Computing Almost Exact Probabilities of Differential Hash Collision Paths by Applying Appropriate Stochastic Methods (v. January 14, 2008)

Max Gebhardt, Georg Illies and Werner Schindler

Bundesamt für Sicherheit in der Informationstechnik (BSI)
Godesberger Allee 185–189
53175 Bonn, Germany
{Maximilian.Gebhardt,Georg.Illies,Werner.Schindler}@bsi.bund.de

Abstract. Generally speaking, the probability of a differential path determines an upper bound for the expected workload and thus for the true risk potential of a differential attack. In particular, if the expected workload seems to be in a borderline region between practical feasibility and non-feasibility it is desirable to know the path probability as exact as possible.
We present a generally applicable approach to determine at least almost exact probabilities of differential paths where we focus on (near-)collision paths for Merkle-Damgard-type hash functions. Our results show both that the number of bit conditions provides only a rough estimate for the true path probability and that the IV may have significant impact on the path probability. For MD5 we verified the effectivity of our approach experimentally. An abbreviated version [GIS4], which in particular omits proofs, technical details and several examples, will appear in the proceedings of the security conference 'Sicherheit 2008'.

Keywords: Hash function, collision path, postaddition, probability, stochastic model.

1 Introduction

The efficiency of differential attacks on cryptographic primitives (block ciphers, stream ciphers, hash functions etc.) is closely related to the probability that pairs of intermediate values follow a particular differential path. From the designer's point of view the efficiency of an attack implies its risk potential. Hence it is clearly desirable to know the probabilities of differential paths as exact as possible, especially if the estimated path probability implies a workload which appears to be "between" practical feasibility and infeasibility.

Primarily, one is interested in conditional probabilities

$$\text{Prob}\big((X_n, X'_n) \in B_n \mid (X_0, X'_0) \in B_0\big). \tag{1}$$

Usually, the exact computation of such probabilities is practically infeasible. Instead, one usually considers the probability of a particular differential path

$$\mathrm{Prob}\big((X_n, X'_n) \in B_n, (X_{n-1}, X'_{n-1}) \in B_{n-1}, \ldots, (X_1, X'_1) \in B_1 \,|\, (X_0, X'_0) \in B_0\big) \quad (2)$$

which provides a lower bound for (1). (Note that different differential paths might end in the set B_n.) Here $X_0, \ldots, X_n, X'_0, \ldots, X'_n$ denote random variables that assume values on a finite set Ω (typically, $\Omega = \{0,1\}^v$) while the subsets $B_0, \ldots, B_n \subseteq \Omega \times \Omega$ characterize conditions that define the differential path. The conditional probability (2) equals a product of conditional probabilities

$$\mathrm{Prob}\big((X_i, X'_i) \in B_i \,|\, (X_{i-1}, X'_{i-1}) \in B_{i-1}, \ldots, (X_0, X'_0) \in B_0\big) \text{ for } i \in \{1, \ldots, n\}. \quad (3)$$

Usually, due to their long 'history' also these conditional probabilities cannot be computed exactly, in particular since the pairs $(X_i, X'_i), (X_{i-1}, X'_{i-1}), \ldots$ are usually not independent, at least not in a strict sense, which causes further computational difficulties. Moreover, the random variables X_i and X'_i are strongly correlated, which complicates concrete calculations additionally unless $|\Omega|$ is very small or the sets B_i are extremely simple. For these reasons the conditional probabilities (3) usually are only roughly estimated. For hash collision paths the subsets B_i typically define conditions on particular bits, and $2^{-(\#\text{affected bits})}$ serves as an approximator for the unknown conditional probability (3).

Generally speaking we propose to study 'primitives'

$$\mathrm{Prob}\big((Z, Z') \in B_3 \mid (X, X') \in B_1, (Y, Y') \in B_2\big) \quad (4)$$

that are tailored to the real-world problem. (Depending on the concrete problem it may be necessary to considered a longer history.) If the range of thses random variables is small (4) can be determined by exhaustion. For hash functions X, X', Y, Y', Z, Z' typically assume values in $\{0,1\}^{32}$ or $\{0,1\}^{64}$, which requires more sophisticated methods.

The understanding of suitable primitives can help to simplify conditional probabilities (3) of random vectors assuming values on the product space $\Omega \times \Omega$ with strongly correlated components to conditional probabilities on Ω, which clearly is an enormous advantage. Another goal is to find sufficient conditions that the conditional random variable $(Y_3, Y'_3) \mid B_3$ is independent of $(Y_2, Y'_2) \mid B_2$. Depending on the concrete situation this may allow to reduce the relevant part of the 'history' in (3), which also simplifies calculations.

Hash functions are used by many cryptographic applications. Strong hash functions should meet the one-way property and the second pre-image property. Many applications (as digital signatures) additionally demand that the hash function shall be *collision resistant*, i.e. it shall not be feasible in practice to find bit strings $M \neq M'$ with identical hash values.

In [WLFCY], [WY] and [WYuY] efficient collision search methods are described for the hash functions HAVAL, RIPEMD, MD4 and MD5 and SHA-0; for improvements see [St], [SNKO], [LiLa], [Kli1], [Kli2], [BCH] , [SLW], [DR]. In [WYiY] a collision attack on SHA-1 is sketched with a predicted workload

of 2^{69} SHA-1 calculations (or more precisely: the workload shall be equivalent to the calculation of 2^{69} hash values) and [WYaYa] announce an improvement with a workload of only 2^{63} SHA-1 calculations. For a reduced SHA-1 version (70 instead of 80 steps) [DMR] presents a collision with workload of 2^{44} hash calculations. In [SLW] and [DR] collision search methods for differing prefixes have been developed.

The core of any dedicated hash function $H\colon \{0,1\}^* \to \{0,1\}^t$ is the *compression function*

$$h : \{0,1\}^t \times \{0,1\}^s \longrightarrow \{0,1\}^t. \tag{5}$$

The compression function h itself consists of a large number of elementary step functions $h_i : \{0,1\}^t \times \{0,1\}^s \to \{0,1\}^t$ (which can be processed efficiently on 32-bit architectures) and a final modular addition of 32-bit words (*postaddition*). All the widely-used dedicated hash functions are of Merkle-Damgard type where the hash value $H(M)$ is computed as follows: According to a specified padding scheme the message M is first expanded to a bit string whose length is a multiple of s bits. The extended bit string is segmented into non-overlapping s-bit blocks: $m_{(1)}||m_{(2)}||\ldots||m_{(r)}$. Beginning with a (fixed) initialization vector IV one iteratively computes

$$h_1 := h(IV, m_{(1)}), h_2 := h(h_1, m_{(2)}), \ldots, h_{j+1} := h(h_j, m_{(j+1)}), \ldots \tag{6}$$

Finally, $H(M) := h_r$. A *2-block collision* is a pair $(m_{(1)}||m_{(2)}, m'_{(1)}||m'_{(2)})$ with $m_{(i)}, m'_{(i)} \in \{0,1\}^s$ and $(m_{(1)}||m_{(2)}) \neq (m'_{(1)}||m'_{(2)})$ with

$$h_2 = h(h(IV, m_{(1)}), m_{(2)}) = h(h(IV, m'_{(1)}), m'_{(2)}) = h'_2 \quad . \tag{7}$$

Note that appending arbitrary blocks $m_{(3)}, m_{(4)}, \ldots$ to both $m_{(1)}||m_{(2)}$ and $m'_{(1)}||m'_{(2)}$ clearly implies $h_j = h'_j$ for all $j \geq 2$. All the attacks mentioned above aim at 2-block collisions. Generically, two-block collision search algorithms (as in [WY] and [WYiY], for instance) work as follows:

1. (first block) Pairs of blocks $(m_{(1)}, m'_{(1)})$ are generated in a specific manner (a part of this procedure is referred to as "message modification") until a pair of chaining value $(h_1 := h(IV, m_{(1)}), h'_1 := h(IV, m'_{(1)}))$ is found that is at least 'similar', satisfying specified conditions.
2. (second block) Depending on the chaining values (h_1, h'_1) message blocks $(m_{(2)}, m'_{(2)})$ are generated in a specific manner until one pair satisfies (7).

[WY] and [WYiY] list sets of conditions on the intermediate results during the step-by-step calculation of the value $h_2 = h(h(IV, m_{(1)}), m_{(2)})$ (called 'sufficient conditions' in [WY] and [WYiY]). If these conditions are fulfilled h_1 and h'_1 meet specific bit conditions, and finally $h_2 = h'_2$. Additionally, these conditions (shall) ensure that intermediate values follow a specified differential path (a so-called collision path, resp. a near-collision path for the first block). A more precise definition of the term 'differential path' will be given later. Basically it is a combination of a 32-bit word modular differential path and a kind of signed XOR differential path.

We point out that, at least for fixed differential schemes (cf. Sects. 2 and 4), the IV may influence the success probability considerably ($\rightarrow$ postadditions). This phenomenon was first quantified in [GIS2] (and almost at the same time qualitatively mentioned in [St]) although it is non-negligible. The impact of the IV may be relevant for 'prefix' attacks as described in [DL] and [GIS1].

In Section 3 we prove three technical theorems that will turn out to be very useful later. In Section 4 the effectiveness of our approach is confirmed by practical experiments with three near-collision paths (specified in the appendix) for the MD5 hash function. Based on a stochastic model with mild assumptions on the mixing properties of the MD5 step function the before-mentioned theorems on the primitives are applied. The 'theoretically' derived path probabilities matched with empirical results. Compared with the 'straight-forward' approximators for the path probabilities (obtained by 'classical' bit counting) we obtained non-negligible 'correction factors' between 1/12 and 5, which in turn imply 'correction factors' between 1/5 and 12 on the expected workload of the collision attack.

We mention that our approach can be adjusted to compute the actual expected workload (e.g.) for specific SHA-1 collision paths (cf. Subsect. 4.4), for instance. In this case 'correction factors' of, let's say, one or two (positive or negative) powers of 2 were surely relevant.

Reference [GIS3] is a pre-version of this paper. An abbreviated version [GIS4], which omits technical details and all proofs from Section 3 and several examples from Section 4 but focuses on the understanding and the application of our approach, will appear in the proceedings of the security conference 'Sicherheit 2008' which will be held in Saarbrücken (Germany) in April 2008.

2 The Goal

Generically, the compression function $h\colon \{0,1\}^t \times \{0,1\}^s \rightarrow \{0,1\}^t$ of a dedicated hash function H consists of the following steps:

1. (Input) chaining value $r_{(0)}$ (first block: IV) and message block m
2. (Message Expansion) $m = (m_1, \ldots, m_{s/32}) \mapsto \widetilde{m} = (\widetilde{m}_1, \ldots, \widetilde{m}_N)$
3. (Initialization of the registers) for $i = 1$ to k do
   $r_{-k+i} := r_{(0),i} \in \{0,1\}^{32}$
   where $r_{(0),i}$ denotes a particular word of the IV, resp. the chaining value.
4. (Step functions) for $i = 1$ to N do
   $r_i := F_i(r_{i-1}, \ldots, r_{i-k}, \widetilde{m}_i)$
5. (Postadditions) for $i = N - k + 1$ to N do
   $r_i^p := r_i + r_{i-N} (\mathrm{mod} 2^{32})$
6. (Output) $(r_{N-k+1}^p, \ldots, r_N^p)$ (new chaining value)

Remark 1. (i) (Example) MD5: $(s,t,N,k) = (512,128,64,4)$, SHA-1: $(s,t,N,k) = (512,160,80,5)$, SHA-256: $(s,t,N,k) = (512,256,64,8)$.
(ii) The step function F_i usually depends on the Step number i.

(iii) The widespread dedicated hash functions usually perform arithmetic on 32-bit words. Although our results can immediately be transferred from $Z_{2^{32}}$ to any other modulus Z_{2^v} we assume $v = 32$ in the following. For the sake of readability we do not introduce a further parameter v.

