The Fletcher Checksums in ZFS

Alan Somers alans@spectralogic.com

April 12, 2013

Introduction

One of the goals of ZFS, the Zettabyte File System [1], is robust data integrity. It is among the first filesystems to include error detection codes (EDC) for every block in the filesystem, including indirect blocks. Every block's EDC is stored in its parent block, all the way up to the root of the filesystem (the überblock). That ensures that each block is always verified against a known-good EDC. As the only orphan block in the filesystem, the überblock must store its own EDC. This makes it somewhat more vulnerable to corruption; to compensate, multiple copies of the überblock are stored.

As of zpool version 28, ZFS offers 4 choices for the checksum function: off (not recommended), Fletcher-2 (deprecated), Fletcher-4, or SHA-256. The two Fletcher options are based on the well-known Fletcher Checksum [2], but differ in the blocksize, checksum length, and checksum number. This paper will discuss the suitability of Fletcher-2 and Fletcher-4 for ensuring data integrity in ZFS.

Algorithms

ZFS's Fletcher-2 and Fletcher-4 checksums operate over a block of data ranging from 1 byte to 128KB. Fletcher-2 produces a 128-bit result consisting of two 64-bit numbers. It is defined by the following recurrence relations:

$$\begin{aligned} a_i &= a_{i-1} + f_{i-1} \pmod{2^{64}} \\ b_i &= b_{i-1} + a_i \pmod{2^{64}} \\ a_0 &= 0 \\ b_0 &= 0 \end{aligned}$$

where $\{f_{i=0}^{n-1}\}$ are the input data, taken in 64-bit words, and the variables a_i and b_i are 64-bit accumulators. As series, Fletcher-2 can be expressed as:

$$\begin{aligned} a_n &= \sum_{i=1}^{n} f_{n-i} \pmod{2^{64}} \\ b_n &= \sum_{i=1}^{n} i \times f_{n-i} \pmod{2^{64}} \end{aligned}$$

Fletcher-4 produces a 256-bit result consisting of four 64-bit numbers. The input data are taken in 32-bit words to prevent overflow of a and b. Because Fletcher-2 overflows a and b by a simple 2^{64} divisor, it is very easy for errors in the high bits to go undetected. For example, the sequences $(0, 0)$ and $(2^{63}, 2^{63})$ will have the same value for the Fletcher-2 checksum. For that reason, Fletcher-2 is deprecated and Fletcher-4 is recommended for newly created volumes. The recurrence relations and series expressions for Fletcher-4 are:

$$
\begin{array}{rcl}
a_i & = & a_{i-1} + f_{i-1} \pmod{2^{64}} \\
b_i & = & b_{i-1} + a_i \pmod{2^{64}} \\
c_i & = & c_{i-1} + b_i \pmod{2^{64}} \\
d_i & = & d_{i-1} + c_i \pmod{2^{64}} \\
a_0 & = & 0 \\
b_0 & = & 0 \\
c_0 & = & 0 \\
d_0 & = & 0 \\
a_n & = & \sum_{i=1}^{n} f_{n-i} \pmod{2^{64}} \\
b_n & = & \sum_{i=1}^{n} i \times f_{n-i} \pmod{2^{64}} \\
c_n & = & \sum_{i=1}^{n} \frac{i(i+1)}{2} f_{n-i} \pmod{2^{64}} \\
d_n & = & \sum_{i=1}^{n} \frac{i(i+1)(i+2)}{6} f_{n-i} \pmod{2^{64}}
\end{array}
$$

1 Overflow in Fletcher-4

Zpool version 28 supports block sizes of any power of 2 between 512 bytes and 128 KB. File tails will always be padded out to the next greatest multiple of four bytes. Thus, the checksum could operate over a sequence of up to $n = 32768$ words.

The maximum possible value for the checksums will be achieved when the input values are all $2^{32} - 1$. So we can see that a_n will never overflow and furthermore $a_n < 2^{47}$ for all $n \leq 32768$. Similarly, b_n will never overflow and $b_n < 2^{61}$ for all $n \leq 32768$. However, c_n does overflow starting at $n = 2953$; it can overflow a maximum of 682 times. d_n overflows at $n = 566$; it can overflow a maximum of 5593429 times.

