

# Advanced Lecture on Internet Infrastructure

## 5. IPv6, ND, NWGN

Masataka Ohta

[mohita@necom830.hpcl.titech.ac.jp](mailto:mohita@necom830.hpcl.titech.ac.jp)

<ftp://chacha.hpcl.titech.ac.jp/infra5e.ppt>

# Merit and Demerit of Monopolization

- merit
  - success of the Internet monopolize IT to make phone and broadcast networks disappear
    - cost reduction and speed increase of of IT environment
- demerit
  - success of IPv4 monopolize IT to make IPv6 disappear
    - future break down of the Internet?

# IPv6 won't be Popular?

- any IPv6 capable applications work with IPv4
  - occupy same ecological niche
- IPv4 is commercially necessary
  - IPv4 has scale merit
- IPv6 won't be popular with free competition in private sectors
  - political intervention should be necessary

# IPv6 is Definitely Necessary (?)

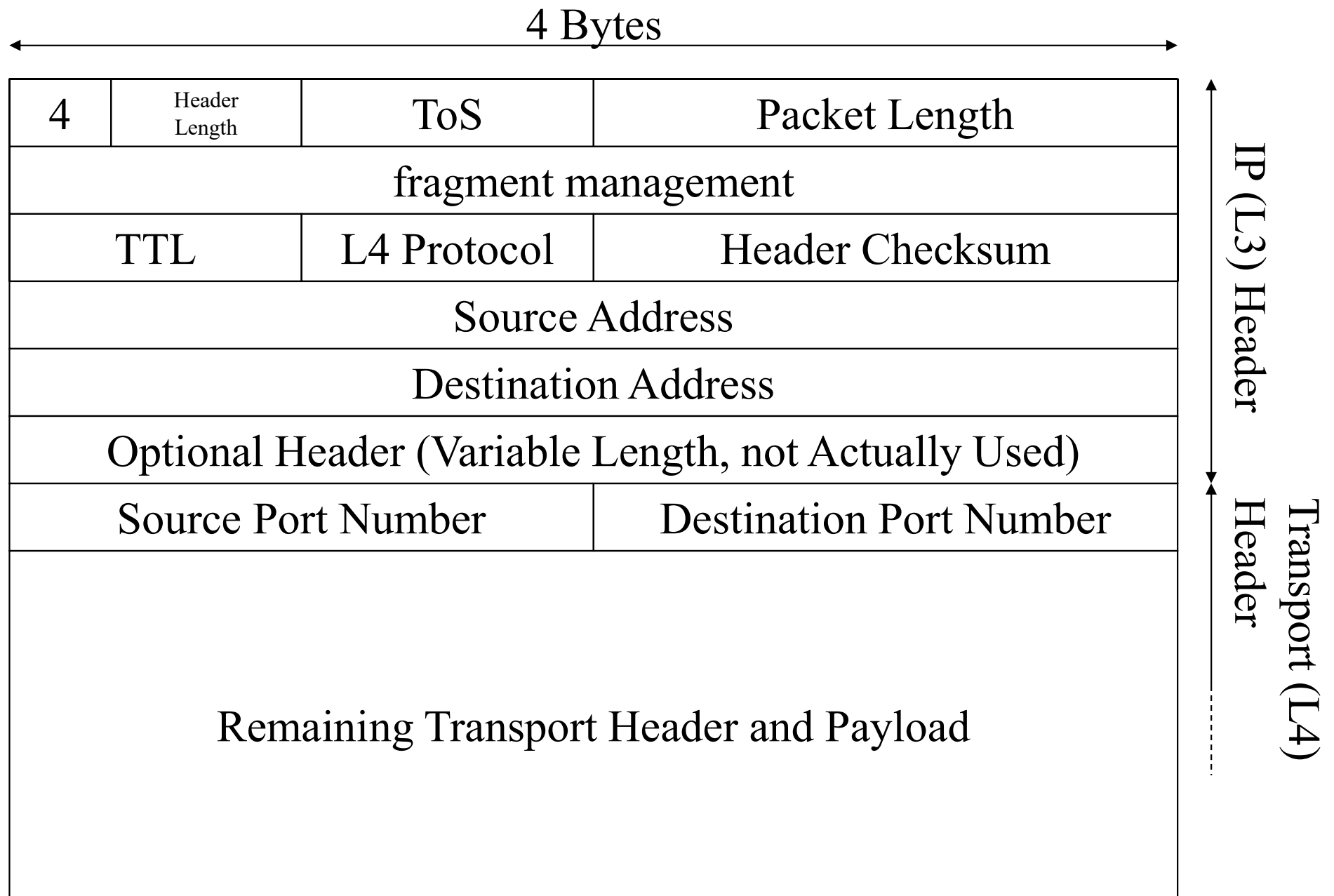
- with IPv4, only 4G hosts may exist
- with IPv6, each of 4G people may have 4G hosts
- when will IPv6 be available?
  - should be before IPv4 address exhausted
    - when will it be exhausted?
      - IPv4 address space was not used efficiently
      - as IPv4 address becomes scarce, efficiency improves
      - some has been saying “will expire in 10 years” since 20 years ago

# IPv4 (Internet Protocol Version 4, rfc791)

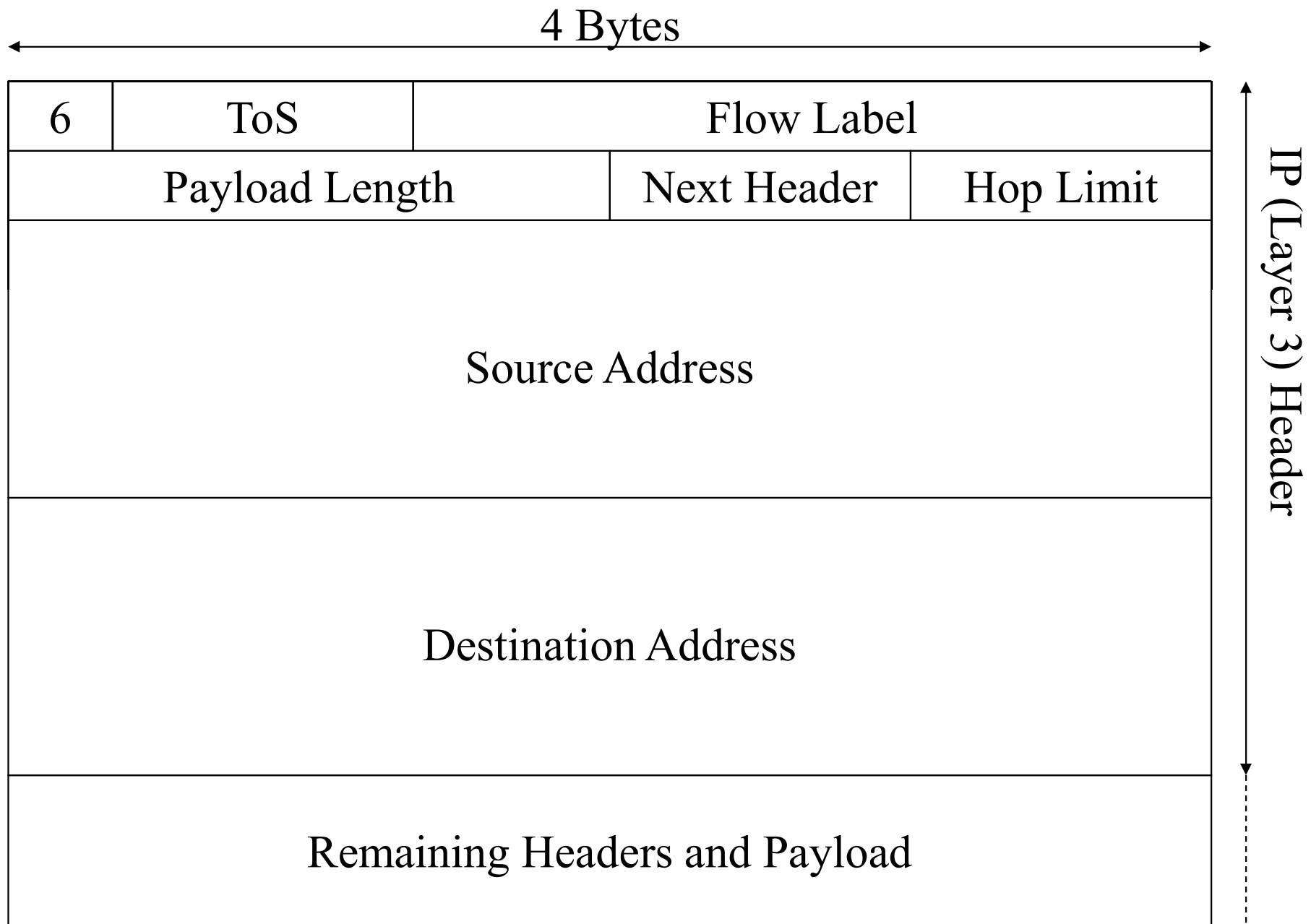
- do almost nothing in the network
  - deliver packets to their destinations
  - fragmentation
  - maintain TTL (time to live)

# IPv6 (Internet Protocol Version 6, rfc2460)

- do almost nothing in the network
  - deliver packets to their destinations
  - **no fragmentation** in the network
  - maintain TTL (time to live) (renamed to be Hop Limit)
  - (**IP option**)
  - (QoS guarantee)



Format of IPv4 Packets (rfc791)



IPv6 Packet Format



# Header Fields Proper to IPv6

- Flow Label
  - identify communication by source address and flow label
    - ease QoS guarantee?
- Next Header
  - equivalent to IPv4 option + (L4) protocol
  - headers are chained, terminated by transport layer header
    - major design flaw of IPv6

# Reason Why Optional OP Header is not (can not be) Used

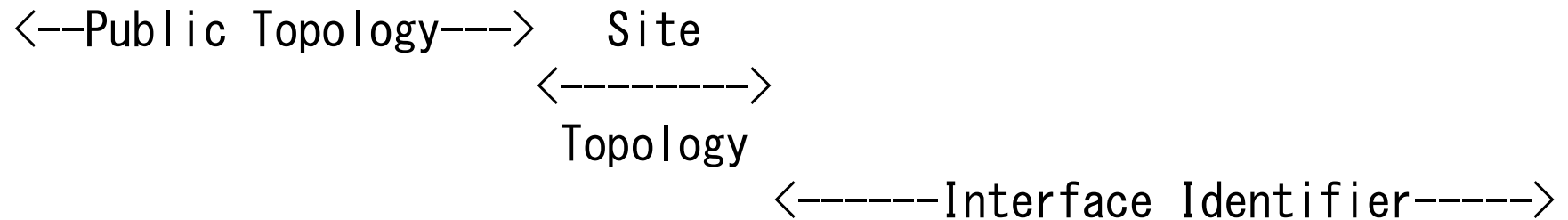
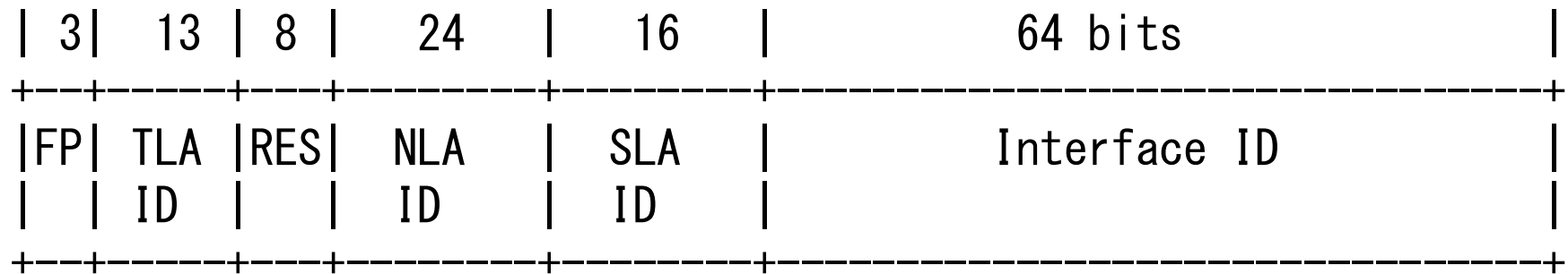
- Options Processed by Routers
  - Router Processing become Complex
    - routers become slower
    - routers may crash
  - Is processing by routers necessary at all?
    - according to the end to end principle, options are harmful and useless
- Options not Processed by Routers
  - Options at or above the transport layer

# Problems of IPv6 Next Headers

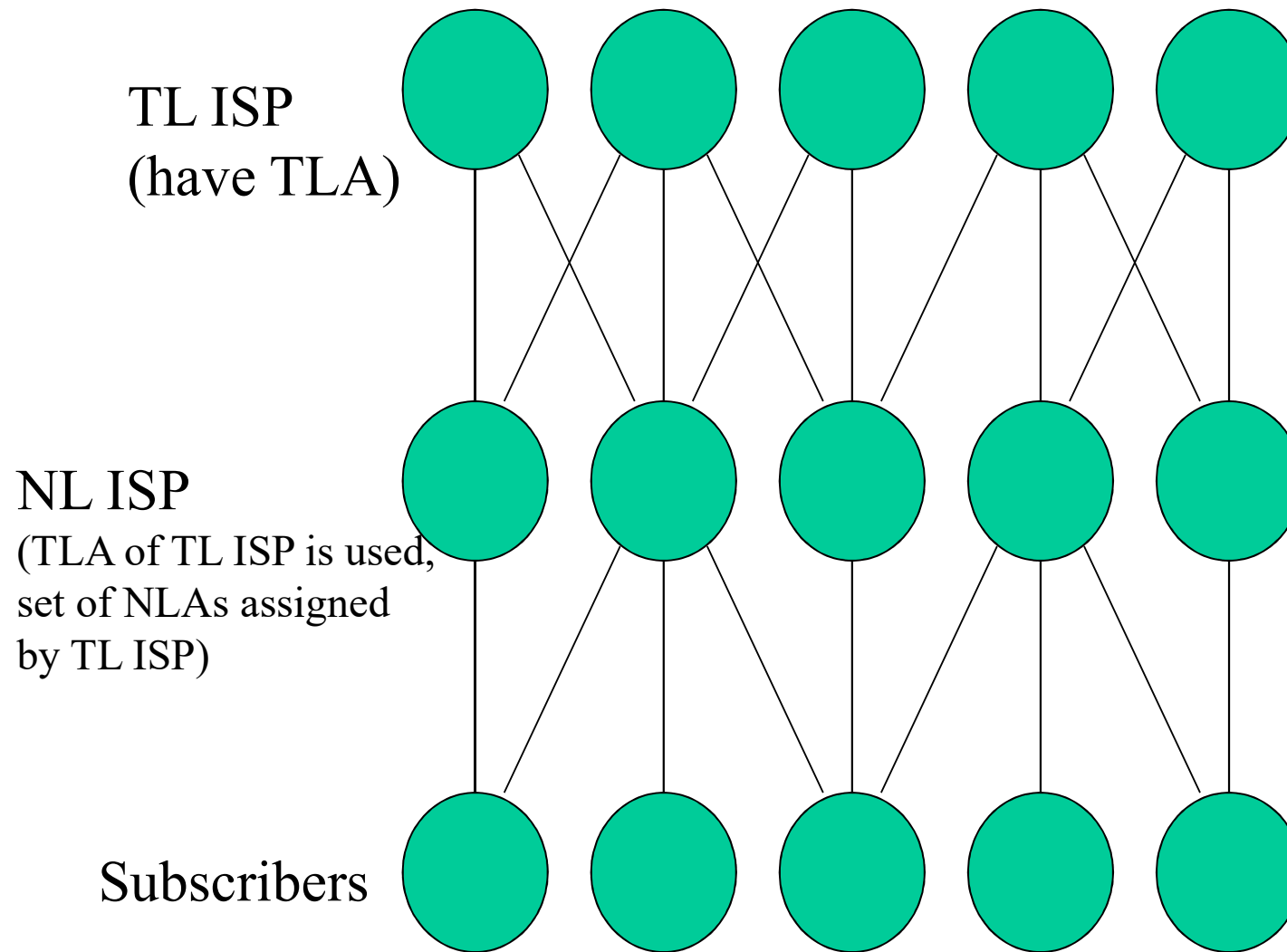
- no limit on the number or length of headers
  - 1280B accepted by any IPv6 capable host
    - IPv4 header  $\leq 60\text{B}$ , 576B accepted by any
  - router processing slow & complicated
    - QoS guarantee by port numbers practically impossible
  - ICMPv6 error may not contain port numbers
    - ICMPv6 can only be 1280B long
  - there is no guaranteed minimum payload length
    - headers can be 1280B long or even longer

# Initial Proposal (rfc2374) of IPv6 Address Structure

- have strong hierarchy
- two layers at ISP level
  - TLA (Top Level Aggregator)
  - NLA (Next Level Aggregator)
- Subscribers can have 65536 links (subnets)
  - SLA (Subscriber Level Aggregator)
- 64 bit Interface ID within each link



Structure of IPv6 address



Typical Scenario of IPv6 ISPs with Multihoming

# Why IPv6 is (not) better than IPv4?

- no fragmentation
- PMTUD
- ND (Neighbor Discovery)
  - configurationless
  - automatic renumbering
- QoS guarantee
- better security
- better mobility

# Fragmentation of IPv6

- fragmentations by intermediate routers are prohibited
  - always results in ICMP error
  - source hosts may fragment with 32bit ID
  - reduce router load?
- to reduce needs for fragmentations
  - minimum MTU of links is 1280B
    - IP over IPsec over IPsec over Ethernet (1500B)
- Discover minimum MTU of a path by PMTUD

