15-441 Computer Networking
Lecture 14: Router Design

Peter Steenkiste

Fall 2010
www.cs.cmu.edu/~prs/15-441-F10

Based on slides from Dave Andersen and Nick Feamster

Router Architecture

- Data Plane
  - Moving the data, i.e., the packets
  - How packets get forwarded
- Control Plane
  - How routing protocols establish routes/etc.

Today's Lecture: Data Plane

- The design of big, fast routers
- Partridge *et al.*, *A 50 Gb/s IP Router*
- Design constraints
  - Speed
  - Size
  - Power consumption
- Components
- Algorithms
  - Lookups and packet processing (classification, etc.)
  - Packet queuing
  - Switch arbitration

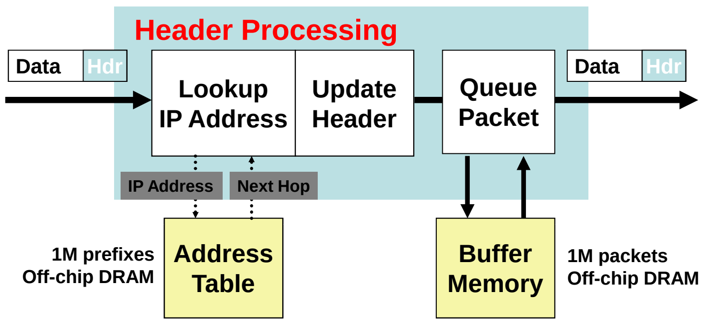
3

Summary of Routing Functionality

- Router gets packet
- Looks at packet header for destination
- Looks up routing table for output interface
- Modifies header Why?
- Passes packet to output interface

4

Generic Router Architecture



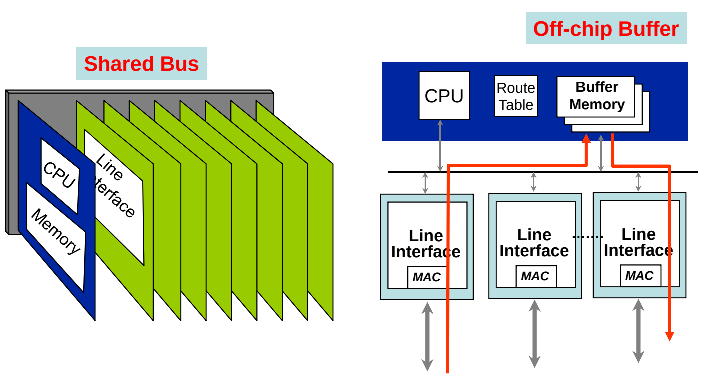
5

What's In A Router

- Interfaces
  - Input/output of packets
- Switching fabric
  - Moving packets from input to output
- Software
  - Routing
  - Packet processing
  - Scheduling
  - Etc.

