Network Layer

1. Introduction

Main functions

- 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

1.1 Forwarding and Routing (转发和路由选择)

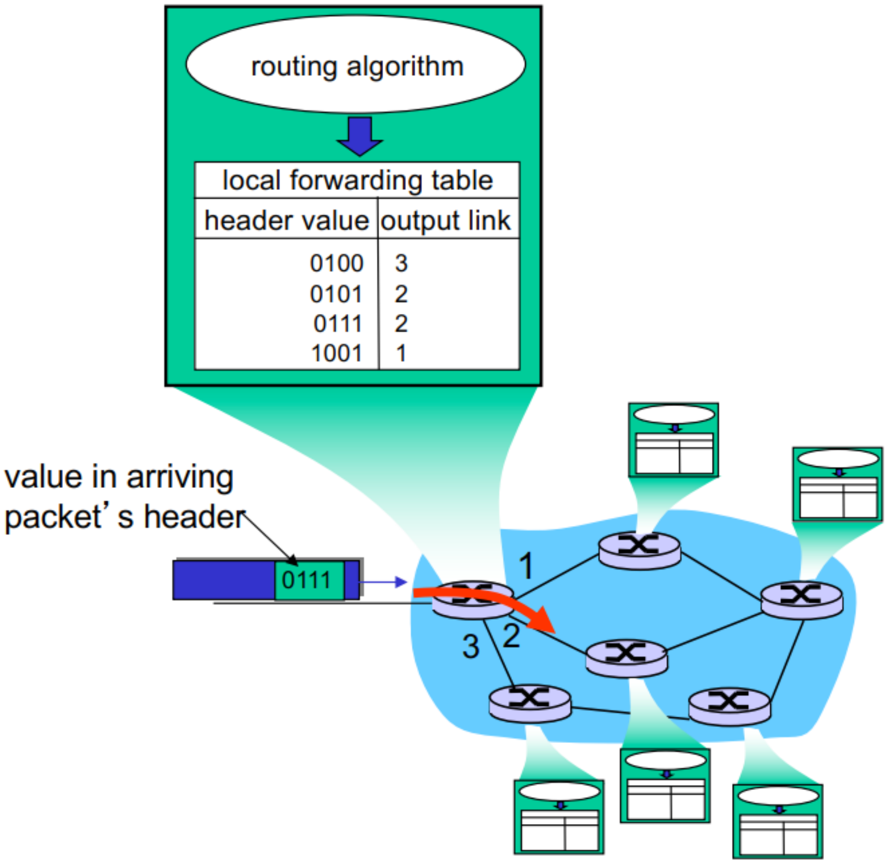
Two key network-layer functions

- **Forwarding** refers to the router-local action of transferring a packet from an input link interface to the appropriate output link interface. (将分组从输入链路移动至合适的输出链路)
- **Routing** refers to the network-wide process that determines the end-to-end paths that packets take from source to destination. (找路径)

analogy:

- **routing**: process of planning trip from source to destination
- **forwarding**: process of getting through single interchange

Forwarding table (转发表)



1.2 Connection Setup (建立连接)

- important function in some network architectures:
  - ATM (Asynchronous Transfer Mode) from telecom world
  - not in the Internet
- before datagrams flow, two end hosts and intervening routers establish virtual connection
  - routers get involved
  - must also keep state for each connection
  - not in the Internet

1.3 Network Service Model

- Guaranteed delivery
- Guaranteed delivery with bounded delay
- In-order packet delivery
- Guaranteed minimal bandwidth
- Guaranteed maximum jitter
- Security services

2 Virtual Circuit and Datagram Networks (虚拟电路和数据报网络)

Computer networks that provide only a **connection service** at the network layer are called **virtual-circuit (VC) networks**; computer networks that provide only a **connectionless service** at the network layer are called **datagram networks.**

2.1 Virtual Circuit

While the Internet is a datagram network, many alternative network architectures use connections at the network layer. These network-layer connections are called **virtual circuits (VCs)**.

A VC consists of:

1. **a path** (that is, a series of links and routers) between the source and destination hosts
2. **VC numbers**, one number for each link along the path
3. **entries** in the forwarding table in each router along the path (沿着该路径的每台路由器的转发表的表项)

   A packet belonging to a virtual circuit will carry a VC number in its header. Because a virtual circuit may have a different VC number on each link, each intervening router (中间路由器) must replace the VC number of each traversing packet with a new VC number. The new VC number is obtained from the forwarding table.

