南京大學
NANJING UNIVERSITY

网络层: 数据平面

殷亚凤
智能软件与工程学院
苏州校区南雍楼东区225
yafeng@nju.edu.cn，https://yafengnju.github.io/

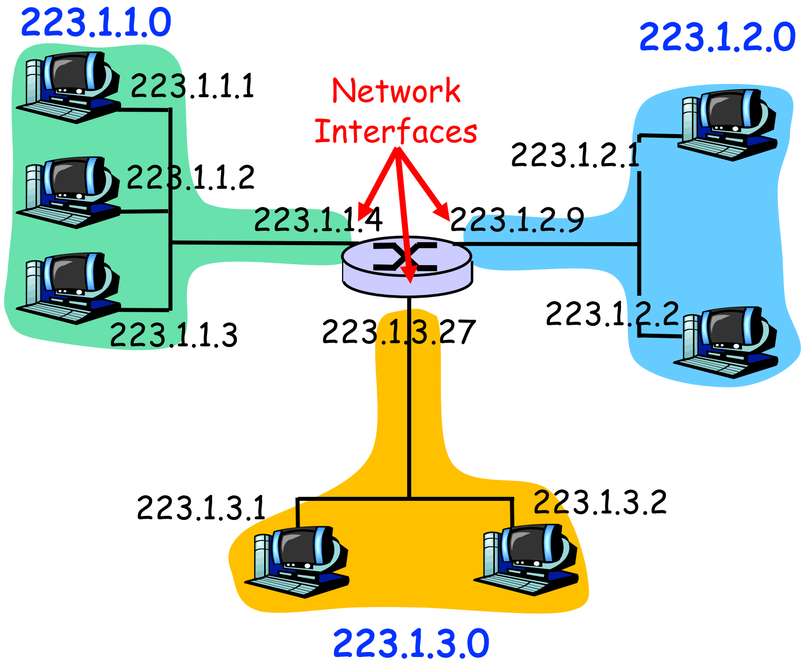
Outline

- IP Addressing
- Network Address Translation
- IPv6
- Generalized Forwarding and SDN
- Middleboxes

IP Addressing

- IP address
  - 32 bit global internet address for each interface
  - Network part (high order bits)
  - Host part (low order bits)

- Physical network (from IP perspective)
  - Can reach each other without intervening router

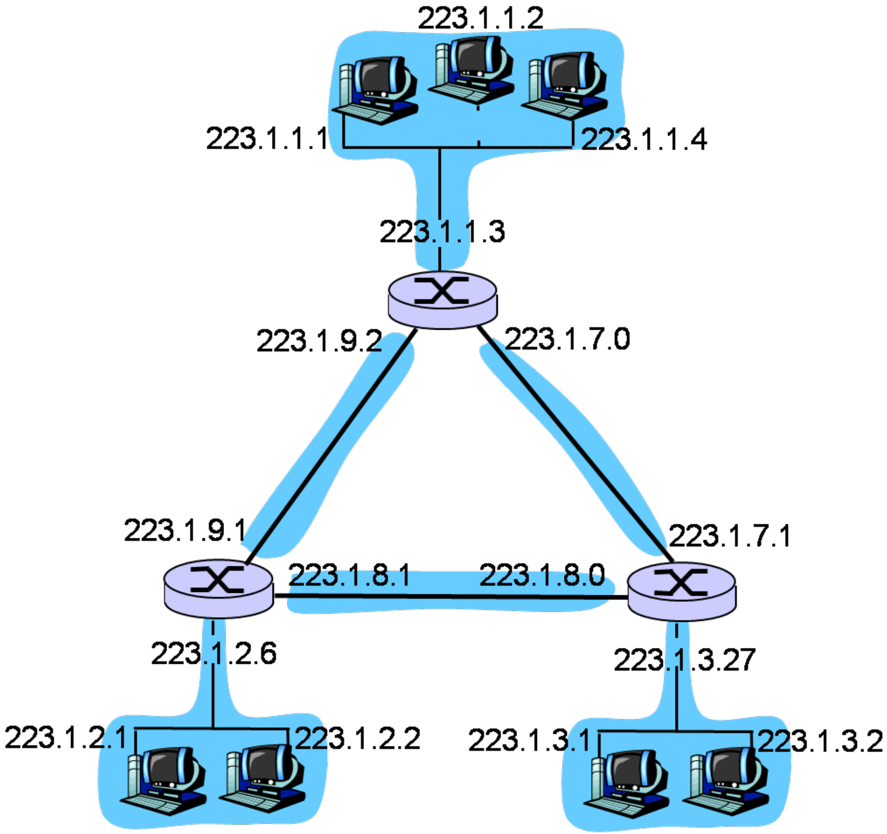


223.1.1.1 = 11011111 00000001 00000001 00000001

223	1	1	1

Count the Physical Networks

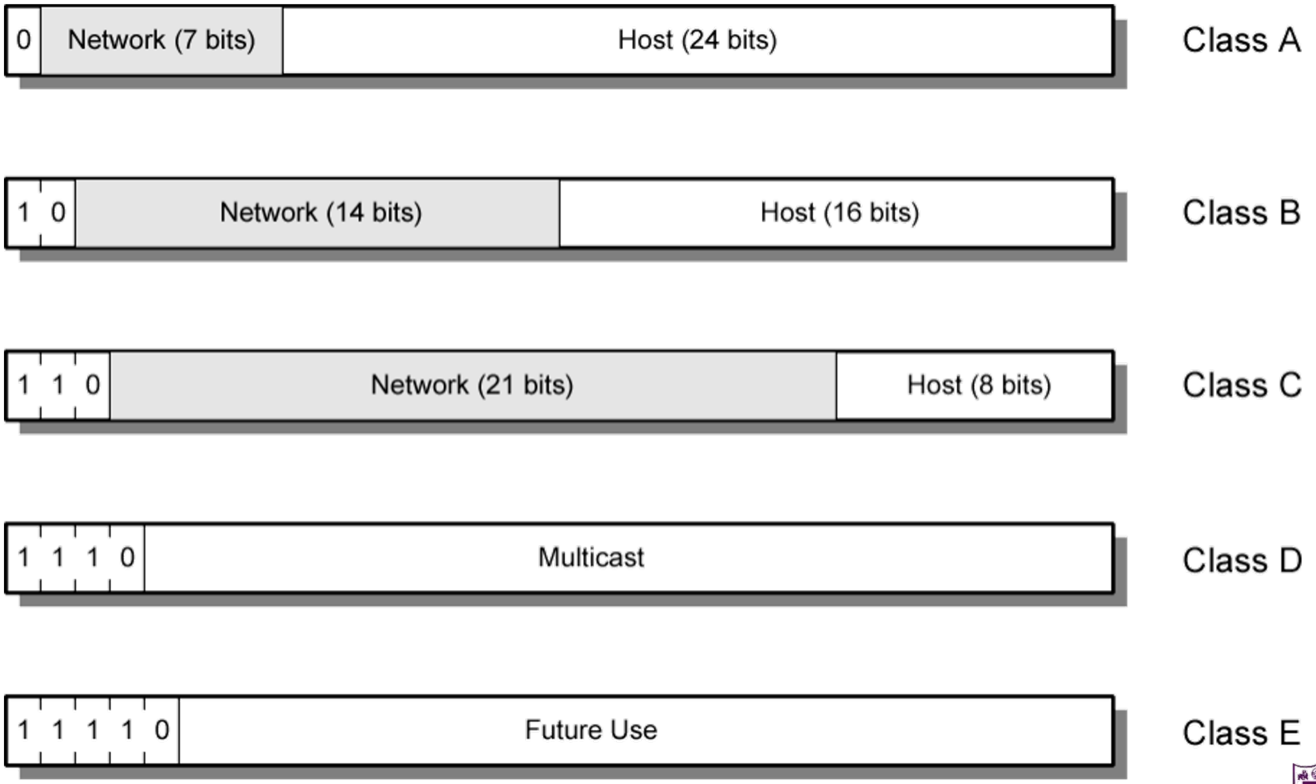
- How many ?



IP Address

- A separate address is required for each physical interface of a host/router to a network
  - Facilitates routing
- Use Dotted-Decimal Notation
- netid unique & administered by
  - American Registry for Internet Numbers (ARIN)
  - Reseaux IP Europeens (RIPE)
  - Asia Pacific Network Information Centre (APNIC)
- hostid assigned within designated organization

IPv4 Address Formats



IP Addresses - Class A

Class A

- Start with binary 0
- Reserved netid
  - All 0 reserved
  - 01111111 (127) reserved for loopback
- Range 1.x.x.x to 126.x.x.x
- Up to 16 million hosts
- All allocated

A类地址：
- 首位为0；
- 支持2^7-2=126个网段；
- 每个网段支持主机数为2^{24}-2 =16777214（全0和全1的地址要扣除，全0是网络号，全1是广播号）

- 127.*.*.*: 回环测试，用于测试本地网卡。127.0.0.1 “localhost”

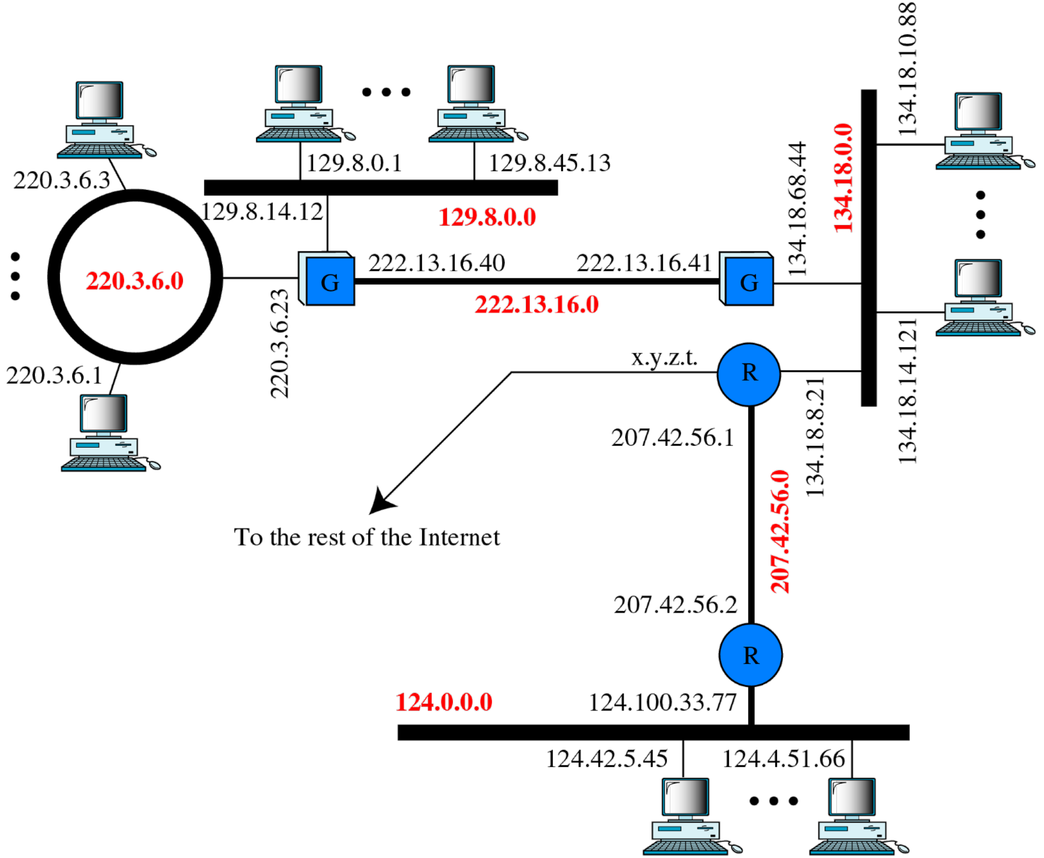
IP Addresses - Class B

- Start with 10
- Range 128.0.x.x to 191.255.x.x
- Second Octet also included in network address
- 2^{14} = 16,384 class B networks
- Up to 65,000 (=2^{16}-2) hosts
- *All allocated*

