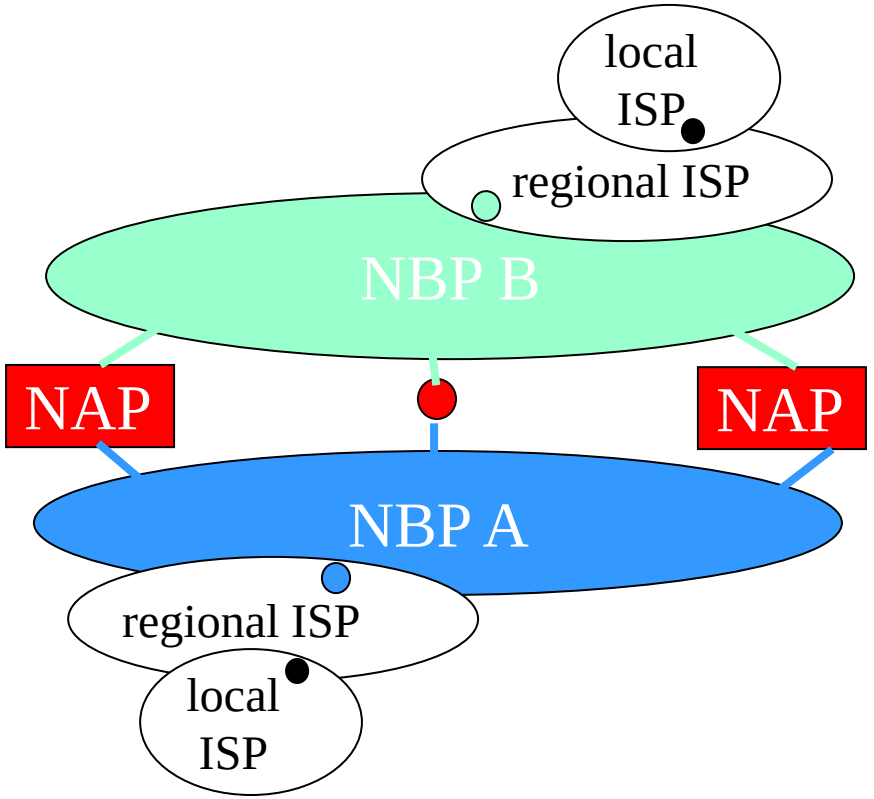
Distributed Systems
分布式系统

Internetworking
网际路由

Internet Structure: Network of Networks

Internet is in a hierarchical structure

- National/ International Backbone Providers (NBP)
  - e.g. BBN/GTE, Sprint, AT&T, IBM, UUNet
  - interconnect (peer) with each other privately, or at public Network Access Point (NAPs)
- Regional ISPs
  - connect into NBPs
- Local ISP
  - connect into regional ISPs



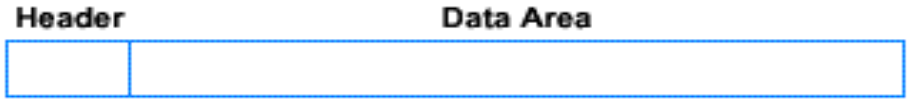
Internet Protocol (IP)

The IP protocol, defined in RFC 791 (1981) at Layer 3, specifies:

- Internet packet format
- Internet addressing
- Internet routing

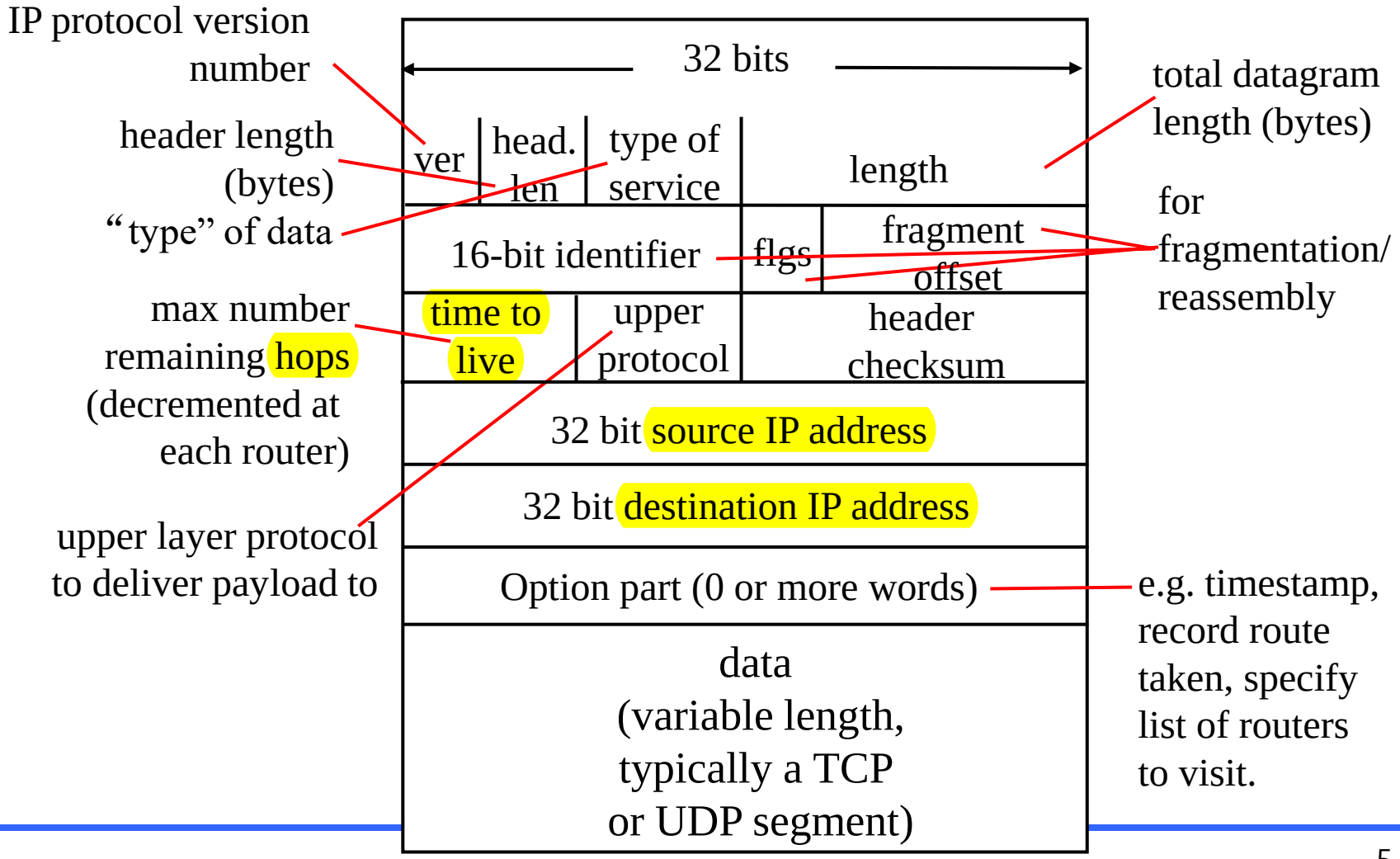
Internet Packets

An IP packet, called *IP datagram*, consists of two parts:



- **Header**
  - Contains source and destination address
  - Fixed-size fields
- **Data Area** (Payload)
  - Variable size up to 64K
  - No minimum size

