pNFS Update
A standard for parallel file systems

HPC Advisory Council
Lugano, March 2011

Brent Welch
welch@panasas.com
Panasas, Inc.

panasas

Why a Standard for Parallel I/O?

- NFS is the only network file system standard
  - Proprietary file systems have unique advantages, but can cause lock-in
- NFS widens the playing field
  - Panasas, IBM, EMC want to bring their experience in large scale, high-performance file systems into the NFS community
  - Sun/Oracle and NetApp want a standard HPC solution
  - Broader market benefits vendors
  - More competition benefits customers
- What about open source
  - NFSv4 Linux client is very important for NFSv4 adoption, and therefore pNFS
  - Still need vendors that are willing to do the heavy lifting required in quality assurance for mission critical storage

NFSv4 and pNFS

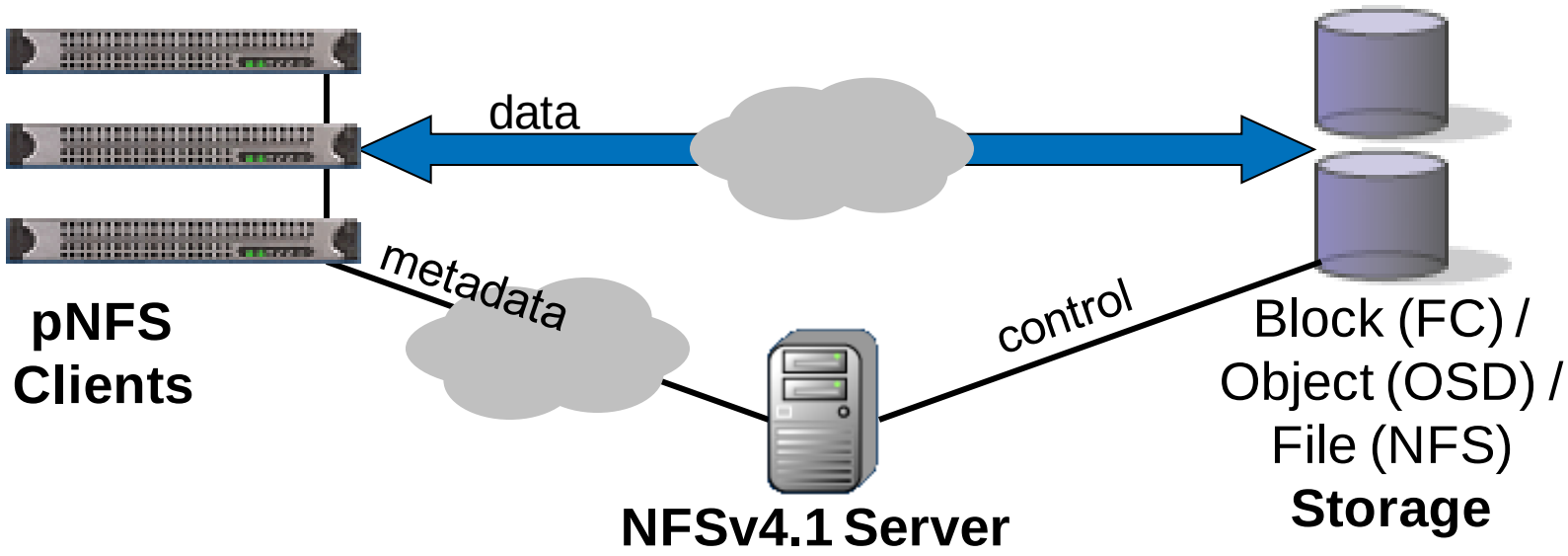
- NFS created in ’80s to share data among engineering workstations
- NFSv3 widely deployed
- NFSv4 several years in the making, lots of new stuff
  - Integrated Kerberos (or PKI) user authentication
  - Integrated File Locking and Open Delegations (stateful server!)
  - ACLs (hybrid of Windows and POSIX models)
  - Official path to add (optional) extensions
- NFSv4.1 adds even more
  - pNFS for parallel IO
  - Directory Delegations for efficiency
  - RPC Sessions for robustness, better RDMA support