For any hash function H a (one-block) collision can be found with complexity $O(2^{t/2})$ ("birthday paradox"). Roughly speaking, the goal of a collision attack is to determine sufficient conditions on related message blocks (m, m') and on the intermediate register values $(r_1, r'_1), \ldots, (r_N, r'_N)$, $(r^p_{N-k+1}, r^{p\prime}_{N-k+1}), \ldots, (r^p_N, r^{p\prime}_N)$ such that $h(c, m) = h(c, m')$ (collision) or at least that $h(c, m)$ and $h(c, m')$ are 'similar', assuming a determined difference (near-collision) and 'preparing' a collision in one of the next blocks. Usually, there exists a number $N_1 < N$ such that a suitable (random) choice of (m, m') guarantees the conditions on the register values (r_j, r'_j) and the expanded message blocks $(\widetilde{m}_j, \widetilde{m}'_j)$ in steps $j \leq N_1$ (message modification). The conditions specified *after step* N_1 shall be satisfied with a considerably larger probability than $2^{-t/2}$.

From Step $N_1 + 1$ to N (including the postadditions) the attacker just checks whether the intermediate register values (and possibly the expanded message blocks) fulfil the given sufficient conditions (with the option of stopping the calculation of $(h(c, m), h(c, m'))$ early), or at least whether $h(c, m)$ and $h(c', m')$ meet certain properties. In *fixed differential schemes* ([WY], [Kli1] etc.) the sufficient conditions for all blocks are determined before the attack is started and remain fixed for any repetition of the attack. In contrast in *variable differential schemes* ([DR],[DMR]) the (near-)collision path in block i depends on the chaining values after step $i - 1$. This saves bit conditions on the chaining values (i.e. on the postadditions) but requires the search of a new (near-)collision paths whenever the attack is applied.

In this paper we are interested in the probabilities of (near-)collision paths, or more precisely, in the probability that the sufficient conditions after Step N_1 (end of the message modification) are fulfilled. We interpret the register values and the extended message blocks as values that are assumed by random variables, which we denote with the respective capital letters. By the example MD5 we formulate and justify a stochastic model and demonstrate how to apply the theorems from Section 3 to determine (almost) exact path probabilities. Note that the exact probability of a collision path follows from the conditional probabilities (= transition probabilities)

$$\text{Prob}((R_i, R'_i) \mid R_{i-1}, R'_{i-1}, \ldots, \widetilde{M}_i, \widetilde{M}'_i, \ldots) \quad \text{and} \tag{8}$$

$$\text{Prob}((R^p_i, R'^p_i) \mid R_i, R'_i, R_{i-N}, R'_{i-N}, \ldots) \text{ (postaddition)}, \tag{9}$$

where the random vectors are contained in particular subsets (cf. (3)). The conditional parts comprise the history up to Step i where the random variables $R_i, R'_i, \ldots$ meet specific path-dependent requirements. The following section is rather technical but provides three very useful theorems that support our goals which were formulated after (4).

3 Three Useful Theorems

In this section we prove three theorems that will be very useful in Section 4. Although our results can immediately be transferred to any other additive group Z_{2^v} we restrict our attention to $\mathrm{Z}_{2^{32}}$ (cf. Remark 1).

During the first reading the reader may skip the lemmata and the proofs within this section. To follow Section 4 it suffices to be familiar with the notation and the statements of the three theorems.

Definition 1. *For $M \in \mathbb{N}$ the term Z_M stands for $\{0, 1, \ldots, M-1\}$. In the following $w[j]$ denotes the j^{th} bit of a 32-bit word w. Numbering starts at the least significant bit with 1.*

For $a, b \in \mathrm{Z}_{2^{32}}$ the term $\Delta(a, b)$ denotes the modulo 2^{32}-difference of a and b, i.e. $\Delta(a, b) := (b - a) (\mathrm{mod}\, 2^{32})$. Similarly as in [WY] we define $\Delta_B(a, b) := [\pm j_1, \ldots, \pm j_k]$ where $j_1, \ldots, j_k$ denote those bit positions where a and b are different. Here '$+j$', resp. simply 'j', means that $(a[j], b[j]) = (0, 1)$ while '$-j$' means that $(a[j], b[j]) = (1, 0)$.

Let X denote a random variable that assumes values on $\mathrm{Z}_{2^{32}}$, and assume $\mathrm{Prob}(X \in A) > 0$. Then $X \mid A$ denotes the conditional random variable, which is given by $\mathrm{Prob}((X \mid A) = x) = \mathrm{Prob}(X = x)/\mathrm{Prob}(X \in A)$ for all $x \in A$ and $= 0$ else. If $\mathrm{Prob}(X = a) = \mathrm{Prob}(X \in A)/|A|$ for each $a \in A$ then $(X \mid A)$ is uniformly distributed on A. If it is non-ambiguous we also loosely say that X is uniformly distributed on A.

In the following $F_+, F_-, F_0, F_1 \subseteq \{1, \ldots, 32\}$ and $F_{32,\neq} \subseteq \{32\}$ denote disjoint subsets. Further, $F_= := \{1, \ldots, 32\} \setminus (F_+ \cup F_- \cup F_0 \cup F_1 \cup F_{32,\neq})$.

There exist obvious 1-1-correspondences between the index sets $F_+, \ldots, F_=$ and the subsets $S_+, \ldots, S_= \subseteq \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}}$ defined below:

$$S_+ := \{(m, m') \in \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \mid (m[j], m'[j]) = (0, 1) \text{ for all } j \in F_+\} \quad (10)$$

$$S_- := \{(m, m') \in \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \mid (m[j], m'[j]) = (1, 0) \text{ for all } j \in F_-\} \quad (11)$$

$$S_0 := \{(m, m') \in \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \mid (m[j], m'[j]) = (0, 0) \text{ for all } j \in F_0\} \quad (12)$$

$$S_1 := \{(m, m') \in \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \mid (m[j], m'[j]) = (1, 1) \text{ for all } j \in F_1\} \quad (13)$$

$$S_{32,\neq} := \begin{cases} \{(m, m') \in \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \mid m[32] \neq m'[32]\} & \text{if } F_{32,\neq} = \{32\} \\ \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} & \text{if } F_{32,\neq} = \{\} \end{cases} \quad (14)$$

$$S_= := \{(m, m') \in \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \mid m[j] = m'[j] \text{ for all } j \in F_=\} \quad (15)$$

In the notation of [WY] the index sets $F_+, F_-, F_0, F_1, F_{32,\neq}, F_=$ express bit conditions. Note that $(a, b) \in S(F_+, F_-, F_0, F_1, F_{32,\neq}, F_=)$
$:= S_+ \cap S_- \cap S_0 \cap S_1 \cap S_{32,\neq} \cap S_=$ iff (a, b) meets the bit conditions implied by $F_+, F_-, F_0, F_1, F_{32,\neq}, F_=$. Example 1 and Table 1 illustrate the connection between bit conditions and the notion of F sets.

Example 1. The bit conditions ($\Delta_B(a, b) = [30, -26]$; $a[4] = b[4] = 1$) correspond to $F_+ = \{30\}, F_- = \{26\}, F_0 = \{\}, F_1 = \{4\}, F_{32,\neq} = \{\}, F_= = \{1, \ldots, 32\} \setminus \{4, 26, 30\}$.

	F_+	F_-	F_0	F_1	$F_{32,\neq}$	$F_=$
$[32]$	$\{32\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\{1,\ldots,32\}\setminus\{32\}$
$[-30]$	$\emptyset$	$\{30\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\{1,\ldots,32\}\setminus\{30\}$
$[32,-30]$	$\{32\}$	$\{30\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\{1,\ldots,32\}\setminus\{32,30\}$
$a[2]=b[2]=0$	$\emptyset$	$\emptyset$	$\{2\}$	$\emptyset$	$\emptyset$	$\{1,\ldots,32\}\setminus\{2\}$
$a[7]=b[7]=1$	$\emptyset$	$\emptyset$	$\emptyset$	$\{7\}$	$\emptyset$	$\{1,\ldots,32\}\setminus\{7\}$
$[*32]$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\{32\}$	$\{1,\ldots,32\}\setminus\{32\}$

Table 1. Bit conditions vs. the notion of F-sets

Definition 2. *Let*

$$\Delta(F_+,F_-,F_{32,\neq 32}) := \sum_{j\in F_{32,\neq}} 2^{31} + \sum_{j\in F_+} 2^{j-1} - \sum_{j\in F_-} 2^{j-1} (\mathrm{mod}\, 2^{32}). \tag{16}$$

In analogy to the sets $F_+,\ldots,F_=$ we assume that the subsets $G_0,G_1\subseteq\{1,\ldots,32\}$ are disjoint.

Similarly as above, for $q\in\{0,1\}$

$$T_q := \{m\in \mathrm{Z}_{2^{32}} \mid m[j]=q \text{ for all } j\in G_q\} \tag{17}$$

implies a 1-1-correspondence between the index set G_q and $T_q\subseteq \mathrm{Z}_{2^{32}}$. Further, $T(G_0,G_1) := T_0\cap T_1$.

The following lemma collects important facts that will be needed later in this section.

Lemma 1. *(i) The mappings*

$$(F_+,F_-,F_0,F_1,F_{32,\neq},F_=)\mapsto \tag{18}$$
$$S(F_+,F_-,F_0,F_1,F_{32,\neq},F_=) = S_+\cap S_-\cap S_0\cap S_1\cap S_{32,\neq}\cap S_=$$
and
$$(G_0,G_1)\mapsto T(G_0,G_1) = T_0\cap T_1. \tag{19}$$

are injective.
(ii) For any $(a,b)\in S(F_+,F_-,F_0,F_1,F_{32,\neq},F_=)$

$$\begin{aligned}\Delta(a,b)\equiv b-a &\equiv \sum_{j\in F_{32,\neq}} 2^{31} + \sum_{j\in F_+} 2^{j-1} - \sum_{j\in F_-} 2^{j-1} (\mathrm{mod}\, 2^{32}) \\ &= \Delta(F_+,F_-,F_{\neq}).\end{aligned} \tag{20}$$

In particular, the function $\Delta(\cdot,\cdot)$ is constant on the set $S_+\cap S_-\cap S_{32,\neq}\supseteq S(F_+,F_-,F_0,F_1,F_{32,\neq},F_=)$, assuming a value $\Delta(F_+,F_-,F_{32,\neq})$.
(iii) Let $\mathrm{pr}_1\colon \mathrm{Z}_{2^{32}}\times\mathrm{Z}_{2^{32}}\to\mathrm{Z}_{2^{32}}$ denote the projection onto the first component. Then

$$\mathrm{pr}_1(S(F_+,F_-,F_0,F_1,F_{32,\neq},F_=)) = T(F_+\cup F_0, F_-\cup F_1). \tag{21}$$

(iv) The mapping

$$(F_+, F_-, F_0, F_1, F_{32,\neq}, F_=) \mapsto (\Delta(F_+, F_-, F_{32,\neq}), (F_+ \cup F_0, F_- \cup F_1)) \quad (22)$$

is injective.

$$(v) \qquad (a,b) \in S(F_+, F_-, F_0, F_1, F_{32,\neq}, F_=) \Longleftrightarrow \quad (23)$$
$$\big((b-a \equiv \Delta(F_+, F_-, F_{32,\neq}) (\mathrm{mod}\ 2^{32})) \,,\, (a \in T(F_+ \cup F_0, F_- \cup F_1)) \big) \,.$$

(vi) Let X, X' denote random variables that assume values on $\mathrm{Z}_{2^{32}}$, and let $S := S(F_+, F_-, F_0, F_1, F_{32,\neq}, F_=)$, $\Delta := \Delta(F_+, F_-, F_{32,\neq})$ and $T := T(F_+ \cup F_0, F_- \cup F_1)$ for the moment. Then

$$\mathrm{Prob}((X, X') \in S) = \mathrm{Prob}(X' - X \equiv \Delta (\bmod\ 2^{32}) \mid X \in T) \cdot \mathrm{Prob}(X \in T). \quad (24)$$

Proof. Assertions (i) and (ii) are obvious since $+2^{31} \equiv -2^{31} (\mathrm{mod} 2^{32})$. Assertion (iii) is true since $j \in F_+ \cup F_0$ implies $m[j] = 0$ for all $(m, m') \in S_+ \cap S_0 \supseteq S(F_+, F_-, F_0, F_1, F_{32,\neq}, F_=)$ etc. The "$\Rightarrow$" direction of (v) is obvious from (ii) and (iii). To verify the inverse direction note that because of (iii) for every $a \in T(F_+ \cup F_0, F_- \cup F_1)$ there is at least one b with $(a, b) \in S(F_+, F_-, F_0, F_1, F_{32,\neq}, F_=)$, but (ii) implies that there is only one, namely, with $b - a \equiv \Delta(F_+, F_-, F_{32,\neq}) (\bmod\ 2^{32})$. Assertion (iv) follows from (v) and also (vi) is an immediate consequence of (v) and the definition of conditional probabilities.

