Advanced Lecture on Internet Infrastructure

4. IPv4, ICMP, ARP

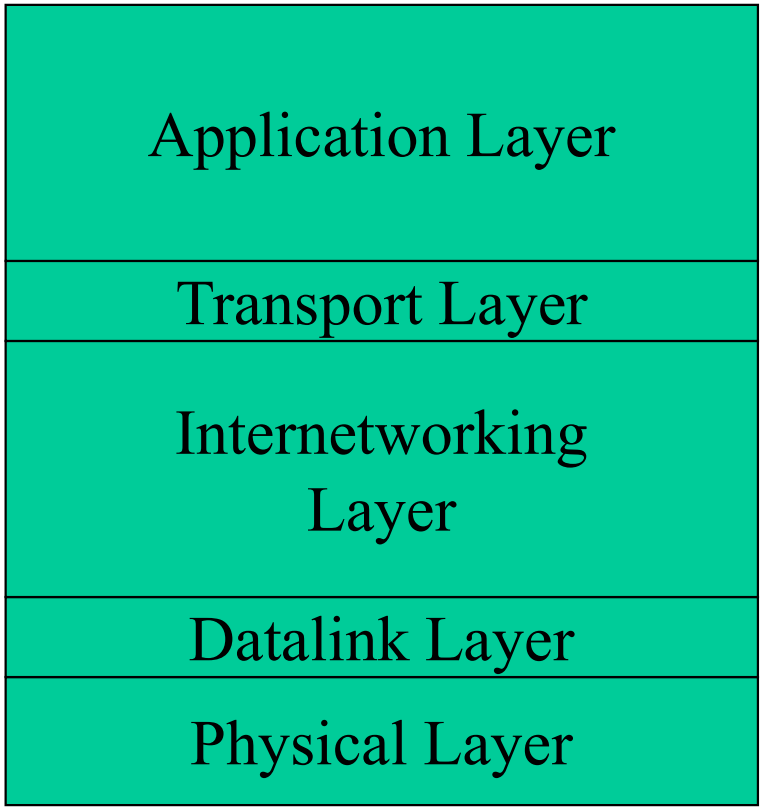
Masataka Ohta

mohta@necom830.hpcl.titech.ac.jp

ftp://chacha.hpcl.titech.ac.jp/infra4e.ppt

Layering of the Internet

- Physical and Application Layers are Essential
- The Internetworking Layer does as Much Things as Possible
- Datalink and Transport Layers should Avoid to do Thing

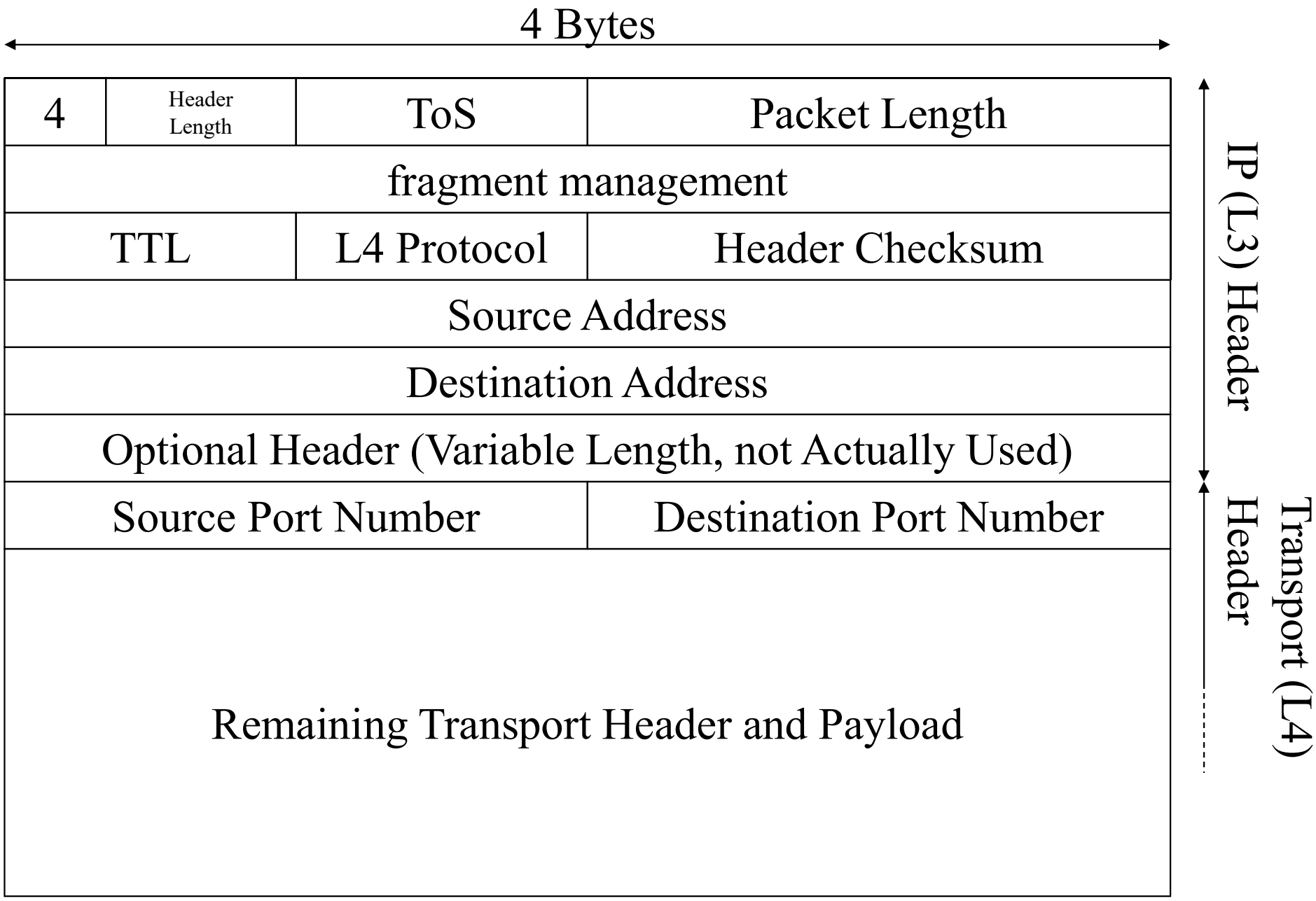


Here is the Essence of the Internet

Layering Structure of the Internet

IPv4 (Internet Protocol Version 4, rfc791)

- do almost nothing in the network
  - deliver packets to their destinations
  - fragmentation
  - maintain TTL (time to live)



Format of IPv4 Packets (rfc791)

Header Length

- 4 bit field
- header length measured by 4B
  - minimum 5 (20B), maximum 15 (60B), always 5 in practice

ToS (Type of Service)

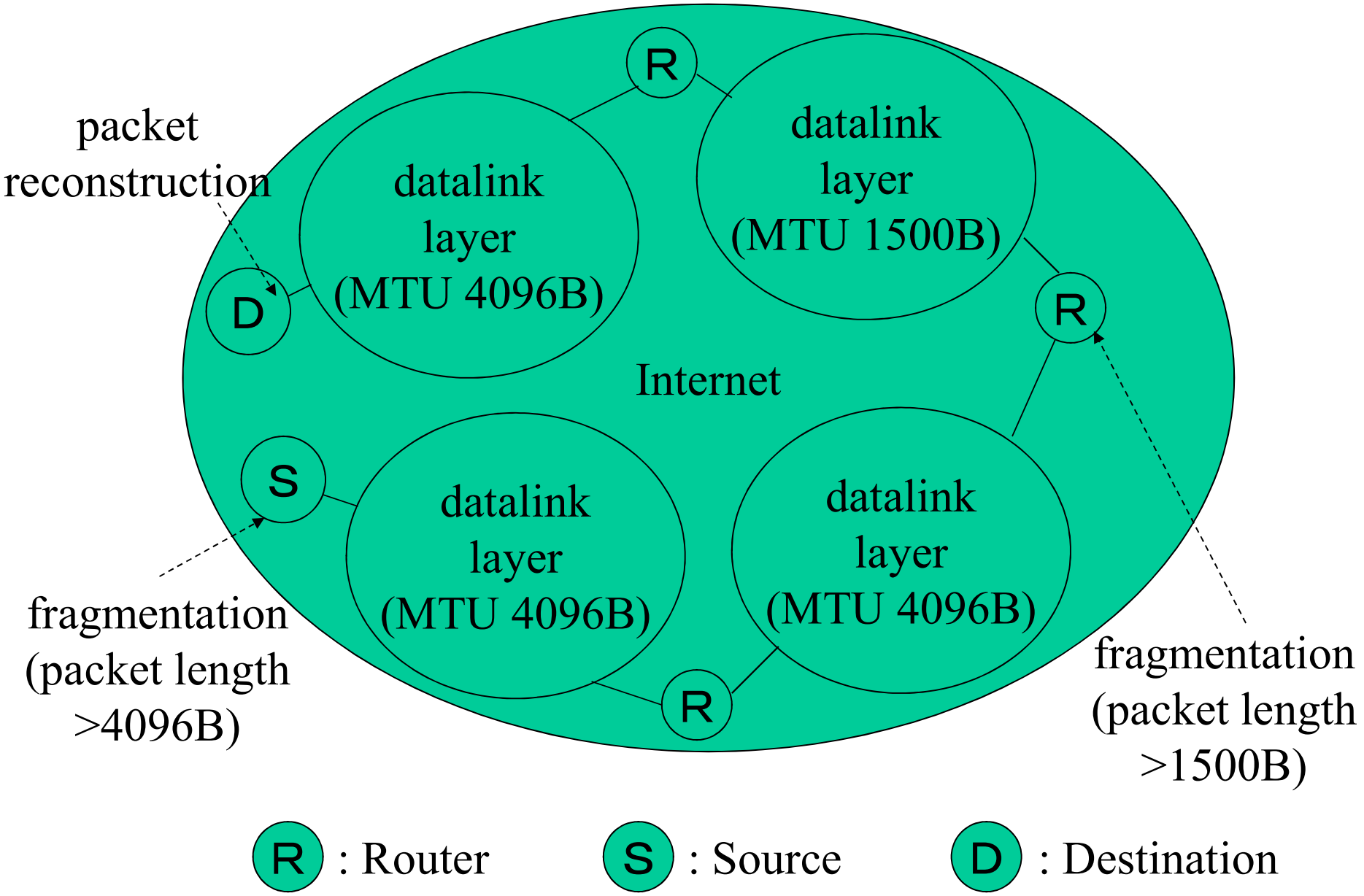
- 8 bit field
- packet priority, etc.
- not really used
- may be used for DiffServe (Differentiated Service) or ECN (Explicite Congestion Notification)

