Scientific Discovery Through Advanced Computing (SciDAC): Petascale Data Storage Institute

CMU: Garth Gibson (PI)
UCSC: Darrell Long (co-PI)
U.Michigan: Peter Honeyman (co-PI)
Los Alamos National Lab: Gary Grider (co-PI)
Lawrence Berkeley Labs (NERSC): Bill Kramer (co-PI)
Oak Ridge National Lab: Philip Roth (co-PI)
Pacific Northwest National Lab: Evan Felix (co-PI)
Sandia National Lab: Lee Ward (co-PI)

SciDAC Institutes

"The SciDAC Institutes are university-led centers of excellence intended to complement [other] efforts. This will be achieved by focusing on major software issues through a range of collaborative research interactions. It may:

- Concentrate efforts
- Focus on a single method or technique
- Be a focal point for bringing together a critical mass of leading experts
- Forge relationships between between experts in software, applications, HPC, industry
- Reach out to engage a broader community
- Have a dimension of training and outreach"

Petascale Data Storage Institute

Petascale computing makes petascale demands on storage:

Capacity, performance, concurrency, reliability, availability, and manageability.

Parallel file systems are barely keeping pace pre-petascale.

PDSI focuses on petascale scientific computing storage problems,

With special attention to community issues such as interoperability, community buy-in, and shared tools.

Bringing together diverse experience in applications and file and storage systems, its members will:

Collaborate on requirements, standards, algorithms, and development and performance tools.

PDSI Thrusts

Three overall thrusts, six projects:

Dissemination

- Community building: workshops, tutorials, course tools
- APIs & standards: incubate, prototype, validate

Collection

- Failure data collection, analysis & publication
- Performance trace collection & benchmark publication

Innovation

- IT automation applied to HEC systems & problems
- Novel mechanisms for core HEC storage problems

PDSI: Dissemination

Outreach

- Led by Garth Gibson, CMU
- Multiple annual workshops:
  - e.g. SC06 petascale data storage workshop, Nov 17
- Training material: conference tutorials, university course, etc
- A resource to support other SciDAC centers/projects

Protocol/API extensions

- Led by Gary Grider, LANL
- Facilitate standards development and deployment
- Validate and demonstrate new extensions/protocols
  - E.g. POSIX extensions (group open, weak consistency in data & metadata, vector read/write), rich metadata definitions, archive interfaces, compute in disk, Parallel NFS, QoS for storage, storage networking topology, data layout specifications

PDSI: Collection

Performance analysis

- Led by William Kramer, NERSC
- Capture traces, characterize, model/replay, benchmark
  - E.g. BLAST, ScalaBLAST (biology) · CCSM (climate) · CTH, Sierra/Salinas/Calore (materials) · EVH-1, RAGE (astrophysics) · Firetec (wildfire prediction) · GYRO, GTC (plasma turbulence) · MADNESS, NWChem, GA-Tools (chemistry) · MCNP (nuclear physics) · Sierra/Alegra (high-energy physics) · QCD (quantum chromodynamics)
- Publish data, analysis and tools

Failure analysis

- Led by Gary Grider, LANL
- Capture failure/error/usage event records
  - E.g. LANL 9 years, 23,000 outages, 22 systems: DSN06
    Other HEC sites (PSC,…) and internet service sites to come
- Publish data, analysis and tools

PDSI: Innovation

IT automation

- Led by Greg Ganger, CMU
- Extensive instrumentation for machine learning techniques
- Visualization and other human engaging techniques
- Planning data layout and access scheduling for performance
- Automating diagnosis, tuning, failure/ill-performance healing

Novel mechanisms

- Led by Darrell Long, UCSC
- Innovation in core HEC storage problems
  - E.g. global/WAN access to parallel storage/grids, security in federated systems, collective operations, predictable performance when big workloads share, rich metadata at scale, metadata search, integration with para-virtualization

PDSI Team - Part 1

Carnegie Mellon Univ.

