Data Protection

Jorge L. Villar

MCYBERS, UPC, Fall 2025

Outline

Hash Functions

A map $H : \{0,1\}^* \to \{0,1\}^\ell$ behaving like a **random function**.

Hash Functions

A map $H : \{0,1\}^* \to \{0,1\}^\ell$ behaving like a **random function**.

Definition (Preimage Resistance)

Given a random $y \in \{0,1\}^\ell$, it is infeasible to compute $x \in \{0,1\}^*$ such that $H(x) = y$.

Definition (Second Preimage Resistance)

Given a random $x_1 \in \{0,1\}^*$, it is infeasible to compute a different $x_2 \in \{0,1\}^*$ such that $H(x_1) = H(x_2)$.

Definition (Collision Resistance)

It is infeasible to compute different $x_1, x_2 \in \{0,1\}^*$ such that $H(x_1) = H(x_2)$.

Ideal Behavior

Definition (Random Function)

For any different $x_1, x_2, \ldots \in \{0,1\}^*$, $H(x_1), H(x_2), \ldots$ are independent uniform random variables.

Ideal Behavior

Definition (Random Function)

For any different $x_1, x_2, \ldots \in \{0,1\}^*$, $H(x_1), H(x_2), \ldots$ are independent uniform random variables.

Preimage Resistance: Expected 2^ℓ hash computations to break it.

Second Preimage Resistance: Expected 2^ℓ hash computations to break it.

Collision Resistance: Expected $2^{\ell/2}$ hash computations to break it.

Ideal Behavior

Definition (Random Function)

For any different $x_1, x_2, \ldots \in \{0,1\}^*$, $H(x_1), H(x_2), \ldots$ are independent uniform random variables.

Preimage Resistance: Expected 2^ℓ hash computations to break it.

Second Preimage Resistance: Expected 2^ℓ hash computations to break it.

Collision Resistance: Expected $2^{\ell/2}$ hash computations to break it.

Birthday paradox: The probability to find a collision in n random values in $\{0,1\}^\ell$ is approx. $n^2 2^{-(\ell+1)}$.

Random Oracle Model

Random Oracle Model (ROM): The security of a cryptosystem is analyzed by replacing a Hash function by a truly random function available to all parties.

Random Oracle Model

Random Oracle Model (ROM): The security of a cryptosystem is analyzed by replacing a Hash function by a truly random function available to all parties.

No real random function (available to all parties) exists.

The random function is simulated with a dynamically generated random table $(x_i, y_i = H(x_i))$.

Random Oracle Model

Random Oracle Model (ROM): The security of a cryptosystem is analyzed by replacing a Hash function by a truly random function available to all parties.

No real random function (available to all parties) exists.

The random function is simulated with a dynamically generated random table $(x_i, y_i = H(x_i))$.

The security proofs in the ROM are just heuristic.

There are known weak systems that are secure in the ROM.

Practical Constructions

Hash Functions are not Symmetric Key objects, but the tools used to build them are very similar:

- Diffusion
- Iteration
- (No key expansion)

Practical Constructions

Hash Functions are not Symmetric Key objects, but the tools used to build them are very similar:

- Diffusion
- Iteration
- (No key expansion)

Construction strategies:

