Advanced Indexing Techniques in Big Data

Δρ. Τιάκας Ελευθέριος

Σχετικά με το μάθημα...

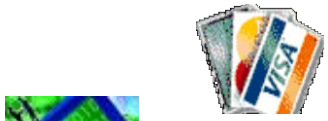
- email: tiakas@csd.auth.gr
- Δικτυακός Τόπος Μαθήματος: http://delab.csd.auth.gr/~tiakas/Advanced-Indexing.htm
- Ώρες Μαθήματος:
  - Παρασκευή 15:30-18:00 (Αίθουσα 1, Forum)
- Αξιολόγηση:
  - Μικρές Εργασίες (Προγραμματιστικές)
  - Ερευνητική Εργασία με Παρουσίαση
  - Γραπτή Εξέταση στη Θεωρία

Ενδεικτική Ύλη Μαθήματος

- Ιεραρχία Μνήμης - Ο Σκληρός Δίσκος
- Μοντέλα Δευτερεύουσας Μνήμης - Parallel Disk Models
- I/O-αποδοτικοί Αλγόριθμοι και Δομές Δεδομένων
- Cache-Oblivious Αλγόριθμοι και Δομές Δεδομένων
- Weight-balanced και Persistent B-trees, Buffer Trees
- Hashing (Chains, Perfect, Universal, Hash Functions)
- Τεχνικές Δόμησης Δεδομένων σε δίκτυα (P2P)
- Αλγόριθμοι και Δομές Δεδομένων για Ροές
- Δομές για Μετρικούς Χώρους

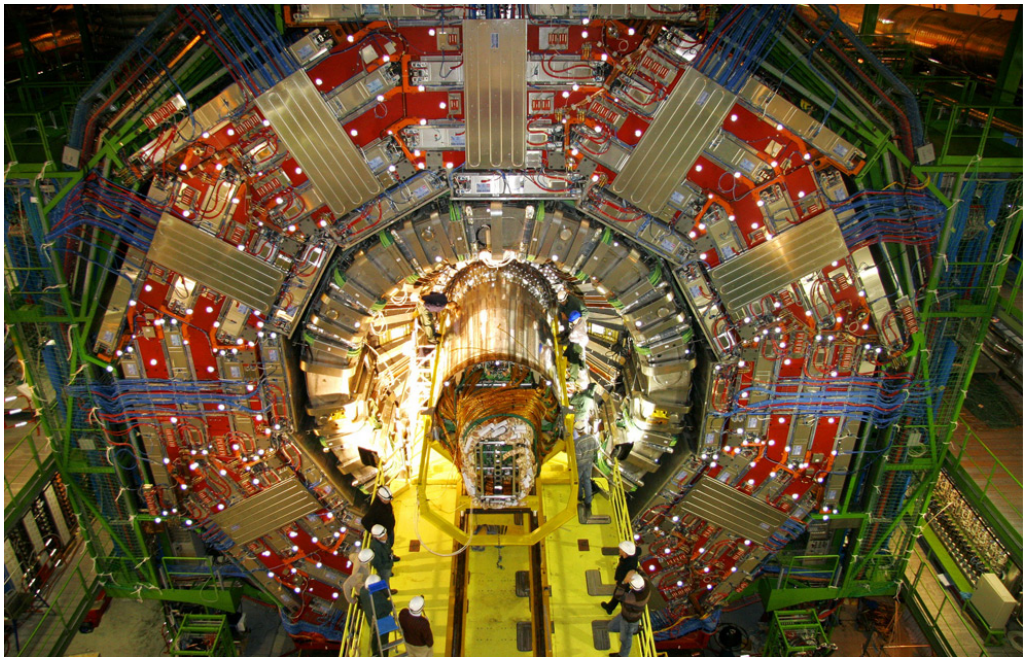
Μεγάλος Όγκος Δεδομένων Παντού!

- Συλλογή Μεγάλου Όγκου Δεδομένων από:
  - Διαδίκτυο
  - Ηλεκτρονικό Εμπόριο
  - Τραπεζικές συναλλαγές
  - Κοινωνικά Δίκτυα
  - Δορυφόροι, GPS
  - Remote Sensors



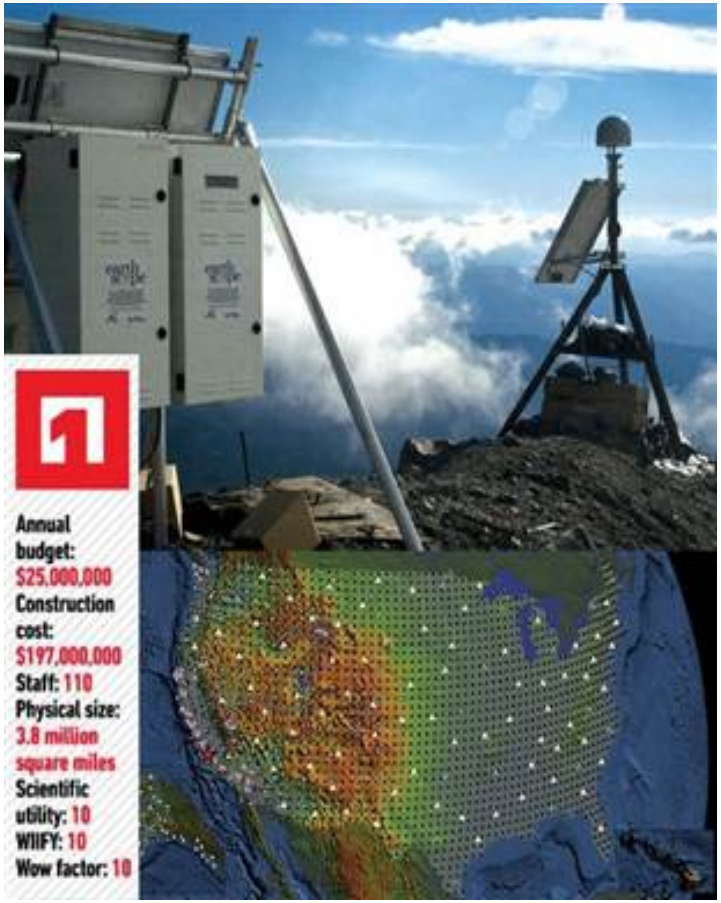
Πόσο Μεγάλος;

- Η Google το 2008 επεξεργαζόταν 20 PB(10^{15}B)την ημέρα
- Το Facebook το 2009 είχε 2.5 PB δεδομένα χρηστών με αύξηση κατά 15 TB/ημέρα
- Το eBay το 2009 είχε 6.5 PB δεδομένα χρηστών με αύξηση κατά 50 TB/ημέρα
- Το Large Hydron Collider (LHC) του CERN δημιουργεί 15 PB δεδομένα κάθε χρόνο



