NetApp™

Go further, faster™

Parallel NFS
pNFS

Javier Martínez
Technical Manager
NetApp Spain

Agenda

- NFS, history and status
- Parallel NFS servers
- NFSv4.1 and pNFS
- pNFS MetaData Server
- pNFS Data Servers
- pNFS Clients
- References

NFS: Network File System

"NFS protocol provides transparent remote access to shared files across networks"

- NFS v2 1989 – RFC1094 (SUN)
- NFS v3 1995 – RFC1813 (IETF)
- NFS v4 2003 – RFC3530 (IETF)

NFS v3 is the most commonly used today (2008)

Comparison of NFSv3 and NFSv4

NFSv3

- A collection of protocols (file access, mount, lock, status)
- Stateless
- UNIX-centric, but seen in Windows too
- UNIX permissions
- Deployed with weak authentication
- 32 bit numeric uids/gids
- Ad-hoc caching
- Works over UDP, TCP
- Needs a-priori agreement on character sets

NFSv4

- One protocol to a single port (2049)
- Lease-based state
- Supports UNIX and Windows file semantics
- Windows-like access
- Mandates strong authentication
- String-based identities
- Real caching handshake
- Bans UDP
- Uses a universal character set for file names

Some performance numbers of NFS v3

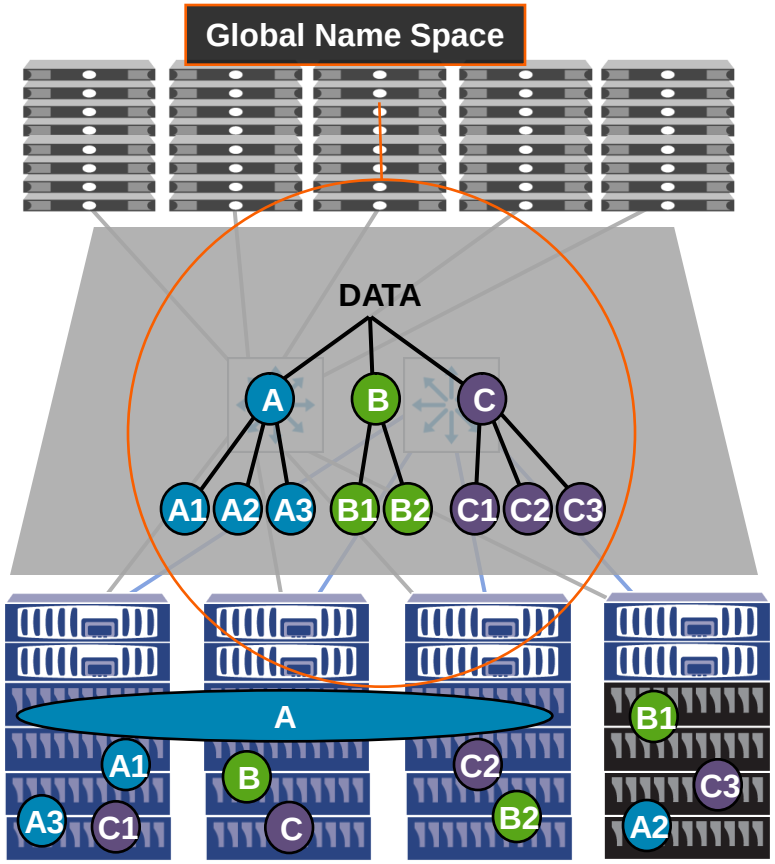
- Most clients limited to 100-200 MB/s due to protocol implementation
- 1 or 2 GbE ports are OK on the clients

- Some servers scale more:
  - 800MB/s sequencial read using 1x 10GbE
  - 82K ops/sec SFS97
    single server, one filesystem, 24 clients
    aprox. 160MB/s read + 60MB/s write