IPv4 datagram format

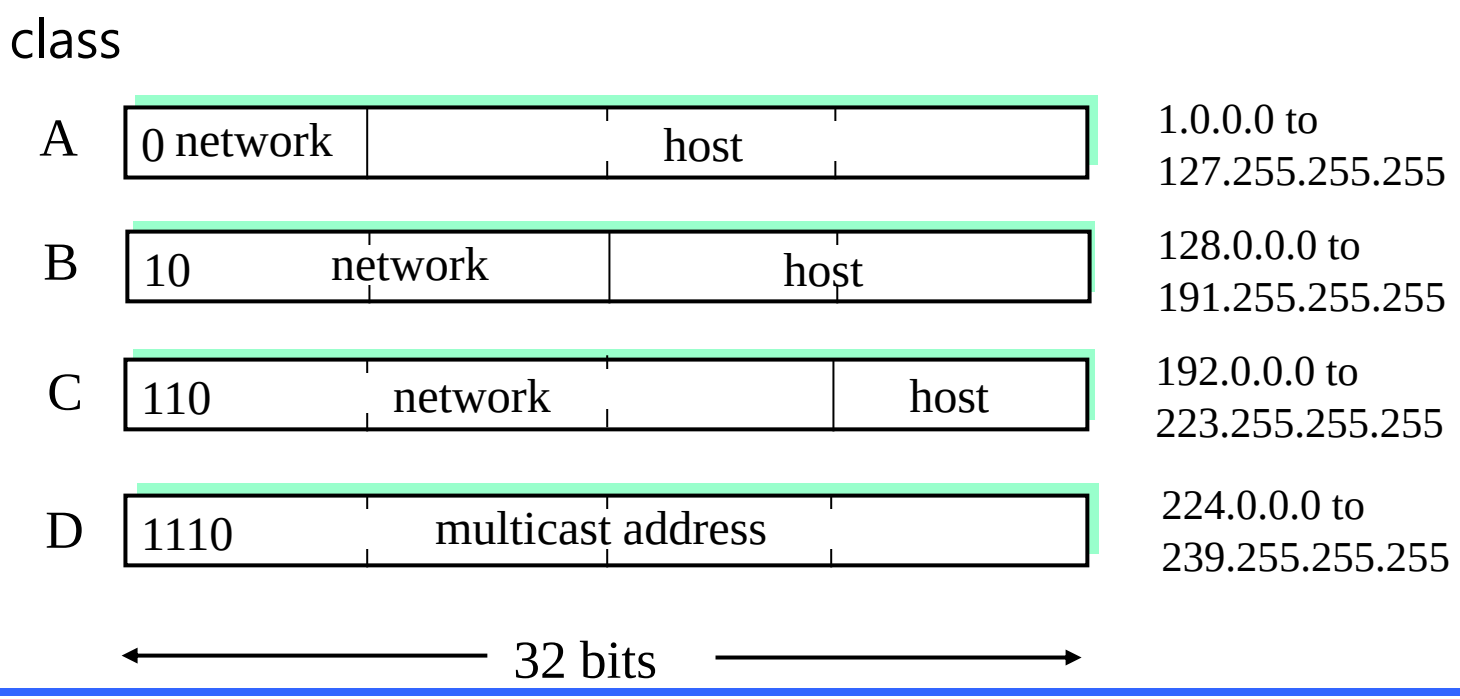


IP Address Details

- An IPv4 address has 32 bits (4 bytes) – written like 192.0.2.53
  - It consists of two parts: prefix identifies network and suffix identifies host
  - It was announced in Jan 2011 that all IPv4 addresses were allocated out.
- An IPv6 address is 128 bits, written in hexadecimal. It looks like
  - 2001:0db8::53. The mark "::" means all zeros in between. So 2001:0db8::53 is 2001:0db8:0000:0000:0000:0000:0000:53
- Global authority assigns unique prefix to network (ICANN)
  - Internet Corporation for Assigned Names and Numbers http://www.icann.org/
- Local administrator assigns unique suffix to host

IPv4 Addresses

The Classful Addressing:



Classes and Network Sizes

Address Class	Prefix Bits	Max Nets	Suffix Bits	Max Hosts Per Net
A	7	128	24	16777216
B	14	16384	16	65536
C	21	2097152	8	256

- Network size is determined by number of bits for hosts
  - Class *A* large
  - Class *B* medium
  - Class *C* small

Are there enough addresses?

- Unfortunately No!
  - 32 bits $\rightarrow$ 4 billion unique addresses.
  - but addresses are assigned in chunks, most of them are not used (particularly class D addresses for multicast).
  - techniques, such as CIDR, NAT, and DHCP are introduced to relief the shortage of IP addresses.
- Expanding the address space!
  - IPv6 128 bit addresses
  - Difficult to deploy (require changes to the core of the Internet)
  - The transition from IPv4 to IPv6 is very slow

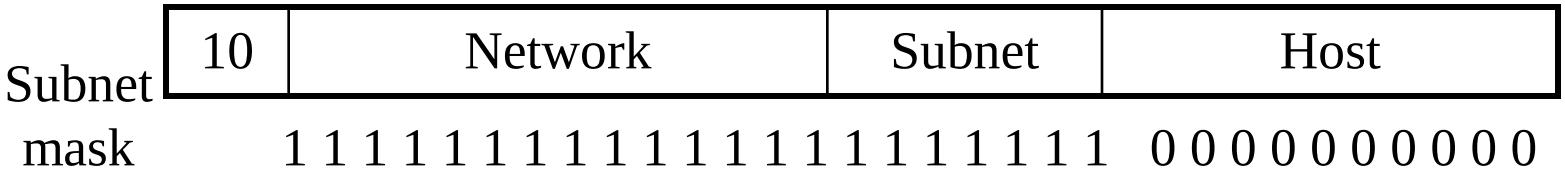
IP address allocation by Geographical Regions

- To simplify routing tables, since 1993, IP address allocation follows the geographical regions:
  - 194.0.0.0 to 195.255.255.255 for Europe
  - 198.0.0.0 to 199.255.255.255 for North America
  - 200.0.0.0 to 201.255.255.255 for Central & South America
  - 202.0.0.0 to 203.255.255.255 for Asia Pacific
- Routing table size can be greatly reduced, e.g., routers outside Europe can have a single table entry for range: 194.0.0.0 - 195.255.255.255 to be routed to the nearest Europe gateway.

