SciDAC
Scientific Discovery through Advanced Computing

ASCR

Office of Science

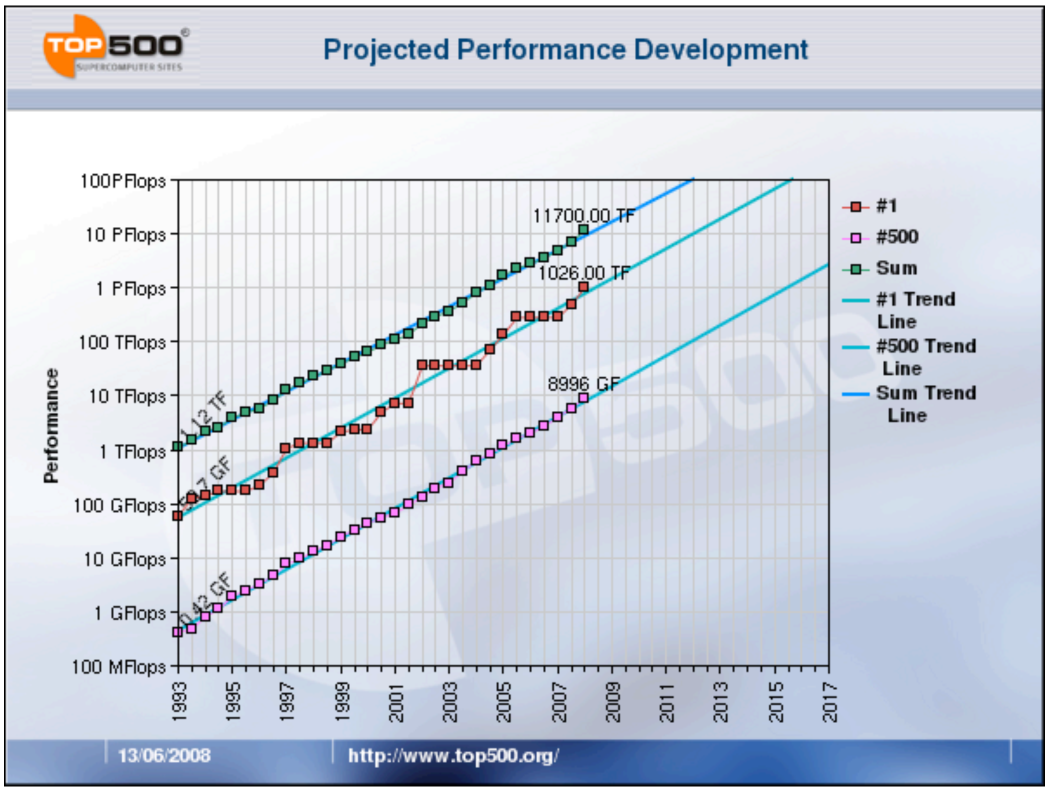
# Exa & Yotta Scale Data

## SC’08 Panel November 21 2008, Austin, TX

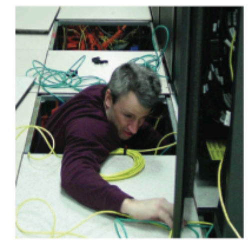
Garth Gibson
Carnegie Mellon University and Panasas Inc.
SciDAC Petascale Data Storage Institute (PDSI)
www.pdsi-scidac.org

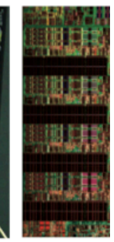
# Charting the Path thru Exa- to Yotta-scale

- Top500.org scaling 100%/yr; Exa in 2018, Zetta in 2028, Yotta in 2038
  - Hard to make engineering predictions out 10 years, but 30 years?




