SciDAC
Scientific Discovery through Advanced Computing

ASCR

Office of Science

SciDAC Petascale Data Storage Institute

Advanced Scientific Computing Advisory Committee Meeting

October 29 2008, Gaithersburg MD

Garth Gibson

Carnegie Mellon University and Panasas Inc.

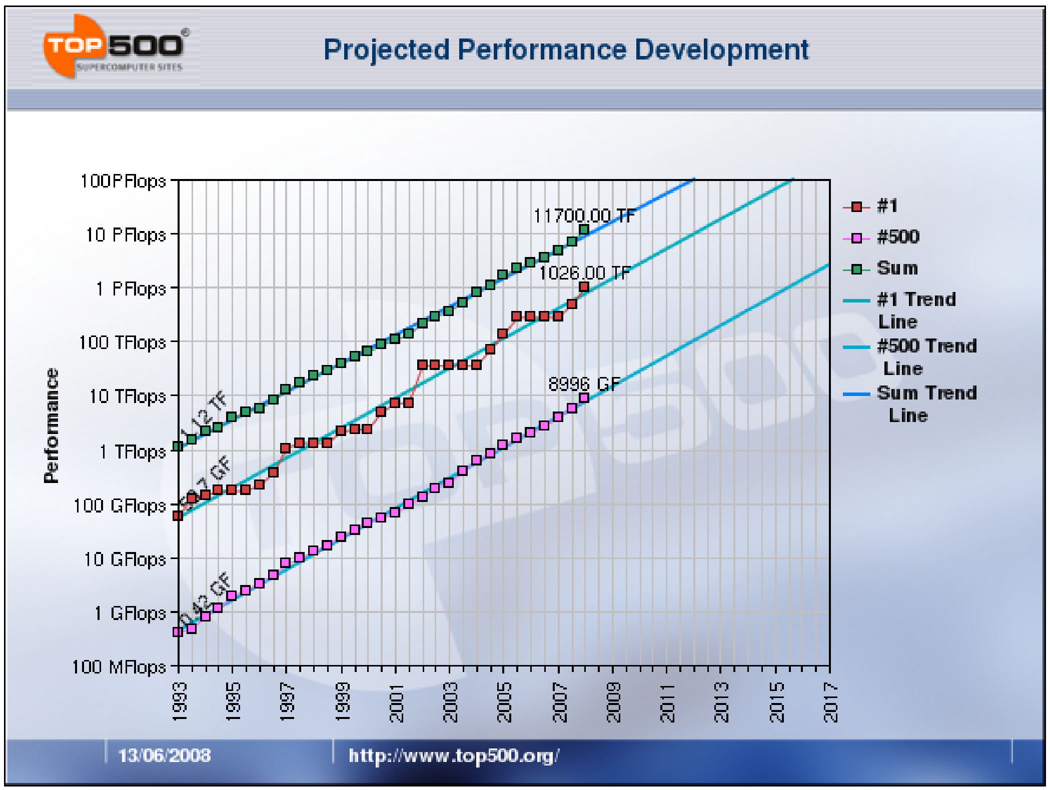
SciDAC Petascale Data Storage Institute (PDSI)

www.pdsi-scidac.org

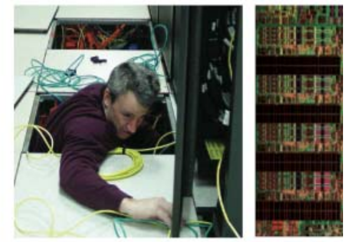
w/ LANL (Gary Grider), LBNL (William Kramer), SNL (Lee Ward),
ORNL (Phil Roth), PNNL (Evan Felix),
UCSC (Darrell Long), U.Mich (Peter Honeyman)

Charting Storage Path thru Peta- to Exa-scale

- Top500.org scaling 100% per year; storage required to keep up
  - This is hard for disks: MB/sec +20% per year, accesses/sec +5% per year
  - Increases number of disks much faster than processor chips or nodes



Los Alamos NATIONAL LABORATORY EST. 1943



Roadrunner

First to break the "petaflop" barrier

At 3:30 a.m. on May 26, 2008, Memorial Day, the "Roadrunner" supercomputer exceeded a sustained speed of 1 petaflop/s, or 1 million billion calculations per second. The sustained performance makes Roadrunner more than twice as fast as the current number 1 system on the TOP500 list. The best sustained performance to date is 74.5% efficiency, 1.026 petaflop/s.

SciDAC Petascale Data Storage Institute

- High Performance Storage Expertise & Experience
  - Carnegie Mellon University, Garth Gibson, lead PI
  - U. of California, Santa Cruz, Darrell Long
  - U. of Michigan, Ann Arbor, Peter Honeyman
  - Lawrence Berkeley National Lab, William Kramer
  - Oak Ridge National Lab, Phil Roth
  - Pacific Northwest National Lab, Evan Felix
  - Los Alamos National Lab, Gary Grider
  - Sandia National Lab, Lee Ward

SciDAC Petascale Data Storage Institute

- Efforts divided into three primary thrusts
- Outreach and leadership
  - Community building: ie. PDSW @ SC08, FAST, FSIO
  - APIs & standards: ie., Parallel NFS, POSIX Extensions
  - SciDAC collaborations: applications, centers, institutes
- Data collection and dissemination
  - Failure data collection, analysis: ie., cfdr.usenix.org
  - Performance trace collection & benchmark publication
- Mechanism innovation
  - Scalable metadata, archives, wide area storage, etc
  - IT automation applied to HEC systems & problems

Outreach: Sponsored Workshops

- SC08: PDSW, Mon Nov 17, 8-5

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

HEC FSIO '07 — August

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 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, Understadning, 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	

www.pdsi-scidac.org/SC07/

Supercomputing '07 — November

SC07

Petascale Data Storage Workshop
Session Chair: Garth Gibson, CMU

Sunday, November 11, 2007
Reno, Nevada

WORKSHOP ABSTRACT

Petascale computing infrastructures make petascale demands on informati and manageability. The last decade has shown that parallel file systems ca dimensions; this poses a critical challenge when near-future petascale requ the data storage problems and emerging solutions found in petascale scien community collaboration can be crucial, problem identification, workload ca 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 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

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

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

FAST '08 — February

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

FAST '08 — FEBRUARY 26–29, 2008, SAN JOSE, CALIFORNIA — Sponsored by USENIX in cooperation with ACM SIGOPS, IEEE MSSTC, and IEEE TCOS — USENIX

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 Shalf, National Energy Research Scientific Computing Center (NERSC), LBNL