  } 1 NetApp FAS 6080

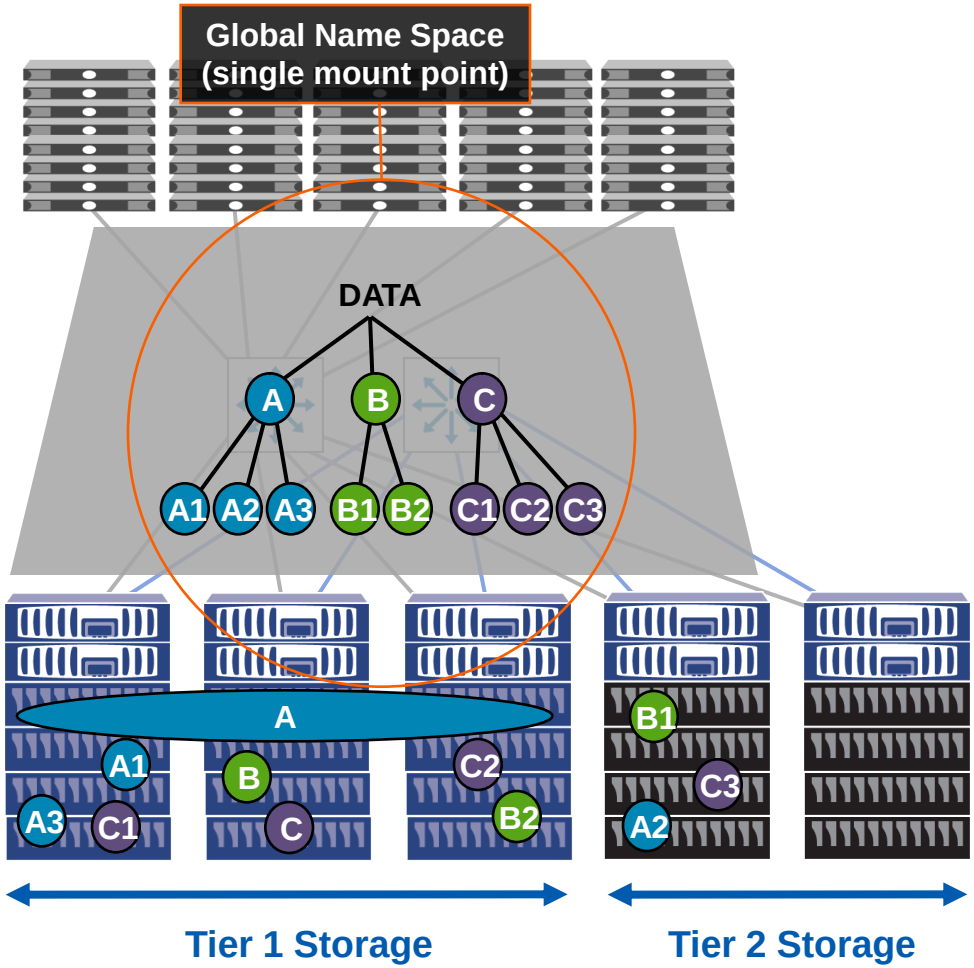
- Parallel servers are available:
  - 1000K ops/sec SFS97
    24 FAS6080 nodes and 216 clients

