CIS 5371, Fall 2024

Hash Functions

Viet Tung Hoang

Agenda

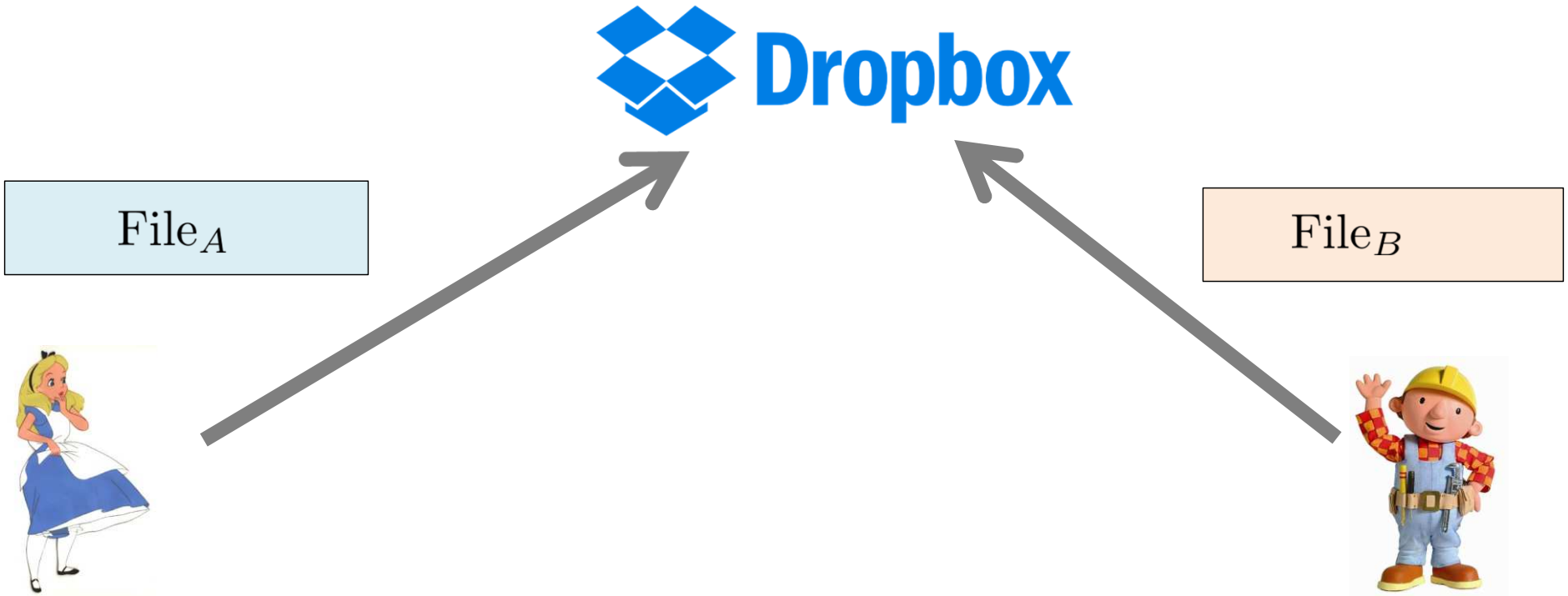
1. **Security Modeling for Hash Functions**

2. **Building Hash Function: MD Transform**

3. **Generic Attacks**

4. **Applications**

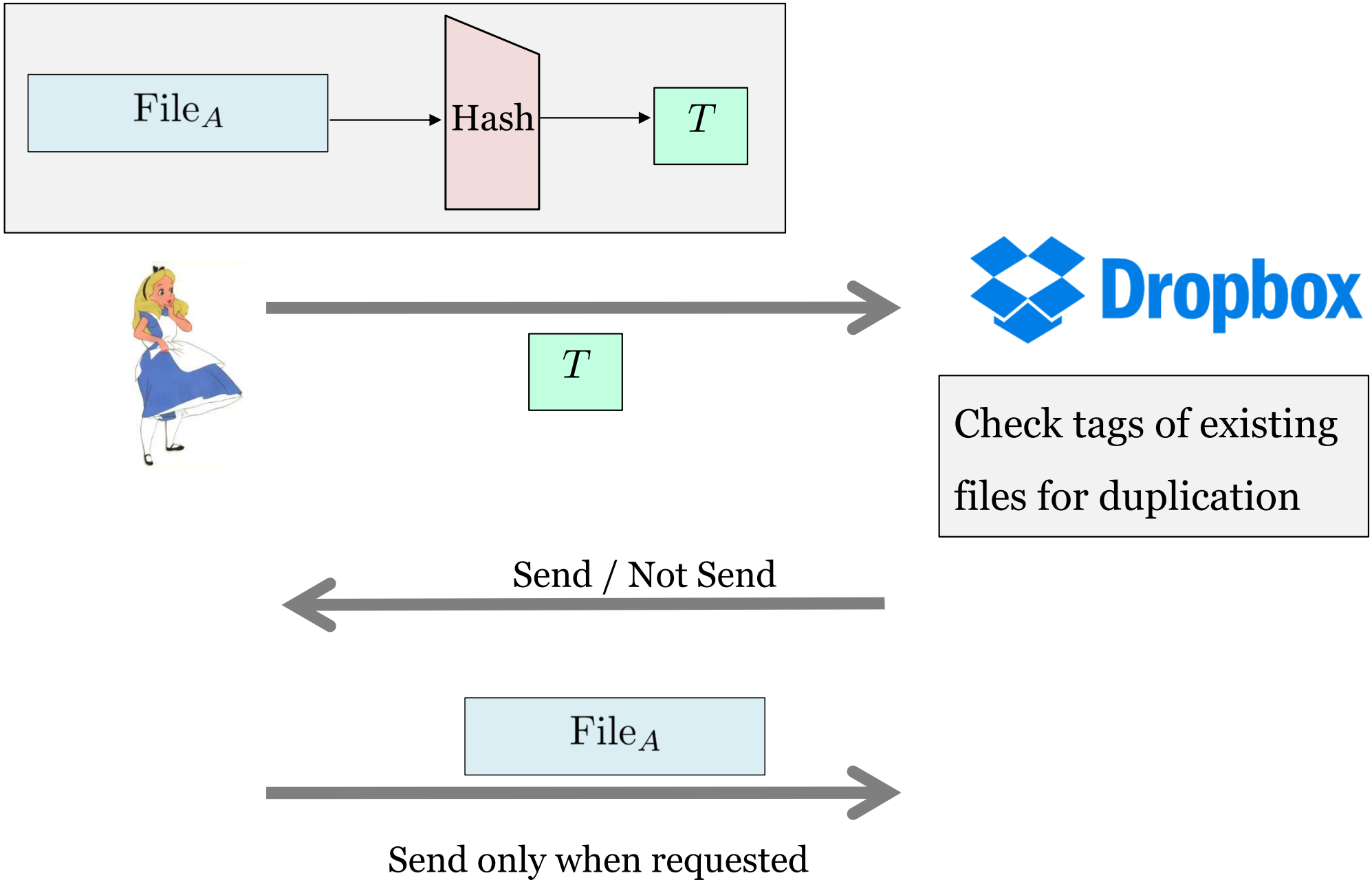
Motivating Application: Data Deduplication



Dropbox's goals:

- If many users store the same file, keep only a **single** copy
- Minimize bandwidth usage

Motivating Application: Data Deduplication

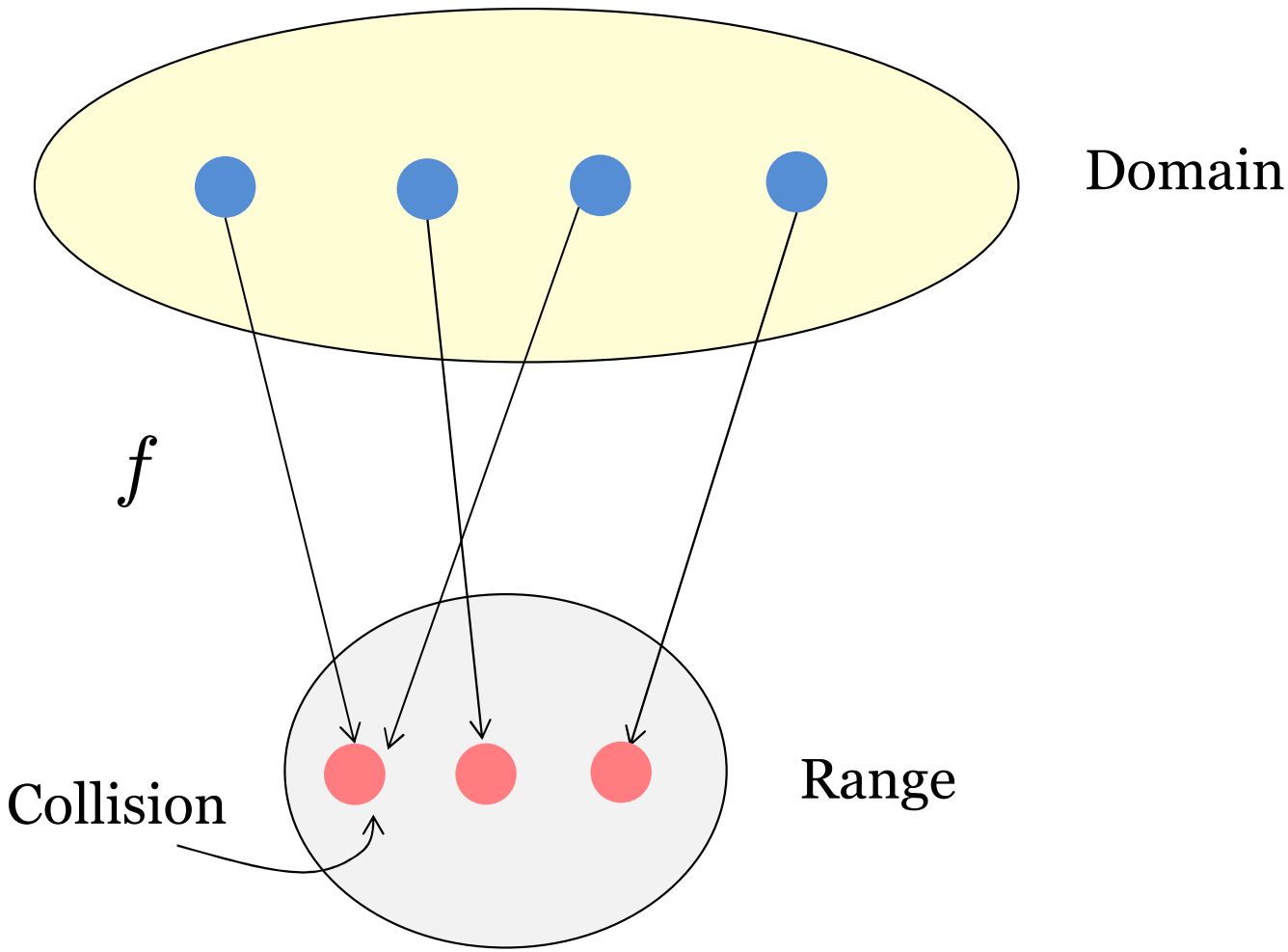


What property

do we need for the hash?

Collision-Resistance

$f : \text{Domain} \rightarrow \text{Range}$



By Pigeonhole Principle, if |Domain| > |Range| then collision exists

Want: collisions are hard to find, although they exist

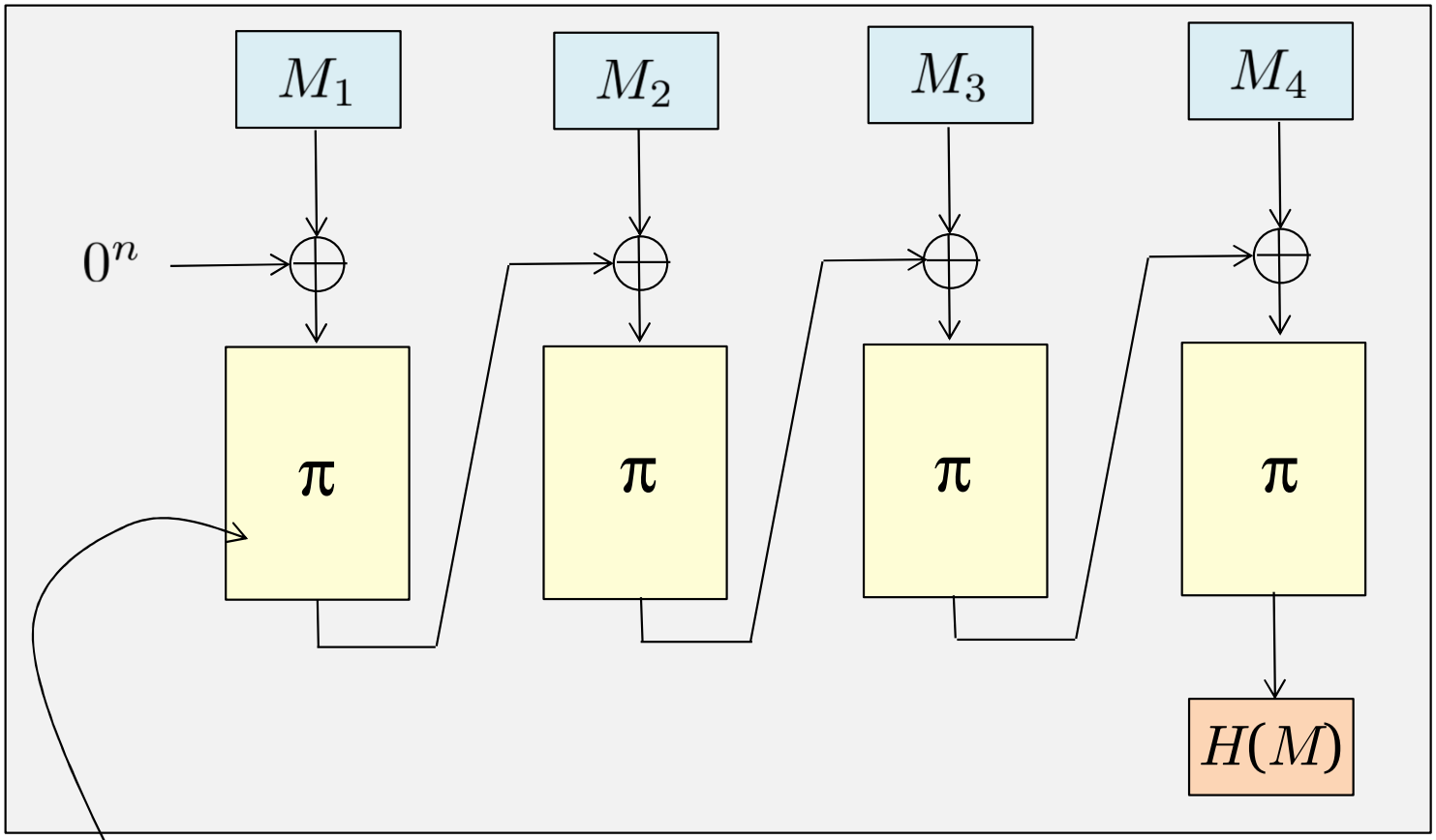
Defining Collision-Resistance

$$A \longrightarrow (X_1, X_2)$$

Must be distinct

$$\mathbf{Adv}_H^{\mathrm{cr}}(A) = \Pr[H(X_1) = H(X_2)]$$

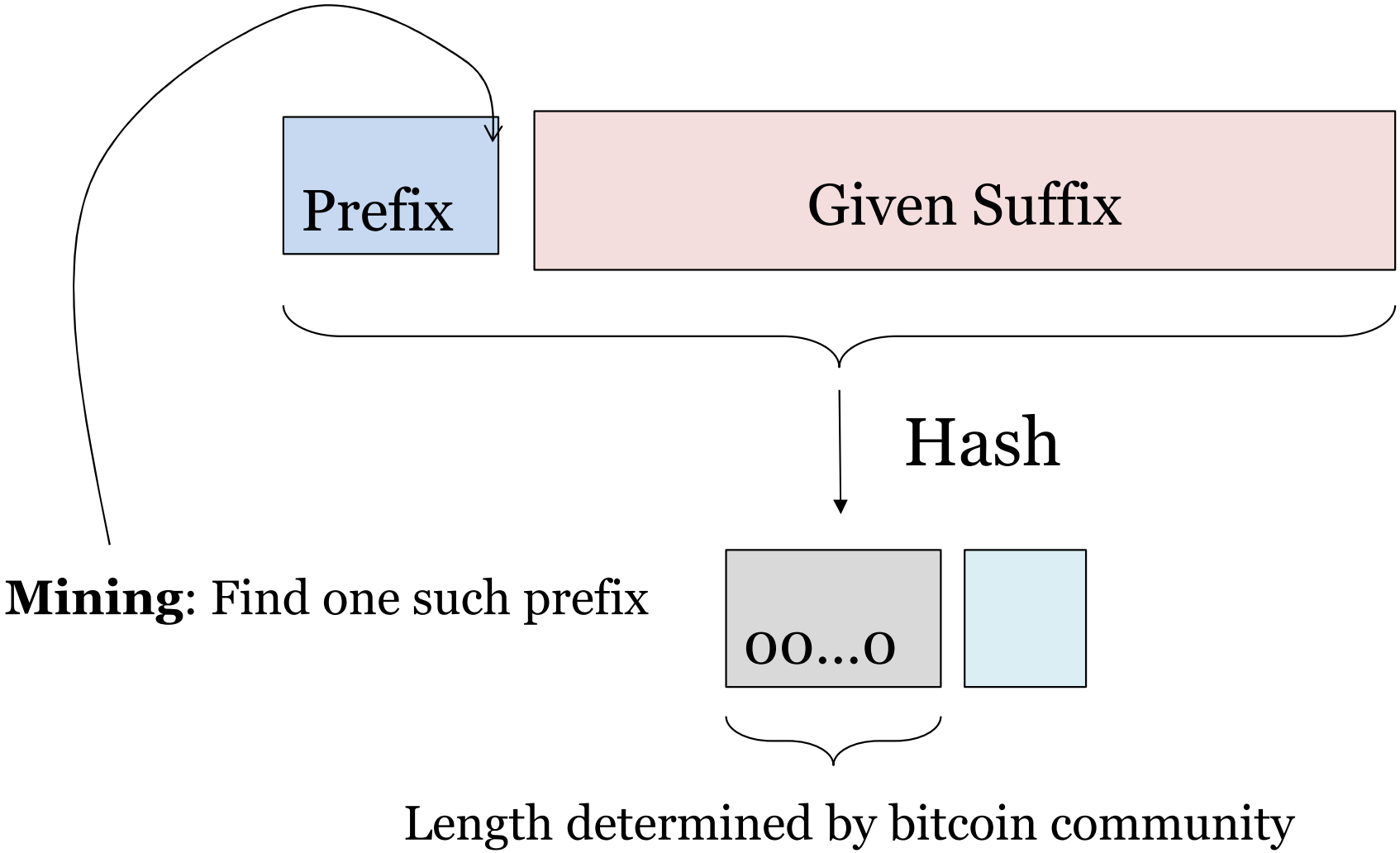
Exercise: Break Collision Resistance



π, π^{-1} are public

Public permutation

CR Is Not Enough: Bitcoin Mining



Mining: Find one such prefix

Want: Can’t mine faster than brute-force

Modeling Security of Hash Functions

The Random Oracle Model

RO

Maintain a secret random $f : \{0,1\}^n \to \{0,1\}^m$

Everybody, including the adversary, has access to RO

Agenda

1. Security Modeling for Hash Functions

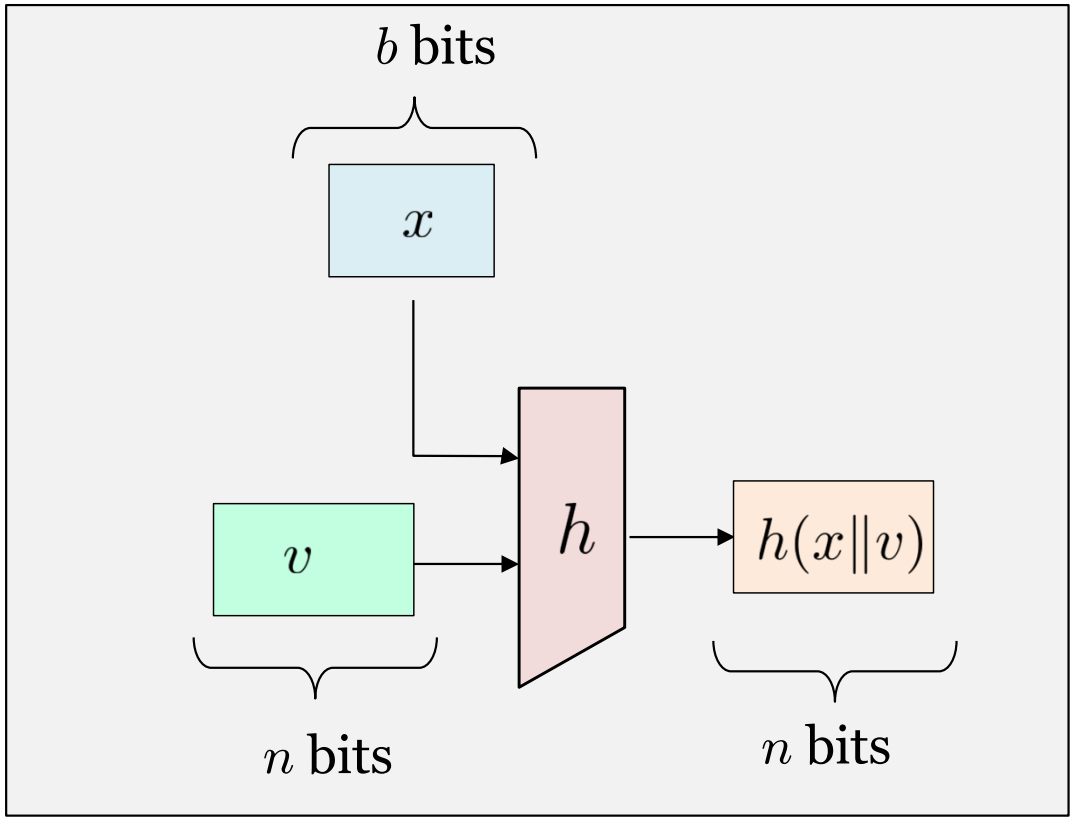
2. **Building Hash Function: MD Transform**