PDSI POSTER AT THE FAST '08 POSTER SESSION

PDSI Data Releases and Repositories

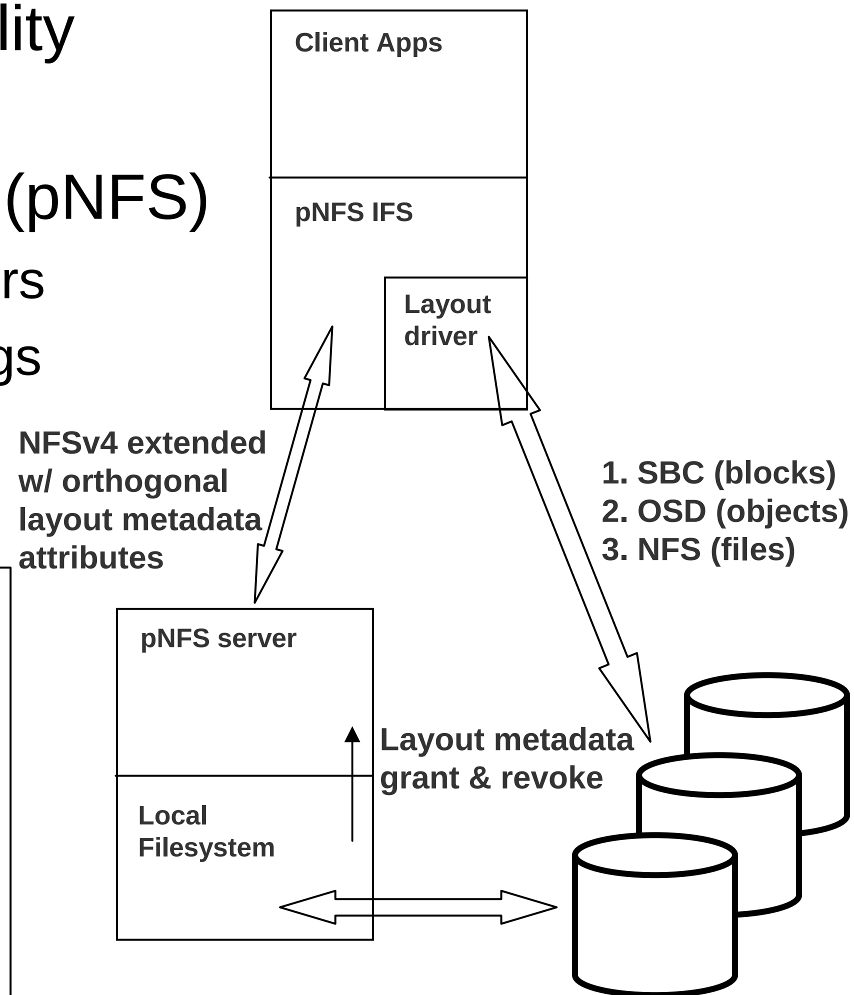
Standards: Multi-vendor, Scalable Parallel NFS

- Persistent investment in scalability
  - Share costs with commercial R&D
- Next generation NFS is parallel (pNFS)
  - Responds to growing role of clusters
  - Open source & competitive offerings
  - NetApp, Sun, IBM, EMC, Panasas, BlueArc, DESY/dCache

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).

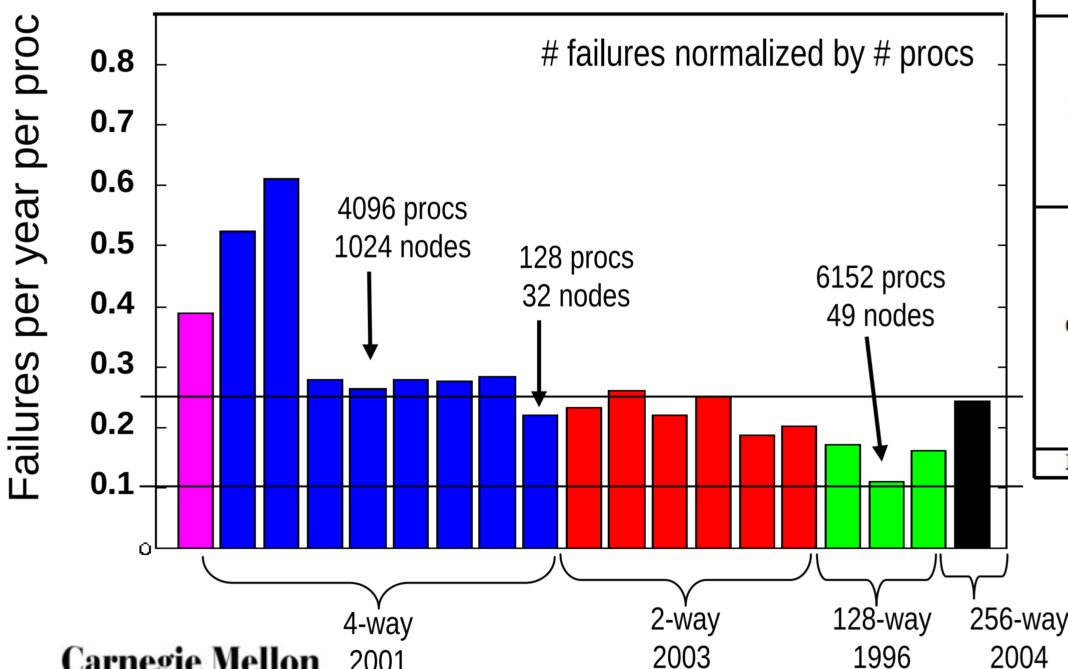


center for information technology integration
UNIVERSITY OF MICHIGAN

pdsi

Dissemination: Fault Data

- Los Alamos root cause logs
  - 22 clusters & 5,000 nodes
  - covers 9 years & continues
  - cfdr.usenix.org publication + PNNL, NERSC, Sandia, PSC, …



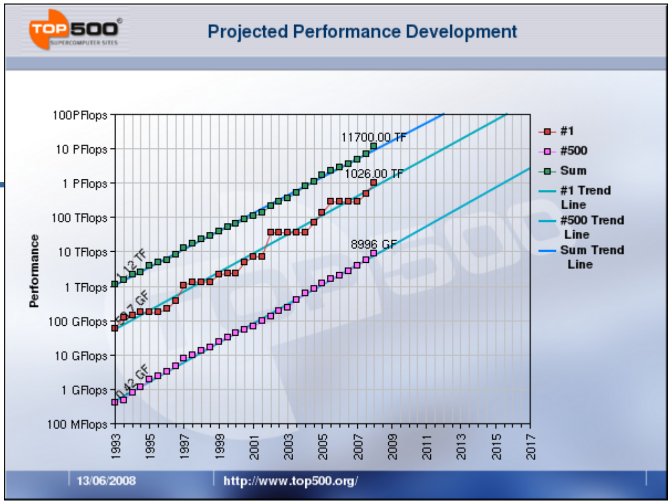
(I) High-level system information				(II) Information per node category			
HW	ID	Nodes	Procs	Procs /node	Production Time	Mem (GB)	NICs
A	1	1	8	8	N/A – 12/99	16	0
B	2	1	32	32	N/A – 12/03	8	1
C	3	1	4	4	N/A – 04/03	1	0
D	4	164	328	2	04/01 – now	1	1
				2	12/02 – now	1	1
E	5	256	1024	4	12/01 – now	16	2
	6	128	512	4	09/01 – 01/02	16	2
	7	1024	4096	4	05/02 – now	8	2
				4	05/02 – now	16	2
				4	05/02 – now	32	2
				4	05/02 – now	352	2
	8	1024	4096	4	10/02 – now	8	2
				4	10/02 – now	16	2
				4	10/02 – now	32	2
	9	128	512	4	09/03 – now	4	1
	10	128	512	4	09/03 – now	4	1
	11	128	512	4	09/03 – now	4	1
	12	32	128	4	09/03 – now	4	1
				4	09/03 – now	16	1
F	13	128	256	2	09/03 – now	4	1
	14	256	512	2	09/03 – now	4	1
	15	256	512	2	09/03 – now	4	1
	16	256	512	2	09/03 – now	4	1
	17	256	512	2	09/03 – now	4	1
	18	512	1024	2	09/03 – now	4	1
				2	03/05 – 06/05	4	1
G	19	16	2048	128	12/96 – 09/02	32	4
				128	12/96 – 09/02	64	4
	20	49	6152	128	01/97 – now	128	12
				128	01/97 – 11/05	32	12
				80	06/05 – now	80	0
	21	5	544	128	10/98 – 12/04	128	4
				32	01/98 – 12/04	16	4
				128	11/02 – now	64	4
				128	11/05 – 12/04	32	4
H	22	1	256	256	11/04 – now	1024	0

