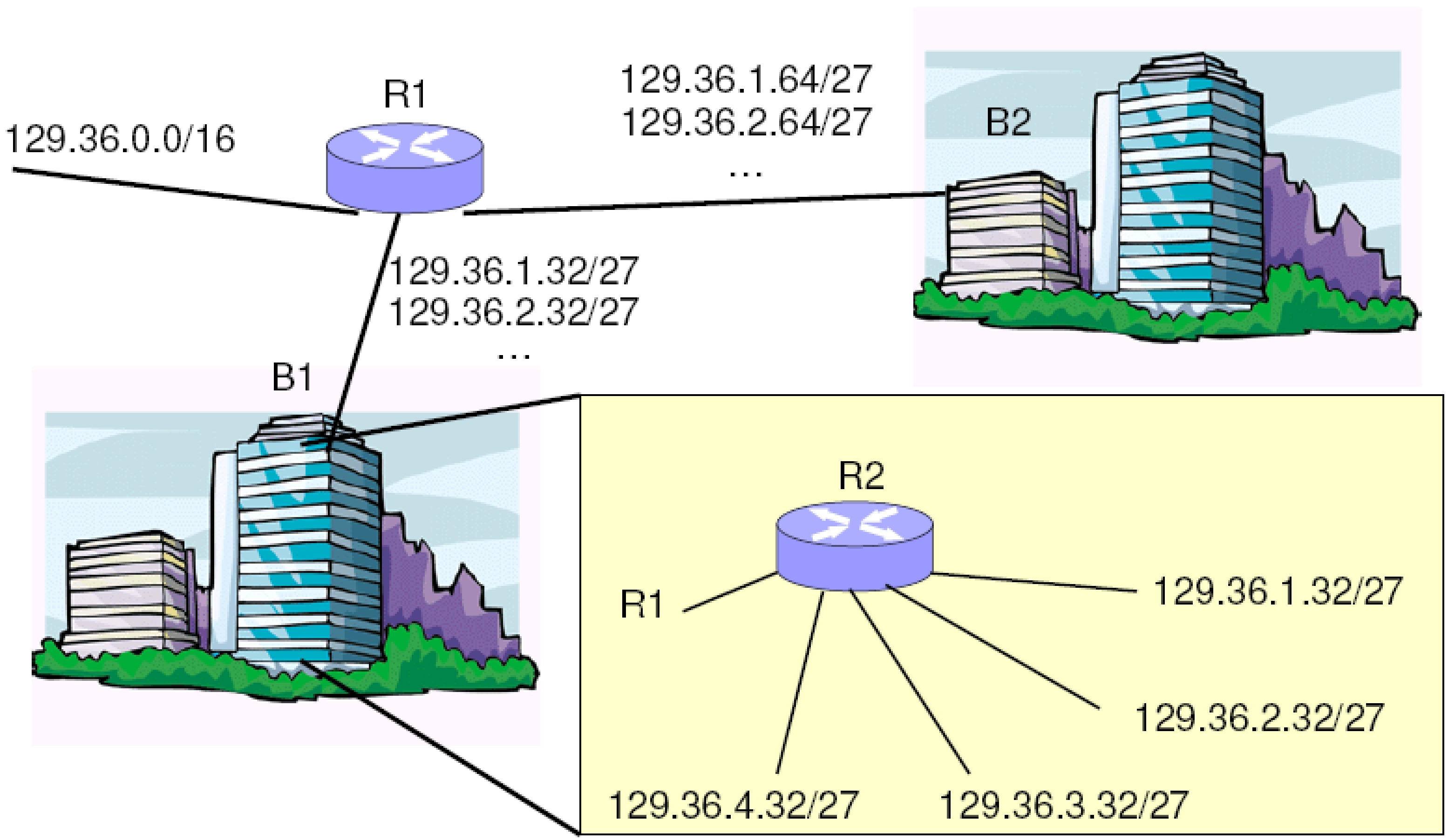
EE 542

Lecture 5: The Internet Networking Methods and Usage

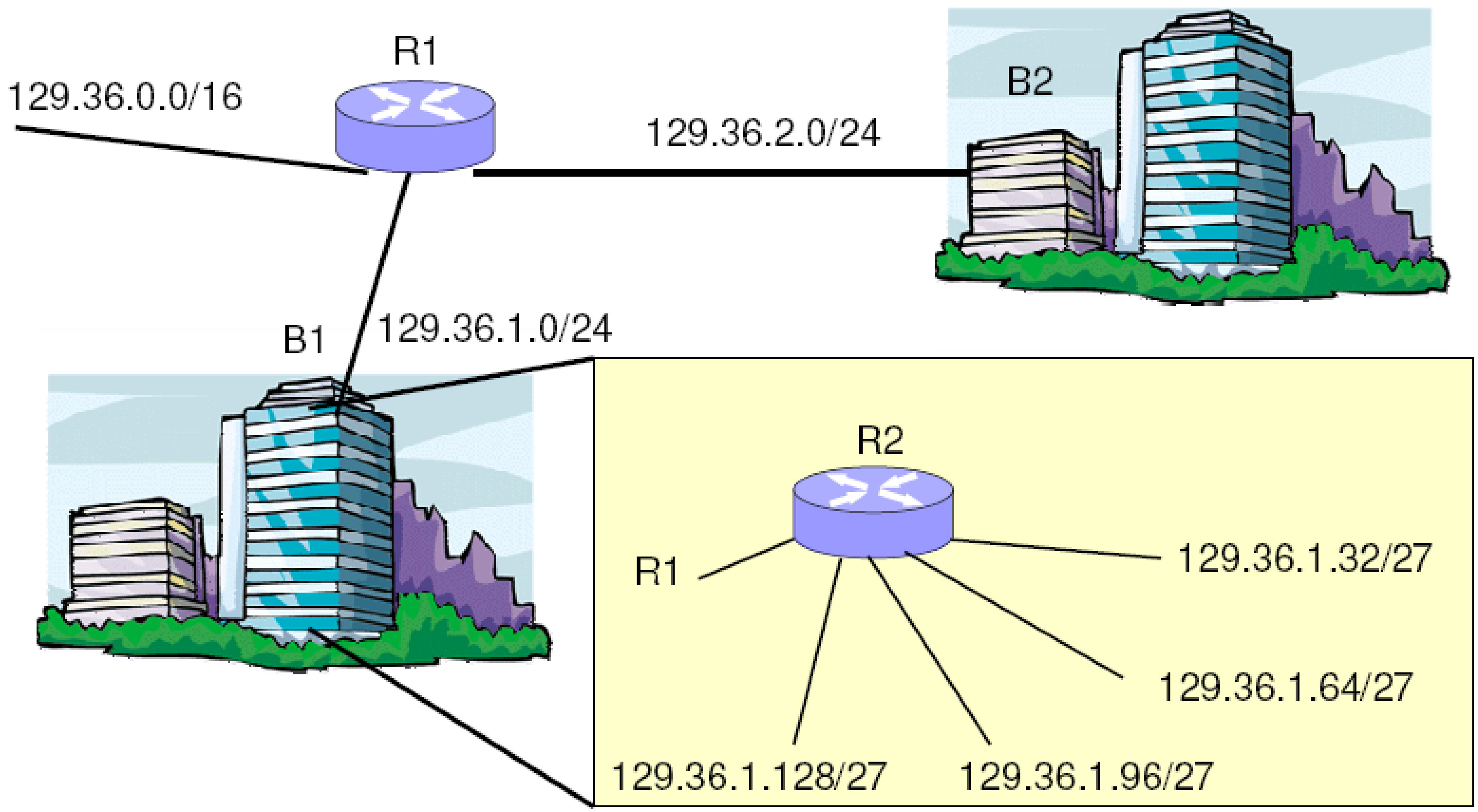
Internetworking and Cloud Computing

Young Cho
Department of Electrical Engineering
University of Southern California

Address Distribution



Route Aggregation



Longest Prefix Match

- Allow more specific entries to supersede more general ones
  - 128.42.8/24
    - Route this traffic to Italy
  - 128.42/16
    - Route this traffic to Houston
    - Except for addresses that match a route with a longer prefix (i.e., 128.42.8/24)
- Allows significantly more route aggregation
- Simplifies things if companies move (physically or to another ISP) their block of IP addresses

Longest Prefix Match

- Software Methods
  - Hashing
  - Predictive
  - Mostly not used
  - May make a come back for IPv6
- Ternary CAM
  - Use "don't care" for bits outside the prefix
  - All matching prefixes will yield a hit
  - Pick the one with the longest prefix
- Simple Improvement
  - Sort CAM entries based on decreasing prefix length (use lowest matching entry)
- Also does not scale for larger tables

IP Routers

- IP routers determine next hop using LPM
  - Given a destination IP address, what is the IP address of the router to which the datagram should be forwarded next?
- Why do they need the next hop IP address?
  - Destination IP address remains unchanged in the datagram
  - Actually need the next hop *Ethernet address*
- Review
  - Ethernet address
    - Unique identifier for a piece of hardware
  - IP address
    - Unique identifier for a system performing a function
- Address Resolution Protocol

Address Resolution Protocol

- Find link layer using network layer
  - i.e., what is the Ethernet address for a given IP address?
- Every IP hosts/routers has an ARP table
  - Map IP to Ethernet addresses on their LAN
  - May be incomplete
