pNFS Performance

October 2010 Results

Halevy, Horrosh, Welch

pNFS Performance Testing

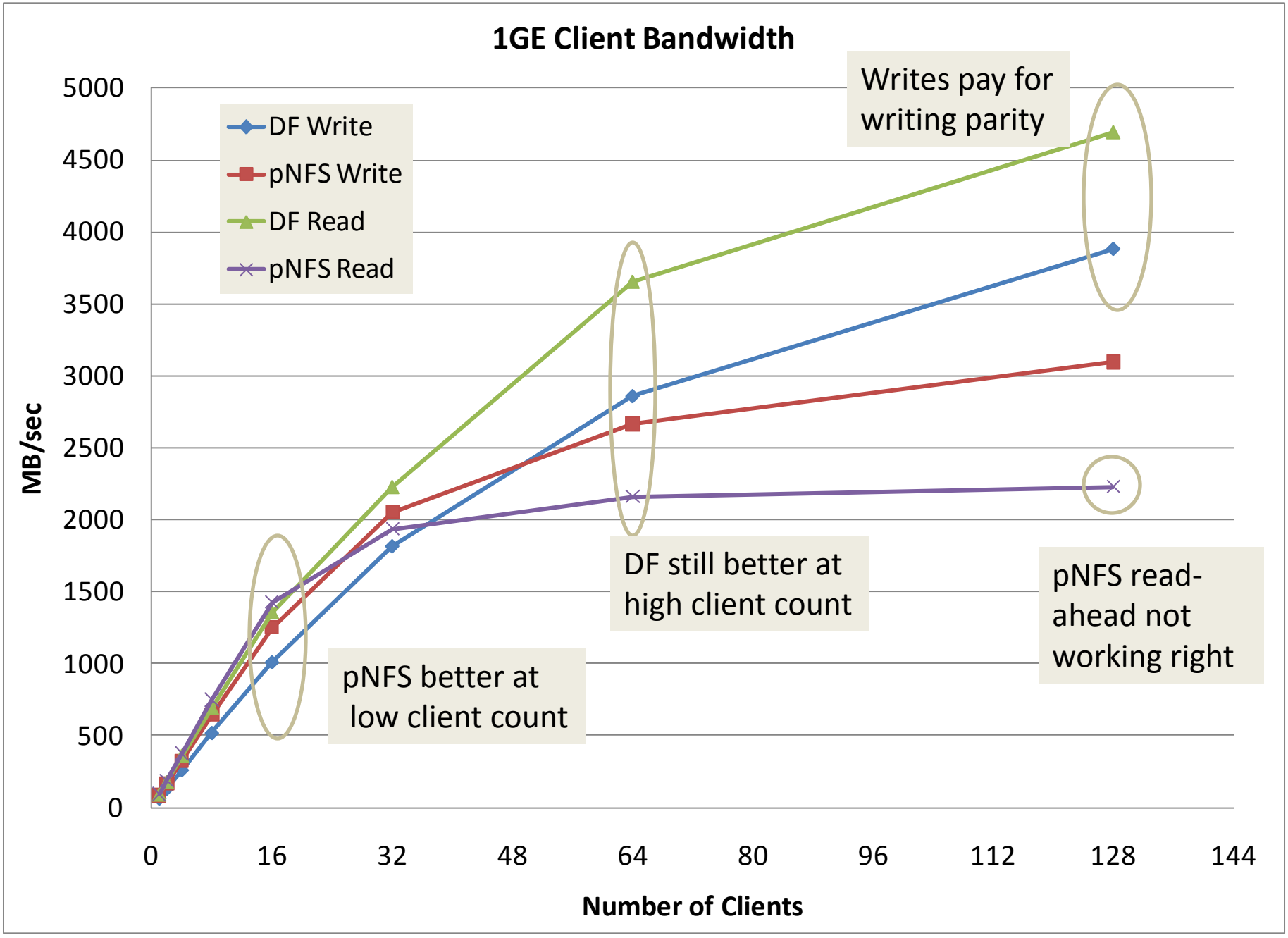
- Testing in Panasas Labs
  - October 2010
  - Benny Halevy, Boaz Harrosh
- Compare pNFS with DirectFLOW same setup
  - Medium sized PanFS storage cluster (4.8 GB/sec)
  - Modest number of clients (128)
  - A few fast clients
  - N-to-N streaming I/O tests

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
0
16
32
48
64
80
96
112
128
144
Number of Clients
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

pNFS Implementation

- 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 now RAID-5 in progress
  - “Most stable and scalable implementation”
- Files (NFSv4 data server) implementation
  - Server based on GFS
  - Layout recall not required due to nature of underlying cluster file system
- Blocks implementation
  - Server in user-level process, FUSE 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 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 Merged December 2010
  - Some client side patches adopted, pNFS still disabled

Linux Release Cycle 2011

- 2.6.X (X > 37)
  - 290 patches represent pNFS functionality divided into 4 waves (at least)
  - Wave 1 is in 2.6.37 but isn't sufficient by itself
  - All four waves represent just the files-based functionality
  - The blocks and object support is ready to go, but is waiting its turn
- Current prediction (feb 2011)
  - Takes all of 2011 to get the rest of the patches, including blocks and objects
  - There is a good chance that blocks and objects slip into early 2012
  - Redhat, however, will continue to pull agressively to make Fedora rpms

How to use pNFS today

- Benny's git tree <bhalevy@panasas.com>:

  git://linux-nfs.org/~bhalevy/linux-pnfs.git

- The rpms <steved@redhat.com>:

  http://fedorapeople.org/~steved/repos/pnfs/i686

  http://fedorapeople.org/~steved/repos/pnfs/x86_64

  http://fedorapeople.org/~steved/repos/pnfs/source/

- Bug database <pnfs@linux-nfs.org>

  https://bugzilla.linux-nfs.org/index.cgi

- OSD target

  http://open-osd.org/