Table 1. *Overview oj SMP-based, and systen*

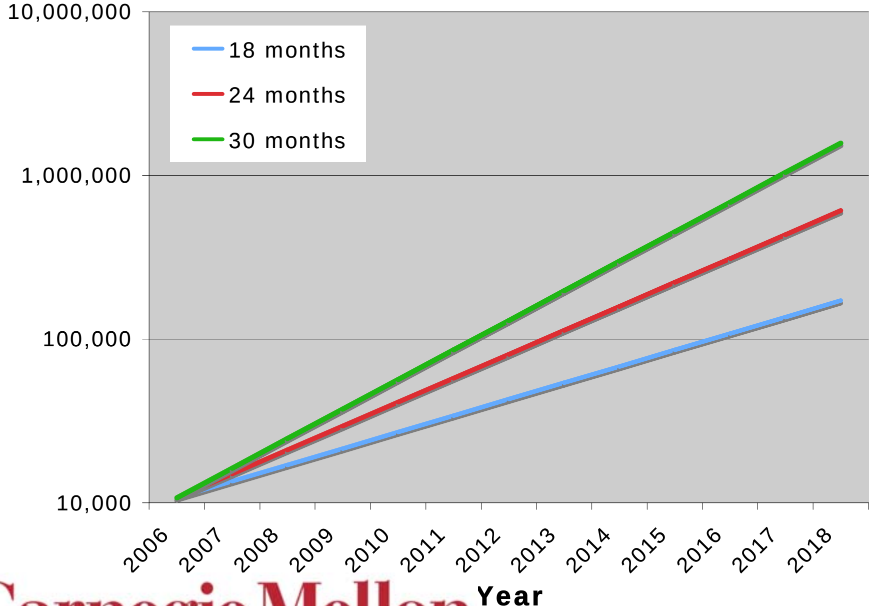
Carnegie Mellon
Parallel Data Laboratory

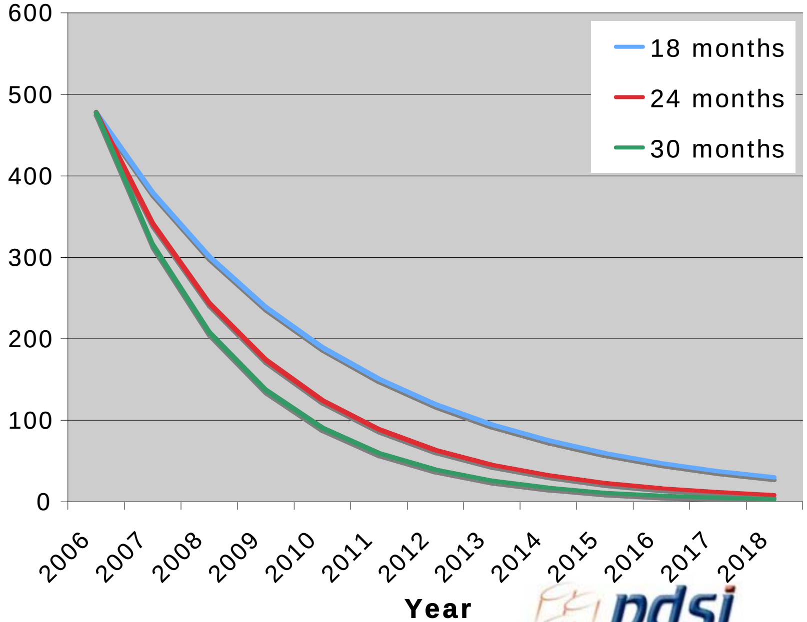
Los Alamos NATIONAL LABORATORY EST. 1943

Projections: More Failures



- Con't top500.org 2X annually
  - 1 PF Roadrunner, May 2008
- Cycle time flat, but more of them
  - Moore's law: 2X cores/chip in 18 mos
- # sockets, 1/MTTI = failure rate up 25%-50% per year
  - Optimistic 0.1 failures per year per socket (vs. historic 0.25)



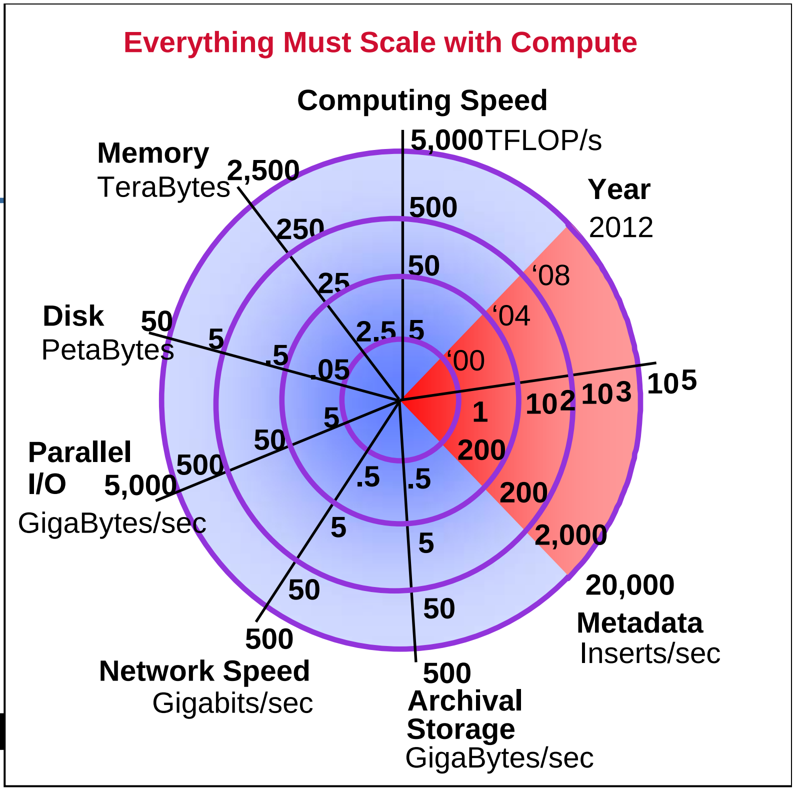


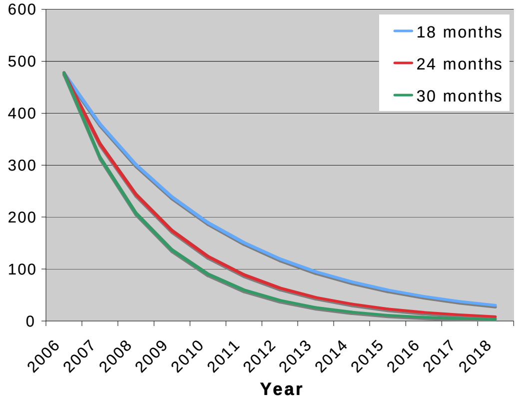
Carnegie Mellon

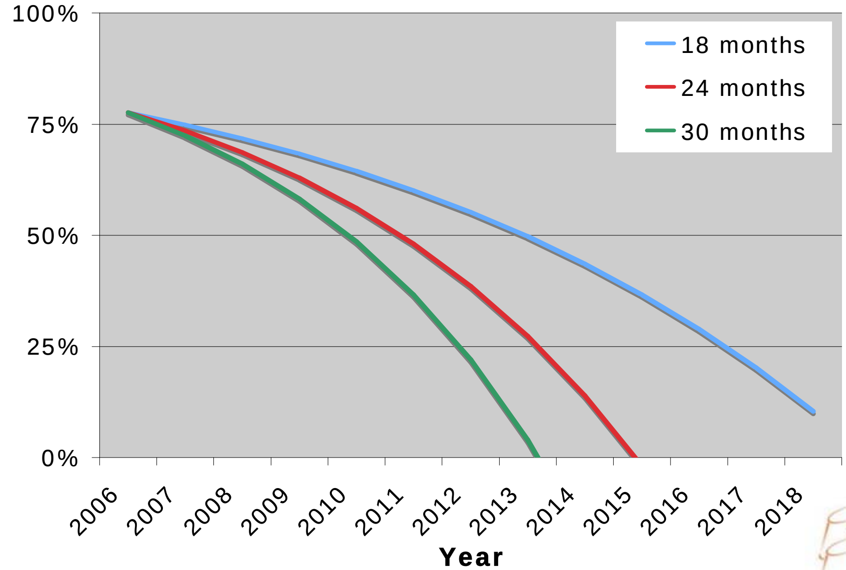
Fault Tolerance Challenges

- Periodic (p) pause to checkpoint (t)
  - Major need for storage bandwidth
- Balanced systems
  - Storage speed tracks FLOPS, memory so checkpoint capture (t) is constant
  - 1 – AppUtilization = t/p + p/(2*MTTI)

