Evaluation Method

- Interim and Final Report
- Attendance is not Checked, but, ...
- Questions or Comments are Mandated
  - In the quater, questions or comments with technical content must be made at least twice during lecture (may be in Japanese)
  - Good questions and comments will be awarded with points
  - Declare your name and student ID after each lecture, if you make questions or comments

Advanced Lecture on Internet Applications
2. Transport Layer: TCP, Congestion Control, Long Fat Pipe, Multihoming

Masataka Ohta

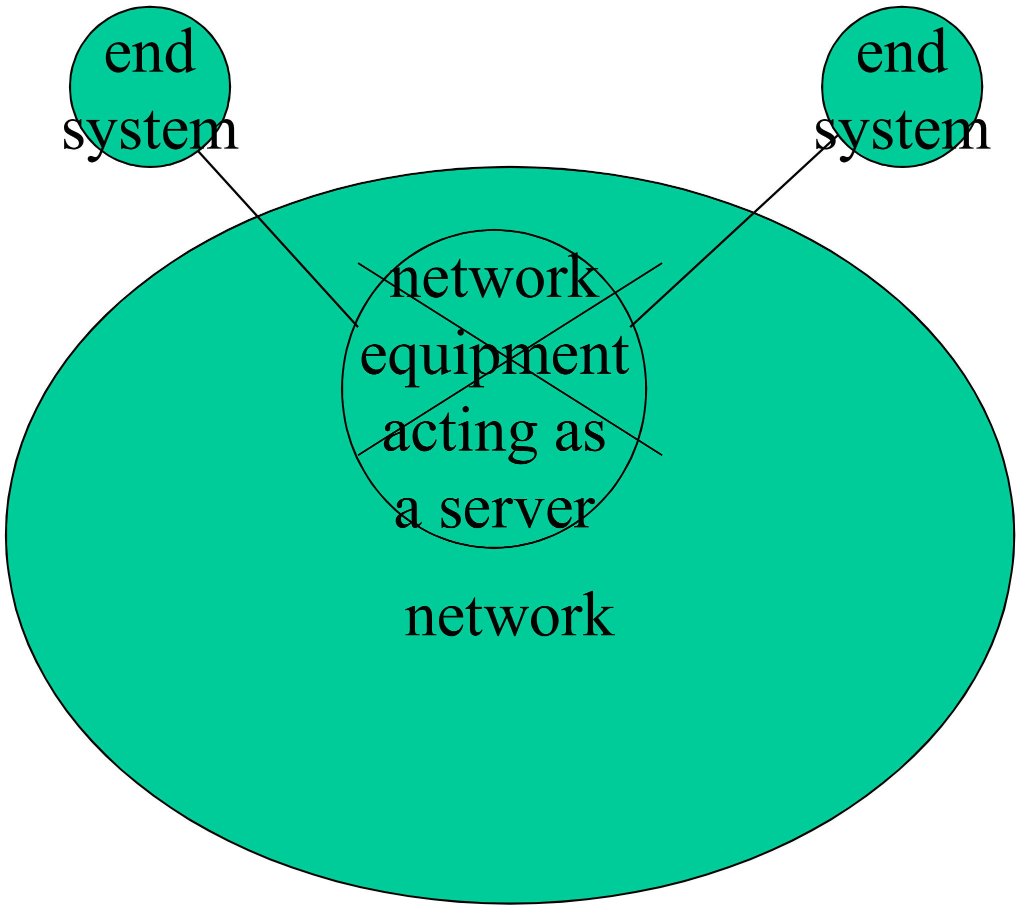
mohta@necom830.hpcl.titech.ac.jp

ftp://ftp.hpcl.titech.ac.jp/appli2e.ppt

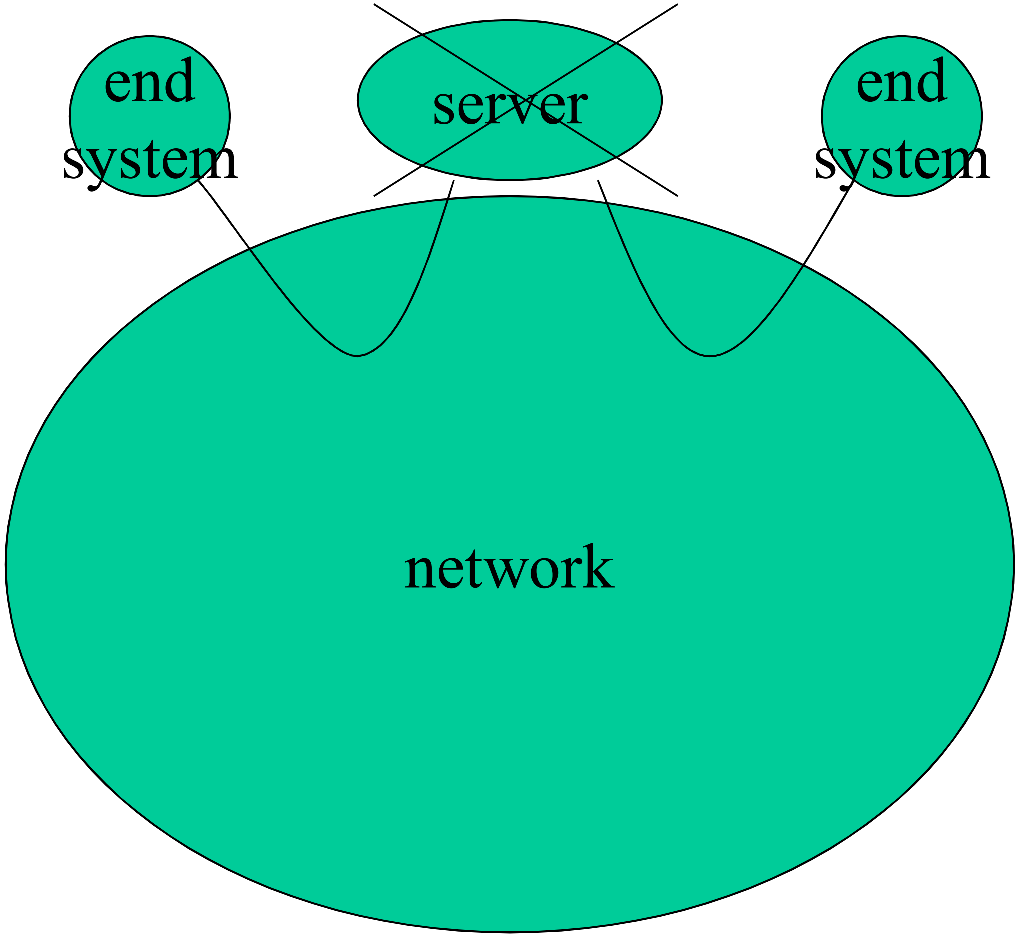
End to End Principle

The Fundamental Principle of the Internet

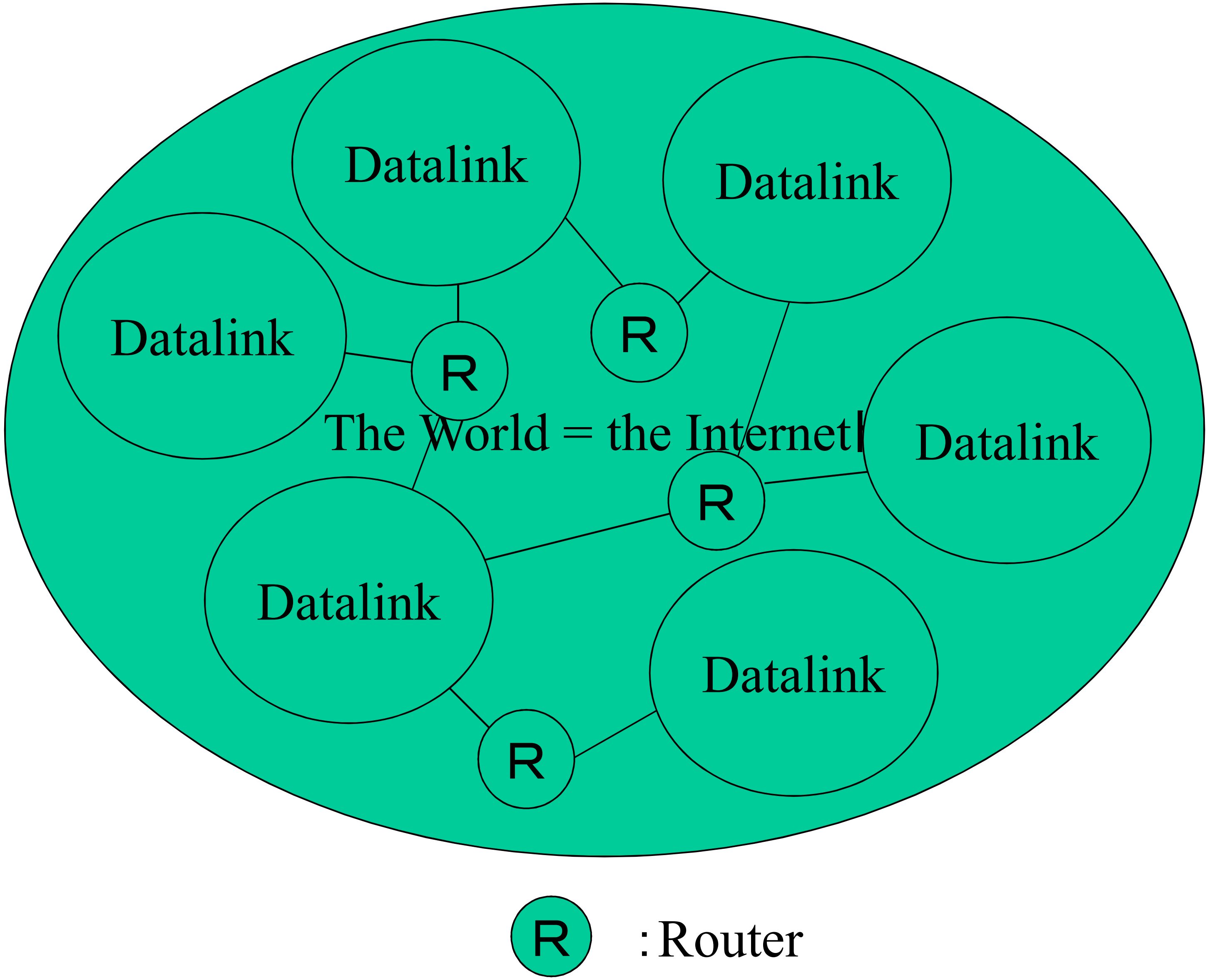
- network do as little as possible
  - only carry packets to destinations
  - no congestion control
- related end systems directly communicate
  - no intermediate intelligent servers



