Master Course
Computer Networks
IN2097

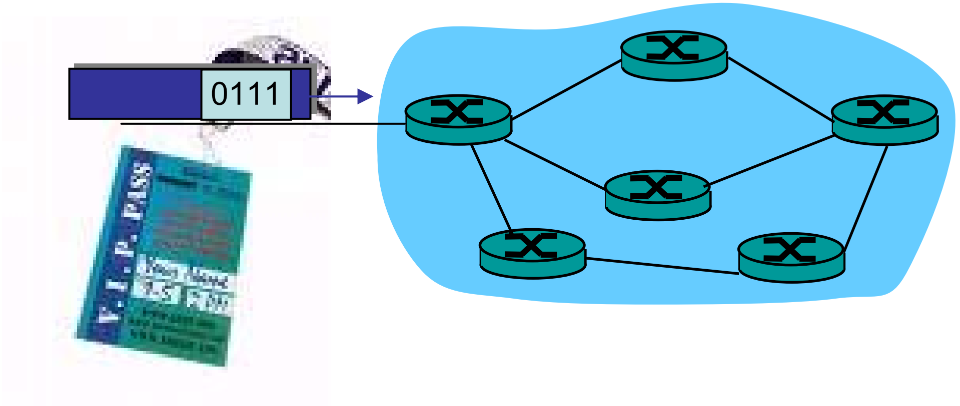
Prof. Dr.-Ing. Georg Carle
Christian Grothoff, Ph.D.

Chair for Network Architectures and Services
Institut für Informatik
Technische Universität München
http://www.net.in.tum.de

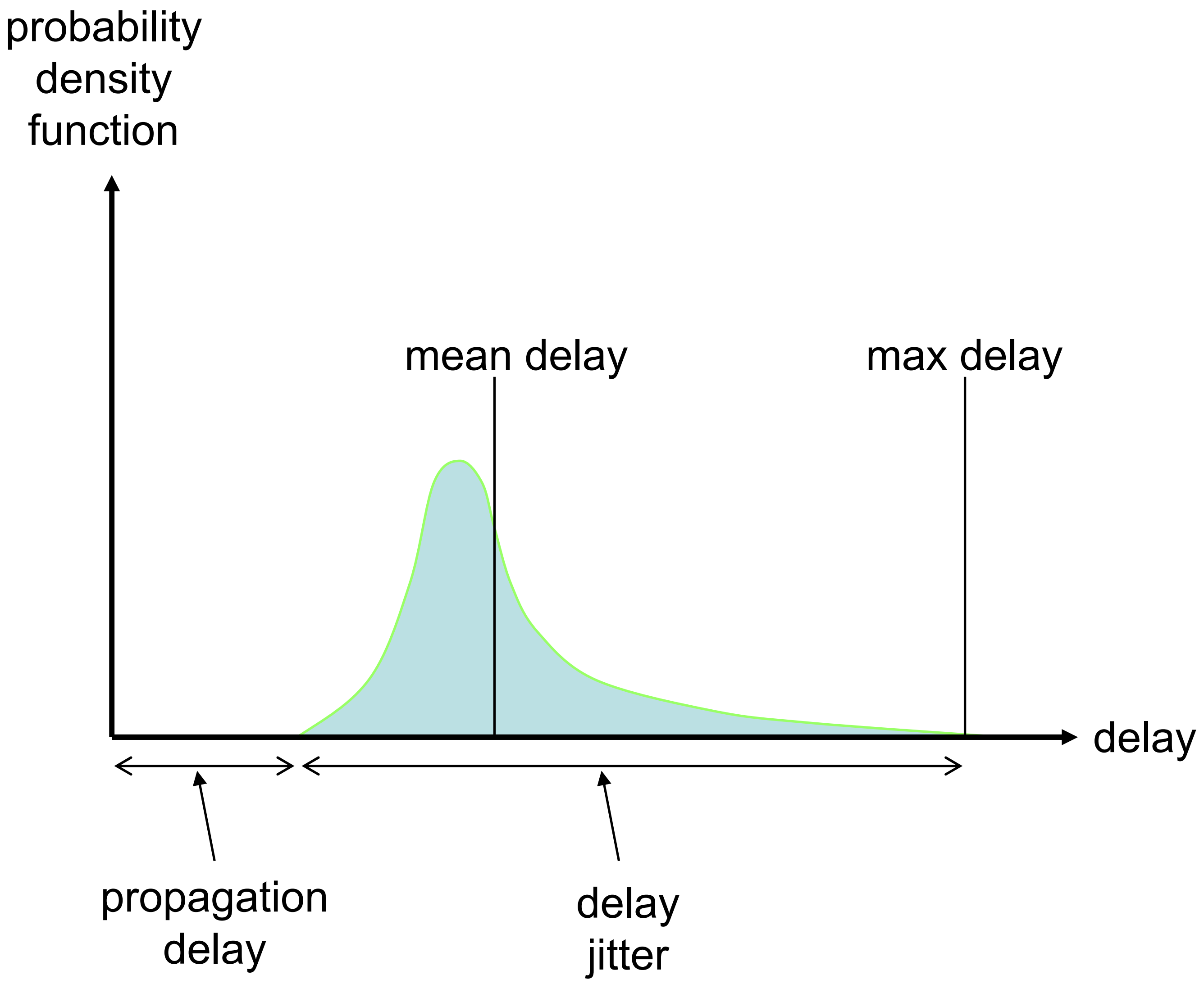
Providing multiple Classes of Service

Providing Multiple Classes of Service

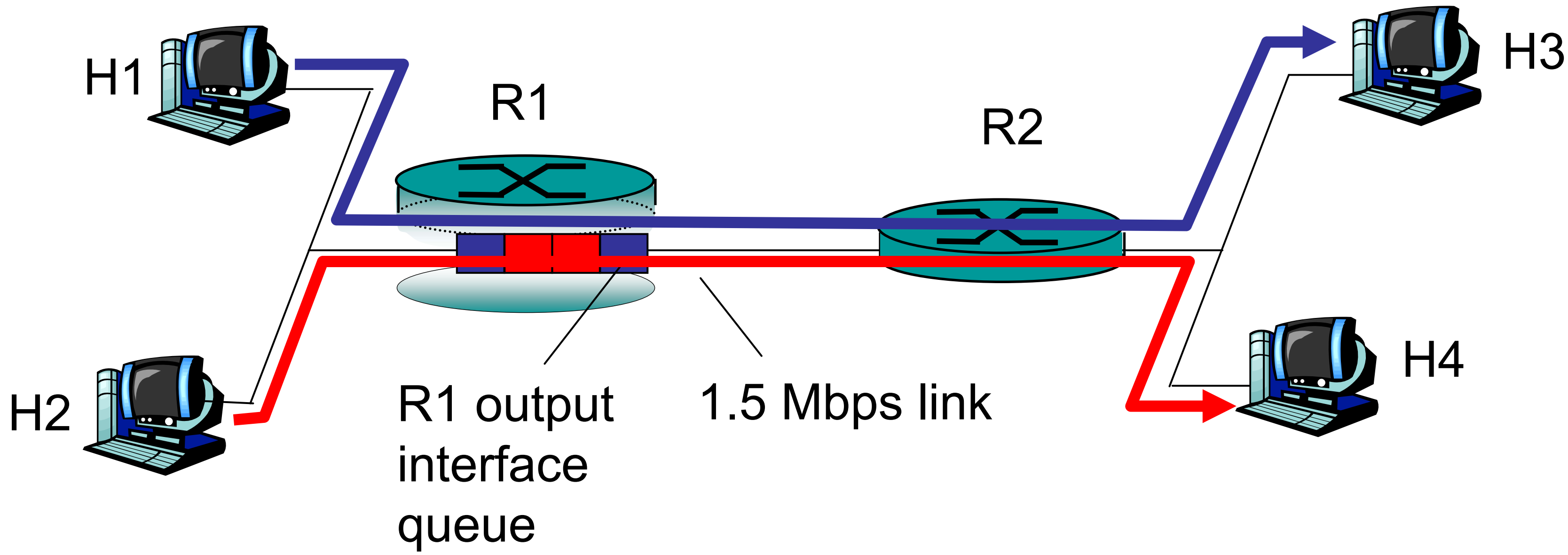
- Traditional Internet approach: making the best of best effort service
  - one-size fits all service model
- Alternative approach: multiple classes of service
  - partition traffic into classes
  - network treats different classes of traffic differently (analogy: VIP service vs regular service)
- granularity: differential service among multiple classes, not among individual connections
- history: ToS bits in IP header



Delay Distributions

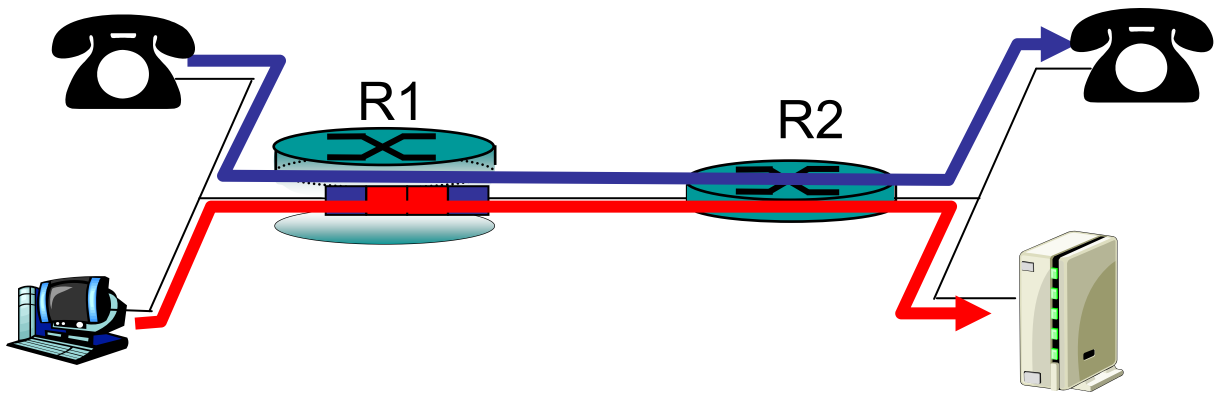


Multiple classes of service: scenario



Scenario 1: mixed FTP and audio

- Example: 1Mbps IP phone, FTP or NFS share 1.5 Mbps link.
  - bursts of FTP or NFS can congest router, cause audio loss
  - want to give priority to audio over FTP

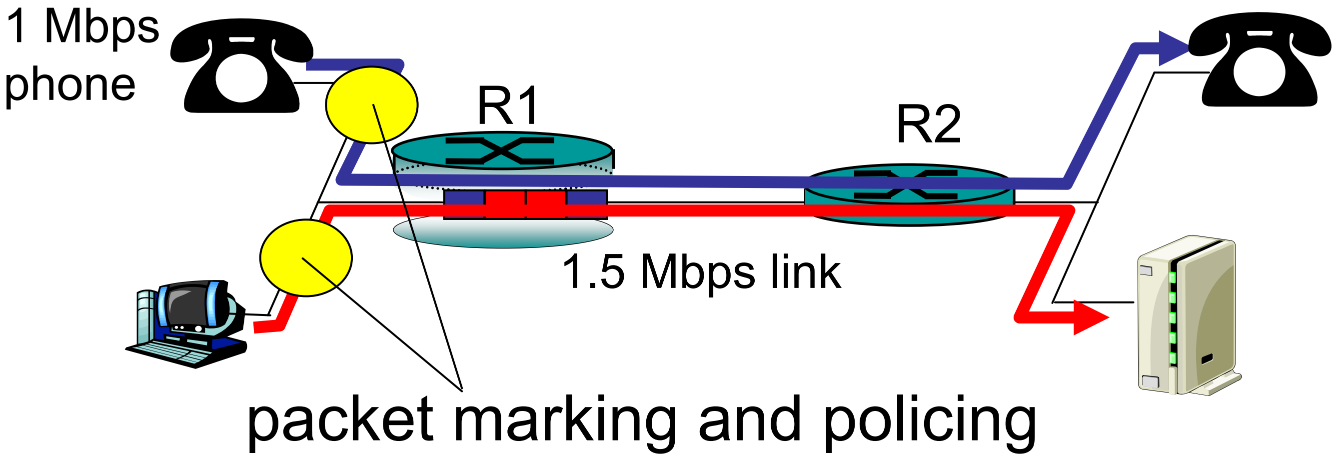


Principle 1

packet marking needed for router to distinguish between different classes; and new router policy to treat packets accordingly