- Garth Gibson (PI)
  - RAID, NASD/OSD, Active disks, Panasas CTO, SNIA-TC, RBJ IEEE Tech Field Award
- Greg Ganger
  - Ursa Major, Soft updates, exokernel, CFFS, Self-*, survivable storage
- Anastassia Ailamaki
  - Cache & disk scheduling for DB performance
- Dave O'Hallaron
  - Quake earthquake simulator, Gordon Bell Award, co-author Computer Systems: A Programmer's Perspective
- Bianca Schroeder
  - Failure data collection & statistics, DB & web services, SC & QoS scheduling

Los Alamos Nat. Lab

- Gary Grider (co-PI)
  - ASC FS technology initiative coord, networking for parallel storage, SGPFS incubation, IO for ASCI Q & Blue Mtn
- James Nunez
  - Networking & IO team leader, NFSv4 & metadata oversight, benchmarking & app perf

Nat. Energy Res. Sci. Center

- William Kramer (co-PI)
  - NERSC GM, HPC performance analysis, HPC for meteorology & physics, integration of 1st large T3E

Oak Ridge Nat. Lab

- Philip Roth (co-PI)
  - Automated perf diagosis & problem search, event trace reduction tools

PDSI Team - Part 2

Pacific Northwest Nat. Lab

- Evan Felix (co-PI)
  - Integrated Lustre on HPCS2 & 300 TB archive, AutoRAID development, active storage processing
- Robert Farber
  - 10^15 B/s particle tracking system, 6B agent epidemic modeling, neural net compiler & use in DNA sequencing

Univ. of Michigan

- Peter Honeyman (co-PI)
  - NFSv4 replication, pNFS, Kerberos, USENIX board
- William Adamson
  - NFSv4 Linux contributor, GridNFS, pNFS
- J. Bruce Fields
  - NFSv4 Linux contributor, rpcsec_gss, ACLs

Sandia Nat. Lab

- Lee Ward (co-PI)
  - Catamount VFS for Cray XT3, ENFS for Cplant, UniTree HSM, X/DSM virtual disk manager

Univ. of California, Santa Cruz

- Darrell Long (co-PI)
  - Nat. Security Panelist, storage reliability & synch, MEMS, object FS, network RAID
- Scott Brandt
  - Petascale metadata, object FS, MEMs storage, storage QoS,
- Ethan Miller
  - Object security, petascale metadata, disk reliability
- Carlos Maltzahn
  - Rich metadata, web proxy IO

PDSI Next Steps

Waiting for official confirmation of funding

Kickoff meeting collocated with this workshop

SC06 Petascale Data Storage Workshop: Nov 17
- CFP to appear soon, suggestions welcome