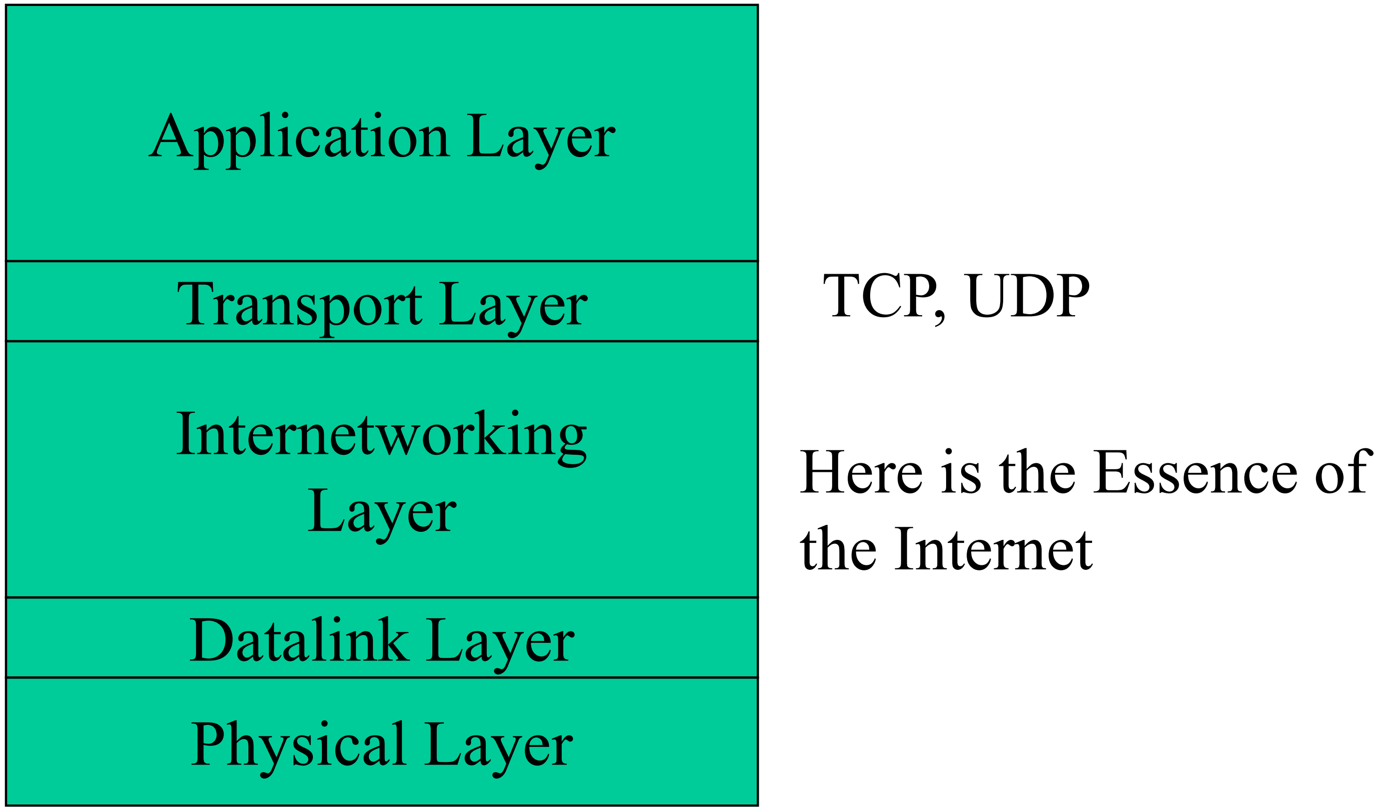
network do nothing



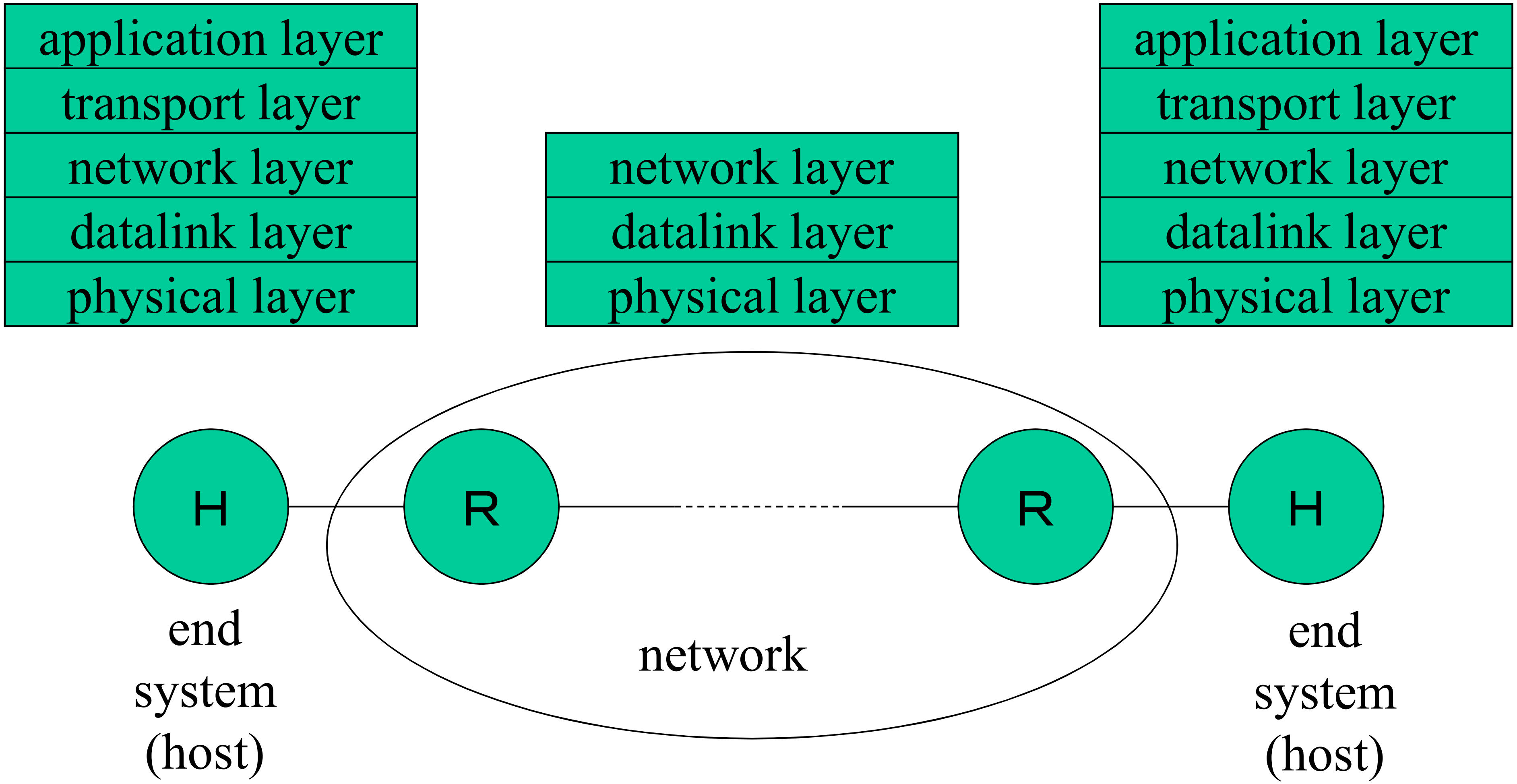
no intermediate server even outside of the network



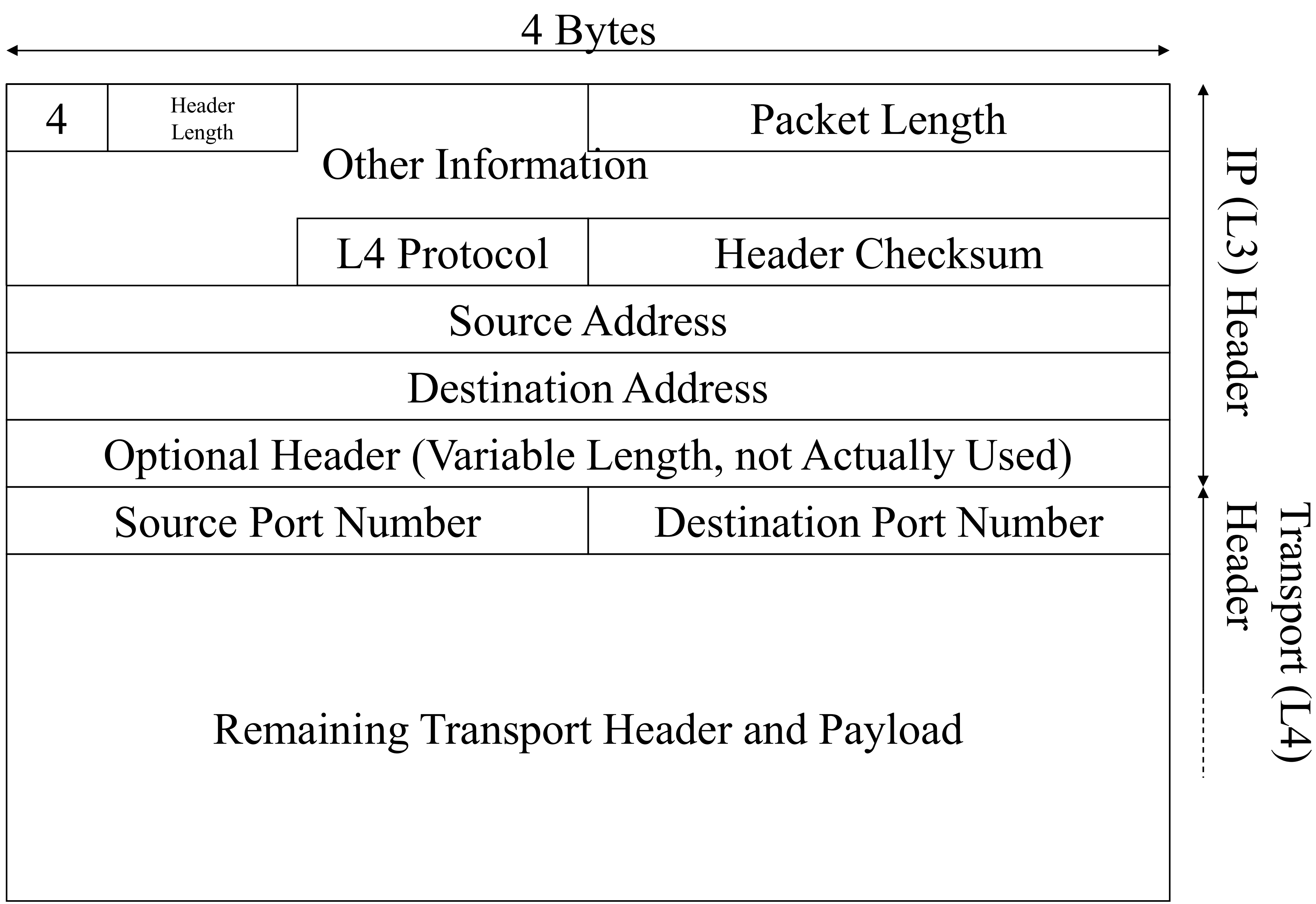
CATENET Model



Layering Structure of the Internet



best effort internet



Format of IPv4 Packets

Function of IP Routers

- decrement TTL and forward packet based on destination address
  - routing table is constructed in advance by routing protocols
  - no advance signaling, no BW guarantee
- with IPv4, may divide packets for datalinks with small MTU (fragmentation)
- IPv4 (rfc791) address is 32bit long
  - transition to IPv6 with 128bit address planned

TCP and UDP

- TCP (rfc793)
  - Transmission Control Protocol
  - retransmit when data error or drop is detected
  - adjust transmission rate
- UDP (rfc768)
  - User Datagram Protocol
  - do nothing (let applications do something)
    - nothing except for delivery to applications

Phone Network vs the Internet

- which is more error prone?

