



Bảng băm

Hash tables



Nội dung

- Kiểu kết hợp/ánh xạ
- Bảng địa chỉ
- Bảng băm
- Giải quyết xung đột

Kiểu kết hợp

- Kiểu dữ liệu kết hợp (ánh xạ)
 - Bộ gồm chỉ mục và dữ liệu (*key, value*)
- Mảng kết hợp
 - Tập hợp động các phần tử kiểu kết hợp
 - Hỗ trợ các phép toán *insert/delete/search* dựa trên chỉ mục (*key*)
- Các dạng cài đặt: dictionary, map, symbol table...



Ví dụ

- Từ điển: chỉ mục từ + định nghĩa từ
- CSDL: khóa + các trường dữ liệu
 - Dữ liệu dân cư: số căn cước công dân + thông tin
 - Sinh viên: mã sv + ...
- Webpage: link + content

S s

side *n* one of the surfaces of object that is to right or left; aspect; faction — *adj* at, in the side; subordinate — *v* (usu. with with) take up cuse of siding *n* short line of rails from main line sideboard *n* piece of dining room furniture sideburns *n* *n* man's side whiskers side effect additional undesirable effect side-kick *n* Inf close associate sidelong *adj* not directly forward — *adv* obliquely sidestep *v* avoid sidetrack *v* divert from main topic sideways *adv* to or from the side. — *n* border, edge, limit, margin, perimeter, rim, verge; aspect, face, facet, flank, part, surface, view; camp, faction, party, sect, team
sidetrack *v* deflect, distract, divert
side *v* move in furtive or stealthy manner; move sideways
siege *n* besieging of town
siesta *n* rest, sleep in afternoon
sieve *n* device with perforated bottom — *v* sift; strain
sift *v* separate coarser portion from finer. — *adj* filter, separate, sieve
sigh *n* (utter) long audible breath
sight *n* faculty of seeing; thing seen; glimpse; device for guiding eye; spectacle — *v* catch sight of; adjust sight on gun etc.
sightseeing *n* visiting places of interest.
— *n* eye, eyes, seeing, vision; display, exhibition, scene, show, spectacle, vista.
— *v* observe, perceive, see, spot
sign *n* mark, gesture etc. to convey some meaning; (board bearing) notice etc.; symbol; omen — *v* put one's signature to; make sign or gesture. — *n* clue, evidence, gesture, hint, indicate, proof signal, symptom, token; board, notice, placard; badge, device, emblem, ensign, logo, mark, symbol; augury, auspice, omen, portent, warning. — *v* autograph, endorse, initial; beckon, gesticulate, gesture, indicate
signal *n* sign to convey order or information; Radio etc. sequence of electrical

impulses transmitted or received — *adj* remarkable — *v* -nalling, -nalled make signals to; give orders etc. by signals.
— *n* beacon, cue, gesture, indication, mark, sign. — *v* beckon, gesture, indicate, motion, sign
signatory *n* one of those who signs agreements, treaties
signature *n* person's name written by himself signature tune tune used to introduce television or radio programme
signal *n* small seal
significance *n* force, import, meaning, message, point; consequence, importance, relevance, weight
significant *adj* revealing; designed to make something known; important significance *n* — *adj* expressive, indicative, meaningful; critical, important, momentous, vital, weighty
signify *v* -fying, -fied mean, indicate; imply; be of importance
silage *n* fodder crop stored in state of partial fermentation
silence *n* absence of noise; refraining from speech — *v* make silent; put a stop to silencer *n* device to reduce noise of engine exhaust, gun etc. silent *adj*. — *n* calm, hush, peace, quiet, stillness; dumbness, muteness, reticence, taciturnity. — *v* cut off, cut short, gag, muffle, quieten, still
silent *adj* hushed, quiet, soundless, still; dumb, mute, speechless, taciturn, voiceless, wordless
silhouette *n* outline of object seen against light — *v* shown in silhouette. — *n* form, outline, profile, shape
silica *n* naturally occurring dioxide of silicon
silicon *n* brittle metal-like element found in sand, clay, stone silicon chip tiny wafer of silicon used in electronics
silk *n* fibre made by silkworms; thread, fabric made from this silky *adj* silkworm *n*

larva of certain moth
sill *n* ledge beneath window
silly *adj* foolish; trivial. — *adj* absurd, asinine, fatuous, foolhardy, foolish, idiotic, insane, irresponsible, ridiculous, stupid
silo *n* (sil-los) pit, tower for storing fodder
silt *n* mud deposited by water — *v* fill, be choked with silt
silver *n* white precious metal; silver coins; cutlery — *adj* made of silver; resembling silver or its colour silvery *adj*
similar *adj* resembling, like similarity *n* likeness. — *adj* alike, comparable, resembling, uniform
similarity *n* affinity, closeness, resemblance likeness, correspondence
simile *n* comparison of one thing with another
simmer *v* keep or be just below boiling point; be in state of suppressed rage
simper *v* smile, utter in silly or affected way
simple *adj* not complicated; plain; not complex; ordinary; stupid simpleton *n* foolish person simplicity *n* simplify *v* -fying, -fied make simple, plain or easy simply *adv*. — *adj* clear, easy easy-peasy *sl*, intelligible, lucid, plain, uncomplicated, understandable; natural, plain, unfussy; elementary, pure, single, uncombined, unmixed; brainless, dense, feeble, foolish, obtuse, slow, stupid, think
simplicity *n* clarity, clearness, ease; naturalness, plainness, purity
simulate *v* make pretence of; reproduce
simulation *n*
simultaneous *adj* occurring at the same time. — *adj* at the same time, coinciding, concurrent, contemporaneous
sin *n* breaking of divine or moral law — *v* sinning, sinned commit sin sinful *adj*
sinner *n*. — *n* crime, evil, guilt, iniquity, misdeed, offence, trespass, iniquitousness, wickedness — *v* err, fall, lapse, offend, transgress
since *prep* during period of time after — *conj* from time when; because — *adj* from that time
sincere *adj* not hypocritical; genuine
sincerity *n*. — *adj* artless, candid, earnest, frank, genuine, guileless, honest, open,

