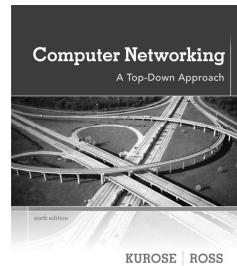


Chapter 4 Network Layer



*Computer
Networking: A Top
Down Approach*
6th edition
Jim Kurose, Keith Ross
Addison-Wesley
March 2012

All material copyright 1996-2013
J.F Kurose and K.W. Ross, All Rights Reserved

Network Layer 4-1

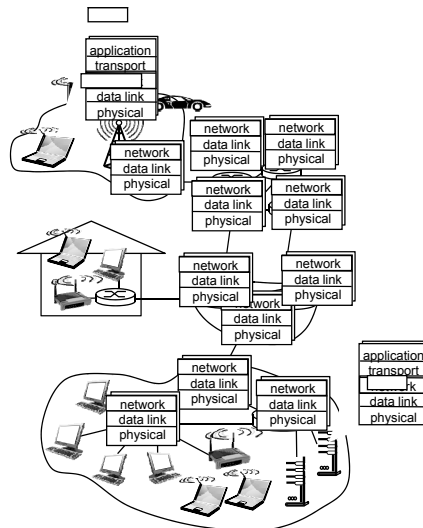
Chapter 4: outline

- 4.1 introduction
- 4.2 virtual circuit and datagram networks
- 4.3 what's inside a router
- 4.4 IP: Internet Protocol
 - datagram format
 - IPv4 addressing
 - ICMP
 - IPv6
- 4.5 routing algorithms
 - link state
 - distance vector
 - hierarchical routing
- 4.6 routing in the Internet
 - RIP
 - OSPF
 - BGP
- 4.7 broadcast and multicast routing

Network Layer 4-2

Network layer

- ❖ transport segment from sending to receiving host
- ❖ on sending side encapsulates segments into datagrams
- ❖ on receiving side, delivers segments to transport layer
- ❖ network layer protocols in every host, router
- ❖ router examines header fields in all IP datagrams passing through it



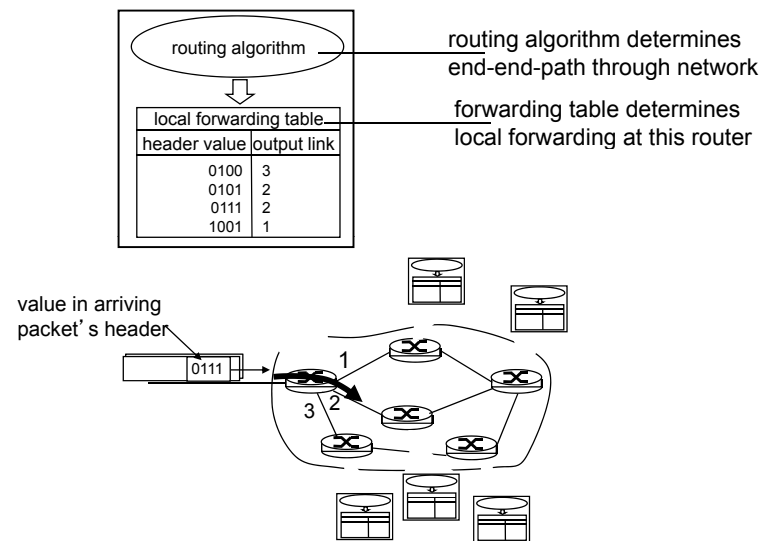
Network Layer 4-3

Two key network-layer functions

- ❖ *forwarding*: move packets from router's input to appropriate router output
 - ❖ *routing*: determine route taken by packets from source to dest.
 - *routing algorithms*
- analogy:*
- ❖ *routing*: process of planning trip from source to dest
 - ❖ *forwarding*: process of getting through single interchange

Network Layer 4-4

Interplay between routing and forwarding



Network Layer 4-5

Connection setup

- ❖ 3rd important function in *some* network architectures:
 - ATM, frame relay, X.25
- ❖ before datagrams flow, two end hosts *and* intervening routers establish virtual connection
 - routers get involved
- ❖ network vs transport layer connection service:
 - *network*: between two hosts (may also involve intervening routers in case of VCs)
 - *transport*: between two processes

Network Layer 4-6

Network service model

Q: What *service model* for “channel” transporting datagrams from sender to receiver?

example services for individual datagrams:

- ❖ guaranteed delivery
- ❖ guaranteed delivery with less than 40 msec delay

example services for a flow of datagrams:

- ❖ in-order datagram delivery
- ❖ guaranteed minimum bandwidth to flow
- ❖ restrictions on changes in inter-packet spacing

Network Layer 4-7

Network layer service models:

Network Architecture	Service Model	Guarantees ?				Congestion feedback
		Bandwidth	Loss	Order	Timing	
Internet	best effort	none	no	no	no	no (inferred via loss)
ATM	CBR	constant rate	yes	yes	yes	no congestion
ATM	VBR	guaranteed rate	yes	yes	yes	no congestion
ATM	ABR	guaranteed minimum	no	yes	no	yes
ATM	UBR	none	no	yes	no	no

Network Layer 4-8

Chapter 4: outline

- 4.1 introduction
- 4.2 virtual circuit and datagram networks
- 4.3 what's inside a router
- 4.4 IP: Internet Protocol
 - datagram format
 - IPv4 addressing
 - ICMP
 - IPv6
- 4.5 routing algorithms
 - link state
 - distance vector
 - hierarchical routing
- 4.6 routing in the Internet
 - RIP
 - OSPF
 - BGP
- 4.7 broadcast and multicast routing

Network Layer 4-9

Connection, connection-less service

- ❖ *datagram* network provides network-layer *connectionless* service
- ❖ *virtual-circuit* network provides network-layer *connection* service
- ❖ analogous to TCP/UDP connecton-oriented / connectionless transport-layer services, but:
 - *service*: host-to-host
 - *no choice*: network provides one or the other
 - *implementation*: in network core

Network Layer 4-10

Virtual circuits

“source-to-dest path behaves much like telephone circuit”

- performance-wise
 - network actions along source-to-dest path
-
- ❖ call setup, teardown for each call *before* data can flow
 - ❖ each packet carries VC identifier (not destination host address)
 - ❖ every router on source-dest path maintains “state” for each passing connection
 - ❖ link, router resources (bandwidth, buffers) may be *allocated* to VC (dedicated resources = predictable service)

Network Layer 4-11

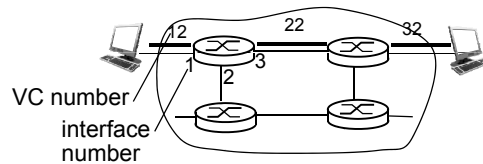
VC implementation

a VC consists of:

1. *path* from source to destination
 2. *VC numbers*, one number for each link along path
 3. *entries in forwarding tables* in routers along path
- ❖ packet belonging to VC carries VC number (rather than dest address)
 - ❖ VC number can be changed on each link.
 - new VC number comes from forwarding table

Network Layer 4-12

VC forwarding table



forwarding table in
northwest router:

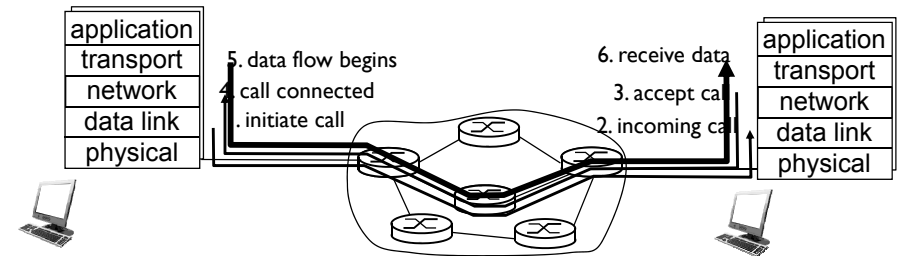
Incoming interface	Incoming VC #	Outgoing interface	Outgoing VC #
1	12	3	22
2	63	1	18
3	7	2	17
1	97	3	87
...	...	...	...

VC routers maintain connection state information!

Network Layer 4-13

Virtual circuits: signaling protocols

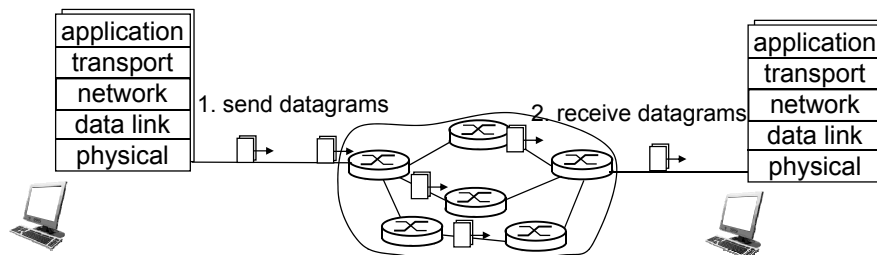
- ❖ used to setup, maintain teardown VC
- ❖ used in ATM, frame-relay, X.25
- ❖ not used in today's Internet



Network Layer 4-14

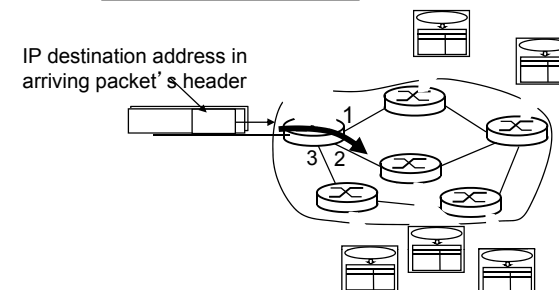
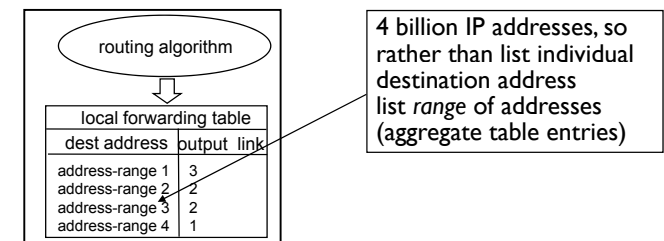
Datagram networks

- ❖ no call setup at network layer
- ❖ routers: no state about end-to-end connections
 - no network-level concept of "connection"
- ❖ packets forwarded using destination host address



Network Layer 4-15

Datagram forwarding table



Network Layer 4-16

Datagram forwarding table

Destination Address Range	Link Interface
11001000 00010111 00010000 00000000 through 11001000 00010111 00010111 11111111	0
11001000 00010111 00011000 00000000 through 11001000 00010111 00011000 11111111	1
11001000 00010111 00011001 00000000 through 11001000 00010111 00011111 11111111	2
otherwise	3

Q: but what happens if ranges don't divide up so nicely?

Network Layer 4-17

Longest prefix matching

longest prefix matching

when looking for forwarding table entry for given destination address, use *longest* address prefix that matches destination address.

Destination Address Range	Link interface
11001000 00010111 00010*** *****	0
11001000 00010111 00011000 *****	1
11001000 00010111 00011*** *****	2
otherwise	3

examples:

DA: 11001000 00010111 00010110 10100001

which interface?

DA: 11001000 00010111 00011000 10101010

which interface?

Network Layer 4-18

Datagram or VC network: why?

