Mestrado em Informática

2010/11

A.J.Proença

Tema

Arquitecturas Paralelas (3)

Adaptado e estendido de

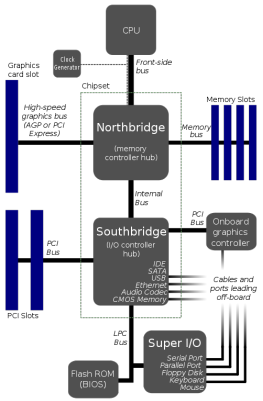
Computer Organization and Design, 4th Ed, Patterson & Hennessy, © 2009, MK

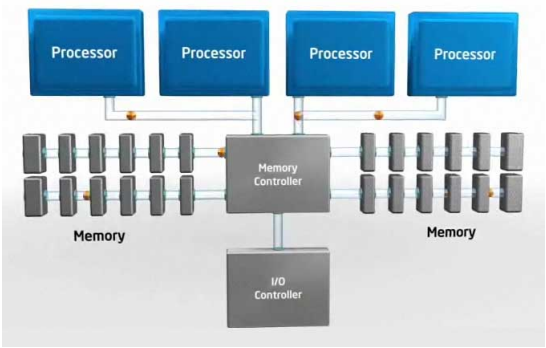
Estrutura do tema AP

1. A evolução das arquitecturas pelo paralelismo
2. Multiprocessadores (SMP e MPP)
3. Data Parallelism: SIMD, Vector, GPU, …
4. Topologias de interligação

External connections on current CPU chips (1)

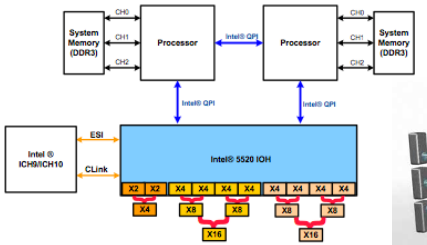
- Motherboards typically have 1-4 CPU chips
  - Intel examples with pre-Nehalem microarchitecture (1 & 4 CPU chips)

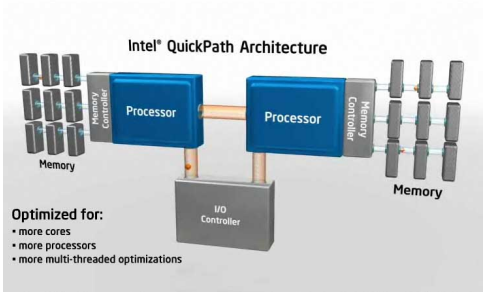




External connections on current CPU chips (2)

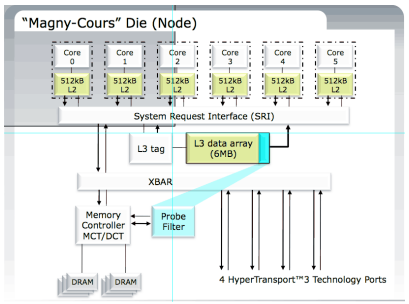
- Motherboards typically have 1-4 CPU chips
  - Intel examples with Nehalem microarchitecture (2 CPU chips)

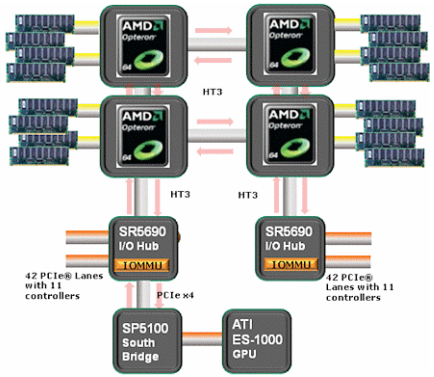




External connections on current CPU chips (3)

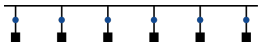
- Motherboards typically have 1-4 CPU chips
  - AMD examples based (1st on half Magny-Cours)



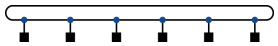


Multiprocessor Network Topologies: Interconnection Networks

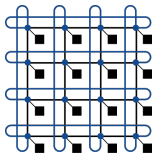
- Network topologies
  - Arrangements of processors, switches, and links