NFS Parallel Server: Data ONTAP GX



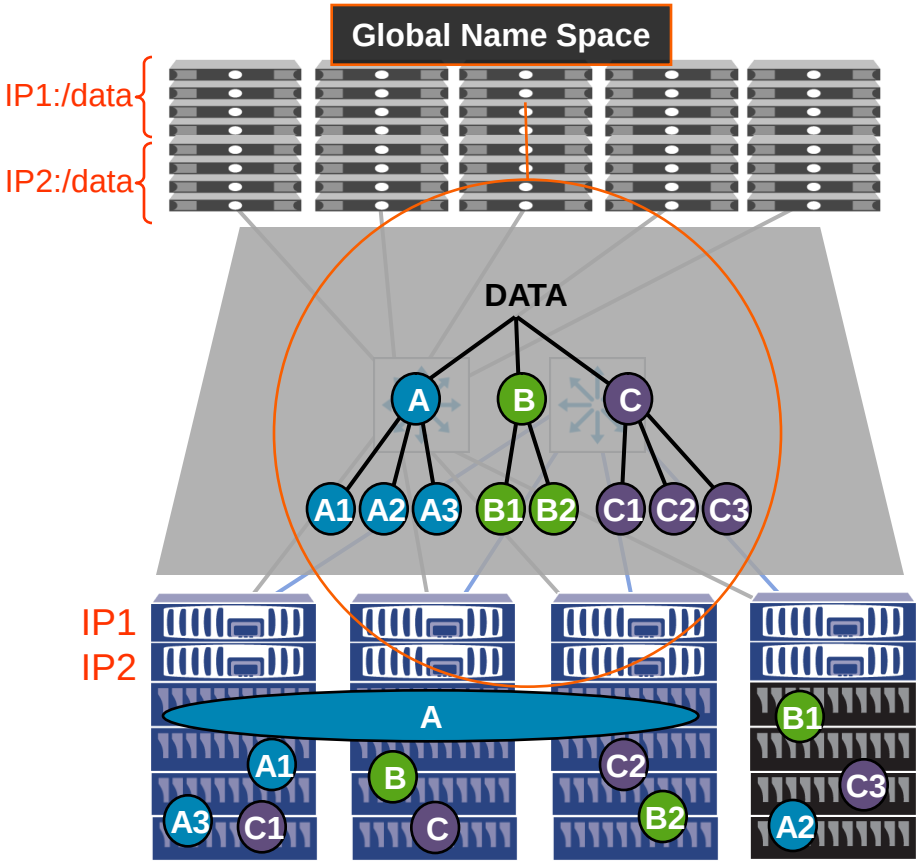
- Several storage controllers working as single system
- Integrated disk management, RAID, filesystem and NFS server in a single device
- HA via controller pairs and shared disks
- Global Name Space to access all servers' data via a single mount point for all clients

Data ONTAP GX features



- Every volume or filesystem is linked in the GNS
- 2 types of volumes:
  - Standard
  - Stripped (A)
- New servers can be added online
- Volumes can be moved or resized transparently
- Different types of controllers and disks for tiering
- + RAID6, Snapshots, …

Data flow in Data ONTAP GX



- Each storage controller exports one or several IPs
- Each client mounts the GNS using one of those IPs
- Local data is served directly
- Remote data is served using the cluster interconnect
- NFS v3 or v4 is a client to one server protocol

What is NFSv4.1?

- A minor version of NFSv4
- Does not modify NFSv4.0
- Delegations on directories, symbolic links, …
- Session model
- pNFS
- Fixes/Cleanups Relative to NFSv4.0
  - Can re-acquire a delegation without re-opening file
  - ACLs even more closely track Windows
  - Exclusive open fixes
  - Referrals Clarifications
- Planned to be closed on Dec 2008

