MD5 해쉬함수의 부울함수 특성 고찰 및 개선

°이 원준, 이 국희, 문 상재

경북대학교 전자공학과

The Study and Improvement of Boolean Function Property of MD5 Hash Function

°Won-Jun Lee, Kook-Heui Lee, Sang-Jae Moon

Dept. of Electronics, Kyungpook National University

요 약 문

일방성 해쉬함수는 임의 길이의 입력메세지를 일정한 길이의 출력메세지로 축약하는 함수로서, 디지탈서명에서 서명을 생성하는 시간을 단축하고 메세지 인증을 위한 암호학적 도구로 사용되고 있다. 본 논문에서는 부울함수를 기초로 하는 해쉬함수 중에서 MD5의 부울함수를 정보이론의 관점에서 분석하여 암호학적으로 강하고 기존의 MD5에서 사용된 부울함수의 성질을 이용한 공격을 막을 수 있는 새로운 부울함수를 제안한다.

1. 서 론

일방성 해쉬함수는 1976년 Diffie와 Hellman의 일방성함수와 일방 trapdoor 함수의 개념을 바탕으로 Rabin과 Merkle에 의하여 시작되었다. 일방성 해쉬함수는 임의 길이의 메세지를 일정한 길이의 메세지로 축약하는 함수로서 컴퓨터 시스템에서 사용자 및 데이타의 인증과 데이타 무결성을 보장하기 위한 암호학적 도구로 사용되고 있다.

해쉬함수는 계산효율이 좋고 일방향성(one-wayness)과 충돌회피성(collision freeness)을 만족하여야 한다[1]. 일방향성이란 출력메세지로부터 입력메세지를 찾는 것이 계산상 불가능함을 가리키며 충돌회피성이란 동일한 출력메세지를 생성하는 다른 입력메세지를 찾는 것이 계산상 불가능함을 가리킨다.

지금까지 제안된 해쉬함수는 블럭암호, 모듈라연산, 부울함수 등에 기초한 해쉬함수로 분류할 수 있다[2]. 블럭암호에 기초한 해쉬함수는 기존의 DES(data encryption standard)와 같은 블럭함수를 이용할 수 있는 장점이 있고 모듈라연산에 기초한 해쉬함수는 RSA와 같이 디지탈서명 함수 그 자체가 모듈라 연산에 사용될 경우 더 적합할 수 있다. 그러나 이들 해쉬함수는 연산속도가 떨어지므로 계산효율이 좋은 부울함수에 기초한 해쉬함수가 더 실용적일 수 있다.

본 논문에서는 부울함수에 기초한 해쉬함수 중에서 계산효율이 좋으며 R.Rivest에 의해 제안된 MD5(message digest 5)[3]의 부울함수의 암호학적 강도와 부울함수 특성을 이용한 기존의 공격을 분석하고 이들 성질을 개선한 새로운 부울함수를 제안한다.

2. MD5 알고리듬

MD5 알고리듬은 임의 길이의 입력메세지를 512비트 블럭으로 나누어 반복연산한 후 128비트

의 메세지로 축약하는 함수로서 디지탈서명에 적용하기 위하여 개발되었다. MD5 알고리듬은 32비트 단위로 처리되며 32비트를 워드라고 한다.

먼저 임의 길이의 입력메세지를 길이가 488 mod 512가 되도록하기 위하여 메세지 다음의 한 비트를 1로 한 후 그 뒤의 비트를 0으로 패딩(padding)시킨다. 그리고 메세지가 패딩되기 전의 길이를 64비트 값으로 나타내어 패딩된 메세지 다음에 더한다. 4개의 워드 버퍼인 MD가 메세지 축약을 위하여 사용되며 이것은 상수로 초기화되어 있다. 패딩된 입력메세지는 16-워드 블럭 M_1, M_2, …, M_{N-1}, M_N으로 나누어지며 각 블럭은 4 라운드(round)를 수행하며 다시 각 라운드는 16단계의 연산을 수행한다. 모든 블럭을 반복연산한 다음의 4-워드 버퍼가 MD5의 출력이 된다.

4-워드 버퍼인 MD의 초기값은 MD_0이며 각 블럭의 과정동안 갱신되어진다. j번째 블럭 처리과정은 4 라운드 함수 *FF, GG, HH, II*를 포함하며 다음 식과 같이 나타낼 수 있다.

$$MD_j = MD_{j-1} + II(M_j, HH(M_j, GG(M_j, FF(M_j, MD_{j-1}))))\quad (1)$$

4-워드 버퍼는 4개의 워드로 구성된 shift 레지스터로서 A,B,C,D로 표시할 수 있다. 각 라운드는 이 버퍼에 대한 16단계로 구성되어지며 각 단계는 다음 식과 같이 나타낼 수 있다.

$$A = B + ((A + f(B,C,D) + x[k] + t) <<< s)\quad (2)$$

여기서 +는 mod 2^{32} 덧셈을 표시하고 f는 라운드마다 달라지는 부울함수이다. 또한 $x[k]$는 M_j가 16개의 워드로 나누어질때 k번째의 워드이며 k,t와 s는 각 단계의 파라미터이고 $<<<s$ 는 한 워드에 대해 s비트만큼의 왼쪽 순환 shift를 표시한다. 각 단계에 사용되는 부울함수는 입력으로 3개의 32비트 워드를 가지고 하나의 32비트 워드를 출력한다. 이때 출력워드의 각 비트는 입력워드의 대응되는 비트에만 의존한다. MD5 부울함수는 표 1에 나타난다.

