EMC²

TECHNICAL NOTES

Parallel NFS Feature in NFS v4.1

Technical Notes

P/N 300-000-258
REV 01
August, 2013

This Parallel NFS (pNFS) technical notes document contains information on these topics:

Overview

Parallel NFS (pNFS) comes as an optional feature in the NFS v4.1 standard that allows clients to access storage devices directly and in parallel. The pNFS architecture increases the scalability and performance associated with NFS servers in deployment today. pNFS achieves this by the separating data and metadata, and by moving the metadata server out of the data path.

pNFS brings together the benefits of parallel I/ O with the current standards for network file systems (NFS). This allows users to experience increased performance and scalability in their storage infrastructure with the added assurance that their ability to choose best-of-breed solutions remains intact.

pNFS removes a performance bottleneck in traditional NAS systems by allowing the compute clients to read and write data directly and in parallel, to and from the physical disks. The NFS server is used only to control metadata and coordinate access, allowing incredibly fast access to very large data sets from many clients.

When a client wants to access a file it first queries the metadata server which provides it with a map of where to find the data and with credentials regarding its rights to read, modify, and write the data. After the client has those two components, it communicates directly to the storage devices when accessing the data. With traditional NFS all data flows through the NFS server –pNFS removes the NFS server from the primary data path allowing free and fast access to data. pNFS maintains all the advantages of NFS but eliminates bottlenecks to allow users to access data can in parallel. This allows very fast throughput rates; system capacity scaling that does not impact overall performance.

Terminology

Readers should have general knowledge about the terms listed in Table 1.

Table 1. Terminology

Term	Definition
Metadata	Information about a file system object, such as its name, location within the namespace, owner, ACL, and other attributes. Metadata may also include storage location information, and this will vary based on the underlying storage mechanism that is used.
Metadata Serve	An NFSv4.1 server that supports the pNFS feature. A variety of architectural choices exist for the metadata server and its use of file system information held at the server. Some servers may contain metadata only for file objects residing at the metadata server, while the file data resides on associated storage devices. Other metadata servers may hold both metadata and a varying degree of file data. The VNX Datamover would be the metadata server when using pNFS
pNFS Client	An NFSv4.1 client that supports pNFS operations and at least one storage protocol.
Storage Device	A storage device stores a regular file's data, but leaves metadata management to the metadata server. A storage device could be another NFSv4.1 server, an object-based storage device (OSD), a block device accessed over a System Area Network (SAN, e.g., VNX via Fibre channel or iSCSI).

pNFS on VNX

VNX now has Parallel NFS (pNFS) support for Block via ISCSI and FC, which allows direct client access to the VNX disks containing file data. Significantly higher file access performance can be seen when file data for an NFSv4 server is stored on multiple and/ or higher throughput VNX disks. The relationship among multiple clients, a single server, and multiple VNX disks for pNFS (server and clients have access to all storage devices) is shown in Figure 1.

