

Ιόνιο Πανεπιστήμιο – Τμήμα Πληροφορικής
Αρχιτεκτονική Υπολογιστών
2021-22

Κρυφές Μνήμες

(οργάνωση, λειτουργία και απόδοση)

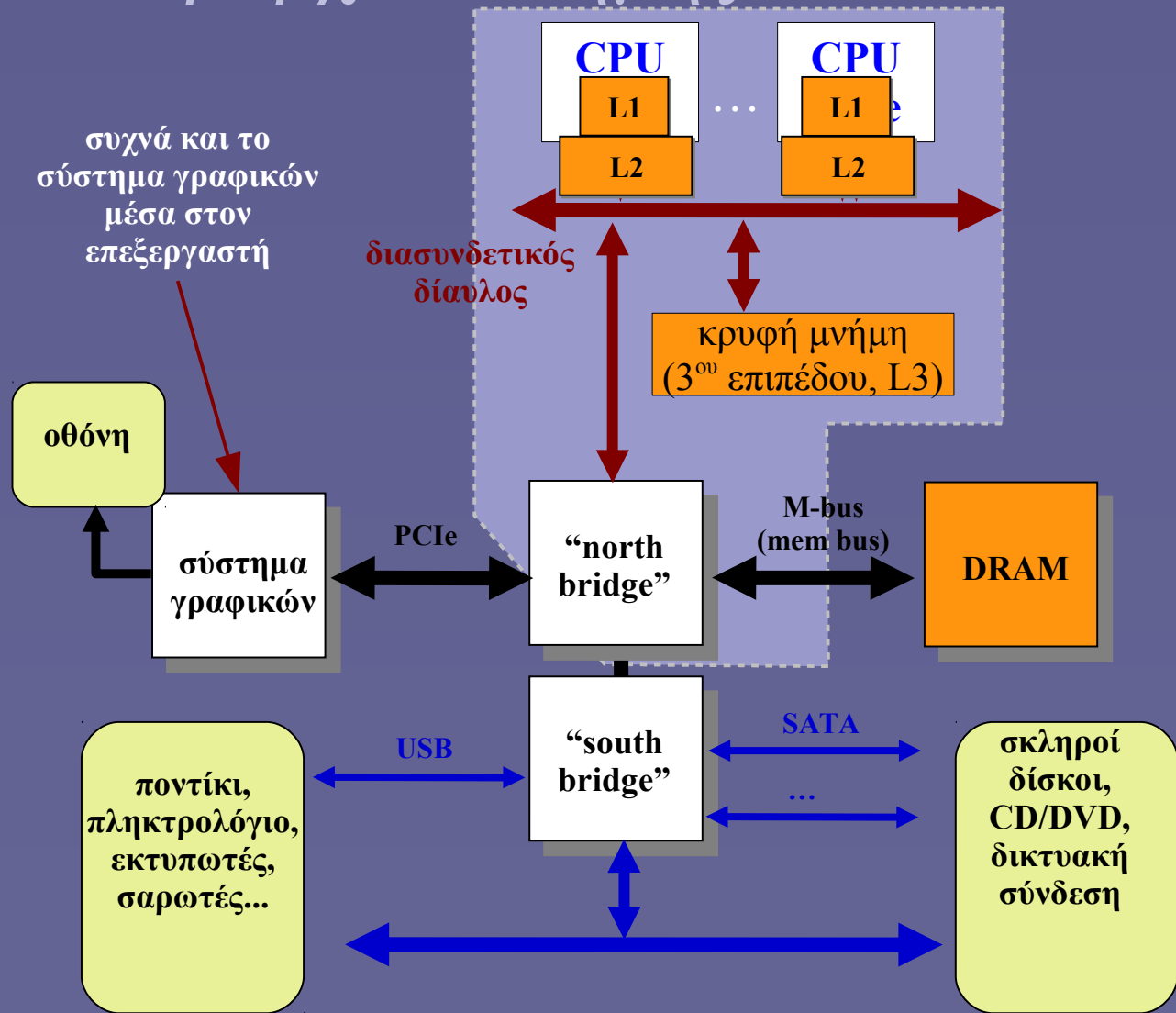
<http://mixstef.github.io/courses/comparch/>