2 Observations

Analysis of Fletcher-4's properties is considerably simplified by making use of a few observations. Firstly, each sum $a_n, ..., d_n$ is a linear function of the input, if we treat the input as one long vector. Therefore, if $\{f_{i=0}^{n-1}\} = \{g_{i=0}^{n-1}\} + \{h_{i=0}^{n-1}\}$, then

$$\text{fletcher4}(\{f_{i=0}^{n-1}\}) \equiv \text{fletcher4}(\{g_{i=0}^{n-1}\}) + \text{fletcher4}(\{h_{i=0}^{n-1}\}) \pmod{2^{64}} \tag{1}$$

Secondly, the checksum of a block of zeros is zero:

$$f_i = 0 \text{ for all } i < n \Rightarrow \text{fletcher4}(\{f_{i=0}^{n-1}\}) = 0 \tag{2}$$

Thirdly, if an input block is prefixed by a run of zeros, then the checksum will be the same whether or not that run of zeros is summed:

$$m < n, f_i = 0 \text{ for all } i < m \Rightarrow \text{fletcher4}(\{f_{i=0}^{n-1}\}) = \text{fletcher4}(\{f_{i=m}^{n-1}\}) \tag{3}$$

Finally, if two data blocks are suffixed by runs of zeros, then their checksums will be the same if and only if their checksums are the same before the trailing zeros are summed:

$$\begin{aligned} & m < n, f_i = g_i = 0 \text{ for all } m < i < n \\ & \qquad \Rightarrow \Big((\text{fletcher4}(\{f_{i=0}^{n-1}\}) = \text{fletcher4}(\{g_{i=0}^{n-1}\}) \Leftrightarrow \text{fletcher4}(\{f_{i=0}^{m}\}) = \text{fletcher4}(\{g_{i=0}^{m}\}) \Big) \end{aligned} \tag{4}$$

3 Hamming distance

The Hamming distance of a checksum is the smallest number of bit errors for which there is at least one undetected case. It is a very important measure of the performance of a checksum used in communication or storage systems where corruption events tend to affect small numbers of bits. A lower limit for the Hamming distance of Fletcher-4 is easily calculated if we ignore sums that overflow.

If $\{f_{i=0}^{n-1}\}$ is the original data block and $\{g_{i=0}^{n-1}\}$ is the corrupted data block, then we can represent it as $\{g_{i=0}^{n-1}\} = \{f_{i=0}^{n-1}\} + \{s_{i=0}^{n-1}\}$ where $\{s_{i=0}^{n-1}\}$ are the values to add (or subtract) from the original data block. The values of $\{s_{i=0}^{n-1}\}$ lie in the interval $(-2^{32}, 2^{32})$. Thanks to the linearity of Fletcher-4, we know that $\{f_{i=0}^{n-1}\}$ and $\{g_{i=0}^{n-1}\}$ will have the same checksums if and only if the checksum of $\{s_{i=0}^{n-1}\}$ is zero.

$$\begin{aligned}
\text{fletcher4}(\{f_{i=0}^{n-1}\}) &\equiv \text{fletcher4}(\{g_{i=0}^{n-1}\}) \pmod{2^{64}} && (5)\\
\text{fletcher4}(\{f_{i=0}^{n-1}\}) &\equiv \text{fletcher4}(\{f_{i=0}^{n-1}\}) + \text{fletcher4}(\{s_{i=0}^{n-1}\}) \pmod{2^{64}} && (6)\\
0 &\equiv \text{fletcher4}(\{s_{i=0}^{n-1}\}) \pmod{2^{64}} && (7)
\end{aligned}$$

Let us assume that $\{f_{i=0}^{n-1}\}$ and $\{g_{i=0}^{n-1}\}$ have the same checksum, and that they differ in only two words. If both words were increased or both were decreased relative to $\{f_{i=0}^{n-1}\}$, then it's obvious that equation (7) cannot be satisfied. Therefore, there must be exactly one positive word and one negative word in $\{s_{i=0}^{n-1}\}$. Let j_1 be the index of the positive word in and j_2 be the index of the negative word. Then we have

$$\begin{aligned}
0 &= \text{fletcher4}(\{s_{i=0}^{n-1}\})\\
0 &= \sum_{i=1}^{n} s_{n-i}\\
0 &= s_{j_1} + s_{j_2}\\
-s_{j_1} &= s_{j_2}\\
0 &= \sum_{i=1}^{n} i s_{n-i}\\
0 &= j_1 s_{j_1} + j_2 s_{j_2}\\
j_1 s_{j_2} &= j_2 s_{j_2}\\
j_1 &= j_2
\end{aligned}$$

This is a contradiction; the same word cannot be both increased and decreased. Therefore the Fletcher-4 checksum is immune from all 2-word errors (at least for 128KB blocks where the first two sums do not overflow). So its Hamming distance is at least 3.

