第18章 网际互联协议

南京大学计算机系　　黄皓教授

2007年9月21日 星期五
|
2007年9月25日 星期二

Reference

- **TCP/IP Tutorial and Technical Overview, ibm.com/**redbooks

Contents

- 协议的基本功能
- 网络层提供的服务
- IP协议
  - IP
  - ICMP
  - ARP
  - RARP
  - DHCP
- 路由算法
- 路由协议
  - IGMP
  - RIP
  - EGP
  - BGP
  - OSPF
- 拥塞控制
- 服务质量

18.1 Protocol Functions

- Small set of functions that form basis of all protocols
- Not all protocols have all functions
  - Reduce duplication of effort
  - May have same type of function in protocols at different levels

18.1 Protocol Functions

(1) Encapsulation

(2) Fragmentation and reassembly

(3) Connection control

(4) Ordered delivery

(5) Flow control

(6) Error control

(7) Addressing

(8) Multiplexing

(9) Transmission services

(1) Encapsulation

- Data usually transferred in blocks
  - Protocol data units (PDUs)
  - Each PDU contains data and control information
  - Some PDUs only control

(1) Encapsulation

- Three categories of control
    - Address
        - Of sender and/or receiver
    - Error-detecting code
        - E.g. frame check sequence
    - Protocol control
        - Additional information to implement protocol functions

(1) Encapsulation

- Addition of control information to data is encapsulation
- Data accepted or generated by entity and encapsulated into PDU
  - Containing data plus control information
  - e.g. TFTP, HDLC, frame relay, ATM, AAL5, LLC, IEEE 802.3, IEEE 802.11

(2) Fragmentation and Reassembly

- Exchange data between two entities
- Characterized as sequence of PDUs of some bounded size
  - Application level message

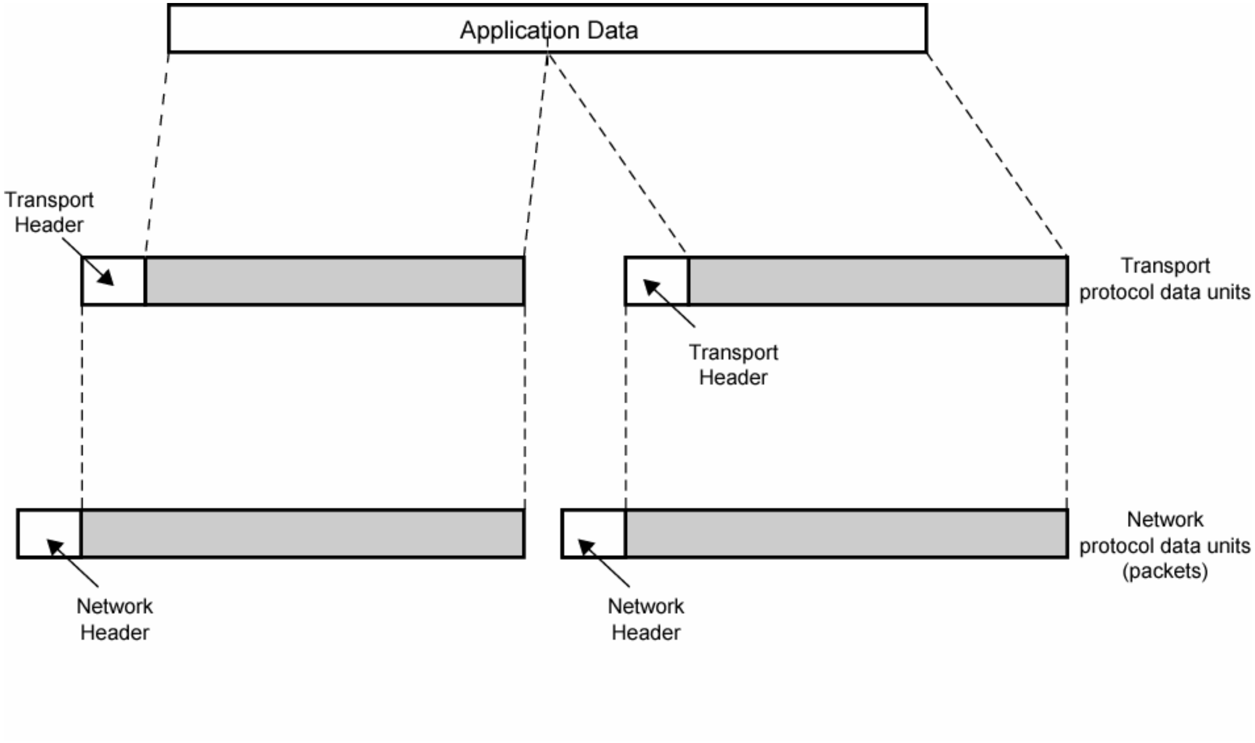
- Lower-level protocols may need to break data up into smaller blocks. Advantages of Fragmentation:
  - Communications network may only accept blocks of up to a certain size
    - ATM 53 octets
    - Ethernet 1526 octets
  - More efficient error control
    - Smaller retransmission
  - Fairer
    - Prevent station monopolizing medium
  - Smaller buffers
  - Provision of checkpoint and restart/recovery operations

(2) Fragmentation and Reassembly

- Disadvantages of Fragmentation
  - Make PDUs as large as possible because
    - PDU contains some control information
    - Smaller block, larger overhead
  - PDU arrival generates interrupt
    - Smaller blocks, more interrupts
  - More time processing smaller, more numerous PDUs

(2) Fragmentation and Reassembly

- Segmented data must be reassembled into messages
- More complex if PDUs out of order



(3) Connection Control

- Connectionless data transfer
  - Each PDU treated independently
  - E.g. datagram
- Connection-oriented data transfer
  - E.g. virtual circuit
- Connection-oriented preferred (even required) for
  - lengthy exchange of data
  - Or if protocol details must be worked out dynamically

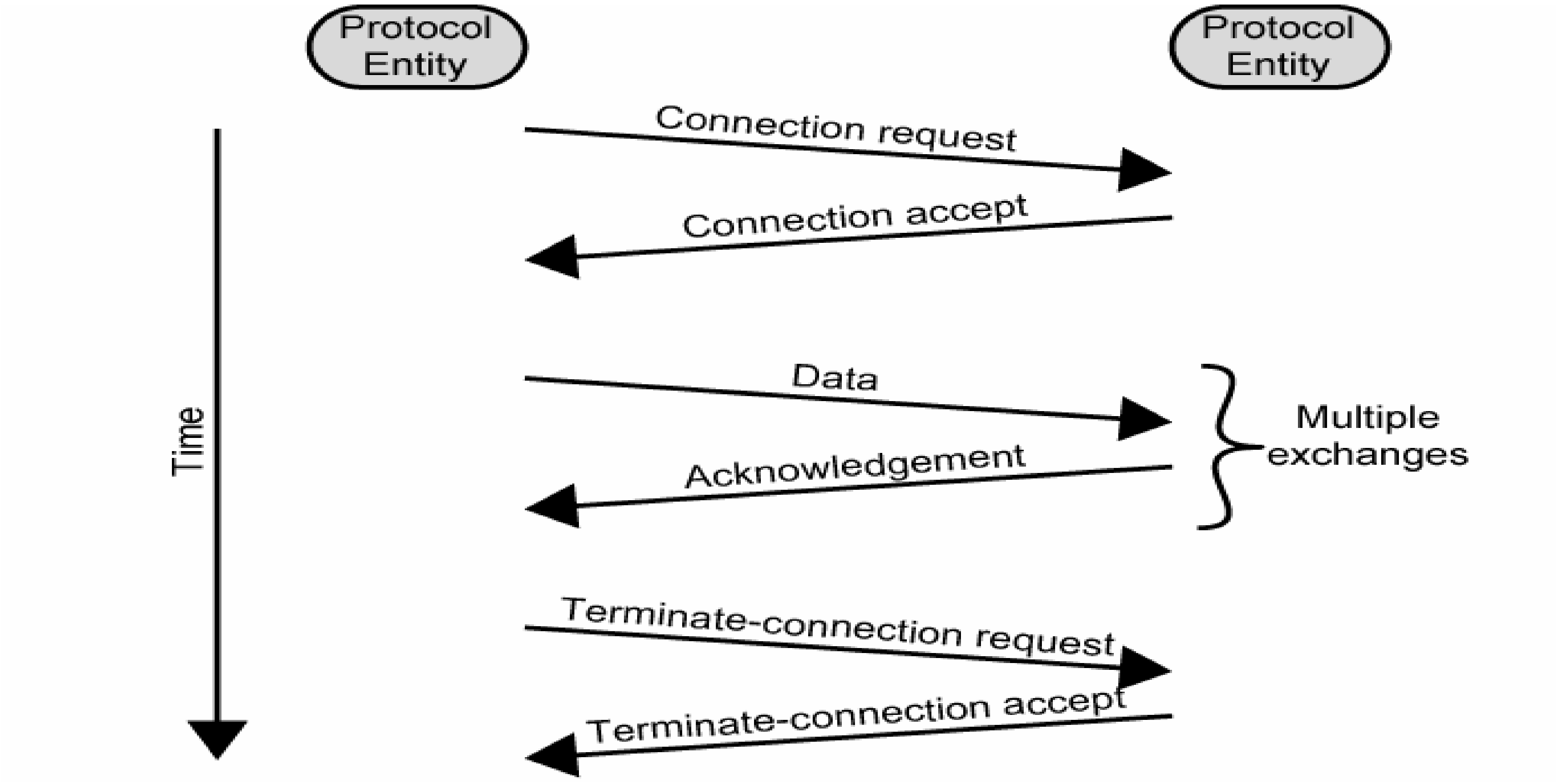
(3) Connection Control

- Logical association, or connection, established between entities
- Three phases occur
  - Connection establishment
  - Data transfer
  - Connection termination

May be interrupt and recovery phases to handle errors

(3) Connection Control

- Phases of Connection Oriented Transfer



(3) Connection Control

- Connection Establishment
  - Entities agree to exchange data
  - Typically, one station issues connection request (In connectionless fashion)
    - May involve central authority
    - Receiving entity accepts or rejects (simple)
  - May include negotiation
    - Syntax, semantics, and timing
    - Both entities must use same protocol
    - May allow optional features
    - Must be agreed E.g. protocol may specify max PDU size 8000 octets; one station may wish to restrict to 1000 octets

(3) Connection Control

- Data Transfer and Termination
  - Both data and control information exchanged
    - e.g. flow control, error control
  - Data flow and acknowledgements may be in one or both directions
  - One side may send termination request
  - Or central authority might terminate

(3) Connection Control

- Sequencing
  - Many connection-oriented protocols use sequencing
    - e.g. HDLC, IEEE 802.11, tcp
  - PDUs numbered sequentially
  - Each side keeps track of outgoing and incoming numbers
  - Supports three main functions
    - Ordered delivery
    - Flow control
    - Error control
  - Not found in all connection-oriented protocols
    - E.g.frame relay and ATM
  - All connection-oriented protocols include some way of identifying connection
    - Unique connection identifier
    - Combination of source and destination addresses

(4) Ordered Delivery

- PDUs may arrive out of order
  - Different paths through network
- PDU order must be maintained
- Number PDUs sequentially
- Easy to reorder received PDUs
- Finite sequence number field
  - Numbers repeat modulo maximum number
  - Maximum sequence number greater than maximum number of PDUs that could be outstanding
  - In fact, maximum number may need to be twice maximum number of PDUs that could be outstanding
    - e.g. selective-repeat ARQ

(5) Flow Control

- Performed by receiving entity to limit amount or rate of data sent
- Stop-and-wait
  - Each PDU must be acknowledged before next sent
- Credit
  - Amount of data that can be sent without acknowledgment
  - E.g. HDLC sliding-window
- Must be implemented in several protocols
  - Network traffic control
  - Buffer space
  - Application overflow
    - E.g. waiting for disk access

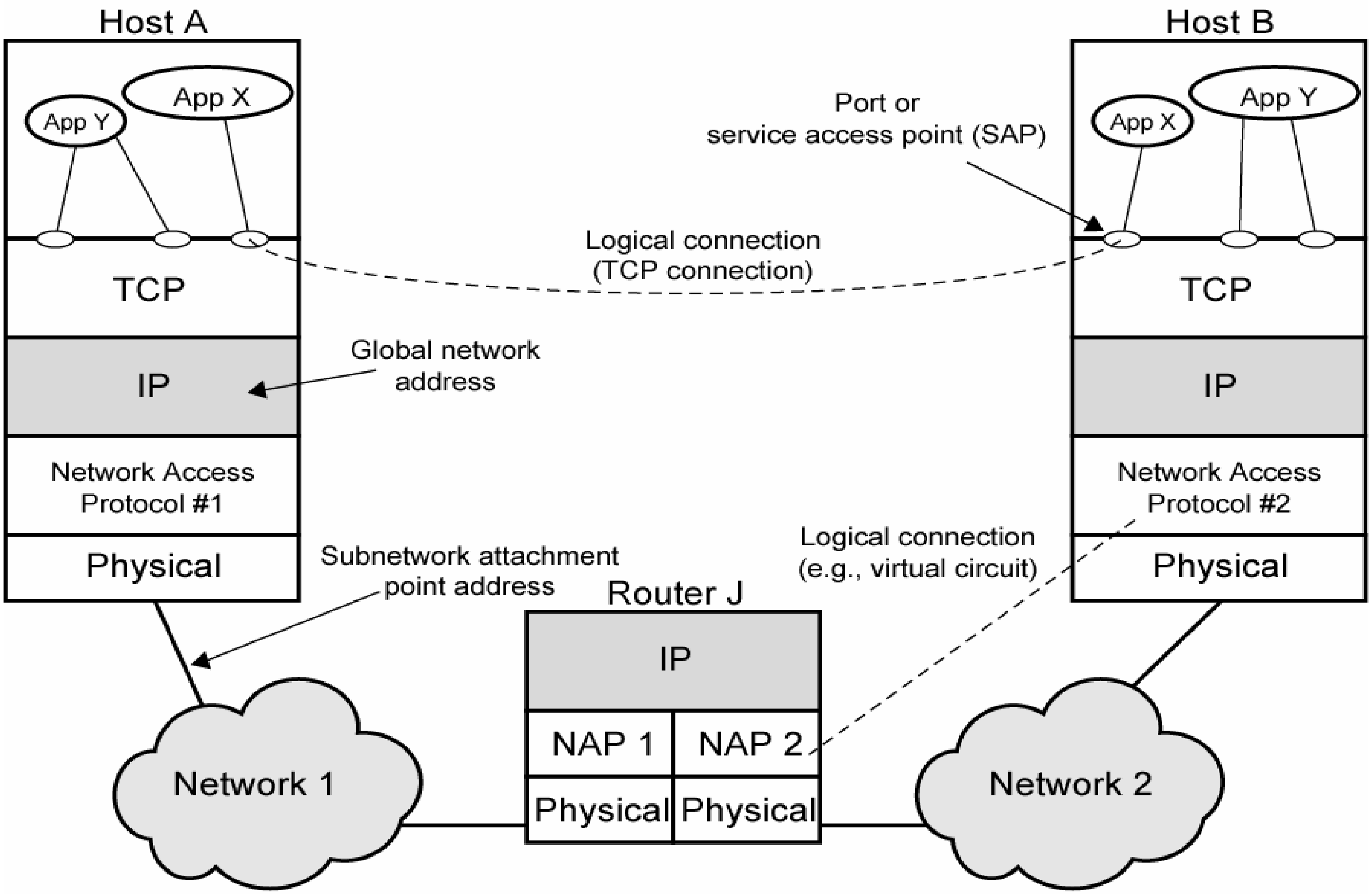
(6) Error Control

- Guard against loss or damage
- Error detection and retransmission
  - Sender inserts error-detecting code in PDU
    - Function of other bits in PDU
  - Receiver checks code on incoming PDU
  - If error, discard
  - If transmitter doesn’t get acknowledgment in reasonable time, retransmit
- Error-correction code
  - Enables receiver to detect and possibly correct errors
- Error control is performed at various layers of protocol
  - Between station and network
  - Inside network

(7) Addressing

- Addressing level
- Addressing scope
- Connection identifiers
- Addressing mode

TCP/IP Concepts


Host A
App X
App Y
TCP
IP
Network Access Protocol #1
Physical
Port or service access point (SAP)
Logical connection (TCP connection)
Global network address
Subnetwork attachment point address
Router J
IP
NAP 1
NAP 2
Physical
Physical
Logical connection (e.g., virtual circuit)
Host B
App Y
App X
TCP
IP
Network Access Protocol #2
Physical
Network 1
Network 2

Addressing Level

- Level in comms architecture at which entity is named
- Unique address for each end system
  - e.g. workstation or server
- And each intermediate system
  - (e.g., router)
- Network-level address
  - IP address or internet address
  - OSI - network service access point (NSAP)
  - Used to route PDU through network
- At destination data must routed to some process
  - Each process assigned an identifier
  - TCP/IP port
  - Service access point (SAP) in OSI

Addressing Scope

- Global address
  - Global nonambiguity
  - Identifies unique system
  - Synonyms permitted
  - System may have more than one global address
  - Global applicability
  - Possible at any global address to identify any other global address, in any system, by means of global address of other system
  - Enables internet to route data between any two systems

Addressing Scope

- Need unique address for each device interface on network
  - MAC address on IEEE 802 network and ATM host address
  - Enables network to route data units through network and deliver to intended system
  - Network attachment point address
- Addressing scope only relevant for network-level addresses
- Port or SAP above network level is unique within system
  - Need not be globally unique
  - E.g port 80 web server listening port in TCP/IP

Connection Identifiers

- Connection identifier used by both entities for future transmissions
  - Reduced overhead
    - Generally shorter than global identifiers
  - Routing
    - Fixed route may be defined
    - Connection identifier identifies route to intermediate systems
  - Multiplexing
    - Entity may wish more than one connection simultaneously
    - PDUs must be identified by connection identifier
  - Use of state information
    - Once connection established, end systems can maintain state information about connection
      - Flow and error control using sequence numbers

Addressing Mode

- Usually address refers to single system or port
  - □ Individual or unicast address
- Address can refer to more than one entity or port

  Multiple simultaneous recipients for data
  - □ Broadcast for all entities within domain
  - □ Multicast for specific subset of entities

(8) Multiplexing

- Multiple connections into single system
  - E.g. frame relay, can have multiple data link connections terminating in single end system
  - Connections multiplexed over single physical interface
- Can also be accomplished via port names
  - Also permit multiple simultaneous connections
  - E.g. multiple TCP connections to given system
    - Each connection on different pair of ports

Multiplexing Between Levels

- Upward or inward multiplexing
  - Multiple higher-level connections share single lower-level connection
    - More efficient use of lower-level service
    - Provides several higher-level connections where only single lower-level connection exists
- Downward multiplexing, or splitting
  - Higher-level connection built on top of multiple lower-level connections
  - Traffic on higher connection divided among lower connections
    - Reliability, performance, or efficiency.

