HalftimeHash: a repaired collision bound and a refuted variant

Thomas Dybdahl Ahle

September 2026

Abstract

The published HalftimeHash composition proof treats a difference between vectors as a difference in every coordinate. That implication fails, so its multiplication of coordinate-tree collision bounds is unjustified. We repair the intended construction using integer matrix fibres and a projection-polynomial inequality. For a distance-k encoder and maximum determinant valuation t, the resulting bound is $2^{-32k}(h+2)^{k-1}(h+1+2^t)$ at forest height h. It gives $6804 \cdot 2^{-96}$, or 83.27 collision bits, for the intended three-word construction at height 16. The shipped `Encode3` has distance one instead of three: an explicit equal-length pair collides in the reachable 24-byte core with probability at least 2^{-32}. The four 64-bit Style wrappers use a valid distance-two encoder; their overlapping key layout admits the sharp length-normalized coefficient $2^{-63} - 2^{-128}$ on explicit finite domains. A corrected distance-three encoder and terminal length word restore the three-word theorem, including unequal lengths. We state the ideal-key and implementation contracts and distinguish paper proofs, exact certificates, and completed Lean theorems.

1 Construction, probability model, and scope

Apple's HalftimeHash[3] uses encode–hash–combine compression, a forest of NH trees, and a final NH reduction. It belongs to the class of fast universal constructions also represented by Khuong's UMASH[5]; the repair here concerns HalftimeHash's own arithmetic and composition. The algebraic and implementation distinctions are developed below.

Put $Q = 2^{32}$, $R = 2^{64} = Q^2$, $\varepsilon = Q^{-1}$, and $\delta = R^{-1} = \varepsilon^2$. All key words are independent uniform 64-bit words. Messages are fixed independently of the key. An η-almost-universal family has collision probability at most η for each distinct pair. An η-almost-difference-universal family has probability at most η for every fixed prescribed output difference. For vectors the latter condition prescribes every coordinate; it does not mean that output coordinates are independent or uniform.

A block contains $b \in \{1, 2, 4, 8\}$ 64-bit words, processed in parallel. Each encoder input symbol is a triple of blocks. An encoder J maps d symbols to e symbols and should have minimum symbol distance k: every distinct input pair differs in at least k encoded symbols. With independently keyed symbol hashes and an integer $k \times e$ matrix T,

$$\mathrm{EHC}(x) = T\big(H_0(J(x)_0), \ldots, H_{e-1}(J(x)_{e-1})\big)^{\mathsf{T}} \quad \text{over } \mathbb{Z}_R.$$

Each scalar position within a block undergoes the same encoder and matrix. Symbol keys can be broadcast over those positions and reused across groups.

The elementary NH product on a 64-bit input with 32-bit halves (x, y) and key halves (s, t) is

$$N_{s,t}(x, y) = ((x + s) \bmod Q)((y + t) \bmod Q) \pmod R. \tag{1}$$

NH sums these products using a separate key per pair. Tree nodes pass the first child through unchanged as an accumulator and NH-hash the remaining children. Distinct keys are used at

different heights and output coordinates, but nodes at one height may reuse keys. Fresh final NH compresses each coordinate's ordered forest roots, with separate keys for root positions and scalar positions within a block. A separate pool hashes the padded raw tail using shifted, overlapping windows of whole NH pairs. Equal byte lengths have the same forest shape, padding support, and key offsets.

The intended theorem requires this distance property, the matrix conditions below, and independence between these specified key pools. The shipped code needs a separate analysis because its encoder and address layout do not meet every informal assumption in the original proof. Throughout, arithmetic means unsigned modular arithmetic with little-endian loads, eight-bit bytes, 64-bit `uint64_t`, and 64-bit `size_t`. The fixed implementation additionally requires valid contiguous key storage of the reported extent and disjoint output storage. There is no claim for a short-seed pseudorandom key expansion.

2 Why the published Lemma 3 fails

Lemma 3 in the published paper asserts a tree-stage collision bound $(h2^{-32})^k$, conditional on EHC not having collided[3, Lemma 3]. The premise only says the two EHC vectors are distinct. For example, $(0,0)$ and $(1,0)$ differ in the first coordinate, while the second coordinate receives identical input and collides with probability one. For a height-one NH comparison with low input half zero versus one, the first coordinate collides exactly when the other key half is zero, of probability ε. The joint event therefore has probability ε, rather than ε^2, despite independent tree keys. Additional unchanged coordinates give the same obstruction at larger k.

This refutes the conditional inference. It does not refute the intended construction's final bound, because its EHC distribution controls how many coordinates can agree. The repair must retain that distribution and its subset probabilities through the tree calculation.

The matrix argument has separate algebraic defects. Write $\det A = q_0 2^{t'}$ with q_0 odd. The exponent is the valuation t', not the power of two $2^{t'}$. Multiplication by an adjugate yields a necessary condition. After removing the odd factor, a soluble congruence $2^{t'}z = \beta \pmod R$ becomes

$$z \equiv \beta/2^{t'} \pmod{2^{64-t'}},$$

with the required division and divisibility condition. Counting these solutions gives an upper bound, not equality of probabilities. Matrix fibres provide a sharper and more direct repair than treating all coordinates independently after this adjugate step.

