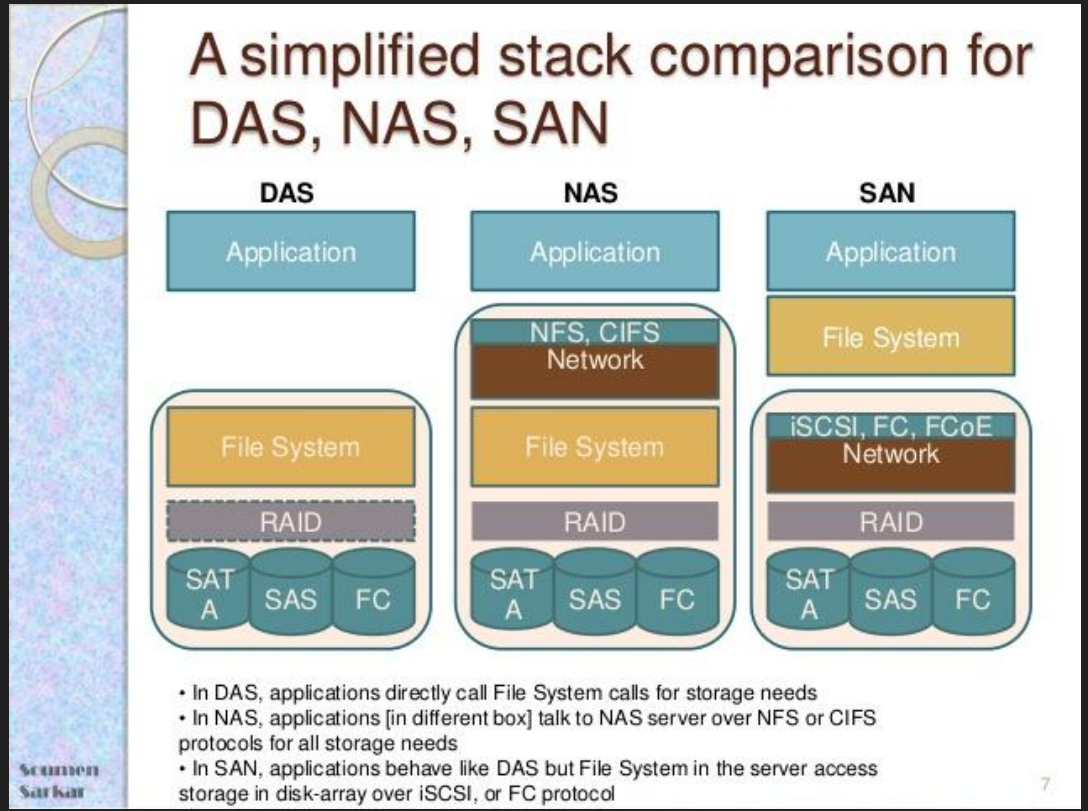
# Storage and then

# Storage

# JBOD(Just a Bunch Of Disk)

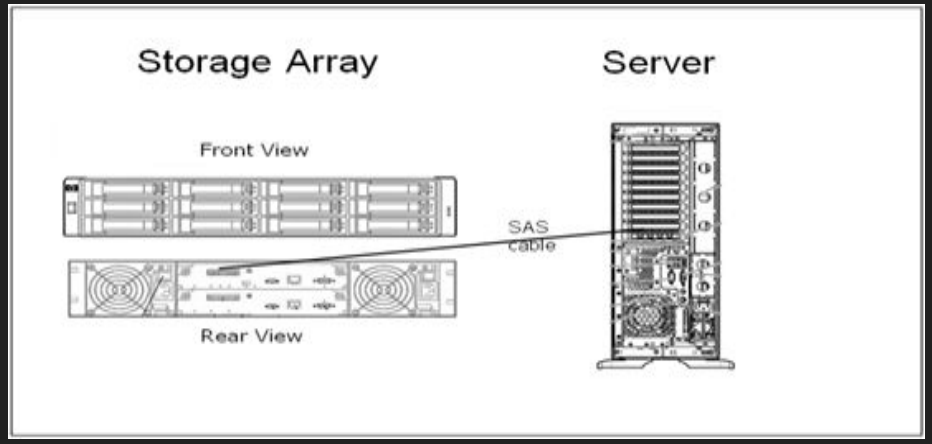
