- PETASCALE DATA STORAGE WORKSHOP
  - 8:30 am - 5 pm Sunday Nov 11, Atlantis Ballroom E
  - www.pdsi-scidac.org/SC07

Petascale computing infrastructures make petascale demands on information storage capacity, performance, concurrency, reliability, availability, and manageability. The last decade has shown that parallel file systems can barely keep pace with high performance computing along these dimensions; this poses a critical challenge when near-future petascale requirements are considered. This recurring one-day workshop focuses on the data storage problems and emerging solutions found in petascale scientific computing environments, with special attention to issues in which community collaboration can be crucial, problem identification, workload capture, solution interoperability, standards with community buy-in, and shared tools.

- Principle Petascale Storage issue is Scale
  - Up to Terabytes/sec bandwidth
  - Widely concurrent write sharing; non-aligned small strided
  - Trillions of files needing to do “ls -l”, “du -s”, backup
  - Billions of files in a directory
  - Millions of files creates and written per minute
  - Increasing need for brute force search
  - An order of magnitude or two more disks
  - Many more frequent failures, multiple failures
  - Operational staff costs not increasing
  - Weak programming for storage skills

- PETASCALE DATA STORAGE INSTITUTE
  - Sponsor and Program Committee:
  - Garth Gibson, Carnegie Mellon University & Panasas
  - Darrell Long, University of California, Santa Cruz
  - Peter Honeyman, University of Michigan, Ann Arbor, Center for Information Technology Integration
  - Gary Grider, Los Alamos National Lab
  - William Kramer, National Energy Research Scientific Computing Center, Lawrence Berkeley National Lab
  - Philip Roth, Oak Ridge National Lab
  - Evan Felix, Pacific Northwest National Lab
  - Lee Ward, Sandia National Lab

- PETASCALE DATA STORAGE WORKSHOP
  - Competitive extended abstract/paper selection (ACM DL will publish)
  - www.pdsi-scidac.org/SC07 for papers, presentations, posters as provided
- 22 submissions, 12 selected:
  - On Application-level Approaches to Avoiding TCP Throughput Collapse ….
  - pNFS/PVFS2 over Infiniband: Early Experiences
  - Integrated Systems Models for Reliability Petascale Storage Systems
  - Scalable Locking and Recovery for Network File Systems
  - Searching and Navigating Petabyte Scale File Systems Based on Facets
  - Scalable Directories for Shared File Systems
  - End-to-end performance management for scalable distributed storage
  - A Fast, Scalable, and Reliable Storage Service for Petabyte-scale ….
  - A Result-Data Offloading Service for HPC Centers
  - Characterizing the I/O Behavior of Scientific Applications on the Cray XT
  - A Universal Taxonomy for Categorizing Trace Frameworks
  - A Data Placement Service for Petascale Applications

- PETASCALE DATA STORAGE WORKSHOP AGENDA
  - 8:30-9:00: Introduction by Garth Gibson
  - 9:00-10:20: Paper Session 1
    - E. Krevat, Ranjit Noronha, Brent Welch, Peter Braam
  - 10:30-11:00: Poster Session 1
    - Ethan Miller or Garth Gibson have easels/boards/clips for posters
  - 11:00-12:20: Paper Session 2
    - Jonathan Koren, Swapnil Patil, Richard Golding, Sage Weil
  - 12:30-2:00: Lunch (on your own)
  - 2:00-3:20: Paper Session 3
    - Henry Monti, Phil Roth, Andrew Konwinski, Ann Chervenak
  - 3:30-3:00: Short Annoucements
    - Sign up with Garth; Announce availability of data, code, working groups etc
  - 4:00-5:00: Poster Session 2
  - 5:00: Closing by Garth Gibson

- PETASCALE DATA STORAGE INSTITUTE
  - 3 universities, 5 labs, G. Gibson, CMU, PI
- SciDAC @ Petascale storage issues
  - www.pdsi-scidac.org
  - Community building: ie. PDSW-SC07
  - APIs & standards: ie., Parallel NFS, POSIX
  - Failure data collection, analysis: ie., CFDR
  - Performance trace collection & benchmark publication
  - IT automation applied to HEC systems & problems
  - Novel mechanisms for core (esp. metadata, wide area)

The University of California · 1868 · Let There Be Light

Parallel Data Laboratory · Carnegie Mellon University

Carnegie Mellon

center for information technology integration
University of Michigan

Los Alamos National Laboratory EST. 1943

Berkeley Lab

Sandia National Laboratories

Oak Ridge National Laboratory

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

pdsi

SciDAC
Scientific Discovery through Advanced Computing

- PDSI Primary Early Emphasis:
  - Data Collection
    - Failure (next: Workload static/dyn)
    - Gather widely (LANL, NERSC, PNNL, ….)
    - Publish widely (CFDR w/ USENIX)

The computer failure data repository (CFDR)

With the growing scale of todays IT installations, component failure is becoming an ever larger problem. Yet, virtually no data on failures in real systems is publicly available, forcing researchers working on system reliability to base their work on anecdotes and back of the envelope calculations, rather than empirical data.