Principles for QOS Guarantees (more)

- what if applications misbehave (audio sends higher than declared rate)
  - policing: force source adherence to bandwidth allocations
- marking and policing at network edge:
  - similar to ATM UNI (User Network Interface)

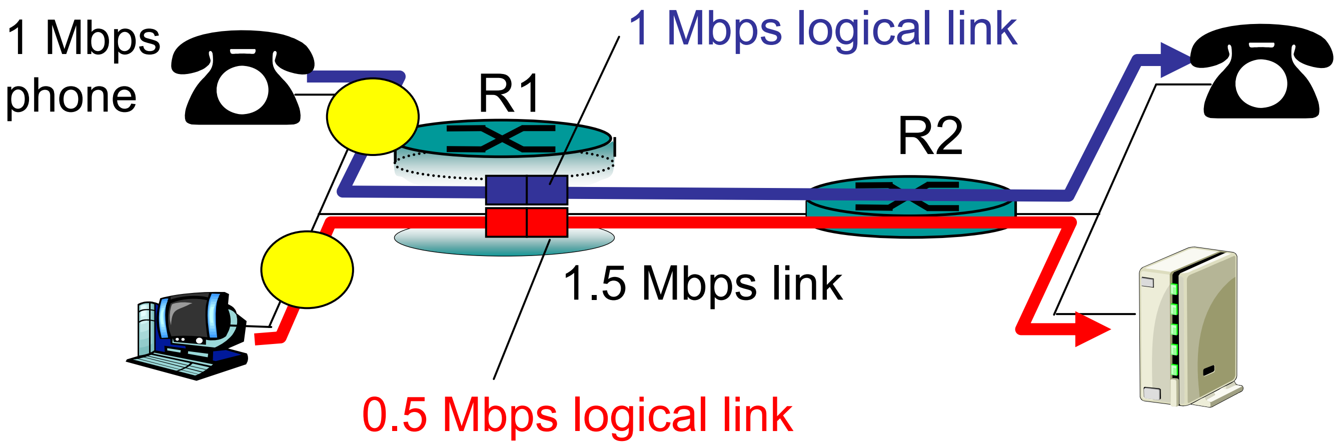


Principle 2

provide protection (*isolation*) for one class from others

Principles for QOS Guarantees (more)

- Allocating *fixed* (non-sharable) bandwidth to flow: *inefficient* use of bandwidth if flows doesn't use its allocation

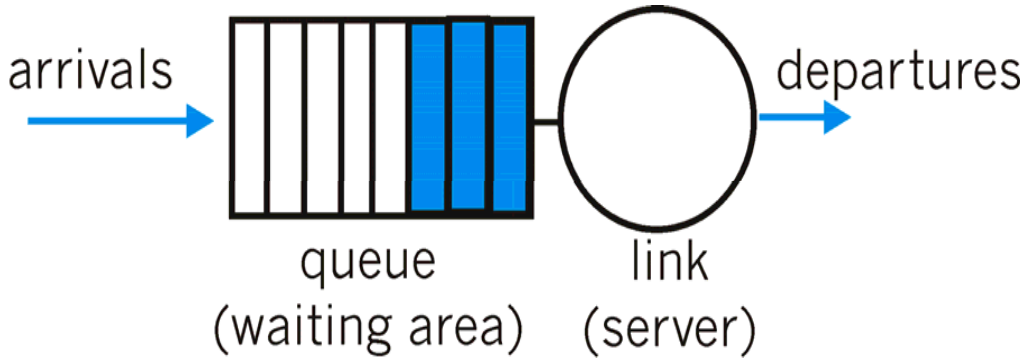


Principle 3

While providing **isolation**, it is desirable to use resources as efficiently as possible

Scheduling And Policing Mechanisms

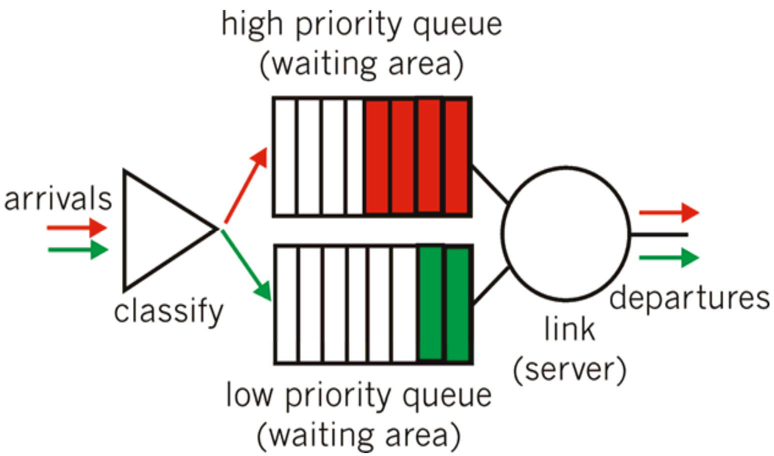
- scheduling: choose next packet to send on link
- FIFO (first in first out) scheduling: send in order of arrival to queue
  - ⇨real-world example?
  - discard policy: if packet arrives to full queue: who to discard?
    - Tail drop: drop arriving packet
    - priority: drop/remove on priority basis
    - random: drop/remove randomly

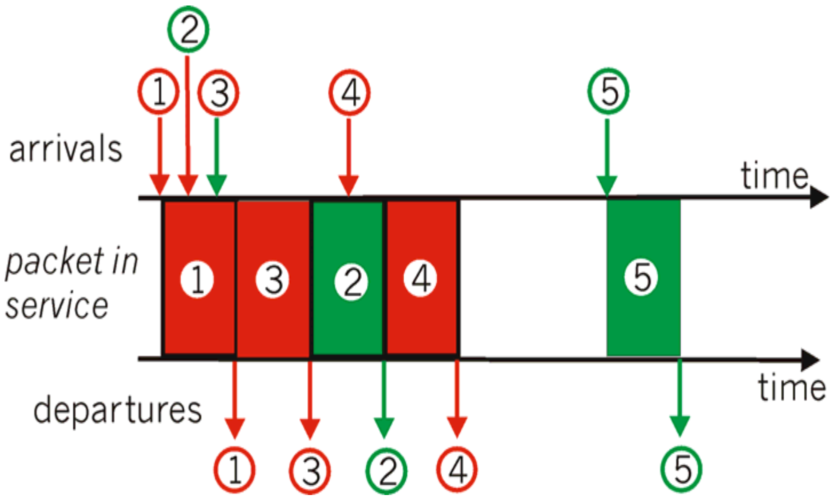


Scheduling Policies: more

Priority scheduling: transmit highest priority queued packet

- multiple *classes*, with different priorities
  - class may depend on marking or other header info, e.g. IP source/dest, port numbers, etc..

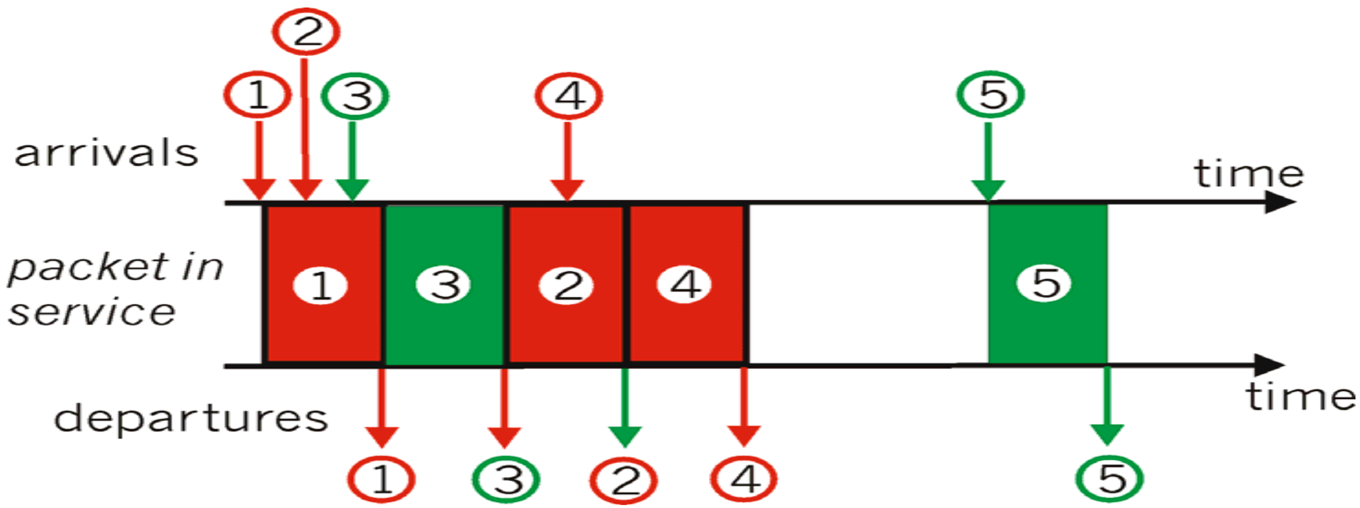




Scheduling Policies: still more

round robin scheduling:

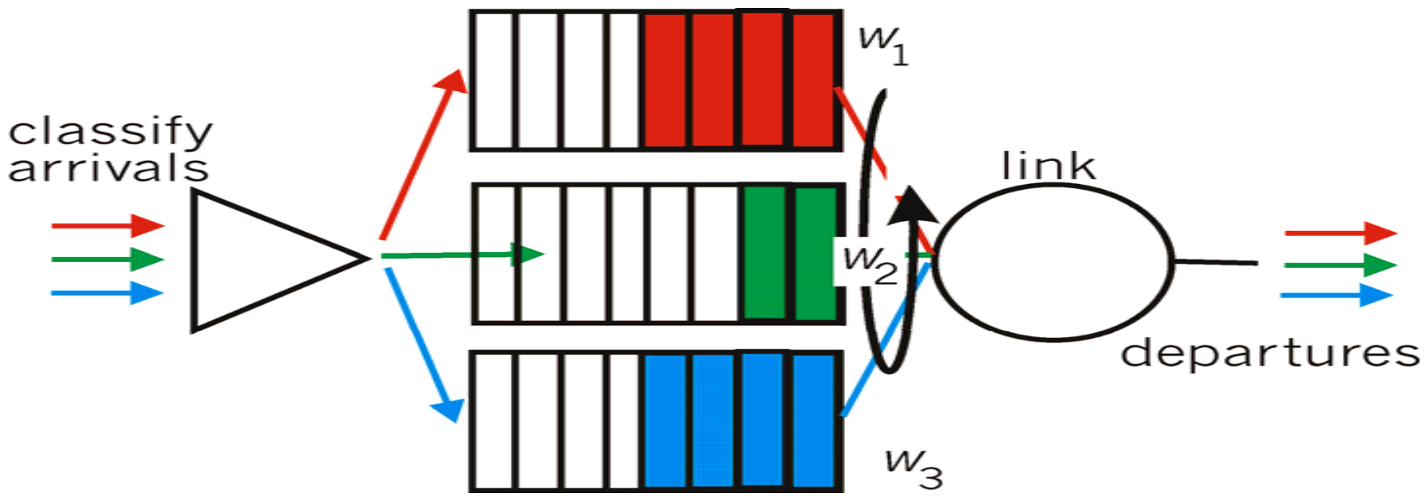
- multiple classes
- cyclically scan class queues, serving one from each class (if available)



Scheduling Policies: still more

Weighted Fair Queuing:

- generalized Round Robin
- each class gets weighted amount of service in each cycle
- when all classes have queued packets, class i will receive a bandwidht ratio of $w_i/\Sigma w_j$



