The Last Bottleneck: How Parallel I/O can improve application performance

HPC ADVISORY COUNCIL STANFORD WORKSHOP; DECEMBER 6TH 2011

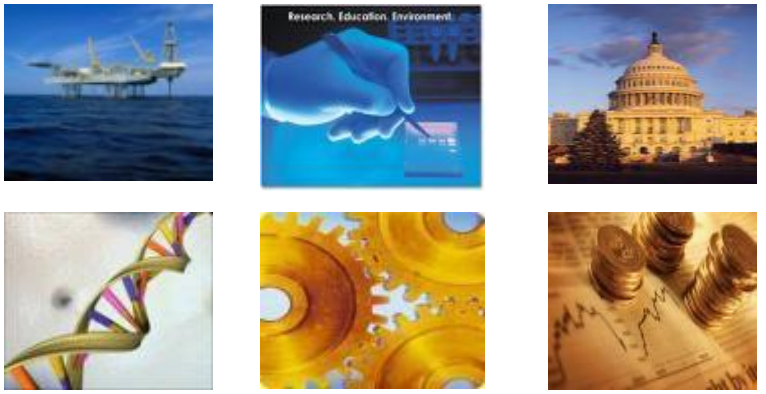
REX TANAKIT DIRECTOR OF INDUSTRY SOLUTIONS

AGENDA

- **Panasas Overview**
  - Who we are, what we do
- **pNFS (Parallel NFS)**
  - What is it?
  - How does it work?
- **Benefits of Parallel I/O**
  - Higher Performance

PANASAS OVERVIEW

- Founded by Dr. Garth Gibson in 1999. First Customer Ship in 2004
- Over 330+ WW customers; many with petabytes of data
- HQ – “Silicon Valley”, CA, USA
- Market Focus:
  - Energy
  - Academia
  - Government
  - Life Sciences
  - Manufacturing
  - Finance



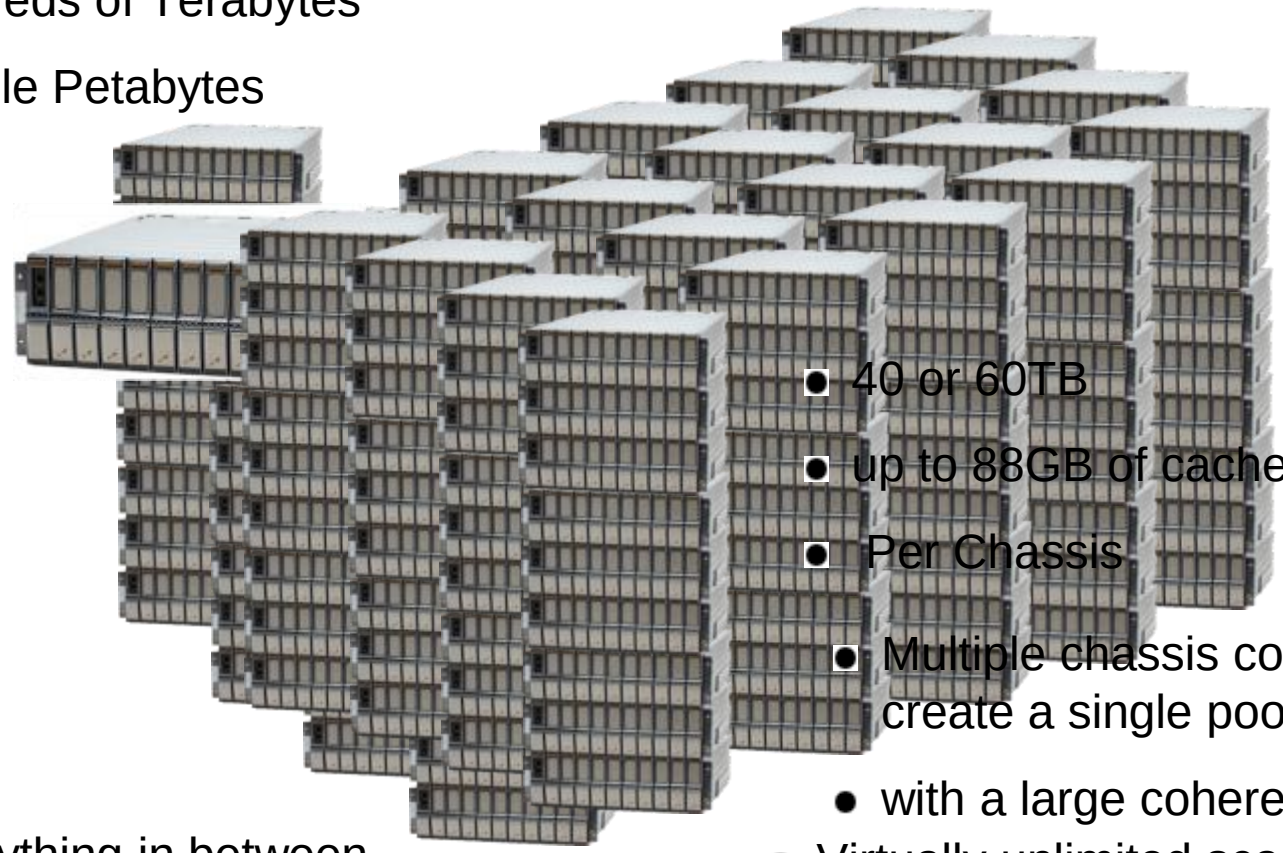
- Technologies: parallel file system and parallel storage appliance

