Linux Filesystems in the Enterprise

Stephen Tweedie
Red Hat Software
sct@redhat.com

Contents

- Linux Filesystems: Past, Present and Future
  - A Brief History: where have we come from?
  - The current "state of the art" in Linux
  - Challenges for the future
- A Roadmap: Works-in-Progress on multiple fronts
- Summary

Linux Filesystems: a Brief History

- A Distant Memory: Linux as an experimental, "hobby" O/S
- The Minix fs: because Linus had Minix
- The "Extended Filesystem" by Remy Card:
  - Big filenames: >14 chars!
  - Big partitions: >64MB!
  - Very, very, very slow...

Linux Filesystems: the next batch

Ext2fs and Xiafs were developed to solve the previous shortcomings. They looked very similar:

- Traditional, FFS-style update-in-place filesystems
- Files indexed by static inodes
- Ext2 uses "block groups" for locality
- Directories are linear unsorted lists
- Bitmaps for free space control

Advancements in current Ext2fs

The current Ext2fs has evolved incrementally since the initial release:

- Fragmentation-resistant fast allocator
- Optimised directory readahead
- Feature maps for compatibility control
- Sparse superblocks
- 64-bit file support on 64-bit systems
- Ports: HURD, Masix
- Quotas

Advancements in current VFS

The underlying VFS has also been continually improved:

- Generic quota support
- The Page Cache in 2.0
- The "dentry" directory cache in 2.2
- Advanced, RAID-aware readahead

State of the Art in Linux

So just where are we at right now?

- A fast, efficient but very traditional design
- Capable of some good high-end usability
- Works very well under its design limits:
  - Moderate-sized files (<2GB)
  - Moderate-sized directories (100s of entries)
  - Moderate-sized filesystems (10s of GB)
- No systematic tuning for the very high end.

Enterprise Challenges

What is so hard about "Enterprise Computing" anyway?

- Scalability:
  - Performance: scale gracefully under load
  - Size: support very large systems
- Availability:
  - Automatic data replication
  - Fast recovery from failures
- Serviceability:
  - Nebulous "TCO" issues: how much time do you invest keeping it running?

Challenges: Performance

Current efforts are focusing on key performance issues:

- Directory performance
  - Ted Ts'o's Btree directory work
- Apply btree code for fast file indirection
- Fast allocation algorithms for very large filesystems
- Virtual Memory bottlenecks: managing the VFS cache on large memory systems

Challenges: Availability

Availability comes in two parts:

- Make sure you don't lose your data
  - Redundant copies of data
  - Clustered storage
  - Fast backup
- Make sure you can get it back quickly after a failure
  - Journaling filesystems
  - Clustered servers

Redundant storage

The first option: Make local copies

- Software RAID: now supports
  - RAID-0145, 0+1
  - Arbitrary numbers of disks including hot spares
  - Background reconstruct
- Dynamic addition/removal of disks
- RAID hardware support

Redundant storage

The second option: make remote copies

- Use a dedicated cluster filesystem
  - CODA: Dedicated client/server with clustered servers
  - GFS: Shared disk filesystem
- RAID1 over NBD: a possible future avenue

New functionality

The extras which make the filesystem easier to live with:

- 64-bit file support on 32-bit platforms
  - The "Large File Summit" standard is supported in glibc-2.1
  - Matti Aarnio has prototype kernel patches for support on linux-2.2
  - Allows developers to use native 64-bit file sizes or 32-bits plus extensions (for compatibility)

New functionality (2)

- LVM: At least two projects under way
- Very Large Filesystems: How to get over the 2TB device size limit?
  - Prototype patches for dynamic binding of multiple ext2 filesystems
- HSM?
- ACLs?

Summary

- Ext2 (ext3?) can receive incremental tune-ups for enterprise scalability
- Prospects for merging networking and storage together
- The future may not include ext2fs: ReiserFS, CODA, GFS are offering alternatives
- Works is already in hand for the key problems:
  - Large file support
  - Fast recovery
  - Directory performance