Κεφάλαιο 6

Αποθήκευση και είσοδος/έξοδος

Μία δε πάντων είσοδος εις τον βίον, έξοδός τε ίση.
---Παλαιά Διαθήκη (Σοφία Σολομώντος)

Η κατοικία του Warren Buffett

1958: $31.500

2010: $650.000

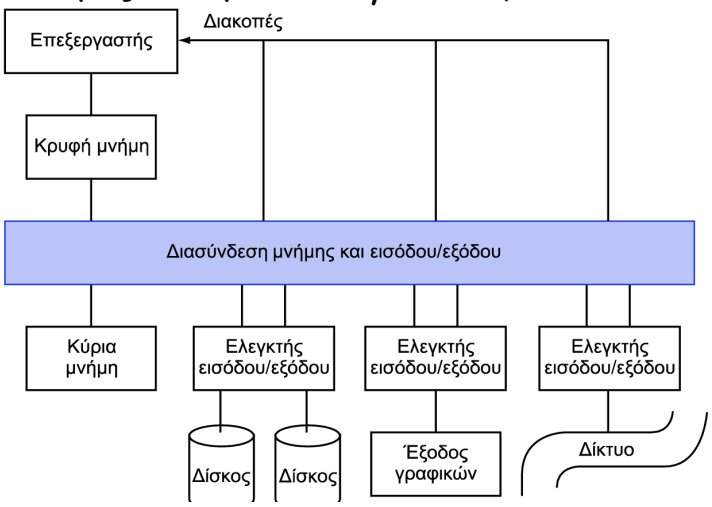
Μονάδα σκληρού δίσκου

1956
5 MB
$35.000

2021
500 GB
$90

Εισαγωγή

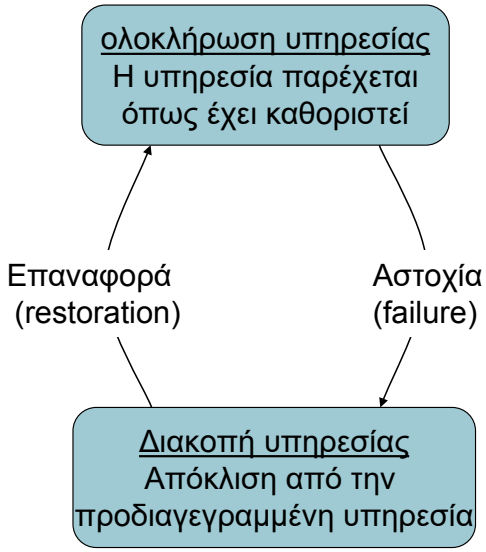
- Οι συσκευές εισόδου/εξόδου
  - Συμπεριφορά: είσοδος, έξοδος, αποθήκευση
  - Εταίροι: άνθρωπος ή μηχανή
  - Ρυθμός δεδομένων: byte/sec, transfers/sec



Χαρακτηριστικά συστήματος Ε/Ε

- Η φερεγγυότητα (dependability) είναι σημαντική
  - Ειδικά για συσκευές αποθήκευσης
- Μέτρα απόδοσης
  - **Λανθάνων** χρόνος (latency) ή **χρόνος απόκρισης** (response time)
  - **Διεκπεραιωτική ικανότητα** (throughput) ή **εύρος ζώνης** (bandwidth)
  - **Επιτραπέζια** και **ενσωματωμένα** συστήματα
    - Ενδιαφέρουν κυρίως για το χρόνο απόκρισης και την ποικιλομορφία των συσκευών
  - **Διακομιστές**
    - Ενδιαφέρουν κυρίως για τη διεκπεραιωτική ικανότητα και την επεκτασιμότητα των συσκευών

Φερεγγυότητα (dependability)



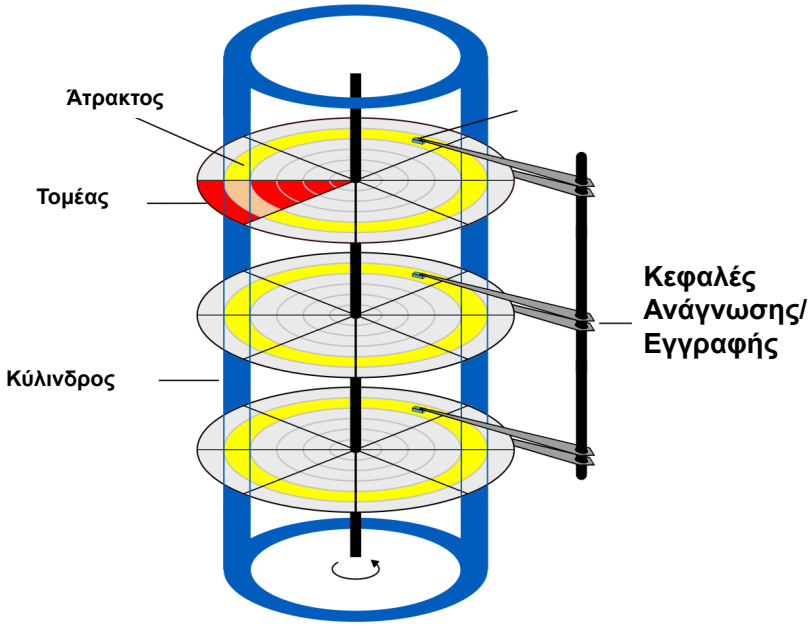
- Ελάττωμα (fault): αστοχία ενός συστατικού
- Μπορεί να οδηγήσει ή να μην οδηγήσει σε αστοχία του όλου συστήματος

Μέτρα φερεγγυότητας

- **Αξιοπιστία** (reliability): μέσος χρόνος πρώτης αστοχίας (MTTF)
- **Διακοπή υπηρεσίας** (service interruption): μέσος χρόνος επιδιόρθωσης (MTTR)
- **Μέσος χρόνος μεταξύ αστοχιών** (MTBF)
  - MTBF = MTTF + MTTR
- **Διαθεσιμότητα** (availability) = MTTF / (MTTF + MTTR)
- Βελτίωση διαθεσιμότητας
  - Αύξηση MTTF: αποφυγή ελαττώματος, ανοχή ελαττωμάτων, πρόβλεψη ελαττωμάτων
  - Μείωση MTTR: βελτιωμένα εργαλεία και διαδικασίες διάγνωσης και επιδιόρθωσης

Αποθήκευση στο δίσκο

- Μη πτητική μαγνητική αποθήκευση



Τομείς δίσκου και προσπέλαση

- Κάθε τομέας (sector) καταγράφει:
  - Την ταυτότητα τομέα (**sector ID**)
  - Δεδομένα (512 byte, 4096 byte προτεινόμενη τιμή)
  - Κώδικας διόρθωσης σφαλμάτων (error correcting code – **ECC**)
  - Πεδία συγχρονισμού και κενά (gaps)

- Η **προσπέλαση** ενός τομέα περιλαμβάνει:
  - **Καθυστέρηση** αναμονής σε ουρά αν εκκρεμούν άλλες προσπελάσεις
  - **Αναζήτηση** (seek): μετακίνηση των κεφαλών
  - **Λανθάνων χρόνος περιστροφής** (rotational latency)
  - **Μεταφορά δεδομένων**
  - **Επιβάρυνση ελεγκτή** (controller)

Παράδειγμα προσπέλασης δίσκου

- Τομέας των 512B, 15.000rpm (περιστροφές ανά λεπτό), μέσος χρόνος αναζήτησης 4ms, ρυθμός μεταφοράς 100MB/s, επιβάρυνση ελεγκτή 0.2ms, δίσκος αδρανής
- Μέσος χρόνος ανάγνωσης:

  4ms χρόνος αναζήτησης
  + ½ / (15,000/60) = 2ms λανθάνων χρόνος περιστροφής
  + 512 / 100MB/s = 0.005ms χρόνος μεταφοράς
  + 0.2ms καθυστέρηση ελεγκτή
  = 6.2ms
- **Αν** ο μέσος χρόνος αναζήτησης ήταν 1ms, τότε:
  - Μέσος χρόνος ανάγνωσης θα ήταν **3.2ms**

Ζητήματα απόδοσης δίσκου

- Οι κατασκευαστές αναφέρουν το **μέσο χρόνο αναζήτησης**, με βάση **όλες** τις πιθανές αναζητήσεις
- Η τοπικότητα και ο χρονοπρογραμματισμός του ΛΣ οδηγούν σε πολύ **μικρότερους** μέσους χρόνους αναζήτησης

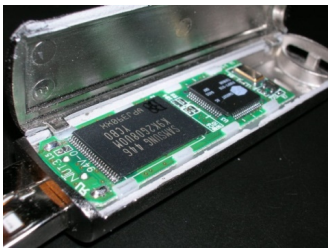
Ελεγκτές δίσκων

- «Έξυπνος» ελεγκτής δίσκου κατανέμει τους φυσικούς τομείς του δίσκου
  - Εμφανίζει τη διασύνδεση των λογικών τομέων (logical sector interface) στον υπολογιστή
  - SCSI, ATA, SATA ελεγκτές

- Οι μονάδες δίσκου περιλαμβάνουν **και κρυφές μνήμες**
  - Εκ των προτέρων προσκόμιση (prefetch) τομέων, με αναμονή προσπέλασής τους σύντομα
  - Αποφυγή αναζήτησης και καθυστέρησης περιστροφής

