Solving Data Loss in Massive Storage Systems

Jason Resch
Cleversafe

In the beginning

- There was replication
  - Long before advanced data protection techniques were known, data was copied
- Replication is wasteful
  - To survive N faults, N+1 copies were needed
  - Applied to disks, (N+1) times the hardware, power, floor space, and cooling are required
    - Not Cheap
    - Not Green
    - Not Performant

Enter RAID

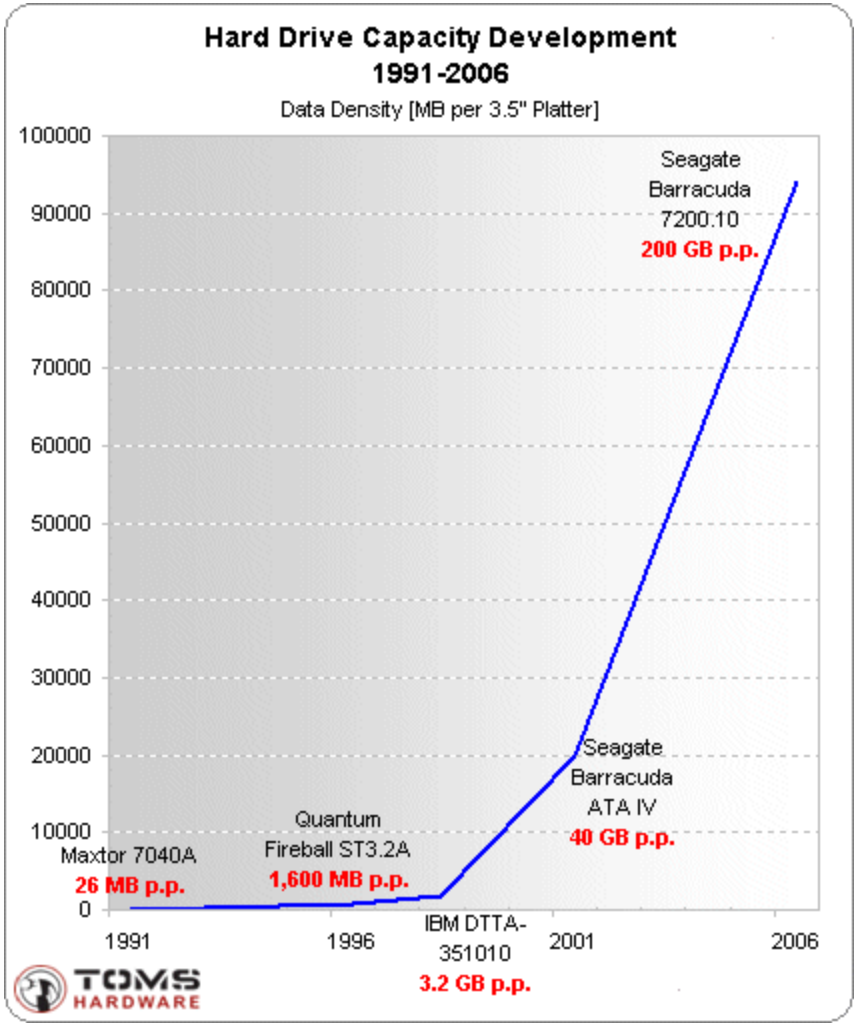
- In the 1980s, RAID was invented
- By storing a little extra information, regarding a larger set of information, errors can be corrected
  - RAID 5 stores parity information:
    - Parity is the property denoting even or odd
    - If the number of 1's across a set of drives is even parity bit is set to 0, if odd it is set to 1
      - If any disk is lost, the parity along with the bits on the surviving disks will yield the content of the lost disk
    - Example RAID 5 recovery:
      - 0 0 1 P = 1
      - 0 X 1 P = 1

