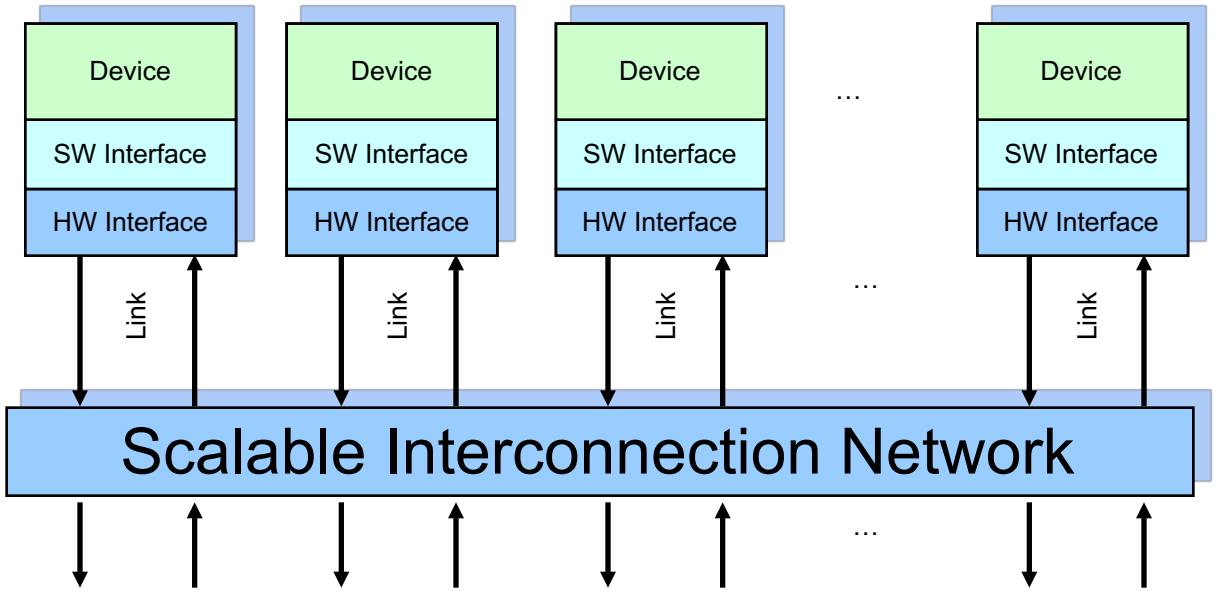
Interconnection Networks

Connecting Multiple Devices

- Connecting multiple devices
- Performance metrics
- Shared vs. switched media
- Network characteristics
- Network topologies
- Torus topologies
- Performance and cost comparisons
- Network routing
- Flow control and switching

Connecting Devices

- Devices:
  - Components within a computer (CPUs, memory, cache, I/O, …)
  - Computers (single board, workstations, supercomputers, …)
  - Systems of computers (grid, clusters, WSC, …)
- Types of elements:
  - End nodes, links, interconnection network, internetworking



New Computer Architecture

- Networking affects the scalability of the system.
  - How many and how powerful end nodes?
  - How fast and scalable network?
  - Performance vs. cost, bandwidth vs. latency
  - Energy efficiency
    - Computation vs. communication
- New computer architecture
  - Traditional architecture + interconnection

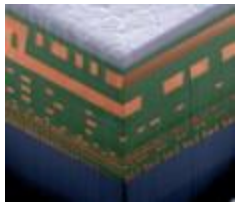
Network Domains

WANs

LANs

SANs

OCNs



Distance (meters)

10^6

10^3

1

10^{-3}

1 10 100 1,000 10,000 >100,000

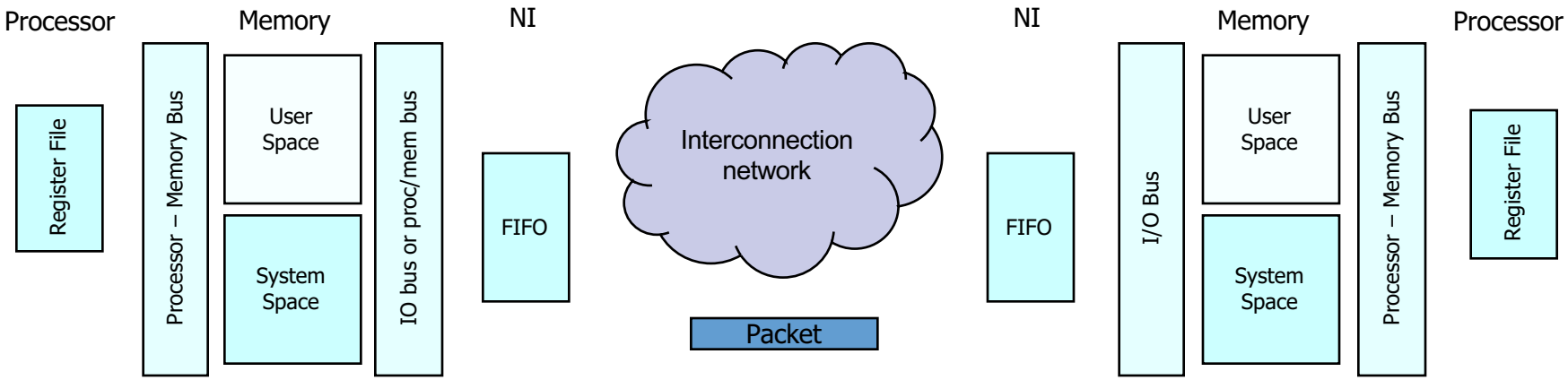
Number of devices interconnected

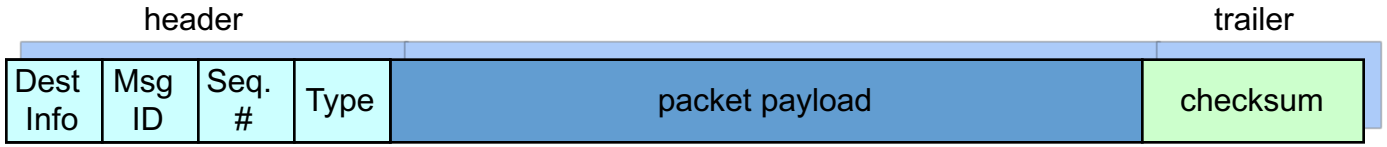
OCN, SAN, LAN, WAN

- On-chip networks (or NoCs)
  - Microarchitectural elements (CPUs, cache, registers, …)
  - 100s of devices in the order of centimeters, metal layer
  - IBM Cell, Sun's Niagara, Intel Teraflops
- System/storage area networks:
  - Multiprocessor and multicomputer systems
  - 100s–1000s of devices on the order of 10s of meters, copper wire
  - InfiniBand (120 Gbps), Myrinet
- Local area networks:
  - Interconnect autonomous computer systems,
  - 100s of devices within a few kilometers
  - Ethernet, copper
- Wide area networks:
  - Across the globe, many millions of devices interconnected
  - ATM, optical

Networking of Two Devices

- Two end-node devices on a dedicated link
  - Basic protocol: request, reply
  - Integrated with operating system (OS)





00 = request
01 = reply
10 = request acknowledge
11 = reply acknowledge

Networking Basics

- Performance metrics
  - Transmission time, transport latency, overhead
- Topology
  - How switches are wired
  - Affects routing, reliability, throughput, latency
- Routing
  - How a message travels
- Buffering and flow control
  - What to store within the network
  - How to throttle during oversubscription