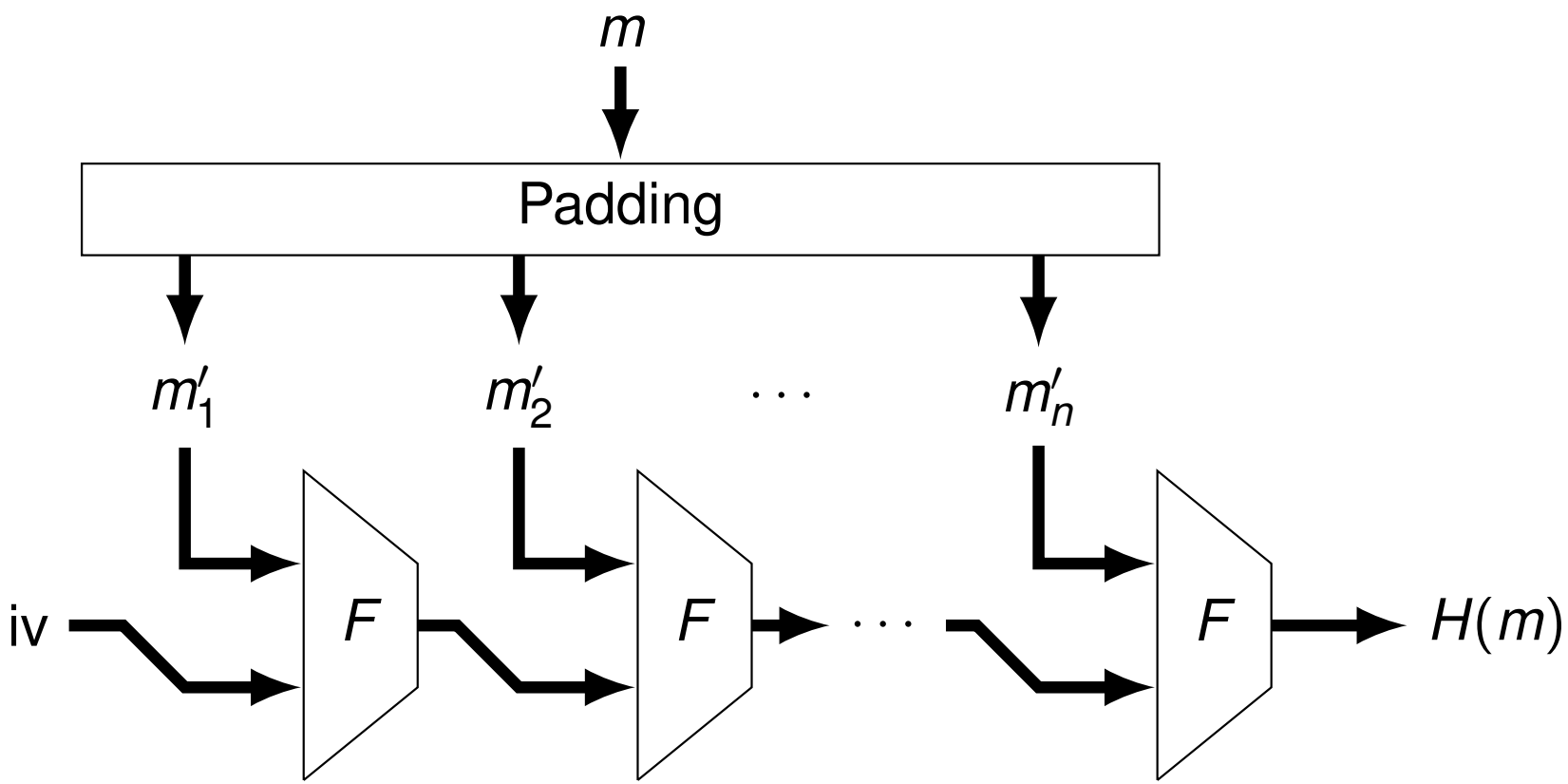
- Merkle-Damgård (e.g. MD5, SHA-1, SHA-2)
- Sponge (e.g. SHA-3)

Merkle-Damgård Construction

Generic construction of H from a compression function $F : \{0,1\}^r \times \{0,1\}^\ell \to \{0,1\}^\ell$.

- The message m is padded to a length multiple of r.
- The padded message m' is split into r-bit blocks $(m'_1, \ldots, m'_n)$.
- A IV is fixed by design.
- The iteration is
  $s_0 = IV$
  $s_i = F(m'_i, s_{i-1})$.
- $H(m)$ is set to $G(s_n)$, for some finalization function G.

