

Tendencias de Uso y Diseño de Redes de Interconexión en Computadores Paralelos

14 de Abril, 2016
Universidad Complutense de Madrid

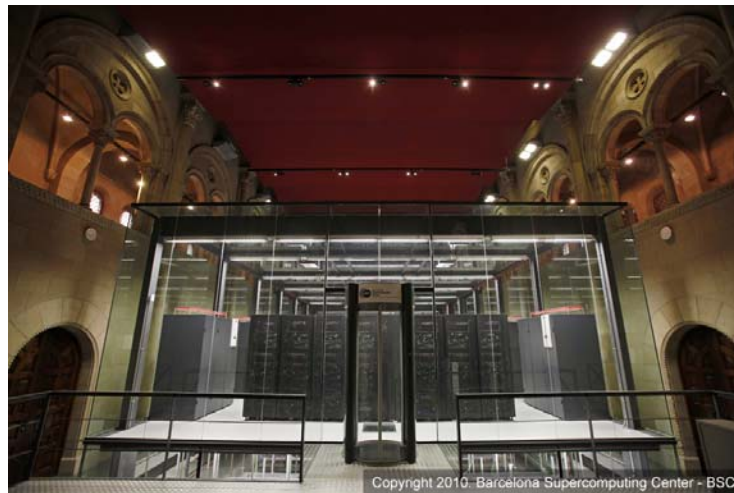
Ramón Beivide
Universidad de Cantabria



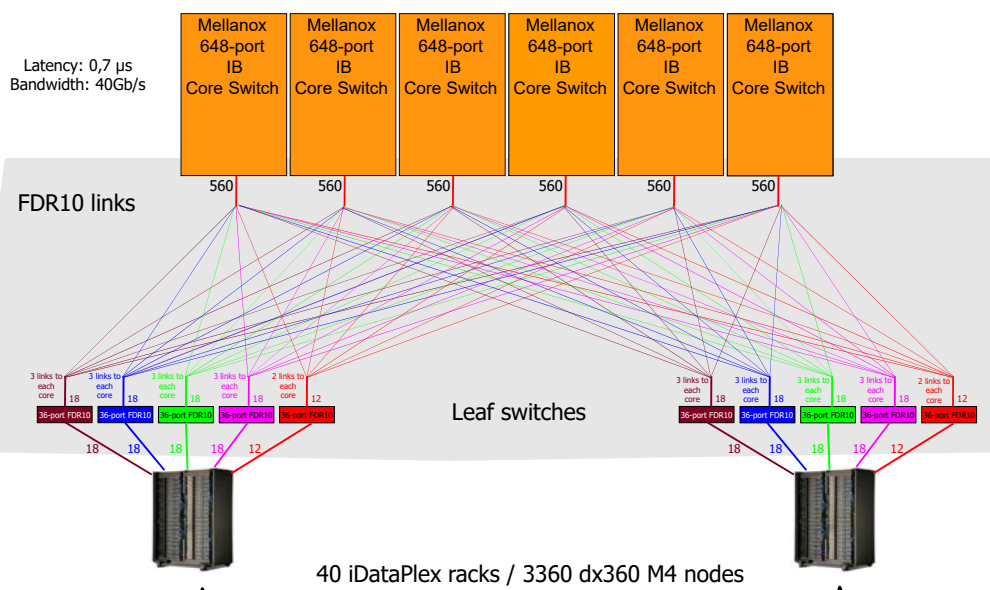
Outline

- 1. Introduction
- 2. Network Basis
- 3. System networks
- 4. On-chip networks (NoCs)
- 5. Some current research

1. Intro: MareNostrum



1. Intro: MareNostrum BSC, Infiniband FDR10 non-blocking Folded Clos (up to 40 racks)



1. Intro: Infiniband core switches



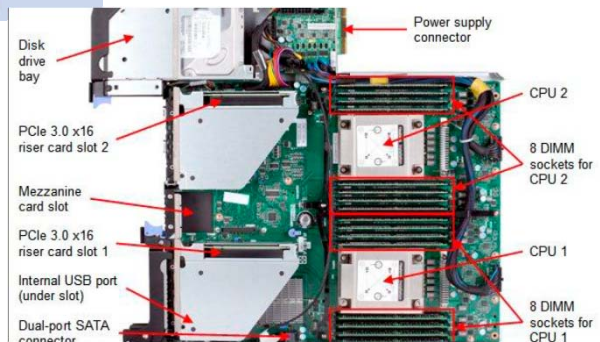
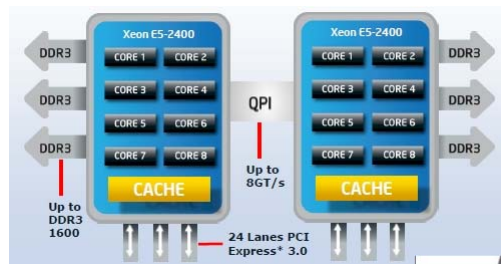
1. Intro: Cost dominated by (optical) wires



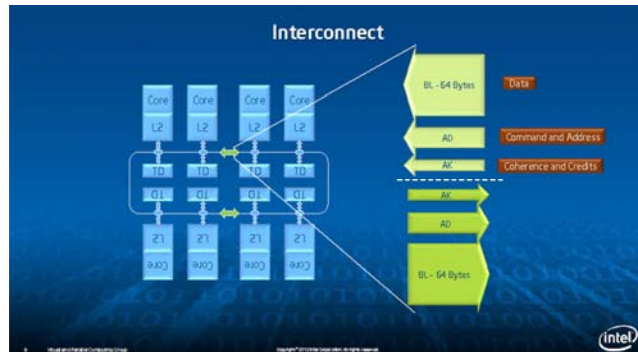
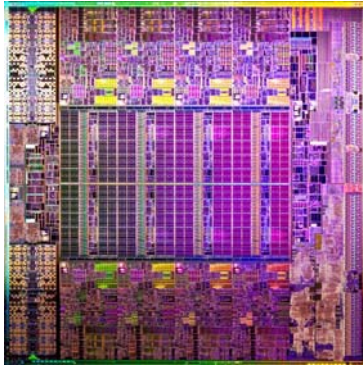
1. Intro: Blades



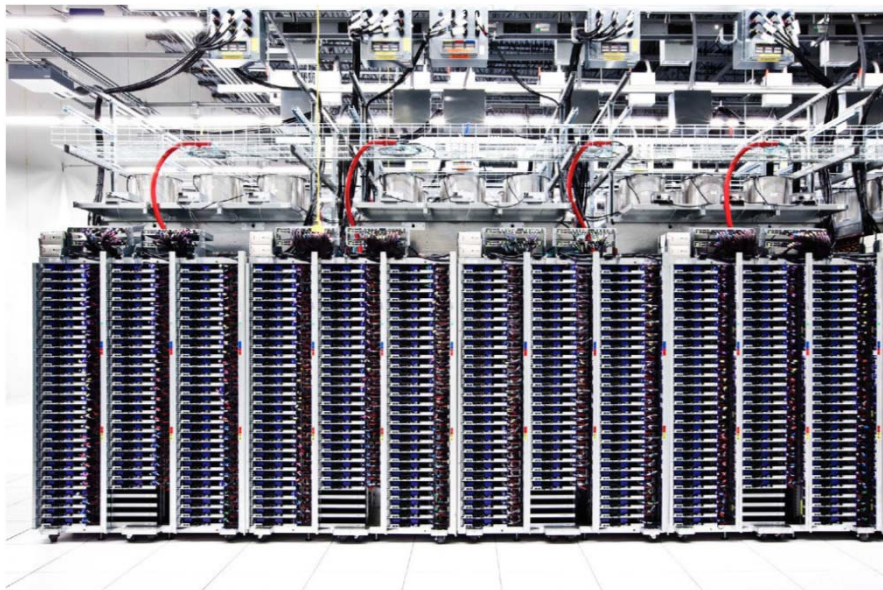
1. Intro: Blades



1. Intro: Multicore E5-2670 Xeon Processor



1. Intro: A row of servers in a Google DataCenter, 2012.

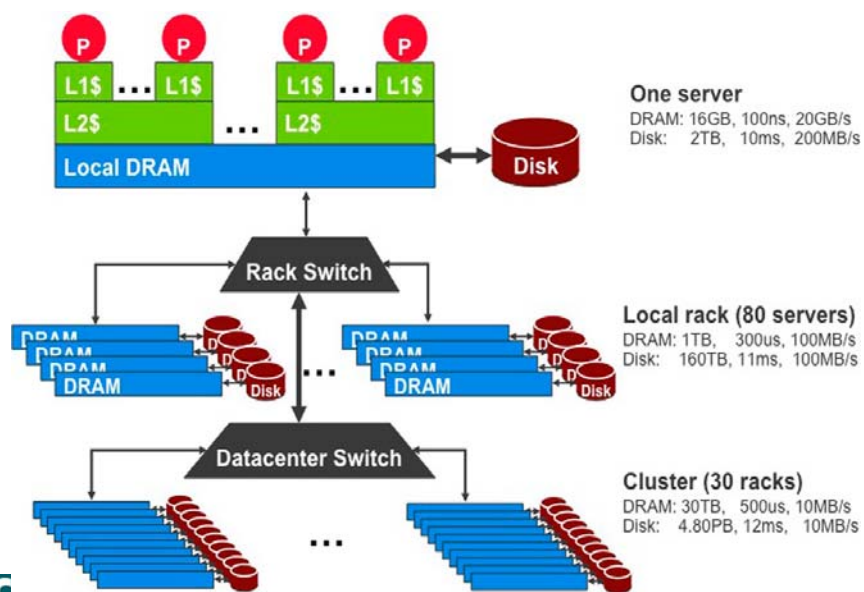


3. WSCs Array: Enrackable boards or blades + rack router

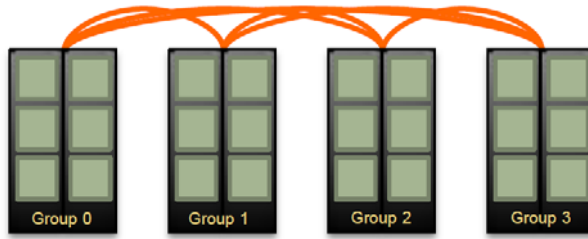


Figure 1.1: Sketch of the typical elements in warehouse-scale systems: 1U server (left), 7' rack with Ethernet switch (middle), and diagram of a small cluster with a cluster-level Ethernet switch/router (right).