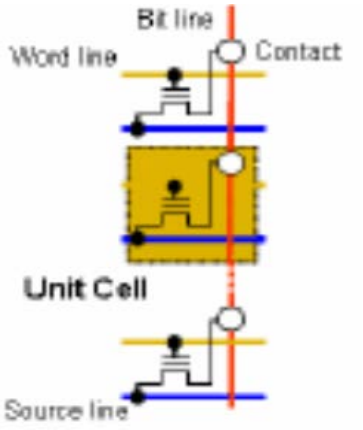
Αποθήκευση σε μνήμη φλας

- Μη πτητική, ηλεκτρονική αποθήκευση
  - 100× – 1000× ταχύτερη από το δίσκο
  - Μικρότερο φυσικό μέγεθος
  - Χαμηλότερη κατανάλωση ισχύος
  - Πιο εύρωστη, ανθεκτική σε μηχανικές δονήσεις
  - Αλλά κοστίζει περισσότερα €/GB (ανάμεσα στο δίσκο και την DRAM)



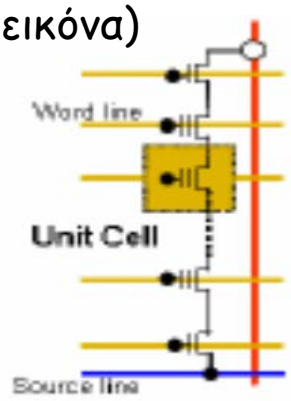
Μνήμη φλας NOR

- Κελί bit μοιάζει με πύλη NOR
  - Τυχαία προσπέλαση ανάγνωσης/εγγραφής
  - Χρησιμοποιείται για μνήμη εντολών σε ενσωματωμένα συστήματα



Μνήμη Φλας NAND

- Κελί bit μοιάζει με πύλη NAND
  - Πιο πυκνή (bit/επιφάνεια), αλλά προσπέλαση ενός ολόκληρου μπλοκ κάθε φορά
  - (Δεν έχει μεταλλική επαφή)
  - Φθηνότερη ανά GB
  - Χρήση σε USB keys, αποθήκευση μέσων (ήχος, εικόνα)



Χρήση - Φθορά Μνημών Φλάς

- Ραγδαία αύξηση της χρήσης μνημών Φλάς
  - Σε φορητές συσκευές
  - Σε διακομιστές υψηλών προδιαγραφών
- Τα bit της μνήμης φλας φθείρονται μετά από κάποιες χιλιάδες προσπελάσεις
  - Δεν είναι κατάλληλη να αντικαταστήσει πλήρως τη RAM ή το δίσκο
  - Εξισορρόπηση φθοράς (wear leveling): επαναχαρτογράφηση δεδομένων, σε λιγότερο χρησιμοποιημένα μπλοκ

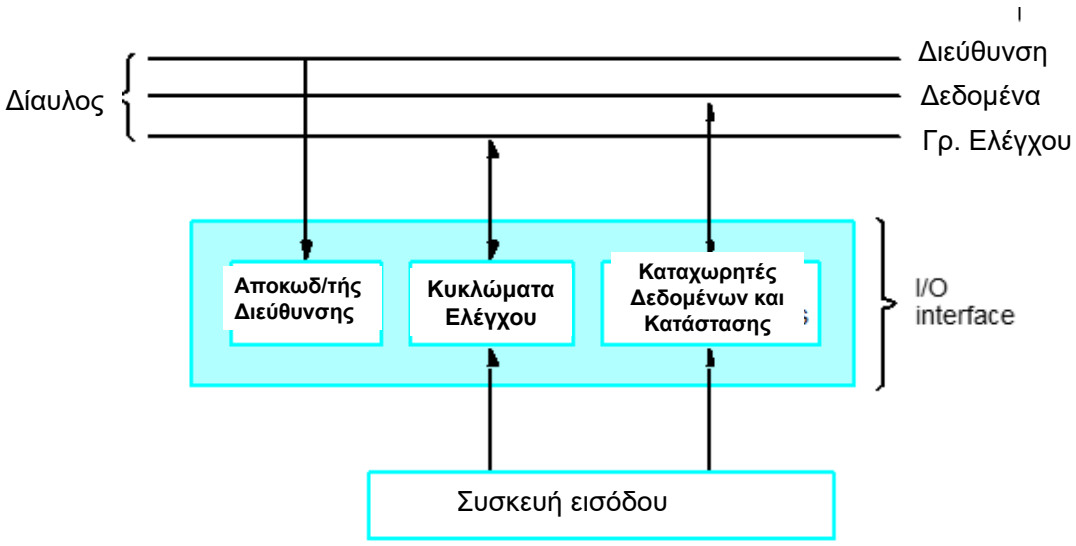
Συστατικά διασύνδεσης

- Ανάγκη διασύνδεσης μεταξύ
  - CPU, μνήμης, ελεγκτών Ε/Ε
- Δίαυλος: κοινόχρηστο κανάλι επικοινωνίας
  - Παράλληλο σύνολο αγωγών για δεδομένα και συγχρονισμό της μεταφοράς τους
  - Μπορεί να αποτελέσει σημείο συμφόρησης
- Η απόδοση περιορίζεται από φυσικούς παράγοντες
  - Μήκος αγωγού, αριθμός συνδέσεων
- Πιο πρόσφατη εναλλακτική: σειριακές συνδέσεις υψηλής ταχύτητας με μεταγωγούς
  - Όπως στα δίκτυα

Τύποι διαύλου

- Δίαυλοι **επεξεργαστή-μνήμης**
  - Κοντοί, μεγάλη ταχύτητα
  - Η σχεδίαση ταιριάζει με την οργάνωση μνήμης
- Δίαυλοι **εισόδου/εξόδου**
  - Μακρύτεροι, επιτρέπουν πολλές συνδέσεις
  - Προδιαγράφονται με **πρότυπα** για λόγους διαλειτουργικότητας (interoperability)
  - Σύνδεση με το δίαυλο επεξεργαστή-μνήμης μέσω μιας γέφυρας (bridge)

Δίαυλος Ε/Ε

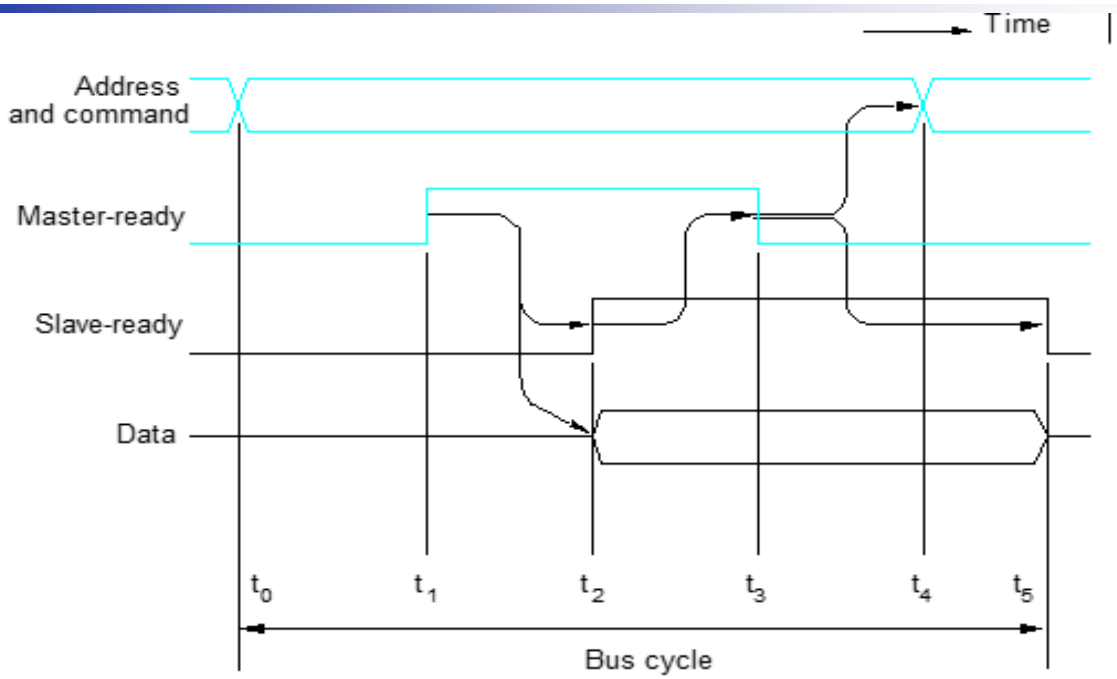

Δίαυλος
Διεύθυνση
Δεδομένα
Γρ. Ελέγχου
Αποκωδ/τής Διεύθυνσης
Κυκλώματα Ελέγχου
Καταχωρητές Δεδομένων και Κατάστασης
I/O interface
Συσκευή εισόδου

