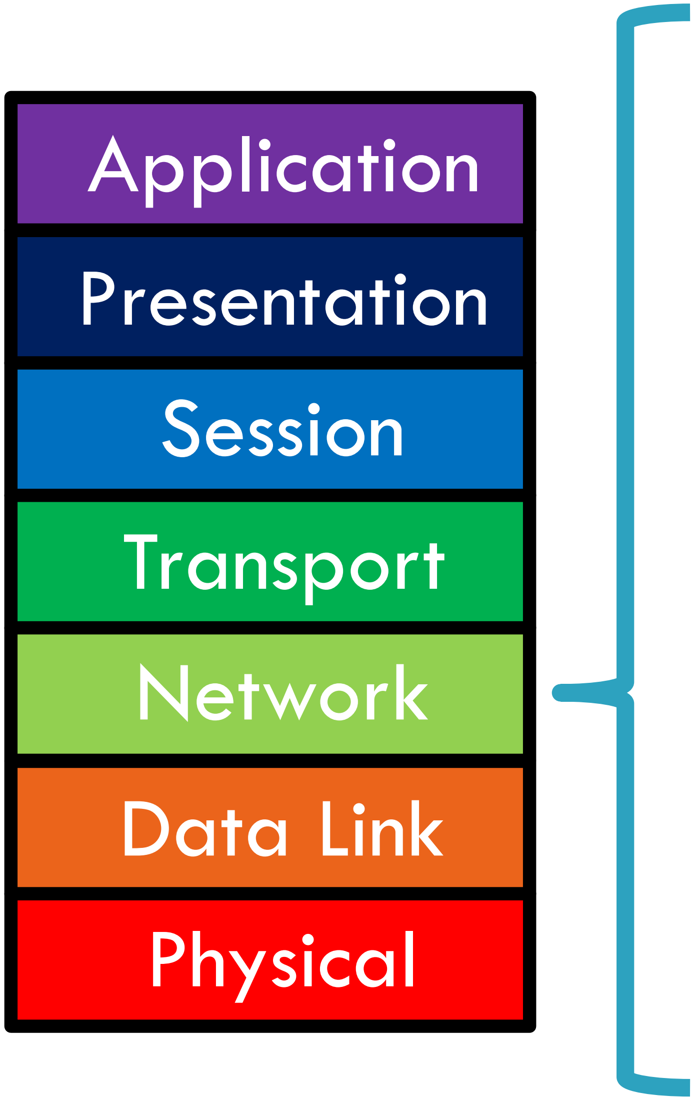
CSCI-351
Data communication and Networks

Lecture 8: Network Layer
(Putting the Net in Internet)

The slide is built with the help of Prof. Alan Mislove, Christo Wilson, and David Choffnes's class

Recap

Network Layer



Application

Presentation

Session

Transport

Network

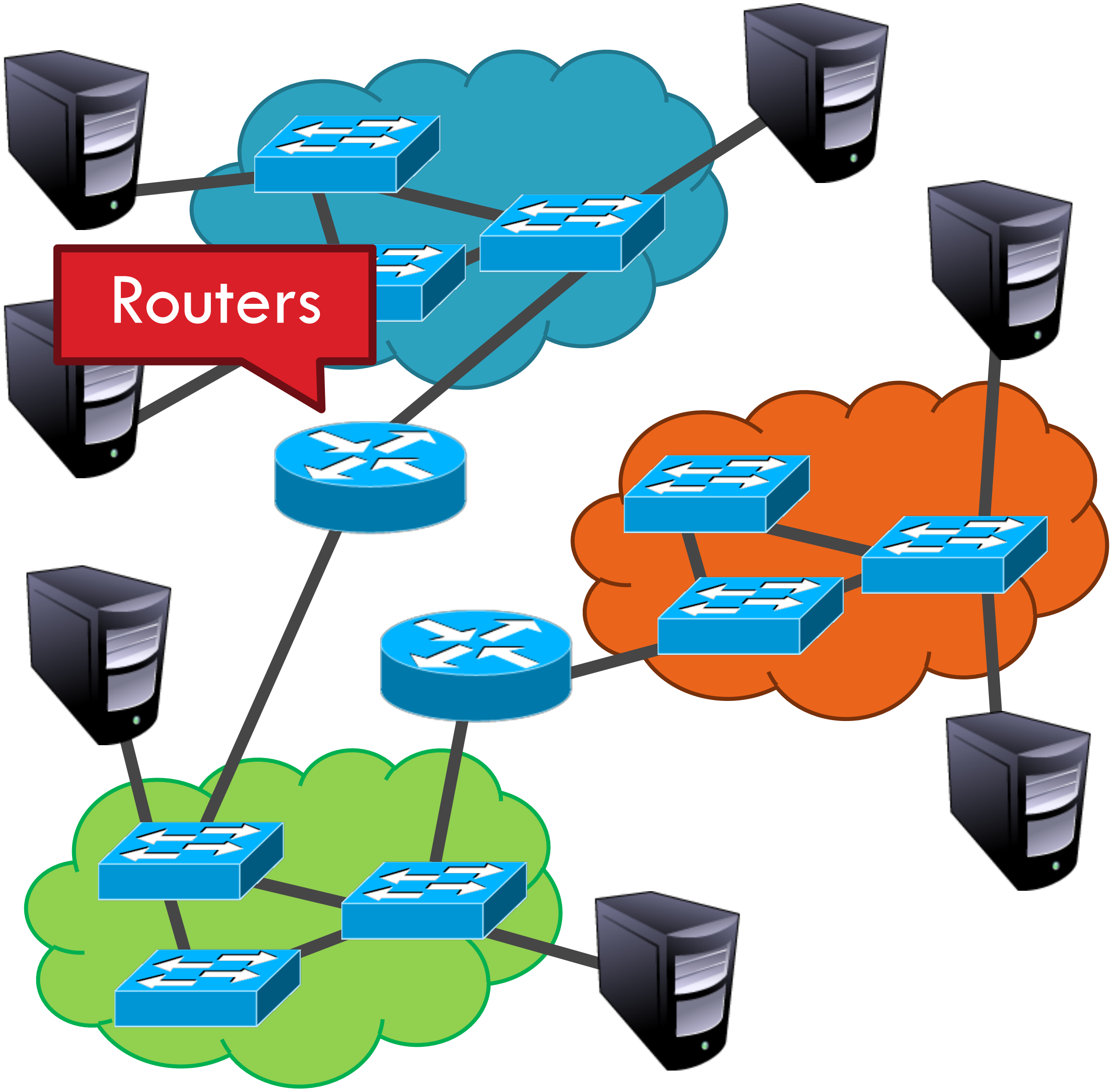
Data Link

Physical

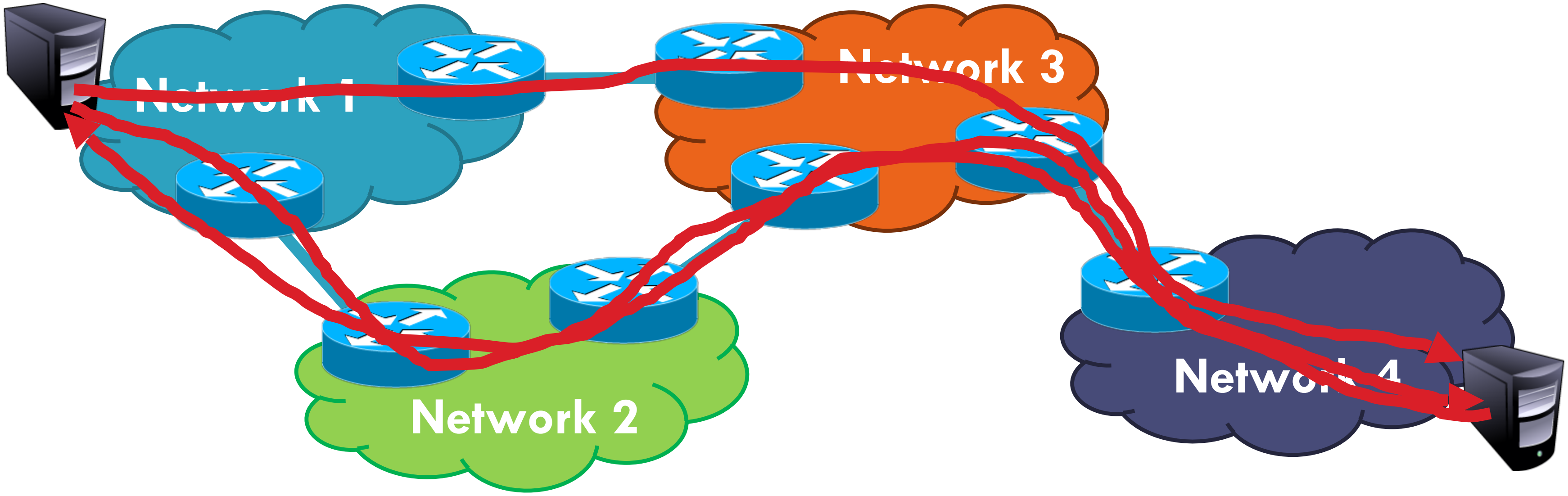
- Function:
  - Route packets end-to-end on a network, through multiple hops
- Key challenge:
  - How to represent addresses
  - How to route packets
    - Scalability
    - Convergence

Routers, Revisited

- How to connect multiple LANs?
- LANs may be incompatible
  - Ethernet, Wifi, etc…
- Connected networks form an internetwork
  - The Internet is the best known example


Routers

Structure of the Internet



- Ad-hoc interconnection of networks
  - No organized topology
  - Vastly different technologies, link capacities
- Packets travel end-to-end by hopping through networks
  - Routers “peer” (connect) different networks
  - Different packets may take different routes

Internetworking Issues

- Naming / Addressing
  - How do you designate hosts?
- Routing
  - M
- Ser
  - W
  - H
  - What happens if there are failures?
  - Must deal with heterogeneity
    - Remember, every network is different

Internet Service Model

- Best-effort (i.e. things may break)
- Store-and-forward datagram network

Outline

- Addressing
  - Class-based
  - CIDR
- IPv4 Protocol Details
  - Packed Header
  - Fragmentation
- IPv6

Possible Addressing Schemes

- Flat
  - e.g. each host is identified by a 48-bit MAC address
  - Router needs an entry for every host in the world
    - Too big
    - Too hard to maintain (hosts come and go all the time)
    - Too slow (more later)
- Hierarchy
  - Addresses broken down into segments
  - Each segment has a different level of specificity

