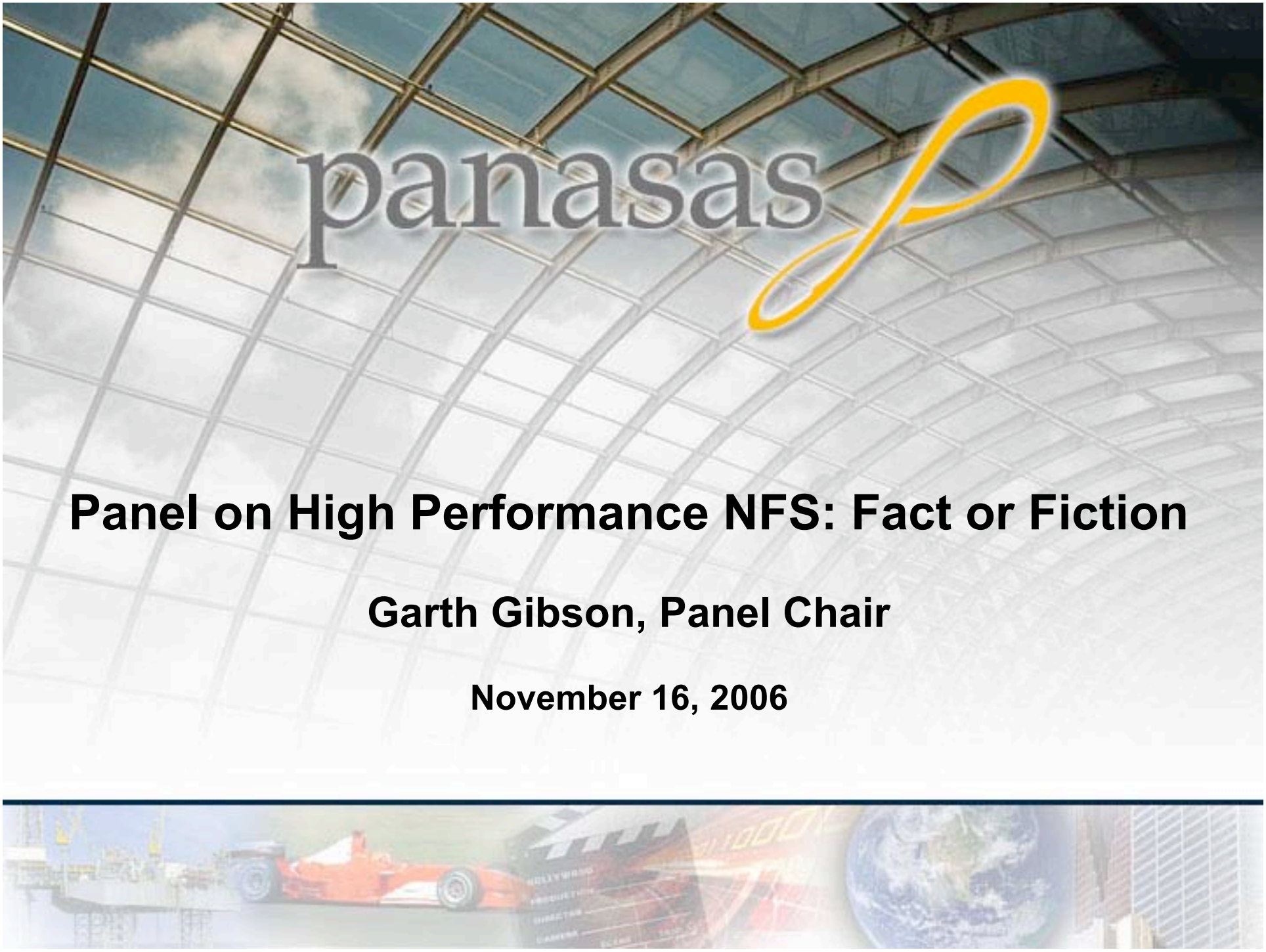


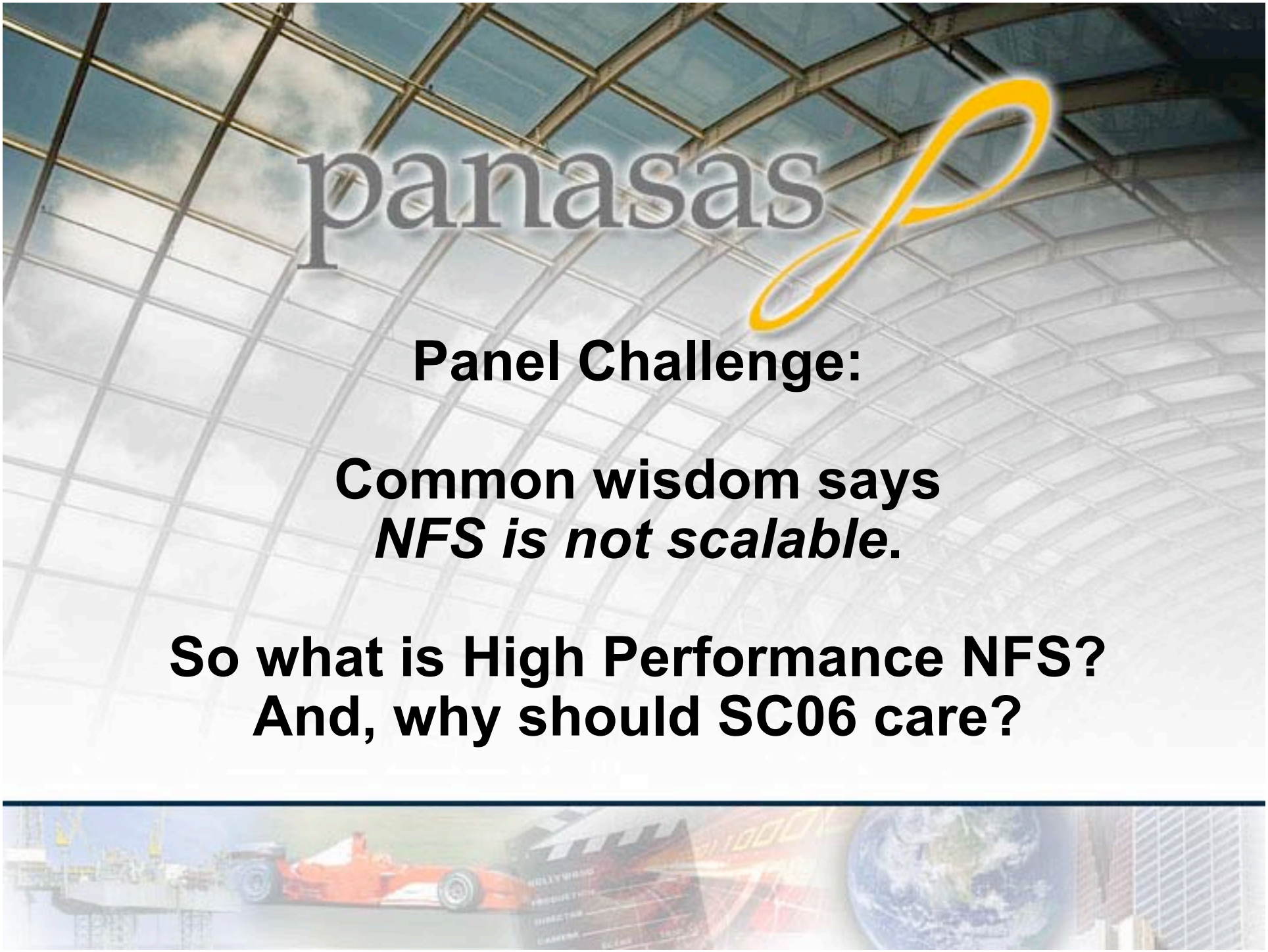
Panel on High Performance NFS: Fact or Fiction