3. WSC Hierarchy



1. Intro: Cray Cascade (XC30, XC40)



1. Intro: Cray Cascade (XC30, XC40)

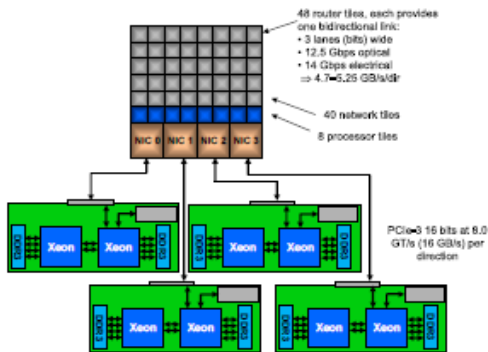
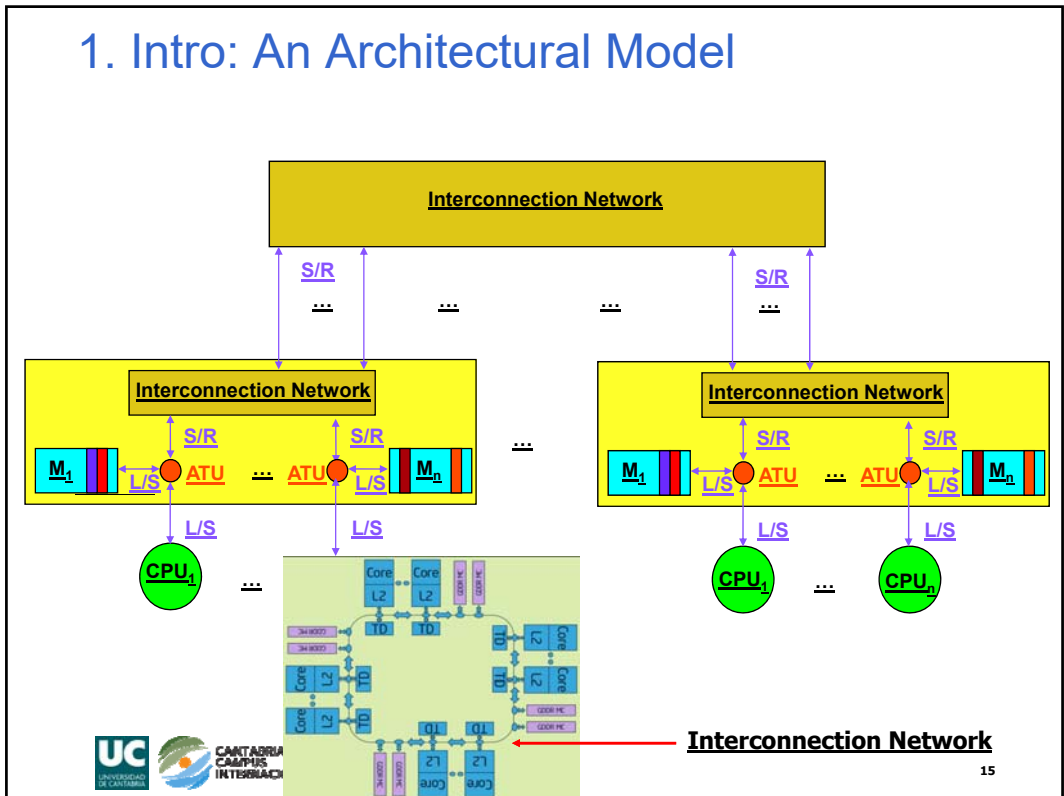


Figure 2. A single Aries system-on-a-chip device provides network connectivity for the four nodes on a Cascade blade.

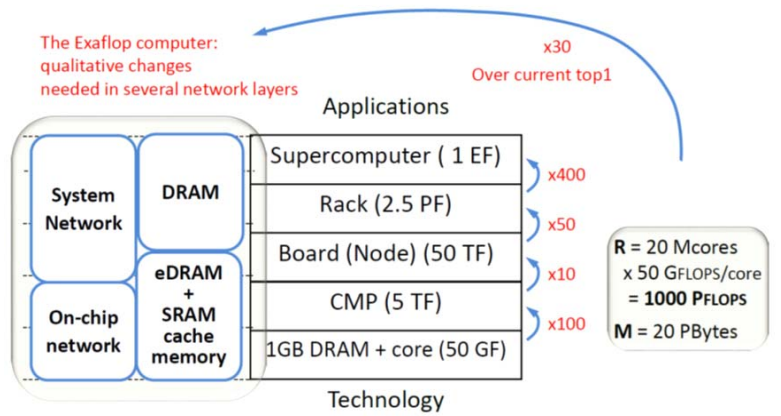


Figure 1. A Cascade blade which consists of 8 Xeon sockets, organized as 4 dual-socket nodes, and a single Aries ASIC.

1. Intro: An Architectural Model



1. Intro: What we need for one ExaFlop/s



Networks are pervasive and critical components in Supercomputers, Datacenters, Servers and Mobile Computers.

Complexity is moving from system networks towards on-chip networks: less nodes but more complex

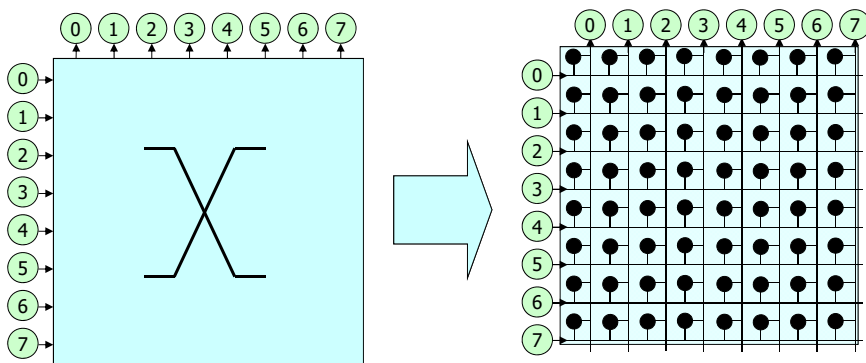
Outline

- 1. Introduction
- 2. Network Basis
 - Crossbars & Routers
 - Direct vs Indirect Networks
- 3. System networks
- 4. On-chip networks (NoCs)
- 5. Some current research

2. Network Basis

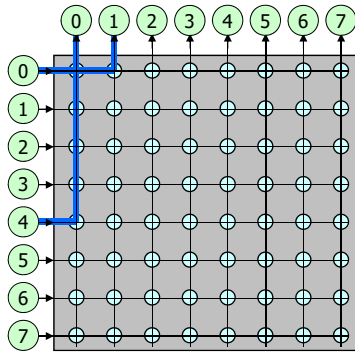
All networks based on **Crossbar switches**

- Switch complexity increases quadratically with the number of crossbar input/output ports, N , i.e., grows as $O(N^2)$
- Has the property of being *non-blocking* ($N!$ I/O permutations)
- *Bidirectional* for exploiting communication locality
- Minimize *latency* & maximize *throughput*

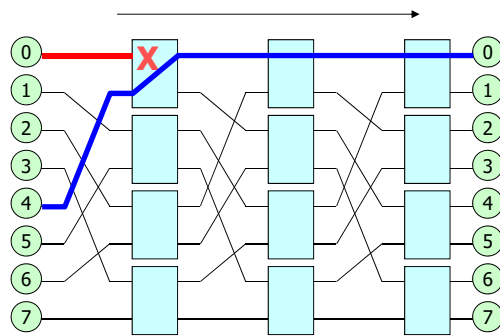


2. Blocking vs. Non-blocking

- Reduction cost comes at the price of performance
 - Some networks have the property of being *blocking (Not N!)*
 - *Contention* is more likely to occur on network links
 - › Paths from different sources to different destinations share one or more links



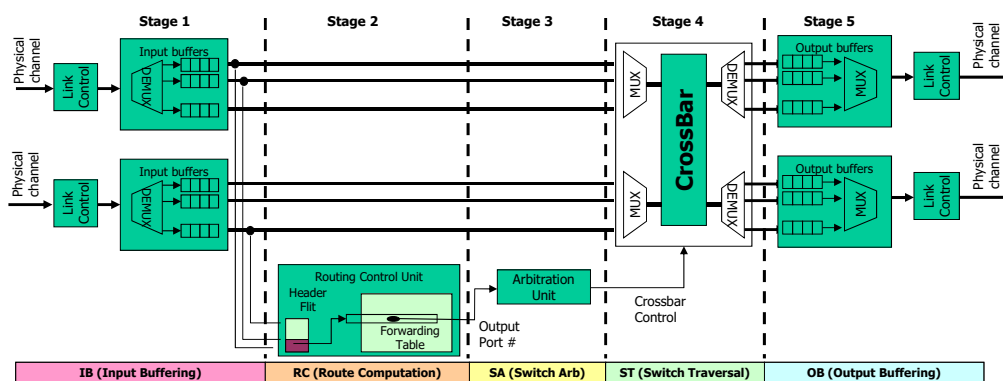
non-blocking topology



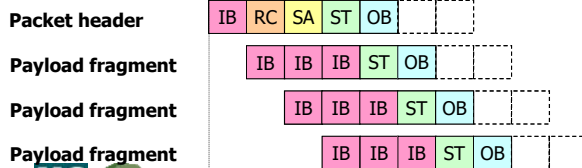
blocking topology

2. Switch or Router Microarchitecture

Pipelined Switch Microarchitecture



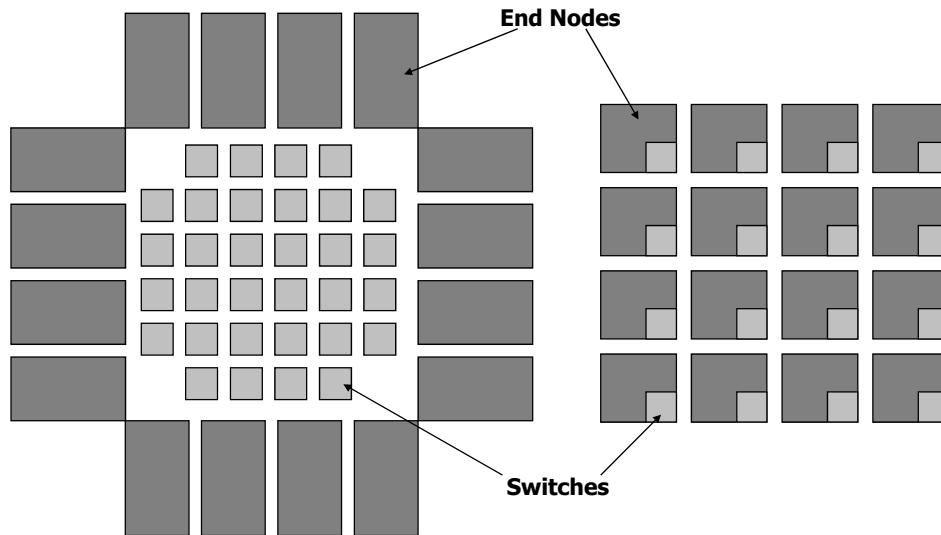
IB (Input Buffering) RC (Route Computation) SA (Switch Arb) ST (Switch Traversal) OB (Output Buffering)