Σήματα διαύλου και συγχρονισμός

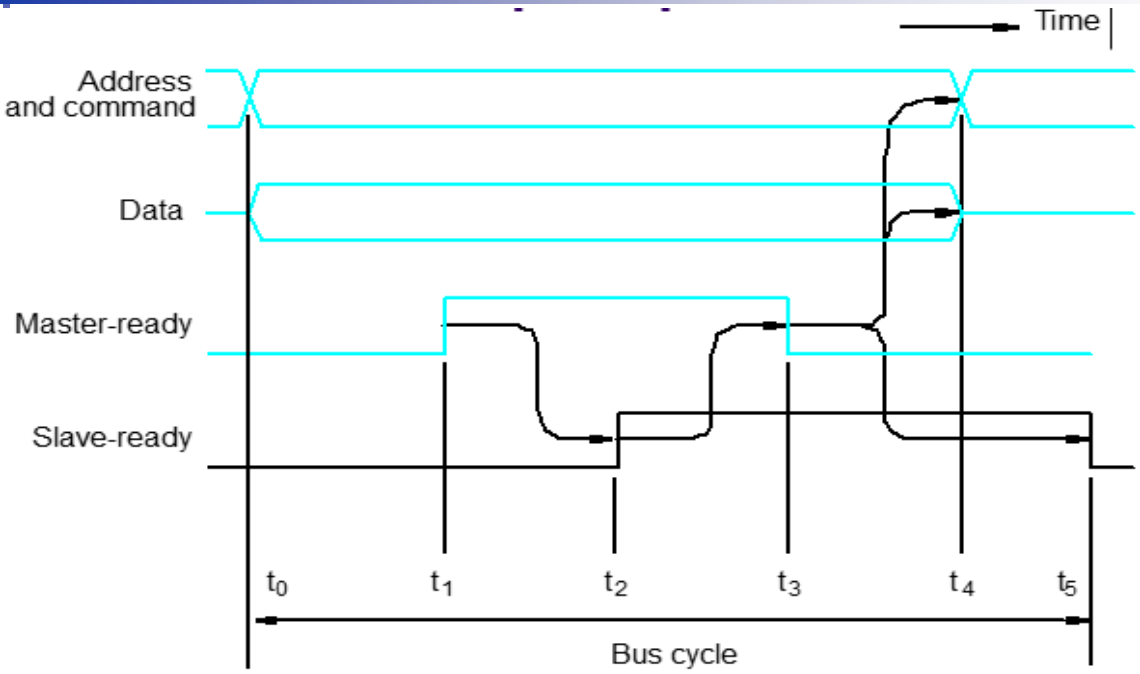
- Γραμμές δεδομένων
  - Μεταφέρουν διεύθυνση και δεδομένα
  - Με πολύπλεξη ή ξεχωριστά
- Γραμμές ελέγχου
  - Δείχνουν τον τύπο δεδομένων, συγχρονίζουν τις συναλλαγές (transactions)

- Σύγχρονη μεταφορά
  - Χρησιμοποιεί ρολόϊ διαύλου
- Ασύγχρονη μεταφορά
  - Χρησιμοποιεί γραμμές ελέγχου αίτησης/επιβεβαίωσης (request/acknowledge) για χειραψία (handshaking)

Ασύγχρονη Είσοδος Δεδομένων



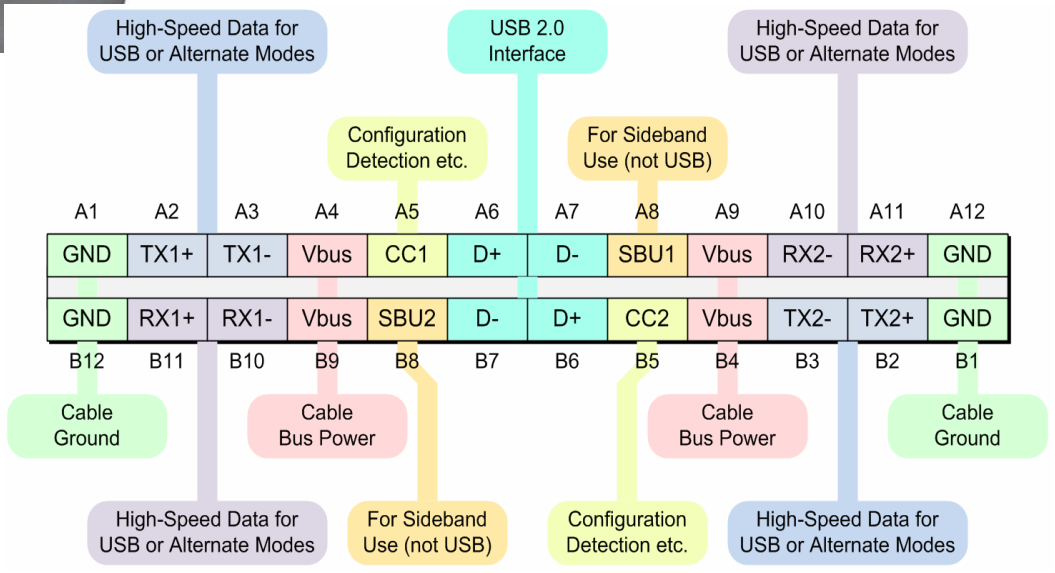
Ασύγχρονη Έξοδος Δεδομένων



Αρχαία Ιστορία

Το Μέλλον (USB Type-C)

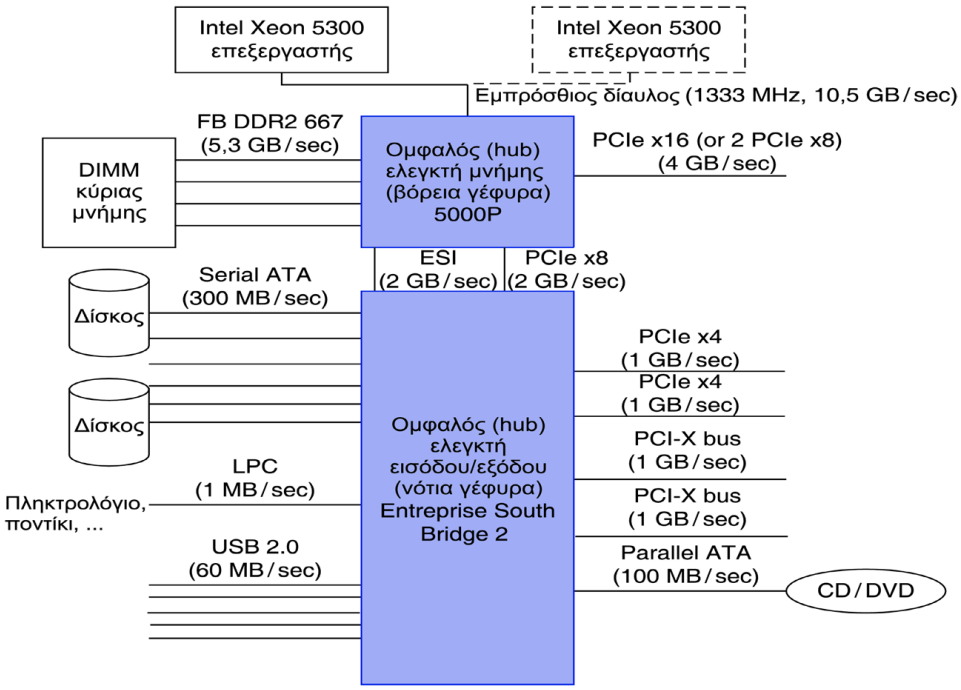
20 Gbps, 100W



Παραδείγματα διαύλων Ε/Ε

	Firewire	USB 2.0	PCI Express	Serial ATA	Serial Attached SCSI
Πρόθεση χρήσης	Εξωτερική	Εξωτερική	Εσωτερική	Εσωτερική	Εξωτερική
Συσκευές ανά κανάλι	63	127	1	1	4
Εύρος δεδομένων	4	2	2/lane	4	4
Μέγιστο εύρος ζώνης	50MB/s ή 100MB/s	0.2MB/s, 1.5MB/s, ή 60MB/s	250MB/s/lane 1×, 2×, 4×, 8×, 16×, 32×	300MB/s	300MB/s
Σύνδεση «εν θερμώ»	Ναι	Ναι	Εξαρτάται	Ναι	Ναι
Μέγιστο μήκος	4.5m	5m	0.5m	1m	8m
Πρότυπο	IEEE 1394	USB Implementers Forum	PCI-SIG	SATA-IO	INCITS TC T10

Τυπικό σύστημα Ε/Ε x86 PC



Διαχείριση εισόδου/εξόδου

- Το ΛΣ είναι ο ενδιάμεσος για την Ε/Ε
  - Πολλά προγράμματα μοιράζονται πόρους εισόδου/εξόδου
    - Χρειάζεται προστασία και χρονοπρογραμματισμός
  - Η Ε/Ε προκαλεί ασύγχρονες διακοπές
    - Ίδιος μηχανισμός με τις εξαιρέσεις
  - Ο προγραμματισμός Ε/Ε είναι περίπλοκος
    - Το ΛΣ παρέχει αφαιρέσεις στα προγράμματα

Διαταγές εισόδου/εξόδου

- Συσκευές Ε/Ε: τις διαχειρίζεται το υλικό των ελεγκτών Ε/Ε
  - Μεταφέρουν δεδομένα από/προς τη συσκευή
  - Συγχρονίζουν τις λειτουργίες με λογισμικό
- Καταχωρητές διαταγών (command registers)
  - Αναγκάζουν τη συσκευή να κάνει κάτι
- Καταχωρητές κατάστασης (status registers)
  - Δείχνουν τι κάνει η συσκευή και την εμφάνιση σφαλμάτων
- Καταχωρητές δεδομένων (data registers)
  - Εγγραφής: μεταφέρουν δεδομένα σε μια συσκευή
  - Ανάγνωσης: μεταφέρουν δεδομένα από μια συσκευή