Under mild and reasonably justifiable stochastic assumptions Theorem 1 to Theorem 3 below allow to move the calculation of transition probabilities of hash collision paths from the product space $\mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}}$ to $\mathrm{Z}_{2^{32}}$ (cf. (8), (9) and (44)), which constitutes an enormous improvement; see Section 4 for details. For the moment we merely mention that in Section 4 the sets $S_{(\cdot)}$ and $T_{(\cdot)}$ will characterize bit conditions while the random variables X, Y and Z correspond to intermediate values or to register values. Note that under the certain conditions the step transition probability indeed equals $2^{-\#\,\text{bit conditions}}$, the value obtained by simple bit condition counting, (cf. Theorem 1(ii) and Theorem 2(ii)).

Definition 3. *For the remainder of this section we use the abbreviations $S_{(i)} := S(F_{(i)+}, F_{(i)-}, F_{(i)0}, F_{(i)1}, F_{(i)32,\neq}, F_{(i)=})$, $\Delta_{(i)} := \Delta(F_{(i)+}, F_{(i)-}, F_{(i)32,\neq})$ and $T_{(i)} := T(F_{(i)+} \cup F_{(i)0}, F_{(i)-} \cup F_{(i)1})$. The index i ranges from 1 to 3.*

For $0 \leq sh < 32$ the term $w^{<<<sh}$ denotes the cyclic shift of the word w by sh positions to the left. Similarly $(w, w')^{<<<sh}$ stands for $(w^{<<<sh}, w'^{<<<sh})$. Analogously, $F_*^{<<<sh}$ results from adding the integer sh to each element in F_*, where the integers $33, 34, \ldots$ are interpreted as $1, 2, \ldots$.

Remark 2. (i) Clearly, if $F_{32,\neq} = \{\}$ the image $S(F_+, F_-, F_0, F_1, F_{32,\neq}, F_=)^{<<<sh}$ equals $S(F_+^{<<<sh}, F_-^{<<<sh}, F_0^{<<<sh}, F_1^{<<<sh}, \{\}, F_=^{<<<sh})$. We have $j \in F_*$ iff $j + sh$, resp. $j + sh - 32 \in F_*^{<<<sh}$. The latter term means the set This

condition is not very restrictive since the set $S(F_+, F_-, F_0, F_1, \{32\}, F_=)$ equals the disjoint union $S(F_+ \cup \{32\}, F_-, F_0, F_1,$
$\{\}, F_=) \cup S(F_+, F_- \cup \{32\}, F_0, F_1, \{\}, F_=)$.
(ii) Note that $\Delta(F_+, F_-, \{\}) = \Delta(F'_+, F'_-, \{\})$ does not necessarily imply $\Delta(F_+^{<<<sh}, F_-^{<<<sh}, \{\}) = \Delta(F'^{<<<sh}_+, F'^{<<<sh}_-, \{\})$.
Counterexample: $F_+ = \{20\}, F'_+ = \{21\}, F'_- = \{20\}, sh = 12$. Then $\Delta(\{20\}, \{\}, \{\})$ $= 2^{19} = \Delta(\{21\}, \{20\}, \{\})$ but $\Delta(\{20\}^{<<<12}, \{\}, \{\}) = \Delta(\{32\}, \{\}, \{\}) = 2^{31}$ whereas $\Delta(\{21\}^{<<<12}, \{20\}^{<<<12}, \{\}) = \Delta(\{1\}, \{32\}, \{\}) \equiv -2^{31} + 1 \equiv 2^{31} + 1 (\mathrm{mod}\, 2^{32})$.

Theorem 1. *Let X, X', Y, Y' denote random variables that assume values in $\mathrm{Z}_{2^{32}}$, where (X, X') and (Y, Y') are independent. Further, assume that $0 \le sh < 32$ and $F_{(1)32,\neq} = \{\}$.*
(i) Setting $\widetilde{\Delta}_{(1)} := \Delta(F_{(1)+}^{<<<sh}, F_{(1)-}^{<<<sh}, \{\})$ the conditional probability

$$\mathrm{Prob}\left(\left[(X, X')^{<<<sh} + (Y, Y')\right] (\mathrm{mod}\, 2^{32}) \in S_{(3)} \mid (X, X') \in S_{(1)}, (Y, Y') \in S_{(2)}\right) \tag{25}$$

equals

$$\begin{cases} \mathrm{Prob}\left(\left[X^{<<<sh} + Y\right] (\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid (X, X') \in S_{(1)}, (Y, Y') \in S_{(2)}\right) \\ \qquad\qquad\qquad\qquad\qquad\qquad \textit{if } \Delta_{(3)} \equiv \widetilde{\Delta}_{(1)} + \Delta_{(2)} (\mathrm{mod}\, 2^{32}) \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad \textit{else} \end{cases} \tag{26}$$

(ii) If (X, X') and X are uniformly distributed on $S_{(1)}$ and $T_{(1)}$, resp., the condition '$(X, X') \in S_{(1)}$' in (26) may be replaced by '$X \in T_{(1)}$'.
If additionally $T_{(1)} = \mathrm{Z}_{2^{32}}$ under the conditions of (26) the random variable $Z := \left(\left[X^{<<<sh} + Y\right] (\mathrm{mod} 2^{32}\right.$ is uniformly distributed on $\mathrm{Z}_{2^{32}}$, and (Z, Z') is uniformly distributed on $S_{(3)}$. Further, $Z \mid T_{(3)}$ and $Y \mid T_{(2)}$ as well as $(Z, Z') \mid S_{(3)}$ and $(Y, Y') \mid S_{(2)}$ are independent, and the first line in (26) equals $2^{-|F_{(3)+} \cup F_{(3)0} \cup F_{(3)-} \cup F_{(3)1}|}$.
The corresponding assertions (with interchanged roles of X and Y) hold if (Y, Y') and Y are uniformly distributed on $S_{(2)}$ and $T_{(2)}$, respectively.
(iii) Assume that in (i) $(X, X'), X, (Y, Y'), Y$ are uniformly distributed on the sets $S_{(1)}, T_{(1)}, S_{(2)}$ and $T_{(2)}$, respectively. Then (26) simplifies to

$$\begin{cases} \mathrm{Prob}\left(\left[X^{<<<sh} + Y\right] (\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid X \in T_{(1)}, Y \in T_{(2)}\right) \\ \qquad\qquad\qquad\qquad\qquad\qquad \textit{if } \Delta_{(3)} \equiv \widetilde{\Delta}_{(1)} + \Delta_{(2)} (\mathrm{mod}\, 2^{32}) \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad \textit{else} \end{cases} \tag{27}$$

Proof. To simplify notation we use the abbreviations $(\widetilde{X}, \widetilde{X}') := (X, X')^{<<<sh}$ and $\widetilde{S}_{(1)} := \widetilde{S}_{(1)}^{<<<sh}$ and $\widetilde{T}_{(1)} := \mathrm{pr}_1(\widetilde{S}_{(1)})$ within this proof. We note that $(\widetilde{X}, \widetilde{X}') \in \widetilde{S}_{(1)}$ iff $(X, X') \in S_{(1)}$ since $F_{32,\neq} = \{\}$ (cf. Remark 2(i)), and similarly $\widetilde{X} \in \widetilde{T}_{(1)}$ iff $X \in T_{(1)}$.
We assume $\mathrm{Prob}((X, X') \in S_{(1)}, (Y, Y') \in S_{(2)}) > 0$ since otherwise the conditional probability (25) may be defined arbitrarily. Obviously, this conditional

probability is zero if $\Delta_{(3)} \not\equiv \widetilde{\Delta}_{(1)} + \Delta_{(2)} (\mathrm{mod}\, 2^{32})$. We assume $\Delta_{(3)} \equiv \Delta_{(1)} + \Delta_{(2)} (\mathrm{mod}\, 2^{32})$ in the remainder of this proof. Using the above equivalences Lemma 1(v) verifies (26), which equals

$$\sum_{(\widetilde{x},\widetilde{x}') \in \widetilde{S}_{(1)}, (y,y') \in S_{(2)}} \mathrm{Prob}\left(\left[\widetilde{X} + Y \right] (\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid (\widetilde{X}, \widetilde{X}') = (\widetilde{x}, \widetilde{x}'), (Y, Y') = (y, y') \right) \times$$

$$\times \frac{\mathrm{Prob}((\widetilde{X}, \widetilde{X}') = (\widetilde{x}, \widetilde{x}'), (Y, Y') = (y, y'))}{\mathrm{Prob}((\widetilde{X}, \widetilde{X}') \in \widetilde{S}_{(1)}, (Y, Y') \in S_{(2)})}.$$

To any $(\widetilde{x}, y) \in \widetilde{T}_{(1)} \times T_{(2)}$ there exists a unique quadruple $(\widetilde{x}, \widetilde{x}', y, y') \in \widetilde{S}_{(1)} \times S_{(2)}$. Assume that (X, X') and X are uniformly distributed on $S_{(1)}$ and $T_{(1)}$, repectively. Then $(\widetilde{X}, \widetilde{X}')$ and $\widetilde{X}$ are uniformly distributed on $\widetilde{S}_{(1)}$ and $\widetilde{T}_{(1)}$, respectively. Since $(\widetilde{X}, \widetilde{X}')$ and (Y, Y') are independent the above term simplifies to

$$\begin{aligned}
&\sum_{\widetilde{x} \in T_{(1)}, y \in T_{(2)}} \mathrm{Prob}\left(\left[\widetilde{X} + Y \right] (\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid \widetilde{X} = \widetilde{x}, Y = y \right) \times \\
&\qquad \times \frac{\mathrm{Prob}((\widetilde{X}, \widetilde{X}') = (\widetilde{x}, \widetilde{x}'))}{\mathrm{Prob}((\widetilde{X}, \widetilde{X}') \in \widetilde{S}_{(1)})} \cdot \frac{\mathrm{Prob}((Y, Y') = (y, y'))}{\mathrm{Prob}((Y, Y') \in S_{(2)})} \\
&= \sum_{\widetilde{x} \in T_{(1)}, \widetilde{y} \in T_{(2)}} \mathrm{Prob}\left(\left[\widetilde{X} + Y \right] (\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid \widetilde{X} = \widetilde{x}, Y = y \right) \times \\
&\qquad \times \frac{\mathrm{Prob}(\widetilde{X} = \widetilde{x})}{\mathrm{Prob}(\widetilde{X} \in \widetilde{T}_{(1)})} \cdot \frac{\mathrm{Prob}((Y, Y') = (y, y'))}{\mathrm{Prob}((Y, Y') \in S_{(2)})} \\
&= \mathrm{Prob}\left(\left[\widetilde{X} + Y \right] (\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid \widetilde{X} \in \widetilde{T}_{(1)}, (Y, Y') \in S_{(2)} \right)
\end{aligned}$$

If additionally $T_{(1)} = \mathrm{Z}_{2^{32}}$ $(= \widetilde{T}_{(1)})$ for any $(y, y') \in S_{(2)}$ the random variable $Z_y := \widetilde{X} + y$ is uniformly distributed on $\mathrm{Z}_{2^{32}}$ since X and (Y, Y') are independent. Since $Z'_y = Z_y + \widetilde{\Delta}_{(1)} + \Delta_{(2)} (\mathrm{mod}\, 2^{32})$ by Lemma 1(v) the random vector (Z_y, Z'_y) is uniformly distributed on $S_{(3)}$ (with total mass zero if $\Delta_{(3)} \not\equiv \widetilde{\Delta}_{(1)} + \Delta_{(2)} (\mathrm{mod}\, 2^{32})$). Since (y, y') was arbitrary, $(Z, Z') \mid S_{(3)}$ and $(Y, Y') \mid S_{(2)}$ are independent, and for similar reasons also $Z \mid T_{(3)}$ and $Y \mid T_{(2)}$ are independent. Since Z is uniformly distributed on $\mathrm{Z}_{2^{32}}$ the conditional probability in (26) equals $|T_{(3)}|/|\mathrm{Z}_{2^{32}}|$. Mimicking this proof verifies the second part of (ii) and (iii).