Garth Gibson, Panel Chair

November 16, 2006

Panelists

- Garth Gibson, CTO, Panasas Inc, & Prof., Carnegie Mellon Univ.
- Mike Kazar, VP & Chief Architect, Network Appliance
- Paul Rutherford, Sr. Director, SW Engineering, Isilon
- Michael Callahan, CTO, PolyServe
- Raju Bopardikar, CTO, Crosswalk
- Uday Gupta, CTO, NAS, EMC
- Peter Honeyman, Scientific Director, CITI, Univ. of Michigan
- Roger Haskin, Sr. Manager, File Systems, IBM



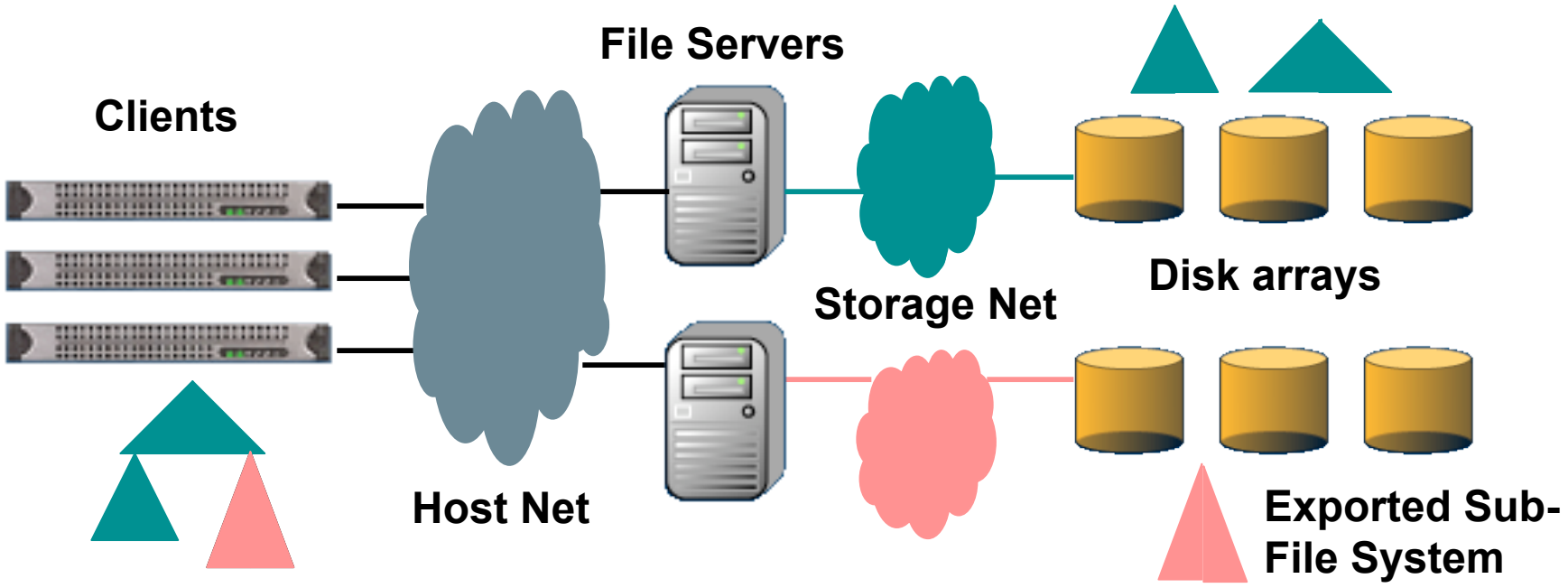
Panel Challenge:

Common wisdom says
***NFS is not scalable*.**

So what is High Performance NFS?
And, why should SC06 care?

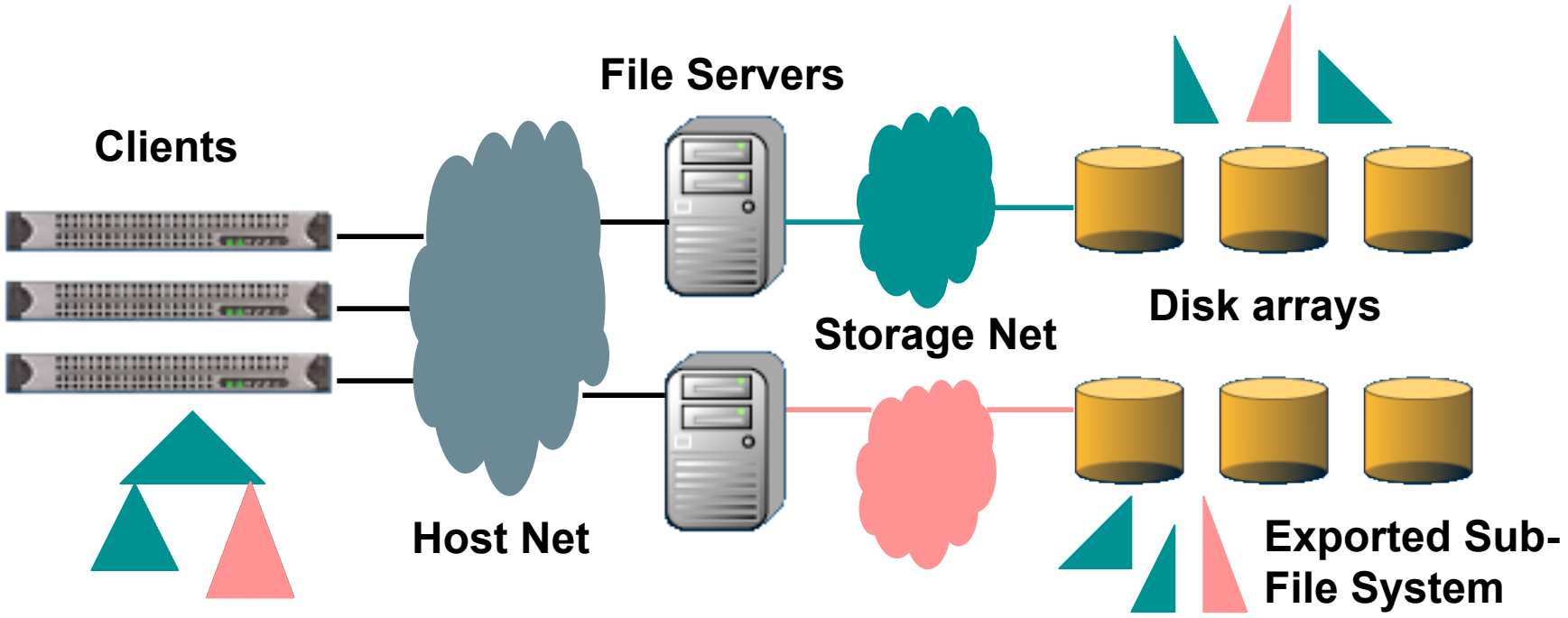
Today's Ubiquitous NFS

- ADVANTAGES
  - Familiar, stable & reliable
  - Widely supported by vendors
  - Competitive market
- LIMITATION
  - Client moves all data and metadata for a sub-file system through one network endpoint (server)



