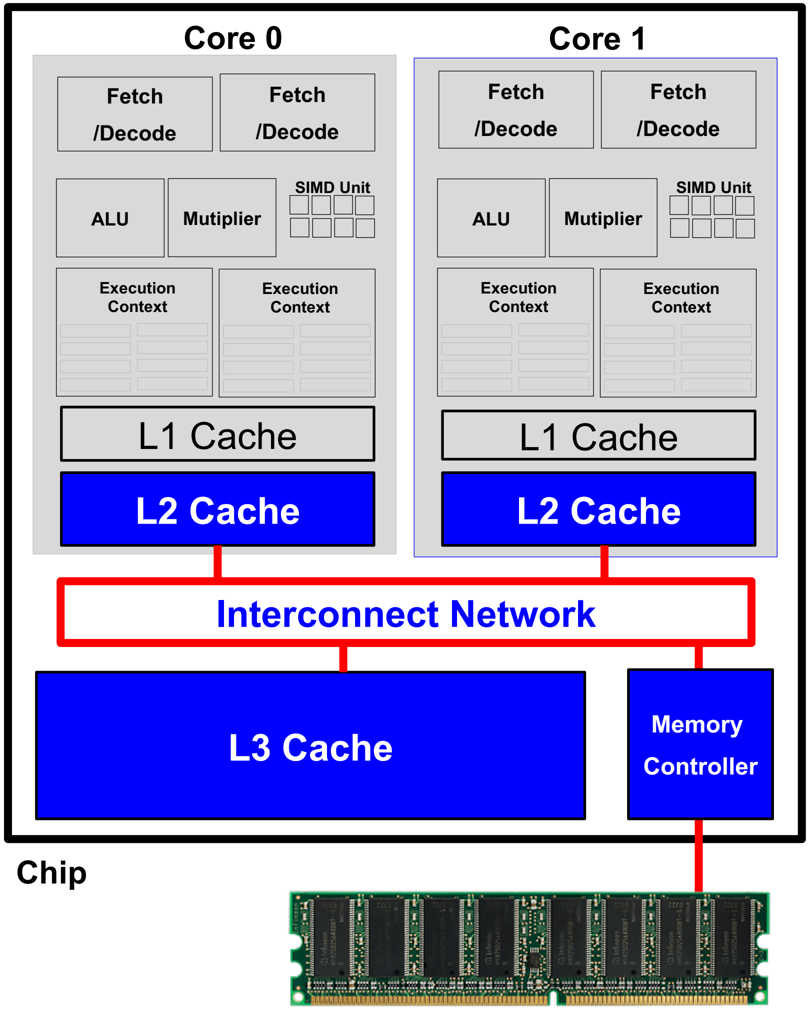
Interconnection Network

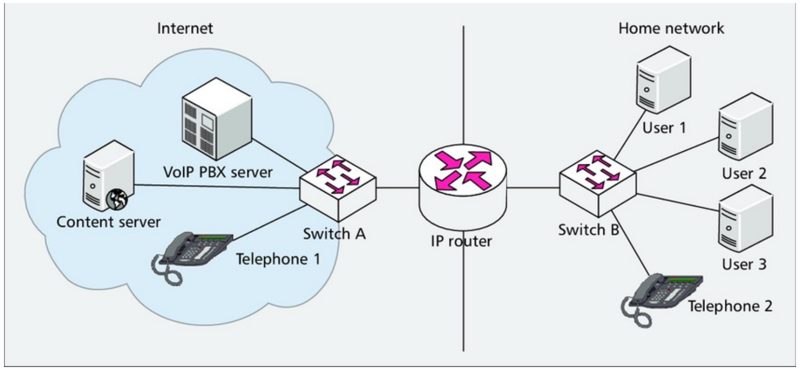
Prof. Seokin Hong

Some slides are modified form CMU 15-418

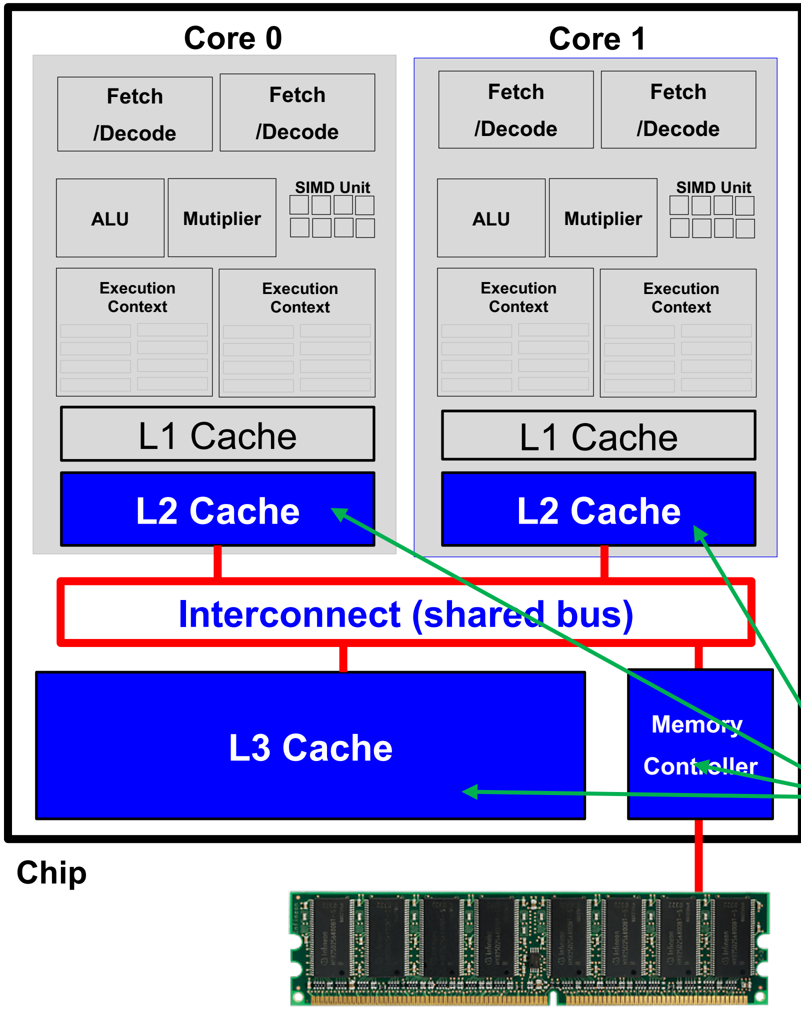
Interconnection Network



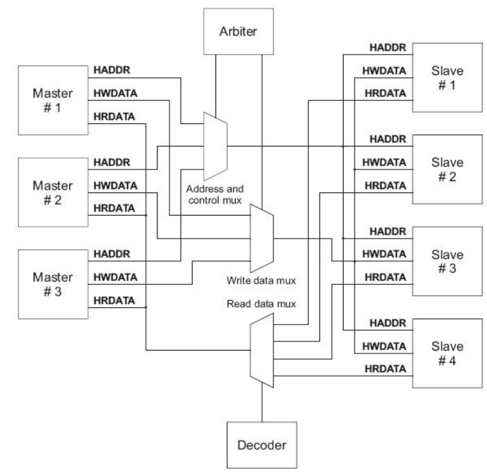
- **The job of an Interconnection Network**
  - ✓ Transfer infromation from any source node to any desired destination node