3. Generic Attacks

4. Applications

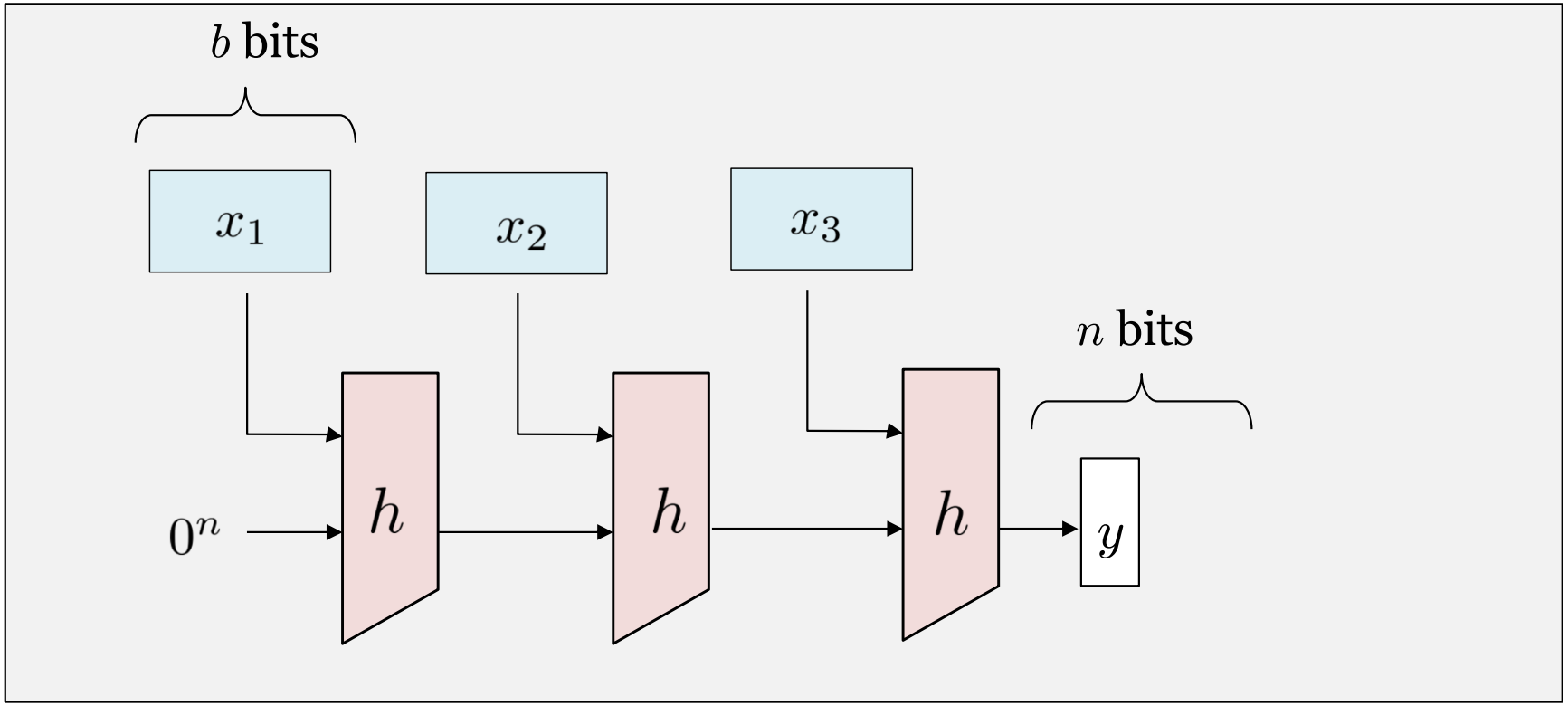
Compression Functions

$$h : \{0,1\}^{b+n} \rightarrow \{0,1\}^n$$



For SHA-2, b = 512 and n = 256

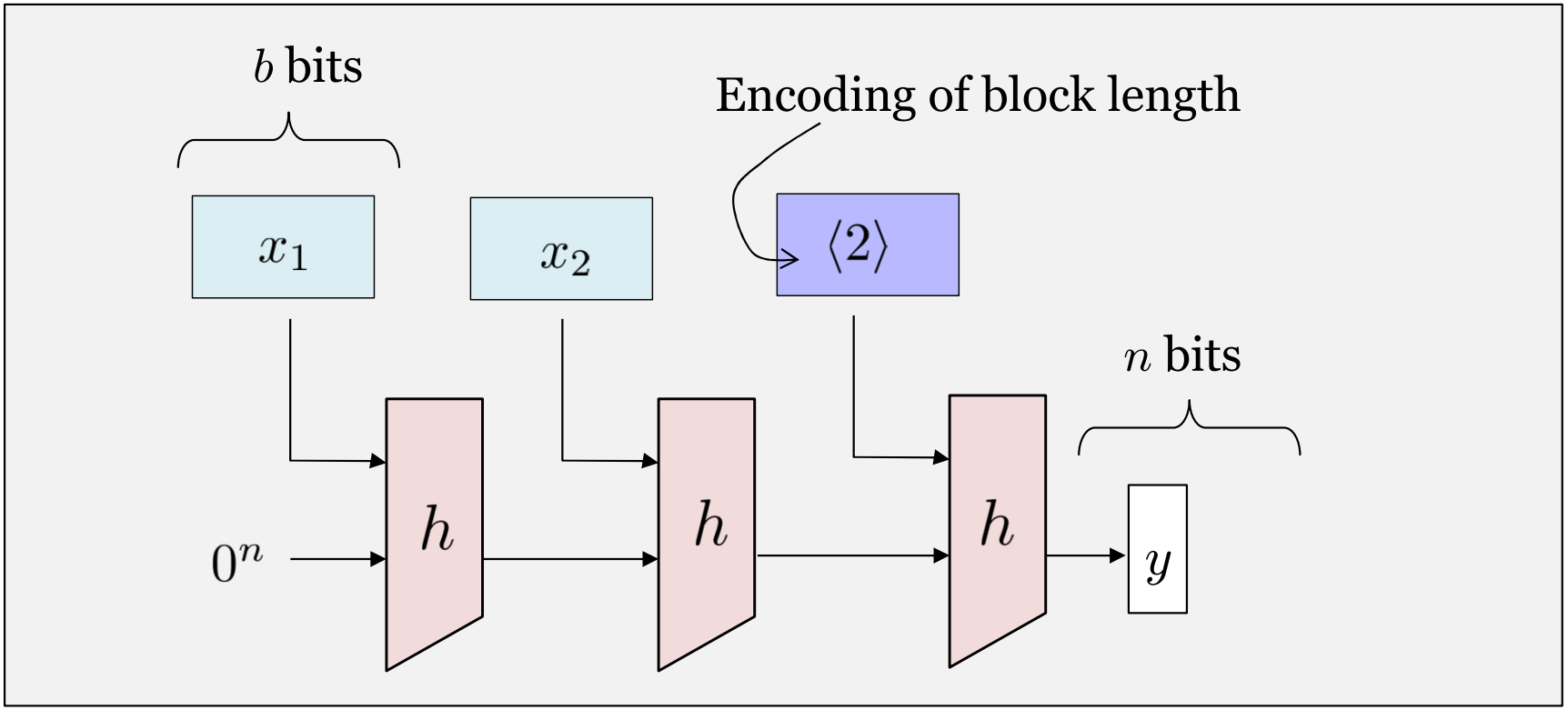
First Attempt



Question: Suppose that $h(0^b \| 0^n) = 0^n$

Break the collision resistance of H

Second Attempt: Plain Merkle-Damgard



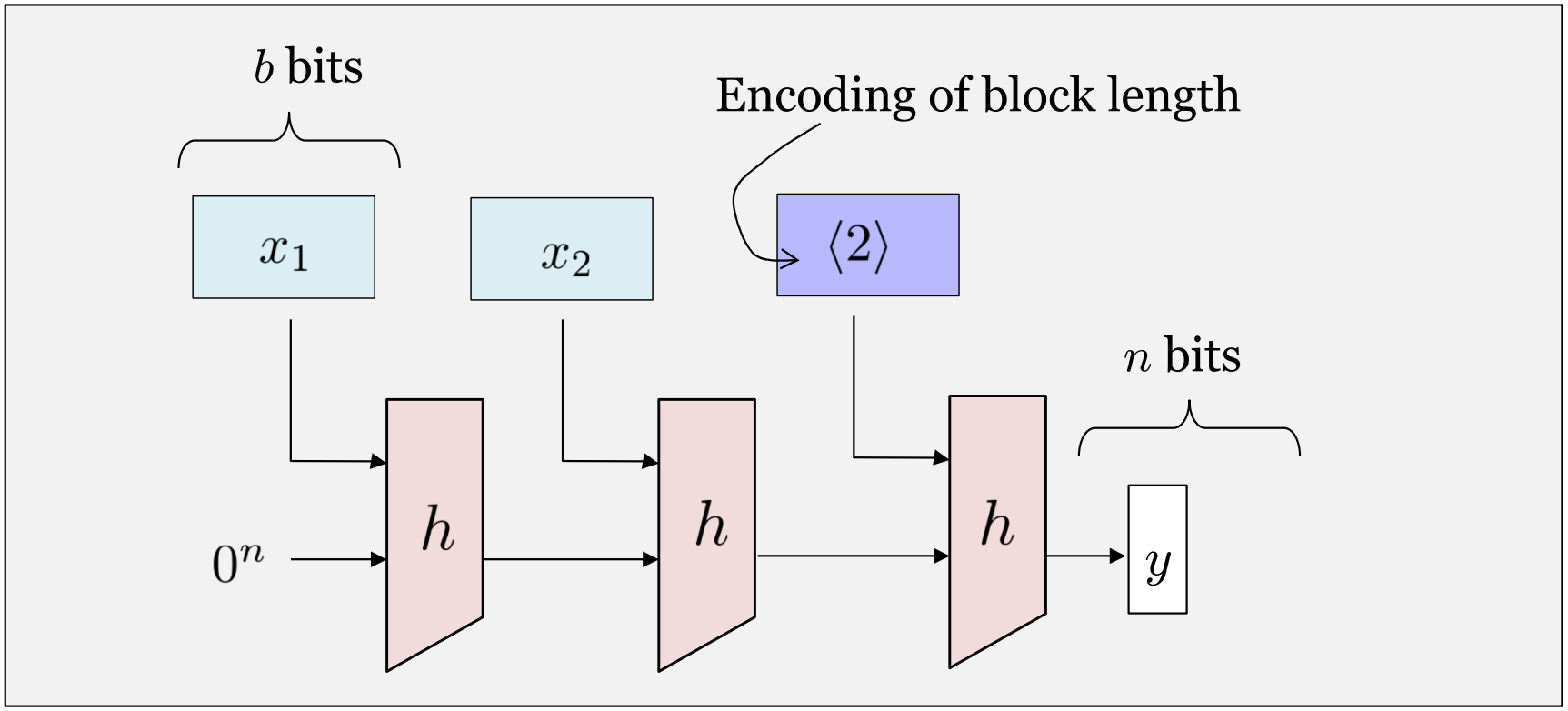
This is the structure of SHA-256

Theorem: If h is CR then $H = \mathbf{MD}(h)$ is also CR

Can't attack H if h has no weakness

Plain MD Is Not Enough for All Applications

Length-Extension Attack



Question: Consider the following MAC F

$$F_K(x) = H(K \| x)$$

Break the MAC security of F using a single Tag query

The Damage of Length Extension Attack

Hacking Trick: Bypass Authentication

K

bank.com

K

bank.com/api?token=ad6613c382&user=alice&cmd=NoOp

$H(K \| \text{“user=alice\&cmd=NoOp”})$

Adversary tricks Alice to perform a harmless command to learn an authentication token

The Damage of Length Extension Attack

Hacking Trick: Bypass Authentication

K

bank.com

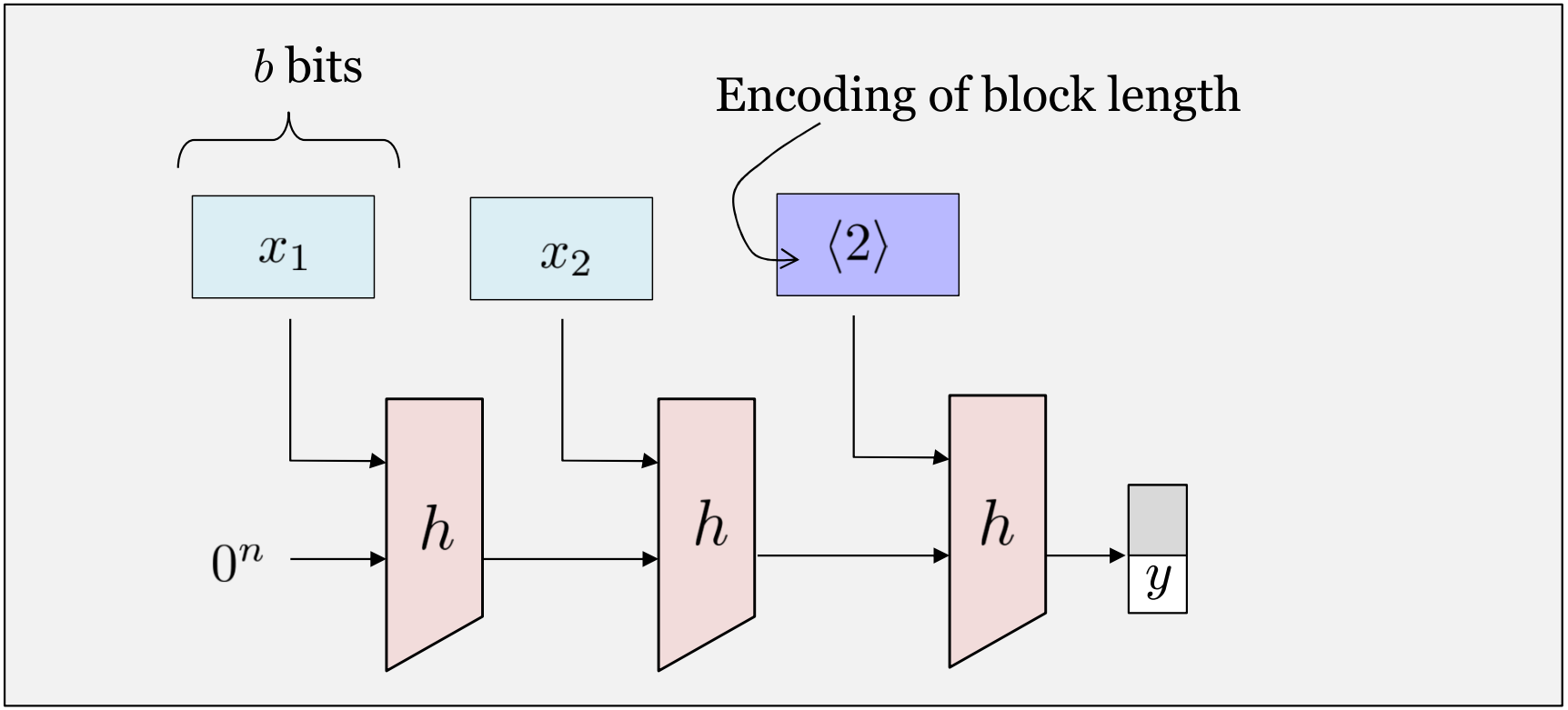
K

bank.com/api?token=dbb78b593f&user=alice&cmd=NoOp&cmd=OpenSafe

$H(K \| \text{“user=alice\&cmd=NoOp\&cmd=OpenSafe”})$

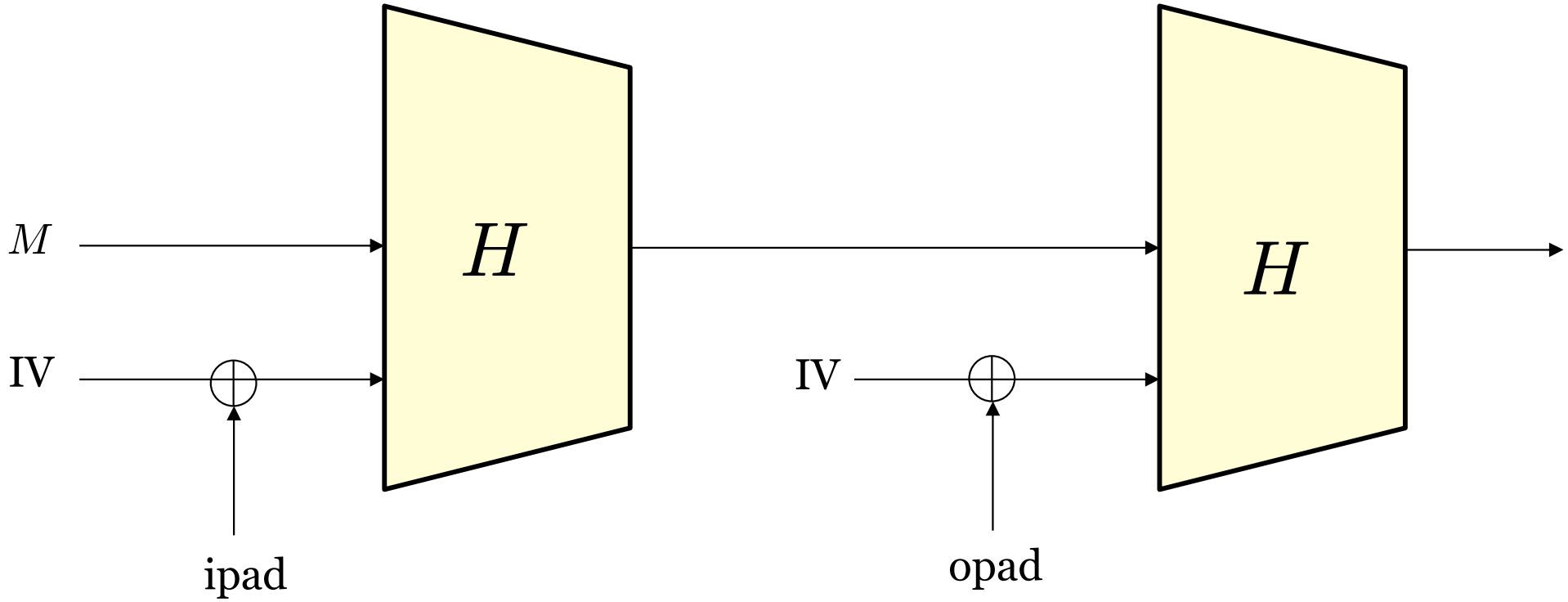
Adversary can compute the authentication token for a damaging command