Corollary 1. *For $sh = 0$ we may drop the condition $F_{(1)32\neq} = \{\}$ in Theorem 1.*

The proof of Theorem 1 can be adapted in a straight-forward way to verify Corollary 1. The case $sh = 0$ is of particular interest ($\rightarrow$ postadditions; cf. Section 4).

Definition 4. *For $a, b, c \in \mathrm{Z}_{2^{32}}$ and $0 \leq sh < 32$ we define the set $M(a, b, c, sh)$* $:= \{u \in \mathrm{Z}_{2^{32}} \mid \Delta((u, u + a (\mathrm{mod}\, 2^{32}))^{<<<sh}) + b \equiv c (\mathrm{mod}\, 2^{32})\}$.

Theorem 2. *Let X, X', Y, Y' denote random variables that assume values in $\mathrm{Z}_{2^{32}}$, where (X, X') and (Y, Y') are independent. Further, $0 \leq sh < 32$.*
(i) Let $\Delta_{[1]} \in \mathrm{Z}_{2^{32}}$. Assume further that (X, X') and X are uniformly distributed on $\{(x, x + \Delta_{[1]} (\mathrm{mod}\, 2^{32})) \mid x \in \mathrm{Z}_{2^{32}}\}$ and on $\mathrm{Z}_{2^{32}}$, respectively. Then

$$\begin{aligned}&\mathrm{Prob}\left(\left[(X, X')^{<<<sh} + (Y, Y')\right] (\mathrm{mod}\, 2^{32}) \in S_{(3)} \mid \Delta(X, X') = \Delta_{[1]}, (Y, Y') \in S_{(2)}\right)\\ &= \mathrm{Prob}\left(\left[X^{<<<sh} + Y\right] (\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid X \in M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh), (Y, Y') \in S_{(2)}\right) \times\\ &\quad \times \mathrm{Prob}(X \in M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh)). \qquad (28)\end{aligned}$$

(ii) If $M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh) = \mathrm{Z}_{2^{32}}$ the random vector $(Z := X^{<<<sh} + Y (\mathrm{mod}\, 2^{32}), Z' := X'^{<<<sh} + Y' (\mathrm{mod}\, 2^{32}))$ and the random variable Z are uniformly distributed on $S_{(3)}$ and $T_{(3)}$, respectively. In particular, $Z \mid T_{(3)}$ and $Y \mid T_{(2)}$ as well as $(Z, Z') \mid S_{(3)}$ and $(Y, Y') \mid S_{(2)}$ are independent, and (28) equals $2^{-|F_{(3)+} \cup F_{(3)0} \cup F_{(3)-} \cup F_{(3)1}|}$.
(iii) Assume that in (i) the random vector (Y, Y') and the random variable Y are uniformly distributed on $S_{(2)}$ and $T_{(2)}$, respectively. For any $M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh)$ the right-hand-side of (28) simplifies to

$$\begin{aligned}&\mathrm{Prob}\left(\left[X^{<<<sh} + Y\right] (\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid X \in M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh), Y \in T_{(2)}\right) \times\\ &\quad \times \mathrm{Prob}(X \in M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh)). \qquad (29)\end{aligned}$$

If $T_{(2)} = \mathrm{Z}_{2^{32}}$ the random vector (Z, Z') and Z are uniformly distributed on $S_{(3)}$ and $\mathrm{Z}_{2^{32}}$. In particular, (29) further simplifies to

$$2^{-|F_{(3)+} \cup F_{(3)0} \cup F_{(3)-} \cup F_{(3)1}|} \cdot \mathrm{Prob}(X \in M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh)). \qquad (30)$$

Proof. We first note that the set $M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh)$ is well-defined. As in the proof of Theorem 1(i) we may assume that $\mathrm{Prob}((Y, Y') \in S_{(2)}) > 0$ in the remainder. Due to Lemma 1(v) the left-hand side in (28) equals

$$\begin{aligned}&\mathrm{Prob}\left(\left[X^{<<<sh} + Y\right] (\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid \Delta(X, X') = \Delta_{[1]}, X \in M(\cdots), (Y, Y') \in S_{(2)}\right) \times\\ &\qquad \times \frac{\mathrm{Prob}(\Delta(X, X') = \Delta_{[1]}, X \in M(\cdots))}{\mathrm{Prob}(\Delta(X, X') = \Delta_{[1]})}\end{aligned}$$

Due to the uniformity assumptions on (X, X') and X the second factor equals $\mathrm{Prob}(X \in M(\cdots))$. If additionally (Y, Y') and Y are uniformly distributed on $S_{(2)}$ and $T_{(2)}$, resp., the above term simplifies to

$$\begin{aligned}&\sum_{x \in M(\cdots), y \in T_{(2)}} \mathrm{Prob}\left(\left[X^{<<<sh} + Y\right] (\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid X = x, Y = y\right) \times\\ &\quad \times \frac{\mathrm{Prob}((X, X') = (x, x + \Delta_{[1]}))}{\mathrm{Prob}((X, X') : X \in M(\cdots), \Delta(X, X') = \Delta_{[1]})} \cdot \frac{\mathrm{Prob}((Y, Y') = (y, y + \Delta_{(2)}))}{\mathrm{Prob}((Y, Y') \in S_{(2)})} \times\\ &\quad \times \mathrm{Prob}(X \in M(\cdots))\\ &= \sum_{x \in M(\cdots), y \in T_{(2)}} \mathrm{Prob}\left(\left[X^{<<<sh} + Y\right] (\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid X = x, Y = y\right) \times\\ &\qquad \times \frac{\mathrm{Prob}(X = x)}{\mathrm{Prob}(X \in M(\cdots))} \cdot \frac{\mathrm{Prob}(Y = y)}{\mathrm{Prob}(Y \in T_{(2)})} \cdot \mathrm{Prob}(X \in M(\cdots))\end{aligned}$$

which proves (29). Apart from the fact that the ratio $\mathrm{Prob}((Y,Y') = (y, y + \Delta_{(2)}))/\mathrm{Prob}((Y,Y') \in S_{(2)})$ is not replaced by $\mathrm{Prob}(Y = y)/\mathrm{Prob}(Y \in T_{(2)})$ formula (28) follows analogously. The remaining assertions can be proved with similar techniques as used in the proof of Theorem 1.

In the remainder of this section we derive a characterization of the set $M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh)$ which is more suitable for concrete computations (cf. Sect. 4).

Definition 5. *For $a \in \mathrm{Z}$ and $n \in N$ we set a div n to be $\lfloor a/n \rfloor$ where $\lfloor r \rfloor$ denotes the largest integer that is $\leq r$.*
For $a \in \mathrm{Z}$ the term $a(\mathrm{mod}\, M)$ stands for the representative of $a + \mathrm{Z}/M\mathrm{Z}$ in Z_{M}, i.e. for that element in Z_{M} that is congruent to the integer a modulo M. For $0 \leq sh < 32$ we define $\mathrm{ca}(a_1, \ldots, a_k; sh) := (a_1 + \cdots + a_k)$ div 2^{32-sh} ('carry').

Lemma 2. *Within this lemma let $x' \in \mathrm{Z}_{2^{32}}$ with $x' \equiv x + \Delta (\mathrm{mod}\, 2^{32})$ for fixed $\Delta \in \mathrm{Z}$. Assume further that $0 \leq sh < 32$ and $x = x_1 \cdot 2^{32-sh} + x_0$ and $x' = x'_1 \cdot 2^{32-sh} + x'_0$ with $0 \leq x_0, x'_0 < 2^{32-sh}$ and $0 \leq x_1, x'_1 < 2^{sh}$. Further, we decompose $\Delta = \Delta_1 \cdot 2^{32-sh} + \Delta_0$ where Δ_0 and Δ_1 may assume arbitrary integer values. This implies:*
(i) For $a \in \mathrm{Z}$ and $n \in N$ we have $a(\mathrm{mod} n) = a - (a$ div $n)n$.
(ii) For $a \in \mathrm{Z}$ and $n, m \in \mathbb{N}$ we have $(a(\mathrm{mod} n))m = am(\mathrm{mod}\, nm)$.
(iii) $x^{<<<sh} = x_0 \cdot 2^{sh} + x_1$.
(iv) $x' = ((x_1 + \Delta_1 + \mathrm{ca}(x_0, \Delta_0; sh))(\mathrm{mod}\ 2^{sh})) \cdot 2^{32-sh} + (x_0 + \Delta_0)(\mathrm{mod}\ 2^{32-sh})$ (integer equation!)
(v) $x'^{<<<sh} \equiv \left[x_0 + \Delta_0 - \left((x_1 + \Delta_1 + \mathrm{ca}(x_0, \Delta_0; sh))\ \mathrm{div}\ 2^{sh}\right)\right] \cdot 2^{sh} + x_1 + \Delta_1 + \mathrm{ca}(x_0, \Delta_0; sh)\ (\mathrm{mod}\, 2^{32})$.
(vi) Let $k \cdot 2^{32-sh} \leq \Delta_0 < (k+1)2^{32-sh}$ for a particular $k \in Z$. Then $\mathrm{ca}(x_0, \Delta_0; sh) \in \{k, k+1\}$.

Proof. Assertion (i) follows from its definition, and $(a(\mathrm{mod} n))m = (a - [a$ div $n]n)m = am - [a$ div $n]nm = am(\mathrm{mod}\ nm)$. Assertions (iii), (iv) and (vi) are obvious. From (iv) we immediately obtain

$$x'^{<<<sh} \equiv \left[(x_0 + \Delta_0)\,(\mathrm{mod}\, 2^{32-sh})\right] \cdot 2^{sh} + (x_1 + \Delta_1 + \mathrm{ca}(x_0, \Delta_0; sh))\,(\mathrm{mod}\ 2^{sh}).$$

Applying (i) to the right-hand summand and (ii) to the left-hand summand proves (v).