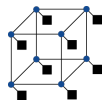
Bus



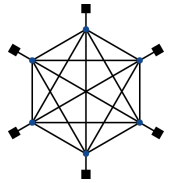
Ring



2D Mesh

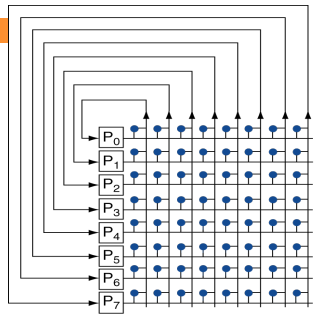


N-cube (N = 3)

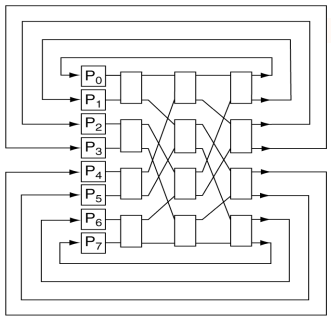


Fully connected

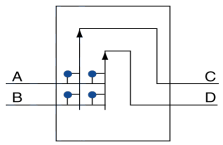
Multistage Networks



a. Crossbar



b. Omega network



c. Omega network switch box

Network Characteristics

- Performance
  - Latency per message (unloaded network)
  - Throughput
    - Link bandwidth
    - Total network bandwidth
    - Bisection bandwidth
  - Congestion delays (depending on traffic)
- Cost
- Power
- Routability in silicon

On-Chip Networks (NoCs)

- Why Network on Chip?
  - Ad-hoc wiring does not scale beyond a small number of cores
    - Prohibitive area
    - Long latency
- NoC offers
  - scalability
  - efficient multiplexing of communication
  - often modular in nature (ease verification)

Slides: Natalie Jerger

Differences between on-chip and off-chip networks

- Off-chip: I/O bottlenecks
  - Pin-limited bandwidth
  - Inherent overheads of off-chip I/O transmission
- On-chip
  - wiring density (function of # metal layers, directionality)
  - router complexity (time, area)
  - need to embed topology in 2D space
    - for topologies > 2D, topological adjacency ≠ physical adjacency

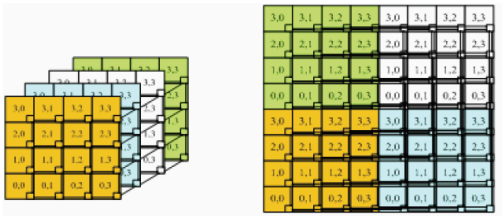
Slides: Natalie Jerger + …

On-chip Network Topologies

- **On-chip topologies will be**
  - **—fixed degree**
  - **—low dimension (1-2) for the forseeable future**
- **Candidate topologies: 2D mesh or torus, rings, and variants**

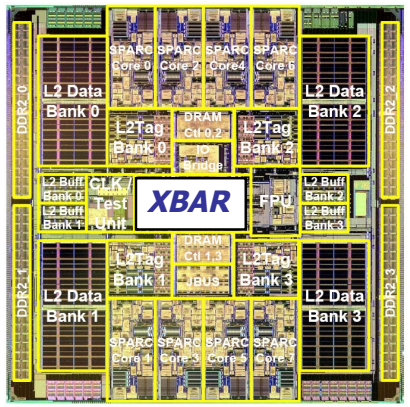
Example: 2D embedding of 64-node 3D mesh

network diameter = 9, but tile span = 18

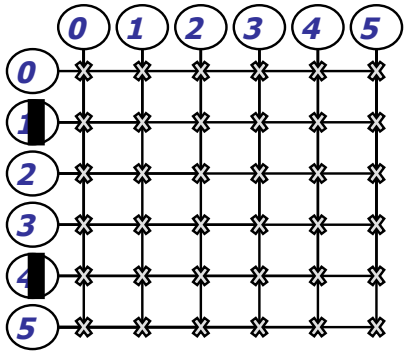


Slides: John Mellor-Crummey, Rice University

NoC on Multicore: Examples (1)