Merkle-Damgård Construction



Merkle-Damgård Security

The padding scheme must fulfil:

- m is a prefix of its padded version m'.
- If x and y have equal lengths, then so do x' and y'.
- If x and y have different lengths, then the last blocks in x' and y' are different.

E.g., length-padding: $m' = m||10\ldots0||d$, where d is a 64-bit representation of the bit-length of m.

Merkle-Damgård Security

The padding scheme must fulfil:

- m is a prefix of its padded version m'.
- If x and y have equal lengths, then so do x' and y'.
- If x and y have different lengths, then the last blocks in x' and y' are different.

E.g., length-padding: $m' = m||10\ldots0||d$, where d is a 64-bit representation of the bit-length of m.

Proposition

If F is collision resistant then H is also collision resistant.

MD5

Message block size: $r = 512$ bits

Hash result and internal state size: $\ell = 128$ bits

Number of rounds: 64

Round structure:

- Split the internal state into (A, B, C, D).
- $A = A + G_i(B, C, D) + M_i + K_i \mod 2^{32}$.
- $A = \text{rotl}(A, s_i) + B \mod 2^{32}$.
- Cyclically permute (A, B, C, D) one position to the left.

The resulting internal state after processing a data block is added to the original one (the state before starting the iteration).

No finalization function is applied (the hash value is just the last internal state).

SHA-1 and SHA-2

Similar to MD5, but with improved security.

- SHA-1 uses $\ell = 160$, the internal state is split in 5 words (A, B, C, D, E) in a bit more complex round structure.
- SHA-2 family uses a larger ℓ (224, 256, 384 and 512), and a similar round design (with more complex nonlinear functions).

SHA-1 and SHA-2

Similar to MD5, but with improved security.

- SHA-1 uses $\ell = 160$, the internal state is split in 5 words (A, B, C, D, E) in a bit more complex round structure.
- SHA-2 family uses a larger ℓ (224, 256, 384 and 512), and a similar round design (with more complex nonlinear functions).

Known attacks:

- MD5: Easy to find lots of collisions.
- SHA-1: Some collisions known but still considered as second preimage resistant.
- SHA-2: No practical attack known to date.

Sponge Construction

Design elements:

- Asymmetric accumulator (R, C) as internal state.
- Stirring function $F : \{0,1\}^{r+c} \to \{0,1\}^{r+c}$.
- Only the R part interacts with input/output values.