(9) Transmission Services

- Protocol may provide additional services to entities
- E.g.:
  - Priority
    - Connection basis
    - On message basis
      - E.g. terminate-connection request
  - Quality of service
    - E.g. minimum throughput or maximum delay threshold
  - Security
    - Security mechanisms, restricting access
  - These services depend on underlying transmission system and lower-level entities

18.2 Principle of Internetworking

- Internetworking Terms
- Requirements of Internetworking
- Architectural Approaches

(1) Internetworking Terms (1)

- Communications Network
  - Facility that provides data transfer service
- An internet
  - Collection of communications networks interconnected by bridges and/or routers
- The Internet - note upper case I
  - The global collection of thousands of individual machines and networks
- Intranet
  - Corporate internet operating within the organization
  - Uses Internet (TCP/IP and http)technology to deliver documents and resources

(1) Internetworking Terms (2)

- End System (ES)
  - Device attached to one of the networks of an internet
  - Supports end-user applications or services
- Intermediate System (IS)
  - Device used to connect two networks
  - Permits communication between end systems attached to different networks

(1) Internetworking Terms (3)

- Bridge
  - IS used to connect two LANs using similar LAN protocols
  - Address filter passing on packets to the required network only
  - OSI layer 2 (Data Link)
- Router
  - Connects two (possibly dissimilar) networks
  - Uses internet protocol present in each router and end system
  - OSI Layer 3 (Network)

(2) Requirements of Internetworking

- Link between networks
  - Minimum a physical and link layer
- Routing and delivery of data between processes on different networks
- Accounting services and status info
- Independent of network architectures

Network Architecture Features

- Addressing
- Packet size
- Access mechanism
- Timeouts
- Error recovery
- Status reporting
- Routing
- User access control
- Connection based or connectionless

(3) Architectural Approaches

- Connection oriented
- Connectionless

Connection Oriented

- Assume that each network is connection oriented
- IS connect two or more networks
  - IS appear as ES to each network
  - Logical connection set up between ESs
    - Concatenation of logical connections across networks
  - Individual network virtual circuits joined by IS
- May require enhancement of local network services
  - 802, FDDI are datagram services

Connection Oriented IS Functions

- Relaying
- Routing

- e.g. X.75 used to interconnect X.25 packet switched networks

- Connection oriented not often used
  - □ (IP dominant)

Connectionless Operation

- Corresponds to datagram mechanism in packet switched network
- Each NPDU treated separately
- Network layer protocol common to all DTEs and routers
  - Known generically as the internet protocol
- Internet Protocol
  - One such internet protocol developed for ARPANET
  - RFC 791 (Get it and study it)
- Lower layer protocol needed to access particular network

Connectionless Internetworking

- Advantages
  - Flexibility
  - Robust
  - No unnecessary overhead
- Unreliable
  - Not guaranteed delivery
  - Not guaranteed order of delivery
    - Packets can take different routes
  - Reliability is responsibility of next layer up (e.g. TCP)

18.3 Internetworking of Connectionless

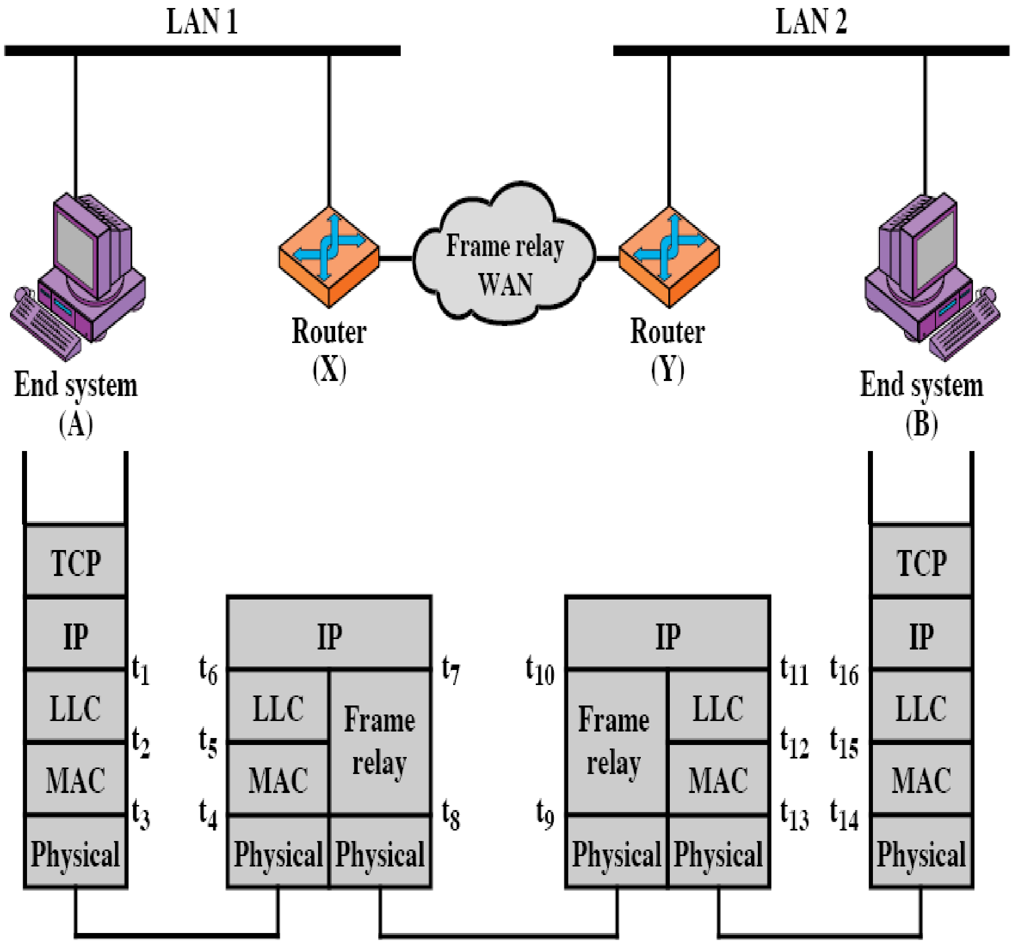
(1) IP operation

(2) Design Issues

(1) IP Operation

$t_1, t_6, t_7, t_{10}, t_{11}, t_{16}$			IP-H	TCP-H	Data	
t_2, t_5		LLC1-H	IP-H	TCP-H	Data	
t_3, t_4	MAC1-H	LLC1-H	IP-H	TCP-H	Data	MAC1-T
t_8, t_9		FR-H	IP-H	TCP-H	Data	FR-T
t_{12}, t_{15}		LLC2-H	IP-H	TCP-H	Data	
t_{13}, t_{14}	MAC2-H	LLC2-H	IP-H	TCP-H	Data	MAC2-T

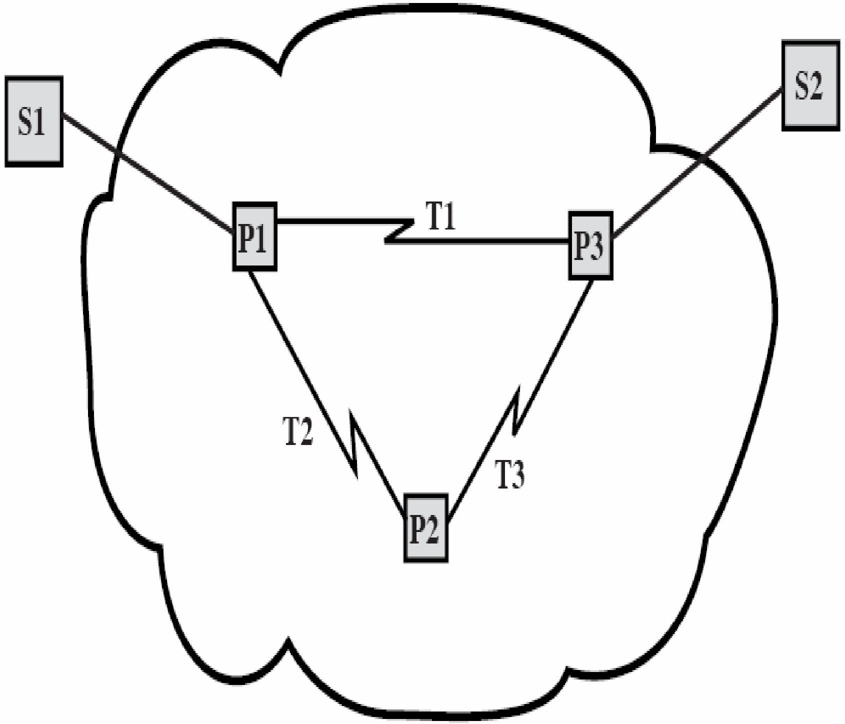
TCP-H	=	TCP header	MACi-T	=	MAC trailer
IP-H	=	IP header	FR-H	=	Frame relay header
LLCi-H	=	LLC header	FR-T	=	Frame relay trailer
MACi-H	=	MAC header			



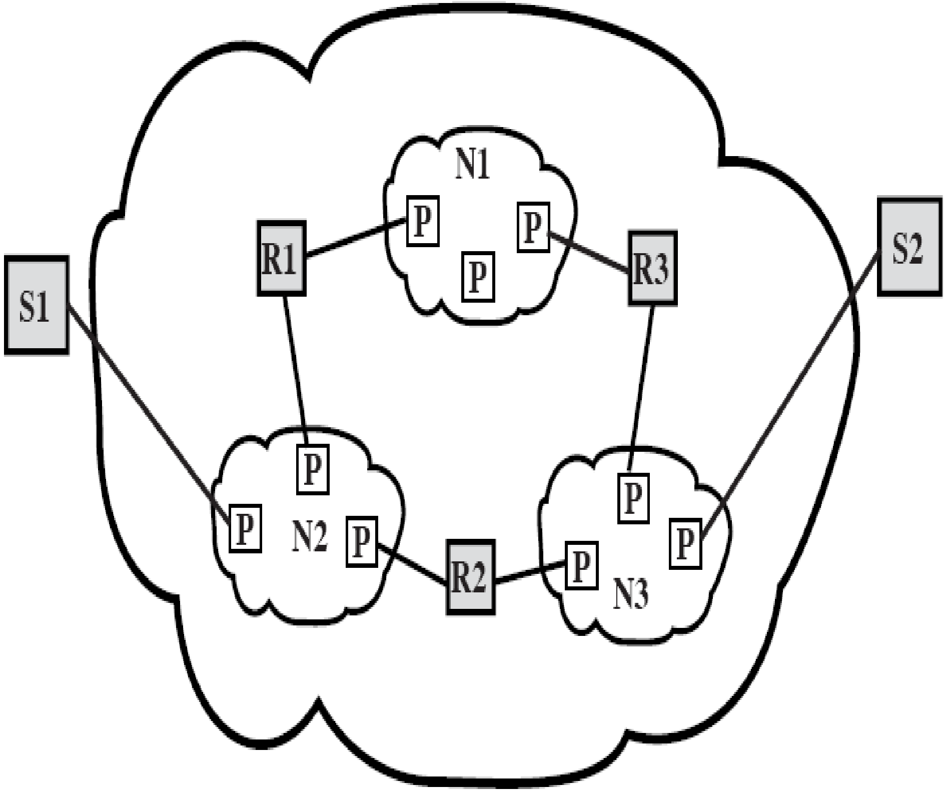
(2) Design Issues

- Routing
- Datagram lifetime
- Fragmentation and re-assembly
- Error control
- Flow control

(2) Design Issues
— The Internet as a Network



(a) Packet-switching network architecture



(b) Internetwork architecture

(2) Design Issues — Routing

- End systems and routers maintain routing tables
  - Indicate next router to which datagram should be sent
  - Static
    - May contain alternative routes
  - Dynamic
    - Flexible response to congestion and errors
- Source routing
  - Source specifies route as sequential list of routers to be followed
  - Security
  - Priority
- Route recording

(2) Design Issues — Datagram Lifetime

- Datagrams could loop indefinitely
  - Consumes resources
  - Transport protocol may need upper bound on datagram life
- Datagram marked with lifetime
  - Time To Live field in IP
  - Once lifetime expires, datagram discarded (not forwarded)
  - Hop count
    - Decrement time to live on passing through a each router
  - Time count
    - Need to know how long since last router
- (Aside: compare with Logan's Run)

(2) Design Issues — Fragmentation and Re-assembly

- Different packet sizes
- When to re-assemble
  - At destination
    - Results in packets getting smaller as data traverses internet
  - Intermediate re-assembly
    - Need large buffers at routers
    - Buffers may fill with fragments
    - All fragments must go through same router
      - Inhibits dynamic routing

(2) Design Issues — IP Fragmentation (1)

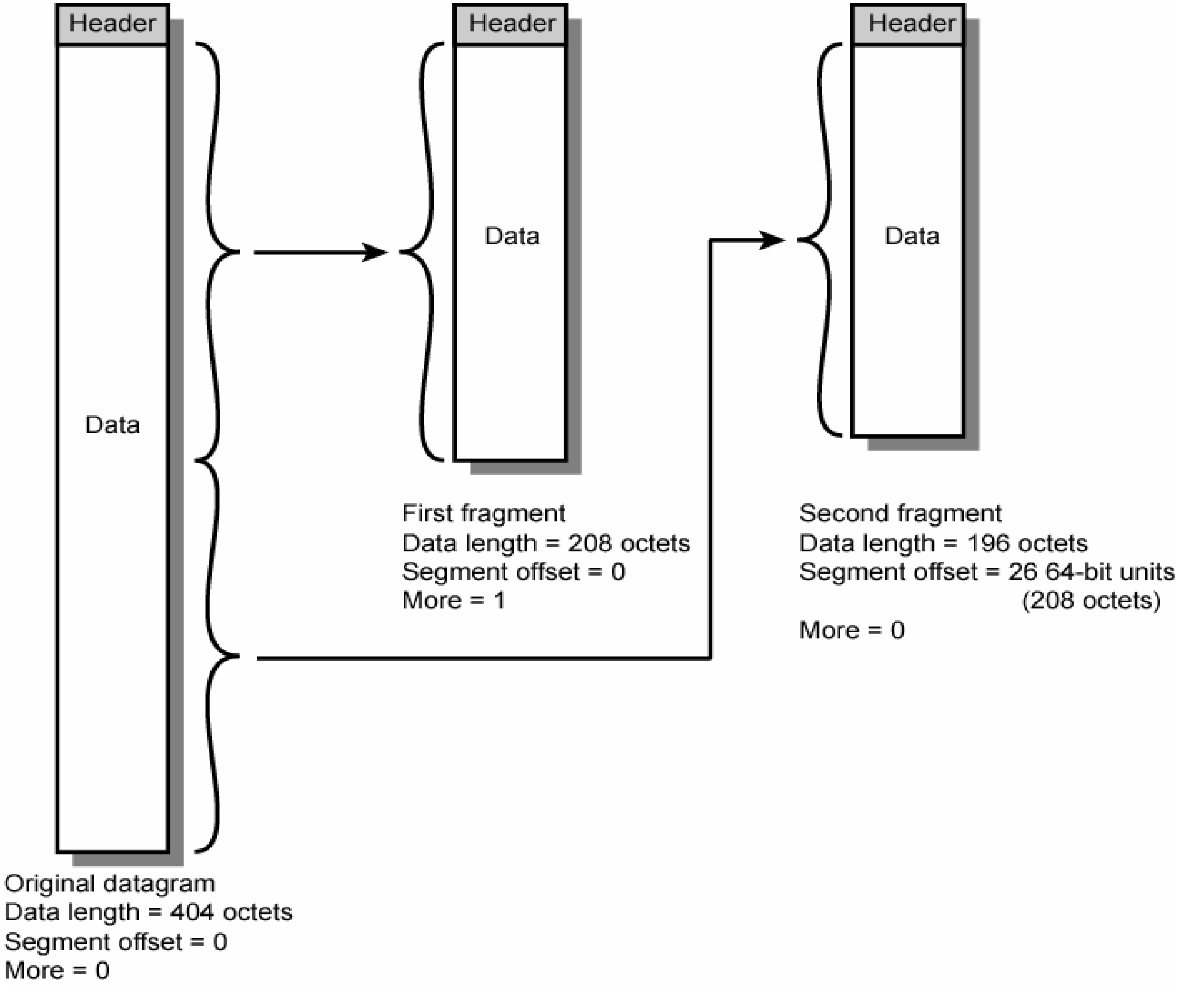
- IP re-assembles at destination only Uses fields in header
  - Data Unit Identifier (ID)
    - Identifies end system originated datagram
      - Source and destination address
      - Protocol layer generating data (e.g. TCP)
      - Identification supplied by that layer
  - Data length
    - Length of user data in octets

(2) Design Issues — IP Fragmentation (2)

- Offset
  - Position of fragment of user data in original datagram
  - In multiples of 64 bits (8 octets)
- *More* flag
  - Indicates that this is not the last fragment

(2) Design Issues — Fragmentation （3）

Example



(3) Design Issues — Fragmentation (4)

- Dealing with Failure
  - Re-assembly may fail if some fragments get lost
  - Need to detect failure
  - Re-assembly time out
    - Assigned to first fragment to arrive
    - If timeout expires before all fragments arrive, discard partial data
  - Use packet lifetime (time to live in IP)
    - If time to live runs out, kill partial data

(4) Design Issues — Error Control

- Not guaranteed delivery
- Router should attempt to inform source if packet discarded
  - e.g. for time to live expiring
- Source may modify transmission strategy
- May inform high layer protocol
- Datagram identification needed

