ΗΜΥ 312
ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Εαρινό Εξάμηνο 2006

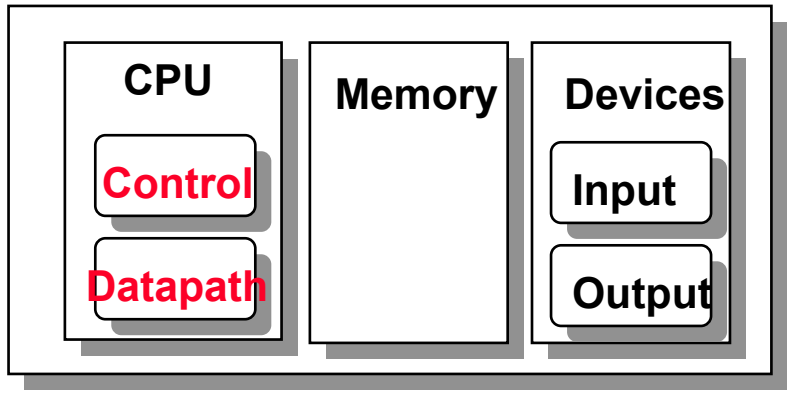
ΔΙΑΛΕΞΗ 03: ΜΝΗΜΗ

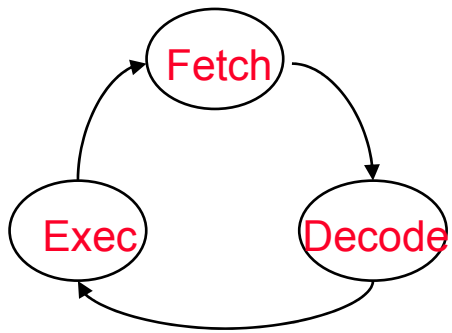
ΘΕΟΧΑΡΗΣ ΘΕΟΧΑΡΙΔΗΣ (charisth@ucy.ac.cy)

[Προσαρμογή από *Computer Architecture*,
Patterson & Hennessy, © 2005, UCB]

(vonNeumann) Οργάνωση Επεξεργαστή

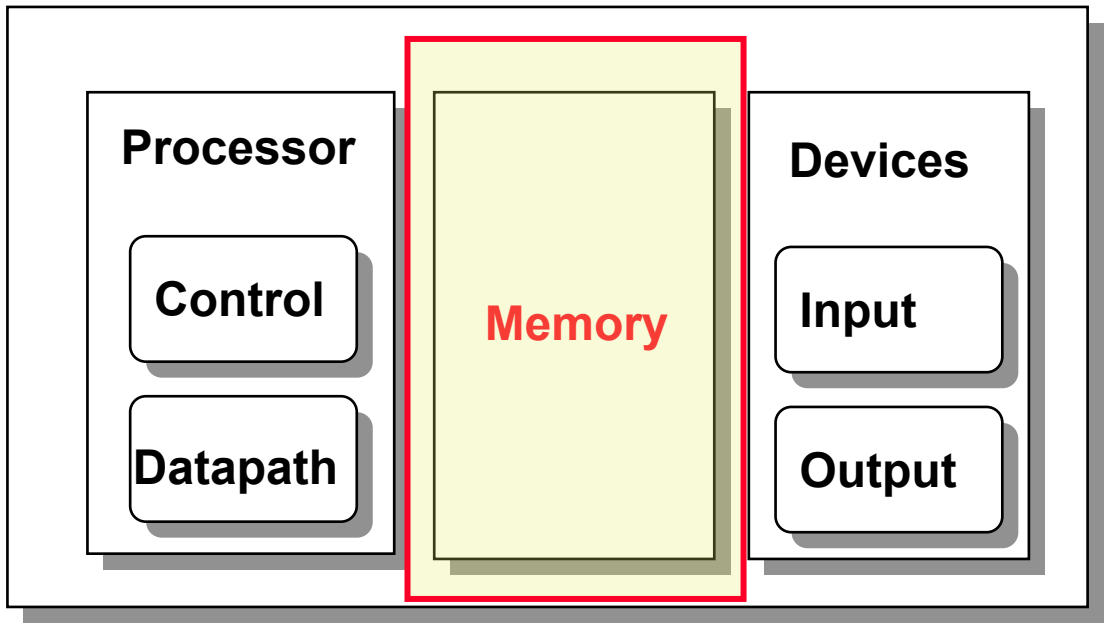
- Control (‘Ελεγχος) χρειάζεται
  1. Να λάβει εντολές απο την Μνήμη
  2. Να εκδώσει τα σήματα που χρειάζονται για μεταφορά δεδομένων μεταξύ συστημάτων του επεξεργαστή και οδηγιών ελέγχου
  3. Να διατηρά την σωστή σειρά οδηγιών ελέγχου.



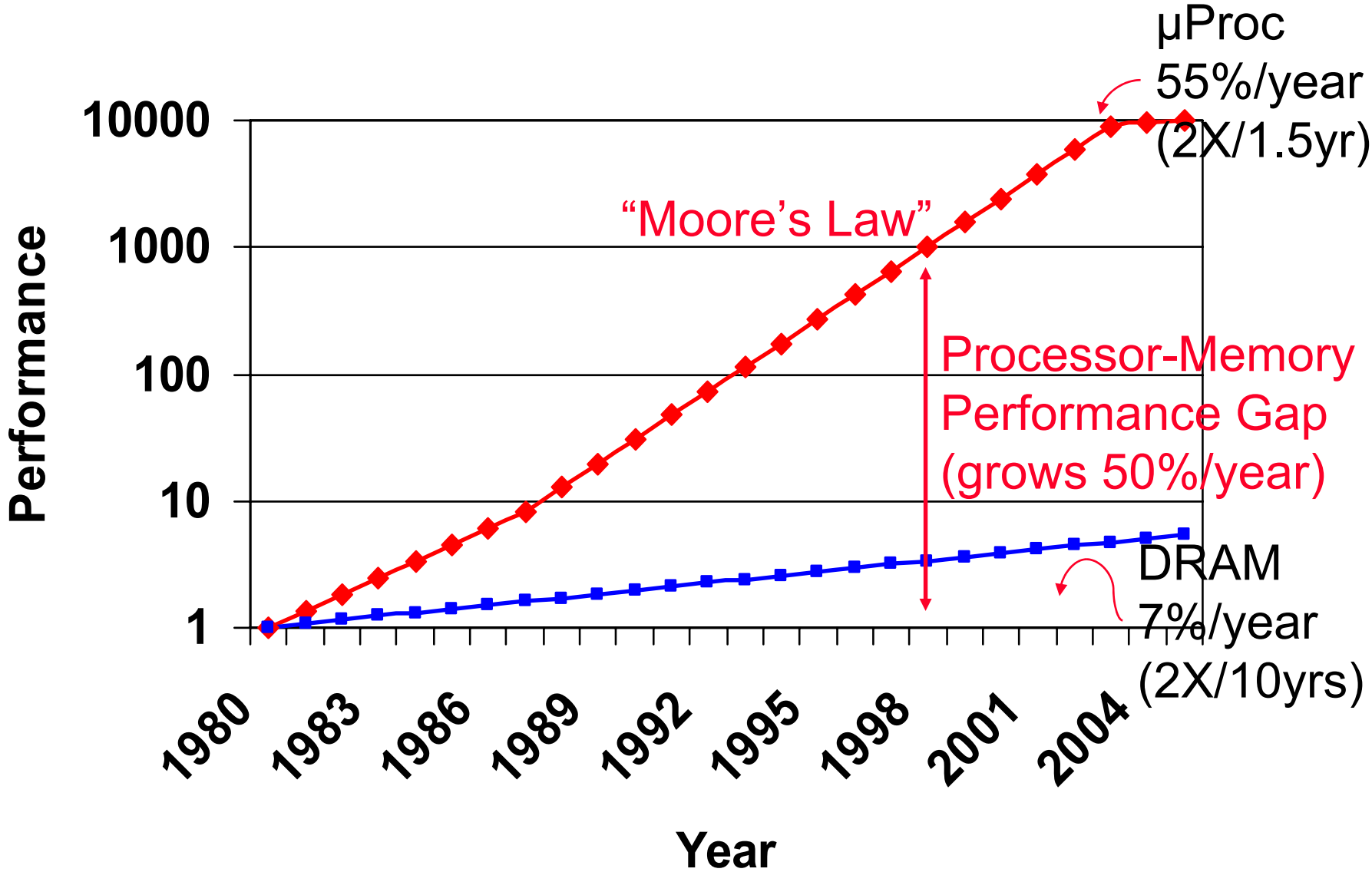


- Datapath (Δίοδος Δεδομένων) χρειάζεται
  - Τα συστήματα – τα εκτελεστικά και αριθμητικά συστήματα καθώς και την απαιτούμενη μνήμη για να εκτέλά εντολές
  - Το σύστημα επικοινωνίας ώστε τα δεδομένα να μπορούν να ταξιδεύουν μεταξύ εκτελεστικών συστημάτων και μνήμης

