Your File System

Upcoming OpenAFS Releases
Feature and Performance Enhancements

Jeffrey Altman, President

Your File System Inc.

20 April 2010

OpenAFS Roadmap! Not a Wish List

- At Fall HEPIX OpenAFS committed to a road map of deliverables over the next two years.
  - 1.6 Spring/Summer 2010
  - 1.8 Fall/Winter 2010
  - 2.0 Spring/Summer 2011
  - 2.x Fall/Winter 2011
- An aggressive schedule to say the least. Especially given the commitments. After yesterday's presentations let us discuss the contents of the deliverables.

OpenAFS 1.4

- The last major update of OpenAFS (1.4) was announced on its 5th birthday, 1 November 2005 four years after 1.2.
- The release took almost four years to develop and included:
  - Significant performance and stability improvements
  - Server support for mobile clients and NATs
  - Audit logging
  - vos copy, vos convertROtoRW, parallel attach on restart
  - Windows clients that worked
  - AIX5, HPUX11.23, Solaris 10, Linux 2.6, MacOS 10.4

OpenAFS 1.6

- Its been more than four years. 1.4.x releases have received many bug fixes and even some new features and performance improvements but major change has all been held back for 1.6.
- Other than Windows which is always using the 1.5.x series for production.
- What has taken so long?
- Source Code Quality and Demand Attach File Service

1.6: Source Code Quality

- When 1.5 was branched there were close to 20,000 warnings produced as part of the x86 MacOS X build
- Today it is possible to build the entire source tree excluding 21 files without warnings
- In the process hundreds of real bugs were fixed
- As was evident from 1.2 instability, there were many lock safety issues resulting in race conditions. Today there are many fewer.
- Prior to the release of 1.6, YFS Inc. will complete a regression test harness that will permit the testing of failure cases in addition to those that are expected to succeed.

1.6: Rx Performance Improvements

- Packet leaks, free packet queue management
- MTU size negotiation failures
- RTT calculation errors
- Unnecessary lock contention
  - Rx statistics
  - NewCall vs EndCall
  - All Write and Read paths
- Races due to improper locking
- Window size errors
- Transmit queues dumped packets on the floor
- NAT Keep-alive support
- > 260MB/second per Rx connection

1.6: Linux Cache Manager

- Performance improvements
  - See accompanying slide deck from Simon Wilkinson
- Dynamic allocation of AFS kernel cache entries to support inotify()-pinned entries
- Path MTU detection

1.6: MacOS X Cache Manager

- Many Finder Improvements
  - Authentication events now refresh
  - Insert only dropboxes
- Improved installation experience
  - GUI queries for local cell information
- AFS Command Preferences Pane
  - Kerberos v5 ticket renewal
- Growl notification service integration
- Significant Rx event handling improvements
- Bulk-stat RPC support for faster directory enumeration

1.6: Demand Attach File Service

- an enhanced volume management library that supports:
  - lock-less I/O
  - on-demand attachment of volumes
  - parallel shutdown of the file server
  - on-line salvaging of volumes
  - automatic detachment of inactive volumes
- a new salvageserver daemon which can salvage volumes on-demand
- a modified bos and bosserver
  - fileserver state saving and restoration
    - host and callback state

1.6: Other

- Major Documentation Improvements
- NFS -> AFS translator for Linux
- DNS SRV record support (replaces AFSDB records)
- /afs/.:mount/cell:volume[:vnode:uniq] direct object access
- Larger than 2TB partitions (1.4 backport)
- Tivoli X/Open Backup API
- Libuafs (userland afs cache manager library)
- AIX6, FreeBSD7.x,8.x, Solaris11, …

1.6: Microsoft Windows

- Nothing new for 1.6. Everything is already in 1.5.74
- Support for all existing operating systems from Windows 2000 to Win7/2008-R2
- Fine grained locking everywhere
- Performance is bound by the SMB implementation
- Unicode character set support

- Native client running on my Win7 laptop to be integrated into 1.7.

What happens Post 1.6?

- When 1.6 branch is cut for release candidates, the master branch becomes 1.7
- All major submissions ready for 1.8 will begin to merge onto the master
- In order for this to happen in an orderly fashion, projects must be able to break their code into small patch sets for submission to http://gerrit.openafs.org/
  - One change per patchset
  - Each patchset reviewable in less than an hour
  - No patchset may break the build or reduce stability
- Documentation to reviewers describing the protocol changes, architecture, and patch submission plan is strongly advised.

