그리드 미들웨어와 WLCG

GSDC 데이터 그리드 컴퓨팅 스쿨

2016. 2. 1.

미들웨어?

- Condor
- 그리드 자원을 네트워크로 연결하고 계산 자원 사용, 저장 공간 관리 등

유럽 그리드 프로젝트

- European Data Grid (EDG)

  - 2000-2003

- Enabling Grids for E-sciencE (EGEE)

  - 2004-2010

- European Grid Infrastructure (EGI)

유럽 미들웨어 프로젝트

- EDG - LCG-2 미들웨어
- EGEE - gLite 미들웨어
  - EGEE-I, -II, -III
- EGI - EMI 미들웨어

gLite 미들웨어

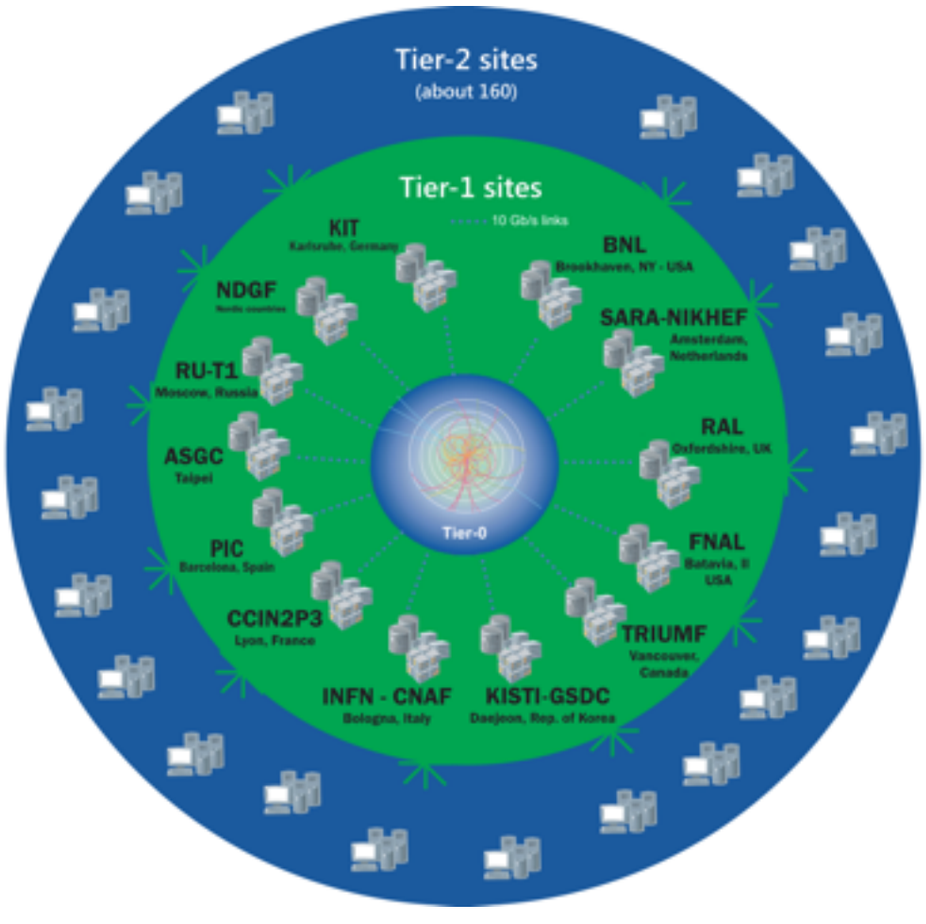
- 여러 그리드 프로젝트 산출물의 집합체
- DataGrid, DataTag, Globus, EGEE, EMI, WLCG 등
- 주로 WLCG/EGI 에서 많이 사용됨

WLCG (1)

- Worldwide LHC Computing Grid 프로젝트
- LHC 실험 데이터를 처리하기 위한 그리드 인프라 운영
- EGI와 인프라 운영에 있어서 공동 운영
- WLCG/EGI Infrastructure
  - + Open Science Grid (VDT 미들웨어)
  - + NorduGrid (ARC 미들웨어)

WLCG (2)