The computer failure data repository (CFDR) aims at accelerating research on system reliability by filling the nearly empty collection of public data with detailed failure data from a variety of large production systems.

Please join us, either by contributing data, downloading data, or joining our mailing lists.

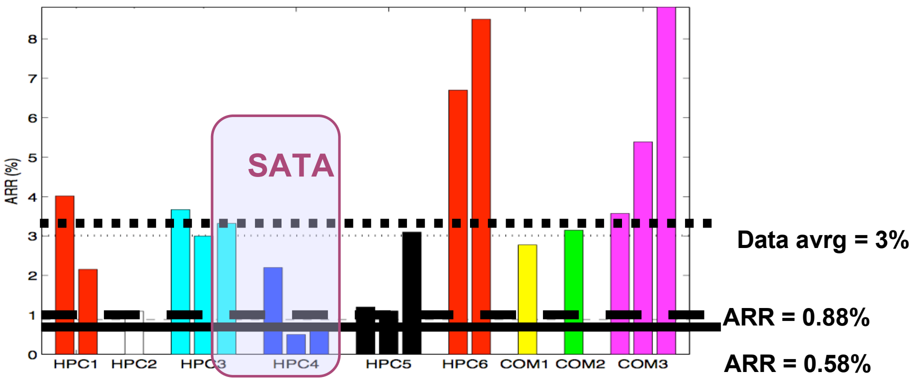
News

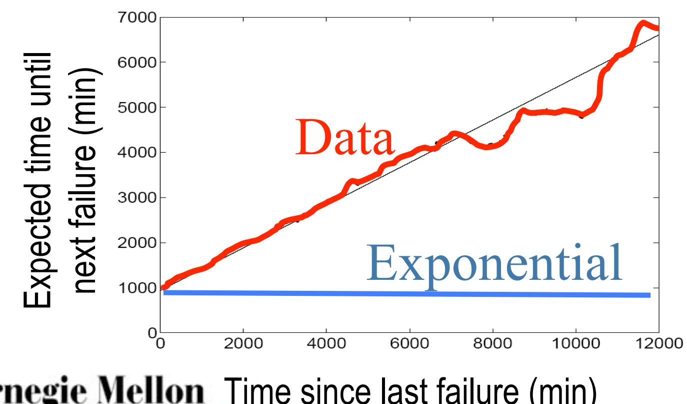
You are viewing a first draft of the CFDR. For feedback and comments please contact the moderators.

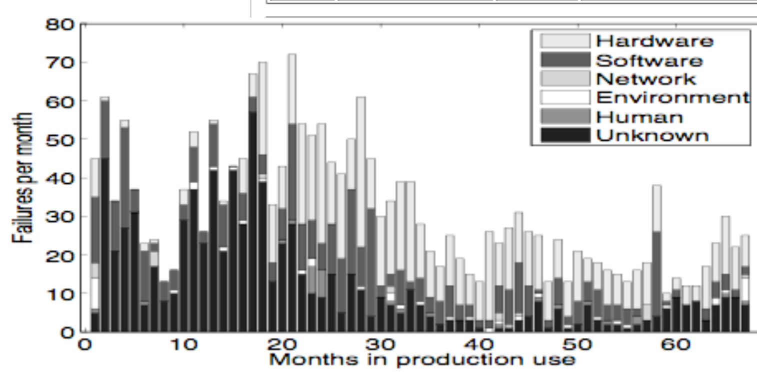
Available data

The table below provides an overview over the available data sets.

Name	Time period	System type	Type of data
LANL	Dec 96 - Nov 05	HPC clusters	The data covers node outages at 22 cluster systems at LANL, including a total of 4,750 nodes and 24,101 processors. Some job logs and error logs are available as well.
HPC1	Aug 01 - May 06	HPC cluster	The data covers hardware replacements at a 765 node cluster with more than 3,000 hard drives.
HPC2	Jan 04 - Jul 06	HPC cluster	Hard drive replacements in a 256 node cluster with 520 drives.
HPC3	Dec 05 - Nov 06	HPC cluster	Hard drive replacements observed in a 1,532-node HPC cluster with more than 14,000 drives.
PNNL	Nov 03 - Sep 07	HPC cluster	Hardware failures recorded on the MPP2 system (a 980 node HPC cluster) at PNNL.
COM1	May 2006	Internet services cluster	Hardware failures recorded by an internet service provider and drawing from multiple distributed sites.
COM2	Sep 04 - Apr 06	Internet services cluster	Warranty service log of hardware failures aggregating events in multiple distributed sites.
COM3	Jan 05 - Dec 05	Internet services cluster	Aggregate quarterly statistics of disk failures at a large external storage system.