Example: Telephone Numbers

1-585-475-3278

Very General

585
NEW YORK
OHIO
PENNSYLVANIA
NEW JERSEY
WEST VIRGINIA
worldatlas

R·I·T

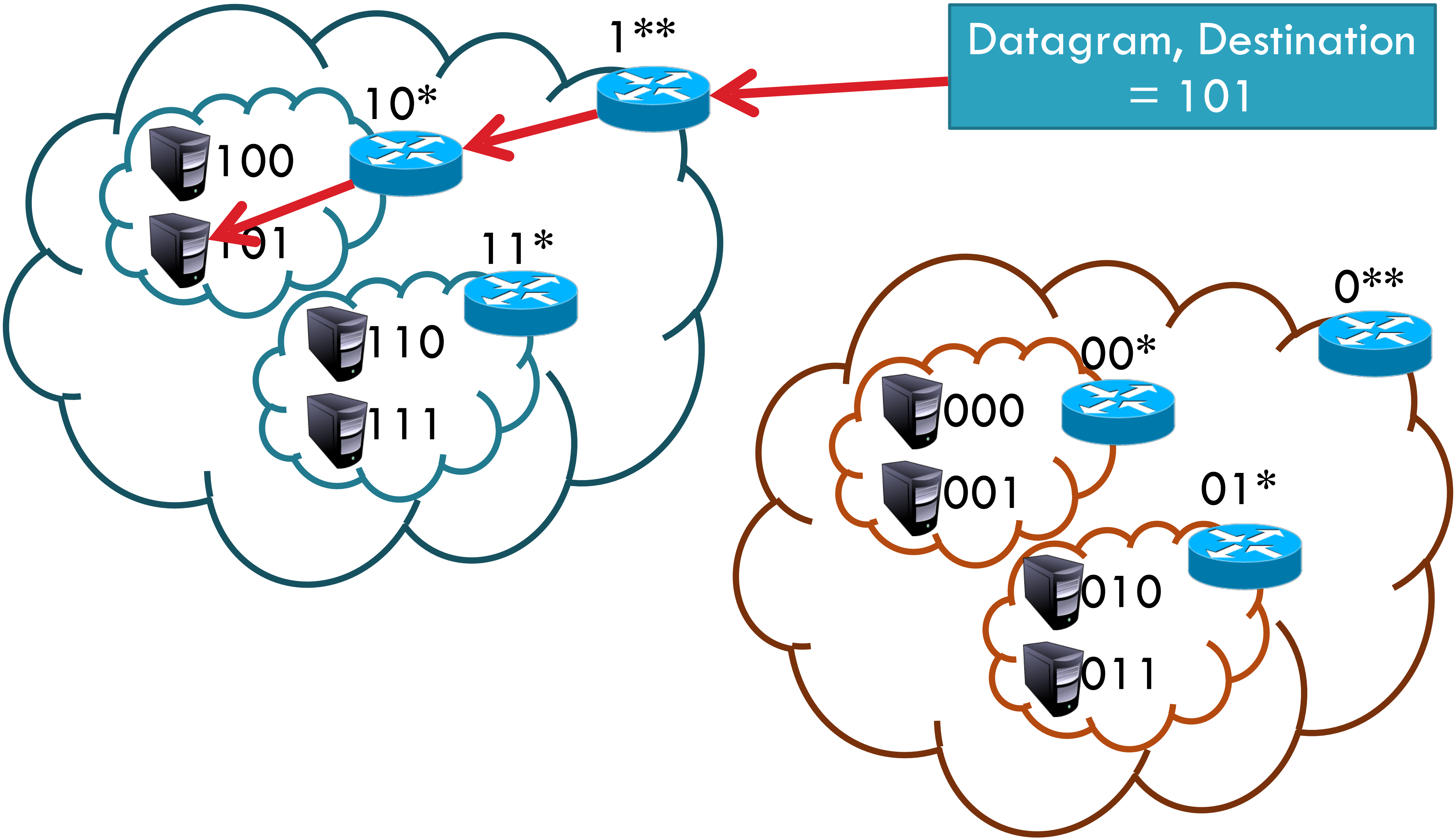
RIT GCCIS
Room 4567

RIT GCCIS
Room 1234

Updates are Local

Very Specific

Binary Hierarchy Example



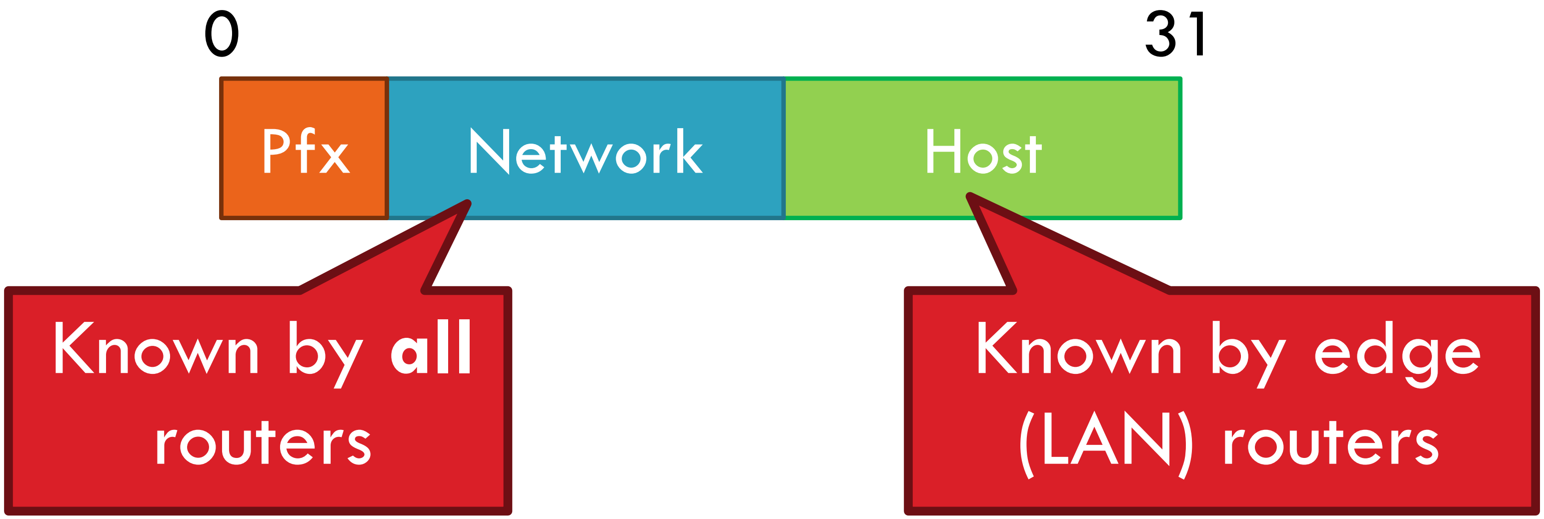
IP Addressing

- IPv4: 32-bit addresses
  - Usually written in dotted notation, e.g. 192.168.21.76
  - Each number is a byte
  - Stored in Big Endian order

	0	8	16	24	31
Decimal	192	168	21	76	
Hex	C0	A8	15	4C	
Binary	11000000	10101000	00010101	01001100	

IP Addressing and Forwarding

- Routing Table Requirements
  - For every possible IP, give the next hop
  - But for 32-bit addresses, 2^{32} possibilities!
  - **Too slow:** 48GE ports and 4x10GE needs 176Gbps bandwidth **DRAM:** ~1-6 Gbps; **TCAM** is fast, but 400x cost of DRAM
- Hierarchical address scheme
  - Separate the address into a network and a host



0

31

Pfx

Network

Host

Known by **all** routers

Known by edge (LAN) routers

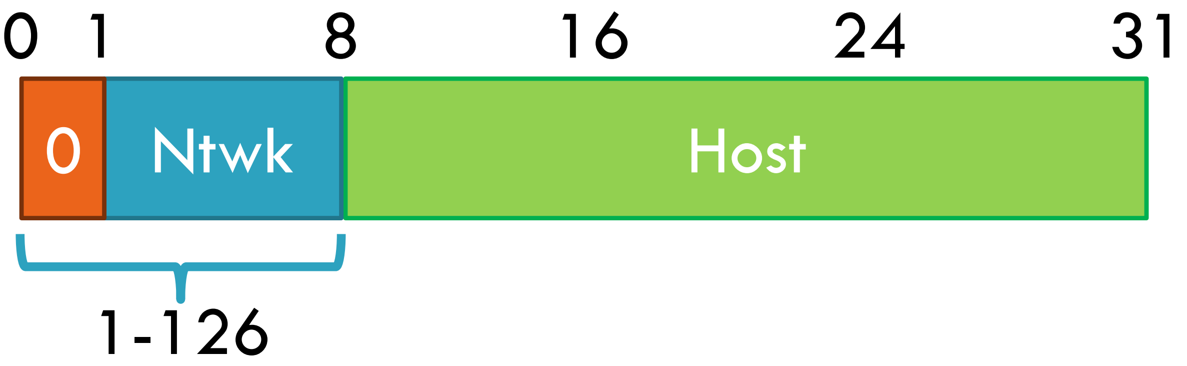
Recap

IP address

- IPv4: 32-bit addresses
  - Too big to fit into a routing table (2^32)
  - Let's make a hierarchy!
    - Class-based IP address
      - Some address spaces are wasted
      - Still too many hosts in one network
  - Let's make another hierarchy!
    - Subnet Mask
  - Classless Routing

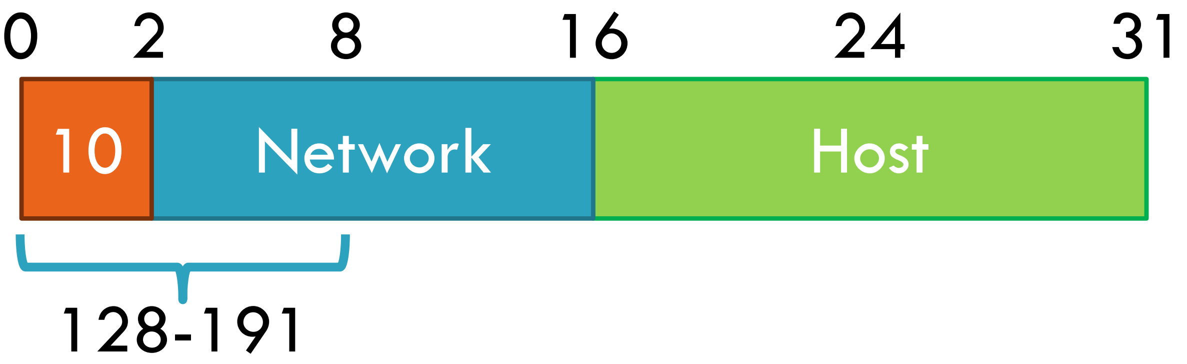
Classes of IP Addresses

- □ Class A



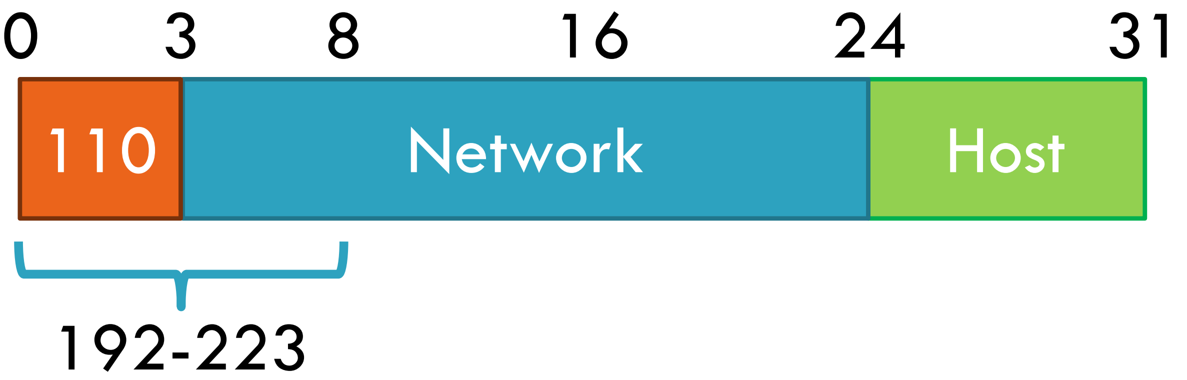
Example: MIT
18.*.*.*

- □ Class B



Example: RIT
172.31.*.*

- □ Class C



Example:
216.63.78.*

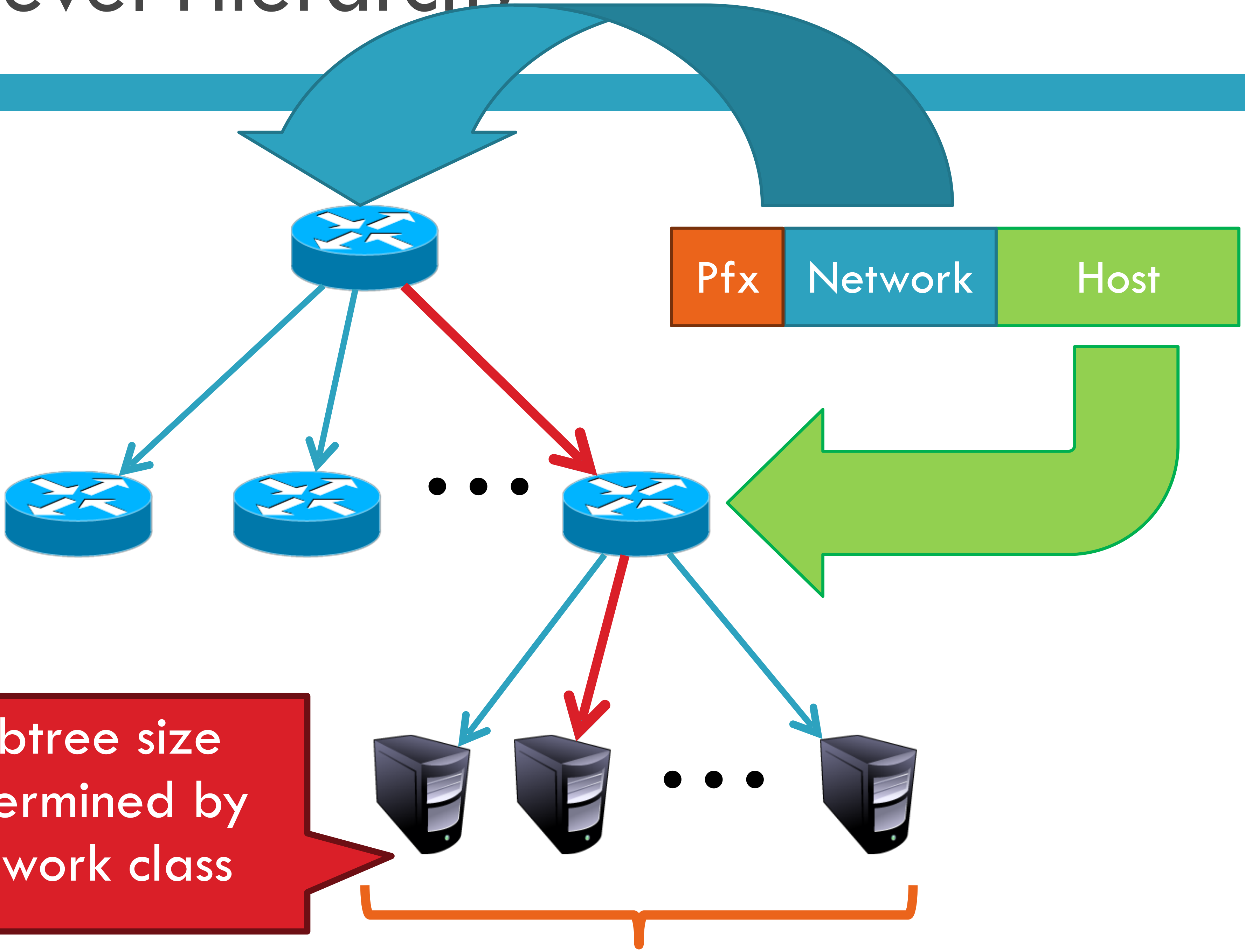
How Do You Get IPs?