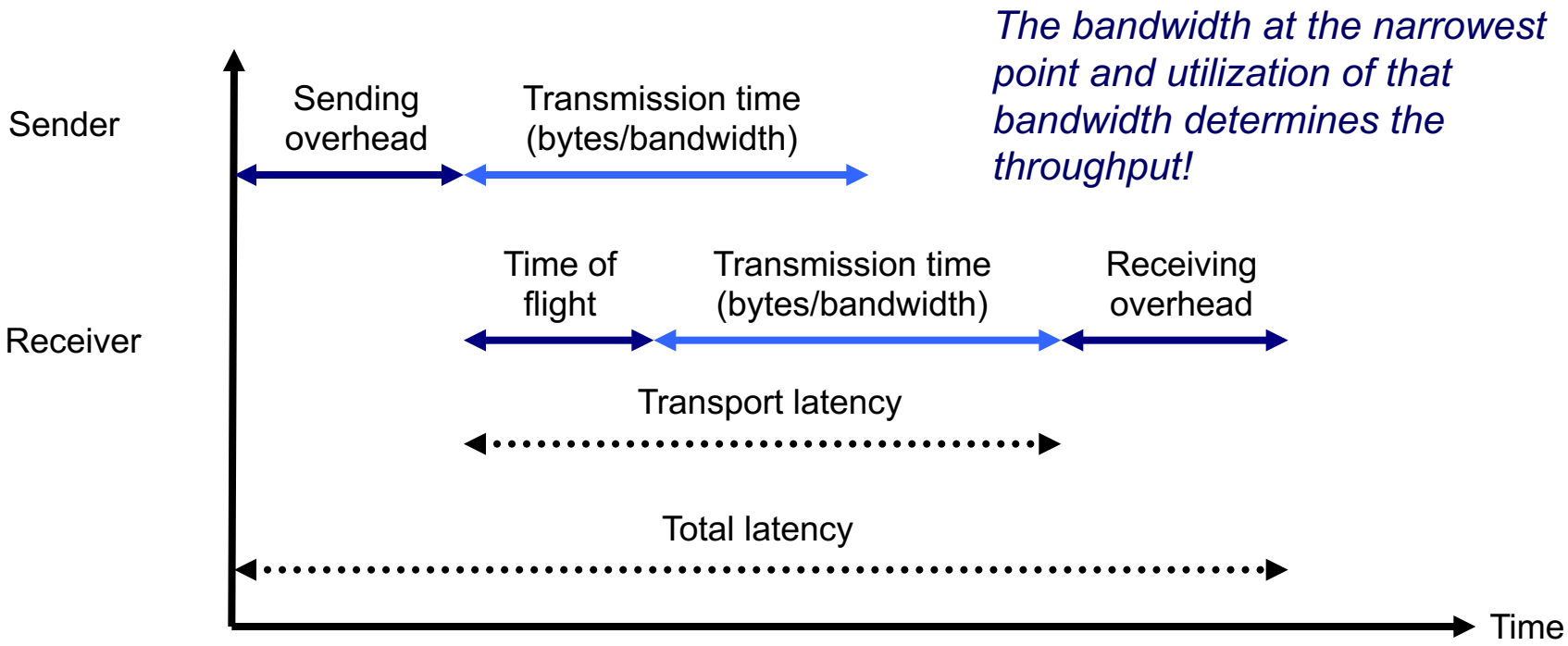
ENGINEERING@SYRACUSE

Performance Metrics

Performance Metrics

- Transmission time
  - The time to pass through the network, not including the time of flight
  - Equal to the packet size divided by the data bandwidth of the link
- Transport latency
  - Sum of the time of flight and the transmission time
  - Measures the time that a packet spends in the network
- Sending overhead (latency)
  - Time to prepare a packet for injection, including hardware/software
  - A constant term (packet size) plus a variable term (buffer copies)
- Receiving overhead (latency)
  - Time to process an incoming packet at the end node
  - A constant term plus a variable term
  - Includes cost of interrupt, packet reorder, and message reassembly

Effective Bandwidth

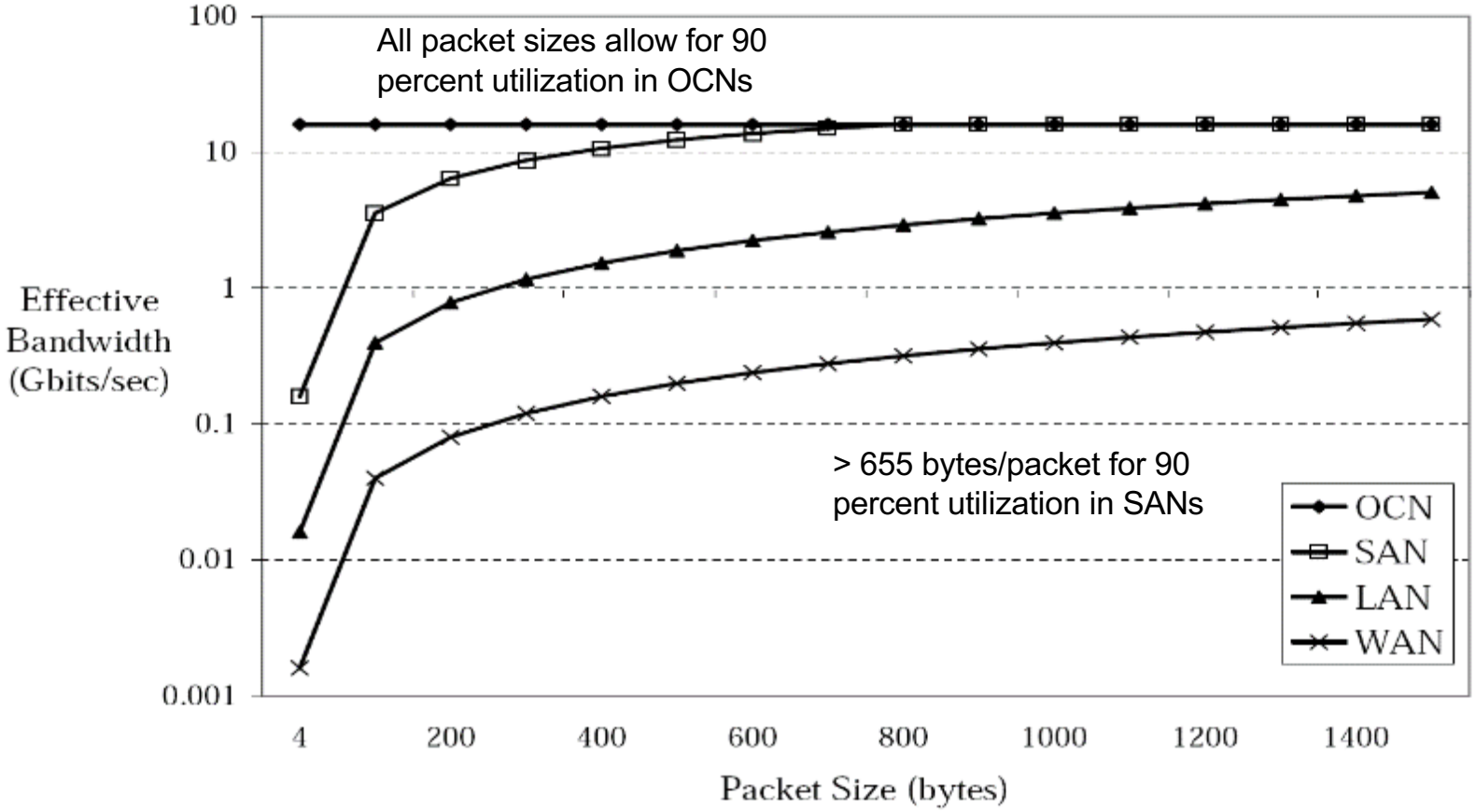


The bandwidth at the narrowest point and utilization of that bandwidth determines the throughput!

