클라우드 스토리지 서비스 N드라이브

NHN Business Platform
저장시스템개발팀/전성원

NHN BUSINESS PLATFORM

목차

N드라이브 서비스를 위해 풀어야 할 숙제들

NHN BUSINESS PLATFORM

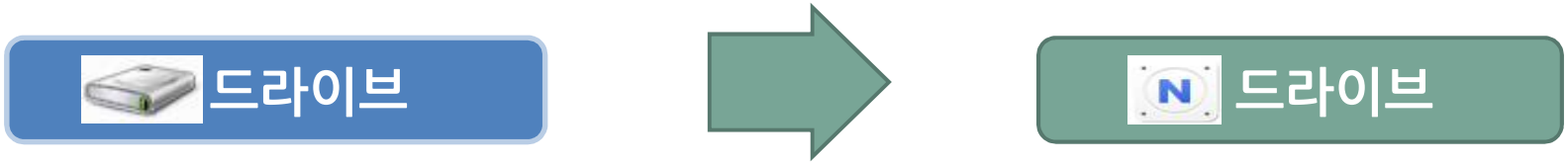
N드라이브 서비스

공지사항

- [이벤트] N드라이브 탐색기 동기화 클로즈베... ›
- [개선] N드라이브 서비스 개편 안내 (국내... ›
- [안내] 2/24 기본 접속 버전 변경 안내 ›
- [오픈] 네이버톡 서비스 웹/PC/모바일용 ... ›

N드라이브 : 규모

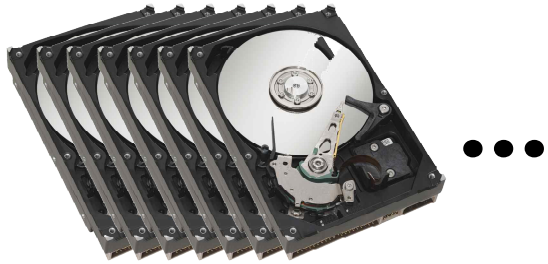
● 30GB



- 1.5MB 고화질 사진 20,480 장
- 6MB 고음질(192k) mp3 음악 5120 곡
- 700MB 동영상 파일 44 편

● 700만 사용자

- 사용자 x 용량 = 7,000,000 x 30GB =



N드라이브 : 사업 측면 당면한 문제들...

1. 모든 사용자에게 준다고 ??
2. 얼마나 쓸 지 모른다고 ??
3. 얼마나 빨리 증가할 지 예측할 수 없다고 ??
4. 어떤 경우든 서비스 중단은 없다고 ??
5. 데이터는 절대 잃어버려서는 안 된다고 ??
6. 그러면서도, 제일 싸게 만들라고

N드라이브 : 개발/인프라 측면 당면한 문제들...

저장할 파일이 엄청 많고,
계속 늘어나요...
그것도 **폭발적**으로...

파일 시스템용량이 다
차면 알아서 파일들을
재 배치 해 주면 좋겠어요

24시간 X 7일
가용성을
보장해 주세요
...

NAS, SAN...
너무 **비싸요**...

분산하면 빠를 것 같고,
복잡하게 어디에 있는지
알고 싶지 않아요? 그냥
알아서 해주면 안되나요?

혹, 디스크 오류라도 나면
알아서 척척 복구해 주면
좋겠어요...

Scalable, Fault Tolerant, Distributed File System !!
OwFS (Owner Based File System)

NHN 파일스토리지 OwFS

NHN BUSINESS PLATFORM

온라인 파일서비스의 특성

● 파일의 특성

- 파일의 개수가 수십억개 이상으로 늘어남
- 개개의 파일의 크기는 작음 (수 KB ~ 수십 MB이 대부분)
- 단일 서비스의 저장공간이 수십 Petabyte 이상으로 늘어남

● 파일 접근 패턴

- WORM (Write-Once-Read-Many)
  - 변경이 거의 없음
  - 단순 연산
    - Create, Read, Delete
- 새롭게 생성된 파일에 대한 참조 지역성 있음

● 24x7 가용성 요구

어떻게 저장하는 것이 효과적일까?

- **개별 파일의 저장**
  - 파일은 분할해서 저장하지 않고 전체를 저장

- **파일은 3개의 복제본 가짐**
  - HDD/서버 고장에 대응
  - 파일 쓰기 부담은 늘어나지만, 읽기 부하는 분산 가능

- **파일의 복제본 정보 관리 (파일시스템의 메타데이터)**
  - 개별 파일 단위로 복제본 정보를 저장하지 말고 모아서 관리
  - 서로 관련된 파일들을 모아 놓은 container를 “Owner”로 정의

Owner는 분산과 복제의 기본 단위

OwFS (Owner-based File System)에서 파일의 경로

(Owner이름, Pathname)

Owner 개념


내문서
사진
2007년
2008년
2009년
문서
다운로드
매뉴얼
PT자료
. . . . . .

