CSE 123A
Computer Networks

Fall 2005

Lecture 14:
IP Router Design
Good news
- I gave up on project #3 (I forgot about TG)
- I have your midterms
  » Mean = 38, median = 37, stdev=10
  » Weird distribution… near linear from min to max

Bad news
- One more homework (will be up tonight)
- Due next Thursday (extra time for vacation)

Misc news
- Changed the order of lectures so today is Router Design
- Probably going to be a shorter lecture (good? bad?)
Overview

- Router basics
- Interconnection architecture
  - Input Queuing
  - Output Queuing
  - Virtual output Queuing
  - Scheduling
- Future bottlenecks
- Case Studies
What’s in a router?

- Physical components
  - One or more input interfaces that receive packets
  - One or more output interfaces that transmit packets
  - A chassis (box + power) to hold it all

- Functions
  - Forward packets
  - Drop packets (congestion, security, QoS)
  - Delay packets (QoS)
  - Transform packets? (Encapsulation, Tunneling)
What a router does:
the normal case

1. Receive incoming packet from link input interface
2. Lookup packet destination in forwarding table (destination, output port(s))
3. Validate checksum, decrement ttl, update checksum
4. Buffer packet in input queue
5. Send packet to output interface (interfaces?)
6. Buffer packet in output queue
7. Send packet to output interface link
What a router looks like?

Cisco 2500

Capacity: <10Mbps

Linksys DEFSR81

Capacity: <10Mbps
What a router looks like (2)

Cisco GSR 12416
- Capacity: 160Gb/s
- Power: 4.2kW
- 6ft x 2ft x 2ft

Juniper M160
- Capacity: 80Gb/s
- Power: 2.6kW
- 2.5ft x 3ft x 19"
High-performance routers

- Geared to core and distribution service needs
  - Requirements: high speed & high density
- Why do we care?
  - Moore’s Law slower than link speed growth (and BW demand)
    - OC48c (2.5Gbps), common, 128ns/packet
    - OC192c (10Gbps), in backbone deployment, 33ns/packet
    - OC768c (40Gbps), emerging, 8ns/packet
  - Need high density/low power for POP deployments
    - Points-of-Presence (POP) – places where a network service provider provides dense connectivity
      - $20-100k & 2-400W per port, 50% ports frequently for internal connectivity (why?)
**Functional architecture**

Control Plane
- Complex
- Per-control action
- May be slow

Data plane
- Simple
- Per-packet
- Must be fast
Interconnect architecture

- Input & output connected via switch fabric
- Kinds of switch fabric
  - Bus
  - Crossbar
  - Shared Memory
- How to deal with transient contention?
  - Input queuing
  - Output queuing
  - Combination
First Generation Routers

- Single CPU and shared memory;
- All classification by main CPU
Second Generation Routers

- CPU
- Route Table
- Shared Bus(s)
- Direct DMA on cache hit
- Line Card
- Buffers
- Forwarding Cache
- MAC
- Cache of recent routes
Third Generation Routers

- Shared interconnect (frequently crossbar)
- Centralized scheduler
- Full forwarding table in line card
- Fixed cells
Output queuing

- Output interfaces buffer packets

- Pro
  - Simple algorithms
  - Single congestion point

- Con
  - N inputs may send to the same output
  - Requires speedup of N
Input queuing

- Input interfaces buffer packets
- Pro
  - Single congestion point
  - Simple to design algorithms
- Con
  - Must implement flow control
  - Low utilization due to Head-of-Line (HoL) Blocking
    » Utilization limited to \(2 - 2^{\frac{1}{2}} = 58\%\)
Head-of-Line Blocking
Virtual Output Queues
IQ + Virtual Output Queuing

- Input interfaces buffer packets in per-output virtual queues

- **Pro**
  - Solves blocking problem

- **Con**
  - More resources per port
  - Complex arbiter at switch
  - Still limited by input/output contention (scheduler)
  - RR: $1/e = 63\%$
Switch scheduling

- Problem
  - Match inputs and outputs
  - Resolve contentions, no packet drops
  - Maximize throughput
  - Do it in constant time…
- If traffic is uniformly distributed its easy
  - Lots of algorithms (approximate matching)
- Recent result (Dai et al, 2000)
  - Maximal size matching + speedup of two guarantees 100% utilization for most traffic assumptions
Modern high-performance router

- IQ + VoQ + OQ
  - Speedup of 2
  - Central scheduler
  - Fixed-sized internal cells

- Pro
  - Can achieve utilization of 1
  - Can scale to > Tb/s

- Con
  - Multiple congestion points
  - Complexity
Next bottlenecks

- Buffering at high speed
  - SRAM density too low for BW*D of 40Gbps link
  - DRAM too slow
  - SRAM memory management as cache for DRAM

- Scheduler overhead
  - Hard to do central scheduler much over 1Tbps
  - Multi-stage load-balanced switches

- High density (100’s-1000’s of line cards)
  - Physical distance to support density; electrical links degrade
  - Optical links; optical cross connect (MEMs,tunable lasers)

- Time to market, Power/Heat
Conclusion

- It is feasible to build very high speed IP routers
  - 40Gbps link speeds
  - Multi Tb/s aggregate capacity
- But…
  - Limited programmability
  - High complexity, slow time to market
    - Juniper I2 ASIC 2.5M gates
    - Typical OC192 LC ~30M gates!
    - Starting to require significant on-chip SRAM
  - Next gen (OC3072 160Gbps LC) may be close to cross-over point for CMOS (luckily, not clear there is demand)
For next time...

- Enjoy your Thanksgiving
- Next Tuesday we’ll cover Quality of Service issues