

An Overview of Parallel Architectures

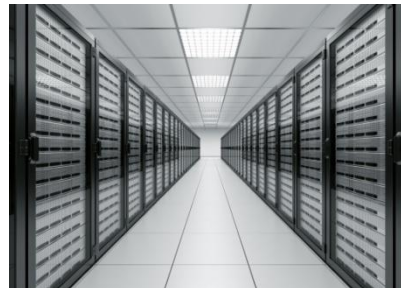
Figures, examples από

1. Αρχιτεκτονική Υπολογιστών, Ποσοτική Προσέγγιση, J.L.Hennessy, A. Patterson
2. An Introduction to the Intel® QuickPath Interconnect:

<http://www.intel.com/content/www/us/en/io/quickpath-technology/quick-path-interconnect-introduction-paper.html>

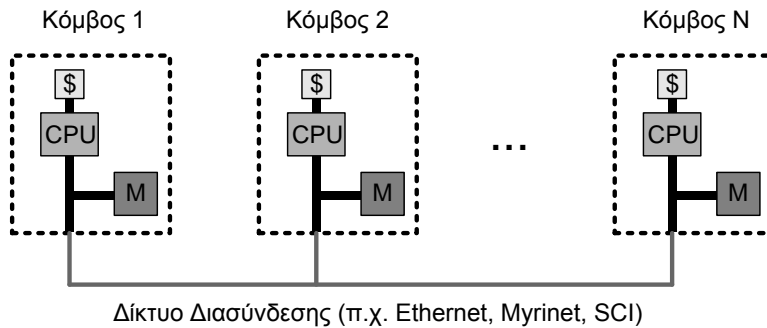
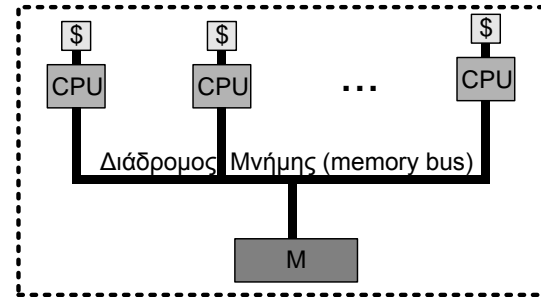
Συστήματα με πολλούς επεξεργαστές: Λίγη αριθμητική

- **2-4 πυρήνες** σε προσωπικούς/φορητούς υπολογιστές και σε κινητά τηλέφωνα
- **Δεκάδες πυρήνες** σε έναν cloud server, σε μία κάρτα γραφικών, σε έναν computation accelerator
- **Εκατοντάδες/Χιλιάδες/Εκατομμύρια(!?) πυρήνες** σε ένα data center, IaaS provider, supercomputer



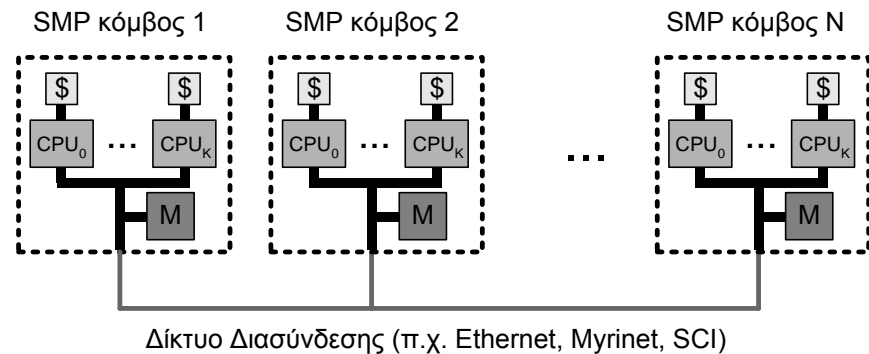
Συστήματα με πολλούς επεξεργαστές: Βασικές αρχιτεκτονικές

Κοινής Μνήμης



Κατανεμημένης Μνήμης

Υβριδική



Συστήματα με πολλούς επεξεργαστές: Ζητήματα

- **Αρχιτεκτονική:**
 - Πώς επηρεάζονται οι ιεραρχίες μνημών;
 - Πώς διασυνδέονται οι επεξεργαστές;
- **Λογισμικό:**
 - Πώς προγραμματίζουμε αυτά τα συστήματα;
 - Λειτουργικό σύστημα: Πώς θα πρέπει να λειτουργεί ο χρονοδρομολογητής;
 - Πώς θα συγχρονίσουμε αποδοτικά πολλαπλά νήματα;

Συστήματα με πολλούς επεξεργαστές: Ζητήματα

■ Αρχιτεκτονική:

- Πώς επηρεάζονται οι ιεραρχίες μνημών; **Προηγμένα Θέματα Αρχιτεκτονικής Υπολογιστών**
- Πώς διασυνδέονται οι επεξεργαστές; **Σημερινό μάθημα και Συστήματα Παράλληλης Επεξεργασίας**

■ Λογισμικό:

- Πώς προγραμματίζουμε αυτά τα συστήματα; **Συστήματα Παράλληλης Επεξεργασίας**
- Λειτουργικό σύστημα: Πώς θα πρέπει να λειτουργεί ο χρονοδρομολογητής; **Συστήματα Παράλληλης Επεξεργασίας**
- Πώς θα συγχρονίσουμε αποδοτικά πολλαπλά νήματα; **Συστήματα Παράλληλης Επεξεργασίας**

Μια ματιά στα συστήματα μεγάλης κλίμακας Υπερυπολογιστές

- Τεράστια συστήματα με χιλιάδες/εκατομμύρια επεξεργαστές
- Χρησιμοποιούνται για επιστημονικές εφαρμογές
 - Life sciences
 - Earth Sciences
 - Engineering
 - Etc.
- Top500:
 - 2 φορές το χρόνο ανακοινώνεται η λίστα με τους 500 ισχυρότερους επεξεργαστές
 - » Supercomputing, Νοέμβριος, <http://sc13.supercomputing.org/> (ΗΠΑ)
 - » ISC, Ιούνιος, <http://www.isc-events.com/isc13/> (Γερμανία)

Supercomputers

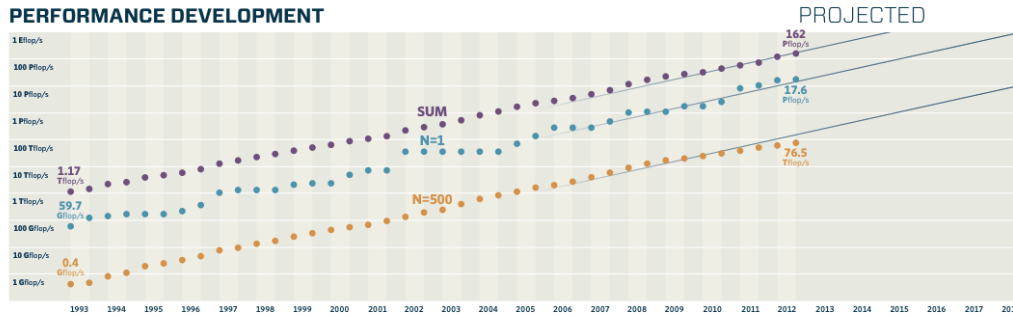


rendered on SuperMUC by LRZ

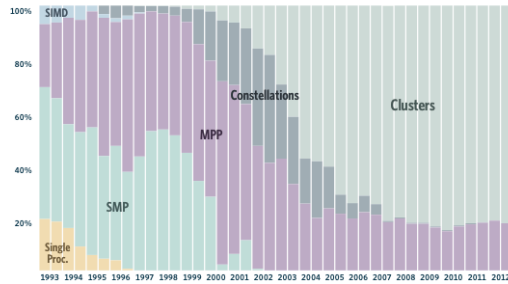


	NAME	SPECS	SITE	COUNTRY	CORES	R _{MAX} PFLOP/S	POWER MW
1	TITAN	Cray XK7, Operon 6274 16C 2.2 GHz + Nvidia Kepler GPU, Custom interconnect	DOE/OS/ORNL	USA	560,640	17.6	8.3
2	SEQUOIA	IBM BlueGene/Q, Power BQC 16C 1.60 GHz, Custom interconnect	DOE/NNSA/LLNL	USA	1,572,864	16.3	7.9
3	K COMPUTER	Fujitsu SPARC64 VIIIfx 2.0GHz, Custom interconnect	RIKEN AICS	Japan	705,024	10.5	12.7
4	MIRA	IBM BlueGene/Q, Power BQC 16C 1.60 GHz, Custom interconnect	DOE/OS/ANL	USA	786,432	8.16	3.95
5	JUQUEEN	IBM BlueGene/Q, Power BQC 16C 1.60 GHz, Custom interconnect	Forschungszentrum Jülich	Germany	393,216	4.14	1.97

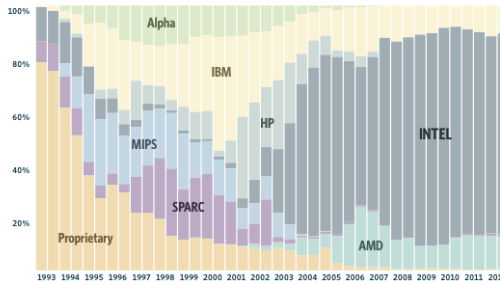
PERFORMANCE DEVELOPMENT



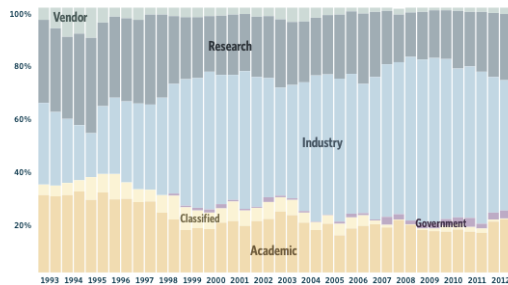
ARCHITECTURES



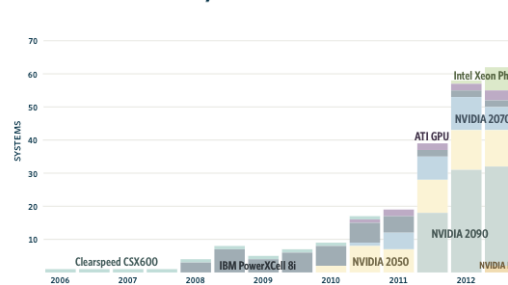
CHIP TECHNOLOGY



INSTALLATION TYPE



ACCELERATORS/CO-PROCESSORS



Top 500 (November 2012 list)

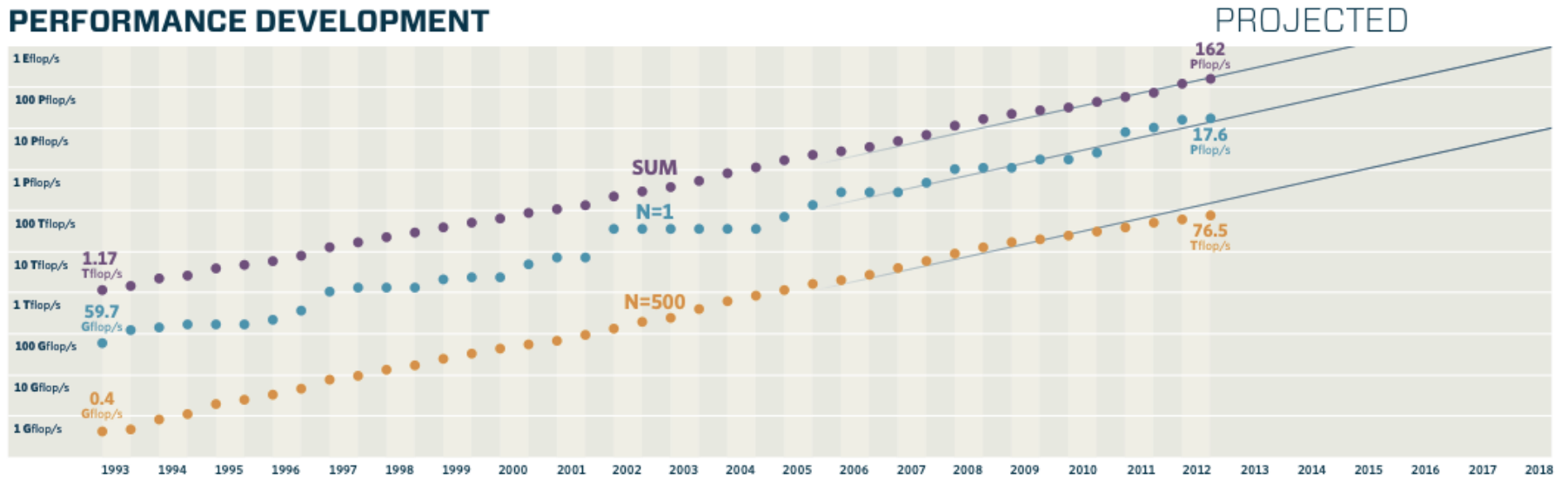
Top 5

	NAME	SPECS	SITE	COUNTRY	CORES	R _{MAX} PFLOP/S	POWER MW
1	TITAN	Cray XK7, Operon 6274 16C 2.2 GHz + Nvidia Kepler GPU, Custom interconnect	DOE/OS/ORNL	USA	560,640	17.6	8.3
2	SEQUOIA	IBM BlueGene/Q, Power BQC 16C 1.60 GHz, Custom interconnect	DOE/NNSA/LLNL	USA	1,572,864	16.3	7.9
3	K COMPUTER	Fujitsu SPARC64 VIIIfx 2.0GHz, Custom interconnect	RIKEN AICS	Japan	705,024	10.5	12.7
4	MIRA	IBM BlueGene/Q, Power BQC 16C 1.60 GHz, Custom interconnect	DOE/OS/ANL	USA	786,432	8.16	3.95
5	JUQUEEN	IBM BlueGene/Q, Power BQC 16C 1.60 GHz, Custom interconnect	Forschungszentrum Jülich	Germany	393,216	4.14	1.97