Subnets and Subnet Masks

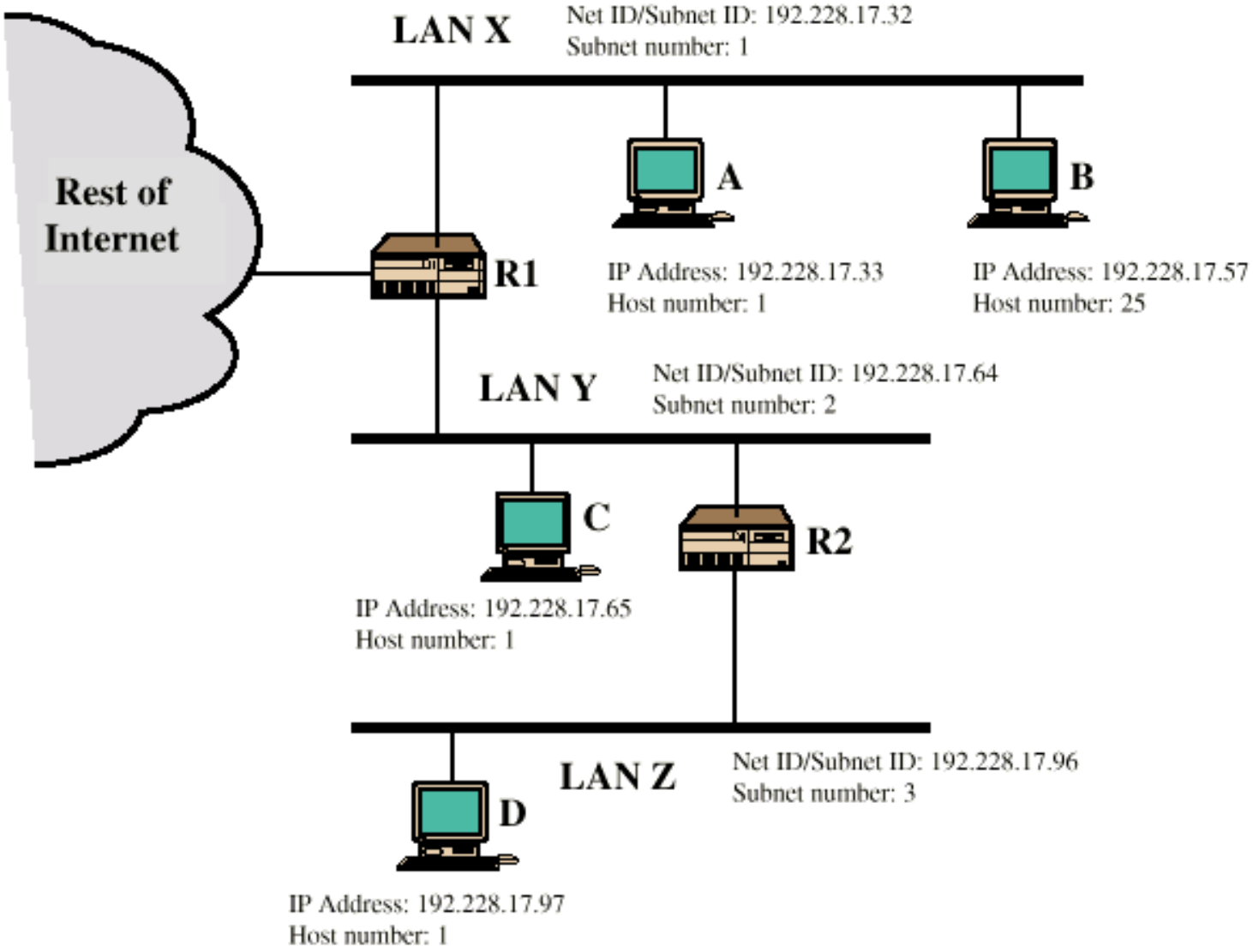
Subnets allow a big organization (e.g., university), who received a class A or B network address, to further divide its hosts into subnets (LANs):

- The host portion of address is further partitioned into subnet number and host number.
- Each LAN is assigned a subnet number (subnet address).
- To route an IP packet, its dest-IP addr is AND-ed with the subnet-mask, obtaining a subnet number. It is forwarded to the router of the dest-subnet:

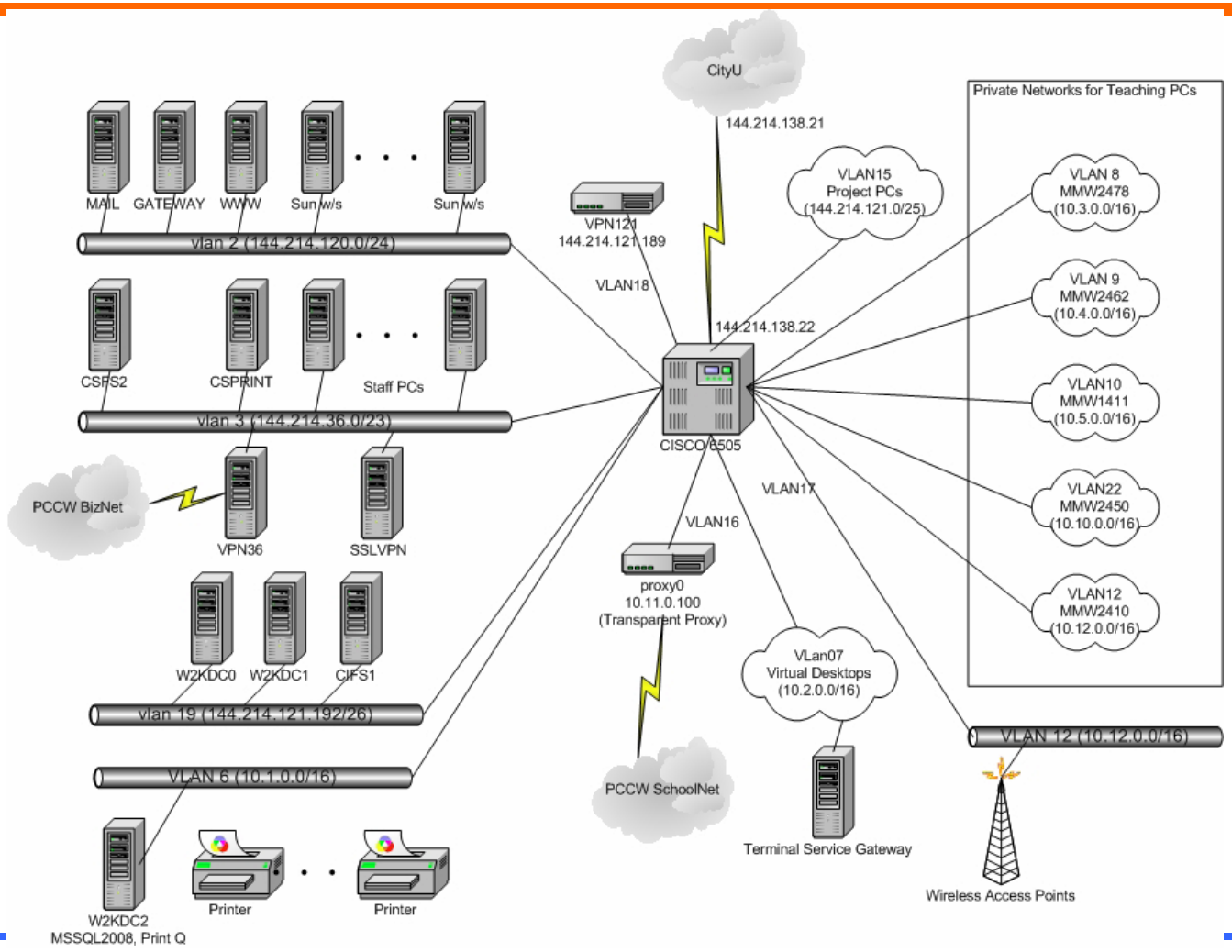


- The dest-subnet router routes the packet to dest-host using host #.

Routing Using Subnets



Cslab Subnet Interconnction



Routing Table in CSlab

- CSlab has 22 LANs connected to a single router. The router is connected to CSC's router.
- The routing table is very small, only contains 22 VLANs and an entry pointing to CSC router 144.214.138.21 (0.0.0.0 for *default routing*).

Default routing

```
S    169.254.0.0/16 is directly connected, Null0
     172.18.0.0/24 is subnetted, 1 subnets
C       172.18.120.0 is directly connected, Vlan21
     144.214.0.0/16 is variably subnetted, 8 subnets, 4 masks
C       144.214.122.0/26 is directly connected, Vlan5
C       144.214.121.0/25 is directly connected, Vlan15
……. (8 subnets)
C       144.214.121.192/26 is directly connected, Vlan19
     10.0.0.0/8, unregistered address, variably subnetted, 11 subnets, 2 masks
C       10.10.0.0/16 is directly connected, Vlan22
C       10.9.0.0/16 is directly connected, Vlan14
…….. (11 subnets)
S*   0.0.0.0/0 [1/0] via 144.214.138.21 // default router
```

CIDR-Classless InterDomain Routing

The classful addressing

- inefficient in use of address space, resulting in address space exhaustion
- e.g., it has to allocate a class B net address (for up to 65K hosts) to a network with only 1K host.

CIDR: subdivide a network addr (class A, B or C) into any length to make efficient use of IP addr space (this subdividing is visible to entire Internet).