Sponge Construction

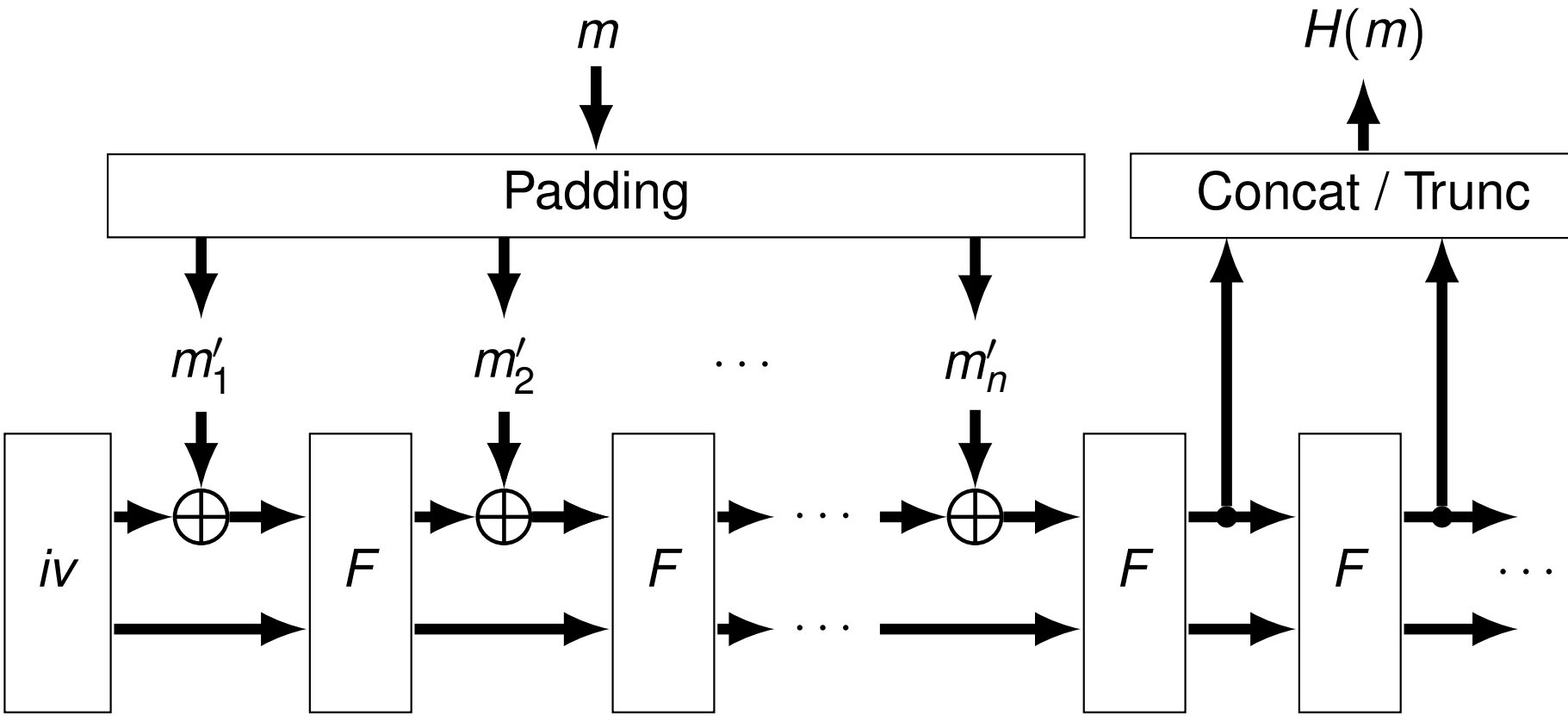
Design elements:

- Asymmetric accumulator (R, C) as internal state.
- Stirring function $F : \{0,1\}^{r+c} \to \{0,1\}^{r+c}$.
- Only the R part interacts with input/output values.

Operation:

- The message m is padded to a length multiple of r.
- The padded message is split into r-bit blocks $(m'_1, \ldots, m'_n)$.
- An IV is fixed by design: $(R_0, C_0) = IV$.
- The "input" iteration is: $(R_i, C_i) = F(R_{i-1} \oplus m'_i, C_{i-1})$
- The "output iteration" is: $(R_{n+j}, C_{n+j}) = F(R_{n+j-1}, C_{n+j-1})$
- The output value is $H(m) = R_n \| \cdots \| R_{n+u}$, $u = \lfloor \ell / r \rfloor$
- The last block can be truncated to obtain the desired ℓ.

Sponge Construction



SHA-3

Sponge construction with $r + c = 1600$ bits

Sizes: $u = 0$, $c = 2\ell$ and $r = 1600 - 2\ell$

Number of rounds: 24

Round structure:

- (R, C) is organized as a 5×5 64-bit words matrix.
- XOR elements with the XOR of adjacent columns.
- Apply different bit-rotations to the 25 words.
- Permute the 25 words.
- XOR rows with a nonlinear function of the other rows.
- XOR one of the 25 words with the output of an LFSR.

Different members in the family: $\ell = 224, 256, 384, 512$

Outline

HMAC

$H(m)$ is a digest of message m (no key involved).

Turn it into a MAC: Append the key and the message
$HMAC(k, m) = H(k \| m)$

Insecure for some Merkle-Damgård based hashes!
(e.g. MD5, SHA1, SHA2)

HMAC

$H(m)$ is a digest of message m (no key involved).