Packet Length

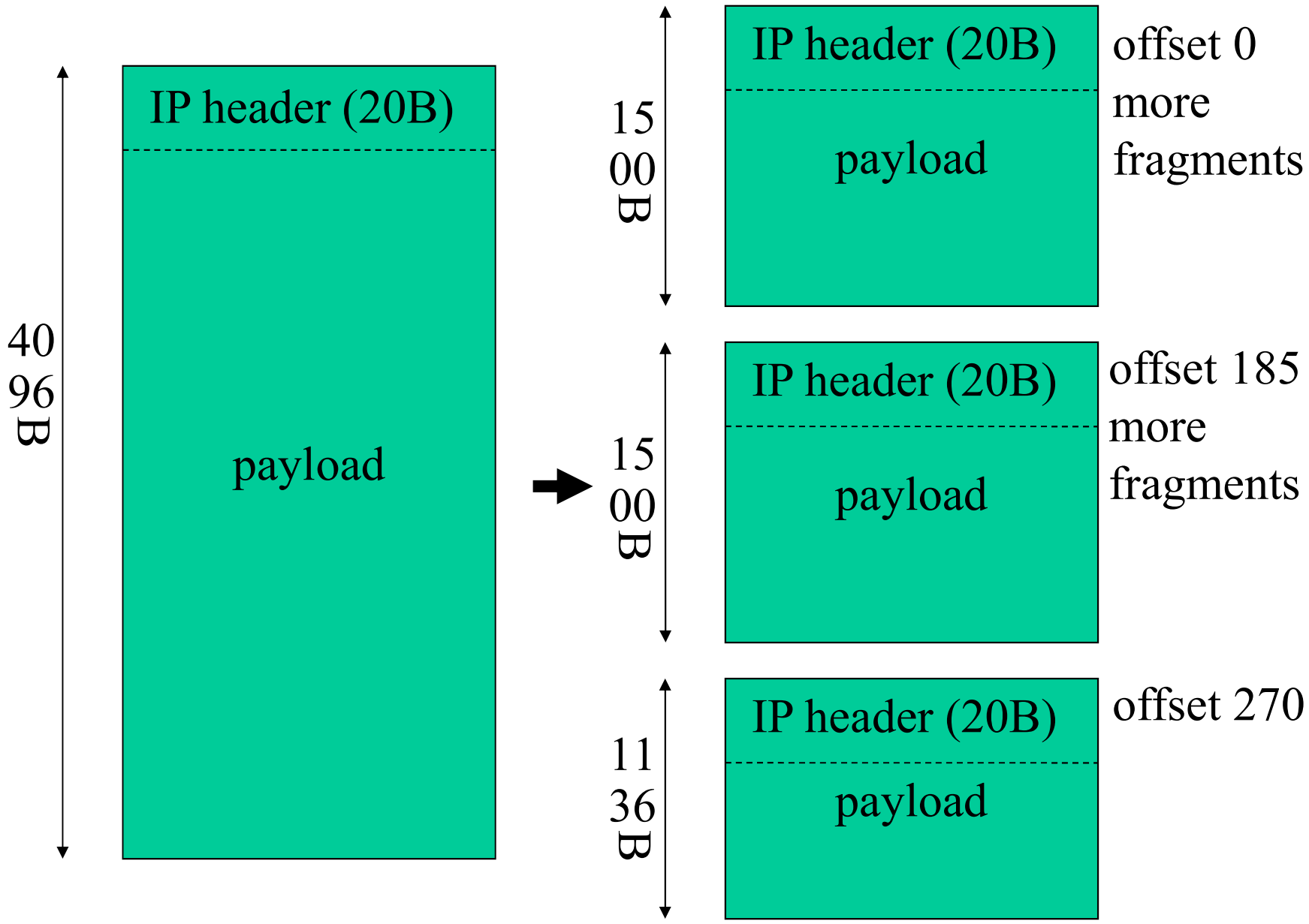
- 16 bit field
- packet length measured by byte
- minimum 20, maximum 65535
- packet length received by any host is 576B (516B (556B) excluding header)
  - beyond that, prior negotiation is necessary, in theory

Fragment Management

- packets longer than MTU is divided en route
  - minimum MTU (Maximum Transmmision Unit) is 68B
- 16 bits of ID field
  - identify original packet of fragmented packets
- 3 bits of flag field
  - zero, don’t fragment, more fragments
- 13 bits offset field
  - fragment location (8B unit) in original packet



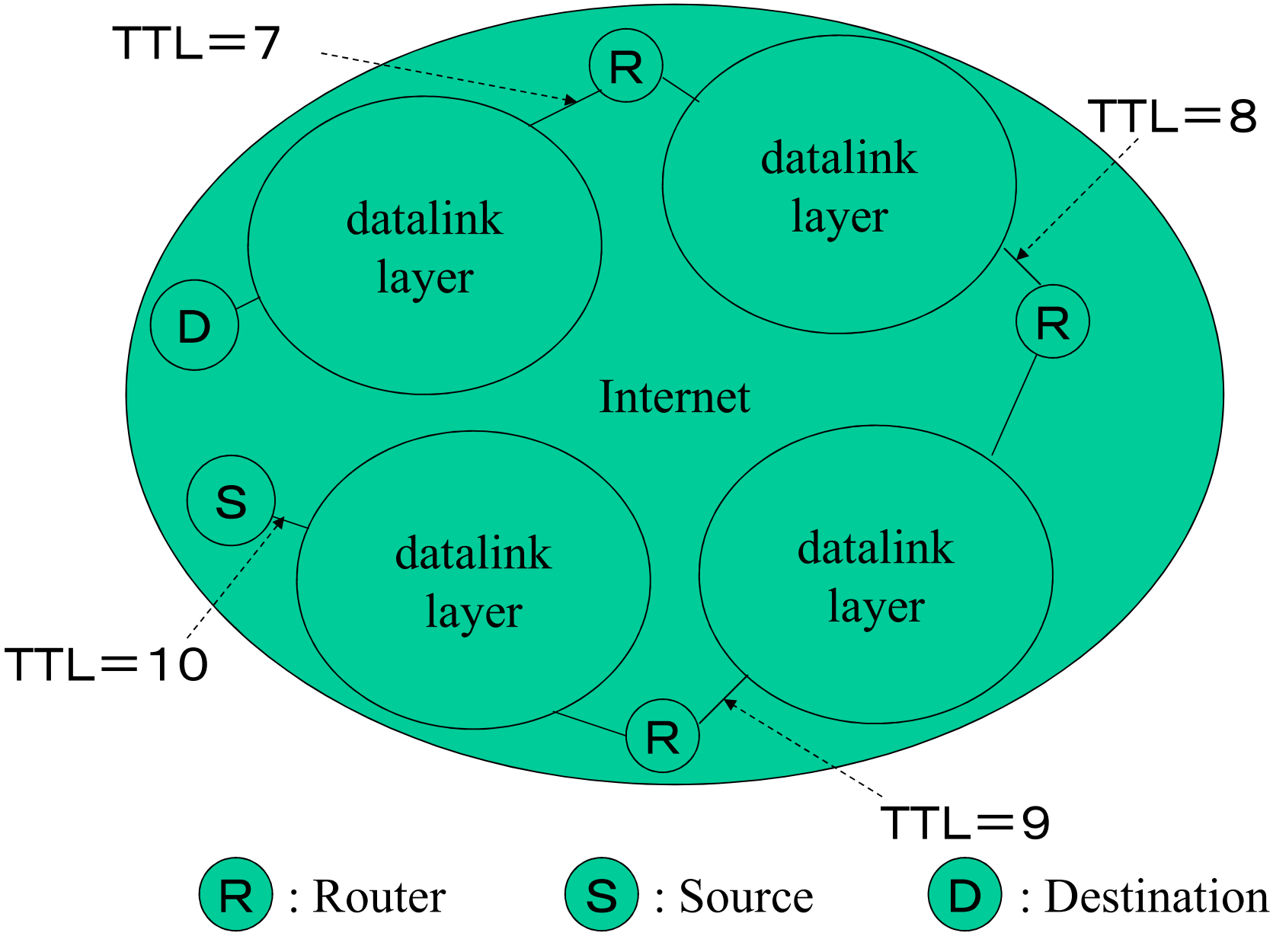
How Fragmentations Occur



not a very good example of fragmentation

TTL (Time to Live)