# DAS NAS SAN

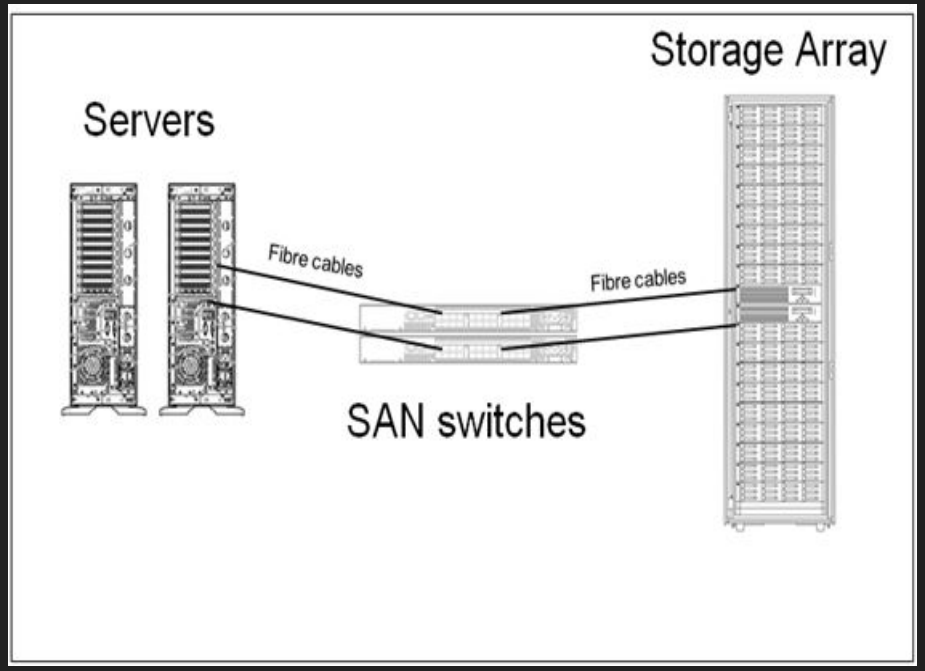
- Direct Attached Storage
- Network Attached Storage
  - NFS, CIFS
- Storage Area Network
  - FC, Ethernet
  - ISCSI
  - HBA