Los Alamos
NATIONAL LABORATORY
EST. 1943





## Roadrunner

### First to break the "petaflop" barrier

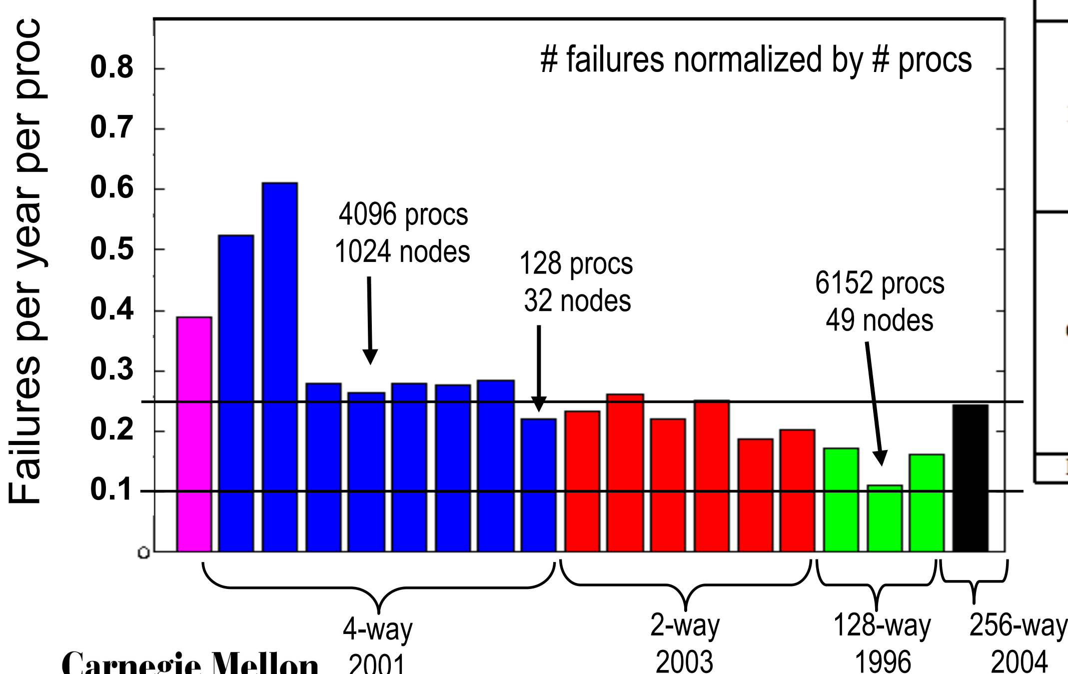
At 3:30 a.m. on May 26, 2008, Memorial Day, the "Roadrunner" supercomputer exceeded a sustained speed of 1 petaflop/s, or 1 million billion calculations per second. The sustained performance makes Roadrunner more than twice as fast as the current number 1 system on the TOP500 list. The best sustained performance to date is 74.5% efficiency, 1.026 petaflop/s.

# Storage Scaling

- Trends are quoted in capacity & performance
- Balance calls for linear scaling with FLOPS
- Disk capacity grows near Moore's Law
  - Disk capacity track compute speed
  - Parallelism grows no better or worse than compute
- But disk bandwidth +20%/yr < Moore's Law
  - Parallelism for BW grows faster than compute!
  - Revisit reason for BW balance: fault tolerance
- And random access? +7%/yr is nearly no growth
  - Coupled with BW parallelism, good growth
  - But new workloads, analytics, more access intensive
  - Solid state storage looks all but inevitable here

# Fault Data & Trends

- Los Alamos root cause logs
  - 22 clusters & 5,000 nodes
  - covers 9 years & continues
  - cfdr.usenix.org publication + PNNL, NERSC, Sandia, PSC, …



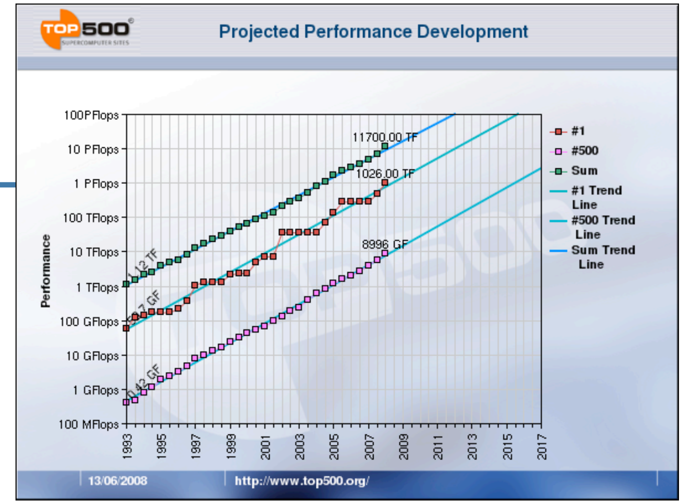

| (I) High-level system information | | | | (II) Information per node category | | | |
|---|---|---|---|---|---|---|---|
| HW | ID | Nodes | Procs | Procs /node | Production Time | Mem (GB) | NICs |
| A | 1 | 1 | 8 | 8 | N/A – 12/99 | 16 | 0 |
| B | 2 | 1 | 32 | 32 | N/A – 12/03 | 8 | 1 |
| C | 3 | 1 | 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 |

