Switching

How a message crosses the network from one node to another

Circuit switching

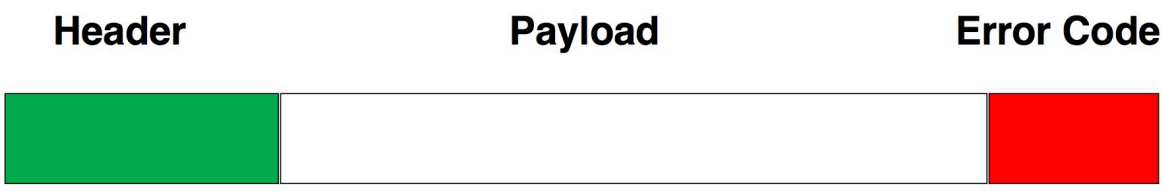
- A path is established from the source to the destination
- (no one else can use those links)
- All packets will take this path
- Phone Analogy (wired)
  + Faster and higher bandwidth
  - setting up and bringing down links slow

Packet switching

- A message is split into a sequence of packets that can be sent on different paths
- Better use of network resources
  + No setup, bring down time
  - Potentially slower (must dynamically switch)

Packet Switching, Packet Format

- **Header**
  - routing and control information
- **Payload**
  - carries data (non HW specific information)
  - can be further divided (framing, protocol stacks...)
- **Error Code**
  - generally at tail of packet so it can be generated on the way out



Packet Switching, routing

Two basic approaches to routing packets, based on what a switch does when a packet arrives

1) Store-and-forward

2) Cut-through
   - Virtual cut-through
   - Wormhole

Packet Switching: Store-and-Forward

- A packet is stored entirely before being forwarded

- **Drawbacks**
  - Need of a lot of memory to store incoming packets

- **Advantage**
  - Switching is done step by step.
  - Little danger of blocking

Packet Switching: Cut Through

A packet can arrive partially in a switch and leave its tail on the other nodes

- It can be on more than two switches

The re-send decision shall be taken immediately

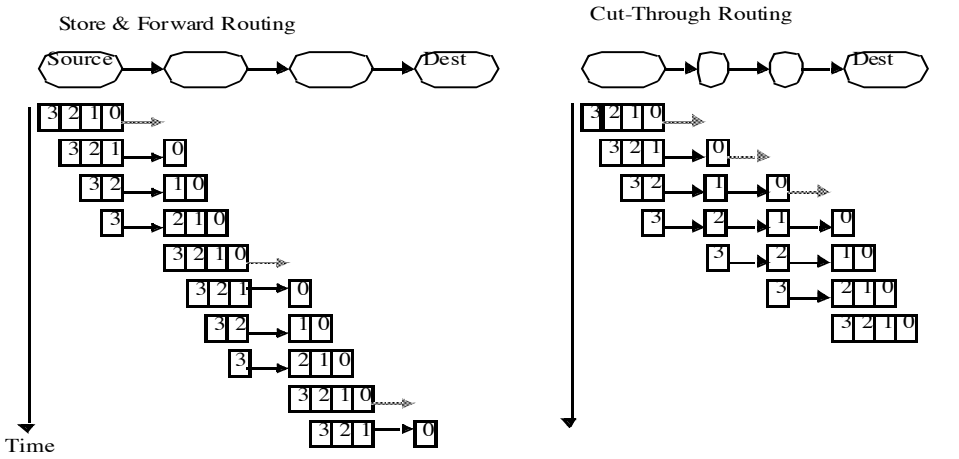
What happens if the head blocks?

- ***Cut-through***: collect the rest of the message where is the head

  Tends towards the store-and-forward model in case of strong contention

- ***Wormhole***: If the head blocks, the whole message hangs

Store & Forward vs Cut-Through



$h(n/b + \Delta)$ vs $n/b + h\,\Delta$

h: number of hops

n: message's size

b: bandwidth

Δ: routing delay per hop

Routing Algorithms

How do I know where a packet should go?