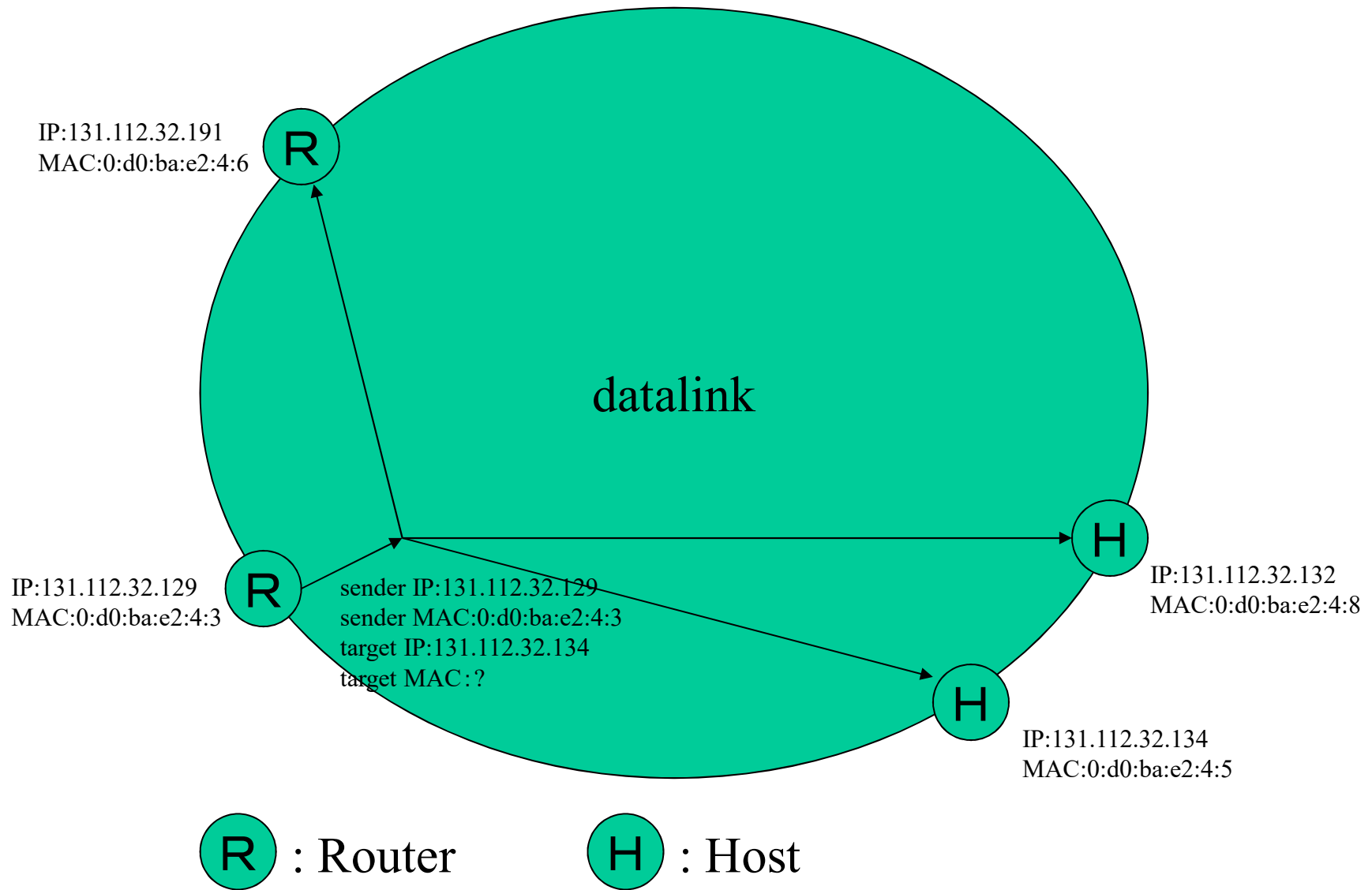


# Path MTU Discovery

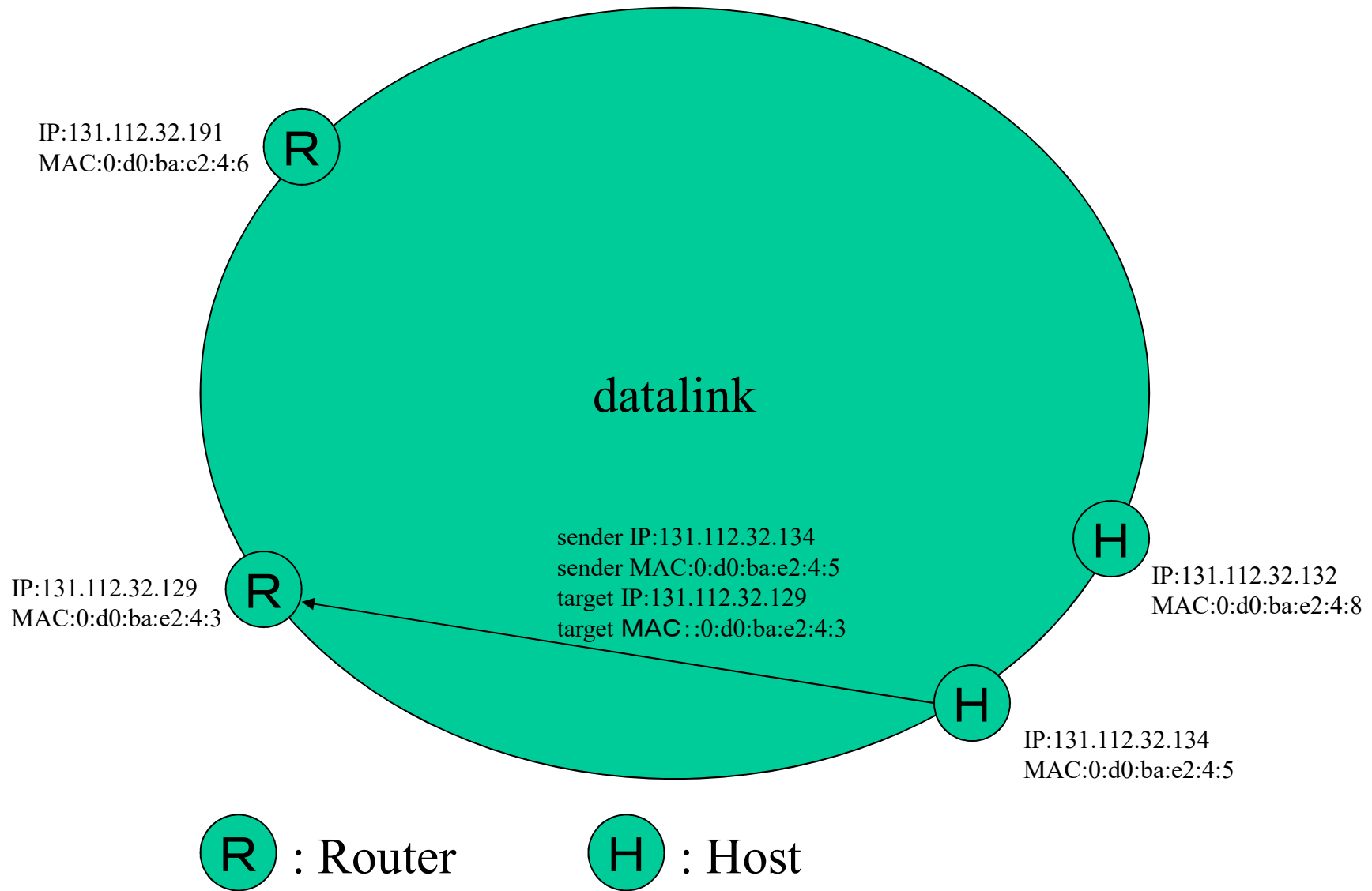
- for each **transport (not IP layer) connection**
  - try to send a packet of certain length
    - if no error is detected (**how?**), PMTU is not less than the length
      - try to send a little longer packet, next time
    - if error is detected, PMTU is less than the length
      - reduce packet size, appropriately (**how much?**)
  - try periodically, as path dynamically changes
    - periodic burden on routers
  - without connection, not usable
    - not applicable for DNS, multicast, etc.

# ARP (Address Resolution Protocol, 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



ARP Query



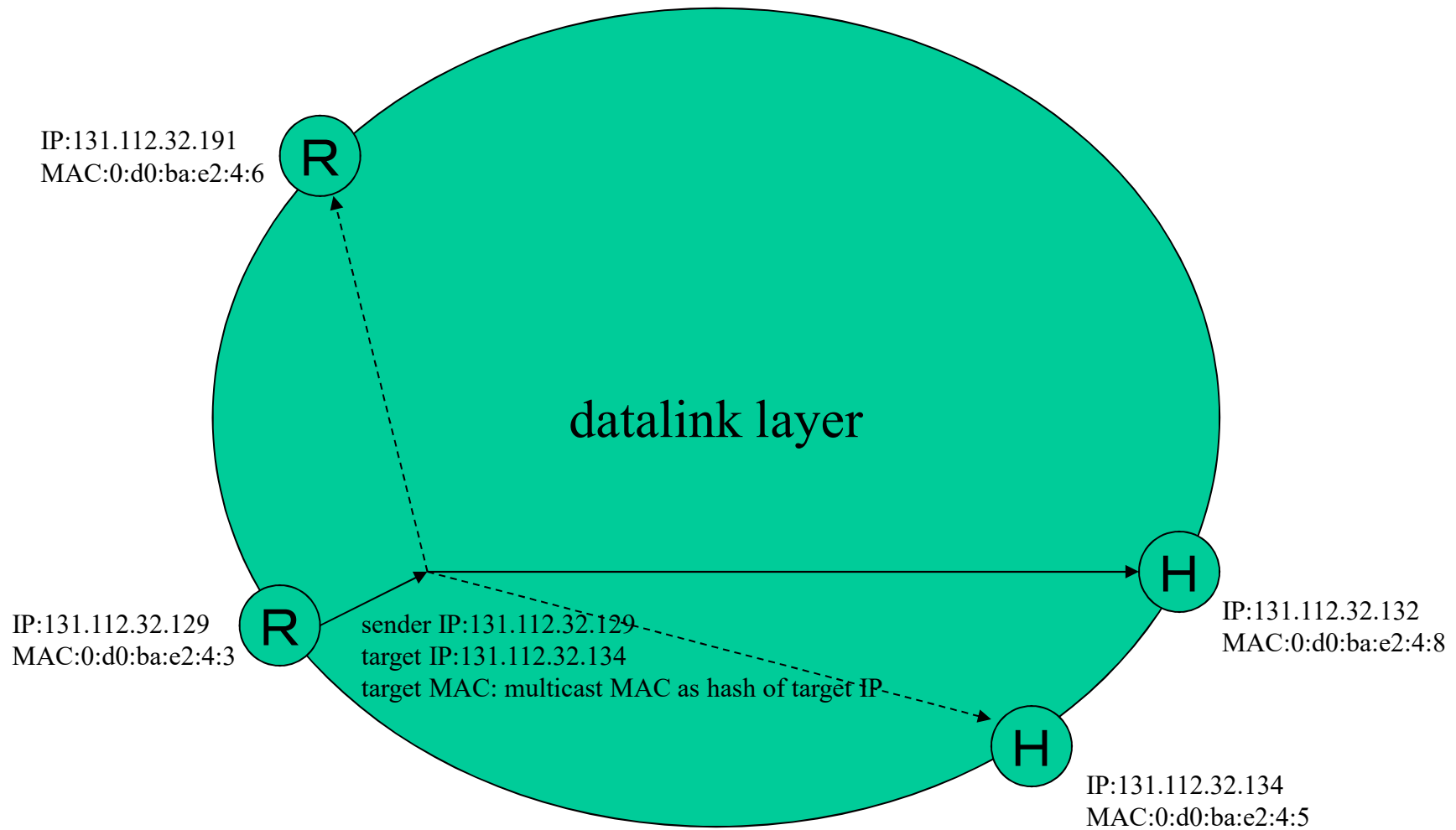
ARP Reply

# Problems (?) of ARP

- if ARP is used, original IP packet causing ARP will be discarded (rfc826)?
  - may be retained
- depends on broadcast
  - cannot be used over large datalink
    - large datalink!?
      - violation of CATENET model was considered necessary for efficient IP over ATM (IP over large cloud)

# Neighbor Discovery (rfc2461)

- use lots of link multicast
  - initial proposal send IP packets with unresolved MAC to pseudo random multicast MAC
- routers are intelligent, hosts are not
  - hosts are automatically configured
  - renumbering of hosts easy?
  - intelligent intermediate entities (routers) is against the e2e principle
    - hosts must be intelligent



**R** : router      **H** : host

packet sent to pseudo random MAC (initial proposal of ND)

# Features Expected for Neighbor Discovery

- general purpose adaptation layer between IP and any datalink layer
  - IP is general purpose, adaptation should depends on datalink layer
- SLAAC (stateless address auto configuration)
- automatic renumbering
  - necessary to change ISP easily even with CIDR



# SLAAC (Stateless Address Auto Configuration)?

- imitate that of Macintosh (apple)
  - suitable for closed LAN at Mac era
  - reason of lengthy (16B) address?
- not practical
  - no security (security needs configuration)
    - not a problem for home LAN, but
  - no registration to DNS
    - not usable as servers

# SLAAC is Full of State with Worst Possible Manner

- stateless means no stateful central server
  - still, address configuration state is state
    - RA (router advertisement) contain (64bit) prefix
      - hosts generate lower (64 (was 48)) bits from MAC address or randomly
      - must use DAD (duplicated address detection)
- state should be centrally managed
  - should discard ND replaced by DHCP(v6)
  - DHCP server can take care of DNS registration

# Automatic Renumbering

- CIDR is assumed by IPv6
  - ISP change means address change (renumbering)
    - unless one have unaggregated address
- upon renumbering, content of DNS should be automatically updated
  - almost impossible as DNS need DNS server addresses as raw addresses
- abandoned

# QoS Guarantee

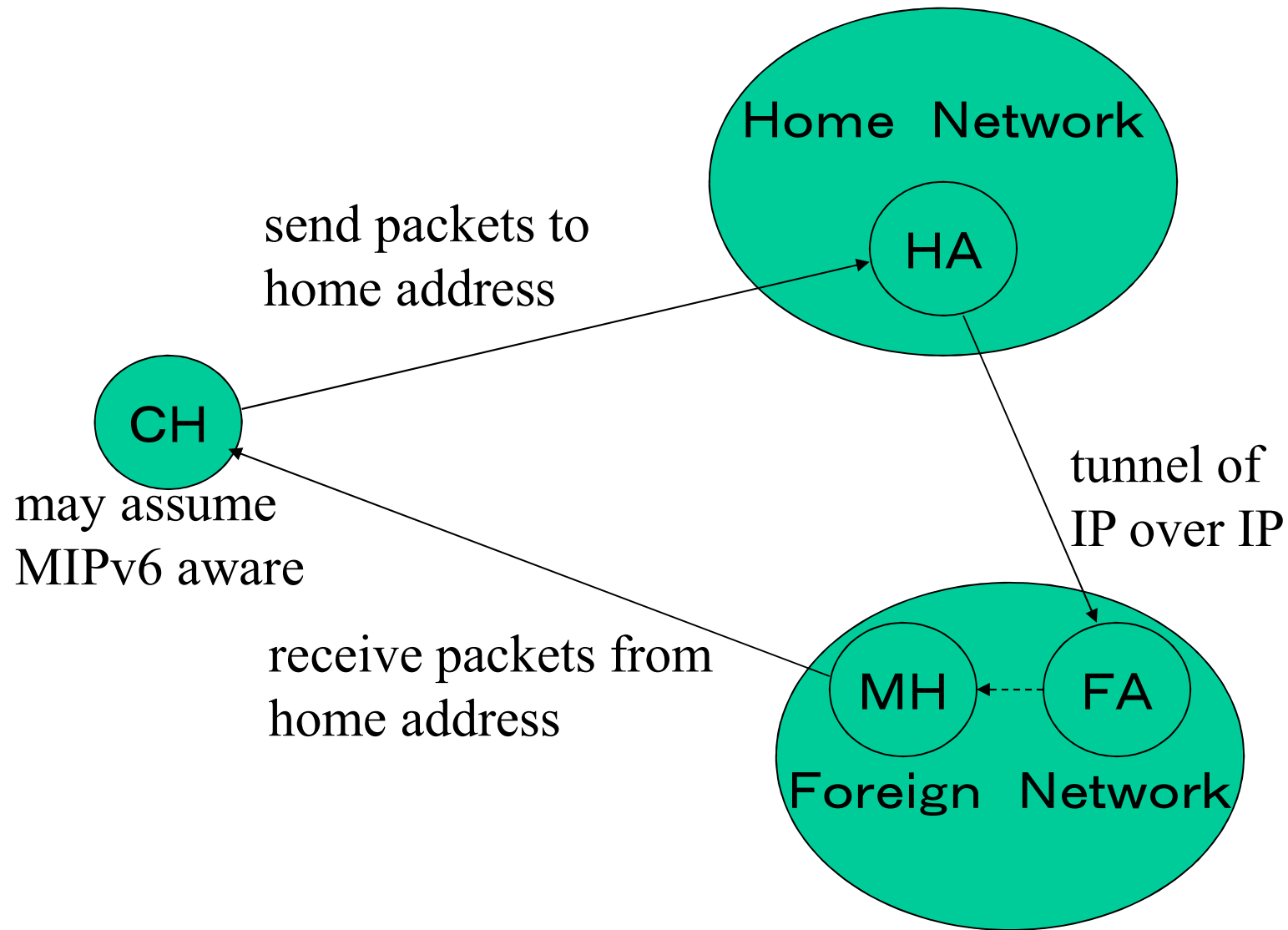
- QoS guarantee by flow labels
- is signaling protocol (RSVP) available?
- how flow labels should be used?
- isn't IPv4 better for QoS guarantee?

# Better Security

- IPv6 mandates IPsec
  - should be able to disable DoS with IPsec authentication
- IPsec needs cryptographic keys configured
  - not useful for packets from unknown origins

# Better Mobility

- IPv6 is designed with mobility in mind
  - all the hosts should be mobility aware
    - with IPv4, some hosts are mobility unaware
  - if CH (corresponding host) is mobility aware
    - triangle elimination may be possible



triangle exchanges of packets between CH(corresponding host), HA (home agent), FA (foreign agent) and MH (mobile host)

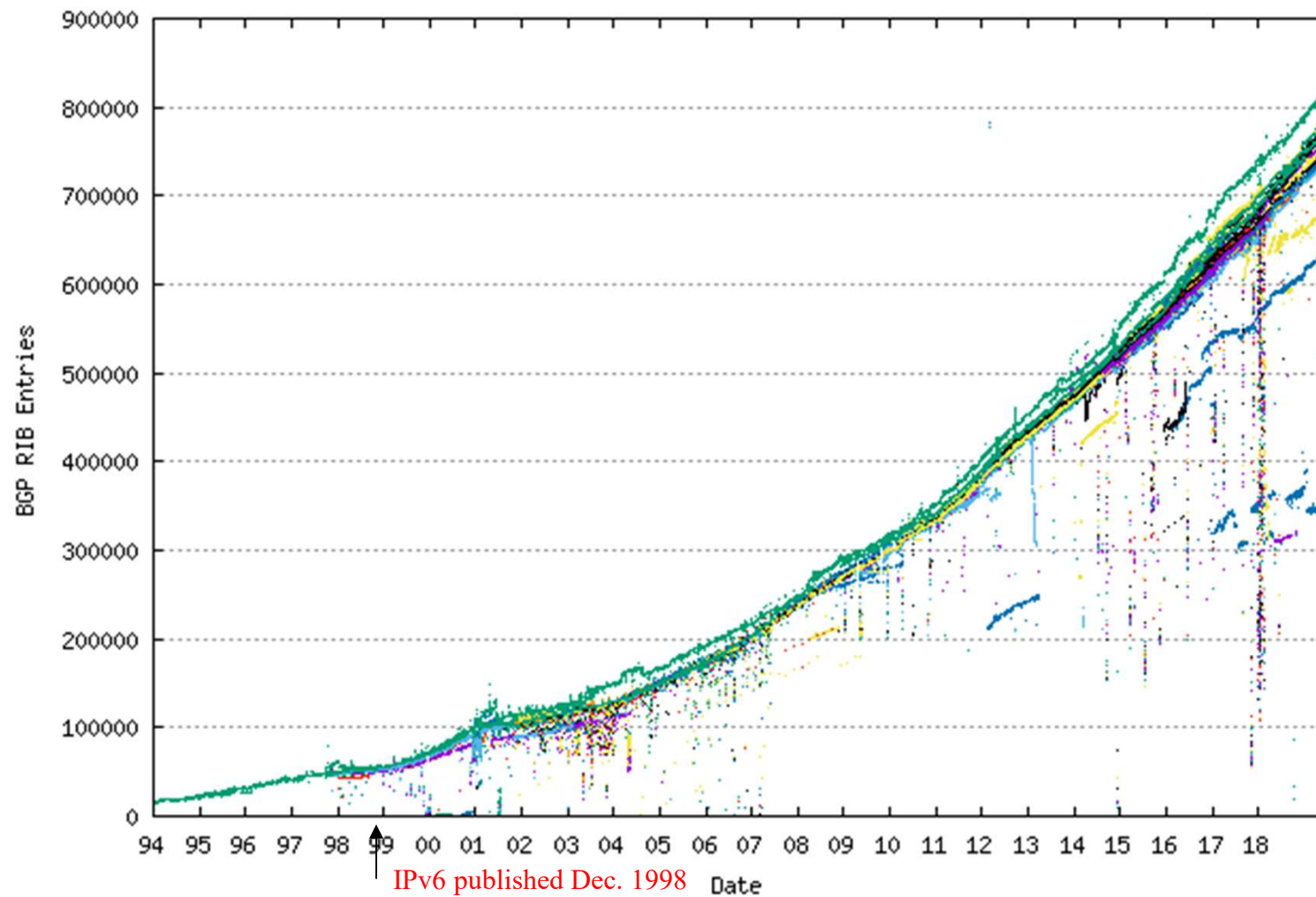
# True Purpose of IPv6

- larger address space?
- smaller routing table size
  - if multihoming is properly treated