Carnegie Mellon
Parallel Data Laboratory

pdsi

Static File System Statistics: New

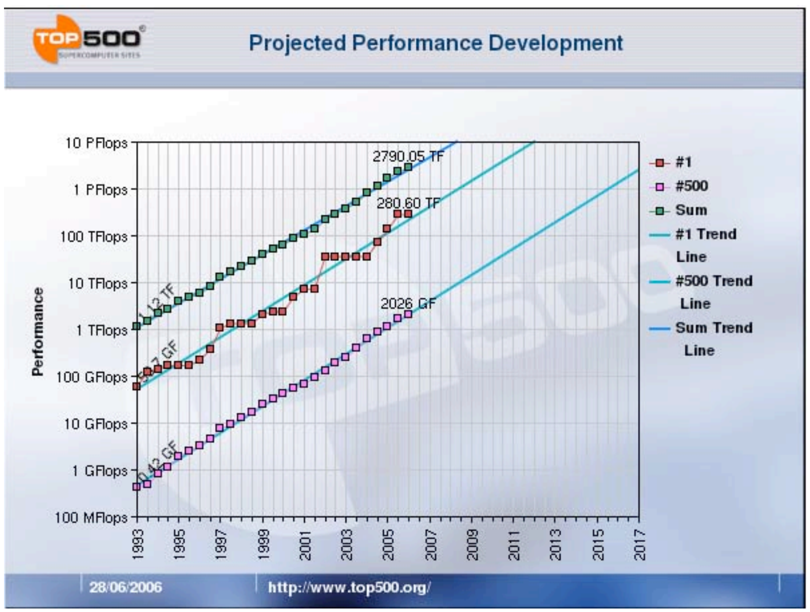
- Understanding File Systems at Rest: www.pdsi-scidac.org/fsstats

