Αποθήκευση και Δομή Αρχείων

Database System Concepts, 6th Ed.

Ταξινόμηση των μέσων φυσικής αποθήκευσης

- **Ταχύτητα** με την οποία είναι δυνατή η πρόσβαση στα δεδομένα
- **Κόστος** ανά μονάδα δεδομένων
- **Αξιοπιστία**
  - απώλεια δεδομένων σε περίπτωση διακοπής ρεύματος ή κατάρρευσης
  - φυσική αστοχία της συσκευής αποθήκευσης
- Διαφοροποιούμε το μέσο αποθήκευσης σε:
  - **πτητικό:** χάνει το περιεχόμενο όταν απενεργοποιηθεί η παροχή ρεύματος
  - **Μη πτητικό**:
    - Τα περιεχόμενα εξακολουθούν να υφίστανται ακόμη και όταν η συσκευή είναι απενεργοποιημένη.
    - Περιλαμβάνει δευτερεύουσα και τριτεύουσα αποθήκευση και κύρια μνήμη μπαταρίας.

Μέσα φυσικής αποθήκευσης

- **Cache** – η ταχύτερη και πιο δαπανηρή μορφή αποθήκευσης. Πτητική. Τη διαχειρίζεται το υλικό του συστήματος.
- **Main memory**:
  - Γρήγορη πρόσβαση (10s με 100s nanoseconds. 1 nanosecond = 10^{-9} seconds)
  - Γενικά είτε μικρή (ή πολύ ακριβή) για να αποθηκεύσουμε όλη τη ΒΔ σε αυτήν
    - Μέχρι κάποια Gigabytes συνήθως
    - Μέγεθος και κόστος αλλάζουν (περίπου με ρυθμό επί 2 κάθε 2 με 3 χρόνια)
  - **Πτητική**.