# IPv4 Routing Table Size

<http://bgp.potaroo.net/>



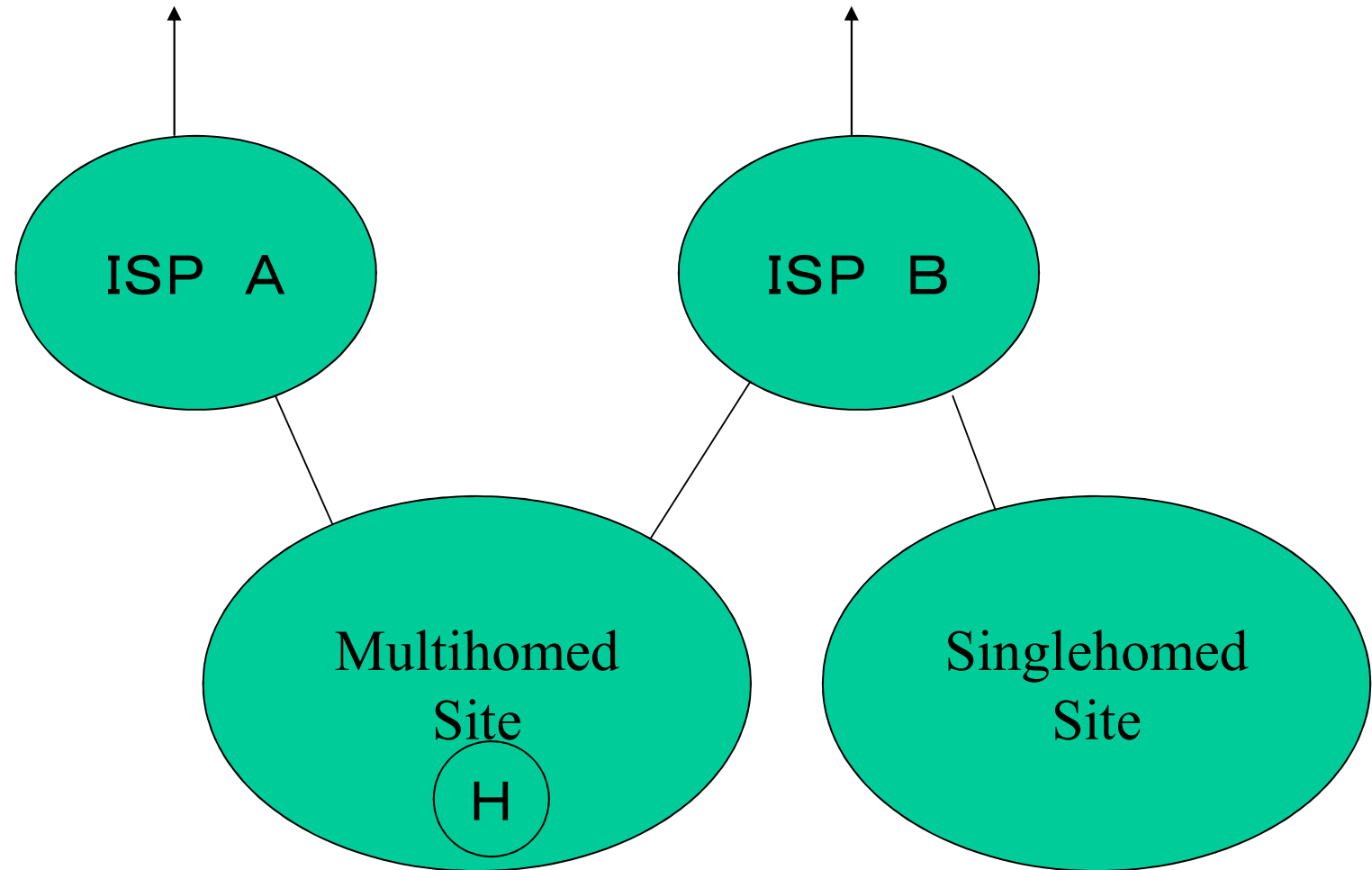
# Cases When Route Aggregation Impossible

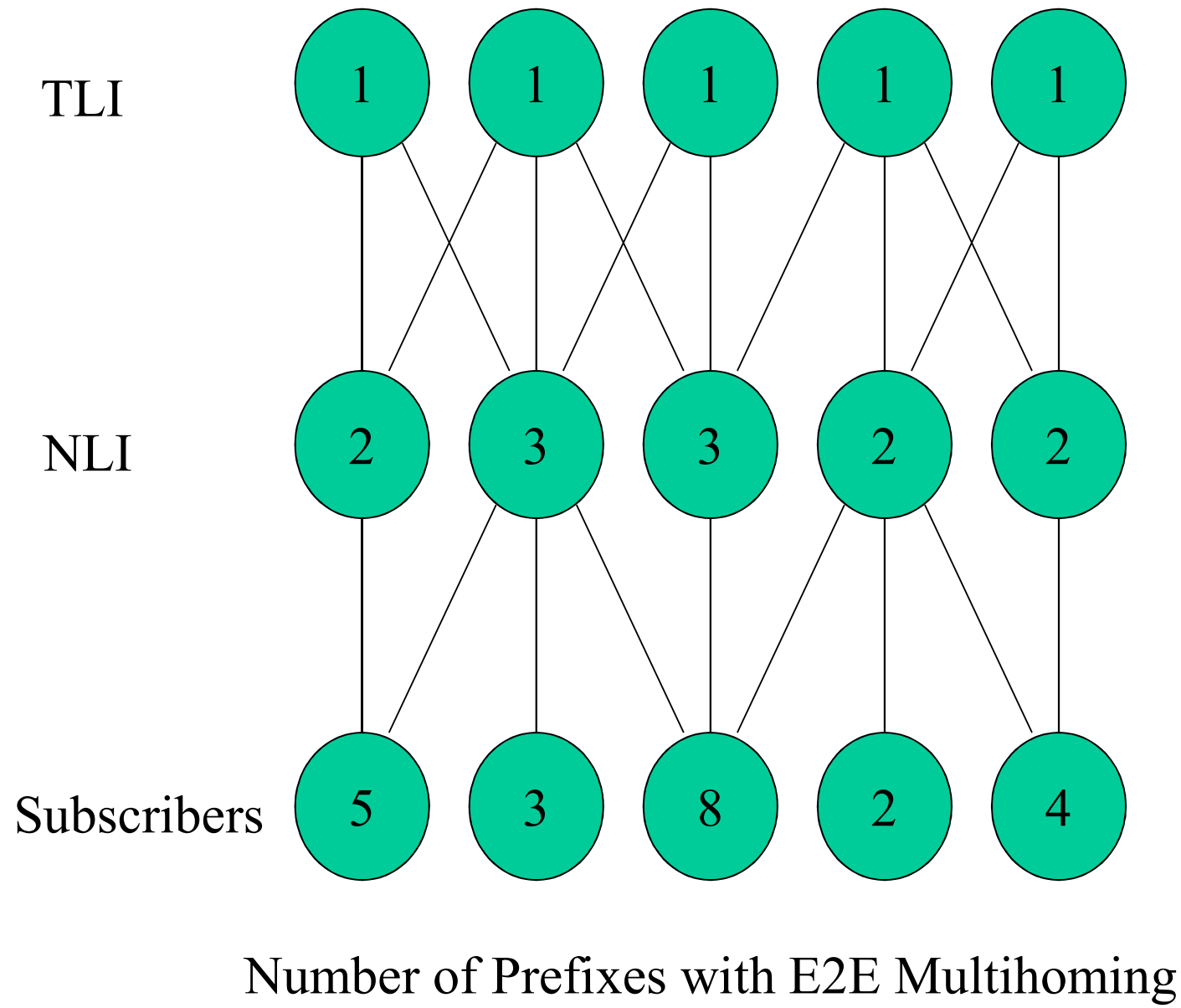
- aggregation possible, if route is shared by addresses sharing a pattern
- route not by destination address only
  - QoS routing depends on required QoS
- destination address not designate location
  - multicast address designate set of locations
- random IP addresses within a region
  - initial allocations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

# Multihoming

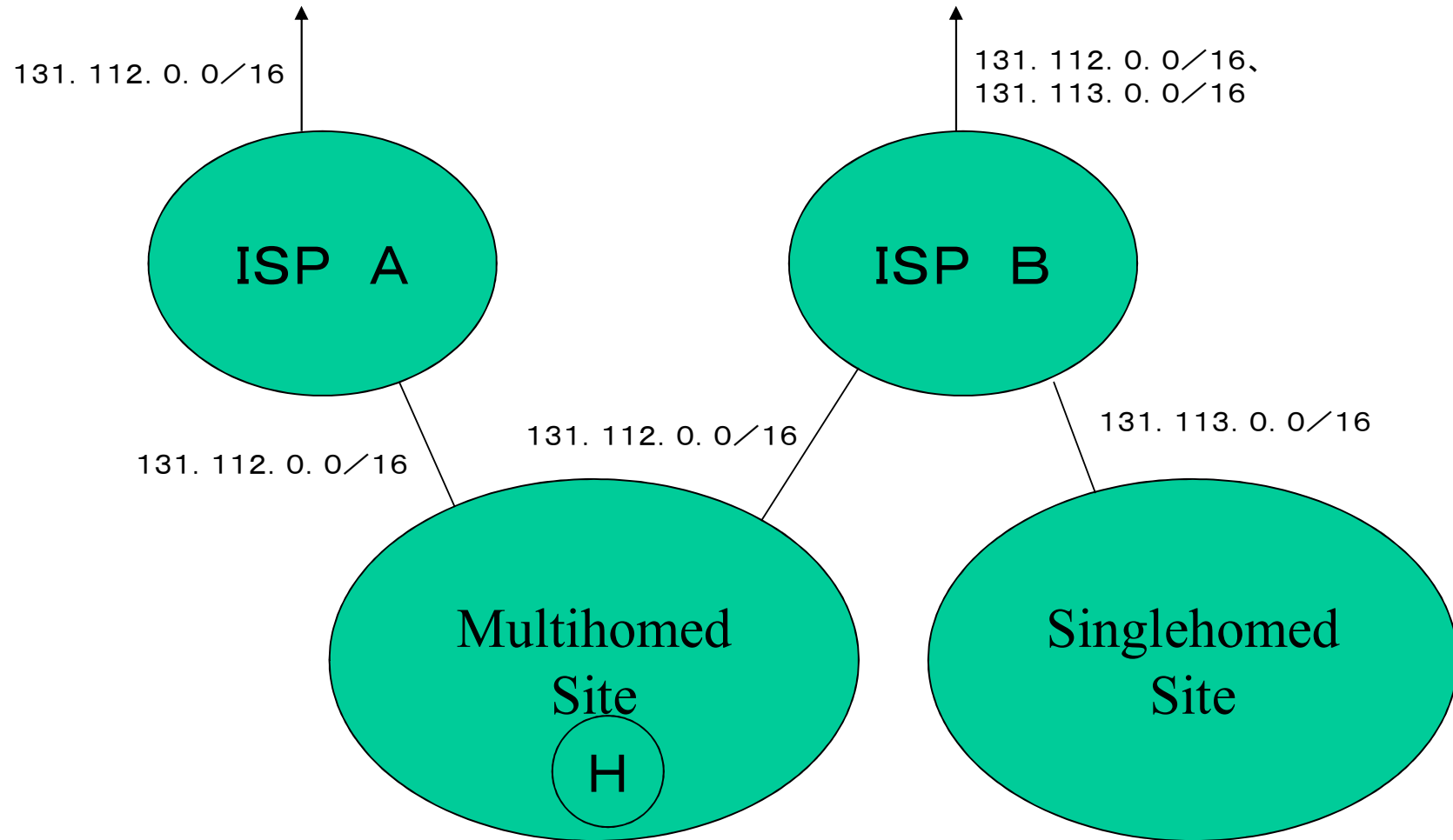
- have multiple upstream ISPs
  - safe even if some ISPs fail
- necessary for reliable service (incl. ISP)
  - IPv6 NLISP want to have multiple TLISPs
- multihoming by routing assumes single address with single TLA regardless of TLISP changes

