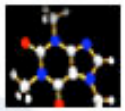
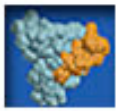
SciDAC
Scientific Discovery through Advanced Computing





pdsi

Failure in Supercomputers and Supercomputer Storage

Los Alamos Computer Science Symposium, Oct 15, 2008

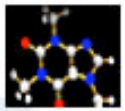
Garth Gibson

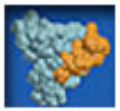
School of Computer Science, Carnegie Mellon University
DOE SciDAC Petascale Data Storage Institute
Chief Technology Officer, Panasas Inc.

garth@cmu.edu, garth@panasas.com

Carnegie Mellon
Parallel Data Laboratory

panasas®





PETASCALE DATA STORAGE INSTITUTE

pdsi

Grow HPC data systems from Tera to Peta & Exa scale

3 universities, 5 US Department of Energy National Labs

www.pdsi-scidac.org

Challenge: Larger = more complexity, more analytics & more failures

center for information technology integration
UNIVERSITY OF MICHIGAN

Carnegie Mellon

OAK RIDGE National Laboratory

BERKELEY LAB Parallel Data Laboratory

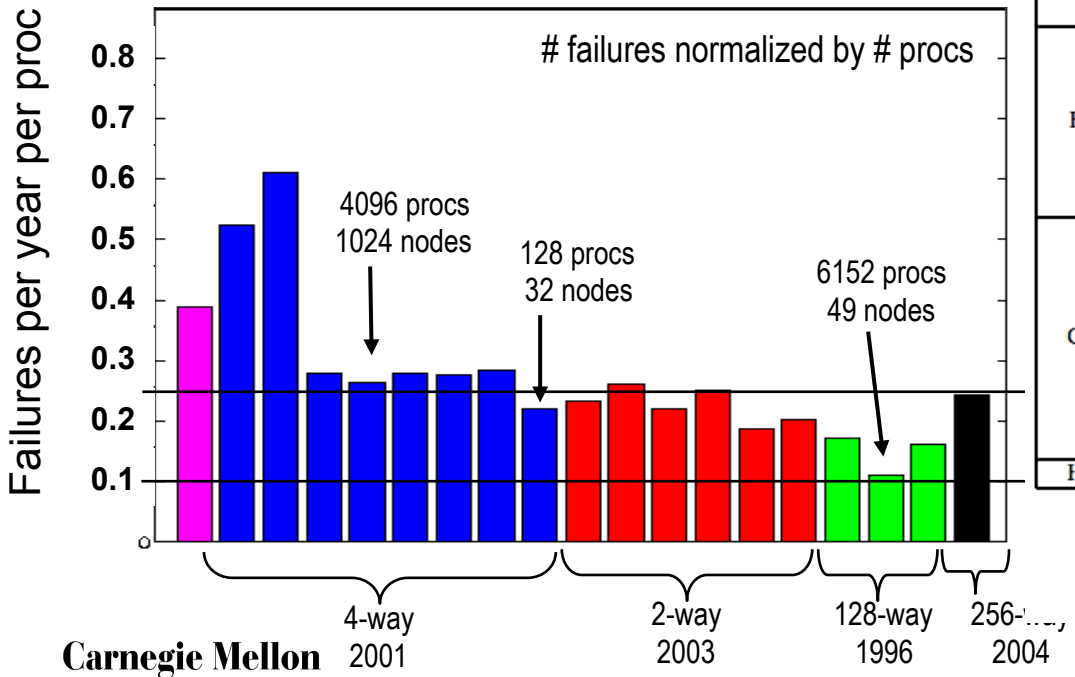
Sandia National Laboratories

Los Alamos NATIONAL LABORATORY EST.1943

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

LANL interrupt history

- Los Alamos root cause logs
  - releases 23,000 events causing application interruption
  - 22 clusters & 5,000 nodes
  - covers 9 years & continues



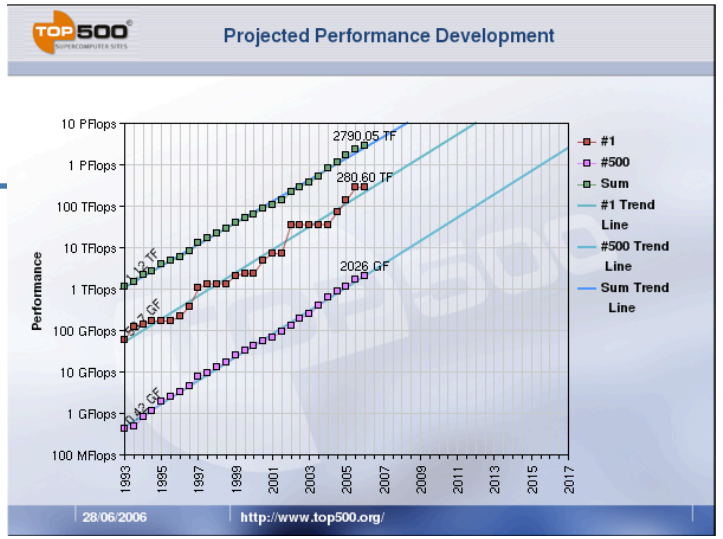
(I) High-level system information				(II) Information per node category			
HW	ID	Nodes	Procs	Procs /node	Production Time	Mem (GB)	NICs
A	1	1	8	8	N/A – 12/99	16	0
B	2	1	32	32	N/A – 12/03	8	1
C	3	1	4	4	N/A – 04/03	1	0
D	4	164	328	2 2	04/01 – now 12/02 – now	1 1	1 1
E	5	256	1024	4	12/01 – now	16	2
	6	128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

Table 1. *Overview of systems. Systems 1–18 are SMP-based, and systems 19–22 are NUMA-based.*