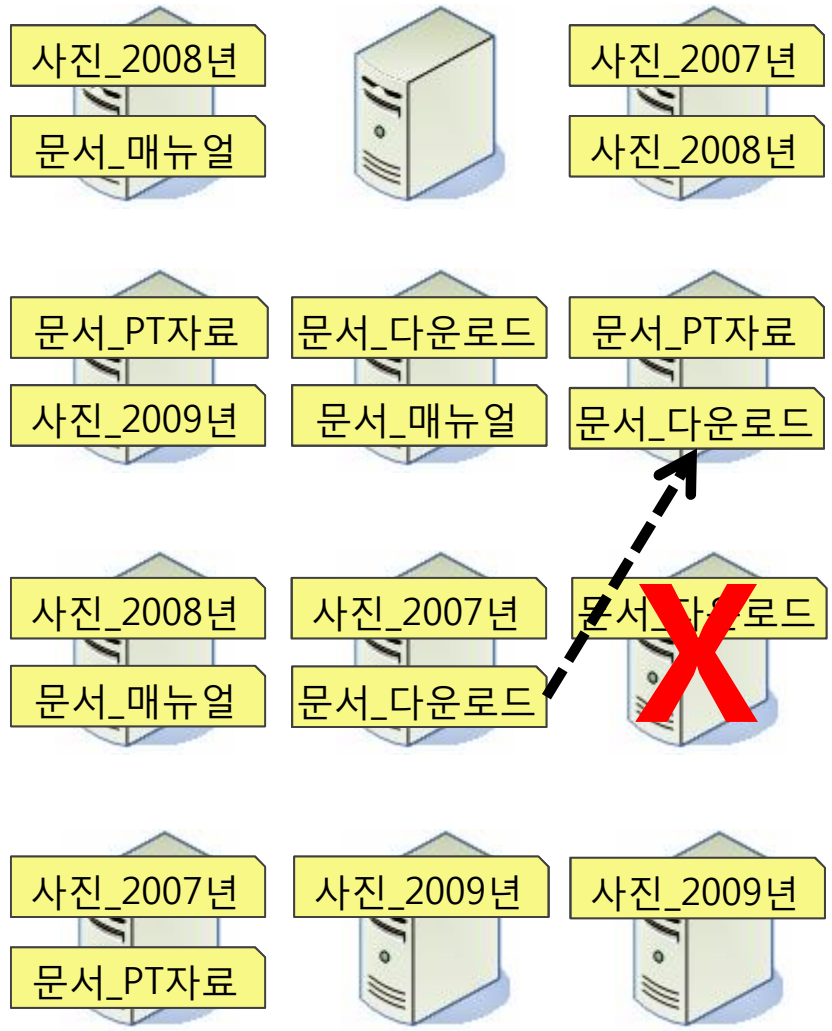


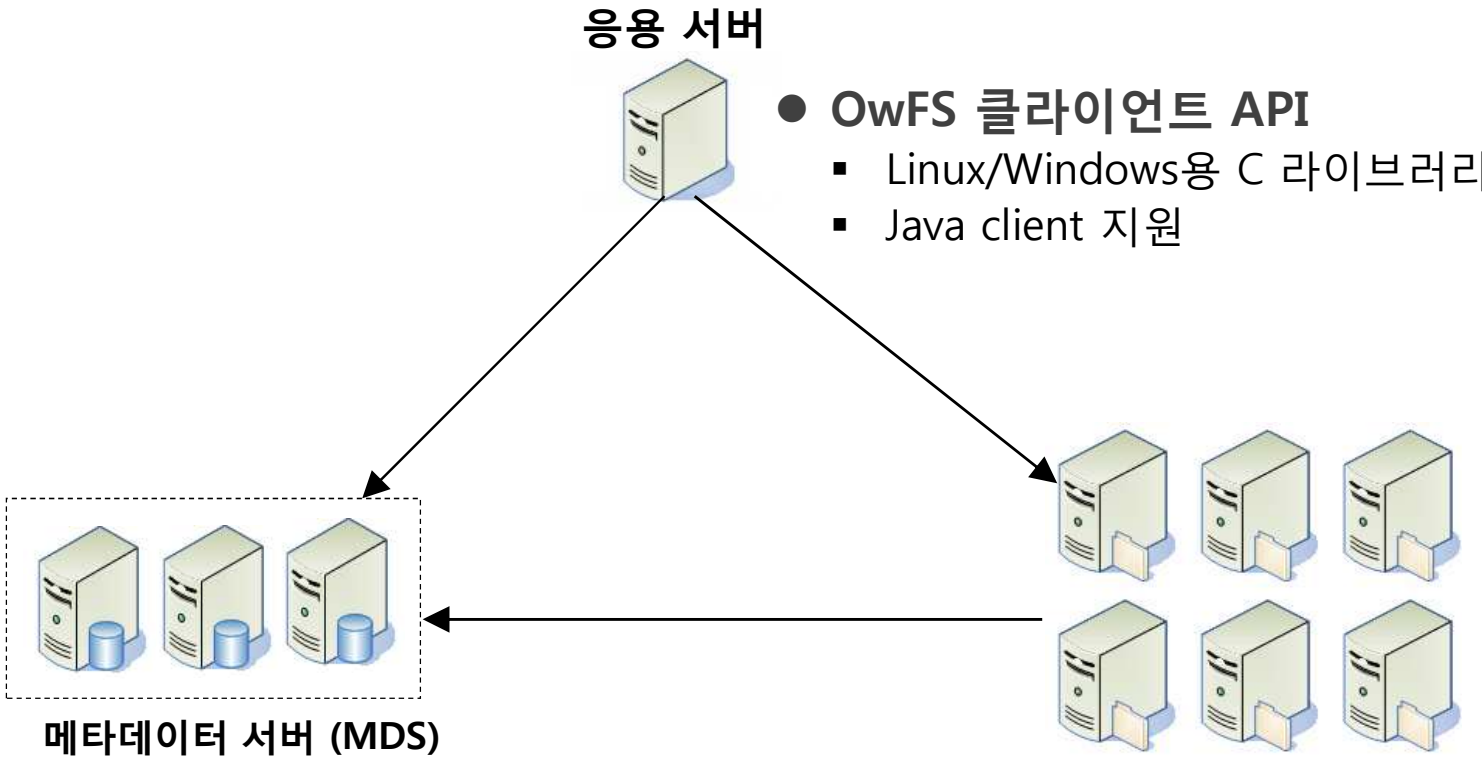
사진_2008년
문서_매뉴얼
사진_2007년
사진_2008년
문서_PT자료
사진_2009년
문서_다운로드
문서_매뉴얼
문서_PT자료
문서_다운로드
사진_2008년
문서_매뉴얼
사진_2007년
문서_다운로드
X
사진_2007년
문서_PT자료
사진_2009년
사진_2009년

분산 파일시스템의 구현 방법

- **고려사항**
  - 성능
  - Legacy 응용과 호환성
  - 개발 및 시험 기간
  - 코드 유지보수
  - 파일시스템 버그로 인한 영향도
  - 플랫폼 업그레이드 용이성
  - 이식성

