Computer Architecture
计算机体系结构

Lecture 11. Interconnection Networks
第十一讲、片上互联网络简介

Chao Li, PhD.
李超 博士

SJTU-SE346, Spring 2019

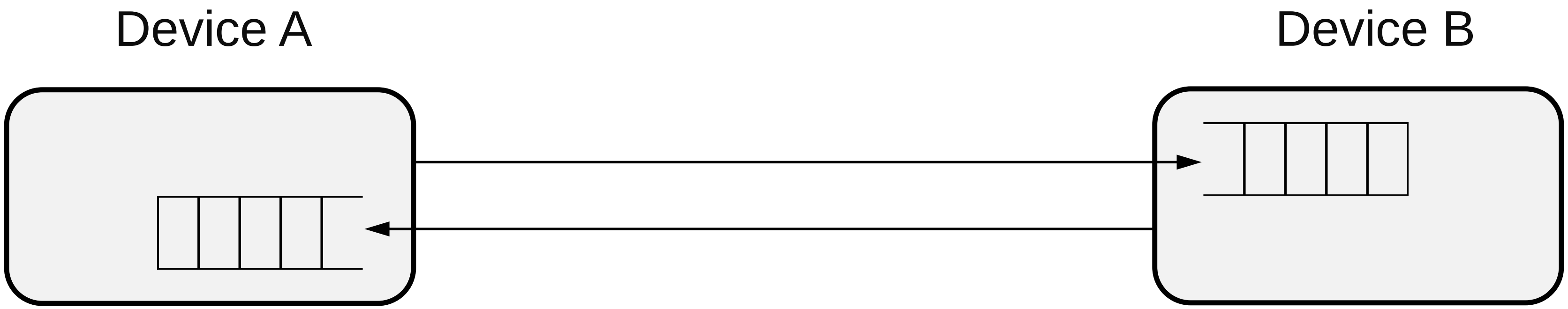
Review

- Data-level parallelism
- Vector processor and vector instruction
- VMPIS components, DAXPY, execution latency
- SIMD lanes, chaining, vector length register
- GPGPU, CUDA programming model
- TPC, SM, SP, warp scheduling, branch divergence
- CPU-GPU system

Outlines

- Introduction and Terminology
- Interconnection Topologies

A Simple Dedicated Link Network



The transmitter, link, and receiver collectively form a channel

- **Links**:
  - Bundle of wires that carries a signal
- **Channel**:
  - A single link between host or switch elements
- **Buffer**:
  - To hold data as it is being transferred

Terminology

- **Node:**
  - A network endpoint connected to a router/switch
- **Switch/Router**:
  - Connects a fixed number of input channels to a fixed number of output channels; this number is called **switch degree or radix**
- **Route/Path**:
  - A sequence of channels and switches; the number of switches through which packets had to traverse is referred to as **hop count**
- **Message**:
  - The unit of information sent or received by network clients; it can be broken into a sequence of **packets**
- **Network Interface**:
  - Compose and process messages

Network Characters

- A network is generally characterized by its
  - **Topology:**
    - Describes the physical interconnection structure of the network graph
  - **Routing algorithm:**
    - Determines which routes messages may flow through the network graph
  - **Switching strategy:**
    - Determines how the data in a message traverses its route
  - **Flow control mechanism:**
    - Determines when the message, or portions of it, move along its route

