Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Εθνικό Μετσόβιο Πολυτεχνείο

Προγραμματιστικές Τεχνικές

Δυναμικές δομές χειρισμού συνόλων: δομές ένωσης-εύρεσης – ουρές προτεραιότητας – κατακερματισμός

Union-Find, Heaps, Hashing

Διδάσκοντες

Βασίλης Βεσκούκης
Νίκος Λεονάρδος
Άρης Παγουρτζής
Νίκος Παπασπύρου
Πέτρος Ποτίκας
Σωτήρης Κοκόσης
Γιώργος Σιόλας

progtech@cslab.ece.ntua.gr

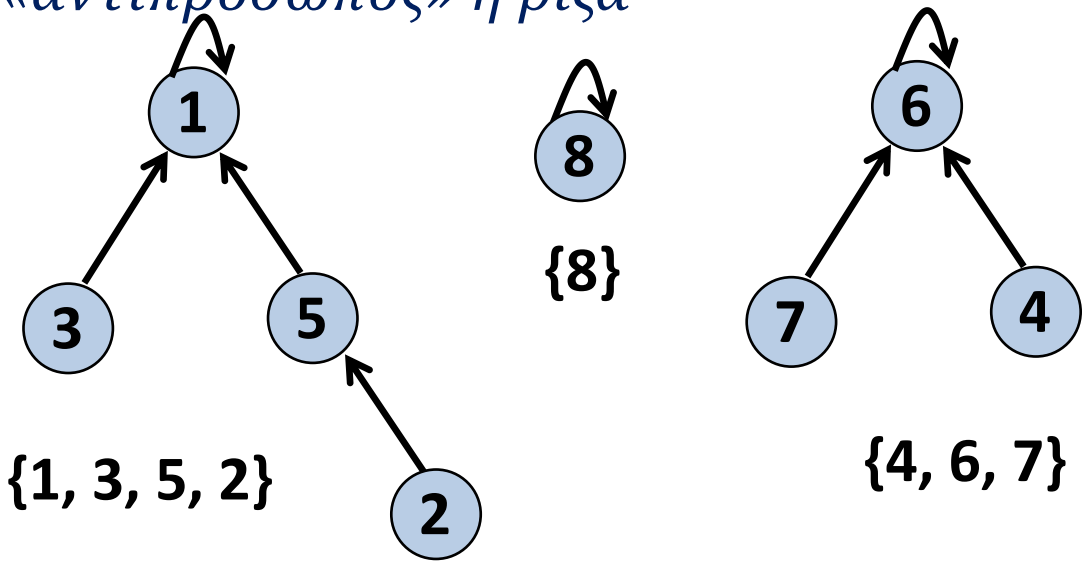
Επιμέλεια διαφανειών: Άρης Παγουρτζής (ΕΜΠ)

Δομές UNION-FIND

- Υποστηρίζουν πράξεις συνόλων:
  - Δημιουργία συνόλου [Makeset]
  - Εύρεση (αναζήτηση) [Find]
  - Ένωση [Union]
- Εφαρμογές: ανίχνευση κοινοτήτων (community detection), (δυναμική) συνεκτικότητα - συνεκτικές συνιστώσες (connected components), αλγόριθμος Kruskal για ελάχιστο συνδετικό δέντρο (minimum spanning tree), και πολλές άλλες

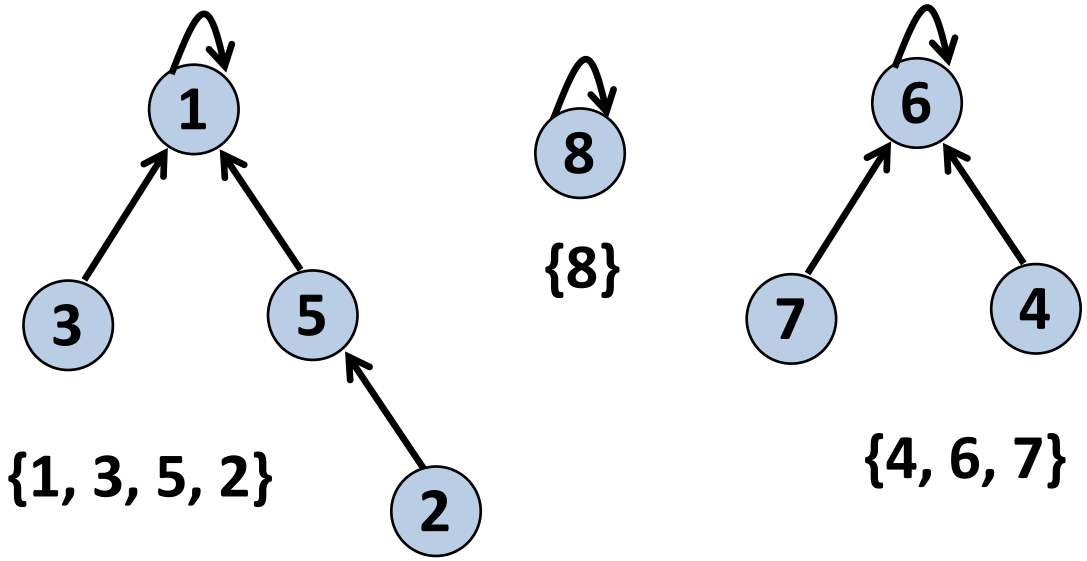
Δομές UNION-FIND

- Υποστηρίζουν πράξεις συνόλων:
  - Δημιουργία συνόλου [Makeset]
  - Εύρεση (αναζήτηση) [Find]
  - Ένωση [Union]
- Υλοποίηση συνόλων με δέντρα: *ένα δέντρο για κάθε σύνολο, «αντιπρόσωπος» η ρίζα*



Δομές UNION-FIND

- Υποστηρίζουν πράξεις συνόλων:
  - Δημιουργία συνόλου [Makeset]
  - Εύρεση (αναζήτηση) [Find]
  - Ένωση [Union]
- Απλή υλοποίηση: με pointers ή array



Δομές UNION-FIND

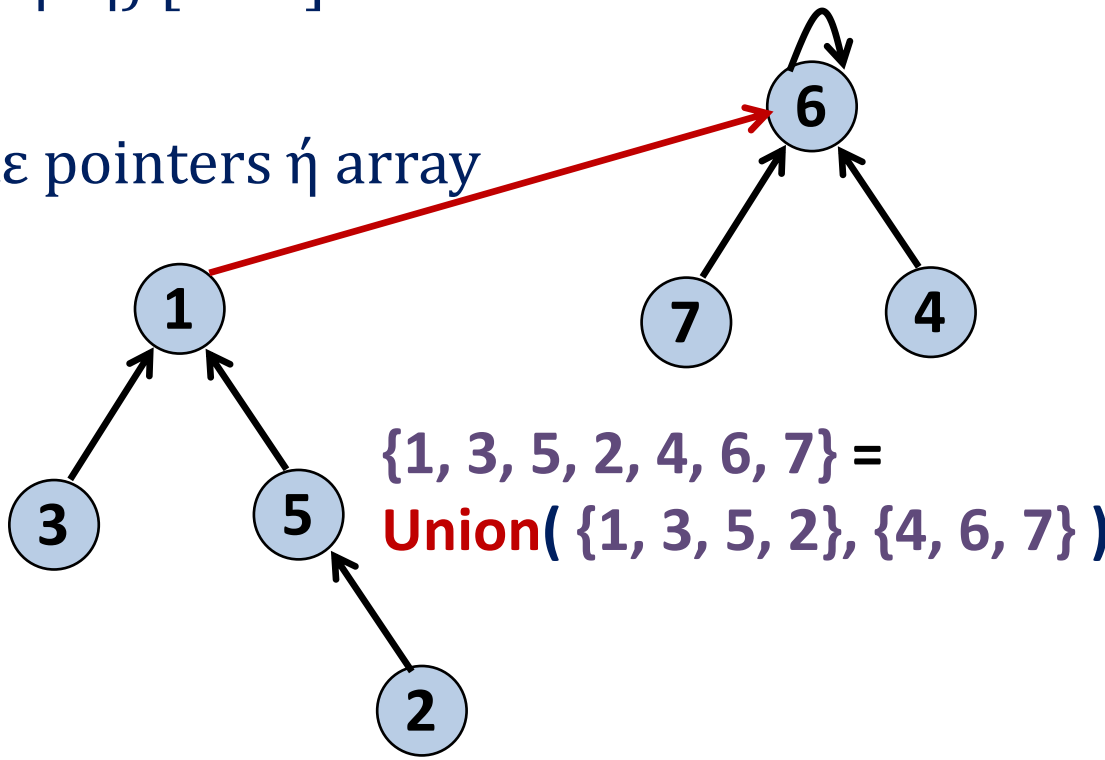
- Υποστηρίζουν πράξεις συνόλων:
  - Δημιουργία συνόλου [Makeset]
  - Εύρεση (αναζήτηση) [Find]
  - Ένωση [Union]
- Απλή υλοποίηση: με pointers ή array
- Δημιουργία: O(1)
- Ένωση: ?

Makeset(8)

{8}

Δομές UNION-FIND

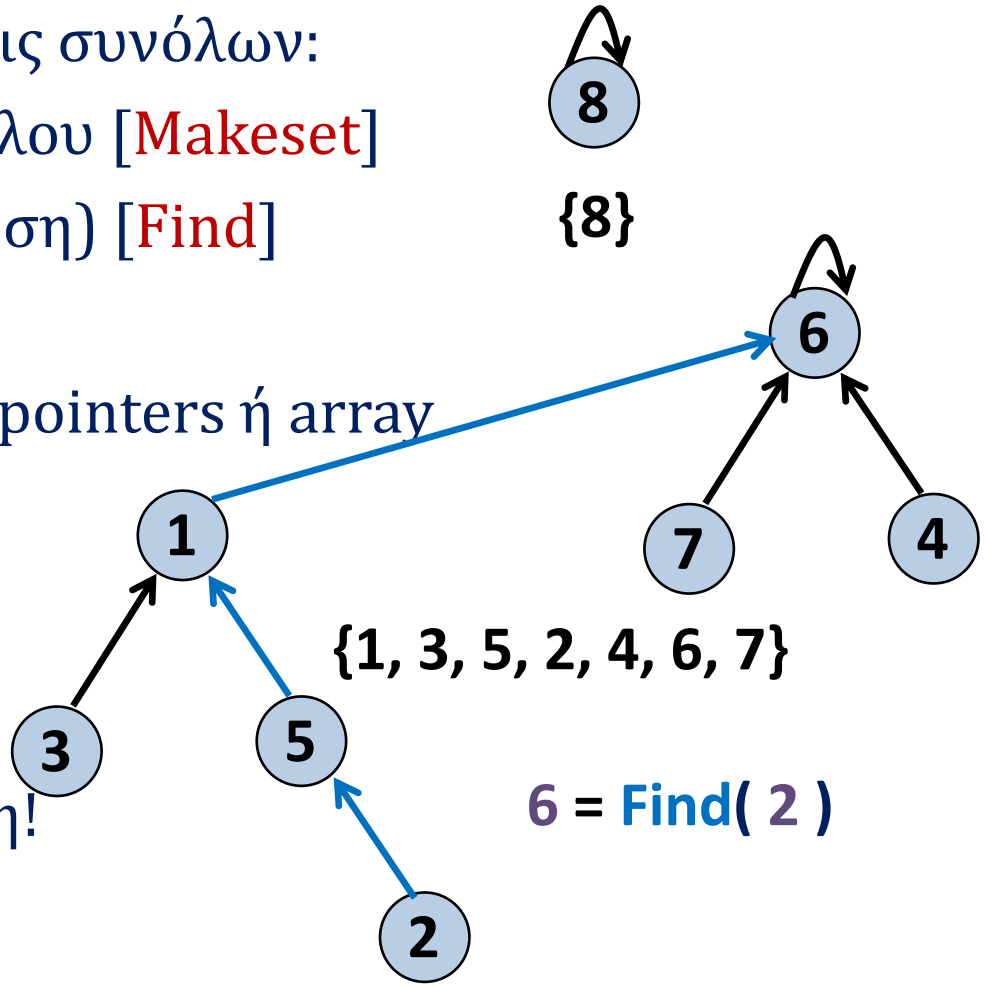
- Υποστηρίζουν πράξεις συνόλων:
  - Δημιουργία συνόλου [Makeset]
  - Εύρεση (αναζήτηση) [Find]
  - Ένωση [Union]
- Απλή υλοποίηση: με pointers ή array
- Δημιουργία: O(1)
- Ένωση: O(1)
- Εύρεση: ?