- wlcg.web.cern.ch
- 42개국
- 168개 컴퓨팅 센터 (T0, T1, T2)
- 일 평균 2백만개 작업 수행



gLite 미들웨어 구조

인증

- Grid Security Infrastructure (GSI) 기반
- 공개키 암호화
- X.509 인증서
- Secure Sockets Layer (SSL) 통신
- 권한 위임 (delegation)

그리드 자원 접근 제어 (1)

- VOMS를 사용하지 않는 경우
- 로컬 계정으로 mapping
  - grid-mapfile : 그리드 사용자와 로컬 계정 1:1 대응
  - 인증서의 Subject Name만 확인

그리드 자원 접근 제어 (2)

- VOMS를 사용하는 경우
- LCAS (Local Centre Authorization Service)
  - Subject Name과 VO 정보를 확인
- LCMAPS (Local Credential Mapping Service)
  - Pool 계정 기반으로 그룹 단위 권한 인증 가능
  - 그리드 인증서에 대한 유효성 검증
- SCAS (Site Central Authorization Service)
  - 사이트 접근 권한 관리를 로컬에서 중앙 서버로 집중
- gLExec
  - LCAS/LCMAPS와 함께 사용되며, 일차적으로 특정 리소스에 대해 획득한 권한을 다른 리소스 접근시에도 전달 가능
  - Pilot Job (Job Agent)이 우선적으로 작업 수행 권한을 획득 후 실제 수행될 작업이 나중에 들어오는 형태 “late binding of workload”

User Interface (UI)

- 그리드 접근 지점
- 개별 사용자 계정
- 개인 인증서
  - $HOME/.globus/usercert.pem
  - $HOME/.globus/userkey.pem
- 프록시 생성 (Proxy generation)
- 그리드 자원 정보 획득 : 계산 노드와 스토리지 정보, 데이터 경로 등
- 작업 제출과 작업 관리

Computing Element (CE) (1)

- 계산 자원을 총칭
  - 클러스터, 컴퓨팅 팜 등으로 불림
- Grid Gate (Gatekeeper)
  - 그리드와 클러스터간 인터페이스
- LRMS (Local Resource Management System)
  - 또는 배치시스템(Batch System)으로도 불림
  - 적절한 스케줄링을 통한 계산 작업 분배 관리
- Worker Nodes (WNs)
  - 실제 작업을 수행하는 컴퓨팅 노드

Computing Element (CE) (2)

- CE의 구성은 Grid Gate와 LRMS에 따라 다양함
- Grid Gate
  - LCG-CE/CREAM-CE (gLite)
  - OSG-CE (OSG), ARC-CE (ARC)
- LRMS
  - OpenPBS/PBS Pro, LSF, Torque/Maui, Condor, GE…

Computing Element (CE) (3)

- gLite 미들웨어의 CE 구성
  - Grid Gate : CREAM-CE
  - LRMS : Torque/Maui (SLURM, LSF, GridEngine)

Computing Element (CE) (4)

- CEId = <gg_hostname>:<port>/<gg_type>-<LRMS_type>-<batch_queue_name>
- 외부 관점에서는 1 CE = 1 큐 (배치시스템)
- 즉, 큐가 다르면 다른 CE로 인식
- 예)
  - ce101.cern.ch:2119/jobmanager-lcglsf-grid_alice
  - cmssrv25.fnal.gov:2119/condor-condor-cms
  - gridce0.pi.infn.it:8443/cream-lsf-cms4

Computing Element (CE) (5)

- 사용자 어플리케이션 제공
- 로컬 설치 및 관리, NFS로 공유
- CernVM-FS
  - Proxy 서버에 caching
  - 네트워크 파일 시스템 형태로 공유