Whence pNFS

- Gary Grider (LANL) and Lee Ward (Sandia)
  - Spoke with Garth Gibson about the idea of parallel IO for NFS in 2003
- Garth Gibson (Panasas/CMU) and Peter Honeyman (UMich/CITI)
  - Hosted pNFS workshop at Ann Arbor in December 2003
- Garth Gibson, Peter Corbett (NetApp), Brent Welch
  - Wrote initial pNFS IETF drafts, presented to IETF in July and November 2004
- Andy Adamson (CITI), David Black (EMC), Garth Goodson (NetApp), Tom Pisek (Sun), Benny Halevy (Panasas), Dave Noveck (NetApp), Spenser Shepler (Sun), Brian Pawlowski (NetApp), Marc Eshel (IBM), …
  - Dean Hildebrand (CITI) did pNFS prototype based on PVFS
  - NFSv4 working group commented on drafts in 2005, folded pNFS into the 4.1 minorversion draft in 2006
- *Many others*

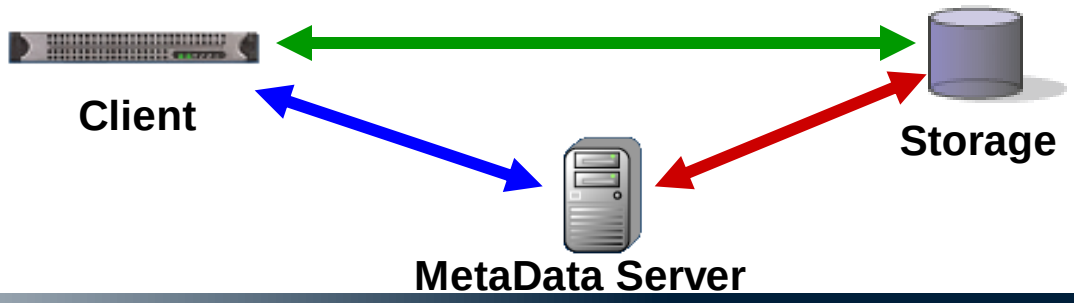
pNFS: Standard Storage Clusters

- pNFS is an extension to the Network File System v4 protocol standard
- Allows for parallel and direct access
  - From Parallel Network File System clients
  - To Storage Devices over multiple storage protocols
  - Moves the NFS (metadata) server out of the data path