Phone Network

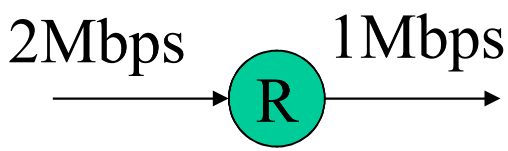
- network for voice transmission (conversation)
- guarantee bandwidth
- minimize delay
  - less than 0.1s of delay desirable for conversations
  - no time for retransmission of lost data (almost impossible with analog circuit)
- error or loss of data causes noise
  - if not very frequent, not a serious problem for conversation

Internet

- network for communication between computers
- a bit of error is often fatal
  - character code, program
  - detection of error or loss required
  - reliable communication over TCP is extensively used
- UDP may be used for voice
- light weight protocols (DNS, tftp, etc.) also use UDP

Reason of Packet Drop

- packet is lost upon transmission error
  - not so common
- routers must drop packet if buffer is full
  - primary reason of packet drop in the Internet



Congestion Control in Phone Network

- no control necessary because bandwidth is reserved?
- connections fail during bandwidth reservation
  - you will here busy tone
    - not your destination but the network is busy

TCP

- connection oriented
- principle is simple
- provisions for various exceptions are complex
- theory and implementation for rate control to avoid congestions is also complex
  - taken care of end systems only

TCP and the Internet

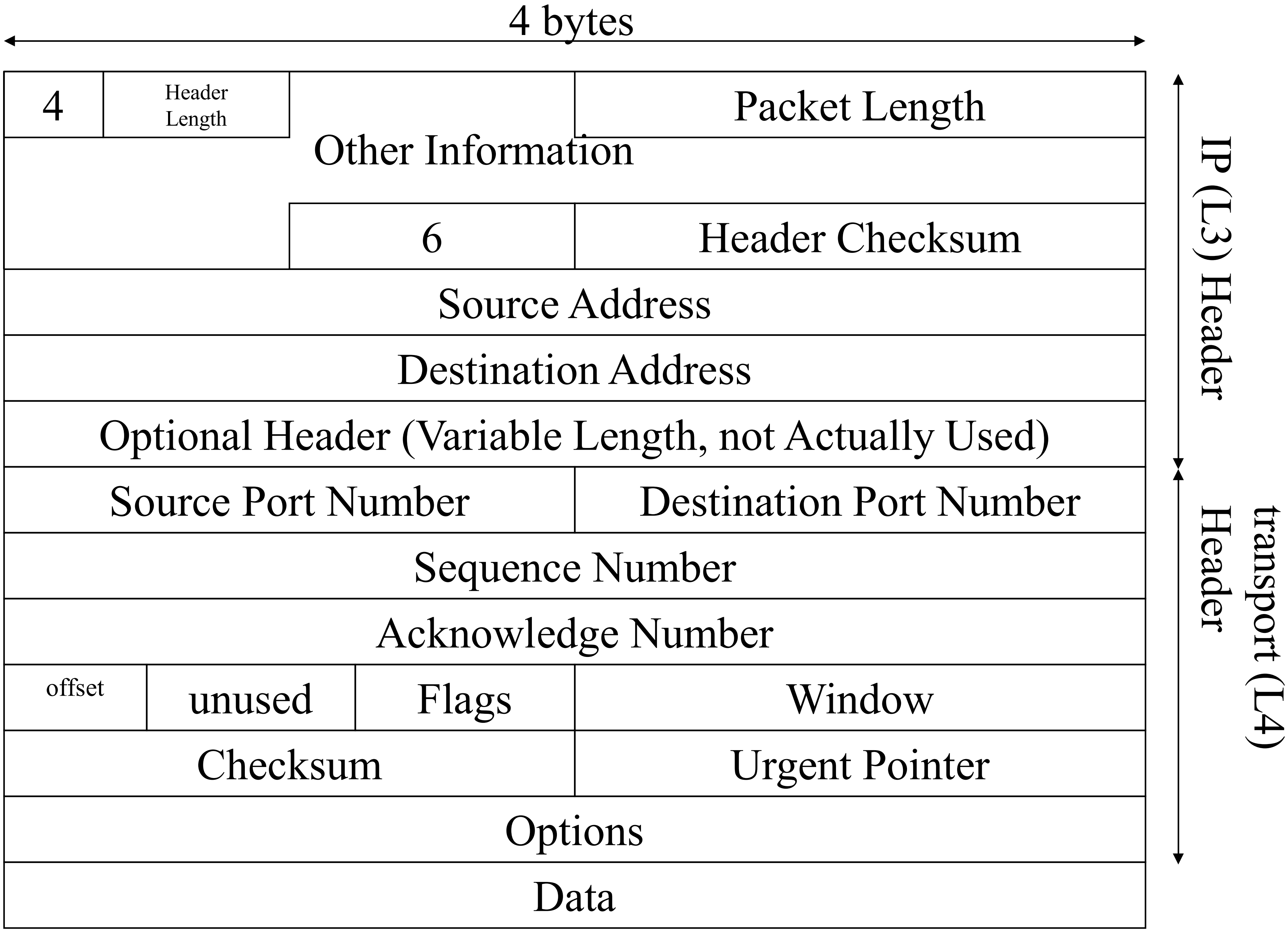
- TCP is the most important transport protocol of the Internet
  - used by almost all applications
  - congestion avoidance of TCP makes the Internet barely operational
    - will collapse?
- TCP/IP protocol suites
  - acronym of set of protocols related to the Internet

Basic Operation of TCP

- initiate connection
  - share sequence numbers to count bytes
- transmit data
  - transmit data within window size
- acknowledge reception of data
  - acknowledge data by sequence number
- retransmit data
  - retransmit unacknowledged data

Reliability and E2E Principle

- there is no error free network
  - retransmission is necessary to recover lost information
  - buffering within network for retransmission
    - fails if the buffer is involved in the error
    - can not adopt to route changes
    - maybe useful if error rate is high "as a performance enhancement"
  - end must hold
    - do not have to make network so much reliable



packet format of TCP over IPv4

Source Port Number, Destination Port Number, Checksum

- same as UDP
  - except for the location of checksum field
- connection is identified by a quadruple of (source address, source port number, destination address, destination port number)
- checksum is mandated

Sequence Number

- sequence number of data (in byte) assigned by source
- initial value is randomly determined
- wraps around to 0 after ffffffff

Acknowledge Number

- sequence number reception of which is acknowledged by destination

Offset

- TCP header length (32bit wise)

Flags

- URG: Urgent Pointer field significant
- ACK: Acknowledgment field significant
- PSH: Push Function
- RST: Reset the connection
- SYN: Synchronize sequence numbers
- FIN: No more data from

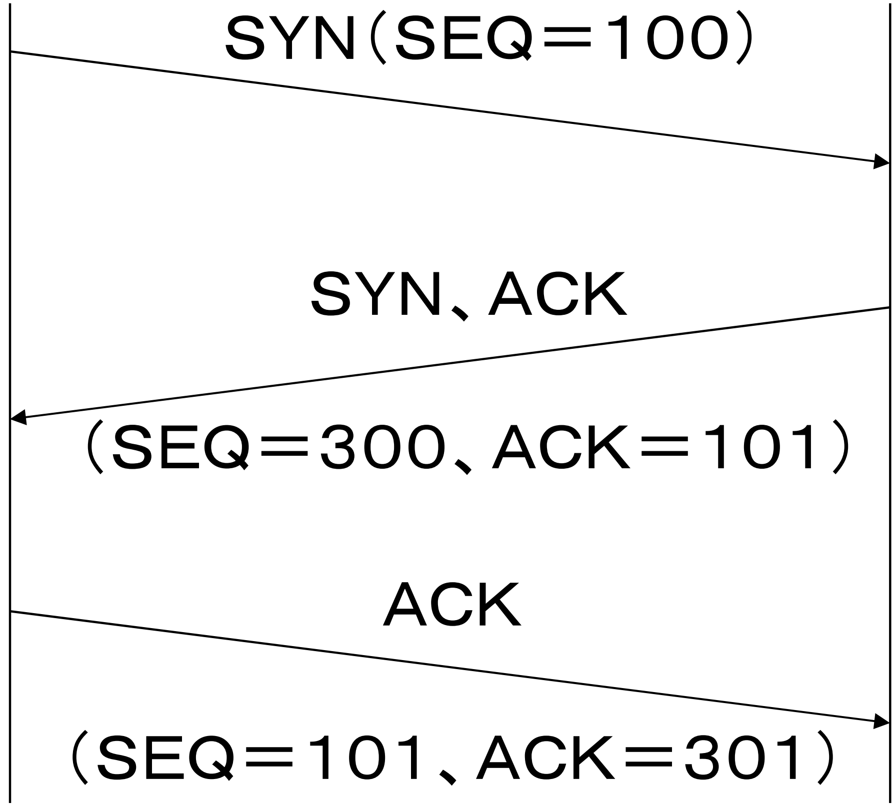
Window

- amount of data source can send before acknowledgement
  - amount of buffer at receiver
- to send data in high speed, window must be large
  - speed=(window size)/(round trip time)
- to avoid congestion, source further limits window size

Urgent Pointer

- location of urgent data in data

Connection Establishment of TCP
3-way Hand Shaking



Congestion Control

- BW is not managed in the Internet
- if everyone send packet at will, large amount of packet loss may occur
- everyone will be happy if packets are sent at rate a little below link BW
- though merely gentlemen's agreement
  - combined with TCP, widely spread
  - can not break the agreement unless both sides cooperate

Practice of Congestion Control by TCP

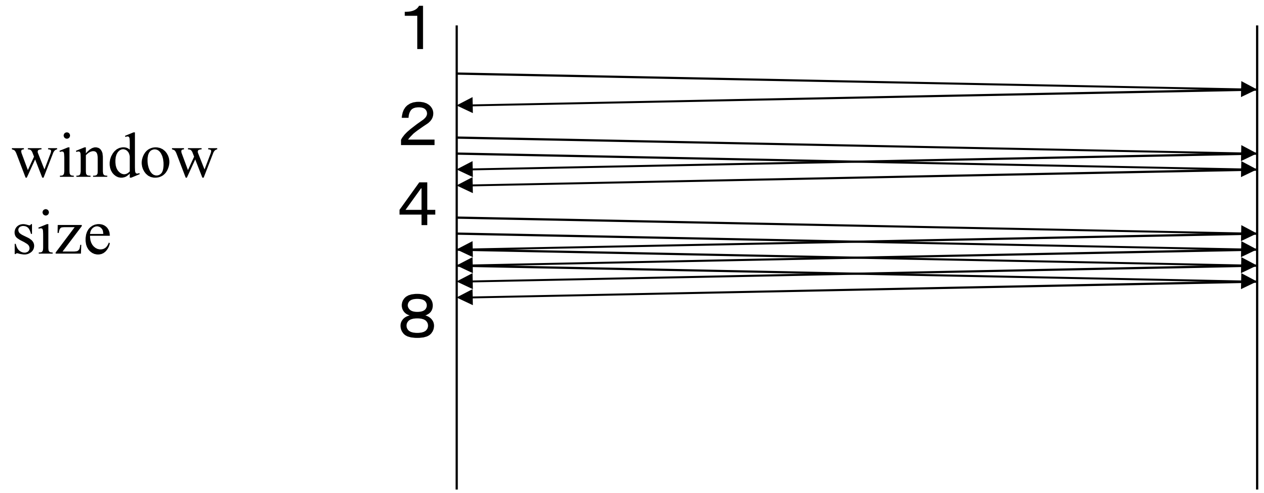
- control TCP rate according to congestion situation of the Internet
  - if congested, reduce transmitter's window size
- what is congestion?
  - packet drop
- packet drop is detected by timeout or acknowledge number not increasing

Congestion Control and E2E Principle

- action for congestion by TCP
  - upon congestion, routers tries buffering and, if buffer is full, just drop packets
    - large buffer is harmful
  - end systems estimate congestion situation in the network and perform complex control (rfc2001)