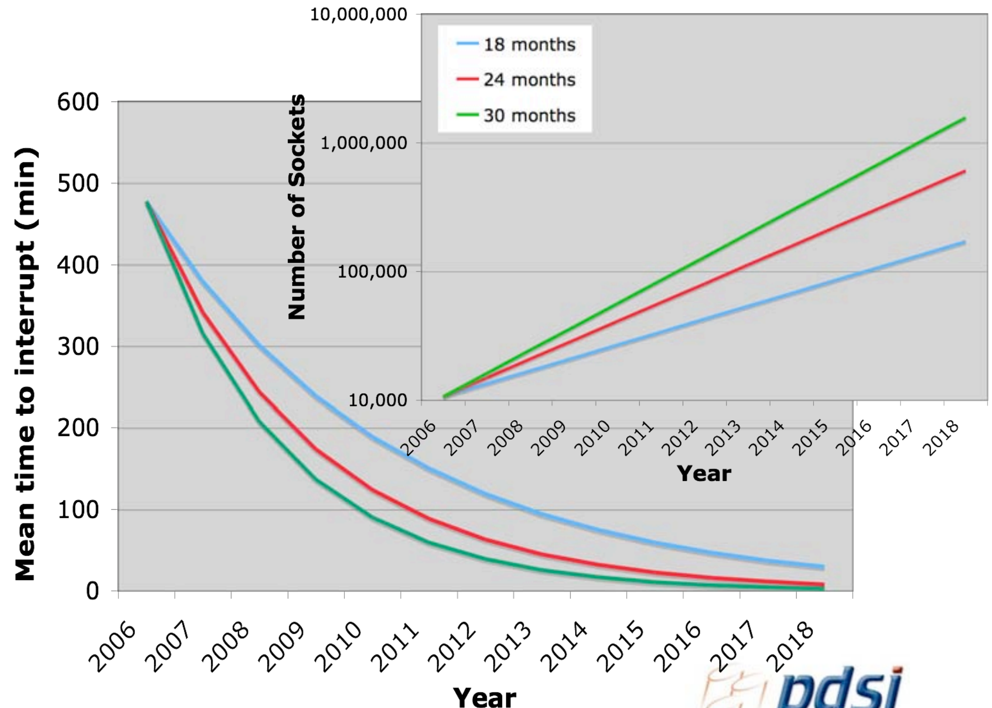
Pacific Northwest National Lab, SATA disks, RAID 5, ext3 FS

```
skipped special files:19135 skipped duplicate hardlinks:21 skipped snapshot dirs:0 total capacity used:233670721104 KB
total user data:233932402222 KB percent overhead:-0.1120
file size Histo:
count=12338926 average=18958.589506 KB
min=0 KB max=757630040 KB
  [          0-         2 KB): 3303866 (26.78%) ( 26.78% cumulative)      1996763.67 KB ( 0.00%) (  0.00% cumulative)
  [          2-         4 KB):  883060 ( 7.16%) ( 33.93% cumulative)      2534585.51 KB ( 0.00%) (  0.00% cumulative)
  [          4-         8 KB):  917461 ( 7.44%) ( 41.37% cumulative)      5182409.88 KB ( 0.00%) (  0.00% cumulative)
  [          8-        16 KB):  744358 ( 6.03%) ( 47.40% cumulative)      8591734.47 KB ( 0.00%) (  0.01% cumulative)
  [         16-        32 KB):  731235 ( 5.93%) ( 53.33% cumulative)     16534655.55 KB ( 0.01%) (  0.01% cumulative)
  [         32-        64 KB):  669568 ( 5.43%) ( 58.75% cumulative)     30855148.03 KB ( 0.01%) (  0.03% cumulative)
  [         64-       128 KB):  757320 ( 6.14%) ( 64.89% cumulative)     70214295.14 KB ( 0.03%) (  0.06% cumulative)
  [        128-       256 KB):  631071 ( 5.11%) ( 70.01% cumulative)    114050978.13 KB ( 0.05%) (  0.11% cumulative)
  [        256-       512 KB):  558914 ( 4.53%) ( 74.54% cumulative)    189985048.43 KB ( 0.08%) (  0.19% cumulative)
  [        512-      1024 KB):  597161 ( 4.84%) ( 79.37% cumulative)    443400973.63 KB ( 0.19%) (  0.38% cumulative)
  [       1024-      2048 KB):  479472 ( 3.89%) ( 83.26% cumulative)    676898557.71 KB ( 0.29%) (  0.67% cumulative)
  [       2048-      4096 KB):  363371 ( 2.94%) ( 86.21% cumulative)   1019631931.23 KB ( 0.44%) (  1.10% cumulative)
  [       4096-      8192 KB):  255781 ( 2.07%) ( 88.28% cumulative)   1534778534.48 KB ( 0.66%) (  1.76% cumulative)
  [       8192-     16384 KB):  256358 ( 2.08%) ( 90.36% cumulative)   2894041905.64 KB ( 1.24%) (  3.00% cumulative)
  [      16384-     32768 KB):  230819 ( 1.87%) ( 92.23% cumulative)   5245575759.34 KB ( 2.24%) (  5.24% cumulative)
  [      32768-     65536 KB):  223892 ( 1.81%) ( 94.04% cumulative)  10337335940.35 KB ( 4.42%) (  9.66% cumulative)
  [      65536-    131072 KB):  584808 ( 4.74%) ( 98.78% cumulative)  52004123186.77 KB (22.23%) ( 31.89% cumulative)
  [     131072-    262144 KB):   42167 ( 0.34%) ( 99.12% cumulative)   7784126469.45 KB ( 3.33%) ( 35.22% cumulative)
  [     262144-    524288 KB):   31868 ( 0.26%) ( 99.38% cumulative)  11411821832.03 KB ( 4.88%) ( 40.09% cumulative)
  [     524288-   1048576 KB):   39972 ( 0.32%) ( 99.70% cumulative)  27336893196.49 KB (11.69%) ( 51.78% cumulative)
  [    1048576-   2097152 KB):   17726 ( 0.14%) ( 99.85% cumulative)  25773260950.03 KB (11.02%) ( 62.80% cumulative)
  [    2097152-   4194304 KB):   13237 ( 0.11%) ( 99.96% cumulative)  37985398325.45 KB (16.24%) ( 79.04% cumulative)
  [    4194304-   8388608 KB):    4336 ( 0.04%) ( 99.99% cumulative)  23511276177.30 KB (10.05%) ( 89.09% cumulative)
  [    8388608-  16777216 KB):     783 ( 0.01%) (100.00% cumulative)   8739054420.16 KB ( 3.74%) ( 92.82% cumulative)
  [   16777216-  33554432 KB):     168 ( 0.00%) (100.00% cumulative)   3598648498.69 KB ( 1.54%) ( 94.36% cumulative)
  [   33554432-  67108864 KB):     111 ( 0.00%) (100.00% cumulative)   5587776404.47 KB ( 2.39%) ( 96.75% cumulative)
  [   67108864- 134217728 KB):      25 ( 0.00%) (100.00% cumulative)   2318337024.21 KB ( 0.99%) ( 97.74% cumulative)
  [ 134217728- 268435456 KB):       9 ( 0.00%) (100.00% cumulative)   1559281156.26 KB ( 0.67%) ( 98.41% cumulative)
  [ 268435456- 536870912 KB):       8 ( 0.00%) (100.00% cumulative)   2969396082.00 KB ( 1.27%) ( 99.68% cumulative)
  [ 536870912-1073741824 KB):       1 ( 0.00%) (100.00% cumulative)    757630040.00 KB ( 0.32%) (100.00% cumulative)
```