ΕΠ: ΣΥΣΤΗΜΑ ΚΑΙ ΚΥΡΙΑ ΜΕΡΗ Η/Υ

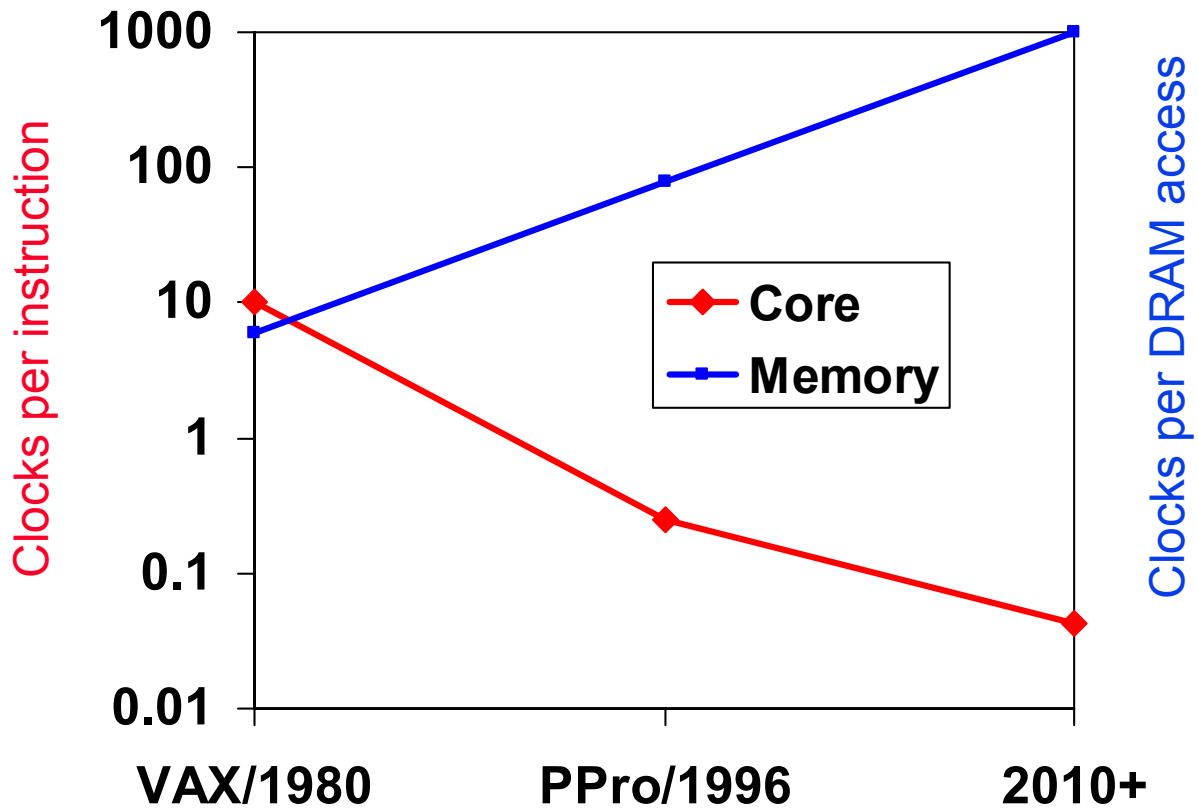


ΔΙΑΦΟΡΑ ΣΤΗΝ ΜΝΗΜΗ ΚΑΙ ΤΟΝ ΕΠΕΞ/ΣΤΗ



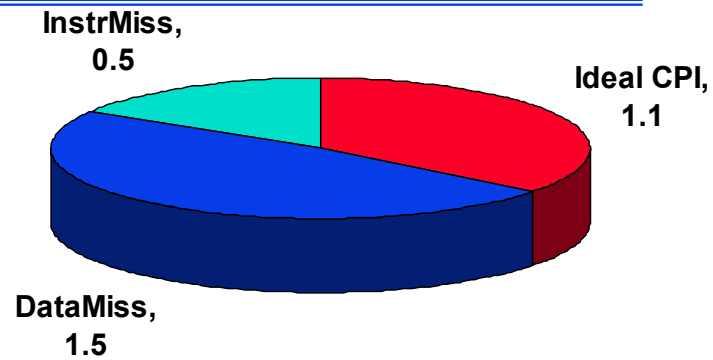
ΤΟ ΦΡΑΓΜΑ ΤΗΣ ΜΝΗΜΗΣ

- Υπολογιστικά Μέρη (Logic) vs DRAM (Μνήμη) – Η διαφορά στην ταχύτητα μεγαλώνει.



ΑΠΟΔΩΣΗ ΜΝΗΜΗΣ ΚΑΙ ΕΠΙΡΡΟΗ ΣΤΗΝ ΣΥΝΟΛΙΚΗ ΑΠΟΔΩΣΗ

- Ας πούμε ότι ένας επεξεργαστής δυλεύει στα ιδεώδες
  - CPI = 1.1
  - 50% arith/logic, 30% ld/st, 20% control



και ότι 10% των δεδομένων μνήμης που δεν βρίσκονται στην μνήμη (miss) έχουν 50 κύκλους ποινή (miss penalty)

- CPI = ιδεώδες CPI + μέσος όρος stalls για κάθε εντολή
  = 1.1(cycle) + (0.30 (datamemops/instr)
  x 0.10 (miss/datamemop) x 50 (cycle/miss))
  = 1.1 cycle + 1.5 cycle = 2.6

άρα 58% της ώρας λειτουργίας του επεξεργαστή περνιέται στο να περιμένει την μνήμη!!!

- 1% instruction miss rate θα πρόσθετε *ακόμα* 0.5 στο CPI!

