Compression Function Design Principles Supporting Variable Output Lengths from a Single Small Function*

Donghoon Chang[1], Mridul Nandi[2], Jesang Lee[1], Jaechul Sung[3], Seokhie Hong[1]

[1] Center for Information and Security Technologies
Korea University, Seoul, Korea
{dhchang, jslee, hsh}@cist.korea.ac.kr
[2] CINVESTAV-IPN, Mexico City
mridul.nandi@gmail.com
[3] Department of Mathematics, University of Seoul, Korea
jcsung@uos.ac.kr

Abstract. In this paper, we introduce new compression function design principles supporting variable output lengths (multiples of size n). They are based on a function or block cipher with an n-bit output size. In the case of the compression function with a $(t+1)n$-bit output size, in the random oracle and ideal cipher models, their maximum advantages from the perspective of collision resistance are $O(\frac{t^2 q}{2^{tn}} + \frac{q^2}{2^{(t+1)n}})$. In the case of $t = 1$, the advantage is near-optimal. In the case of $t > 1$, the advantage is optimal.

Keywords : Hash function, Random oracle, Ideal cipher model.

1 Introduction.

In 2004 and 2005, Wang *et al.* [19–22] introduced a new strategy for finding a collision of widely used hash functions such as MD5 [13], SHA-1 [6] etc. Since fatal weaknesses of MD5 and SHA-1 were exposed by Wang *et al.*, many cryptographers have recognized the necessity of new hash functions as their replacements. Impelled by this recognition, NIST announced the development of one or more additional hash algorithms via a public competition similar to AES [7]. NIST also announced that the algorithm must support 224, 256, 384, and 512-bit message digests, and a maximum message length of at least 2^{64} bits. Therefore, it is important to develop provably secure design principles supporting variable length output. As one method, for each output length we can design hash functions independently similar to SHA-family. We can also design stream cipher-style hash functions such as RadioGatún [1] and RC4-Hash [4], which use an iterative function to obtain the desired size of hash output. As an alternative method, we can design variable output length-hash functions with one small output-length

* This paper was accepted to IEICE Fundamentals. It will appear in Sep. 2008.

algorithm. Double Block Length (DBL) hash functions by Nandi [11] and Hirose [8, 9] are representative of this. Nandi [11] proved that his construction has the near-optimal collision resistance in the random oracle model. Based on his idea, Hirose [8, 9] proposed block cipher based DBL hash functions and proved its near-optimality of the collision resistance. However, since they only considered DBL hash functions, their constructions have the limitation that they can not support variable sizes of hash outputs.

In this paper, we focus on designing compression functions because the Merkle-Damgård construction can be used to construct a hash function from a compression function. Via the generalization of Nandi's result we show that we can handle arbitrary output lengths of compression functions from a small function. Especially, in the case of our construction with at least a $3n$-bit output size where n-bit is the output size of the underlying function, we prove the optimal collision resistance in the random oracle model. Furthermore, we propose a new block cipher based compression function with variable output length and prove its near-optimal collision resistance (in the case of $2n$-bit output size) and optimal collision resistance (in the case of at least $3n$-bit output size). Based on the results, with a 128-bit RC6 [14] and 64-bit Blowfish [15] (where RC6 has a maximum key size of 2,040 bits, and Blowfish has that of 448 bits.) we can design a hash function to handle a maximum of 1,920-bit and 384-bit compression function outputs, respectively.

2 Definitions and Previous Works

In this section, we define the notation and symbols and describe previous work.

Random Oracle. Let $\mathsf{Func}(m, n)$ denote the set of all functions from $\{0,1\}^m$ to $\{0,1\}^n$. In the random oracle model, a function f is chosen at random from $\mathsf{Func}(m, n)$ and any adversary in the random oracle model can access the function f as if it is a black-box. It is clear that for any $x_i \in \{0,1\}^m$, (it can be any function of $x_1, \cdots, x_{i-1} \in \{0,1\}^m$ and $y_1, \cdots, y_{i-1} \in \{0,1\}^n$) and $y_i \in \{0,1\}^n$ such that $x_i \neq x_j$ for all $j < i$ the following is valid:

$$\mathsf{Pr}[f(x_i) = y_i | f(x_1) = y_1, \cdots, f(x_{i-1}) = y_{i-1}] = 1/2^n.$$

Note that x_i can be any random variable independent of the random oracle f such that $x_i \neq x_j$ for all $j < i$ with probability 1. Let A^f be any adversary with access to the random oracle f, and suppose x_i is the i-th query and y_i is the i-th response of the random oracle. In this paper, we assume that all queries are distinct, that is, $x_i \neq x_j$ for all $i \neq j$. This is obviously a reasonable assumption. Under this assumption, the conditional probability distribution of the i-th response is uniformly and independently distributed over $\{0,1\}^n$.

Ideal Cipher. The ideal cipher E has an n-bit block size with a k-bit key size. For any key $a \in K$, $E_a(\cdot)$ is a random permutation. Equivalently, the ideal cipher E is selected randomly from $\mathsf{Block}(k, n)$, which denotes the set of all block ciphers with an n-bit block size and a k-bit key size. For a key-plaintext query $(1, a, x)$, the ideal cipher outputs $y = E_a(x)$. For a key-ciphertext query $(-1, a, y)$, the ideal cipher outputs $x = E_a^{-1}(y)$. We denote the j-th query-response pair by (w_j, a_j, x_j, y_j), where $w_j = 1$ means the encryption query, $w_j = -1$ means the decryption query, a_j is a key, x_j is a plaintext, and y_i is a ciphertext.

