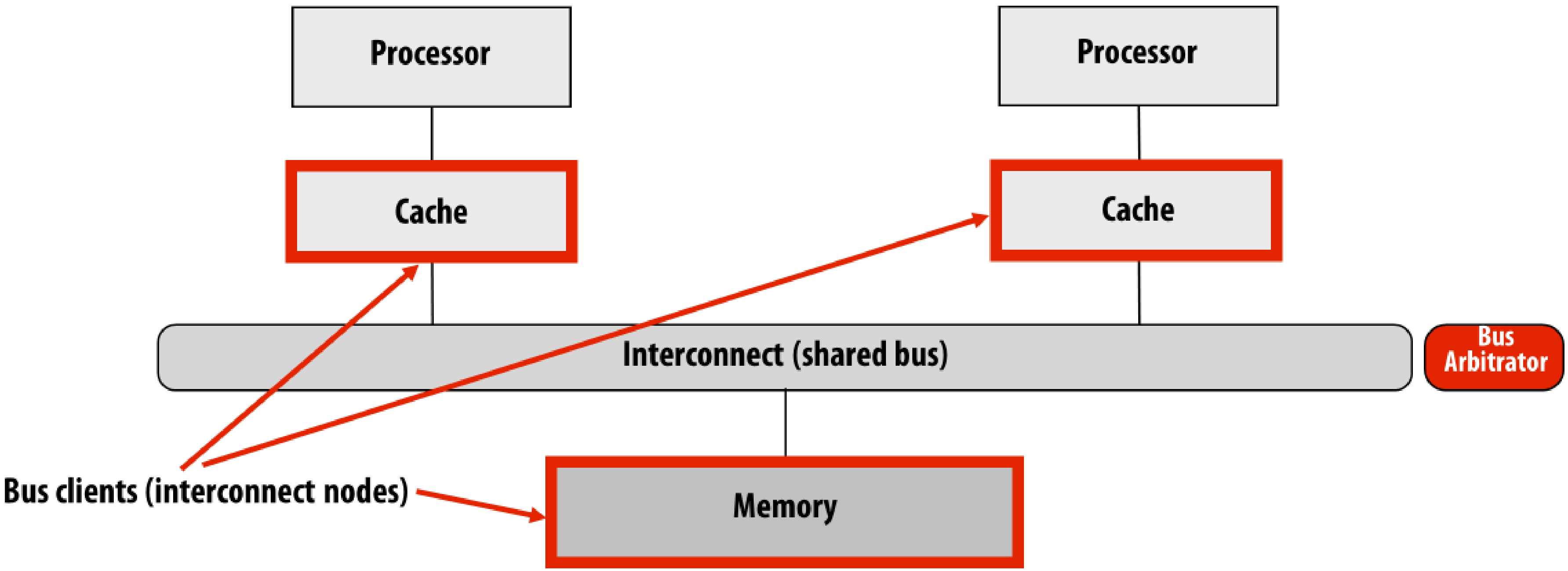
Interconnection Networks

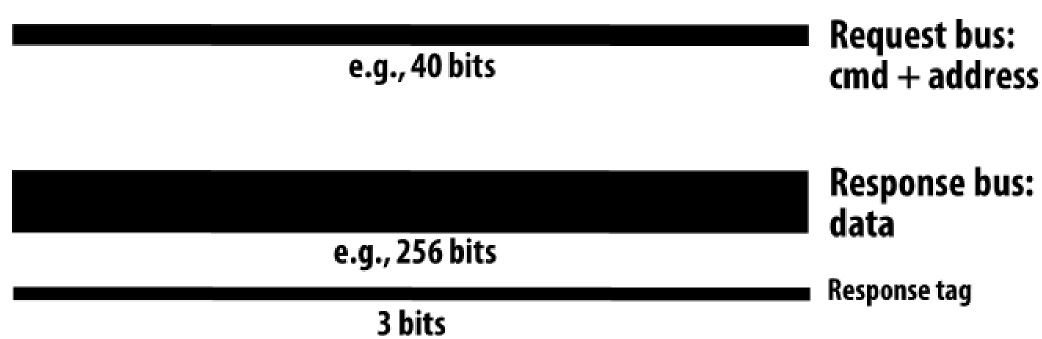
Computer Architecture: Design and Simulation
CMU 15-346, Spring 2022