Web site to appear: www.pdl.cmu.edu/PDSI

Q&A

Eg., LANL HEC FAULT DATA

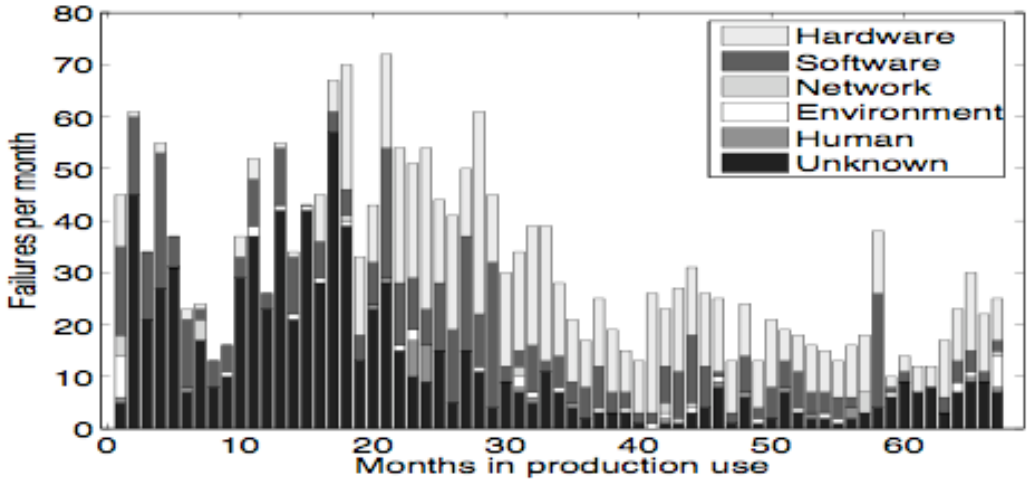
Expand data collection

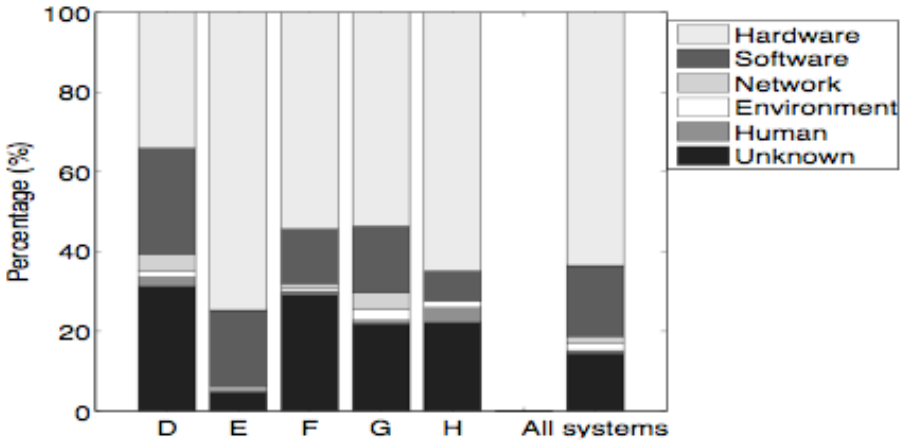
- Usage/load data
- Event/error data
- Other HEC sites

Wide open sharing of data

- Data files go on web

www.pdl.cmu.edu/FailureData/





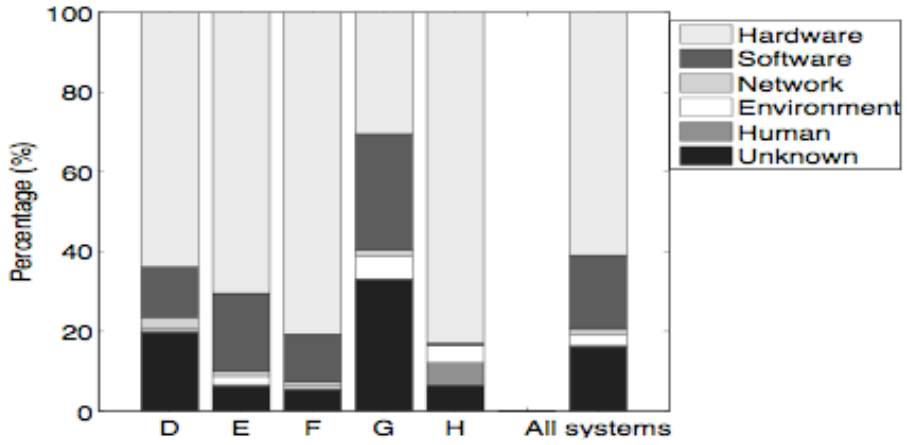


Figure 1: *The breakdown of failures into root causes (left) and the breakdown of downtime into root causes*

Eg. POSIX Ext: Lazy I/O data integrity

Specify O_LAZY in *flags* argument to **open**(2)

Requests lazy I/O data integrity

- Allows network filesystem to relax data coherency requirements to improve performance for shared-write file
- Writes may not be visible to other processes or clients until **lazyio_propagate**(2), **fsync**(2), or **close**(2) is called
- 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 actions

This is a hint only

- if file system does not support lazy I/O integrity, it does not have to

www.pdl.cmu.edu/POSIX/

Eg., pNFS: Scalability into Mainstream

IETF NFSv4.1

- draft-ietf-nfsv4-minorversion1-02.txt 3/06
- Includes pNFS, stronger security, sessions/RDMA, directory delegations
- U.Mich/CITI impl'g Linux client/server

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-00.txt)
- OBJECTS: OSD/iSCSI/TCP/IP/GE for files built on objects Panasas, Sun (-pnfs-obj-00.txt)