The (Strengthened) MD Transform



The output needs to be truncated

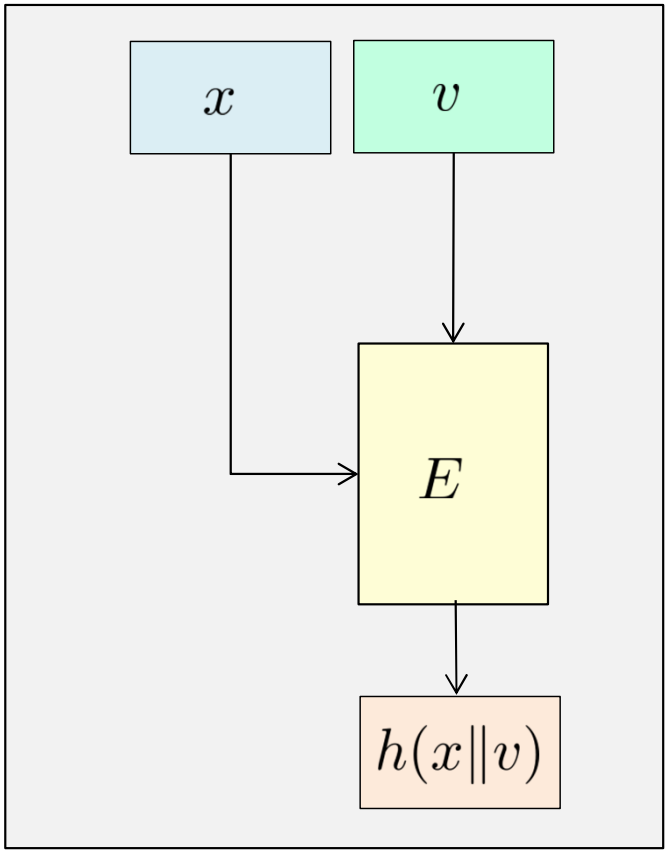
How To Have Large Output: HMAC



On large input, HMAC is only a bit more expensive than SHA-256

Designing Compression Functions

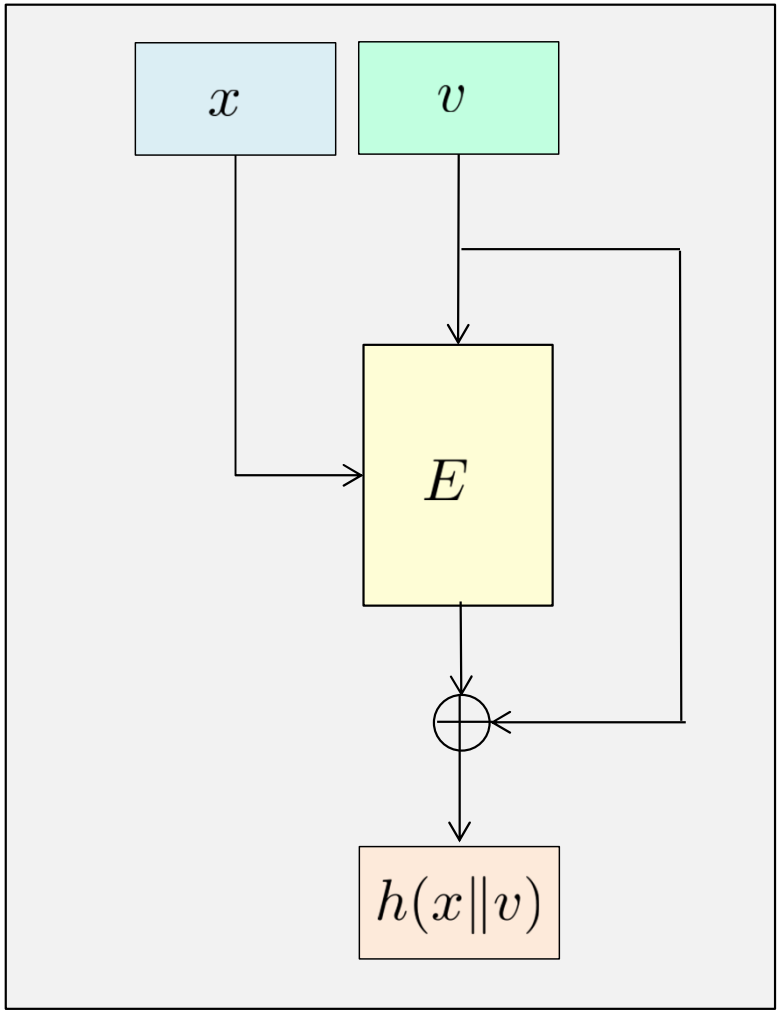
Often designed from a blockcipher $E : \{0,1\}^b \times \{0,1\}^n \to \{0,1\}^n$