  - Can include both static and dynamic entries

Static ARP Entries

- Example
  - IP address 74.23.121.45 has Ethernet address 00:E0:81:A4:23:EF
- Must be managed by the sysadmin
  - Actively managing IP to Ethernet address mappings for all nodes in a LAN would be difficult
  - What if that NIC fails and is replaced?
  - What if that system's IP address is changed?

Dynamic ARP Entries

- Systems "discover" IP Ethernet address mappings, as needed
- Each entry has an IP address, an Ethernet address, and a timeout (typically around 20 minutes)
- ARP packets are broadcast on the LAN to discover mappings
  - ARP packets are encapsulated in Ethernet frames

ARP Request

Destination MAC Address (ff:ff:ff:ff:ff:ff)		
Destination MAC Address	Source MAC Address	
Source MAC Address		
Type (0x0806)	HW Type (Ethernet: 0x0001)	
Protocol Type (IP: 0x0800)	HW AddrLen (6)	Prot AddrLen (4)
Opcode (Request: 1)	Source HW Address	
Source HW Address		
Source IP Address		
Destination HW Address		
Destination HW Address	Destination IP Address	
Destination IP Address	Padding	
Ethernet CRC		

Sending a Packet from a Host

- Host setup
  - IP address
  - Subnet – what IP addresses are on the same LAN
  - Gateway – where to send traffic outside the LAN
- Destination on LAN
  - Create ARP request for destination IP
  - Broadcast to everyone on the LAN
  - Destination should reply with its MAC address
- Destination not on LAN
  - Create ARP request for gateway IP
  - Broadcast to everyone on the LAN
  - Gateway should reply with its MAC address

Learning MAC Addresses

- Hosts learn IP to Ethernet address map
  - ARP responses are stored in ARP tables
  - ARP requests are stored in ARP tables (whether the host is the target or not!)
- ARP entries time out
  - Allow machines to change IP and/or MAC addresses transparently
  - Eliminate stale entries (machines turn off, move, crash, etc.)