to rest of the Internet



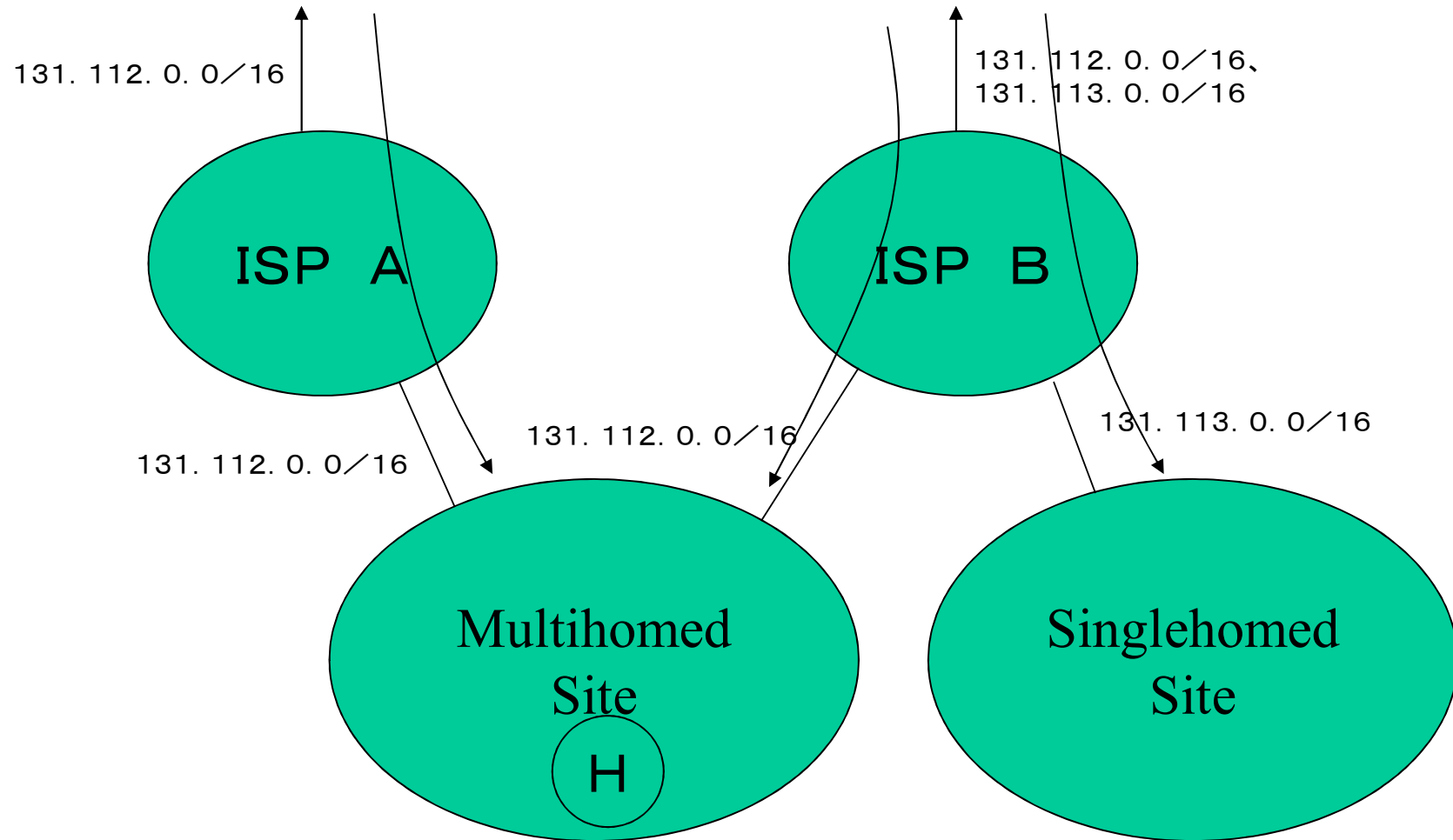


to rest of the Internet



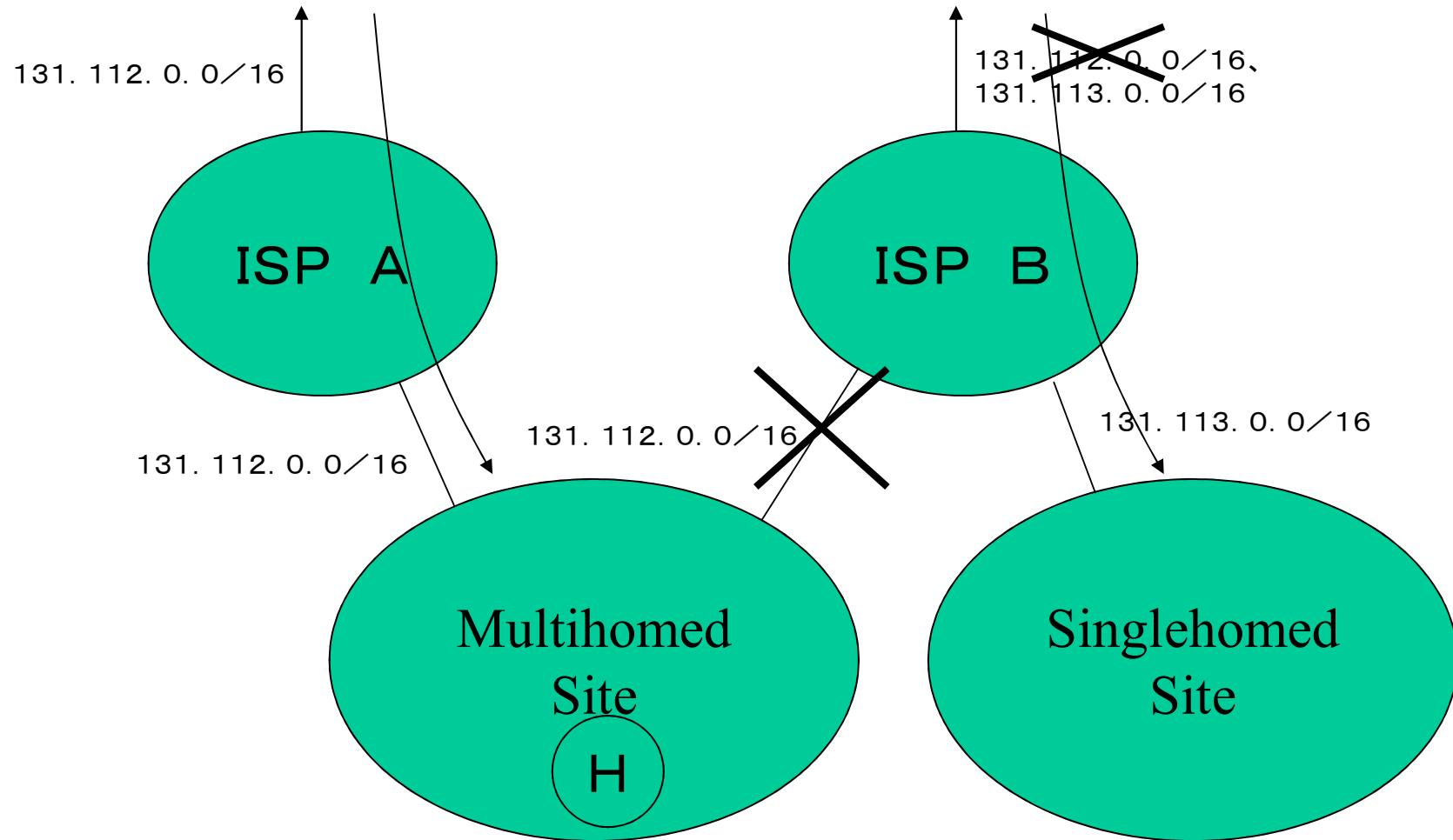
multihoming by routing

to rest of the Internet



multihoming by routing

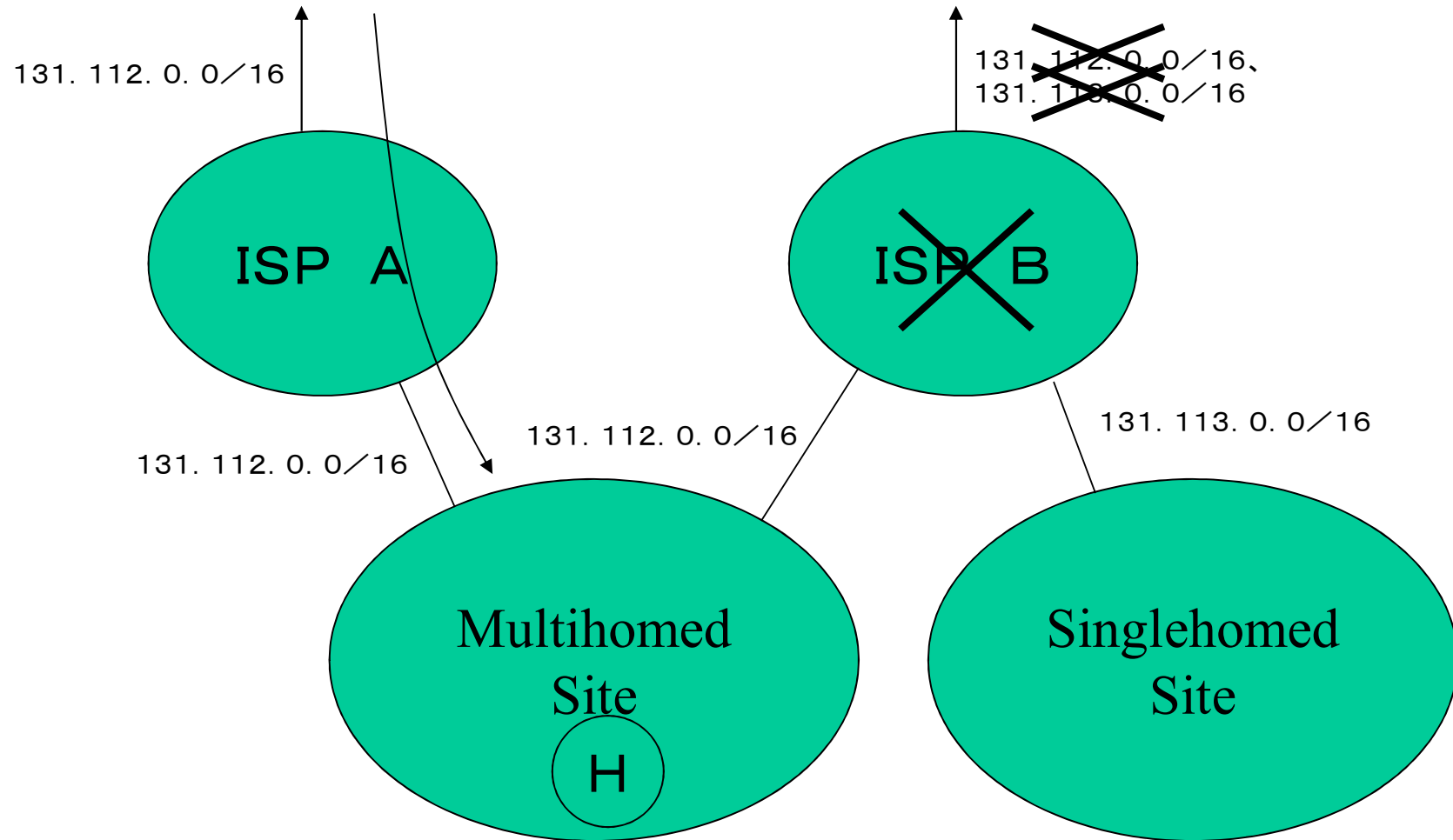
to rest of the Internet



multihoming by routing

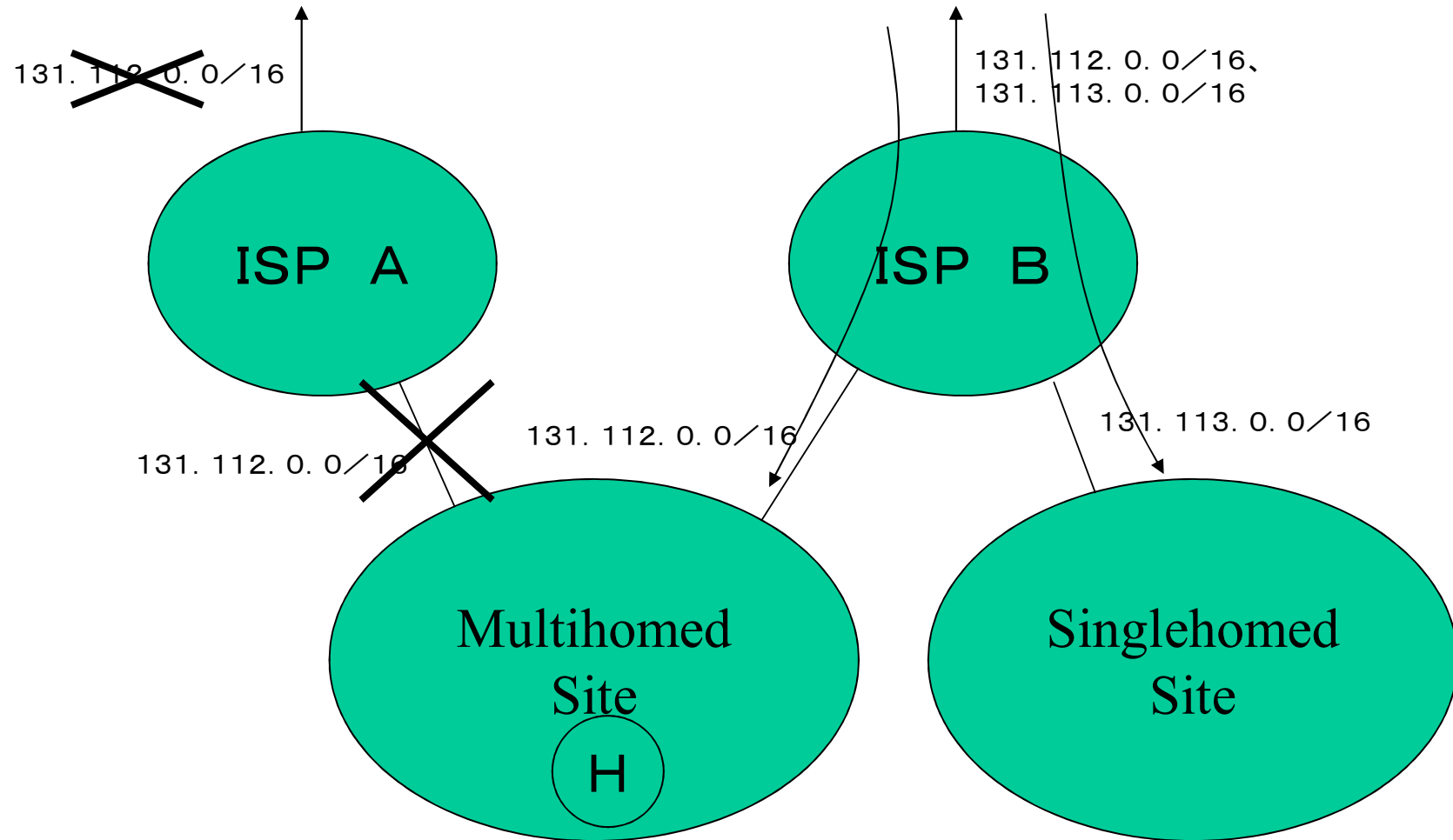


to rest of the Internet



multihoming by routing

to rest of the Internet

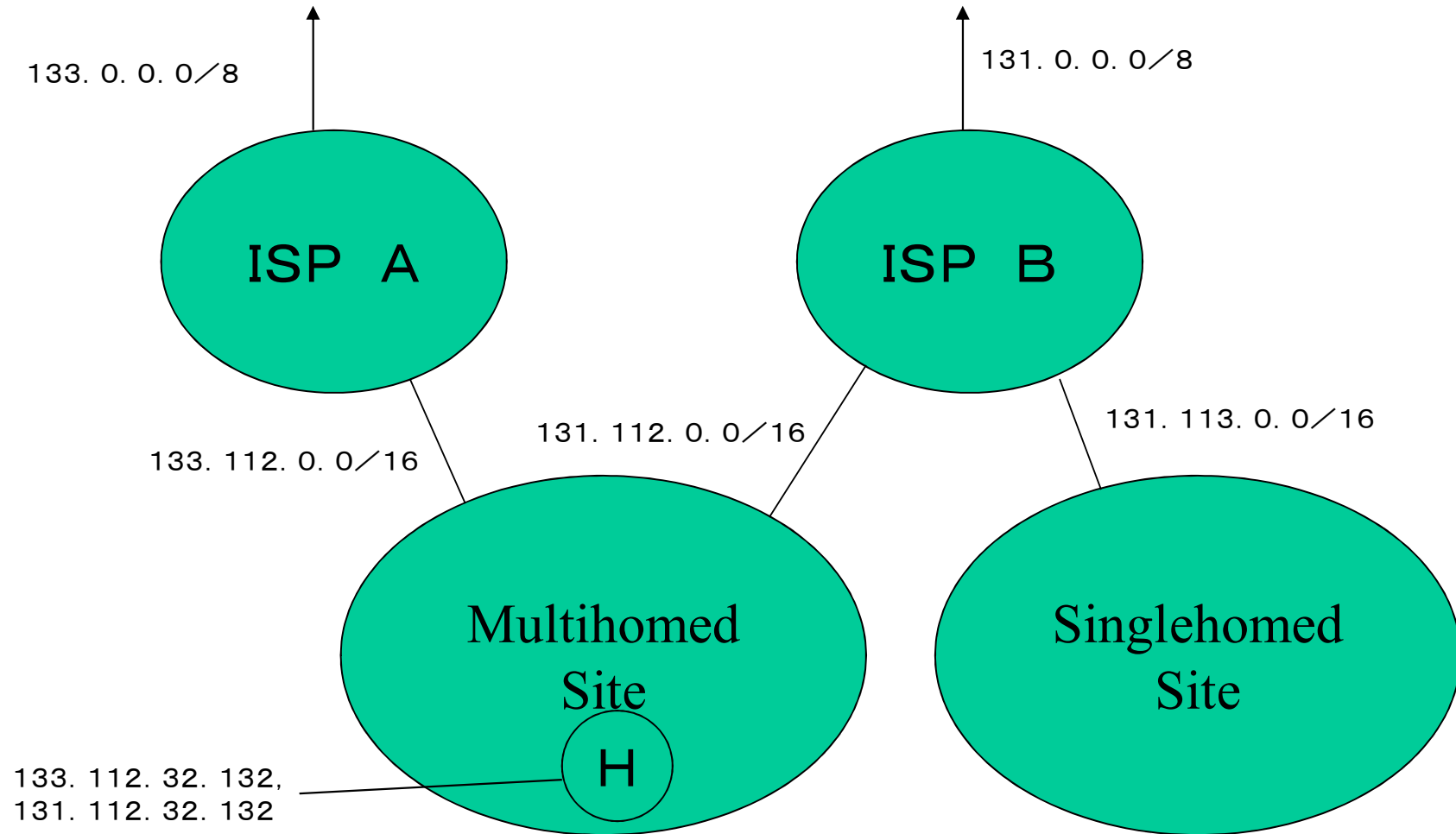


multihoming by routing

# End to End Multihoming

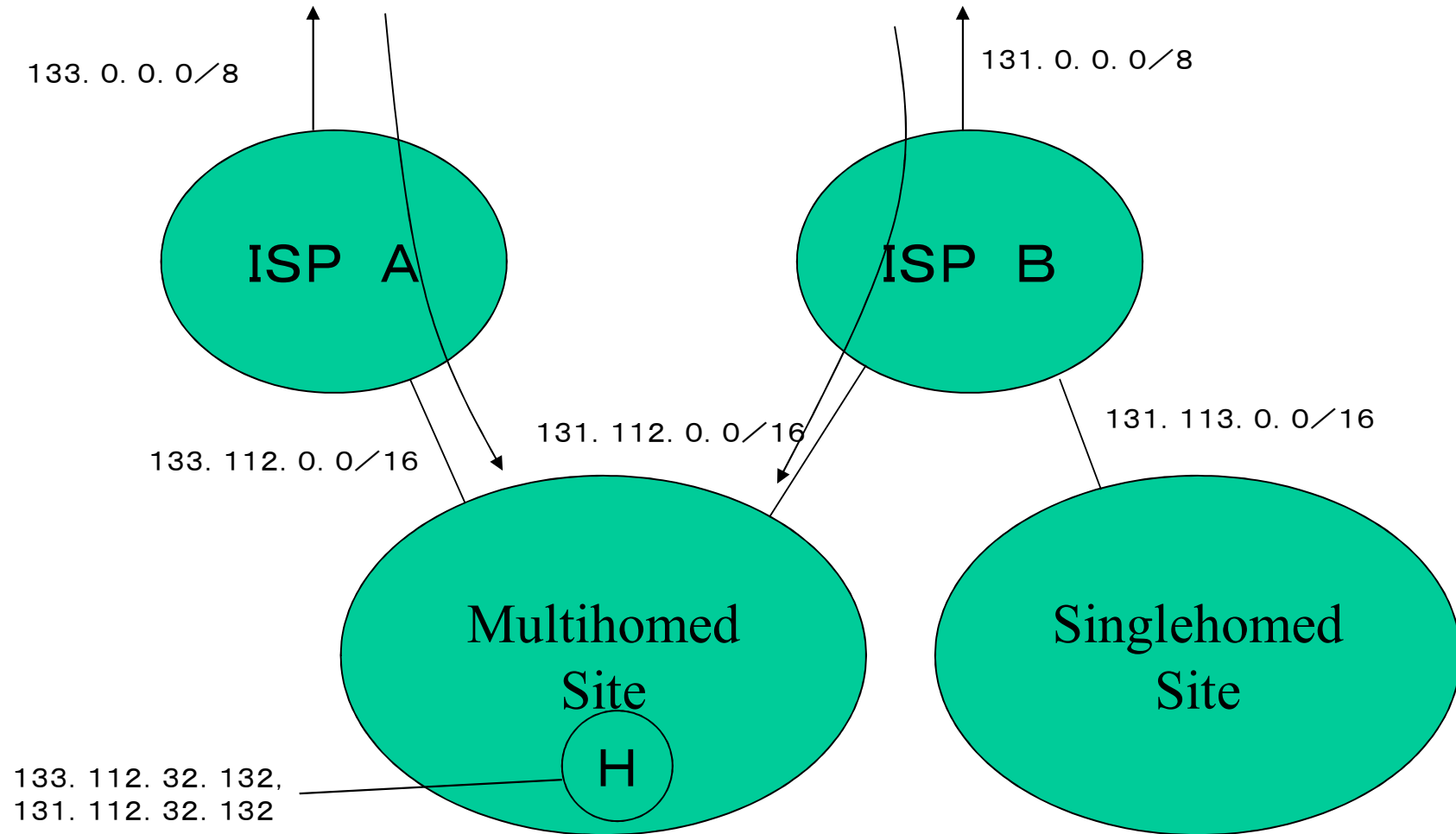
- a host has multiple IP addresses
- **peer of a host** try to use multiple addresses of the host
  - rough unreachability by global routing table
  - if some address works, communication starts
  - if timeout occurs, other addresses are tried
- multihoming by routing is not necessary