- 8 bit field
- manage lifetime of packets
  - reduced at least 1 on every router
  - reduced N if N seconds passed in a router
    - only in theory
  - packets with zero TTL is discarded
  - prevent infinite loop
    - also useful for traceroute (tracert)



Decrementing TTL

(L4) Protocol

- 8 bit field
- distinguish L4 (?) protocol
  - 1 for ICMP (Internet Control Message Protocol, may not be L4)
  - 6 for TCP (Transmission Control Protocol)
  - 17 for UDP (User Datagram Protocol)

Header Chechsum

- 16 bit field
- 16 bit wise sum (1's complement, zero is represented with all bits 1) of remaining part of header

Source Address

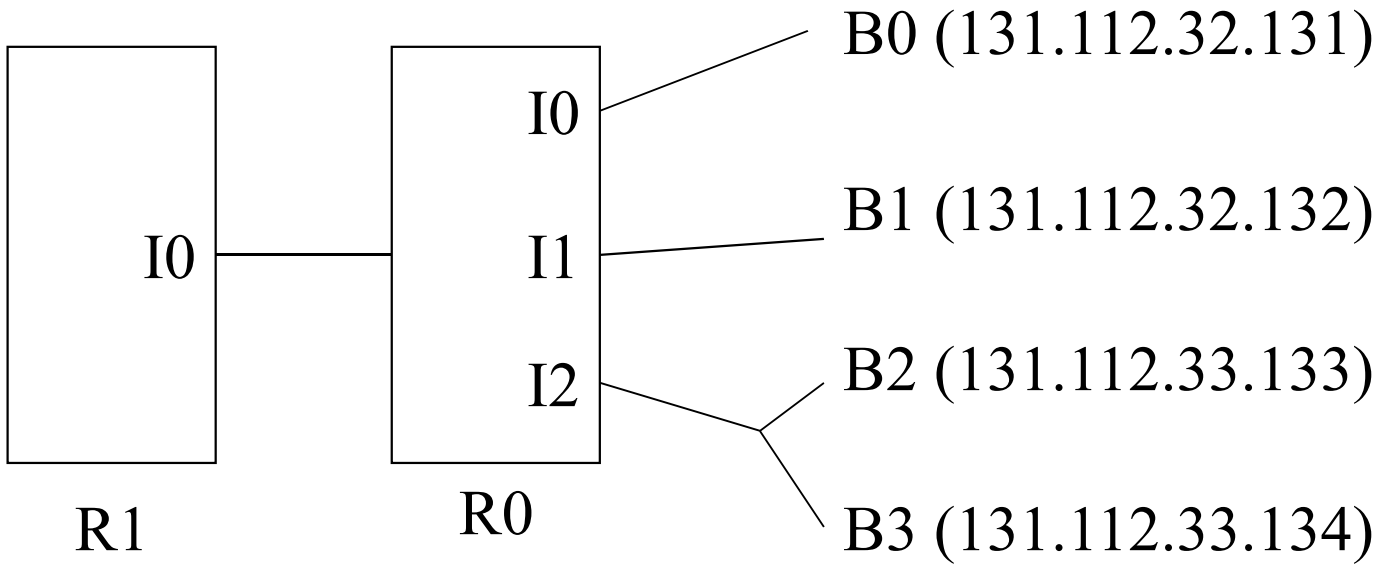
- IP address of packet source
- used for reply
  - though not reliable
- combined with source port number, identify peer at transport layer

Destination Address

- IP address of packet destination
- routers determine next hop router by looking up routing table using destination address as the key
  - variable length mask is used for the look up
- hosts (incl. routers) receive packets, if destination addresses of the packets are that of the hosts

Routing Table

- routers send packets to next hop routers based on look up results of routing table
  - key of the look up is destination address
- same entry may be shared if similar(?) addresses occur only in some remote region
  - route aggregation
    - 1 entry shared by many addresses
  - like phone numbers, may be hierarchical
    - +81-3-5734-3299



routing table at R0

destination	next hop
131.112.32.131	I0
131.112.32.132	I1
131.112.33.*	I2

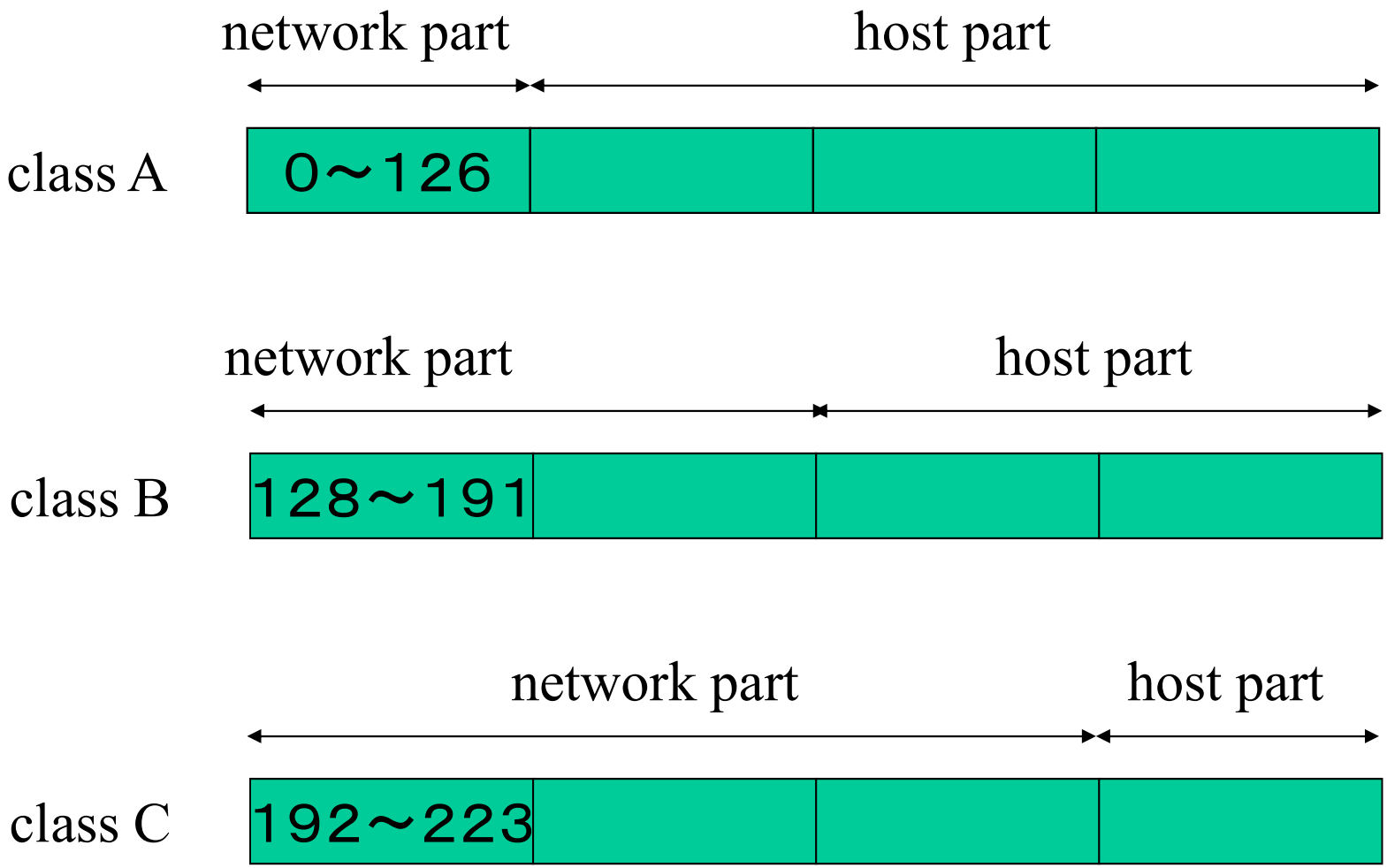
routing table at R1

destination	next hop
131.112.*	I0

route aggregation

Class based Routing

- IPv4 addresses are divided into 5 classes
  - Class A, B and C for unicast
    - class D for multicast, E reserved
- unicast IP address is divided to network part and host part
  - routing is by network part (no hierarchy)
  - host part of all 1 means broadcast within the network
  - host part of all 0 is address of the network



