KiSTi
www.kisti.re.kr
한국과학기술정보연구원
Korea Institute of Science and Technology Information

데이터 분산 관리 시스템

XRootD

안상언
GSDC

GSDC2016, KIRD

2016.12.28

목표

- XRootD기반의 데이터 분산 관리를 위한 클러스터 구축

- 장점
  - 공개 SW이기 때문에 쉽게 구할 수 있음
  - 간편한 설치 및 설정
  - 단순한 구조로 인해 관리 용이

- 단점
  - 파일 단위로 저장하기 때문에 데이터 안전성 보장의 어려움 및 낮은 성능

XRootD는 무엇인가?

- 스탠퍼드 선형 가속기 연구소(SLAC)에서 BaBar 실험을 위해 개발

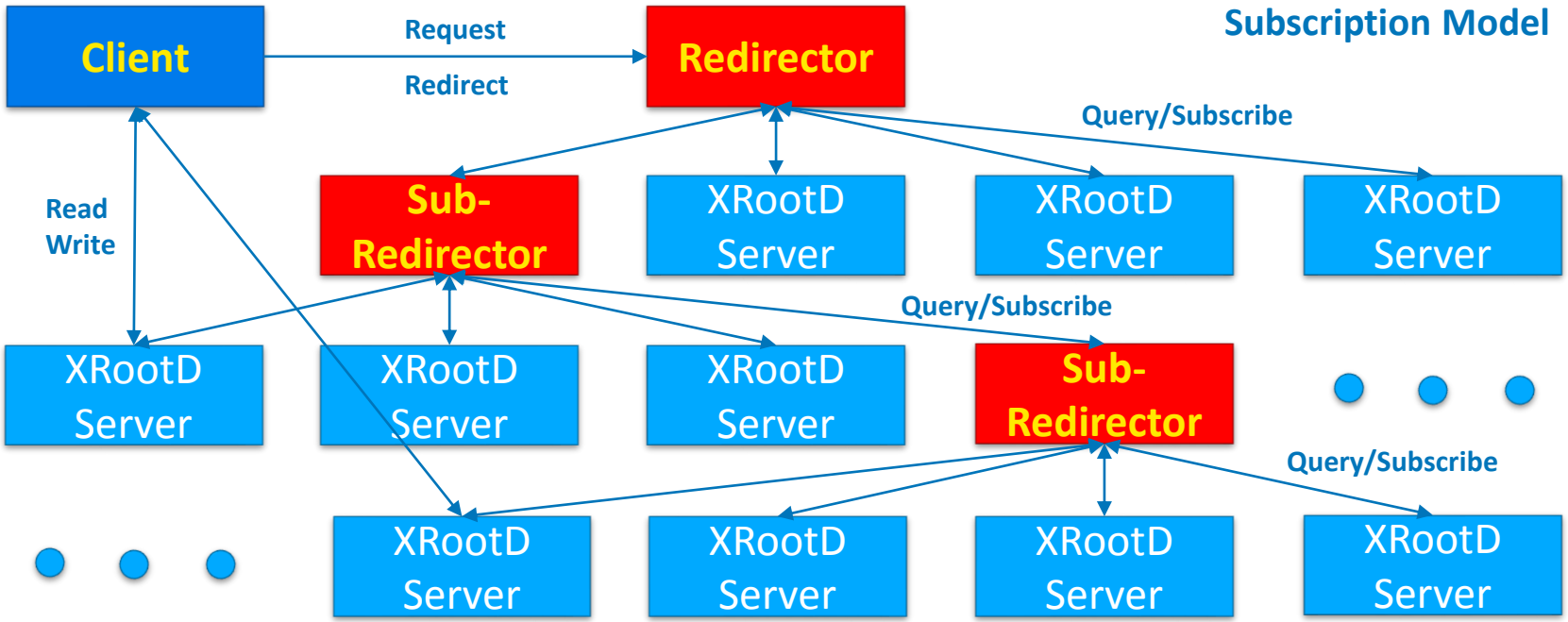
- eXtended Root Daemon
  - 기존 RootD를 대체
  - ROOT라는 HENP 분야에서 많이 활용되는 Analysis Framework의 파일(.root)을 관리하기 위해 개발
  - 파일 타입에 대한 제약이 없기 때문에 일반적인 데이터 관리 시스템으로 활용