- IP address ranges controlled by IANA

Internet Assigned Numbers Authority

  - Internet Assigned Number Authority
  - Roots go back to 1972, ARPANET, UCLA
  - Today, part of ICANN
- IANA grants IPs to regional authorities
  - ARIN (American Registry of Internet Numbers) may grant you a range of IPs
  - You may then advertise routes to your new IP range
  - There are now secondary markets, auctions, …

Two Level Hierarchy



Pfx

Network

Host

Subtree size determined by network class

Class Sizes

Class	Prefix Bits	Network Bits	Number of Classes	Hosts per Class
A	1	7	$2^7 - 2 = 126$ (0 and 127 are reserved)	$2^{24} - 2 = 16{,}777{,}214$ (All 0 and all 1 are reserved)
B	2	14	$2^{14} = 16{,}398$	$2^{16} - 2 = 65{,}534$ (All 0 and all 1 are reserved)
C	3	21	$2^{21} = 2{,}097{,}512$	$2^8 - 2 = 254$ (All 0 and all 1 are reserved)
			Total: 2,114,036	

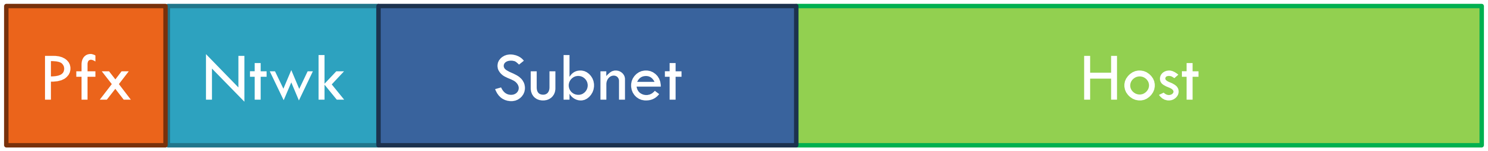
Way too big

Too many network IDs

Too small to be useful

Subnets

- Problem: need to break up large A and B classes
- Solution: add another layer to the hierarchy
  - From the outside, appears to be a single network
    - Only 1 entry in routing tables
  - Internally, manage multiple subnetworks
    - Split the address range using a subnet mask



Subnet Mask:

Subnet Example

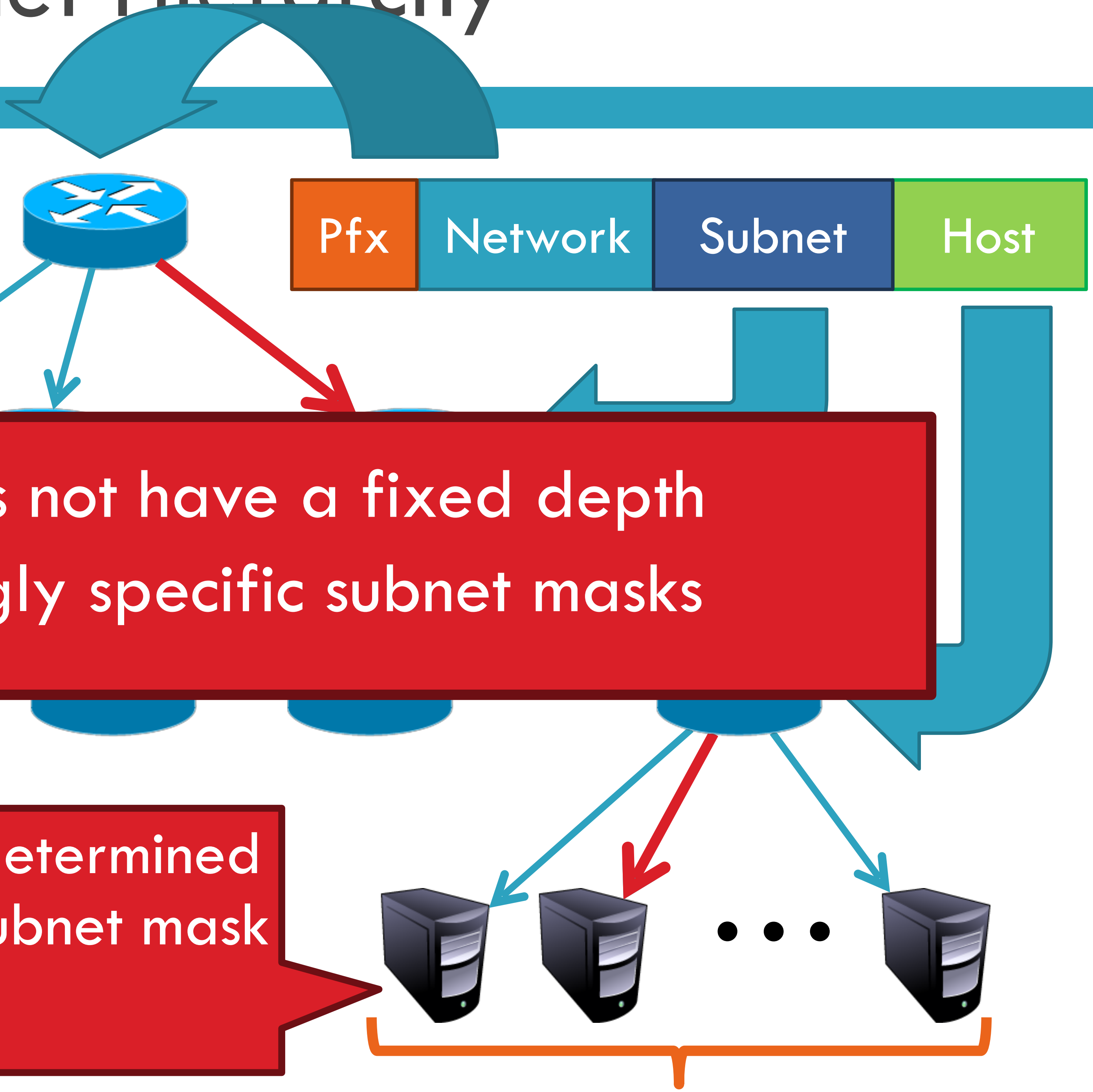
- Extract network:

IP Address:	10110101 11011101 01010100 01110010
Subnet Mask:	& 11111111 11111111 11000000 00000000
Result:	10110101 11011101 01000000 00000000

- Extract host:

IP Address:	10110101 11011101 01010100 01110010
Subnet Mask:	& ~(11111111 11111111 11000000 00000000)
Result:	00000000 00000000 00010100 01110010

N-Level Subnet Hierarchy



- Tree does not have a fixed depth
- Increasingly specific subnet masks

Subtree size determined by length of subnet mask

Example Routing Table

Address Pattern	Subnet Mask	Destination Router
0.0.0.0	0.0.0.0	Router 4
18.0.0.0	255.0.0.0	Router 2
128.42.0.0	255.255.0.0	Router 3
128.42.128.0	255.255.128.0	Router 5
128.42.222.0	255.255.255.0	Router 1

- Question: 128.42.222.198 matches four rows
  - Which router do we forward to?
- Longest prefix matching
  - Use the row with the longest number of 1's in the mask
  - This is the most specific match

Subnetting Revisited

- Question: does subnetting solve all the problems of class-based routing?

NO

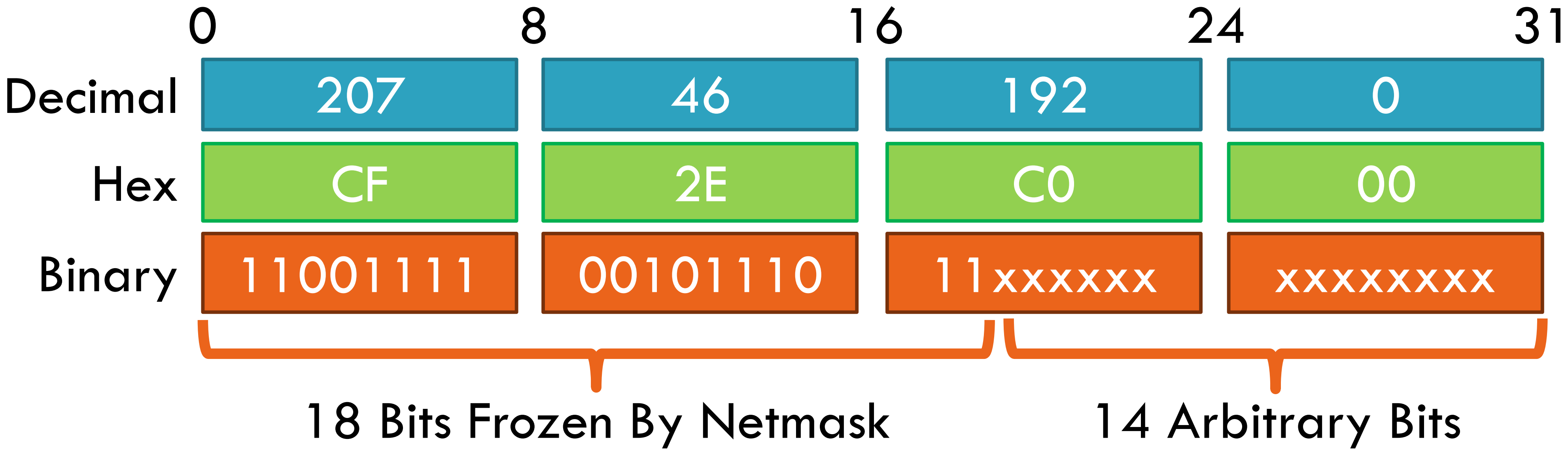
- Classes are still too coarse
  - Class A can be subnetted, but only 126 available
    - What if MIT does not use all IPs?
  - Class C is too small
  - Class B is nice, but there are only 16,398 available
- Routing tables are still too big
  - 2.1 million entries per router

Classless Inter Domain Routing

- CIDR, pronounced ‘cider’
- Key ideas:
  - Get rid of IP classes
  - Use bitmasks for all levels of routing
  - **Aggregation** to minimize FIB (forwarding information base)
- Arbitrary split between network and host
  - Specified as a bitmask or prefix length
  - Example: Rochester Institute of Technology
    - 172.31.0.0 with netmask 255.255.0.0
    - 172.31.0.0 / 16

Aggregation with CIDR

- Original use: aggregating class C ranges
- One organization given contiguous class C ranges
  - Example: Microsoft, 207.46.192.* – 207.46.255.*
  - Represents $2^6 = 64$ class C ranges
  - Specified as CIDR address 207.46.192.0/18



	0	8	16	24
Decimal	207	46	192	0
Hex	CF	2E	C0	00
Binary	11001111	00101110	11xxxxxx	xxxxxxxx

18 Bits Frozen By Netmask

14 Arbitrary Bits

Example CIDR Routing
Insert an routing entry to cover all addresses under 207.46.0.0/16

Address	Netmask	Third Byte	Byte Range
207.46.0.0	19	000xxxxx	0 – 31
207.46.32.0	19	001xxxxx	32 – 63
207.46.64.0	19	010xxxxx	64 – 95
207.46.128.0	18	10xxxxxx	128 – 191
207.46.192.0	18	11xxxxxx	192 – 255

Hole in the Routing Table: No coverage for 96 – 127
207.46.96.0/19

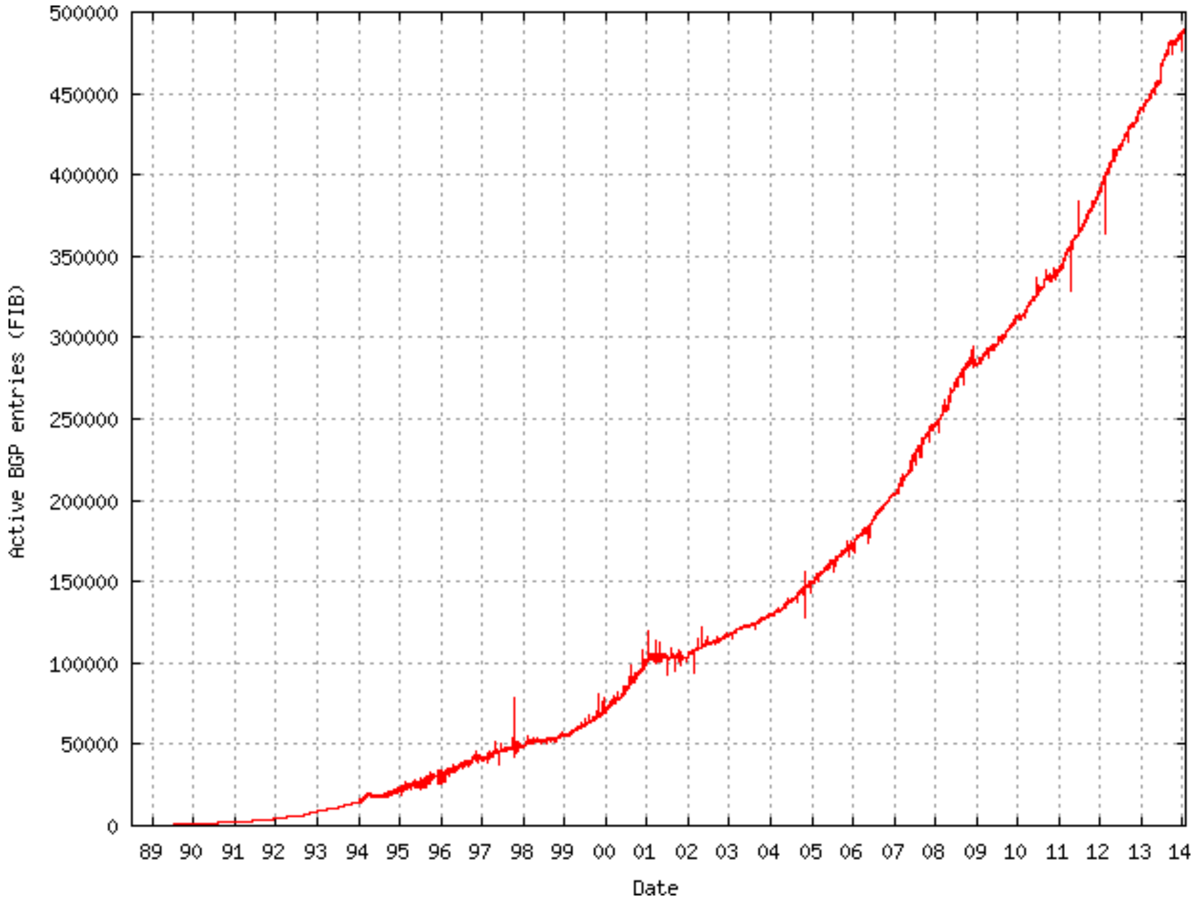
Example CIDR Routing
Insert an routing entry to cover all addresses under 207.46.0.0/16