Policing Mechanisms

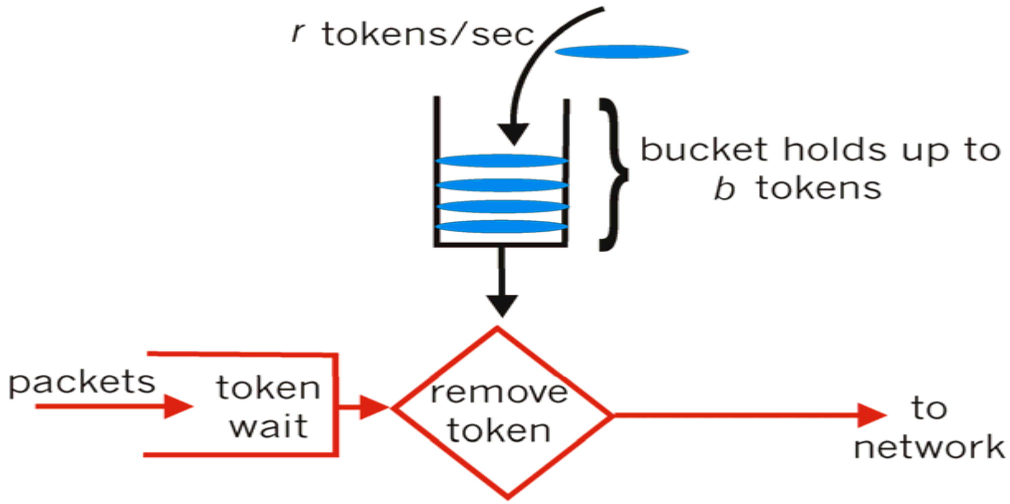
Goal: limit traffic to not exceed declared parameters

Three common-used criteria:

- *(Long term) Average Rate:* how many packets can be sent per unit time (in the long run)
  - crucial question: what is the interval length:
    100 packets per sec
    or 6000 packets per min have same average!
- *Peak Rate:* e.g., 6000 packets per min. (ppm) avg.;
  1500 pps peak rate
- *(Max.) Burst Size:* max. number of packets sent consecutively

Policing Mechanisms

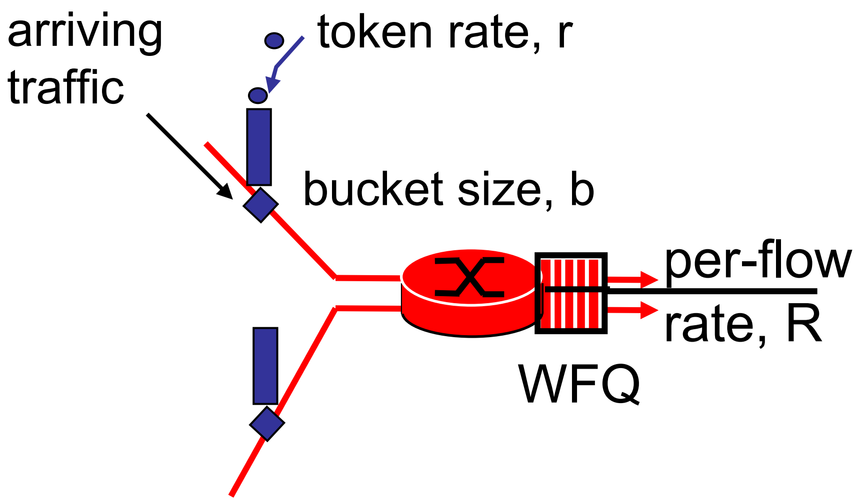
Token Bucket: limit input to specified Burst Size and Average Rate.



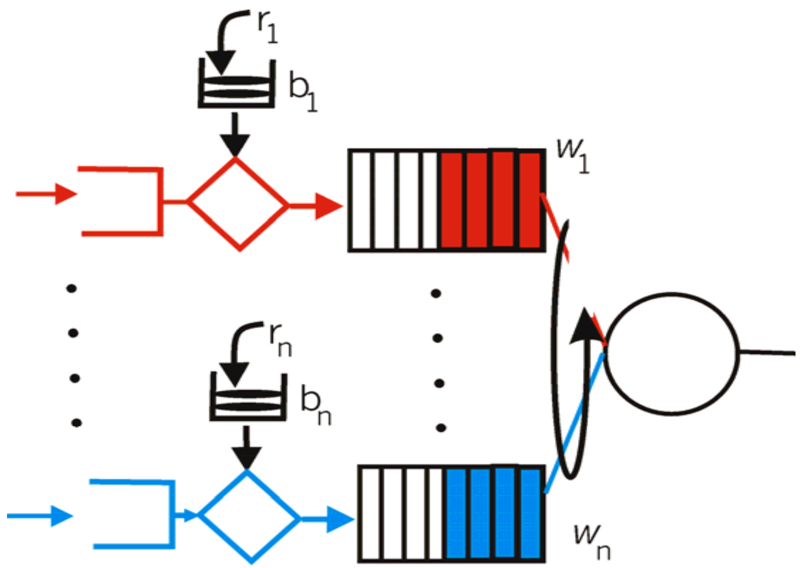
- bucket can hold b tokens $\Rightarrow$ limits maximum burst size
- tokens generated at rate *r token/sec* unless bucket full
- *over interval of length t: number of packets admitted less than or equal to (r t + b).*

Policing Mechanisms (more)

- token bucket, WFQ combined provide guaranteed upper bound on delay, i.e., *QoS guarantee*



$D_{max} = b/R$



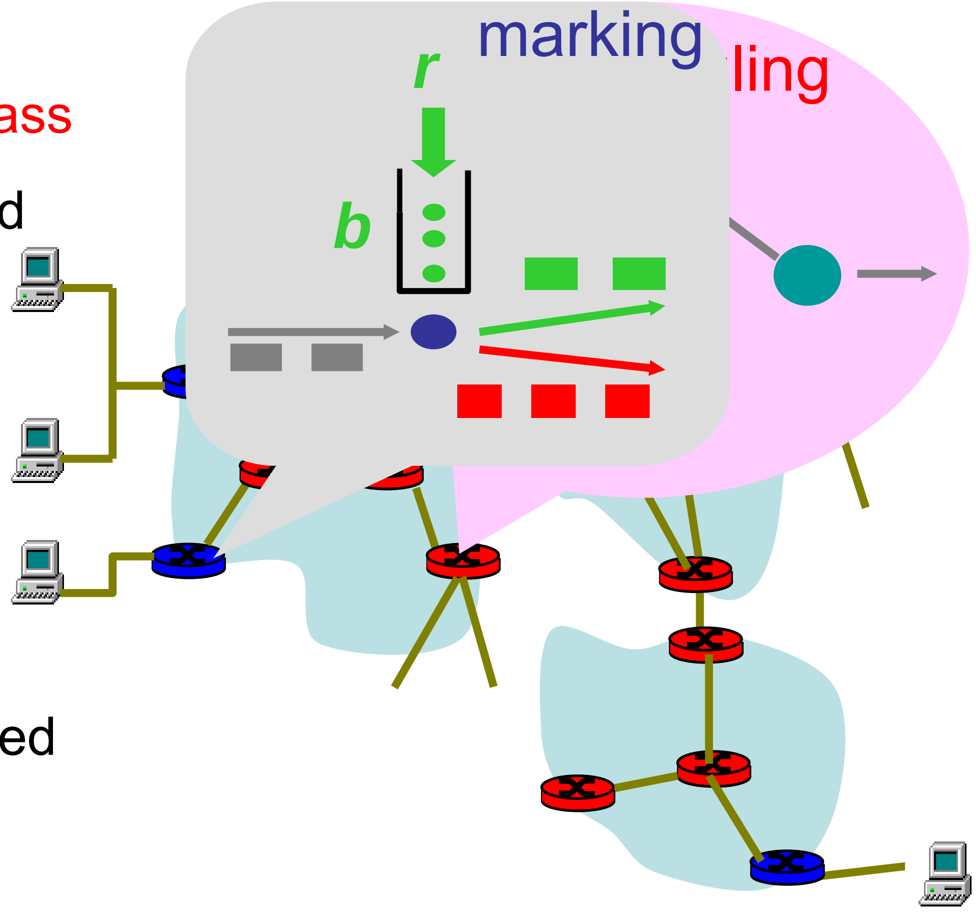
IETF Differentiated Services

- want "qualitative" service classes
  - "behaves like a wire"
  - relative service distinction: Platinum, Gold, Silver
- *scalability:* simple functions in network core, relatively complex functions at edge routers (or hosts)
  - in contrast to IETF Integrated Services: signaling, maintaining per-flow router state difficult with large number of flows
- don't define define service classes, provide functional components to build service classes

Diffserv Architecture

Edge router:

- per-flow traffic management
- marks packets according to class
- marks packets as in-profile and out-profile

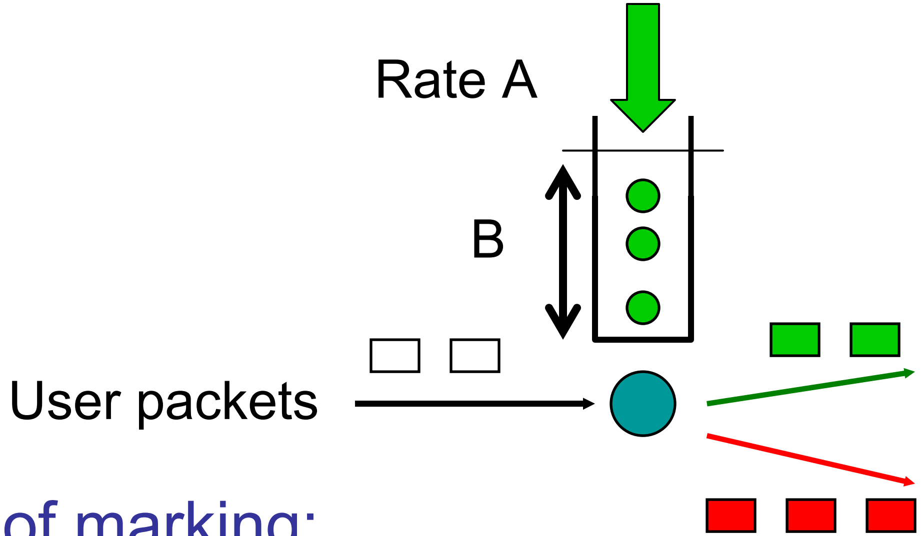


Core router:

- per class traffic management
- buffering and scheduling based on marking at edge
- preference given to in-profile packets

Edge-router Packet Marking

- profile: pre-negotiated rate A, bucket size B
- packet marking at edge based on per-flow profile

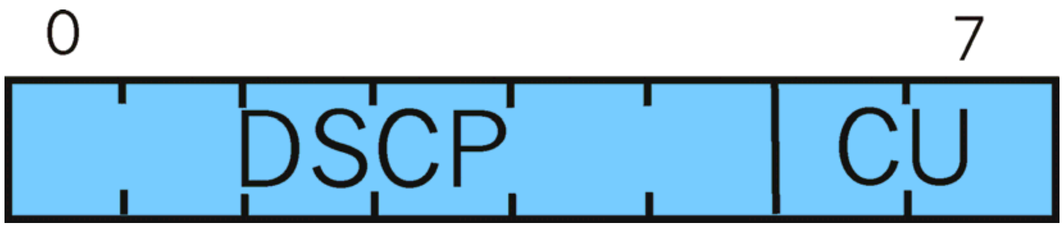


Possible usage of marking:

- class-based marking: packets of different classes marked differently
- intra-class marking: conforming portion of flow marked differently than non-conforming one

