SciDAC
Scientific Discovery through Advanced Computing

ASCR

Office of Science

SciDAC PDSI Update (part 1)

CS/VIS PI Meeting, October 23, Germantown, MD

Garth Gibson
Carnegie Mellon University and Panasas Inc.

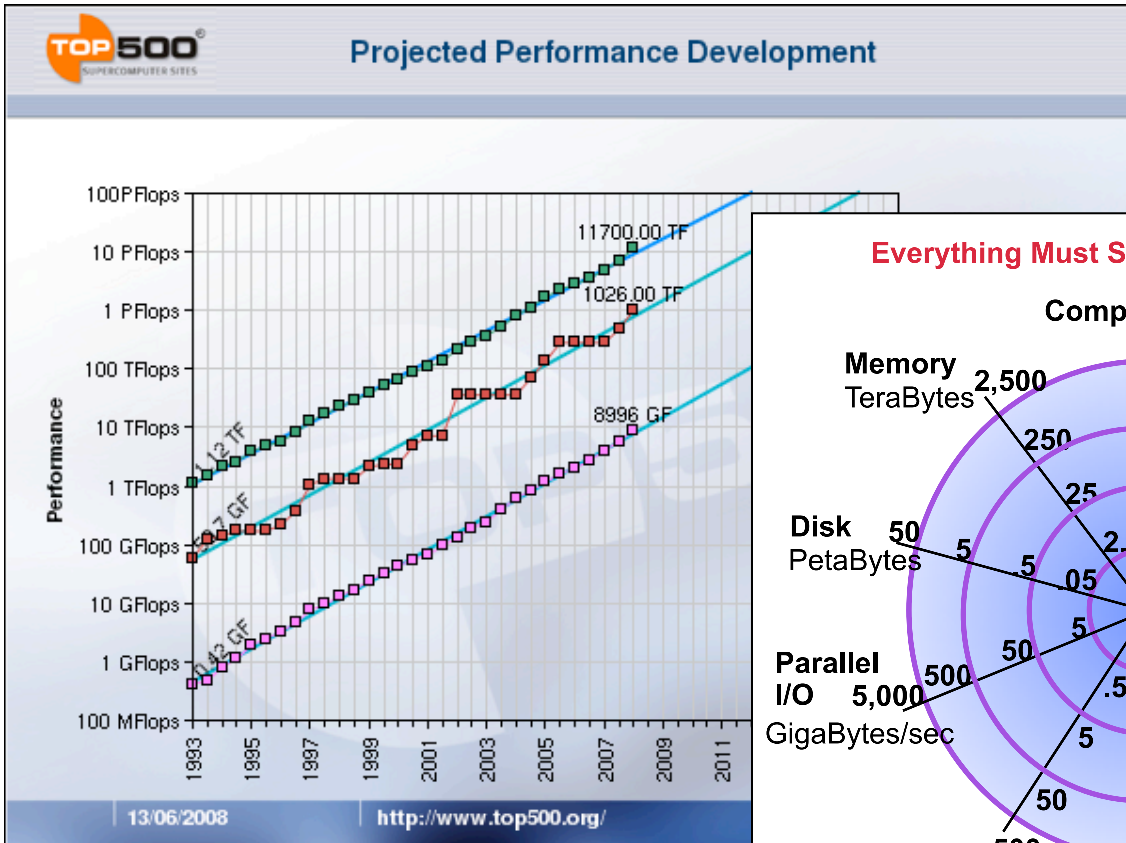
SciDAC Petascale Data Storage Institute (PDSI)

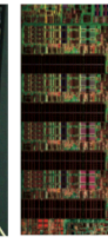
www.pdsi-scidac.org

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

Clearing Path thru Petascale to Exascale

- Scaling 100%/yr given disk realities is hard
  - Disk BW @ 20%/yr, IO/s @ 5%/yr
  - Storage problem renews itself each year

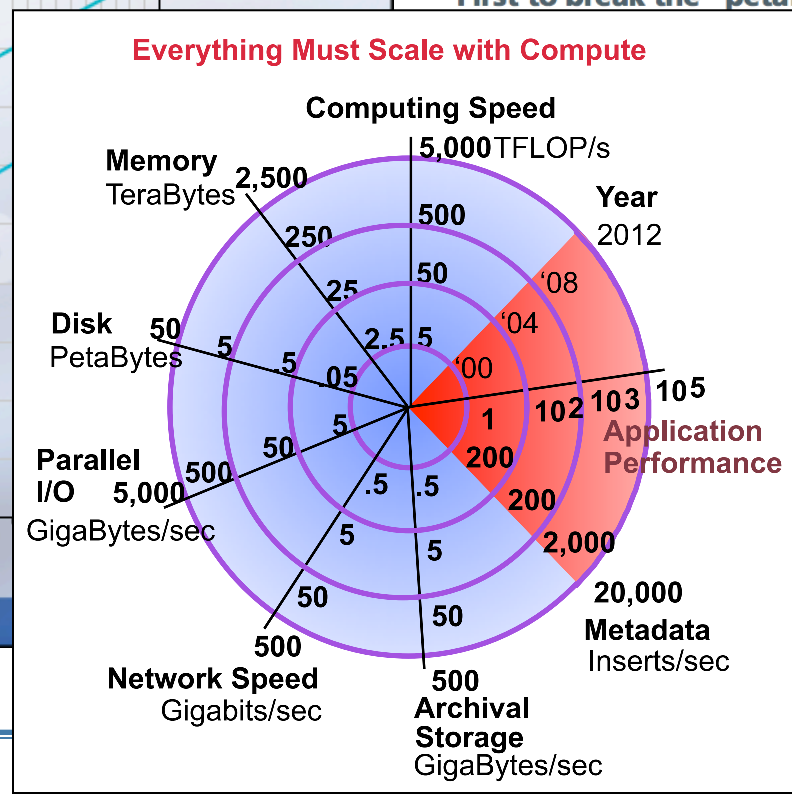




Roadrunner

First to break the "petaflop" barrier

emorial Day, the
eeded a sustained speed
n calculations per second.
Roadrunner more than
r 1 system on the TOP500
nce to date is 74.5% ef-



