Classic Internet Protocols

Dr. Ying-Dar Lin

High Speed Network Lab.

Department of Computer and Information Science

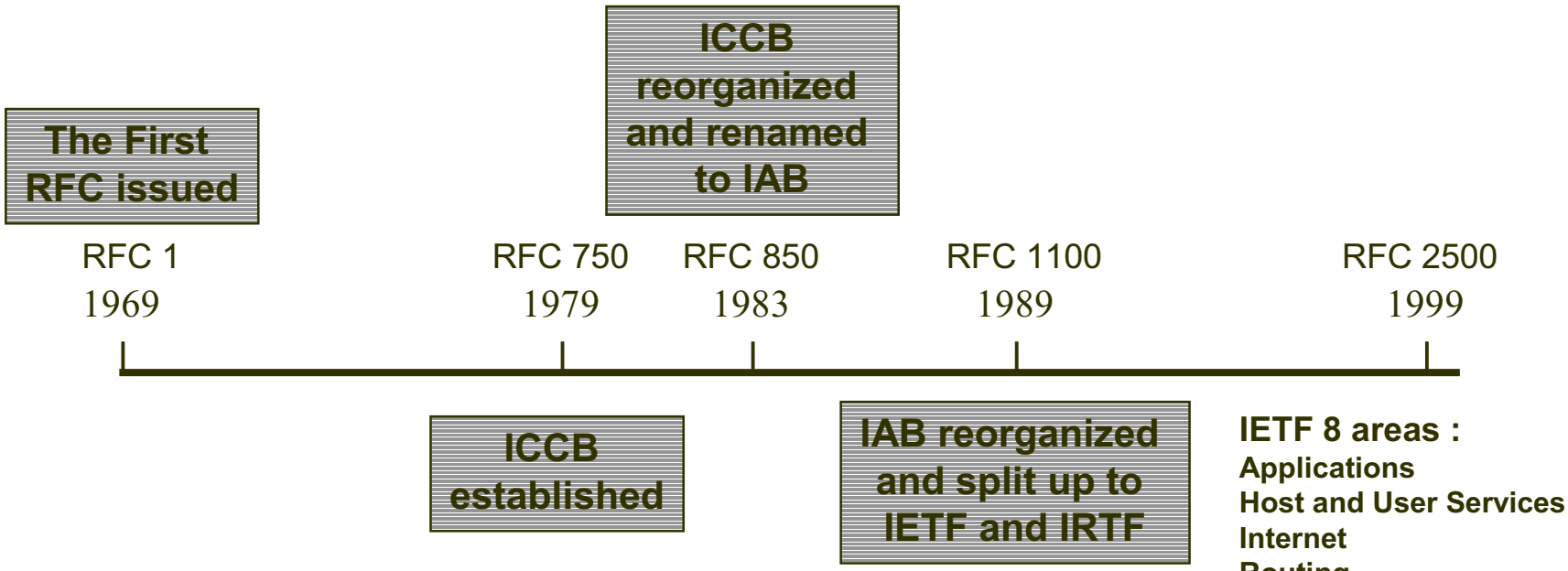
National Chiao Tung University

May 28, 1999

Content

- 30 Years of RFCs
- IP Design Philosophy
- Internet Protocol
- Internet Control Protocols
- Routing Protocols

30 Years of RFCs

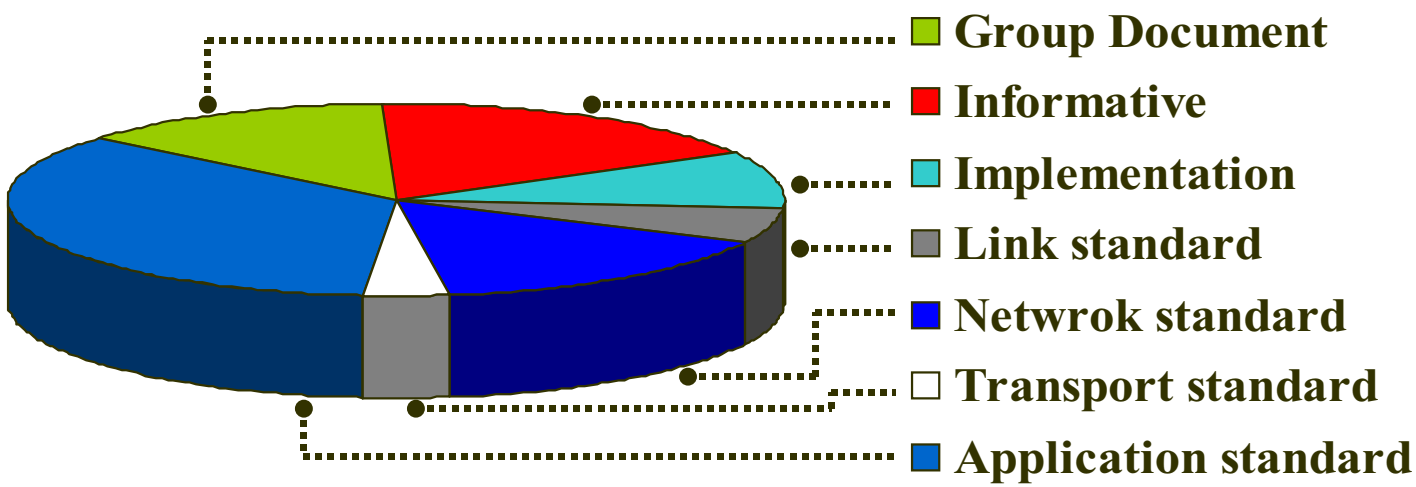


ICCB : Internet Configuration Control board
IAB : Inter Activities Board
IETF : Internet Engineering Task Force
IRTF : Internet Research Task Force

RFC Classification

- Group Document(~350)
- Informative(~450)
  - ideas , discussion
- Implementation(~250)
  - experiences, measurement

- Standard
  - Link(~150)
    - Data-link layer
  - Network(~400)
    - Network layer
  - Transport(~100)
    - Transport, Session and Presentation layers
  - Application(~900)
    - Application layer



Link Standards

RFCs	Name	Content
1055	SLIP	IP on the Serial Line
1661, 1662, 1663	PPP	Point-to-point protocol

Network Standards

RFCs	Name	Content
760, 791, 815, 1154	IP	Internet Protocol
777, 792	ICMP	Internet Control Message Protocol
1245, 1246 , 1247	OSPF	Open Shortest Path Finder Routing Protocol

Transport Standards

RFCs	Name	Content
768	UDP	User Datagram Protocol
675, 761, 793, 896	TCP	Transmission Control Protocol

Application Standards

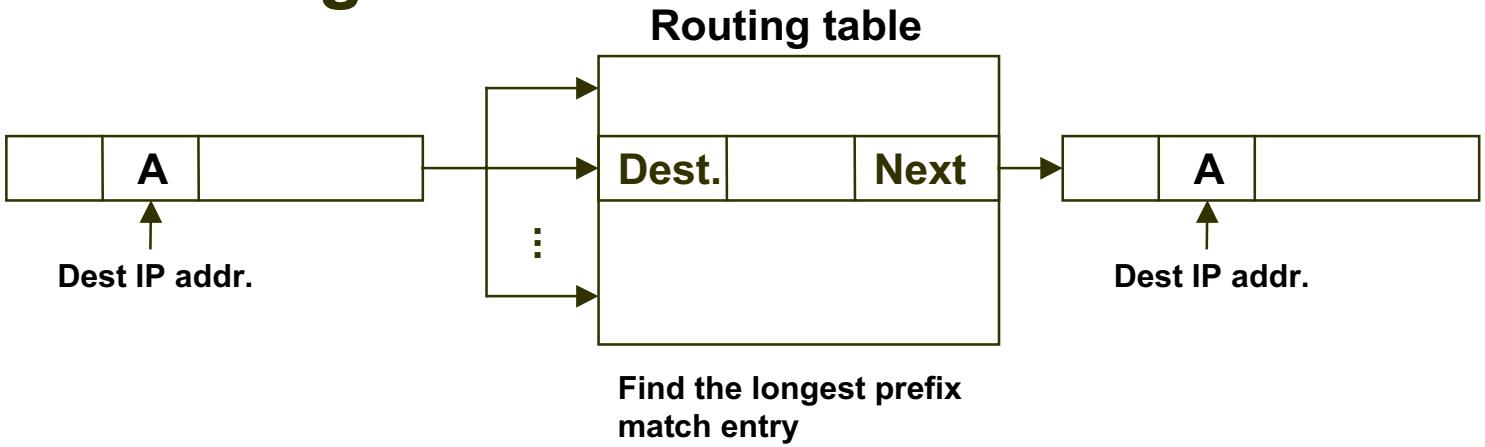
RFCs	Name	Content
318, 435, 495	Telnet	Remote login to host
114, 354, 959	FTP	File transfer protocol
1866, 2068, 1045, 2069	HTTP	Hypertext transfer protocol

IP Design Philosophy

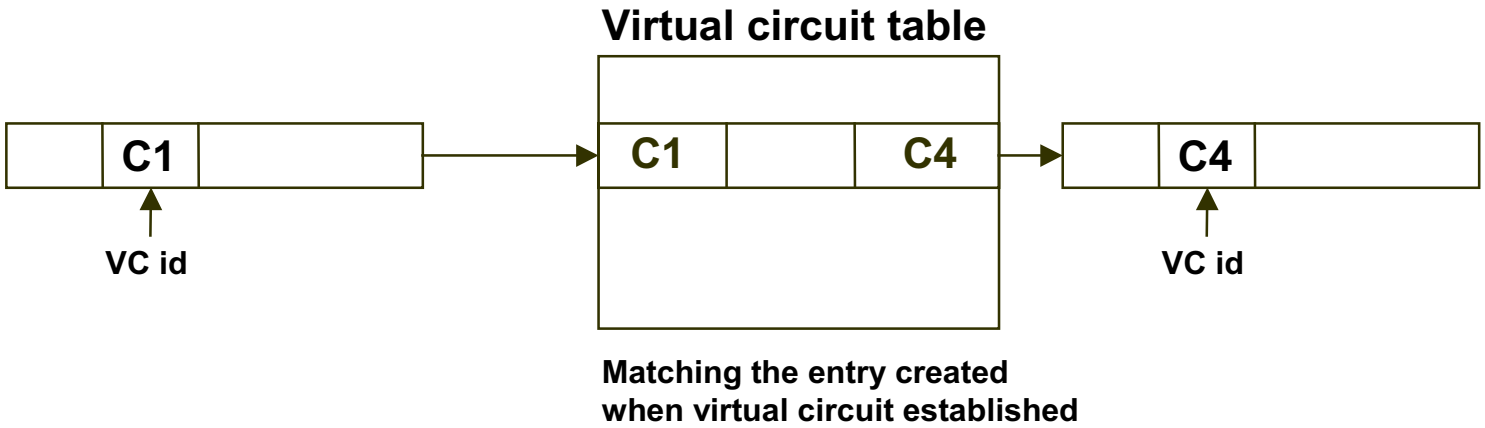
- Transmitting blocks of data called datagrams
- Connectionless: source and destination hosts identified by address in datagram headers
- Stateless : datagrams routed indendently
- Fragment and reassembly for long datagrams
- Support networks to provide various types of service.

