디스크 가상화 – VxFS/ZFS

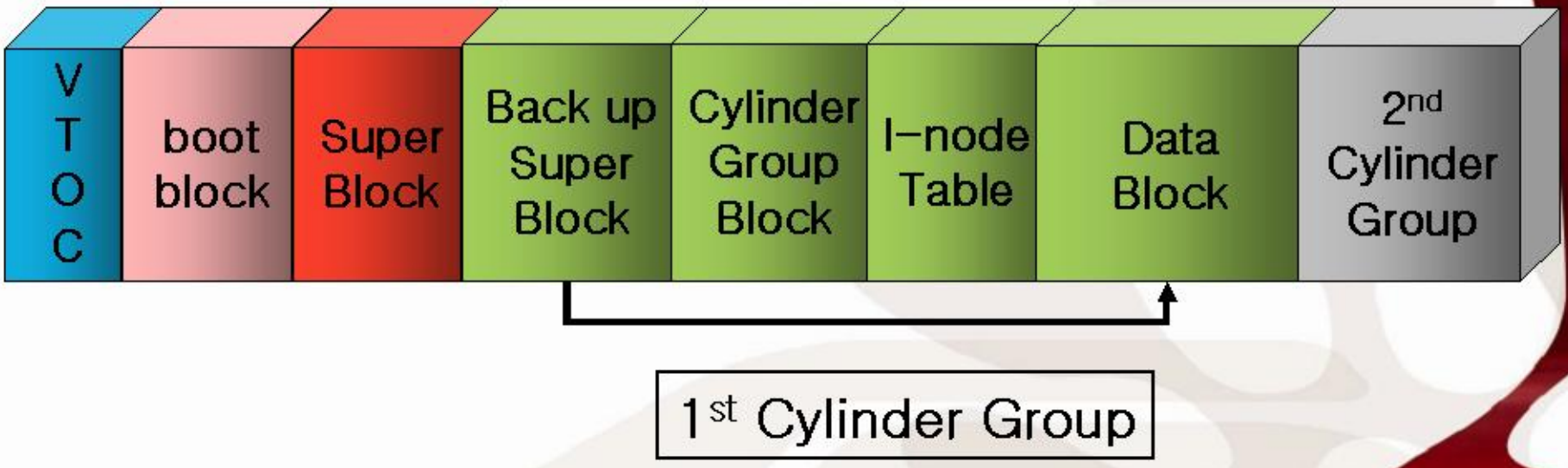
금 동 훈
IT-Lec Consulting
http://cafe.naver.com/solatech
hoonykd@naver.com

Physical Disk 기반의 관리 형태

- 물리적 디스크 개별적인 관리 형태
- 개별 디스크 파티션에 Cylinder group을 이용한 관리 형태
- 개별 디스크에 대한 정보는 개별 디스크에서 관리되며, redundant가 존재하지 않음.
- 초기 설정된 크기 정보 등에 대한 변경이 불가능
- inode, data block과 free inode, data block에 대한 bitmap 정보를 갖는다.
- inode에 의한 data block 지정 방식