Μ.Στεφανιδάκης



Ιεραρχία Μνήμης

- Ιεραρχία Μνήμης



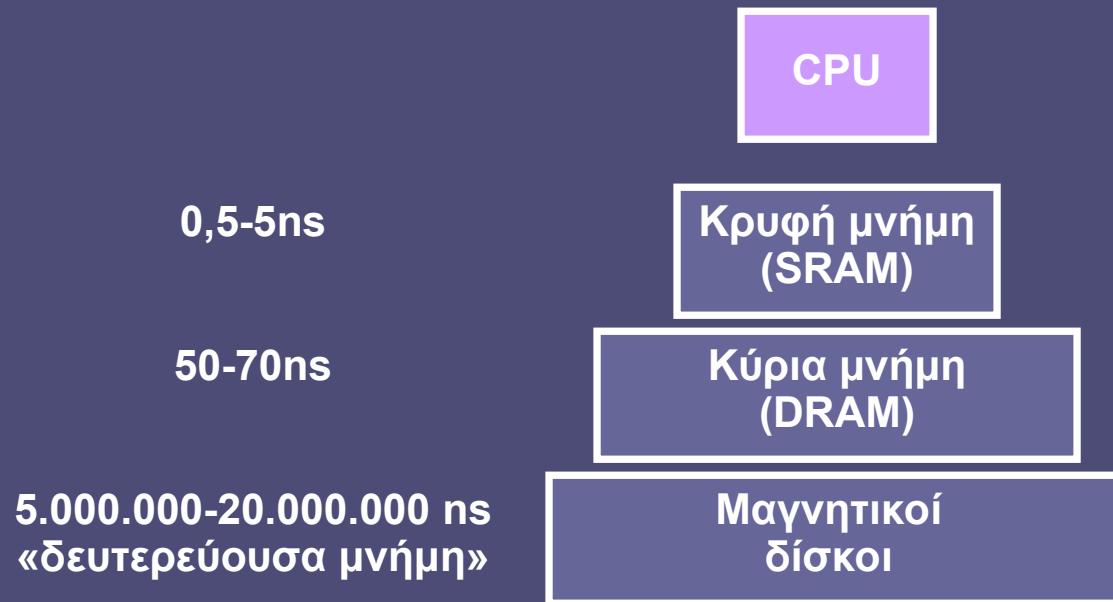
Για βελτίωση της απόδοσης του συστήματος μνήμης (και του συνολικού υπολογιστικού συστήματος)

Σκοπός της Ιεραρχίας Μνήμης

- Ιεραρχία Μνήμης

- Προσέγγιση της ιδανικής μνήμης
 - Ο επεξεργαστής να βλέπει “μνήμη”
 - Με την **ταχύτητα** του υψηλότερου επιπέδου
 - Και το **μέγεθος** του χαμηλότερου επιπέδου

Για να επιτύχει τον σκοπό της η ιεραρχία μνήμης εκμεταλλεύεται την αρχή της **τοπικότητας**



Ιεραρχία μνήμης και τοπικότητα

- Ιεραρχία Μνήμης

«ένα πρόγραμμα εκτελεί το 90% των εντολών του μέσα στο 10% του κώδικά του»

- **Χρονική Τοπικότητα**
 - Εάν προσπελαστεί μια θέση μνήμης, είναι πολύ πιθανό να προσπελαστεί ξανά στο άμεσο μέλλον
 - Παράδειγμα: οι εντολές ενός βρόχου (loop)
- **Εφαρμογή:**
 - Δεδομένα και εντολές που βρίσκονται ήδη κοντύτερα στον επεξεργαστή (π.χ. στην κρυφή μνήμη) θα προσπελαστούν πολύ γρηγορότερα

Ιεραρχία μνήμης και τοπικότητα

- Ιεραρχία Μνήμης

- Χωρική Τοπικότητα

- Εάν προσπελαστεί μια θέση μνήμης, είναι πολύ πιθανό να προσπελαστούν και οι γειτονικές θέσεις στο άμεσο μέλλον
 - Εντολές προγραμμάτων
 - Δεδομένα σε πίνακες κλπ

- Εφαρμογή:

- Εάν προσπελαστεί μια θέση μνήμης, μεταφέρονται και οι διπλανές της λέξεις στη μνήμη του υψηλότερου επιπέδου
 - Μεταφορά σε μπλοκ (πολλαπλές λέξεις μνήμης)
- Γρηγορότερη προσπέλαση όταν ζητηθούν και αυτές

Κρυφές μνήμες

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη

- Σημαντικό τμήμα στην ιεραρχία μνήμης
- Εξέλιξη συστημάτων κρυφής μνήμης
 - 1962: οι πρώτες ιεραρχίες μνήμης (Atlas computer)
 - Όχι όμως κρυφή μνήμη
 - 1965: η πρώτη περιγραφή κρυφής μνήμης (Wilkes)
 - Ο πρώτος υπολογιστής με κρυφή μνήμη (IBM 360/85)
 - 1968: η πρώτη χρησιμοποίηση του όρου “**cache memory**”
 - Στη συνέχεια:
 - Πολλαπλά επίπεδα κρυφής μνήμης (L1, L2, L3)
 - Βελτιωμένες αρχιτεκτονικές κρυφής μνήμης