(5) Design Issues — Flow Control

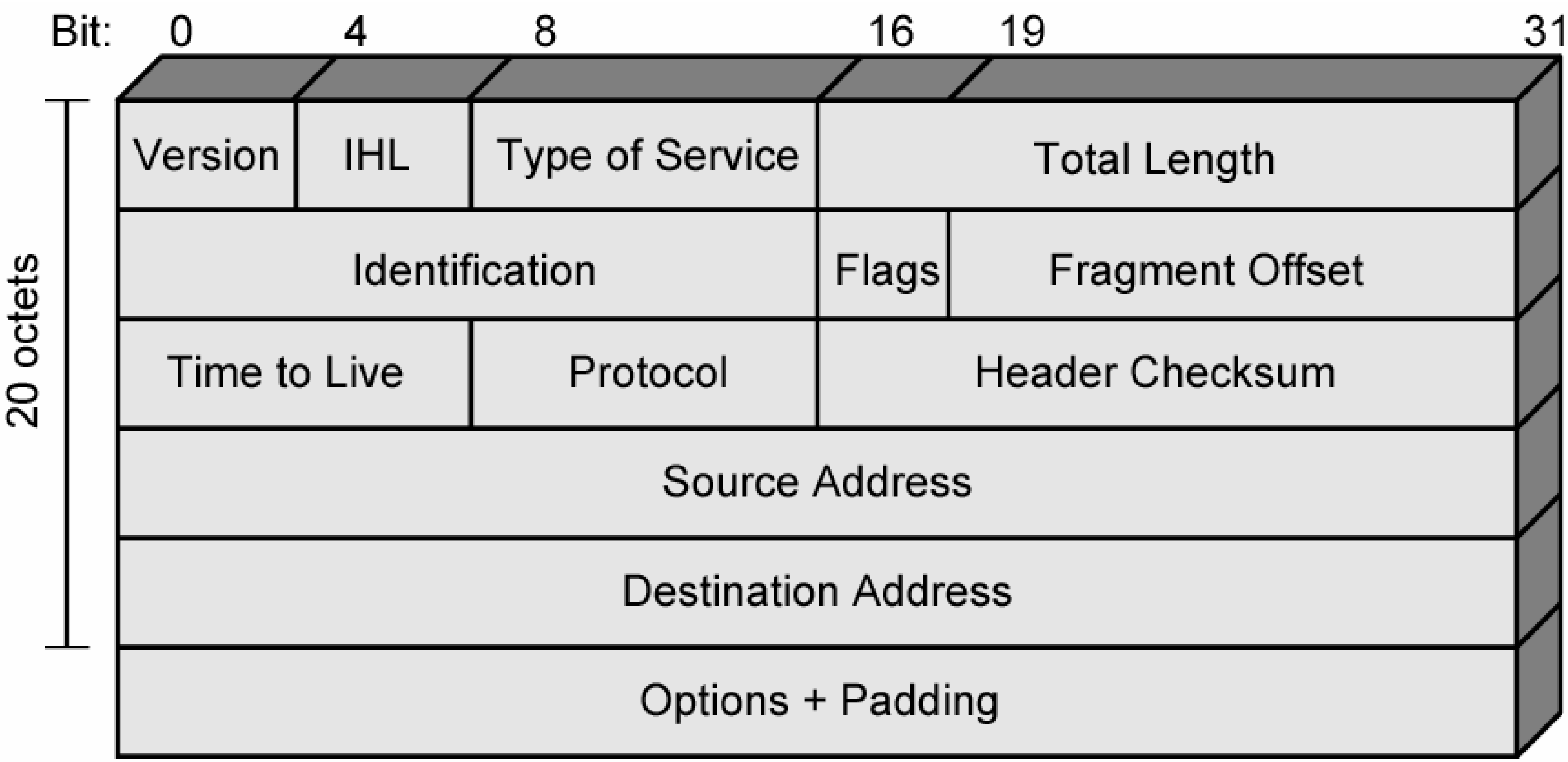
- (5) Allows routers and/or stations to limit rate of incoming data
- (6) Limited in connectionless systems
  - (5) Send flow control packets Requesting reduced flow，e.g. ICMP

18.4 Internet Protocol

Internet Protocol (IP) Version 4

- Part of TCP/IP
  - Used by the Internet
- Specifies interface with higher layer
  - e.g. TCP
- Specifies protocol format and mechanisms
- RFC 791
  - Get it and study it!
  - www.rfc-editor.org
- Will (eventually) be replaced by IPv6 (see later)

IPv4 Header



Bit: 0 4 8 16 19 31

Version | IHL | Type of Service | Total Length

Identification | Flags | Fragment Offset

Time to Live | Protocol | Header Checksum

Source Address

Destination Address

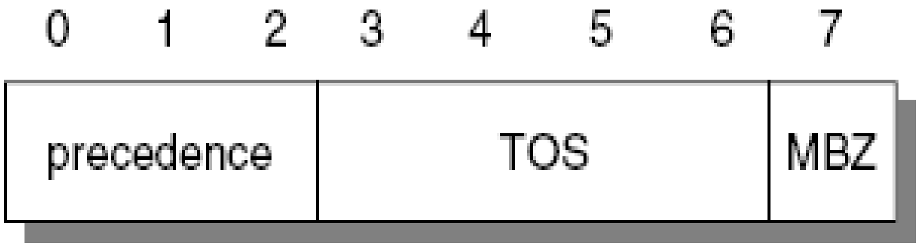
Options + Padding

20 octets

Header Fields (1)

- VERS: The field contains the IP protocol version.
  - The current version is 4. 5 is an experimental version. 6 is the version for IPv6.
- HLEN
  - The length of the IP header counted in 32-bit quantities.
  - This doesnot include the data field.
- Type Of Service
  - The service type is an indication of the quality of service requested for this IP datagram.
  - This field contains the following information:

Header Fields (2)



- Precedence: This field specifies the nature and priority of the datagram:
  - 000: Routine
  - 001: Priority
  - 010: Immediate
  - 011: Flash
  - 100: Flash override
  - 101: Critical
  - 110: Internetwork control
  - 111: Network control

- TOS: Specifies the type of service value:
  - 1000: Minimize delay
  - 0100: Maximize throughput
  - 0010: Maximize reliability
  - 0001: Minimize monetary cost
  - 0000: Normal service

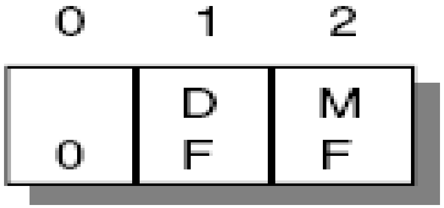
- MBZ*:* Reserved for future use.

Header Fields (3)

- Total Length
  - The total length of the datagram (or fragment), header and data.
- Identification
  - A unique number assigned by the sender to aid in reassembling a fragmented datagram.
  - Each fragment of a datagram hasthe same identification number.
- Time to Live
  - This field indicates the maximum time the datagram is allowed to remain in the internet system.
  - The time is measured in units of seconds, In practise, a router processes the datagram in less than 1 second.
  - The intention is to cause undeliverable datagrams to be discarded, and to bound the maximum datagram lifetime.

Header Fields (4)

- Flags
  - □ More bit
  - □ Don′t fragment



- Fragmentation offset
  - □ This field indicates where in the datagram this fragment belongs. The fragment offset is measured in units of 8 octets (64 bits). The first fragment has offset zero

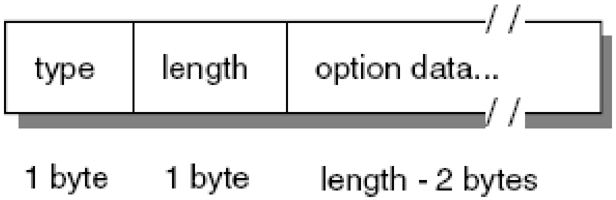
Header Fields (5)

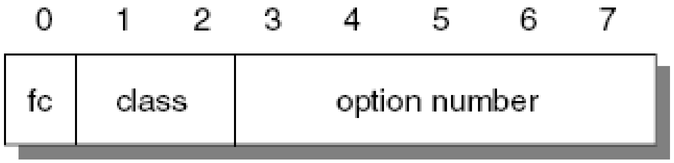
- Protocol Number: This field indicates the higher level protocol to which IP should deliver the data in this datagram.
  - 0: Reserved
  - 1: Internet Control Message Protocol (ICMP)
  - 2: Internet Group Management Protocol (IGMP)
  - 3: Gateway-to-Gateway Protocol (GGP)
  - 4: IP (IP encapsulation)
  - 5: Stream
  - 6: Transmission Control Protocol (TCP)
  - 8: Exterior Gateway Protocol (EGP)
  - 9: Private Interior Routing Protocol
  - 17: User Datagram Protocol (UDP)
  - 41: IP Version 6 (IPv6)
  - 50: Encap Security Payload for IPv6 (ESP)
  - 51: Authentication Header for IPv6 (AH)
  - 89: Open Shortest Path First

Header Fields (6)

- Header checksum
  - Reverified and recomputed at each router
  - 16 bit ones complement sum of all 16 bit words in header
  - Set to zero during calculation
  - If the header checksum does not match the contents, the datagram is discarded.
- Source address
- Destination address

Header Fields (7)





- fc (Flag copy): This field indicates whether (1) or not (0) the option field is copied when the datagram is fragmented.
- class: The option class is a 2-bit unsigned integer:
  - 0: control
  - 1: reserved
  - 2: debugging and measurement
  - 3: reserved

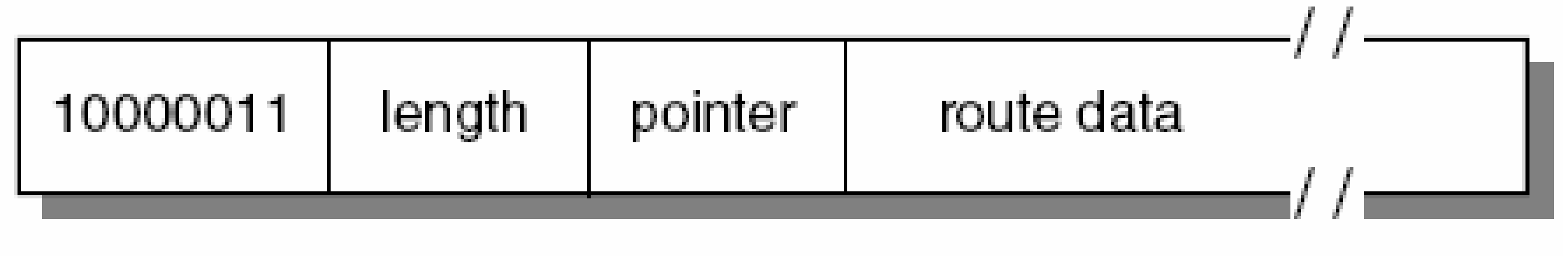
Header Fields (8)

- option number
  - 0: End of option list. That is, the option list is terminated by a X'00' byte. It is only required if the IP header length (which is a multiple of 4 bytes) does not match the actual length of the options.
  - 1: No operation. It has a class of 0, the fc bit is not set and there is no length byte or data. That is, a X'01' byte is a NOP. It may be used to align fields in the datagram.
  - 2: Security. It has a class of 0, the fc bit is set and there is a length byte with a value of 11 and 8 bytes of data). It is used for security information needed by U.S. Department of Defense requirements.
  - 3: Loose source routing. It has a class of 0, the fc bit is set and there is a variable length data field.
  - 4: Internet timestamp. It has a class of 2, the fc bit is not set and there is a variable length data field. The total length may be up to 40 bytes.
  - 7: Record route. It has a class of 0, the fc bit is not set and there is a variable length data field.

- 8: Stream ID. It has a class of 0, the fc bit is set and there is a length byte with a value of 4 and one data byte. It is used with the SATNET system.
- 9: Strict source routing. It has a class of 0, the fc bit is set and there is a variable length data field.

fc	class	number	length	DESCRIPTION
0	0	0	-	End of Option list, occupies only 1 octet，no length octet
0	0	1	-	No Operation. occupies only 1 octet，no length octet
1	0	2	11	Security.
1	0	3	var.	Loose Source Routing.
1	0	9	var.	Strict Source Routing.
0	0	7	var.	Record Route.
0	0	8	4	Stream ID.
0	2	4	var.	Internet Timestamp.

Loose source routing



- The loose source routing option, also called the loose source and record route (LSRR) option, provides a means for the source of an IP datagram to supply explicit routing information.
- This information is used by the routers when forwarding the datagram to the destination. It is also used to record the route.

Loose source routing

- Length
  - Contains the length of this option field, including the type and length fields.
- Pointer
  - Points to the option data at the next IP address to be processed.
  - It is counted relative to the beginning of the option, so its minimum value is 4. If the pointer is greater than the length of the option, the end of the source route is reached and further routing is to be based on the destination IP address (as for datagrams without this option).
- route data
  - This field contains a series of 32-bit IP addresses.

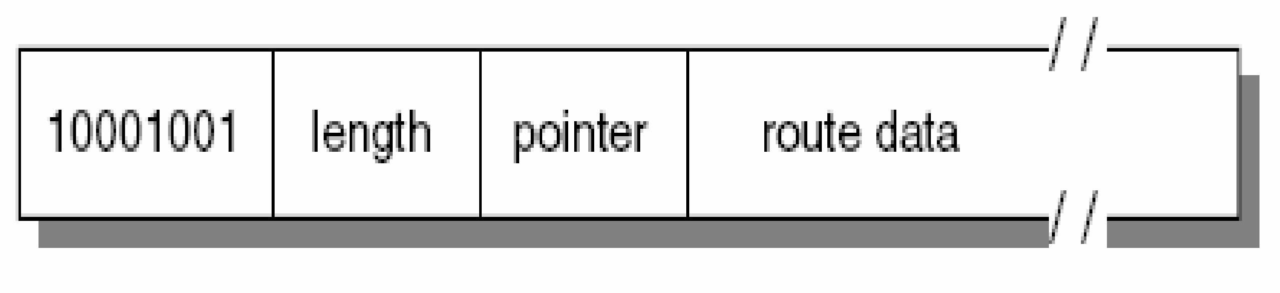
Loose source routing

- When a datagram arrives at its destination and the source route is not empty (pointer < length) the receiving host:
  - □ Takes the next IP address in the route data field (the one indicated by the pointer field) and puts it in the destination IP address field of the datagram.
  - □ Puts the local IP address in the source list at the location pointed to by the pointer field. The IP address for this is the local IP address corresponding to the network on which the datagram will be forwarded. (Routers are attached to multiple physical networks and thus have multiple IP addresses.)
  - □ Increments the pointer by 4.
  - □ Transmits the datagram to the new destination IP address.

Loose source routing

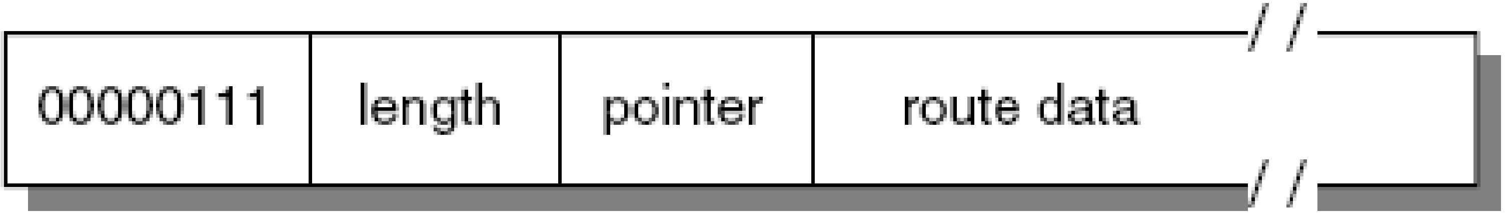
- The originating host puts the IP address of the first intermediate router in the destination address field
- The IP addresses of the remaining routers in the path, including the target destination are placed in the source route option.
- The recorded route in the datagram when it arrives at the target contains the IP addresses of each of the routers that forwarded the datagram.
- Each router has moved one place in the source route, and normally a different IP address will be used, since the routers record the IP address of the outbound interface but the source route originally contained the IP address of the inbound interface.

Strict source routing



- The strict source routing option, also called the strict source and record route (SSRR) option
- uses the same principle as loose source routing except the intermediate router *must* send the datagram to the next IP address in the source route via a directly connected network.
- It cannot use an intermediate router.
- If this cannot be done, ICMP Destination Unreachable error message is issued.

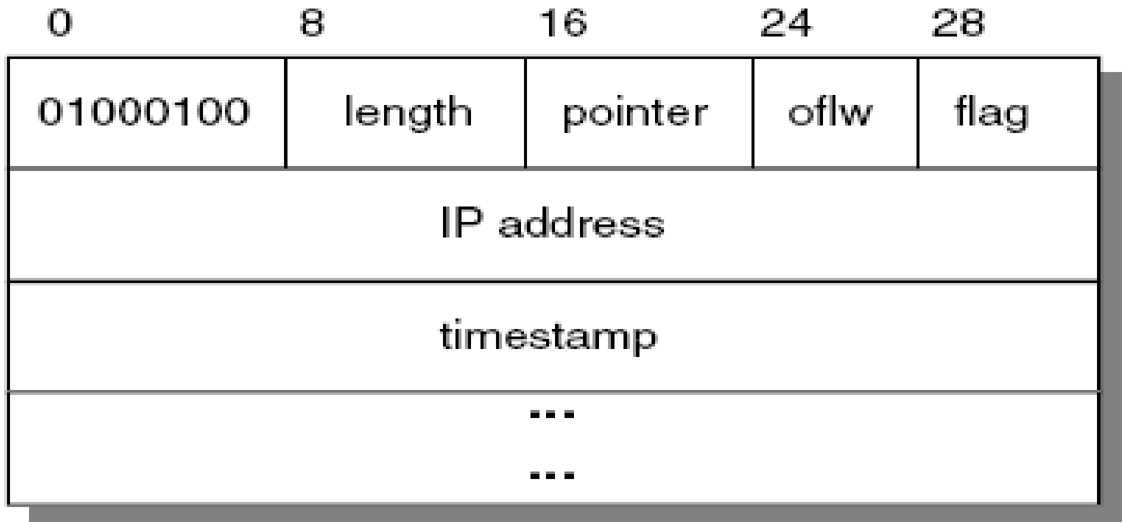
Record route



00000111 | length | pointer | route data

- This option provides a means to record the route traversed by an IP datagram.
- It functions similarly to the source routing option.
- However, this option provides an empty routing data field.
- This field is filled in as the datagram traverses the network.
- Sufficient space for this routing information must be provided by the source host.
- If the data field is filled before the datagram reaches its destination, the datagram is forwarded with no further recording of the route.

Internet timestamp



0	8	16	24	28
01000100	length	pointer	oflw	flag
IP address				
timestamp				
...				
...				

- They cannot be used for performance measurement for two reasons:
  - □ Because most IP datagrams are forwarded in less than one second, the timestamps are not precise.
  - □ Because IP routers are not required to have synchronized clocks, they may not be accurate.

Internet timestamp

- Length
  - Contains the total length of this option, including the type and length fields.
- Pointer
  - Points to the next timestamp to be processed (first free time stamp).
- oflw (overflow)
  - This field contains the number of devices that cannot register timestamps due to a lack of space in the data field.
- Flag
  - Is a 4-bit value which indicates how timestamps are to be registered:
  - 0: Timestamps only, stored in consecutive 32-bit words.
  - 1: Each timestamp is preceded by the IP address of the registering device.
  - 2 The IP address fields are pre-specified, an IP device only registers when it finds its own address in the list.
  - Timestamp: A 32-bit timestamp recorded in milliseconds since midnight UT (GMT).