6

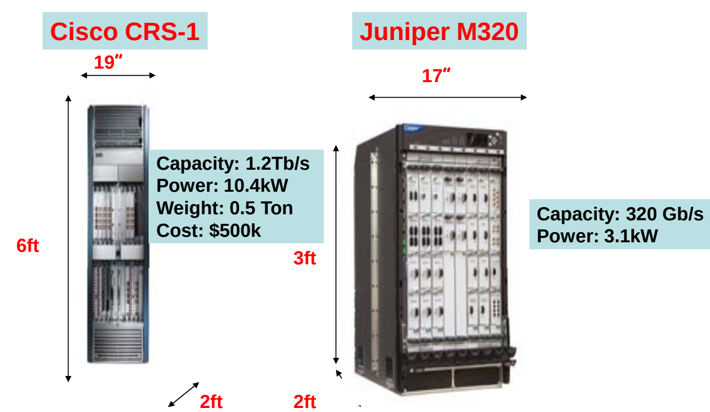
First Generation Routers



Typically <0.5Gb/s aggregate capacity

7

What a Router Chassis Looks Like

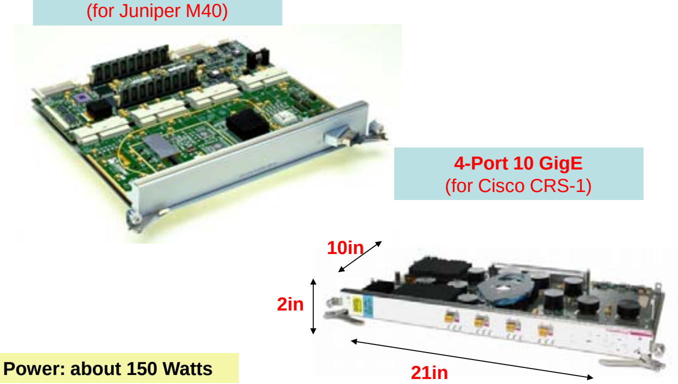


8

What a Router Line Card Looks Like

1-Port OC48 (2.5 Gb/s)
(for Juniper M40)

4-Port 10 GigE
(for Cisco CRS-1)



Power: about 150 Watts

9

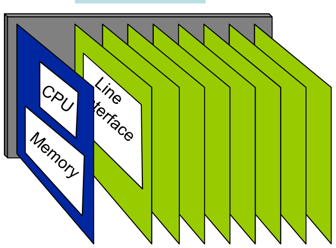
Big, Fast Routers: Why Bother?

- Faster link bandwidths
- Increasing demands
- Larger network size (hosts, routers, users)
- More cost effective

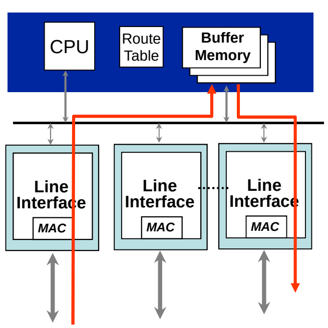
10

First Generation Routers

Shared Bus



Off-chip Buffer



Typically <0.5Gb/s aggregate capacity

11

Innovation #1: Each Line Card Has the Routing Tables

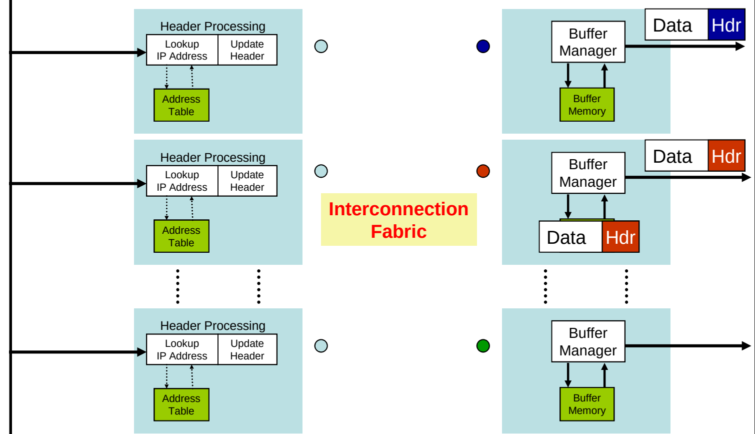
- Prevents central table from becoming a bottleneck at high speeds
- **Complication:** Must update forwarding tables on the fly.

12

Control Plane & Data Plane

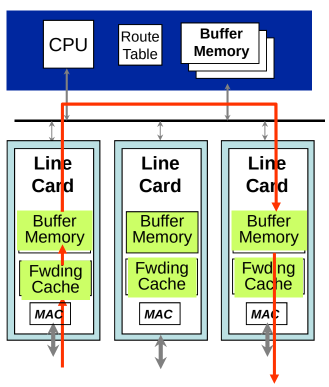
- Control plane must remember lots of routing info (BGP tables, etc.)
- Data plane only needs to know the “FIB” (Forwarding Information Base)
  - Smaller, less information, etc.
  - Simplifies line cards vs the network processor

Generic Router Architecture



14

Second Generation Routers



Typically <5Gb/s aggregate capacity

Bypasses memory bus with direct transfer over bus between line cards

Moves forwarding decisions local to card to reduce CPU pain

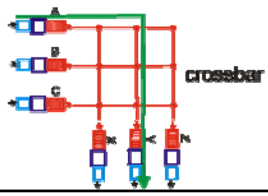
Punt to CPU for “slow” operations

Bus-based

- Some improvements possible
  - Cache bits of forwarding table in line cards, send directly over bus to outbound line card
- But shared bus was big bottleneck
  - E.g., *modern* PCI bus (PCIx16) is only 32Gbit/sec (in theory)
  - Almost-modern cisco (XR 12416) is 320Gbit/sec.
  - Ow! How do we get there?

