15-744: Computer Networking

L-8 Routers

Forwarding and Routers

- Forwarding
- IP lookup
- High-speed router architecture
- Readings
  - [McK97] A Fast Switched Backplane for a Gigabit Switched Router
  - [KCY03] Scaling Internet Routers Using Optics
  - Know RIP/OSPF
- Optional
  - [D+97] Small Forwarding Tables for Fast Routing Lookups
  - [BV01] Scalable Packet Classification

Outline

- IP router design
- IP route lookup
- Variable prefix match algorithms
- Packet classification

IP Router Design

- Different architectures for different types of routers
- High speed routers incorporate large number of processors
- Common case is optimized carefully

What Does a Router Look Like?

- Currently:
  - Network controller
  - Line cards
  - Switched backplane
- In the past?
  - Workstation
  - Multiprocessor workstation
  - Line cards + shared bus

5

Line Cards

- Network interface cards
- Provides parallel processing of packets
- Fast path per-packet processing
  - Forwarding lookup (hardware/ASIC vs. software)

6

Network Processor

- Runs routing protocol and downloads forwarding table to line cards
  - Some line cards maintain two forwarding tables to allow easy switchover
- Performs “slow” path processing
  - Handles ICMP error messages
  - Handles IP option processing

7

Switch Design Issues

- Have N inputs and M outputs
  - Multiple packets for same output – output contention
  - Switch contention – switch cannot support arbitrary set of transfers
    - Crossbar
    - Bus
      - High clock/transfer rate needed for bus
    - Banyan net
      - Complex scheduling needed to avoid switch contention
- Solution – buffer packets where needed

8

Switch Buffering

- Input buffering
  - Which inputs are processed each slot – schedule?
  - Head of line packets destined for busy output blocks other packets
- Output buffering
  - Output may receive multiple packets per slot
  - Need speedup proportional to # inputs
- Internal buffering
  - Head of line blocking
  - Amount of buffering needed

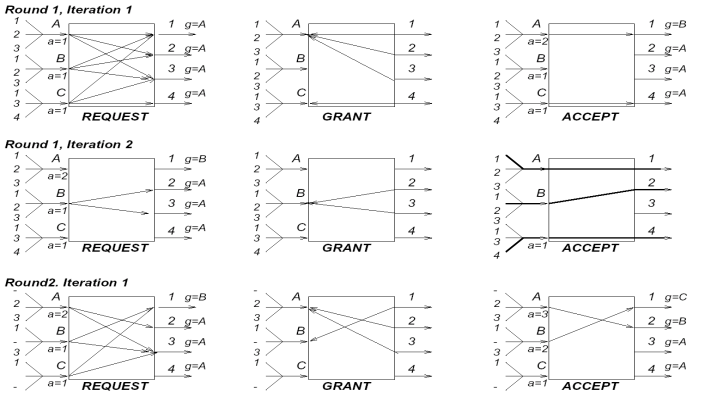
9

Line Card Interconnect

- Virtual output buffering
  - Maintain per output buffer at input
  - Solves head of line blocking problem
  - Each of MxN input buffer places bid for output
- Crossbar connect
- Challenge: map of bids to schedule for crossbar

