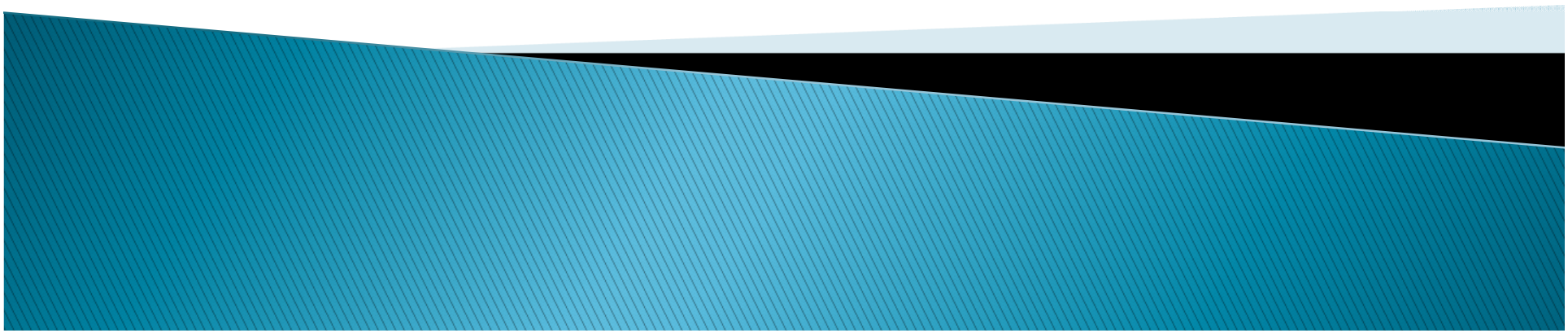
Fixity: problems with fixity at scale

Henry Newman
Scott Rife
9/18/2017



What is Fixity?

- File Fixity is a digital preservation term referring to the property of a digital file being fixed, or unchanged
- Fixity is typically checked using a cryptographic hash function such as SHA256
- Cryptographic hash is a mathematical algorithm that maps data of arbitrary size to a bit string of a fixed size (a hash function)
  - It is more suitable to ensuring fixity than a checksum because it is infeasible to reverse engineer

What is a Checksum?

- A checksum is a small-sized datum derived from a block of digital data for the purpose of detecting errors which may have been introduced during transmission and/or storage.
- Checksums like CRC and parity are used to identify errors as packets and blocks of files are moved through a server or network
- A Cyclic Redundancy Check (CRC) is an error-detecting code used to detect accidental changes to data
- A parity bit, or check bit, is a bit added to a string of binary digits to ensure that the total number of 1-bits in the string is even or odd. Parity bits are used as the simplest form of error detecting code.

What is a Collision?

- It occurs when the same value is generated for 2 different sets of source bits
  - Take the well-known checksum function CRC32
  - If you feed this function the two strings “plumless” and “buckeroo”, it generates the same value. This is known as a collision.
    - “plumless” => CRC32 => 0x4ddb0c25
    - “buckeroo” => CRC32 => 0x4ddb0c25

Scaling Fixity

- Scale refers growth in number of bytes and/or number of files
- CRC codes in network packets are 32 bit.
  - Originally implemented 1 (Mb/sec) Megabit Ethernet networks
- Parity bit in memory
  - Early implementations Kilobytes (KiB) of memory
- Parity bit in bus communication on the backplane of servers
  - Early implementations 80 Megabytes/sec SCSI