Δομές UNION-FIND

- Υποστηρίζουν πράξεις συνόλων:
  - Δημιουργία συνόλου [Makeset]
  - Εύρεση (αναζήτηση) [Find]
  - Ένωση [Union]
- Απλή υλοποίηση: με pointers ή array
- Δημιουργία: O(1)
- Ένωση: O(1)
- Εύρεση: O(n)

χειρότερη περίπτωση!
Γίνεται καλύτερα;



Δομές UNION-FIND

- Υποστηρίζουν πράξεις συνόλων:
  - Δημιουργία συνόλου [Makeset]
  - Εύρεση (αναζήτηση) [Find]
  - Ένωση [Union]
- Υλοποίηση συνόλων με δένδρα: *ένα δένδρο για κάθε σύνολο, «αντιπρόσωπος» η ρίζα, κόμβοι δείχνουν προς τον γονέα*
- Απλή υλοποίηση: με pointers ή array
- Δημιουργία: $O(1)$
- Ένωση: $O(1)$
- Εύρεση: $O(\log n)$ με Union-by-Size

Δομές UNION-FIND

- Εύρεση: $O(\log n)$ με **Union-by-Size** :

  το δέντρο με τα λιγότερα στοιχεία «κρεμιέται» στη ρίζα του δέντρου με τα περισσότερα στοιχεία

  - παραλλαγές: **Union-by-Height**, **Union-by-Rank**

Ισχυρισμός: κάθε δένδρο n κόμβων που προκύπτει με διαδοχικές εφαρμογές Union-by-Size έχει ύψος

$$h \leq \log_2 n$$

Απόδειξη: με *επαγωγή* στην *κατασκευή του δένδρου*

Δομές UNION-FIND

Θεώρημα: κάθε δένδρο n κόμβων που προκύπτει με διαδοχικές εφαρμογές Union-by-Size έχει ύψος

$$h \leq \log_2 n$$

Απόδειξη: με *επαγωγή* στην *κατασκευή του δένδρου* :

Θεωρήστε δένδρο ύψους h με n κόμβους, από ένωση δύο δένδρων με ύψη $h_1 \leq \log_2 n_1$, , $h_2 \leq \log_2 n_2$

Έστω χ.β.γ. ότι $n_1 \leq n_2$

Αν $h_1 < h_2$ τότε $h = h_2 \leq \log_2 n_2 < \log_2 n$

Αν $h_1 \geq h_2$ τότε $h = h_1+1 \leq \log_2 (2n_1) \leq \log_2 n$ □

Δομές UNION-FIND

- Πολυπλοκότητα πράξεων συνόλων:
  - Δημιουργία: $O(1)$
  - Ένωση: $O(1)$
  - Εύρεση: $O(\log^* n)$ ***αποσβετικά*** (amortized analysis) σε σύνολο m πράξεων

με **Union-by-Rank** + **Path Compression** (σε κάθε εύρεση όλα τα στοιχεία του μονοπατιού προς την ρίζα τα «κρεμάμε» κατευθείαν στη ρίζα)

($\log^* n \leq 5$ για $n \leq 2^{65636}$!!)

Δομές UNION-FIND

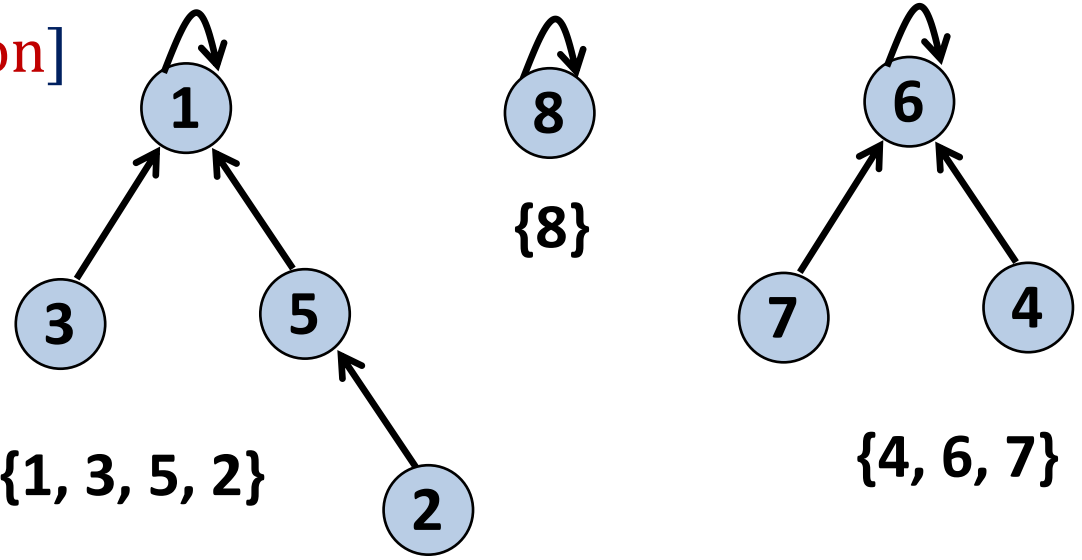
- Πολυπλοκότητα πράξεων συνόλων :
  - Δημιουργία: $O(1)$
  - Ένωση: $O(1)$
  - Εύρεση: $O(\alpha(n))$ ***αποσβετικά*** (amortized analysis) σε σύνολο m πράξεων (με *πιο πολύπλοκη ανάλυση*)

με Union-by-Rank + Path Compression

$\alpha(n) \leq 4$ για $n \leq 2^{2^{2^{65636}}}$-3 , αντίστροφη Ackermann

Δομές UNION-FIND

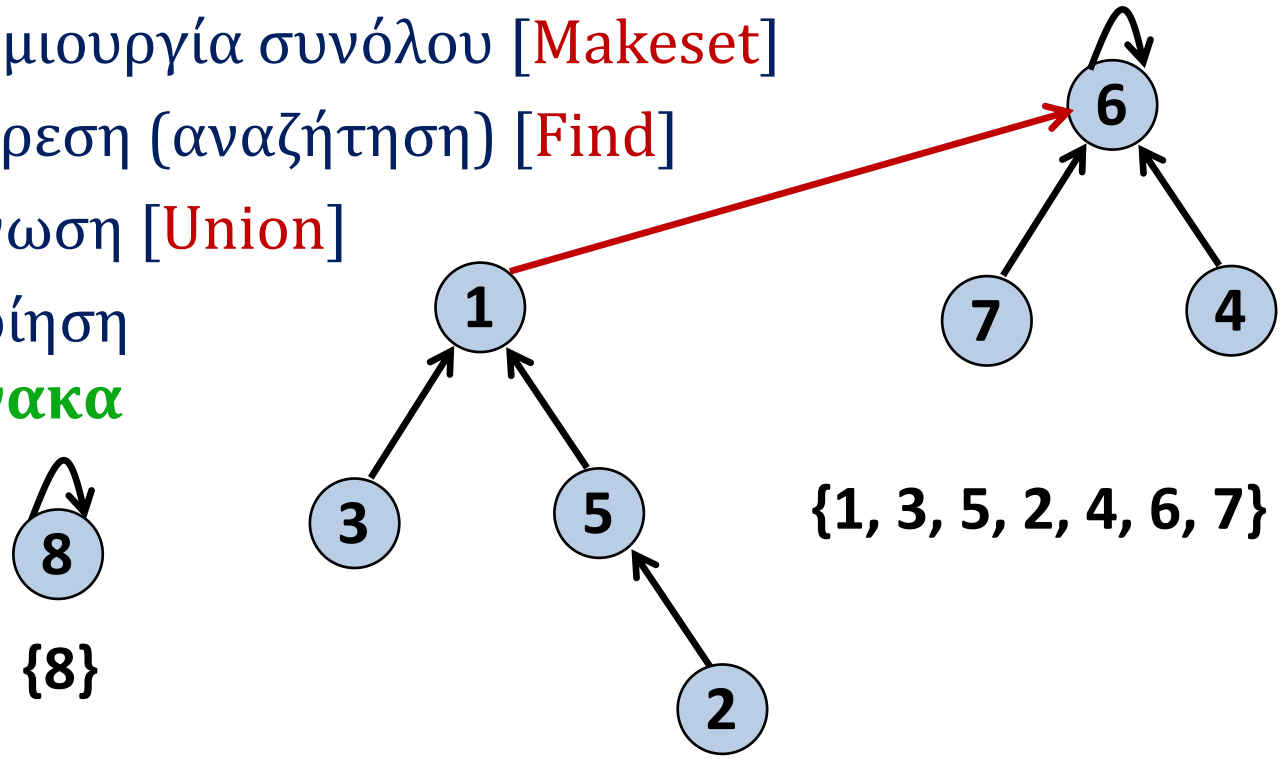
- Υποστηρίζουν πράξεις συνόλων:
  - Δημιουργία συνόλου [Makeset]
  - Εύρεση (αναζήτηση) [Find]
  - Ένωση [Union]
- Υλοποίηση με **πίνακα**



1	2	3	4	5	6	7	8
1	5	1	6	1	6	6	8

Δομές UNION-FIND

- Υποστηρίζουν πράξεις συνόλων:
  - Δημιουργία συνόλου [Makeset]
  - Εύρεση (αναζήτηση) [Find]
  - Ένωση [Union]
- Υλοποίηση με **πίνακα**



1	2	3	4	5	6	7	8
6	5	1	6	1	6	6	8

Ουρά προτεραιότητας (priority queue)

- Δομή για αποθήκευση συνόλου στοιχείων με σχέση προτεραιότητας ανάμεσα στα στοιχεία του (ολική διάταξη)
  - π.χ. προηγούνται τα μεγαλύτερα στοιχεία
- Εφαρμογές:
  - εξυπηρέτηση κατά προτεραιότητα
  - διαχείριση bandwidth
  - ταξινόμηση
  - αλγόριθμος Dijkstra (shortest paths) και αλγόριθμος Prim (minimum spanning tree)
  - κωδικοποίηση Huffman

Ουρά προτεραιότητας (priority queue)

