Reliability Modeling for Large Scale Declustered Storage

Lin Xiao, Garth Gibson

Motivation

More disks in large scale systems

- 1988 RAID paper: 10 disks/array, maybe 100
- 2009 OpenCloud: 256 disks, M45: 1920 disks

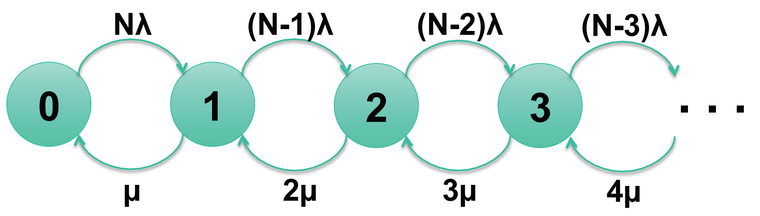
Mean time to data loss is not everything

- Clustered system: lose whole disks and whole array
- Declustered system: partial data loss
  - MTTDL = 1 month, data loss = 1B or 1 file
  - MTTDL = 1 year, data loss = 1Tb or 1000 files
- Declustered system with double correcting code
  - Three failures ≠ data loss
  - Many three-failures will not have any RAID sets in common

Reliability Modeling

Markov Model

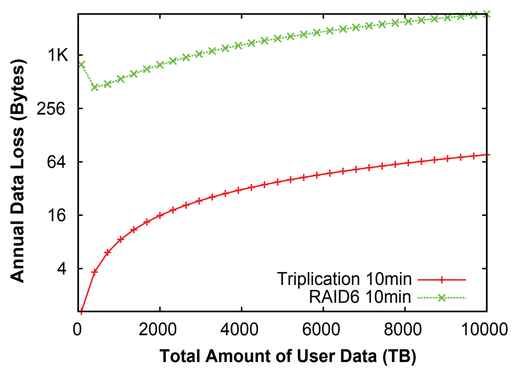
- Time to fail/repair is assumed to be exponentially distributed
- State i means there are i disk failures in the system
- λ = disk failure rate μ = disk repair rate
- N = total number of disks in the system
- Expected data loss with each failure



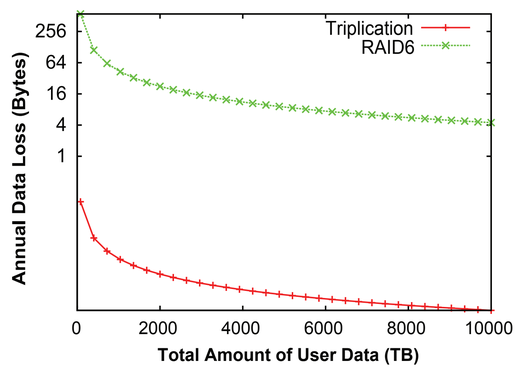
Triplication Has a Lower Data Loss Rate

1TB/disk 80% full, 64MB/chunk, 8+2, 25MB/s/disk repair

10 minute failure detection



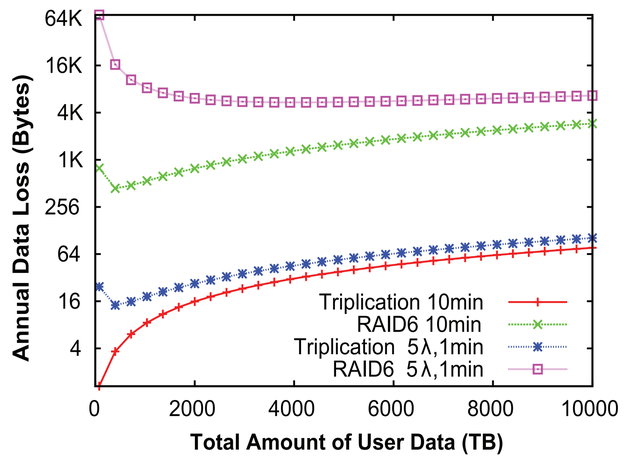
instant failure detection



- With the size of systems increasing, expected annual data loss increases with non-zero detection time and decreases with zero detection time (because repair is linearly faster in larger arrays)..

Faster Detection is NOT Free

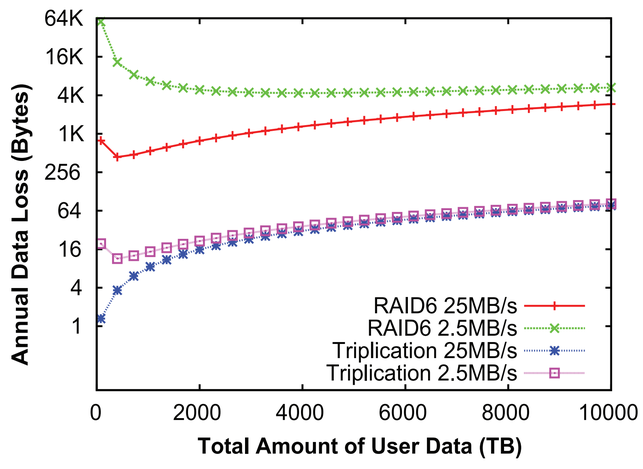
1TB/disk 80% full, 64MB/chunk, 8+2, 25MB/s/disk repair



- Google [Ford10] : More than 90% of unavailable events are shorter than 10 minutes
- 10x faster detection + 5x more failures is not a win

Repair Bandwidth DOESN'T Matter Much

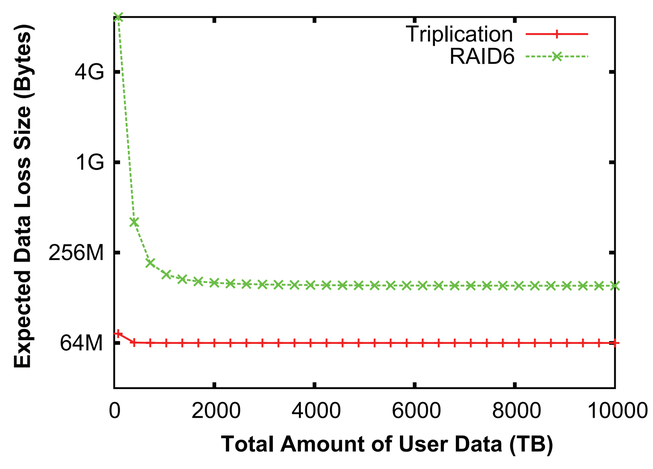
1TB/disk 80% full, 64MB/chunk, 8+2, 10 min detection



- Slower repair better for user performance during repair
- 10X slower repair has little impact on data loss rate

Expected Data Loss Size

1TB/disk 80% full, 64MB/chunk, 8+2,10min detection, 25MB/s disk repair



In larger systems:

- More disk combinations for a RAID set pattern
- For a given set of disks, less likely to lose data
- Expect to lose three chunks in a RAID set

Carnegie Mellon

Parallel Data Laboratory

PDLR10