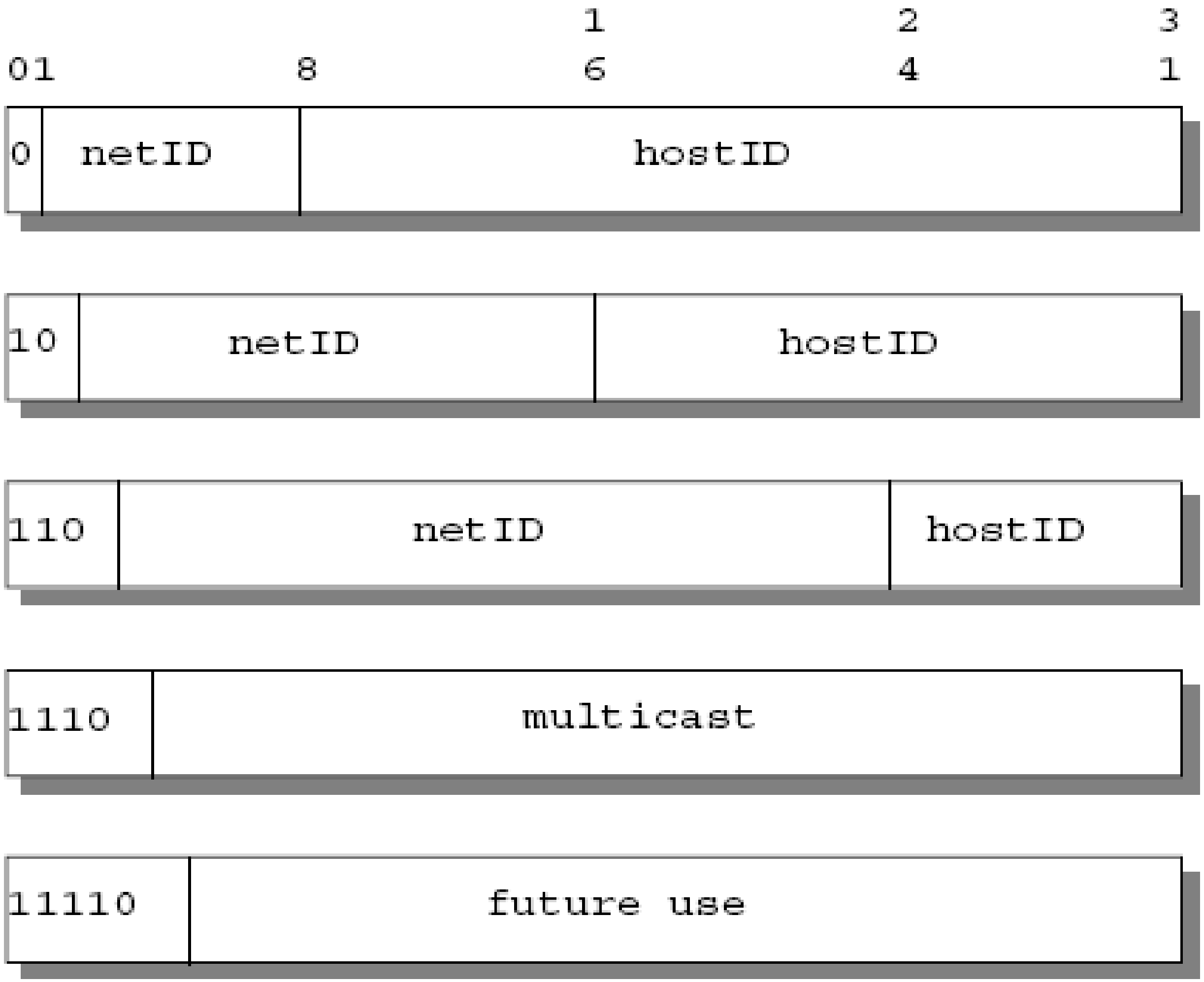
The IP address

- IP addressing standards are described in RFC 1166 – Internet Numbers.
- When the host is attached to more than one network, it is called *multi-homed* and has one IP address for each network interface.
- IP address = <network number><host number>

The IP address

- The *network number* portion of the IP address is administered by one of three Regional Internet Registries (RIR)
- American Registry for Internet Numbers (ARIN)
  - North America, South America, the Caribbean and sub-Saharan Africa.
- Reseaux IP Europeens (RIPE)
  - Europe, Middle East, parts of Africa.
- Asia Pacific Network Information Centre (APNIC)
  - Asia Pacific region.

Class-based IP addresses



0-126	Class A	0	netID	hostID
128-191	Class B	10	netID	hostID
192-223	Class C	110	netID	hostID
224-239	Class D	1110	multicast	
240-254	Class E	11110	future use	

Reserved IP addresses

- All bits 0:
  - An address with all bits zero in the host number portion is interpreted as *this* host
- All bits 1
  - An address with all bits one is interpreted as *all* networks or *all* hosts.
- Loopback
  - The class A network 127.0.0.0 is defined as the loopback network.

IP subnets

- A new type of physical network is installed at a location.
- Growth of the number of hosts requires splitting the local network into two or more separate networks.
- Growing distances require splitting a network into smaller networks, with gateways between them.

- To avoid having to request additional IP network addresses, the concept of IP subnetting was introduced.
- The assignment of subnets is done locally.
- The entire network still appears as one IP network to the outside world.

IP subnets

- The host number part of the IP address is subdivided into a second network number and a host number.
- This second network is termed a *subnetwork* or *subnet*.
- The main network now consists of a number of subnets. The IP address is interpreted as:

<network number><subnet number><host number>

- the local administrator chose a subnet number and host number
- The division is done using a 32-bit *subnet mask*.
- Bits with a value of one indicate positions ascribed to the subnet number.
- Bits with a value of zero bits in the subnet mask indicate positions ascribed to the host number.

IP routing

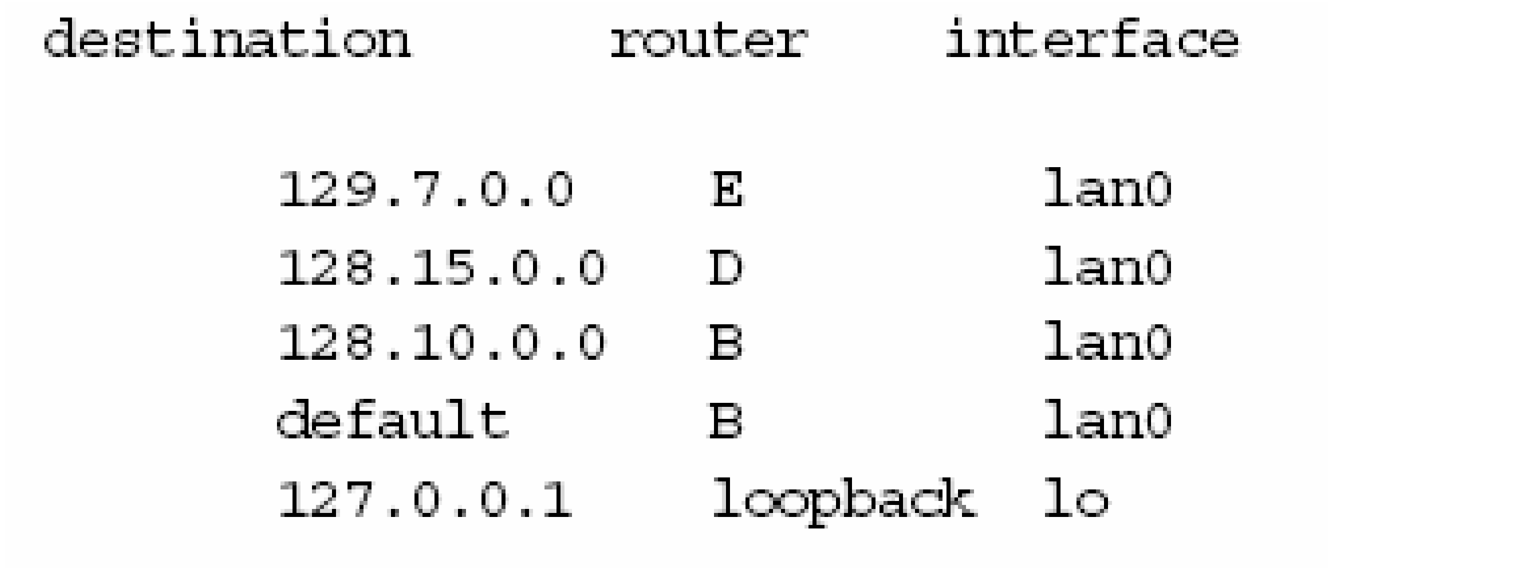
- An important function of the IP layer is *IP routing*.
- A device can simultaneously function as both a normal host and a router.

types of IP routing

- **Direct routing**
  - If the destination host is attached to the same physical network as the source host, IP datagrams can be directly exchanged.
- **Indirect routing**
  - the destination host is not connected to a network directly attached to the source host.
  - The only way to reach the destination is via one or more IP gateways.
  - The address of the first gateway (the first hop) is called an indirect route in the IP routing algorithm.
  - The address of the first gateway is the only information needed by the source host to send a packet to the destination host.

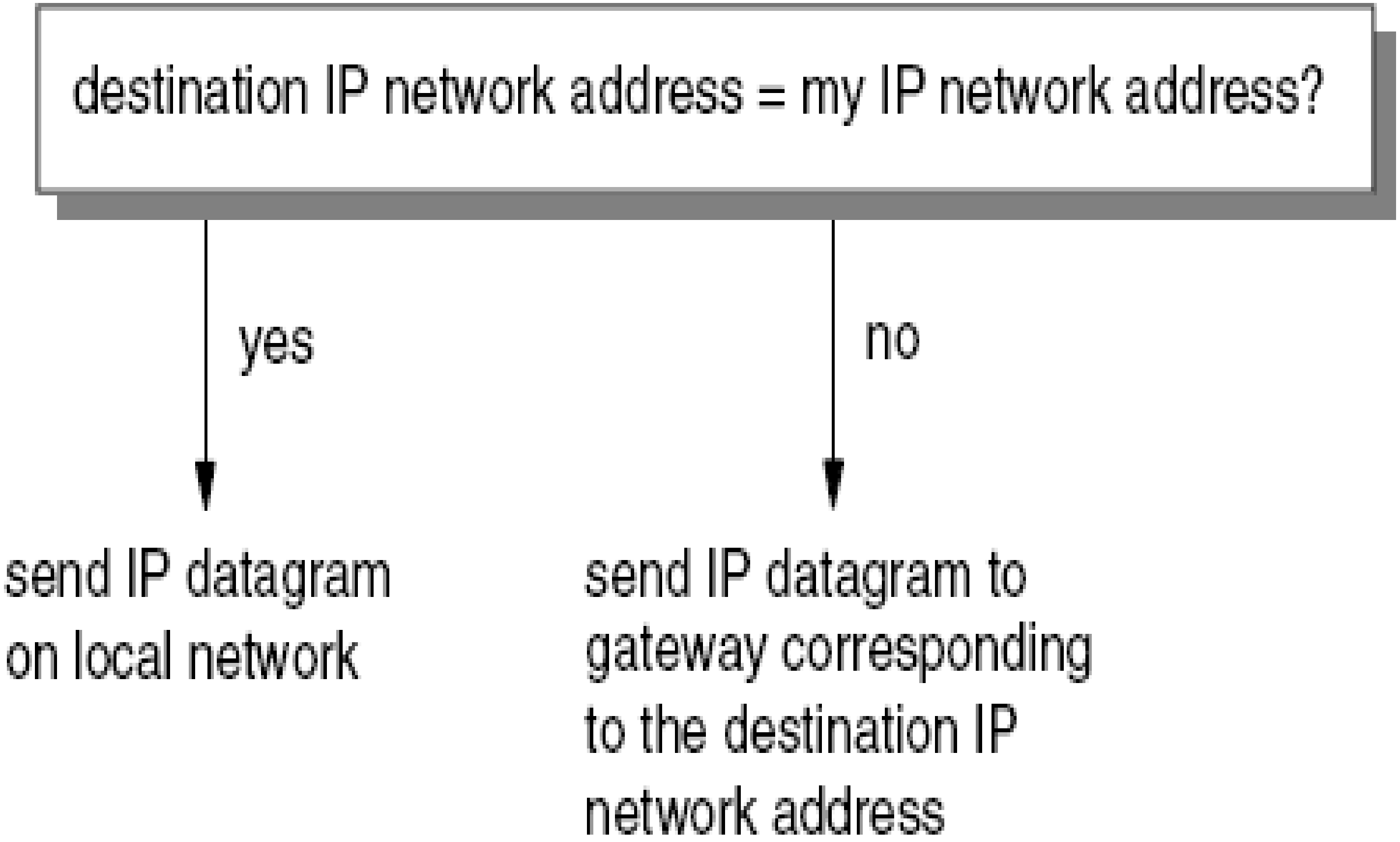
IP routing table

- Each host keeps the set of mappings between the following:
  - □ Destination IP network address(es)
  - □ Route(s) to next gateway(s)



destination	router	interface
129.7.0.0	E	lan0
128.15.0.0	D	lan0
128.10.0.0	B	lan0
default	B	lan0
127.0.0.1	loopback	lo

IP routing algorithm



destination IP network address = my IP network address?

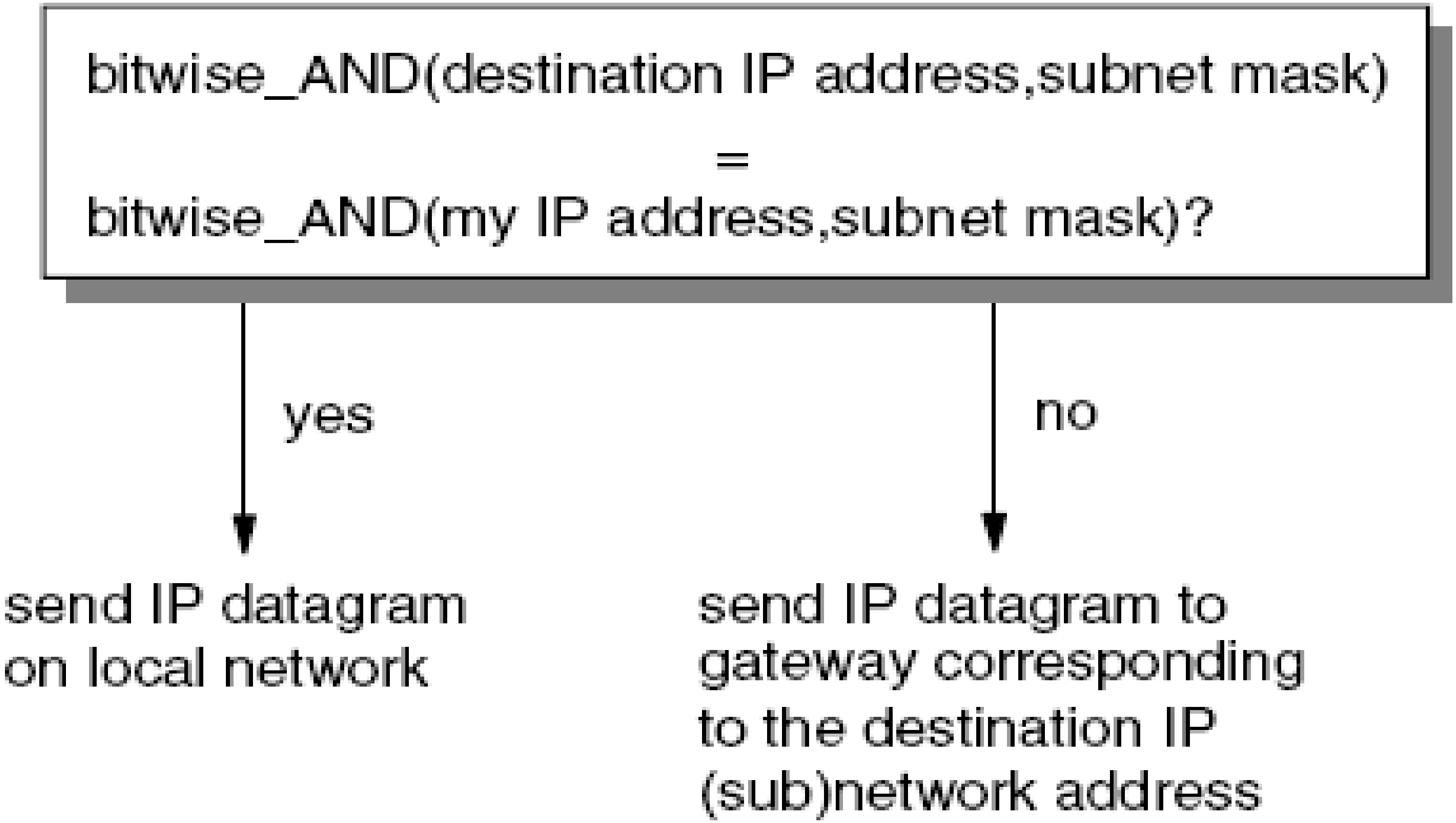
yes

no

send IP datagram on local network

send IP datagram to gateway corresponding to the destination IP network address

IP routing algorithm



bitwise_AND(destination IP address,subnet mask)

=

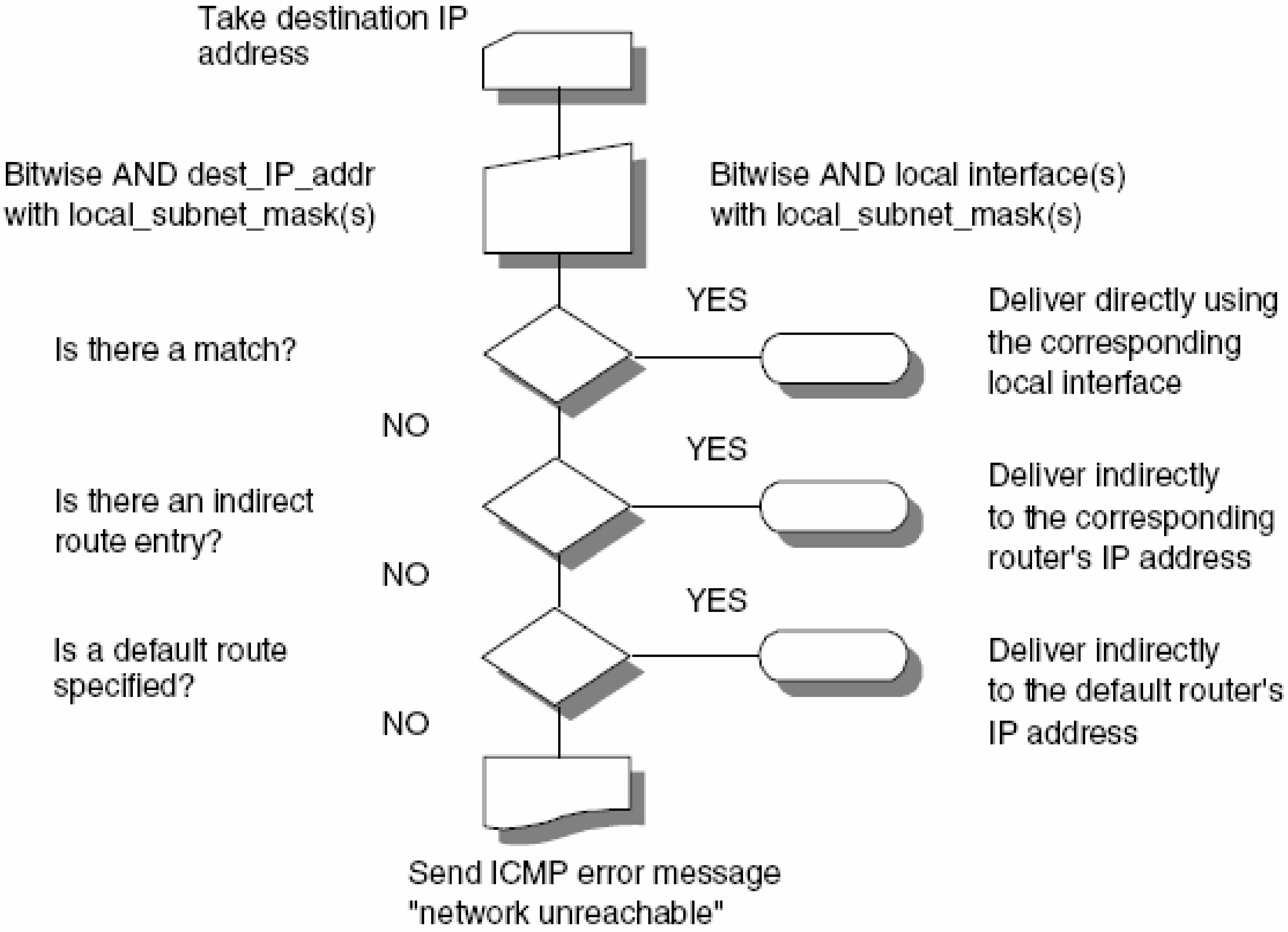
bitwise_AND(my IP address,subnet mask)?

yes

no

send IP datagram on local network

send IP datagram to gateway corresponding to the destination IP (sub)network address



Take destination IP address

Bitwise AND dest_IP_addr with local_subnet_mask(s)

Bitwise AND local interface(s) with local_subnet_mask(s)

Is there a match?