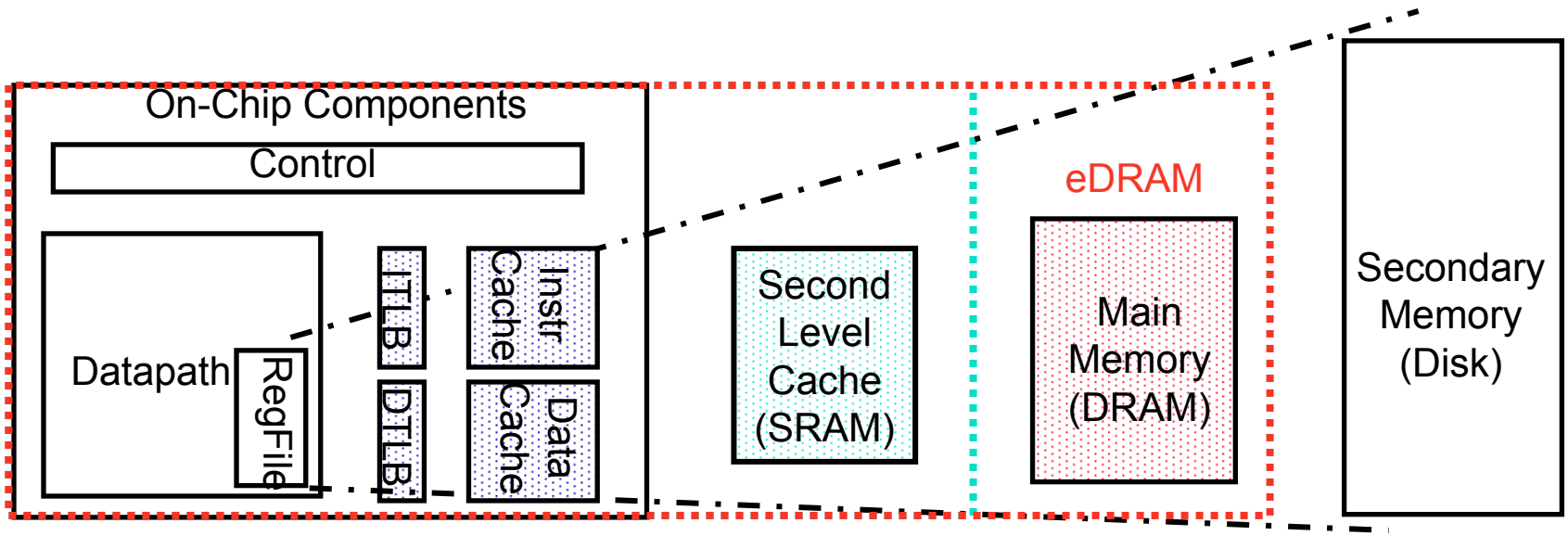
Ο ΣΚΟΠΟΣ ΤΗΣ ΙΕΡΑΡΧΙΑΣ ΤΗΣ ΜΝΗΜΗΣ

- ΓΕΓΟΝΟΣ: Οι μνήμες μεγάλου μεγέθους είναι αργές ενώ οι μνήμες μικρού μεγέθους είναι γρήγορες.

- Πως μπορούμε να δημιουργήσουμε μια μνήμη που θα φαίνεται οτι ειναι ταυτόχρονα μεγάλη, φθηνή και (τις περισσότερες φορές) γρήγορη?
- Με ιεραρχία
  - Με παραλληλισμό

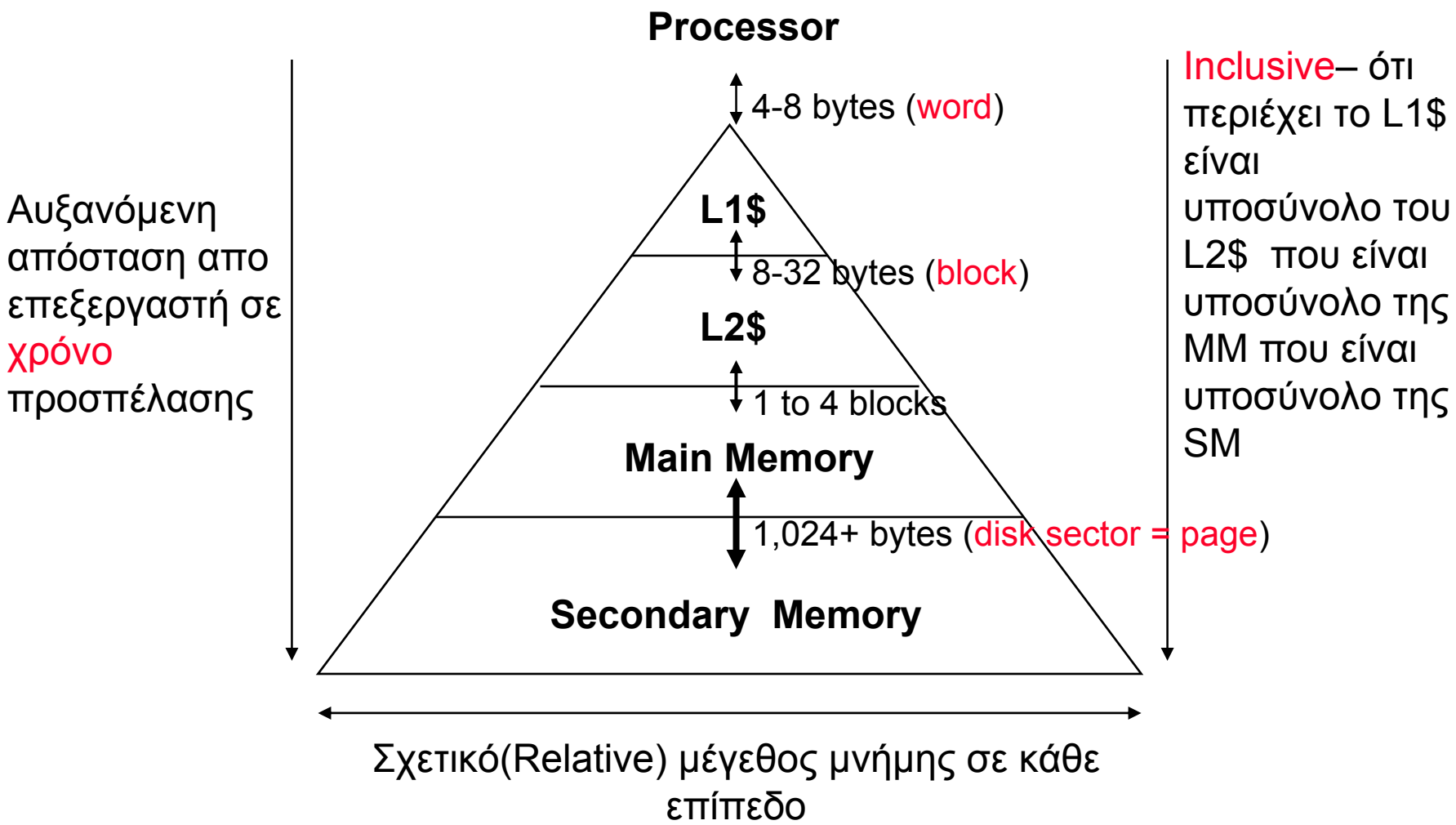
ΤΥΠΙΚΗ ΙΕΡΑΡΧΙΑ ΜΝΗΜΗΣ Η/Υ

- Χρησημοποίηση πλεονεκτίματος τοπικότητας(locality)
  - Να παρουσιάσουμε στον χρήστη όση μνήμη μπορούμε στην πιο φθηνή τεχνολογία
  - Στην ταχύτητα που προσφέρει η πιο γρήγορη τεχνολογία

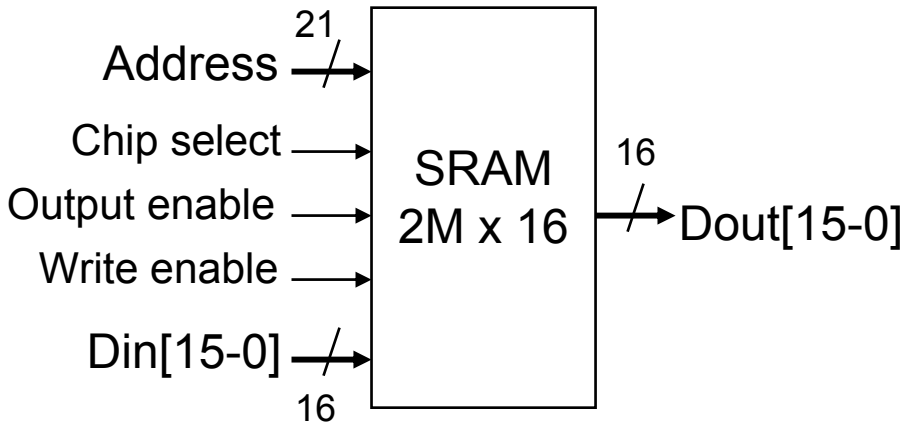


Speed (%cycles):	½'s	1's	10's	100's	1,000's
Size (bytes):	100's	K's	10K's	M's	G's to T's
Cost:	highest				lowest

ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΙΕΡΑΡΧΙΑΣ ΜΝΗΜΗΣ



ΤΕΧΝΟΛΟΓΙΕΣ ΙΕΡΑΡΧΙΑΣ ΜΝΗΜΗΣ



- Τα Caches χρησημοποιούν *SRAM* για ταχύτητα και συμβατότητα τεχνολογίας
  - Χαμήλή πυκνότητα (6 transistor cells), ψηλή ενέργεια, ακριβή, γρήγορη
  - Στατική: τα περιεχόμενα παραμένουν μέχρις ότου να διακοπεί το ρεύμα.