Next let us consider $\{f_{i=0}^{n-1}\}$ and $\{g_{i=0}^{n-1}\}$ that have the same checksum but differ in three words. For $n < 2953$, c_n will never overflow so we can easily analyze its contribution to the Hamming distance. Let $\{s_{i=0}^{n-1}\}$ have three nonzero words with word indices j_1, j_2, and j_3.

$$\begin{aligned}
0 &= \sum_{i=1}^{n} s_{n-i}\\
0 &= s_{j_1} + s_{j_2} + s_{j_3}\\
0 &= \sum_{i=1}^{n} i s_{n-i}\\
0 &= j_1 s_{j_1} + j_2 s_{j_2} + j_3 s_{j_3}\\
0 &= \sum_{i=1}^{n} \frac{i(i+1)}{2} s_{n-i}\\
0 &= \frac{j_1(j_1+1)}{2} s_{j_1} + \frac{j_2(j_2+1)}{2} s_{j_2} + \frac{j_3(j_3+1)}{2} s_{j_3}
\end{aligned}$$

$$
\begin{aligned}
k(s_{j_1}+s_{j_2}) &= \\
k &= \frac{j_1 s_{j_1}+j_2 s_{j_2}}{s_{j_1}+s_{j_2}} \\
\sum_{i=1}^{n}\frac{i(i+1)}{2}r_{n-i} &= \sum_{i=1}^{n}\frac{i(i+1)}{2}s_{n-i} \\
\frac{k(k+1)}{2}r_k &= \frac{j_1(j_1+1)}{2}s_{j_1}+\frac{j_2(j_2+1)}{2}s_{j_2} \\
&\text{, if } s_{j_1}\neq s_{j_2}\text{or } j_1\neq j_2 \\
(k(k+1))(s_{j_1}+s_{j_2}) &= (j_1(j_1+1))s_{j_1}+(j_2(j_2+1))s_{j_2} \\
\left(\frac{j_1 s_{j_1}+j_2 s_{j_2}}{s_{j_1}+s_{j_2}}\right)\left(\frac{j_1 s_{j_1}+j_2 s_{j_2}+s_{j_1}+s_{j_2}}{s_{j_1}+s_{j_2}}\right)(s_{j_1}+s_{j_2}) &= \\
(j_1 s_{j_1}+j_2 s_{j_2})(j_1 s_{j_1}+j_2 s_{j_2}+s_{j_1}+s_{j_2}) &= \left((j_1^2+j_1)s_{j_1}+(j_2^2+j_2)s_{j_2}\right)(s_{j_1}+s_{j_2}) \\
j_1^2 s_{j_1}^2+2j_1j_2 s_{j_1}s_{j_2}+j_2^2 s_{j_2}^2+j_1 s_{j_1}^2+j_1 s_{j_1}s_{j_2}+j_2 s_{j_1}s_{j_2}+j_2 s_{j_2}^2 &= j_1^2 s_{j_1}^2+j_1^2 s_{j_1}s_{j_2}+j_2^2 s_{j_1}s_{j_2}+j_2^2 s_{j_2}^2 \\
&\quad + j_1 s_{j_1}^2 + j1 s_{j_1}s_{j_2}+j_2 s_{j_1}s_{j_2}+j_2 s_{j_2}^2 \\
2j_1j_2 s_{j_1}s_{j_2} &= j_1^2 s_{j_1}s_{j_2}+j_2^2 s_{j_1}s_{j_2} \\
2j_1j_2 &= j_1^2+j_2^2 \\
0 &= j_1^2-2j_1j_2+j_2^2 \\
0 &= (j_1-j_2)^2 \\
j_1 &= j_2
\end{aligned}
$$