Paradise Lost

- RAID 5 was great
  - Gave similar protection to making 1 copy, yet overhead was significantly less
  - For example: Using 3 disks for data and 1 for parity, the overhead was only 33%
- However, two factors would conspire to destroy the practical usefulness of RAID 5
  - Disk capacity outpacing performance
  - The growing chance of Latent Sector Errors (LSEs) which increased with disk capacity

Hard drive capacity growth

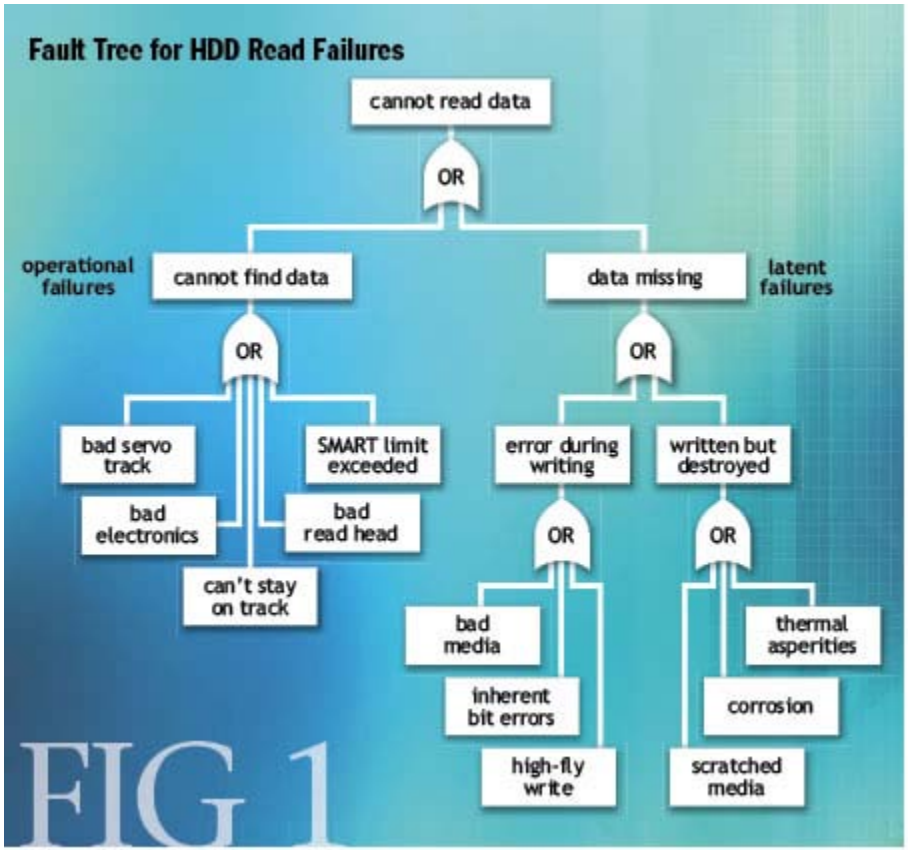
- Typical hard drive in 1991 was 40 MB and took 57 seconds to read entirely
- In 2006 an typical hard drive was 750 GB
  - 19,200-fold increase!
  - Took 3.27h to read [1]
- Today's 2 TB drives can take up to 8 hours



Impact on RAID 5

- RAID 5 can tolerate only one error at a time
- After first failure, data is in a vulnerable state
  - No additional redundancy exists
  - Secondary disk failure causes irrecoverable loss
- This was exceedingly unlikely when a disk could be rebuilt in minutes (as was the case in 1991)
- Today, disks can take hours or days to rebuild
  - Longer rebuild time means the chance of a secondary failure is ~500 times greater

Disks can fail in many ways



Jon Elerath 2007 [2]

- Outright disk failure is just one possibility
- More commonly, one or more sectors may be found unreadable at some future time
- A latent failure while rebuilding RAID 5 will cause data loss