- Topology does not determine routing

Routing algorithms

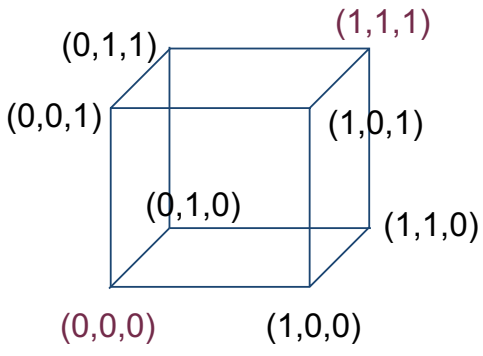
1) Arithmetic
2) Based on source
3) With a table (*table lookup*)
4) Adaptive: The route is determined by the state of the network (taking into account the contention)

Arithmetic Routing

On a regular topology, use simple arithmetic to determine the path

E.g., XY routing in a 3D torus

- The packet header contains a signed offset to the destination (by dimension)
- For each jump, switch +/- to reduce the offset in the dimension
- When x == 0 and y = 0, then we have reached the destination



Source-Based and Table-Based Routing

Source-based routing

- The source specifies the output port for each switch on the route
- Simple switches
- No control state
- Header removed whenever switch is crossed
- Used by Myrinet
- Can not become adaptive

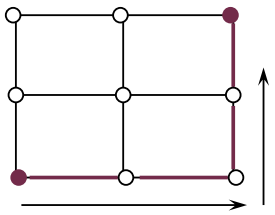
Table-based routing (*Table Lookup*)

- Small header: Contains a field that is an index in a table for the output port
- Large tables that must be kept up to date

Deterministic or Adaptive Routing

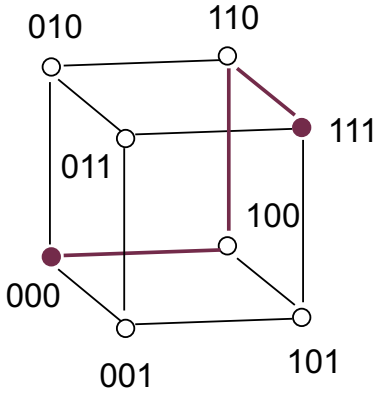
Deterministic

- Follows a pre-defined path
  - K-ary d-cube: dimensional routing
    (x1, y1) → (x2, y2)
    First Dx = x2 - x1,
    Then Dy = y2 - y1,
  - Tree: common ancestor
  - Simple algorithms can become blocking



Adaptive

- Route determined by contentions on the output port
  - Essential for fault tolerance
  - At least multi-paths
  - Can improve network utilization



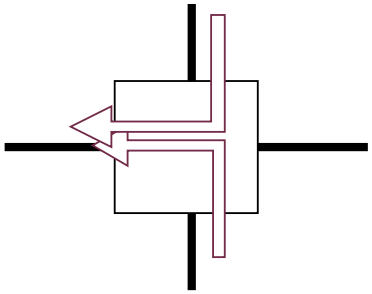
Contention

Two packets trying to use the same link at the same time

- Limited bufferization
- Loss?

Most networks of parallel machines hang

- Traffic can be returned to the source



Communication Performance: Latency

$Time(n)_{s-d}$ = overhead + routing delay + channel occupation + contention delay

- Overhead: time required to initiate sending and receiving a message
- Occupation = $(n + n_e) / b$
  - n: data size
  - n_e: packet envelop's size
- Routing delay
- Contention

Communications Performance: Bandwidth

What affects the local bandwidth?

- Packet density $b \times n/(n + n_e)$
- Routing delay $b \times n / (n + n_e + w\Delta)$

