HIGH PERFORMANCE INNOVATION

Errors in the Datapath
April 21, 2009

Henry Newman

Instrumental Inc/CTO

hsn@instrumental.com

What is Silent Corruption

- Communication on computer systems is protected by error detection and correction encoding
  - Reed Solomon encoding is the common method
- Silent corruption happens when there is an error in both the error encoding section and an error in the data section of the packet
- Error rates on most channels are $10E^{12}$ bits
- Depending on Channel type, channels are corrected to higher values

Types of Silent Corruption

- This can result in 2 different types of silent corruption
  - Mis-corrected
    - The error encoding determines that there is an error either rightly or wrongly and incorrectly changes the data
  - Undetected error
    - The error encoding and the data packet both have an error such that the error detection mechanism does not detect the error(s)

More Background

- Different channels have different amounts and protection schemes
  - FC
  - SATA
  - PCIe
  - FICON
- Are all different
- This presentation focuses on FC and SATA given that is what is mostly used today

FC and SATA Background

- Fibre Channel originally developed at 25 MB/sec; today it is 64 times faster at 1600 MB/sec (8 Gb/sec FC full duplex)
  - Channel error rate is1 in $10E^{12}$ bits
  - No one expected SANs with many 100s of switch ports
- IDE channel originally was .625 MB/sec and it is now 480 times faster at 300 MB/sec
  - The channel error rate is 1 in $10E^{12}$ bits
  - No one expected the performance we have today

Media is not the Problem

- Both disk and tape have far more error encoding than the channels
  - The undetectable or mis-corrected error rate for LTO tape is
    - $10E^{28}$ bits
  - T10000B tape is at claimed
    - $10E^{33}$ bits
  - Disk above the channel rate but not public information
- The channel error encode was not a consideration 20 years ago as things were too slow and too expensive to have lots of channels
  - This is no longer true

Error Encoding

- Robustness of error encoding has not changed for either storage channel type over the life of the channel
  - Between 20 (FC) and 25+ (SATA) years
- This has resulted in a situation where organizations are starting to see actual data loss as we have hit the wall with error encoding
- Fun fact: USB cables are rated at around 1 in $10E^9$ bits corrected from what I can find
  - There was no change in this rate with a 40x increase in the data rate from USB 1 to 2

Examples of Silent Corruption

- CERN
  - http://www.nsc.liu.se/lcsc2007/presentations/LCSC_2007-kelemen.pdf
- NetApp
  - http://www.usenix.org/event/fast08/tech/full_papers/bairavasundaram/bairavasundaram.pdf
  - http://institutes.lanl.gov/hec-fsio/conferences/2008/presentations/day3/Schroeder-HEC-FSIO.pdf
- Netflix
  - http://blog.netflix.com/2008/08/shipping-delay-recap.html

CERN Conclusions

- Silent corruptions are a fact of life
  - First step towards a solution is detection
  - Complete elimination seems impossible
- Existing datasets are at the mercy of Murphy
- Effort has to start now (if not started already)
  - Correction will cost time AND money
  - Multiple cost-schemes exist:
    - Trade time and storage space (à la Google)
    - Trade time and CPU power (correction codes)
- The best protection is probably a combination of redundancy and correction codes

NetApp Conclusions

- Big differences between disk models
  - 2 orders of magnitude difference in median
- Silent data corruption happens!
- More than 400,000 instances in our study
- For nearline drives, 8% discovered during RAID reconstruction
- Nearline drives are affected an order of magnitude more often than enterprise
- Affected enterprise drives develop more corruptions than nearline drives
  - Strong spatial locality
  - Strong dependence in time
- Some disk models can be really bad
  - Model E-1: 3% of disks have corruption and 25% of those have > 1000 errors (all within 17 months)