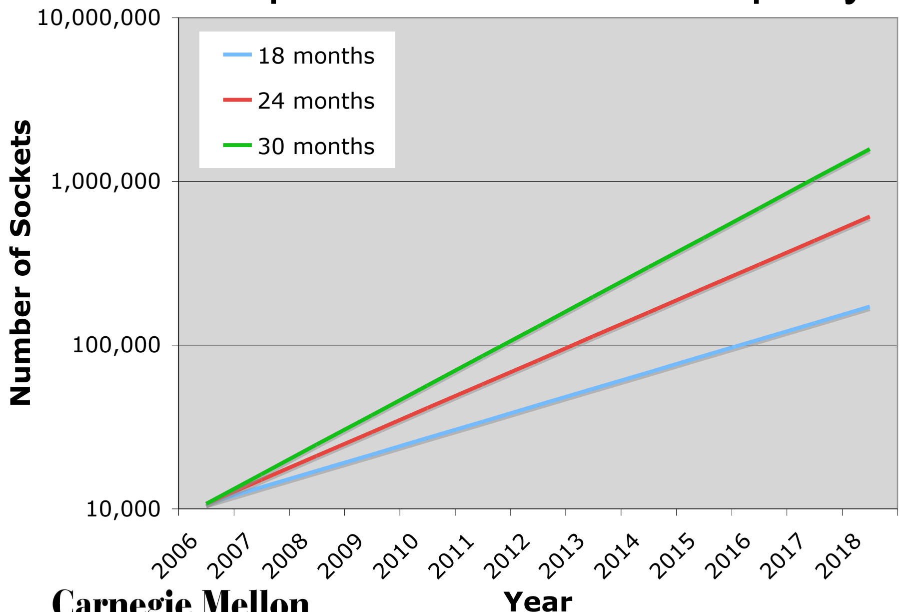
**Table 1.** *Overview o*
*SMP-based, and systen*

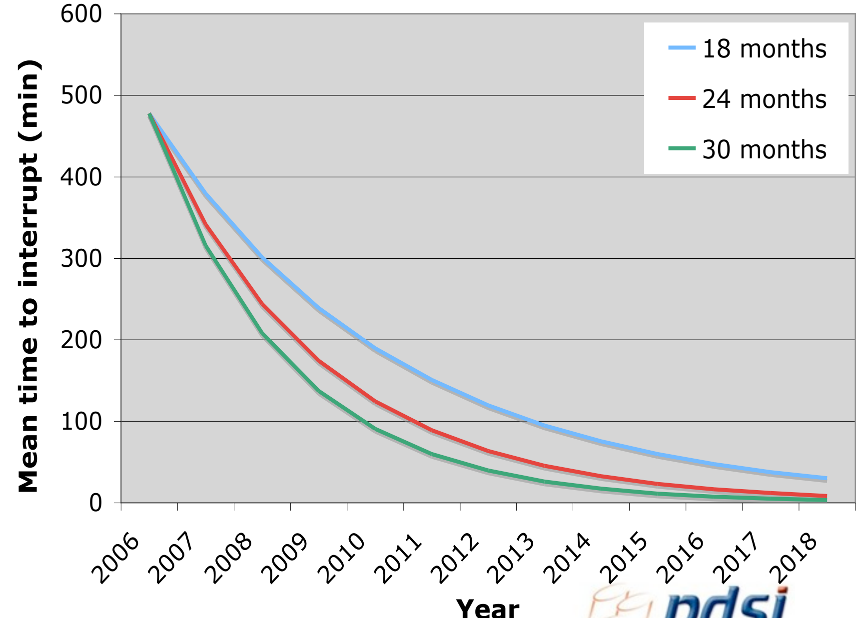
Los Alamos
NATIONAL LABORATORY
EST. 1943

