Sun microsystems

Lustre OST Migration & RAID-1 SNS

Andreas Dilger

Overview

- RAID-1 Server Network Striping
- RAID-1 Recovery Issues
- OST Migration

What is RAID-1 SNS?

- Server Network Striping

What is RAID-1 SNS?

- Server Network Striping
- Mirroring of File Data
- NOT the same as replication feature

What is RAID-1 SNS?

- Server Network Striping
- Mirroring of File Data

- Data Redundancy
- High Availability
- Read Performance Improvement

What is RAID-1 SNS?

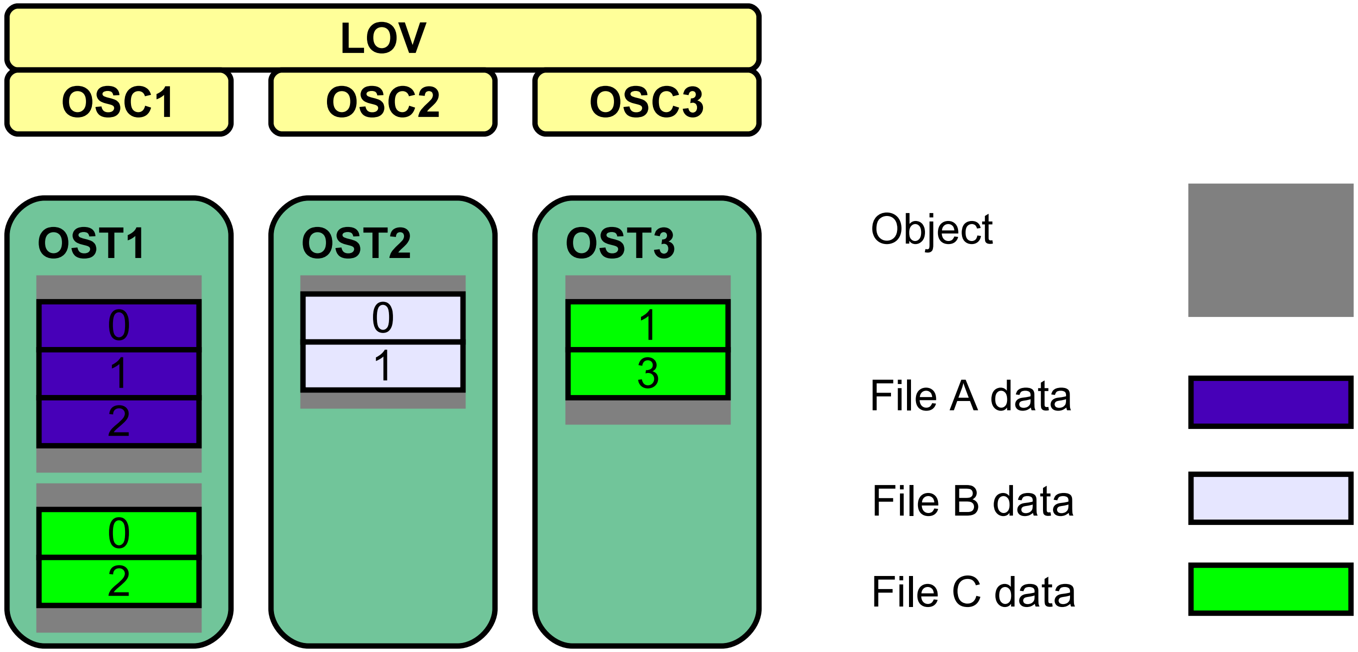
- Server Network Striping
- Mirroring of File Data