Chance of a LSE during a rebuild

- Drive manufacturers often report LSE rates of 1 per every 10^{14} to 10^{15} bits (11 – 113 TB) read
  - When disks were only a couple of MB or GB, this probability was negligible
  - Consider a RAID 5 array using 2 TB disks:
    - After a disk failure, all other disks need to be read flawlessly, without encountering a LSE.
    - For a 4 disk array, 6 TB of data must be read
    - This works out to a 41% chance of a LSE during rebuild assuming LSE rate of 10^{-14} (5% if 10^{-15})

Impact of a LSE during rebuild

- A disk sector is corrupted (usually 512 bytes)
  - Effect may be minor, even unnoticed
  - Other times it may lead to corruption of a file
  - If the sector contained critical metadata, it may result in severe file system corruption
  - In some cases, especially with desktop-class drives, the drive may spend many minutes in a recovery mode, causing it to be kicked from the array and thus failing the whole rebuild

Quantifying Risk

- We now know: *bigger disks* = *increased risks*
  - But how significant is this risk?
  - How much data is expected to be lost?

- Fortunately there are techniques for calculating these risks if one knows the Disk's:
  - Mean Time to Failure (MTTF)
  - Capacity and performance
  - Rate of Latent Sector Errors

Mean Time To Failure (MTTF)

- Average time between failures
  - Over *useful life* of a component
  - Not to be confused with expected life
    - A 30-year old human has a MTTF of 900 years [3]
      - This doesn't imply they will live another 900 years
      - It implies a 1 in 900 chance of failing over 1 year
- Example application of MTTF:
  - Assume a drive has a MTTF of 20 years
  - We operate 1,000 such drives over 6 months
  - This works out to 500 drive-years
    - We should therefore expect (500 / 20) = 25 failures

Mean Time To Repair (MTTR)

- Average time to fully repair a failed component
  - Includes:
    - Time for operator to replace failed drive
    - Time to rebuild lost data on the new drive
  - Time to replace can vary significantly
    - May be hours or days, or zero with hot spares
  - Time to rebuild is often estimated
    - Take a drive's capacity and divide by its throughput
    - This is a best case scenario: in practice rebuilds may compete with normal I/O requests
      - (1/3) * (Capacity / Throughput) is more realistic [4]

Estimating time to data loss

- The MTTFs of sub-components can be combined to yield the MTTF for the system as a whole:
  - $MTTF_{computer} = \left(MTTF_{cpu}^{-1} + MTTF_{mem}^{-1} + MTTF_{psu}^{-1}\right)^{-1}$
    - Essentially, the inverse of the sum of the inverses
      - Also known as the Harmonic Sum
    - When the MTTFs are identical, a shortcut exists:
      - $MTTF_{sys} = MTTF_{sc} / N$
        - Where N is the number of sub-components
  - This explains why RAID 0 is so unreliable
    - $MTTDL_{RAID0} = MTTF_{disk} / NumDisks$
    - Has only a fraction the reliability of an individual disk

Estimating time to loss in RAID 5

- There are two paths to data loss in RAID 5:
  - Disk Failure followed by another during rebuild
  - Disk Failure followed by a LSE while rebuilding
- We know how to predict the time to the first failure
  - $MTTFirstFailure = MTTF_{disk} / NumDisks$
    - This doesn't imply data loss, only that a rebuild must occur
    - We must estimate the likelihood of a secondary failure
  - Assume the array had N disks to start
    - After the first failure N-1 disks remain
    - One of these must fail during the rebuild to cause data loss

Chance of Secondary Disk Failure

- Disk Failure followed by another during rebuild
  - $MTTFirstFailure = MTTF_{disk} / NumDisks$
  - $MTTSecondFailure = MTTF_{disk} / (NumDisks - 1)$
    - Second Failure must happen within the Rebuild Time
  - Therefore chance of second failure during rebuild is:
    - $SecondaryFailureChance = MTTR_{disk} / MTTSecondFailure$
- Putting it all together [5]:
  - $MTTDL_{RAID5\_DF} = MTTFirstFailure / SecondaryFailureChance$
  - $$MTTDL_{RAID5\_DF} = \frac{MTTF_{disk}^2}{MTTR_{disk} \cdot N \cdot (N-1)}$$