Slow Start

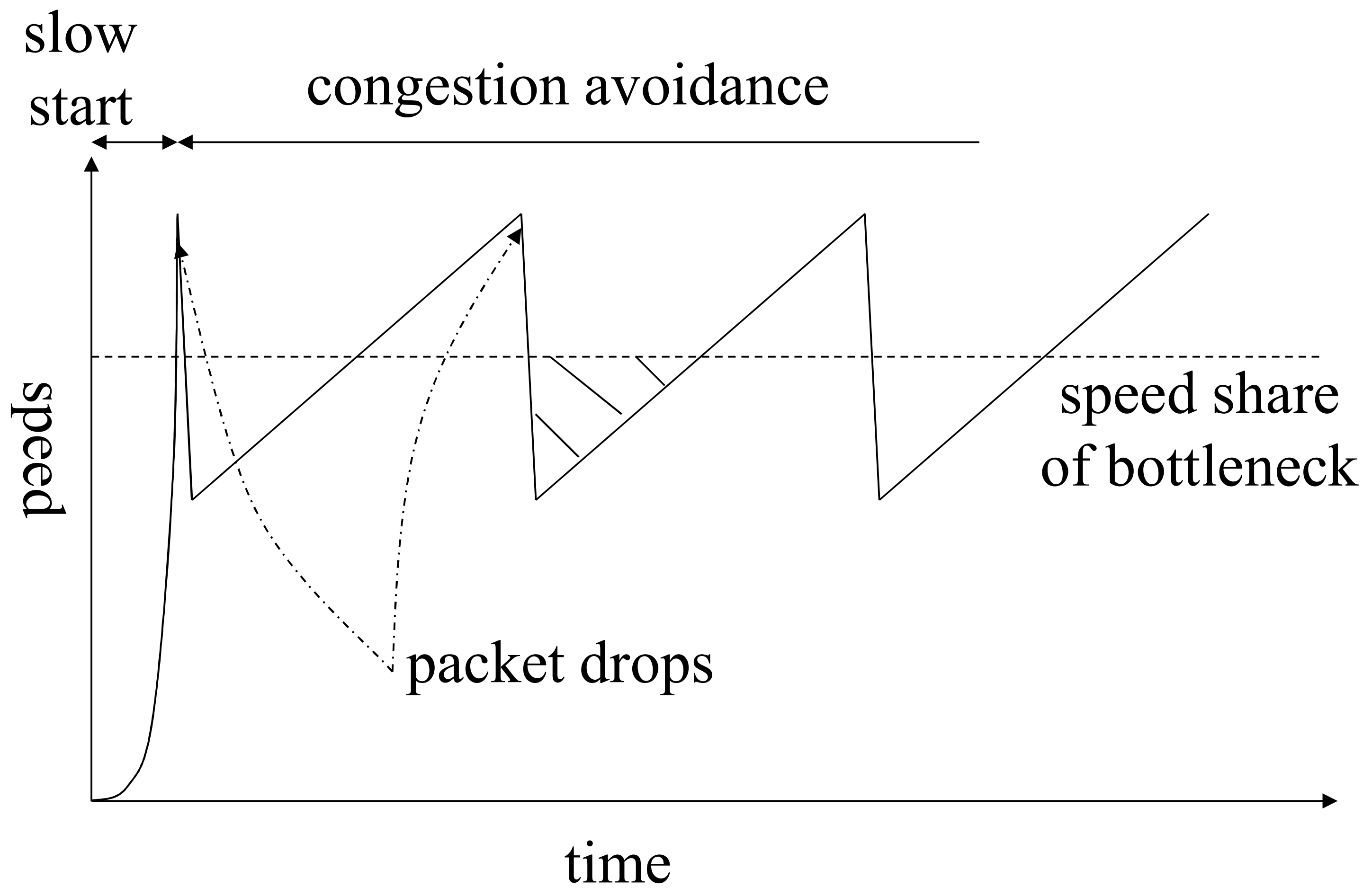
- initial window size is minimum
- if ACK is smoothly replied, increase window size with a fixed amount
  - speed increase exponentially



Congestion Avoidance

- if packet drop is detected
  - window size is halved
- then, window size is increased inversely proportional to the current window size
  - speed increases linearly

Traffic Variation of TCP



Path MTU Discovery

- MTU
  - Maximum Transfer Unit
  - maximum packet size carried by datalink
    - different datalink by datalink (1500B for Ethernet)
- path MTU
  - minimum MTU of path between ends
- path MTU discovery (PMTUD)
  - estimate PMTU

Why PMTUD Necessary?

- with IPv4, packets larger than MTU is divided (fragmented) by intermediate routers
  - heavy weight processing
    - not allowed with IPv6
- larger packet size means less overhead by headers (BW and processing)
- packets of PMTU size is most efficient

Reality of PMTUD

- set “don’t fragment” bit in IPv4 header
  - ICMP error is returned if MTU is exceeded
    - ICMP packet contains MTU of next hop (rfc1191)
- periodically try to send larger packets
  - because route changes may mean MTU changes
- not very practical
  - ICMP packets are often filtered
  - not usable with multicast
  - periodic ICMP error generation increases load

Extension to Congestion Control

- by modifying routers
- RED
  - Random Early Drop
- ECN
  - Explicit Congestion Notification

RED

- tail drop
  - drop packets if buffer is full
- random early drop
  - drop packet with low probability if buffer is occupied to some extent
- can initiate TCP congestion control in early stage

ECN

- mark packets if packet drop is likely to occur
  - use 2 bits of ToS field of IP header
- questionable effect
  - RTT amount of delay to source, anyway
    - not very different from packet drop
    - though packets are not dropped

Collapse of the Internet (game theoretic instability)

- lost packet may be restored by ECC
  - FEC (Forward Error Correction)
- under severe congestion
  - using strong FEC makes the user happy
    - though the FEC increases traffic
  - if all use the FEC, congestion becomes worse
    - beyond correction capability of FEC
- BW guarantee in network inevitable?

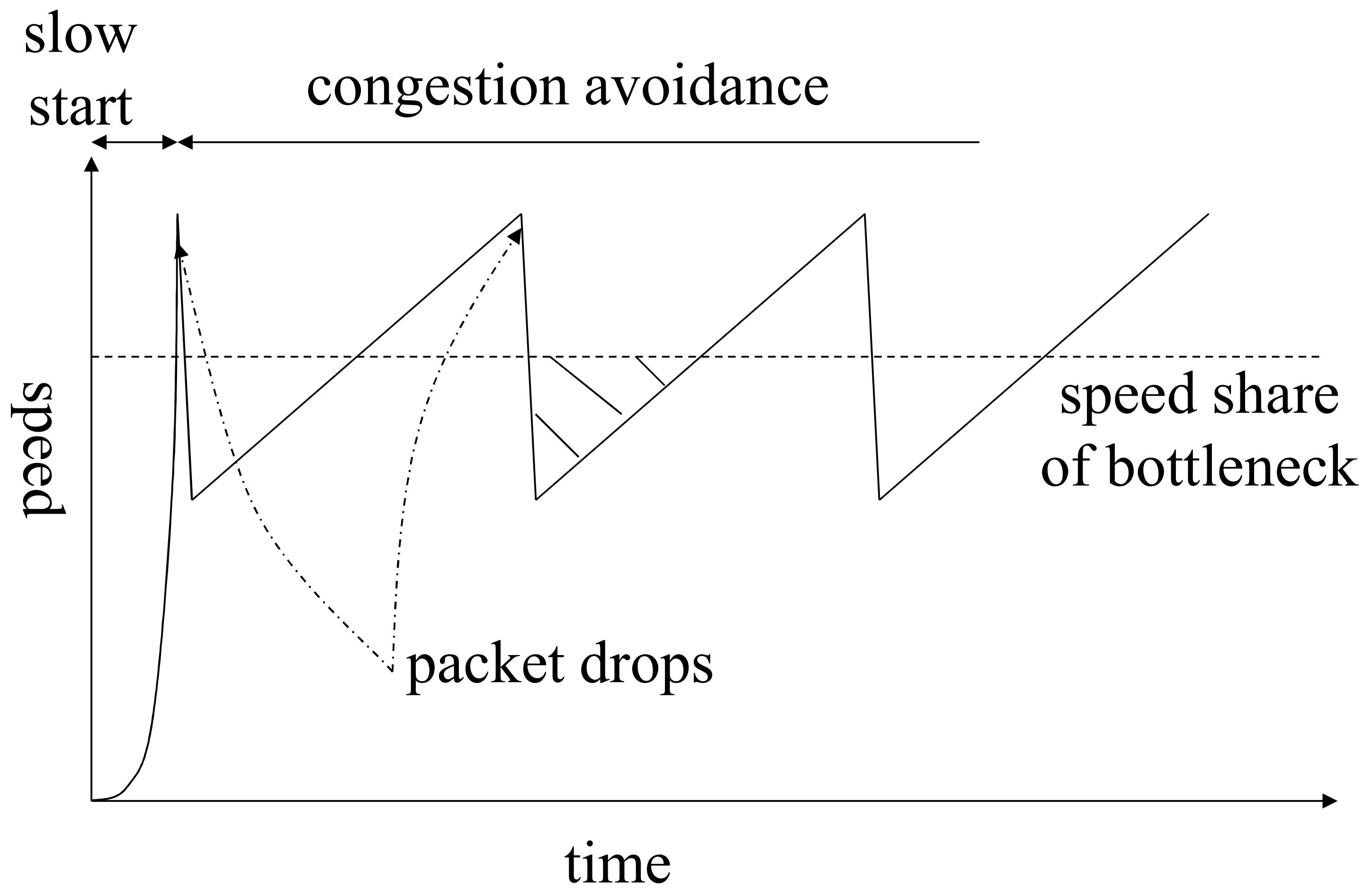
Long Fat Pipe

- TCP performance is limited by RTT
  - speed=(window size)/RTT
- RTT is large for long distance link
  - performance degrages
    - 5Mbps with 0.1s RTT and 64kB window
- packet drop can be disastrous
  - all data within RTT must be resent
- various workaround such as SACK (rfc2018)
  - ultimate solution is BW guarantee

TCP and Router Buffer

- CA makes traffic variation like saw tooth
- without buffer, link speed can not be used up
  - (link(?) delay)*(link BW) of buffer necessary
- backbone routers needs large buffer?
  - backbone is fast
  - backbone is long