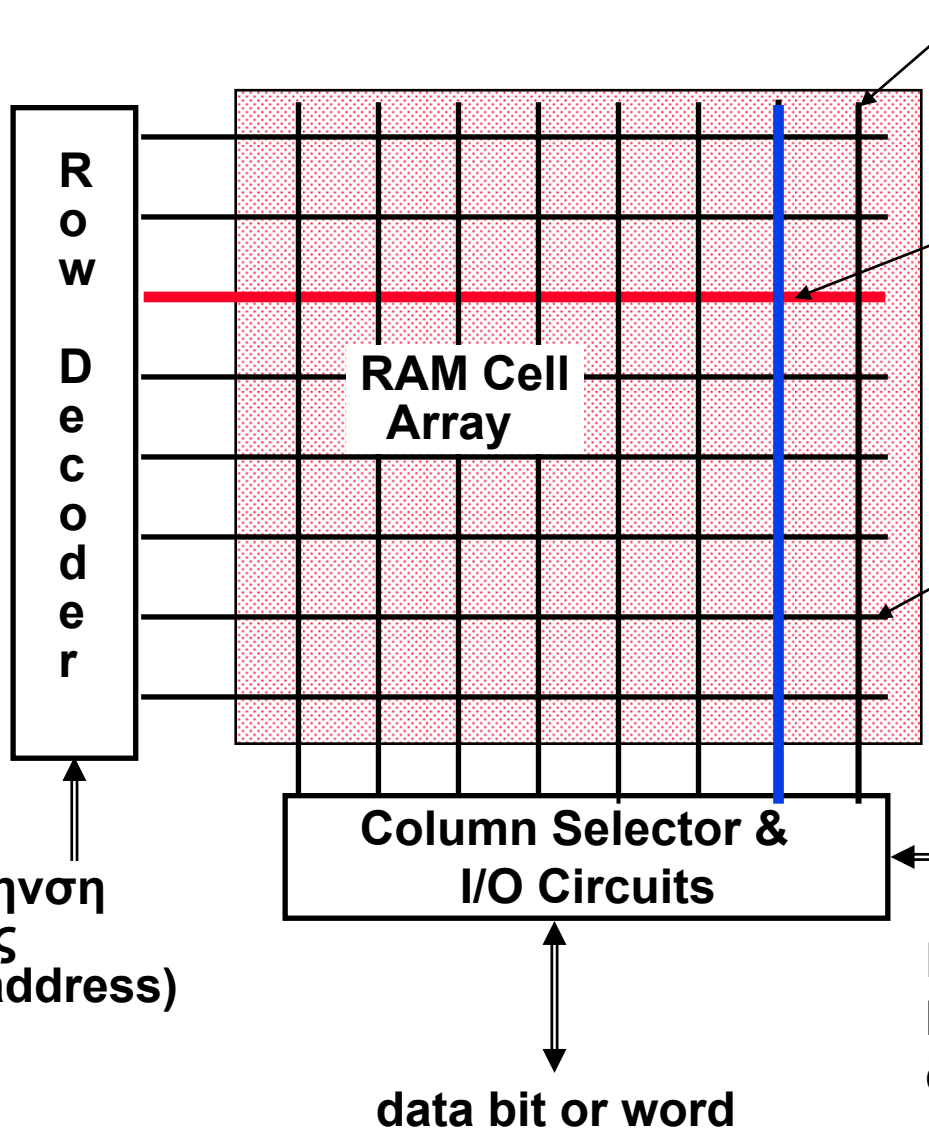
ΤΕΧΝΟΛΟΓΙΑ ΜΝΗΜΗΣ (Συν)

- Η κυρίως μνήμη χρησημοποιά *DRAM* λόγω μεγέθους (πυκνότητα)
  - Ψηλή πυκνότητα (1 transistor cells), χαμήλή ενέργεια, φθηνή, αργή
  - Δυναμική (Dynamic): Χρειάζεται ανανέωση (“refresh”) τακτικά (~ κάθε 8 ms) → 1% εως 2% από τους ενεργείς κύκλους του DRAM
  - Οι διευθήνσεις μοιράζονται σε 2 “μέρη” (σειρά/στήλη) (row and column)
    - *RAS* ή *Row Access Strobe* ενεργοποιούν τον αποκωδικοποιητή σειράς (row decoder)
    - *CAS* ή *Column Access Strobe* ενεργοποιούν τον αποκωδικοποιητή στήλης (column selector)

ΜΕΤΡΙΚΑ ΑΠΟΔΩΣΗΣ ΜΝΗΜΗΣ

- (Χρόνος) Latency: Ο χρόνος για να διαβαστεί μια *λέξη (word).*
  - *Χρόνος προσπέλασης (Access time)*: ο χρόνος μεταξύ της ώρας ζήτησης και της ώρας που τα δεδομένα είναι διαθέσιμα.
  - *Χρόνος σε κύκλους (Cycle time)*: ο χρόνος μεταξύ ζητήσεων (requests)
  - Συνήθως, χρόνος σε κύκλους > *Χρόνος προσπέλασης*
  - Συνηθησμένοι χρόνοι προσπέλασης SRAMs στο 2004 ήταν 2 με 4 ns για τις γρήγορες και μικρές μνήμες και 8 με 20ns συνηθησμένες πιο μεγάλες.
- Εύρος ζώνης (Bandwidth): Πόσα δεδομένα μπορεί η μνήμη να προσφέρει στον επεξεργαστή σε ένα κύκλο χρόνου
  - το μεγεθος σε μπιτ του data channel * η ροή από την μνήμη (rate)
- *Φυσικό Μέγεθος*: DRAM to SRAM - *4 to 8*
- *Κόστος/Χρόνος σε κύκλους*: SRAM σε DRAM - *8 σε 16*

ΚΛΑΣΣΙΚΗ ΟΡΓΑΝΩΣΗ RAM (~Τετράγωνη)