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

Annual PDSI Sponsored Workshops

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

open mic avail

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 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, Understanding, 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

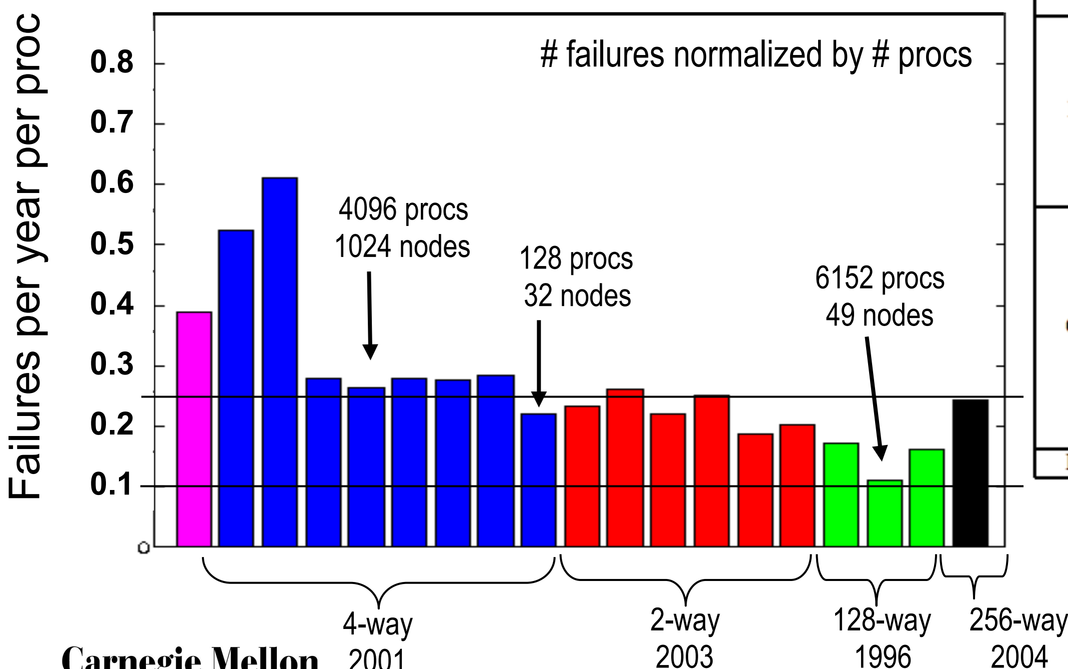
Special Presentation on HPC User Requirements:
I/O Requirements for HPC Applications: A User Perspective
John Shalf, National Energy Research Scientific Computing Center (NERSC), LBNL

PDSI POSTER AT THE FAST '08 POSTER SESSION

PDSI Data Releases and Repositories

Failure Data Collection

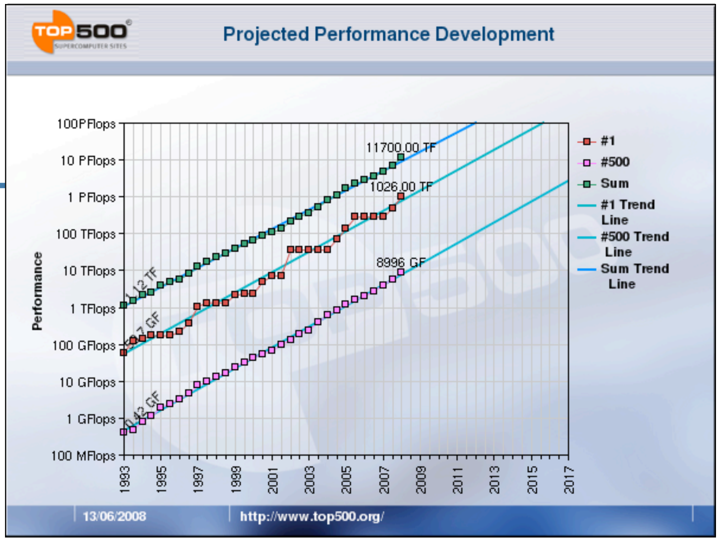
- Los Alamos root cause logs
  - 22 clusters & 5,000 nodes
  - covers 9 years & continues
  - cfdr.usenix.org publishes this and many other failure datasets