Routing vs Switching

- **Routing: stateless**

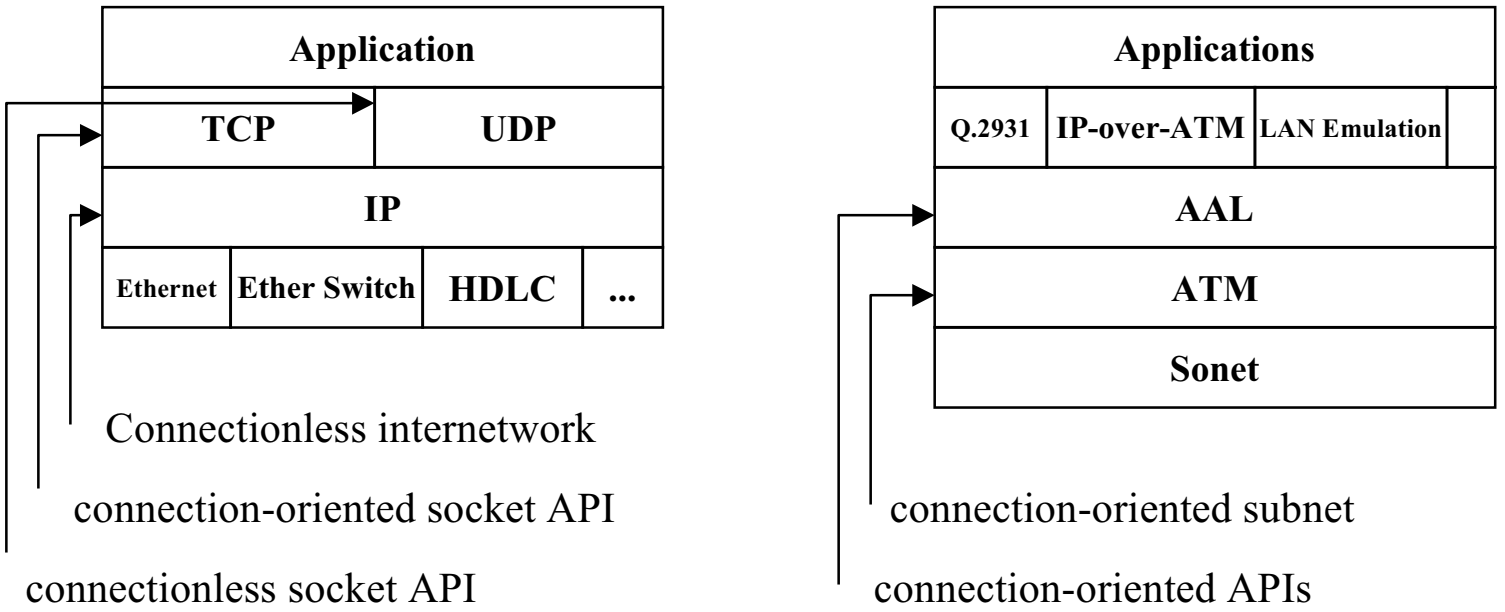


- **Switching: stateful**

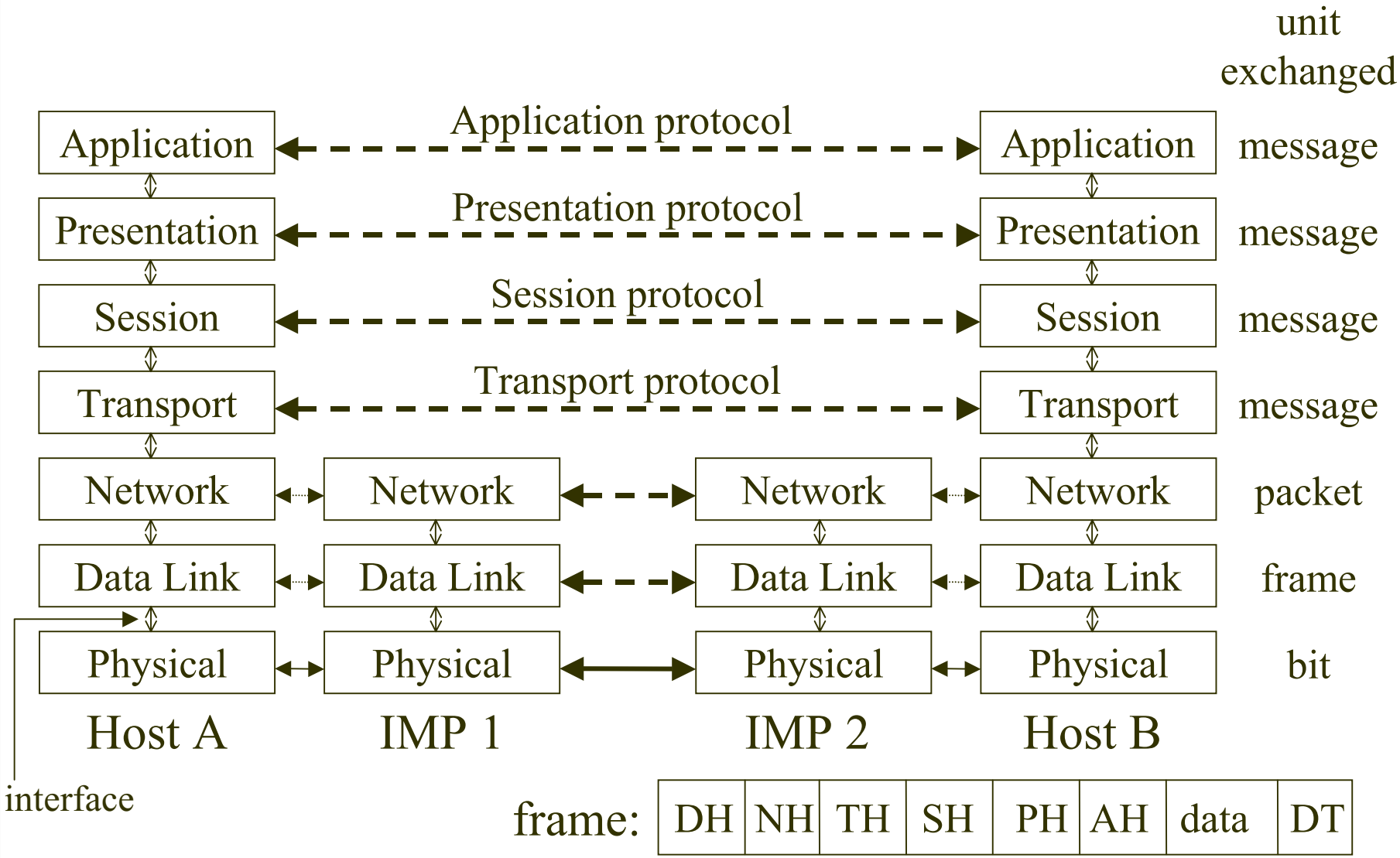


IP vs ATM

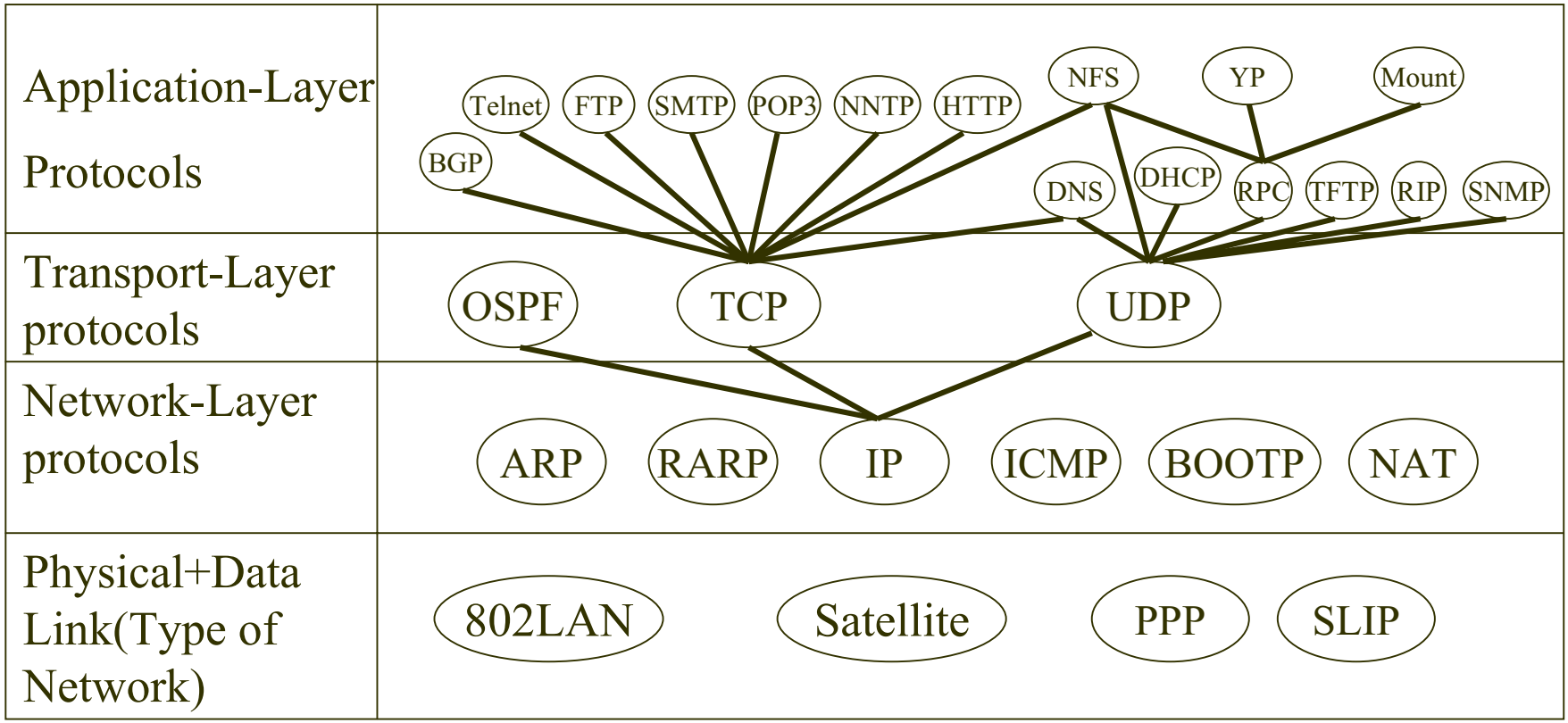
- datagram vs virtual-circuit
- stateless flow vs hard-state connection

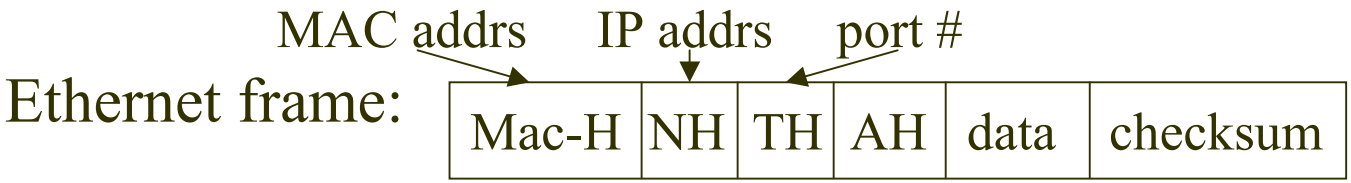


OSI Architecture

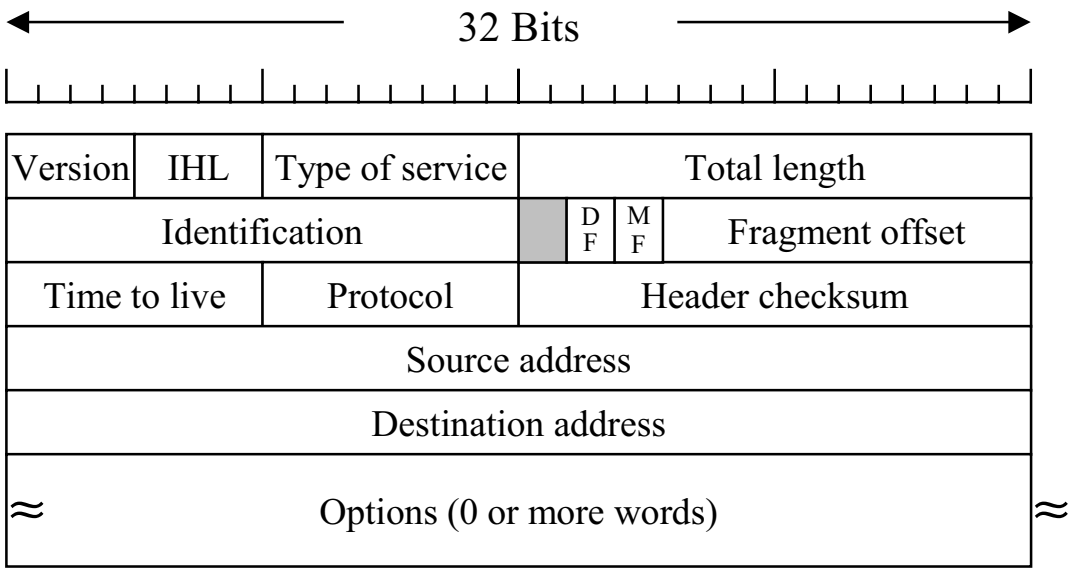


Internet Protocol Hierarchy





IP Header



Option	Description
Security	Specifies how secret the data gram is
Strict source routing	Gives the complete path to be followed
Loose source routing	Gives a list of routers not to be missed
Record route	Makes each router append its IP address
Timestamp	Makes each router append its address and timestamp

IP Address and Netmask

- IP Address is 32 bits long.
- Netmask: all 1s except the host field, determine the subnet of the address

- IP Address with all 0s host field means the network address.