An **incorrect** design $h(x\|v) = E_x(v)$

Question: Break the collision resistance of h

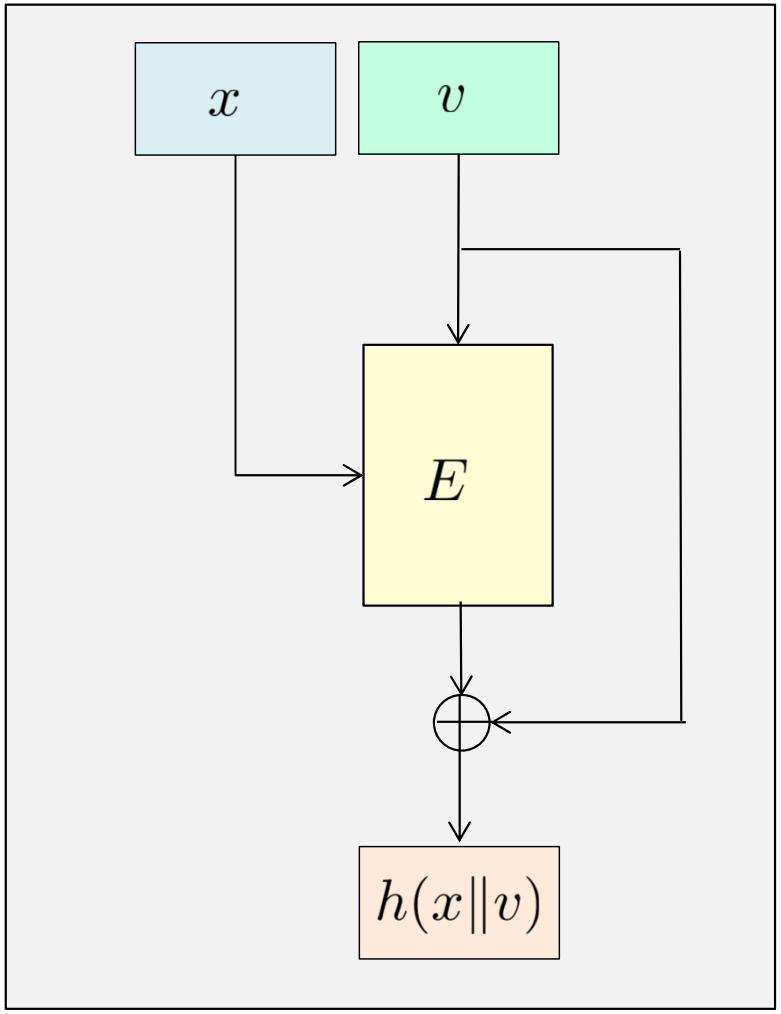
The Davies-Meyer Construction



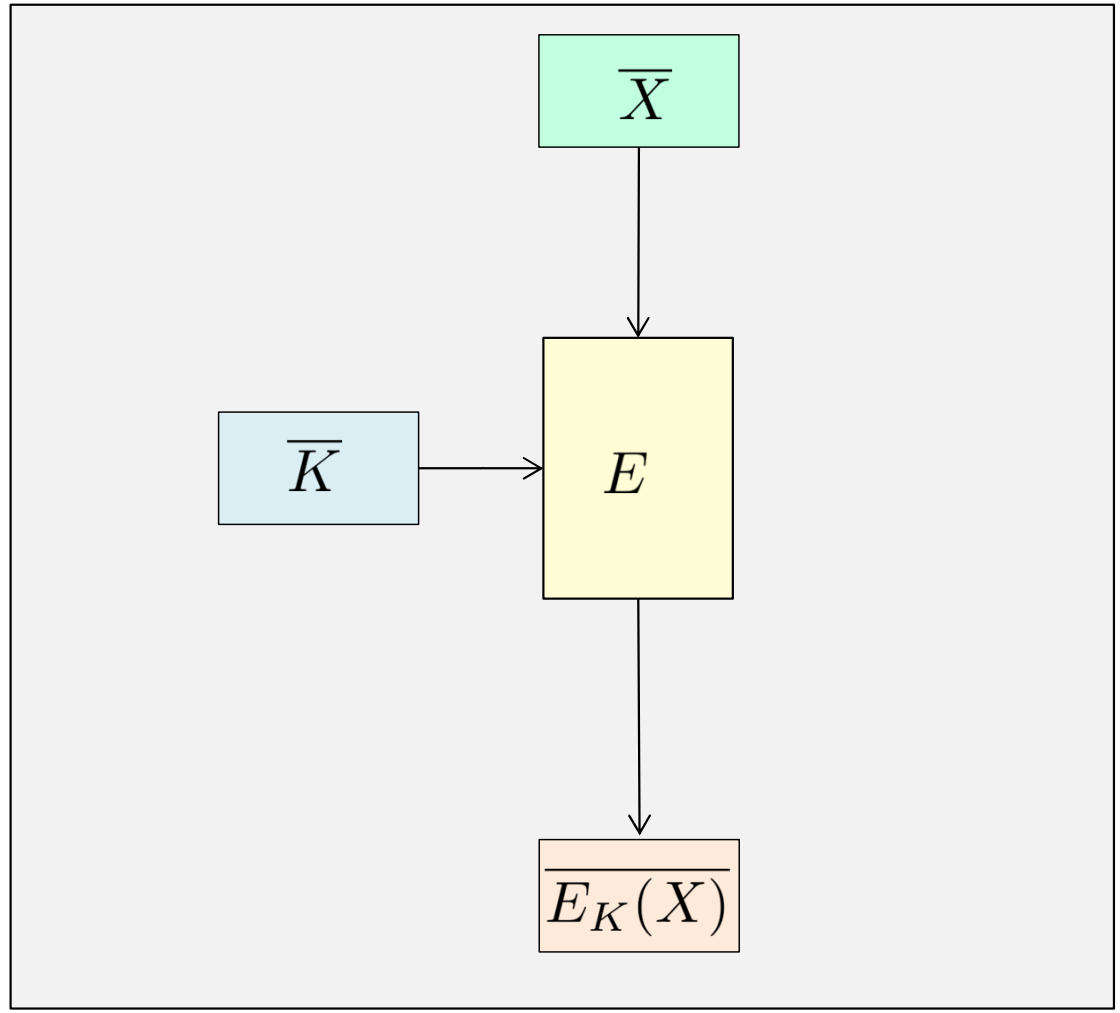
Work if we model E as **ideal**

$$h(x\|v) = E_x(v) \oplus v$$

Exercise: Breaking Davies-Meyer on DES



Davies-Meyer



A property of E that is true for DES

Agenda

1. Security Modeling for Hash Functions

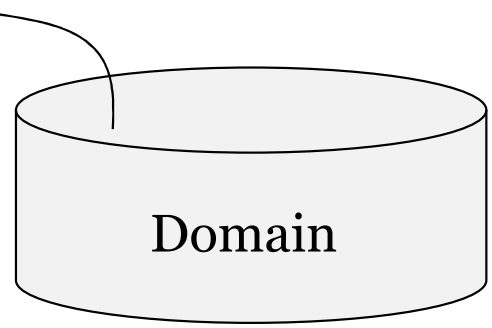
2. Building Hash Function: MD Transform

3. **Generic Attacks**

4. Applications

Birthday Attack

distinct $x_1, \ldots, x_q$



$H(x_1)$ … $H(x_q)$

Look for duplicate values

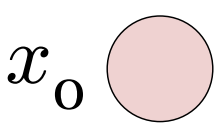
If H is modeled as a random oracle

$$\Pr[\text{collision}] \approx \frac{q^2}{2^n}$$

Output length

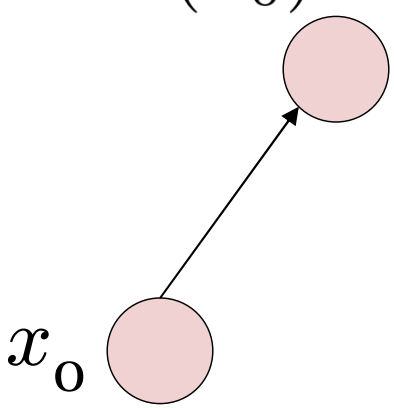
Problem: Use $O(q)$ space

Rho Attack: Start Arbitrarily



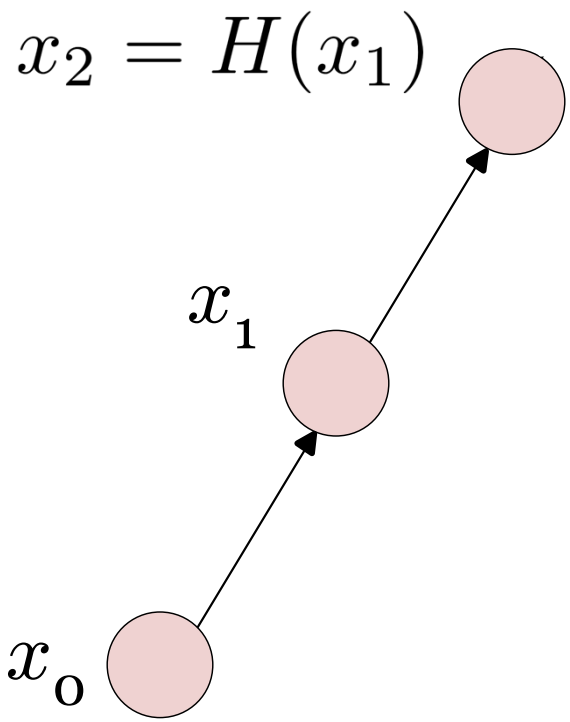
Rho Attack

$$x_1 = H(x_0)$$

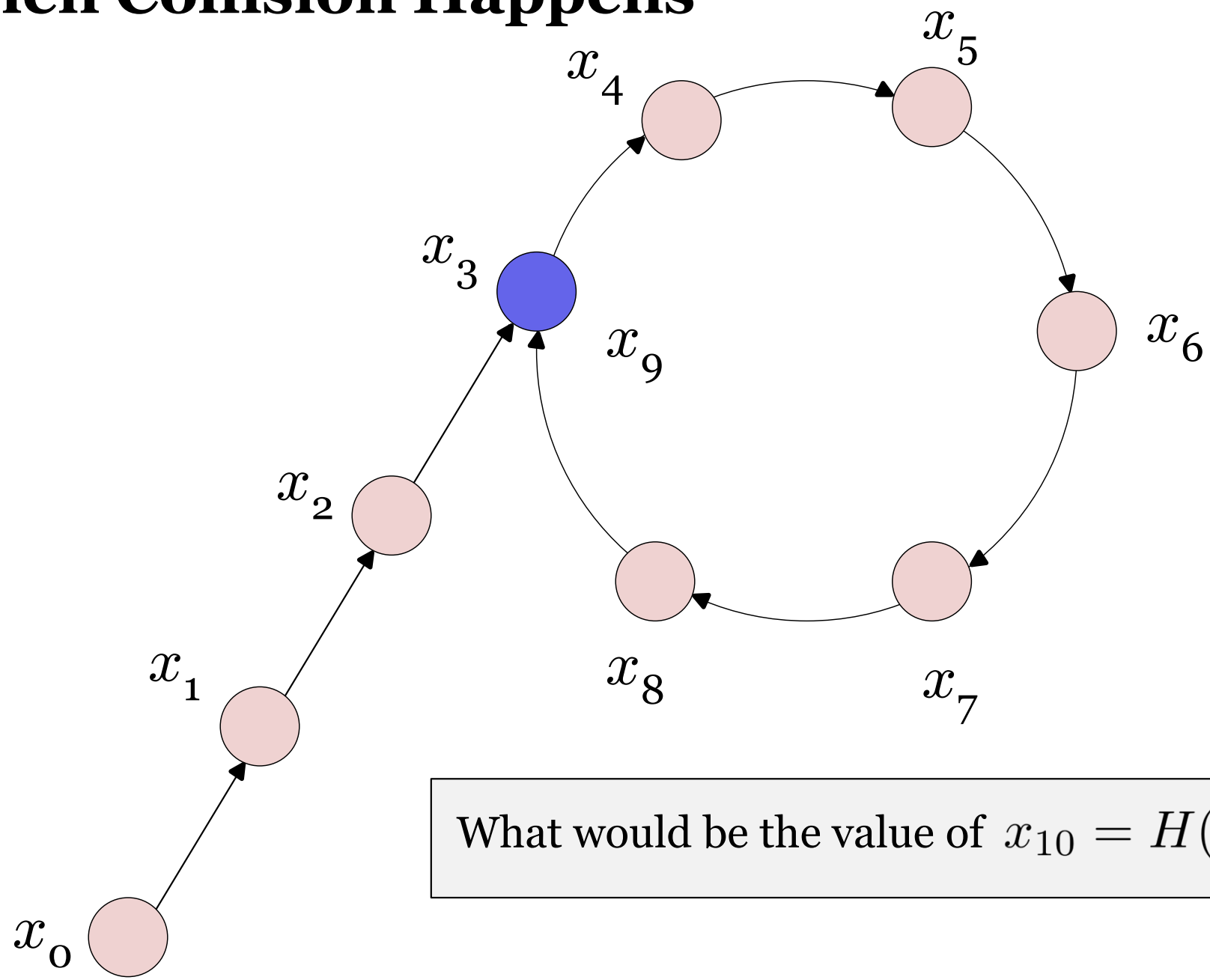


Rho Attack

Sooner or later a collision must happen

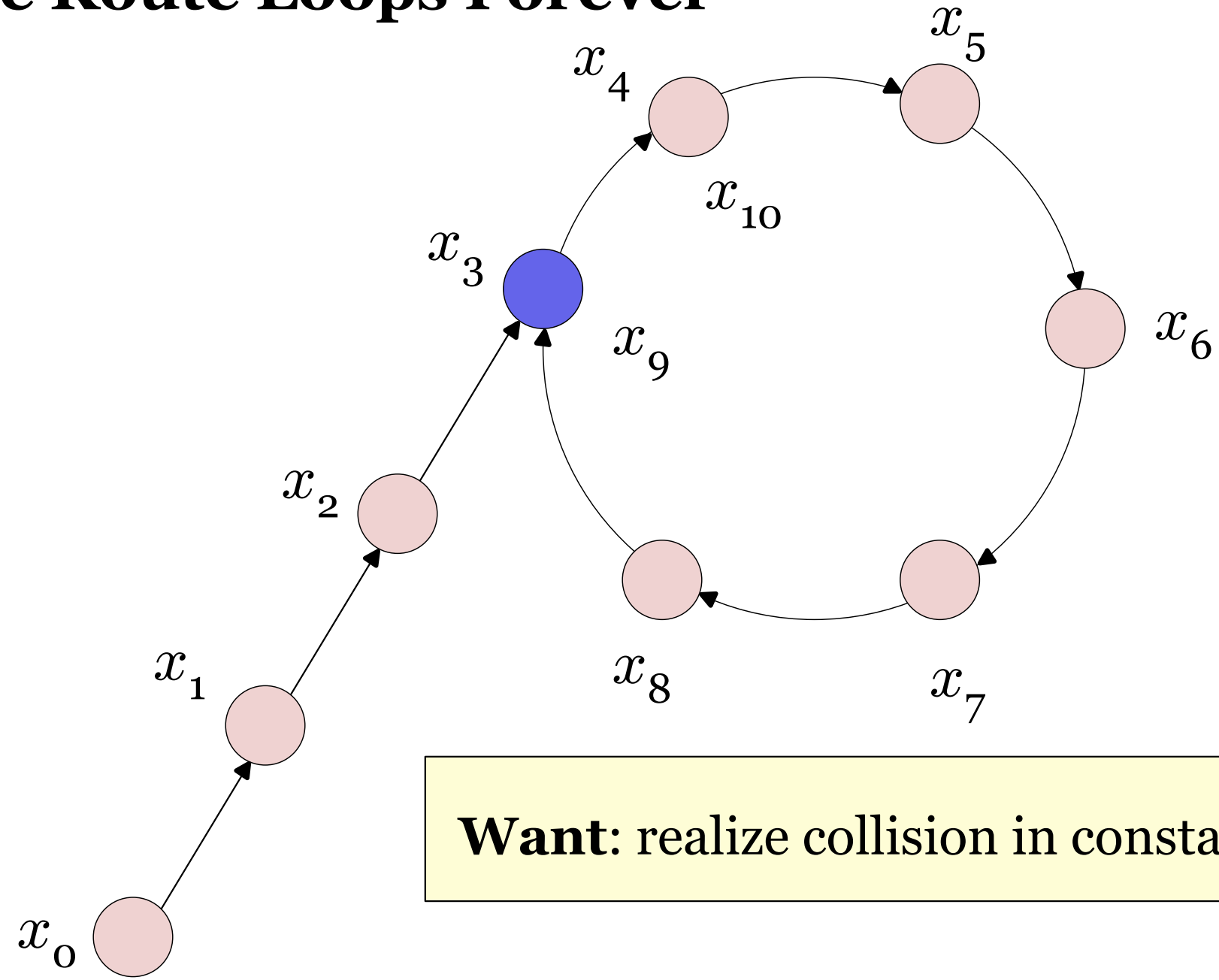


When Collision Happens



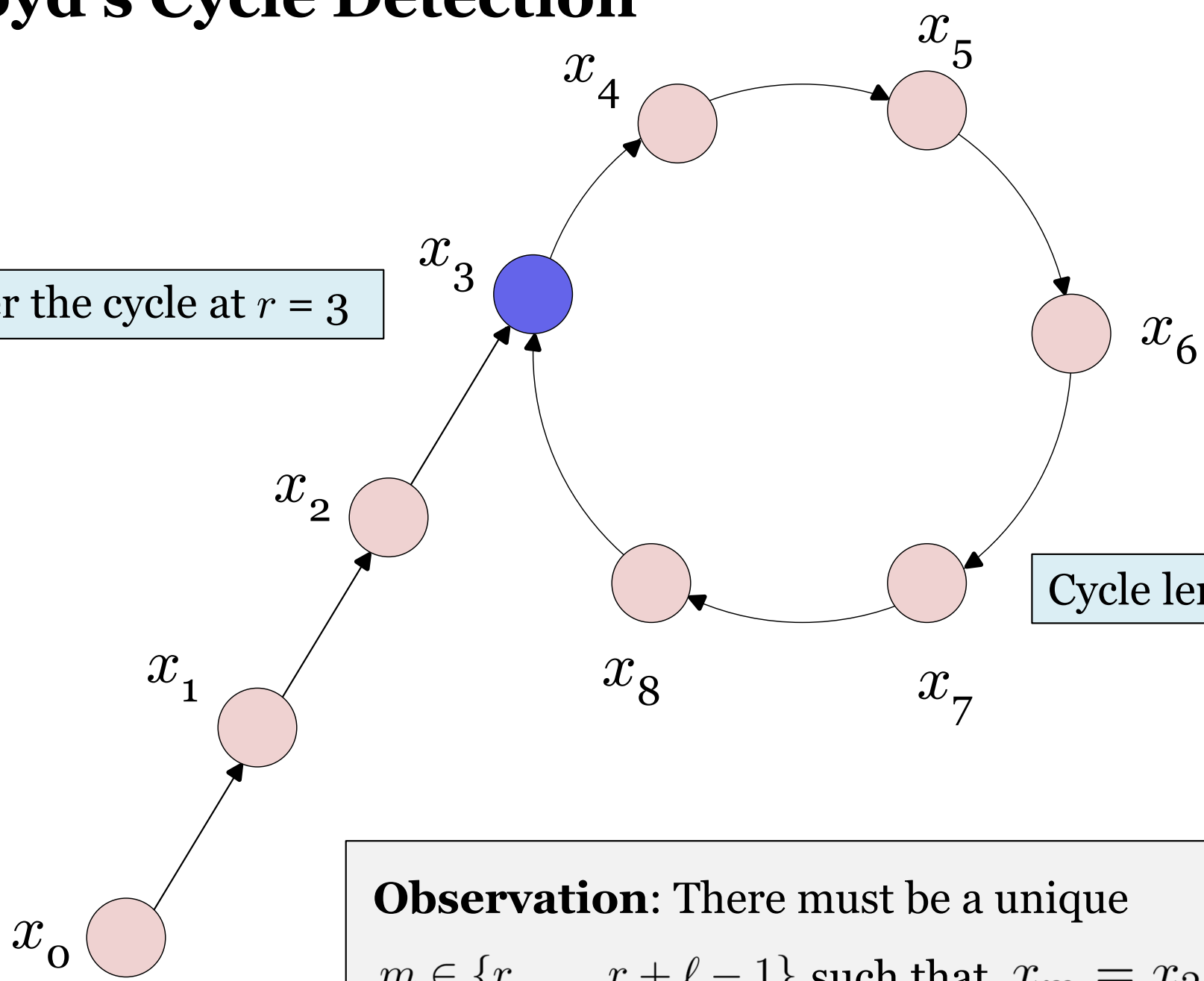
What would be the value of $x_{10} = H(x_9)$?

The Route Loops Forever



Want: realize collision in constant space

Floyd's Cycle Detection

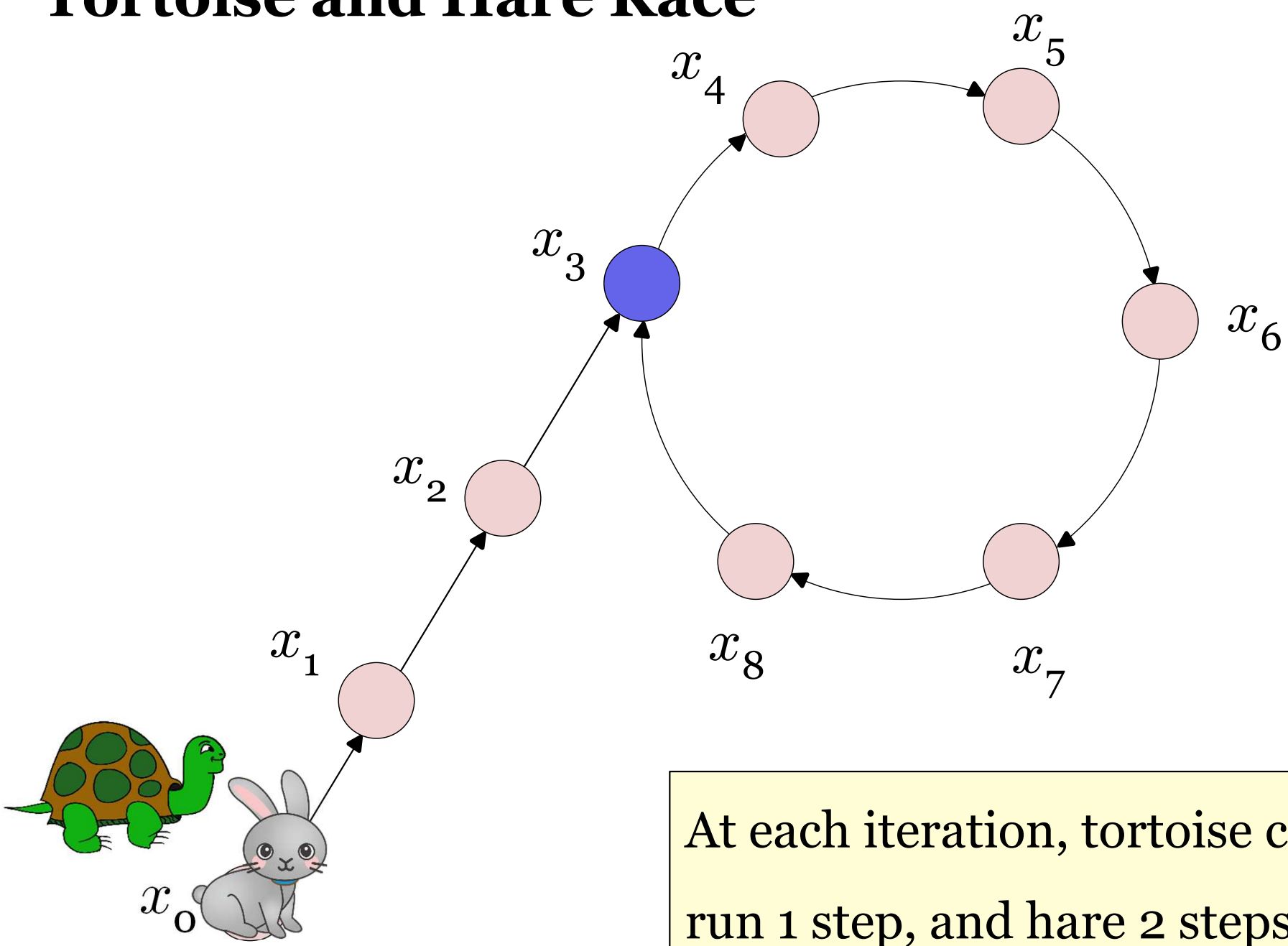


Enter the cycle at $r = 3$

Cycle length $\ell = 6$

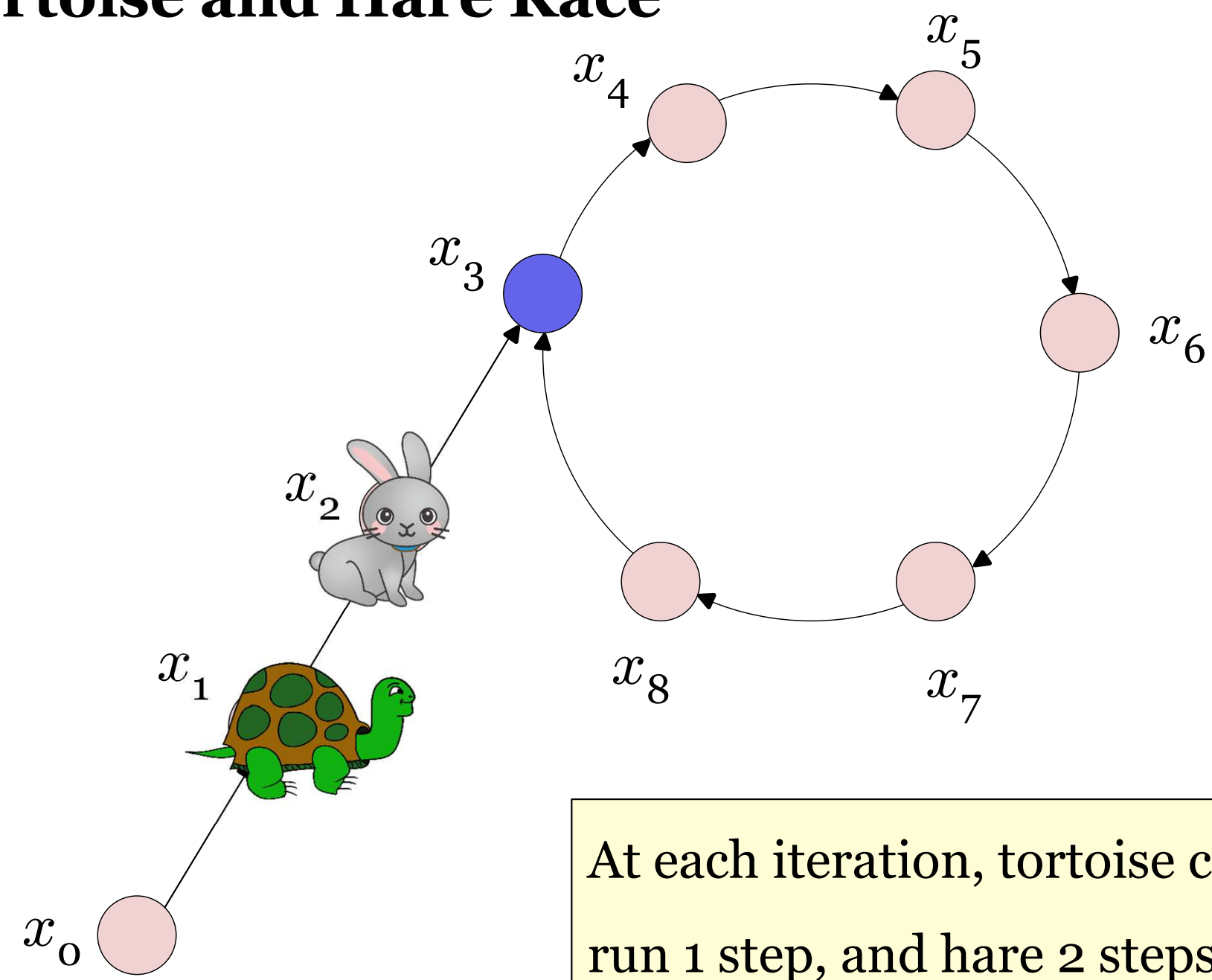
Observation: There must be a unique $m \in \{r, \ldots, r + \ell - 1\}$ such that $x_m = x_{2m}$

Tortoise and Hare Race



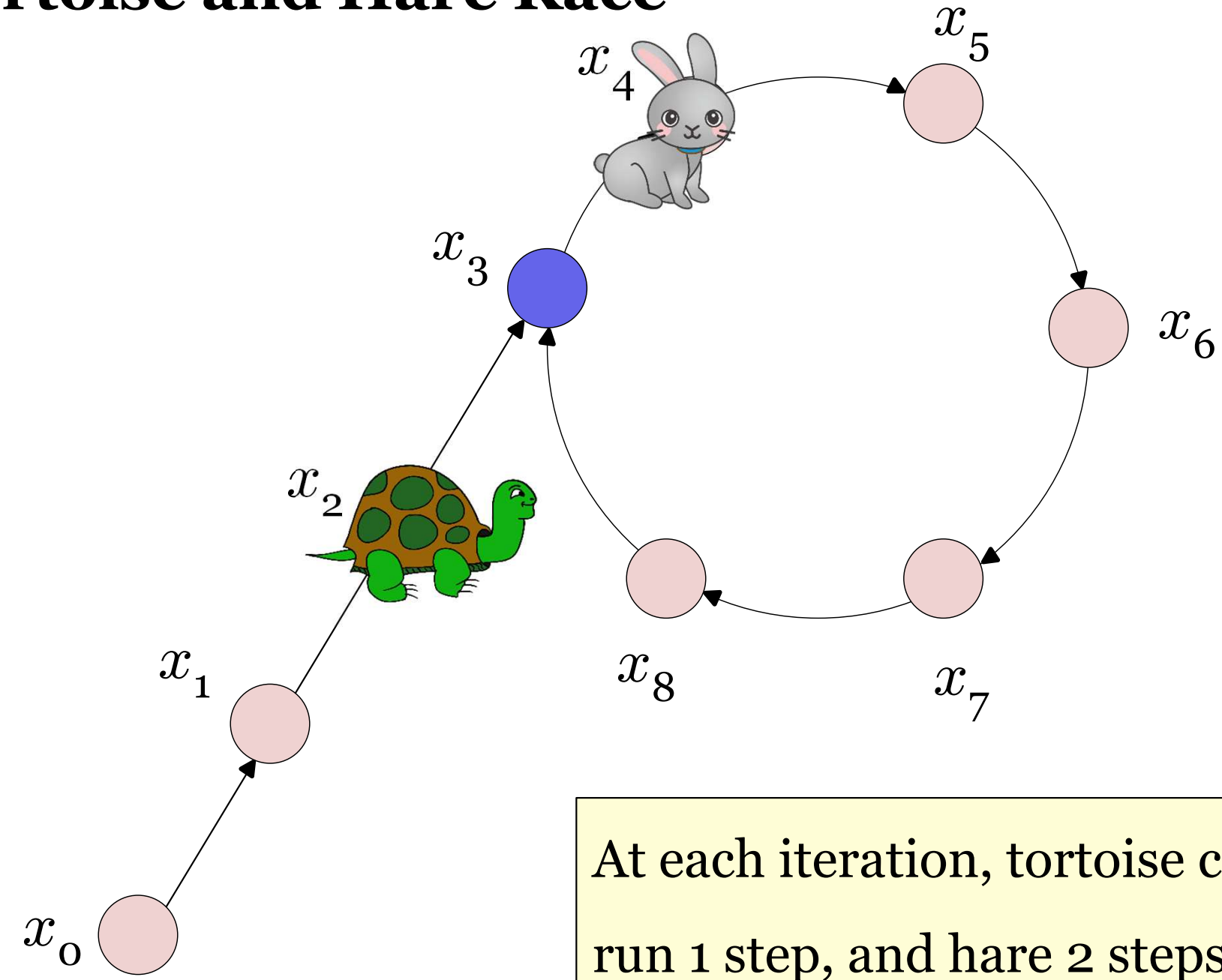
At each iteration, tortoise can run 1 step, and hare 2 steps

Tortoise and Hare Race



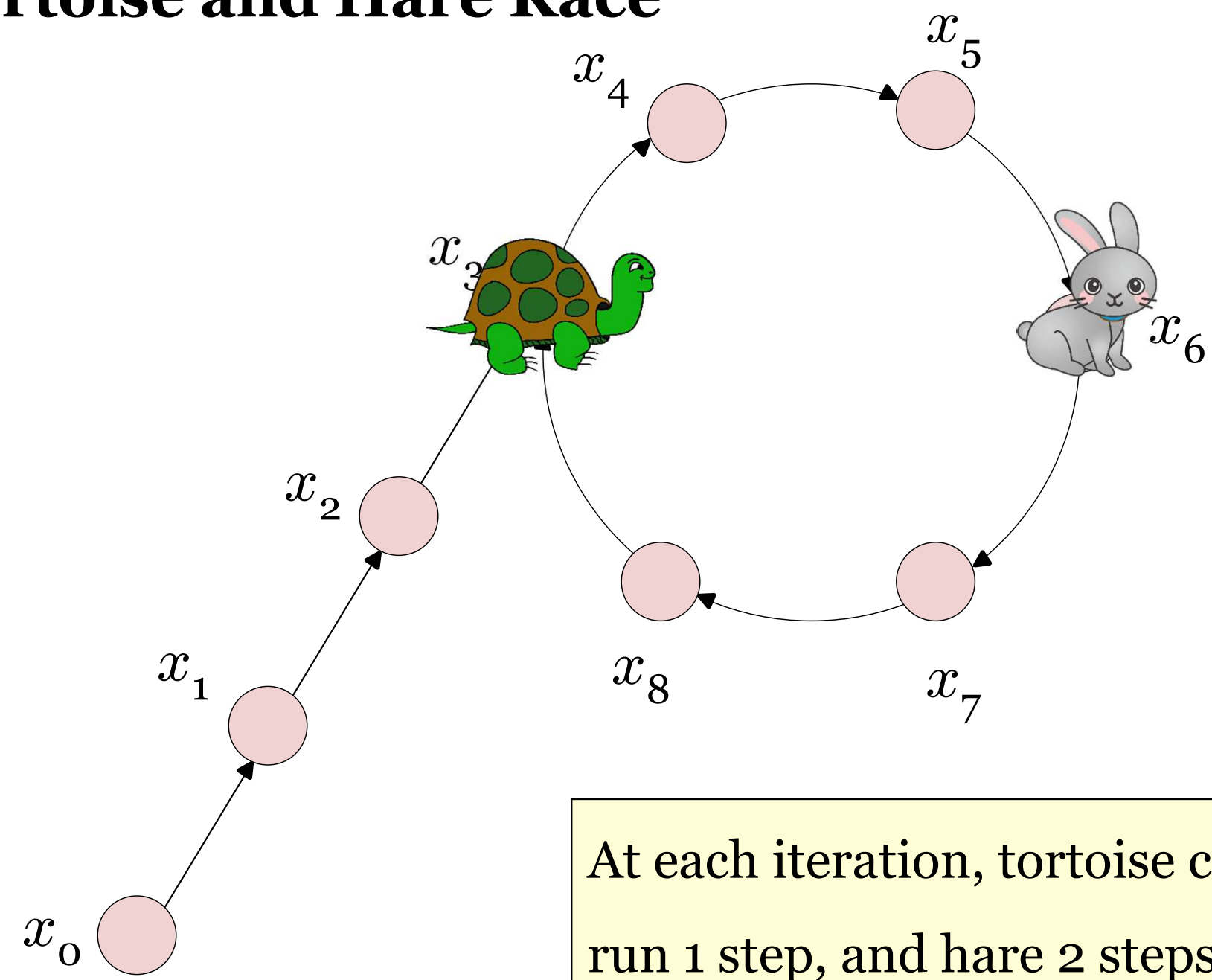
At each iteration, tortoise can run 1 step, and hare 2 steps