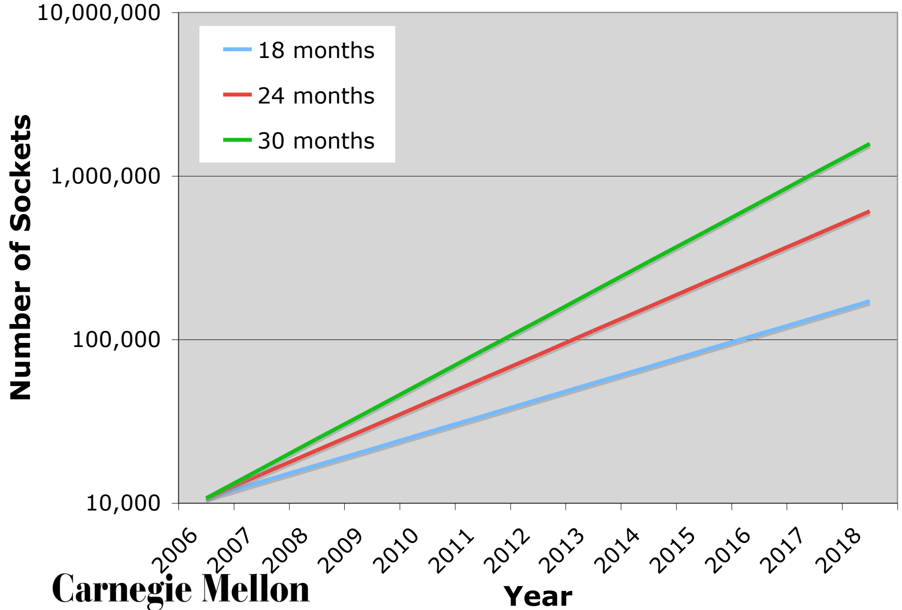
(I) High-level system information				(II) Information per node category			
HW	ID	Nodes	Procs	Procs /node	Production Time	Mem (GB)	NICs
A	1	1	8	8	N/A – 12/99	16	0
B	2	1	32	32	N/A – 12/03	8	1
C	3	1	4	4	N/A – 04/03	1	0
D	4	164	328	2	04/01 – now	1	1
				2	12/02 – now	1	1
E	5	256	1024	4	12/01 – now	16	2
	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

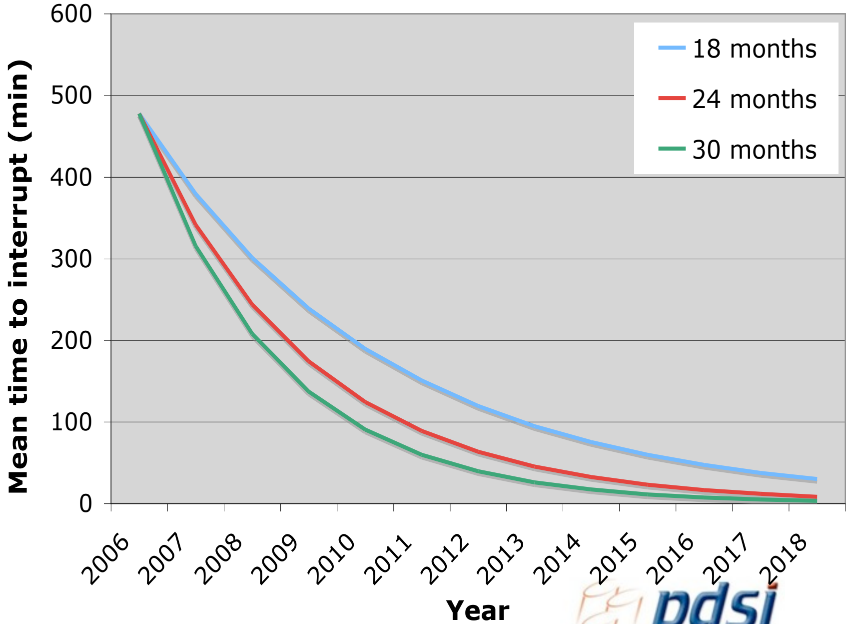
Table 1. *Overview oj SMP-based, and systen*

Projections: More Failures



- Con't top500.org 2X annually
  - 1 PF Roadrunner in 2008 (May 26)
- 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)

