Προηγμένη Αρχιτεκτονική Υπολογιστών

Δίκτυα Διασύνδεσης

Νεκτάριος Κοζύρης & Διονύσης Πνευματικάτος

{nkoziris,pnevmati}@cslab.ece.ntua.gr

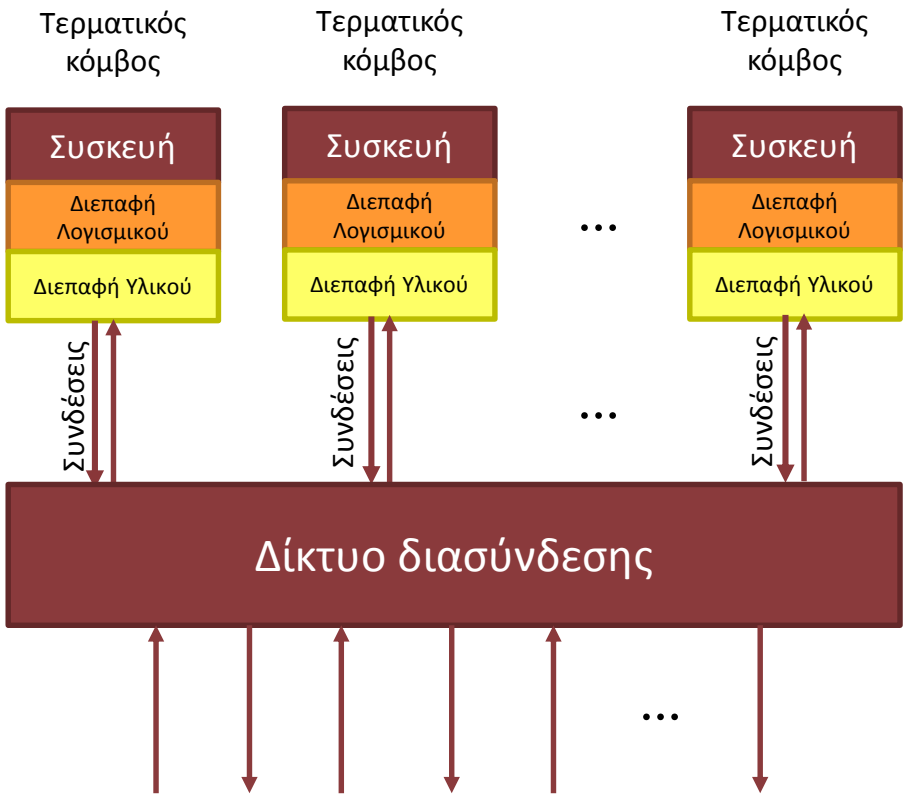
8ο εξάμηνο ΣΗΜΜΥ – Ακαδημαϊκό Έτος: 2019-20

http://www.cslab.ece.ntua.gr/courses/advcomparch/

Εισαγωγή

- Δίκτυα διασύνδεσης για τη σύνδεση πολλών «συσκευών» σε ένα σύστημα

- Από τι αποτελείται ένα δίκτυο;
  - Τερματικοί κόμβοι (end nodes)
  - Συνδέσεις (links)
  - Δικτυακά στοιχεία (routers, switches)



- Παραδείγματα:
  - Επεξεργαστές και επεξεργαστές
  - Επεξεργαστές και μνήμες
  - Επεξεργαστές και κρυφές μνήμες
  - Κρυφές μνήμες και κρυφές μνήμες
  - Επεξεργαστές και συσκευές Ε/Ε

Σημασία δικτύων διασύνδεσης

- Τα δίκτυα διασύνδεσης σχεδιάζονται ώστε:
  - να μεταφέρουν τη μέγιστη δυνατή πληροφορία (εύρος ζώνης - bandwidth)
  - στον ελάχιστο δυνατό χρόνο (χρόνος απόκρισης - latency)
  - χωρίς να δημιουργούν bottlenecks (κλιμακωσιμότητα - scalability)
  - με το ελάχιστο δυνατό κόστος και την ελάχιστη δυνατή κατανάλωση ενέργειας

- Επηρεάζουν την **κλιμακωσιμότητα** του συστήματος
  - Πόσο μεγάλα συστήματα μπορούμε να φτιάξουμε;
  - Πόσο εύκολα μπορούμε να προσθέσουμε επιπλέον συσκευές;

- Επηρεάζουν την **επίδοση** και την **ενεργειακή απόδοση**
  - Πόσο γρήγορα μπορούν να επικοινωνήσουν επεξεργαστές, κρυφές μνήμες, μνήμη;
  - Πόσο διαρκεί η πρόσβαση στη μνήμη;
  - Πόση ενέργεια καταναλώνεται στην επικοινωνία;

Ορολογία

- **Endpoint**
  - Ακραίο σημείο στο δίκτυο από όπου εισάγονται/εξάγονται δεδομένα από/προς το δίκτυο
- **Διεπαφή δικτύου** (network interface)
  - Η μονάδα που συνδέει τα endpoints (ακραία σημεία) στο δίκτυο
- **Σύνδεσμος** (link)
  - Το «καλώδιο» που μεταφέρει ένα σήμα
- **Κανάλι** (channel)
  - Η λογική σύνδεση μεταξύ διακοπτών/δρομολογητών
- **Διακόπτης/δρομολογητής** (switch/router)
  - Η μονάδα που συνδέει πεπερασμένο αριθμό καναλιών εισόδου με πεπερασμένο αριθμό καναλιών εξόδου
- **Κόμβος** (node)
  - Διακόπτης/δρομολογητής εντός του δικτύου
- **Μήνυμα** (message)
  - Μονάδα μεταφοράς δεδομένων για τις συσκευές του δικτύου
- **Πακέτο** (packet)
  - Μονάδα μεταφοράς για το δίκτυο

Μετρικές για την αξιολόγηση των δικτύων διασύνδεσης

- **Επίδοση**
  - *Χρόνος απόκρισης (Latency):* ο χρόνος που απαιτείται για να φτάσει το πρώτο byte πληροφορίας από τον αποστολέα στον παραλήπτη
  - *Εύρος ζώνης (Bandwidth):* ο ρυθμός μετάδοσης της πληροφορίας

- **Κόστος / Κατανάλωση ενέργειας**
  - Αριθμός ports στους διακόπτες
  - Αριθμός διακοπτών
  - Αριθμός συνδέσεων

- **Κλιμακωσιμότητα/Επεκτασιμότητα**
  - Η δυνατότητα του δικτύου να υποστηρίξει επέκταση σε μεγαλύτερο πλήθος συσκευών χωρίς bottlenecks

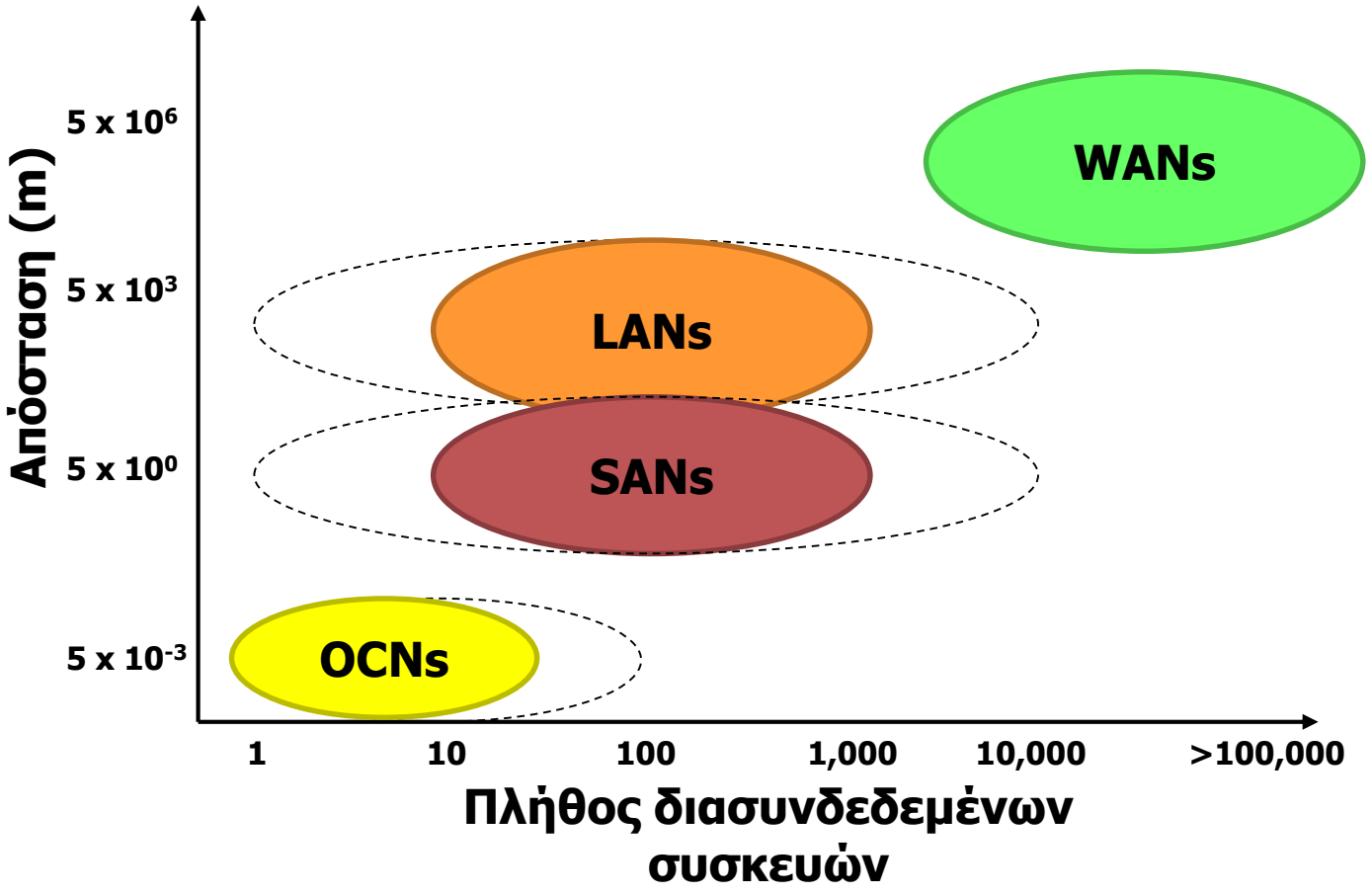
Κατηγορίες δικτύων διασύνδεσης

- Τα δίκτυα διασύνδεσης κατηγοριοποιούνται με τρεις τρόπους:
  1. Με βάση το πλήθος και την απόσταση των συσκευών προς διασύνδεση
  2. Με βάση την τοπολογία
  3. Με βάση τον τύπο των συνδέσεων

Κατηγορίες δικτύων με βάση την απόσταση των συσκευών

- **On-Chip Network (OCN)** ή **Network-On-Chip (NoC)**
  - Caches
  - Cores
  - CMPs
- **System/Storage Area Networks (SAN)**
  - Επεξεργαστές με μονάδες μνήμης
  - Υπολογιστές μεταξύ τους
  - Υπολογιστές με συσκευές Ε/Ε
- **Local Area Networks (LAN)**
  - Υπολογιστές σε ένα τοπικό δίκτυο
- **Wide Area Networks (WAN)**
  - Υπολογιστές οπουδήποτε στον κόσμο

Κατηγορίες δικτύων με βάση την απόσταση των συσκευών

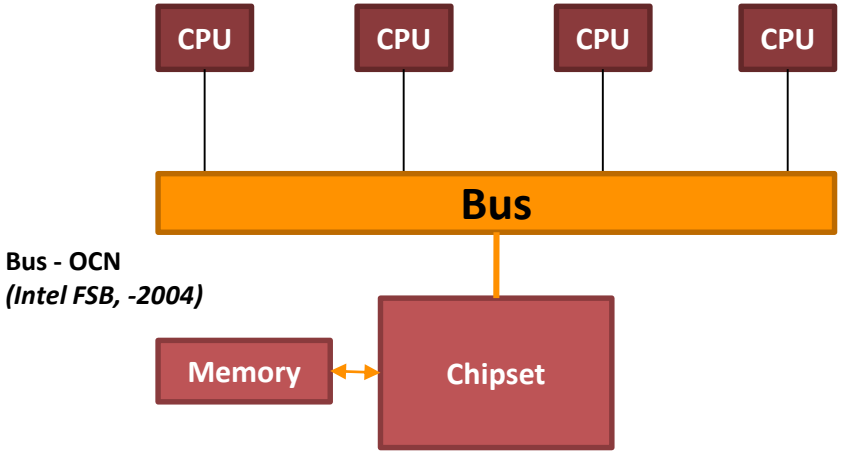


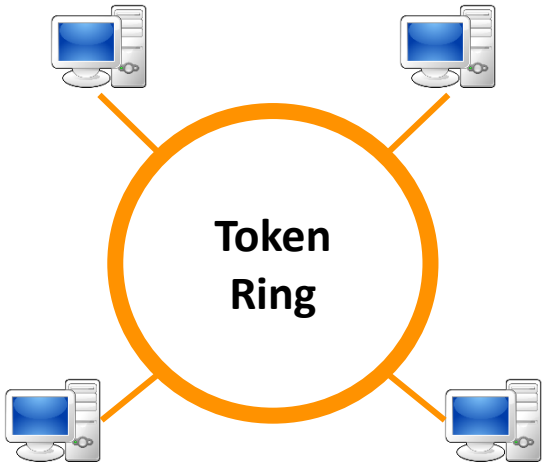
Κατηγορίες δικτύων με βάση την τοπολογία

- **Shared-medium Networks**
  - Όλες οι συσκευές (endpoints) *μοιράζονται* το *ίδιο* μέσο διασύνδεσης
  - Μόνο μία συσκευή μπορεί να χρησιμοποιεί το μέσο διασύνδεσης κάθε φορά
- **Direct / Point-to-point Networks**
  - Κάθε συσκευή (endpoint) συνδέεται σε ένα router
  - Κάθε router συνδέεται με συνδέσεις με ένα ή περισσότερα άλλα routers
  - Οι συσκευές μπορούν να χρησιμοποιούν το δίκτυο ταυτόχρονα
- **Indirect / Switch-based Networks**
  - Δεν υπάρχει απευθείας σύνδεση μεταξύ δύο συσκευών (endpoints)
  - Κάθε συσκευή έχει έναν προσαρμογέα δικτύου και μία σύνδεση προς έναν διακόπτη
  - Ένας ή περισσότεροι διακόπτες αναλαμβάνουν τη σύνδεση μεταξύ δύο συσκευών