Peta/Exa-scale projections: more failures

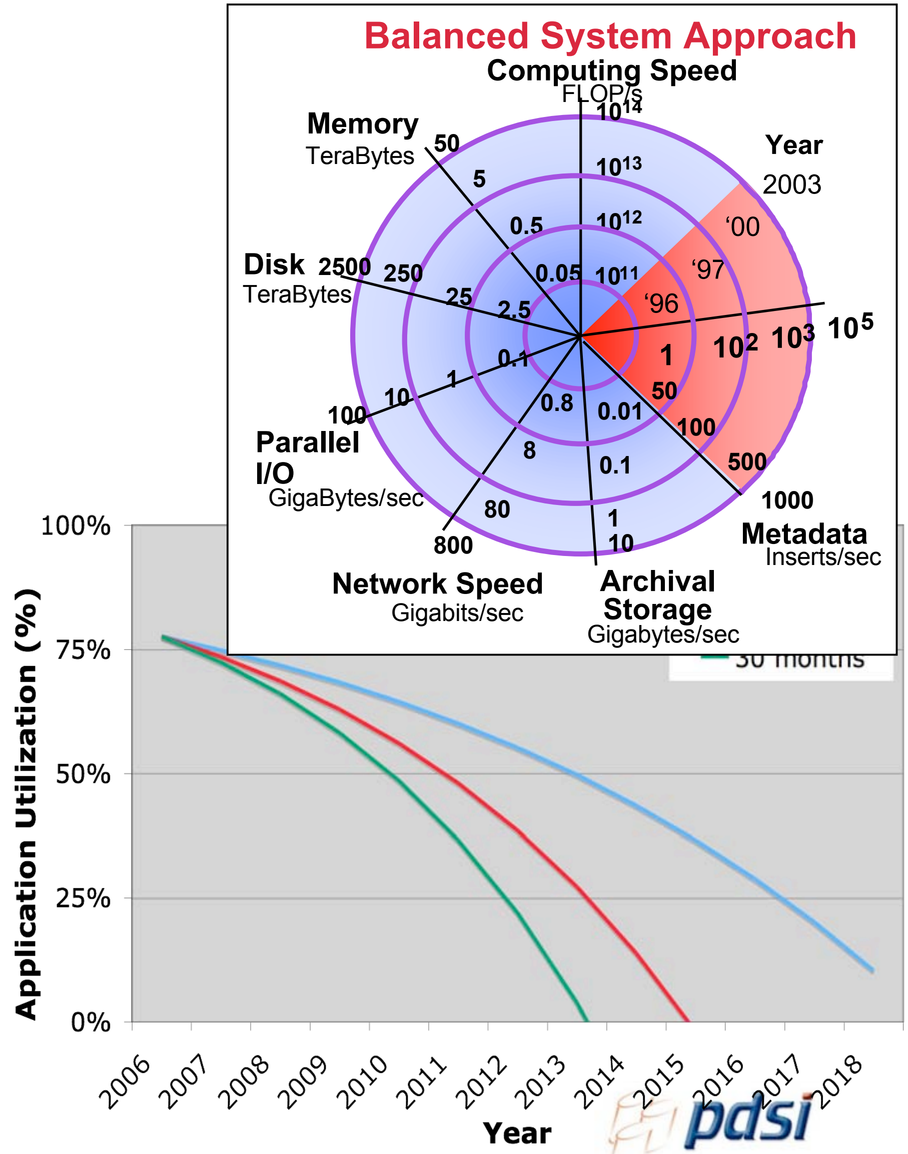
- Continue top500.org annual 2X peak FLOPS
  - Talks: SciDAC07, ICPP07 Keynote, SEG (Oil&Gas), HECURA
- Cycle time flat; Cores/chip reaching for Moore's law
  - 2X cores per chip every 18-30 mos
- # sockets, 1/MTTI = failure rate up 25%-50% per year
  - Optimistic 0.1 failures/yr per chip (vs. LANL historic 0.25)