3. MD5 부울함수의 암호학적 강도 분석

차동 암호분석(differential cryptanalysis), 선형 암호분석(linear cryptanalysis) 등과 같은 통계적 암호분석(statistic cryptanalysis)의 발달에 의해 암호시스템에 대한 공격이 성공할 가능성이 커지고 있다. 부울함수에 기초한 암호시스템에서는 부울함수가 암호학적으로 강해야함이 필수적이다. 암호학적으로 강한 부울함수의 측정기준으로 0-1 balance, 비선형도(nonlinearity), completeness, strict avalanche criterion(SAC)[4], correlation immunity[5], propagation criterion[6], 고차의 SAC[7] 등이 있다. 먼저 위의 측정기준을 살펴본다.

GF(2)의 원소로 구성된 벡터공간을 V_n으로 표시할때 V_n에서 GF(2)로 가는 함수를 부울함수라 한다. V_n상의 부울함수는 n개의 독립변수 $X_1, X_2, \cdots, X_{n-1}, X_n$을 사용한 다항식 $f(X_1, X_2, \cdots, X_{n-1}, X_n)$으로 표현할 수 있다. $X_1, X_2, \cdots, X_{n-1}, X_n$이 0,0,…,0에서 1,1,…,1까지 변할 때 함수 f의 2^n개 출력 비트를 연쇄하여 이를 f의 시퀀스(sequence)라 한다. $a_i \in GF(2)$이고 $f(X_1, X_2, \cdots X_{n-1}, X_n) = a_n X_n \oplus a_{n-1} X_{n-1} \oplus \cdots \oplus a_1 X_1 \oplus a_0$의 형태를 가질 때 함수 f를 아핀(affine)함수라 하고, 특히 $a_0 = 0$ 이면 선형함수라 한다.

V_n상의 함수 f의 시퀀스에서 0과 1의 개수가 2^{n-1}이면 0-1 balance라고 한다. 이것을 만족하면 엔트로피(entropy) $H(f(X_1, X_2, \cdots X_{n-1}, X_n))$가 최대이므로 입력에 대한 출력의 결과를 예측하기가 어려워진다. g를 또 다른 부울함수라 하면 함수 f와 g 사이의 거리는 두 함수 시퀀스간의 해밍거리(Hamming distance)를 의미한다. 그리고 함수의 비선형도는 f와 V_n에서 GF(2)로 가는 모든 아핀함수 사이의 최소거리를 의미한다. 비선형도는 좋은 암호설계를 위하여 중요한 평가기준이다. 비선형도가 높을수록 암호시스템은 선형함수 집합을 사용하여 공격하는 것이 어려워진다. $V_{2k+1}(k \geq 1)$상의 부울함수는 최대 $2^{2k} - 2^k$, $V_{2k}(k \geq 2)$상의 부울함수는 최대 $2^{2k-1} - 2^{k-1}$의 비선형도를 갖는다[4].

만약 암호적 변환이 completeness를 만족하면 암호문(cipher text)의 각 비트들은 평문(plain text)의 모든 비트들에 의존한다. 즉, 암호문의 각 비트를 표현하는 가장 간단한 부울함수들은 입력의 모든 비트들을 포함하여야 한다. 어떤 암호적 변환에서 입력 중 1비트만 변했을 때 출력 비트들의 1/2 이 변할 경우 암호적 변환은 애벌런치 효과(avalanche effect)를 나타낸다고 한다.

V_n상의 부울함수 f가 모든 가능한 입력벡터에 대하여 입력 비트 중 x_i ($1 \le i \le n$)만 바뀌었을 때 출력의 1/2이 바뀌면 함수는 SAC을 만족한다고 한다. SAC는 위에서 정의한 completeness와 애벌런치 효과의 개념을 결합한 것이다. 이 조건을 만족하는 함수 f에 대한 V_{n-1}상의 가장 좋은 근사함수 f'는 f의 출력과 다를 확률이 1/4 이기 때문에 함수 f'는 함수 f의 비교적 나쁜 근사함수이다. SAC를 만족하는 V_n상의 함수 f의 입력 중 임의의 k개 비트를 1 또는 0로 고정시켰을 때도 함수 f가 SAC를 만족한다면 함수 f는 k차 SAC를 만족한다고 한다. 한편 V_n상의 함수 f의 입력 중 임의의 m개($0 < m \le k$)비트를 보수화하였을때 함수 출력이 1/2의 확률로 변한다면 그 함수는 k차 propagation criterion(PC-k)를 만족한다고 한다. V_n상의 독립변수 $X_1, X_2, \cdots X_{n-1}, X_n$로부터 k개 변수를 선택하여 얻어지는 k-tuple이 통계적으로 함수 f의 출력에 독립일때 함수 f는 k차 correlation-immunity를 만족한다고 한다.

본 절에서는 부울함수의 암호학적 강도를 정보누설량과 관련지어 분석하고[8] MD5 부울함수의 보수 속성을 이용한 공격을 살펴본다[9].

3.1. 부울함수의 정보누설량

함수의 입력에 대한 완전한 정보 없이는 출력과 출력의 변화에 대한 예측(predictability)이 어렵다면 그 함수는 통계적 분석에 저항한다고 할 수 있다. 통계적 분석의 목적은 정적 분석(static analysis)과 동적 분석(dynamic analysis)에서 입력과 출력간의 통계적 관련성(statistical relationship)을 결정하는 것으로 정적 분석은 입력과 출력사이의 통계적 관련을 결정하는 것을 가리키고, 동적 분석은 입력 변화와 출력 변화사이의 통계적 관련성을 결정하는 것을 나타낸다.