Worldwide support with over 25 global resellers

CTCSP Corporation · DELL · Penguin Computing · ServiWare · sgi

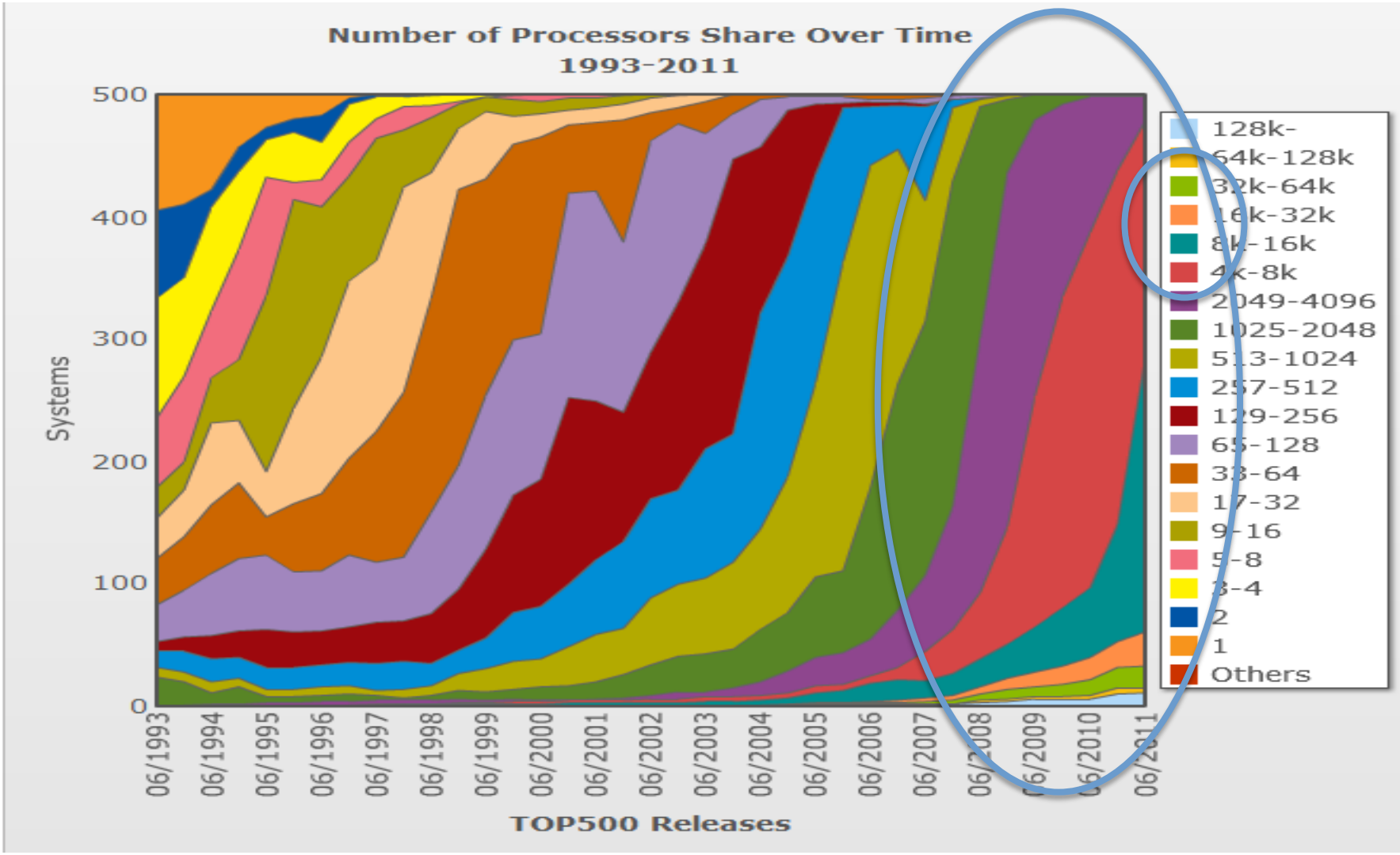
WHAT DO WE DO?