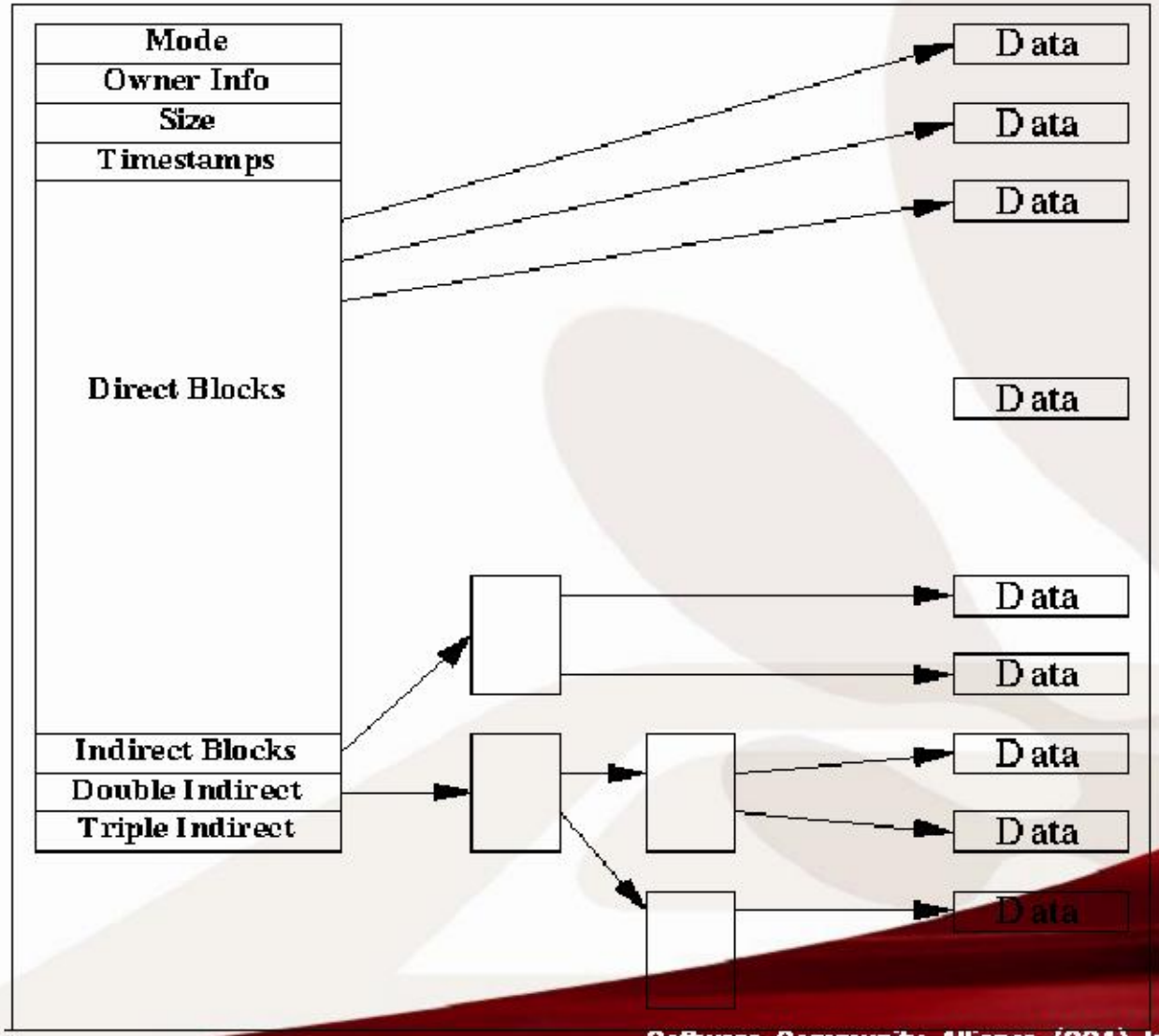
UFS File System 구조 1

UFS File system diagram



UFS File System 구조 2

i-node diagram



새로운 File System의 필요성

전통적인 UFS File system의 한계성

- 디스크 관리의 불편성
- UFS File system의 한계는 1Tbyte
- Random I/O에 적합한 설계로 대용량 데이터를 위한 Sequential I/O 관리에 부적합
- i-node의 개수로 파일 데이터 생성이 제한
- 데이터 점검 시 전체 파일시스템을 비교하는 방식으로 지연 시간이 길어짐

새로운 File System의 필요성

새로운 File system 요구

- DB와 같은 대용량 데이터를 처리하기 위한 Sequential I/O
- File system 관리 시 시간의 단축
- File system 관리의 편의성

VxFS – Veritas File System

VxFS

대용량 데이터를 처리하기 위한 고가용성, 고성능을 제공하기 위한 파일 시스템으로서 다음 내용을 제공한다.

- Fast file system recovery
- Online system administration
- Online backup
- Enhanced file system performance
- Extent-based allocation

VxFS – Veritas File System

Fast File System Recovery

- Intent log를 이용한 file system 복구 시간 단축
- Mount 되어 있는 시스템에서도 전체적인 구조 체크 없이 점검이 가능

- 하드웨어적인 장애나 intent log의 정보가 부족한 경우 수동적인 점검 방식 또는 fsck등의 명령을 사용 점검한다.

VxFS – Veritas File System

Online Administration

Defragmentation

- 디스크 unmount 없이 fsadm 명령을 이용한 조각모음 기능

Resizing

- File system 접속 시에 별도의 조치 없이 file system의 확장 감소가 가능

Online Back Up

- Online 중 snapshot 기능을 이용한 read-only image 백업 기능을 제공
- snapshot 기능을 이용할 경우 원본 file system은 서비스의 정지 없이 백업 가능
- snapshot 생성 후 snapshot 자체를 read-only 형태의 file system으로 이용이 가능