Basic Interconnect : Bus



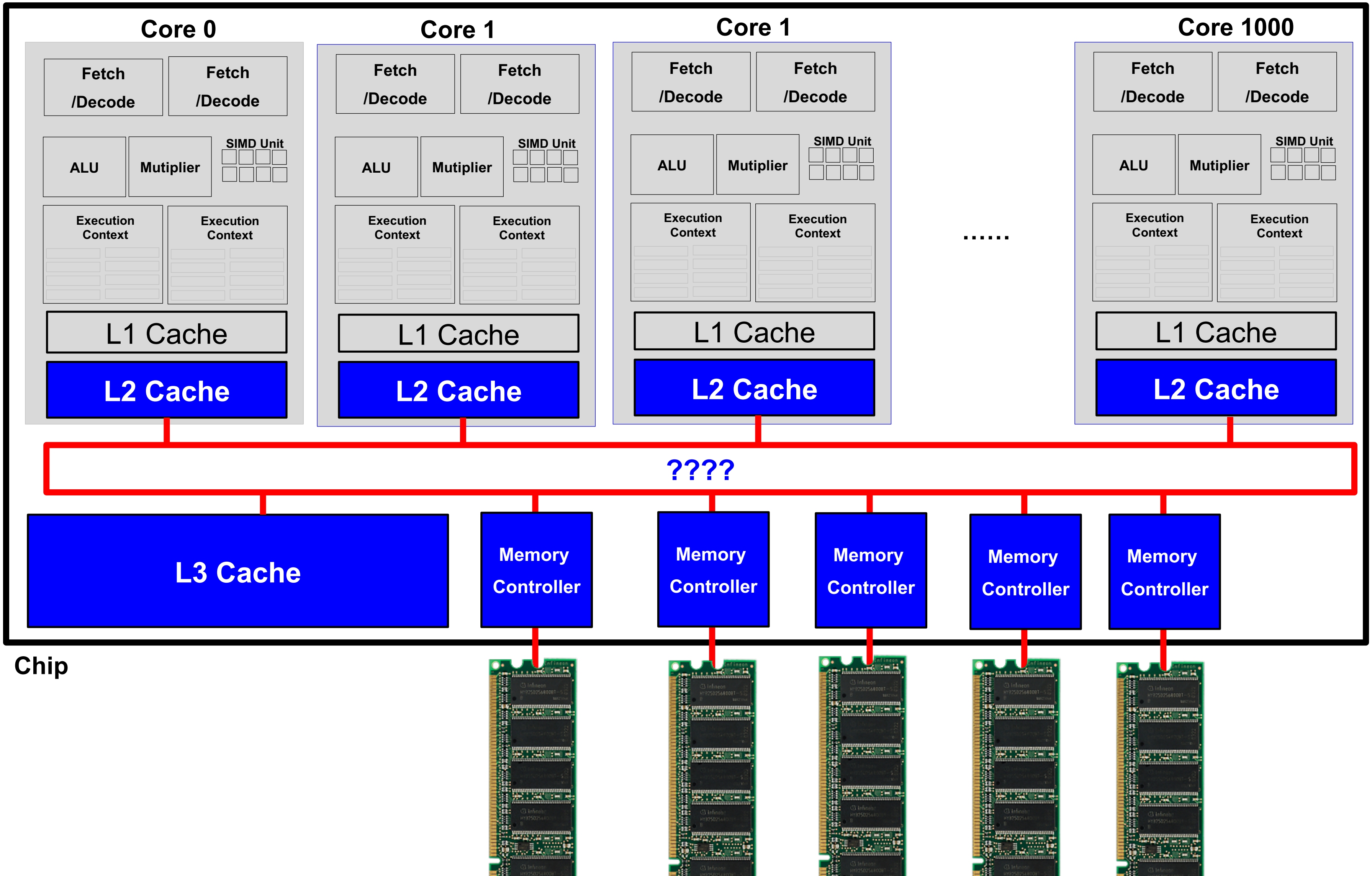
- **Bus: a set of wires that conects all nodes**
  - Address
  - Command
  - Data



Bus clients (interconnect nodes)

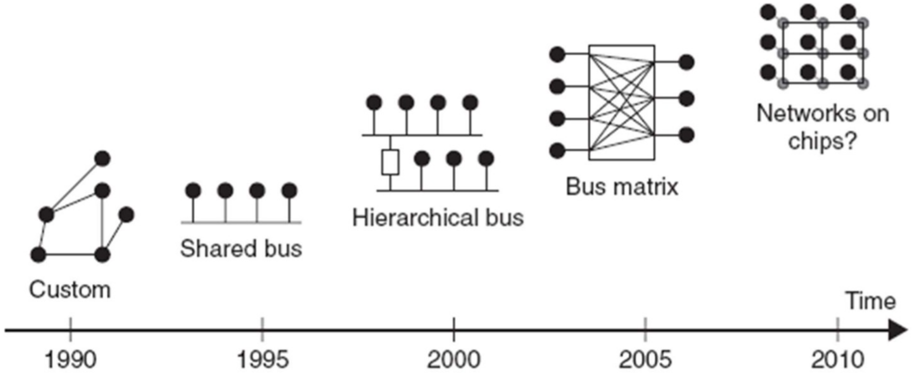
Today's topic

- High-performance and/or scalable interconnection network!!



Agenda

- **Basic Concept**
- **Topology**
- **Flow control**



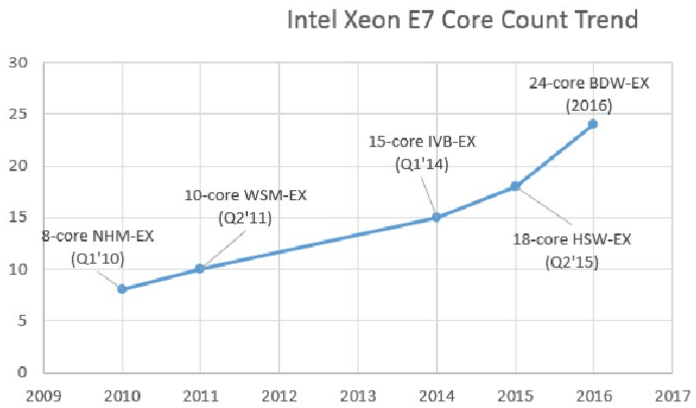
What is it, Why important?

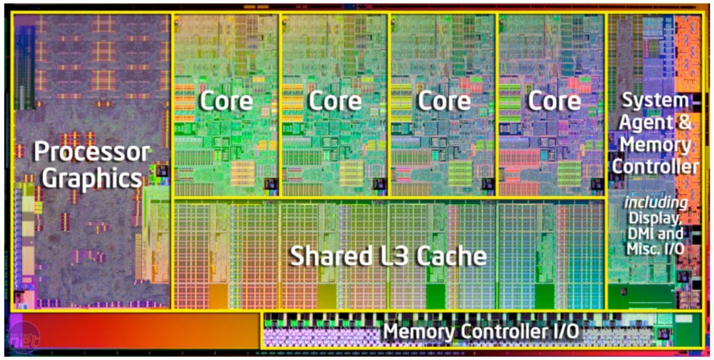
- **What are connected?**
  - Processor Cores
  - Caches
  - Memory Controller
  - I/O Interface Controller
  - ….