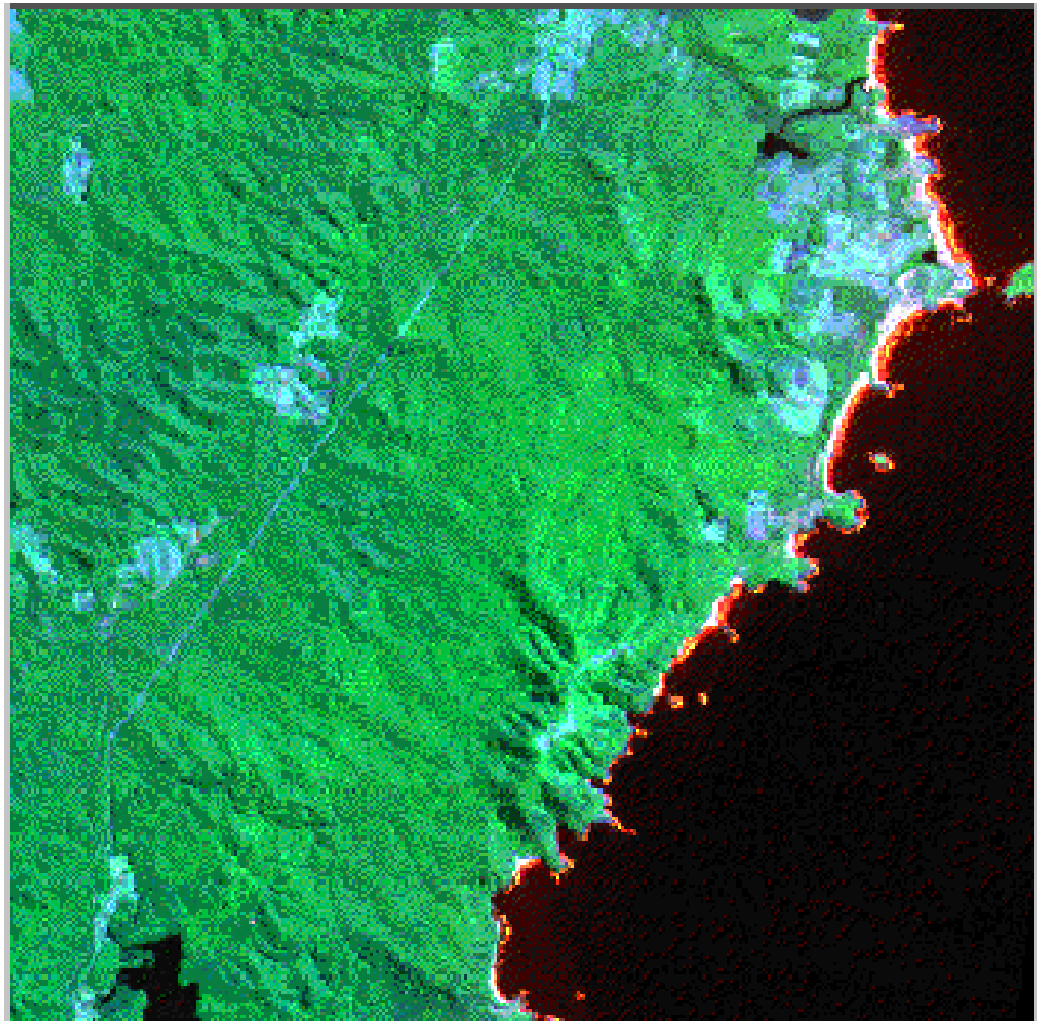
Project Earthscope

- Είναι ένα από τα μεγαλύτερα παγκόσμια επιστημονικά project
- Σχεδιάστηκε για να καταγράφει τη σεισμική δραστηριότητα της Βορείου Αμερικής
- Καταγράφει τα δεδομένα μίας έκτασης πάνω από 3,8 εκατομμυρίων τετραγωνικών μιλίων, σε 67 terabytes κάθε στιγμή
- Αναλύει τη σεισμικές τάσεις του ρήγματος του Άγιου Ανδρέα, την ποσότητα του μάγματος κάτω από το Yellowstone και πολλά άλλα



Μεγάλος Όγκος Δορυφορικών Δεδομένων

Το 1980 η ακρίβεια ήταν 1 pixel = 30m x 30m



Το 2000 η ακρίβεια ήταν 1 pixel = 2.5m x 2.5m

Airborne Sensor Platforms

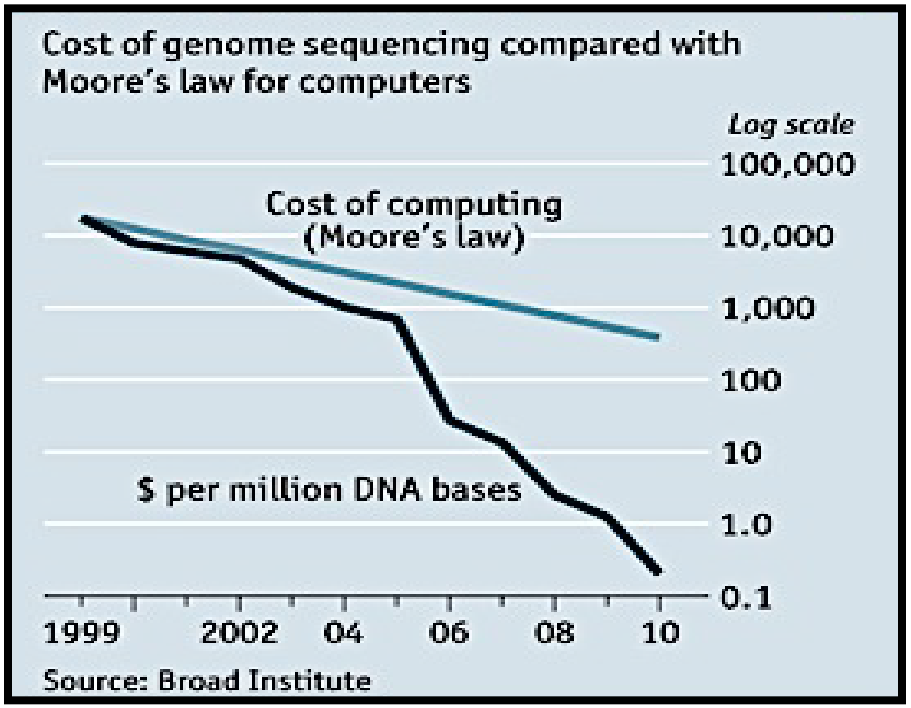
(φθηνότερα και πιο ευέλικτα από τους δορυφόρους)

Remote Sensing Platforms

(μη επανδρωμένα)

Πάλι Μεγάλοι Όγκοι

- Το σύστημα της NASA Earth Observation System (EOS) το 2010 είχε καταγράψει 4.2 PB δεδομένων
- Μάλιστα αυτός ο όγκος δεν αποτελείται από δεδομένα δορυφορικών εικόνων (που είναι πολύ περισσότερα) αλλά δεδομένα τοπογραφικών μοντέλων και χαρτών
- Τα δεδομένα ανάλυσης γονιδιωμάτων διπλασιάζονται σχεδόν κάθε 6 μήνες



Τύποι Δεδομένων

- Σχεσιακά Δεδομένα (Tables/Transaction/Legacy Data)
- Δεδομένα Κειμένων και Πολυμεσικά (Web)
- Ημι-δομημένα Δεδομένα (XML)
- Δεδομένα Γράφων
  - Social Network, Semantic Web (RDF), …
- Δεδομένα Ροών
- κλπ.