real, true, unaffected
sincerity *n* candour, frankness, genuineness, honesty, truth
sine *n* in a right-angled triangle, ratio of opposite side to hypotenuse
sinew *n* tough, fibrous cord joining muscle to bone
sinful *adj* bad, corrupt, guilty, immoral, iniquitous, unrighteous, wicked
sing *v* singing, sang, sung utter (sounds, words) with musical modulation; hum, ring; celebrate in song singer *n*. — *adj* chant, croon, trill, warble
singe *v* singeing, singed burn surface of
single *adj* one only; unmarried; for one; denoting ticket for outward journey only — *n* single thing — *v* pick (out) single file persons in one line single-handed *adj* without assistance single-minded *adj* having one aim only. — *adj* individual, lone, one, sole, solitary; free, unattached, unmarried, unwed
single-minded *adj* dedicated, determined, dogged, fixed, steadfast
singlet *n* sleeveless underest
singular *adj* remarkable; unique; denoting one person or thing. — *adj* exceptional, notable, noteworthy, outstanding, remarkable, unparalleled; individual, separate, single
sinister *adj* threatening; evil-looking; wicked. — *adj* menacing, ominous, threatening
sink *v* sinking, sank, sunk or sunken become submerged; drop; decline; penetrate (into); cause to sink; make be digging out; invest — *n* fixed basin with waste pipe. — *v* decline, descend, dip, disappear, drop, ebb, fall, lower, plunge, submerge, subside; decay, decline, die, diminish, dwindle, fade, lessen
sinuous *adj* curving
sinus *n* cavity in bone, esp. of skull
sip *v* sipping, sipped drink in very small portions — *n* amount sipped
siphon, **syphon** *n* (device to) draw liquid from container
sir *n* polite term of address for a man
sire *n* male parent, esp. of horse or domestic animal — *v* father
siren *n* device making loud wailing noise
sirloin *n* prime cut of beef
sissy *adj* weak, cowardly (person)

Mục đích

- Cần cấu trúc dữ liệu để lưu trữ khóa giúp **INSERT/DELETE/SEARCH** nhanh hơn.