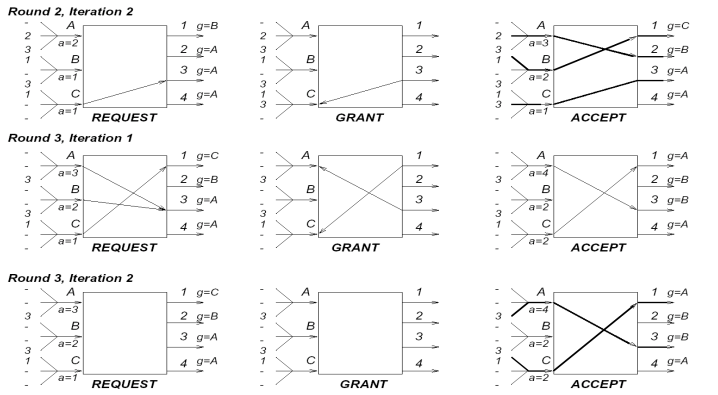
10

ISLIP



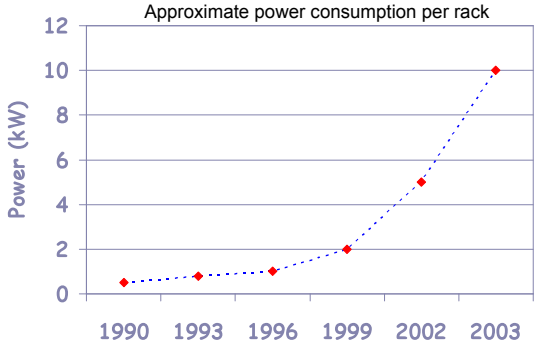
11

ISLIP (cont.)



12

What Limits Router Capacity?

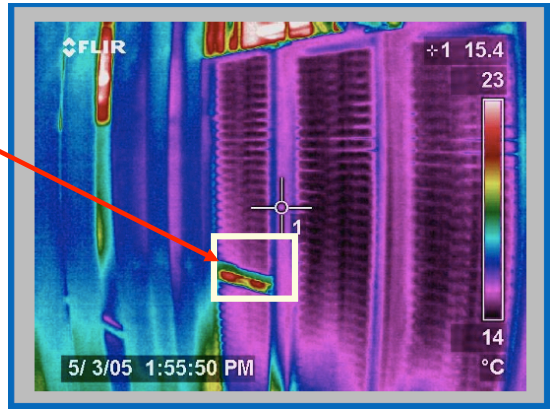


Power density is the limiting factor today

13

Thermal Image of Typical Cluster

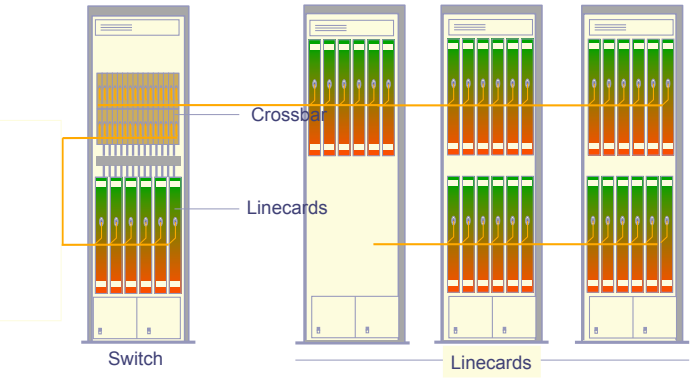
Rack
Switch



M. K. Patterson, A. Pratt, P. Kumar,
"From UPS to Silicon: an end-to-end evaluation of datacenter efficiency", Intel Corporation

14

Multi-rack Routers Reduce Power Density



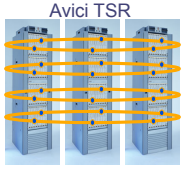
15

Examples of Multi-rack Routers

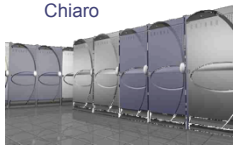
Alcatel 7670 RSP

Juniper TX8/T640

Avici TSR



Chiaro



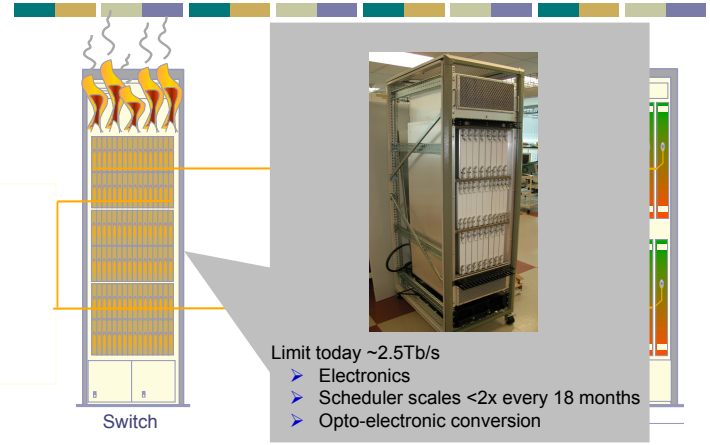
16

Limits to Scaling

- Overall power is dominated by linecards
  - Sheer number
  - Optical WAN components
  - Per packet processing and buffering.
- But power *density* is dominated by switch fabric