IP Addresses - Class C

- Start with 110
- Range 192.0.0.x to 223.255.255.x
- Second and third octet also part of network address
- 2^{21} = 2,097,152 networks
- Up to 254 (=2^8-2) hosts
- Nearly all allocated

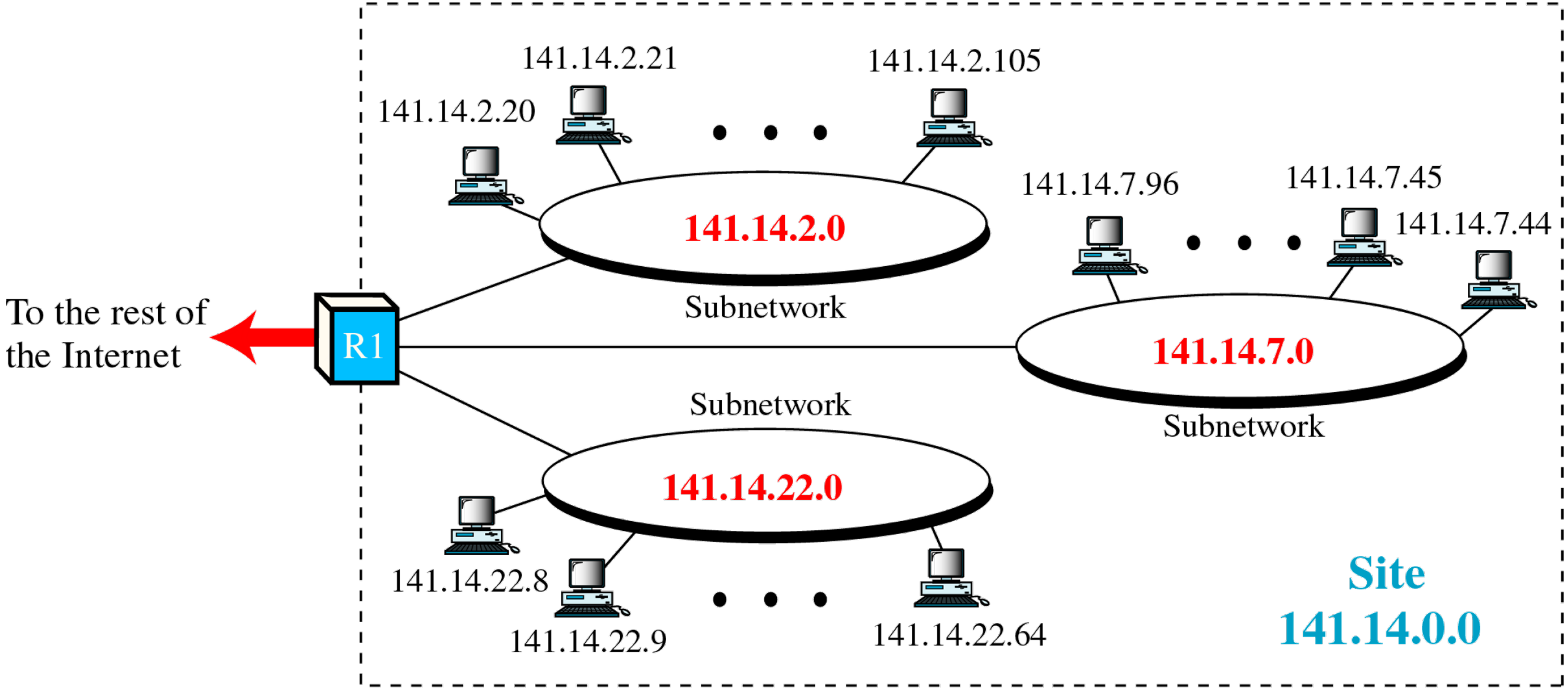
Inter-Networks with Addresses



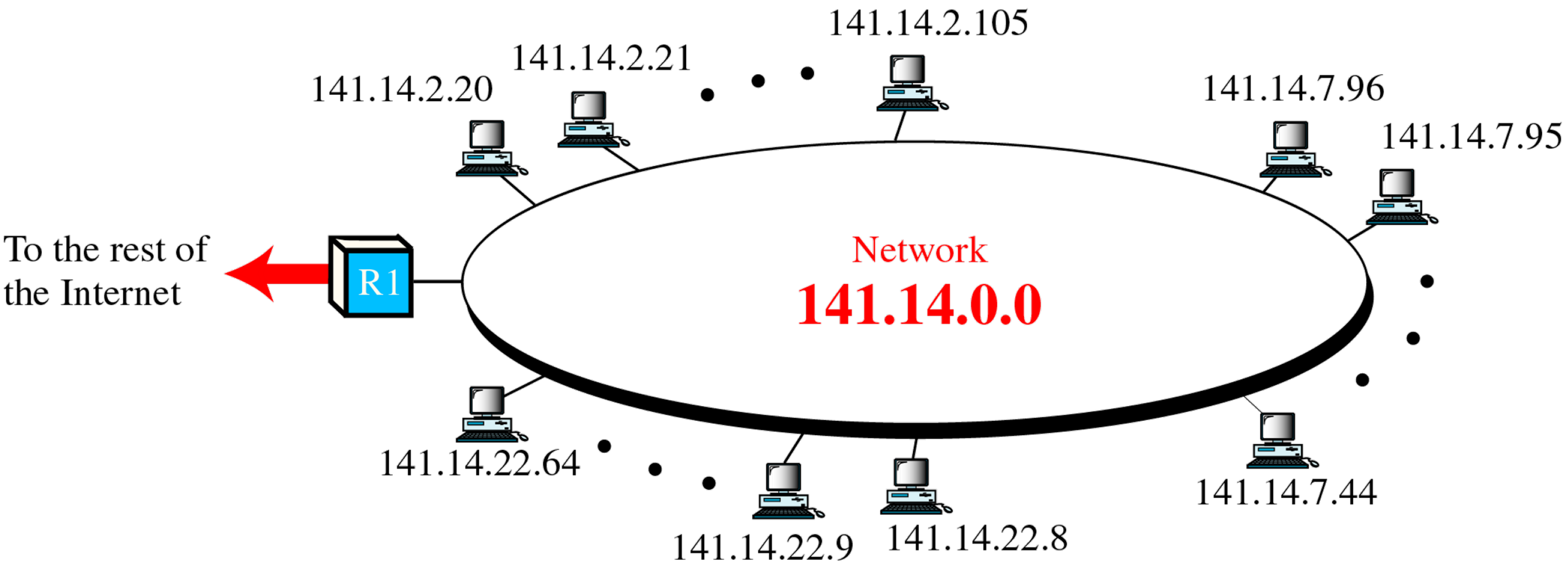
Subnets and Subnet Masks

- Handle problem of network address inadequacy
- Host portion of address partitioned into subnet number and host number
  - Subnet mask indicates which bits are subnet number and which are host number
  - Each LAN assigned a subnet number, more flexibility
  - Local routers route within subnetted network
- Subnets looks to rest of internet like a single network
  - Insulate overall Internet from growth of network numbers and routing complexity

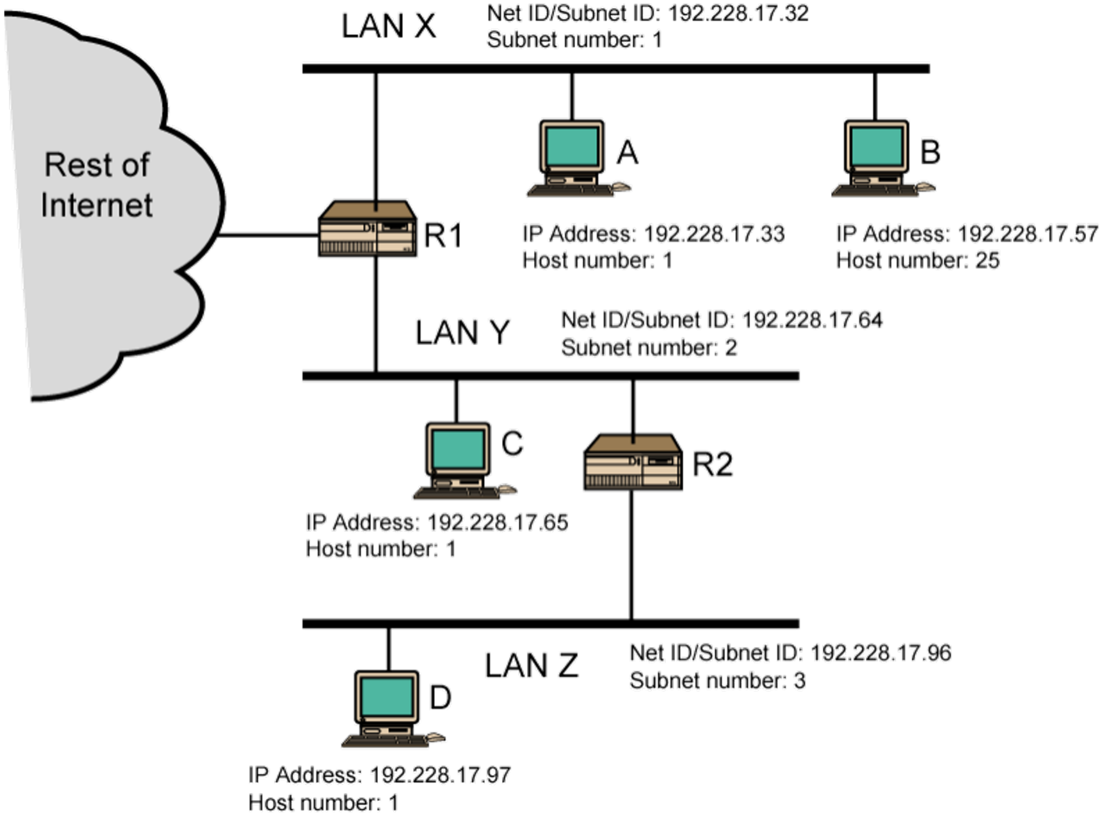
Subnets Example



Subnets to the Rest



Routing Using Subnets (1)



Routing Using Subnets (2)

(a) Dotted decimal and binary representations of IP address and subnet masks

	Binary Representation	**Dotted Decimal**
IP address	11000000.11100100.00010001.00111001	192.228.17.57
Subnet mask	11111111.11111111.11111111.11100000	255.255.255.224
Bitwise AND of address and mask (resultant network/subnet number)	11000000.11100100.00010001.00100000	192.228.17.32
Subnet number	11000000.11100100.00010001.001	1
Host number	00000000.00000000.00000000.00011001	25