# Projections: More Failures

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



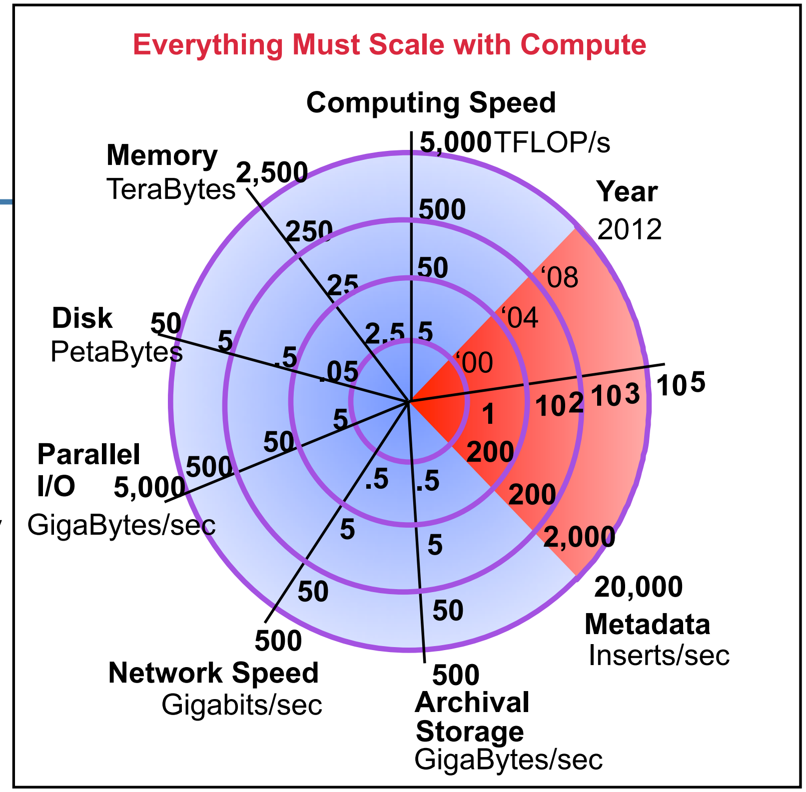



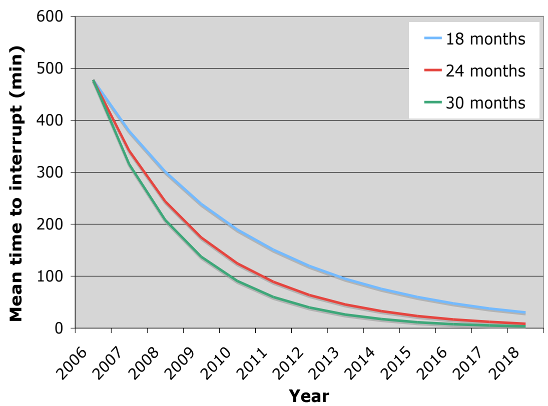



# Fault Tolerance Challenge

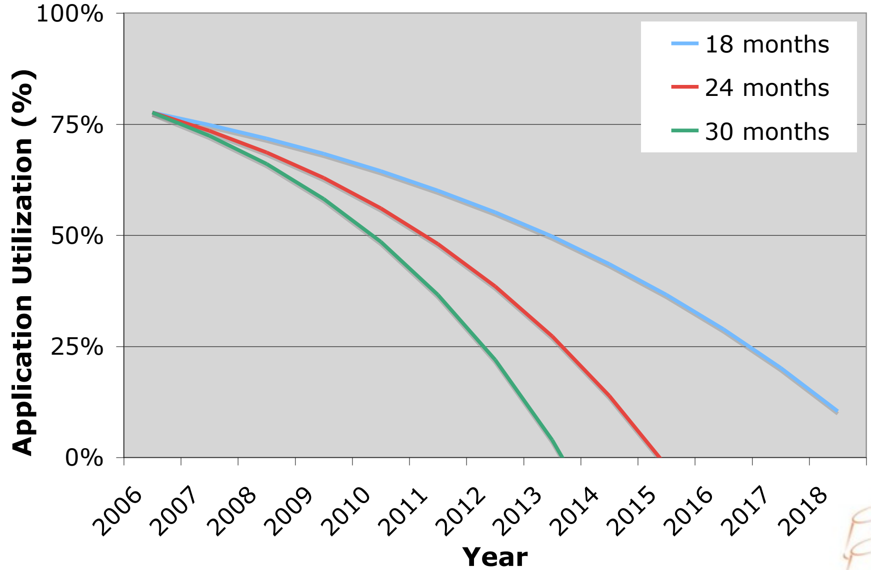
- Periodic (p) pause to checkpoint (t)
  - Major need for storage bandwidth
- Balanced systems
  - Storage speed tracks FLOPS, memory so checkpoint capture (t) is constant
  - 1 – AppUtilization = t/p + p/(2*MTTI)

