Reliability and Availability at Scale

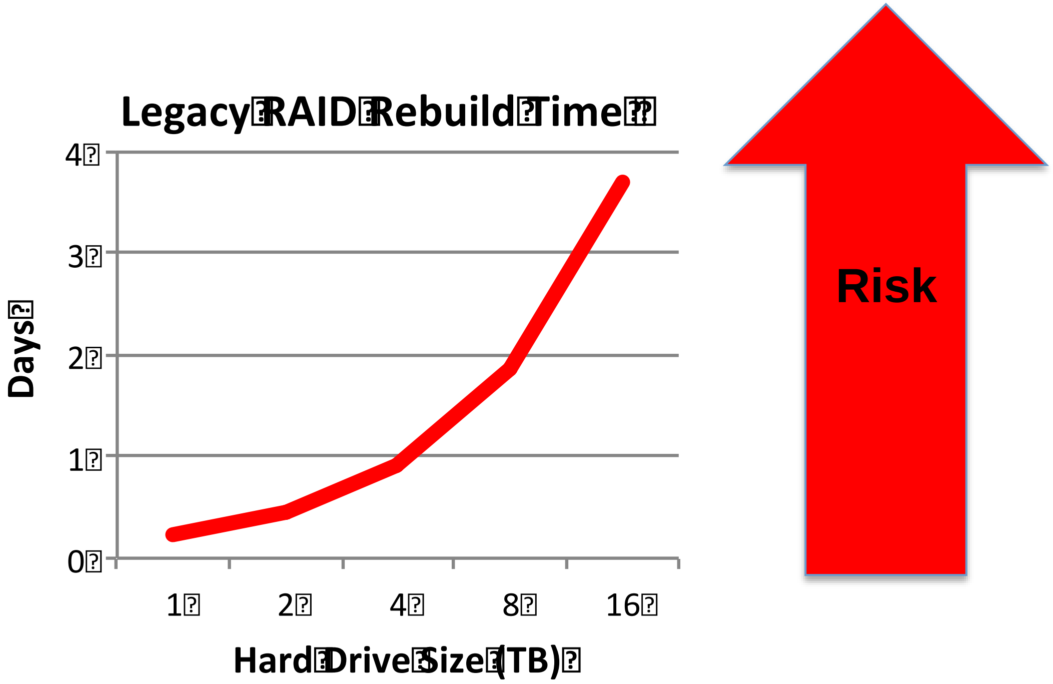
MAKING THE ODDS WORK IN YOUR FAVOR

IDC HPC USER FORUM – SEPT. 16, 2014

PANASAS PRODUCT MARKETING

HARDWARE RAID REACHING ITS LIMITS

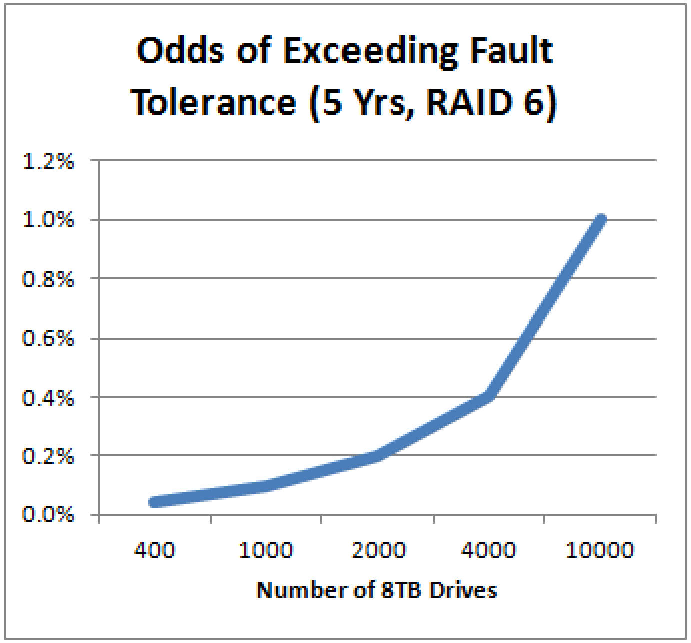
- **Large Deployments Exacerbate Existing Vulnerabilities in Traditional Data Protection Schemes**
  - Reliability gets worse with scale
  - Slow rebuild times
  - Lengthy disaster recovery
  - Unnecessary availability outages



At 50MB/s RAID rebuild rate

HARDWARE RAID 6 THEORETICAL

- All hardware RAID volumes risk exceeding fault tolerance
- 100 hardware RAID volumes = 100x the risk
- What are the odds for a typical hardware RAID 6 system?