Basic system design from previous lectures



Bus interconnect:

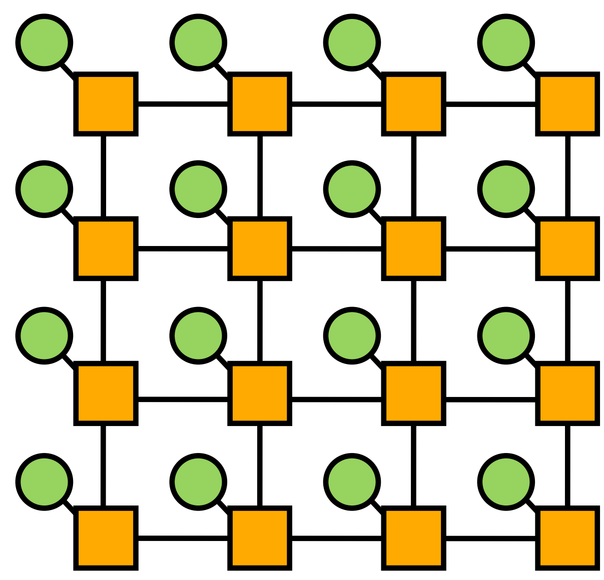
All nodes connected by a shared set of wires



Properties of interconnect topology

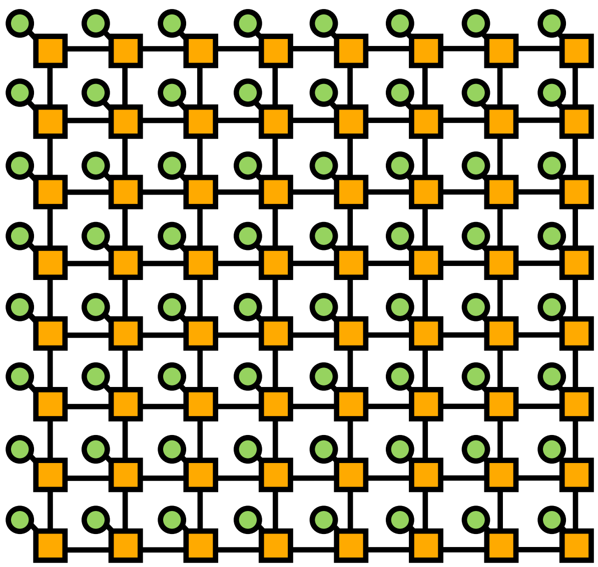
- **Routing distance**
  - Number of links ("hops") along a route between two nodes
- **Diameter: the maximum routing distance**
- **Average distance: average routing distance over all valid routes**

Example:
diameter = 6

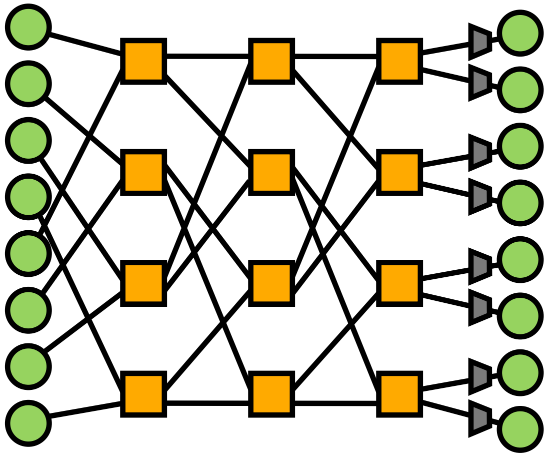


Properties of interconnect topology

- Direct vs. indirect networks
  - Direct network: endpoints sit "inside" the network
  - e.g., mesh is direct network: every node is both an endpoint and a switch