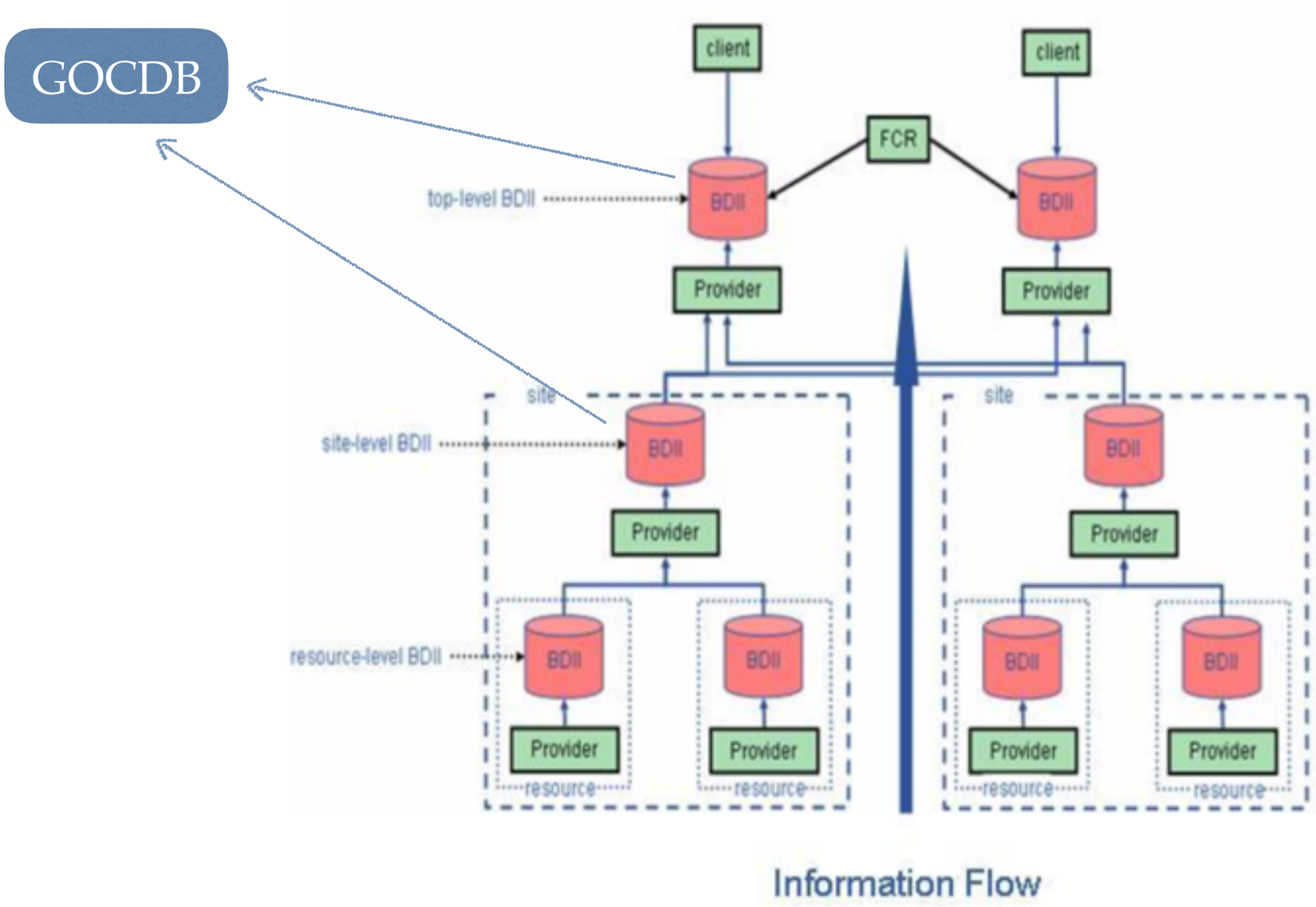
Storage Element (SE)

- SE의 역할은 사용자들에게 저장 장치에 대해 일관된 접근 인터페이스를 제공하는 것
- 한 개의 디스크 ~ 수 PB에 이르는 디스크 Array 또는 테잎 기반의 MSS (Mass Storage System)
- 데이터 접근 프로토콜
  - GSIFTP 또는 xrootd
- 다양한 저장 자원 관리 시스템
  - SRM (Storage Resource Manager)
  - DPM (Disk Pool Manager)
  - CASTOR (CERN Advanced STORage manager)
  - dCache
  - XRootD

Information Service (IS) (1)

- 그리드 자원 정보 제공
- 모니터링 또는 통계 자료로 활용
- OpenLDAP (Lightweight Directory Access Protocol) 기반
  - Entry - Attributes
  - DN (Distinguished Name) Entry의 고유이름
  - 하나의 Entry는 여러 Attribute를 가질 수 있음

Information Service (IS) (2)



WLCG Information Service 구조

데이터 관리 (1)