Traffic Variation of TCP



TCP and Backbone Router Buffer

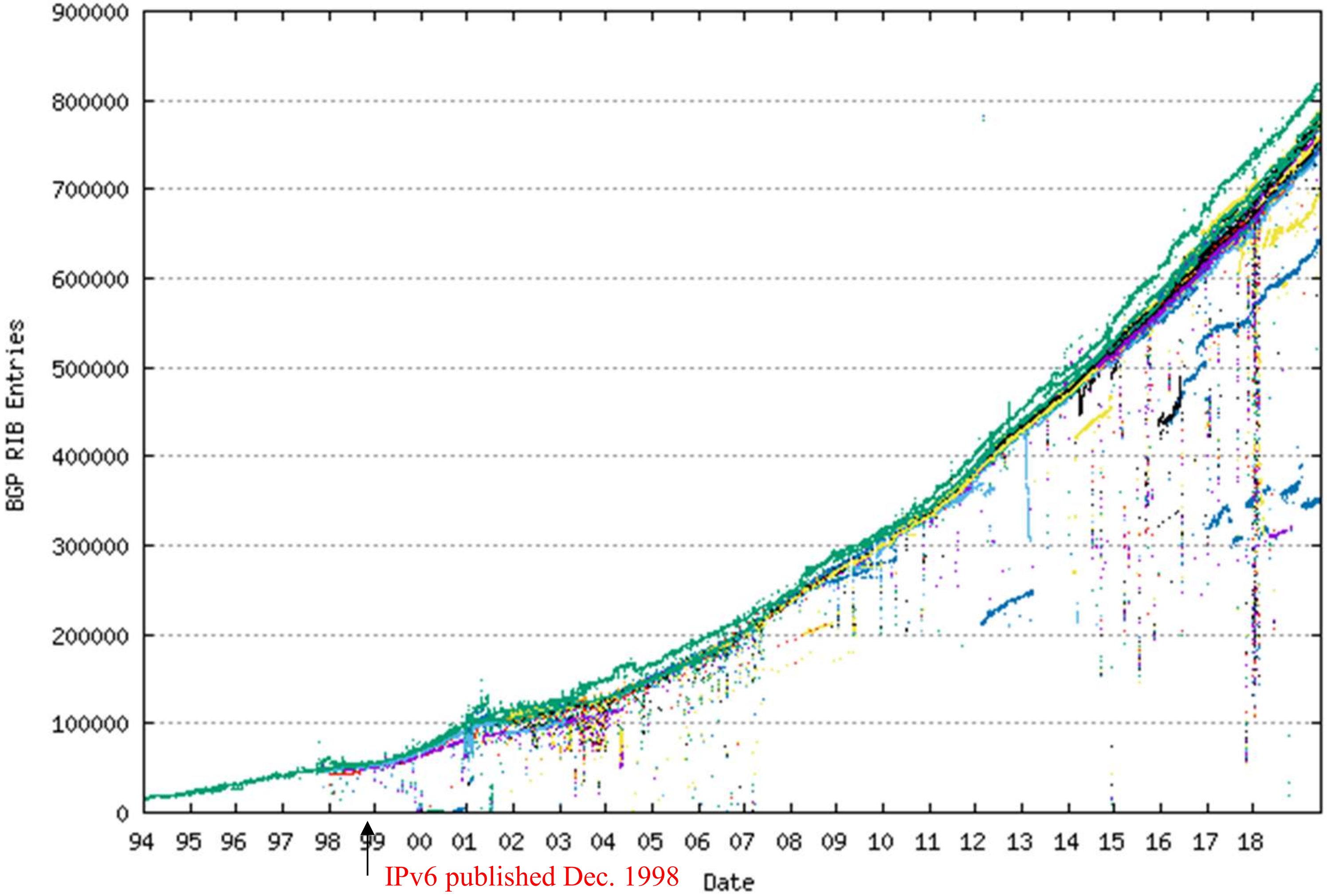
- backbone routers need huge buffer?
  - at backbone, variations of many (*N*) TCP are averaged (if variations are independent)
  - variation is 1/sqrt(*N*)
    - need 1/sqrt(*N*) less buffer?
    - if several times of 1/sqrt(*N*) of link speed is sacrificed
      - traffic rarely to exceed link capacity (if poisson)
    - buffering of several tens of packets to absorb short term variation is enough
      - » optical packet router is plausible

Function of IP Routers

- decrement TTL and forward packet based on destination address
  - routing table is constructed in advance by routing protocols
  - no advance signaling, no BW guarantee
- with IPv4, may divide packets for datalinks with small MTU (fragmentation)