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



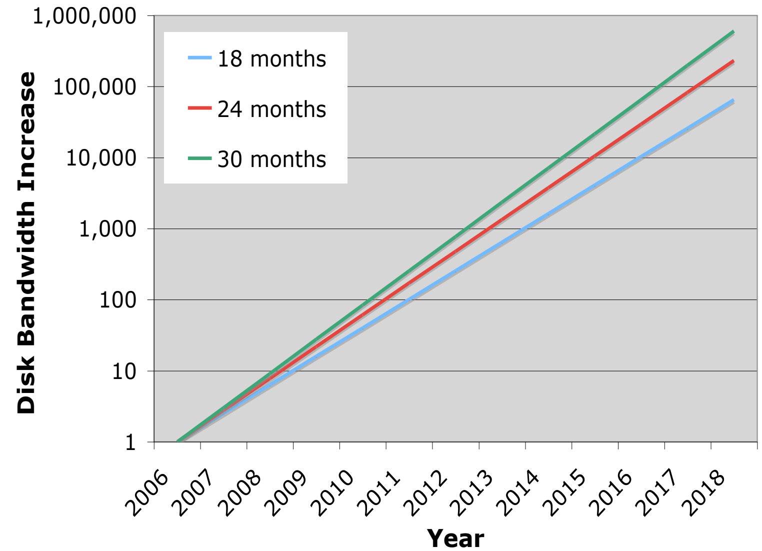



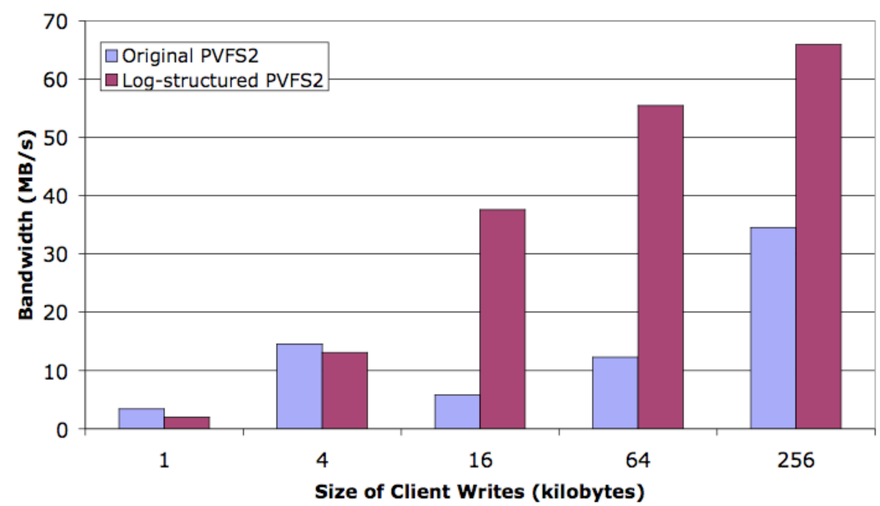

- ***but dropping MTTI kills app utilization!***



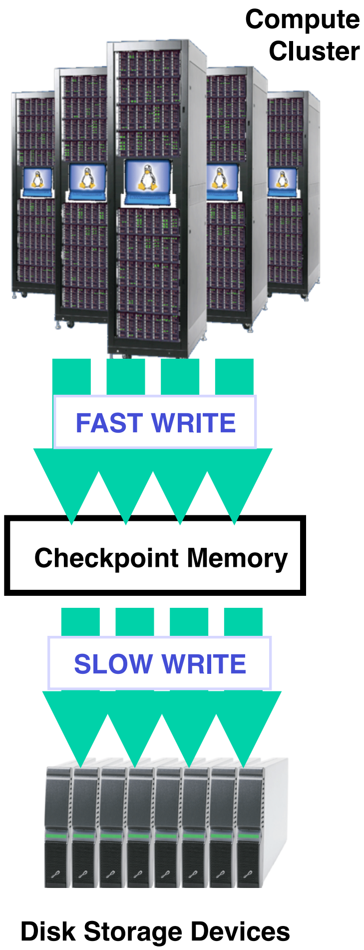
# Fault Tolerance Drives Bandwidth

- More storage bandwidth?
  - disk speed 1.2X/yr
    - – # disks +67%/y just for balance !
  - to also counter MTTI
    - – # disks +130%/yr !
  - Little appetite for the cost
- N-1 checkpoints hurt BW
  - Concurrent strided write
  - Will fix with internal file structure: write optimized
  - See Zest, ADIOS, ….