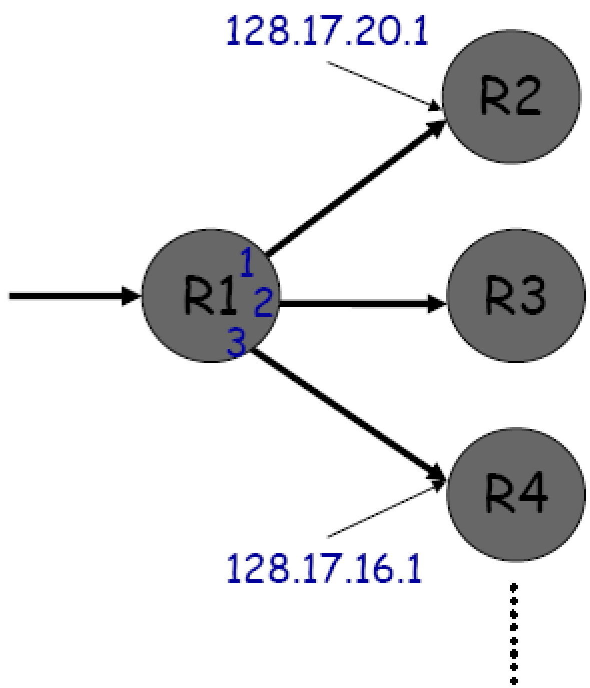
Forwarding a Packet in a Router

- Lookup dest IP address in forwarding table
  - Yields a next hop port and IP address
  - If not found, drop packet
- Lookup packet DA in forwarding table.
  - If known, forward to correct port.
  - If not found (in particular: no default router), drop packet
- Decrement TTL, update heade Checksum.
- Forward packet to outgoing interface
- Transmit packet onto link

Forwarding Decision

1. Determine the network prefix of the destination
2. Use own address and subnet mask if on the same network
3. If found, immediate destination = final destination
4. Else, use routing table to find immediate destination
5. Use ARP to find datalink (MAC) address
6. Send packet over to datalink immediate destination

Routing Table

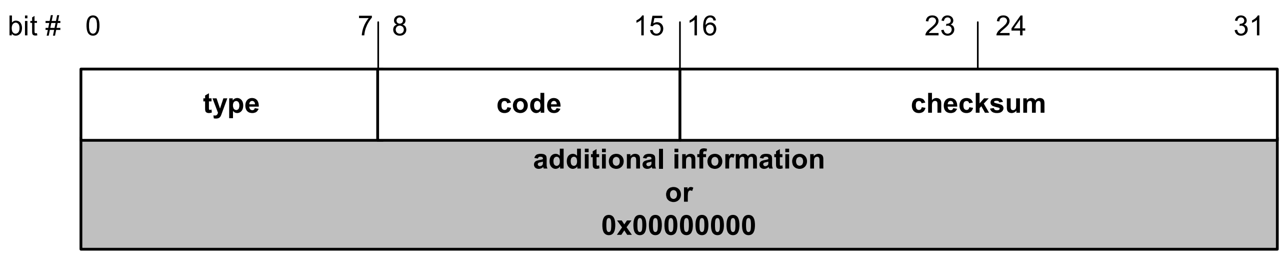


e.g. 128.9.16.14 => Port 2

Prefix	Next-hop	Port
65/8	128.17.16.1	3
128.9/16	128.17.14.1	2
128.9.16/20	128.17.14.1	2
128.9.19/24	128.17.10.1	7
128.9.25/24	128.17.14.1	2
128.9.176/20	128.17.20.1	1
142.12/19	128.17.16.1	3

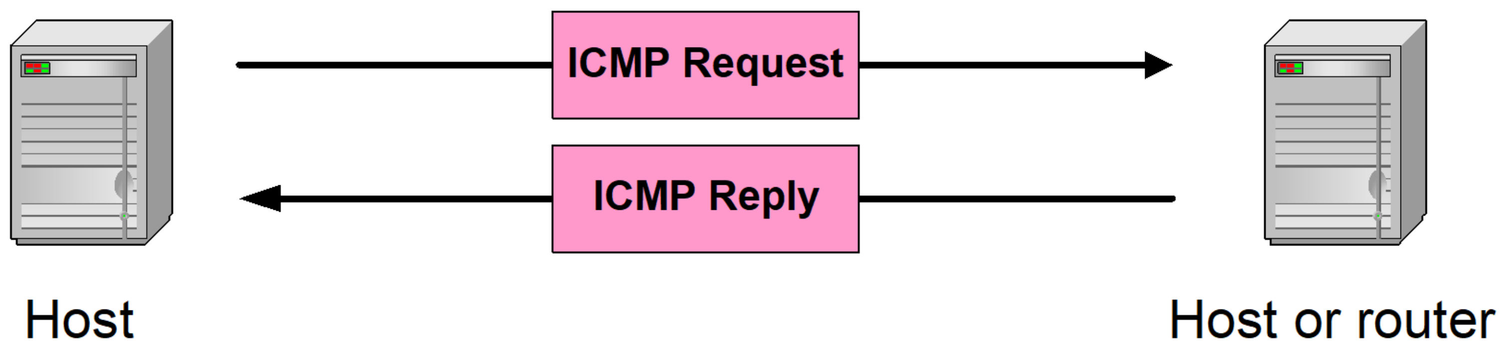
Forwarding/routing table

Internet Control Message Protocol



- ICMP is a helper protocol for IP
  - Error reporting
  - Simple queries
  - Encapsulated as IP datagrams
- Header
  - Type (1 byte): type of ICMP message
  - Code (1 byte): subtype of ICMP message
  - Checksum (2 bytes): checksum is calculated over entire ICMP message
- If there is no additional data, 4 bytes are set to zero.
  - Each ICMP messages is at least 8 bytes long