Latency = sending overhead + time of flight + packet size/bandwidth + receiving overhead

Effective bandwidth = min ($BW_{NetworkInjection}$, $BW_{Network}$, $BW_{NetworkReception}$)

Effective Bandwidth (cont.)



Transmission time is the limiter for OCNs.

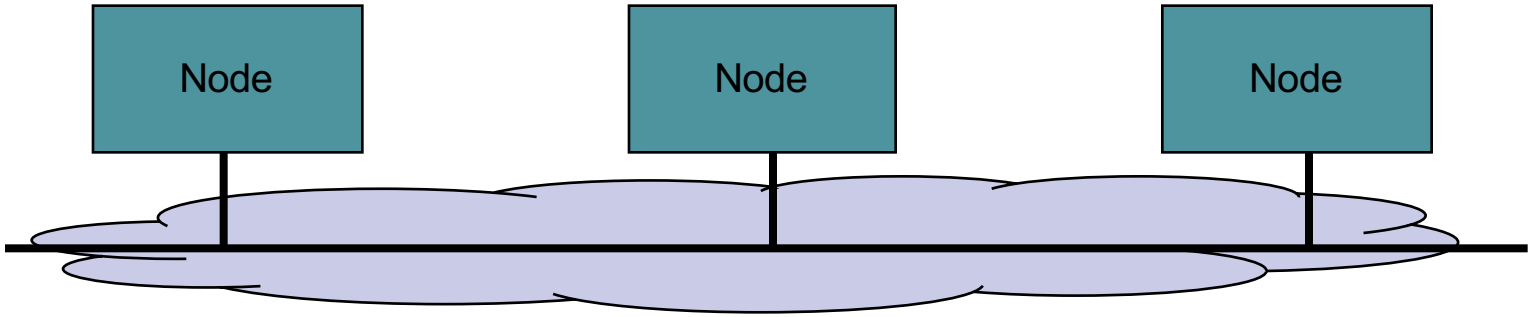
Overhead limits SANs for packets sizes < 800 bytes.

ENGINEERING@SYRACUSE

Shared vs. Switched Media

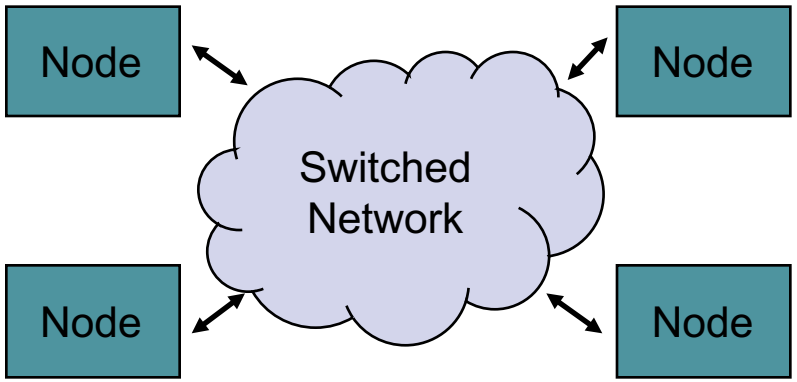
Shared Media Networks

- Network media is shared by all the devices.
  - Broadcasting network: bus, crossbar, …
  - Half-duplex or full-duplex
- Arbitration:
  - Small scale: dedicated control lines
  - Large scale: carrier sensing + collision detection + retransmission
- Straightforward switching and routing.



Switched Media Networks

- Disjointed parts of the media are shared via switching.
- Switch fabric components.
  - Passive point-to-point links
  - Active switches
- Dynamically establish communication between sets of source-destination pairs.
- Aggregate bandwidth can be many times higher than that of shared-media networks.



Switched Media Networks (cont.)

- Routing
  - Every time a packet enters the network, it is routed.
- Arbitration
  - Centralized or distributed
  - Resolves conflicts among concurrent requests
- Switching
  - Once conflicts are resolved, the network “switches in” the required connections.
- Established order: routing, arbitration, and then switching

Shared vs. Switched

- Shared-media networks
  - Low cost.
  - Aggregate network bandwidth does not scale with number of devices.
  - Global arbitration scheme required (a possible bottleneck).
  - Time of flight increases with the number of end nodes.
- Switched-media networks
  - Aggregate network bandwidth scales with number of devices.
  - Concurrent communication.
  - Potentially much higher network effective bandwidth.

ENGINEERING@SYRACUSE

Network Characteristics

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

Network Topology

- One switch suffices to connect a small number of devices.
  - Number of switch ports limited by VLSI technology, power consumption, packaging, and other such cost constraints.
- A fabric of interconnected switches is needed when the number of devices is much larger.
  - The topology must make a path(s) available for every pair of devices—property of connectedness or full access.
- Topology defines the connection structure across all components.
  - Bisection bandwidth: the minimum bandwidth of all links crossing a network split into two roughly equal halves
- Topology is constrained primarily by local chip/board pin-outs; secondarily, (if at all) by global bisection bandwidth.

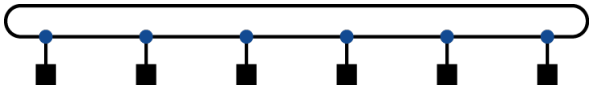
Network Topology (cont.)

- 1970s and 1980s
  - Topologies to reduce hop count
- 1990s
  - Pipelined transmission and switching.
  - Packet latency became decoupled from hop count.
- 2000s
  - Topology still important (especially OCNs, SANs) when N is high.
  - Topology impacts performance and has a major impact on cost.

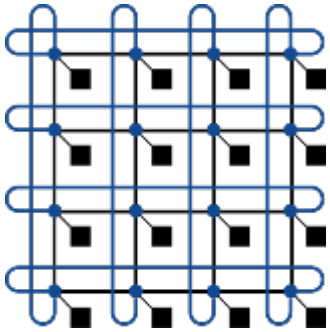
Topologies

- Arrangements of processors, switches, and links

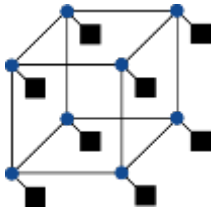
Bus



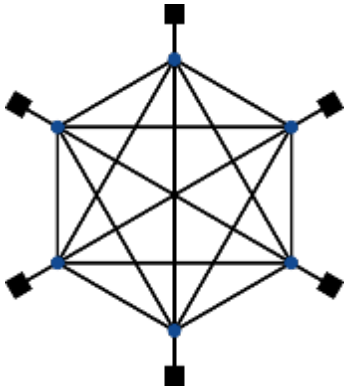
Ring



2-D Torus

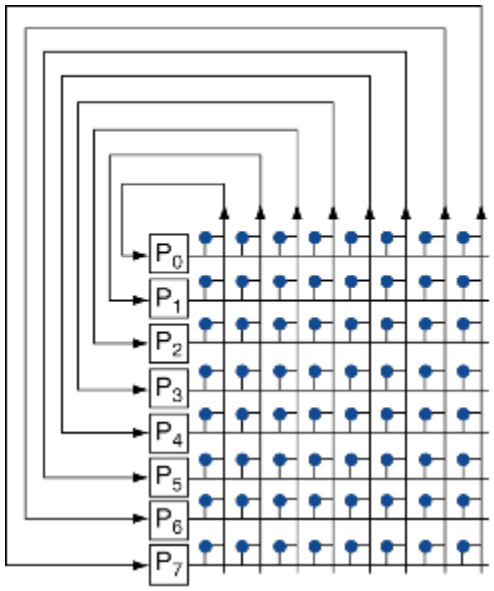


N-cube (N = 3)