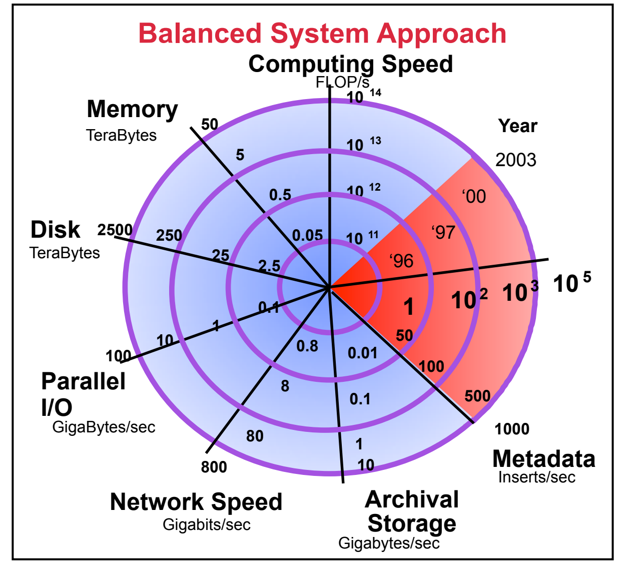


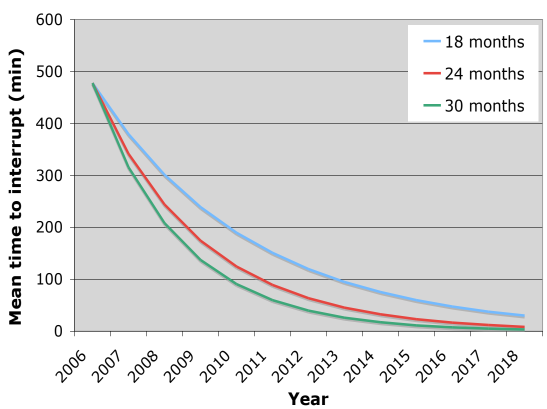


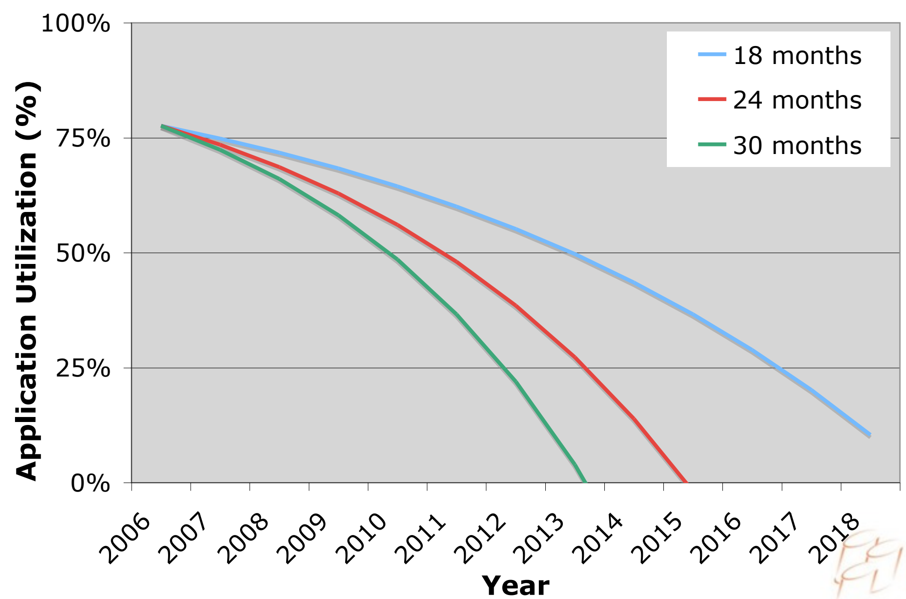
Checkpointing Failure Tolerance in Trouble

- Periodic (p) pause to checkpoint (t)
  - On failure, roll back & restart
- Balanced systems
  - disk speed tracks FLOPS & mem size, so checkpoint capture (t) is constant time
  - 1 – AppUtilization = t/p + p/(2*MTTI)

$$p^2 = 2*t*MTTI$$



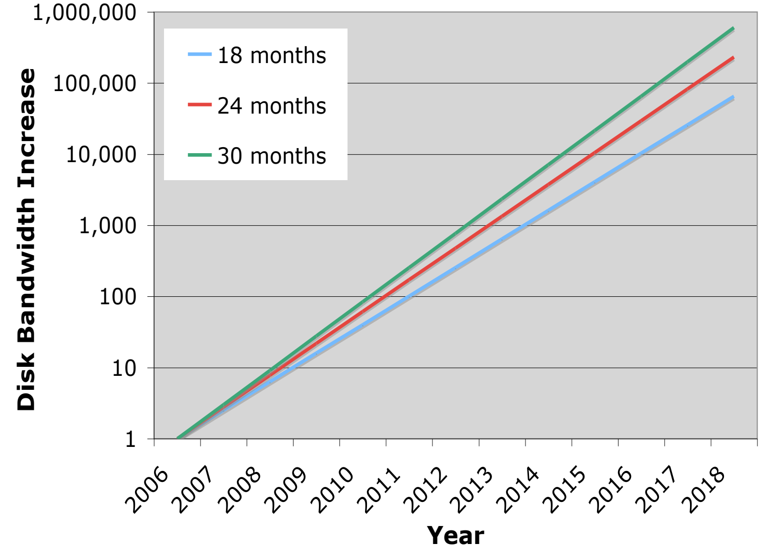




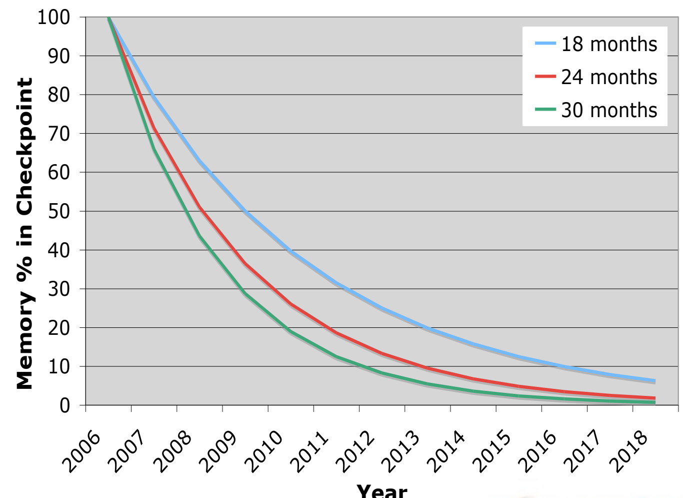
- ***but dropping MTTI kills app utilization!***

Bolster HEC Fault Tolerance

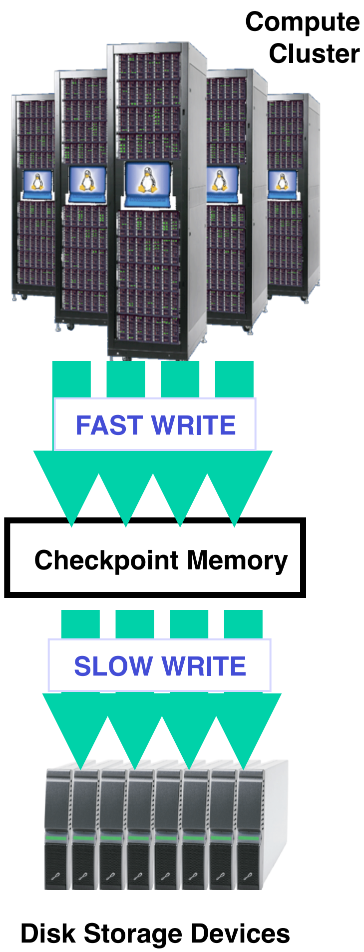
- More storage bandwidth
  - disk speed 1.2X/yr
    - – # disks +67%/y just for balance !
  - to also counter MTTI
    - – # disks +130%/yr !
  - poor appetite for the cost