Classification and Conditioning

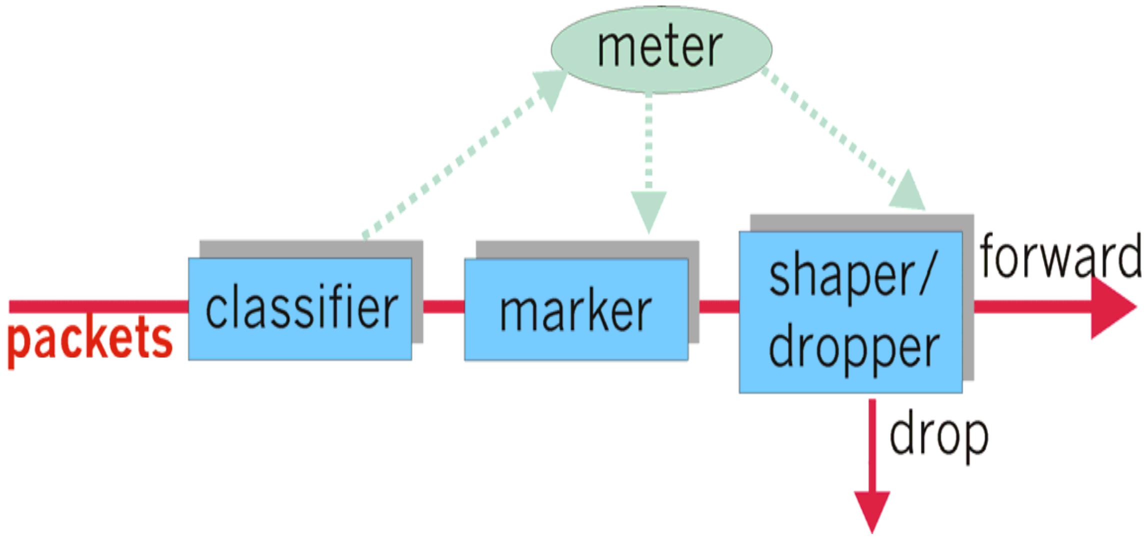
- Packet is marked in the Type of Service (TOS) in IPv4, and Traffic Class in IPv6
- 6 bits used for Differentiated Service Code Point (DSCP) and determine PHB that the packet will receive
- 2 bits can be used for congestion notification: Explicit Congestion Notification (ECN), RFC 3168



Classification and Conditioning

May be desirable to limit traffic injection rate of some class:

- user declares traffic profile (e.g., rate, burst size)
- traffic metered, shaped or dropped if non-conforming



Forwarding (PHB)

- PHB result in a different observable (measurable) forwarding performance behavior
- PHB does not specify what mechanisms to use to ensure required PHB performance behavior
- Examples:
  - Class A gets x% of outgoing link bandwidth over time intervals of a specified length
  - Class A packets leave first before packets from class B

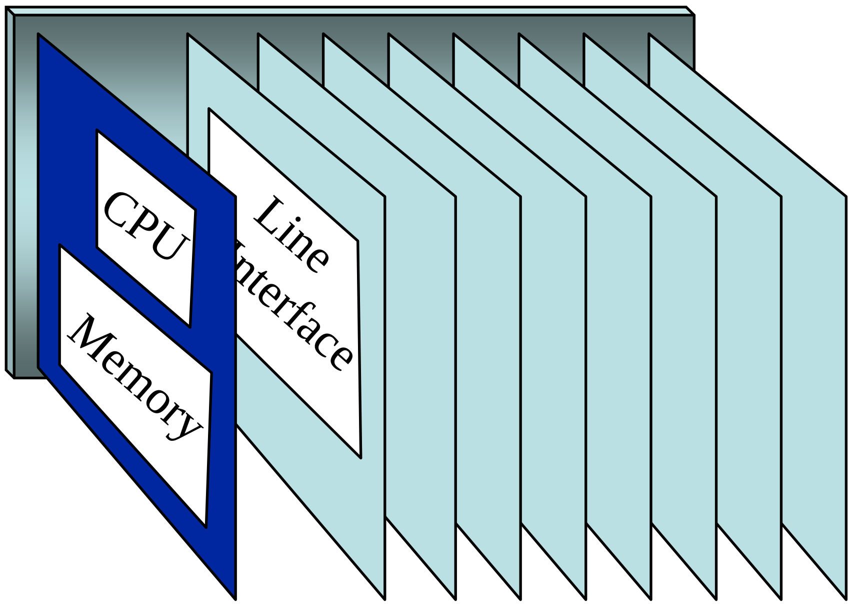
Forwarding (PHB)

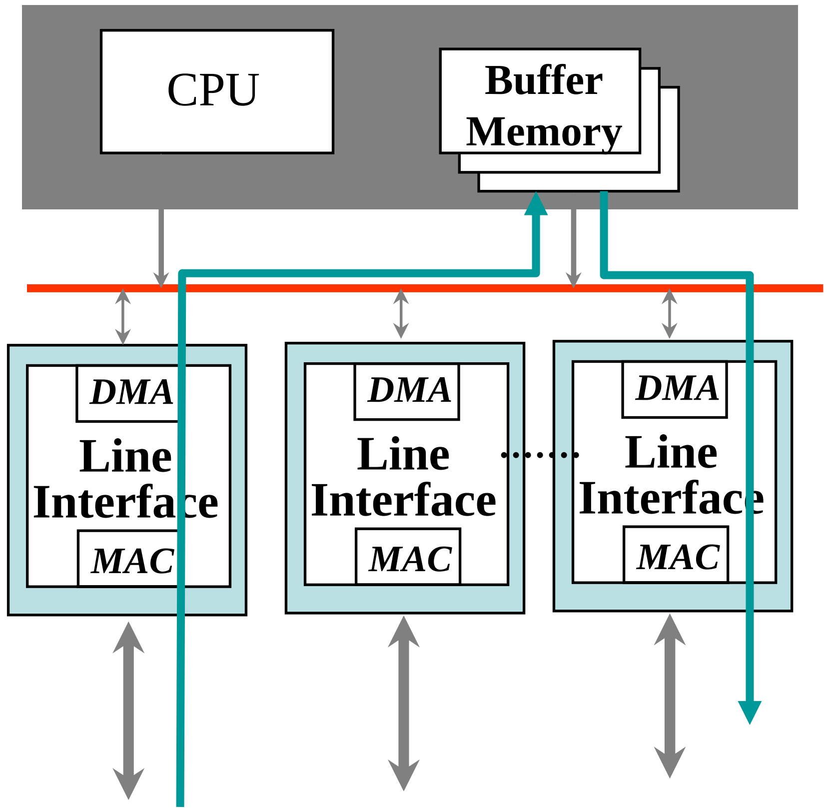
PHBs being developed:

- Expedited Forwarding: packet departure rate of a class equals or exceeds specified rate
  - logical link with a minimum guaranteed rate
- Assured Forwarding: e.g. 4 classes of traffic
  - each class guaranteed minimum amount of bandwidth and a minimum of buffering
  - packets each class have one of three possible drop preferences; in case of congestion routers discard packets based on drop preference values