입력의 부분적인 정보에 의하여 부울함수 출력의 불확실성(uncertainty)이 감소하는 양을 부울함수의 정보누설량이라 한다. 부울함수의 정보누설량은 크게 정적 정보누설량(static information leakage, SL)과 동적 정보누설량(dynamic information leakage, DL)으로 나누어진다. 정적 정보누설량은 입력이 변화하지 않을 때 알고 있는 입력의 부분적인 정보에 의해 부울함수 출력의 불확실성이 감소하는 양이고 동적 정보누설량은 입력의 변화에 대한 부분적인 정보에 의해 부울함수의 출력 변화의 불확실성이 감소하는 양이다. 정보이론의 관점에서 예측의 어려운 정도(unpredictability)은 출력 변수 또는 출력 변수 변화의 조건부 엔트로피(conditional entropy)에 의해 구해진다.

n개의 변수를 가지는 부울함수를 $Y = f(\mathbf{X})$로 나타낼때 출력 변화는 $\Delta Y = f(\mathbf{X} \oplus \Delta \mathbf{X}) \oplus f(\mathbf{X})$로 정의된다. 여기서 $\mathbf{X}=(X_1, X_2, \cdots, X_n)$이고 $\Delta \mathbf{X}=(\Delta X_1, \Delta X_2, \cdots, \Delta X_n)$이다. 또한 k ($0<k \le n$)개의 입력 변수는 $\mathbf{X_k}=(X_{i1}, X_{i2}, \cdots, X_{ik})$로 그리고 입력 변수 변화는 $\Delta \mathbf{X_k}=(\Delta X_{i1}, \Delta X_{i2}, \cdots, \Delta X_{ik})$로 표시된다. $\mathbf{x_k}$와 $\Delta \mathbf{x_k}$는 각각 $\mathbf{X_k}$, $\Delta \mathbf{X_k}$의 값이다.

입력 벡터 $\mathbf{x_k}$가 주어질때 부울함수 출력 Y의 정적 정보누설량은 다음 식과 같이 정의된다.

$$SL(Y;\ \mathbf{X_k} \mid \mathbf{x_k}) = 1 - H(Y \mid \mathbf{X_k}=\mathbf{x_k}) \quad (3)$$

유사하게, 입력 변화 벡터 $\Delta \mathbf{x_k}$가 주어질때 출력 Y의 동적 정보누설량은 다음 식과 같이 정의된다.

$$DL(\Delta Y;\ \Delta \mathbf{X_k} \mid \Delta \mathbf{x_k}) = 1 - H(\Delta Y \mid \Delta \mathbf{X_k}=\Delta \mathbf{x_k}) \quad (4)$$

각 입력 변수 $X_1, X_2, \cdots, X_n$과 입력 변수의 변화량 $\Delta X_1, \Delta X_2, \cdots, \Delta X_n$은 서로 독립적인 확률변수이기에 $\mathbf{X_k}$ 또는 $\Delta \mathbf{X_k}$와 출력사이의 정보누설량은 다음 식과 같다.

$$SL(Y;\ \mathbf{X}_k) = 2^{-k} \sum_{x_k \in Z_2^k} SL(Y;\ \mathbf{X}_k \mid \mathbf{x}_k) \tag{5}$$

$$DL(\Delta Y;\ \Delta \mathbf{X}_k) = 2^{-k} \sum_{x_k \in Z_2^k} DL(\Delta Y;\ \Delta \mathbf{X}_k \mid \Delta \mathbf{x}_k) \tag{6}$$

단, k=n인 경우는 $SL(Y;\ \mathbf{X}_k \mid \mathbf{x}_k)$는 $SL(Y;\ \mathbf{X} \mid \mathbf{x})$되고 이 값은 항상 1이 된다. 즉, $SL(Y;\ \mathbf{X})=1$ 로서 입력이 완전히 알려지기 때문에 출력에서 어떠한 불확실성도 없다.

입력 변수의 합 $\Sigma_k\mathbf{X}=\beta$가 주어질때 Y의 정적 정보누설량은 다음과 같이 정의된다.

$$SL(Y;\ \Sigma_k\mathbf{X} \mid \beta) = 1 - H(Y \mid \Sigma_k\mathbf{X}=\beta) \tag{7}$$

여기서 $\Sigma_k\mathbf{X} = X_{i1} \oplus X_{i2} \oplus \cdots \oplus X_{ik}$ 이다.

3.2. 부울함수의 정적 정보누설량 고찰

입력의 부분적인 정보와 출력사이에 중대한(significant) 정적 정보누설량이 없다면 그 함수는 정적 암호분석에 저항한다고 할 수 있다. 이것을 보이기 위하여 정적 정보누설량의 특성을 알아본다.

$\mathbf{X}_k$가 $\mathbf{x}_k$로 고정되어 있을때 $f(\mathbf{X})$의 부함수를 $Y' = f(\mathbf{X}_{n-k} \mid \mathbf{x}_k)$로 나타낼 수 있다. 여기서 $\mathbf{X}_{n-k}$는 $\mathbf{X}$에서 $\mathbf{X}_k$를 제외한 변수의 집합이다. 고정된 $\mathbf{X}_k$에 의해 $f(\mathbf{X})$로부터 얻어지는 모든 부함수가 0-1 balance이면 $SL(Y;\ \mathbf{X}_k) = 0$ 이다. 즉, 출력과 주어진 입력사이의 정적 정보누설량은 부함수의 imbalance에 의해 일어난다.