Direct network



Indirect network

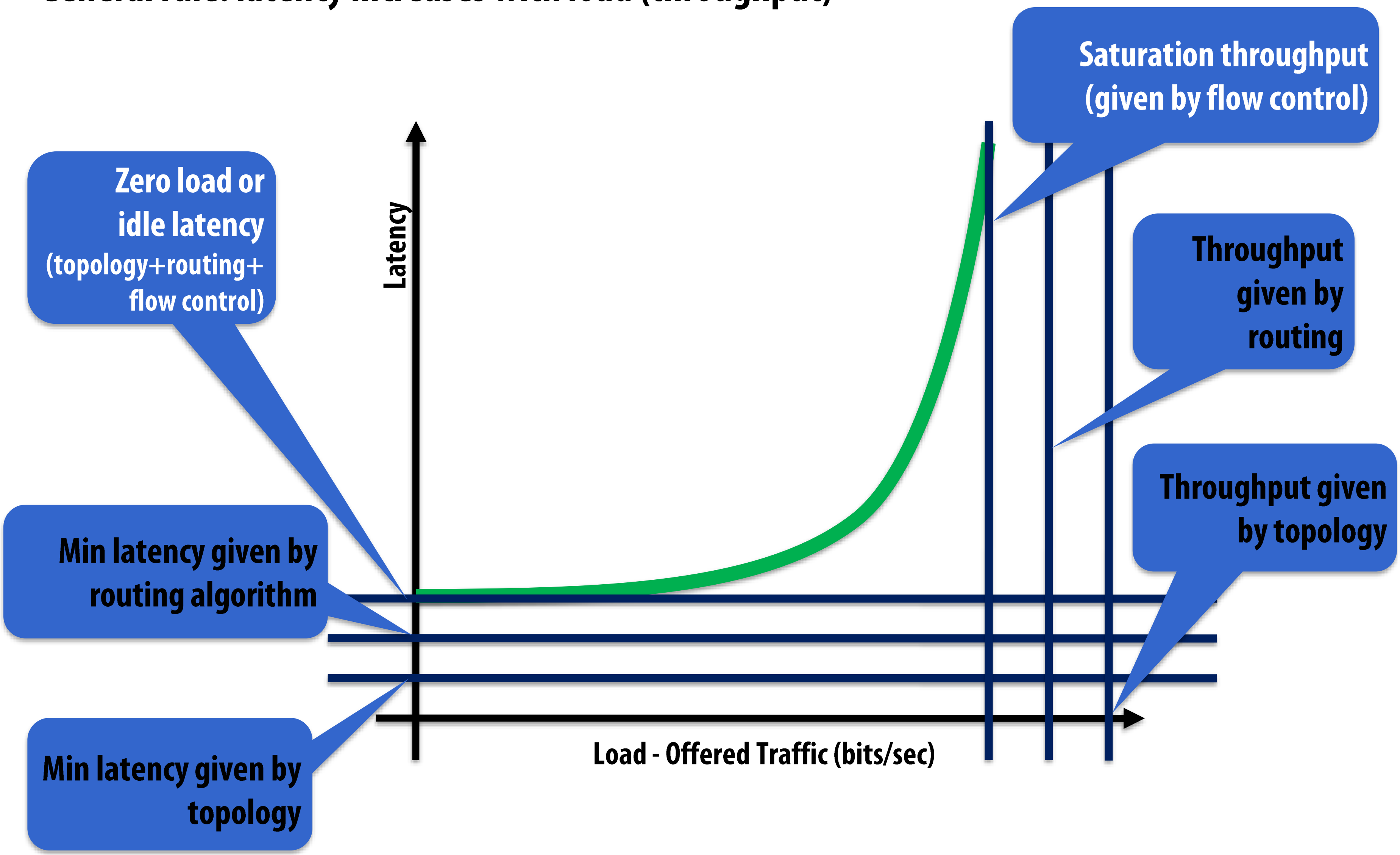
Properties of an interconnect topology

- **Bisection bandwidth:**
  - Common metric of performance for recursive topologies
  - Cut network in half, sum bandwidth of all severed links
  - Warning: can be misleading as it does not account for switch and routing efficiencies

- **Blocking vs. non-blocking:**
  - If connecting any pairing of nodes is possible, network is non-blocking (otherwise, it's blocking)

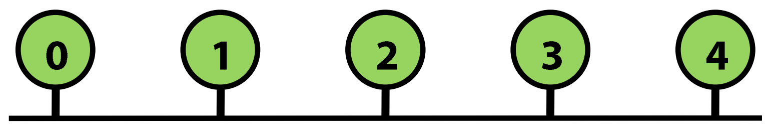
Load-latency behavior of network

General rule: latency increases with load (throughput)



