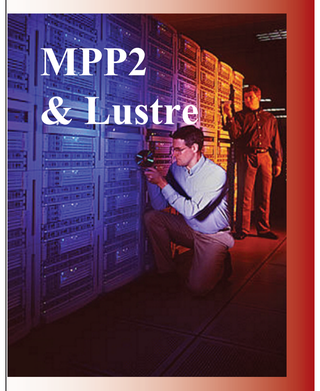
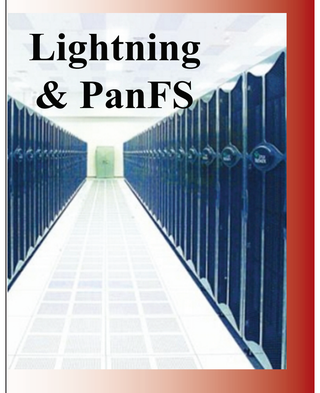
Petascale Data Management: Guided by Measurement

Office of Science U.S. DEPARTMENT OF ENERGY

SciDAC Scientific Discovery through Advanced Computing





petascale data storage institute

www.pdsi-scidac.org/

pdsi www.pdsi-scidac.org

MEMBER ORGANIZATIONS

- Los Alamos National Laboratory – institute.lanl.gov/pdsi/
- Oak Ridge National Laboratory – www.csm.ornl.gov/
- National Energy Research Scientific Computing Center pdsi.nersc.gov/
- Pacific Northwest National Laboratory – www.pnl.gov/
- Parallel Data Lab, Carnegie Mellon University – www.pdl.cmu.edu/
- Sandia National Laboratories – www.sandia.gov/
- Center for Information Technology Integration, U. of Michigan www.citi.umich.edu/projects/pdsi/
- University of California at Santa Cruz – www.pdsi.ucsc.edu/

The Computer Failure Data Repository

- Goal: to collect and make available failure data from a large variety of sites
  - Better understanding of the characteristics of failures in the real world
- Now maintained by USENIX at cfdr.usenix.org/

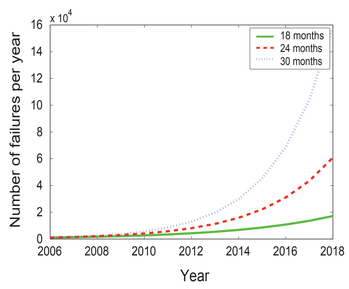
NAME	SYSTEM TYPE	SYSTEM SIZE	TIME PERIOD	TYPE OF DATA
Los Alamos NATIONAL LABORATORY EST.1943	22 HPC clusters	5000 nodes	9 years	Any node outage
PITTSBURGH SUPERCOMPUTING CENTER	1 HPC cluster	765 nodes 3,400 disks	5 years	Hardware/ disk drive replacements
1 Internet service, Various HPC sites	3 storage, many HPC clusters	>10,000 nodes >100,000 disks	1 mth - 5 yrs	
Pacific Northwest National Laboratory Operated by Battelle for the U.S. Department of Energy	MPP2 system HPC cluster	980 nodes	4 yrs	Hardware failures
NERSC NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER	HPC cluster	A number of production systems	5 years	I/O specific failures
COM 1	Internet services cluster	Multiple distributed sites	1 mth	Hardware failures
COM 2	Internet services cluster	Multiple distributed sites	20 mths	Warranty service log of hardware failures
COM 3	Internet services cluster	Large external storage system	1 yr	Aggregate quarterly stats of disk failures

BASE ASSUMPTIONS

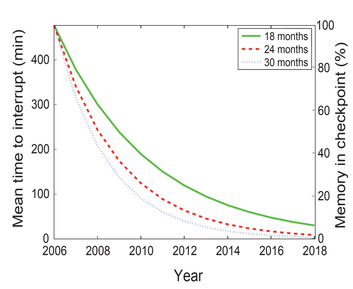
- Top500.org: Performance of top installations doubles every year
- Technology trends:
- Little/no increase in clock speed → cycles coming from more sockets and more cores per socket
- # cores per socket doubles every 18-30 months
- Failure data:
- Hardware is not getting more reliable over time
- Failure rate proportional to # sockets (optimistically 0.1 failures/socket/year)