This is a linear system of three equations with three unknowns, s_{j_1}, s_{j_2}, and s_{j_3}, if we treat the indices j_1, j_2, and j_3 as constant terms. We can solve it by using Gaussian elimination, noting any constant values that will result in degenerate solutions.

$$
\begin{pmatrix} 1 & 1 & 1 \\ j_1 & j_2 & j_3 \\ \frac{j_1(j_1+1)}{2} & \frac{j_2(j_2+1)}{2} & \frac{j_3(j_3+1)}{2} \end{pmatrix}\begin{pmatrix} s_{j_1} \\ s_{j_2} \\ s_{j_3} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}
$$

$$
\left(\begin{array}{ccc|c} 1 & 1 & 1 & 0 \\ j_1 & j_2 & j_3 & 0 \\ \frac{j_1(j_1+1)}{2} & \frac{j_2(j_2+1)}{2} & \frac{j_3(j_3+1)}{2} & 0 \end{array}\right)
$$

$$
\left(\begin{array}{ccc|c} 1 & 1 & 1 & 0 \\ 0 & -j_1+j_2 & -j_1+j_3 & 0 \\ \frac{j_1(j_1+1)}{2} & \frac{j_2(j_2+1)}{2} & \frac{j_3(j_3+1)}{2} & 0 \end{array}\right)
$$

$$
\left(\begin{array}{ccc|c} 1 & 1 & 1 & 0 \\ 0 & -j_1+j_2 & -j_1+j_3 & 0 \\ j_1(j_1+1) & j_2(j_2+1) & j_3(j_3+1) & 0 \end{array}\right)
$$

$$
\left(\begin{array}{ccc|c} 1 & 1 & 1 & 0 \\ 0 & -j_1+j_2 & -j_1+j_3 & 0 \\ 0 & -j_1-j_1^2+j_2+j_2^2 & -j_1-j_1^2+j_3+j_3^2 & 0 \end{array}\right)
$$

$$
\left(\begin{array}{ccc|c} 1 & 1 & 1 & 0 \\ 0 & -j_1+j_2 & -j_1+j_3 & 0 \\ 0 & 0 & -j_1-j_1^2+j_3+j_3^2+\frac{(j_1+j_1^2-j_2-j_2^2)(-j_1+j_3)}{-j_1+j_2} & 0 \end{array}\right)
$$

$$
\left(\begin{array}{ccc|c} 1 & 1 & 1 & 0 \\ 0 & -j_1+j_2 & -j_1+j_3 & 0 \\ 0 & 0 & (j_1-j_3)(j_2-j_3) & 0 \end{array}\right)
$$

The only solutions for s_{j_3} are the trivial $j_1 = j_3$, $j_2 = j_4$, or $s_{j_3} = 0$. These contradict the problem statement. Therefore, Fletcher-4 is immune from all 3-word errors for block sizes of less than 2953 words. For block sizes up to the maximum, observe that c_n will still not overflow when only three words are changed:

$$\sum_{i=1}^{n} \frac{i(i+1)}{2} s_{n-i} < \frac{1}{2} \times 3 \times 2^{15} \times (2^{15}+1) \times (2^{32}-1))) = 13835480264525905920 \approx 3 \times 2^{62}$$

Therefore the previous section still applies, and Fletcher-4 is immune from all three bit errors at blocksizes up to the maximum.

When four words have been changed, we must consider the d_n checksums to have enough equations for the Gaussian elimination. However, since the d_n checksum can overflow, collisions are still possible. For example:

word index	value	prime factorization
1	-805306368	$-1 \times 2^{28} \times 3$
4097	2013265920	$2^{27} \times 3 \times 5$
8193	-1342177280	$-1 \times 2^{28} \times 5$
20481	134217728	2^{27}

Therefore, the Hamming distance of Fletcher-4 is 4 words.

Independent Collisions

The last section dealt with Fletcher-4's resistance to low error rate gaussian noise, which affects a small number of bits. On modern hard drives, it is also possible (though unlikely) to experience silent data corruption that affects entire blocks. For example, the hard drive may return the contents of one LBA when the user requested a read of a different LBA. The Fletcher-4 checksum will almost always detect this error, since a block's checksum is stored in its parent indirect block. The only way for this type of error to go unnoticed by ZFS would be for two different, independent blocks to have the exact same checksum. In this section, we will attempt to calculate the likelihood of that collision.