- XRootD는 xrootd 서버와 Objectivity/DB 서버의 Load Balancing을 위해 개발된 olbd 서버로 구성
  - Open Load Balancing Daemon
  - 현재는 cmsd (Cluster Management Service Daemon)

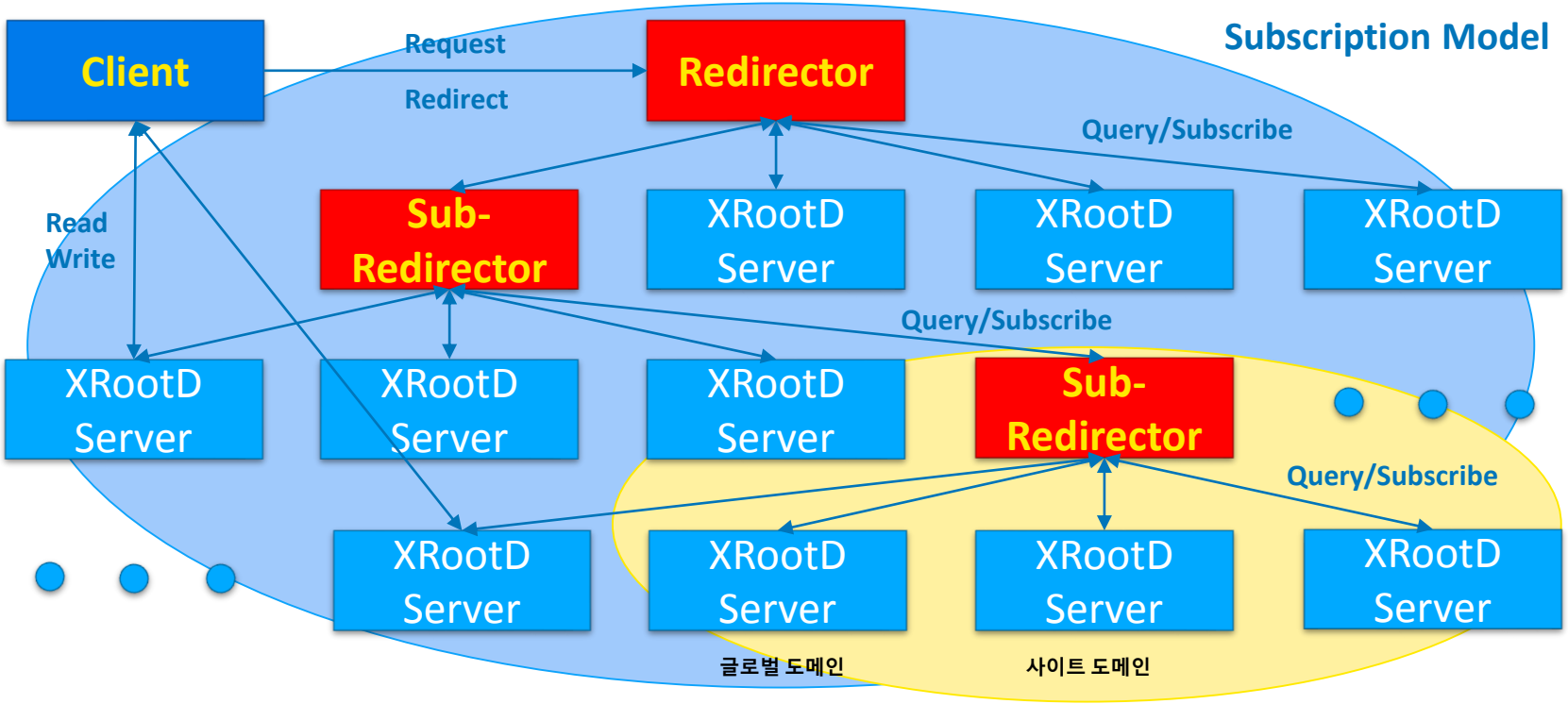
XRootD 현황

- CERN LHC 가속기 실험 등 HENP 분야에서 많이 활용
  - 대용량 데이터 분산 처리 목적
- ALICE 실험에서는 실험 초기 디자인 단계에서 글로벌 스토리지 시스템으로 XRootD를 도입
- 현재 ATLAS와 CMS 실험에서도 XRootD를 도입 중
- 특히, FRM(File Redundancy Manager)을 통한 Storage Federation 프로젝트를 진행 중
  - ATLAS FAX, CMS AAA