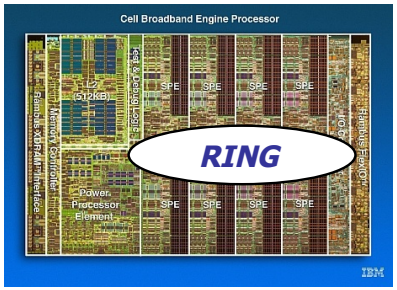


Sun Niagara



Slides: Natalie Jerger

NoC on Multicore: Examples (2)

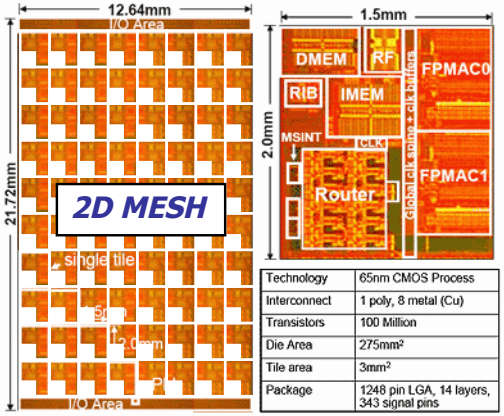


IBM Cell

- Element Interconnect Bus
  - 4 rings
  - Packet size: 16B-128B
  - Credit-based flow control
  - Up to 64 outstanding requests
  - Latency: 1 cycle/hop

Slides: Natalie Jerger

NoC on Multicore: Examples (3)



Technology	65nm CMOS Process
Interconnect	1 poly, 8 metal (Cu)
Transistors	100 Million
Die Area	275mm²
Tile area	3mm²
Package	1248 pin LGA, 14 layers, 343 signal pins

- Intel Polaris
  - 80 core prototype
- Academic Research ex:
  - MIT Raw, TRIPs
    - 2-D Mesh Topology
    - Scalar Operand Networks

Slides: Natalie Jerger

Some Alternative Topologies

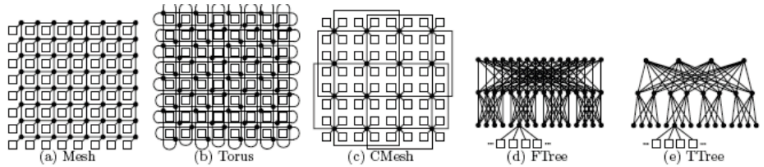


Figure 8: Network Topologies

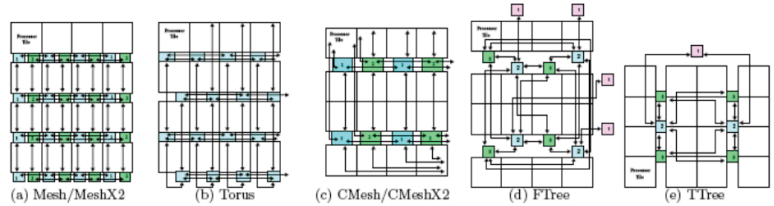


Figure 9: Floorplan for Estimating Area (Lower Left Quadrant of Die Shown)

Balfour and Daily, Design Tradeoffs for Tiled CMP Networks, ICS 2006

Next generation On-Chip Networks

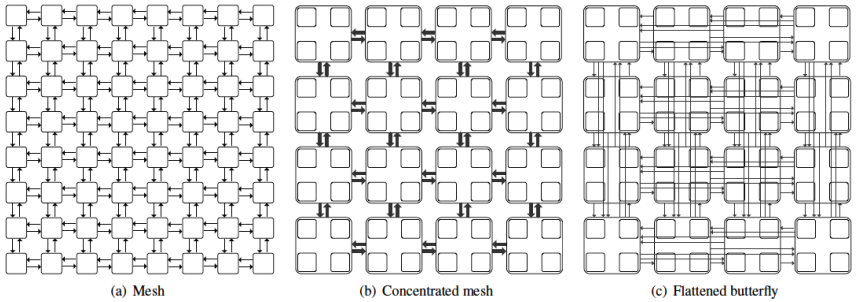


Figure 1. Mesh, Concentrated Mesh and Flattened Butterfly topologies for a 64-node network.

Boris Grot and Stephen W. Keckler, "Scalable On-Chip Interconnect Topologies",
2nd Workshop on Chip Multiprocessor Memory Systems and Interconnects, Beijing, China, 2008