Shared-medium Networks

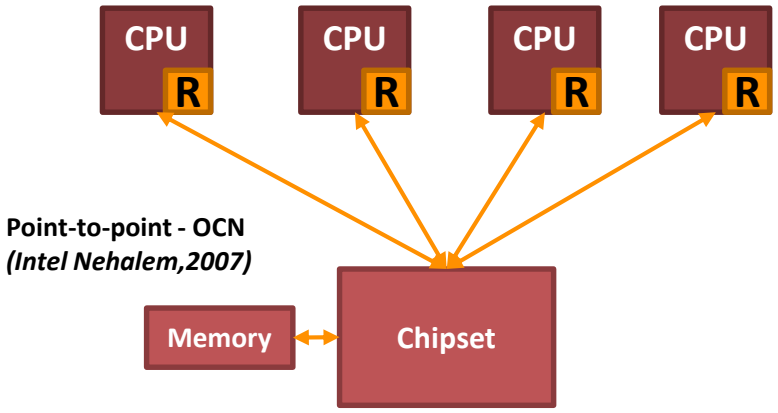
- **Πλεονεκτήματα**
  - Ελάχιστη πολυπλοκότητα
  - Δυνατότητα broadcast
    - Σημαντικό για snoop cache coherence protocols, barriers...
- **Μειονεκτήματα**
  - Περιορισμένο bandwidth
  - Υποστήριξη μικρού αριθμού συσκευών



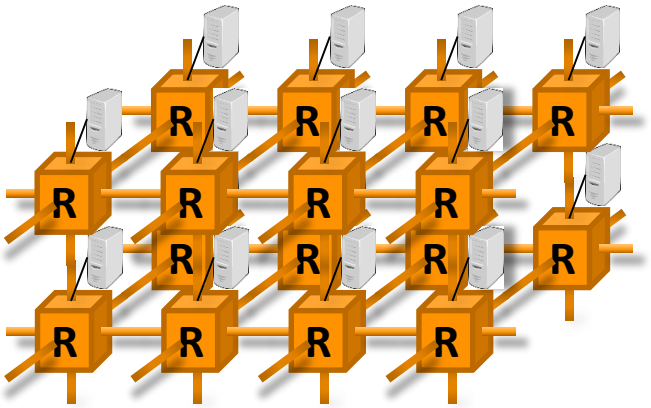


Direct Networks

- **Πλεονεκτήματα**
  - Είναι κλιμακώσιμα
  - Είναι επεκτάσιμα
- **Μειονεκτήματα**
  - Χρειάζονται routers
    - Ο επεξεργαστής μπορεί να κάνει το routing σε μικρή κλίμακα
  - Χρειάζονται περισσότερους συνδέσμους
- Χρησιμοποιούνται ευρέως σε συστήματα μεγάλης κλίμακας
  - Τοπολογίες n-διάστατου πλέγματος, τόρου, υπερκύβου



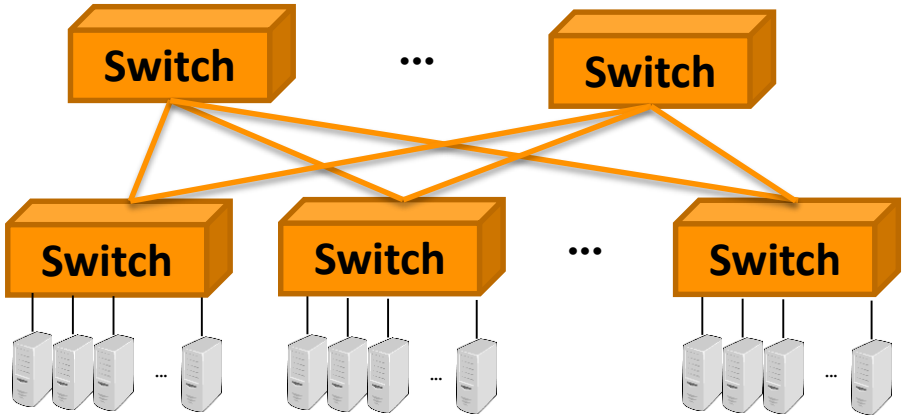
Point-to-point - OCN
(Intel Nehalem,2007)



3D-mesh - SAN
(Cray XT3 "Seastar" 2004)

Indirect Networks

- **Πλεονεκτήματα**
  - Είναι κλιμακώσιμα
  - Είναι επεκτάσιμα
  - Προσφέρουν μεγαλύτερο εύρος ζώνης
- **Μειονεκτήματα**
  - Χρειάζονται switches (μεγάλο κόστος)



Fat Tree - SAN
(InfiniBand fat tree)

Κατηγορίες δικτύων με βάση τον τύπο των συνδέσεων

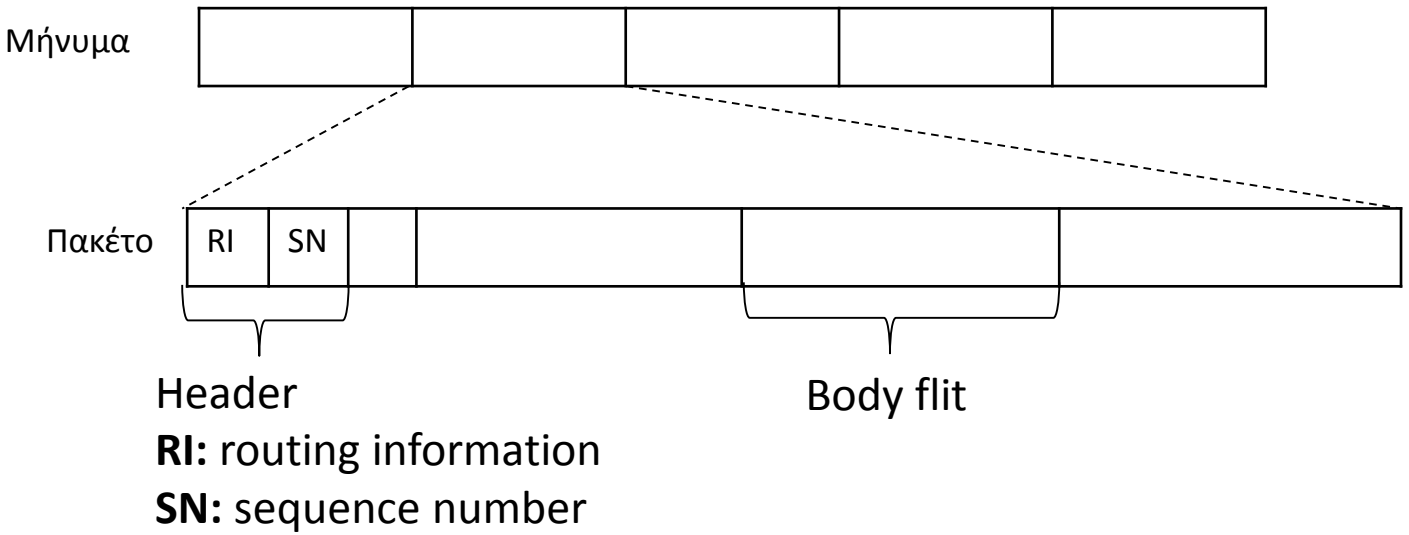
- **Στατικά Δίκτυα**
  - Οι συνδέσεις μεταξύ των συσκευών είναι προκαθορισμένες και στατικές
  - Τα direct networks είναι στατικά
    - Κάθε συσκευή είναι ένας κόμβος
    - Οι συνδέσεις μεταξύ των κόμβων είναι στατικές και καθορίζονται από την τοπολογία
- **Δυναμικά Δίκτυα**
  - Οι συνδέσεις δημιουργούνται από το δίκτυο όταν χρειάζονται
  - Τα indirect networks είναι δυναμικά
    - Κάθε διακόπτης είναι ένας κόμβος
    - Οι διακόπτες του δικτύου καθορίζουν τις συνδέσεις δυναμικά

Βασικές έννοιες στα δίκτυα διασύνδεσης

- **Τοπολογία**
  - Καθορίζει τον τρόπο της σύνδεσης των κόμβων/συσκευών
  - Επηρεάζει τη δρομολόγηση (routing), το throughput, το χρόνο απόκρισης, την ευκολία κατασκευής του δικτύου διασύνδεσης
- **Δρομολόγηση** (Routing)
  - Αφορά το πώς διαβιβάζεται ένα μήνυμα από την αφετηρία στον προορισμό
  - Μπορεί να είναι στατική ή δυναμική
- **Αποθήκευση** (Buffering) και **έλεγχος ροής** (Flow control)
  - Τι αποθηκεύεται στο δίκτυο; (τίποτα, ολόκληρα πακέτα, κάποια πακέτα κλπ.)
  - Τι γίνεται αν ένα πακέτο δεν μπορεί να μεταφερθεί λόγω συμφόρησης; Πώς κάνουμε throttling;
  - Το buffering και το flow control σχετίζονται με τη στρατηγική δρομολόγησης

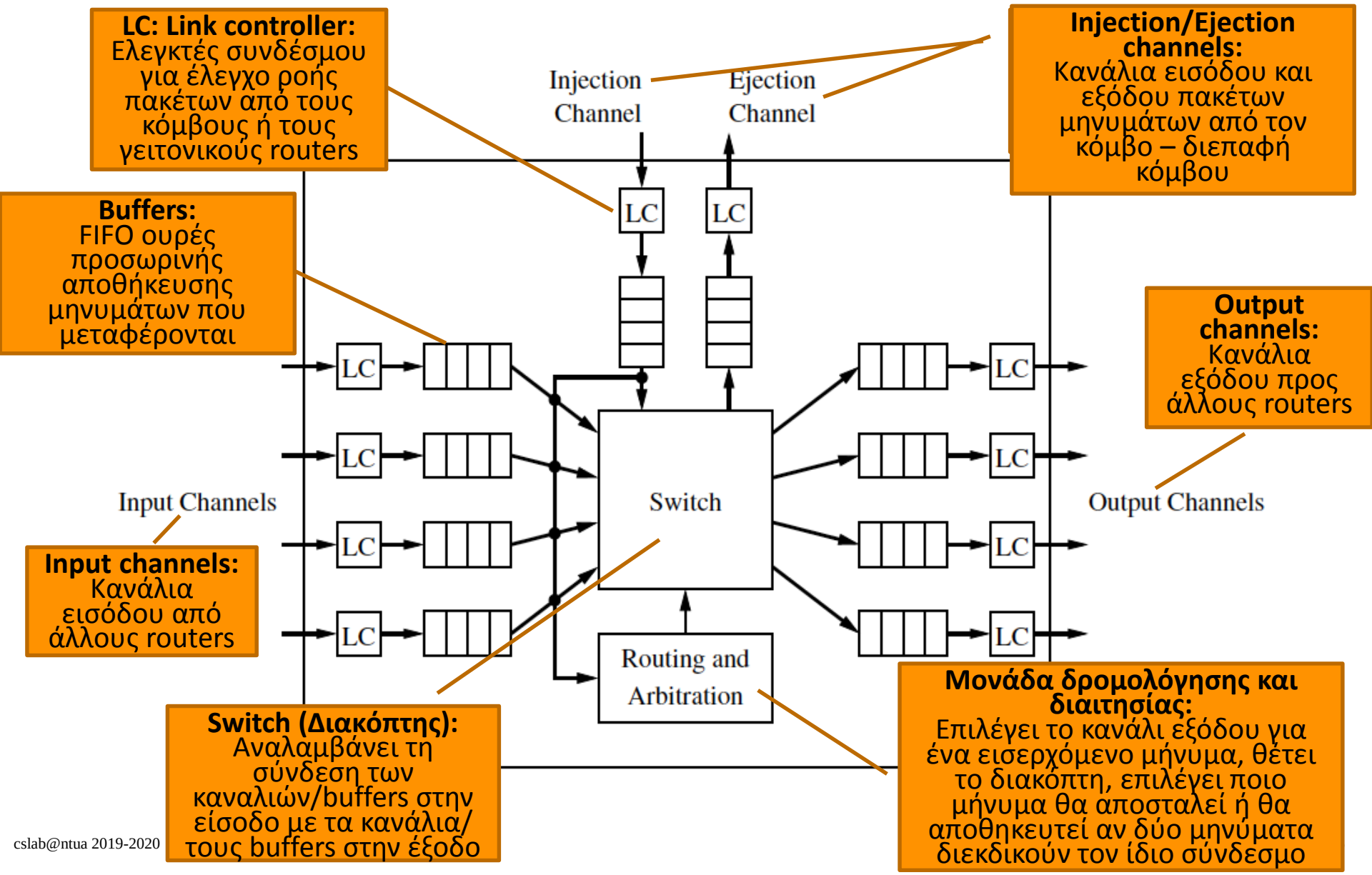
Μηνύματα και Πακέτα

- Το **μήνυμα** είναι μία ακολουθία από bytes που πρέπει να μεταφερθεί από την αφετηρία στον προορισμό
- Αν ένα μήνυμα είναι πολύ μεγάλο, διαιρείται σε μονάδες με περιορισμένο μέγιστο μήκος, που λέγεται **πακέτο**



- Το πακέτο μπορεί να διαιρείται περαιτέρω σε **flits**
  - βασικές μονάδες του ελέγχου ροής

Γενική αρχιτεκτονική ενός δρομολογητή (router)



LC: Link controller: Ελεγκτές συνδέσμου για έλεγχο ροής πακέτων από τους κόμβους ή τους γειτονικούς routers

Injection/Ejection channels: Κανάλια εισόδου και εξόδου πακέτων μηνυμάτων από τον κόμβο – διεπαφή κόμβου

Buffers: FIFO ουρές προσωρινής αποθήκευσης μηνυμάτων που μεταφέρονται

Output channels: Κανάλια εξόδου προς άλλους routers

Input channels: Κανάλια εισόδου από άλλους routers

Switch (Διακόπτης): Αναλαμβάνει τη σύνδεση των καναλιών/buffers στην είσοδο με τα κανάλια/ τους buffers στην έξοδο

Μονάδα δρομολόγησης και διαιτησίας: Επιλέγει το κανάλι εξόδου για ένα εισερχόμενο μήνυμα, θέτει το διακόπτη, επιλέγει ποιο μήνυμα θα αποσταλεί ή θα αποθηκευτεί αν δύο μηνύματα διεκδικούν τον ίδιο σύνδεσμο