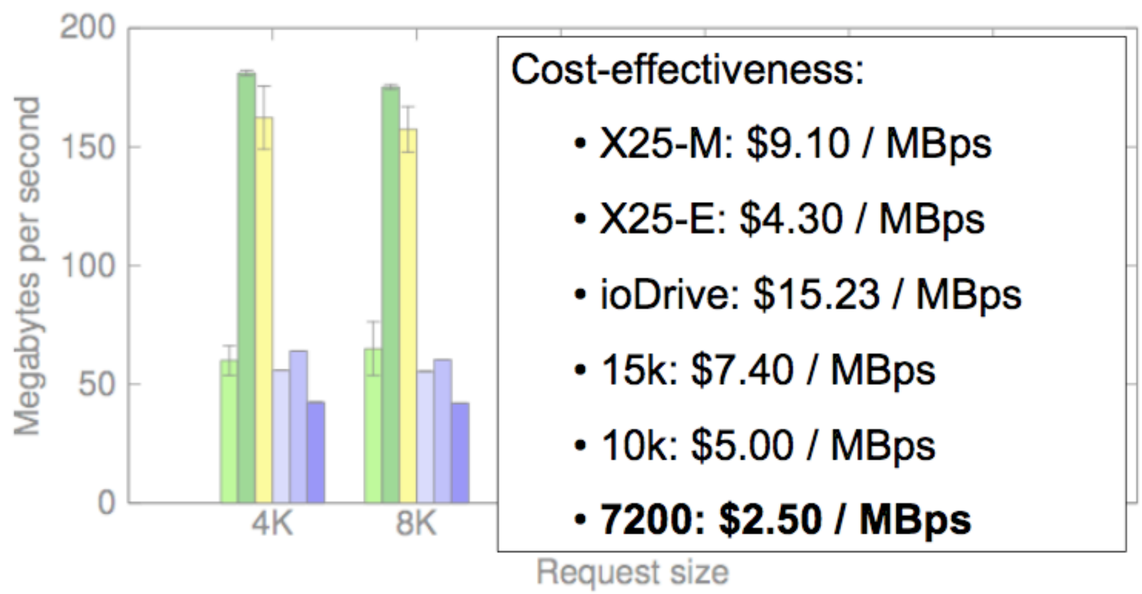


# Alternative: Specialize Checkpoints

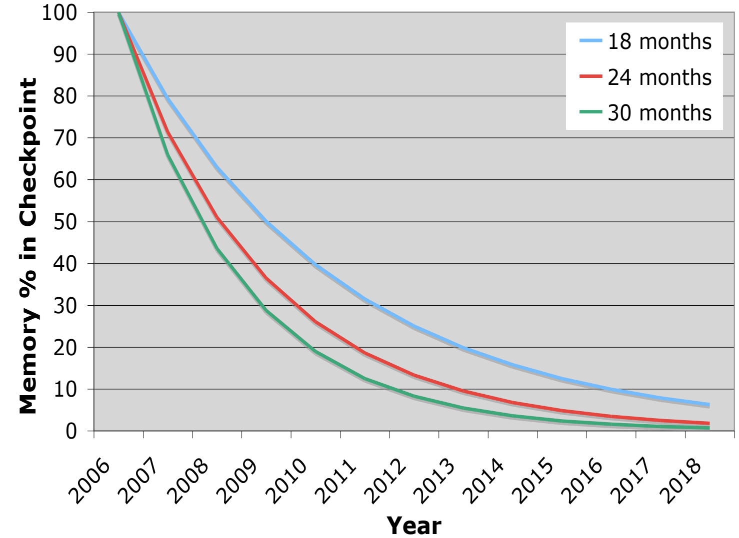



- Dedicated checkpoint device (ie., PSC Zest)
  - Stage checkpoint through fast memory
  - Cost of dedicated memory large fraction of total
  - Cheaper SSD (flash?) now bandwidth limited
  - There is hope: 1 flash chip == 1 disk BW …..

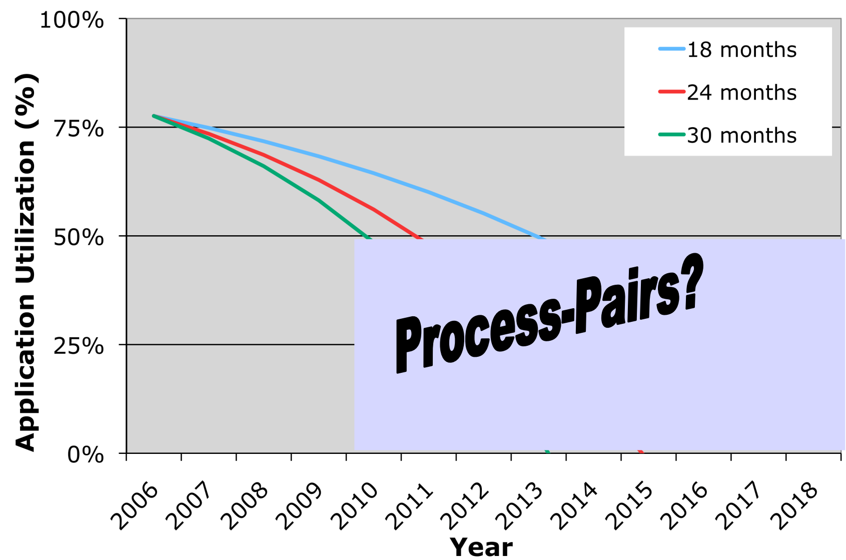


# Application Level Alternatives

- Compress checkpoints!
  - plenty of cycles available
  - smaller fraction of memory each year (application specific)
    – 25-50% smaller per year