Χαρτογράφηση καταχωρητών Ε/Ε

- Ε/Ε με χαρτογράφηση μνήμης (memory mapped I/O)
  - οι καταχωρητές προσπελάζονται στον ίδιο «χώρο δ/νσεων» με την μνήμη
  - ο αποκωδικοποιητής δ/νσεων κάνει το διαχωρισμό
  - Το ΛΣ χρησιμοποιεί μηχανισμό μετάφρασης δ/νσεων ώστε να τους κάνει προσπελάσιμους μόνο στον πυρήνα (kernel) του ΛΣ
- Εντολές Ε/Ε
  - Ξεχωριστές εντολές για προσπέλαση καταχωρητών Ε/Ε
  - Μπορούν να εκτελεστούν μόνο σε **κατάσταση πυρήνα**
  - Παράδειγμα: x86

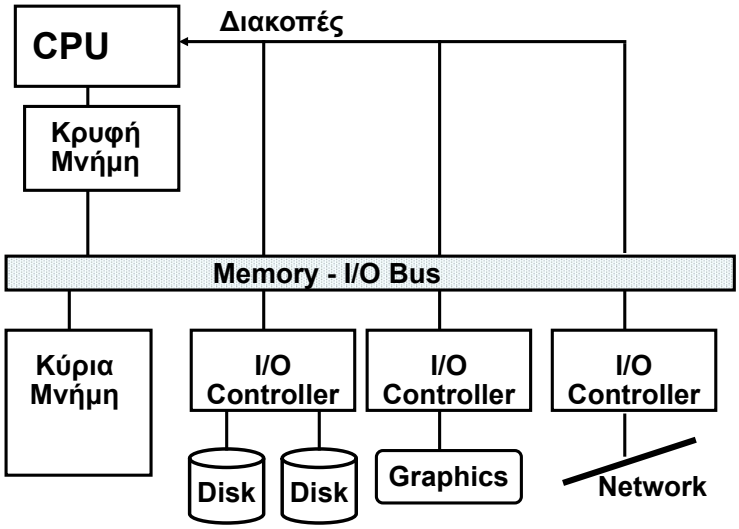
Περιόδευση (polling)

- Περιοδικός έλεγχος του καταχωρητή κατάστασης (status register) της Ε/Ε
  - Αν η συσκευή είναι έτοιμη, καμία λειτουργία
  - Αν υπάρχει σφάλμα, ανάληψη δράσης
- Χρησιοποιείται σε μικρά ή απλά ενσωματωμένα συστήματα
- Προβλέψιμος χρονισμός
  - Χαμηλό κόστος υλικού

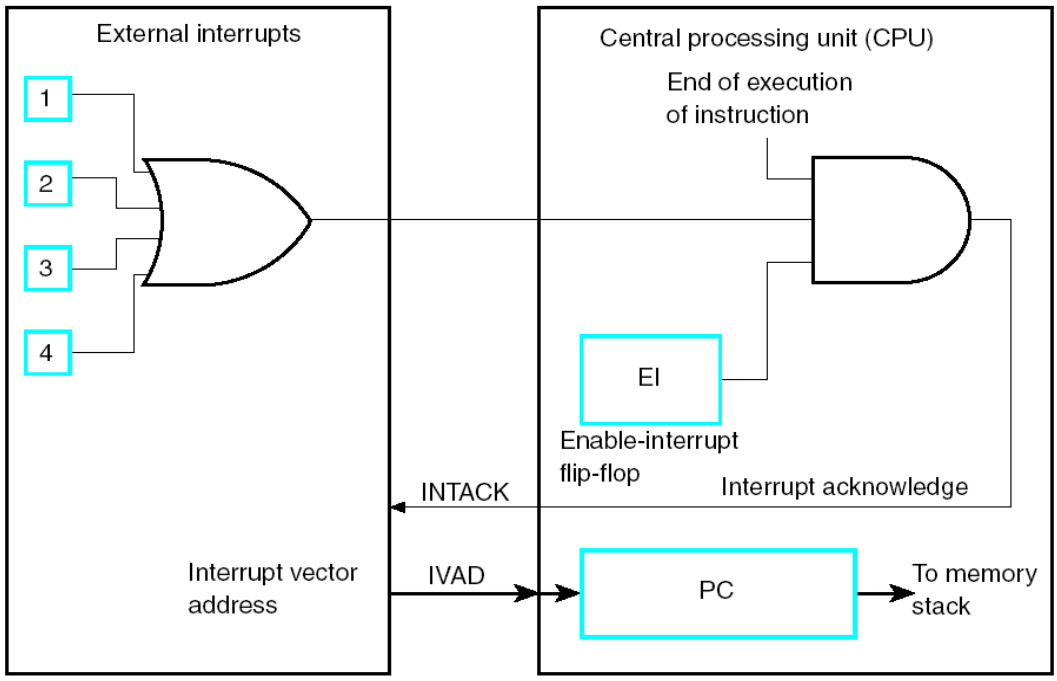
Διακοπές (interrupts)

- Όταν μια συσκευή είναι έτοιμη ή όταν συμβεί σφάλμα
  - ο ελεγκτής **διακόπτει** τη CPU
- Η διακοπή είναι σαν την **εξαίρεση** (exception)
  - Αλλά δεν συγχρονίζεται με την εκτέλεση των εντολών
  - Μπορεί να ζητήσει να κληθεί ο **χειριστής** (handler) μεταξύ εντολών
  - Πληροφορία για το αίτιο (cause) προσδιορίζει συχνά τη συσκευή που προκαλεί διακοπή
- Διακοπές με προτεραιότητες
  - οι συσκευές που χρειάζονται πιο επείγουσα προσοχή λαμβάνουν υψηλότερη προτεραιότητα
  - Μπορούν να διακόψουν το χειριστή μιας διακοπής χαμηλότερης προτεραιότητας

Τυπικό σύστημα Ε/Ε



Υλικό Διακοπών



Χειρισμός Διακοπής

Μικρολειτουργίες CPU

SP ← SP – 1

M[SP] ← PC

SP ← SP – 1

M[SP] ← PSR

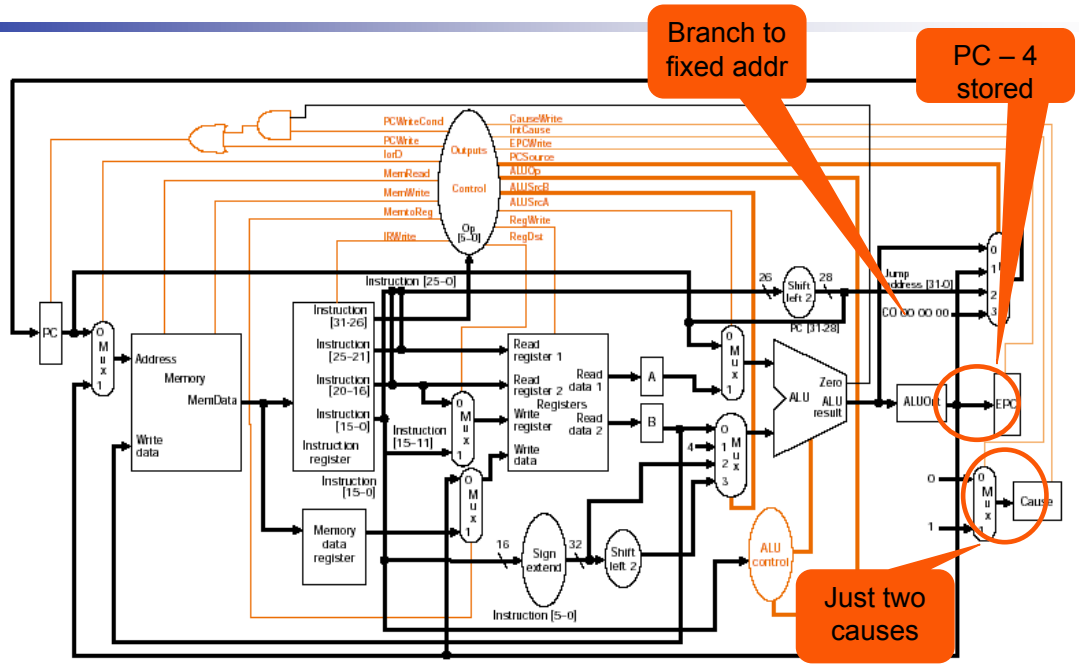
- PSR is “processor status register”

EI ← 0

INT_ACK ← 1

PC ← IVAD

Διακοπή σφάλματος εντολής MIPS



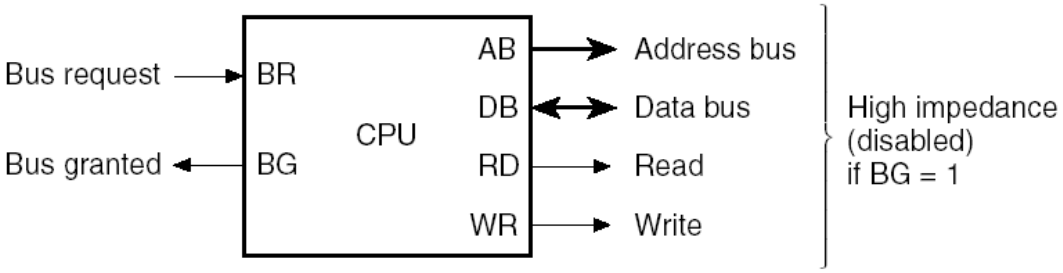
Undefined instruction and arithmetic overflow