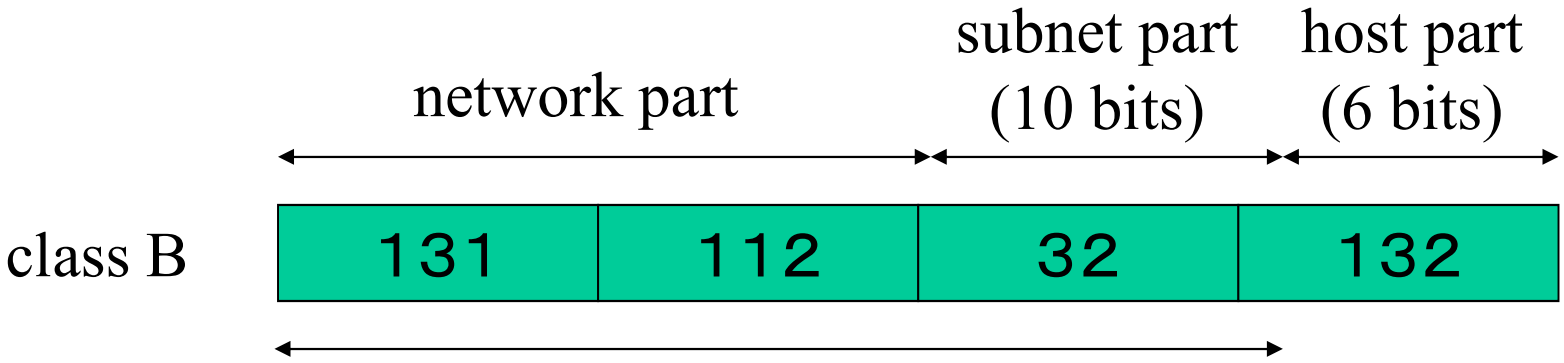
address structure of IPv4 unicast classes

Problems of Class based Routing

- each link has, at most, several tens of hosts
  - though some operated with thousands of hosts
    - only to find it inoperational
  - even class C is too large
    - unnecessary increase of route information
    - unnecessary consumption of IPv4 addresses
- finer subdivision of IPv4 address necessary
  - subnet

Subnet

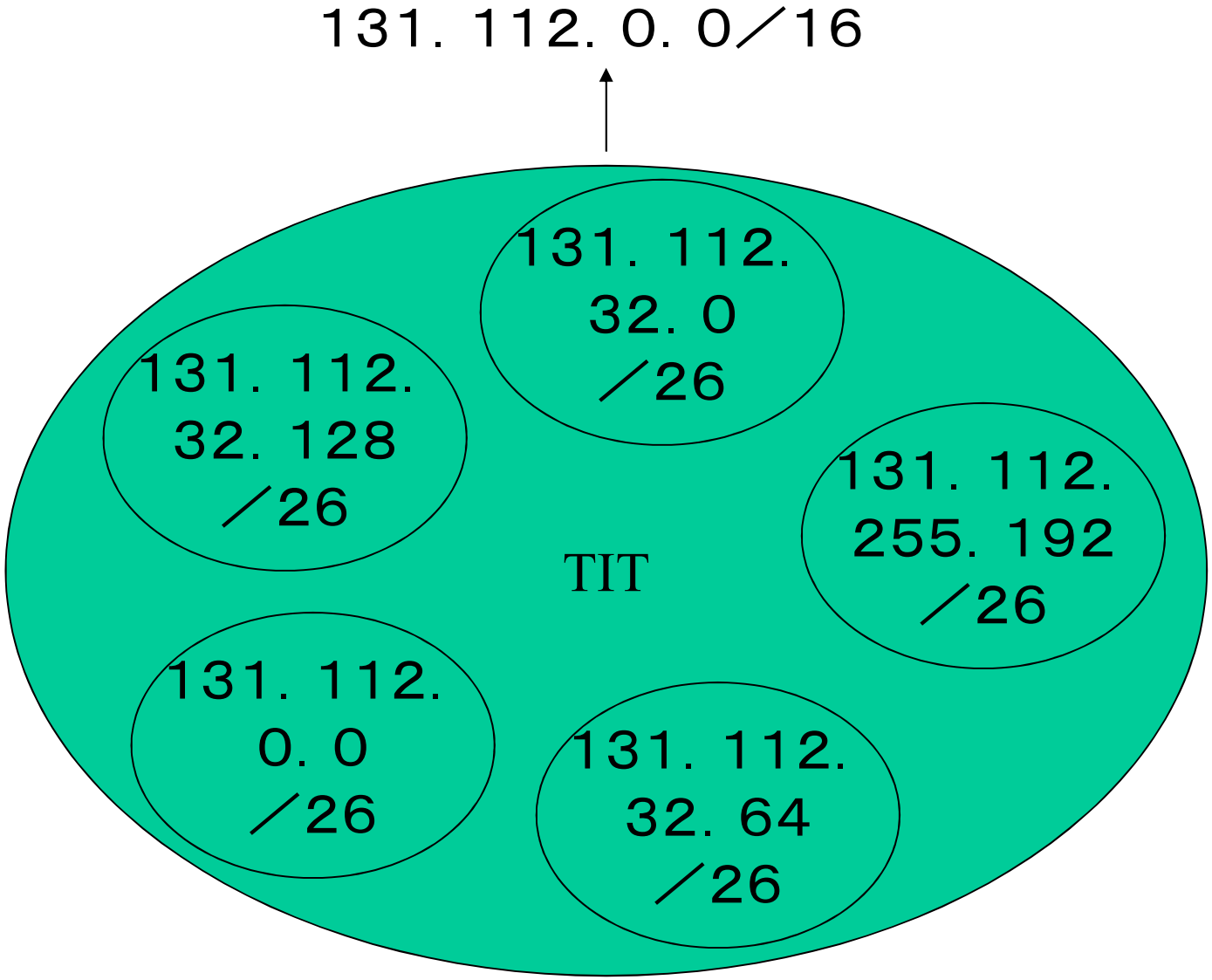
- divide host part into subnet part and host part
- subnet-wise routing within a network
  - 1 class B address is mostly enough for each organization
- network-wise routing outside of the network
  - only 1 route information for each organization externally



(26 bits for TIT, same within a network)

131.112.32.128/26 identify a subnet

example of structure of subnetted IP address of TIT


131. 112. 0. 0／16
131. 112.
32. 0
／26
131. 112.
32. 128
／26
131. 112.
255. 192
／26
TIT
131. 112.
0. 0
／26
131. 112.
32. 64
／26

CIDR (Classless Inter-Domain Routing) (rfc1519)

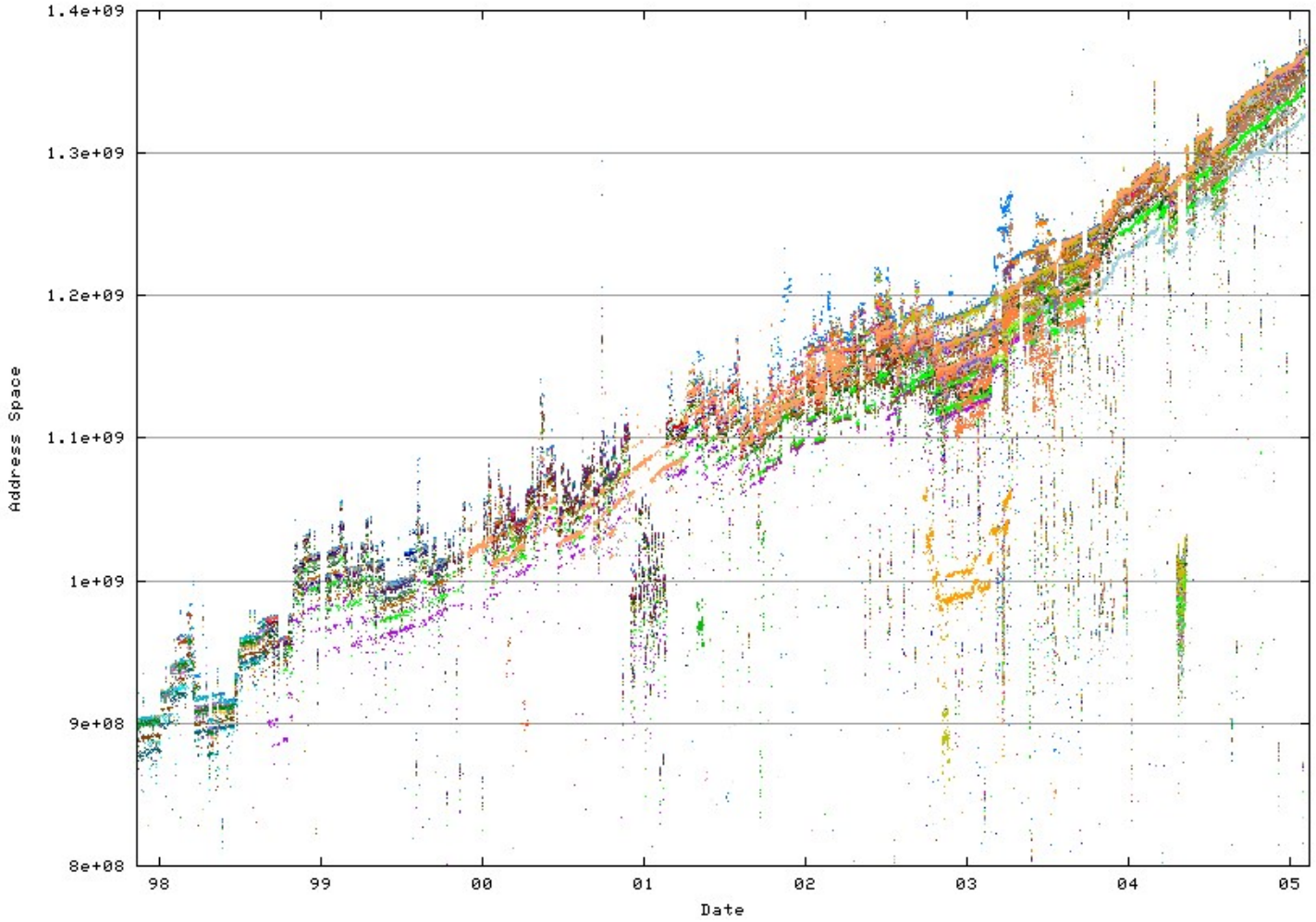
- classes are totally abandoned
  - routing protocols carry netmask for each routing table entry
    - must upgrade routing protocols
- example of hierarchical address allocation
  - ISP is allocated a block of 256 addresses
    - routing table entry with netmask /24 outside of the ISP
  - the ISP allocate 8 addresses to each customer
    - 32 routing table entries with netmask /29 in the ISP for the block