- **What is important in the design of Interconnection Network**?
  - **System Scalability**
    - Connect more nodes (e.g., cores, caches, memory controller)
  - **System Performance and Energy-efficiency**
    - Enable fast communication between nodes
      - High bandwidth, low latency
    - Consume less energy

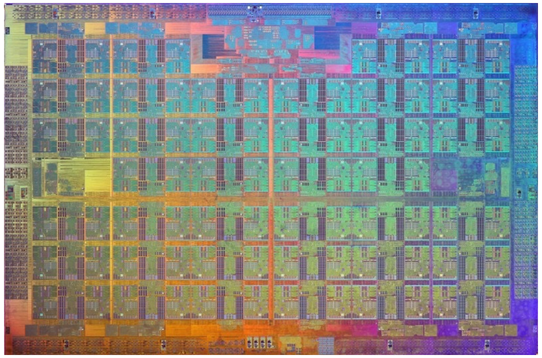
Interconnection Network Becomes More Important

- **Core counts keep increasing…**

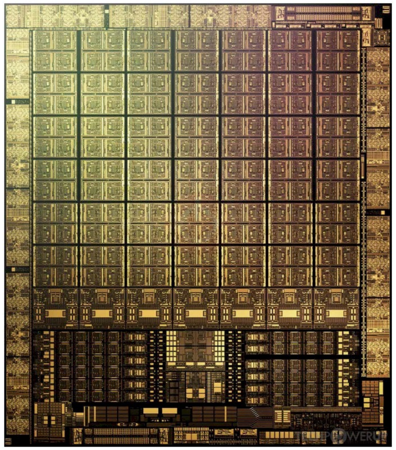




Intel core i7 (4 CPU cores + GPU)



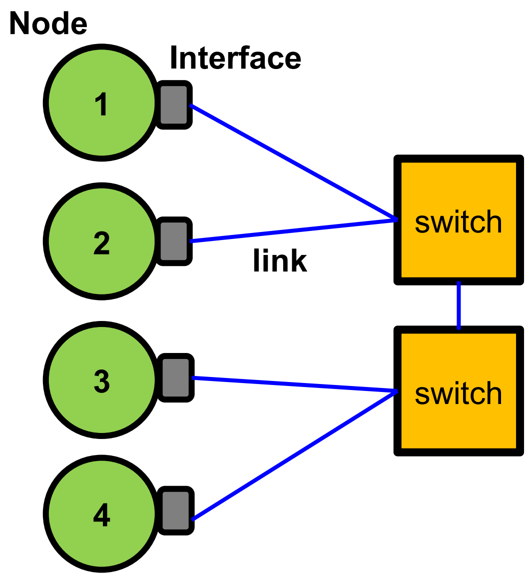
Intel Xeon PhI (72 CPU Cores)

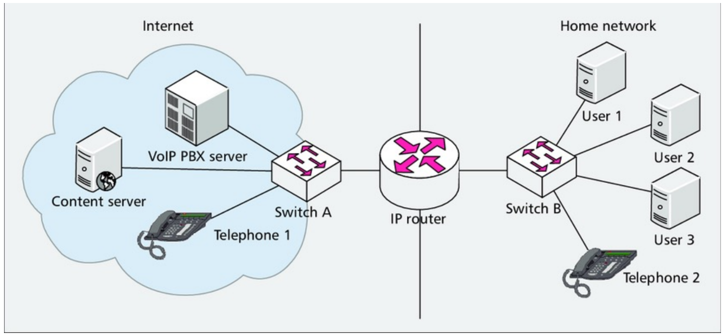


Nvidia Ampere
(10368 CUDA Cores)

Terminology

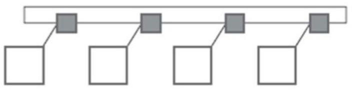
- **Network Node**
  - A Network endpoint connected to a router/switch
    - Example: processor cores, caches, memory controller
- **Network Interface**
  - Connects nodes to the network
- **Links**
  - A bundle of wires carring a signal
- **Switch**
  - Connects a fixed number of input links to a fixed number of output links





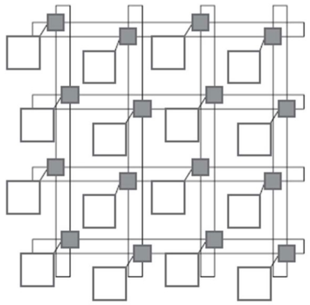
Basic concepts and design issues

- **Topology**
  - Physical interconnection structure of the network graph
  - How switches are connected via link



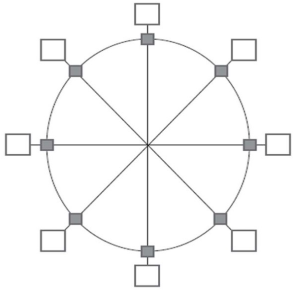
1-D torus

- **Routing Algorithm**
  - Restricts the set of possible paths to a smaller set of legal paths
  - How a message gets from its source node to its destination node



2-D torus

- **Flow control**
  - Determine when the message moves along its route
  - It is necessary whenever two or more message attept to use the same network resources

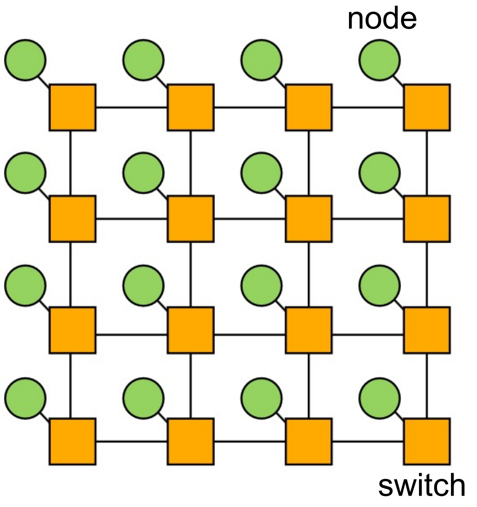


Octagon

Topologies

Properties of interconnect topology

- **Routing distance**
  - Number of links ("hop") along a route between two nodes
- **Diameter**
  - the maxumum routing distance
- **Average distance**
  - average routing distance over all valid routes
- **Blocking vs non-blocking**
  - If connecting any pairing of nodes is possible, the network is non-blocking

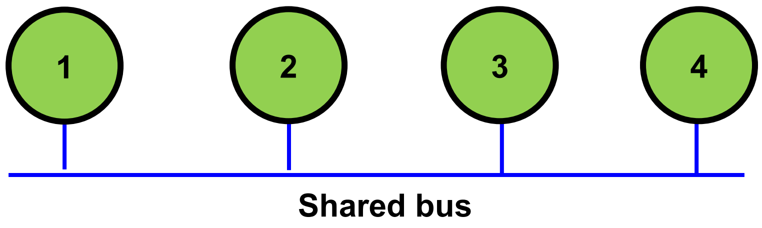


Diameter? 6

Bus Interconnect

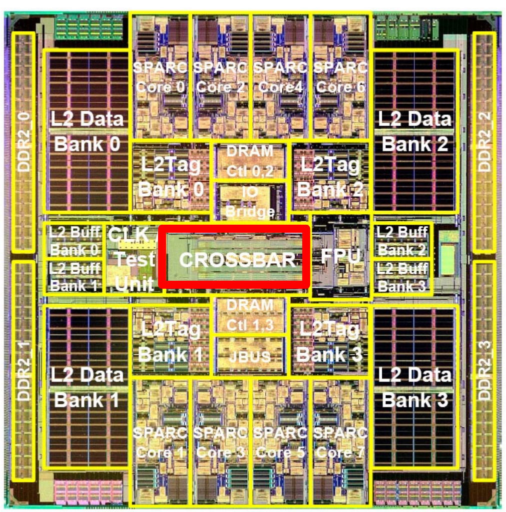
- **Pros**
  - Simple
  - Cost effective for a small number of nodes
  - Easy to implement snooping-based coherence

- **Bad**
  - **Contention**: all nodes contend for shared bus
  - **Limited bandwidth**: all nodes communicate over same wire
  - **High electrical load** = low frequency, high energy-consumption