Padding Rule. A padding rule g has an input of arbitrary length and an output of a multiple of $d-s$ ($d \geqslant s+64$) which are defined in MD_g^F in the next paragraph. There are many kinds of padding rules but, here, we fix g for any $M \in \{0,1\}^*$ as follows:

$$g(M) = M||10^t||\mathsf{bin}_{64}(|M|),$$

Where t is the smallest non-negative integer such that $g(x)$ is a multiple of $d-s$ and $\mathsf{bin}_{64}(x)$ means the 64-bit binary representation of x.

MD_g^F Construction. $\mathsf{MD}_g^F : \{0,1\}^* \rightarrow \{0,1\}^s$ is the design principle proposed by Merkle and Damgård. It is the method of designing a hash function from a compression function $F : \{0,1\}^d \rightarrow \{0,1\}^s$ with the padding rule g [5, 10]. They proved that if F is collision resistant then MD_g^F is also collision resistant. MD_g^F is defined as follows.

$$\begin{aligned}\mathsf{MD}_g^F(M) &= \mathsf{MD}^F(g(M)) \\ &= F(\cdots F(F(F(IV, m_0), m_1), m_2) \cdots, m_t)\end{aligned}$$

Where $g(M) = (m_0||m_1||\cdots||m_t)$, IV (or h_0) is the s-bit initial value, $h_i = F(h_{i-1}, m_i)$, and each m_i is $d-s$ bits.

Non-adaptive and Adaptive Models. When the adversary A is permitted to make q queries to a given oracle, in the non-adaptive model A can only make a maximum of q queries simultaneously, then, A can obtain all responses. In the adaptive model A can make the i-th query after he obtains $i-1$ query-responses. In this paper, we consider the adaptive model, which is the strongest model from the security perspective. Moreover, an adaptive adversary A can be reasonably considered for a public compression function or public block cipher as a random oracle model. An adversary A can adaptively compute the outputs of these.

We assume that adversary A can make a maximum of q queries, and that he is deterministic and computationally unbounded. It is easy to prove that if a scheme is secure from all deterministic and computationally unbounded adversaries, then, it is secure against all probabilistic adversaries. Let the i-th query be x_i and y_i be the oracle response $\mathcal{O}(x_i)$. We define the view $\mathcal{V}_A^{\mathcal{O}}(i) = ((x_1, y_1), (x_2, y_2), \cdots, (x_i, y_i))$ which is all A's information. Here, since A is adaptive and deterministic, his i-th query is uniquely determined from the view $\mathcal{V}_A^{\mathcal{O}}(i-1)$. Equivalently,

$$A^{\mathcal{O}}((x_1, y_1), (x_2, y_2), \cdots, (x_{i-1}, y_{i-1})) = x_i.$$

In the ideal cipher model, it can be similarly defined.

Collision Resistance. We only focus on security against collision resistance. Informally, collision resistance means the difficulty of finding two different inputs X and X' such that their hash outputs are identical. Firstly, we define the collision resistant measurement of compression function F and hash function MD_g^F in the random oracle and ideal cipher models. We assume that F is constructed from the random oracle f or ideal cipher E. We assume that the adversary A is deterministic and computationally unbounded and he can make a maximum of q queries. Then, we can define the collision resistance of the compression function F to the adversary A in the random oracle model as follows:

$$\begin{aligned}\mathsf{Adv}_F^{coll}(A(q)) = \mathsf{Pr}[&f \leftarrow_R \mathsf{Func}(m,n); X, Y \leftarrow A^f(q) \\ &: (X \neq Y) \wedge (F(X) = F(Y))].\end{aligned}$$

Similarly, we can define the collision resistance of compression function F to the adversary A in the ideal cipher model as follows:

$$\begin{aligned}&\mathsf{Adv}_F^{coll}(A(q)) = \mathsf{Pr}[E \leftarrow_R \mathsf{Block}(k,n); \\ &\quad X, Y \leftarrow A^{E,E^{-1}}(q) : (X \neq Y) \wedge (F(X) = F(Y))].\end{aligned}$$

The collision resistance of MD_g^F is similarly defined as follows:

$$\begin{aligned}\mathsf{Adv}_{\mathsf{MD}_g^F}^{coll}(A(q)) &= \mathsf{Pr}[f \leftarrow_R \mathsf{Func}(m,n); M, M' \leftarrow A^f(q) \\ &\quad : (M \neq M') \wedge (\mathsf{MD}_g^F(M) = \mathsf{MD}_g^F(M'))]. \\ \mathsf{Adv}_{\mathsf{MD}_g^F}^{coll}(A(q)) &= \mathsf{Pr}[E \leftarrow_R \mathsf{Block}(k,n); \\ &\quad\quad M, M' \leftarrow A^{E,E^{-1}}(q) : \\ &\quad\quad\quad (M \neq M') \wedge (\mathsf{MD}_g^F(M) = \mathsf{MD}_g^F(M'))].\end{aligned}$$