Problems of IPv4

- address space is too small
  - with 32 bits of address, only 4G terminals
- increased routing table size
  - was 50k when IPv6 published (1998)
  - >800k even with aggregation by CIDR (2019)

IPv4 Address Span

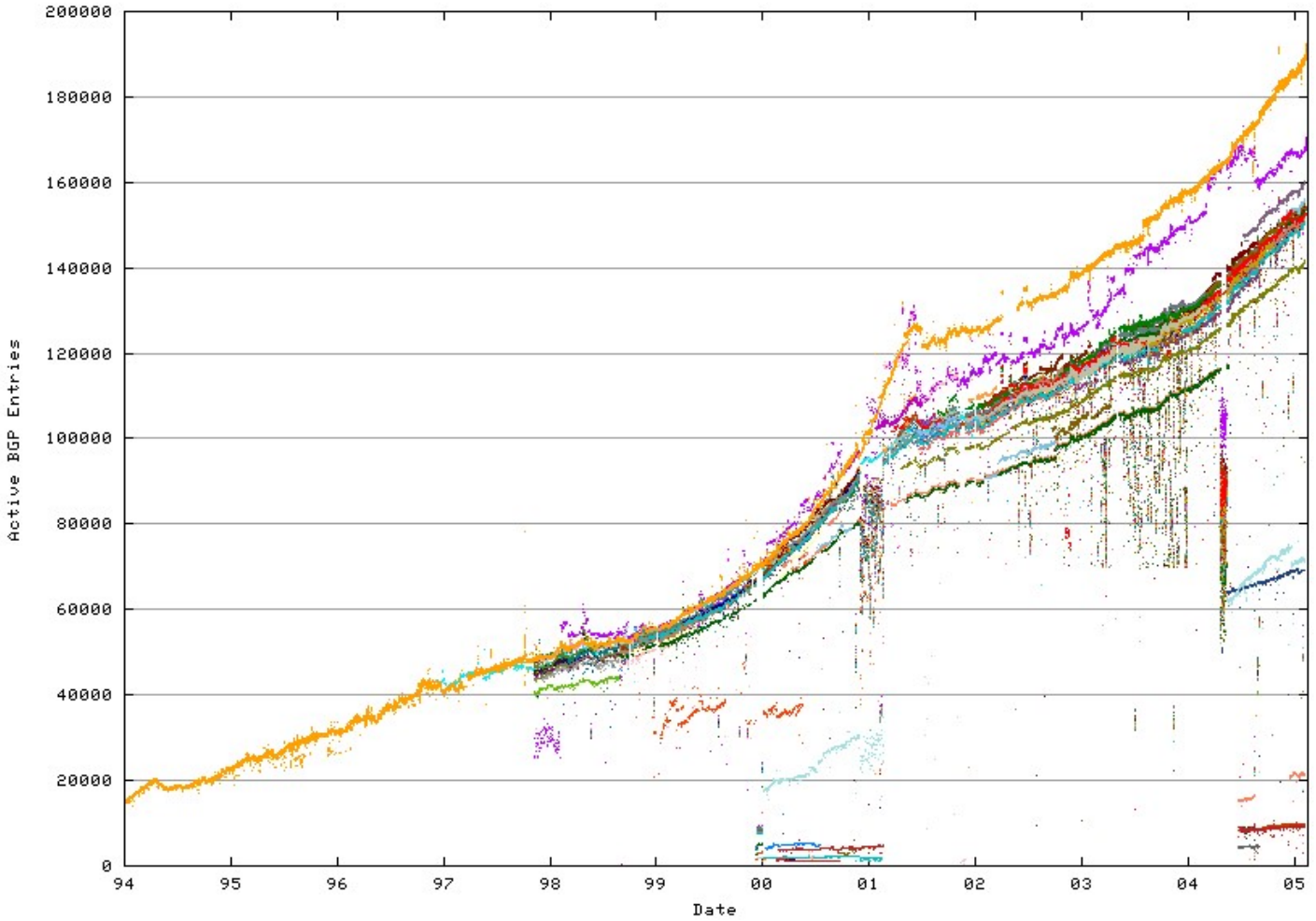
http://www.apnic.net/meetings/19/docs/sigs/routing/routing-pres-info-huston-routing-table.ppt



This figure shows the total amount of address space spanned by the routing table. This is a view derived from the Route-Views archive, where each AS has a single colour. The snapshots are at two-hourly intervals, and span from early 2000 until the present. The strong banding in the figure is spaced 16.7M units apart, or the size of a /8 advertisement There appear to be 3 /8 advertisements that are dynamic. Not every AS sees the same address range, and this is long term systemic, rather than temporary. This is probably due to routing policy interaction, coupled with some cases of prefix length filtering of routing information. The rate of growth declined sharply across 2002 and the first half of 2003, resuming its 2000 growth levels in 2004.

IPv4 Routing Table Size

http://www.apnic.net/meetings/19/docs/sigs/routing/routing-pres-info-huston-routing-table.ppt



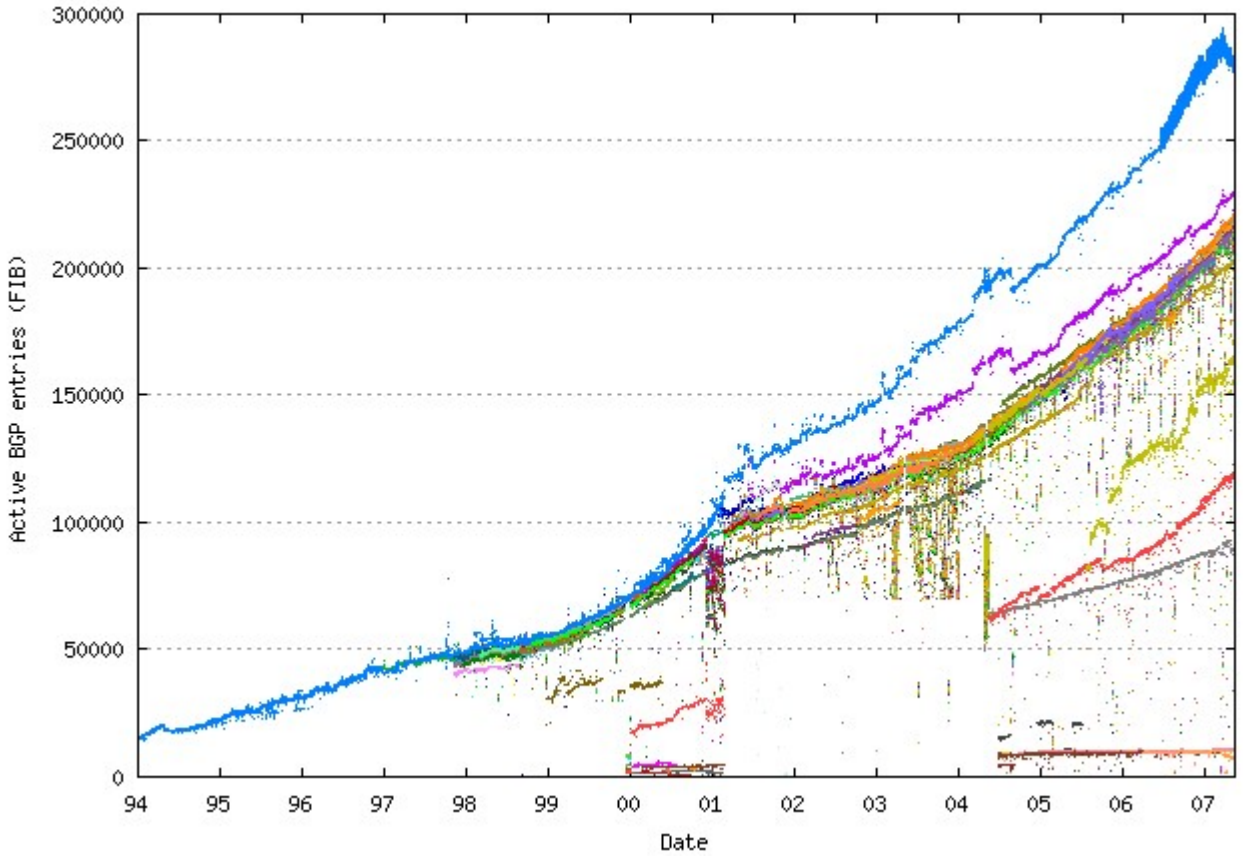
Data assembled from a variety of sources, Including Surfnet, Telstra, KPN and Route Views. Each colour represents a time series for a single AS.