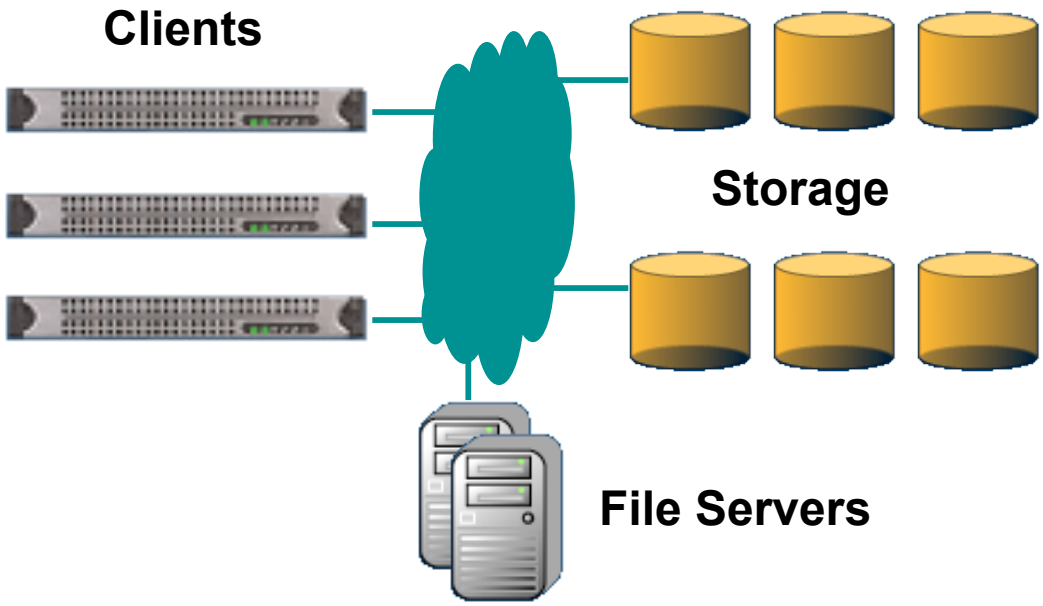
Today's Ubiquitous NFS Doesn't Scale

- ADVANTAGES
  - Familiar, stable & reliable
  - Widely supported by vendors
  - Competitive market
- DISADVANTAGES
  - Capacity doesn't scale
  - Bandwidth doesn't scale
  - "Cluster" by customer-exposed namespace partitioning



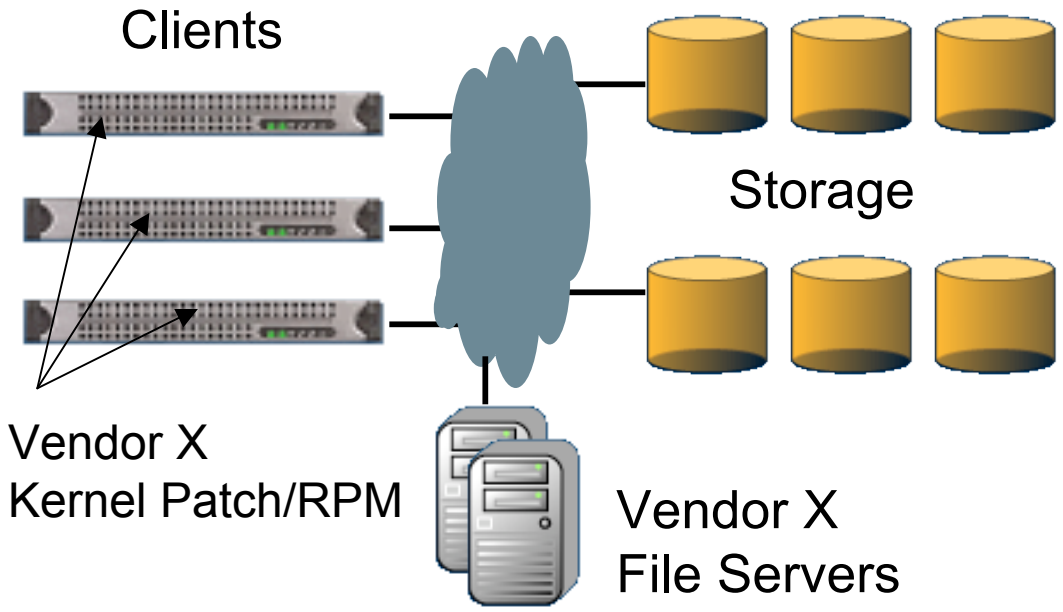
Scale Out File Service w/ Out-of-Band

- Client sees many storage addresses, accesses in parallel
  - Zero file servers in data path allows high bandwidth thru scalable networking
  - A.K.A. SAN file systems and parallel file systems
  - NOT NFS