We also define their maximum advantages over all adversaries as follows:

$$\mathsf{Adv}_F^{coll}(q) = \mathrm{Max}_A[\mathsf{Adv}_F^{coll}(A(q))].$$

$$\mathsf{Adv}_{\mathsf{MD}_g^F}^{coll}(q) = \mathrm{Max}_A[\mathsf{Adv}_{\mathsf{MD}_g^F}^{coll}(A(q))].$$

We state that F (or MD_g^F) is collision resistant (has collision resistance) if the maximum advantage is negligible. Especially, we state that F (or MD_g^F) is optimally collision resistant (has optimal collision resistance) if the maximum advantage is $O(q^2/2^s)$, where the output length is s-bit. We know the following by [5, 10].

$$\mathsf{Adv}_{\mathsf{MD}_g^F}^{coll}(q) \leqslant \mathsf{Adv}_F^{coll}(q).$$

The above relation means that if F is collision resistant (optimally collision resistant) then MD_g^F is also collision resistant (optimally collision resistant). So, we focus on designing a compression function F and showing that the upper bound of $\mathsf{Adv}_F^{coll}(q)$ is negligible. Especially, we show that our compression functions are near-optimally or optimally collision resistant in the random oracle or ideal cipher models.

Note 1. We assume that the adversary does not repeatedly make identical queries. In the random oracle model, all queries x_i's are different. In the ideal cipher model, once he obtains (a, x, y) such that $E_a(x) = y$ (or $E_a^{-1}(y) = x$), he does not make a decryption query (a, y) or encryption query (a, x). Secondly, as mentioned in [3], we assume that when the adversary A finally outputs two message X and X', A has already computed $F(X)$ and $F(X')$, in the sense that A has made necessary oracle queries to compute $F(X)$ and $F(X')$. As described in [18], if the second assumption is not made, the adversary can output two very long messages (which are not related to A's view) which will collide with a high probability. As mentioned in [3], an adversary A not obeying these assumptions can easily be modified to given an adversary A' having similar computational complexity that obeys these assumptions and has the same advantage as A.

Note 2. Our goal is to show that the maximum advantage of our design principles are negligible in the random oracle and ideal cipher models. According to the definition of the advantage from the perspective of collision resistance, the advantage arises from the probability that final outputs of the adversary collide. Based on the second assumption in the Note 1, if the adversary finds collisions, the collisions should be constructed from the final view. Equivalently, without considering final outputs of the adversary, we can directly obtain the upper bound of the advantage from the final view. So, we focus on the probability that there is a collision constructed from the final view.

Nandi [11]. Nandi proposed the following compression function F from a function f of small output size n. Since the output size of F is double of that of f, we call F a Double Block Length (DBL) compression function:

$$F(X) = f(X)||f(P(X)),$$

where $f : \{0,1\}^m \rightarrow \{0,1\}^n$ $(m > 2n)$ and P is a permutation with no fixed point such that P^2 is the identity permutation. Also Nandi proved that F is near-optimally collision resistant in the random oracle model, where f is a random oracle. The hash function MD_g^F based on the DBL compression function F is called the DBL hash function.

Hirose [8, 9]. Hirose constructed f with a block cipher as follows [8]:

$$f(h||g||m) = E_{h||m}(g) \oplus g,$$

where $|h| = |g| = |m| = n$ and the block cipher E has a $2n$-bit key size and an n-bit block size. Hirose proved that if f is applied to Nandi's construction, MD_g^F is near-optimally collision resistant in the ideal cipher model. Hirose also proposed five other constructions [9] and proved their near-optimal collision resistance. These six constructions are classified as DBL hash functions.

Hash Rate The term Hash Rate is used to indicate the efficiency of the hash function. A rate is defined as follows:

$$\mathsf{Rate} = \frac{\mathrm{R}}{\mathrm{T} \times \mathrm{S}},$$

where R is the size of the message used in a compression function, T is the number of an atomic function used in a compression function, and S is the output size of an atomic function. For example, the rate of Nandi's construction (the atomic function is f) is $\frac{m-2n}{2n}$. In the case of Hirose's construction (the atomic function is E), the rate is $1/2$.

3 Compression Function with Variable Output Size in the Random Oracle Model

In this section, we explain Nandi's construction [11] and reformulate its proof for easy generalization.

3.1 Nandi's Construction and Its Security

In [11], Nandi proved that F is near-optimally collision resistant in the random oracle model as follows:

Theorem 1. *In the random oracle model, an upper bound of the maximum advantage from the perspective of the collision resistance of F is expressed as follows:*

$$\mathsf{Adv}_F^{coll}(q) \leqslant \frac{q-1}{2^n} + \frac{q^2-1}{2^{2n+1}}.$$

Proof : We prove the theorem in four steps. Let A be any deterministic and computationally unbounded adversary. We assume that A makes q queries to the oracle.

1. For any final view $\mathcal{V}_A^f(q) = ((x_1, y_1), (x_2, y_2), \cdots, (x_q, y_q))$ generated from the random oracle f, a maximum of q input-output pairs of F can be constructed. For an even q, there is an adversary to construct q input-output pairs of F from q input-output pairs of f.

