Λειτουργικά Συστήματα (HY321)

Διάλεξη 12:
Συστήματα RAID

Χρήστος Δ. Αντωνόπουλος
8/5/2017

Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Οι Καθυστερήσεις των Δίσκων

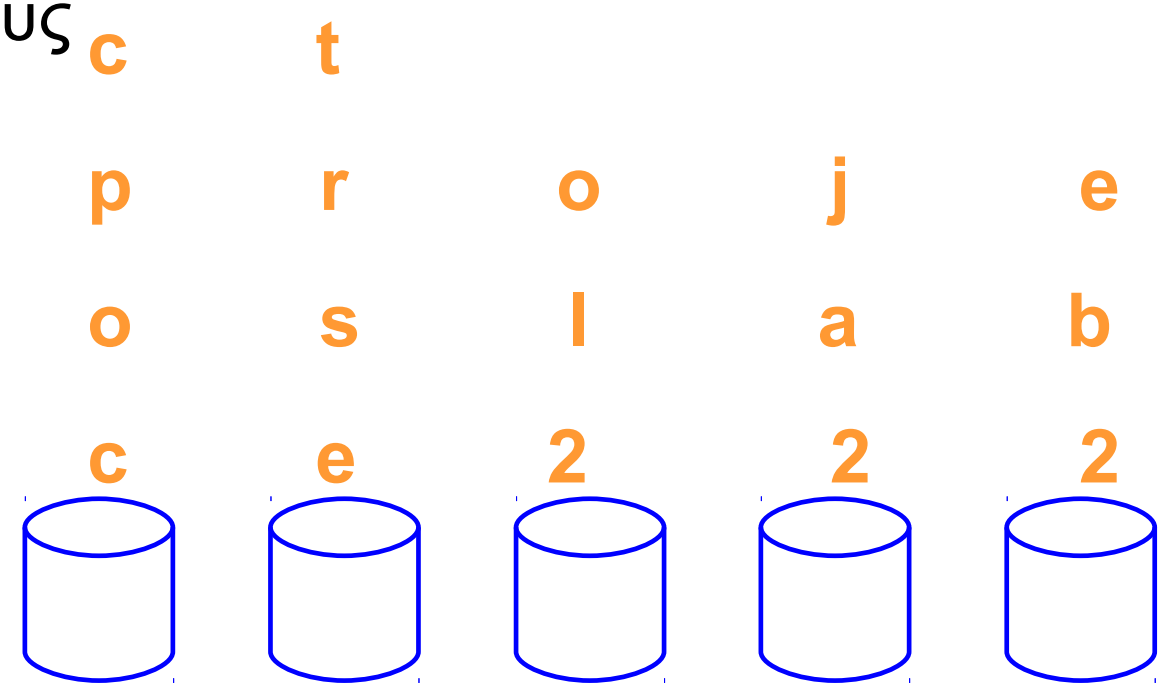
- Χρόνος αναζήτησης
  - Μάζα της κεφαλής / βραχίονα
  - Καθυστέρηση για την σταθεροποίηση / τοποθέτηση με ακρίβεια
  - Δύσκολο να βελτιωθεί…
    - Ακόμα και για τους κυριλέ («λεφτάδες») πελάτες μας
- Καθυστέρηση περιστροφής
  - Πόσο γρήγορα να στρίψει και αυτός ο κινητήρας…
  - Ανάγνωση / εγγραφή δεδομένων => αναλογικό σε ψηφιακό (και αντίστροφα)
    - Ούτε και αυτό μπορεί να γίνει τρομερά γρήγορα
  - Εδώ τα χρήματα μπορούν να αγοράσουν λίγη επίδοση
    - Λίγη = x1.5, x2.0 αλλά όχι x100.0
- Χρόνος μεταφοράς
  - Και πάλι έχει να κάνει βασικά με την καθυστέρηση περιστροφής

Οι Λύσεις σε Υλικό: Παράλληλες Μεταφορές

- Να μειώσουμε το χρόνο μεταφοράς
  - Χωρίς ο δίσκος να γυρίζει γρηγορότερα
- Διάβασε «παράλληλα» από πολλαπλές κεφαλές
  - Μα χρειάζεται πολλαπλές κεφαλές
  - … ενδεχομένως και βραχίονες
  - … και αντίγραφα των ηλεκτρονικών για τη μετατροπή αναλογικού ⇔ ψηφιακό
  - Είπαμε, μας ενδιαφέρουν οι «κυριλέ» πελάτες
- Μήπως μπορούμε να χρησιμοποιήσουμε τους υπάρχοντες δίσκους;
  - RAID (Redundant Array of Inexpensive Disks)