Top 500 (November 2012 list)

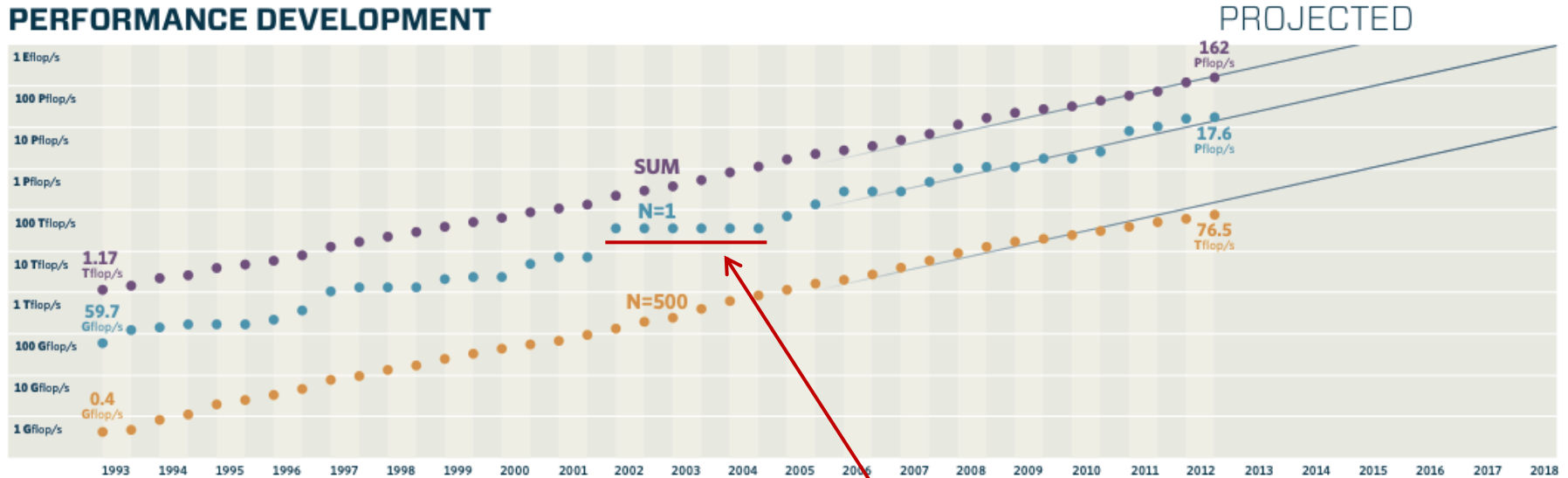
Performance development

PERFORMANCE DEVELOPMENT



Top 500 (November 2012 list)

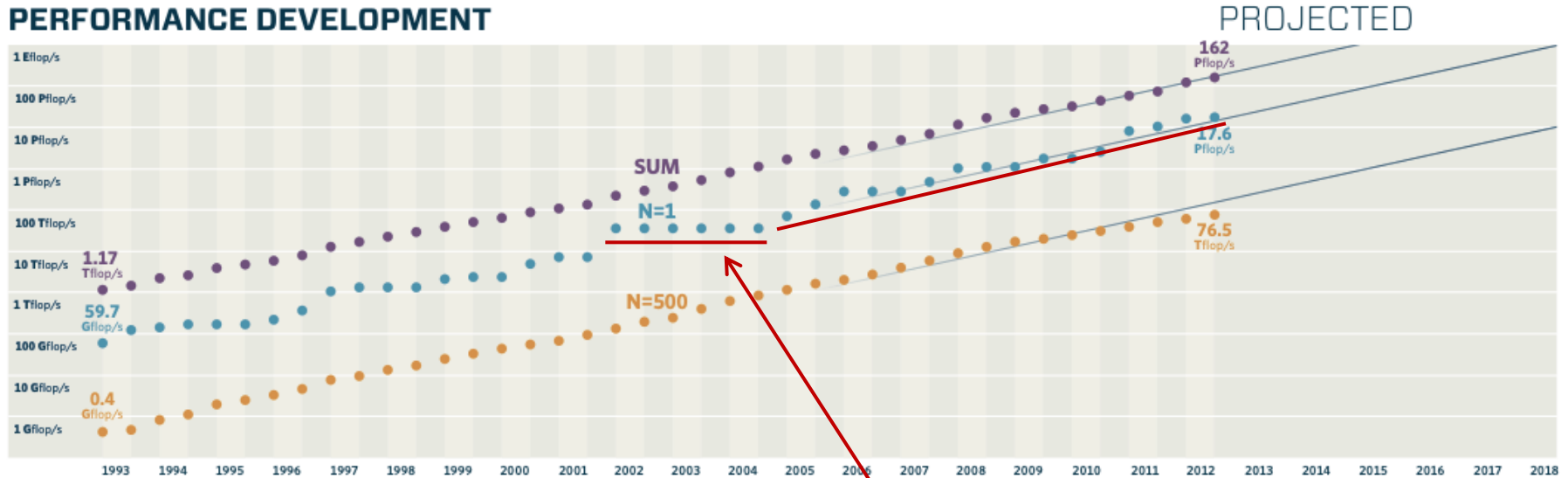
Performance development



3 years in Top1!
Earth Simulator
Japan

Top 500 (November 2012 list)

Performance development

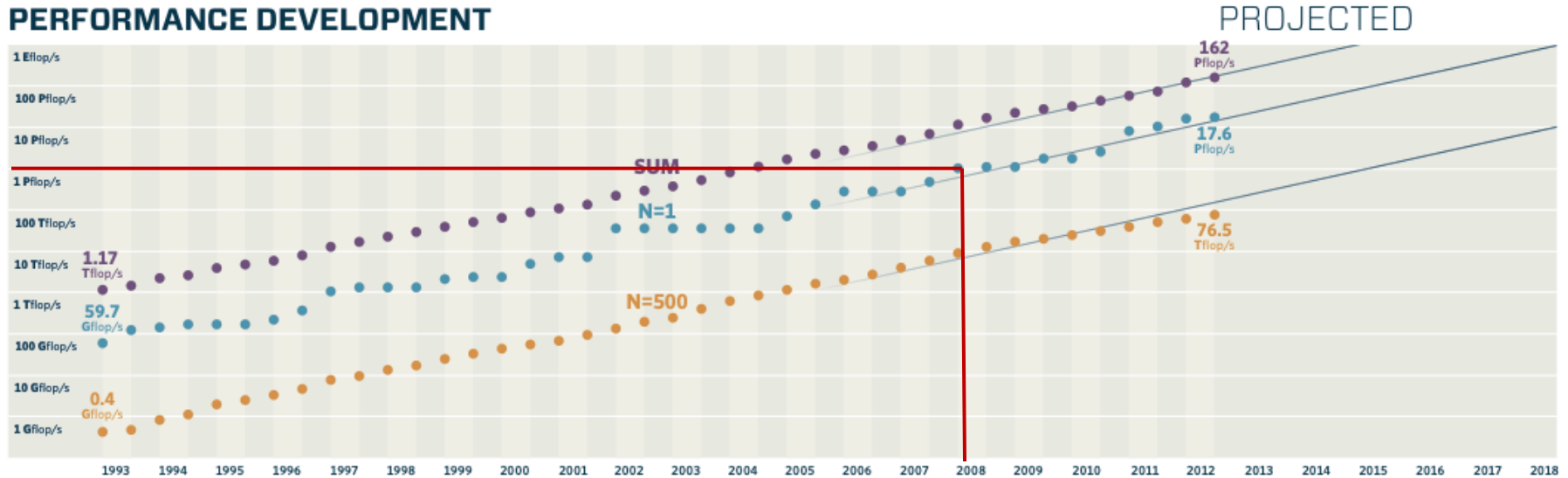


3 years in Top1!
Earth Simulator
Japan

The race begins!

Top 500 (November 2012 list)

Performance development



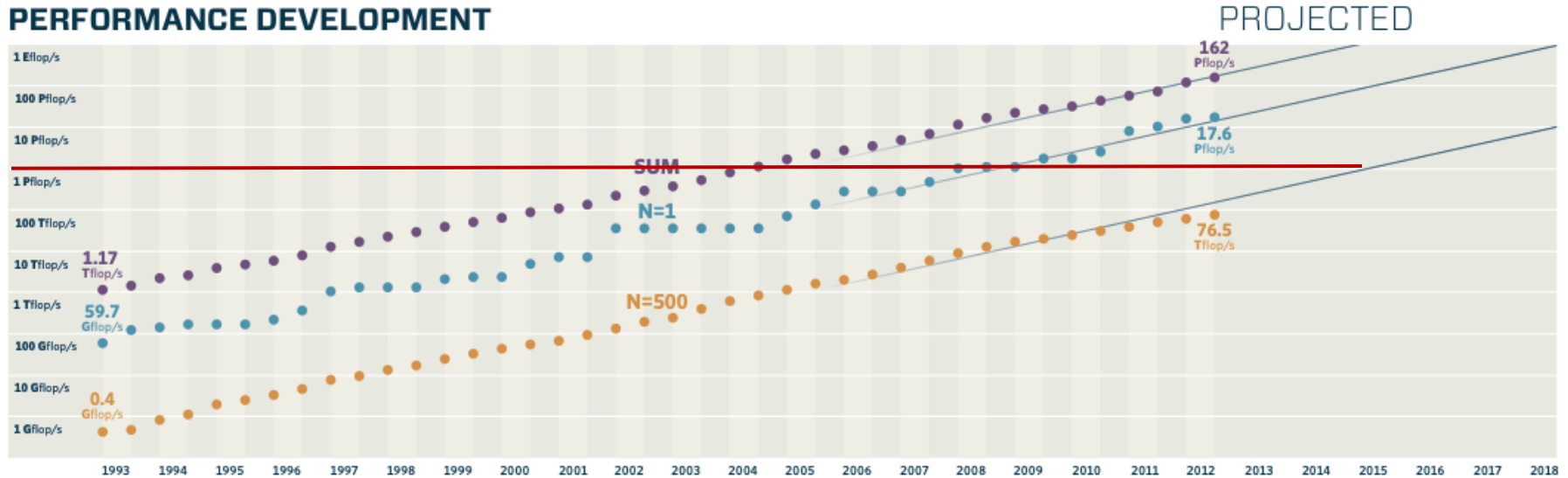
Petaflop barrier

Roadrunner, Los Alamos National
Laboratory, USA

Decommissioned 31st March 2013 ☹

Top 500 (November 2012 list)

Performance development



Why?

Roadrunner is still high in Top500 (rank 22!)

