Reliability and Availability at Scale

MAKING THE ODDS WORK IN YOUR FAVOR
**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*
- 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 (Lustre)

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 rebuilds always complete
- Latent Sector Errors are becoming a big problem
  - Hard drive vendors say: 1 in $10^{15}$ sectors
  - Large-scale studies say they’re much more common: 1.9% of enterprise drives had LSE’s, even with background disk scrubbing – probably a higher % today…
- LSE’s typically count as a drive failure for hardware RAID, bringing actual RAID 6 reliability much closer to theoretical RAID 5 levels:

DELIVERING RELIABILITY AT SCALE

- Replace hardware RAID with software-based per-file RAID using erasure coding
- Protect 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 to protect against Latent Sector Errors
- 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

---

![Diagram](image)

Files distributed over ten drives  X = DRIVE FAILURES

- 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

---

Same files distributed over 20 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

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:

  ![Graph showing reliability improvement with RAID 6+](image)

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

http://www.panasas.com

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

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

http://www.youtube.com/PanasasHPC