  Δ: number of cycles to wait for a routing decision

  w: channel width
- Contention

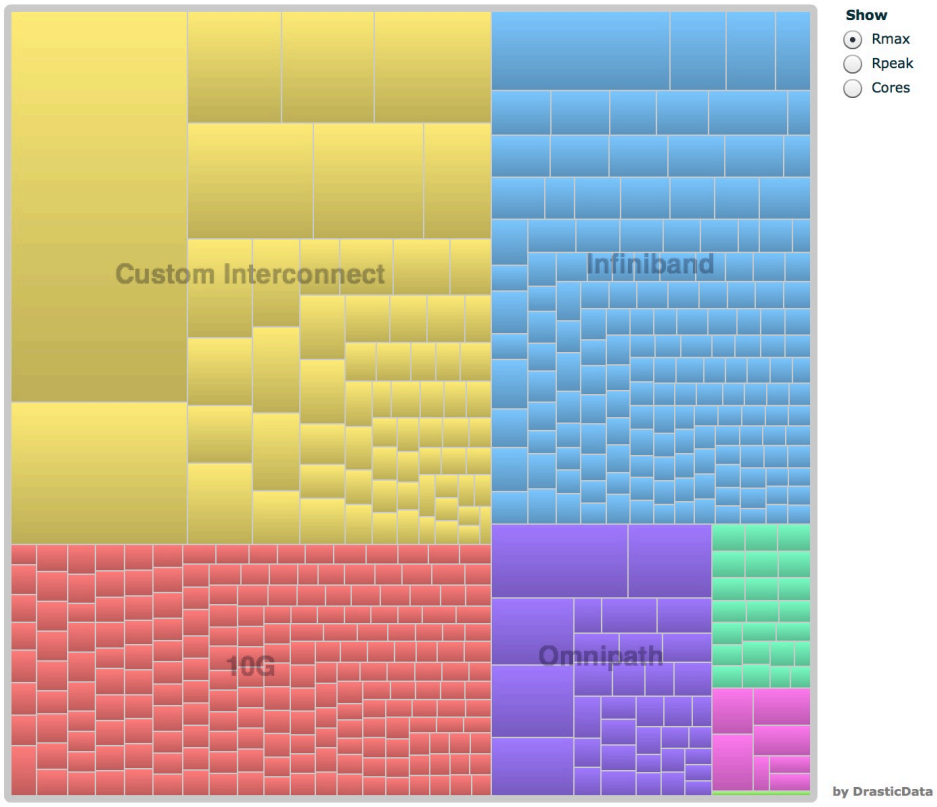
Aggregate bandwidth

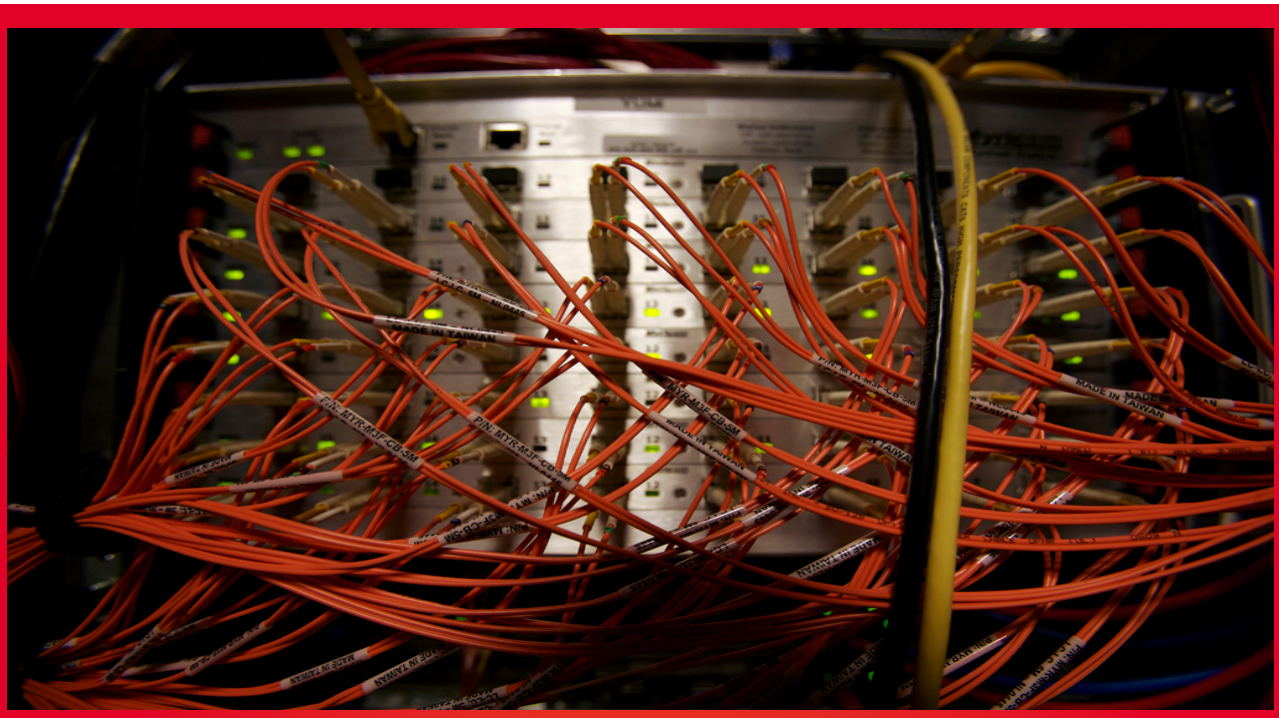
- Bisection width
  - Sum of the bandwidths of the smallest set of links that partition the network
  - Poor if non-uniform distribution of communications
- Total bandwidth of all channels

Ethernet, Infiniband, Omnipath

	Ethernet	InfiniBand	Omnipath
Commonly used in what kinds of network	Local area network(LAN) or wide area network(WAN)	Interprocess communication (IPC) network	Interprocess communication (IPC) network
Transmission medium	Copper/optical	Copper/optical	Copper/optical
Bandwidth	1Gb/10Gb	2.5Gb~120Gb	100Gb
Latency	High	Low	Low
Cost	Low	High	High

Top500





INVENTEURS DU MONDE NUMÉRIQUE