IPv4 Routing Table Size

http://bgp.potaroo.net/

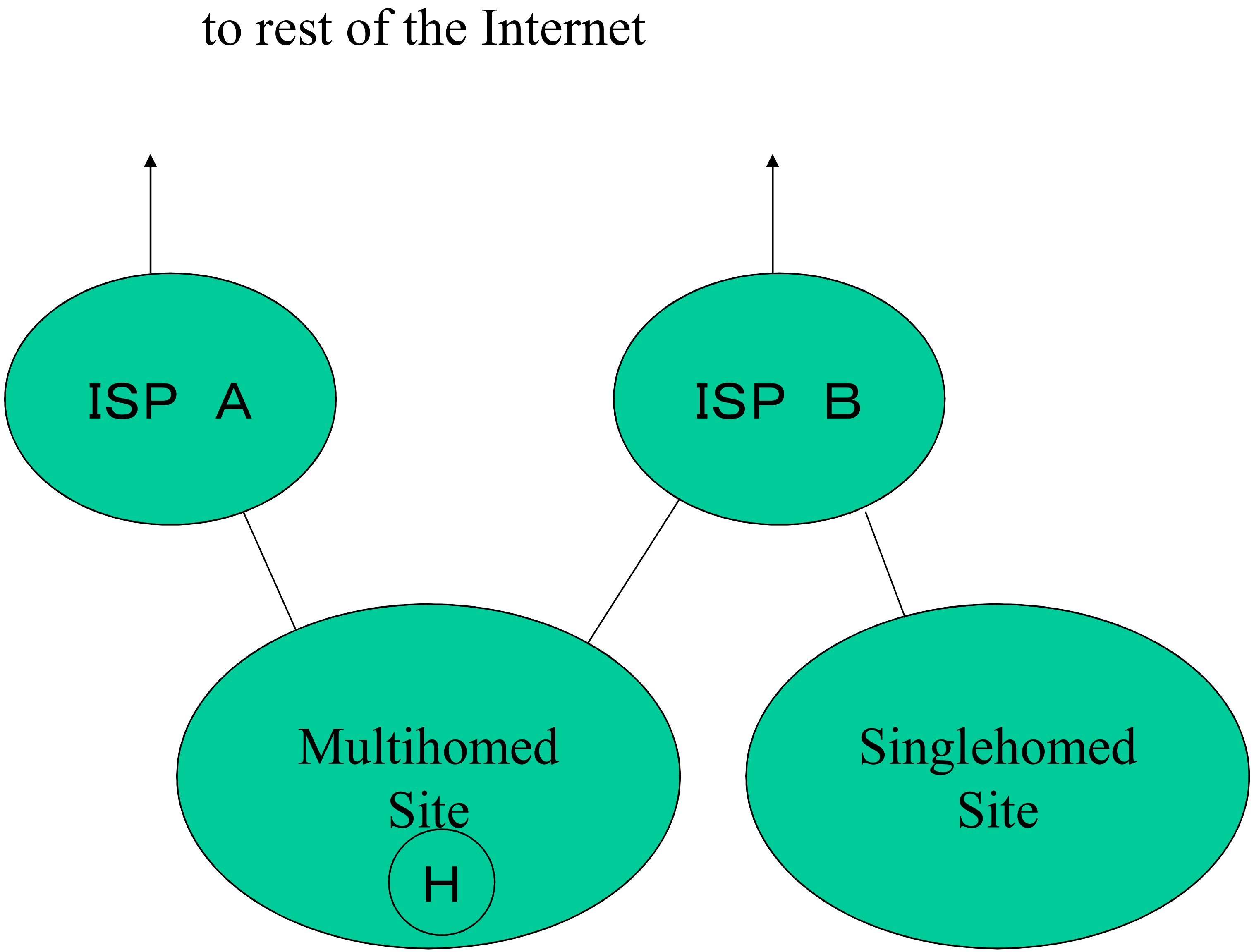


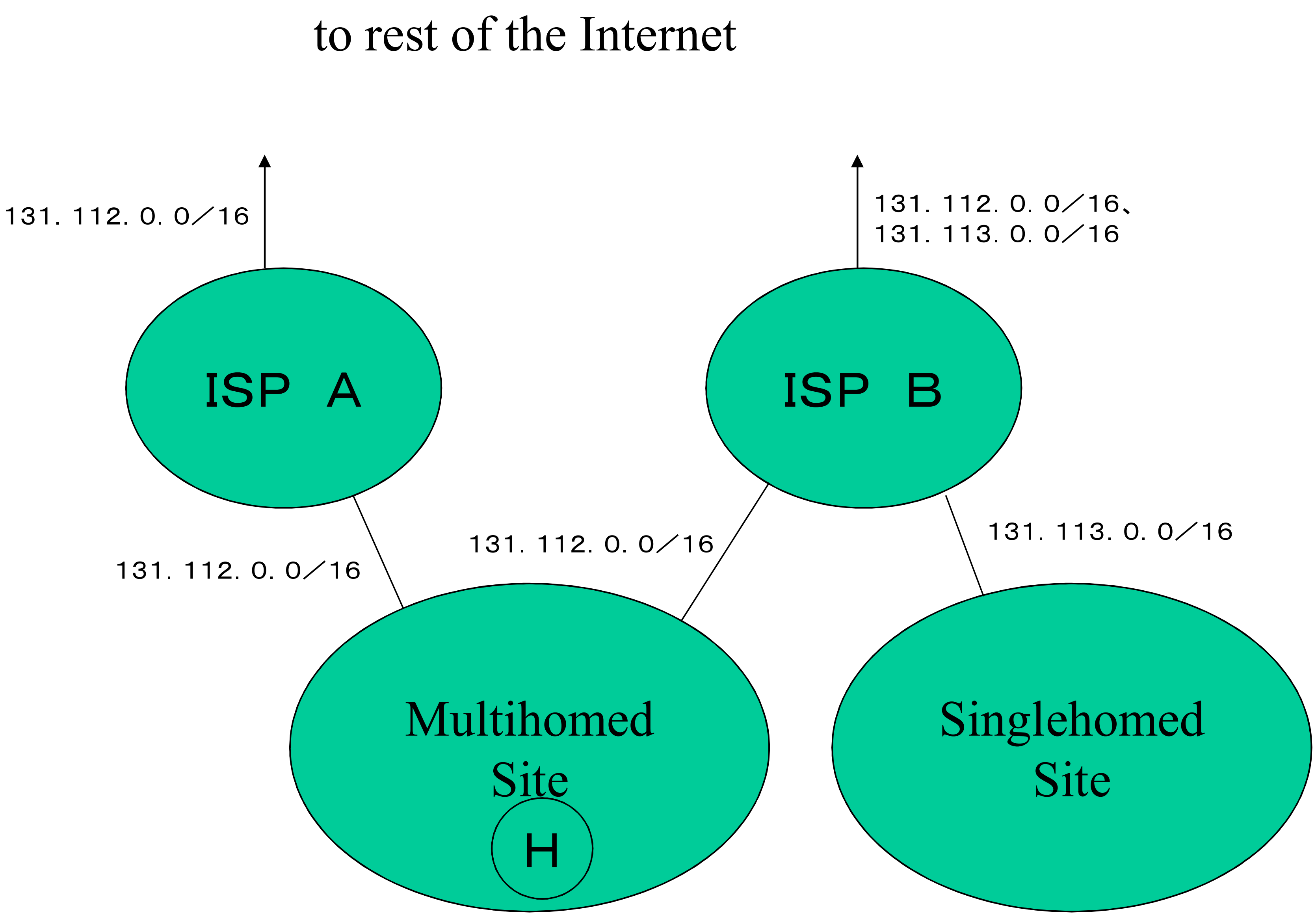
Cases When Route Aggregation Impossible

- 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

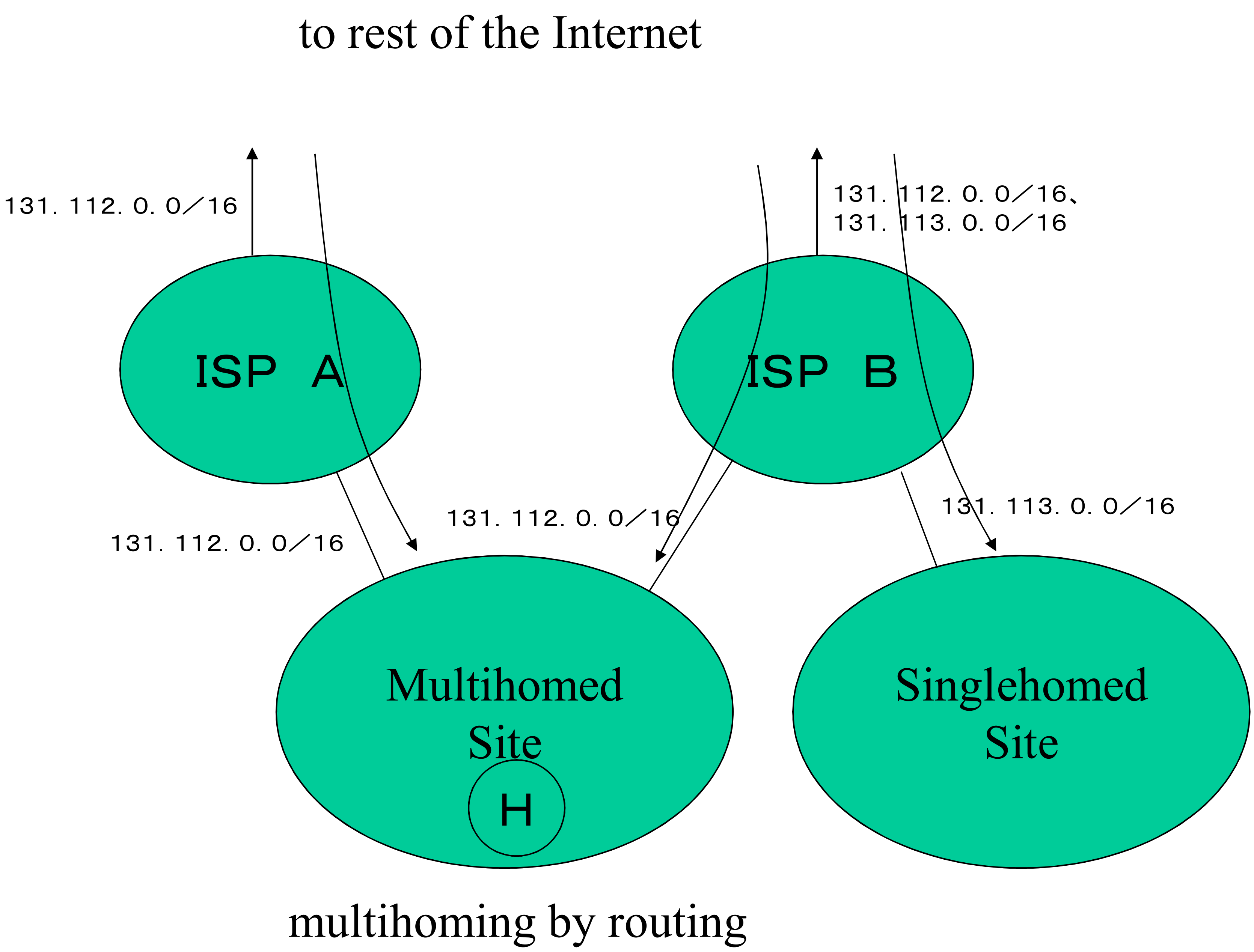
Multihoming

- have multiple upstream ISPs
  - safe even if some ISPs fail
- necessary for reliable service
  - police, ISP, banks, large corporations etc.
- multihoming by routing send single address range to multiple pathes
  - let routing protocol of the network choose the better (or available) one

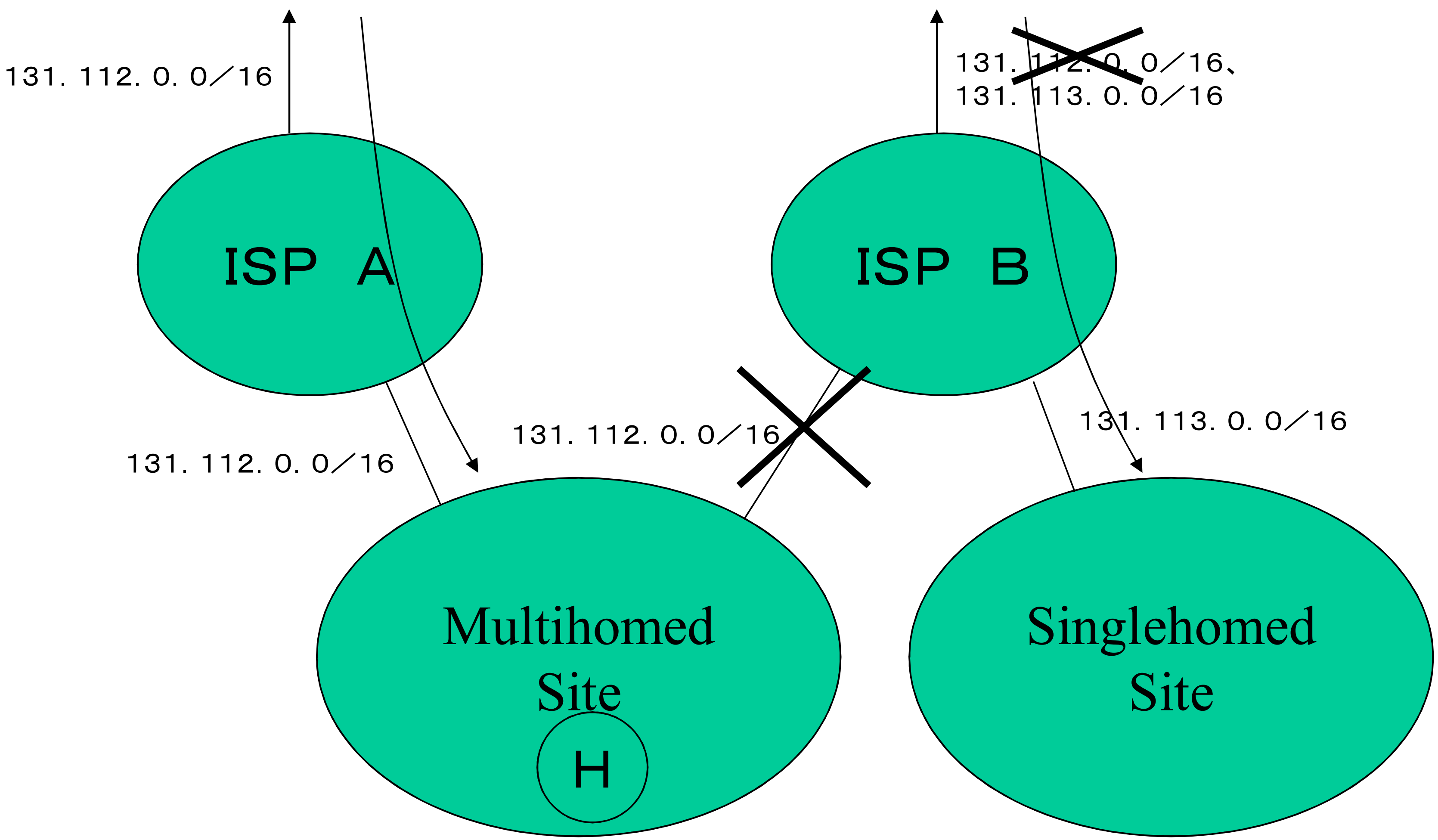

to rest of the Internet
ISP A
ISP B
Multihomed
Site
H
Singlehomed
Site



multihoming by routing

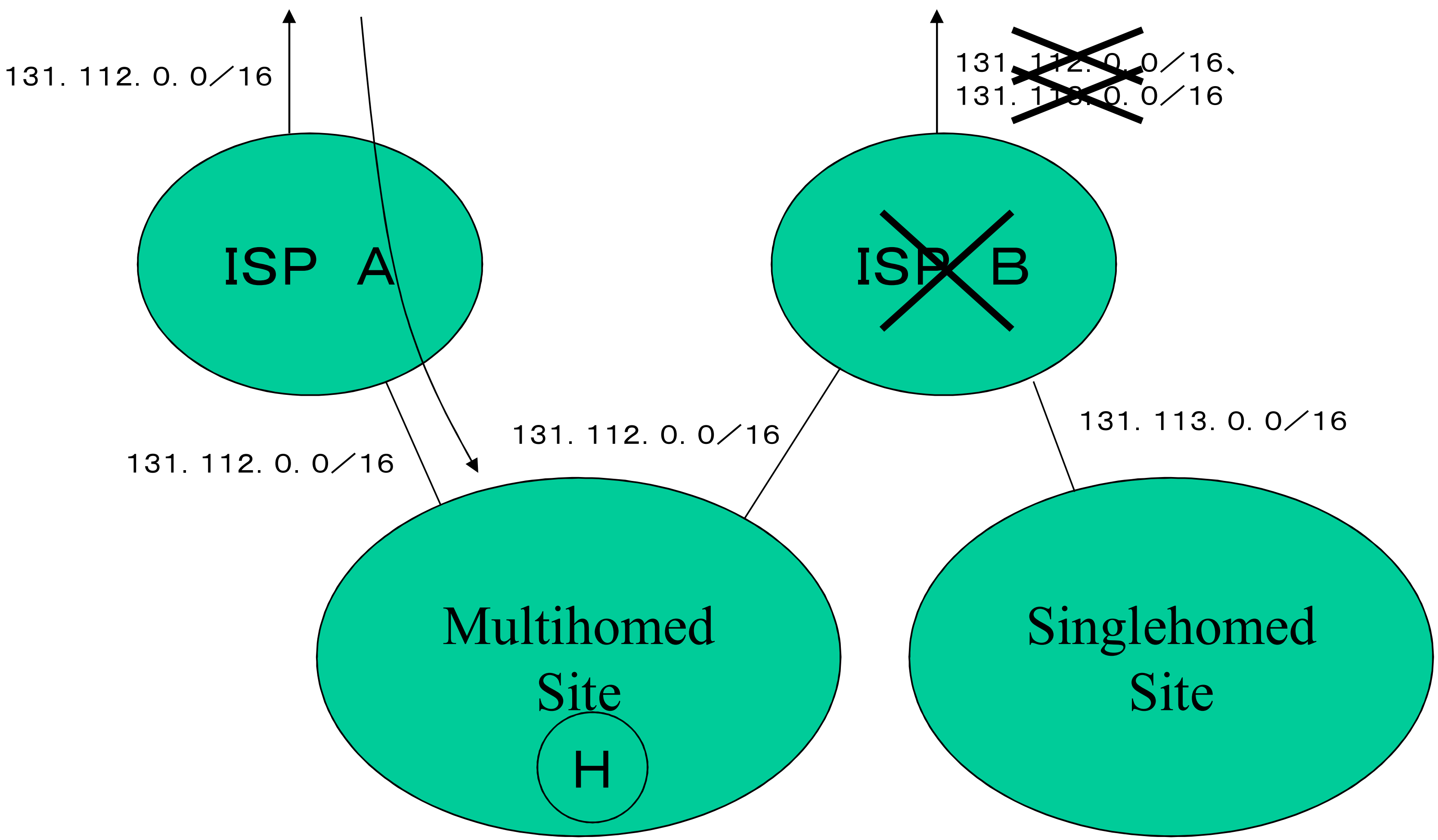

to rest of the Internet
131．112．0．0／16
131．112．0．0／16、
131．113．0．0／16
ISP A
ISP B
131．112．0．0／16
131．112．0．0／16
131．113．0．0／16
Multihomed
Site
H
Singlehomed
Site
multihoming by routing