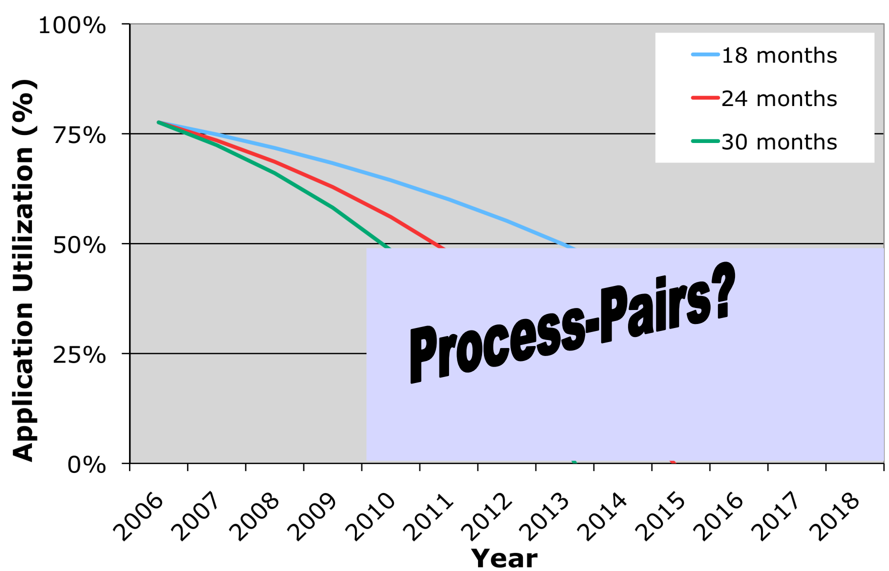
- Compress checkpoints
  - plenty of cycles available
  - smaller fraction of memory each year
    - – 25-50% smaller / yr



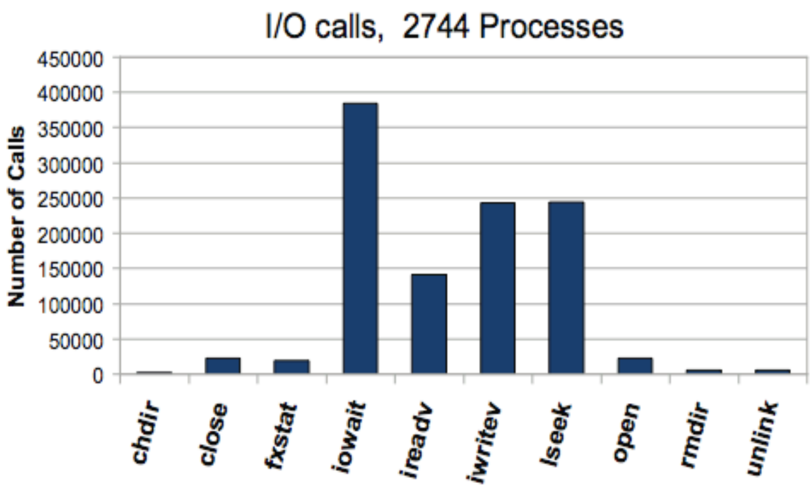
Different Approaches



- 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

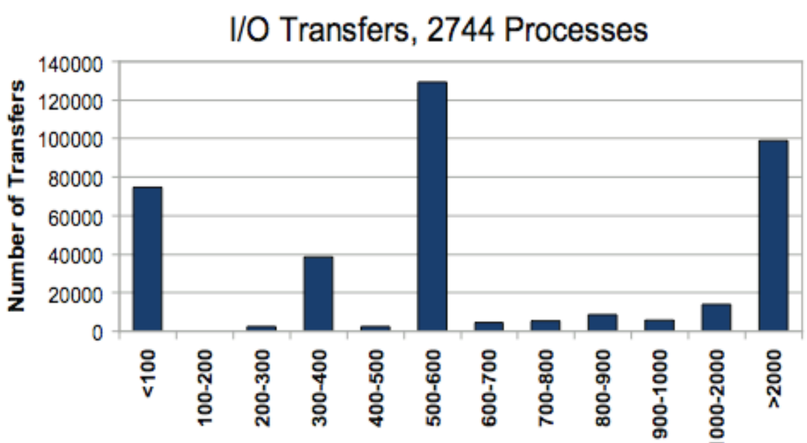


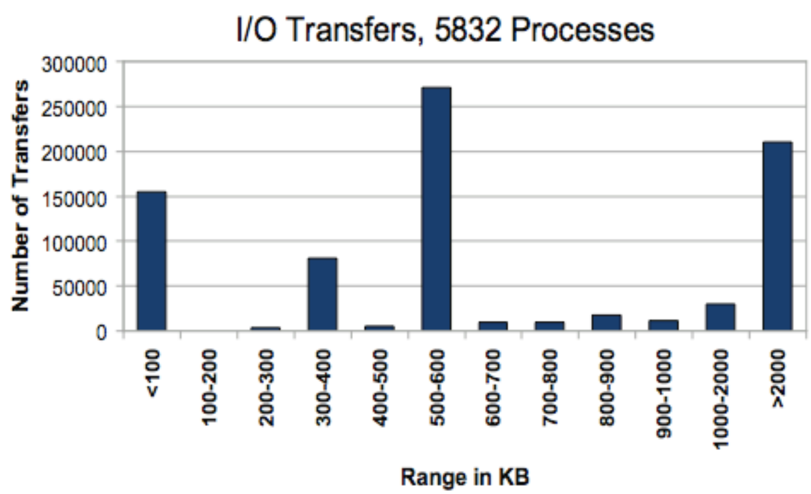
Tools for Understanding IO in Apps



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.