Γενική μορφή του pipeline του δρομολογητή

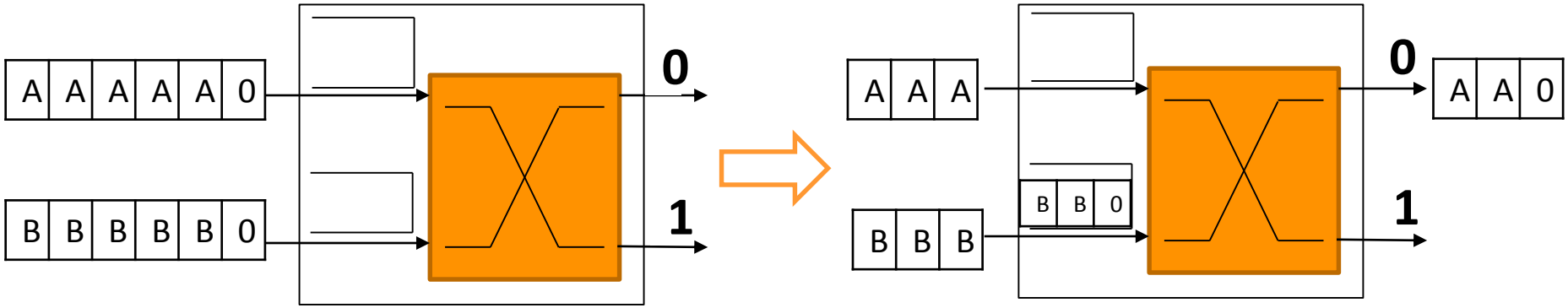
- Pipeline 5 σταδίων
  1. BW: Buffer write – Προσωρινή αποθήκευση στην ουρά των buffers
     - Ένα πακέτο εισέρχεται στο δρομολογητή και αποθηκεύεται στον buffer εισόδου
     - Το στάδιο παραλείπεται αν δεν υπάρχουν buffers (bufferless routers)
  2. RC: Routing computation – Υπολογισμός προορισμού
     - Η μονάδα δρομολόγησης και διαιτησίας προσδιορίζει τον προορισμό του πακέτου
  3. SA: Switch allocation – Δέσμευση του διακόπτη
     - Η μονάδα δρομολόγησης και διαιτησίας δεσμεύει το διακόπτη και θέτει κατάλληλα την είσοδο και την έξοδο
  4. ST: Switch traversal – Διάσχιση του διακόπτη
     - Το πακέτο διασχίζει το διακόπτη
  5. LT: Link traversal – Διάσχιση του συνδέσμου
     - Το πακέτο διασχίζει το σύνδεσμο εξόδου προς τον προορισμό

Έλεγχος ροής

- Έστω ότι δύο πακέτα θέλουν να χρησιμοποιήσουν τον ίδιο σύνδεσμο την ίδια στιγμή
- Τι επιλογές υπάρχουν;

Έλεγχος ροής

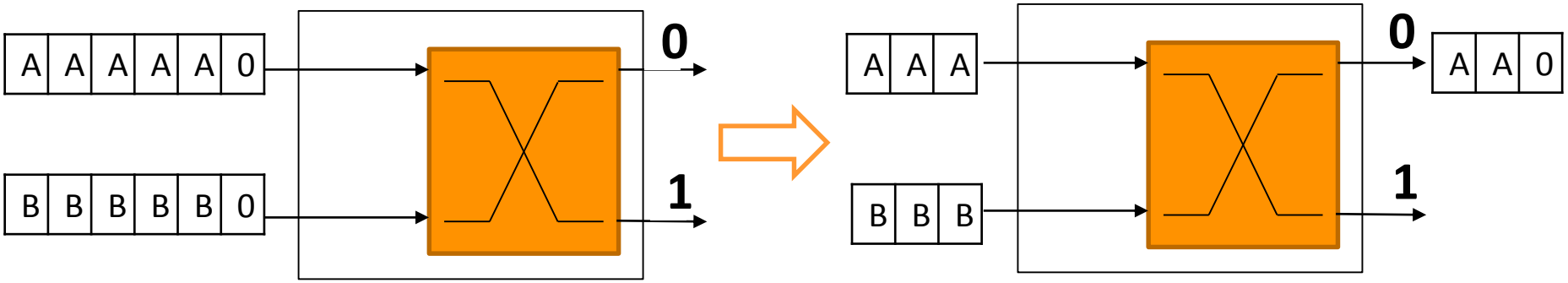
- Έστω ότι δύο πακέτα θέλουν να χρησιμοποιήσουν τον ίδιο σύνδεσμο την ίδια στιγμή
- Τι επιλογές υπάρχουν;
  1. **Buffering** - στέλνω το ένα πακέτο, αποθηκεύω προσωρινά το άλλο
     - Μόνο για buffered networks



- *Απαιτείται έλεγχος ροής για το buffering*

Έλεγχος ροής

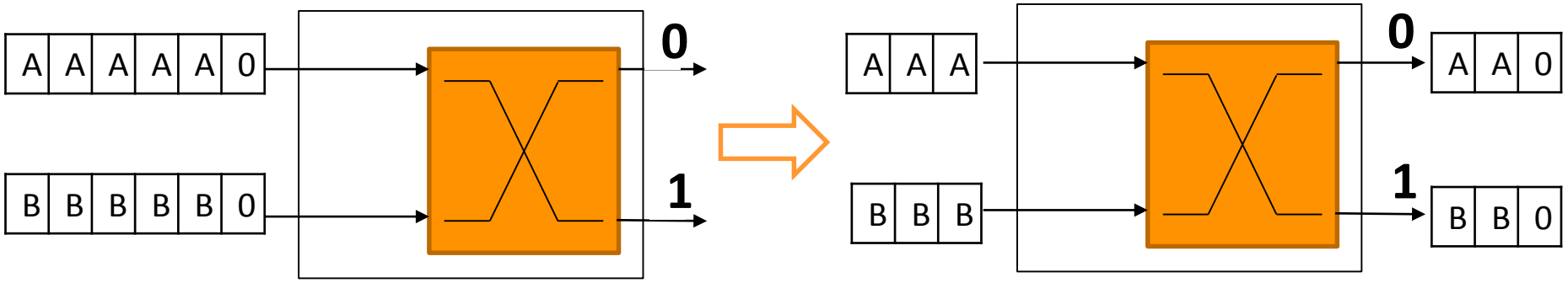
- Έστω ότι δύο πακέτα θέλουν να χρησιμοποιήσουν τον ίδιο σύνδεσμο την ίδια στιγμή
- Τι επιλογές υπάρχουν;
  1. **Buffering** - στέλνω το ένα πακέτο, αποθηκεύω προσωρινά το άλλο
     - Μόνο για buffered networks
  2. **Drop** – στέλνω το ένα πακέτο, πετάω το άλλο



- *Ο αποστολέας πρέπει να ξαναστείλει το μήνυμα B*

Έλεγχος ροής

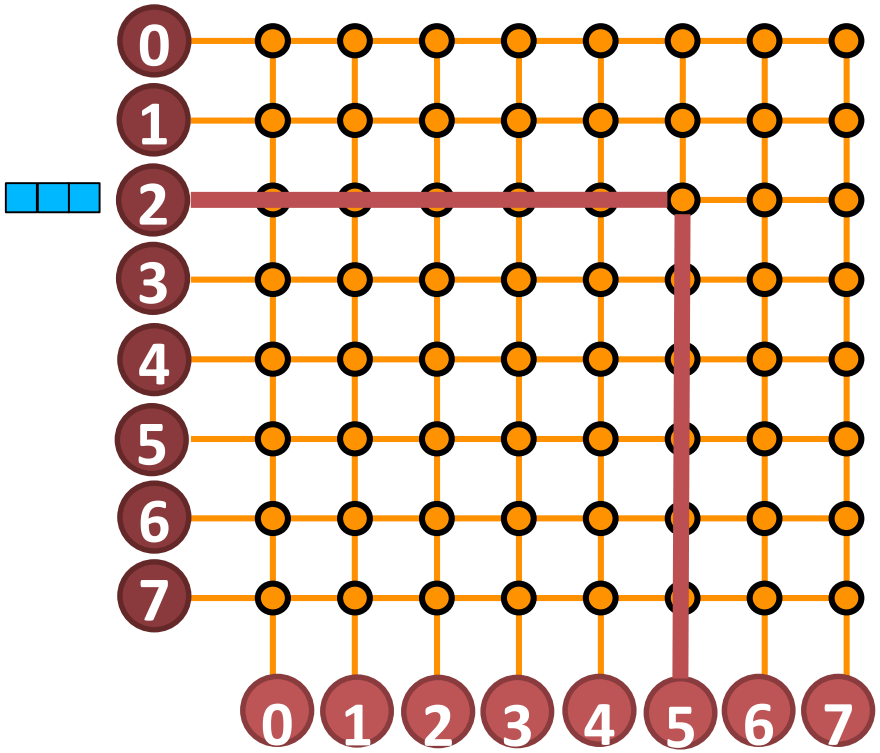
- Έστω ότι δύο πακέτα θέλουν να χρησιμοποιήσουν τον ίδιο σύνδεσμο την ίδια στιγμή
- Τι επιλογές υπάρχουν;
  1. **Buffering** - στέλνω το ένα πακέτο, αποθηκεύω προσωρινά το άλλο
     - Μόνο για buffered networks
  2. **Drop** – στέλνω το ένα πακέτο, πετάω το άλλο
  3. **Deflect** – στέλνω το ένα πακέτο, στέλνω το δεύτερο πακέτο κάπου αλλού



- *Το πακέτο του μηνύματος Β πρέπει να ξαναδρομολογηθεί αργότερα στο δίκτυο*
- *Γνωστό και ως **hot-potato routing***

Circuit switching

- Έστω ότι θέλω να στείλω ένα μήνυμα από το 2 στο 5
- Με **circuit switching**:
  - Το μονοπάτι κατασκευάζεται πριν ξεκινήσει η αποστολή
    - Δημιουργείται ένα κύκλωμα από το 2 στο 5
  - Το μονοπάτι καταστρέφεται αφού τελειώσει η αποστολή

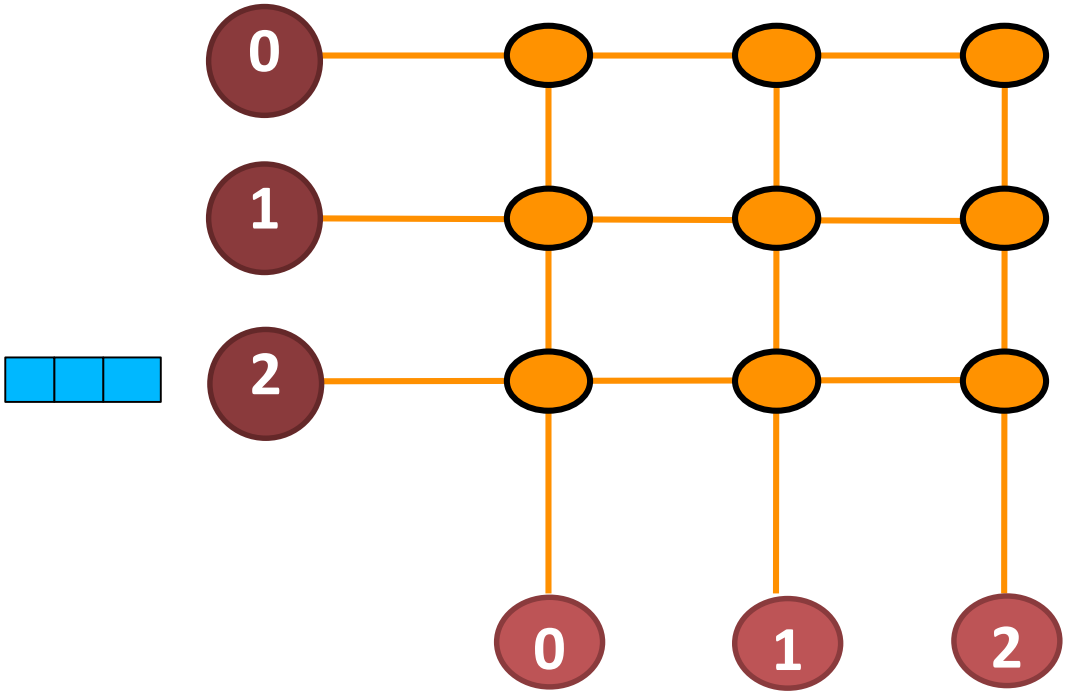


Circuit switching

- **Circuit switching:** Μέθοδος ελέγχου ροής για bufferless δίκτυα
  - Πριν την αποστολή, δημιουργείται ένα κύκλωμα (circuit) από την αφετηρία ως τον προορισμό
  - Όλα τα πακέτα από τη συγκεκριμένη αφετηρία στο συγκεκριμένο προορισμό στέλνονται μέσω του κυκλώματος
  - Αν δεν υπάρχουν άλλα πακέτα προς αποστολή, το κύκλωμα απελευθερώνεται
  - Οι σύνδεσμοι που απαρτίζουν το κύκλωμα δεν μπορούν να χρησιμοποιηθούν για άλλα κυκλώματα
- **Πλεονεκτήματα**
  - Γρήγορη διαιτησία (π.χ. σε crossbar δίκτυα)
  - Δεν απαιτείται buffering
  - Δε με ενδιαφέρει το μέγεθος του μηνύματος
- **Μειονεκτήματα**
  - Η δημιουργία του κυκλώματος κοστίζει σε χρόνο
  - Οι σύνδεσμοι του κυκλώματος μπορεί να υπο-χρησιμοποιούνται

Store-and-forward switching

- Έστω ότι θέλω να στείλω ένα μήνυμα από το 2 στο 1
- Με **store-and-forward switching**:
  - Ένα πακέτο δρομολογείται από την αφετηρία σε έναν κόμβο του δικτύου
  - Το πακέτο αποθηκεύεται ολόκληρο σε αυτό τον κόμβο
  - Το πακέτο δρομολογείται στον επόμενο κόμβο
  - Η διαδικασία επαναλαμβάνεται μέχρι να φτάσει στον προορισμό

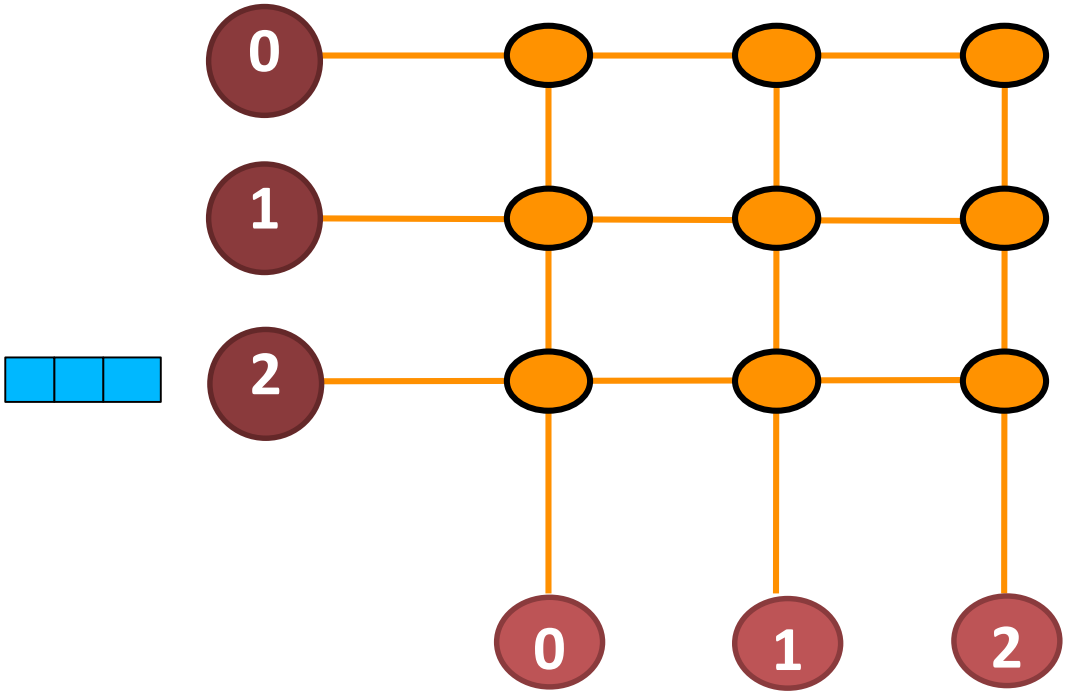


Store-and-forward packet switching

- **Store-and-forward packet switching:** Μέθοδος ελέγχου ροής για buffered δίκτυα
  - Κάθε πακέτο δρομολογείται ξεχωριστά
  - Το πακέτο αποθηκεύεται *πάντα* πριν προωθηθεί
  - Αν ένας σύνδεσμος δε χρησιμοποιείται, οποιοδήποτε πακέτο μπορεί να μεταφερθεί πάνω από αυτόν
- **Πλεονεκτήματα**
  - Δεν απαιτείται χρόνος για τη δημιουργία κυκλώματος
  - Οι σύνδεσμοι χρησιμοποιούνται καλύτερα
  - Λειτουργεί καλύτερα για μικρά, συχνά μηνύματα
- **Μειονεκτήματα**
  - Η ροή είναι δυναμική
    - Θεωρητικά ο έλεγχος ροής μπορεί να κοστίζει περισσότερο
    - Η διαίρεση σε πακέτα έχει κάποιο κόστος
  - Ο χρόνος απόκρισης ανά πακέτο είναι υψηλός