- **Kernel level vs. User level**
  - User level 구현의 장점이 더 많음
  - API를 이용하여 접근

OwFS 구성요소



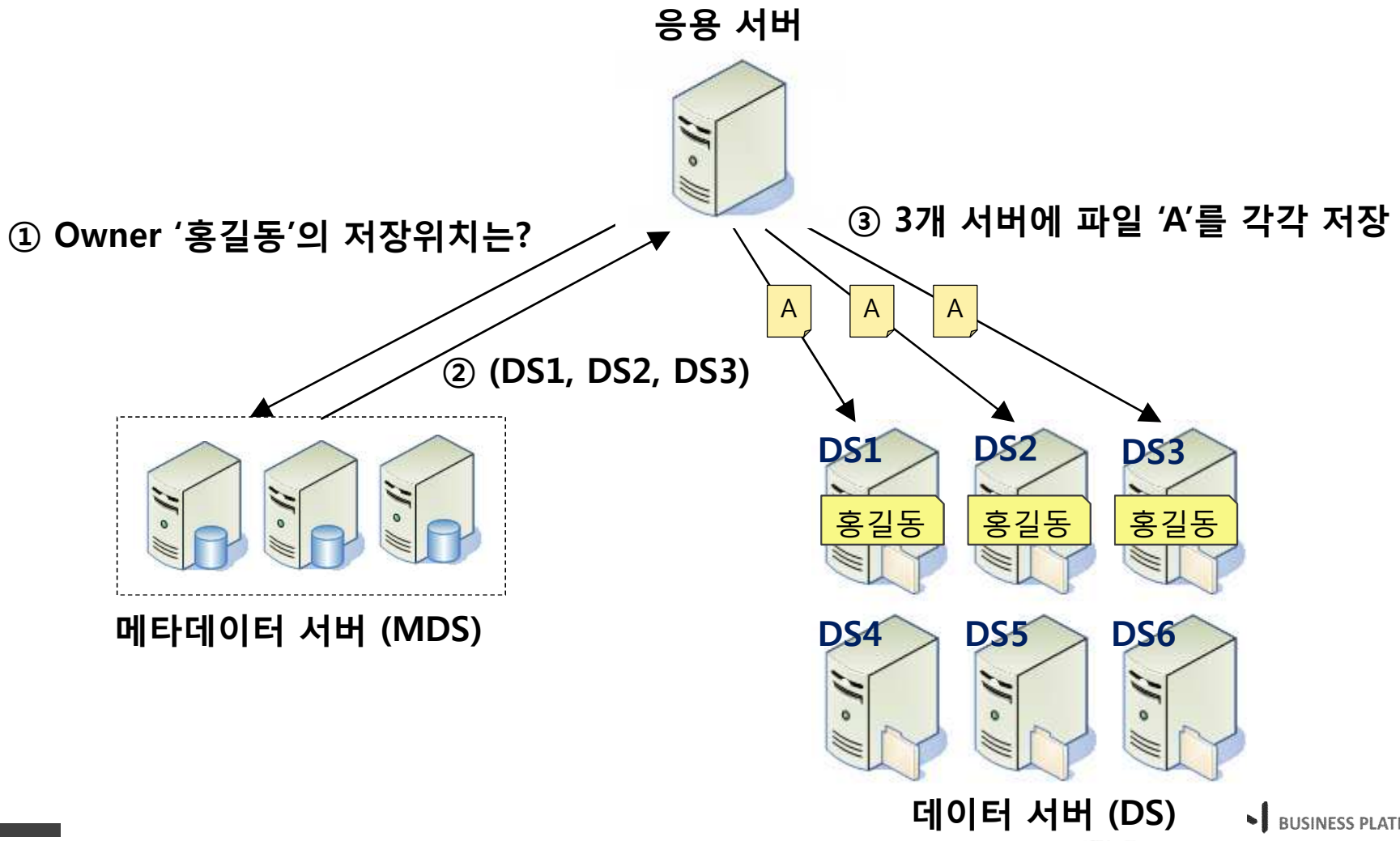
- **OwFS 클라이언트 API**
  - Linux/Windows용 C 라이브러리 지원
  - Java client 지원

- **메타데이터 서버 (MDS)**
  - MDS는 owner별 복제본 위치, 상태 정보 유지
  - 각 owner 당 3개의 복제본 서버 할당
    - 동일한 owner에 속한 파일은 동일한 서버들에 저장
    - 장애 발생 및 복구에 따라 owner의 복제본 상태 존재