17

Multi-rack Routers Reduce Power Density



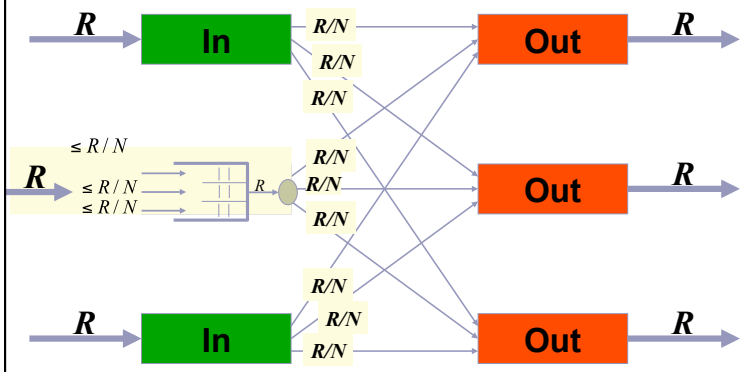
18

Question

- Instead, can we use an **optical** fabric at 100Tb/s with 100% throughput?
- Conventional answer: **No**
  - Need to reconfigure switch too often
  - 100% throughput requires complex electronic scheduler.

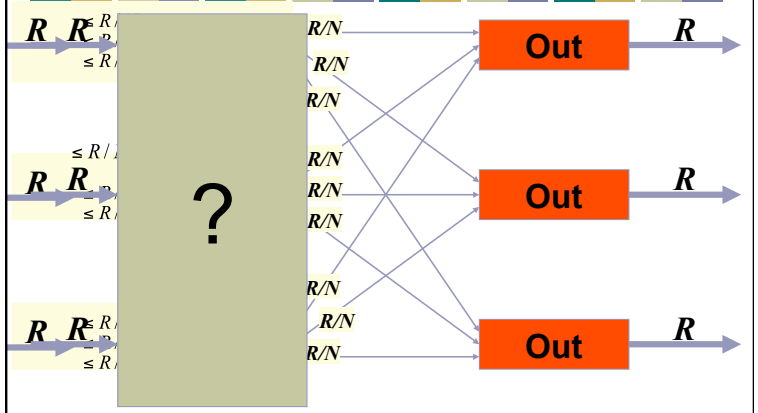
19

If Traffic is Uniform…



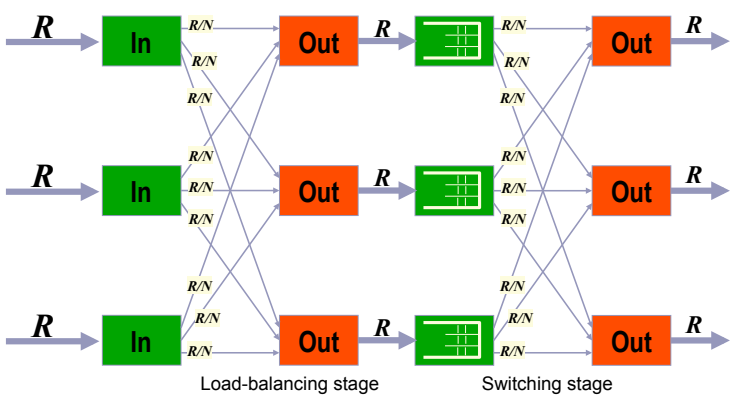
20

Real Traffic is Not Uniform