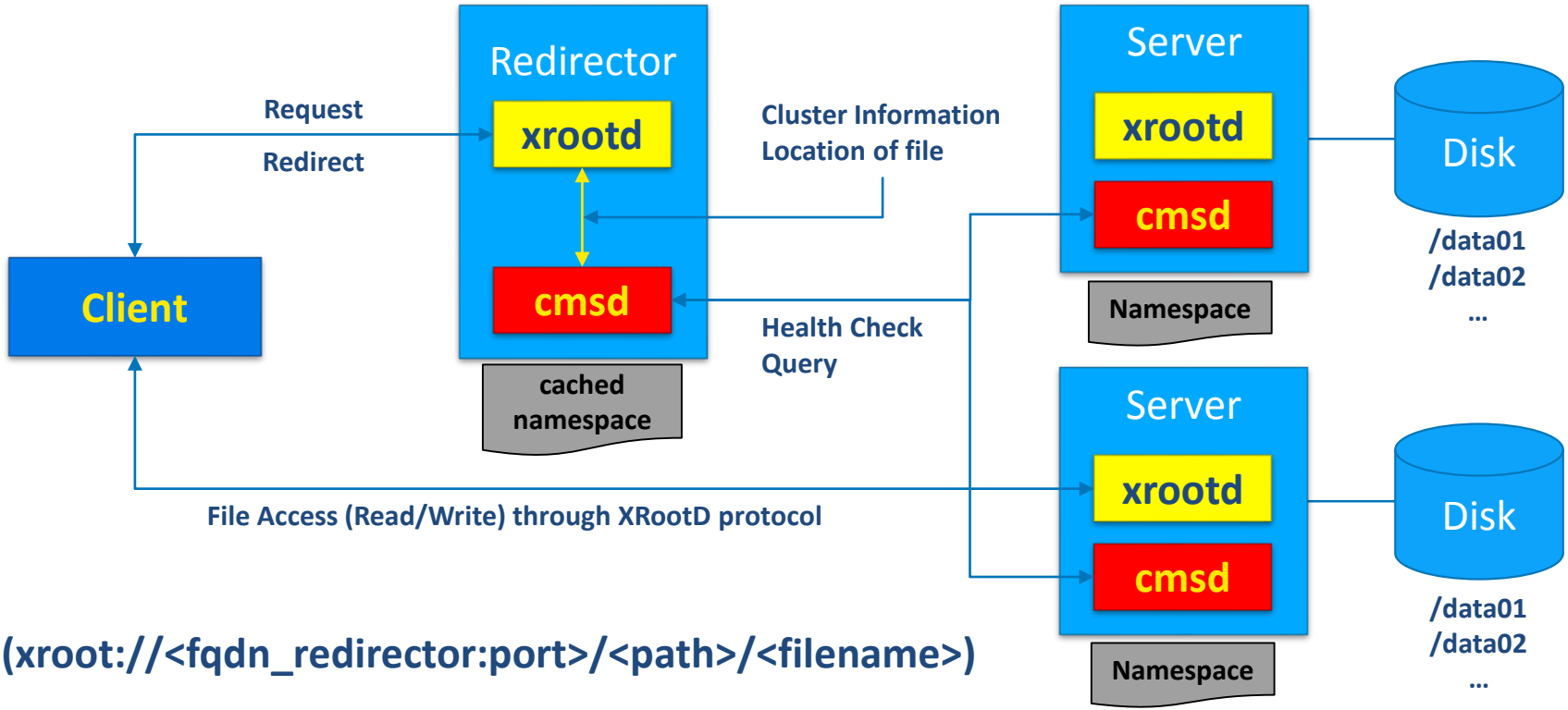
XRootD 기본 구조


Client
Request
Redirect
Redirector
Subscription Model
Query/Subscribe
Read
Write
Sub-
Redirector
XRootD
Server
XRootD
Server
XRootD
Server
Query/Subscribe
XRootD
Server
XRootD
Server
XRootD
Server
Sub-
Redirector
Query/Subscribe
XRootD
Server
XRootD
Server
XRootD
Server
XRootD
Server