# DAS & SAN

- LUN
- Zone
  - WWN
- SAN Switch
  - Brocade



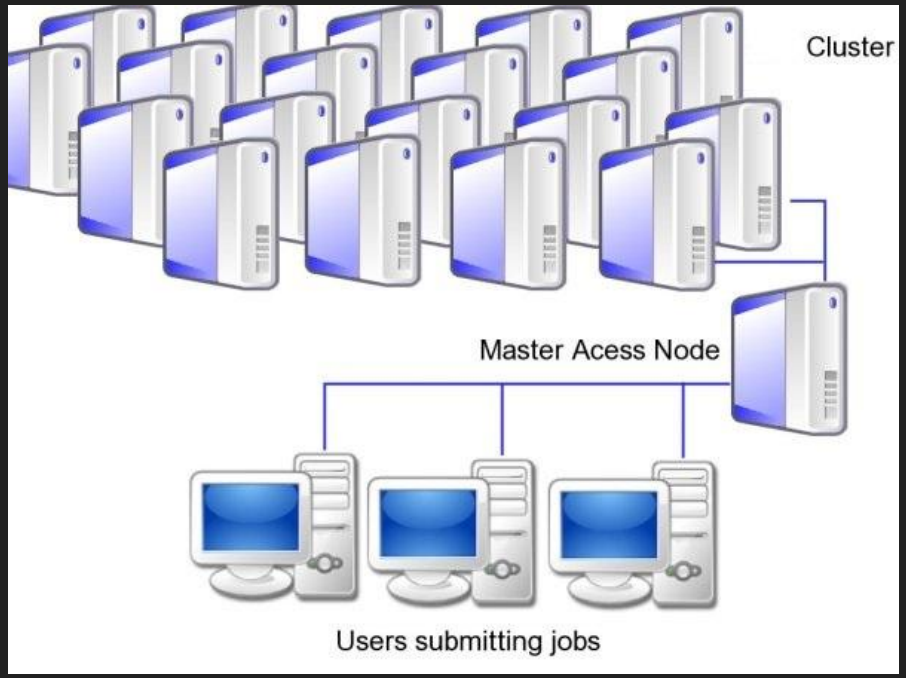
# Clustered File System

- Shared-disk file system
  - Veritas Cluster File System (VCFS)
  - Microsoft Cluster Shared Volumes (CSV)
  - Oracle Cluster File System (OCFS)
  - Redhat GFS …
  - 비싼것들
- Distributed file system
  - HDFS
  - Ceph
  - GlusterFS
  - Windows Distributed File System(DFS)

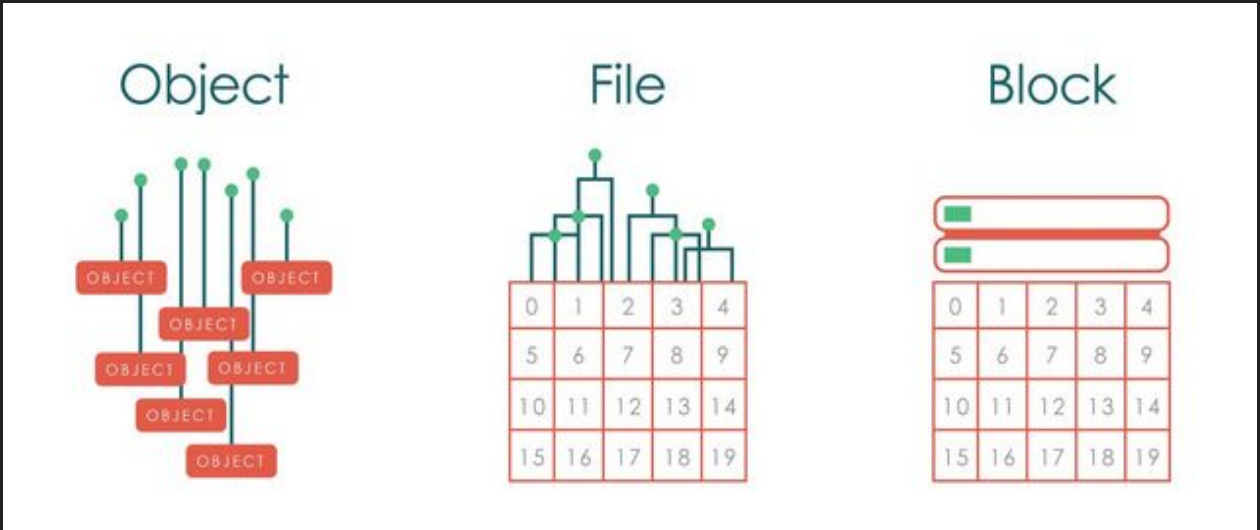
# Distributed File System

# Distributed File System

- 기존 형태로는 확장, 동기화, 가용성이 떨어짐
- 서버 여러대를 하나처럼...
- 객체 기반 파일시스템
  - Object Based File System
- Meta와 Data 분리
  - Meta - 파일명, 크기, path, 접근시간, ACL
  - Data - 실 data
- Replica!!!