21

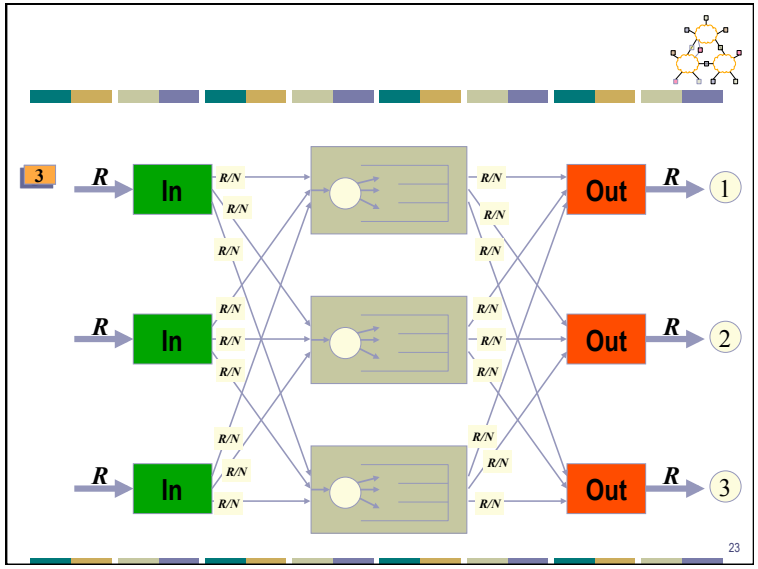
Two-stage Load-Balancing Switch



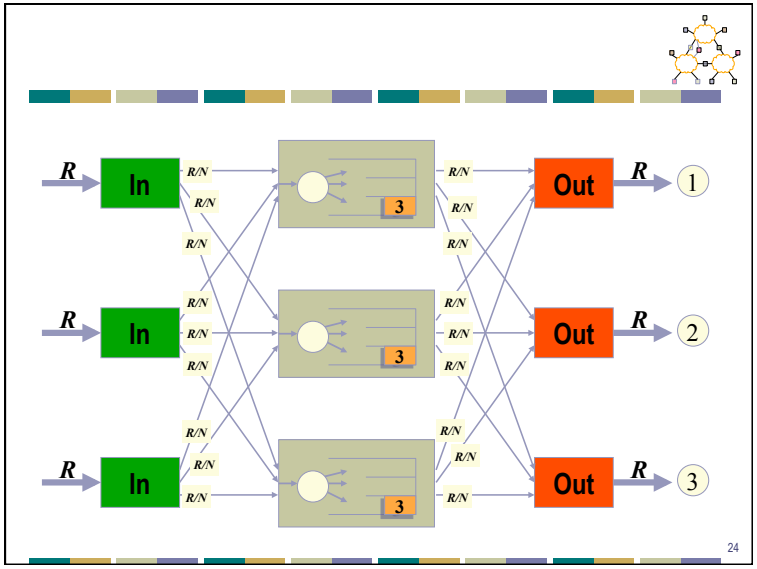
100% throughput for weakly mixing, stochastic traffic

[C.-S. Chang, Valiant]

22

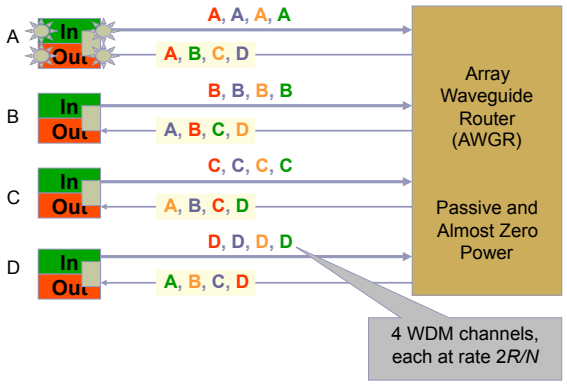


23



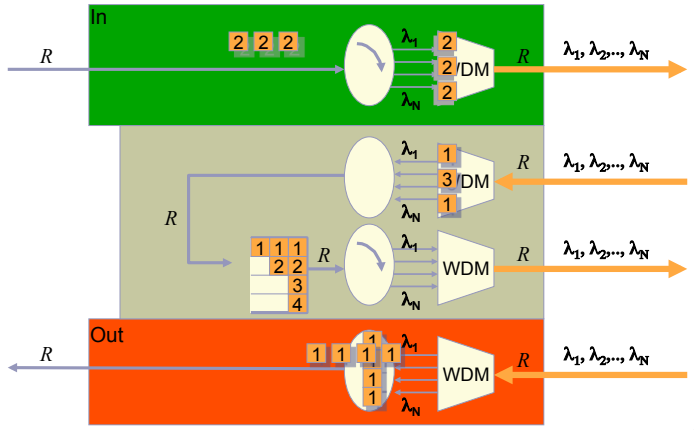
24

Static WDM Switching



25

Linecard Dataflow



26

Outline

- IP router design
- IP route lookup
- Variable prefix match algorithms
- Packet classification