XRootD 기본 구조

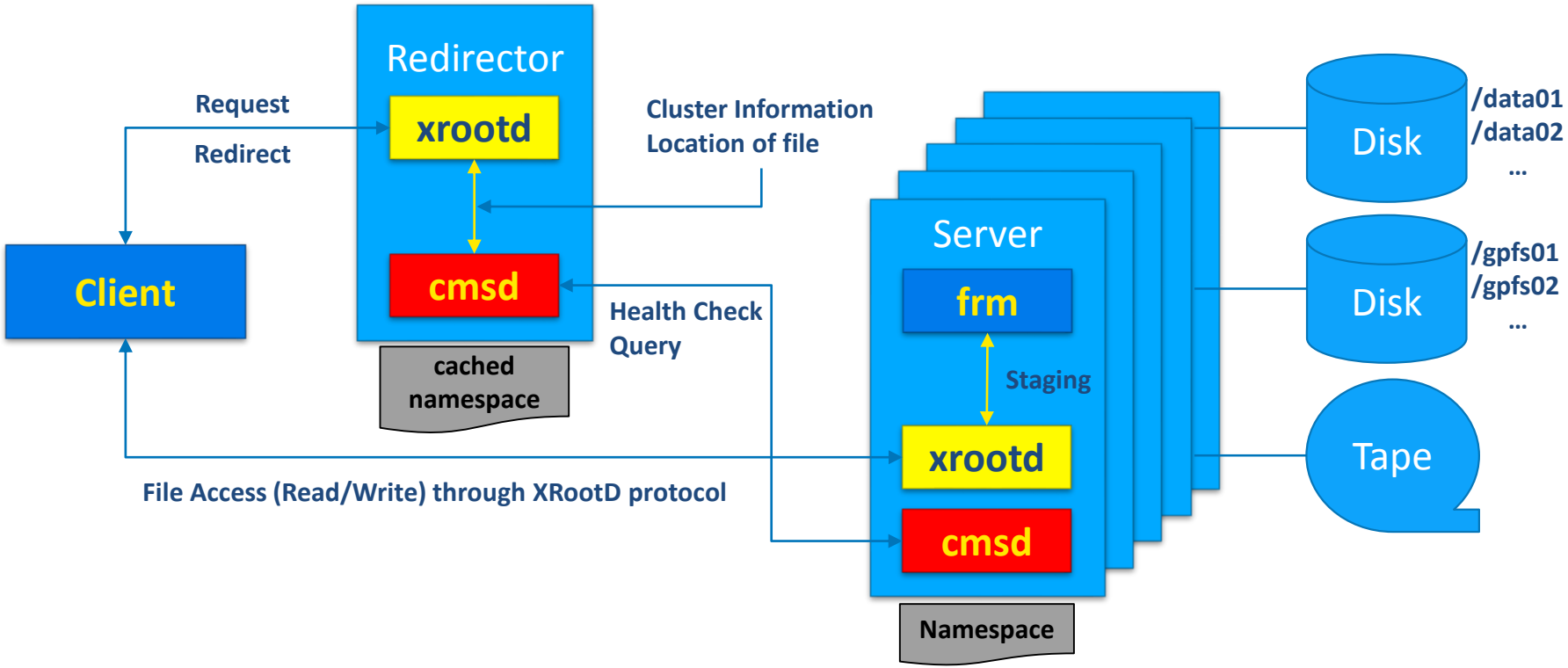

Client
Request
Redirect
Redirector
Subscription Model
Query/Subscribe
Read
Write
Sub-
Redirector
XRootD
Server
XRootD
Server
XRootD
Server
Query/Subscribe
XRootD
Server
XRootD
Server
XRootD
Server
Sub-
Redirector
Query/Subscribe
XRootD
Server
XRootD
Server
XRootD
Server
XRootD
Server
글로벌 도메인
사이트 도메인

