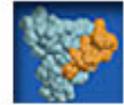
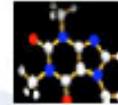




SciDAC

Scientific Discovery through Advanced Computing



SciDAC PDSI Update

FSIO Workshop 2008, August 6, Arlington VA

Garth Gibson

Carnegie Mellon University and Panasas Inc.

SciDAC Petascale Data Storage Institute (PDSI)

www.pdsi-scidac.org

w/ LANL (G. Grider), LBNL (W. Kramer), SNL (L. Ward),
ORNL (P. Roth), PNNL (E. Felix),
UCSC (D. Long), U.Mich (P. Honeyman)

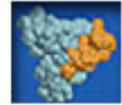
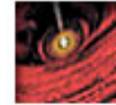
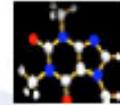
Carnegie Mellon
Parallel Data Laboratory

 **pdsi**



SciDAC

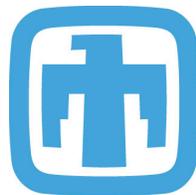
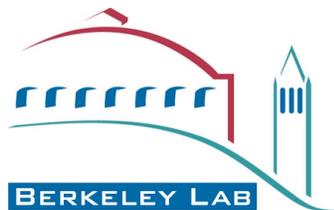
Scientific Discovery through Advanced Computing



- PETASCALE DATA STORAGE INSTITUTE 06-11
 - 3 universities, 5 labs, G. Gibson, CMU, PI
- SciDAC @ Petascale storage issues
 - www.pdsi-scidac.org
 - Community building: ie. PDSW @ SCxy
 - APIs & standards: ie., Parallel NFS, POSIX
 - Failure data collection, analysis: ie., cfduserenix.org
 - Performance trace collection & benchmark publication
 - IT automation applied to HEC systems & problems
 - Novel mechanisms for core (esp. metadata, wide area)



Carnegie Mellon



Sandia National Laboratories



Pacific Northwest National Laboratory

Operated by Battelle for the U.S. Department of Energy

Carnegie Mellon Parallel Data Laboratory

<http://www.pdsi-scidac.org/>



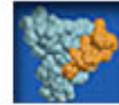
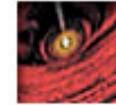
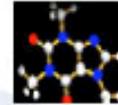
UNIVERSITY OF MICHIGAN





SciDAC

Scientific Discovery through Advanced Computing



- PDSI @ FSIO08 so far
 - Mgmt: Greg and Priya
 - Metadata: Garth and Ethan
- PDSI @ FSIO07
 - Failure data: cfd.r.usenix.org, and SciDAC07



Carnegie Mellon



Sandia National Laboratories



Pacific Northwest National Laboratory
Operated by Battelle for the U.S. Department of Energy

Carnegie Mellon
Parallel Data Laboratory

<http://www.pdsi-scidac.org/>



UNIVERSITY OF MICHIGAN



Annual PDSI Sponsored Workshops

institutes.lanl.gov/hec-fsio/workshops/2007/

HEC FSIO '07

HEC FSIO R&D Workshop/HECURA FSIO PI Meeting '07 AGENDA

Workshop Location: National S

Session

Monday 8/6/2007

Welcome Review of HEC FSIO 06 outcomes, F 2007 Workshop Overview

Welcome from NSF

NSF Vision

Research Session 1 QoS

Quality of Service Guarantee for Scalable For scalable Parallel Storage Systems

End-to-End Performance Management for Large Distributed Storage

Open review of gaps/progress

LANL ISSDM and IRPIT

LANL New Data Available

Research Session 2

Measurement, Understading, Cache Mgmt

File System Tracing, Replaying, Profiling, and Analysis on HEC Systems

Memory caching and prefetching

Open review of gaps/progress

Research Session 3 Metadata

Petascale I/O for High End Computing

Techniques for Streaming File Systems And Databases

Microdata Storage Systems for High-End Computing

SAM^2 Toolkit: Scalable and Adaptive Metadata Management for High End Computing

Open review of gaps/progress

Research Session 4 Security and Archive

Asymmetry in Performance and Security Requirements for I/O in HEC

Integrated Infrastructure for Secure and Efficient Long-Term Data Management

Open review of gaps/progress

Posters for all Day 1 talks

Tuesday 8/7/2007

Use of Xen for Testing File Systems At Scale

Research Session 5 Next Generation I/O Architectures

Deconstructing Clusters for High End Biometrics

August

www.pdsi-scidac.org/sc07/



Supercomputing '07

Petascale Data Storage Workshop
Session Chair: Garth Gibson, CMU

Sunday, November 11, 2007
Reno, Nevada

WORKSHOP ABSTRACT

Petascale computing infrastructures make petascale demands on information and manageability. The last decade has shown that parallel file systems can scale to exa-dimensions; this poses a critical challenge when near-future petascale requires the data storage problems and emerging solutions found in petascale scientific community collaboration can be crucial, problem identification, workload characterization, and shared tools.

Petascale Data Storage Workshop Introduction
Garth Gibson

SESSION I: Scalable Systems

E. Krevat (presenter), V. Vasudevan, A. Phanishayee, D. Andersen, G. Ganger, G. Gibson, S. Seshan, Carnegie Mellon University
On Application-level Approaches to Avoiding TCP Throughput Collapse in Cluster-Based Storage Systems
Paper / Slides / Poster

Lei Chai, Xiangyong Quyang, Ranjit Noronha (presenter) and Dhabaleswar K. Panda, Ohio State University
pNFS/PVFS2 over InfiniBand: Early Experiences
Paper / Slides

Brent Welch (presenter), Panasas, Inc.
Integrated System Models for Reliable Petascale Storage Systems
Paper / Slides

Peter Braam, Byron Neitzel (presenter), Sun/Cluster File Systems
Scalable Locking and Recovery for Network File Systems
Paper / Slides

POSTER SESSION 1 - see info below

SESSION II: Scalable Services

Jonathan Koren (presenter), Yi Zhang, Univ. of California, Santa Cruz
Searching and Navigating Petabyte Scale File Systems Based on Facets
Paper / Slides

Swapnil V. Patil (presenter), Garth A. Gibson, Sam Lang, Milo Polte, Carnegie Mellon University
GIGA+: Scalable Directories for Shared File Systems
Paper / Slides / Poster

D. Bigelow, S. Iyer, T. Kaldewey, R. Pineiro, A. Povzner, S. Brandt, R. Golding (presenter), T. Wong, C. Maltzahn, Univ. of California, Santa Cruz, IBM-Almaden
End-to-end Performance Management for Scalable Distributed Storage
Paper / Slides

Sage A. Weil (presenter), Andrew W. Leung, Scott A. Brandt, Carlos Maltzahn, Univ. of California, Santa Cruz
RADOS: A Fast, Scalable, and Reliable Storage Service for Petabyte-scale Storage Clusters
Paper / Slides

November

www.pdsi-scidac.org/events/FAST08BOF/

FAST '08

Wednesday, February 27, 2008
Petascale Data Storage BoF Session at FAST '08

Organizer: Garth Gibson, Carnegie Mellon University and Panasas
Co-organizers: Peter Honeyman, U. Michigan/CITI; Darrell Long, U.C. Santa Cruz; Gary Grider, Los Alamos NL; Lee Ward, Sandia NL; Evan Felix, Pacific Northwestern NL; Phil Roth, Oak Ridge NL; Bill Kramer, Lawrence Berkeley NL

The Petascale Data Storage Institute is a DOE-funded collaboration of three universities and five national labs with the objective of anticipating the challenges of data storage for computing systems operating in the peta-operations per second to exa-operations per second and working toward the resolution of these challenges in the community as a whole. An important part of our agenda is outreach to other researchers and practitioners to share our resources and gather better understanding of the petascale issues ahead from all.