Matching the throughput of the internal switch datapath to the external link BW is the goal

2. Network Organization

Indirect (Centralized) and Direct (Distributed) Networks



2. Previous Myrinet core switches (Indirect, Centralized)

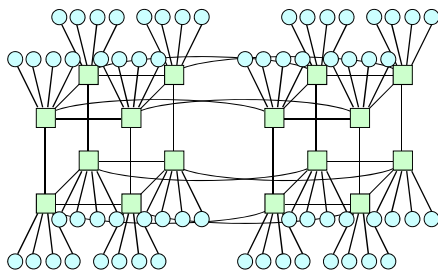


2. IBM BG/Q (Direct, Distributed)

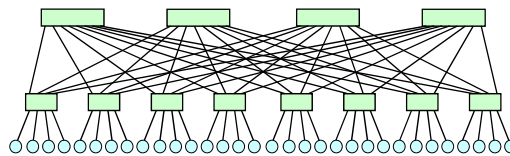


2. Network Organization

- As **crossbars** do not scale they need to be **interconnected** for servicing an increasing number of endpoints.
- **Direct** (Distributed) vs **Indirect** (Centralized) Networks
 - **Concentration** can be used to reduce network costs
 - “*c*” end nodes connect to each switch
 - Allows larger systems to be built from fewer switches and links
 - Requires larger *switch degree*



64-node system with 8-port switches, $c = 4$

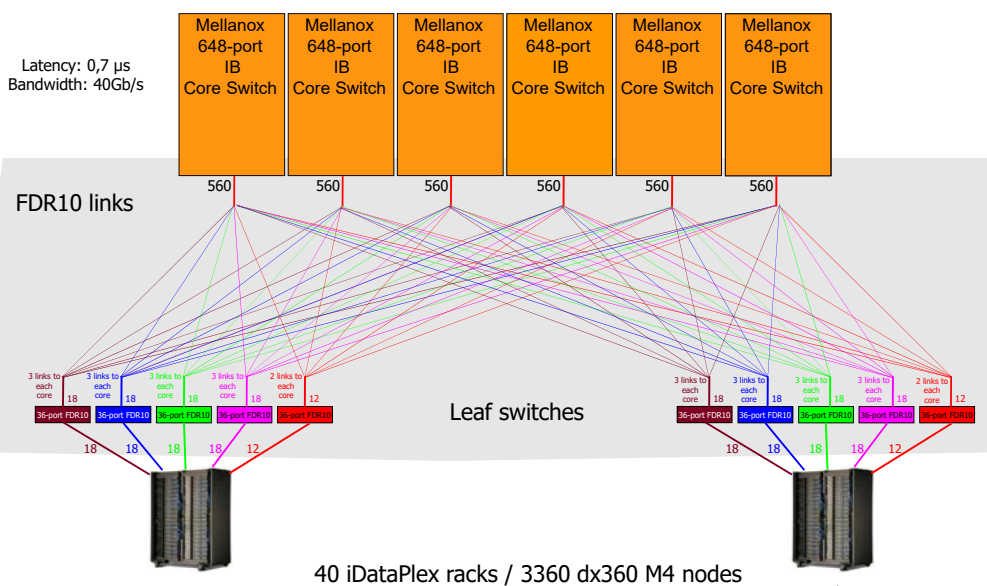


32-node system with 8-port switches

Outline

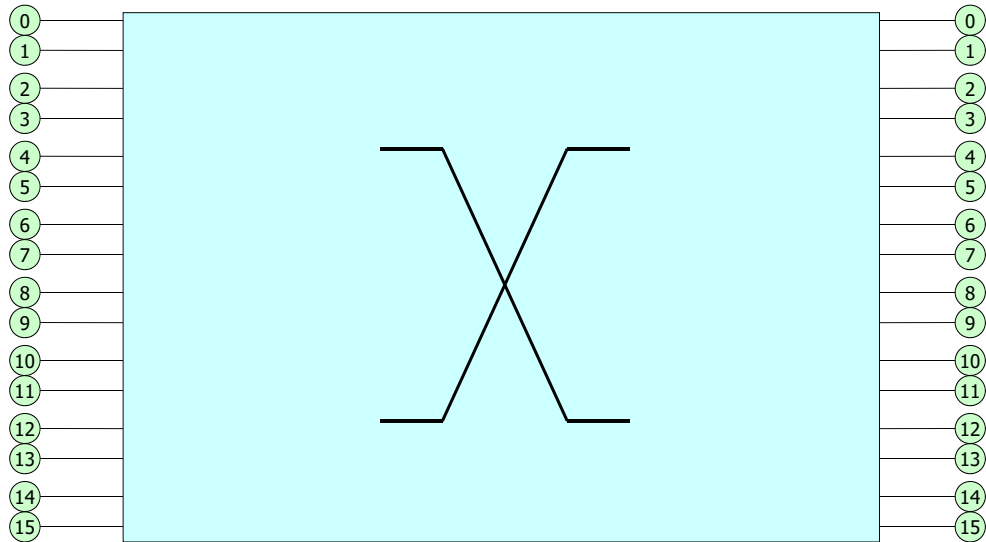
- 1. Introduction
- 2. Network Basis
- 3. System networks
 - Folded Clos
 - Tori
 - Dragonflies
- 4. On-chip networks (NoCs)
- 5. Some current research

