Hash Function

- Also known as
  - Message digest
  - One-way transformation
  - One-way function
  - Hash

- Length of $H(m)$ much shorter than length of $m$
- Usually fixed lengths: 128 or 160 bits
Requirements for a Hash Function

- Consider a hash function H
  - **Flexibility**: Can be applied to a block of data of any size
  - **Convenience (for check)**: produce a fixed-length short output.
  - **Performance**: Easy to compute H(m)
  - **One-way property**: Given H(m) but not m, it’s difficult to find m
  - **Weak collision resistance (free)**: Given H(m), it’s difficult to find m' such that H(m') = H(m).
  - **Strong collision resistance (free)**: Computationally infeasible to find m₁, m₂ such that H(m₁) = H(m₂)

Birthday Paradox

- **Question:**
  - What is the minimum value of k such that the probability is greater than 0.5 that at least two people in a group of k people have the same birthday?
    - Ignore February 29 and assume each birthday is equally likely.
    - Probability of k people having k different birthdays:
      \[ Q(365,k) = \frac{365!}{(365-k)!365^k} \]
    - Probability that at least two people have the same birthday:
      \[ P(365,k) = 1 - Q(365,k) \]
    - K is about 23.
Generalization of Birthday Paradox

- Given a random variable that is an integer with uniform distribution between 1 and \( n \) and a selection of \( k \) instances of the random variables, what is the least value of \( k \) such that the probability \( P(n,k) \) is greater than 0.5 that there is at least one duplicate?
  - \( P(n,k) > 1 - e^{k(k-1)/2n} \)
  - For large \( n \) and \( k \), we have
    \[
    k = \sqrt[2]{(\ln 2)n} = 1.18\sqrt{n} = \sqrt[2]{n}
    \]
  - Intuition: How many \( k \) do we need to have a collision with \( P=0.5 \)?

- Implication
  - For a hash function \( H \) with \( 2^m \) possible outputs, if we apply \( H \) to \( k=(2^m)/2=2^{m/2} \) random inputs, the probability that there is at least one duplicate is greater than 0.5.

Birthday Attack

- The source, A, is prepared to sign a message
- The opponent generates \( 2^{m/2} \) variations on the message, and prepares \( 2^{m/2} \) variations on the fraudulent message.
- The opponent compares the two sets of messages to find a pair of messages that produces the same hash value. The probability of success is greater than 0.5. The opponent repeats generating variations until a match is found.
- The opponent offers the valid variation to A for signature, but attaches the signature to the fraudulent variation.
How Many Bits for Hash?

- $m$ bits, takes $2^{m/2}$ to find two with the same hash at the probability 0.5
- 64 bits, takes $2^{32}$ messages to search duplicate
- Need at least 128 bits

Building Hash Using Block Chaining Techniques

- Divide M into fixed-size blocks $M_1, M_2, \ldots, M_n$
- Compute the hash as follows
  - $H_0$=Initial value
  - $H_i=E_{M_i}(H_{i-1})$
  - Hash value $G=H_n$
- Weakness
  - Birthday attack (reason: hash value is too short)
  - Meet-in-the-middle attack
Building Hash Using Block Chaining Techniques (Cont’d)

- Meet-in-the-middle attack
  - Get the correct hash value G
  - Construct any message in the form Q₁, Q₂, … , Qₙ₋₂
  - Compute \( H_i = E_{Q_i}(H_{i-1}) \) for \( 1 \leq i \leq (n-2) \).
  - Generate \( 2^{m/2} \) random blocks; for each block X, compute \( E_X(H_{n-2}) \).
  - Generate \( 2^{m/2} \) random blocks; for each block Y, compute \( D_Y(G) \).
  - With high probability there will be an X and Y such that \( E_X(H_{n-2}) = D_Y(G) \).
  - Form the message \( Q_1, Q_2, \ldots, Q_{n-2}, X, Y \). It has the hash value G.

Modern Hash Functions

- MD5
  - Previous versions (i.e., MD2, MD4) have weaknesses.
- SHA (Secure Hash Algorithm)
  - Weaknesses were found
- SHA-1
  - Broken; collisions were published in February 2005.
- RIPEMD-160
MD5: Message Digest Version 5

input Message

Output 128 bits Digest

MD5: A High-Level View

K bits
Padding (1 to 512 bits)
Message Length (K mod 2^64)

Message
100…0

Y_0

Y_1

…

Y_{L-1}

IV

128 bits

MD5

CV_1

MD5

CV_{L-1}

128-bit digest
Padding

- Given original message M, add padding bits “10*” such that resulting length is 64 bits less than a multiple of 512 bits.
- Append \((\text{original length in bits} \mod 2^{64})\), represented in 64 bits to the padded message.
- Final message is chopped 512 bits a block.
- Exercise:
  - How to add padding bits to a message that is already a multiple of 512 bits?

MD5 (Intermediate) Buffer

- Used to hold intermediate and final result of MD5.
- 128 bits
- Represented as four 32-bit words
  - \((A,B,C,D)\)
  - Initially, \(A=0x67452301, B=0xEFCDAB89, C=0x98BADCFE, D=0x10325476\)
  - Stored in little-endian format, \(A=0x01234567, B=0x89ABCDEF, C=0xFEDCBA98, D=0x76543210\). 
Processing of A Single Block

512-bit message block (16 words)

128-bit vector
(Initial or from
the previous
block)

MD5

128-bit result

Primitive operations
used in MD5:

\[
\begin{align*}
F(x,y,z) &= (x \land y) \lor (\neg x \land z) \\
G(x,y,z) &= (x \land z) \lor (y \land \neg z) \\
H(x,y,z) &= x \oplus y \oplus z \\
I(x,y,z) &= y \oplus (x \land \neg z)
\end{align*}
\]

+: addition mod 2^{32}

x_{\text{ly}}: x \text{ left rotate } y \text{ bits}

Processing of A Single Block (Cont’d)

- Every message block contains 16 32-bit words:
- Every stage consists of 4 rounds over the message block, each modifying the MD5 buffer (A,B,C,D).
  - The four rounds use functions F, G, H, I, respectively.
- Each round uses one-fourth of a 64-element table T[1…64].
  - T[i] = 2^{32}\#abs(sin(i)) represented in 32 bits.
- The output of the fourth round is added to the input to the first round.
Processing of Block $m_i$ : 4 Rounds

Logic of Each Round

- Each round consists of 16 steps
- Each step is of the form
  - $A \leftarrow B + ((A + g(B,C,D) + X[k] + T[i]) \ll s)$
    - Function $g$ is one of $F$, $G$, $H$, $I$
    - $X[k]$ is the word in the input
    - $T[i]$ is the $i$th word in $T$
    - $\ll s$: circular left shift by $s$ bits.
  - Followed by a word level circular right shift of one word.
Logic of Each Step

- Within a round, each of the 16 words of X[i] is used exactly
  - First round, X[i] are used in the order of I
  - Round 2, in the order of \( \rho_2(i) \), where \( \rho_2(i) = (1+5i) \text{mod} 16 \);
  - Round 3, in the order or \( \rho_3(i) \), where \( \rho_3(i) = (5+3i) \text{mod} 16 \);
  - Round 4, in the order or \( \rho_4(i) \), where \( \rho_2(i) = 7i \text{mod} 16 \);
- Each word of T[i] is used exactly once.
Security of MD5

• A recently discovered method can find a collision in a few hours
  – A few collisions were published on 8/17/04
  – Exact method has not been published yet
  – Can find many collisions for two 1024-bit messages
  – SHA-1 was also broken; collisions were published on 2/13/05.

• Birthday attack
  – $2^{64}$