(b) Default subnet masks

	Binary Representation	**Dotted Decimal**
Class A default mask	11111111.00000000.00000000.00000000	255.0.0.0
Example Class A mask	11111111.11000000.00000000.00000000	255.192.0.0
Class B default mask	11111111.11111111.00000000.00000000	255.255.0.0
Example Class B mask	11111111.11111111.11111000.00000000	255.255.248.0
Class C default mask	11111111.11111111.11111111.00000000	255. 255. 255.0
Example Class C mask	11111111.11111111.11111111.11111100	255. 255. 255.252

CIDR Notation

- Classless Inter Domain Routing (CIDR)
  - An IP address is represented as "A.B.C.D/n", where n is called the IP (network) prefix

IP Address	10 .	217 .	123 .	7
	00001010	11011001	01111011	00000111
Subnet	255 .	255 .	240 .	0
	11111111	11111111	11110000	00000000
Network ID	00001010	11011001	01110000	00000000
CIDR	10.217.112.0/20			

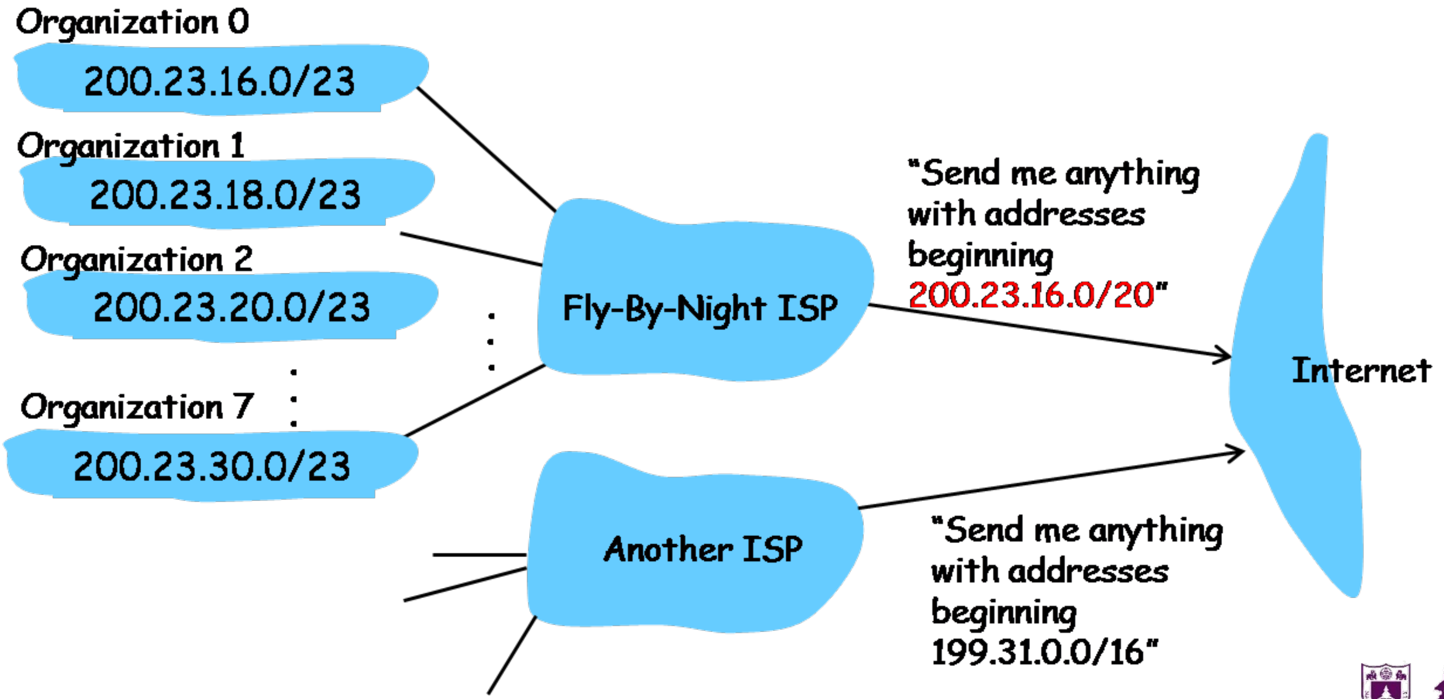
More General Case

- An ISP can be looked as a set of subnets
  - Support many organizations (Intranets)
  - Hierarchical addressing

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

Route Aggregation

- Allows efficient advertisement of routing information



IP addresses: how to get one?

That's actually two questions:

1. Q: How does a host get IP address within its network (host part of address)?
2. Q: How does a network get IP address for itself (network part of address)

How does host get IP address?

- hard-coded by sysadmin in config file (e.g., /etc/rc.config in UNIX)
- DHCP: Dynamic Host Configuration Protocol: dynamically get address from as server
  - "plug-and-play"

DHCP: Dynamic Host Configuration Protocol

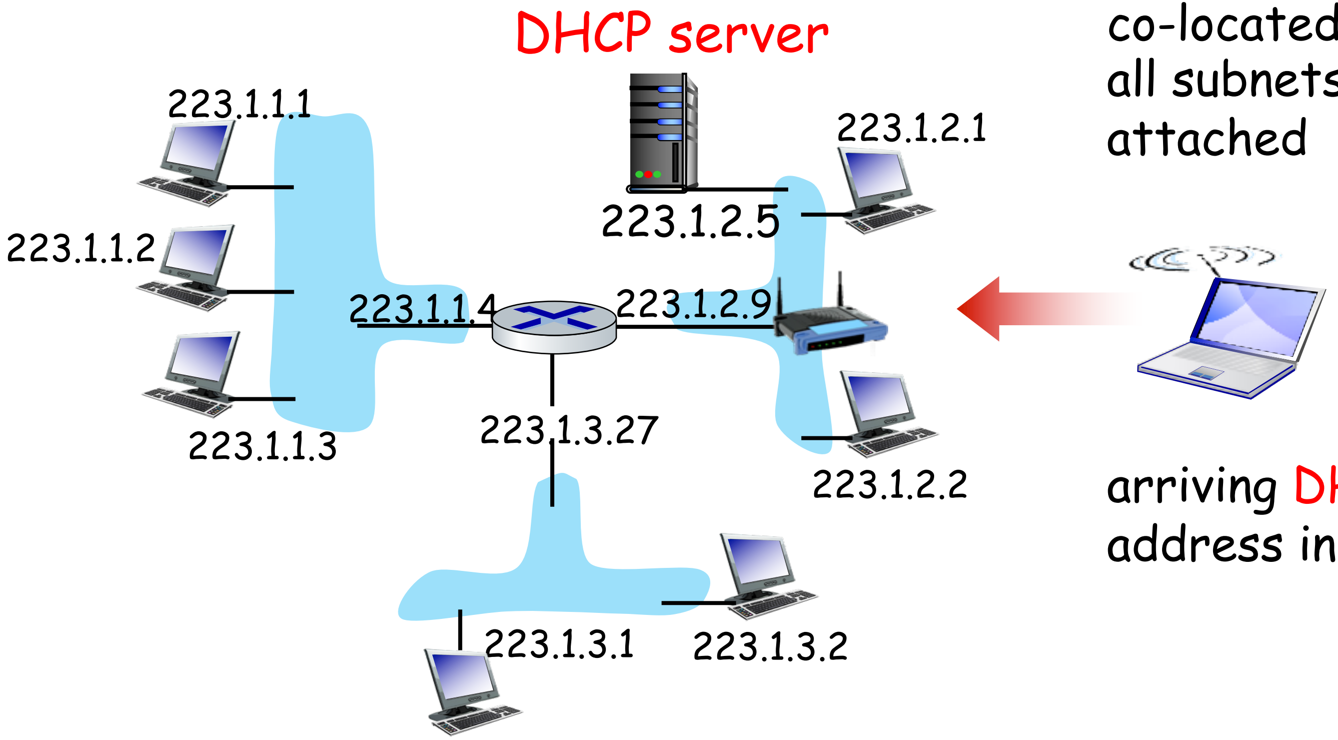
Goal: host dynamically obtains IP address from network server when it "joins" network

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

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

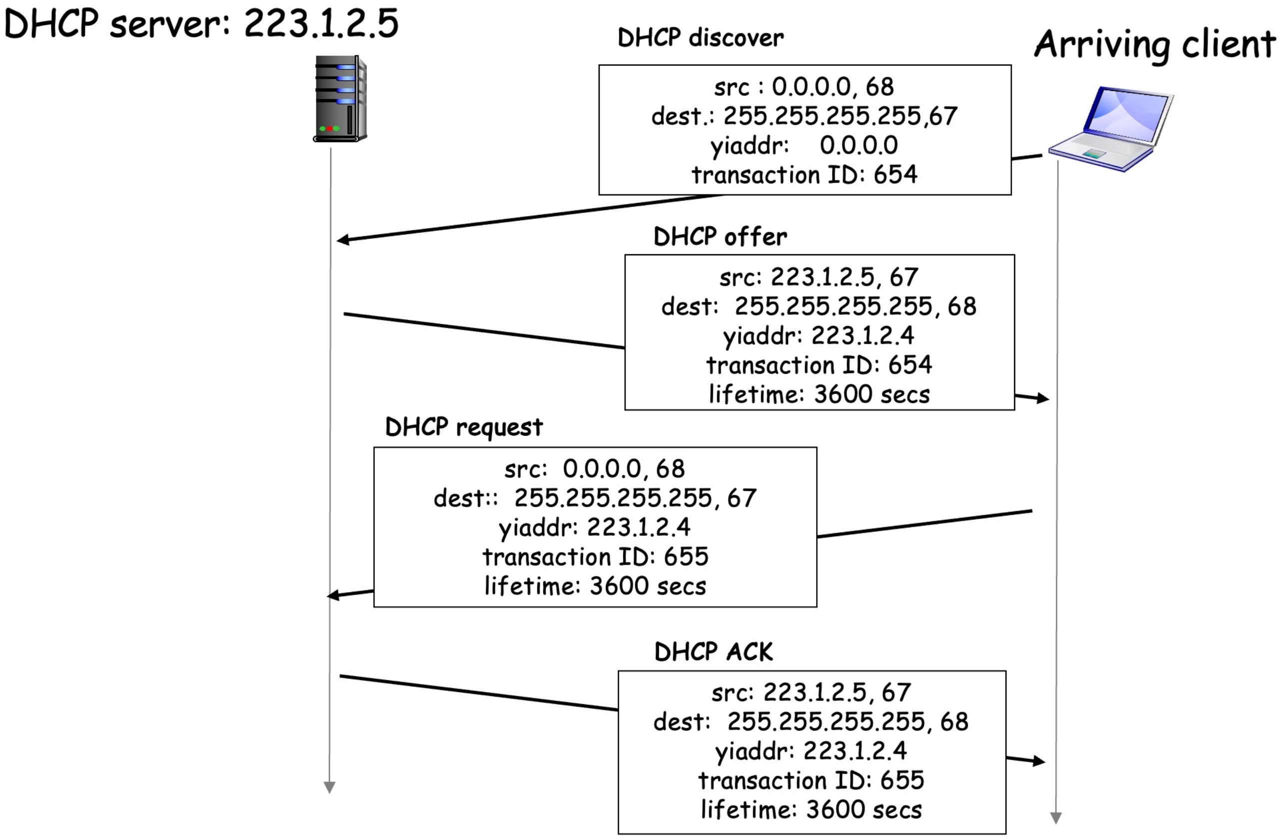
DHCP client-server scenario



Typically, DHCP server will be co-located in router, serving all subnets to which router is attached