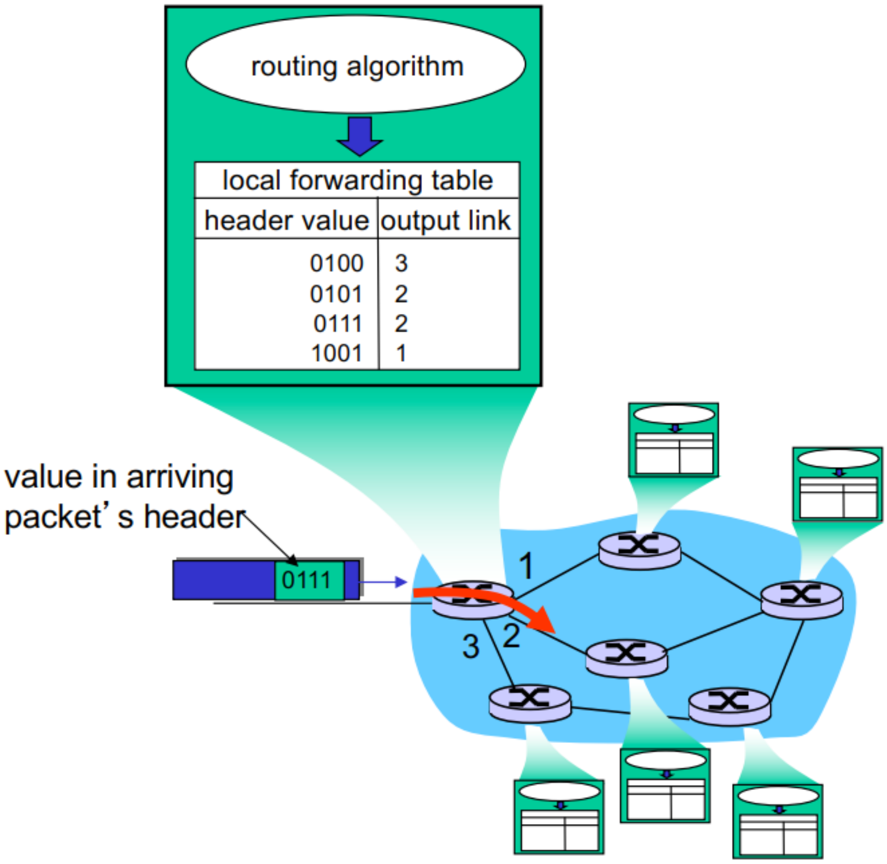
2.2 Datagram Networks

In a datagram network, each time an end system wants to send a packet, it stamps the packet with the address of the destination end system and then pops the packet into the network (加上目的端系统的地址).

2.2.1 Datagram forwarding table

As a packet is transmitted from source to destination, it passes through a series of routers. Each of these routers uses the packet’s destination address to forward the packet.

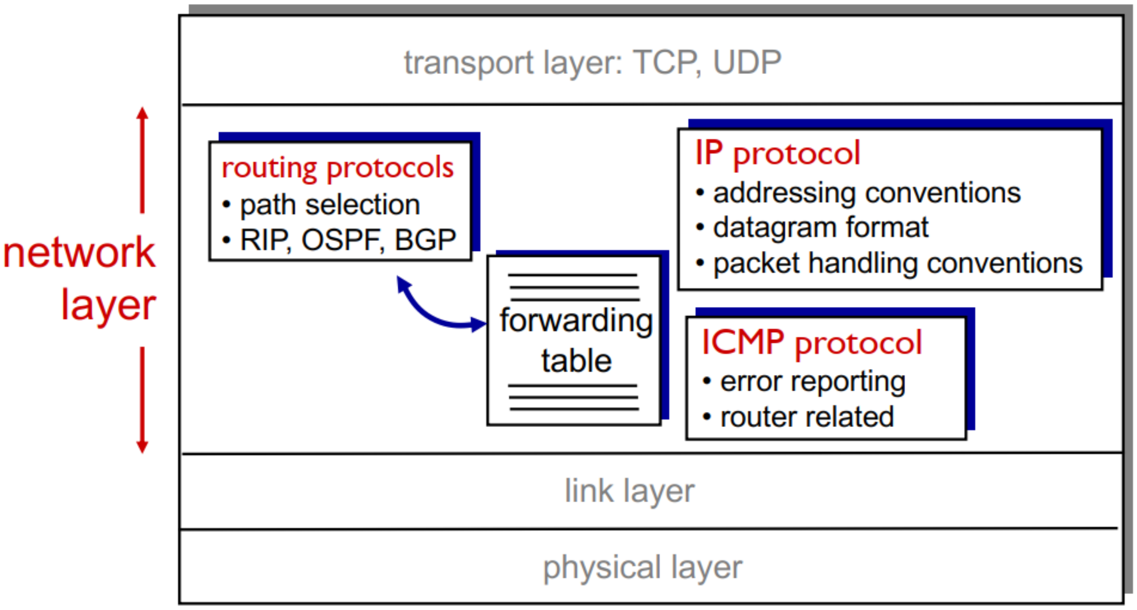
Specifically, each router has a forwarding table that maps destination addresses to link interfaces; when a packet arrives at the router, the router uses the packet’s destination address to look up the appropriate output link interface in the forwarding table. The router then intentionally forwards the packet to that output link interface.