Theorem 3. *(Continuation of Theorem 2) Assume that $\Delta_{(3)} - \Delta_{(2)} \equiv \left(\widetilde{\Delta}_0 \cdot 2^{sh} + \widetilde{\Delta}_1\right)(\mathrm{mod}\, 2^{32})$ and $\Delta_{[1]} \equiv \left(\Delta_1 \cdot 2^{32-sh} + \Delta_0\right)(\mathrm{mod}\, 2^{32})$ with integers $\widetilde{\Delta}_0, \widetilde{\Delta}_1, \Delta_0, \Delta_1$, which need not be nonnegative.*
(i) Then

$$x = x_1 \cdot 2^{32-sh} + x_0 \in M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh) \tag{31}$$

iff

$$\widetilde{\Delta}_0 \cdot 2^{sh} + \widetilde{\Delta}_1 \equiv \left(\Delta_0 - [x_1 + \Delta_1 + \mathrm{ca}(x_0, \Delta_0; sh)]\ \mathrm{div}\ 2^{sh}\right) \cdot 2^{sh} + [\Delta_1 + \mathrm{ca}(x_0, \Delta_0; sh)]\ (\mathrm{mod}\, 2^{32}) \tag{32}$$

(ii) In particular,

$$\begin{aligned} &\mathrm{ca}(x_0, \Delta_0; sh) \equiv \widetilde{\Delta}_1 - \Delta_1 (\mathrm{mod}\, 2^{sh}) \quad \textit{and} \\ &\mathrm{ca}(x_0, \Delta_0; sh) \in \{\Delta_0 \text{ div } 2^{32-sh}, \Delta_0 \text{ div } 2^{32-sh} + 1\}. \end{aligned} \tag{33}$$

(iii) For $0 < sh < 32$ *the relations (33) determine* $\mathrm{ca}(x_0, \Delta_0; sh)$ *uniquely.*
(iv) For $sh = 0$ *trivially* $M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, 0) = \mathrm{Z}_{2^{32}}$ *iff* $\Delta_{[1]} + \Delta_{(2)} \equiv \Delta_{(3)} (\mathrm{mod}\, 2^{32})$ *and* $= \emptyset$ *else.*

Proof. To prove Theorem 3(i) recall the definition of $M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh)$. By assumption, the left-hand side of (32) equals $\Delta_{(3)} - \Delta_{(2)}$. The right-hand side follows immediately from Lemma 2(iii) and (v), which are applied to the first and the second component of $\Delta(u := x_1 \cdot 2^{32-sh} + x_0, u + \Delta_{[1]})$, respectively. Applying Lemma 2(vi) to (32) yields the second assertion of (33), reducing this congruence modulo 2^{sh} proves the first assertion. The proof of (iii) and (iv) is obvious.

Theorem 3 provides the promised alternative characterization of the set $M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh)$, which is more convenient for concrete computations. In particular, (33) provides an inequality for x_0. Due to Theorem3(iii) $ca(x_0, \Delta_0, sh)$ is unique. Substituting this value into (32) provides a relation that determines the 'upper' part x_1; see Section 4 for illustrating examples. We mention that the term $\mathrm{ca}(x_0, \Delta; sh)$ compensates the 'non-uniqueness' of the values Δ_0, Δ_1.

Remark 3. All three theorems will be very useful in the next section. We point out that they can be extended to handle bit conditions that affect (Y, Y') and $(Z := X^{<<<sh} + Y (\mathrm{mod}\, 2^{32}), Z' := X'^{<<<sh} + Y (\mathrm{mod} 2^{32}))$ simultaneously. For instance, the (additional) condition $Y[3] = Y'[3] = Z[3] = Z'[3]$ can be decomposed into two disjoint cases, namely into $Y[3] = Y'[3] = 0 = Z[3] = Z'[3]$, and $Y[3] = Y'[3] = 1 = Z[3] = Z'[3]$, respectively. Both cases can be expressed in the form $S(F_+, F_-, F_0 \cup \{3\}, F_1, F_{32,\neq}, F_=)$ and $S(F_+, F_-, F_0, F_1 \cup \{3\}, F_{32,\neq}, F_=)$ with suitable subsets $F_+, F_-, F_0, F_1, F_{32,\neq}, F_=$.

4 Example: Concrete Collision paths in MD5

In this section we demonstrate the use and the usefulness of the theorems proved in Section 3 by three MD5 near-collision paths for which an experimental verification is possible. These paths may not be optimal (i.e., not the most probable ones) but this is irrelevant for our purpose.

The 512-message block $m_{(1)}$ (resp., $m_{(2)}$) is segmented into 16 words $m_1, \ldots, m_{16}$ of length 32. (We omit the index (1) to simplify notation.) This sequence is extended to 64 words $\widetilde{m}_1, \ldots, \widetilde{m}_{64}$ as follows (message extension). For $i \leq 16$ we set $\widetilde{m}_i := m_i$, and $\widetilde{m}_{17}, \ldots, \widetilde{m}_{32}$, resp. $\widetilde{m}_{33}, \ldots, \widetilde{m}_{48}$, resp. $\widetilde{m}_{49}, \ldots, \widetilde{m}_{64}$ are permutations of $m_1, \ldots, m_{16}$. After the initialization of four registers by the $IV = (IV_0, IV_1, IV_2, IV_3)$

$$r_{-3} := IV_0, \quad r_{-2} := IV_3, \quad r_{-1} := IV_2, \quad r_0 := IV_1 \tag{34}$$

the MD5 algorithm processes 64 steps. (For the second block $m_{(2)}$ the IV has to be replaced by the chaining value h_1.) In Step i the MD5 step function has the form

$$\text{(Step i)} \quad r_i \equiv r_{i-1} + (\Phi_i(r_{i-1}, r_{i-2}, r_{i-3}) + r_{i-4} + \widetilde{m}_i + const_i)^{<<<sh(i)} (\bmod\ 2^{32}) \tag{35}$$

where $\Phi_i \colon \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \to \mathrm{Z}_{2^{32}}$ is a bit-oriented, step-dependent function (cf. Appendix). Also the constant $const_i$ and the number of shift positions $sh(i)$ depend on the particular step. Finally, the four registers are updated by

$$\text{(postaddition)} \quad r_i^p \equiv r_i + r_{i-64} (\bmod 2^{32}) \quad i \in \{61, 62, 63, 64\} \tag{36}$$

The known MD5 attacks are two-block attacks (see, e.g. [WY,Kli2,YaSh]), i.e. after block 1 the pairs $(r_{61}^p, r'^p_{61}), (r_{62}^p, r'^p_{62}), (r_{63}^p, r'^p_{63}), (r_{64}^p, r'^p_{64})$ shall meet specified bit conditions that shall 'prepare' a collision after the compression of the second block. E.g. in [WY,Kli2,YaSh,Th] conditions on the message blocks, the register bits and intermediate values are formulated that shall ensure this goal. The conditions for the first 20 steps can be guaranteed by a (sophisticated) random choice of the message blocks, the so-called message modification ([WY,Kli2] etc.). Our goal is to compute the probability for concrete (near-)collision paths from Step 21 to Step 64 (including the postadditions).

4.1 Step Transition Probabilities

Definition 6. *In this section we denote random variables by capital letters, their realizations, i.e., values assumed by these random variables, by the respective small letters.*

Since at least large parts of the message blocks $m_1, \ldots, m_{16}$ are chosen randomly we interpret the register values $r_{-3}, \ldots, r_0, r_1, \ldots, r_{64}^p$ and the extended message blocks $\widetilde{m}_1, \ldots, \widetilde{m}_{64}$ as realizations of random variables $R_{-3}, \ldots, R_0, R_1, \ldots, R_{64}^p$ and $\widetilde{M}_1, \ldots, \widetilde{M}_{64}$. (The random variables $R_{-3}, \ldots, R_0$ assume constant values (cf. (34).) In the notion of random variables (35) and (36) read

$$R_i \equiv R_{i-1} + (\Phi_i(R_{i-1}, R_{i-2}, R_{i-3}) + R_{i-4} + \widetilde{M}_i + const_i)^{<<<sh(i)} (\bmod\ 2^{32}) \tag{37}$$

and

$$R_i^p \equiv R_i + R_{i-64} (\bmod 2^{32}) \quad i \in \{61, 62, 63, 64\} \tag{38}$$

Note that if we replaced $const_i$ by an independent random variable C_i that is uniformly distributed on $\mathrm{Z}_{2^{32}}$ the terms R_{i-1} and $(\cdots)^{<<<sh(i)}$ were independent and the latter was uniformly distributed on $\mathrm{Z}_{2^{32}}$. Although $const_i$ assumes a constant value the following stochastic model is reasonable.

Definition 7. *In the following we use the abbreviations from Definition 3 and Definition 4 but the indices (i) now denote the number of the step of the compression function (i.e., $i \in \{-3, \ldots, 64\}$).*

Stochastic Model. For $i \leq 64$ we assume that the pairs of random variables $(R_{i-1}, R'_{i-1}), (R_{i-2}, R'_{i-2}), \ldots$ follow a particular near-collision path, i.e. that they meet specified sufficient conditions. Let
$X_i := \left(\Phi_i(R_{i-1}, R_{i-2}, R_{i-3}) + R_{i-4} + \widetilde{M}_i + const_i\right) \pmod{2^{32}}$ and
$X'_i := \left(\Phi_i(R'_{i-1}, R'_{i-2}, R'_{i-3}) + R'_{i-4} + \widetilde{M}'_i + const_i\right) \pmod{2^{32}}$.
We assume that

- (a) the pairs (X_i, X'_i) and (R_{i-1}, R'_{i-1}) are independent
- (b) X_i is uniformly distributed on $\mathrm{Z}_{2^{32}}$
- (c) $(X_i, X'_i) \mid \{(x, x + \Delta_{[i]} (\mathrm{mod}\, 2^{32})) \mid x \in \mathrm{Z}_{2^{32}}\}$ is uniformly distributed.

(The difference $\Delta_{[i]} \in \mathrm{Z}_{2^{32}}$ is determined by the (near-)collision path.)

Justification of the Stochastic Model. (i) We add R_{i-4} and $\widetilde{M}_i$, which have no 'obvious' (at least no linear) dependencies with R_{i-1}, to $\Phi(R_{i-1}, R_{i-2}, R_{i-3})$ (merging the last three register values in a non-linear manner), while the modular addition is a $\mathrm{Z}_{2^{32}}$-linear operation on $\mathrm{Z}_{2^{32}}$. As the same argumentation holds for the related message M' instead of M this justifies Condition (a).
(ii) Even under weak heuristic assumptions the modular sum of three random variables is very close to the uniform distribution, justifying (b). (Note that at least R_{i-4} and $\Phi_i(R_{i-1}, R_{i-2}, R_{i-3})$ should be nearly uniformly distributed on 'large' subsets of $\mathrm{Z}_{2^{32}}$ (determined by the collision path), and also the extended message block $\widetilde{M}_i$ contains some randomness.)
(iii) Assumption (c) grounds on the fact that the register values 'spread' rapidly for different messages. For 'purely' random input M and M' (without message modification) and neglecting any bit condition up to step $i-1$ we would assume that (X, X') is uniformly distributed on $\mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}}$. In our scenario, i.e. where we focus on the small subset of (near-collision) paths that fulfil a sequence of bit conditions, it is reasonable only to assume the weaker assumption which is formulated in (c).

Recall that we are interested in the computation of conditional probabilities that were defined in (8) and in (9). Since (X_i, X'_i) and in particular the difference $\Delta_{[i]}$ result from $(\widetilde{M}_i, \widetilde{M}'_i)$ and the sets $S_{(i-1)}, \ldots, S_{(i-4)}$ we may extend the conditional part in (8) to

$$\mathrm{Prob}((R_i, R'_i) \in S_{(i)} \,|\, (X_i, X'_i) \in \{(u, u + \Delta_{[i]}) |\ u \in \mathrm{Z}_{2^{32}}\}, (R_{i-1}, R'_{i-1}) \in S_{(i-1)}, \times \\ \times \ldots, \widetilde{M}_i, \widetilde{M}'_i, \ldots). \qquad (39)$$

By assertion (a) of the stochastic model the random vectors (X_i, X'_i) and (R_{i-1}, R'_{i-1}) are independent, and (b), (c) specify the distribution of X_i and (X_i, X'_i) . Since (R_i, R'_i) computes from (R_{i-1}, R'_{i-1}) and (X_i, X'_i) we omit the remainder of the conditional part, and in place of (39) we consider the conditional probability

$$\mathrm{Prob}((R_i, R'_i) \in S_{(i)} \,|\, (X_i, X'_i) \in \{(u, u + \Delta_{[i]}) \,|\, u \in \mathrm{Z}_{2^{32}}\}, (R_{i-1}, R'_{i-1}) \in S_{(i-1)}) \quad (40)$$

where (X_i, X_i') and (R_{i-1}, R_{i-1}') satisfy the conditions formulated in the stochastic model. This allows to apply Theorem 2 and Theorem 3 in the following. Remark 4 and Lemma 3(ii) collect useful facts.