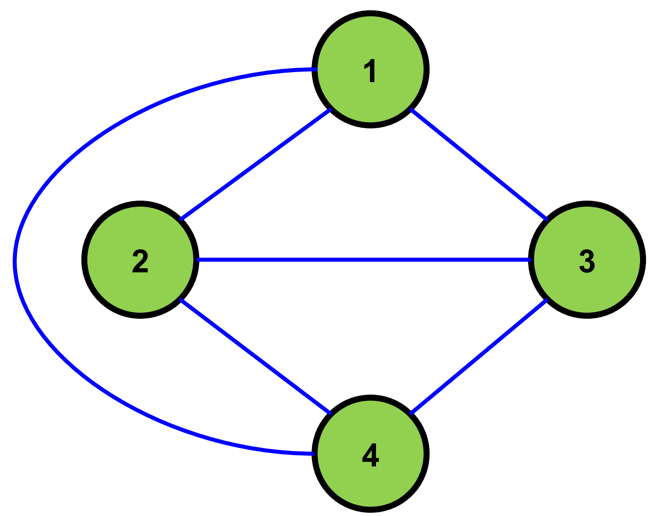


Crossbar Interconnect

- Every node is directly connected to every other node
  - Connects n inputs to m outputs without intermediate stages
- Diameter: 1 link
- Non-blocking
- **Good:** O(1) distance(latency) and high bandwidth
- **Bad:** O(N^2) cost → Not scalable

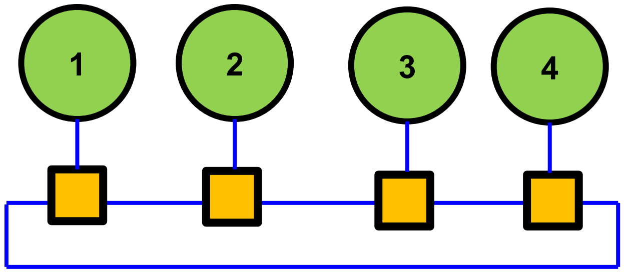


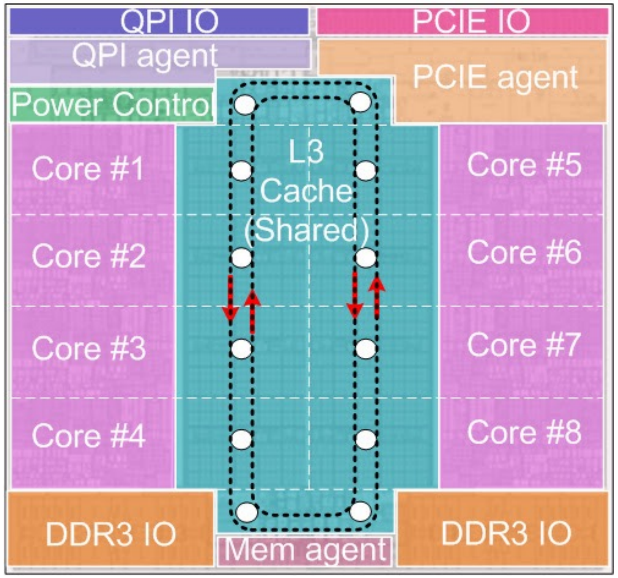
The Sun Microsystems OpenSPARC T2



Ring

- A linear array of nodes connected by bidirectional links
- Diameter: N-1
- Average distance: N/2
- Good
  - Simple and Cheap: O(N)
- Bad
  - High latency (i.e., distance): O(N)

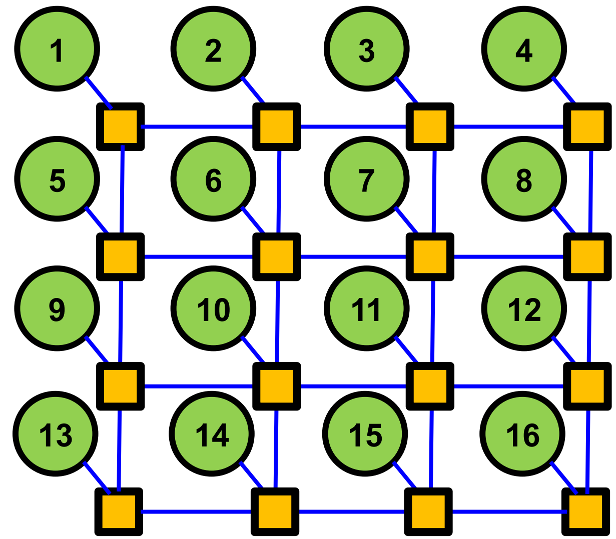


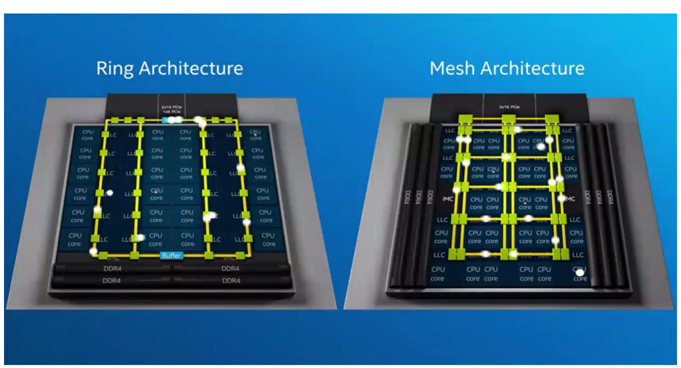


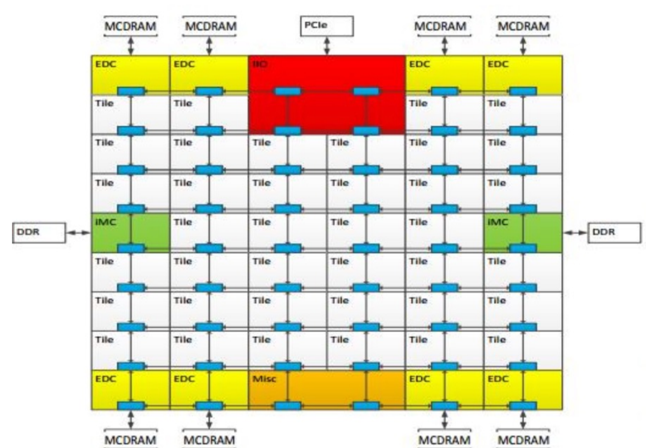
Intel Sandy Bridge with 8cores

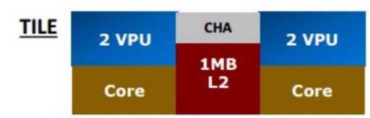
Mesh

- Rings generalize to higher dimensions
- O(N) cost
- Average distance: O(sqrt(N))
- Good when traffic has locality
- Fixed-length link→ Easy to layout on chip
- Path diversity → Many ways for message to travel from one node to another node