3. MareNostrum BSC, Infiniband FDR10 non-blocking Folded Clos (up to 40 racks)



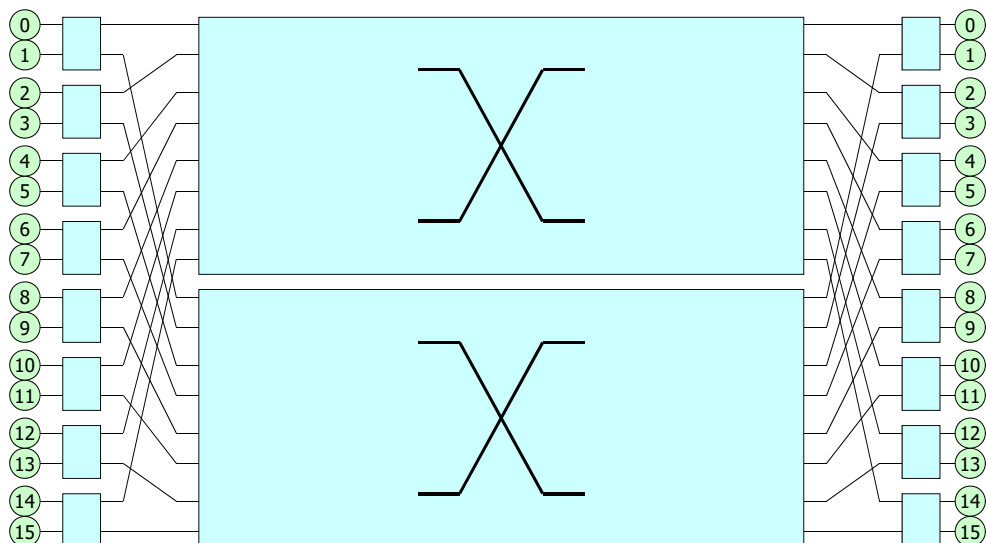
3. Network Topology

Centralized Switched (Indirect) Networks



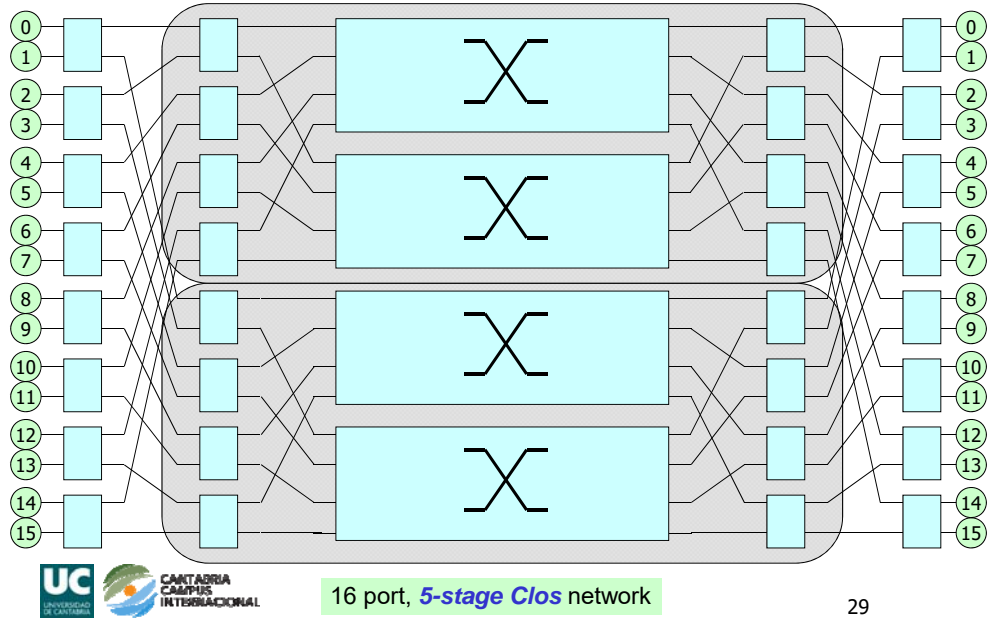
3. Network Topology

Centralized Switched (Indirect) Networks



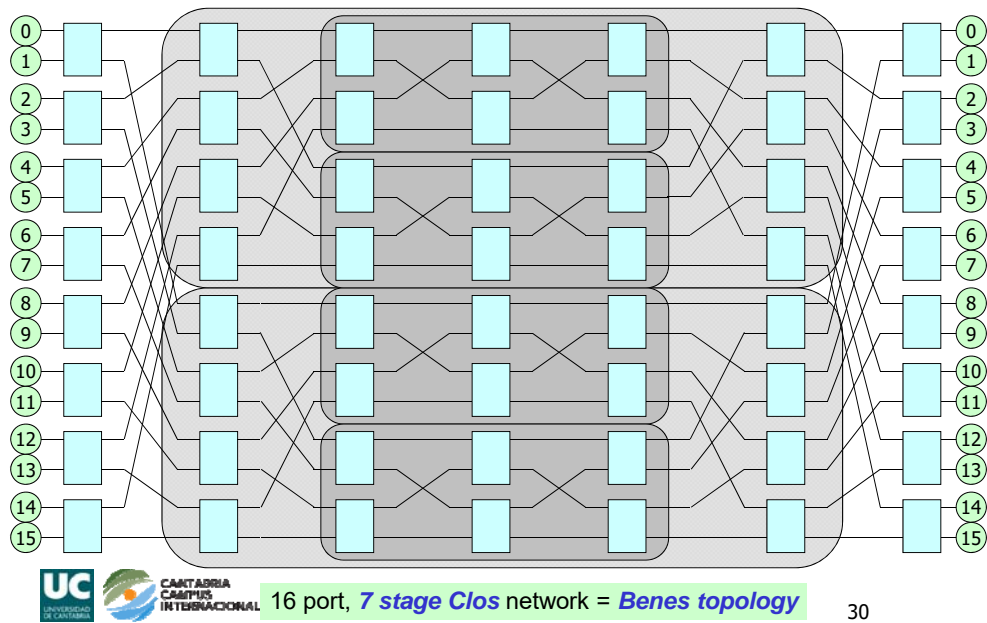
3. Network Topology

Centralized Switched (Indirect) Networks



3. Network Topology

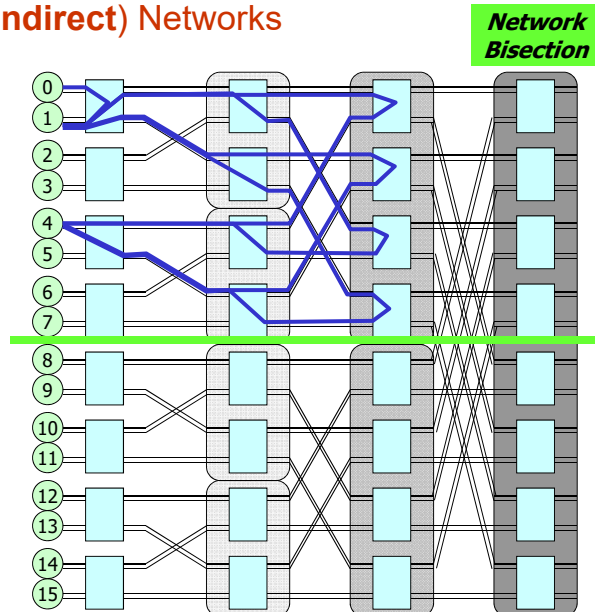
Centralized Switched (Indirect) Networks



3. Network Topology

Centralized Switched (Indirect) Networks

- *Bidirectional MINs*
- Increase modularity
- Reduce hop count, d
- **Folded Clos network**
 - Nodes at tree leaves
 - Switches at tree vertices
 - Total link bandwidth is constant across all tree levels, with **full bisection bandwidth**



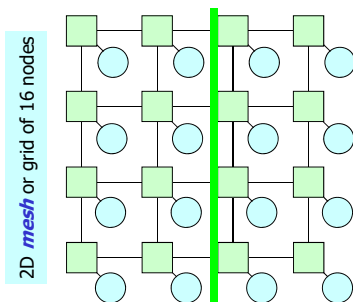
Network Bisection

Folded Clos = Folded Benes <-> Fat tree network !!!

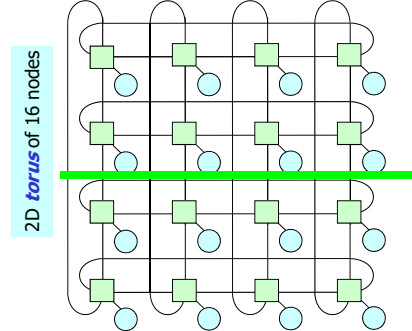


3. Other **DIRECT** System Network Topologies

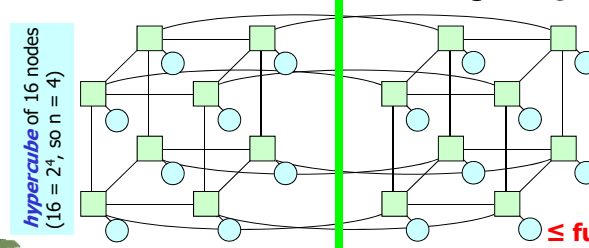
Distributed Switched (Direct) Networks



2D mesh or grid of 16 nodes



2D torus of 16 nodes



hypercube of 16 nodes
($16 = 2^4$, so $n = 4$)

Network Bisection

≤ full bisection bandwidth!

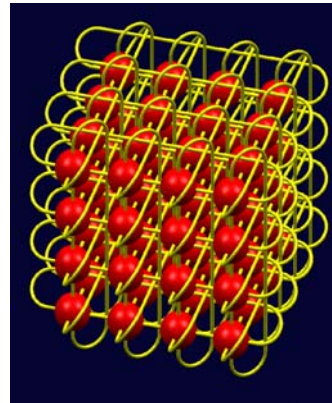


3. IBM BlueGene/L/P Network

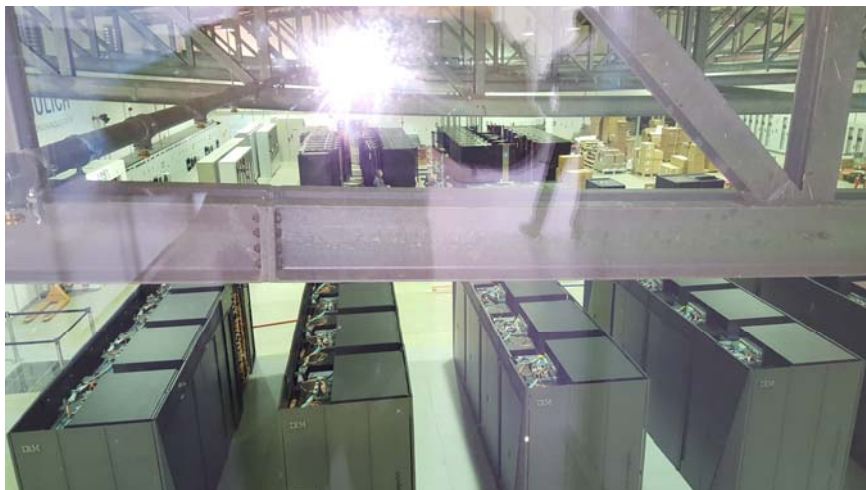
Prismatic 32x32x64 Torus (mixed-radix networks)