- **데이터 서버 (DS)**
  - 실제 파일을 저장하는 서버
  - 파일 서비스의 주체

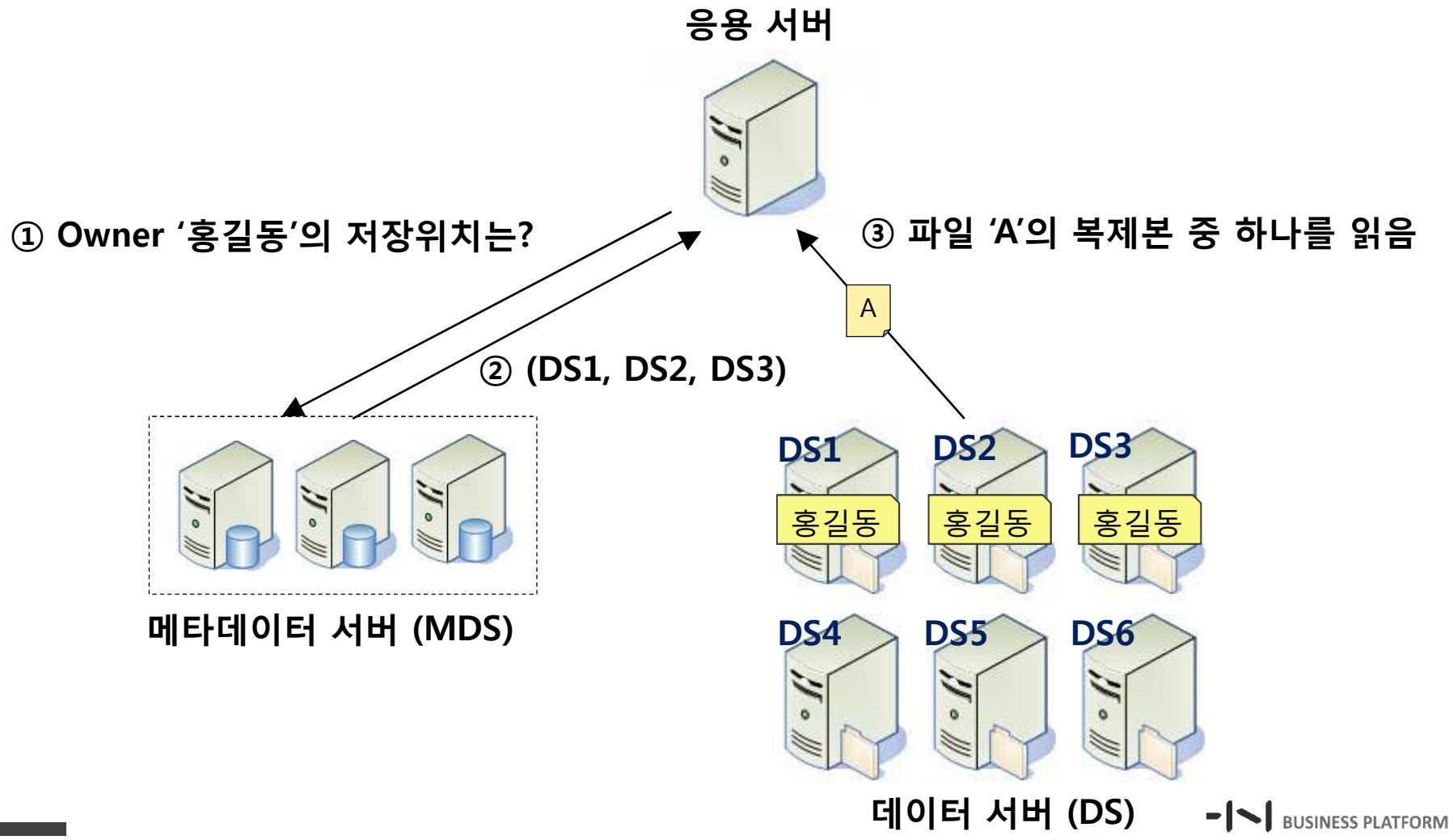
파일 쓰기 동작

홍길동 owner에 파일 A를 쓰기

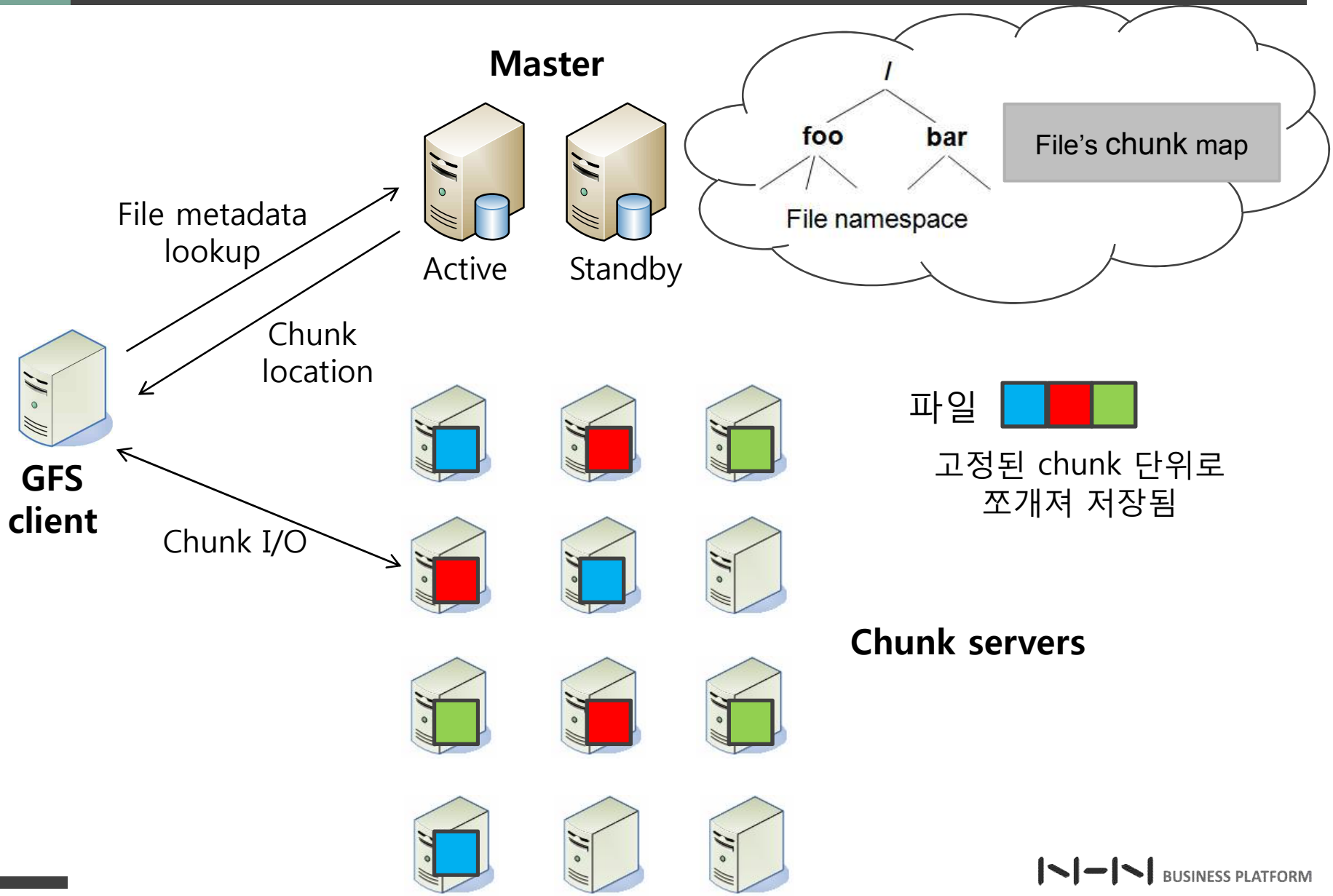


파일 읽기 동작

홍길동 owner에서 파일 A를 읽기



타 파일시스템과 비교 (GFS: Google File System)



타 파일시스템과 비교 (GFS: Google File System)

GFS(Google File System)

- 파일을 chunk 단위로 나누어 chunk 서버에 저장
  - 작은 파일 저장에 불리
- Master 서버가 directory 구조 및 file의 정보를 metadata로 관리
  - File chunk map을 master 서버가 metadata로 관리
- 파일 수가 많아지면 metadata 량도 많아짐
  - Master 서버에 부담이 커질 수 있음
- 적은 수의 대용량 파일에 적합

OwFS

- 파일을 DS에 그대로 저장
- MDS가 owner의 위치 정보만 관리
- 파일 수가 많아져도 metadata의 변화가 거의 없음
  - Owner의 복제본 정보만 matadata로 관리
- 다수의 적은 파일 저장에도 적합