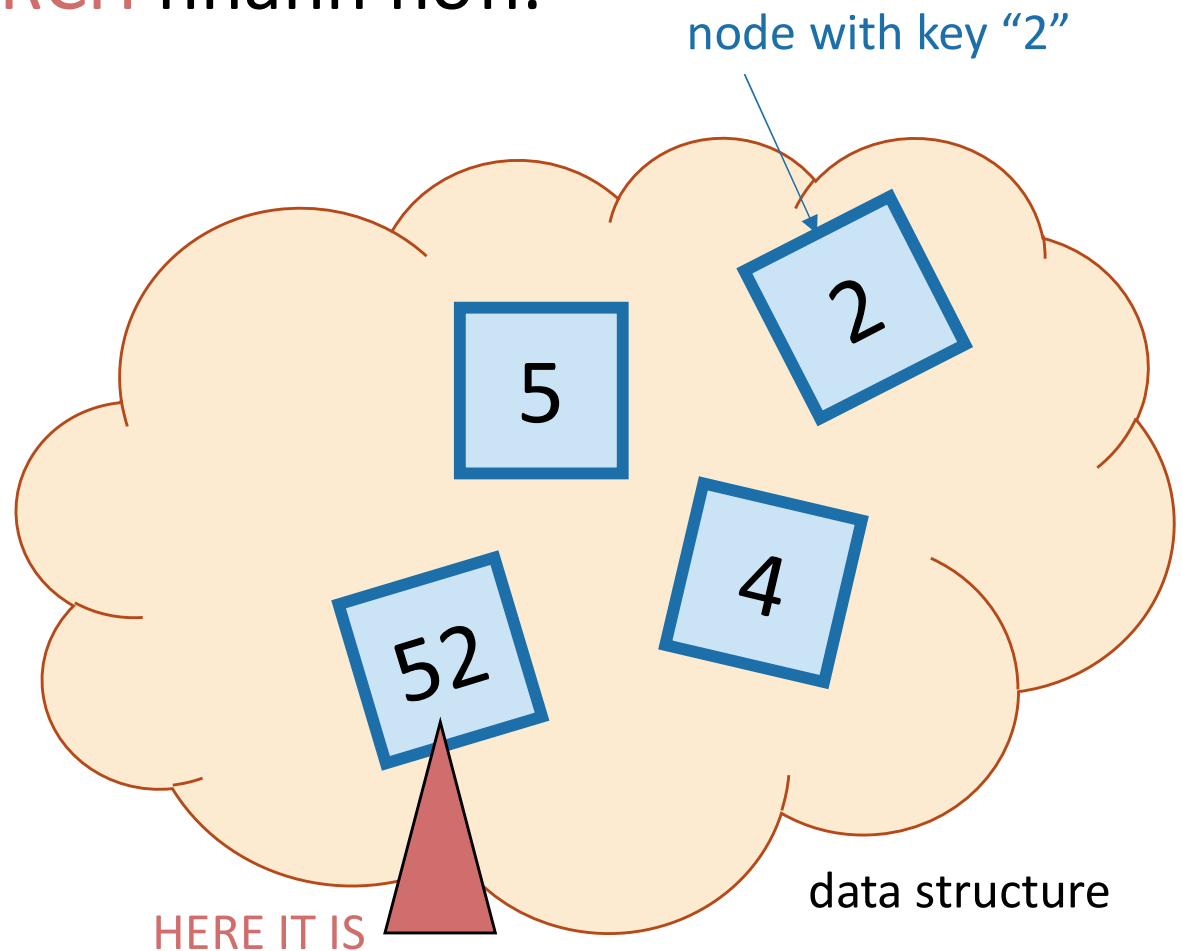
- INSERT



- DELETE



- SEARCH



Độ phức tạp

- Cần lưu trữ và tìm kiếm dữ liệu theo khóa hoặc chỉ mục (key/index)
- Lưu trữ bằng mảng, danh sách liên kết, cây nhị phân?

OPERATION	SORTED ARRAY	UNSORTED LINKED LIST	BST (WORST CASE)	BST (BALANCED)
SEARCH	$O(\log(n))$	$O(n)$	$O(n)$	$O(\log(n))$
DELETE	$O(n)$	$O(n)$	$O(n)$	$O(\log(n))$
INSERT	$O(n)$	$O(1)$	$O(n)$	$O(\log(n))$

$\log(n)$ đã đủ tốt với với CSDL lớn và các ứng dụng hệ thống chưa?

Ví dụ: làm thế nào truy vấn nhanh?

1. Cả nước có gần 11.000 xã/phường được đánh số tuần tự, tìm thông tin khi biết mã số
2. Tìm thông tin sinh viên UET biết mã sinh viên
3. Có 10.000 hội viên được đánh số ngẫu nhiên trong khoảng: 1 ~ 1.000.000
4. Tra cứu từ điển có
 - ☐ 10.000 mục từ
 - ☐ 100.000 mục từ
 - ☐ 300.000 mục từ (Oxford Dic. ~ 273k)

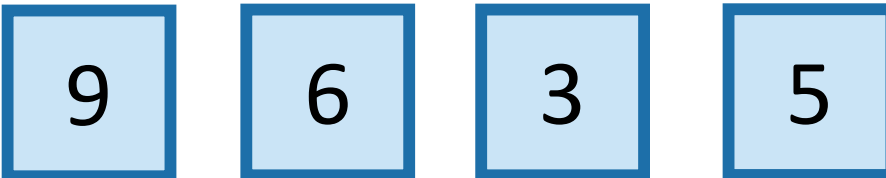
...

Tìm kiếm nhị phân: số phép tính?

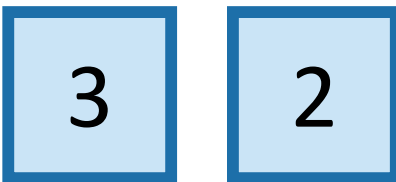
- Độ phức tạp: $\log(n)$
- Số phép tính thực tế?
 - ☐ Tìm một mã sinh viên?
 - ☐ Tìm một mục từ trong từ điển?
- Giải pháp: dùng mảng địa chỉ trực tiếp?
 - ☐ Dùng chính giá trị của khóa tìm kiếm (key) là địa chỉ

Mảng địa chỉ trực tiếp

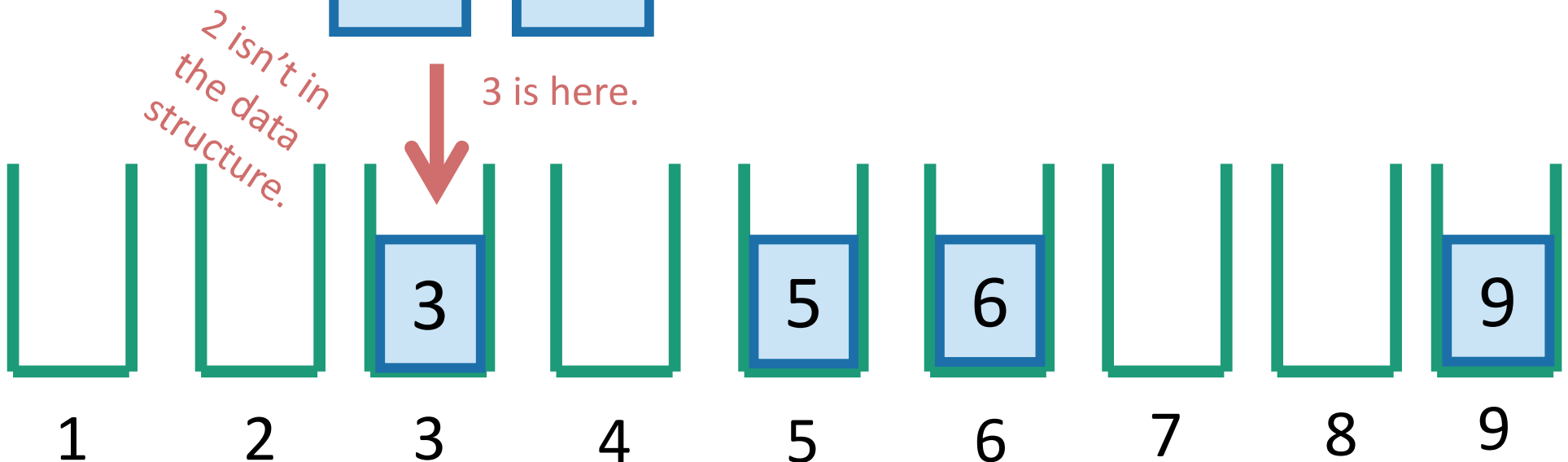
- Giả sử khóa là tập: {1,2,3,4,5,6,7,8,9}.