- 파일(File) 기반
- 그리드의 모든 파일은 하나 이상의 복제본(Replica)을 갖는다
- 복제 파일들 간 일관성을 유지 하기 위해 그리드 상에서 한번 생성된 파일은 수정 불가
- 오직 읽기/삭제만 가능
- 그리고 기본적으로 사용자는 파일이 그리드상에 어디에 존재하는지 알 필요가 없다
- 데이터 관리 시스템이 파일의 위치 확인과 접근을 제공

데이터 관리 (2)

- 그리드에서 하나의 파일은 4가지 이름으로 참조됨
- GUID (Grid Unique IDentifier)
- LFN (Logical File Name)
- PFN (Physical File Name)
  - SURL (Storage URL)
  - TURL (Transport URL)

데이터 관리 (3)

- GUID, LFN
  - 파일의 위치와 무관하게 파일 식별
  - 리눅스의 파일/장치 등의 UUID(MAC+시간)와 비슷한 고유 번호
    - guid:93bd7772a-b282-4332-a0c5-c79e99fc2e9c
    - lfn:/alice/cern.ch/data/2012/LHC12f/000197342/pass2/013/root_archive.zip
- PFN (SURL, TURL)
  - 파일과 그 복제본의 물리적 위치 식별
  - 접근 방법 제공
  - srm://<SE_hostname>:<port>/srm/managerv2?SFN=<path>
  - xrootd://xhs11.sdfarm.kr:3124/00/12345/… <local file path>

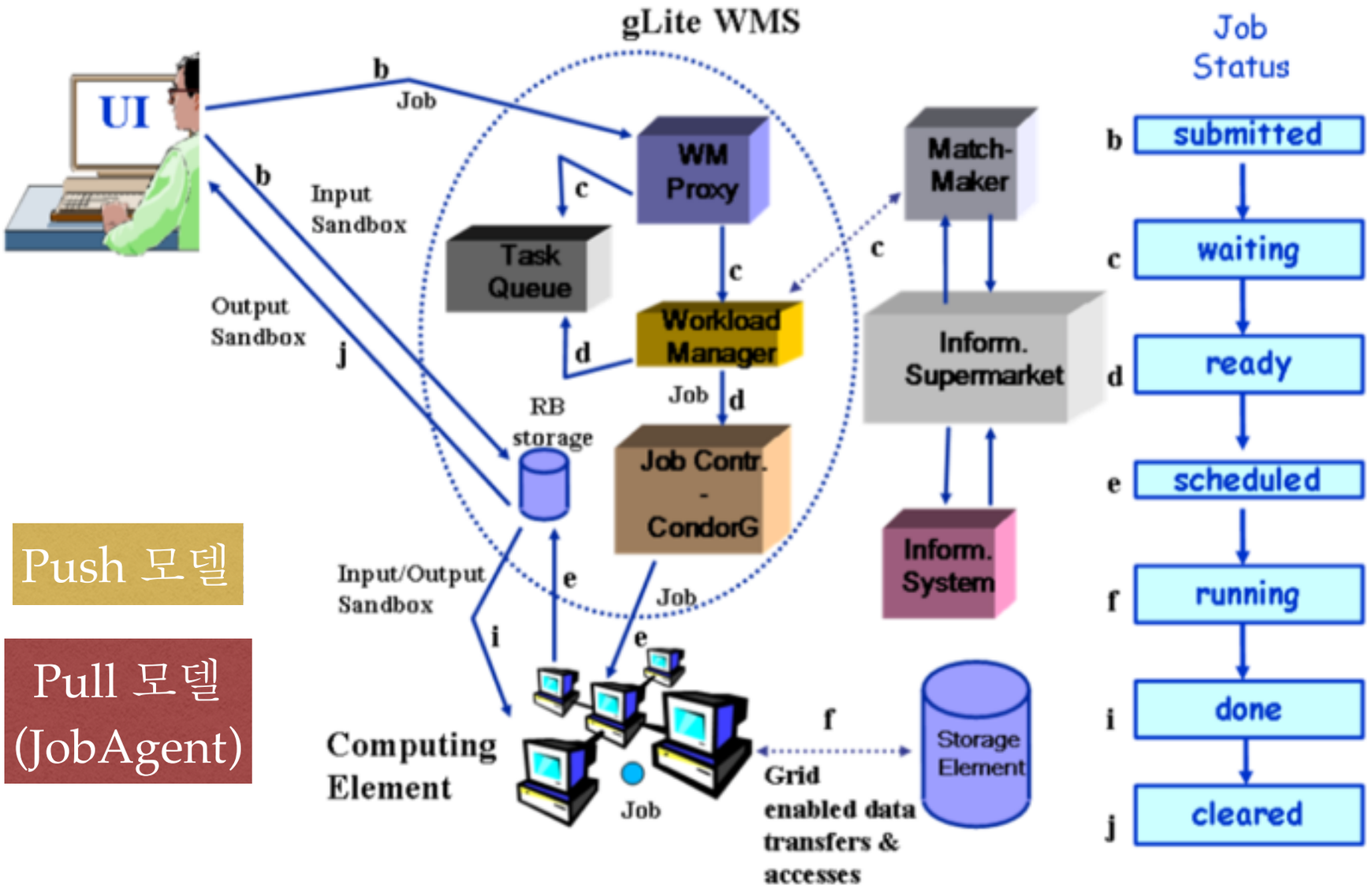
데이터 관리 (4)