Extent-Based allocation

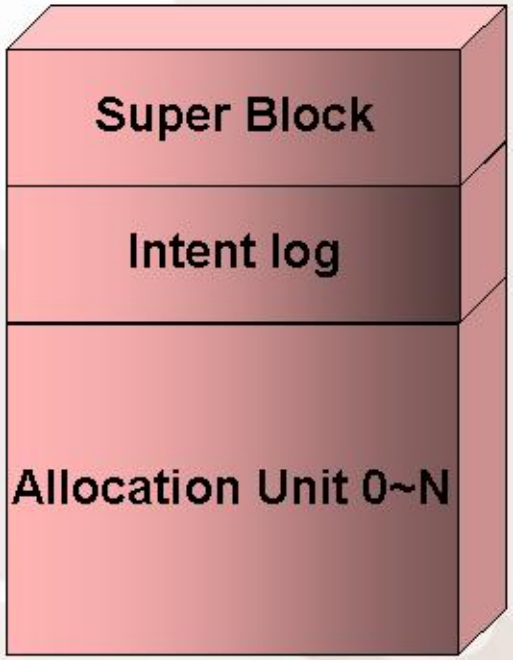
extent는 file system내의 한 개 이상의 연속정인 block을 의미하며 대용량 데이터 기록시 연속적인 공간을 할당하기 위해 사용되는 단위

- 대용량 데이터 접근에 용의한 sequential I/O를 위한 extent-base 도입
- inode가 extent를 참조하여 multiple block에 접근 하는 방식을 사용

VxFS – Veritas File System

Disk Layout

- Super block
- Object-location table
- intent log
- Replica of the object location table
- Allocation unit

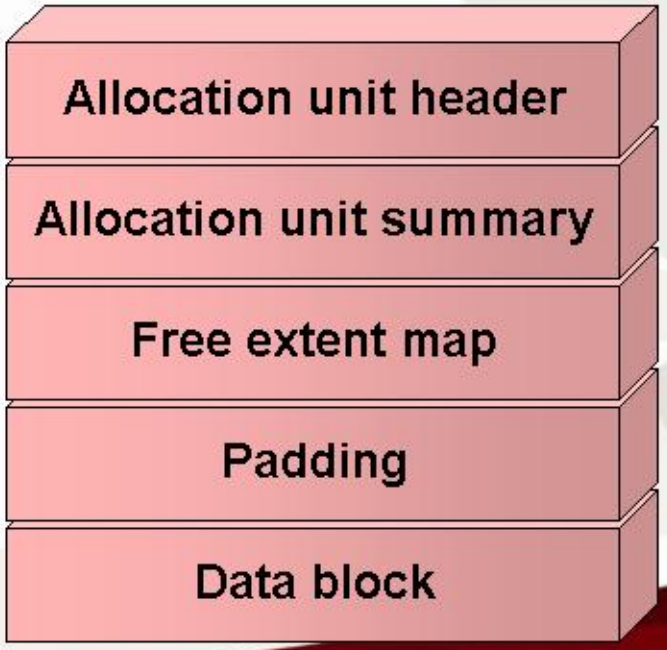


Intent log

- File system의 meta data update후 로그를 기록 하는 방식으로 log 정보 속에 시스템의 파일 시스템 구조 update record를 포함한다.
- system 장애 발생 시 file system transaction의 pending change는 무효화 되거나, fsck명령 등에 의해 완료된다.
- 기본 log 크기는 512byte이며 파일 시스템이 4Mbye 미만인 경우 크기는 감소한다.

Allocation Unit 1

연속되는 block의 그룹으로 resource summary, free-resource map, data block, super block 복사본을 저장하고 있다.



Zettabyte File System - ZFS

ZFS란?