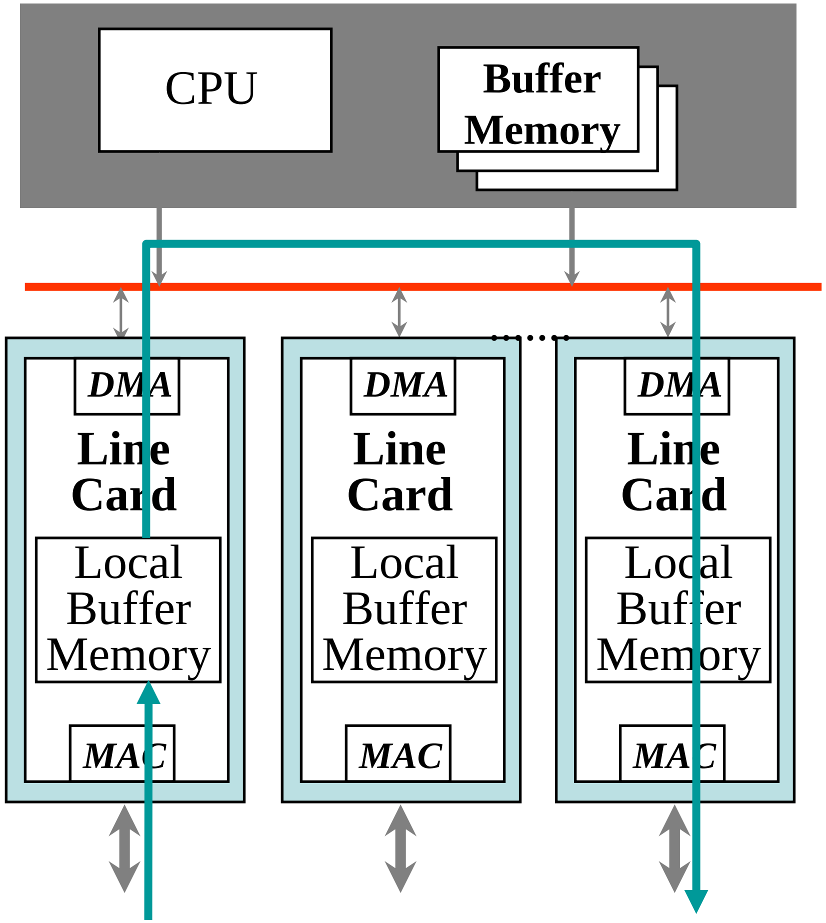
The Evolution of IP Routers

First-Generation IP Routers

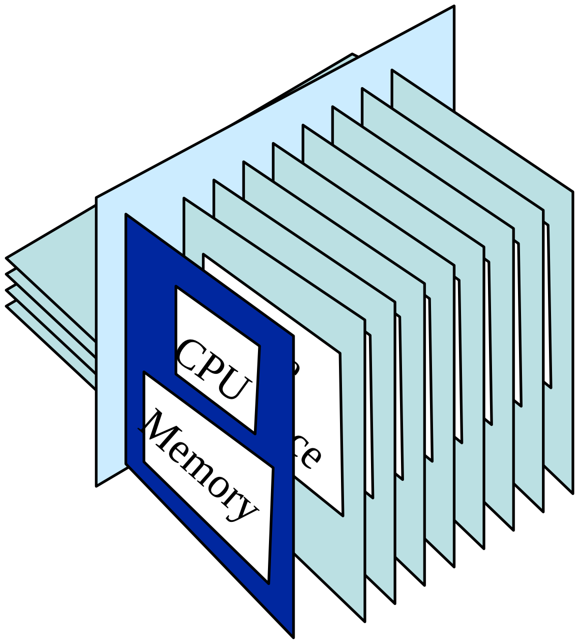


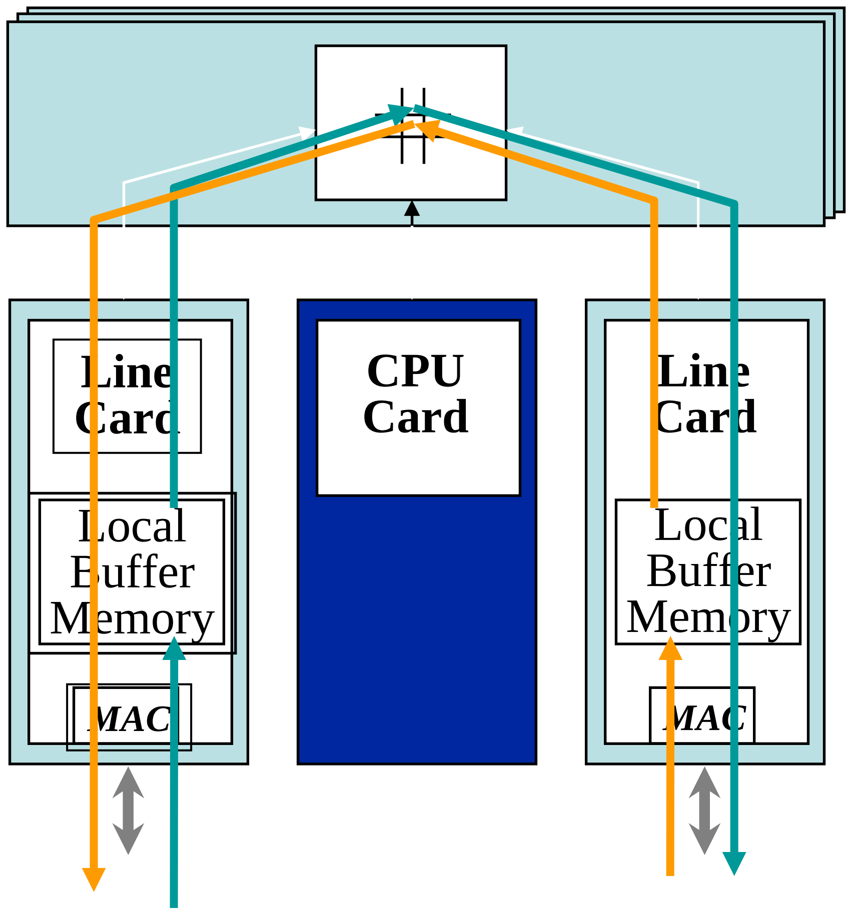


Second-Generation IP Routers



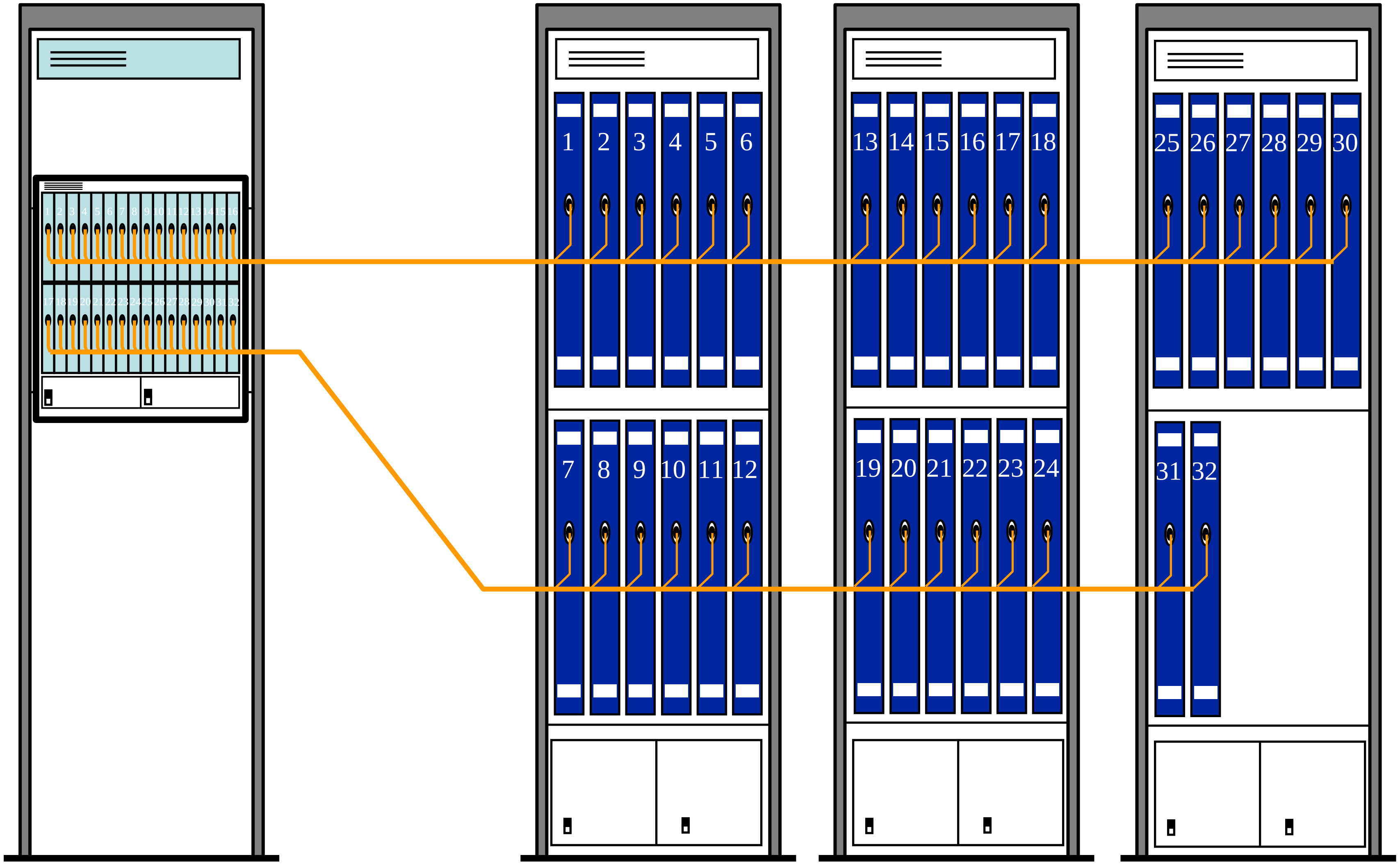
Third-Generation Switches/Routers





Fourth-Generation Switches/Routers

Clustering and Multistage



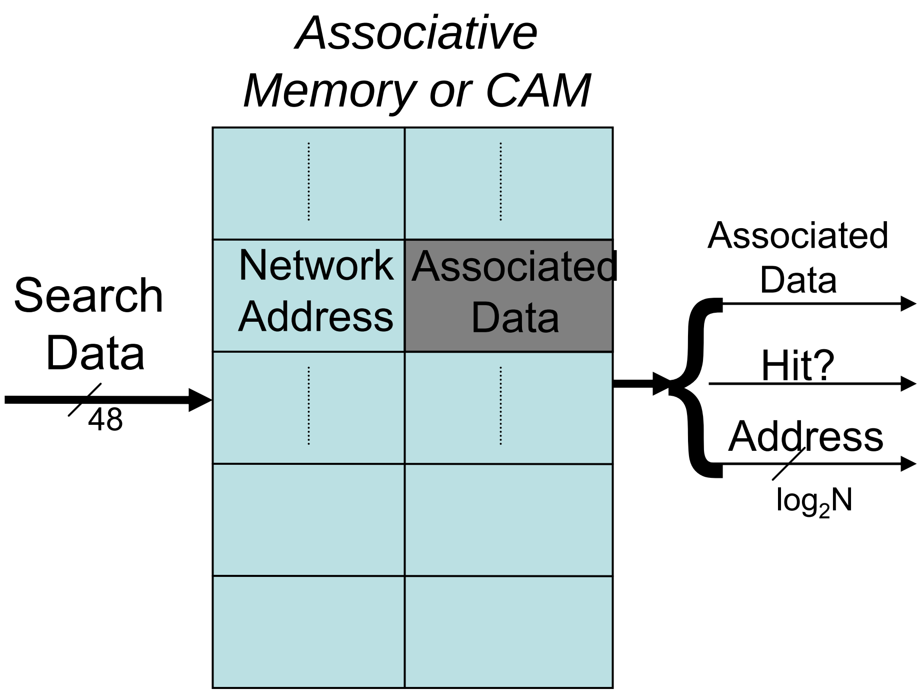
Forwarding Implementations

Principles

Credits:
Nick McKeown, Stanford University

Forwarding Implementation (1)

Associative Lookups



Advantages:

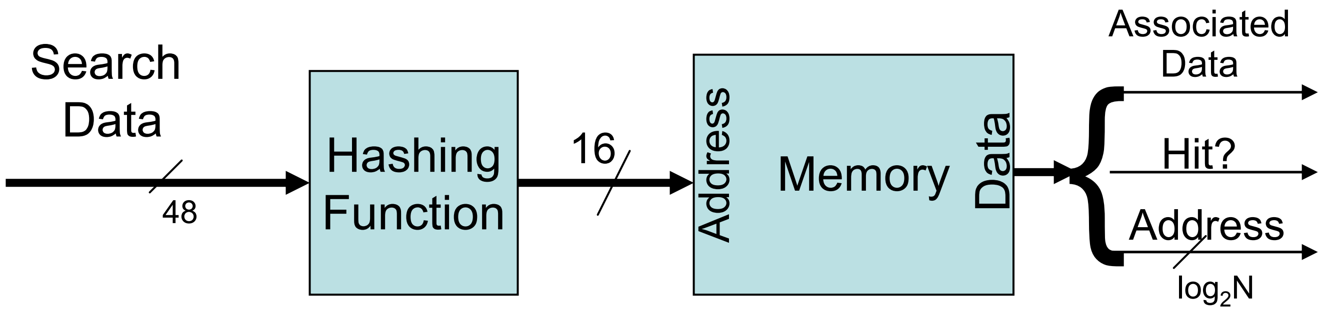
- Simple

Disadvantages

- Relatively slow
- High power consumption
- Small
- Expensive

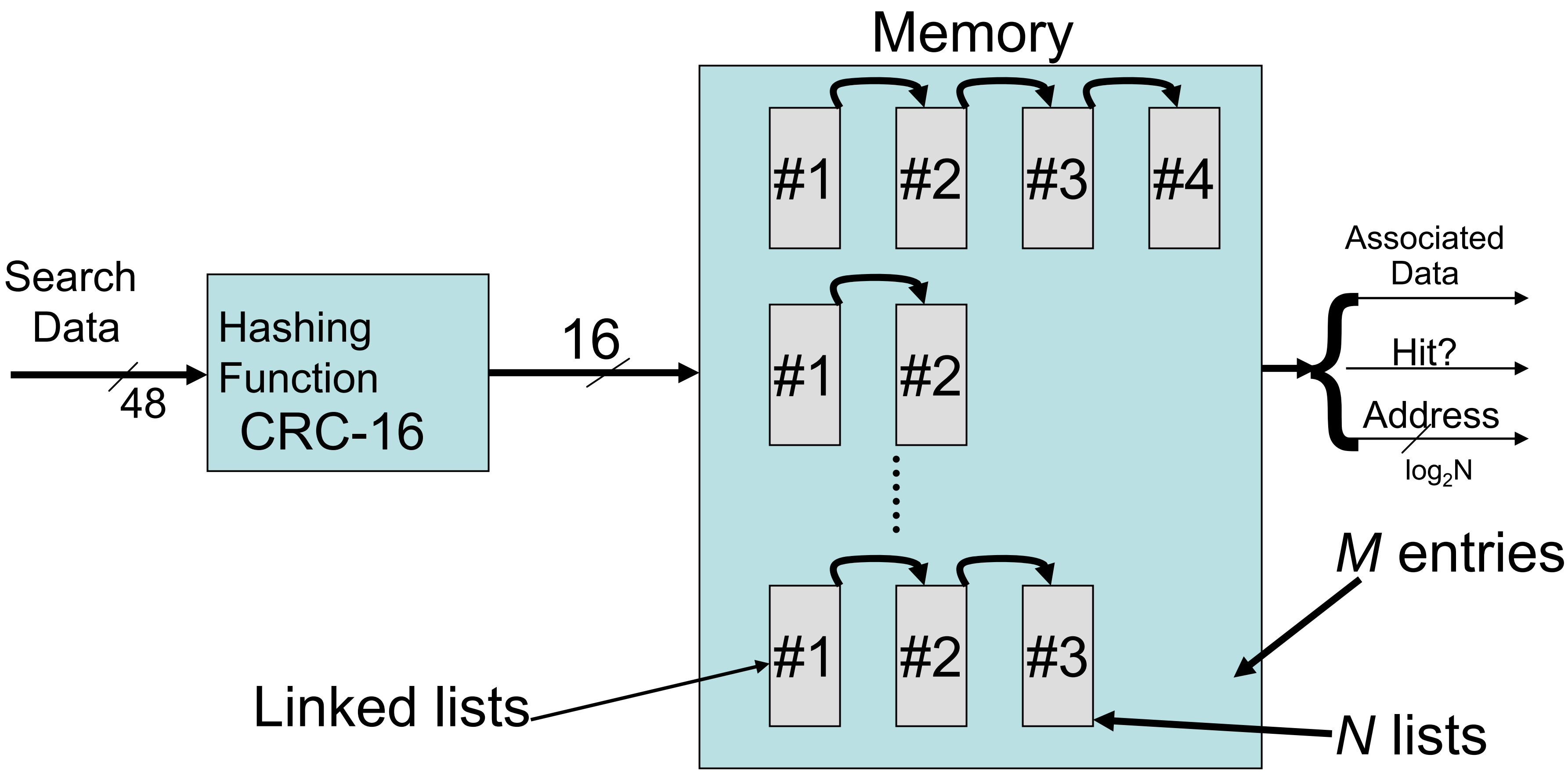
Forwarding Implementation (2)