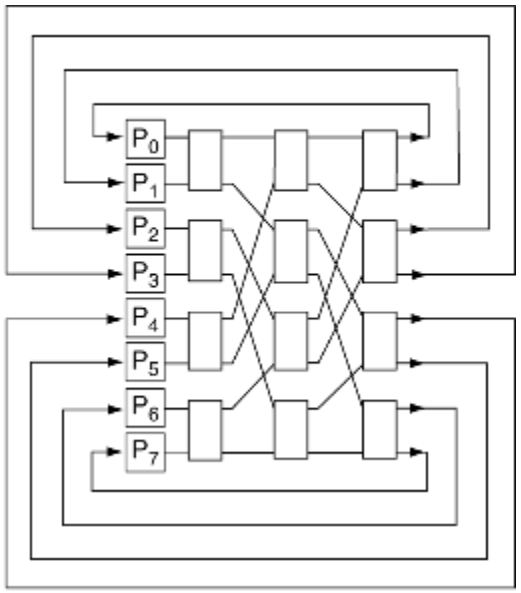


Fully connected

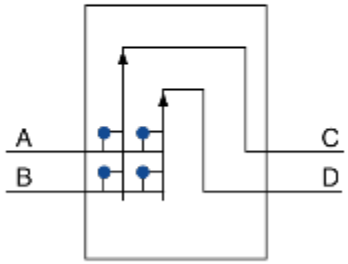
Multistage Networks



a. Crossbar



b. Omega network



c. Omega network switch box

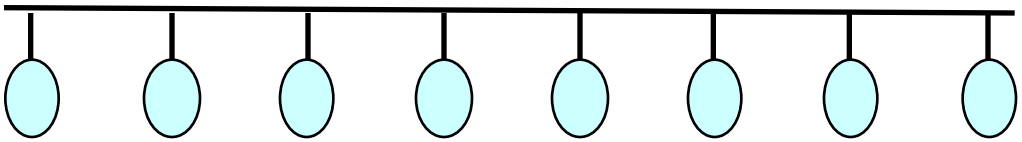
ENGINEERING@SYRACUSE

Network Topologies

Buses

- Simple, cost effective, easy coherence
  - Snooping and serialization
- Not scalable to large number of nodes
  - Limited bandwidth, high contention

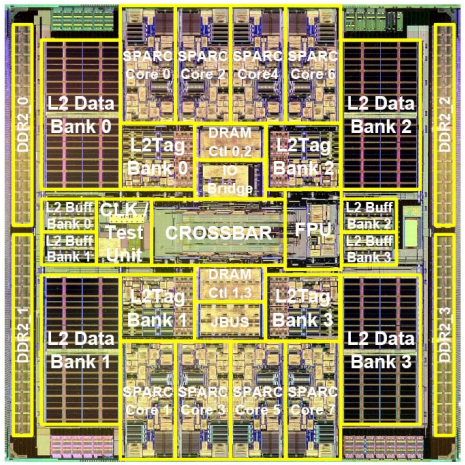
➔ For a small number of nodes



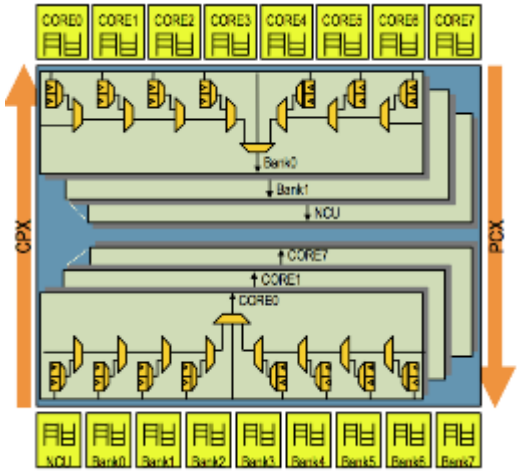
Crossbar (Indirect) Networks

- Concurrent messages to nonconflicting destinations.
- Low latency and high throughput.
- Expensive and not scalable. Cost: $O(N^2)$.
- Used in core-to-cache-bank networks.
- Indirect: Nodes and switches are separate.

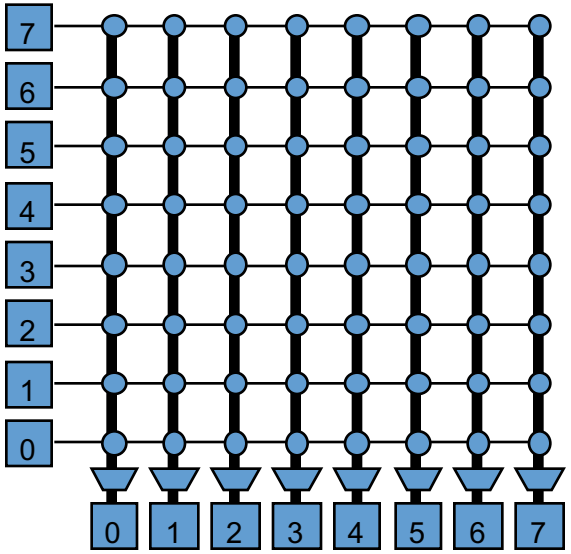
➔ Good for small number of devices.



Sun Niagara

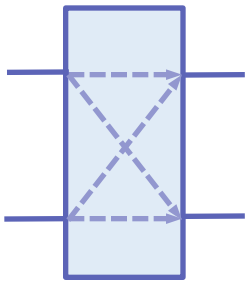


Ultra Sparc T2

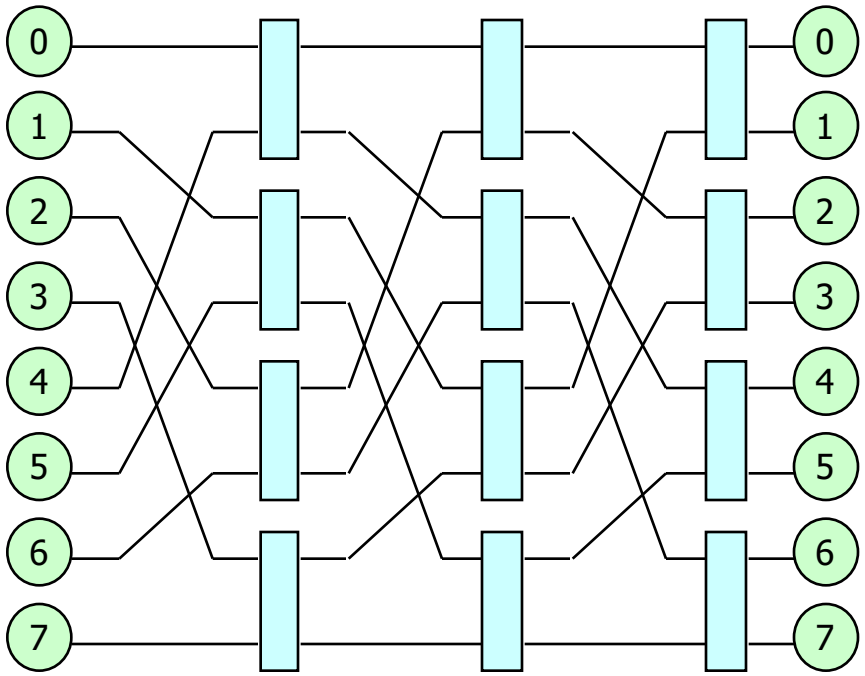


Multistage Interconnection

- Crossbar split into several stages of smaller crossbars.
- Cost: O(N×log N), latency: O(logN).
- Variations: omega, butterfly,…
- Interstage connections represented by a set of permutation functions.



