An Overview of NFSv4

NFSv4.0, NFSv4.1, pNFS, and proposed NFSv4.2 features

June 2012

SNIA

Table of Contents

Figures

Tables

SNIA

Introduction

NFSv4 has been a standard file sharing protocol since 2003, but has not been widely adopted. Yet, NFSv4 improves on NFSv3 in many important ways. In this white paper, we explain how NFSv4 is better suited to a wide range of datacenter and HPC use than its predecessor NFSv3, as well as providing resources for migrating from v3 to v4. And, most importantly, we make the argument that users should, at the very least, be evaluating and deploying NFSv4.1 for use in new projects; and ideally, should be using it wholesale in their existing environments.

The Background of NFSv4.1

NFSv2 and its popular successor NFSv3 (specified in RFC-1813[1], but never an Internet standard) was first released in 1995 by Sun. It has proved a popular and robust protocol over the 17 years it has been in use, and with wide adoption it soon eclipsed some of the early competitive UNIX-based filesystem protocols such as DFS and AFS.

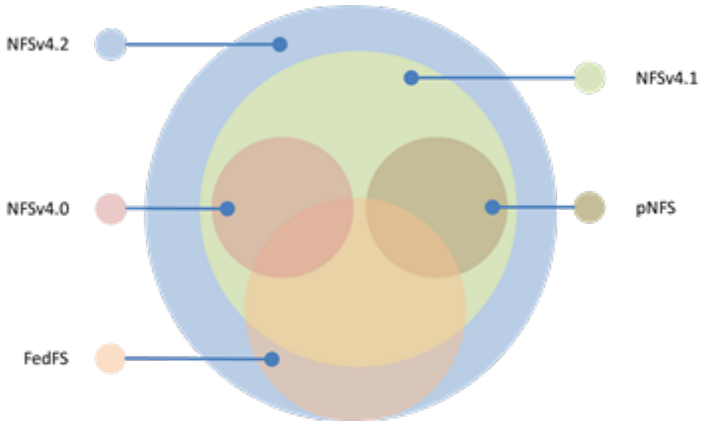

Figure 1; Relationship between NFS Versions

NFSv3 was extensively adopted by storage vendors and OS implementers beyond Sun's Solaris; it was available on an extensive list of systems, including IBM's AIX, HP's HP-UX, Linux and FreeBSD. Even non-UNIX systems adopted NFSv3; Mac OS, OpenVMS, Microsoft Windows, Novell NetWare, and IBM's AS/400 systems. In recognition of the advantages of interoperability and standardization, Sun relinquished control of future NFS standards work, and work leading to NFSv4 was by agreement between Sun and the Internet Society (ISOC), and is undertaken under the auspices of the Internet Engineering Task Force (IETF).

In April 2003, the Network File System (NFS) version 4 Protocol was ratified as an Internet standard, described in RFC-3530, which superseded NFSv3. This was the first open filesystem and networking protocol from the IETF. NFSv4 introduces the concept of state to ameliorate some of the less desirable features of NFSv3, and other enhancements to improve usability, management and performance.

But shortly following its release, an Internet draft written by Garth Gibson and Peter Corbett outlined several problems with NFSv4[2]; specifically, that of limited bandwidth and scalability, since NFSv4 like NFSv3 requires that access is to a single server. NFSv4.1 (as described in RFC-5661, ratified in January 2010) was developed to overcome these limitations, and new features such as parallel NFS (pNFS) were standardized to address these issues.

---

[1] NFSv3 specification: http://tools.ietf.org/html/rfc1813. Other IETF RFCs mentioned in the text can be found at the same site.
[2] The "pNFS Problem Statement": http://tools.ietf.org/html/draft-gibson-pnfs-problem-statement-01

Now NFSv4.2 is moving towards ratification[3]. In a change to the original IETF NFSv4 development work, where each revision took a significant amount of time to develop and ratify, the workgroup charter was modified to ensure that there would be no large standards documents that took years to develop, such as RFC-5661, and that additions to the standard would be an on-going yearly process. With these changes in the processes leading to standardization, features that will be ratified in NFSv4.2 (expected in 2012) are available from many vendors and suppliers today. These relationships are shown in Figure 1.