Top 500 (November 2012 list)

Performance development

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
21	Information Technology Center, The University of Tokyo Japan	Oakleaf-FX - PRIMEHPC FX10, SPARC64 IXfx 16C 1.848GHz, Tofu interconnect Fujitsu	76800	1043.0	1135.4	1177
22	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband IBM	122400	1042.0	1375.8	2345
23	University of Edinburgh United Kingdom	DIRAC - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	98304	1035.3	1258.3	493

Top 500 (November 2012 list)

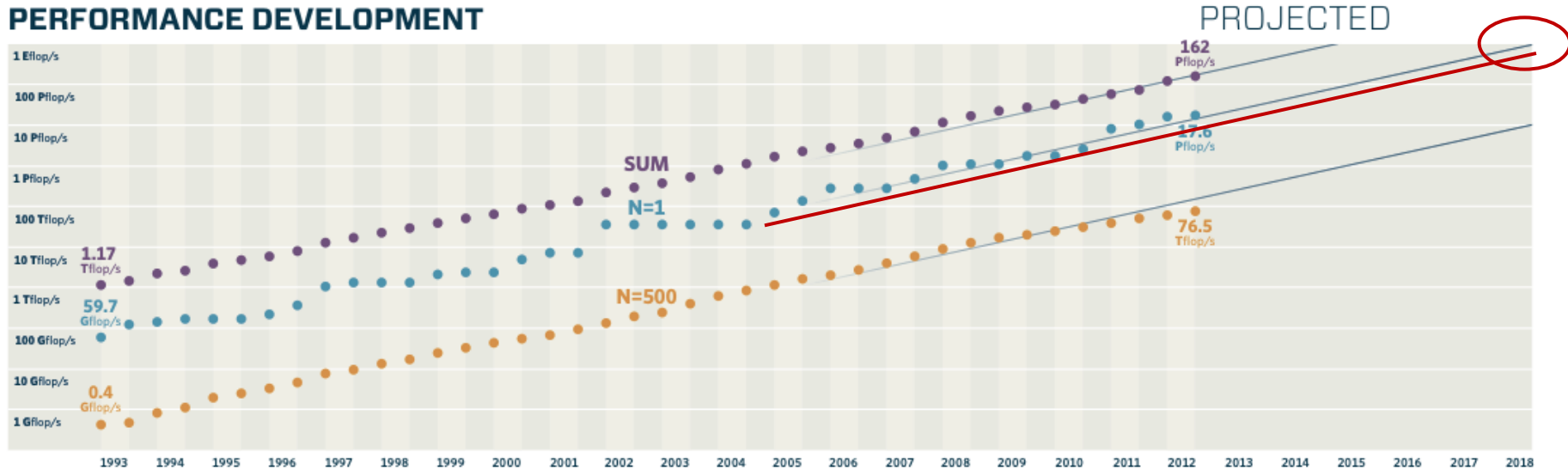
Performance development

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
21	Information Technology Center, The University of Tokyo Japan	Oakleaf-FX - PRIMEHPC FX10, SPARC64 IXfx 16C 1.848GHz, Tofu interconnect Fujitsu	76800	1043.0	1135.4	1177
22	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband IBM	122400	1042.0	1375.8	2345
23	University of Edinburgh United Kingdom	DIRAC - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	98304	1035.3	1258.3	493

**Low power efficiency!
(Flop/Watt)**

Top 500 (November 2012 list)

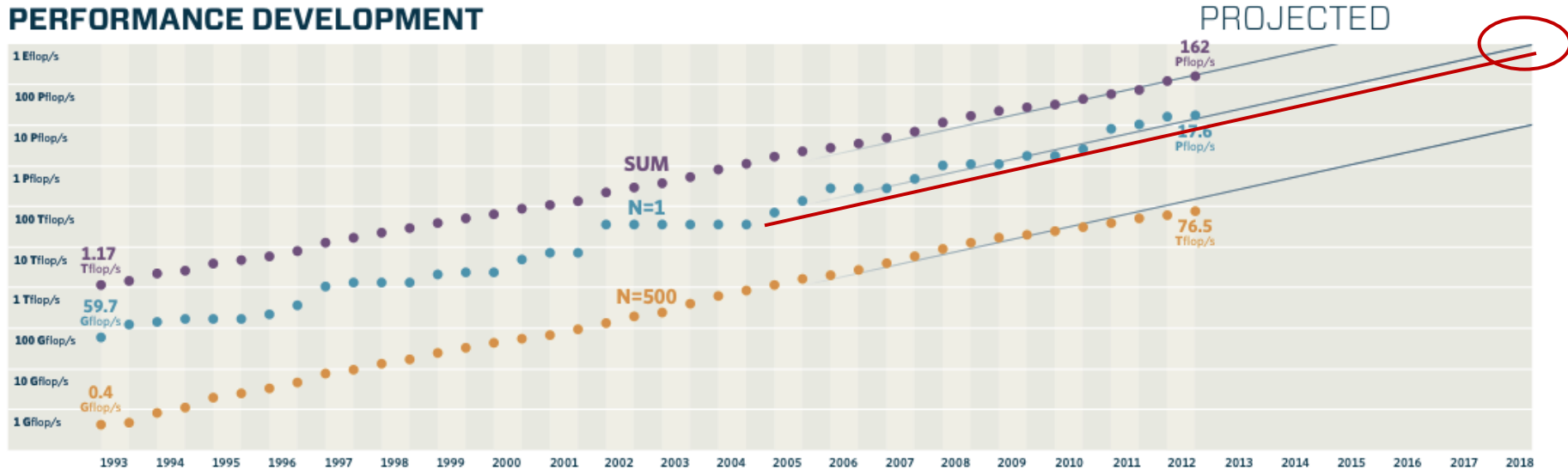
Performance development



Shall we reach “Exaflop computing” by the end of this decade?

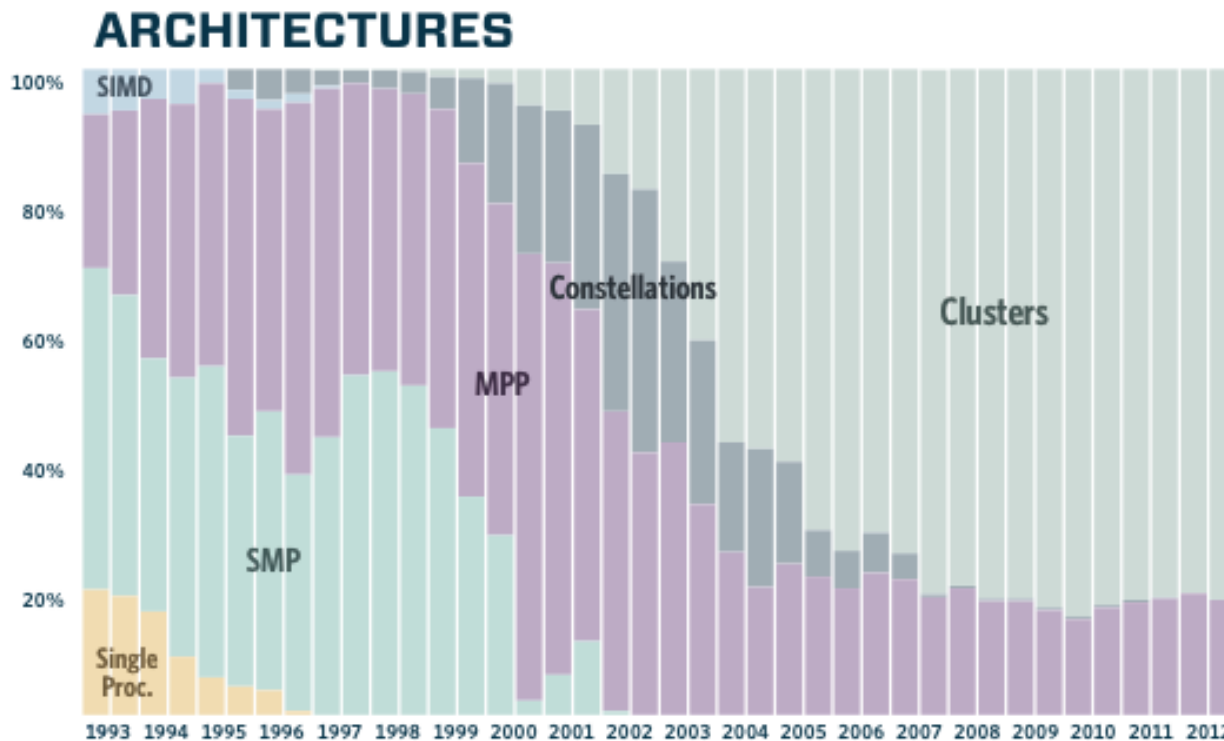
Top 500 (November 2012 list)

Performance development

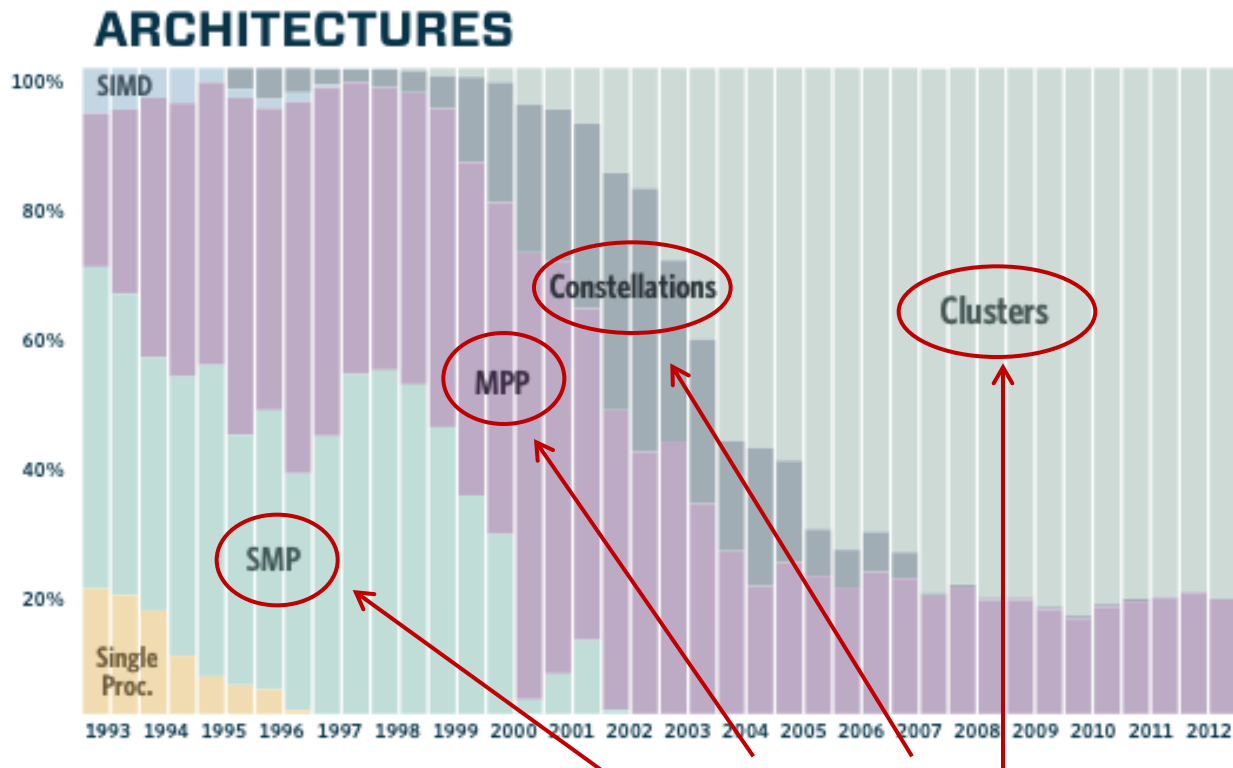


Two major problems:
a) power
b) data movements

Top 500 (November 2012 list) Architecture



Top 500 (November 2012 list) Architecture



What are these?