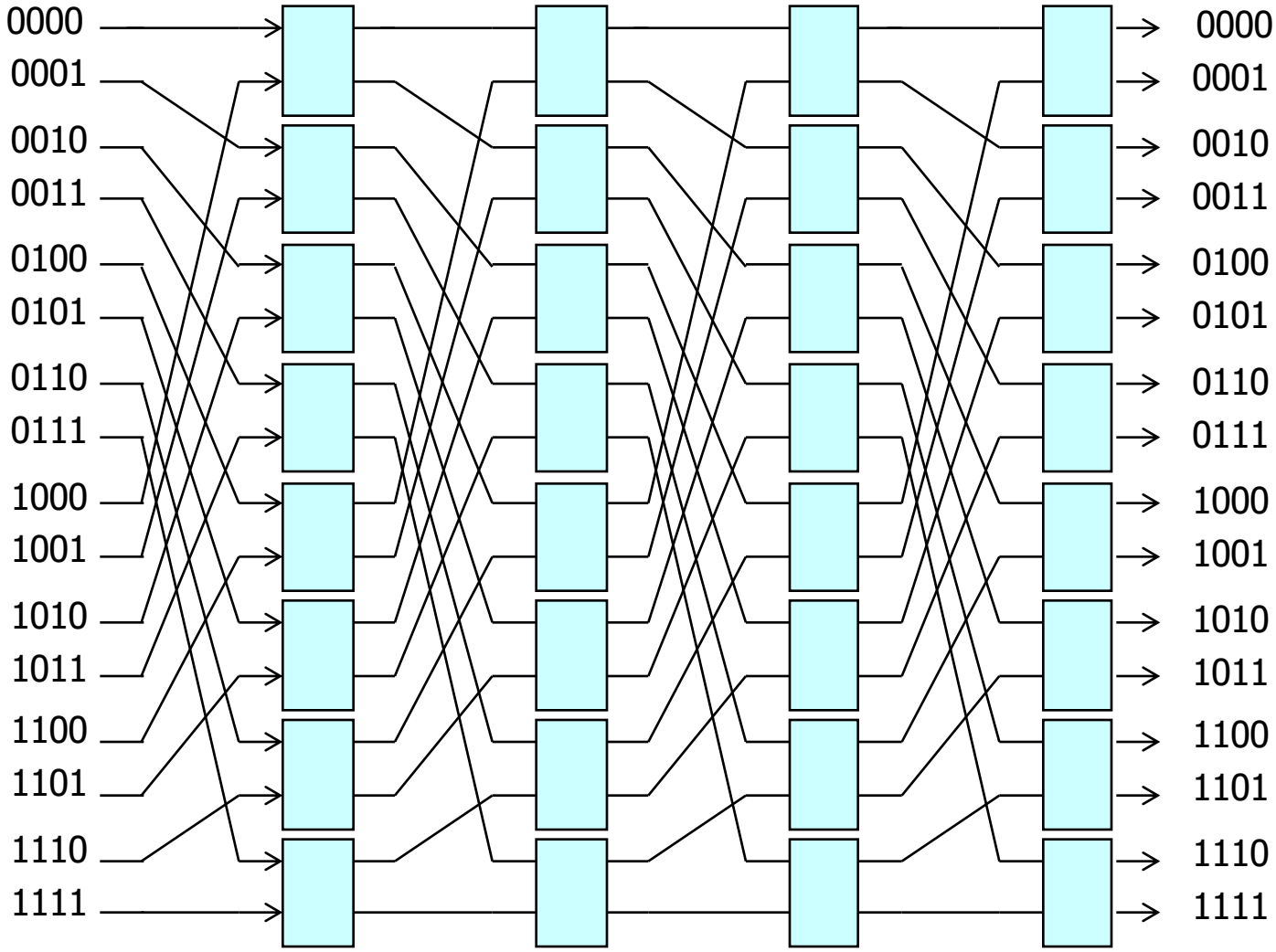
Switch box



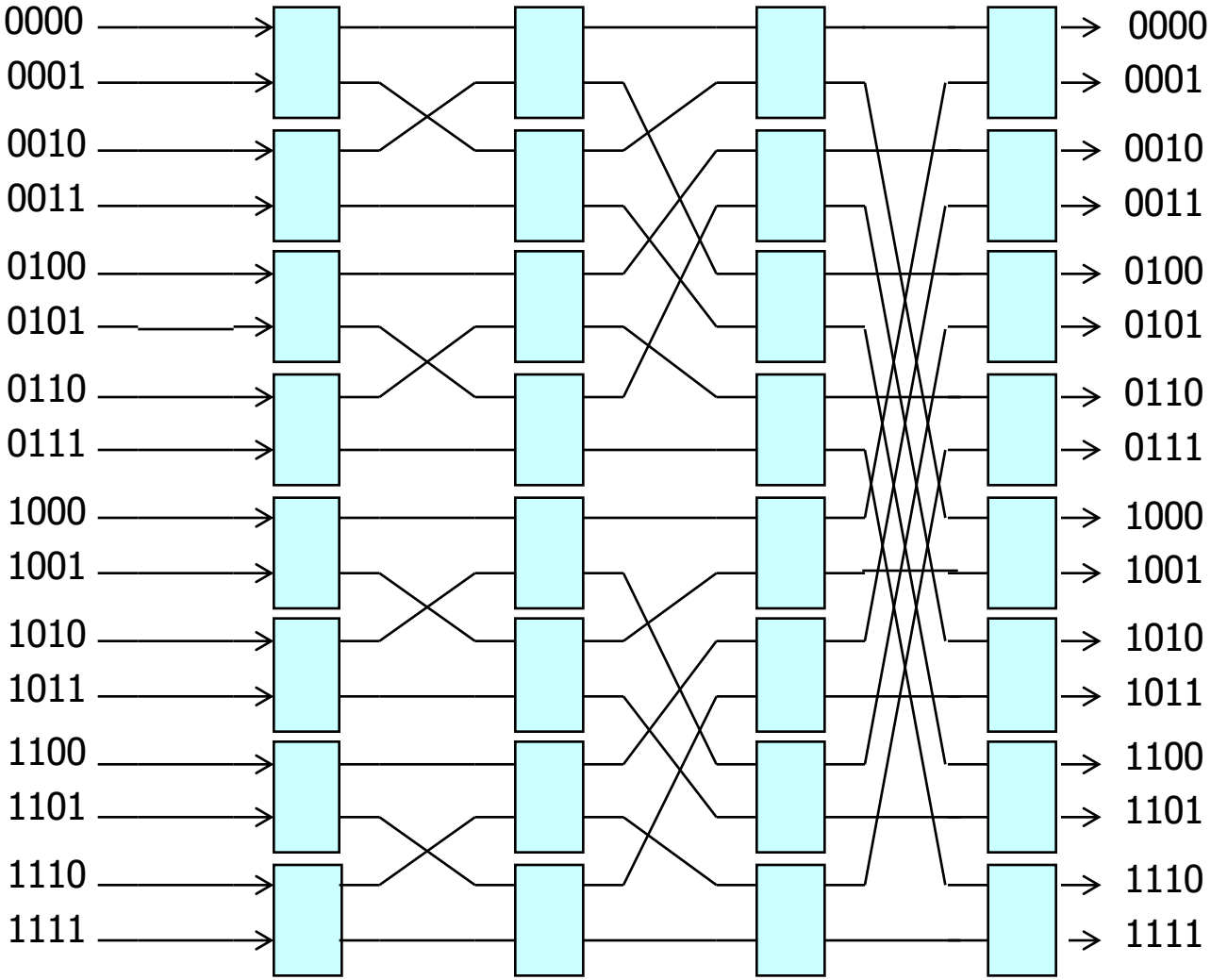
- ***N*** input/output ports
- ***k* x *k*** switches
- $\log_k N$ switch stages, each with ***N/k*** switches
- $N/k(\log_k N)$ total number of switches