Turn it into a MAC: Append the key and the message
$HMAC(k, m) = H(k \| m)$

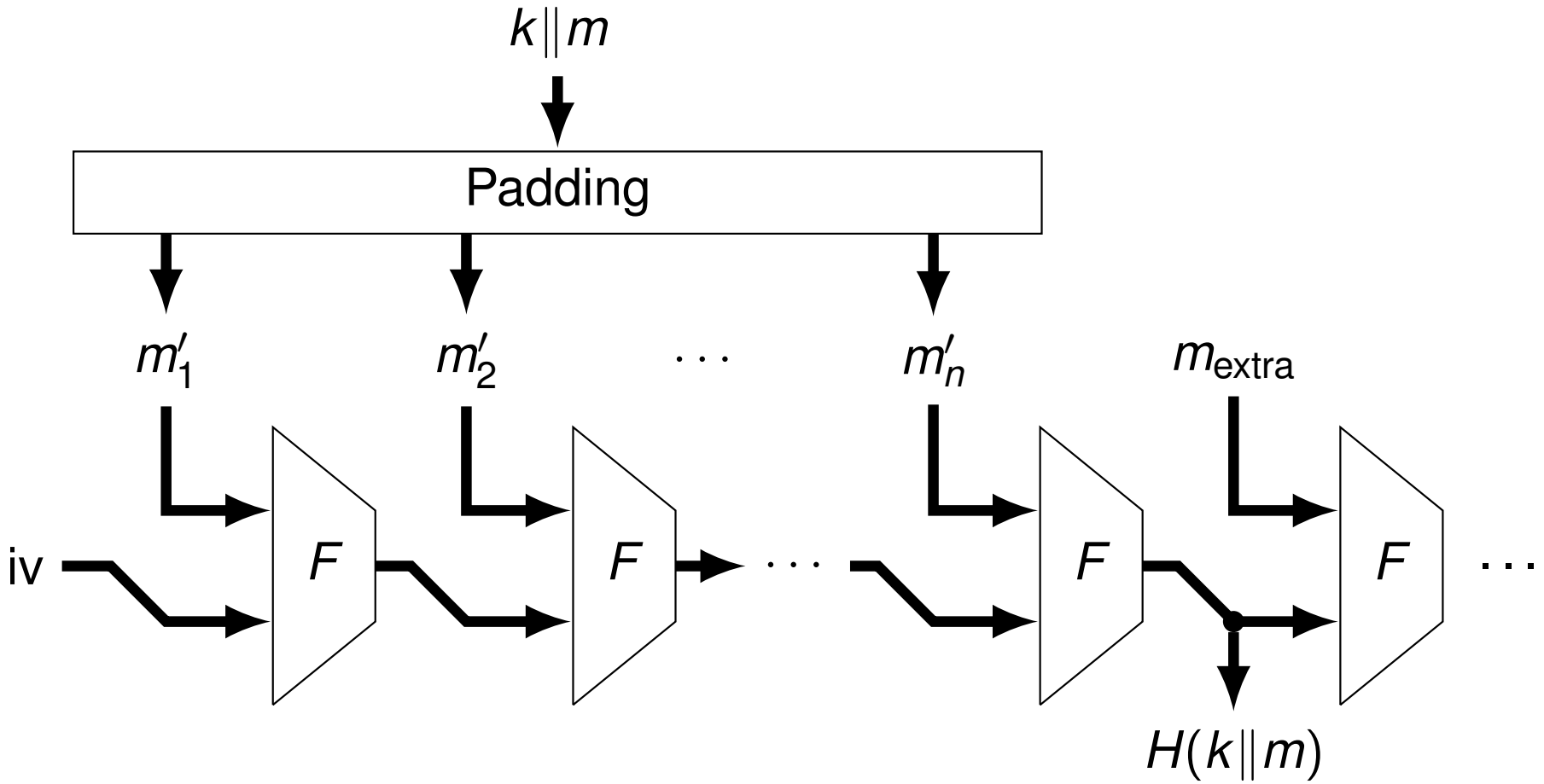
Insecure for some Merkle-Damgård based hashes!
(e.g. MD5, SHA1, SHA2)