Tortoise and Hare Race



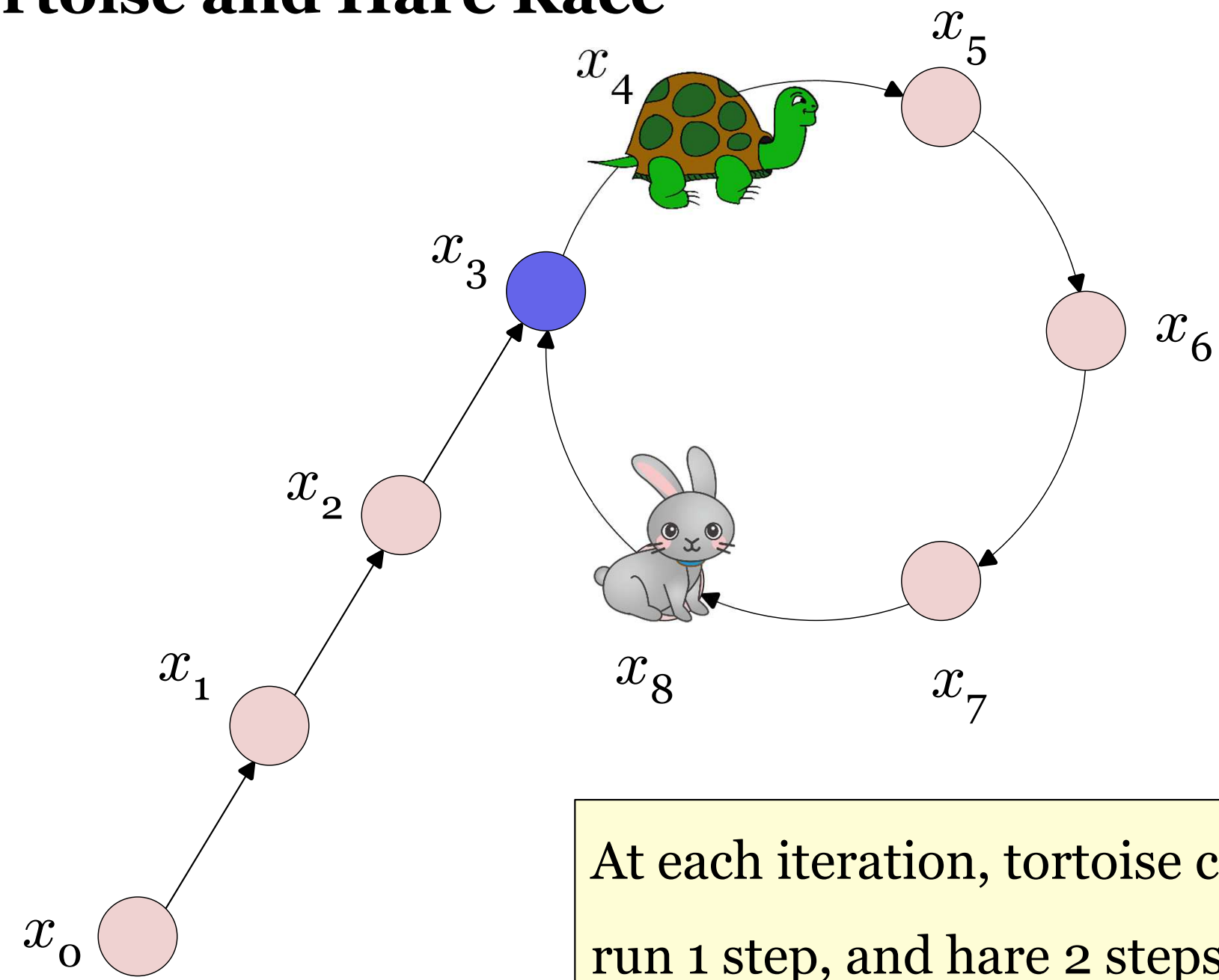
At each iteration, tortoise can run 1 step, and hare 2 steps

Tortoise and Hare Race



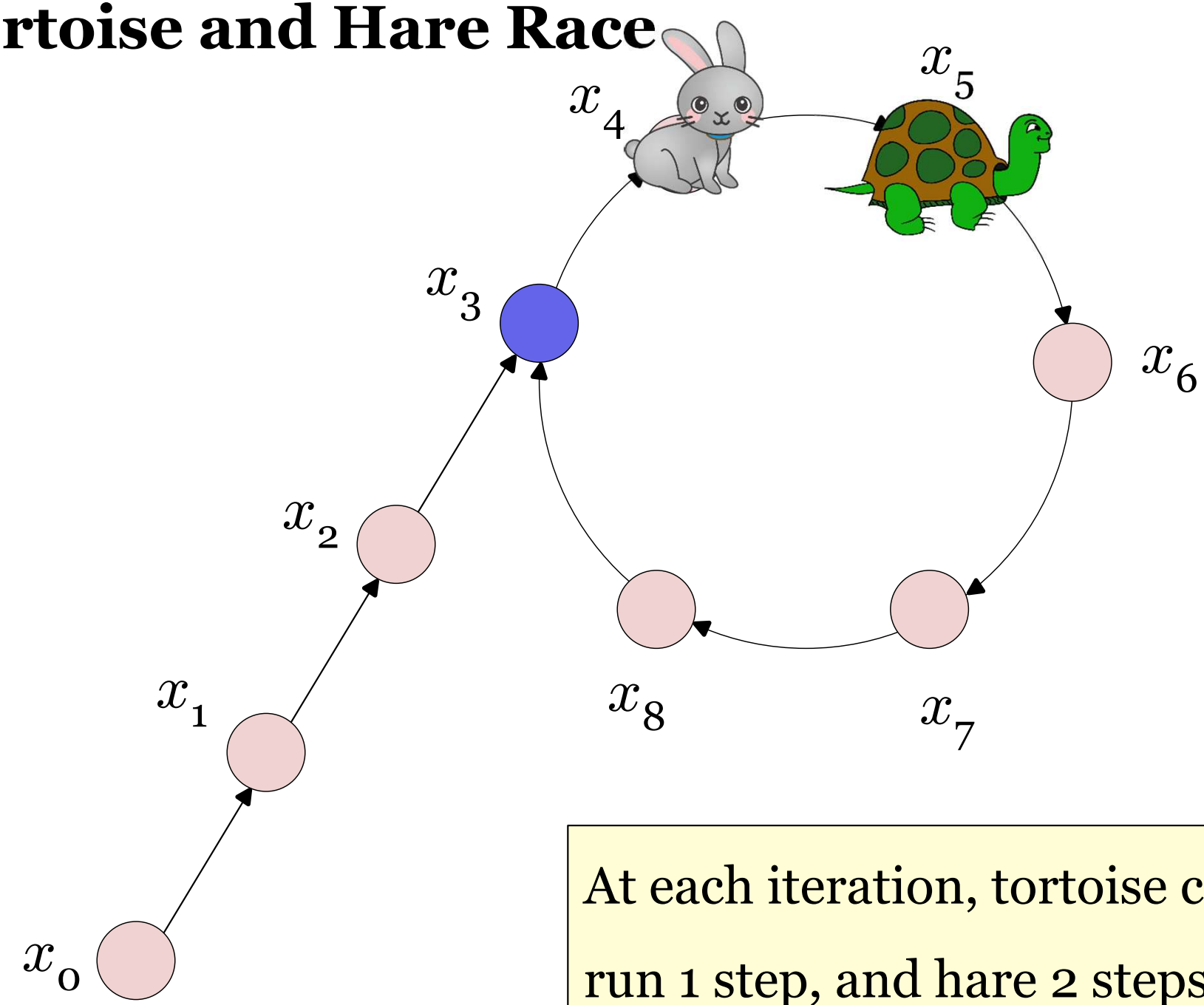
At each iteration, tortoise can run 1 step, and hare 2 steps

Tortoise and Hare Race



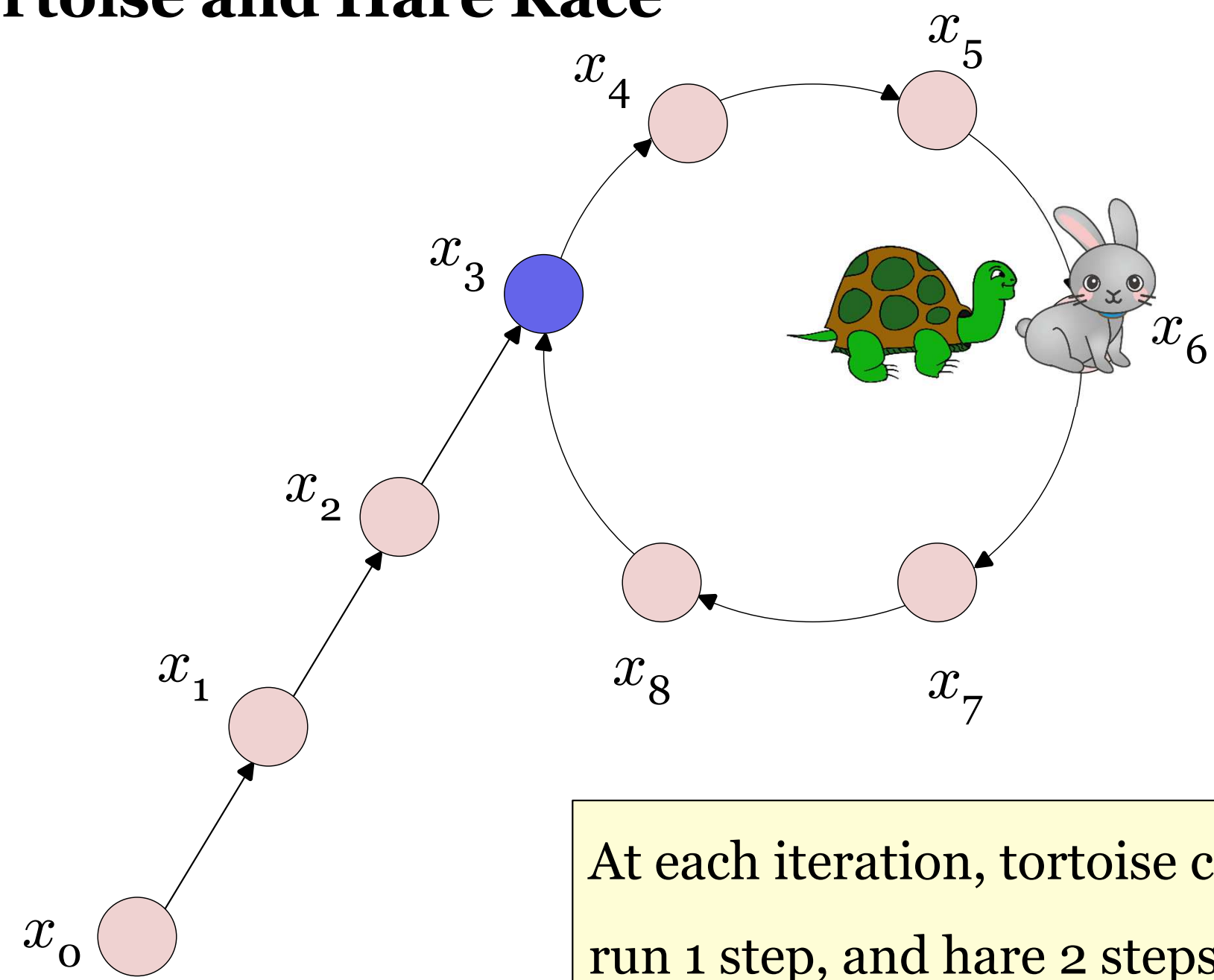
At each iteration, tortoise can run 1 step, and hare 2 steps

Tortoise and Hare Race



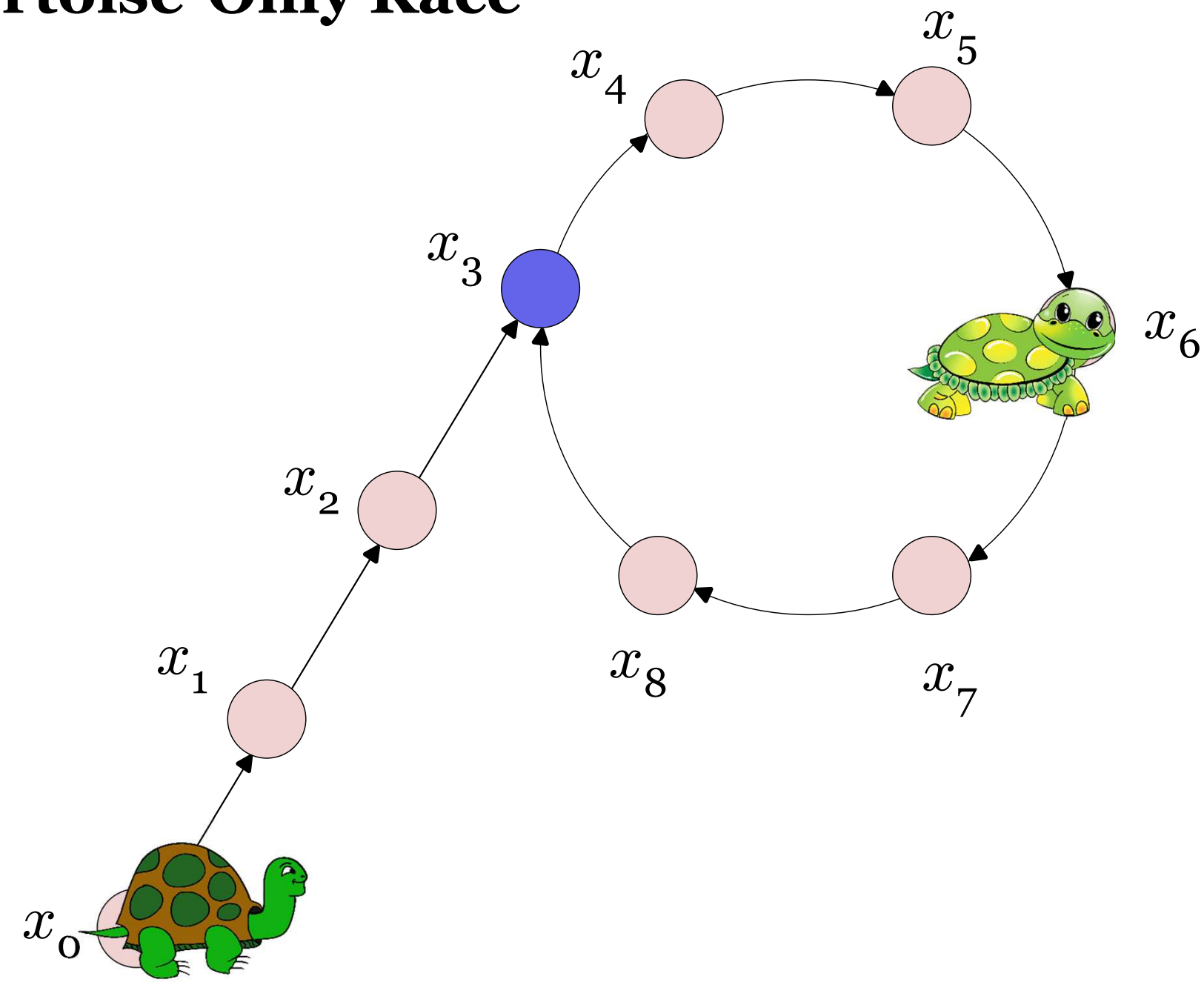
At each iteration, tortoise can run 1 step, and hare 2 steps