27

Original IP Route Lookup

- Address classes
  - A: 0 | 7 bit network | 24 bit host (16M each)
  - B: 10 | 14 bit network | 16 bit host (64K)
  - C: 110 | 21 bit network | 8 bit host (255)
- Address would specify prefix for forwarding table
  - Simple lookup

28

Original IP Route Lookup – Example

- www.cmu.edu address 128.2.11.43
  - Class B address – class + network is 128.2
  - Lookup 128.2 in forwarding table
  - Prefix – part of address that really matters for routing
- Forwarding table contains
  - List of class+network entries
  - A few fixed prefix lengths (8/16/24)
- Large tables
  - 2 Million class C networks
- 32 bits does not give enough space encode network location information inside address – i.e., create a structured hierarchy

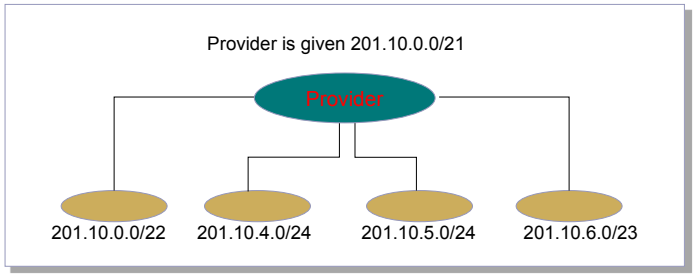
29

CIDR Revisited

- Supernets
  - Assign adjacent net addresses to same org
  - Classless routing (CIDR)
- How does this help routing table?
  - Combine routing table entries whenever all nodes with same prefix share same hop
  - Routing protocols carry prefix with destination network address
  - Longest prefix match for forwarding

30

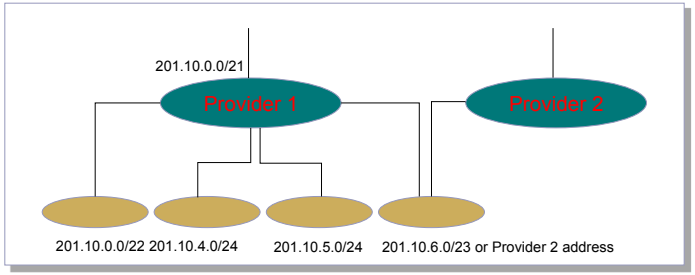
CIDR Illustration



31

CIDR Shortcomings

- Multi-homing
- Customer selecting a new provider



32

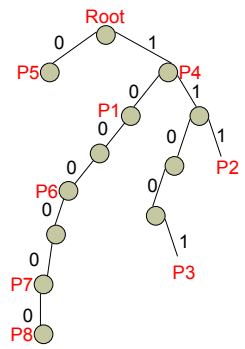
Outline

- IP router design
- IP route lookup
- Variable prefix match algorithms
- Packet classification

33

Trie Using Sample Database

Trie



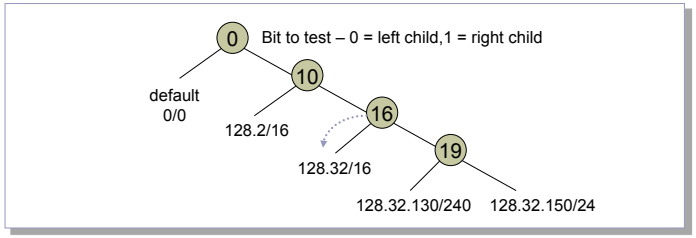
Sample Database

- P1 = 10*
- P2 = 111*
- P3 = 11001*
- P4 = 1*
- P5 = 0*
- P6 = 1000*
- P7 = 100000*
- P8 = 1000000*

34

How To Do Variable Prefix Match

- Traditional method – Patricia Tree
  - Arrange route entries into a series of bit tests
- Worst case = 32 bit tests
  - Problem: memory speed is a bottleneck