- FC (File Catalogue)
  - GUID, LFN과 PFN (SURL, TURL)간 mapping
    - LFC (LCG File Catalogue)
- FTS (File Transfer Service)
  - SE간 데이터 전송 서비스
  - 서로 다른 VO간 데이터 전송도 가능

Workload Management

- WMS (Workload Management System)
- 사용자 작업을 접수, 가장 적절한 CE에 할당, 작업 현황 추적 및 결과 전송 역할
- 사용자 작업은 JDL(Job Description Language)로 작성
  - 실행 파일 이름과 파라미터들
  - Input 파일 리스트
  - 작업 수행에 필요한 여러 환경변수 및 조건들
- WMS는 사용자 JDL과 가장 적합한 CE를 추려낸 후 등급(Rank)을 부여
  - Match-making
  - 사용자가 분석하고자 하는 데이터와 가장 근접한 CE
  - 일반적으로 CE의 running job과 queued job 갯수에 관한 함수로 CE의 상태 정보를 표현
- LB (Logging and Bookkeeping service)가 WMS에 의해 관리되는 작업들을 추적

Job Flow



Push 모델

Pull 모델
(JobAgent)

WLCG 구축 사례

Batch System: Torque/MAUI
Grid Middleware: EMI

ALICE

isolated

perfSONAR

WLCG Mesh

LHCOPN

10G

8G

xrootd

Xrootd RDR

xrootd

apmon

ALICE::KISTI_GSDC::SE2

raw replication

Xrootd Servers

xrootd

apmon

/data

1000TB

SAN

EMC

squid

VOBOX

emi

alien

apmon

cvmfs

xrootd

Xrootd RDR

xrootd

apmon

ALICE::KISTI_GSDC::TAPE

xrootd

SE

emi

srm

gridftp

srm, gridftp

CE

emi

cream

torque

xrootd

Xrootd Servers

xrootd

apmon

frm

SAN

400TB

200TB

/data

/gpfs[01..08]

