NFSv4.1/pNFS
Ready for Prime Time Deployment

December 7, 2011
Usenix LISA ’11 - Boston

NFSv4.1 pNFS product community
Value of NFSv4.1 / pNFS

- Industry Standard
- Secure
- Performance and Scale
  - Throughput
  - Increased Storage Capacity (pNFS)
- Manageable
  - Separates namespace (metadata) from data
  - Allows for data movement, tiering, manipulation while providing direct access to the client
pNFS Vendors Status

- EMC
- NetApp
- Panasas
- IBM
- BlueArc
- Microsoft
- dCache
- Tonian
- RedHat
- Novell
- Oracle (Solaris)
Linux Client

- Linux has the first commercial implementation of NFSv4.1 client
- Basic client implementation of NFSv4.1 and pNFS in the upstream mainline kernel
  - Supports all 3 pNFS layouts
  - Emphasis on scalability and feature stability
    - More performance optimisations to come
    - Some features still missing:
      - O_DIRECT over pNFS (coming soon!)
Client supported in 2 distributions:

- Fedora 16 has support for all 3 pNFS layout types (files, objects, blocks)
- Red Hat Enterprise Linux 6.2 has support for the files pNFS client
Linux pNFS project is actively maintained by Tonian.
- Development tree: git://linux-nfs.org/projects/bhalevy/linux-pnfs.git

The project includes the reference implementation of the pnfs server for:
- files: Exporting GFS2 and OCFS2 (DLM based clustered file system)
  - supporting parallel I/O for read access
- objects: Exporting the EXOFS file system.
- blocks: Exporting block-based file systems, such as ext4, xfs, btrfs, etc.

Development appears to be accelerating now that the client is done

Server code to be submitted to the kernel in the coming months
Client support only
pNFS file layout
Insert module into kernel
  - Create `/etc/modprobe.d/dist-nfs41.conf`
  - Add ‘alias nfs-layouttype4-1 nfs_layout_nfsv41_files’
  - Reboot
  - Note: with RHEL6.3 above will not be needed
Mount the file system with “minorversion” mount option
  - E.g. `mount --o minorversion=1 server:/export /mnt`
Client support only
GA end of February 2012
EMC pNFS Block Server Status

- Support for pNFS block server since 2010 – first GA product
- Next EMC VNX release will include pNFS server optimized for performance

pNFS block server performance (from multiple clients with iSCSI) – 900MB/sec
Funding CITI to implement Linux pNFS block client

New pNFS block client patches by EMC developers provide optimizations for performance in Linux Kernel 3.2

pNFS block client performance over iSCSI –
read-100MB/sec; write-90MB/sec
## NetApp NFS Support Matrix

**Announced 21 Nov; ONTAP 8.1 RC2**


<table>
<thead>
<tr>
<th></th>
<th>7.3.x</th>
<th>8.1 7-mode</th>
<th>8.1 C-Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NFS v3</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>NFS v4.0</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>NFS v4.0 with Delegations</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>NFS v4.0 with Referrals</strong></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>NFS v4.1</strong></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>NFS v4.1 with pNFS</strong></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>NFS v4.1 with Referrals</strong></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>NFS v4.1 with Delegations</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>NFS v4.1 with pNFS and Delegations</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
Cluster-Mode – Optimized Data Path with pNFS

- Direct network path to volume
- Layout recalls trigger new network path computation
- Automatic provisioning
- Minimum cluster traffic between nodes
- Faster response time

D-Blade
Network stack, WAFL, Caching
Meta-data, data, Locking

N-Blade
NFS, Network, M-Host
Conversion to SpinNP
VLDB/Name Lookup

SpinNP

Memory Bus

Gigabit Cluster interconnect

Storage

Storage

Meta-data Path

Data Path

pNFS BoF- LISA 2011-12-07
Panasas to ship pNFS in 2012

- Panasas a founding advocate of pNFS standards process, has contributed to Linux client & server code, especially object layout code
- Panasas systems designed from the ground up, anticipating pNFS
  - True scale-out architecture backed by high-performance PanFS file system
  - Today shipping with DirectFlow, precursor to pNFS with 8 years of production use
  - pNFS Objects will be ideal for high throughput applications
Panasas pNFS Scaling

- Panasas has already demonstrated pNFS scaling to 128 clients at multiple gigabytes per second
IBM GPFS

- Fully-symmetric GPFS architecture - scalable data and metadata
  - pNFS client can mount and retrieve layout from any GPFS node
  - Metadata requests load balanced across cluster
  - Direct data access from any GPFS server

- pNFS server and native GPFS clients can share the same file system
  - Backup, deduplication, and other management functions don’t need to be done over NFS

- Beyond client access, will be key part of SONAS Active Cloud Engine
Feature support (not native Windows)
✓ NFSv4.1 sessions
✓ Mandatory and named attributes
✓ Security: RPCSEC-GSS, SECINFO, ACLs
✓ Referrals
✓ Reboot recovery
✓ Locking
✓ Delegations
✓ pNFS sparse and dense layouts

Client GbE performance:
100 MB/sec read, 80 MB/sec write
Features missing

- Session security
  - Machine creds or SSV
- Segmented layouts (whole file only)
- Session trunking on client
Windows Server 8 (native server)

- Base NFSv4.1 only
  - Mandatory aspects of RFC 5661
- Integrated with Windows Failover clustering
- Identity Mapping Support
  - Passwd/group file mapping
  - Active Directory
  - ADLDS or 3rd party LDAP stores (RFC 2307 compliant)
  - User name mapping (legacy)
- RPCSEC_GSS support
  - Krb5, Krb5i, and Krb5p
- Multiprotocol access (SMB / NFS) to same share
- Volume Mount Point Support
"Oracle strongly supports NFSv4.1 and pNFS file and will deliver implementations of both in future releases of Solaris."
Tonian (new vendor) Status

- Tonian is a VC-backed start up founded in 2011 (Charles River Ventures and Cedar Fund)
- Tonian is developing a pNFS-based products for the enterprise market.
- For more information: Benny Halevy <bhalevy@tonian.com>
Getting Started with NFSv4.1/pNFS

Assist user community as NFSv4.1 is tested and deployed
Gather NFSv4.1 practical deployment information on a shared web site
E.g. Opensource toolset for evaluation