Internet (datagram)

- ❖ data exchange among computers
 - “elastic” service, no strict timing req.
- ❖ many link types
 - different characteristics
 - uniform service difficult
- ❖ “smart” end systems (computers)
 - can adapt, perform control, error recovery
 - **simple inside network, complexity at “edge”**

ATM (VC)

- ❖ evolved from telephony
- ❖ human conversation:
 - strict timing, reliability requirements
 - need for guaranteed service
- ❖ “dumb” end systems
 - telephones
 - **complexity inside network**

Network Layer 4-19

Chapter 4: outline

4.1 introduction

4.2 virtual circuit and datagram networks

4.3 what's inside a router

4.4 IP: Internet Protocol

- datagram format
- IPv4 addressing
- ICMP
- IPv6

4.5 routing algorithms

- link state
- distance vector
- hierarchical routing

4.6 routing in the Internet

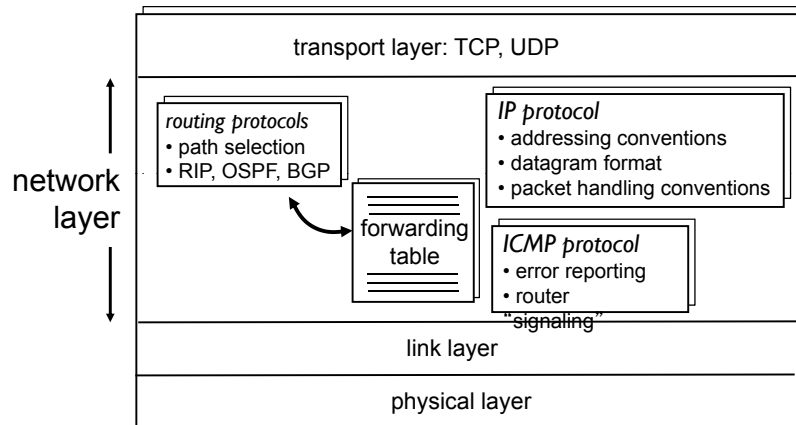
- RIP
- OSPF
- BGP

4.7 broadcast and multicast routing

Network Layer 4-20

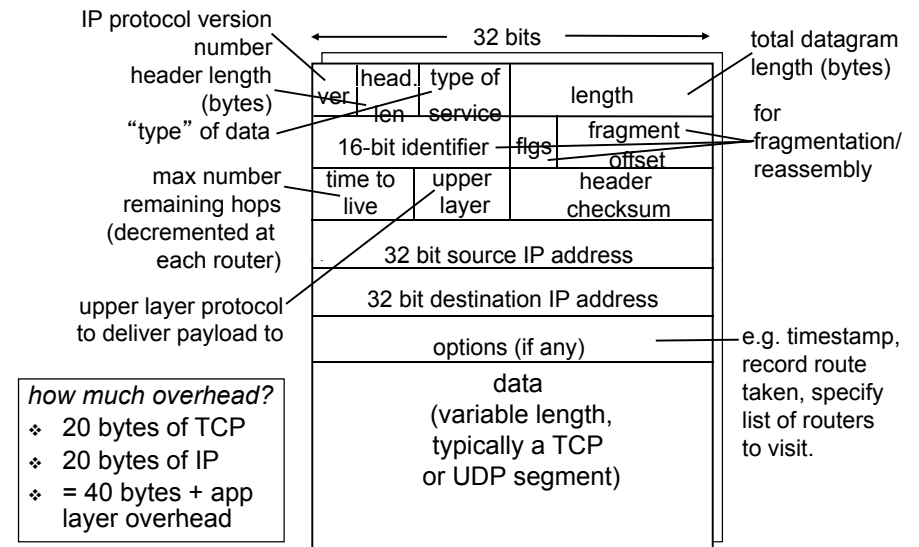
The Internet network layer

host, router network layer functions:



Network Layer 4-21

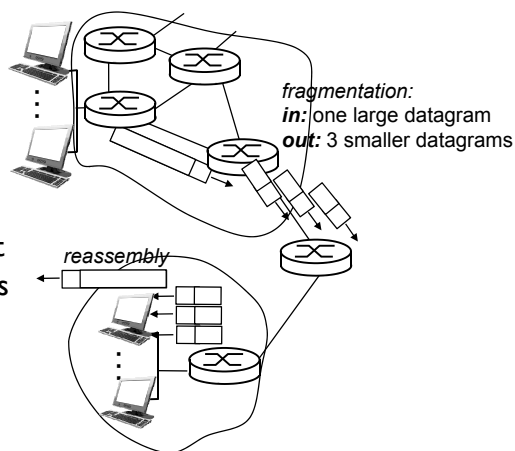
IP datagram format



Network Layer 4-22

IP fragmentation, reassembly

- ❖ network links have MTU (max.transfer size) - largest possible link-level frame
 - different link types, different MTUs
- ❖ large IP datagram divided ("fragmented") within net
 - one datagram becomes several datagrams
 - "reassembled" only at final destination
 - IP header bits used to identify, order related fragments

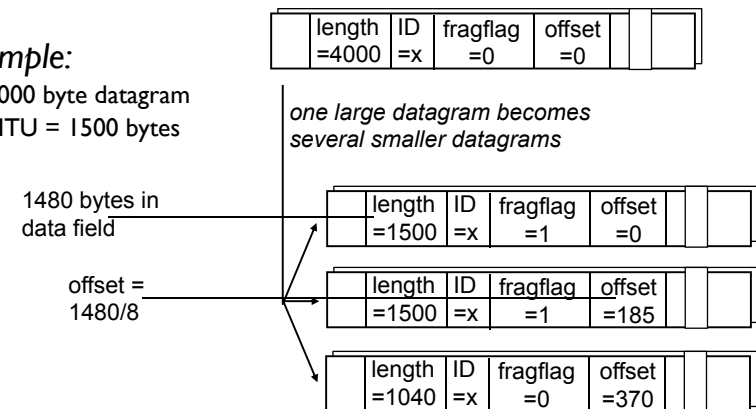


Network Layer 4-23

IP fragmentation, reassembly

example:

- ❖ 4000 byte datagram
- ❖ MTU = 1500 bytes



Network Layer 4-24

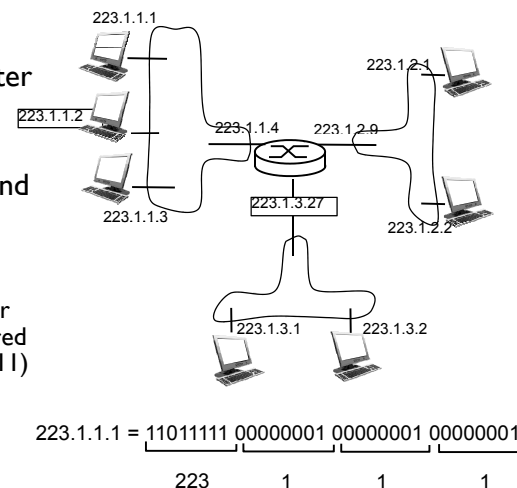
Chapter 4: outline

- 4.1 introduction
- 4.2 virtual circuit and datagram networks
- 4.3 what's inside a router
- 4.4 IP: Internet Protocol
 - datagram format
 - IPv4 addressing
 - ICMP
 - IPv6
- 4.5 routing algorithms
 - link state
 - distance vector
 - hierarchical routing
- 4.6 routing in the Internet
 - RIP
 - OSPF
 - BGP
- 4.7 broadcast and multicast routing

Network Layer 4-25

IP addressing: introduction

- ❖ **IP address:** 32-bit identifier for host, router interface
- ❖ **interface:** connection between host/router and physical link
 - router's typically have multiple interfaces
 - host typically has one or two interfaces (e.g., wired Ethernet, wireless 802.11)
- ❖ **IP addresses associated with each interface**



Network Layer 4-26

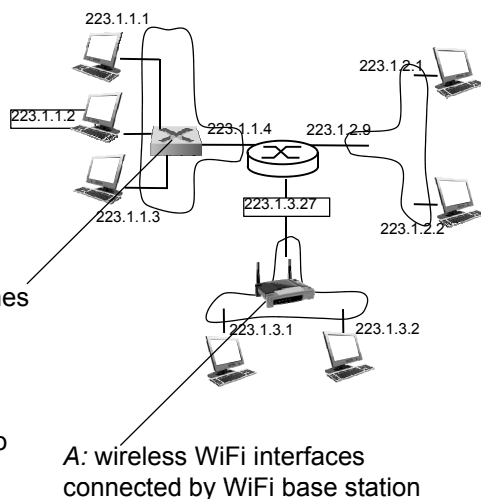
IP addressing: introduction

Q: how are interfaces actually connected?

A: we'll learn about that in chapter 5, 6.

A: wired Ethernet interfaces connected by Ethernet switches

For now: don't need to worry about how one interface is connected to another (with no intervening router)

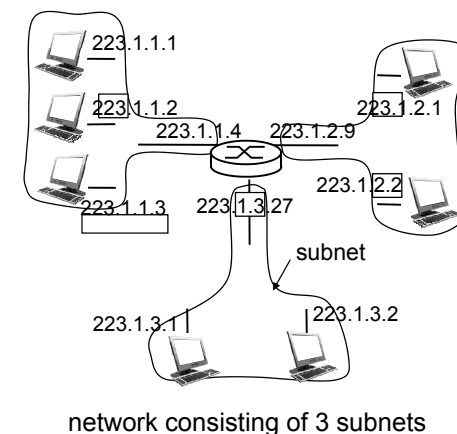


A: wireless WiFi interfaces connected by WiFi base station

Network Layer 4-27

Subnets

- ❖ **IP address:**
 - subnet part - high order bits
 - host part - low order bits
- ❖ **what's a subnet?**
 - device interfaces with same subnet part of IP address
 - can physically reach each other *without* intervening router

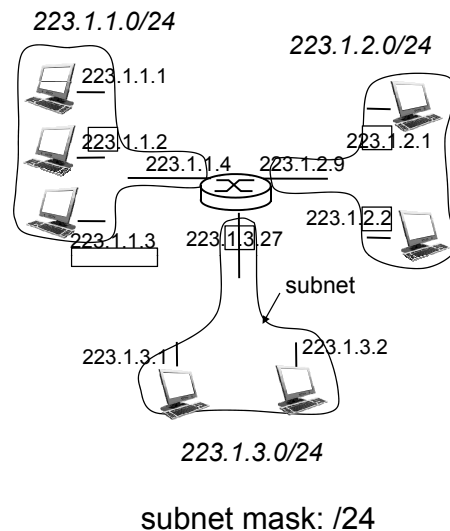


Network Layer 4-28

Subnets

recipe

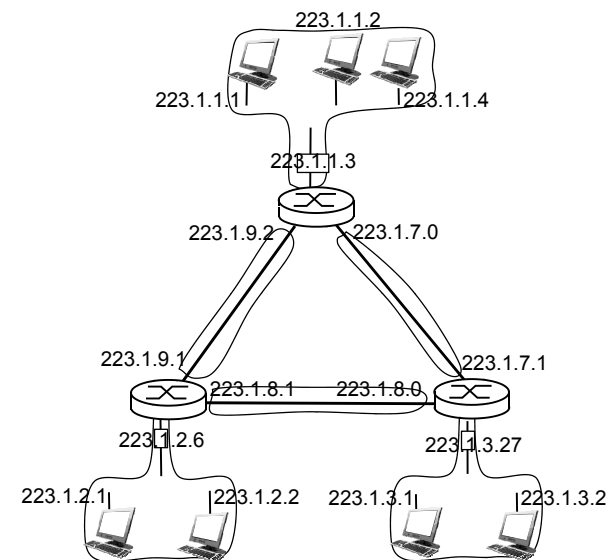
- ❖ to determine the subnets, detach each interface from its host or router, creating islands of isolated networks
- ❖ each isolated network is called a *subnet*



Network Layer 4-29

Subnets

how many?