$p^2 = 2*t*MTT$







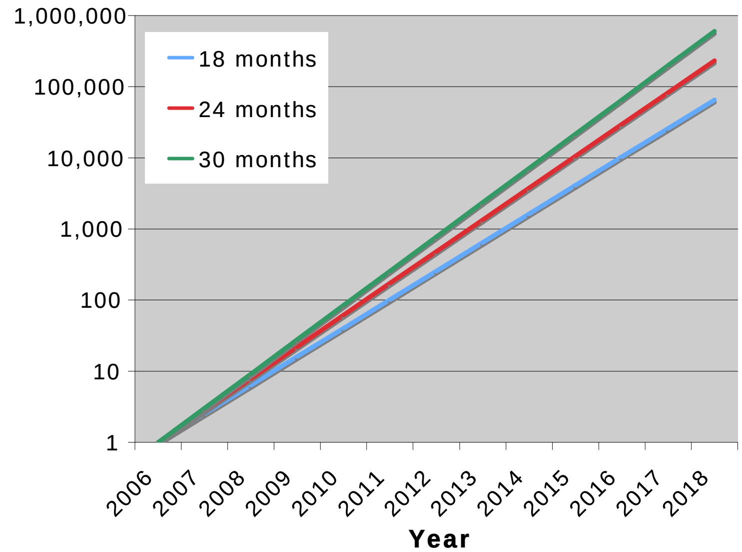
- ***but dropping MTTI kills app utilization!***

Carnegie Mellon

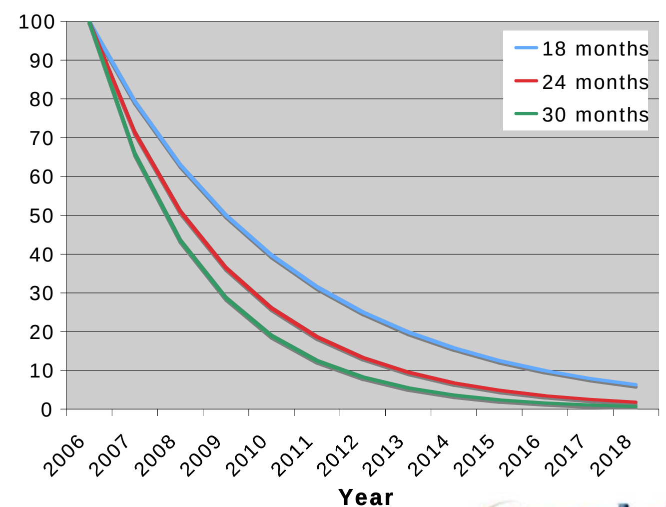
pdsi

Bolster HEC Fault Tolerance

- More storage bandwidth?
  - disk speed 1.2X/yr
    - – # disks +67%/y just for balance !
  - to also counter MTTI
    - – # disks +130%/yr !
  - Little appetite for the cost



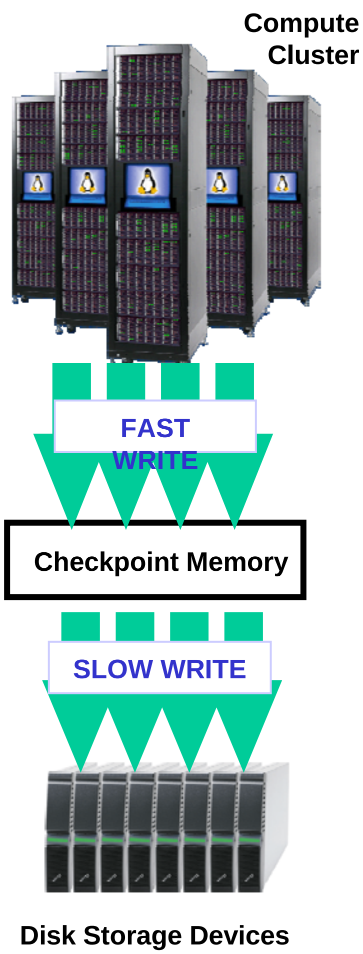
- Compress checkpoints!
  - plenty of cycles available
  - smaller fraction of memory each year (application specific)
    - – 25-50% smaller per year



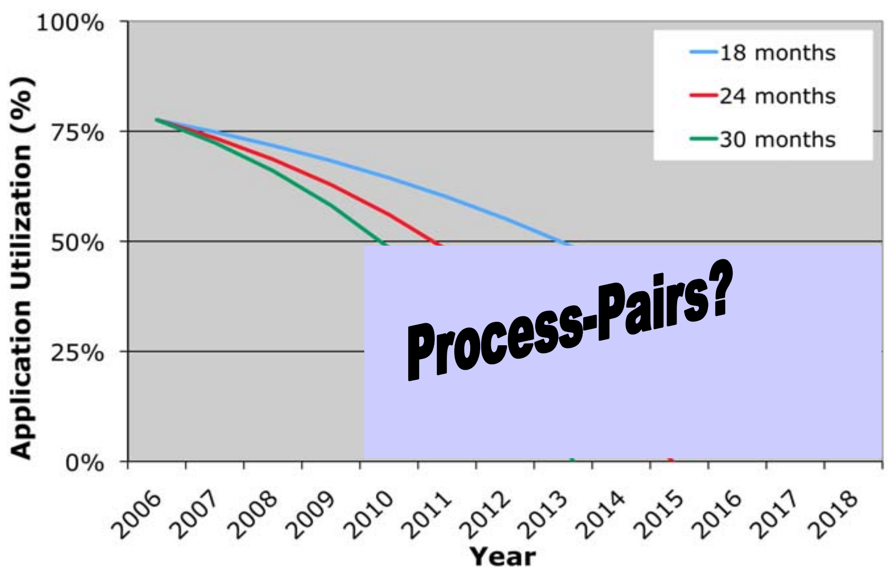
Carnegie Mellon

pdsi

Alternative Approaches



- Dedicated checkpoint device (ie., PSC Zest)
  - Stage checkpoint through fast memory
  - Cost of dedicated memory large fraction of total
  - Cheaper memory (flash?) now bandwidth limited
- Classic enterprise process pairs duplication
  - Flat 50% efficiency cost, plus message duplication



Storage Suffers Failures Too

		Type of drive	Count	Duration
Pittsburgh Supercomputing Center	HPC1	18GB 10K RPM SCSI 36GB 10K RPM SCSI	3,400	5 yrs
Los Alamos National Laboratory EST.1943	HPC2	36GB 10K RPM SCSI	520	2.5 yrs
Supercomputing X	HPC3	15K RPM SCSI 15K RPM SCSI 7.2K RPM SATA	14,208	1 yr
Various HPCs	HPC4	250GB SATA 500GB SATA 400GB SATA	13,634	3 yrs
Internet services Y	COM1	10K RPM SCSI	26,734	1 month
	COM2	15K RPM SCSI	39,039	1.5 yrs
	COM3	10K RPM FC-AL 10K RPM FC-AL 10K RPM FC-AL 10K RPM FC-AL	3,700	1 yr