- We build high performance scalable storage systems using parallel FS
- From a single Panasas Shelf
- To Hundreds of Terabytes
- To Multiple Petabytes



- 40 or 60TB
- up to 88GB of cache
- Per Chassis
- Multiple chassis combine to create a single pool of storage
- with a large coherent cache
- Virtually unlimited scaling capability
- And everything in between

PARALLEL SYSTEM GROWTH (TOP500)



Source: http://top500.org/overtime/list/37/procclass

PARALLEL SYSTEM

System: Hardware + File system + Storage + Application +…..

Amdahl's law: "*The speedup of a program using multiple processors in parallel computing is limited by the time needed for the sequential fraction of the program.*" *Source: Wikipedia*

- Multi-core and cpu are the norm 
- Parallel File System: 10+ years with Panasas + pNFS 
- Parallel programming: OpenMP, MPI, 15+years 
- Applications are moving from serial to parallel I/O 

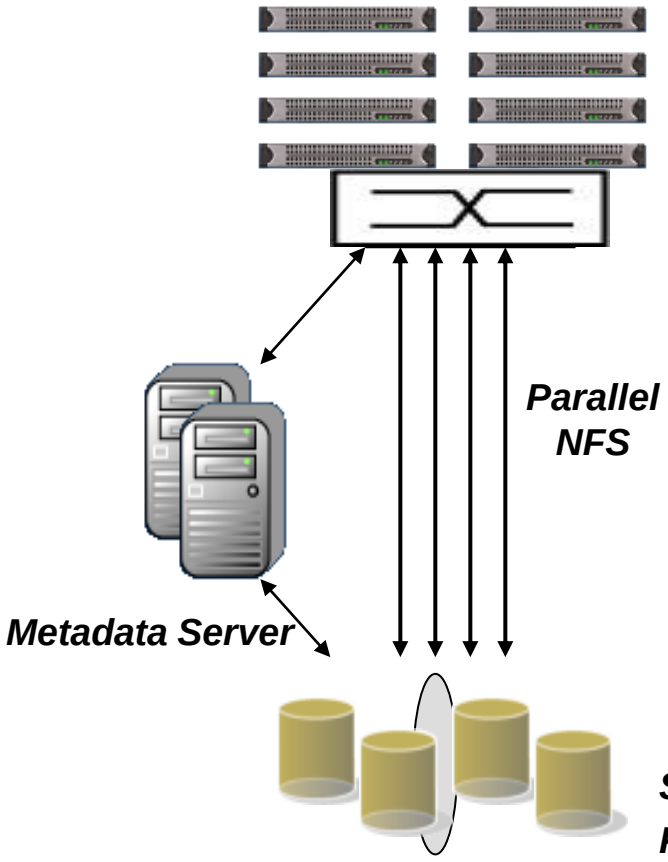
A standard like pNFS encourages moving to parallel I/O

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 KEY CONCEPTS

- **Separate metadata and data**
  - Compute clients can access data directly and in parallel
  - Add metadata server
- **Introduce the concept of data layout**
  - Layout is a map for clients to access data on the storage
- **Backend storage**
  - Supports multiple types of back-end storage systems, including traditional block storage, other file servers, and object storage systems

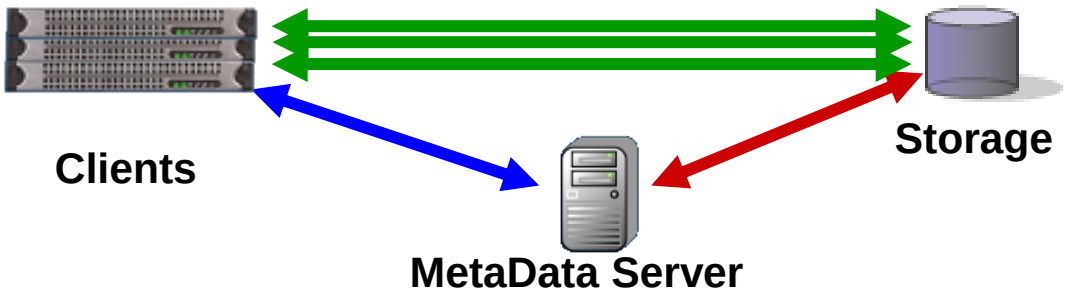


Parallel Storage:

File server not in data path.
Performance bottleneck eliminated.