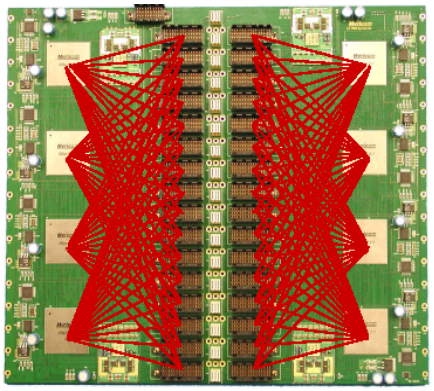
Omega topology
perfect-shuffle exchange

16-Port, Four-Stage Omega Network



16-Port, Four-Stage Butterfly Network



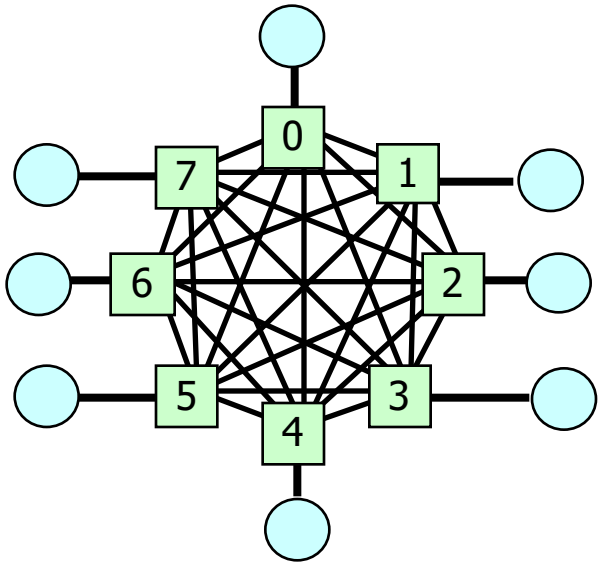


Backplante of M3-E128 switch

Myrinet-2000 clos network for 128 hosts

Distributed (Direct) Networks

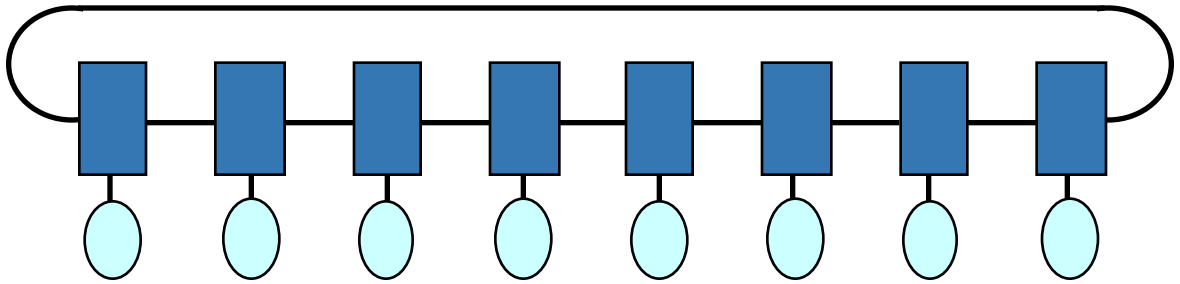
- Tight integration of end-node devices with network resources.
  - Network switches distributed among end nodes
- A “node” now consists of a network switch with one or more end node devices directly connected to it.
  - Nodes are directly connected to other nodes.
- **Fully connected network:**
  - All nodes are directly connected to all other nodes using bidirectional dedicated links.



Lowest contention/lowest latency
Highest cost: $O(N^2)$ links

Bidirectional Ring Networks

- Low cost, O(N)
- High latency, O(N)
- Limited scalability, constant bisection bandwidth
- Simultaneous packet transport over disjoint paths
- Popular, also used in Intel Haswell, IBM Cell,…



ENGINEERING@SYRACUSE

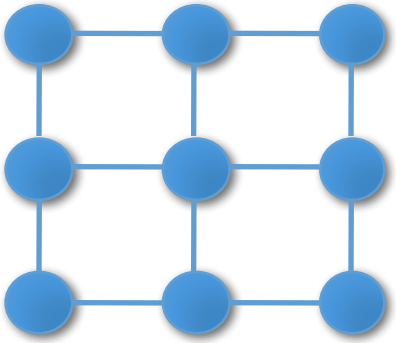
Torus Topologies

Ideal Topology

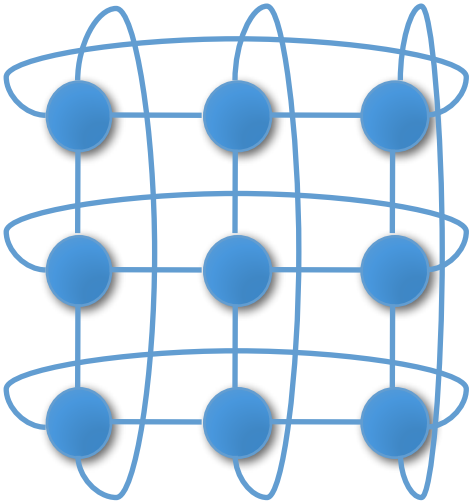
- Cost approaching a ring, performance approaching a fully connected topology
- More practical topologies:
  - k-ary n-cubes (meshes, tori, hypercubes)
  - k nodes connected in each dimension, with n total dimensions
  - Symmetry and regularity
    - Simple implementation and routing