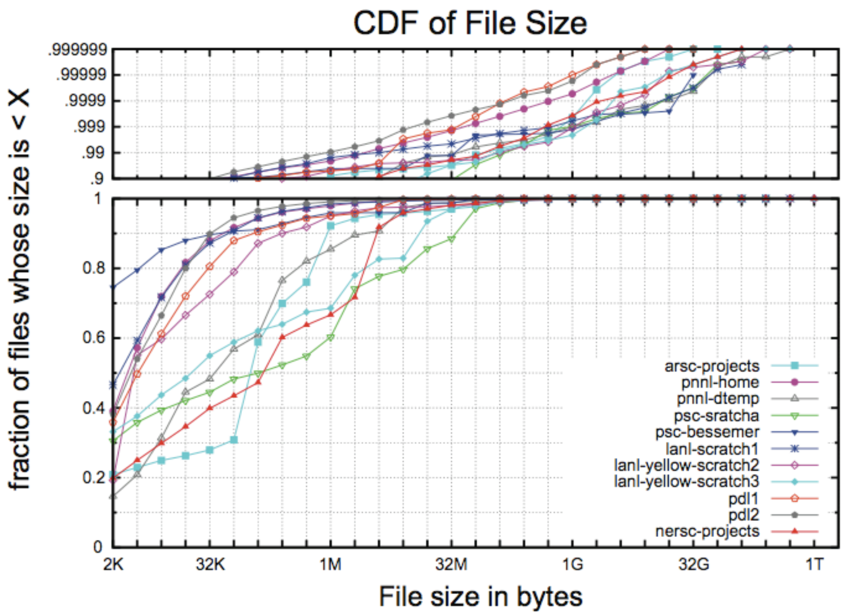


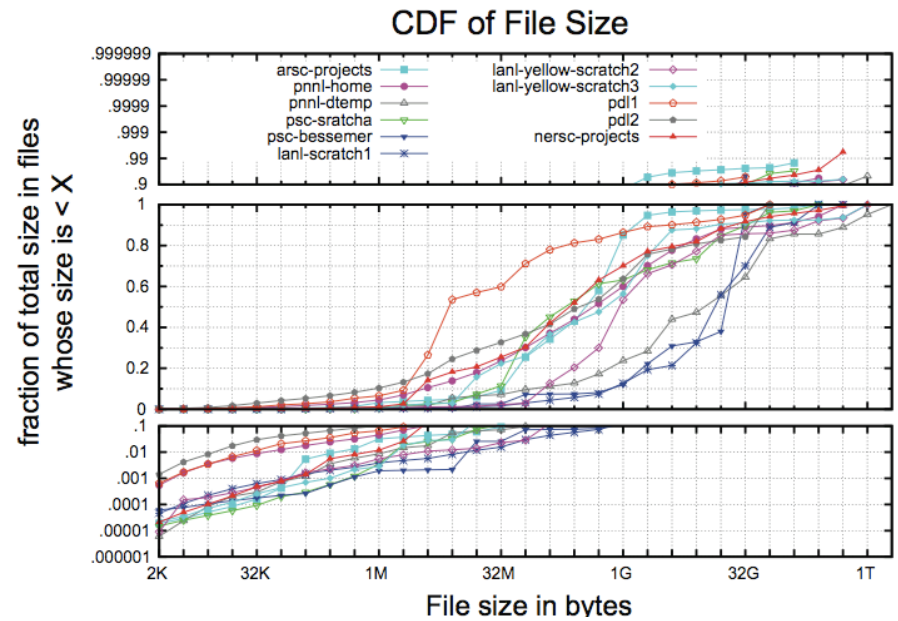
Sandia National Laboratories

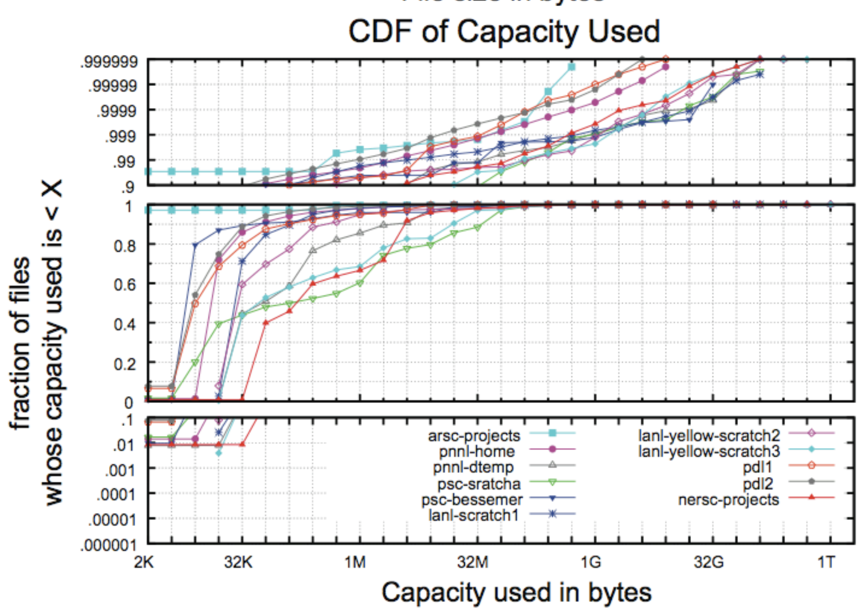
sourceforge.net/projects/libsysio

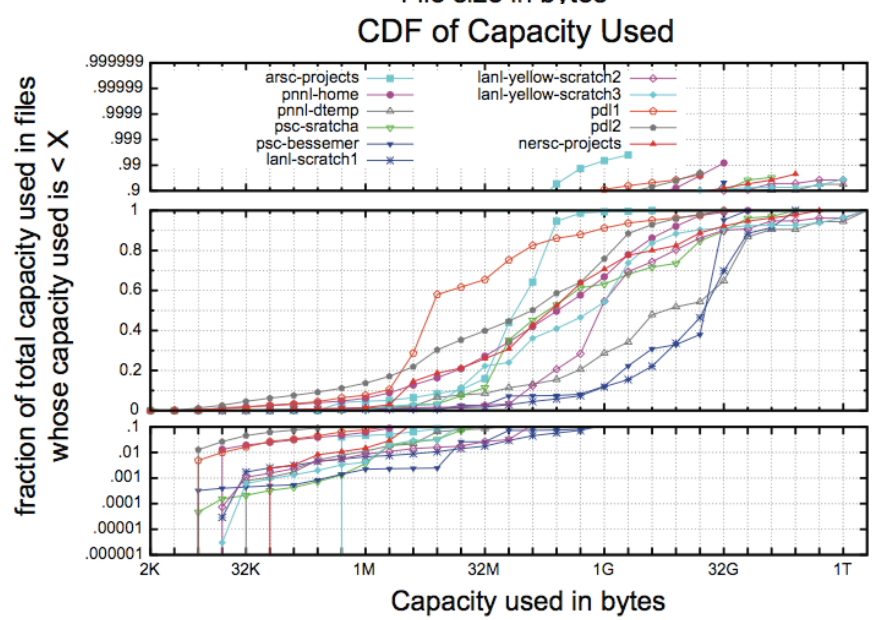
pdsi

Datasets: File Statistics