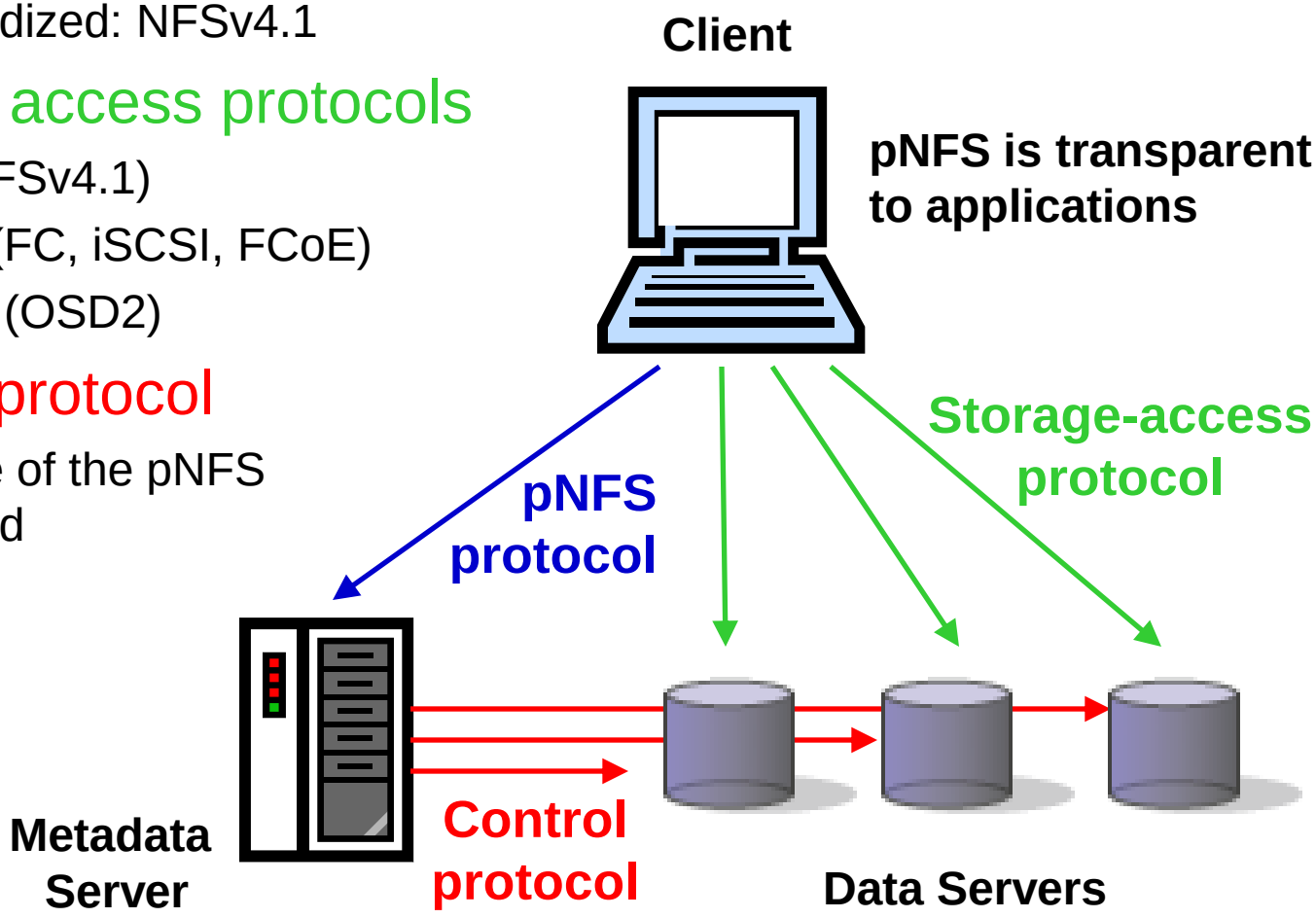
pNFS, a new arquitecture and protocol

- Multiple data servers will provide parallel access to a given filesystem or individual files
  - A single filesystem might be stripped across several servers, either at the file or block level
  - A cluster filesystem is “needed”

- NFS clients will be aware of which data servers have the data for file’s given byte range
  - Eliminates a single server as a throughput and latency bottleneck

pNFS data flow

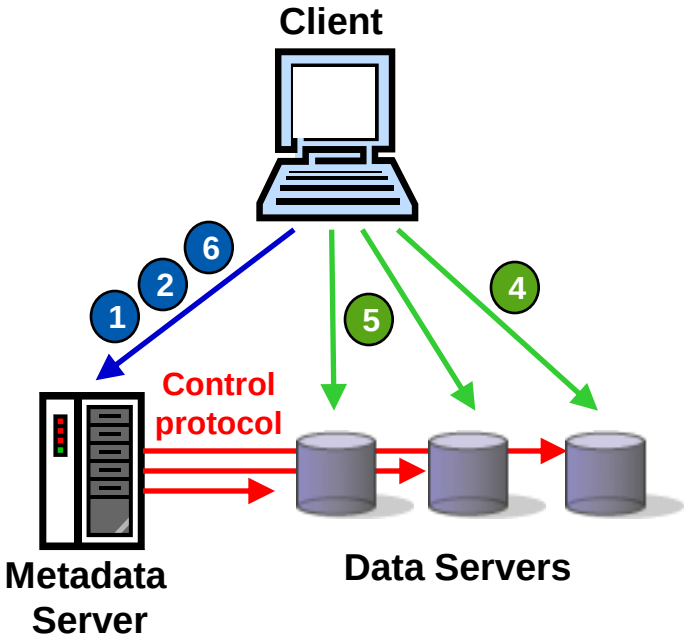
- pNFS protocol
  - standardized: NFSv4.1
- Storage access protocols
  - files (NFSv4.1)
  - blocks (FC, iSCSI, FCoE)
  - objects (OSD2)
- Control protocol
  - Outside of the pNFS standard