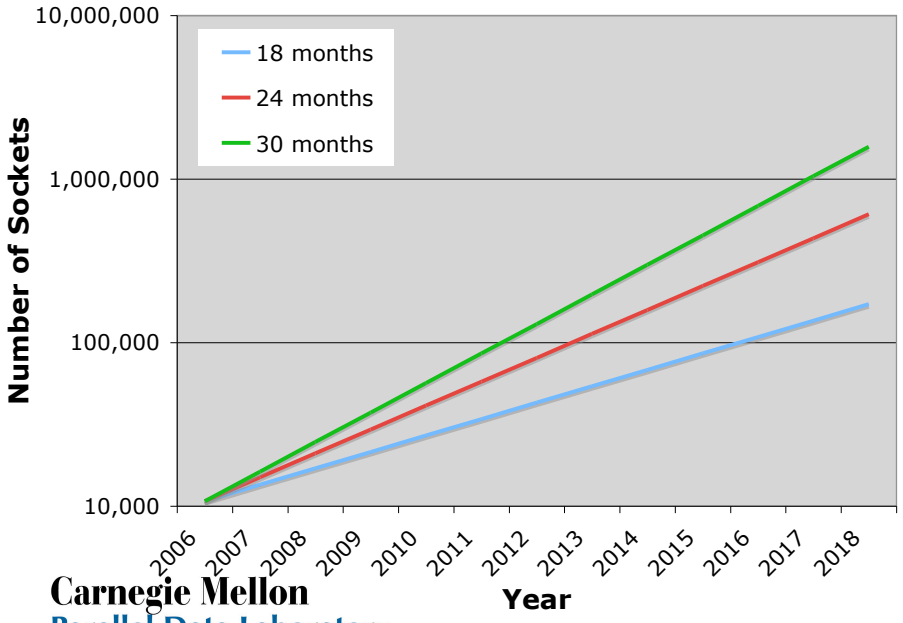
Carnegie Mellon
Parallel Data Laboratory

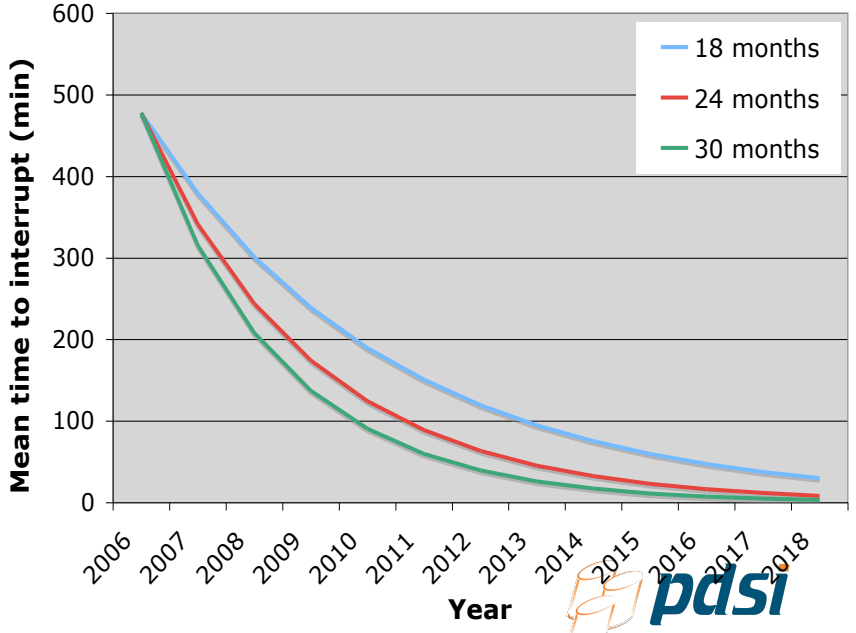
pdsi

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)





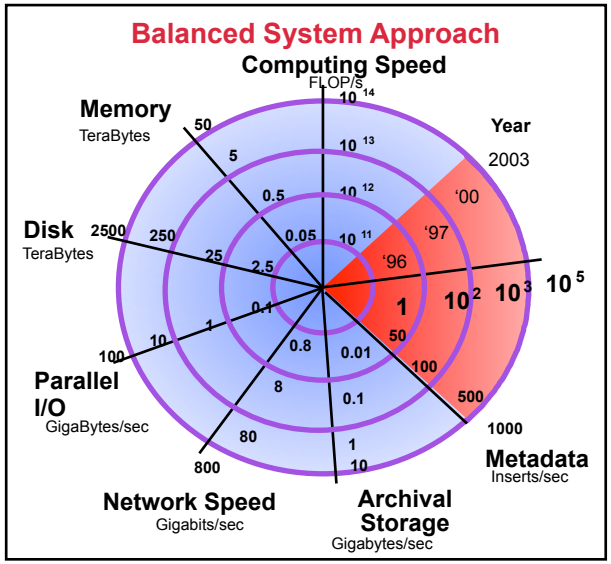
Carnegie Mellon
Parallel Data Laboratory

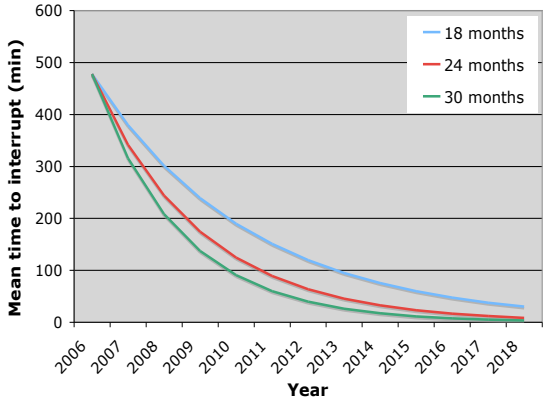
pdsi

Checkpointing failure tolerance in trouble

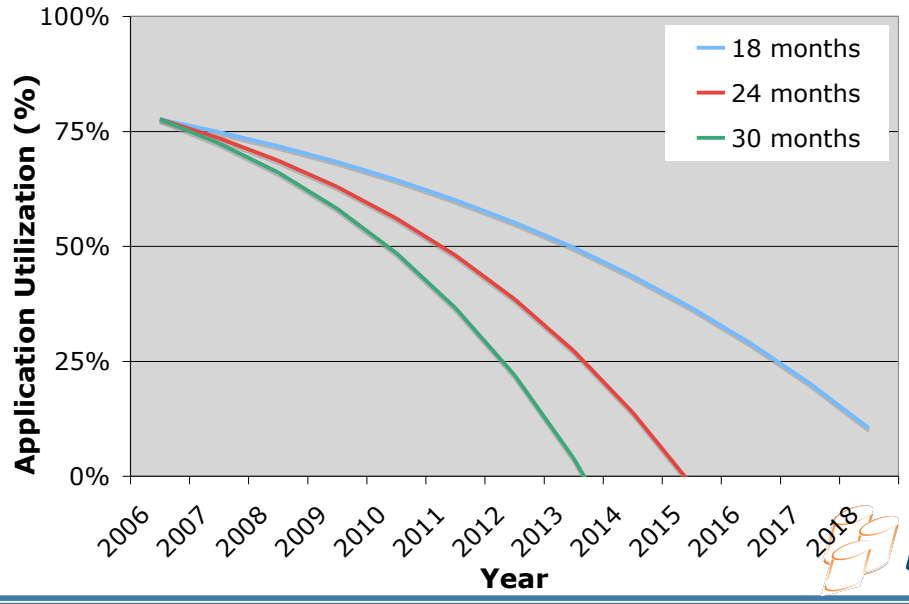
- Periodic (p) pause to checkpoint (t)
  - On failure, roll back & restart
- Balanced systems
  - disk speed tracks FLOPS & mem size, so checkpoint capture (t) is constant time
  - 1 – AppUtilization = t/p + p/(2*MTTI)