BlueGene/P: 32x32x72 in maximum configuration

Mixed-radix prismatic Tori also used by Cray



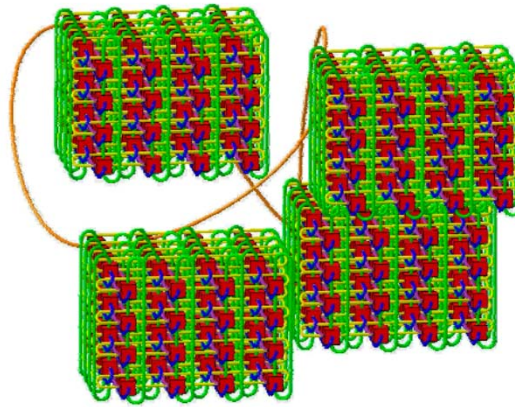
3. IBM BG/Q



3. IBM BG/Q



Blue Gene/Q Network: 5D Torus



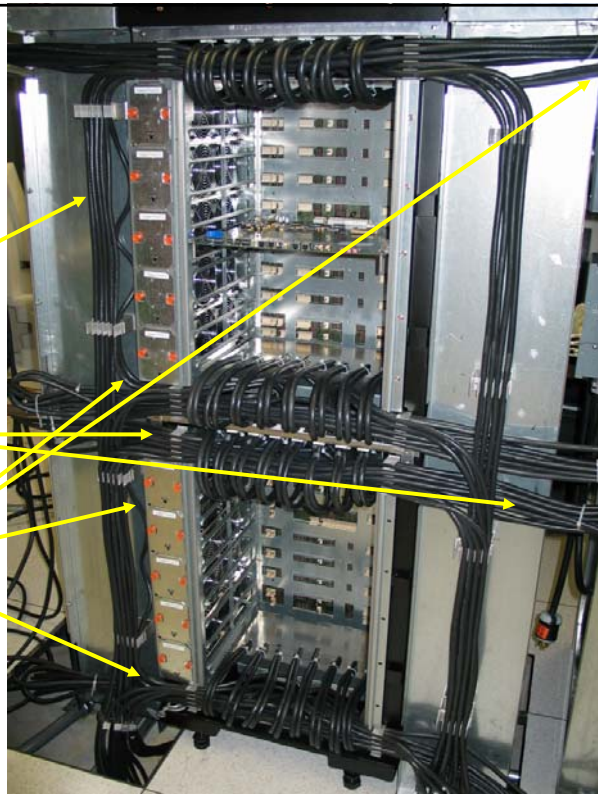
3 .BG Network Routing

Y Wires

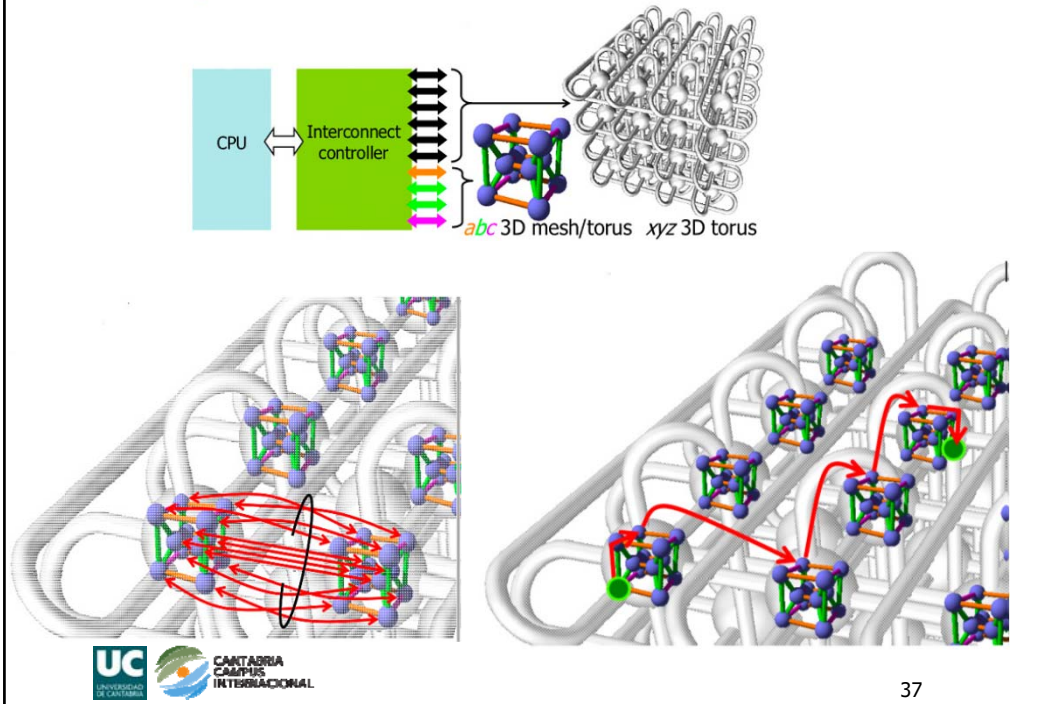
X Wires

Z Wires

Adaptive Bubble Routing
ATC-UC Research Group



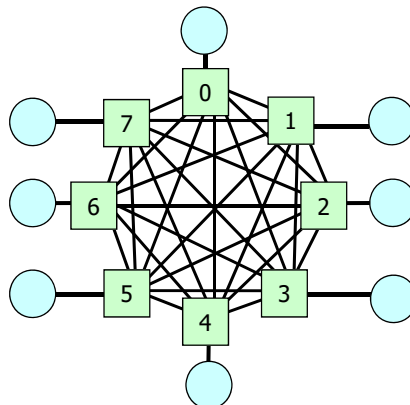
3. Fujitsu Tofu Network



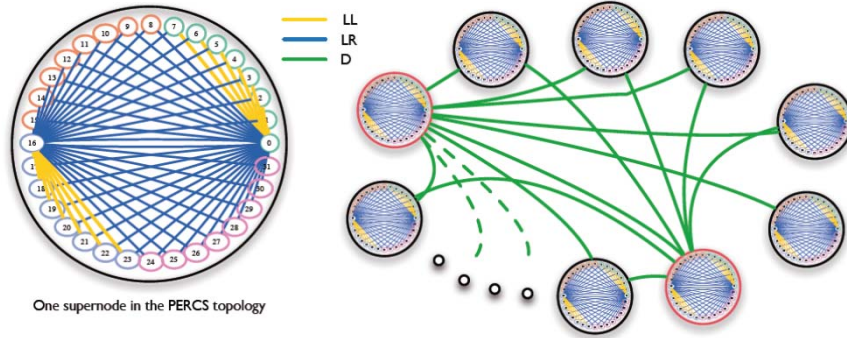
3. More Recent Network Topologies

Distributed Switched (**Direct**) Networks

- **Fully-connected network**: all nodes are directly connected to all other nodes using bidirectional dedicated links

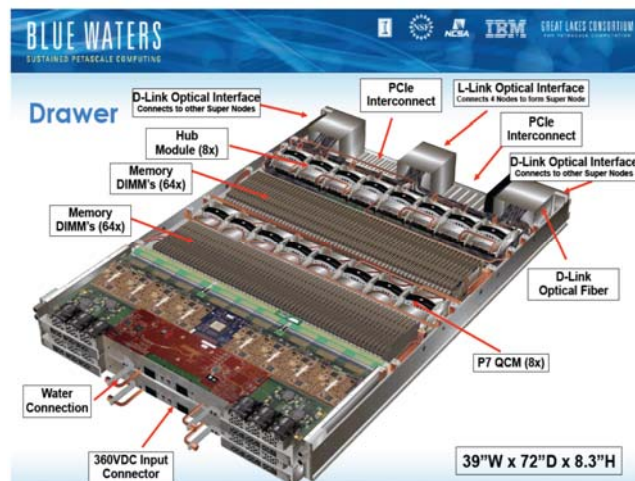


3. IBM PERCS



One supermode in the PERCS topology

3. IBM PERCS




3. IBM PERCS

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

IBM GREAT LAKES CONSORTIUM

Interconnection Network

- Hub/Switch (one per SMP node)
peak bi-directional bandwidths
 - 192 GB/s to host node (WXYZ links, 24 GB/s each way)
 - 336 GB/s to 7 other nodes in same drawer (L-local links, 24 GB/s each way)
 - 240 GB/s to 24 nodes in other 3 drawers in same SuperNode (L-remote links, 5 GB/s each way)
 - 320 GB/s to hubs in other SuperNodes (Up to 16 D-links, 10 GB/s each way)
(Note that BW is not fully populated)



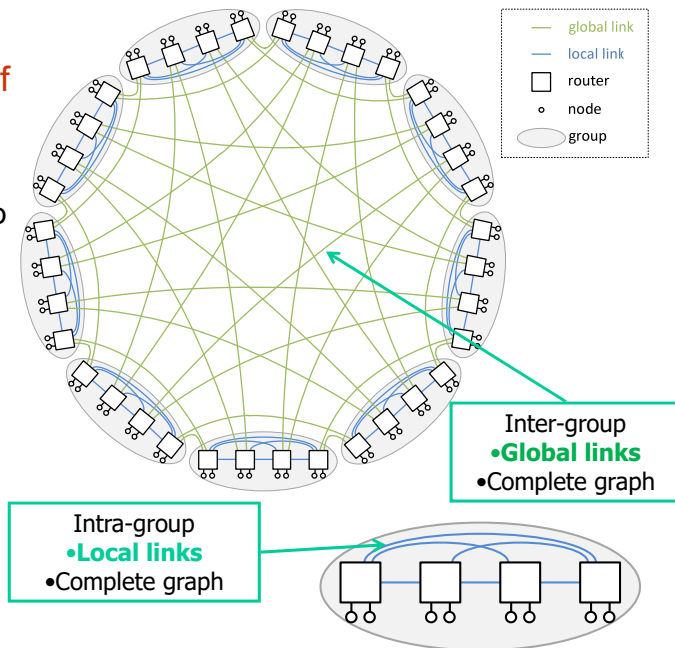
3. Dragonfly Interconnection Network

Organized as groups of routers

Parameters:

- a**: Routers per group
- p**: Node per router
- h**: Global link per router
- Well-balanced dragonfly [1]

$$a = 2p = 2h$$



3. Dragonfly Interconnection Network

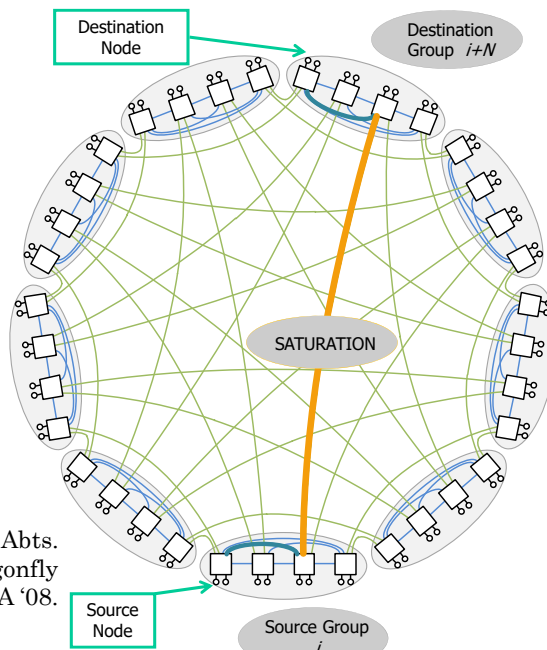
Minimal routing

- Longest path 3 hops:
 - local-global-local**
- Good performance under UN traffic

Adversarial traffic [1]

- ADV+N: Nodes in group i send traffic to group $i+N$
- Saturation of the global link

[1] J. Kim, W. Dally, S. Scott, and D. Abts. "Technology-driven, highly-scalable dragonfly topology." ISCA '08.

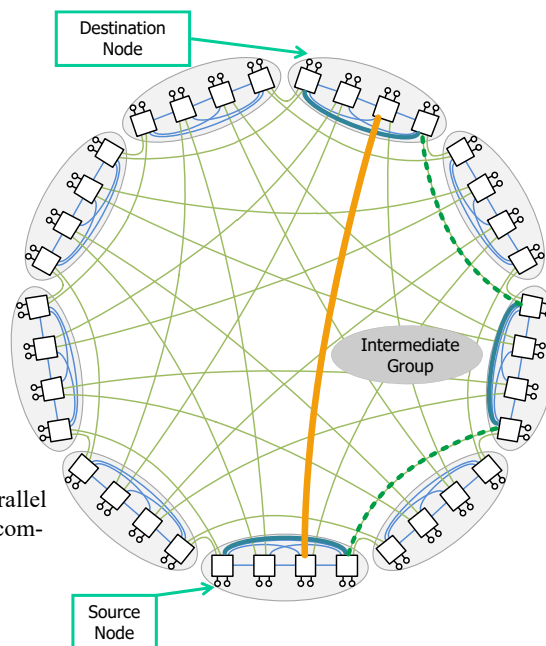


3. Dragonfly Interconnection Network

Valiant Routing [2]

- Randomly selects an intermediate group to misroute packets
- Avoids saturated channel
- Longest path 5 hops
 - local-global-local-global-local**

[2] L. Valiant, "A scheme for fast parallel communication," SIAM journal on computing, vol. 11, p. 350, 1982.