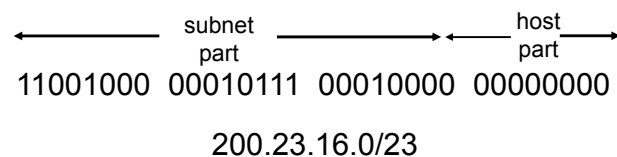


Network Layer 4-30

IP addressing: CIDR

CIDR: Classless InterDomain Routing

- subnet portion of address of arbitrary length
- address format: a.b.c.d/x, where x is # bits in subnet portion of address



Network Layer 4-31

IP addresses: how to get one?

Q: How does a *host* get IP address?

- ❖ hard-coded by system admin in a file
 - Windows: control-panel->network->configuration->tcp/ip->properties
 - UNIX: /etc/rc.config
- ❖ DHCP: Dynamic Host Configuration Protocol: dynamically get address from a server
 - “plug-and-play”

Network Layer 4-32

DHCP: Dynamic Host Configuration Protocol

goal: allow host to *dynamically* obtain its IP address from network server when it joins network

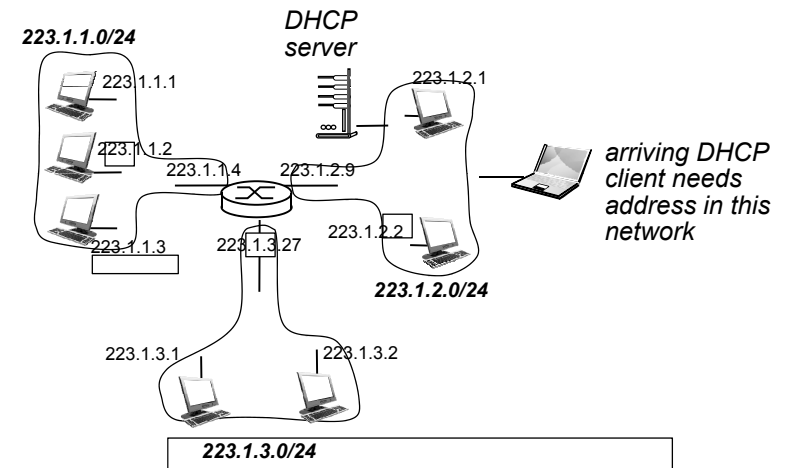
- can renew its lease on address in use
- allows reuse of addresses (only hold address while connected/"on")
- support for mobile users who want to join network (more shortly)

DHCP overview:

- host broadcasts "DHCP discover" msg [optional]
- DHCP server responds with "DHCP offer" msg [optional]
- host requests IP address: "DHCP request" msg
- DHCP server sends address: "DHCP ack" msg

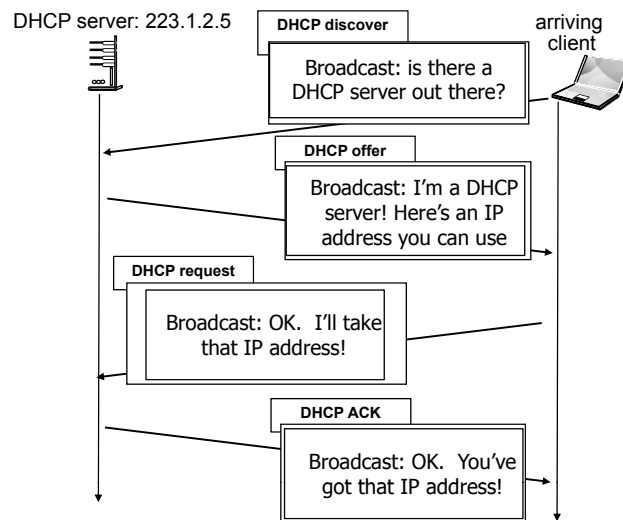
Network Layer 4-33

DHCP client-server scenario



Network Layer 4-34

DHCP client-server scenario



Network Layer 4-35

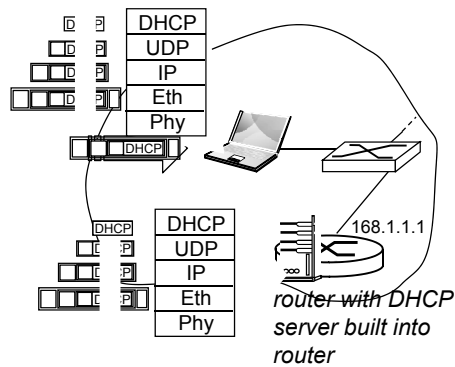
DHCP: more than IP addresses

DHCP can return more than just allocated IP address on subnet:

- address of first-hop router for client
- name and IP address of DNS sever
- network mask (indicating network versus host portion of address)

Network Layer 4-36

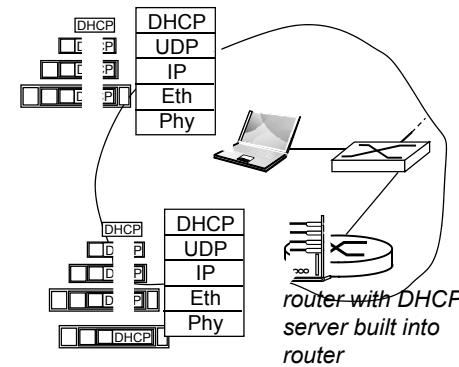
DHCP: example



- ❖ connecting laptop needs its IP address, addr of first-hop router, addr of DNS server: use DHCP
- ❖ DHCP request encapsulated in UDP, encapsulated in IP, encapsulated in 802.1 Ethernet
- ❖ Ethernet frame broadcast (dest: FFFFFFFF) on LAN, received at router running DHCP server
- ❖ Ethernet demuxed to IP demuxed, UDP demuxed to DHCP

Network Layer 4-37

DHCP: example



- ❖ DCP server formulates DHCP ACK containing client's IP address, IP address of first-hop router for client, name & IP address of DNS server
- ❖ encapsulation of DHCP server, frame forwarded to client, demuxing up to DHCP at client
- ❖ client now knows its IP address, name and IP address of DSN server, IP address of its first-hop router

Network Layer 4-38

DHCP: Wireshark output (home LAN)

Message type: **Boot Request (1)**
 Hardware type: Ethernet
 Hardware address length: 6
 Hops: 0
Transaction ID: 0x6b3a11b7
 Seconds elapsed: 0
 Bootp flags: 0x0000 (Unicast)
 Client IP address: 0.0.0.0 (0.0.0.0)
 Your (client) IP address: 0.0.0.0 (0.0.0.0)
 Next server IP address: 0.0.0.0 (0.0.0.0)
 Relay agent IP address: 0.0.0.0 (0.0.0.0)
Client MAC address: Wistron_23:68:8a (00:16:d3:23:68:8a)
 Server host name not given
 Boot file name not given
 Magic cookie: (OK)
 Option: (t=53,i=1) **DHCP Message Type = DHCP Request**
 Option: (61) Client identifier
 Length: 7; Value: 010016D323688A;
 Hardware type: Ethernet
 Client MAC address: Wistron_23:68:8a (00:16:d3:23:68:8a)
 Option: (t=50,i=4) Requested IP Address = 192.168.1.101
 Option: (t=12,i=5) Host Name = "nomad"
Option: (55) Parameter Request List
 Length: 11; Value: 010F03062C2E2F1F21F92B
 1 = Subnet Mask; 15 = Domain Name
 3 = Router; 6 = Domain Name Server
 44 = NetBIOS over TCP/IP Name Server

request

Message type: **Boot Reply (2)**
 Hardware type: Ethernet
 Hardware address length: 6
 Hops: 0
Transaction ID: 0x6b3a11b7
 Seconds elapsed: 0
 Bootp flags: 0x0000 (Unicast)
Client IP address: 192.168.1.101 (192.168.1.101)
 Your (client) IP address: 0.0.0.0 (0.0.0.0)
Next server IP address: 192.168.1.1 (192.168.1.1)
 Relay agent IP address: 0.0.0.0 (0.0.0.0)
 Client MAC address: Wistron_23:68:8a (00:16:d3:23:68:8a)
 Server host name not given
 Boot file name not given
 Magic cookie: (OK)
Option: (t=53,i=1) DHCP Message Type = DHCP ACK
Option: (t=54,i=4) Server Identifier = 192.168.1.1
Option: (t=1,i=4) Subnet Mask = 255.255.255.0
Option: (t=3,i=4) Router = 192.168.1.1
Option: (6) Domain Name Server
 Length: 12; Value: 445747E2445749F244574092;
 IP Address: 68.87.71.226;
 IP Address: 68.87.73.242;
 IP Address: 68.87.64.146
Option: (t=15,i=20) Domain Name = "hsd1.ma.comcast.net."

reply

Network Layer 4-39

IP addresses: how to get one?

Q: how does *network* get subnet part of IP addr?

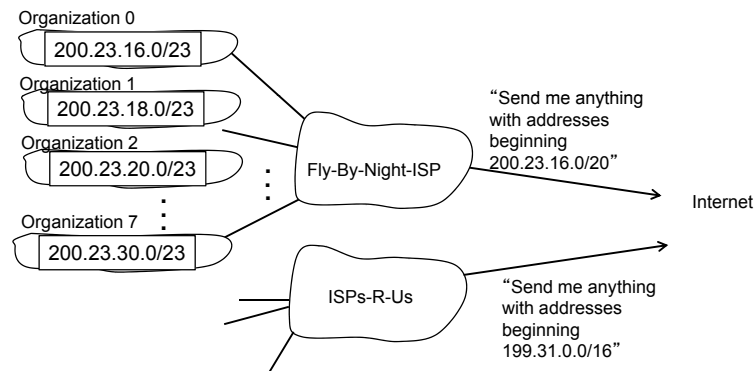
A: gets allocated portion of its provider ISP's address space

ISP's block	11001000 00010111 00010000 00000000	200.23.16.0/20
Organization 0	11001000 00010111 00010000 00000000	200.23.16.0/23
Organization 1	11001000 00010111 00010010 00000000	200.23.18.0/23
Organization 2	11001000 00010111 00010100 00000000	200.23.20.0/23
...		
Organization 7	11001000 00010111 00011110 00000000	200.23.30.0/23

Network Layer 4-40

Hierarchical addressing: route aggregation

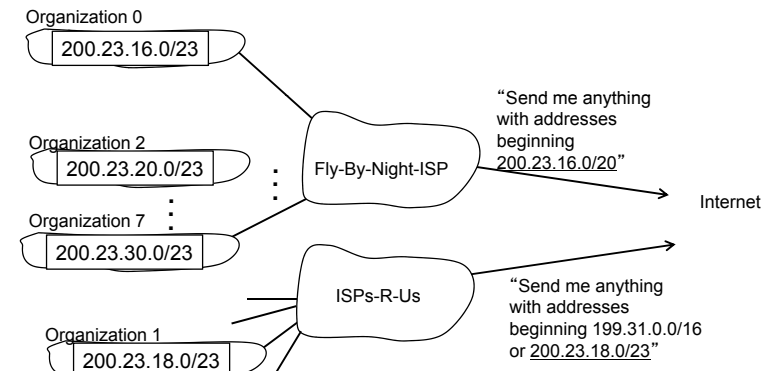
hierarchical addressing allows efficient advertisement of routing information:



Network Layer 4-41

Hierarchical addressing: more specific routes

ISPs-R-U's has a more specific route to Organization 1



Network Layer 4-42

IP addressing: the last word...