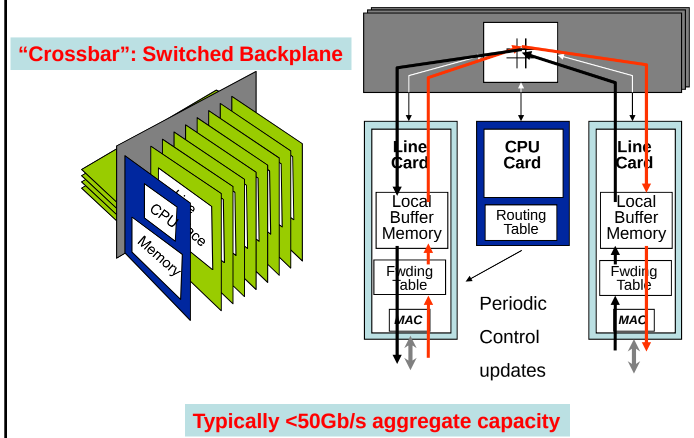
Innovation #2: Switched Backplane

- Every input port has a connection to every output port
- During each timeslot, each input connected to zero or one outputs
- **Advantage:** Exploits parallelism
- **Disadvantage:** Need scheduling algorithm



17

Third Generation Routers



What's so hard here?

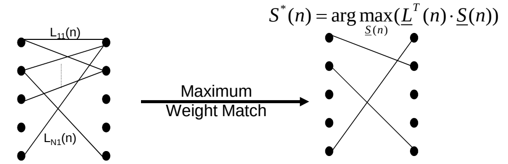
- Back-of-the-envelope numbers
  - Line cards can be 40 Gbit/sec today (OC-768)
    - Undoubtedly faster in a few more years, so scale these #s appropriately!
  - To handle minimum-sized packets (~40b)
    - 125 Mpps, or 8ns per packet
    - But note that this can be deeply pipelined, at the cost of buffering and complexity. Some lookup chips do this, though still with SRAM, not DRAM. Good lookup algos needed still.
- For every packet, you must:
  - Do a routing lookup (where to send it)
  - Schedule the crossbar
  - Maybe buffer, maybe QoS, maybe filtering by ACLs

Crossbar Switching

- **Conceptually:** *N* inputs, *N outputs*
  - Actually, inputs are also outputs
- In each timeslot, one-to-one mapping between inputs and outputs.
- Crossbar constraint: If input I is connected to output j, no other input connected to j, no other output connected to input I
- **Goal:** Maximal matching

Traffic Demands **Bipartite Match**

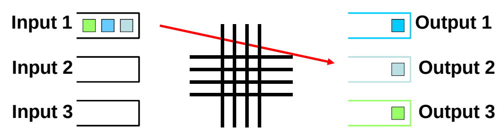
$$S^*(n) = \arg\max_{\underline{S}(n)}(\underline{L}^T(n) \cdot \underline{S}(n))$$



20

Head-of-Line Blocking

Problem: **The packet at the front of the queue experiences contention for the output queue, blocking all packets behind it.**



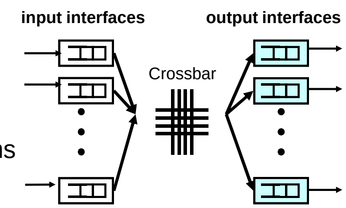
Maximum throughput in such a switch: 2 – sqrt(2)

M.J. Karol, M. G. Hluchyj, and S. P. Morgan, "Input Versus Output Queuing on a Space-Division Packet Switch," *IEEE Transactions On Communications*, Vol. Com-35, No. 12, December 1987, pp. 1347-1356.