- Δομή για αποθήκευση συνόλου στοιχείων με σχέση προτεραιότητας ανάμεσα στα στοιχεία του (ολική διάταξη)
  - π.χ. προηγούνται τα μεγαλύτερα στοιχεία
- Απαραίτητες λειτουργίες:
  - δημιουργία ουράς
  - εισαγωγή στοιχείου
  - διαγραφή στοιχείου
  - ανάκτηση (και διαγραφή) στοιχείου *μέγιστης προτεραιότητας*

Ουρές προτεραιότητας: απλοϊκή υλοποίηση

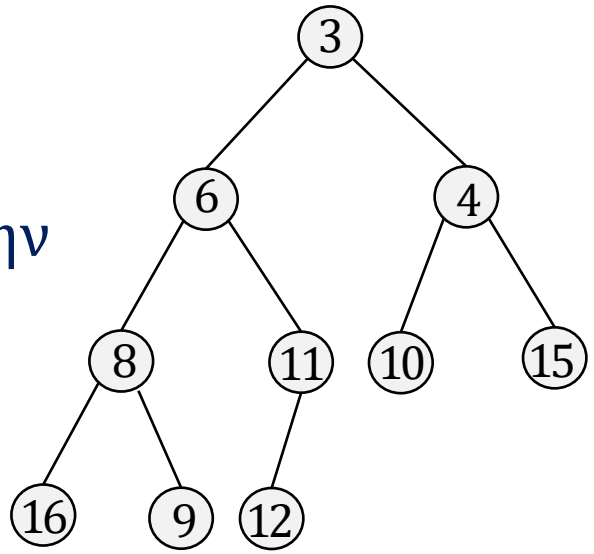
- Με γραμμική δομή (list):
  - εισαγωγή σε χρόνο $O(1)$
  - ανάκτηση (και διαγραφή) στοιχείου μέγιστης προτεραιότητας: χρόνος $O(n)$

Ουρές προτεραιότητας: απλοϊκή υλοποίηση

- Με γραμμική δομή (list):
  - εισαγωγή σε χρόνο $O(1)$
  - ανάκτηση (και διαγραφή) στοιχείου μέγιστης προτεραιότητας: χρόνος $O(n)$
- Εναλλακτικά (list):
  - εισαγωγή σε χρόνο $O(n)$ (για ταξινόμηση)
  - ανάκτηση (και διαγραφή) στοιχείου μέγιστης προτεραιότητας: χρόνος $O(1)$
- Παρόμοιοι χρόνοι με χρήση array

Σωρός (heap)

- Ένας *καλύτερος τρόπος* υλοποίησης ουράς προτεραιότητας
- Δυαδικός σωρός (binary heap):
  - πλήρες δυαδικό δένδρο
  - ελάχιστο (ή μέγιστο) στοιχείο στην ρίζα
  - κάθε υποδένδρο είναι επίσης δυαδικός σωρός
  - απλή υλοποίηση με πίνακα

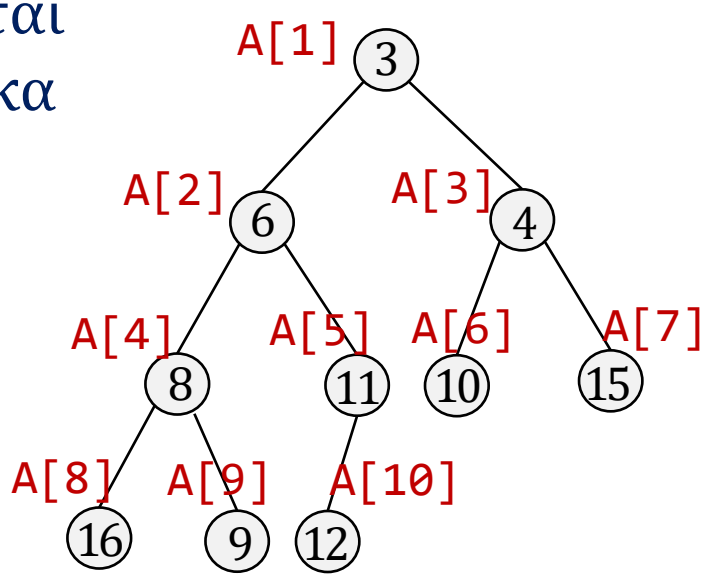


Δυαδικός σωρός ελαχίστου

Υλοποίηση σωρού με πίνακα

- Παιδιά του A[i]: A[$2i$], A[$2i$+1]
- Γονέας του A[i]: A[$i/2$]
- Τα στοιχεία του σωρού γράφονται σε συνεχόμενες θέσεις του πίνακα σε σειρά αντίστοιχη της εξερεύνησης BFS

Δυαδικός σωρός ελαχίστου



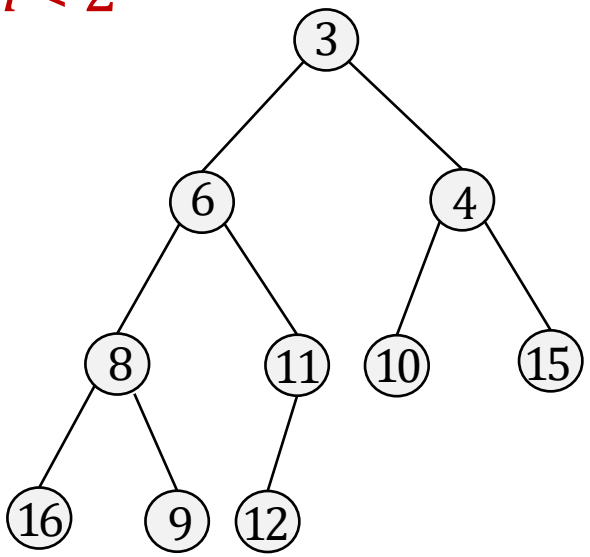
A[1]	A[2]	A[3]	A[4]	A[5]	A[6]	A[7]	A[8]	A[9]	A[10]
3	6	4	8	11	10	15	16	9	12

Σημαντικές ιδιότητες δυαδικών σωρών

- ύψος δένδρου: $\mathbf{O(\log_2 n)}$

  #κόμβων σε δένδρο ύψους h: $2^h \le n < 2^{h+1}$

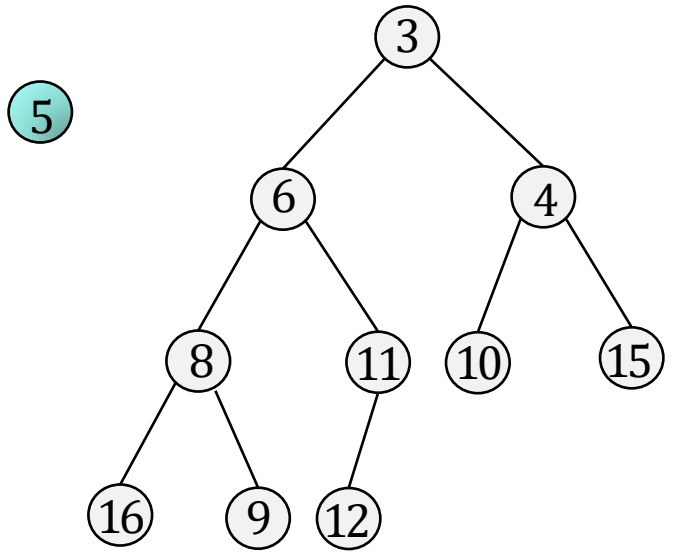
- άμεση ανάκτηση στοιχείου μέγιστης προτεραιότητας: $\mathbf{O(1)}$
- εισαγωγή και διαγραφή σε χρόνο $\mathbf{O(\log_2 n)}$
- τώρα θα δούμε πώς



Δυαδικός σωρός ελαχίστου

Εισαγωγή σε δυαδικό σωρό

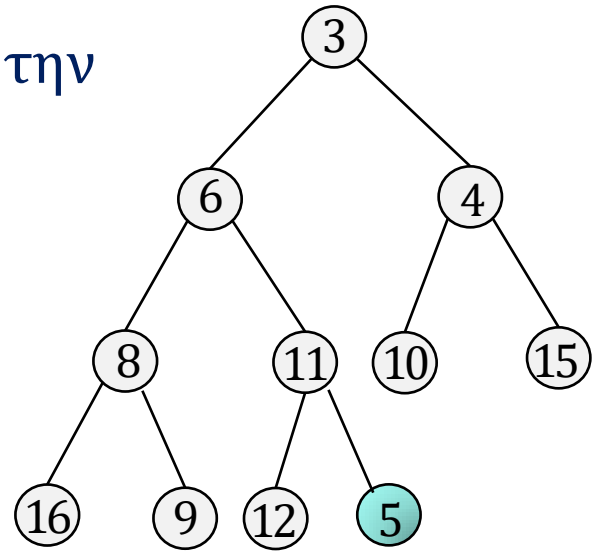
- Έστω ότι θέλουμε να εισαγάγουμε ένα νέο στοιχείο με τιμή 5



Δυαδικός σωρός ελαχίστου

Εισαγωγή σε δυαδικό σωρό

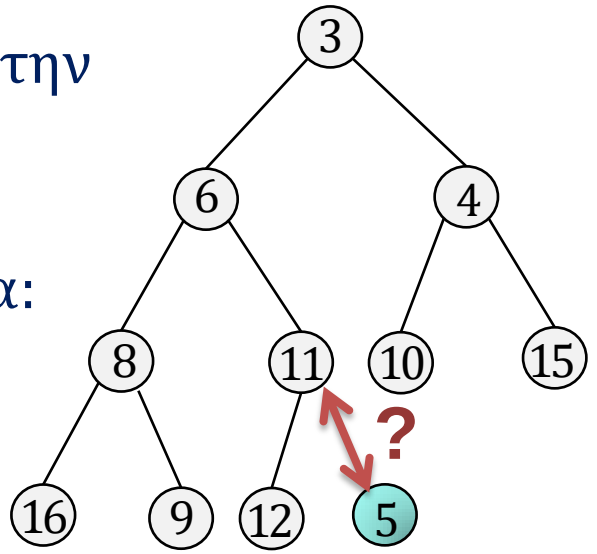
- Έστω ότι θέλουμε να εισαγάγουμε ένα νέο στοιχείο με τιμή 5
- Αρχικά τοποθετούμε το στοιχείο μετά την τελευταία θέση (A[n+1])



Δυαδικός σωρός ελαχίστου

Εισαγωγή σε δυαδικό σωρό

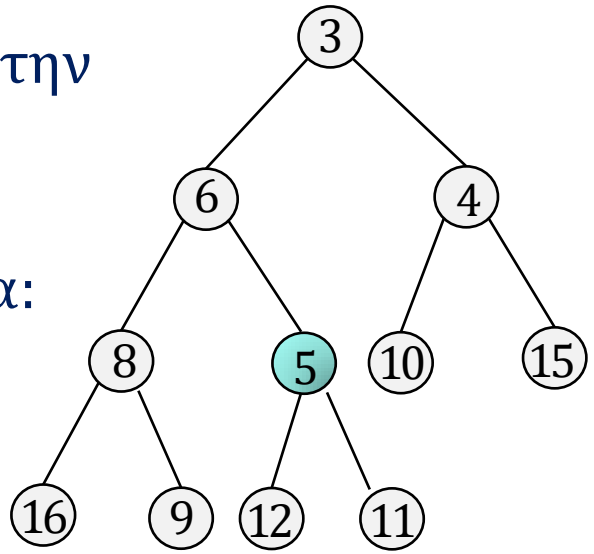
- Έστω ότι θέλουμε να εισαγάγουμε ένα νέο στοιχείο με τιμή 5
- Αρχικά τοποθετούμε το στοιχείο μετά την τελευταία θέση (A[n+1])
- Συγκρίνουμε με τον γονέα και αν χρειάζεται ανταλλάσσουμε τα στοιχεία: `swap(A[i],A[i/2])`