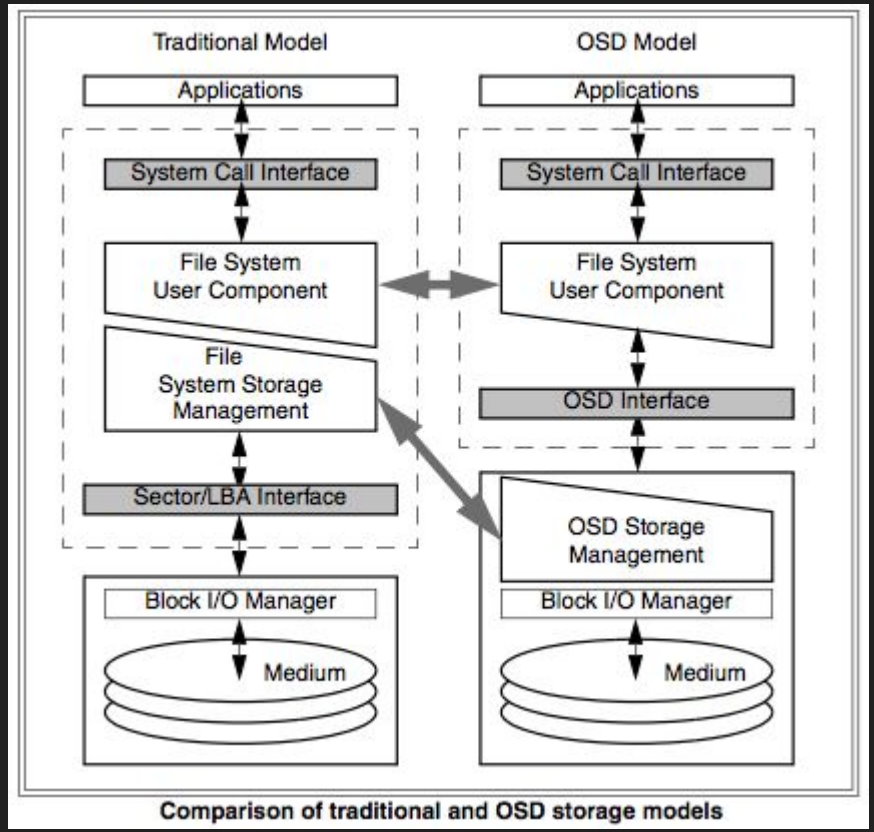


# Object File System



# 객체 기반 스토리지

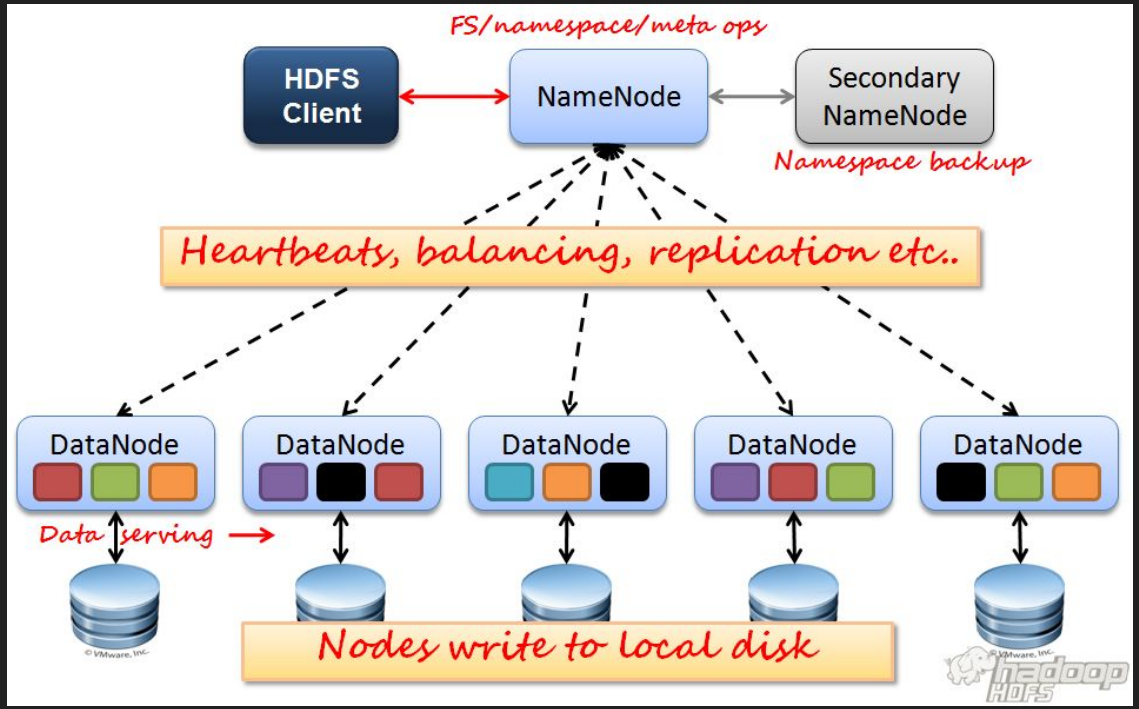
1. 파일시스템
   a. Kernel
   b. FUSE(File system in User SpacE)
2. 메타데이터 서버
   a. 클라이언트의 파일관련 요청조정
   b. 인증 및 권한 관리
   c. 객체 스토리지 상태 모니터링 관리
   d. Cache coherency 관리 (분산 lock 등)
   e. 용량 관리
3. 네트워크
   a. 프로토콜 - RPC, ISCSI, RDMA
   b. 물리 장치 - Ethernet, FC, Infiniband