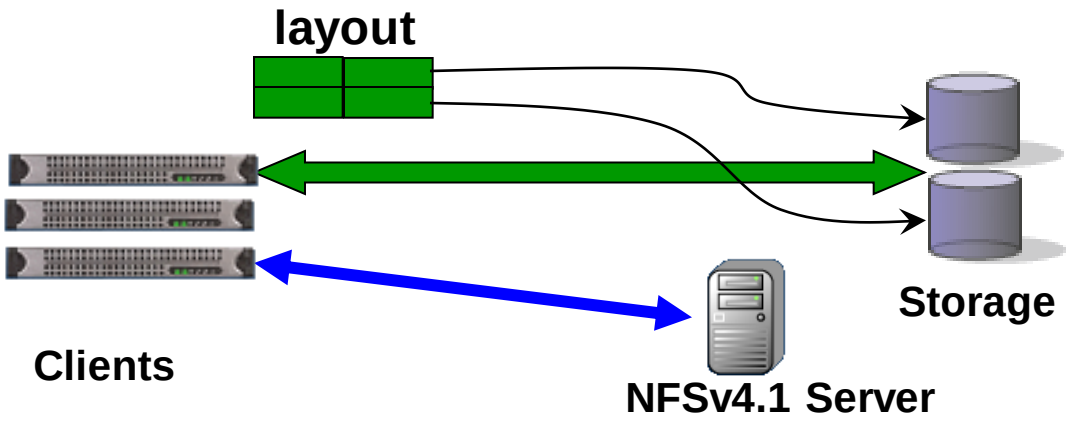
The pNFS Standard

- The **pNFS** standard defines the NFSv4.1 protocol extensions between the **server** and **client**
- The **I/O** protocol between the **client** and **storage** is specified elsewhere, for example:
  - SCSI **Block** Commands (**SBC**) over Fibre Channel (**FC**)
  - SCSI **Object**-based Storage Device (**OSD**) over iSCSI
  - Network **File** System (**NFS**)
- The **control** protocol between the **server** and **storage** devices is also specified elsewhere, for example:
  - SCSI **Object**-based Storage Device (**OSD**) over iSCSI



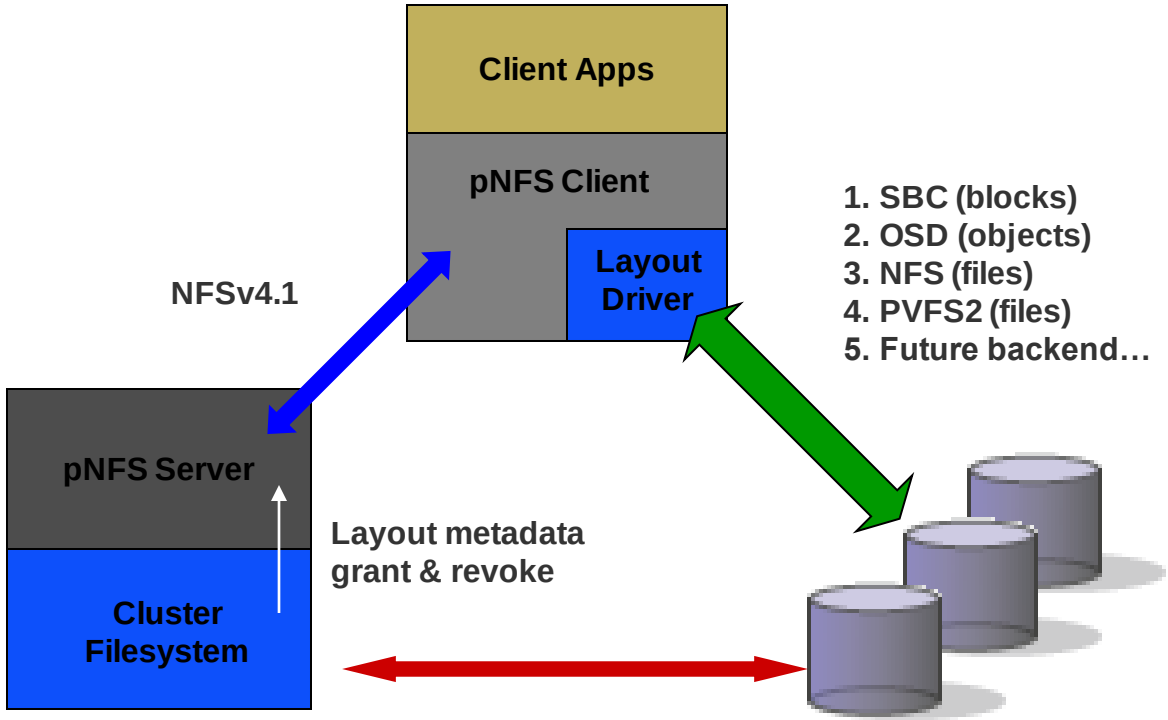
pNFS Layouts

- Client gets a *layout* from the NFS Server
- The layout maps the file onto storage devices and addresses
- The client uses the layout to perform direct I/O to storage
- At any time the server can recall the layout
- Client commits changes and returns the layout when it's done
- pNFS is optional, the client can always use regular NFSv4 I/O



pNFS Client

- Common client for different storage back ends
- Wider availability across operating systems
- Fewer support issues for storage vendors