   Proof) We assume that q input-output pairs of F are given. Equivalently, we have $\{(X_i, Y_i)\}_{1 \leqslant i \leqslant q}$ where $F(X_i) = Y_i$ and $X_i \neq X_j$ for all i and j ($i \neq j$). Since $F(X) = f(X)||f(P(X))$, we need to make at least q queries X_i ($1 \leqslant i \leqslant q$) to the random oracle f to obtain q input-output pairs of F. Therefore, a maximum of q input-output pairs of F can be constructed from q

input-output pairs of f. Next, we need to construct an adversary to construct q input-output pairs of F from q input-output pairs of f. This is simple. In order to obtain $F(X) = f(X)||f(P(X))$, we need to make two queries X and $P(X)$, to the random oracle f. Once we obtain $F(X) = f(X)||f(P(X))$, it is clear that $F(P(X)) = f(P(X))||f(X)$ without additional queries. Therefore, we can obtain two input-output pairs of F from two input-output pairs of f. Likewise, when q is even, we can obtain q input-output pairs of F from q input-output pairs of f.

2. Let $F[\mathcal{V}_A^f(q)]$ be the set of input-output pairs of F to be generated from the final view $\mathcal{V}_A^{\mathcal{O}}(q)$. Based on the result of Step 1, we state that $F[\mathcal{V}_A^f(q)] = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_p, Y_p)\}$, where $p \leqslant q$ and $X_i \neq X_j$ for all i and j $(i \neq j)$. Here, we need to compute the probability that $F(X_i) = F(X_j)$ for any i and j (where $i \neq j$). The following is valid for any i and j, where $j < i \leqslant p$.

   (a) If $X_i = P(X_j)$: $\mathsf{Pr}[F(X_i) = F(X_j)] = \mathsf{Pr}[f(P(X_j)) = f(X_j)] = 1/2^n$.

   (b) If $X_i \neq P(X_j)$: Since $\{X_i, X_j\} \cap \{P(X_i), P(X_j)\} = \varnothing$,

   $$\begin{aligned}&\mathsf{Pr}[F(X_i) = F(X_j)]\\&= \mathsf{Pr}[f(X_i) = f(X_j) \wedge f(P(X_i)) = f(P(X_j))]\\&= \mathsf{Pr}[f(X_i) = f(X_j)|f(P(X_i)) = f(P(X_j))]\\&\quad \times \mathsf{Pr}[f(P(X_i)) = f(P(X_j))] = \mathsf{Pr}[f(X_i) =\\&\quad f(X_j)] \times \mathsf{Pr}[f(P(X_i)) = f(P(X_j))]\\&= \tfrac{1}{2^n} \times \tfrac{1}{2^n} = \tfrac{1}{2^{2n}}.\end{aligned}$$

3. Let the event C_i be the event for which there is a j (where $j < i$) such that $F(X_i) = F(X_j)$. Then, $\mathsf{Pr}[C_2] \leqslant \frac{1}{2^n}$, and, for $i > 2$, $\mathsf{Pr}[C_i] \leqslant \frac{1}{2^n} + \frac{i-1}{2^{2n}}$.

   Proof) Based on the result of Step 2-(a) and (b), $\mathsf{Pr}[C_2] = \mathsf{Pr}[F(X_2) = F(X_1)] \leqslant Max(\frac{1}{2^n}, \frac{1}{2^{2n}})$. For $i > 2$, the case of Step 2-(a) occurs at most one time, and, the case of Step 2-(b) occurs a maximum of $i-1$ times. Therefore, $\mathsf{Pr}[C_i] \leqslant \frac{1}{2^n} + \frac{i-1}{2^{2n}}$.

4. From the above results, we can compute the upper bound of the advantage of the collision resistance of F.

$$\begin{aligned}\mathsf{Adv}_F^{coll}(q) &= Max_A[\mathsf{Adv}_F^{coll}(A(q))]\\&= Max_A[\mathsf{Pr}_A[C_2 \vee C_3 \cdots \vee C_q]]\\&\leqslant Max_A[\mathsf{Pr}_A[C_2] + \textstyle\sum_{i=3}^{q} \mathsf{Pr}_A[C_i]]\\&\leqslant Max_A[\tfrac{1}{2^n} + \textstyle\sum_{i=3}^{q}(\tfrac{1}{2^n} + \tfrac{i-1}{2^{2n}})]\\&\leqslant \tfrac{q-1}{2^n} + \tfrac{q^2-1}{2^{2n+1}}. \blacksquare\end{aligned}$$

3.2 Generalization

In this subsection, we generalize the result by Nandi. Firstly, we propose the generalized construction, then, we prove its optimal collision resistance.

Generalized Construction. We need to construct F from a function f which has an m-bit input and an n-bit output such that $m > (t+1)n$.

$$F(X) = f(P_0(X))||f(P_1(X))||f(P_2(X))||\cdots||f(P_t(X))$$

Where P_0 is the identity permutation, P_i is a permutation with no fixed point such that P_i^2 is the identity permutation. For any i and j (where $i \neq j$), $P_iP_j = P_jP_i$. For all $(i_1, i_2, \cdots, i_t) \in \{0,1\}^t \setminus \{0\}^t$, $P_1^{i_1}P_2^{i_2}\cdots P_t^{i_t}$ has no fixed point. For example, in the case of $t \leqslant n$, we can define $P_i(x) = x \oplus (1000\cdots 0)^{\lll i}$ where $(1000\cdots 0)$ denotes only zero-bits except that the left-most bit is one, and $\lll i$ means the i-bit left-rotation. Then, we can prove the following theorem for $t \geqslant 2$.