Δυαδικός σωρός ελαχίστου

Εισαγωγή σε δυαδικό σωρό

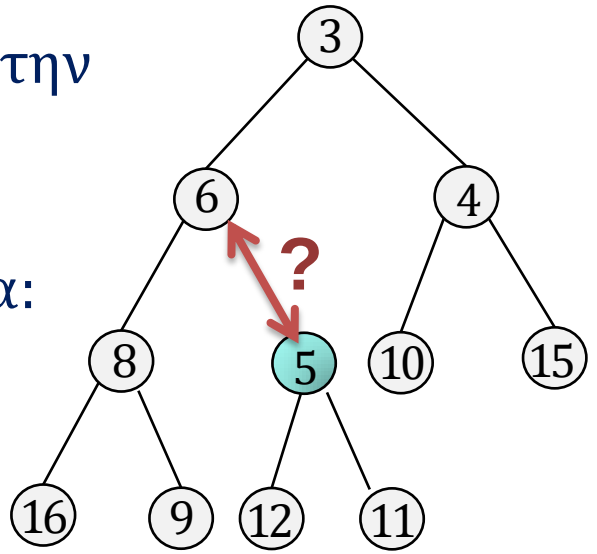
- Έστω ότι θέλουμε να εισαγάγουμε ένα νέο στοιχείο με τιμή 5
- Αρχικά τοποθετούμε το στοιχείο μετά την τελευταία θέση (A[11])
- Συγκρίνουμε με τον γονέα και αν χρειάζεται ανταλλάσσουμε τα στοιχεία: `swap(A[i],A[i/2])`
- Επαναλαμβάνουμε …



Δυαδικός σωρός ελαχίστου

Εισαγωγή σε δυαδικό σωρό

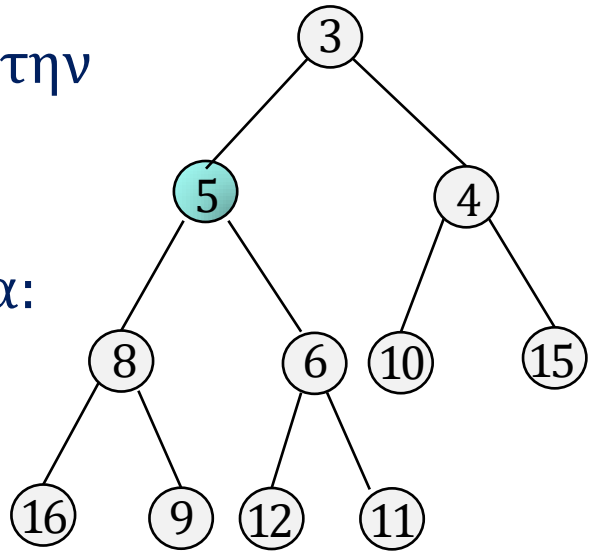
- Έστω ότι θέλουμε να εισαγάγουμε ένα νέο στοιχείο με τιμή 5
- Αρχικά τοποθετούμε το στοιχείο μετά την τελευταία θέση (A[11])
- Συγκρίνουμε με τον γονέα και αν χρειάζεται ανταλλάσσουμε τα στοιχεία: `swap(A[i],A[i/2])`
- Επαναλαμβάνουμε …



Δυαδικός σωρός ελαχίστου

Εισαγωγή σε δυαδικό σωρό

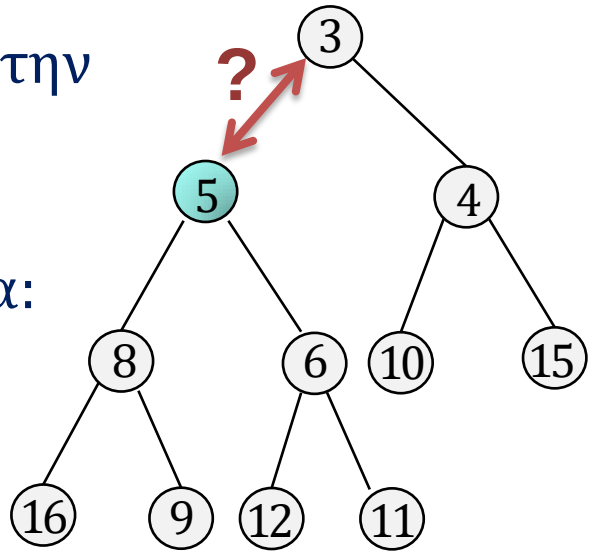
- Έστω ότι θέλουμε να εισαγάγουμε ένα νέο στοιχείο με τιμή 5
- Αρχικά τοποθετούμε το στοιχείο μετά την τελευταία θέση (A[11])
- Συγκρίνουμε με τον γονέα και αν χρειάζεται ανταλλάσσουμε τα στοιχεία: `swap(A[i],A[i/2])`
- Επαναλαμβάνουμε …



Δυαδικός σωρός ελαχίστου

Εισαγωγή σε δυαδικό σωρό

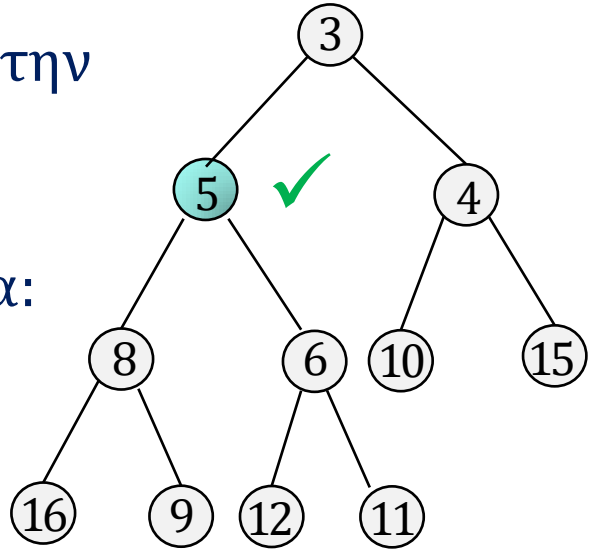
- Έστω ότι θέλουμε να εισαγάγουμε ένα νέο στοιχείο με τιμή 5
- Αρχικά τοποθετούμε το στοιχείο μετά την τελευταία θέση (A[11])
- Συγκρίνουμε με τον γονέα και αν χρειάζεται ανταλλάσσουμε τα στοιχεία: `swap(A[i],A[i/2])`
- Επαναλαμβάνουμε …



Δυαδικός σωρός ελαχίστου

Εισαγωγή σε δυαδικό σωρό

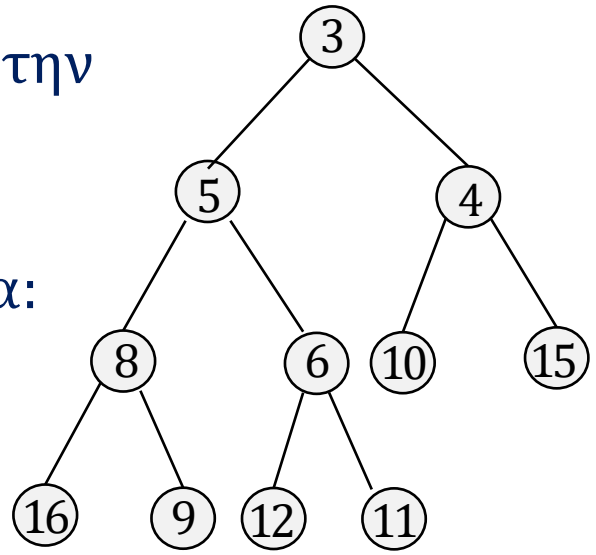
- Έστω ότι θέλουμε να εισαγάγουμε ένα νέο στοιχείο με τιμή 5
- Αρχικά τοποθετούμε το στοιχείο μετά την τελευταία θέση (`A[n+1]`)
- Συγκρίνουμε με τον γονέα και αν χρειάζεται ανταλλάσσουμε τα στοιχεία: `swap(A[i],A[i/2])`
- Επαναλαμβάνουμε …



Δυαδικός σωρός ελαχίστου

Εισαγωγή σε δυαδικό σωρό

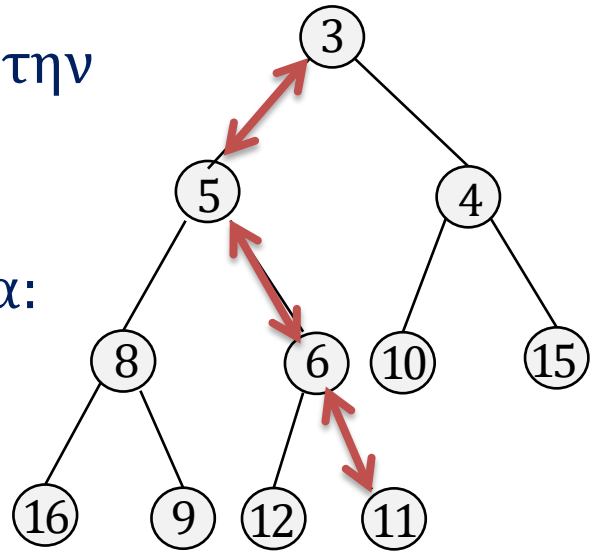
- Έστω ότι θέλουμε να εισαγάγουμε ένα νέο στοιχείο με τιμή 5
- Αρχικά τοποθετούμε το στοιχείο μετά την τελευταία θέση (`A[n+1]`)
- Συγκρίνουμε με τον γονέα και αν χρειάζεται ανταλλάσσουμε τα στοιχεία: `swap(A[i],A[i/2])`
- Επαναλαμβάνουμε …
- Χρόνος εισαγωγής: **O(log *n*)** – γιατί;



Δυαδικός σωρός ελαχίστου

Εισαγωγή σε δυαδικό σωρό

- Έστω ότι θέλουμε να εισαγάγουμε ένα νέο στοιχείο με τιμή 5
- Αρχικά τοποθετούμε το στοιχείο μετά την τελευταία θέση (`A[n+1]`)
- Συγκρίνουμε με τον γονέα και αν χρειάζεται ανταλλάσσουμε τα στοιχεία: `swap(A[i],A[i/2])`
- Επαναλαμβάνουμε …
- Χρόνος εισαγωγής: **O(log *n*)** ~ ύψος δένδρου