Key pNFS Participants

- Panasas (Objects)
- ORNL and ESSC/DoD funding Linux pNFS development
- Network Appliance (Files over NFSv4)
- IBM (Files, based on GPFS)
- BlueArc (Files over NFSv4)
- EMC (Blocks, HighRoad MPFSi)
- Sun/Oracle (Files over NFSv4)
- U of Michigan/CITI (Linux maint., EMC and Microsoft contracts)
- DESY – Java-based implementation

pNFS Standard Status

- IETF approved Internet Drafts in December 2008
- RFCs for NFSv4.1, pNFS-objects, and pNFS-blocks published January 2010
  - RFC 5661 - Network File System (NFS) Version 4 Minor Version 1 Protocol
  - RFC 5662 - Network File System (NFS) Version 4 Minor Version 1 External Data Representation Standard (XDR) Description
  - RFC 5663 - Parallel NFS (pNFS) Block/Volume Layout
  - RFC 5664 - Object-Based Parallel NFS (pNFS) Operations

pNFS Implementation Status

- NFSv4.1 mandatory features have priority
  - RPC session layer giving reliable at-most-once semantics, channel bonding, RDMA
  - Server callback channel
  - Server crash recovery
  - Other details
- EXOFS object-based file system (file system over OSD)
  - In kernel module since 2.6.29 (2008)
  - Export of this file system via pNFS server protocols
  - Simple striping (RAID-0), mirroring (RAID-1), and RAID-5
  - “Most stable and scalable implementation”
- Files (NFSv4 data server) implementation
  - Open source server based on GFS
  - Layout recall not required due to nature of underlying cluster file system
- Blocks implementation
  - Server in user-level process, Ganesha/NFS support desirable
  - Sponsored by EMC

Calibrating My Predictions

- 2006
  - "TBD behind adoption of NFS 4.0 and pNFS implementations"
- 2007 September
  - Anticipate working group "last call" this October
  - Anticipate RFC being published late Q1 2008
  - Expect vendor announcements after the RFC is published
- 2008 November (SC08)
  - IETF working group last call complete, area director approval
  - *(Linux patch adoption process really just getting started)*
- 2009 November (SC09)
  - Basic NFSv4.1 features 2H2009
  - NFSv4.1 pNFS and layout drivers by 1H2010
  - Linux distributions shipping supported pNFS in 2010, 2011

Linux Release Cycle 2009

- 2.6.30
  - Merge window March 2009
  - RPC sessions, NVSv4.1 server, OSDv2 rev5, EXOFS
- 2.6.31
  - Merge window June 2009
  - NFSv4.1 client, sans pNFS
- 2.6.32
  - Merge window September 2009
  - 130 server-side patches add back-channel
- 2.6.33
  - Merge window December 2009, released Feb 2010
  - 43 pNFS patches

Linux Release Cycle 2010

- 2.6.34
  - Merge window February 2010, Released May 2010
  - 21 NFS 4.1 patches
- 2.6.35
  - Merge window May 2010, release August? 2010
  - 1 client and 1 server patch (4.1 support)
- 2.6.36
  - Merge window August 2010
  - 16 patches accepted into the merge
- 2.6.37
  - Merge window November 2010
  - Includes first chunk of pNFS beyond generic 4.1 infrastructure
  - Still disabled