First, we approximate the data words as real-valued from the interval $[0, 2^{32})$ instead of as discrete integers. Assume two blocks, each of length n real-valued words, filled with random, independent data uniformly distributed. We call these blocks $f_{i=0}^{n-1}$ and $g_{i=0}^{n-1}$. Each one is a sequence of samples of the random variable $\mathcal{U}(0, 2^{32})$. The first Fletcher-4 sum, $A(f_{i=0}^{n-1})$, is simply a sum of n samples of $\mathcal{U}(0, 2^{32})$. It has an Irwin-Hall distribution, which is approximately Gaussian for any allowed ZFS blocksize. So

$$A(f_{i=0}^{n-1}) \sim \mathcal{N}\left(2^{31}n, \frac{2^{62}}{3}n\right)$$

Under our real-valued approximation, the probability that two blocks' checksums will collide is the probability that the difference in their sums will be less than one half:

$$P(\text{collision in } a_n) = P\left(\left|A(f_{i=0}^{n-1}) - A(g_{i=0}^{n-1})\right| < \frac{1}{2}\right)$$

Since $A(f_{i=0}^{n-1})$ is normally distributed, the difference is also normally distributed:

$$A(f_{i=0}^{n-1}) - A(g_{i=0}^{n-1}) \sim \mathcal{N}\left(0, \frac{2^{63}}{3}n\right) \tag{8}$$

So we can calculate the probability that two blocks' A checksums collide:

$$\begin{aligned}
P\left(\left|A(f_{i=0}^{n-1}) - A(g_{i=0}^{n-1})\right| < \frac{1}{2}\right) &= \Phi\left(\frac{\frac{1}{2}}{\sqrt{\frac{2^{63}}{3}n}}\right) - \Phi\left(\frac{-\frac{1}{2}}{\sqrt{\frac{2^{63}}{3}n}}\right) \\
&= \frac{1}{2}\left(\operatorname{erf}\left(\frac{\frac{1}{2}}{\sqrt{\frac{2^{63}}{3}n}\sqrt{2}}\right) - \operatorname{erf}\left(\frac{-\frac{1}{2}}{\sqrt{\frac{2^{63}}{3}n}\sqrt{2}}\right)\right) \\
&= \frac{1}{2}\left(\operatorname{erf}\left(\frac{1}{2^{33}}\sqrt{\frac{3}{n}}\right) - \operatorname{erf}\left(-\frac{1}{2^{33}}\sqrt{\frac{3}{n}}\right)\right) \\
&= 1.25694 \times 10^{-12}\big|n = 32768
\end{aligned}$$

where Φ is the cumulative distribution function of the standard normal distribution.

The B checksum is a sum of n random variables, each of which is distributed uniformly across a different range. $B(f_{i=0}^{n-1}) \sim \mathcal{U}(0, 2^{32}) + 2\mathcal{U}(0, 2^{32}) + ... + n\mathcal{U}(0, 2^{32})$. If we approximate each r.v. as $\mathcal{U}(0, 2^{32}) \sim \mathcal{N}(2^{31}, \frac{2^{62}}{3})$ (an approximation which is supported by simulation), then

$$\begin{aligned}
B(f_{i=0}^{n-1}) &\sim n\mathcal{N}\left(2^{31}, \frac{2^{62}}{3}\right) + (n-1)\mathcal{N}\left(2^{31}, \frac{2^{62}}{3}\right) + ... + \mathcal{N}\left(2^{31}, \frac{2^{62}}{3}\right) \tag{9} \\
&= \mathcal{N}\left(\sum_{i=1}^{n} 2^{31}i, \sum_{i=1}^{n} \frac{2^{62}}{3}i^2\right) \tag{10} \\
&= \mathcal{N}\left(n(n+1)2^{30}, \left(\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}\right)\frac{2^{62}}{3}\right) \tag{11} \\
B(f_{i=0}^{n-1}) - B(g_{i=0}^{n-1}) &\sim \mathcal{N}\left(0, \left(\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}\right)\frac{2^{63}}{3}\right) \tag{12}
\end{aligned}$$