Chance of LSE During Rebuild

- Disk Failure followed by a Latent Sector Error
  - $MTTFirstFailure = MTTF_{disk} / N$
  - (N – 1) Disks remain and must be read entirely
  - Therefore chance of a LSE during rebuild is:
    - $ErrorDuring\,\mathrm{Re}\,build = 1 - \left(1 - LSE_{rate}\right)^{BitsPerDisk \cdot (N-1)}$
- Putting it all together:
  - $MTTDL_{RAID5\_LSE} = MTTFirstFailure / ErrorDuring\,\mathrm{Re}\,build$
  - $$MTTDL_{RAID5\_LSE} = \frac{MTTF_{disk}}{ErrorDuring\,\mathrm{Re}\,build \cdot N}$$

Combining paths to loss

- There are two paths to data loss in RAID 5:
  - Disk Failure followed by another during rebuild
  - Disk Failure followed by a LSE while rebuilding

- We can now calculate the MTTF for each path, but how can they be combined into a single estimate?
  - $$MTTDL_{RAID5} = \left(MTTDL^{-1}_{RAID5\_DF} + MTTDL^{-1}_{RAID5\_LSE}\right)^{-1}$$
    - We simply use the Harmonic sum, as we learned before

What good is a MTTF number?

- The MTTF statistic on its own is not very meaningful
  - However it can be used to generate actionable information, such as chance of data loss or expected amount of data loss over a period of time.

- Failures can be assumed to be random processes
  - Constant failure rates imply a Poisson distribution
  - $$FailureChanceOverTime(t) = 1 - e^{-t/MTTDL}$$
    - Where e is Euler's number ~= 2.718281828459045235360…

Estimating amount of Data Lost

- Another useful statistic is Expected Data Loss (EDL)
  - $EDL(t) = (Z \cdot t)/MTTDL$
    - Z is the amount of data lost in a data loss event
      - For Disk Failures, it is the usable capacity of the RAID array
      - For LSEs is theoretically the sector size, but more practically it may be the average file size, as often a whole file can become unusable
        - Depends on data format and associated application resiliency
- Assume an array with 6 TB usable and 500 MB files:
  - $EDL_{DF}(t) = (6TB \cdot t)/MTTDL_{DF}$
  - $EDL_{LSE}(t) = (500MB \cdot t)/MTTDL_{LSE}$
  - $EDL_{total}(t) = EDL_{DF}(t) + EDL_{LSE}(t)$

Example RAID 5 Calculation

- Let's calculate expected data loss and chance of loss for a RAID 5 configuration using 2 TB disks, assume:
  - RAID 5 configuration of 3 data disks, 1 parity disk
  - MTTF of disk is 220,000 hours (~4% AFR) [6]
  - LSE rate is 10^{-14} per bit
  - Disks Rebuild at 30 MB/s → MTTR is 19.41 hours
  - Average weighted file size is 500 MB
    - Picking a random sector on the disk, what is the average size of the file that contains that sector?

Why RAID 5 is dead

- Using the same formulae described previously:
  - FailureChanceOverTime(10 years) = 47.98%
    - $MTTF_{total}$ = 15.30 years
      - $MTTF_{DF}$ = 23,704.84 years
      - $MTTF_{LSE}$ = 15.31 years
  - EDL_{total}(10 years) = 2,980.68 MB
    - EDL_{DF}(10 years) = 2,654.08 MB
    - EDL_{LSE}(10 years) = 326.60 MB

The successor, RAID 6

- RAID 6 can recover from *two* simultaneous failures
  - Loss requires one of:
    - 3 disk failures during a rebuild window
    - 2 disk failures during a rebuild window plus a LSE
- Reliability formulas for RAID systems:
  - Notice a pattern?

