

## ***An Analysis of State-of-the-Art Parallel File System for Linux***

**Martin W. Margo, Patricia A. Kovatch,  
Phil Andrews, Bryan Banister  
San Diego Supercomputer Center**



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER



## ***San Diego Supercomputer Center***

- **University of California - San Diego**
- **National Science Foundation (NSF) Supercomputer Center**
- **Scientists and researchers nationwide apply for computational time on our resources**
- **Allocations board reviews applications and grants time**
- **SDSC Resources:**
  - 20 TeraFlops computational resources (Linux, AIX)
  - 600 TeraBytes SAN rotating disk (T3, T4, Minnow)
  - 1400 SAN switch ports (Brocade)
  - 1 PetaByte data stored in archival storage (HPSS, SAMQFS)
  - 6 PetaByte archival storage capacity (5 StorageTek silos)
  - 30 Gb/s connection to 40 Gb/s TeraGrid backbone



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER



## ***Linux HPC trend***

- **Linux is getting a lot of interest as OS of choice in High Performance Computing (HPC) clusters**
- **As commodity processors get faster, memory gets larger, storage gets bigger, and network achieves bigger bandwidth...**
- **There is a need for high performance parallel file systems for Linux to fully utilize all available storage resources**



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER



## ***Parallel File System***

- **Definition:**  
**A file system which enables each process from each cluster node to read from and write to a common storage target**
- **We need parallel file systems to accommodate I/O intensive scientific applications. Some applications, such as Enzo (an astrophysics application) outputs massive amount of data (1 TB / hour)**



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER



## ***Parallel File System for Linux***

- **Here are some parallel file systems for Linux:**
  1. Clemson University and Argonne National Laboratories (ANL) PVFS (Parallel Virtual File System)
  2. IBM's GPFS (General Parallel File System)
  3. Cluster File System Lustre (linux cluster)
  4. Sestina GFS (Global File System)
- **We chose to evaluate the first 3**



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER



## ***PVFS***

- **Open source, Linux only**
- **Uses any kind of file systems supported by Linux, can not use block devices**
- **Stripes data over all the disks**
- **3 components**
  - Metadata server (MGR)
  - Input Output Daemon (IOD)
  - PVFS client
- **All transactions (files, etc) travel over Gigabit ethernet**



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER



## **GPFS**

- **Proprietary**
- **Use any kind of block devices**
- **SAN mode**
  - All nodes are servers and can see all disks
  - Metadata traffic travels over Gigabit ethernet
  - Actual I/O travels over SAN
- **NSD mode**
  - NSD servers see all disks
  - NSD servers exports file system to all nodes over Gigabit ethernet
  - Metadata and actual I/O travel over Gigabit ethernet



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER



## **Lustre**

- **Open source**
- **Can create file system from any block device**
- **3 components:**
  - Metadata Server (MDS)
  - OST (Object Storage Target) servers
  - Lustre client
- **Portals lightweight messaging layer**



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER



## *Analysis criteria*

- **We think that the following criteria are important for a parallel file system:**
  - Ease of installation and administration
  - Redundancy
  - Performance
  - Scalability
  - Special Features



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER



## *Testbed*

- **To analyze each parallel file system, we built a test cluster of 20 nodes with these specs:**
  - Dual Intel Itanium 2 (Madison) @ 1.5 GHz
  - 4 GB of RAM
  - 2 10,000 RPM 73 GB IBM SCSI drives
  - Dual built-in gigabit ethernet adapters
  - Qlogic 2340 HBA (Host Bus Adapter)
  - Myrinet adapter
- **Storage Area Network (SAN) infrastructure**
  - Brocade SilkWorm 3900 32 port Fibre Channel switch
  - 5 brick pairs of Sun StorEdge 3510 “minnow”

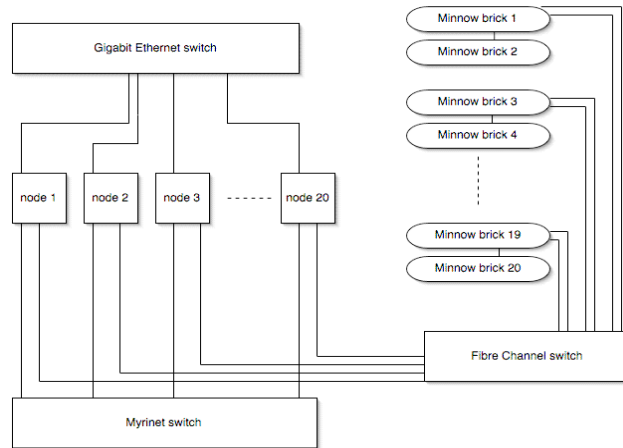


NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER



## Testbed Diagram



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER



## Installation & Administration

- **PVFS**
  - Straightforward to install
- **GPFS**
  - Straightforward to install
- **Lustre**
  - A challenge to install, requires custom kernel



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER



## *Redundancy*

- **PVFS**
  - Depends on underlying storage system (i.e. RAID), no built in redundancy feature
- **GPFS**
  - SAN mode: independent to node failures with core quorum, disk failures are prevented with RAID-5
  - NSD mode: failure groups prevents file system failures
- **Lustre**
  - Failover MDS server prevent single point of failure
  - Standby backup OST provides redundant path to storage target