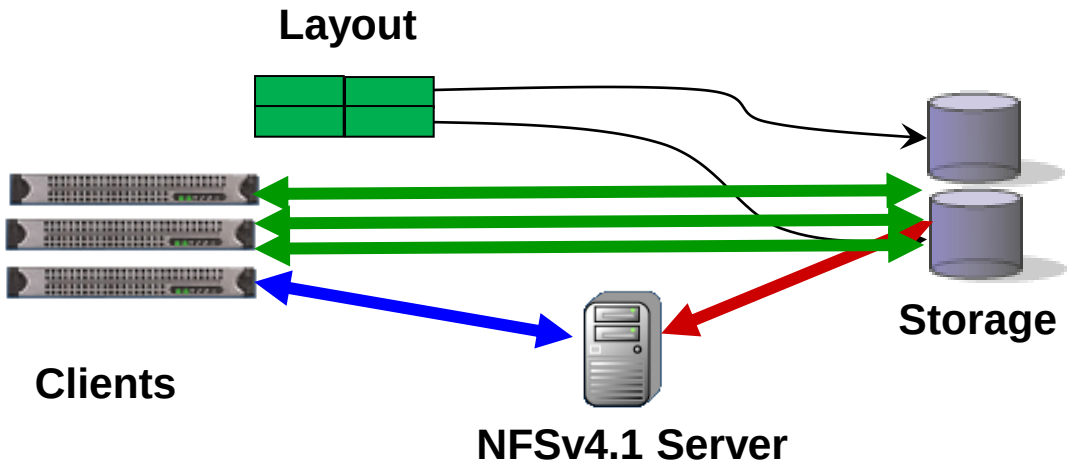
THE PNFS STANDARD

- **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 **Object**-based Storage Device (**OSD**) over iSCSI
  - SCSI **Block** Commands (**SBC**) over Fibre Channel (**FC**)
  - 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