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

Example of CIDR subdividing a class B net address: 194.24.0.0

University	Base address	Last address	# of hosts	Written as
Cambridge	194.24.0.0	194.24.7.255	2048	194.24.0.0/21
Edinburgh	194.24.8.0	194.24.11.255	1024	194.24.8.0/22
Oxford	194.24.16.0	194.24.31.255	4096	194.24.16.0/20

CIDR Routing

- Each entry in CIDR routing table contains a base address and a subnet mask (same as routing in subnets).
- For routing an IP packet, its destination IP addr is Boolean ANDed with the subnet mask to see if it matches the entry's base addr (the longest match is used if multiple matches found).
- The packet is routed out via the matched entry's outgoing link.
- Not all routers perform CIDR. (if all routers adopt CIDR, network portion of addresses can be of any length!)

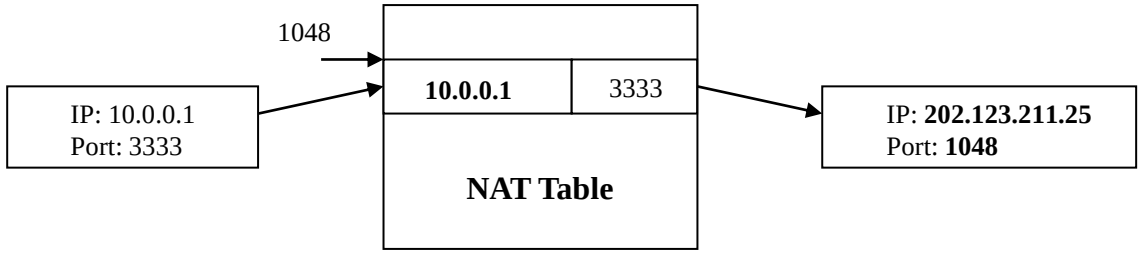
Routing table

	Base address	Mask	Out link
C	194.24.0.0/21	ff.ff.f8.00	Cambridge's router
E	194.24.8.0/22	ff.ff.fc.00	Edinburgh's router
O	194.24.16.0/20	ff.ff.f0.00	Oxford's router

NAT- Network Address Translation

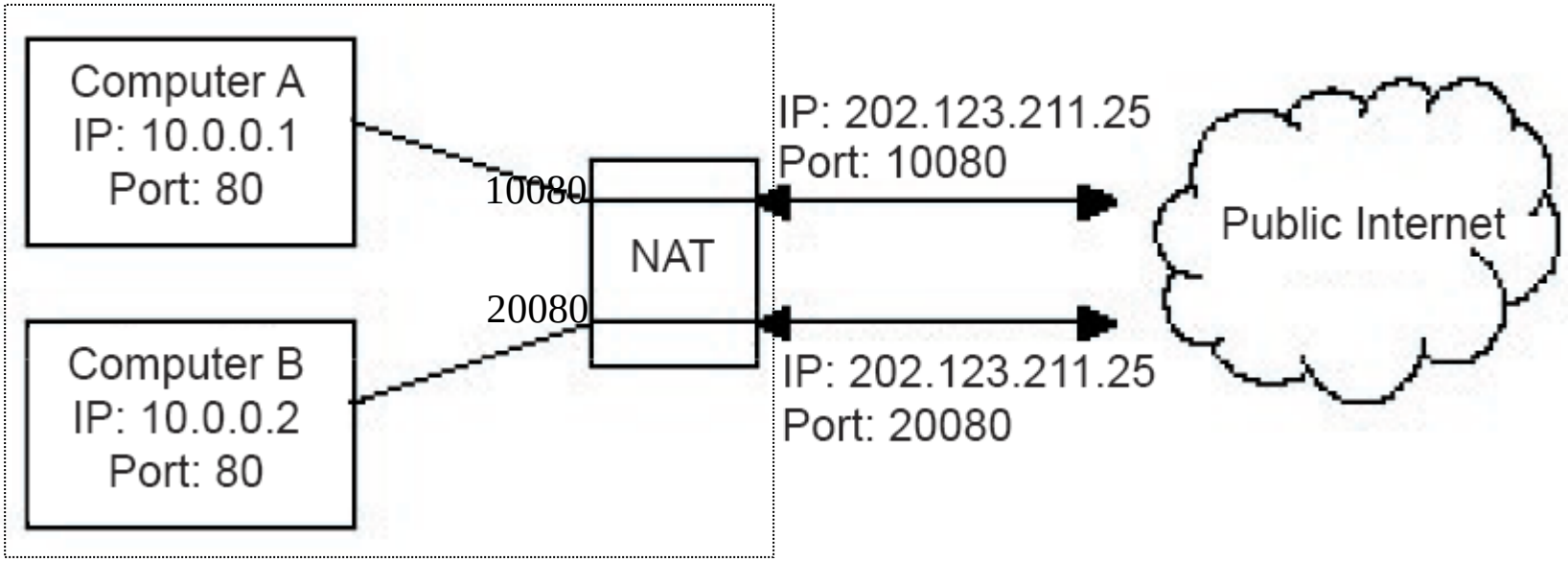
To relief the shortage of IP addresses, NAT technique uses only a small number of external IP addresses for communicating with outside, and hide local IP addresses from outside (so these local IP addresses can be re-used elsewhere):