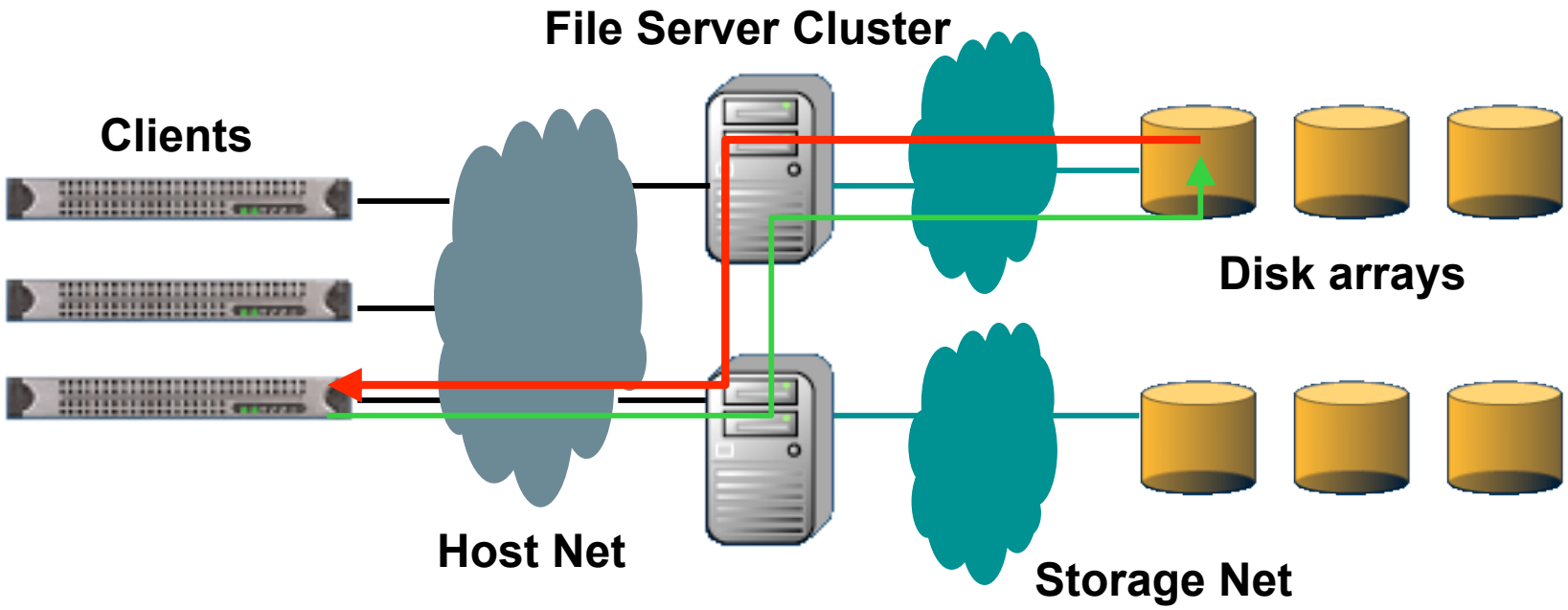
Out-of-Band Interoperability Issues

- ADVANTAGES
  - Capacity scaling
  - Faster bandwidth scaling
- DISADVANTAGES
  - Requires client kernel addition
  - Many non-interoperable solutions
  - Not necessarily able to replace NFS



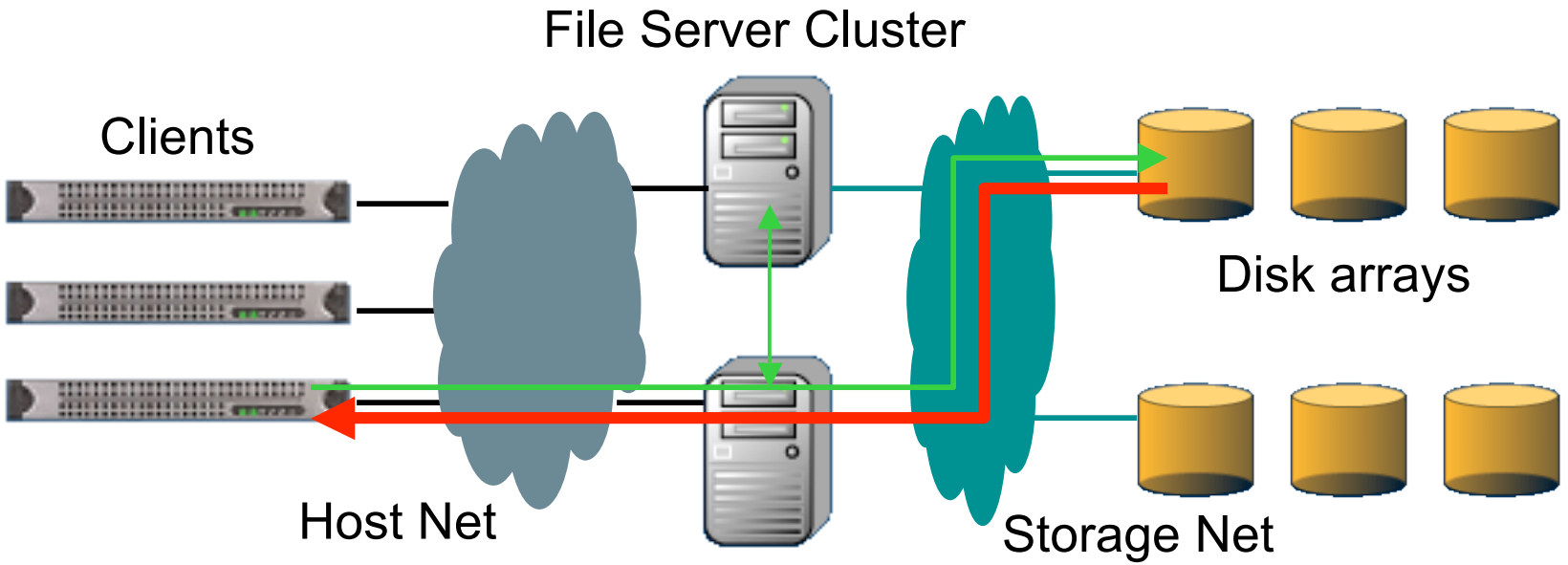
Scale Out: Cluster NFS Servers (1)

- Bind many file servers into single system image with forwarding
  - Mount point binding less relevant, allows DNS-style balancing, more manageable



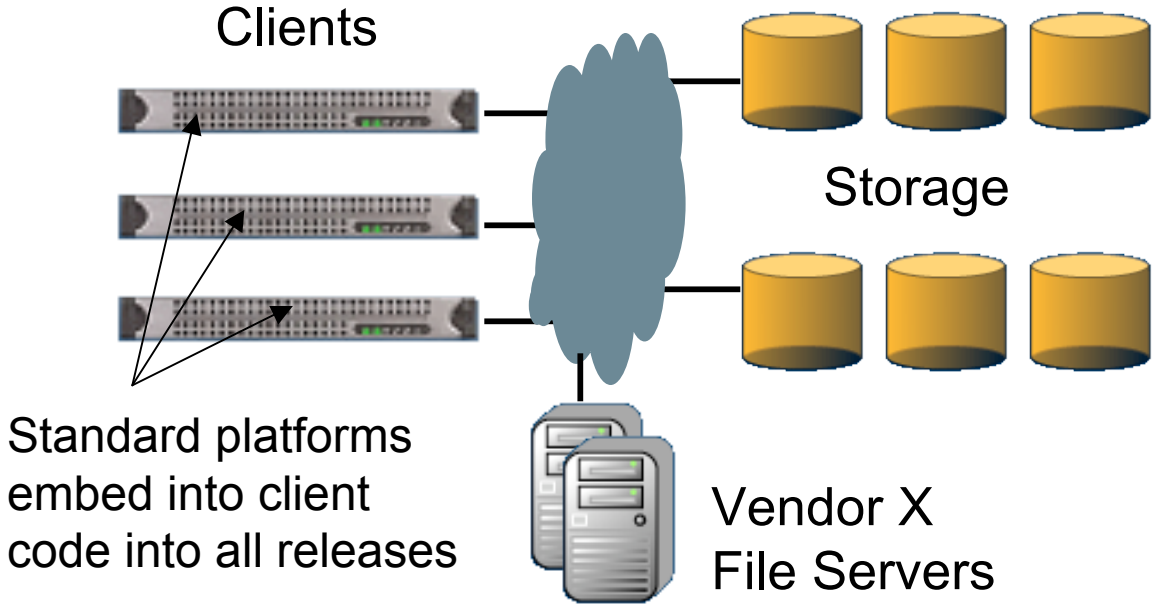
Scale Out: Cluster NFS Server (2)