YES

Deliver directly using the corresponding local interface

NO

Is there an indirect route entry?

YES

Deliver indirectly to the corresponding router's IP address

NO

Is a default route specified?

YES

Deliver indirectly to the default router's IP address

NO

Send ICMP error message "network unreachable"

- If the IP implementation on any of the hosts does not support subnetting, that host will be able to communicate with any host in its own subnet but not with any machine on another subnet within the same network.

Broadcasting

- Limited broadcast address
  - This uses the address 255.255.255.255
  - It refers to all hosts on the local subnet.
  - Routers do not forward this packet.
- Network-directed broadcast address
  - This is used in an unsubnetted environment.
  - The network number is a valid network number and the host number is all ones (for example, 128.2.255.255).
  - This address refers to all hosts on the specified network.
  - Routers should forward these broadcast messages.
- Subnet-directed broadcast address:
  - the network number is a valid network number, the subnet number is a valid subnet number and the host number is all ones,
  - the address refers to all hosts on the specified subnet.

Multicasting

- 224.0.0.0-241.255.255.255
- Each group is represented by a Class D IP address.
- For each multicast address, a set of zero or more hosts are listening for packets addressed to the address.
- This set of hosts is called the *host group*.
- Packets sent to a multicast address are forwarded only to the members of the corresponding host group.

Anycasting

- Sometimes, the same IP services are provided by different hosts.
- For example, a user wants to download a file via FTP and the file is available on multiple FTP servers.
- Hosts that implement the same service provide ananycast address to other hosts that require the service.
- Connections are made to the first host in the anycast address group to respond.
- This process is used to guarantee the service is provided by the host with the best connection to the receiver.

private IP addresses

- 10.0.0.0
  - A single Class A network
- 172.16.0.0 - 172.31.0.0
  - 16 contiguous Class B networks
- 192.168.0.0 - 192.168.255.0
  - 256 contiguous Class C networks

- Any organization can use any address in these ranges.

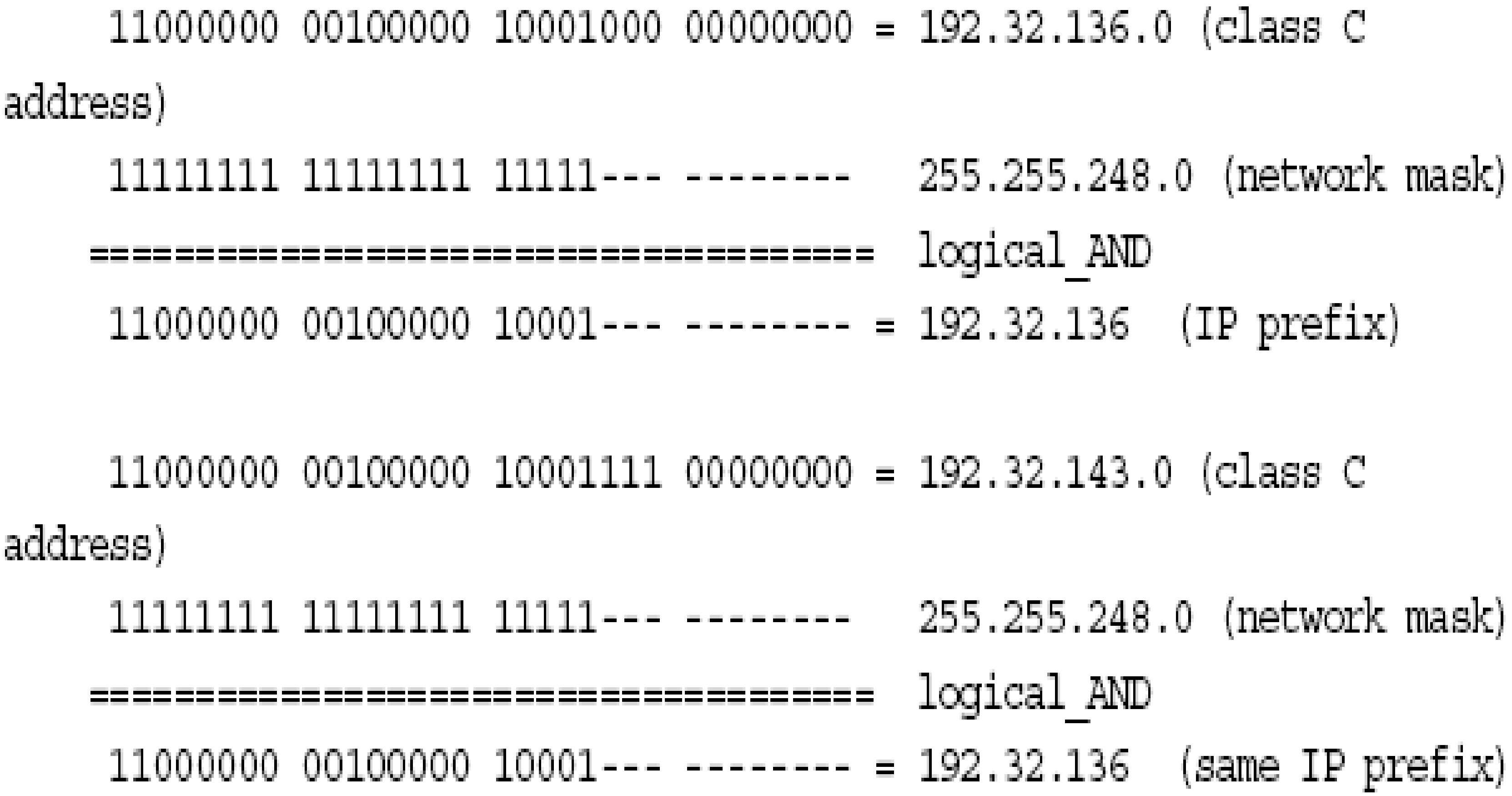
- Routers in networks not using private addresses are expected to quietly discard all routing information regarding these addresses.

- All connectivity to external Internet hosts must be provided with application gateways

Classless Inter-Domain Routing (CIDR)

- *routing table explosion* problem
  - A Class B network of 3000 hosts requires one routing table entry at each backbone router.
  - The same environment, if addressed as a range of Class C networks, requires 16 entries.
- The solution to this problem is called Classless Inter-Domain Routing (CIDR). CIDR is described in RFCs 1518 to 1520.

Classless Inter-Domain Routing (CIDR)



```
     11000000 00100000 10001000 00000000 = 192.32.136.0 (class C
address)
     11111111 11111111 11111---  --------   255.255.248.0 (network mask)
    ====================================  logical_AND
     11000000 00100000 10001--- -------- = 192.32.136  (IP prefix)

     11000000 00100000 10001111 00000000 = 192.32.143.0 (class C
address)
     11111111 11111111 11111---  --------   255.255.248.0 (network mask)
    ====================================  logical_AND
     11000000 00100000 10001--- -------- = 192.32.136  (same IP prefix)
```

18.5 **Internet Control Message Protocol (ICMP)**

- It is described in RFC 792 with updates in RFC 950.
- A router or a destination host uses ICMP to inform the source host about errors in datagram processing.

Characteristic of ICMP

- ICMP uses IP **as if ICMP were a higher level protocol**. However, ICMP is an integral part of IP and must be implemented by every IP module.

- ICMP is used to report errors, *not* to make IP reliable. Datagrams may still be undelivered without any report on their loss. Reliability must be implemented by the higher-level protocols using IP services.

- ICMP cannot be used to report errors with ICMP messages. This avoids infinite repetitions.

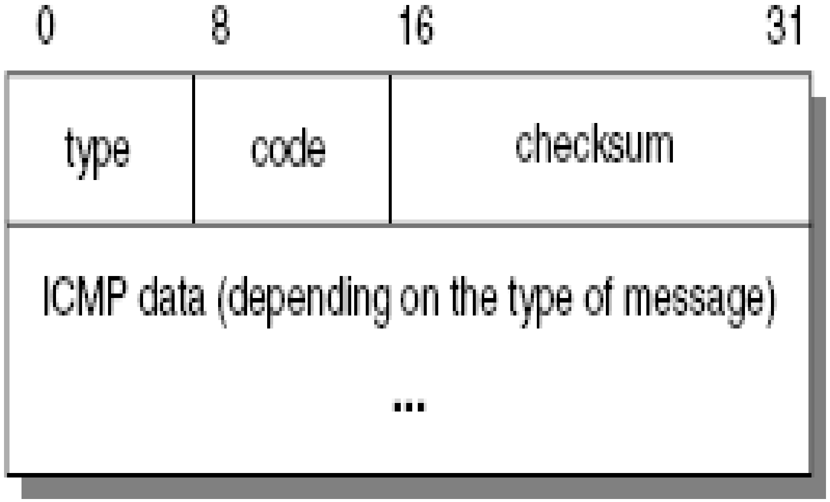
- For fragmented datagrams, ICMP messages are only sent about errors with the first fragment.

Characteristic of ICMP

- ICMP messages are never sent in response to datagrams with a broadcast or a multicast destination address.
- ICMP messages are never sent in response to a datagram that does not have a source IP address representing a unique host.

ICMP — type

- 0: Echo reply
- 3: Destination unreachable
- 4: Source quench
- 5: Redirect
- 8: Echo
- 9: Router advertisement
- 10: Router solicitation
- 11: Time exceeded
- 12: Parameter problem
- 13: Timestamp request
- 14: Timestamp reply
- 15: Information request (obsolete)
- 16: Information reply (obsolete)
- 17: Address mask request
- 18: Address mask reply
- 30: Traceroute



0 8 16 31

type code checksum

ICMP data (depending on the type of message)

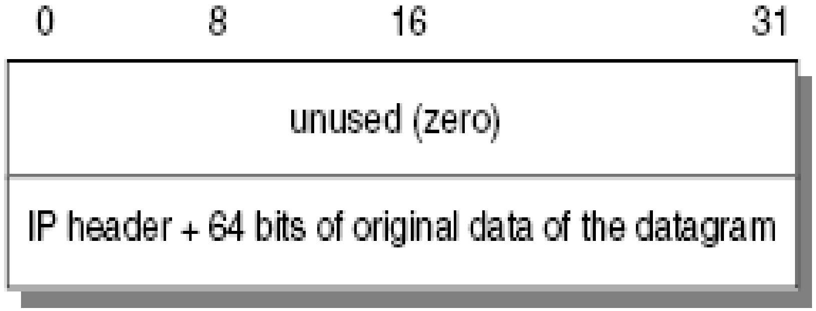
...

- 31: Datagram conversion error
- 32: Mobile host redirect
- 33: IPv6 Where-Are-You
- 34: IPv6 I-Am-Here
- 35: Mobile registration request
- 36: Mobile registration reply
- 37: Domain name request
- 38: Domain name reply
- 39: SKIP
- 40: Photuris

ICMP

- Code
  - Contains the error code for the datagram reported by this ICMP message. The interpretation is dependent upon the message type.
- Checksum
- Data
  - Contains information for this ICMP message. Typically it will contain the portion of the original IP message for which this ICMP message was generated.

Destination Unreachable (3)



0 8 16 31

unused (zero)

IP header + 64 bits of original data of the datagram

- received from an intermediate router
  - it means that the router regards the destination IP address as unreachable.
- received from the destination host
  - the protocol or theport is inactive.

0： Network unreachable

1： Host unreachable

2： Protocol unreachable

3： Port unreachable

4： Fragmentation needed but the Do Not Fragment bit was set

5： Source route failed

6： Destination network unknown

7： Destination host unknown

8： Source host isolated (obsolete)

9： Destination network administratively prohibited

10：Destination host administratively prohibited

11：Network unreachable for this type of service

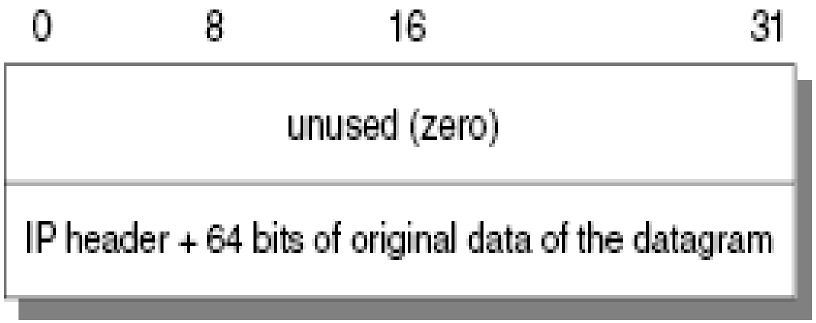
12：Host unreachable for this type of service

13：Communication administratively prohibited by filtering

14：Host precedence violation

15：Precedence cutoff in effect

Source Quench (4)



0 8 16 31

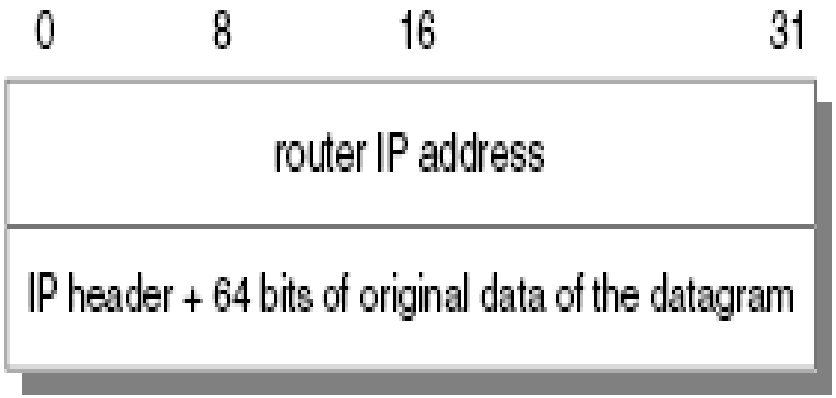
unused (zero)

IP header + 64 bits of original data of the datagram

- code field is always zero.

- from an intermediate router
  - the router did not have the buffer space needed to queue the datagram.
- from the destination host
  - it means that the incoming datagrams are arriving too quickly to be processed.

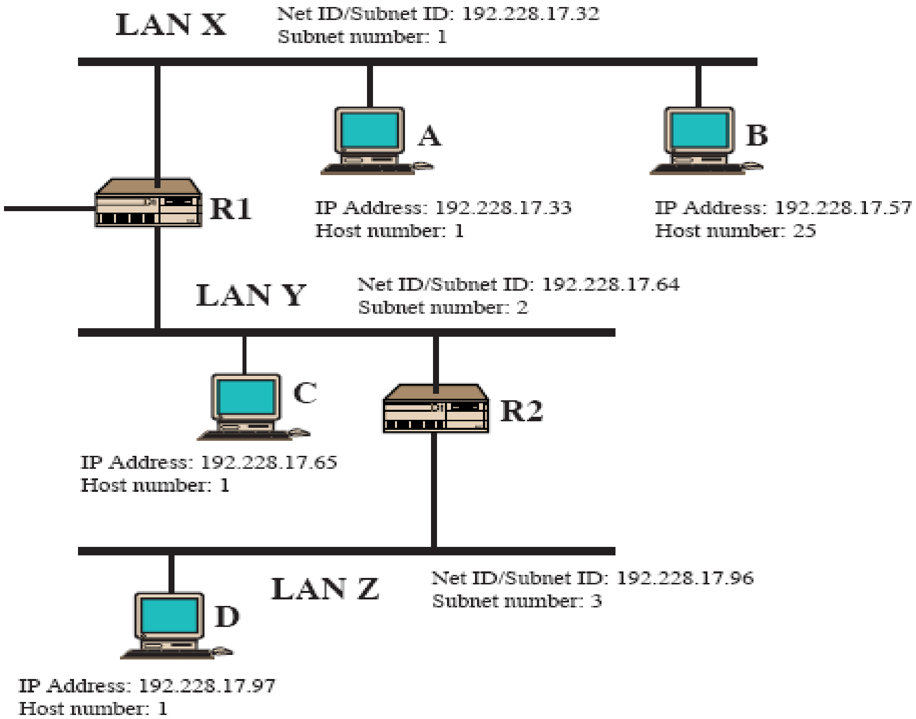
Redirect (5)



0 8 16 31

router IP address

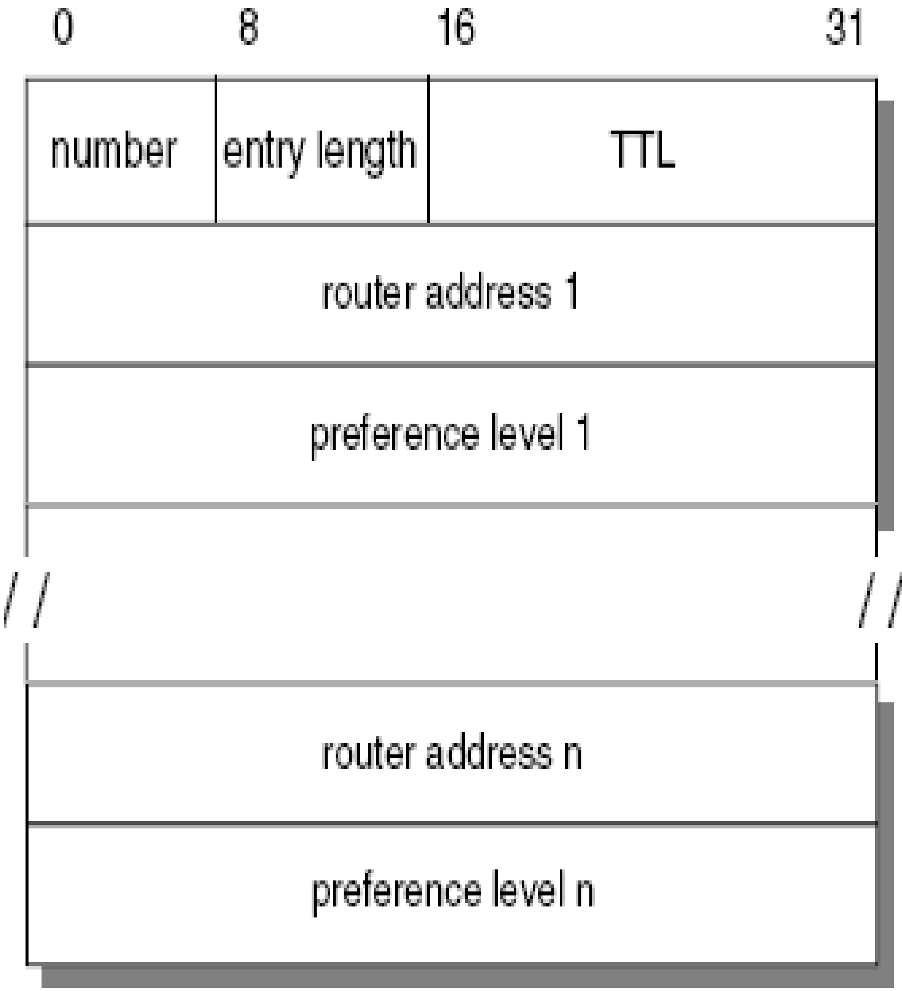
IP header + 64 bits of original data of the datagram


LAN X
Net ID/Subnet ID: 192.228.17.32
Subnet number: 1
A
B
R1
IP Address: 192.228.17.33
Host number: 1
IP Address: 192.228.17.57
Host number: 25
LAN Y
Net ID/Subnet ID: 192.228.17.64
Subnet number: 2
C
R2
IP Address: 192.228.17.65
Host number: 1
LAN Z
Net ID/Subnet ID: 192.228.17.96
Subnet number: 3
D
IP Address: 192.228.17.97
Host number: 1

- received from an intermediate router
  - it means that the host should send future datagrams for the network to the router whose IP address is specified in the ICMP message. T
  - his preferred router will always be on the same subnet as the host that sent the datagram and the router that returned the IP datagram.
  - The router forwards the datagram to its next hop destination.
  - This message will not be sent if the IP datagram contains a source route.