21

Combined Input-Output Queuing

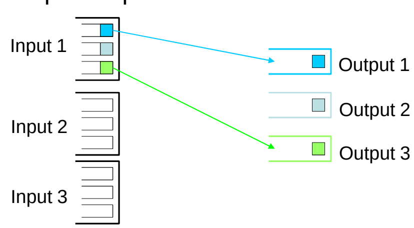
- **Advantages**
  - Easy to build
  - Better throughput
- **Disadvantages**
  - Harder to design algorithms
    - Two congestion points



22

Solution: Virtual Output Queues

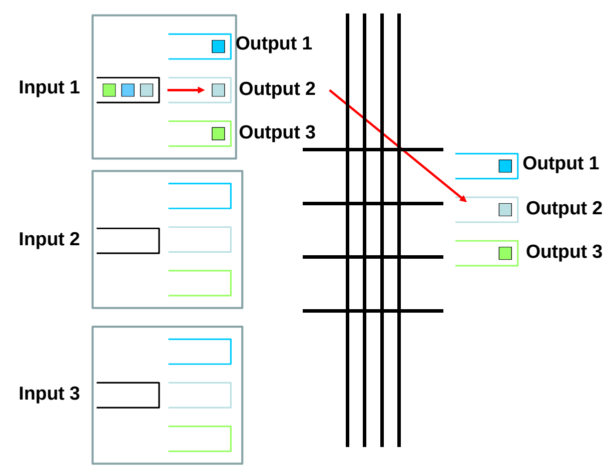
- Maintain N virtual queues at each input
  - one per output



N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand, "Achieving 100% Throughput in an Input-Queued Switch," IEEE Transactions on Communications, Vol. 47, No. 8, August 1999, pp. 1260-1267.

23

Virtual Output Queues



24

Quality of Service (QoS)

- Ensure that every network customer gets quality service and their fair share of the network
- Might need to reorder packages
  - Complicates router design
- More on this later…

25

Why QoS?

- Internet currently provides one single class of **“best-effort” service**
  - No assurances about delivery
- Existing applications are ***elastic***
  - Tolerate delays and losses
  - Can adapt to congestion
- Future “real-time” applications may be ***inelastic***

26

Router Components and Functions

- Route processor
  - Routing
  - Installing forwarding tables
  - Management
- Line cards
  - **Packet processing and classification**
  - Packet forwarding
- Switched bus (“Crossbar”)
  - Scheduling

27

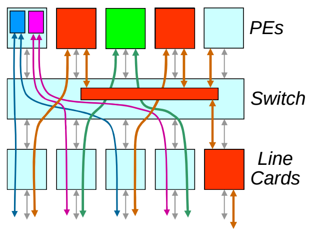
Processing: Fast Path vs. Slow Path

- **Optimize for common case**
  - BBN router: 85 instructions for fast-path code
  - Fits entirely in L1 cache
- Non-common cases handled on slow path
  - Route cache misses
  - Errors (*e.g.,* ICMP time exceeded)
  - IP options
  - Fragmented packets
  - Mullticast packets

28

Recent Trends: Programmability

- NetFPGA: 4-port interface card, plugs into PCI bus (Stanford)
  - Customizable forwarding
  - Appearance of many virtual interfaces (with VLAN tags)
- Programmability with Network processors (Washington U.)