- IP Address with all 1s host field means limited broadcast address in the same subnet.

```
                IP address : 10001100 01110001 01011000 10101100
                & netmask : 11111111 11111111 11111111 00000000
= network address  : 10001100 01110001 01011000 00000000
              | inverse netmask : 00000000 00000000 00000000 11111111
= limited broadcast address : 10001100 01110001 01011000 11111111
```

Netmask Operations (1/3)

- Host sends to an other host
  - 140.113.88.1 netmask 255.255.255.0
  - send to 140.113.88.2
    - 140.113.88.1 & 255.255.255.0 = 140.113.88.0
    - 140.113.88.2 & 255.255.255.0 = 140.113.88.0
    - 140.113.88.0 = 140.113.88.0 matched ,
    - 140.113.88.1 sends packets to 140.113.88.2 directly
  - send to 140.113.89.2
    - 140.113.88.1 & 255.255.255.0 = 140.113.88.0
    - 140.113.89.2 & 255.255.255.0 = 140.113.89.0
    - 140.113.88.0 ≠ 140.113.89.0 unmatched ,
    - 140.113.88.1 sends packets to router

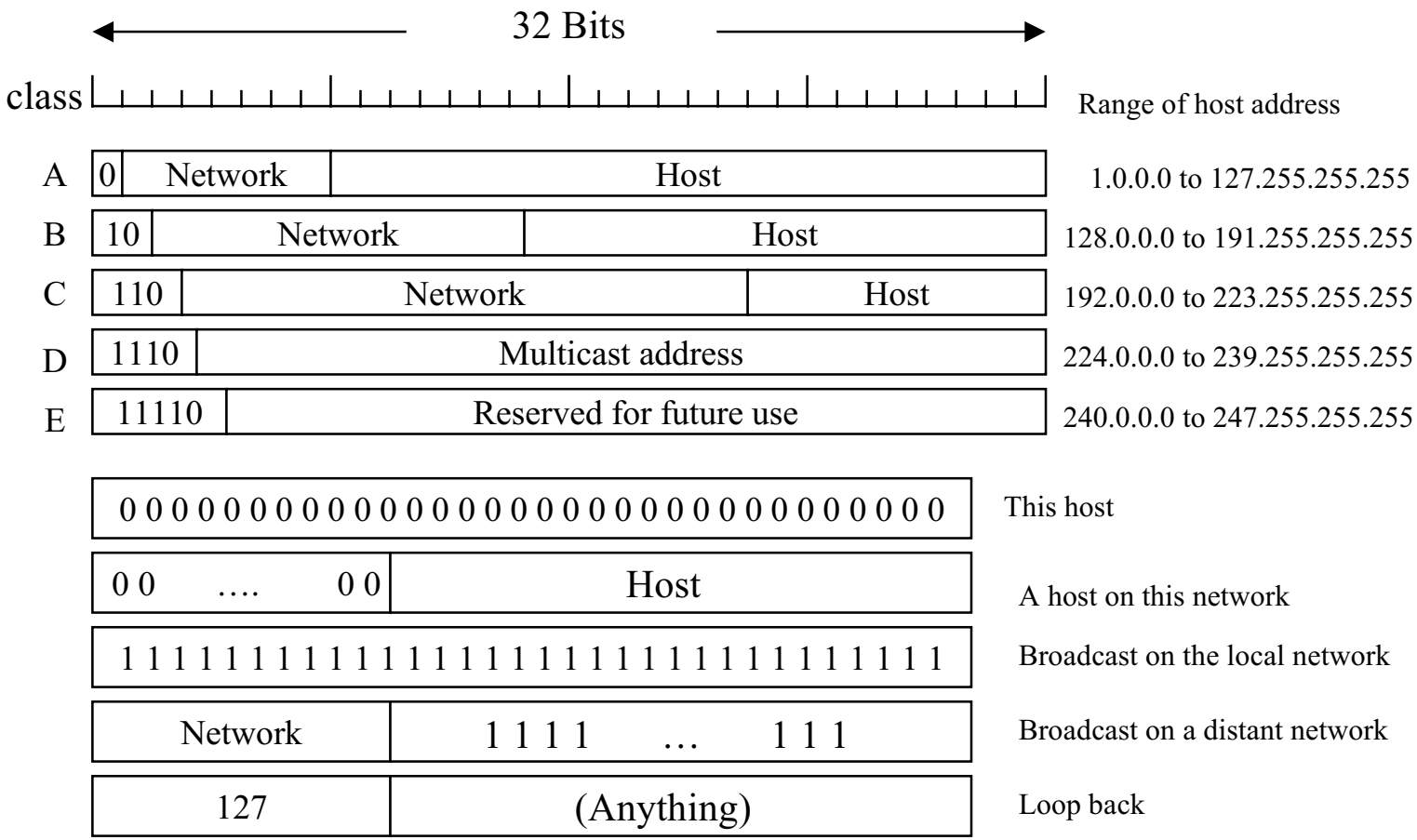
Netmask Operations (2/3)

