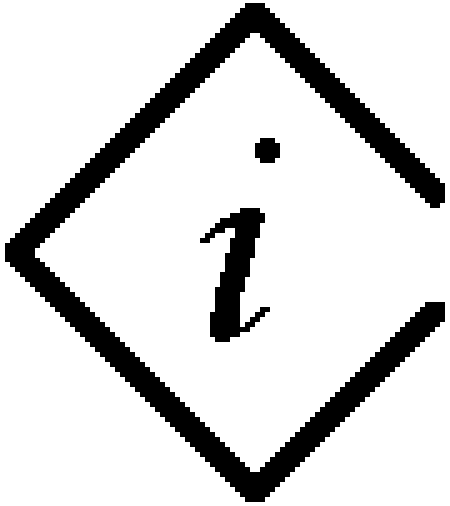
Recovering from iSCSI Digest Errors



Mark Bakke

June 28, 2001

Expected Digest Error Rates

- Layer 2 network – local
  - Already covered by layer 2 CRC
- Layer 3 network – local
  - Need latency and p(cksum fail)
- Metro network
  - Need latency and p(cksum fail)
- Internet (better error rate)
- Internet (worst error rate)

From Stone & Partridge

- Some numbers adjusted by factor of 10
  - Eliminated ack-of-fin bug errors
- “Good” and “Bad” internet connections
  - Data on number of TCP checksum failures
  - Computed range of expected iSCSI CRC failures
  - Used TCP cksum failure as p(packet loss)
    - P(loss) often will be higher

Assumptions

- Too many to enumerate ☺
- All packets dropped are due to TCP checksum mismatch
  - Assumes no congestion loss
- Achievable bandwidth assumes current TCP slow-start
- 1500-byte MTU
- 8192-byte iSCSI PDU

More notes

- Assumptions made in favor of seeing *more* digest errors
  - Actual rates are probably at the low end of the ranges given
- 1 Gbps and 10 Gbps should have virtually no digest errors if run on a layer 2 network

“Good” internet

- TCP checksum mismatch 1 in 90,000
- Checksum escape 1 in 135 million to 10 billion
- Bandwidth 30 Mbits/second @ 100ms RTT
  - 8 to 600 digest errors per year
  - 1 header digest error every 2 months to 10 years
- Bandwidth 300 Mbits/sec @ 10ms RTT
  - 80 to 6,000 digest errors per year
  - 1 to 70 header digest errors per year

“Bad” internet

- TCP checksum mismatch 1 in 11,000
- TCP checksum escape 1 in 16 million to 1 in 1 billion
- Max bandwidth 10 Mbits/second @ 100ms
  - 0.5 to 33 digest errors per week
  - 26 to 1650 digest errors per year
  - 0.3 to 20 header digest errors per year
- Max bandwidth 100 Mbits/second @ 10ms
  - 5 to 335 digest errors per week
  - 260 to 16,500 digest errors per year
  - 3 to 200 header digest errors per year

1 Gbps iSCSI connection

- Assuming current TCP, 70% bandwidth utilization
  - Can get worst error rate at 70% utilization????
- Expected digest errors at 100 ms
  - Packet loss less than 1 in 50 million
  - 0.5 to 40 digest errors per year
  - 1 header digest error every 2..100 years
- Expected digest errors at 10 ms
  - Packet loss less than 1 in 500,000
  - Less than 60 to 4000 digest errors per year
  - 1 to 40 header digest errors per year

10 Gbps iSCSI connection

- Assuming current TCP, 70% bandwidth utilization
- Expected digest errors at 100 ms RTT
  - Packet loss less than 1 in 5 billion
  - 1 digest error every 3 months to 10 years
  - 1 header digest error every 20 to 1000 years
- Expected digest errors at 10 ms RTT
  - Packet loss less than 1 in 50 million
  - Less than 6 to 400 digest errors per year
  - 1 header digest error every 3 months to 12 years

Digest Error rate and Bandwidth

- If digest error rate is high
  - Packet loss rate will be much higher
  - Causing achievable bandwidth to plummet
  - Due to TCP congestion avoidance
  - Which brings down digest error rate
- 10x higher latency tolerates 100x fewer errors
- 10x higher bandwidth tolerates 10x fewer errors

Tape Jobs

- iSCSI CRC errors fail tape jobs
  - If no iSCSI recovery done
- Failures based on tape job size, not bandwidth
- 30,000 MB tape job at 10 ms
  - Bad Internet fails most jobs
  - Good Internet fails around 1 in 400 jobs
  - 1 Gbps connection can fail 1 in 400 jobs
  - 10 Gbps connection can fail 1 in 40,000 jobs
  - 100x fewer fails tolerated if can do at 100ms

How often will it happen?

- Often enough to worry about
  - Without it, data will be corrupted
- Seldom enough to keep recovery simple
  - 1..10Gbps cxns will see few errors per year
  - Performance of error recovery irrelevant
  - Disk can take the performance hit
- Tape can be a problem

Fixing up tape

- Who would do big tapes over bad links?
- IPsec can help shore up TCP
  - Might be needed for security anyway
- SSC-2
  - Hosts and remote tape servers could implement
  - Tape devices would not have to implement

Do we need iSCSI CRC?

- Data WILL be corrupted eventually
- 10s to 100s of disk blocks per year could be corrupted over internet-like cxn
- 1 tape backup job in hundreds..thousands can be corrupted over internet-like cxn

Recovery Scheme

- Trust-and-verify TCP
- Let SCSI recover (or not) from errors
- When will this not be good enough?
  - Too many tape job failures
    - 1 in ???…??? considered unacceptable
  - Enough disk failures to affect performance
    - How many times per day/week are OK?
    - Different for MPIO and non-MPIO scenarios

Recovery from Header CRC

- Probably 1-2% of total digest errors
  - Extremely seldom
  - Tens per year on a bad connection
- Take the big hammer approach
  - Kill the connection
  - Kill the session
  - Abort outstanding tasks

Recovery from Data CRC

- 98-99% of digest errors
  - Normally less than one or two per week
- Little hammer approach
  - Keep connection/session alive
  - Propagate parity error or similar to SCSI
- Write data CRC (detected by target)
- Read data CRC (detected by initiator)
  - How to deal with ordering???

When to use iSCSI CRC

- Over any internet or similar connection
- When iSCSI proxies or gateways are used
- Negotiate away on layer 2 networks
  - Target & initiator may negotiate away
  - iSCSI proxy should insist on it

Deploying iSCSI CRC

- Offload card
  - Only on well-tested, reliable host hardware
  - Same goes for devices and gateways

Other reasons for error recovery

- Interface or path failure
  - Many systems use multipath I/O for disk
  - Should iSCSI provide another layer??
  - Nothing yet for tape (SSC-2)
- iSCSI format errors
  - Treat these as bugs; kill the session!

Questions

- iSCSI CRC errors is just one type of error
  - Do other errors require finer-grained recovery?
- Will other connection errors happen?
- With what frequency?
  - Keep in mind the bandwidth limits if they do

Really Low-End

- What about Coke® machines?
  - Diskless DOS over 56k line
  - 700 byte window size at 100 ms
  - One digest error every 1..6 years
  - Single outstanding command at a time
  - Recovery: reboot the machine

Framing Methods

- Each method adds its own 8-byte header
  - These need to be protected, too

Framing Method	Framing Header	Escapes per iSCSI header escape
Markers	2k intervals	0.3
Framing w/o chunks	Per 1k iSCSI PDU	1.5
Framing w/chunks	Per 1460-byte packet	0.5