Torus

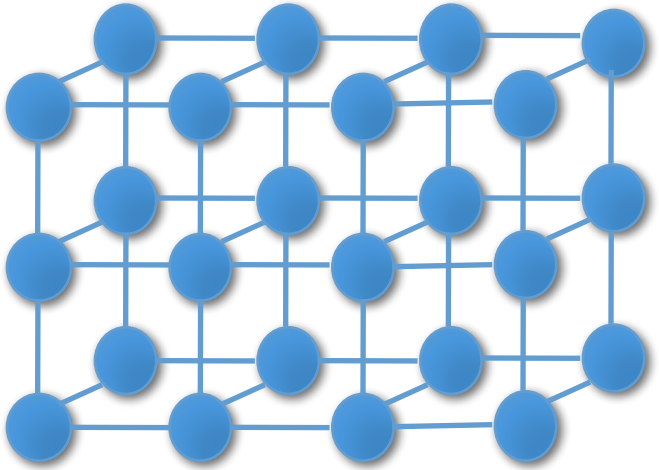
- k-ary n-cube: k^n network nodes
- n-dimensional grid with k nodes in each dimension



3-ary 2-mesh
2-D Mesh

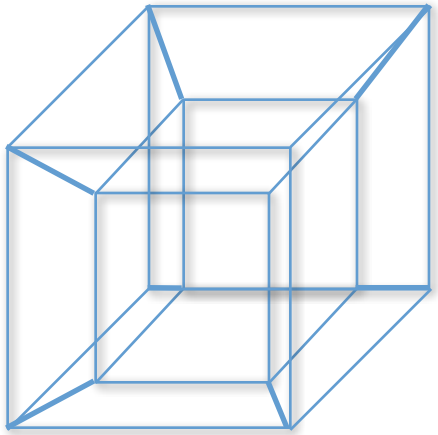


3-ary 2-cube
2-D Torus



2,3,4-ary 3-Mesh

Torus Topologies



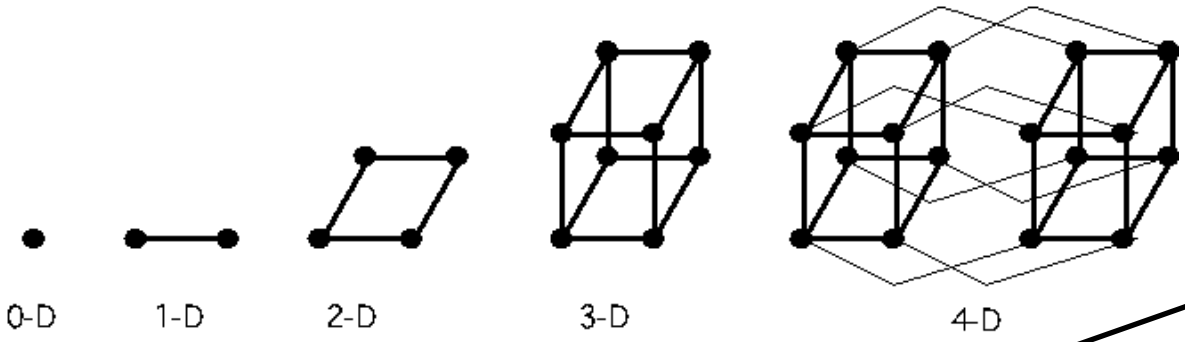
- Ring: k-ary 1-cube.
- Hypercubes: 2-ary n-cube.
- Edge symmetric.
  - Good for load balancing
  - Removing wrap-around links for mesh loses edge symmetry
    - More traffic concentrated on center channels
- Good path diversity.
- Exploit locality for near-neighbor traffic.

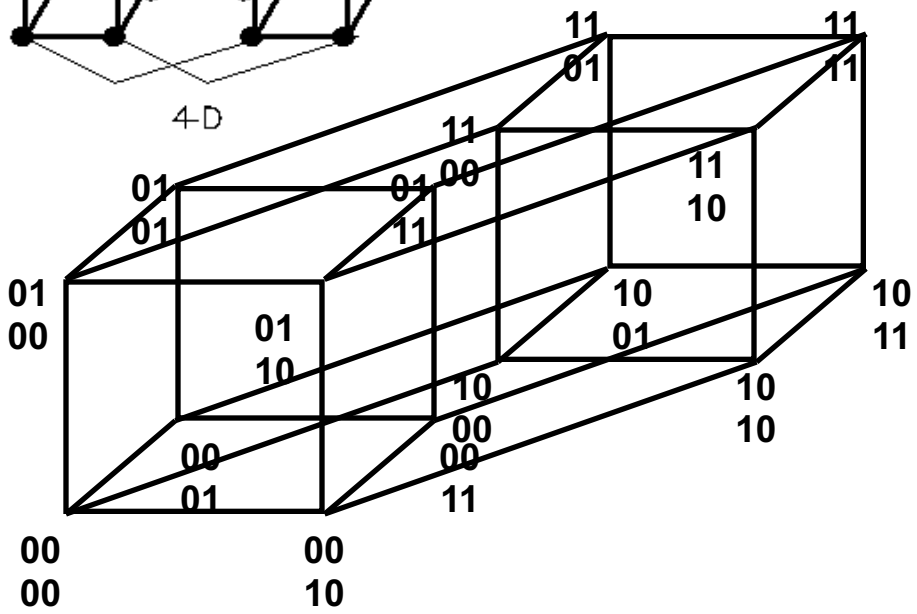
2-D Mesh

- Low cost, O(N)
- Average latency, O(sqrt(N))
- Easy to implement: regular length links
- Path diversity: many ways to get from one node to another
- Used in Tilera 100-core
- And many on-chip network prototypes

Hypercube

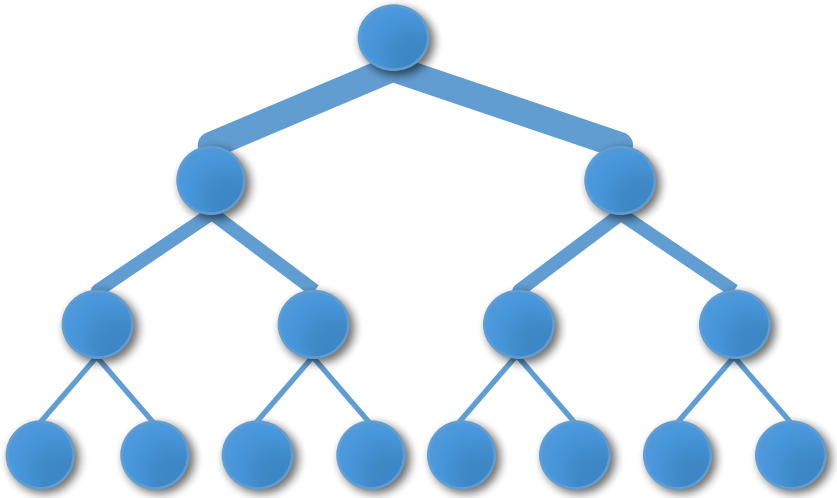
- Low latency: O(logN)
- Number of links: O(NlogN)





Fat Tree

- Planar, hierarchical topology
- Low latency, O(logN)
- Low cost, O(N)
- Good for local traffic