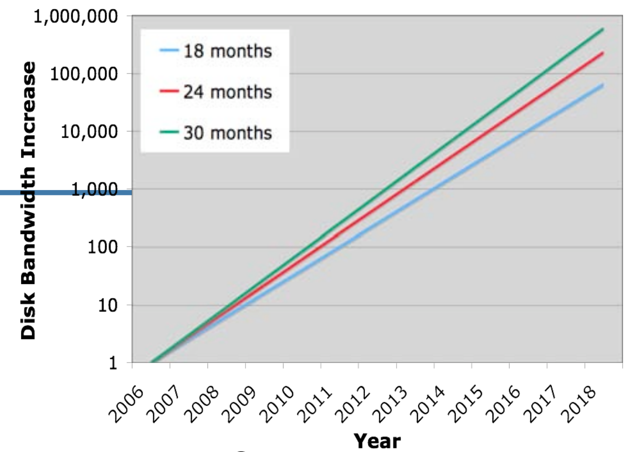


Checkpointing failure tolerance faltering

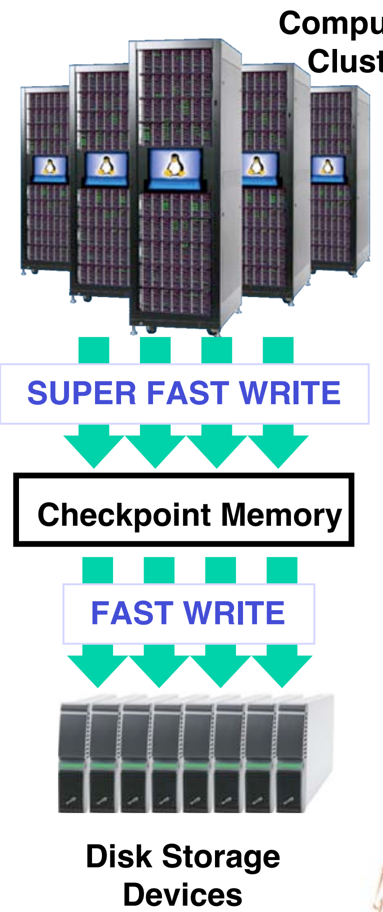
- Periodic (p) checkpoint (t)
- On failure, rollback & restart
- Balanced systems
  - Memory size tracks FLOPS
  - Disk speed tracks both
  - Checkpoint capture (t) constant
  - 1 - App util = t/p + p/(2*MTTI)
    $p^2 = 2*t*MTTI$
  - If MTTI was constant,
    app utilization would be too
  - But MTTI & app utilization drop
- Half machine gone soon
  and exascale era bleak



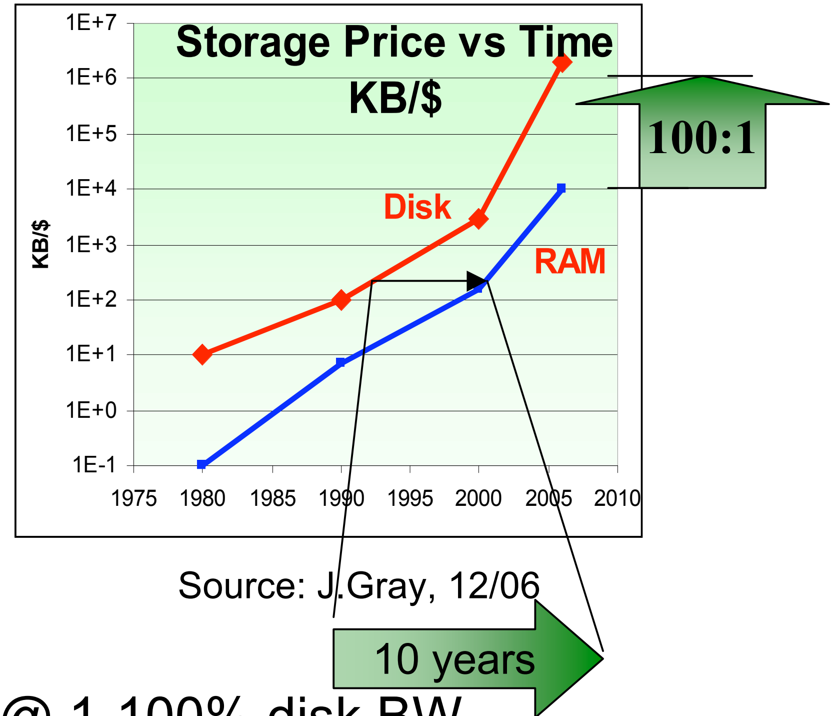
Fixes for Checkpoint/Restart



- Fix with more disk bandwidth?
  - Disk BW +20%/yr: Balance = +67% disks/yr
  - If MTTI drops, need +130% disks/yr !
- Smaller apps don't care
  - Constant HW & MTTI, so balance sufficient
- Compress memory image
  - 25%-50% smaller per byte per year
    fixes MTTI trend given balanced BW
- Process pairs: duplex all calculations
  - At 50% machine effectiveness,
    change to compute-thru-no-restart model
- Special purpose checkpoint devices
  - Fast memory to memory copy, offline to disk
  - Make copy "cheaper", say Flash



Storage Trends: What impact Flash?