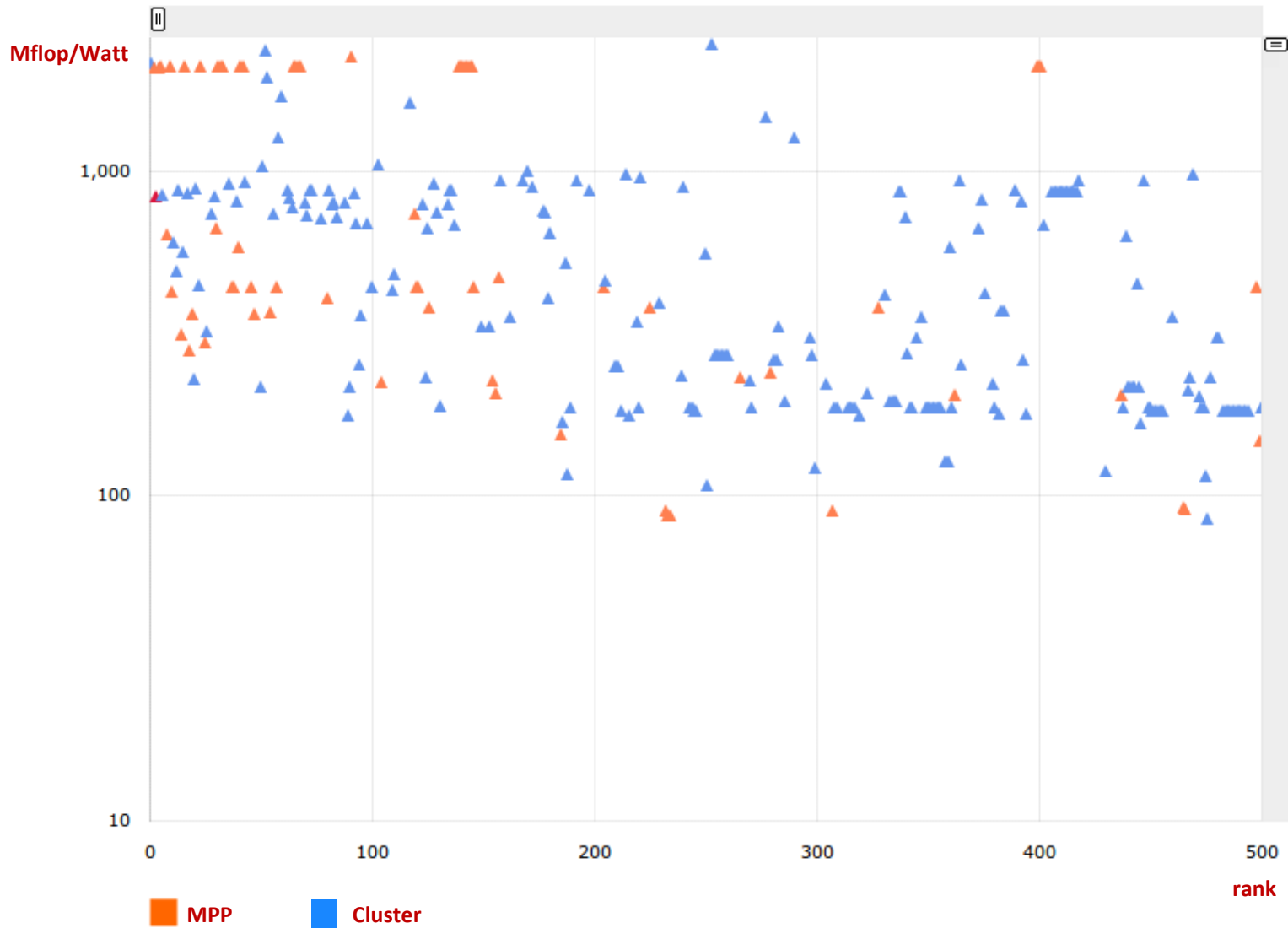
Top 500 (November 2012 list)

Architecture

- **SMP:** Symmetric Multiprocessor
- **Constellation:** Συλλογή από διαφορετικά συστήματα
- **MPP:** Massively Parallel Processing
 - Μαζικά παράλληλα συστήματα
 - Βασίζονται σε ειδικά κατασκευασμένα (custom made) στοιχεία
 - » Δίκτυα διασύνδεσης (κατά κύριο λόγο)
 - » Επεξεργαστικές μονάδες
 - Π.χ. Blue Gene/Q, Cray XK7
 - Καταλαμβάνουν τις υψηλότερες θέσεις του Top500
 - Χαμηλότερη κατανάλωση ενέργειας
 - Υψηλότερο κόστος
- **Clusters:** Συστοιχίες συστημάτων
 - Όλα τα στοιχεία τους είναι ήδη εμπορικά διαθέσιμα
 - Δίκτυα διασύνδεσης: Infiniband, 10G Ethernet, Gbit Ethernet

Top 500 (November 2012 list)

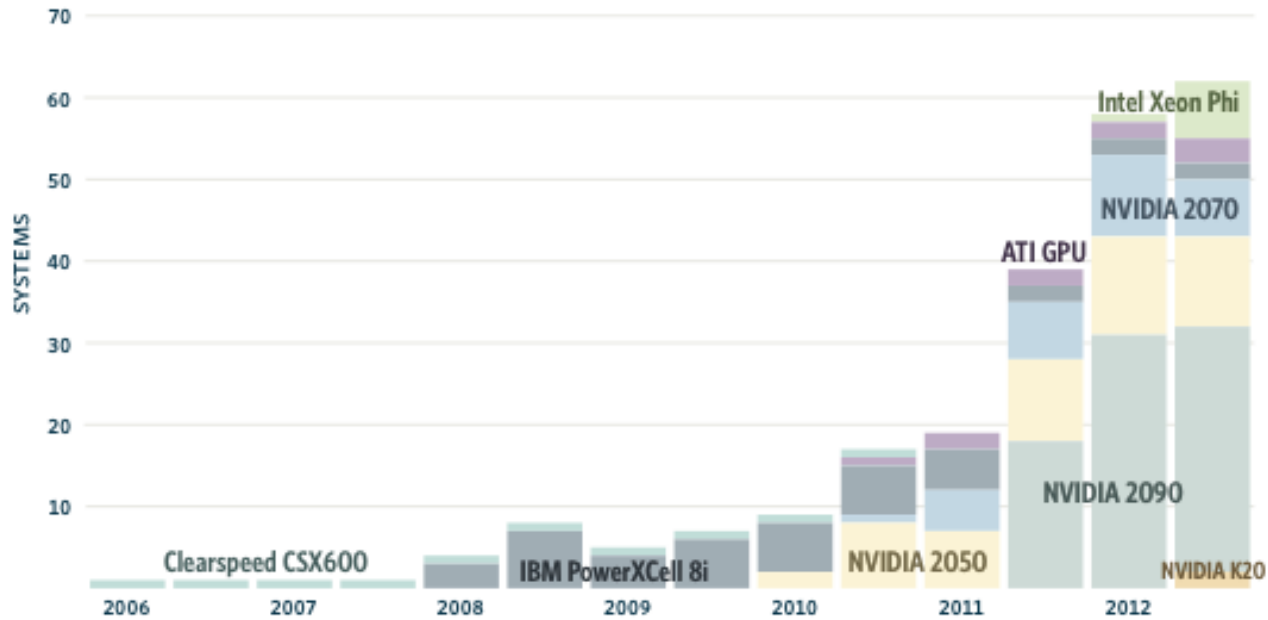
Power efficiency



Top 500 (November 2012 list)

Architecture – The accelerator trend

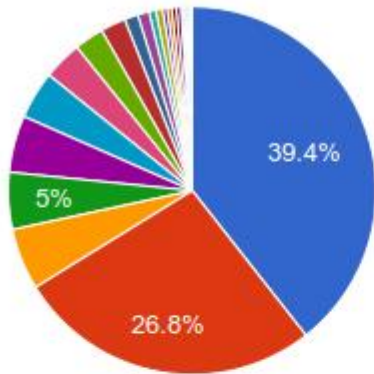
ACCELERATORS/CO-PROCESSORS



Top 500 (November 2012 list)

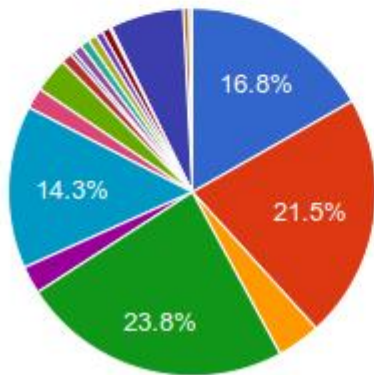
Processor Architecture

Processor Generation System Share



▲ 1/4 ▼

Processor Generation Performance Share

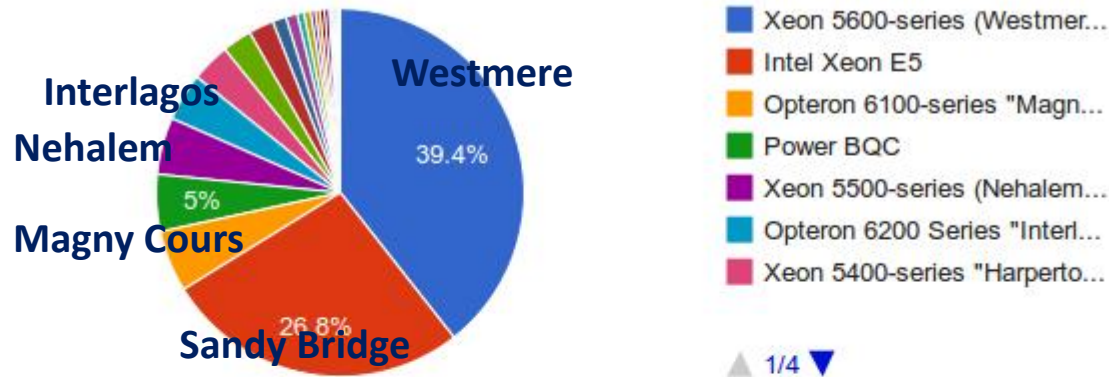


▲ 1/3 ▼

Top 500 (November 2012 list)

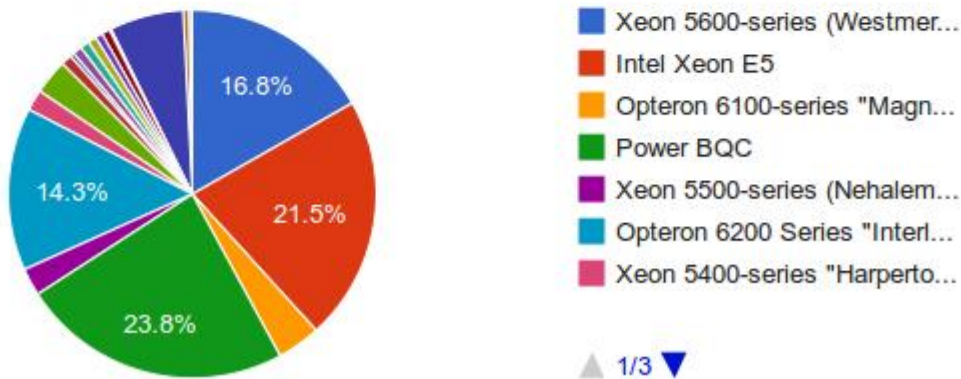
Processor Architecture

Processor Generation System Share



Commodity processors

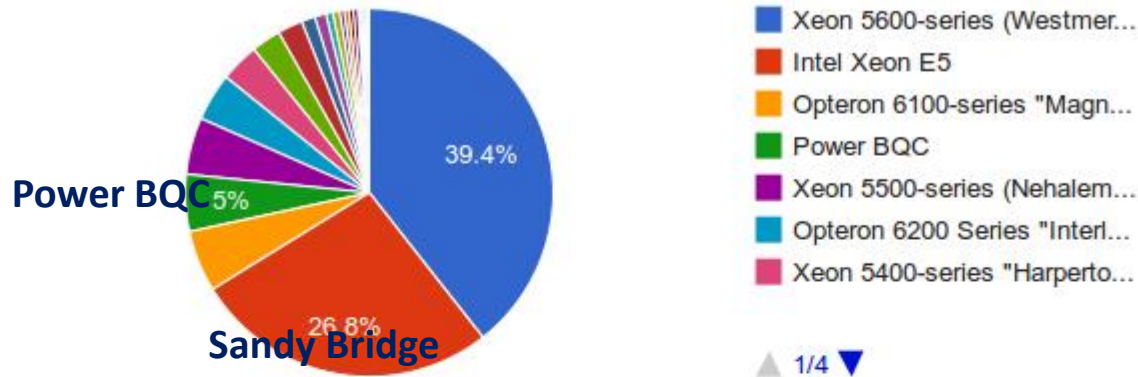
Processor Generation Performance Share



Top 500 (November 2012 list)