Longest prefix matching (最长前缀匹配原则)

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

3. The Internet Protocol (IP) 网际协议



Internet's network layer has three major components.

1. IP protocol

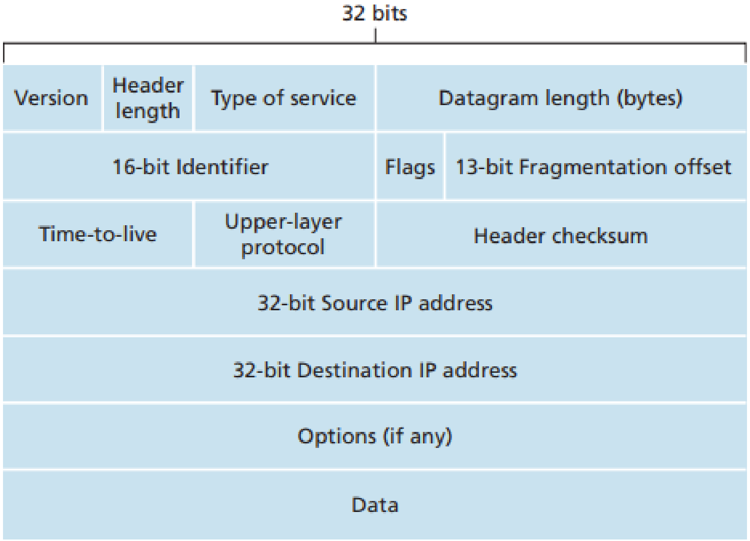
2. Routing component, which determines the path a datagram follows from source to destination.

   Routing protocols compute the forwarding tables that are used to forward packets through the network.

3. A facility (ICMP protocol) to report errors in datagrams and respond to requests for certain network-layer information

3.1 Datagram Format 数据报格式

Take IPv4 datagram format for example:



The key fields in the IPv4 datagram:

- **Version number (4 bits) 版本号**

  specify the IP protocol version of the datagram

  By looking at the version number, the router can determine how to interpret the remainder of the IP datagram.

- **Header length (bytes) 首部长度**
- **Type of service**
- **Datagram length**

  This is the total length of the IP datagram (header plus data), measured in bytes.

  Since this field is 16 bits long, the theoretical maximum size of the IP datagram is 65,535 bytes.

- **Identifier, flags, fragmentation offset**
- **Time-to-live**

  The time-to-live (**TTL**) field is included to ensure that datagrams do not circulate forever (due to, for example, a long-lived routing loop) in the network. **This field is decremented by one each time the datagram is processed by a router. If the TTL field reaches 0, the datagram must be dropped.**

- **Protocol**

  The value of this field indicates the specific transport-layer protocol to which the data portion of this IP datagram should be passed.

- **Header checksum**

  aids a router in detecting bit errors in a received IP datagram

- **Source and destination IP addresses**

- **Options**

  The options fields allow an IP header to be extended.

  the amount of time needed to process an IP datagram at a router can vary greatly. These considerations become particularly important for IP processing in high-performance routers and hosts.

  For these reasons and others, IP options were dropped in the IPv6 header.

- **Data (payload)**

  In most circumstances, the data field of the IP datagram contains the transport-layer segment (TCP or UDP) to be delivered to the destination.

3.2 IP Datagram Fragmentation, Reassembly (IP数据报分片组装)

- network links have MTU (max.transfer unit), 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 and order fragments to reassemble

Example:

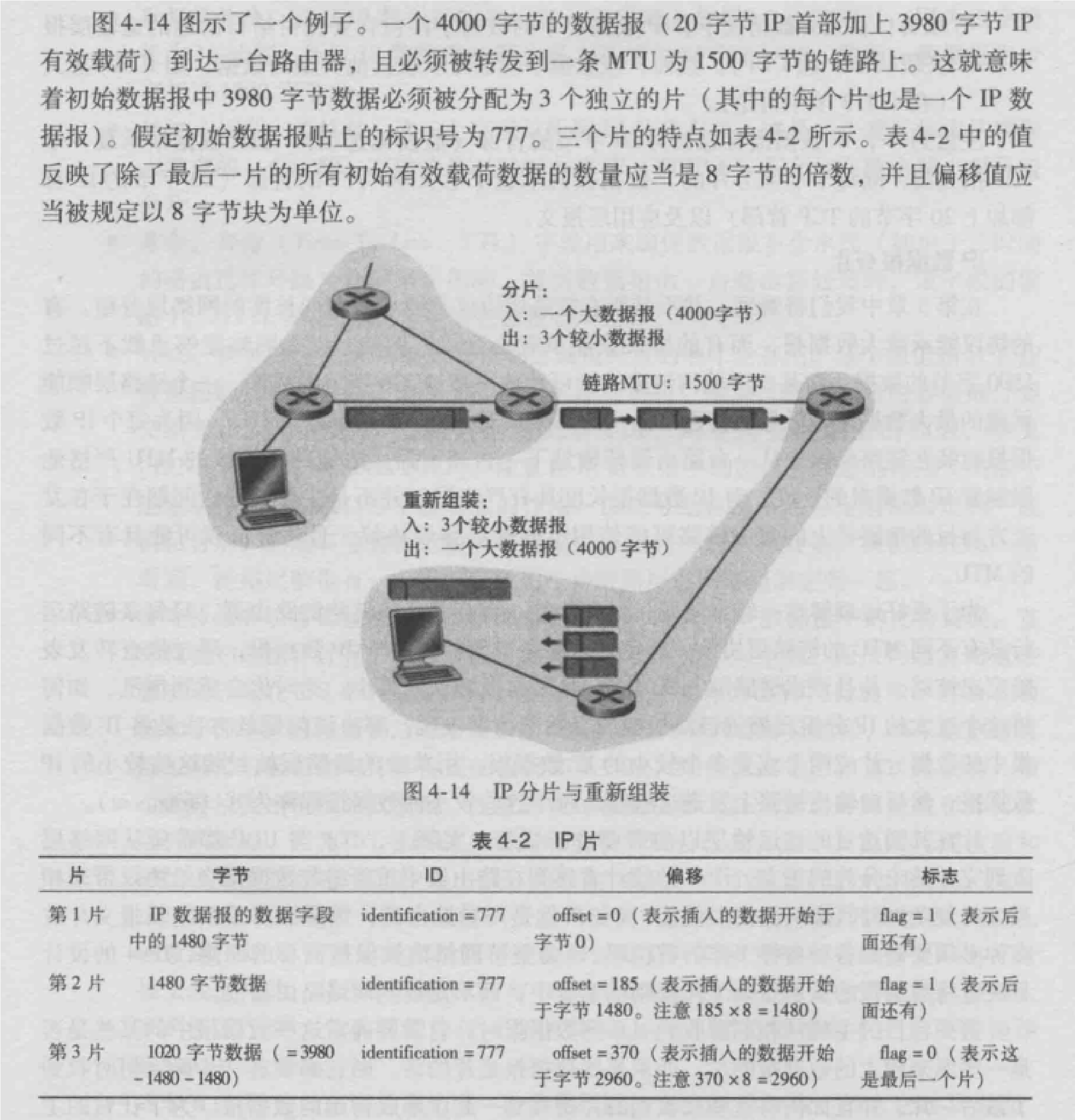


图 4-14 图示了一个例子。一个 4000 字节的数据报（20 字节 IP 首部加上 3980 字节 IP 有效载荷）到达一台路由器，且必须被转发到一条 MTU 为 1500 字节的链路上。这就意味着初始数据报中 3980 字节数据必须被分配为 3 个独立的片（其中的每个片也是一个 IP 数据报）。假定初始数据报贴上的标识号为 777。三个片的特点如表 4-2 所示。表 4-2 中的值反映了除了最后一片的所有初始有效载荷数据的数量应当是 8 字节的倍数，并且偏移值应当被规定以 8 字节块为单位。