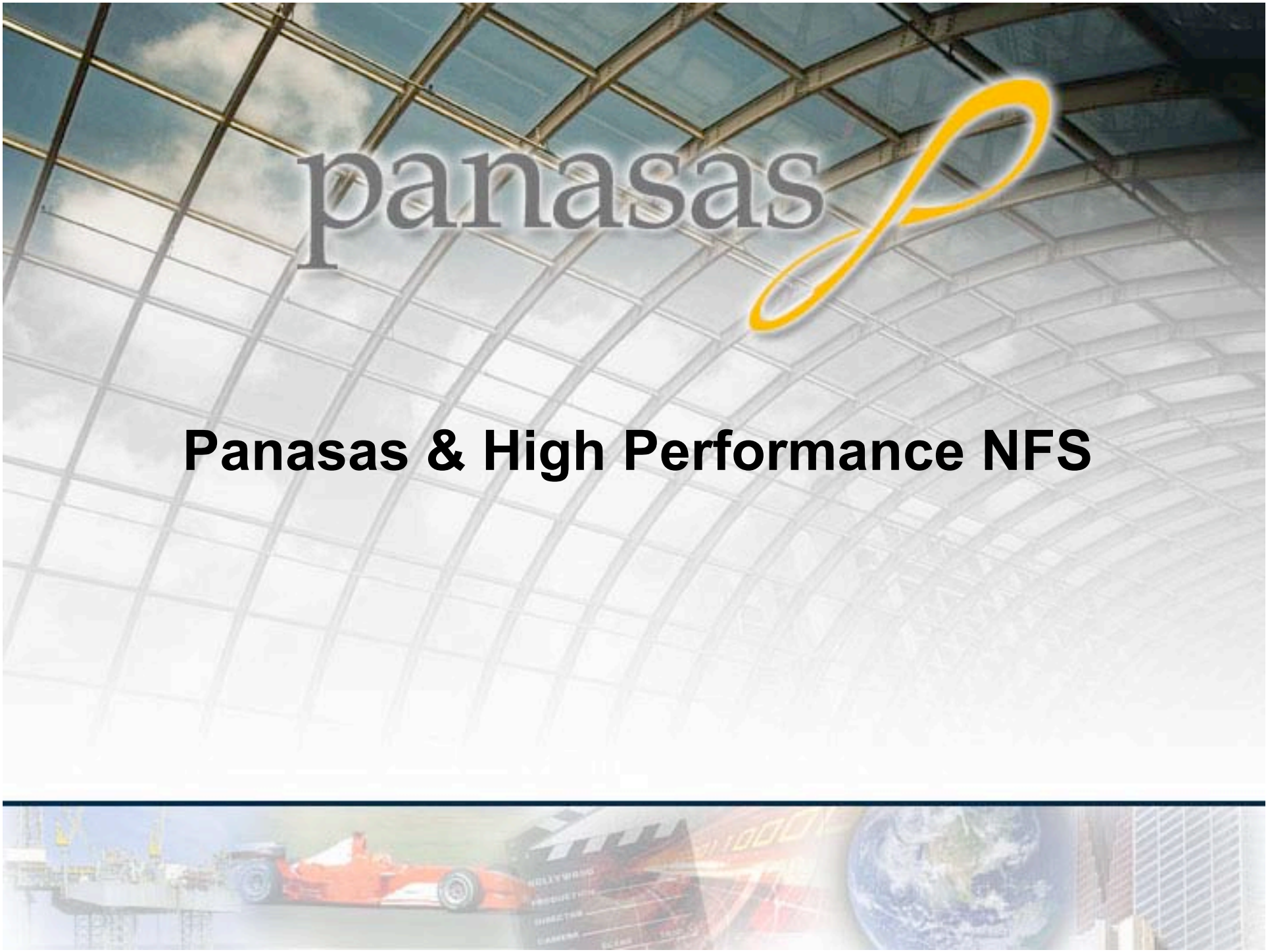
- Single server does all data transfer in single system image
  - Servers share access to all storage and “hand off” role of accessing storage
  - Control and data traverse mount point path (in band) passing through one server
  - Typically built on top of a SAN file system or parallel file system



pNFS: Out-of-Band Added to NFS

- ADVANTAGES
  - Capacity scaling
  - Faster bandwidth scaling
- Work to be done
  - Get widespread agreement on semantics
  - Build multiple reference implementations
  - Test interoperability constantly
  - Compete on SW, server implementations