Σε τι τα χρησιμοποιούμε;

- Αποθήκευση, Συγκέντρωση, Στατιστική Ανάλυση
  - Data warehouse and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Ανακάλυψη Γνώσης (Knowledge discovery)
  - Data Mining
  - Statistical Modeling
- Οπτικοποίηση και Ανάλυση

Όμως Προσοχή

- Η γνώση συνεχώς εξελίσσεται και αλλάζει
- Η ακρίβεια και η αντικειμενικότητα των δεδομένων είναι πολλές φορές σχετική
- Τα περισσότερα δεδομένα δεν σημαίνουν απαραίτητα και καλύτερα δεδομένα
- Δεν είναι όλα τα δεδομένα ισάξια
- Δεν πρέπει να είναι όλα τα δεδομένα προσβάσιμα από όλους

Πως αλλιώς συγκεντρώνονται;

- Από web browsers και μηχανές αναζήτησης
- Από Smartphones, Tablets και τις εφαρμογές τους
- Από συστήματα GPS και Internet Service Providers
- Από σύγχρονες συσκευές ψηφιακής τεχνολογίας (τηλεοράσεις, κινητά τηλέφωνα κλπ.)

Προκλήσεις στη Διαχείριση Μεγάλου Όγκου Δεδομένων

- Περιορισμένη πρόσβαση στα δεδομένα
- Χρόνος προσπέλασης των δεδομένων
- Ελάχιστο μέρος των δεδομένων μπορεί να παραμείνει στην κύρια μνήμη
- Πολυπλοκότητα των δεδομένων
- Χρόνοι μεταφοράς και επεξεργασίας των δεδομένων:

Προκλήσεις στη Διαχείριση Μεγάλου Όγκου Δεδομένων

Για παράδειγμα 1 TB δεδομένων απαιτεί:

Disks	MB / s	Time
Seagate Barracuda	78	3.6 hours
Seagate Cheetah	125	2.2 hours
Networks	**MB / s**	**Time**
Home Internet	< 0.625	> 18.5 days
Gigabit Ethernet	< 125	> 2.2 hours
PSC TeraGrid Connection	< 3,750	> 4.4 minutes

Λύση οι Υπερυπολογιστές;

- Παράλληλα συστήματα βάσεων δεδομένων
  - Προτάθηκαν στα τέλη του ‘80
  - Ωρίμασαν τις τελευταίες δύο δεκαετίες
  - Ιδιωτικές Εταιρίες προσέφεραν αποδοτικά συστήματα για αποθήκες μεγάλων βάσεων δεδομένων επί πληρωμή
- Map Reduce
  - Χρησιμοποιήθηκε από την Google (20PB/ημέρα)
  - Χρησιμοποιήθηκε πιο πολύ ως open-source Hadoop

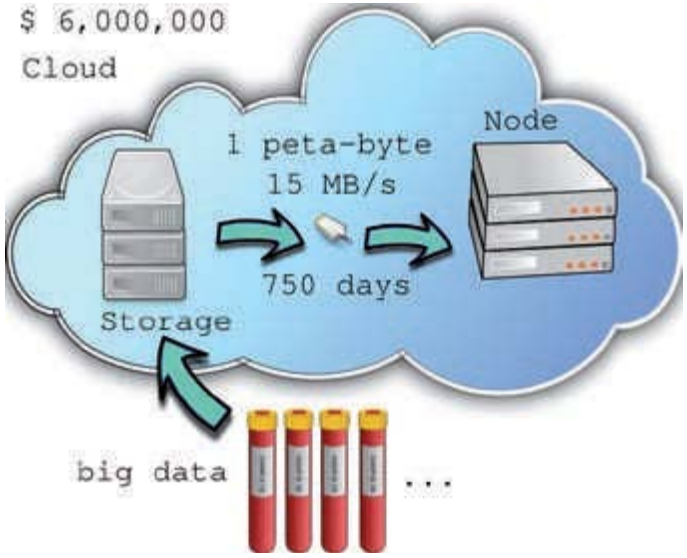
Πλεονεκτήματα

- Παράλληλα συστήματα βάσεων δεδομένων
  - Ώριμη Τεχνολογία
  - Υποστήριξη Σχεσιακού Μοντέλου, Indexing, SQL
  - Advanced query optimization
- Map Reduce
  - Αυτόματος παράλληλος προγραμματισμός
  - Τμηματοποίηση δεδομένων και χειρισμός προβλημάτων επικοινωνίας και λειτουργίας
  - Απόλυτη διαφάνεια στον προγραμματιστή/χρήστη

Όμως, μία ιδέα για το κόστος

Τι κόστος έχει η επεξεργασία 1 Petabyte δεδομένων σε cloud σύστημα με 1000 κόμβους;

- 1 PB = 10^{15} bytes = 1 million gigabytes
- Κάθε κόμβος σε 9 ώρες επεξεργάζεται 500GB με ταχύτητα 15MB/sec
- Απαιτούνται 750 ημέρες
- Συνολικό κόστος: 6,12M $