FedFS[4] (a "federated file system") is a currently unratified standards proposal that provides a set of open protocols that permit the construction of a scalable, federated file system namespace accessible to unmodified NFSv4 clients. Work is in progress; again, some of the features expected in FedFS are already available.

Adoption of NFSv4

While there have been many advances and improvements to NFS, many users have elected to continue with NFSv3. NFSv4 is a mature and stable protocol with many advantages in its own right over its predecessors NFSv3 and NFSv2, yet adoption remains slow. Adequate for some purposes, NFSv3 is a familiar and well understood protocol; but with the demands being placed on storage by exponentially increasing data and compute growth, NFSv3 has become increasingly difficult to deploy and manage.

So What's The Problem with NFSv3?

In essence, NFSv3 suffers from problems associated with statelessness. While some protocols such as HTTP and other RESTful APIs see benefit from not associating state with transactions – it considerably simplifies application development if no transaction from client to server depends on another transaction – in the NFS case, statelessness has led, amongst other downsides, to performance and lock management issues.

NFSv4.1 and parallel NFS (pNFS) address well-known NFSv3 "workarounds" that are used to obtain high bandwidth access; users that employ (usually very complicated) NFSv3 automounter maps and modify them to manage load balancing should find pNFS provides comparable performance that is significantly easier to manage.

Extending the use of NFS across the WAN is difficult with NFSv3. Firewalls typically filter traffic based on well-known port numbers, but if the NFSv3 client is inside a firewalled network, and the server is outside the network, the firewall needs to know what ports the portmapper, `mountd` and `nfsd` servers are listening on. As a result of this promiscuous use of ports, the multiplicity of "moving parts" and a justifiable wariness on the part of network administrators to punch random holes through firewalls, NFSv3 is not practical to use in a WAN environment. By contrast, NFSv4 integrates many of

[3] NFSv4.2 proposed specification: http://www.ietf.org/id/draft-ietf-nfsv4-minorversion2-06.txt; the draft as of November 2011
[4] An overview of FedFS: http://people.redhat.com/steved/Bakeathon-2010/fedfs_fast10_bof.pdf

these functions, and mandates that all traffic (now exclusively TCP) uses the single well-known port 2049.

One of the most annoying NFSv3 "features" has been its handling of locks. Although NFSv3 is stateless, the essential addition of lock management (NLM) to prevent file corruption by competing clients means NFSv3 application recovery is slowed considerably. Very often stale locks have to be manually released, and the lock management is handled external to the protocol. NFSv4's built-in lock leasing, lock timeouts, and client-server negotiation on recovery simplifies management considerably.

In a change from NFSv3, these locking and delegation features make NFSv4 stateful, but the simplicity of the original design is retained through well-defined recovery semantics in the face of client and server failures and network partitions. These are just some of the benefits that make NFSv4.1 desirable as a modern datacenter protocol, and for use in HPC, database and highly virtualized applications.

The Advantages of NFSv4.1

The Gibson and Corbett paper identified some issues with NFSv4 that were successfully addressed in NFSv4.1, and NFSv4.1 is where the focus for end-user evaluation and implementation should be.

References to features in NFSv4.0 apply equally to NFSv4.1, since it was a minor version update, unlike the changes from NFSv3 to NFSv4.