3. Cray Cascade, electrical supernode

backplanes connected with copper cables in a group: "Black Network"

Optical cables interconnect groups "Blue Network"

Aries connected by backplane "Green Network"

4 nodes connect to a single Aries

UC UNIVERSITY OF CANTABRIA
CANTABRIA CAMPUS INTERNACIONAL

45

3. Cray Cascade, system and routing

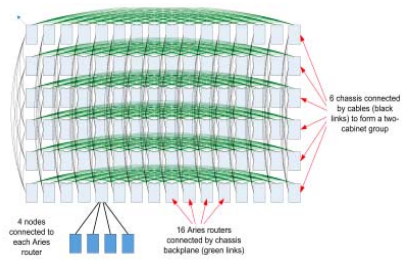


Figure 4. Structure of a Cascade electrical group. Each row represents the 16 Aries in a chassis, with 4 nodes attached to each, and connected by the chassis backplane (green links). Each column represents an Aries in one of the six chassis of a two-cabinet group, connected by electrical cables (black links).

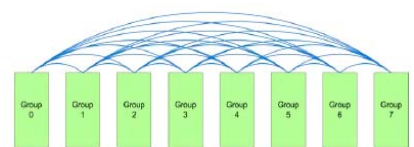
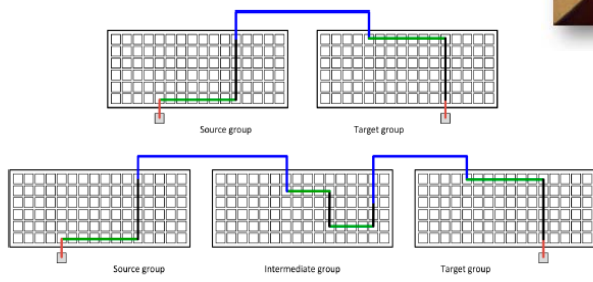


Figure 5. The global (blue) links connect Dragonfly groups together. In a large system these links are active optical cables.

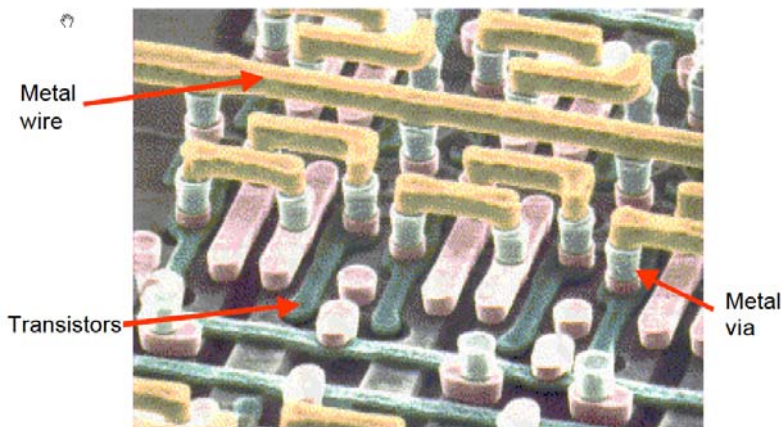


Outline

- 1. Introduction
- 2. Network Basis
- 3. System networks
- 4. On-chip networks (NoCs)
 - Rings
 - Meshes
- 5. Some current research

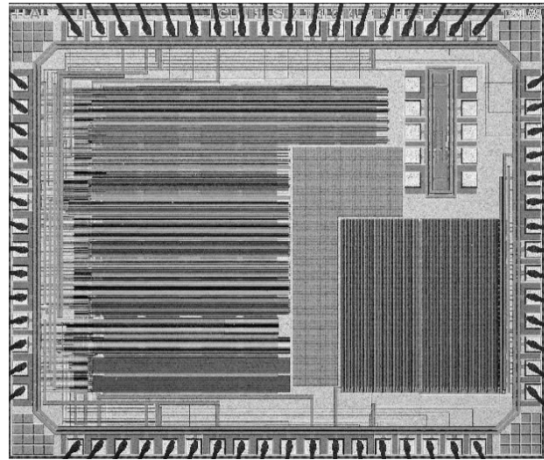
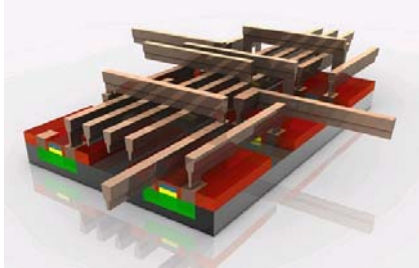
4. On-Chip local interconnects

SEM photo of local levels interconnect

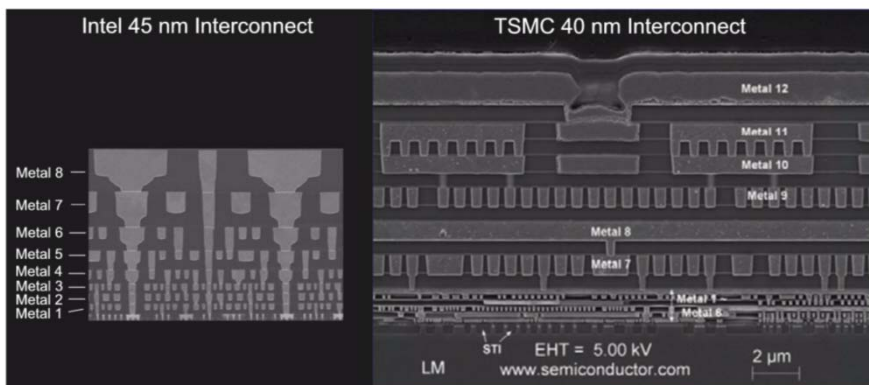


4. On-Chip global interconnects

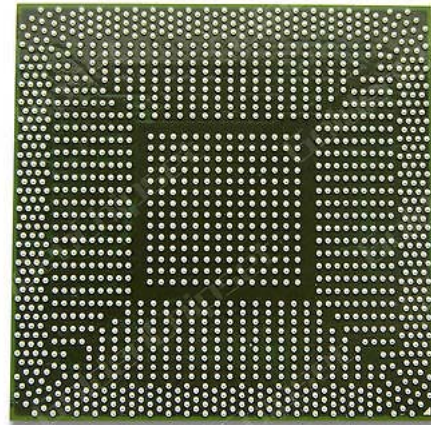
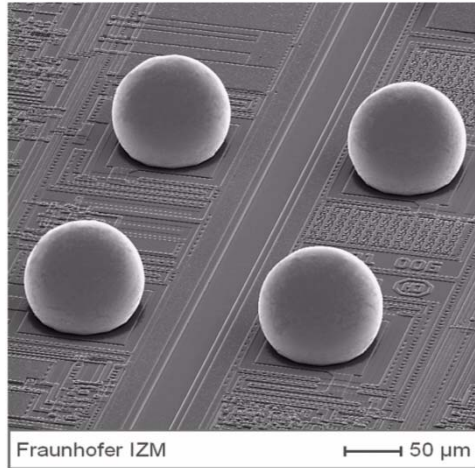
Global levels interconnect



4. Metal Layers



4. Bumps & Balls

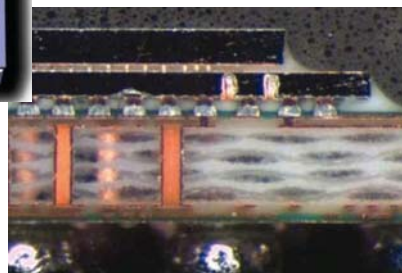
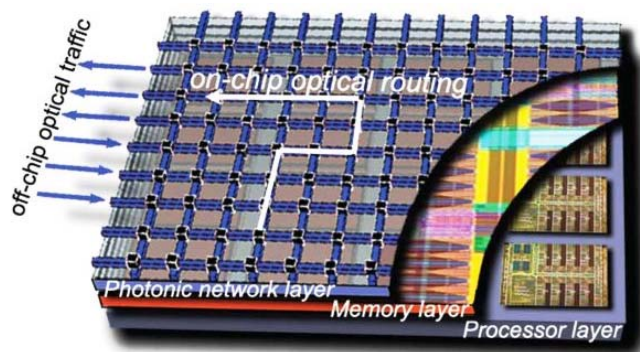


goldenchip.net



4. 3D (& 2.5D) Stacking & Silicon Photonics

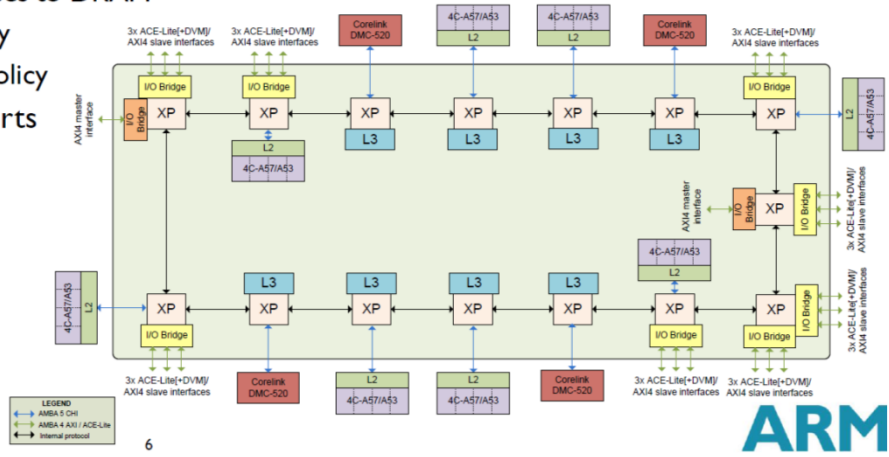
Multiple integration with 3D stacking...



