Failures in hard disks and SSDs

Tyler Bletsch
Duke University

Slides include material from Vince Freeh (NCSU), some material adapted from “Hard-Disk Drives: The Good, the Bad, and the Ugly” by Jon Elerath (Comm. ACM, Vol. 52 No. 6, Pages 38-45)
HDD/SSD failures

- Hard disks are the weak link
  - A mechanical system in a silicon world!
- SSDs better, but still fallible

- RAID: Redundant Array of Independent Disks
  - Helps compensate for the device-level problems
  - Increases reliability and performance
  - Will be discussed in depth later
Failure modes

• Failure: cannot access the data

  • **Operational**: faults detected when they occur
    • Does not return data
    • Easy to detect
    • Low rates of occurrence

  • **Latent**: undetected fault, only found when it’s too late
    • Returned data is corrupt
    • Hard to detect
    • Relatively high rates of occurrence
To learn more about individual failure modes for HDD, see "Hard-Disk Drives: The Good, the Bad, and the Ugly" by Jon Elerath (Comm. ACM, Vol. 52 No. 6, Pages 38-45)
Fault tree for SSD

- Controller failure
- Whole flash chip failure

Operational Failures
- cannot find data

Latent Failures
- data missing
  - error during writing
  - written but destroyed

Degradation loss due to write cycles (probabilistic) – gate lost ability to ever hold data

Loss of gate state over time ("bit rot") – gate lost its current data (due to time or adjacent writes)
What to do about failure

- Pull disk out
- Throw away
- Restore its data from parity (RAID) or backup
The danger of latent errors

- Operational errors:
  - Detected as soon as they happen
  - When you detect an operational error, the total number of errors is likely one

- Latent errors:
  - Accrue in secret over time!
  - In the darkness, little by little, your data is quietly corrupted
  - When you detect a latent error, the total number of errors is likely many

- In the intensive I/O of reconstructing data lost due to latent errors, more likely to encounter operational error
  - Now you’ve got multiple drive failure, data loss more likely
Minimizing latent errors

• Catch latent errors earlier (so fewer can accrue) with this highly advanced and complex algorithm known as **Disk Scrubbing**:

  Periodically, read everything
Disk reliability

- MTBF (Mean Time Between Failure): a useless lie you can ignore

1,000,000 hours = 114 years

“Our drives fail after around a century of continuous use.”

-- A Huge Liar
### Data from BackBlaze

- **BackBlaze**: a large scale backup provider
  - Consumes thousands of hard drives, publishes health data on all of them publically
  - Data presented is a little old – newer data exists (but didn’t come with pretty graphs)
- Other large-scale studies of drive reliability:
  - “Failure Trends in a Large Disk Drive Population” by Pinheiro et al (Google), FAST’07
  - “Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?” by Schroeder et al (CMU), FAST’07
General Predicted Failure Rates

- Decreasing Failure Rate
- Constant Failure Rate
- Increasing Failure Rate

- Early "Infant Mortality" Failure
- Observed Failure Rate
- Constant (Random) Failures
- Wear Out Failures

Time

Failure Rate
Annual Failure Rate Each Quarter

![Annual Failure Rate Chart](chart.png)
Drives Have 3 Distinct Failure Rates

Interesting observation: The industry standard warranty period is 3 years...
80% of Drives Last Four Years

Hard Drive Survival Rates - Chart 2

Survival Rate

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%

1 Year 2 Year 3 Year 4 Year

80% drives live
What about SSDs?

• From recent paper at FAST’16: **Flash Reliability in Production: The Expected and the Unexpected** by Schroeder et al (feat. data from Google)

• **KEY CONCLUSIONS**
  - Ignore Uncorrectable Bit Error Rate (UBER) specs. A meaningless number.
  - **Good news:** Raw Bit Error Rate (RBER) increases slower than expected from wearout and is not correlated with UBER or other failures.
  - High-end SLC drives are no more reliable that MLC drives.
  - **Bad news:** SSDs fail at a lower rate than disks, but UBER rate is higher (see below for what this means).
  - **SSD age, not usage, affects reliability.**
    - Bad blocks in new SSDs are common, and drives with a large number of bad blocks are much more likely to lose hundreds of other blocks, most likely due to die or chip failure.
    - 30-80 percent of SSDs develop at least one bad block and 2-7 percent develop at least one bad chip in the first four years of deployment.

Drive replacements

- Percentage of drives replaced annually due to suspected hardware problems over the first 4 years in the field:

  - MLC-A
  - MLC-B
  - MLC-C
  - MLC-D
  - SLC-A
  - SLC-B
  - SLC-C
  - SLC-D

  Average annual replacement rates for hard disks (2-20%)

- ~1-2% of drives replaced annually, much lower than hard disks!
- 0.5-1.5% of drives developed bad chips per year
  - Would have been replaced without methods for tolerating chip failure

Slide from "Flash Reliability in Production: The Expected and the Unexpected" by Schroeder et al. FAST'16.
Errors experienced during a drive’s lifecycle

- Non-transparent errors common:
  - 26-60% of drives with uncorrectable errors
  - 2-6 out of 1,000 drive days experience uncorrectable errors
  - Much worse than for hard disk drives (3.5% experiencing sector errors)
Overall conclusions on drive health

- **HDD:**
  - Usually just die, sometimes have undetected bit errors.
  - Need to protect against drive data loss!

- **SSD:**
  - Usually have undetected bit errors, sometimes just die.
  - Need to protect against drive data loss!

- **Overall conclusion?**
  Need to protect against drive data loss!