Pseudo Filesystem

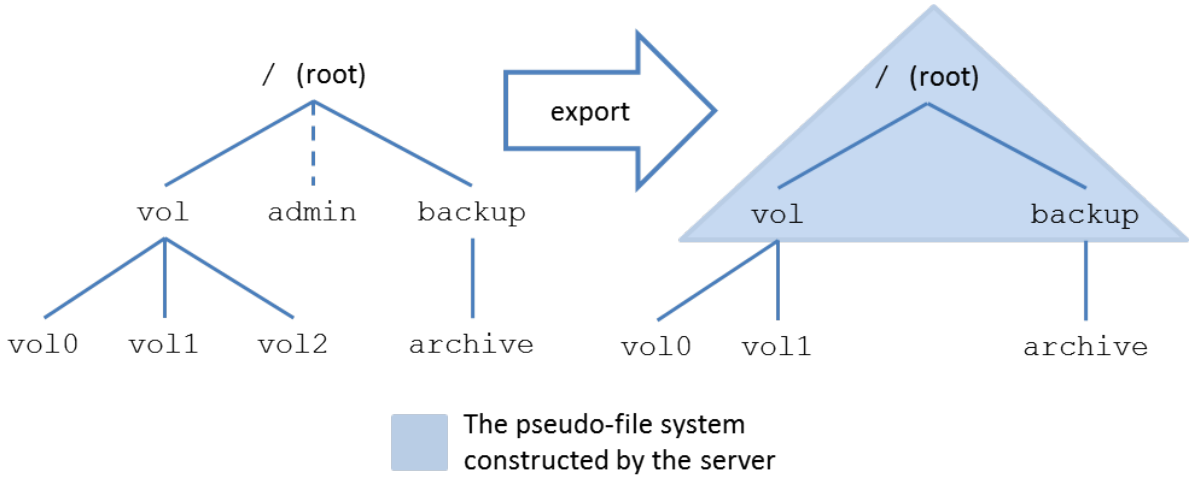

Figure 2; Pseudo File System

On most operating systems, the name space describes the set of available files arranged in a hierarchy. When a system acts as a server to share files, it typically exports (or "shares") only a portion of its name space, excluding perhaps local administration and temporary directories. Consider a file server that exports the following directories:

```
/vol/vol0
/vol/vol1
/backup/archive
```

The server provides a single view of the exported file systems to the client as shown in Figure 2.

In NFSv4, a server's shared name space is a single hierarchy. In the example illustrated in Figure 2, the export list and the server hierarchy is disjoint, and not connected. When a server chooses to export a disjoint portion of its name space, the server creates a pseudo-file system (the area shown in grey) to bridge the unexported portions of the name space allowing a client to reach the export points from the single common root. A pseudo-file system is a structure containing only directories, created by the server, having a unique filesystem id (*fsid*) that allows a client to browse the hierarchy of exported file systems.

The flexibility of the pseudo filesystem as presented by the server can be used to limit the parts of the name space that the client can see, a powerful feature that can be used to considerable advantage. For example, to contrast the differences between NFSv3 and NFSv4 name spaces, consider the mount of the root filesystem `/` in Figure 1. A mount of `/` over NFSv3 allows the client to list the contents of **`/vol/vol2`** as the *fsid* for `/` and **`/vol/vol2`** is the same. An NFSv4 mount of `/` over NFSv4 generates a pseudo *fsid*. As **`/vol/vol2`** has not been exported and the pseudo filesystem does not contain it, it will not be visible. An explicit mount of **`vol/vol2`** will be required.

The flexibility of pseudo-filesystems permits easier migration from NFSv3 directory structures to NFSv4, without being overly concerned as to the server directory hierarchy and layout. However, if there are applications that traverse the filesystem structure or assume the entire filesystem is visible, caution should be exercised before moving to NFSv4 to understand the impact presenting a pseudo filesystem, especially when converting NFSv3 mounts of `/` to NFSv4.

TCP for Transport

Although NFSv3 supports both TCP and UDP, UDP is employed for applications that support it because it is perceived to be lightweight and faster in comparison with TCP. The downside of UDP is that it's an unreliable protocol. There is no guarantee that the datagrams will be delivered in any given order to the destination host -- or even delivered at all -- so applications must be specifically designed to handle missing, duplicate or incorrectly ordered data. UDP is also not a good network citizen; there is no concept of congestion or flow control, nor the ability to apply quality of service (QoS) criteria.

The NFSv4.0 specification requires that any transport used provides congestion control. The easiest way to do this is via TCP. By using TCP, NFSv4 clients and servers are able to adapt to known frequent spikes in unreliability on the Internet; and retransmission is managed in the transport layer instead of in the application layer, greatly simplifying applications and their management on a shared network.

NFSv4.0 also introduces strict rules about retries over TCP in contrast to the complete lack of rules in NFSv3 for retries over TCP. As a result, if NFSv3 clients have timeouts that are too short, NFSv3

servers may drop requests. NFSv4.0 relies on the timers that are built into the connection-oriented transport.