Source: J.Gray, 12/06

- Disk technology trends hold >10 yrs
  - Perpendicular then HAMR
  - 2.5" disk for enterprise soon
  - 2.5" SATA in ~2 years
- Disk vs DRAM
  - "Access Gap" isn't closing
- Flash may change the game
  - 100%/year capacity growth recently
  - ~25X disk $/GB & closing
  - Only 10^4-10^6 write cycles is 3 years @ 1-100% disk BW
  - Probable for log-structured disk caches, checkpoint devices
- And then holographic, phase change NVRAM, nanotube wires …
- Good place for PDSW to explore more

Eg. POSIX Ext: Lazy I/O data integrity

- O_LAZY in *flags* argument to **open**(2)
- Requests lazy I/O data integrity
  - Allows filesystem to relax data coherency to improve performance for shared-write file
  - Writes may not be visible to other processes or clients until after **lazyio_propagate**(2), **fsync**(2), or **close**(2)
  - Reads may come from local cache (ignoring changes to file on backing storage) until **lazyio_synchronize**(2) is called
  - Does not provide synchronization across processes or nodes – program must use external synchronization (e.g., pthreads, XSI message queues, MPI) to coordinate

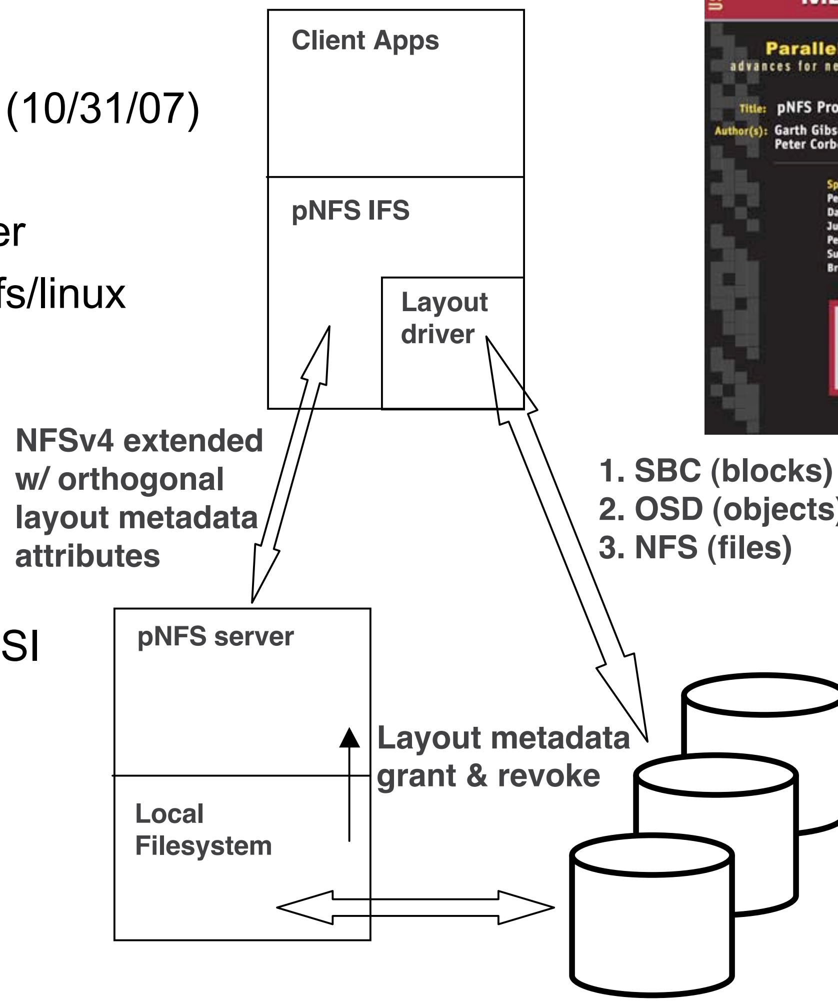
THE Open GROUP Making standards work® High End Computing Extensions Working Group