Address	Bit-represented Address	
207.46.0.0/19	11001111.00101110.00000000.00000000/19	
207.46.32.0/19	11001111.00101110.00100000.00000000/19	
207.46.64.0/19	11001111.00101110.01000000.00000000/19	
207.46.128.0/18	11001111.00101110.10000000.00000000/18	
207.46.192.0/18	11001111.00101110.11000000.00000000/18	

Example CIDR Routing
Insert an routing entry to cover all addresses under 207.46.0.0/16

Address	Bit-represented Address	
207.46.0.0/19	11001111.00101110.00000000.00000000/19	
207.46.32.0/19	11001111.00101110.00100000.00000000/19	
207.46.64.0/19	11001111.00101110.01000000.00000000/19	
207.46.96.0/19	11001111.00101110.01100000.00000000/19	
207.46.128.0/18	11001111.00101110.10000000.00000000/18	
207.46.192.0/18	11001111.00101110.11000000.00000000/18	

Size of CIDR Routing Tables



- From www.cidr-report.org
- CIDR has kept IP routing table sizes in check
  - Currently ~450,000 entries for a complete IP routing table
  - Only required by backbone routers

Takeaways

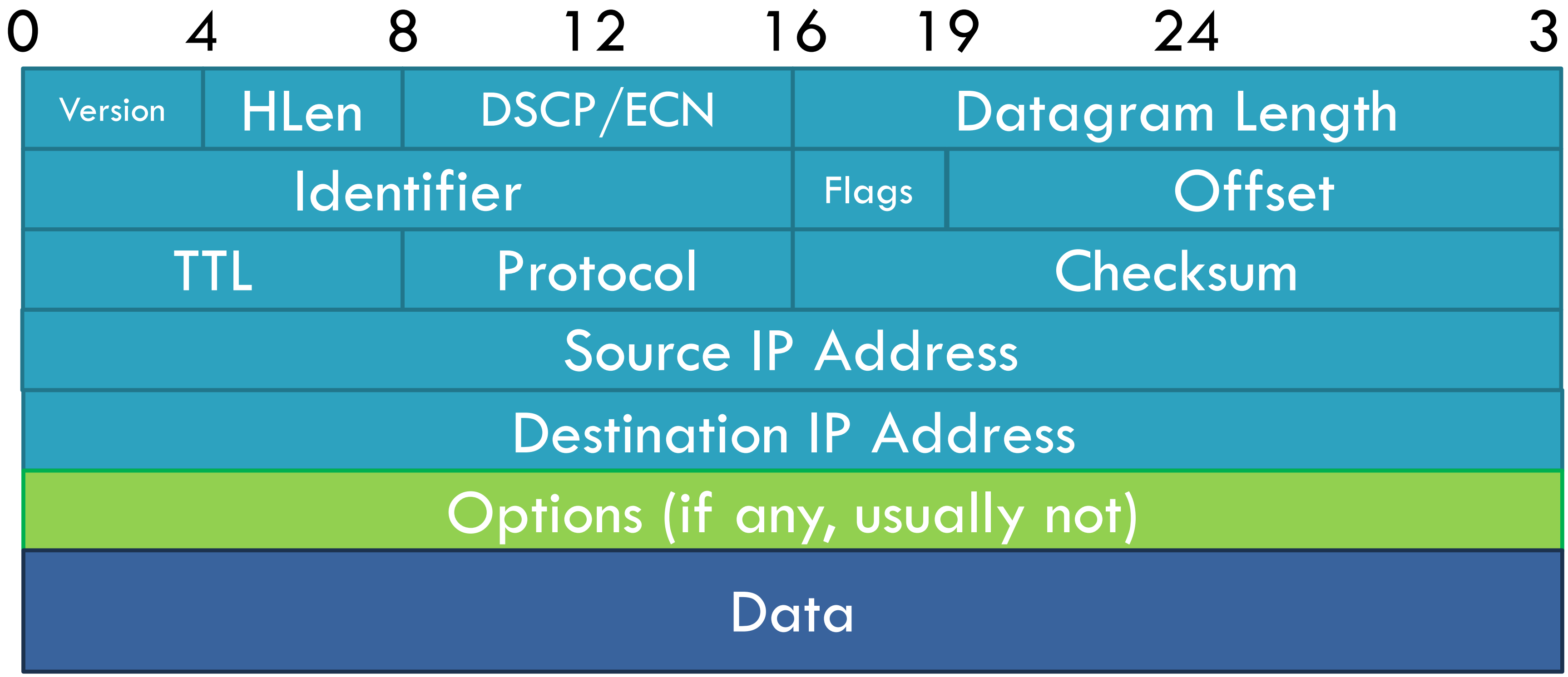
- Hierarchical addressing is critical for scalability
  - Not all routers need all information
  - Limited number of routers need to know about changes
- Non-uniform hierarchy useful for heterogeneous networks
  - Class-based addressing is too course
  - CIDR improves scalability and granularity
- Implementation challenges
  - Longest prefix matching is more difficult than schemes with no ambiguity

Outline

- Addressing
  - Class-based
  - CIDR
- IPv4 Protocol Details
  - Packed Header
  - Fragmentation
- IPv6

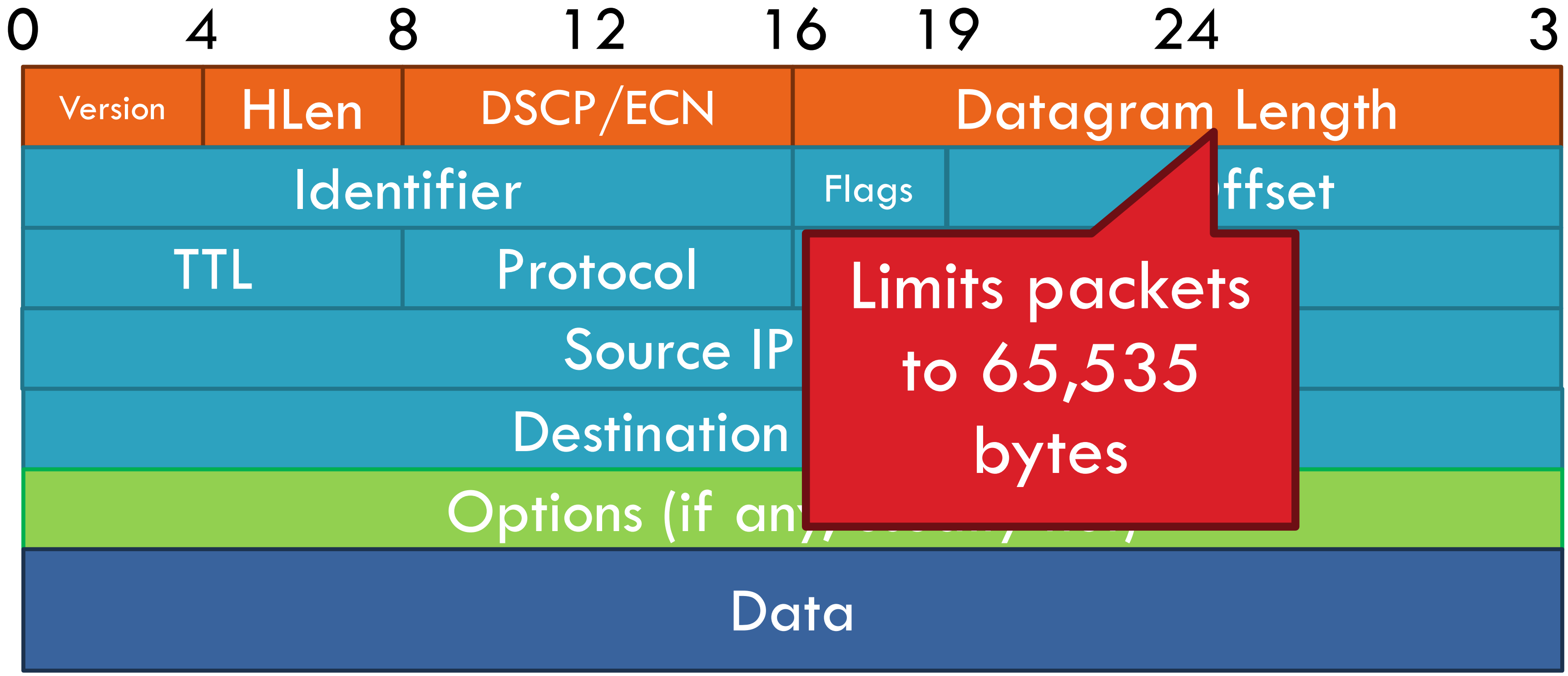
IP Datagrams

- IP Datagrams are like a letter
  - Totally self-contained
  - Include all necessary addressing information
  - No advanced setup of connections or circuits



IP Header Fields: Word 1

- Version: 4 for IPv4
- Header Length: Number of 32-bit words (usually 5)
- Type of Service: Priority information (unused)
- Datagram Length: Length of header + data in bytes



0 4 8 12 16 19 24 31

Version

HLen

DSCP/ECN

Datagram Length

Identifier

Flags

ffset

TTL

Protocol

Source IP

Destination

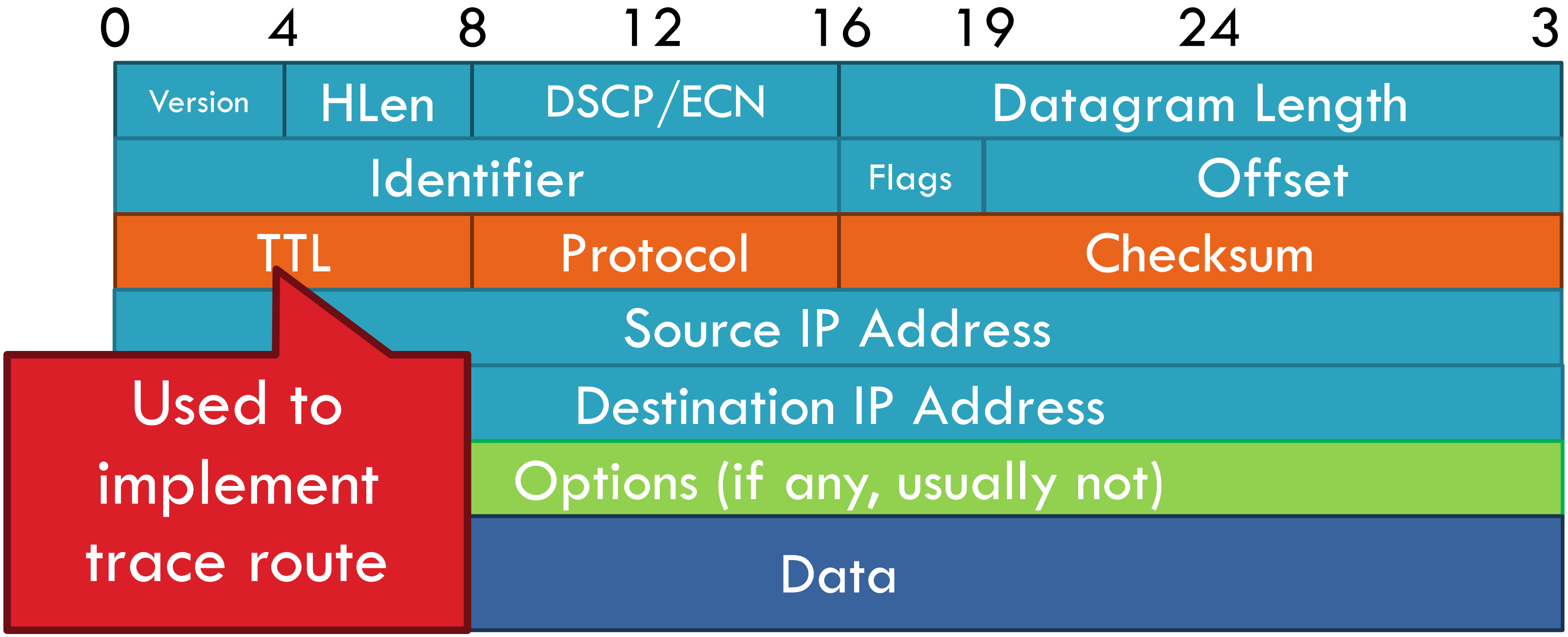
Options (if any

Data

Limits packets to 65,535 bytes

IP Header Fields: Word 3

- Time to Live: decremented by each router
  - Used to kill looping packets
- Protocol: ID of encapsulated protocol
  - 6 = TCP, 17 = UDP
- Checksum