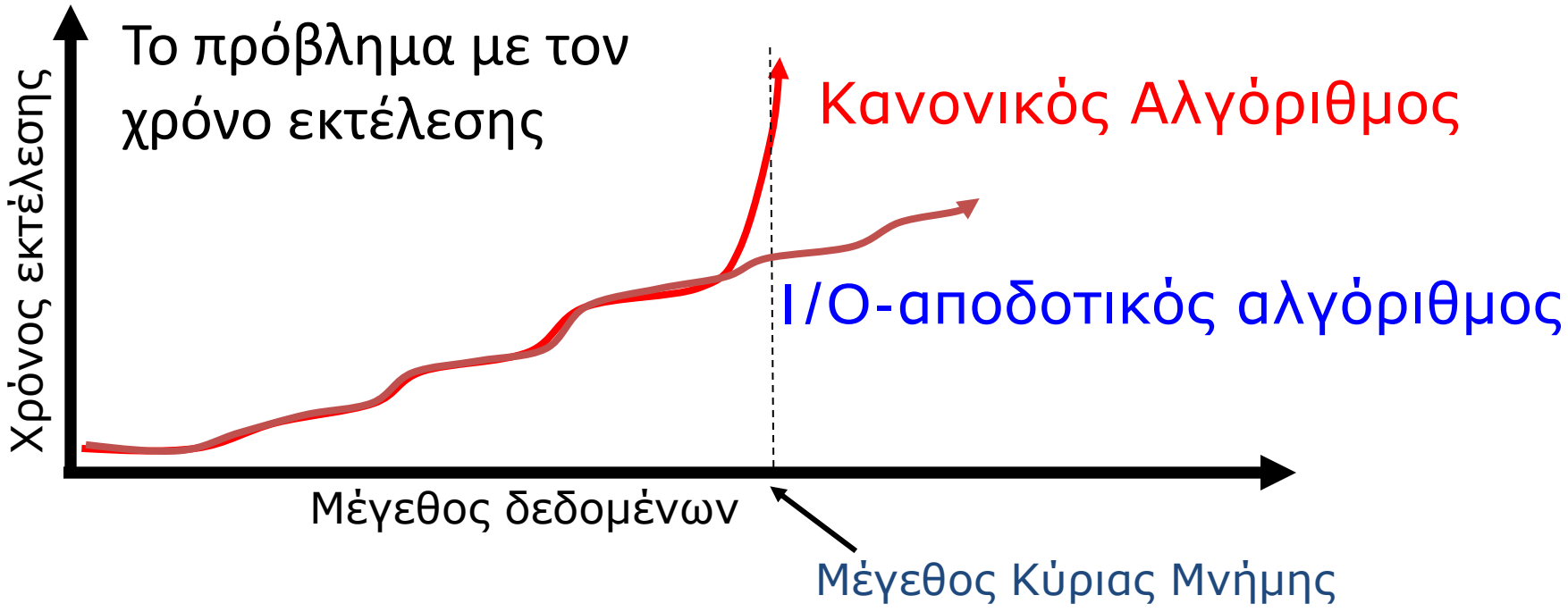
Στόχοι

- Η ανάπτυξη αλγοριθμικών τεχνικών και μοντέλων που να λαμβάνουν υπόψη τους παραπάνω περιορισμούς και κυρίως τον τεράστιο όγκο των δεδομένων
- Να είναι σχετικά απλές και γενικές ώστε να καλύπτουν όσο το δυνατόν περισσότερες περιπτώσεις
- Να μπορούν να εφαρμοστούν στην πράξη

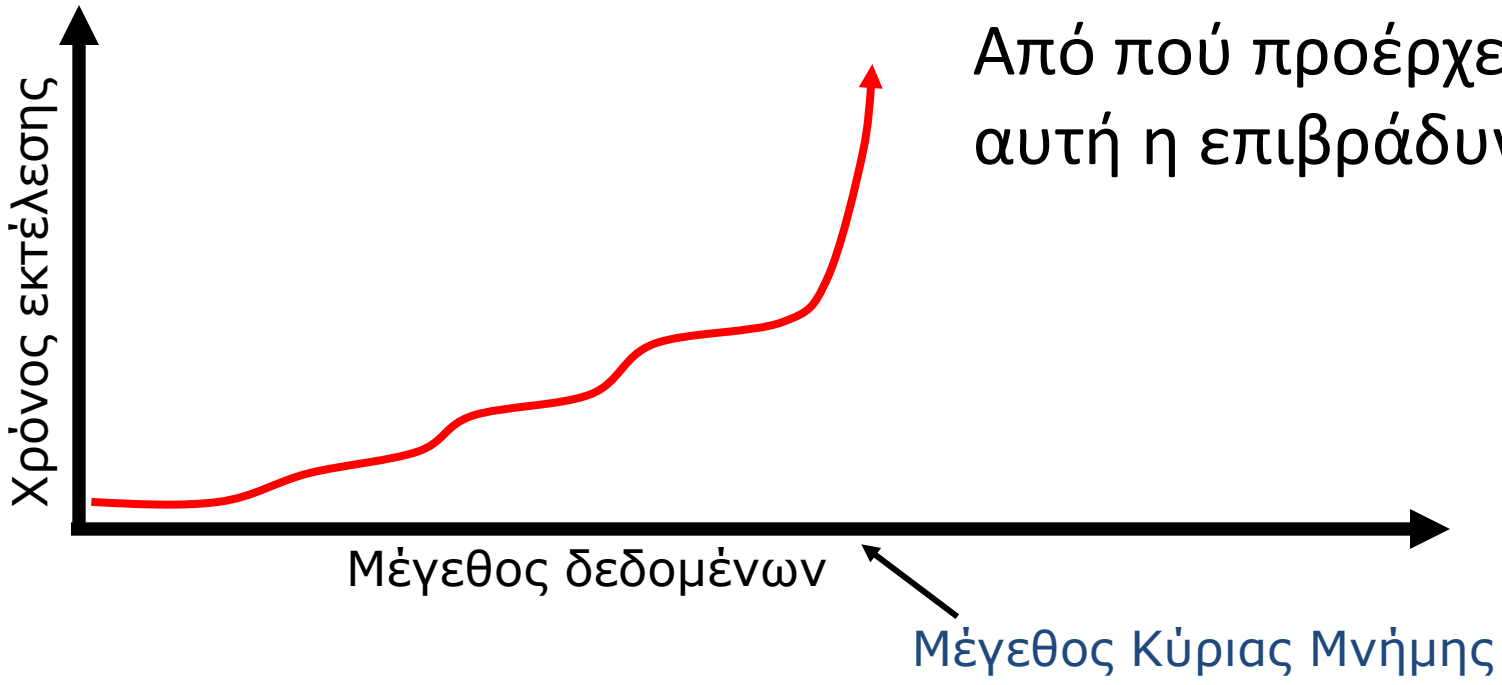
Κάποια Βασικά Στοιχεία

- Η επίδραση της δευτερεύουσας μνήμης και η λειτουργία των Cache
- Ορισμένοι περιορισμοί σε δίκτυα
- Peer-to-Peer δίκτυα (P2P)