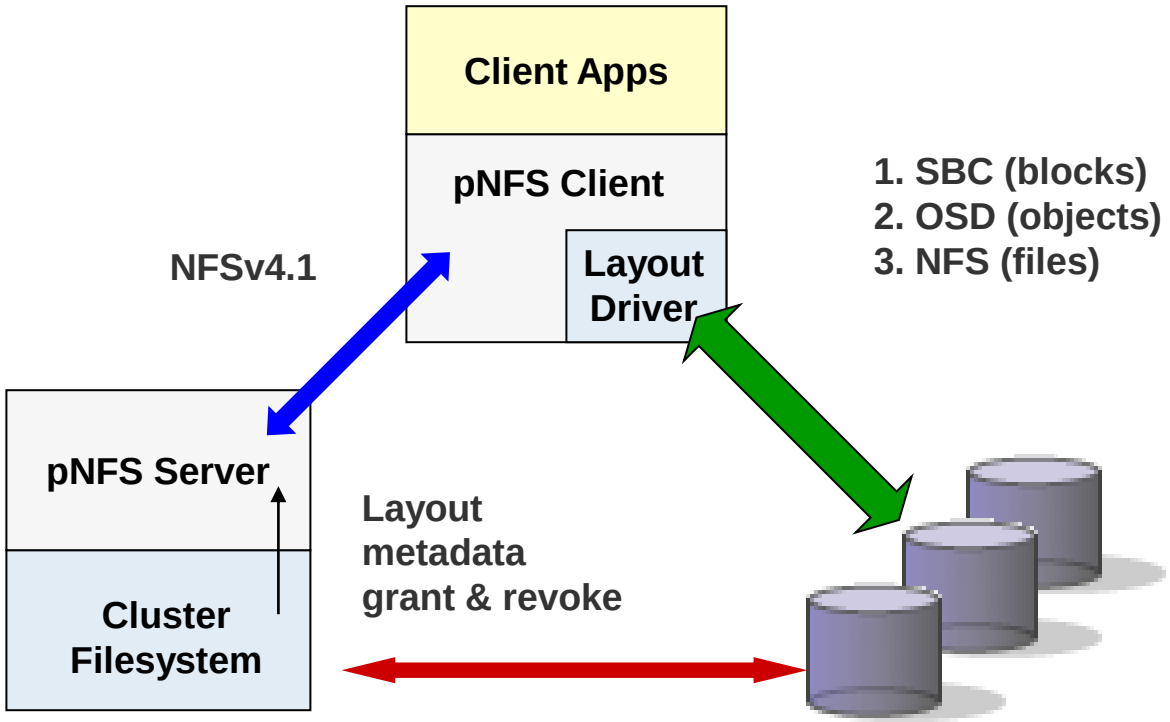
PNFS LAYOUTS

- **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**
- **Client commits changes and returns the layout when it's done**
- **At any time the server can recall the layout**
- **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



LINUX CLIENT RELEASE CYCLE 2011

Kernel	Merge Window Date	What's New
2.6.38	Jan 2011	More generic pNFS code, still disabled, not fully functional
2.6.39	Apr 2011	Files-based back end, read, write, commit on the client. Linux server is read-only via pNFS.
~~2.6.40~~ 3.0 (Fedora 15)	Jun 2011	RAID0/1 Object-based back end
3.1	Oct 22nd 2011	Block-based back end
3.2	Dec 2011 (?)	RAID5 Object (Already in Linus's Tree)

- RHEL 6 and SLES 11 based on 2.6.32
  - Backporting pNFS for files will be attempted
- RHEL 7 and SLES 12 based on 3.*
  - Integrated pNFS of all flavors – timeline 2012+

PNFS SERVER TODAY

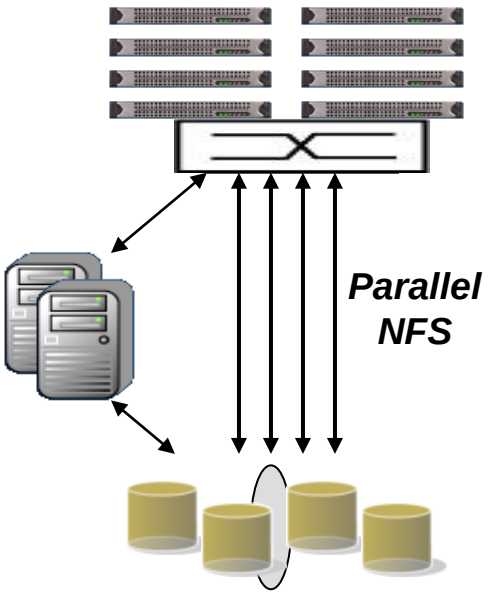
panasas

- **Up-to-date GIT tree from Linux pNFSD server**
  - git://linux-nfs.org/~bhalevy/linux-pnfs.git
  - Files and Blocks Simple Server (spNFS)

- **pNFS Object Open-source Server: http://open-osd.org**
  - Useful to get to OSD target, the user level program
  - Exofs uses kernel initiator, need the target

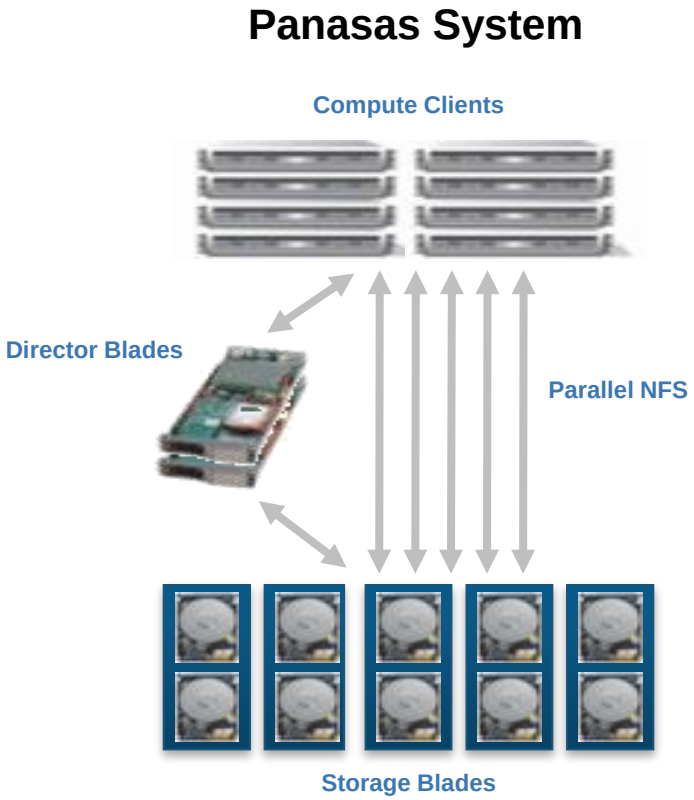
- **Questions: NFS mailing lists**
  - linux-nfs@kernel.org, nfsv4@ietf.org

“PNFS” TODAY WITH PANASAS



Parallel Storage:

File server not in data path.
Performance bottleneck eliminated.