$$MTTDL_{RAID0\_DF} = \frac{MTTF_{disk}^{1}}{MTTR_{disk}^{0} \cdot N}$$

$$MTTDL_{RAID5\_DF} = \frac{MTTF_{disk}^{2}}{MTTR_{disk}^{1} \cdot N \cdot (N-1)}$$

$$MTTDL_{RAID6\_DF} = \frac{2 \cdot MTTF_{disk}^{3}}{MTTR_{disk}^{2} \cdot N \cdot (N-1) \cdot (N-2)}$$

Why RAID 6 is so much better

- Every additional tolerated failure increases MTTF by:
  - MTTF / (MTTR × N)
    - MTTF is usually many years, while MTTR is a time in hours
    - With current disk MTTF and MTTR times, each additional tolerated failure increases reliability by a factor of several hundred to a few thousand!
- Reliability metrics for a RAID 6 array (6+2):
  - FailureChanceOverTime(10 years) = 0.13%
  - EDL_{total}(10 years) = 7.20 MB

Problem Solved?

- For that RAID 6 system, the chance of data loss over 10 years is about 1 in 780
  - It would seem the data loss daemon has been slain
- However, there are two factors not accounted for:
  - Some storage systems are massive (in the PB scale)
  - Disk capacities keep doubling

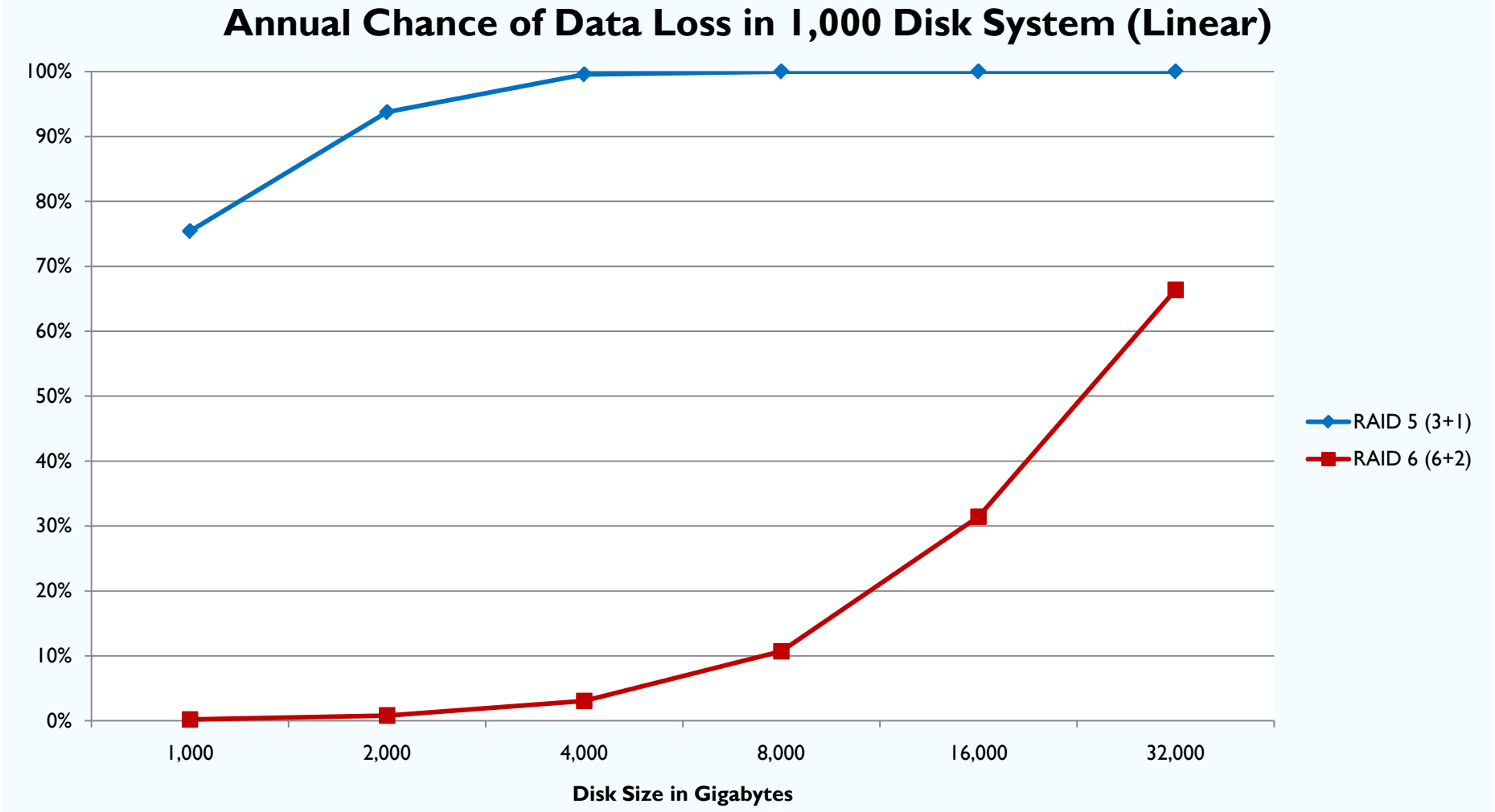
Issues of Scale

- Large systems require a large number of arrays
  - One cannot create a 998+2 RAID 6 array
    - Too many disks would have to be touched for each update