Router Advertisement (9) and Router Solicitation (10)

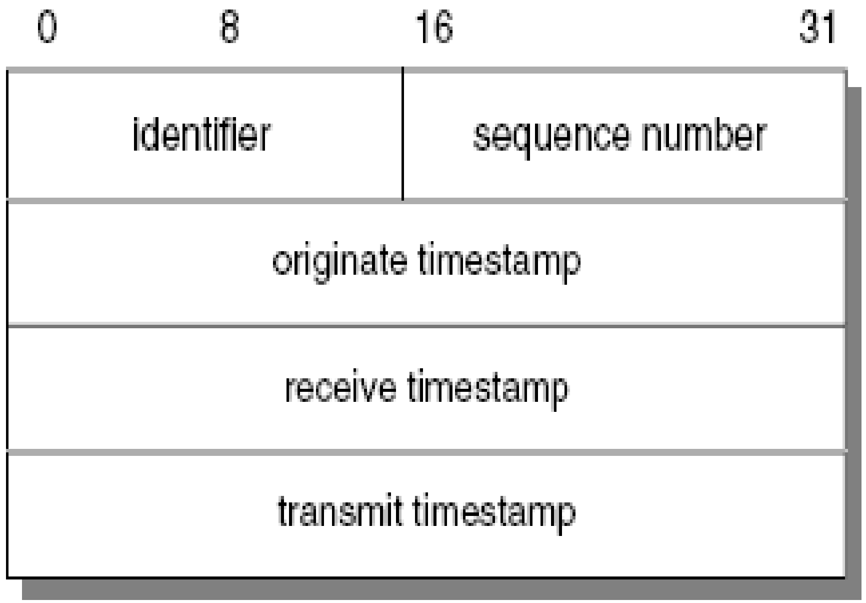
- RFC 1256
- Number
  - □ The number of entries in the message.
- entry length
  - □ The length of an entry in 32-bit units.
- TTL
  - □ The number of seconds that an entry will be considered valid.
- router address
  - □ One of the sender's IP addresses.
- preference level
  - □ A signed 32-bit level indicating the preference to be assigned to this address when selecting a default router.



number	entry length	TTL
router address 1		
preference level 1		
router address n		
preference level n		

Timestamp Request (13) and Timestamp Reply (14)

- The sender initializes the identifier and sequence number
- The sender sets the originate timestamp.
- The receiving host fills in the receive and transmit timestamps.



Echo (8) and Echo Reply (0)

- Echo is used to detect if another host is active on the network.
- The sender initializes the identifier, sequence number, and data field.
- The datagram is then sent to the destination host.
- The recipient changes the type to Echo Reply and returns the datagram to the sender.
- It is used by the Ping command.

ICMP applications

- **Traceroute**
  - The Traceroute program is used to determine the route IP datagrams follow through the network.
  - Traceroute is based upon ICMP and UDP.

- It sends an IP datagram with a TTL of 1 to the destination host.

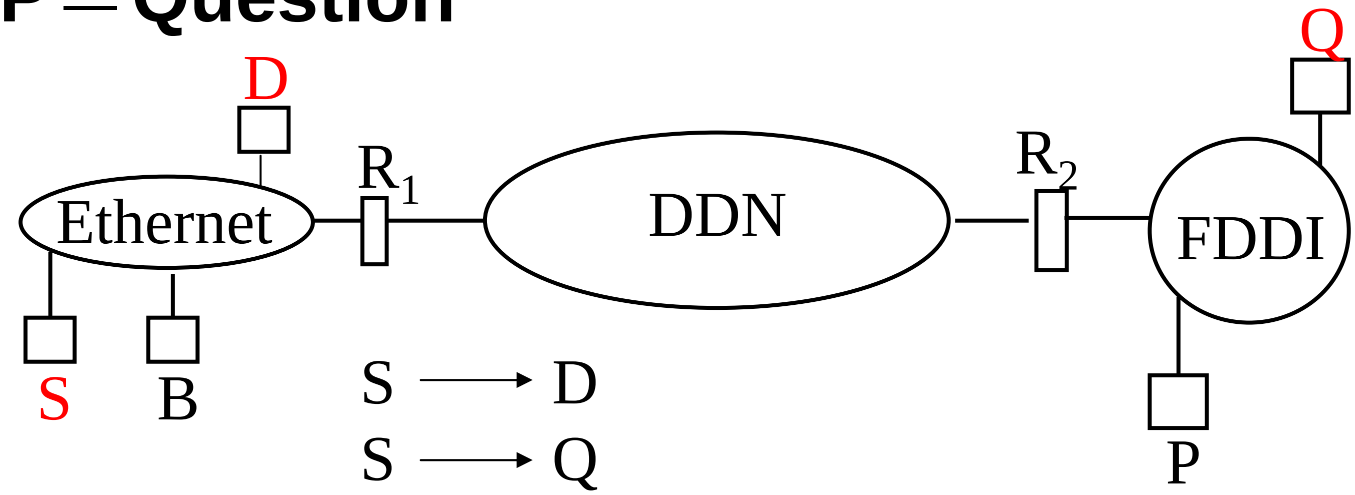
- The first router decrements the TTL to 0, discards the datagram and returns an ICMP Time Exceeded message to the source.

- This process is repeated with successively larger TTL values to identify the exact series of routers in the path to the destination host.

18.6 **Address Resolution Protocol (ARP)**

- ARP is responsible for converting the higher level protocol addresses (IP addresses) to physical network addresses.
- It is described in RFC 826.

ARP – Question



What address is used to send packets?

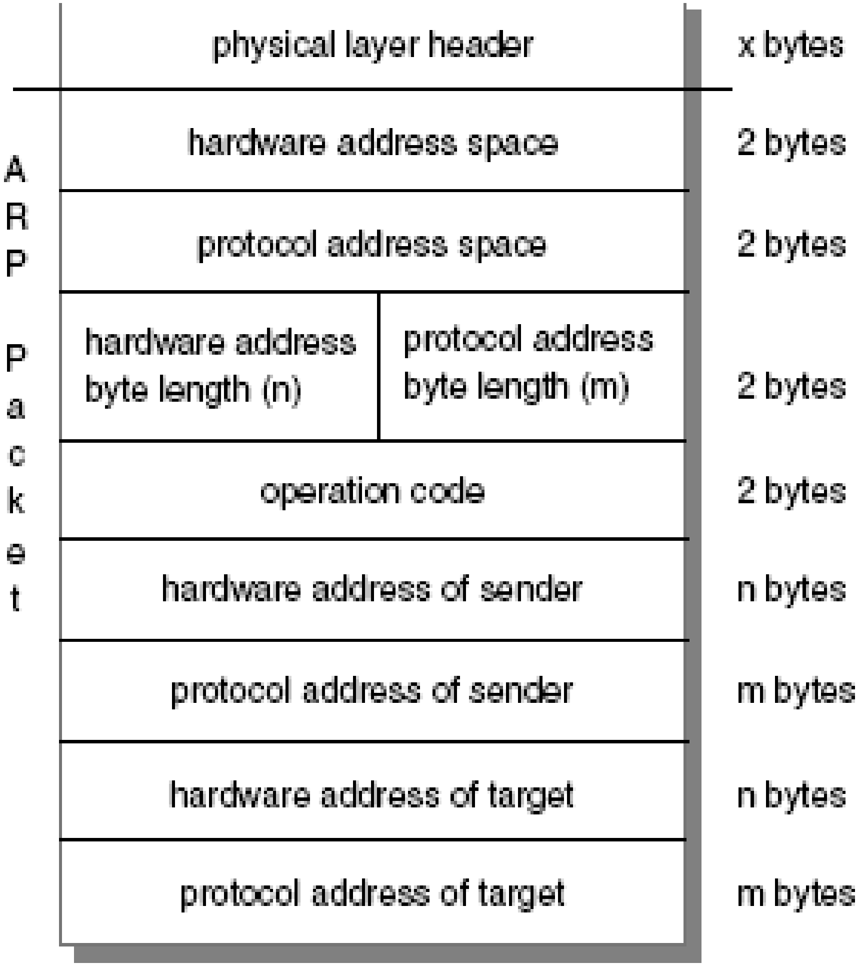
Machine	IP address	Hardware address
S	IPs	ETH_S
B	IP_B	ETH_B
D	IP_D	ETH_D
R_1	IP_{R1} , IP_{R11}	ETH_{R1} , DDN_{R1}
P	IP_P	$FDDI_P$
Q	IP_Q	$FDDI_Q$
R_2	IP_{R2}， IP_{R21}	$FDDI_{R2}$， DDN

ARP

- The ARP module tries to find the address in this ARP cache.
- If it finds the matching pair, it gives the corresponding 48-bit physical address back to the caller (the device driver), which then transmits the packet.
- If it doesn't find the pair in its table, it *discards the packet* (assumption is that a higher level protocol will retransmit) and generates a network *broadcast* of an ARP request.

ARP

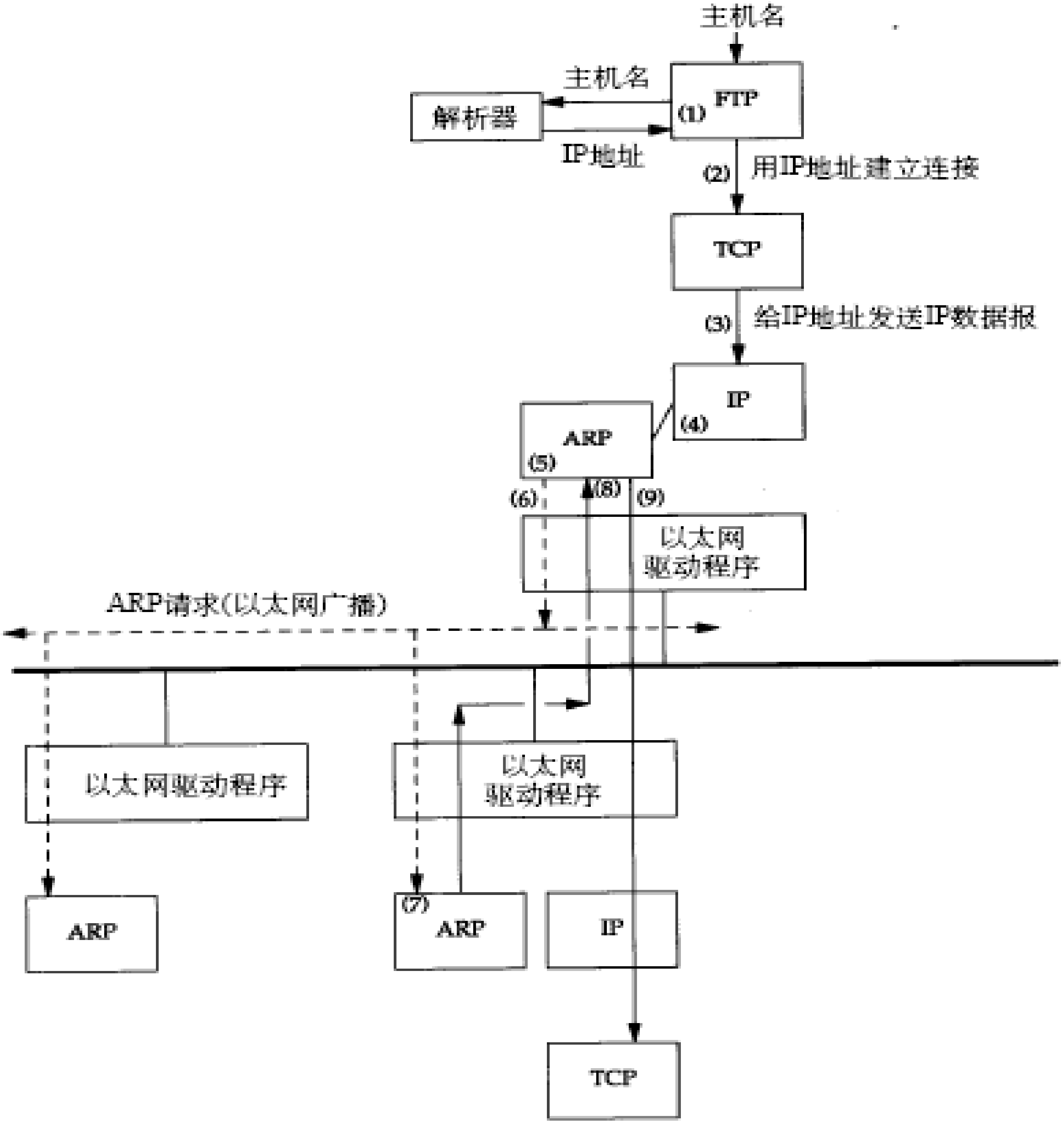
- Hardware address space: Specifies the type of hardware; examples are Ethernet or Packet Radio Net.
- Protocol address space: Specifies the type of protocol, same as the EtherType field in the IEEE 802 header (IP or ARP).
- Operation code: Specifies whether this is an ARP request (1) or reply (2).



ARP Packet

physical layer header	x bytes
hardware address space	2 bytes
protocol address space	2 bytes
hardware address byte length (n) / protocol address byte length (m)	2 bytes
operation code	2 bytes
hardware address of sender	n bytes
protocol address of sender	m bytes
hardware address of target	n bytes
protocol address of target	m bytes

用户输入ftp hostname时，ARP的操作

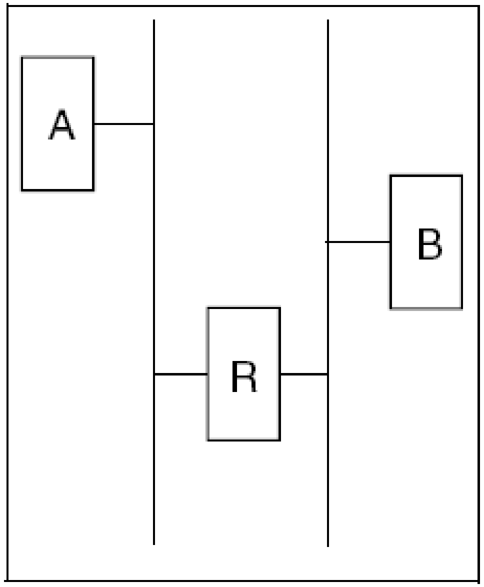


ARP　Experiment

- ARP cache
  - Command “arp –a ” will list the cached mapings.
- Tcpdump
  - Using the tcpdump you can lookup the process of ARP operations

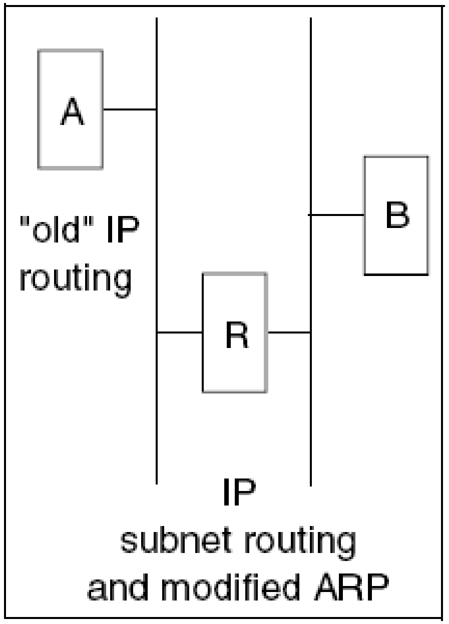
Proxy-ARP or transparent subnetting

- When host A wants to send an IP datagram to host B, it first has to determine the physical network address of host B through the use of the ARP protocol.
- As host A cannot differentiate between the physical networks, its IP routing algorithm thinks that host B is on the local physical network and sends out a broadcast ARP request.
- Host B doesn‘t receive this broadcast, but router R does.
- If router R‘s routing tables specify that the next hop to that other network is through a different physical device, it will reply to the ARP as if it were host B, saying that the network address of host B is that of the router R itself.

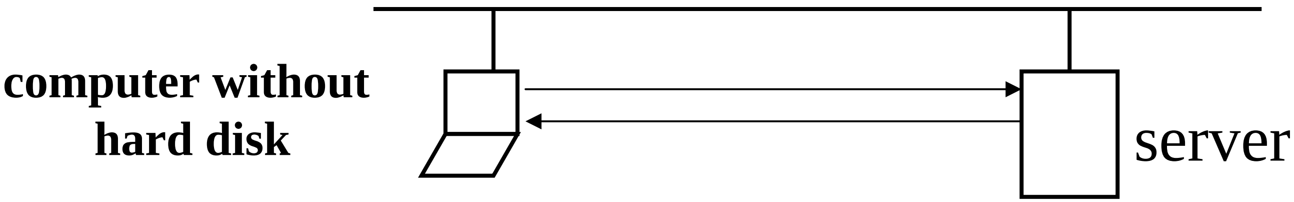


Proxy-ARP or transparent subnetting

- When host A wants to send an IP datagram to host B, it first has to determine the physical network address of host B through the use of the ARP protocol.
- As host A cannot differentiate between the physical networks, its IP routing algorithm thinks that host B is on the local physical network and sends out a broadcast ARP request.
- Host B doesn‘t receive this broadcast, but router R does.
- If router R‘s routing tables specify that the next hop to that other network is through a different physical device, it will reply to the ARP as if it were host B, saying that the network address of host B is that of the router R itself.



Reverse Address Resolution Protocol (RARP)



Where is IP address?   How get it?

Method:

- **broadcast a request with its hardware address**
- **server replies with required IP address**

Reverse Address Resolution Protocol (RARP)

- The RARP protocol is a network-specific standard protocol. It is described in RFC 903.
- Some network hosts, such as diskless workstations, do not know their own IP address when they are booted.
- The hardware address of the host is the known parameter, and the IP address the queried parameter.
- RARP server must exist on the network that maintains that a database of mappings from hardware address to protocol address must be pre-configured.