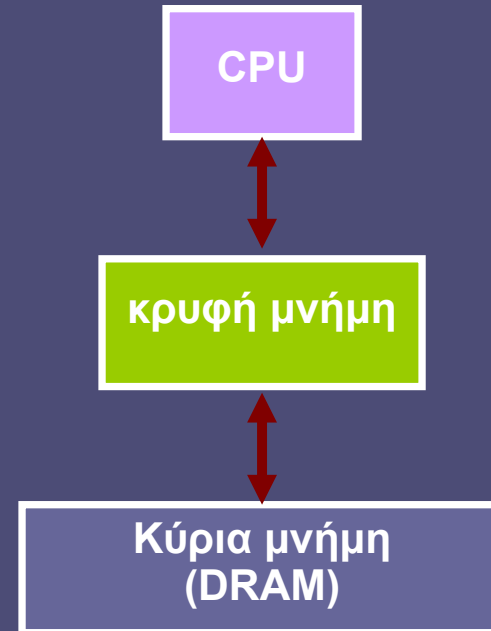
Απλό μοντέλο ιεραρχίας μνήμης

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη

Οι αρχές λειτουργίας της απλής ιεραρχίας μπορούν να επεκταθούν σε πολλαπλά επίπεδα (κρυφή μνήμη L1, L2 και L3)

Η διαχείριση της κρυφής μνήμης γίνεται από το υλικό

διαφανώς προς τις εφαρμογές



- Τα δεδομένα βρίσκονται αρχικά στην κύρια μνήμη
- Η κρυφή μνήμη περιέχει **υποσύνολο** των δεδομένων
- Μεταφορά μεταξύ επιπέδων μνήμης σε **μπλοκ** λέξεων

Αποθήκευση δεδομένων στην Ιεραρχία Μνήμης

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη

- **Αποθήκευση δεδομένων**
 - Τα υψηλότερα επίπεδα της ιεραρχίας μνήμης (πιο κοντά στις ΚΜΕ) είναι **υποσύνολα** των χαμηλότερων
 - Όλα τα δεδομένα αποθηκεύονται τελικά στο χαμηλότερο επίπεδο (κύρια μνήμη)
- **Μεταφορά δεδομένων**
 - Αντιγραφή από επίπεδο σε επίπεδο
 - Το ελάχιστο σύνολο δεδομένων που μεταφέρεται μεταξύ δύο επιπέδων ονομάζεται **μπλοκ**
 - Πολλαπλά bytes (πολλές λέξεις μαζί)

Αναζήτηση δεδομένων στην Ιεραρχία Μνήμης

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη

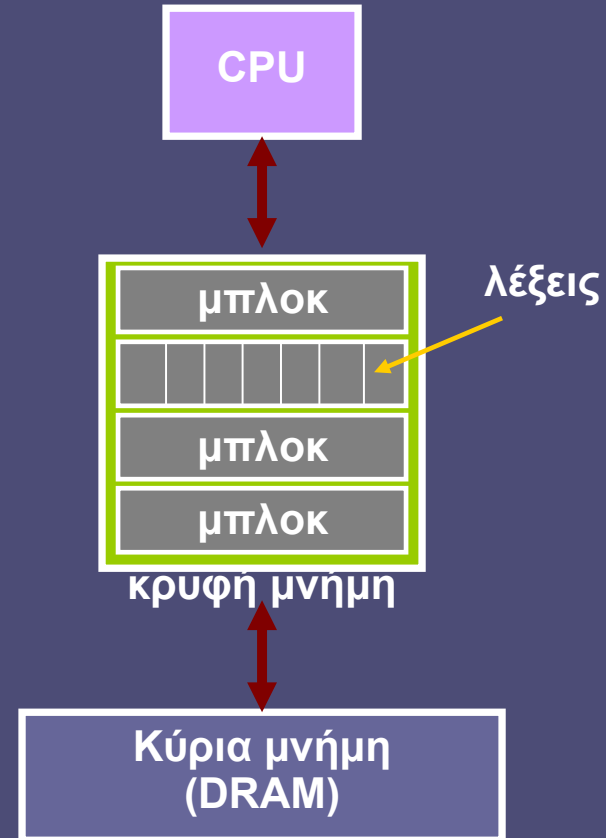
- Αναζήτηση δεδομένων
 - Ο επεξεργαστής ζητά **πάντοτε** τα δεδομένα/εντολές από το κοντινότερο σε αυτόν επίπεδο
 - Τα δεδομένα υπάρχουν στο επίπεδο αυτό: **hit**
 - Τα δεδομένα δεν βρίσκονται στο επίπεδο αυτό: **miss**
 - Η αίτηση προωθείται στο επόμενο (χαμηλότερο) επίπεδο
 - Όταν βρεθεί, το **μπλοκ** που περιέχει τα δεδομένα **αντιγράφεται** στο ανώτερο επίπεδο

Μπλοκ (γραμμές) κρυφής μνήμης

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη

- Για την εκμετάλλευση της **τοπικότητας**
- Όταν πρέπει να μεταφερθεί μια λέξη, μεταφέρεται **το μπλοκ που την περιέχει**
- Το μέγεθος του μπλοκ είναι καθοριστικό για την **απόδοση της ιεραρχίας μνήμης**
- Το σύστημα κύριας μνήμης έχει **βελτιστοποιηθεί αρχιτεκτονικά για μεταφορές μπλοκ**

Οι σύγχρονοι επεξεργαστές διαθέτουν κρυφές μνήμες με μέγεθος μπλοκ ίσο με 64 bytes