• INSERT: 

• DELETE: 

• SEARCH: 

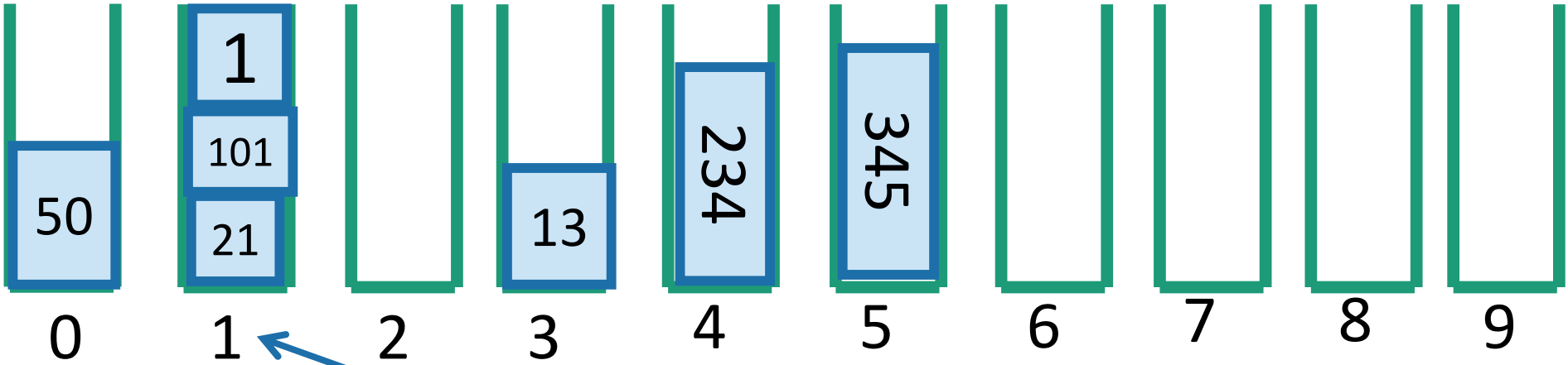
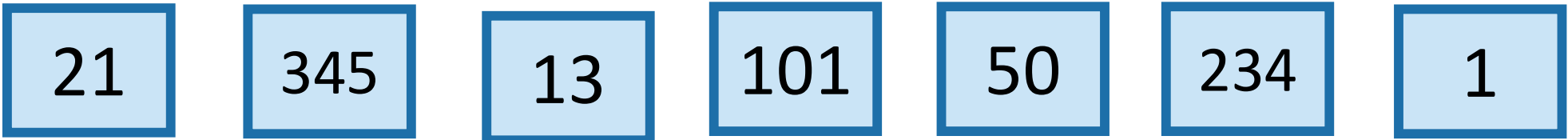
Không thực tế nếu
không gian dữ liệu lớn
Vd. INSERT 1000000



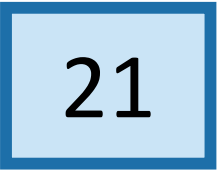
Giải pháp: chia vào các rổ

Nén vào không gian nhỏ hơn!!!

INSERT:



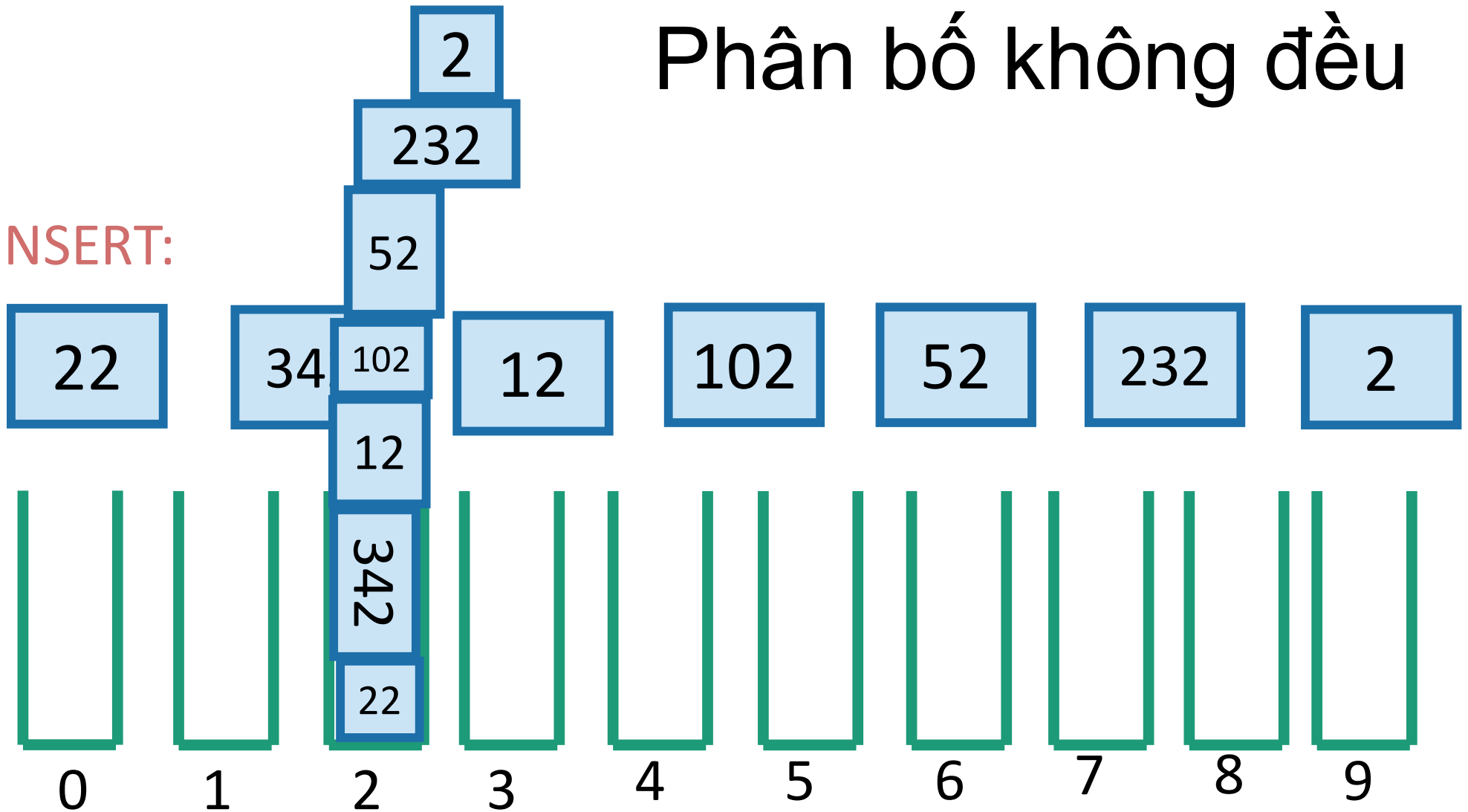
Now SEARCH



Trong cùng rổ thì duyệt hết

Phân bố không đều

INSERT:



Now SEARCH

102

Băm (hashing)

- Bảng băm

- Bảng địa chỉ để giúp tăng tốc truy vấn dữ liệu

- Hàm băm (hash function)

- “Ánh xạ” để tạo ra bảng băm từ các khóa

- Cần hàm băm tốt

- Phân bổ đều
 - Hạn chế xung đột (trùng địa chỉ)
 - Tính toán đơn giản

Khái niệm

- U là không gian dữ liệu kích thước M (**cực lớn**)
- Thực tế chỉ **n** dữ liệu xuất hiện và **$M \gg n$**
- Chúng ta không biết rõ cụ thể tập dữ liệu nào



Chỉ một số khóa
cụ thể tồn tại

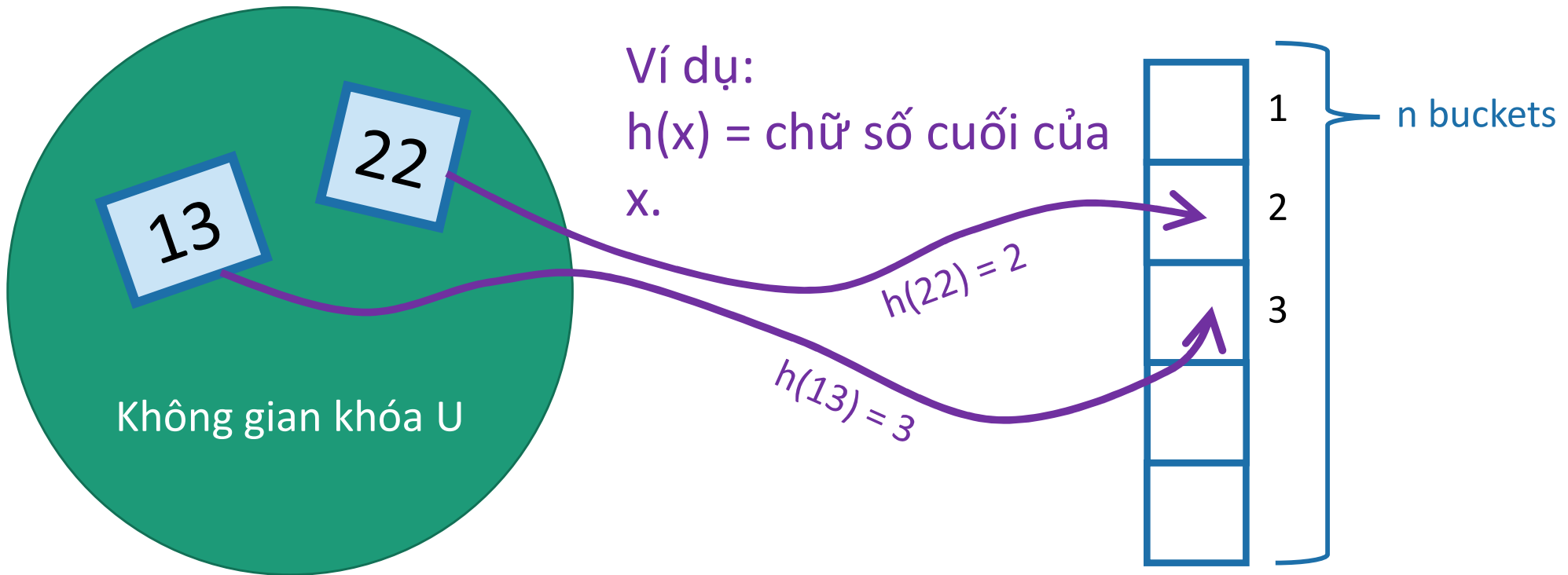
Ví dụ: U là tập các nhãn dài không quá 100 ký tự. (128^{100}).