分片：
入：一个大数据报（4000字节）
出：3个较小数据报
链路MTU：1500 字节
重新组装：
入：3个较小数据报
出：一个大数据报（4000 字节）

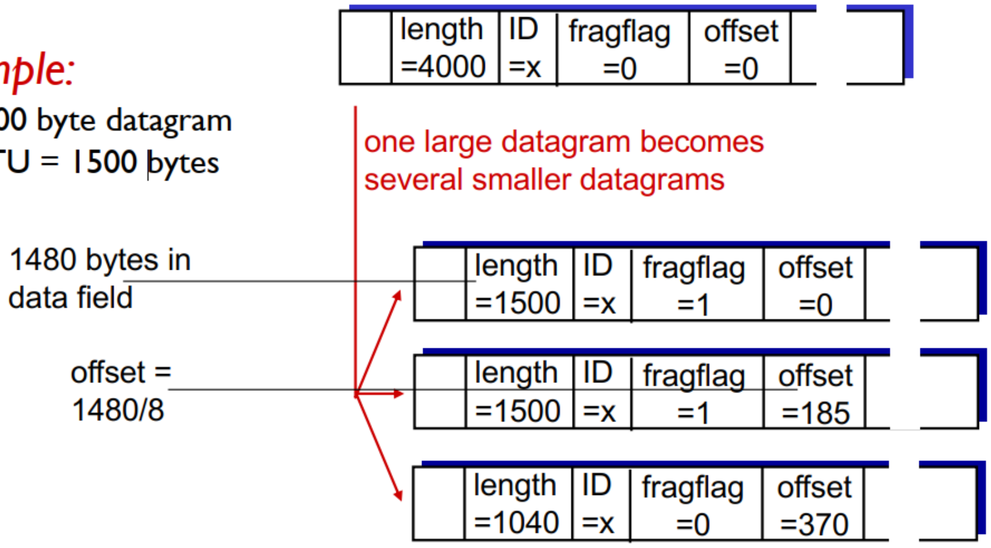
图 4-14　IP 分片与重新组装

表 4-2　IP 片

片	字节	ID	偏移	标志
第 1 片	IP 数据报的数据字段中的 1480 字节	identification = 777	offset = 0（表示插入的数据开始于字节 0）	flag = 1（表示后面还有）
第 2 片	1480 字节数据	identification = 777	offset = 185（表示插入的数据开始于字节 1480。注意 185 × 8 = 1480）	flag = 1（表示后面还有）
第 3 片	1020 字节数据（= 3980 - 1480 - 1480）	identification = 777	offset = 370（表示插入的数据开始于字节 2960。注意 370 × 8 = 2960）	flag = 0（表示这是最后一个片）

example:

- 4000 byte datagram
- MTU = 1500 bytes



At the destination, the payload of the datagram is passed to the transport layer only after the IP layer has fully reconstructed the original IP datagram. If one or more of the fragments does not arrive at the destination, the incomplete datagram is discarded and not passed to the transport layer.

However, if TCP is being used at the transport layer, then TCP will recover from this loss by having the source retransmit the data in the original datagram.

3.3 IPv4 Addressing 编址

A host typically has only a single link into the network; when IP in the host wants to send a datagram, it does so over this link.

The boundary between the host and the physical link is called an **interface**.

- host typically has one or two interfaces (e.g., wired Ethernet, wireless 802.11)

The boundary between the router and any one of its links is also called an **interface**.

- routers typically have multiple interfaces

IP addresses are associated with each interface.

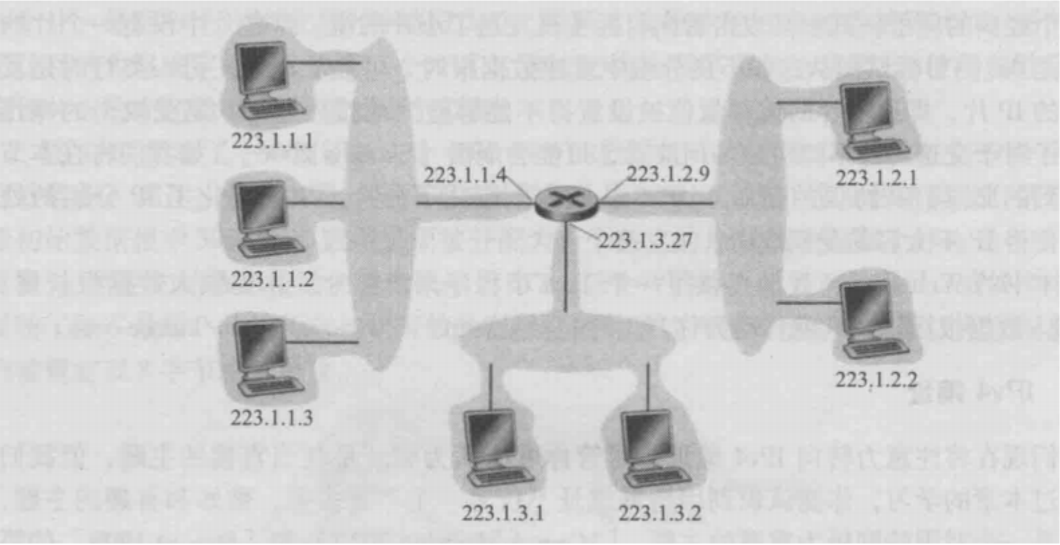
Each IP address is 32 bits long (4 bytes). In total, there are 2^32 possible IP addresses.