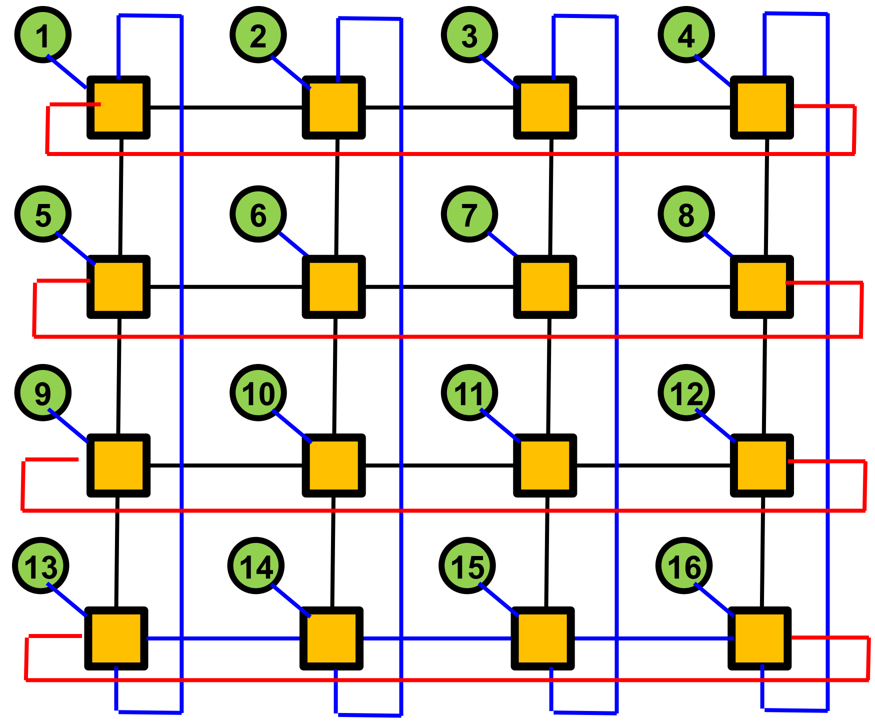




Intel Phi (Knights Landing)
72 cores, 6x6 mesh of tiles

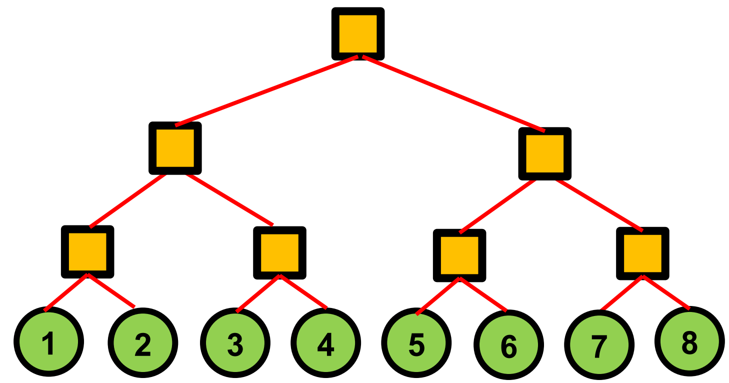
Torus

- Limitation of mesh
  - Network characteristics of the nodes are different depending on their position in the network
- Torus solve this problem by using new links
- Cost: still O(N)
- Higher path diversity
- Higher complexity
  - Difficult to layout on chip

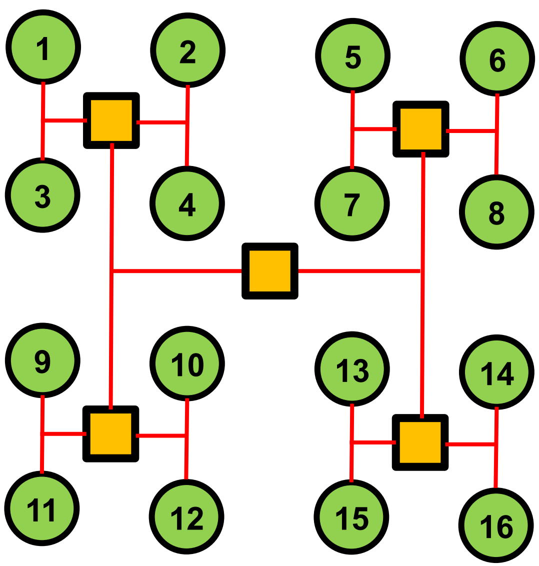


Tree

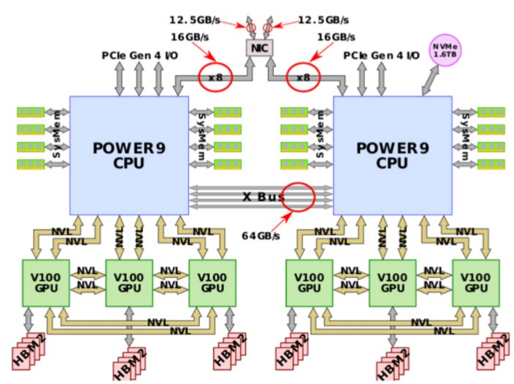
- Limitation of mesh
  - More node → Diameter and average distance increases
- Tree can mitigate this problem
- Good when traffic has locality

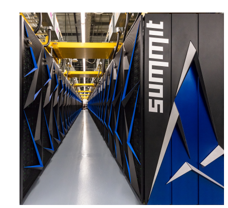


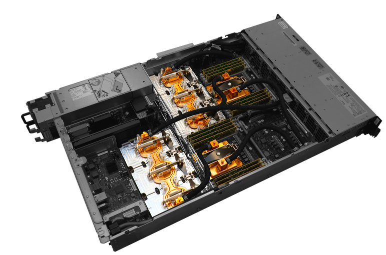
Fat Tree



H Tree







Oak Ridge Summit Node
3-level fat tree

Flow Control

Granularity of Communication

- **Message**
  - Unit of transfer between node (e.g., processor cores)
  - Can be transmitted using many packets

Packet Format

Header	Payload	Tail

- **Packet**
  - Unit of transfer for network
  - Can be transmitted using many flits