3 A repaired theorem for the intended construction

Lemma 3.1 (Integer NH)**.** *On equal-length distinct pair strings, NH is ε-almost-difference-universal. The modified tree node is ε-almost-universal.*

Proof. Choose a differing half in a differing input pair, and fix its key half. The corresponding two factors $c, c' \in [0, Q)$ are distinct. Condition all other pairs; the remaining uniform half z must solve

$$cz - c'((z+d) \bmod Q) = v \pmod{Q^2}.$$

Within each wrap interval the nonzero slope gives at most one solution. If $z_1 < z_2$ were solutions on different sides, their difference would require $(c-c')(z_2-z_1) + Qc' = 0 \pmod{Q^2}$. This integer is strictly between zero and Q^2, whether $c > c'$ or $c < c'$, and cannot vanish. Thus at most one of Q half-keys works. This also proves the sum version by conditioning all other pairs, the standard NH property[4].

For a tree node, if a hashed child differs the same proof incorporates the passed-through child's difference into the target. If only the passed-through child differs, collision is impossible. The

latter case is an equality guarantee; it is not an almost-difference-universal assertion for every target on all distinct node inputs. □

Lemma 3.2 (Matrix fibres)**.** *If an integer square matrix A is nonsingular with Smith invariants d_i, every nonempty fibre over $(\mathbb{Z}_{2^{64}})^r$ has size*

$$|\ker A| = \prod_i 2^{\min(64,\nu_2(d_i))} \le 2^{\nu_2(\det A)}.$$

If $\nu_2(\det A) < 64$, equality holds in the last inequality.

Proof. Integer Smith normal form gives unimodular U, V with $UAV = \operatorname{diag}(d_i)$. Their determinants ± 1 remain units modulo 2^{64}, so the two changes of variables preserve fibre size. The scalar kernel has $\gcd(d_i, 2^{64})$ elements. Multiply these counts; every soluble fibre is a translate of the kernel. Since determinant valuations add, the final assertion follows. The valuation restriction matters for the equality formula. □

Fix any k differing encoded symbols, a set denoted F. For each output-row subset S, put $c_{\varnothing,F} = 1$ and

$$c_{S,F} = \min_{\substack{J\subseteq F,\ |J|=|S| \\ \det T[S,J]\neq 0}} 2^{\nu_2(\det T[S,J])}. \tag{2}$$

If T_F is nonsingular, every row subset has full row rank, so this minimum exists. Condition all symbol keys except a minimizing J. Each prescribed joint atom of the selected symbol differences costs at most $\varepsilon^{|S|}$ by independence and Lemma 3.1. Counting matrix fibres therefore gives, for every target z_S,

$$\Pr[\mathrm{EHC}(x)_S - \mathrm{EHC}(y)_S = z_S] \le c_{S,F}\varepsilon^{|S|}. \tag{3}$$

For whole blocks, choose one scalar position where the original group differs. The distance assumption must hold on that position. Equality of whole blocks implies equality there. Broadcast keys therefore require no independence between scalar positions.

Theorem 3.3 (Corrected EHC bound)**.** *If every k-column submatrix of T is nonsingular and $t = \max_F \nu_2(\det T_F)$, a distance-$k$ encoder with independent ε-almost-difference-universal symbol hashes gives a $2^t\varepsilon^k$-almost-difference-universal EHC family.*

Proof. Choose k differing symbols and apply (3) to all output rows. The argument is confined to this fixed group and permits key reuse across other groups. □

Lemma 3.4 (Projection polynomial)**.** *For nonsingular $A = T_F$ and $u, v \ge 0$,*

$$\sum_{S\subseteq[k]} c_{S,F} u^{k-|S|} v^{|S|} \le (u+v)^{k-1}(u + 2^{\nu_2(\det A)} v). \tag{4}$$

Proof. Integer unimodular column operations triangularize A: $B = AV$ is upper triangular. This follows recursively by the Euclidean algorithm on the last row, leaving only its last entry nonzero and then treating the upper-left submatrix. If $e_i = \nu_2(B_{ii})$, then $e_i \ge 0$ and $\sum_i e_i = \nu_2(\det A)$.