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



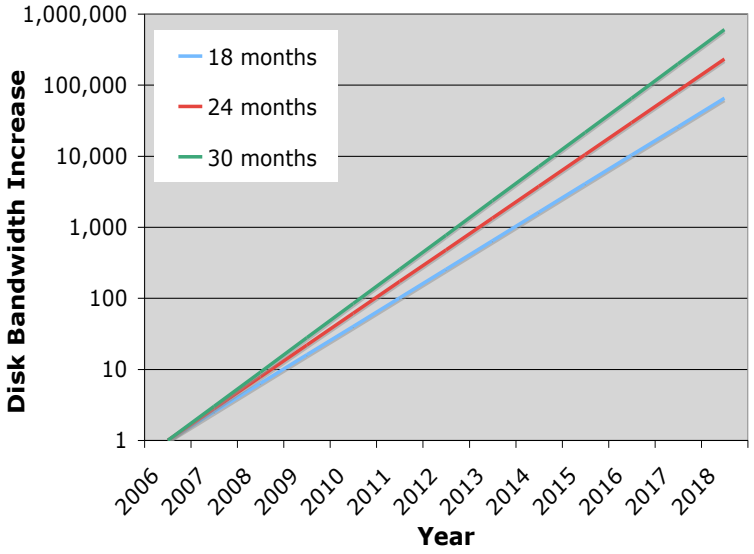


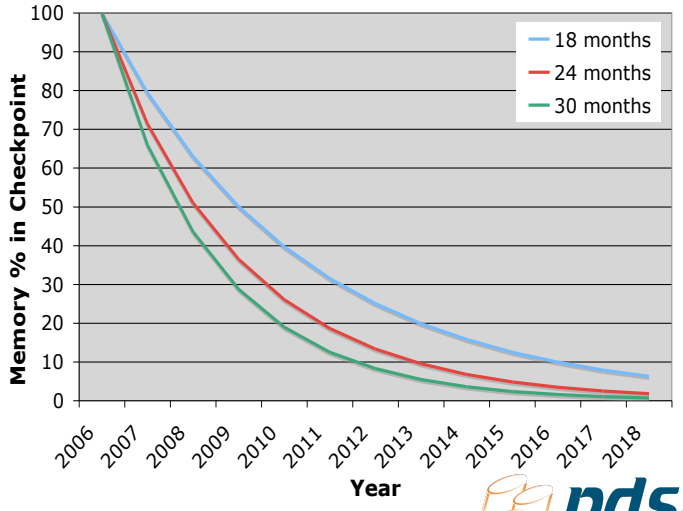
- ***but dropping MTTI kills app utilization!***