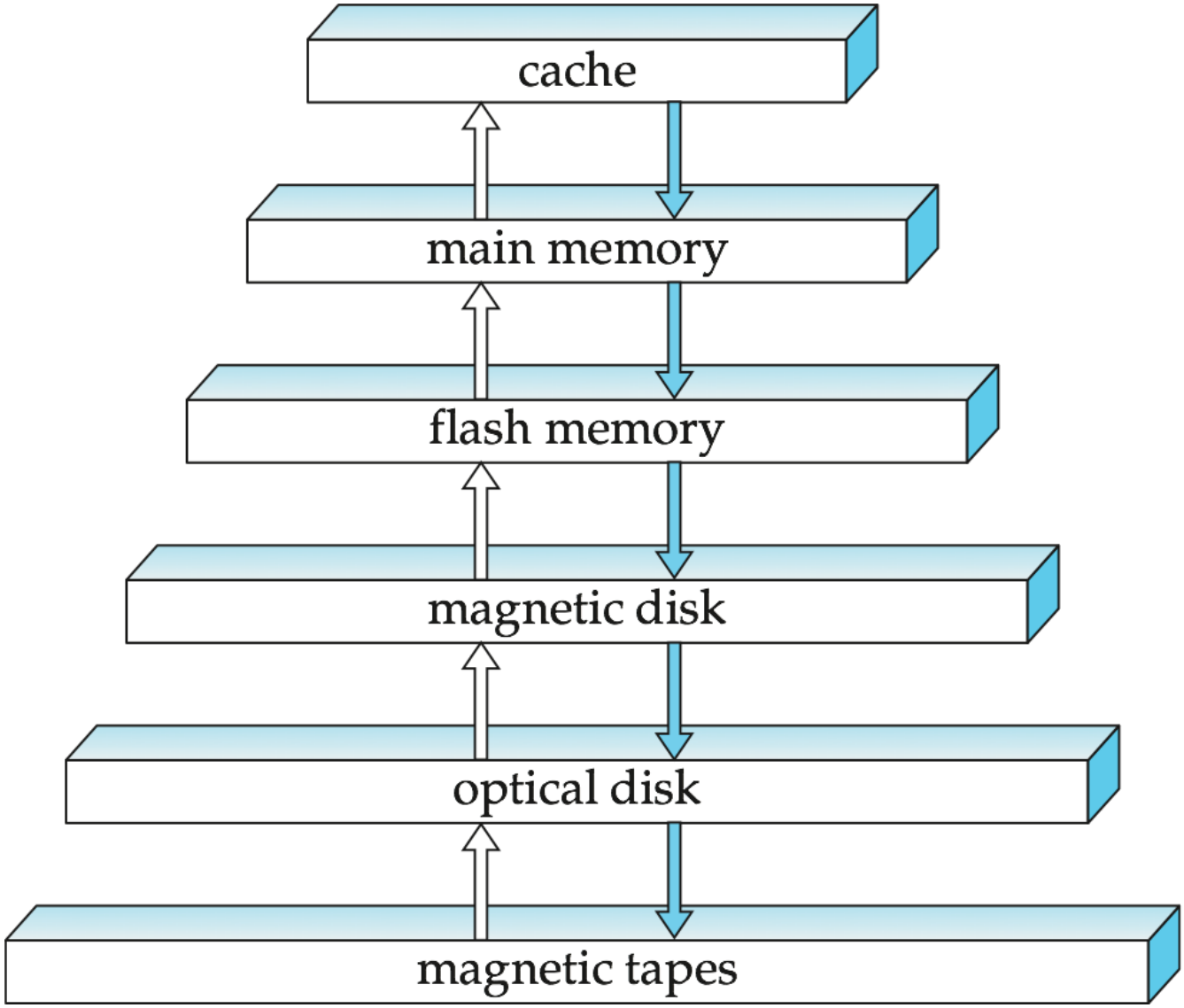
Μέσα φυσικής αποθήκευσης

- **Flash memory**
  - Τα δεδομένα επιβιώνουν με διακοπή ρεύματος
  - Τα δεδομένα μπορούν να γραφτούν σε μια τοποθεσία μόνο μία φορά, αλλά η θέση μπορεί να διαγραφεί και να γραφτεί ξανά
    - Μπορεί να υποστηρίξει μόνο έναν περιορισμένο αριθμό κύκλων εγγραφής / διαγραφής (10K – 1M)
    - Η διαγραφή της μνήμης πρέπει να γίνει σε μια ολόκληρη τράπεζα μνήμης
  - Οι αναγνώσεις είναι περίπου τόσο γρήγορες όσο η κύρια μνήμη
  - Αλλά οι εγγραφές είναι αργές (λίγα μικροδευτερόλεπτα), η διαγραφή είναι πιο αργή
  - Χρησιμοποιείται ευρέως σε ενσωματωμένες συσκευές όπως ψηφιακές φωτογραφικές μηχανές, τηλέφωνα και κλειδιά USB