1.8 Feature Targets

- Heimdal crypto replaces OpenAFS crypto
- rxk5 security class
- Object storage
- Native AFS redirector client for Microsoft Windows (no support for Windows 2000)
- Rx UDP performance improvements
  - Window Size Negotiation*
  - Dynamic Retransmit Calculation*
  - Path MTU Discovery
  - Large Data Buffers
  - Improved Jumbograms
  - Max Call Negotiation

1.8: More Feature Targets

- PTS authentication name extensions
  - Kerberos v5 and extendible to other name forms (GSS, X.509, SCRAM, …)
- Extended callbacks
  - Significant reductions in network traffic
- More Linux Cache Manager enhancements
  - Byte Range Locking
  - Direct and Synchronous I/O
  - Demand Prefetching
- Pthreaded Ubik servers

2.0: Feature Targets

- rxgk security class
  - Kerberos v5, X.509 and SCRAM authentication
- Protection of anonymous connections
- Protection of the server to client callback connection
  - Permitting full use of Extended Callbacks
    - Metadata changes can be sent from server to clients as part of the notification avoiding even more network traffic and reducing cross-client change contention
- File server coordinated byte range locking
- Whatever else is ready based on work from YFS,Inc and others

Lets not forget about kafs

- David Howell’s kafs is the AFS client built into the Linux kernel that is integrated with fscache and is license compatible
- OpenAFS wants kafs to succeed and replace the OpenAFS cache manager for Linux for many use cases
- A goal is to permit use of either kafs or OpenAFS kernel modules in combination with OpenAFS servers and userland tools
- kAFS is in the Linux kernel repository where it won’t be broken and can take advantage of all future Linux kernel file system improvements
- OpenAFS is once again supporting kafs development through Google Summer of Code
- Tell Red Hat that you want kafs

EU Wish List for OpenAFS

- Many things on the EU wish list are funded by U.S. Dept of Energy but may not be available in OpenAFS until 2012:
  - Large Directory Support
  - TCP based Rx transport
  - Read/write replication
  - Faster replication between read/write site and readonly sites

Unfunded Wish List

- Many things are not funded and not on the roadmap
  - Direct vicep access for Lustre or dCache
  - dCache as an OSD backend
  - Faster metadata performance in the file server backend
  - Improved Fetch/Store Data RPCs
    - Scatter / gather variants
    - Fetch Data with Hash
      - Avoid retransmitting data that is already valid in the cache
      - Multiple writers use-case
  - More File Servers per cell
  - Unix CM Profiling and use of Fine Grained Locking to improve concurrency
  - Direct to object mount points
  - On-the-fly volume splitting and / or striping
  - LDAP backend for Protection Server
  - Native Windows client
    - Initial version in 1.8 but there are many improvements that can be implemented
  - AFS Explorer Shell integration
  - AFS PAGs for MacOS X
  - ZFS specific backend for AFS File Server
  - Disconnected AFS Usability Improvements
  - Performance Monitoring Instrumentation
  - Extended Attributes and Multiple Data Streams

How to Move from Wish List to Road Map Targets?

- There is not enough money nor developers to implement all of the functionality in the next two years
- Implementation designs and Cost/Time estimates for each of the proposals must be developed
- Priorities need to be determined not only by the funders desires but should include what the OpenAFS leadership believes is necessary to further adoption
- This must include client side usability improvements
  - User Shell integration (Explorer, Finder, Gnome, …)
  - Porting Network Identity Manager to Linux and MacOS

OpenAFS Governance is Key

- Incorporation or Joining an Umbrella organization is blocked by the IBM trademarks of “AFS” and “OpenAFS”
- Once the necessary permissions for use are obtained, the not-for-profit corporation must be formed so that funds can be raised and pooled efficiently
- Priorities would be set via a Technical Advisor Board (TAB) consisting of all large contributors, representatives of medium sized contributors, and representatives of individual users and developers
- Gatekeepers would be advisors to the TAB providing expert review of proposals and producing architecture design documents
- The corporation would issue RFQs to find developers to implement the approved designs, communicate with the standards communities, and manage the contractors
- The Gatekeepers would be compensated for their time and an Executive Director would be hired to handle administrator functions

In the meantime, Your File System, Inc. is available

- To research implementation designs
- To develop cost and time estimates
- To implement proposals that are agreed upon by the community
- To enter into support agreements to assist in-house developers, system administrators, and end user help desks

YFS Suggested Order of Implementation

1. All implementation estimates
2. vicep access
3. Extended Attributes and Multiple Data Streams
4. Windows AFS redirector enhancements
5. AFS Explorer Shell extensions
6. OSX pag implementation
7. Userspace Cache Manager for use with FUSE
8. Improved performance of metadata operations
9. dCache as OSD backend
10. Improved ZFS server integration
11. Direct-to-object mount points
12. Unix cache manager profiling and performance improvements
13. Other

Contact Info

- Jeffrey Altman
- President
- Your File System Inc.
- jaltman@your-file-system.com
- +1 212 769-9018

YourFileSystem