OwFS가 지원하는 API

Owner 연산

- Owner 생성/삭제 (undelete도 가능)/이름변경/리스트조회

파일 연산

- 파일 생성/덮어쓰기/append
  - 파일의 중간 부분변경은 지원하지 않음
- 파일 읽기
- 파일 삭제 (undelete도 가능)
- 파일 이름 변경
- 파일 속성 읽기
- 파일 존재 여부 확인

디렉토리 연산

- 디렉토리 생성/삭제/이름변경/파일리스트 조회

고장 및 모니터링

● OwFS의 고장모델

- Fail-Stop : 조금이라도 이상이 있는 장비는 서비스에서 제외

● 개별 서버 자체 모니터링

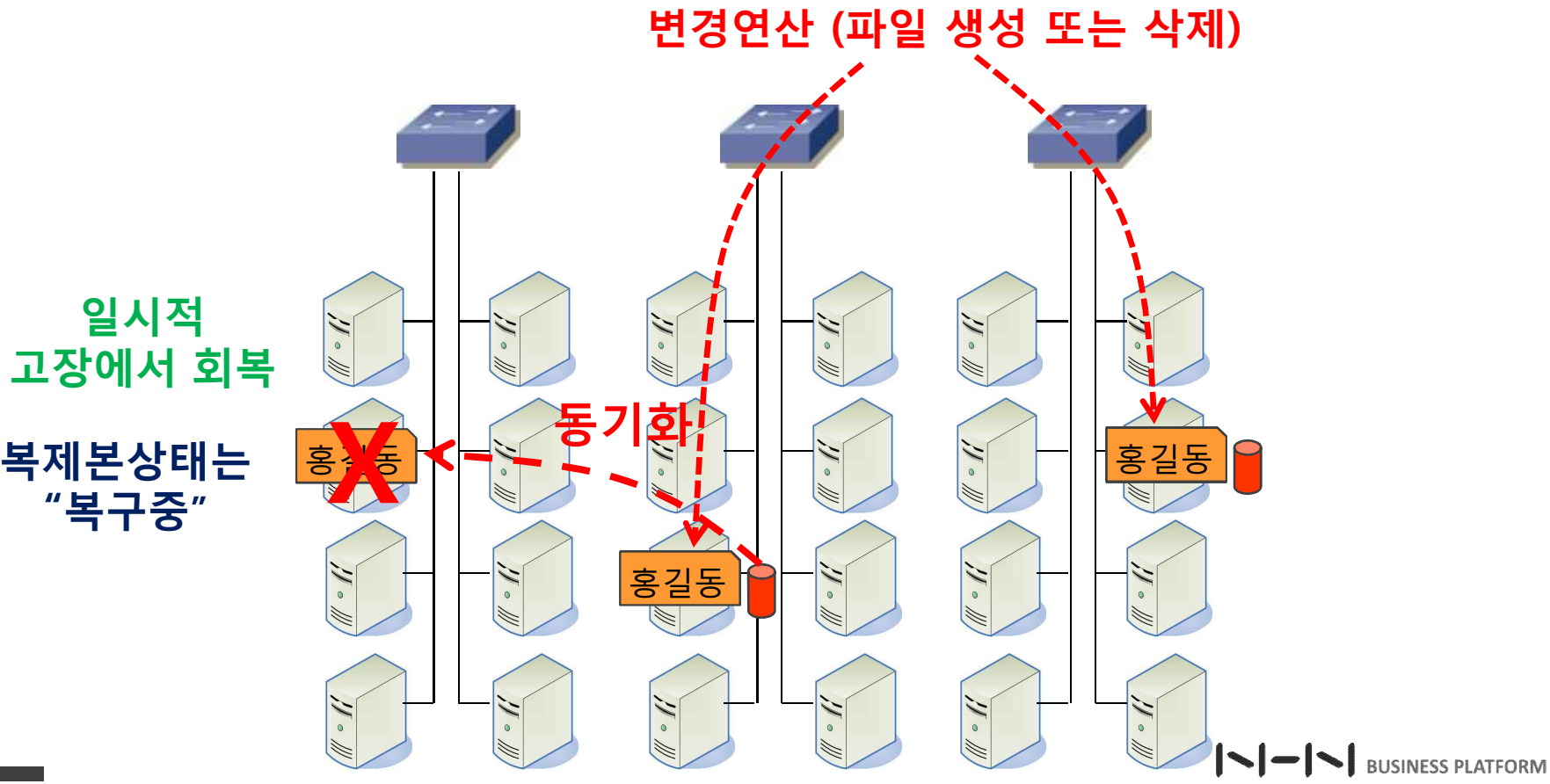
종류	내용
고장모니터링	HDD고장, 커널/파일시스템 고장
	네트워크 포트
	데몬 비정상 종료
성능모니터링	CPU/네트워크/디스크 사용률, 처리량
	메모리, 쓰레드, 네트워크 연결
	파일 연산별 누적 카운터