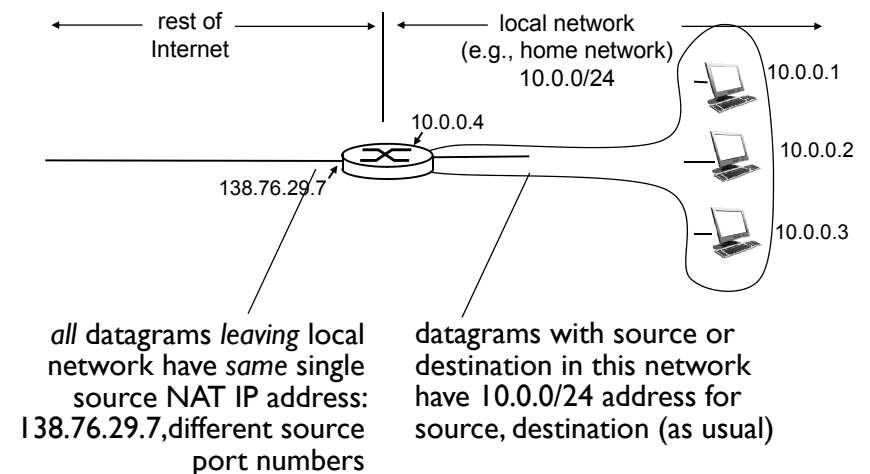
Q: how does an ISP get block of addresses?

A: ICANN: Internet Corporation for Assigned Names and Numbers <http://www.icann.org/>

- allocates addresses
- manages DNS
- assigns domain names, resolves disputes

Network Layer 4-43

NAT: network address translation



Network Layer 4-44

NAT: network address translation

motivation: local network uses just one IP address as far as outside world is concerned:

- range of addresses not needed from ISP: just one IP address for all devices
- can change addresses of devices in local network without notifying outside world
- can change ISP without changing addresses of devices in local network
- devices inside local net not explicitly addressable, visible by outside world (a security plus)

Network Layer 4-45

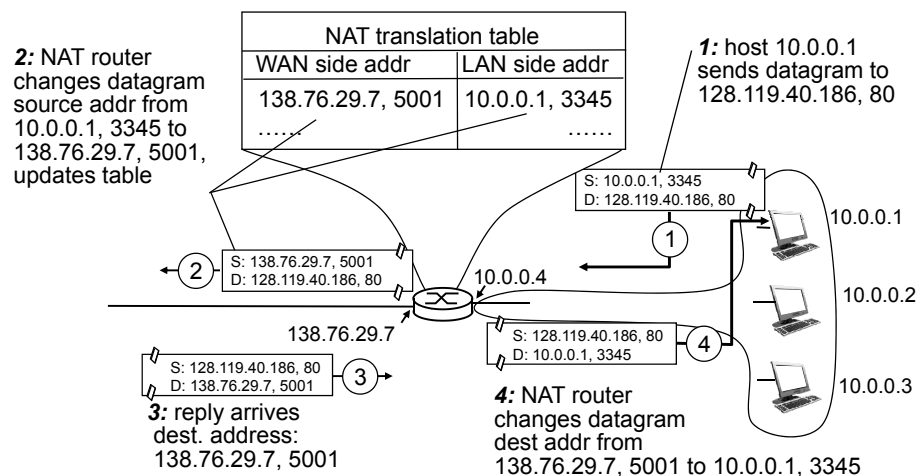
NAT: network address translation

implementation: NAT router must:

- *outgoing datagrams:* replace (source IP address, port #) of every outgoing datagram to (NAT IP address, new port #)
... remote clients/servers will respond using (NAT IP address, new port #) as destination addr
- *remember (in NAT translation table)* every (source IP address, port #) to (NAT IP address, new port #) translation pair
- *incoming datagrams:* replace (NAT IP address, new port #) in dest fields of every incoming datagram with corresponding (source IP address, port #) stored in NAT table

Network Layer 4-46

NAT: network address translation



Network Layer 4-47

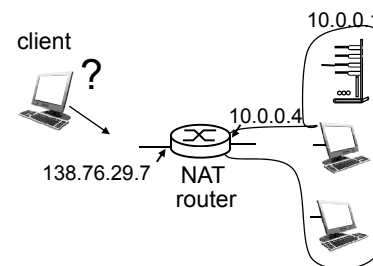
NAT: network address translation

- ❖ 16-bit port-number field:
 - 60,000 simultaneous connections with a single LAN-side address!
- ❖ NAT is controversial:
 - routers should only process up to layer 3
 - violates end-to-end argument
 - NAT possibility must be taken into account by app designers, e.g., P2P applications
 - address shortage should instead be solved by IPv6

Network Layer 4-48

NAT traversal problem

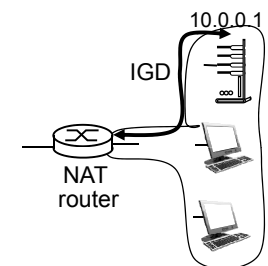
- ❖ client wants to connect to server with address 10.0.0.1
 - server address 10.0.0.1 local to LAN (client can't use it as destination addr)
 - only one externally visible NATed address: 138.76.29.7
- ❖ *solution 1*: statically configure NAT to forward incoming connection requests at given port to server
 - e.g., (138.76.29.7, port 2500) always forwarded to 10.0.0.1 port 25000



Network Layer 4-49

NAT traversal problem

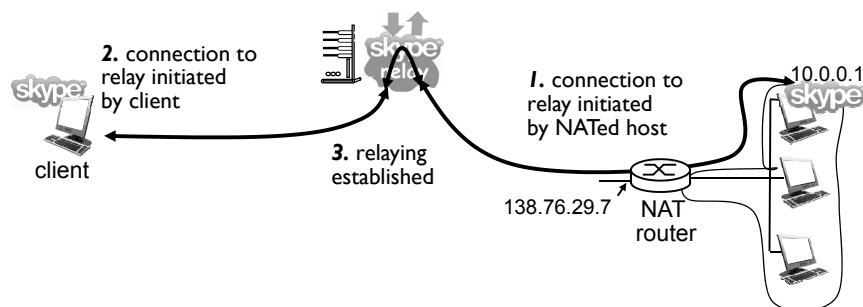
- ❖ *solution 2*: Universal Plug and Play (UPnP) Internet Gateway Device (IGD) Protocol. Allows NATed host to:
 - ❖ learn public IP address (138.76.29.7)
 - ❖ add/remove port mappings (with lease times)
- i.e., automate static NAT port map configuration



Network Layer 4-50

NAT traversal problem

- ❖ *solution 3*: relaying (used in Skype)
 - NATed client establishes connection to relay
 - external client connects to relay
 - relay bridges packets between to connections



Network Layer 4-51

Chapter 4: outline

- 4.1 introduction
- 4.2 virtual circuit and datagram networks
- 4.3 what's inside a router
- 4.4 IP: Internet Protocol
 - datagram format
 - IPv4 addressing
 - ICMP
 - IPv6
- 4.5 routing algorithms
 - link state
 - distance vector
 - hierarchical routing
- 4.6 routing in the Internet
 - RIP
 - OSPF
 - BGP
- 4.7 broadcast and multicast routing

Network Layer 4-52

ICMP: internet control message protocol

- ❖ used by hosts & routers to communicate network-level information

- error reporting: unreachable host, network, port, protocol
- echo request/reply (used by ping)

- ❖ network-layer “above” IP:

- ICMP msgs carried in IP datagrams

- ❖ ICMP message: type, code plus first 8 bytes of IP datagram causing error

Type	Code	description
0	0	echo reply (ping)
3	0	dest. network unreachable
3	1	dest host unreachable
3	2	dest protocol unreachable
3	3	dest port unreachable
3	6	dest network unknown
3	7	dest host unknown
4	0	source quench (congestion control - not used)
8	0	echo request (ping)
9	0	route advertisement
10	0	router discovery
11	0	TTL expired
12	0	bad IP header

Network Layer 4-53

Traceroute and ICMP

- ❖ source sends series of UDP segments to dest

- first set has TTL = 1
- second set has TTL=2, etc.
- unlikely port number

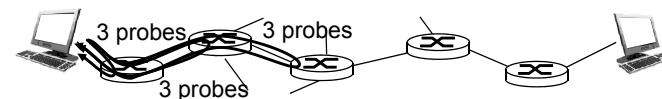
- ❖ when *n*th set of datagrams arrives to *n*th router:

- router discards datagrams
- and sends source ICMP messages (type 11, code 0)
- ICMP messages includes name of router & IP address

- ❖ when ICMP messages arrives, source records RTTs

stopping criteria:

- ❖ UDP segment eventually arrives at destination host
- ❖ destination returns ICMP “port unreachable” message (type 3, code 3)
- ❖ source stops



Network Layer 4-54

IPv6: motivation

- ❖ *initial motivation:* 32-bit address space soon to be completely allocated.
- ❖ additional motivation:
 - header format helps speed processing/forwarding
 - header changes to facilitate QoS

IPv6 datagram format:

- fixed-length 40 byte header
- no fragmentation allowed

Network Layer 4-55

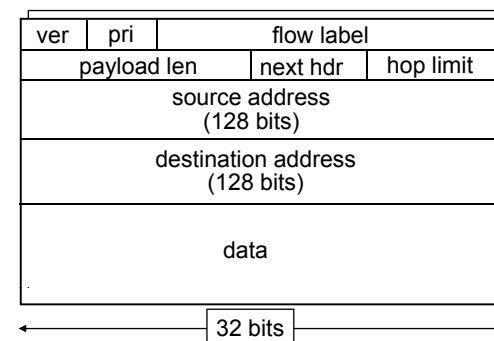
IPv6 datagram format

priority: identify priority among datagrams in flow

flow Label: identify datagrams in same “flow.”

(concept of “flow” not well defined).

next header: identify upper layer protocol for data



Network Layer 4-56

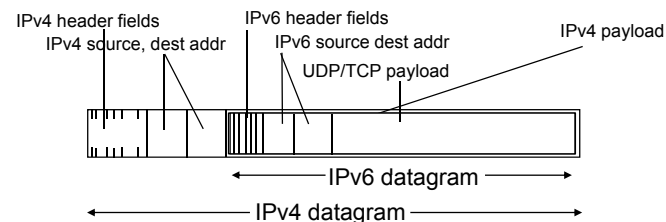
Other changes from IPv4

- ❖ *checksum*: removed entirely to reduce processing time at each hop
- ❖ *options*: allowed, but outside of header, indicated by “Next Header” field
- ❖ *ICMPv6*: new version of ICMP
 - additional message types, e.g. “Packet Too Big”
 - multicast group management functions

Network Layer 4-57

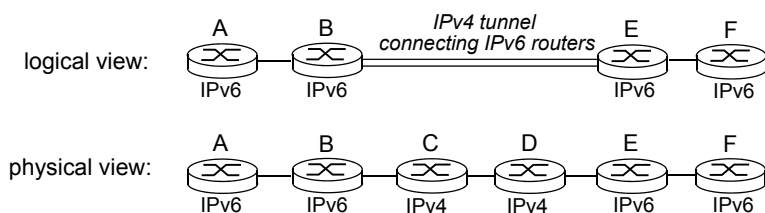
Transition from IPv4 to IPv6

- ❖ not all routers can be upgraded simultaneously
 - no “flag days”
 - how will network operate with mixed IPv4 and IPv6 routers?
- ❖ *tunneling*: IPv6 datagram carried as *payload* in IPv4 datagram among IPv4 routers