Bus interconnect

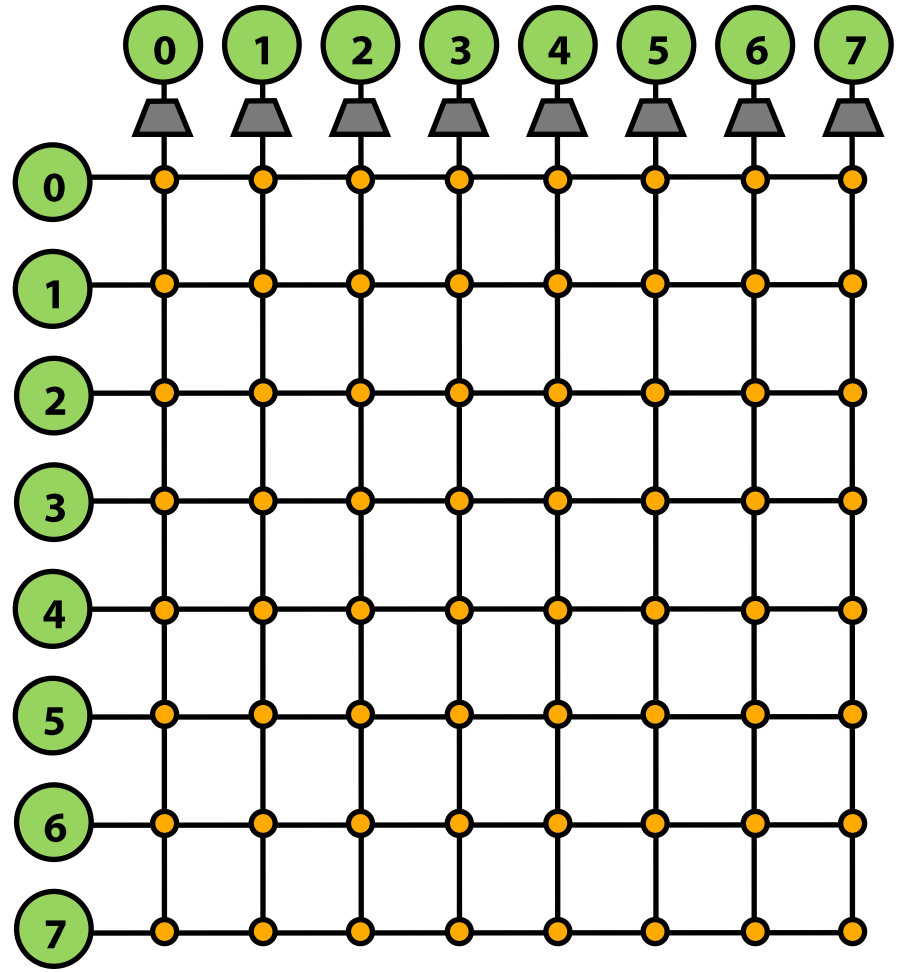
- Good:
  - Simple design
  - Cost effective for a small number of nodes
  - Easy to implement coherence (via snooping)



- Bad:
  - Contention: all nodes contend for shared bus
  - Limited bandwidth: all nodes communicate over same wires (one communication at a time)
  - High electrical load = low frequency, high power

Crossbar interconnect

- Every node is connected to every other node (non-blocking, indirect)
- Good:
  - $O(1)$ latency and high bandwidth
- Bad:
  - Not scalable: $O(N^2)$ switches
  - High cost
  - Difficult to arbitrate at scale