These addresses are typically written in **dotted-decimal notation (点分十进制记方法)**.

e.g.

223.1.1.1 = 11011111 00000001 00000001 00000001

3.3.1 Subnet (子网)



The three hosts in the upper-left portion of the figure above, and the router interface to which they are connected, all have an IP address of the form **223.1.1.xxx**. That is, they all have the **same leftmost 24 bits** in their IP address.

In IP terms, this network interconnecting three host interfaces and one router interface forms a **subnet** . (A subnet is also called an IP network or simply a network in the Internet literature.)

IP addressing assigns an address to this subnet: 223.1.1.0/24, where the **/24 notation**, sometimes known as a **subnet mask (子网掩码)**, indicates that the **leftmost 24 bits of the 32-bit quantity define the subnet address**.

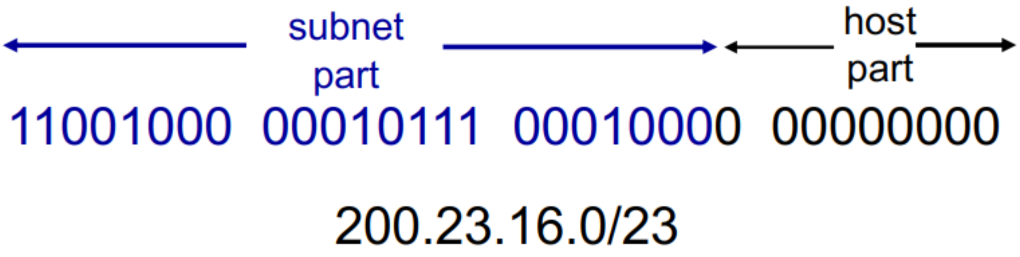
- IP address
    - subnet part: high order bits
    - host part: low order bits
- what's a subnet ?
    - device interfaces with same subnet part of IP address
    - can physically reach each other **without intervening router**

3.3.2 Classless Interdomain Routing (CIDR) 无类别域间路由选择

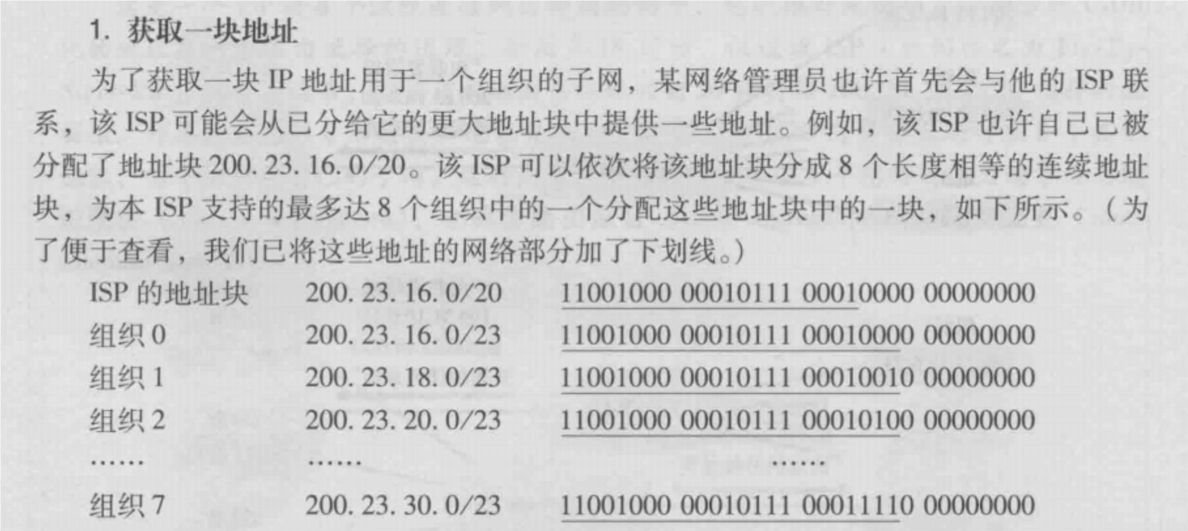
CIDR is the Internet's address assignment strategy.