For a row subset S, the principal minor $B[S,S]$ has valuation $\sum_{i\in S} e_i$. By Cauchy–Binet,

$$\det B[S,S] = \sum_{|J|=|S|} \det A[S,J] \det V[J,S].$$

If every nonzero A-minor in this sum had larger valuation, the sum would too. Hence $c_{S,F} \le 2^{\sum_{i\in S} e_i}$, and the left side of (4) is at most $\prod_i (u + 2^{e_i} v)$. For $a, b \ge 1$,

$$(u+v)(u+abv) - (u+av)(u+bv) = uv(a-1)(b-1) \ge 0.$$

Repeatedly combine two exponents into one. Their total becomes $\nu_2(\det A)$, proving (4) for arbitrary dimension, without an independence assumption on EHC coordinates. □

Lemma 3.5 (Forest and Toeplitz tail)**.** *For different fixed coordinate sequences with a common forest shape of maximum height h, the probability that all ordered roots agree is at most $1-(1-\varepsilon)^h \le h\varepsilon$. After fresh final NH, each prescribed output difference has probability at most*

$$a_h = 1 - (1-\varepsilon)^{h+1} \le (h+1)\varepsilon.$$

For different equally supported padded tails, k whole-pair-shift Toeplitz outputs have joint difference probability at most ε^k.

Proof. Select a tree with different leaf sequences. At each height, condition lower keys and, if a difference survives, select a differing child tuple. The fresh key at that height merges it with probability at most ε. Thus a difference survives to a root with probability at least $(1-\varepsilon)^h$. Using only one selected tuple per height permits key reuse among nodes at that height. If the root lists differ, fresh final NH costs ε for any target; if they agree, bound the event by one. This gives a_h. With blocks, fix all but one differing scalar position's independent final key before applying NH.

For the tail, choose its last differing NH pair. In the last shifted output, its key lies beyond every key relevant to earlier output differences; identical trailing pairs cancel. Condition all other keys and use the one-pair NH bound on an exposed half-key. Remove that output and repeat backwards through the shifts. This gives one factor ε per output, despite overlap of the full key windows. □

Theorem 3.6 (Repaired collision bound)**.** *Under the construction, distance, matrix, and key assumptions above, distinct equal-length inputs with at least one complete EHC group satisfy*

$$\Pr[H(x) = H(y)] \le \min\{1,\ 2^{-32k}(h+2)^{k-1}(h+1+2^t)\}. \tag{5}$$

The same bound holds for every prescribed vector difference. If no complete group exists, or if the equally supported raw tails differ, 2^{-32k} suffices.

Proof. For equal tails, select a differing group and let I_j indicate equality of its complete EHC block in coordinate j. Conditional on all EHC keys, an identical whole coordinate sequence implies $I_j = 1$. A differing sequence costs at most $\rho = (h+1)\varepsilon$ by Lemma 3.5. The remaining tree and final key sets are independent across coordinates. If $\rho \ge 1$ the clipped result is immediate. Otherwise the conditional product and expansion give

$$\begin{aligned}\Pr[H(x) - H(y) = z] &\le \mathbb{E}\prod_{j=1}^{k}(\rho + (1-\rho)I_j)\\ &\le \sum_{S\subseteq[k]} c_{S,F}\rho^{k-|S|}\varepsilon^{|S|}\\ &\le (\rho+\varepsilon)^{k-1}(\rho+2^t\varepsilon),\end{aligned}$$

using (3) and Lemma 3.4. We dropped only factors $(1-\rho)^{|S|} \le 1$. The I_j need not be independent.

For different tails, condition on every non-tail key. Their required difference is a fixed vector and Lemma 3.5 supplies ε^k. With no groups, distinct equal-length inputs have different padded tails by injectivity of padding at a fixed length. These are disjoint cases; no tail error is added to the forest bound. □

The intended matrices have maximal determinant valuations $t = 2, 2, 3, 3$ for $k = 2, 3, 4, 5$, respectively. Exact arithmetic checks all $21 + 84 + 210 + 126 = 441$ maximal minors; none vanishes. The resulting bounds are

Output bytes	(d, e, k)	t	Bound with a group
16	$(6, 7, 2)$	2	$(h+2)(h+5)2^{-64}$
24	$(7, 9, 3)$	2	$(h+2)^2(h+5)2^{-96}$
32	$(7, 10, 4)$	3	$(h+2)^3(h+9)2^{-128}$
40	$(5, 9, 5)$	3	$(h+2)^4(h+9)2^{-160}$

For $k = 3$, the subset coefficient sums are at most 6, 9, and 4 at subset sizes one, two, and three, agreeing with $(u+v)^2(u+4v)$. At $h = 16$, Theorem 3.6 gives

$$6804 \cdot 2^{-96}, \qquad 96 - \log_2(6804) = 83.2678325743\ldots. \tag{6}$$

This is a pointwise collision guarantee. It is distinct from the word-normalized score in Section 5.

4 The shipped 24-byte core is a refuted variant

The implementation considered here is the saved header at commit `caf7924ceab4721f4e0cc33442b185558ba7f1c4`. The entry points `advanced::V1<3>`, `V2<3>`, `V3<3>`, and `V4<3>` return three 64-bit words and reach `Encode3`. The ordinary Style wrappers instead use output width two. A Style number such as 128 describes the block width, not the hash output width.

