Future Internet Forum

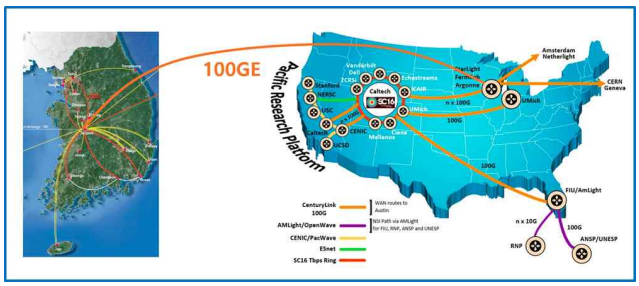
Newsletter
< 87호 >
2017.09.15

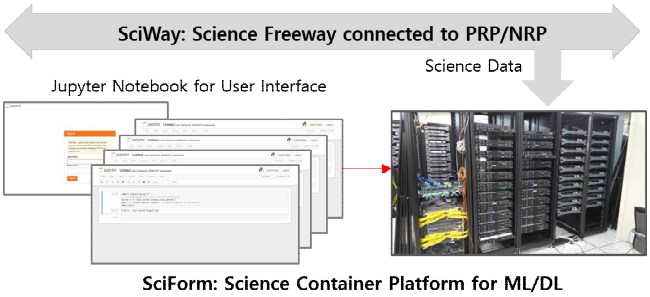
Science Platform for Bulk Data Transfer and ML Analysis

석우진, 문정훈, 홍원택, 곽재승(한국과학기술정보연구원 계산과학공학센터) / FIF Testbed WG

과학 데이터는 일반적인 빅데이터와 달리, 하나의 데이터가 수 테라(Tera) 혹은 수백 테라에 이르는 이른바 벌크 데이터 혹은 엘리펀트(Elephant) 데이터로 규정하고 있다[1]. 이러한 벌크데이터의 고성능 전송을 위하여, 미국 NSF에서는 슈퍼컴퓨터 혹은 클라우드 등의 주요 과학 데이터 출처인 Superfacility 사이트들의 고성능 연계를 위한 PRP(Pacific Research Platform) 프로젝트를 추진하고 있으며, 주요 5개 과학 분야(입자물리과학, 천체과학, 바이오/의과학, 지구과학, 가시화과학)의 데이터 전달과 분석을 목적으로 하고 있다. 최근 2017년 8월 PRP 공식미팅에서는 미국 전역 사이트(NRP: National Research Platform)의 확대를 발표하였다. KISTI는 2015년부터 국제 협력 기관으로 참여하고 있으며[2], PRP 핵심 참여기관인 미국 캘리포니아공대(Caltech)와 2016 슈퍼컴퓨팅 컨퍼런스(SC2016)에서 한국-미국 간의 9.5Gbps 성능을 시연한바 있다.

PRP 프로젝트에서는 Superfacility들 간의 고성능 전송을 위하여, DTN(Data Transfer Node)이라는 TCP 전송에 최적화된 고성능 서버를 활용하고 있으며, 현재 KISTI를 포함한 글로벌 24개 사이트에 PRP/NRP DTN이 배치되어 있다.





[그림] SC2016 시연: 9.5Gbps 전송성능(좌), KISTI SciForm and SciWay (우)

고성능 데이터의 전송의 연장으로, DTN으로 전송된 데이터들의 분석과 처리를 위하여, DTN과 연계된 슈퍼컴퓨터, 클라우드, 혹은 클러스터에서 과학데이터를 Seamless하게 처리하는 방안이 주목된다. 이상의 목적과 부합하는 파일럿 모델로써, KISTI에서는 컨테이너(Container) 플랫폼 기반의 DTN 연계방식의 데이터 처리 방안을 제시하고자 한다. 이는 최근 ML(Machine Learning)/DL(Deep Learning)을 활용한 분석 방법을 과학 분야에 적용하고자 하는 요구와 부합하기도 한다[3].

오픈 소스 SW인 오픈스택(Openstack)과 메소스(Mesos)를 기반으로 주피터 노트북(Jupyter Notebook)을 구동할 수 있는 컨테이너 기반의 플랫폼을 설계함으로써, 웹기반의 주피터 노트북과 텐서플로우(Tensorflow), 카페(Caffe)등의 ML/DL 프로그래밍 혹은 데이터 분석의 수행 작업(instance)들이 메소스 기반의 컨테이너에서 처리될 수 있도록 구축될 것이다. 컨테이너 기반의 작업수행을 위한 스케쥴러로는 마라톤(Marathon) 스케쥴러를 사용하며, 주피터 허브(Jupyter Hub)와의 연계를 위하여 마라톤 스포너(Spawner)가 사용된다.

과학 응용연구 분야의 분석 용도로 활용될 것이며, 특히 ML/DL 기반의 분석을 위한 컨테이너 이미지를 기본 제공함으로써 연구의 효율성 증진을 기대하고 있다. 컨테이너 이미지들의 저장을 위하여 오픈스택 스위프트(Swift)가 사용되어 다양한 컨테이너의 연구 환경 이미지들을 제공함으로써 연구결과의 공유 및 재현성을 향상시킬 것이다. 기존 악성코드 및 DDoS 공격 등의 보안 연구나 네트워크 프로토콜 연구를 위하여 테스트베드로 제공하던 베어메탈 클라우드(emulab)와 더불어, ML/DL을 활용한 과학 응용 분석용 실험환경을 제공하게 될 것이다.

참고문헌

[1] 석우진, 문정훈, 홍원택, 곽재승, “엑사바이트 과학빅데이터 전송을 위한 DTN기반 저장-후-전송 데이터 전송 방안”, 통신학회 10월호(예정)
[2] Jeonghoon Moon, Woojin Seok, Minsun Lee, “KISTI Participation for PRP”, PRP Kickoff Meeting, San Diego USA, Oct. 2015.
[3] Kento Aida, “Application-Centric Overlay Cloud Utilization Inter-Cloud”, APAN44 Meeting, Dalian China, Aug. 30 2017.