Ví dụ: #hashinghashtags

Số lượng nhãn thực tế nhỏ hơn 128^{100} nhiều lần.

Hàm băm

- Hàm băm $h: U \rightarrow \{1, \dots, n\}$ là ánh xạ phần tử của U vào tập $\{1, \dots, n\}$



Chọn hàm băm

■ Hàm băm tốt

- ☐ Tính toán dễ dàng
- ☐ Hạn chế xung đột (phân bố đều)

■ Cần khảo sát

- ☐ Không gian dữ liệu (M), kích thước bảng n
- ☐ Phân bố dữ liệu (tính chất của tập dữ liệu)

Ví dụ

- Cần tính bảng băm kích thước N cho tập nhân (xâu ký tự) $S = \{ S_1, S_2 \dots, S_N \}$
- $H(S_i) = (\sum S_i[j]) \bmod N$
 - Có đủ tốt không?
- $H(S) = (\sum S_i[j] \cdot d^j) \bmod N$
 - Độ phức tạp?

Horner's Rule

- $H(S) = (\sum S_i [j] \cdot d^j) \bmod N$

$$\sum a_i \cdot x^i = (\dots((a_n \cdot x + a_{n-1}) \cdot x + \dots) \cdot x + a_0)$$

Chọn hàm băm: Chia dư

- Ánh xạ vào không gian m ô bằng cách chia lấy số dư:

$$h(x) = x \bmod n$$

- Ưu điểm: nhanh
- Hạn chế: khả năng xung đột cao, đặc biệt là với một số giá trị m đặc thù
 - m là lũy thừa của 2
 - m không phải là số nguyên tố

Hàm băm tốt: Universal Hashing

- Chọn số nguyên tố lớn $p \geq M$.
- Chọn ngẫu nhiên a, b ($0 < a, b < p - 1$)

$$f_{a,b}(x) = (ax + b) \bmod p$$

$$h_{a,b}(x) = f_{a,b}(x) \bmod n$$

- Đảm bảo $f_{a,b}(x)$ đơn nhất (ánh xạ không gian U vào một không gian lớn hơn)
- Cần chọn p không quá lớn so với M