- Data Redundancy
- High Availability
- Read Performance Improvement

- Per-file Layouts
- Leverage OST Pools

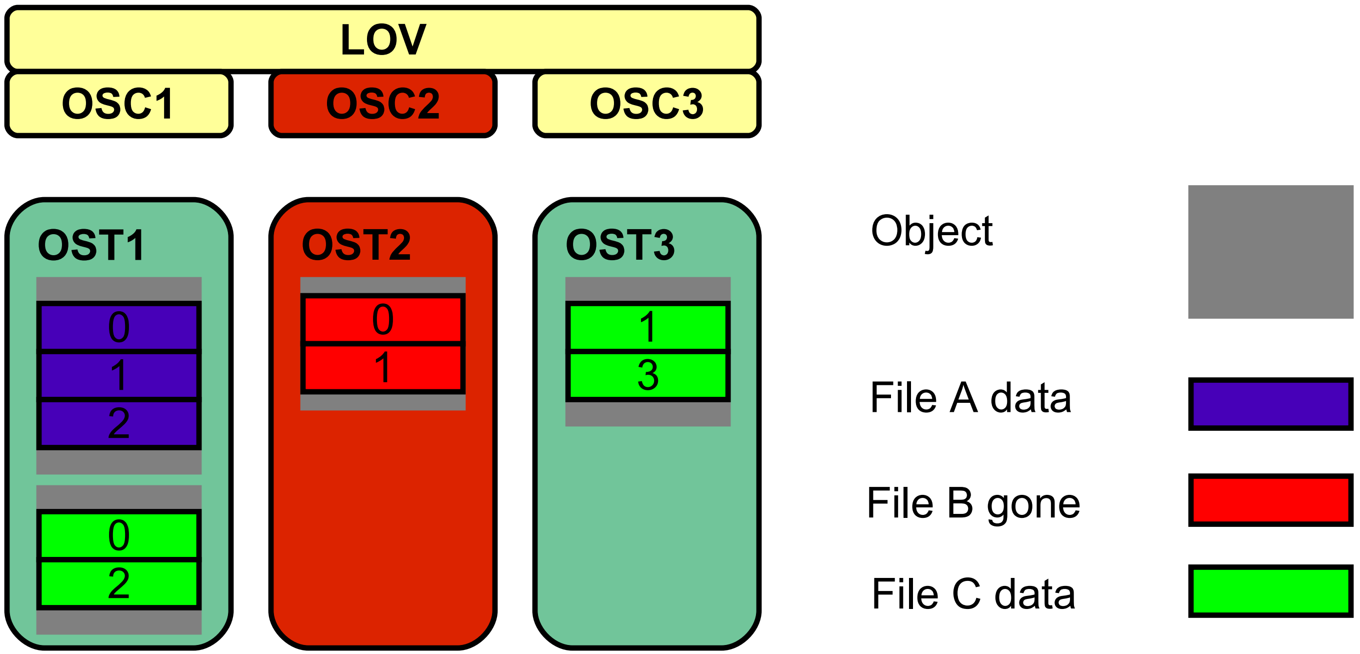
Current RAID-0 Layout

- One copy of each chunk



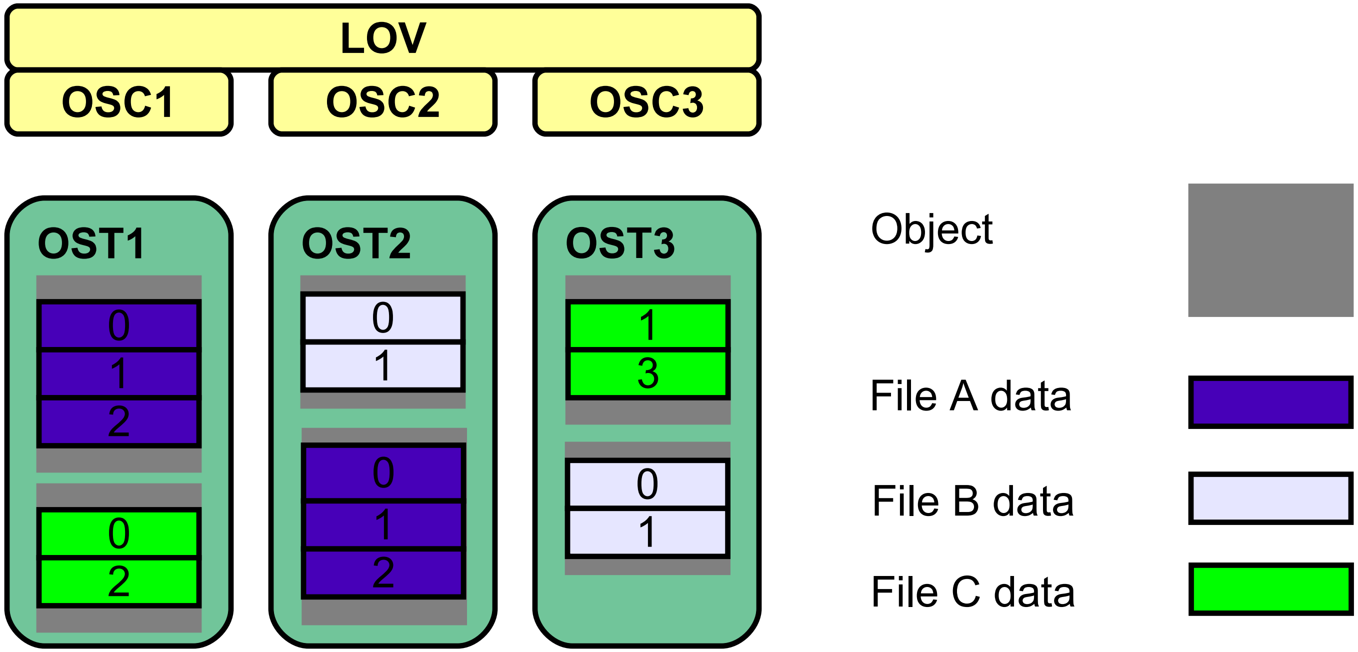
OST2 Is Gone

- Network/server/disk failure
- File B temporarily/permanently gone