ENGINEERING@SYRACUSE

Performance and Cost Comparisons

Performance and Cost of 64-Node Networks

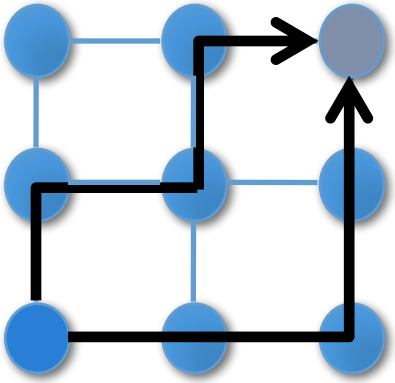
	Evaluation category	Bus	Ring	2-D mesh	2-D torus	Hypercube	Fat tree	Fully connected
Perf.	$BW_{Bisection}$ in # links	1	2	8	16	32	32	1024
	Max (ave.) hop #	1 (1)	32 (16)	14 (7)	8 (4)	6 (3)	11 (9)	1 (1)
Cost	I/O ports per switch	NA	3	5	5	7	4	64
	Number of switches	NA	64	64	64	64	192	64
	Number of net. links	1	64	112	128	192	320	2016
	Total # of links	1	128	176	192	256	384	2080

ENGINEERING@SYRACUSE

Network Routing

Routing

- Routing algorithm determines path(s) from source to destination.
  - **Deterministic:** always chooses the same path for a communicating source-destination pair
  - **Oblivious:** chooses different paths, without considering network state
  - **Adaptive:** can choose different paths, adapting to the state of the network



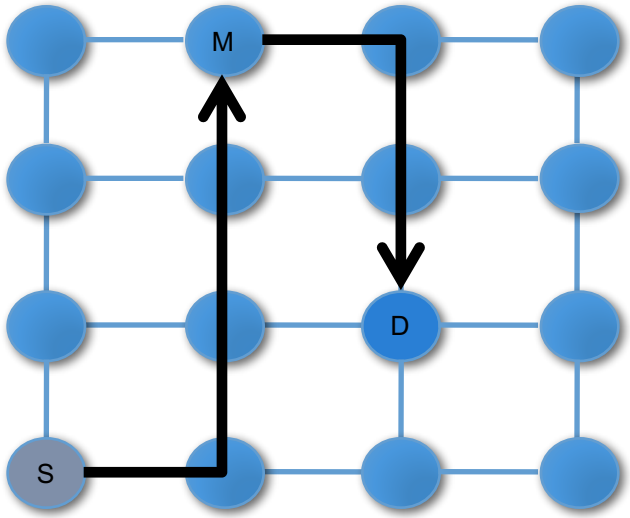
Deadlock: A set of packets being blocked waiting for network resources held by other packets in the set.
Livelock: Unbounded number of allowed nonminimal hops → Solution: Restrict the number of nonminimal (mis)hops allowed.

Deterministic Algorithm

- All messages from source to destination will traverse the same path.
  - Common example: dimension order routing (DOR).
  - Message traverses network dimension by dimension.
  - Aka XY routing, first x- then y-dimension.
- Simple, inexpensive, deadlock free
- Poor load balancing

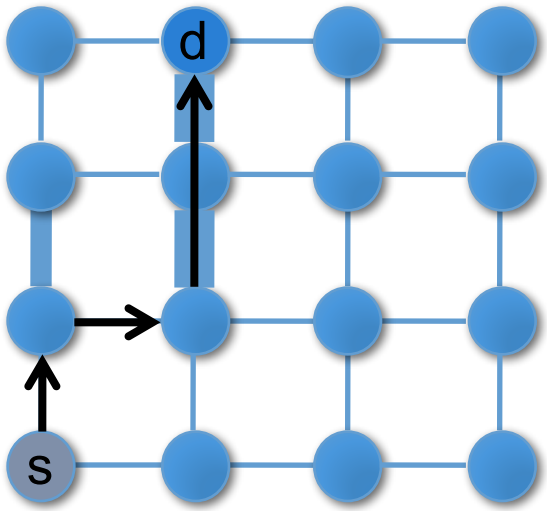
Oblivious Routing

- An example, Valiant's algorithm.
- To route from S to D, randomly choose intermediate node M.
- Route from S to M and from M to D.
- Randomizes any traffic pattern.
  - All patterns appear to be uniform random.
  - Balances network load.
- Nonminimal.
  - Latency can increase.
- Do this on high load.



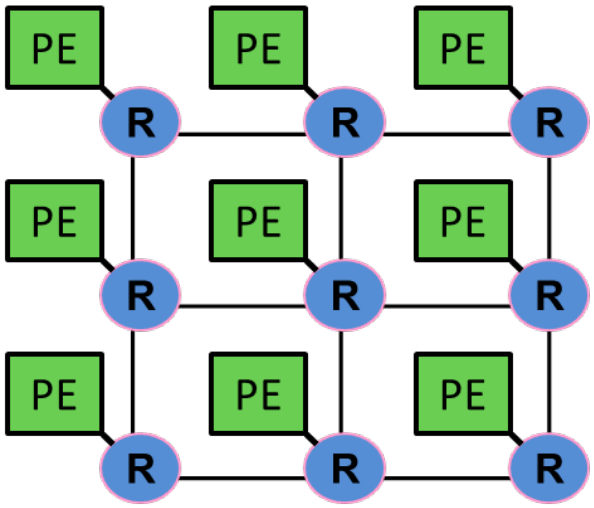
Minimal Adaptive Routing

- Router uses network state (e.g., downstream buffer occupancy) to pick which "productive" output port to send a packet to.
- Productive output port: port that gets the packet closer to its destination.
- Aware of only local congestion.
  - Suboptimal choices
- Fully adaptive.
  - Better utilization, load balance
  - Livelock potential
    - Traversing without ever reaching



On-Chip Networks

- Connects processing elements (PE).
  - Cores, caches, memory controllers, …
- Serves cache misses and memory requests.
- Buses and crossbars are not scalable.
- Packet switched.
- 2-D mesh.
  - Most commonly used topology
- Dimension order routing.
  - Simplicity and deadlock freedom



ENGINEERING@SYRACUSE

Flow Control and Switching

Flow Control

- Topology
  - Determines connectivity of network
- Routing
  - Determines paths through network
- Flow control
  - Determines allocation of resources to messages as they traverse network
    - Buffers and links
    - Impact on throughput and latency of network

Packets

- Messages
  - Composed of one or more packets
- Packets
  - Composed of one or more flits
- Flit
  - Flow control digit
- Phit
  - Physical digit
  - Subdivides flit into chunks = to link width
- In on-chip networks, flit size == phit size.
- Due to very wide on-chip channels

Switching

- Circuit switching sets up full path.
  - Establish route, then send data.
  - No one else can use those links.
  - Faster arbitration, but setting up and bringing down links takes time.
- Packet switching routes per packet.
  - Route each packet individually (possibly via different paths).
  - If link is free; any packet can use it.
  - Potentially slower but no setup or bring down time and more flexible.
- Choice is independent of topology.
  - But maybe more preferable for one type over the other

Flow Control Types

- Store and forward (for LAN, WAN)
  - Links and buffers are allocated to entire packet.
  - Not suitable for on-chip.
    - Requires large buffers to hold entire packet
    - High latencies (pays serialization at each hop)
- Wormhole (for NoC, SAN)
  - Flit can proceed to next router when there is buffer space available for that flit.
  - Good for on-chip.
    - More efficient buffer utilization
    - Low latency

ENGINEERING@SYRACUSE

Conclusions

In Conclusion

- Devices: components, computers, systems
- Domains: OCN, SAN, LAN, WAN
- Metrics: latency, effective BW
- Media: shared vs. switched, low cost vs. scalable
- Topologies: bus, k-ary n-cube, multistage
- Comparisons: bisection BW, links, switches
- Routing: deterministic, adaptive
- Flow: store and forward, wormhole

ENGINEERING@SYRACUSE