Storage Failure Recovery is On-the-fly

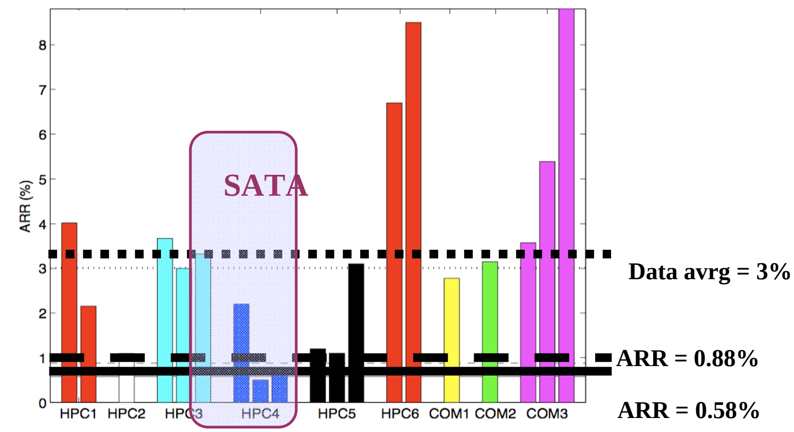
- Scalable performance = more disks
- But disks are getting bigger
- Recovery per failure increasing
- Hours to days on disk arrays
- Consider # concurrent disk recoveries

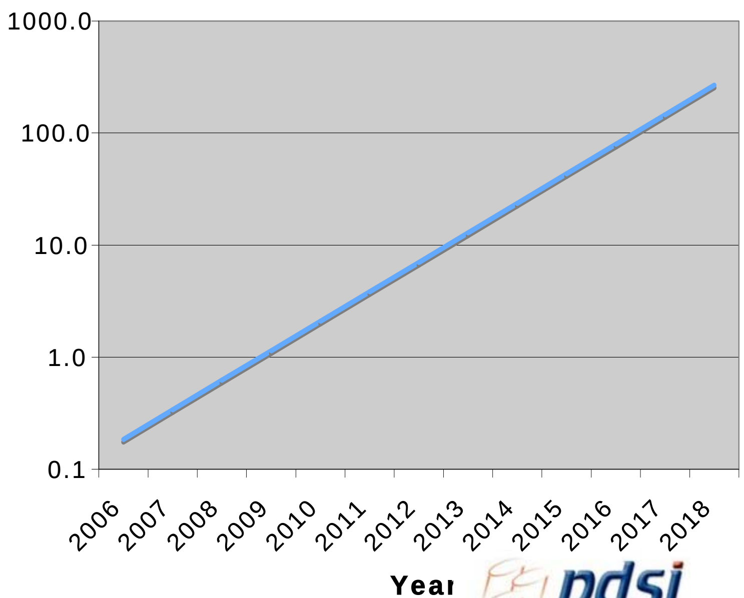
  e.g. 10,000 disks

  3% per year replacement rate

  1+ day recovery each

  Constant state of recovering ?
- Maybe soon 100s of concurrent recoveries (at all times!)
- Design normal case for many failures (huge challenge!)



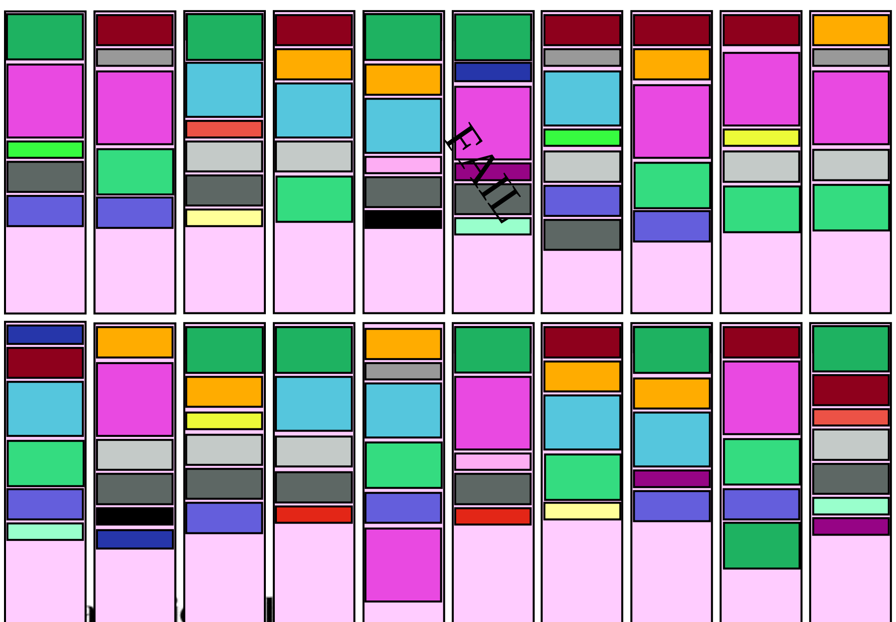


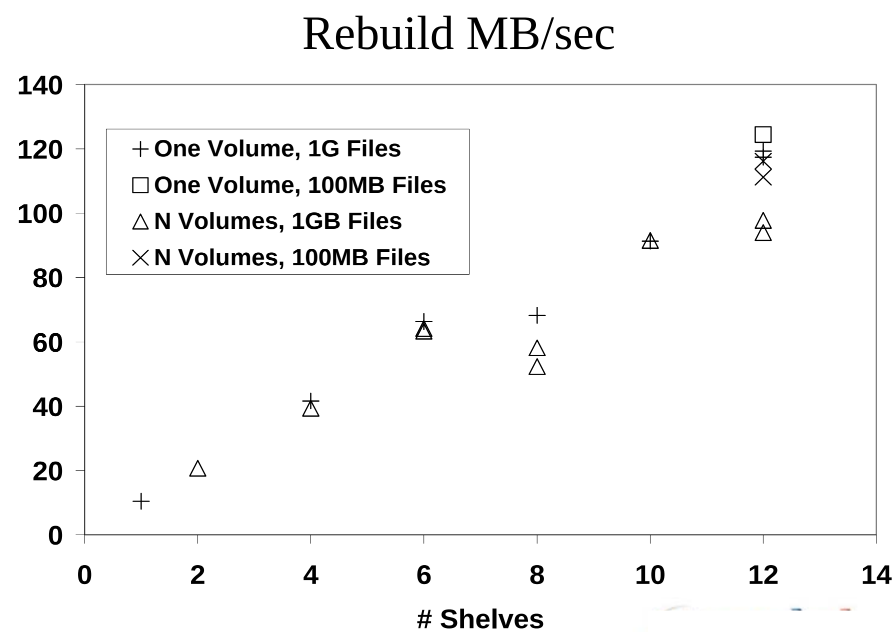
Carnegie Mellon

pdsi

Parallel Scalable Repair

- Defer the problem by making failed disk repair a parallel app
- File replication and, more recently, object RAID can scale repair
  - “decluster” redundancy groups over all disks (mirror or RAID)
  - use all disks for every repair, faster is less vulnerable
- Object (chunk of a file) storage architecture dominating at scale
  PanFS, Lustre, PVFS, … GFS, HDFS, … Centera, …





PDSI Collaborations

- Primary partner: Scientific Data Management (SDM)
  – PVFS, metadata, checkpoint, failure diagnosis
- Storage IO & Performance Engineering (PERI)
  – Ocean (POP), Combustion (S3D), P. Roth
- Storage trouble shooting with trace tools
  – Climate Change (CCSM), L. Ward
- Vendor partnerships & centers of excellence
  – pNFS, IBM GPFS, Sun Lustre, Panasas, EMC
- Leadership computing facilities partnerships
  – Roadrunner, Redstorm, Jaguar, Franklin, ...
- Base program cooperation: FASTOS IO Forwarding

Its All About Data, Scale & Failure

- Continual gathering of data on data storage
  - Failures, distributions, traces, workloads
- Nurturing of file systems to HPC scale, requirements
  - pNFS standards, benchmarks, testing clusters, academic codes
- Checkpoint specializations
  - App-compressed state, special devices & representations
- Failure as the normal case?
  - Risking 100s of concurrent disk rebuilds (need faster rebuild)
  - Quality of service (performance) during rebuild in design
- Correctness at increasing scale?
  - Testing using virtual machines to simulate larger machines
  - Formal verification of correctness (performance?) at scale
- HPC vs Cloud Storage Architecture
  - Common software? Share costs with new technology wave

SciDAC
Scientific Discovery through Advanced Computing

Office of Science

SciDAC Petascale Data Storage Institute

Advanced Scientific Computing Advisory Committee Meeting, October 29 2008, Gaithersburg MD

Garth Gibson

Carnegie Mellon University and Panasas Inc.