18.7 BOOTP

BOOTP— Introduction

1. What should a new computer know before using Internet?

- Its IP address and network mask
- default router
- DNS server
- other servers

2. How get them?

- manually
- automatically

many computers? MH?

where and how exchange message?

Drawbacks of RARP

- operates at a low level. application server?
- a small piece of information. waste
- use hardware address. not suit to some network

BOOTP

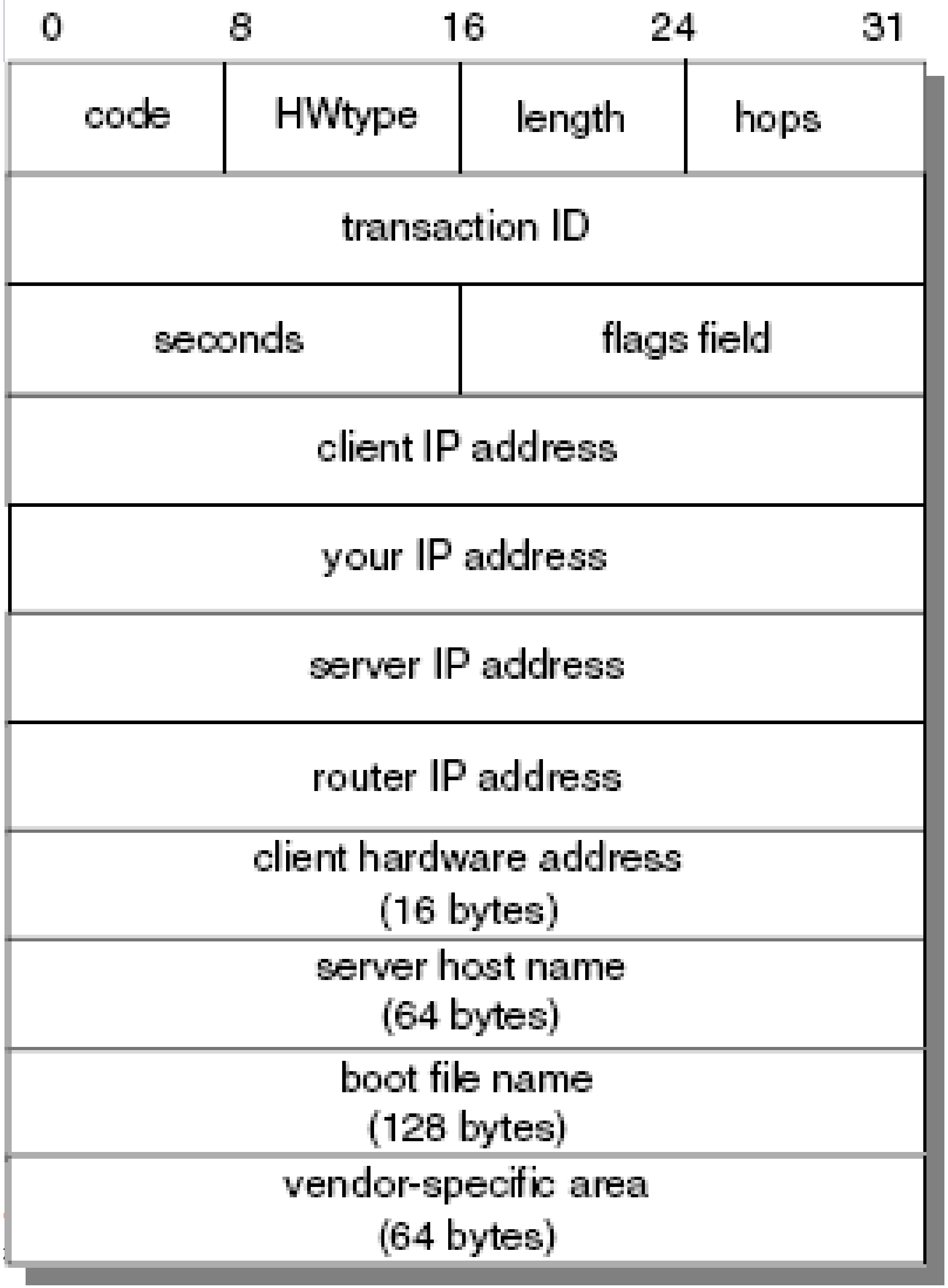
- The bootstrap protocol (BOOTP) enables a client workstation to initialize with a minimal IP stack and request
  - it’s IP address
  - a gateway address
  - the address of a name server from a BOOTP server.
- The BOOTP specifications can be found in RFC 951 – Bootstrap Protocol.

The BOOTP process

- The client determines its own hardware address; this is normally in a ROM on the hardware.
- A BOOTP client sends its hardware address in a UDP datagram to the server.
  - If the client does not know its own IP address, it uses 0.0.0.0.
  - If the client does not know the server's IP address, it uses the limited broadcast address (255.255.255.255).
  - The UDP port number is 67.

- The server receives the datagram and looks up the hardware address of the client in its configuration file, which contains the client's IP address.
- The server fills in the remaining fields in the UDP datagram and returns it to the client using UDP port 68.

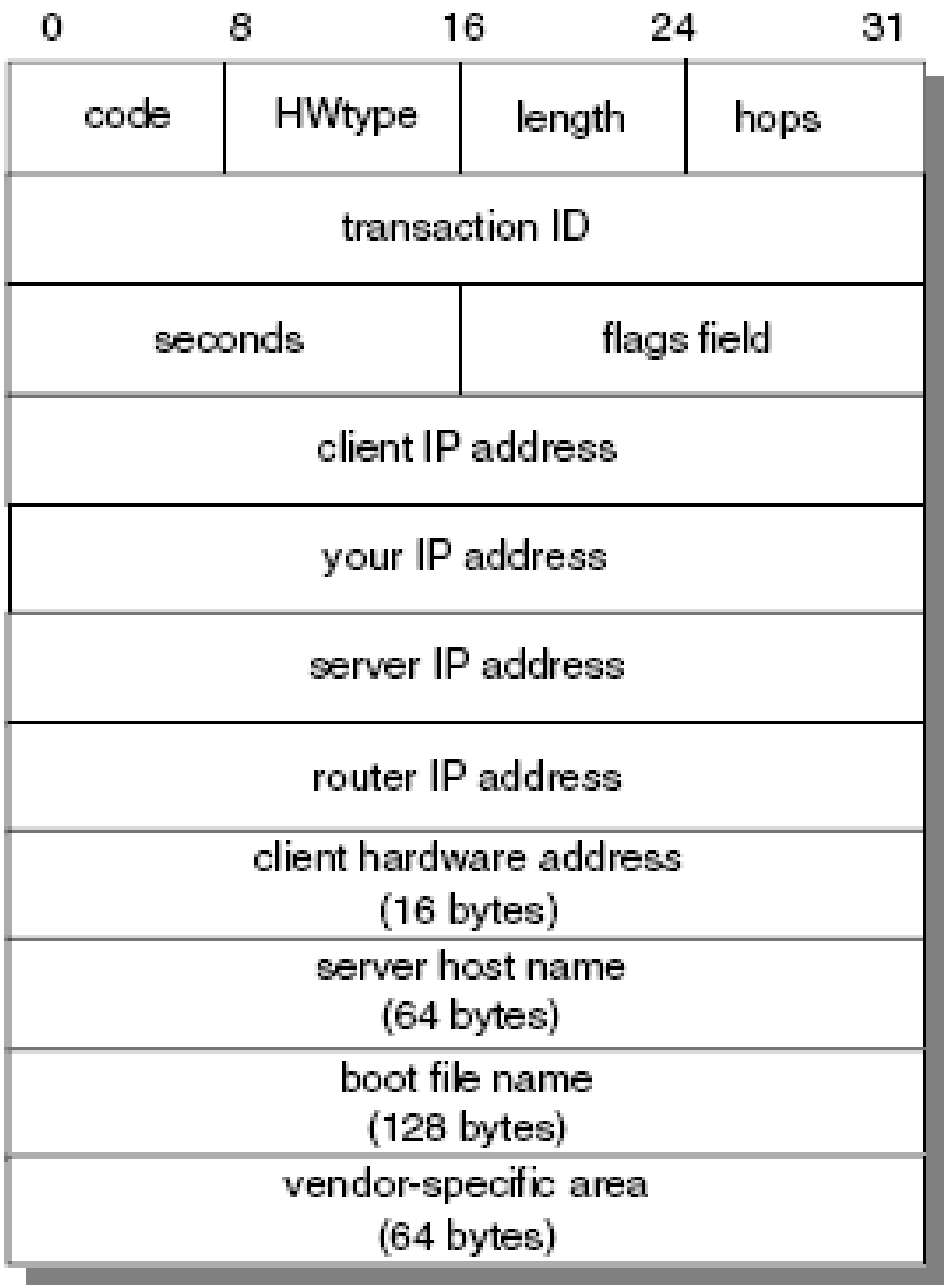
- code: Indicates a request or a reply.
  - 1: Request
  - 2: Reply
- HWtype: The type of hardware, for example:
  - 1: Ethernet
  - 6: IEEE 802 Networks
- Length
  - Hardware address length in bytes. Ethernet and token-ring both use 6, for example.
- Hops
  - The client sets this to 0.
  - It is incremented by a router that relays the request to another server and is used to identify loops.



0 8 16 24 31

code	HWtype	length	hops
transaction ID			
seconds		flags field	
client IP address			
your IP address			
server IP address			
router IP address			
client hardware address (16 bytes)			
server host name (64 bytes)			
boot file name (128 bytes)			
vendor-specific area (64 bytes)			

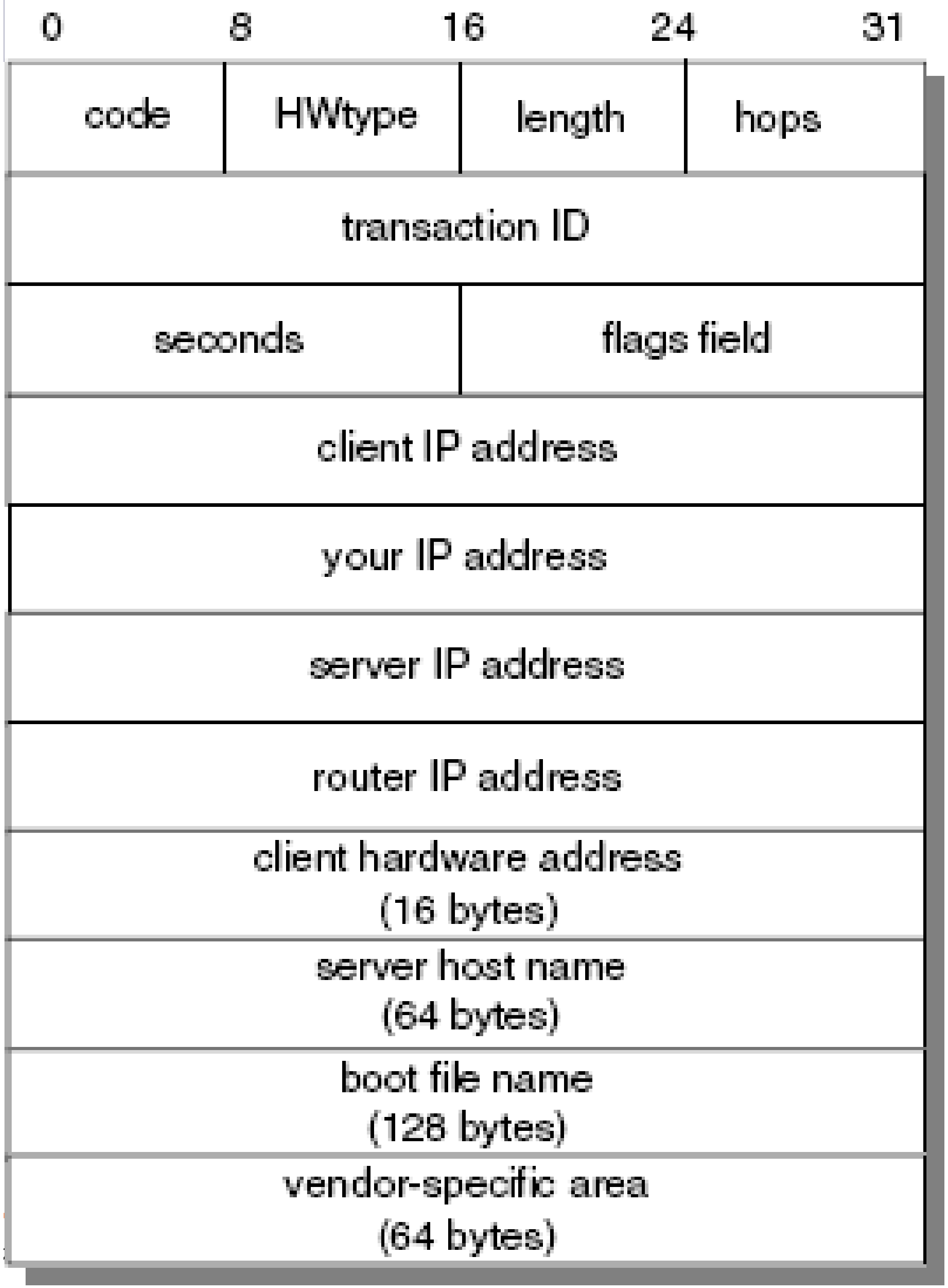
- Transaction ID
  - A random number used to match this boot request with the response it generates.
- Seconds
  - Set by the client. It is the elapsed time in seconds since the client started its boot process.
- Flags field
  - The most significant bit of the flags field is used as a broadcast flag. All other bits must be set to zero;
  - If a host is unable to receive a unicast IP datagram until it knows its IP address, then this broadcast bit must be set to indicate to the server that the BOOTREPLY must be sent as an IP and MAC broadcast. Otherwise this bit must be set to zero



0 8 16 24 31

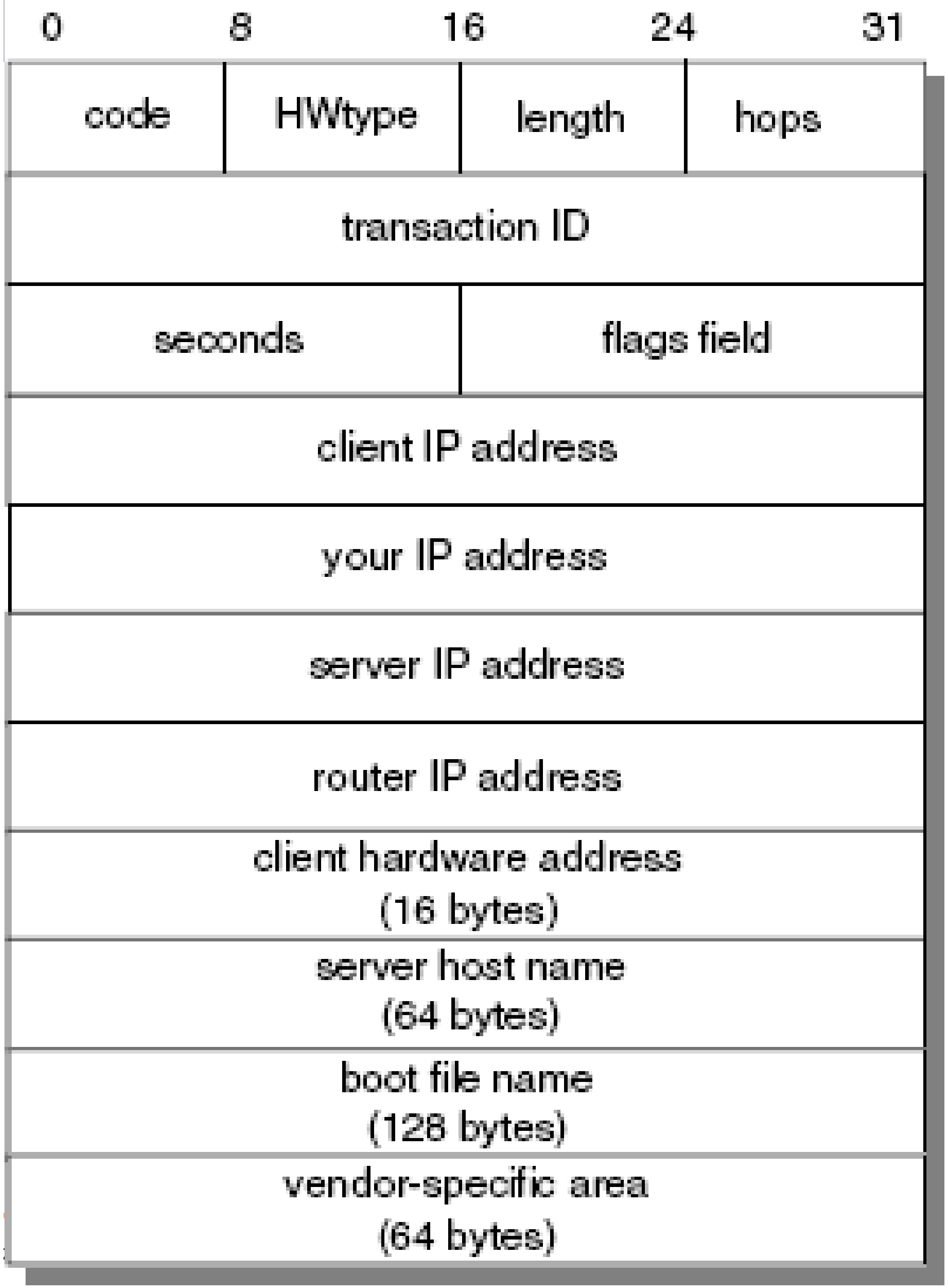
code	HWtype	length	hops
transaction ID			
seconds		flags field	
client IP address			
your IP address			
server IP address			
router IP address			
client hardware address (16 bytes)			
server host name (64 bytes)			
boot file name (128 bytes)			
vendor-specific area (64 bytes)			

- Client IP address
  - □ Set by the client, either to its known IP address or 0.0.0.0.
- Your IP address
  - □ Set by the server if the client IP address field was 0.0.0.0.
- Server IP address
  - □ Set by the server.
- Router IP address
  - □ This is the address of a BOOTP relay agent.
- Client hardware address
  - □ Set by the client and used by the server to identify which registered client is booting.



0	8	16	24	31
code	HWtype	length	hops	
transaction ID				
seconds		flags field		
client IP address				
your IP address				
server IP address				
router IP address				
client hardware address (16 bytes)				
server host name (64 bytes)				
boot file name (128 bytes)				
vendor-specific area (64 bytes)				

- Server host name
  - Optional server host name terminated by X'00'.
- Boot file name
  - The client either leaves this null or specifies a generic name, such as router indicating the type of boot file to be used.
  - The server returns the fully qualified file name of a boot file suitable for the client.
  - The value is terminated by X'00'.



