



Your File System

Next Generation Storage Clouds

Jeffrey Altman, President

Your File System Inc.

29 September 2009

OpenAFS Roadmap? Or Wish List?

- At every Workshop and Conference a roadmap is presented
 - but its not a roadmap
 - No commitments
 - No delivery dates
 - How are you supposed to plan your rollout schedule?
- The problem is lack of resources
- The Gatekeepers and Elders compile a list of requests but have little influence on what people work on

HEPiX Fall 2007 was a Wake Up Call

- OpenAFS was unable to provide:
 - Commitments
 - A Delivery Schedule
 - A list of what was being worked on
- The message was received loud and clear:
 - Do not ask others to help you until you can prove that you can help yourselves

YFS Inc. Founded to Drive Demand Globally Accessible File Systems

- Open source projects are funded by organizations that are dependent upon the technologies
- The benefits of AFS are lost of the vast majority of the world
- MobileMe, BigVault, and similar sync and access cloud storage services are far behind the capabilities of AFS
- YFS will provide services directly to home, small business, and enterprise users and indirectly through telecommunication companies
- With hundreds of millions of users, there is a business case for enhancing the software on a regular basis

The Mission

- Develop, deploy, and operate “Write once, Access Manywhere™” global storage solutions
- Support the on-going development of critical path open source technologies

U.S. Department of Energy Small Business Innovative Research Grant

- The U.S. DoE labs are large users of AFS:
 - support their HEPiX research
 - provide global access to home and project data
 - distribute and manage applications
- YFS Inc. applied for a grant in 2007
- In 2008, received \$99,000 to fund Rx improvements and a feasibility study
- In August 2009, awarded US\$648,000 to design, standardize, and implement core protocol enhancements
- All grant work to be contributed to OpenAFS

YFS Requirements

- Server scalability (~60,000 clients per server vs ~1000)
- Networking Improvements
 - 10Gbit networks
 - IPv6
 - TCP and/or SCTP in addition to UDP communications
- Optimized file change notification protocol
- Read/write replication in addition to read-only replication
- Server based virtual query volumes

YFS Requirements

- Directory improvements
 - Internationalization, Extended Attributes, Multiple Data Streams per Object
- Mandatory locking
- End-to-end Security
 - AES-256 encryption
 - Both Kerberos and X.509 certificates for authentication
 - Per Service Keys
 - Anonymous Client Access is Protected
 - Secure Callback Channels

YFS Phase I Success

- See openafs-info archive 10/2/2008 e-mail
- Rx Packet Management Issues addressed in 1.4.8 and 1.5.53
- 1.4.8 Rx stack is capable of 124MB/sec over a 10Gbit link

YFS Phase II First Year Road Map (August 2010 deliverables)

- Rx Improvements
 - Path MTU Discovery
 - Large Data Buffers
 - Improved Jumbograms
 - Window Size Negotiation
 - Dynamic Retransmit Calculation
 - Max Call Negotiation
 - Asynchronous API
 - TCP Transport
- Protection Service
 - Anonymous Machine Accounts
- Ubik enhancements
- RxGK Security Class
- Client Improvements
 - Byte Range Locking
 - Direct and Synchronous I/O
 - Demand Prefetching

YFS Phase II Second Year Road Map (August 2011 Deliverables)

- Server Improvements
 - Event driven workflow
 - Posix Ext. Attr. backend
 - Service Port Independence
 - Split Horizon Support
 - Volume Release Optimizations
 - Read Write Replication
- IPv6 Support
- Extended Attributes
- Partition UUIDs
- Long Volume Names
- Per File ACLs
- Directory Format Improvements
 - Unicode
 - Alternate Streams
 - DOS Names
 - DOS Attributes

File System Comparison

| CRITERIA | Volume Management | Filesystem snapshots | POSIX Extended Attributes | Transport | Scalability | Performance |
|----------------------|--------------------------------|----------------------------|---------------------------|----------------------------|---|--|
| OPENAFS | Yes | Limited | No | UDP IPv4 | Yes | Moderate |
| OPENAFS NOTES | Transparent movement of data. | Typically one "backup". | | TCP support planned. | Thousands of clients per server in practice. | No parallel access today. Limited by transport. |
| LUSTRE | No | No | Yes | TCP IPv4 | Yes | High |
| LUSTRE NOTES | Online data migration planned. | Planned for 3.0. | | | 30000 clients per node. | Optimized; Uses object-based storage. |
| NFS V4 | Extension | No | Yes | TCP | Yes | Varies |
| NFS V4 NOTES | Not always available | | | IPv6 not widely available. | | pNFS extension, TCP allow good performance. |
| ZFS | Yes | Yes | Yes | N/A | N/A | High |
| ZFS NOTES | | | | Local only. | | Uses mirroring and striping to achieve high bandwidth. |
| YFS | Yes | Limited | No | UDP, TCP; IPv4, IPv6 | Yes | High |
| YFS NOTES | Striping; Q3 2011 | More than OpenAFS; Q3 2010 | Q3 2011 | TCP Q3 2010; IPv6 Q3 2011 | Asynchronous threading model; 60,000 clients / server Q3 2010 | Transport, threading, OSD; Q3 2010-11 |

File System Comparison (cont.)

| CRITERIA | Locking | Replication | Object Storage Integration | Security | Authentication | Open Source | Commercial Support |
|----------------------|-------------------------|-----------------------------|----------------------------|----------------------------------|--------------------------------|---------------------------------------|---|
| OPENAFS | Advisory | Read-Only | No | Yes | Yes | Yes | Yes |
| OPENAFS NOTES | Whole file only. | Read-Write planned. | Integration to begin soon. | 56 bit fcrypt. K5crypto, 2010 | Kerberos 4 and Kerberos 5. | IBM Public License V1.0. | Linux Box Secure Endpoints Sine Nomine Associates |
| LUSTRE | Yes | Local | Yes | No | No | Yes | Yes |
| LUSTRE NOTES | No lockf / flock yet. | RAID, not multi-server yet. | That's largely the point! | Planned for 1.8. | Kerberos support in Lustre 1.8 | GPL. | ClusterFS (now Sun). |
| NFSV4 | Yes | Extension | Extension | Yes | Yes | Available | Yes |
| NFSV4 NOTES | Mandatory and Advisory. | Not widely available. | In pNFS/NFS v4.1. | GSSAPI RPC. | GSSAPI / Kerberos 5. | Citi reference implementation is GPL. | Typically from OS vendor. |
| ZFS | Yes | Manual | Extension | N/A | N/A | Available | Yes |
| ZFS NOTES | Mandatory and Advisory. | Using zfs send/receive. | Block-based ZFS. | | | | Typically from OS vendor. |
| YFS | Yes | Read-Write & Read-Only | No | Yes | Yes | Yes | Yes |
| YFS NOTES | Q3 2010 | Q3 2011 | Q3 2011 | k5crypto, Q3 2010 | GSSAPI / Kerberos 5; Q3 2010 | IBM Public License V1.0 + BSD | YFS |

The Development Team

- Jeffrey Altman
- Matt Benjamin (Linux Box)
- Derrick Brashear
- Chris Clausen
- Tracy Di Marco White
- Ken Hornstein
- Peter Scott
- Marshall Vale
- Simon Wilkinson

Open Source is a Commitment

- Open Design
- Open Standardization
- Open Implementation
- Open Contributions

- Public git and gerrit instances will be provided

- All externally funded projects will be contributed to OpenAFS upon completion under a BSD license

Contact Info

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



Your File System