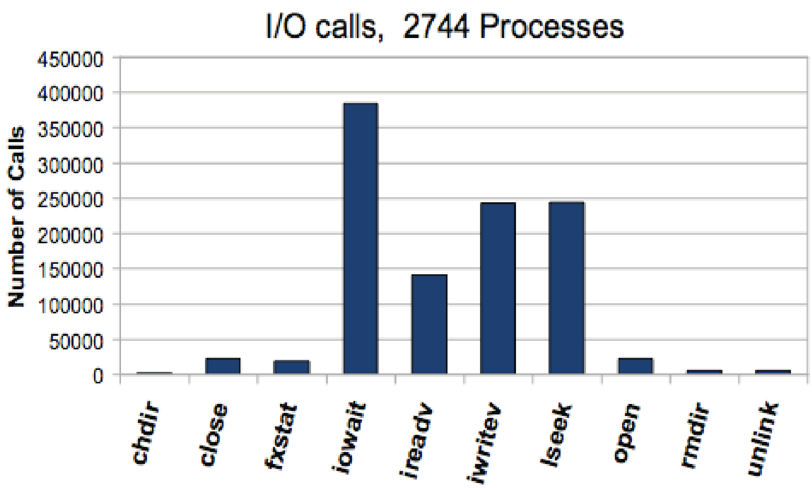
SciDAC Petascale Data Storage Institute (PDSI)

www.pdsi-scidac.org

w/ LANL (Gary Grider), LBNL (William Kramer), SNL (Lee Ward), ORNL (Phil Roth), PNNL (Evan Felix), UCSC (Darrell Long), U.Mich (Peter Honeyman)

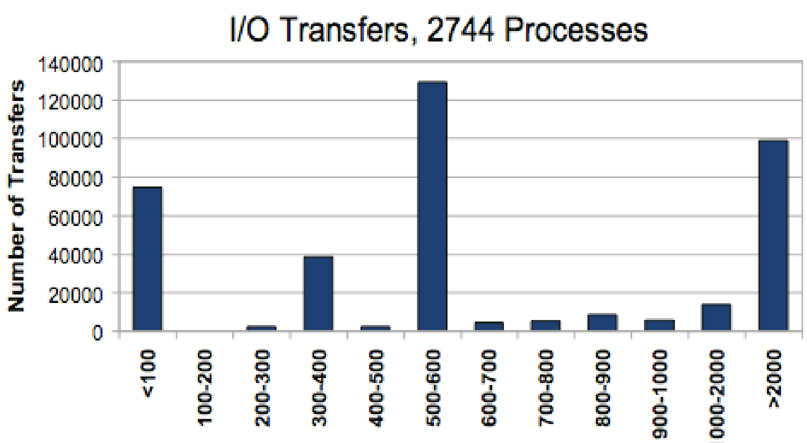
Backup

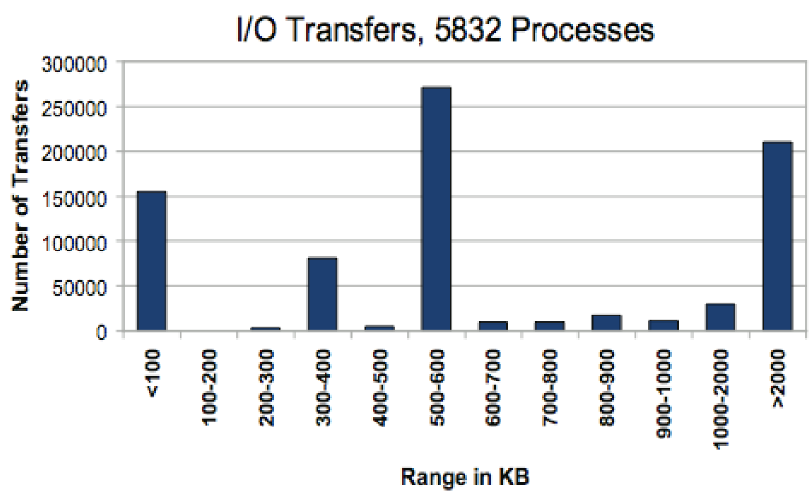
Tools: Understanding IO in Apps



NEWEST TRACE DATA, REDSTORM, SANDIA NAT'L LAB

- A physics simulation problem for a common Sandia application, Alegra
  - 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.



