Network devices

George Bosilca
bosilca@cs.utk.edu

Long time ago …

- First message ?
  - 24 May 1844 Washington to Baltimore
  - Wire connections and train rails accelerate the world (west) conquest.
- Everything is about connecting …

Networks

- What really generate the performances:
  - Direct (p2p) vs. indirect (multi-hop)
  - Topology (bus, ring, DAG, …)
  - Routing algorithms
  - Switching (aka multiplexing)
  - Choice of media (copper, coax, fiber)
- What really matter …
  - Latency
  - Bandwidth
  - Cost
  - Reliability

System Area Network

- Connecting close computers
- Target high bandwidth and low latency
- May offer “in order” packet delivery
- Technologies:
  - Infiniband
  - Myrinet
  - SCI
  - …

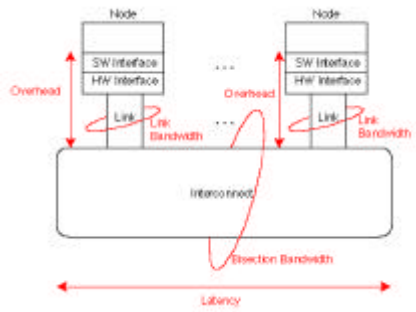
Local Area Network

- Basically single protocol: Ethernet
  - 10Mbps original version
  - 100Mbps & 1Gbps
  - 10Gbps (tomorrow?)
- Carrier Sense Multiple Access
  - Collision detection + random retransmission
  - Shared media (wires topology)
  - Ethernet switches (star topology)
- Large packet + broadcast mode
- Destination based routing
- Connectionless

Wide Area Network

- 155Mbps * factor
- 53 byte packet (48 payload + 5 header)
- Virtual circuit
- ATM protocols quite complex but switching is simple
- SONET: packaging multiple ATM packet in a single longer SONET packet
  - Internet core actually 100% SONET
- Could be replaced by IP/MPLS.

Network Performance

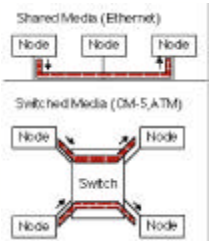


Connection-Based vs. Connectionless

- Telephone: operator sets up connection between the caller and the receiver
  - Once the connection is established, conversation can continue for hours
- Share transmission lines over long distances by using switches to multiplex several conversations on the same lines
  - "Time division multiplexing" divide B/W transmission line into a fixed number of slots, with each slot assigned to a conversation
- Problem: lines busy based on number of conversations, not amount of information sent
- Advantage: reserved bandwidth

Connecting Multiple Computers

- Shared Media vs. Switched: pairs communicate at same time: "point-to-point" connections
- Aggregate BW in switched network is many times shared
  - point-to-point faster since no arbitration, simpler interface
- Arbitration in Shared network?
  - Central arbiter for LAN?
  - Listen to check if being used ("Carrier Sensing")
  - Listen to check if collision ("Collision Detection")
  - Random resend to avoid repeated collisions; not fair arbitration;
  - OK if low utilization



(A. K. A. data switching interchanges, multistage interconnection networks, interface message processors)

Connection-Based vs. Connectionless

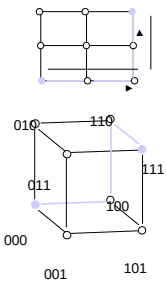
- Connectionless: every package of information must have an address => packets
  - Each package is routed to its destination by looking at its address
  - Analogy, the postal system (sending a letter)
  - also called "Statistical multiplexing"
  - Note: "Split phase buses" are sending packets

Routing Messages

- Shared Media
  - Broadcast to everyone
- Switched Media needs real routing. Options:
  - Source-based routing: message specifies path to the destination (changes of direction)
  - Virtual Circuit: circuit established from source to destination, message picks the circuit to follow
  - Destination-based routing: message specifies destination, switch must pick the path
    - deterministic: always follow same path
    - adaptive: pick different paths to avoid congestion, failures
    - Randomized routing: pick between several good paths to balance network load

Deterministic Routing Examples

- mesh: dimension-order routing
  - (x_1, y_1) -> (x_2, y_2)
  - first $\Delta x = x_2 - x_1$,
  - then $\Delta y = y_2 - y_1$,
- hypercube: edge-cube routing
  - $X = x_o x_1 x_{2\ldots} x_n$ -> $Y = y_o y_1 y_{2\ldots} y_n$
  - $R = X$ xor Y
  - Traverse dimensions of differing address in order
- tree: common ancestor
- Deadlock free?