SL(Y)=0 이면 함수 $f(\mathbf{X})$는 0-1 balance이다. 즉 자신 정보누설량(self information leakage)은 입력에 대한 어떠한 정보도 주어지지 않을때 통계적 분석의 취약성을 측정한다.

입력의 모든 변수들이 함수의 출력 Y에 영향을 줄때 completeness를 만족한다고 한다. 랜덤 입력벡터 $\mathbf{X}$중에서 임의의 변수 X_{in}이 보수(complement)가 될때 출력 Y가 p_{i_n}의 확률로 변한다고 하면 다음 식이 성립한다.

$$SL(Y;\ \mathbf{X}_{n-1}) = 1 - p_{i_n} \qquad (1 \le i_n \le n) \tag{8}$$

여기서 $\mathbf{X}_{n-1}$는 $\mathbf{X}$에서 X_{in}을 제거시킬때 얻어지는 변수의 집합이다. 이때 모든 X_{in}에 대하여 $p_{i_n}>0$인 경우 즉, 출력과 모든 n-1 입력 변수들과의 정적 정보누설량이 1보다 작다면 completeness를 만족한다. 이 성질에 의하여 n차원의 함수가 n-1차원의 함수로 정확히 근사화될 수 없다. 만약 $p_{i_n} = 1/2$ 일때 즉, X_{in}이 보수가 될때 출력 Y가 1/2의 확률로 변한다면 이 함수는 SAC를 만족한다.

k차 correlation-immunity는 모든 $\mathbf{X}_k$에 대하여 $I(Y;\ \mathbf{X}_k)=0$인 것과 동치이다. 함수가 k차 correlation-immunity을 만족하면 정적 정보누설량 $SL(Y;\ \mathbf{X}_k)$를 Y의 자신 정적 정보누설량(self static information leakage)으로 줄일 수 있다.

만약 $f(\mathbf{X})$가 0-1 balance이고 k차 correlation-immunity를 만족하면 k개의 임의의 입력 비트가 알려져있고 나머지 n-k개 비트를 임의로 선택하였을때 모든 출력이 발생할 확률이 동일하다. 이때 이 함수를 k-resilient 함수라고 한다. k-resilient 함수는 k차의 정적 정보누설량을 0으로 한다. 그러나 고차의 정적 정보누설량이 0인 경우는 입력 변수의 합 $\Sigma_k\mathbf{X}=\beta$와 출력 Y사이의 높은 정적 정보누설량을 야기시키기 때문에 낮은 차수의 정적 정보누설량이 0인 함수보다 반드시 통계적 암호분석에 더 강한 것은 아니다.

함수 $f(\mathbf{X})$와 선형함수 $b(\mathbf{X}_k)=\Sigma_k\mathbf{X}$와 보수 $\overline{b}(\mathbf{X}_k)$사이에 적당한 거리가 없다면 입력 변수의 합 $\Sigma_k\mathbf{X}=\beta$와 출력 Y사이의 정적 정보누설량을 제한하는 것은 불가능하다. 두 부울함수 간의 거리는 두 함수의 시퀀스간의 해밍거리를 의미한다. 모든 경우에서의 입력 변수의 합 $\Sigma_k\mathbf{X}=\beta$와 출력 Y사이의 정적 정보

누설량을 제한하기 위하여 함수 $f(X)$와 모든 아핀함수간에 정당한 거리를 가지는 것이 필요하다. 그러므로 높은 비선형도는 입력 변수의 합 $\Sigma_k X=\beta$ 와 출력 Y사이의 높은 정적 정보누설량을 피하기위한 필요조건이다.

3.3. 부울함수의 동적 정보누설량 고찰

동적 정보누설량은 차동 암호분석에 대한 함수의 취약성을 측정할 수 있는 척도이다. 먼저 자기상관함수(autocorrelation)를 이용하여 동적 정보누설량의 원인을 살펴본다.

$(-1)^{f(X)}$ 의 자기상관함수는 다음 식과 같이 정의된다.

$$\pi(\Delta X)=2^{-n}\sum_{x\in Z_2^n}(-1)^{f(x\oplus\Delta X)\oplus f(x)} \tag{9}$$

간략히 $\pi(\Delta X)$를 $f(X)$의 자기상관함수라 하고 $\pi(\Delta x)$를 $f(X)$의 자기상관계수라 한다.

ΔX_k가 Δx_k로 고정될 때 얻어지는 $\pi(\Delta X_{n-k} \mid \Delta x_k)$를 $\pi(\Delta X)$의 부함수라 하면 ΔX_{n-k}의 모든 가능한 값을 부함수에 적용시켜 $f(X)$의 2^{n-k}자기상관계수를 얻는다. 이때 $\hat{\pi}(\Delta x_k)$를 2^{n-k}자기상관계수에 대한 평균 자기상관계수의 절대치라 하며 다음 식과 같다.

$$\hat{\pi}(\Delta x_k) = 2^{-(n-k)} \mid \sum_{\Delta x_{n-k}} \pi(\Delta X_{n-k} \mid \Delta x_k) \mid \tag{10}$$

$\hat{\pi}(\Delta x_k)$을 이용하여 Δx_k가 주어질때 Y의 동적 정보누설량은 다음 식으로 표현될 수 있다.

$$DL(\Delta Y;\ \Delta X_k \mid \Delta x_k) = 1-h\left(\frac{1}{2}-\frac{\hat{\pi}(\Delta x_k)}{2}\right) \tag{11}$$

여기서 $h(t) = -t\log_2 t - (1-t)\log_2(1-t)$인 이진 엔트로피 함수이다.