Κάθε διασταύρωση αντιπροσωπεύει 6-T SRAM cell η ένα 1-T DRAM cell

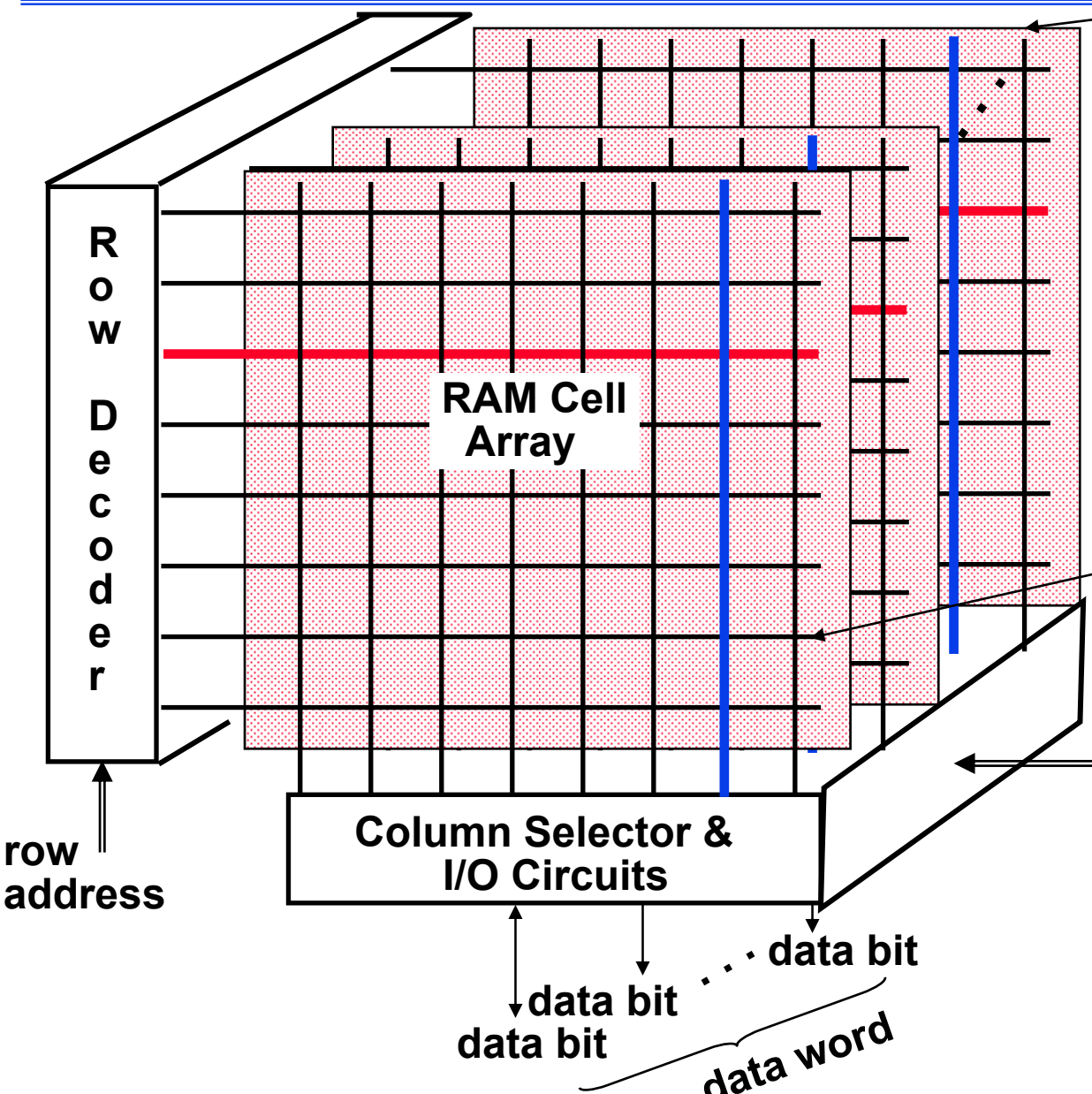
Λέξη (word) (row) line

Διεύθυνση σειράς (row address)

Διεύθυνση στήλης (column address)

Μια σειρά μνήμης κρατάει ενα block δεδομένων, ώστε η διεύθυνση στήλης να διαλέξει το ζητούμενο μπιτ (bit) ή την ζητούμενη λέξη (word) από το διαλεγμένο block.

ΚΛΑΣΣΙΚΗ ΟΡΓΑΝΩΣΗ DRAM (~Τετράγωνα Επίπεδα)

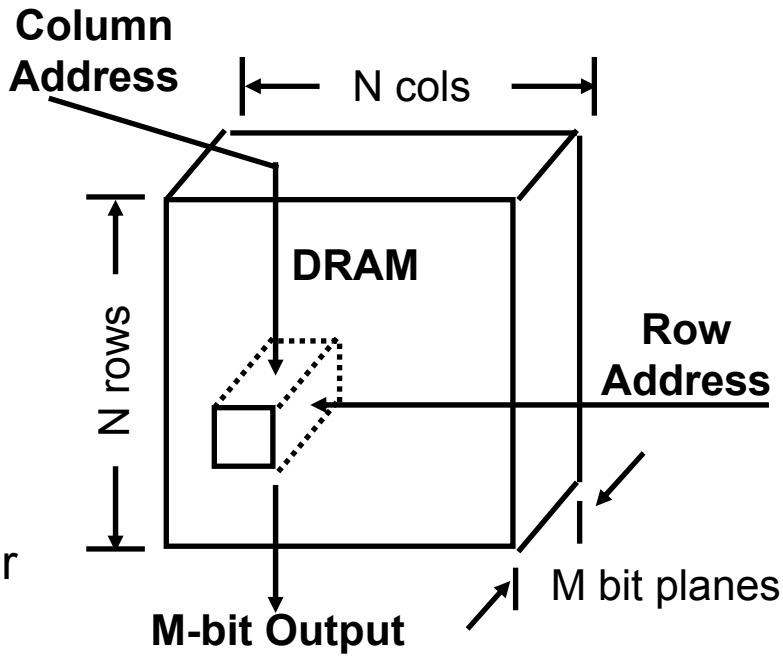


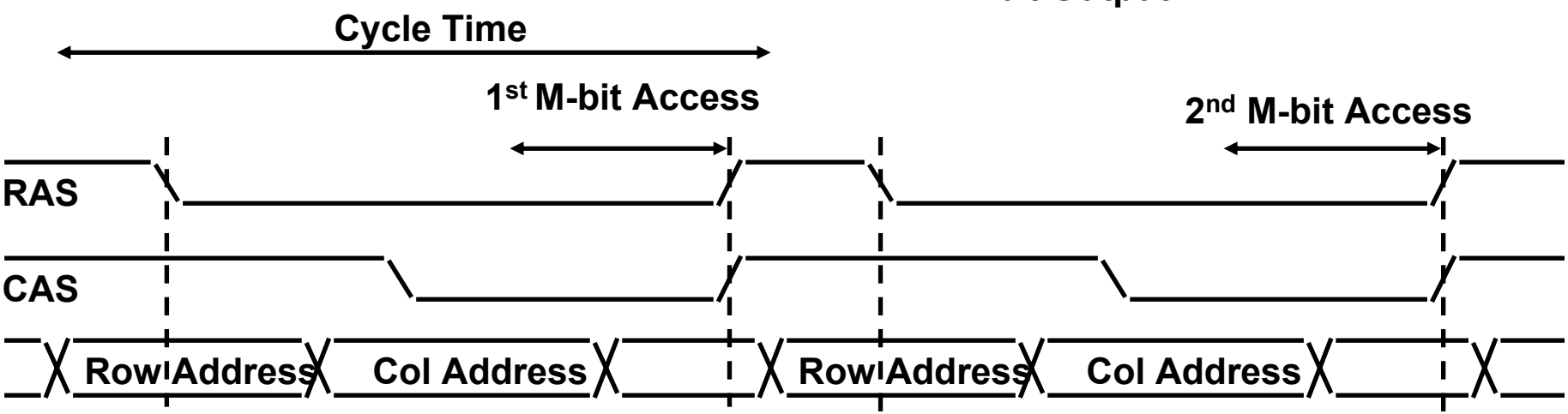
Κάθε διασταύρωση αντιπροσωπεύει ένα 1-T DRAM cell

Η διεύθηνση στήλης διαλέγει το bit από την σειρά σε κάθε επίπεδο (plane)

ΚΛΑΣΣΙΚΗ ΛΕΙΤΟΥΡΓΙΑ DRAM

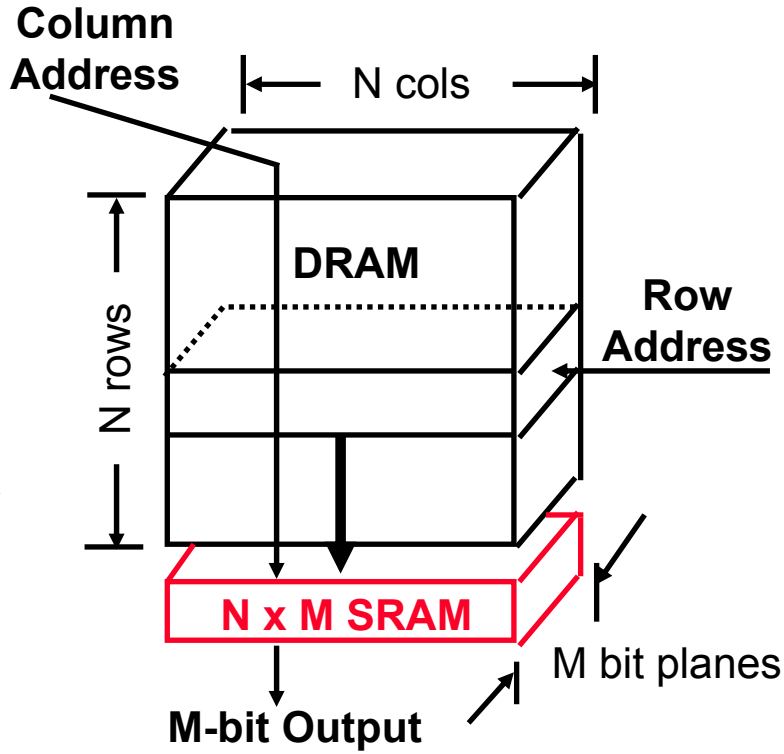
- Οργάνωση DRAM:
  - N σειρές x N στείλες x M-bit
  - Ανάγνωση ή Γραφή M-bit κάθε φορά
  - Κάθε M-bit προσπέλαση απαιτεί ένα RAS / CAS κύκλο
    - RAS – Row Address Selector
    - CAS – Column Address Selector