PREDICTIONS

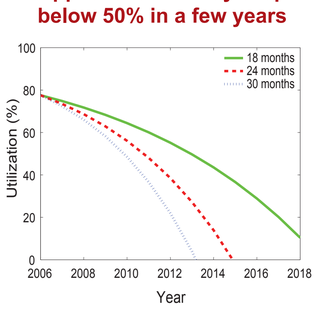
Dramatic increase in failure rates



Significant drop in MTTI

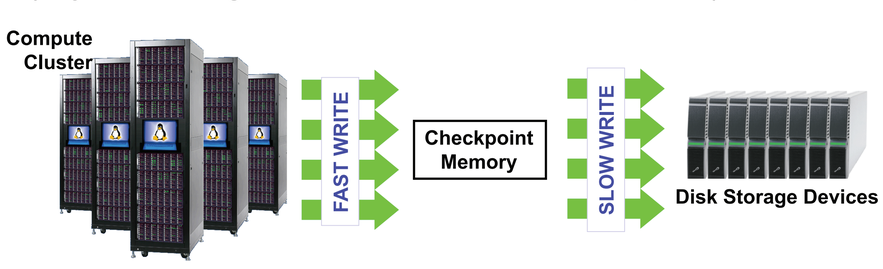


App. utilization may drop below 50% in a few years



POSSIBLE SOLUTIONS

- Make more reliable chips (not likely given history data we have seen)
- Use non-growing portion of bigger machines (not an answer for demanding applications)
- Checkpoint smaller fraction of memory (compression)
- Increase storage bandwidth (significantly more drives)
- Specialized checkpoint devices – use memory to stage checkpoint (maybe flash if organization for write bandwidth is improved)



- Investigate alternative methods, e.g. process pairs
- Expensive, but will pay off once application utilization drops below 50%

Filesystems Statistics Survey

GOALS

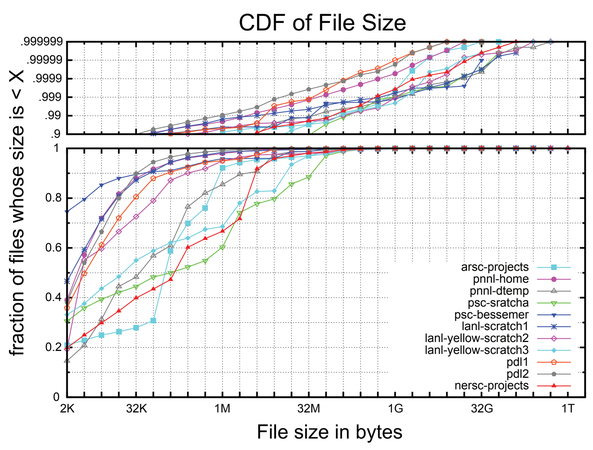
- Gather & build large DB of static filetree summary
  - Build small, non-invasive, anonymizing stats gather tool
  - Distribute fsstats tool via easily used web site
  - Encourage contributions (output of tool) from many FSs
  - Offer uploaded statistics & summaries to public