Fix: Use two passes
$HMAC(k, m) = H((k' \oplus P_{\text{out}}) \| H((k' \oplus P_{\text{in}}) \| m))$

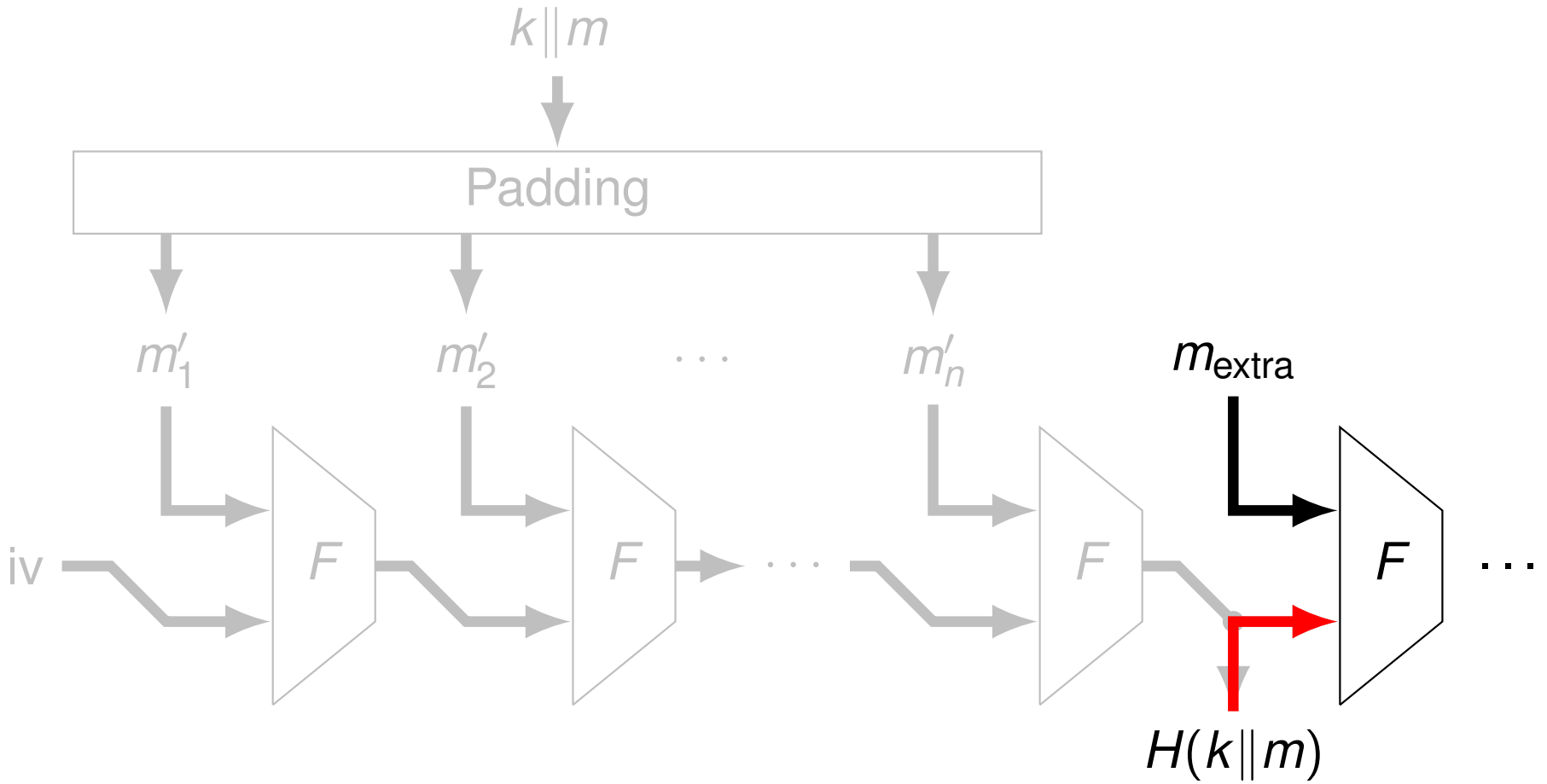
If k is longer than a hash block, then $k' = H(k)$.
Otherwise, $k' = k$, or $k' = k \| 0 \cdots 0$.