to rest of the Internet



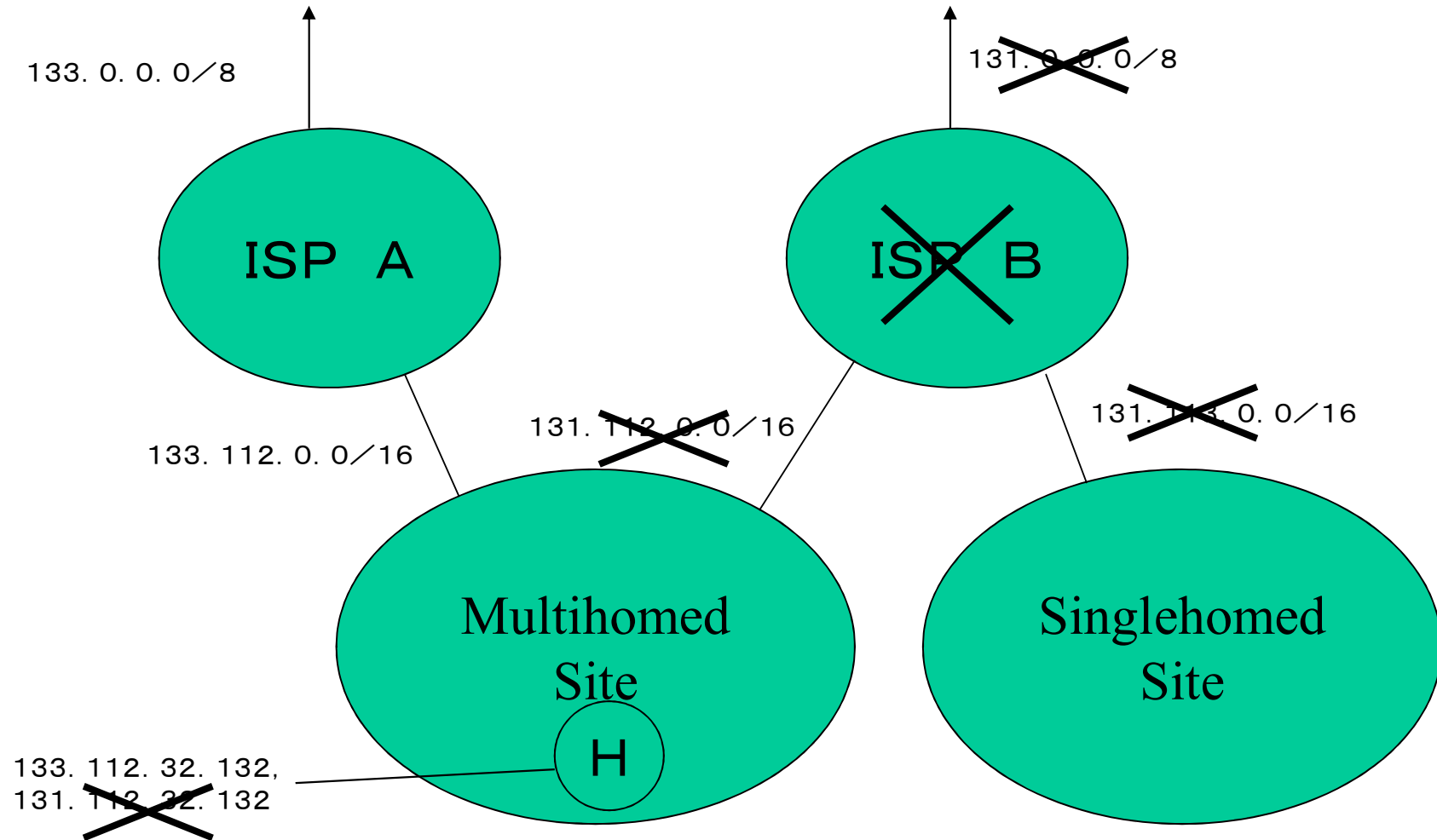
end to end multihoming

to rest of the Internet



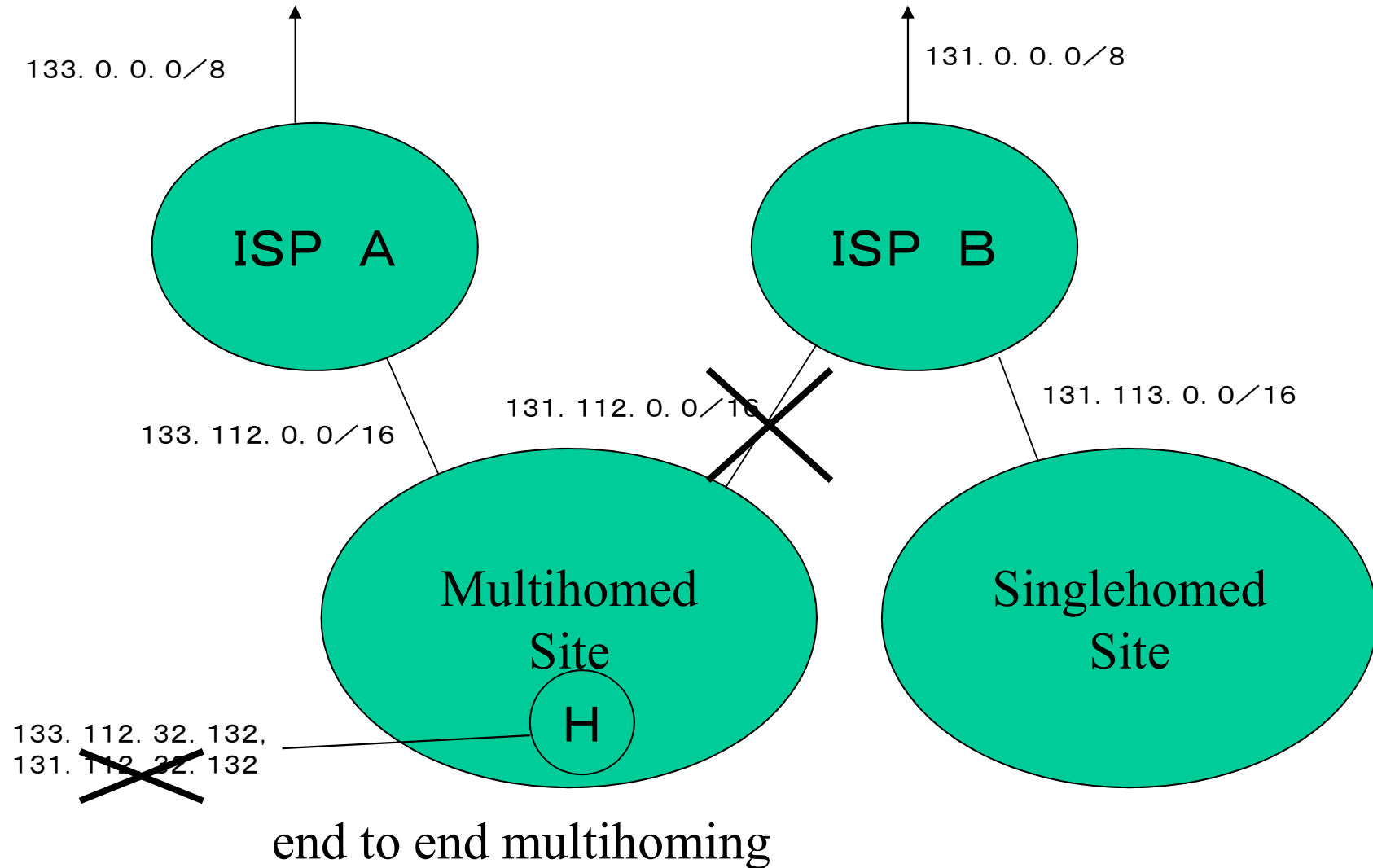
end to end multihoming

to rest of the Internet

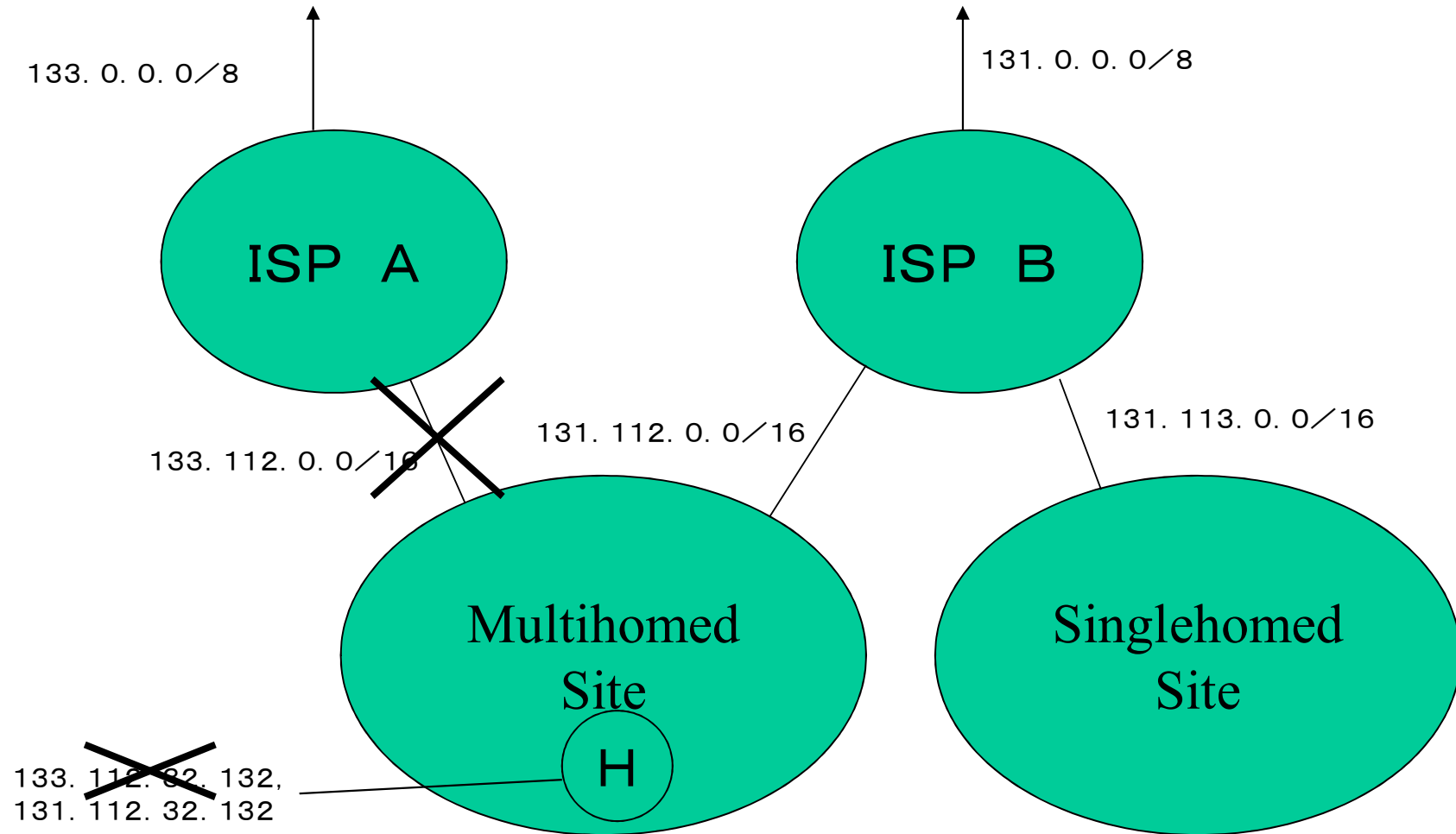


end to end multihoming

to rest of the Internet



to rest of the Internet



end to end multihoming



# ID/Locator Separation

- separate (IPv6) address (16B) into ID (8B) and locator (8B)
  - ID globally identifies a host
    - something like fixed length domain name
  - locator is used for routing
    - with hierarchy of TL, NL, etc.
- not deployed with IPv6
  - 8B address was enough for IPv6

# Why ID/Locato Separation Useful?

- about half amount of data for multiple (many) addresses
  - may be useful for e2e multihoming
- locators may be rewritten en route
  - source locator, rewritten by ISPs, could be reliable
  - rewriting destination locator makes tunneling unnecessary (e.g. for mobile IP)

# Wrap-up

- IPv6 was necessary
  - especially for better multihoming, but...
- IPv6 is not usable
  - broken in several ways
    - some features needs serious redesigning
  - simplifications necessary
- ND should be deprecated

# Inapplicability of Neighbor Discovery over Wireless LAN

Masataka Ohta

Tokyo Institute of Technology

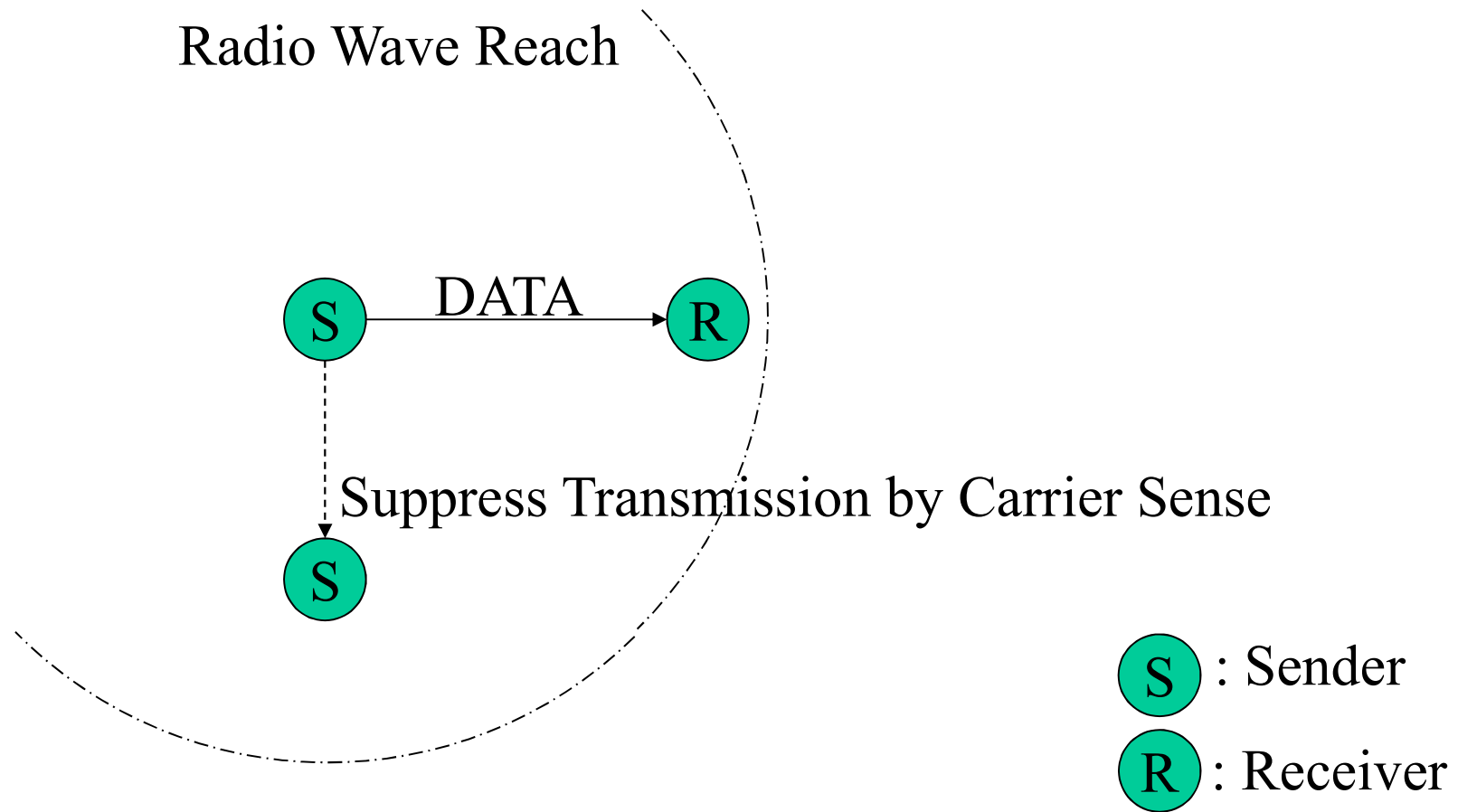
[mohta@necom830.hpcl.titech.ac.jp](mailto:mohta@necom830.hpcl.titech.ac.jp)

# Wireless LAN of IEEE 802.11

- Relies on CSMA/CA
  - because of undetectable collisions
    - ACKs are the MUST for reliable communication
      - or packets are lost upon collisions
    - Unicast packets are ACKed and delivered reliably