XRootD 상세 구조 1

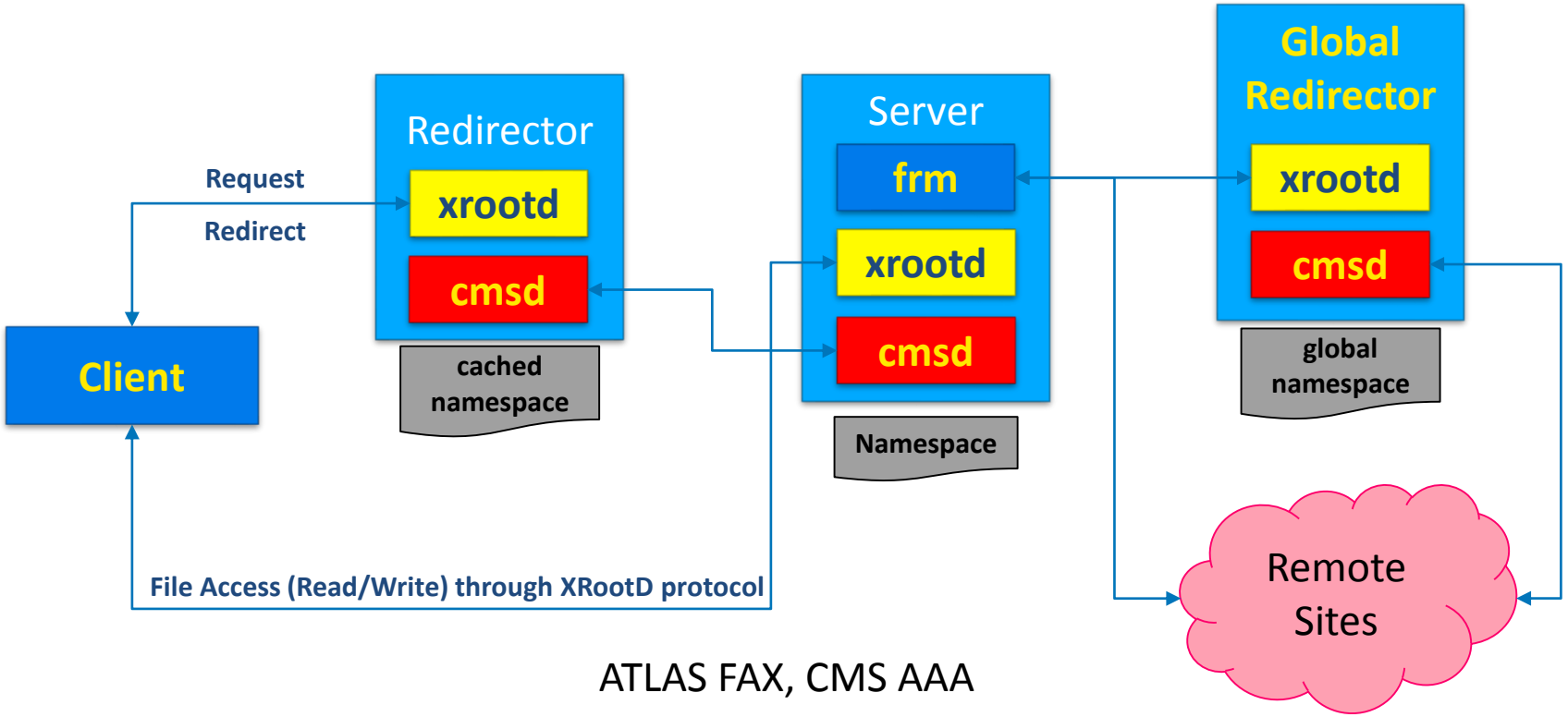


(xroot://<fqdn_redirector:port>/<path>/<filename>)

XRootD 상세 구조 2



XRootD 기반 Storage Federation



XRootD 특징 1

- 데이터 분산 관리 시스템
  - 대용량의 데이터 관리 및 분산 처리 목적
  - 파일 단위 입출력: 낮은 성능, 파일 안전성 취약
  - 병렬 파일시스템으로 보완 가능(로컬)
    - gpfs, lustre, glusterfs, ceph, hdfs 등

병렬파일시스템
- 파일을 Object 형태로 다중노드에 분산 저장
- 노드간 상호 백업을 통한 데이터 안전성 확보
- 여러 클라이언트에서 동시 입출력 가능
- 고성능 컴퓨팅을 목적으로 활용
- 사이트 도메인 내에서 효과적

- Subscription 모델
  - XRootD 서버 설정시 Redirector에 대한 정보만 필요
  - 하나의 설정 파일을 공통으로 사용 가능