- Classic enterprise answer: process pairs duplication
  - Flat 50% efficiency cost, plus message duplication



# Storage Suffers Failures Too

| | | Type of drive | Count | Duration |
|---|---|---|---|---|
| Pittsburgh Supercomputing Center | HPC1 | 18GB 10K RPM SCSI<br>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<br>15K RPM SCSI<br>7.2K RPM SATA | 14,208 | 1 yr |
| Various HPCs | HPC4 | 250GB SATA<br>500GB SATA<br>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<br>10K RPM FC-AL<br>10K RPM FC-AL<br>10K RPM FC-AL | 3,700 | 1 yr |

# Storage Failure Recovery is On-the-fly

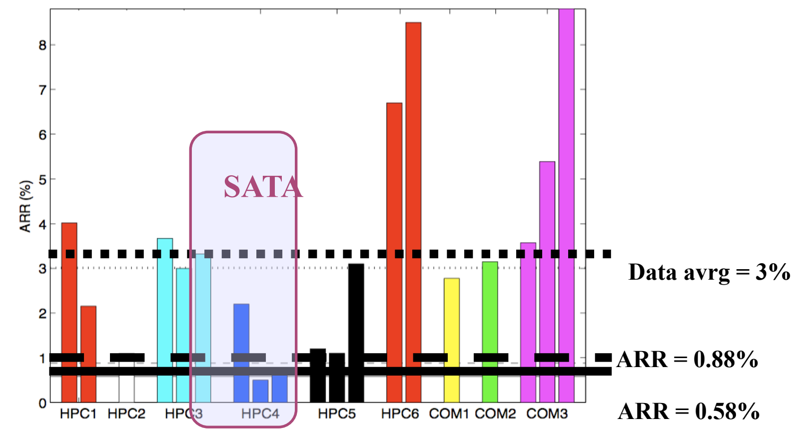
- Scalable performance = more disks
- But disks are getting bigger
- Recovery per failure increasing
- Hours to days on disk arrays
- Consider # concurrent disk recoveries