$\hat{\pi}(\Delta x_k) = 0$ 일때 $DL(\Delta Y;\ \Delta X_k \mid \Delta x_k) = 0$이다. 그리고 $\hat{\pi}(\Delta x_k)$가 증가할수록 $DL(\Delta Y;\ \Delta X_k \mid \Delta x_k)$도 증가한다. 그러므로 Y와 어떤 $k(1\le k<n)$ 입력 변수들 사이의 동적 정보누설량은 $f(X)$의 부함수 $f(X_{n-k} \mid x_k)$의 평균 자기상관계수에 의해 발생한다. 그리고 Y와 X사이의 동적 정보누설량은 $f(X)$의 자기상관에 의해 발생한다.

PC-k는 해밍거리가 k이하인 모든 $\Delta x(0<wt(\Delta x)\le k)$에 대해 $DL(\Delta Y;\ \Delta X \mid \Delta x) = 0$ 으로 표현될 수 있다. SAC는 PC-1과 동치이다. 정보누설량의 관점에서 PC-k를 만족하는 함수는 특정한 입력 변화 $(0<wt(\Delta x)\le k)$에서 동적 정보누설량을 0으로 하므로 다른 입력의 변화량에 대한 동적 정보누설량이 증가되어 차동 분석에 취약할 수 있다. 입력의 차원 n이 홀수인 함수에서 PC의 최대 차수는 n-1이 되며, PC-(n-1)을 만족하는 모든 함수에서 입력이 모두 변할때 동적 정보누설량은 1이 된다. 즉 $DL(\Delta Y;\ \Delta X \mid \underline{1}) = 1$이 되어 차동 분석에 취약하다. 그러므로 암호학적으로 강한 함수는 propagation criterion이 필수적이지 않다. 여기서 $\underline{1}$는 모든 성분이 1인 벡터이다.

k개의 입력변수를 고정할때 얻어지는 모든 $f(X)$의 부함수 $Y' = f(X_{n-k} \mid x_k)$가 SAC를 만족한다면 $f(X)$는 k차 SAC를 만족한다고 한다. 이 성질은 해밍거리가 1인 모든 ΔX_{n-k}에 대하여 $DL(\Delta Y';\ \Delta X_{n-k} \mid \Delta x_{n-k}) = 0$으로 표현될 수 있다. 그러나 고차의 SAC는 지나치게 함수를 제한하고 차동 분석에 취약할 수 있다. 그러므로 암호학적으로 강한 함수는 고차의 SAC가 필수적이지 않다.

이상의 논의에서 어떤 특정한 경우에서의 정보누설량을 0으로 하는 것은 다른 경우의 정보누설량을

증가시키는 결과를 초래할 수 있으므로 바람직하지 않다. 그러므로 정보누설량의 관점에서 암호학적으로 강한 부울함수는 전체적으로 정보누설량이 크지 않은 함수임을 알 수 있다.

3.4. MD5 부울함수의 보수속성을 이용한 공격

MD5에서 각 단계는 이전 단계의 결과를 더하는데 이것이 MD5의 압축함수(compression function)에 대한 충돌[9]을 찾을 수 있게 한다. 여기서 압축함수이란 4-워드 버퍼 (A,B,C,D)와 16-워드 메세지 블럭 $(x[0], x[1],\cdots,x[15])$을 입력으로 하여 4-워드 출력 (AA,BB,CC,DD)를 생성하는 함수로서 Φ로 표시하며 다음 식과 같다.

$$(AA,BB,CC,DD)=\Phi((A,B,C,D),(x[0],x[1],\cdots,x[15])) \tag{12}$$

충돌을 찾기위한 개념은 4-워드 버퍼 (A,B,C,D) 각각의 최상위비트를 보수화하여도 압축함수 Φ의 출력에 영향을 주지 않는 Φ의 입력을 찾는 것이다. 즉, 다음 식을 만족하는 (A,B,C,D)와 $(x[0],x[1],\cdots,x[15])$를 찾는 것이다.

$$\Phi((A,B,C,D)\oplus(2^{31},2^{31},2^{31},2^{31}),(x[0],x[1],\cdots,x[15]))=\Phi((A,B,C,D),(x[0],x[1],\cdots,x[15])) \tag{13}$$

압축함수 Φ는 4번째 라운드 수행 후의 4-워드 A,B,C,D에 첫번째 라운드 시작에서 가졌던 4-워드를 더하므로 다음 식과 같이 나타난다.

$$\Phi(A,B,C,D)=\Psi(A,B,C,D)+(A,B,C,D) \tag{14}$$

여기서 Ψ는 4개의 16-단계 라운드로 구성된다.

식 (13)의 Φ를 식 (14)로 대체하면 식 (13)은 다음 식과 같이 표현될 수 있다.

$$\Psi((A,B,C,D)\oplus(2^{31},2^{31},2^{31},2^{31}),(x[0],x[1],\cdots,x[15]))= \Psi((A,B,C,D),(x[0],x[1],\cdots,x[15]))\oplus(2^{31},2^{31},2^{31},2^{31}) \tag{15}$$

만약 식 (15)의 B,C,D 각각의 최상위비트가 보수화될 때 $f(B,C,D)$의 최상위비트가 보수화된다면 식 (2)의 좌변에서 갱신되는 A의 최상위비트도 보수화될 것이다. 이때 압축함수의 20-워드 입력이 라운드함수 f의 모든 단계에서 다음 식 (16)을 만족한다면 압축함수의 20-워드 입력과 그것의 처음 4-워드의 최상위비트가 보수화된 20-워드 입력은 같은 출력값을 가질 것이다.

$$f(\bar{x},\bar{y},\bar{z})=\overline{f(x,y,z)} \tag{16}$$

MD5 부울함수 F, G, H, I에서 위의 식을 만족하는 입력 (x, y, z)는 다음과 같다.