- Routing table lookup
  - forward a packet to 140.113.88.1
    - Entry 1 140.113.0.0 netmask 255.255.0.0 to gw 1
      - 140.113.88.1 & 255.255.0.0 = 140.113.0.0
      - 140.113.0.0 = 140.113.0.0, matched
    - Entry 2 140.113.87.0 netmask 255.255.255.0 to gw 2
      - 140.113.88.1 & 255.255.255.0 = 140.113.88.0
      - 140.113.88.0 ≠ 140.113.87.0, unmatched
    - Entry 3 140.113.88.0 netmask 255.255.255.0 to gw 3
      - 140.113.88.1 & 255.255.255.0 = 140.113.88.0
      - 140.113.88.0 = 140.113.88.0, matched
    - Match entry 1 & entry 3
      - entry 3 have longer 1s netmask than entry 1
    - packet forward to gw 3

Netmask Operations (3/3)

- Limited Broadcast
  - IP 140.113.88.1 netmask 255.255.255.0
  - limited broadcast address is 140.113.88.255
  - send packet to 140.113.88.255
    - all hosts have network address 140.113.88.0 will get the packet

Classes & Loop back



Classless InterDomain Routing

- More than half of all class B networks have fewer than 50 hosts. Too many small class C networks would enlarge routing table dramatically.
- CIDR: allocate the remaining class C networks (about 2 million) in variable-sized blocks and exercise longest prefix matching

194.0.0.0~195.255.255.255: for Europe (194.xx.xx.xx and 195.xx.xx.xx entries)
198.0.0.0~199.255.255.255: for North America
200.0.0.0~201.255.255.255: for Central and South America
202.0.0.0~203.255.255.255: for Asia and the Pacific
204.0.0.0~223.255.255.255: reversed for future use

Internet Control Protocols

- ICMP (Internet Control Message Protocol)
- DNS (Domain Name System)
  - Symbolic name 

 IP address
- ARP (Address Resolution Protocol) & RARP (Reverse Address Resolution Protocol)
  - IP address 

 MAC address
- BOOTP & DHCP
- NAT (Network Address Translation)
- PPP & SLIP

Internet Control Message Protocol

- To report an error in datagram processing
  - For example: ping, traceroute

Message Type	Description
Destination unreachable	Packet could not be delivered
Time exceeded	Time to live field hit 0
Parameter problem	Invalid header field
Source quench	Choke packet
Redirect	Teach a router about geography
Echo request	Ask a machine if it is alive
Echo reply	Yes, I am alive
Timestamp request	Same as Echo request, but with timestamp
Timestamp reply	Same as Echo reply, but with timestamp

DNS (Domain Name System)

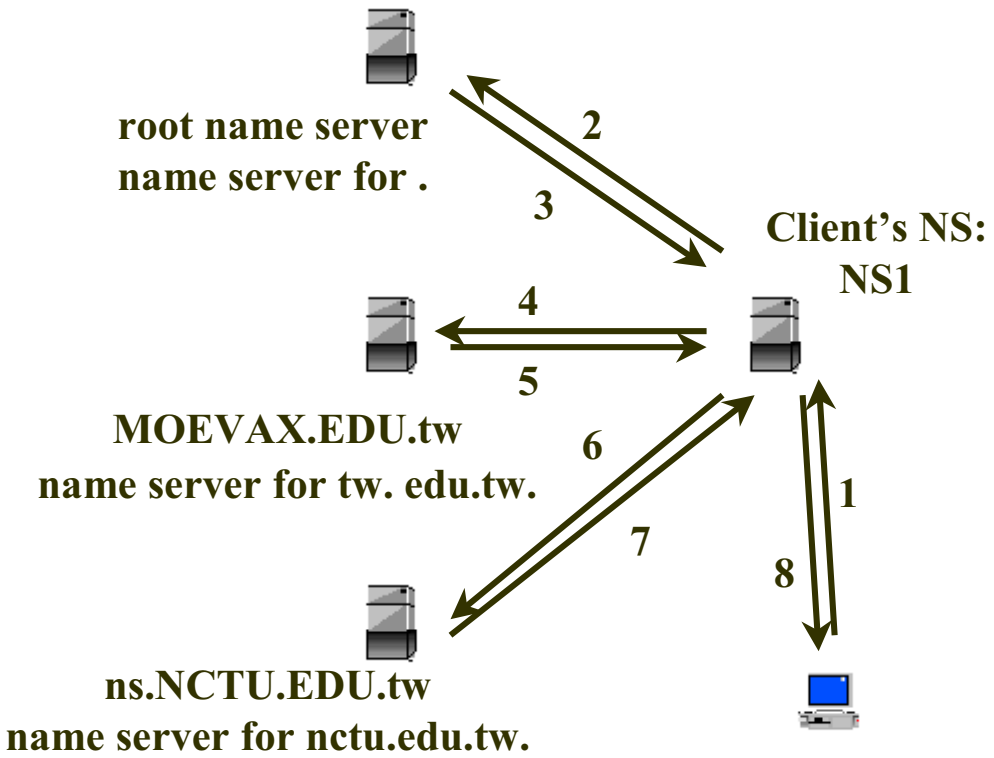
- Mapping the name of host to its IP address
- The DNS name space
  - Internet is divided into several top-level domains
  - Each domain is named by the path upward from it to the (unnamed) root
- Name Server
  - Nonoverlapping zones in DNS name space
  - One primary name server and one or more secondary name servers

DNS Resource Records

Type	Meaning	Value
SOA	Start of Authority	Parameters of this zone
A	IP address of a host	32-Bit integer
MX	Mail exchange	Priority, domain willing to accept email
NS	Name Server	Name of a server for this domain
CNAME	Canonical name	Domain name
PTR	Pointer	Alias for an IP Address
HINFO	Host description	CPU and OS in ASCII
TXT	Text	Uninterpreted ASCII text

```
@                IN    SOA      ns.NCTU.edu.tw.   hostmaster.nctu.edu.tw. (
                                  1997071001   ; Serial
                                  86400        ; Refresh - 1 days
                                  1800         ; Retry
                                  1728000      ; Expire - 20 days
                                  259200       ); Minimum TTL - 3 days
                 IN    NS       NCTU.edu.tw.
;
NCTU.edu.tw.     IN    A        140.113.1.1
                 IN    MX       0  ns1.NCTU.edu.tw.
nctu             IN    CNAME    NCTU.edu.tw.
```

DNS query sequence



1. Client query www.nctu.edu.tw's IP on its NS1
2. NS1 query on root NS
3. Root NS tell NS1 to query on moevax.edu.tw
4. NS query on moevax.edu.tw
5. moevax.edu.tw tell NS1 to query on ns.nctu.edu.tw
6. NS1 query on ns.nctu.edu.tw
7. ns.nctu.edu.tw tell NS1 www.nctu.edu.tw's IP
8. NS1 tell theclient www.nctu.edu.tw's IP

ARP & RARP

- ARP (Address Resolution Protocol)
  - The Data Link Layer doesn't understand the IP address
  - Use ARP to ask the MAC address of destination host
- RARP (Reverse Address Resolution Protocol)
  - Send a limited broadcast message to get its IP address
  - Allow a diskless host to get its IP address from other host or network device

BOOTP & DHCP

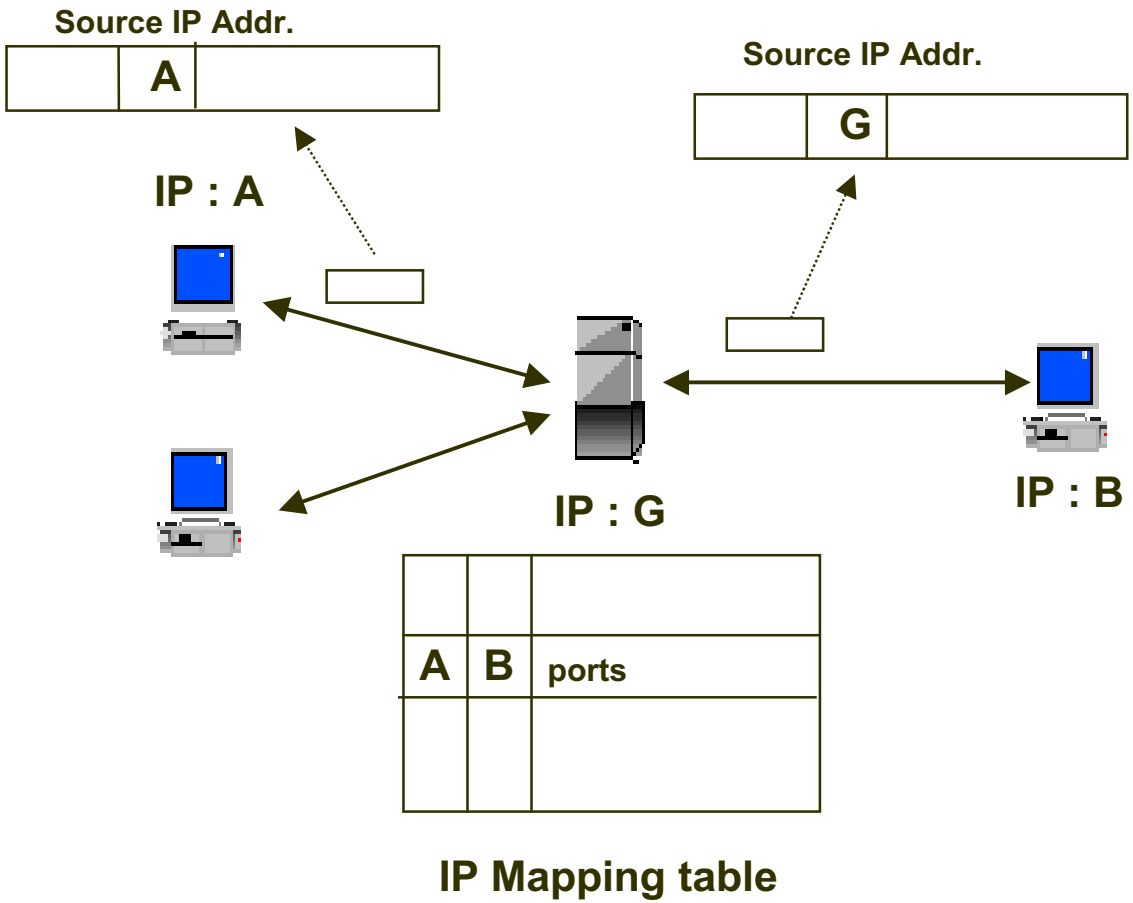
- BOOTP (Bootstrap Protocol)
  - Allow a diskless client to get its IP Address.
  - RARP server must be in the same network.
  - BOOTP runs over data link or UDP message(can be forwarded by router)
- DHCP (Dynamic Host Configuration Protocol)
  - Derived on BOOTP, runs over UDP
  - Automatic allocation of reusable network addresses