Store and Forward vs. Cut-Through

- Store-and-forward policy: each switch waits for the full packet to arrive in switch before sending to the next switch (good for WAN)
- Cut-through routing or worm hole routing: switch examines the header, decides where to send the message, and then starts forwarding it immediately
  - In worm hole routing, when head of message is blocked, message stays strung out over the network, potentially blocking other messages (needs only buffer the piece of the packet that is sent between switches).
  - Cut through routing lets the tail continue when head is blocked, accordioning the whole message into a single switch. (Requires a buffer large enough to hold the largest packet).

Cut-Through vs. Store and Forward

- Advantage
  - Latency reduces from function of:

    number of intermediate switches X by the size of the packet
    to
    time for 1st part of the packet to negotiate the switches
    + the packet size ÷ interconnect BW

Congestion Control

- Packet switched networks do not reserve bandwidth; this leads to *contention* (connection based limits input)
- Solution: prevent packets from entering until contention is reduced (e.g., freeway on-ramp metering lights)
- Options:
  - Packet discarding: If packet arrives at switch and no room in buffer, packet is discarded (e.g., UDP)
  - Flow control: between pairs of receivers and senders; use feedback to tell sender when allowed to send next packet
    - Back-pressure: separate wires to tell to stop
    - Window: give original sender right to send N packets before getting permission to send more; overlaps latency of interconnection with overhead to send & receive packet (e.g., TCP), adjustable window
  - Choke packets: aka "rate-based"; Each packet received by busy switch in warning state sent back to the source via choke packet. Source reduces traffic to that destination by a fixed % (e.g., ATM)

Protocols: HW/SW Interface

- Internetworking: allows computers on independent and incompatible networks to communicate reliably and efficiently;
  - Enabling technologies: SW standards that allow reliable communications without reliable networks
  - Hierarchy of SW layers, giving each layer responsibility for portion of overall communications task, called protocol families or protocol suites
- Transmission Control Protocol/Internet Protocol (TCP/IP)
  - This protocol family is the basis of the Internet
  - IP makes best effort to deliver; TCP guarantees delivery
  - TCP/IP used even when communicating locally: NFS uses IP even though communicating across homogeneous LAN

TCP/IP

- Application send a message
- TCP break it on 64KB segments + 20 bytes header
- IP add 20 bytes header and send it to the network
- Ethernet break it again in 1500B (MTU) packets with headers, trailers
- Headers, trailers have length field, destination, window number, version, …

Fast Path

- Initial idea allowing inside the kernel a fast path from the code directly to the network device
  - Come from distributed shared memory computer
  - Removing the overhead for small messages (as a page request).
  - Small pages of memory from the network interface directly mapped into the memory. Data is copied inside and then an ioctl raised.
- Same idea can be used for long messages ?

Fast Path

- Locking a memory area and allowing to the network interface direct access to that area
  - Lock to avoid swapping the area from the memory on context switch.
  - Limited amount of available memory for locking
- Drastically improve the performances for data transfer as there is no additional memory copy operations (called zero memory copy).
- Decrease the capacity of the network interface to be shared between several applications (dedicated mode).

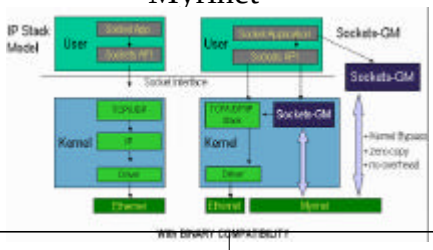
SCI

- Scalable Coherent Interface (Dolphin)
- Very low latency and high bandwidth (~5.3Gb/s)
- Offer support for both the shared-memory and message passing paradigms.
- Bus share
- Remote writes 10 times faster than remote reads

Myrinet

•2GB/s bidirectional links (optical or not)
•16 crossbar switches
•Wormhole routing with
  •Link flow control
  •Automatic dead-lock detection
•Programmable interfaces
•Drivers GM, BIP, PM

Myrinet



Sustained one way	248MB/s
Sustained 2way	489MB/s
Latency	6.3 µs
CPU utilization	Less than 0.5 µs