arriving DHCP client needs address in this network

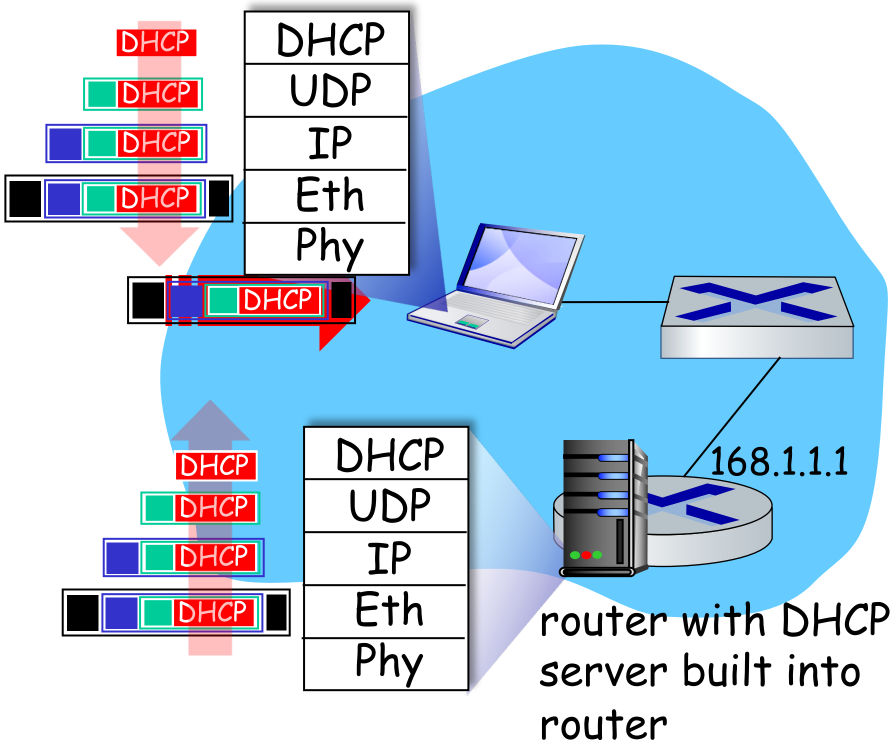
DHCP client-server scenario



DHCP: more than IP addresses

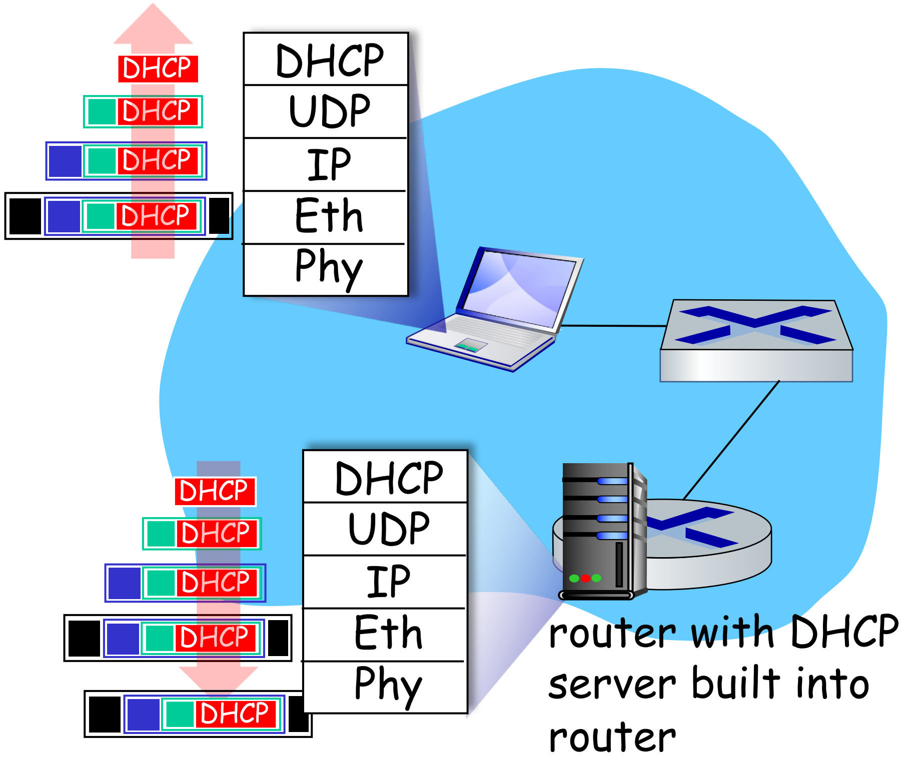
- DHCP can return more than just allocated IP address on subnet:
  - address of first-hop router for client
  - name and IP address of DNS sever
  - network mask (indicating network versus host portion of address)

DHCP: example



- Connecting laptop will use DHCP to get IP address, address of first-hop router, address of DNS server.
- DHCP REQUEST message encapsulated in UDP, encapsulated in IP, encapsulated in Ethernet
- Ethernet frame broadcast (dest: FFFFFFFFFFFF) on LAN, received at router running DHCP server
- Ethernet de-mux'ed to IP de-mux'ed, UDP de-mux'ed to DHCP

DHCP: example



- DHCP server formulates DHCP ACK containing client's IP address, IP address of first-hop router for client, name & IP address of DNS server
- encapsulated DHCP server reply forwarded to client, de-muxing up to DHCP at client
- client now knows its IP address, name and IP address of DNS server, IP address of its first-hop router

Outline

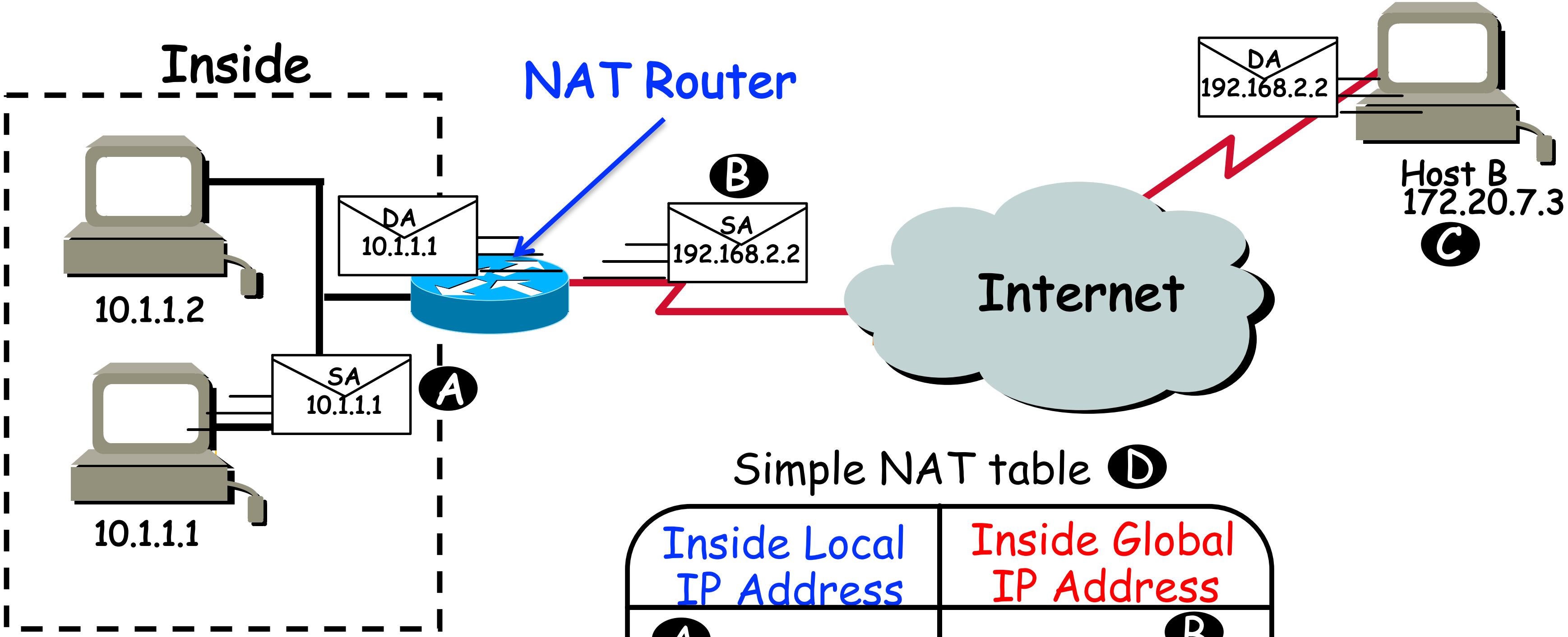
- IP Addressing
- Network Address Translation
- IPv6
- Generalized Forwarding and SDN
- Middleboxes

Network Address Translation

- NAT
  - Enables different sets of IP addresses for internal and external traffic
  - The IP address translations occur where the Intranet interfaces with the broader Internet

- Purposes
  - Acts as a firewall by hiding internal IP addresses
  - Enables an enterprise (organization) to use more internal IP addresses
  - Isolate the (organization / ISP) changes

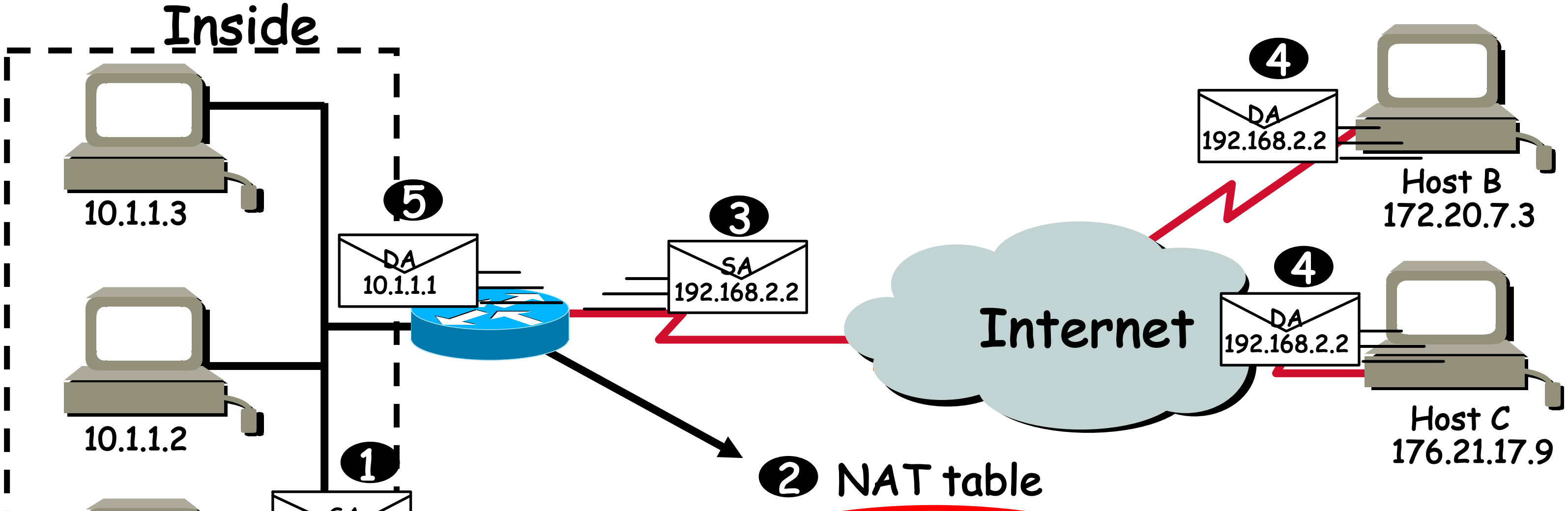
Illustration of NAT



Simple NAT table D

Inside Local IP Address	Inside Global IP Address
A 10.1.1.2	B 192.168.2.3
10.1.1.1	192.168.2.2