Theorem 2. *In the random oracle model, an upper bound of the maximum advantage from the perspective of the collision resistance of F is expressed as follows:*

$$\mathsf{Adv}_F^{coll}(q) \leqslant \frac{t(t+3)(q-1)}{2^{tn+1}} + \frac{q^2-1}{2^{(t+1)n+1}} \quad \textit{where } t \geqslant 2.$$

Proof : Its proof is similar to that of Theorem 1. Here, A is any deterministic and computationally unbounded adversary. We assume that A makes q queries to the oracle.

1. For any final view $\mathcal{V}_A^f(q) = ((x_1, y_1), (x_2, y_2), \cdots, (x_q, y_q))$ generated from the random oracle f, a maximum of q input-output pairs of F can be constructed.

   Proof) We assume that q input-output pairs of F are given. Equivalently, we have $\{(X_i, Y_i)\}_{1 \leqslant i \leqslant q}$ where $F(X_i) = Y_i$ and $X_i \neq X_j$ for all i and j (where $i \neq j$). Since $F(X) = f(X)||f(P_1(X))||\cdots||f(P_t(X))$, we must make at least q queries X_i (where $1 \leqslant i \leqslant q$) to the random oracle to obtain q input-output pairs of F. Therefore, a maximum of q input-output pairs of F can be constructed from q input-output pairs of f.

2. Let $F[\mathcal{V}_A^f(q)]$ be the set of input-output pairs of F to be generated from final view $\mathcal{V}_A^{\mathcal{O}}(q)$. Based on the result of Step 1, we state $F[\mathcal{V}_A^f(q)] = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_p, Y_p)\}$ where $p \leqslant q$ and $X_i \neq X_j$ for all i and j (where $i \neq j$). Here, we want to compute the probability that $F(X_i) = F(X_j)$ for any i and j (where $i \neq j$). The following is valid for any i and j where $j < i \leqslant p$:

(a) If $X_i = P_u(X_j)$ (for a u, $1 \leqslant u \leqslant t$): Firstly, we compute the number of elements of $T_u = \{\{P_r P_u(X_j), P_r(X_j)\}\}_{0 \leqslant r \leqslant t}$. $r = 0$ indicates the element $\{P_u(X_j), X_j\}$ of T_u and $r = u$ indicates the element $\{X_j, P_u(X_j)\}$ of T_u. That is, $r = 0$ and $r = u$ correspond to an identical element. And by the relations among P_i's, $|T_u| = t$ and for any l, k ($l \neq k$,$\{l,k\} \neq \{0,u\}$), $\{P_l P_u(X_j), P_l(X_j)\} \cap \{P_k P_u(X_j), P_k(X_j)\} = \varnothing$. Therefore, $\mathsf{Pr}[F(X_i) = F(X_j)] = (\frac{1}{2^n})^t = \frac{1}{2^{tn}}$.

(b) If $X_i = P_v P_u(X_j)$ (for some v and u, where $1 \leqslant v < u \leqslant t$): Firstly, we compute the number of elements of $T_{v,u} = \{\{P_r P_v P_u(X_j), P_r(X_j)\}\}_{0 \leqslant r \leqslant t}$. $r = v$ indicates the element $\{P_u(X_j), P_v(X_j)\}$ of T_u and $r = u$ indicates the element $\{P_v(X_j), P_u(X_j)\}$ of T_u. That is, $r = 0$ and $r = u$ correspond to an identical element. And by the relations among P_i's, $|T_u| = t$ and for any l, k ($l \neq k$ and $\{l, k\} \neq \{u, v\}$), $\{P_l P_v P_u(X_j), P_l(X_j)\} \cap \{P_k P_v P_u(X_j), P_k(X_j)\} = \varnothing$. Therefore, $\mathsf{Pr}[F(X_i) = F(X_j)] = (\frac{1}{2^n})^t = \frac{1}{2^{tn}}$.

(c) If $X_i \neq P_v P_u(X_j)$ (for all v and u, where $0 \leqslant v < u \leqslant t$): If $T = \{\{P_r(X_i), P_r(X_j)\}\}_{0 \leqslant r \leqslant t}$, by the relations among P_i's, $|T| = t + 1$ and the intersection of any two elements of T is the empty set. Therefore, $\mathsf{Pr}[F(X_i) = F(X_j)] = (\frac{1}{2^n})^{t+1} = \frac{1}{2^{(t+1)n}}$.

3. Let C_i be the event for which there is a j (where $j < i$) such that $F(X_i) = F(X_j)$. Then, $\mathsf{Pr}[C_2] \leqslant \frac{1}{2^{tn}}$ and for $i > 2$, $\mathsf{Pr}[C_i] \leqslant \frac{t(t+3)}{2^{tn+1}} + \frac{i-1}{2^{(t+1)n}}$.

   Proof) Based on the result of Step 2-(a), (b) and (c), $\mathsf{Pr}[C_2] = \mathsf{Pr}[F(X_2) = F(X_1)] \leqslant Max(\frac{1}{2^{tn}}, \frac{1}{2^{(t+1)n}})$. Step 2-(a) contains a maximum of t cases, and Step 2-(b) contains a maximum of $\frac{t(t+1)}{2}$ cases, and Step 2-(c) contains a maximum of $i - 1$ cases. Therefore, $\mathsf{Pr}[C_i] \leqslant \frac{t}{2^{tn}} + \frac{t(t+1)}{2^{tn+1}} + \frac{i-1}{2^{(t+1)n}} \leqslant \frac{t(t+3)}{2^{tn+1}} + \frac{i-1}{2^{(t+1)n}}$.