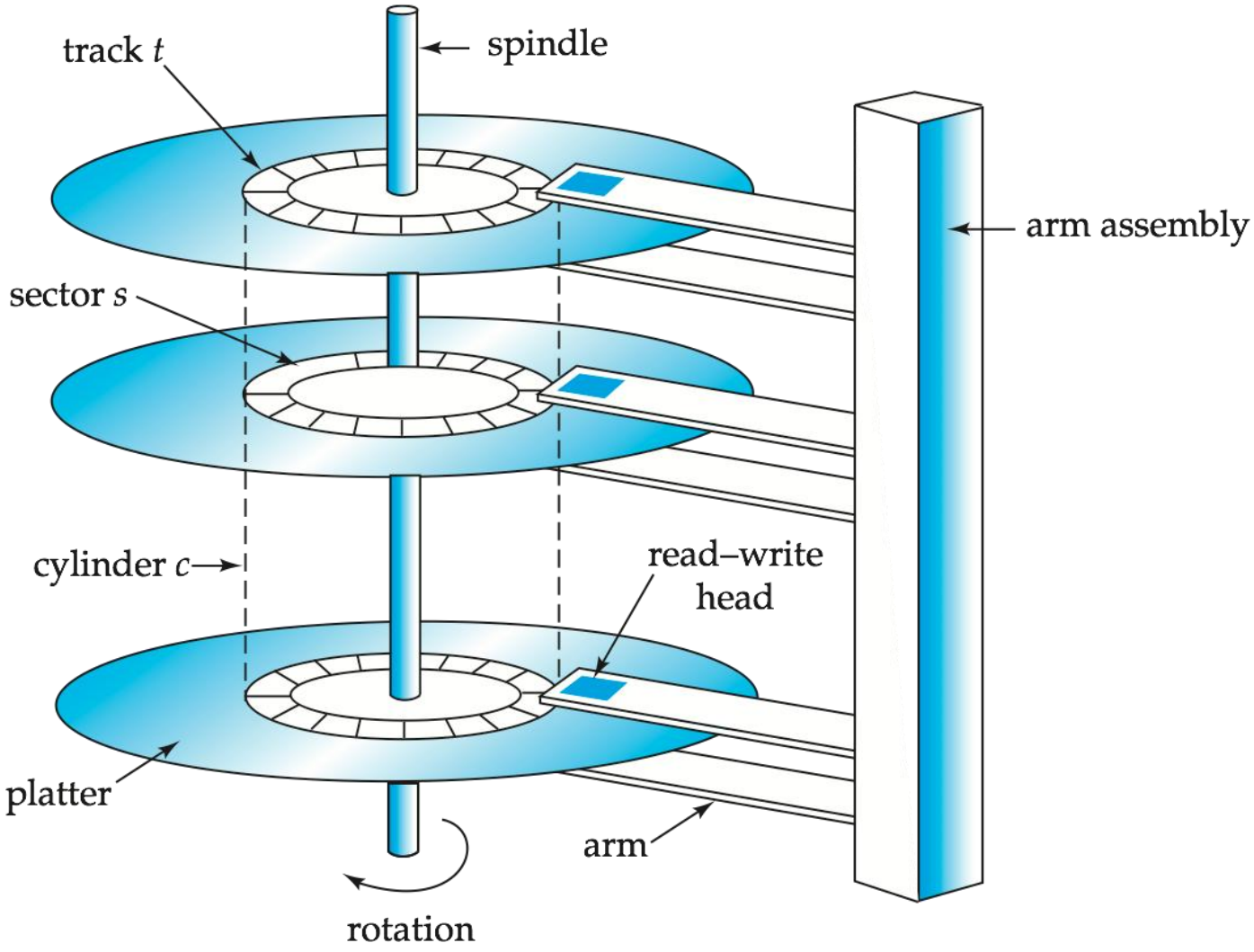
Μέσα φυσικής αποθήκευσης

- **Μαγνητικός Δίσκος**
  - Τα δεδομένα αποθηκεύονται στον περιστρεφόμενο δίσκο και διαβάζονται / γράφονται μαγνητικά
  - Κύριο μέσο για τη μακροπρόθεσμη αποθήκευση δεδομένων. Συνήθως αποθηκεύει ολόκληρη τη ΒΔ.
  - Τα δεδομένα πρέπει να μετακινούνται από το δίσκο στην κύρια μνήμη για πρόσβαση και να γράφονται πίσω για αποθήκευση
    - Πολύ πιο αργή πρόσβαση από την κύρια μνήμη
  - **άμεση πρόσβαση** - δυνατότητα ανάγνωσης δεδομένων σε δίσκο με οποιαδήποτε σειρά, σε αντίθεση με μαγνητική ταινία
  - Χωρητικότητες μέχρι κάποια ΤΒ
    - Πολύ μεγαλύτερη χωρητικότητα και μικρότερο κόστος / byte από την κύρια μνήμη / μνήμη flash
    - Μεγαλώνει συνεχώς
  - Επιζεί από διακοές ρεύματος και crashes
    - η αποτυχία δίσκου μπορεί να καταστρέψει δεδομένα, αλλά είναι σπάνια