0 4 8 12 16 19 24 31

Version | HLen | DSCP/ECN | Datagram Length

Identifier | Flags | Offset

TTL | Protocol | Checksum

Source IP Address

Destination IP Address

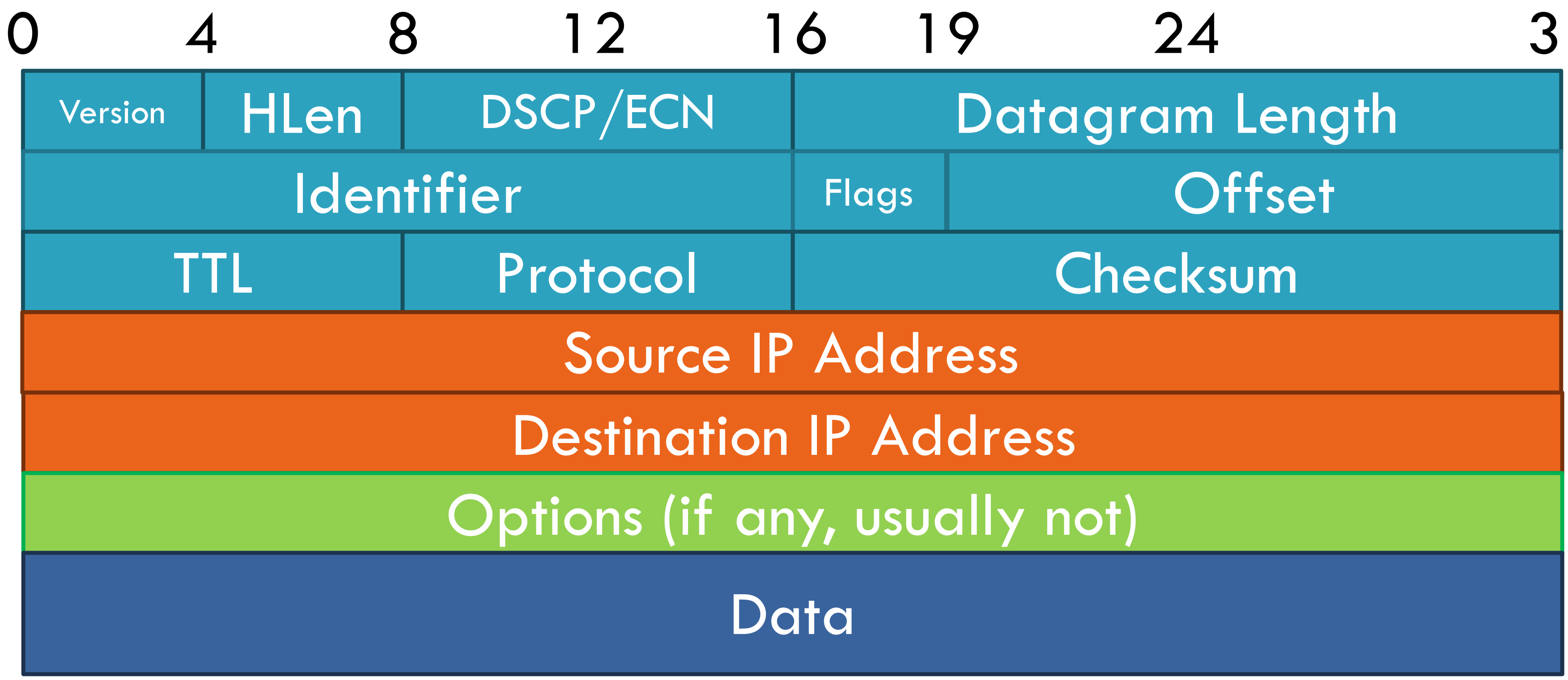
Options (if any, usually not)

Data

Used to implement trace route

IP Header Fields: Word 4 and 5

- Source and destination address
  - In theory, must be globally unique
  - In practice, this is often violated



0 4 8 12 16 19 24 31

Version | HLen | DSCP/ECN | Datagram Length

Identifier | Flags | Offset

TTL | Protocol | Checksum

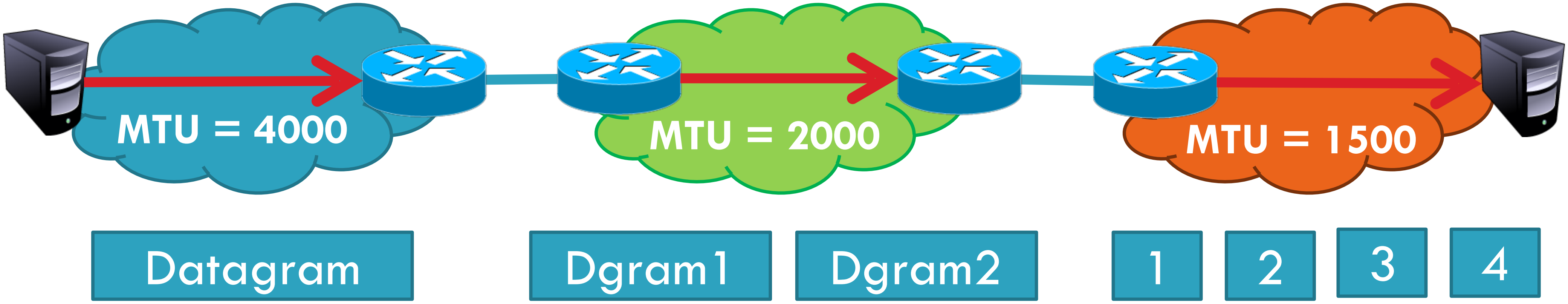
Source IP Address

Destination IP Address

Options (if any, usually not)

Data

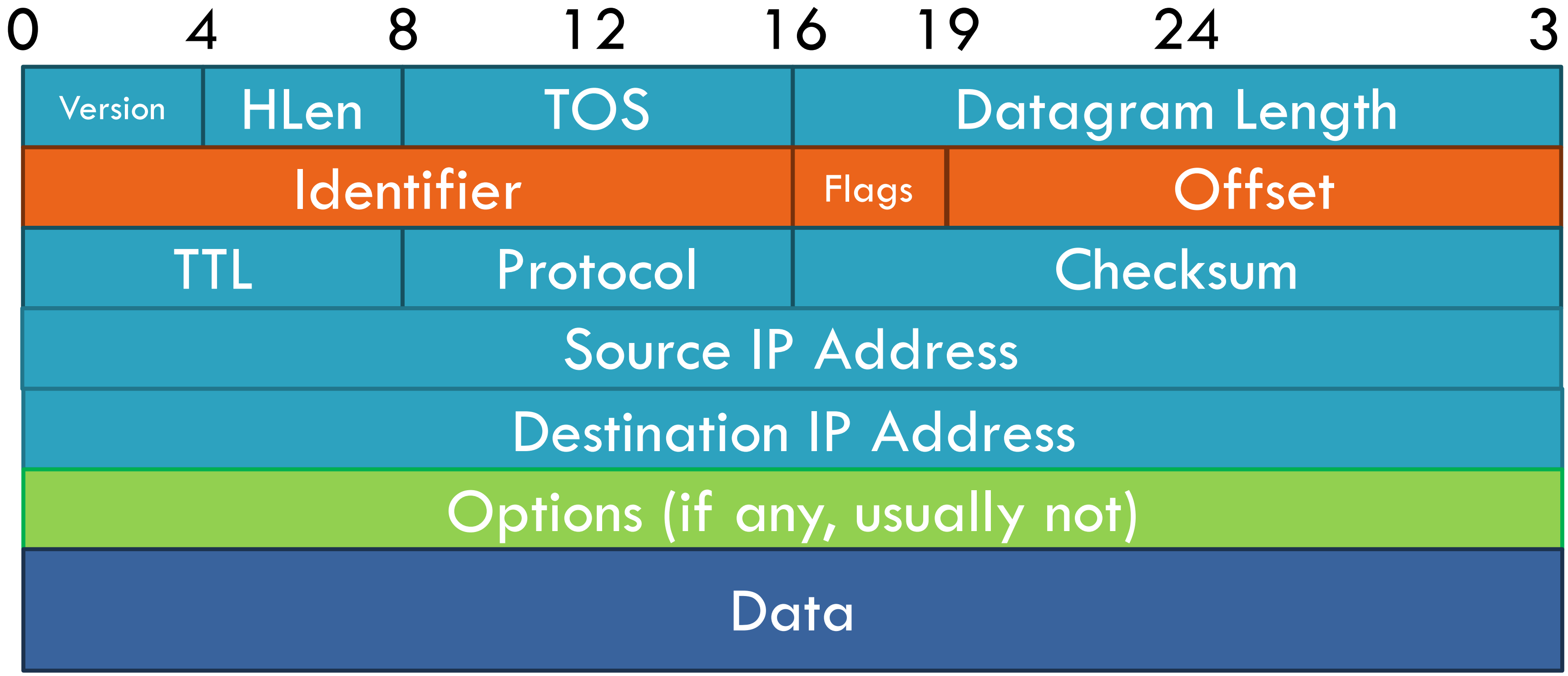
Problem: Fragmentation



- Problem: each network has its own MTU
  - DARPA principles: networks allowed to be heterogeneous
  - Minimum MTU may not be known for a given path
- IP Solution: fragmentation
  - Split datagrams into pieces when MTU is reduced
  - Reassemble original datagram at the receiver

IP Header Fields: Word 2

- Identifier: a unique number for the original datagram
- Flags: M flag (3 bits)
- Offset: byte position of the first byte in the fragment
  - Divided by 8



0 4 8 12 16 19 24 31

Version HLen TOS Datagram Length

Identifier Flags Offset

TTL Protocol Checksum

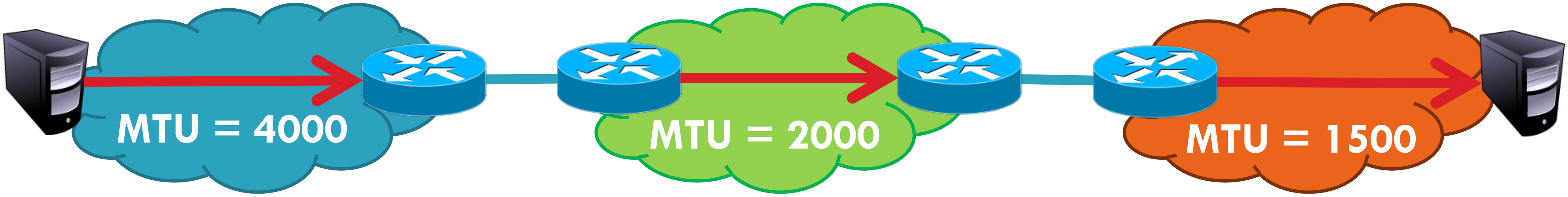
Source IP Address

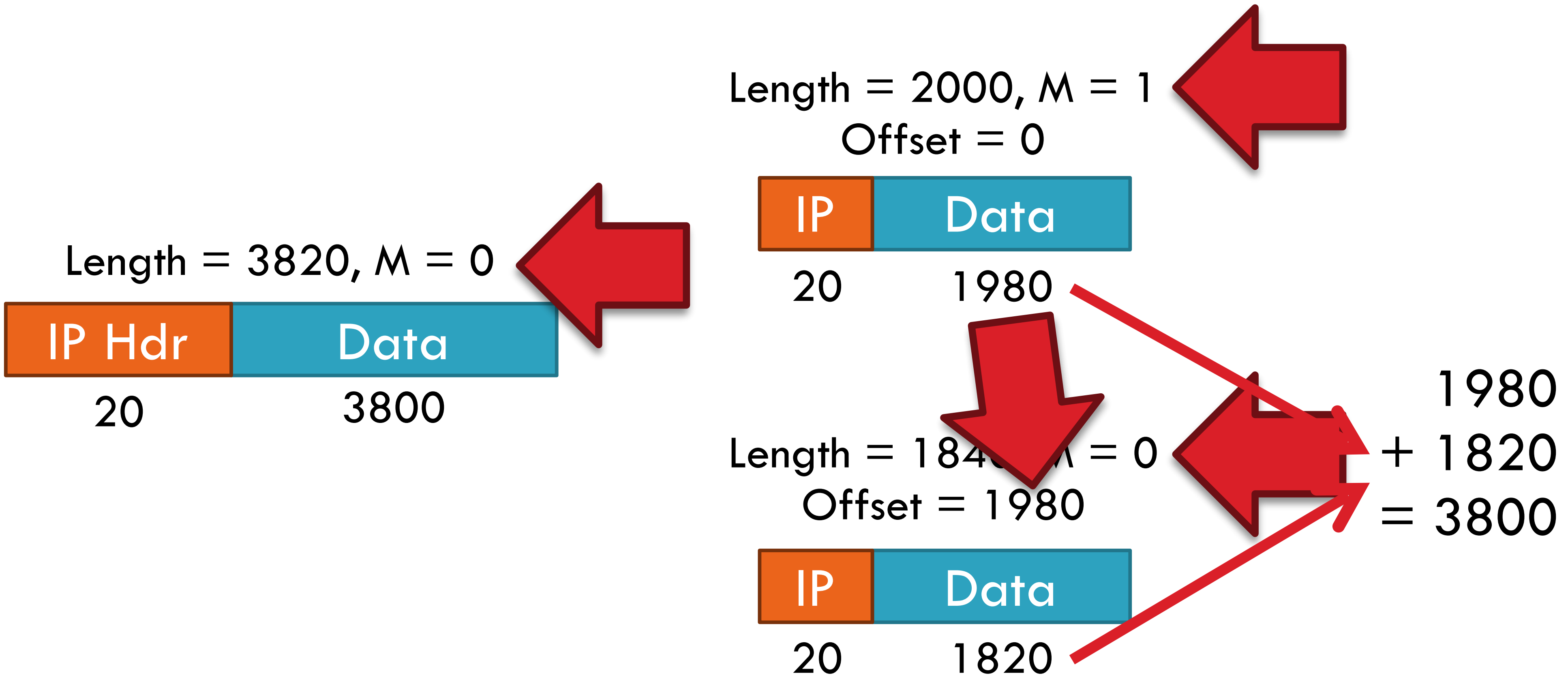
Destination IP Address

Options (if any, usually not)

