Linux Powered Storage:

Building a Storage Server with Linux

Ric Wheeler
Architect & Senior Manager
rwheeler@redhat.com

June 6, 2012

Linux Based Systems are Everywhere

- Used as the base for commercial appliances
  - Enterprise class appliances
  - Consumer home appliances
  - Mobile devices
- Used under most cloud storage providers
  - Google, Facebook, Amazon
- Most common client for high performance computing
  - Lustre, Panasas

What Goes into a Linux Storage Server?

- Server components
  - Kernel NFS server
  - Samba (user space) server
  - Target mode support for block (iSCSI, FCoE)
- Local file systems
  - Ext4, XFS and Btrfs
- Clustered file systems
  - GFS2, OCFS2, etc
- Block layer
  - LVM, RAID code, remote replication
  - Support for new devices types (SSD's, etc).

Things We Get Right

Linux NFS Servers

- Support most of the 4.1 NFS specification
  - Client supports all of the mandatory features and pNFS
  - Server supports most of the mandatory features (not pNFS)
- Reasonably good performance for streaming and small file workloads
- Supports pretty much any transport
  - Ethernet, IB, ...
- Increased involvement in IETF
  - Direct community member engagement and traditional standards people also reach out to Linux developers

CIFS (aka Samba) Support

- Samba is robust and widely used
  - Has cluster support with CTDB
- In kernel CIFS client provides an alternative solution for Linux
  - Performance now approaches NFS performance
- Good community relations with Microsoft
  - Multiple plugfest like events per year
  - Regular engineering calls
- Large, multi-vendor development community
  - SambaXP conference each year just for Samba (and CIFS client) development

Ext3 File System

- Ext3 ~~is~~ was the most common file system in Linux
  - Most distributions historically used it as their default
  - Applications tuned to its specific behaviors (fsync...)
  - Familiar to most system administrators
- Ext3 challenges
  - File system repair (fsck) time can be extremely long
  - Limited scalability - maximum file system size of 16TB
  - Can be significantly slower than other local file systems
  - direct/indirect, bitmaps, no delalloc ...

The Ext4 filesystem

- Ext4 has many compelling new features
  - Extent based allocation
  - Faster fsck time (up to 10x over ext3)
  - Delayed allocation, preallocation
  - Higher bandwidth
- Users still on EXT3 have an easy migration path
  - The same commands and utilities
- Large and active developer community that crosses multiple vendors

The XFS File System

- XFS is very robust and scalable
  - Very good performance for large storage configurations and large servers
  - Many years of use on large (> 16TB) storage
- XFS is the most common file system used in serious storage appliances
- Reasonable sized and active developer community that crosses a few vendors

The BTRFS File System

- BTRFS is the newest local file system
- More integrated approach to the storage stack
  - Has its own internal RAID and snapshot support
  - Does full data integrity checks for metadata and user data
  - Compression support
  - Can dynamically grow and shrink
- Ships in multiple enterprise and community distrbutions
- Large and active developer community that crosses multiple vendors

Active Maintenance and Development

- Since kernel v2.6.18 (~RHEL5):
  - Ext3 : 556 commits, ~136 authors
  - Ext4 : 1649 commits, ~213 authors
  - XFS : 1857 commits, ~136 authors
  - Btrfs : 2228 commits, ~139 authors
- Each file system has relatively few very active authors

New Features Tend to be Widely Supported

- Ext4, XFS, btrfs all have:
  - Delayed allocation
  - Per-file space preallocation
  - Hole punch (not on btrfs yet)
  - Trim / discard
  - Barrier (now flush/FUA) support
  - Defragmentation
- Ongoing work to unify the mount options across all file systems

LVM and Block Layer

- LVM and device mapper has had an activity spurt
  - New support for thin provisioned target
  - LVM can manage MD RAID devices
  - Native multipath increasingly used in high end accounts
- New open source drivers for PCI-e SSD cards
  - Micron driver is now upstream
  - Intel has been promoting the NVM express standard and driver
- Multiple ways to use SSD devices as a cache
  - Bcache, vendor specific, fscache, ???

Very Active Developer Community – LSF/MM 2012

What Do We Get Wrong?