Storage Hierarchy



Μηχανισμός μανγητικού σκληρού δίσκου



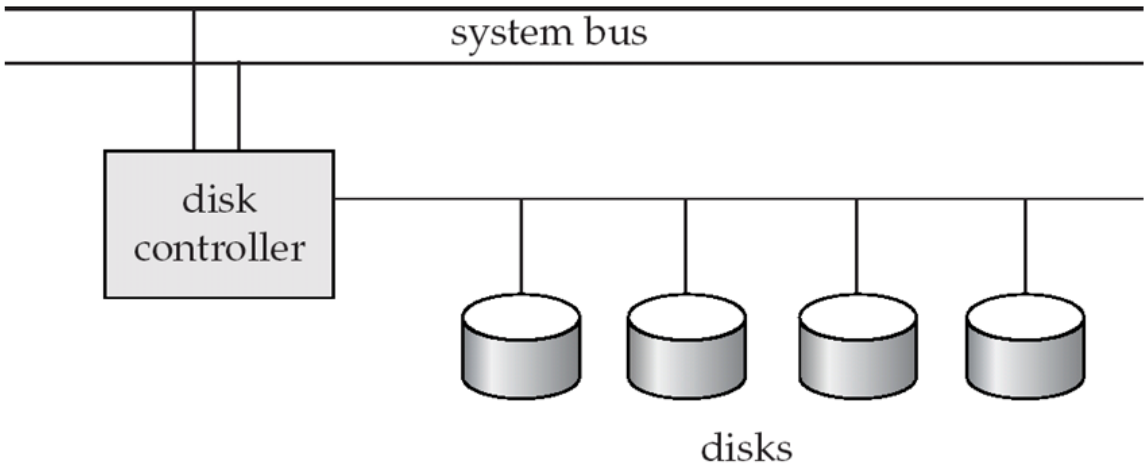
Μαγνητικοί Σκληροί Δίσκοι

- **Κεφαλή Read-write**
  - Τοποθετείται πολύ κοντά στην επιφάνεια - σχεδόν αγγίζει
  - Διαβάζει ή γράφει μαγνητικά κωδικοποιημένες πληροφορίες.
- Η επιφάνεια διαιρείται σε κυκλικά **tracks**
  - Πάνω από 50K-100K tracks ανά επιφάνεια
- Κάθε track διαιρείται σε **sectors.**
  - Είναι η μικρότερη μονάδα δεδομένων που μπορεί να διαβαστεί ή να γραφτεί.
  - Το μέγεθός του είναι συνήθως 512 byte
  - Τυπικοί sectors ανά track: 500 έως 1000 (σε εσωτερικά tracks), 1000 έως 2000 (σε εξωτερικά tracks).
- Για ανάγνωση/εγγραφή ενός sector
  - ο βραχίονας του δίσκου μετακινείται για να τοποθετήσει την κεφαλή στο σωστό track
  - Η επιφάνεια περιστρέφεται συνεχώς. Τα δεδομένα διαβάζονται / εγγράφονται καθώς ο sector περνά κάτω από την κεφαλή

Μαγνητικοί Σκληροί Δίσκοι

- **Ελεγκτής Δίσκου Disk controller** – διασύνδεση μεταξύ του συστήματος υπολογιστή και του υλικού της μονάδας δίσκου.
  - δέχεται εντολές υψηλού επιπέδου για ανάγνωση/εγγραφή sector
  - ξεκινάει ενέργειες όπως η μετακίνηση του βραχίονα του δίσκου στο σωστό track
  - Υπολογίζει και αποδίδει αθροίσματα ελέγχου **checksums** σε κάθε sector για να επαληθεύσει ότι τα δεδομένα διαβάζονται σωστά
  - Εξασφαλίζει την επιτυχημένη εγγραφή μέσω απανανάγνωσης του τομέα μετά την εγγραφή του
  - Εκτελεί remapping κακών bad sectors

Υποσύστημα Δίσκου



- Οικογένειες διαπροσωπείας Δίσκων
  - ATA (AT adaptor)
  - SATA (Serial ATA)
  - SCSI (Small Computer System Interconnect)
  - SAS (Serial Attached SCSI)