- Panasas is designed to help remove the last bottleneck
- Compute clients access storage blades directly
- Full drive performance is delivered to clients

BASIC BUILDING BLOCKS

4U Shelf

1 + 10 Configuration

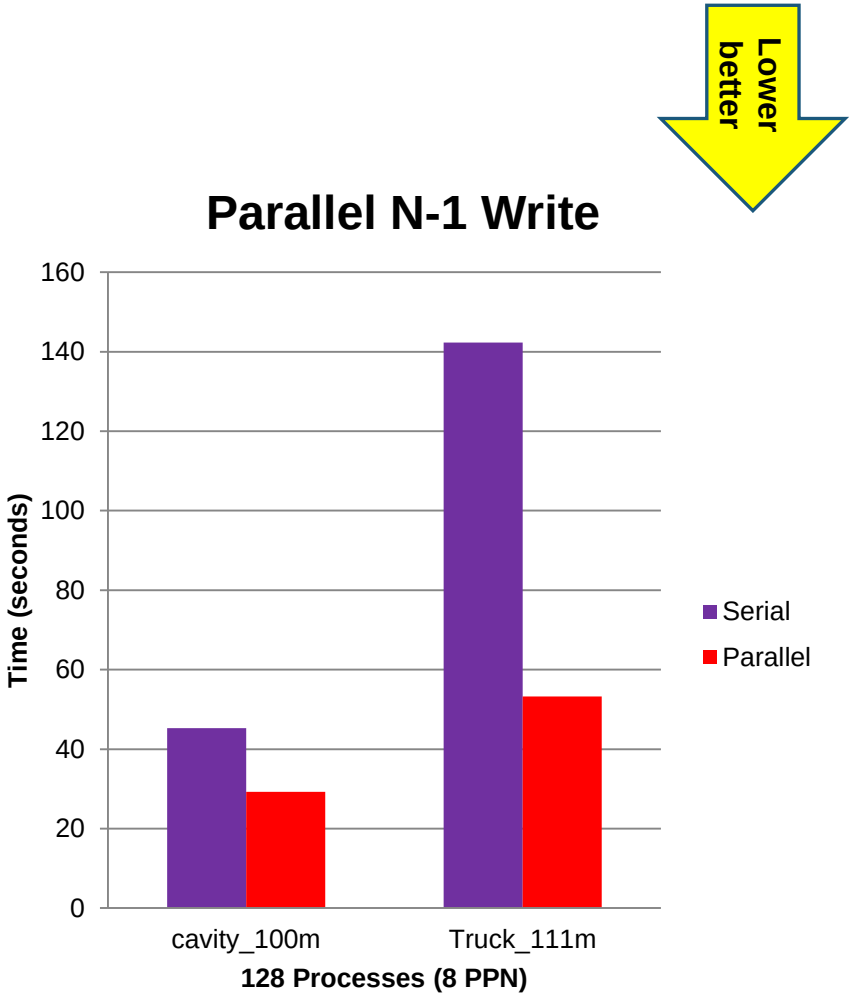
DirectorBlade

StorageBlade

ANSYS FLUENT V 13.0 PARALLEL (N-1 WRITE)