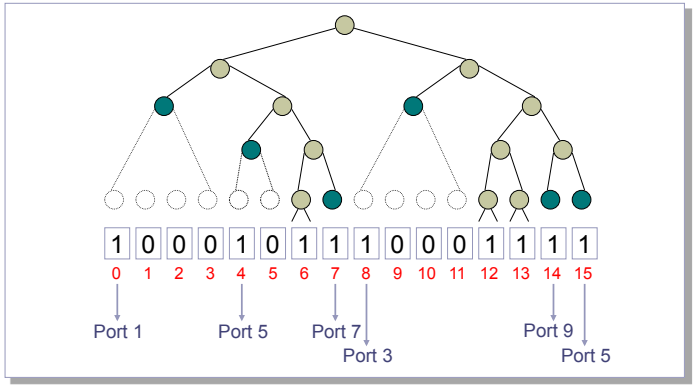
35

Speeding up Prefix Match (P+98)

- Cut prefix tree at 16 bit depth
  - 64K bit mask
  - Bit = 1 if tree continues below cut (root head)
  - Bit = 1 if leaf at depth 16 or less (genuine head)
  - Bit = 0 if part of range covered by leaf

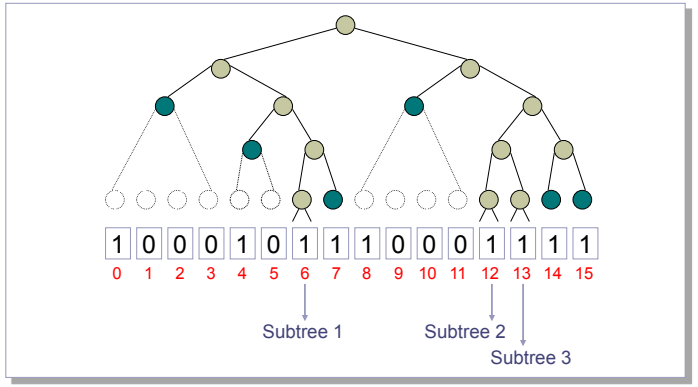
36

Prefix Tree



37

Prefix Tree



38

Speeding up Prefix Match (P+98)

- Each 1 corresponds to either a route or a subtree
  - Keep array of routes/pointers to subtree
  - Need index into array – how to count # of 1s
  - Keep running count to 16bit word in base index + code word (6 bits)
  - Need to count 1s in last 16bit word
    - Clever tricks
- Subtrees are handled separately

39

Speeding up Prefix Match (P+98)

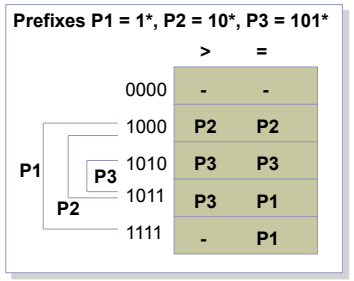
- Scaling issues
  - How would it handle IPv6
- Update issues
- Other possibilities
  - Why were the cuts done at 16/24/32 bits?
  - Improve data structure by shuffling bits

40

Speeding up Prefix Match - Alternatives

- Route caches
  - Temporal locality
  - Many packets to same destination
- Other algorithms
  - Waldvogel – Sigcomm 97
    - Binary search on prefixes
    - Works well for larger addresses
  - Bremler-Barr – Sigcomm 99
    - Clue = prefix length matched at previous hop
    - Why is this useful?
  - Lampson – Infocom 98
    - Binary search on ranges

Binary Search on Ranges



- Encode each prefix as range and place all range endpoints in binary search table or tree. Need two next hops per entry for > and = case. [Lampson, Srinivasan, Varghese]

- Problem: Slow search ($\log_2$ N+1 = 20 for a million prefixes) and update (O(n)).
  - Some clever implementation tricks to improve on this

Speeding up Prefix Match - Alternatives

- Content addressable memory (CAM)
  - Hardware based route lookup
  - Input = tag, output = value associated with tag
  - Requires exact match with tag
    - Multiple cycles (1 per prefix searched) 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 in CAM

Outline

- IP router design
- IP route lookup
- Variable prefix match algorithms
- Packet classification

Packet Classification

- Typical uses
  - Identify flows for QoS
  - Firewall filtering
- Requirements
  - Match on multiple fields
  - Strict priority among rules
    - E.g 1. no traffic from 128.2.*
      2. ok traffic on port 80