Διαμοίραση σε «Ταινίες» (Stripping) – RAID 0

- Βάλε πολλούς ίδιους δίσκους
  - Μοίρασε τα δεδομένα σε αυτούς
  - Διάβασε/γράψε παράλληλα από/σε πολλαπλούς δίσκους



Ταινιοποίηση

- Μονάδα ταινιοποίησης
  - Πόσα δεδομένα πάνε σε κάθε δίσκο (πόσο μεγάλοι ορμαθοί);
    - Byte
    - Bit
    - Τομέας (συνήθως)
- Μέγεθος ταινίας: μονάδα ταινιοποίησης x # δίσκων
- Συμπεριφορά: «μεγάλοι» τομείς
  - Το σύστημα αρχείων δημιουργεί μια «μεγάλη» αίτηση: N ενέργειες στο δίσκο
  - Κάθε δίσκος διαβάζει / γράφει έναν τομέα

Παράδειγμα

- Μονάδα: Τομέας (512 bytes)
  - 5 δίσκοι
  - Ταινία: 5x512 bytes (N x sector size) = 2.5 Kbytes
- Τι πετύχαμε;
  - Χρόνος αναζήτησης: Ίδιος με του ενός δίσκου
  - Ρυθμός μεταφοράς: 5x (Nx) του ενός δίσκου
- Κανένα πρόβλημα;
  - Τι συμβαίνει με την καθυστέρηση περιστροφής;

Ταινιοποίηση υψηλής επίδοσης

- Η καθυστέρηση περιστροφής χειροτερεύει
  - Κάθε ενέργεια δεν ολοκληρώνεται έως ότου γραφτεί/αναγνωσθεί ο τομέας και στον τελευταίο δίσκο
  - Σε 1 δίσκο μέση καθυστέρηση:
    - ½ περιστροφή
    - Σε N δίσκους τείνει σε 1 περιστροφή
- Λύση;
  - Συγχρόνισε τα μοτέρ των N δίσκων
    - Βεβαιώσου ότι ο τομέας 0 θα περνάει κάτω από όλες τις κεφαλές την ίδια στιγμή
  - Τι χρειάζεται;
    - Απλοί δίσκοι με λίγο παραπάνω λογική για το συγχρονισμό
      - Σχετικά φθηνό
      - Χρησιμοποιείται σε μεγάλους υπολογιστές

Ένας Πιο Γήινος Στόχος: Χωρητικότητα

- Μπορώ να έχω…
  - … όσο χώρο θέλω …
  - … για πάντα;
- Εύκολο: Φτιάξε μεγαλύτερους δίσκους
  - Αρχαία ιστορία: IBM 3380 (80's)
    - Δίσκος διαμέτρου 14'' (όσο μια μικρή οθόνη)
    - Μαζί με τα «συνοδευτικά» όσο ένα μικρό ψυγείο
    - Αλλά είχε μεγάλη χωρητικότητα
      - 1-3 GBytes (ώπααα!)
  - Πρέπει να είναι ακριβός ε;
  - Δε μπορούμε να βάλουμε πολλούς απλούς δίσκους μαζί;
    - Όπως κάναμε για να κερδίσουμε ταχύτητα

Ταινιοποίηση (και) για Χωρητικότητα

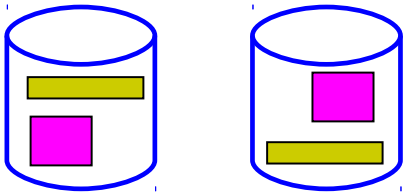
- 5 δίσκοι, μονάδα ταινίας 512 Bytes, μέγεθος ταινίας 2.5 KB
- Τι πετύχαμε;
  - Χρόνος αναζήτησης: Όσο στον 1 δίσκο (καλούλι)
  - Καθυστέρηση περιστροφής: Όσο στον 1 δίσκο (και πάλι καλούλι)
  - Ρυθμός μεταφοράς: 5x του ενός δίσκου (όχι κακό…)
  - Χωρητικότητα: 5x του ενός δίσκου (καθόλου κακό…)
- Τι μπορεί να «στραβώσει»;