  - XRootD 서버 확장 또는 축소 용이 (클러스터 shutdown 필요없음)

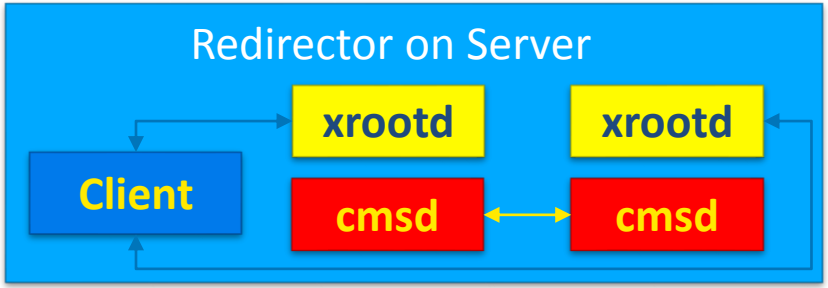
XRootD 특징 2

- Namespace 형태의 메타데이터 관리
  - XRootD 서버별 Namespace 관리
  - Redirector에 Caching, Query 시간 감소

- 간단한 구조, 관리 용이, Scalable

XRootD 구축 시나리오 1

- All-in-One 모드
  - Redirector와 Server를 한 노드에
  - 최소 설치, 최소 구성, 테스트 용도



- Normal 모드
  - Redirector와 Server 노드 분리
  - 가장 일반적인 구성
  - Server를 가능할 경우 추가 확장

XRootD 구축 시나리오 2

- 다중 계층 모드
  - Sub-redirector를 구성하여 Main-redirector의 Load 분산
    - Server 노드의 수 O(100)
  - 이중 (또는 다중) Sub-redirector를 통해 Fault-tolerant 구성