4. From the above results, we can compute the upper bound of the advantage of the collision resistance of F as follows:

$$
\begin{aligned}
\mathsf{Adv}_F^{coll}(q) &= Max_A[\mathsf{Adv}_F^{coll}(A(q))] \\
&= Max_A[\mathsf{Pr}_A[C_2 \vee C_3 \cdots \vee C_q]] \\
&\leqslant Max_A[\mathsf{Pr}_A[C_2] + \textstyle\sum_{i=3}^{q} \mathsf{Pr}_A[C_i]] \\
&\leqslant Max_A[\frac{1}{2^{tn}} + \textstyle\sum_{i=3}^{q}(\frac{t(t+3)}{2^{tn+1}} + \frac{i-1}{2^{(t+1)n}})] \\
&= \frac{t(t+3)(q-1)}{2^{tn+1}} + \frac{q^2-1}{2^{(t+1)n}}. \ \blacksquare
\end{aligned}
$$

4 Compression Function with Variable Output Size in the Ideal Cipher Model

4.1 Limitation of Proofs in Previous Work Using Only a Single Block Cipher

There are several papers which proved the security of hash functions based on a block cipher. For example, Black *et al.* [3] proved the optimal collision resistance of 20 PGV schemes in the ideal cipher model. The security of MDC-2 [18] and Hirose's constructions [8, 9] were also proved in the ideal cipher model. All their proofs share commonalities; the upper bound of the number of queries is $q < 2^n$. This is because, in the case of a fixed key, the block cipher is a random permutation. For example, for a key a, we assume that we have query-response pairs (a, x_i, y_i) $(1 \leqslant i \leqslant t)$ such that the block size is an n-bit and $E_a(x_i) \oplus x_i = y_i$. Then, if we make a new encryption query (a', x_{t+1}) to the ideal cipher (where a' may not equal to a), we know that y_{t+1} is selected randomly from among at least $2^n - t$ candidates. That is, when $q < 2^n$, we can consider a block cipher-based function as a random function in the set of at least size $2^n - q$. This insight helps us to prove the security of hash functions based on the block cipher. The restriction $q < 2^n$ is significant in double block length hash functions, because with a high probability the adversary can find a collision with q (close to 2^n) queries. However, in the case of hash functions of at least three block output size ($3n$-bit), $q < 2^n$ is not sufficient. This is because we can only guarantee that the security of a hash function with a $3n$-bit output is at least a $2n$-bit security. In fact, the optimal security is a $3n$-bit security. In this paper, we give a solution to the problem of overcoming this barrier to prove the security of a hash function with at least triple block output size. In the case of the construction described in the Section 4.3, q is any value and collision resistance is optimal in the case of $t \geqslant 2$.

4.2 New DBL Compression Function based on a Block Cipher.

We consider the following function f based on a block cipher E:

$$f(X) = E_X(c) \text{ and } F(X) = f(X)||f(P(X)),$$

where X is an m-bit input and c is an n-bit constant, $m > 2n$ and F is Nandi's construction explained in Section 3.1. Then, based on Lemma 1, we can prove the near-optimal collision resistance of F in Theorem 3.

The goal of the collision finding adversary is to find X and X', where $F(X) = F(X')$ and $X \neq X'$. In the ideal cipher model, the adversary can make queries to both oracles E and E^{-1}. In our construction, plaintexts of meaningful query-response pairs should be c, because, in our construction the plaintext is always the fixed c as $f(X) = E_X(c)$. Therefore, we prove the following equality:

Lemma 1. *For any A who can have query-response pairs such that the plaintext is not c, there is B such that:*

$$\mathit{Adv}_F^{coll}(A(q)) = \mathit{Adv}_F^{coll}(B(q)),$$

where B is any adversary who can only make encryption queries for which the plaintext is always c.

Proof : Let A be a collision-finding adversary with access to both oracles E and E^{-1}. We can define an adversary B^E which only makes an encryption query with plaintext c.

Adversary $B = (B_1, B_2)$**.**

$B = (B_1, B_2)$ first runs A and communicates with A as follows:

- When the A's i-th query is an encryption query $(1, x, y)$ to B_1 where x is a key and y is a plaintext, B keeps z^* which is the response of the oracle E for the query $(1, x, c)$, and then does as the followings:
  - If $y = c$, B forwards z^* to A.
  - If $y \neq c$, B chooses an element z randomly from the set $\{0,1\}^n \setminus \{z^*\} \cup \{z' | (w', x, y', z') \in \mathcal{V}_A^{B_1,B_2}(i-1)\}$. Then, B forwards z to A.
- When the A's i-th query is an decryption query $(-1, x, z)$ to B_2 where x is a key and z is a ciphertext, B keeps z^* which is the response of the oracle E for the query $(1, x, c)$, and then does as the followings:
  - If $z = z^*$, B forwards c to A.
  - If $z \neq z^*$, B chooses an element y randomly from the set $\{0,1\}^n \setminus \{c\} \cup \{y' | (w', x, y', z') \in \mathcal{V}_A^{B_1,B_2}(i-1)\}$. Then, B forwards y to A.
- B's final output is that of A.