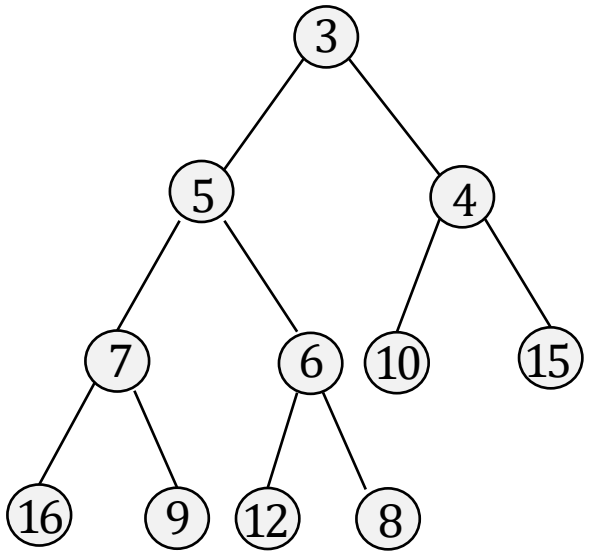


Δυαδικός σωρός ελαχίστου

Μικρή βελτίωση: *δεν χρειάζεται να κάνουμε swap , απλά* `A[i] ← A[i/2]` *ώσπου να "αδειάσει" η σωστή θέση*

Διαγραφή από δυαδικό σωρό

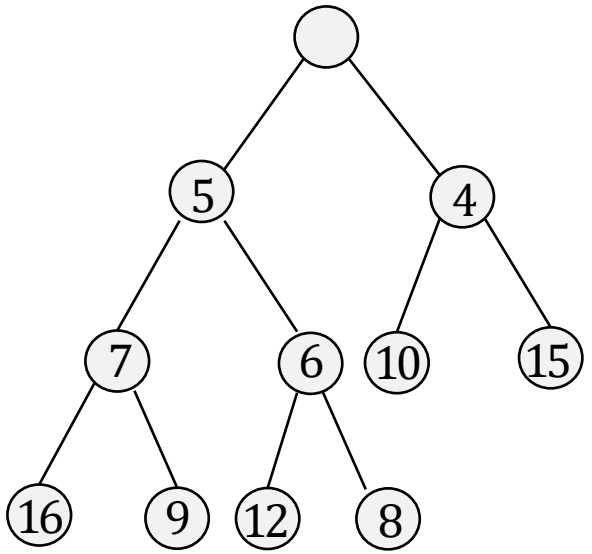
- Έστω ότι θέλουμε να διαγράψουμε το ελάχιστο στοιχείο (ρίζα)



Δυαδικός σωρός ελαχίστου

Διαγραφή από δυαδικό σωρό

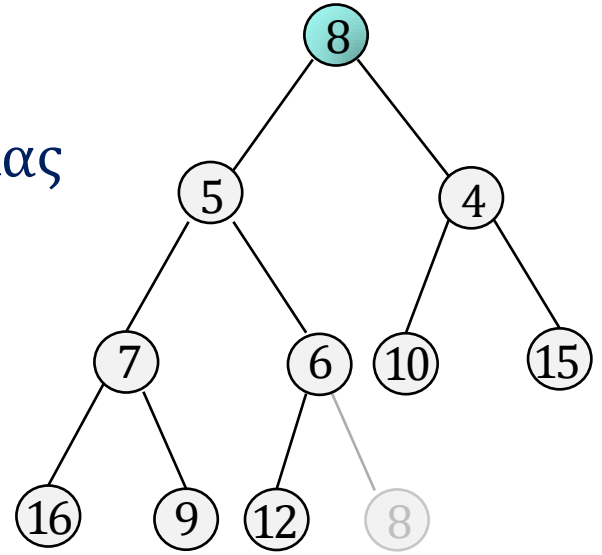
- Έστω ότι θέλουμε να διαγράψουμε το ελάχιστο στοιχείο (ρίζα)
- Η τελευταία θέση «περισσεύει»



Δυαδικός σωρός ελαχίστου

Διαγραφή από δυαδικό σωρό

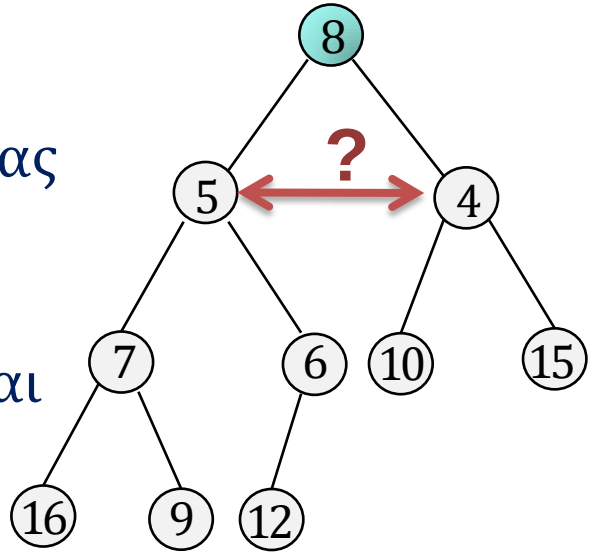
- Έστω ότι θέλουμε να διαγράψουμε το ελάχιστο στοιχείο (ρίζα)
- Η τελευταία θέση «περισσεύει»
- Τοποθετούμε το στοιχείο της τελευταίας θέσης (A[n]) στη ρίζα



Δυαδικός σωρός ελαχίστου

Διαγραφή από δυαδικό σωρό

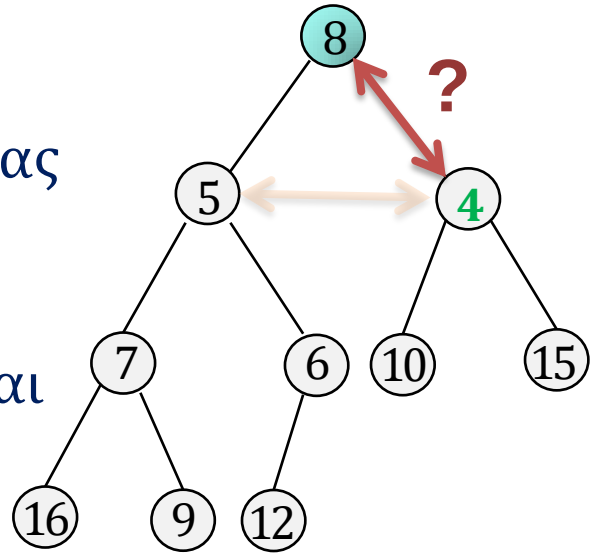
- Έστω ότι θέλουμε να διαγράψουμε το ελάχιστο στοιχείο (ρίζα)
- Η τελευταία θέση «περισσεύει»
- Τοποθετούμε το στοιχείο της τελευταίας θέσης (A[n]) στη ρίζα
- «Κυλάμε» το στοιχείο προς τα κάτω συγκρίνοντας με το μικρότερο παιδί και ανταλλάσσοντας (swap) αν χρειάζεται



Δυαδικός σωρός ελαχίστου

Διαγραφή από δυαδικό σωρό

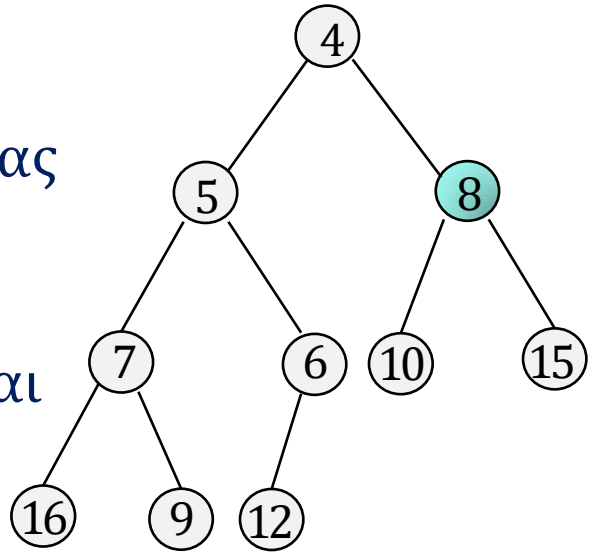
- Έστω ότι θέλουμε να διαγράψουμε το ελάχιστο στοιχείο (ρίζα)
- Η τελευταία θέση «περισσεύει»
- Τοποθετούμε το στοιχείο της τελευταίας θέσης (A[n]) στη ρίζα
- «Κυλάμε» το στοιχείο προς τα κάτω συγκρίνοντας με το μικρότερο παιδί και ανταλλάσσοντας (swap) αν χρειάζεται



Δυαδικός σωρός ελαχίστου

Διαγραφή από δυαδικό σωρό

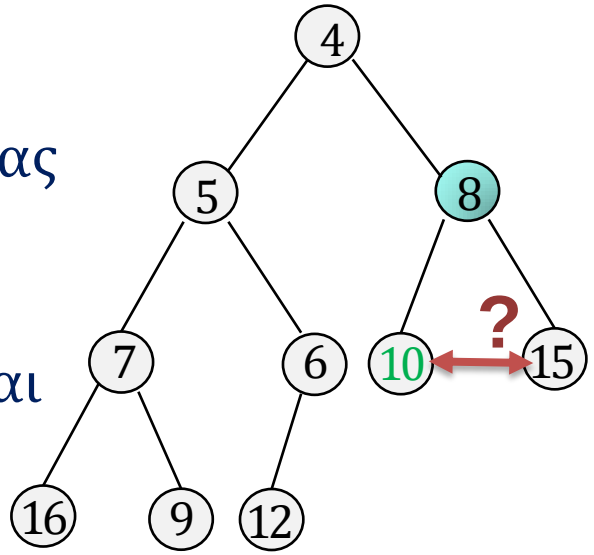
- Έστω ότι θέλουμε να διαγράψουμε το ελάχιστο στοιχείο (ρίζα)
- Η τελευταία θέση «περισσεύει»
- Τοποθετούμε το στοιχείο της τελευταίας θέσης (A[n]) στη ρίζα
- «Κυλάμε» το στοιχείο προς τα κάτω συγκρίνοντας με το μικρότερο παιδί και ανταλλάσσοντας (swap) αν χρειάζεται



Δυαδικός σωρός ελαχίστου

Διαγραφή από δυαδικό σωρό

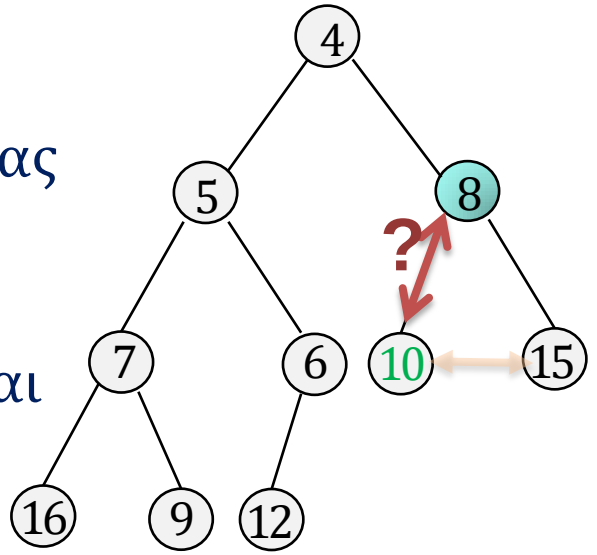
- Έστω ότι θέλουμε να διαγράψουμε το ελάχιστο στοιχείο (ρίζα)
- Η τελευταία θέση «περισσεύει»
- Τοποθετούμε το στοιχείο της τελευταίας θέσης (A[n]) στη ρίζα
- «Κυλάμε» το στοιχείο προς τα κάτω συγκρίνοντας με το μικρότερο παιδί και ανταλλάσσοντας (swap) αν χρειάζεται
- Επαναλαμβάνουμε …