Data

Fragmentation Example





Length = 3820, M = 0

IP Hdr | Data

20 3800

Length = 2000, M = 1
Offset = 0

IP | Data

20 1980

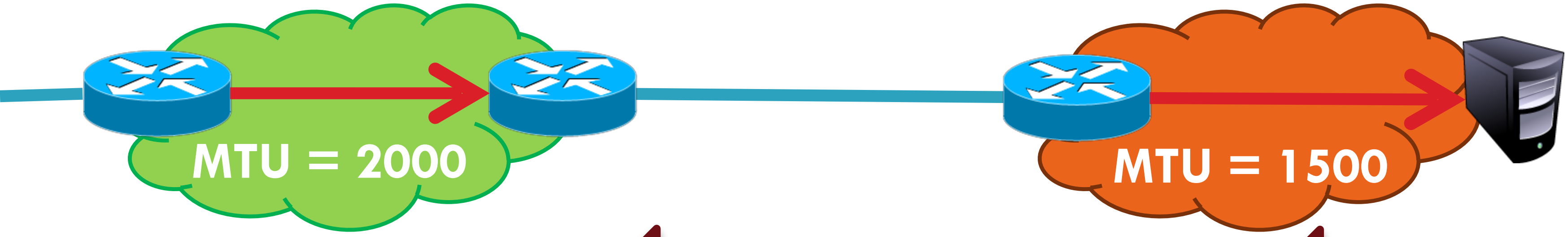
Length = 1840, M = 0
Offset = 1980

IP | Data

20 1820

1980
+ 1820
= 3800

Fragmentation Example



Length = 2000, M = 1
Offset = 0

IP | Data
20 | 1980

Length = 1840, M = 0
Offset = 1980

IP | Data
20 | 1820

ength = 1500, M = 1
Offset = 0

IP | Data
20 | 1480

Length = 52 = 1
Offset = 1480

IP | Data
20 | 500

Length = 1500, M = 1
Offset = 1980

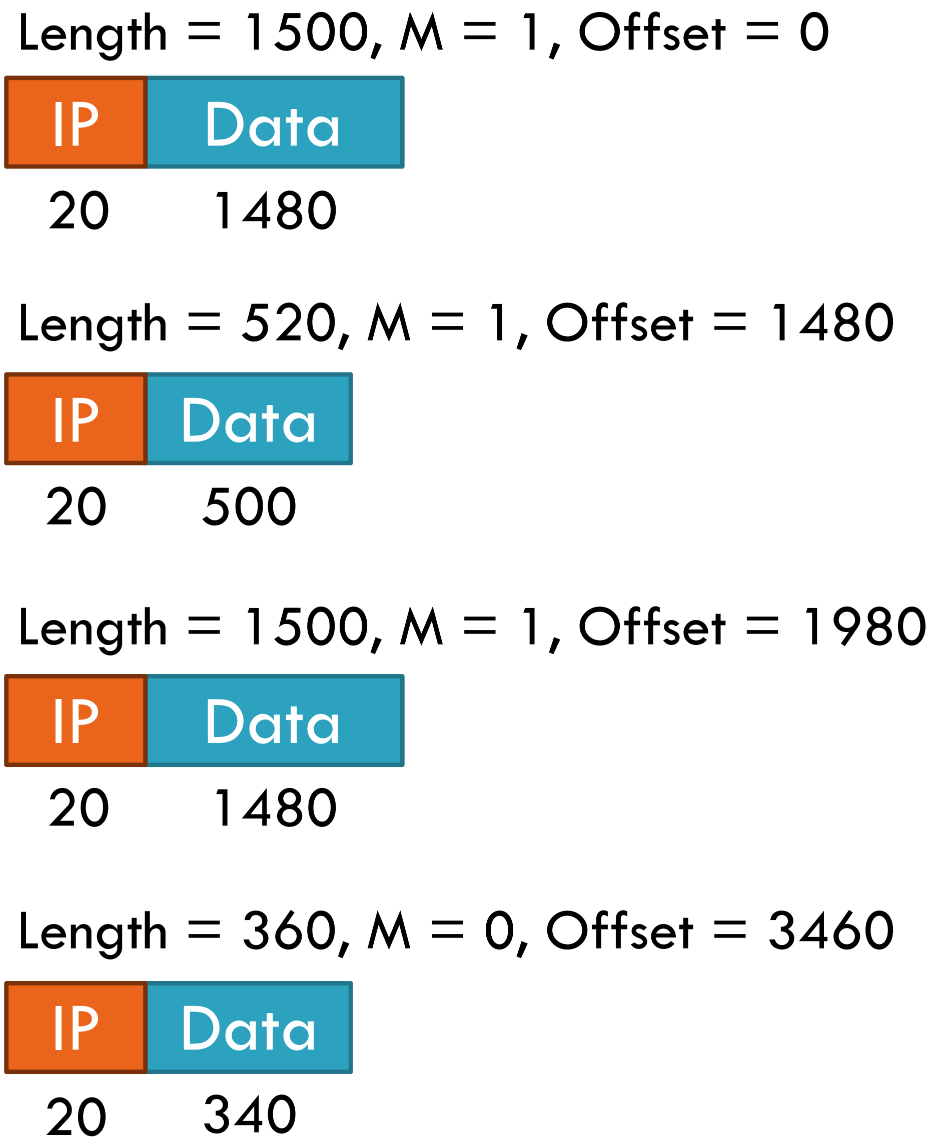
IP | Data
20 | 1480

1480
+ 500
= 1980

Length = 360, M = 0
Offset = 3460

IP | Data
20 | 340

IP Fragment Reassembly


Length = 1500, M = 1, Offset = 0

Length = 520, M = 1, Offset = 1480

Length = 1500, M = 1, Offset = 1980

Length = 360, M = 0, Offset = 3460

- Performed at destination
- M = 0 fragment gives us total data size
  - 360 – 20 + 3460 = 3800
- Challenges:
  - Out-of-order fragments
  - Duplicate fragments
  - Missing fragments
- Basically, memory management nightmare

Fragmentation Concepts

- Highlights many key Internet characteristics
  - Decentralized and heterogeneous
    - Each network may choose its own MTU
  - Connectionless datagram protocol
    - Each fragment contains full routing information
    - Fragments can travel independently, on different paths
  - Best effort network
    - Routers/receiver may silently drop fragments
    - No requirement to alert the sender
  - Most work is done at the endpoints
    - i.e. reassembly

Fragmentation in Reality

- Fragmentation is expensive
  - Memory and CPU overhead for datagram reconstruction
  - Want to avoid fragmentation if possible
- MTU discovery protocol
  - Send a packet with “don’t fragment” bit set
  - Keep decreasing message length until one arrives
  - May get “can’t fragment” error from a router, which will explicitly state the supported MTU
- Router handling of fragments
  - Fast, specialized hardware handles the common case
  - Dedicated, general purpose CPU just for handling fragments
- Examples
  - Censorship?

Outline

- Addressing
  - Class-based
  - CIDR
- IPv4 Protocol Details
  - Packed Header
  - Fragmentation
- IPv6

The IPv4 Address Space Crisis

- Problem: the IPv4 address space is too small
  - 2^{32} = 4,294,967,296 possible addresses
  - Less than one IP per person
- Parts of the world have already run out of addresses
  - IANA assigned the last /8 block of addresses in 2011

Region	Regional Internet Registry (RIR)	Exhaustion Date
Asia/Pacific	APNIC	April 19, 2011
Europe/Middle East	RIPE	September 14, 2012
North America	ARIN	September 24, 2015
South America	LACNIC	June 10, 2014
Africa	AFRINIC	2019 (Projected)

IPv6

- IPv6, first introduced in 1998(!)
  - 128-bit addresses
  - 4.8 * 10^{28} addresses per person
- Address format
  - 8 groups of 16-bit values, separated by ':'
  - Leading zeroes in each group may be omitted
  - Groups of zeroes can be omitted using '::'

2001:0db8:0000:0000:0000:ff00:0042:8329

2001:0db8:0:0:0:ff00:42:8329

2001:0db8::ff00:42:8329

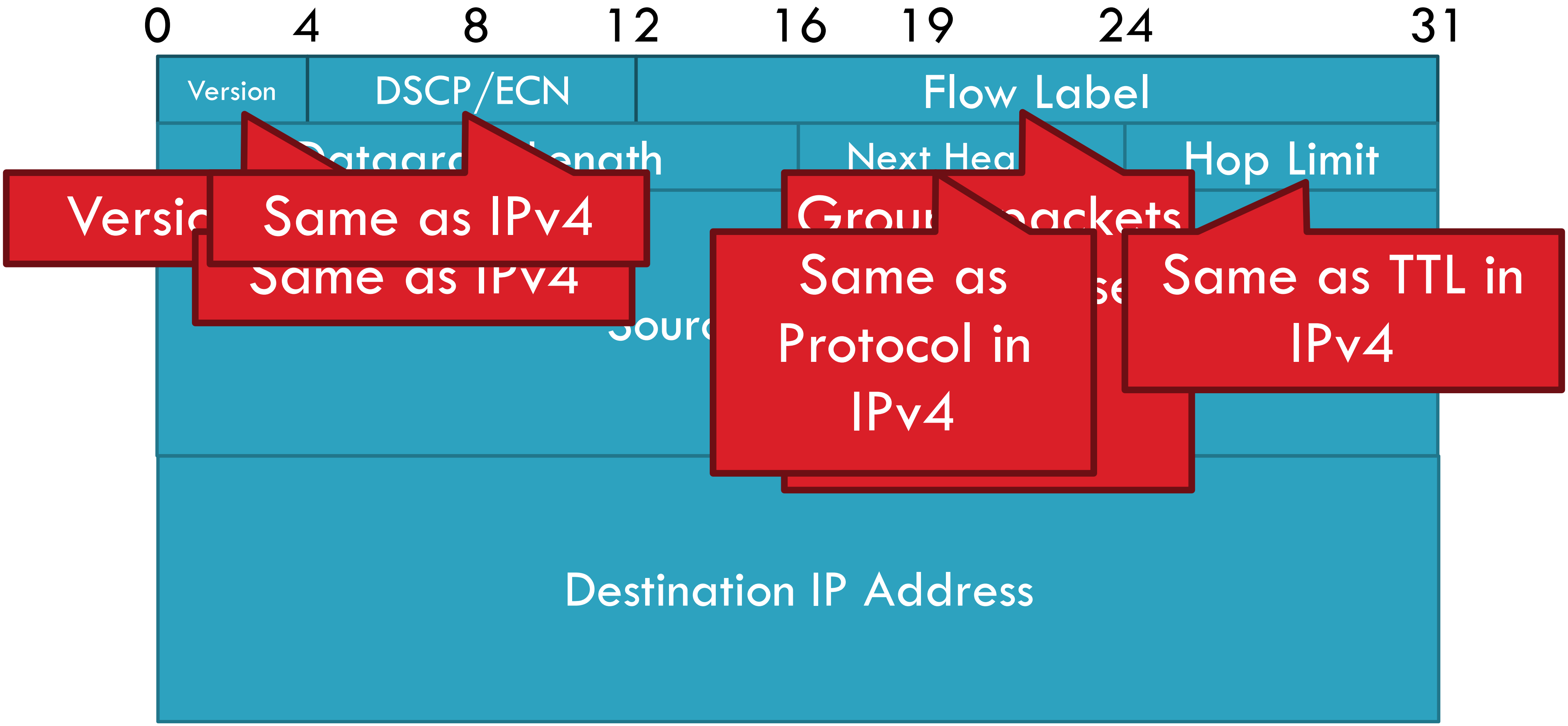
IPv6 Trivia

- Who knows the IP for localhost?
  - 127.0.0.1

- What is localhost in IPv6?
  - ::1

IPv6 Header

- Double the size of IPv4 (320 bits vs. 160 bits)



0 4 8 12 16 19 24 31

Version

DSCP/ECN

Flow Label

Datagra Length

Next Hea

Hop Limit

Sourc

Destination IP Address

Versio

Same as IPv4

Same as IPv4

Grou packets

se

Same as Protocol in IPv4

Same as TTL in IPv4

Differences from IPv4 Header

- Several header fields are missing in IPv6
  - Header length – rolled into Next Header field
  - Checksum – was useless, so why keep it
  - Identifier, Flags, Offset
    - IPv6 routers do not support fragmentation
    - Hosts are expected to use path MTU discovery
- Reflects changing Internet priorities
  - Today's networks are more homogeneous
  - Instead, routing cost and complexity dominate
- No security vulnerabilities due to IP fragments