Get IP From Network

	RARP (kernel)	BOOTP (kernel & daemon)	DHCP (daemon)
Range	Only on local network	Can pass through gateway	Can pass through gateway
Underlying protocol	data link	UDP	UDP
IP address	Fixed	Fixed	Automatic dynamic Allocated

NAT (Network Address Translation)

e.g. IP masquerade in Linux



- The IP addresses in a stub domain are not globally unique, they can be reused in the other domains.
- When communicate with outside x, translate the local address to the globally unique address.
- It can be installed incrementally without changing either hosts or routers.
- Special support needed for some protocols : ftp, irc, ICQ, Real Audio, VDO live

PPP & SLIP

- PPP (Point-to-Point Protocol)
  - Encapsulating datagrams over serial links.
  - An extensible Link Control Protocol (LCP)
  - Network Control Protocol (NCP) for establishing and configuring different network-layer protocols.
- SLIP (Serial Line IP)
  - Addressing
  - Type identification
  - Error detection / correction
  - Compression

Routing Protocols

- RIP (Routing Information Protocol)
- OSPF (Open Shortest Path First)
- BGP (Border Gateway Protocol)
- IS-IS (Intermediate System to Intermediate System Routing Protocol)
- Mobile IP

RIP (Routing Information Protocol)

Interior Gateway Routing Protocol for AS

- Distance vector routing
- Use hop count to be metric
- Broadcasting of responses
- Update Routing table either at regular 30-second intervals or triggered
- RIP-2 support for CIDR, authentication and multicast transmission

RIP Message Formats

Command(1)	Version(1)	Must be Zero(2)
Address Family Identifier(2)		Must be Zero(2)
IP Address(4)		
Must be Zero(4)		
Must be Zero(4)		
Metric(4)		

RIP-1

Command(1)	Version(1)	Routing Domain(2)
Address Family Identifier(2)		Route Tag(2)
IP Address(4)		
Subnet Mask(4)		
Next Hop(4)		
Metric(4)		

RIP-2

Routing Domain : group routers in domains which share routing information
Route Tag : propagate the information acquired from an EGP
Subnet Mask : subnet mask for CIDR implementation
Next Hop : next hop addresses that allows for optimization routes

OSPF (Open Shortest Path First)

Interior Gateway Routing Protocol for AS

- **Link state routing**
- **Requirements: open, variety of distance metrics, dynamic, based on type of service, load balancing, hierarchical security, tunnel allowed**
- **Each AS, divided into areas, has a backbone area connecting all other areas, by point-to-point links, broadcast links, or tunnels in WAN**
- **Three kinds of routes:**
  - intra-area: link state shorting path routing
  - inter-area: (1) from source to backbone,
    (2) across backbone to destination area,
    (3) to destination
  - inter-AS: BGP - exterior gateway routing protocol

OSPF Messages

- **The Common Header:**

Version #	Type	Packet length
Router ID		
Area ID		
Checksum		Autype
Authentication		
Authentication		

- **Hello message**
  - Used to discover who the neighbors are

OSPF packet header, type = 1 (hello)		
Network mask		
Hello Interval	Options	Priority
Dead Interval		
Designated router		
Backup designated router		
Neighbor		

Neighbor		

- **Database description**
  - Announces which updates the sender has

OSPF packet header, type = 2 (dd)			
0	0	Options	0 IMMs
DD sequence number			
Link State Type			
Link State ID			
Advertising router			
Link State sequence number			
Link State checksum		Link State age	

- **Link state request**
  - Requests information from the partner

OSPF packet header, type = 3 (rq)
Link State Type
Link State ID
Advertising router

- **Link state update**
  - Provides the sender's costs to its neighbors

OSPF packet header, type = 4 (uqd.)
Number of advertisements
Link State advertisements

- **Link state update**
  - Provides the sender's costs to its neighbors

OSPF packet header, type = 5 (ack.)
Link State advertisements headers

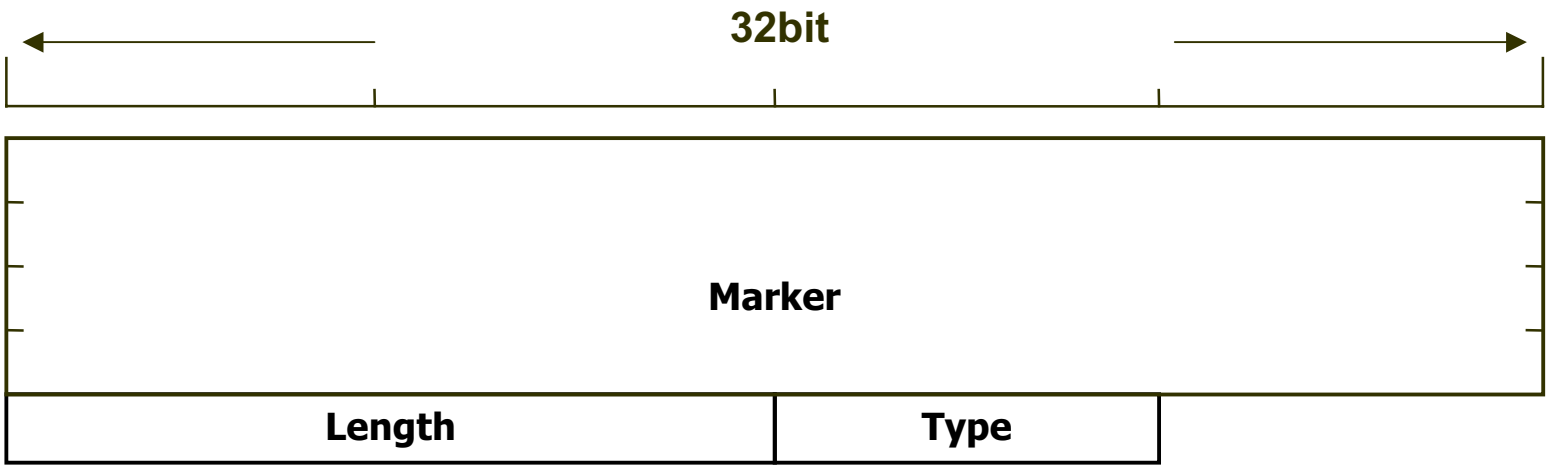
BGP (Border Gateway Protocol)

Exterior Gateway Routing Protocol for Inter-AS

- Designed to allow many kinds of routing policies to be enforced in the inter AS traffic, by route scoring function
- A distance vector protocol where BGP router keeps track of the exact path used in routing table and tells its neighbor the exact path, instead of cost, it is using
- Pairs of BGP routers communicate with each other by establishing TCP connections
- No count-to-infinity problem