- 그리드 모드
  - 글로벌 Redirector를 통해 공동의 데이터 분산 관리 시스템 구성
  - 기관별 보유 데이터를 공동 활용 또는 상호 백업

XRootD 파일시스템

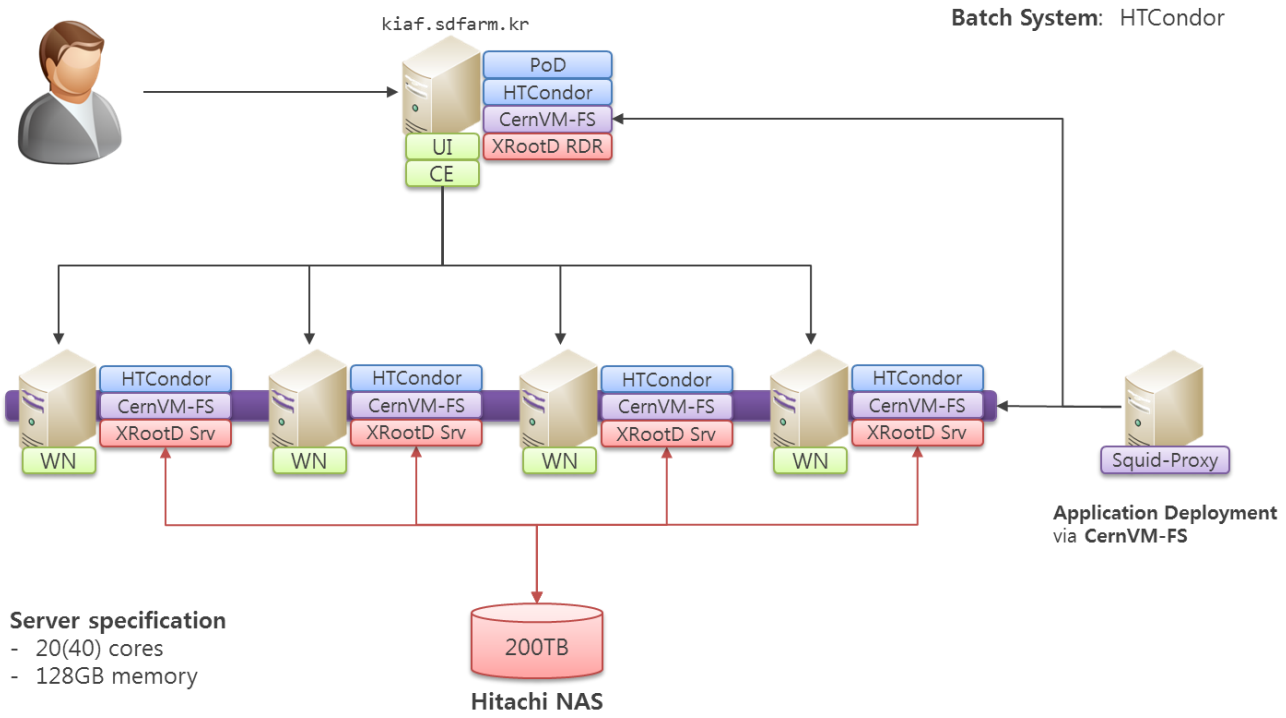
- XRootD를 위한 POSIX 파일시스템
  - XRootD로 구축한 스토리지를 파일시스템으로 마운트 (FUSE 기반)
  - cd, ls, rm 등 POSIX 명령 사용 가능
    - Portable Operating System Interface: 운영체제간 공통 API를 위한 표준

- 사용자 어플리케이션에서 XRootD에 대한 인지 없이 데이터 I/O 가능
  - 예) HTCondor 클러스터로 Job 처리시 XRootD로부터 데이터 입/출력

XRootD 활용 Analysis Facility 구축

- HTCondor + XRootD
- 단일 클러스터로 데이터 관리 및 분석 작업 처리까지

예제: KISTI Analysis Facility



References

- XRootD Webpage
  - http://www.xrootd.org
- XRootD Documentation
  - http://www.xrootd.org/docs.html
- FUSE
  - https://github.com/libfuse/libfuse