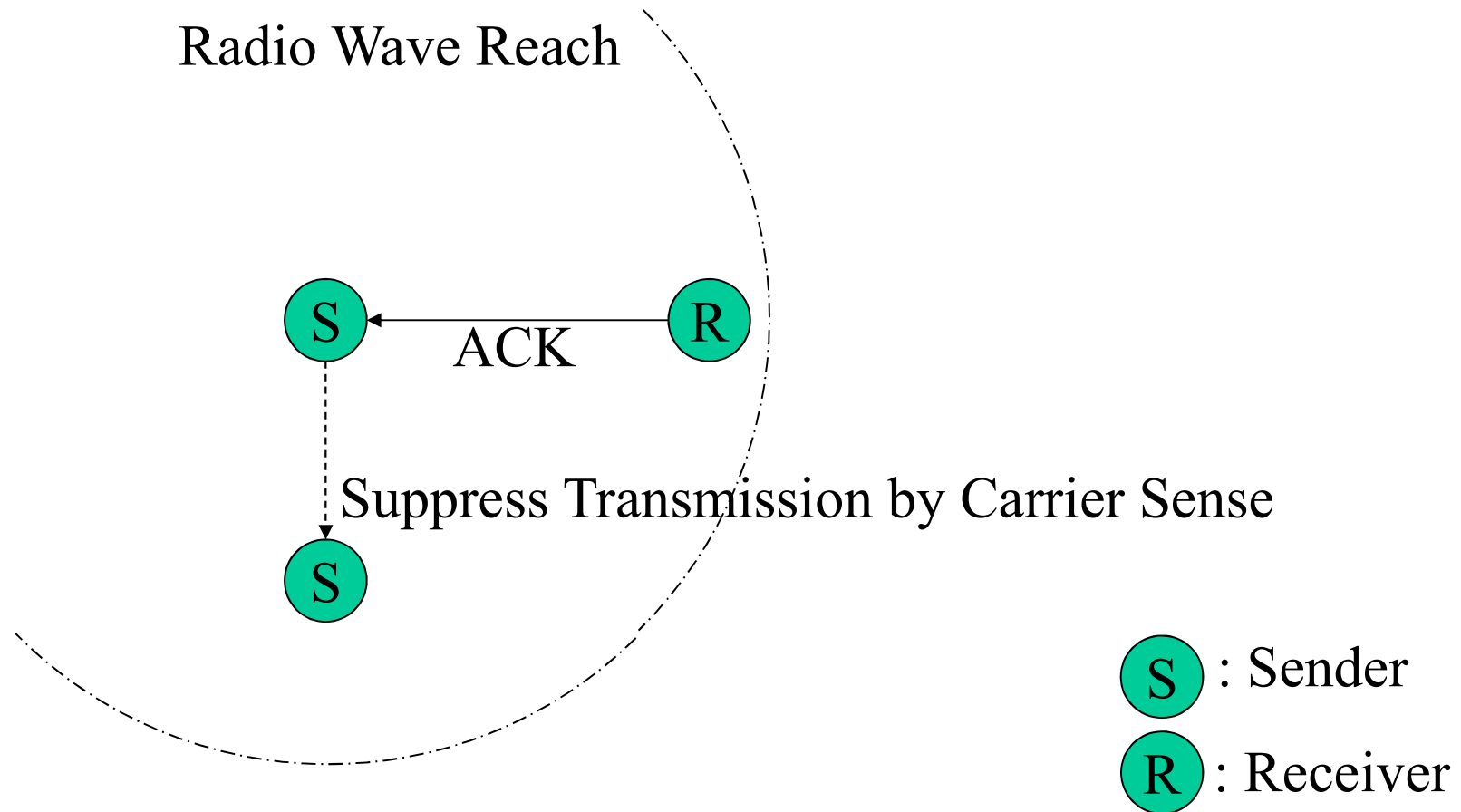
# CSMA/CA and ACK

## Suppression by Carrier Sense



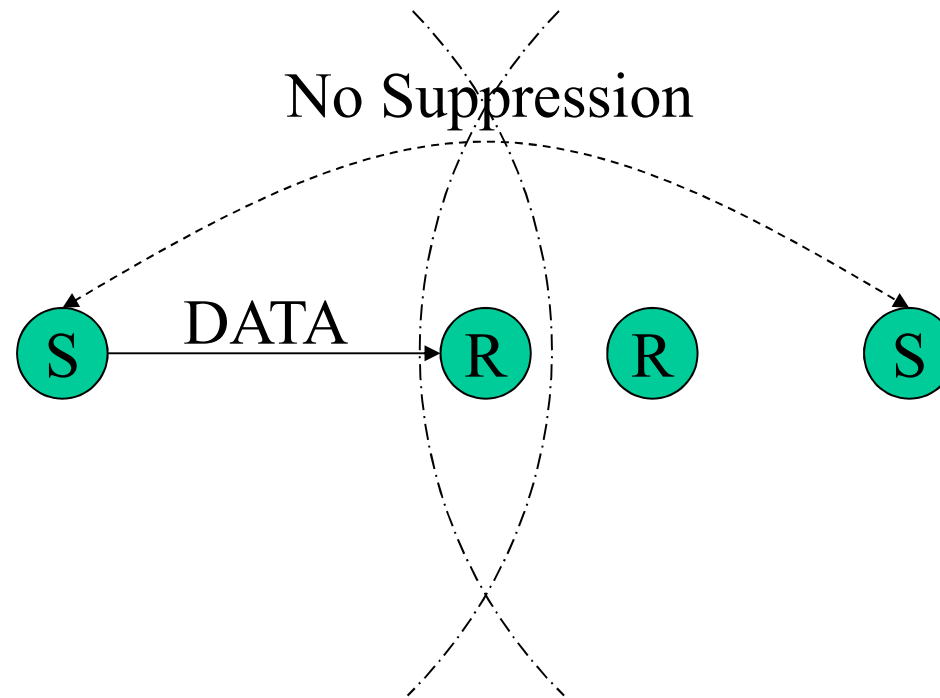
# CSMA/CA and ACK

## Suppression by Carrier Sense



# CSMA/CA and ACK

## Collision by Hidden Terminal



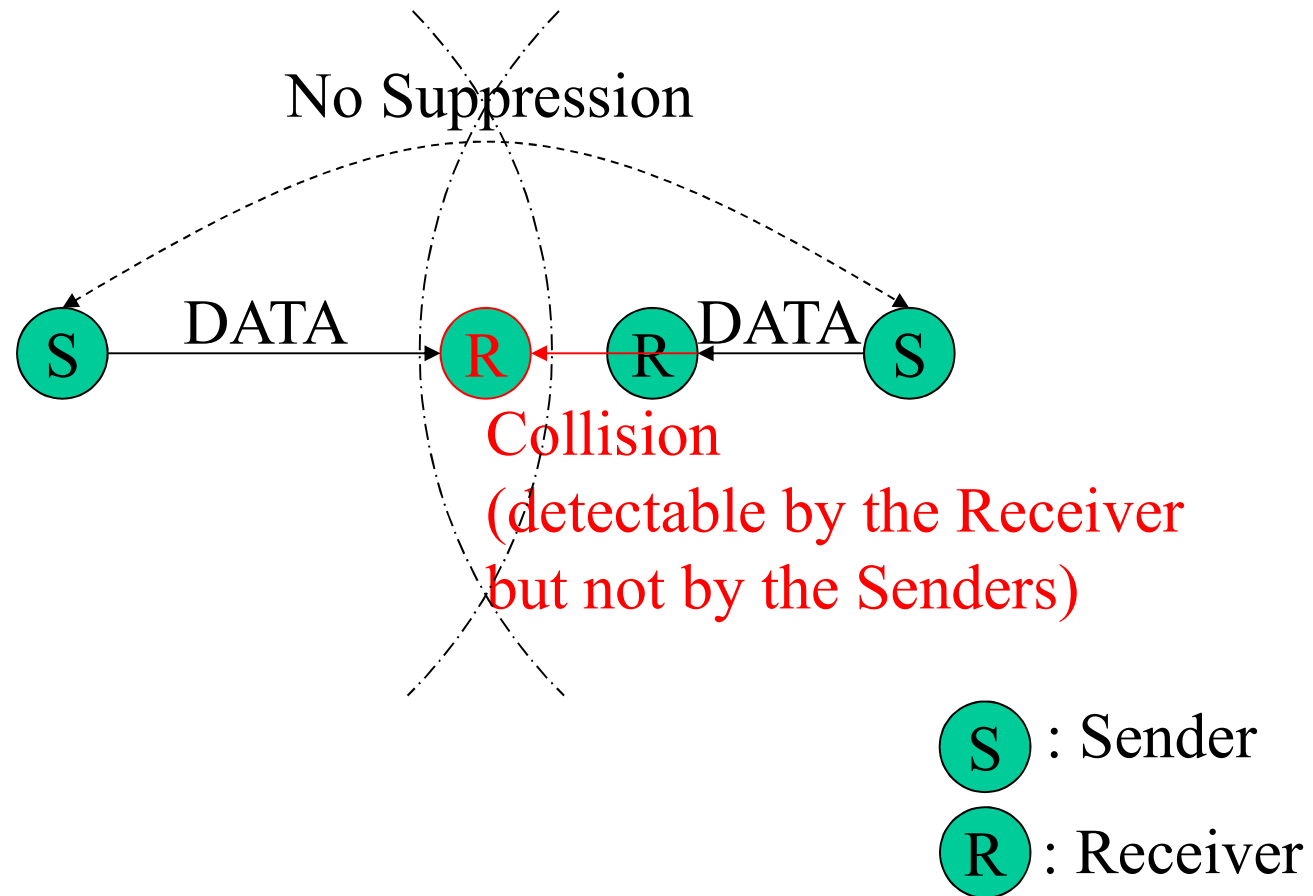
 : Sender

 : Receiver



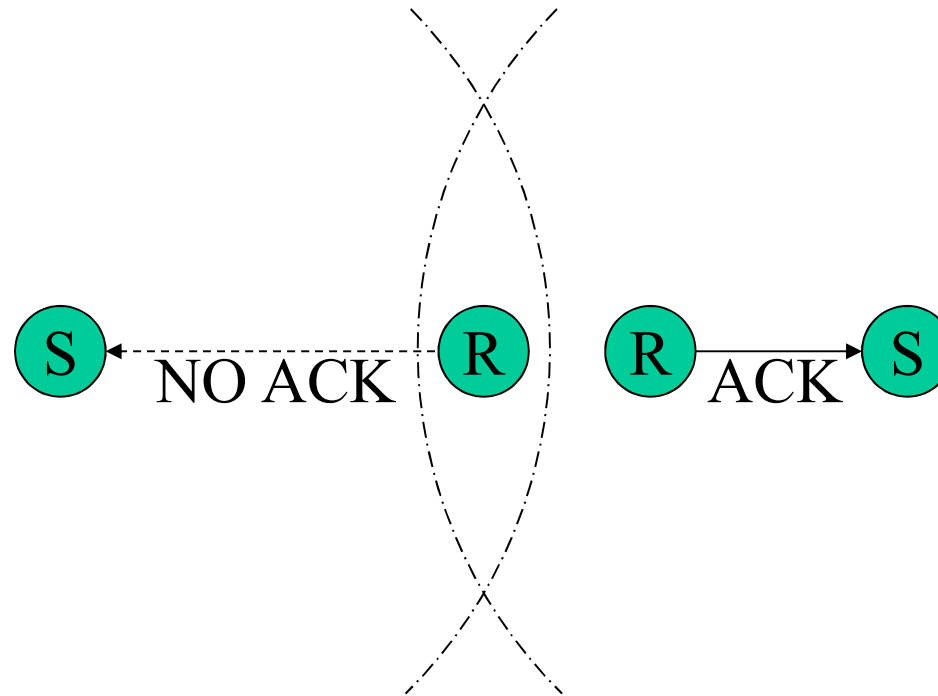
# CSMA/CA and ACK



## Collision by Hidden Terminal



# CSMA/CA and ACK

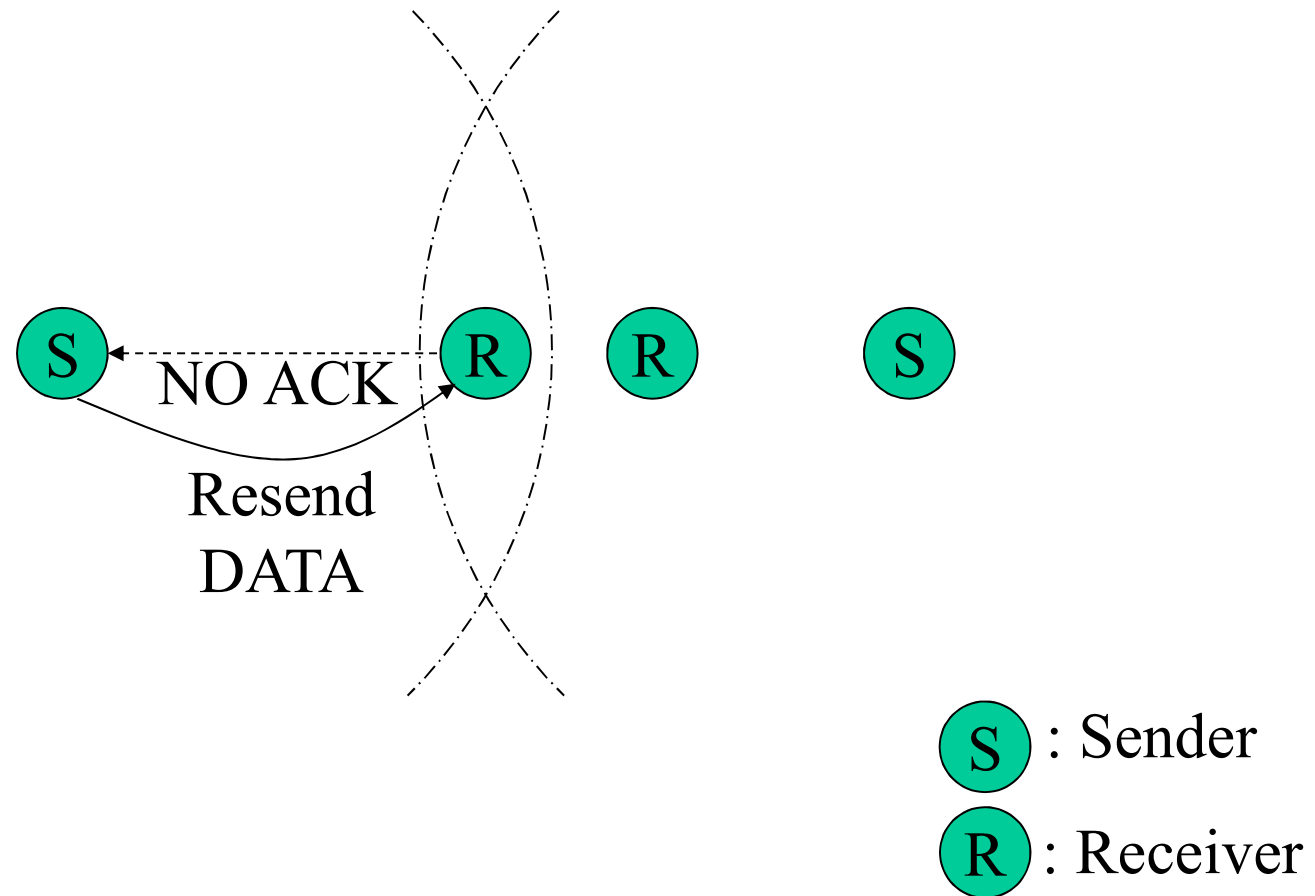
## Collision by Hidden Terminal



 : Sender  
 : Receiver

# CSMA/CA and ACK

## Collision by Hidden Terminal

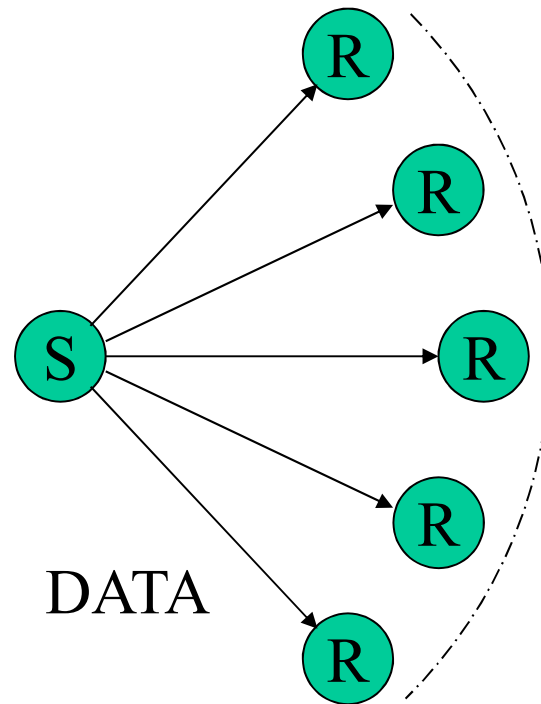


# Wireless LAN of IEEE 802.11

- Relies on CSMA/CA
  - because of undetectable collisions
    - ACKs are the MUST for reliable communication
      - or packets are lost upon collisions
    - Unicast packets are ACKed and delivered reliably
  - Broadcast/multicast packets *can not* be ACKed
    - Broadcast/multicast packets are delivered *unreliably*
      - *The major difference to Ethernet*

# CSMA/CA and ACK

## Multicast and Lack of ACK

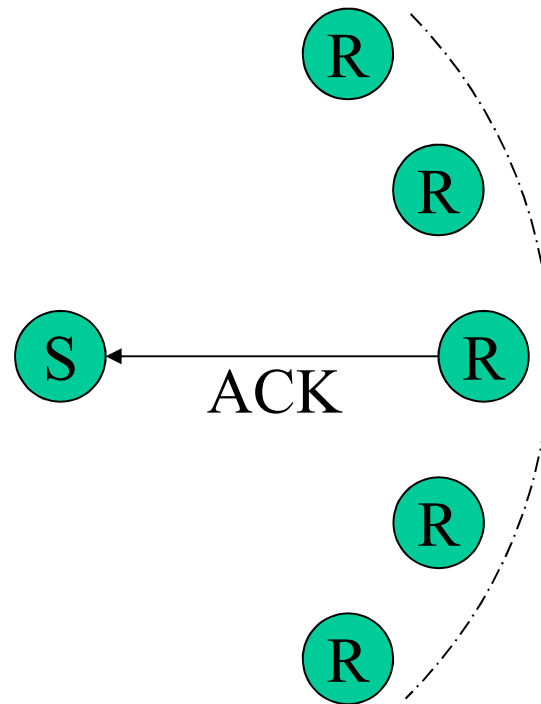



**S** : Sender

**R** : Receiver

# CSMA/CA and ACK

## Multicast and Lack of ACK

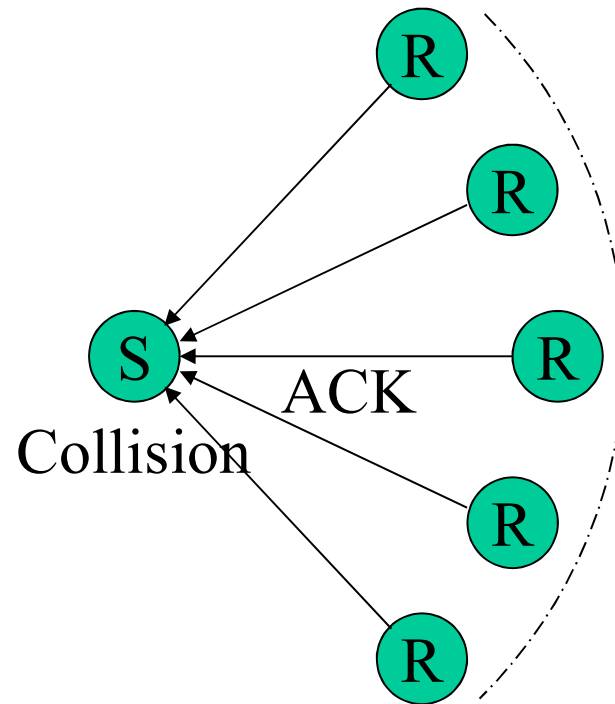


 : Sender

 : Receiver

# CSMA/CA and ACK

## Multicast and Lack of ACK

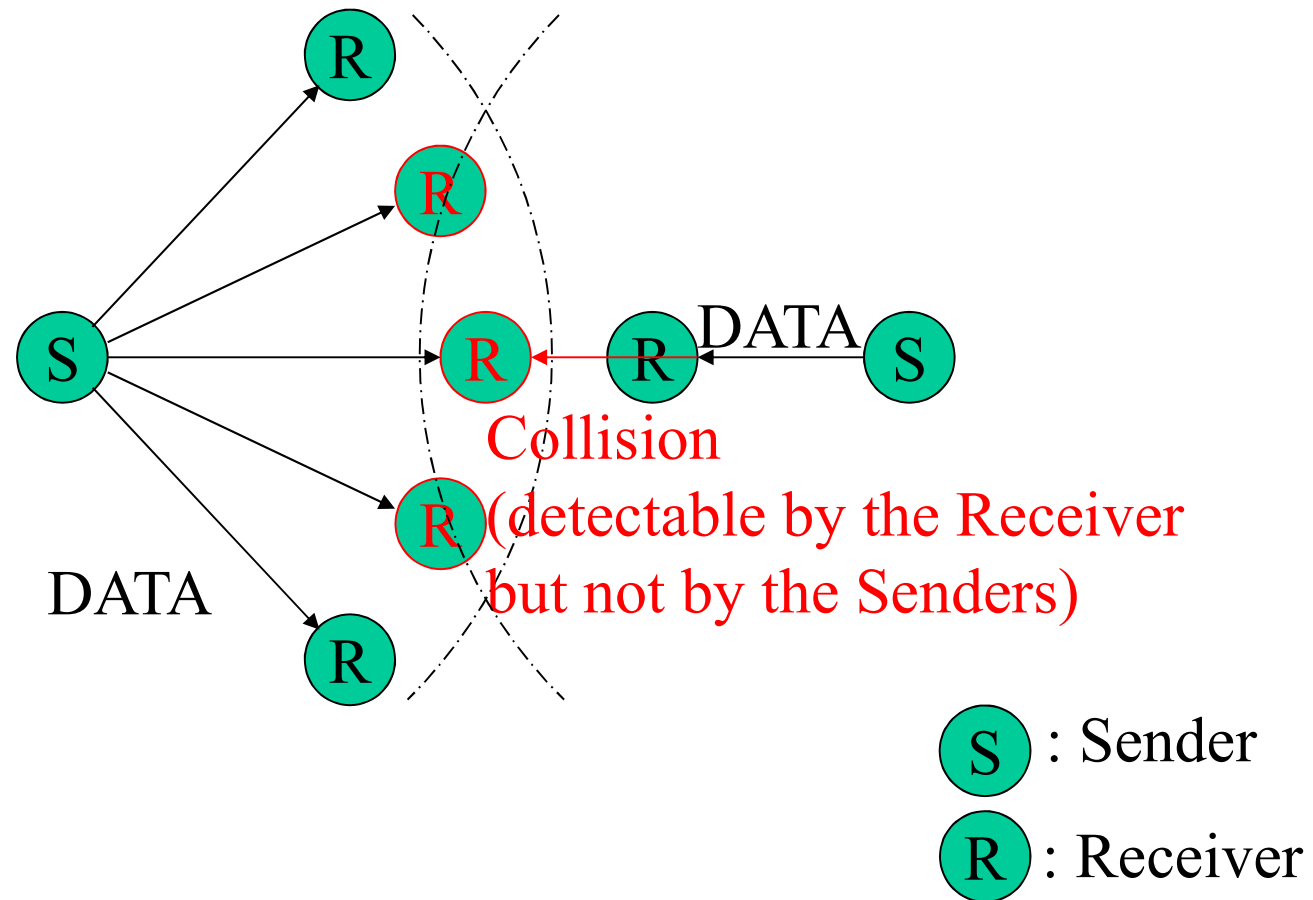


 : Sender

 : Receiver

# CSMA/CA and ACK

## Multicast and Unreliability





# Wireless LAN of IEEE 802.11

- Relies on CSMA/CA
  - because of undetectable collisions
    - ACKs are the MUST for reliable communication
      - or packets are lost upon collisions
    - Unicast packets are ACKed and delivered reliably
  - Broadcast/multicast packets *can not* be ACKed
    - Broadcast/multicast packets are delivered *unreliably*
      - *The major difference to Ethernet*
  - Reliable broadcast is by *frequent* beacons

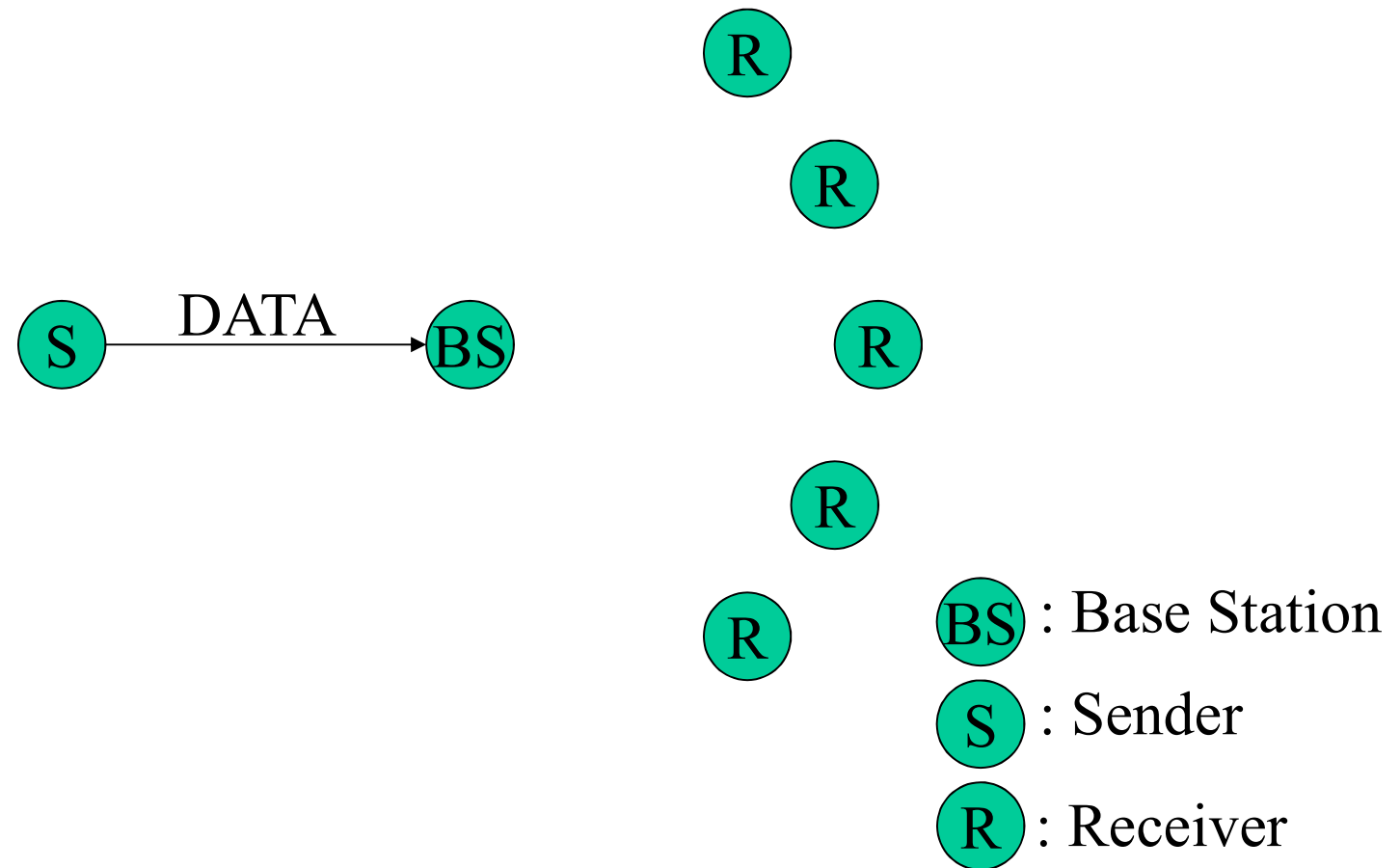
# Reliability by Frequent Beacons

- If broadcast is received 20% of the time
  - repeated beacons will finally be received
    - with 10 repetitions, 90% of the time
    - with 20 repetitions, 99% of the time
- If broadcast is received 10% of the time
  - repeated beacons will finally be received
    - with 10 repetitions, 65% of the time
    - with 20 repetitions, 88% of the time
    - with 40 repetitions, 99.5% of the time

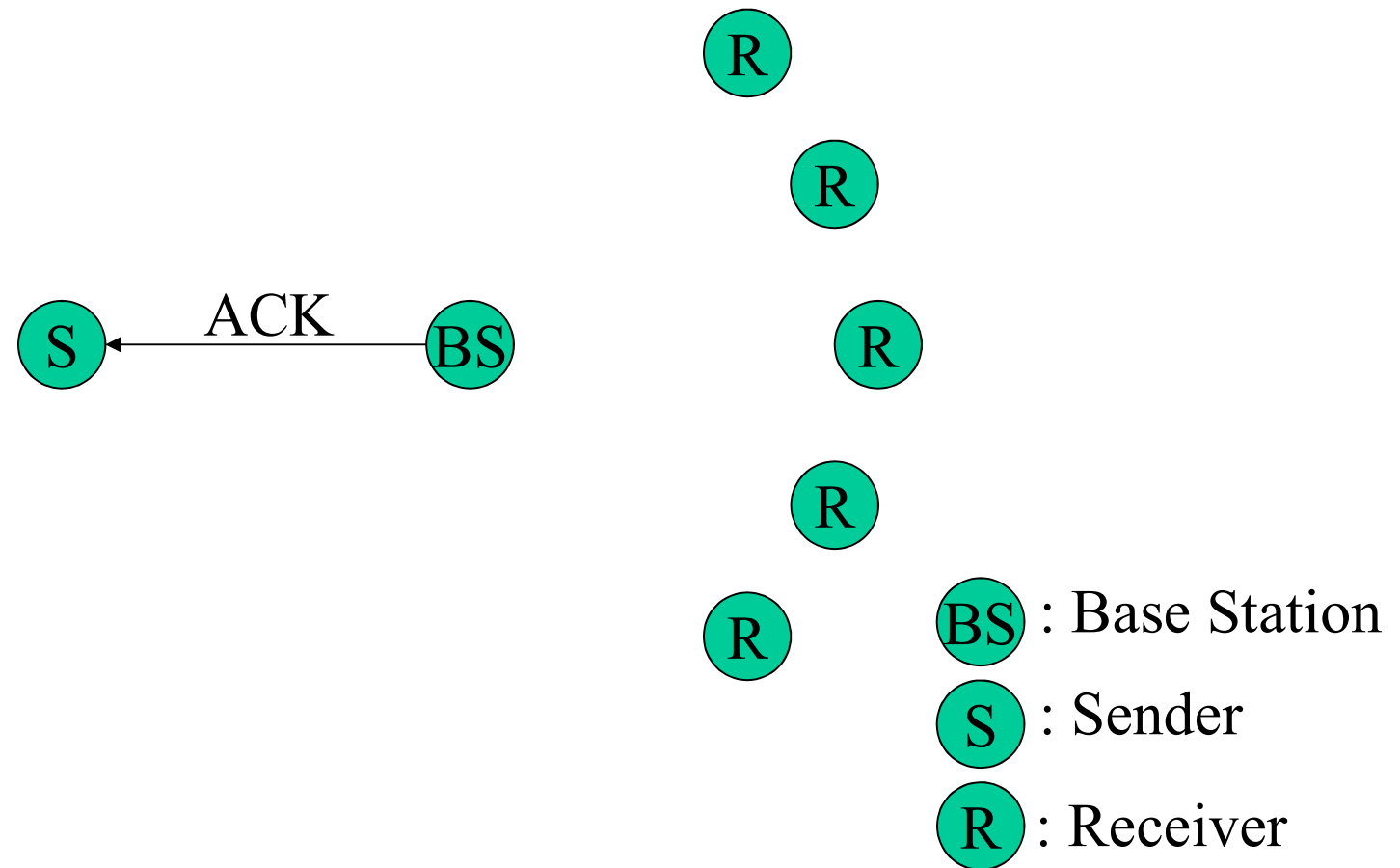
# Broadcast over Wireless LAN of Infrastructure IEEE 802.11

- Stations (STAs) send broadcast packets to the base station (BS) through link unicast
  - delivery is ACKed and reliable
    - Broadcast from STAs is received by BS reliably
  - BS, then, broadcast the packet to all the STAs
    - Broadcast from the STAs to non-BS STAs are unreliable