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

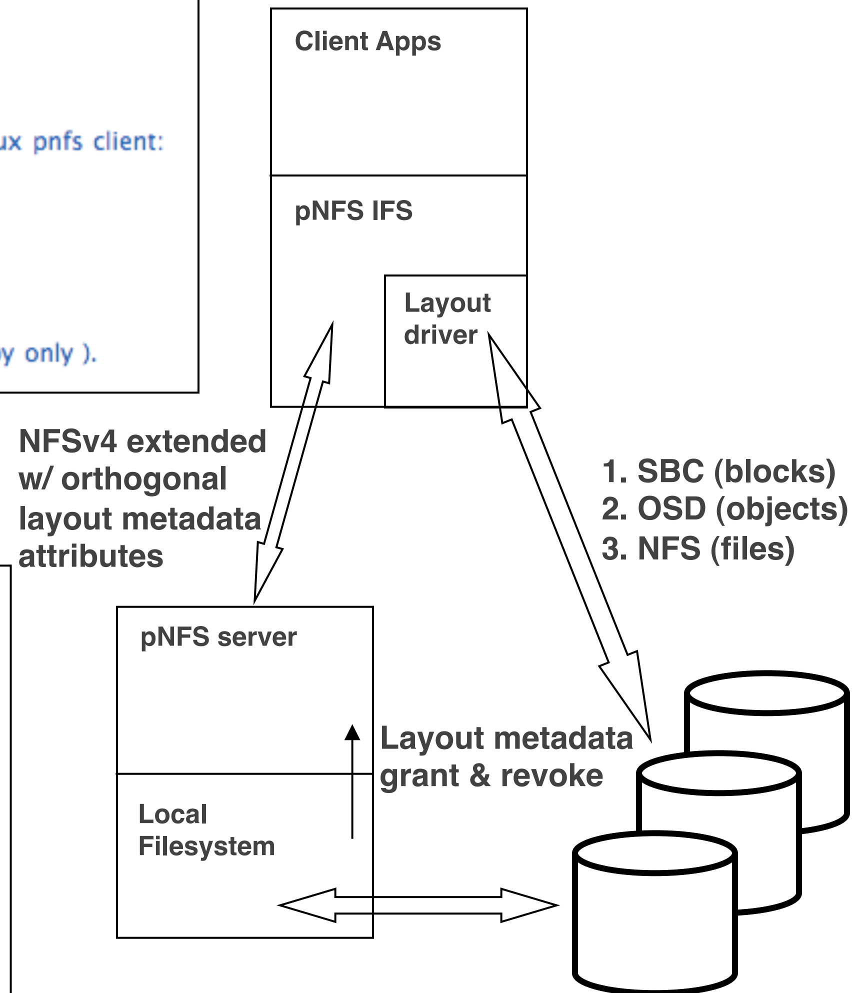
Anyway, fist time we can show that NFSv4.1 is something real (and not my hobby only).

- Open source & competitive offerings!
- NetApp, Sun, IBM, EMC, Panasas ….