Hashing



Lookups Using Hashing

An example



Lookups Using Hashing

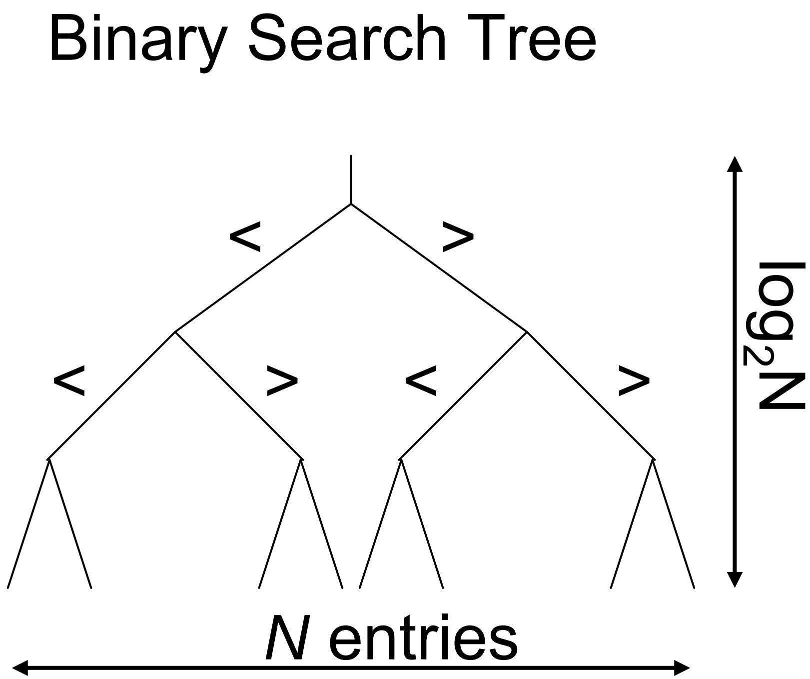
Advantages:

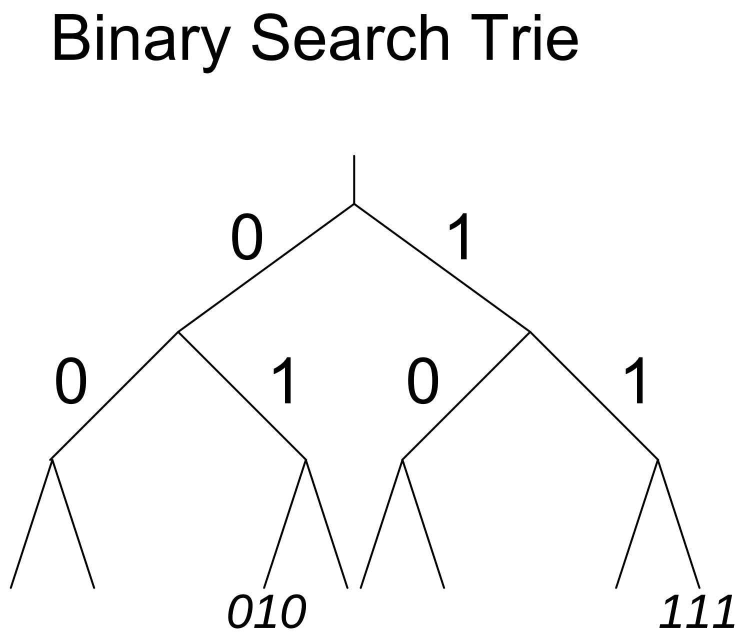
- Simple
- Expected lookup time can be small

Disadvantages

- Non-deterministic lookup time
- Inefficient use of memory

Forwarding Implementation: Trees and Tries



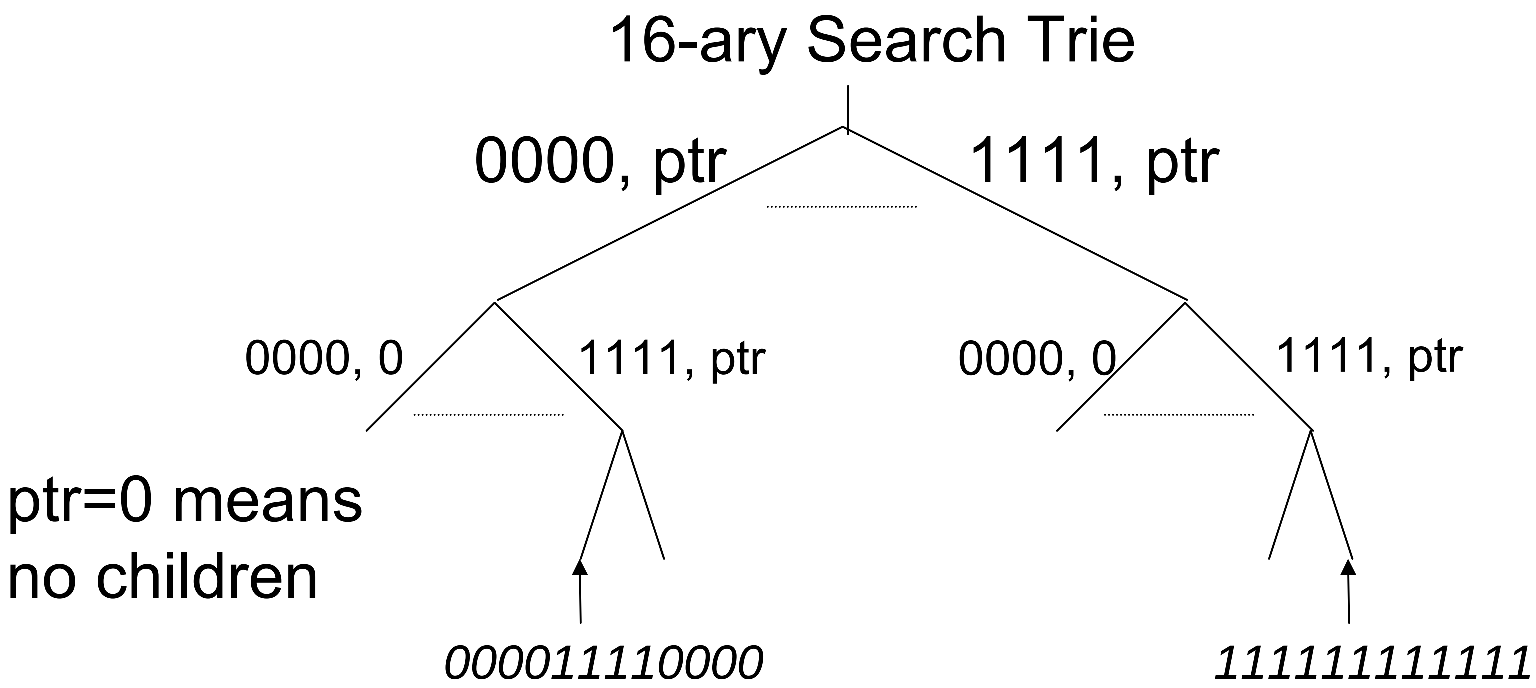


A *trie* (from re**trie**val), is a multi-way tree structure useful for storing strings over an alphabet.

Simple Tries and Exact Matching in Ethernet

- Each address in the forwarding table is of fixed length
  - E.g. using a simple binary search trie it takes 48 steps to resolve an MAC address lookup
  - When the table is much smaller than 2^48 (!), this seems like a lot of steps
  - Instead of matching one bit per level, why not m bits per level?

Trees and Tries: *Multiway tries*

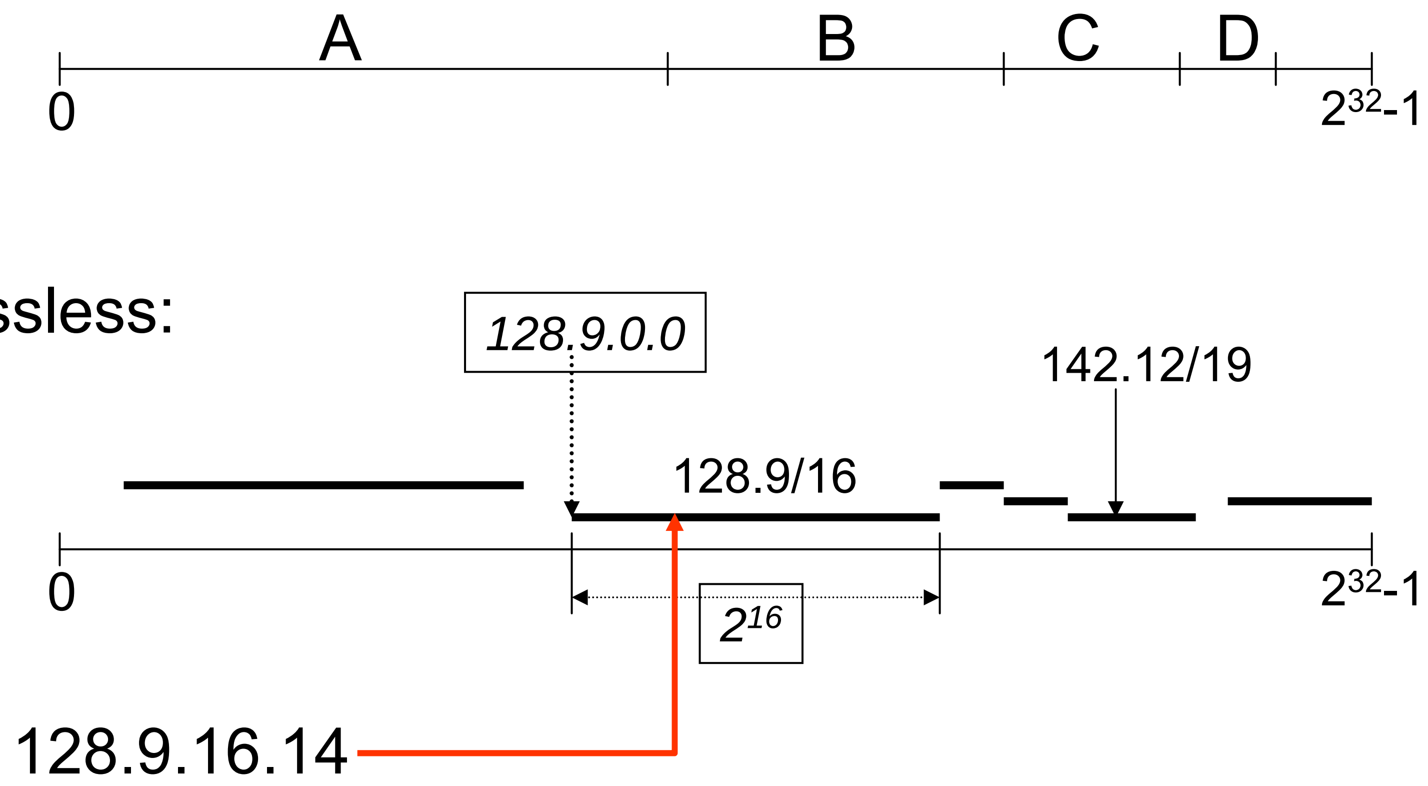


IP Forwarding

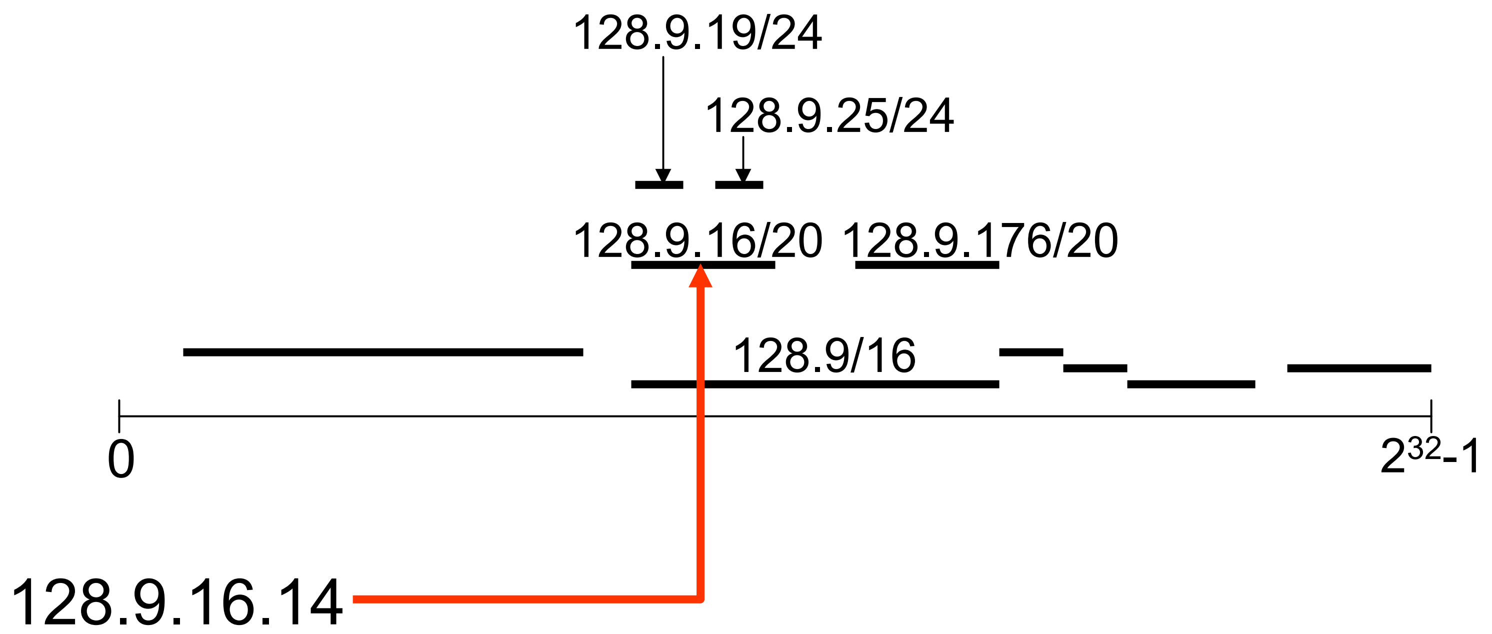
IP Routers: *CIDR*

Class-based:

Classless:



IP Routers: *CIDR*



Most specific route = “longest matching prefix”

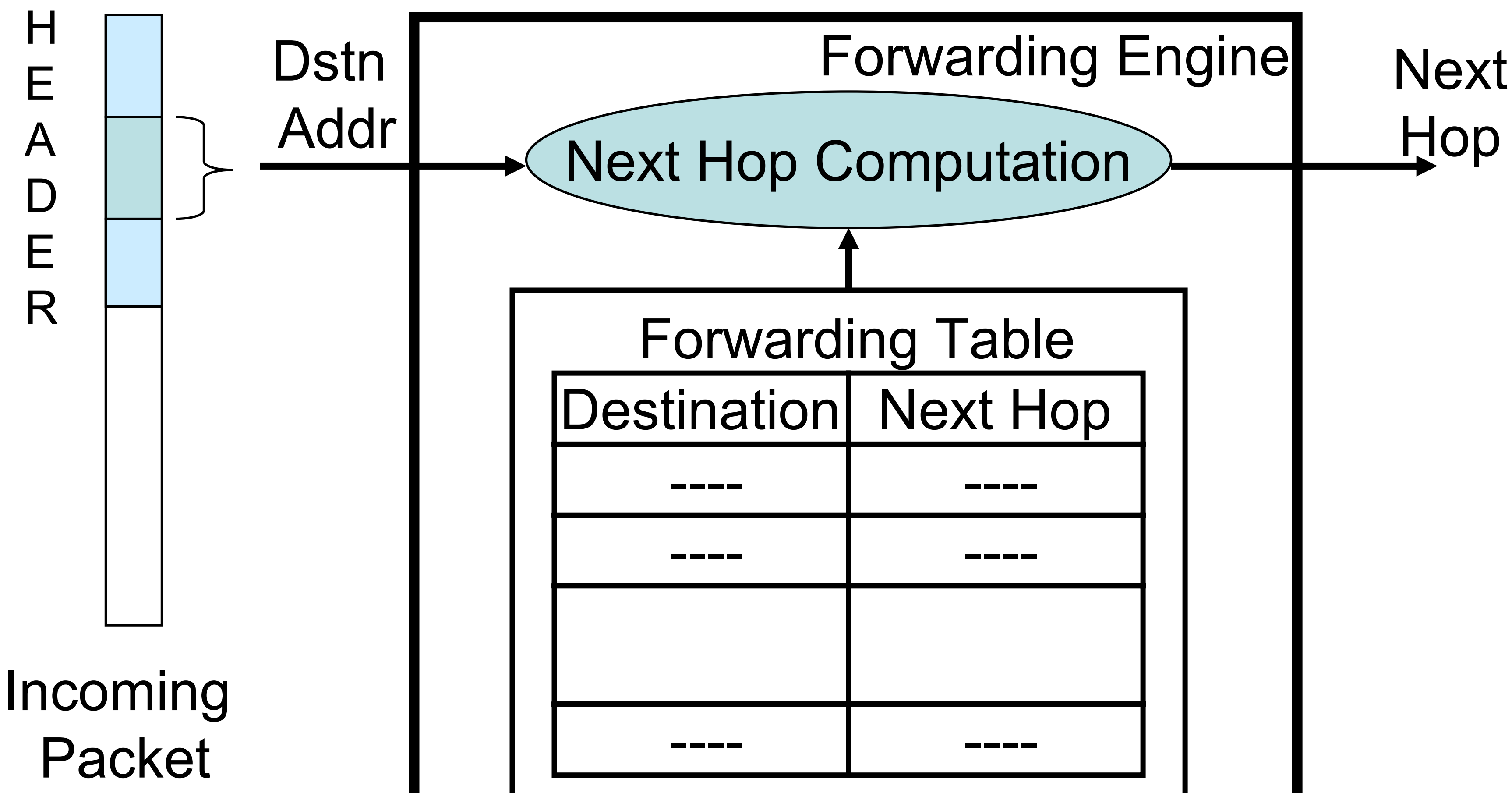
IP Routers: *Metrics for Lookups*

128.9.16.14 → 128.9.16/20

Prefix	Port
65/24	3
128.9/16	5
128.9.16/20	2
128.9.19/24	7
128.9.25/24	10
128.9.176/20	1
142.12/19	3

- Lookup time
- Storage space
- Update time

IP Router: *Lookup*



IPv4 unicast destination address based lookup

Need more than IPv4 unicast lookups

- Multicast
  - PIM-SM (Protocol-Independent Multicast, Sparse Mode)
    - Longest Prefix Matching on the source (S) and group (G) address
    - Start specific, subsequently apply wildcards:
      try (S,G) followed by (*,G) followed by (*,*,RP)
    - Check Incoming Interface
  - DVMRP:
    - Incoming Interface Check followed by (S,G) lookup

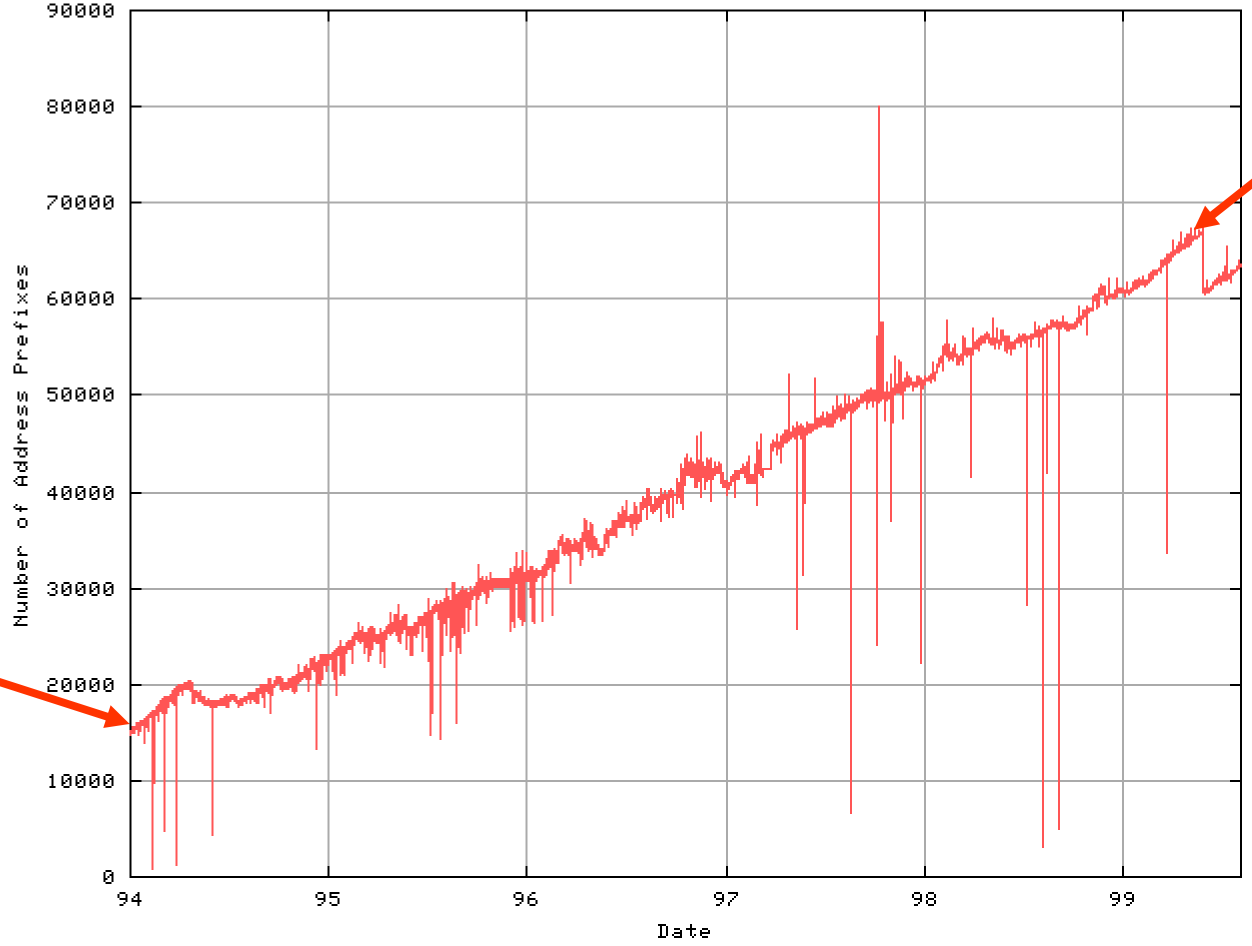
- IPv6
  - 128-bit destination address field

Required Lookup Performance

Line	Line Rate	Pkt-size=40B	Pkt-size=240B
T1	1.5Mbps	4.68 Kpps	0.78 Kpps
OC3	155Mbps	480 Kpps	80 Kpps
OC12	622Mbps	1.94 Mpps	323 Kpps
OC48	2.5Gbps	7.81 Mpps	1.3 Mpps
OC192	10 Gbps	31.25 Mpps	5.21 Mpps