    - The chance of tertiary failures would be too great
- Each array has its own independent chance of failure
  - Recall that $MTTF_{sys} = MTTF_{sc} / N$
  - Its true whether the component is a disk or an array
  - Consider a 5 PB storage system
    - This requires 427 individual RAID 6 arrays
      - Assuming 2 TB disks in a 6+2 configuration
    - Failure of any array causes irrecoverable data loss

Why RAID 6 is dead (in big systems)

- For a 5 Petabyte RAID 6 system:
  - FailureChanceOverTime(10 years) = 42.19%
    - $MTTF_{total}$ = 18.24 years
      - $MTTF_{DF}$ = 44,944.69 years
      - $MTTF_{LSE}$ = 18.25 years
  - EDL_{total}(10 years) = 3,073.56 MB
    - EDL_{DF}(10 years) = 2,799.64 MB
    - EDL_{LSE}(10 years) = 273.91 MB
- This is essentially as bad as the single RAID 5 array…

Why RAID 6 is dead (for big disks)



Is Replication The Answer?

- When spending millions of dollars for a storage system, who wants to double or triple that cost?

- Instead, we can take the same path that was taken from RAID 5 to RAID 6
  - Scale out fault tolerance
  - Maintain same level of storage efficiency
  - Only additional cost:
    - Increased processing