ICMP Query message

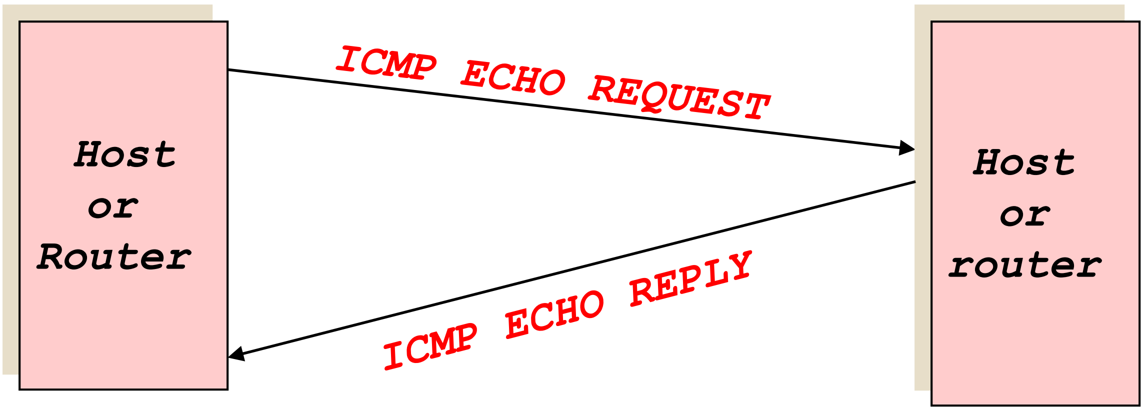


ICMP query:

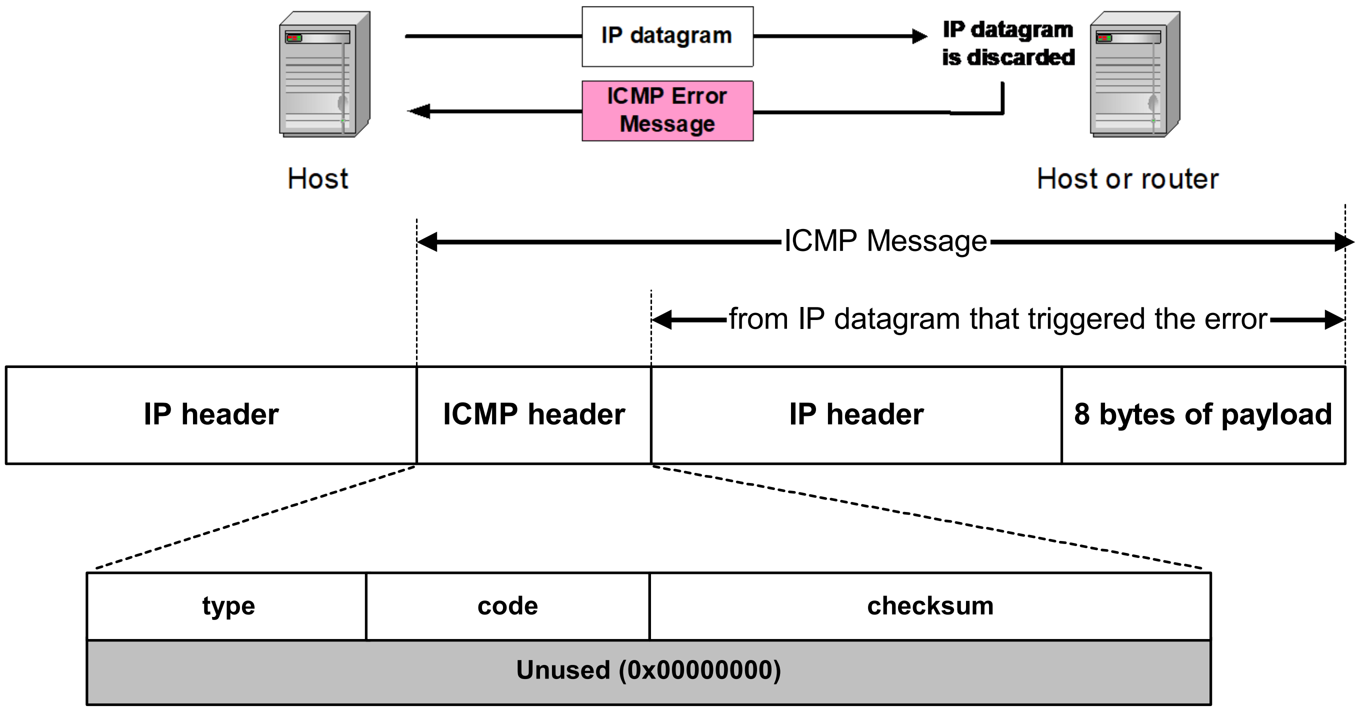
- Request sent by host to a router or host
- Reply sent back to querying host

Example of a Query: Echo Request and Reply

- Ping's are handled directly by the kernel
- Each Ping is translated into an ICMP Echo Request
- The Ping'ed host responds with an ICMP Echo Reply