Regarding Adversary B, whenever A finds a collision, B can also obtain a collision. Moreover, B perfectly simulates the ideal block cipher for A. Thus, the following is true:

$$\mathsf{Adv}_F^{coll}(A(q)) = \mathsf{Adv}_F^{coll}(B(q)). \qquad \blacksquare$$

Based on Lemma 1, we need to compute the upper bound of $\mathsf{Adv}_F^{coll}(B(q))$ for any Adversary B defined in Lemma 1. Since the queries are different, in our construction the key of the block cipher should be different. Therefore, in the ideal cipher model, the response of the query is random. Therefore, we can prove the following theorem in a similar manner to that used in Section 3.

Theorem 3. *In the ideal cipher model, an upper bound of the maximum advantage from the perspective of the collision resistance of F is expressed as follows:*

$$\mathit{Adv}_F^{coll}(q) \leqslant \frac{q-1}{2^n} + \frac{q^2-1}{2^{2n+1}}.$$

We can also generalize the above result as the following Section 4.3.

4.3 Block Cipher based Generalized Construction I

We consider the following function f based on a block cipher E:

$$f(X) = E_X(c) \text{ and}$$
$$F(X) = f(P_0(X))||f(P_1(X))||\cdots||f(P_t(X)),$$

where X is an m-bit input and c is an n-bit constant, $m > (t+1)n$, and F is the general construction explained in Section 3.2. Then, for $t \geqslant 2$ we can prove the optimal security of F in the following theorem. In the case of 128-bit RC6 with a key size of 2,040 bits, $X = h_{i-1}||m_i$ is 2,040-bit such that h_{i-1} is 1,920-bit, m_i is 120-bit, and $t = 14$.

Theorem 4. *In the ideal cipher model, an upper bound of the maximum advantage from the perspective of the collision resistance of F is expressed as follows:*

$$\mathsf{Adv}_F^{coll}(q) \leqslant \frac{t(t+3)(q-1)}{2^{tn+1}} + \frac{q^2-1}{2^{(t+1)n+1}} \quad \textit{where } t \geqslant 2.$$

4.4 Block Cipher based Generalized Construction II

We consider the following function F based on a block cipher E:

$$F(X) = E_X(c_1)||E_X(c_2)||\cdots||E_X(c_t),$$

where X is an m-bit input, and c_i's (where for $i \neq j$, $c_i \neq c_j$) are n-bit constants, and $m > tn$. Then, for any positive integer t we can prove the near-optimal security of F in Theorem 5 using Lemma 2.

Lemma 2. *For any A who can have query-response pairs such that plaintext is not one among c_i's, there is B such that*

$$\mathsf{Adv}_F^{coll}(A(q)) = \mathsf{Adv}_F^{coll}(B(tq)),$$

where B is any adversary who can only make encryption queries for which the plaintext is always one among c_i's.

Proof : Let A be a collision-finding adversary with access to both oracles E and E^{-1}. We can define an adversary B^E, which only makes an encryption query such that the plaintext is one among c_i's (where $1 \leqslant i \leqslant t$).

Adversary $B = (B_1, B_2)$**.**

$B = (B_1, B_2)$ first runs A and communicates with A as follows:

- When the A's i-th query is an encryption query $(1, x, y)$ to B_1, where x is a key and y is a plaintext, for $1 \leqslant i \leqslant t$ B keeps z_i^* which is the response of the oracle E for the query $(1, x, c_i)$ and then does as the followings:

- If $y = c_i$ for a i, B forwards z_i^* to A.
- If $y \neq c_i$ for all i, B chooses an element z randomly from the set $\{0,1\}^n \setminus \{z_1^*, \cdots, z_t^*\} \cup \{z' | (w', x, y', z') \in \mathcal{V}_A^{B_1,B_2}(i-1)\}$. Then, B forwards z to A.

– When the A's i-th query is an decryption query $(-1, x, z)$ to B_2 where x is a key and z is a ciphertext, for $1 \leqslant i \leqslant t$ B keeps z_i^* which is the response of the oracle E for the query $(1, x, c_i)$ and then does as the followings:

- If $z = z_i^*$ for a i, B forwards c_i to A.
- If $z \neq z_i^*$ for all i, B randomly chooses an element y from the set $\{0,1\}^n \setminus \{c_1, \cdots, c_t\} \cup \{y' | (w', x, y', z') \in \mathcal{V}_A^{B_1,B_2}(i-1)\}$. Then, B forwards y to A.

– B's final output is that of A.

Regarding Adversary B, whenever A finds a collision, B can also obtain a collision. Moreover, B perfectly simulates the ideal block cipher for A. Thus, the following is true:

$$\mathsf{Adv}_F^{coll}(A(q)) = \mathsf{Adv}_F^{coll}(B(tq)). \qquad \blacksquare$$

Theorem 5. *In the ideal cipher model, an upper bound of the maximum advantage from the perspective of the collision resistance of F is expressed as follows:*

$$\mathsf{Adv}_F^{coll}(q) \leqslant \frac{t^2q^2 - tq}{2^{tn+1}} \quad \text{where } t \geqslant 1.$$