Υποσύστημα Δίσκου

- Οι δίσκοι συνήθως συνδέονται απευθείας με το σύστημα υπολογιστή
- Στα **Storage Area Networks (SAN)**, ένας μεγάλος αριθμός δίσκων συνδέεται μέσω δικτύου υψηλής ταχύτητας σε έναν αριθμό από servers
- Το **Network Attached Storage (NAS)** παρέχει μια διασύνδεση συστήματος αρχείων που χρησιμοποιεί πρωτόκολλο δικτύου αρχείων δικτύου, αντί να παρέχει διεπαφή συστήματος δίσκου

Μέτρα απόδοσης των δίσκων

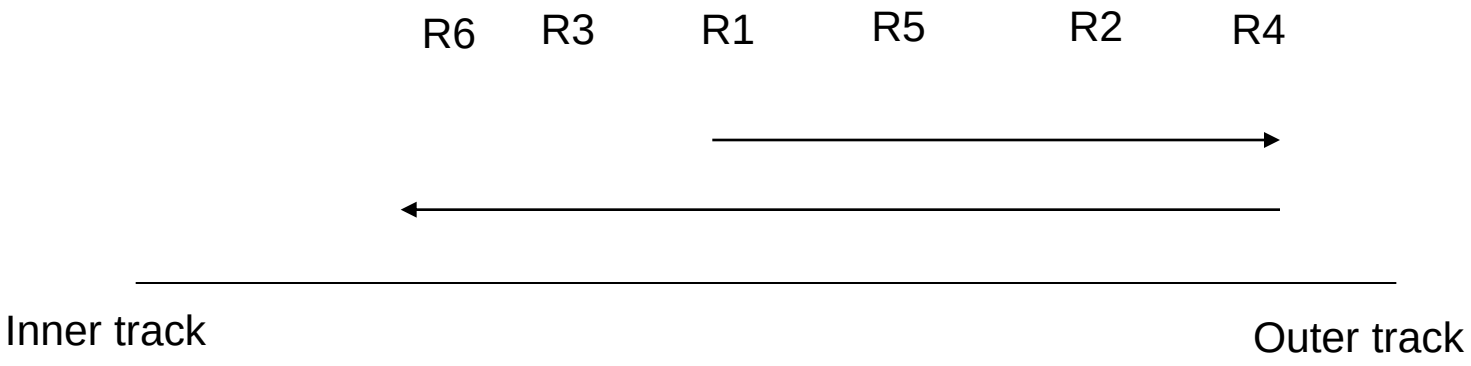
- **Χρόνος πρόσβασης** - ο χρόνος από όταν εκδίδεται μια αίτηση ανάγνωσης ή εγγραφής έως την έναρξη της μεταφοράς δεδομένων. Αποτελείται από:
  - **Χρόνος αναζήτησης(Seek time)** – χρόνος που απαιτείται για επανατοποθέτιση του βραχίονα στο σωστό track.
    - Μέσο seek time είναι το 1/2 της χειρότερης περίπτωσης.
      - Θα ήταν 1/3 αν όλα τα tracks είχαν ίδιο αριθμό sectors και αγνοούσαμε το χρόνο για το σταμάτα-ξεκίνα του βραχίονα
    - 4 με 10 milliseconds
  - **Rotational latency** – χρόνος που απαιτείται για έναν sector να εμφανιστέι κάτω από την κεφαλή
    - Μέση τιμή είναι το 1/2 της χειρότερης περίπτωσης.
    - 4 με 11 milliseconds (5400 έως15000 r.p.m.)
- **Ρυθμός μετάδοσης(Data-transfer rate)** – ο ρυθμός με τον οποίο τα δεδομένα μπορούν να ανακτηθούν ή να αποθηκευτούν στο δίσκο.
  - 25 έως 100 MB per second το μέγιστο
  - Πολλαπλοί δίσκοι μπορούν να μοιράζονται έναν ελεγκτή, οπότε ο ρυθμός που μπορεί να χειριστεί ο ελεγκτής είναι επίσης σημαντικός
  - E.g. SATA: 150 MB/sec, SATA-II 3Gb (300 MB/sec)
    - Ultra 320 SCSI: 320 MB/s, SAS (3 to 6 Gb/sec)
    - Fiber Channel (FC2Gb or 4Gb): 256 to 512 MB/s