- $F(x, y, z) = xy \vee (\sim x)z$: (0, 0, 0), (1, 0, 0)과 그의 보수 (1, 1, 1)과 (0, 1, 1)
- $G(x, y, z) = xz \vee y(\sim z)$: (0, 0, 0), (0, 0, 1)과 그의 보수 (1, 1, 1)과 (1, 1, 0)
- $H(x, y, z) = x \oplus y \oplus z$: 모든 입력
- $I(x, y, z) = y \oplus x(\sim z)$: (0, 0, 0), (0, 1, 0)과 그의 보수 (1, 1, 1)과 (1, 0, 1)

$\vee$: bitwise OR $\sim$: bitwise complement
$\oplus$: bitwise XOR xy : bitwise AND of x and y

라운드 1의 각 단계에서 A, B, C, D중에 하나만이 갱신되므로 어떤 단계에서 (1, 0, 0)이 입력되면 다음 단계에서는 $(x, 1, 0)$이 입력될 것이다. 여기서 x는 0또는 1이다. 이때 $F(\bar{x},0,1)$과 $\overline{F(x,1,0)}$은 같지 않으므로 (1, 0, 0)은 라운드 1에서 충돌을 찾기 위한 함수 F의 입력으로 사용될 수 없다. 이것은

(1, 0, 0)의 보수 (0, 1, 1)에도 적용된다. 마찬가지로 라운드 2의 함수 G에서 (0, 0, 1), (1, 1, 0)도 충돌을 찾기 위한 입력으로 사용될 수 없다. 그러므로 충돌을 찾기 위해서 (1, 1, 1)또는 (0, 0, 0)만이 라운드 1과 2의 함수 입력으로 사용될 수 있다. 한편, 라운드 3은 입력에 어떠한 제약도 없으며 라운드 4의 모든 단계에서 임의의 20-워드 입력이 $I(\bar{x}, \bar{y}, \bar{z}) = \overline{I(x, y, z)}$를 만족할 확률은 2^{-16}이 된다. 그러므로 라운드 1과 2에서 함수 F와 G의 입력 워드들의 최상위비트가 1을 유지하도록 초기 버퍼값 A,B,C,D와 16-워드 $x[0], x[1], \cdots x[15]$를 구하면 4-라운드 전체에 대하여 2^{16}개의 후보들마다 하나의 충돌쌍을 찾을 수 있다.

4. MD5를 위한 부울함수 개발

부울함수에 기초한 MD5의 안전성은 부울함수의 안전성이 필수적이다. MD5에서 사용될 수 있는 부울함수의 암호학적 강도를 높이고 보수속성을 이용한 공격을 막을 수 있는 새로운 부울함수를 개발한다. 먼저 정보누설량의 관점에서 기존의 MD5에 사용된 부울함수의 암호학적 강도를 살펴보기로 한다. MD5 부울함수의 정보누설량을 살펴보면 표 1과 같다.

Table 1. Information leakage of MD5 boolean functions.

f(X_1, X_2, X_3) / SL	F $X_1X_2 \vee (\sim X_1)X_3$	G $X_1X_3 \vee X_2(\sim X_3)$	H $X_1 \oplus X_2 \oplus X_3$	I $X_2 \oplus (X_1 \vee (\sim X_3))$
SL(Y; X_1)	0.0000	0.1887	0.0000	0.0000
SL(Y; X_2)	0.1887	0.1887	0.0000	0.1887
SL(Y; X_3)	0.1887	0.0000	0.0000	0.0000
SL(Y; X_1, X_2)	0.5000	0.5000	0.0000	0.5000
SL(Y; X_1, X_3)	0.5000	0.5000	0.0000	0.0000
SL(Y; X_2, X_3)	0.5000	0.5000	0.0000	0.5000
SL(Y; $X_1 \oplus X_2$)	0.1887	0.0000	0.0000	0.1887
SL(Y; $X_1 \oplus X_3$)	0.1887	0.1887	0.0000	0.0000
SL(Y; $X_2 \oplus X_3$)	0.0000	0.1887	0.0000	0.1887
SL(Y; $X_1 \oplus X_2 \oplus X_3$)	0.0000	0.0000	1.0000	0.1887
SL(Y)	0.0000	0.0000	0.0000	0.0000
DL(Y; $\triangle X_1$)	0.0000	0.0456	0.0000	0.0000
DL(Y; $\triangle X_2$)	0.0456	0.0456	0.0000	0.0456
DL(Y; $\triangle X_3$)	0.0456	0.0000	0.0000	0.0000
DL(Y; $\triangle X_1$, $\triangle X_2$)	0.0944	0.0944	0.0000	0.0944
DL(Y; $\triangle X_1$, $\triangle X_3$)	0.0944	0.0944	0.0000	0.0000
DL(Y; $\triangle X_2$, $\triangle X_3$)	0.0944	0.0944	0.0000	0.0944
DL(Y; $\triangle X_1, \triangle X_2, \triangle X_3$)	0.2500	0.2500	1.0000	0.2500

표 1에서 모든 함수의 SL(Y)가 0이므로 모든 함수가 0-1 balance를 만족한다. 함수 F, G의 모든 SL(Y; $\mathbf{X}_{n-1}$)이 1/2이므로 F, G는 SAC와 completeness 성질을 만족한다.