The probability that two blocks' B checksums collide is then:

$$\begin{aligned}
P\left(\left|B(f_{i=0}^{n-1}) - B(g_{i=0}^{n-1})\right| < \frac{1}{2}\right) &= \Phi\left(\frac{\frac{1}{2}}{\sqrt{\left(\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}\right)\frac{2^{63}}{3}}}\right) - \Phi\left(\frac{-\frac{1}{2}}{\sqrt{\left(\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}\right)\frac{2^{63}}{3}}}\right) \\
&= \frac{1}{2}\left(\operatorname{erf}\left(\frac{\frac{1}{2}}{\sqrt{\left(\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}\right)\frac{2^{64}}{3}}}\right) - \operatorname{erf}\left(\frac{-\frac{1}{2}}{\sqrt{\left(\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}\right)\frac{2^{64}}{3}}}\right)\right) \\
&= \frac{1}{2}\left(\operatorname{erf}\left(\frac{1}{2^{33}}\sqrt{\frac{3}{\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}}}\right) - \operatorname{erf}\left(-\frac{1}{2^{33}}\sqrt{\frac{3}{\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}}}\right)\right) \\
&= 6.64357 \times 10^{-17}\big|n = 32768
\end{aligned}$$

We can calculate the probability that two blocks' C checksums collide using the same method that we did for the B checksums. The only difference is that the C checksum can wrap around the 2^{64} boundary.

$$
\begin{aligned}
C(f_{i=0}^{n-1}) &\sim \frac{n(n+1)}{2}\mathcal{N}\left(2^{31}, \frac{2^{62}}{3}\right) + \frac{(n-1)(n)}{2}\mathcal{N}\left(2^{31}, \frac{2^{62}}{3}\right) + ... + \mathcal{N}\left(2^{31}, \frac{2^{62}}{3}\right) \\
&= \mathcal{N}\left(\sum_{i=1}^{n} 2^{31}\frac{i(i+1)}{2}, \sum_{i=1}^{n} \frac{2^{62}}{3}\frac{i^2(i+1)^2}{4}\right) \\
&= \mathcal{N}\left(2^{30}\left(\sum_{i=1}^{n} i^2 + \sum_{i=1}^{n} i\right), \frac{2^{60}}{3}\left(\sum_{i=1}^{n} i^4 + \sum_{i=1}^{n} 2i^3 + \sum_{i=1}^{n} i^2\right)\right) \\
&= \mathcal{N}\Bigg(2^{30}\left(\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6} + \frac{n(n+1)}{2}\right) \\
&\qquad , \frac{2^{60}}{3}\left(\frac{n^5}{5} + \frac{n^4}{2} + \frac{n^3}{3} - \frac{n}{30} + 2\left(\frac{n^4}{4} + \frac{n^3}{2} + \frac{n^2}{4}\right) + \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}\right)\Bigg) \\
&= \mathcal{N}\left(2^{30}\left(\frac{n^3}{3} + n^2 + \frac{2n}{3}\right), \frac{2^{60}}{3}\left(\frac{n^5}{5} + n^4 + \frac{5n^3}{3} + n^2 + \frac{2n}{15}\right)\right) \\
C(f_{i=0}^{n-1}) - C(g_{i=0}^{n-1}) &\sim \mathcal{N}\left(0, \frac{2^{61}}{3}\left(\frac{n^5}{5} + n^4 + \frac{5n^3}{3} + n^2 + \frac{2n}{15}\right)\right)
\end{aligned}
$$

For blocks of $n \leq 2952$ where the C checksums never overflow, the probability that two blocks' C checksums collide is

$$
\begin{aligned}
\sigma &= \sqrt{\frac{2^{61}}{3}\left(\frac{n^5}{5} + n^4 + \frac{5n^3}{3} + n^2 + \frac{2n}{15}\right)} \\
P\left(\left|C(f_{i=0}^{n-1}) - C(g_{i=0}^{n-1})\right| < \frac{1}{2}\right) &= \Phi\left(\frac{1}{2\sigma}\right) - \Phi\left(-\frac{1}{2\sigma}\right) \\
&= \frac{1}{2}\left(\operatorname{erf}\left(\frac{1}{2\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{1}{2\sqrt{2}\sigma}\right)\right) \\
&= 2.9719 \times 10^{-18}\big|n = 2592
\end{aligned}
$$