In this BOF we will:

- 1) Introduce the Petascale Data Storage Institute (PDSI),
- 2) Advertise PDSI gathered and released sources of useful data, including
 - data sets of node and storage failures in large scale computing
 - file access traces of non-trivial petascale computing applications
 - collections of file systems statistics gathered from petascale computing systems and other systems,
- 3) Discuss requirements for one or more petascale data storage systems and applications, and
- 4) Lead an open discussion of these and other issues for large scale data storage systems.

PRESENTATIONS

PDSI FAST 2008 BOF Introduction - Garth Gibson, CMU

The Computer Failure Data Repository (CFDR) - Bianca Schroeder, University of Toronto

File System Statistics - Shobhit Dayal, CMU, Garth Gibson, CMU, Marc Unangst, Panasas

PNNL - Petascale Data Storage Institute Data release Update - Evan Felix, PNNL

NERSC Reliability Data - Bill Kramer, Jason Hick, Akbar Mokhtarani, NERSC

LANL SciDAC Petascale Data Storage Institute Operational Data Releases - James Nunez, Gary Grider, John Bent, HB Chen, Meghan Quist, Alfred Torrez, Los Alamos National Lab

Ceph: An Open-Source Petabyte-Scale File System - Ethan Miller, Storage Systems Research Center, UCSanta Cruz

Special Presentation on HPC User Requirements:

I/O Requirements for HPC Applications: A User Perspective
John Shaif, National Energy Research Scientific Computing Center (NERSC), LBNL

PDSI POSTER AT THE FAST '08 POSTER SESSION

PDSI Data Releases and Repositories

February



Sponsored by USENIX
in cooperation with
ACM SIGOPS,
IEEE MISTC,
and IEEE TCOS

USENIX

PDSW07 papers are published

THE ACM DIGITAL LIBRARY

 [Feedback](#)

Conference on High Performance Networking and Computing [archive](#)

Proceedings of the 2nd international workshop on Petascale data storage: held in conjunction with Supercomputing '07
2007, Reno, Nevada November 11 - 11, 2007

Additional Information: [full citation](#)

Conference Chair [Garth A. Gibson](#) Carnegie Mellon University and Panasas Inc.

Front matter

 [pdf](#)

Front matter (Title page, TOC, Committee, Author index)

Table of Contents

SESSION: [Scalable systems](#)

[On application-level approaches to avoiding TCP throughput collapse in cluster-based storage systems](#)

Elie Krevat, Vijay Vasudevan, Amar Phanishayee, David G. Andersen, Gregory R. Ganger, Garth A. Gibson, Srinivasan Seshan

Pages 1-4

Full text available:  [Pdf](#)(115 KB)

Additional Information: [full citation](#), [abstract](#), [references](#), [index terms](#)

[pNFS/PVFS2 over InfiniBand: early experiences](#)

Lei Chai, Xiangyong Ouyang, Ranjit Noronha, Dhableswar K. Panda

Pages 5-11

Full text available:  [Pdf](#)(128 KB)

Additional Information: [full citation](#), [abstract](#), [references](#), [index terms](#)

[Integrated system models for reliable petascale storage systems](#)

Brent Welch

Pages 12-16

Full text available:  [Pdf](#)(105 KB)

Additional Information: [full citation](#), [abstract](#), [references](#), [index terms](#)

[Scalable locking and recovery for network file systems](#)

Peter J. Braam

Pages 17-20

Full text available:  [Pdf](#)(317 KB)

Additional Information: [full citation](#), [abstract](#), [index terms](#)

SESSION: [Scalable services](#)

Carnegie Mellon
Parallel Data Laboratory



PDSW07 talks are online



9:00am - 10:20am **SESSION I: Scalable Systems**

E. Krevat (presenter), V. Vasudevan, A. Phanishayee, D. Andersen, G. Ganger, G. Gibson, S. Seshan, Carnegie Mellon University
On Application-level Approaches to Avoiding TCP Throughput Collapse in Cluster-Based Storage Systems
[Paper](#) / [Slides](#) / [Poster](#)

Lei Chai, Xiangyong Ouyang, Ranjit Noronha (presenter) and Dhabaleswar K. Panda,
Ohio State University
pNFS/PVFS2 over InfiniBand: Early Experiences
[Paper](#) / [Slides](#)

Brent Welch (presenter), Panasas, Inc.
Integrated System Models for Reliable Petascale Storage Systems
[Paper](#) / [Slides](#)

Peter Braam, Byron Neitzel (presenter), Sun/Cluster File Systems
Scalable Locking and Recovery for Network File Systems
[Paper](#) / [Slides](#)

10:30am - 11:00am **POSTER SESSION 1 - see info below**

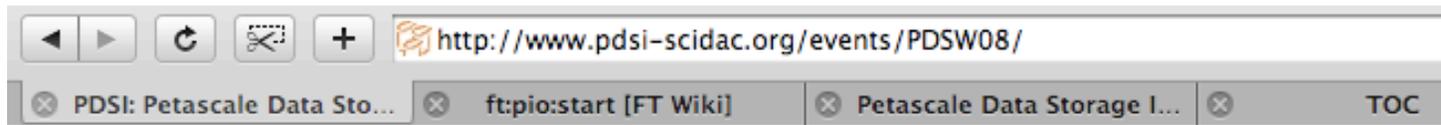
11:00am - 12:20pm **SESSION II: Scalable Services**

Jonathan Koren (presenter), Yi Zhang, Sasha Ames, Andrew Leung, Carlos Maltzahn, Ethan Miller, Univ. of California, Santa Cruz
Searching and Navigating Petabyte Scale File Systems Based on Facets
[Paper](#) / [Slides](#)

Swapnil V. Patil (presenter), Garth A. Gibson, Sam Lang, Milo Polte, Carnegie Mellon University
GIGA+: Scalable Directories for Shared File Systems
[Paper](#) / [Slides](#) / [Poster](#)

D. Bigelow, S. Iyer, T. Kaldewey, R. Pineiro, A. Povzner, S. Brandt, R. Golding (presenter), T. Wong, C. Maltzahn, Univ. of California, Santa Cruz, IBM-Almaden
End-to-end Performance Management for Scalable Distributed Storage
[Paper](#) / [Slides](#)

PDSW08 call for papers is out



CALL FOR PAPERS

This workshop seeks contributions on relevant topics, including but not limited to: performance and benchmarking results and tools, failure tolerance problems and solutions, APIs for high performance features, parallel file systems, high bandwidth storage architectures, wide area file systems, metadata intensive workloads, autonomies for HPC storage, virtualization for storage systems, archival storage advances, resource management innovations, etc.

Paper Submission Webpage: [link coming soon](#)

Paper (extended abstract in pdf format) due Fri Sept. 26, 2008

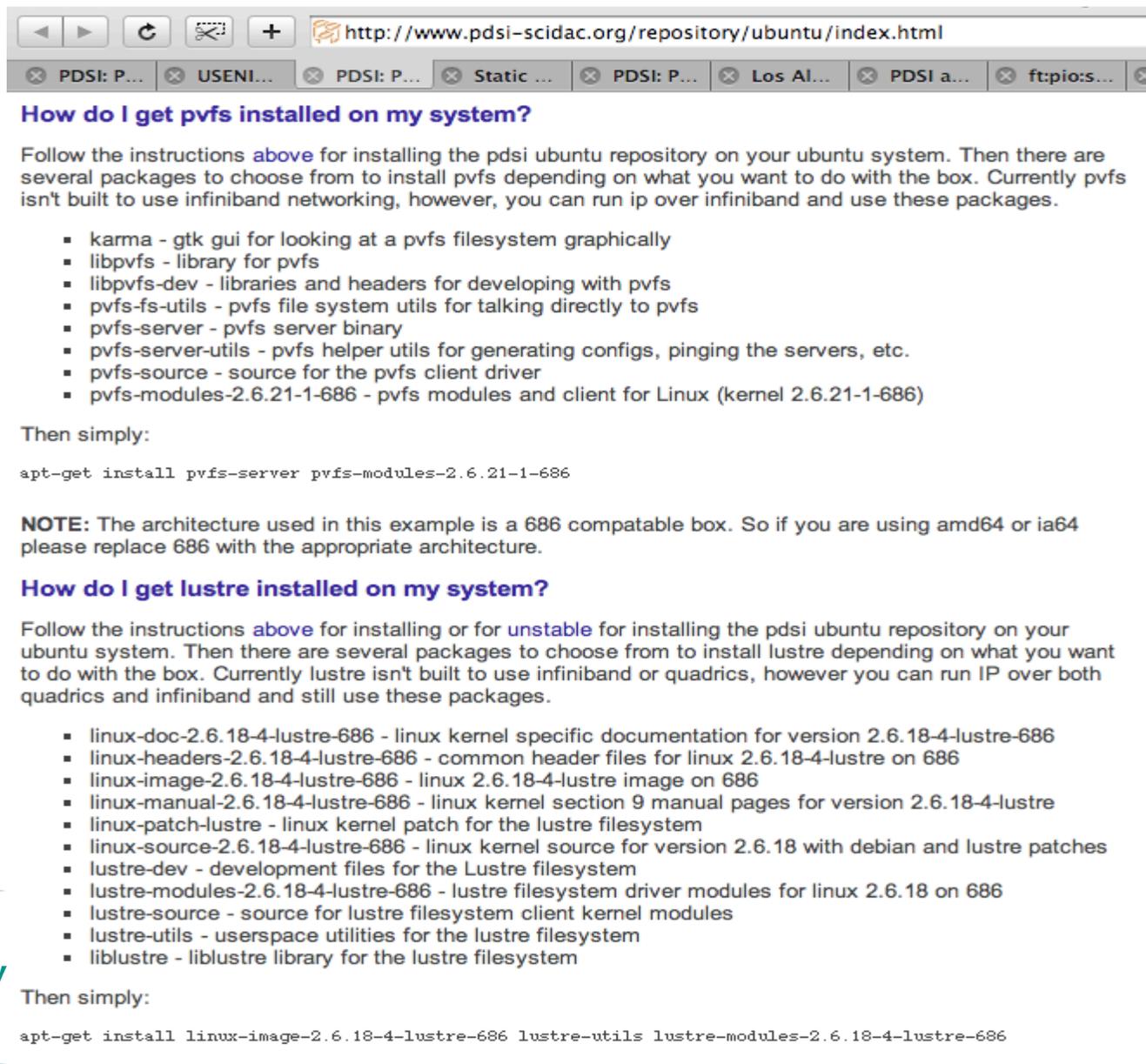
Notification: Mon Oct. 20, 2008

Softcopy and slides due Nov. 16 *BEFORE* the workshop

Paper Submission Details:

The petascale data storage workshop holds a peer reviewed competitive process for selecting extended abstracts and short papers. Submit a not previously published extended abstract of 2 to 5 pages, not less than 10 point font, in a PDF file as instructed on the workshop web site. Submitted papers will be reviewed under the supervision of the workshop program committee. Submissions should indicate authors and affiliations. Selected papers and associated talks will be made available on the workshop web site. Selected final papers may be longer than in submission, but not longer than 10 pages. The workshop proceedings will be published in association with SC08 in the IEEE digital library and talk slides will be made available on the workshop web site.

PDSI distributes convenient packages



How do I get pvfs installed on my system?

Follow the instructions [above](#) for installing the pdsi ubuntu repository on your ubuntu system. Then there are several packages to choose from to install pvfs depending on what you want to do with the box. Currently pvfs isn't built to use infiniband networking, however, you can run ip over infiniband and use these packages.

- karma - gtk gui for looking at a pvfs filesystem graphically
- libpvfs - library for pvfs
- libpvfs-dev - libraries and headers for developing with pvfs
- pvfs-fs-utils - pvfs file system utils for talking directly to pvfs
- pvfs-server - pvfs server binary
- pvfs-server-utils - pvfs helper utils for generating configs, pinging the servers, etc.
- pvfs-source - source for the pvfs client driver
- pvfs-modules-2.6.21-1-686 - pvfs modules and client for Linux (kernel 2.6.21-1-686)

Then simply:

```
apt-get install pvfs-server pvfs-modules-2.6.21-1-686
```

NOTE: The architecture used in this example is a 686 compatible box. So if you are using amd64 or ia64 please replace 686 with the appropriate architecture.

How do I get lustre installed on my system?

Follow the instructions [above](#) for installing or for [unstable](#) for installing the pdsi ubuntu repository on your ubuntu system. Then there are several packages to choose from to install lustre depending on what you want to do with the box. Currently lustre isn't built to use infiniband or quadrics, however you can run IP over both quadrics and infiniband and still use these packages.

- linux-doc-2.6.18-4-lustre-686 - linux kernel specific documentation for version 2.6.18-4-lustre-686
- linux-headers-2.6.18-4-lustre-686 - common header files for linux 2.6.18-4-lustre on 686
- linux-image-2.6.18-4-lustre-686 - linux 2.6.18-4-lustre image on 686
- linux-manual-2.6.18-4-lustre-686 - linux kernel section 9 manual pages for version 2.6.18-4-lustre
- linux-patch-lustre - linux kernel patch for the lustre filesystem
- linux-source-2.6.18-4-lustre-686 - linux kernel source for version 2.6.18 with debian and lustre patches
- lustre-dev - development files for the Lustre filesystem
- lustre-modules-2.6.18-4-lustre-686 - lustre filesystem driver modules for linux 2.6.18 on 686
- lustre-source - source for lustre filesystem client kernel modules
- lustre-utils - userspace utilities for the lustre filesystem
- liblustre - liblustre library for the lustre filesystem

Then simply:

```
apt-get install linux-image-2.6.18-4-lustre-686 lustre-utils lustre-modules-2.6.18-4-lustre-686
```

PDSI distributes parallel workloads

MPI-IO Test

Although there are a host of existing file system and I/O test programs, most are not designed with parallel I/O in mind and are not useful at the clusters at Los Alamos National Lab (LANL). LANL's MPI-IO Test was designed with parallel I/O and scale in mind. The MPI-IO test is built on top of MPI and is used to gather timing information for reading from and writing to using a variety of I/O profiles; N processes writing to N files, N processes writing to one file, N processes sending data to M processes writing to M files, or N processes sending data to M processes to one file. These diagrams illustrate various I/O access patterns. A data aggregation capability is available and can pass down MPI-IO, ROMIO and file system specific hints. The MPI-IO Test can be used for performance benchmarking and, in some cases, to diagnose problems with file systems or I/O networks.

The MPI-IO Test is open sourced under LA-CC-05-013.

Release	Date	Source	Documentation
1.000.21	8 July 2008	mpi_io_test_21.tgz	README
1.000.20	13 November 2007	mpi_io_test_20.tgz	README
1.000.09	15 December 2006	mpi_io_test_09.tgz	README
1.000.08	2 March 2006	mpi_io_test_08.tgz	README

MPI_IO_TEST traces

These traces were collected using LANL-Trace (V 1.0.0) on the LANL MPI-IO test (V 1.00.020) application. These traces are all from system data machine number 25 on this [computer systems table](#). Here is the README and FAQ that explains how LANL-Trace works and what the output files look like:

[TRACE README](#),
[TRACE FAQ](#).

N-to-N

	64 KB	256 KB	448 KB	512 KB	1024 KB	4096 KB	8192 KB	16386 KB	32772 KB	65544 KB
32 Procs		TGZ	TGZ	TGZ	TGZ	TGZ	TGZ	TGZ	TGZ	TGZ
96 Procs		TGZ	TGZ	TGZ	TGZ	TGZ		TGZ	TGZ	TGZ

N-to-1 nonstrided

	64 KB	256 KB	448 KB	512 KB	1024 KB	4096 KB	8192 KB	16386 KB	32772 KB	65544 KB
32 Procs	TGZ		TGZ	TGZ	TGZ	TGZ	TGZ	TGZ	TGZ	TGZ
96 Procs	TGZ	TGZ		TGZ	TGZ	TGZ		TGZ	TGZ	TGZ

N-to-1 strided

	64 KB	256 KB	448 KB	512 KB	1024 KB	4096 KB	8192 KB	16386 KB	32772 KB	65544 KB
32 Procs	TGZ	TGZ	TGZ	TGZ	TGZ	TGZ	TGZ	TGZ	TGZ	TGZ
96 Procs	TGZ	TGZ	TGZ	TGZ	TGZ	TGZ		TGZ	TGZ	TGZ

PDSI distributes parallel workloads

MADBench: Microwave Anisotropy Dataset Computational Analysis Package Benchmark

The benchmark code MADBench is a "stripped-down" version of MADCAP, a Microwave Anisotropy Dataset Computational Analysis Package [...more>>>](#)

IPM benchmarks: [Medium](#), [Large](#) and [X-large](#) datasets.

MILC: MIMD Lattice Computation

The benchmark code MILC represents part of a set of codes written by the MIMD Lattice Computation (MILC) collaboration used to study quantum chromodynamics (QCD), the theory of the strong interactions of subatomic physics [...more>>>](#)

IPM benchmarks: [Medium](#) and [Large](#) datasets.

PMEMD: Particle Mesh Ewald Molecular Dynamics

The benchmark code PMEMD (Particle Mesh Ewald Molecular Dynamics (MD), NMR Refinement and minimizations [...more>](#)

IPM benchmarks: [Medium](#) and [Large](#) datasets

IO Benchmarks with IPM*

The new version of IPM integrates the standard POSIX IO call runs are made with this new feature on [Jacquard](#) (courtesy of

MADBench:

- 256 tasks, POSIX one file per task [\[plots\]](#) [\[stats\]](#)
- 64 tasks, POSIX one file per task [\[plots\]](#) [\[stats\]](#)
- 16 tasks, POSIX shared file [\[plots\]](#) [\[stats\]](#)

Chombo:

- 256 tasks, 2 components [\[plots\]](#) [\[stats\]](#)
- 32 tasks, 2 components [\[plots\]](#) [\[stats\]](#)
- 32 tasks, 10 components [\[plots\]](#) [\[stats\]](#)

AMRScalingXfer: 128 tasks, small run [\[plots\]](#) [\[stats\]](#)

*Note: This is development software, and the runs/plots are profiling in IPM.

Trace Data

Here are files containing trace data for some of the applications. These traces are generated by invoking the "strace" utility on every task and piping the data for each task to a separate file. Process ID is used to create unique file names. All applications were run on [Jacquard](#). The files are compressed tar files of the trace data

[PMEMD 16 tasks small dataset run](#)

[MADbench 64 tasks medium dataset run](#)

[MILC 16 tasks medium dataset run](#)

I/O Benchmark and Characterization Links:

I/O Performance for HPC Platform using IOR [PDF ppt](#)

This study analyzes the I/O practices and requirements of current HPC applications and use them as criteria to select a subset of microbenchmarks that reflect workload requirements.

FLASH I/O Benchmark [PDF](#)

This code from 'The Center for Astrophysical Thermonuclear Flashes' can test either HDF5, Parallel NetCDF, or a direct Fortran write. The I/O benchmarks are compared for Seaborg and Bassi systems

Performance Effect of Multi-core on Scientific Applications (PDF) [paper slides](#)

Presents performance measurements of several complete scientific applications on single and dual core Cray XT3 and XT4 systems.

MADBench - IPM of a Cosmology Application on Leading HEC Platforms [PDF](#)

Presents MADBench, a lightweight version of MADCAP CMB power spectrum estimation code, and uses the Integrated Performance Monitoring (IPM) package to extract MPI message-passing overheads

MADBench2 [PDF](#)

Presents I/O analyses of modern parallel filesystems and examines a broad range of system architectures and configurations. It also describes use of Luster striping to improve concurrent file access performance.

Effective I/O Bandwidth Benchmark [PDF](#)

This paper describes the design and implementation of a parallel I/O benchmark useful for comparing filesystem performance on a variety of architectures, including, but not limited to cluster systems.

Efficient Parallel I/O on the Cray XT3/XT4 [PDF](#)

Provides an overview of I/O methods for three different applications

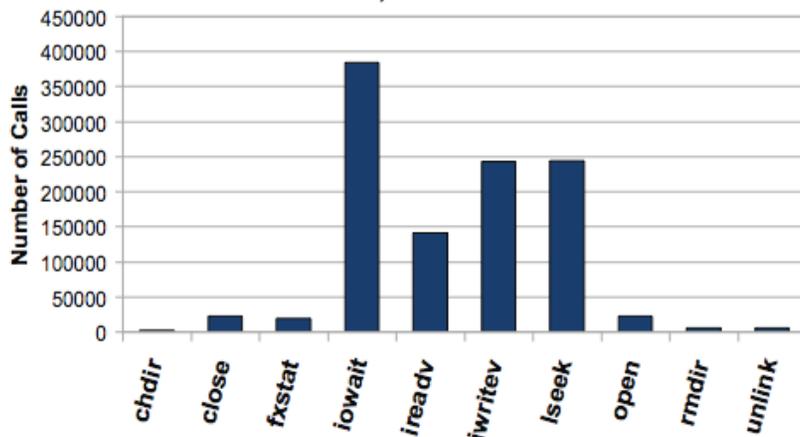


PDSI distributes parallel workloads

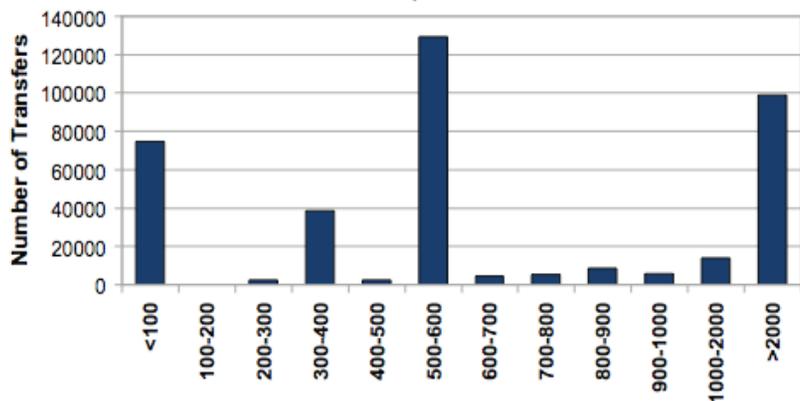
NEWEST TRACE DATA, REDSTORM, SANDIA NAT'L LAB

- A physics simulation problem for a common Sandia application, Alegria
 - Runs were performed alongside regular user runs
- Each run generated 4 restart dumps, and ran for 20 simulation cycles
- Both single core per node, and 2 core (virtual node mode) per node
 - Repeated with and without tracing enabled
- The single core per node jobs ran at a client size of 2744 processes
 - Non-tracing elapsed run time 10:42 minutes
 - Tracing elapsed run time 11:07 minutes
- The 2 core per node jobs ran at 2916 nodes, 5832 processes.
 - Non-tracing elapsed run time 15:52 minutes
 - Tracing elapsed run time 16:37 minutes
- Raw trace file sizes 30K-50K per MPI rank, except rank zero (600KB-700KB)
 - Rank 0 I/O to terminal records progress in the job.

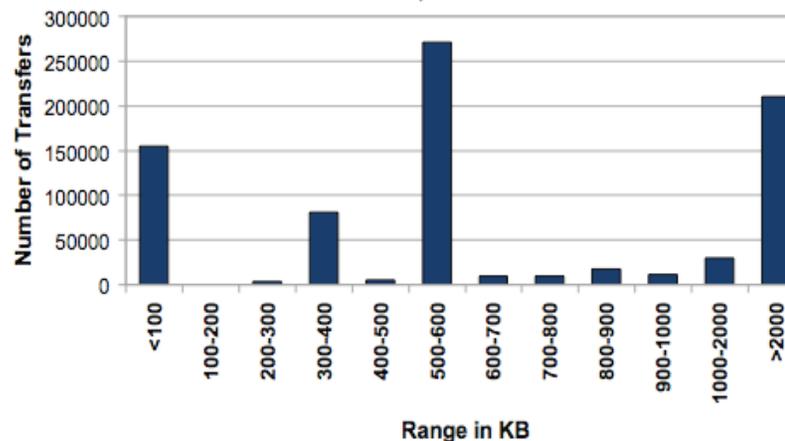
I/O calls, 2744 Processes



I/O Transfers, 2744 Processes



I/O Transfers, 5832 Processes



Sandia
National
Laboratories

sourceforge.net/projects/libsysio



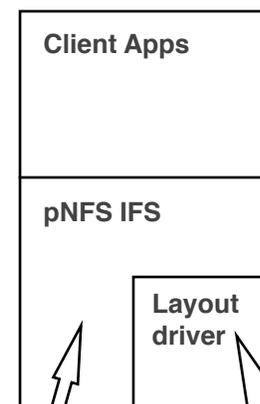
pNFS: scalable NFS standard & code soon

From: Tigran Mkrtchyan <tigran.mkrtchyan@desy.de>
Date: July 16, 2008 4:18:13 AM PDT
To: pnfs@linux-nfs.org
Subject: [pnfs] pnfs becomes real!

today we ran the first real physics analysis job using dCache-pnfs server and linux pnfs client:

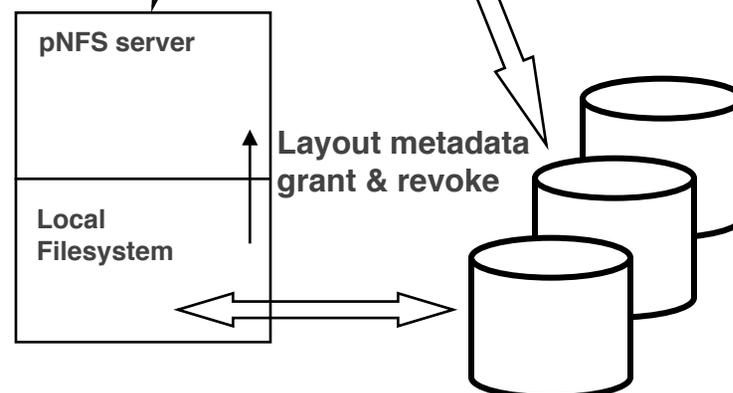
```
tigran@nairi:~/work/linux-pnfs> git show | head -5  
commit 6ae52464ba2c77f1bf2365e415305dfd9b51dd20  
Author: Benny Halevy <bhalevy@panasas.com>  
Date: Tue Jul 15 20:22:51 2008 +0300
```

Anyway, fist time we can show that NFSv4.1 is something real (and not my hobby only).



NFSv4 extended
w/ orthogonal
layout metadata
attributes

1. SBC (blocks)
2. OSD (objects)
3. NFS (files)



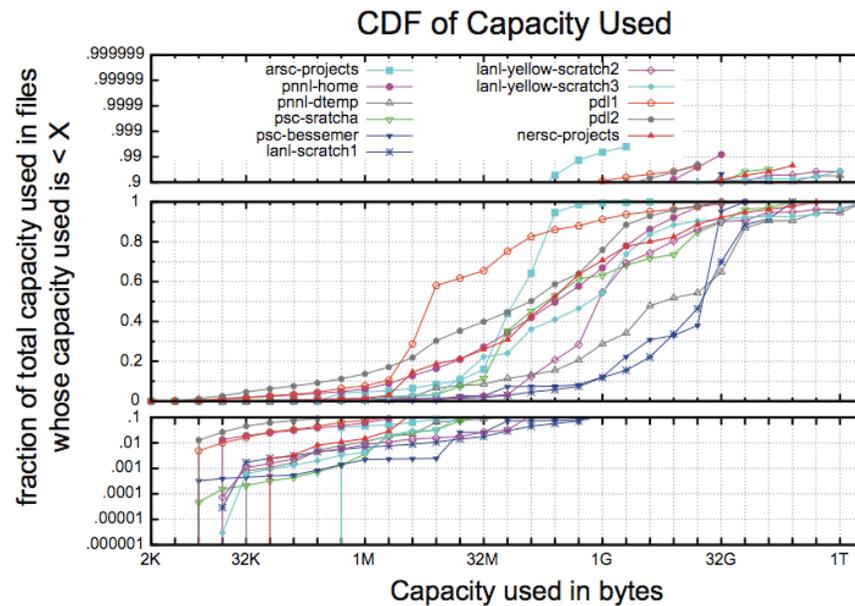
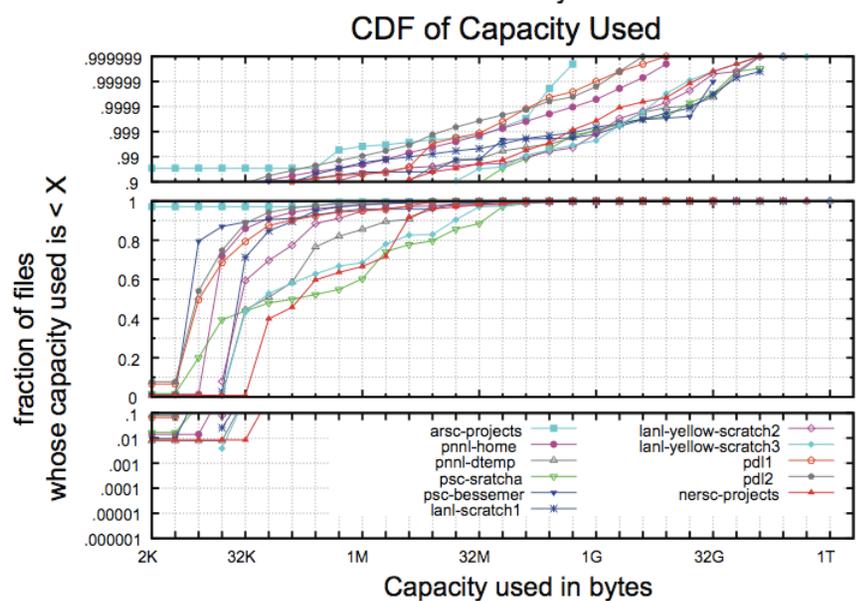
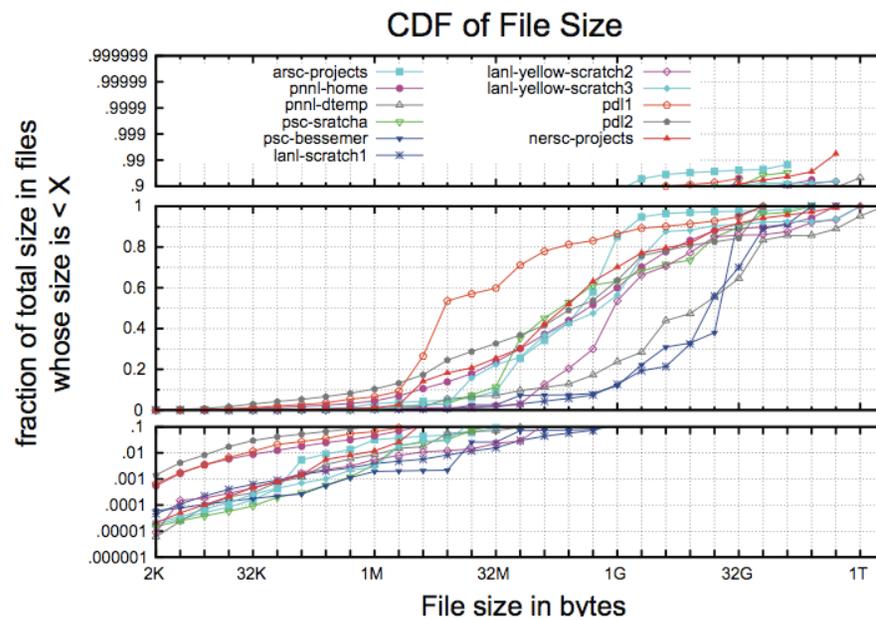
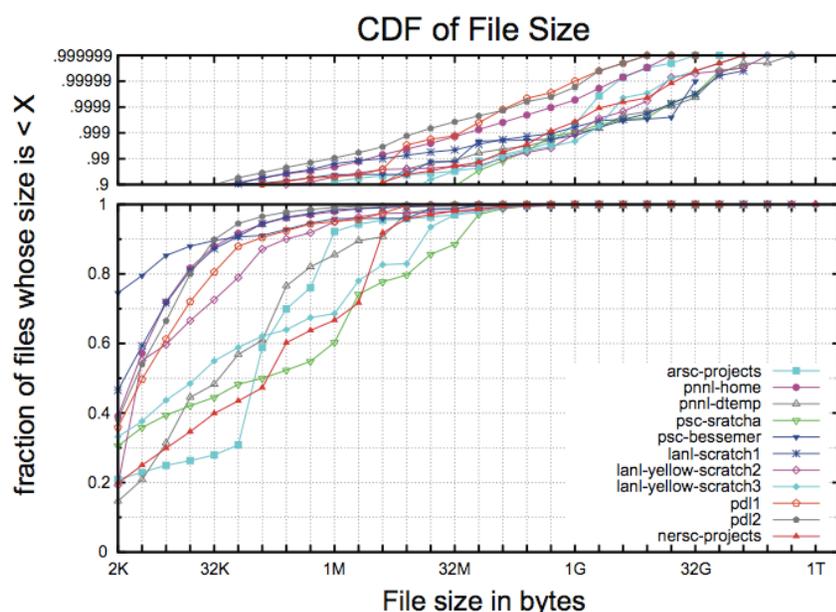
From: Spencer Shepler <Spencer.Shepler@Sun.COM>
Date: August 1, 2008 4:34:46 PM GMT-04:00

2. IETF status

All of the current working group internet drafts are moving forward for publication. This means that they have submitted to the area director and will start their way through the process (IETF last call and IESG review).