Label	Date (2008)	Type	File System	Total Size TB	Total Space TB	# files M	# dirs K	max size GB	max space GB	max dir ents	max name bytes	avg file MB	avg dir ents
Satyanarayanan81	<1981	Home	TOPS10		<.0016	.086						.012	
Irlam93	Nov 1993				.259	12						.022	
SFS97	<1997		NFS					.001		30		.027	30
Douceur99	Sept 98	Desktops	NTFS	10.5		141						.079	
VU2005	2005	Home	UNIX			1.7		2				.327	
SFS2008	<2008		NFS					0.32		30		.531	30
CMU gg1	4/10	OS	HFS+	.044	.046	1.0	258	2.1	2.1	100344	252	.046	5
CMU gg2	4/10	Home	HFS+	.0098	.0099	.028	3.2	.328	.328	448	123	.37	10
CMU gg3	4/10	Media	HFS+	.065	.066	.042	2.6	2.2	2.2	536	129	1.6	17
CMU pdl1	4/9	Project	WAFL	3.93	3.68	11.3	821	37.7	23.4	56960	255	.37	15
CMU pdl2	4/9	Project	WAFL	1.28	1.09	8.11	694	37.7	23.4	89517	255	.17	14
NERSC	4/8	Project	GPFS	107	107	20.5	917	616	523	143365	152	5.3	23
PNNL nwfs	3/17	Archival	Lustre	265	264	13.7	1824	1074	1074	57114	232	19.3	9
PNNL home	3/17	Home	ADVFS	4.7	4.3	10.1	682	268	35	23556	255	.46	16
PNNL dtemp	3/17	Scratch	Lustre	22.5	19.2	2.2	51	1074	1075	8004	89	10.3	44
PCS scratch	3/27	Scratch	Lustre	32	32	2.07	451	173	173	64010	160	15.6	6
PSC bessemer	3/27	Project	Lustre	3.7	3.7	0.38	15	51	51	8226	89	9.6	26
LANL scratch1	4/1	Scratch	PanFS	9.2	10.7	1.52	120	134	154	14420	90	6.0	14
LANL scratch2	4/10	Scratch	PanFS	25	26	3.30	241	978	1076	50000	73	8.2	15
LANL scratch3	4/10	Scratch	PanFS	26	29	2.58	374	998	1099	45002	65	10.9	8
ARSC seau1	3/13	Archival	SAM-QFS	305	4.3	10.5	326	386	13.7	62803	245	29	34
ARSC seau2	3/14	Archival	SAM-QFS	115	4.6	5.3	116	366	7.0	25008	144	21.7	47
ARSC nanu1	3/12	Archival	SAM-QFS	69	4.5	6.7	338	601	13.6	58648	234	10.4	21
ARSC projects	3/13	Archival	SAM-QFS	32	.93	6.2	898	171	3.7	24153	81	5.2	8

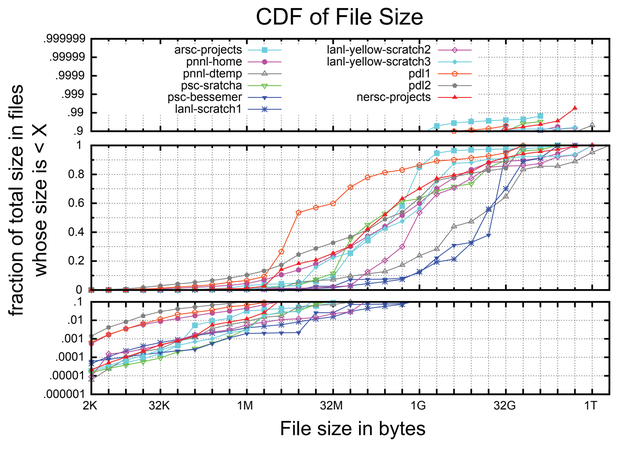
THE FSSTATS TOOL

- Authored by Shobhit Dayal, CMU and Marc Unangst, Panasas
  - Portable perl code, GPL license
  - One file, built-in doc (perldoc)
- Gathers size, links and age histograms
  - User data, blocks used, directory entries, name length, symlink lengths
- Outputs ASCII histogram, CSV, checkpoints
  - Can restart from checkpoint so long captures can be killed
- Upload the CSV output
- Cumulative distribution graphs show % of "metric" below given size
- Graphs special case "five 9s" at the bottom or top end of scale as appropriate

RESULTS



- Wide variation in file size distributions between sites
- Huge fraction of files very small even in supercomputers: 20-80% are < 2KB
- Long tail of small numbers of really huge files



- By space used, however, small files consume almost no space
- Very wide variation in file size which divides space in half: 8MB - 16GB

Better Data Management Comes From Better Understanding of Data
Monitor Your Data Systems and Publish their Statistics!

Los Alamos NATIONAL LABORATORY EST.1943 · OAK RIDGE National Laboratory · NERSC NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER · Pacific Northwest NATIONAL LABORATORY · PARALLEL DATA LABORATORY CARNEGIE MELLON UNIVERSITY · Sandia National Laboratories · center for information technology integration UNIVERSITY OF MICHIGAN · UNIVERSITY OF CALIFORNIA