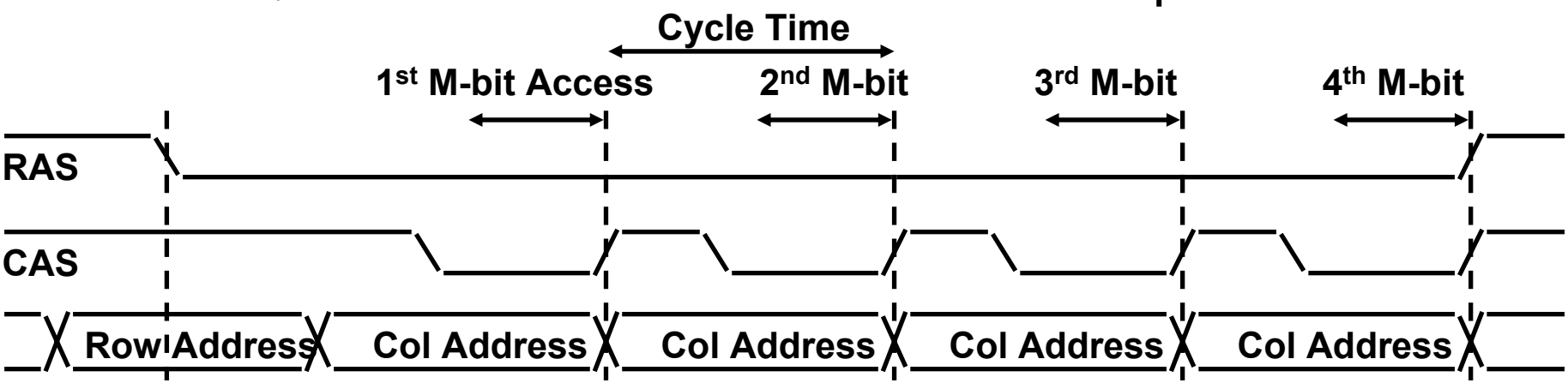




Page Mode DRAM Λειτουργία

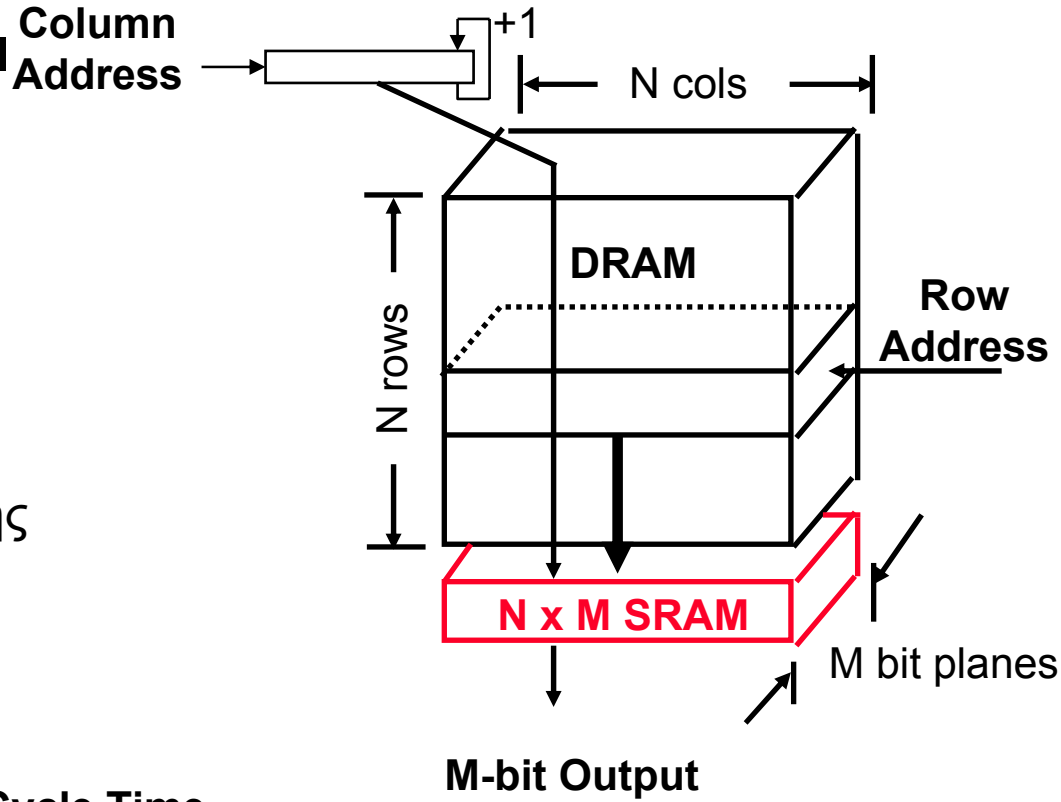
- Page Mode DRAM
  - N x M SRAM (γλύτωνουμε μια σειρά)
- Όταν διαβαστεί μια σειρά στο SRAM "register"
  - Μόνο το CAS χρειάζεται για να προσπελαστούν M-bit λέξεις στην ίδια σειρά.
  - Το RAS παραμένει ενεργό καθώς αλλάζει το CAS

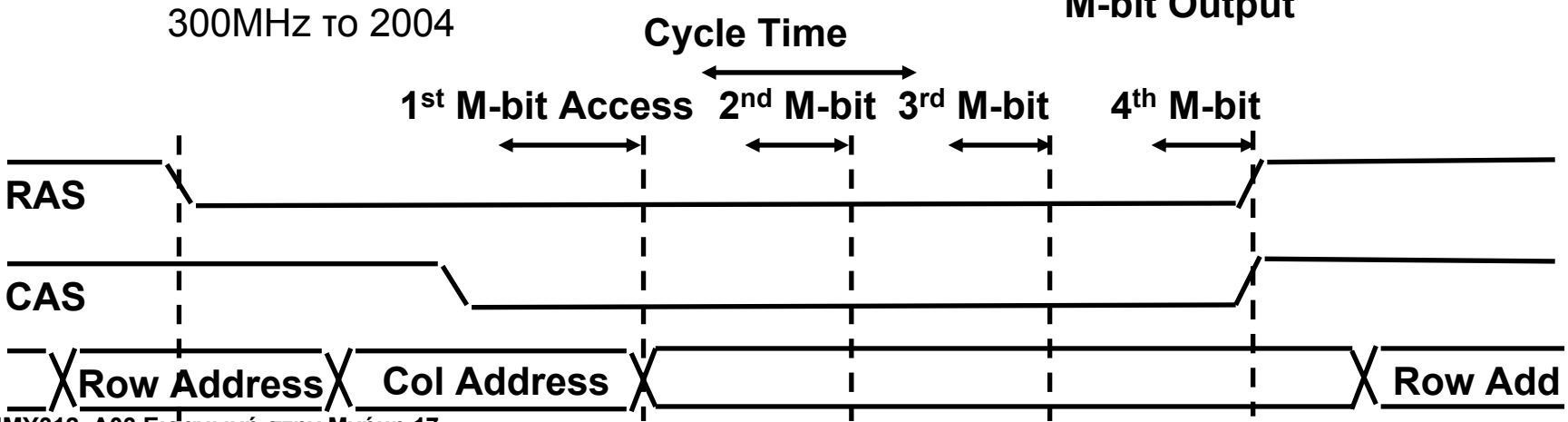




ΣΥΧΡΟΝΙΣΜΕΝΗ DRAM (SDRAM) ΛΕΙΤΟΥΡΓΙΑ

- Όταν μια σειρά διαβάζεται στο SRAM register
  - Βάζει το CAS σαν την αρχική “burst” διεύθηνση μαζί με το μήκος του burst (“Μπλόκ”)
  - Μεταφέρει ένα burst δεδομένων από μια σειρά συνεχομένων διευθήνσεων της σειράς αυτής.
    - Η Ωρολογιακή συχνότητα καθορίζει τον ρυθμό μεταφοράς δεδομένων – 300MHz το 2004