For larger blocks we must consider the overflow. When $k = 0$, it is easy to accurately evaluate the cumulative distribution function. However, when $k \neq 0$, the floating-point rounding error is too great. For those cases we will numerically integrate the probability density function (PDF)instead. Because the PDF is so flat and the interval so short, we can accurately evaluate it using the rectangle rule with one sample point. Also, we combine the positive and negative integrals since the PDF is an even function.

$$\begin{aligned}
P\left(\left|C(f_{i=0}^{n-1}) - C(g_{i=0}^{n-1})\right| < \frac{1}{2} \pmod{2^{64}}\right) &= \sum_{k=-\infty}^{\infty} \Phi\left(\frac{2^{64}k + \frac{1}{2}}{\sigma}\right) - \Phi\left(\frac{2^{64}k - \frac{1}{2}}{\sigma}\right) \\
&= \frac{1}{2}\sum_{k=-\infty}^{\infty} \operatorname{erf}\left(\frac{2^{64}k + \frac{1}{2}}{\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{2^{64}k - \frac{1}{2}}{\sqrt{2}\sigma}\right) \\
&= \frac{1}{2}\left(\operatorname{erf}\left(\frac{1}{2\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{1}{2\sqrt{2}\sigma}\right)\right) \\
&\quad + \sum_{k=-\infty}^{-1} \int_{2^{64}k-\frac{1}{2}}^{2^{64}k+\frac{1}{2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-t^2/(2\sigma^2)} dt \\
&\quad + \sum_{k=1}^{\infty} \int_{2^{64}k-\frac{1}{2}}^{2^{64}k+\frac{1}{2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-t^2/(2\sigma^2)} dt \\
&\approx \frac{1}{2}\left(\operatorname{erf}\left(\frac{1}{2\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{1}{2\sqrt{2}\sigma}\right)\right) \\
&\quad + 2\sum_{k=1}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(2^{64}k\right)^2/(2\sigma^2)} \\
&= 5.42101 \times 10^{-20} \big| n = 32768
\end{aligned}$$

The calculation for the D checksum is very similar to the C checksum.

$$\begin{aligned}
D(f_{i=0}^{n-1}) &\sim \frac{n(n+1)(n+2)}{6}\mathcal{N}\left(2^{31}, \frac{2^{62}}{3}\right) + \frac{(n-1)(n)(n+1)}{6}\mathcal{N}\left(2^{31}, \frac{2^{62}}{3}\right) + ... + \mathcal{N}\left(2^{31}, \frac{2^{62}}{3}\right) \\
&= \mathcal{N}\left(\sum_{i=1}^{n} 2^{31}\frac{i(i+1)(i+2)}{6}, \sum_{i=1}^{n} \frac{2^{62}}{3}\frac{i^2(i+1)^2(i+1)^6}{36}\right) \\
&= \mathcal{N}\left(\frac{2^{30}}{3}\sum_{i=1}^{n} i^3 + 3i^2 + 2i, \frac{2^{60}}{27}\sum_{i=1}^{n} i^6 + 6i^5 + 13i^4 + 12i^3 + 4i^2\right) \\
&= \mathcal{N}\Bigg(\frac{2^{30}}{3}\frac{n^4}{4} + \frac{n^3}{2} + \frac{n^2}{4} + 3\left(\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}\right) + 2\left(\frac{n^2}{2} + \frac{n}{2}\right), \\
&\quad \frac{2^{60}}{27}\Bigg(\frac{n^7}{7} + \frac{n^6}{2} + \frac{n^5}{2} - \frac{n^3}{6} + \frac{n}{42} + 6\left(\frac{n^6}{6} + \frac{n^5}{2} + \frac{5n^4}{12} - \frac{n^2}{12}\right) + 13\left(\frac{n^5}{5} + \frac{n^4}{2} + \frac{n^3}{3} - \frac{n}{30}\right) \\
&\quad +12\left(\frac{n^4}{4} + \frac{n^3}{2} + \frac{n^2}{4}\right) + 4\left(\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}\right)\Bigg)\Bigg) \\
&= \mathcal{N}\left(\frac{2^{30}}{3}\left(\frac{n^4}{4} + \frac{3n^3}{2} + \frac{11n^2}{4} + \frac{3n}{2}\right), \frac{2^{60}}{27}\left(\frac{n^7}{7} + \frac{3n^6}{2} + \frac{61n^5}{10} + 12n^4 + \frac{23n^3}{2} + \frac{9n^2}{2} + \frac{9n}{35}\right)\right)
\end{aligned}$$