Μεταφορά δεδομένων Ε/Ε

- Περιόδευση και Ε/Ε οδηγούμενη από διακοπές
  - Η CPU μεταφέρει δεδομένα μεταξύ μνήμης και καταχωρητών δεδομένων Ε/Ε
  - Χρονοβόρα διαδικασία για συσκευές υψηλής ταχύτητας
- Άμεση προσπέλαση μνήμης (direct memory access – DMA)
  - Το ΛΣ παρέχει την αρχική δ/νση μνήμης
  - ο Ελεγκτής Ε/Ε κάνει μεταφορά προς/από τη μνήμη αυτόνομα
  - ο Ελεγκτής προκαλεί διακοπή όταν ολοκληρώσει τη μεταφορά ή σε περίπτωση σφάλματος

Λειτουργία DMA

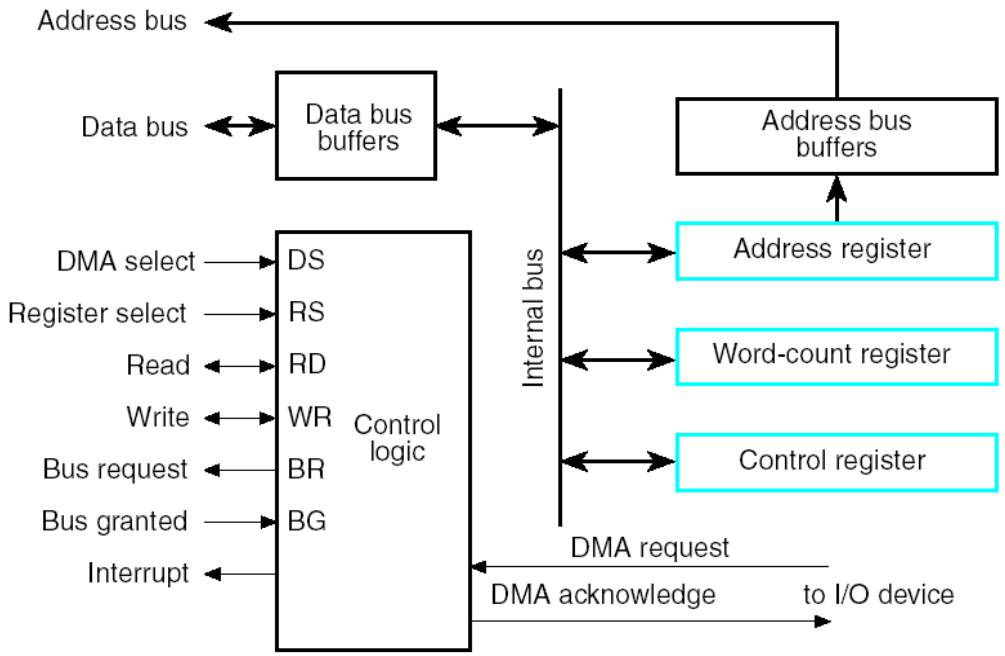
- DMA ζητά τον δίαυλο (assert BR)
- CPU αποδέχεται το αίτημα (assert BG)
- CPU οδηγεί τα σήματά της σε υψηλή αντίσταση Hi-Z



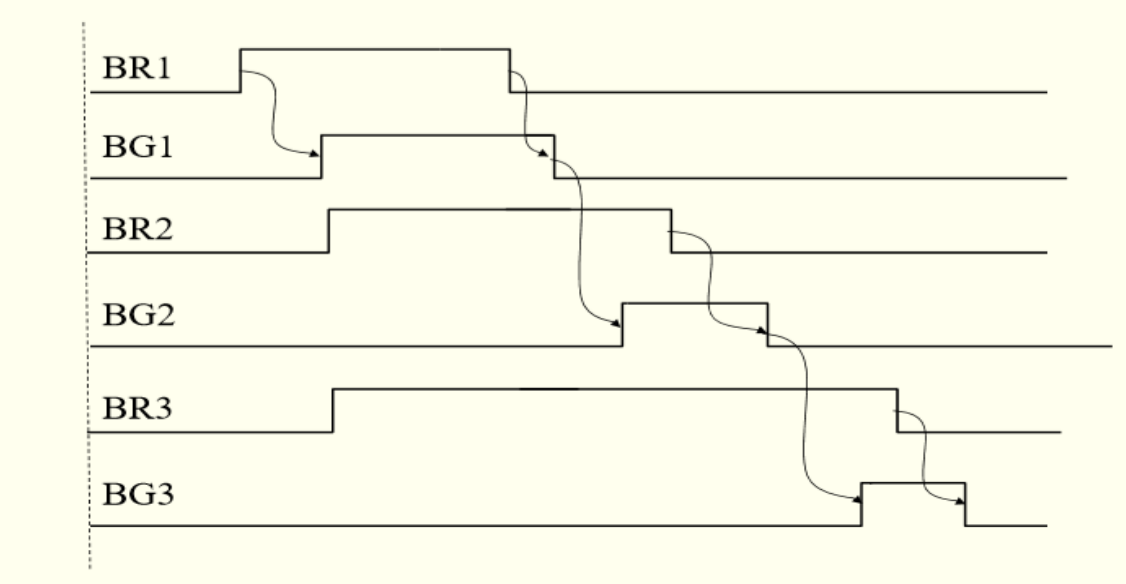
Συνεχής – DMA μεταφέρει όλα τα δεδομένα (π.χ. ένα τμήμα δίσκου) αδιάκοπα

Με Ριπές – DMA «κλέβει» ελεύθερους κύκλους (*cycle stealing*), όταν η CPU δεν χρησιμοποιεί τον δίαυλο.

DMA Controller



Πολλαπλές αιτήσεις DMA



Αλληλεπίδραση DMA/Cache

- **Αν** DMA γράφει σε μπλοκ μνήμης που βρίσκεται & στην κρυφή μνήμη
  - Το αντίγραφο της κρυφής μνήμης γίνεται «παλιό»
- Αν η κρυφή μνήμη είναι **ετερόχρονης** εγγραφής και το μπλοκ είναι «ακάθαρτο», και η DMA **διαβάζει** το μπλοκ της μνήμης
  - Διαβάζει τα «παλιά» δεδομένα. Απαιτείται διόρθωση.
- Πρέπει να εγγυηθούμε τη Συνοχή (coherence) της κρυφής μνήμης. Επιλογές:
  - «**Εκκένωση**» (flush) των μπλοκ από τη κρυφή μνήμη αν πρόκειται αυτά να χρησιμοποιηθούν σε DMA, ή
  - Χρήση θέσεων μνήμης που **δεν** αποθηκεύονται στη κρυφή μνήμη (non-cacheable) για τις λειτουργίες Ε/Ε

Μέτρηση απόδοσης Ε/Ε

- Η απόδοση Ε/Ε εξαρτάται από
  - Υλικό: CPU, μνήμη, ελεγκτές, δίαυλοι
  - Λογισμικό: λειτουργικό σύστημα, σύστημα διαχείρισης βάσης δεδομένων, εφαρμογή
  - Φορτίο εργασίας: ρυθμοί και μοτίβα αιτήσεων
- Η σχεδίαση του συστήματος Ε/Ε μπορεί να κάνει συμβιβασμούς μεταξύ χρόνου απόκρισης και ρυθμού διεκπεραίωσης

Μετροπρογράμματα επεξεργασίας συναλλαγών

- Συναλλαγές (Transactions)
  - Μικρές προσπελάσεις δεδομένων σε ένα σύστημα διαχείρισης βάσης δεδομένων (DBMS)
  - Το ενδιαφέρον είναι στο ρυθμό Ε/Ε, όχι στον ρυθμό δεδομένων
- Μέτρηση ρυθμού διεκπεραίωσης (throughput)
  - Υπόκειται σε περιορισμούς χρόνου απόκρισης και χειρισμό αστοχιών
  - ACID (Atomicity/Ατομικότητα, Consistency/Συνέπεια, Isolation/Απομόνωση, Durability/Αντοχή)
  - Συνολικό κόστος ανά συναλλαγή
- Μετροπρογράμματα του Transaction Processing Council (TPC, www.tcp.org)
  - TPC-APP: διακομιστής εφαρμογών και υπηρεσιών ιστού
  - TCP-C: περιβάλλον καταχώρισης παραγγελιών
  - TCP-E: επεξεργασία συναλλαγών μεσιτικού γραφείου
  - TPC-H: υποστήριξη αποφάσεων — κατά περίπτωση (ad-hoc) ερωτήματα με προσανατολισμό επιχειρήσεις

Μετροπρογράμματα συστήματος Αρχείων και Ιστού

- SPEC System File System (SFS)
  - Συνθετικό φορτίο εργασίας για διακομιστή NFS, με βάση παρακολούθηση πραγματικών συστημάτων
  - Αποτελέσματα
    - Ρυθμός διεκπεραίωσης, throughput (λειτουργίες/sec)
    - Χρόνος απόκρισης (μέσο ms/λειτουργία)
- SPEC Web Server benchmark
  - Μετράει τις ταυτόχρονες συνεδρίες (sessions) χρηστών, με βάση τον απαιτούμενο ρυθμό διεκπεραίωσης ανά συνεδρία
  - Τρία φορτία εργασίας: Τραπεζική, Ηλεκτρονικό εμπόριο, και Υποστήριξη