ICMP Error message

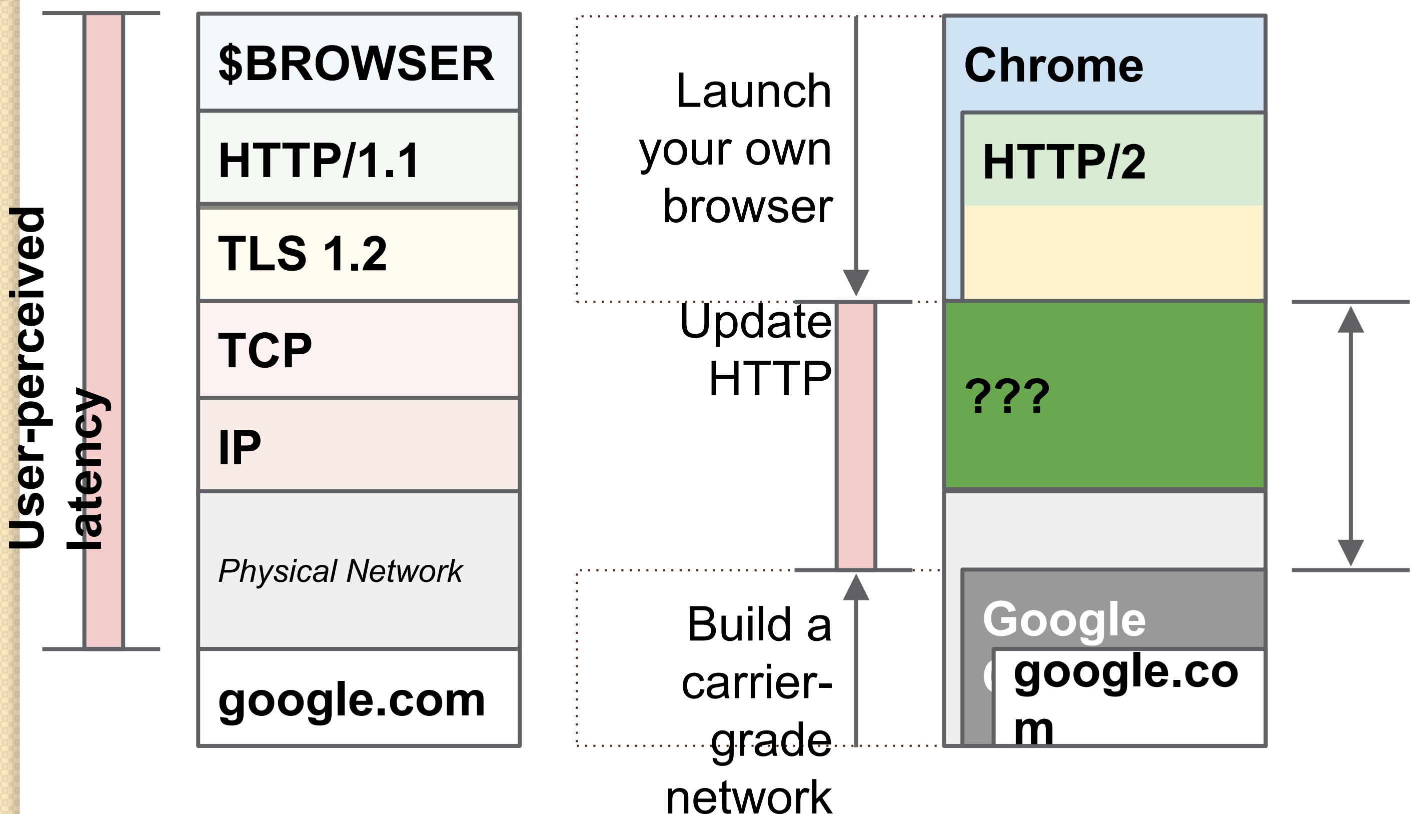


- **ICMP error messages report error conditions**
  - **ICMP error messages include the complete IP header and the first 8 bytes of the payload (typically: UDP, TCP)**
- **Typically sent when a datagram is discarded**
- **Error is often passed from ICMP to the apps**

Frequent ICMP Error message

Type	Code	Description	
3	0–15	Destination unreachable	Notification that an IP datagram could not be forwarded and was dropped. The code field contains an explanation.
5	0–3	Redirect	Informs about an alternative route for the datagram and should result in a routing table update. The code field explains the reason for the route change.
11	0, 1	Time exceeded	Sent when the TTL field has reached zero (Code 0) or when there is a timeout for the reassembly of segments (Code 1)
12	0, 1	Parameter problem	Sent when the IP header is invalid (Code 0) or when an IP header option is missing (Code 1)

Can HTTP be Faster?



QUIC Features

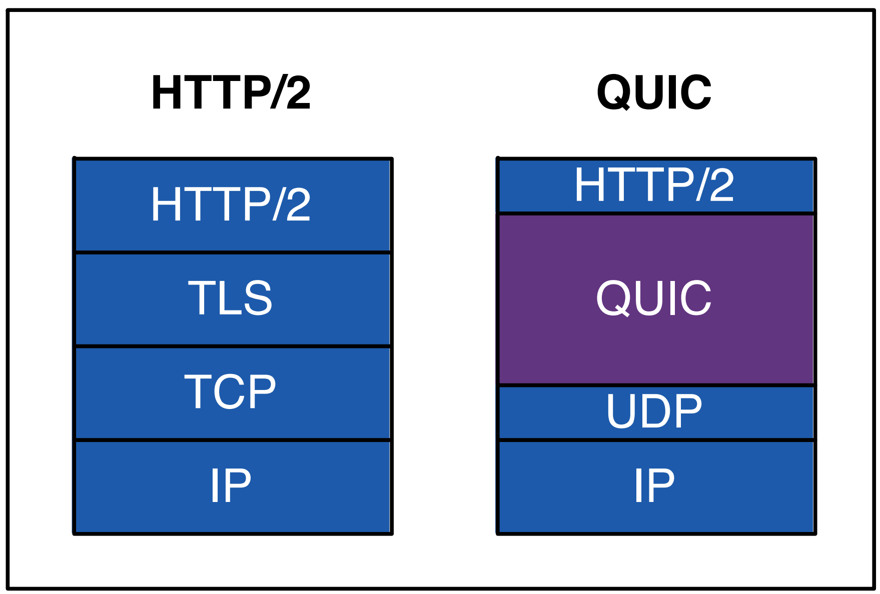
- 0-RTT connection handshake
- 0-RTT encryption handshake
- Connections survive IP address change
- Enhanced packet loss recovery
- Always encrypted
- Mostly fixes head of line blocking
- FEC (Forward Error Correction) data recovery