Η επίδραση της Δευτερεύουσας Μνήμης

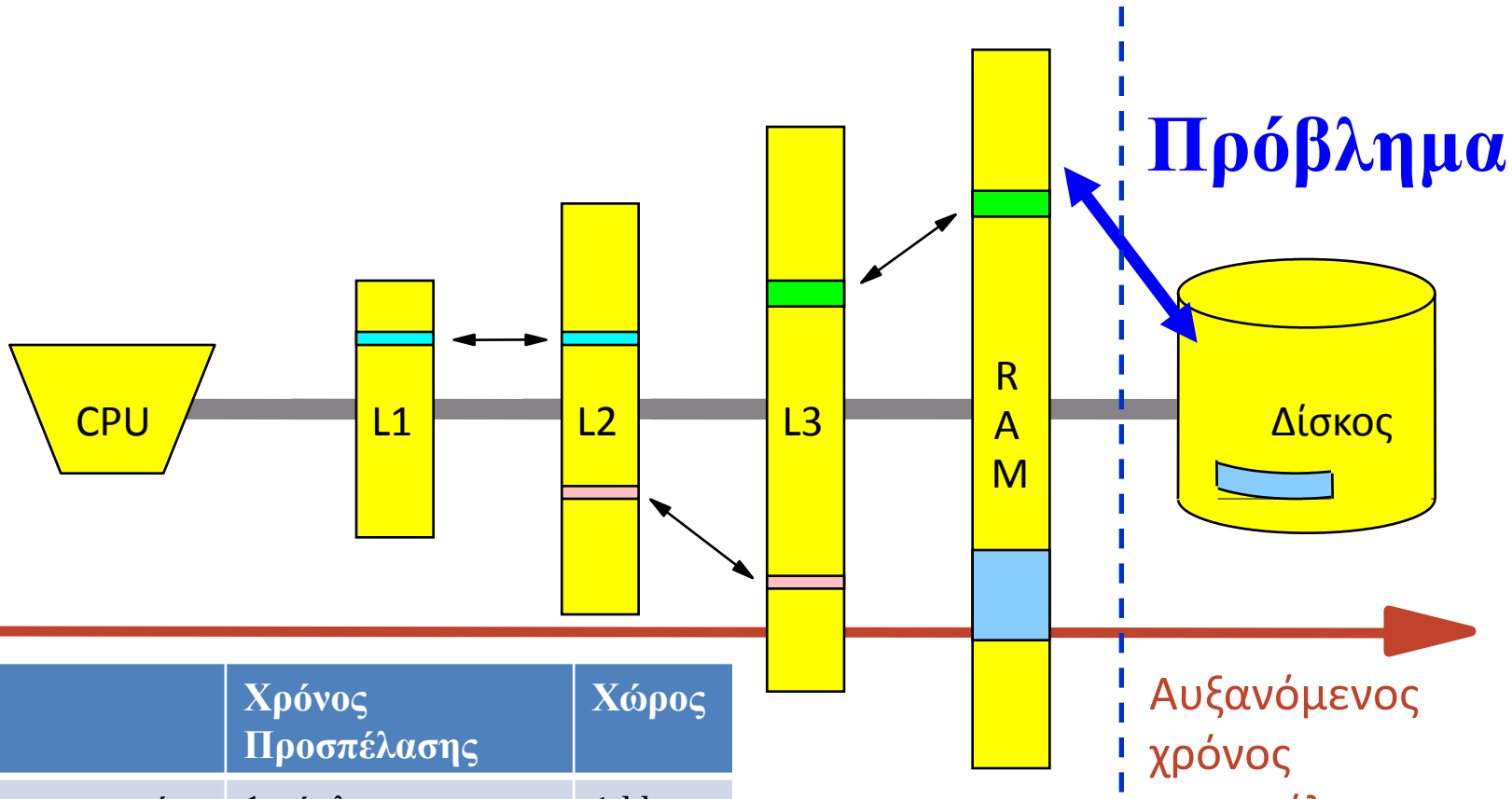


Η επίδραση της Δευτερεύουσας Μνήμης



Από πού προέρχεται αυτή η επιβράδυνση;

Οι Ιεραρχικές Μνήμες



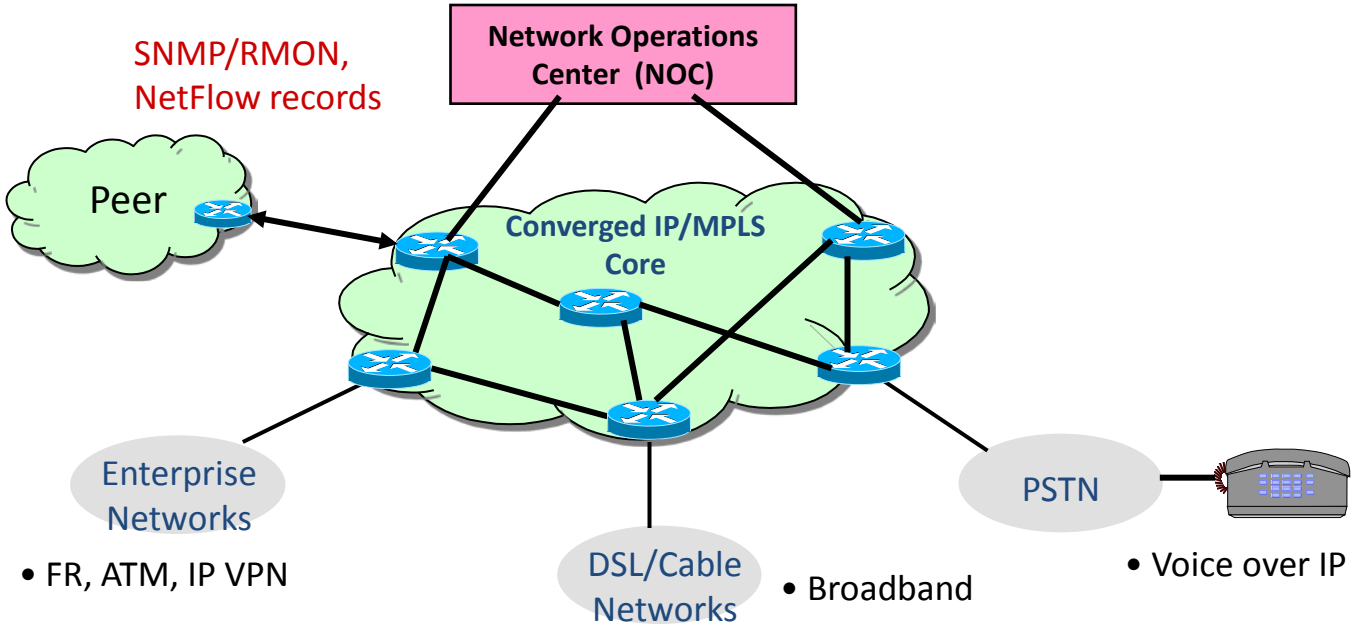
	Χρόνος Προσπέλασης	Χώρος
Καταχωρητές	1 κύκλος	1 kb
Cache	5 κύκλοι	512 kb
RAM	50 κύκλοι	1Gb
Δίσκος	10.000.000 κύκλοι	160 Gb