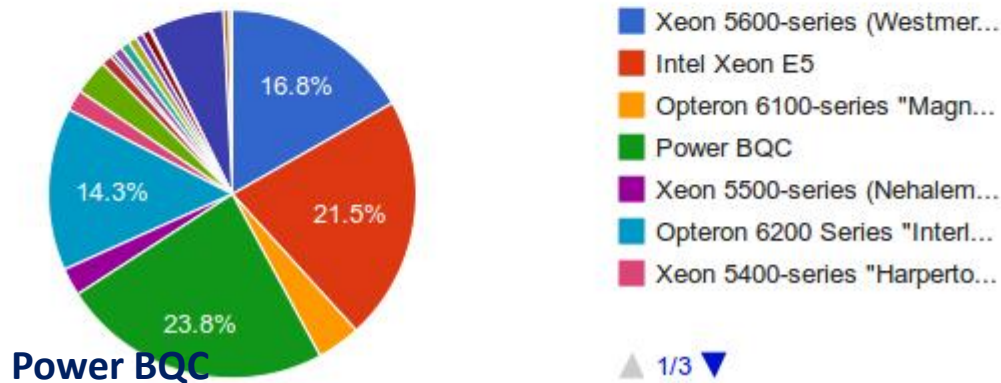
Processor Architecture

Processor Generation System Share

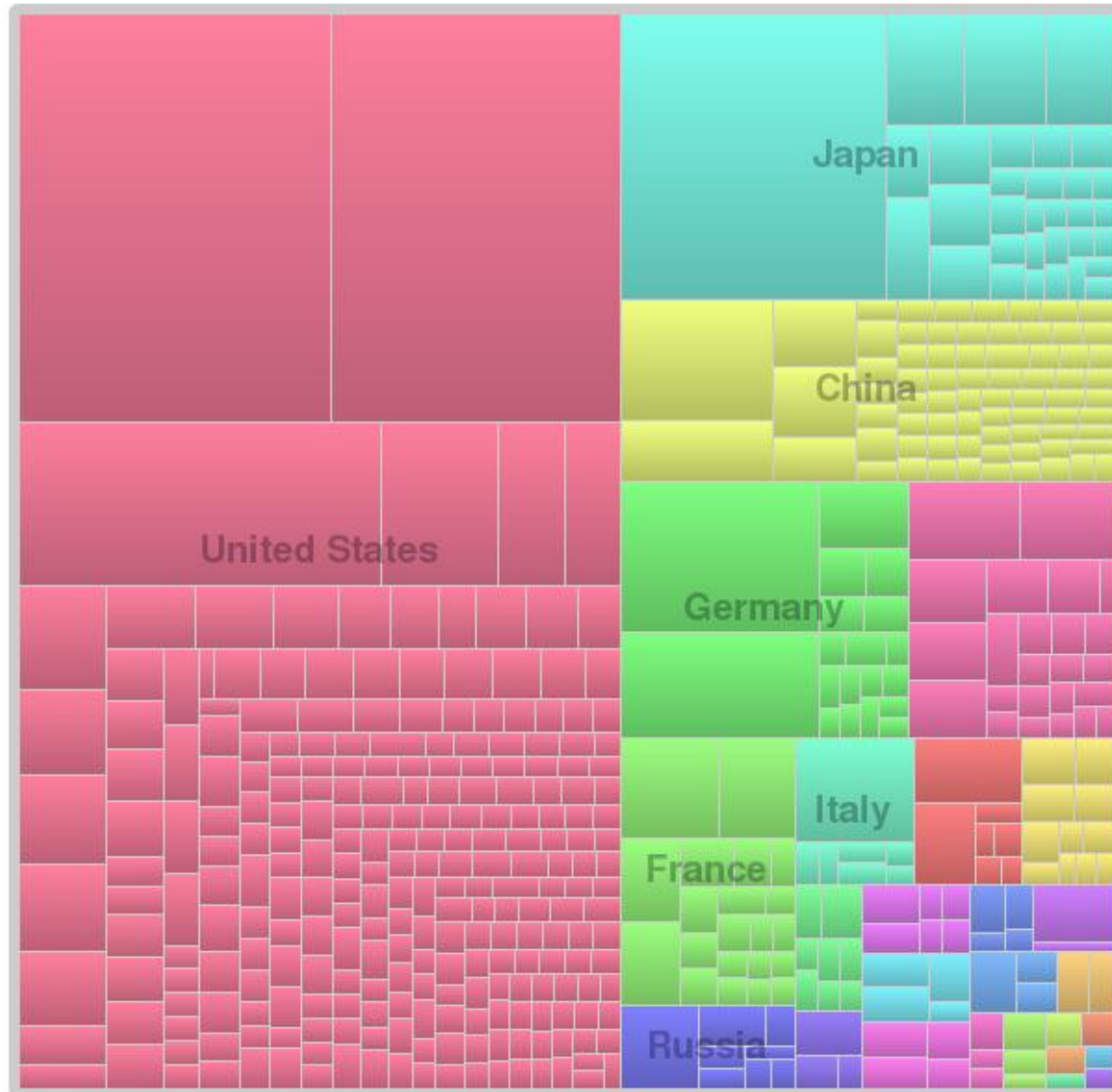


Custom processors

Processor Generation Performance Share



Top 500 (November 2012 list) Countries



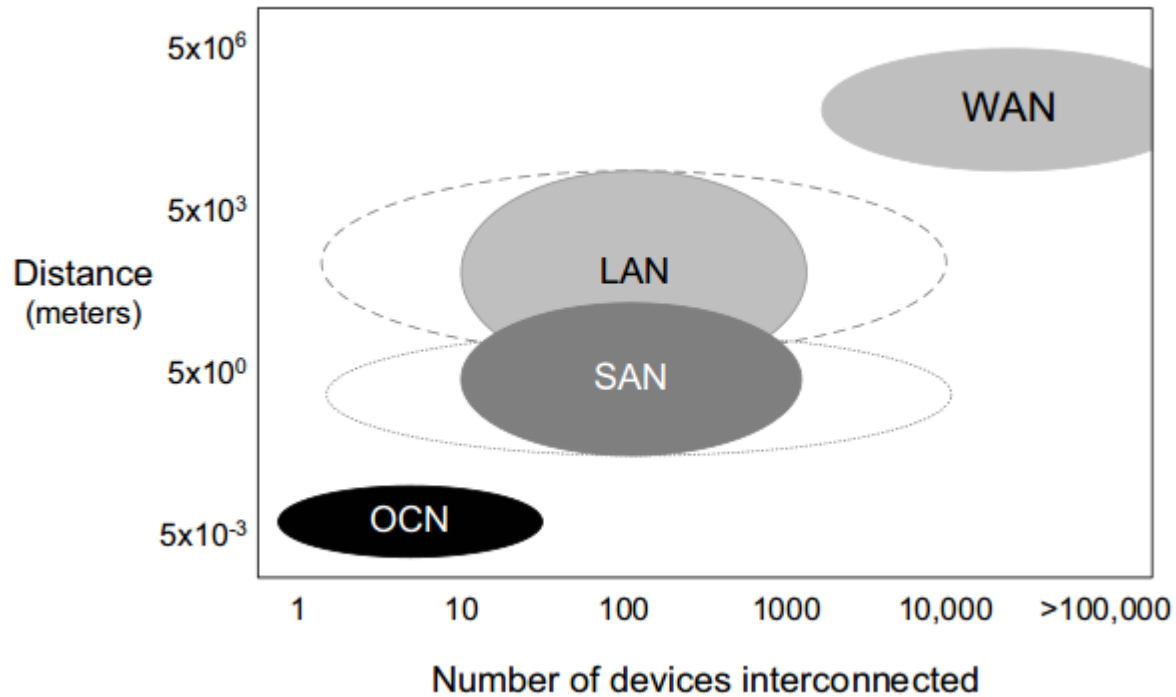
Δίκτυα διασύνδεσης

- Διασυνδέουν δομικές μονάδες ενός σύνθετου συστήματος
- **On-Chip Network (OCN) or Network-on-Chip (NoC):**
 - Caches
 - Processing cores
 - CMPs.
- **System/Storage Area Networks (SAN):**
 - Επεξεργαστές με μονάδες μνήμης
 - Υπολογιστές μεταξύ τους
 - Υπολογιστές με συσκευές αποθήκευσης
- **Local Area Networks (LAN):**
 - Υπολογιστές σε ένα τοπικό δίκτυο
- **Wide Area Networks (WAN):**
 - Υπολογιστές σε οποιοδήποτε σημείο του πλανήτη

Δίκτυα διασύνδεσης

- Διασυνδέουν δομικές μονάδες ενός σύνθετου συστήματος
- **On-Chip Network (OCN) or Network-on-Chip (NoC):**
 - Caches
 - Processing cores
 - CMPs.
- **System/Storage Area Networks (SAN):**
 - Επεξεργαστές με μονάδες μνήμης
 - Υπολογιστές μεταξύ τους
 - Υπολογιστές με συσκευές αποθήκευσης
- **Local Area Networks (LAN):**
 - Υπολογιστές σε ένα τοπικό δίκτυο
- **Wide Area Networks (WAN):**
 - Υπολογιστές σε οποιοδήποτε σημείο του πλανήτη

Δίκτυα διασύνδεσης



Κρίσιμες μετρικές για την αξιολόγηση ενός δικτύου διασύνδεσης

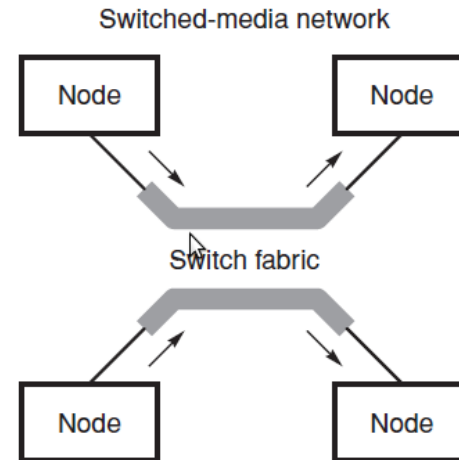
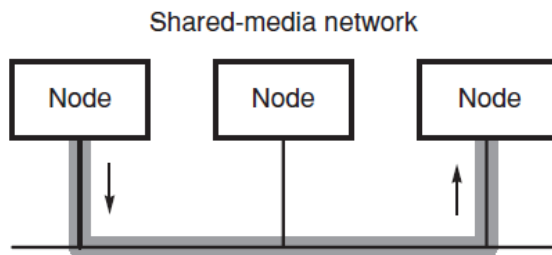
- **Επίδοση:**
 - **Latency:** Χρόνος που απαιτείται για να φτάσει το πρώτο byte πληροφορίας από τον αποστολέα στον παραλήπτη
 - **Bandwidth:** Ο ρυθμός με τον οποίο μεταδίδεται η πληροφορία
- **Κόστος:**
 - Αριθμός ports στα switches
 - Αριθμός switches
 - Αριθμός συνδέσεων
- **Επεκτασιμότητα:** Η δυνατότητα του δικτύου να υποστηρίξει επέκταση σε μεγαλύτερο αριθμό διασυνδεόμενων μονάδων

Χαρακτηριστικά συνδεσμολογιών

- **Βαθμός κόμβου (node degree) d :** αριθμός συνδέσμων σε ένα κόμβο
 - πρέπει να είναι
 - » μικρός (λόγω κόστους)
 - » σταθερός (για επεκτασιμότητα)
- **Διάμετρος δικτύου D :** μέγιστο ελάχιστο μονοπάτι μεταξύ δύο οποιωνδήποτε κόμβων
 - Όσο μικρότερη, τόσο καλύτερη η χειρότερη περίπτωση επικοινωνίας
- **Εύρος τομής (bisection width) b :** ο ελάχιστος αριθμός ακμών που κόβουμε, χωρίζοντας το δίκτυο στα δύο
 - Αποτελεί ένα καλό δείκτη του μέγιστου εύρους ζώνης επικοινωνίας σε ένα δίκτυο

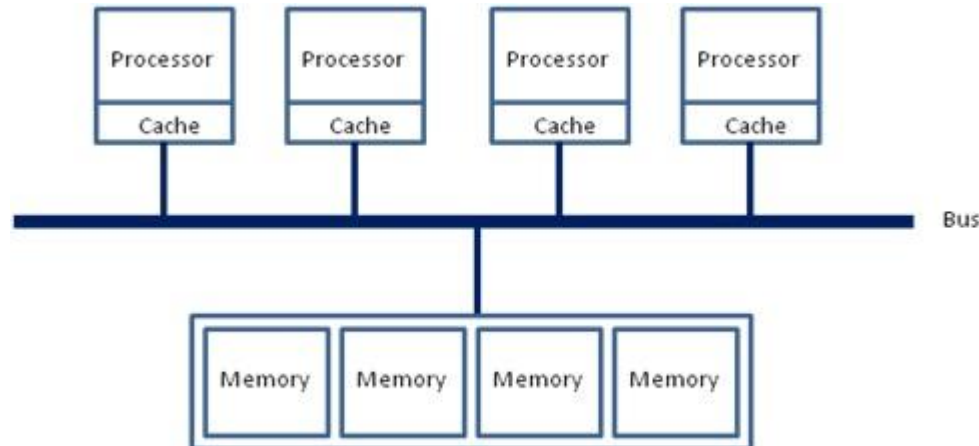
Κατηγορίες δικτύων

- **Shared-media networks:** Το μέσο είναι διαμοιραζόμενο από όλους τους κόμβους, π.χ.
 - Δίαυλος (bus) σε μονοεπεξεργαστικά και πολυεπεξεργαστικά συστήματα
 - Το παραδοσιακό Ethernet
- **Switched-media networks:** Υπάρχουν διακοπτόμενα μονοπάτια που μπορούν να υποστηρίξουν την ταυτόχρονη επικοινωνία ανάμεσα σε διαφορετικά ζεύγη κόμβων

