

Linux File & Storage Systems

Enabling the Latest Storage Hardware in Linux

Ric Wheeler
Red Hat

Overview

- ❑ What is Linux?
- ❑ Evolution of Three Features
- ❑ Effectively Working with the Linux Community

Kernel Summit 2010

SDC 
STORAGE DEVELOPER CONFERENCE
SNIA ■ SANTA CLARA, 2011



What is Linux

Not just a monolithic body

- Various free and fee-based distributions
- Hardware vendors from handsets up to mainframes
- Many different development communities

No single party can promise your favorite feature will be in all distributions

- Getting into upstream makes it much more likely though!

The Linux Community is Huge

- Most active companies in 3.0
 - No affiliation - 12.0%
 - Red Hat - 11.1%
 - Intel - 9.3%
 - Unknown - 6.3%
 - Novell /SuSE - 4.9%
 - IBM - 4.2%
 - Microsoft - 4.0%
 - Atheros Communications - 2.7%
 - Texas Instruments - 2.6%
 - Broadcom - 2.5%
- Companies 11-20 also missing traditional storage companies
 - Oracle appears at 11 and Fujitsu at 14
 - Statistics from: <http://lwn.net/Articles/451243/>

The Life Span of a Linux Enhancement

- ❑ Origin of a feature
 - ❑ Driven through standards like T10 or IETF
 - ❑ Pushed by a single vendor
 - ❑ Created by a developer
- ❑ Proposed in the upstream community
 - ❑ Prototype patches posted
 - ❑ Feedback and testing
 - ❑ Advocacy for inclusion
- ❑ Move into a “free” distribution
- ❑ Shipped and supported by an enterprise distribution

The Long Road: pNFS and Linux

- ❑ Original work done in research groups
 - ❑ Original commercial versions from Panasas and EMC ten years ago
- ❑ Standardized by IETF
 - ❑ Massive specification
 - ❑ Size and complexity a challenge for vendors
- ❑ Overview of pNFS:
 - ❑ [Http://www.pnfs.com](http://www.pnfs.com)

pNFS and Linux

- ❑ Started development as a very large, out of tree patchset
 - ❑ Pushed actively by Panasas, CITI, NetApp and EMC
- ❑ Met lots of upstream resistance
 - ❑ Complicated code
 - ❑ Resistance to supporting multiple layouts
 - ❑ No open source server for testing
- ❑ Developers responded to criticism and pulled in a broader coalition
 - ❑ Fedora integration of prototype kernel & tools
 - ❑ Red Hat is pushing pNFS file layout into a tech preview item for RHEL6
 - ❑ Community working with industry partners to test both upstream and enterprise versions

pNFS Lessons Learned

- Specification was massive
 - IETF working group going for smaller (yearly) spec updates
 - Too large to digest easily in a minor update
 - Chicken and egg dependencies between vendors and operating systems vendors delayed deployment
- Working with linux kernel developers requires allowing them to test
 - Upstream, open source server helps a lot
 - Aggressive testing and involvement by vendors critical
- Vendors carry a lot of responsibility when expensive, pre-production hardware is required
 - Need to test and give feedback actively
 - Best to have kernel developers on staff to help engage and motive other upstream developers

Community Driven Feature: Discard Support

Linux Discard: TRIM/UNMAP

- ❑ Feature informs storage devices of unused ranges
 - ❑ Many SSD devices support ATA TRIM
 - ❑ Several enterprise SCSI arrays support either UNMAP or WRITE_SAME/UNMAP
- ❑ Linux support introduced in several phases
 - ❑ User space tool by Mark Lord (wiper.sh)
 - ❑ Fine grain discard for ext4
 - ❑ Bulk discard

History of Discard Support

- ❑ First appeared in the kernel in 2.6.28 (December 2008)
 - ❑ Look for `blkdev_issue_discard()` and `sb_issue_discard()` in `include/linux/blkdev.h`
 - ❑ Only used by ext4 and FAT code in 28 kernels
 - ❑ Very few devices to test on
- ❑ Many enterprise distros branched from 2.6.32
 - ❑ Added support for discard in btrfs and gfs2
 - ❑ Stable .32 series is close to branch & lacks new features
- ❑ New provisioning types supported in 2.6.39
 - ❑ Explicit checks added by Martin Petersen for updated specifications
 - ❑ Backported into RHEL6 and other distros