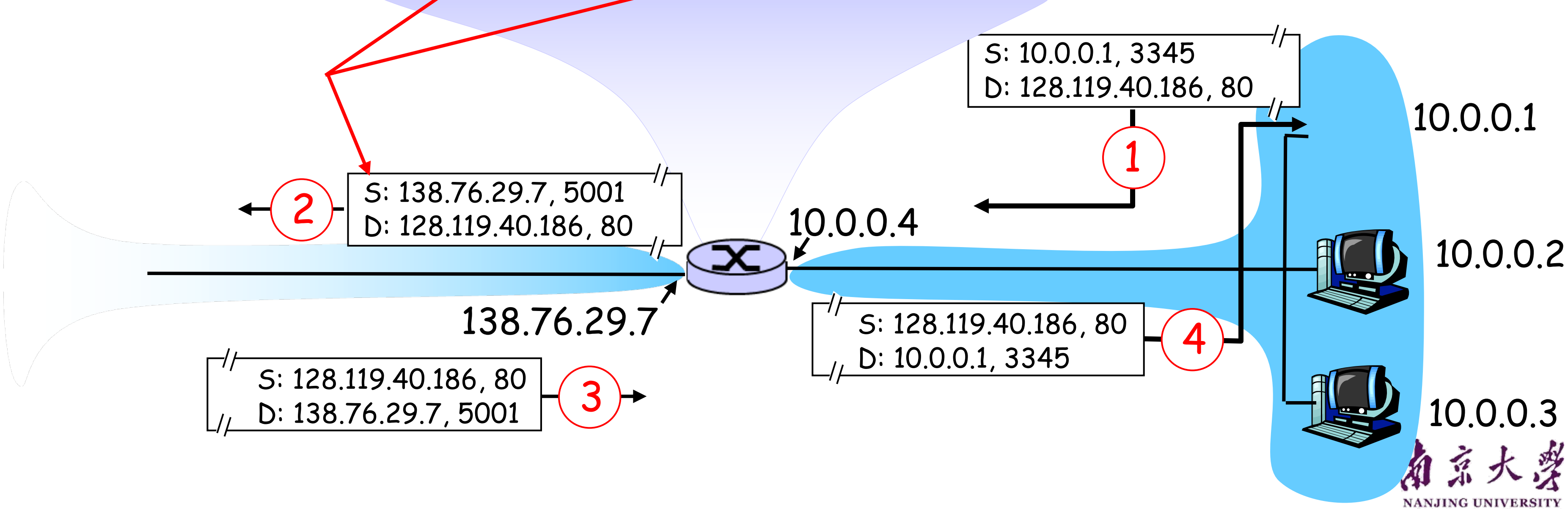
Overloading Global Address



Protocol 10.1.1.1	Inside Local IP Address: Port	Inside Global IP Address: Port	Outside Global IP Address: Port
TCP	10.1.1.3:1723	192.168.2.2:1492	176.21.17.9:23
TCP	10.1.1.2:1723	192.168.2.2:1723	176.21.17.9:23
TCP	10.1.1.1:1024	192.168.2.2:1024	172.20.7.3:23

Network Address Translation

NAT translation table	
WAN side addr	LAN side addr
138.76.29.7, 5001	10.0.0.1, 3345
……	……



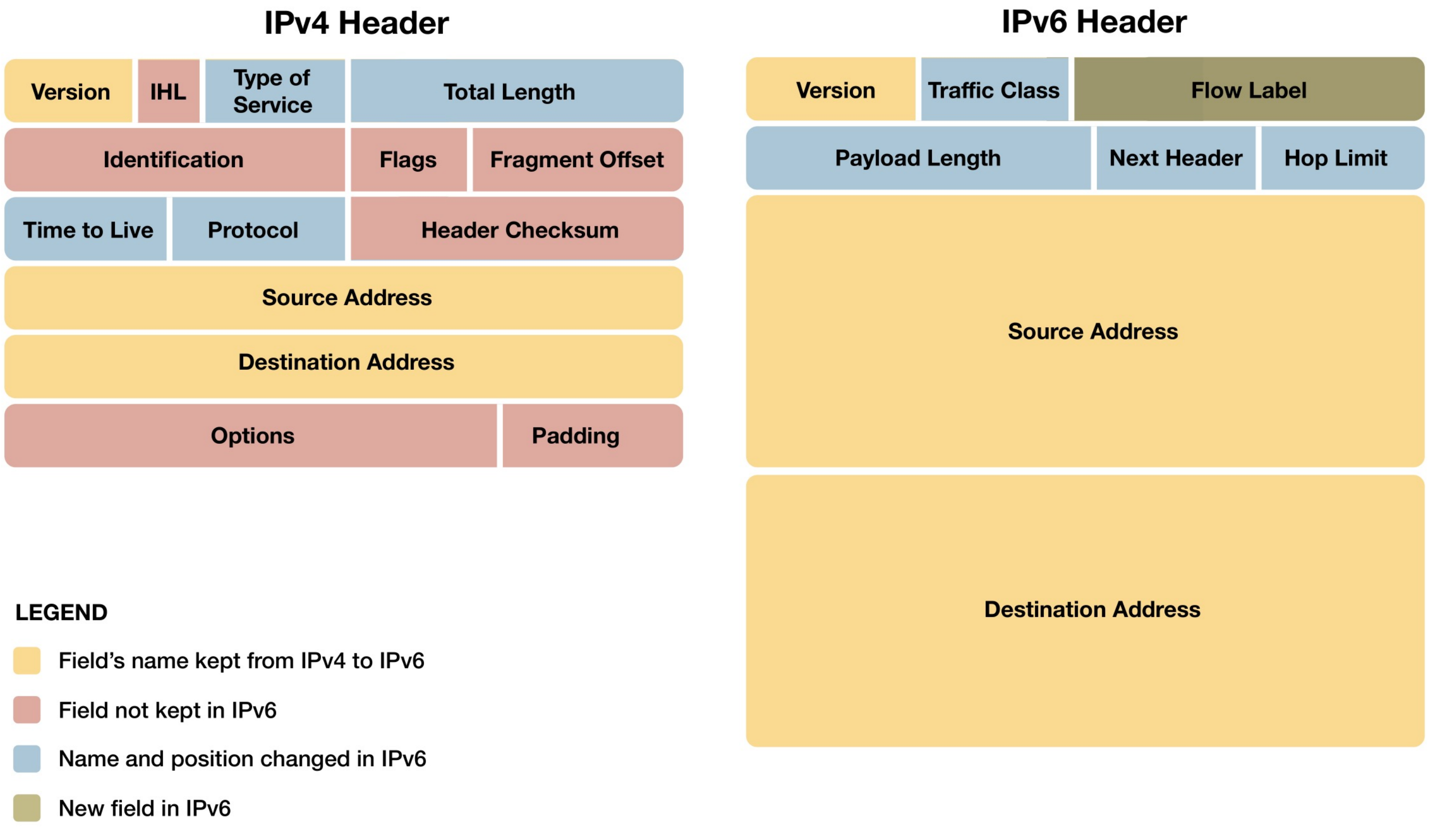
Outline

- IP Addressing
- Network Address Translation
- IPv6
- Generalized Forwarding and SDN
- Middleboxes

IPv6

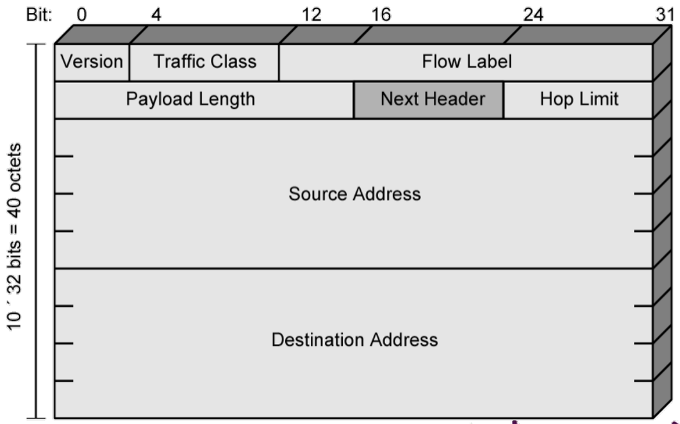
- Initial motivation: address space exhaustion
  - Rapid growth of networks and the Internet
  - 32-bit address space (esp. net address) soon to be completely allocated
- Additional motivation
  - New header format helps speed processing and forwarding
  - Header changes to facilitate QOS
  - No fragmentation at router
  - New address mode: route to "best" of several replicated servers

IPv6 Header VS IPv4 Header

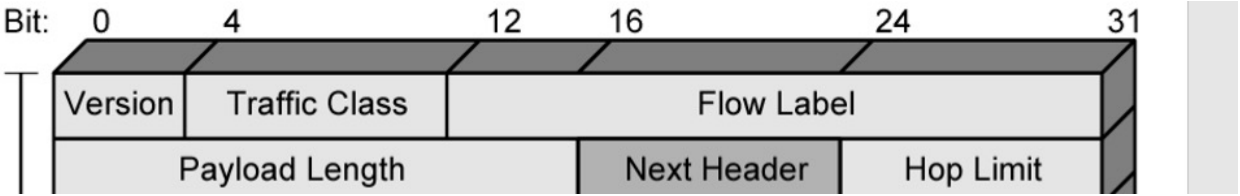


IPv6 Header Fields

- Version (4 bits): 6
- Traffic Class (8 bits)
  - Classes or priorities of packet, identify QoS
- Flow Label (20 bits)
  - Identify datagrams in the same "flow"
- Payload length (16 bits)
  - Includes all extension headers plus user data
- Next Header (8 bits)
  - Identifies type of the next header
  - Extension or next layer up
- Source / Destination Address (128 bits)

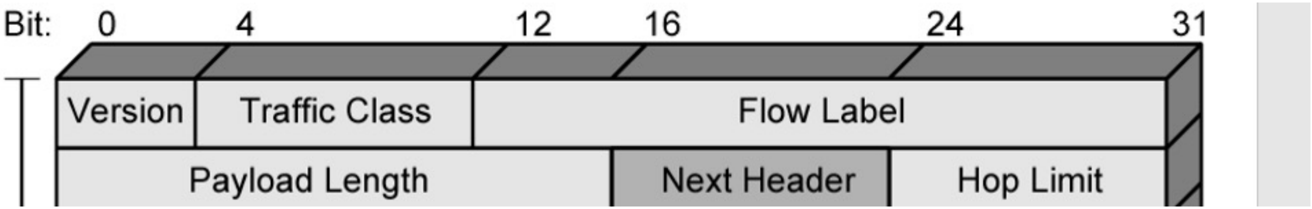


Traffic Class



- The 8-bit field in the IPv6 header is available for use by originating nodes and/or forwarding routers to identify and distinguish between different classes or priorities of IPv6 packets.
  - E.g., used as the codepoint in DiffServ
- General requirements
  - Service interface must provide means for upper-layer protocol to supply the value of traffic class
  - Value of traffic class can be changed by source, forwarder, receiver
  - An upper-layer protocol should not assume the value of traffic class in a packet has not been changed.