You are here: Platform Forum > HECEWG > Documents

Created	Title (see details)	Version (+ implies others)	Formats (download)
17-Aug-2006	Evaluation Criteria for Proposed High End Computing Extensions to the POSIX I/O API	1.2 +	PDF
30-Jun-2006	Manpage - readdirplus	1	PDF
30-Jun-2006	Manpage - lockg (group lock)	1	PDF
30-Jun-2006	Manpage - sutoc (convert file handle to file descriptor)	1	PDF
30-Jun-2006	Manpage - NFSV4acls	1	PDF
30-Jun-2006	Manpage - openg (group open)		PDF
30-Jun-2006	Manpage - statlite and family of light weight stat calls	1	PDF
30-Jun-2006	Manpage - open (O_LAZY flags)	1	PDF
30-Jun-2006	POSIX I/O High Performance Extensions presentation Panasas SC05	1	PDF
30-Jun-2006	POSIX I/O High Performance Computing Extensions ASC SC05 presentation	1	PDF
30-Jun-2006	High End Computing Early Goals for extesions to POSIX I/O API	1	PDF
30-Jun-2006	A Business Case for Extensions to the POSIX I/O API for High End, Clustered, and Highly Concurrent Computing	1	PDF

pNFS: Parallel File System Standards

- IETF NFSv4.1: draft soon!
  - draft-ietf-nfsv4-minorversion1-15.txt (10/31/07)
  - Includes pNFS, sessions
  - U.Mich/CITI impl'g Linux client/server
  - www.citi.umich.edu/projects/asci/pnfs/linux
- Three (or more) flavors of out-of-band metadata attributes:
  - FILES: NFS/ONCRPC/TCP/IP/GE for files built on subfiles NetApp, Sun, IBM, U.Mich/CITI
  - BLOCKS: SBC/FCP/FC or SBC/iSCSI for files built on blocks EMC (-pnfs-blocks-04.txt, 10/4/07)
  - OBJECTS: OSD/iSCSI/TCP/IP/GE for files built on objects Panasas (-pnfs-obj-04.txt, 9/5/07)



Far-reaching Standards Incubation

- Guiding, dogging, driving, …., UMich critical to deploying standard

StorageMojo RSS Feed

- Architecture
- Backup
- Clusters
- Disk
- Enterprise
- Future Tech
- Information Management
- NAS, IP, iSCSI
- Off-Topic
- Price Lists
- SAN, FC
- Security & Public Policy
- SOHO/SMB
- SSD/Flash Disk

pNFS technical intro

October 15th, 2007 by Robin Harris in Future Tech, NAS, IP, iSCSI, Clusters, Architecture

I don't normally link and run but this is a good article on the Next Big Thing in NFS v4.1.

Written by 3 NetApp engineers, Garth Goodson, Sai Susarla, and Rahul Iyer, Standardizing Storage Clusters offers a good overview of what's new. It's on the ACM Queue web site.

If paragraphs like

> *protocol operations*
>
> *The pNFS protocol adds a total of six operations to NFSv4. Four of them support layouts (i.e., getting, returning, recalling, and committing changes to file metadata). The two other operations aid in the naming of data servers (i.e., translating a data server ID into an address and getting a list of data servers). All the new operations are designed to be independent of the type of layouts and data-server names used. This is key to pNFS's ability to support diverse back-end storage architectures.*

get you interested the article is well worth a read.

The StorageMojo take
pNFS is going to commoditize parallel data access. In 5 years we won't know how we got along without it.

Robin Harris

Community Recognition

IEEE Reynold B. Johnson Information Storage Systems Award

Sponsored by: IBM Almaden Research Center

Nomination Form | Recipients | Committee Roster

Nomination Deadline - 31 January

The IEEE Reynold B. Johnson Information Storage Systems Award was established by the Board of Directors in 1991 and may be presented annually for outstanding contributions to information storage systems, with emphasis on computer storage systems.

It may be presented to an individual, multiple recipients or team up to three in number.

It is named in honor of Reynold B. Johnson, who is renowned as a pioneer of magnetic disk technology and was founding manager of the IBM San Jose Research and Engineering Laboratory, San Jose, California in 1952, where IBM research and development in the field was centered.

In the evaluation process, the following criteria are considered: computer storage is emphasized, achievement may relate to materials, concepts, design, hardware or software, may be theoretical or experimental, but will be judged on the impact and the historical significance on the evolution of computer storage systems, and the quality of the nomination.

- PETASCALE DATA STORAGE WORKSHOP AGENDA
  - 8:30-9:00: Introduction by Garth Gibson
  - 9:00-10:20: Paper Session 1
    - E. Krevat, Ranjit Noronha, Brent Welch, Peter Braam
  - 10:30-11:00: Poster Session 1
    - Ethan Miller or Garth Gibson have easels/boards/clips for posters
  - 11:00-12:20: Paper Session 2
    - Jonathan Koren, Swapnil Patil, Richard Golding, Sage Weil
  - 12:30-2:00: Lunch (on your own)
  - 2:00-3:20: Paper Session 3
    - Henry Monti, Phil Roth, Andrew Konwinski, Ann Chervenak
  - 3:30-3:00: Short Annoucements
    - Sign up with Garth; Announce availability of data, code, working groups etc
  - 4:00-5:00: Poster Session 2
  - 5:00: Closing by Garth Gibson