4.1 Distance one and a fixed equal-length witness

Write raw blocks as $r_0, r_1, \ldots$, each with b scalar words. In the shipped encoder, the inner distribution function captures its row iterator by value. Changes to the outer iterator do not change the blocks read by the function; the increments also use a single-block stride. Symbolic XOR expansion gives the two added packets

$$J(x)_7 = (r_0, r_1, r_2), \qquad J(x)_8 = (r_0 \oplus r_1, r_1 \oplus r_2, r_2 \oplus r_3). \tag{7}$$

The first loop repeats the frozen triple 26 times, so its added terms cancel. All parity information is confined to the first four raw blocks. Changing r_6 changes only systematic packet two. Systematicity excludes distance zero, so `Encode3` has distance exactly one. The same fixed-capture defect gives distance one for the shipped `Encode4` and `Encode5`.

Proposition 4.1 (Equal-length lower bound)**.** *For each $b \in \{1, 2, 4, 8\}$ there are two fixed distinct strings of length $168b$ bytes whose shipped three-word core collision probability is at least 2^{-32}.*

Proof. Take the all-zero string and change only word number $6b$ (zero based) to the integer one, equivalently byte $48b$ to `01`. Only the first scalar position of raw block six changes, and by (7) only one encoded packet differs. Write its responsible key word as $A + QZ$. The NH difference is

$$\Delta = \begin{cases} Z, & A < Q - 1, \\ -(Q-1)Z \pmod{R}, & A = Q - 1. \end{cases} \tag{8}$$

Both coefficients of Z are odd, so $\Delta = 0$ if and only if $Z = 0$. Column two of T_3, numbered from zero, is $(1, 0, 1)^{\mathsf{T}}$. This is therefore exactly the EHC-collision condition for the witness. On that event all subsequent stages receive identical data and the full three-word core collides.

Exactly Q of the Q^2 values of the responsible key word have $Z = 0$, independently of every other key. This sufficient event has exact probability $1/Q = 2^{-32}$. Later stages may introduce additional collisions; the conclusion for the complete core is a lower bound. □

The witness lengths are 168, 336, 672, and 1344 bytes. Common additional groups propagate the same sufficient event at longer supported equal lengths. This contradicts the intended distance-three core guarantee at a single group by many orders of magnitude. It is a mathematical key-fibre count, independent of observed frequencies in execution tests.

4.2 Length-free padding

The shipped advanced cores do not encode byte length. Empty input and the one-byte string `00` both produce no full groups and the same zero-padded raw block, including the extra padded block inserted at an aligned or empty remainder. Their complete core outputs agree for every key. Thus an unrestricted unequal-length theorem for those raw cores has collision bound one. The Style wrappers distinguish lengths by tabulation; the repaired three-word core in Section 6 uses a terminal length word.

5 The Style wrappers retain a 63-bit bound

5.1 Distance-two compression and sharper NH estimates

`Encode2` retains six symbols and appends their XOR. A single changed symbol changes itself and the parity; two changed symbols already give distance two. This proves distance exactly two, also on each scalar position within a block. Its matrix is

$$T_2 = \begin{pmatrix} 1 & 0 & 1 & 1 & 2 & 1 & 4 \\ 0 & 1 & 1 & 2 & 1 & 4 & 1 \end{pmatrix}. \tag{9}$$

The general minor method gives singleton sum at most five and joint coefficient four, yielding $(h+2)(h+5)\delta$. A property of integer NH improves this bound.

Lemma 5.1 (Truncated NH)**.** *Every prescribed NH difference modulo 2^{62} or 2^{63} has probability at most 2ε for distinct equal-length pair strings.*

Proof. Expose a differing pair as in Lemma 3.1, now modulo $M = Q^2/4$. If a nonzero half-input difference has valuation at most 30, choose it for the conditioned factor. On each wrap interval the other half-key has slope with gcd with M at most $Q/4$. Solutions are separated by at least $M/(Q/4) = Q$, giving at most one per interval, or two altogether.

Otherwise the two half-input differences lie in $\{0, Q/2\}$, with one nonzero. Select $Q/2$ for the conditioned factor. The solution period is $Q/2$. If the other difference is zero, a single interval of length Q permits at most two solutions. If it is $Q/2$, the two intervals each have length $Q/2$, each permitting at most one. Thus at most two of Q half-keys work. A prescribed residue modulo 2^{63} also prescribes one modulo 2^{62}. □

For two selected differing symbols and their output differences D_1, D_2, the literal matrix gives

$$\Pr[D_1 = 0] + \Pr[D_2 = 0] \le 3\varepsilon, \qquad \Pr[D_1 = D_2 = 0] \le 2\varepsilon^2. \tag{10}$$

An odd selected coefficient in a row costs ε; a row with only even selected coefficients costs at most 2ε by Lemma 5.1. At least one row has an odd coefficient. For the joint bound, determinants of valuation at most one use Lemma 3.2. The only valuation-two column pairs are $\{0, 5\}$ and $\{1, 6\}$, triangular systems containing a coefficient four. That equation costs 2ε modulo 2^{62}, and the other fixes an independently keyed NH difference, costing ε. This proves (10) for arbitrary fixed offsets as well.