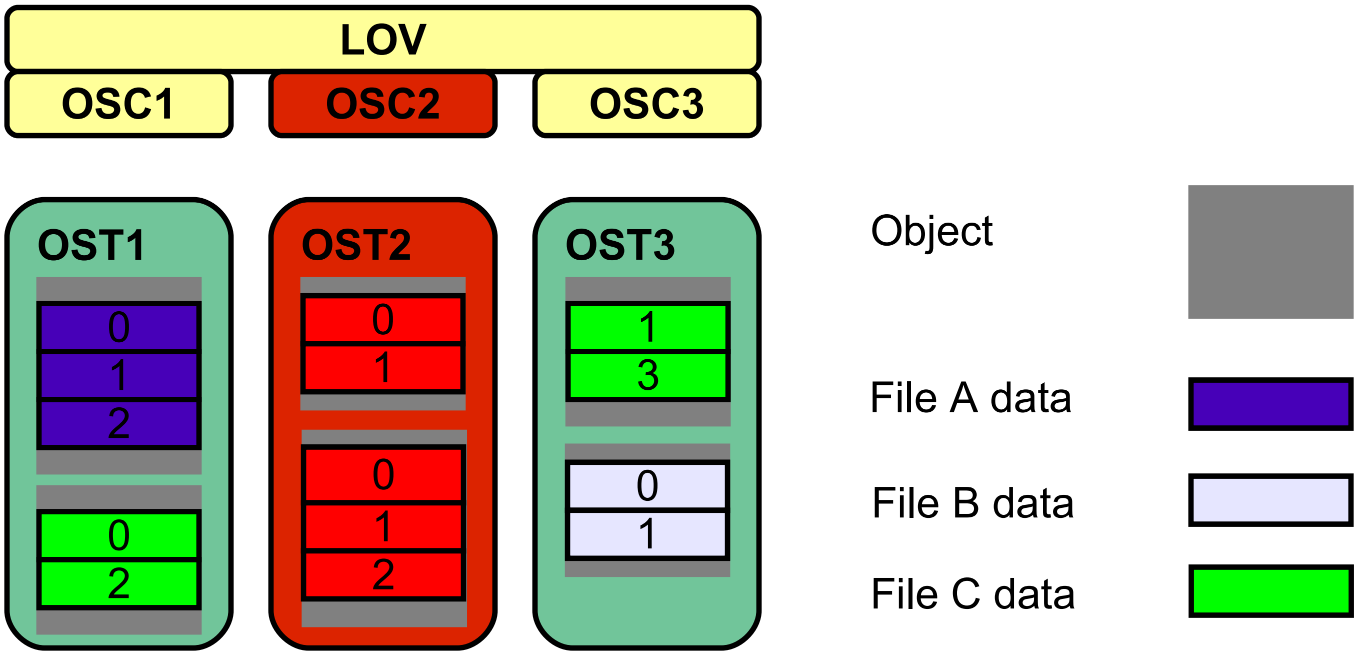
RAID-1 SNS Layout

- Redundant copy of File A, File B
- No extra copy of File C



RAID-1 SNS Layout – OST2 gone

- File A, File B still available
- File C at risk



What is OST Migration?

- Move file data between OSTs
- Change the striping of a file

What is OST Migration?