29

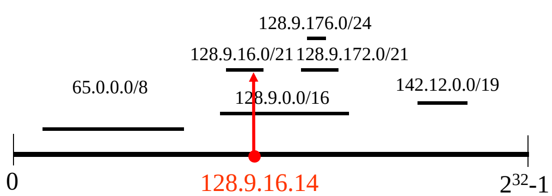
IP Address Lookup Challenges

Challenges:

1. Longest-prefix match (not exact).
2. Tables are large and growing.
3. Lookups must be fast.

30

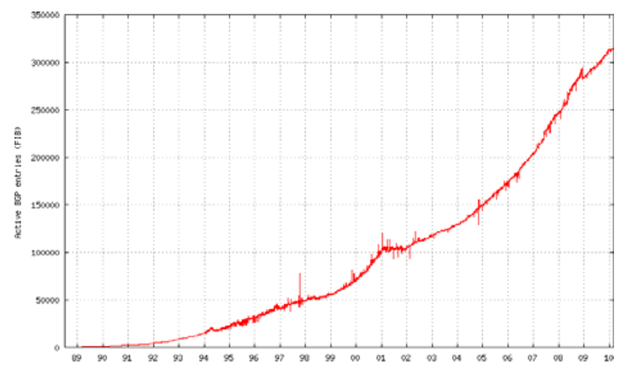
IP Lookups find Longest Prefixes



Routing lookup: Find the longest matching prefix (aka the most specific route) among all prefixes that match the destination address.

31

Address Tables are Large



32

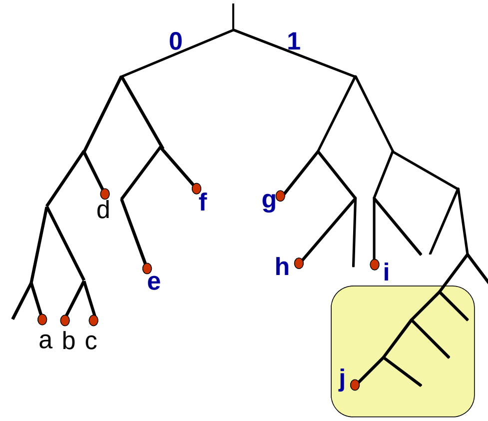
Lookups Must be Fast

Cisco CRS-1 1-Port OC-768C
(Line rate: 42.1 Gb/s)

Year	Line	40B packets (Mpkt/s)	
1997	622Mb/s	1.94	OC-12
1999	2.5Gb/s	7.81	OC-48
2001	10Gb/s	31.25	OC-192
2003	40Gb/s	125	OC-768

33

IP Address Lookup: Binary Tries

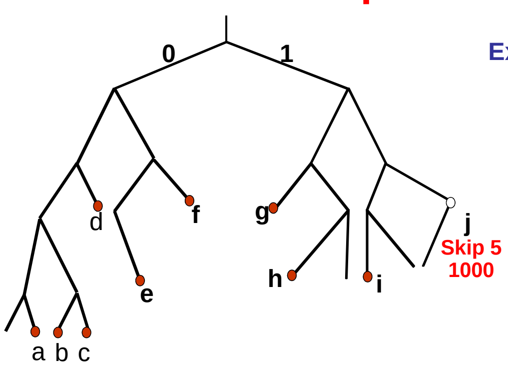


Example Prefixes:

a) 00001
b) 00010
c) 00011
d) 001
e) 0101
f) 011
g) 100
h) 1010
i) 1100
j) 11110000

34

IP Address Lookup: Patricia Trie



Example Prefixes

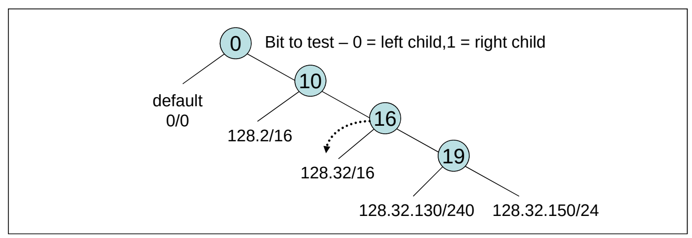
a) 00001
b) 00010
c) 00011
d) 001
e) 0101
f) 011
g) 100
h) 1010
i) 1100
j) 11110000

Problem: Lots of (slow) memory lookups

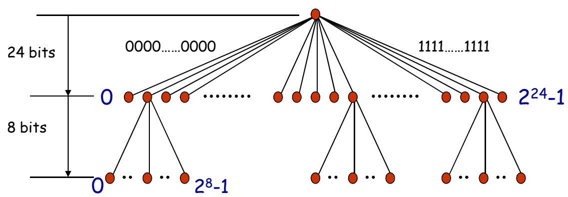
35

LPM with PATRICIA Tries

- Traditional method – Patricia Tree
  - Arrange route entries into a series of bit tests
- Worst case = 32 bit tests
  - Problem: memory speed, even w/SRAM!