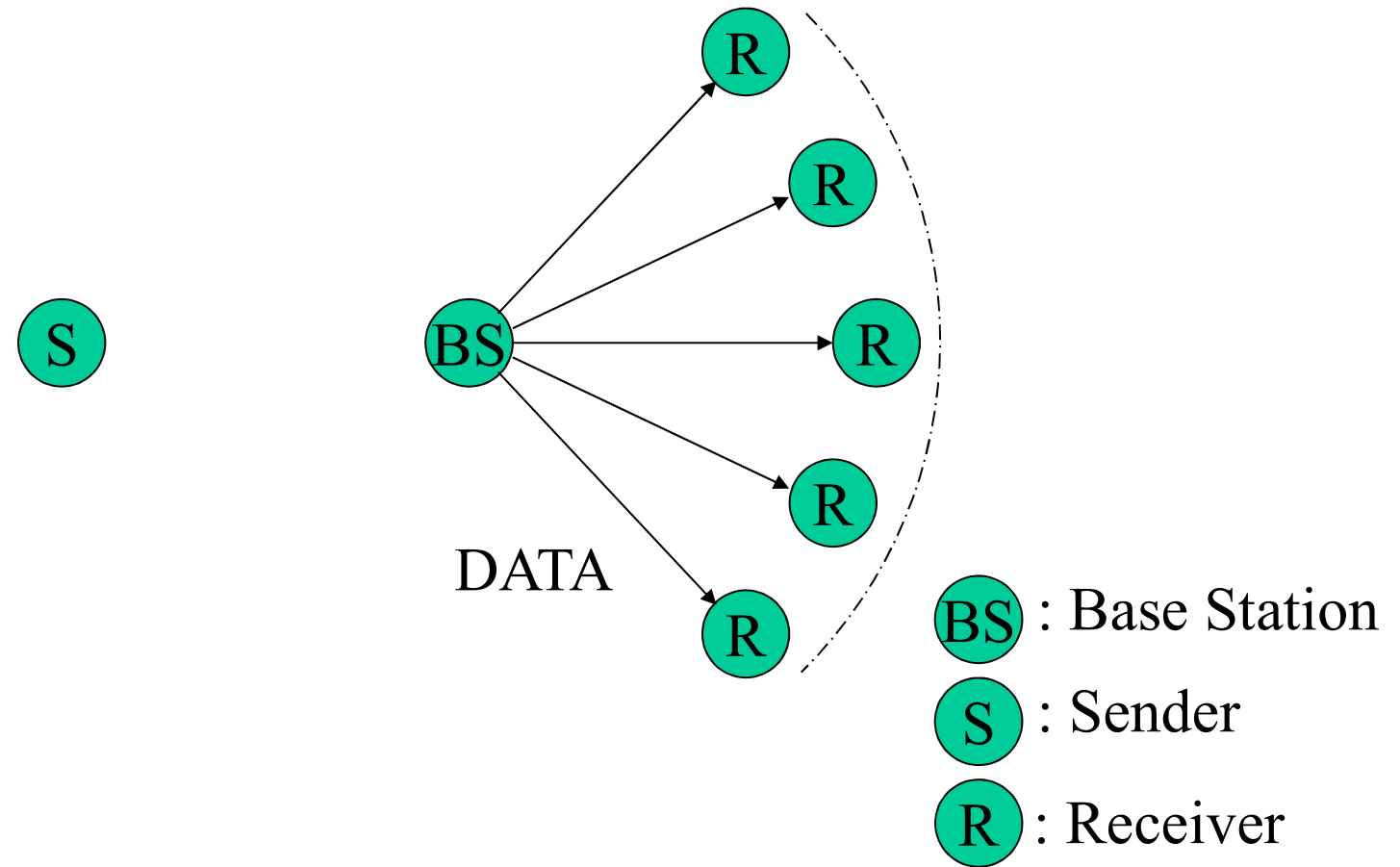
# Broadcast/Multicast over Infrastructure WLAN



# Broadcast/Multicast over Infrastructure WLAN



# Broadcast/Multicast over Infrastructure WLAN



# ARP

## The Way for IP over Ethernet

- IP uber alles! (IP over everything!)
  - IP **MUST** work over any link layers
    - Various adaptation mechanisms take care of matching between L3 and L2s
      - The adaptation mechanisms take care of differences between various L2s
    - ARP (of IPv4) is the adaptation mechanisms between IP and Ethernet

# Neighbor Discovery (ND)

## The Major Design Flaw of Ipv6

- ~~IP uber alles! (IP over everything!)~~
- ND uber alles!?! (ND over everything)
  - ND **MUST** work over any link layers!
    - A single adaptation mechanism **CAN NOT** take care of matching between L3 and various L2s
      - The single adaptation mechanism **CAN NOT** take care of differences between various L2s
  - ND was designed for Ethernet, PPP and ATM
    - but not for Wireless LAN nor other L2s
      - ND **MAY NOT** be able to take care of Wireless LAN



# Wrong Assumptions of ND on L2

- The world will be ATM centric
  - IP over a large L2 cloud of worldwide ATM
    - L1/L2 broadcast is inhibited
      - timeout period of L2 multicast (P2MP) is long
- Terminals are mostly immobile
  - “Routers generate Router Advertisements frequently enough that hosts will learn of their presence *within a few minutes*” (RFC2461)
- L2 broadcast/multicast is reliable

# The Reality of L2s under IP

- The world is IP centric
  - ATM has gone
- L2 is small
  - The CATENET model, of course
- Terminals are highly mobile
  - can't wait a few minutes for network reconf
- L2 broadcast/multicast is **UNRELIABLE** over (congested) WLAN

# How ND was expected to Work over WLAN

- NS (Node Solicitation) is **multicast**
- RS (Router Solicitation) is **multicast**
  - received unreliably (except by BS)
  - Then, RA (Router Advertisement) is  
unicast/**multicast**
- Unsolicited RA is **multicast**

# Other Protocols Affected

- Protocols using broadcast/multicast suffer
  - DHCP, ARP, Routing Protocols, ...,
- However, if BS is the only router
  - DHCP discover to BS is reliable
  - ARP to BS is reliable
  - ARP from BS is unreliable
    - not common in ad hoc environment
  - Routing protocols are not necessary

# Reaction from IETF

- The problem is recognized
- Treat WLAN as NBMA?
  - However, RFC2461 says “The details of how one uses ND on NBMA links is an area for further study.”

# Conclusions

- Wireless LAN and Ethernet are different
- “ND over everything” is a bad idea
  - proven by the most popular (next to Ethernet) L2 technology of wireless LAN
- Further study is necessary
  - to make IPv6 deployable
- IP uber alles!
  - not necessarily IPv6

# How Path MTU Discovery not Work

Masataka Ohta

Tokyo Institute of Technology

[mohita@necom830.hpcl.titech.ac.jp](mailto:mohita@necom830.hpcl.titech.ac.jp)

# Abstract

- Multicast path MTU discovery (PMTUD) is a new feature of IPv6. However, ICMP implosion with multicast PMTUD can be serious when most MTU bottlenecks are located near individual receivers. ICMP Packet Too Big, at least those generated against multicast packets, will be filtered, which is a standard violation, which means there is no reason not to filter unicast ones. Thus, unicast PMTUD is not expected to work. We should not send packet >1280B, except for IP over IP tunnels.



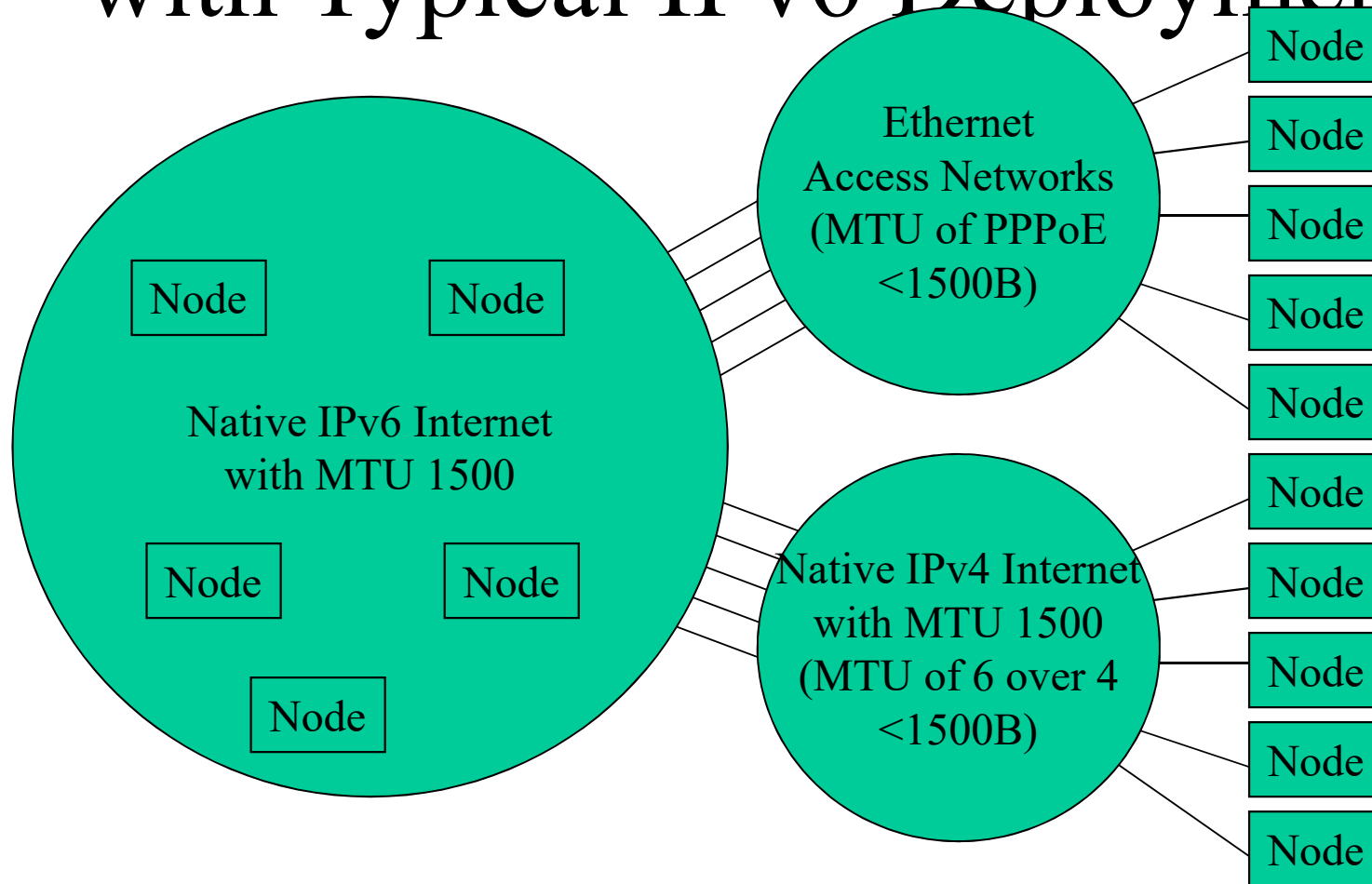
# PATH MTU Discovery

- Measure Path MTU by ICMP Packet Too Big
  - Path MTU is set to the value contained in the ICMP packet
  - does not work if ICMP Packet Too Big is filtered or not generated
- Periodically send larger packet to detect MTU increase by path change
- “SHOULD be supported” (node requirement)
  - ISPs SHOULD NOT filter ICMP Packet Too Big?

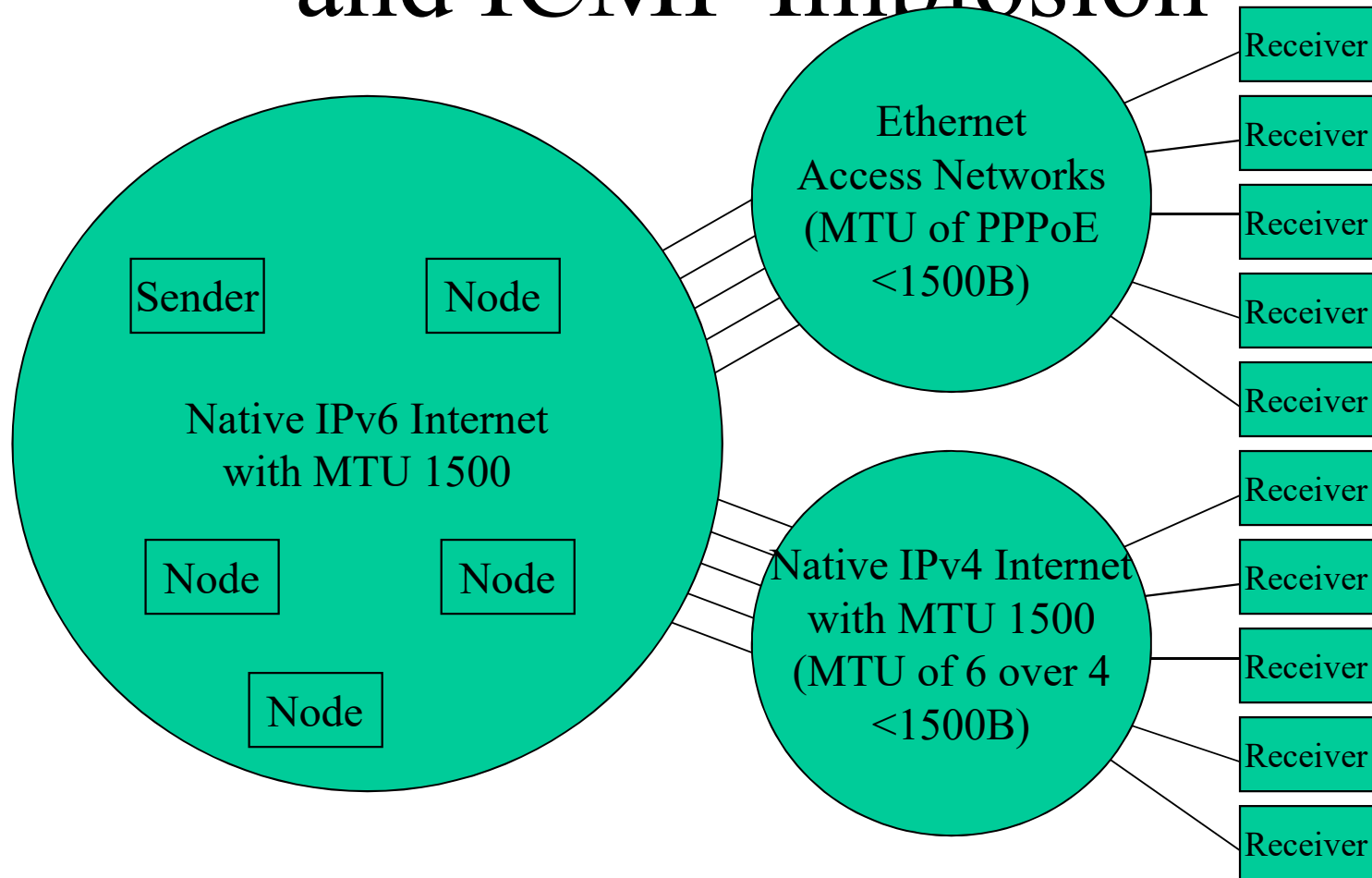
# RFC1981 (Path MTU Discovery for IP version 6)

- The Draft Standard Specifies:
  - Path MTU Discovery supports multicast as well as unicast destinations. In the case of a multicast destination, copies of a packet may traverse many different paths to many different nodes. Each path may have a different PMTU, and a single multicast packet may result in multiple Packet Too Big messages, each reporting a different next-hop MTU. The minimum PMTU value across the set of paths in use determines the size of subsequent packets sent to the multicast destination.
  - In the case of a multicast destination address, copies of a packet may traverse many different paths to reach many different nodes. The local representation of the "path" to a multicast destination must in fact represent a potentially large set of paths.
- How large is “a potentially large set of paths”?

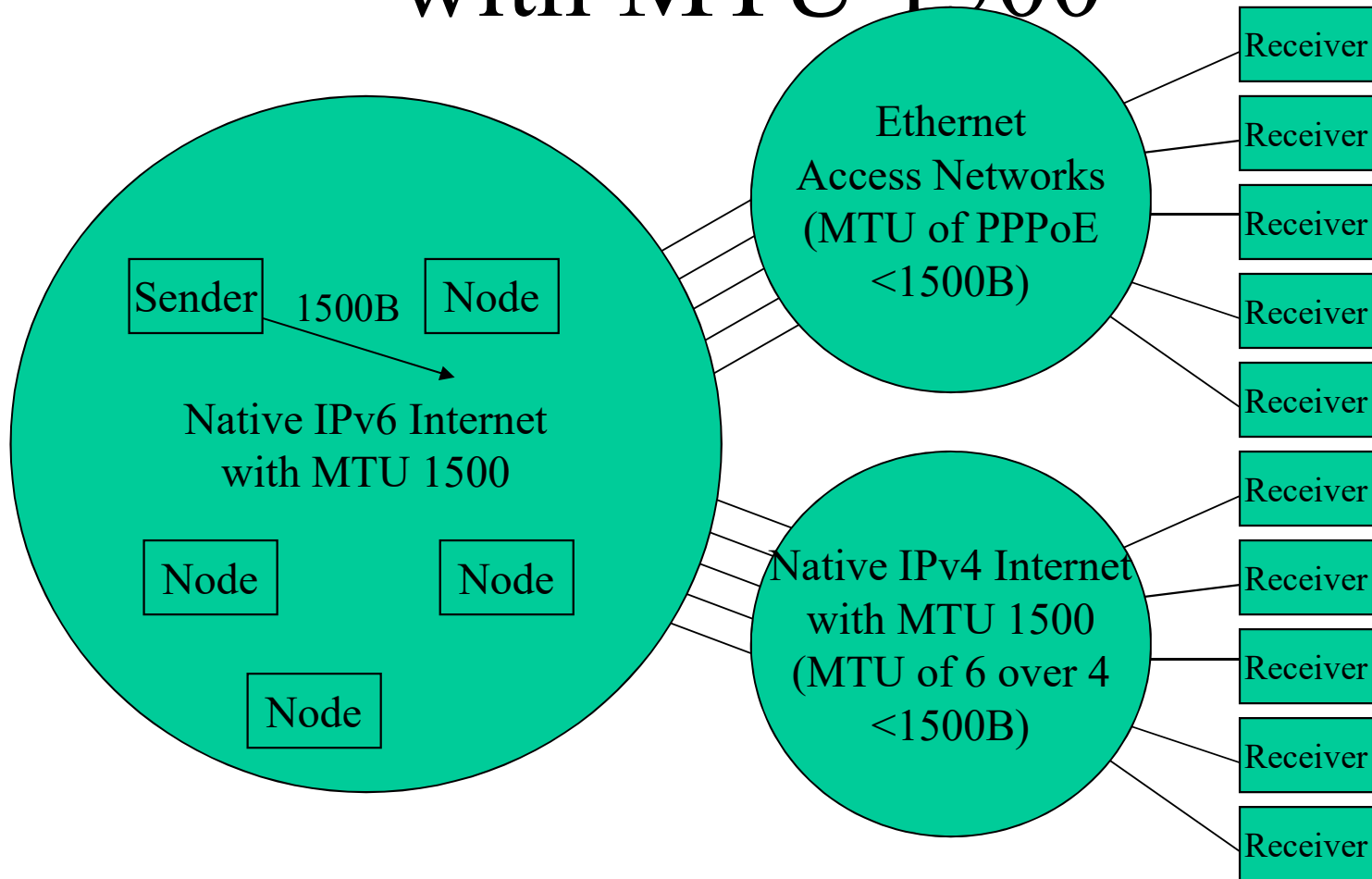
# Tunnels at the Last Hop with Typical IPv6 Deployment