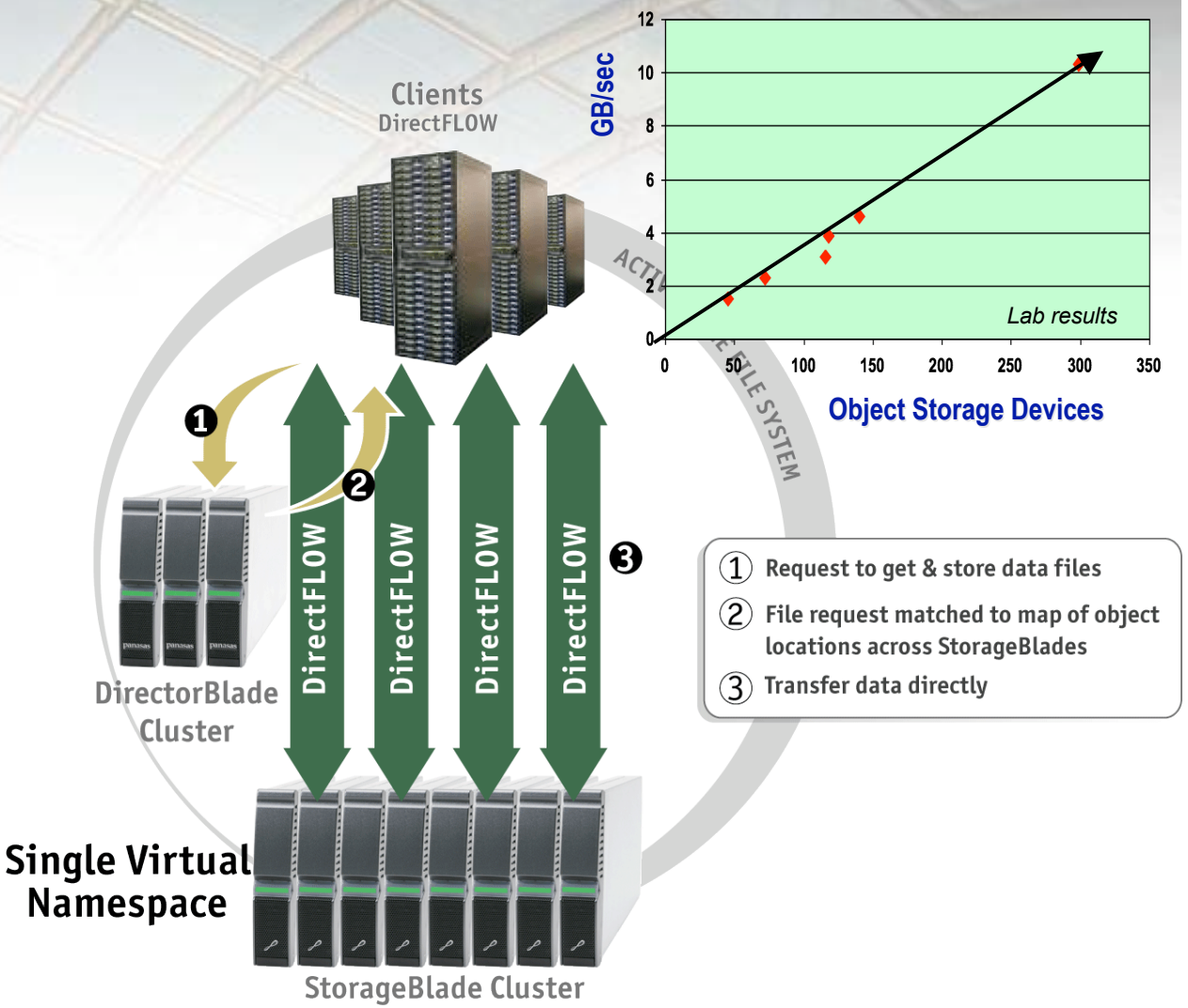




Panasas & High Performance NFS

Panasas Out-of-Band Object Storage

- **Object Based (iSCSI/OSD)**
  - For superior scalability, reliability & manageability
  - Scalable bandwidth
- **DirectFLOW client S/W**
  - Patchless Red Hat, Suse, Fedora, etc. RPM
- **DirectorBlades**
  - Manages & enables metadata scalability
  - Divides single namespace into virtual volumes
  - Clustered NFS & CIFS
- **StorageBlades**
  - Wide striping & smart prefetching
  - Smart caching & write anywhere

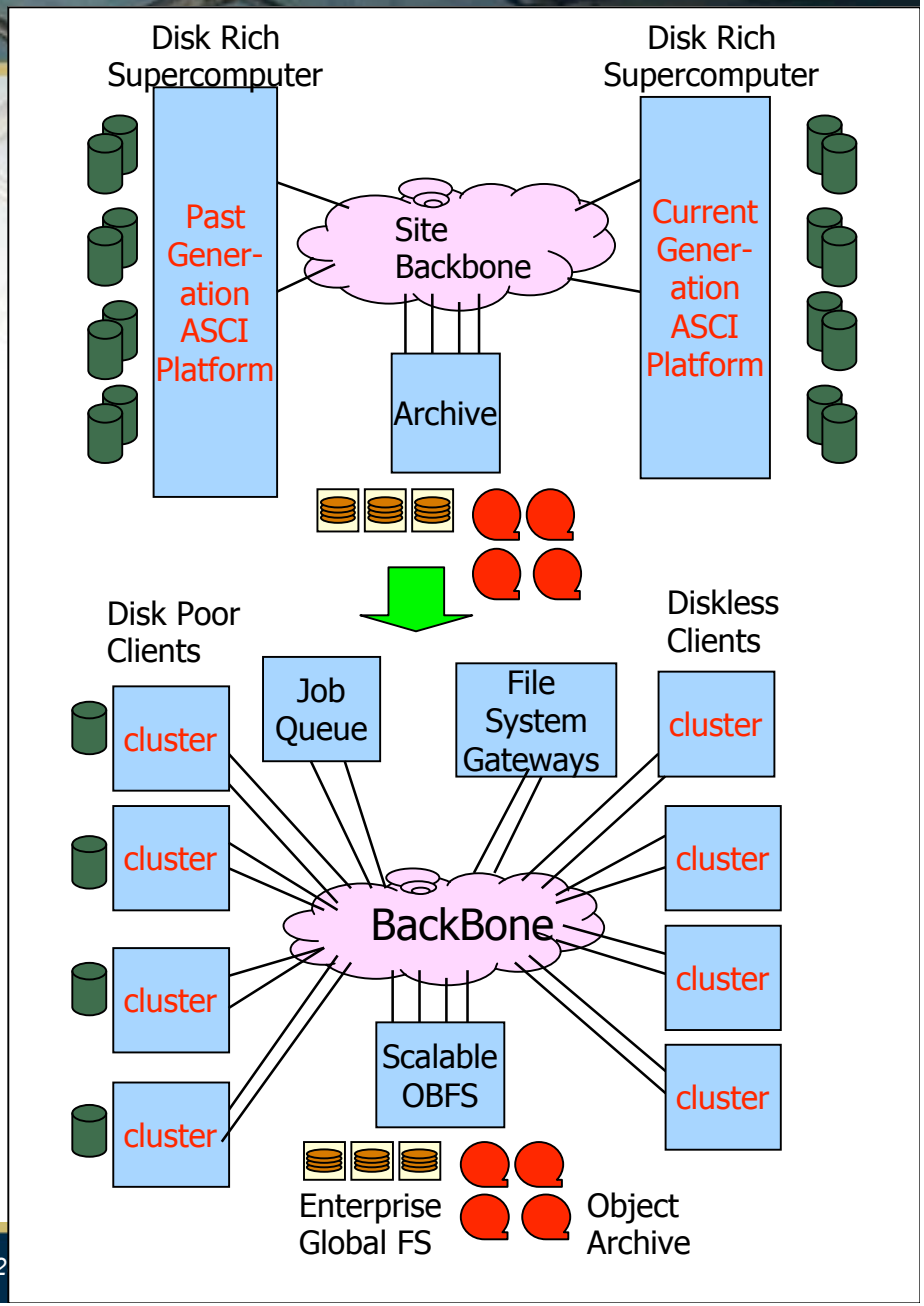


A Shared, Petascale Solution

"We've been using Panasas storage for a long time at LANL to provide scalable and globally-shared storage to multiple terascale clusters. We will leverage our successful, scalable, and stable Panasas storage solution to provide the I/O solution for the Roadrunner system,"

