Hash Function Design Principles Supporting Variable Output Lengths from One 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 hash function design principles with variable output lengths (multiple of n). It is based on a function or a block cipher which has output size n. In the random oracle model it has optimal collision resistance which requires $\Theta(2^{(t+1)n/2})$ queries to find $(t+1)n$-bit hash output collisions, where t is any positive integer. Similarly, in the ideal cipher model, $\Theta(2^{(t+1)n/2})$ queries are required to find $(t+1)n$-bit hash output collisions.

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

1 Introduction.

In 2004 and 2005, Wang *et al.* [17–20] introduced a new strategy to find a collision of widely used hash functions such as MD5 [13], SHA-1 [7] and so on. Since fatal weaknesses of MD5 and SHA-1 were revealed by Wang *et al.*, many cryptographers have recognized the necessity of new hash functions as their replacements. Upon this recognition, NIST announced to develop one or more additional hash algorithms through a public competition like AES [1]. 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 a provably secure design principle to support variable length output. As a method, for each output length we can design hash function independently like SHA-family. We can also design steam cipher-style hash functions such as RadioGatún [2] and RC4-Hash [5] which use a function repeatedly till we get the size of hash output we want. As another method, we can design variable output length-hash functions with a small output-length algorithm. Double block length (DBL) hash functions by Nandi [11] and Hirose [8, 9] are such a case. Nandi [11] proved that his construction has the 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 optimality of the collision resistance. However, since they considered only DBL hash functions, their constructions have a limitation that they can not support variable sizes of hash outputs.

In this paper, through the generalization of Nandi's result we show that we can handle arbitrary hash output with a function. Also we prove the optimal security of its collision resistance in the random oracle model. Furthermore, we propose a new block cipher based hash function with variable output length and we prove its optimal security from the viewpoint of collision resistance. Based on the result of this paper, with a 128-bit RC-6, a 64-bit Blowfish and SHA-256 we can design a hash function to handle maximum 1920-bit, 384-bit and 512-bit hash outputs, respectively.

2 Definitions and Known Results

In this section, we define the notations and symbols and describe the known results.

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 have access to the function f as a black-box manner. It is easy to see 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 $j < i$ we have

$$\mathsf{Pr}[f(x_i) = y_i | f(x_1) = y_1, f(x_2) = y_2, \cdots, f(x_{i-1}) = y_{i-1}] = 1/2^n$$

Note that x_i can be any random variable independent to the random oracle f. Let A^f be any adversary which has access of the random oracle f and suppose x_i is the i-th query and y_i is 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 condition probability distribution of i-th response is uniformly and independently distributed on $\{0,1\}^n$.

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. In other words, 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 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 and 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 part.

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 $\mathsf{bin}_i(x)$ means the i-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 to design a hash function from a compression function $F : \{0,1\}^d \rightarrow \{0,1\}^s$ with the padding rule g [6, 10]. They proved that if F is collision resistant then MD_g^F is also collision resistant. MD_g^F is defined as follows.

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

where $g(M) = (m_0||m_1||\cdots||m_t)$ and IV is the s-bit initial value and each m_i is $d-s$ bits.

Non-adaptive and Adaptive Models. When the adversary is permitted to make q queries to a given oracle, in the non-adaptive model he can ask only maximum q queries simultaneously and then he can get all responses. And in the adaptive model he can ask the i-th query after he gets $i-1$ query-responses. In this paper, we consider the adaptive model and it is the strongest model in security point of view. Moreover, adaptive adversary is a reasonable consideration as we consider public compression function or a public block cipher as a random oracle model. Adversary can compute the outputs of these adaptively.

We assume that the adversary A can make q queries at most. And we assume that he is deterministic and computationally unbounded. It is easy to prove that if a scheme is secure against 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 information he has. Here, since he is adaptive and deterministic, his i-th query is uniquely determined from the view $\mathcal{V}_A^{\mathcal{O}}(i-1)$. In other words,

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

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

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

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

Similarly, we can define the collision resistant measurement of compression function F against the adversary A in the ideal cipher model as follows.

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

The collision resistant measurement of MD_g^F is defined similarly as follows.

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

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 say that F (or MD_g^F) is collision resistant (or has collision resistance) if the maximum advantage is negligible. Especially, we say that F (or MD_g^F) is optimally collision resistant (or has optimal collision resistance) if $\Theta(2^{s/2})$ queries are required to make the maximum advantage '1' where the output length is a s-bit. We know the followings by [6, 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 showing that the upper bound of $\mathsf{Adv}_F^{coll}(q)$ is negligible.