LAB III & IV

LAB III: XRootD 클러스터 구축

- HTCondor 배치시스템이 구축된 클러스터에 XRootD 클러스터 구축
- xrootdfs를 이용 워커노드에 파일시스템 마운트
- 데이터 샘플을 XRootD 클러스터에 전송
  - xrootdfs로 확인

시나리오

- 데이터 분산 관리 시스템을 구축해야만 하는 상황
  - 교수님, Post-Doc, 박사/석사 선배의 요구
  - 타이트한 일정 (내일? 3일 뒤? …)
- 주어진 조건
  - 매뉴얼 또는 매뉴얼 URL
- 어디서부터 시작해야 할까?

요약

- 패키지 배포 저장소 (repository) 설치
- 패키지 설치
- 설정
- 서비스 시작
- 트러블슈팅

매뉴얼

- URL : http://www.xrootd.org/docs.html
- xrootd, cmsd, frm 등 데몬 설정 커맨드 상세 설명
- 패키지 다운로드
- 설치가이드?

xrootd-clustered.cfg

```
all.export /data
set xrdr=(fqdn of xrootd redirector)
all.manager $(xrdr) 3121
if $(xrdr)
      all.role manager
else
      all.role server
      cms.space min 2g 5g
fi
```

xrootdfs

```
sudo mkdir /data
sudo chown xrootd:xrootd /data
sudo mkdir /xrootdfs
sudo chown xrootd:xrootd /xrootdfs

sudo xrootdfs -o rdr=xroot://<xrdr>:1094//data,uid=xrootd /xrootdfs

sudo umount /mnt
sudo mkfs –t xfs /dev/vdb
sudo mount /dev/vdb /mnt
sudo chown xrootd:xrootd /mnt
```

xrdcp

xrdcp testfile xroot://<xrdr>:1094//mnt/testfile

xrdcp xroot://<xrdr>:1094//mnt/testfile testfile2

xrdcp xroot://<xrdr>:1094//mnt/testfile
xroot://<xrdr>:1094//mnt/testfile2

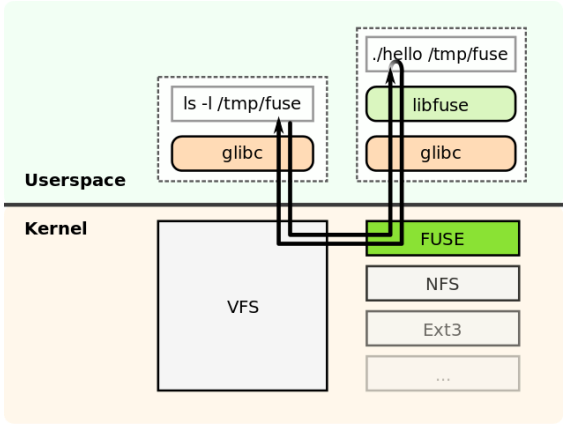
LAB IV: Analysis Facility 구축

- HTCondor & XRootD 클러스터에 사용자 어플리케이션 배포
- Analysis 예제를 HTCondor에서 수행 후 결과 확인

BACKUP

FUSE

- Filesystem in Userspace
- 커널 수정없이 파일시스템 생성
- Linux 2.6.14부터 FUSE 모듈 기본 탑재



<그림=Wikipedia>

- glibc: read(), write() -> sys_read (system-call)
- VFS: sys_read -> vfs_read (각 파일시스템의 read operation 수행)
  - 여럿 파일시스템의 연산을 호출할 수 있는 인터페이스 제공
  - struct file_operations에 등록된 read(특정 파일시스템에서 제공) 호출
- FUSE의 경우, FUSE에서 제공하는 file_operations의 read 호출