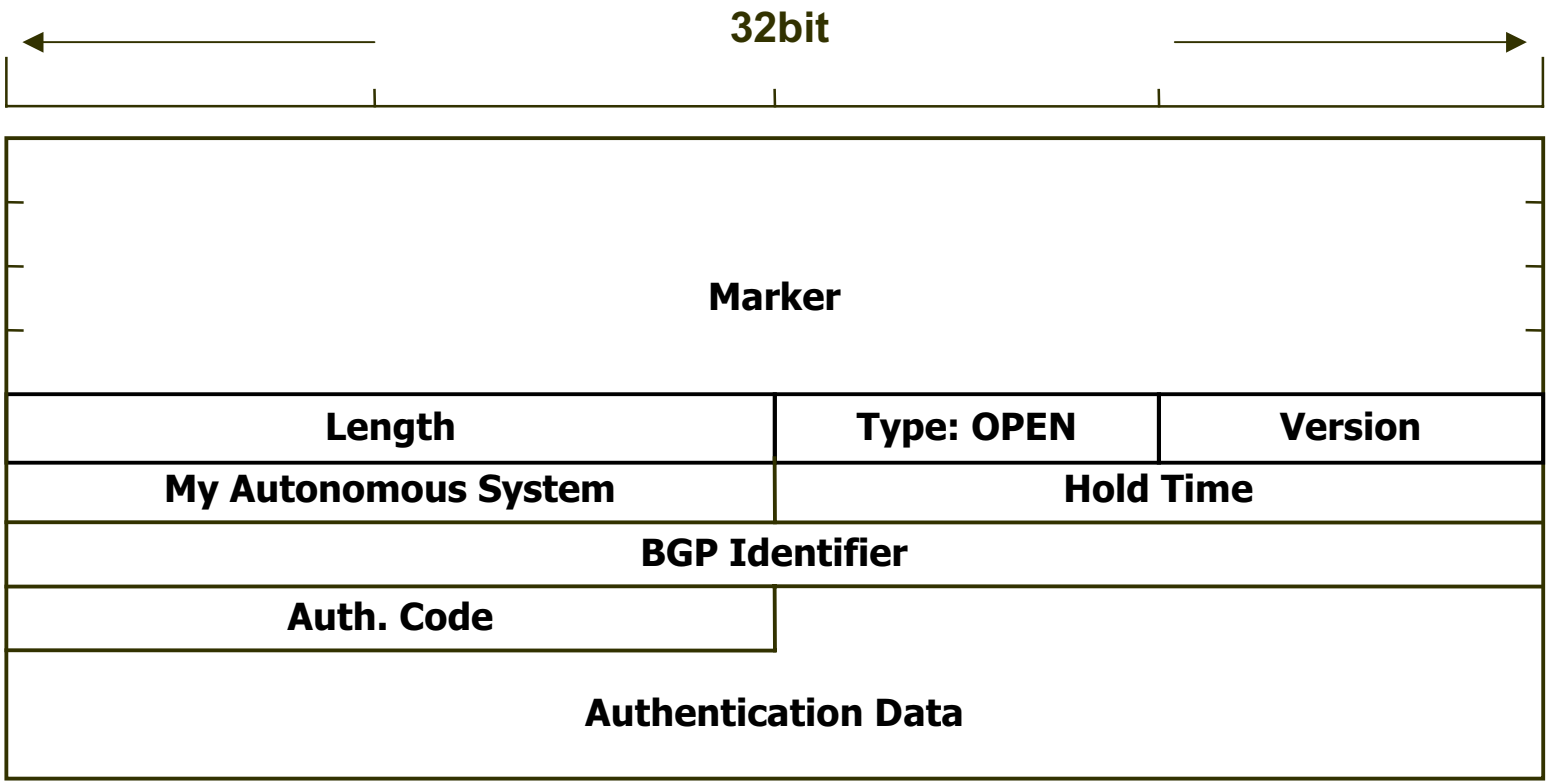
BGP messages

- Common Header

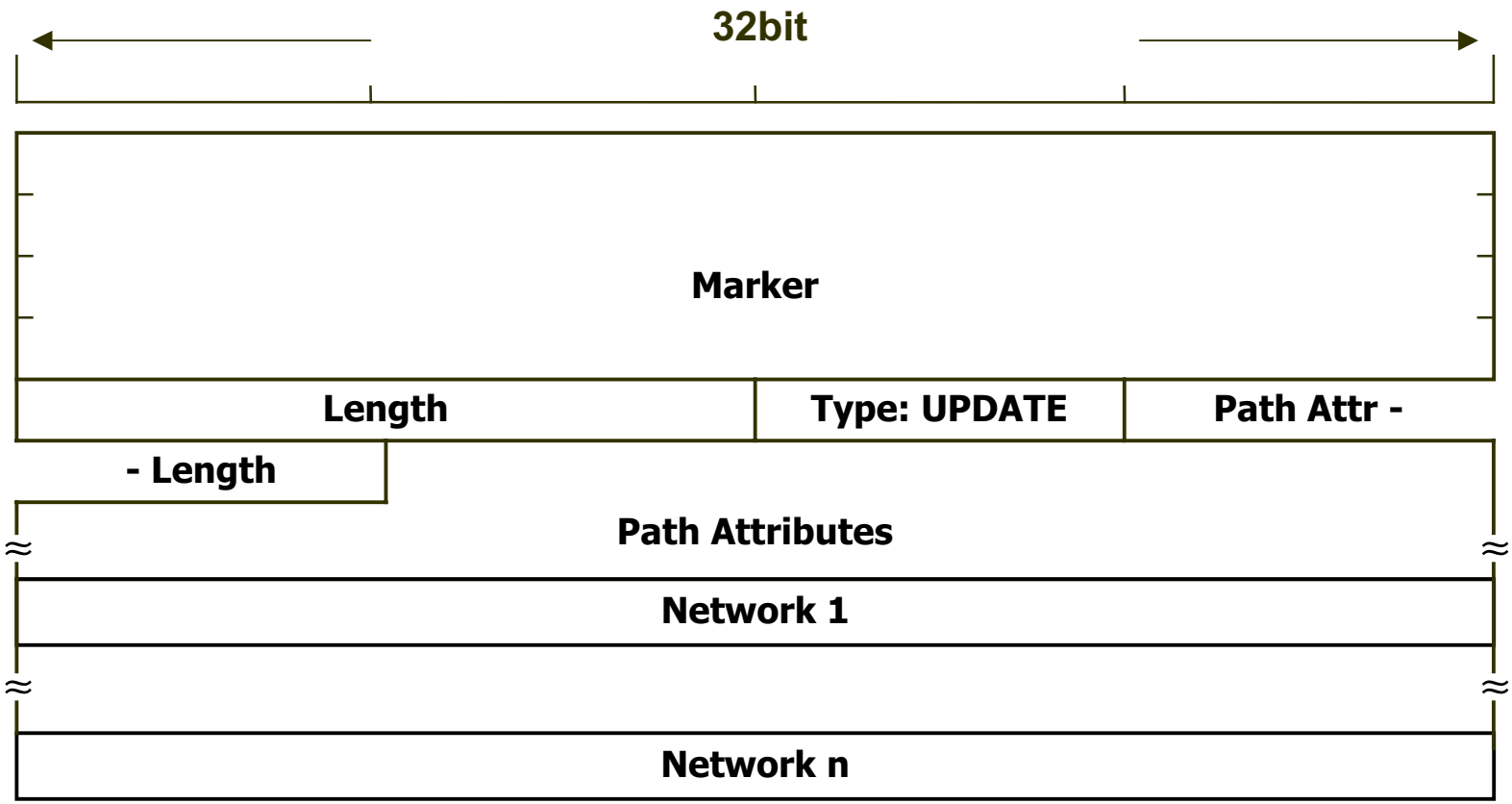


  - Marker : detect the loss of synchronization and authenticate incoming message
  - Length : total length of message
  - Type : type code of message

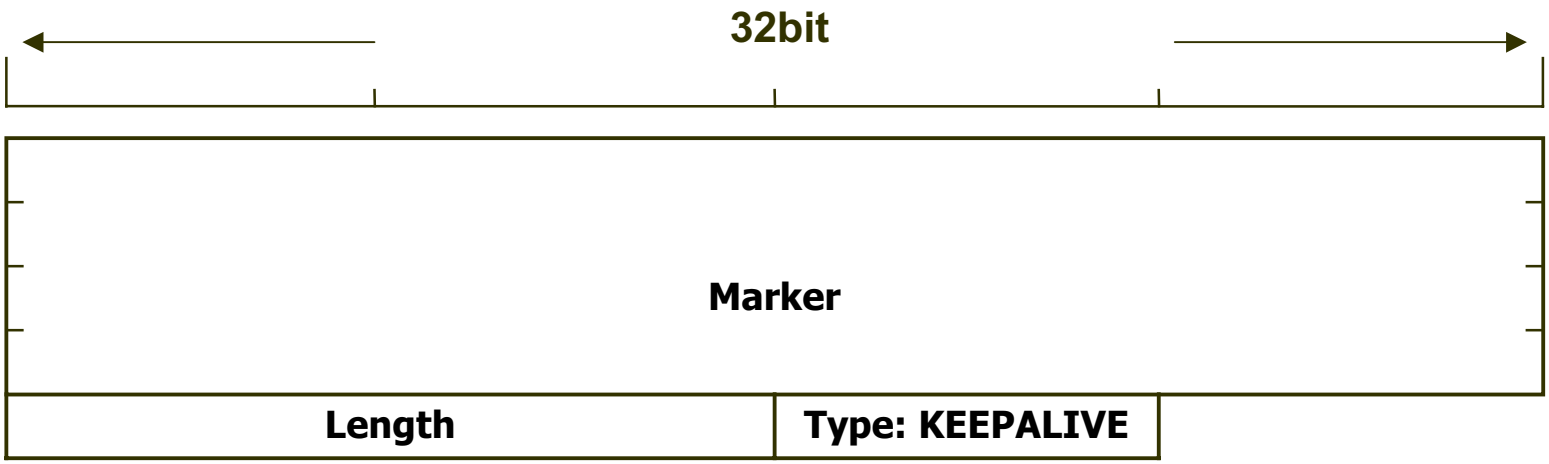
- **Open**
  - initial exchange



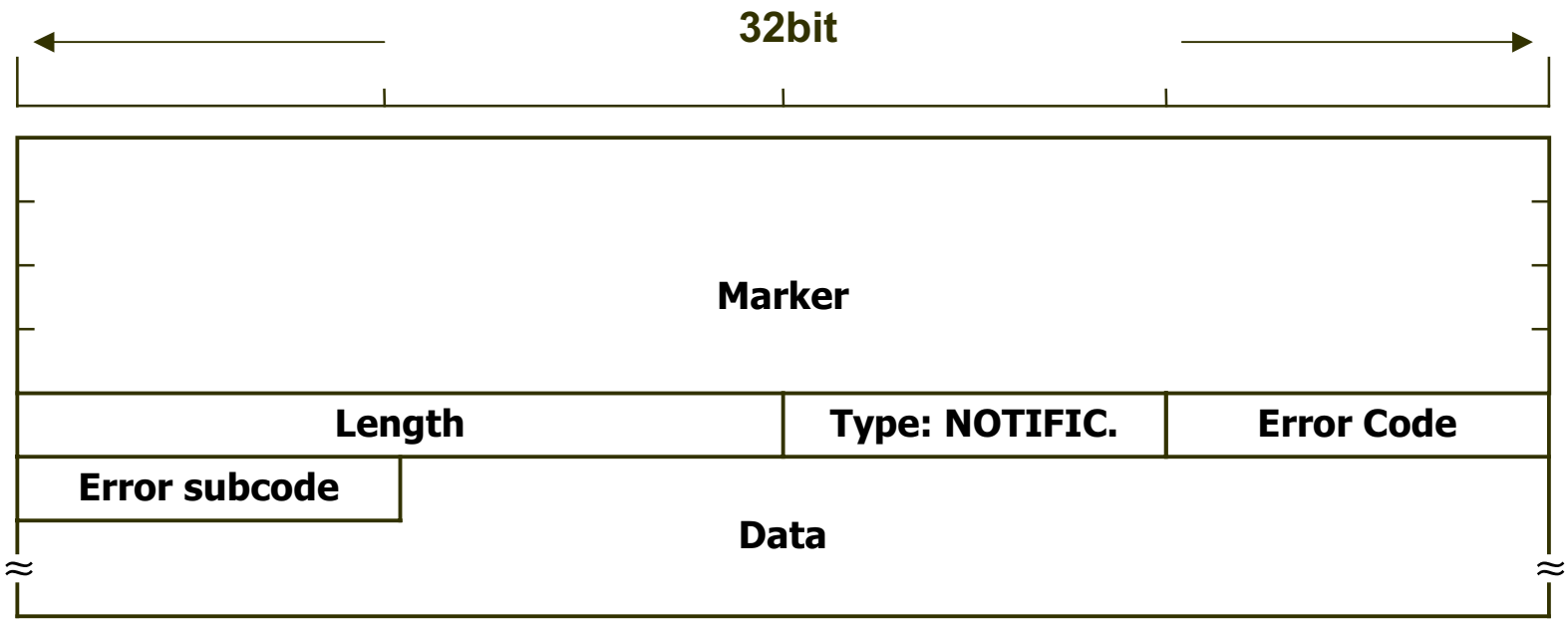
- **Update**
  - transfer routing information