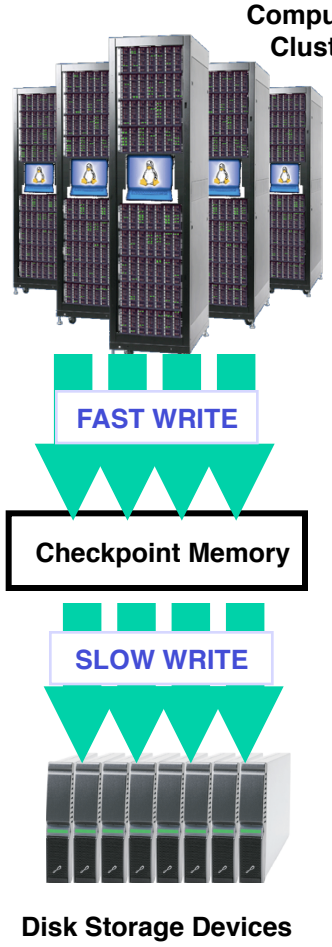
Bolster HEC fault tolerance

- 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

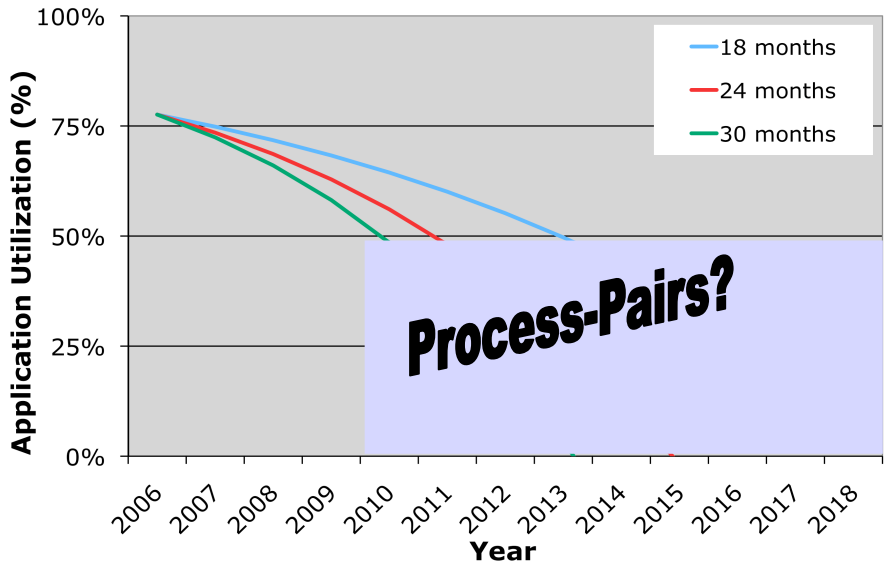




Different approaches

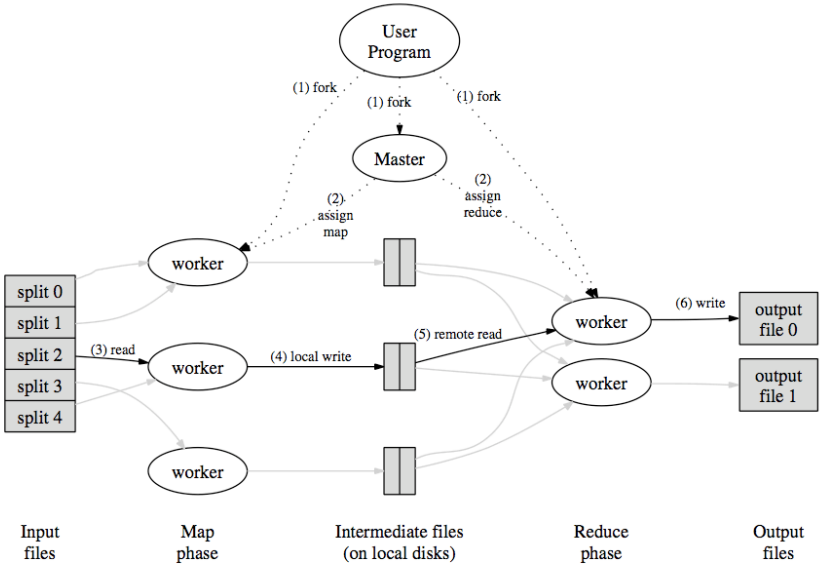


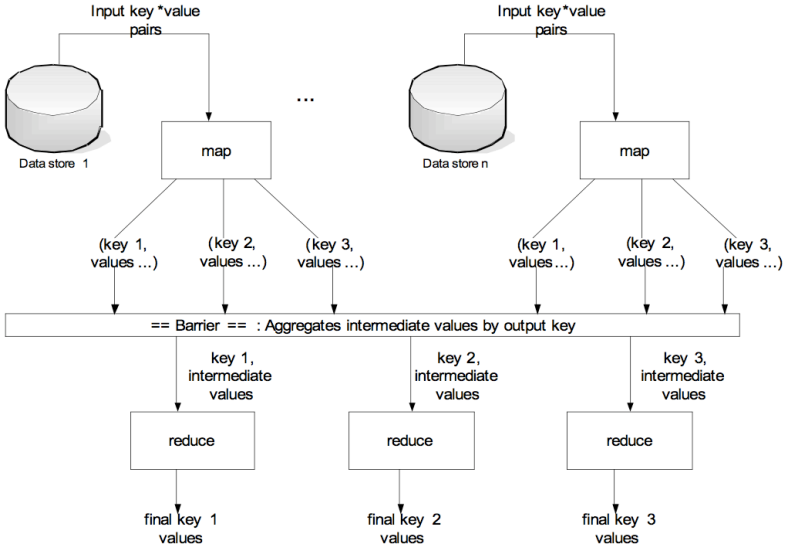
- Dedicated checkpoint device (ie., PSC Zest)
  - Stage checkpoint through fast memory
    - Steal another node's or use cheap memory?
    - Flash $3-4/MBps == 15Krpm disk (raw)?
- Classic enterprise process pairs duplication
  - Flat 50% efficiency cost, plus message duplication



Programming paradigm: MapReduce

- For data-intensive apps that do not have fine grain, lockset processing reqs: MapReduce
  - Divide input, independent transform, then reduce
- Built-in restart of chunks of work





Recap so far

- Failure rates proportional to number of components
  - Specifically, growing # sockets in parallel computer
- If peak compute continues to outstrip Moore's law
  - MTTI will drop, forcing more checkpoints & restarts
- Hero apps, wanting all the resources, bear burden
  - Storage won't keep up b/c cost; dedicated device similar
  - Squeezing checkpoint not believable; process pairs is

- ***Effective fault tolerance increasing challenge***

- Schroeder, B., G. A. Gibson, "Understanding Failures in Petascale Computers," *Journal of Physics: Conference Series* **78** (2007), SciDAC 2007.