Address Lookup: Direct Trie



- When pipelined, one lookup per memory access
- **Inefficient use of memory**

37

Faster LPM: Alternatives

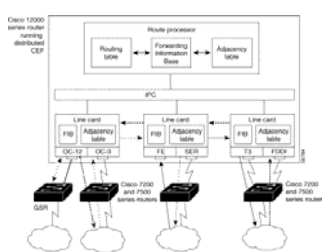
- Content addressable memory (CAM)
  - **Hardware-based** route lookup
  - Input = tag, output = value
  - Requires exact match with tag
    - Multiple cycles (1 per prefix) with single CAM
    - Multiple CAMs (1 per prefix) searched in parallel
  - Ternary CAM
    - (0,1,don't care) values in tag match
    - Priority (*i.e.*, longest prefix) by order of entries

Historically, this approach has not been very economical.

38

Faster Lookup: Alternatives

- Caching
  - Packet trains exhibit temporal locality
  - Many packets to same destination
- Cisco Express Forwarding



39

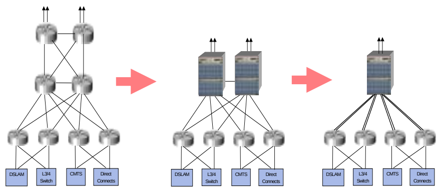
IP Address Lookup: Summary

- Lookup limited by memory bandwidth.
- Lookup uses high-degree trie.
- State of the art: 10Gb/s line rate.
- Scales to: 40Gb/s line rate.

40

Fourth-Generation: Collapse the POP

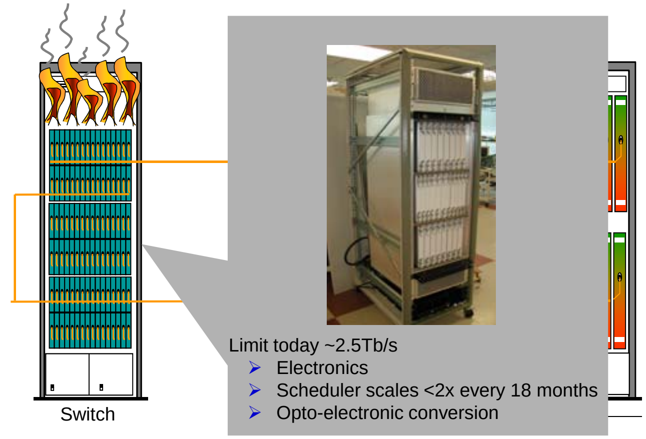
High Reliability and Scalability enable “vertical” POP simplification



Reduces CapEx, Operational cost
Increases network stability

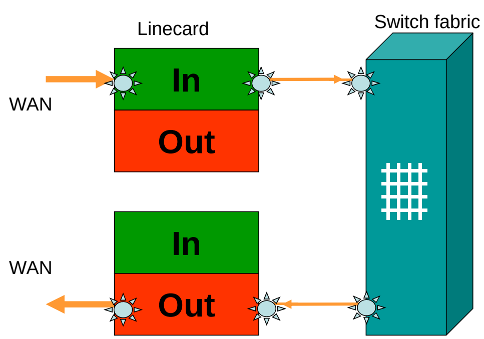
41

Fourth-Generation Routers



42

Multi-rack routers



43

Router Design

- Many trade-offs: power, $$$, throughput, reliability, flexibility
- Move towards distributed architectures
  - Line-cards have forwarding tables
  - Switched fabric between cards
  - Separate Network processor for “slow path” & control
- Important bottlenecks on fast path
  - Longest prefix match
  - Cross-bar scheduling
- Beware: lots of feature creep

44