pNFS data flow and calls

- Client mounts filesystem
  1. GETDEVINFO/LIST

     Enumerates the data servers and access path (IP, WWN, …)
- Client to read a file
  2. LAYOUTGET
  3. pNFS client knows were to read
  4. READ file or byte range
- Client writes to file
  5. Client writes to data server
  6. LAYOUTCOMMIT



pNFS MetaData Server

- MetaData server mantains data map and updates Data Servers
- New role/server in the NFS world
- The control protocol is not defined in the RFC and will be specific to each implentation
- Part of the Data Servers will also be particular to each implementation
- In the current definition only one MetaData server is used
  Planned to add support for more servers in the future

pNFS Data Servers

- Only File based layout is included in NFS v4.1.
  Other protocols will be standarized later.

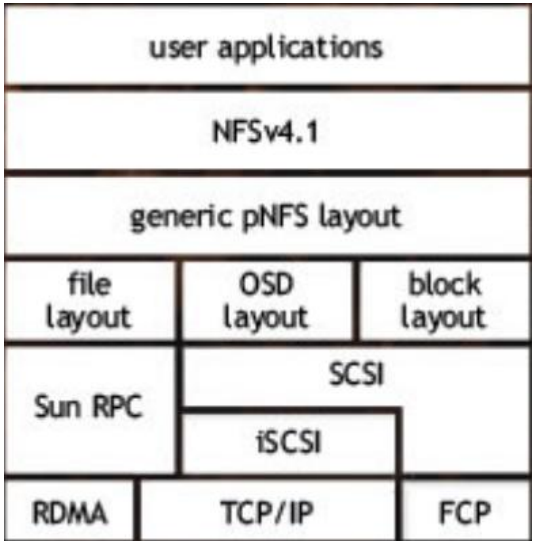
- File based layout (a NFS server with extensions)
  - A cluster filesystem is needed. Its features and limitations will be inherited by the pNFS access
  - Linux (PVFS2, Lustre, GPFS), NetApp (ONTAP), SUN (Lustre), IBM (GPFS), …

- Block based layout (SCSI: FCP, iSCSI, …)
  - Access to data is done via SCSI read/writes to LUNs
  - EMC (Highroad)

- Object based layout (OSD storage protocol)
  - More complex protocol to access data object over SCSI
  - SUN and Panasas

pNFS Clients

- Linux (file): U. Michigan+NetApp+IBM+Panasas
- Linux (Block): U. Michigan+EMC
- Linux (Object): SUN+Panasas
- Solaris: http://opensolaris.org/os/project/nfsv41/



The fit for pNFS

- Parallel access to a cluster filesystem
- Use of IP and Ethernet for storage access
- Performance ? Scalability ? Stability ?
  - Bandwith, latency, metadata, locking, …
- Might become a standard for several OS. Interoperatibility
- High performance WAN access

  http://www.linuxclustersinstitute.org/conferences/archive/2008/PDF/Hildebrand_98265.pdf

  San Diego Super Computing Center ===== Reno Super Computing Center
  10 Gb connection with 18 ms latency
  3 pNFS clients and 3 pNFS servers (GPFS) with 10GbE cards
  9.28 Gb/sec data rate

References

- IETF web & RFCs — NFSv2: http://www.ietf.org/rfc/rfc1094.txt
  NFSv3: http://www.ietf.org/rfc/rfc1813.txt
  NFSv4: http://www.ietf.org/rfc/rfc3530.txt
- SPEC NFS results — http://www.spec.org/sfs97r1/results
- NFS v4 web: http://www.nfsv4.org/
- pNFS web: http://www.pnfs.com/
- pNFS report Thijs Stuurman — https://www.os3.nl/_media/2007-2008/courses/rp2/ts-report.pdf
- pNFS Univ. Michigan — http://www.citi.umich.edu/projects/asci/pnfs/linux/
- Mike Eisler's blog: http://blogs.netapp.com/eislers_nfs_blog