Ε/Ε έναντι απόδοσης CPU

- Νόμος του Amdahl
  - Δεν αγνοούμε την επίδοση της Ε/Ε, καθώς η παραλληλία αυξάνει την απόδοση των υπολογισμών
- Παράδειγμα
  - Το μετροπρόγραμμα διαρκεί 90s χρόνο CPU, 10s χρόνο Ε/Ε
  - Διπλάσιες CPU κάθε 2 χρόνια
    - Ε/Ε αμετάβλητη

Έτος	Χρόνος CPU	Χρόνος Ε/Ε	Παρελθών χρόνος	% Χρόνος Ε/Ε
Τώρα	90s	10s	100s	10%
+2	45s	10s	55s	18%
+4	23s	10s	33s	31%
+6	11s	10s	21s	47%

Δίσκοι RAID

- Πλεονασματικές συστοιχίες φθηνών (ανεξάρτητων) δίσκων – Redundant Array of Inexpensive (Independent) Disks
  - Χρήση πολλών μικρότερων δίσκων (σε σχέση με ένα μεγάλο)
  - Η παραλληλία βελτιώνει την απόδοση
  - Και πρόσθετοι δίσκοι για αποθήκευση πλεονασματικών δεδομένων
- Παρέχουν σύστημα αποθήκευσης με ανοχή σε ελαττώματα (fault tolerant)

- RAID 0
  - Χωρίς πλεονασμό
    - Τα δεδομένα απλώς μοιράζονται σε πολλούς δίσκους
  - Αλλά βελτιώνει την απόδοση

RAID 1 & 2

- RAID 1: Δημιουργία ειδώλων (mirroring)
  - N + N δίσκοι, επανάληψη δεδομένων
    - Εγγραφή δεδομένων και στο δίσκο δεδομένων και στο δίσκο είδωλο
    - Σε περίπτωση αστοχίας δίσκου, ανάγνωση από το είδωλο
- RAID 2: Κώδικας διόρθωσης σφαλμάτων(Error correcting code – ECC)
  - N + E δίσκοι (π.χ., 10 + 4)
  - Χωρισμός δεδομένων σε επίπεδο bit στους N δίσκους
  - Δημιουργία ECC των E bit
  - Υπερβολικά πολύπλοκο, δε χρησιμοποιείται στην πράξη

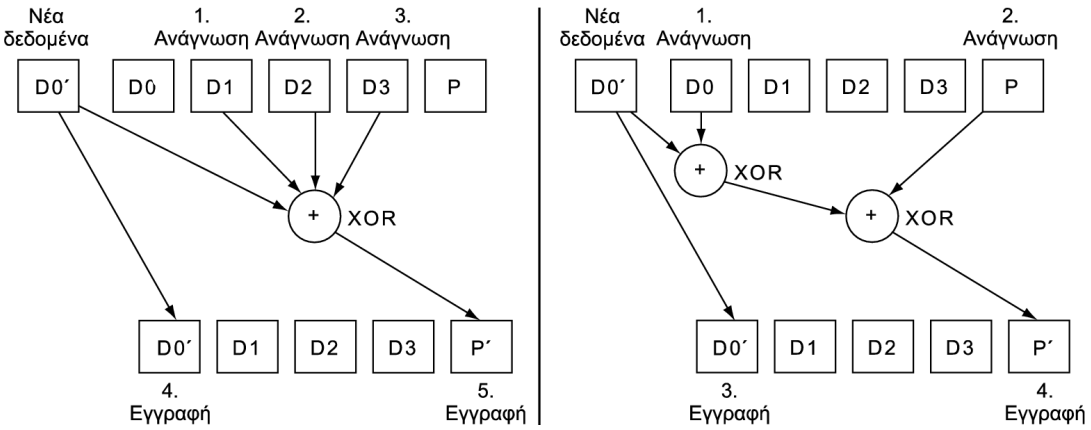
RAID 3: Ισοτιμία πλέξης bit

- Bit-Interleaved Parity
- N + 1 δίσκοι
  - Δεδομένα μοιράζονται σε N δίσκους σε επίπεδο byte
  - Πλεονασματικός δίσκος αποθηκεύει την ισοτιμία
  - Προσπέλαση ανάγνωσης
    - Ανάγνωση όλων των δίσκων
  - Προσπέλαση εγγραφής
    - Δημιουργία νέας ισοτιμίας και ενημέρωση όλων των δίσκων
  - Σε περίπτωση αστοχίας
    - Χρήση ισοτιμίας για επανασύσταση των χαμένων δεδομένων
- Δεν χρησιμοποιείται ευρέως

RAID 4: Ισοτιμία πλέξης μπλοκ

- Block-Interleaved Parity
- N + 1 δίσκοι
  - Τα δεδομένα μοιράζονται σε N δίσκους σε επίπεδο μπλοκ
  - Πλεονασματικός δίσκος αποθηκεύει την ισοτιμία για μια ομάδα μπλοκ
  - Προσπέλαση ανάγνωσης
    - Διαβάζει μόνο το δίσκο που περιέχει το ζητούμενο μπλοκ
  - Προσπέλαση εγγραφής
    - Απλώς διαβάζει το δίσκο οι οποίος περιέχει το μπλοκ που τροποποιείται, και το δίσκο ισοτιμίας
    - Υπολογισμός νέας ισοτιμίας, ενημέρωση δίσκου δεδομένων και δίσκου ισοτιμίας
  - Σε περίπτωση αστοχίας
    - Χρήση ισοτιμίας για την επανασύσταση των χαμένων δεδομένων
- Δεν χρησιμοποιείται ευρέως

RAID 3 έναντι RAID 4



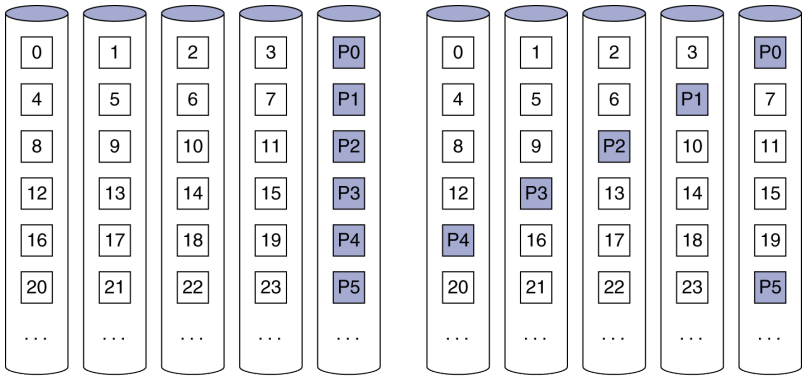
$P=D0 \oplus D1 \oplus D2 \oplus D3$, $A \oplus A=0$, $0 \oplus X=X$

$P' = D0' \oplus D1 \oplus D2 \oplus D3 =$

$= D0' \oplus D0 \oplus D0 \oplus D1 \oplus D2 \oplus D3 = D0' \oplus D0 \oplus P$

RAID 5: Κατανεμημένη ισοτιμία

- N + 1 δίσκοι
  - Όπως το RAID 4, αλλά τα μπλοκ ισοτιμίας κατανέμονται στους δίσκους
    - Αποφεύγει τη δημιουργία σημείου συμφόρησης (bottleneck) στον δίσκο ισοτιμίας
- Ευρεία χρήση



RAID 6: Πλεονασμός P + Q

- P + Q Redundancy
- N + 2 δίσκοι
  - Σαν το RAID 5, αλλά με δύο «παρτίδες» ισοτιμίας
  - Μεγαλύτερη ανοχή σε ελαττώματα μέσω περισσότερου πλεονασμού
- Πολλαπλά RAID
  - Πιο προηγμένα συστήματα δίνουν παρόμοια ανοχή σε ελαττώματα με καλύτερη απόδοση

Περίληψη RAID

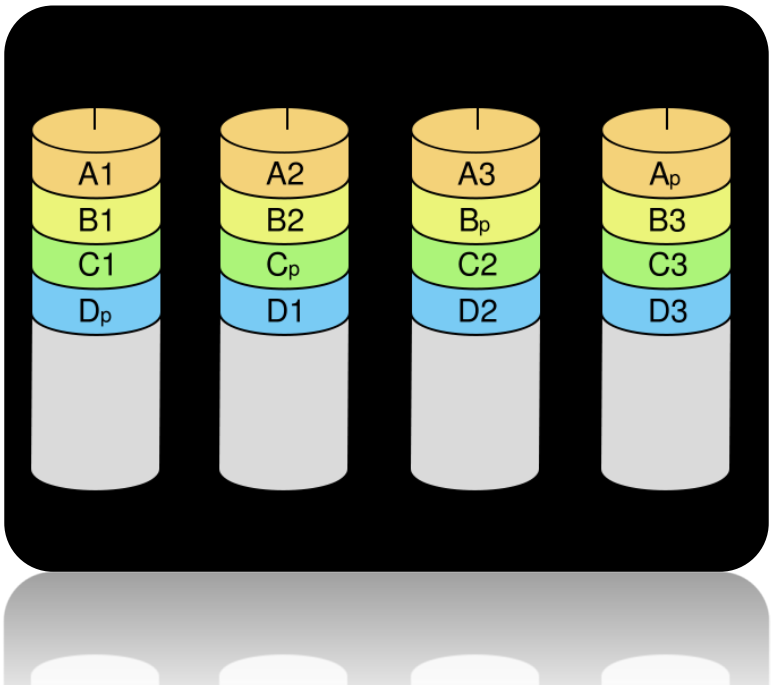
- Το RAID μπορεί να βελτιώσει την απόδοση και τη διαθεσιμότητα (availability)
  - Η υψηλή διαθεσιμότητα απαιτεί «εν θερμώ» εναλλαγή (hot swapping)
- Υποθέτει ότι οι αστοχίες δίσκων είναι ανεξάρτητες
- Δείτε το "Hard Disk Performance, Quality and Reliability"
  - http://www.pcguide.com/ref/hdd/perf/index.htm

RAID 5 – Παράδειγμα

Συνεχόμενα blocks γράφονται εναλλάξ στους δίσκους, ενώ κατανέμεται σε αυτούς και ένα block ισοτιμίας.