Ποιος Πήρε την Αξιοπιστία;

- Μέσος χρόνος μεταξύ βλαβών (Mean Time To Failure – MTTF)
- MTTF (συστοιχίας) = MTTF (δίσκου) / # δίσκων
- Δίσκος χωρητικότητας 300 GB, MTTF=200.000 ώρες
  - Συστοιχία 100 δίσκων:
    - Χωρητικότητα 30 TBytes (εξαιρετικά)
    - MTTF = 200.000 ώρες / 100 = 2.000 ώρες (3 μήνες – όχι και τόσο καλά…)
      - Δηλαδή κάθε 3 μήνες θα φορτώνω το σύστημα αρχείων από το backup;
        - Πόσο είπαμε είναι το σύστημα αρχείων;
        - Δε λέει…

Αντιγραφή (Mirroring): RAID 1

- Όταν ένας δίσκος «παραδώσει πνεύμα»
  - Ας υποθέσουμε ότι ένας άλλος δίσκος έχει τα ίδια δεδομένα
  - Άρα αντέγραψε τα δεδομένα από τον άλλο δίσκο σε ένα νέο δίσκο
    - Πικρή ιστορία, αλλά λιγότερο πικρή από την αποκατάσταση δίσκου από backup (ταινία)
    - Και τα δεδομένα στο αντίγραφο είναι πάντα επίκαιρα



Η Χρήση των Αντιγράφων

- Ενέργεια:
  - Εγγραφή: Γράψε και στους 2 δίσκους
  - Ανάγνωση: Διάβασε από οποιονδήποτε δίσκο
- Επίδοση:
  - Εγγραφές: Λίγο πιο αργές
  - Αναγνώσεις: Έως και 2x πιο γρήγορες
- Αξιοπιστία:
  - Πολύ πολύ καλύτερα…
- Κόστος:
  - 2πλάσιο κόστος / byte
    - Συμπαθητικό αλλά ακριβό
    - Μπορώ να έχω κάτι ενδιάμεσο;

Κωδικοποίηση σε 5 Απλές Διαφάνειες…

- Δεδομένα και Μηνύματα:
  - Δεδομένα: Αυτό που θέλουμε να μεταδώσουμε
  - Μήνυμα: Δεδομένα + πληροφορία ελέγχου (επιπλέον bits)
- Ανίχνευση σφάλματος
  - Κάπου κάτι κάπως χάλασε…
  - … τουλάχιστον μπορώ να το καταλάβω
- Διόρθωση σφάλματος
  - Το bit 12 θα έπρεπε να είναι 1 και όχι 0
    - Σαφώς πιο χρήσιμο
    - Και μάλλον πιο «ακριβό»

Ένα Παράδειγμα

- Αντί για bits στέλνω 4άδες bits
  - Bit 0 => λέξη 0000
  - Bit 1 => λέξη 1111
- Μετάδοση:
  - Στέλνω 0000, λαμβάνω 0100
- Ανίχνευση σφάλματος
  - Το 0100 δεν είναι νόμιμη λέξη!!!
- Διόρθωση σφάλματος
  - Το 0100 μάλλον μοιάζει με το 0000 και όχι με το 1111

Που θα την Πατήσω;

- Αν λάβω 0110;
  - Είναι 0000 ή 1111;
- Πολλαπλά σφάλματα:
  - 0000 => 0010 => 1010 => 1011
  - Μάλλον μοιάζει με 1111, σωστά;
- Οι κώδικες συνήθως αναγνωρίζουν περισσότερα σφάλματα από όσα μπορούν να διορθώσουν
  - Π.χ. αναγνώριση 1-4 λαθών, διόρθωση μονών λαθών
  - Αν έχω 5 λάθη θα τα αναγνωρίσω πιθανότατα σα μια σωστή (πλην όμως διαφορετική) λέξη