Τοποθέτηση ενός μπλοκ

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη

- Η κύρια μνήμη περιέχει πολύ περισσότερα «μπλοκ» από όσα χωρούν στην κρυφή μνήμη
 - Συνεπώς, στην ίδια θέση της κρυφής μνήμης πρέπει να τοποθετηθούν (κατά τη διάρκεια εκτέλεσης) περισσότερα από ένα μπλοκ
 - Σύγκρουση μπλοκ
- Πώς αποφασίζεται η θέση ενός μπλοκ στην κρυφή μνήμη;
 - Η απλή λύση: άμεση απεικόνιση (direct mapped caches)
 - Κάθε μπλοκ πηγαίνει σε μία μόνο θέση
(αριθμός μπλοκ) mod (θέσεις στην κρυφή μνήμη)
 - Υπολογίζεται πολύ εύκολα αν οι θέσεις είναι δύναμη του 2

Άμεση απεικόνιση θέσης μπλοκ

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη

Αριθμός μπλοκ = 0x22
 $0x226 = 001\ 00010\ 0110$
θέση μπλοκ = 2

$0x7E9 = 011\ 11110\ 1001$

$0x821 = 100\ 00010\ 0001$



κρυφή μνήμη

Κάθε **διεύθυνση** μπορεί να βρίσκεται σε μία μόνο θέση στην κρυφή μνήμη

Ποιο μπλοκ βρίσκεται σε κάθε θέση;

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη

0x226 = 001 00010 0110

tag (ετικέτα)

	μπλοκ 0
	μπλοκ 1
001	μπλοκ 2
	μπλοκ 3
	μπλοκ ..
	μπλοκ 30
	μπλοκ 31

κρυφή μνήμη

0x821 = 100 00010 0001

Σύγκριση με **ετικέτα (tag)** της επιλεγμένης θέσης στην κρυφή μνήμη

Ανάγνωση: Cache Hit

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη

Σε περίπτωση εύρεσης των δεδομένων στην κρυφή μνήμη, η ΚΜΕ μπορεί να τα λάβει ακόμα και σε 1 κύκλο

read 0x226 = 001 00010 0110

hit



Κύρια μνήμη
(DRAM)

•Αιτήσεις για ανάγνωση: εντολές και δεδομένα

Ανάγνωση: Cache Miss

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη

Miss penalty:

ο χρόνος για την μεταφορά του μπλοκ από κύρια μνήμη και επιστροφή δεδομένων στον επεξεργαστή

read 0x821 = 100 00010 0001

miss

read 0x820...0x82F

CPU

μπλοκ 0

μπλοκ 1

100

μπλοκ 3

μπλοκ ..

Κύρια μνήμη
(DRAM)

• Αιτήσεις για ανάγνωση: εντολές και δεδομένα

Εγγραφή στην κρυφή μνήμη

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη

Συνοχή δεδομένων:

Πώς επηρεάζουν οι κρυφές μνήμες τη σχεδίαση πολυεπεξεργαστικών συστημάτων;

- Μόνο για δεδομένα
- Write Hit – Ενημέρωση κρυφής μνήμης
 - Η νέα τιμή βρίσκεται **μόνο** στην κρυφή μνήμη
 - Η τιμή στην κύρια μνήμη (ή γενικότερα, στο χαμηλότερο επίπεδο) ενημερώνεται **όταν το μπλοκ εκτοπίζεται από την κρυφή μνήμη (victim)**
 - Απαιτείται επιπλέον λογική (hardware) για τον έλεγχο της **συνοχής** των δεδομένων
 - Όλοι οι πυρήνες πρέπει να βλέπουν τα ίδια δεδομένα
- Write Miss
 - Πρέπει το μπλοκ να έρθει (**ανάγνωση!**) πρώτα στην κρυφή μνήμη από την κύρια μνήμη

Τι δημιουργεί cache misses;

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη
- Απόδοση κρυφής μνήμης

- Η πρώτη φορά προσπέλασης ενός μπλοκ
 - Μπλοκ που δεν βρέθηκαν **ποτέ μέχρι τώρα** στην κρυφή μνήμη
- Λόγω χωρητικότητας της κρυφής μνήμης
 - Η κρυφή μνήμη **δεν χωράει** όλα τα μπλοκ (ταυτόχρονα)
 - Μπλοκ που τοποθετούνται στην **ίδια θέση** στην κρυφή μνήμη, συναγωνίζονται για τη θέση αυτή
 - ανάλογα με τη μέθοδο τοποθέτησης
 - ακόμα κι αν άλλο μέρος της κρυφής μνήμης είναι ελεύθερο...