Παρέχει υψηλή απόδοση στις αναγνώσεις, αφού αυτές μπορούν να γίνουν από πολλούς δίσκους εναλλάξ.

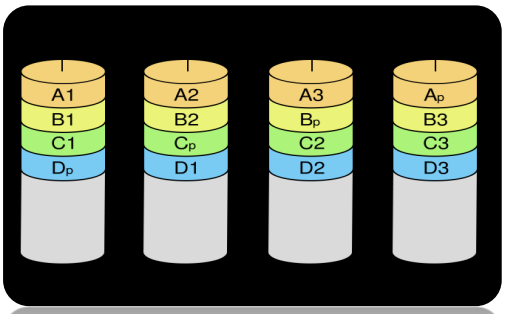
Παρέχει αξιοπιστία, αφού τα δεδομένα μπορούν να ανακτηθούν από τους υπόλοιπους δίσκους.



RAID 5 - Παράδειγμα

Έστω ότι διαθέτουμε 4 δίσκους.

Ας θεωρήσουμε ότι οι δίσκοι έχουν τα παρακάτω δεδομένα:



	DISK0	DISK1	DISK2	DISK3
STRIPE0	0100	0101	0010	
STRIPE1	0010	0000		0100
STRIPE2	0011		1010	1000
STRIPE3		0001	1101	1010

XOR		
Είσοδος		Έξοδος
0	0	0
0	1	1
1	0	1
1	1	0

Στα κίτρινα σημεία, τοποθετούνται τα δεδομένα ισοτιμίας, ως το Exclusive-OR (XOR) του ίδιου stripe όλων των δίσκων.

RAID 5 - Παράδειγμα

	DISK0	DISK1	DISK2	DISK3
STRIPE0	**0100**	**0101**	**0010**	**0011**
STRIPE1	0010	0000		0100
STRIPE0	0011		1010	1000
STRIPE3		0001	1101	1010

STRIPE0.DISK3 = 0100 XOR 0101 XOR 0010 = 0011

RAID 5 - Παράδειγμα

	DISK0	DISK1	DISK2	DISK3
STRIPE0	0100	0101	0010	0011
STRIPE1	**0010**	**0000**	**0110**	**0100**
STRIPE2	0011		1010	1000
STRIPE3		0001	1101	1010

STRIPE0,DISK3 = 0100 XOR 0101 XOR 0010 = 0011
STRIPE1,DISK2 = 0010 XOR 0000 XOR 0100 = 0110

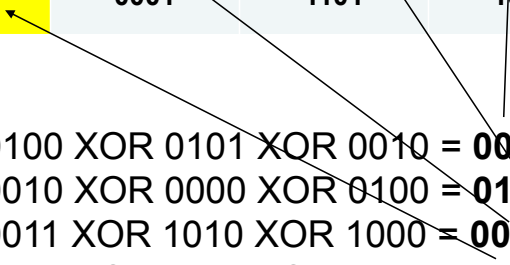
RAID 5 - Παράδειγμα

	DISK0	DISK1	DISK2	DISK3
STRIPE0	0100	0101	0010	0011
STRIPE1	0010	0000	0110	0100
STRIPE2	**0011**	**0001**	**1010**	**1000**
STRIPE3		0001	1101	1010

STRIPE0,DISK3 = 0100 XOR 0101 XOR 0010 = 0011
STRIPE1,DISK2 = 0010 XOR 0000 XOR 0100 = 0110
STRIPE2,DISK1 = 0011 XOR 1010 XOR 1000 = 0001

RAID 5 - Παράδειγμα

	DISK0	DISK1	DISK2	DISK3
STRIPE0	0100	0101	0010	0011
STRIPE1	0010	0000	0110	0100
STRIPE2	0011	0001	1010	1000
STRIPE3	**0110**	**0001**	**1101**	**1010**



STRIPE0,DISK3 = 0100 XOR 0101 XOR 0010 = **0011**
STRIPE1,DISK2 = 0010 XOR 0000 XOR 0100 = **0110**
STRIPE2,DISK1 = 0011 XOR 1010 XOR 1000 = **0001**
STRIPE3,DISK0 = 0001 XOR 1101 XOR 1010 = **0110**

RAID 5 – Παράδειγμα (εγγραφή)

Έστω ότι γίνεται η εγγραφή του στοιχείου 1101 στο block 2

	DISK0	DISK1	DISK2	DISK3
STRIPE0	0100	0101	~~0010~~ **1101**	~~0011~~ **1100**
STRIPE1	0010	0000	0110	0100
STRIPE2	0011	0001	1010	1000
STRIPE3	0110	0001	1101	1010

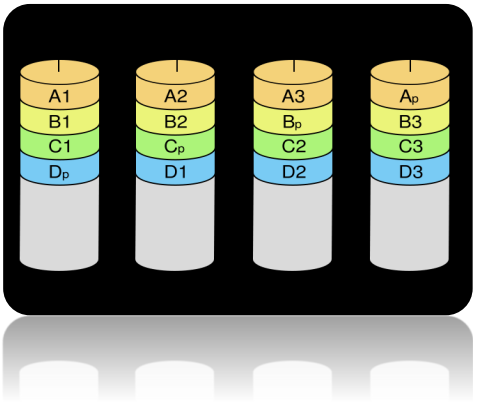
ο ελεγκτής RAID κάνει την εγγραφή του στοιχείου στο αντίστοιχο block ... και ταυτόχρονα ξαναδημιουργεί την ισοτιμία για το συγκεκριμένο stripe, χρησιμοποιώντας την **παλιά τιμή**, την **νέα τιμή** και την **παλιά ισοτιμία**

STRIPE0,DISK3 = 0010 XOR 1101 XOR 0011 = 1100

RAID 5 – Παράδειγμα (βλάβη)

Τι γίνεται αν χαλάσει ένας δίσκος;

Ας θεωρήσουμε ότι οι 4 δίσκοι έχουν τα παρακάτω δεδομένα:



	DISK0	DISK1	DISK2	DISK3
STRIPE0	0100	0101	1101	1100
STRIPE1	0010	0000	0110	0100
STRIPE2	0011	0001	1010	1000
STRIPE3	0110	0001	1101	1010

Έστω ότι χαλάει ο DISK2…

RAID 5 - Παράδειγμα

	DISK0	DISK1	DISK2	DISK3
STRIPE0	0100	0101	1101	1100
STRIPE1	0010	0000	0110	0100
STRIPE2	0011	0001	1010	1000
STRIPE3	0110	0001	1101	1010

Ο ελεγκτής RAID εξυπηρετεί τις αιτήσεις για τις πληροφορίες του DISK2, χρησιμοποιώντας **όλους** τους άλλους δίσκους + την ισοτιμία.

STRIPE0,DISK2 = 0100 XOR 0101 XOR 1100 = 1101
STRIPE1,DISK2 = 0010 XOR 0000 XOR 0100 = 0110
STRIPE2,DISK2 = 0011 XOR 0001 XOR 1000 = 1010
STRIPE3,DISK2 = 0110 XOR 0001 XOR 1010 = 1101

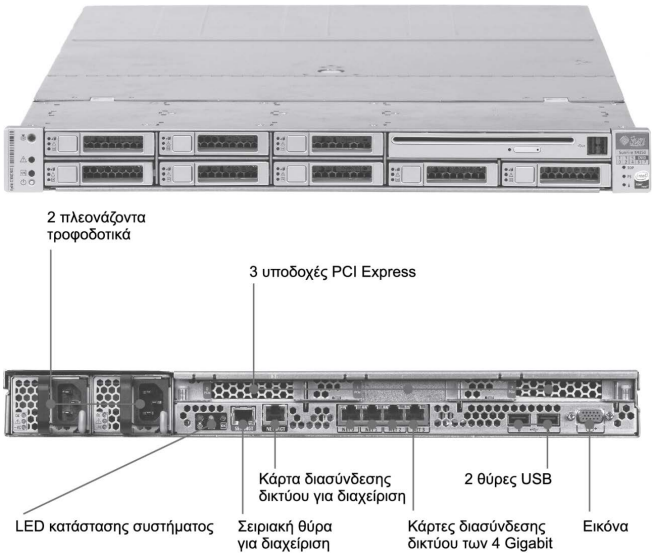
*Κάθε ανάγνωση του χαλασμένου δίσκου, αντιστοιχεί σε αναγνώσεις σε **όλους** τους υπόλοιπους δίσκους.*