Ισοτιμία

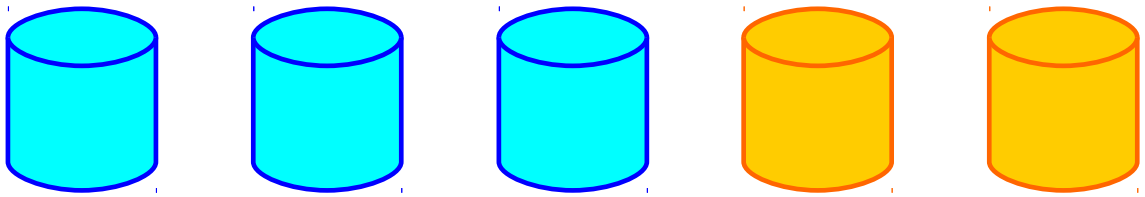
- Bit ισοτιμίας : XOR των bits δεδομένων
  - $a \oplus b = a'b + ab'$
  - $0 \oplus 1 \oplus 1 = 0$
  - Parity bit τέτοιο ώστε οι 1 στη λέξη να είναι ζυγές
  - $x \oplus x = 0$,
  - Αν $x \oplus y = z$ τότε $z \oplus y = x$ και $z \oplus x = y$
- Ανιχνεύει μονά λάθη
  - Ο αποστολέας στέλνει δεδομένα + parity bit
  - 011,0 σωστό – 011,1 λάθος
    - Ανιχνεύονται μονά, 3πλά, 5πλά κοκ λάθη
    - Δεν υπάρχει η δυνατότητα διόρθωσης
    - Δεν ανιχνεύονται 2πλά, 4πλά, 6πλά κοκ λάθη

Κώδικες Διόρθωσης Λαθών (Error Correcting Codes - ECC)

- Λέξη = Δεδομένα + πολλαπλά bits ελέγχου
- «Μαγικά» μαθηματικά
  - Κώδικας Hamming, Reed-Solomon κοκ
  - Μπορεί να ανιχνεύσει έως N σφάλματα (το N εξαρτάται από τον αριθμό bits ελέγχου / bit δεδομένων)
  - Μπορεί να διορθώσει M σφάλματα (M<N, συχνά M~N/2)

Τι Ψάχναμε; RAID 2

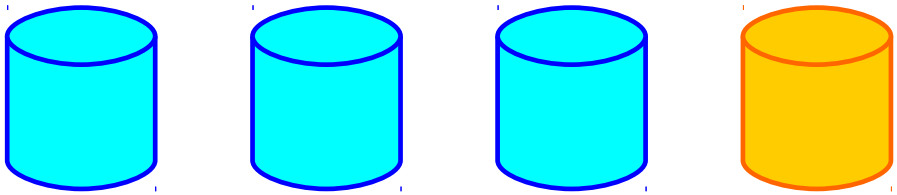
- Κάτι ανάμεσα στο RAID 0 και το RAID 1
  - Ε, η λύση είναι η μερική αντιγραφή
- RAID 2:
  - Μονάδα ταινιοποίησης 1 bit / δίσκο
  - N δίσκοι δεδομένων, M δίσκοι parity
  - Ανίχνευση/διόρθωση πολλαπλών λαθών



  - Ξεχάστε το…
    - Δε χρησιμοποιείται συχνά

RAID 3

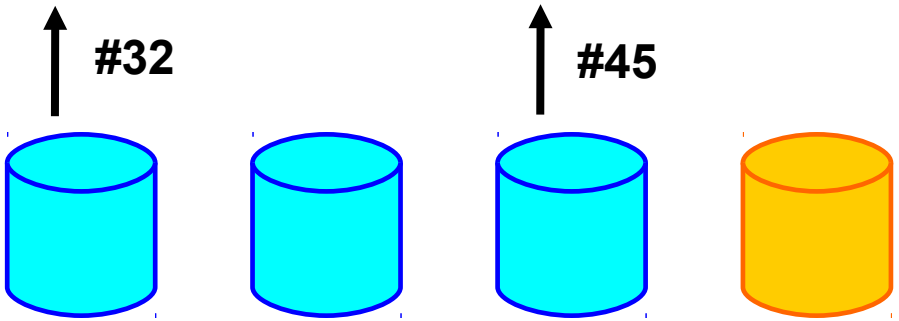
- RAID 3:
  - Μονάδα ταινιοποίησης 1 byte / δίσκο
  - Ν δίσκοι δεδομένων, 1 δίσκος parity
  - Ανίχνευση/διόρθωση απλού λάθους
    - Αρκεί να ξέρουμε ποιος δίσκος είναι χαλασμένος



  - Χρησιμοποιείται σε κάποιες εφαρμογές υψηλής επίδοσης

RAID 4

- RAID 4:
  - Παρόμοιο με το RAID 3
  - Όμως μονάδα ταινιοποίησης 1 τομέας (αντί 1 byte)
  - Οι αναγνώσεις 1 τομέα «ακουμπούν» μόνο 1 δίσκο
    - Άρα πολλές αναγνώσεις 1 τομέα μπορούν να γίνουν παράλληλα
    - Καλό για χειρισμό συναλλαγών, μικρά αρχεία