- Move file data between OSTs
- Change the striping of a file

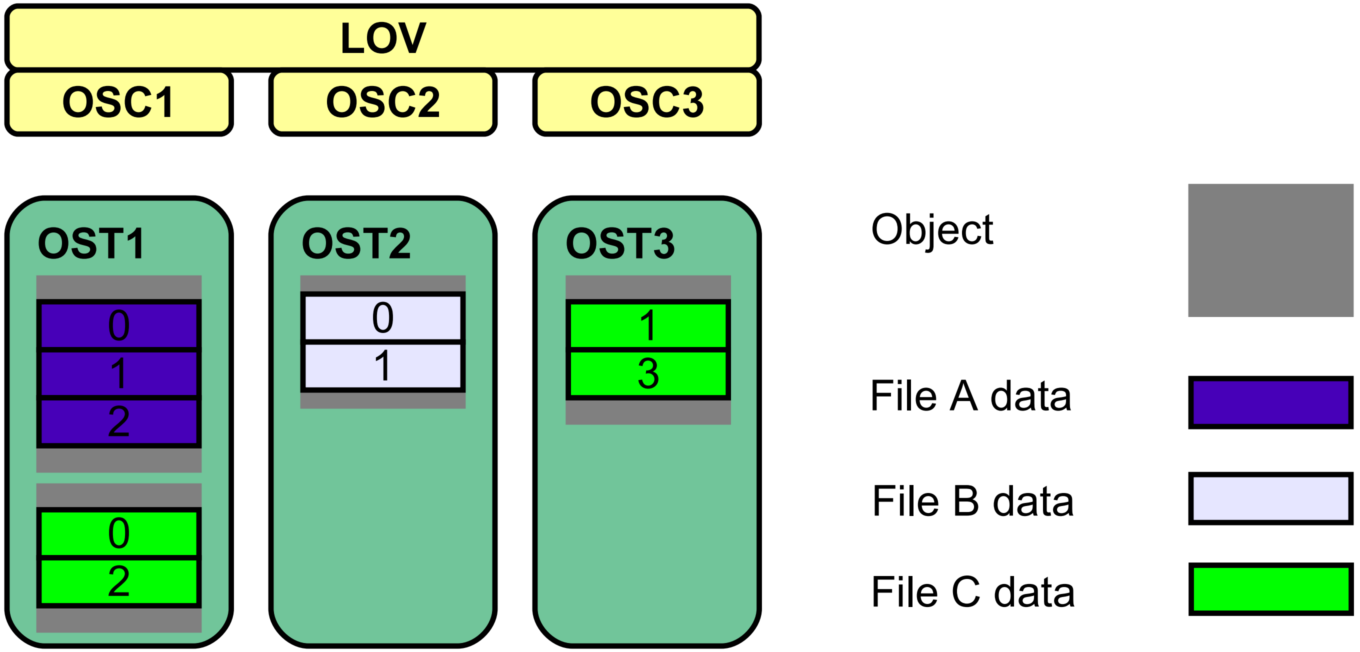
- File accessible during migration
- Robust in the face of failure

Migration Uses

- Space Balancing of OSTs
- Load Balancing of OSTs
- Evacuation of OSTs
- Heirarchical Storage Tiers