Remark 1. We assume that the adversary does not make a same query repeatedly. In the random oracle model, all queries x_i's are different. In the ideal cipher model, once he gets (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 an encryption query (a, x)). Secondly, we assume that the adversary's final outputs which are expected to collide should be able to be constructed from his final view. As described in [16], if there is no second assumption, the adversary can output two very long messages (which is not related to his view) to collide with a high probability.

Remark 2. Our goal is to show the maximum advantage of our design principle is negligible in the random oracle model and ideal cipher model. According to the definition of the advantage from the viewpoint of the collision resistance, the advantage is from the probability that final outputs of the adversary collide. Recall that, according to the second assumption in the Remark 1, if the adversary find collisions, the collisions should be constructed from the final view. In other words, without considering final outputs of the adversary, we can directly get the upper bound of the advantage from the final view. So, we focus on the probability that there exists collisions 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 and P^2 is the identity permutation. Also he proved that F is 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 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. He proved that if f is applied to the Nandi's construction, MD_g^F is optimally collision resistant in the ideal cipher model. He also proposed five other constructions [9] and proved their optimal collision resistance. These six constructions belong to DBL hash functions.

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

$$\mathsf{Rate} = \frac{\text{size of message used in comp. func.}}{(\sharp\text{ of atomic function used in comp. func.}) \times (\text{output size of atomic func.})}$$

For example, the rate of Nandi's construction (the atomic function is f) is $\frac{m-2n}{2n}$. In the case of the Hirose's construction (the atomic function is E), the rate becomes 1/2.

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

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

3.1 Nandi's Construction and Its Security [11]

As mentioned before, Nandi proposed a DBL compression construction $F(X) = f(X)||f(P(X))$ from f of a small output size, where $f : \{0,1\}^m \rightarrow \{0,1\}^n$ $(m > 2n)$ and P is a permutation with no fixed point and P^2 is the identity permutation. He proved that F is optimally collision resistant in the random oracle model.

Theorem 1. *In the random oracle model, an upper bound of the maximum advantage from the viewpoint of the collision resistance of F is described 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 asks 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, at most q input-output pairs of F can be constructed. For an even q, there exists 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. In other words, 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 have to ask at least q queries X_i $(1 \leqslant i \leqslant q)$ to the random oracle f to get q input-output pairs of F. Therefore, at most q input-output pairs of F can be constructed from q input-output pairs of f. Next, we want 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 get $F(X) = f(X)||f(P(X))$, we need to ask two queries X and $P(X)$ to the random oracle f. Once we get $F(X) = f(X)||f(P(X))$, we can know $F(P(X)) = f(P(X))||f(X)$ without any additional queries. Therefore, we can get two input-output pairs of F from two input-output pairs of f. Likewise, when q is even, we can get 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 $\mathcal{V}_A^{\mathcal{O}}(q)$. According to the result of step 1, we write $F[\mathcal{V}_A^f(q)] = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_p, Y_p)\}$ where $p \leqslant q$ and $X_i \neq X_j$ for all i and j $(i \neq j)$. Here, we want to compute the probability that $F(X_i) = F(X_j)$ for any i and j $(i \neq j)$. The following holds for any i and j where $j < i \leqslant p$.

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

   (b) When $X_i \neq P(X_j)$: Since $\{X_i, X_j\} \cap \{P(X_i), P(X_j)\} = \varnothing$,
   $\mathsf{Pr}[F(X_i) = F(X_j)] = \mathsf{Pr}[f(X_i) = f(X_j) \wedge f(P(X_i)) = f(P(X_j))]$

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

3. Let the event C_i be the event that there exists j ($j < i$) 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 one time at most and the case of step 2-(b) occurs $i - 1$ times at most. 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 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 and then we prove its optimal collision resistance.

Generalized Construction. We want to construct F from a function f which has a 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 and P_i is a permutation with no fixed point and P_i^2 is the identity permutation. For any i and j ($i \neq j$), $P_iP_j = P_jP_i$. And 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 $P_i(x) = x \oplus (1000\cdots0)^{\lll i}$ where $(1000\cdots0)$ has all zero-bit except that the left most bit is one. 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 in the viewpoint of the collision resistance of F is described as follows:*

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

Proof. Its proof is similar to that of theorem 1. Here, A is any deterministic and computationally unbounded adversary. We assume that A asks 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, at most q input-output pairs of F can be constructed.