● 서버간 모니터링

- Heartbeat timeout

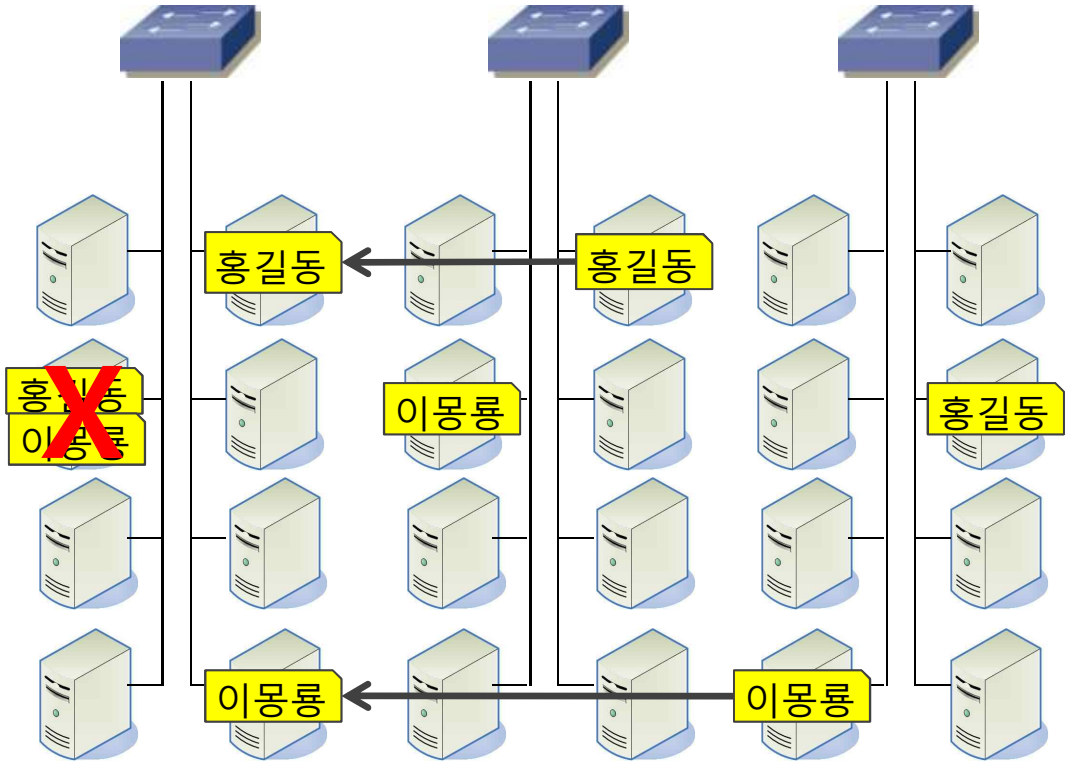
DS 고장에 대한 대응: 일시적 장애

- **DS 장애 발생 직전의 데이터를 이용할 수 있는 경우**
  - 예) 네트워크의 일시적 장애, 관리자에 의한 시스템 재시작, …
- **장애발생기간 중 수행된 파일 연산을 수행하여 복구**



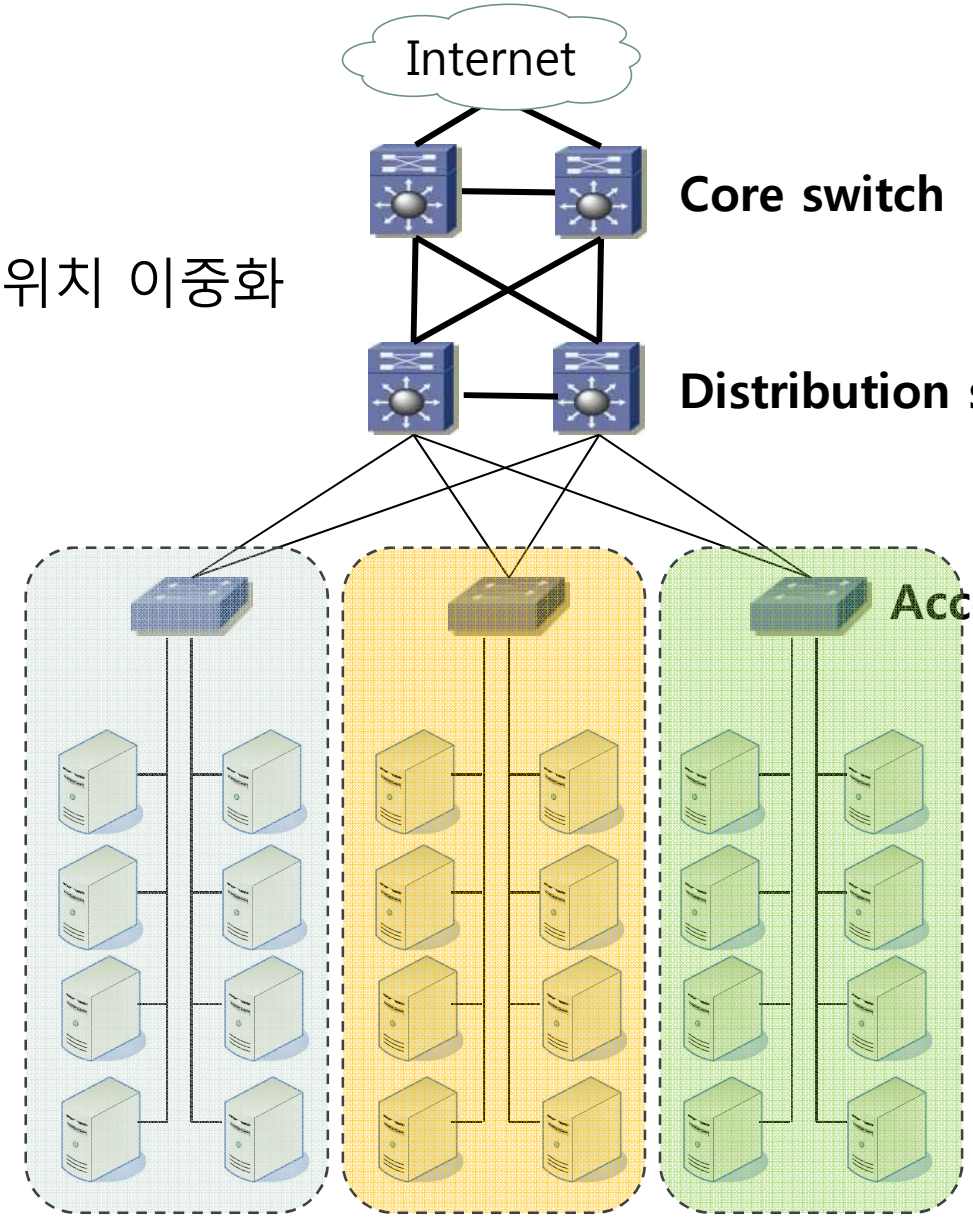
DS 고장에 대한 대응: 영구적 장애

- **장애 발생 직전의 데이터를 이용할 수 없는 경우**
  - 예) 디스크 고장, 파일 시스템 고장, 서버 장애, ...
- **복제본을 다른 DS에 재생성함**