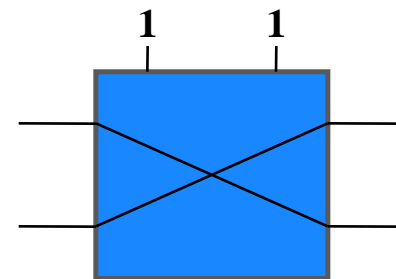
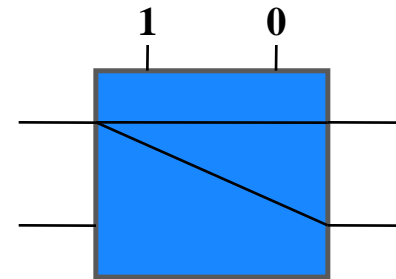
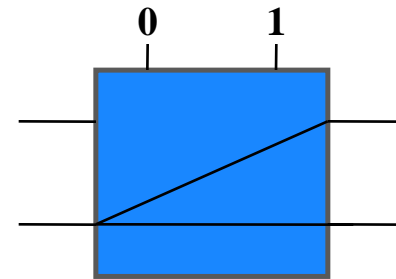
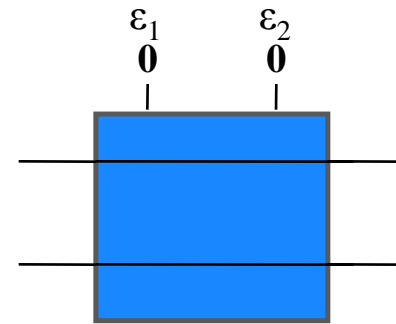
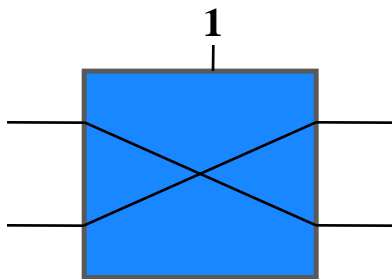
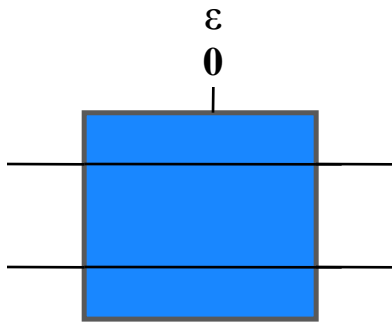


Διάδρομος (Bus)

- Παραδοσιακός τρόπος διασύνδεσης σε ένα NoC
- Απλή υλοποίηση με χαμηλό κόστος
 - Data, address, control buses
 - Διαιτησία (Arbitration)
- Υποστηρίζει εύκολα broadcast και multicast
- Εύκολη υλοποίηση cache coherence με snooping
- **Αλλά:** δεν είναι επεκτάσιμος (τυπικά λίγες δεκάδες στοιχεία)
 - Περιορισμένο συνολικό bandwidth
 - Δυσκολία στη διαιτησία



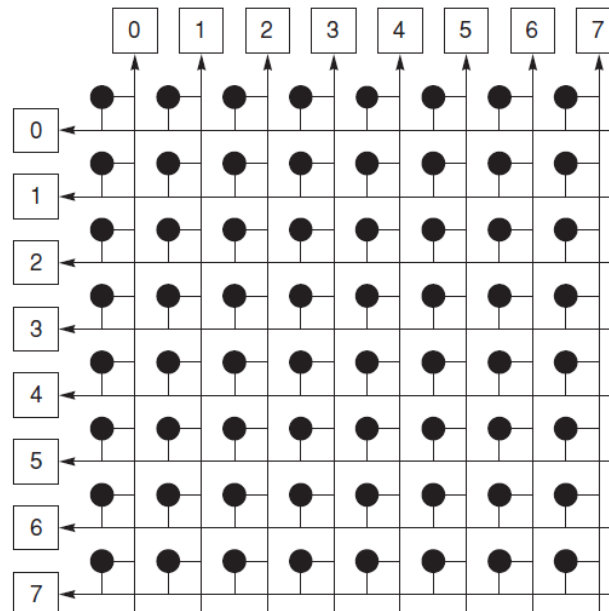
Διακόπτες



Centralized Switched Networks

Crossbar Switch

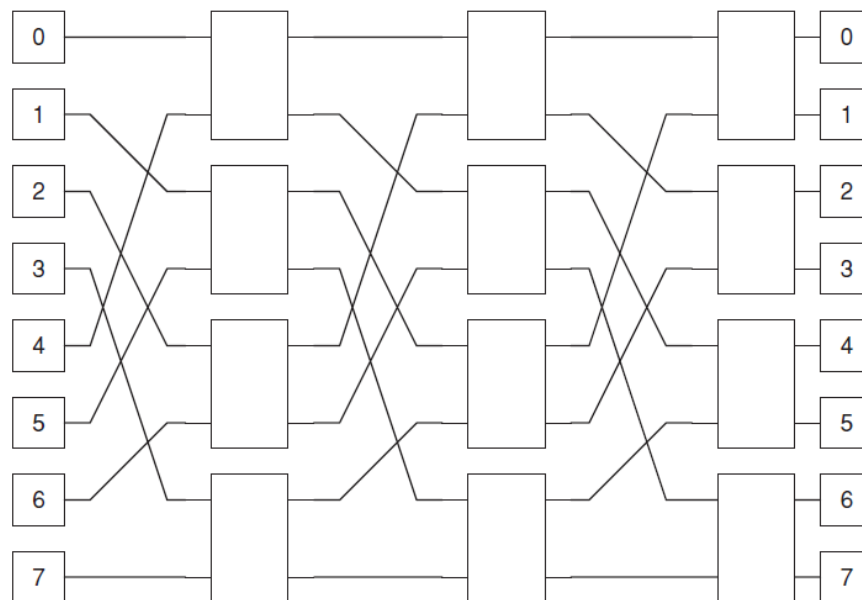
- Απλούστερη, ταχύτερη αλλά και ακριβότερη λύση για τη διασύνδεση N στοιχείων
- Απαιτεί N^2 διακόπτες, δεν κλιμακώνει λόγω κόστους
- Χρησιμοποιείται σε NoC για τη διασύνδεση λίγων δεκάδων στοιχείων



Centralized Switched Networks

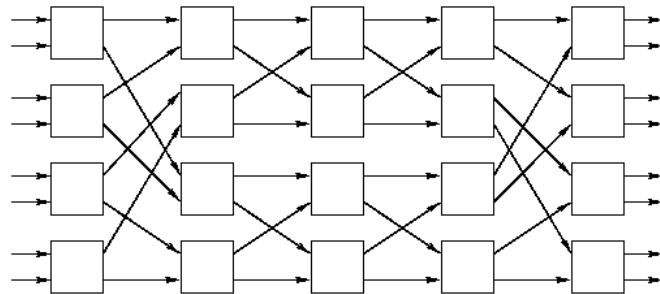
Multistage Interconnection Networks

- Διασυνδέουν N στοιχεία με τη χρήση πολυεπίπεδων διακοπών
- Αν χρησιμοποιηθούν $k \times k$ διακόπτες, χρειάζονται $\log_k N$ στάδια με N/k διακόπτες ανά στάδιο (σύνολο $N/k \log_k N$ διακόπτες)
- Ανάλογα με τη διασύνδεση των διακοπών έχουν προκύψει διαφορετικά δίκτυα που ανταποκρίνονται σε διαφορετικά patterns επικοινωνίας

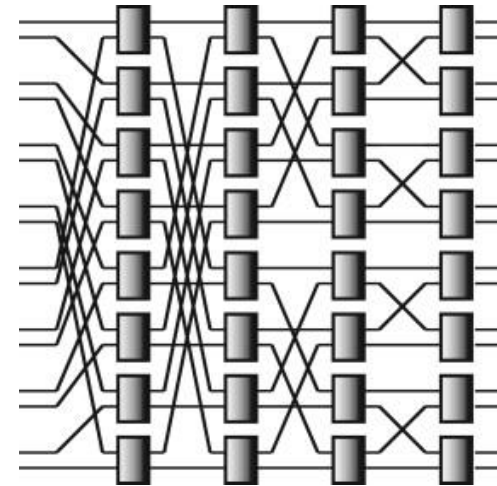


Centralized Switched Networks

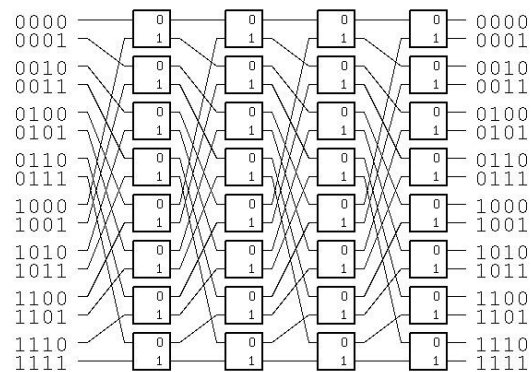
Multistage networks



Benes network



Butterfly network

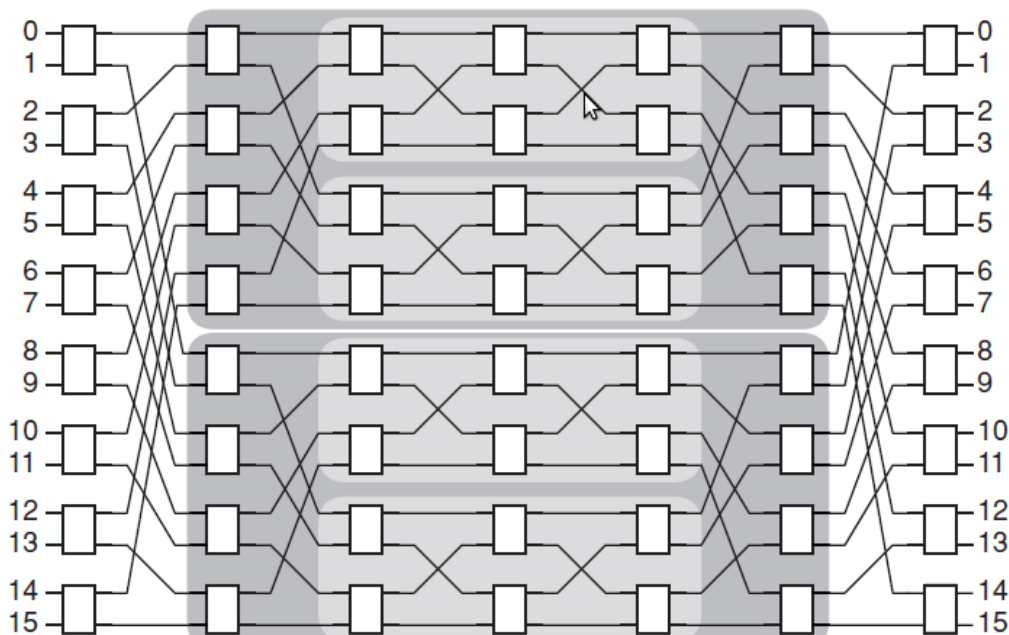
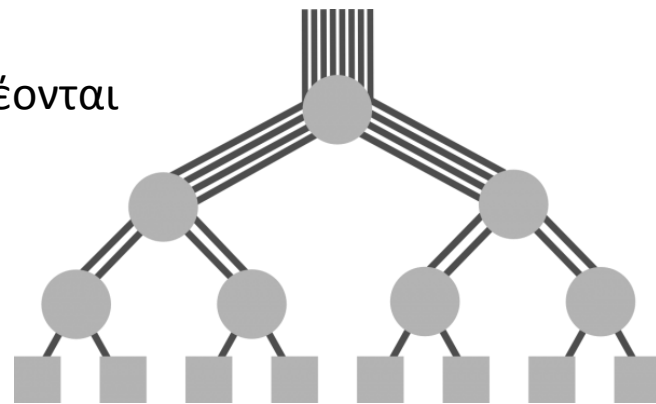