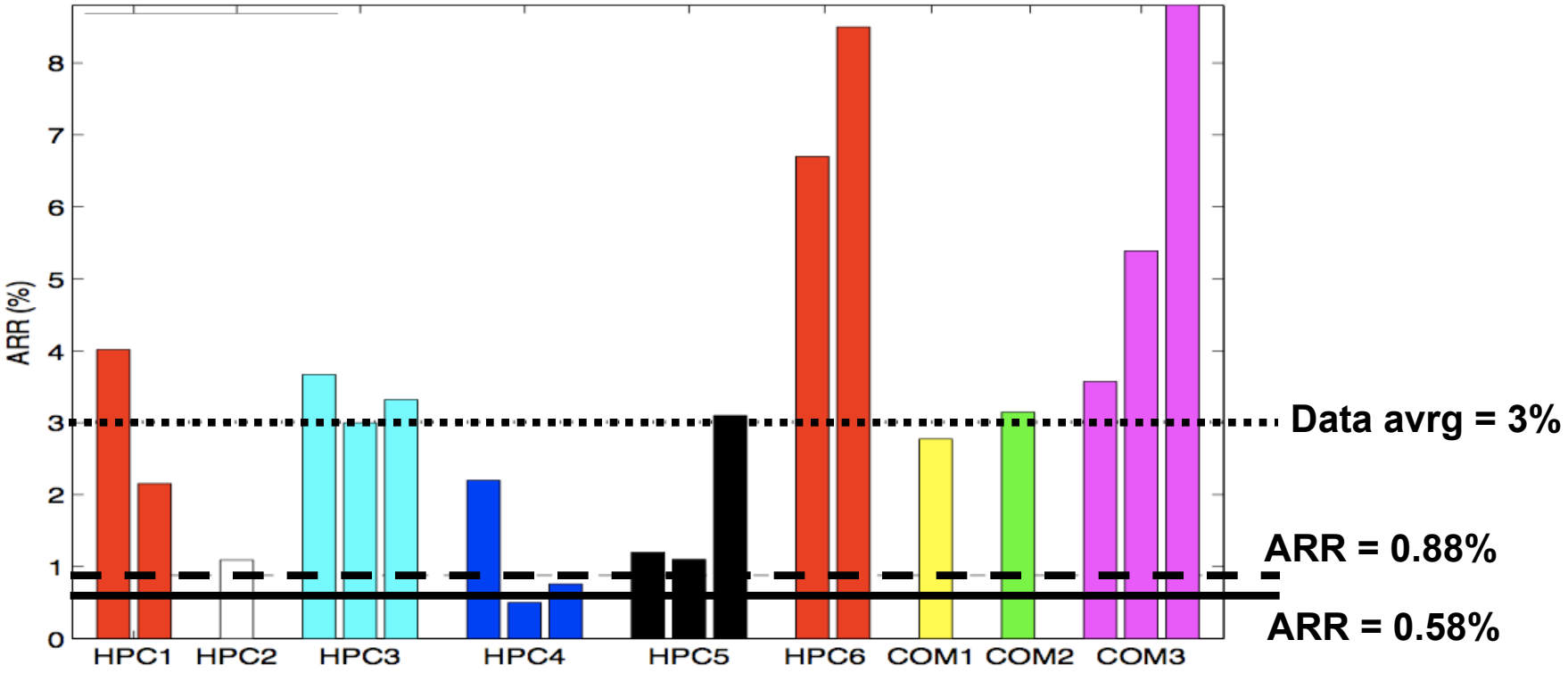
Storage suffers failures too

		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

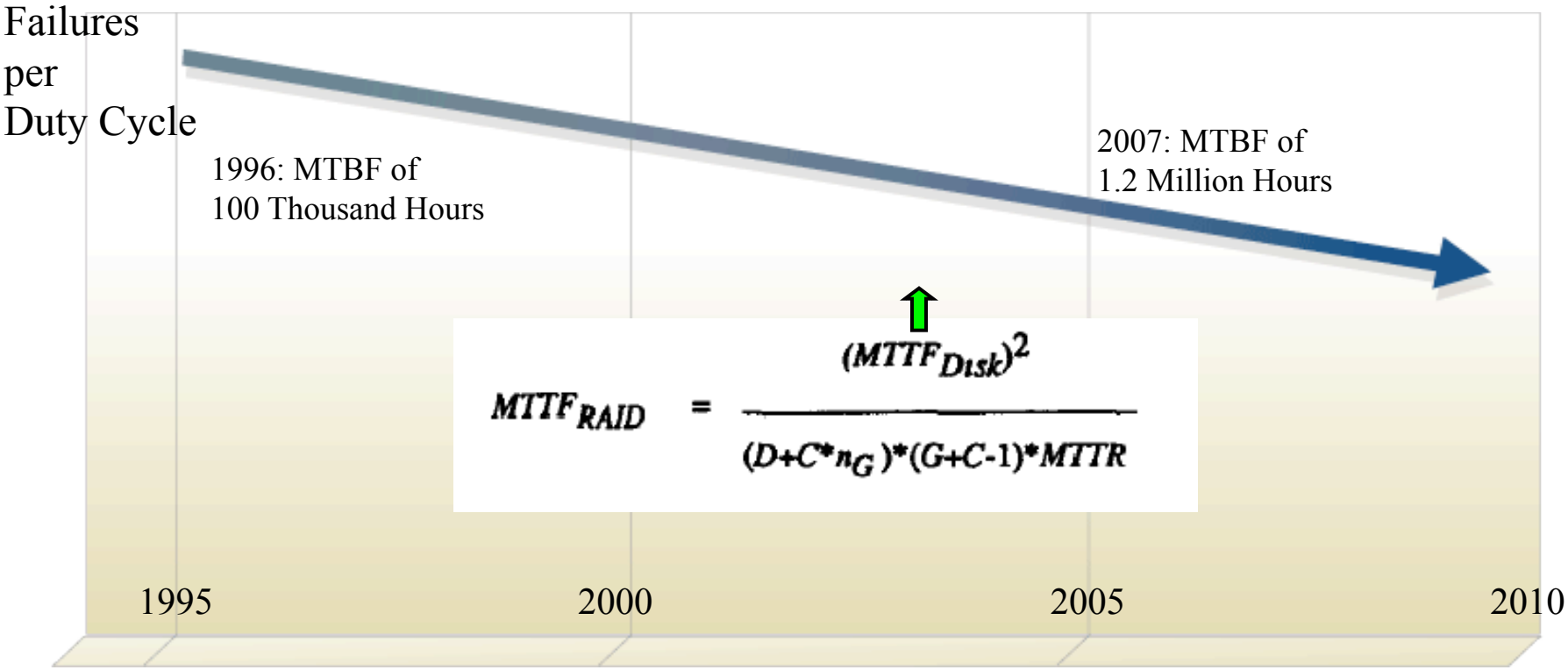
Annual disk replacement rate (ARR)

- Datasheet MTTFs are 1,000,000 to 1,500,000 hours.

=> Expected annual replacement rate (ARR): 0.58 - 0.88 %.