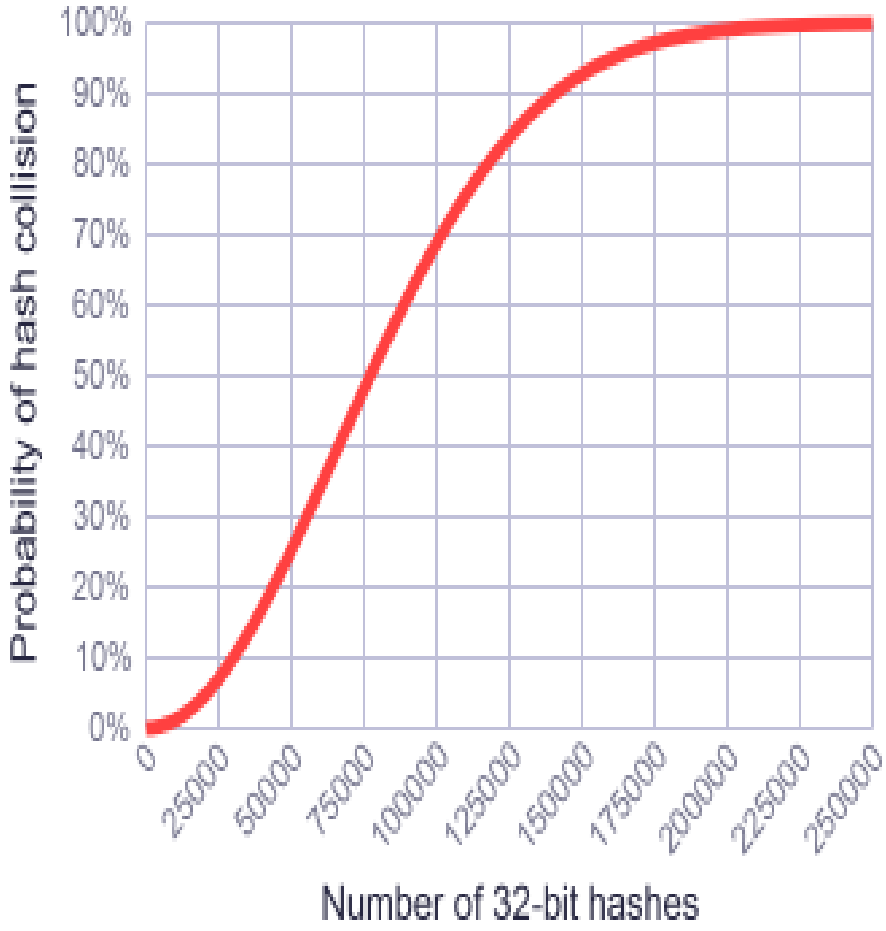
Collisions increase at scale

- Collisions can occur
- 87.3 to 238.3 packets/second in Megabit Ethernet. Call it 100 to make the math easier.
- 1,000,000 packets/second in 10 Gigabit Ethernet
- Assuming every packet has an error, at 100 packets a second, the likelihood of a collision is near 0.
  - At 25,000 it is 10%
  - At 75,000 it is 50%
  - At 200,000 it is 100%

http://preshing.com/20110504/hash-collision-probabilities/

Probability of collisions in CRC

Here is a graph for N =2^32. This illustrates the probability of collision when using 32-bit CRC. It's worth noting that a 50% chance of collision occurs when the number of packets is 77,163. Also note that the graph takes the same S-curved shape for any value of N.



http://preshing.com/20110504/hash-collision-probabilities/

Probability of a collision is reduced by increasing the number of bits in hash

Number of **32-bit** hash values	Number of **64-bit** hash values	Number of **160-bit** hash values	**Odds of a hash collision**
77163	5.06 billion	1.42×10^{24}	1 in 2
30084	1.97 billion	5.55×10^{23}	1 in 10
9292	609 million	1.71×10^{23}	1 in 100
2932	192 million	5.41×10^{22}	1 in 1000
927	60.7 million	1.71×10^{22}	1 in 10000
294	19.2 million	5.41×10^{21}	1 in 100000
93	6.07 million	1.71×10^{21}	1 in a million
30	1.92 million	5.41×10^{20}	1 in 10 million
10	607401	1.71×10^{20}	1 in 100 million
	192077	5.41×10^{19}	1 in a billion
	60740	1.71×10^{19}	1 in 10 billion
	19208	5.41×10^{18}	1 in 100 billion
	6074	1.71×10^{18}	1 in a trillion
	1921	5.41×10^{17}	1 in 10 trillion
	608	1.71×10^{17}	1 in 100 trillion
	193	5.41×10^{16}	1 in 10^{15}
	61	1.71×10^{16}	1 in 10^{16}
	20	5.41×10^{15}	1 in 10^{17}
	7	1.71×10^{15}	1 in 10^{18}

Odds of a full house in poker
1 in 693

Odds of four-of-a-kind in poker
1 in 4164

Odds of being struck by lightning
1 in 576000

Odds of winning a 6/49 lottery
1 in 13.9 million

Odds of dying in a shark attack
1 in 300 million

Odds of a meteor landing on your house
1 in 182 trillion

http://preshing.com/20110504/hash-collision-probabilities/

Hash collisions and packet errors increase at scale

- This does not mean that there is a packet with hash collision every second because not all packets are in error
- It is more likely that a network, with a component in distress / generating many error packets, can send a packet with bad payload through
- Multiple levels of checking are required to ensure fixity
- Note: If the header is corrupted then the packet is resent

Engineer and Budget solutions

- Fixity must be checked at different levels
  - File / Object
  - Shard / Stripe
  - Block / Packet
  - Bit / Byte (SECDED – Single error correction Double error detection)
- Systems must be engineered to perform these checks while keeping up with ingest and access
- Budgets must be justified to engineer systems to perform fixity and monitor errors

References

- http://www.digitalpreservation.gov/documents/NDSA-Fixity-Guidance-Report-final100214.pdf
- https://research.google.com/pubs/pub35162.html
- http://preshing.com/20110504/hash-collision-probabilities/
- https://www.caida.org/research/traffic-analysis/pkt_size_distribution/graphs.xml