IPv6 Flow



- A sequence of packets sent from a particular source to a particular destination
- From hosts point of view
  - Generated from one application and have the same transfer service requirements
  - May comprise a single or multiple TCP connections
  - One application may generate a single flow or multiple flows
- From routers point of view
  - Share attributes that affect how these packets are handled by the router
  - e.g. routing, resource allocation, discard requirements, accounting, and security

Flow Label

- A flow is uniquely identified by the combination of
  - Source and destination address
  - A non-zero 20-bit Flow Label
- Flow requirements are defined prior to flow commencement
  - Then a unique Flow Label is assigned to the flow
- Router decide how to route and process the packet by
  - Simply looking up the Flow Label in a table and without examining the rest of the header

Advantages of IPv6 over IPv4

- Expanded addressing capabilities
  - 128 bit
  - Scalability of multicast addresses
  - Anycast – delivered to one of a set of nodes
  - Address auto-configuration
- Improved option mechanism
  - Separate optional headers between IPv6 header and transport layer header
  - Most are not examined by intermediate routers
  - Easier to extend options
  - Checksum removed to further reduce processing time at each router

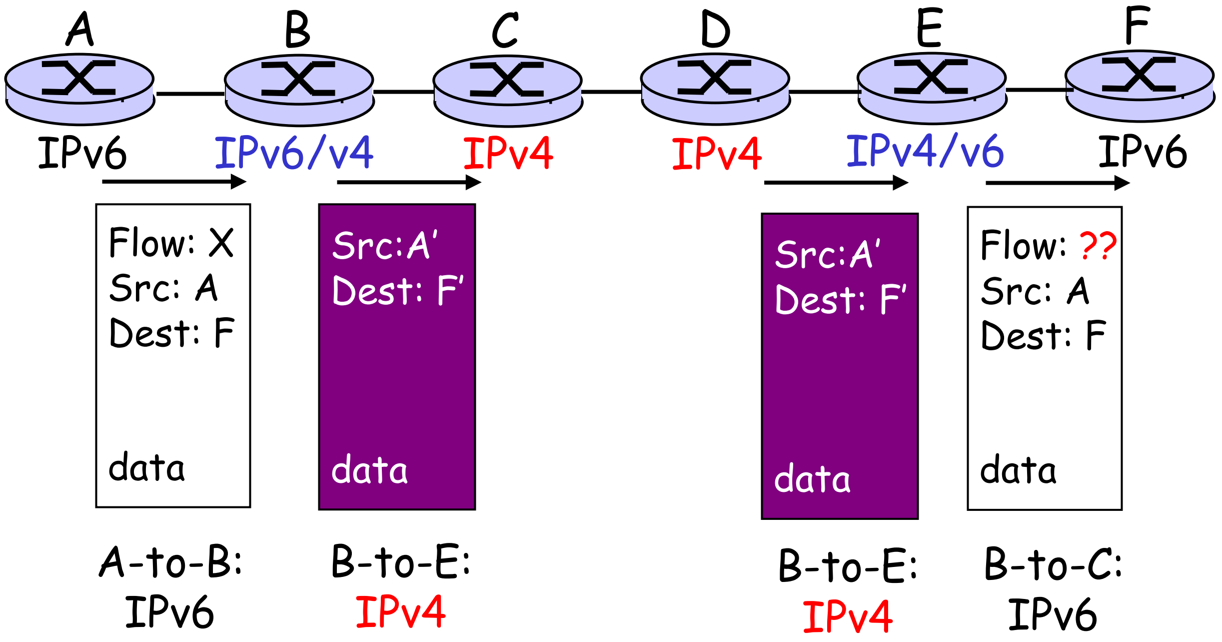
Advantages of IPv6 over IPv4

- Support for resource allocation
  - Uses traffic class
  - Grouping packets to particular traffic flow
  - Allows QoS handling other than best-effort, e.g. real-time video
- More efficient and robust mobility mechanism
- More security: Built-in, strong IP-layer encryption and authentication

Transition From IPv4 To IPv6

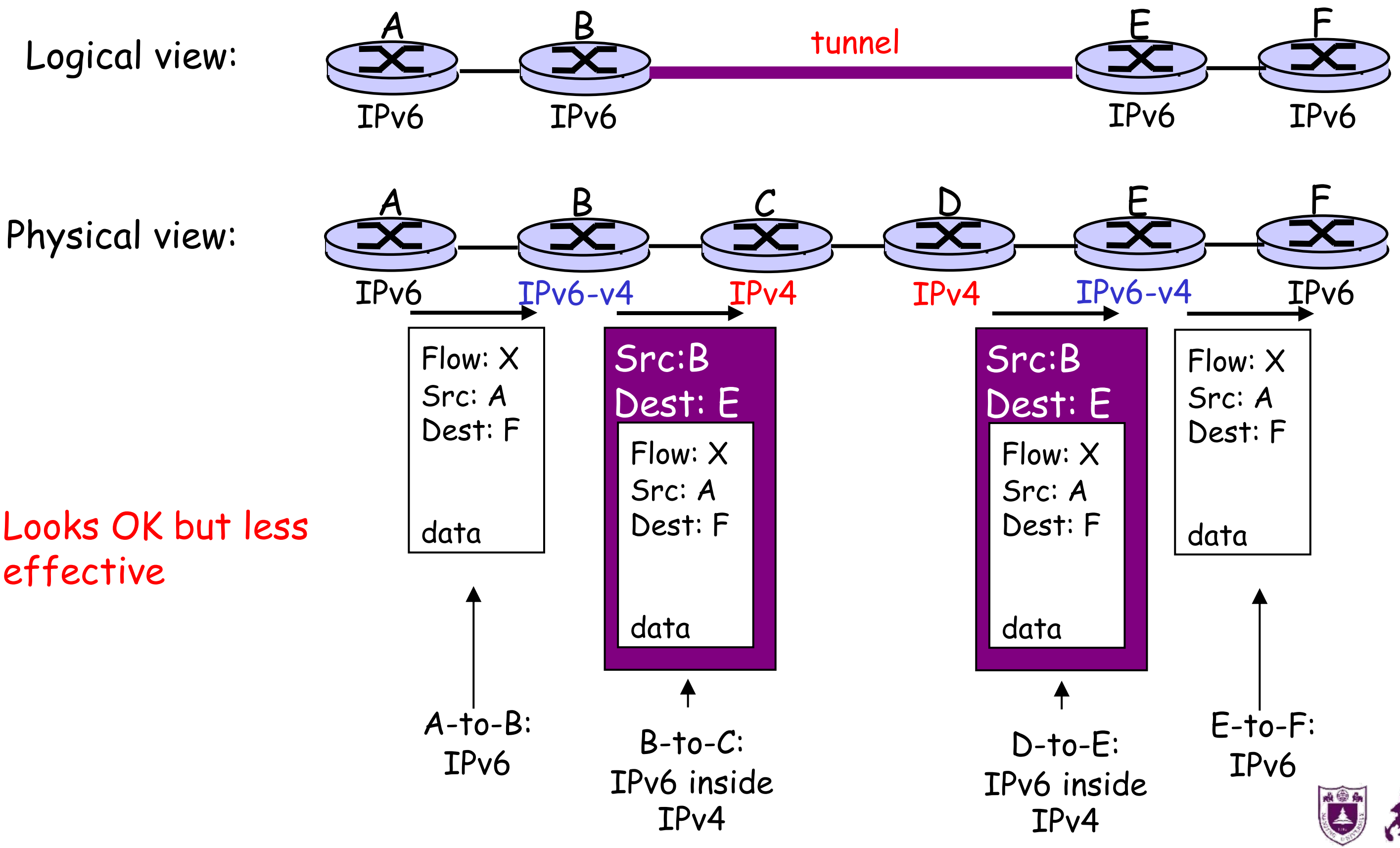
- Not all routers can be upgraded simultaneously
  - How will the network operate with mixed IPv4 and IPv6 routers
- Two proposed approaches
  - Dual Stack - some routers with dual stack (IPv6, IPv4) can translate between formats
  - Tunneling - IPv6 carried as payload in IPv4 datagram among IPv4 routers

Dual Stack Approach



- Address translation between IPv4 and IPv6 is needed
- Some IPv6 features is lost

Tunneling



Looks OK but less effective

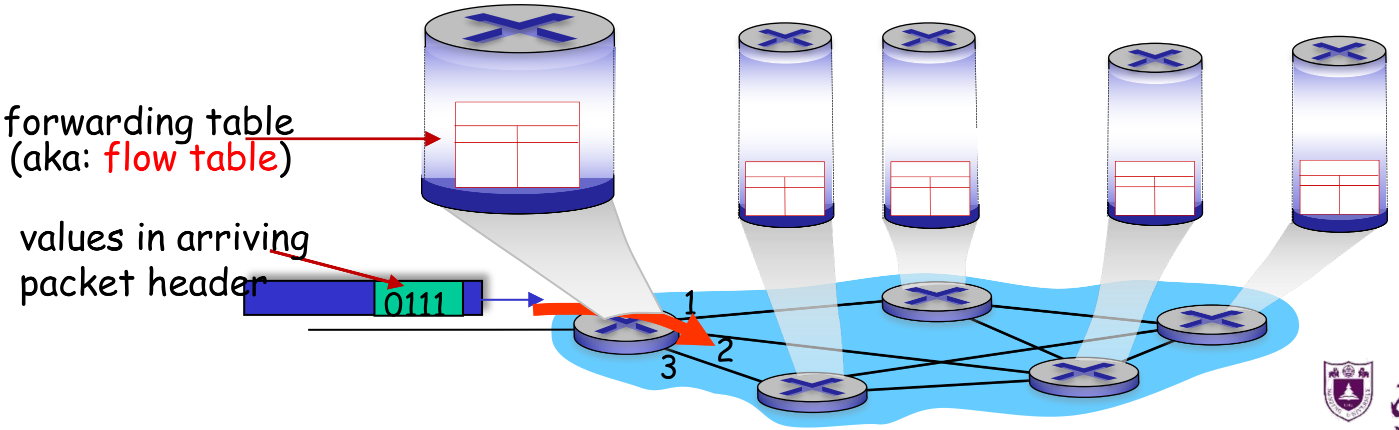
Outline

- IP Addressing
- Network Address Translation
- IPv6
- Generalized Forwarding and SDN
- Middleboxes