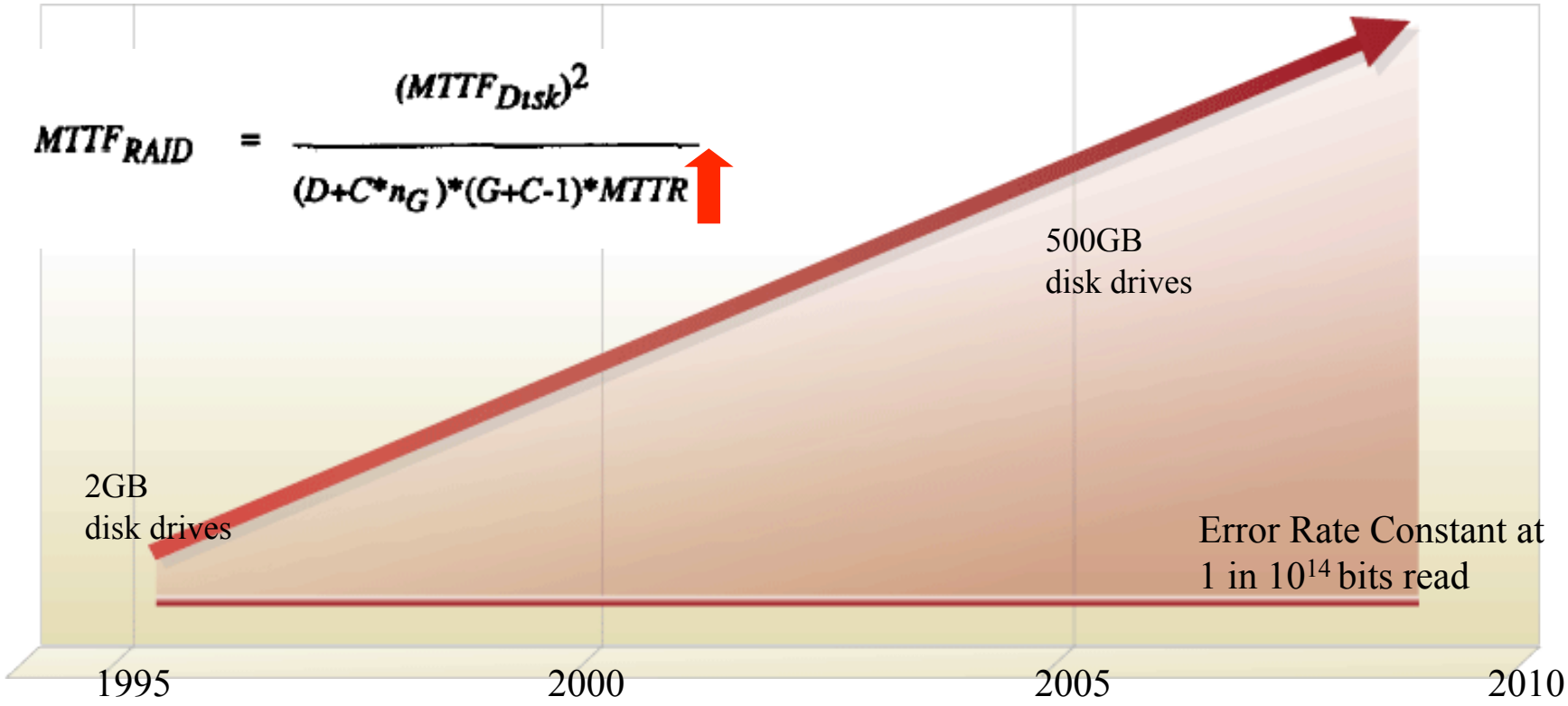
And yet disk reliability has improved



- RAID 5 protection sufficient when MTBF was 100,000 hours
- Today's disks are 10 times more reliable per mfr specs

RAID reliability should be 100X greater

But size of disk reconstruction grows faster

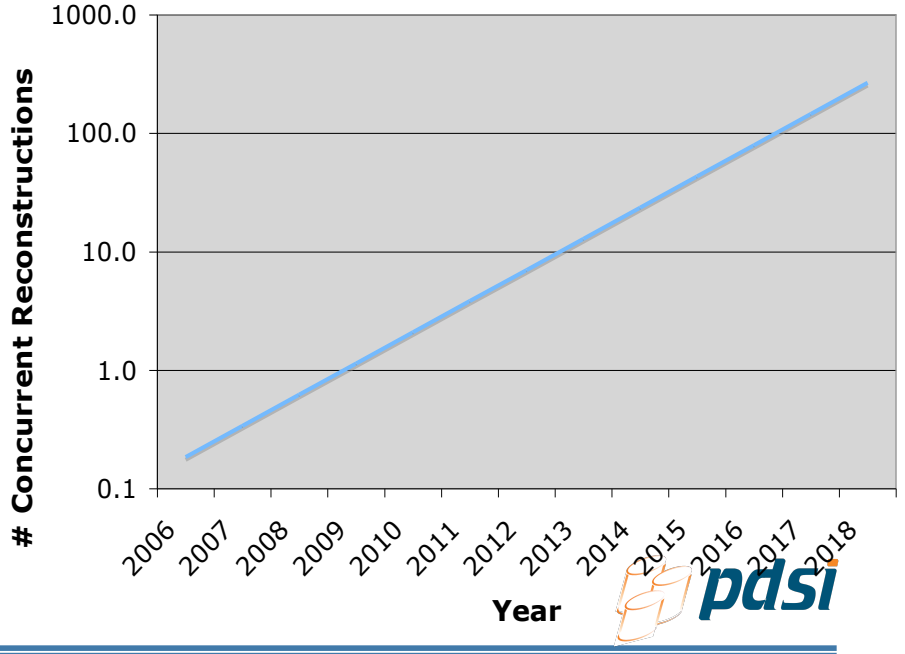


- Disks construction is 250 times more work, taking much more time
- Media errors proportional to reading, so longer reconstructions fail more

More Media = Longer, Riskier Reconstruction

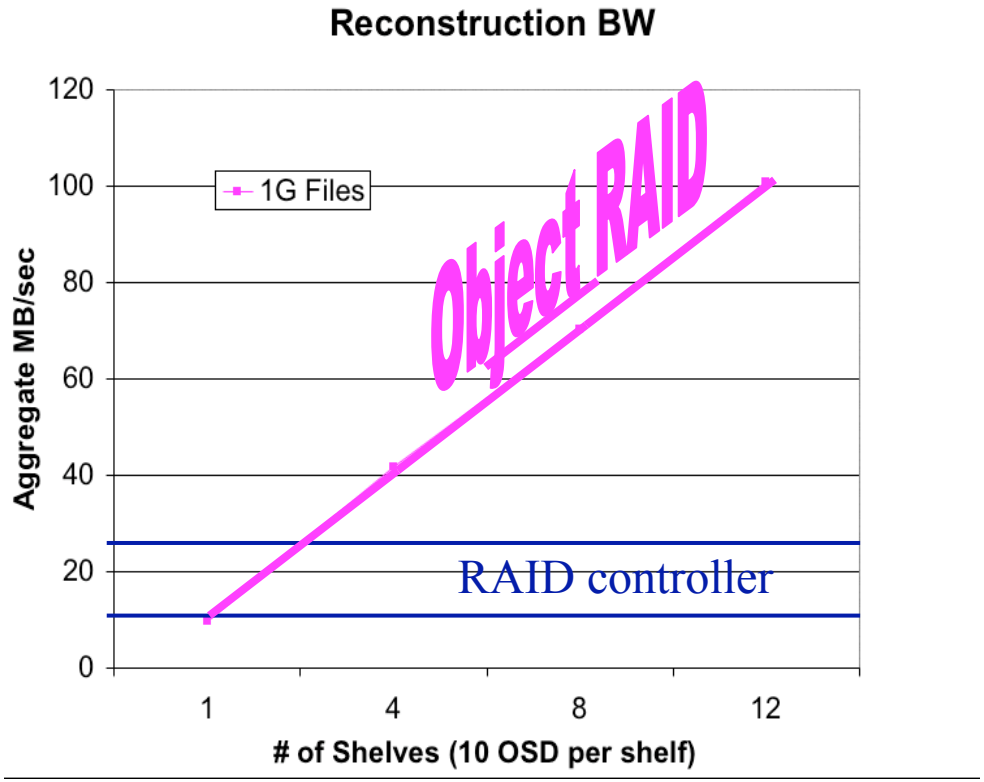
Projection of concurrent reconstructions