## *Performance*

- **IOR benchmark utility from LLNL (MPI-IO over Myrinet)**
  1. Read/write 1 GB from each client with varying block sizes (16 KB - 8 MB)
  2. Use the optimum block size from previous test, each node read/write from (4 MB - 8 GB)
- **2 rounds of tests**
  1. All nodes are clients and servers
    - PVFS and GPFS SAN mode
  2. Two nodes are servers and the rest are clients
    - PVFS, GPFS NSD mode, and Lustre





NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

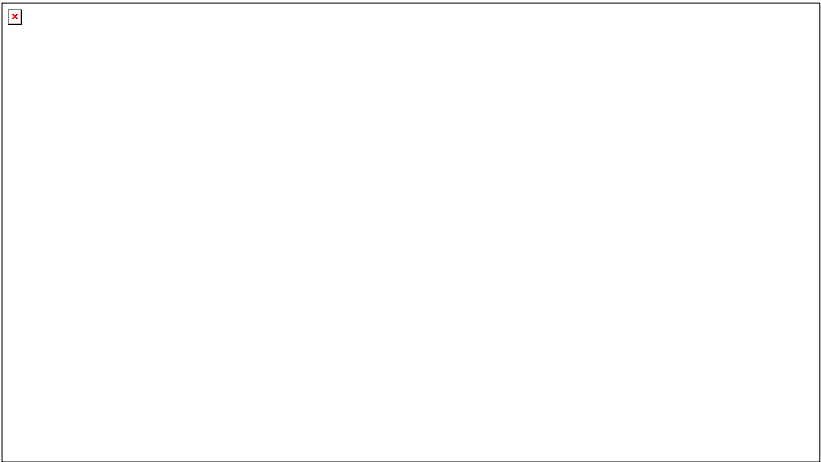
SAN DIEGO SUPERCOMPUTER CENTER



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER





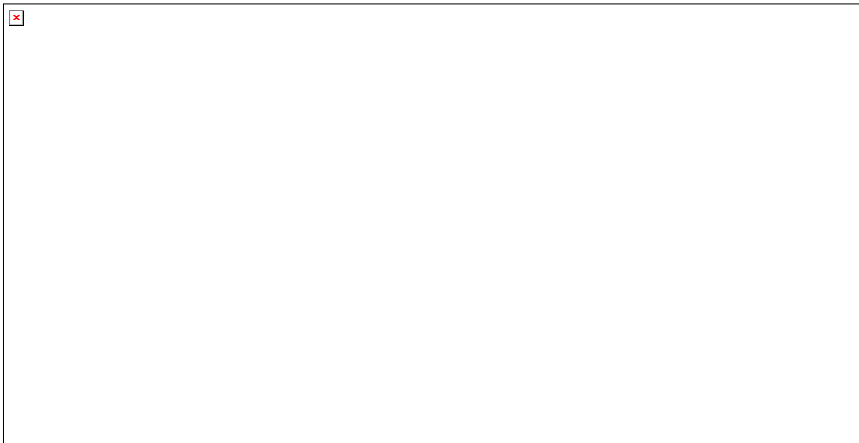
## *Scalability*

- We use mpirun on top of IOR to scale our clients from 2 nodes to 18 nodes, (17 in case of Lustre). Each client reads and writes 1 GB using the parallel file system's optimal block size
- We ran 2 rounds of tests



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

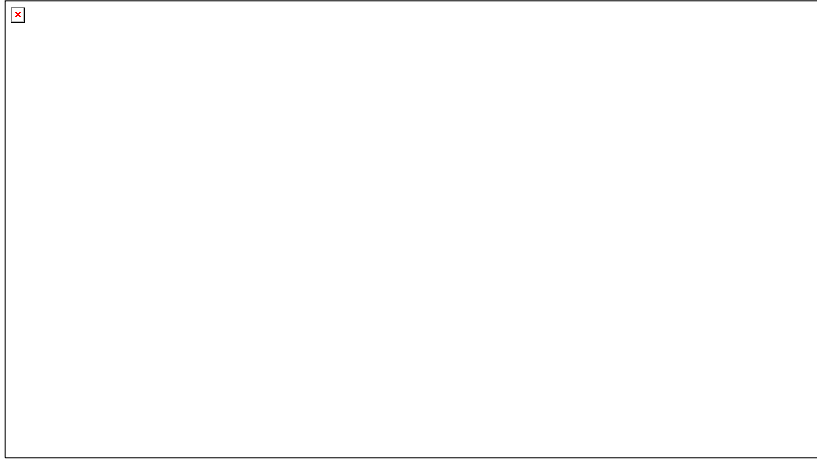
SAN DIEGO SUPERCOMPUTER CENTER



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER





## *Special Features*

- **PVFS**
  - Built in Romio (MPI-IO implementation) support
- **GPFS**
  - Native SAN support
  - Linux-AIX interoperability
- **Lustre**
  - Intent-based locking system
  - Portal lightweight messaging layer

## *Conclusion*

- **Something for everyone?**
  - PVFS and Lustre are open source
  - PVFS and GPFS are easy to install
  - GPFS and PVFS are mature
  - GPFS and Lustre have many redundancy features
  - Performance is reasonable for all 3 file systems
  - Lustre is designed to scale up to tens of thousands of clients
  - All file systems have WAN capabilities



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER



## *Acknowledgement*

- **Special thanks to**
  - Donald Thorp, SAN manager, SDSC
  - Haisong Cai, SAN engineer, SDSC
  - Christopher Jordan, HPC System Engineer, SDSC



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER



# *Questions & Answers*



NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER