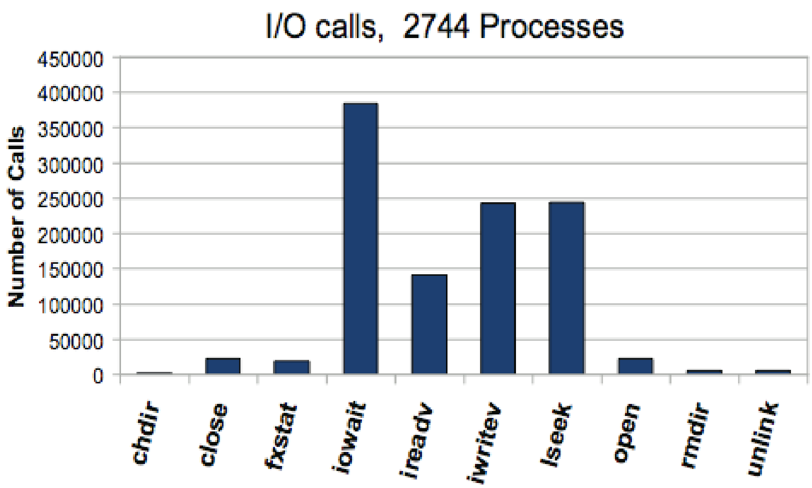


Sandia National Laboratories

sourceforge.net/projects/libsysio

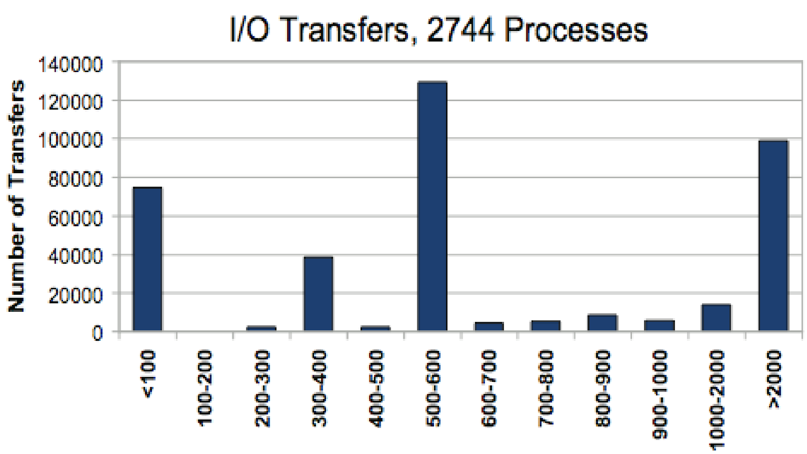
pdsi

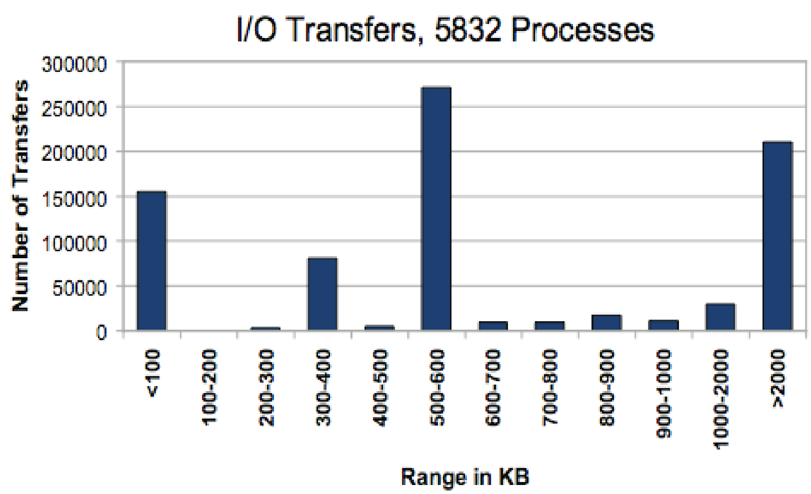
Dissemination: Parallel Workloads



NEWEST TRACE DATA, REDSTORM, SANDIA NAT'L LAB

- A physics simulation problem for a common Sandia application, Alegra
  - 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.





Sandia National Laboratories

sourceforge.net/projects/libsysio

pdsi

Dissemination: Parallel Workloads

MPI-IO Test

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

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

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

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

Dissemination: 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) collaboratoration 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 I
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 cal
runs are made with this new feature on Jacquard (courtesy o

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 arer
profiling in IPM.

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

Performance Effect of Multi-core on Scientific Applicationa (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.

Effcient Parallel I/O on thee Cray XT3/XT4 PDF
Provides an overview of I/O methods for three different applications

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 where run on Jacquard . The files are compressed tar files of the trace data

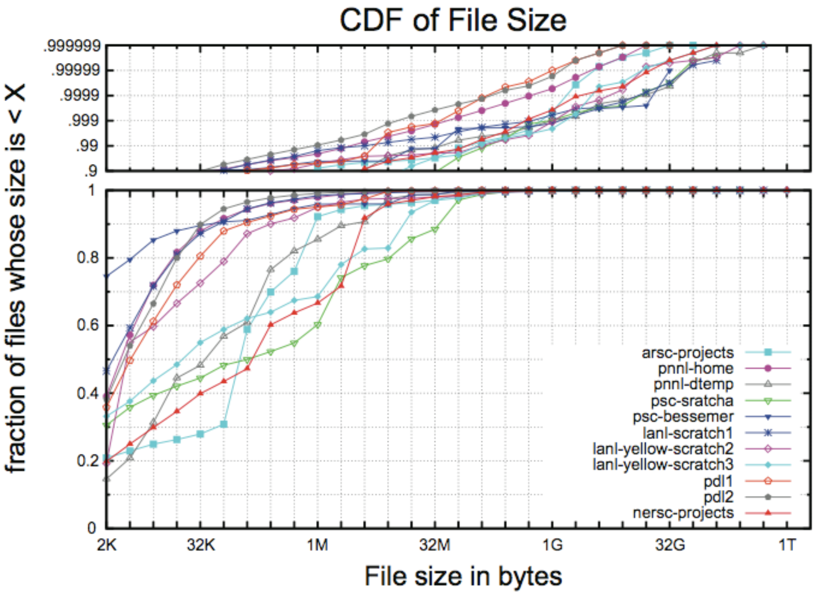
PMEMD 16 tasks small dataset run

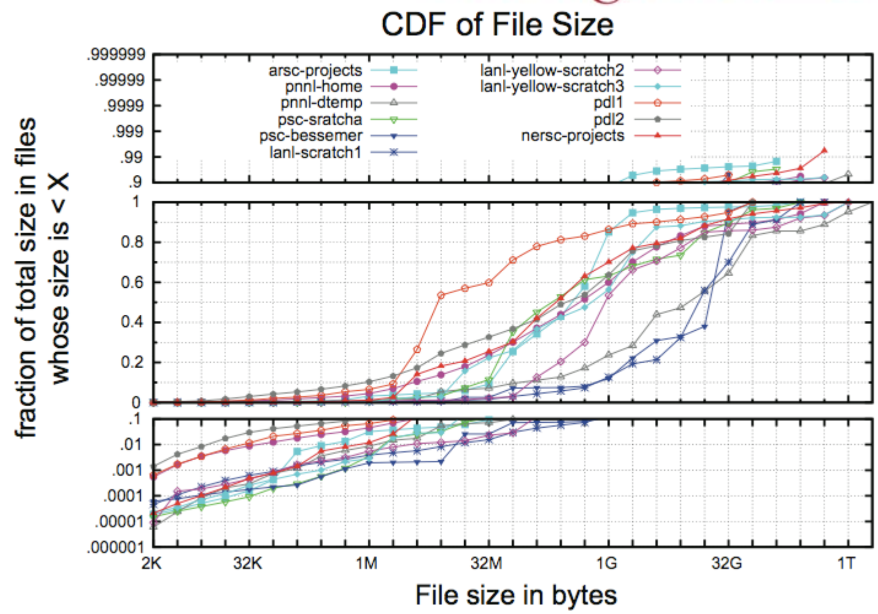
MADbench 64 tasks medium dataset run

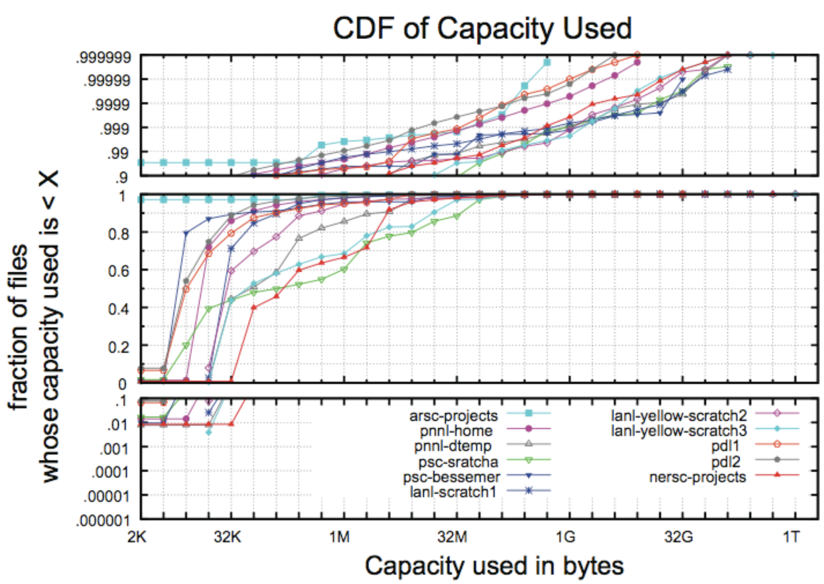
MILC 16 tasks medium dataset run

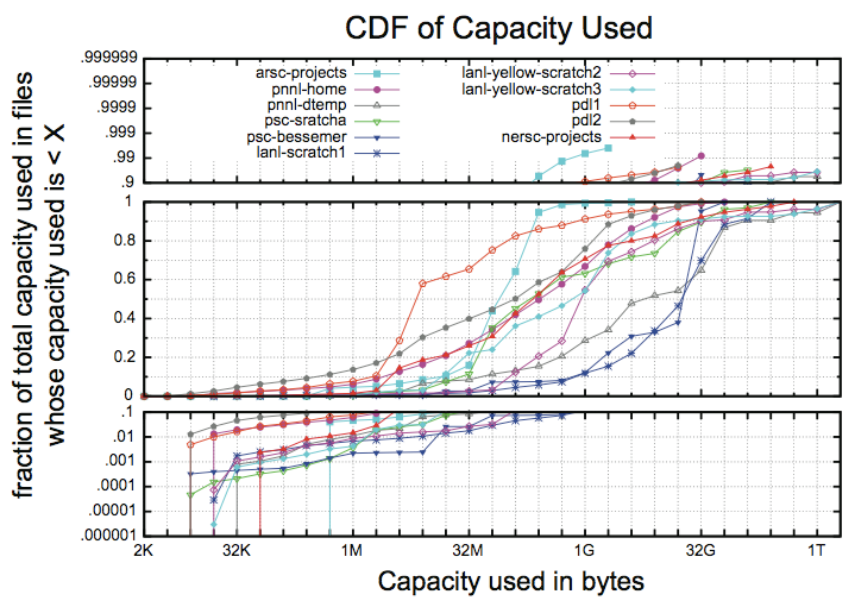
pdsi @ NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Dissemination: Statistics