0	8	16	24 31
code	HWtype	length	hops
transaction ID			
seconds		flags field	
client IP address			
your IP address			
server IP address			
router IP address			
client hardware address (16 bytes)			
server host name (64 bytes)			
boot file name (128 bytes)			
vendor-specific area (64 bytes)			

BOOTP forwarding

- BOOTP relay agent checks the hops field
- it checks the contents of the router IP address field. If this field is zero, it fills this field with the IP address of the interface on which the BOOTPREQUEST was received. If this field already has an IP address of another relay agent, it is not touched.
- The value of the hops field is incremented.
- The relay agent then forwards the BOOTPREQUEST to one or more BOOTP servers. The address of the BOOTP server(s) is preconfigured at the relay agent.

18.8 DHCP —— the reason

- The use of BOOTP allows centralized configuration of multiple clients.

- It requires a static table to be maintained with an IP address preallocated for every client, even if the client is seldom active.

- This means that there is no relief on the number of IP addresses required.

- A client will only be allocated an IP address by the server if it has a valid MAC address.

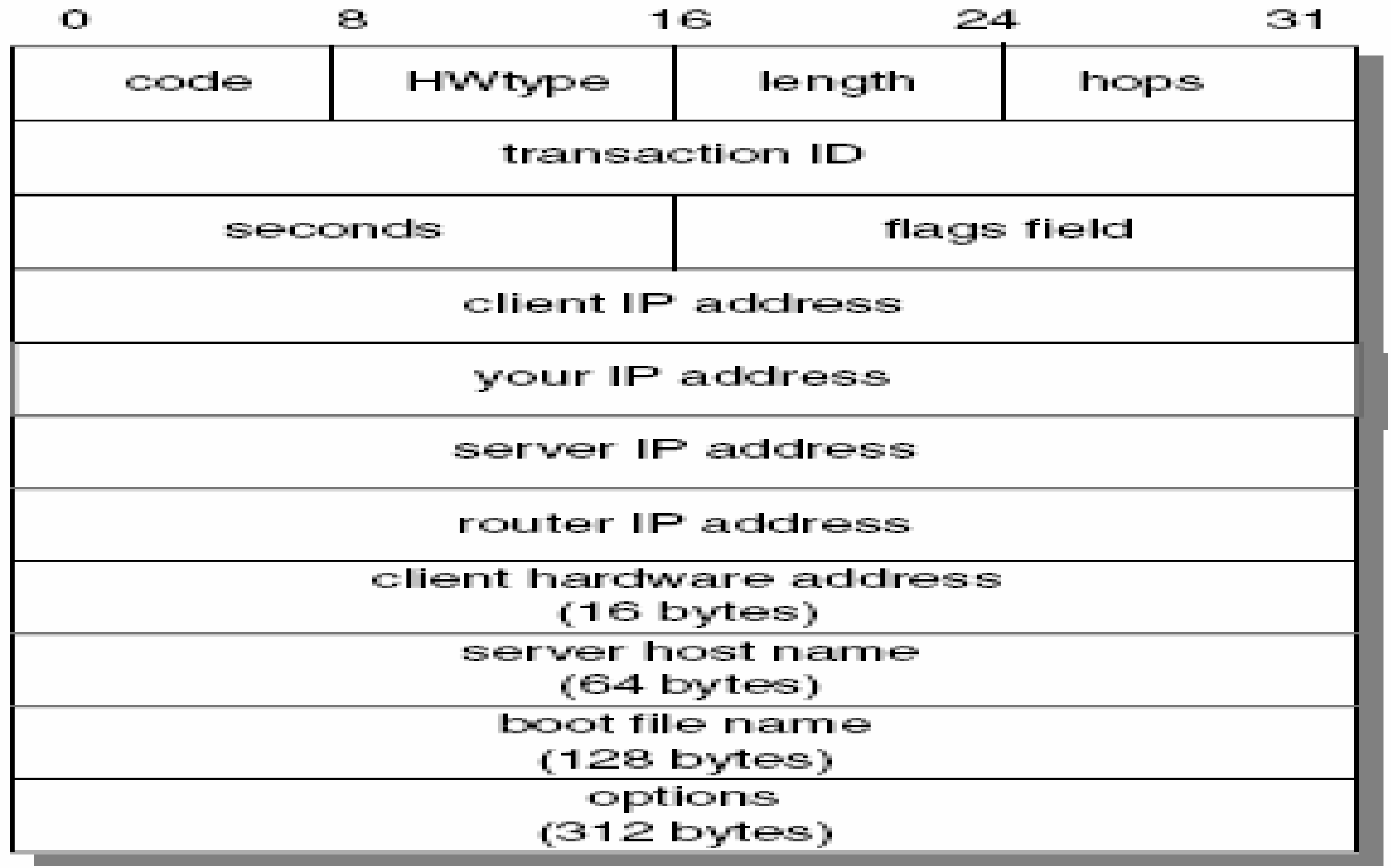
DHCP — specification

- RFC 2131：“ Dynamic Host Configuration Protocol”
- RFC 2132 ：“ DHCP Options and BOOTP Vendor Extensions”.
- DHCP is based on the BOOTP protocol
- Automatic allocation of reusable network addresses
- DHCP messages use UDP port 67, 68.
- DHCP participants can interoperate with BOOTP participants.

DHCP — three mechanisms for IP address allocation

- Automatic allocation
  - DHCP assigns a permanent IP address to the host.
- Dynamic allocation
  - DHCP assigns an IP address for a limited period of time. Such a network address is called a *lease*. This is the only mechanism that allows automatic reuse of addresses that are no longer needed by the host to which it was assigned.
- Manual allocation
  - The host's address is assigned by a network administrator.

DHCP — the message format



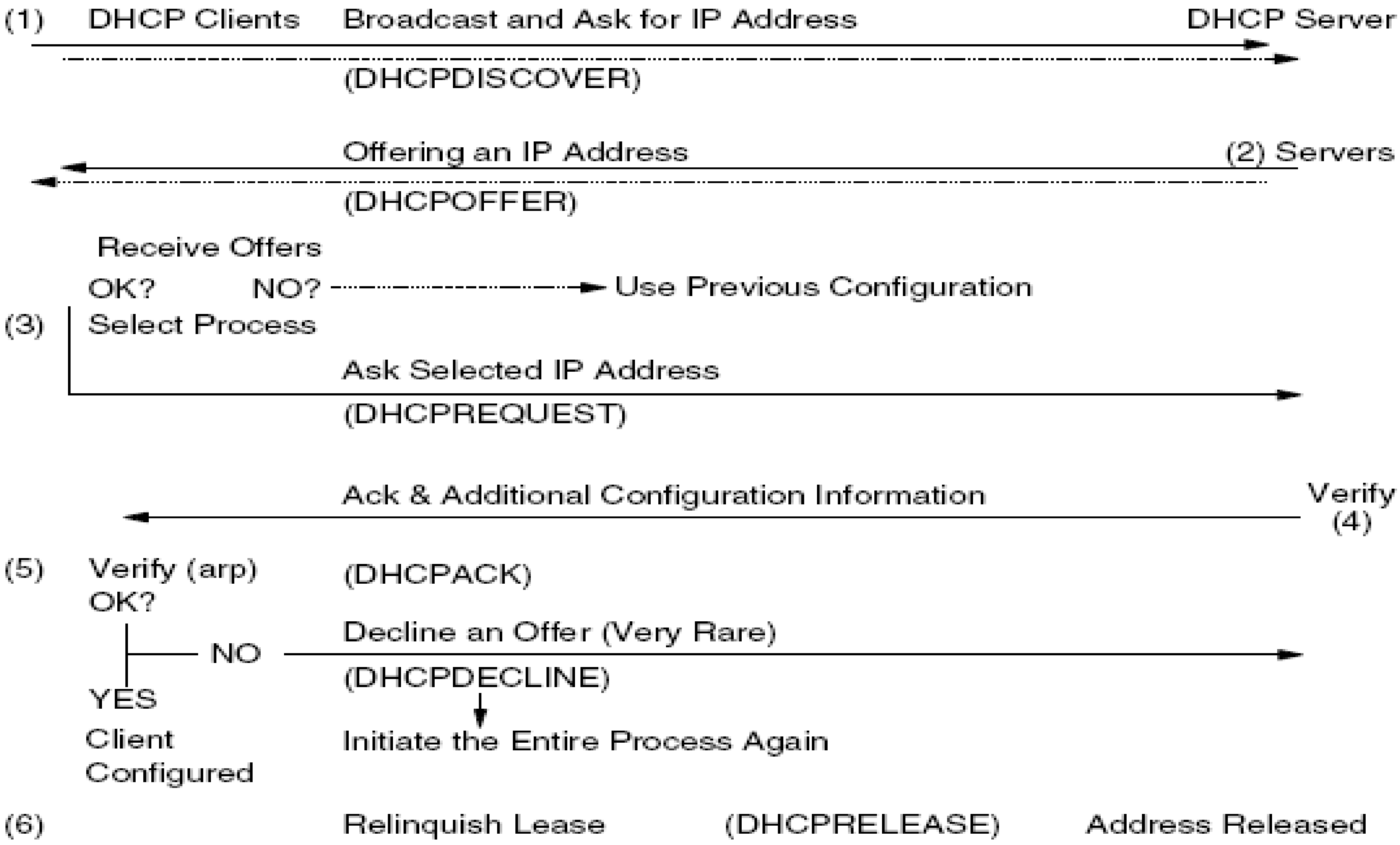
0 8 16 24 31

<table>
<tr><td>code</td><td>HWtype</td><td>length</td><td>hops</td></tr>
<tr><td colspan="4">transaction ID</td></tr>
<tr><td colspan="2">seconds</td><td colspan="2">flags field</td></tr>
<tr><td colspan="4">client IP address</td></tr>
<tr><td colspan="4">your IP address</td></tr>
<tr><td colspan="4">server IP address</td></tr>
<tr><td colspan="4">router IP address</td></tr>
<tr><td colspan="4">client hardware address
(16 bytes)</td></tr>
<tr><td colspan="4">server host name
(64 bytes)</td></tr>
<tr><td colspan="4">boot file name
(128 bytes)</td></tr>
<tr><td colspan="4">options
(312 bytes)</td></tr>
</table>

DHCP — message types

- DHCP**DISCOVER**
- DHCP**OFFER**
- DHCP**REQUEST**
- DHCP**ACK**.
- DHCP**NACK**
- DHCP**DECLINE**
- DHCP**RELEASE**
- DHCP**INFORM**

DHCP — Allocating a new network address



(1) DHCP Clients Broadcast and Ask for IP Address DHCP Server

(DHCPDISCOVER)

Offering an IP Address (2) Servers

(DHCPOFFER)

Receive Offers

OK? NO? Use Previous Configuration

(3) Select Process

Ask Selected IP Address

(DHCPREQUEST)

Ack & Additional Configuration Information Verify (4)

(5) Verify (arp) OK? (DHCPACK)

Decline an Offer (Very Rare)

NO

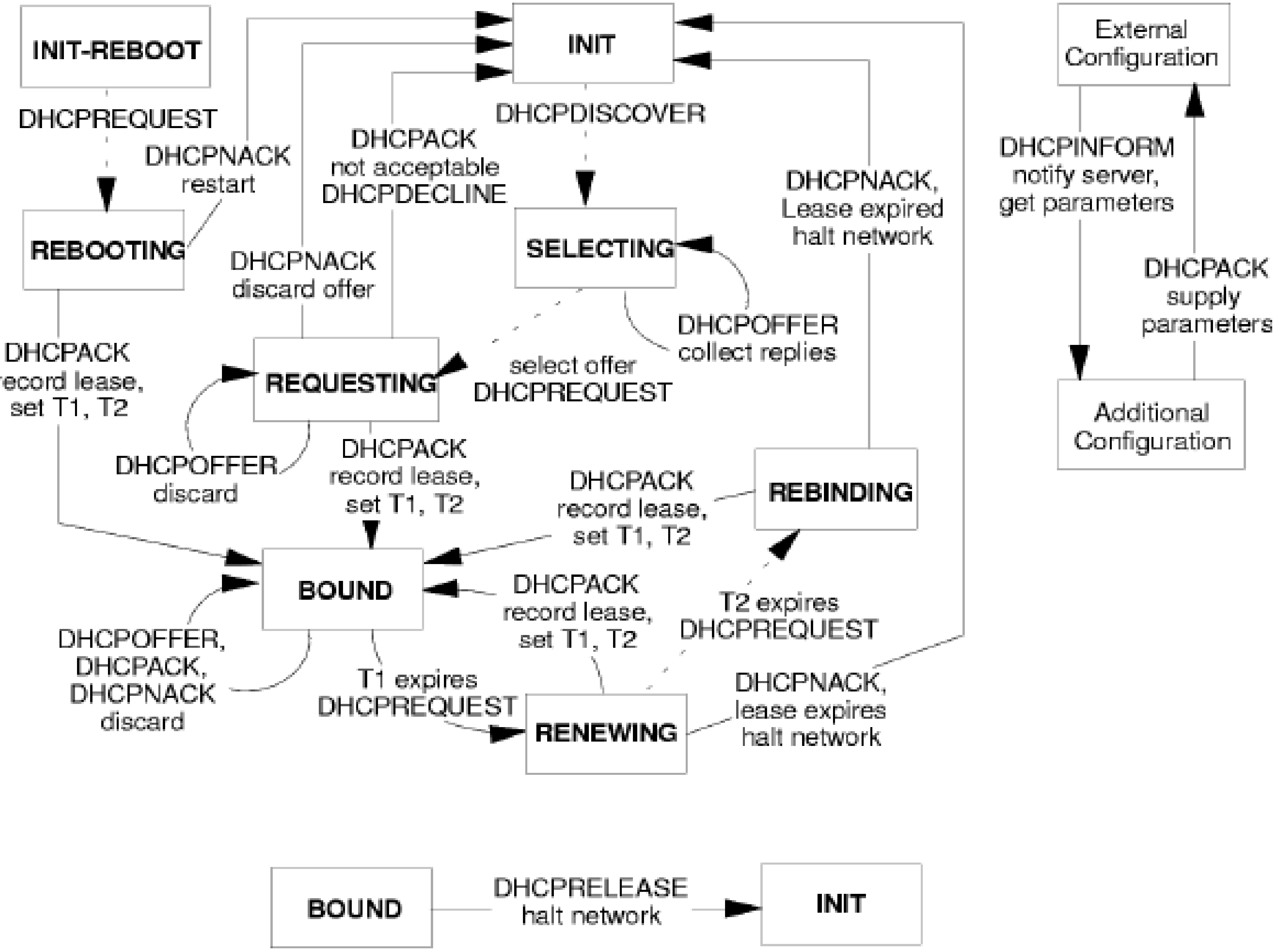
(DHCPDECLINE)

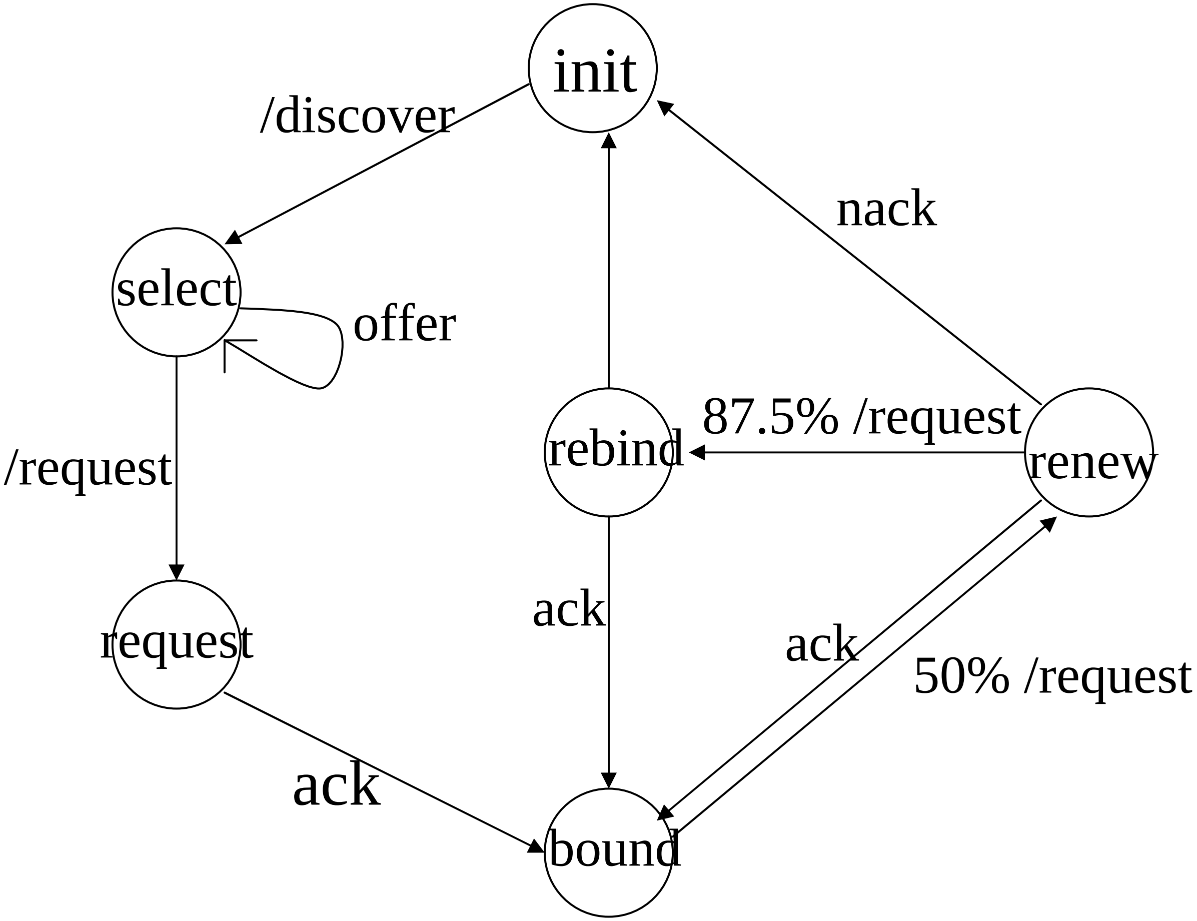
YES

Initiate the Entire Process Again

Client Configured

(6) Relinquish Lease (DHCPRELEASE) Address Released


INIT-REBOOT
INIT
External Configuration
DHCPREQUEST
DHCPNACK restart
DHCPDISCOVER
DHCPACK not acceptable DHCPDECLINE
DHCPINFORM notify server, get parameters
DHCPNACK, Lease expired halt network
REBOOTING
SELECTING
DHCPNACK discard offer
DHCPACK supply parameters
DHCPOFFER collect replies
DHCPACK ecord lease, set T1, T2
REQUESTING
select offer DHCPREQUEST
Additional Configuration
DHCPOFFER discard
DHCPACK record lease, set T1, T2
DHCPACK record lease, set T1, T2
REBINDING
BOUND
DHCPACK record lease, set T1, T2
T2 expires DHCPREQUEST
DHCPOFFER, DHCPACK, DHCPNACK discard
T1 expires DHCPREQUEST
RENEWING
DHCPNACK, lease expires halt network
BOUND
DHCPRELEASE halt network
INIT


init
/discover
select
offer
nack
/request
rebind
87.5% /request
renew
request
ack
ack
50% /request
ack
bound