Header: contains routing and control information

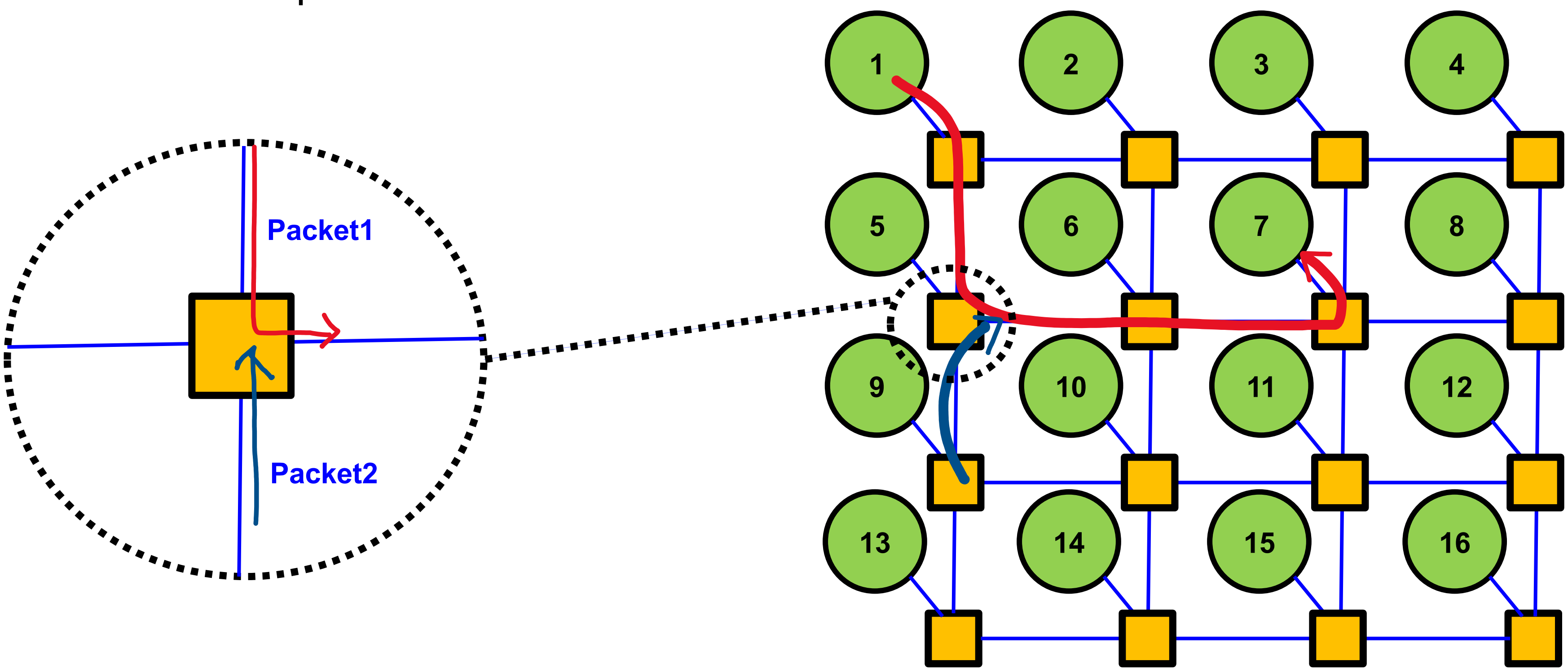
Payload: contain the data

Tail: contain control information
 - checksum code for error detection

- **Flit (flow control unit)**
  - A unit of flow control in the network
  - Flits are minimum granularity of routing

How to handle contention?

- What happen if two packets need to be routed onto the same output port of a switch at the same time?
- **Three options**
  - Buffer one of packets and send it later
  - Drop one of packets
  - Reroute the packet

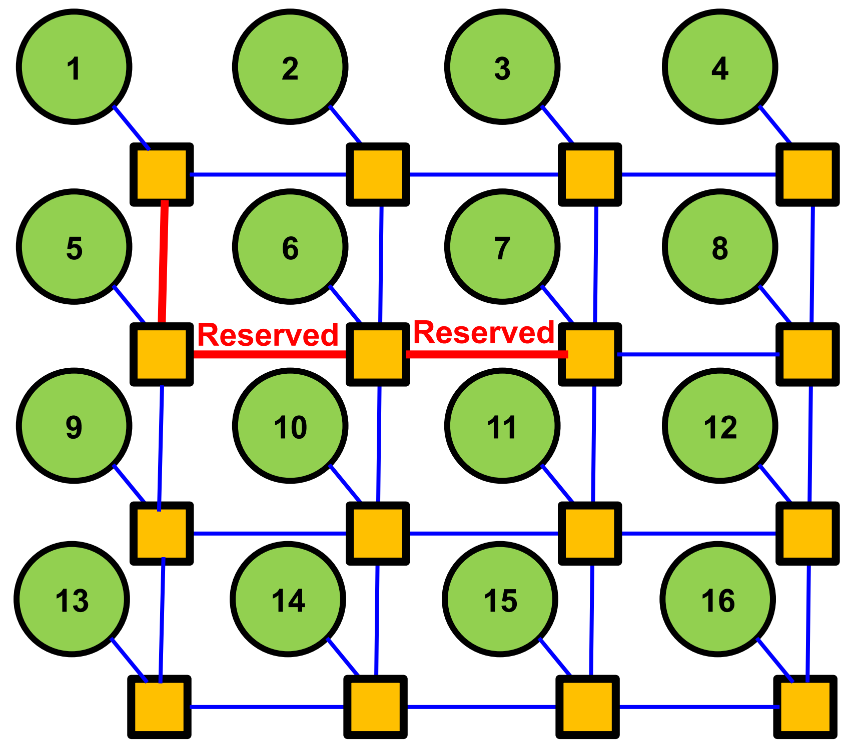


Circuit-switched routing

- **Idea:** Pre-allocate all resources (links across multiple switches) along entire path for a message

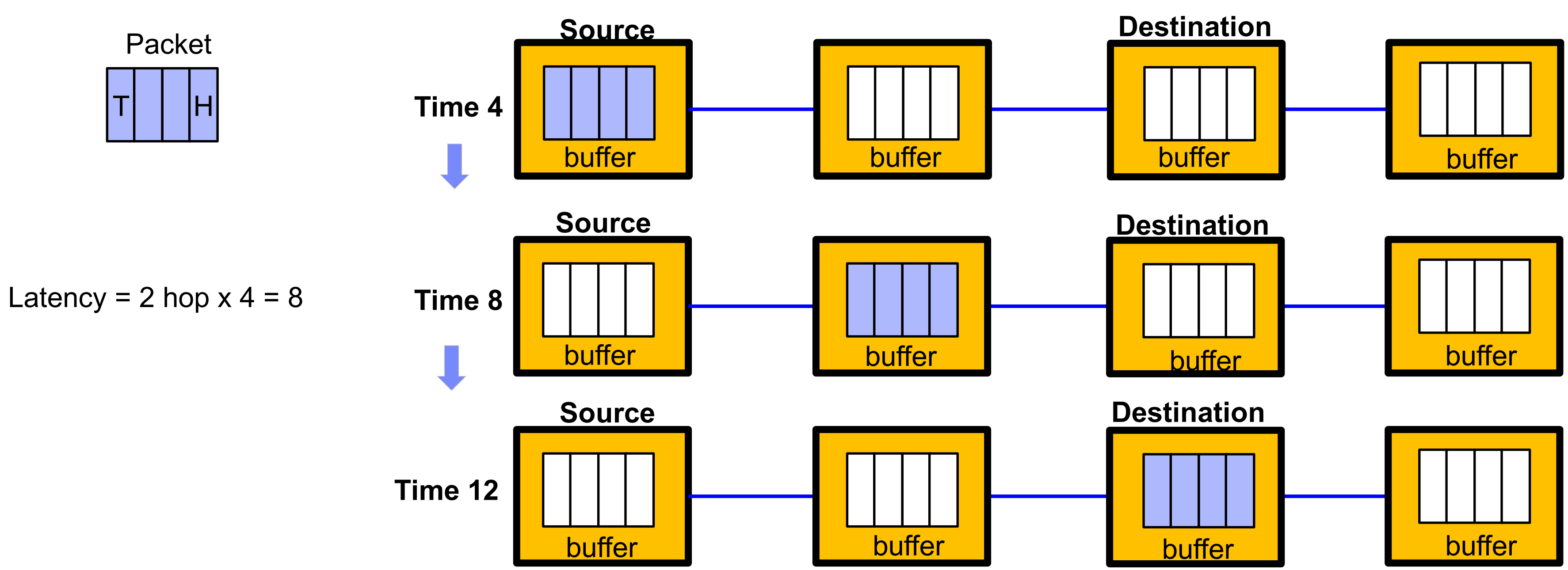
- **Pros**
  - No contention
  - No need for buffering