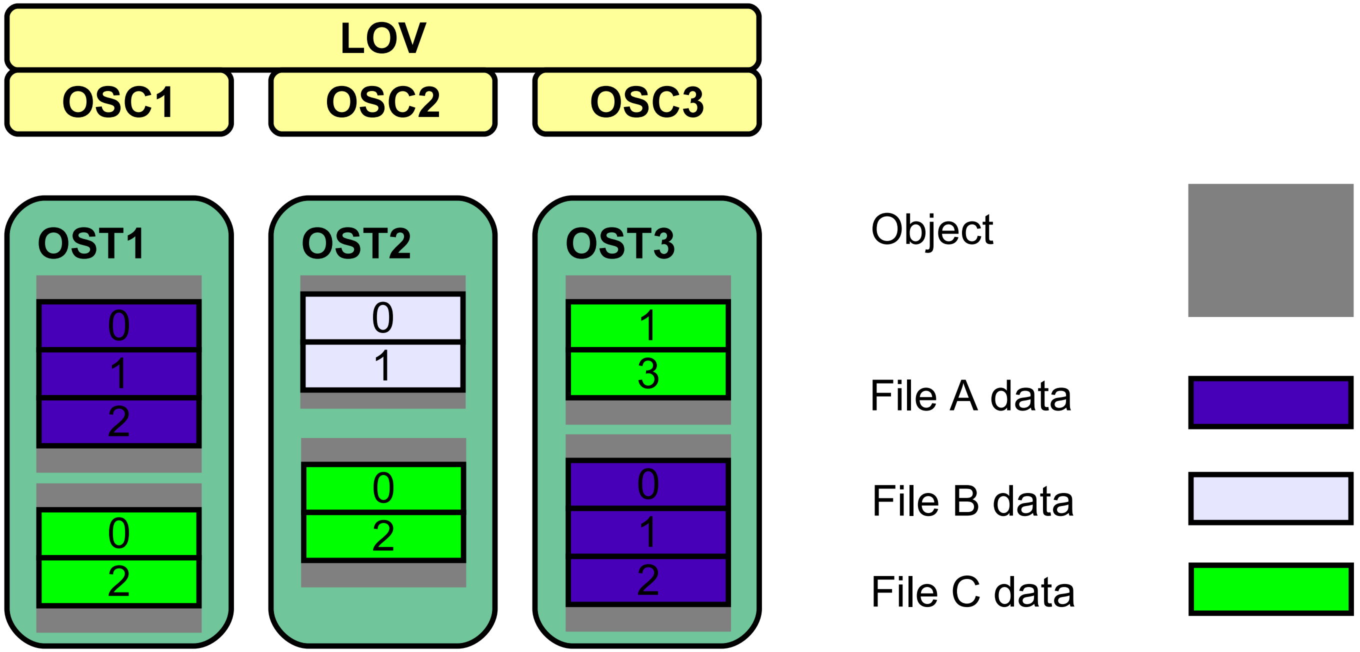
Migration Implementation

- Create SNS RAID-1 mirror of file
- Resync data to new file mirror
- Remove original copy of file



Migration Implementation

- Create SNS RAID-1 mirror of file
- Resync data to new file mirror
- Remove original copy of file