Not necessary for SHA-3:
$HMAC\text{-}SHA3(k, m) = SHA\text{-}3(k \| m)$

Merkle-Damgård Prolongation Attack



Merkle-Damgård Prolongation Attack



Hash Puzzles

Hash preimage resistance $\longrightarrow$ Problem set with **tunable** hardness.

Problem (Hash Puzzle)

Given t, find $x \in \{0,1\}^*$ such that $H(x)$ has at least t leading zeros.

Hash Puzzles

Hash preimage resistance $\longrightarrow$ Problem set with **tunable** hardness.

Problem (Hash Puzzle)

Given t, find $x \in \{0,1\}^*$ such that $H(x)$ has at least t leading zeros.

- The expected number of hash computations is 2^t.
- No possible speed-up.
- No possible precomputation if x must have a given fresh prefix.
- Applications as proof-of-work in antispam tools or blockchain.

Merkle Trees

Efficient hashing of a collection of documents.

- Documents are the leaves of a binary tree.
- The value of each node in the tree is the hash of its childrens' values.
- The value of the root is the hash of the collection.
- Efficient proof that a document is part of the collection.
- Efficient document update.
- Efficient addition of new documents.

Merkle Trees

Root hash:

$h_\emptyset = H(1 \| h_0 \| h_1)$

Intermediate node hash:

$h_x = H(1 \| h_{x0} \| h_{x1})$, if node $x1$ exists.

$h_x = H(1 \| h_{x0})$, otherwise

Leave:

$h_x = H(0 \| m_x)$

Merkle Trees

Root hash:
$h_\emptyset = H(1 \| h_0 \| h_1)$

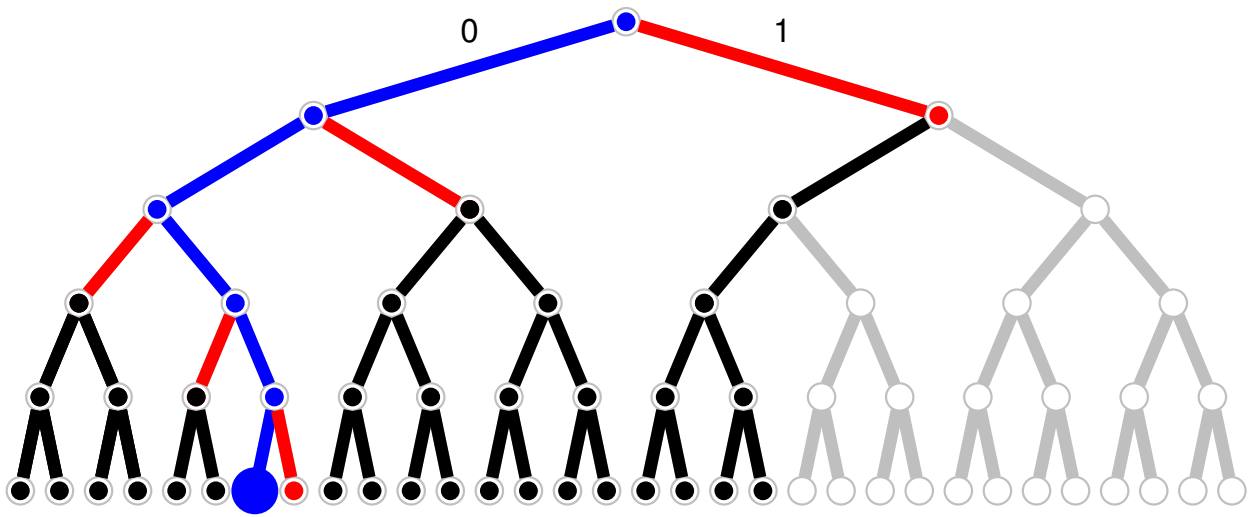
Intermediate node hash:
$h_x = H(1 \| h_{x0} \| h_{x1})$, if node $x1$ exists.
$h_x = H(1 \| h_{x0})$, otherwise