RAID 4: Προβλήματα

- Εγγραφές 1 τομέα:
  - Πρέπει να διαβάσω την παλιά έκδοση του τομέα δεδομένων (και του τομέα ισοτιμίας)
    - 2 αναγνώσεις
  - Πρέπει να προσαρμόσω την τιμή του τομέα ισοτιμίας για την αντίστοιχη ταινία
  - Πρέπει να γράψω τις νέες τιμές του τομέα δεδομένων και ισοτιμίας
    - 2 εγγραφές
- Οι εγγραφές 1 τομέα (μικρές) γίνονται ακολουθιακά
  - Καθεμία χρειάζεται το δίσκο ισοτιμίας
    - 2 φορές
  - Ο δίσκος ισοτιμίας σημείο συμφόρησης

RAID 5

- Όπως το RAID 4
  - Όμως η πληροφορία ισοτιμίας μοιράζεται σε όλους τους δίσκους
  - Δεν υπάρχει πια δίσκος ισοτιμίας (και άρα και σημείο συμφόρησης)
    - Κάθε μικρή εγγραφή εξακολουθεί να απαιτεί ανάγνωση από και εγγραφή σε 2 δίσκους
    - Όμως μπορούμε να κάνουμε εγγραφές παράλληλα αν τα σύνολα δίσκων που «ακουμπάμε» είναι διαφορετικά
      - Δηλαδή αν είμαστε τυχεροί
- Χρησιμοποιείται συχνά
  - Γρηγορότερες μικρές αναγνώσεις, μεγάλες αναγνώσεις, μεγάλες εγγραφές

Εφαρμογές

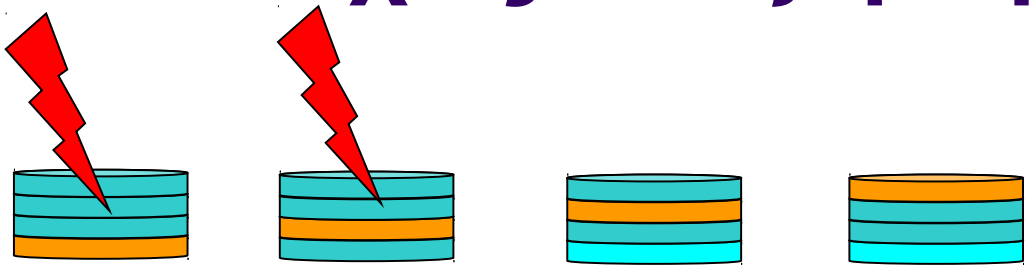
- RAID 0
  - Προσωρινή αποθήκευση / swap για συστήματα υψηλής επίδοσης
  - Αναξιόπιστο
- RAID 1
  - Καλή επίδοση, καλή αξιοπιστία
  - Εφαρμογές που απαιτούν υψηλή αξιοπιστία (π.χ. τράπεζες)
- RAID 5
  - «Φθηνή» αξιοπιστία και επίδοση
  - Άπειρες εφαρμογές

- Μπορεί να δείτε και άλλες, εξειδικευμένες υλοποιήσεις RAID (6, 7, 10, 53, 0+1 κλπ)

Είναι οι Αστοχίες Ανεξάρτητες;