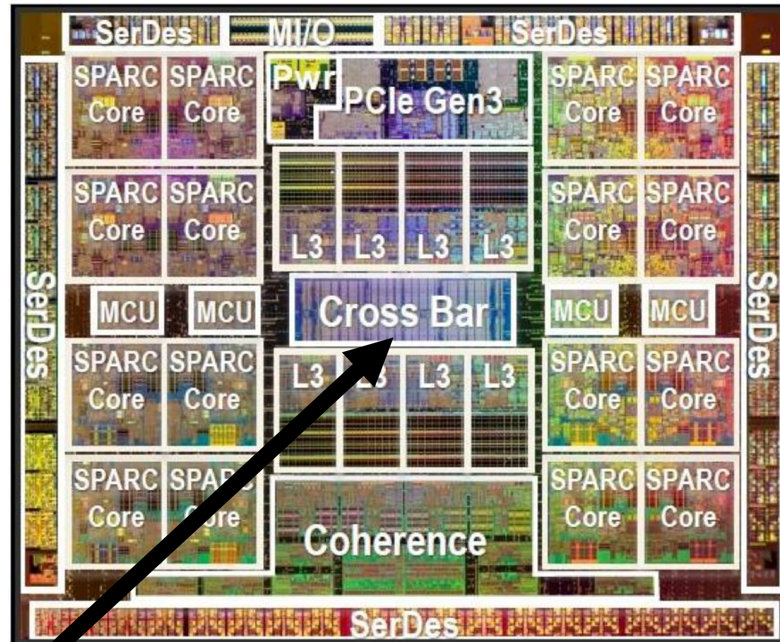


8-node crossbar network (N=8)

Note: in this network illustration, each node is drawn twice for clarity (at left and at top)

Crossbars were used in recent multi-core processing from Oracle (previously Sun)

Sun SPARC T2 (8 cores, 8 L2 cache banks)

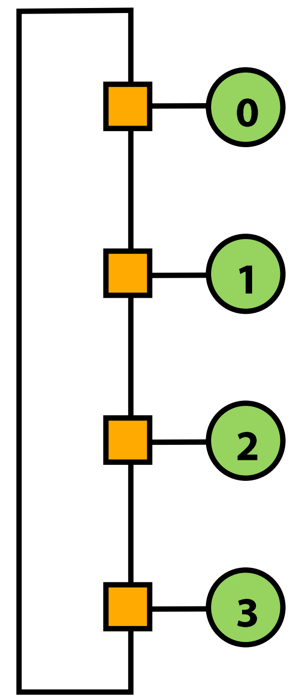


Oracle SPARC T5 (16 cores, 8 L3 cache banks)

Note that crossbar (CCX) occupies about the same chip area as a core

Ring

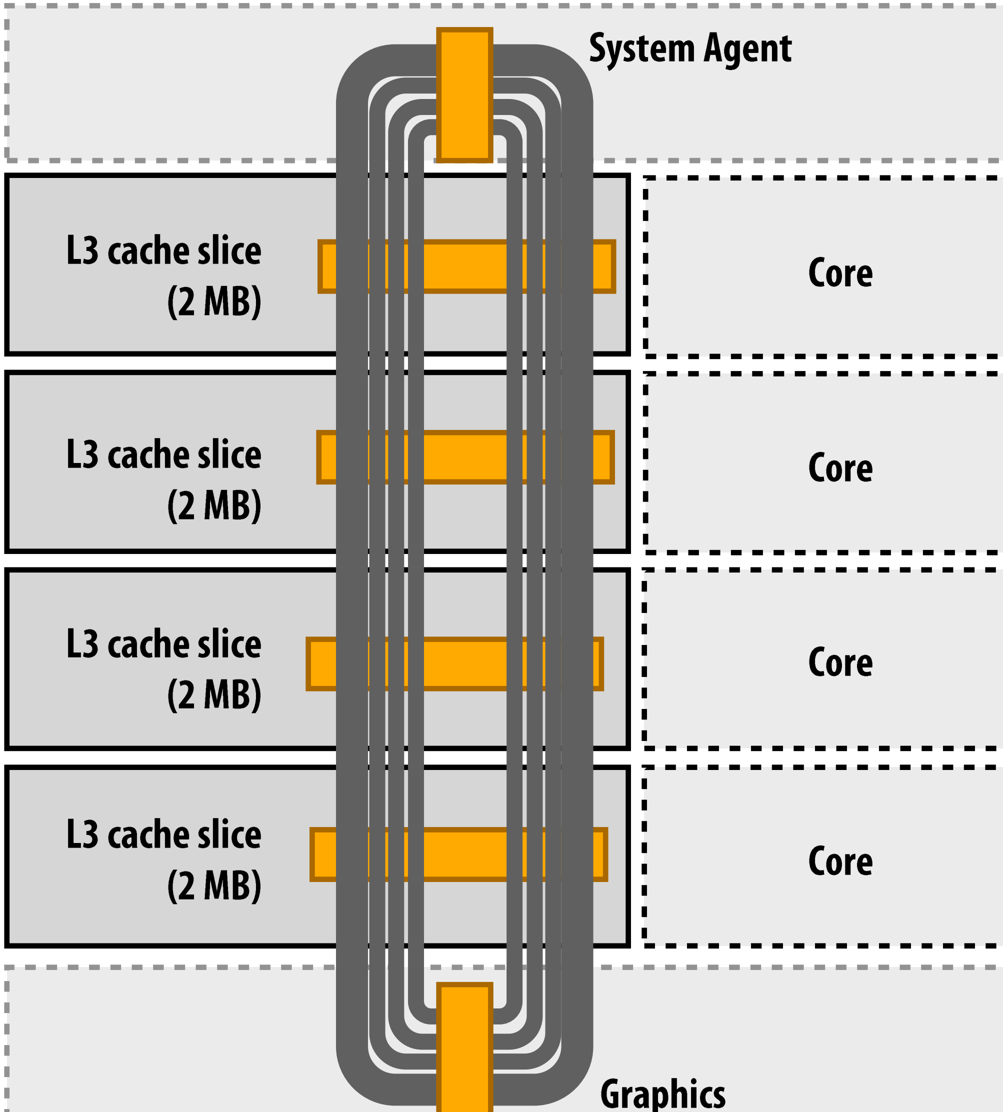
- Good:
  - Simple
  - Cheap: O(N) cost
- Bad:
  - High latency: O(N)
  - Bisection bandwidth remains constant as nodes are added (scalability issue)



