OCFS2 / ASM / NFS Storage Options for Oracle on Linux

Sunil Mushran

May 2007

Oracle

What is OCFS2?

- General purpose cluster file system
  - Shared disk model
  - Symmetric architecture
  - Almost POSIX compliant
    - shared writeable mmap (expected in 2.6.23)
    - fcntl (2) locking (post 1.4)
- Cluster Stack
  - Small, suitable for a file system

History

- Release 1.0 in August 2005
- Accepted into the mainline Linux kernel with 2.6.16 (January 2006)
- Release 1.2 certified with Oracle RAC (April 2006)
- Support for general purpose usage announced with Release 1.2.5 (April 2007)

Design Principles

- Many learned from kernel community
- Avoid useless abstraction layer
    - Use VFS object life times
    - Mimic the kernel API
    - Keep necessary abstractions as thin layers
- Reuse good ideas
    - JBD, ext3 directory code, group allocation
- Make good ideas resuable
    - configfs
- Keep file system operations local

Features

- Easy setup (2 config files)
  - one for cluster layout and one for timeouts
  - both files are the same across all the nodes
- GUI console to configure and manage volumes
  - Propagates config files to all the nodes
- Full set of tools – mkfs, fsck, tunefs, debugfs
- Integrated cluster stack with DLM
- POSIX compliant (almost)

Distributions

- OCFS2 1.2 packages are currently available for:
    - Oracle Enterprise Linux 4 for x86 & x86-64 on linux.oracle.com
    - Red Hat's RHEL4 for x86, x86-64, ia64, ppc64 & s390x on oss.oracle.com
    - Novell's SLES9 & SLES10 for x86, x86-64, ia64, ppc64 & s390x from novell.com
- OCFS2 1.3 (mainline) shipped with:
    - ubuntu 7.04 “feisty fawn” (2.6.20) for x86, x86-64 and UltraSPARC

Release 1.4

- Features added recently into mainline:
  - sys_splice() (2.6.19)
  - Local mounts (2.6.20)
  - Sparse Files (2.6.22)
- Features in the works:
  - Unwritten extents (posix_fallocate())
  - Shared writeable mmap
  - Freeze/Thaw

Release 1.4

- Features in our TODO list:
  - Data in the inode
  - Online Resize
  - Extended Attributes
  - Global disk heartbeat
  - Integration with CLVM2
- Release 1.4 is being planned for late 2007

What is ASM?

- Logical Volume Manager + File System
- Built into the Oracle kernel (10g +)
- Works only with the Oracle database
- Cross platform
- Can be used with both local and clustered (RAC) databases

ASM

- Simplifies disk administration
    - Automatically manages all storage given to it
- Automatic optimization of data
    - Striped performance
        - No more having to balance data and index datafiles
    - Multi copy redundancy
- Automatic rebalance within a disk group

What is ASMLIB?

- Optional kernel driver provided with ASM on Linux
- Improves manageability
  - Provides disk labeling service for automatic volume discovery
- Improves performance
  - Provides vectored io interface allowing Oracle to bundle multiple ios in one syscall
- Especially useful in large JBOD deployment

NFS

- Runs on Ethernet
    - Low cost commodity network hardware
    - Leverages pre-existing network infrastructure
- Widely implemented IETF protocol
- Popular in environments having mixed storage needs
- Easy setup
- Flexible
    - Some NFS servers also provide iSCSI and Fiber channel targets

Sample Deployments

- Existing NFS infrastructure
    - NFS
- Large JBOD deployment
    - ASM + ASMLIB
- Raw performance with FS interface
    - OCFS2
- Shared Oracle home
    - NFS or OCFS2
- Can mix any of these approaches with Oracle!