Generalized forwarding: match plus action

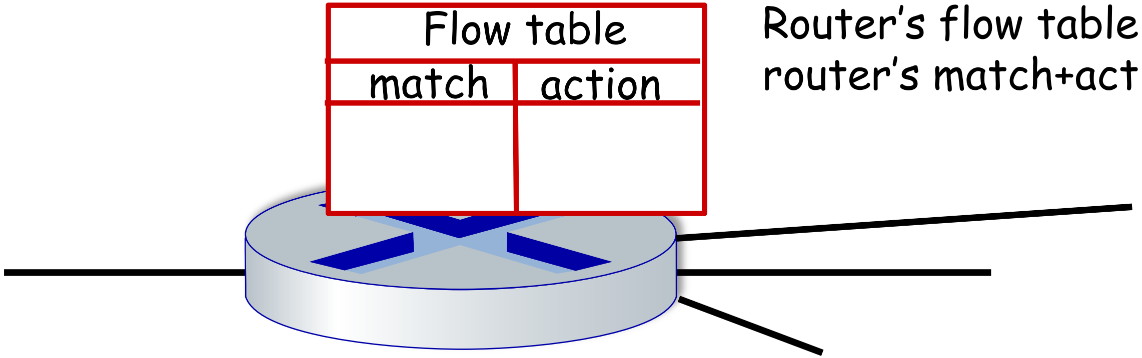
Review: each router contains a forwarding table (aka: flow table)

- "match plus action" abstraction: match bits in arriving packet, take action
- destination-based forwarding: forward based on dest. IP address
- generalized forwarding:
  - many header fields can determine action
  - many action possible: drop/copy/modify/log packet



Flow table abstraction

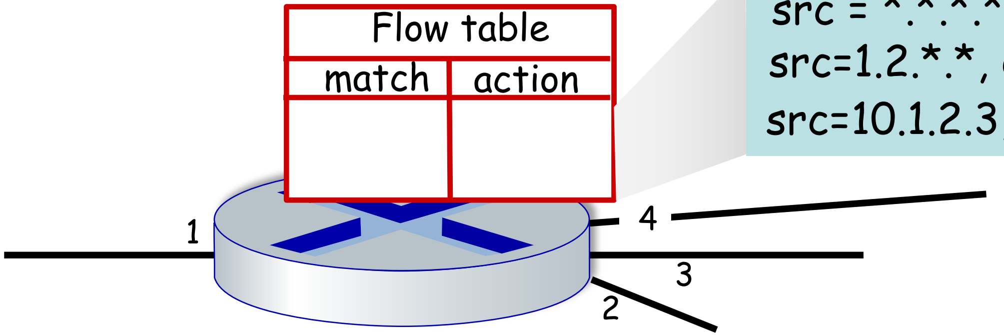
- flow: defined by header field values (in link-, network-, transport-layer fields)
- generalized forwarding: simple packet-handling rules
  - match: pattern values in packet header fields
  - actions: for matched packet: drop, forward, modify, matched packet or send matched packet to controller
  - priority: disambiguate overlapping patterns
  - counters: #bytes and #packets



Router's flow table define router's match+action rules

Flow table abstraction

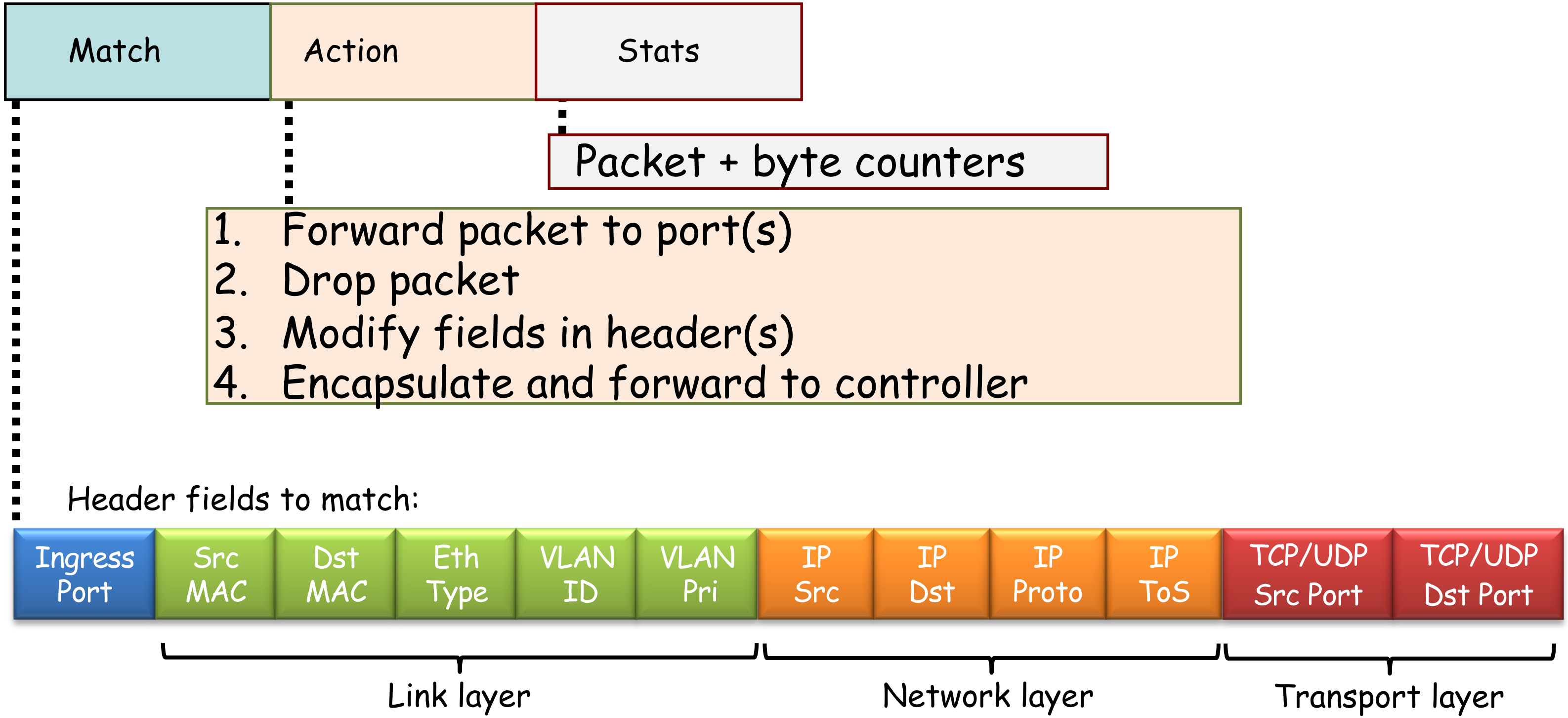
- flow: defined by header fields
- generalized forwarding: simple packet-handling rules
  - match: pattern values in packet header fields
  - actions: for matched packet: drop, forward, modify, matched packet or send matched packet to controller
  - priority: disambiguate overlapping patterns
  - counters: #bytes and #packets



src = *.*.*.*, dest=3.4.*.*	forward(2)
src=1.2.*.*, dest=*.*.*.*	drop
src=10.1.2.3, dest=*.*.*.*	send to controller

* : wildcard

OpenFlow: flow table entries



OpenFlow abstraction

- match+action: abstraction unifies different kinds of devices

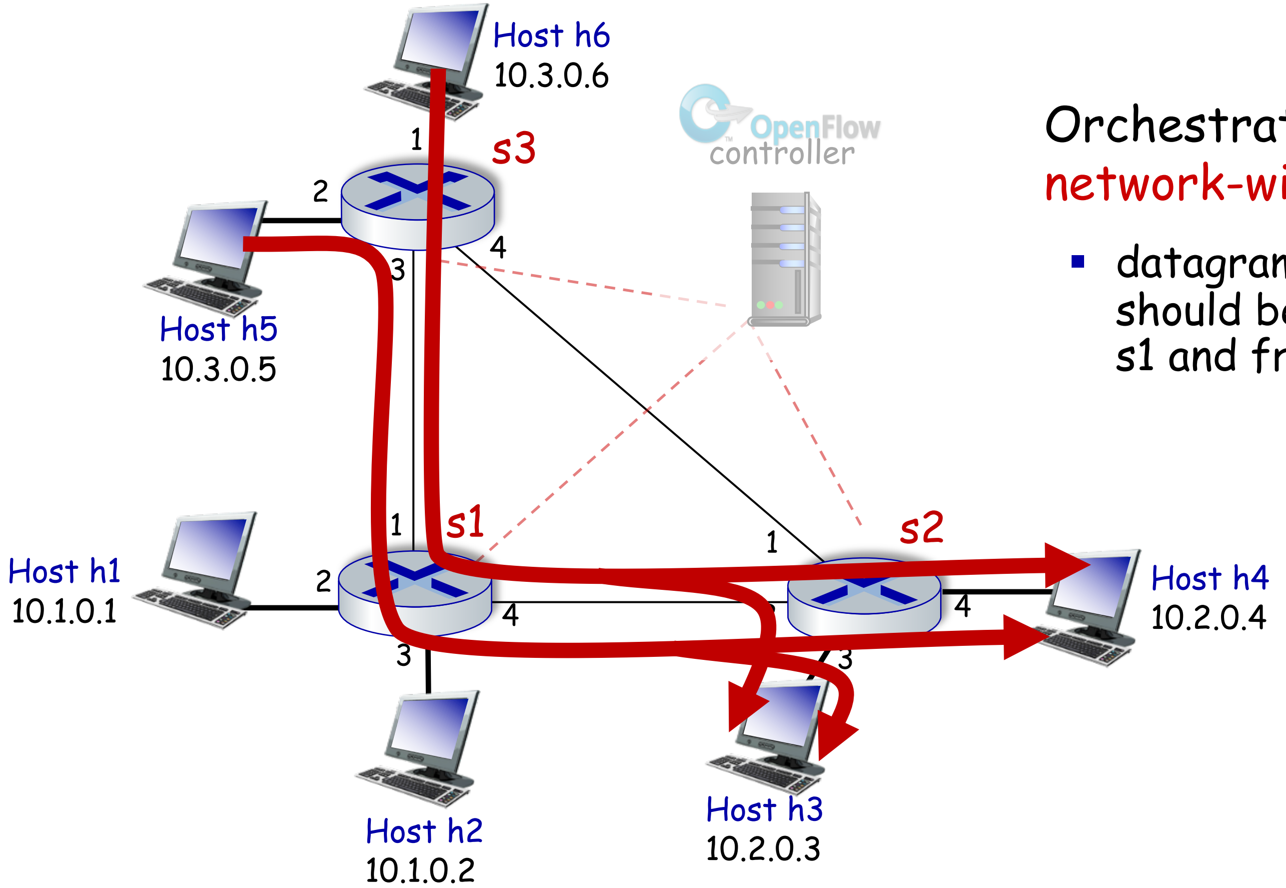
Router
- match: longest destination IP prefix
- action: forward out a link