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

Example:



3.3.3 IP addresses: how to get one?



1. **获取一块地址**

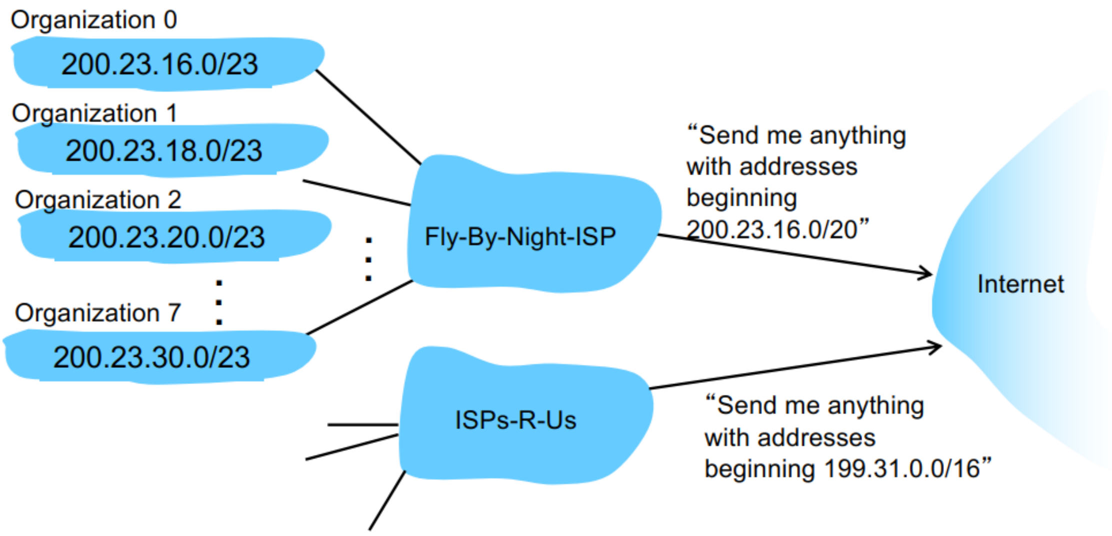
为了获取一块 IP 地址用于一个组织的子网，某网络管理员也许首先会与他的 ISP 联系，该 ISP 可能会从已分给它的更大地址块中提供一些地址。例如，该 ISP 也许自己已被分配了地址块 200. 23. 16. 0/20。该 ISP 可以依次将该地址块分成 8 个长度相等的连续地址块，为本 ISP 支持的最多达 8 个组织中的一个分配这些地址块中的一块，如下所示。（为了便于查看，我们已将这些地址的网络部分加了下划线。）

ISP 的地址块	200. 23. 16. 0/20	11001000 00010111 00010000 00000000
组织 0	200. 23. 16. 0/23	11001000 00010111 00010000 00000000
组织 1	200. 23. 18. 0/23	11001000 00010111 00010010 00000000
组织 2	200. 23. 20. 0/23	11001000 00010111 00010100 00000000
……	……	……
组织 7	200. 23. 30. 0/23	11001000 00010111 00011110 00000000

In order to obtain a block of IP addresses for use within an organization's subnet, a network administrator might first contact its ISP, which would provide addresses from a larger block of addresses that had already been allocated to the ISP. For example, the ISP may itself have been allocated the address block 200.23.16.0/20. The ISP, in turn, could divide its address block into eight equal-sized contiguous address blocks and give one of these address blocks out to each of up to eight organizations that are supported by this ISP. (We have underlined the subnet part of these addresses for your convenience.)

3.3.4 Hierarchical addressing: route aggregation

Example:



This example of an ISP that connects eight organizations to the Internet nicely illustrates how carefully allocated CIDRized addresses facilitate routing. Suppose, as shown in the Figure, that the ISP (which we'll call Fly-By-Night-ISP) advertises to the outside world that it should be sent any datagrams whose first 20 address bits match 200.23.16.0/20. The rest of the world need not know that within the address block 200.23.16.0/20 there are in fact eight other organizations, each with its own subnets. This ability to use a single prefix to advertise multiple networks is often referred to as address aggregation (also route aggregation or route summarization).