Properties of Network Topology

- **Diameter**
  - The maximum shortest path between any two nodes

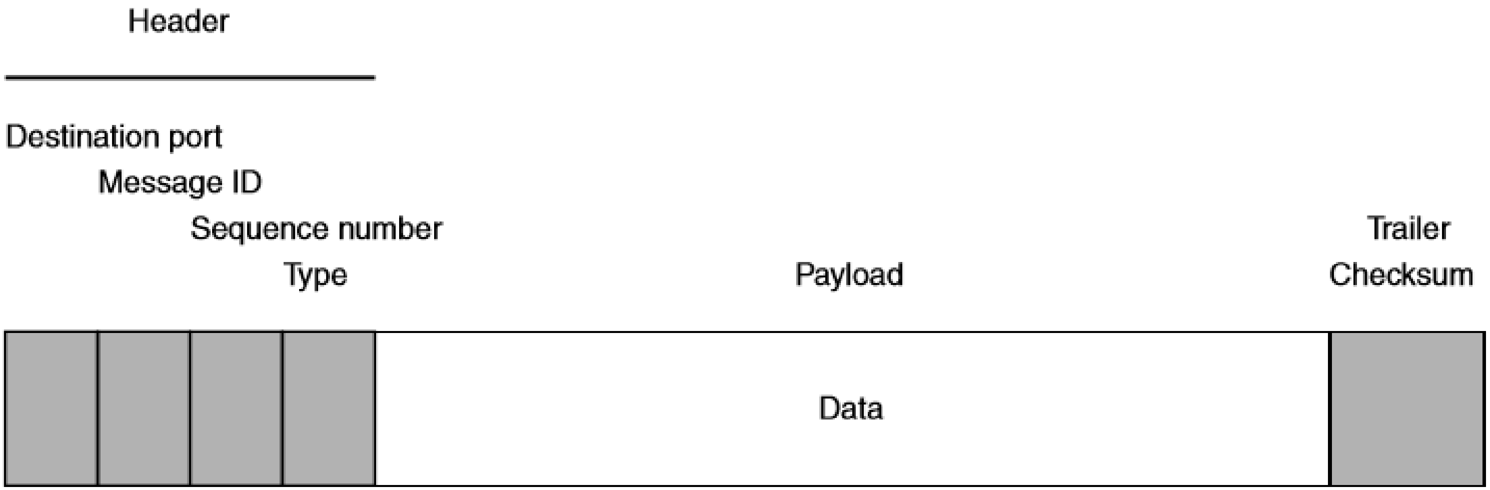
- **Routing distance**
  - Number of links/hops along the route

- **Average Distance**
  - Average number of hops across all valid routes

- **Non-Blocking (Blocking)**
  - Can connect any idle input to any idle output
  - Blocking arises due to paths sharing one or more links

Packet Format

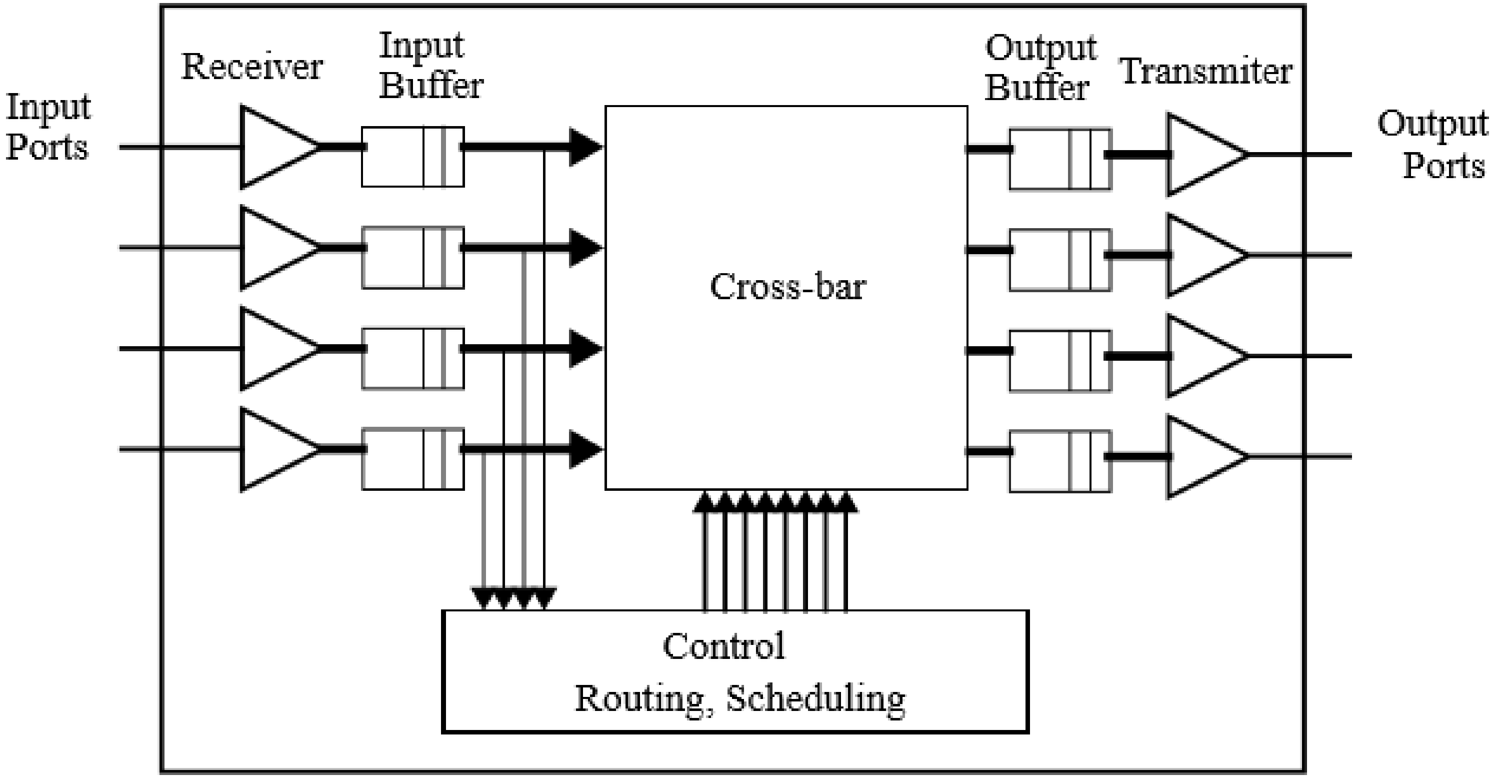
- **Packet**:
  - Consists of a header, a data payload, and a trailer