Comparison of traditional and OSD storage models

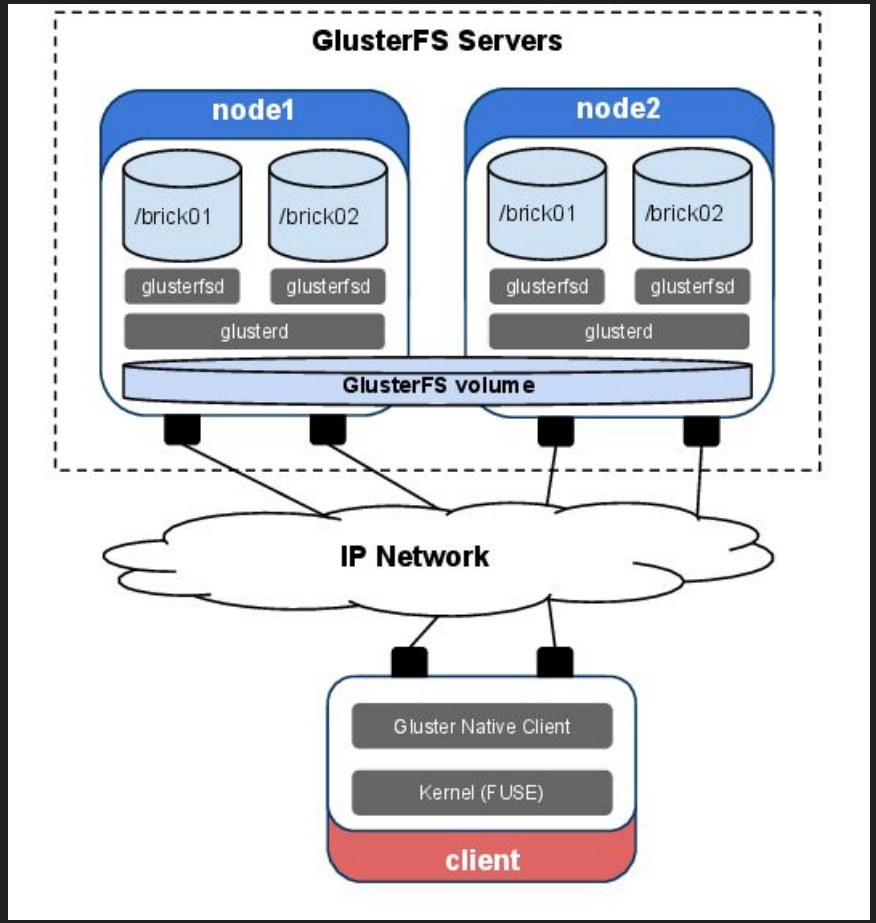
# HDFS


- Google File System 클론
- 구성
  - NameNode(메타서버)
  - Secondary
  - Master & Slave
  - DataNode
- 64MB block chunk
  - HDFS 2.0 - 128MB
- random access는 불가



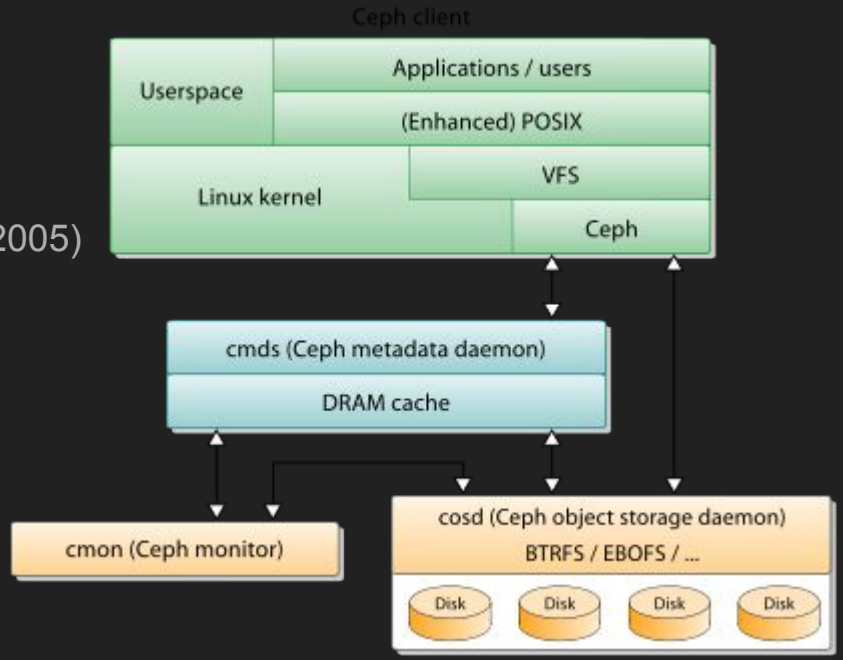


- No MDS
- DHT(동적 해시테이블 알고리즘)
  - Consistent Hashing Table
- Brick
- FUSE, NFS, CIFS
- GlusterFS Volume
- 복제
  - 파일기반
  - 주브릭과 복제브릭에서 동시에
  - 실패시 일단 놔둔다
  - IO요청시 싱크





- 2007. Sage Weil 의 논문에서 시작
- Linux kernel 2.6.34 포함
- 주요기술
  - RADOS – distributed object storage cluster (2005)
  - EBOFS – local object storage (2004/2006)
  - CRUSH – hashing for the real world (2005)
  - Paxos monitors – cluster consensus (2006)



# 메타데이터 관리

- Dynamic subtree
  - 알지는 못함
- Static subtree 모델보다 느리지만 대규모 서비스에서 변경, 장애에 대해 더 나은 방식