Network Ports

To access an NFS server, an NFSv3 client must contact the server's portmapper to find the port of the **`mountd`** server. It then contacts the mount server to get an initial file handle, and again contacts the portmapper to get the port of the NFS server. Finally, the client can access the NFS server.
This creates problems for using NFS through firewalls, because firewalls typically filter traffic based on well-known port numbers. If the client is inside a firewalled network, and the server is outside the network, the firewall needs to know what ports the portmapper, **`mountd`** and **`nfsd`** servers are listening on. The mount server can listen on any port, so telling the firewall what port to permit is not practical. While the NFS server usually listens on port 2049, sometimes it does not. While the portmapper always listens on the same port (111), many firewall administrators, out of excessive caution, block requests to port 111 from inside the firewalled network to servers outside the network. As a result, NFSv3 is not practical to use through firewalls.

NFSv4 uses a single port number by mandating the server will listen on port 2049. There are no "auxiliary" protocols like **`statd`**, **`lockd`** and **`mountd`** required as the mounting and locking protocols have been incorporated into the NFSv4 protocol. This means that NFSv4 clients do not need to contact the portmapper, and do not need to access services on floating ports.

As NFSv4 uses a single TCP connection with a well-defined destination TCP port, it traverses firewalls and network address translation (NAT) devices with ease, and makes firewall configuration as simple as configuration for HTTP.

Mounts and Automounter

The automounter daemons and the utilities on different flavors of UNIX and Linux are capable of identifying different NFS versions. However, using the automounter will require at least port 111 to be permitted through any firewall between server and client, as it uses the portmapper. This is undesirable if you are extending the use of NFSv4 beyond traditional NFSv3 environments, so in preference the widely available *"mirror mount"* facility can be used. It enhances the behavior of the NFSv4 client by creating a new mountpoint whenever it detects that a directory's *fsid* differs from that of its parent and automatically mounts filesystems when they are encountered at the NFSv4 server[5].

This enhancement does not require the use of the automounter and therefore does not rely on the content or propagation of automounter maps, the availability of NFSv3 services such as **`mountd`**, or opening firewall ports beyond the single port 2049 required for NFSv4.

[5] The Linux "mirror mounting" feature is not specific to NFSv4; NFSv2 and NFSv3 mounts can be configured to act in the same way.

Internationalization Support; UTF-8

In a welcome recognition that the ASCII character set no longer provides the descriptive capabilities demanded by languages with larger alphabets or those that use an extensive range of non-Roman glyphs, NFSv4 uses UTF-8 for file names, directories, symlinks and user and group identifiers. As UTF-8 is backwards compatible with 7 bit encoded ASCII, any names that are 7 bit ASCII will continue to work.

Compound RPCs

Latency in a WAN is a perennial issue, and is very often measured in tenths of a second to seconds. NFS uses RPC to undertake all its communication with the server, and although the payload is normally small, meta-data operations are largely synchronous and serialized. Operations such as file lookup (LOOKUP), the fetching of attributes (GETATTR) and so on, makes up the largest percentage by count of the average workload (see table 1).

NFSv3 Operation	SPECsfs2008
GETATTR	26%
LOOKUP	24%
READ	18%
ACCESS	11%
WRITE	10%
SETATTR	4%
READDIRPLUS	2%
READLINK	1%
READDIR	1%
CREATE	1%
REMOVE	1%
FSSTAT	1%

Table 1; SPECsfs2008 percentages for NFSv3 operations[6]

This mix of a typical NFS set of RPC calls in versions prior to NFSv4 requires each RPC call is a separate transaction over the wire. NFSv4 avoids the expense of single RPC requests and the attendant latency issues and allows these calls to be bundled together. For instance, a lookup, open, read and close can be sent once over the wire, and the server can execute the entire compound call as a single entity. The effect is to reduce latency considerably for multiple operations.

Delegations

Servers are employing ever more quantities of RAM and flash technologies, and very large caches in the orders of terabytes are not uncommon. Applications running over NFSv3 can't take advantage of these

[6] http://www.spec.org/sfs2008/docs/usersguide.html#_Toc191888936 gives a typical mix of RPC calls from NFSv3.

caches unless they have specific application support. With increasing WAN latencies doing every IO over the wire introduces significant delay.

NFSv4 allows the server to delegate certain responsibilities to the client, a feature that allows caching locally where the data is being accessed. Once delegated, the client can act on the file locally with the guarantee that no other client has a conflicting need for the file; it allows the application to have locking, reading and writing requests serviced on the application server without any further communication with the NFS server. To prevent deadlocking conditions, the server can recall the delegation via an asynchronous callback to the client should there be a conflicting request for access to the file from a different client.