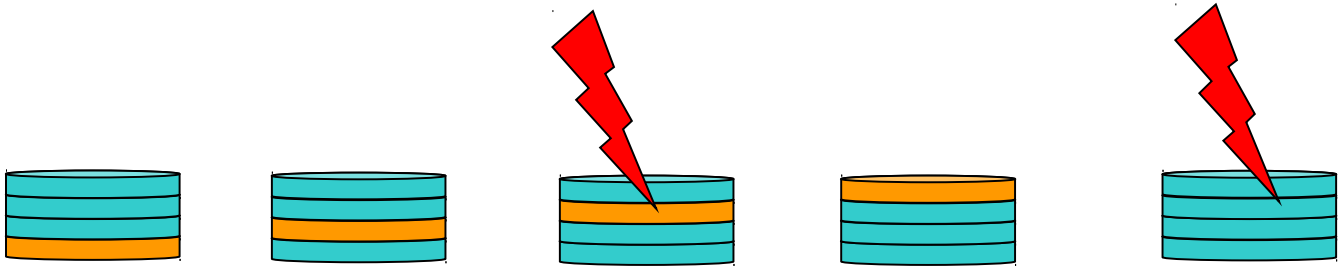
- Με τα RAID 1-5 οι αστοχίες δίσκων ανεκτές
- Αλλά όχι οι αστοχίες όλης της συστοιχίας
  - Πότε; Όταν αστοχήσουν πολλοί δίσκοι σε μικρό διάστημα
    - Τότε δεν είναι δυνατό να ανασυντεθούν τα δεδομένα του χαμένου δίσκου με XOR των υπολοίπων
  - Κάντε το σταυρό σας οι ταινίες του backup να μην έχουν πρόβλημα…
  - … και να έχετε ένα καλό και γρήγορο σύστημα backup
- Ευτυχώς πολλαπλές αστοχίες είναι «πολύ σπάνιες»
  - Επειδή οι αστοχίες δίσκων είναι ανεξάρτητες μεταξύ τους

Είναι οι Αστοχίες Ανεξάρτητες;



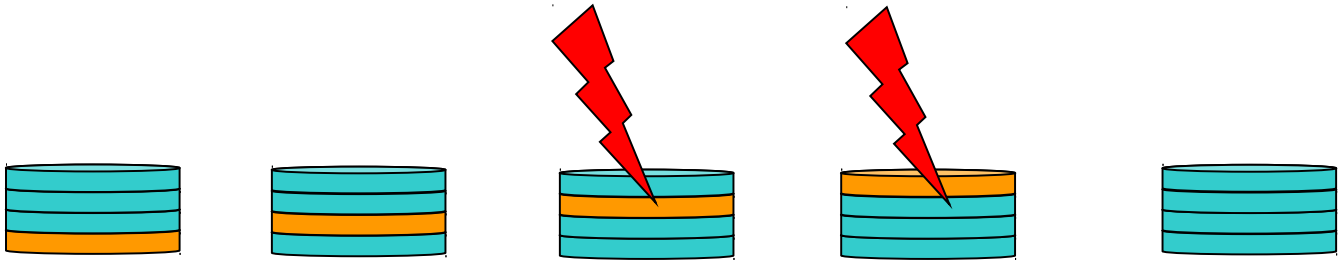
- 2 δίσκοι / καλωδιοταινία (π.χ. EIDE)

Είναι οι Αστοχίες Ανεξάρτητες;



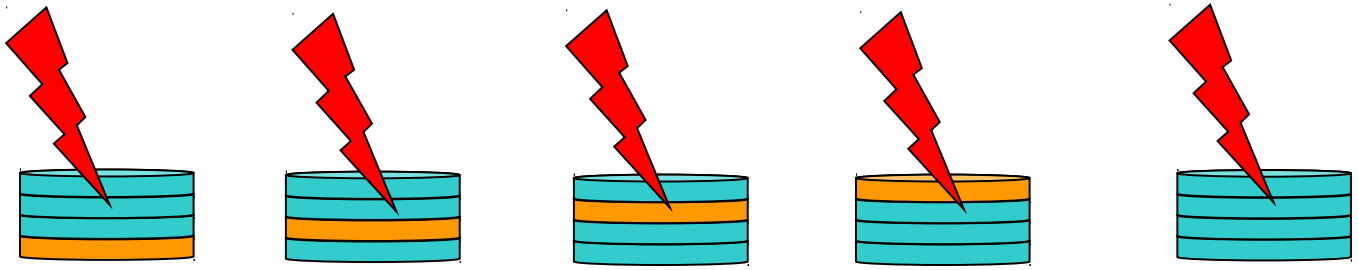
- Αν δε χρησιμοποιείς το δίσκο (έξτρα δίσκος για την αποκατάσταση), πού ξέρεις αν δουλεύει;

Είναι οι Αστοχίες Ανεξάρτητες;



- Μερικές μέρες απλά είναι κακές…
  - Καλύτερα να μένεις στο κρεβάτι σου

Είναι οι Αστοχίες Ανεξάρτητες;



- Η μπορεί να μας πέσει ο ουρανός στο κεφάλι
  - Ή ένας κεραυνός στο κεφάλι
    - Ή κάπου γύρω
  - Ή μια αιχμή τάσης
  - Ή μια φωτιά
  - Ή …