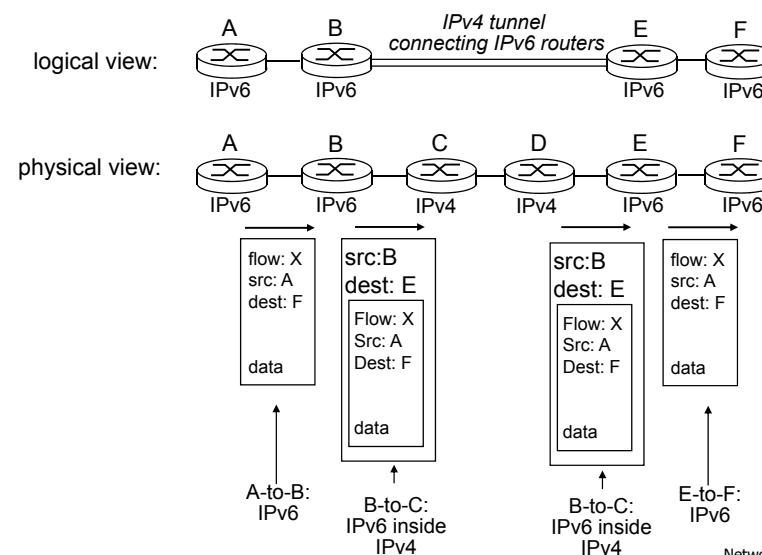
Network Layer 4-58

Tunneling



Network Layer 4-59

Tunneling



Network Layer 4-60

IPv6: adoption

- ❖ US National Institutes of Standards estimate [2013]:
 - ~3% of industry IP routers
 - ~11% of US gov't routers
- ❖ *Long (long!) time for deployment, use*
 - 20 years and counting!
 - think of application-level changes in last 20 years: WWW, Facebook, ...
 - *Why?*

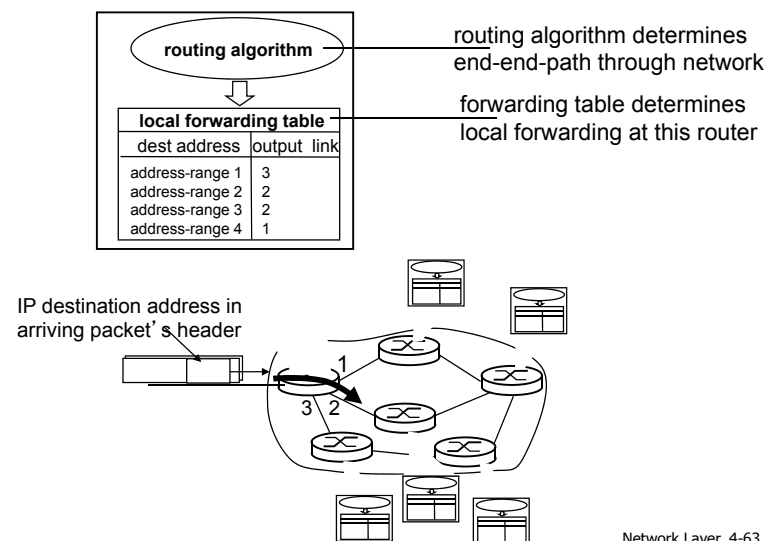
Network Layer 4-61

Chapter 4: outline

- 4.1 introduction
- 4.2 virtual circuit and datagram networks
- 4.3 what's inside a router
- 4.4 IP: Internet Protocol
 - datagram format
 - IPv4 addressing
 - ICMP
 - IPv6
- 4.5 routing algorithms
 - link state
 - distance vector
 - hierarchical routing
- 4.6 routing in the Internet
 - RIP
 - OSPF
 - BGP
- 4.7 broadcast and multicast routing

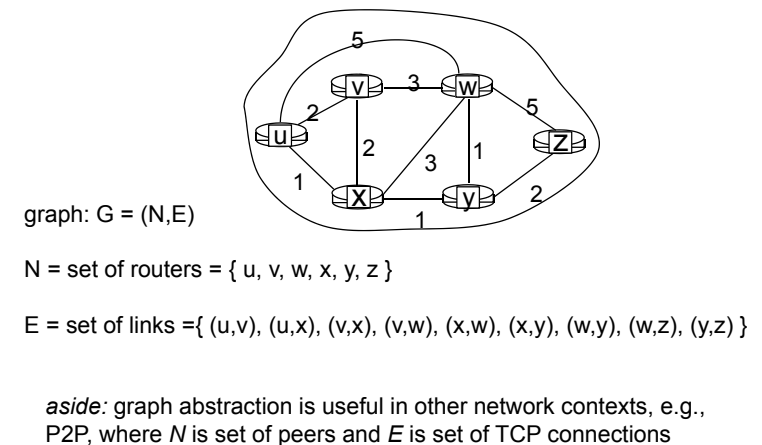
Network Layer 4-62

Interplay between routing, forwarding



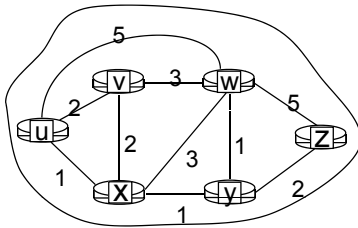
Network Layer 4-63

Graph abstraction



Network Layer 4-64

Graph abstraction: costs



$c(x, x') = \text{cost of link } (x, x')$
e.g., $c(w, z) = 5$

cost could always be 1, or
inversely related to bandwidth,
or inversely related to
congestion

cost of path $(x_1, x_2, x_3, \dots, x_p) = c(x_1, x_2) + c(x_2, x_3) + \dots + c(x_{p-1}, x_p)$

key question: what is the least-cost path between u and z ?

routing algorithm: algorithm that finds that least cost path

Network Layer 4-65

Routing algorithm classification

Q: global or decentralized information?

global:

- ❖ all routers have complete topology, link cost info
- ❖ “link state” algorithms

decentralized:

- ❖ router knows physically-connected neighbors, link costs to neighbors
- ❖ iterative process of computation, exchange of info with neighbors
- ❖ “distance vector” algorithms

Q: static or dynamic?

static:

- ❖ routes change slowly over time

dynamic:

- ❖ routes change more quickly
 - periodic update
 - in response to link cost changes

Network Layer 4-66

Chapter 4: outline

4.1 introduction

4.2 virtual circuit and datagram networks

4.3 what's inside a router

4.4 IP: Internet Protocol

- datagram format
- IPv4 addressing
- ICMP
- IPv6

4.5 routing algorithms

- link state
- distance vector
- hierarchical routing

4.6 routing in the Internet

- RIP
- OSPF
- BGP

4.7 broadcast and multicast routing

A Link-State Routing Algorithm

Dijkstra's algorithm

- ❖ net topology, link costs known to all nodes
 - accomplished via “link state broadcast”
 - all nodes have same info
- ❖ computes least cost paths from one node (“source”) to all other nodes
 - gives *forwarding table* for that node
- ❖ iterative: after k iterations, know least cost path to k dest.'s

notation:

- ❖ $C(x, y)$: link cost from node x to y; $= \infty$ if not direct neighbors
- ❖ $D(v)$: current value of cost of path from source to dest. v
- ❖ $p(v)$: predecessor node along path from source to v
- ❖ N' : set of nodes whose least cost path definitively known

Network Layer 4-67

Network Layer 4-68

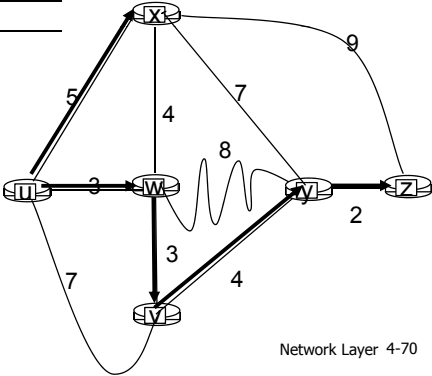
Dijkstra's Algorithm

- 1 **Initialization:**
- 2 $N' = \{u\}$
- 3 for all nodes v
- 4 if v adjacent to u
- 5 then $D(v) = c(u,v)$
- 6 else $D(v) = \infty$
- 7
- 8 **Loop**
- 9 find w not in N' such that $D(w)$ is a minimum
- 10 add w to N'
- 11 update $D(v)$ for all v adjacent to w and not in N' :
- 12 $D(v) = \min(D(v), D(w) + c(w,v))$
- 13 /* new cost to v is either old cost to v or known
- 14 shortest path cost to w plus cost from w to v */
- 15 **until all nodes in N'**

Dijkstra's algorithm: example

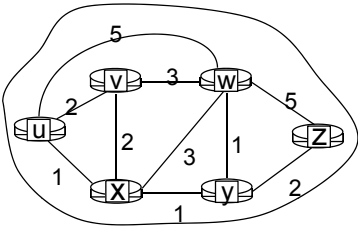
Step	N'	D(v) p(v)	D(w) p(w)	D(x) p(x)	D(y) p(y)	D(z) p(z)
0	u	7,u	(3,u)	5,u	∞	∞
1	uw	6,w		(5,u)	11,w	∞
2	uwx		(6,w)		11,w	14,x
3	uwxv				(10,y)	14,x
4	uwxvy					(12,y)
5	uwxvyz					

- notes:
- ❖ construct shortest path tree by tracing predecessor nodes
 - ❖ ties can exist (can be broken arbitrarily)



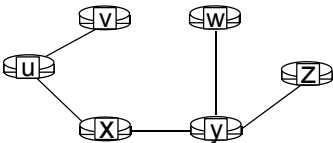
Dijkstra's algorithm: another example

Step	N'	D(v),p(v)	D(w),p(w)	D(x),p(x)	D(y),p(y)	D(z),p(z)
0	u	2,u	5,u	1,u	∞	∞
1	ux	2,u	4,x		2,x	∞
2	uxy	2,u	3,y			4,y
3	uxyv		3,y			4,y
4	uxyvw					4,y
5	uxyvwz					



Dijkstra's algorithm: example (2)

resulting shortest-path tree from u:



resulting forwarding table in u:

destination	link
v	(u,v)
x	(u,x)
y	(u,x)
w	(u,x)
z	(u,x)

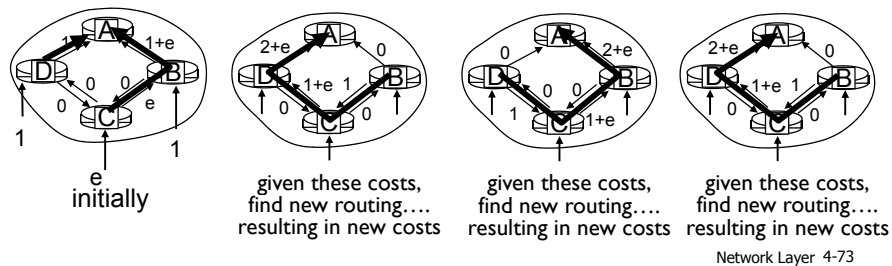
Dijkstra's algorithm, discussion

algorithm complexity: n nodes

- ❖ each iteration: need to check all nodes, w, not in N
- ❖ $n(n+1)/2$ comparisons: $O(n^2)$
- ❖ more efficient implementations possible: $O(n \log n)$

oscillations possible:

- ❖ e.g., support link cost equals amount of carried traffic:



Chapter 4: outline

4.1 introduction

4.2 virtual circuit and datagram networks

4.3 what's inside a router

4.4 IP: Internet Protocol

- datagram format
- IPv4 addressing
- ICMP
- IPv6

4.5 routing algorithms

- link state
- distance vector
- hierarchical routing

4.6 routing in the Internet

- RIP
- OSPF
- BGP

4.7 broadcast and multicast routing

Network Layer 4-74

Distance vector algorithm

Bellman-Ford equation (dynamic programming)

let

$d_x(y) :=$ cost of least-cost path from x to y

then

$$d_x(y) = \min_v \{ c(x,v) + d_v(y) \}$$

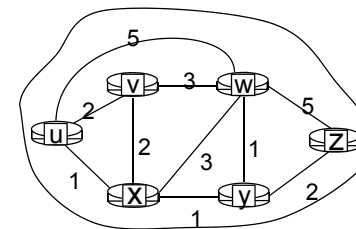
min taken over all neighbors v of x

cost to neighbor v

cost from neighbor v to destination y

Network Layer 4-75

Bellman-Ford example



clearly, $d_v(z) = 5$, $d_x(z) = 3$, $d_w(z) = 3$

B-F equation says:

$$\begin{aligned} d_u(z) &= \min \{ c(u,v) + d_v(z), \\ &\quad c(u,x) + d_x(z), \\ &\quad c(u,w) + d_w(z) \} \\ &= \min \{ 2 + 5, \\ &\quad 1 + 3, \\ &\quad 5 + 3 \} = 4 \end{aligned}$$

node achieving minimum is next hop in shortest path, used in forwarding table

Network Layer 4-76

Distance vector algorithm

- ❖ $D_x(y)$ = estimate of least cost from x to y
 - x maintains distance vector $\mathbf{D}_x = [D_x(y): y \in N]$
- ❖ node x:
 - knows cost to each neighbor v: $c(x,v)$
 - maintains its neighbors' distance vectors. For each neighbor v, x maintains $\mathbf{D}_v = [D_v(y): y \in N]$

Network Layer 4-77

Distance vector algorithm

key idea:

- ❖ from time-to-time, each node sends its own distance vector estimate to neighbors
- ❖ when x receives new DV estimate from neighbor, it updates its own DV using B-F equation:

$$D_x(y) \leftarrow \min_v \{c(x,v) + D_v(y)\} \text{ for each node } y \in N$$

- ❖ under minor, natural conditions, the estimate $D_x(y)$ converge to the actual least cost $d_x(y)$

Network Layer 4-78

Distance vector algorithm

iterative, asynchronous:

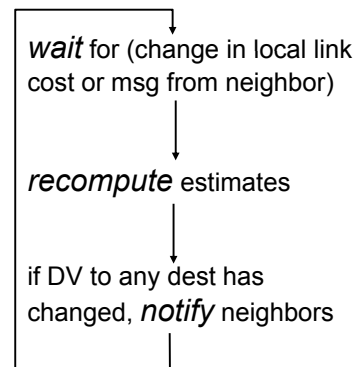
each local iteration caused by:

- ❖ local link cost change
- ❖ DV update message from neighbor

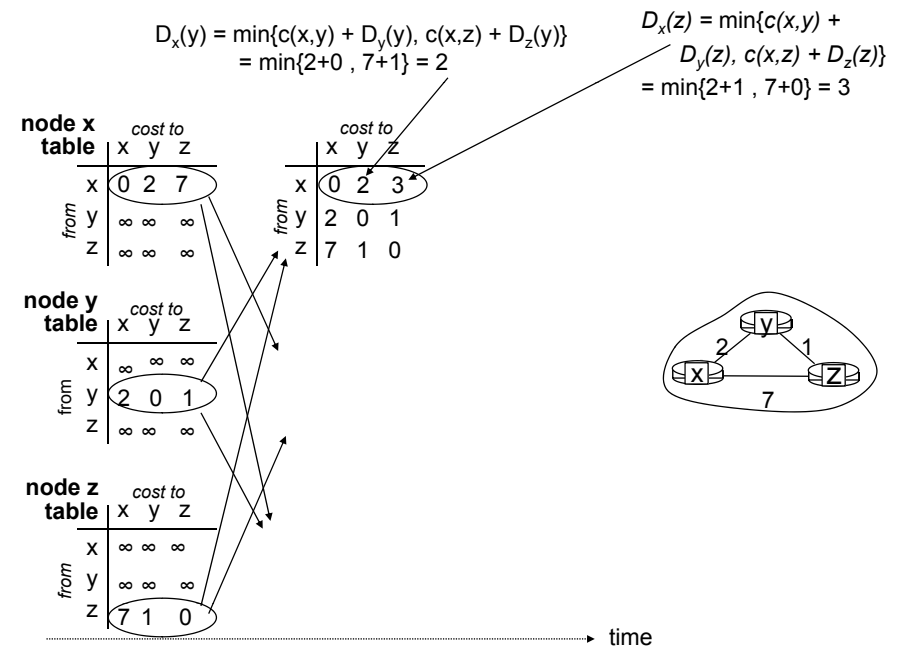
distributed:

- ❖ each node notifies neighbors *only* when its DV changes
 - neighbors then notify their neighbors if necessary

each node:



Network Layer 4-79

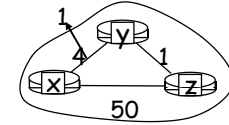


Network Layer 4-80

Distance vector: link cost changes

link cost changes:

- ❖ node detects local link cost change
- ❖ updates routing info, recalculates distance vector
- ❖ if DV changes, notify neighbors

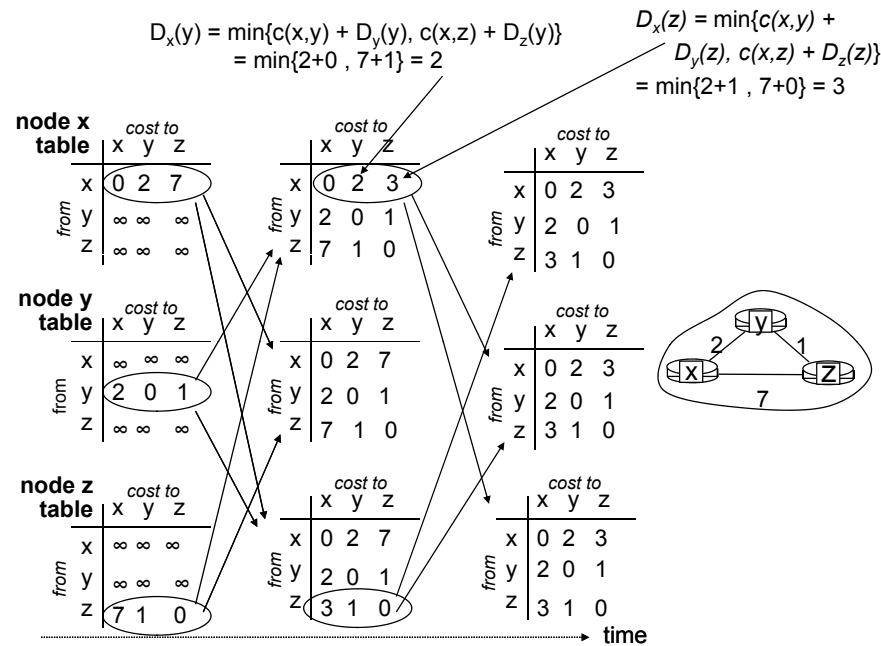


“good news travels fast”

t_0 : y detects link-cost change, updates its DV, informs its neighbors.

t_1 : z receives update from y, updates its table, computes new least cost to x, sends its neighbors its DV.

t_2 : y receives z's update, updates its distance table. y's least costs do *not* change, so y does *not* send a message to z.



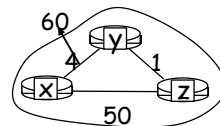
Network Layer 4-81

Network Layer 4-82

Distance vector: link cost changes

link cost changes:

- ❖ node detects local link cost change
- ❖ *bad news travels slow* - “count to infinity” problem!
- ❖ 44 iterations before algorithm stabilizes: see text



poisoned reverse:

- ❖ If Z routes through Y to get to X :
 - Z tells Y its (Z's) distance to X is infinite (so Y won't route to X via Z)
- ❖ will this completely solve count to infinity problem?

Network Layer 4-83

Comparison of LS and DV algorithms

message complexity

- ❖ **LS**: with n nodes, E links, $O(nE)$ msgs sent
- ❖ **DV**: exchange between neighbors only
 - convergence time varies

speed of convergence

- ❖ **LS**: $O(n^2)$ algorithm requires $O(nE)$ msgs
 - may have oscillations
- ❖ **DV**: convergence time varies
 - may be routing loops
 - count-to-infinity problem

robustness: what happens if router malfunctions?

LS:

- node can advertise incorrect link cost
- each node computes only its own table

DV:

- DV node can advertise incorrect *path* cost
- each node's table used by others
 - error propagate thru network

Network Layer 4-84

Chapter 4: outline

- 4.1 introduction
- 4.2 virtual circuit and datagram networks
- 4.3 what's inside a router
- 4.4 IP: Internet Protocol
 - datagram format
 - IPv4 addressing
 - ICMP
 - IPv6
- 4.5 routing algorithms
 - link state
 - distance vector
 - hierarchical routing
- 4.6 routing in the Internet
 - RIP
 - OSPF
 - BGP
- 4.7 broadcast and multicast routing

Network Layer 4-85

Hierarchical routing

our routing study thus far - idealization

- ❖ all routers identical
- ❖ network “flat”
- ... *not* true in practice

scale: with 600 million destinations:

- ❖ can't store all dest's in routing tables!
- ❖ routing table exchange would swamp links!

administrative autonomy

- ❖ internet = network of networks
- ❖ each network admin may want to control routing in its own network

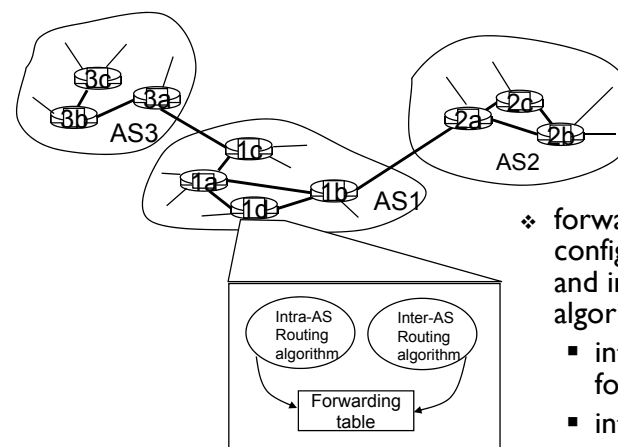
Network Layer 4-86

Hierarchical routing

- ❖ aggregate routers into regions, “autonomous systems” (AS)
 - ❖ routers in same AS run same routing protocol
 - “intra-AS” routing protocol
 - routers in different AS can run different intra-AS routing protocol
- gateway router:*
- ❖ at “edge” of its own AS
 - ❖ has link to router in another AS

Network Layer 4-87

Interconnected ASes



- ❖ forwarding table configured by both intra- and inter-AS routing algorithm
 - intra-AS sets entries for internal dests
 - inter-AS & intra-AS sets entries for external dests

Network Layer 4-88

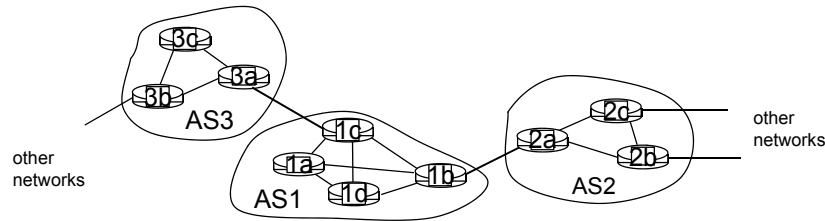
Inter-AS tasks

- ❖ suppose router in AS1 receives datagram destined outside of AS1:
 - router should forward packet to gateway router, but which one?