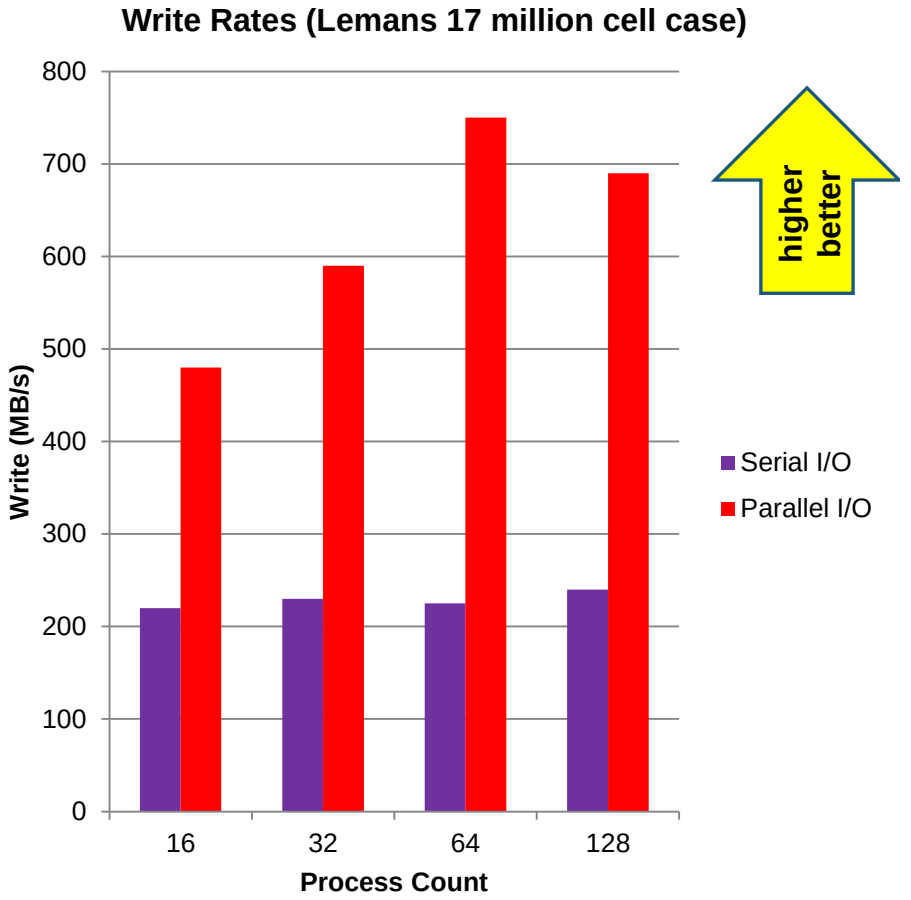
- **128 Processes running on 16 Compute Nodes (8 PPN)**
- **Benchmark (Test cases)**
  - Cavity_100m (synthetic benchmark)
  - Truck_111m (realistic workload)
- ***> 2.5 X over serial I/O with Truck_111m data set***

- **Clients:**
  - 16 Compute Nodes, each with Dual Socket Intel Xeon X5650
  - 24GB memory
  - 10GigE network
- **Storage –**
  - Single ActiveStor-11 shelf
  - 10GigE



STAR-CCM+ WRITE RESULTS (N-1 ACCESS)

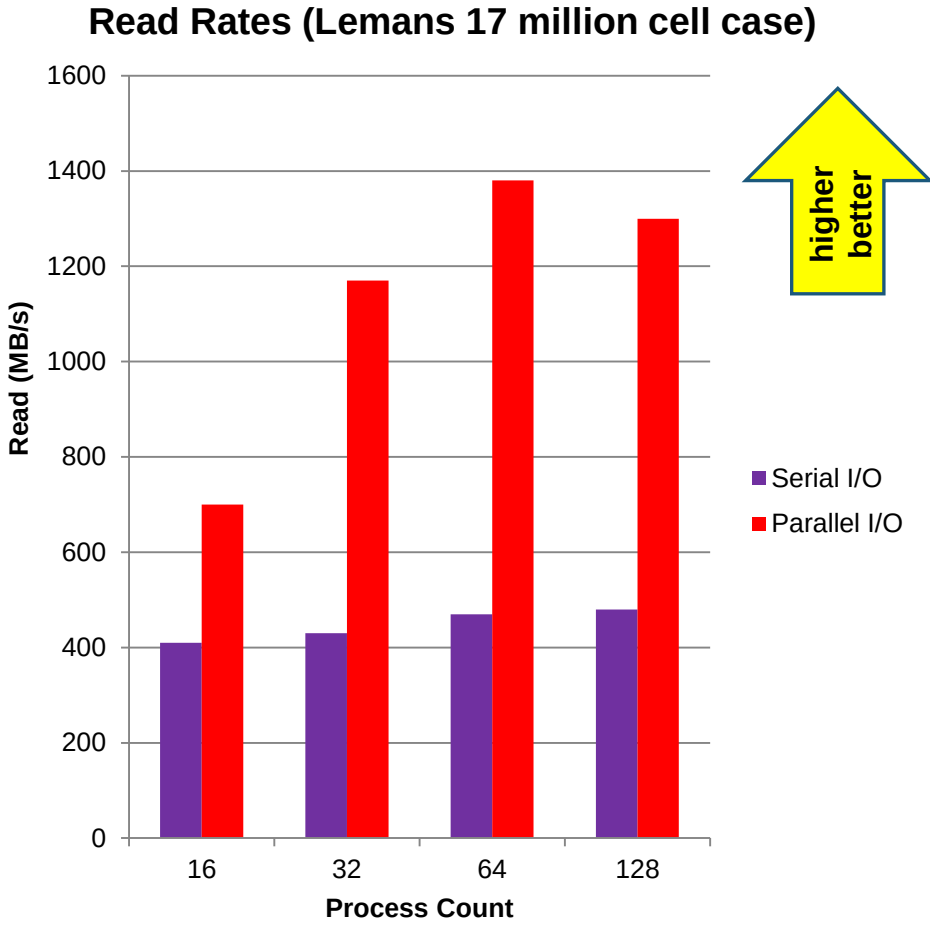
- **Serial I/O: constant as process count increases**
- **Parallel I/O: performance increases**
  - > 3X serial I/O at 64P
  - Low np limited by # of clients
  - High np limited by # of disks
- **Storage subsystem:**
  - PAS8: 4 shelves, 33 SBs
  - 10GigE
  - PanFS™ version 3.5