Χαρακτηριστικά απόδοσης κρυφής μνήμης

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη
- Απόδοση κρυφής μνήμης

- **Hit Rate**

- Ποσοστό προσπελάσεων μνήμης, όπου τα δεδομένα βρίσκονται στην κρυφή μνήμη

- **Miss Rate**

- Ποσοστό προσπελάσεων μνήμης, όπου τα δεδομένα δεν βρίσκονται στην κρυφή μνήμη
 - (1-hit rate)

- **Hit Time**

- Ο χρόνος για την προσπέλαση δεδομένων σε hit

- **Miss Penalty**

- Ο χρόνος για την προσπέλαση, μεταφορά και τοποθέτηση των δεδομένων miss από την κύρια στην κρυφή μνήμη και στον επεξεργαστή

Το κόστος των cache misses

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη
- Απόδοση κρυφής μνήμης

- **Χαμένοι κύκλοι ρολογιού**
 - Σε αναμονή για προσπέλαση κύριας μνήμης

Κύκλοι Αναμονής =

*Προσπελάσεις μνήμης * Miss Rate * Miss Penalty*
- **Είναι απλουστευμένο μοντέλο γιατί:**
 - Διαφορετικό Miss Rate ανά κατηγορίες εντολών
 - Διαφορετικό Miss Rate για ανάγνωση-εγγραφή
 - Σύνθετη ανάλυση για **εκτέλεση εκτός σειράς**
 - Ο επεξεργαστής “κρύβει” την καθυστέρηση εκτελώντας κάτι άλλο: **πώς υπολογίζεται το miss penalty τότε;**
- **Βελτίωση της απόδοσης**
 - Μείωση του miss rate
 - Μείωση του miss penalty

Τεχνικές μείωσης miss rate

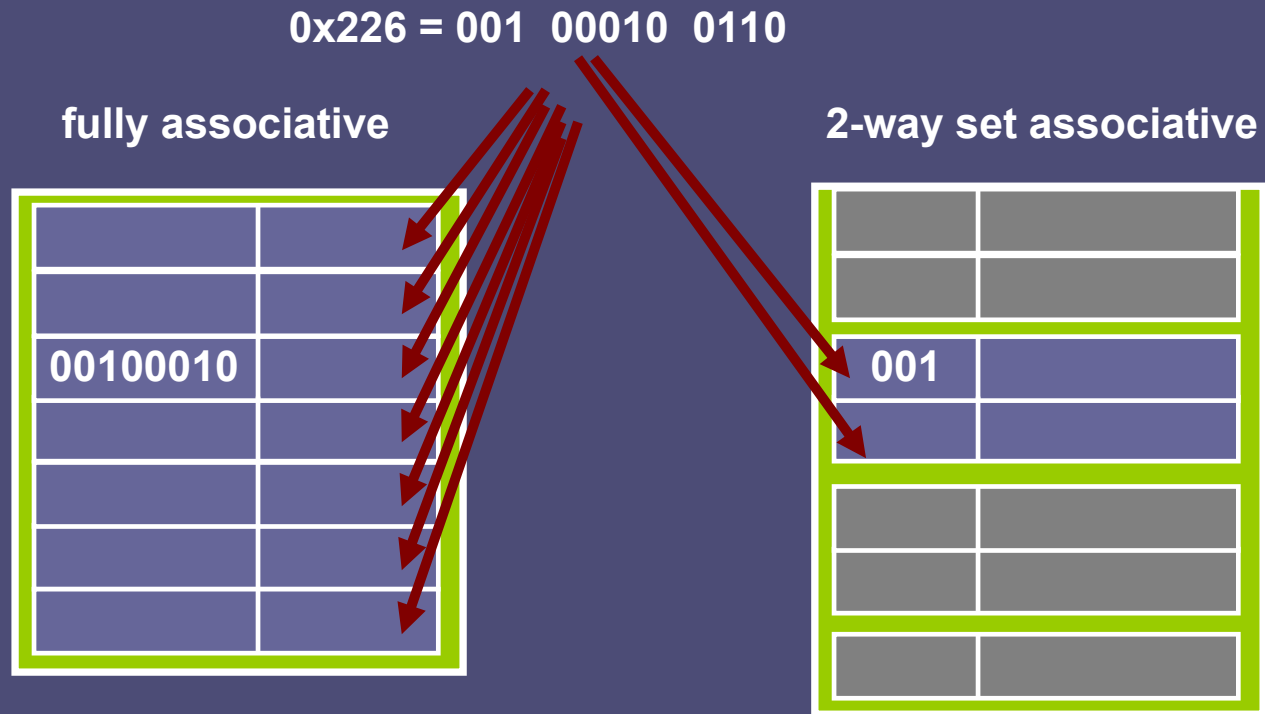
- Ιεραρχία Μνήμης
- Κρυφή Μνήμη
- Απόδοση κρυφής μνήμης

- Αντιμετώπιση αιτιών που προκαλούν misses
- **Αύξηση χωρητικότητας κρυφής μνήμης**
 - Αλλά: μια μεγάλη κρυφή μνήμη μπορεί να είναι πιο αργή (αύξηση hit time)
- **Αύξηση του μεγέθους του μπλοκ**
 - Προσπάθεια εκμετάλλευσης της τοπικότητας
 - Αλλά: αυξάνει το miss penalty
 - Πιθανόν να αυξάνει τελικά το miss rate, λόγω των λιγότερων μπλοκ στην κρυφή μνήμη
- **Ευέλικτες τεχνικές τοποθέτησης των μπλοκ**
 - Όστε να παραμένουν περισσότερο στην κρυφή μνήμη