Μέτρα απόδοσης των δίσκων

- **Mean time to failure (MTTF) Μέσος χρόνος εμφάνισης βλάβης** – ο μέσος χρόνος που ο δίσκος αναμένεται να λειτουργεί συνεχώς χωρίς αποτυχία
  - Συνήθως 3 με 5 χρόνια
  - Η πιθανότητα αποτυχίας είναι μικρή, με “θεωρητικό MTTF” 500,000 έως 1,200,000 ωρών για ένα νέο δίσκο
  - Το MTTF μειώνεται καθώς ένας δίσκος «γερνά»

Βελτιστοποίηση της πρόσβασης σε μπλοκ δίσκου

- **Block** – μια συνεχόμενη ακολουθία από σταθερό αριθμό sectors από ένα μόνο track
  - τα δεδομένα μεταφέρονται μεταξύ του δίσκου και της κύριας μνήμης σε μπλοκ
  - Μεγέθη ποικίλουν από 12 bytes έως αρκετά kilobytes
    - Μικρότερα blocks: πιο πολλές μεταφορές
    - Μεγαλύτερα blocks: περισσότερος χώρος σπαταλιέται λόγω μερικώς γεμισμένων μπλοκ
    - Συνήθως 4 με 16 kilobytes
- **Αλγόριθμοι δρομολόγησης δίσκου-βραχίονα** προσπαθούν να ελαχιστοποιήσουν την κίνηση του βραχίονα
  - **elevator algorithm**:



Βελτιστοποίηση της πρόσβασης σε μπλοκ δίσκου

- **Οργάνωση αρχείων -** βελτιστοποίηση του χρόνου πρόσβασης μπλοκ με την οργάνωση των μπλοκ ώστε να αντιστοιχούν στον τρόπο πρόσβασης στα δεδομένα
  - Π.χ. Αποθήκευση σχετικών δεδομένων στον ίδιο ή κοντινούς κυλίνδρους.
  - Τα αρχεία ενδέχεται να κατακερματιστούν (fragmented) με την πάροδο του χρόνου
    - Π.χ. Αν δεδομένα εισάγονται/διαγράφονται από το αρχείο
    - 'Η ελεύθερα blocks στο δίσκο διασκορπίζονται οπότε ένα νέο αρχείο έχει τα μπλοκ του διασκορπισμένα
    - Η σειριακή πρόσβαση σε ένα κατακερματισμένο αρχείο προκαλεί αυξημένη κίνηση βραχίονα
  - Κάποια συστήματα έχουν εφαρμογές για defragment του συστήματος αρχείων για επιτάχυνση

Βελτίωση της απόδοσης μέσω παραλληλισμού

- Δύο κύριοι στόχοι του παραλληλισμού σε ένα σύστημα δίσκων:

1. Ισορροπία φορτίου για πολλαπλές μικρές προσβάσεις για την αύξηση της απόδοσης
2. Παραλληλοποίηση μεγάλων προσπελάσεων για μείωση του χρόνου απόκρισης.

- Βελτίωση του ρυθμού μετάδοσης δεδομένων καταμερίζοντας (**striping**) τα δεδομένα σε πολλαπλούς δίσκους.
- **Bit-level striping** – μοίρασμα των ψηφίων κάθε byte σε πολλούς δίσκους
  - Σε συστοιχία 8 δίσκων, γράψε το bit *i* από κάθε byte στον δίσκο *i.*
  - Κάθε πρόσβαση μπορεί να διαβάσει 8 φορές πιο γρήγορα.
  - Αλλά ο χρόνος αναζήτησης/πρόσβασης χειρότερος από ενός δίσκου
    - Ο καταμερισμός σε επίπεδο Bit δεν χρησιμοποιείται πλέον
- **Block-level striping** – με *n* δίσκους, το block *i* ενός αρχείου πάει στο δίσκο (*i* mod *n*) + 1
  - Αιτήματα για διαφορετικά block εκτελούνται παράλληλα αν τα block βρίσκονται σε διαφορετικούς δίσκους
  - Αίτημα για ακολουθία από πολλά block μπορεί να χρησ/σει όλους τους δίσκους παράλληλα