Omega network

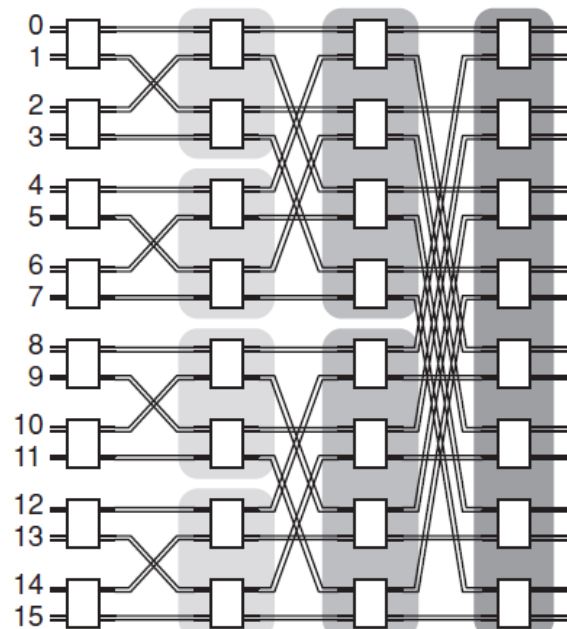
Centralized Switched Networks

Fat tree

- Τα φύλλα του δέντρου είναι τα στοιχεία που διασυνδέονται
- Οι εσωτερικοί κόμβοι είναι διακόπτες
- Χρησιμοποιείται κατά κόρον σε SANs και κυρίως σε Supercomputers (Infiniband, Myrinet, κλπ)



Benes network

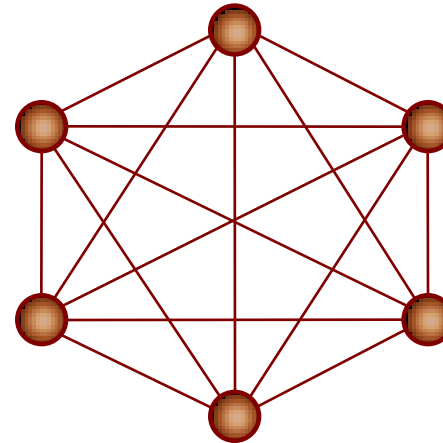


Folded Benes network

Distributed Switched Networks

Fully connected

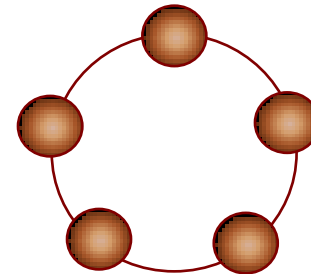
- N κόμβοι
- $N(N-1)/2$ σύνδεσμοι
- Βαθμός κόμβου $d=N-1$
- Διάμετρος $D=1$
- Εύρος τομής $b=(N/2)^2$
- Είναι συμμετρικό



Distributed Switched Networks

Ring

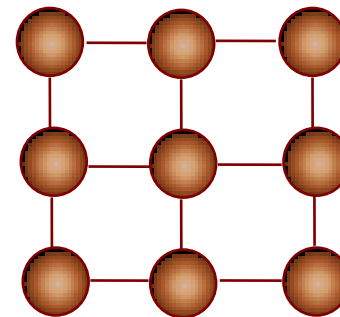
- N κόμβοι
- N σύνδεσμοι
- Βαθμός κόμβων $d=2$
- Διάμετρος: $N/2$
- Εύρος τομής $b=2$
- Είναι συμμετρικό



Distributed Switched Networks

Mesh

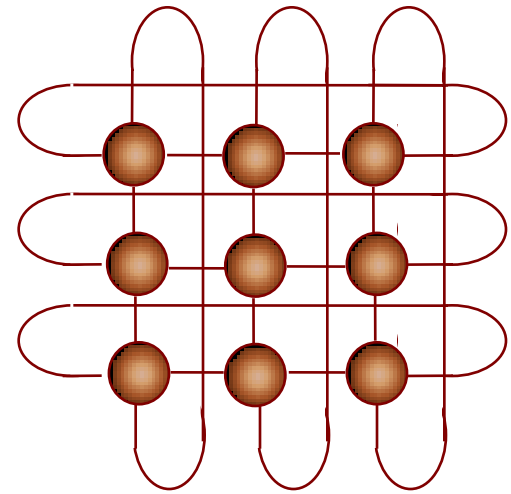
- $N=n^k$ κόμβοι
- k -διάστατο mesh με n κόμβους ανά διεύθυνση
- βαθμός κόμβου $d=2k$
- διάμετρος δικτύου $D=k(n-1)$
- Για ένα 2-διάστατο mesh:
 - » $N=n^2$ κόμβοι
 - » $2N-2n=2n^2-2n$ σύνδεσμοι
 - » Βαθμός εσωτερικών κόμβων $d=4$
 - » Διάμετρος $D=2(n-1)$
 - » Εύρος τομής $b=n$
 - » Δεν είναι συμμετρικό



Distributed Switched Networks

Torus

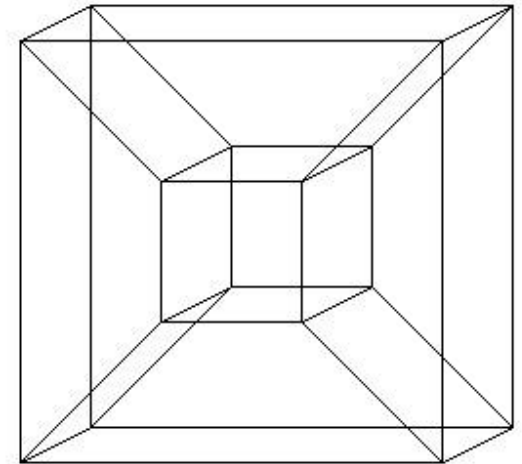
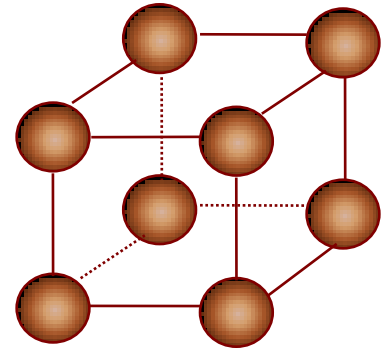
- Υποδιπλασιάζεται η διάμετρος σε σχέση με το Mesh
- για έναν $n \times n$ δυαδικό torus ($k=2$):
 - » $N=n^2$ κόμβοι
 - » $2N$ σύνδεσμοι
 - » βαθμός κόμβου $d=4$
 - » Διάμετρος $D = 2 \left\lfloor \frac{n}{2} \right\rfloor$
 - » Εύρος τομής $2n$
 - » Είναι συμμετρικό



Distributed Switched Networks

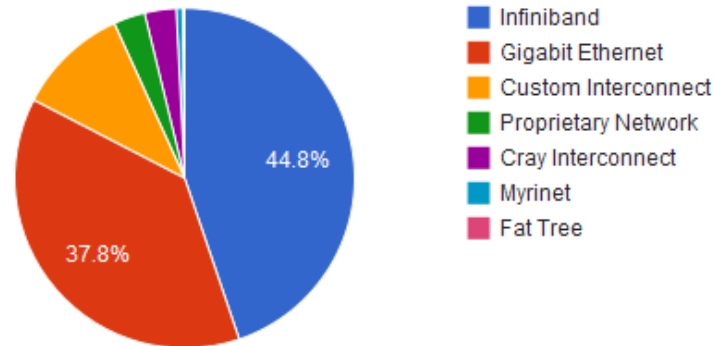
Hypercube

- $N=2^n$ κόμβοι
- $nN/2$ σύνδεσμοι
- Βαθμός κόμβου $d=n$
- Διάμετρος $D=n$
- Εύρος τομής $b=N/2$
- Είναι συμμετρικό
- Άμεσος προσδιορισμός διαδρομής

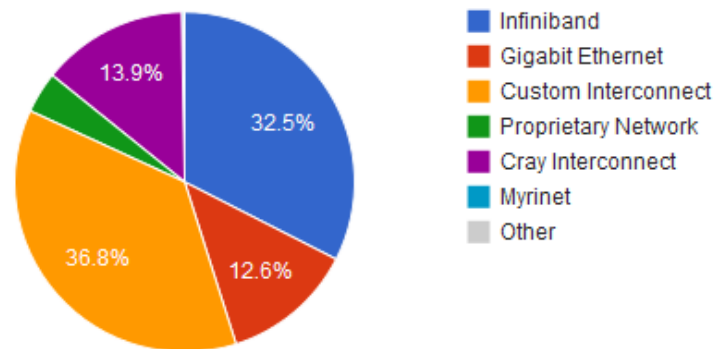


Δίκτυα διασύνδεσης στους υπερυπολογιστές Top500, November 2012

Interconnect Family System Share



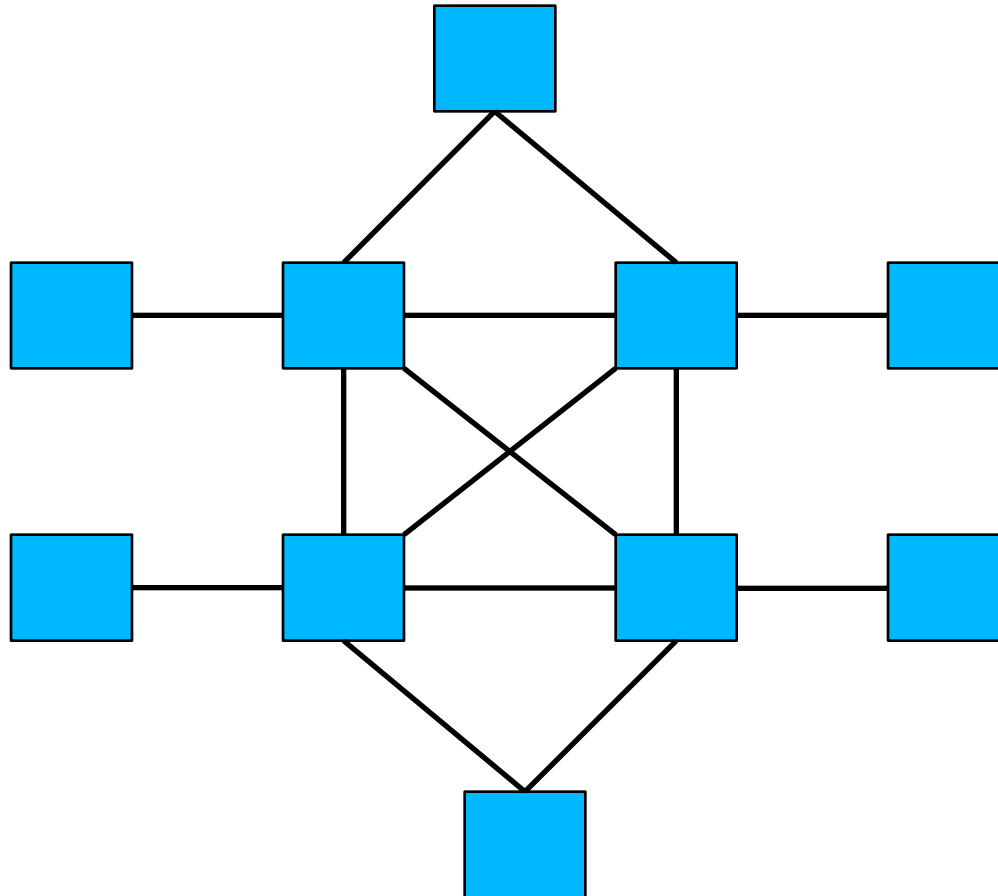
Interconnect Family Performance Share



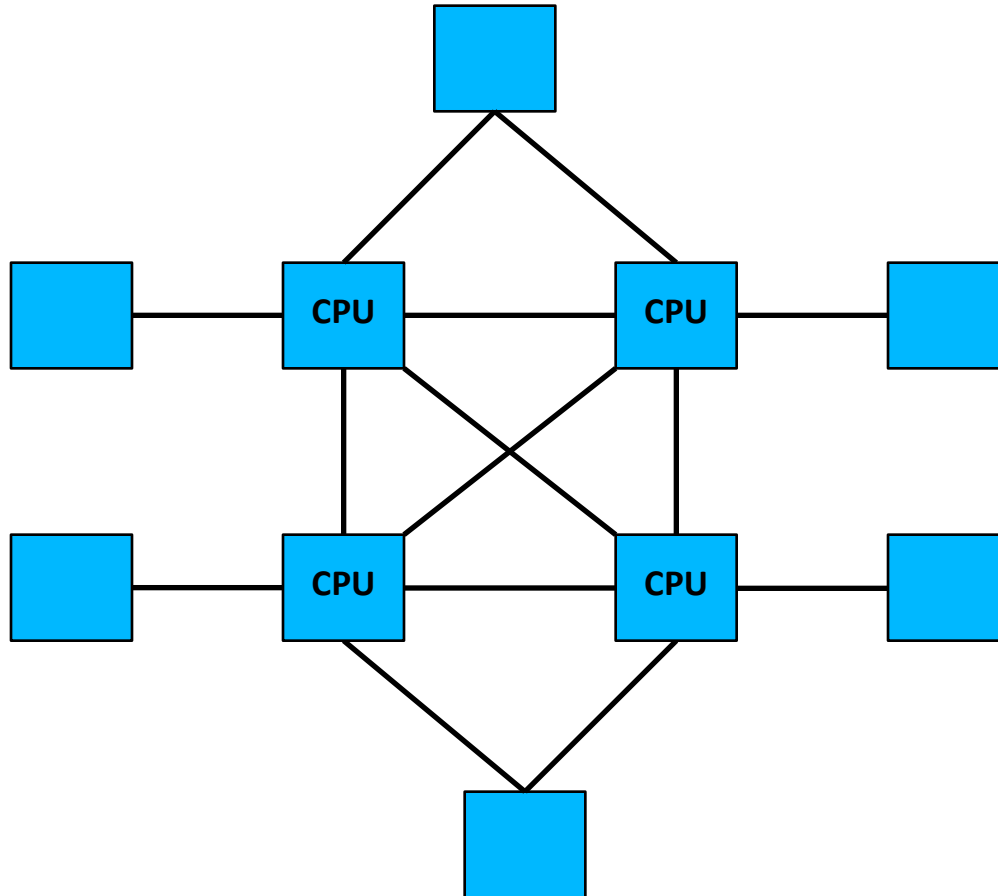
Supercomputers

- **BlueGene/Q** : 5D torus
- **BlueGene/P** : binary tree, 3D torus
- **K computer**: 6D torus
- **Infiniband configuration**: fat tree
- **Historical note (1987)**: Connection Machine CM-2, 8192 nodes,
hypercube