Leave:
$h_x = H(0 \| m_x)$

Example of **proof of membership** for $m = m_{00110}$
$\pi_{00110} = (h_1, h_{01}, h_{000}, h_{0010}, h_{00111})$

If some sibling does not exist the hash is replaced by $\perp$.

Example



Example of a proof of membership for document m_{00110}, in an incomplete binary tree.

$\pi_{00110} = (h_1, h_{01}, h_{000}, h_{0010}, h_{00111})$

Outline

Definition

Providing both **confidentiality** and **integrity** at once.

Trivial solutions:

- Encrypt-then-MAC:
  (c, t) where $c = \mathrm{Enc}(k_1, m); t = \mathrm{MAC}(k_2, c)$
- MAC-then-Encrypt:
  $c = \mathrm{Enc}(k_1, m \| t)$ where $t = \mathrm{MAC}(k_2, m)$

Definition

Providing both **confidentiality** and **integrity** at once.

Trivial solutions:

- Encrypt-then-MAC:
  (c, t) where $c = \mathrm{Enc}(k_1, m); t = \mathrm{MAC}(k_2, c)$
- MAC-then-Encrypt:
  $c = \mathrm{Enc}(k_1, m \| t)$ where $t = \mathrm{MAC}(k_2, m)$

Improvements:

- Compact secret keys
  Ideally, a single short key
- Small computational overhead
  Similar cost as encryption
- Design flexibility
  Choice of the block cipher, auxiliary unencrypted data

Encrypt-then-MAC vs. MAC-then-Encrypt

Using independent keys:

- **MAC-then-Encrypt**
  Insecure if MAC and Decrypt produce different error messages.
- **Encrypt-then-MAC**
  Generically secure (the IV must be included in the MAC!).
  The ciphertext is authenticated (but not the message itself).

Encrypt-then-MAC vs. MAC-then-Encrypt

Using independent keys:

- **MAC-then-Encrypt**
  Insecure if MAC and Decrypt produce different error messages.
- **Encrypt-then-MAC**
  Generically secure (the IV must be included in the MAC!).
  The ciphertext is authenticated (but not the message itself).

Never use the same key for Encrypt and MAC!
Do it only if you know a specific security proof.

E.g., same key for CBC-MAC and Enc-CBC is insecure.

CCM

Same key for both MAC and Enc. (Particular construction.)

MAC-then-Encrypt design:

$t = \text{CBC-MAC}(k, m)$
$c = \text{Enc-CTR}(k, m \| t)$

The IV used in CBC-MAC (IV=0) must never be used as a value of the counter in Enc-CTR.

It costs two block cipher encryptions per message block.

OCB

- No extra ciphertext blocks due to padding.
- Single key.
- Almost one block cipher call per message block.
- Special design (it uses an “improper” MAC and a particular mode of operation).
- Authenticated unencrypted data can also be added to the tag computation.

Ciphertext blocks: $c_i = E_k(m_i \oplus \Delta_i) \oplus \Delta_i, \quad i = 1, \ldots, n$

Last incomplete block: $c_* = \mathrm{trunc}(E_k(\Delta_*)) \oplus m_*$

Authentication tag: $t = \mathrm{trunc}(E_k(m_1 \oplus \cdots \oplus m_n \oplus m'_* \oplus \Delta_\$))$

m'_* is 0 if no incomplete block exists, or m_* padded with a 1 and zero or more 0, otherwise.

Constants $\Delta_1, \ldots, \Delta_n, \Delta_*, \Delta_\$$ are computed from k and IV.

Data Protection

Jorge L. Villar

MCYBERS, UPC, Fall 2025

END