$$\begin{aligned}
\sigma &= \sqrt{\frac{2^{61}}{27}\left(\frac{n^7}{7} + \frac{3n^6}{2} + \frac{61n^5}{10} + 12n^4 + \frac{23n^3}{2} + \frac{9n^2}{2} + \frac{9n}{35}\right)} \\
P\left(\left|D(f_{i=0}^{n-1}) - D(g_{i=0}^{n-1})\right| < \frac{1}{2} \pmod{2^{64}}\right) &\approx \frac{1}{2}\left(\operatorname{erf}\left(\frac{1}{2\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{1}{2\sqrt{2}\sigma}\right)\right) \\
&\quad + 2\sum_{k=1}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(2^{64}k\right)^2/(2\sigma^2)} \\
&= 5.42101 \times 10^{-20} \big| n = 32768
\end{aligned}$$

So far we have calculated the independent collision probabilities for each of the 4 checksums. But the checksums are not all independent: numerical simulation shows that A and B are strongly correlated. Our method will be to numerically fit a multivariate normal distribution to the combination of A and B, and approximate C and D as being uncorrelated to the other sums. Then we will simply multiply the collision probabilites for our three independent random variables.

Let Δ_{AB} be the two dimensional random variable formed by the combination of Δ_A and Δ_B. From (8) and (12) we know its mean vector and the diagonal of its covariance matrix. The off-diagonal entries of the covariance matrix must be calculated numerically. Once we have done that, we can calculate the probability that Δ_{AB} falls in the range ([-.5, .5], [-.5, .5]) by numerically integrating the PDF. As before, the PDF is sufficiently flat that it can be accurately integrated using the rectangle rule with one sample.

$$\begin{aligned}
\Delta_{AB} &\sim \mathcal{N}_2\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} \sigma^2_{\Delta_A} & \rho\sigma_{\Delta_A}\sigma_{\Delta_B} \\ \rho\sigma_{\Delta_A}\sigma_{\Delta_B} & \sigma^2_{\Delta_B} \end{array}\right)\right) \\
\Delta_{AB} &\sim \mathcal{N}_2\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} \frac{2^{63}}{3}n & \rho\sigma_{\Delta_A}\sigma_{\Delta_B} \\ \rho\sigma_{\Delta_A}\sigma_{\Delta_B} & \frac{2^{62}}{3}\left(\frac{n^3}{3}+\frac{n^2}{2}+\frac{n}{6}\right) \end{array}\right)\right) \\
&\sim \mathcal{N}_2\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} 1.0061\times 10^{-23} & 1.6459\times 10^{-27} \\ 1.6459\times 10^{-27} & 3.5947\times 10^{-31} \end{array}\right)\right)\Bigg|n = 32768 \\
P\left(|\Delta_A| < 0.5 \text{ and } |\Delta_B| < 0.5\right) &\approx \frac{1}{2\pi\sigma_{\Delta_A}\sigma_{\Delta_B}\sqrt{1-\rho^2}}e^0 \\
&= 1.671\times 10^{-28}
\end{aligned}$$

Since Δ_C and Δ_D are uncorrelated with each other and with Δ_{AB}, we can simply multiply their collision probabilities together.

$$\begin{aligned}
P(\text{collision}) &= P\left(|\Delta_A| < 0.5 \text{ and } |\Delta_B| < 0.5\right) \times P\left(|\Delta_C| < 0.5\right) \times P\left(|\Delta_D| < 0.5\right) \\
&\approx 1.671\times 10^{-28} \times 5.42101\times 10^{-20} \times 5.42101\times 10^{-20}\big|n = 32768 \\
&= 4.9106\times 10^{-67}
\end{aligned}$$

References

[1] TODO

[2] J. G. Fletcher. An arithmetic checksum for serial transmissions. *IEEE Transactions on Communications*, COM-30(1):247252, Jan. 1982.