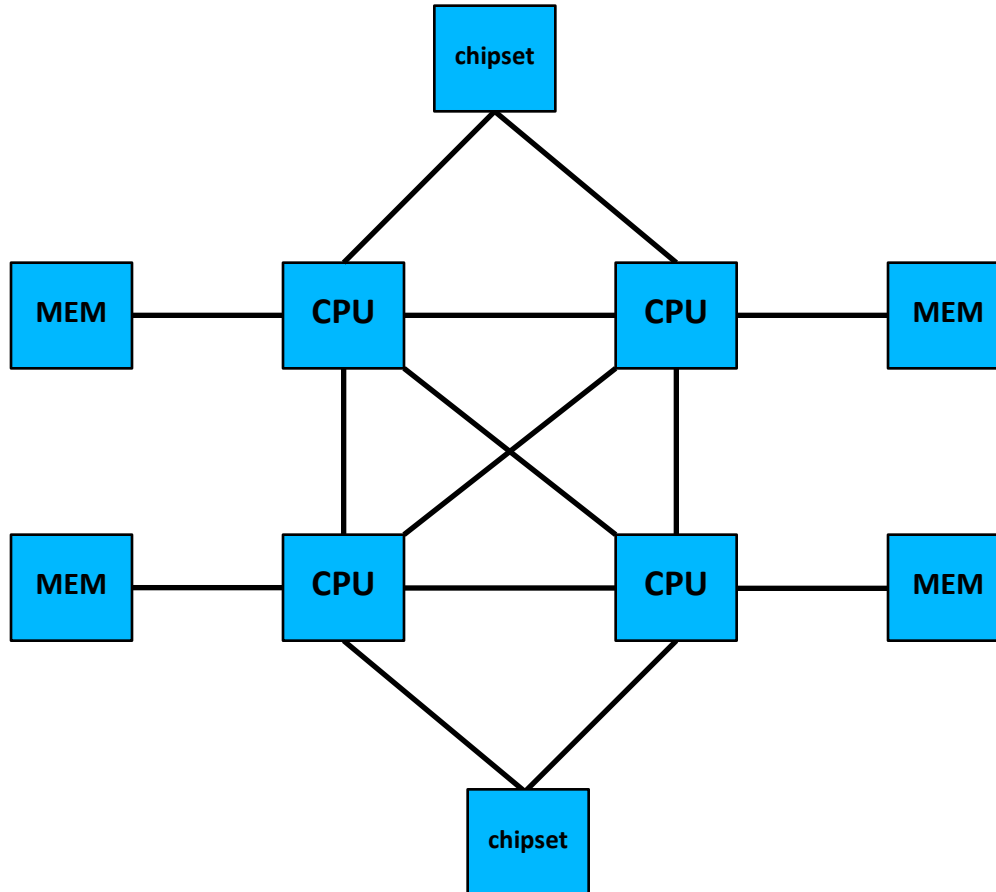
Δίκτυα διασύνδεσης



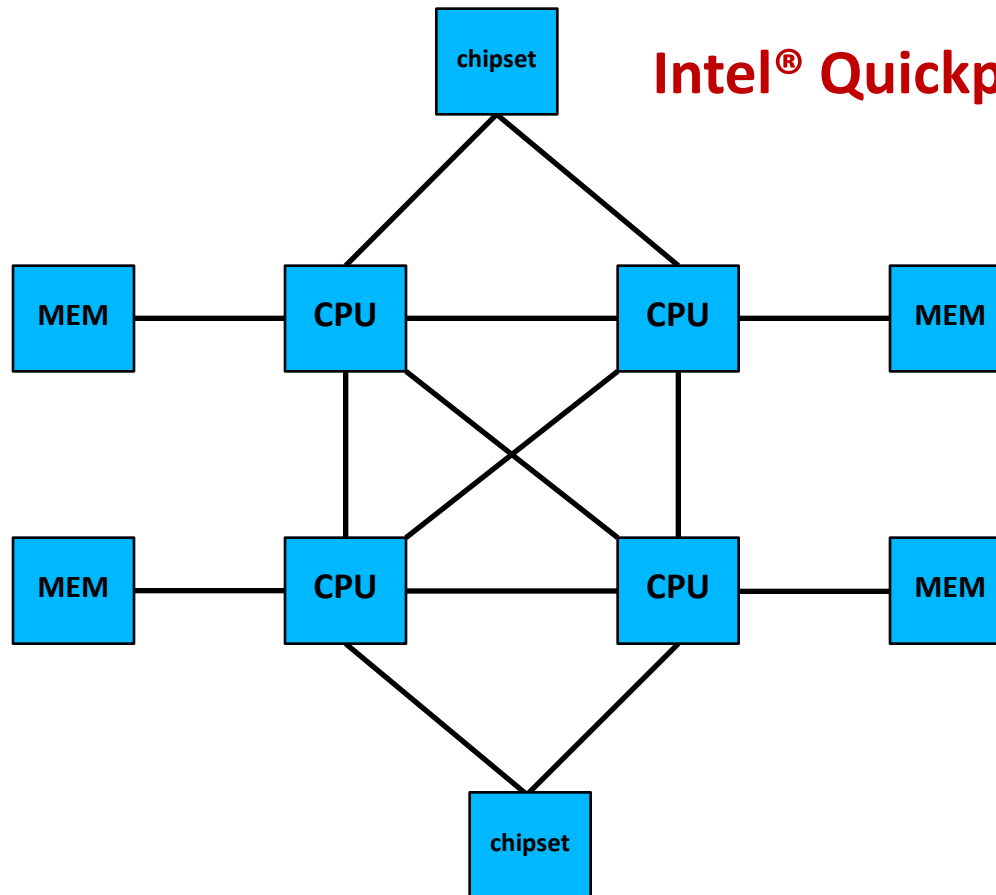
Δίκτυα διασύνδεσης



Δίκτυα διασύνδεσης



Δίκτυα διασύνδεσης



Intel® Quickpath Interconnect

Intel® Quickpath Interconnect

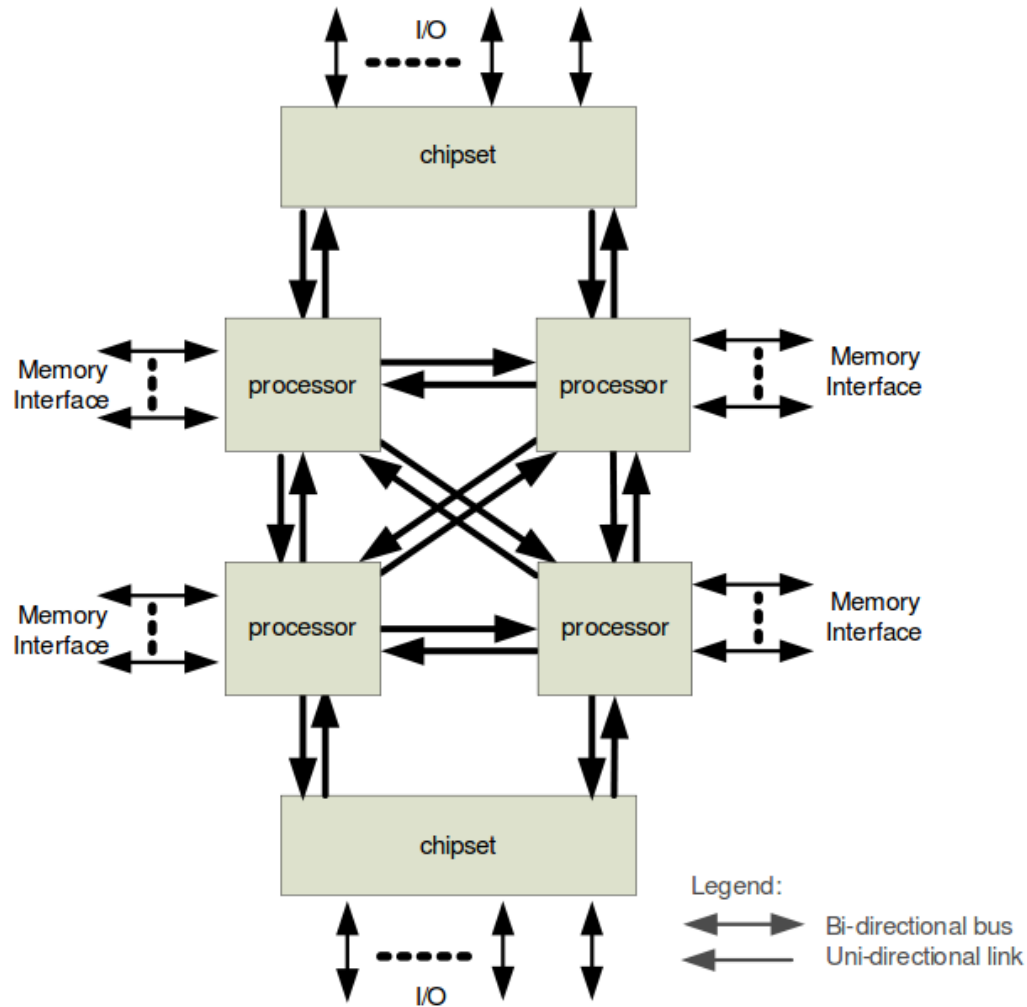


Image taken from: An Introduction to the Intel® QuickPath Interconnect:

<http://www.intel.com/content/www/us/en/io/quickpath-technology/quick-path-interconnect-introduction-paper.html>

Λίγη διαφήμιση...

- Συστήματα Παράλληλης Επεξεργασίας (9^ο Εξάμηνο)
- Αντικείμενο μαθήματος:
 - Αρχιτεκτονικές Παράλληλων Συστημάτων
 - Δίκτυα διασύνδεσης
 - Μοντελοποίηση επίδοσης
 - Σχεδιασμός παράλληλων προγραμμάτων
 - Υλοποίηση παράλληλων προγραμμάτων
 - Συγχρονισμός
 - Χρονοδρομολόγηση
 - Εφαρμογές

Λίγη διαφήμιση...

- Εργαστηριακές ασκήσεις:
 - Προγραμματισμός για συστήματα μεγάλης κλίμακας (Message Passing Interface – MPI)
 - Προγραμματισμός για πολυπύρρηνα συστήματα (OpenMP, Cilk, TBBs)
 - Προγραμματισμός για κάρτες γραφικών (CUDA) και επιταχυντές (Xeon Phi)
 - Εκτέλεση εφαρμογών σε:
 - » 8-node, Intel Xeon Cluster (16 cores, 32 threads)
 - » 32-node, 8-core Intel Harpertown Cluster (256 cores)
 - » 24-core Intel Dunnington
 - » 32-core/64-thread Intel Sandy Bridge
 - » Fermi GPUs
 - 2^η Εργαστηριακή Άσκηση Ακ. Έτος 2011-2012:
Students of the Parallel Processing Systems course, “Parallelizing the Floyd-Warshall Algorithm on Modern Multicore Platforms: Lessons Learned”, 5th Conference of Electrical and Computer Engineering Students, University of Thrace, Xanthi, Greece, April 2012.