Ví dụ

- Giả sử $M = 1000$, $n = 101$, ta chọn $p = 1009$, $a = 12$, $b = 55$
- Ta có

$$hf(50) = 49$$

$$hf(100) =$$

$$hf(155) =$$

$$hf(205) =$$

$$hf(330) =$$

$$hf(620) =$$

$$hf(730) =$$

$$hf(850) =$$

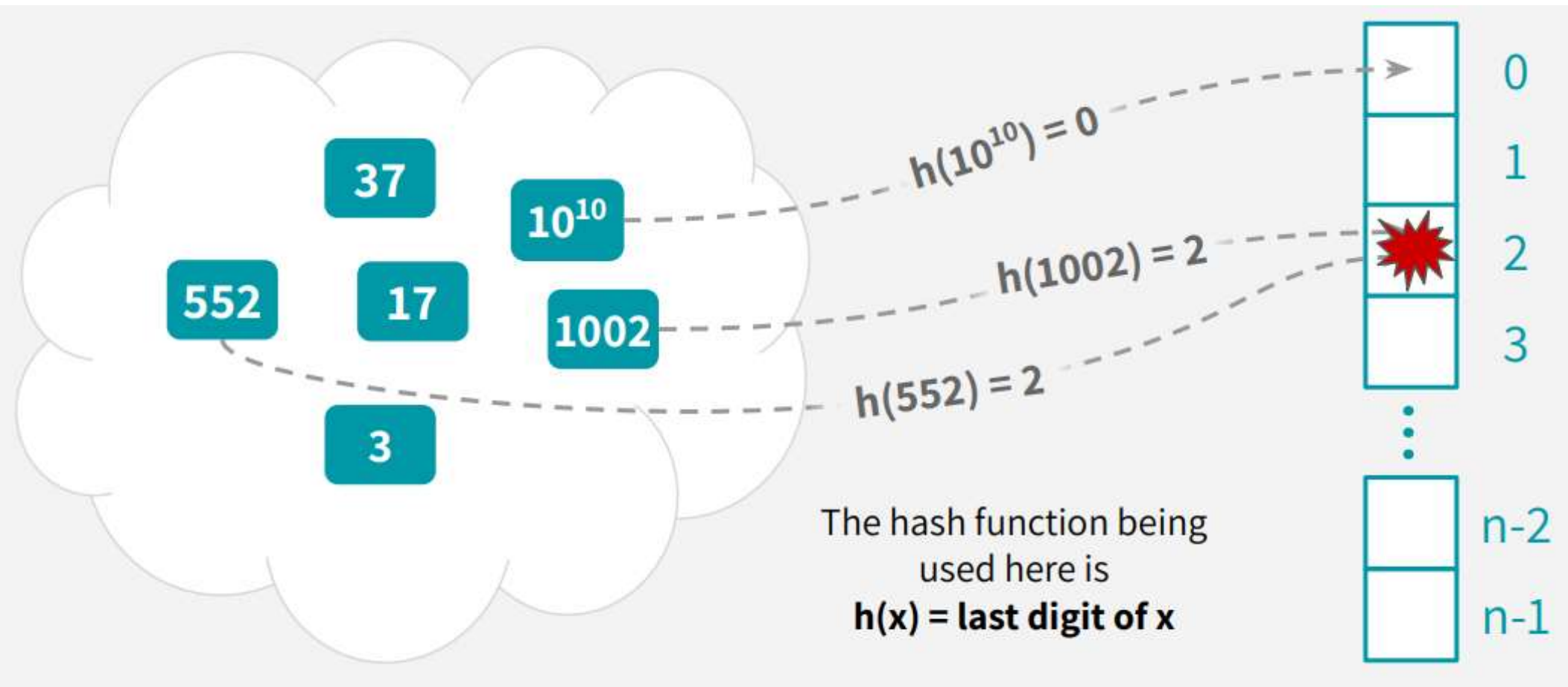
$$hf(999) =$$

Ví dụ: mã sinh viên UET

- Chọn số p, n ?
- Chọn a, b
- Ví dụ
 - $p = 30.000.011$
 - $n = 10.007$
 - $a = 5, b = 19$
 - MSV 25052004 \Rightarrow 6320

Xung đột

- Không tránh khỏi xung đột khi 2 khóa (hoặc nhiều hơn) bị ánh xạ vào cùng một rổ



Giải quyết xung đột

- Chuỗi tách biệt (separate chaining)
 - Xem mỗi vị trí của bảng băm là một chuỗi phần tử
- Cơ chế địa chỉ mở (open addressing)
 - Nếu xung đột thì dò tìm địa chỉ chưa được sử dụng

Chuỗi tách biệt - separate chaining

- Mảng n vị trí (rổ)
- Mỗi rổ là một danh sách liên kết
 - Thêm vào danh sách với thời gian $O(1)$
 - Tìm kiếm với thời gian $O(\text{length}(\text{list}))$.

INSERT:

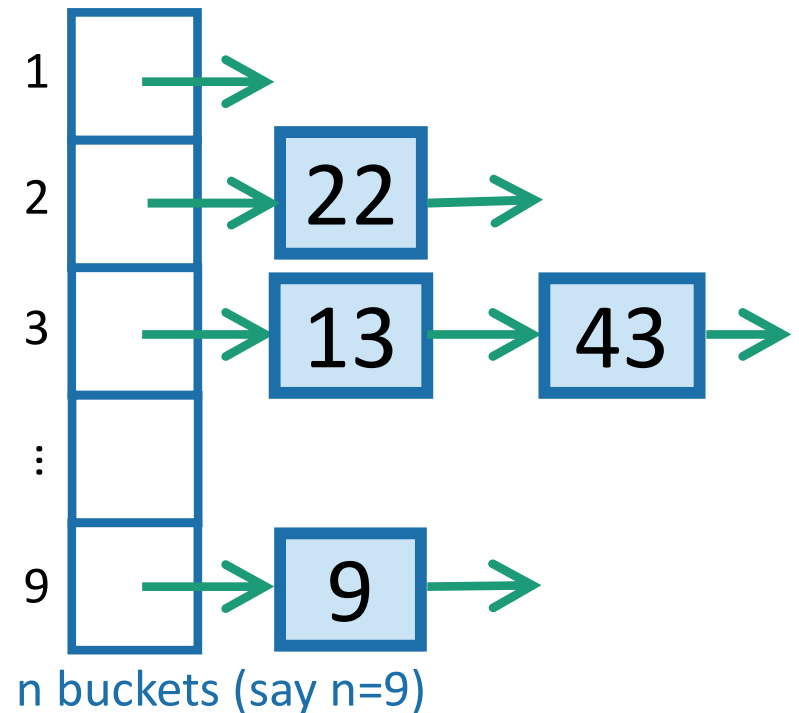


SEARCH 43:

Duyệt trong rổ $h(43) = 3$.