- The minimum unit of information that can be transferred across a link is called a flow control unit: ***flit***

Network Switch

- Network switches implement the routing, arbitration, and switching functions of switched-media networks



internal structure of a network switch

Switching Strategy

- **Switching strategy:**
  - Determines how the data in a message traverses its route
  - e.g. **circuit switching** establishes a dedicated communications channel (circuit)
  - e.g. **packet switching**: divides the data into packets that share the network

- **Circuit vs. Packet:**
  - Circuit switching has better bandwidth
  - Circuit switching has longer setup time

Performance Evaluation

- Link width: $\boldsymbol{w}$
- Unit interval: $\boldsymbol{\tau}$
- Signaling rate: $\boldsymbol{f} = \boldsymbol{1} / \boldsymbol{\tau}$
- Channel bandwidth: $\boldsymbol{b} = \boldsymbol{w} \cdot \boldsymbol{f}$
- Total bandwidth of all the channels (or links)
  - the number of channels times the bandwidth per channel

Latency (Lower Bound)

- Sending overhead: $\boldsymbol{Overhead_s}$
- Receiving overhead: $\boldsymbol{Overhead_r}$
- Total routing time: $\boldsymbol{T_R}$
- Arbitration time: $\boldsymbol{T_A}$
- Switching time: $\boldsymbol{T_S}$
- Total time of flight of the packet $\boldsymbol{T_{TotalProp}}$

$$\boldsymbol{Latency =}$$

$$\boldsymbol{Overhead_s + (T_{TotalProp} + T_R + T_A + T_S) + \frac{Packet\ size}{Bandwidth} + Overhead_r}$$

Outlines

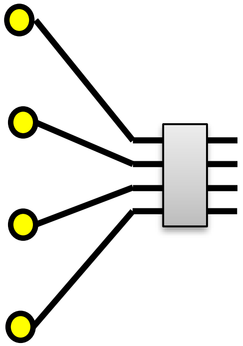
- **Introduction and Terminology**
- **Interconnection Topologies**

Switch Network Topologies

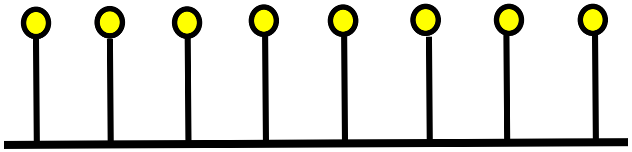
- View switched network as a graph
  - Vertices: nodes or switches
  - Edges: communication paths
- Describes the structure of the network graph
  - e.g. Direct network have a host node connected to each switch
    - a.k.a. distributed switch
  - e.g. Indirect network have hosts connected only to certain switches
    - a.k.a. centralized switch
- Regular vs Irregular
  - Regular network are widely used graphs such as grid and tree, etc.

Bus

- Connects all inputs to all outputs using a single switch ?
  - Diameter is 1
  - Degree is *N*
  - No fault tolerance: single point of failure
  - Low performance: bandwidth is *O*(1)