Discard Lessons

- ❑ Code was created from spec
 - ❑ Introduced when few devices supported low level commands
- ❑ Discard requirements
 - ❑ SPC-3 support for SCSI
 - ❑ ATA devices must report compliance with at least ATA/ATAPI-7 and have bit 0 of the IDENTIFY DEVICE word 169 set
 - ❑ <http://oss.oracle.com/~mkp/docs/linux-advanced-storage.pdf>
- ❑ Lots of testing done on donated S-ATA SSD's
 - ❑ Many generous donations from vendors
- ❑ Very few enterprise arrays tested early
 - ❑ Kernel developers have limited access, especially to new firmware revisions
- ❑ Testing and feedback on early kernels is key!
 - ❑ Send email to lists with results or patches

Vendor Driven Feature: IO Stack Performance for PCI-e SSDs

Support for PCI-e SSD Devices

- ❑ Sparked by Fusion-io
 - ❑ First card that got the communities attention that could drive hundreds of thousands of IOPS
 - ❑ Binary driver hides most of the functionality
 - ❑ It is a block level driver above the SCSI stack
- ❑ Other vendors joined in soon
 - ❑ Micron's has an open source mtip32xx driver for their p320h part are queued up for 3.2
 - ❑ Some attempt to be AHCI devices, others stayed at the block level
- ❑ Getting consensus on driver placement would be great
 - ❑ NVM Express to the rescue?

Performance Bottlenecks

- ❑ Several performance bottlenecks need resolved
 - ❑ Jens Axboe overview of issues in <http://lwn.net/Articles/408428>
- ❑ Performance tweaks for non-rotational devices
 - ❑ IO scheduler reworked to avoid plugging queue
 - ❑ Timeout per queue instead of timeout per IO request
 - ❑ Avoid per IO calls to `add_timer_randomness()` (used to help generate random numbers)
- ❑ Lots of work still ongoing
 - ❑ Multiqueue work might provide a true block driver model

High Speed SSD Lessons Learned

- ❑ An example of storage vendors taking the lead
 - ❑ Intel dedicated several full time kernel developers
 - ❑ Fusion-io hired senior kernel engineers (Jens Axboe and Nick Piggin) to drive performance scalability
- ❑ Vendors provided analysis and improved code
 - ❑ Tested and collaborated with commercial distros
 - ❑ Pushed generic work through upstream process
 - ❑ Donation of hardware to community for testing

Getting Your Companies Features into Linux!

Storage Companies & Linux

- ❑ It can be really challenging for storage companies to work in the open source world
 - ❑ Ferocious competition in the storage space between vendors
 - ❑ History of secrecy and proprietary development
 - ❑ Lots of litigation between vendors!
- ❑ At the same time, many storage companies leverage Linux aggressively
 - ❑ Lots of Linux based storage appliances
 - ❑ Some vendors are quite effective at moving their concerns through Linux

Learn to Test & Respond

- ❑ Most companies can be convinced to let developers test new upstream kernels
 - ❑ Test new features for compatibility with your products
 - ❑ Respond to proposed patches with basic performance results
- ❑ Important to stay informed
 - ❑ Read lwn.net
 - ❑ Subscribe to linux-scsi, linux-ide, etc
 - ❑ Participate in community events like LinuxCon or the Linux Storage and File System workshop

Active Partnership

- ❑ Contribute hardware to key developers or Linux distributions
 - ❑ If developers do not have hands on access, your parts will not get much attention
 - ❑ Clear cost benefit to providing hardware to developers
- ❑ Develop formal partnerships with Linux distributions
 - ❑ Can get you access to alpha and beta releases
 - ❑ Access to roadmaps
 - ❑ Kind of like working with any OS company!

Become Kernel Developers

- Work with Corporate Legal
 - Need clear legal guidelines on how to contribute to open source
 - Need to be able to “sign off” on contributions to upstream
- Respond to postings or new features
 - Provide patches early in the discussion
 - Provide open source drivers
- Participate actively in Linux specific events
 - Present results at conferences
 - Buy beer and pizza for other developers!

Moving Code to Customer

- ❑ Upstream Code is not always in Distros
 - ❑ Red Hat forks every 2-3 years from upstream
 - ❑ Red Hat developers monitor their areas of interest upstream and backport
- ❑ Inclusion in RHEL
 - ❑ Work in the Fedora community to integrate and test
 - ❑ Develop formal partnership with RHEL
 - ❑ Talk to mutual customers about new features you want in
 - ❑ Loan equipment to Red Hat or provide active feedback
- ❑ Other distros have similar paths (SuSE, Canonical, etc)

Questions?

- ❑ Resources for kernel developers:
 - ❑ Lwn.net
 - ❑ Mailing lists like linux-scsi, linux-ide, linux-fsdevel, etc
- ❑ Linux focused events
 - ❑ LSF workshop
 - ❑ LinuxCon open invitation conferences
 - ❑ Linux Plumbers