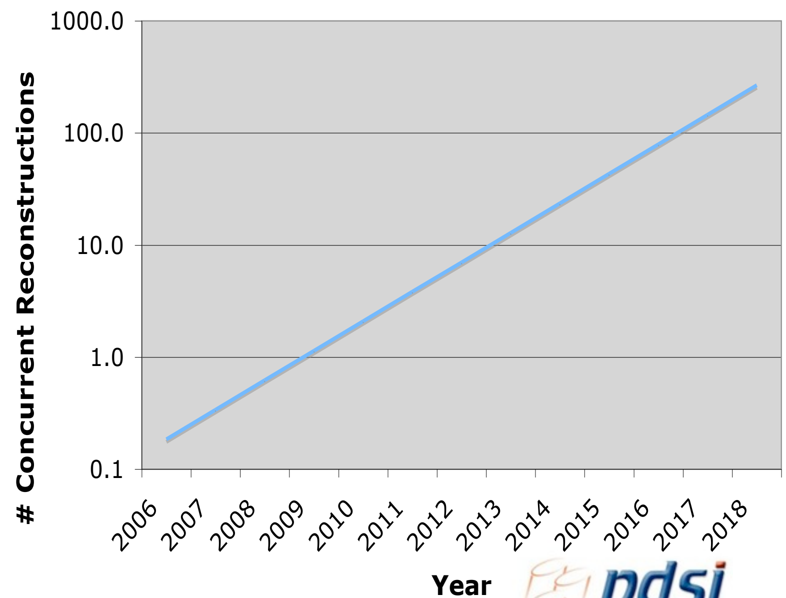
  e.g. 10,000 disks

  3% per year replacement rate

  1+ day recovery each

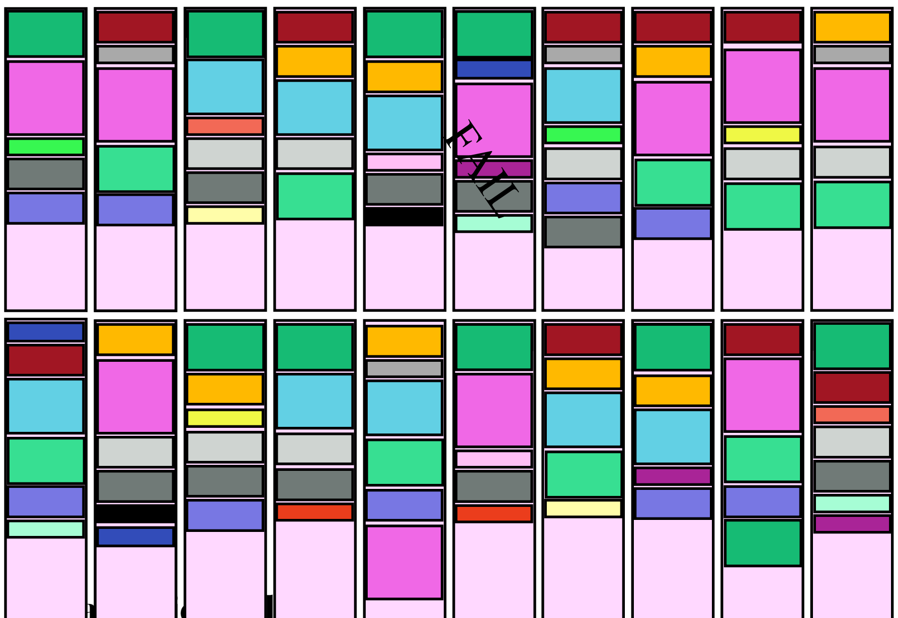
  Constant state of recovering ?
- Maybe soon 100s of
  concurrent recoveries (at all times!)
- Design normal case
  for many failures (huge challenge!)

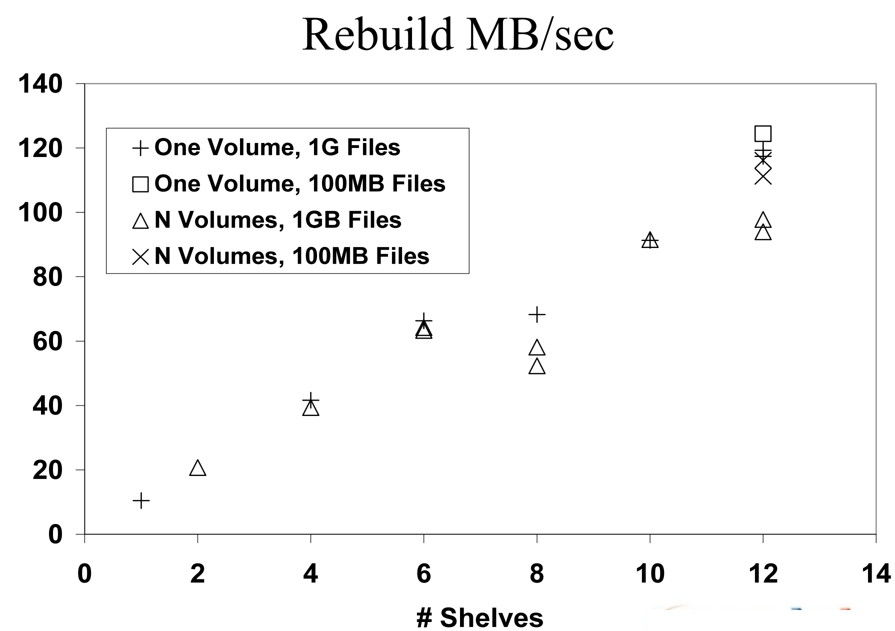


# Parallel Scalable Repair

- Defer the problem by making failed disk repair a parallel app
- File replication and, more recently, object RAID can scale repair
  - "decluster" redundancy groups over all disks (mirror or RAID)
  - use all disks for every repair, faster is less vulnerable
- Object (chunk of a file) storage architecture dominating at scale
  PanFS, Lustre, PVFS, … GFS, HDFS, … Centera, …



# Scaling Exa- to Yotta-Scale

- Exascale capacity parallelism not worse than compute parallelism
  - But internal fault tolerance harder for storage than compute
- Exascale bandwidth a big problem, but dominated by checkpoint
  - Specialize checkpoint solutions to reduce stress
  - Log-structured files, dedicated devices, Flash memory …..
  - Application alternatives: state compression, process pairs
- Long term: 20%/yr bandwidth growth serious concern
  - Primary problem is economic: what is value of data vs compute?
- Long term: 7%/yr access rate growth threatens market size
  - Solid state will replace disk for small random access

# SciDAC Petascale Data Storage Institute

- High Performance Storage Expertise & Experience
  - Carnegie Mellon University, Garth Gibson, lead PI
  - U. of California, Santa Cruz, Darrell Long
  - U. of Michigan, Ann Arbor, Peter Honeyman
  - Lawrence Berkeley National Lab, William Kramer
  - Oak Ridge National Lab, Phil Roth
  - Pacific Northwest National Lab, Evan Felix
  - Los Alamos National Lab, Gary Grider
  - Sandia National Lab, Lee Ward

www.pdsi-scidac.org