Χρόνοι Προσπέλασης (2 Cache Levels)

	Χρόνος Προσπέλασης	Σε σχέση με CPU
Register	0.5 ns	1
L1 cache	0.5 ns	1-2
L2 cache	3 ns	2-7
DRAM	150 ns	80-200
Δίσκος	10 ms	10^7

Αυξημένο

Επιτήρηση Δικτύου



- Τεράστιες ροές δεδομένων έρχονται με γρήγορους ρυθμούς (π.χ. η AT&T συλλέγει 600-800 GB πακέτων κάθε μέρα)
- Συνήθως κάποια από τα δεδομένα αναλύονται αργότερα

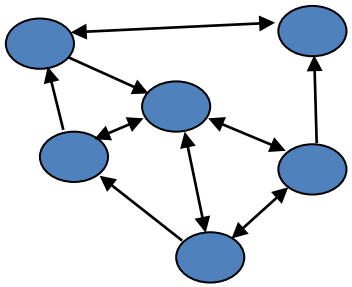
Προβλήματα Επιτήρησης

- Τα πακέτα φθάνουν κάθε 25ns σε μία γραμμή 40 Gbps
- Η παραδοσιακή λύση της δειγματοληψίας δεν λειτουργεί λόγω της μικρής ταχύτητας δειγματοληψίας εξαιτίας των περιορισμών του υλικού (π.χ. ταχύτητα της DRAM)
- Πρέπει να χρησιμοποιηθεί SRAM για την επεξεργασία κάθε πακέτου
- Αλλά η SRAM είναι αρκετά μικρή για να χωρέσει τα πάντα

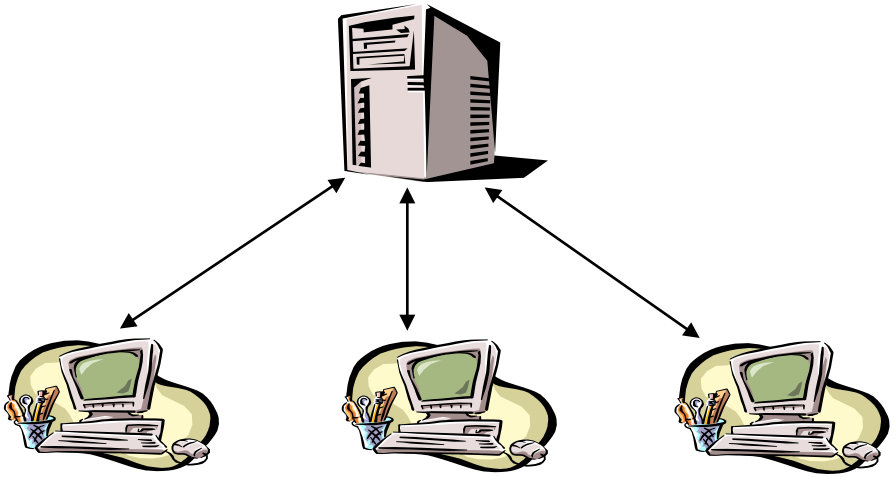
P2P Δίκτυα

- Ο P2P υπολογισμός και αποθήκευση ξεκίνησε από το Napster
- Αντί η πληροφορία να είναι αποθηκευμένη σε μερικούς κεντρικούς Η/Υ, ο P2P υπολογισμός επιτρέπει την πρόσβαση σε δεδομένα σε οποιονδήποτε Η/Υ στο δίκτυο

P2P



Όχι P2P



P2P Δίκτυα

- Το 2003 στο P2P δίκτυο Kazaa υπήρχαν διαθέσιμα 5 Petabytes
  - Οι αλγόριθμοι εύρεσης όμως δεν ήταν τόσο αποδοτικοί

- Το 2007 το P2P δίκτυο Skype έπεσε για 48 ώρες!
  - Είχε περίπου 12 εκατομμύρια χρήστες
  - Λόγω μαζικών επανεκκινήσεων Η/Υ (από Windows Update) προκλήθηκε υπερχείλιση από log-in requests

Το Πρόβλημα των Διακριτών Στοιχείων

- Δοθείσης μίας ακολουθίας $A=\alpha_1,\ldots,\alpha_n$, που διαβάζεται από μία πηγή, με $\alpha_i \in \{1,\ldots,m\}$, να υπολογιστεί το πλήθος των διαφορετικών στοιχείων στο A ($|A|$)
- Πιστεύετε ότι είναι εύκολο πρόβλημα;
- Προσοχή: Ανάλογα με το μοντέλο που εφαρμόζεται υπάρχουν και διαφορετικές λύσεις!
- Έχετε κάποιες ιδέες;

Λύση στην Κύρια Μνήμη

- Χρήση ενός μετρητή c=0
- Κατασκευή Πίνακα Χ στην μνήμη μεγέθους n
- Κάνουμε κατακερματισμό με αλυσίδες (hash with chains) στον πίνακα Χ
- Κάθε α_i που διαβάζεται τοποθετείται στον Χ και αν δεν υπάρχει εκεί αυξάνεται η τιμή του μετρητή: c=c+1
- Πολυπλοκότητα: O(n) αν η συνάρτηση hash είναι καλή (randomness)
- Αν όμως ο Χ δεν χωράει στην μνήμη;

Λύση στην Δευτερεύουσα Μνήμη

- Έστω M το μέγεθος της κύριας μνήμης και n>>M
- Τα δεδομένα του πίνακα X αποθηκεύονται στον δίσκο
- Ο χώρος του δίσκου διαιρείται σε blocks μεγέθους B<=M
- Η μεταφορά ενός block από το δίσκο στην μνήμη απαιτεί μία μονάδα χρόνου (π.χ. 1msec)
- Οι πράξεις στην κύρια μνήμη γίνονται ταχύτατα (οπότε θεωρούμε μηδενικό το κόστος τους)

Είναι Ρεαλιστικό το Μοντέλο;