Hitachi USP

tsm

TSM Server

frm

siteBDII

emi

bdii

Worker Nodes

emi

glexec

cvmfs

torque

GPFS Servers

gpfs

tsm

SAN

1.5PB

IBM TS3500

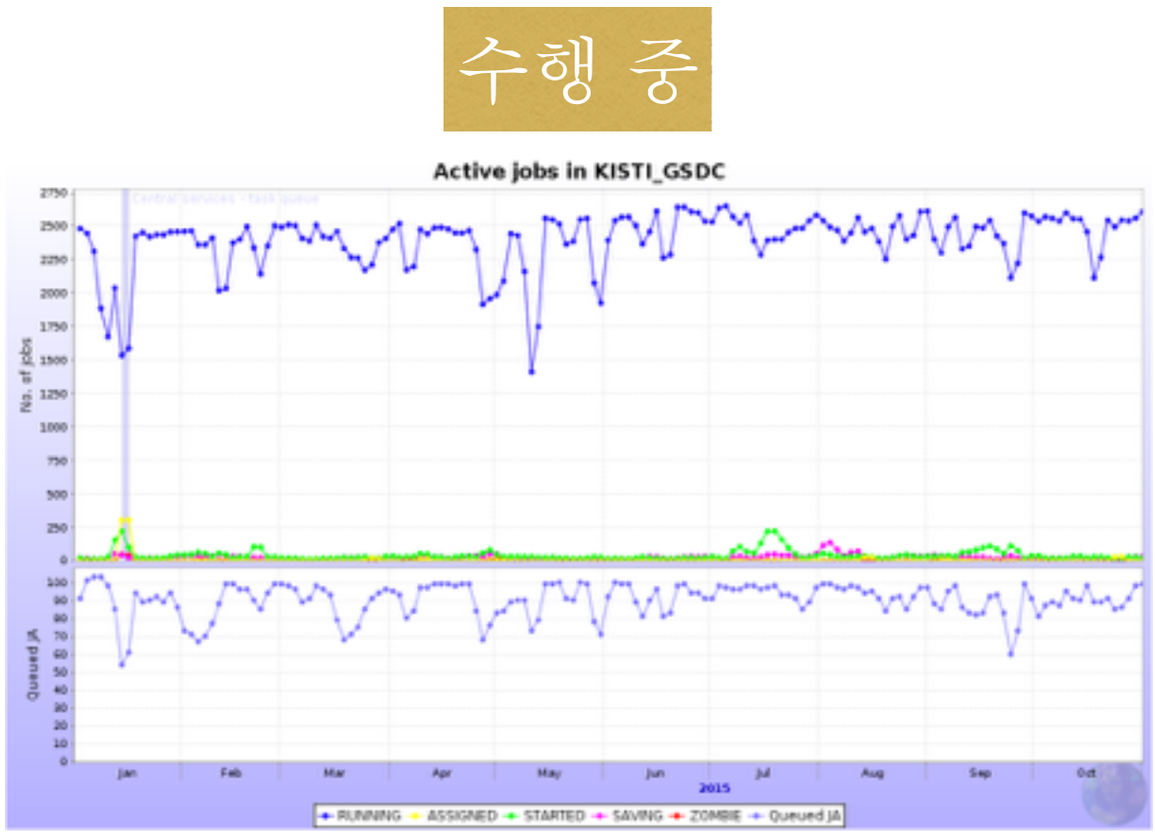
PS, Proxy, SE, BDII

Computing Resources (2688 core)

Storage Resources (3.1PB)