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

2. IETF status

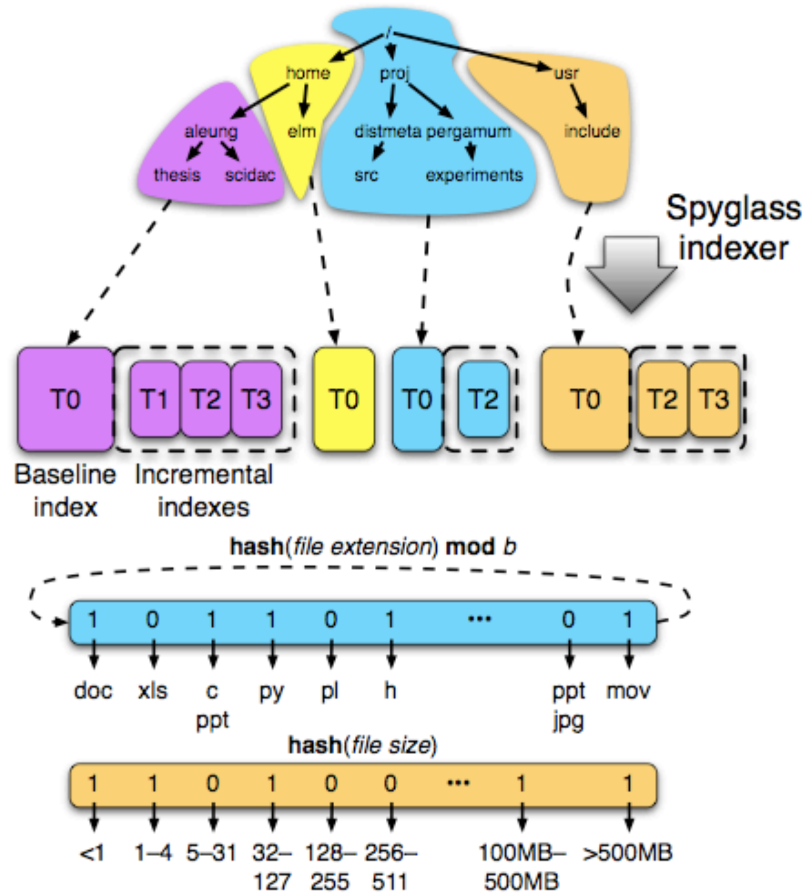
All of the current working group internet drafts are moving forward for publication. This means that they have submitted to the area director and will start their way through the process (IETF last call and IESG review).



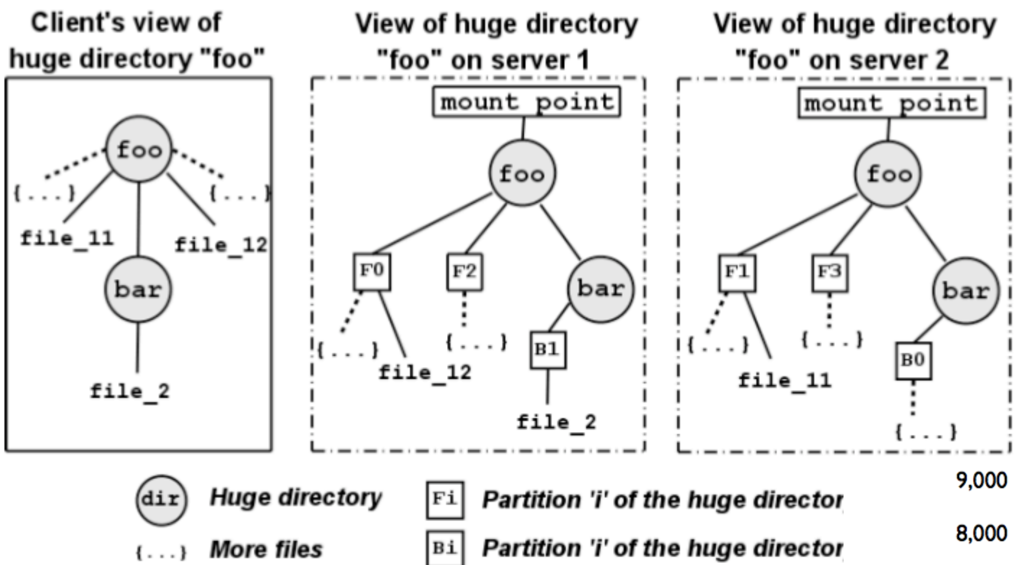
Metadata: Subindexing by Names

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



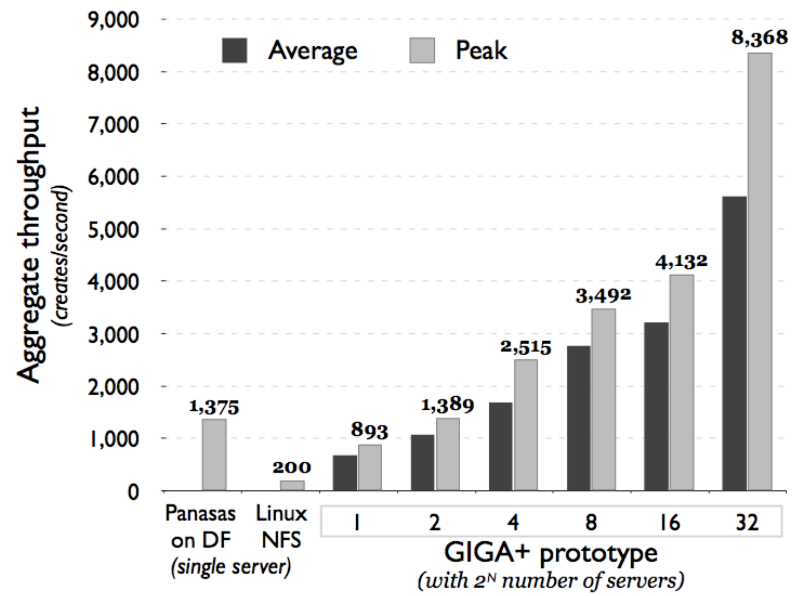
Metadata: Huge Directories (PVFS, FUSE)



Local representation of huge directory in Giga

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

- Eliminate serialization
  - All servers grow directory independently, in parallel, without any co-ordinator



Scale and performance of Giga+ using UCAR Metarates benchmark.

SciDAC PDSI Update (part 1)

CS/VIS PI Meeting, October 23, Germantown, MD

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)