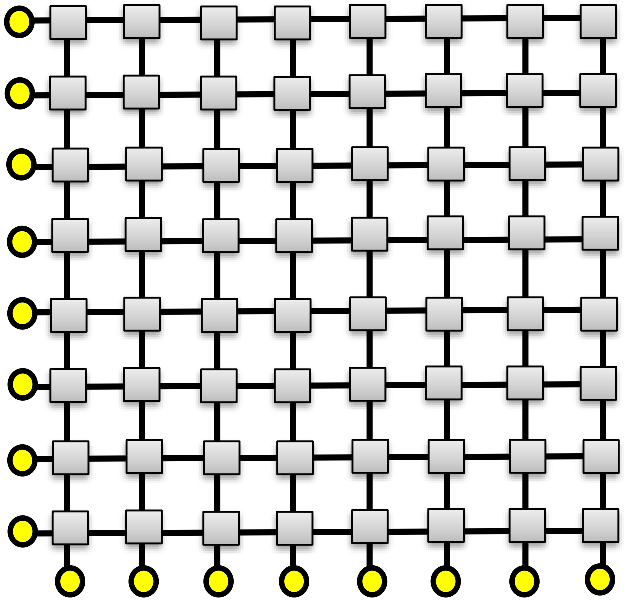


- Bus: the interconnect is mostly just wires
  - Only one transaction in progress at a time
  - Frequency affected by physical limitations



Crossbar

- Crossbar switch
  - A type of fully-connected network
  - Every node connected to all others
  - O(N) bandwidth
  - Cost of interconnect: O(N^2）
  - Good for small number of nodes

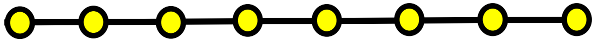


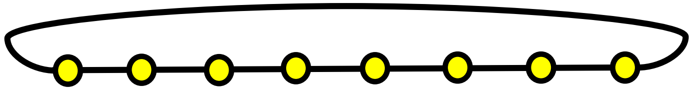
- Crosspoint switch complexity increases quadratically with the number of ports

- Multistage interconnection networks reduces complexity

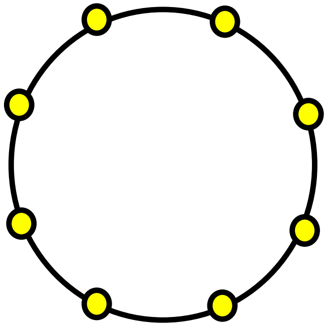
Linear Arrays and Rings

- Array
  - Connected by bidirectional links
  - Diameter is *N-1*
  - Average distance is about $^{2}/_{3}\ N$
  - Bisection width is 1





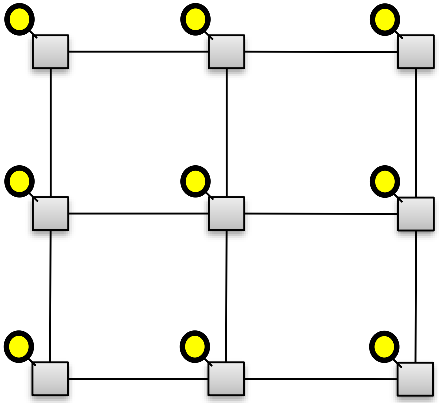
- Ring
  - adding end-round direct connections
  - Connecting the two ends of an array
  - Assume unidirectional links:
    - Diameter is N-1
    - Average distance is N/2
    - Bisection width is 1 link



Multidimensional Topology: 2D Mesh

- Arrays and rings generalize naturally to higher dimensions

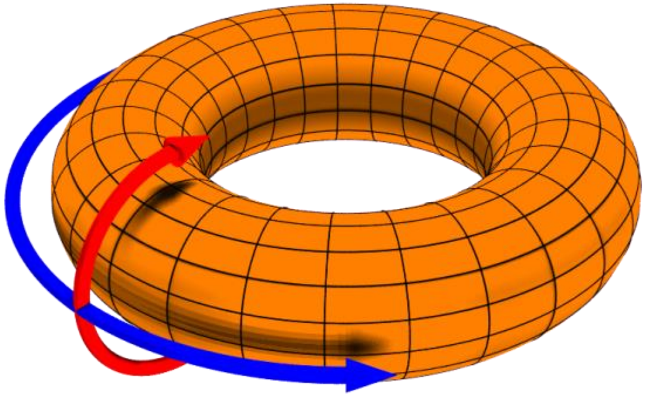
- 2D Mesh
  - Nodes are connected as a grid
  - Multiple routes between a pair of nodes
  - Non-uniform node-to-node latency
  - Cost of interconnect: *O*(N)

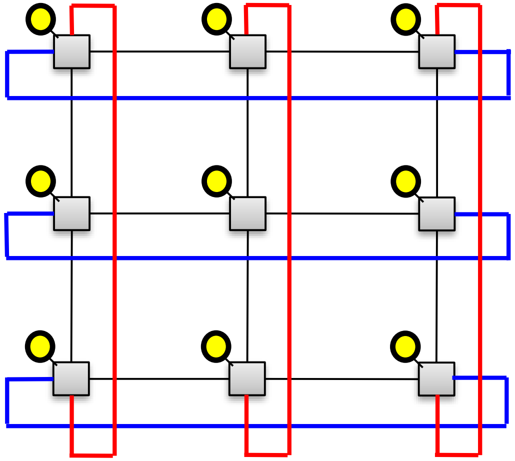


- Example:
  - Intel Paragon 2D mesh (64x64)