Διακομιστές

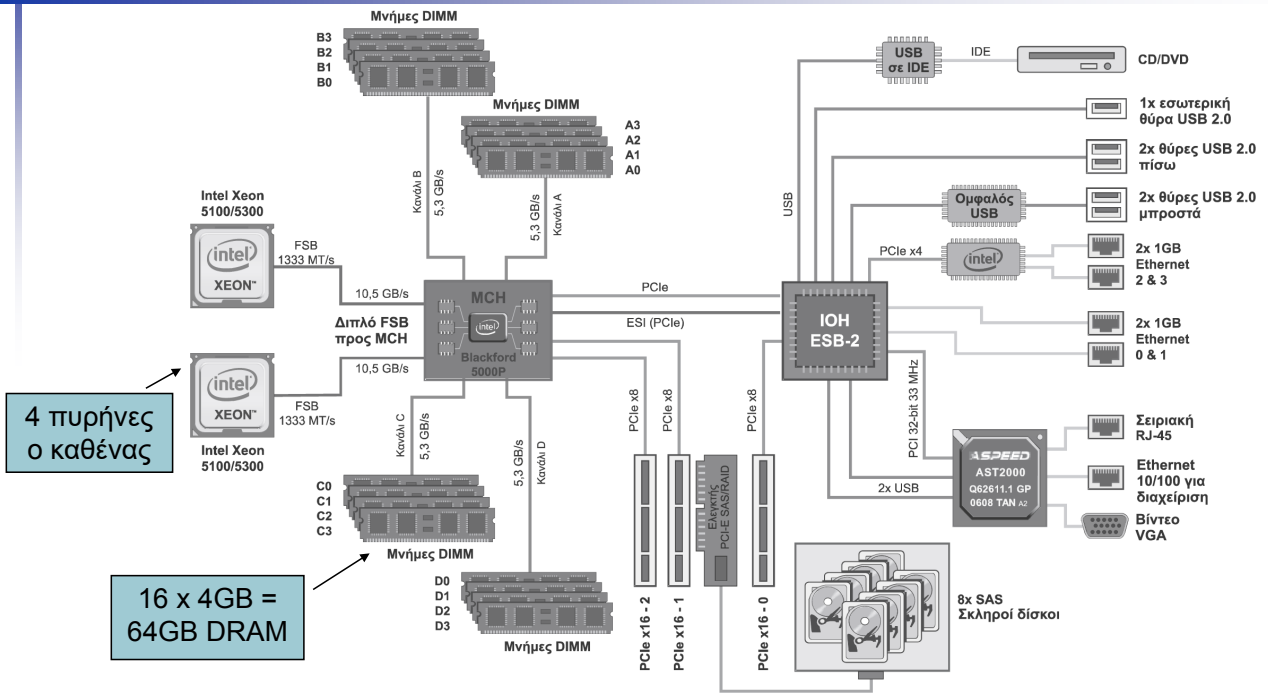
- οι εφαρμογές εκτελούνται όλο και περισσότερο σε διακομιστές (servers)
  - Αναζήτηση στον Ιστό, εφαρμογές γραφείου, εικονικοί κόσμοι, …
- Απαιτούνται μεγάλοι διακομιστές κέντρων δεδομένων
  - Πολλοί επεξεργαστές, συνδέσεις δικτύου, μαζική αποθήκευση
  - Περιορισμοί χώρου και ηλεκτρικής ισχύος
- Εξοπλισμός διακομιστών για ικριώματα (racks) των 19 ιντσών
  - Ύψος σε πολλαπλάσια 1.75 ιντσών (1U)

Διακομιστές για ικρίωμα

Διακομιστής Sun Fire x4150 1U



Διακομιστής Sun Fire x4150 1U



Παράδειγμα σχεδίασης συστήματος Ε/Ε…

- Σύστημα Sun Fire x4150
  - Φορτίο εργασίας: αναγνώσεις δίσκου των **64KB**
    - Κάθε λειτουργία Ε/Ε απαιτεί 200.000 εντολές κώδικα χρήστη και 100.000 εντολές του ΛΣ
  - Κάθε CPU: 10^9 εντολές/sec
  - FSB (Front Side Bus, Εμπρόσθιος δίαυλος): 10.6 GB/sec (max)
  - DRAM DDR2 στα 667MHz: 5.336 GB/sec
  - PCI-E 8× δίαυλος: 8 × 250MB/sec = 2GB/sec
  - Δίσκοι: 15.000 rpm, 2,9ms μέσος χρόνος αναζήτησης, 112MB/sec ρυθμός μεταφοράς
- Ποιος είναι ο ρυθμός Ε/Ε που μπορεί να διατηρηθεί;
  - Για τυχαίες αναγνώσεις, και για ακολουθιακές αναγνώσεις

Παράδειγμα σχεδίασης (συνέχεια)

- Ρυθμός Ε/Ε για τις CPU
  - Ανά πυρήνα: 10^9 / (100.000 + 200.000) = 3333 λειτουργίες/sec
  - Για 8 πυρήνες: 26667 λειτουργίες/sec
- **Τυχαίες** αναγνώσεις, ρυθμός Ε/Ε για τους δίσκους
  - Έστω ότι ο πραγματικός χρόνος αναζήτησης είναι το 1/4 τού αναφερόμενου
  - Χρόνος/λειτουργία = αναζήτηση + λανθάνων χρόνος + μεταφορά
    = 2.9ms/4 + 4ms/2 + 64KB/(112MB/s) = 3,3ms
  - Άρα, έχουμε 1000/3,3 = 303 λειτουργίες/sec ανά δίσκο,
    - Ήτοι: 8x303=2.424 λειτουργίες/sec για 8 δίσκους
- **Ακολουθιακές** αναγνώσεις
  - 112MB/s / 64KB = 1750 λειτουργίες/sec ανά δίσκο
  - Ήτοι: 8x1.750 = 14.000 λειτουργίες/sec για 8 δίσκους

Παράδειγμα σχεδίασης (συνέχεια)

- Ρυθμός Ε/Ε του PCI-E
  - 2GB/sec / 64KB/λειτουργία = 31.250 λειτουργίες/sec
- Ρυθμός Ε/Ε της DRAM
  - 5.336 GB/sec / 64KB/λειτουργία = 83.375 λειτουργίες/sec
- Ρυθμός Ε/Ε του FSB
  - Έστω ότι μπορούμε να διατηρήσουμε το 50% του μέγιστου ρυθμού
  - 5.3 GB/sec / 64KB = 81.540 λειτουργίες/sec ανά FSB
  - 163.080 λειτουργίες/sec για 2 FSB
- ο αδύναμος κρίκος: **οι δίσκοι**
  - 2.424 λειτουργίες/sec τυχαίες, 14.000 λειτουργίες/sec ακολουθιακές
  - Τα άλλα συστατικά έχουν άφθονο περιθώριο για να αντέξουν αυτόν τον ρυθμό λειτουργιών

Πλάνη: Φερεγγυότητα δίσκων

- Αν ένας κατασκευαστής δίσκων δίνει ότι το MTTF είναι 1.200.000 ώρες (140 χρόνια)
  - Ένας δίσκος θα έχει τόσο μεγάλη διάρκεια;
- Λάθος: αυτός είναι ο **μέσος** χρόνος της πρώτης αστοχίας
  - Ποια είναι η κατανομή των αστοχιών;
  - Τι θα γίνει αν έχουμε 1000 δίσκους;
    - Πόσοι θα αστοχούν κάθε χρόνο; 365x24 = 8.760 h/year

$$\text{Annual Failure Rate (AFR)} = \frac{1000 \text{ disks} \times 8760 \text{ hrs/disk}}{1200000 \text{ hrs/failure}} = 0.73\%$$

Ήτοι, αναμένονται περί τις **7 βλάβες ανά έτος**, από το πρώτο έτος!

Παγίδα: μέγιστη απόδοση

- Οι μέγιστοι ρυθμοί Ε/Ε είναι σχεδόν αδύνατον να πραγματωθούν
  - Συνήθως, κάποια άλλα συστατικά του συστήματος περιορίζουν την απόδοση
  - Π.χ., μεταφορές στη μνήμη μέσω ενός διαύλου
    - Σύγκρουση με την ανανέωση (refresh) της DRAM
    - Συναγωνισμός διαιτησίας με άλλου κύριους (masters) του διαύλου
  - Π.χ., δίαυλος PCI: μέγιστο εύρος ζώνης ~133 MB/sec
    - Στην πράξη, μπορεί να διατηρηθεί το 80MB/sec κατά μέγιστο

Συμπερασματικές παρατηρήσεις

- Μέτρα απόδοσης Ε/Ε
  - Ρυθμός διεκπεραίωσης, χρόνος απόκρισης
  - Η φερεγγυότητα και το κόστος είναι επίσης σημαντικά
- Χρησιμοποιούνται δίαυλοι για τη σύνδεση CPU, μνήμης, ελεγκτών Ε/Ε
  - Περιόδευση, διακοπές, DMA
- Μετροπρογράμματα Ε/Ε
  - TPC, SPECSFS, SPECWeb
- RAID
  - Βελτιώνει την απόδοση και τη φερεγγυότητα