Virtual cut-through (packet) switching

- Έστω ότι θέλω να στείλω ένα μήνυμα από το 2 στο 1
- Με **virtual cut-through switching**:
  - Αν υπάρχει αποθηκευτικός χώρος στην έξοδο, το πρώτο πακέτο (header) δρομολογείται από την αφετηρία προς τον προορισμό
  - Τα υπόλοιπα πακέτα ακολουθούν με pipelined τρόπο



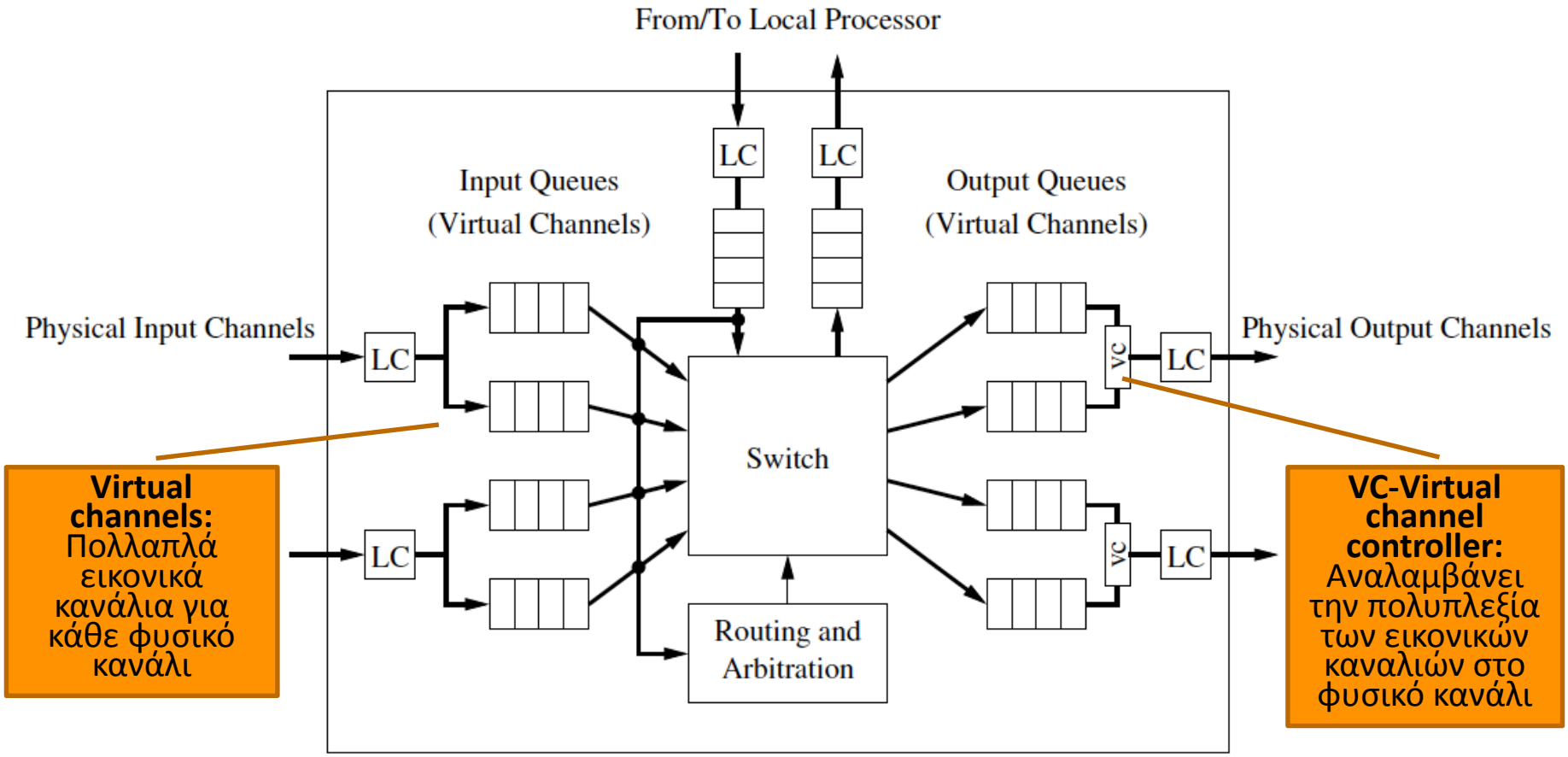
Virtual cut-through (packet) switching

- **Virtual cut-through switching:** Μέθοδος ελέγχου ροής για buffered δίκτυα
  - Κάθε πακέτο δρομολογείται ξεχωριστά
  - Αν ο buffer στο κανάλι εξόδου είναι ελεύθερος, ξεκινά η δρομολόγηση του πακέτου
    - Ο header του πακέτου περιέχει την πληροφορία δρομολόγησης και αποστέλλεται πρώτος
    - Τα υπόλοιπα bytes μεταφέρονται με pipelined τρόπο
  - Αν ένας σύνδεσμος δε χρησιμοποιείται, οποιοδήποτε πακέτο μπορεί να μεταφερθεί πάνω από αυτόν
- **Πλεονεκτήματα**
  - Δεν απαιτείται χρόνος για τη δημιουργία κυκλώματος
  - Μειώνεται ο χρόνος απόκρισης σε σχέση με το store-and-forward
    - Όχι οι απαιτήσεις σε buffers!
  - Οι σύνδεσμοι χρησιμοποιούνται καλύτερα
- **Μειονεκτήματα**
  - Η ροή είναι δυναμική
  - Αν **ο header μπλοκάρει** στην έξοδο, απαιτείται αρκετός αποθηκευτικός χώρος για όλο το πακέτο
  - Εκφυλίζεται σε store-and-forward σε περιπτώσεις υψηλού ανταγωνισμού
  - Απαιτεί διαχείριση deadlocks

Wormhole (flit) switching

- **Wormhole switching:** Μέθοδος ελέγχου ροής για buffered δίκτυα
  - Κάθε flit ενός πακέτου δρομολογείται ξεχωριστά
  - Ο έλεγχος ροής για κάθε flit είναι αντίστοιχος με τον έλεγχο ροής πακέτων στο virtual cut-through switching
    - Αν ο buffer στο κανάλι εξόδου είναι ελεύθερος, ξεκινά η δρομολόγηση του header flit
    - Τα υπόλοιπα flits δρομολογούνται με pipelined τρόπο
  - Τα διάφορα flits που αποτελούν ένα πακέτο μπορούν να προωθούνται στο δίκτυο
  - Flits του ίδιου πακέτου μπορεί να βρίσκονται αποθηκευμένα σε buffers διαφορετικών routers πριν την ολοκλήρωση της αποστολής του πακέτου
- **Πλεονεκτήματα**
  - Απαιτείται μικρότερος αποθηκευτικός χώρος ανά router
    - Flit bytes << Packet bytes
  - Ο χρόνος απόκρισης καθορίζεται από το μέγεθος του flit
- **Μειονεκτήματα**
  - Αν **το header flit μπλοκάρει**, ολόκληρο το πακέτο μπλοκάρει
  - Μεγαλύτερη πολυπλοκότητα στην αποφυγή deadlocks
    - Πολλά flits του ίδιου πακέτου σε διαφορετικούς routers

Virtual channels – Παραλληλισμός στην αρχιτεκτονική



- Χωρίς εικονικά κανάλια, αν ένα μήνυμα καταλάβει μία ουρά από buffers, κανένα άλλο μήνυμα δεν μπορεί να χρησιμοποιήσει το κανάλι
- Τα εικονικά κανάλια λειτουργούν σαν πολλαπλά φυσικά κανάλια μικρότερης ταχύτητας
- Βελτιώνουν το χρόνο απόκρισης και το throughput του δικτύου

Μηχανισμοί δρομολόγησης

- Υπάρχουν τρεις μηχανισμοί δρομολόγησης
  - **Αριθμητική:** Βασίζεται σε απλή αριθμητική για να προσδιορίσει ένα μονοπάτι σε κανονικές τοπολογίες, ή στη διάσταση σε πιο σύνθετες τοπολογίες (dimension-order routing)
  - **Με βάση την αφετηρία:** Η αφετηρία επιλέγει την πόρτα εξόδου σε κάθε διακόπτη στο μονοπάτι
  - **Με αναζήτηση σε πίνακα (table lookup):** Η πόρτα εξόδου αναζητείται σε έναν πίνακα
    - Χρησιμοποιείται κυρίως στο packet switching

Αλγόριθμοι δρομολόγησης

- Υπάρχουν τρεις τύποι αλγορίθμων δρομολόγησης
  - **Deterministic:** Για ένα συγκεκριμένο ζεύγος αφετηρίας-προορισμού, επιλέγεται πάντα το ίδιο μονοπάτι
  - **Oblivious:** Για ένα συγκεκριμένο ζεύγος αφετηρίας-προορισμού, επιλέγονται διαφορετικά μονοπάτια, άσχετα με την κατάσταση του δικτύου
  - **Adaptive:** Για ένα συγκεκριμένο ζεύγος αφετηρίας-προορισμού, επιλέγονται διαφορετικά μονοπάτια, ανάλογα με την κατάσταση του δικτύου
    - Για την προσαρμογή στην κατάσταση του δικτύου απαιτείται τροφοδότηση από το δίκτυο (τοπικά ή συνολικά)
    - Τα διαφορετικά μονοπάτια μπορεί να είναι ελάχιστα ή μη ελάχιστα

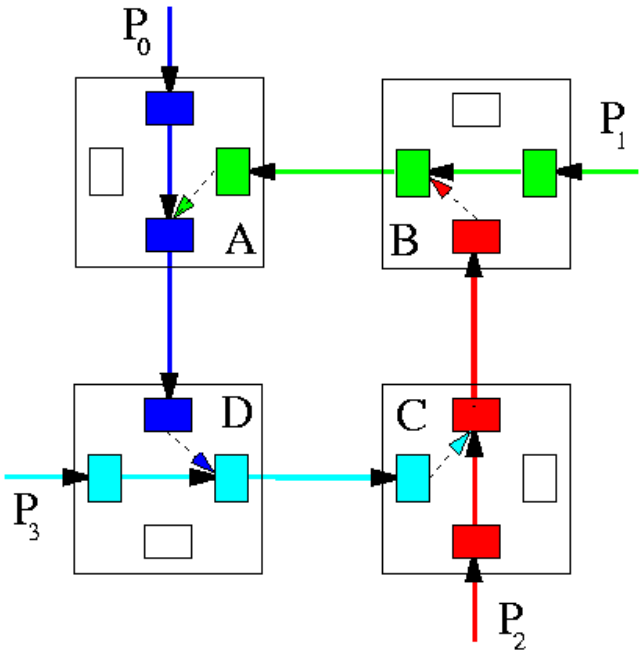
Deterministic routing

- Στη ντετερμινιστική δρομολόγηση, όλα τα πακέτα από μία συγκεκριμένη αφετηρία σε έναν συγκεκριμένο προορισμό ακολουθούν την ίδια διαδρομή
  - *Δεν απαιτεί circuit switching*
- **Dimension-order routing:** Αλγόριθμος ντετερμινιστικής δρομολόγησης
  - Διάσχιση του δικτύου ανά διάσταση
    - Π.χ. σε ένα 2D-mesh, πρώτα διάσχιση κατά X, μετά διάσχιση κατά Y
  - **Πλεονεκτήματα**
    - Απλός αλγόριθμος
    - Δε δημιουργεί deadlocks *στις περισσότερες τοπολογίες*
  - **Μειονεκτήματα**
    - Μπορεί να δημιουργήσει φαινόμενα ανταγωνισμού
    - Δεν αξιοποιεί τα διαφορετικά μονοπάτια στο δίκτυο

Deadlocks στη ντετερμινιστική δρομολόγηση

- **Deadlock:** Μια κατάσταση όπου δεν υπάρχει πρόοδος προς τα εμπρός
  - Προκαλείται από κυκλικές εξαρτήσεις στο δίκτυο
  - Ένα πακέτο περιμένει για έναν buffer που καταλαμβάνεται από κάποιο άλλο πακέτο που κινείται προς την ίδια κατεύθυνση

- Επίλυση deadlocks:
  - Αποφυγή κύκλων στη δρομολόγηση
    - Dimension-order routing
    - Κίνηση με συγκεκριμένη κατεύθυνση
  - Επιπλέον προσωρινός αποθηκευτικός χώρος
  - Εντοπισμός των deadlocks και επίλυση



Oblivious routing

- Στην oblivious δρομολόγηση, τα πακέτα μπορούν να κινηθούν από διαφορετικά μονοπάτια από την αφετηρία στον προορισμό, χωρίς να λαμβάνουν υπόψη την κατάσταση του δικτύου
- **Αλγόριθμος Valiant:** Αλγόριθμος για oblivious δρομολόγηση
  1. Τυχαία επιλογή ενός ενδιάμεσου προορισμού
  2. Δρομολόγηση από την αφετηρία ως τον ενδιάμεσο προορισμό
  3. Δρομολόγηση από τον ενδιάμεσο προορισμό στον τελικό προορισμό
     - Η ενδιάμεση δρομολόγηση μπορεί να είναι διαφορετική - π.χ. dimension-order
  - **Πλεονεκτήματα**
    - Η επιλογή τυχαίων προορισμών κατανέμει περισσότερο ομοιόμορφα το φορτίο στο δίκτυο
      - Random pattern -> Uniform traffic
  - **Μειονεκτήματα**
    - Τα μονοπάτια που επιλέγονται δεν είναι ελάχιστα
  - **Εναλλακτικά**
    - Αξίζει να χρησιμοποιηθεί όταν το φορτίο στο δίκτυο είναι υψηλό

Adaptive routing

- Στην προσαρμοστική δρομολόγηση, τα πακέτα μπορούν να κινηθούν από διαφορετικά μονοπάτια από την αφετηρία στον προορισμό, λαμβάνοντας υπόψη την τρέχουσα κατάσταση του δικτύου
- **Ελάχιστη προσαρμογή:** Αλγόριθμος για προσαρμοστική δρομολόγηση
  - Ο δρομολογητής ελέγχει την κατάσταση του δικτύου για να αποφασίσει πού θα στείλει ένα πακέτο
  - Επιλέγει το port που είναι πιο κοντά στον προορισμό -> ελάχιστα μονοπάτια
  - **Πλεονεκτήματα**
    - Έχει επίγνωση των τοπικών φαινομένων ανταγωνισμού
  - **Μειονεκτήματα**
    - Η επιλογή ελάχιστου μονοπατιού μειώνει τη δυνατότητα εξισορρόπησης του φορτίου
- **Μη-ελάχιστη προσαρμογή:** Αλγόριθμος για προσαρμοστική δρομολόγηση
  - Ο δρομολογητής στέλνει πακέτα σε κάποιο σημείο στο δίκτυο, άσχετα με την απόστασή του από τον προορισμό
  - **Πλεονεκτήματα**
    - Μπορεί να επιτύχει εξισορρόπηση φορτίου και καλύτερη χρήση του δικτύου
  - **Μειονεκτήματα**
    - Πρέπει να διασφαλίζει ότι δε θα εμφανιστεί livelock

Τοπολογία

- Bus
  - η απλούστερη τοπολογία
- Point-to-point συνδέσεις
  - η ταχύτερη αλλά πιο ακριβή τοπολογία
- Crossbar
  - λιγότερο ακριβή από τις point-to-point συνδέσεις
- Ring
- Tree
- Omega
- Hypercube
- Mesh
- Torus
- Butterfly
- Dragonfly
- Slimfly
- ...