Courtesy of CD-adapco, Inc

STAR-CCM+ READ RESULTS (N-1 ACCESS)

- **Similar trends to writes:**
  - Serial I/O: constant as process count increases
  - Parallel I/O: performance increases
  - Low np limited by # of clients
  - High np limited by # of disks
- **Read rates > write rates**
  - N-1 write has higher overhead due to coherency



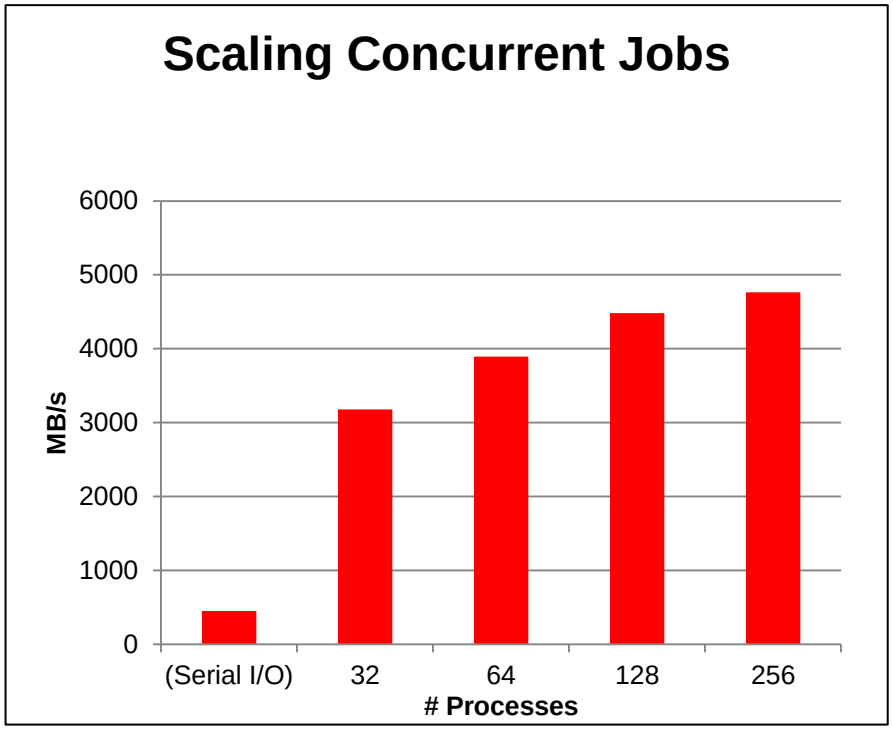
Courtesy of CD-adapco, Inc

LANDMARK PROMAX/SEISSPACE

HALLIBURTON

panasas

- **Parallel I/O throughput: multiple concurrent jobs**
- **Hardware:**
  - 4 Panasas AS12 shelves
  - 32 8-core compute nodes
- **Landmark ProMAX/SeisSpace generating geoseismic trace data**
- **Parallel I/O is done using JavaSeis (Opensource)**



SUMMARY

- **As cluster computing (HW) continues to grow, software is catching up to avoid slowest link (Amdahl's law)**
  - Software = File system, applications, I/O
  - ISVs are coming
- **Parallel system provides higher applications performance**
  - Faster time to market
  - Increase ROI
- **Panasas gives you parallel file system and parallel I/O today.**

Thank You