DELETE 43:

Tìm 43 trong chuỗi và xóa

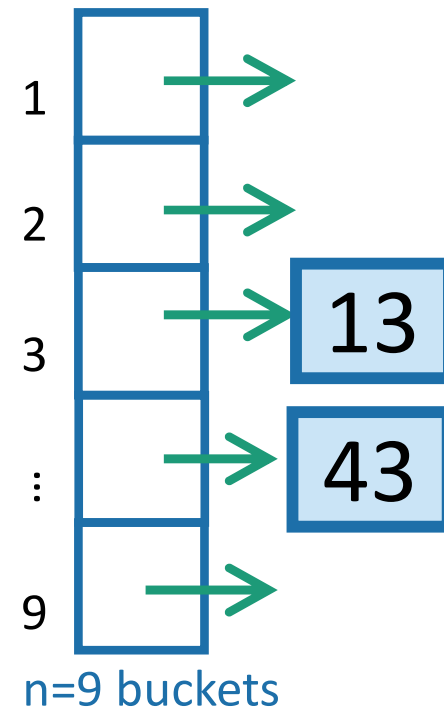
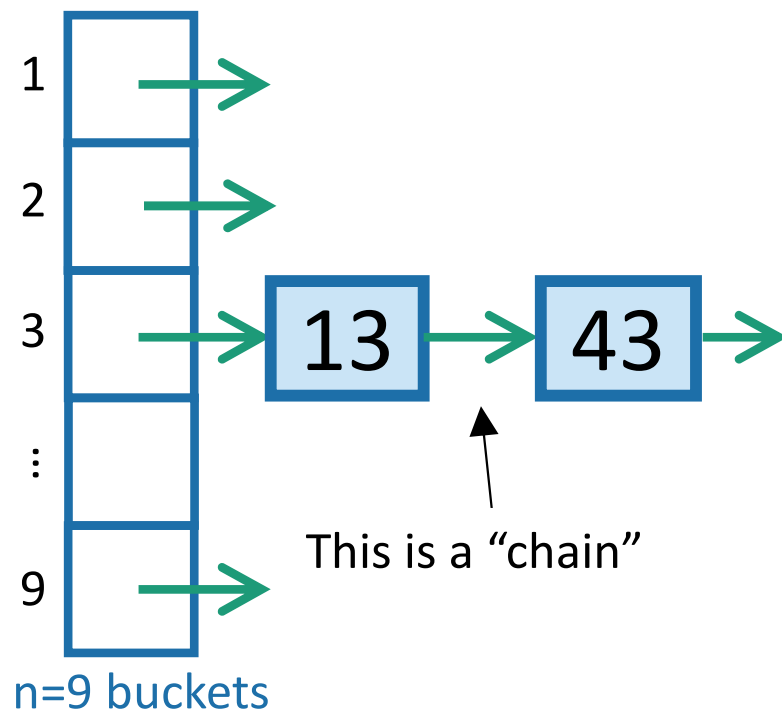


Chuỗi tách biệt

- Trong trường hợp nhiều chuỗi con có kích thước lớn, và không gian dữ liệu tương đối ổn định, có thể cải thiện hiệu năng bằng
 - Tổ chức cây tìm kiếm nhị phân
 - Tạo bảng băm thứ cấp
- Ví dụ, tra cứu từ điển giấy
 - *Kết hợp băm và tìm kiếm nhị phân?*

Địa chỉ mở - open addressing

- Nếu trùng địa chỉ thì dò tìm địa chỉ chưa sử dụng tiếp theo
- Đánh dấu địa chỉ đó đã được dùng



Thăm dò tuyến tính

- Phương pháp đơn giản nhất là thăm dò tuyến tính, bằng cách cải tiến hàm băm bổ sung thêm chỉ số thăm dò

$$h(x) = f(x) \bmod N$$



$$h(x, i) = (f(x) + i) \bmod N$$

- Thăm dò lần lượt $i=0$ cho đến khi tìm được rỗng hoặc giới hạn thăm dò m cho trước

Địa chỉ mở - xóa phần tử

- Khi xóa phần tử, cần đánh dấu ô tương ứng là đã xóa (khác với chưa sử dụng) để không ảnh hưởng tới tìm kiếm các phần tử liên quan



then, Insert 53 ???

So sánh


Operation	Sorted array	Linked list	BST	Hash table (lý tưởng)
SEARCH	$O(\log n)$	$O(n)$	$O(\log n)$	$O(1)$
DELETE	$O(n)$	$O(n)$	$O(\log n)$	$O(1)$
INSERT	$O(n)$	$O(1)$	$O(\log n)$	$O(1)$

Ứng dụng

- Các ứng dụng cần truy vấn tốc độ cao
 - Các hệ quản trị CSDL
 - Search engine
 - Các ứng dụng thời gian thực với dữ liệu lớn
- Là một kỹ thuật mã hóa một chiều
 - Ví dụ: lưu mã hóa mật khẩu
 - $h(\text{"vùng ới mở cửa ra"}) = 0xABC123$

Bài tập/thực hành

- Cài đặt bảng băm cho khóa số nguyên với kích thước không gian M , n tùy ý
 - Cài đặt 2 phương thức chuỗi tách biệt và địa chỉ mở
- Xây dựng hàm băm cho khóa là chuỗi ký tự (ví dụ: tập từ vựng tiếng Anh)
- So sánh hiệu năng với mảng, BST trên tập dữ liệu lớn



Chuẩn bị

- Tìm hiểu về đồ thị
 - ☐ Khái niệm, đồ thị vô hướng, có hướng
 - ☐ Miền liên thông
 - ☐ Đường đi ngắn nhất
 - ☐ Cây khung