Job 모니터링

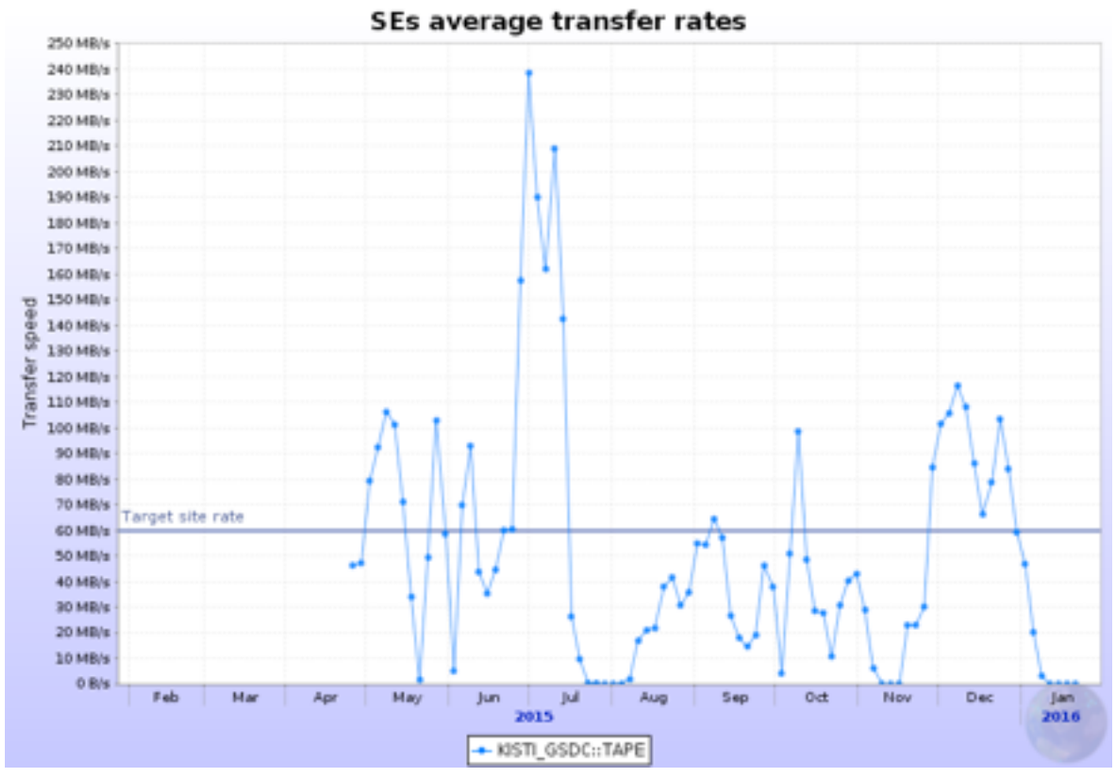
수행 중



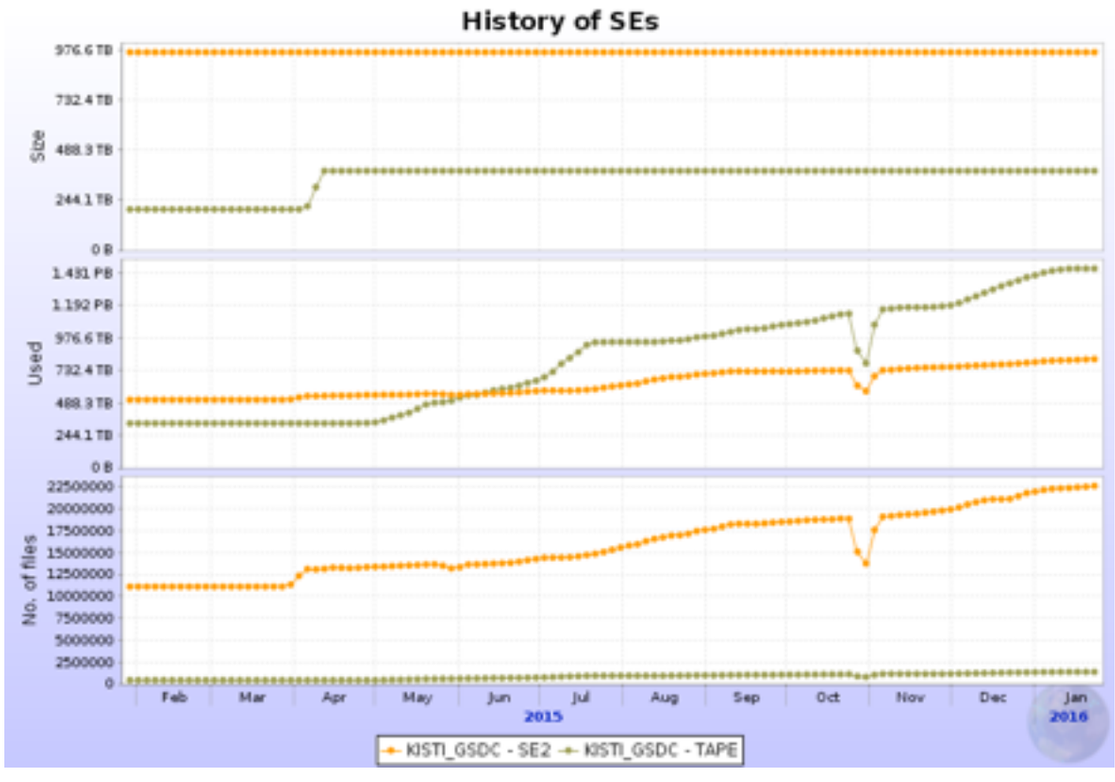
완료

SE 모니터링

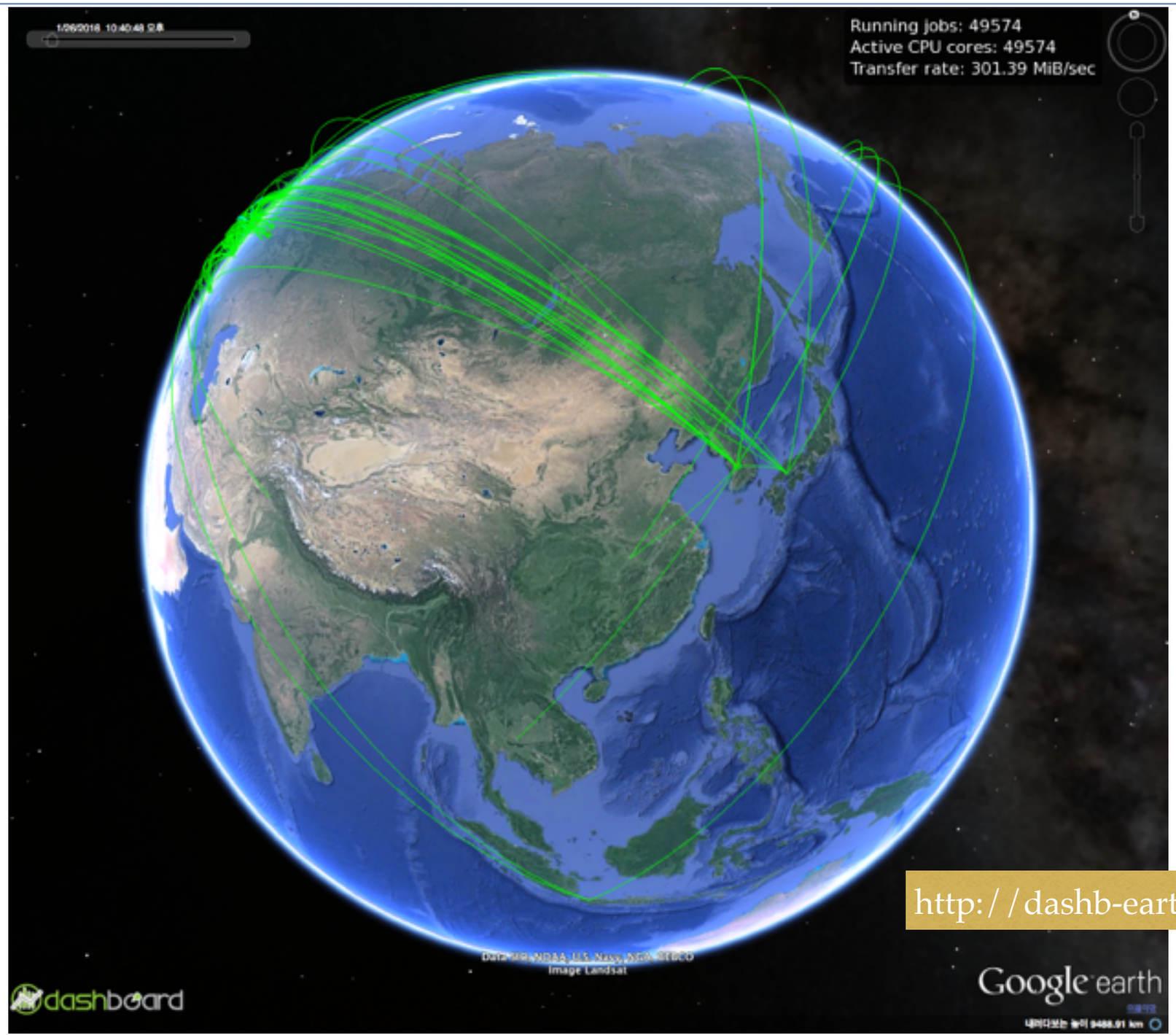
데이터 전송률



SE 사용률



데이터 전송 Live



http://dashb-earth.cern.ch/?vo=alice

요약

- gLite 미들웨어의 구성 요소는
  - User Interface (UI), Computing Element (CE), Storage Element (SE), Information Service (IS), File Catalogue (FC), Workload Management System (WMS)
- 사용자 작업 흐름은
  - 사용자 인증서 발급 -> UI 접속 -> 프록시 생성 -> WMS에 작업 제출 -> Match-making (데이터 위치 확인) -> CE에 작업 제출 -> CE에서 작업 수행 -> 결과파일 WMS로 전송 (대용량의 경우 SE에 저장) -> 사용자 결과 확인

참고문헌

- “gLite 3.2 User Guid Manuals Series”, CERN-LCG-GDEIS-722398

- “DataGrid Project Status”, Fabrizio Gagliardi

- wlcg.web.cern.ch

- alimonitor.cern.ch