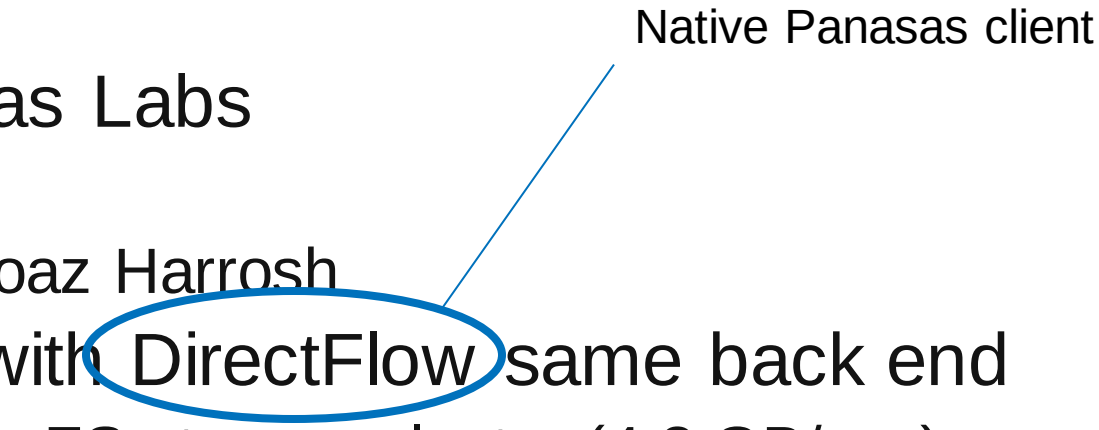
Linux Release Cycle 2011

- 2.6.XX
  - 250 patches of remaining pNFS functionality divided into 4 batches by the developers and maintainers
  - Remaining files-based pNFS patches
  - Object-based pNFS
  - Block-based pNFS
- Push from RedHat for 6.1 and beyond

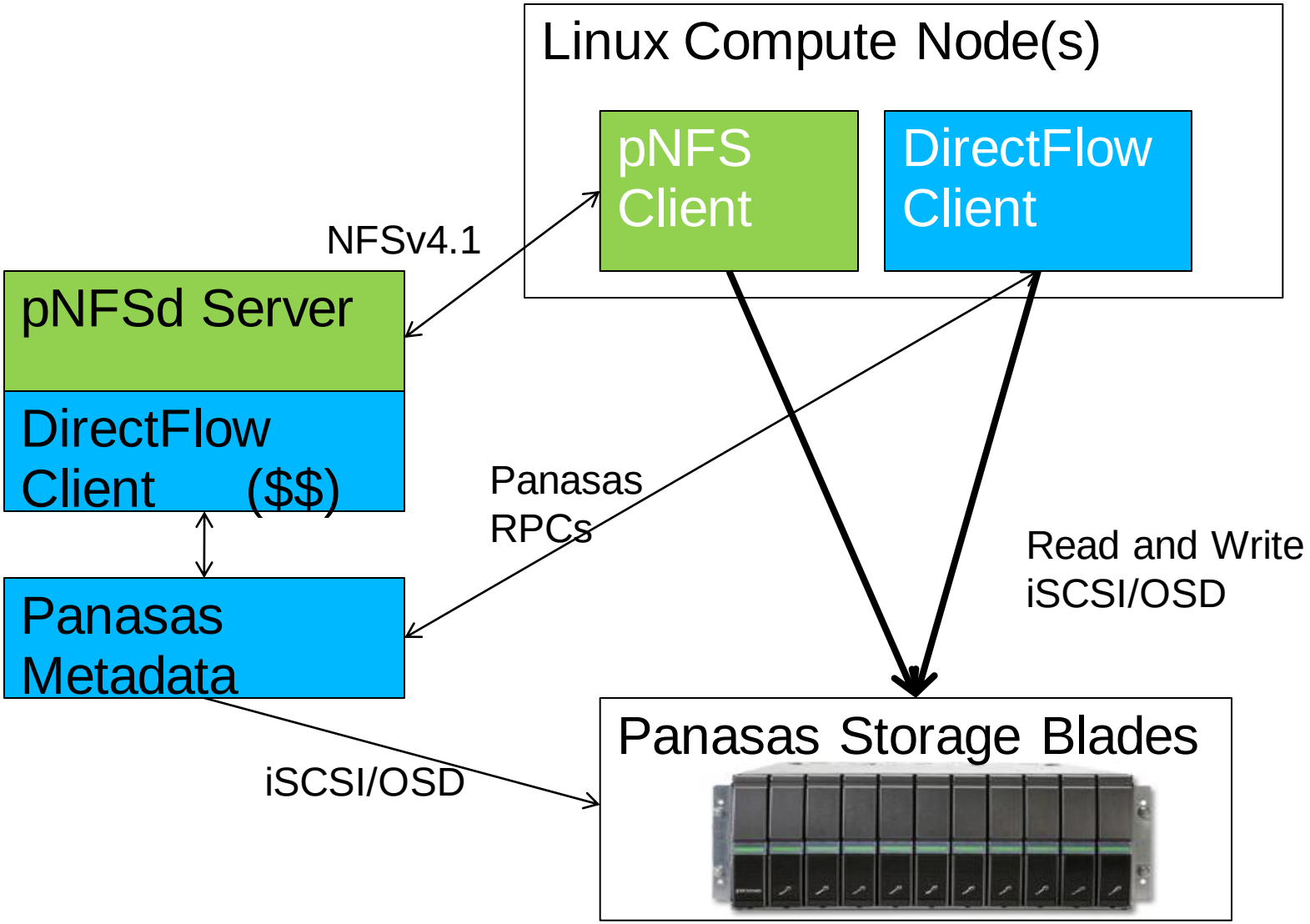
- Key IETF working group meeting in February that resolved a fine point with LAYOUT_COMMIT and files backend
- Patch adoption process will stretch to the end of 2011