Migration, Replicas and Referrals

For broader use within a datacenter, and in support of high availability applications such as databases and virtual environments, copying data for backup and disaster recovery purposes, or the ability to migrate it to provide VM location independence are essential. NFSv4 provides facilities for both transparent replication and migration of data, and the client is responsible for ensuring that the application is unaware of these activities. A NFSv4 referral allows servers to redirect clients from this server's namespace to another server; it allows the building of a global namespace while maintaining the data on discrete and separate servers.

Sessions

Sessions bring the advantages of correctness and simplicity to NFS semantics. In order to improve the correctness of NFSv4, NFSv4.1 sessions introduce "exactly-once" semantics. Servers maintain one or more session states in agreement with the client; a session maintains the server's state relative to the connections belonging to a client. Clients can be assured that their requests to the server have been executed, and that they will never be executed more than once. Sessions extend the idea of NFSv4 delegations, which introduced server-initiated asynchronous callbacks; clients can initiate session requests for connections to the server. For WAN based systems, this simplifies operations through firewalls.

Security

An area of great confusion, many believe that NFSv4 *requires* the use of strong security. The NFSv4 specification simply states that *implementation* of strong RPC security by servers and clients is mandatory, not the *use* of strong RPC security. This misunderstanding may explain the reluctance of users from migrating to NFSv4 due to the additional work in implementing or modifying their existing Kerberos security.

Security is increasingly important as NFSv4 makes data more easily available over the WAN. This feature was considered so important by the IETF NFS working group that the security specification using Kerberos v5[7] was "retrofitted" to NFSv2 and NFSv3 and specified in RFC-2623.

[7] "Kerberos Overview– An Authentication Service for Open Network Systems http://www.cisco.com/application/pdf/paws/16087/1.pdf

Although access to an NFSv2, 3 or 4 filesystem without strong security such as provided by Kerberos is possible, across a WAN it should really be considered only as a temporary measure. In that spirit, it should be noted that *NFSv4 can be used without implementing Kerberos security*[8]. The fact that it is possible does not make it desirable! A fuller description of the issues and some migration considerations can be found in the SNIA White Paper "Migrating from NFSv3 to NFSv4".

Many of the practical issues faced in implementing robust Kerberos security in a UNIX environment can be eased by using a Windows Active Directory (AD) system. Windows uses the standard Kerberos protocol as specified in RFC 1510; AD user accounts are represented to Kerberos in the same way as accounts in UNIX realms. This can be a very attractive solution in mixed-mode environments[9].

Parallel NFS (pNFS) and Layouts

Parallel NFS (pNFS) represents a major step forward in the development of NFS. Ratified in January 2010 and described in RFC-5661, pNFS depends on the NFS client understanding how a clustered

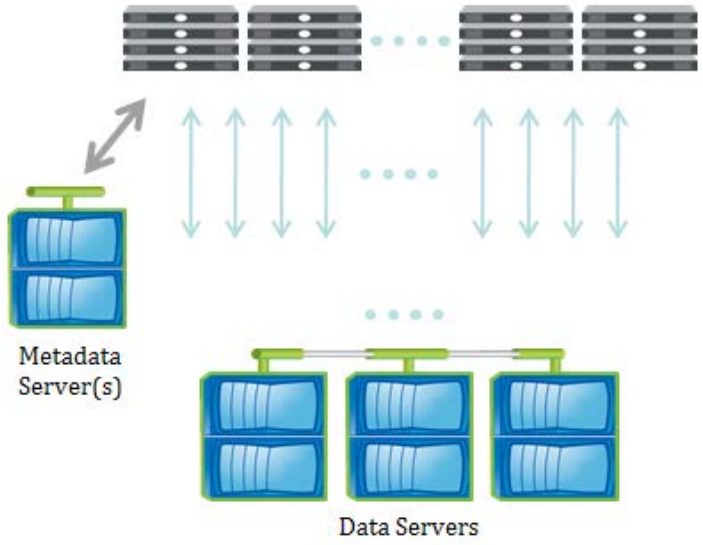

Figure 3; pNFS Conceptual Data Flow

filesystem stripes and manages data. It's not an attribute of the data, but an arrangement between the server and the client, so data can still be accessed via non-pNFS and other file access protocols. pNFS benefits workloads with many small files, or very large files, especially those run on compute clusters requiring simultaneous, parallel access to data.