Switch
- match: destination MAC address
- action: forward or flood

Firewall
- match: IP addresses and TCP/UDP port numbers
- action: permit or deny

NAT
- match: IP address and port
- action: rewrite address and port

OpenFlow example



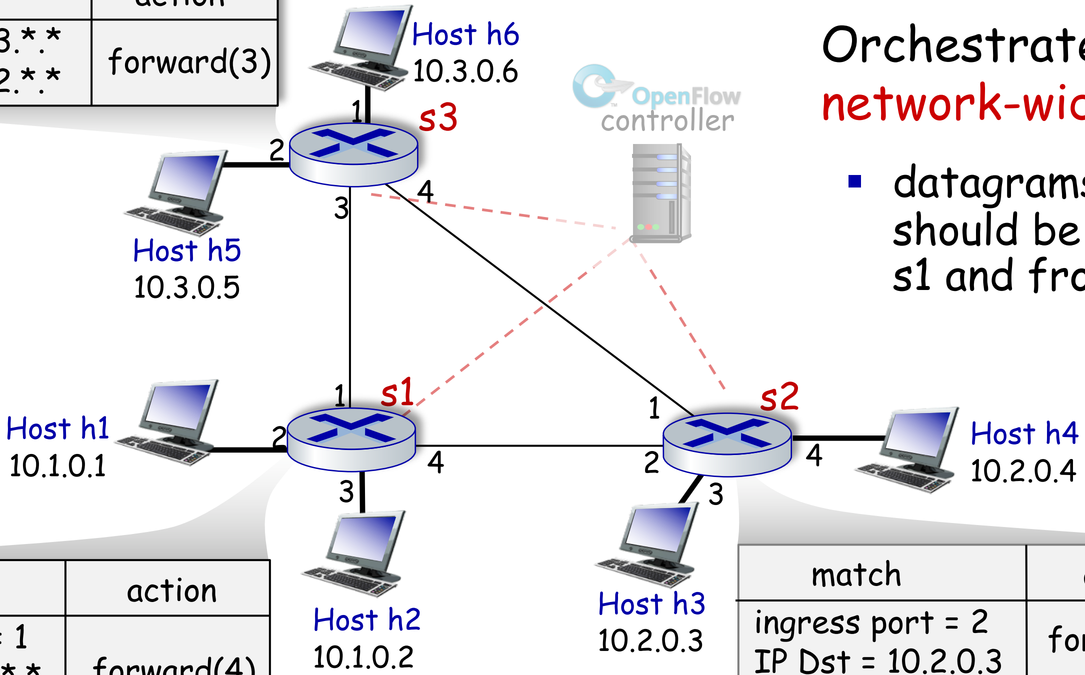
Orchestrated tables can create network-wide behavior, e.g.,:

- datagrams from hosts h5 and h6 should be sent to h3 or h4, via s1 and from there to s2

OpenFlow example

Orchestrated tables can create network-wide behavior, e.g.,:

- datagrams from hosts h5 and h6 should be sent to h3 or h4, via s1 and from there to s2



s3 flow table:

match	action
IP Src = 10.3.*.* IP Dst = 10.2.*.*	forward(3)

s1 flow table:

match	action
ingress port = 1 IP Src = 10.3.*.* IP Dst = 10.2.*.*	forward(4)

s2 flow table:

match	action
ingress port = 2 IP Dst = 10.2.0.3	forward(3)
ingress port = 2 IP Dst = 10.2.0.4	forward(4)

Outline

- IP Addressing
- Network Address Translation
- IPv6
- Generalized Forwarding and SDN
- Middleboxes

Middleboxes

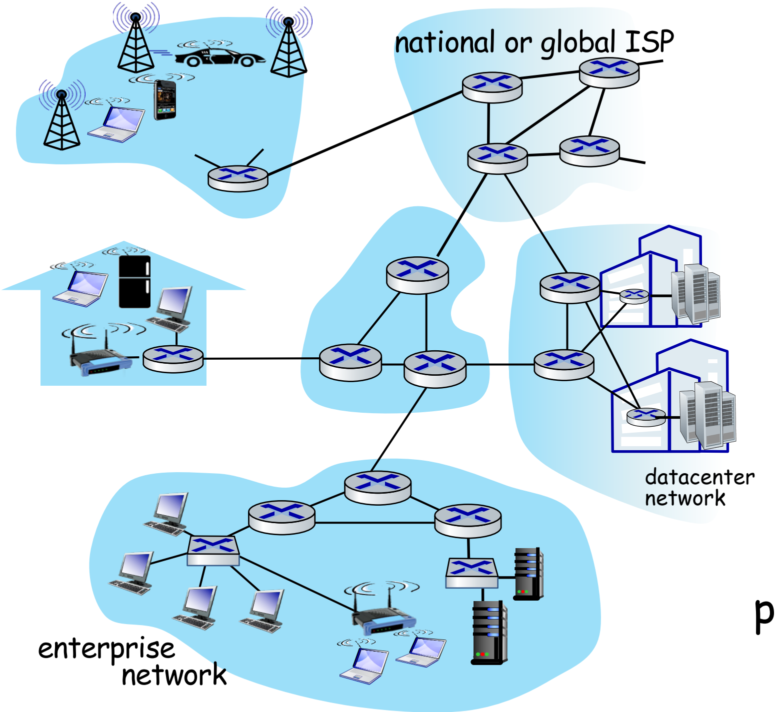
Middlebox (RFC 3234)

"any intermediary box performing functions apart from normal, standard functions of an IP router on the data path between a source host and destination host"

Middleboxes everywhere!

Firewalls, IDS: corporate, institutional, service providers, ISPs

NAT: home, cellular, institutional



Load balancers: corporate, service provider, data center, mobile nets

Application-specific: service providers, institutional, CDN

Caches: service provider, mobile, CDNs

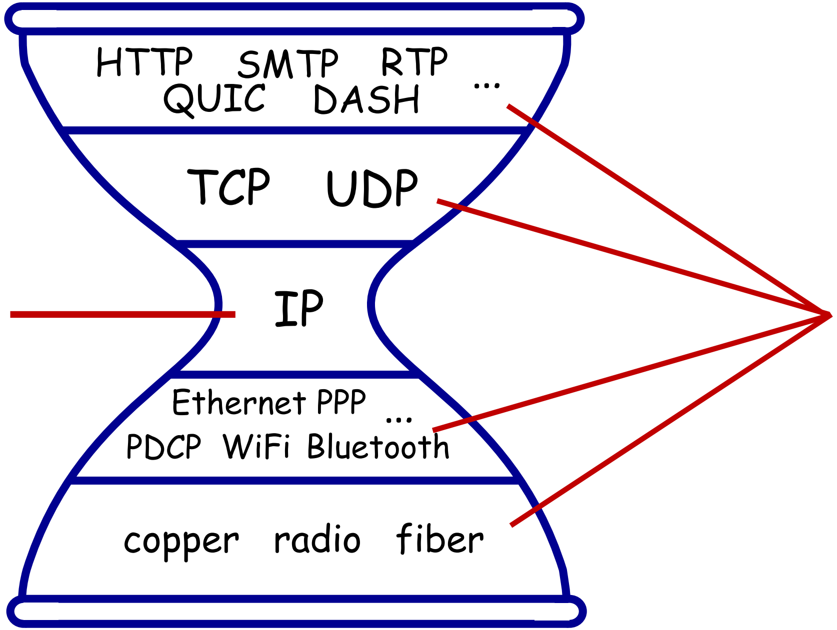
Middleboxes

- initially: proprietary (closed) hardware solutions
- move towards "whitebox" hardware implementing open API
  - move away from proprietary hardware solutions
  - programmable local actions via match+action
  - move towards innovation/differentiation in software
- SDN: (logically) centralized control and configuration management often in private/public cloud
- network functions virtualization (NFV): programmable services over white box networking, computation, storage

The IP hourglass

Internet's "thin waist":

- one network layer protocol: IP
- must be implemented by every (billions) of Internet-connected devices

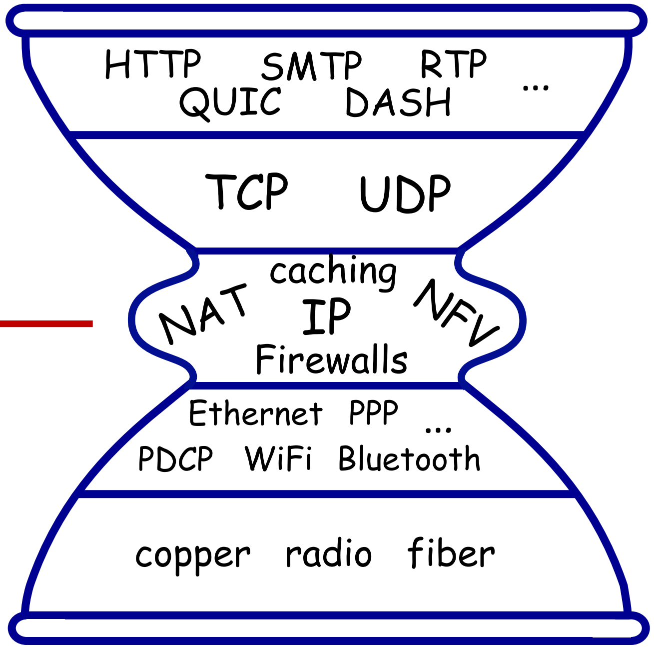


many protocols in physical, link, transport, and application layers

The IP hourglass, at middle age

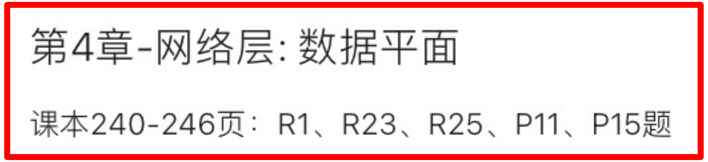
Internet's middle age "love handles"?

- middleboxes, operating inside the network



课程习题（作业）——截止日期：4月15日晚23:59

- **课本240-246页**：R1、R23、R25、P11、P15题
- 提交方式：https://selearning.nju.edu.cn/（教学支持系统）



- 命名：学号+姓名+第*章。
- 若提交遇到问题请及时发邮件或在下一次上课时反馈。

课程习题（作业）——截止日期：4月15日晚23:59

R1. 我们回顾在本书中使用的某些术语。前面讲过运输层的分组名字是报文段，数据链路层的分组名字是帧。网络层的分组名字是什么？前面讲过路由器和链路层交换机都被称为分组交换机。路由器与链路层交换机间的根本区别是什么？

R23. 考察使用 DHCP 的主机，获取它的 IP 地址、网络掩码、默认路由器及其本地 DNS 服务器的 IP 地址。列出这些值。

R25. 假设某应用每 20ms 生成一个 40 字节的数据块，每块封装在一个 TCP 报文段中，TCP 报文段再封装在一个 IP 数据报中。每个数据报的开销有多大？应用数据所占百分比是多少？

P11 考虑互联 3 个子网（子网 1、子网 2 和子网 3）的一台路由器。假定这 3 个子网的所有接口要求具有前缀 223.1.17/24。还假定子网 1 要求支持多达 60 个接口，子网 2 要求支持多达 90 个接口，子网 3 要求支持多达 12 个接口。提供 3 个满足这些限制的网络地址（形式为 *a.b.c.d/x*）。

课程习题（作业）——截止日期：4月15日晚23:59

P15 考虑图 4-20 中显示的拓扑。（在 12:00 以顺时针开始）标记具有主机的 3 个子网为网络 A、B 和 C，标记没有主机的子网为网络 D、E 和 F。

a. 为这 6 个子网分配网络地址，要满足下列限制：所有地址必须从 214.97.254/23 起分配；子网 A 应当具有足够地址以支持 250 个接口；子网 B 应当具有足够地址以支持 120 个接口；子网 C 应当具有足够地址以支持 120 个接口。当然，子网 D、E 和 F 应当支持两个接口。对于每个子网，分配采用的形式是 *a. b. c. d/x* 或 *a. b. c. d/x ~ e. f. g. h/y*。

b. 使用你对（a）部分的答案，为这 3 台路由器提供转发表（使用最长前缀匹配）。

Q & A

殷亚凤
智能软件与工程学院
苏州校区南雍楼东区225
yafeng@nju.edu.cn，https://yafengnju.github.io/