- **Keep Alive**
  - avoid the exchange to be expired



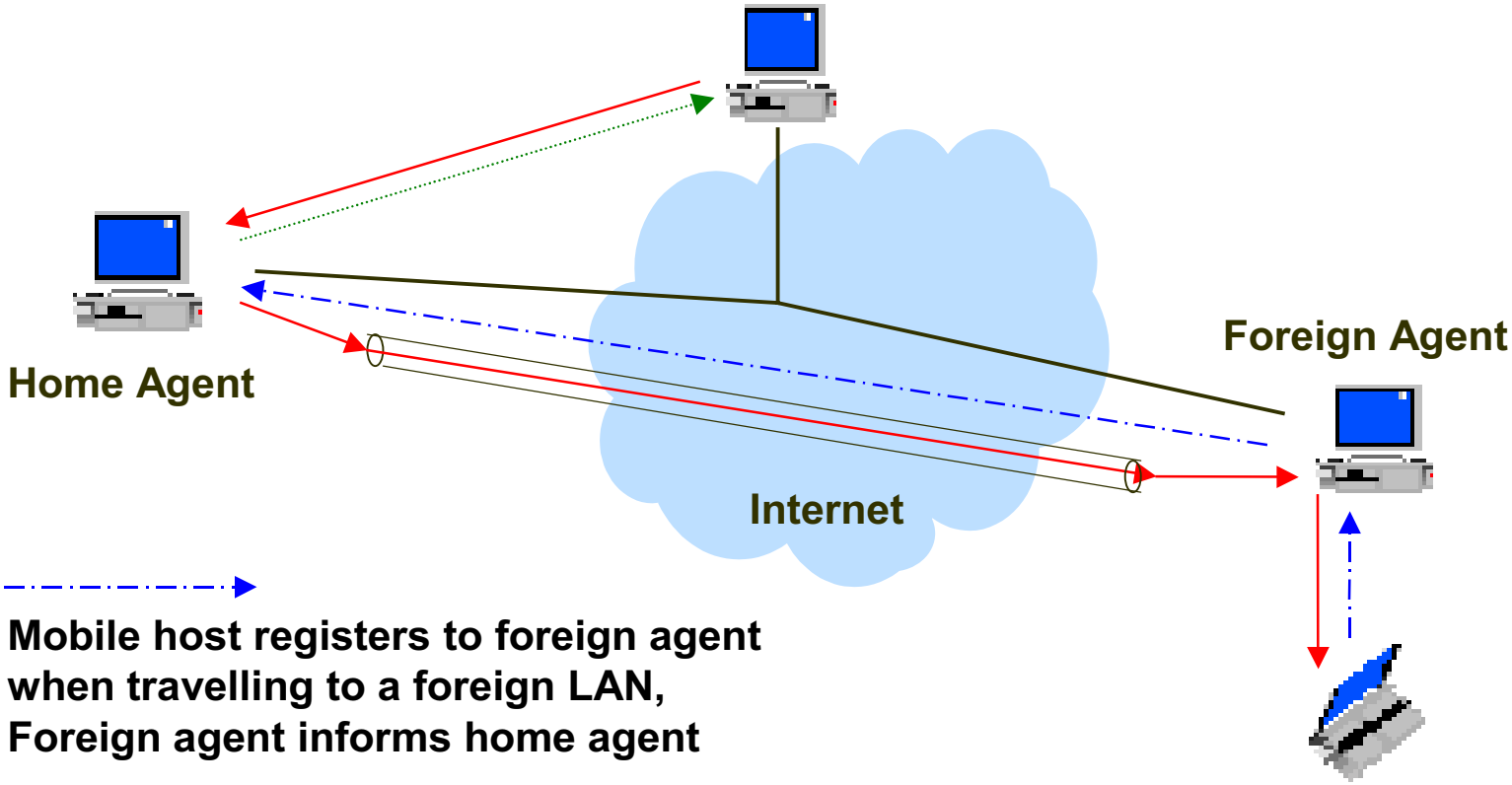
- **Error Notifications**
  - report the error



IS-IS (Intermediate System to Intermediate System Routing Protocol)

- support large routing domains consisting of combinations of many types of subnetworks
  - point-to-point links
  - multipoint links
  - X.25 subnetworks
  - broadcast subnetworks
- divided into areas administratively, organize Intra-domain routing hierarchically
  - Level 1 routing : routing within an area
  - Level 2 routing : routing between areas

Mobile IP



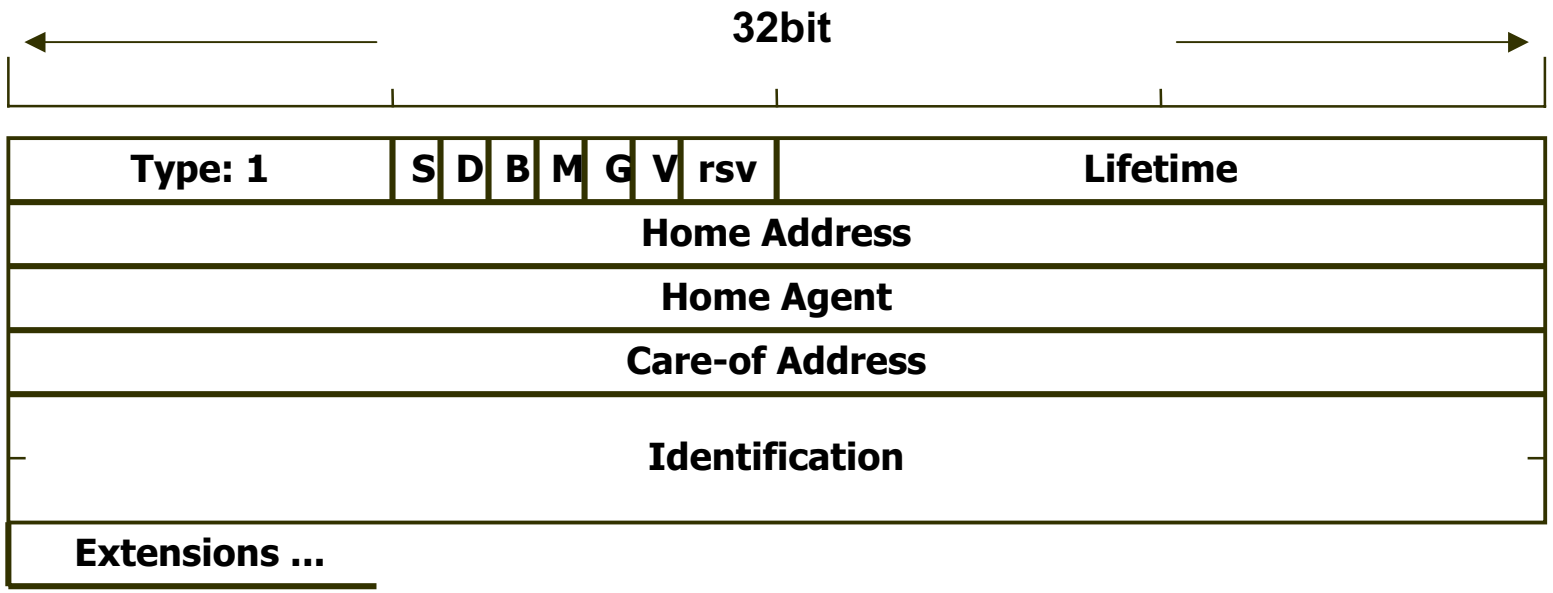
Mobile host registers to foreign agent
when travelling to a foreign LAN,
Foreign agent informs home agent

Packets sent to the mobile host are intercepted
by home agent and tunneled to foreign agent