Δυαδικός σωρός ελαχίστου

Διαγραφή από δυαδικό σωρό

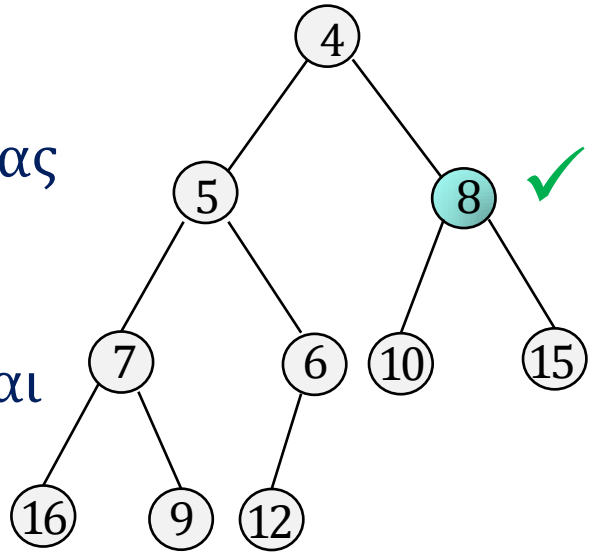
- Έστω ότι θέλουμε να διαγράψουμε το ελάχιστο στοιχείο (ρίζα)
- Η τελευταία θέση «περισσεύει»
- Τοποθετούμε το στοιχείο της τελευταίας θέσης (A[n]) στη ρίζα
- «Κυλάμε» το στοιχείο προς τα κάτω συγκρίνοντας με το μικρότερο παιδί και ανταλλάσσοντας (swap) αν χρειάζεται
- Επαναλαμβάνουμε …



Δυαδικός σωρός ελαχίστου

Διαγραφή από δυαδικό σωρό

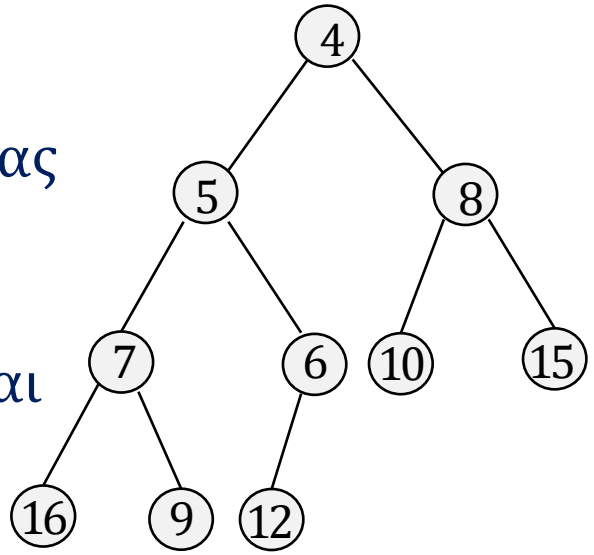
- Έστω ότι θέλουμε να διαγράψουμε το ελάχιστο στοιχείο (ρίζα)
- Η τελευταία θέση «περισσεύει»
- Τοποθετούμε το στοιχείο της τελευταίας θέσης (A[n]) στη ρίζα
- «Κυλάμε» το στοιχείο προς τα κάτω συγκρίνοντας με το μικρότερο παιδί και ανταλλάσσοντας (swap) αν χρειάζεται
- Επαναλαμβάνουμε …



Δυαδικός σωρός ελαχίστου

Διαγραφή από δυαδικό σωρό

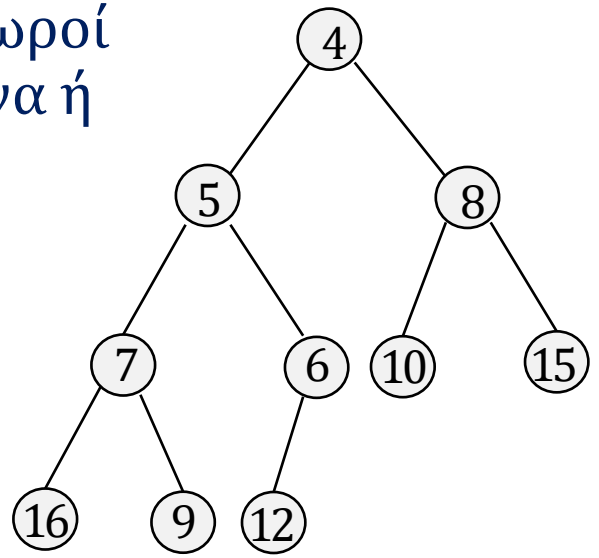
- Έστω ότι θέλουμε να διαγράψουμε το ελάχιστο στοιχείο (ρίζα)
- Η τελευταία θέση «περισσεύει»
- Τοποθετούμε το στοιχείο της τελευταίας θέσης (A[n]) στη ρίζα
- «Κυλάμε» το στοιχείο προς τα κάτω συγκρίνοντας με το μικρότερο παιδί και ανταλλάσσοντας (swap) αν χρειάζεται
- Επαναλαμβάνουμε …
- Χρόνος ανακατασκευής: **O(log *n*)** – γιατί;



Δυαδικός σωρός ελαχίστου

Διαγραφή από δυαδικό σωρό => μέθοδος (ανα)κατασκευής

- Η διαδικασία διαγραφής ουσιαστικά υποδεικνύει διαδικασία **ανακατασκευής** του σωρού όταν τα υποδένδρα είναι σωροί αλλά η ρίζα είναι μεγαλύτερη από το ένα ή και τα δύο παιδιά της
- Αυτή η διαδικασία λέγεται Heapify ή Combine στη βιβλιογραφία και είναι χρήσιμη γενικότερα
- Για παράδειγμα, μπορούμε με χρήση της διαδικασίας αυτής να κατασκευάσουμε σωρό με *n* στοιχεία σε **συνολικό χρόνο O(*n*)** (αντί για **O(*n* log*n*)** με χρήση διαδοχικών εισαγωγών) – *πώς*;
- Ερώτημα: πώς κάνουμε **ενημέρωση** τιμής;



Δυαδικός σωρός ελαχίστου

Ταξινόμηση με σωρό (HeapSort)

- Έστω ότι θέλουμε να ταξινομήσουμε n αριθμούς (ή στοιχεία)
- Τα εισάγουμε σε σωρό με n διαδοχικές εισαγωγές: **$O(n \log n)$** [ή χρησιμοποιούμε Heapify, **$O(n)$**]
- Επαναλαμβάνουμε n φορές:
  - διαγραφή ελαχίστου (και ανακατασκευή σωρού) και εισαγωγή του σε λίστα
- Συνολικός χρόνος διαγραφών: **$O(n \log n)$**
- Πολυπλοκότητα HeapSort: **$O(n \log n)$**

- Εναλλακτικά: **σωρός μεγίστου**, διαγράφουμε διαδοχικά μέγιστο και το εισάγουμε στην τελευταία θέση (ταξινόμηση ***επί τόπου*** **!**)

Κατακερματισμός (Hashing)

- Έστω ότι θέλουμε να καταγράψουμε την εμφάνιση δεδομένων (ή *κλειδιών / κωδικών*) σε ευρετήριο (λεξικό)
- …ώστε να μπορούμε να ελέγξουμε αν κάποιο κλειδί έχει εμφανιστεί στα δεδομένα μας (και πόσες φορές)
- Τα δεδομένα μας έρχονται από ένα σύνολο N δυνατών κλειδιών, όπου το N μπορεί να είναι *πολύ μεγάλο*
- Εκτιμούμε ότι θα λάβουμε $m << N$ κλειδιά
- Πόσο καλά μπορούμε να το κάνουμε;

Κατακερματισμός (Hashing)

Εφαρμογές

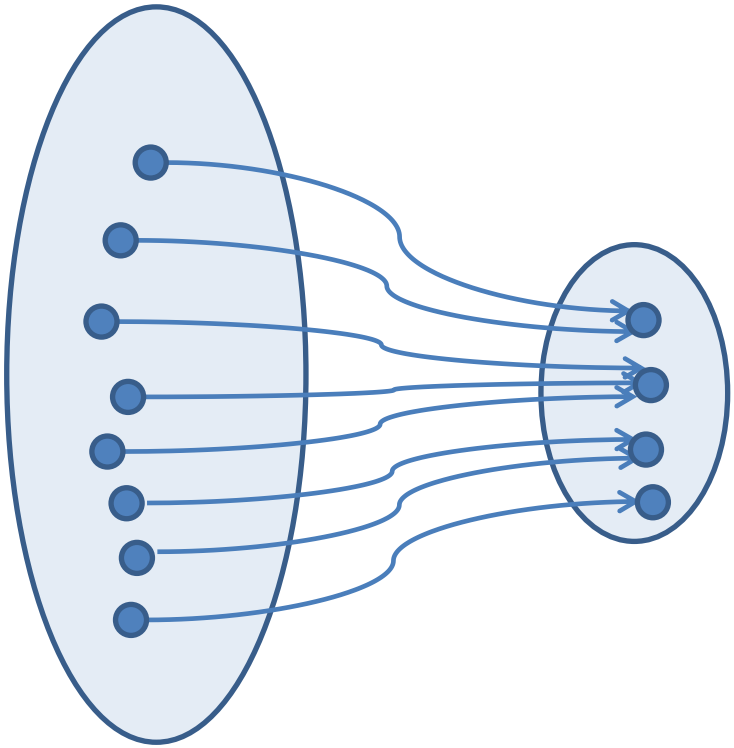
- Καταμέτρηση διαφορετικών ή επαναλαμβανόμενων δεδομένων / γεγονότων από ένα stream (την τελευταία μέρα, εβδομάδα, κ.λπ. ή σε ένα «παράθυρο» k στοιχείων, π.χ. $k=10^6$): παραγγελίες προϊόντων, αλληλεπιδράσεις σε κοινωνικά δίκτυα, αστρονομικές παρατηρήσεις, τιμές μετοχών, κ.λπ.
- Δίκτυα: δρομολόγηση, εξισορρόπηση φορτίου
- Κρυπτογραφία: passwords, ψηφιακές υπογραφές, proof-of-work (blockchains), ακεραιότητα δεδομένων
- Πάρα πολλές διαφορετικές χρήσεις
- Κοινός τόπος: αποφυγή συγκρούσεων (στην κρυπτογραφία εντελώς)

Κατακερματισμός (Hashing)

- Δεδομένα από σύνολο N δυνατών κλειδιών, όπου το N μπορεί να είναι *πολύ μεγάλο*
- Εκτιμούμε ότι θα λάβουμε $m << N$ κλειδιά
- Πόσο καλά μπορούμε να το κάνουμε;
- Θέλουμε γρήγορη καταχώρηση και ανάκτηση, ιδανικά **O(1)**
- Εφικτό με χρήση array – αλλά «σπάταλο» σε μνήμη (ιδέες από **bucket sort**)
- Εφικτό με λίγη μνήμα, με χρήση **συναρτήσεων κατακερματισμού** (hash functions)

Κατακερματισμός (Hashing)