PDSI gathers HPC file system statistics



www.pdsi-scidac.org/fsstats

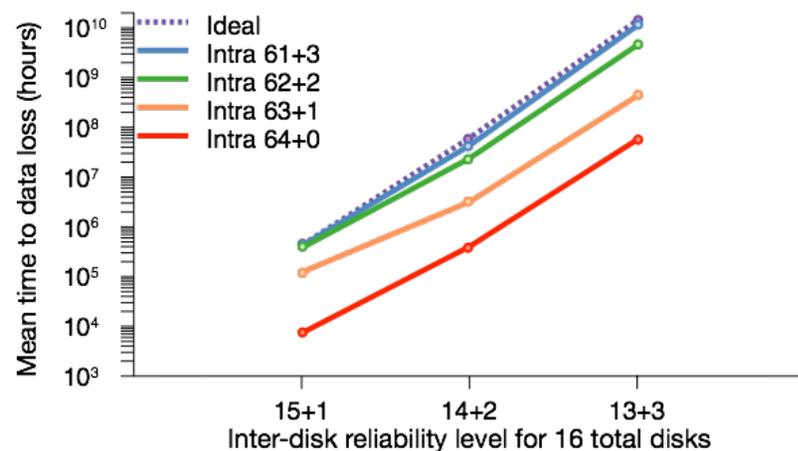
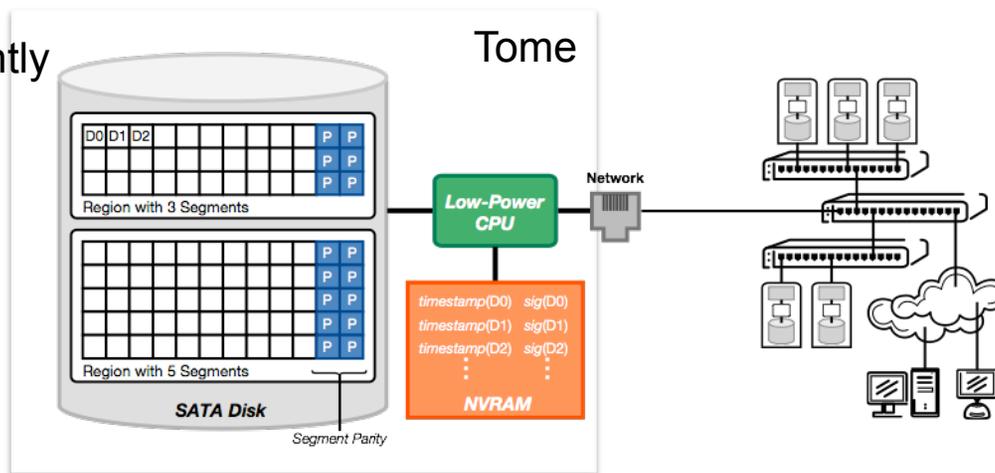
Results

Uploaded File in CSV format	Organization	Date	Data Size	System Name	Form Questions	Formatted Result	Graphs	Graphs	fsstats Version
BradHavel_nanu1.csv	ARSC	Mar122008	69TB	SAMQFS	Form	Histograms	PS	PDF	1.4
BradHavel_seau2.csv	ARSC	Mar132008	115TB	SAMQFS	Form	Histograms	PS	PDF	1.4
BradHavel_seau1.csv	ARSC	Mar132008	305TB	SAMQFS	Form	Histograms	PS	PDF	1.4
BradHavel_nanprojects.csv	ARSC	Mar132008	32TB	SAMQFS	Form	Histograms	PS	PDF	1.4
JamesNunez_panscratch1.csv	LANL	Apr012008	9TB	Panasas	Form	Histograms	PS	PDF	1.4
JamesNunez_yellowscratch2.csv	LANL	Apr042008	25TB	Panasas	Form	Histograms	PS	PDF	1.4
JamesNunez_yellowscratch3.csv	LANL	Apr042008	26TB	Panasas	Form	Histograms	PS	PDF	1.4
AkbarMokhtarani_NGFFsstats.csv	NERSC	Apr082008	107TB	GPFS	Form	Histograms	PS	PDF	1.4
PhilRoth_fsstats.csv	ORNL	Oct102007	305GB	Panasas	Form	Histograms	PS	PDF	1.4
MichaelStroucken_pdl2.csv	PDL	Apr092008	1TB	WAFI	Form	Histograms	PS	PDF	1.4
MichaelStroucken_pdl1.csv	PDL	Apr092008	4TB	WAFI	Form	Histograms	PS	PDF	1.4
EvanFelix_dtemp.csv	PNNL	Mar172008	23TB	Lustre	Form	Histograms	PS	PDF	1.4
EvanFelix_nwfs.csv	PNNL	Mar172008	265TB	Lustre	Form	Histograms	PS	PDF	1.4
EvanFelix_home.csv	PNNL	Mar172008	5TB	ADVFS	Form	Histograms	PS	PDF	1.4
EvanFelix_mpp2dtemp.csv	PNNL	Oct102007	12TB	ext3	Form	Histograms	PS	PDF	1.4
EvanFelix_nwfs.csv	PNNL	Oct102007	233TB	ext3	Form	Histograms	PS	PDF	1.4
EvanFelix_mpp2home.csv	PNNL	Oct102007	4TB	advfs	Form	Histograms	PS	PDF	1.4
Katie_scratch1.csv	PSC	Mar272008	32TB	Lustre	Form	Histograms	PS	PDF	1.4
Katie_bessemer1.csv	PSC	Mar272008	4TB	Lustre	Form	Histograms	PS	PDF	1.4

[A Comparative graph of some of the above results EPS](#)
[A Comparative graph of some of the above results PDF](#)
[A Comparative graph of archival file system EPS](#)
[A Comparative graph of archival file systems PDF](#)

