OrangeFS : Advancing PVFS

Michael Moore David Bonnie Becky Ligon Mike Marshall
Walt Ligon Nicholas Mills Elaine Quarles Sam Sampson
Shuangyang Yang Boyd Wilson
Clemson University *Omnibond Systems*

1 Work In Progress

PVFS [6] is a frequently used high performance, parallel file system that was developed to meet research and high performance data transfer needs. The resources to develop PVFS into a production quality file system were not available. A second code branch of PVFS, called Orange, was created and cooperatively developed with PVFS over the last two years. The work on the Orange branch lead to the release of OrangeFS [4] which aims to add production quality features to the open source, community driven code base. The features of a production quality file system are not only needed to meet changing research and high performance data transfer demands but also to support large scale file systems and make parallel file systems available in more traditional computing environments. The features currently under development and planned include:

Metadata and data redundancy

Online configuration and file system repair

Capability-based access control

Windows and WebDAV clients

Distributed directories

Commercial grade services

As computing environments have grown the ability to meet storage demands has required dynamically growing or modifying file systems while they remain available. For file systems with the capability to scale out, such as OrangeFS, this means not only the addition of new servers but the ability to dynamically change servers already participating in the file system. In addition to allowing for maintenance and administration this ability is important to mitigate the impact of a single server failure as the number of servers comprising the file system increases. In order to provide resiliency OrangeFS has begun development of file system level redundant metadata and data. The goal is to move OrangeFS from a fault tolerant to a fault accepting architecture. Filesystem level redundancy mitigates the risk that a single server failure leads to a file system failure. Redundancy, specifically providing replication of objects within the file system, allows for future development to address use cases such as off-site mirroring and tiered storage. The underlying implementation required to support redundant metadata and data will also allow for the online changes and repair of the file system.

The need for trusted access to data has become a requirement as the notion of high performance computers has grown to include disparate systems connected together via an untrusted network. In the confines of a single, large system, individual components are equally trusted and usually share a common interconnect. In more traditional computing environments and large scale, shared file systems the clients accessing the file system may not be equally trusted and may not communicate via a single network. To provide trusted access in these environments, OrangeFS has developed an enhanced security implementation. This implementation allows servers to securely sign capabilities for clients, which can be verified when requests are received, to assure the client's authorization to perform a given action.

OrangeFS's Linux implementation is thorough including kernel VFS, FUSE, shared library, and MPI/ROMIO support. However, platforms other than Linux exist and to allow new and different computing environments to leverage OrangeFS we are working to add support for other platforms. Currently, we are developing a Windows OrangeFS client, using Dokan [1], and a WebDAV implementation based on mod_dav [2].

One challenge facing many file systems is supporting large numbers of files, on the order of millions, within a single directory. Although OrangeFS has distributed metadata, the granularity of the metadata distribution is

currently at a directory level. This means that a single server within the file system is responsible for all metadata operations related to a given directory. This situation can cause performance issues for the clients accessing the directory and uneven load across the servers comprising the file system. OrangeFS's experimental release includes an implementation of distributed directories, spreading a directory's metadata amoung all metadata servers in the file system, via the techniques used in Giga+ [5].

One final step in responding to community feedback is commercial grade services around OrangeFS. Support of the open source product is now available via Omnibond Systems [3]. In addition to providing commercial support Omnibond Systems is providing development resources in order to help bring the features discussed above and future features to OrangeFS.

The goal of OrangeFS is to meet the changing demands of file systems in research, high performance, and commodity computing environments. We have begun work on implementing the new features outlined above. We hope the feedback received in presenting our initial work can help guide the technical implementation of, and inform selection of, new features in OrangeFS.

2 Availability

OrangeFS is an open source project and is available for download at

`http://www.orangefs.org/download/`

References

[1] Dokan. `http://http://dokan-dev.net/en/`.

[2] mod_dav: a DAV modules for Apache. `http://www.webdav.org/mod_dav/`.

[3] Omnibond - From Concept to Solution. `http://www.omnibond.com`.

[4] Orange Filesystem. `http://www.orangefs.org/`.

[5] PATIL, S., AND GIBSON, G. GIGA+ : Scalable Directories for Shared File Systems. Tech. rep., Carnegie Mellon University Parallel Data Lab, Oct 2008.

[6] The Parallel Virtual Filesystem. `http://www.pvfs.org`.