Gigabit Ethernet (84B packets): 1.49 Mpps

Size of the Routing Table



Exponential growth before CIDR

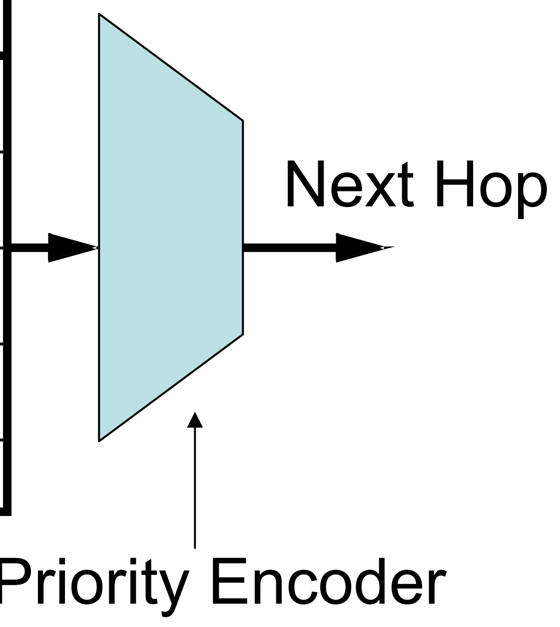
About 10k new prefixes per year

Source: http://www.telstra.net/ops/bgptable.html

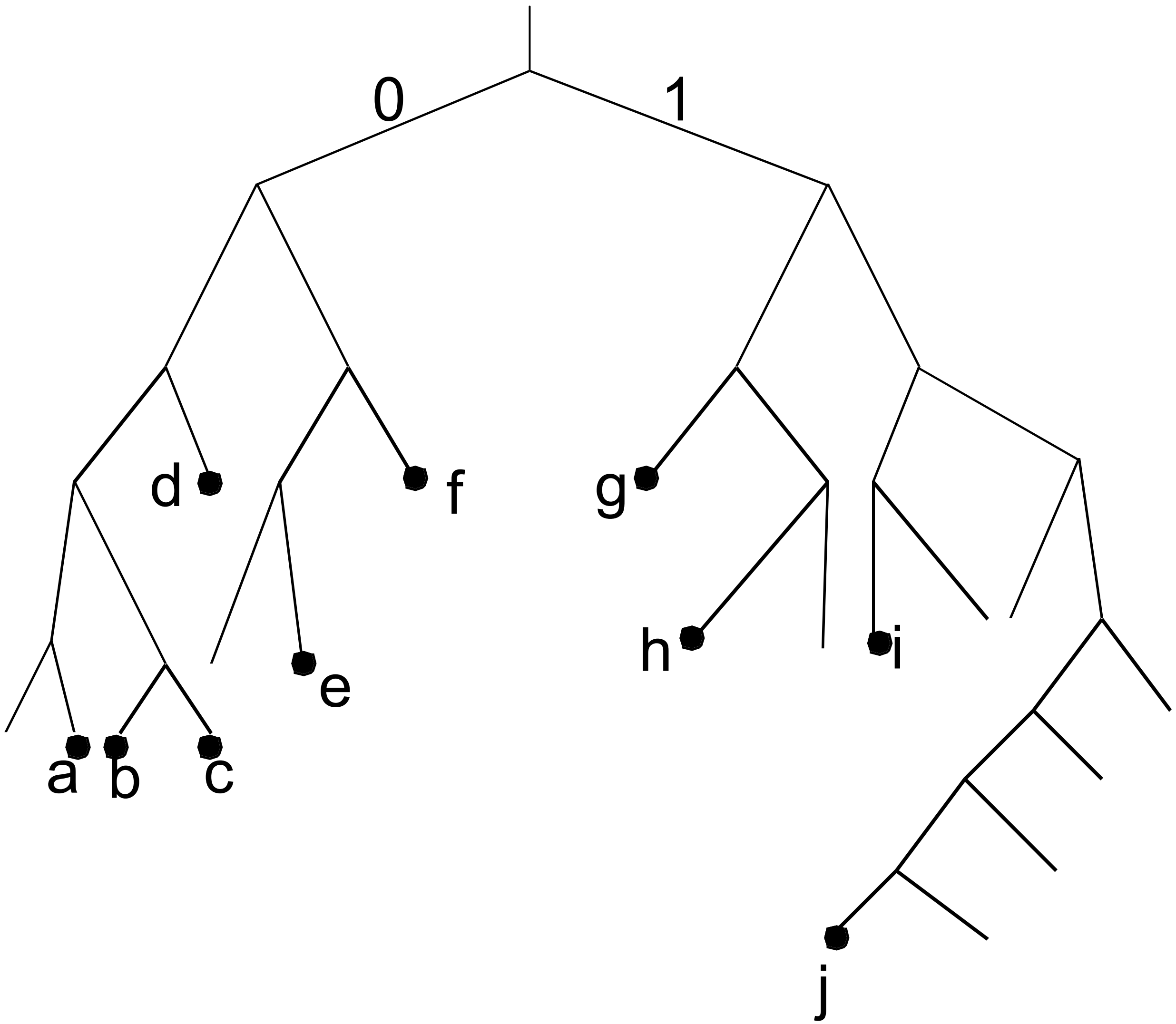
Method: CAMs (Content-Addressable Memories)

Associative Memory

Value	Mask	
10.0.0.0	255.0.0.0	R1
10.1.0.0	255.255.0.0	R2
10.1.1.0	255.255.255.0	R3
10.1.3.0	255.255.255.0	R4
10.1.3.1	255.255.255.255	R4



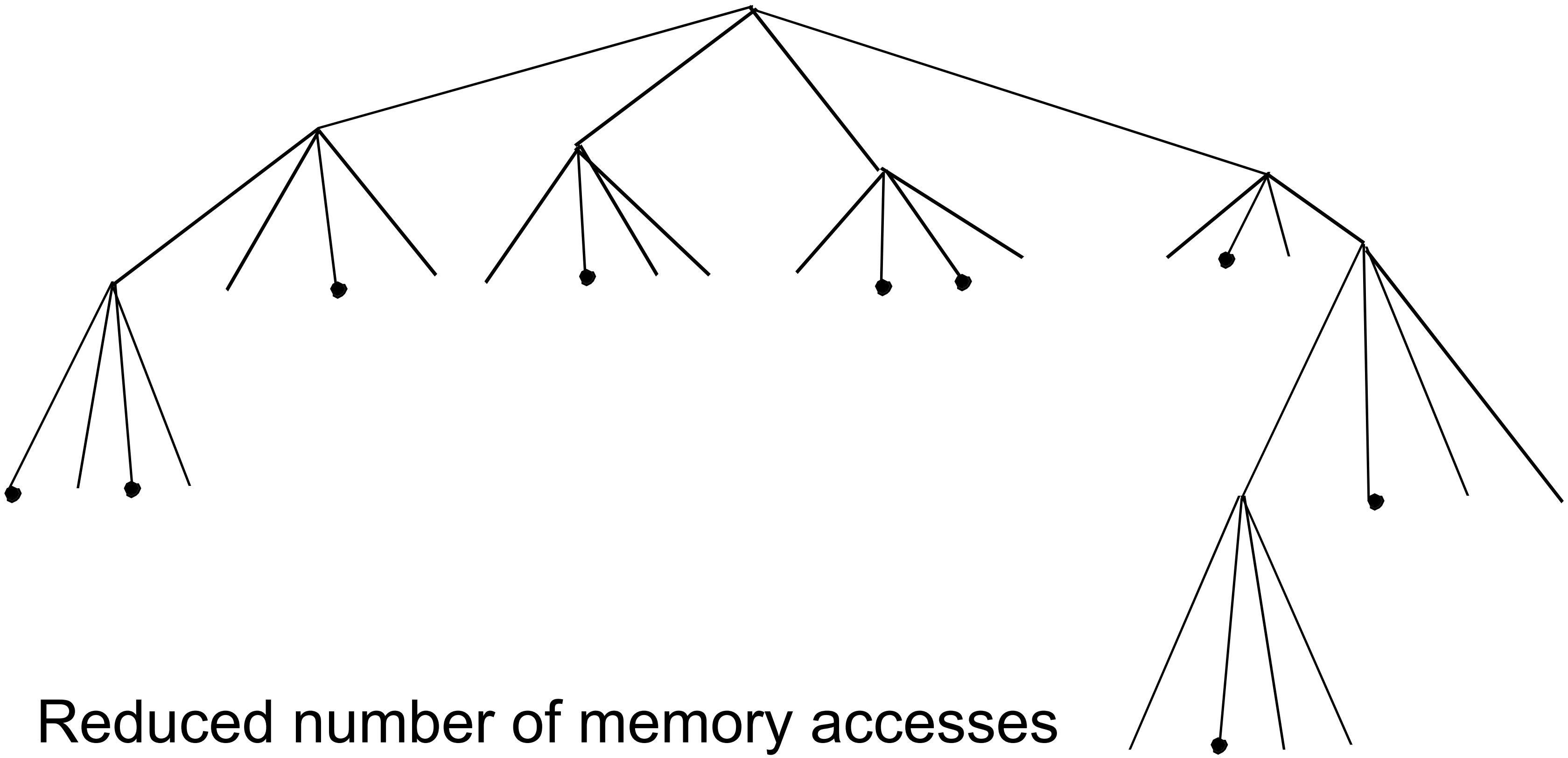
Method: Binary Tries



Example Prefixes

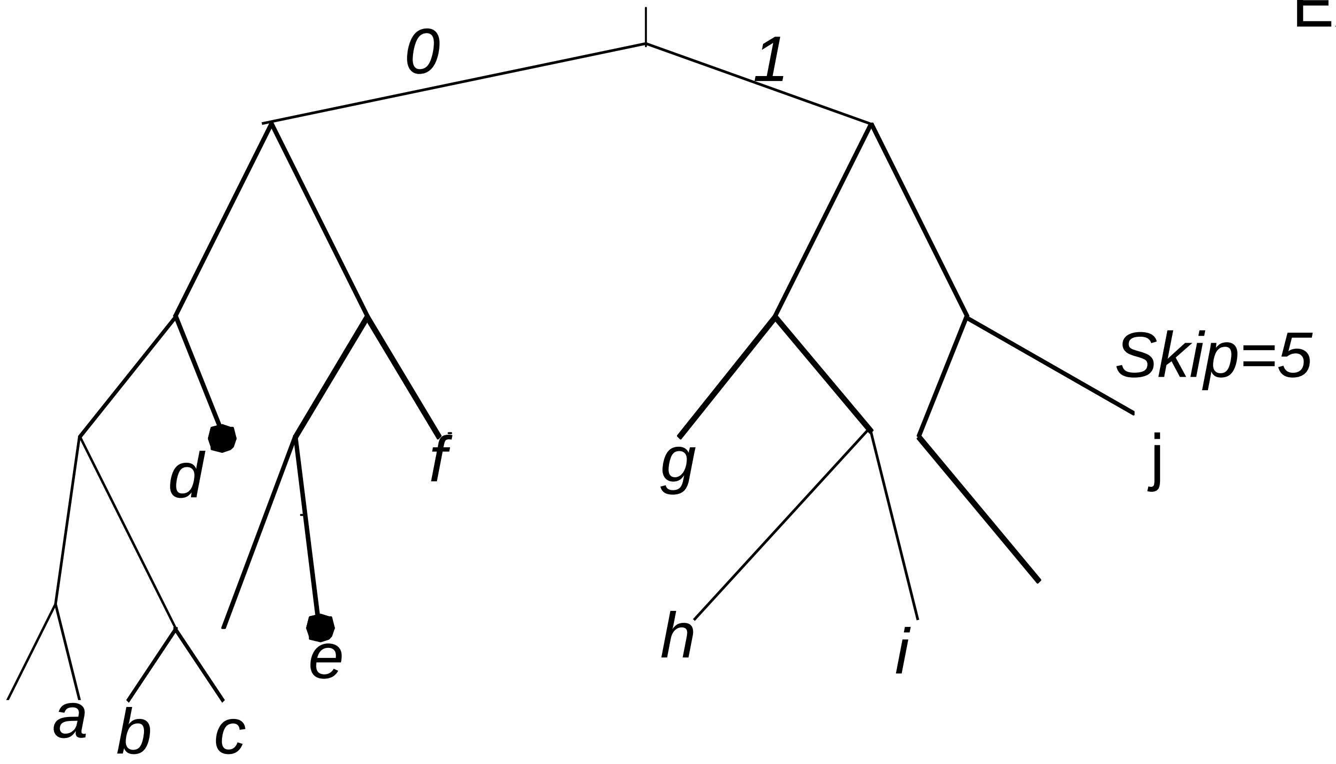
a) 00001
b) 00010
c) 00011
d) 001
e) 0101
f) 011
g) 100
h) 1010
i) 1100
j) 11110000

Four-way tries



Reduced number of memory accesses
But greater wasted space...

Method: Patricia Tree



Example Prefixes

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

Disadvantages

- Many memory accesses
- Pointers take a lot of space
  (Total storage for 40K entries is 2MB)

Advantages

- General solution
- Extensible to wider fields

Method: Compacting Forwarding Tables

- Optimize the data structure to store 40,000 routing table entries in about 150-160kBytes.
- Rely on the compacted data structure to be residing in the primary or secondary cache of a fast processor.
- Achieves e.g. 2Mpps on a Pentium.

Disadvantages	*Advantages*
• Only 60% actually cached • Scalability to larger tables • Handling updates is complex	• Good software solution for low speeds and small routing tables.

Forwarding Decisions

Packet Classification

Providing Value-Added Services: Some examples

- Differentiated services
  - Regard traffic from AS#33 as `platinum-grade'
- Access Control Lists
  - Deny udp host 194.72.72.33 194.72.6.64 0.0.0.15 eq snmp
- Committed Access Rate
  - Rate limit WWW traffic from sub-interface#739 to 10Mbps
- Policy-based Routing
  - Route all voice traffic through the ATM network
- Peering Arrangements
  - Restrict the total amount of traffic of precedence 7 from
  - MAC address N to 20 Mbps between 10 am and 5pm
- Accounting and Billing
  - Generate hourly reports of traffic from MAC address M

Flow Classification