Proof : Since we showed in Lemma 2 that $\mathsf{Adv}_F^{coll}(A(q)) = \mathsf{Adv}_F^{coll}(B(tq))$, we have only to show that $\mathsf{Adv}_F^{coll}(B(tq)) \leqslant \frac{t^2q^2-tq}{2^{tn+1}}$, where $t \geqslant 1$ and B is any adversary B^E which only makes an encryption query with a plaintext c_j (where c_j is one among c_i's). We let $f_i(X) = E_X(c_i)$. So, $F(X) = f_1(X)||f_2(X)|| \cdots ||f_t(X)$. Since each f_i's output distribution is uniform and random and independent of other f_j's by the property of the ideal cipher, the probability that there is a collision of F is at most $\frac{t^2q^2-tq}{2^{tn+1}}$ by the birthday probability. $\blacksquare$

5 Conclusion

In this paper, we investigated the means to design compression functions with variable lengths from an atomic function of a fixed output length. Our results are significant because we can make hash functions with variable output sizes with only a single function. Recently, several constructions were suggested, where some independent and uniform random functions were used [12, 16, 17]. We hope that our results are applied to the reduction of the number of random functions required to guarantee the collision resistance of constructions proposed in [12, 16, 17].

Acknowledgement

We would like to thank Dr. Jongsung Kim and the anonymous referee for his valuable comments and for suggesting the second construction in Section 4.4. This research was supported by the MKE(Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Advancement)(IITA-2008-(C1090-0801-0025)).

References

1. G. Bertoni, J. Daemen, M. Peeters and G. V. Assche,*RadioGatún, a belt-and-mill hash function*, Cryptology ePrint Archive: Report 2006/369.
2. M. Bellare and P. Rogaway, *Random Oracles Are Practical : A Paradigm for Designing Efficient Protocols*, 1st Conference on Computing and Communications Security, ACM, pages 62–73. 1993.
3. J. Black, P. Rogaway and T. Shrimpton, *Black-box analysis of the block-cipher-based hash function constructions from PGV*, CRYPTO 2002, LNCS 2442, pp. 320-335. Springer-Verlag, 2002.
4. D. Chang, K. C. Gupta and M. Nandi, *RC4-Hash : A New Hash Function based on RC4*, INDOCRYPT 2006, LNCS 4329, pp. 80-94, Springer-Verlag, 2006.
5. I. B. Damgård, *A design principle for hash functions*, CRYPTO 1989, LNCS 435, pp. 416-427, Springer-Verlag, 1989.
6. FIPS 180-1, *Secure Hash Standard*, US Department of Commerce, Washington D. C, 1996.
7. FIPS 197, *Advanced Encryption Standard (AES)*, 2001.
8. S. Hirose, *Some Plausible Constructions of Double-Block-Length Hash Functions*, FSE 2006, LNCS 4047, pp. 210-225, Springer-Verlag, 2006.
9. S. Hirose, *How to Construct Double-Block-Length Hash Functions*, In second Hash Workshop, 2006.
10. R. C. Merkle, *One way hash functions and DES*, CRYPTO 1989, LNCS 435, pp. 428-446, Springer-Verlag, 1990.
11. M. Nandi, *Towards Optimal Double-Length Hash Functions*, INDOCRYPT 2005, LNCS 3797, pp. 77-89, Springer-Verlag, 2005.
12. T. Peyrin, H. Gilbert, F. Muller and M. Robshaw, *Combining Compression Functions and Block Cipher-Based Hash Functions*, ASIACRYPT 2006, LNCS 4284, pp. 315–331, Springer-Verlag, 2006.
13. Ronald L. Rivest, *The MD5 message-digest algorithm*, Request for comments (RFC 1321), Internet Activities Board, Internet Privacy Task Force, 1992.
14. R.L. Rivest, M.J.B. Robshaw, R. Sidney, Y.L. Yin, *The RC6 Block Cipher. v1.1*, AES Proposal, 1998., available via http://people.csail.mit.edu/rivest/Rc6.pdf.
15. B. Schneier, *Description of a New Variable-Length Key, 64-Bit Block Cipher (Blowfish)*, FSE 1994, LNCS 809, pp. 191-204, Springer-Verlag, 1994.
16. Y. Seurin and T. Peyrin, *Security Analysis of Constructions Combining FIL Random Oracles*, FSE 2007, LNCS 4593, pp. 119-136, Springer-Verlag, 2007.
17. T. Shrimpton and M. Stam, *Building a Collision-Resistant Compression Function from Non-Compressing Primitives*, Cryptology ePrint Archive: Report 2007/409.
18. J. P. Steinberger, *The Collision Intractability of MDC-2 in the Ideal-Cipher Model*, EUROCRYPT 2007, LNCS 4515, pp. 34-51, Springer-Verlag, 2007.

19. X. Wang, X. Lai, D. Feng, H. Chen and X. Yu, *Cryptanalysis of the Hash Functions MD4 and RIPEMD*, EUROCRYPT 2005, LNCS 3494, pp. 1-18, Springer-Verlag, 2005.
20. X. Wang and H. Yu, *How to Break MD5 and Other Hash Functions*, EUROCRYPT 2005, LNCS 3494, pp. 19-35, Springer-Verlag, 2005.
21. X. Wang, H. Yu and Y. L. Yin, *Efficient Collision Search Attacks on SHA-0*, CRYPTO 2005, LNCS 3621, pp. 1-16, Springer-Verlag, 2005.
22. X. Wang, Y. L. Yin and H. Yu, *Finding Collisions in the Full SHA-1*, CRYPTO 2005, LNCS 3621, pp. 17-36, Springer-Verlag, 2005.