스위치 고장에 대한 대응

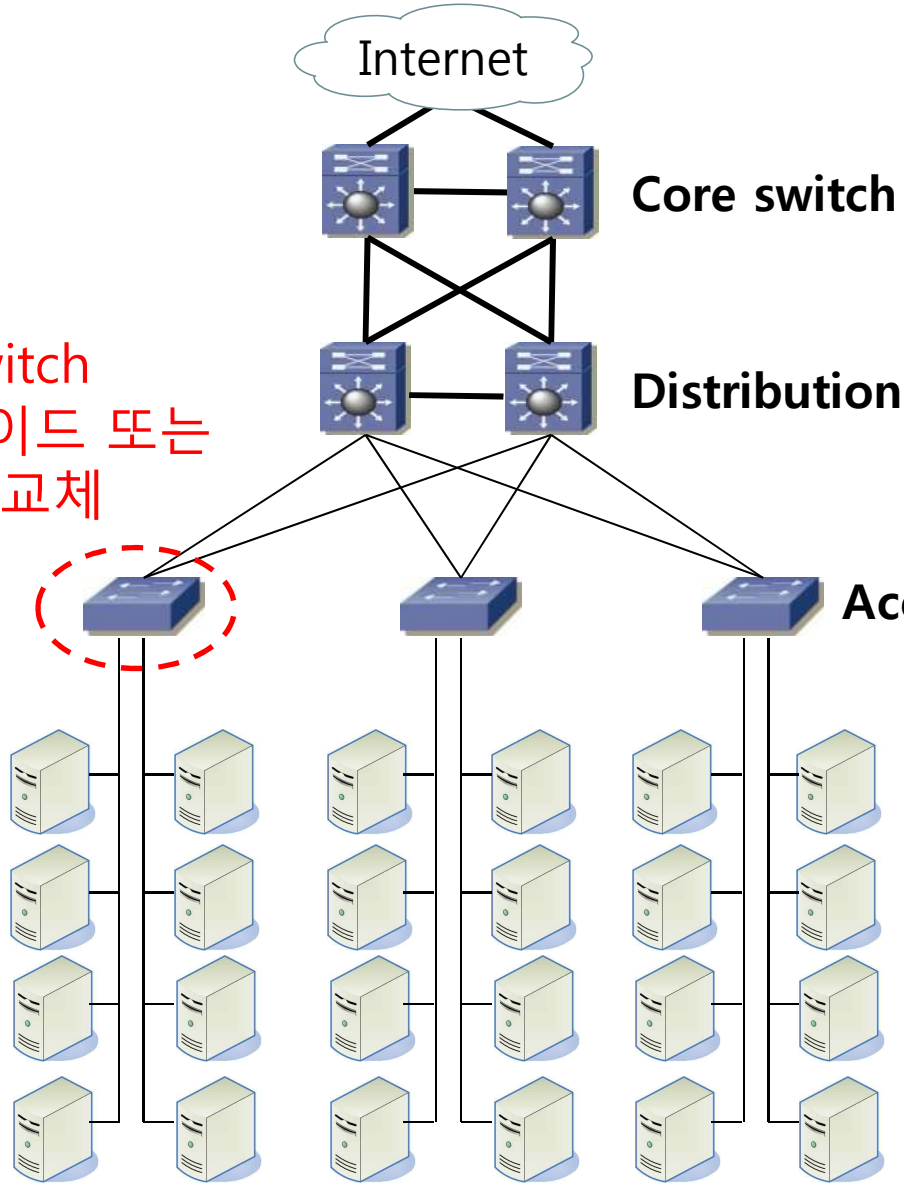
네트워크 스위치 이중화



3벌의 복제본을
fault domain별로
분리하여 배치

OwFS 적용으로 얻은 이점

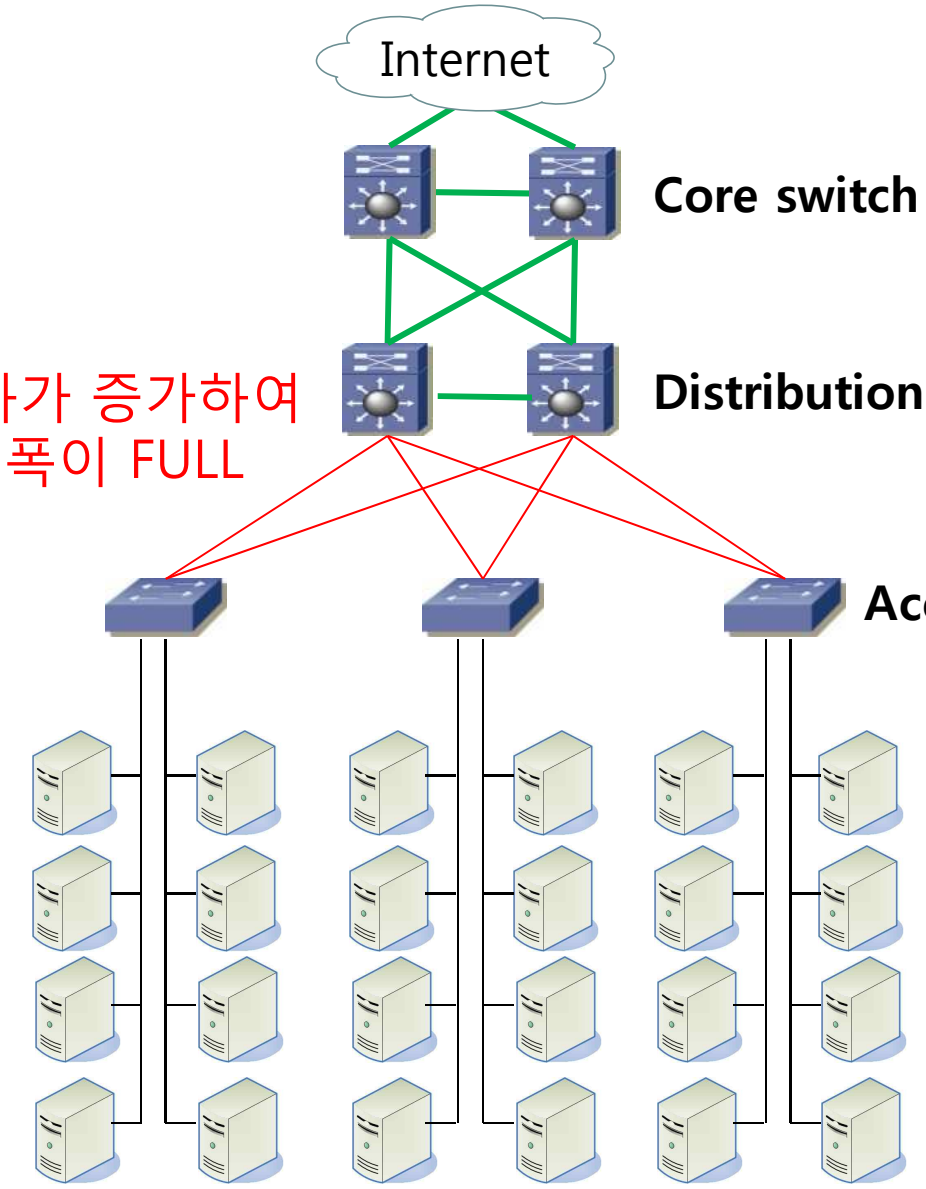
서비스 중단 없이 인프라 유지보수



Access switch
펌웨어 업그레이드 또는
불량 포트 교체

파일시스템 플랫폼
온라인 업그레이드
가능

서비스 중단 없이 네트워크 구성 변경 (1/2)