to rest of the Internet

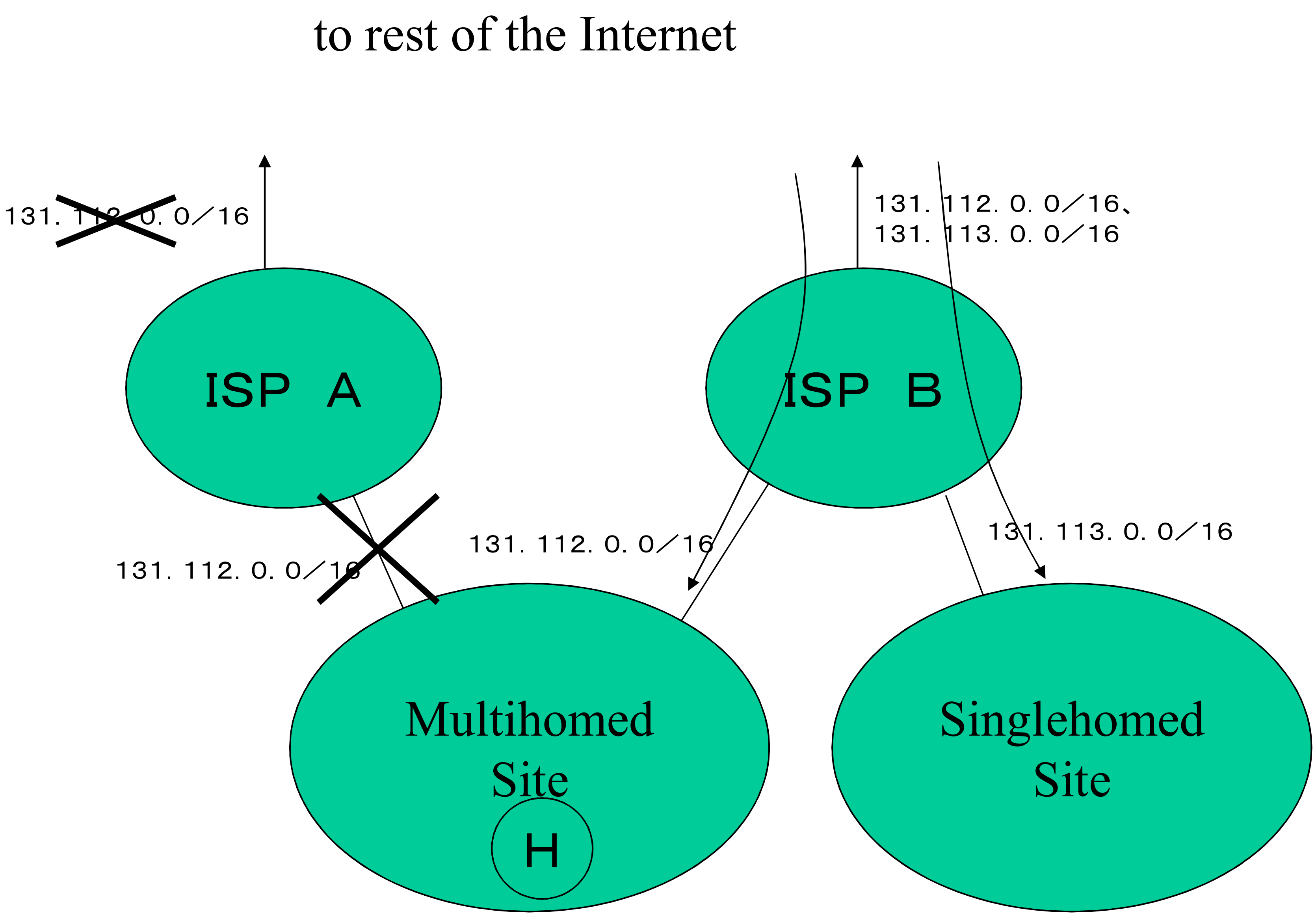


multihoming by routing

to rest of the Internet



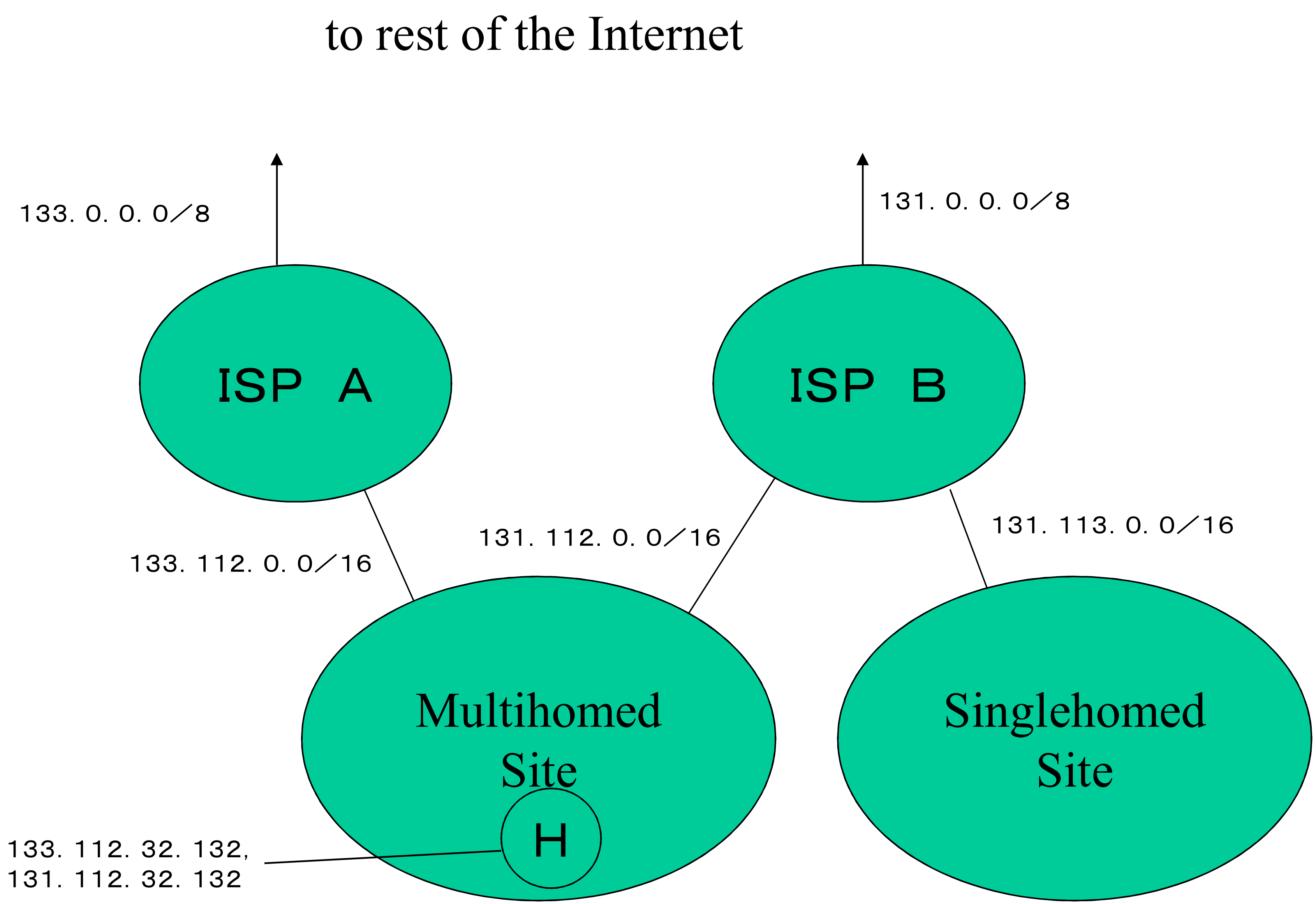
multihoming by routing



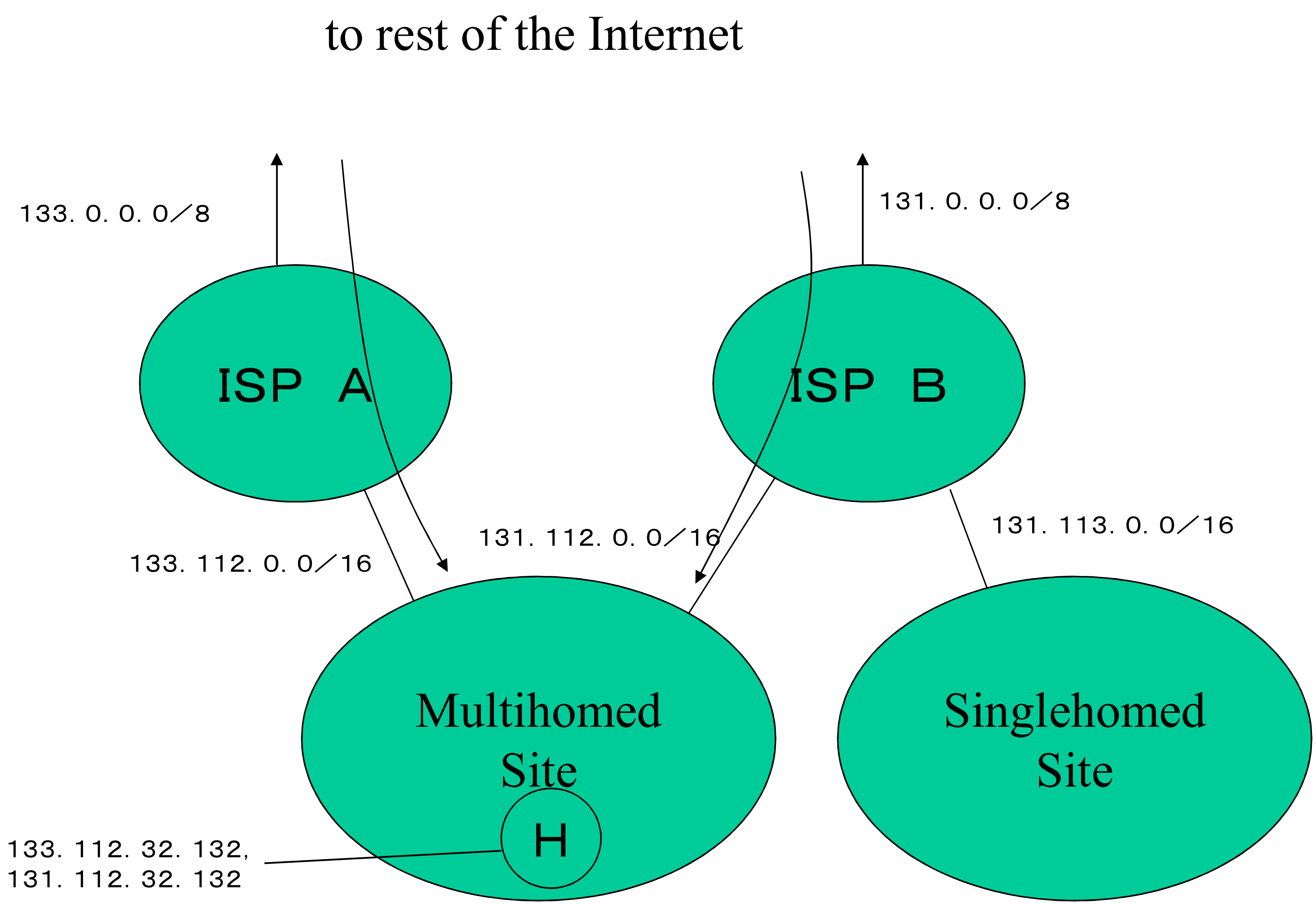
multihoming by routing

End to End Multihoming

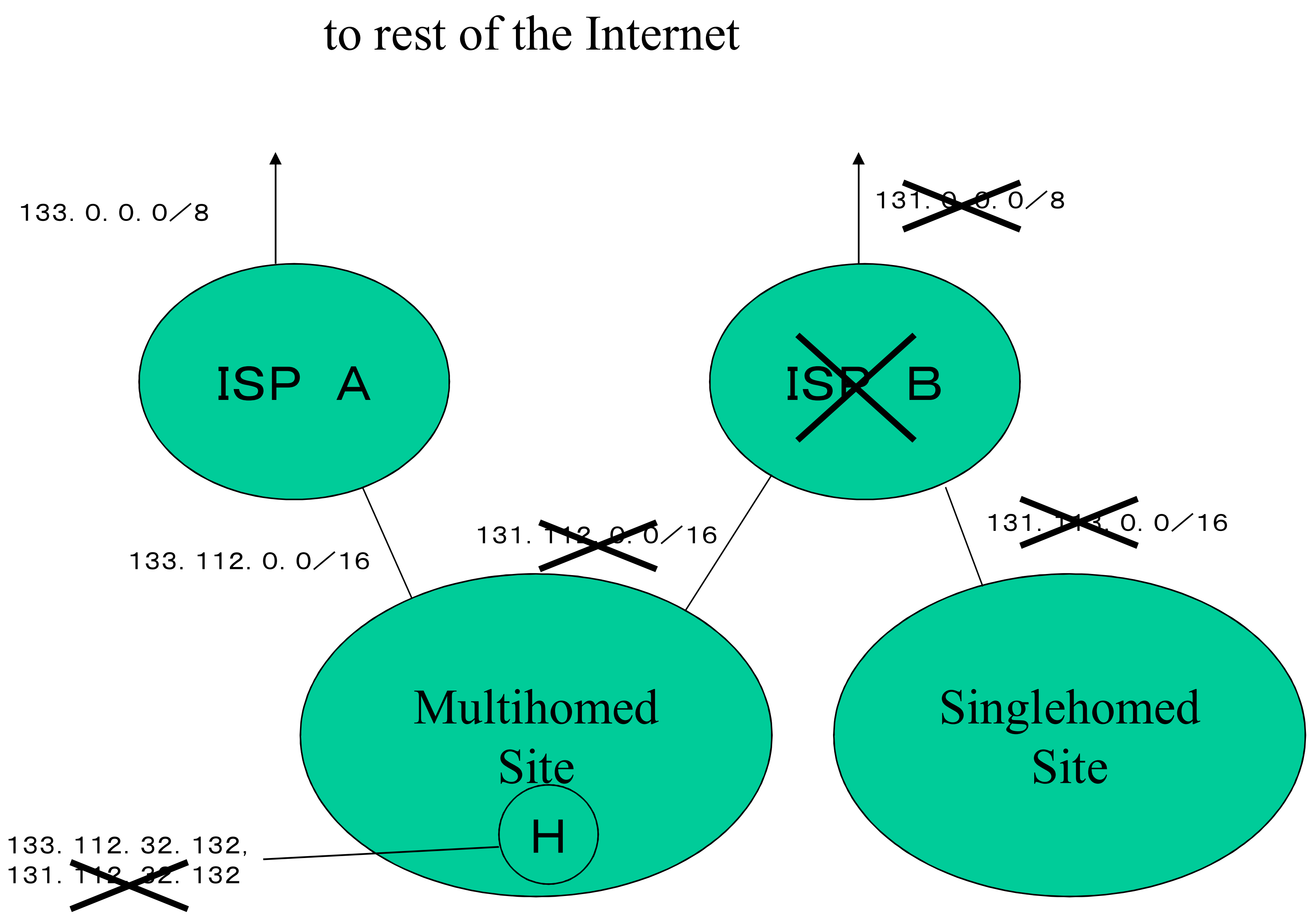
- a host has multiple IP addresses
- transport or application layer of peer of the host try to use best address of the host
  - rough unreachability by global routing table
  - if some address works, communication starts
  - if timeout occurs, other addresses are tried
- multihoming by routing is not necessary



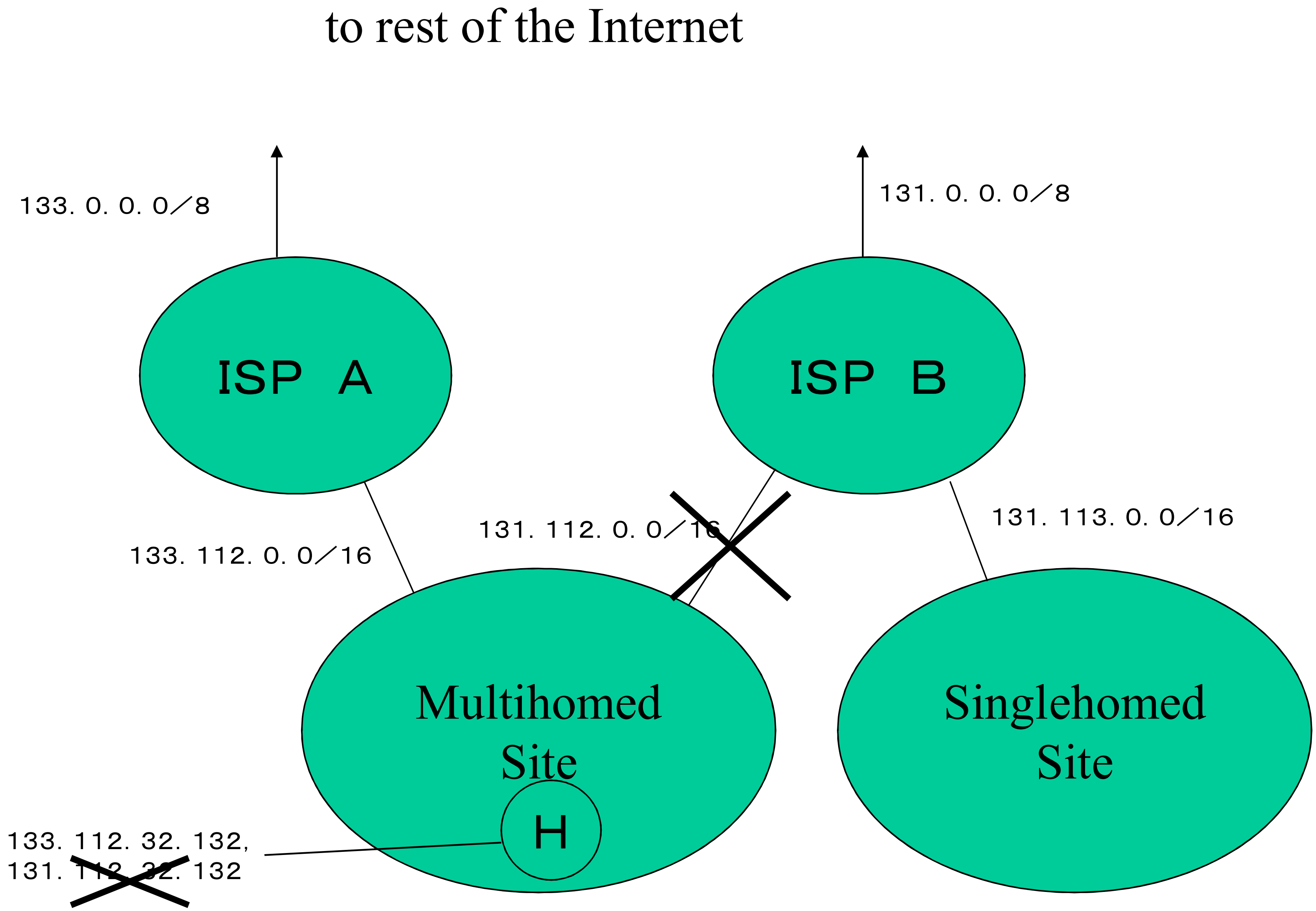
end to end multihoming



end to end multihoming