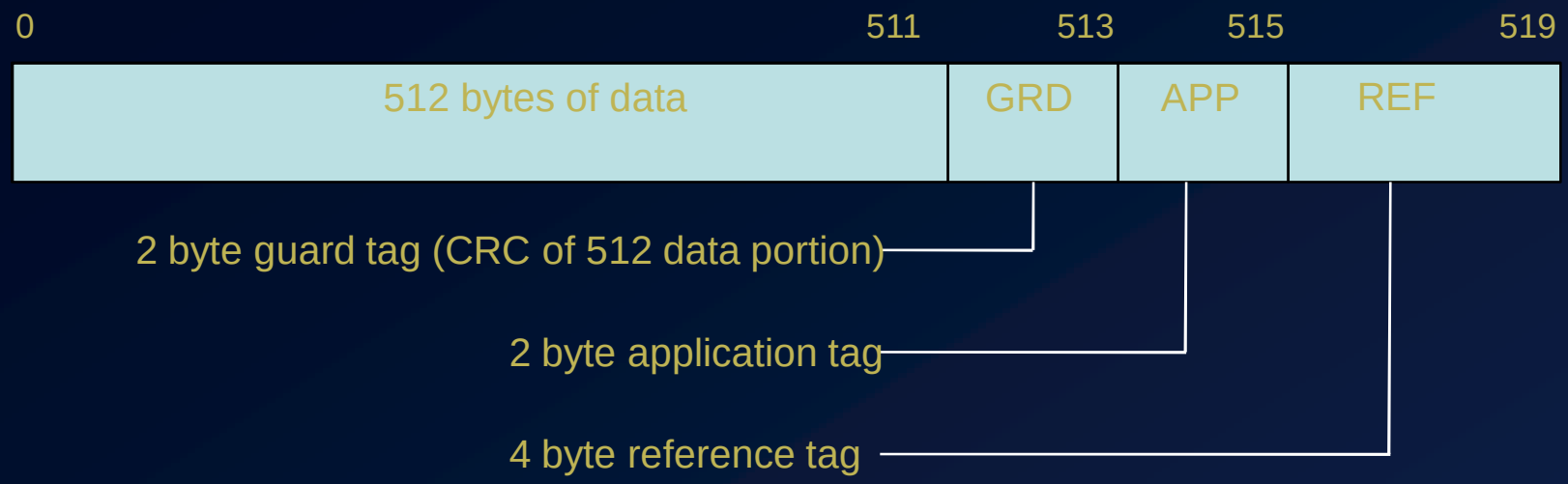
NetFlix Conclusions

- Remember the shipping database corruption from August 2008?
- Here is what Netflix engineering put in their blog
  - With some great forensic help from our vendors, root cause was identified as a faulty key hardware component
  - It definitively caused the problem, yet reported no detectable errors
- Netflix uses only enterprise storage from IBM (Shark), EMC (DMX), HDS (USPV)

Results

- Silent corruption is real
- We are getting to a point where the increased channel speed and size of the configurations is causing us to hit the error encoding limits
- Some people recognized this was coming a few years ago
- The ANSI T10 group and other recognized this was coming and created a new standard about 5 years ago
  - It is just coming to fruition likely this year
    - Disk drives already support this

The T10 PI Field



- PI adds an application checksum but:
  - What standard framework exists to allow the application to do this?
    - None is the answer as it is not supported in user space
    - No file system or OS support for application tag
- Oracle HARD is a special case
  - Done via ioctl and special file system interface

Increased Reliability in Datapath

- T10 Protection Information (PI) allows a checksum to be transmitted from the HBA and application to the disk drive
- It detects errors and does not correct them
  - If an error is found, a SCSI retry is initiated
- PI significantly (according to some $10E^{28}$) improves reliability in the datapath
  - No support for SATA or tapes
  - This is a concern given the channel is at risk

What the Rate Means to you

	Annual Failure Rates at Different Sustained Transfer Rates Per Second.						
UDBER	0.5 GB/sec	1 GB/sec	10 GB/sec	100 GB/sec	1 TB/sec	10 TB/sec	100 TB/sec
1.E-28	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.E-27	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.E-26	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.E-25	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.E-24	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.E-23	0.0	0.0	0.0	0.0	0.0	0.0	0.3
1.E-22	0.0	0.0	0.0	0.0	0.0	0.3	2.7
1.E-21	0.0	0.0	0.0	0.0	0.3	2.7	27.1
1.E-20	0.0	0.0	0.0	0.3	2.7	27.1	270.9
1.E-19	0.0	0.0	0.3	2.7	27.1	270.9	2708.9
1.E-18	0.1	0.3	2.7	27.1	270.9	2708.9	27089.2
1.E-17	1.4	2.7	27.1	270.9	2708.9	27089.2	270892.2
1.E-16	13.5	27.1	270.9	2708.9	27089.2	270892.2	2708921.8
1.E-15	135.4	270.9	2708.9	27089.2	270892.2	2708921.8	27089217.7

These annual failure rates are for a perfect world where the channels are operating at the specified rate of $10E^{12}$ and corrected to$10E^{17/19}$

What happens when the world is not perfect?

CRC Errors

- As hardware starts to fail CRC errors
- If multiple components are getting CRC I speculate that the chances of silent corruption can increase dramatically
  - A number of places I am aware of have reported corruptions when multiple components were reporting intermittent CRCs
    - HBA, switch ports, RAID front ends
    - This is not definitive proof but to me it seems logical that the more errors the higher likelihood that you will get an error at the same time you are checking the incoming packet for errors
    - Do something wrong with the packet

My Recommendations

- Monitor hardware and replace failing hardware as quickly as possible
- Use FC wherever possible for communication and keep SATA drives to a minimum
  - Parity checking RAID devices will help address the SATA reliability between the RAID controller and drive, but does not impact the reliability between the host and RAID
- If you really care about your data then T10 PI is the only standards based protection mechanism

My Final Thoughts

- If there is corruption, most people blame the file system first and the hardware last
  - That might have been a good plan in the 1970s - 1990s but it is no longer true in many cases
- Some questions I have are:
  - Does error correction belong in the file system or hardware?
  - What should be done about errors in SATA drives?
  - What are the error rates expected for 40 Gb and 100 Gb ethernet?
  - What should be done about tape?

Instrumental

Thank You