- Used in recent Intel architectures
  - Core i7
- Also used in IBM CELL Broadband Engine (9 cores)

Intel's ring interconnect

Introduced in Sandy Bridge microarchitecture



- Four rings
  - request
  - snoop
  - Ack
  - data (32 bytes)
- Six interconnect nodes: four "slices" of L3 cache + system agent + graphics
- Each bank of L3 connected to ring bus twice
- Theoretical peak BW from cores to L3 at 3.4 GHz is approx. 435 GB/sec
  - When each core is accessing its local slice

Routing and Flow Control

- Consider a bus
  - Routing is simple
  - Flow control is arbitration

Granularity of communication

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

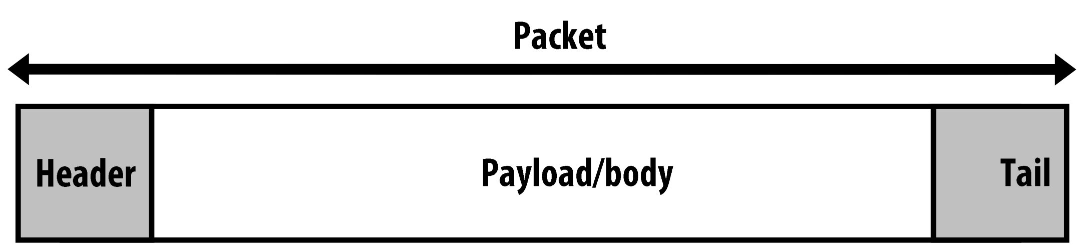
- **Packet**
  - Unit of transfer for network
  - Can be transmitted using multiple flits (will discuss later)

- **Flit (flow control digit)**
  - Packets broken into smaller units called "flits"
  - Flit: ("flow control digit") a unit of flow control in the network
  - Flits become minimum granularity of routing/buffering

