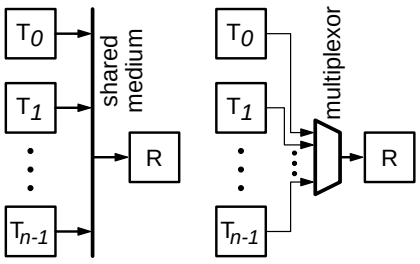
Packet Switch Architecture

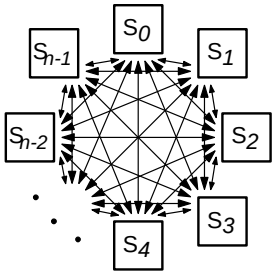
The Hardware Architect's perspective on
High-Speed Networking Problems

Manolis Katevenis
University of Crete & FORTH

(1984-now: course includes 20+ years of research…)

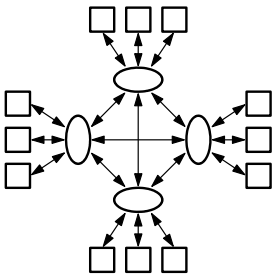
Communication Networks





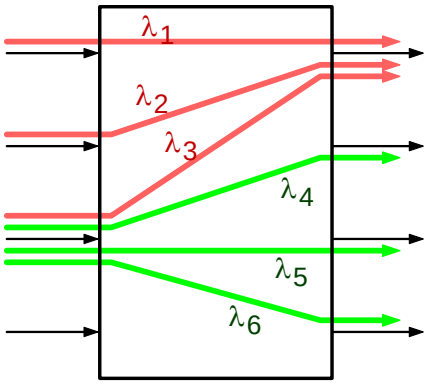
all-to-all interconnection

- Multi-party communication under resource constraints (total link length, link throughput, multiplexors, buffers,…)
- Receiver capacity is far below the aggregate capacity (rate) of all transmitters (and symmetrically for each transmitter relative to all rcvrs)



multi-stage interconnection

Interdependent Constraints



- $\lambda_1+\lambda_2+\lambda_3 \leq 100\%$ (output contention): 33%+33%+33% ? (fairness)
- $\lambda_3+\lambda_4+\lambda_5+\lambda_6 \leq 100\%$ (input rate limitation): 25%+25%+25%+25% ?
- $\lambda_3=25\% \Rightarrow \lambda_1+\lambda_2 = 75\% \Rightarrow$ $\lambda_1+\lambda_2 = 37.5\% + 37.5\%$? (max-min fairness) ?
- or: $\lambda_1+\lambda_2+\lambda_3 = 50+50+0\%$, $\lambda_3+\lambda_4+\lambda_5+\lambda_6 = 0+33+33+33\%$? (maximum utilization) ?

Distributed Control Problem

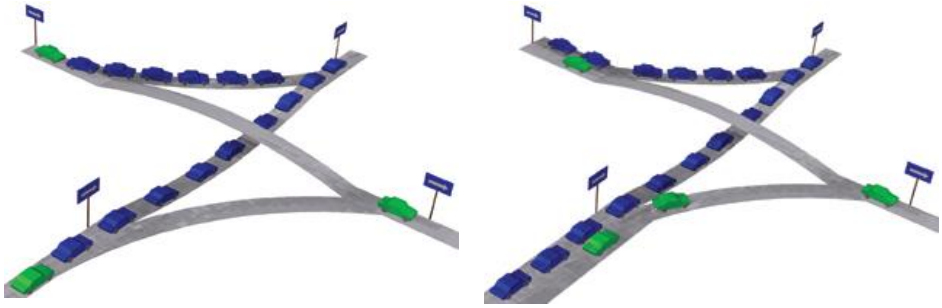
- Who determines the solution? When? Where? How?
- Geographically distributed (traditional networking), or…
  (long time scales may allow software solution)
- Microelectronic cores in a chip or in a box or in a room
  (short time scales demand hardware speed)
- Inputs do not know of each other's intentions when they start transmitting
- Distance, speed, complexity preclude centralized solution

Reactions to Output Contention

- Short-Term (within a round-trip time)
  - buffer conflicting packets (in the network), or…
  - drop conflicting packets (and retransmit?)
- Long-Term
  - flow control (congestion management) (after the fact)
  - admission control (beforehand)

⇒ High-speed Memories + Distributed Control

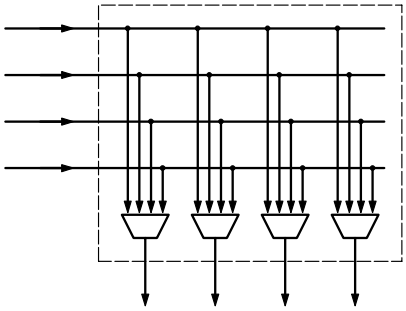
Head-of-Line Blocking – Multiple Queues

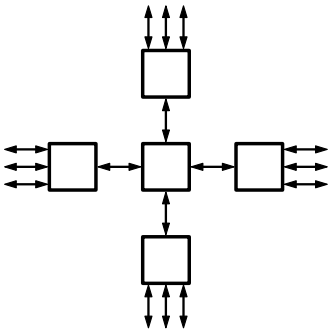


Head packets in a queue may block other packets behind them, even if those behind are destined elsewhere ⇒

⇒ want *Multiple Queues* (within High-Speed Memories)

Interconnection Fabrics





- Single-stage all-to-all (crossbar) performs well but its N^2 cost is too high for large N
- Multi-stage (hierarchical) fabrics trade cost for internal blocking ("all lines are busy –please try later") (locality of traffic?)

Technology Outlook: NoC, Commodity Switches

- Ubiquitous, Switch-based Interconnection Networks
  - buses have inherent performance limitations
  - switch-based interconnects proliferate from WAN to LAN, then to SAN (storage-area or system-area), then to processor-memory-I/O interconnects and to Networks-on-Chip (NoC)
- New Market: Next Generation IT Infrastructure
  - chip multiprocessors (CMP), using networks-on-chip (NoC)
  - cluster/blade-based systems and servers

$\Rightarrow$ Commodity Switches: Mass Market, sharp price drop

  - fabrics of inexpensive, mass-produced switches will replace the current very expensive, custom-made telco switches/routers (analogous to workstation clusters replacing supercomputers)
  - what should be their ("RISC-style") architecture ???
- NoC switches need to be quite less expensive – how ?