Remark 4. (i) If

$$M_{(i)} := M(\Delta_{[i]}, \Delta_{(i-1)}, \Delta_{(i)}, sh(i)) \tag{41}$$

equals $\mathrm{Z}_{2^{32}}$ by Theorem 2(ii) (R_i, R_i') and R_i are uniformly distributed on $S_{(i)}$ and $T_{(i)}$, respectively. Additionally, $(R_i, R_i') \mid S_{(i)}$ and $(R_{i-1}, R_{i-1}') \mid S_{(i-1)}$ are independent.
(ii) For our three near-collision paths the difference $\Delta_{[i]}$ follows deterministically from $\widetilde{M}_i, \widetilde{M}_i', S_{(i-1)}, \ldots, S_{(i-4)}$. Our approach can be adjusted to more general situations where a certain difference $\Delta_{[i]}$ is only assumed with a particular probability.

Lemma 3. *(i) The differential paths specified in Table 3 satisfy*

$$M_{(i)} = \mathrm{Z}_{2^{32}} \text{ for } \mathrm{i} \in \{21, \ldots, 64\} \setminus \{23, 35, 62\} \text{ and} \tag{42}$$
$$M_{(i)} \neq \mathrm{Z}_{2^{32}} \text{ for } \mathrm{i} \in \{23, 35, 62\} \tag{43}$$

(ii) Theorem 2(ii) can be applied in Step $i \in \{21, \ldots, 64\} \setminus \{23, 35, 62\}$. In particular, for these i the exact transition probabilities coincide with the value obtained by bit condition counting.
(iii) Theorem 2(iii) (resp., formula (29)) can applied in Step $i \in \{23, 35, 62\}$.

Proof. Applying Theorem 3 to Steps 21 to 64 verifies (i); see Example 2 and Example 3 for illustration. Theorem 2 clearly can be applied in all steps. Assertion (ii) then follows immediately from (42). In particular, (R_{i-1}, R_{i-1}') and R_{i-1} are uniformly distributed on $S_{(i-1)}$ and $T_{(i-1)}$, respectively, for $i \in \{23, 35, 62\}$, which proves (iii).

All examples in Section 4 refer to the three near-collision paths, which are given in Table 3. In the present subsection we consider the step transition probabilities, i.e. the conditional probabilities in (8). Our goal is to apply Theorem 2 with $\Delta_{[1]} := \Delta_i$, $S_{(2)} = S_{(i-1)}$, $S_{(3)} = S_{(i)}$, $(X, X') := (X_i, X_i')$ and $(Y, Y') := (R_{i-1}, R_{i-1}')$. The bit conditions from Step 21 to Step 64 are listed in Table 3. For $i > 21$ the sets $S_{(i)}$, and for $i \geq 61$ also the $S_{(i)p}$ can be expressed in the form $S_{(i)} := S(F_{(i)+}, F_{(i)-}, F_{(i)0}, F_{(i)1}, F_{(i)32,\neq}, F_{(i)=})$, resp. $S_{(i),p} := S(F_{(i)+,p}, F_{(i)-,p}, F_{(i)0,p}, F_{(i)1,p}, F_{(i)32,\neq,p}, F_{(i)=,p})$ (cf. Sect. 3). In Step 21 we have a specific equality condition $(r_{21}[18] = r_{21}'[18] = r_{20}[18] = r_{20}'[18])$ which yet can also be handled with Theorem 2 and Theorem 3 (see Remark 3).

Example 2. (Step 48) Following Theorem 3 we decompose $x_{(48)} = x_1 \cdot 2^{32-sh(48)} + x_0$ with $0 \leq x_0 < 2^{32-sh(48)}$ and $0 \leq x_1 < 2^{sh(48)}$. Considering the bit conditions in Table 3 elementary considerations give $X_{[48]} = X_{48}' - X_{48} \equiv 0 (\bmod 2^{32})$ since Φ_{48} is given by the bitwise XOR-addition. Further, $\Delta_{(48)} - \Delta_{(47)} \equiv 2^{31} \pm$

$2^{31} \equiv 0 (\mathrm{mod}\, 2^{32})$, where "+" holds for Path 1 and "-" for Path 2 and 3. Using the notation from Theorem 3 (with $(X_i, X_i'), (R_{i-1}, R_{i-1}'), (R_i, R_i')$ corresponding to $(X,X'), (Y,Y'), (Z,Z')$) we conclude $\Delta_0 = \Delta_1 = \widetilde{\Delta}_0 = \widetilde{\Delta}_1 = 0$. In particular, $\mathrm{ca}(x_0, \Delta_0; sh(48)) = \mathrm{ca}(x_0, 0; sh(48)) = 0$ for all x_0, and (32) simplifies to $0 \equiv -(x_1 \text{ div } 2^{sh(48)}) \cdot 2^{sh(48)} + 0 (\mathrm{mod}\, 2^{32})$ which obviously is fulfilled for all $0 \leq x_1 < 2^{sh(48)}$. In other words, $M(\Delta_{[48]}, \Delta_{(47)}, \Delta_{(48)}, sh(48)) = M(0, \pm 2^{31}, 2^{31}, 23)) = M(0, 2^{31}, 2^{31}, 23)) = \mathrm{Z}_{2^{32}}$ since $2^{31} \equiv -2^{31} (\mathrm{mod}\, 2^{32})$, and $S_{(48)} = (\{32\}, \{\}, \{\}, \{\}, \{\}, \{1, \ldots, 31\})$. Theorem 2 (ii) yields the conditional probability (transition probability) $\mathrm{Prob}((R_{48}, R'_{48}) \in S_{(48)} \mid \Delta(X_{48}, X'_{48}) = \Delta_{[48]}, (R_{47}, R'_{47}) \in S_{(47)}) = 2^{-|F_{(48)+}|} = 2^{-1}$. Analogously, one obtains the same transition probablity 2^{-1} for path 2 and path 3.

Since $M_{(48)} = \mathrm{Z}_{2^{32}}$ by Theorem 2(ii) the exact transition probability in Step 48 equals the value that follows from simple 'condition counting'. We point out that the conditions in Steps 36 to 45 are fulfilled with probability 1 (no 'real' bit conditions), which is obvious, resp. can be verified with formula (28). (Note that formally $\Delta_{[i]} = \Delta(X_i, X_i') = 0$ and $\Delta_{(i-1)} = \Delta_{(i)} = \pm 2^{31} = 2^{31}$ in these steps.) The situation in Step 64 is different from that in the other steps since r_{64} has no impact on any other register. Hence only the modulo 2^{32}-difference $(r'_{64} - r_{64}) (\mathrm{mod}\, 2^{32})$ is relevant (cf. Example 4(iv)). Since $\Delta_{[64]} := \Delta(X_{64}, X'_{64}) = 0$ and $\Delta_{(63)} = \Delta_{(64)}$ this modulo 2^{32} condition is fulfilled with probability 1.

Due to (40) the path transition probability from Step 21 to 64 (before postadditions) reads

$$\prod_{i \in \{21,\ldots,63\} \setminus \{23,35,62\}} 2^{-|F_{(i)+} \cup F_{(i)0} \cup F_{(i)-} \cup F_{(i)1}|} \prod_{i=64} 1 \times \tag{44}$$
$$\prod_{i \in \{23,35,62\}} \Big(\mathrm{Prob}(X_i^{<<<sh(i)} + R_{i-1} (\mathrm{mod}\, 2^{32}) \in T_{(i)} \mid X_i \in M_{(i)}, R_{i-1} \in T_{(i-1)}) \times$$
$$\times \mathrm{Prob}(X_i \in M_{(i)}) \Big).$$

Example 3 treats the exceptional steps 23, 35 and 62 (cf. Remark 4 and Lemma 3).

Example 3. (i) (Step 23): As $sh(23) = 14$ following Theorem 3 we decompose $x_{(23)} = x_1 \cdot 2^{18} + x_0$ with $0 \leq x_0 < 2^{18}$ and $0 \leq x_1 < 2^{14}$. Elementary calculations give $\Delta_{[23]} = X'_{(23)} - X_{(23)} \equiv 2^{31} + 2^{31} + 2^{17} \equiv 2^{17} (\mathrm{mod}\, 2^{32})$ and $\Delta_{(23)} - \Delta_{(22)} \equiv 0 - 2^{31} \equiv 2^{31} (\mathrm{mod}\, 2^{32})$. We conclude $\Delta_0 = 2^{17}, \Delta_1 = 0, \widetilde{\Delta}_0 = 2^{17}, \widetilde{\Delta}_1 = 0$. From (33) we obtain the condition $\mathrm{ca}(x_0, \Delta_0; sh(23)) = \mathrm{ca}(x_0, 2^{17}; 14) = 0$, or equivalently, $0 \leq x_0 < 2^{17}$. Substituting into (32) we obtain $2^{31} \equiv (2^{17} - (x_1 + 0 + 0) \text{ div } 2^{14})) \cdot 2^{14} + 0 \equiv 2^{31} + 0 (\mathrm{mod}\, 2^{32})$ for all x_1. In other words, $M_{(23)} := M(\Delta_{[23]}, \Delta_{(22)}, \Delta_{(23)}, sh(23)) = M(0, 2^{31}, 0, 14) = \{x \in \mathrm{Z}_{2^{32}} \mid x[18] = 0\}$. Hence $\mathrm{Prob}(X_{(23)} \in M_{(23)}) = 0.5$. Note that $F_{(22)+} = \{32\}$ and $F_{(23)0} = \{32\}$. To finally apply (28) it remains to determine the conditional probability $\mathrm{Prob}([X^{<<<14} + R_{22}] (\mathrm{mod}\, 2^{32}) < 2^{31} \mid R_{22} < 2^{31}, X \in M_{(23)}\}$