   Proof) We assume that q input-output pairs of F are given. In other words, 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 $F(X) = f(X)||f(P_1(X))|| \cdots ||f(P_t(X))$, we have to ask at least q queries X_i $(1 \leqslant i \leqslant q)$ to the random oracle to get q input-output pairs of F. Therefore, at most 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 $\mathcal{V}_A^{\mathcal{O}}(q)$. According to the result of step 1, we write $F[\mathcal{V}_A^f(q)] = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_p, Y_p)\}$ where $p \leqslant q$ and $X_i \neq X_j$ for all i and j $(i \neq j)$. Here, we want to compute the probability that $F(X_i) = F(X_j)$ for any i and j $(i \neq j)$. The following holds for any i and j where $j < i \leqslant p$.

   (a) When $X_i = P_u(X_j)$ (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. I.e., $r = 0$ and $r = u$ correspond to the same 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) When $X_i = P_v P_u(X_j)$ (for some v and u, $1 \leqslant v < u \leqslant t$): 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 same element. And by the relations among P_i's, $|T_u| = t$ and for any l, k $(l \neq k)$, $\{P_l P_v P_u(X_j), P_l(X_j)\} \cap \{P_k P_v P_u(X_j), P_k(X_j)\} = \varnothing$. Therefore, $\mathsf{Pr}[F(X_i) = F(X_j)] = (\frac{1}{2^n})^t = \frac{1}{2^{tn}}$.

   (c) When $X_i \neq P_v P_u(X_j)$ (for all v and u, $0 \leqslant v < u \leqslant t$): When $T = \{\{P_r(X_i), P_r(X_j)\}\}_{0 \leqslant r \leqslant t}$, 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 that there exists j $(j < i)$ 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 t cases at most. Step 2-(b) contains $\frac{t(t+1)}{2}$ cases at most. Step 2-(c) constains $i - 1$ cases at most. 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 collision resistance of F.

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

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

Limitation of Proofs in Known Results When only One Block Cipher is Used. There are several papers which proved the security of hash functions based on a block cipher. For example, Black *et al.* [4] proved the optimal collision resistance of 20 PGV schemes in the ideal cipher model. The securities of MDC-2 [16] and Hirose's constructions [8, 9] were also proved in the ideal cipher model [16]. In all their proofs, there is something in common; the upper bound of the number of queries :'$q < 2^n$'. This is because for 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 ask a new encryption query (a, x_{t+1}) to the ideal cipher, we know that y_{t+1} will be selected randomly from unknown set of size $2^n - t$. I.e., when $q < 2^n$, we can consider a block cipher-based function as a random function in the set of size $2^n - q$ at least. This trick helps us to prove the security of the hash functions based on the block cipher. The restriction $q < 2^n$ is meaningful in double block length hash functions, because with a high probability the adversary can find a collision with q (near to 2^n) queries. However, in the case of the hash functions of three block output size ($3n$-bit) at least, $q < 2^n$ is not enough. This is because we can guarantee only that the security of hash function with a $3n$-bit output is a $2n$-bit security at least. In fact, the optimal security should be a $3n$-bit security. So, how can we overcome this barrier to prove the security of hash function with triple block output size at least? In this paper, we give an answer. In our construction, q is any value.

New DBL Compression Function based on a Block Cipher. We consider the following function f based on a block cipher E,

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

where X is a m-bit and IV is an n-bit initial value, $m > 2n$ and F is the Nandi's construction explained in section 3.1. Then, based on Lemma 1, we can prove 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 attacker can ask queries to both oracles E and E^{-1}. In our construction, the query-response pair whose plaintext is not IV can not be used to construct X and X' where $F(X) = F(X')$ and $X \neq X'$, because in our construction the plaintext is always the fixed IV as $f(X) = E_X(IV)$. Therefore, we prove the following equality (Lemma 1).

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

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

where B is any adversary who can make only the encryption queries whose plaintext is always IV.

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

Adversary $B = (B_1, B_2)$**.**
$B = (B_1, B_2)$ first runs A and it responses A's query as follows.

- When the A's i-th query is the 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 $(1, x, IV)$ and then
  - If $y = IV$, B forwards z^* to A.
  - If $y \neq IV$, 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 the 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 $(1, x, IV)$ and then
  - If $z = z^*$, B forwards IV to A.
  - If $z \neq z^*$, B chooses an element y randomly from the set $\{0,1\}^n \setminus \{IV\} \cup \{y' | (w', x, y', z') \in \mathcal{V}_A^{B_1,B_2}(i-1)\}$. Then, B forwards y to A.
- B's final output is that of A.

In the adversary B, whenever A finds a collision, B can also get a collision. Moreover, B perfectly simulates 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 want 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, the key of the block cipher should be different in our construction. Therefore, the response of the query is random in the ideal cipher model. Therefore, we can prove the following theorem in the similar way used in section 3.

Theorem 3. *In the ideal cipher model, an upper bound of the maximum advantage in the viewpoint of the collision resistance of F is described as follows:*

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

We can also generalize the above result as follows.