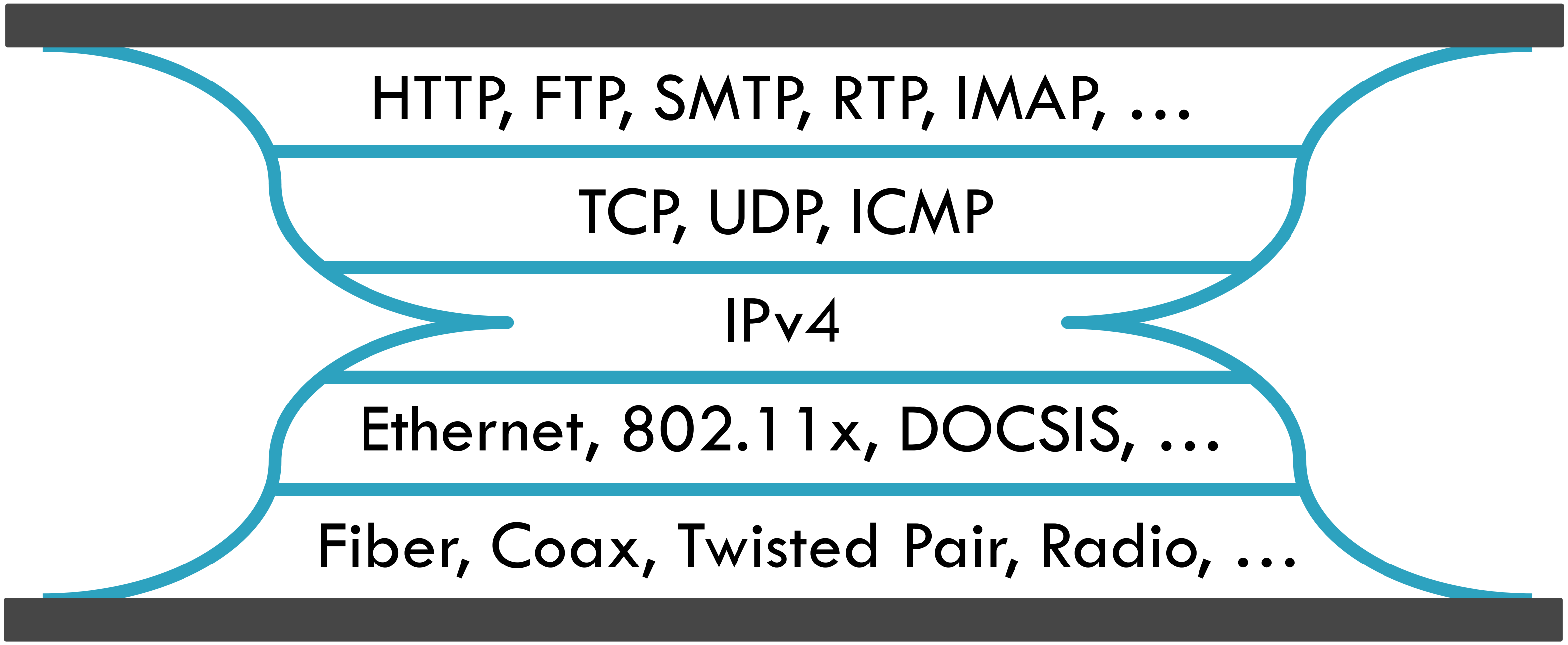
Performance Improvements

- No checksums to verify
- No need for routers to handle fragmentation
- Simplified routing table design
  - Address space is huge
  - No need for CIDR (but need for aggregation)
  - Standard subnet size is 2^{64} addresses
- Simplified auto-configuration
  - Neighbor Discovery Protocol
  - Used by hosts to determine network ID
  - Host ID can be random!

Additional IPv6 Features

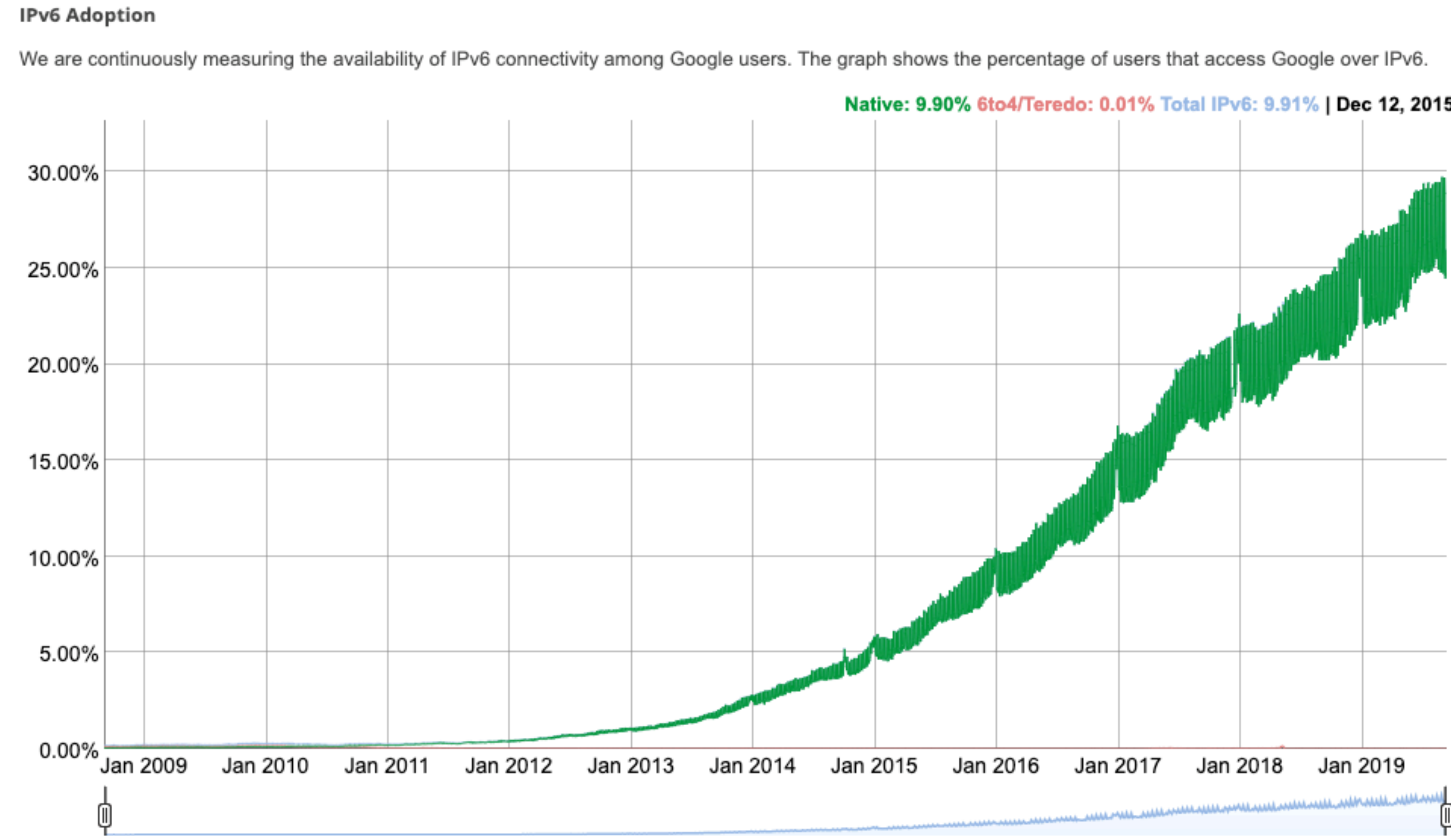
- Source Routing
  - Host specifies the route to wants packet to take
- Mobile IP
  - Hosts can take their IP with them to other networks
  - Use source routing to direct packets
- Privacy Extensions
  - Randomly generate host identifiers
  - Make it difficult to associate one IP to a host
- Jumbograms
  - Support for 4Gb datagrams

Deployment Challenges


HTTP, FTP, SMTP, RTP, IMAP, …
TCP, UDP, ICMP
IPv4
Ethernet, 802.11x, DOCSIS, …
Fiber, Coax, Twisted Pair, Radio, …

- Switching to IPv6 is a whole-Internet upgrade
  - All routers, all hosts
  - ICMPv6, DHCPv6, DNSv6
- 2013: 0.94% of Google traffic was IPv6

https://www.google.com/intl/en/ipv6/statistics.html

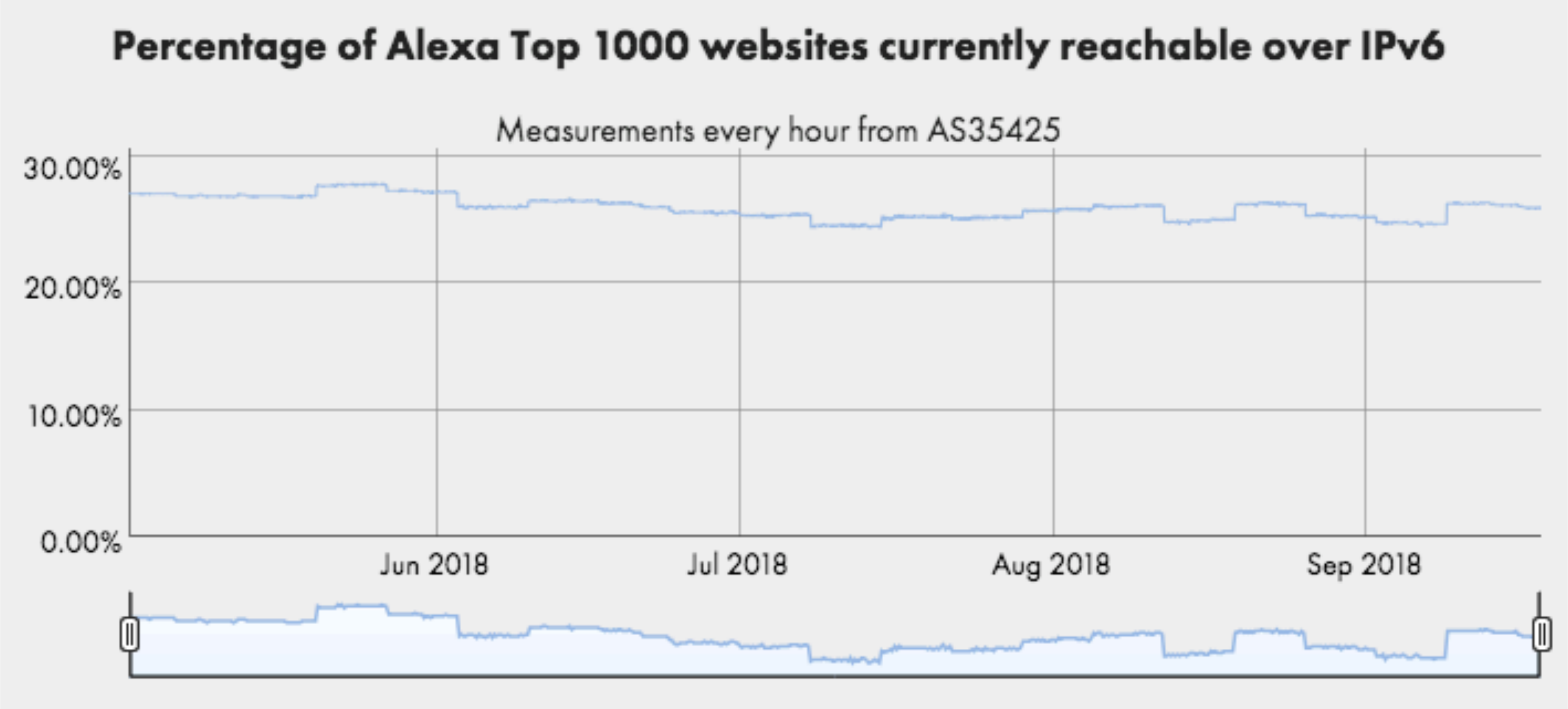


IPv6 Adoption

We are continuously measuring the availability of IPv6 connectivity among Google users. The graph shows the percentage of users that access Google over IPv6.

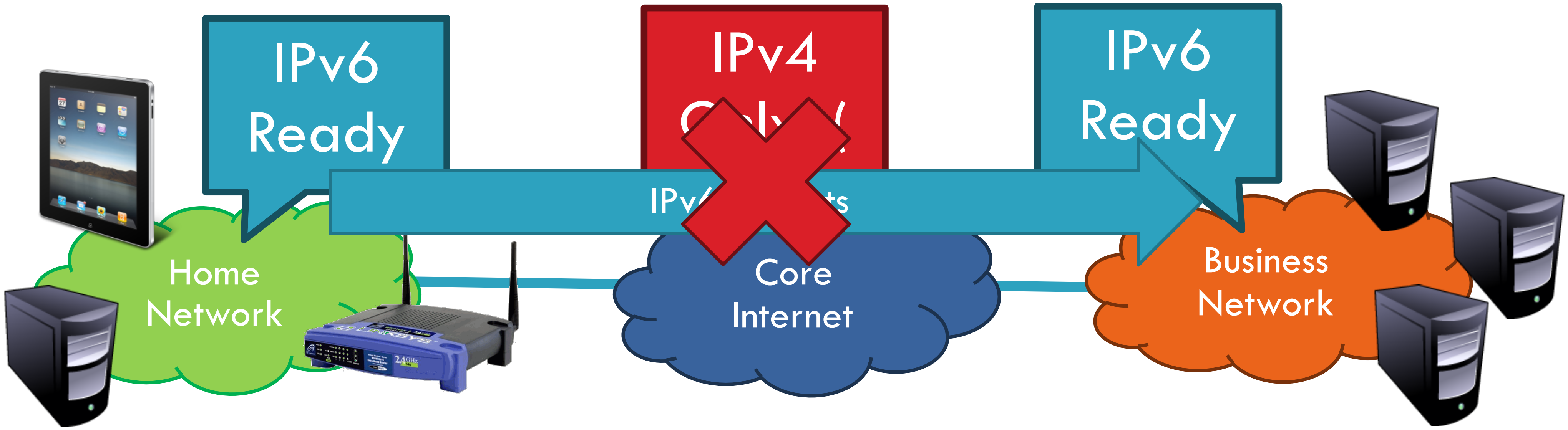
Native: 9.90% 6to4/Teredo: 0.01% Total IPv6: 9.91% | Dec 12, 2015

Server-side reachability



Transitioning to IPv6

- How do we ease the transition from IPv4 to IPv6?
  - Today, most network edges are IPv6 ready
    - Windows/OSX/iOS/Android all support IPv6
    - Your wireless access point probably supports IPv6
  - The Internet core is hard to upgrade
  - … but a IPv4 core cannot route IPv6 traffic

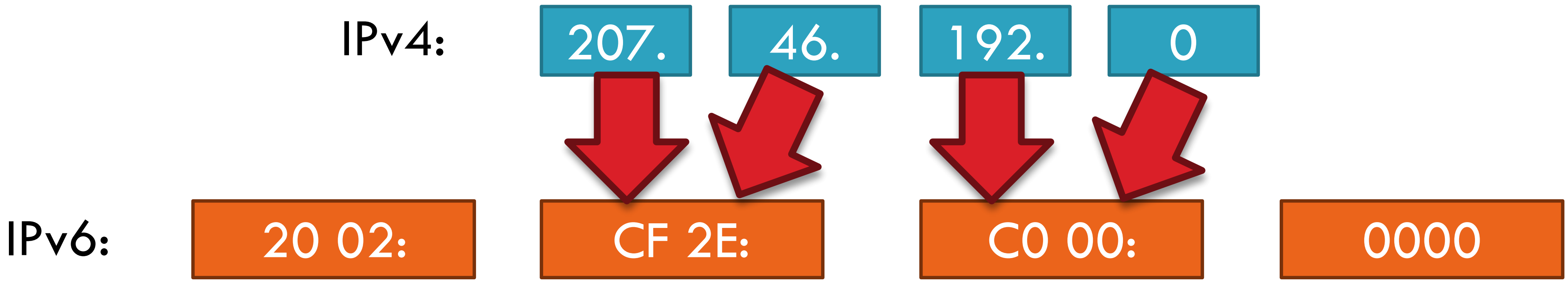


Transition Technologies

- How do you route IPv6 packets over an IPv4 Internet?
- Transition Technologies
  - Use tunnels to encapsulate and route IPv6 packets over the IPv4 Internet
  - Several different implementations
    - 6to4
    - IPv6 Rapid Deployment (6rd)
    - Teredo
    - … etc.