$= \text{Prob}(X_2 + Y_2 (\text{mod}\, 2^{32}) < 2^{31})$ with independent uniformly distributed random variables X_2 and Y_2 with range $\text{Z}_{2^{31}}$. Hence the last term equals 0.5. (To be precise, the precise value is $0.5 + 2^{-32}$, but the correction term 2^{-32} is negligible.) Hence $\text{Prob}((R_{23}, R'_{23}) \in S_{(23)} \mid (R_{22}, R'_{22}) \in S_{(22)}, \Delta(X_{23}, X'_{23}) = 2^{17}) = 2^{-1} \cdot 2^{-1} = 2^{-2}$.
(ii) (Step 35): In Step 35 we have $sh(35) = 16$ and $M_{(35)} := \{x \in \text{Z}_{2^{32}} \mid x[16] = 0\}$. As in (i) we obtain $\text{Prob}(X_{35} \in M_{(35)}) = 0.5$. For $i = 35$ we have $F_{32,\neq} = \{32\}$ and $F_{=} = \{1, \ldots, 31\}$, which gives $T_{(35)} = \text{Z}_{2^{32}}$, i.e. R_{35} need not satisfy any condition. Hence the transition probability from Step 34 to Step 35 equals 2^{-1}.
(iii) (Step 62): $sh(62) = 10$. Similarly as for Step 23 and Step 35 we conclude that $x = x_1 \cdot 2^{22} + x_0 \in M_{(62)}$ iff $0 \leq x_0 < 2^{22} - 2^{15}$. Hence $\text{Prob}(X_{62} \in M_{(62)}) = 1 - 2^{-7}$. Since $F_{(62)+} = \{26, 32\}$ and $F_{(62)=} = \{1, \ldots, 32\} \setminus \{26, 32\}$ we immediately obtain $\text{Prob}([X^{<<<10} + R_{61}] (\text{mod}\, 2^{32}) \in T_{(62)} \mid R_{61} \in T_{(61)}) = 2^{-2}$ by Theorem 2(ii). For path 1 we further compute $\text{Prob}([X^{<<<10} + R_{61}] (\text{ mod } 2^{32}) \in T_{(62)} \mid X \notin M_{(62)}, R_{61} \in T_{(61)}) = \text{Prob}(127 \cdot 2^{25} + X_2 + 2^{27} Y_3 + 2^{25} + Y_2 (\text{mod } 2^{32}) \in T_{(62)}) = \text{Prob}(X_2 + Y_2 + 2^{27} Y_3 (\text{mod}\, 2^{32}) \in T_{(62)}) = \text{Prob}(X_2 + Y_2 < 2^{25}) \cdot \text{Prob}(Y_3 < 16)$ where X_2, Y_2, Y_3 denote independent uniformly distributed random variables with range $\text{Z}_{2^{25}}$, $\text{Z}_{2^{25}}$, and Z_{2^4}, respectively. Hence this probability equals $2^{-1} \cdot 1 = 2^{-1}$. Analogously, for path 2 and 3 the last probability equals $\text{Prob}(Y_3 + 16 (\text{mod}\, 32) < 16) = 0$, and thus the product is 0. Finally, note that $\text{Prob}(\cdot \mid X_{62} \in M_{(62)}, R_{61} \in T_{(61)}) 127/128 = \text{Prob}(\cdot \mid R_{61} \in T_{(61)}) - \text{Prob}(\cdot \mid X_{62} \notin M_{(62)}, R_{61} \in T_{(61)}) 1/128$, yielding the values 63/254, 64/254 and 64/254, respectively. Finally, the respective transition probabilities are 63/256, $64/256 = 1/4$ and $64/256 = 1/4$. Interestingly, for path 2 and 3 the transition probabilities coincide with the values from bit condition counting.

4.2 The Impact of the Postadditions on Path Probabilities

In this subsection we quantify the impact of bit conditions for the chaining values on the probabilities of hash collision paths. In the previous subsection we approximated the conditional probabilities (39) with regard to our stochastic model by conditional probabilities (40), which allowed to apply Theorem 2 and Theorem 3. Using a similar in Step 61 to 63 we consider conditional probabilities

$$\text{Prob}((R_i^p, R_i'^p) \in S_{(i),p} \mid (R_i, R'_i) \in S_{(i)}, (R_{i-64}, R'_{i-64}) = (r_{i-64}, r'_{i-64}) \quad (45)$$

which allows to apply Theorem 1 with $(X, X') = (R_i, R'_i), (Y, Y') = (r_{i-64}, r'_{i-64})$, $S_{(1)} = S_{(i)}$, $S_{(2)} = \{(r_{i-64}, r'_{i-64})\}$ and $S_{(3)} := S_{(i),p}$. In Step 64 we apply Theorem 2(ii) with $\Delta_{[1]} = \Delta(R_{64}, R'_{64})$ and $sh = 0$. In the first message block $r_{i-64} = r'_{i-64}$.

For the last block of a multiblock collision (i.e., the first block in a one-block collision) we have $S_{(i),p} = S(\cdots)$ with $F_{(i)=,p} = \{1, \ldots, 32\}$ and hence $T_{(i),p} = \text{Z}_{2^{32}}$. Consequently, we have

$$\text{Prob}\left((R_i^p, R_i'^p) \in S_{(i),p} \mid (R_i, R'_i) \in S_{(i)}, (R_{i-64}, R'_{i-64}) = (r_{i-64}, r'_{i-64})\right) = 1, \quad (46)$$

provided, of course, that $\Delta_{(i)} + \Delta_{[i-64]} \equiv \Delta_{(i),p} = 0 (\mathrm{mod}\, 2^{32})$. In contrast, for near-collisions $R^p_{(i)} \neq R'^p_{(i)}$ for at least one $i \in \{N-k+1, \ldots, N\}$. Unequal register pairs fulfil certain modulo 2^{32}-conditions and / or bit conditions. The probabilities in Example 4(i) to (iii) refer to the standard IV = (0x 67452301, 0x efcdab89, 0x 98badcfe, 0x 10325476), i.e. $r_{-3} = 0\text{x}\,67452301$, $r_{-2} = 0\text{x}\,10325476$, $r_{-1} = 0\text{x}\,98\text{badcfe}$, and $r_0 = 0\text{x}\,\text{efcdab89}$.

Example 4. (i) (Postaddition in Step 61): We first note that collision path 1 (see Table 3) fulfils $\Delta_{(61)} + \Delta_{(-3)} \equiv \Delta_{(61),p} (\mathrm{mod}\, 2^{32})$, and Lemma 3 and Remark 4 imply that $(R_{(61)}, R'_{(61)})$ is uniformly distributed on $S_{(61)}$. As $S_{(-3)}$ is singleton we may apply (27) from Theorem 1 with $F_{(61)+} = \{32\}, F_{(61)0} = \{27\}, F_{(61)1} = \{26\}$ and $F_{(61)+,p} = \{32\}$. It remains to determine the probability $\mathrm{Prob}([X + r_{-3}] (\mathrm{mod}\, 2^{32}) \in [0, 2^{31} - 1) \mid X[26] = 1, X[27] = X[32] = 0\}$ for uniformly distributed random variable X. Let X_1 and X_3 denote independent uniformly distributed random variables with range $\mathrm{Z}_{2^{25}}$, resp. Z_{2^4}. The last probability can be rewritten to $\mathrm{Prob}([X_1 + 2^{25} + 2^{27} X_3 + r_{-3}] (\mathrm{mod}\, 2^{32}) \in [0, 2^{31}))$. Similarly, $r_{-3} = c_1 + c_2 2^{25} + c_3 2^{27} + c_4 2^{31}$ with $c_1 \in [0, 2^{25})$, $c_2 \in [0, 4)$, $c_3 \in [0, 16)$, and $c_4 \in [0, 2)$. For the standard IV we have $c_2 = 3$, $c_3 = 12$, and $c_4 = 0$. Since $r_{-3} < 2^{31}$ the above probability simplifies to $\mathrm{Prob}((X_1 + c_1) + (X_3 + 12 + 1) 2^{27} \in [0, 2^{31}))$. As $0 < c_1 + X_1 < 2^{26}$ this expression equals $\mathrm{Prob}((X_3 + 12 + 1) 2^{27} \in [0, 2^{31})) = \mathrm{Prob}(X_3 + 13 < 16) = 3/16 = 0.1875$. For collision path 2 and collision path 3 from Table 3 we have $F_{(61)-} = \{32\}$ instead of $F_{(61)+} = \{32\}$. The same argumentation as above then yields $\mathrm{Prob}((X_1 + c_1) + (X_3 + 12 + 1) 2^{27} + 2^{31} \in [2^{32}, 2^{32} + 2^{31}))$ which can be reduced to $\mathrm{Prob}(X_3 + 13 \geq 16) = 13/16 = 0.81250$.
(ii) (Postaddition in Step 62): As $M_{(62)} \neq \mathrm{Z}_{2^{32}}$ (Example 3(iii)) the random vector (R_{62}, R'_{62}) may not be uniformly distributed on $S_{(62)}$. However, since $|M_{(62)}| / |\mathrm{Z}_{2^{32}}| = 1 - 2^{-7}$ and since (R_{61}, R'_{61}) is uniformly distributed on $S_{(61)}$ we may assume that (R_{62}, R'_{62}) is at least 'almost' uniformly distributed on $S_{(62)}$. For this reason we yet applied (27) (instead of (26), which was correct in a strict sense) to simplify calculations. Similar techniques as in (i) yield the transition probability 0.789 for all three paths.
(iii) (Postaddition in Step 63): Similar techniques as in (i) yield the transition probabilities 0.034, 0.148 and 0.516 for path 1, path 2 and path 3.
(iv) (Postaddition in Step 64): Theorem 2(ii) implies that the postaddition transition probability equals 2^{-4} for all paths.
(v) The probabilities for the postadditions change when IVs are used that are not standard-conformant. For collision path 2, for example, for IV=(0x 80000000, 0x efcdab89, 0x 82000000, 0x 00000000) the joint transition probability for the postadditions in Step 61 - 63 equals 0.5. In contrast, IV=(0x 00000000, 0x efcdab89, 0x 80000000, 0x 82000000) gives the joint transition probability 0 (impossible transition).

Example 4 underlines the impact of the IV, or more precisely of the combination of the IV (resp., the previous chaining value) with bit conditions $\Delta_B(R^p_i, R'^p_i)$ on the transition probabilities, at least for fixed differential schemes, favouring prefix attacks. However, also variable differential schemes may not ac-

cept bit differences $\Delta_B(\cdots)$ with large Hamming weight as this complicates the message modification in the next block.

4.3 Overall Collision Path Probabilities

The results from Subsections 4.1 and 4.2 yield the overall probabilities for the near-collision paths 1, 2, and 3 (cf. Table 3) after message modification. Table 2 below contains the probabilities for the exceptional steps 23, 35 and 62 (Example 3) and the postadditions (Example 4), the theoretically computed path probabilities ('theor. prob.') for the MD5 standard IV, the relative frequencies obtained by practical experiments ('rel. frequency') and the number of bit conditions per collision path ('bit cond.). The relative frequencies were computed from $2^{41.866}$ many samples. (Of course, this sample size is too small to provide stable relative frequencies for path 1.)

Table 2 underlines that there are significant differences between the true probabilities and their coarse estimates gained from bit condition counting. Interestingly, although near-collision path 3 demands one bit condition more than the near-collision paths 1 and 2 (39 in place of 38; giving the coarse probability estimate 2^{-39} and 2^{-38}) it is the most probable one.

steps	23	35	62	61p	62p	63p	64p	rest	theor. prob.	rel. frequency	bit cond.
Path 1	2^{-2}	2^{-1}	63/256	0.1875	0.789	0.034	2^{-4}	2^{-25}	$2^{-41.65}$	$2^{-40.86}$	38
Path 2	2^{-2}	2^{-1}	2^{-2}	0.8125	0.789	0.148	2^{-4}	2^{-25}	$2^{-37.40}$	$2^{-37.11}$	38
Path 3	2^{-2}	2^{-1}	2^{-2}	0.8125	0.789	0.516	2^{-4}	2^{-26}	$2^{-36.60}$	$2^{-36.25}$	39

Table 2. Transition probabilities for the three paths in Table 3

Our experiments showed that also other (slightly different) near-collision paths as listed in Table 2 may lead to the near-collisions that satisfy the bit conditions after the postadditions. As already pointed out, the path probabilties of concrete near-collision paths only give upper bounds for the workload of collision attacks. Usually, this effect should relax the impact of the IV.

The probabilities of the collision paths in the second block are significantly larger than the probabilities of the near-collision paths in the first block. This is due to the fact that the modulo 2^{32} differences of the chaining values of the first block and the modulo 2^{32} differences of the register values $61, \ldots, 64$ of the second block shall add up to 0 (cf. (46)), defining unique bit conditions. We just note that a particular sample path after message modification in Steps 1 to 20 occurs with probability $2^{-30.01}$.