- All Internet traffic to outside has to go through a NAT box, where the internal IP address (the source IP addr) is replaced by an external IP address.
- The **source port** # field in the IP packet is replaced by **the index** pointing to the entry in the NAT box's translation table. This entry contains the internal IP addr and the original source port #.
- When an outside IP packet arrives at the NAT box, its destination port # is extracted and used as an index to the translation table, where the original IP addr and port # are extracted and placed back to the packet.



NAT in operation

- Translate addresses when packets traverse through NAT.
- Use port numbers to identify internal IP users.
- This is a dirty way to fix the problem and only works for TCP/UDP traffic.



Dynamic Host Configuration Protocol (DHCP)

To make NAT work, three ranges of IP addresses have been declared for private use (called *unregistered addresses*):

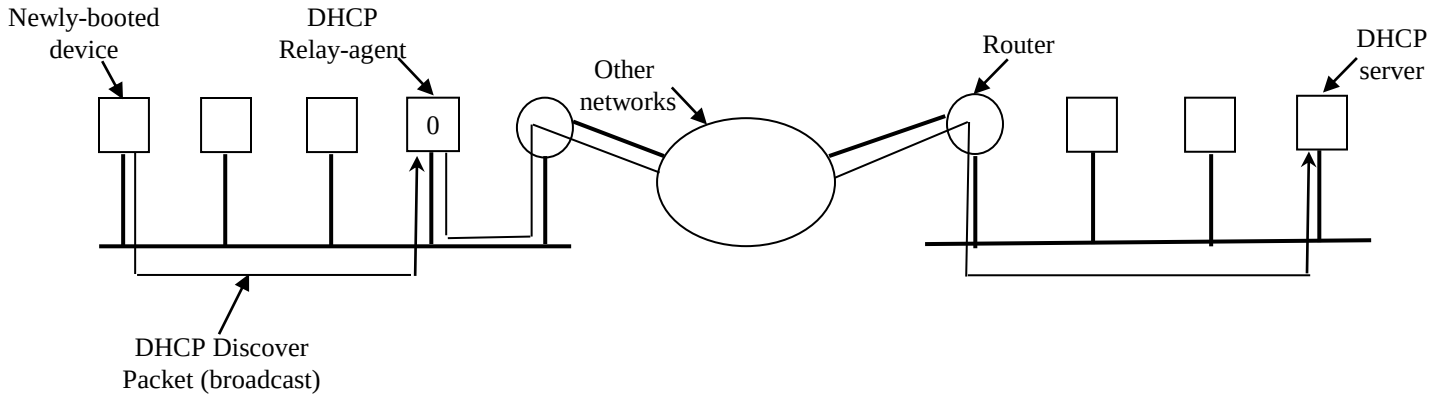
- 10.0.0.0 – 10.255.255.255/8
- 172.16.0.0 – 172.31.255.255/12
- 192.168.0.0 – 192.168.255.255/16

The allocation of the private IP addresses is done by Dynamic Host Configuration Protocol, a service running on a router:

- lease IP addresses for short time period (particularly for mobile users)
- hosts may refresh addresses periodically

DHCP in Operation

- A DHCP server may serve multiple LANs (subnets).
- A DHCP relay-agent is needed on each LAN to enable hosts on this LAN to reach DHCP server via a broadcast message.
- When a device (host) is booted up, it broadcasts a DHCP *discover* packet (to request / discover an IP address).
- DHCP *discover* packet is intercepted by the DHCP agent on the same LAN and forwarded to the DHCP server.



Address Resolution Problem

IP layer specifies the next hop router (destination) by IP addresses. When it gives an IP packet to the underlying network (e.g., Ethernet) for transmission, it needs to insert the "physical addresses" to the packet:

- How to determine the MAC addr of an IP addr?

Network Card Hardware Address (Ethernet address)

- Ethernet addresses (48-bits), in the format of:

  0e: 3c: 24: 3a: 03: 06

  are hard-wired into network cards at factories.
- IEEE allocates Ethernet *addr* to manufactures to ensure the uniqueness of Ethernet *addr*.
- Most network cards understand only hardware *addr*, not IP *addr* at all.

ARP- Address Resolution Protocol

ARP provides a mapping from IP addr to physical addr.