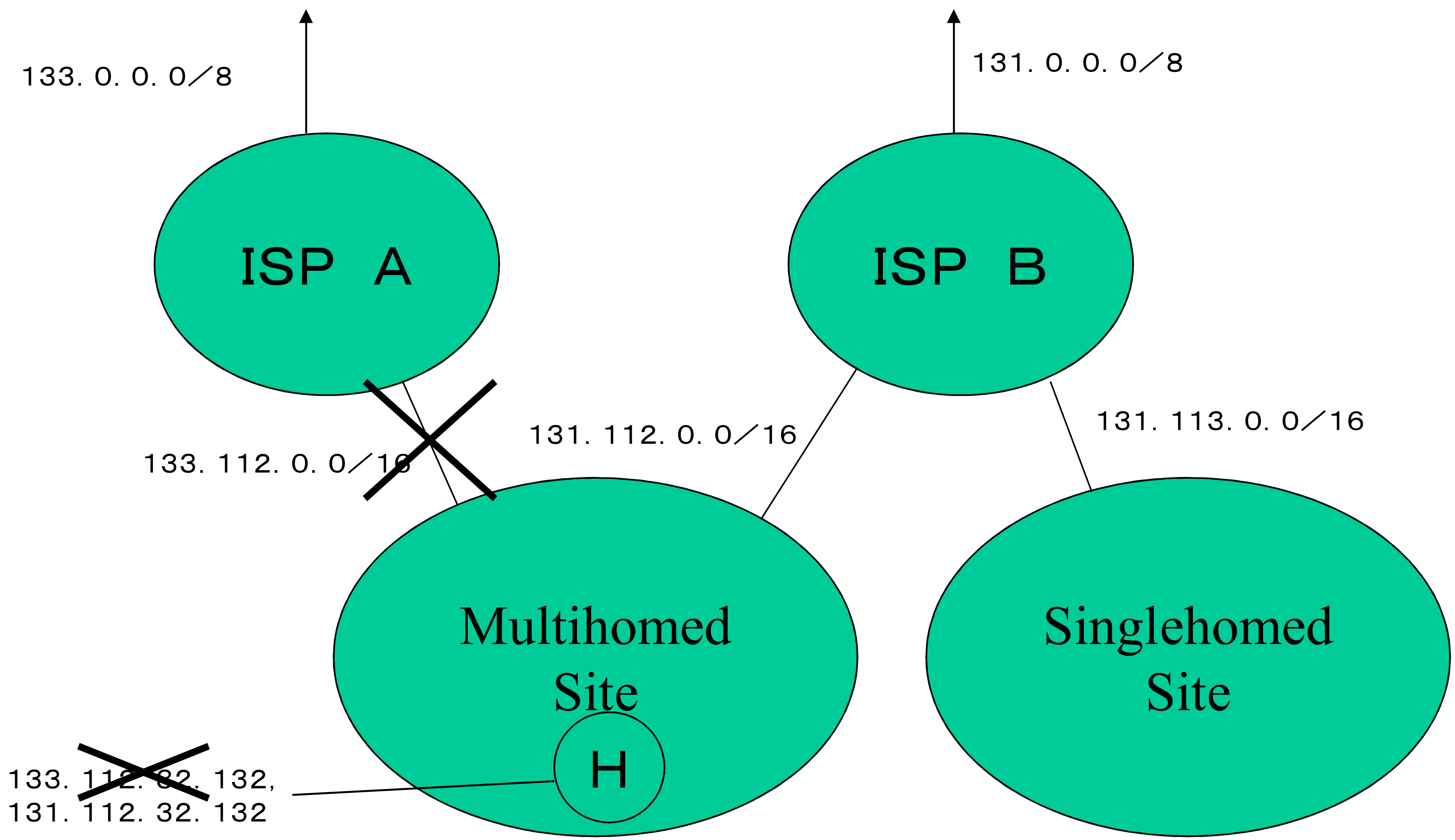


end to end multihoming



end to end multihoming

to rest of the Internet



end to end multihoming

e-mail and E2E Multihoming

- e-mail (SMTP+DNS (rfc974) supports E2E multihoming at application layer
  - if a mail server have multiple addresses
    - all the addresses are tried
  - it is of course as e-mail was the most important application of the Internet
- DNS also support E2E multihoming
  - all the addresses of NSes are tried

Wrap Up

- Internet is as reliable as phone network
- internet does not control BW in network
  - drop packets upon congestion
- dropped packets are retransmitted by TCP
  - speed control is by TCP at ends
- multihoming should also be performed by ends