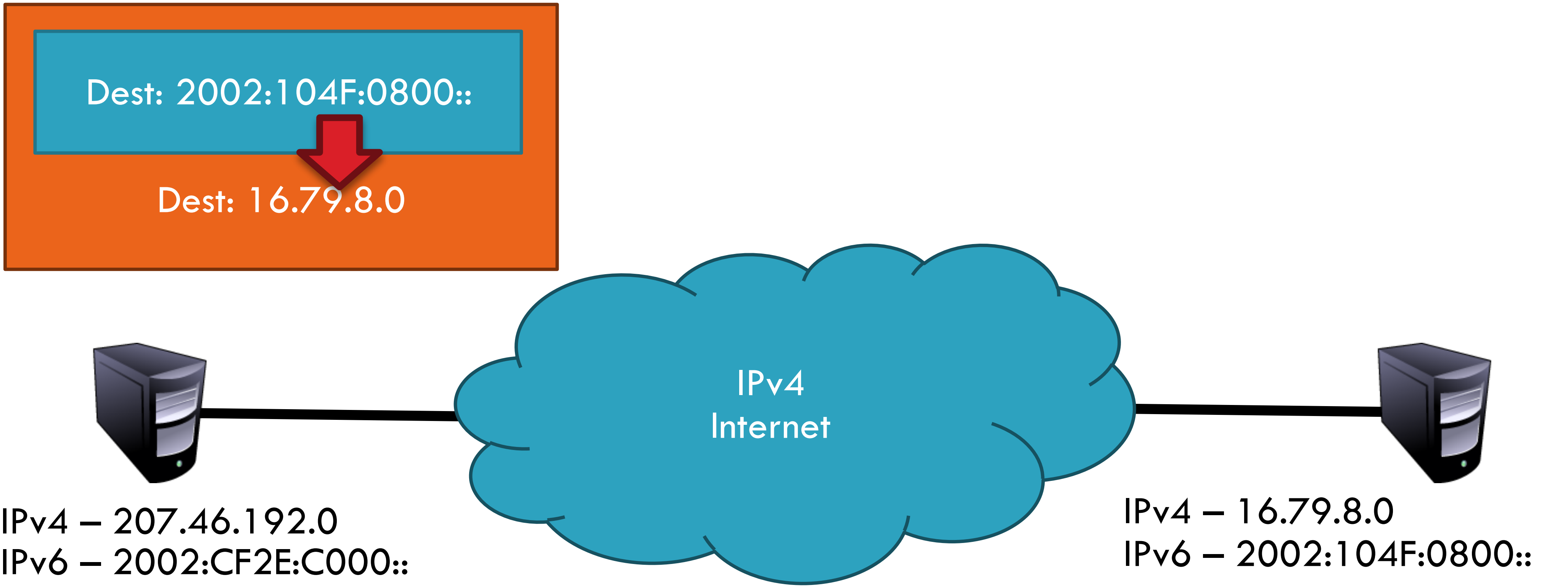
6to4 Basics

- Problem: you've been assigned an IPv4 address, but you want an IPv6 address
  - Your ISP can't or won't give you an IPv6 address
  - You can't just arbitrarily choose an IPv6 address
- Solution: construct a 6to4 address
  - 6to4 addresses always start with 2002::
  - Embed the 32-bit IPv4 inside the 128-bit IPv6 address

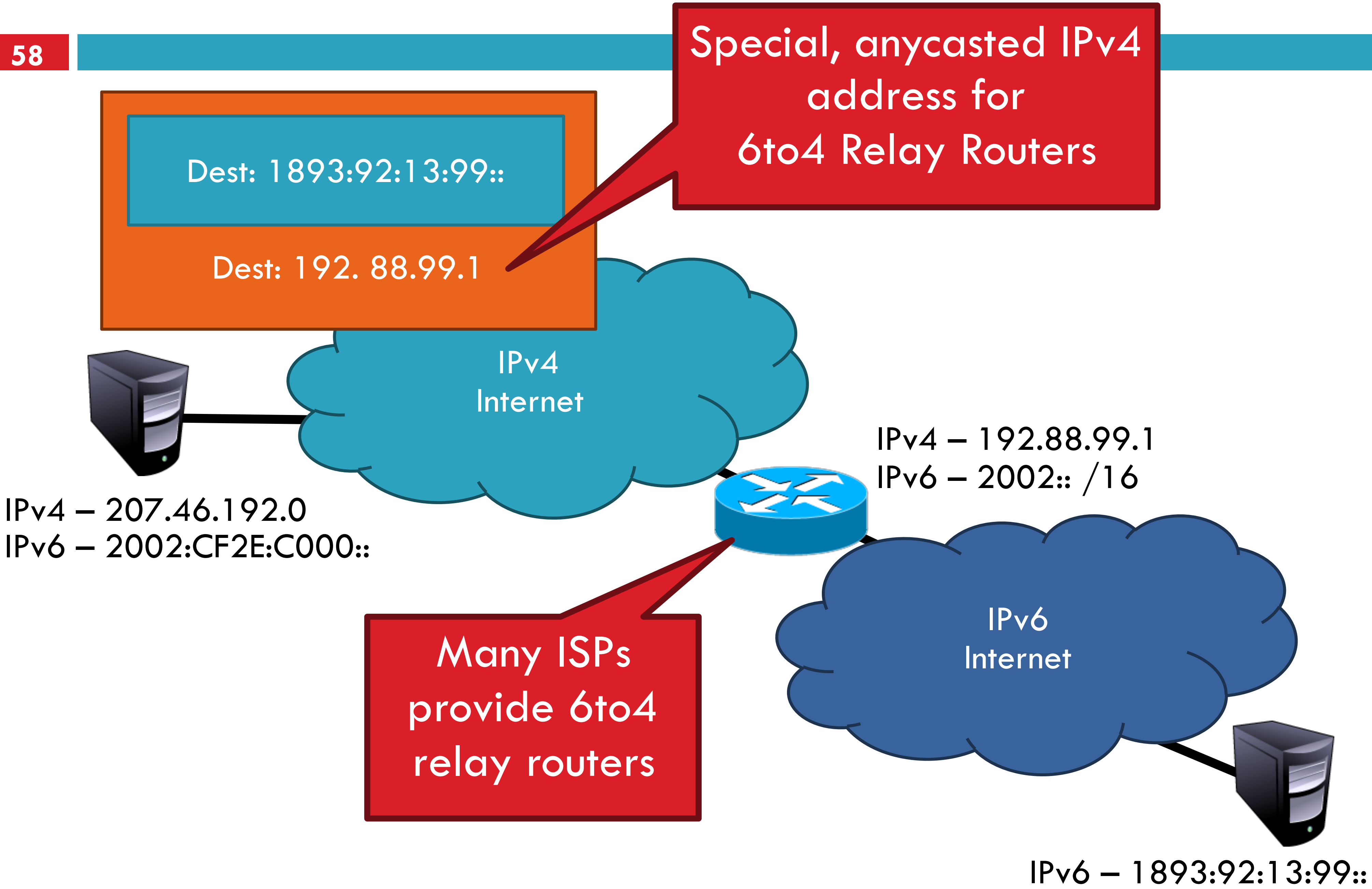


Routing from 6to4 to 6to4

- How does a host using 6to4 send a packet to another host using 6to4?



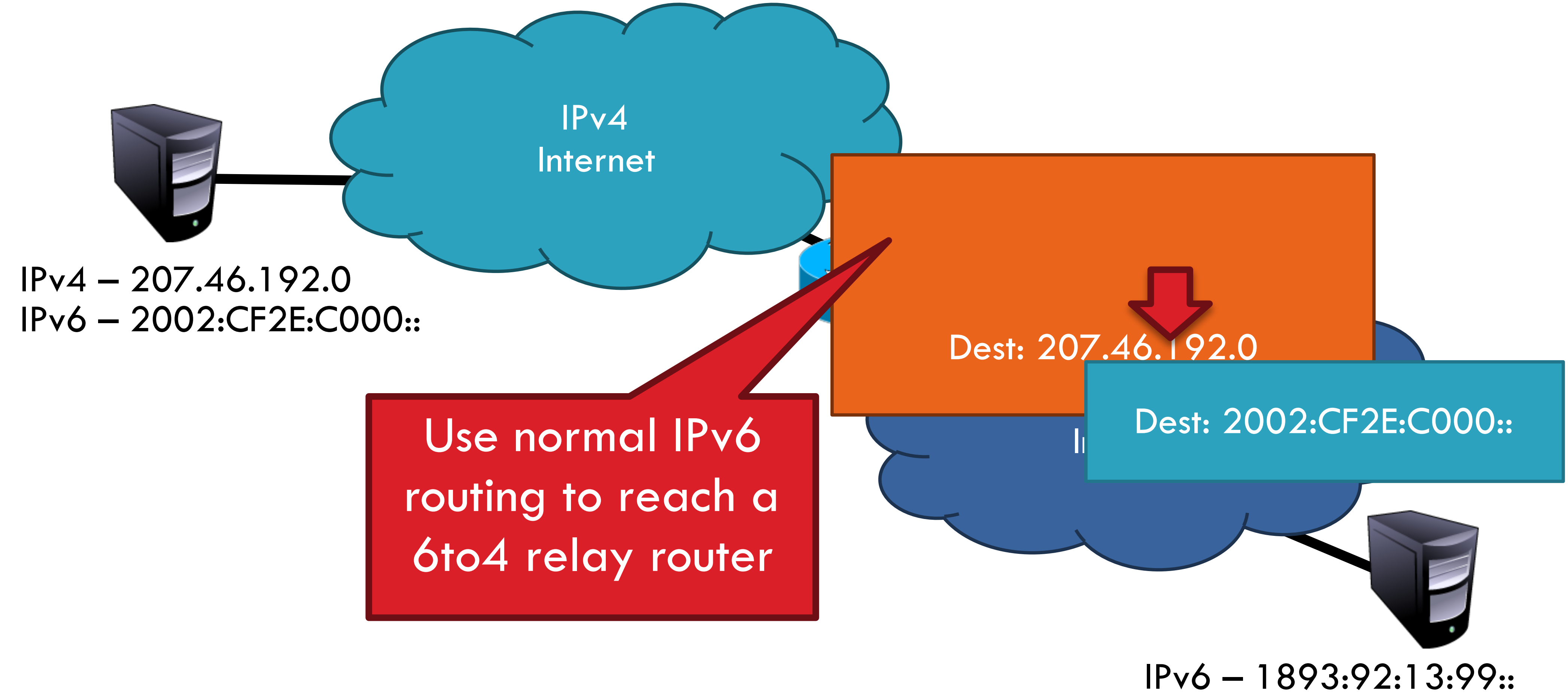
Routing from 6to4 to Native IPv6



Special, anycasted IPv4 address for 6to4 Relay Routers

Many ISPs provide 6to4 relay routers

Routing from Native IPv6 to 6to4



Use normal IPv6 routing to reach a 6to4 relay router

Problems with 6to4

- Uniformity
  - Not all ISPs have deployed 6to4 relays
- Quality of service
  - Third-party 6to4 relays are available
  - …but, they may be overloaded or unreliable
- Reachability
  - 6to4 doesn't work if you are behind a NAT
- Possible solutions
  - IPv6 Rapid Deployment (6rd)
    - Each ISP sets up relays for its customers
    - Does not leverage the 2002:: address space
  - Teredo
    - Tunnels IPv6 packets through UDP/IPv4 tunnels
    - Can tunnel through NATs, but requires special relays

Consequences of IPv6

- Beware unintended consequences of IPv6
- Example: IP blacklists
  - Currently, blacklists track IPs of spammers/bots
  - Few IPv4 addresses mean list sizes are reasonable
  - Hard for spammers/bots to acquire new IPs
- Blacklists will not work with IPv6
  - Address space is enormous
  - Acquiring new IP addresses is trivial