Generalized Construction based on a Block Cipher. We consider the following function f based on a block cipher E.

$$f(X) = E_X(IV) \text{ and } F(X) = f(P_0(X))||f(P_1(X))||f(P_2(X))|| \cdots ||f(P_t(X)),$$

where X is a m-bit and IV is an n-bit initial value, $m > (t+1)n$ and F is the general construction explained in section 3.2. Then for $t \geqslant 2$ we can prove the following theorem.

Theorem 4. *In the ideal cipher model, an upper bound of the maximum advantage in the viewpoint of the collision resistance of F is described as follows:*

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

5 Conclusion

In this paper, we investigated how to design hash functions with variable lengths from an atomic function of a fixed output length. Our results are meaningful because we can make hash functions with variable output sizes with only one function. Recently, several constructions have been suggested where some independent and uniform random functions are used [12, 14, 15]. We hope that our results can be applied to reduce the number of random functions required to guarantee the optimal collision resistance of constructions in [12, 14, 15].

Acknowledgement

We would like to thank Dr. Jongsung Kim for his helpful comments.

References

1. Federal Information Processing Standards Publication 197, *ADVANCED ENCRYPTION STANDARD (AES)*, 2001.
2. G. Bertoni, J. Daemen, M. Peeters and G. V. Assche, *RadioGatún, a belt-and-mill hash function*, Cryptology ePrint Archive: Report 2006/369.

3. M. Bellare and P. Rogaway, *Random Oracles Are Practical : A Paradigm for Designing Efficient Protocols*, 1st Conference on Computing and Communications Security, ACM, pp. 62–73. 1993.
4. J. Black, P. Rogaway and T. Shrimpton, *Black-box analysis of the block-cipher-based hash function constructions from PGV*, Advances in Cryptology - Crypto'02, LNCS 2442, pp. 320–335, Springer-Verlag, 2002.
5. D. Chang, K. C. Gupta and M. Nandi, *RC4-Hash : A New Hash Function based on RC4*, Indocrypt'06, LNCS 4329, pp. 80–94, Springer-Verlag, 2006.
6. I. B. Damgard, *A design principle for hash functions.* Advances in Cryptology - Crypto'89, LNCS 435, pp. 416–427, Springer-Verlag, 1989.
7. FIPS 180-1, Secure Hash Standard, US Department of Commerce, Washington D. C, 1996.
8. S. Hirose, *Some Plausible Constructions of Double-Block-Length Hash Functions*, FSE'06, LNCS 4047, pp. 210–225, Springer-Verlag, 2007.
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*, Advances in Cryptology-Crypto'89, LNCS 435, pp. 428–446. Springer-Verlag, 1990.
11. M. Nandi, *Towards Optimal Double-Length Hash Functions*, Indocrypt'05, 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'06, LNCS 4284, pp. 315–331, Springer-Verlag, 2006.
13. Ronald L. Rivest, *The MD5 message-digest algorithm*, Request for comments (RFC 1320), Internet Activities Board, Internet Privacy Task Force, 1992.
14. Y. Seurin and T. Peyrin, *Security Analysis of Constructions Combining FIL Random Oracles*, FSE'07, LNCS 4593, pp. 119–136, Springer-Verlag, 2007.
15. T. Shrimpton and M. Stam, *Building a Collision-Resistant Compression Function from Non-Compressing Primitives*, Cryptology ePrint Archive: Report 2007/409.
16. J. P. Steinberger, *The Collision Intractability of MDC-2 in the Ideal-Cipher Model*, Advances in Cryptology-Eurocrypt'07, LNCS 4515, pp. 34–51, Springer-Verlag, 2007.
17. X. Wang, X. Lai, D. Feng, H. Chen and X. Yu, *Cryptanalysis of the Hash Functions MD4 and RIPEMD*, Advances in Cryptology-Eurocrypt'05, LNCS 3494, pp. 1–18, Springer-Verlag, 2005.
18. X. Wang and H. Yu, *How to Break MD5 and Other Hash Functions*, Advances in Cryptology-Eurocrypt'05, LNCS 3494, pp. 19–35, Springer-Verlag, 2005.
19. X. Wang, H. Yu and Y. L. Yin, *Efficient Collision Search Attacks on SHA-0*, Advances in Cryptology-Crypto'05, LNCS 3621, pp. 1–16, Springer-Verlag, 2005.
20. X. Wang, Y. L. Yin and H. Yu, *Finding Collisions in the Full SHA-1*, Advances in Cryptology-Crypto'05, LNCS 3621, pp. 17–36, Springer-Verlag, 2005.