Μετρικές αξιολόγησης ενός δικτύου

- Χρόνος απόκρισης (latency)
  - Σχετίζεται με την τεχνολογία του δικτύου και τον αλγόριθμο δρομολόγησης
- Εύρος ζώνης (bandwidth)
  - Σχετίζεται με την τεχνολογία του δικτύου
- Εύρος τομής (bisection bandwidth)
  - Σχετίζεται με την τοπολογία του δικτύου και το μέγεθος του δικτύου
- Διάμετρος
  - Σχετίζεται με την τοπολογία του δικτύου και το μέγεθος του δικτύου
- Κόστος
  - Πλήθος δικτυακών στοιχείων, πλήθος συνδέσμων, μέγεθος/πολυπλοκότητα δικτυακών στοιχείων
- Ανταγωνισμός – Συμφόρηση (Contention - Congestion)
- Κατανάλωση ενέργειας
- Συνολική επίδοση του συστήματος

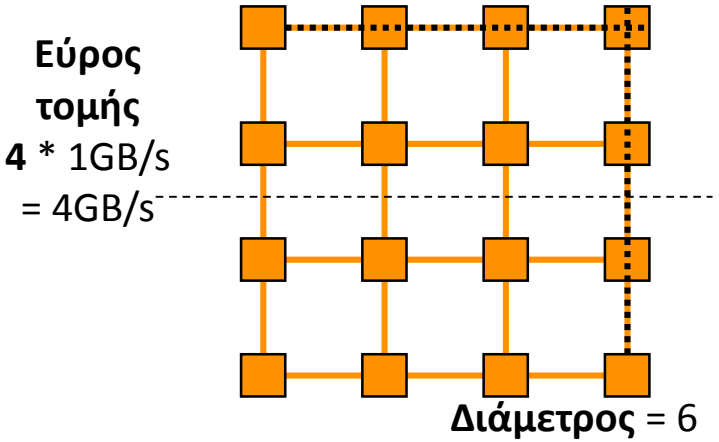
Χρόνος απόκρισης και Εύρος ζώνης

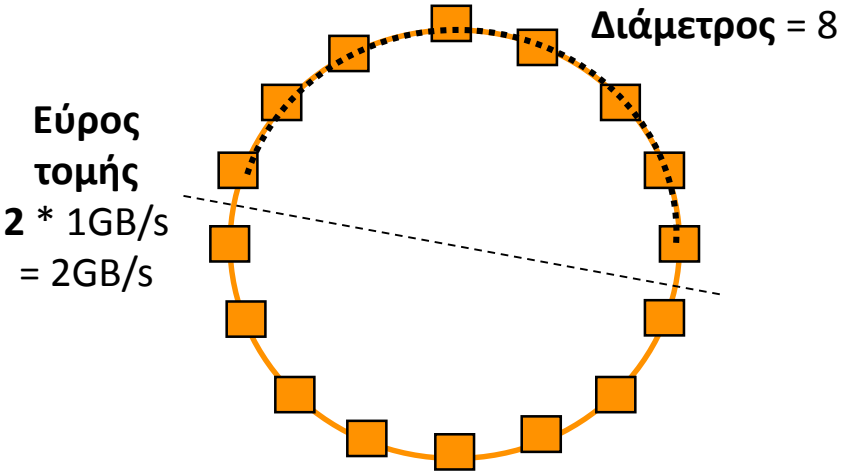
- **Χρόνος απόκρισης:** ο χρόνος που απαιτείται για την αποστολή της πρώτης μονάδας του μηνύματος
  - Ορίζεται ως χρόνος για το πρώτο byte, το πρώτο πακέτο, το πρώτο flit
  - Εξαρτάται από την τεχνολογία του δικτύου αλλά και τον τρόπο και τον αλγόριθμο δρομολόγησης

- **Εύρος ζώνης:** ο ρυθμός μετάδοσης του μηνύματος (bytes/second)
  - Εξαρτάται από την τεχνολογία του δικτύου

Εύρος τομής και διάμετρος

- Εύρος τομής: Το άθροισμα του εύρους ζώνης όλων των συνδέσμων αν διαιρέσουμε το δίκτυο σε δύο μέρη με ίσο αριθμό κόμβων
  - Εξαρτάται από την τοπολογία και το μέγεθος του δικτύου
- Διάμετρος: Το μεγαλύτερο ελάχιστο μονοπάτι μεταξύ οποιωνδήποτε δύο κόμβων στο δίκτυο
  - Εξαρτάται από την τοπολογία και το μέγεθος του δικτύου
- *Παράδειγμα:* Δίκτυα 16 κόμβων με συνδέσμους 1GB/s διπλής κατεύθυνσης

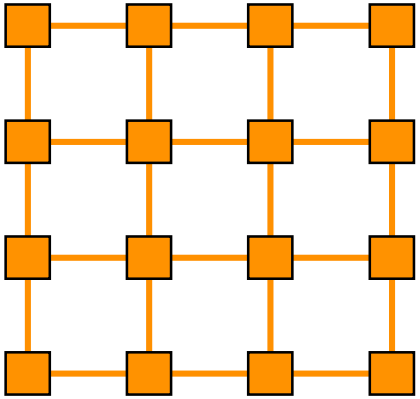




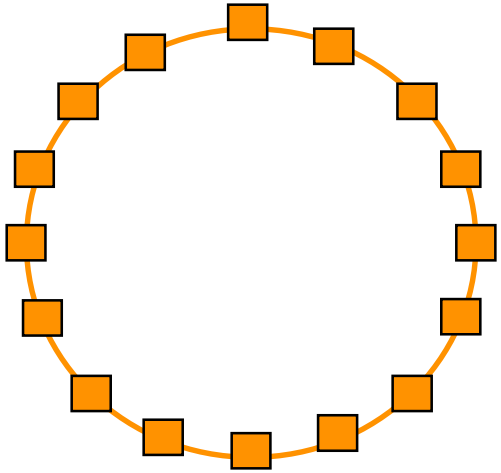
Κόστος

- Το κόστος του δικτύου καθορίζεται από:
  - το πλήθος των δικτυακών στοιχείων
  - το μέγεθος ή/και την πολυπλοκότητα των δικτυακών στοιχείων
  - το πλήθος των συνδέσμων

- *Παράδειγμα:* Δίκτυα 16 κόμβων

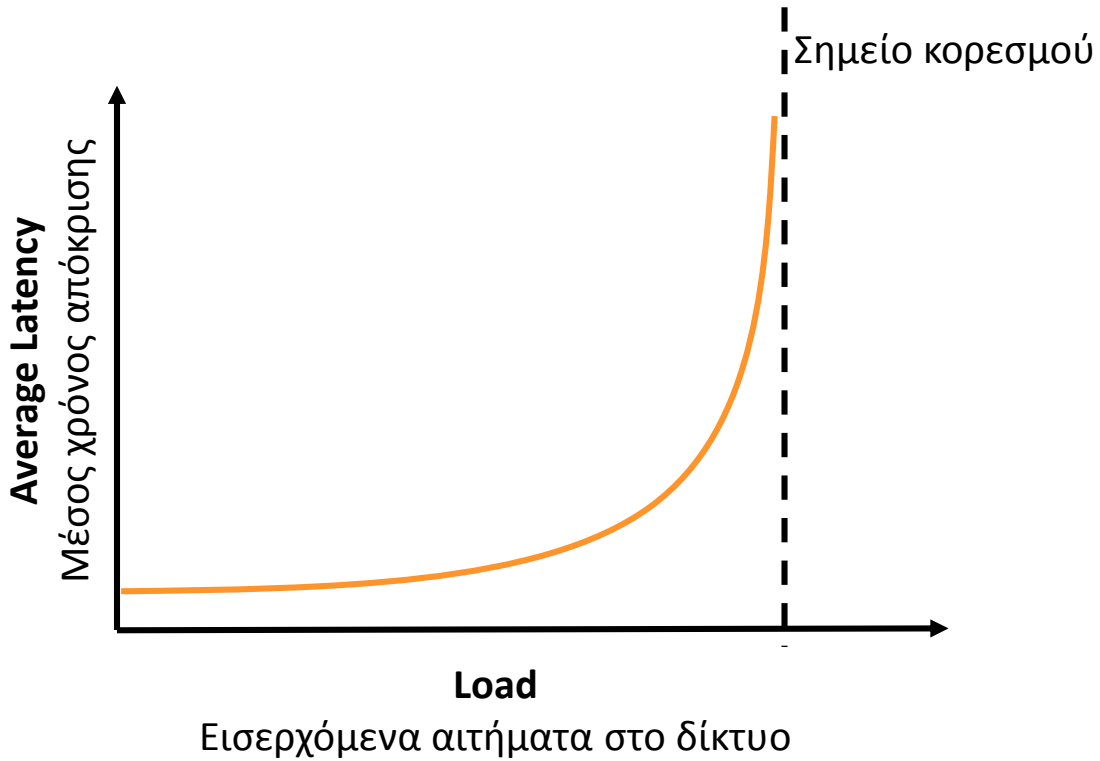


-16 διακόπτες
-4 θύρες/διακόπτη
-24 σύνδεσμοι



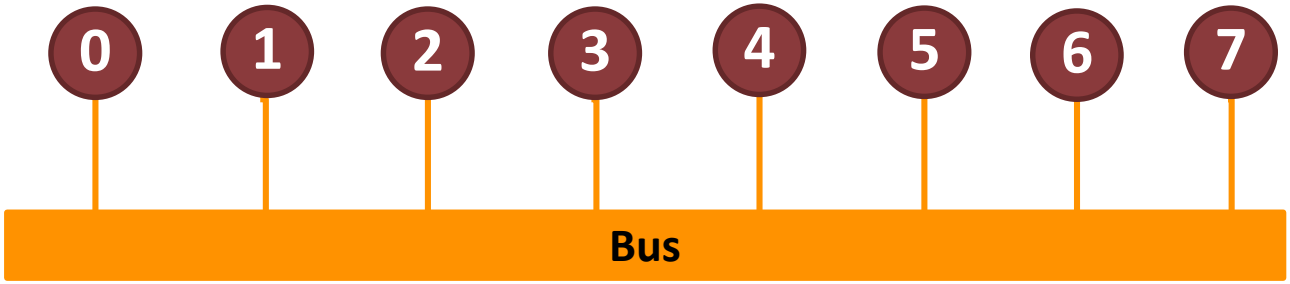
-16 διακόπτες
-2 θύρες/διακόπτη
-16 σύνδεσμοι

Συνολική επίδοση του συστήματος



- Διαφορετικές τοπολογίες παρουσιάζουν διαφορετικά σημεία κορεσμού
- Διαφορετικές τεχνικές δρομολόγησης παρουσιάζουν διαφορετικά σημεία κορεσμού
- Διαφορετικές τεχνικές ελέγχου ροής παρουσιάζουν διαφορετικά σημεία κορεσμού

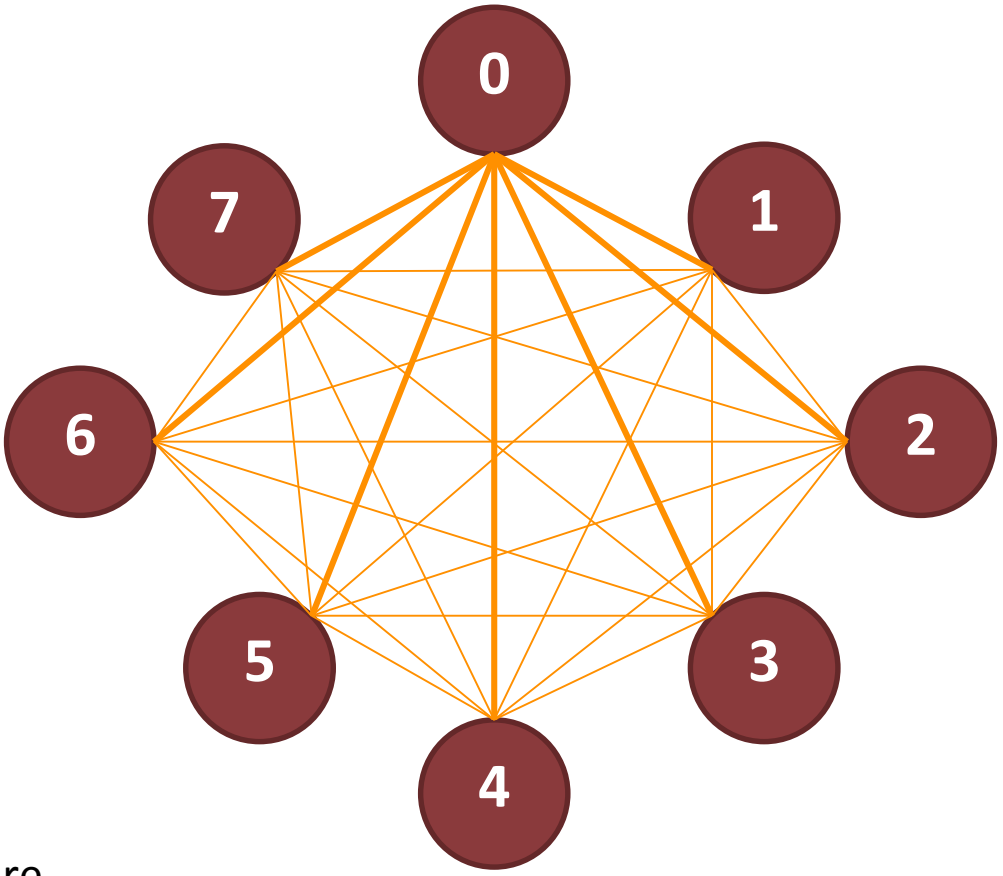
Bus



- **Bus:** Όλοι οι κόμβοι συνδέονται σε έναν σύνδεσμο
- **Πλεονεκτήματα**
  - Απλό δίκτυο με χαμηλό κόστος για μικρό πλήθος κόμβων
  - Απλή δρομολόγηση – δεν απαιτεί switching
  - Η υλοποίηση coherence είναι εύκολη (snooping, serialization)
- **Μειονεκτήματα**
  - Δεν κλιμακώνει σε μεγάλο πλήθος κόμβων
    - Το εύρος ζώνης είναι περιορισμένο
    - Υπάρχουν περιορισμοί στην αύξηση της συχνότητας για αύξηση του εύρους ζώνης
  - Εμφανίζονται φαινόμενα ανταγωνισμού και γρήγορος κορεσμός
    - Δεν μπορούν να επικοινωνήσουν ταυτόχρονα όλα τα δυνατά ζεύγη