4.4 Applicability to SHA-1

The SHA-1 step function reads

$$r_i \equiv {r_{i-1}}^{<<<5} + \Phi_i(r_{i-2}, r_{i-3}, r_{i-4}) + r_{i-5} + \widetilde{m}_i + const_i \pmod{2^{32}} \quad (47)$$
$$r_{i-2} := {r_{i-2}}^{<<<30}.$$

In analogy to the MD5 case we may set $X := \Phi_i(R_{i-2}, R_{i-3}, R_{i-4}) + R_{i-5} + \widetilde{M_i} + const_i \pmod{2^{32}}$ and $Y := R_{i-1}$, and apply a pendant of Theorem 2 with interchanged roles of X and Y (concerning the shift operations). In fact, (X, X') and X may be assumed to be uniformly distributed on $\{(x, x + \Delta_{[1]} \pmod{2^{32}}) \mid x \in \mathrm{Z}_{2^{32}}\}$ and $\mathrm{Z}_{2^{32}}$, resp., whereas (Y, Y') assumes values in a particular set $S_{(2)} := S(\cdots)$. The shift in the second line of (47) just transforms an '$S(\cdots)$'-set into another '$S(\cdots)$'-set (Remark 2(i)) and hence does not cause principal problems.

5 Conclusion

We developed a new methodology to compute probabilities of differential paths that is based on a thorough analysis of two-dimensional random vectors that assume values in the product space $\mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}}$. We proved three stochastic theorems that allow to simplify special types of conditional probabilities of random vectors to conditional probabilities of their projections onto the first component. This facilitates concrete computations considerably, especially since the components of the random vectors are strongly correlated. For MD5 we illustrated the use of this approach, and we confirmed experimentally that these theorems support the effective computation of probabilities for given (near-)collision paths after message modification. In particular, the computed path probabilities were found to be in conformance with experimental results. Our method is not a ready-to-use tool but it is applicable to a wide class of collision attacks. It may be expected that similar calculations deliver reliable results (e.g.) for SHA-1 collision paths, too, where the knowledge of exact probabilities is certainly more relevant. An interesting observation is the significant impact of the postadditions and the IV, especially on fixed differential schemes.

Acknowledgement: We would like to thank Søren Thomsen for making his paper [Th] available to us.

References

[BCH] J. Black, M. Cochran, T. Highland, *A Study of the MD5 Attacks: Insights and Improvements* , FSE 2006, Springer, LNCS 4047 (2006), 262–277

[Daum] M. Daum, *Cryptanalysis of Hash Functions of the MD4-Family*, PhD thesis, Ruhr-Universität Bochum, June 2005

[DL] M. Daum, S. Lucks, *The Story of Alice and Bob*, Presented at the rump session of Eurocrypt '05, May 2005, online at http://www.cits.rub.de/imperia/md/content/magnus/rump_ec05.pdf

[DMR] C. De Canniere, F. Mendel, C. Rechberger, *On the Full Cost of Collision Search for SHA-1* , Workshop Proceedings of the ECRYPT Hash Workshop 2007, 24-25 May 2007, Barcelona, Spain, 174–189

[DR] C. De Canniere, C. Rechberger, *Finding SHA-1 Characteristics: General Results and Applications*, ASIACRYPT 2006, Springer, LNCS 4284 (2006), 1–20

[GIS1] M. Gebhardt, G. Illies, W. Schindler, *A Note on the Practical Value of Single Hash Collisions for Special File Formats*, Sicherheit 2006 — 'Sicherheit — Schutz und Zuverlässigkeit', Köllen, LNI P-77 (2006), 333-344.
Extended version: NIST Cryptographic Hash Workshop 2005, online at http://www.csrc.nist.gov/pki/Hashworkshop/2005/Oct31_Presentations/.. ..Illies_NIST_05.pdf

[GIS2] M. Gebhardt, G. Illies, W. Schindler, *The Impact of the IV on Multiblock Hash Collision Paths*, FSE 2006, rump session, 16 Mar 2006. http://fse2006.iaik.tugraz.at/rumpsession.html

[GIS3] M. Gebhardt, G. Illies, W. Schindler, *Precise Probabilities of Hash Collision Paths*, Second Cryptographic Hash Workshop, NIST http://www.csrc.nist.gov/pki/HashWorkshop/2006/Papers/

[GIS4] M. Gebhardt, G. Illies, W. Schindler, *On An Approach to Compute (at least Almost) Exact Probabilities for Differential Hash Collision Paths*, to appear in: Sicherheit 2008 — 'Sicherheit, Schutz und Zuverlssigkeit', Köllen, LNI series (2008)

[HPR] P. Hawkes, M. Paddon, G. D. Rose, *Musing on the Wang et. al. MD5 Collision*, Cryptology ePrint Archive, Report 2004/264, http://eprint.iacr.org/2004/264

[Kli1] V. Klima, *Finding MD5 Collisions on a Notebook PC Using Multi-messagenModifications*, Cryptology ePrint Archive, Report 2005/102, http://eprint.iacr.org/2005/102

[Kli2] V. Klima, *Tunnels in Hash-Functions: MD5 Collisions Within a Minute*, Cryptology ePrint Archive, Report 2006/105, http://eprint.iacr.org/2006/105

[LiLa] J. Liang, X. Lai, *Improved Collision Attack on Hash Function MD5*, Cryptology ePrint Archive, 23 Nov 2005 , Report 2005/425, http://eprint.iacr.org/2005/425

[MOV] A. Menezes, P. C. van Oorschot, S. A. Vanstone, *Handbook of Applied Cryptography*, CRC Press, Boca Raton, 1997

[MPRR] F. Mendel, N. Pramstaller, C. Rechberger, V. Rijmen, *The Impact of Carries on the Complexity of Collision Attacks*, FSE 2006, Springer, LNCS 4047 (2006), 278–292

[SNKO] Y. Sasaki, Y. Naito, N. Kunihiro, K. Ohta, *Improved Collision Attack on MD5*, Cryptology ePrint Archive, 07 Nov 2005 , Report 2005/400, http://eprint.iacr.org/2005/400

[SO] M. Schläffer, E. Oswald, *Searching for Differential Paths in MD4*, FSE 2006, Springer, LNCS 4047 (2006), 242–261

[SLW] M. Stevens, A.K. Lenstra, B.M.M. de Weger, *Target collisions for MD5 and colliding X.509 certificates for different identities*, Cryptology ePrint Archive, 04 Nov 2006 , Report 2006/360, http://eprint.iacr.org/2006/360

[SLW2] M. Stevens, A.K. Lenstra, B.M.M. de Weger, *Chosen-prefix Collisions for MD5 and Colliding X.509 Certificates for Different Identities*, Eurocrypt 2007, Springer, LNCS 4515 (2007), 1–22

[St] M. Stevens, *Fast Collision Attack on MD5*, Cryptology ePrint Archive, 17 Mar 2006 , Report 2006/104, http://eprint.iacr.org/2006/104

[Th] S. Thomsen, *Cryptographic Hash Functions*, Master thesis, Technical University of Denmark, November 2005

[WLFCY] X. Wang, X. Lai, D. Feng, H. Chen and X. Yu, *Cryptanalysis of the Hash Functions MD4 and RIPEMD*, EuroCrypt 2005, Springer, LNCS 3494 (2005), 1–18

[WY] X. Wang and H. Yu , *How to Break MD5 and Other Hash Functions*, EuroCrypt 2005, Springer, LNCS 3494 (2005), 19–35

[WYaYa] X.Wang, A. Yao, F. Yao, *New Collision Search for SHA-1*, Presented by Adi Shamir at the rump session of Crypto '05, Aug 2005, online at http://www.iacr.org/conferences/crypto2005/rumpSchedule.html

[WYiY] X. Wang, Y. L. Yin, H. Yu, *Collision Search Attacks on SHA-1*, Crypto 2005, Springer LNCS 3621 (2005), 17-36

[WYuY] X. Wang, H. Yu, Y. L. Yin, *Efficient Collision Search Attacks on SHA0*, Crypto 2005, Springer, LNCS 3621 (2005), 1-16

[YaSh] J. Yajima, T. Shimoyama, *Wang' s sufficient conditions on MD5 are not sufficient*, Cryptology ePrint Archive, 10 Aug 2005 , Report 2005/236, http://eprint.iacr.org/2005/236

Appendix

Table 3 contains bit conditions for three MD5-near-collision paths for block 1. If the conditions for Step i are the same for each path the three columns are merged to a single column. The terms [j] and [-j] were already defined in Sect. 3. Further, $r_{i,j}$ denotes the j^{th} bit of r_i, and [*32] stands for $r_{i,32} \neq r'_{i,32}$. The additional conditions in Step 23, Step 35, and Step 62 (e.g., $x_0 < 2^{17}$ in Step 23; cf. Example 3) are not mentioned in Table 3. The conditions for Step 1 to Step 20 are as in [WY] (apart from additional conditions as in Steps 21, 35 and 62). Path 1 corresponds to the published bit conditions in [WY] while their published collision satisfies the bit conditions of path 2.

The mapping $\Phi_i \colon \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \to \mathrm{Z}_{2^{32}}$ is bit-oriented, i.e. $\Phi_i(a,b,c) = (\Phi_{i;b}(a_{32},b_{32},c_{32}),\ldots,\Phi_{i;b}(a_1,b_1,c_1))$ with $\Phi_{i;b} \colon \mathrm{Z}_2 \times \mathrm{Z}_2 \times \mathrm{Z}_2 \to \mathrm{Z}_2$. In particular,

$$\begin{aligned}
\Phi_{i,b}(a_j,b_j,c_j) &:= (a_j \wedge b_j) \vee (\neg a_j \wedge c_j) \text{ for } i = 1,\ldots,16\\
\Phi_{i,b}(a_j,b_j,c_j) &:= (a_j \wedge c_j) \vee (b_j \wedge \neg c_j) \text{ for } i = 17,\ldots,32\\
\Phi_{i,b}(a_j,b_j,c_j) &:= (a_j + b_j + c_j) (\mathrm{mod}\, 2) \text{ for } i = 33,\ldots,48\\
\Phi_{i,b}(a_j,b_j,c_j) &:= b_j + (a_j \vee \neg c_j) (\mathrm{mod}\, 2) \text{ for } i = 49,\ldots,64.
\end{aligned}$$

Step	Shift	Path 1	Path 2	Path 3
19	14	$[18, 32]$		
20	20	$[32]$		
21	5	$[32]$, $r_{21,18} = r_{20,18}$		
22	9	$[32]$		
23	14	$r_{23,32} = 0 = r'_{23,32}$		
24	20	$r_{24,32} = 1 = r'_{24,32}$		
25... 34	...			
35	16	$[*32]$		
36	23	$[*32]$		
37	4	$[*32]$		
38	11	$[*32]$		
39	16	$[*32]$		
40	23	$[*32]$		
41	4	$[*32]$		
42	11	$[*32]$		
43	16	$[*32]$		
44	23	$[*32]$		
45	4	$[*32]$		
46	11	$[32]$		
47	16	$[32]$	$[-32]$	$[-32]$
48	23	$[32]$		
49	6	$[32]$	$[-32]$	$[-32]$
50	10	$[-32]$		
51	15	$[32]$	$[-32]$	$[-32]$
52	21	$[-32]$		
53	6	$[32]$	$[-32]$	$[-32]$
54	10	$[-32]$		
55	15	$[32]$	$[-32]$	$[-32]$
56	21	$[-32]$		
57	6	$[32]$	$[-32]$	$[-32]$
58	10	$[-32]$		
59	15	$[32]$	$[-32]$	$[-32]$
60	21	$[32]$, $r_{60,26} = 0 = r'_{60,26}$		
61	6	$[32]$	$[-32]$	$[-32]$
61		$r_{61,27} = 0 = r'_{61,27}$, $r_{61,26} = 1 = r'_{61,26}$		
62	10	$[32, 26]$		
63	15	$[32, 26]$	$[-32, 26]$	$[-32, 27, -26]$
64	21	$r'_{64} - r_{64} = 2^{31} + 2^{25} \pmod{2^{32}}$		
61,p		$[32]$		
62,p		$[32, 26]$		
63,p		$[32, 27, -26]$		
64,p		$[32, 26]$, $r^p_{64,27} = 0 = r'^p_{64,27}$, $r^p_{64,6} = 0 = r'^p_{64,6}$		

Table 3. Three MD5 near-collision paths in the 1^{st} block (message modification ends with Step 20)