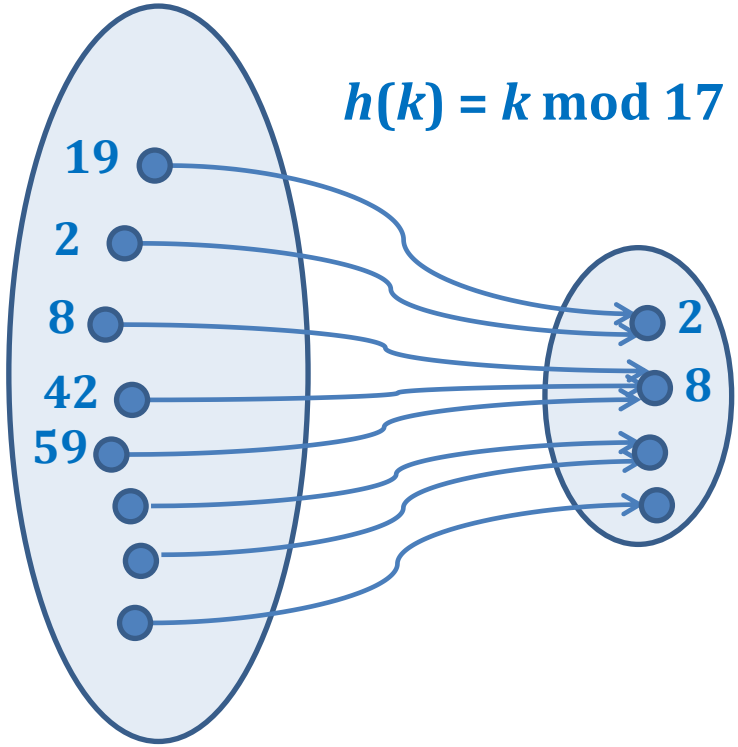
- Απεικόνιση ενός μεγάλου συνόλου σε ένα πολύ μικρότερο
- Όσο το δυνατόν πιο ομοιόμορφα
- Στόχος: ελαχιστοποίηση συγκρούσεων
- (Διαφορά με κρυπτο-γραφικό hashing: εκεί **απαγορεύονται** οι συγκρούσεις)



Κατακερματισμός (Hashing)

Μέθοδος διαίρεσης

- Απλή μορφή: $h(k) = k \bmod m$



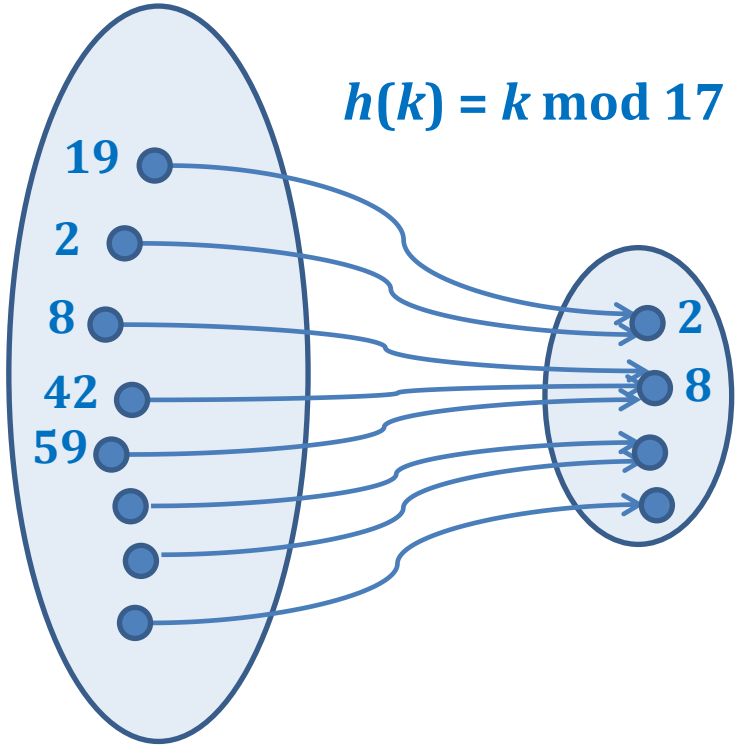
Κατακερματισμός (Hashing)

Μέθοδος διαίρεσης

- Απλή μορφή: $h(k) = k \bmod m$
- Μειονέκτημα: δεν έχει *ιδιότητα* ***καθολικότητας***

Καθολική (universal) συνάρτηση κατακερ/σμού:

απεικονίζει διαφορετικά κλειδιά στην ίδια θέση με πιθανότητα $\leq 1/m$



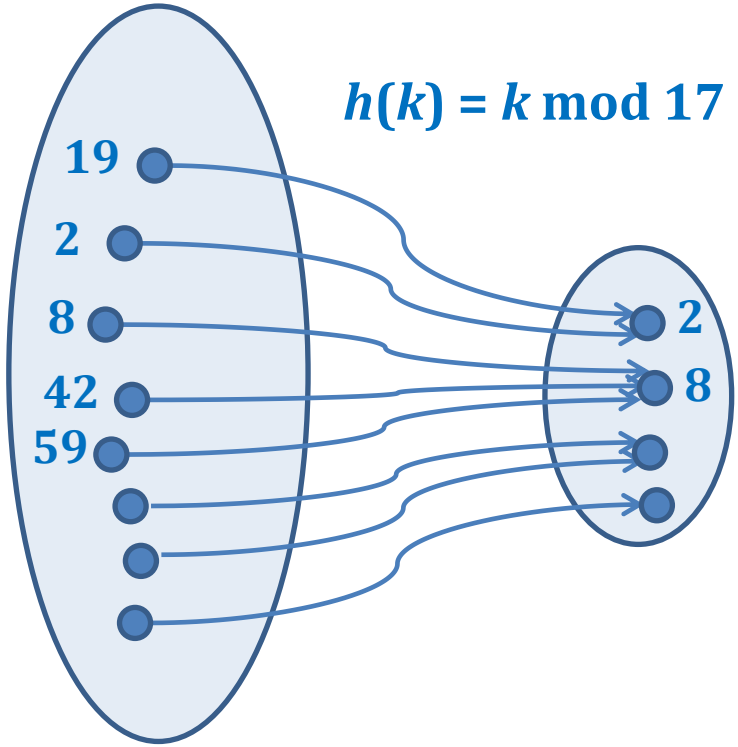
Κατακερματισμός (Hashing)

Μέθοδος διαίρεσης

- Απλή μορφή: $h(k) = k \bmod m$
- Μειονέκτημα: δεν έχει *ιδιότητα* ***καθολικότητας***

  Καθολική (universal) συνάρτηση κατακερ/σμού:

  απεικονίζει διαφορετικά κλειδιά στην ίδια θέση με πιθανότητα $\leq 1/m$
- ***Αδύνατο*** για μία μόνο συνάρτηση! (*γιατί*;)



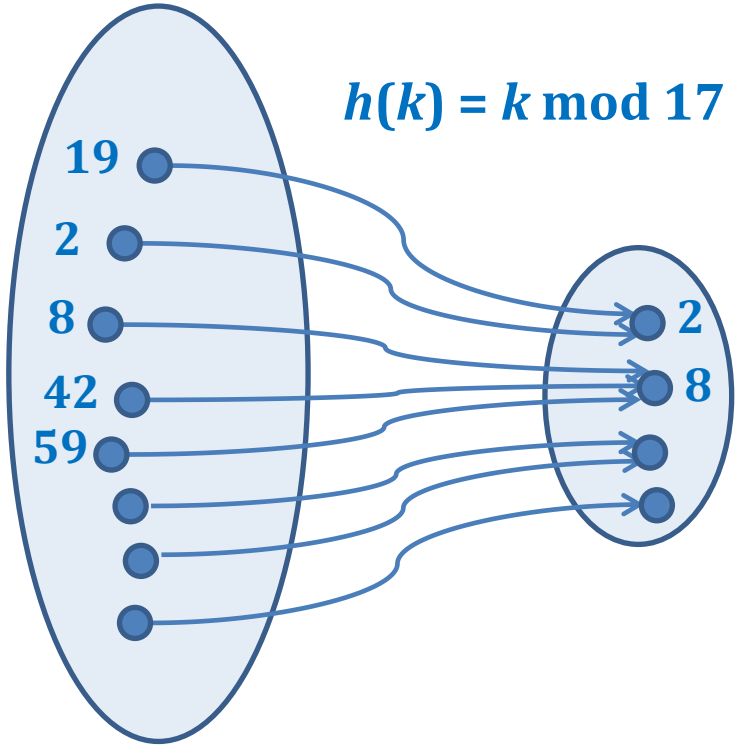
Κατακερματισμός (Hashing)

Μέθοδος διαίρεσης

- Απλή μορφή: $h(k) = k \bmod m$
- Μειονέκτημα: δεν έχει *ιδιότητα* ***καθολικότητας***

  Καθολική (universal) συνάρτηση κατακερ/σμού:

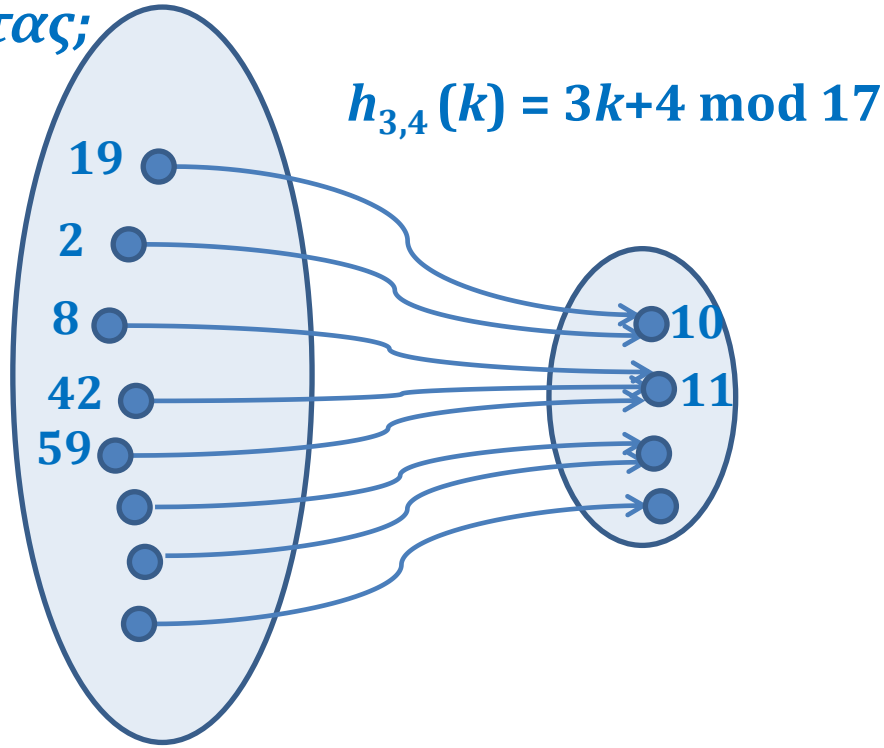
  απεικονίζει διαφορετικά κλειδιά στην ίδια θέση με πιθανότητα $\leq 1/m$
- ***Αδύνατο*** για μία μόνο συνάρτηση: *ντετερμινισμός!*



Κατακερματισμός (Hashing)