- **Cons**
  - Need pre-allocate phase
  - Lower link utilization



Store-and-forward

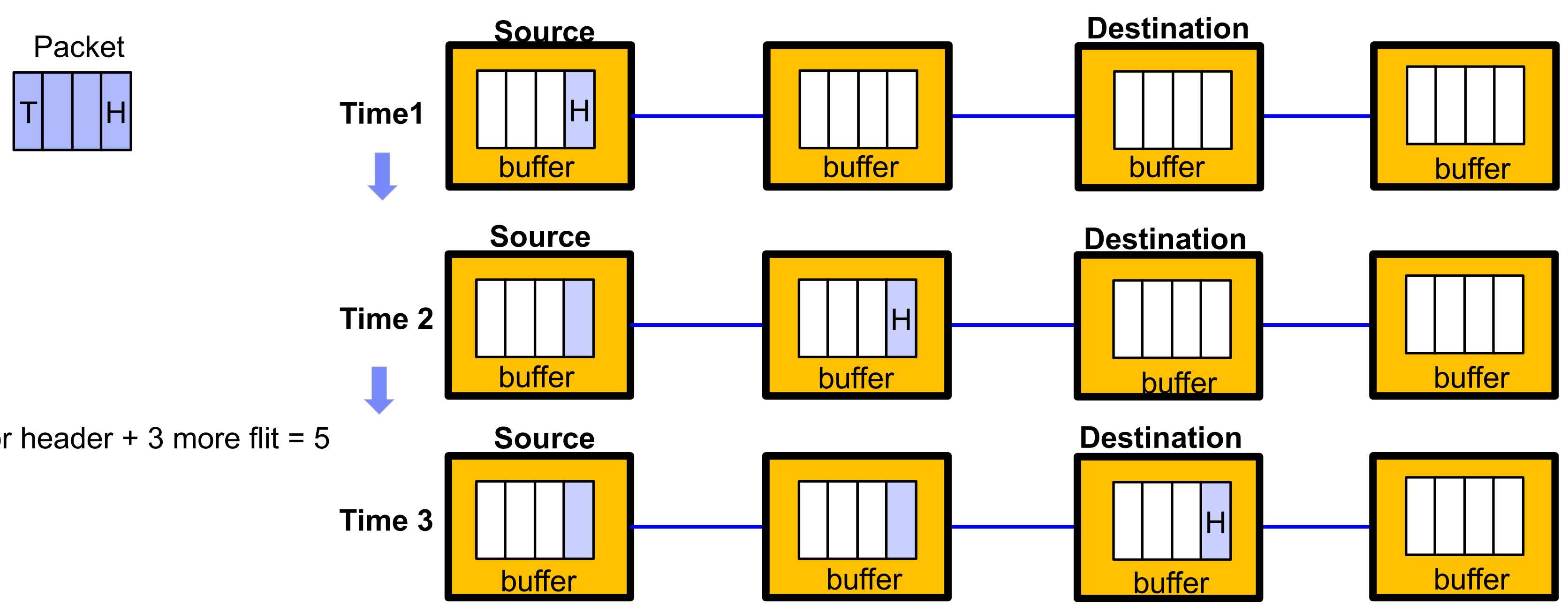
- **Idea**
  - Copy all packet into a network switch before move to next switch
  - Flow control unit is a **packet**
    - Different packet from the same message can go to different routes
- **Pros:** support error checking
- **Cons:** high latency (latency= transmission time on link x distance)



Cut-through flow control

- **Idea:** forward data on next link as soon as packet header is received
  - Packet header contains the destination information

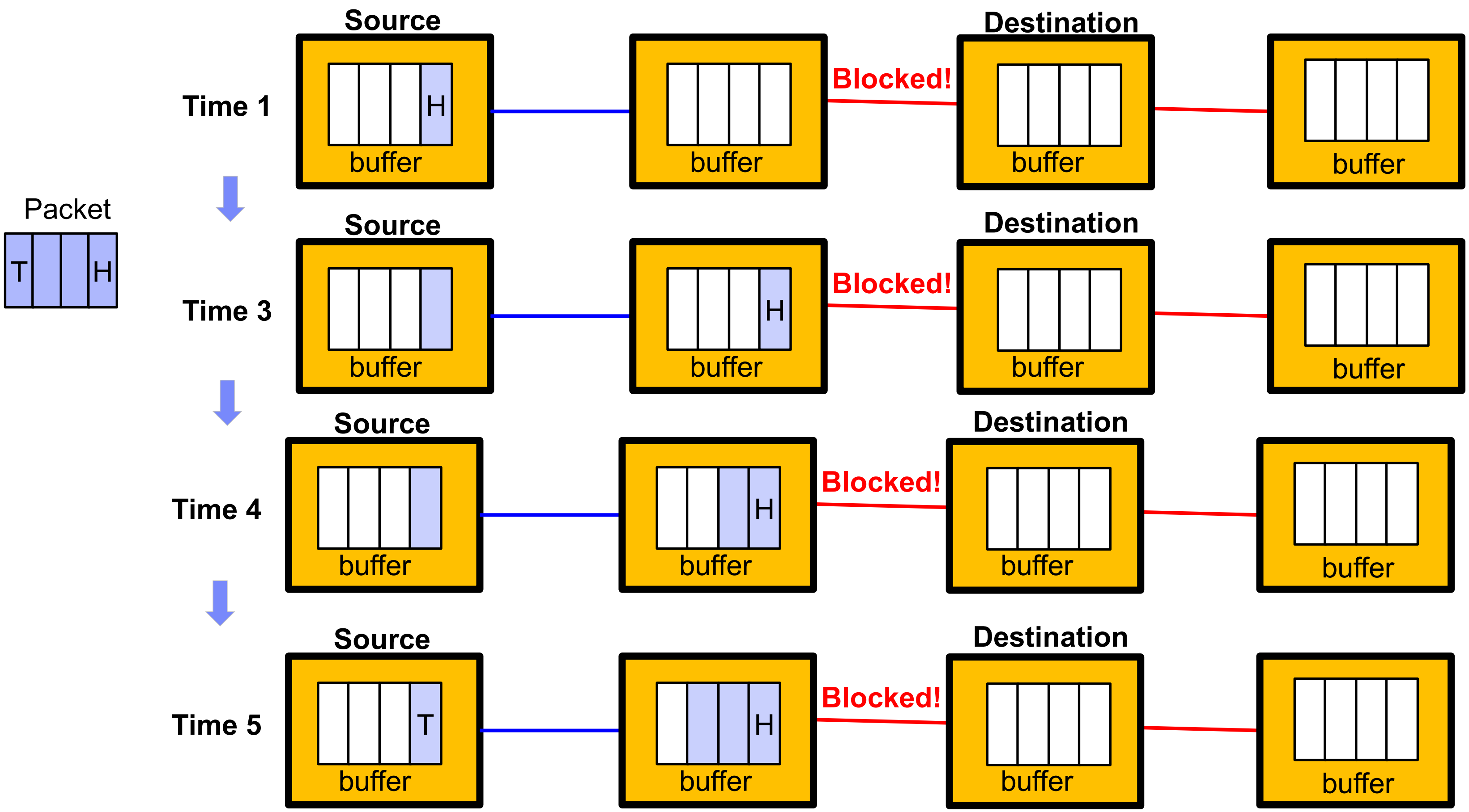
- **Pros:** low latency
- **Cons:** error can be propagated to the destination