The major point here is that there is no single view of routing. Each AS view is based on local conditions, which include some local information and also local filtering policies about external views.

IPv4 Routing Table Size

http://bgp.potaroo.net/

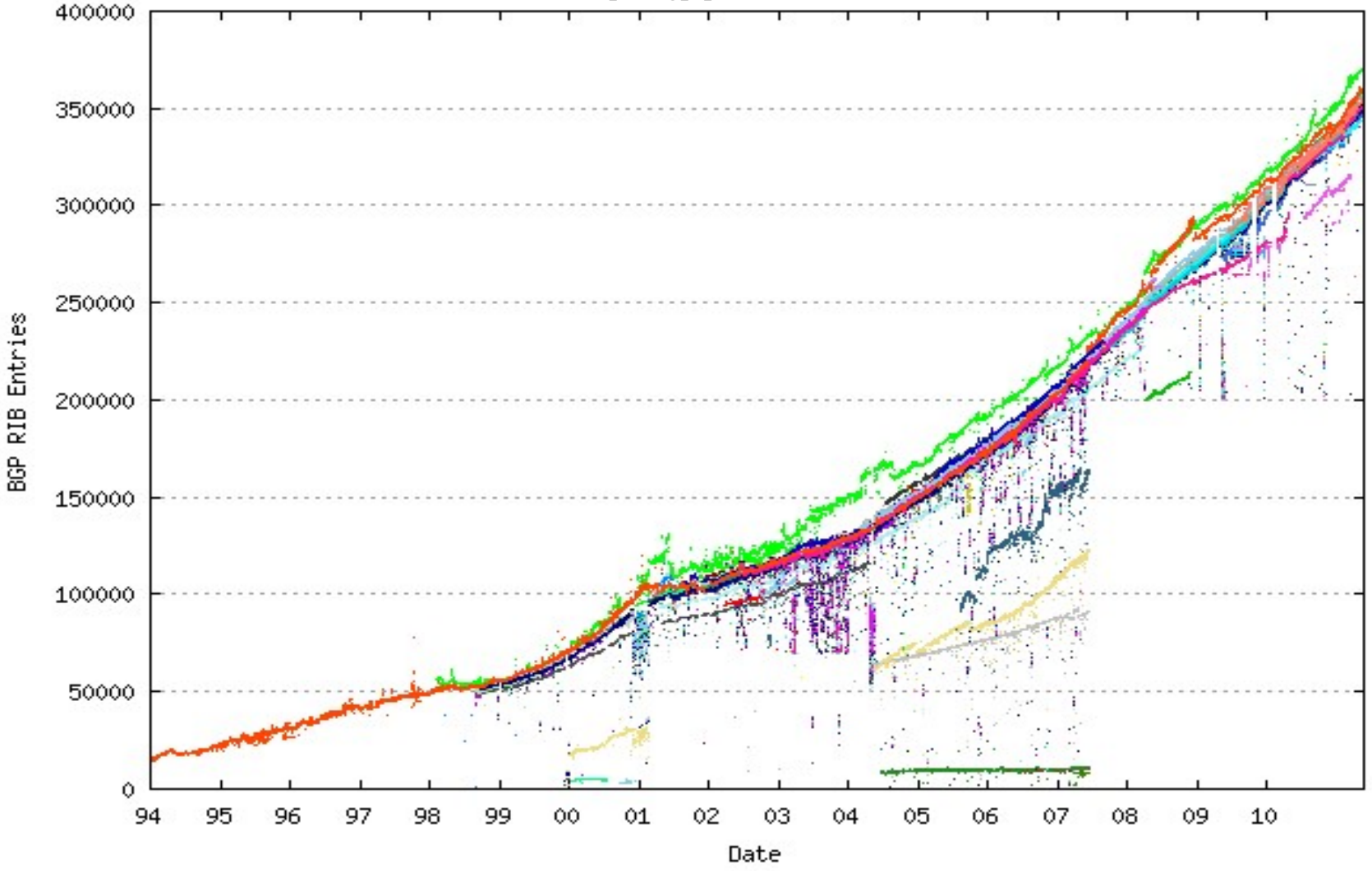
Growth of the BGP Table - 1994 to Present



(Data Gathered from AS1221 and Route-Views)

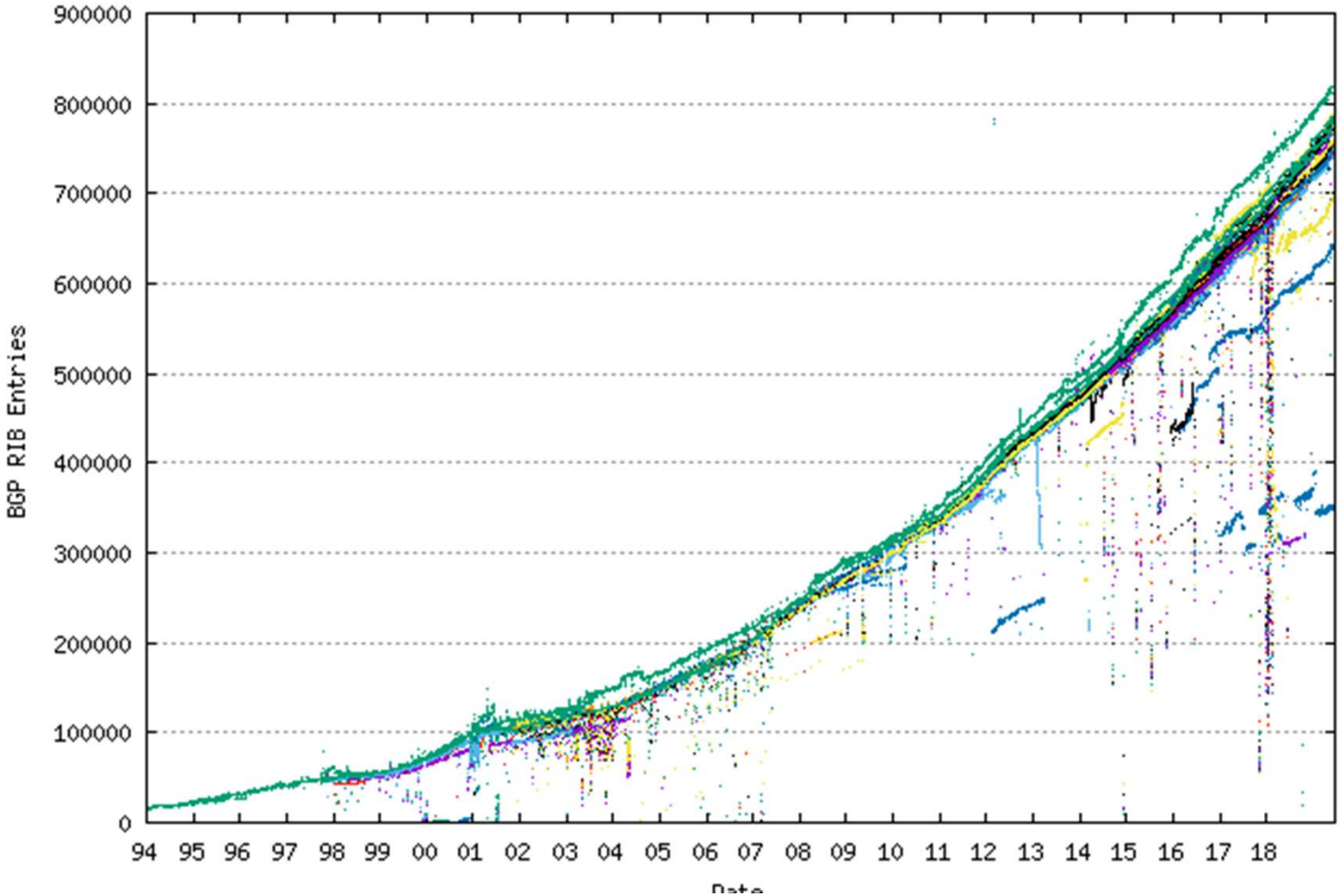
IPv4 Routing Table Size

http://bgp.potaroo.net/



IPv4 Routing Table Size

http://bgp.potaroo.net/



Cases When Route Aggregation Impossible

- aggregation possible, if route is shared by addresses sharing a pattern
- route not by destination address only
  - QoS routing depends on required QoS
- destination address not designate location
  - multicast address designate set of locations
- random IP addresses within a region
  - initial allocations for IPv4
  - multihoming by routing

ICMP (Internet Control Message Protocol, rfc792)

