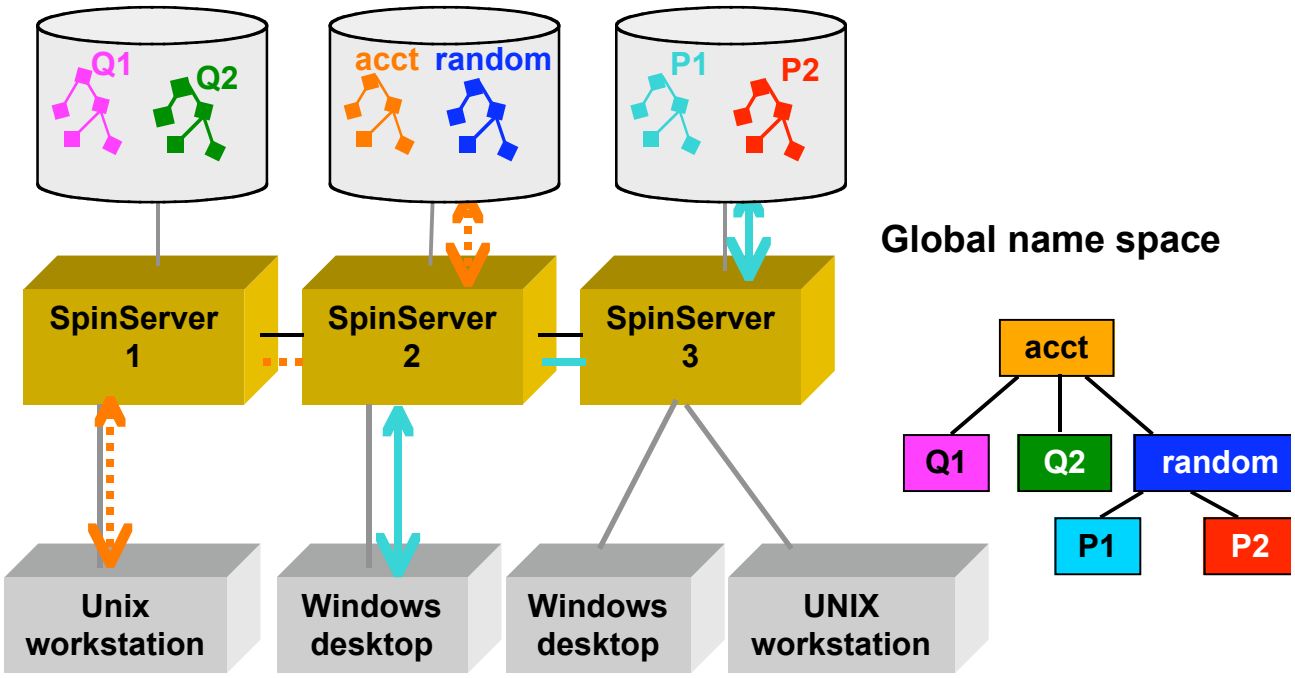
NFS Clustering Position Paper

Michael L. Kazar

Spinnaker Networks' SpinServer system implements a scalable NFS (and CIFS) server based on NAS switching, wherein incoming requests are translated into a back-end file system protocol (SpinFS), and the resulting SpinFS requests are then switched within a cluster of servers to the server holding the accessed data.

The figure below shows an example of a SpinServer cluster, where a global name space is created by "gluing" together several virtual file systems, spread across multiple SpinServers, into a single cluster-wide global name space. Requests enter the cluster at an arbitrary user-facing network port, and are switched among the servers across a private clustering network, typically implemented from the same networking technology as the incoming user-facing network. Today, that networking technology is typically gigabit Ethernet.



This architecture allows the constructing of a scalable cluster, but it requires the standardization of various intra-cluster protocols to allow the construction of heterogeneous clusters, with components provided by different vendors. Note that we expect that support for protocols other than NFS arriving at the user side of the SpinSever cluster will be required, since most NFS servers sold today also include support for protocols such as CIFS at a minimum.

Thus, we expect that additional protocol standardization will have to occur at least in the following arenas:

1. Multi-protocol locking. The cluster protocol should allow the expression of a number of locking protocols, including CIFS operation locks, NFSv4 locks, and those even NFSv3 locks. Because these protocols often store state at the server, in a clustered environment these protocols will need special support to handle the independent failure of individual cluster components. In the example cluster above, if SpinServer1 fails, it will release its CIFS file open state, and notification of this event must be propagated to SpinServer2, to ensure that the corresponding state stored at that server is also released.

2. Data movement. Some of the greatest benefits of a NAS switching system arise from the ability to move data between servers within a cluster. Implementing such move operations requires preserving file handles between the source and target file systems, so that applications running before and after the move operation do not see any differences in the system. In addition, such a move protocol should allow the passing of opaque file system properties between systems, so as to reduce the likelihood of losing individual file attributes if the data when data is moved between homogeneous system with proprietary features.

3. Dynamic state preservation. File systems typically maintain additional state about a file system beyond the actual contents of the files, directories and ACLs. This additional dynamic state includes file lock state and open state. If this additional state can be moved between servers within a cluster, location transparency within a cluster can be leveraged to yield a system that can be configured dynamically online, a very powerful tool for storage administrators.

4. File system location. Protocols need be standardized for locating named resources in a cluster environment, such as virtual network ports, virtual file servers and virtual file systems.

5. Common cluster and client protocols. In the figure above, the clients do not realize that they are communicating with a clustered NAS server. However, if the NFS protocol is enhanced sufficiently, these clients can communicate directly with the server containing the data being referenced at any instant, avoiding the extra switching communication step, eliminating one hop in the data path.

6. Although not shown in the figure above, this architecture also relatively easily supports striping of an individual file or virtual file system across multiple servers. With the standardization of features describing the striping of files across a collection of servers, these files too could be stored in heterogeneous clusters. This feature is especially useful in environments that require very high bandwidth to a small set of files.

Today SpinFS is a proprietary collection of protocols, but it would be preferable if NFS could be extended as described above to allow the easy construction of heterogeneous multi-vendor clusters.

Spinnaker's position is that standardizing protocols such as those described in points 1-4 can increase the acceptance of NAS switching architectures by ensuring interoperability between solutions provided by multiple vendors.