HTTP/2 Features in QUIC

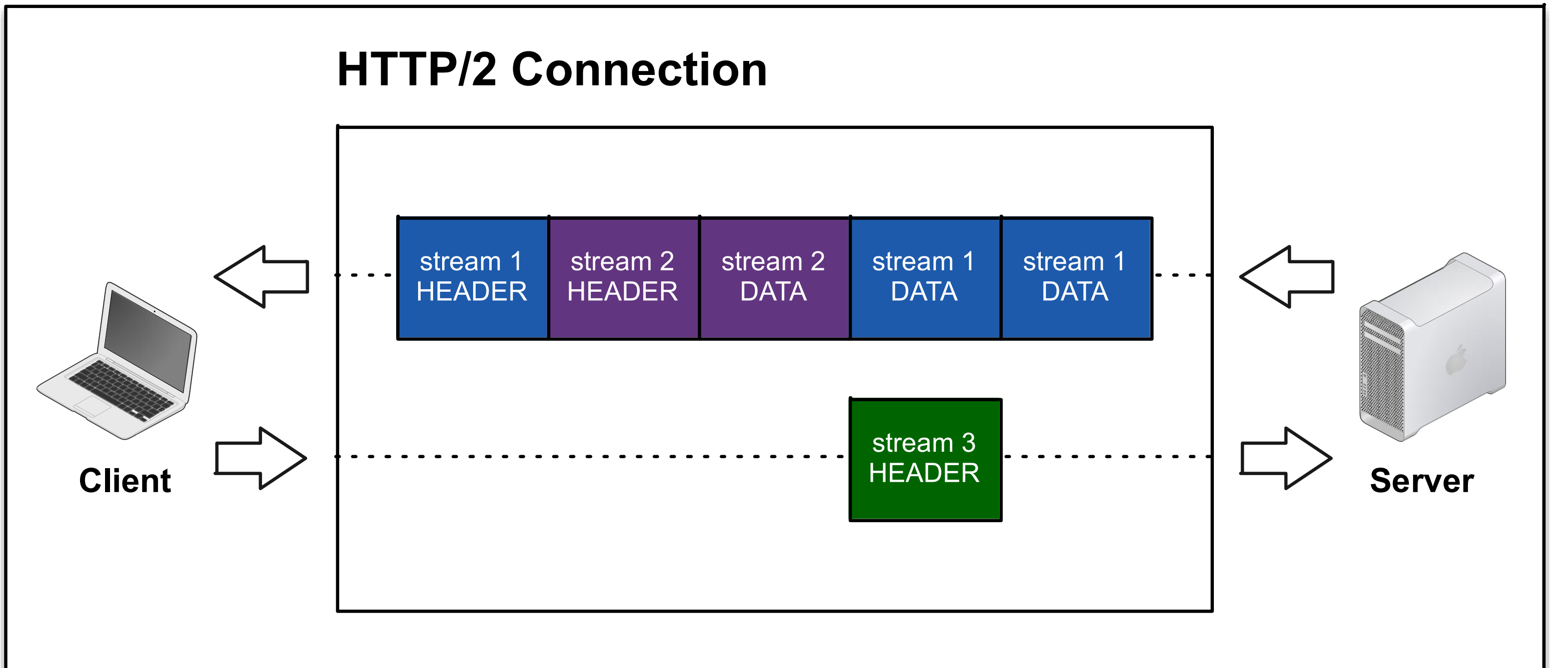
- Multiplexed streams
- Sharing connection across domains
- HPACK header compression
- Stream prioritization
- Flow Control
- Serverinitiated streams

QUIC

- Congestion control, encryption, and some HTTP/2 move to QUIC

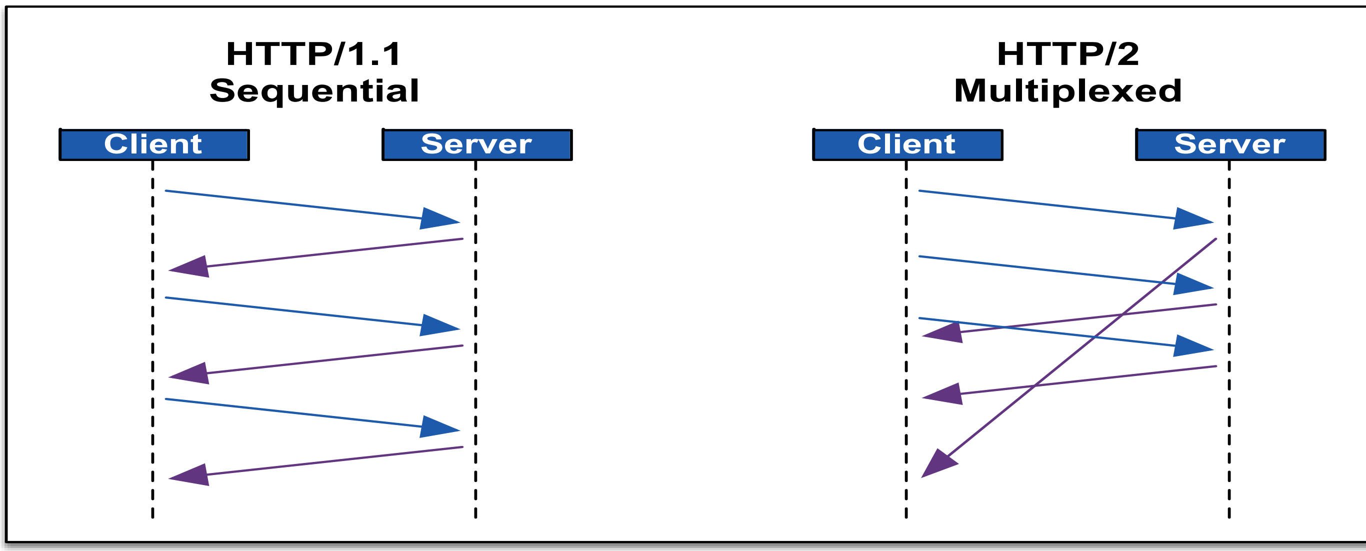


Streams



- One stream per request
- Stream are broken up into frames
- Stream 1 crypto handshake
- Stream 3 is for headers – to serialize headers (HPACK)