Set $s_h = (1-\varepsilon)^{h+1}$ and $a_h = 1 - s_h$. Retaining survival factors in the conditional product gives

$$B_2(h) = (a_h + s_h\varepsilon)(a_h + 2s_h\varepsilon) \le (h+2)(h+3)2^{-64}. \tag{11}$$

Indeed the expansion is $a_h^2 + a_h s_h \mathbb{E}(I_1 + I_2) + s_h^2 \mathbb{E}(I_1 I_2)$, and (10) bounds its last two terms. With no group, or with different equally supported raw tails, the sharper δ bound applies.

5.2 The forest and its supported length domain

The implementation promotes a full level before inserting the next leaf. For $n > 0$ groups the live-root counts are its bijective base-eight digits: emit $1 + (n-1) \bmod 8$ and replace n by $\lfloor (n-1)/8 \rfloor$ repeatedly. Induction on insertion proves this representation, with digits in $\{1, \dots, 8\}$ and no zero holes. The maximum live height is

$$h(n) = \lfloor \log_8(7n+1) \rfloor - 1.$$

A nine-entry stack is retained, and its finalizer requires an in-array zero sentinel. Therefore set

$$C = 8 + 8^2 + \cdots + 8^8 = 19{,}173{,}960. \tag{12}$$

At $n = C$ the digits are eight eights followed by zero; at $n = C + 1$ they are nine ones, leaving no sentinel. The supported contiguous domain is $n \le C$, with $h \le 7$ and at most 64 roots. A Style group uses $144b$ bytes, so:

Wrapper	Exclusive byte limit	Last core key index
Style64	2,761,050,384	805
Style128	5,522,100,768	952
Style256	11,044,201,536	1246
Style512	22,088,403,072	1834

All four wrappers output 64 bits. No theorem here extends the shipped forest beyond these domains.

5.3 Conditioning with the actual overlapping keys

Number the caller's entropy words $K[0], K[1], \dots$. The eight length-byte tables occupy $[0, 2048)$, and the sixteen core-output byte tables occupy $[2048, 6144)$. The core begins at index 512: EHC uses $[512, 533)$, tree storage is reserved through index 658, and final NH begins at 659. For r live roots, root keys use $2br$ words; the tail needs at most $19b$ distinct words, including its always-inserted padded block and the output shift. Every core read is therefore below

$$659 + b(2 \cdot 64 + 19) \le 1835 < 2048.$$

The overlap is entirely with the length tables. The output tables remain fresh after all lower words have been exposed.

Let C_0 denote equality of the complete two-word core signatures, and D the XOR difference of the length-table terms. Condition on all words below 2048, fixing these two objects jointly. If the core signatures differ, one differing byte gives a fresh output-table entry, so equality has probability exactly δ, regardless of D. If the signatures agree, all output-table terms cancel and equality holds exactly when $D = 0$. Thus

$$\Pr[H(x) = H(y)] = \delta \Pr[\neg C_0] + \Pr[C_0 \cap \{D = 0\}]. \tag{13}$$

At equal lengths this equals $\delta + (1 - \delta) \Pr[C_0]$. At unequal lengths, marginally $\Pr[D = 0] = \delta$ by ordinary tabulation of two distinct fixed length encodings. Since $\Pr[C_0] \ge \Pr[C_0 \cap \{D = 0\}]$, equation (13) gives

$$\Pr[H(x) = H(y)] \le 2\delta - \delta^2. \tag{14}$$

No independence between core signatures and length terms is required.

Theorem 5.2 (Sharp normalized Style coefficient)**.** *Let $L \geq 1$ be an integer word cap and let $x \neq y$ each have at most $8L$ bytes and belong to the supported domain of a fixed Style wrapper. Under the stated flat-address modular-arithmetic model,*

$$\Pr[H(x) = H(y)] \leq L(2^{-63} - 2^{-128}). \tag{15}$$

The coefficient is optimal over these caps. Its length-normalized score, with output floor 2^{-64}, is $64 - \log_2(2 - 2^{-64})$, conservatively 63 bits.

Proof. Unequal lengths satisfy (14), which is no larger than (15). Equal lengths below one group have core bound δ, giving the same expression. At height $h \geq 0$, use (11) with $c_h = (h+2)(h+3)$. The group count implies

$$L \geq 18b\frac{8^{h+1} - 1}{7}.$$

At $h = 0$, $c_h + 1 = 7 \leq 18b$. Successive height thresholds grow by more than eight, while $c_h + 1$ grows by less than eight. Therefore $c_h + 1 \leq L$ at every such length. The equal-length probability is at most $\delta(1 + (1 - \delta)c_h) \leq L\delta \leq L(2 - \delta)\delta$.