Point-to-point

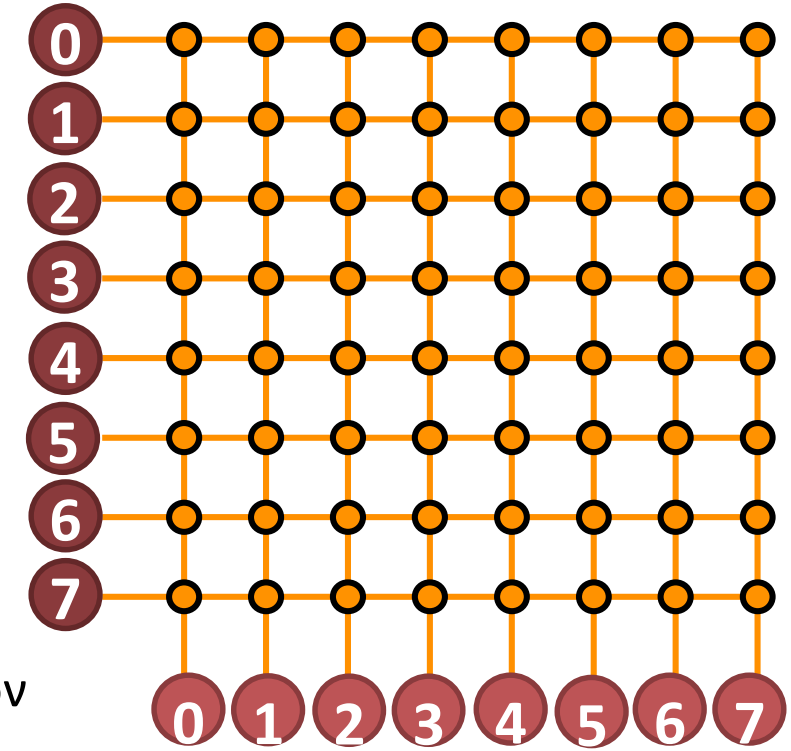
- **Point-to-point:** Κάθε κόμβος συνδέεται με κάθε άλλο κόμβο με έναν απευθείας σύνδεσμο
- **Πλεονεκτήματα**
  - Ελάχιστος ανταγωνισμός
  - Ελάχιστος χρόνος απόκρισης
  - Μέγιστη επίδοση
- **Μειονεκτήματα**
  - Μέγιστο κόστος
    - O(N) συνδέσεις/ports ανά κόμβο
    - $O(N^2)$ σύνδεσμοι
  - Δεν κλιμακώνει λόγω κόστους!
  - Υλοποιείται δύσκολα στο hardware
    - Πολύπλοκη συνδεσμολογία



- *Ιδανικά θέλουμε την επίδοση του point-to-point δικτύου και το κόστος του bus*

Crossbar

- **Crossbar:** Κάθε κόμβος συνδέεται με κάθε άλλο κόμβο με *μοιραζόμενους* συνδέσμους και διακόπτες
  - Δυναμικό και indirect δίκτυο
  - Η πολυπλοκότητα βρίσκεται στους ενδιάμεσους διακόπτες
- Επιτρέπει ***ταυτόχρονες*** μεταφορές αν οι προορισμοί ***δεν είναι ανταγωνιστικοί***
- Έχει χαμηλό κόστος για μικρό πλήθος κόμβων
- **Πλεονεκτήματα**
  - Μικρός χρόνος απόκρισης
  - Μεγάλο throughput
- **Μειονεκτήματα**
  - Δεν κλιμακώνει
    - $O(N^2)$ διακόπτες
  - Δύσκολη διαιτησία (έλεγχος εξόδου - arbitration)



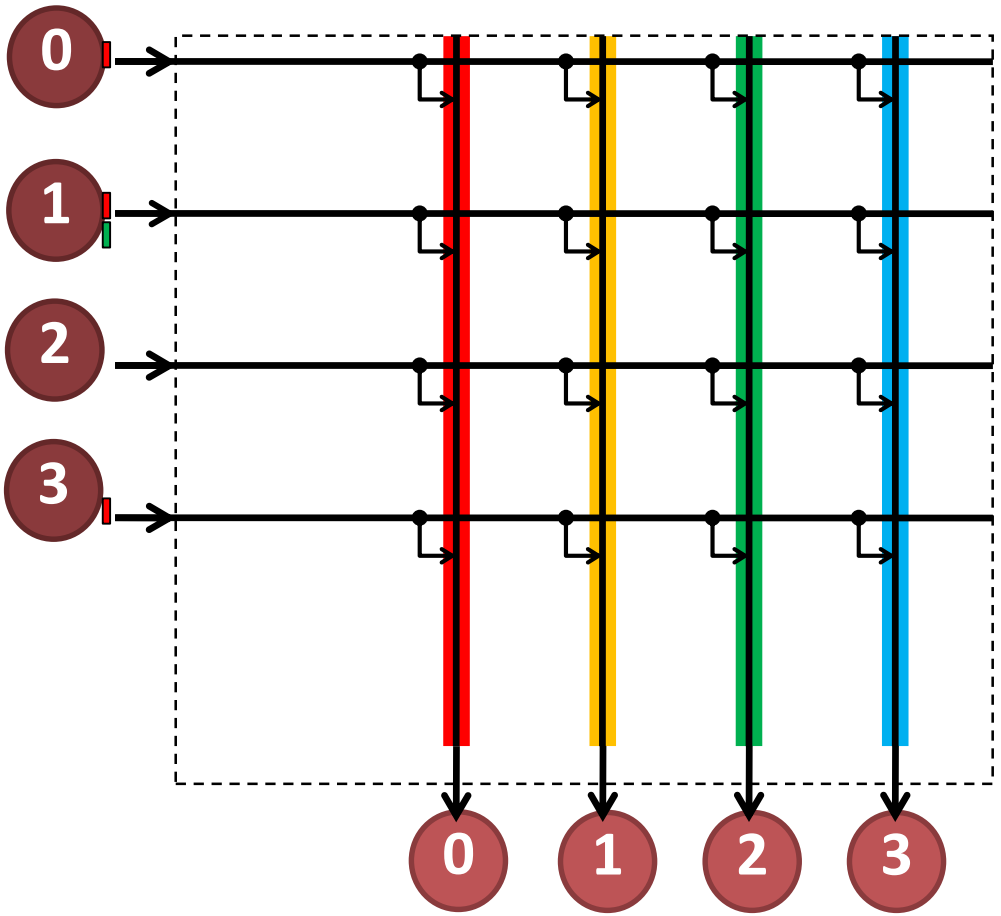
Bufferless vs Buffered crossbar networks

- Το πρόβλημα στα crossbar networks είναι η διαιτησία
  - Κάποιος πρέπει να αποφασίσει πώς θα φτάσει ένα πακέτο από την είσοδο στην έξοδο
    - Πριν την αποστολή του μηνύματος
  - Απαιτείται κεντρικός δρομολογητής
    - Μεγάλη πολυπλοκότητα

- *Εναλλακτικά:* μπορώ να αποθηκεύω πακέτα στο δίκτυο με χρήση buffers
  - Η διαιτησία γίνεται πιο εύκολη

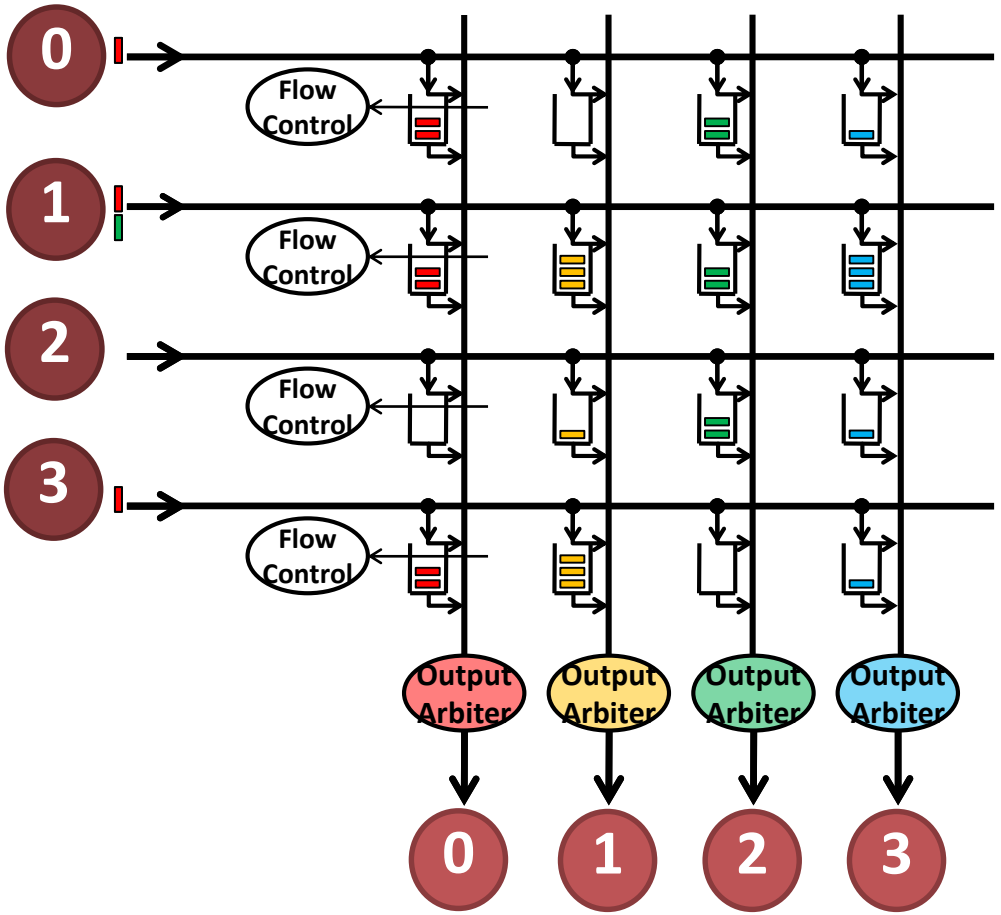
Bufferless crossbar

- Όταν υπάρχουν ανταγωνιστικοί προορισμοί, ο κεντρικός διαιτητής αποφασίζει να σειριοποίησει πακέτα
- Χαμηλό throughput!

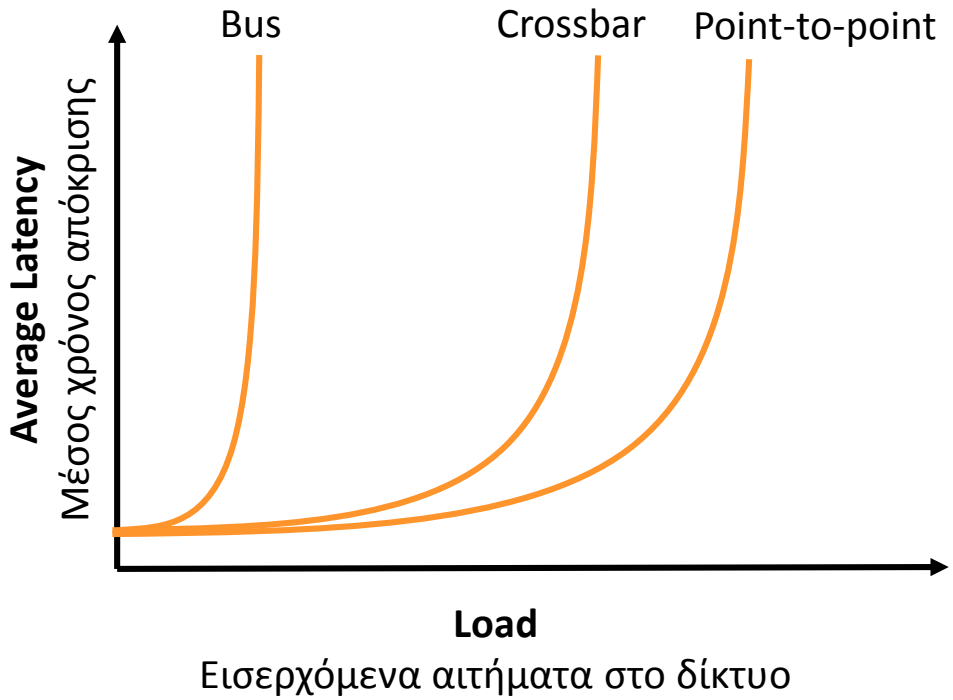


Buffered crossbar

- Προσθήκη ουρών για buffering σε όλους τους ενδιάμεσους διακόπτες
  - **Μεγάλο κόστος:** O(N^2) buffers
- Απαιτείται έλεγχος ροής για τη διαχείριση των ουρών
  - Για αποφυγή υπερχείλισης