pNFS Performance Testing

- Testing in Panasas Labs
  - October 2010
  - Benny Halevy, Boaz Harrosh
- Compare pNFS with DirectFlow same back end
  - Medium sized PanFS storage cluster (4.8 GB/sec)
  - Modest number of clients (128)
  - A few fast clients
  - N-to-N streaming I/O tests



System Structure

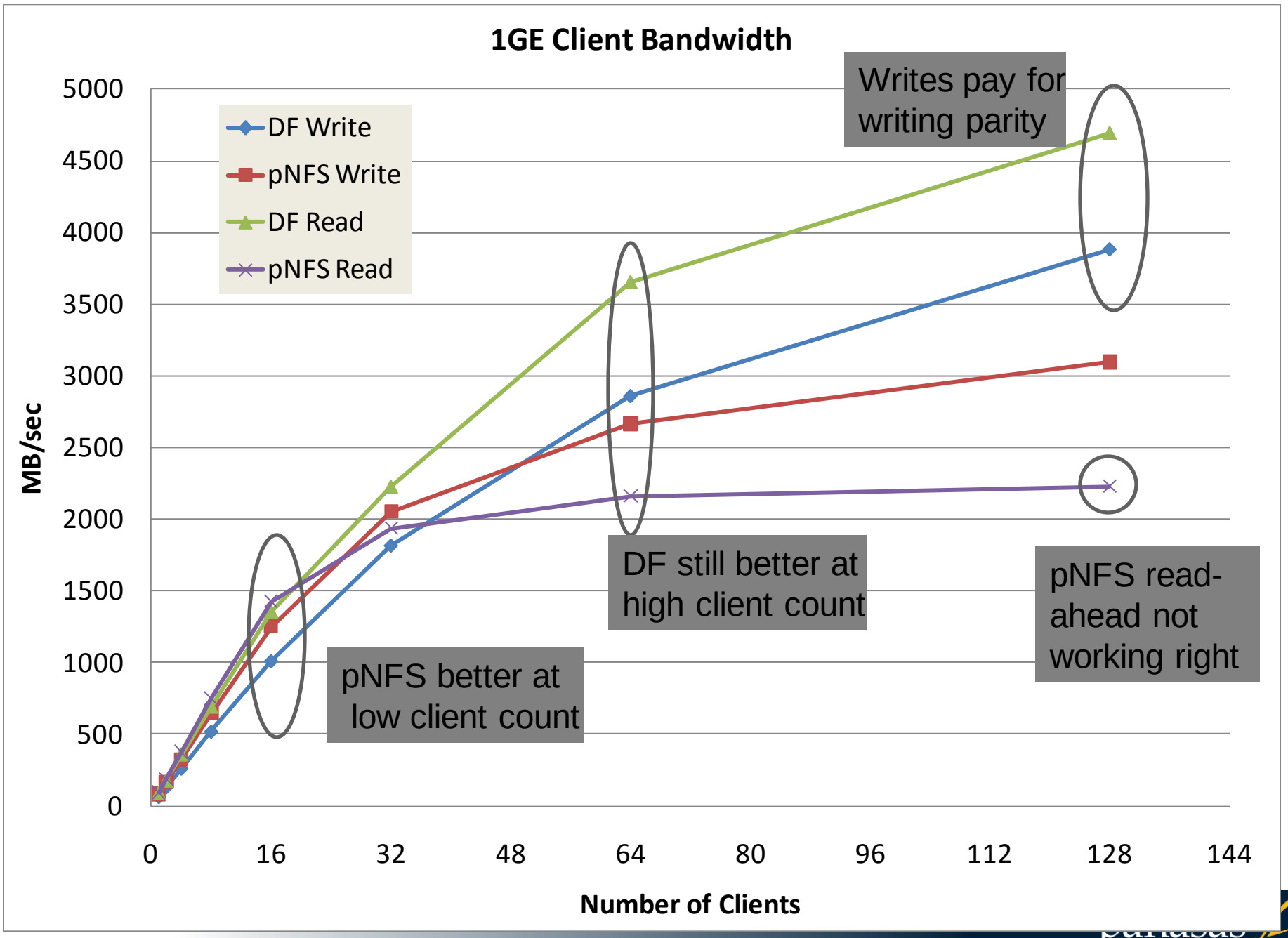


Equipment

- 12 Shelves Pas 7
  - 500 GB Blades
  - 4x 10GE uplink from each shelf
- Force 10 E-1200 switch
- 128 clients (relatively old Nacona)
  - 2 single-core sockets (2.8Gz), 8GB mem, 1GE
- 4 Faster clients (E5530)
  - 4 quad-core sockets (2.4 GHz), 12GB mem, 10GE

Streaming Bandwidth

- Iozone benchmark
- 1GE files
- Per-file Object RAID
  - Client writes data and parity in RAID-5 pattern
  - Feature of object-based pNFS layout


1GE Client Bandwidth
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
MB/sec
DF Write
pNFS Write
DF Read
pNFS Read
Writes pay for writing parity
DF still better at high client count
pNFS read-ahead not working right
pNFS better at low client count
0
16
32
48
64
80
96
112
128
144
Number of Clients

How to use pNFS today

- Up-to-date GIT tree from Linux developers
  - bhalevy@panasas.com manages the source trees
- Red Hat/Fedora RPMs that include pNFS