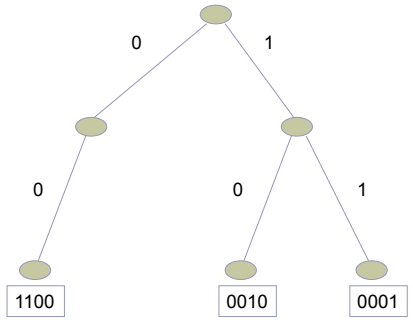
45

Complexity

- N rules and k header fields for k > 2
  - O(log N^{k-1}) time and O(N) space
  - O(log N) time and O(N^k) space
  - Special cases for k = 2 → source and destination
    - O(log N) time and O(N) space solutions exist
- How many rules?
  - Largest for firewalls & similar → 1700
  - Diffserv/QoS → much larger → 100k (?)

46

Bit Vectors

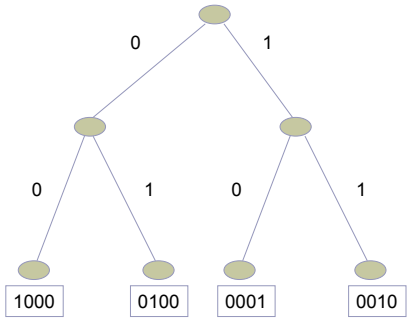


Rule	Field1	Field2
0	00*	00*
1	00*	01*
2	10*	11*
3	11*	10*

Field 1

47

Bit Vectors



Rule	Field1	Field2
0	00*	00*
1	00*	01*
2	10*	11*
3	11*	10*

Field 2

48

Observations [GM99]

- Common rule sets have important/useful characteristics
  - Packets rarely match more than a few rules (rule intersection)
    - E.g., max of 4 rules seen on common databases up to 1700 rules

49

Aggregating Rules [BV01]

- Common case: very few 1's in bit vector → aggregate bits
- OR together A bits at a time → N/A bit-long vector
  - A typically chosen to match word-size
  - Can be done hierarchically → aggregate the aggregates
- AND of aggregate bits indicates which groups of A rules have a possible match
  - Hopefully only a few 1's in AND'ed vector
  - AND of aggregated bit vectors may have false positives
- Fetch and AND just bit vectors associated with positive entries

50

Rearranging Rules [BV01]

- Problem: false positives may be common
- Solution: reorder rules to minimize false positives
  - What about the priority order of rules?
- How to rearrange?
  - Heuristic → sort rules based on single field's values
    - First sort by prefix length then by value
    - Moves similar rules close together → reduces false positives

51

Summary: Addressing/Classification

- Router architecture carefully optimized for IP forwarding
- Key challenges:
  - Speed of forwarding lookup/classification
  - Power consumption
- Some good examples of common case optimization
  - Routing with a clue
  - Classification with few matching rules
  - Not checksumming packets

52

Open Questions

- Fanout vs. bandwidth
- MPLS vs. longest prefix match
- More vs. less functionality in routers
- Hardware vs. software
  - CAMs vs. software
- Impact of router design on network design

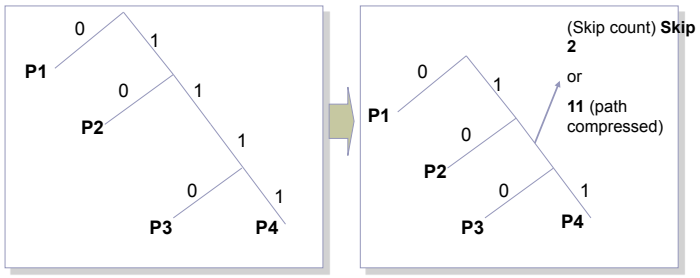
53

Summary

	Source Routing	Global Addresses	Virtual Circuits
Header Size	Worst	OK – Large address	Best
Router Table Size	None	Number of hosts (prefixes)	Number of circuits
Forward Overhead	Best	Prefix matching	Pretty Good
Setup Overhead	None	None	Connection Setup
Error Recovery	Tell all hosts	Tell all routers	Tell all routers and Tear down circuit and re-route

54

Skip Count vs. Path Compression



- Removing one way branches ensures # of trie nodes is at most twice # of prefixes
- Using a skip count requires exact match at end and backtracking on failure → path compression simpler

55