Tortoise and Hare Race

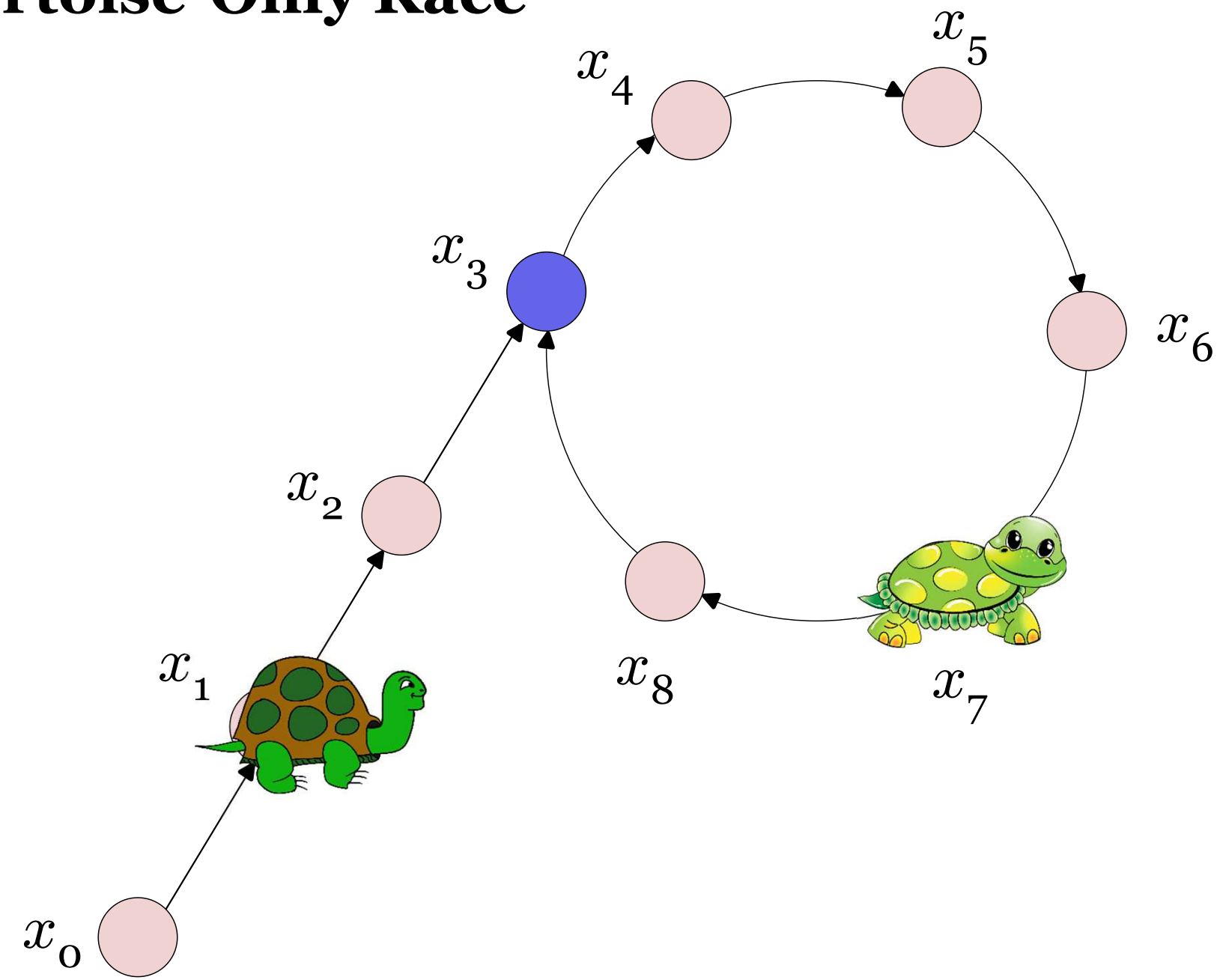


At each iteration, tortoise can run 1 step, and hare 2 steps

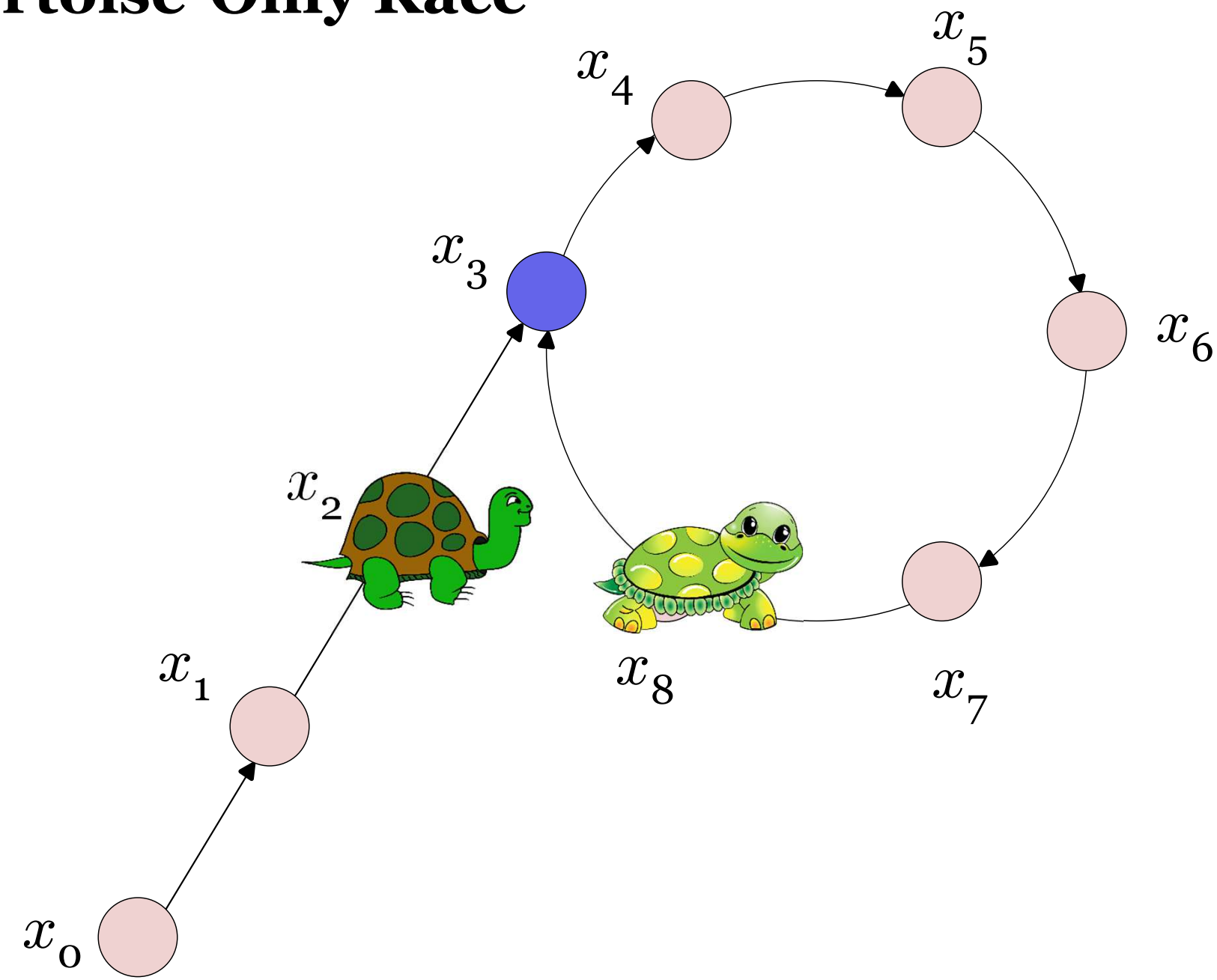
Tortoise-Only Race



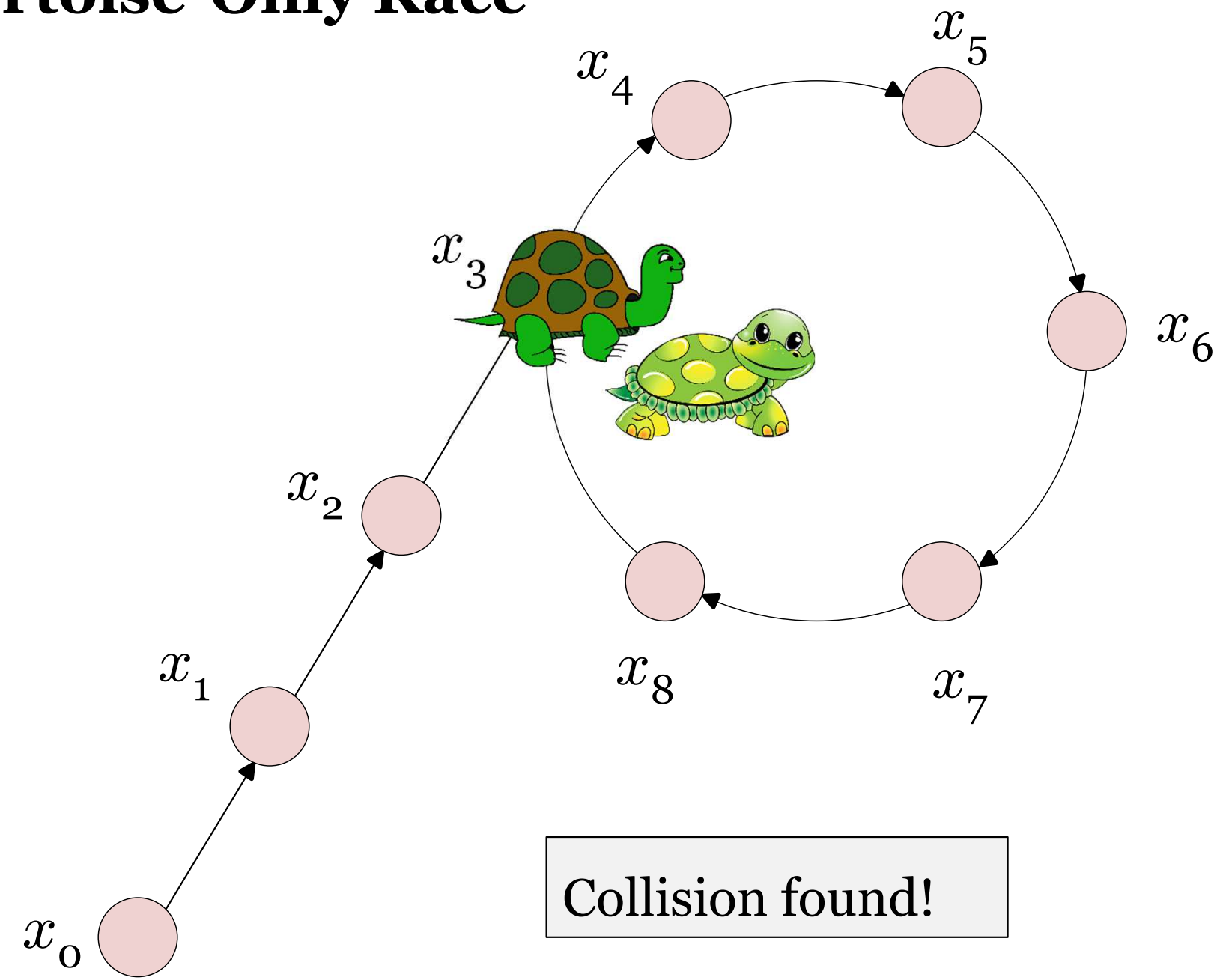
Tortoise-Only Race



Tortoise-Only Race



Tortoise-Only Race



Collision found!

Agenda

1. Security Modeling for Hash Functions

2. Building Hash Function: MD Transform

3. Generic Attacks

4. **Applications**

Application #1: Password Storage

MOTHERBOARD
TECH BY VICE

T-Mobile Stores Part of Customers' Passwords In Plaintext, Says It Has 'Amazingly Good' Security

A T-Mobile Austria customer represe
admission in a Twitter thread.

ars TECHNICA

BIZ & IT —

How an epic blunder by Adobe could strengthen hand of password crackers

Engineers flout universal taboo by *encrypting* 130 million pilfered passwords.

NEWS

Hackers crack more than 60% of breached LinkedIn passwords

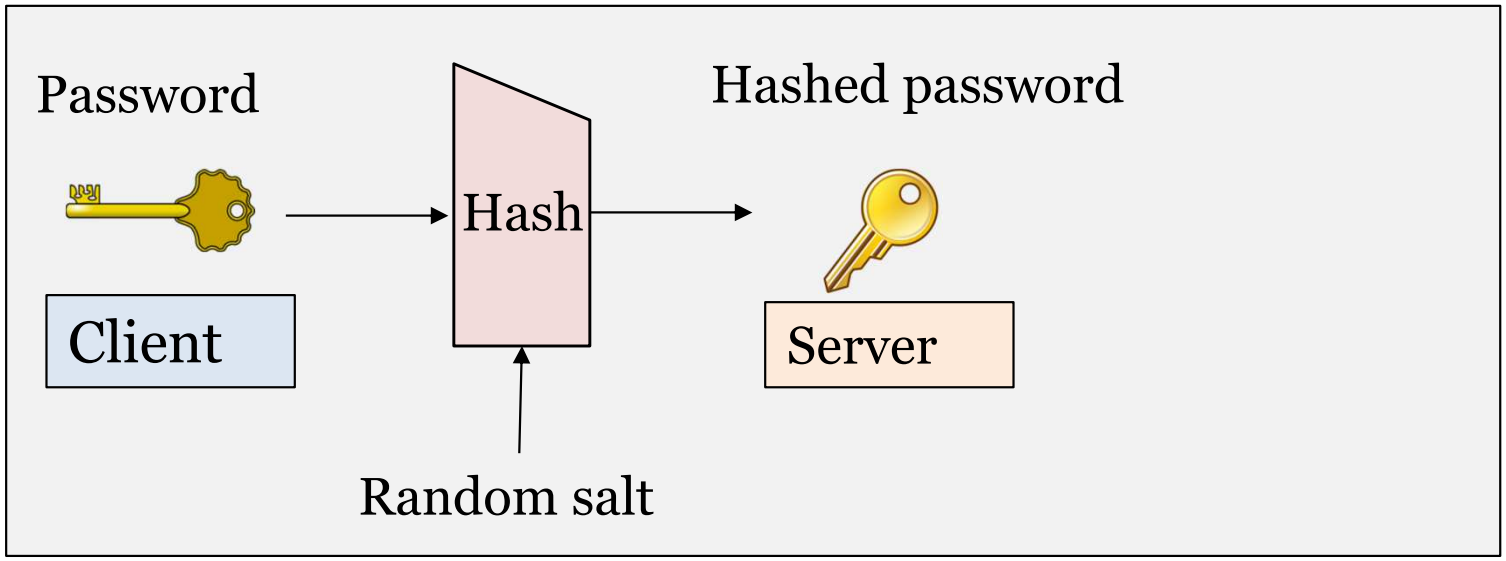
Speed of hackers to crack passwords shows weakness of security scheme used by LinkedIn, researchers say

How Should Servers Store Users' Passwords?

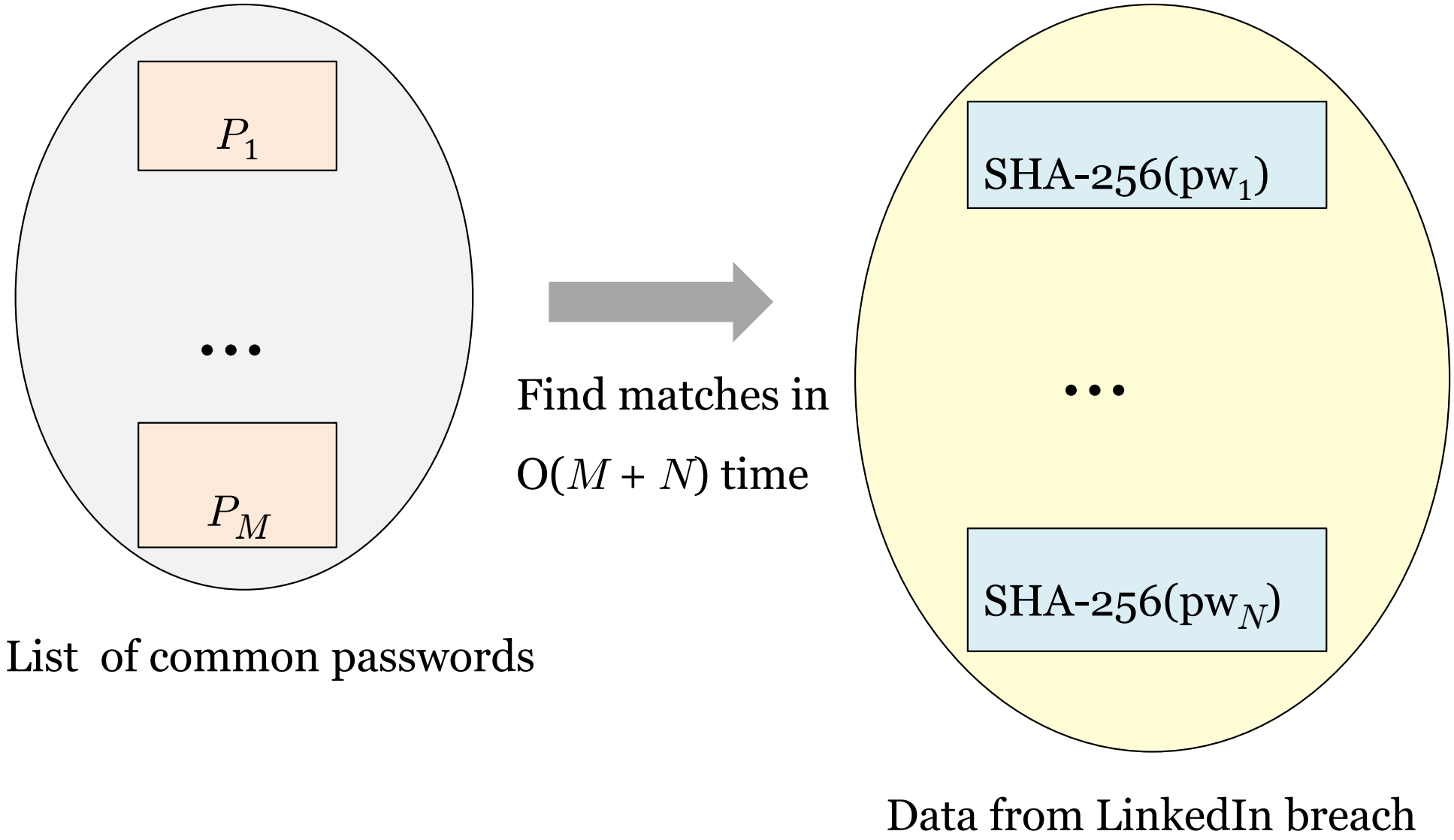
Rule 1: Only store hash outputs of passwords