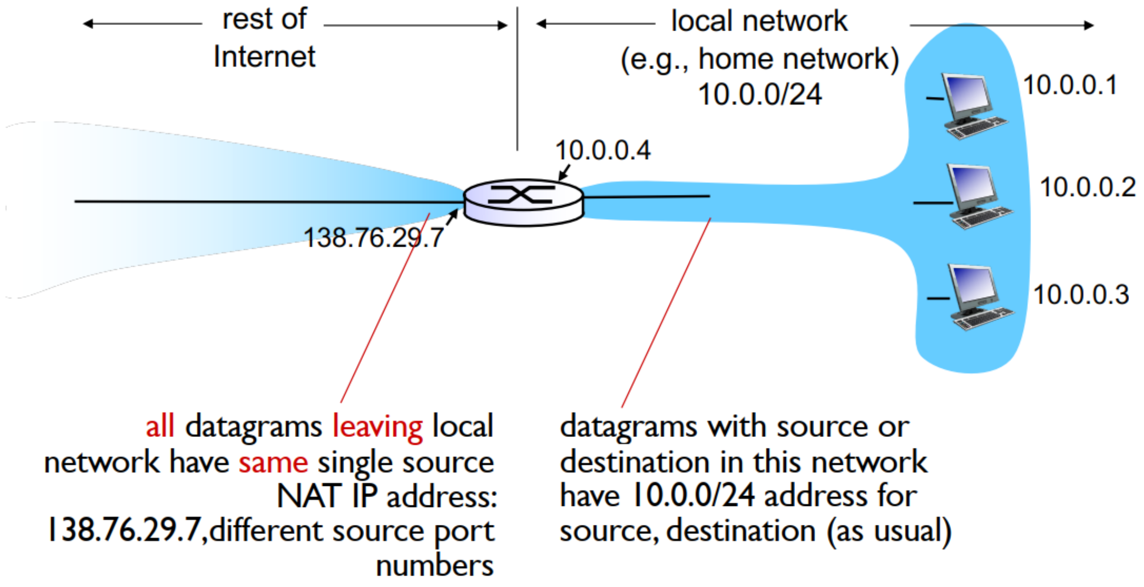
How does an ISP get a block of addresses?
ICANN: Internet Corporation for Assigned Names and Numbers

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

3.4 Network Address Translation (NAT) 网络地址转换

LAN (local area network)&**WAN** (wide area network)

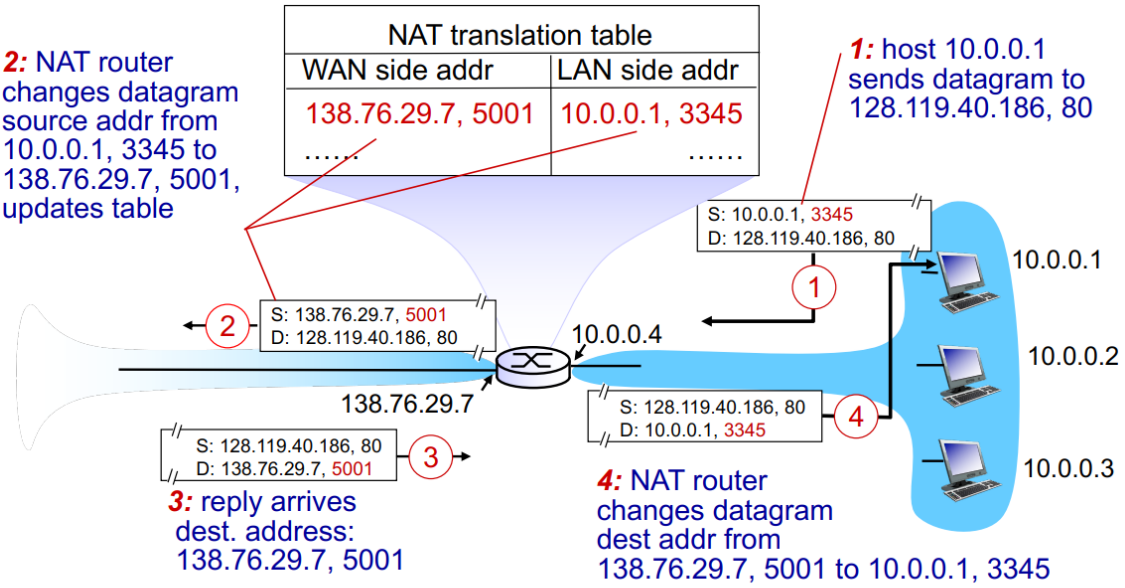
从局域网到广域网


rest of Internet
local network (e.g., home network) 10.0.0/24
10.0.0.1
10.0.0.2
10.0.0.3
10.0.0.4
138.76.29.7
all datagrams leaving local network have same single source NAT IP address: 138.76.29.7,different source port numbers
datagrams with source or destination in this network have 10.0.0/24 address for source, destination (as usual)

- 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

- **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

Example:



NAT translation table

WAN side addr	LAN side addr
138.76.29.7, 5001	10.0.0.1, 3345
……	……

1: host 10.0.0.1 sends datagram to 128.119.40.186, 80

2: NAT router changes datagram source addr from 10.0.0.1, 3345 to 138.76.29.7, 5001, updates table

3: reply arrives dest. address: 138.76.29.7, 5001

4: NAT router changes datagram dest addr from 138.76.29.7, 5001 to 10.0.0.1, 3345