- protocol to control the Internet or report errors in the Internet (incl. destination)
- mixture of IP and transport layer functions
- format of ICMP for packet errors
  - (64B ICMP header)+(IP header and 64 bits after the header of packets casing the error)

Types of ICMP Message

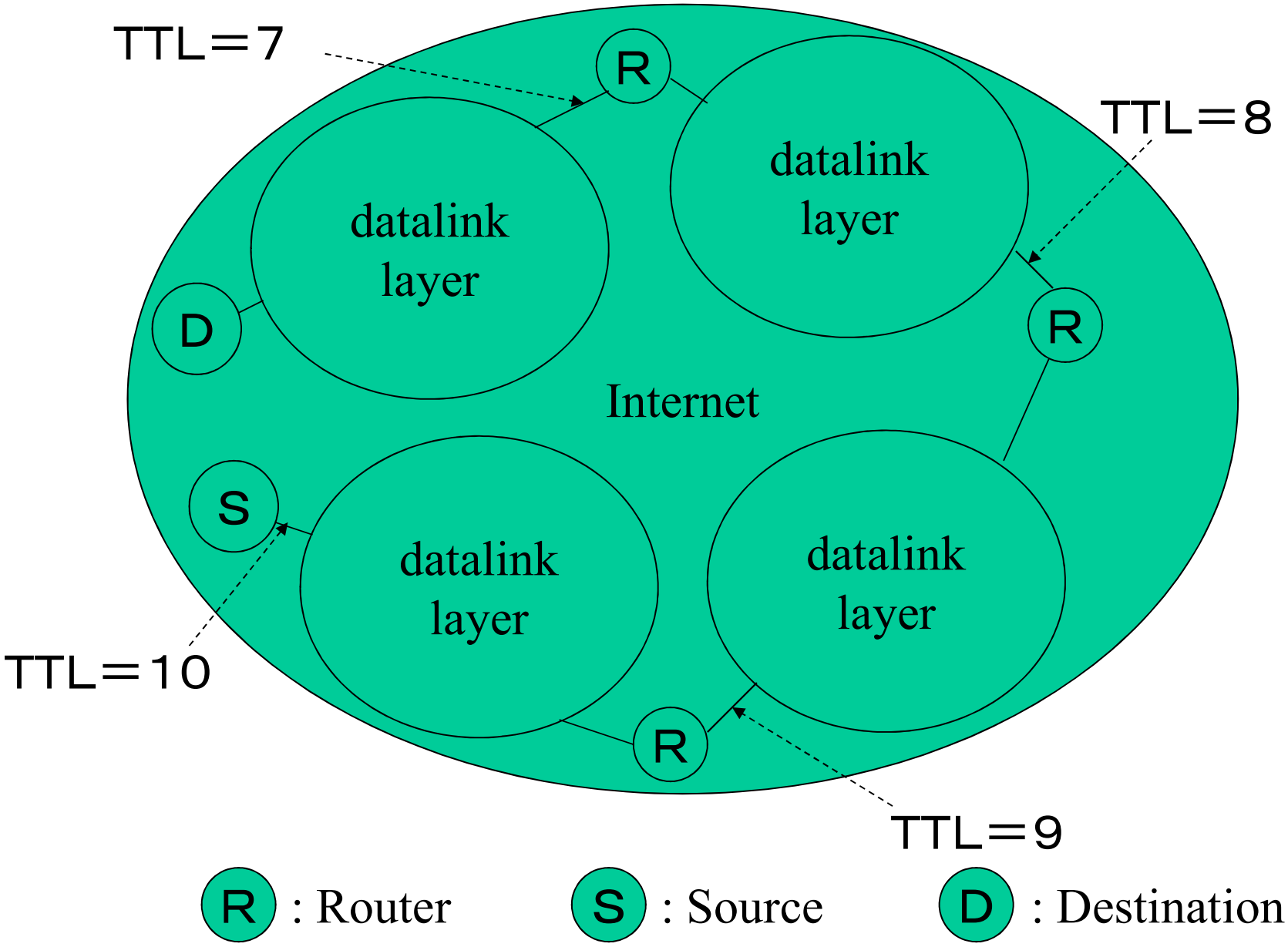
- Errors
  - Destination Unreachable
  - Time Exceeded
  - Parameter Problem
- Control
  - Source Quench
  - Redirect
  - Echo & Echo Reply
  - Time Stamp & Time Stamp Reply
  - Information Request & Reply

ICMP Destination Unreachable

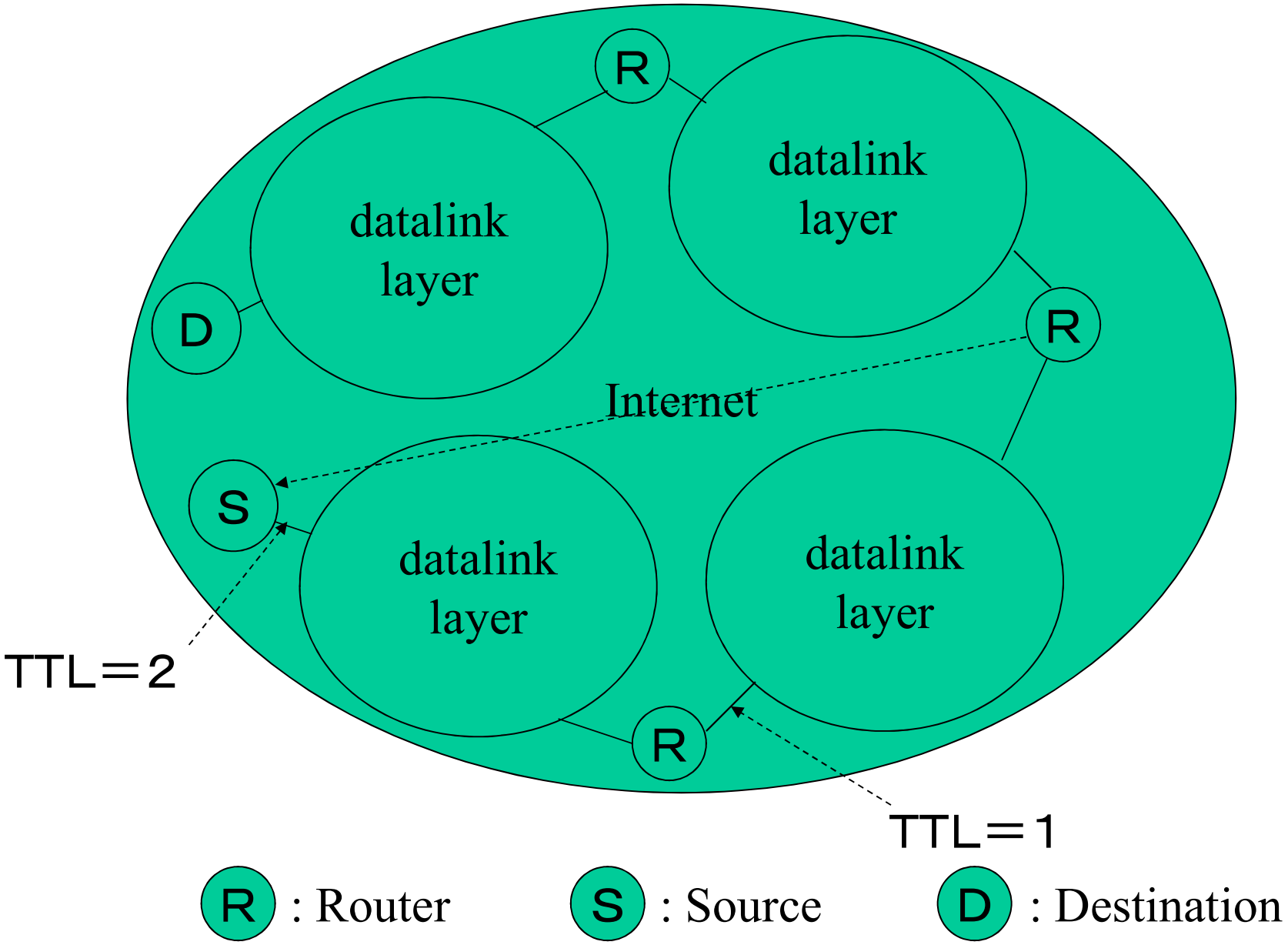
- unreachability of various kind
  - no route to destination network
  - host not reachable (no response to ARP)
  - no protocol supported at destination host
  - destination port is not served
  - fragmentation is necessary but “don’t fragment” bit is set
  - source routing failed

ICMP
Time Exceeded

- TTL becomes 0 en route
  - if TTL is gradually increased and source address of ICMP is checked, route to desination can be detected (traceroute/tracert)