함수 F, G, I는 모든 $\mathbf{X}_k$에 대하여 I(Y; $\mathbf{X}_k$)=0인 경우가 존재하지 않으므로 F, G, I는 correlation-inmmunity를 만족하지는 않는다. 함수 H은 2차의 correlation-inmmunity를 만족하며 특히 0-1 balance 이므로 2차-resilient 함수라고 한다. 그러나 고차의 정적 정보누설량이 0인 것은 입력 변수의 합 $\Sigma_k\mathbf{X}=\beta$와 출력 Y사이의 높은 정적 정보누설량을 야기시키기 때문에 낮은 차수의 정적 정보누설량이 0인 함수보다 반드시 통계적 암호분석에 더 강한 것은 아니다.

함수 F, G, I를 살펴보면 모든 경우에서의 입력 변수의 합 $\Sigma_k\mathbf{X}=\beta$와 출력 Y사이의 정적 정보누설량이 크지 않으므로 높은 비선형도를 가질 것이다. 그러나 함수 H는 SL(Y; $X_1 \oplus X_2 \oplus X_3$)이 1이므로 높은 비선형도를 만족시키지 못할 것이다. 실제 비선형도를 구해보면 F, G, I의 비선형도는 2이고 H의 비선형도는 0이다. H는 높은 비선형도를 가지지 않으므로 바람직하지 않다. 그리고 I는 SAC를 만족하지

않으므로 바람직하지 않다.

함수 F, G는 해밍거리가 1이하인 모든 $\Delta\mathbf{x}(0<wt(\Delta\mathbf{x})\leq 1)$에 대해 $DL(\Delta Y;\ \Delta\mathbf{X}\mid\Delta\mathbf{x})=0$ 으로 PC-1을 만족한다. 그러나 해밍거리가 1인 모든 $\mathbf{X}_{n-k}$에 대하여 $DL(\Delta Y';\ \Delta\mathbf{X}_{n-k}\mid\Delta\mathbf{x}_{n-k})\neq 0$이므로 1차 SAC성질은 만족하지 않는다. 함수 F, G는 전체적으로 동적 정보누설량이 크지 않으므로 차동분석에 취약하지 않다. 함수 H는 특정한 경우에서의 동적 정보 누설량을 0으로 하지만 모든 $\Delta\mathbf{x}$에 대해 $DL(\Delta Y;\ \Delta\mathbf{X}\mid\Delta\mathbf{x})=1$이므로 차동 분석에 매우 취약하다.

MD5의 부울함수를 정보누설량의 관점에서 분석한 결과 함수 H와 I가 통계적 분석에 취약함을 알 수 있었다. 그러므로 이 함수들을 대체할 새로운 부울함수를 구한다.

V_3상에서 가능한 함수의 시퀀스 갯수는 2^8이다. 이들의 정보누설량을 조사하여 0-1 balance, 높은 비선형도, completeness와 SAC를 만족하는 32개의 부울함수를 발견하였다. 이들의 정보누설량은 표 2와 같다. 편의상 해당함수를 시퀀스로 표현한다.

Table 2. Information leakage of boolean functions on V_3 satisfying 0-1 balance maximum nonlinearity, completeness and SAC.

sequence of funtion / SL, DL	[class 1]	[class 2]		
	17H, 2bH, 4dH, 71H, 8eH, b2H, d4H, e8H	caH, c5H, acH, a3H, 5cH, 53H, 3aH, 35H	e2H, d1H, b8H, 8bH, 74H, 47H, 2eH, 1dH	e4H, d8H, b1H, 8dH, 72H, 4eH, 27H, 1bH
$SL(Y;\ X_1)$	0.1887	0.0000	0.1887	0.1887
$SL(Y;\ X_2)$	0.1887	0.1887	0.0000	0.1887
$SL(Y;\ X_3)$	0.1887	0.1887	0.1887	0.0000
$SL(Y;\ X_1,\ X_2)$	0.5000	0.5000		
$SL(Y;\ X_1,\ X_3)$	0.5000	0.5000		
$SL(Y;\ X_2,\ X_3)$	0.5000	0.5000		
$SL(Y;\ X_1\oplus X_2)$	0.0000	0.1887	0.1887	0.0000
$SL(Y;\ X_1\oplus X_3)$	0.0000	0.1887	0.0000	0.1887
$SL(Y;\ X_2\oplus X_3)$	0.0000	0.0000	0.1887	0.1887
$SL(Y;\ X_1\oplus X_2\oplus X_3)$	0.1887	0.0000		
$SL(Y)$	0.0000	0.0000		
$DL(Y;\ \Delta X_1)$	0.0456	0.0000	0.0456	0.0456
$DL(Y;\ \Delta X_2)$	0.0456	0.0456	0.0000	0.0456
$DL(Y;\ \Delta X_3)$	0.0456	0.0456	0.0456	0.0000
$DL(Y;\ \Delta X_1,\Delta X_2)$	0.0944	0.0944		
$DL(Y;\ \Delta X_2,\Delta X_3)$	0.0944	0.0944		
$DL(Y;\ \Delta X_2,\Delta X_3)$	0.0944	0.0944		
$DL(Y;\ \Delta X_1,\Delta X_2,\Delta X_3)$	0.2500	0.2500		

V_3상에서 propagation criterion의 최대 차수는 2이며 분류 1의 함수는 이 성질을 만족하지만 $DL(\Delta Y;\ \Delta\mathbf{X}\mid 1)=1$이므로 바람직하지 않다. 분류 2의 함수는 전체적인 정보누설량이 분류 1의 함수보다 적고 $DL(\Delta Y;\ \Delta\mathbf{X}\mid 1)\neq 1$이므로 새로운 MD5의 부울함수를 이 가운데에서 선택하는 것이 바람직하다. 분류 2의 함수들 중에는 기존의 MD5의 부울함수가 존재하므로 이들을 선택한 후 이 함수들과 상호 output-uncorrelation한 함수들을 구해보면 8개의 함수가 있다. 이들의 시퀀스는 1dH, 3aH, 4eH, 74H, 8bH, b1H, c5H, e2H와 같다. 함수 f와 g가 0-1 balalce일때 $f\oplus g$도 0-1 balance라면 f, g는 상호 output-uncorrelation이라고 한다[10]. 위에서 구한 8개 함수들 간의 상호 output-uncorrelation는 표 3과 같다.