- Model numbers of disks needed to support Petascale fault tolerance, apply failure rates & disk size trends
  - Slower than Moore's law because of move to 2.5" disks
  - Traditional RAID controller reconstruction modeled
- Today reconstructing disks up to 10-20% of time
- Could be soon 100s of concurrent reconstructions!
- Storage does not have checkpoint/restart model
  - Can't forget last hour of data writes and restart
- Design normal case for many failures always



Need reconstruction in parallel

- Scaling disk failure repair: should speed up with size
  - Traditional RAID recovery speed constant with scale
    - Not massively parallel
  - Instead, decluster RAID sets
  - RAID work striped widely
    - All arms & managers used in recovery
  - More storage = faster repair
  - Plus, shorter degraded periods

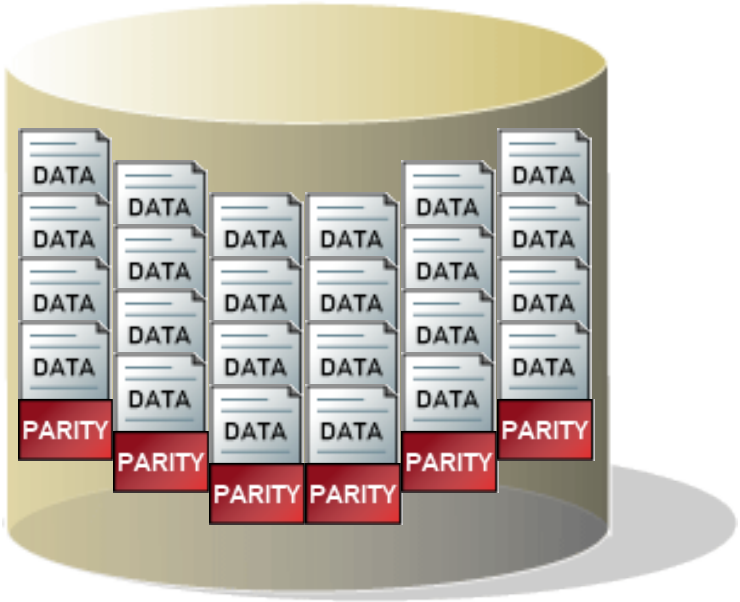


- ***Storage fault tolerance needs even more attention***

And point solutions for narrow problems

- Study media errors
- Devise per disk correcting codes to scale with disk size
  - Improves on internal ECC capabilities (limited by economics
- Independent of traditional cross disk parity schemes
- Avoids using double failed disk codes until double failures are the problem

Example: Panasas Vertical Parity



panasas

Closing

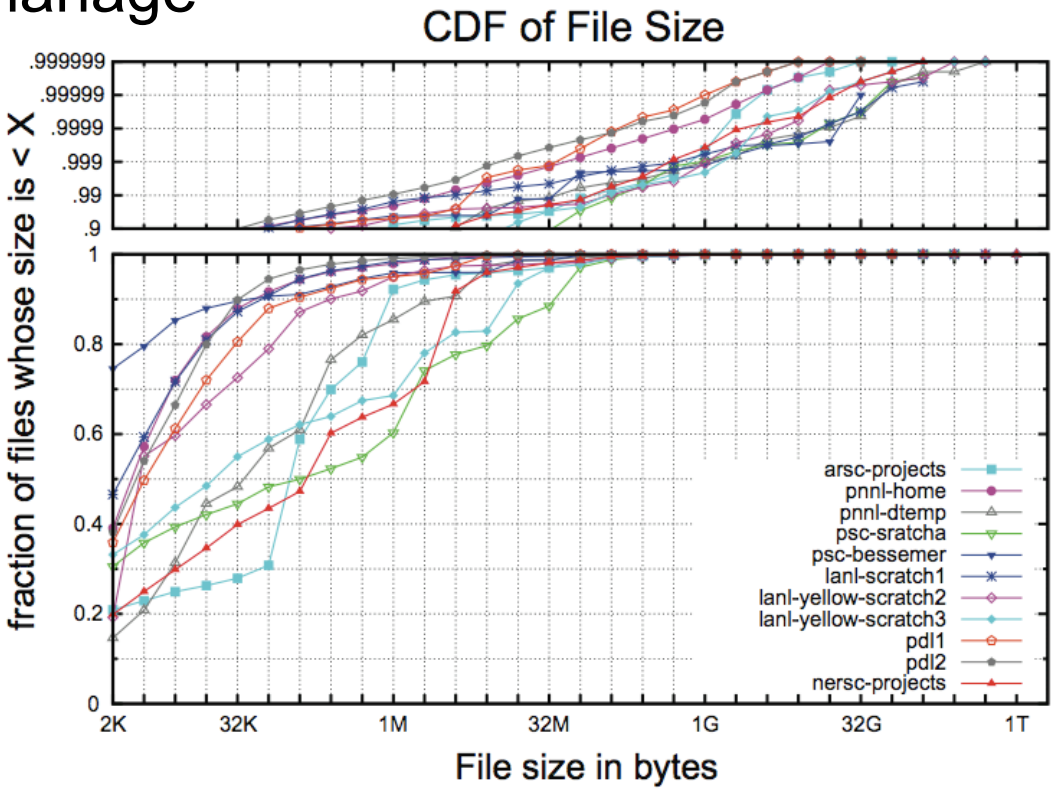
- Future parallel computing increasingly suffers failures
- ***Field data needs to be collected and shared***
  - ***cfdr.usenix.org, pdsi-scidac.org: please contribute!***
- Traditional fault tolerance needs to be revisited
  - Checkpointing needs new paradigms
- Systems need to be designed to operate in repair
  - Reconstruction must be parallel, faster in larger systems
  - Specific failures should be addressed with specific solutions

- Brent Welch, Garth Gibson, et. al., "Scalable Performance of the Panasas Parallel File System," USENIX Conference on File and Storage Technology (FAST), Feb. 2008.

