CS 5224

Architecture of a Packet Switch

Dr. Chan Mun Choon
School of Computing, National University of Singapore

References

- Some slides are taken from the following source:
  - S. Keshav, "An Engineering Approach to Computer Networking", Chapter 8: Switching
- Readings
  - V.P. Kumar, T.V. Lakshman, and D. Stiliadis, "Beyond Best Effort: Router Architectures for the Differentiated Services of Tomorrow's Internet," IEEE Communications Magazine, May 1998, pp. 152-164.

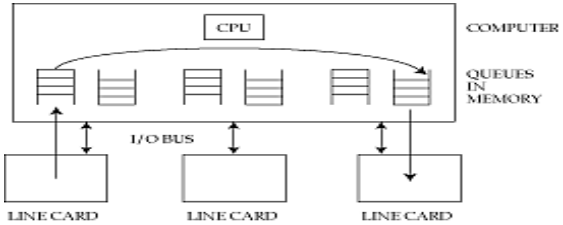
Outline

- Difference generations of packet switch design
- Architecture/Components of a packet switch
  - Flow Identification
  - Routing Lookup
  - Scheduling/Buffer Management

Three generations of packet switches

- Different trade-offs between cost and performance
- Represent evolution in switching capacity, rather than in technology
  - With same technology, a later generation switch achieves greater capacity, but at greater cost
- All three generations are represented in products

First generation switch



- Based on a single general purpose CPU and a real-time OS
- Assumes multi-protocol operations and that routers cannot be optimized for a specific protocol
- Packets arriving at the interface cards are forwarded to the CPU

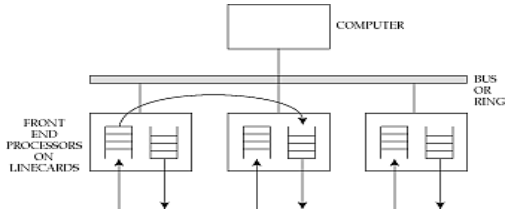
First generation switch

- Packets are transmitted twice over the shared bus
- Performance heavily depends on the throughput of the
  - Shared Bus
  - Forwarding speed of CPU (including operating system overhead)
- Some Ethernet switches and "cheap" packet routers

Example

- First generation router built with 133 MHz Pentium
  - Mean packet size 500 bytes
  - Interrupt takes 10 microseconds, word access take 50 ns
- A copy loop has 4 instructions + 2 memory accesses = 130.08 ns
  - Copying packet takes 500/4 *130.08 = **16.26 µs**
- Interrupt takes **10 µs**
- Per-packet processing time takes 200 instructions = **1.504 µs**
- Total time = 27.764 µs
  - => speed is 144.1 Mbps ((500 * 8)/27.764 Mbps)

Second generation switch



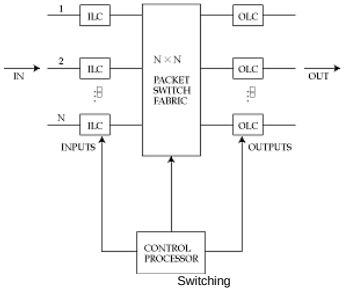
- Distribute forwarding computation by adding routing intelligence to line cards
- Add route cache to interface card so that subsequent packets of the same connection can be forwarded without consulting the controller
  - ATM switch guarantees hit in lookup cache
  - IP routers, cache miss takes more time to process

Second generation switch

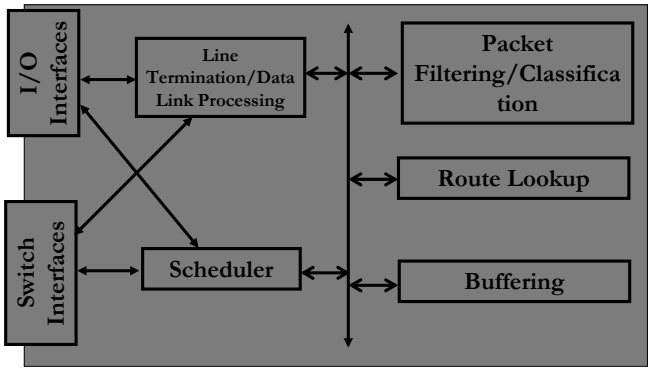
- Cache entry is on a per-connection or per-route basis
- Bottlenecks:
  - Traffic dependent
  - Shared bus architecture
  - General purpose CPU

Third generation switches

- Bottleneck in second generation switch is the bus (or ring)
- Third generation switch provides parallel paths (fabric)