---

[8] For examples of NFSv4 without Kerberos, see Ubuntu Linux; https://help.ubuntu.com/community/NFSv4Howto and SUSE Linux Enterprise; http://www.novell.com/support/dynamickc.do?cmd=show&forward=nonthreadedKC&docType=kc&externalId=7005060&sliceId=1

[9] Windows Security and Directory Services for UNIX Guide v1.0: http://technet.microsoft.com/en-us/library/bb496504.aspx

Clients request information about data layout from a Metadata Server (MDS), and get returned layouts that describe the location of the data. (Although often shown as separate, the MDS may or may not be standalone nodes in the storage system depending on a particular storage vendor's hardware architecture.) The data may be on many data servers, and is accessed directly by the client over multiple paths. Layouts can be recalled by the server, as in the case for delegations, if there are multiple conflicting client requests.

By allowing the aggregation of bandwidth, pNFS relieves performance issues that are associated with point-to-point connections. With pNFS, clients access data servers directly and in parallel, ensuring that no single storage node is a bottleneck. pNFS also ensures that data can be better load balanced to meet the needs of the client (see figure 2).

The pNFS specification also accommodates support for multiple layouts, defining the protocol used between clients and data servers. Currently, three layouts are specified; files as supported by NFSv4, objects based on the *Object-based Storage Device Commands* (OSD) standard (INCITS T10) approved in 2004, and block layouts (either FC or iSCSI access). The layout choice in any given architecture is expected to make a difference in performance and functionality. For example, pNFS object based implementations may perform RAID parity calculations in software on the client, to allow RAID performance to scale with the number of clients and to ensure end-to-end data integrity across the network to the data servers.

So although pNFS is new to the NFS standard, the experience of users with proprietary precursor protocols to pNFS shows that high bandwidth access to data with pNFS will be of considerable benefit.

Potential performance of pNFS is definitely superior to that of NFSv3 for similar configurations of storage, network and server. The management is definitely easier, as NFSv3 automounter maps and hand-created load balancing schemes are eliminated; and by providing a standardized interface, pNFS ensures fewer issues in supporting multi-vendor NFS server environments.

Some Proposed NFSv4.2 features

NFSv4.2 promises many features that end-users have been requesting, and that makes NFS more relevant as not only an "every day" protocol, but one that has application beyond the data center.

Server Side Copy

Server-Side Copy (SSC) removes one leg of a copy operation. Instead of reading entire files or even directories of files from one server through the client, and then writing them out to another, SSC permits the destination server to communicate directly to the source server without client involvement, and removes the limitations on server to client bandwidth and the possible congestion it may cause.

Application Data Blocks (ADB)

ADB allows definition of the format of a file; for example, a VM image or a database. This feature will allow initialization of data stores; a single operation from the client can create a 300GB database or a VM image on the server.

Guaranteed Space Reservation & Hole Punching

As storage demands continue to increase, various efficiency techniques can be employed to give the appearance of a large virtual pool of storage on a much smaller storage system. Thin provisioning, (where space appears available and reserved, but is not committed) is commonplace, but often problematic to manage in fast growing environments. The guaranteed space reservation feature in NFSv4.2 will ensure that, regardless of the thin provisioning policies, individual files will always have space available for their maximum extent.

While such guarantees are a reassurance for the end-user, they don't help the storage administrator in his or her desire to fully utilize all his available storage. In support of better storage efficiencies, NFSv4.2 will introduce support for sparse files. Commonly called "hole punching", deleted and unused parts of files are returned to the storage system's free space pool (see figure 4).