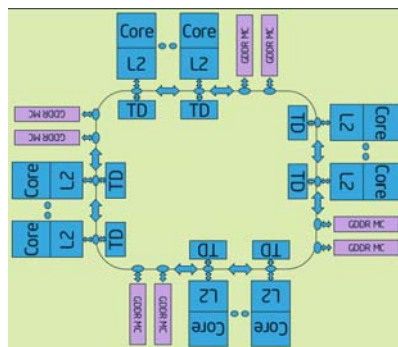
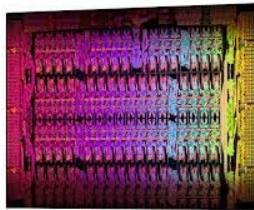
4. Rings from ARM

Access to DRAM

Memory
Policy
Ports



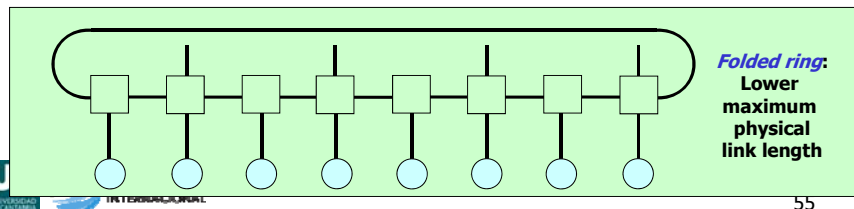
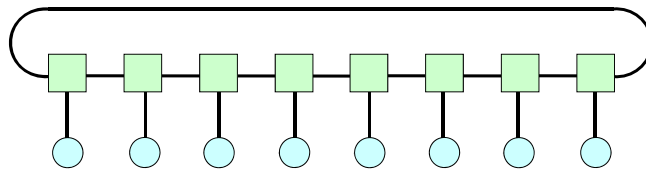
4. Rings from Intel



4. Rings (Direct or Indirect?)

- **Bidirectional Ring networks (folded)**

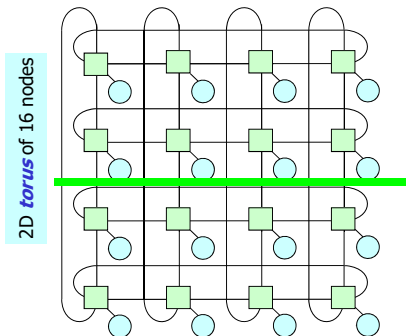
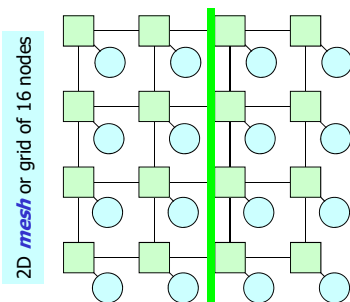
- N switches (3×3) and N bidirectional network links
- Simultaneous packet transport over disjoint paths
- Packets must hop across intermediate nodes
- Shortest direction usually selected ($N/4$ hops, on average)
- Bisection Bandwidth???



55

4. Meshes and Tori

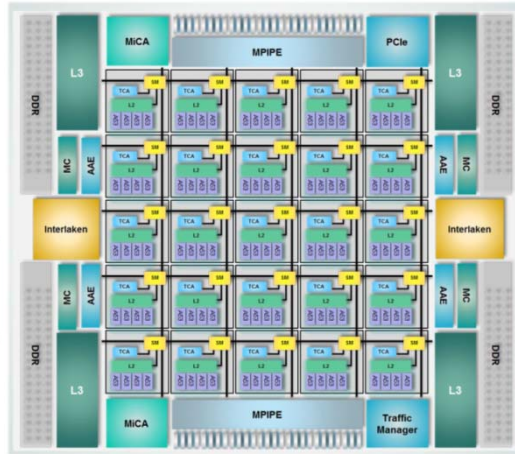
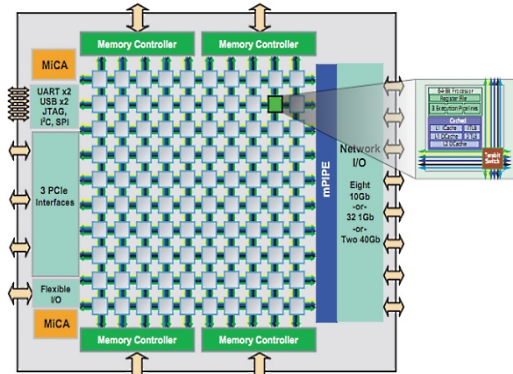
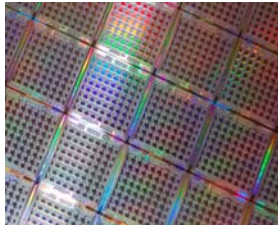
Distributed Switched (Direct) Networks



**Network
Bisection**

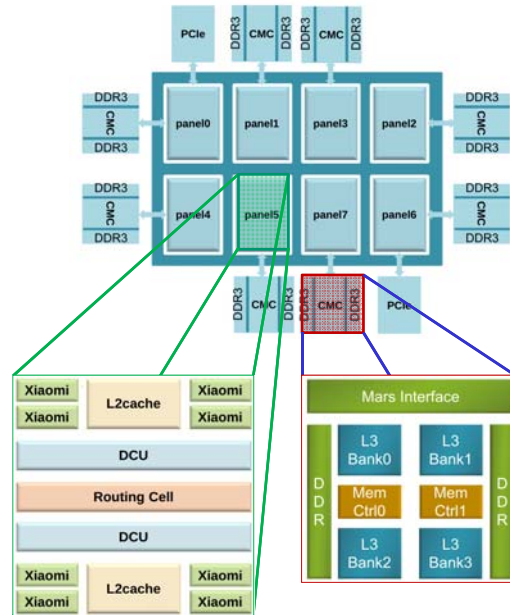
56

4. Meshes from Tileria



4. Mesh from Pythium Mars Architecture

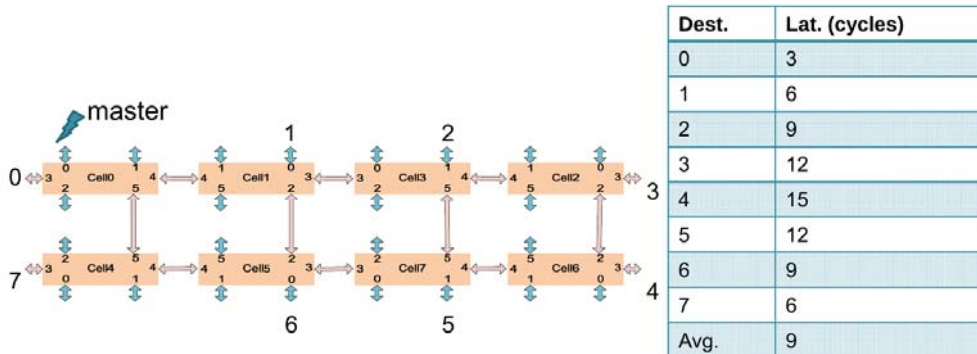
- L1:
 - Separated L1 Icache and L1 Dcache
 - 32 KB Icache
 - 32 KB Dcache
 - 6 outstanding loads
 - 4 cycles latency from load to use
- L2:
 - 16 L2 banks of 4 MB
 - 32 MB of shared L2
- L3:
 - 8 L3 arrays of 16 MB
 - 128 MB of L3
- Memory Controllers:
 - 16 DDR3-1600 channels
- 2x16-lane PCIe-3.0
- Directory based cache coherency
 - 16 Directory Control Unit (DCU)
- MOESI like cache coherence protocol



These images were taken from the slides presented at Hot Chips 2015

4. Pythium Mars NoC

- 6 bi-directional ports switches
- 4 physical channels for cache coherence
- 3 cycles for each hop
- 384 GB/s each cell



This image was taken from the slides presented at Hot Chips 2015

4. Meshes from Intel Knights Landing

Knights Landing: Next Intel® Xeon Phi™ Processor

Intel® Many-Core Processor targeted for HPC and Supercomputing

First **self-boot** Intel® Xeon Phi™ processor that is **binary compatible** with main line IA. Boots standard OS.

Significant improvement in scalar and vector performance

Integration of **Memory on package**: innovative memory architecture for high bandwidth and high capacity

Integration of **Fabric on package**

Three products

KNL Self-Boot	KNL Self-Boot w/ Fabric	KNL Card
(Baseline)	(Fabric Integrated)	(PCIe-Card)



Potential future options subject to change without notice.
All timeframes, features, products and dates are preliminary forecasts and subject to change without further notification.



4. Intel Knights Landing

Knights Landing
Holistic Approach to Real Application Breakthroughs

Platform Memory
NEW Up to **384 GB** DDR4 (6 ch)

Compute

- Intel® Xeon® Processor Binary-Compatible
- 3+ TFLOPS¹, 3X ST²** (single-thread) perf. vs KNC
- 2D Mesh** Architecture
- Out-of-Order** Cores

On-Package Memory

- Over **5x** STREAM vs. DDR4³
- Up to **16 GB** at launch

Omni-Path (optional) ■ **1st** Intel processor to integrate

I/O NEW Up to **36** PCIe 3.0 lanes

Over **60** Cores

Integrated Intel® Omni-Path Processor Package

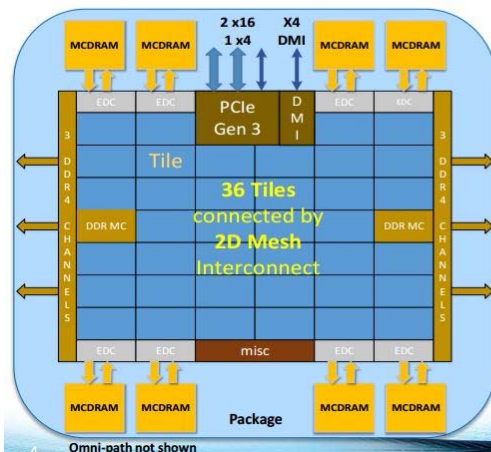
Intel Inside XEON PHI



4. Intel Knights Landing

Knights Landing Overview

TILE		
2 VPU	CHA	2 VPU
Core	1MB L2	Core



Chip: 36 Tiles interconnected by 2D Mesh
Tile: 2 Cores + 2 VPU/core + 1 MB L2

Memory: MCDRAM: 16 GB on-package; High BW
 DDR4: 6 channels @ 2400 up to 384GB

IO: 36 lanes PCIe Gen3. 4 lanes of DMI for chipset

Node: 1-Socket only

Fabric: Omni-Path on-package (not shown)

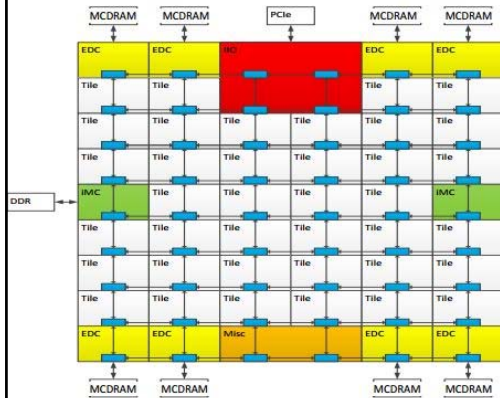
Vector Peak Perf: 3+TF DP and 6+TF SP Flops
Scalar Perf: ~3x over Knights Corner
Streams Triad (GB/s): MCDRAM : 400+; DDR: 90+

Source Intel. All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. KNC data are preliminary based on current expectations and are subject to change without notice. ¹Binary Compatible with Intel Xeon processors using Haswell architecture. ²Single-threaded. ³Bandwidth numbers are based on STREAM-like memory access pattern using MCDRAM. ⁴Up to 36 lanes of DMI for chipset. Results have been estimated based on internal Intel benchmarks and may vary based on system configuration. Any difference in system configuration may affect performance.