AS1 must:

1. learn which dests are reachable through AS2, which through AS3
2. propagate this reachability info to all routers in AS1

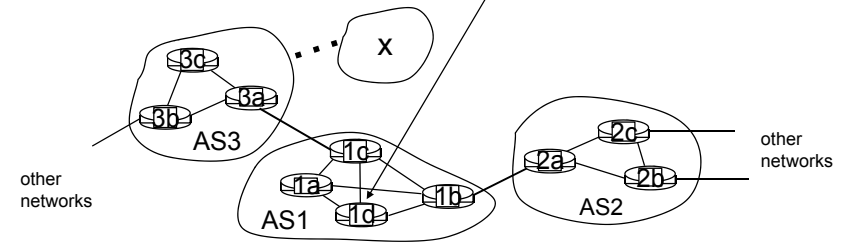
job of inter-AS routing!



Network Layer 4-89

Example: setting forwarding table in router 1d

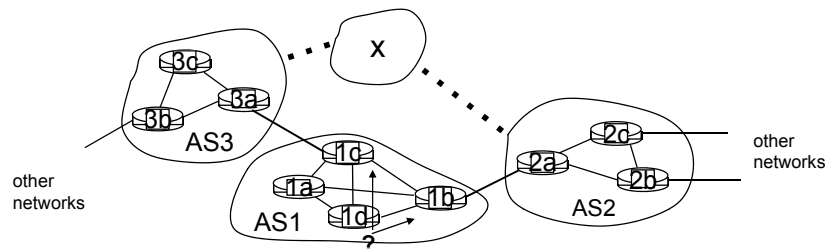
- ❖ suppose AS1 learns (via inter-AS protocol) that subnet x is reachable via AS3 (gateway 1c), but not via AS2
 - inter-AS protocol propagates reachability info to all internal routers
- ❖ router 1d determines from intra-AS routing info that its interface 1 is on the least cost path to 1c
 - installs forwarding table entry (x, 1)



Network Layer 4-90

Example: choosing among multiple ASes

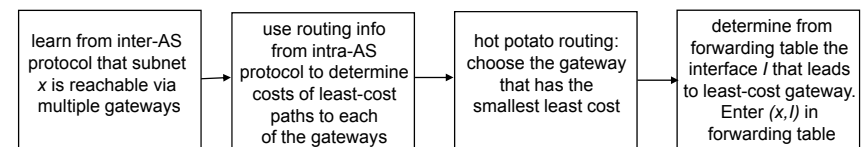
- ❖ now suppose AS1 learns from inter-AS protocol that subnet x is reachable from AS3 and from AS2.
- ❖ to configure forwarding table, router 1d must determine which gateway it should forward packets towards for dest x
 - this is also job of inter-AS routing protocol!



Network Layer 4-91

Example: choosing among multiple ASes

- ❖ now suppose AS1 learns from inter-AS protocol that subnet x is reachable from AS3 and from AS2.
- ❖ to configure forwarding table, router 1d must determine towards which gateway it should forward packets for dest x
 - this is also job of inter-AS routing protocol!
- ❖ *hot potato routing*: send packet towards closest of two routers.



Network Layer 4-92

Chapter 4: outline

- 4.1 introduction
- 4.2 virtual circuit and datagram networks
- 4.3 what's inside a router
- 4.4 IP: Internet Protocol
 - datagram format
 - IPv4 addressing
 - ICMP
 - IPv6
- 4.5 routing algorithms
 - link state
 - distance vector
 - hierarchical routing
- 4.6 routing in the Internet
 - RIP
 - OSPF
 - BGP
- 4.7 broadcast and multicast routing

Network Layer 4-93

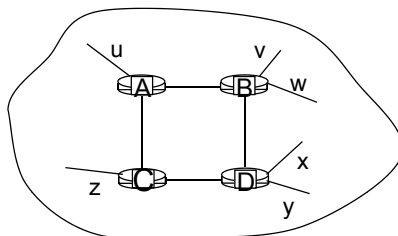
Intra-AS Routing

- ❖ also known as *interior gateway protocols (IGP)*
- ❖ most common intra-AS routing protocols:
 - RIP: Routing Information Protocol
 - OSPF: Open Shortest Path First
 - IGRP: Interior Gateway Routing Protocol (Cisco proprietary)

Network Layer 4-94

RIP (Routing Information Protocol)

- ❖ included in BSD-UNIX distribution in 1982
- ❖ distance vector algorithm
 - distance metric: # hops (max = 15 hops), each link has cost 1
 - DVs exchanged with neighbors every 30 sec in response message (aka advertisement)
 - each advertisement: list of up to 25 destination *subnets* (in IP addressing sense)

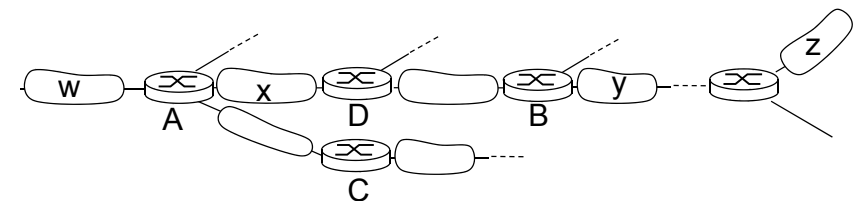


from router A to destination subnets:

subnet	hops
u	1
v	2
w	2
x	3
y	3
z	2

Network Layer 4-95

RIP: example

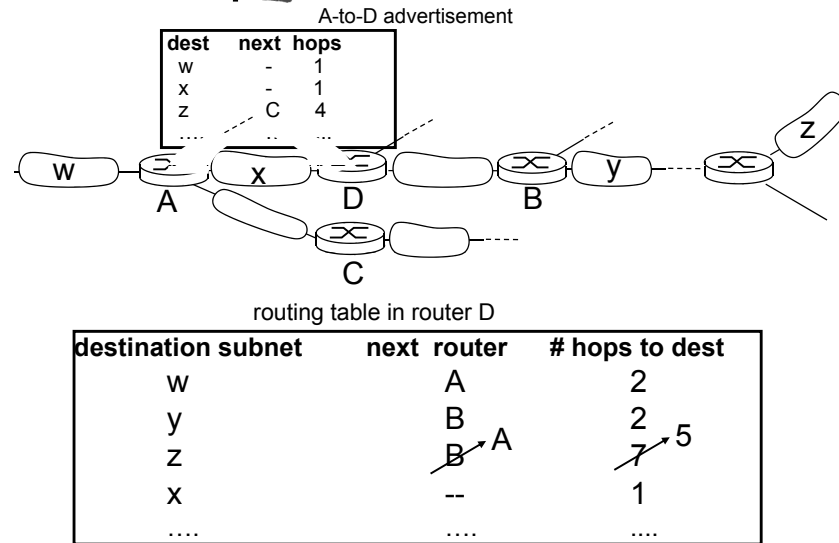


routing table in router D

destination subnet	next router	# hops to dest
w	A	2
y	B	2
x	--	1
....		

Network Layer 4-96

RIP: example



Network Layer 4-97

RIP: link failure, recovery

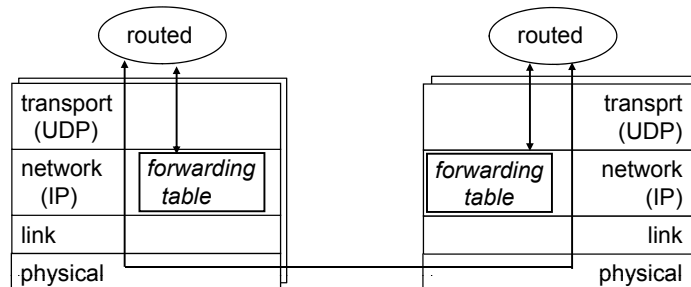
if no advertisement heard after 180 sec -->
neighbor/link declared dead

- routes via neighbor invalidated
- new advertisements sent to neighbors
- neighbors in turn send out new advertisements (if tables changed)
- link failure info quickly (?) propagates to entire net
- *poison reverse* used to prevent ping-pong loops (infinite distance = 16 hops)

Network Layer 4-98

RIP table processing

- ❖ RIP routing tables managed by *application-level* process called route-d (daemon)
- ❖ advertisements sent in UDP packets, periodically repeated



Network Layer 4-99

OSPF (Open Shortest Path First)

- ❖ “open”: publicly available
- ❖ uses link state algorithm
 - LS packet dissemination
 - topology map at each node
 - route computation using Dijkstra’s algorithm
- ❖ OSPF advertisement carries one entry per neighbor
- ❖ advertisements flooded to *entire AS*
 - carried in OSPF messages directly over IP (rather than TCP or UDP)
- ❖ *IS-IS* routing protocol: nearly identical to OSPF

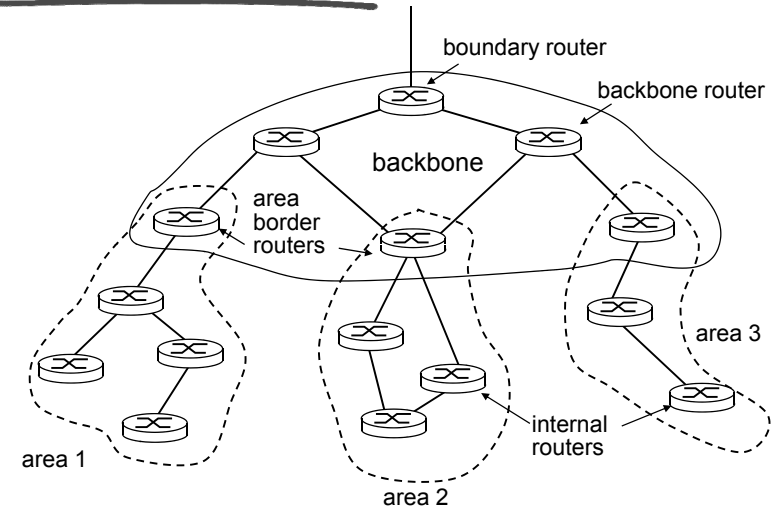
Network Layer 4-100

OSPF “advanced” features (not in RIP)

- ❖ *security*: all OSPF messages authenticated (to prevent malicious intrusion)
- ❖ multiple same-cost paths allowed (only one path in RIP)
- ❖ for each link, multiple cost metrics for different TOS (e.g., satellite link cost set “low” for best effort ToS; high for real time ToS)
- ❖ integrated uni- and multicast support:
 - Multicast OSPF (MOSPF) uses same topology data base as OSPF
- ❖ hierarchical OSPF in large domains.

Network Layer 4-101

Hierarchical OSPF



Network Layer 4-102

Hierarchical OSPF

- ❖ *two-level hierarchy*: local area, backbone.
 - link-state advertisements only in area
 - each nodes has detailed area topology; only know direction (shortest path) to nets in other areas.
- ❖ *area border routers*: “summarize” distances to nets in own area, advertise to other Area Border routers.
- ❖ *backbone routers*: run OSPF routing limited to backbone.
- ❖ *boundary routers*: connect to other AS’ s.