Home agent informs the packet sender
foreign agent's address

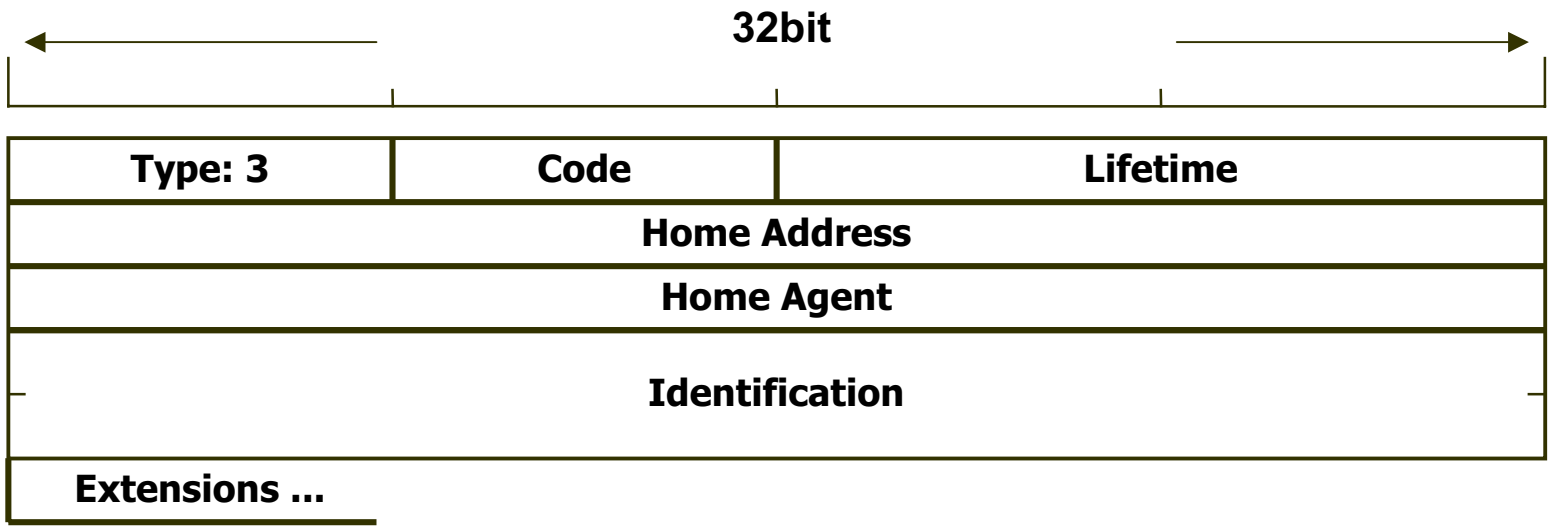
Registration Messages

- **Registration Request**



S Simultaneous bindings
B Broadcast datagrams
D Decapsulation by mobile node
M Minimal encapsulation
G GRE encapsulation
V Van Jacobson header compression
rsv reserved bit

Registration Reply



Code: result of registration

0 : registration accepted

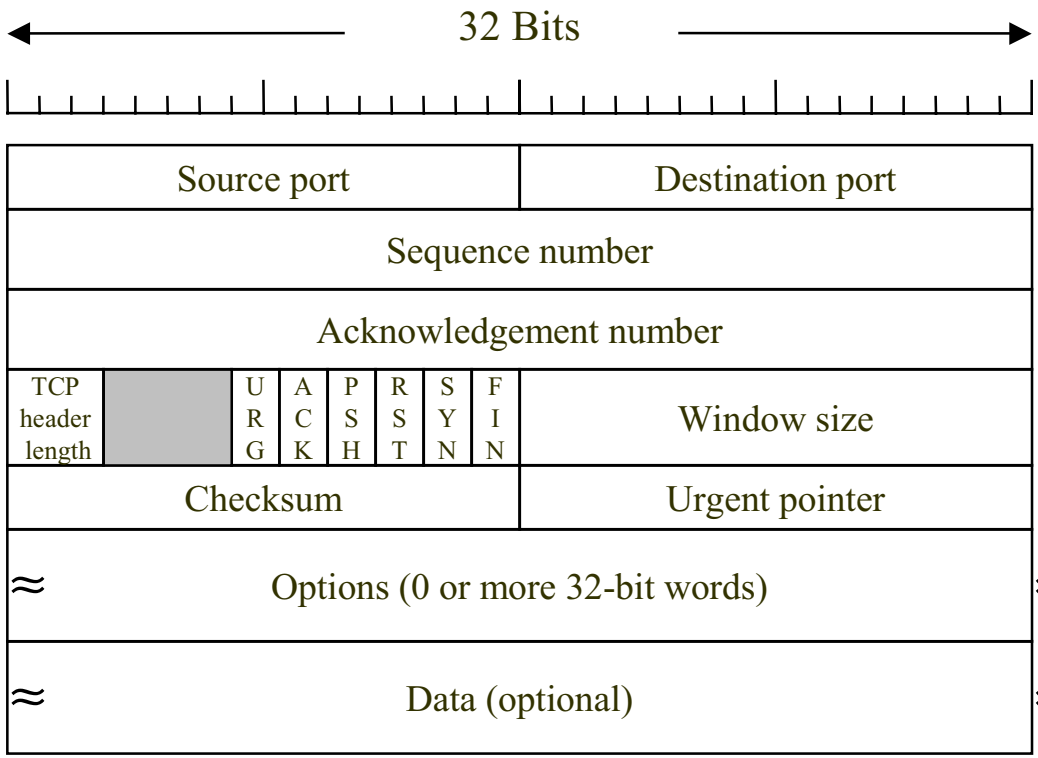
1 : registration accepted, but simultaneous mobility bindings unsupported

other: registration denied by foreign agent

TCP & UDP

- **Transmission Control Protocol(TCP):**
  - **A reliable end-to-end byte stream protocol over an unreliable internetwork**
- **User Datagram Protocol(UDP):**
  - **A protocol to send encapsulated raw IP datagrams without having to establish a connection**

TCP Header



URG: urgent data

Urgent pointer: byte offset

PSH: PUSHed data (not to buffer it)

RST: reset a connection

SYN, ACK: (1,0)--connection request

(1,1)--connection reply

FIN: release a connection

Window size: receiver window size

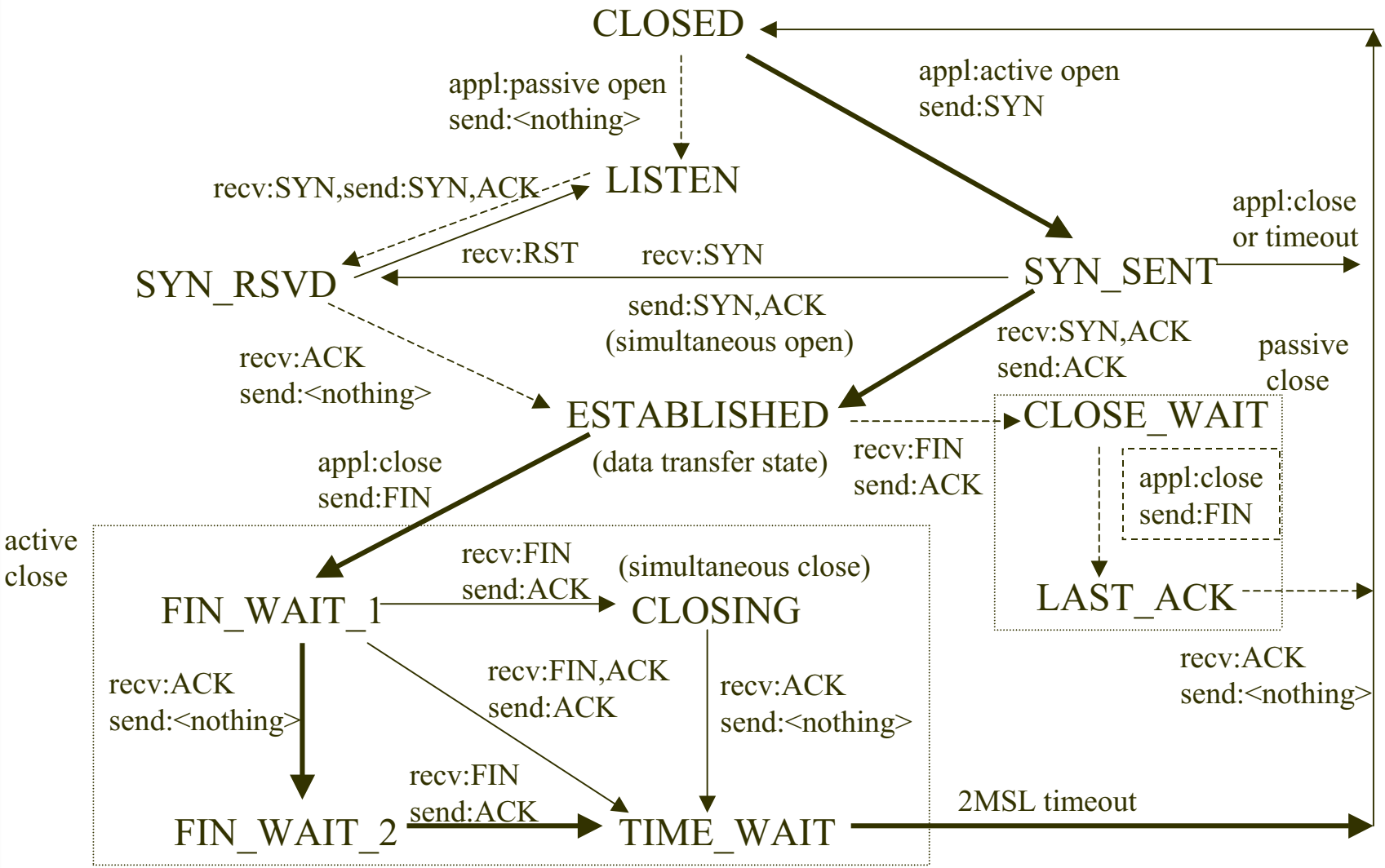
Checksum: on header, data, pseudo header

Options: max TCP payload (default 536 bytes)

window scale factor (up to $\times 2^{16}$)

NAK for selective repeat

TCP Connection Management



TCP Transmission Management

- **Example: a TELNET TCP connection to an interactive editor**
  1. Source→dest: 21-byte TCP data segment (41-byte IP datagram)
  2. Source←dest: acknowledgement segment (40-byte)
  3. Source←dest: window update segment (40-byte)(after editor reads the byte)
  4. Source←dest: echo segment (41-byte) (after editor processes the byted)
  5. Repeat step 1, 162 bytes in 4 segments for each character types!!
- **Silly window syndrome: frequent but mall window updates**
  1. sending application to TCP one byte at a time
  2. receiving application sucks the data up from TCP one byte at a time
- **Nagle's algorithm to solve 1:**
  - When data come into the sender one byte at a time, just send the first byte and buffer all the rest until the outstanding byte is acked.
- **Clark's algorithm to solve 2:**
  - The receiver should not send a window update until it can handle the max segment size it advitised when connection was established, or its buffer is half empty, whichever is smaller.