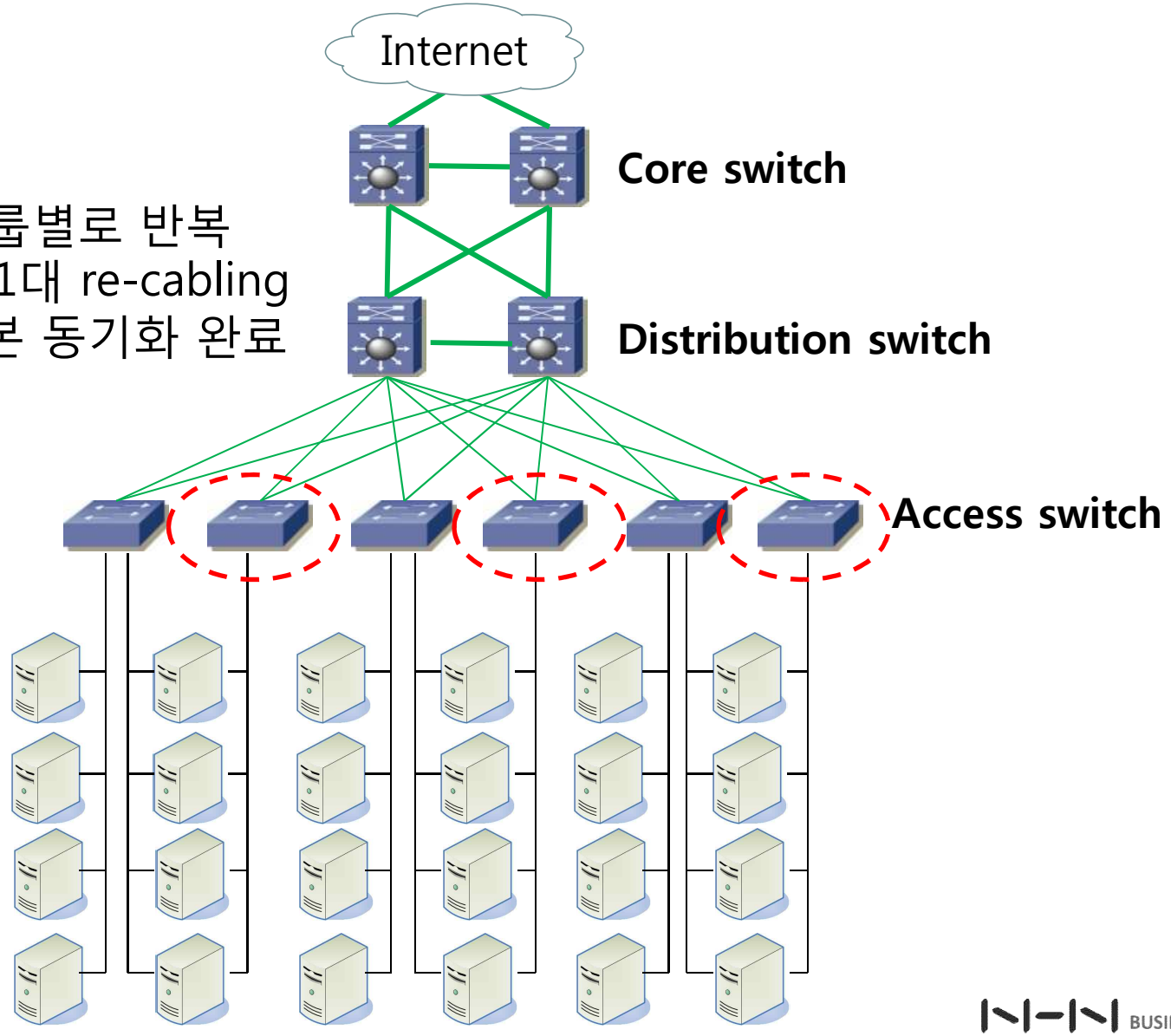


읽기/쓰기 부하가 증가하여
업링크 대역폭이 FULL

각 그룹별로
Access switch를
각각 증설해야 한다면?

서비스 중단 없이 네트워크 구성 변경 (2/2)

스위치그룹별로 반복
Step1. 서버 1대 re-cabling
Step2. 복제본 동기화 완료



확장성

● 용량 확장성

- 저장 공간을 늘리려면 데이터 서버를 추가로 설치
- 데이터 서버가 추가되면 서버당 저장 용량 배분 작업 수행
  - 관리자에 의해 설정된 임계점에 도달하면 자동으로 용량 배분 작업이 기동됨
  - 용량 배분 작업은 데이터 서버에 추가 I/O 부담을 주기 때문에 부하 수준을 제어할 수 있는 방법 제공
- Owner에 대한 이름공간은 그대로 유지됨

● 성능 확장성

- Owner 공간의 분배가 공평하다면, 각 데이터서버는 비슷한 수준의 파일 연산 처리
- 서버를 증설하면 전체 파일연산 수와 처리량이 선형적으로 증가

TCO (Total Cost of Ownership) 개선

- **OwFS는 스토리지 운영상의 다양한 장점을 가지면서 TCO 절감도 가능**

- **Commodity 서버의 내장 SATA 디스크 채용**
  - H/W의 발전으로 추가 TCO 절감

- **표준화된 서버와 네트워크 스위치 구성 관리**

- **기존 네트워크 스토리지에 비해 TB당 TCO 절감**

감사합니다.

NHN BUSINESS PLATFORM