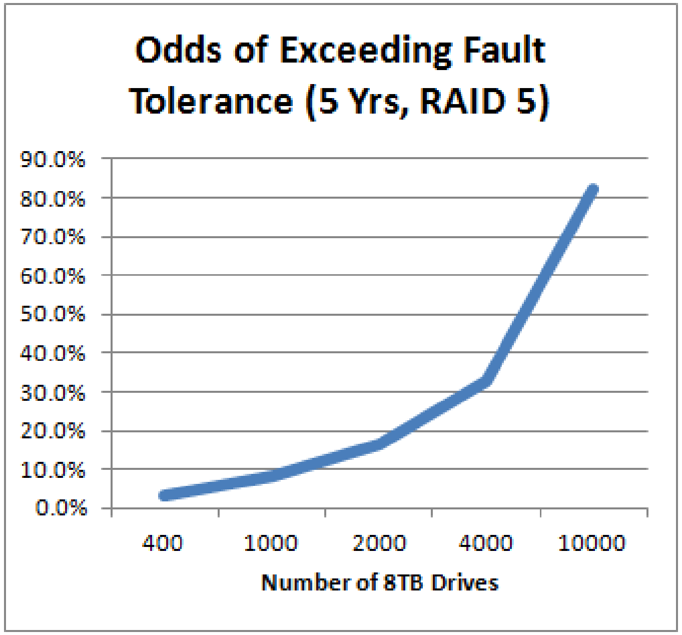


Assumptions:
- 8TB drives
- 10 drive RAID 6 stripes
- 50MB/s rebuild rate
- 3% drive AFR

- This may appear ok, but there’s a problem here…

HARDWARE RAID 6 ACTUAL

- **Previous graph assumes RAID 6 rebuilds always complete**
- **Latent Sector Errors = increasingly a big problem**
  - HDD vendors: 1 in 10^15 to 10^16 sectors
  - U Wisc/NetApp study (2007) of 1.5m HDDs: 3.45% of drives had LSE's, >60% found by data scrubbing, LSE rate increases with time and size of drive
  - Panasas: vertical parity prevented rebuilds on ~7% of deployed drives
- **LSEs in hardware RAID-based approaches can lower actual RAID 6 reliability almost to theoretical RAID 5 levels**



Assumptions:
- 8TB drives
- 10 drive RAID 5 stripes
- 50MB/s rebuild rate
- 3% drive AFR

Source: http://research.cs.wisc.edu/wind/Publications/latent-sigmetrics07.pdf

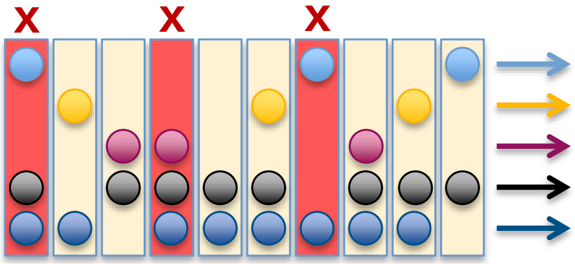
DELIVERING RELIABILITY AT SCALE

- **Replace hardware RAID with software-based, per-file RAID using erasure coding**
- **Protect files (stripes of files), not entire block devices**
- **Limit rebuilds to affected files, not entire drives**
  - Don’t rebuild portions of drives that are ok
  - Don’t rebuild empty space
- **Provide additional parity protection against Latent Sector Errors**
  - And keep background scrubbing which is effective
- **Distribute data on stripes selected from all drives in system**
  - RAID rebuild performance scales linearly
  - Data reliability can increase with system scale instead of decreasing

UNDERSTANDING PER-FILE RAID

Per-file Distribution Reduces Risk at Scale

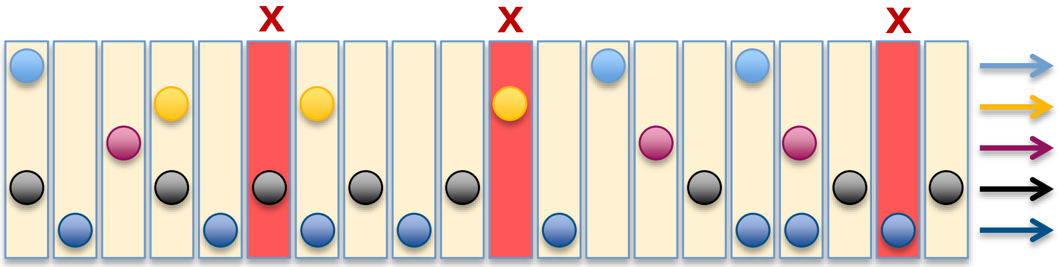
- Small files are triple mirrored, large files are striped
- With more and more drives, three drive failures (exceeding fault tolerance) are less and less likely to affect any given file