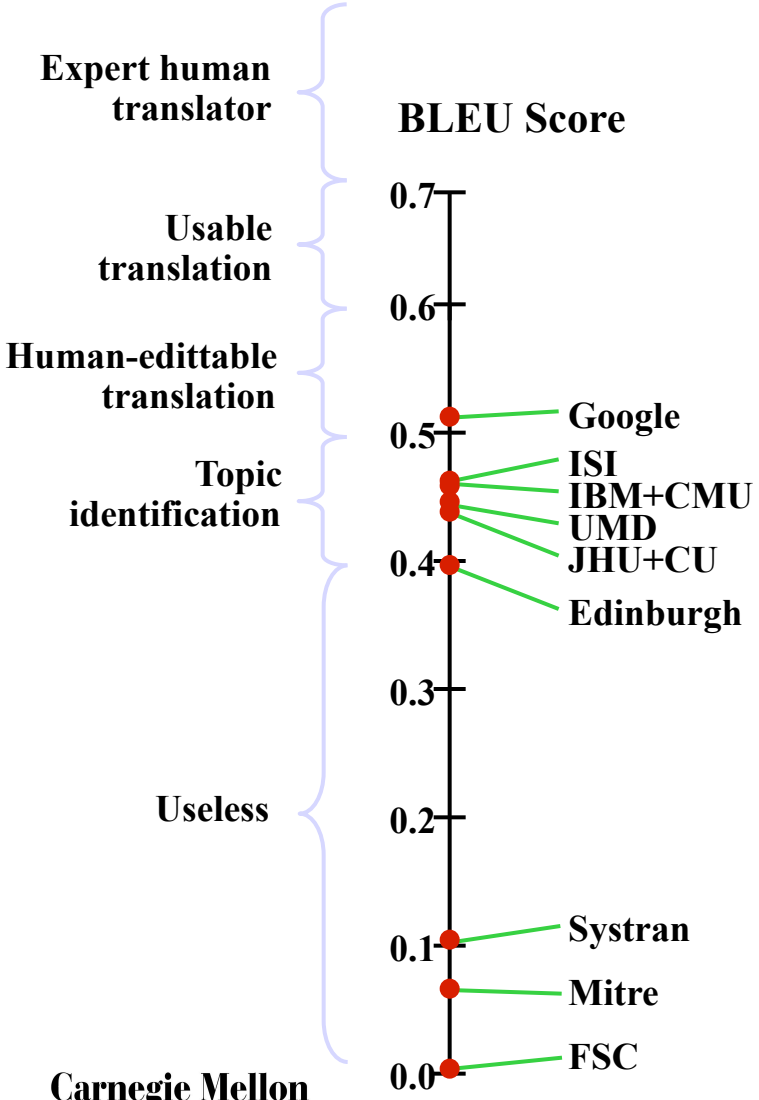
And it ain't all huge data files

- Study data distributions – millions of files – 20-80% tiny
  - Still majority of space in relatively few huge files (like checkpoints)
- Lots more metadata to manage

Label	Date (2008)	Type	File System	Total Size TB	Total Space TB	# files M	# dirs K
Satyanarayanan81	<1981	Home	TOPS10		<.0016	.086	
Irlam93	Nov 1993				.259	12	
SFS97	<1997		NFS				
Douceur99	Sept 98	Desktops	NTFS	10.5		141	
VU2005	2005	Home	UNIX			1.7	
SFS2008	<2008		NFS				
CMU gg1	4/10	OS	HFS+	.044	.046	1.0	258
CMU gg2	4/10	Home	HFS+	.0098	.0099	.028	3.2
CMU gg3	4/10	Media	HFS+	.065	.066	.042	2.6
CMU pdl1	4/9	Project	WAFL	3.93	3.68	11.3	821
CMU pdl2	4/9	Project	WAFL	1.28	1.09	8.11	694
NERSC	4/8	Project	GPFS	107	107	20.5	917
PNNL nwfs	3/17	Archival	Lustre	265	264	13.7	1824
PNNL home	3/17	Home	ADVFS	4.7	4.3	10.1	682
PNNL dtemp	3/17	Scratch	Lustre	22.5	19.2	2.2	51
PCS scratch	3/27	Scratch	Lustre	32	32	2.07	451
PSC bessemer	3/27	Project	Lustre	3.7	3.7	0.38	15
LANL scratch1	4/1	Scratch	PanFS	9.2	10.7	1.52	120
LANL scratch2	4/10	Scratch	PanFS	25	26	3.30	241
LANL scratch3	4/10	Scratch	PanFS	26	29	2.58	374
ARSC seau1	3/13	Archival	SAM-QFS	305	4.3	10.5	326
ARSC seau2	3/14	Archival	SAM-QFS	115	4.6	5.3	116
ARSC nanu1	3/12	Archival	SAM-QFS	69	4.5	6.7	338
ARSC projects	3/13	Archival	SAM-QFS	32	.93	6.2	898



Data intensive computing has many forms



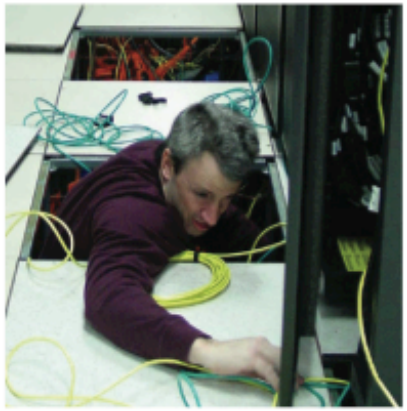
- NIST translate 100 articles
  - Arabic-English competition
- 2005 outcome: Google wins!
- Qualitatively better on 1st entry
- Not most sophisticated approach
  - Brute force statistics
- But more data & compute !!
  - 200M words from UN translations
  - 1 trillion words of English grammar
  - 1000 processor cluster
- Science of all types going to scale
  - Can't do the best science without it

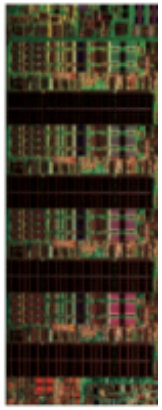
Carnegie Mellon
Parallel Data Laboratory

pdsi

The PetaFLOPS era is here

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.