LOV

OSC1 OSC2 OSC3

OST1: 0 1 2 | 0 2

OST2: 0 1 | 0 2

OST3: 1 3 | 0 1 2

Object

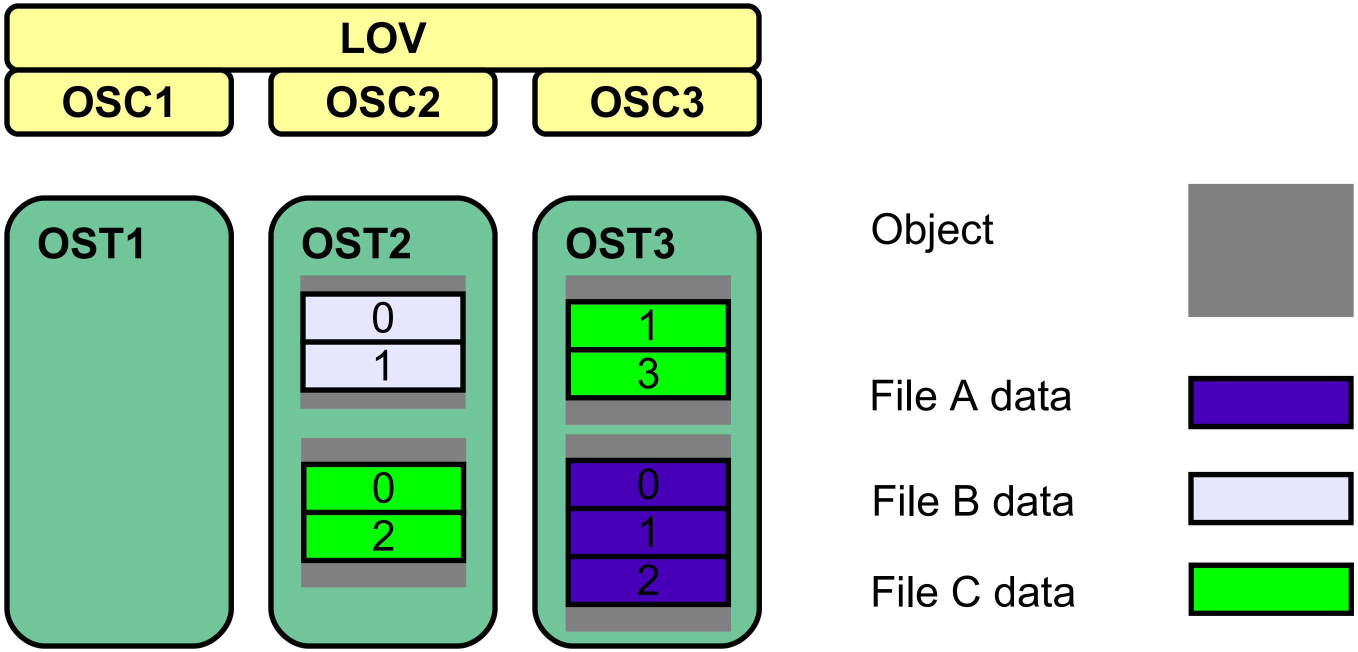
File A data

File B data

File C data

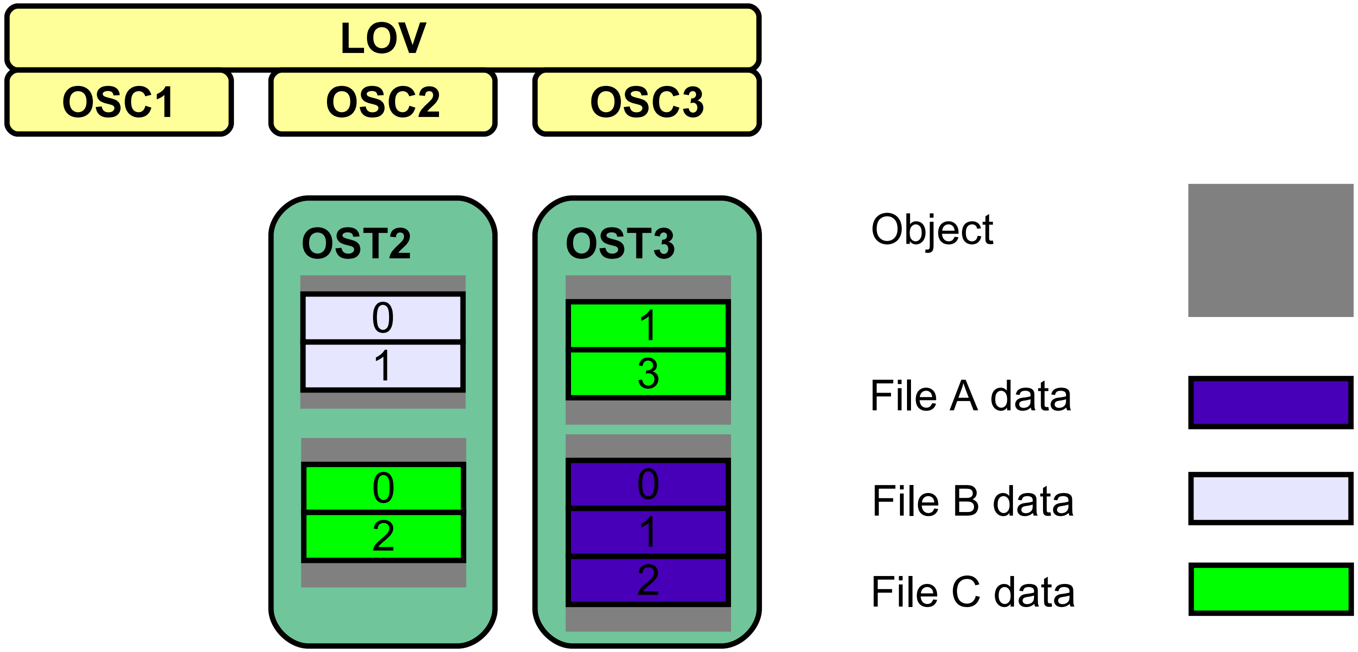
Migration Implementation

- Create SNS RAID-1 mirror of file
- Resync data to new file mirror
- Remove original copy of file



Migration Implementation

- Create SNS RAID-1 mirror of file
- Resync data to new file mirror
- Remove original copy of file



RAID-1 Layout Management

- New LOV_EA IBITS lock on MDS
- Held until EXTENT lock(s) referenced
- Ensures MDS can recall LOV EA
- Write needs all EXTENT locks
- Read needs any EXTENT lock

- New chained LOV EA format
- Allow many mirrors of file

RAID-1 Operation

- Writes go to all stripes of a file
  - > Primary stripe first, then others
  - > Llog cookies returned with first write
  - > Cancelled on write commit at backups
- Reads go to any stripe of a file
  - > EXTENT lock ensures data up-to-date
  - > Clients can load balance over stripes

RAID-1 Recovery & Rebuild

- Failed write marks stripe stale in EA
- Clients will avoid stripe (some/all?)
- Llog on updated stripes drive rebuild
- OST will fetch stale data in recovery
- OST will clear stale flag in LOV EA

Sun microsystems

adilger@sun.com