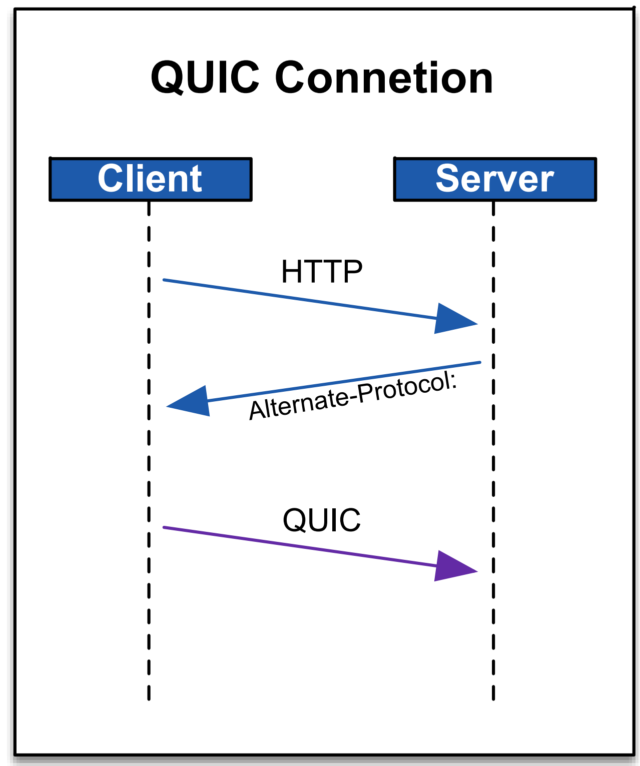
Multiplexed Streams



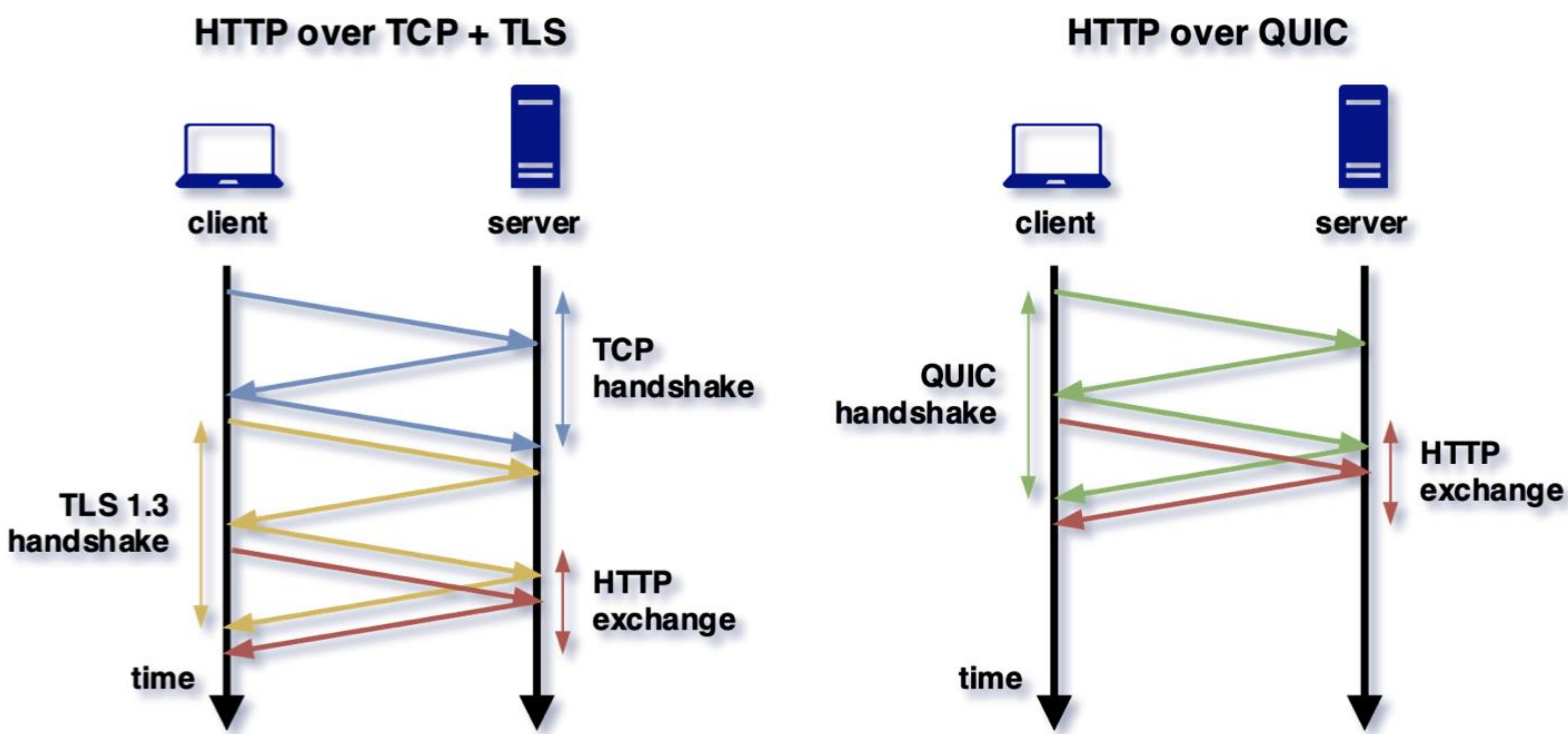
- HTTP/1.1
  - 4-8 outstanding requests on 4-8 connections
  - Resource intensive on the server
- HTTP/2 and QUIC
  - One connection, many concurrent requests
  - Normally limited to 100

Establishing a QUIC Connection

- HTTP response header
  - Alternate-Protocol: 443:quic
- Client establishes QUIC connection in the background
- Client's can cache if server supports QUIC