Suppose *A* knows *B*'s IP addr and wants to know *B*'s physical addr:

- *A* broadcasts ARP query packet, containing *B*'s IP address
  - all machines on the LAN receive ARP query
- *B* receives ARP packet, replies to *A* with its (*B*'s) physical layer address
- *A* caches (saves) IP-to-physical address pairs until information becomes old (times out)
  - soft state: information that times out (goes away) unless refreshed
- ARP shell command: arp

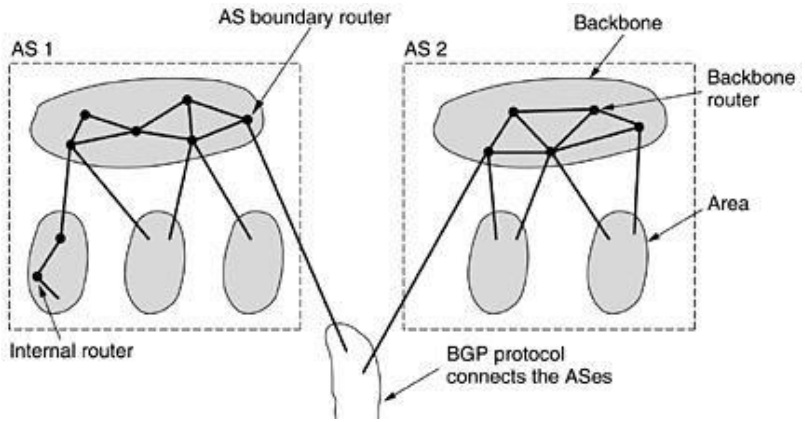
ARP Cache and RARP

- ARP maintains a cache of the recent IP to physical address mappings to make address translation fast
- Each entry is aged (usually the lifetime is 20 minutes) forcing periodic updates of the cache
- ARP replies are often broadcast so that all hosts can update their caches
- RARP (Reverse ARP) protocol allows a newly booted-up host to find its own IP address from its physical address (RARP was largely replaced by DHCP).

OSPF-Interior Gateway Routing Protocol

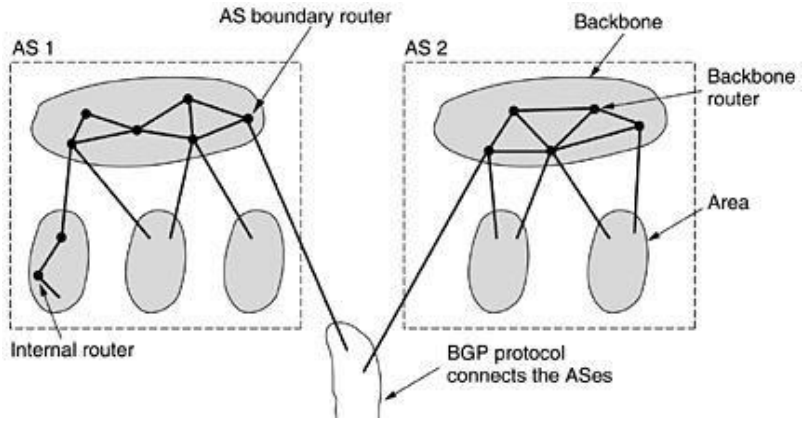
Internet consists of a large number of ASes (autonomous systems). Each AS, operated by an organization (or university), can use its own routing algorithm inside. Nevertheless, OSPF (open shortest path first) is a de facto routing standard for intra-AS routing.

- An AS is often divided into numbered areas (a subnet or a set of contiguous subnets).
- Each AS has a backbone, called area 0. At least one router in each area is connected to the backbone routers.
- Some backbone routers (called AS boundary routers) are connected to other ASes for inter-AS routing.



OSPF Algorithm

- OSPF abstracts the routers within the AS into a connected graph and assign weight to each edge (distance, delay, etc).
- Each router periodically floods *Link-State-Update* message to its adjacent routers, thus each router knows the topology within the area.
- Intra-area routing: each router uses the **link state** database and runs the **shortest path** to the destination router.
- Inter-area routing: the source router → backbone → the destination area → destination router
- Inter-AS routing, the source router → backbone → AS boundary router → destination AS boundary router (via BGP) → dest-router



BGP – Exterior Gateway Routing Protocol

Inter-AS traffics are routed by BGP (Border Gateway Protocol):

- Exterior gateway protocol concerns more on routing policies, such as no transit traffic through certain ASes, no relay for certain ASes, etc.
- BGP is a **distance vector** protocol. Each entry in routing table contains: dest, next-hop-to-dest, distance (# of hops). It always chooses the next hop neighbor with the shortest distance to a destination to forward the packet, called next-hop-routing.
- Each BGP router exchanges routing info with its neighbors (distances & next-hop to all destinations). After collecting the info from neighbors, a BGP router can update (build up) its own routing table.