For optimality, compare the one-byte strings `00` and `01`. Only the first scalar word of one raw pair differs. In each of the two core outputs its NH difference is (8), vanishing exactly when the high half of its key is zero. The two relevant key words, at indices 659 and $659 + b$, are independent. Core equality therefore has probability exactly $\varepsilon^2 = \delta$. The lengths agree and the output tables are independent, so (13) gives exactly $2\delta - \delta^2 = 2^{-63} - 2^{-128}$. This attains the coefficient at $L = 1$. □

The theorem is not a constant 2^{-63} collision bound at every length. It is the stated word-cap guarantee. The literal shipped table view is undersized for its later row accesses, and some scalar sums use signed arithmetic. Consequently the shipped result is about the specified flat-address modular function; it is not an unconditional portable ISO C++ theorem. The fixed header corrects the array representation, unsigned sums, and NEON dispatch while preserving the Style function's intended arithmetic[1].

6 The repaired encoder and an all-length three-word theorem

The corrected header, `HalftimeHash24-fixed.hpp`, retains three 64-bit outputs. Its encoder iterator refers to a whole three-block row, advances by one row, and is captured by reference by the distribution function. The first parity loop stops at the end of the seven data rows, so parity rows are never read as data. The remaining arithmetic uses flat key-array indexing and explicitly unsigned horizontal sums[1].

6.1 Distance three of the corrected encoder

The repaired `Encode3` has systematic symbols $x_0, \ldots, x_6 \in (\mathbb{F}_2)^3$ at each individual bit position. Its two parity symbols are

$$P = \bigoplus_{i=0}^{6} x_i, \qquad Q' = \bigoplus_{i=0}^{6} A_i x_i,$$

where the binary matrices A_i have the following column encodings (a column (a, b, c) is the integer $a + 2b + 4c$):

$$(1,2,4),\ (4,5,2),\ (5,7,6),\ (2,6,5),\ (3,4,1),\ (6,3,7),\ (7,1,3). \tag{16}$$

Each A_i and each $A_i + A_j$ for $i \neq j$ is invertible; direct elimination checks the seven plus 21 binary 3×3 matrices.

Lemma 6.1. *The repaired encoder has minimum packet distance exactly three, for every block width.*

Proof. One changed input packet changes itself, P, and Q' because its A_i is invertible. With exactly two changed input packets, either their differences have nonzero XOR and P changes, or they are equal to the same nonzero vector and Q' changes by $(A_i + A_j)v \neq 0$. Three or more changed systematic packets already suffice. This proves distance at least three; a single changed packet attains three. The XOR identities act separately at every bit and scalar position, so selecting a differing position proves the statement for arbitrary blocks. □

The exact source certificate also extracts the generator from the actual header by inserting distinct basis bits, then checks every subset of surviving packets by GF(2) Gaussian elimination. A survivor contributes all three of its rows. For `Encode3`, every seven-packet survivor set has rank 21; all 512 subsets are recorded, including the 36 critical seven-packet subsets. If a nonzero codeword had at most two nonzero packets, erasing them would leave a full-rank survivor set with zero output, a contradiction. The corresponding repaired encoder certificates are:

Encoder	Data	Encoded	Input rank	Subsets	Distance
Encode2	6	7	18	128	2
Encode3	7	9	21	512	3
Encode4	7	10	21	1024	4
Encode5	5	9	15	512	5

These are exact rank certificates, not sampled tests of selected messages.

6.2 Fixed key layout and the terminal length word

For the corrected three-word core, all indices below are word offsets relative to its key pointer. EHC occupies $[0, 27)$, reserved tree keys $[27, 216)$, and the forest-finalizer keys $[216, 216 + 192b)$, reserving all 64 possible roots and all three output coordinates. A separate Toeplitz pool starts at $A = 216 + 192b$.

If the remainder has $r < 168b$ bytes, it is padded into $t = \lfloor r/(8b) \rfloor + 1 \leq 21$ raw blocks, including an extra zero-padded block even when aligned. These blocks are right aligned in the tail pool: their first key position is $A + (21 - t)b$. Output $j \in \{0, 1, 2\}$ shifts its keys by jb whole NH pairs. After the scalar-position sum, append one scalar NH term containing the original total byte length, with key

$$K[A + q_0 + jb], \qquad q_0 = 21b. \tag{17}$$

The length word is scalar, not broadcast across the block; there are still only three outputs. Modular addition permits this term to be applied after the horizontal sum.

Lemma 6.2 (Terminal length separation)**.** *For unequal lengths in the supported byte domain, every prescribed difference of the repaired three-word core has probability at most* 2^{-96}.

Proof. The lengths are distinct 64-bit integers, so at least one 32-bit half differs. For output j, write its difference as a residual plus

$$\mathrm{NH}(|x|, K[A + q_0 + jb]) - \mathrm{NH}(|y|, K[A + q_0 + jb]).$$

All data-tail indices in output i are at most $A + q_0 - 1 + ib$, and its length key is $A + q_0 + ib$. Thus every earlier output $i < j$ ignores the length key of output j, and output j's residual ignores its own length key. The forest pool is disjoint for both messages even if their root counts differ.