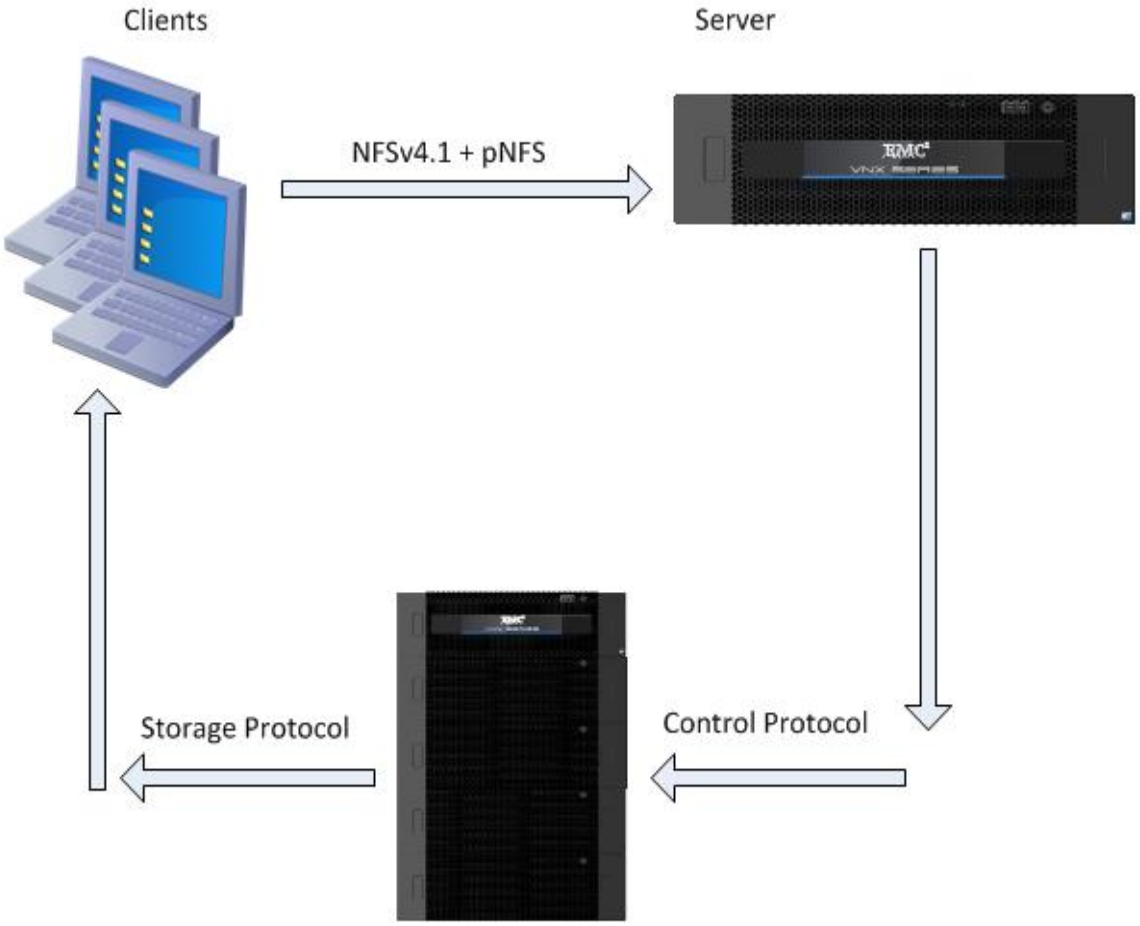

Figure 1. Parallel NFS (pNFS) support for Block via ISCSI and FC

Figure 1 shows the clients, server, and VNX are responsible for managing file access. This is in contrast to NFSv4 without pNFS, where it is primarily the server's responsibility; some of this responsibility may be delegated to the client under strictly specified conditions. The VNX will play the part of the NFS server and the storage system in this model. VNX supports pNFS block server since 2010, and has been optimized for performance.

RFCs

For more information on pNFS it is defined in the following standards:

- RFC 5661--This RFC describes NFS version 4 minor version one, including features retained from the base protocol and protocol extensions made subsequently. Major extensions introduced in NFS version 4.1 include: Sessions, Directory Delegations, and parallel NFS (pNFS).
- RFC 5662--This RFC contains the machine readable XDR definitions for the protocol.
- RFC 5663--This document provides a specification of a block based layout type definition to be used with the NFSv4.1 protocol. As such, this is a companion specification to NFS version 4 Minor Version 1

Supported Clients

Currently only the Fedora Linux operating system has the capability to support pNFS block.

Limitations

These restrictions apply when using pNFS:

- pNFS over iSCSI configurations rely on the iSCSI initiator and the MDS switch or the VNX Block OE as the iSCSI target.
- The Nolock (CIFS) locking policy is the default setting for pNFS; it is also the only locking policy supported on a pNFS-enabled file system.
- Only the Nolock locking policy is compatible with pNFS on a VNX. A server cannot properly mount a file system when a Data Mover is running an incompatible locking policy.
- Before enabling any new features, ensure that the file system is compatible with pNFS.
- pNFS improves performance significantly when large file transfers (sequential I/ Os) are common. pNFS does not greatly benefit a configuration that deals with many small, random I/ Os.
- When both Checkpoint and VNX Replicator™ are active, pNFS system performance is reduced. The performance reduction is caused by additional CPU use and I/ O overhead of the block copy operation.

- A pNFS file system with a stripe size of 256 KB generally achieves optimal performance. To ensure continuous availability of file systems in the unlikely event of a Data Mover failure, configure each pNFS-enabled Data Mover for automatic failover to a standby Data Mover. Refer to the *EMC VNX Series, Release 7.0, Configuring Standbys on EMC VNX* document on the EMC Online Support website at http:/ / Support.EMC.com for more information on configuring a standby Data Mover.
- When exporting a file system on a Data Mover using the server_export command and using the ro (read-only) option, pNFS disregards the read-only option and writes to the file system.
- When a pNFS-enabled file system is extended, using the nas_fs command through the CLI or by using the Unisphere GUI, the server loses the pNFS connection. The server needs to be rezoned to see the added disks; only after the rezoning will the server enable pNFS on the file system. The *EMC VNX Series, Release 7.0, VNX Command Line Interface Reference for File* document provides information on the nas_fs command. The *EMC VNX Series, Release 7.1, Managing Volumes and File Systems on VNX Manually* document provides information on extending VNX file systems. Both documents are located on the *EMC VNX Documentation CD* that is packaged with the VNX and available on support.emc.com.