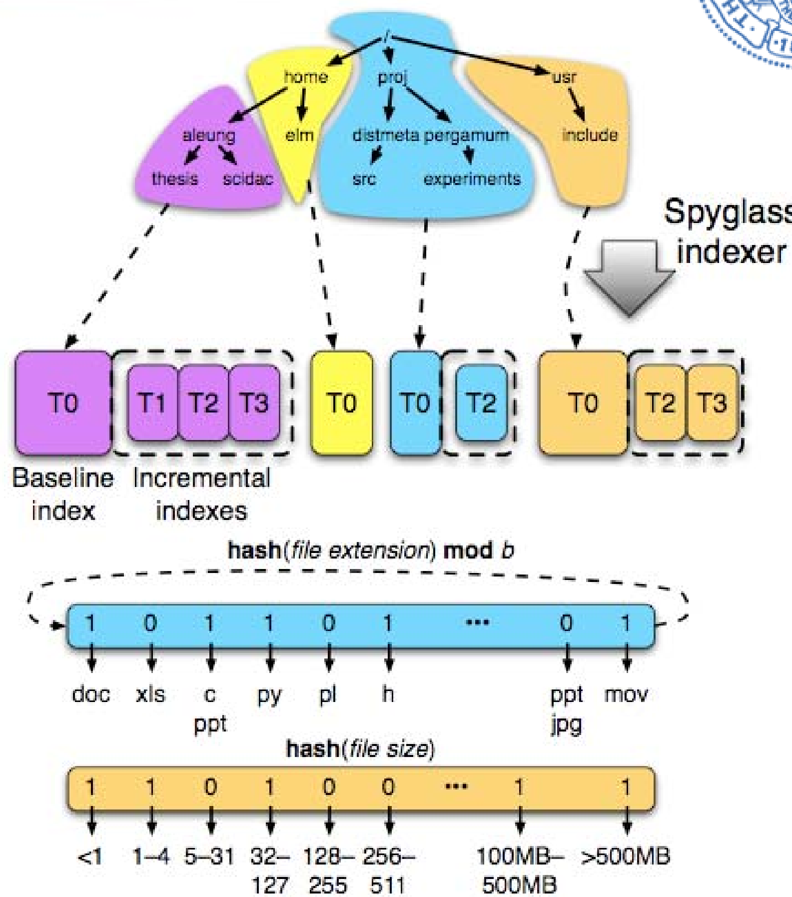




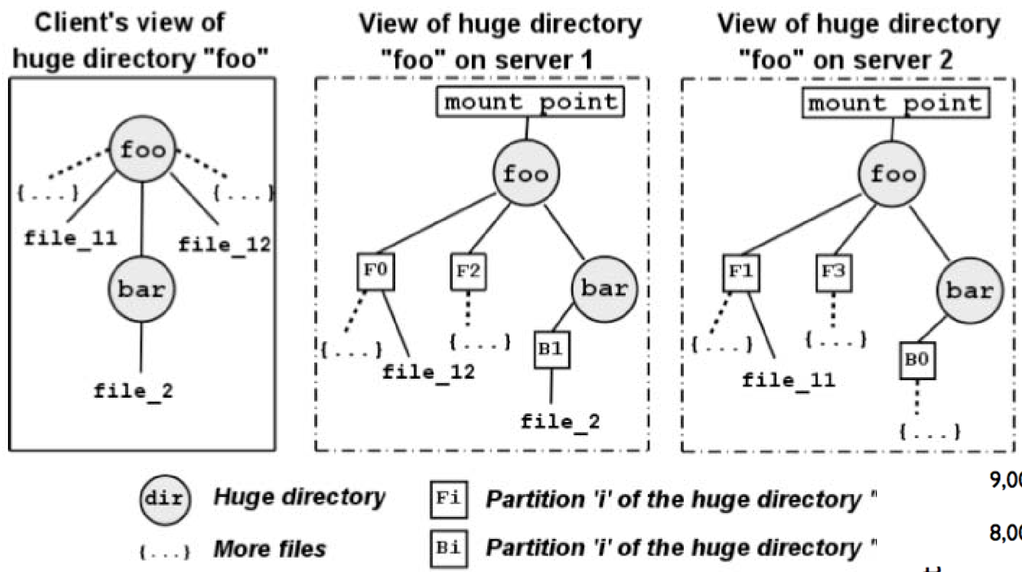
Mechanisms for Scalable Metadata (1)

Spyglass design

- Partition file system hierarchy by subtree
  - Each subtree is an independent subindex
- Summarize contents of each subindex
  - Quickly rule out entire subindexes that can't satisfy the query
- Log incremental changes
  - Rebuild index when there are "enough" changes
- Integrity is much easier
  - Rebuild subindex, not entire index



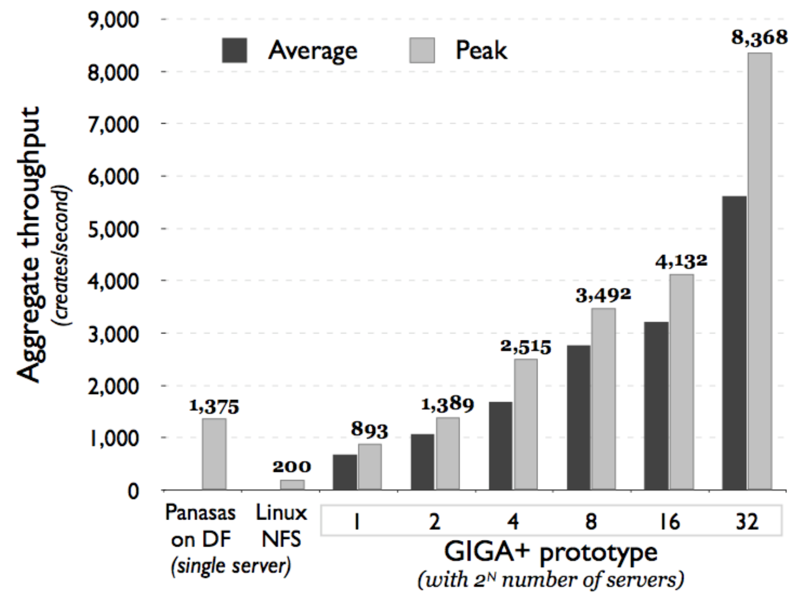
Mechanisms for Scalable Metadata (2)



Local representation of huge directory in Giga+

- Billion+ files in a directory
- Eliminate serialization
  - All servers grow directory independently, in parallel, without any co-ordinator

- No synchronization or consistency bottlenecks
  - Servers only keep local "view", no shared state



Scale and performance of Giga+ using UCAR Metarates benchmark.

Carnegie Mellon