Ευέλικτες τεχνικές τοποθέτησης μπλοκ

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη

Πιθανή η αύξηση του hit time λόγω πολύπλοκου κυκλώματος



→ Πιθανή τοποθέτηση
Παράλληλη αναζήτηση

• Αντικατάσταση: least recently used (LRU)

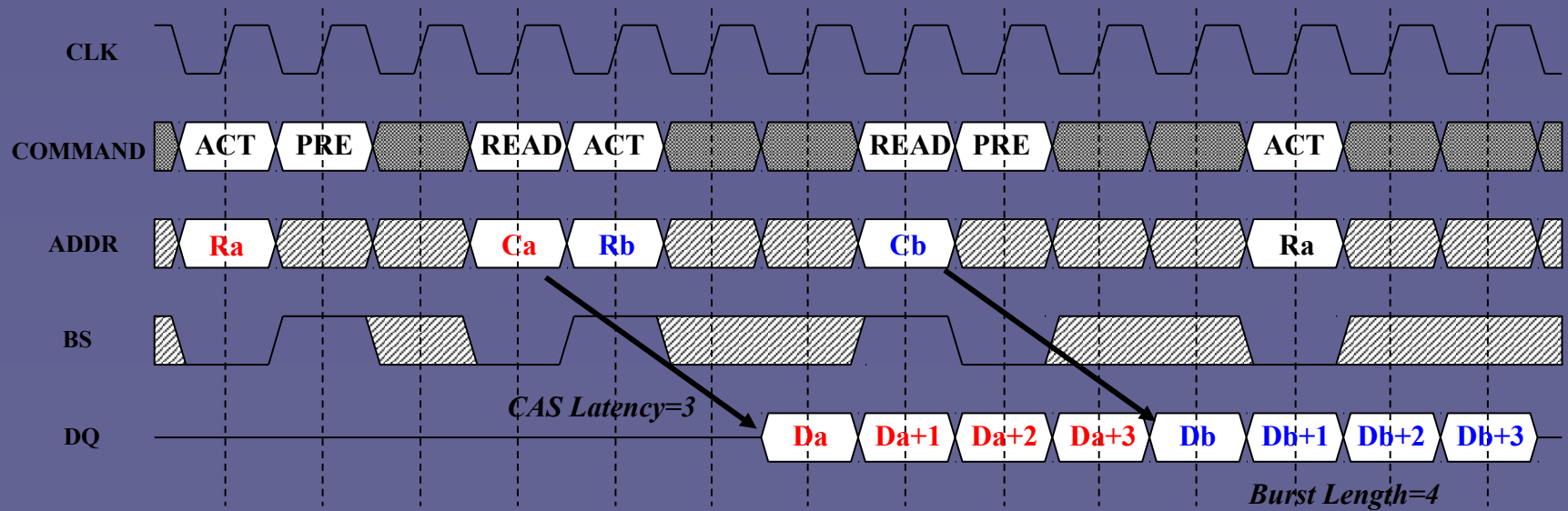
Τεχνικές μείωσης miss penalty

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη
- Απόδοση κρυφής μνήμης

Οι σύγχρονοι επεξεργαστές έχουν L1, L2 και L3 cache μέσα στο ίδιο το chip τους (ίσως και 4ο επίπεδο, ως cache «τελευταίας ευκαιρίας»)

- Μείωση των χρόνων μεταφοράς μπλοκ
- Βελτιστοποιήσεις στην επικοινωνία με την κύρια μνήμη
 - Έτσι ώστε ένα ολόκληρο μπλοκ να μεταφέρεται με τη μικρότερη δυνατή καθυστέρηση (bursts)
- Πολυεπίπεδες ιεραρχίες κρυφής μνήμης
 - Μείωση miss penalty πρώτου επιπέδου (L1)
 - L1: μικρότερο μέγεθος, μεγαλύτερη ταχύτητα
 - Μεγαλύτερο miss rate αλλά miss penalty μικρότερο
 - L2: μεγαλύτερο μέγεθος, μικρότερη ταχύτητα
 - Αργότερη αλλά δεν επηρεάζει hit time επεξεργαστή
 - L3: κοινή για ομάδες πυρήνων

Ανάγνωση από κύρια μνήμη



Πολυεπίπεδη οργάνωση κρυφής μνήμης

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη
- Απόδοση κρυφής μνήμης

Τύπος	Μέγεθος	Χρόνος προσπέλασης	Ρυθμός μεταφοράς
L1	έως 64KB	4ns	50GB/s
L2	έως 8MB	10ns	25GB/s
L3	έως 64MB	20ns	10GB/s

Οι σύγχρονοι επεξεργαστές έχουν ξεχωριστή κρυφή μνήμη L1 για εντολές και δεδομένα. Ποια τα πλεονεκτήματα-μειονεκτήματα;