Μια πρώτη σύγκριση

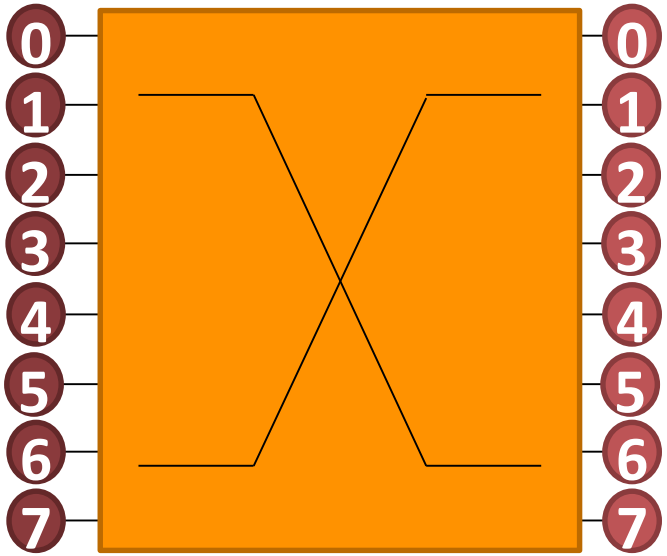


- **Throughput:** Point-to-point >> Crossbar >> Bus
- **Κόστος:** Point-to-point >> Crossbar >> Bus
- Μπορούμε να διατηρήσουμε την καλή επίδοση μειώνοντας το κόστος;

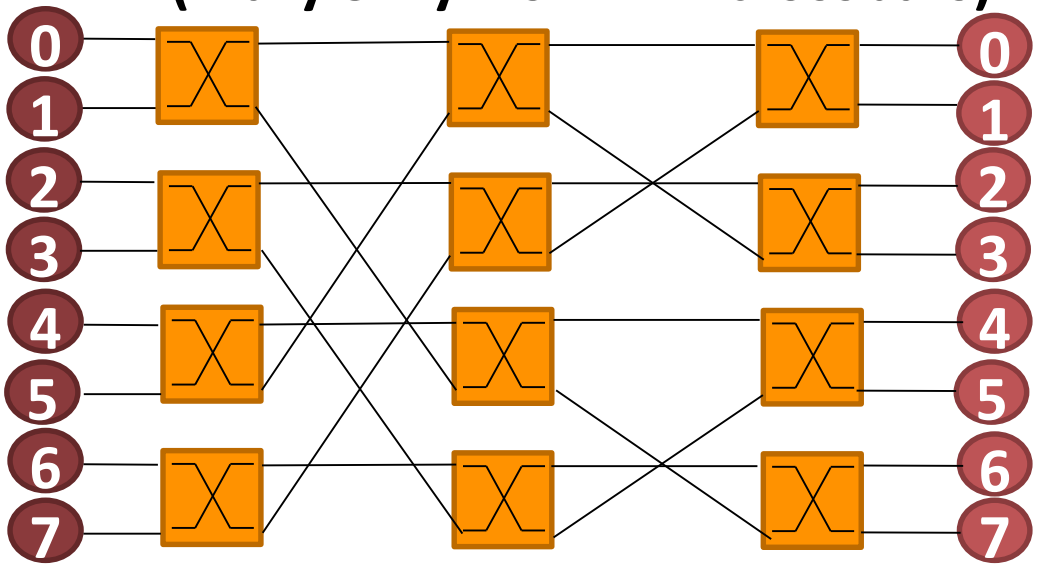
Multistage Interconnection Networks (MINs)

- Αντικατάσταση ενός μεγάλου (ακριβού) διακόπτη με περισσότερους μικρότερους (φθηνότερους) διακόπτες σε επίπεδα/στάδια

8x8 crossbar

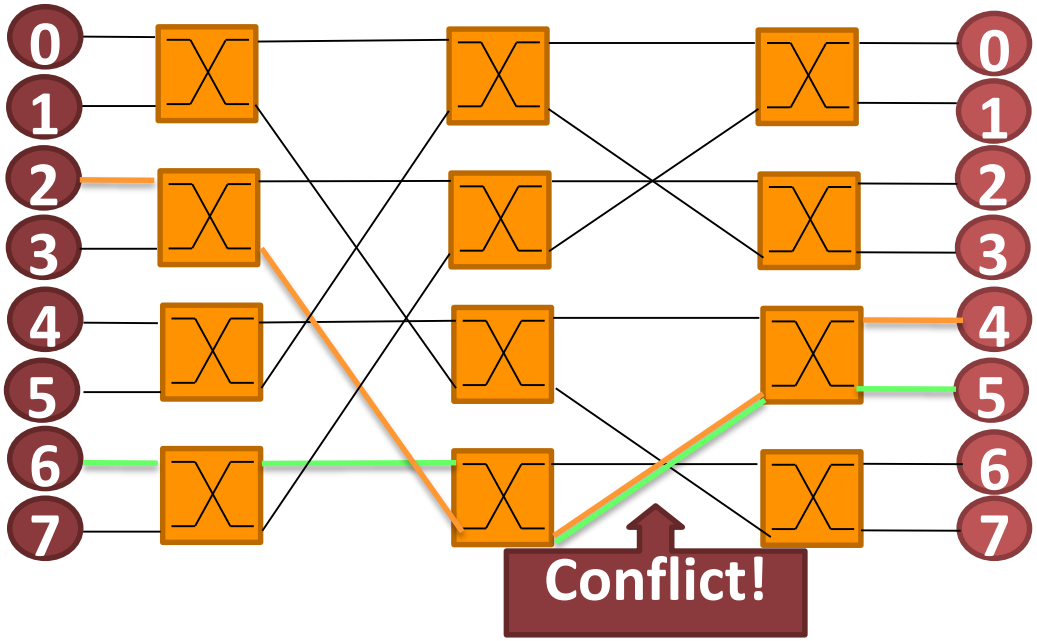


Butterfly
(2-ary 3-fly from 2x2 crossbars)



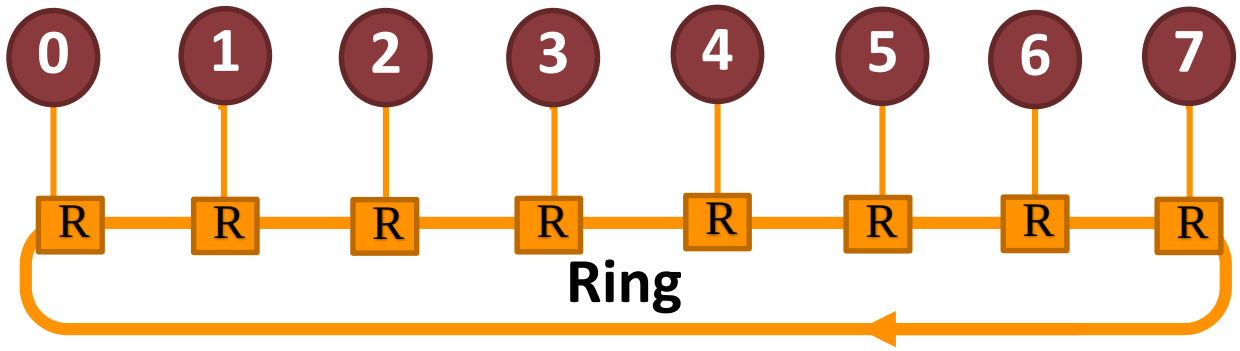
Multistage Interconnection Networks

- **Πλεονεκτήματα**
  - Κόστος: O(NlogN) διακόπτες
- **Μειονεκτήματα**
  - Μοναδικό μονοπάτι από κάθε αφετηρία σε κάθε προορισμό
  - Περισσότερος ανταγωνισμός – μικρότερο throughput σε σχέση με buffered crossbar

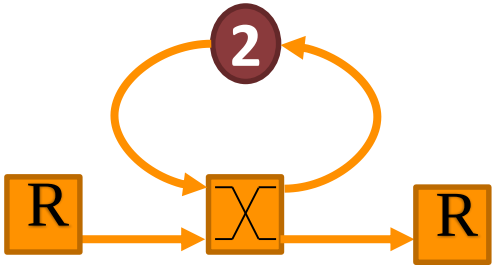


- Πολλές τοπολογίες MIN: Butterfly, Omega, Benes, Banyan κ.ά.

Ring

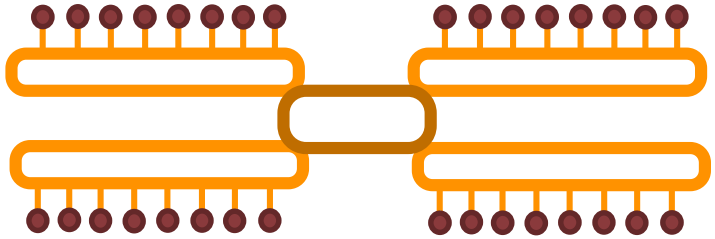


- **Ring:** Κάθε κόμβος συνδέεται με δυο άλλους κόμβους - όλοι οι κόμβοι σχηματίζουν ένα συνεχόμενο μονοπάτι
- Διαφέρει από το bus
  - Χρησιμοποιεί διακόπτες
  - Υπάρχουν point-to-point συνδέσεις μεταξύ των διακοπτών
- **Πλεονεκτήματα**
  - Κόστος: O(N)
  - Απλοί 2x2 διακόπτες σε κάθε κόμβο (π.χ. 2x2 crossbar)
- **Μειονεκτήματα**
  - Μεγάλος χρόνος απόκρισης: O(N)
  - Σταθερό εύρος τομής
  - Δεν κλιμακώνει λόγω εύρους τομής



Παραλλαγές του Ring

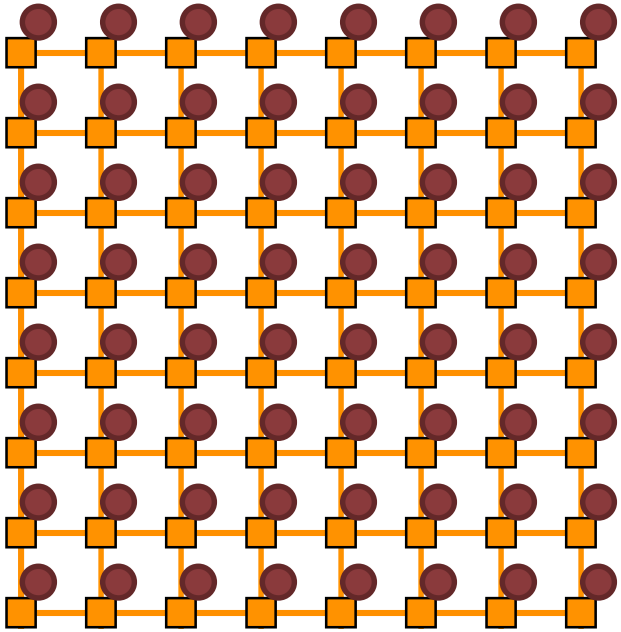
- *Εναλλακτικά* στο απλό unidirectional ring
  - Bidirectional ring
  - Πολλαπλά rings
  - Ιεραρχικά rings



- **Πλεονεκτήματα**
  - Μικρότερος χρόνος απόκρισης
  - Καλύτερη κλιμακωσιμότητα
- **Μειονεκτήματα**
  - Μεγαλύτερη πολυπλοκότητα στον έλεγχο ροής

2D-Mesh

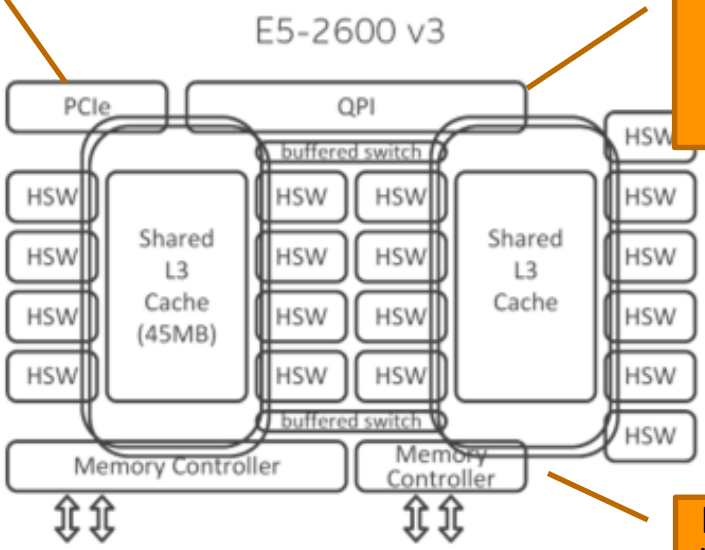
- **2D-Mesh:** Κάθε κόμβος συνδέεται μέσω ενός διακόπτη με τους 4 γειτονικούς του κόμβους στο διδιάστατο πλέγμα
  - Στατικό και direct δίκτυο
- **Πλεονεκτήματα**
  - Κόστος: O(N) διακόπτες
  - Μέσος χρόνος απόκρισης: $O(\sqrt{N})$
  - Εύκολη υλοποίηση σε chip
  - Πολλαπλά μονοπάτια από έναν κόμβο σε έναν άλλο -> μικρότερος ανταγωνισμός
- **Μειονεκτήματα**
  - Πολύπλοκοι διακόπτες (π.χ. 5x5 crossbar)
  - Η επίδοση εξαρτάται από τη θέση στο πλέγμα
    - *Λύση: τοπολογία τόρου*



Χρήση δικτύων σε πραγματικά συστήματα

- Επεξεργαστές Intel Xeon

PCIe **bus** μεταξύ επεξεργαστών και συσκευών(κάρτες γραφικών, δικτύων, FPGAs SSDs κ.ά.)

