 위한 각 공공기관에서의 보안체계 중심에서의 빅데이터 전송체계 개선이 지속적으로 진행되어야 하겠다.

참 고 문 헌

- [1] Kettimuthu, R. (2018). "Transferring a Petabyte in a Day", International Journal of Future Generation Computer Systems, 88(1), 191-198.
- [2] Crichigno, J. (2019). "A Comprehensive Tutorial on Science DMZ", IEEE Communications Surveys & Tutorials, 21(2), 2041-2078.
- [3] Dart, E. Dart, (2014). "The ScienceDMZ: A network design pattern for data-intensive science," Scientific Programming, 22(2), 173-185.