File 1 (small): down 2, rebuild mirror
File 2 (small): unaffected
File 3 (small): down 1, rebuild mirror
File 4 (large): down 2, RAID 6 rebuild
File 5 (large): **down 3, file damaged!**

One file damaged;
Only need to restore File 5

Files distributed over ten drives **X = DRIVE FAILURES**



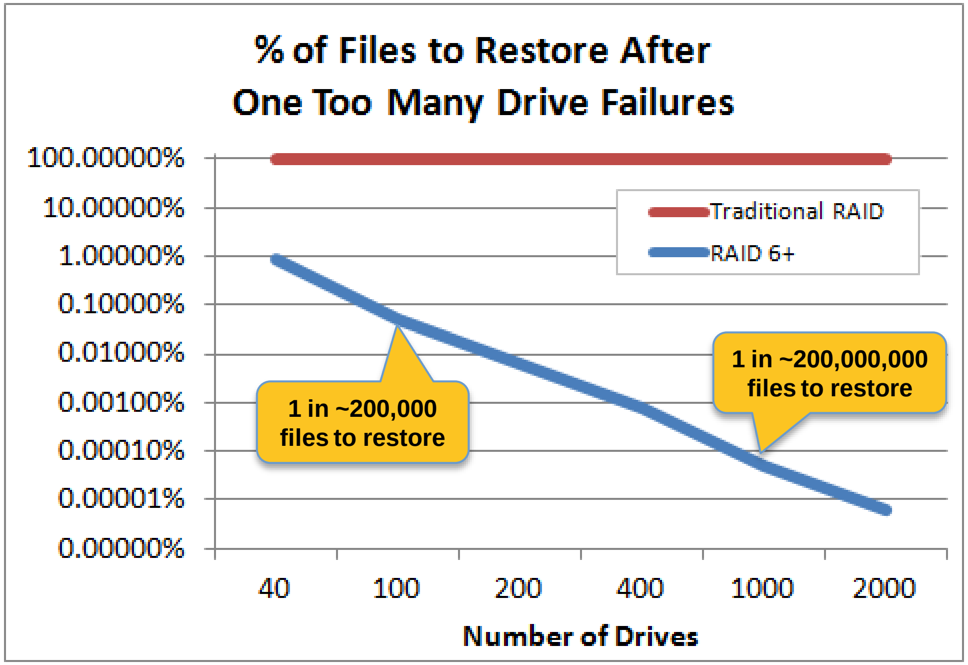
File 1 (small): unaffected
File 2 (small): down 1, rebuild mirror
File 3 (small): unaffected
File 4 (large): down 1, RAID 5 rebuild
File 5 (large): down 1, RAID 5 rebuild

Same files distributed over 20 drives **X = DRIVE FAILURES**

No files damaged;
Can rebuild all data

UNMATCHED DISASTER RECOVERY

- **Fast Time to Restore**
  - Restore specific files instead of entire file system
  - Made possible by extra protection of namespace (directory data) in RAID 6+
- **Percentage of Files to Restore Approaches Zero with Scale**
  - With RAID 6+ (66% small files), a triple simultaneous disk failure means:



Scaling by 10x increases reliability by 1000x!

DELIVERING AVAILABILITY AT SCALE

- **Current availability model for storage is a problem at scale**
  - System goes offline upon exceeding fault tolerance anywhere in system
  - Availability needs to be more granular
- **Instead architect for “Always On”**
  - File system remains available even after exceeding fault tolerance
  - Protect directory structure deeper than data so directory structure stays navigable and all unaffected files can be accessed normally
  - Make it easy to quickly restore damaged files if possible

PANASAS ACTIVESTOR WITH PANFS 6.0

- **ActiveStor 16 with PanFS 6.0: no-compromise hybrid scale-out NAS**
- **Data reliability increases with scale instead of decreasing**
  - RAID 6+ triple parity protection based on erasure codes in software – 150x improvement over dual parity and no hardware RAID controllers
  - New availability model keeps file systems online, even after “one too many drive failures”
- **For more, please visit: http://www.panasas.com**

ActiveStor 16

10 shelves, 1.2PB

THANK YOU!

http://www.panasas.com

http://www.linkedin.com/company/panasas

http://twitter.com/#!/panasas

http://www.youtube.com/PanasasHPC