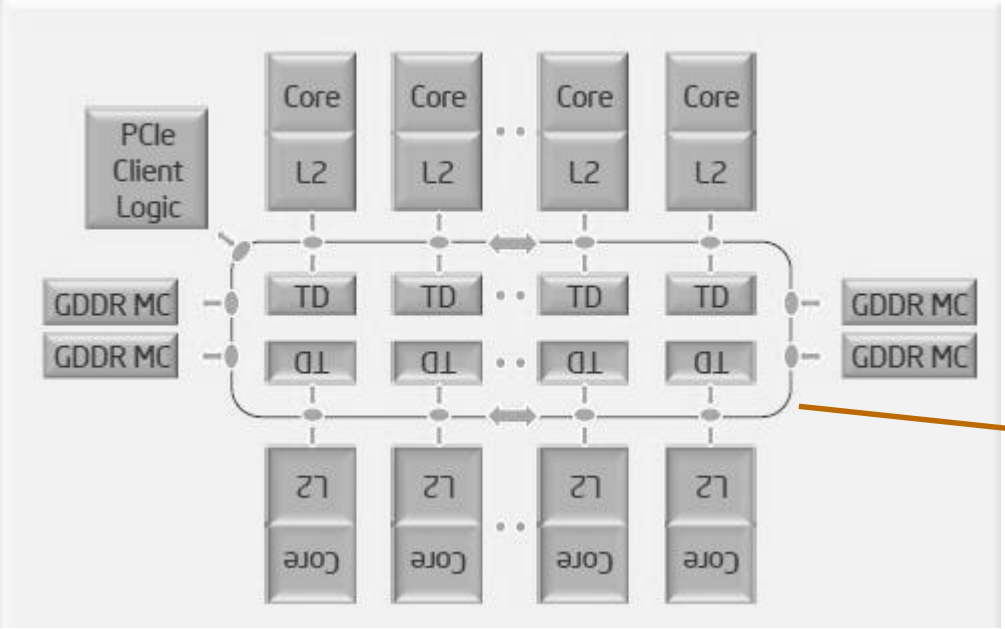


QPI **crossbar** μεταξύ επεξεργαστών και μνημών

Πολλαπλά rings μεταξύ cores και L3 cache

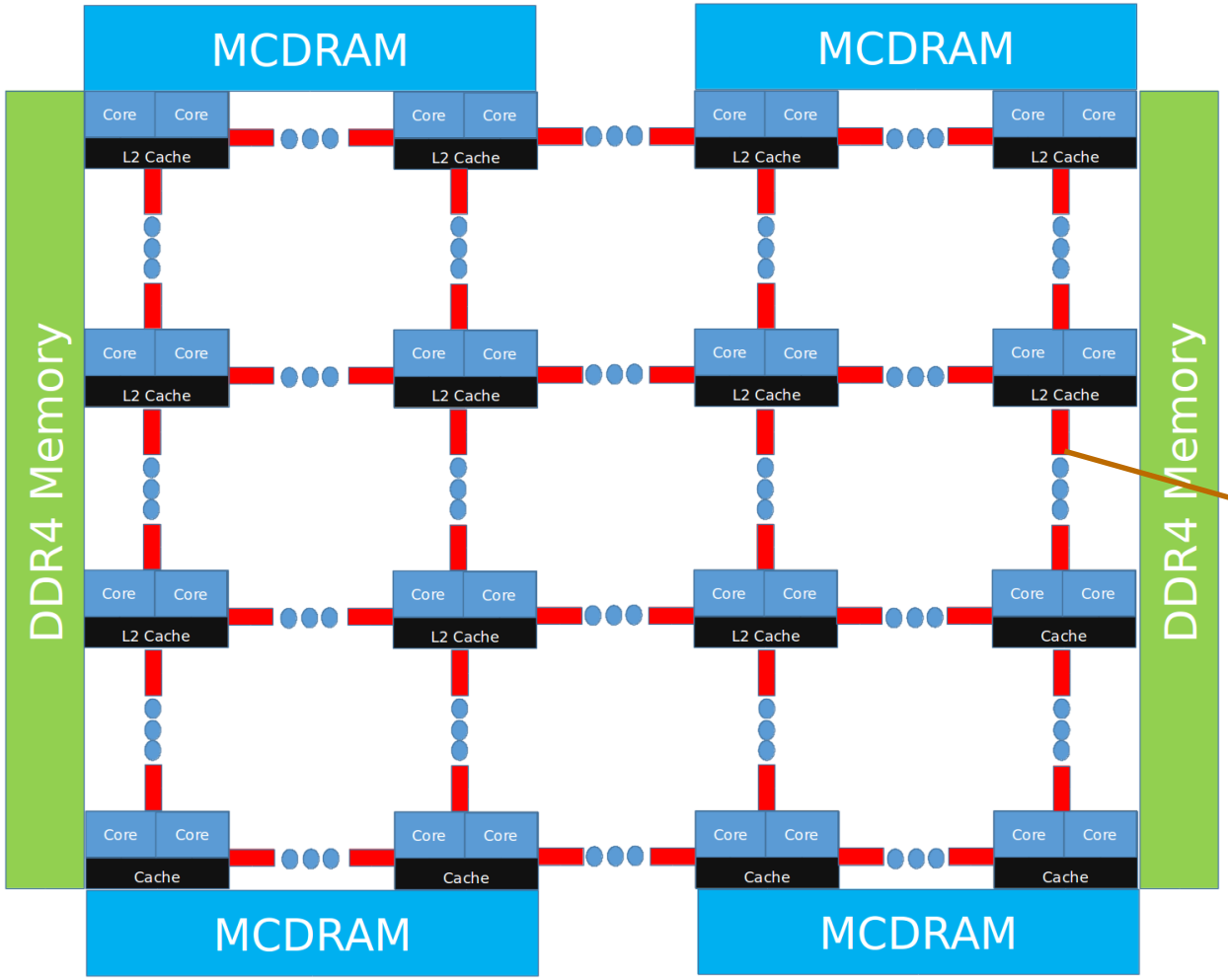
Χρήση δικτύων σε πραγματικά συστήματα

- Συνεπεξεργαστής Intel Xeon Phi Knight's Corner (KNC)



Χρήση δικτύων σε πραγματικά συστήματα

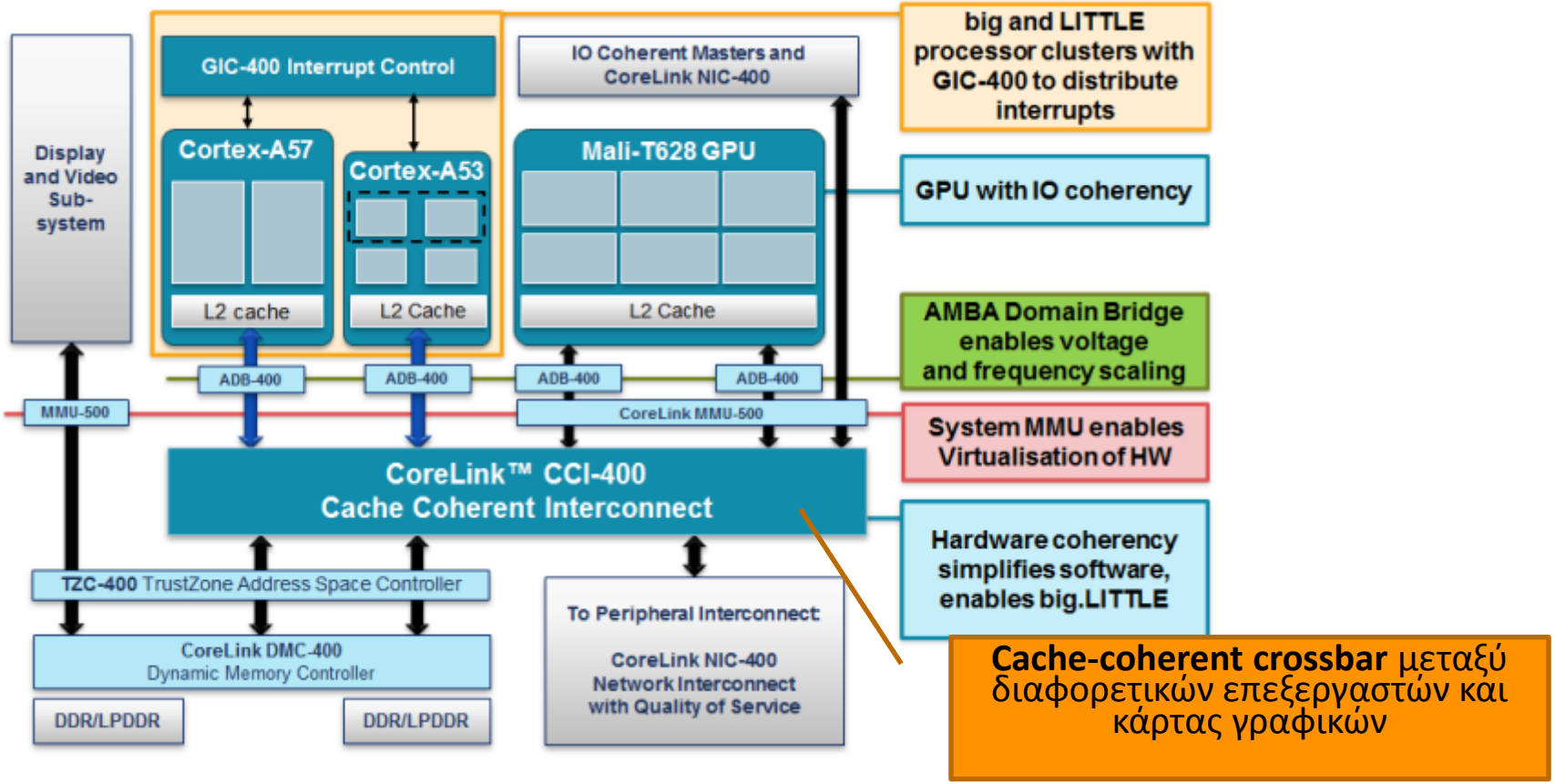
- Συνεπεξεργαστής Intel Xeon Phi Knight's Landing (KNL)



2D-mesh μεταξύ tiles (2 cores ανά tile), MCDRAM (γρήγορες μνήμες) και μνήμης

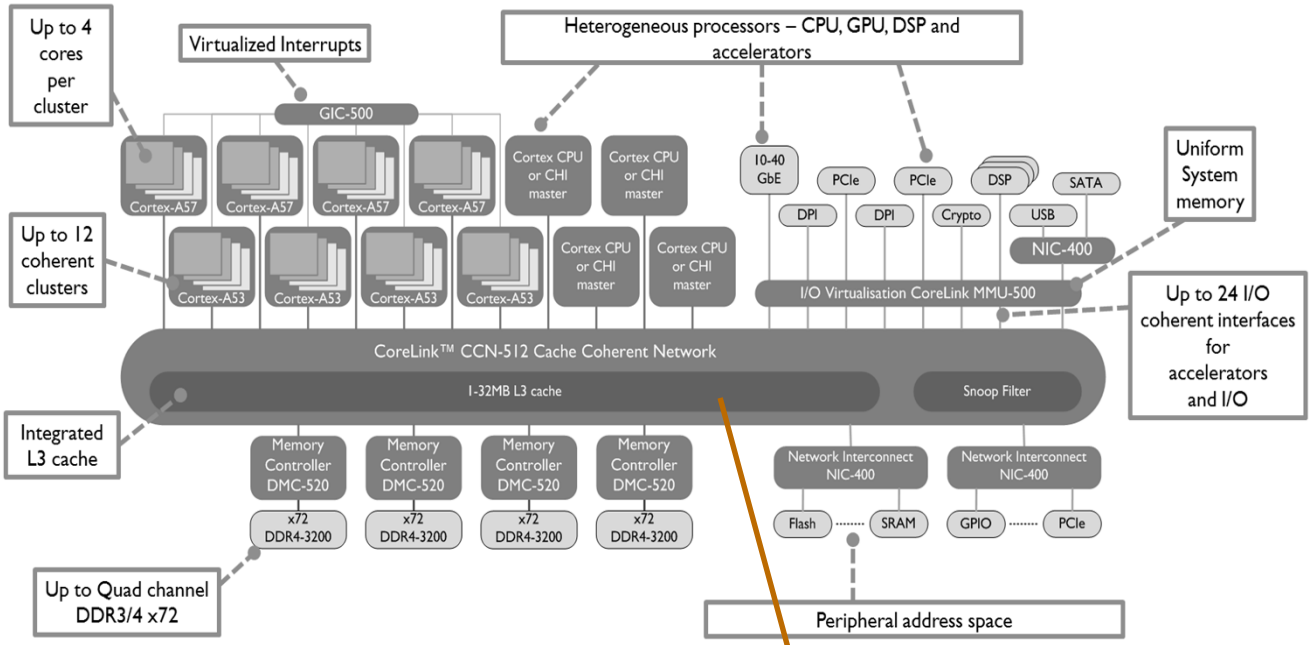
Χρήση δικτύων σε πραγματικά συστήματα

- ARM big.LITTLE - CoreLink CCI400
  - Δίκτυο χαμηλού κόστους και κατανάλωσης ισχύος για mobile πλατφόρμες



Χρήση δικτύων σε πραγματικά συστήματα

- ARM big.LITTLE - CoreLink CCN512
  - Δίκτυο υψηλού κόστους και κατανάλωσης ισχύος για servers



Cache-coherent ring μεταξύ πολλών ετερογενών επεξεργαστών με snoop filter και L3-cache

Τοπολογίες για συστήματα μεγάλης κλίμακας

- Τα συστήματα μεγάλης κλίμακας έχουν διαφορετικές απαιτήσεις από τα δίκτυα διασύνδεσης σε chip
  - Ομοιόμορφη επίδοση, κόστος, **κλιμακωσιμότητα** (σε χιλιάδες κόμβους)
- **Διαδεδομένες τοπολογίες**
- Τόρος
  - Ομοιόμορφη κατανομή της επίδοσης σε σχέση με τα πλέγματα
  - Μικρή διάμετρος σε υψηλές διαστάσεις
- Fat tree
  - Μεγάλο εύρος τομής, μεγάλο κόστος διακοπτών
- Υπερκύβος
  - Μικρός μέσος χρόνος απόκρισης, μικρό κόστος συνδέσμων
- Dragonfly
  - Μεγάλο εύρος τομής, χάρις στα δύο επίπεδα συνδεσμολογίας
- *Περισσότερα στο μάθημα «Συστήματα Παράλληλης Επεξεργασίας»* ☺

Συμπεράσματα

- Τα δίκτυα διασύνδεσης έχουν πολύ σημαντικό ρόλο στην αρχιτεκτονική υπολογιστών
  - Στα παράλληλα συστήματα, κύριος στόχος είναι η μείωση του χρόνου μεταφοράς δεδομένων
- Η σχεδίαση του δικτύου διασύνδεσης γίνεται με στόχο τη μεταφορά της μεγαλύτερης δυνατής πληροφορίας στο μικρότερο δυνατό χρόνο με το μικρότερο δυνατό κόστος και τη μικρότερη δυνατή κατανάλωση ενέργειας, χωρίς να προκαλείται συμφόρηση στο δίκτυο
- Η σχεδίαση του δικτύου διασύνδεσης απαιτεί ολιστική προσέγγιση
  - διεπαφή με συσκευές, τεχνολογία δικτύου, τεχνολογία/διεπαφή συνδέσμων
  - τοπολογία, δρομολόγηση, διαιτησία, έλεγχος ροής
  - εφαρμογές και patterns επικοινωνίας
- Δεν υπάρχει ιδανικό δίκτυο διασύνδεσης για κάθε περίπτωση
  - Απαιτείται ταυτόχρονη αξιολόγηση πολλών μετρικών επίδοσης/κόστους/κατανάλωσης ισχύος
  - Απαιτείται μελέτη των εφαρμογών που θα χρησιμοποιούν το δίκτυο σε κάθε περίπτωση

Πηγές/Βιβλιογραφία

- "Computer Architecture – A quantitative approach", John Hennessy, David Patterson, 5th edition, Morgan Kaufman Publishers, Appendix F
  - http://booksite.mkp.com/9780123838728/references/appendix_f.pdf
- "Principles and Practices of Interconnection Networks", William James Dally, Brian Patrick Towles, Morgan Kaufman Publishers, 2004
- "Interconnection Networks – An engineering approach", José Duato, Sudhakar Yalamanchali, Lionel Ni, Morgan Kaufman Publishers, 2003
- Onur Mutlu, "Interconnects", Computer Architecture - Lecture 21 – ETH, 2017 (slides)
  - https://safari.ethz.ch/architecture/fall2017/lib/exe/fetch.php?media=onur-comparch-fall2017-lecture22-interconnects-ii-afterlecture.pdf