Linux NFS Servers Problems

- Experience with NFS 4.0 and 4.1 still relatively new
  - Expect to get increases in user base
  - Will complete any lingering rough edges on server implementation
- Lacks support for clustered NFS servers
  - Running with a shared file system back end “mostly” works
  - Ongoing work to resolve lock recovery deficiencies
- Missing pNFS server code in upstream
  - Microsoft is likely to have production a pNFS file layout server before the upstream kernel

Samba Challenges

- Microsoft is moving rapidly to SMB3.0
  - Specification will be completely finalized once Windows8 server ships
  - Fixes performance issues
  - Good support for clustered servers
- Samba support for SMB2.1 mostly there
- SMB3.0 development
  - Samba plans for SMB3.0 support underway
  - CIFS client support limited to SMB1

Lack of Rich ACL Support

- Windows and Linux/UNIX are really different
  - Windows locks are mandatory
  - Linux locks are advisory
- Exporting the same file system via both NFS and CIFS leads to data corruption for lock users
- Rich ACL patch provides the missing support
  - Need the “rich ACL” patches to add support for Windows style semantics
  - Currently not actively being worked on

Ext3 Challenges

- Ext3 challenges
  - File system repair (fsck) time can be extremely long
  - Limited scalability - maximum file system size of 16TB
- Major performance limitations
  - Can be significantly slower than other local file systems
  - Dwindling developer pool

Ext4 Challenges

- Ext4 challenges
  - Large device support (greater than 16TB) is relatively new
  - Has different behavior over system failure than ext3 users are used to
- Usability concerns
  - Lots of mount options and tuning parameters
  - Relies on complex and high powered tools to support LVM and RAID configurations

XFS Challenges

- XFS challenges
  - Not as well known by many customers and field support people
  - Until recently, had performance issues with meta-data intensive (create/unlink) workloads (fixed in upstream and recent enterprise releases like RHEL6.2)
- Similar usability concerns
  - Fair number of mount options and tuning parameters
  - Relies on complex and high powered tools to support LVM and RAID configurations

BTRFS Worries

- Repair tool still very young
- Ongoing worries with the hard bits of doing “copy on write” file systems
  - ENOSPC took a while (fixed now!)
  - Encryption yet to come
  - COW can fragment oft-written files
- Performance analysis and testing takes a back seat to XFS and ext4 work

All Things Management Related

- Linux systems have a tradition of relying on third party management tools
  - Lots of power tools for experts
  - Few tools appropriate for casual users
- Many ways to do one thing
  - Multiple RAID, SSD block caching layers

Ongoing Work Worth Following

NFS & Samba

- Advanced support for clustered storage very active
  - Lock recovery work being pushed upstream
  - Multiple (out of tree) parallel NFS servers
  - FedFS support
- All things to do with SMB3.0
- Combinations of NFS servers and Samba with other file systems
- NFS V4.2 adds new support for
  - Copy offload operation, FedFS and Labeled NFS
  - Most of this not yet implemented

Ext4 Scaling & Features

- Bigalloc (since kernel 3.2)
  - Workaround for bitmap scalability issues
  - Allocates *multiples* of 4k blocks at a time
  - Not true large filesystem blocks, but close?
- Inline Data - planned(maybe?)
  - Store data inline in (larger) inodes
  - Mitigate bigalloc waste?
- Metadata Checksumming - planned

XFS Scaling & Features

- “Delayed logging” is done
  - dramatically improved metadata performance
  - default since v2.6.39
  - Last™ big performance issue
- Integrity work is next
  - CRCs on all metadata and log
  - FS UUID to detect misdirected writes
  - Transaction rollback in the face of errors
  - Background scrub

BTRFS Scaling & Features

- Scaling work here and there
- Mostly still fleshing out features
  - Checksumming was done early
  - RAID 5/6
  - Quotas
  - Dedup
  - Encryption

Block Level Convergence

- Active work on converging the SSD block cache layer
  - Proposal to get bcache from Google ported into device mapper
- Ongoing effort to reuse RAID implementations

Management Work

- Libstoragemgmt
  - Provides a library to do common block level operations on storage arrays
  - Full time developers and storage vendor participation
  - http://sourceforge.net/apps/trac/libstoragemgmt
- System Storage Manager
  - Btrfs like “ease of use” for xfs, ext4 on top of LVM
  - http://sourceforge.net/p/storagemanager/home/Home/

Resources & Questions

- Resources
  - Linux Weekly News: http://lwn.net/
  - Mailing lists like linux-scsi, linux-ide, linux-fsdevel, etc
- Storage & file system focused events
  - LSF workshop
  - Linux Foundation events
  - Linux Plumbers
- IRC
  - irc freenode.net
  - irc.oftc.net