ΑΛΛΕΣ ΑΡΧΙΤΕΚΤΟΝΙΚΕΣ DRAM

- Double Data Rate SDRAMs – DDR-SDRAMs (και DDR-SRAMs)
  - Double data rate (διπλή ροή δεδομένων) επειδή μεταφέρονται τα δεδομένα όταν αναιβαίνει και όταν κατεβαίνει η συχνότητα (rising and falling edge of the clock).
  - Είναι η πιο δημοφιλής τεχνολογία μνήμης SDRAM

- DDR2-SDRAMs

http://www.corsairmemory.com/corsair/products/tech/memory_basics/153707/main.swf

DRAM Μνήμη Latency & Bandwidth Ιστορικές Στιγμές

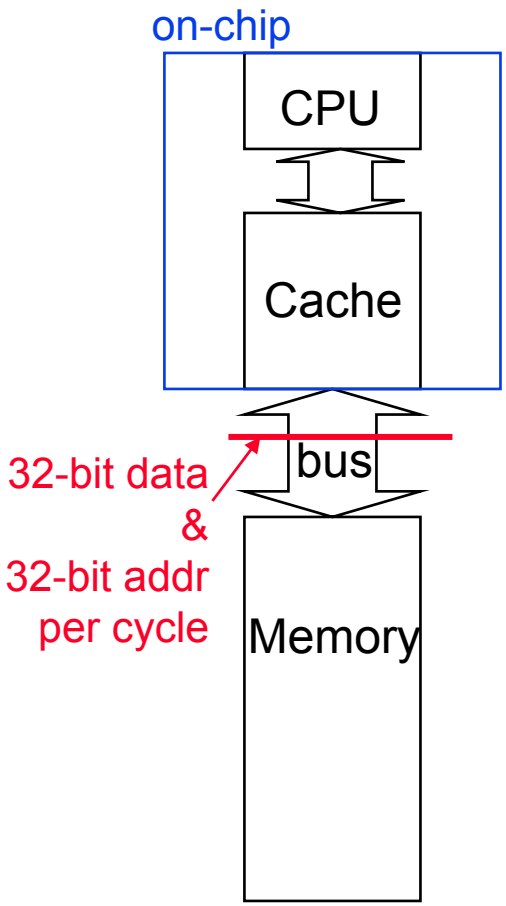
	DRAM	Page DRAM	FastPage DRAM	FastPage DRAM	Synch DRAM	DDR SDRAM
Module Width	16b	16b	32b	64b	64b	64b
Year	1980	1983	1986	1993	1997	2000
Mb/chip	0.06	0.25	1	16	64	256
Die size (mm^2)	35	45	70	130	170	204
Pins/chip	16	16	18	20	54	66
BWidth (MB/s)	13	40	160	267	640	1600
Latency (nsec)	225	170	125	75	62	52

Patterson, CACM Vol 47, #10, 2004

- Στην περίοδο που το εύρος μεταξύ μνήμης και επεξεργαστή (bandwidth) διπλασιαζεται, ο χρόνος προσπέλασης της μνήμης (latency) βελτιώνεται μόνο 1.2 με 1.4 φορές
- Για τόσο ψηλό εύρος, η εσωτερική DRAM πρέπει να οργανωθεί διακλαδωτικά (interleaved) σε memory banks

ΣΥΣΤΗΜΑΤΑ ΜΝΗΜΗΣ ΠΟΥ ΥΠΟΣΤΗΡΙΖΟΥΝ Caches

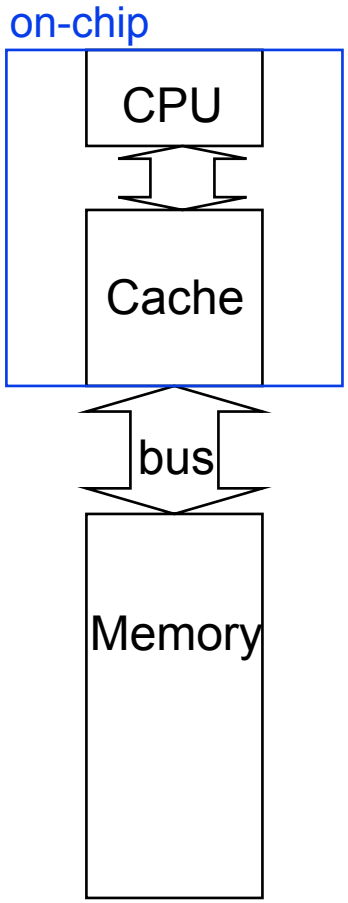
- Η Διασύνδεση (interconnect) μεταξύ του chip του επεξεργαστή και της μνήμης επιρεάζει σε πάρα πολύ μεγάλο βαθμό την απόδοση.



One word wide organization (one word wide bus and one word wide memory)

- Υποθέτουμε οτι,
  1. 1 clock cycle για τη διεύθηνση
  2. 25 clock cycles για DRAM cycle time, 8 clock cycles access time
  3. 1 clock cycle για επιστροφή μιας λέξης δεδομένων.
- Εύρος διαύλου μνήμης με Cache
  - Ο αριθμός των bytes που μεταφέρονται μεταξή μνήμης και cache/CPU κάθε clock cycle

ΟΡΓΑΝΩΣΗ ΜΝΗΜΗΣ ΜΙΑΣ ΛΕΞΗΣ (ONE WORD MEMORY)

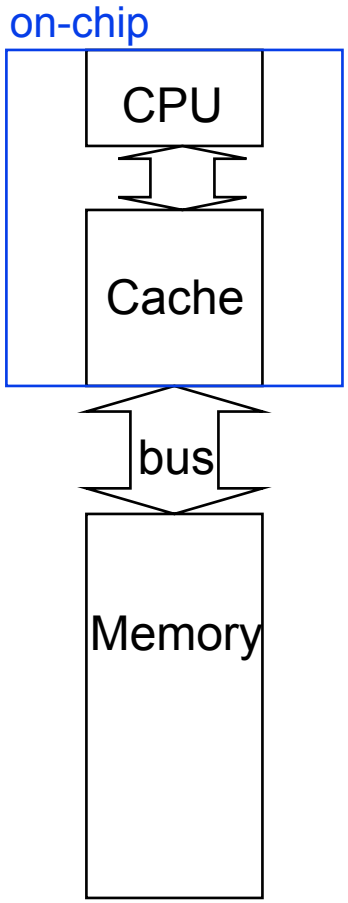


- Όταν το block size είναι μια *λέξη*, τότε όταν το Cache έχει miss,η διασωλήνωση πρέπει να σταματήσει (stall) για τόσους κύκλους όσος και ο αριθμός κύκλων που χρειάζεται για να έρθει η λέξη από την κυρίως μνήμη.