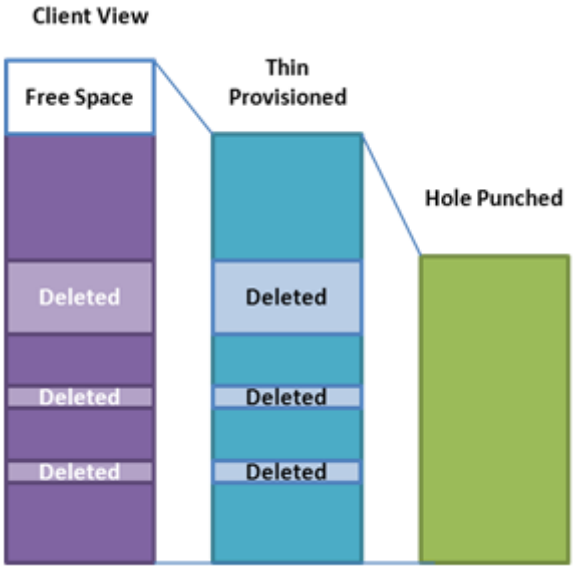

Figure 4; Reservations & Hole Punching

Obtaining Servers and Clients

With this background on the features of NFS, there is considerable interest in the end-user community for NFSv4.1 support from both servers and clients. Many Network Attached Storage (NAS) vendors now support NFSv4, and in recent months, there has been a flurry of activity and many developments in server support of NFSv4.1 and pNFS. For NFS server vendors, refer to their websites, where you will get the latest up-to-date information.

On the client side, there is RedHat Enterprise Linux 6.2 that includes an NFSv4.1 Technical Preview[10], Novell SUSE Linux Enterprise Server 11 SP2 (with NFSv4.1 and pNFS based on the 3.0 Linux kernel), and Fedora 15 and 16, available at fedoraproject.org. Both distributions support NFSv4.1 and files-based pNFS. For the adventurous who wish to build their own kernels and want to explore the latest block or objects access, there is full file pNFS support in the upstream 3.0 Linux kernel, block support in the 3.1 Linux kernel, and object-based pNFS support in the 3.3 Linux kernels.

For Windows, Microsoft has publically indicated that it will be supporting NFSv4.1 in Windows 8. An open-source client implementation preview is available from the Center for Information Technology Integration (CITI) at University of Michigan[11].

Conclusion

NFSv4.1 includes features intended to enable its use in global wide area networks (WANs). These advantages include:

- Firewall-friendly single port operations
- Advanced and aggressive cache management features
- Internationalization support
- Replication and migration facilities
- Optional cryptography quality security, with access control facilities that are compatible across UNIX® and Windows®
- Support for parallelism and data striping

The goal for NFSv4.1 and beyond is to define how you get to storage, not what your storage looks like. That has meant inevitable changes. Unlike earlier versions of NFS, the NFSv4 protocol integrates file locking, strong security, operation coalescing, and delegation capabilities to enhance client performance for data sharing applications on high-bandwidth networks.

NFSv4.1 servers and clients provide even more functionality such as wide striping of data to enhance performance. NFSv4.2 and beyond promise further enhancements to the standard that increase its applicability to today's application requirements. It is due to be ratified in August 2012, and we can expect to see server and client implementations that provide NFSv4.2 features soon after this; in some cases, the features are already being shipped now as vendor specific enhancements.

With careful planning, migration to NFSv4.1 and NFSv4.2 from prior versions can be accomplished without modification to applications or the supporting operational infrastructure, for a wide range of applications; home directories, HPC storage servers, backup jobs and so on.

---

[10] RedHat 6.2 NFSv4.1 Technical Preview: http://docs.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/6-Beta/html/Storage_Administration_Guide/ch12s02.html

[11] NFSv4 Client for Windows at CITI; http://www.citi.umich.edu/projects/nfsv4/windows/

About the Ethernet Storage Forum

The Ethernet Storage Forum (ESF) is the marketing organization within the Storage Networking Industry Association (SNIA) focused on Ethernet-connected storage networking solutions. Through the creation of vendor-neutral educational materials, ESF thought leaders leverage SNIA and Industry events and end-user outreach programs to drive market awareness and adoption of Ethernet-connected storage networking technologies, worldwide. For more information, visit www.snia.org/forums/esf.

About the SNIA

The Storage Networking Industry Association (SNIA) is a not–for–profit global organization, made up of some 400 member companies spanning virtually the entire storage industry. SNIA's mission is to lead the storage industry worldwide in developing and promoting standards, technologies, and educational services to empower organizations in the management of information. To this end, the SNIA is uniquely committed to delivering standards, education, and services that will propel open storage networking solutions into the broader market. For additional information, visit the SNIA web site at http://www.snia.org.

This SNIA whitepaper is based on the article written by Alex McDonald entitled "NFSv4", originally published in ";login: The Magazine of USENIX", vol. 37, no. 1 (Berkeley, CA: USENIX Association, February 2012), pp. 28-35.

SNIA