Decrementing TTL



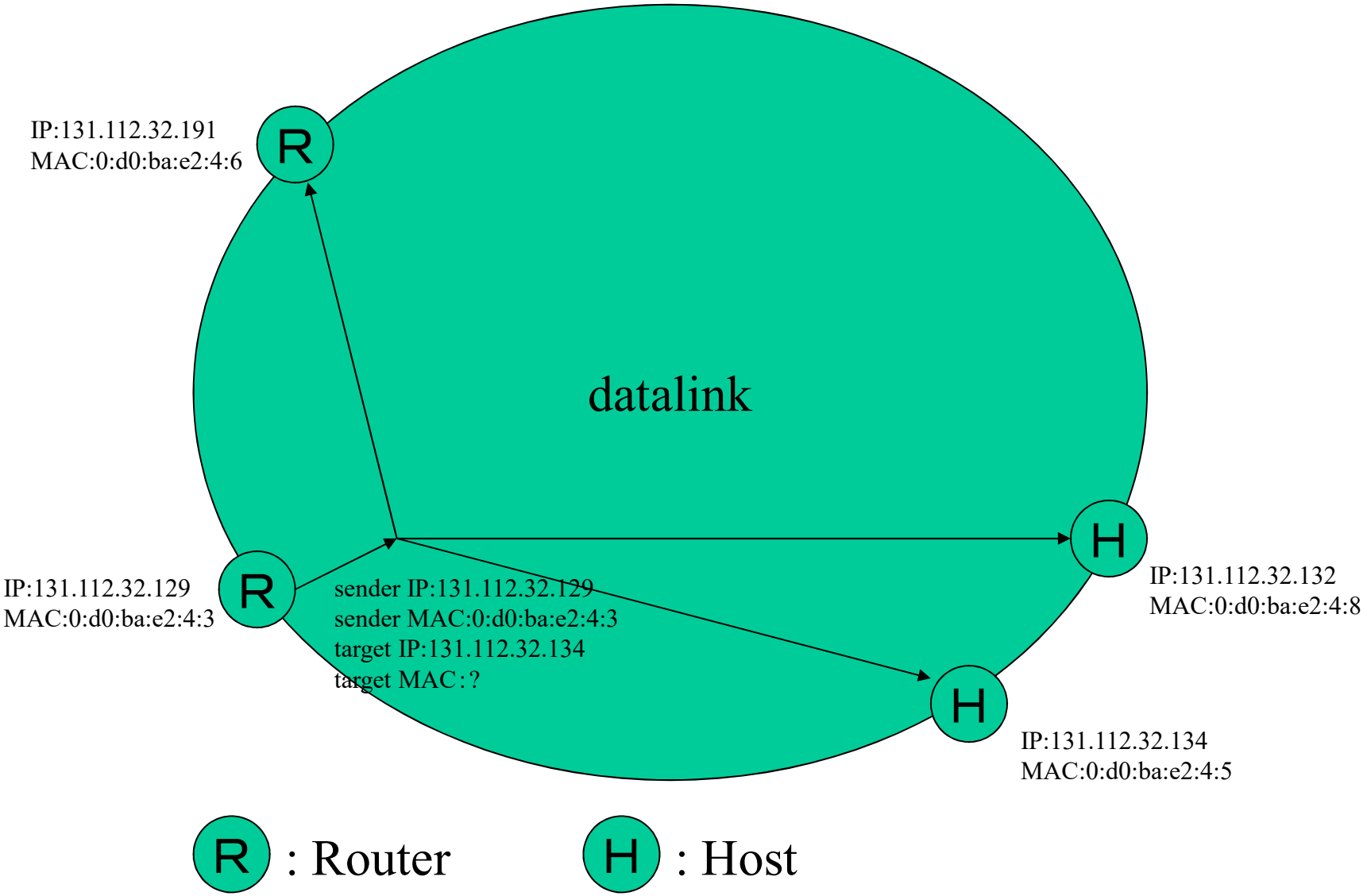
ICMP Time Exceeded

Packet Relaying by Routers to Ethernet by MAC Address

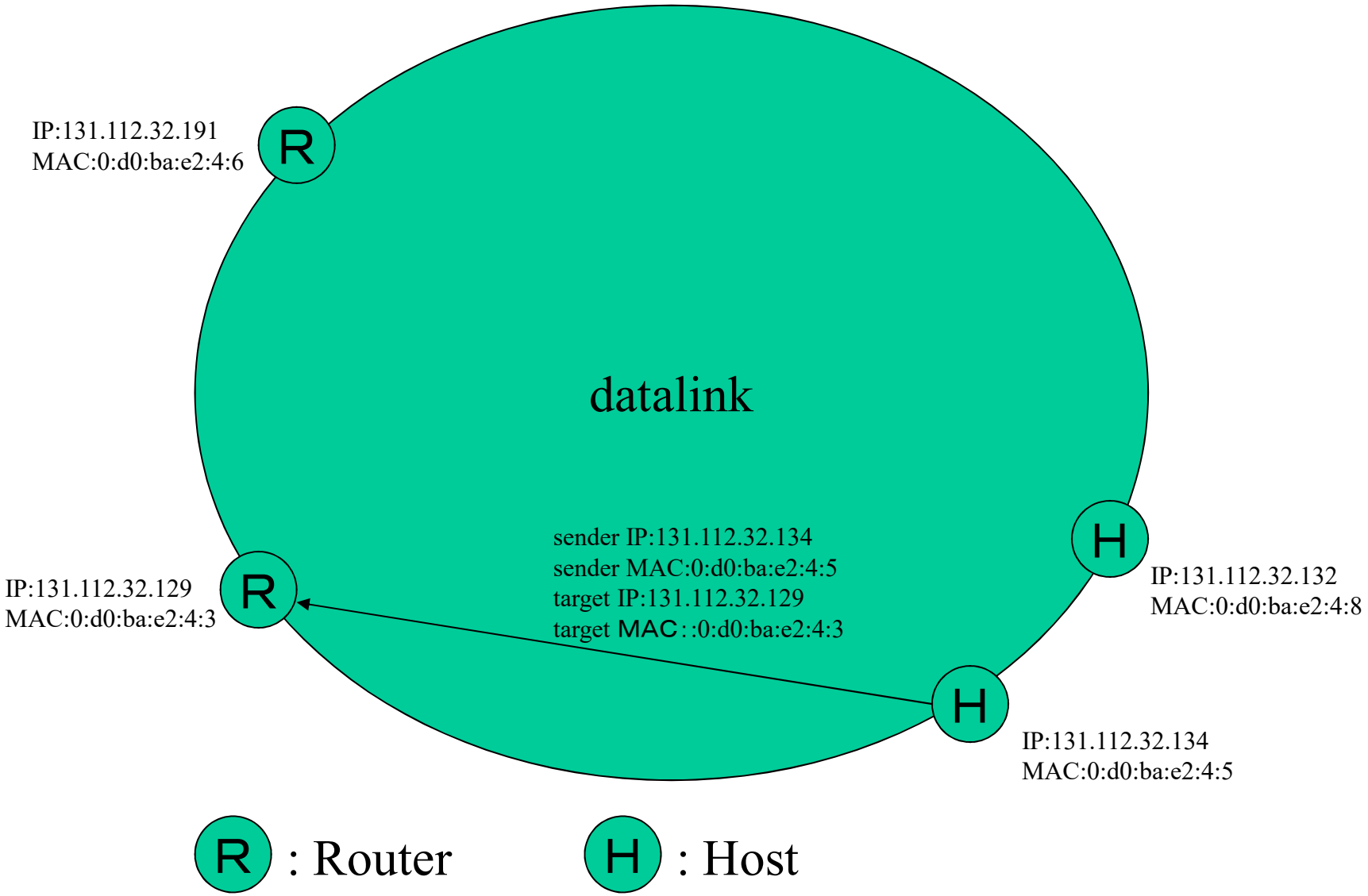
- destination is on datalink directly connected to a router
  - address range of datalink statically configured
  - relay after resolving target MAC address by destination IP address
- destination is on datalink at a distance
  - look up next hop router by destination address
  - relay after resolving target MAC address by IP address of next hop router

ARP (Address Resolution Protocl, rfc826)

- know MAC address of host with known IP address used in a datalink
  - ARP Query including target IP address (and sender MAC and IP address) broadcast over the datalink broadcast
    - ARP Reply by a host with the target IP address
- duplicate IP and MAC address may be detected


IP:131.112.32.191
MAC:0:d0:ba:e2:4:6
R
datalink
H
IP:131.112.32.132
MAC:0:d0:ba:e2:4:8
IP:131.112.32.129
MAC:0:d0:ba:e2:4:3
R
sender IP:131.112.32.129
sender MAC:0:d0:ba:e2:4:3
target IP:131.112.32.134
target MAC:?
H
IP:131.112.32.134
MAC:0:d0:ba:e2:4:5
R : Router
H : Host

ARP Query


IP:131.112.32.191
MAC:0:d0:ba:e2:4:6
R
datalink
sender IP:131.112.32.134
sender MAC:0:d0:ba:e2:4:5
target IP:131.112.32.129
target MAC::0:d0:ba:e2:4:3
H
IP:131.112.32.132
MAC:0:d0:ba:e2:4:8
IP:131.112.32.129
MAC:0:d0:ba:e2:4:3
R
H
IP:131.112.32.134
MAC:0:d0:ba:e2:4:5
R : Router
H : Host

ARP Reply

Wrap Up

- IPv4 do almost nothing in the network
  - deliver packets to their destinations
  - fragmentation
  - maintain TTL (time to live)
- ICMP manages IP and transport layers
- MAC address is dynamically resolved through ARP