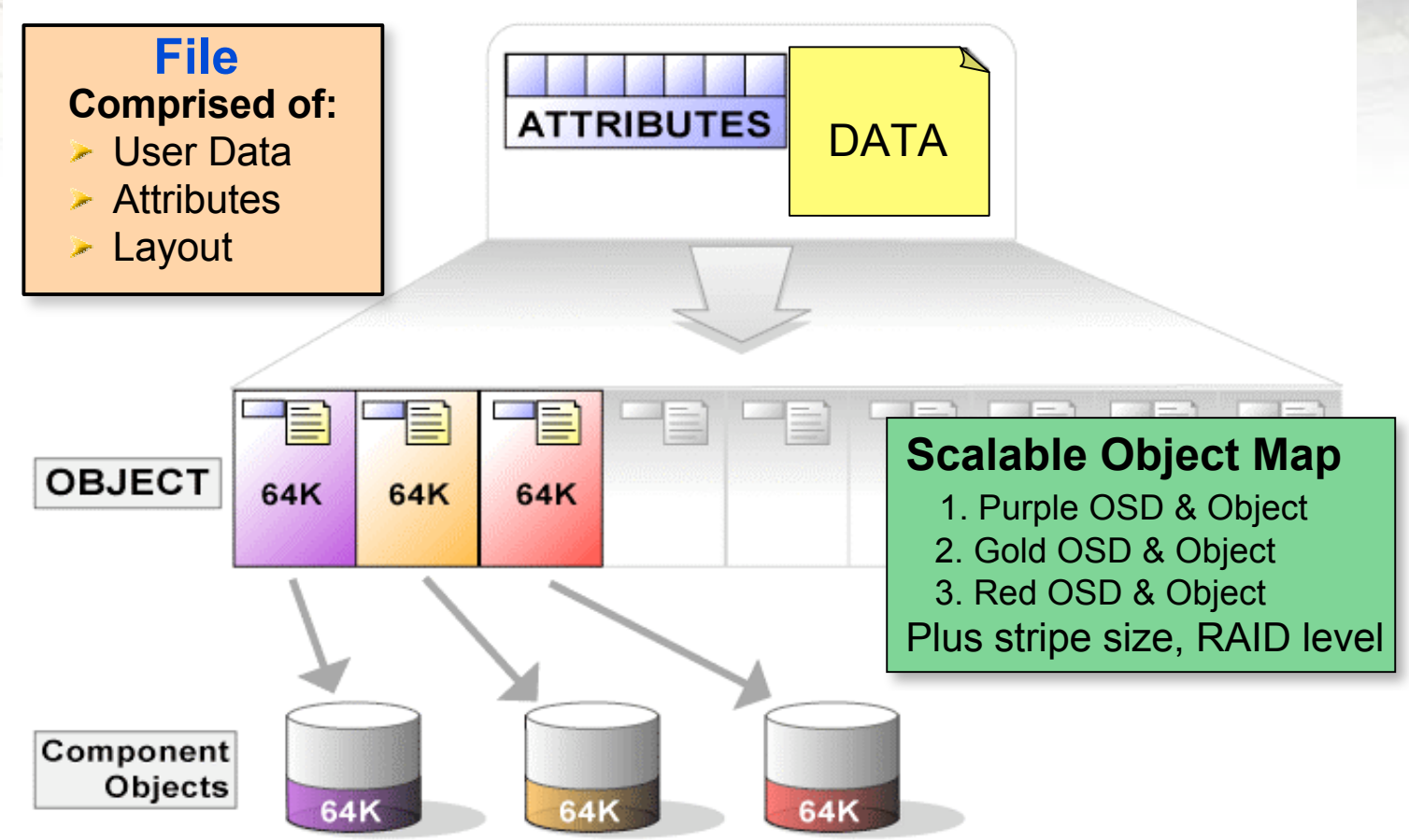
Gary Grider, group leader of
Los Alamos' High Performance Computing
Systems Integration Group
November 13, 2006

- > 1 PB Panasas in 7+ clusters
  - Myrinet: 5600 nodes, 11000+ procs, Lightning, Bolt, Pink, TLC, Flash, Gordon
  - Infiniband: 1856 nodes, 3700+ procs, Blue Steel, Coyote, & soon Roadrunner



How Does Panasas Scale Objects?

Scale capacity, bandwidth, reliability by striping according to small map



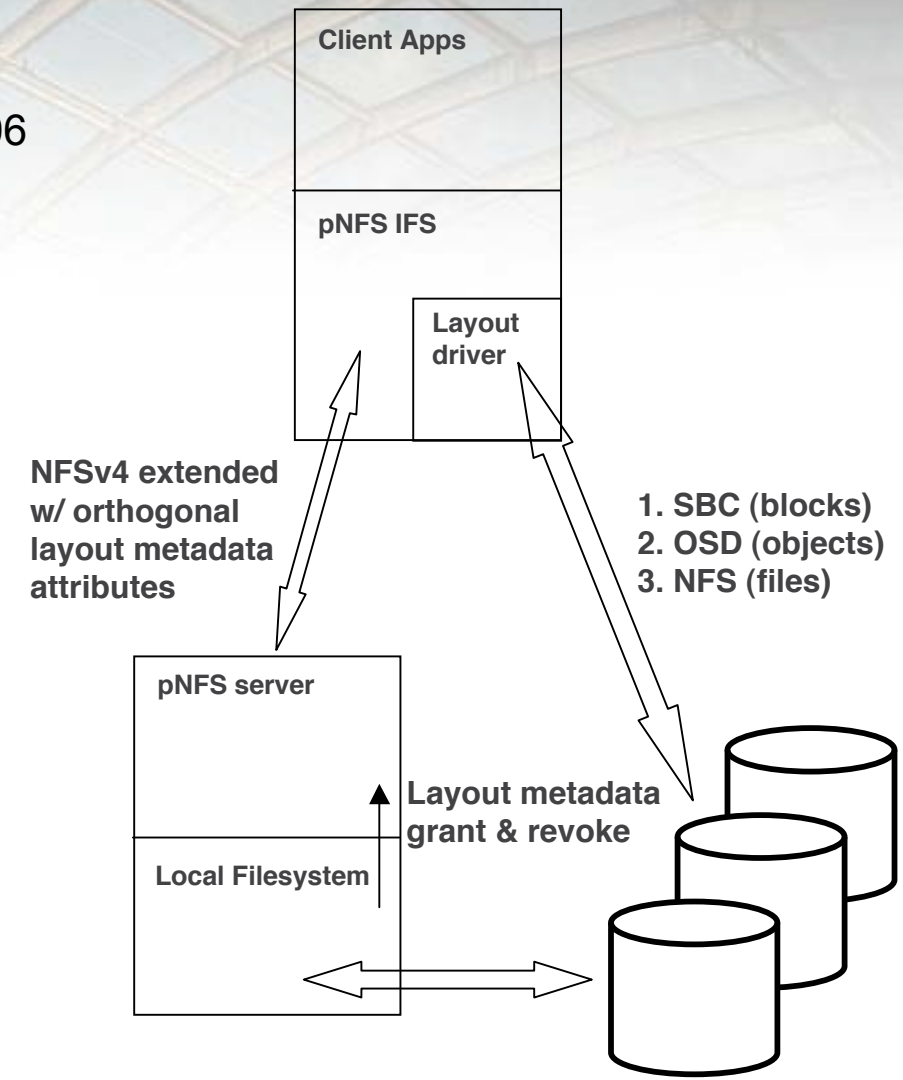
Central idea in pNFS is to enable NFSv4 to delegate maps (layouts)

Highlights of the History of pNFS

- Conversations with Gary Grider, LANL, & Lee Ward, Sandia, 2003
  - How to make HPC investment in High Performance File Systems persistent
- Workshop on NFS Extensions for Parallel Storage, Dec 2003, Ann Arbor
  - Chaired by Peter Honeyman, CITI/U.Mich., & Garth Gibson, CMU
- Initial problem statement, operations proposal to IETF July & Nov 2004
  - Garth Gibson, Peter Corbett, NetApp, Brent Welch, Panasas
- Standards development team in action
  - Andy Adamson, CITI/U.Mich, David Black, EMC, Garth Goodson, NetApp, Tom Pisek, Sun, Benny Halevy, Panasas, Dave Noveck, NetApp, Spencer Shepler, Sun, Brian Pawlowski, NetApp, Marc Eshel, IBM, & many others
  - Dean Hildebrand, CITI/U.Mich, with Lee Ward, did first prototype & paper
- IETF working group folded it into NFSv4.1 minorversion draft in 2006
  - www.ietf.org/html.charters/nfsv4-charter.html

Parallel NFS: Delegate Maps to Clients

- IETF NFSv4.1
  - draft-ietf-nfsv4-minorversion1-08.txt 10/06
  - Includes pNFS, sessions/RDMA, directory delegations
  - U.Mich/CITI impl'g Linux client/server
- Three (or more) flavors:
  - FILES: NFS/ONCRPC/TCP
    NetApp, Sun, IBM, U.Mich/CITI, DESY
  - BLOCKS: SBC/FC or SBC/iSCSI
    EMC (-pnfs-blocks-01.txt)
  - OBJECTS: OSD/iSCSI or OSD/FC
    Panasas, Sun (-pnfs-obj-02.txt)