Fix, in each length key, the half associated with a differing length half. Expose the other halves in output order. Lemma 3.1 permits at most one of Q values for each target equation after

earlier halves are known; earlier equations ignore the next half. The triangular count is at most $Q^{-3} = 2^{-96}$. No independence of residual outputs is required, and absent leading data-tail terms are simply absent, rather than being replaced by NH of zero. □

Theorem 6.3 (Repaired three-word core)**.** *Let $b \in \{1, 2, 4, 8\}$ and C be (12). For distinct inputs satisfying*

$$0 \le |x|, |y| < 168b(C + 1),$$

the fixed core has, for every vector target, difference probability at most 2^{-96} when lengths differ, when equal-length padded tails differ, or when equal lengths have no complete group. In every remaining case, for a common forest-height upper bound h,

$$\Pr[H(x) - H(y) = v] \le \min\{1, (h + 2)^2(h + 5)2^{-96}\}. \tag{18}$$

In particular it has the 83.27-bit guarantee (6). On its actual retained stack $h \le 7$, the stronger uniform bound is $972 \cdot 2^{-96}$, or 86.0751874964... collision bits.

Proof. Unequal lengths use Lemma 6.2. At equal lengths the terminal terms cancel, and the two tails have the same right alignment and support. If the tails differ, use the Toeplitz part of Lemma 3.5; below one group, distinct inputs necessarily have different tails. Otherwise select a differing group and a scalar position where it differs. Lemma 6.1 supplies three differing encoded symbols there. The unchanged T_3 matrix has nonzero minors and maximum valuation two. Conditional on the EHC keys, the three tree/finalizer pools are independent. Theorem 3.6 therefore gives $(h + 2)^2(h + 5)2^{-96}$, also for a nonzero target. The actual forest schedule is fixed by public length and has height at most seven. Substituting $h = 7$ or the conservative abstract value 16 proves the displayed constants. □

An actual height-16 tree cannot be processed by this retained stack; its 83.27-bit label is conservative on the supported domain. The fixed code does not expand that domain or add a runtime length check.

For the 192-bit output floor, the corrected core has normalized score exactly 96 bits under the word-cap convention. One valid envelope is 2^{-96} below $L = 21b$, and otherwise $(h+2)^2(h+5)2^{-96}$ with $h = \min(7, \lfloor \log_8 \lfloor L/(21b) \rfloor \rfloor)$. At the first group, $L/((h + 2)^2(h + 5)) = 21b/20 \ge 1$. Later thresholds grow by eight while the coefficient grows by at most $54/20 < 8$; within each plateau the score increases. The one-byte pair `00`, `01` attains collision probability 2^{-96}: its equal length terms cancel and the three distinct data-key high halves must be zero. This normalized score is consistent with the separate 83.27-bit pointwise statement.

The fixed three-word allocation is exactly the prefix of $216 + 215b + 1$ words: 432, 647, 1077, or 1937 words for the four widths. These extents include reserved holes and the terminal keys; unused words integrate out of the probability experiment. The fixed Style wrappers keep Encode2 and their length-byte tables, preserve intended output values, and satisfy Theorem 5.2. Their flat-array and unsigned-arithmetic repairs support the stated C++ object contract without changing that mathematical function.

7 Machine-checked statements and implementation boundary

The machine-checked statements page[2] reports successful Lean 4.24.0 builds and an audit of 237 new proof declarations plus 57 inherited declarations. The reported theorems use only subsets of `propext`, `Classical.choice`, and `Quot.sound`; finite integer certificates use kernel-checked reduction and arithmetic. The following distinctions are material.

- Integer NH, its wrap argument, the correctly scoped passed-through accumulator, and the 2ε truncated-output theorem are checked.

- The general-dimensional matrix-fibre theorem and corrected EHC theorem are checked. For T_2, T_3, all required minors and subset coefficient bounds are checked. The general arbitrary-dimension projection-polynomial inequality of Lemma 3.4 is a paper proof; the concrete $k = 2, 3$ inequalities needed by the checked constructions are formalized.
- The scalar abstract forest, whole-pair-shift Toeplitz tail, and the $6804 \cdot 2^{-96}$ theorem are checked. The byte-input Style model additionally includes word packing, broadcast keys, the sharp $B_2(h)$ bound, flat key addresses, overlap conditioning, and the instantiated normalized probability theorem `modeledStyleHash_normalized`.
- The added `terminal_length_three_bound` in `FixedLength.lean` checks the terminal-word argument using the same NH and triangular-counting theorems. Its key-index residual condition is discharged for the fixed header by the explicit inequalities preceding Lemma 6.2.

The fixed encoder's source-derived ranks and its address/schedule correspondence are exact certificates and paper arguments. Neither the original abstract distance-three theorem nor the new terminal-word theorem is a complete C++-to-Lean refinement. In particular the abstract model's public schedule conditions are proved mathematically for the stated promotion schedule, but the entire C++ loop, its machine instructions, compiler transformations, object lifetimes, and all optimized paths are not machine-verified by these files. The optimality witnesses for the Style coefficient and the repaired core's score are paper proofs, not reported Lean milestones.