Multidimensional Topology: 2D Torus

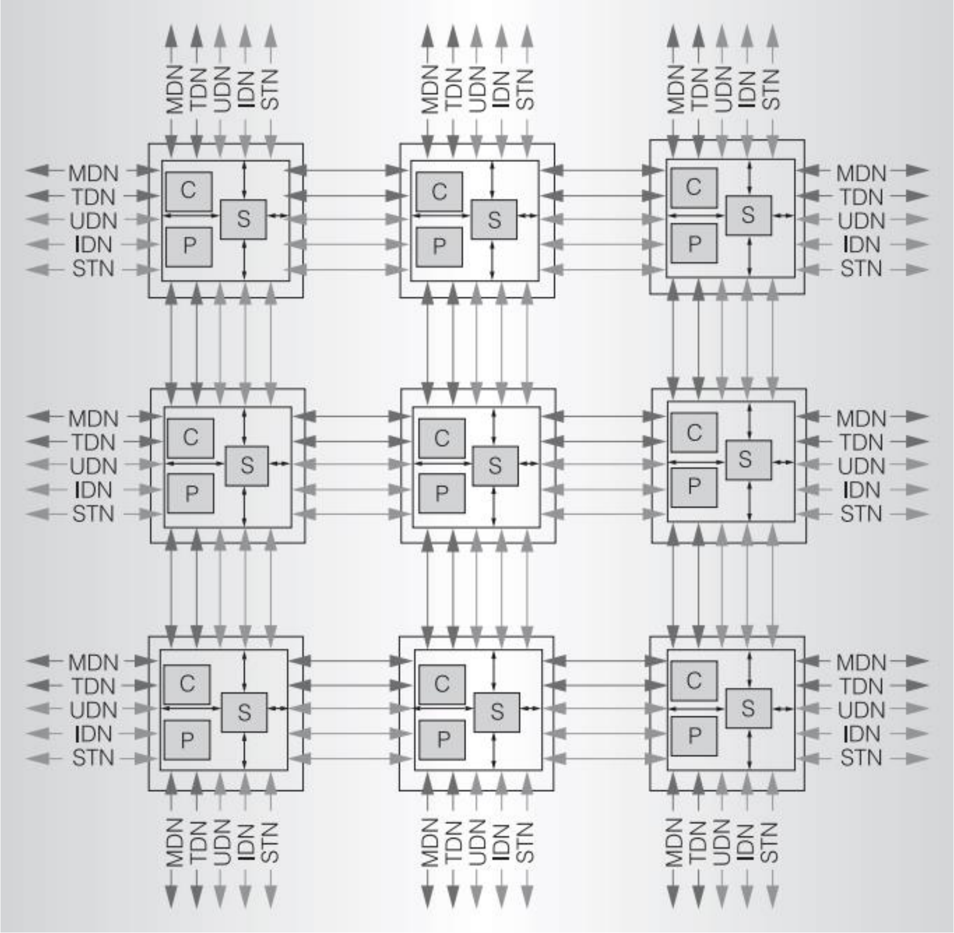
- Reduces the diameter of a mesh network
- Torus: adding end-round direct connections
  - An extension of the 2D mesh
  - A regular torus has long warp-around links
  - Slightly lower latency while higher cost