4. Intel Knights Landing

KNL Mesh Interconnect



Mesh of Rings

- Every row and column is a (half) ring
- YX routing: Go in Y → Turn → Go in X
- Messages arbitrate at injection and on turn

Cache Coherent Interconnect

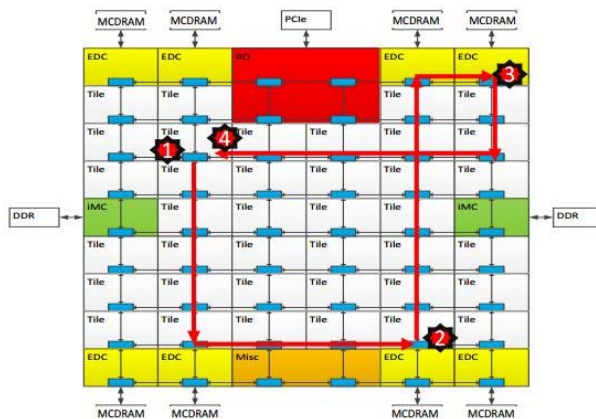
- MESIF protocol (F = Forward)
- Distributed directory to filter snoops

Three Cluster Modes

(1) All-to-All (2) Quadrant (3) Sub-NUMA Clustering

4. Intel Knights Landing

Cluster Mode: All-to-All



Address uniformly hashed across all distributed directories

No affinity between Tile, Directory and Memory

Most general mode. Lower performance than other modes.

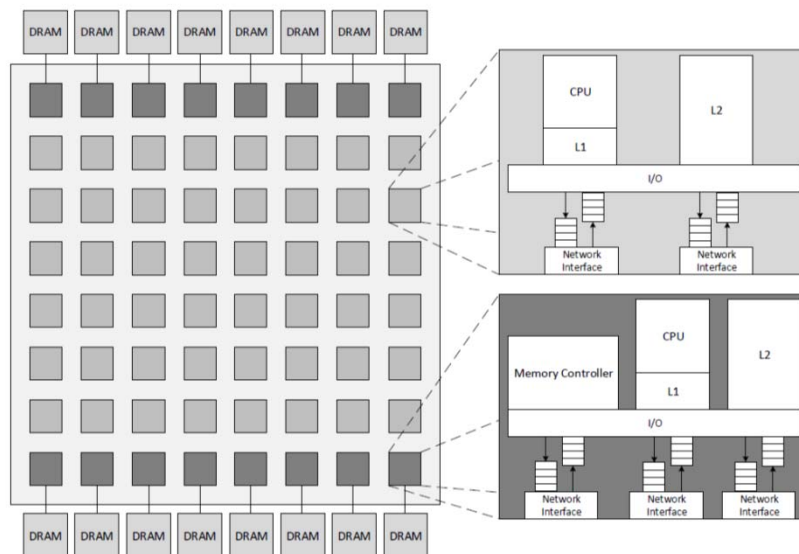
Typical Read L2 miss

1. L2 miss encountered
2. Send request to the distributed directory
3. Miss in the directory. Forward to memory
4. Memory sends the data to the requester

Outline

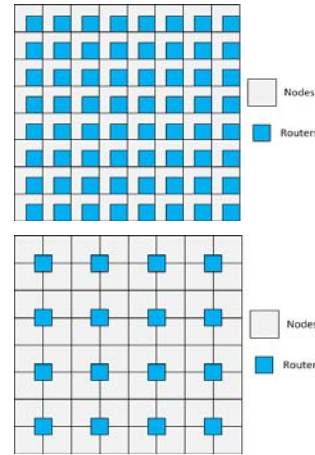
- 1. Introduction
- 2. Network Basis
- 3. System networks
- 4. On-chip networks (NoCs)
- 5. Some current research

5. Some research on NUCA-based CMP Models



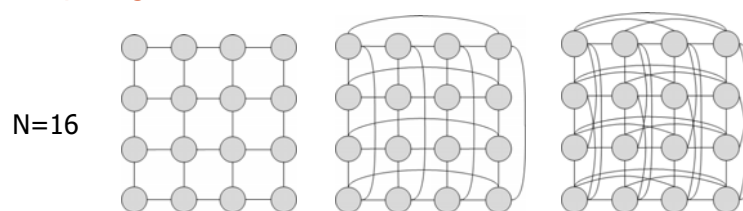
5. Full-system simulation including concentration

GEM5 + BookSim full-system simulation platform parameters	
ISA	X86
Number of Cores	64
CPU Model	Out of Order
CPU Frequency	2 GHz
Cache Coherence Protocol	MESI
L1 Instructions Size	32 KB
L1 Data Size	64 KB
Shared distributed L2	256 KB per Core
# Memory Controllers	4
Network Frequency	1 GHz
Router Pipeline Stages	4
Physical Networks	3
Buffer Size	10 flits
Link Width	64 bits
Topologies	8x8 mesh, torus and FBFLY 4x4 FBFLY with C=4
Applications used	PARSEC benchmarks



5. Topology comparison

Three different topologies are considered:

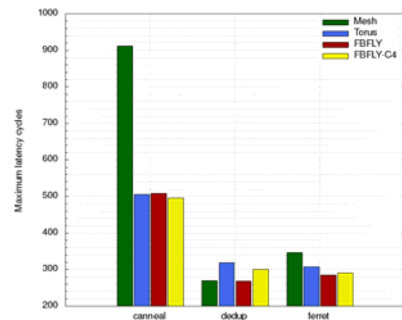
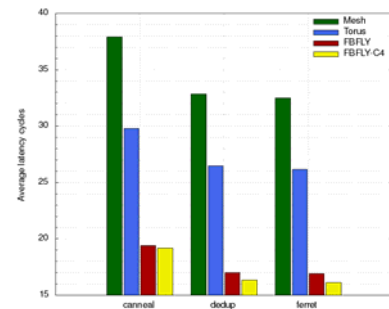
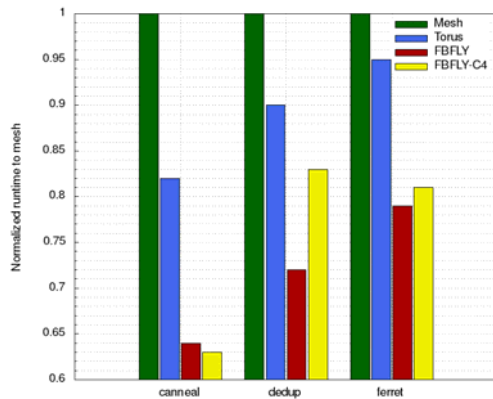


Topology	2D Mesh	2D Torus	2D FBFLY
Degree (ports) ↓	4	4	$2(\sqrt{N} - 1)$
Diameter (max. distance) ↓	$2(\sqrt{N} - 1)$	\sqrt{N}	2
Average distance ↓	$\approx 2\sqrt{N}/3$	$\approx \sqrt{N}/2$	$\approx 2 - 2/\sqrt{N}$
Bisection Bandwidth (links) ↑	\sqrt{N}	$2\sqrt{N}$	$N^{3/2}/4$
Advantages	Low degree Shortest links	Low degree Symmetry Better properties	Symmetry Best properties Larger concentration
Disadvantages	Largest distances Lowest BB	Folding Deadlock	Highest costs Non-uniform link lengths

5. Full-system simulation

Normalized execution time and network latencies:

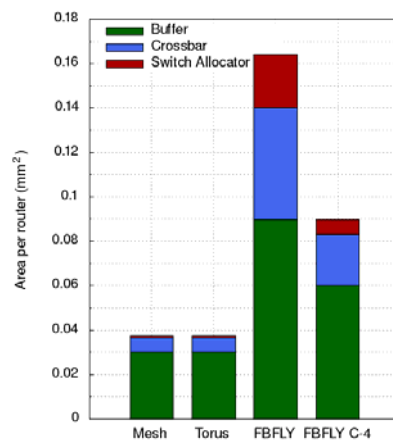
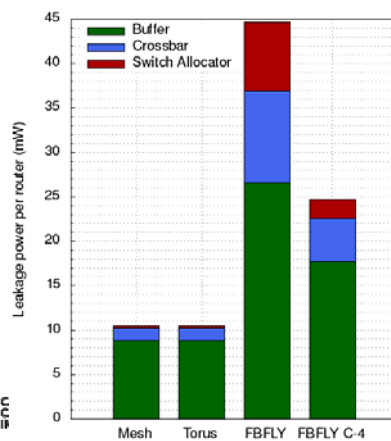
- Average latency has impact in AMAT.
- High latencies can degrade execution times if the affected data are critical.



5. Router Power and Area

Router leakage power and area evaluation:

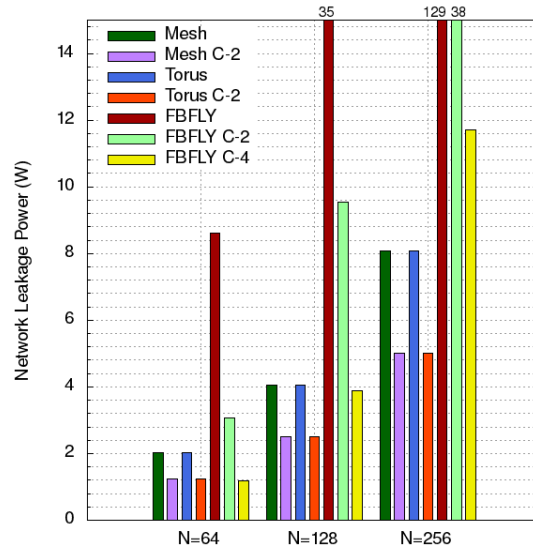
- Buffers are the most consuming part of the router.
- Crossbars and allocators grew quadratically with the number of ports.
- The load in these simulations is low. Hence, the leakage power is the dominant one.



5. Router Power and Area

Network leakage power evaluation:

- FBFLY can manage higher concentrations because its higher BB.



5. OmpSs vs. pThreads