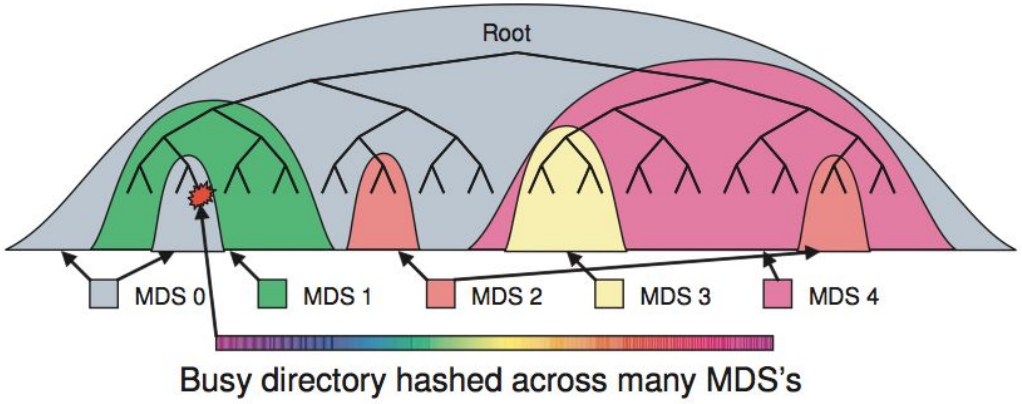
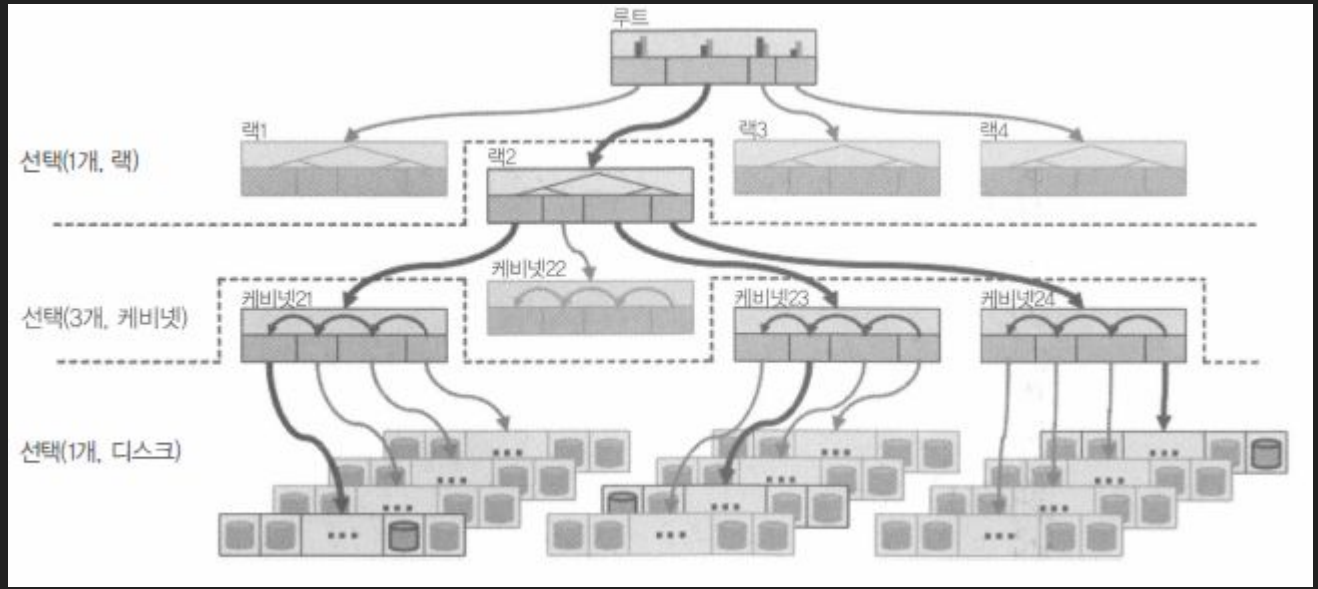



Figure 2: Ceph dynamically maps subtrees of the directory hierarchy to metadata servers based on the current workload. Individual directories are hashed across multiple nodes only when they become hot spots.

# 복제관리

- CRUSH(Controlled Replication Under Scalable Hashing) 알고리즘
  - 이런걸 쓴다카더라.



# References


- 『실전 클라우드 인프라 구축 기술』
- Naver D2 - 어떤 분산 파일 시스템을 사용해야 하는가?