1	cycle to send address
25	cycles to read DRAM
1	cycle to return data
27	total clock cycles miss penalty

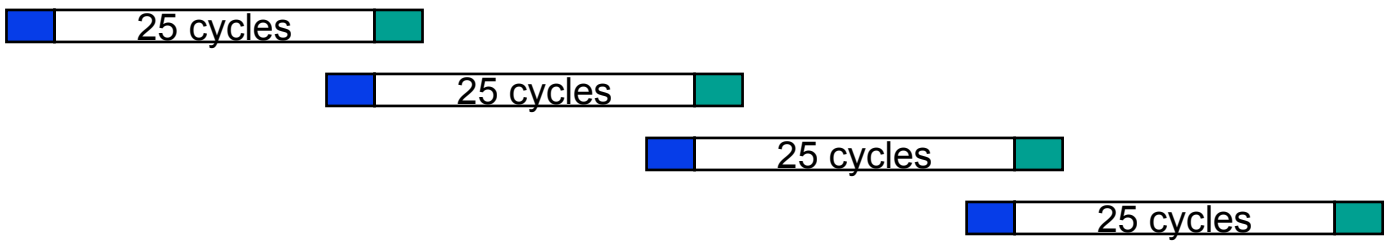
- Ο αριθμός των bytes (bandwidth) που μεταφέρονται κάθε κύκλο για κάθε miss είναι 4/27 = 0.148 bytes per clock

ΟΡΓΑΝΩΣΗ ΜΝΗΜΗΣ ΜΙΑΣ ΛΕΞΗΣ (ONE WORD MEMORY) (Συν)



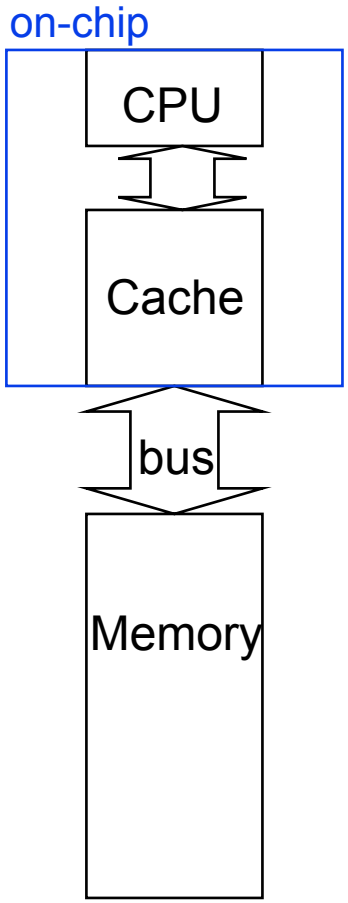
- Τι γίνεται όταν το block size είναι τέσσερις λέξεις?

	1	cycle to send 1st address
4 x 25 =	100	cycles to read DRAM
	1	cycles to return last data word
	102	total clock cycles miss penalty



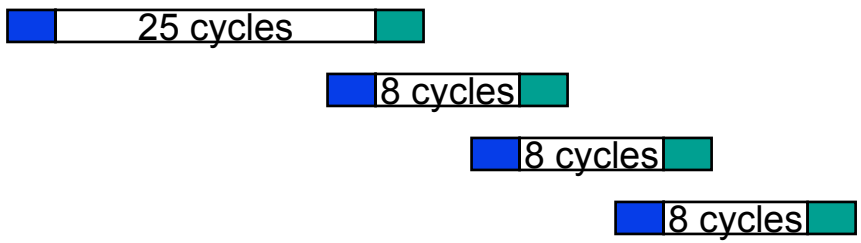
Ο αριθμός των bytes (bandwidth) που μεταφέρονται κάθε κύκλο για κάθε miss είναι (4 x 4)/102 = 0.157 bytes per clock

ΟΡΓΑΝΩΣΗ ΜΝΗΜΗΣ ΜΙΑΣ ΛΕΞΗΣ (ONE WORD MEMORY) (Συν)



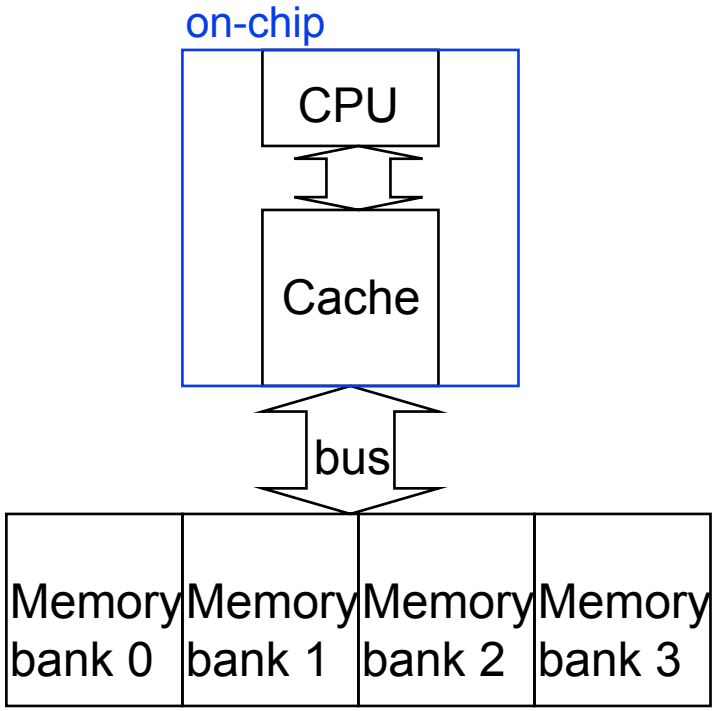
- Τι γίνεται όταν το block size είναι τέσσερις λέξεις και χρησημοποιούμε fast page mode DRAM?

1	cycle to send 1st address
25 + 3*8 = 49	cycles to read DRAM
1	cycles to return last data word
51	total clock cycles miss penalty



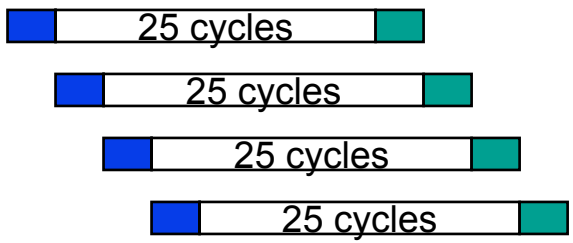
- Ο αριθμός των bytes (bandwidth) που μεταφέρονται κάθε κύκλο για κάθε miss είναι (4 x 4)/51 = 0.314 bytes per clock

ΔΙΑΚΛΑΔΩΜΕΝΗ ΜΝΗΜΗ (Interleaved Memory)



- Για ένα block size τεσσάρων λέξεων

	1	cycle to send 1st address
25 + 3 =	28	cycles to read DRAM
	1	cycles to return last data word
	30	total clock cycles miss penalty



- Ο αριθμός των bytes (bandwidth) που μεταφέρονται κάθε κύκλο για κάθε miss είναι (4 x 4)/30 = 0.533 bytes per clock

ΠΕΡΙΛΗΨΗ ΣΥΣΤΗΜΑΤΟΣ DRAM

- Είναι πολύ σημαντικό να σχεδιάζετε με βάση τα χαρακτηριστικά του cache
  - Τα caches διαβάζουν ένα block κάθε φορά (συνήθως περισσότερο από μια λέξη μόνο)
- με τα χαρακτηριστικά της DRAM
  - Χρήση DRAMs που υποστηρίζουν προσπέλαση πολλαπλών λέξεων, κάτα προτίμηση να διαβάζουν το ίδιο block size με το cache
- με τα χαρακτηριστικά του διάυλου μνήμης
  - σηγουρευόμαστε ότι ο δίαυλος μνήμης DRAM είναι συμβατός με την ροή που υποστηρίζει η μνήμη και τον τρόπο προσπέλασης
  - με απότερο σκοπό να βελτιώνουμε το εύρος του διάυλου μνήμης με το Cache (bandwidth)

Επόμενη διάλεξη και υπενθιμήσεις

- Επόμενη διάλεξη
  - Reading assignment – PH 7.2
- Υπενθιμήσεις
  - Αύριο Εργαστήριο 3-5 (GP006)
  - VHDL Tutorial?