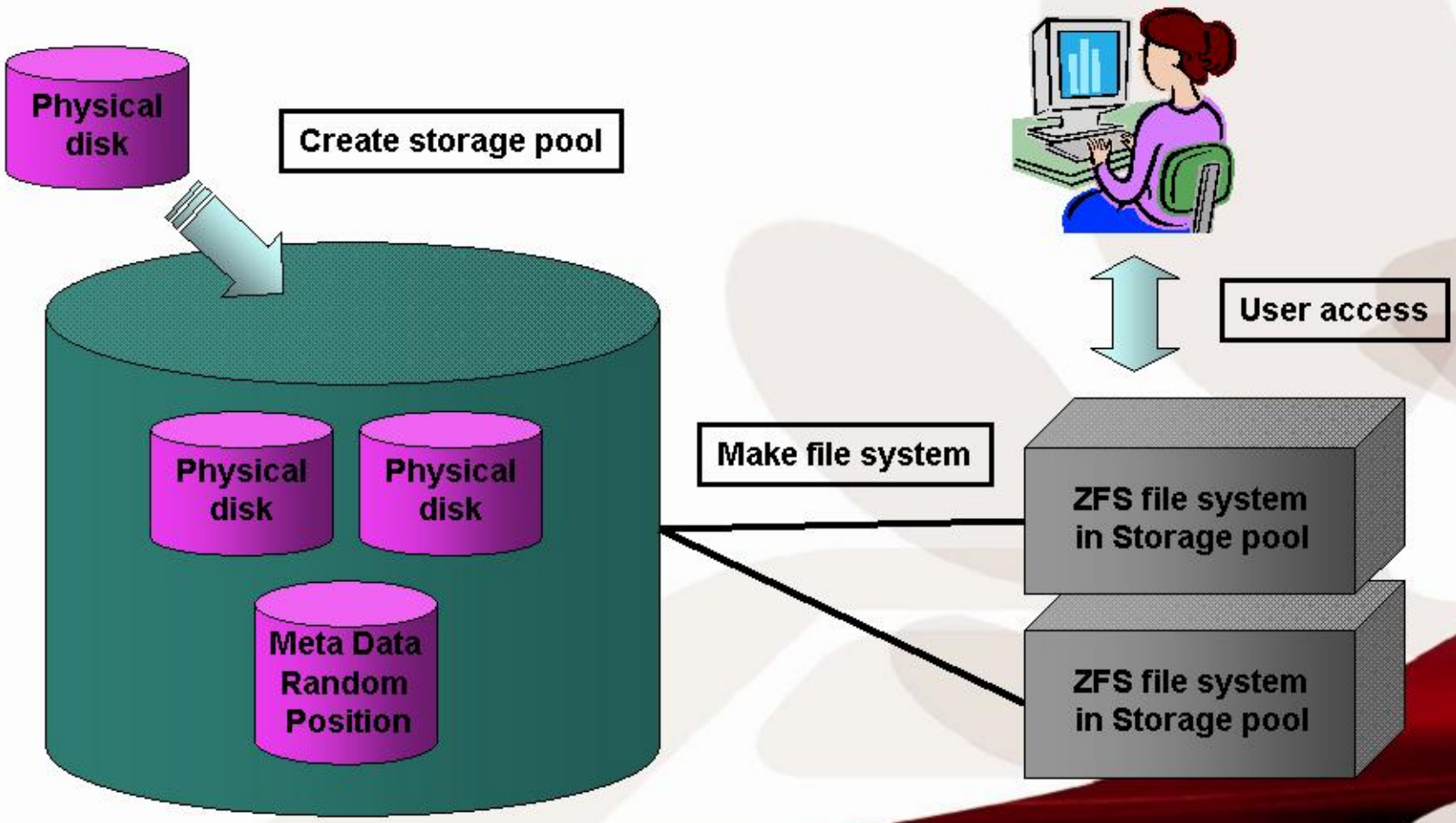
- 기존 발표된 file system과 차별화 된 강하고,
  확장성 및 관리의 간편화를 추구하는 file system
- Pooled storage 이용한 디스크 관리 방식을 사용
- Transactional file system 개념을 도입하여
  디스크 상의 데이터의 일관성이 유지
- Checksum을 이용한 데이터의 self-healing 지원
- snapshot 기능을 이용한 online back up 지원

Pooled Storage

- 물리적 디스크 관리를 위한 storage pool을 생성
- Storage pool에서 file system을 생성
- 다양한 형태의 디스크 가상화를 지원(raidz)
- Storage pool 내의 포함 된 디스크들을 공유하는 형태
- Storage pool에 디스크 추가 시 자동적으로 크기가 확장

Zettabyte File System - ZFS

Storage Pool diagram



Transactional Semantics

- Transactional file system에서 데이터 기록 시 copy-on-write방식 사용한다.
- Copy-on-write 방식이 적용될 경우 데이터는 기존 파일 시스템처럼 원본 위치에 overwrite되지 되지 않으므로 데이터가 손실되지 않는다.

Zettabyte File System - ZFS

Self-Healing Data

- ZFS는 다양한 레벨의 redundancy를 storage pool에 제공
- mirror, raid 기법 등을 적용한 redundancy 사용이 가능
- Bad data 발생 시 다른 복사 본으로부터 올바른 데이터를 복사

Zettabyte File System - ZFS

ZFS와 다른 file system의 차이점 – 관리 편의성

- Space accounting 방식으로 file system 생성 시 storage pool의 디스크들이 공유된다.
- File system 생성 시 사이즈를 지정하지 않는다.
- Meta data를 저장하기 위한 공간이 각 디스크에 할당되며, 동적으로 기록된다.
- 새로운 file system을 /etc/vfstab 파일에 저장하지 않는다.
- 별도의 볼륨 매니저 프로그램이 필요 없다.
- 볼륨 제작을 위한 단계들을 축소 시킴