Μέθοδος διαίρεσης

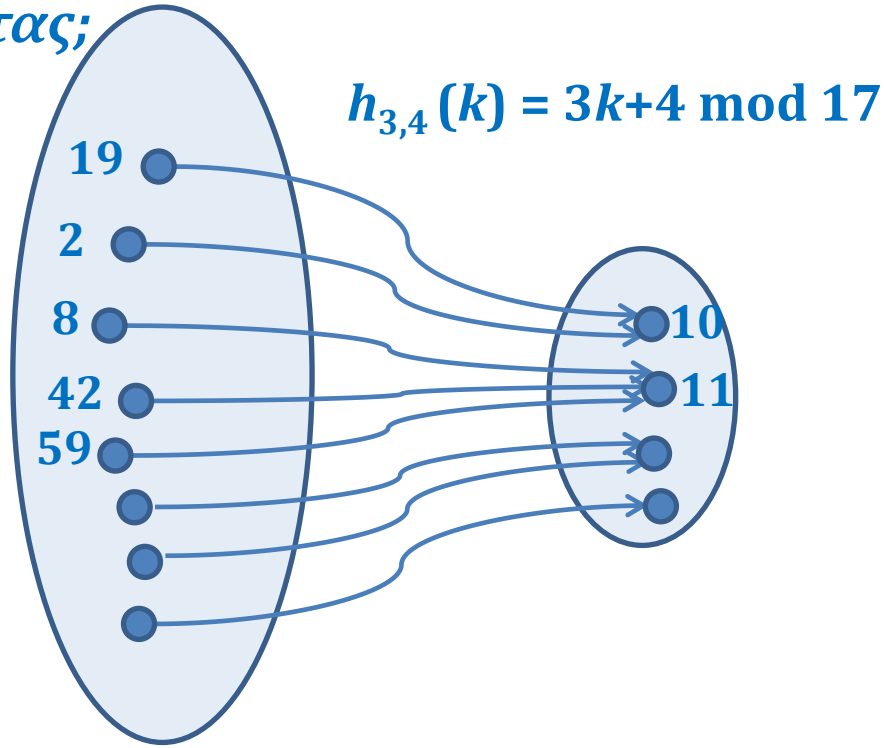
- Οικογένεια συναρτήσεων:
  $h_{a,b}(k) = ak + b \bmod m$
- Έχει *ιδιότητα* ***καθολικότητας;***



Κατακερματισμός (Hashing)

Μέθοδος διαίρεσης

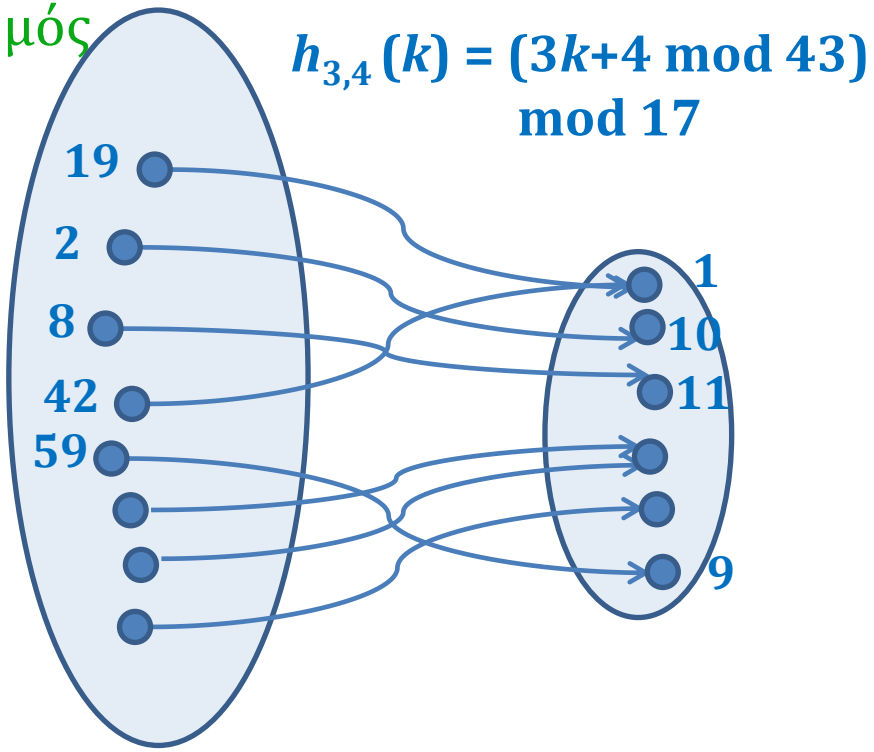
- Οικογένεια συναρτήσεων:
  $h_{a,b}(k) = ak + b \bmod m$
- Έχει *ιδιότητα* ***καθολικότητας;***
- Όχι! (*γιατί*;)



Κατακερματισμός (Hashing)

Μέθοδος διαίρεσης

- Καθολική οικογένεια συναρτήσεων:
  $h_{a,b}(k) = (ak + b \bmod p) \bmod m$
- $p > |U| >> m$, p πρώτος αριθμός
- Έχει *ιδιότητα* ***καθολικότητας!***
- *Απόδειξη*: άσκηση (προαιρετική)

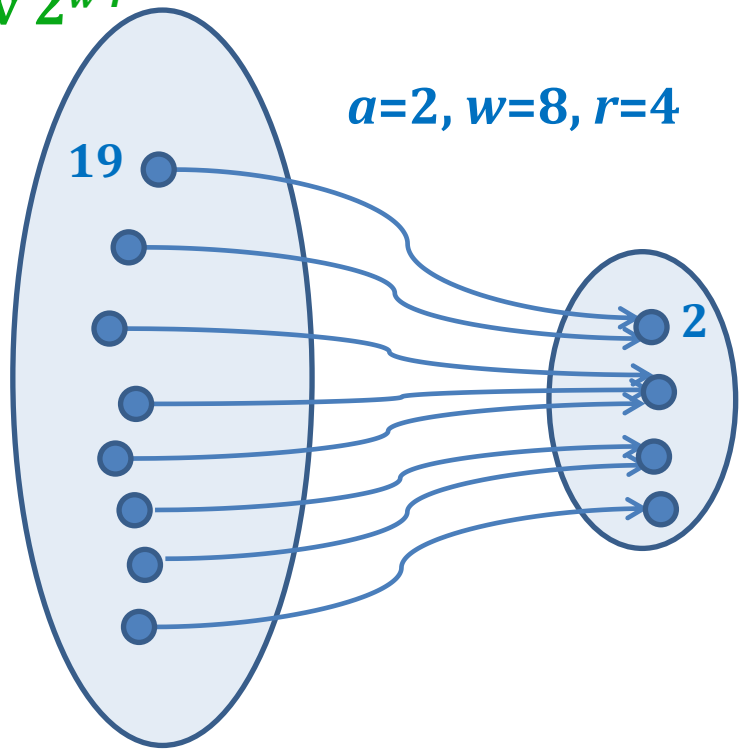


Κατακερματισμός (Hashing)

Μέθοδος πολλαπλασιασμού

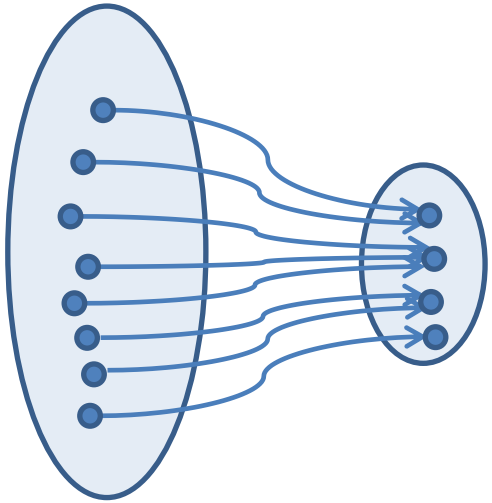
- Απλή μορφή: $h(k) = (ak \bmod 2^w) \text{ div } 2^{w-r}$ $[\, m = 2^r \,]$
- Οικογένεια συναρτήσεων:
  $h(k) = (ak + b \bmod 2^w) \text{ div } 2^{w-r}$
- Πλεονεκτήματα:
  - εύκολη υλοποίηση
  - *ιδιότητα* ***2-καθολικότητας***

    ομ/φα τυχαία επιλογή a, b, απεικονίζει διαφορετικά κλειδιά στην ίδια θέση με πιθανότητα $\leq 2/m$



Κατακερματισμός (Hashing)

- **Άμεση / κλειστή διευθυνσιοδότηση** (direct / closed addressing)
  - Κάθε κλειδί απεικονίζεται στην ίδια πάντα θέση
  - Συγκρούσεις αντιμετωπίζονται με **αλυσίδωση**



  - Μέθοδος διαίρεσης, μέθοδος πολλαπλασιασμού
  - Σημαντικό μέγεθος: **παράγοντας φόρτου**

$$a = n / m > 1$$

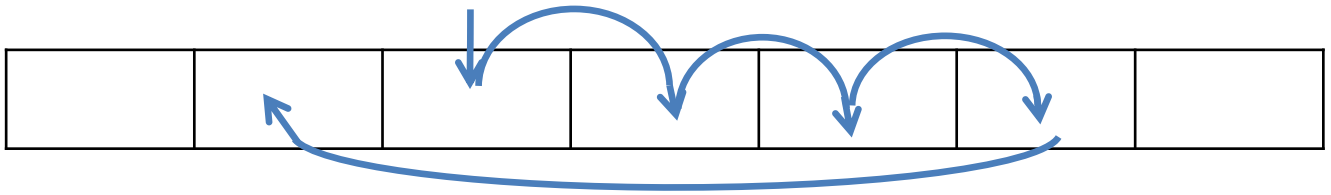
  - Πολυπλοκότητα εισαγωγής, αναζήτησης, διαγραφής: **O(1+*a*)** [αναμεν/νος # βημάτων]

Κατακερματισμός (Hashing)

- **Ανοιχτή διευθυνσιοδότηση** (open addressing)
  - Αποφυγή συγκρούσεων με κατάλλη αλλαγή θέσης
  - Γραμμική και τετραγωνική διερεύνηση, διπλός κατακερματισμός, μέθοδος «κούκου»
  - Σημαντικό μέγεθος: **παράγοντας φόρτου**

    $$a = n / m < 1$$

  - Πολυπλοκότητα εισαγωγής, αναζήτησης, διαγραφής:
    O(1/(1-*a*)) [αναμ/νος # δοκιμών για ομοιόμορφα τυχαία διερεύνηση (διπλός κατακερματισμός την προσεγγίζει καλά)]
    O(1) για σταθερό *a* (π.χ. ≤2 για *a*≤ 0.5)



Σύνοψη

Είδαμε 3 μεθόδους διαχείρισης δεδομένων:

Δομές Union-Find για ομαδοποίηση δεδομένων σε σύνολα με υποστήριξη ταχείας *ένωσης συνόλων και εύρεσης συνόλου* που ανήκει κάθε στοιχείο

Ουρές προτεραιότητας για διαχείριση δεδομένων με τιμές προτεραιότητας και υποστήριξη ταχείας *ανάκτησης στοιχείου μέγιστης προτεραιτότητας*

Συναρτήσεις κατακερματισμού για συνοπτική αποτύπωση δεδομένων με υποστήριξη ταχέος *ελέγχου (επαν)εμφάνισης* στοιχείων