# Multicast Path MTU Discovery and ICMP Implosion

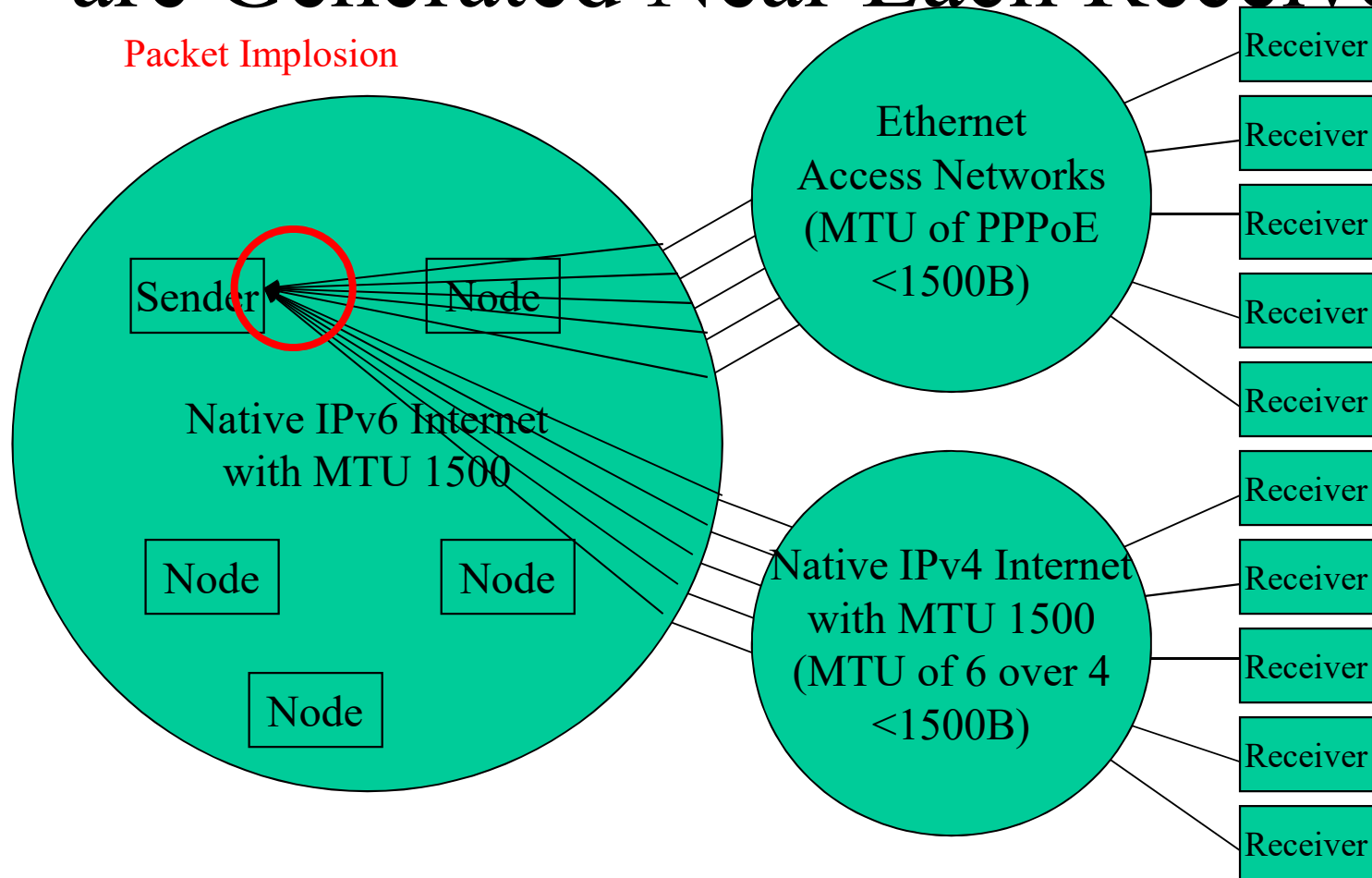


# Sender Periodically Send Packets with MTU 1500



# ICMP Packet Too Big Messages are Generated Near Each Receiver

Packet Implosion



# DOS

- Some multicast routing protocol allows for source address spoofing
  - ICMP may be used for DOS amplifier
  - even if non-link-local multicast is not enabled around a victim

# Not a Problem?

- Because almost all ISPs do not enable multicast routing protocol
- ISPs do not allow ordinary users send multicast packets
  - **still a problem**, because rational ISPs want to avoid to rely on rational operations of other ISPs
  - instead, the multicast PMTUD problem is yet another reason for ISPs to disable multicast
    - multicast PMTUD, to promote multicast and MTUD, ironically killed multicast and MTUD thoroughly



# RFC2463 (ICMPv6) Requires

- A Packet Too Big **MUST be sent** by a router in response to a packet that it cannot forward because the packet is larger than the MTU of the outgoing link. The information in this message is used as part of the Path MTU Discovery process [PMTU].
- Sending a **Packet Too Big Message** makes an **exception** to one of the rules of when to send an ICMPv6 error message, in that unlike other messages, **it is sent in response to a packet received with an IPv6 multicast destination address, or a link-layer multicast or link-layer broadcast address.**
  - Parameter Problem Messages also make an exception

# To Prevent ICMP Implosions

- Violate RFC2463 to
  - stop generating ICMP packet too big and parameter problem for multicast packet
  - filter ICMP packet too big and parameter problem for multicast packet
- Or, as it is already a violation, simply
  - stop generating any ICMP
  - filter all the ICMP
  - “it’s against an RFC” is not a valid criticism

# Fundamental Solution

- Update RFC2463 to prohibit generation of ICMP against multicast packets
- Write an BCP to Force ISPs not Filter ICMP
- Should take another decade or two
  - unrealistic

# Without PMTUD...

- According to RFC2460:
  - It is strongly recommended that IPv6 nodes implement Path MTU Discovery [RFC-1981], in order to discover and take advantage of path MTUs greater than 1280 octets. However, a minimal IPv6 implementation (e.g., in a boot ROM) may **simply restrict itself to sending packets no larger than 1280 octets**, and omit implementation of Path MTU Discovery.
    - Packet larger than 1280B can not be sent
- IP over IP tunnels (e.g. RFC2473 for MIPv6) needs tunnel MTU 1280B, violating RFC2460
  - or, all the 1280B packets are fragmented, because MTU of 1280B tunnel is smaller than 1280B

# Conclusion

- Multicast PMTUD is broken
  - to cause ICMP implosion
- ISPs should filter ICMP Packet Too Big
  - at least against multicast packets but maybe all
- We can't expect unicast PMTUD work
- We shouldn't send packets  $> 1280\text{B}$ 
  - except for tunnels

# 新世代ネットワークアーキテクチャ実現に向けて —AKARIプロジェクトからの報告—

新世代ネットワークワークショップ

2007年6月11日

平原 正樹

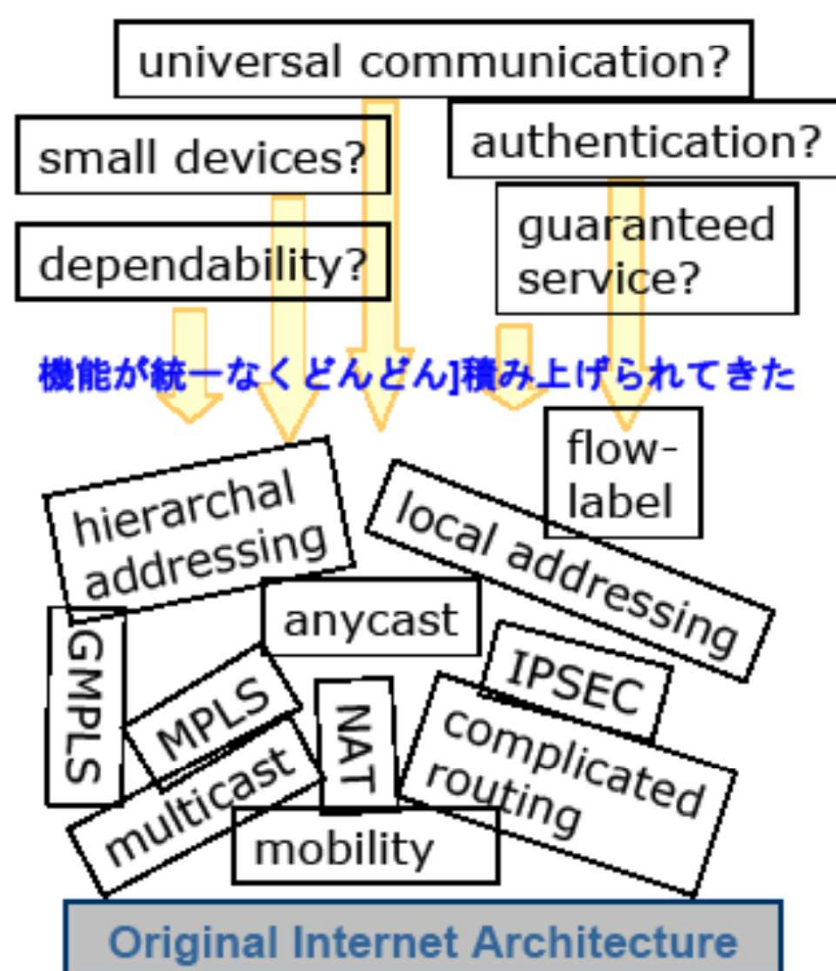
ネットワークアーキテクチャグループ

新世代ネットワーク研究センター

NICT

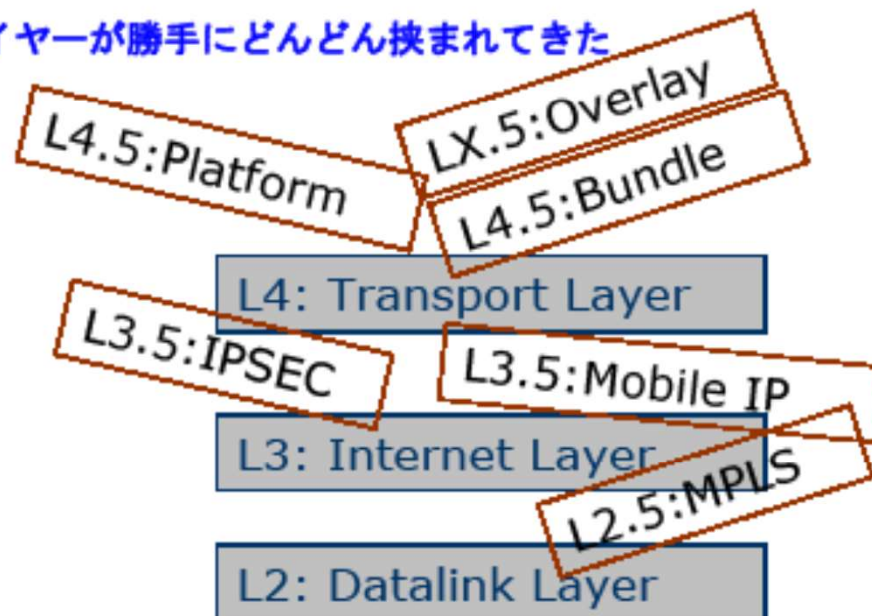
# インターネット – あまりにも複雑/矛盾 → 破綻 第2章1節

新しい機能を積み上げることができない。未来の社会を支えるサービスを提供できない。



- ・命を預けられるか?(遠隔医療、交通、緊急通報)
- ・生活を預けられるか?(防犯、契約行為、金融)
- ・生活を豊かにできるか?(センサー、RFID)
- ・安心した生活が送れるか?(対スパム、耐攻撃)
- ・あと何十年使えるの?(インフラ、持続社会)
- ・未来の変革を受け入れる余裕は?(未知の要求)

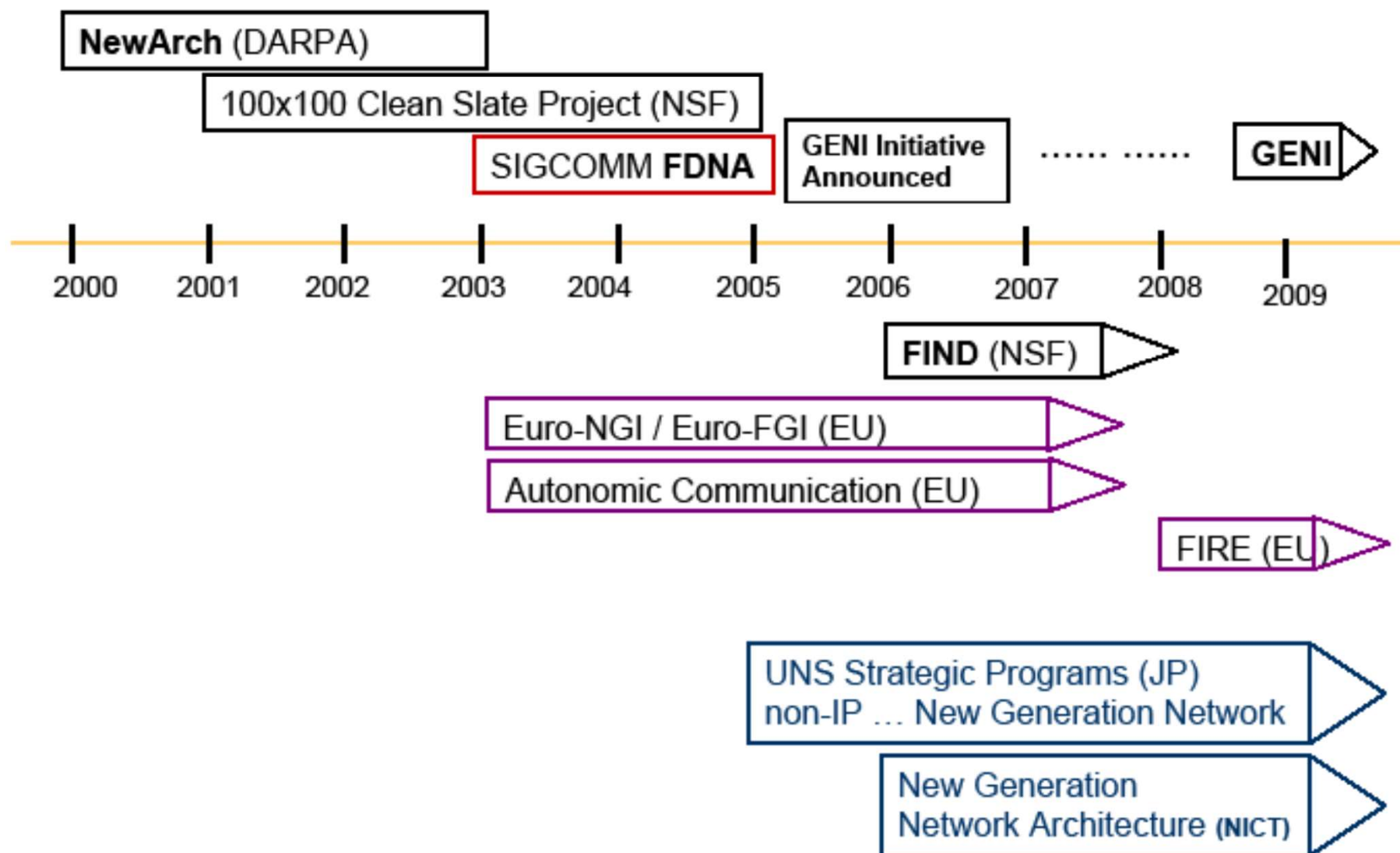
レイヤーが勝手にどんどん挟まれてきた



個々はOKでも全体では×

**一から作り直すべき時期が迫っている!**

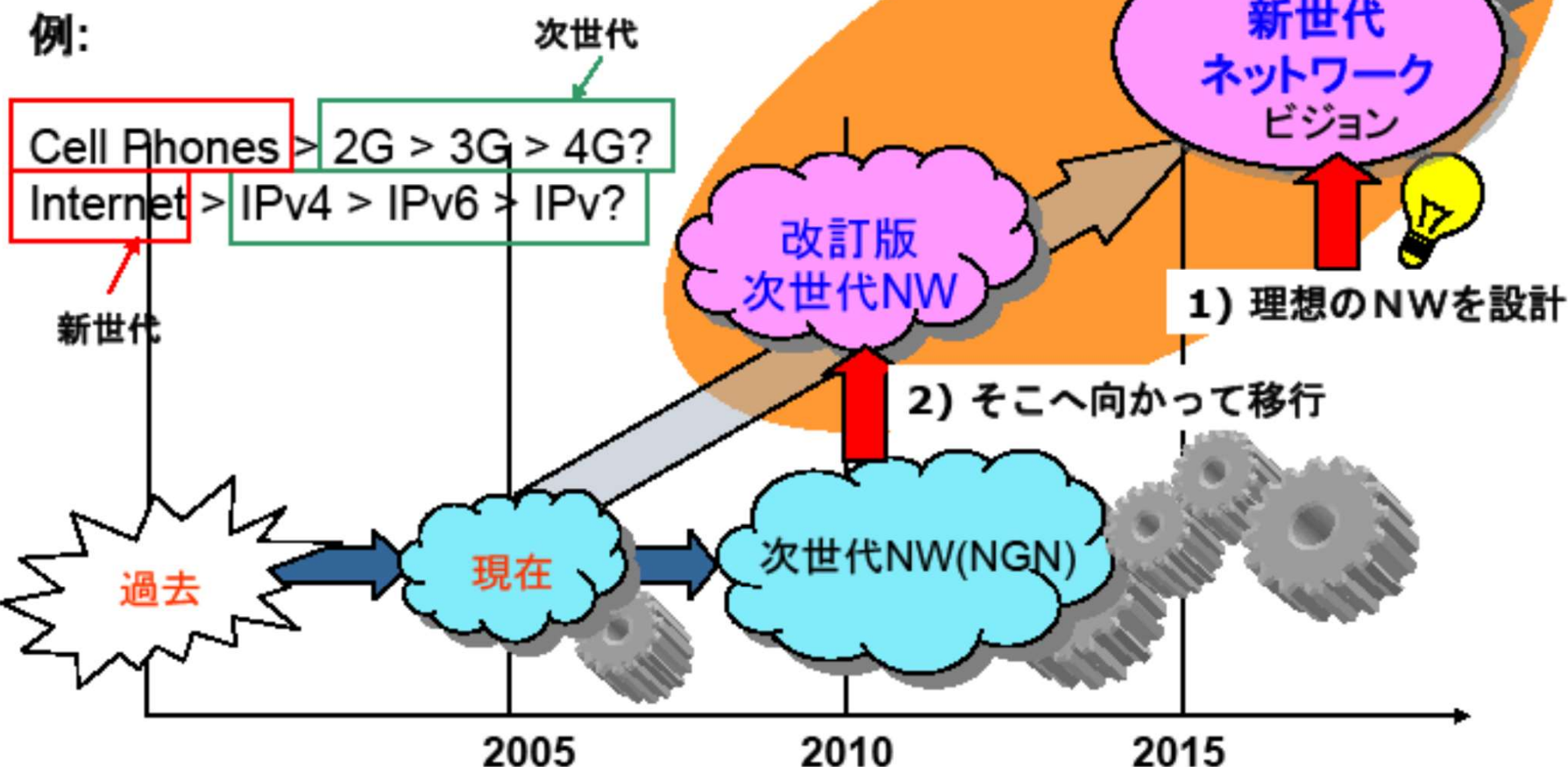
## ネットワークアーキテクチャ“白紙から作り直し”機運





# 新世代ネットワーク研究開発の 位置付け

例:



# AKARIプロジェクト

- a small light in the dark pointing to the future -

## 目標: 2015年の新世代ネットワーク設計図

・現在のしがらみに捕われない。・白紙から理想を追い求める。・その後で現在からの移行を考える。

グループリーダー: 平原 正樹

原井 洋明(光交換)、徐 蘇鋼(光パス)、宮澤 高也

盛岡 敏夫(光伝送)、大槻 英樹(ネットワーク制御)、Jumpot Phuritakul

井上 真杉(アクセス)、中内 清秀(オーバーレイ)、Ved Kafle (アドレッシング)

客員  
研究員

大阪大学 村田教授 (ネットワーク科学)

慶応大学 寺岡教授 (モビリティ)

東京大学 森川教授 (ユビキタス)

東京工業大学 太田講師 (パケット交換)

Masataka Ohta (packet exchange)

アドバイザー:

青山(プログラムディレクタ)

久保田(センター長)

ミーティング 2回/月, 合宿 2回/年

## AKARI アーキテクチャの目的

### 持続可能なネットワークアーキテクチャ

量から質へ (Capacity for Quality)

簡約化 (KISS)

ネット空間の現実化 (Realizable Network Space)

人類の可能性を伸ばす (Future Diverse Society)

## 新世代ネットワークアーキテクチャの原理原則

### 1. KISS原則 (Keep It Simple, Stupid)

- 結晶合成 (選択・統合・単純化)
- 共通レイヤ (レイヤ縮退)
- End-to-End (Original Internet)

### 2. 現実結合原則

- 物理・論理アドレス分離
- 双方向認証
- 追跡可能

### 3. 持続的な進化原則

- 自己創発 (エマージェント)
- 自律分散制御
- スケーラブル
- 社会選択

# What is “packet exchange”?

- PDMA (as explained in 2)
- optical packet (will be explained in 12.)
- IP-- (explanations follow)

# IP--

- originally intended IPv6 (originally named SIP (Simple Internet Protocol))
  - extend address space
  - aggressive address hierarchy
  - simplification

# Properties of IP--

- no IP options
  - header may included source locators (up to 15)
- minimum MTU 9kB, no PMTUD
- support broadcast
- multicast and IPsec not included
- ID locator separation
  - locators has 5 layers of hierarchy (12 bits each)
    - all hosts should have full (4k entries only ) routing table for 5 layers, no default route or router

# Packet Format of IP--

Payload Length	Protocol	HTL
Source Locator List Length	(reserved)	
Source Locator		
Source ID		
Destinatino Locator		
Destination ID		
Source Locator List		
Payload		