Latency
= 2 hops for header + 3 more flit = 5

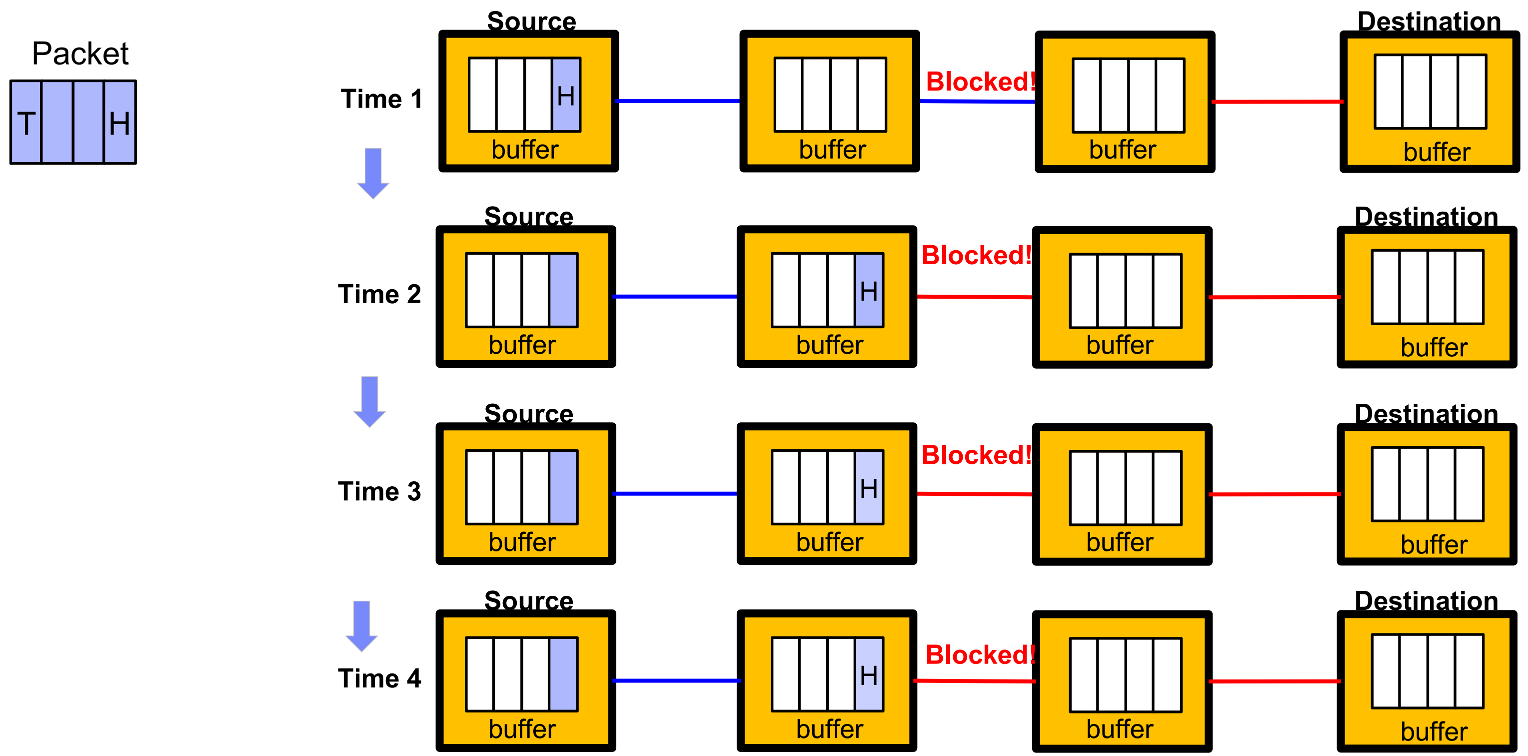
Cut-through flow control (Cont.)

- **What happen if header flit is blocked on a switch?**
- **All flits can be stored in the buffer of a switch**
  - Require a large buffer (at least packet size)



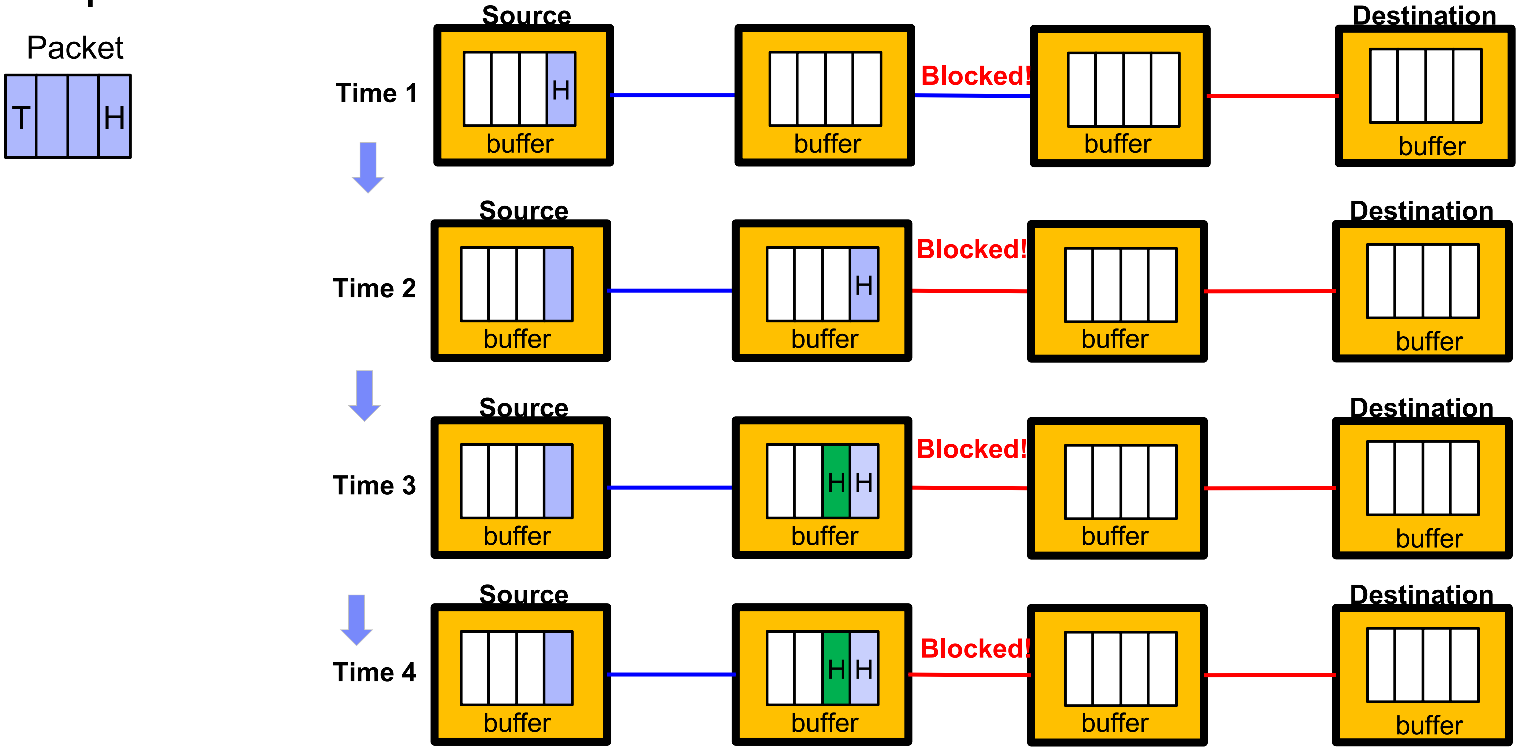
Wormhole flow control

- **Idea**
  - Forward data flits as soon as header flit is received
  - If head flit blocks, other flits of a packet stop and wait on different switchs
- **Pros:** do not require a large buffer (good for On-chip Network)



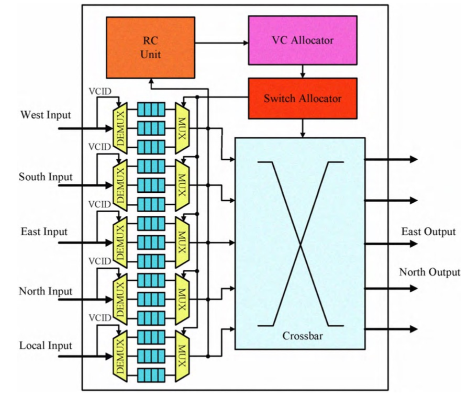
Wormhole flow control (Cont.)

- Idea
  - Forward data flits as soon as header flit is received
  - If head flit blocks, other flits of a packet stop and wait on different switchs
- **Pros:** do not require a large buffer (good for On-chip Network)
- **Cons: head-of-line blocking →** block other packets even if the link for the packet is idle



Virtual channel flow control

- **Idea:** divide input buffer into multiple buffers sharing a single physical channel



Switch supporting virtual channel