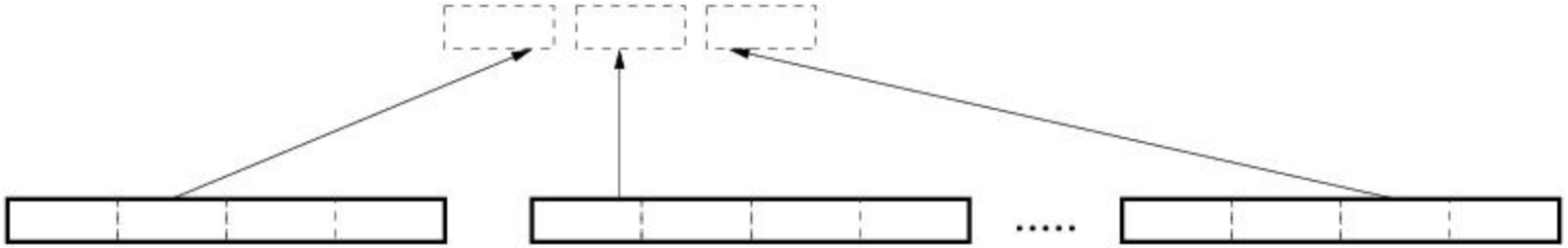
- Το κόστος των πράξεων θεωρείται μηδενικό καθώς:
  - Η ταχύτητα του επεξεργαστή είναι μερικά GHz
  - Η ταχύτητα αναζήτησης στο δίσκο είναι μερικά msecs
- Για τη μεταφορά ενός block:
  - Ταχύτητα διαμεταγωγής: 100MB/sec
  - Αν θέσουμε B = 100KB τότε η μεταφορά γίνεται στην ουσία δωρεάν (1msec)
- Τότε το κόστος του αλγορίθμου εξαρτάται κυρίως από τα πόσα blocks αναζητούνται/μεταφέρονται στον δίσκο (I/O)

Ο προηγούμενος Αλγόριθμος αρκεί;

- Ο αλγόριθμος της κύριας μνήμης θα απαιτούσε Θ(n) προσπελάσεις block (λόγω του κατακερματισμού)
- Αυτό δεν είναι καλό, π.χ. αν $n=10^9$ θα χρειαζόμασταν 10^6 sec ≈ 278 ώρες!
- Απαιτείται λοιπόν μία άλλη προσέγγιση
- Έχετε καμία ιδέα;

Ο Αλγόριθμος Merge-Sort

- Αλγόριθμος Ταξινόμησης με Συγχώνευση:
  - Διαίρεσε τα δεδομένα σε κομμάτια μεγέθους M/B
  - Αναδρομικά ταξινόμησε κάθε κομμάτι
  - Συγχώνευσε τα κομμάτια κάνοντας n/B μεταφορές block



- Χρόνος ταξινόμησης του A: $\mathbf{(n/B)\log_{M/B} n}$
- Έπειτα από την ταξινόμηση του A, υπολογίζουμε το |A| σε ένα σειριακό απλό πέρασμα
- Συνολικό κόστος: $\mathbf{((n/B)+1)\log_{M/B} n}$

Αν έχουμε Ροή δεδομένων;

- Τι γίνεται αν τα δεδομένα είναι σε ένα CD-ROM ή έρχονται απευθείας από την πηγή ή είναι τόσο μεγάλα που δεν τίθεται θέμα δεικτοδότησης ή πολλαπλών περασμάτων;
- Μοντέλο Ροών (περιορισμοί):
  - Το μέγεθος της κύριας μνήμης είναι M<<n
  - Τα δεδομένα έρχονται με τη μορφή ροής: σε κάθε βήμα μπορούμε να ζητήσουμε μόνο το επόμενο α_i
  - Θεωρούμε το n γνωστό (αν και δεν είναι απαραίτητο)
- Σαφώς, οι προηγούμενοι αλγόριθμοι δεν είναι εφαρμόσιμοι!

Αλγοριθμικοί Περιορισμοί

- **Θεώρημα:** Κάθε αιτιοκρατικός αλγόριθμος για αυτό το πρόβλημα απαιτεί O(m) χώρο
- **Απόδειξη:**
  - Έστω F ένας οποιοσδήποτε αλγόριθμος που μετρά το πλήθος των διαφορετικών στοιχείων στην τρέχουσα μνήμη, από μία ροή S με στοιχεία $\alpha_i \in \{1,\ldots,m\}$.
  - Για να ελέγξουμε αν ένα στοιχείο ανήκει στη ροή ή όχι κρατάμε ένα στιγμιότυπο της μνήμης του αλγόριθμου (F is a tester). Αρχικά είναι κενή.
  - Για κάθε στοιχείο $\alpha_i \in \{1,\ldots,m\}$ που διαβάζεται ελέγχεται με τον F αν άλλαξε ή όχι το πλήθος στην τρέχουσα μνήμη:

Αλγοριθμικοί Περιορισμοί

- Αν το πλήθος των διακριτών στοιχείων άλλαξε τότε σημαίνει ότι $x \in S$, διαφορετικά $x \notin S$.
- Η λειτουργία του αλγορίθμου F πρέπει να έχει αρκετή πληροφορία ώστε να διακρίνει τα 2^m δυνατά υποσύνολα του {1,…,m} που θα μπορούσαν να είναι τα διακριτά στοιχεία της S.
- Συνεπώς απαιτείται O(m) bits χώρος για να διακριθούν οι 2^m περιπτώσεις.

- Και αν το m είναι μεγάλο και δεν υπάρχει διαθέσιμη μνήμη;

Μπορούμε να κάνουμε εκτίμηση;

- Μπορούμε να προσεγγίσουμε με κάποιους τρόπους την τιμή του πλήθους των διακριτών στοιχείων της ροής
- Χρειάζονται πιθανοτικοί (randomized) αλγόριθμοι, που κάνουν εκτίμηση με βάση ένα μικρό σφάλμα ε
- Υπάρχει αλγόριθμος που εκτιμά το πλήθος με μεγάλη πιθανότητα και απαιτεί χώρο: poly(log(m),1/ε)

Ένας απλός Αλγόριθμος

- Χρησιμοποιούμε ένα bitset (πίνακας bit) στην μνήμη B[1…$\log_2 m$], αρχικοποιημένο με 0
- Χρησιμοποιούμε μία συνάρτηση κατακερματισμού f: {1…m} → {0…$\log_2 m$}
- Υπολογίζουμε την τιμή $f(\alpha_i)$ για κάθε στοιχείο α_i που έρχεται και θέτουμε το bit $B[f(\alpha_i)]$ σε 1
- Από το τελικό μέγεθος του bitset B (πλήθος bit που είναι 1) μπορούμε να πάρουμε μία προσεγγιστική τιμή του πλήθους των διακριτών στοιχείων της ροής

Ένας απλός Αλγόριθμος

- Για οποιοδήποτε στοιχείο $x \in \{1,\ldots,m\}$, κάθε φορά που έρχεται, πέφτει στο ίδιο κελί του B, και έτσι μετράμε μόνο τα διακριτά στοιχεία της ροής
- Αν υπάρχουν D διαφορετικά στοιχεία στη ροή τότε περιμένουμε ότι $1+\log_2 D$ κελιά του B θα είναι γεμάτα (θα έχουν 1)
- Στο τέλος λοιπόν μετράμε πόσα κελιά του B έχουν 1, έστω R, και η τελική εκτίμηση είναι ο αριθμός 2^{R-1}