Processing Architecture



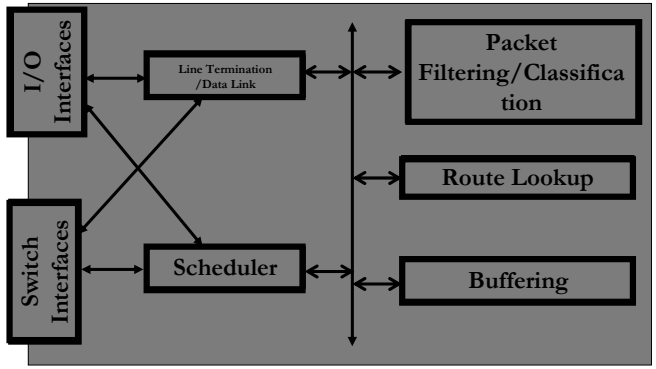
Example

- CRS-1 is the top-of-the-line high speed router from Cisco (www.cisco.com)
  - Interface Modules
    - OC768c (39.82Gbps), OC-192, OC-48, 10GE
    - ***OC - Optical Carrier*** (OC-1 = 51.85 Mbps)
  - Modular Services Card
    - Two processors, one for ingress and one for egress
  - Route Processor
    - System management, control-plane management
  - Switching Fabric: up to 1152 40Gb/s slots

CRS-1 Modular Services Card

- Forwarding features
  - Access control lists (ACLs/xACLs)
  - Quality of service/class of service (QoS/CoS) using Modular QoS CLI (MQC)
  - IP packet classification/marking
  - Queuing (both ingress and egress)
  - Policing (both ingress and egress)
  - Diagnostic and network management support

Processing Architecture



Typical "Path" of a Packet

At the input queue

1. A packet arrived at the input interface
2. Perform line termination and protocol conversion
3. Perform packet classification
4. Perform route lookup
5. Buffer packet
6. Packet schedule for transmission to switching fabric

At the output queue

1. Perform packet classification
2. Buffer packet
3. Schedule packet for transmission to output link

Operation Costs

- Many operations are per-packet
  - Packet classification
  - Route lookup
  - Scheduling
  - Buffering and switching
- Some function costs are also bit sensitive
  - Buffering and switching
- Performance of a packet switch therefore has to scale with both bit-per-second and packet-per-second
- A 1Gbps link can operate at
  - 83.3K pkt/sec (for 1.5K bytes packets)
  - 3.1M pkt/sec (for 40 bytes packets)

Packet Size in the Internet

- This is from a packet trace done in 1997 over the MCI backbone (Keshav'98).
  - 40 bytes ~ 50%
  - 500 bytes ~ 10%
  - 1500 bytes ~ 10%

Packet Filtering/Classification

- Classify packet according to information the packet header
- For ATM
  - 12-bit VPI + 16-bit VCI
- For IP: the 6 tuples (or more …)
  - 32-bit source IP
  - 32-bit destination IP
  - 16-bit source port
  - 16-bit destination port
  - 8-bit protocol
  - 8-bit TOS
  - Total = 112 bits
- Classification cost increases with # of bits classified

Packet Filtering/Classification

- Possible objectives:
  - Allow/Reject: some packets may not be allowed to pass through
    - Access control, firewall
  - Rate control: if there are too many packets of certain types, drop them
    - leaky bucket
  - Accounting
    - Billing, network measurements
  - Differentiation: classify packet and tag them so that they can be treated differently later
    - by the same switch or some other switches downstream)

Packet switching

- Recall that in a circuit switch, path of a sample is determined at time of connection establishment
  - No need for a sample header--position in frame is enough
- In a packet switch, packets carry a destination field
  - Need to look up destination port on-the-fly
- Datagram
  - lookup based on entire destination address
- ATM Cell
  - lookup based on VCI
- Other than that, very similar

Route Lookup

- IP route table lookup was considered one of the most challenging operations during the forwarding process
- Longest Prefix Match
  - Forwarding entries are stored in the form <network address/mask, port>
  - A packet is routed to the port that matches the longest prefix in the forwarding entry
  - Take the entries <128.32.1.5/16,1>,<128.32.225.0/18,3>,<128.0.0.0/8,5>
  - A packet with destination 128.32.195.1 matches all three entries and can be routed to port 1,3 or 5
  - However, the match with the longest match is 128.32.225.0 and the packet will be routed to port 3

Why is IP Lookup Hard?

- Routing tables may contain many thousands of entries
  - 10K – 100K or more
- The number of lookups per second is large
  - There are many small (40 bytes) packets, > 1M per second (up to 3.1M packets/sec) for a 1Gbps link
- A packet can match multiple entries and the entry with the longest prefix match should be found
  - Worst case scenario: # of matches per second is product of number of entries and number of packets arrived per sec
  - Designing an efficient data structure is non-trivial
  - Current trend is towards hardware-implementation using TCAM

ATM vs. IP Lookup

- ATM is designed to enable cheap switching
  - Small and fixed packet header (16-bit address) for lookup
  - Fix packet length minimizes fragmentation by switch and reduces complexity of scheduling algorithm
- IP
  - Large packet header and address space (32-bit) and requires longest prefix match
  - Variable size packet length
- But …
  - Advances in route lookup technology makes IP lookup much cheaper
  - Inside a switch, IP packets are often fragmented into fixed size packets to ease buffering and switching complexity (implemented like an ATM switch)
  - IP routers are much more widely deployment, making it cheaper to build even if the complexity is higher