Even server can't recover the passwords

Rule 2: Use a random salt for each user

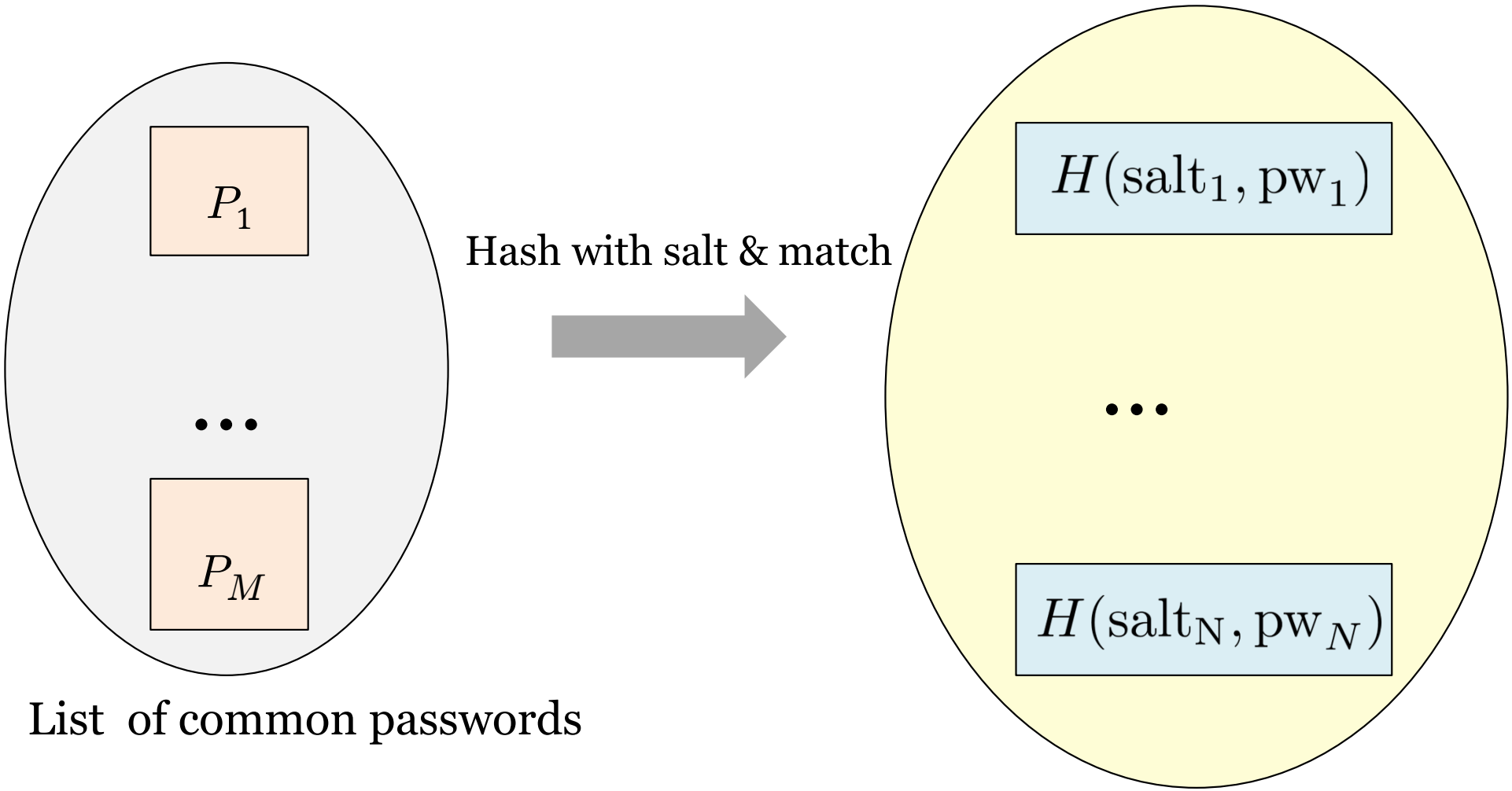


Why Salts: Dictionary Attacks



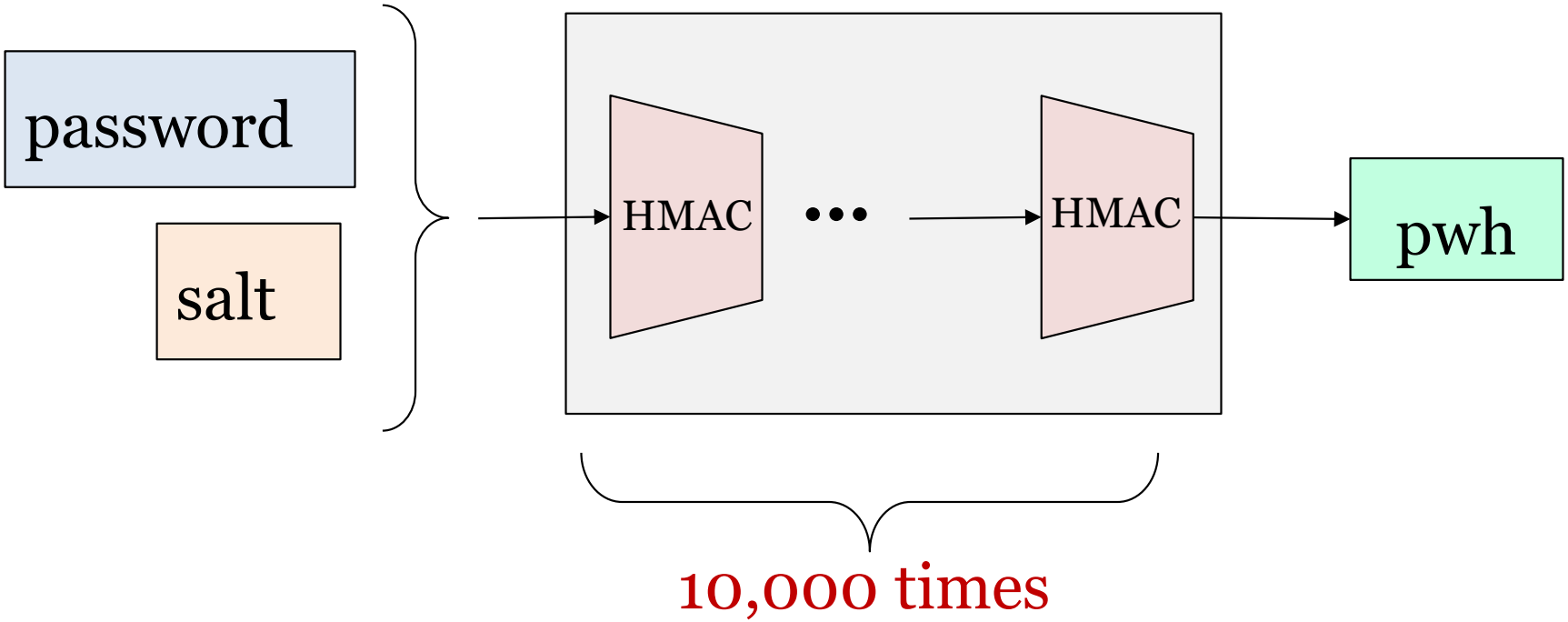
Cost of Dictionary Attacks on Salting

Need $\Theta(Mq)$ calls to H to recover q passwords



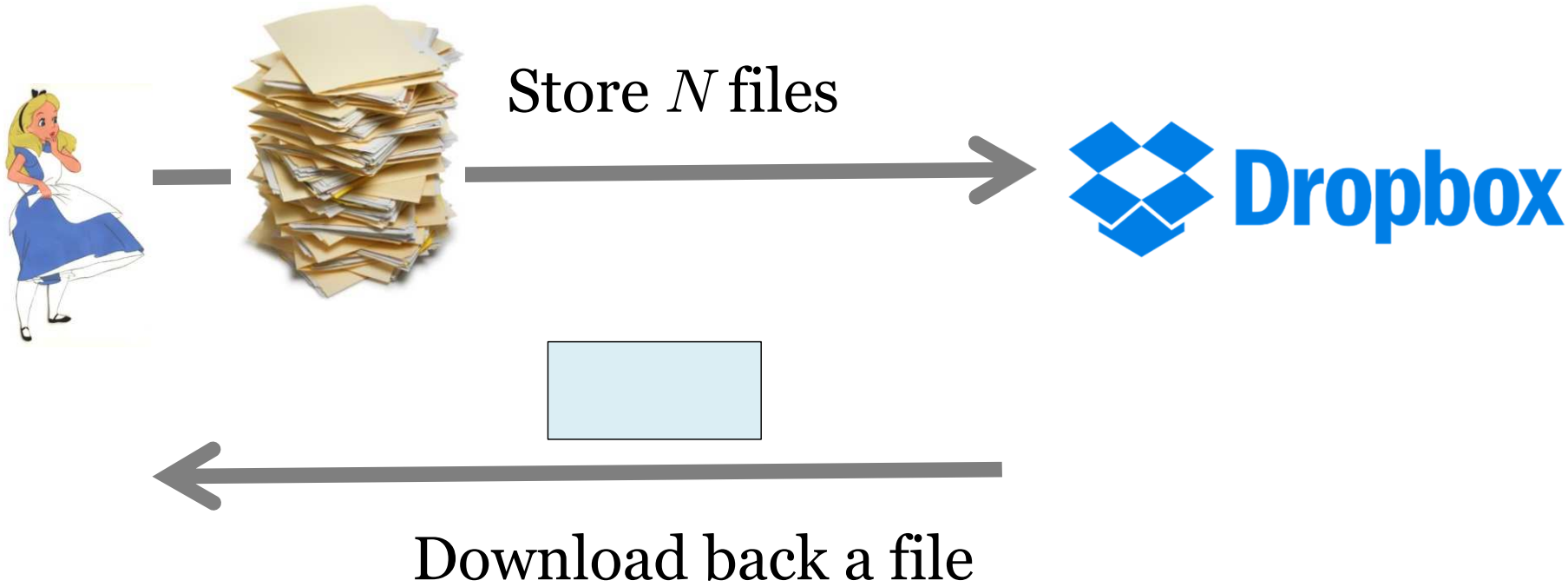
Make It Even More Expensive

Deliberately Slow Hashing



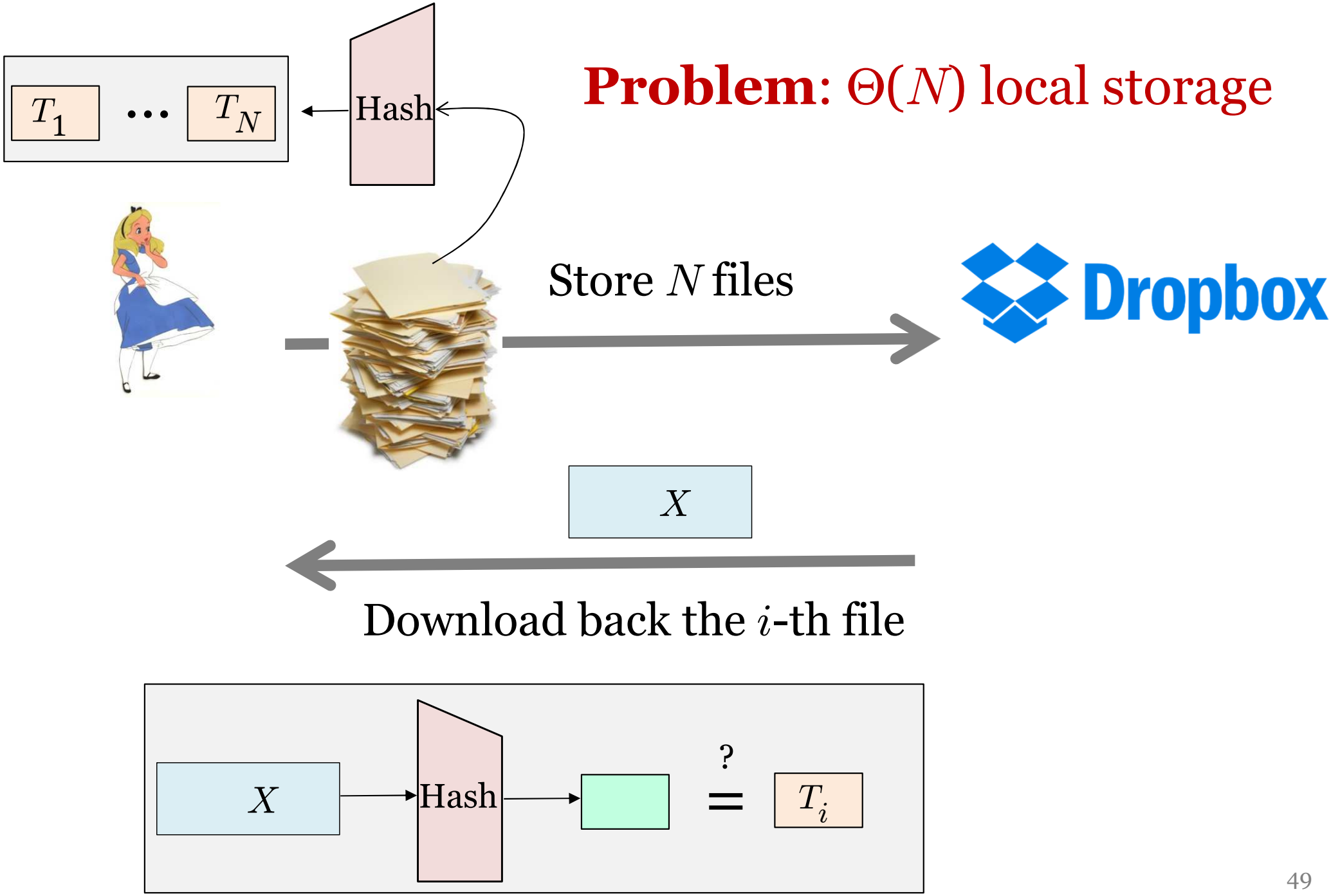
- Makes no difference for human users.
- Increase the cos of attackers for 10,000 times

Application #2: Verifiable Data Storage



Goal: Check if the file was modified

Naïve Solution

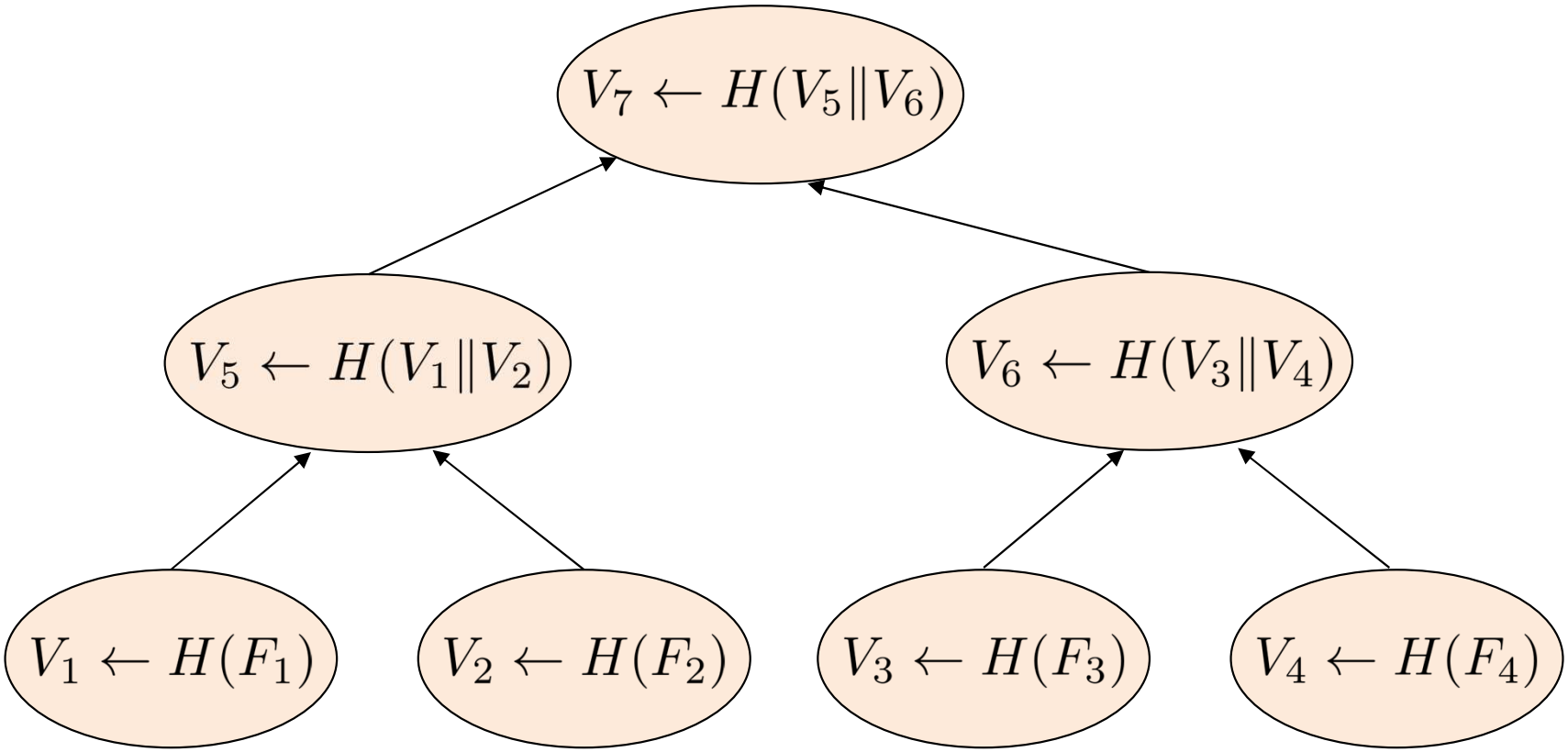


Problem: $\Theta(N)$ local storage

Solution: Merkle Tree

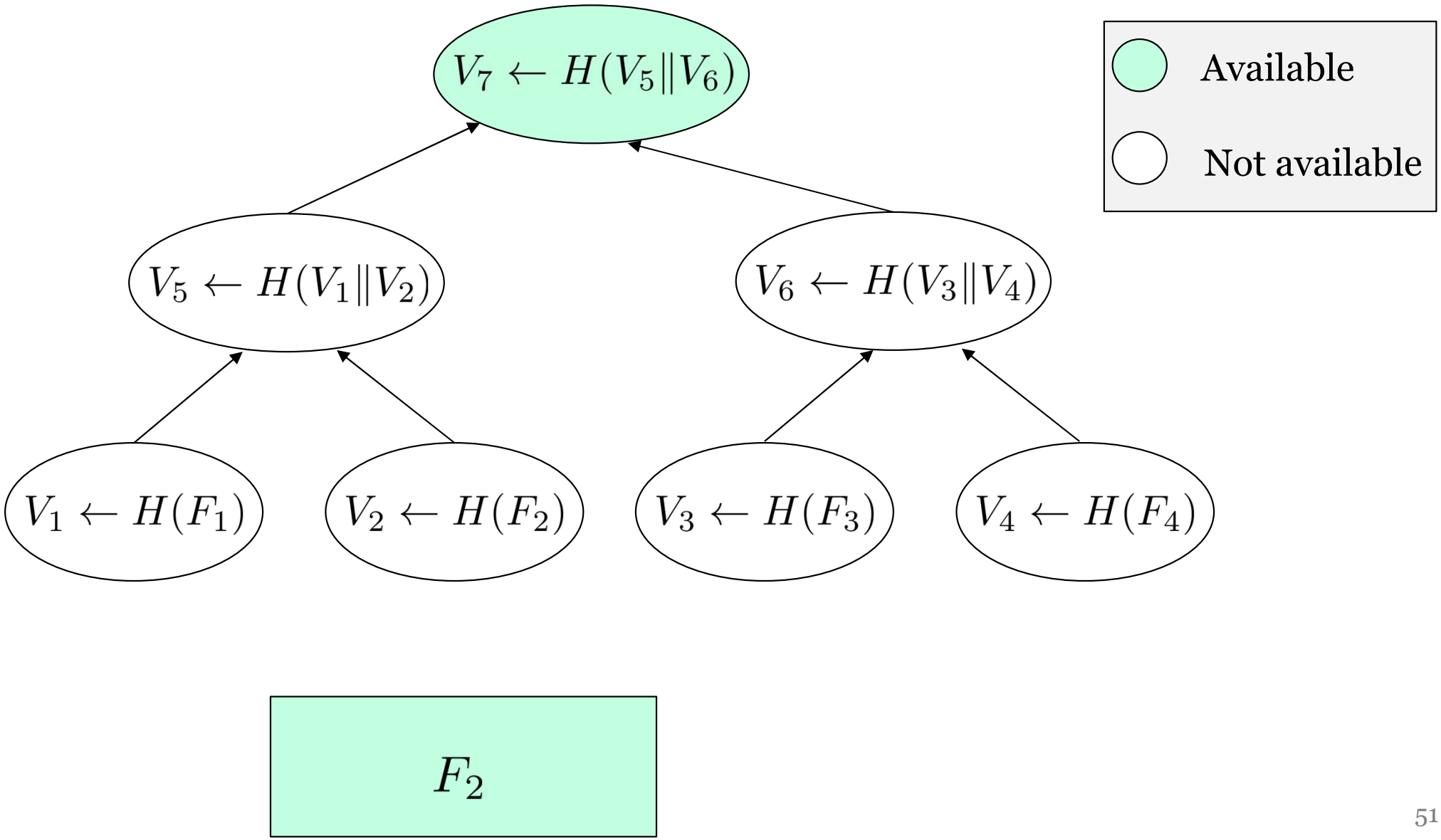
Alice creates Merkle tree and stores it in Dropbox, and store the root locally