Network Layer 4-103

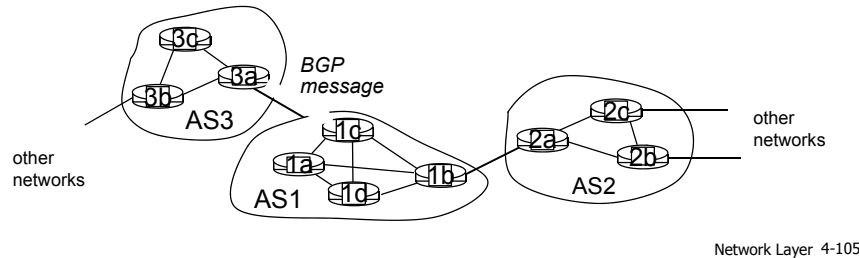
Internet inter-AS routing: BGP

- ❖ BGP (Border Gateway Protocol): *the de facto* inter-domain routing protocol
 - “glue that holds the Internet together”
- ❖ BGP provides each AS a means to:
 - eBGP: obtain subnet reachability information from neighboring ASs.
 - iBGP: propagate reachability information to all AS-internal routers.
 - determine “good” routes to other networks based on reachability information and policy.
- ❖ allows subnet to advertise its existence to rest of Internet: “*I am here*”

Network Layer 4-104

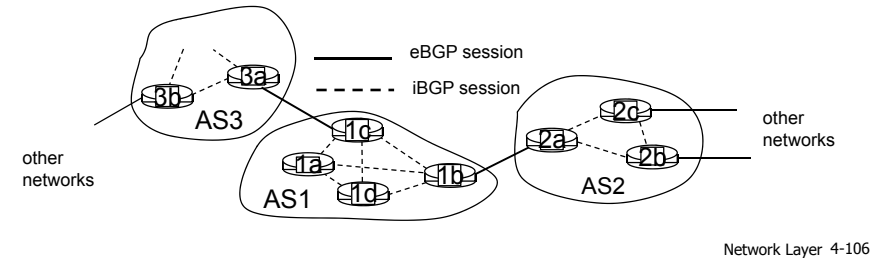
BGP basics

- ❖ BGP session: two BGP routers (“peers”) exchange BGP messages:
 - advertising *paths* to different destination network prefixes (“path vector” protocol)
 - exchanged over semi-permanent TCP connections
- ❖ when AS3 advertises a prefix to AS1:
 - AS3 *promises* it will forward datagrams towards that prefix
 - AS3 can aggregate prefixes in its advertisement



BGP basics: distributing path information

- ❖ using eBGP session between 3a and 1c, AS3 sends prefix reachability info to AS1.
 - 1c can then use iBGP to distribute new prefix info to all routers in AS1
 - 1b can then re-advertise new reachability info to AS2 over 1b-to-2a eBGP session
- ❖ when router learns of new prefix, it creates entry for prefix in its forwarding table.



Path attributes and BGP routes

- ❖ advertised prefix includes BGP attributes
 - prefix + attributes = “route”
- ❖ two important attributes:
 - AS-PATH: contains ASs through which prefix advertisement has passed: e.g., AS 67, AS 17
 - NEXT-HOP: indicates specific internal-AS router to next-hop AS. (may be multiple links from current AS to next-hop-AS)
- ❖ gateway router receiving route advertisement uses import policy to accept/decline
 - e.g., never route through AS x
 - *policy-based* routing

BGP route selection

- ❖ router may learn about more than 1 route to destination AS, selects route based on:
 1. local preference value attribute: policy decision
 2. shortest AS-PATH
 3. closest NEXT-HOP router: hot potato routing
 4. additional criteria

BGP messages

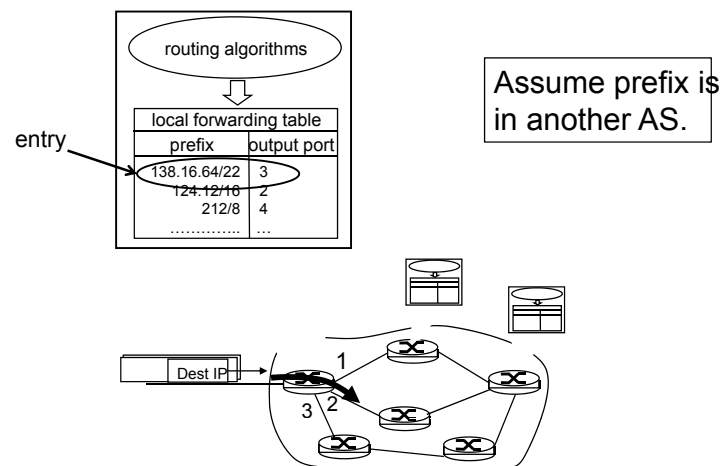
- ❖ BGP messages exchanged between peers over TCP connection
- ❖ BGP messages:
 - OPEN: opens TCP connection to peer and authenticates sender
 - UPDATE: advertises new path (or withdraws old)
 - KEEPALIVE: keeps connection alive in absence of UPDATES; also ACKs OPEN request
 - NOTIFICATION: reports errors in previous msg; also used to close connection

Network Layer 4-109

Putting it Altogether: *How Does an Entry Get Into a Router 's Forwarding Table?*

- ❖ Answer is complicated!
- ❖ Ties together hierarchical routing (Section 4.5.3) with BGP (4.6.3) and OSPF (4.6.2).
- ❖ Provides nice overview of BGP!

How does entry get in forwarding table?

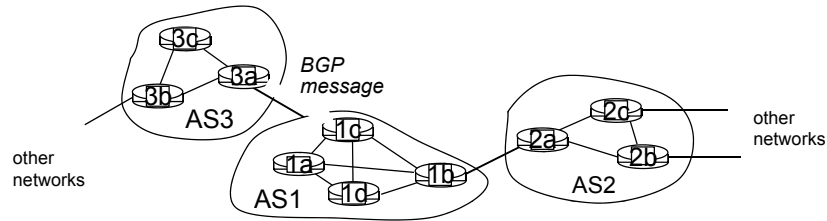


How does entry get in forwarding table?

High-level overview

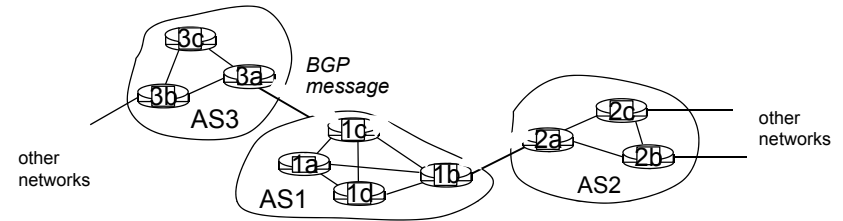
1. Router becomes aware of prefix
2. Router determines output port for prefix
3. Router enters prefix-port in forwarding table

Router becomes aware of prefix



- ❖ BGP message contains “routes”
- ❖ “route” is a prefix and attributes: AS-PATH, NEXT-HOP,...
- ❖ Example: route:
 - ❖ Prefix: 138.16.64/22 ; AS-PATH: AS3 AS131 ;
 - ❖ NEXT-HOP: 201.44.13.125

Router may receive multiple routes



- ❖ Router may receive multiple routes for same prefix
- ❖ Has to select one route

Select best BGP route to prefix

- ❖ Router selects route based on shortest AS-PATH

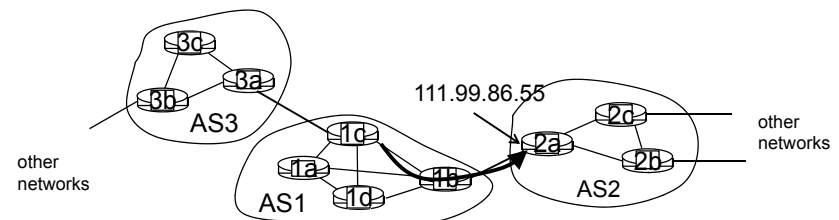
- ❖ Example:

❖ AS2 AS17 to 138.16.64/22 select
 ❖ AS3 AS131 AS201 to 138.16.64/22

- ❖ What if there is a tie? We'll come back to that!

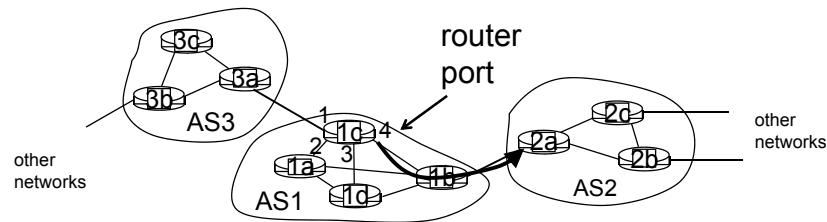
Find best intra-route to BGP route

- ❖ Use selected route's NEXT-HOP attribute
 - Route's NEXT-HOP attribute is the IP address of the router interface that begins the AS PATH.
- ❖ Example:
 - ❖ AS-PATH: AS2 AS17 ; NEXT-HOP: 111.99.86.55
- ❖ Router uses OSPF to find shortest path from 1c to 111.99.86.55



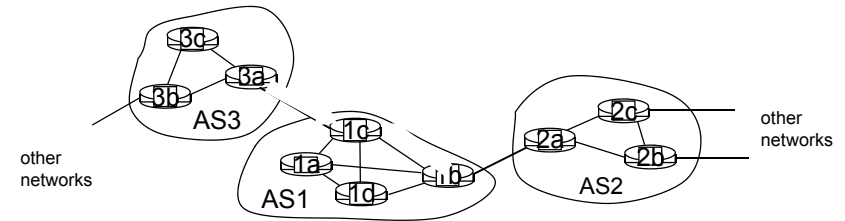
Router identifies port for route

- ❖ Identifies port along the OSPF shortest path
- ❖ Adds prefix-port entry to its forwarding table:
 - (138.16.64/22 , port 4)



Hot Potato Routing

- ❖ Suppose there two or more best inter-routes.
- ❖ Then choose route with closest NEXT-HOP
 - Use OSPF to determine which gateway is closest
 - Q: From 1c, chose AS3 AS131 or AS2 AS17?
 - A: route AS3 AS201 since it is closer

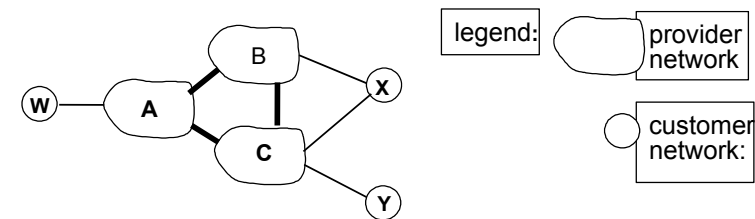


How does entry get in forwarding table?

Summary

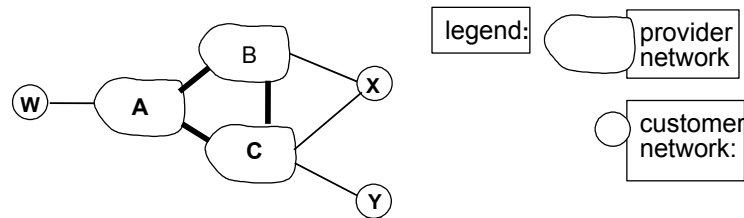
1. Router becomes aware of prefix
 - via BGP route advertisements from other routers
2. Determine router output port for prefix
 - Use BGP route selection to find best inter-AS route
 - Use OSPF to find best intra-AS route leading to best inter-AS route
 - Router identifies router port for that best route
3. Enter prefix-port entry in forwarding table

BGP routing policy



- ❖ A,B,C are provider networks
- ❖ X,W,Y are customer (of provider networks)
- ❖ X is *dual-homed*: attached to two networks
 - X does not want to route from B via X to C
 - .. so X will not advertise to B a route to C

BGP routing policy (2)



- ❖ A advertises path AW to B
- ❖ B advertises path BAW to X
- ❖ Should B advertise path BAW to C?
 - No way! B gets no “revenue” for routing CBAW since neither W nor C are B’s customers
 - B wants to force C to route to w via A
 - B wants to route *only* to/from its customers!

Why different Intra-, Inter-AS routing ?

policy:

- ❖ inter-AS: admin wants control over how its traffic routed, who routes through its net.
- ❖ intra-AS: single admin, so no policy decisions needed

scale:

- ❖ hierarchical routing saves table size, reduced update traffic

performance:

- ❖ intra-AS: can focus on performance
- ❖ inter-AS: policy may dominate over performance