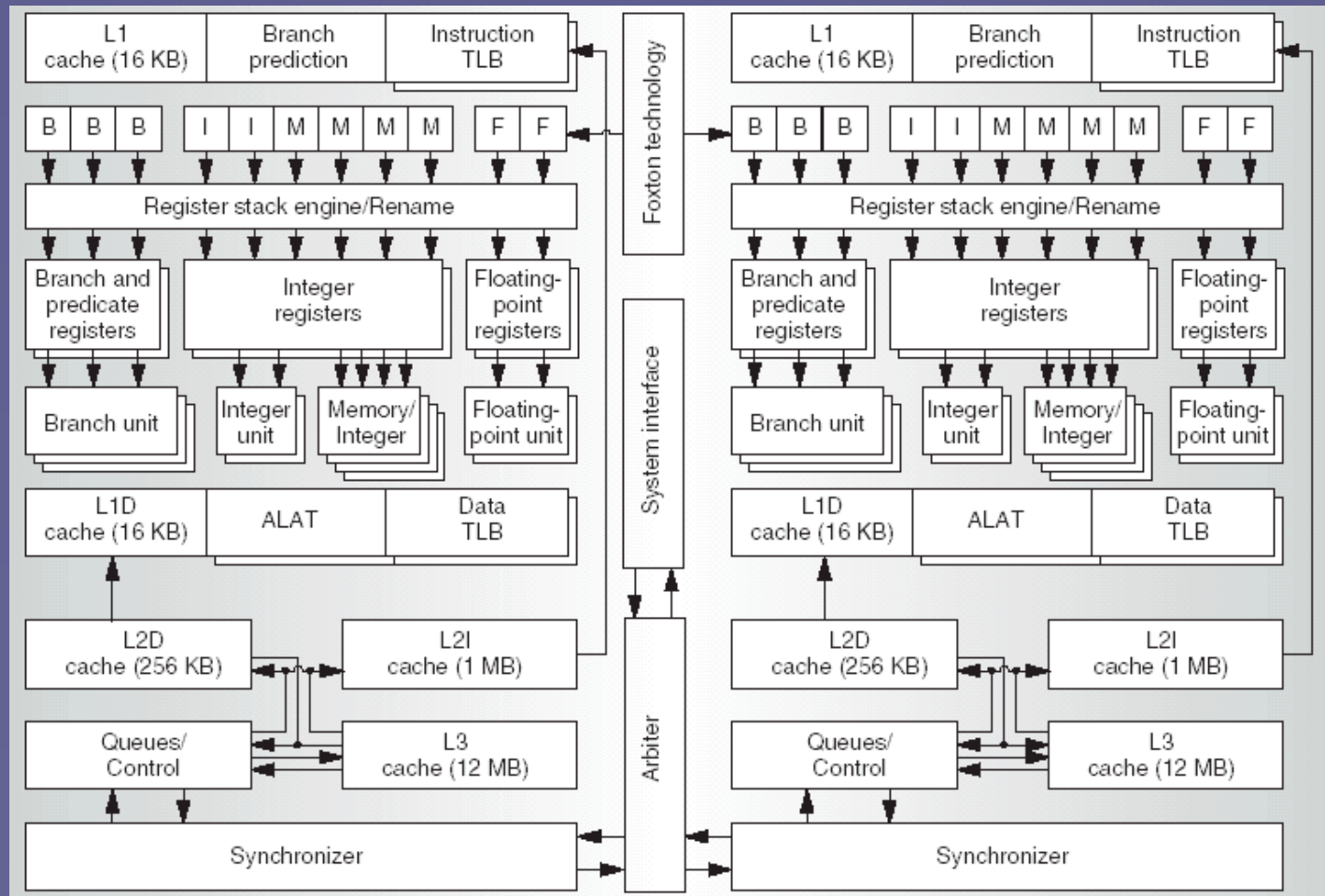
- Παράδειγμα: **Pentium4**
 - L1 cache: 4 κύκλοι ρολογιού (pipelined: 1)
 - L2 cache: 20 κύκλοι ρολογιού
 - Προσπέλαση στη μνήμη: >100 κύκλοι ρολογιού

Intel “Montecito”: Επίπεδα κρυφής μνήμης

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη
- Απόδοση κρυφής μνήμης

Intel Montecito:

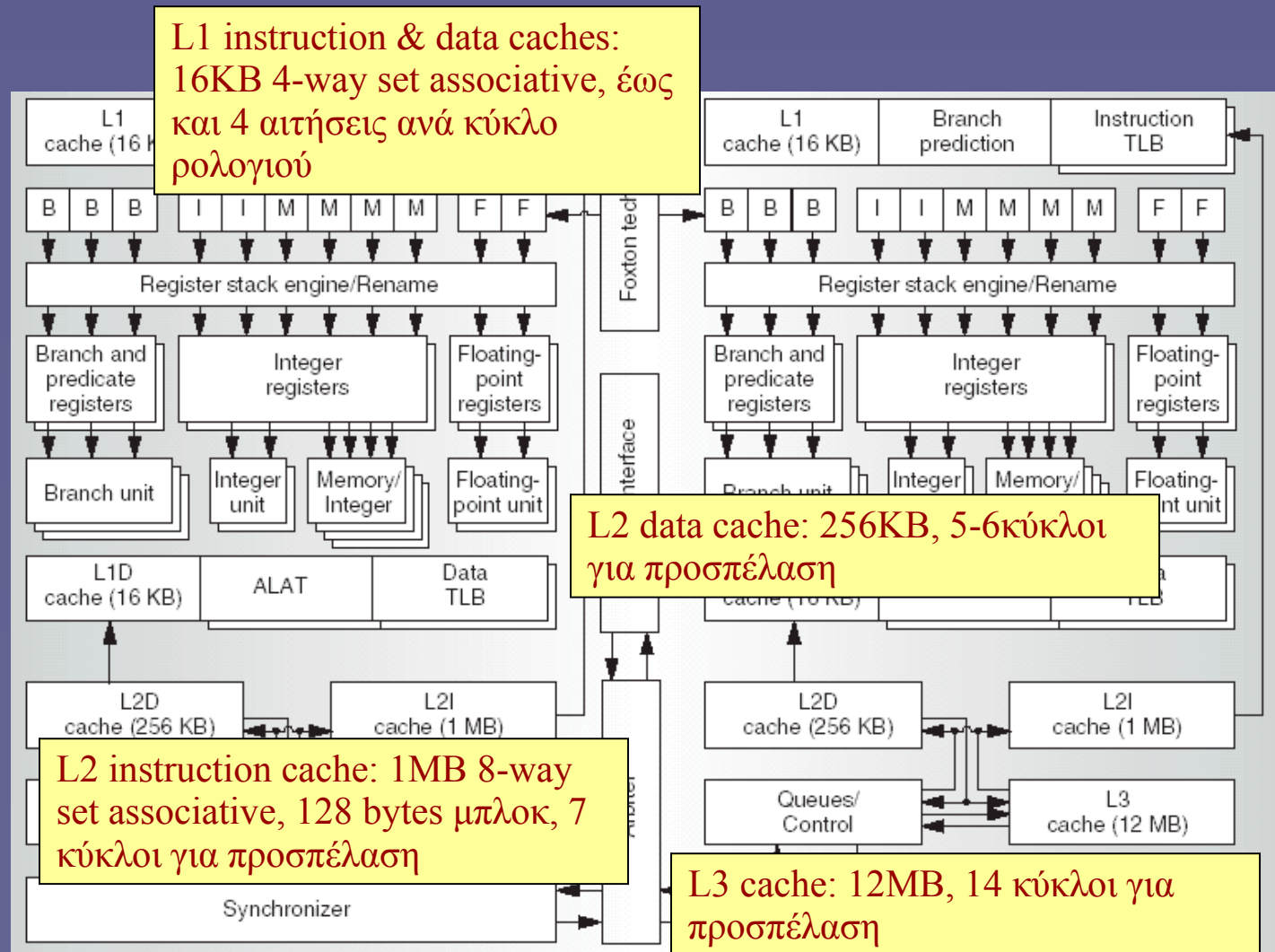
1,72 δις
τρανζίστορ
2 επεξεργαστές
Itanium2
1.8GHz @ 100W



Intel “Montecito”: Επίπεδα κρυφής μνήμης

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη
- Απόδοση κρυφής μνήμης

Intel Montecito:
συνολικά 27MB
κρυφή μνήμη μέσα
στο chip



Βελτιστοποίηση απόδοσης κρυφής μνήμης

- Ιεραρχία Μνήμης
 - Κρυφή Μνήμη
 - Απόδοση κρυφής μνήμης
- Αρχιτεκτονικές βελτιώσεις
 - Pipelining
 - Non-blocking – εξυπηρέτηση πολλαπλών αιτήσεων
 - Πολλαπλά επίπεδα κρυφής μνήμης στο chip του επεξεργαστή
 - Ο ρόλος του λογισμικού (μεταγλωττιστές)
 - Αναδιοργάνωση προγραμμάτων για αύξηση της τοπικότητας (κυρίως στους βρόχους επανάληψης)
 - Prefetching: μετακίνηση δεδομένων στην κρυφή μνήμη πριν αυτά χρειαστούν στον επεξεργαστή

Η απόδοση της κρυφής μνήμης συνοπτικά

- Ιεραρχία Μνήμης
- Κρυφή Μνήμη
- Απόδοση κρυφής μνήμης

- Καθοριστική για τα σύγχρονα υπολογιστικά συστήματα
- Μείωση του miss rate ή του miss penalty
 - Όμως: η συμπεριφορά της ιεραρχίας μνήμης επηρεάζεται από πολλούς παράγοντες
- Η πραγματική συμπεριφορά
 - Είναι σύνθετη – απαιτούνται εξομοιώσεις πριν τη σχεδίαση νέων συστημάτων
 - Είναι διαφορετική ανά εφαρμογή – δεν υπάρχει ένα μόνο αντιπροσωπευτικό πρόγραμμα
 - Είναι διαφορετική ανά υπολογιστικό σύστημα – desktop, server ή embedded