Preprocessing cost: O(N) time, cloud storage: O(N) space



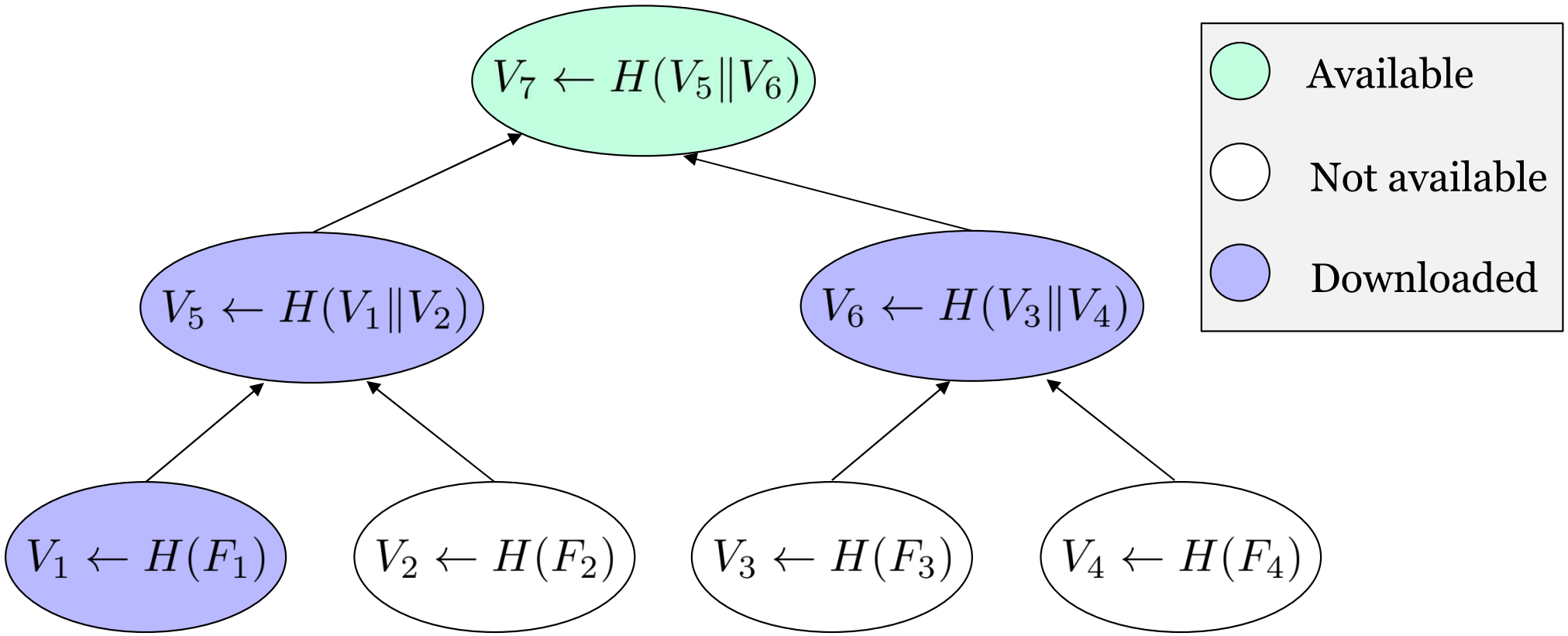
Solution: Merkle Tree

Suppose we want to check the integrity of F_2



Solution: Merkle Tree

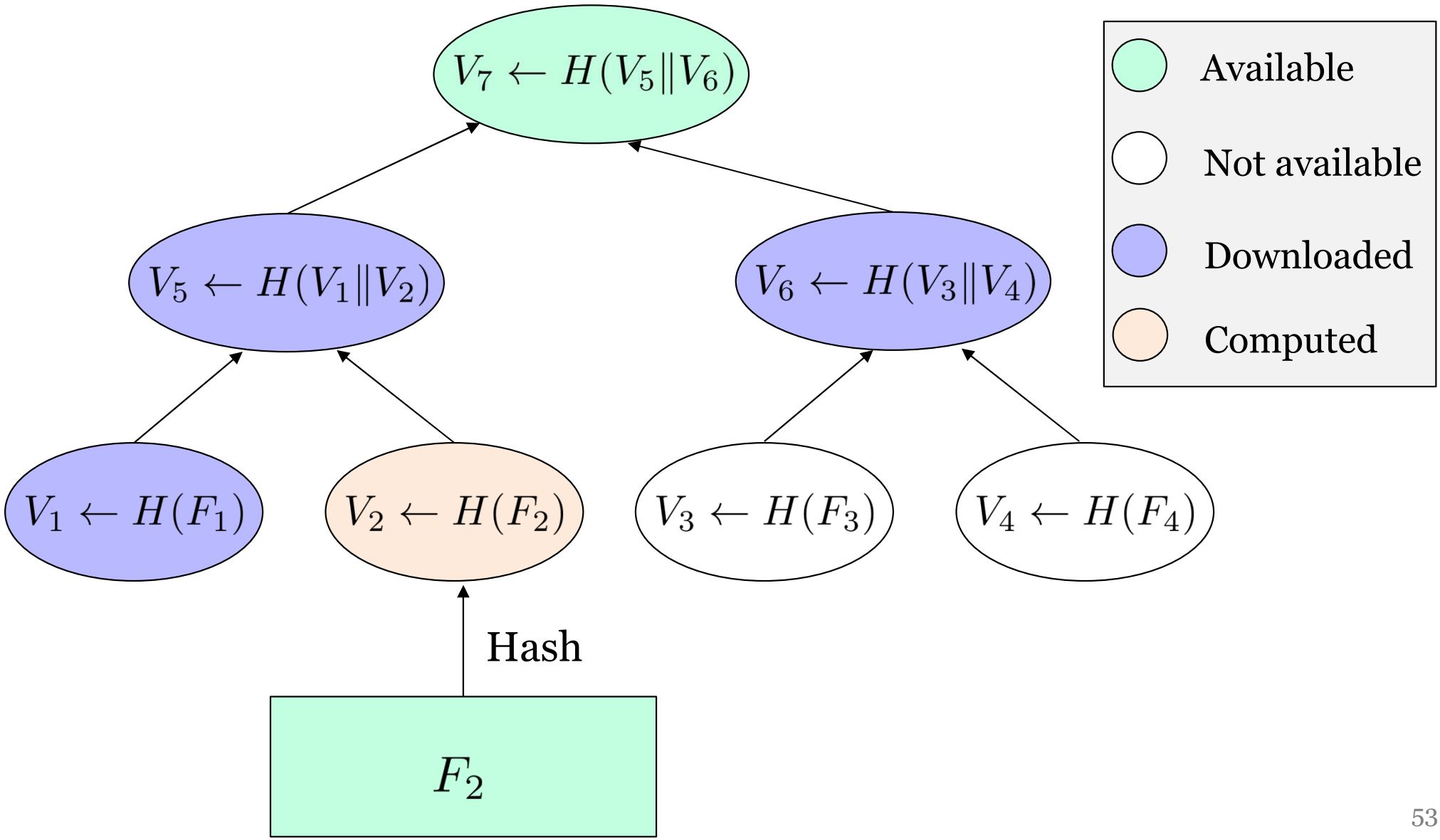
Download V_1, V_5, V_6



F_2

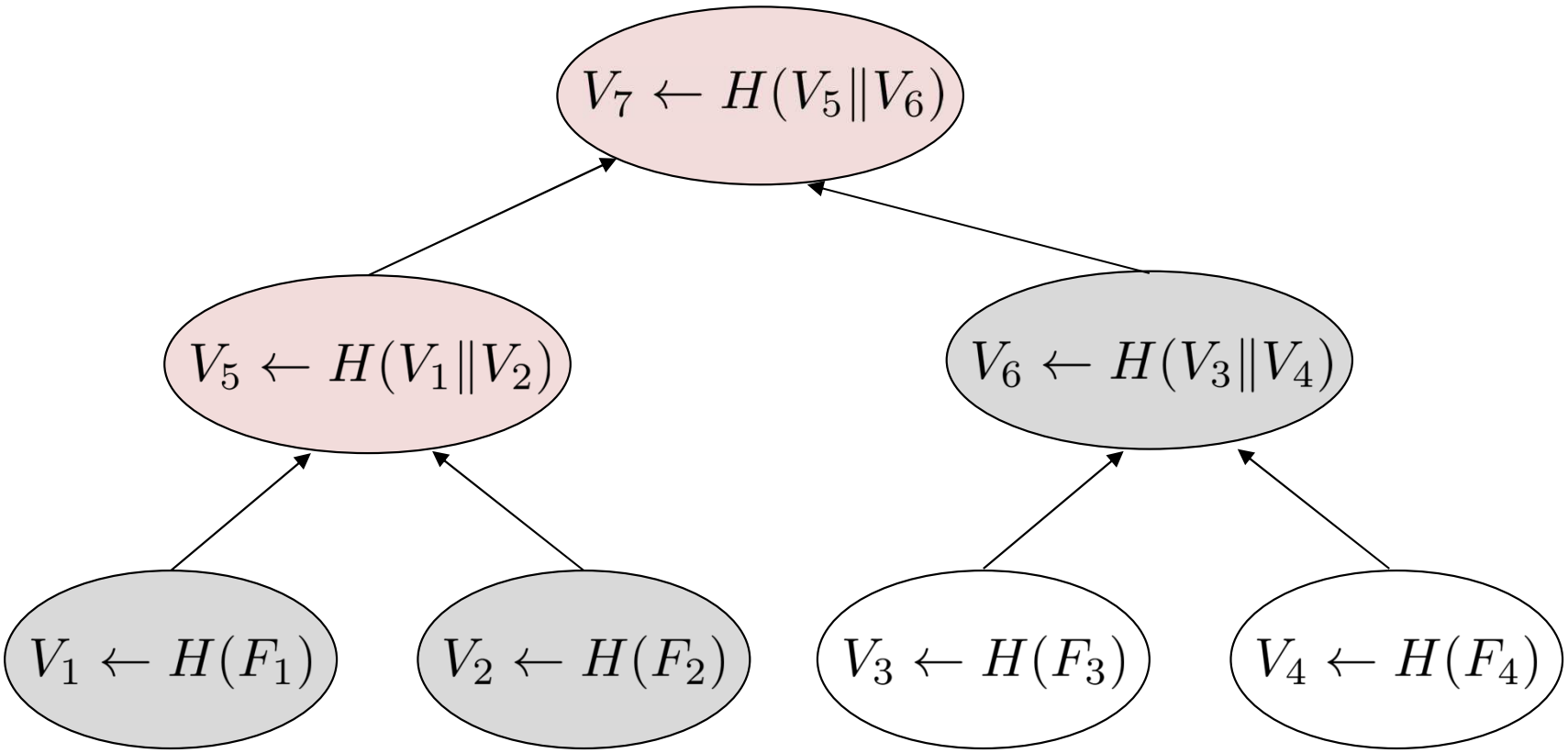
Solution: Merkle Tree

Compute $V_2 \leftarrow H(F_2)$



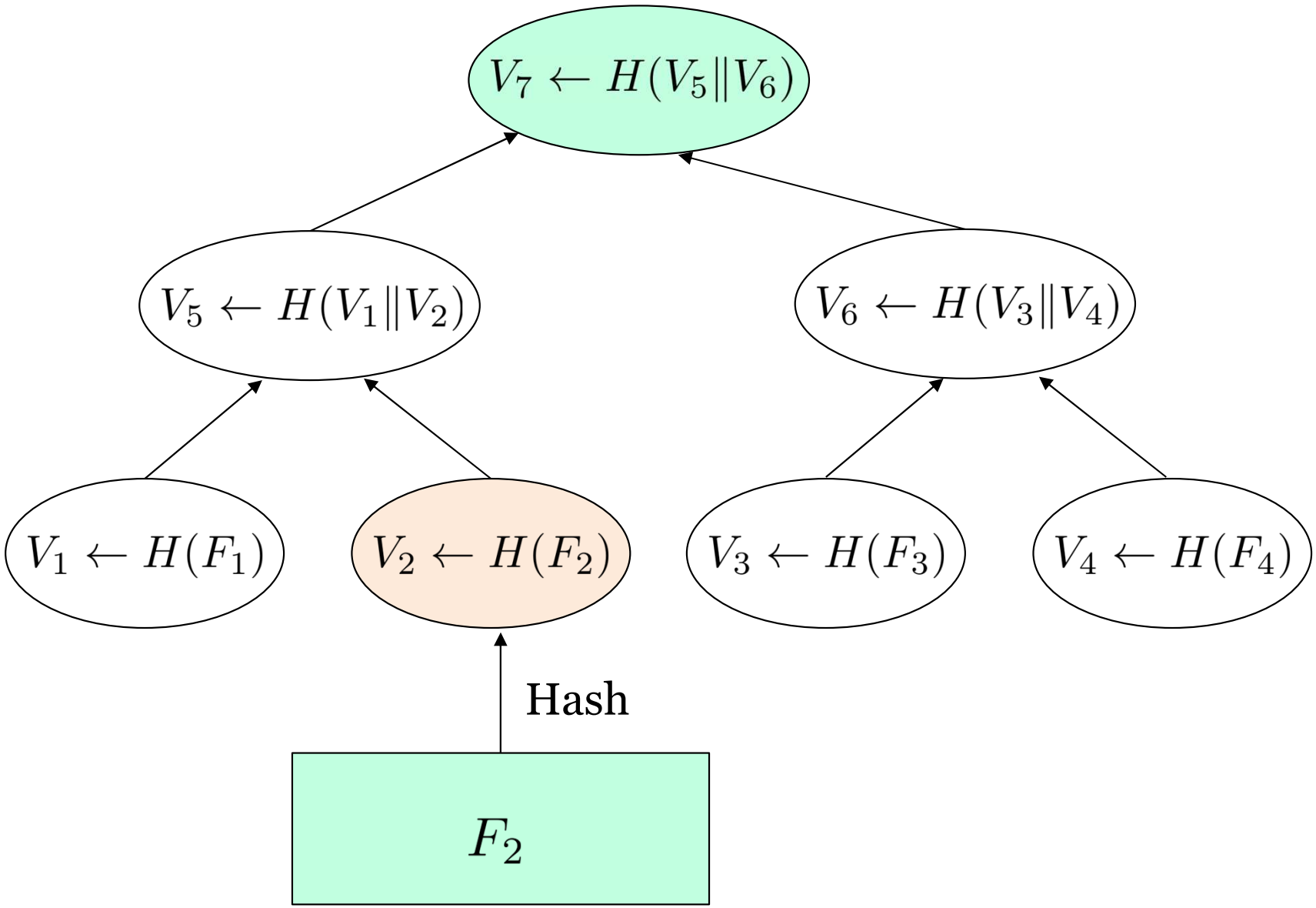
Solution: Merkle Tree

Check if $V_5 = H(V_1 \| V_2), V_7 = H(V_5, V_6)$



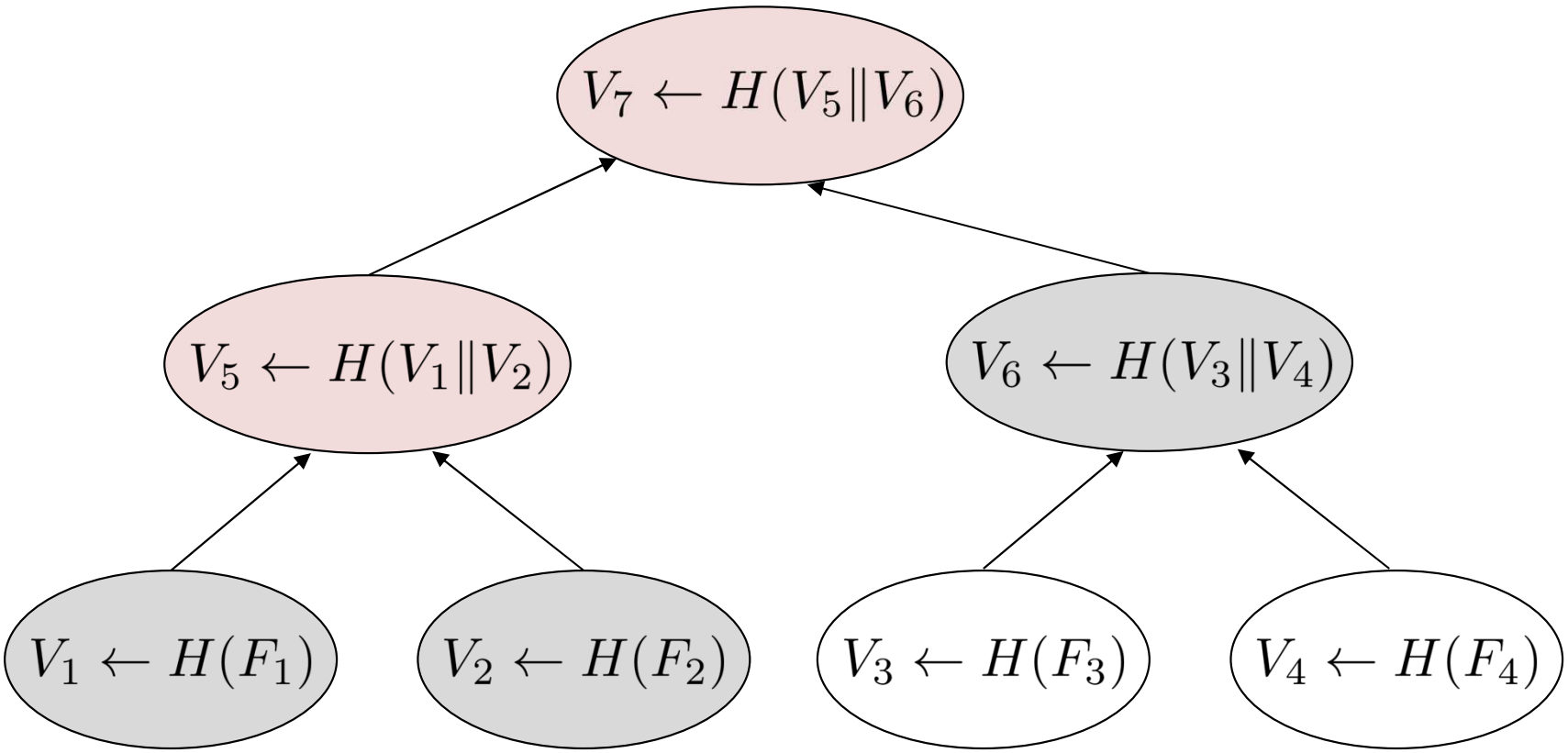
How Much Does it Cost?

O(|File|) time for hashing the file



How Much Does it Cost?

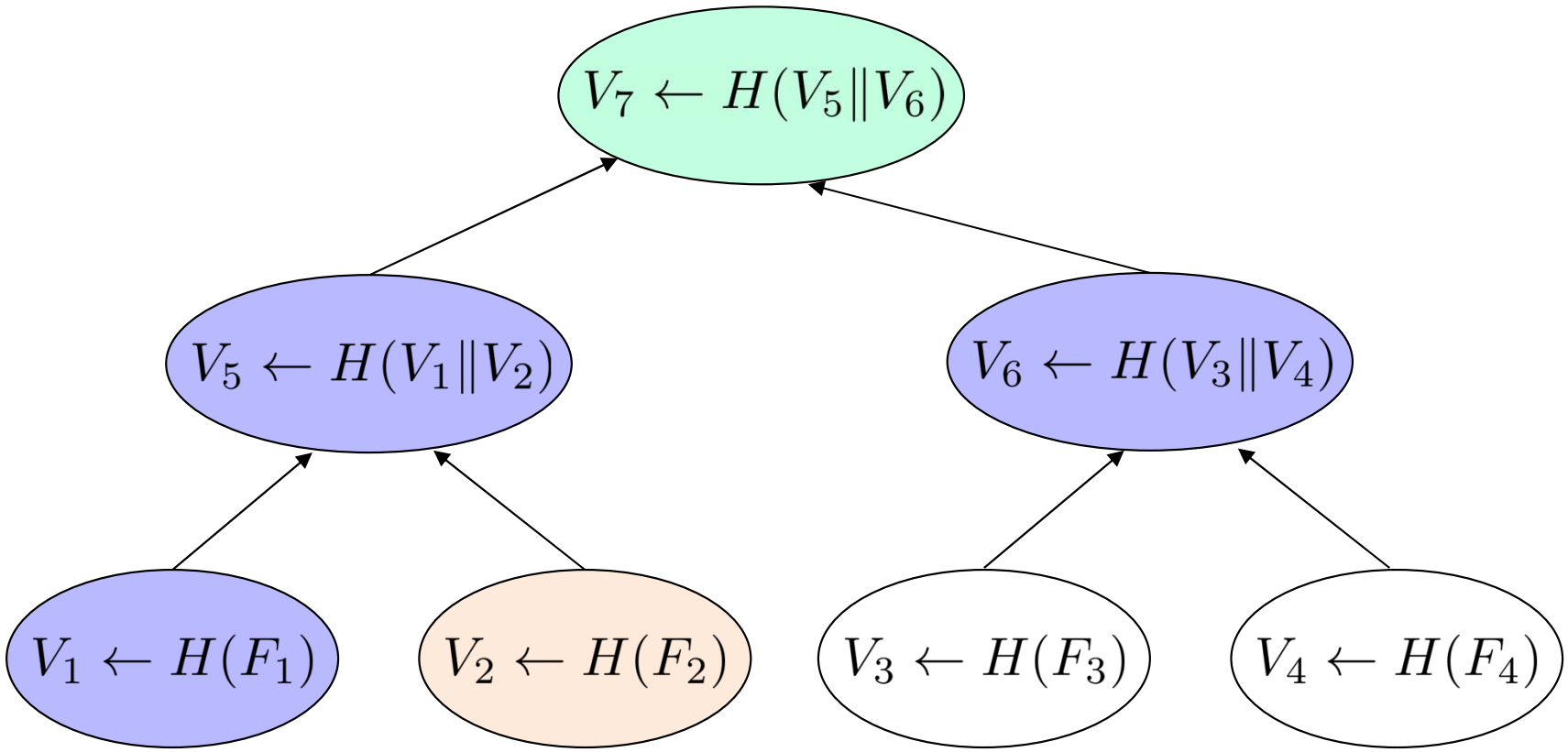
O(log(N)) time for verifying consistency in Merkle tree



These **red** nodes are nodes in the path from V_2 to the root

How Much Does it Cost?

O(log(N)) time and space for downloading



These **purple** nodes are nodes in the path from V_2 to root, and their siblings