Internet-Drafts:

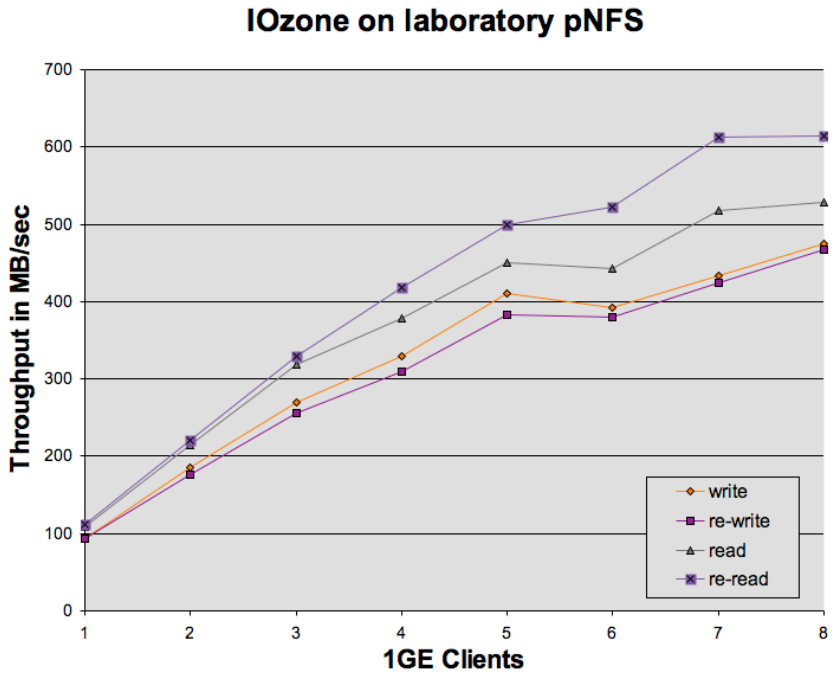
Mapping Between NFSv4 and Posix Draft ACLs (34408 bytes)
NFS RDMA Problem Statement (37522 bytes)
RDMA Transport for ONC RPC (73502 bytes)
NFS Direct Data Placement (22222 bytes)
NFSv4 Minor Version 1 (1070993 bytes)
pNFS Block/Volume Layout (45088 bytes)
Object-based pNFS Operations (51209 bytes)

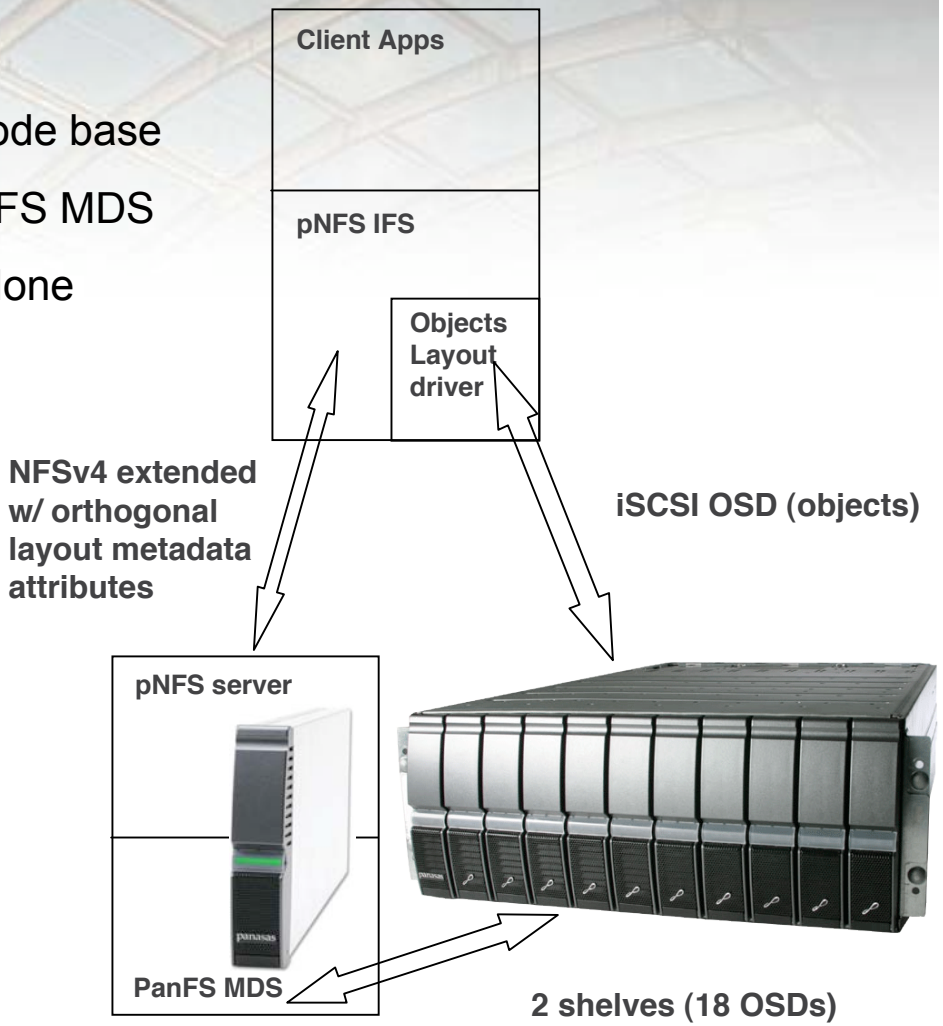
pNFS Protocol Operations

- LAYOUTGET
  - (filehandle, type, byte range) -> type-specific layout
- LAYOUTRETURN
  - (filehandle, range) -> server can release state about the client
- LAYOUTCOMMIT
  - (filehandle, byte range, updated attributes, layout-specific info) -> server ensures that data is visible to other clients
  - Timestamps and end-of-file attributes are updated
- CB_LAYOUTRECALL
  - Server tells the client to stop using a layout
- CB_RECALLABLE_OBJ_AVAIL
  - Delegation available for a file that was not previously available
- GETDEVICEINFO, GETDEVICELIST
  - Map deviceID in layout to type-specific addressing information

Panasas Committed to pNFS

- Promising preliminary results
  - Built on U.Mich/CITI Linux client/server code base
  - Layer NFSv4.1 server on DirectFlow/PanFS MDS
  - Many parts of the pNFS solution not yet done
  - Iozone -c -e -r448k -s 5g -t #clients





Accelerating Time to Results With Clustered Storage

Garth Gibson
garth@panasas.com

Panelists

- Garth Gibson, CTO, Panasas Inc, & Prof., Carnegie Mellon Univ.
- Mike Kazar, VP & Chief Architect, Network Appliance
- Paul Rutherford, Sr. Director, SW Engineering, Isilon
- Michael Callahan, CTO, PolyServe
- Raju Bopardikar, CTO, Crosswalk
- Uday Gupta, CTO, NAS, EMC
- Peter Honeyman, Scientific Director, CITI, Univ. of Michigan
- Roger Haskin, Sr. Manager, File Systems, IBM