These limits concern the type of evidence. The repaired mathematical bounds, the shipped distance-one lower bound, and the supported-domain Style theorem are unconditional within their explicitly stated ideal-key arithmetic models. No result here asserts a collision guarantee for the benchmark's PRNG-derived key distribution or for out-of-domain calls.

A Matrices and exact reproduction

The three-word combine matrix is

$$T_3 = \begin{pmatrix} 0 & 0 & 1 & 4 & 1 & 1 & 2 & 2 & 1 \\ 1 & 1 & 0 & 0 & 1 & 4 & 1 & 2 & 2 \\ 1 & 4 & 1 & 1 & 0 & 0 & 2 & 1 & 2 \end{pmatrix}.$$

For completeness the higher-dimensional matrices used in the general specialization table are

$$T_4 = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 4 & 2 & 4 & 1 & 1 \\ 0 & 1 & 2 & 0 & 0 & 1 & 1 & 2 & 4 & 1 \\ 2 & 0 & 1 & 0 & 4 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 4 & 1 & 2 & 8 \end{pmatrix}, \quad T_5 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 2 & 4 \\ 0 & 1 & 0 & 0 & 0 & 1 & 2 & 1 & 7 \\ 0 & 0 & 1 & 0 & 0 & 1 & 3 & 8 & 5 \\ 0 & 0 & 0 & 1 & 0 & 1 & 4 & 9 & 8 \\ 0 & 0 & 0 & 0 & 1 & 1 & 5 & 3 & 9 \end{pmatrix}.$$

Maximal-valuation examples, with zero-based columns, are $(0, 5)$ with determinant four in T_2, $(0, 1, 3)$ with determinant 12 in T_3, $(0, 1, 5, 6)$ with determinant -24 in T_4, and $(0, 1, 2, 4, 8)$ with determinant -8 in T_5. All maximal minors, the minimum nonzero subset costs (2), and the binary matrices in (16) can be checked with exact integer elimination; floating-point determinants are unnecessary.

All following paths are relative to the post's `records/` directory. The implementation and accompanying theorem are `halftime/HalftimeHash24-fixed.hpp` and `halftime/THEOREM.html`. The exact source certificate map is:

Path within `halftime/certificates/`	Role
`src/matrix.cpp`, `src/rank.py`	Extract the actual encoder generator and check every surviving-packet subset over $\mathbb{F}_2$; the full output is `rank-certificate.json`.
`src/score.py`	Exact supported-domain, key-layout, and word-cap comparisons; results in `scores-layout.json`.
`src/compare_style.py`	Compare intended Style outputs across the source repairs.
`src/check_objects.py`	Check the fixed flat-array and unsigned-arithmetic source conditions.
`src/verify-xeon.sh`, `src/final-xeon.sh`	Build and execute the implementation certificates and collect their reports.
`lean/FixedLength.lean`, `lean/FixedLength.log`	The terminal-length theorem and its axiom report.
`lean/base/`	Pinned mathematical proof development, toolchain, matrix certificates, and audits.

The fixed-header rank extractor tests the actual XOR maps rather than an independently designed replacement. Its all-subset ranks, together with the proof following (16), establish encoder distance. Guard-page checks exercise the reported exact extents at short lengths, group boundaries, and forest promotions; they supplement the symbolic address map and do not replace it. `verification-summary.json` records the archived execution results and source provenance.

The LaTeX source embeds its BibTeX database and compiles with `latexmk -pdf`. The paper's theorems use the proved counting arguments and exact certificates; finite execution comparisons are evidence of source correspondence, rather than an extrapolation to every input or every compiler.

References

[1] Thomas Dybdahl Ahle. The fixed HalftimeHash24 theorem. https://thomasahle.com/blog/adversarial-examples-for-hashes/records/halftime/THEOREM.html, September 2026. Implementation and exact certificates in the same directory.

[2] Thomas Dybdahl Ahle. HalftimeHash formalization: integer arithmetic, matrices, forests, and byte input. https://thomasahle.com/blog/adversarial-examples-for-hashes/records/machine-checked.html, September 2026.

[3] Jim Apple. HalftimeHash: Modern hashing without 64-bit multipliers or finite fields. arXiv:2104.08865v2, 2021.

[4] John Black, Shai Halevi, Hugo Krawczyk, Ted Krovetz, and Phillip Rogaway. UMAC: Fast and secure message authentication. In *Advances in Cryptology—CRYPTO '99*, volume 1666 of *Lecture Notes in Computer Science*, pages 216–233. Springer, 1999.

[5] Paul Khuong. UMASH: a fast almost universal 64-bit string hash. https://github.com/backtrace-labs/umash/blob/master/umash.pdf, 2022. Literate whitepaper, 23 February 2022.