  - steved@redhat.com builds experimental packages
- Linux NFS mailing list, nfs@linux-nfs.org
- http://open-osd.org
  - Useful to get to OSD target, the user level program
  - Exofs uses kernel initiator, need the target

How to use pNFS today

- Benny's git tree:
  git://linux-nfs.org/~bhalevy/linux-pnfs.git
- The kernel rpms can be found at:
  http://fedorapeople.org/~steved/repos/pnfs/i686
  http://fedorapeople.org/~steved/repos/pnfs/x86_64
- The source rpm can be found at:
  http://fedorapeople.org/~steved/repos/pnfs/source/
- Bug database
  https://bugzilla.linux-nfs.org/index.cgi
- OSD target
  http://open-osd.org/

Online References: pNFS

- NFS Version 4.1
  - RFC 5661 - Network File System (NFS) Version 4 Minor Version 1 Protocol
  - RFC 5662 - Network File System (NFS) Version 4 Minor Version 1 External Data Representation Standard (XDR) Description
  - RFC 5663 - Parallel NFS (pNFS) Block/Volume Layout
  - RFC 5664 - Object-Based Parallel NFS (pNFS) Operations
  - http://tools.ietf.org/html/
- pNFS Problem Statement
  - Garth Gibson (Panasas), Peter Corbett (Netapp), Internet-draft, July 2004
  - http://www.pdl.cmu.edu/pNFS/archive/gibson-pnfs-problem-statement.html
- Linux pNFS Kernel Development
  - http://www.citi.umich.edu/projects/asci/pnfs/linux