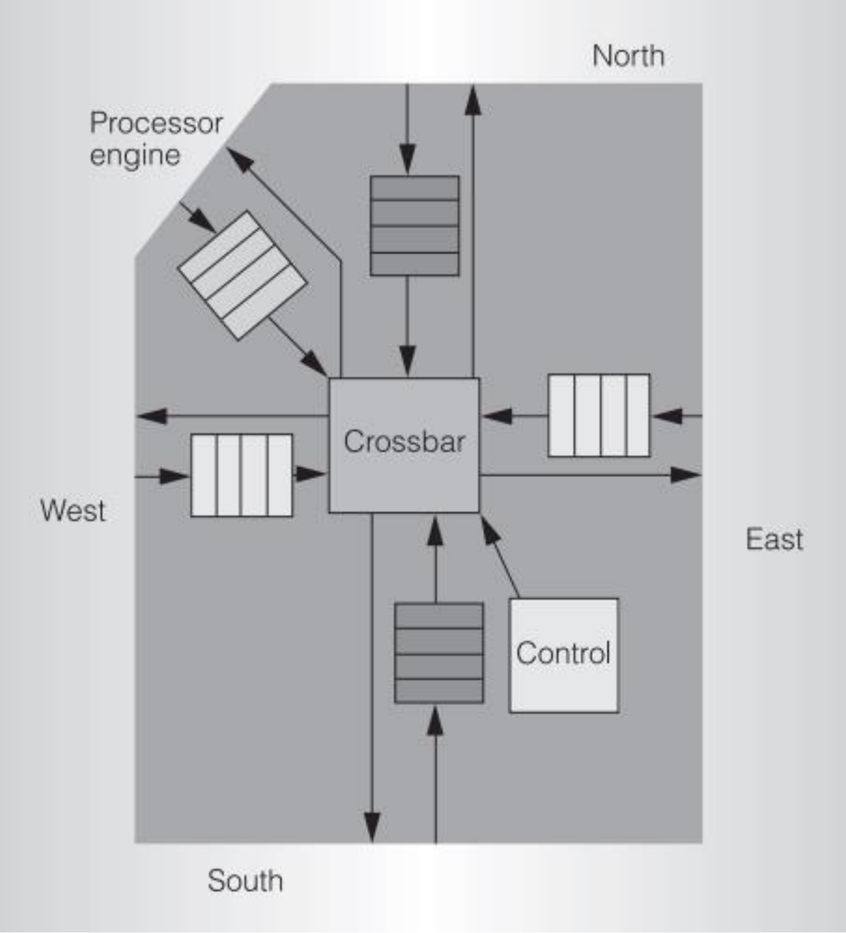


Case Study: Tilera's iMesh

- Coupes tiles (cores) with five 2D mesh on-chip networks



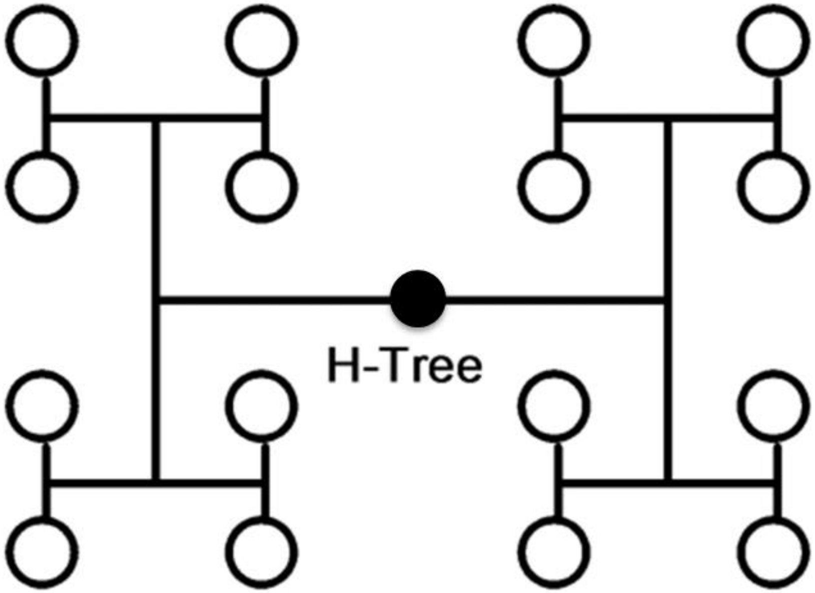
A 3x3 grid of tiles



A single network crossbar

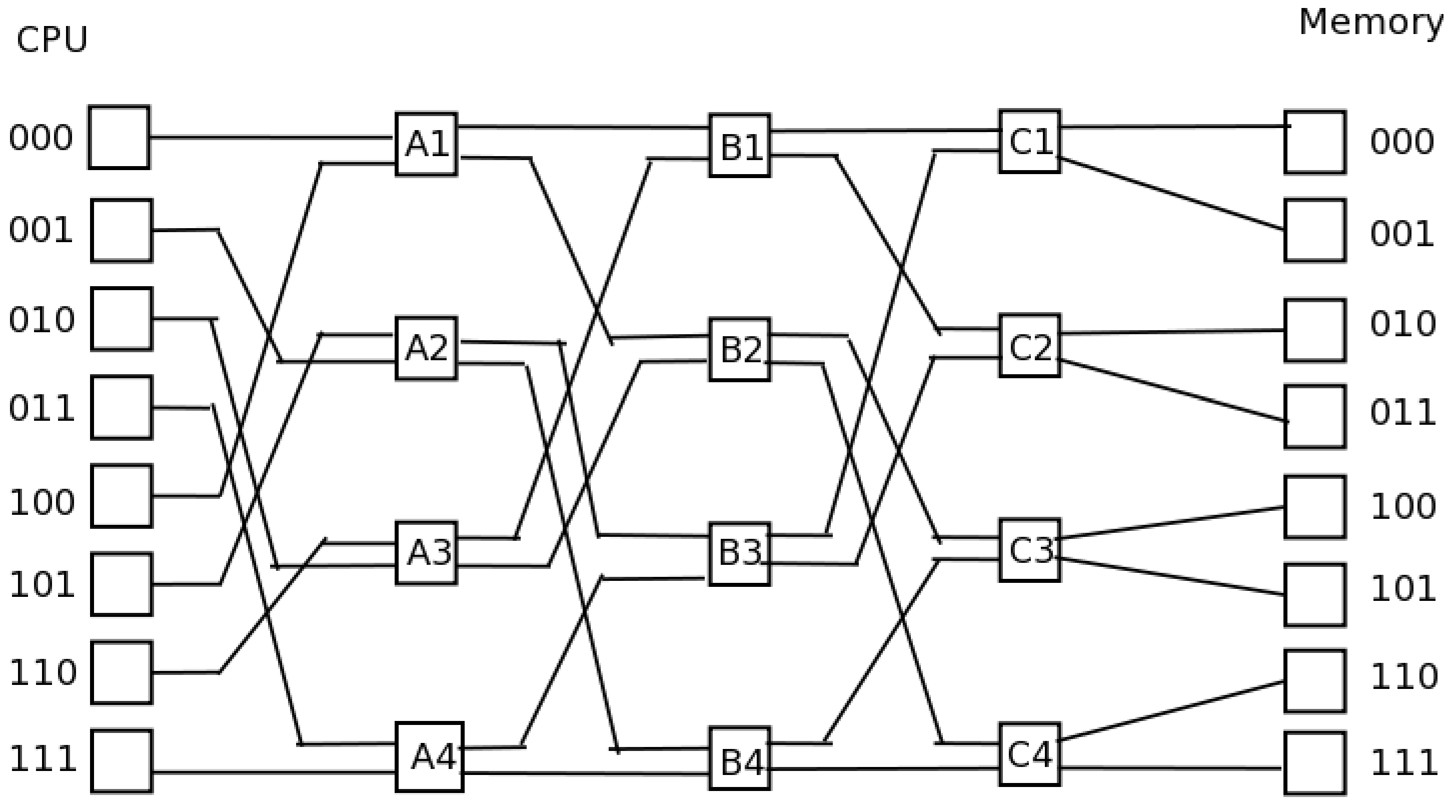
Trees (Optional)

- Trees features planar, hierarchical topology
- Employed as indirect networks with hosts as leaves
- Routing distance grows only logarithmically



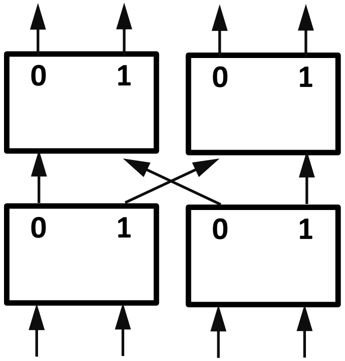
Multistage Interconnection Network (Optional)

- Indirect networks with multiple layers of switches
- Omega network:
  - Log(N) number of stages and N/2 switching units