Developing reliable, evolvable archives

- ▶ Evolvable, distributed network of intelligent, disk-based *tomes*
 - ▶ Smart enough to function independently
 - ▶ Provide inter-disk redundancy
 - ▶ Building blocks for more complex systems
 - ▶ Evolve over time: integrate new technologies
- ▶ Handle errors at multiple levels
 - ▶ Scale response to size of problem
 - ▶ Very high reliability!
- ▶ Control costs
 - ▶ Commodity low-power hardware
 - ▶ Keep disks spun down
 - ▶ Standardized interfaces



ory



Revisiting checkpoint: Log representation

- Fastest checkpoint might be a sequential series of “variable=value”
 - Instead of seeking to serialized location, just append operation to log
 - Each thread writes strictly sequential log of operations
 - “Meaning” of set of logs is applying log to (possibly null) initial database
- Prior: Gatech/ORNL ADIOS, ANL summer project
 - Brainstorming with SciDAC SDM on application to pHDF5/netCDF
- Decouple writing logs from applying logs to serialized database
 - Optimize each separately; pipeline from compute to IO nodes
 - Defer serializing by just storing changelogs for later application
 - Some checkpoints never read, so never serialized
 - If read before serialized, trigger serialization (or something smarter)
 - Represent logs as attributes of database; that is, hide in FS (directory?)

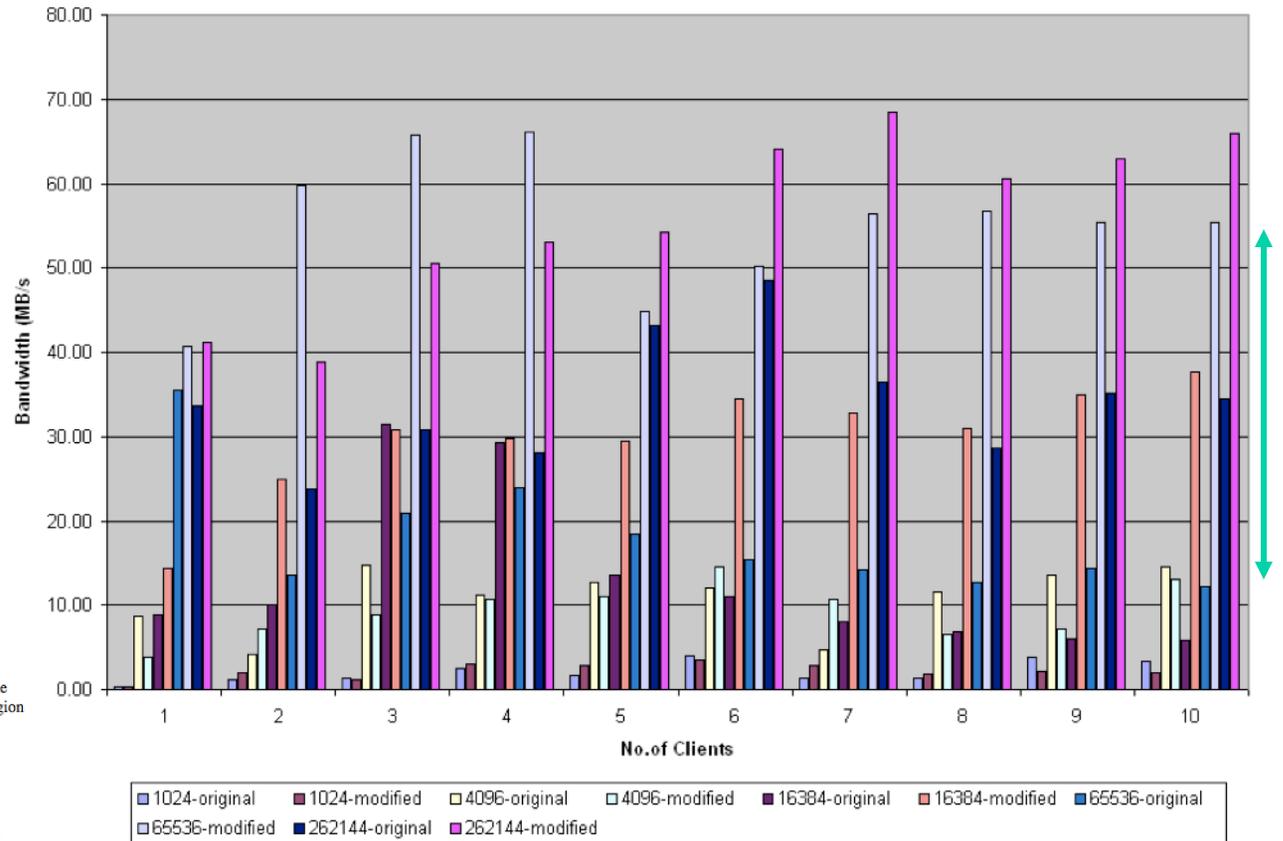
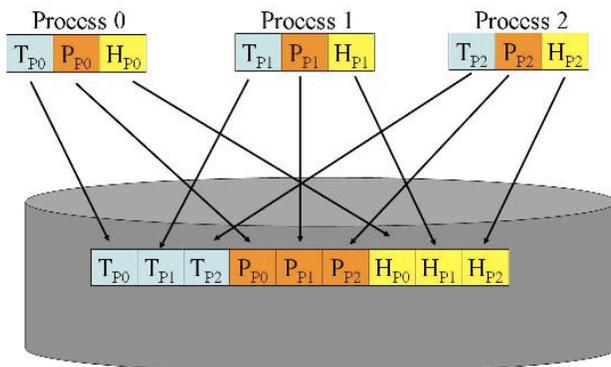
CMU class project: log-structured PVFS files

- HPC checkpoints
 - AMR apps are non-sequential concurrent writers
 - Lousy BW
 - Store file as log of writes
 - Good BW

N to 1 strided

Each process writes each element in a single shared "stride" within a single shared file. The file consists of one region per element (not one region per process as in N-1 non strided). Each region contains "strided" data from each process.

N-to-1 strided example



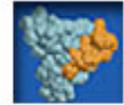
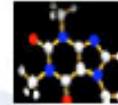
(c) Write Bandwidth of modified and unmodified PVFS with various numbers of clients and block sizes

- Group 8 mpi-io write test (from LANL)
 - S. Dayal, M. Chainani, D.K. Uppugandi, W. Tantosiroj



SciDAC

Scientific Discovery through Advanced Computing



SciDAC PDSI Update

FSIO Workshop 2008, August 6, Arlington VA

Garth Gibson

Carnegie Mellon University and Panasas Inc.

SciDAC Petascale Data Storage Institute (PDSI)

www.pdsi-scidac.org

w/ LANL (G. Grider), LBNL (W. Kramer), SNL (L. Ward),
ORNL (P. Roth), PNNL (E. Felix),
UCSC (D. Long), U.Mich (P. Honeyman)

Carnegie Mellon
Parallel Data Laboratory

 **pdsi**