Packet format

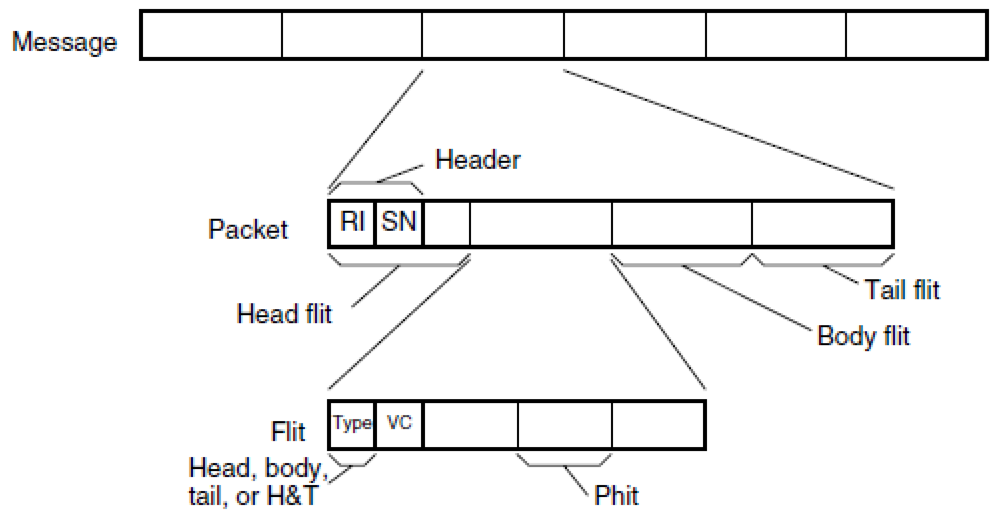
- A packet consists of:
  - Header:
    - Contains routing and control information
    - At start of packet to router can start forwarding early
  - Payload/body: containing the data to be sent
  - Tail
    - Contains control information, e.g., error code
    - Generally located at end of packet so it can be generated “on the way out”

(sender computes checksum, appends it to end of packet)



Different Packets

- **Simple example formats:**



Phit – physical transfer digits

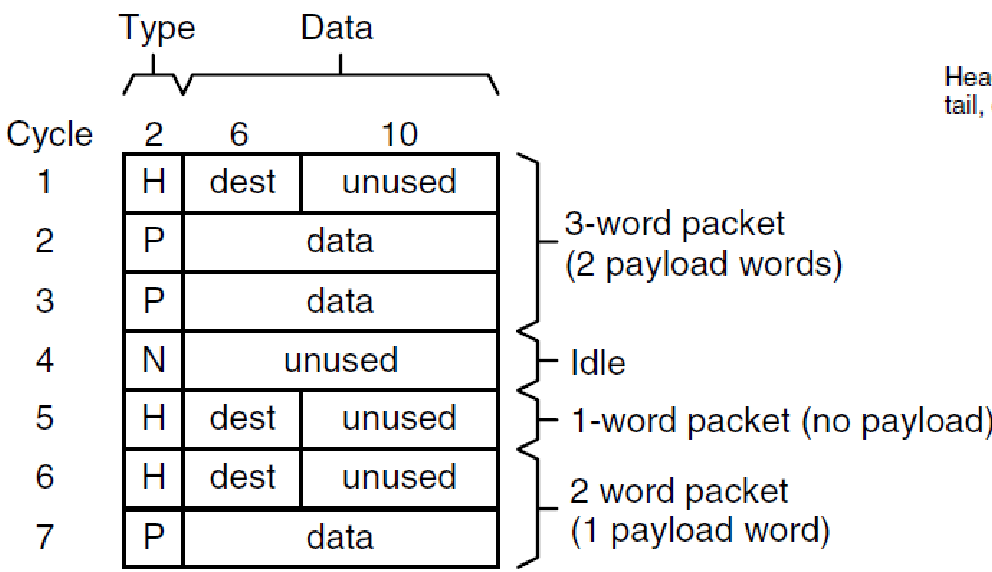


Fig – Principles and Practices of Interconnection Networks, Dally and Towles

Flow Control

- Simple to complex
  - Bufferless – drop or misroute packets
  - Circuit switched – buffer the header as it allocates the path
  - Buffered – store some or all of packet in network

Flow Control with Buffering

- Store and Forward
  - Store packet at switch before sending

- Cut through
  - Send early if possible

- Wormhole
  - Only send on progress

Virtual Channels

- A channel is a physical link
- Virtually create more channels
  - Multiplex between different waiters to make progress

- Can also support traffic types, etc

Routing

- Greedy – shortest path
- Random – pick among paths
- Weighted Random – weight shortest when picking
- Adaptive – pick based on load

- Features:
  - Deterministic – do same thing
  - Oblivious – may do different thing (i.e. random)
  - Adaptive – consider network state

Examples

- Dimension order – XY
- Valiant's Aglorithm
  - Randomly pick an intermediate node to send
- Adaptive
  - Local versus global knowledge
  - Backpressure
  - Danger of livelock