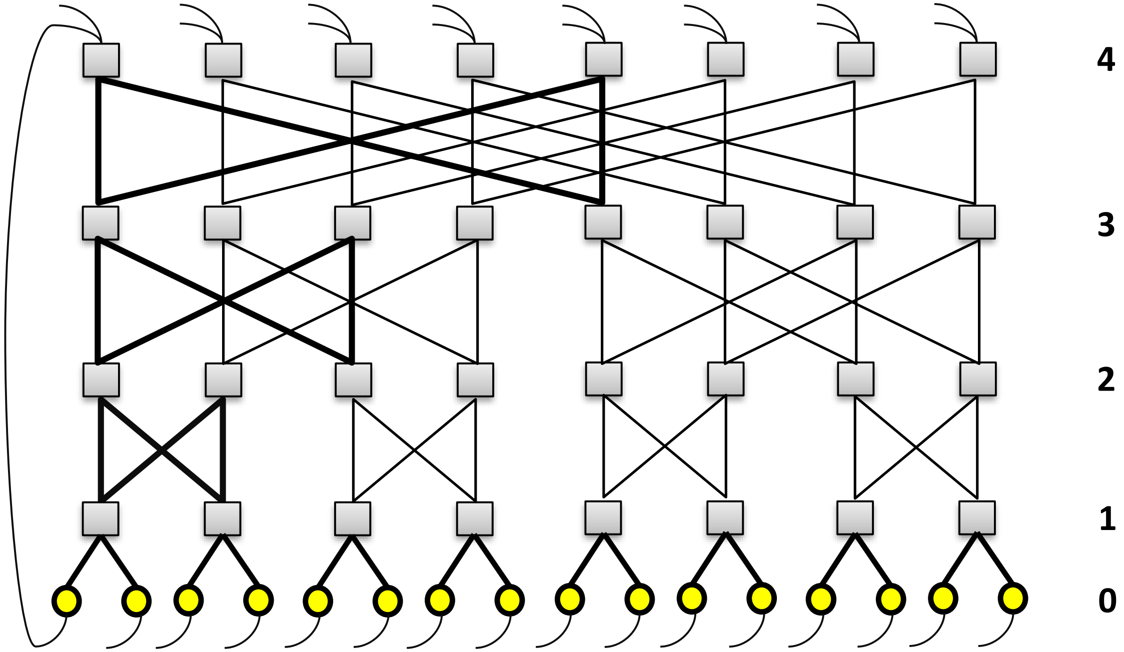


Butterfly Topology (Optional)

- Butterfly is an important logarithmic network
  - Can be viewed as a tree with multiple roots
- A d-dimensional indirect butterfly:
  - Connects $N = 2^d$ nodes (d≥2)
  - $d = \log_2 N$ levels of switches

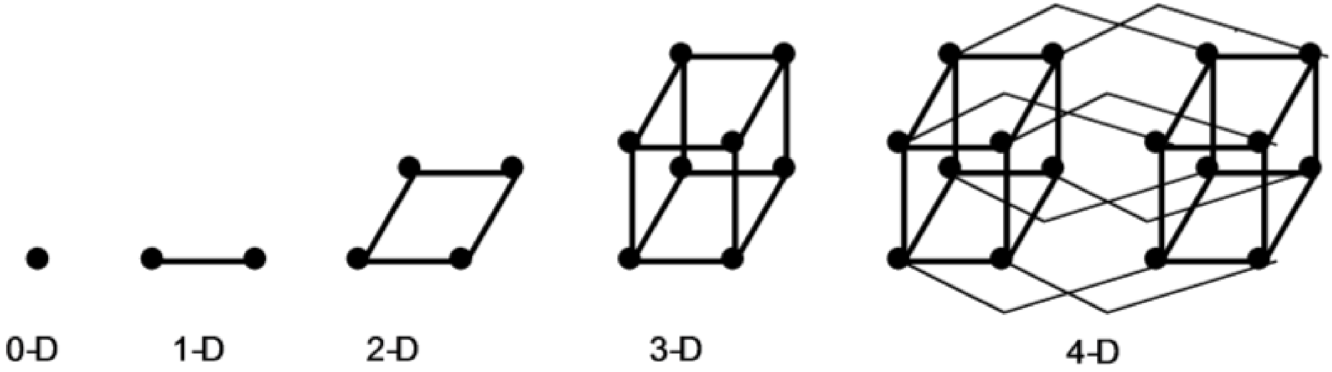


Basic butterfly building block



Hypercube (Optional)

- Also known as binary d-cubes $\mathrm{N} = 2^d$
- Switch degree equals dimension: $d = \log_2 N$
- Good bisection bandwidth
- ***k*-ary *d*-cube** is a d dimensional torus with k elements along each dimension: $\mathrm{N} = k^d$
  - Each node is addressed by a d-vector



Summary

- Basic concepts: link/channel/buffer
- Switch degree, average distance
- Non-blocking network, direct/Indirect network
- Network performance, latency estimation
- Network switch and switching strategy
- Bus and crossbar
- Array, ring, mesh, torus, tree, butterfly, hypercube

References

- 课本参考：J. Hennessy, D. Patterson. Computer Architecture, Fifth Edition: A Quantitative Approach.
  - Appendix F.

- 其它参考：D. Culler et al., 《Parallel Computer Architecture: A Hardware/Software Approach》, Second Edition.
  - Chapters: 10.1, 10.2, 10.3, 10.5

- 课外阅读：D. Wentzlaff et al., On-chip Interconnection Architecture of the Tile Processor, IEEE Micro