Ένα Παράδειγμα

- Έστω η συνάρτηση κατακερματισμού f(x) που επιστρέφει το πλήθος των συνεχόμενων μηδενικών από το LSB και πέρα στην δυαδική αναπαράσταση του (3x mod 5) +1
- Έστω ότι η ροή είναι: 1, 3, 2, 1, 2, 3, 4, 3, 1, 2, 3, 1...
- Τότε οι αντίστοιχες τιμές της f(x) θα είναι: 2, 0, 1, 2, 1, 0, 0, 0, 2, 1, 0, 2,…
- Άρα το bitset θα έχει B[0] = 1, B[1] = 1, B[2] = 1, B[3]=0, B[4]=0
- Οπότε R = 3
- Τελική εκτίμηση: $2^2 = 4$

Αν επιλέξουμε δίκτυο P2P

- Το δίκτυο P2P είναι ένα γράφημα όπου κάθε κορυφή είναι μία υπολογιστική οντότητα και κάθε ακμή ένας δίαυλος επικοινωνίας
- Μετράμε το πλήθος των μηνυμάτων που στέλνονται διαμέσου των διαύλων (διότι κοστίζουν περισσότερο από τους υπολογισμούς στους κόμβους)
- Ο υπολογισμός στους κόμβους θεωρείται δωρεάν
- Τα μηνύματα είναι σταθερού μήκους
- Οι ερωτήσεις ξεκινούν από κάποιον κόμβο και η απάντηση επιστρέφει σε αυτόν
- Υπάρχει Παραλληλισμός

Είναι ρεαλιστικό μοντέλο;

- Σε δίκτυα επικοινωνίας, η αποστολή μηνυμάτων είναι η πιο χρονοβόρα διαδικασία. Βέβαια, εννοείται ότι η επεξεργασία εντός κάθε κόμβου δεν είναι απαιτητική
  - Ταχύτητα επεξεργαστή κόμβων: σε GHz
  - Ταχύτητα αποστολής μηνυμάτων: σε **Mbps**
- Δεν λαμβάνουμε υπόψη λάθη που δημιουργούνται στο δίκτυο
- Έχετε καμία ιδέα αλγοριθμικής λύσης σε P2P;

Μία προσέγγιση σε P2P

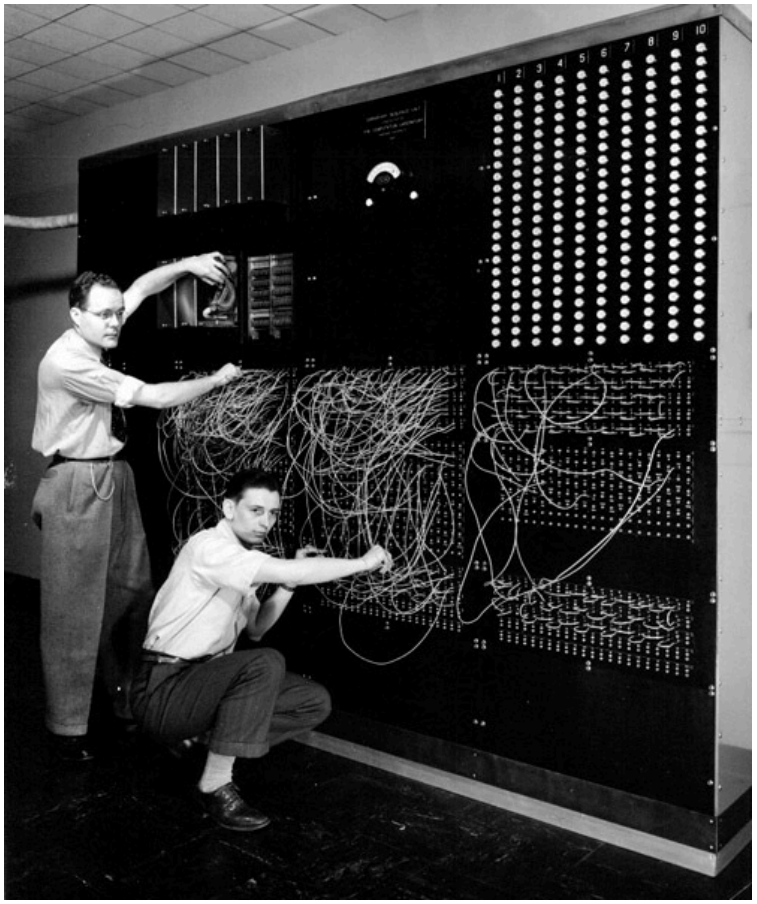
- Κάνουμε μία κατά πλάτος διαπέραση του δικτύου ξεκινώντας από τον κόμβο που κάνει την ερώτηση
- Φτιάχνουμε ένα ζευγνύων δέντρο για το δίκτυο. Πλήθος μηνυμάτων: k (ίσο με το πλήθος των κόμβων)
- Κάθε κόμβος έχει έναν πίνακα από bits (bitset) μεγέθους m που περιλαμβάνει τα διακριτά στοιχεία που βρήκε τοπικά
- Έπειτα με k μηνύματα ξεκινά η διαδικασία συγχώνευσης των πινάκων από τα φύλλα μέχρι να φτάσουμε στη ρίζα. Πλήθος μηνυμάτων: km
- Άρα το πλήθος των μηνυμάτων είναι O(km)

Συμπεράσματα

- Υπάρχουν τεράστιου όγκου δεδομένα παντού
- Οι παραδοσιακοί αλγόριθμοι εμφανίζουν προβλήματα κλιμάκωσης στο χρόνο εκτέλεσης:
  - Λόγω της Ιεραρχικής μνήμης
  - Λόγω του μικρού της μεγέθους σε σχέση με τα δεδομένα
  - Λόγω της επικοινωνίας
- Αλγόριθμοι που λαμβάνουν υπόψη τα παραπάνω βελτιώνουν κατά πολύ την αποδοτικότητά τους

Zetabyte Horizon

- Το 2009 όλος ο παγκόσμιος ιστός εκτιμήθηκε ότι περιλαμβάνει περίπου 500 Exabytes (10^{18}). Αυτό είναι το μισό ενός Zetabyte (10^{21}).
- Το συνολικό πλήθος των δεδομένων αυξήθηκε σε 2,7 Zetabytes μόνο κατά τη διάρκεια του 2012. Αυτό είναι 48% άνοδος σε σχέση με το 2011.



x50

2012

2020

How big is big?