Table 3. Mutual output-uncorrelation of selected functions.

53H	○								
1dH	○	○							
3aH	○	○	○						
4eH	○	○	○	○					
74H	○	○	○	○	○				
8bH	○	○	○	○	○	×			
b1H	○	○	○	○	×	○	○		
c5H	○	○	○	×	○	○	○	○	
e2H	○	○	×	○	○	○	○	○	○
	27H	53H	1dH	3aH	4eH	74H	8bH	b1H	c5H

한편 부울함수의 보수속성을 이용한 공격이 성공하기 위해서는 모든 라운드의 부울함수들이 보수속성을 만족해야 한다. 그러므로 4-라운드 중에서 한 라운드만이라도 보수속성을 만족하지 않으면 이 공격을 막을 수 있다. 표 3의 함수 중에서 상호 output-uncorrelation한 함수 c5H, 53H, 27H, 1dH을 선택하여 차례로 4 라운드 함수로 사용하기로 한다. 여기서 라운드 1의 함수 c5H를 살펴보면 (0, 0, 1), (1, 1, 0)과 그의 보수 (1, 1, 0), (0, 0, 1)은 식 (16)을 만족하지만 그 다음에 이어지는 입력값이 계속해서 보수속성을 만족하지 않으므로 보수속성을 이용한 공격을 막을 수 있다.

개선된 MD5의 부울함수는 표 4와 같다.

Table 4. Improved boolean functions of MD5.

round	function	$f(X_1, X_2, X_3)$
1	F (c5H)	$X_1X_3 \vee (\sim X_1)(\sim X_2)$
2	G (53H)	$X_1X_2 \vee (\sim X_1)X_3$
3	H (27H)	$X_1X_3 \vee X_2(\sim X_3)$
4	I (1dH)	$X_1(\sim X_2) \vee X_2X_3$

5. 결 론

부울함수에 기초한 해쉬함수는 디지탈서명 또는 메세지인증 등의 적용시 블럭함수에 기초한 해쉬함수보다 속도면에서 월등하다. 한편 부울함수에 기초한 해쉬함수는 부울함수가 암호학적으로 강해야함이 필수적이다.

본 논문에서는 MD5의 부울함수로 이용할 수 있는 함수의 측정기준을 정보누설량의 관점에서 분석하여 0-1 balance, 비선형도, completeness와 SAC를 만족하는 한편 전체적으로 정보누설량이 크지 않은 함수를 살펴보았다. 또한 부울함수의 보수속성을 이용한 공격을 분석하였다. 분석한 결과를 바탕으로 부울함수의 암호학적 강도를 향상시키고 보수속성을 이용한 공격을 막을 수 있는 새로운 부울함수를 제안하였다.

정보누설량을 고려한 부울함수 설계는 SHS와 같은 부울함수를 이용한 암호시스템에 적용가능할 것이다.

참 고 문 헌

[1] R. C. Merkle, " One way hash functions and DES," *Crypto'89 Abstract*, Aug. 1989, pp. 407-417.

[2] C.J.Mitchell, F.piper and P.Wild, "Digital signatures," in *Contemporary Cryptology: The Science of Information Integrity*, G.J.Simmons, editor, IEEE Press, 1991, pp.325-378

[3] R. Rivest and S. Dusse, "The MD5 message digest algorithm," *Internet-draft, July, 1991*

[4] Jennifer Seberry and Xian-Mo Zhang,"Highly Nonlinear 0-1 Balanced Boolean Functions Satisfying Strict Avalanche Criterion," *AusCrypt'92 Abstracts*, 1992.

[5] T. Siegenthaler. "Correlation-Immunity of Nonlinear Combining Functions for cryptographic Applocations," *IEEE Trans. on info. theory*, Vol.IT-30, No.5:776-780, Sept. 1984.

[6] B. Preneel, W. V. Leekwijk, L. V. Linden, R. Govaerts, and J. Vandewalle. "Propagation Characteristics of Boolean Functions," *Advances in cryptology, Proceedings of Eurocrypt'90, Springer-verlag, Berlin*, pp 161-173, 1991.

[7] R. Forre. "The Strict Avalanche Criterion : Spectral Properties of Boolean Functions and an Extended Definition," *Advances in cryptology, Proceedings of Eurocrypt'88, Springer-verlag, Berlin*, pp 450-468, 1989.

[8] M. Zhang, S.E. Tavares, and L. L. Campbell, "Information Leakage of Boolean Functions as a Measure of Cryptographic Stength," *Proc. of SAC 94, Kingston*, pp. 40-51, 1994.

[9] B. den Boer and A. Bosselaes, "Collisions for the Compression Function of MD5," *EuroCrypt'93 Abstracts*, 1993.

[10] Y. Zheng, J. Pieprzyk, and J. Seberry, "HAVAL - A One-Way Hashing Algorithm with Variable Length of Output," *AusCrypt'92* Abstracts, 1992