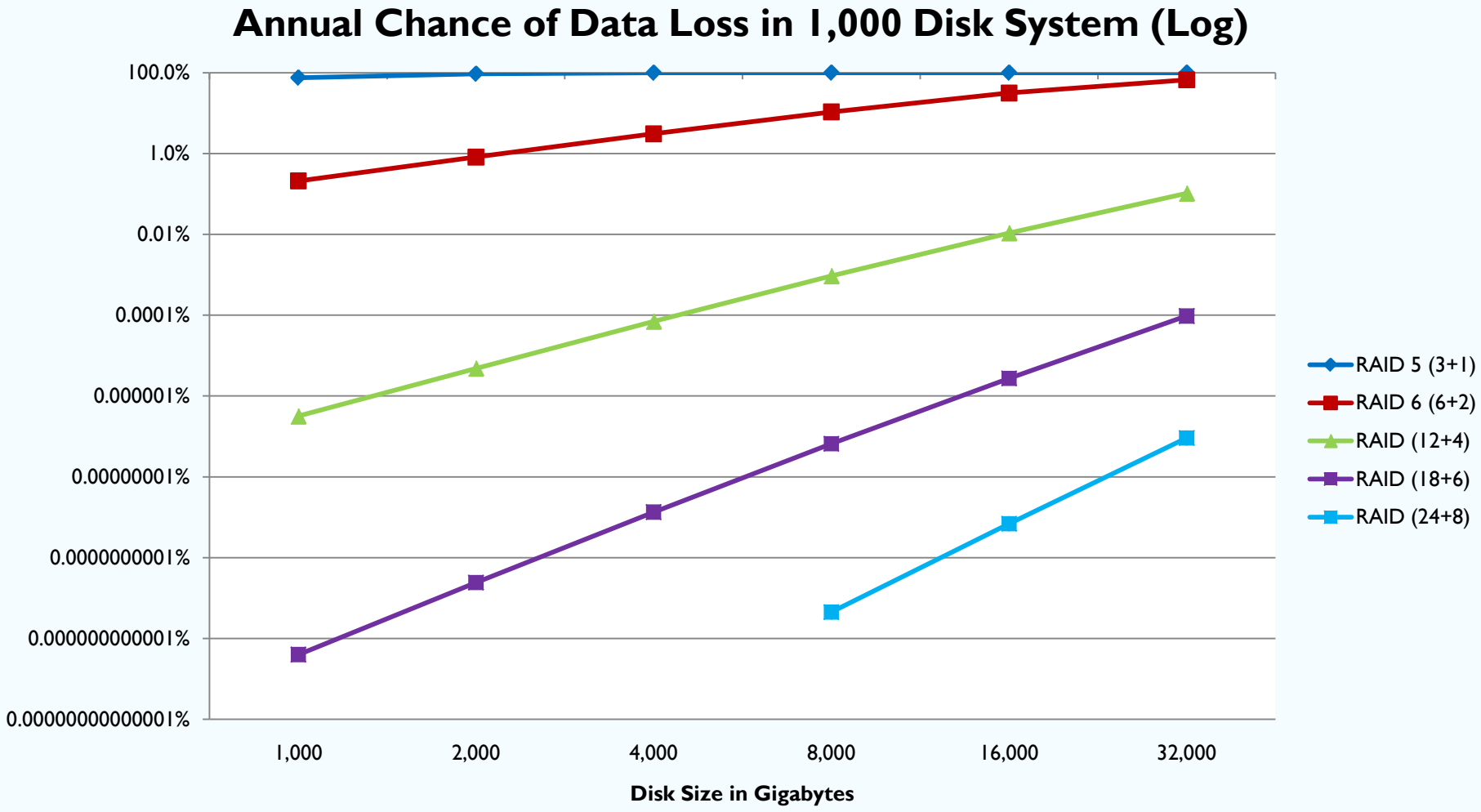
Reliability for arbitrary K-of-N

- Where K is the number of data Disks, and N is the total number of disks in the array
- System tolerates N – K failures without loss [7]

- $MTTF_{DF} = \frac{MTTF}{k * \binom{n}{k}} \times \left(\frac{MTTF}{MTTR}\right)^{n-k}$

- $MTTF_{LSE} = \frac{MTTF}{(k+1) * \binom{n}{k+1}} \times \left(\frac{MTTF}{MTTR}\right)^{n-(k+1)} \times \frac{1}{P(LSE)}$

Solution: Scale Fault Tolerance



References

- [1] http://www.tomshardware.com/reviews/15-years-of-hard-drive-history,1368.html
- [2] http://queue.acm.org/detail.cfm?id=1317403
- [3] http://www.faqs.org/faqs/arch-storage/part2/section-151.html
- [4] http://storageadvisors.adaptec.com/2005/11/01/raid-reliability-calculations/
- [5] RAID: High-Performance, Reliable Secondary Storage (1994 Chen et al.)
- [6] Failure Trends in a Large Disk Drive Population (2007 Pinherio et al.)
- [7] On Computing MTBF for a k-out-of-n:G Repairable System