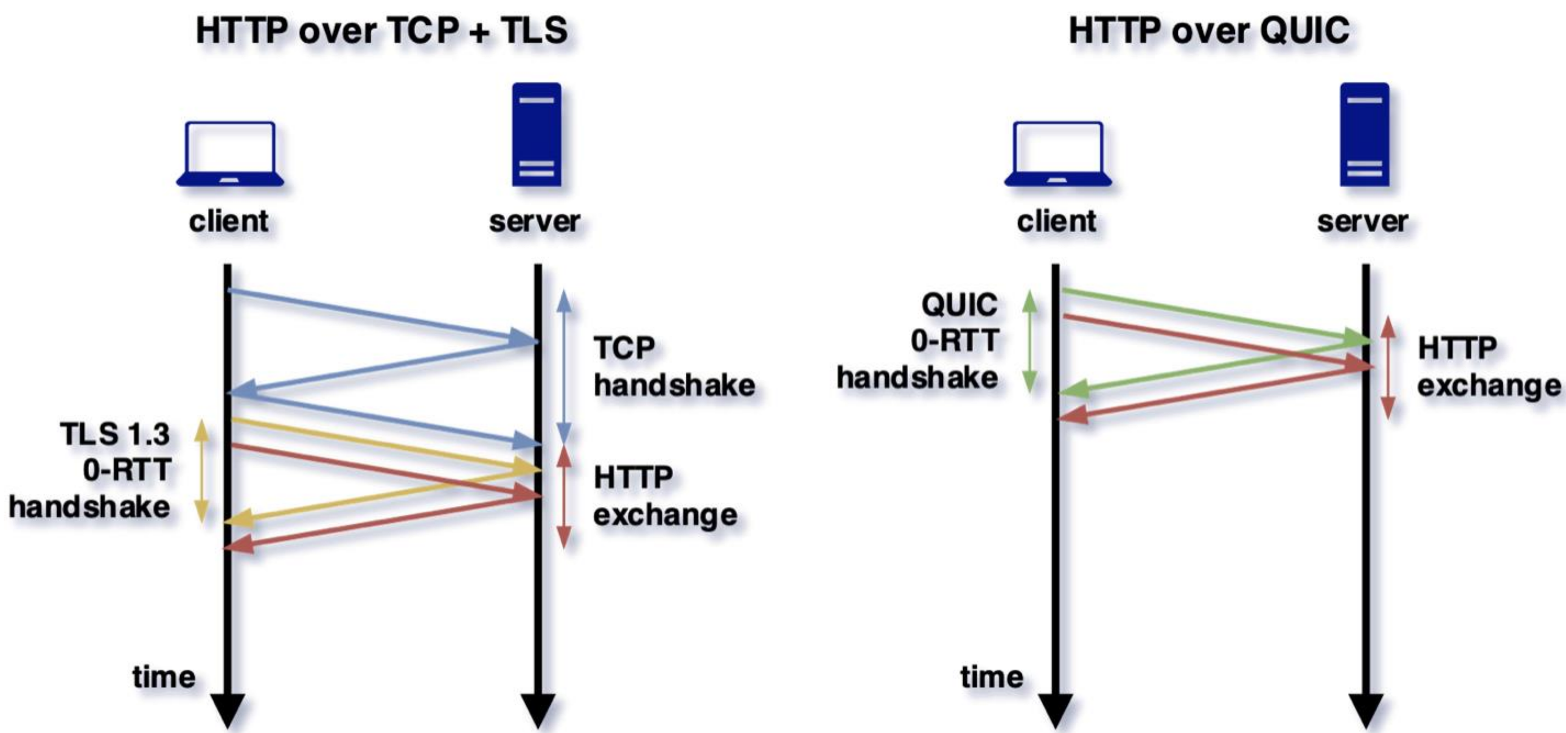


TCP+TLS vs. QUIC



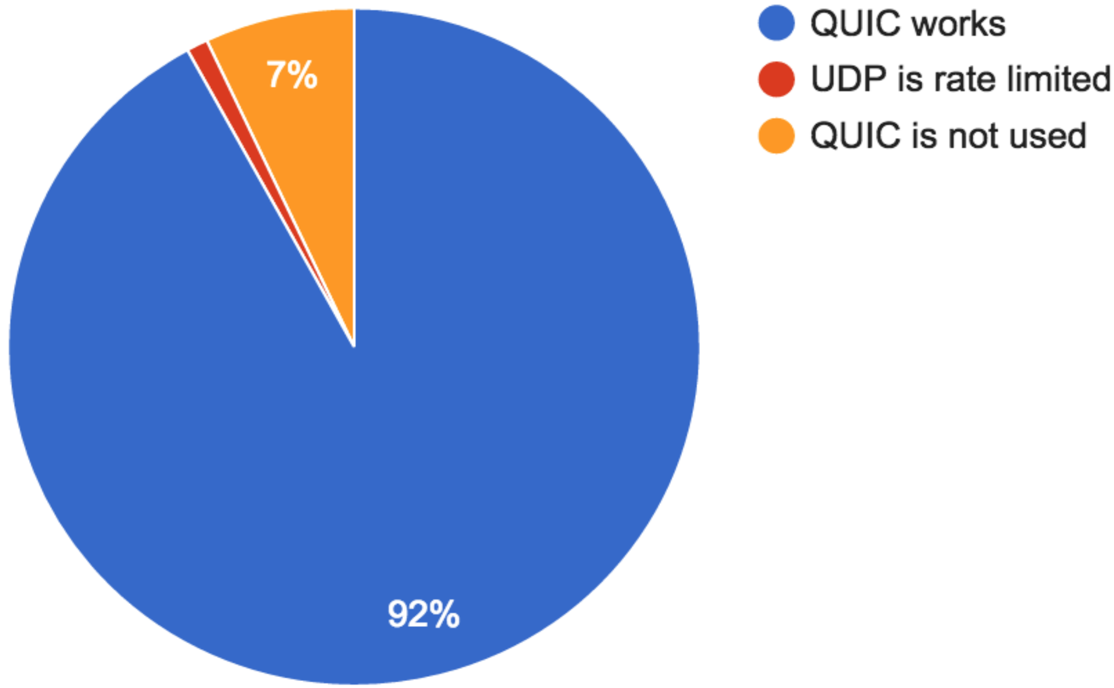
Initial Connection

TCP+TLS vs. QUIC



Reconnection

QUIC Success Rate

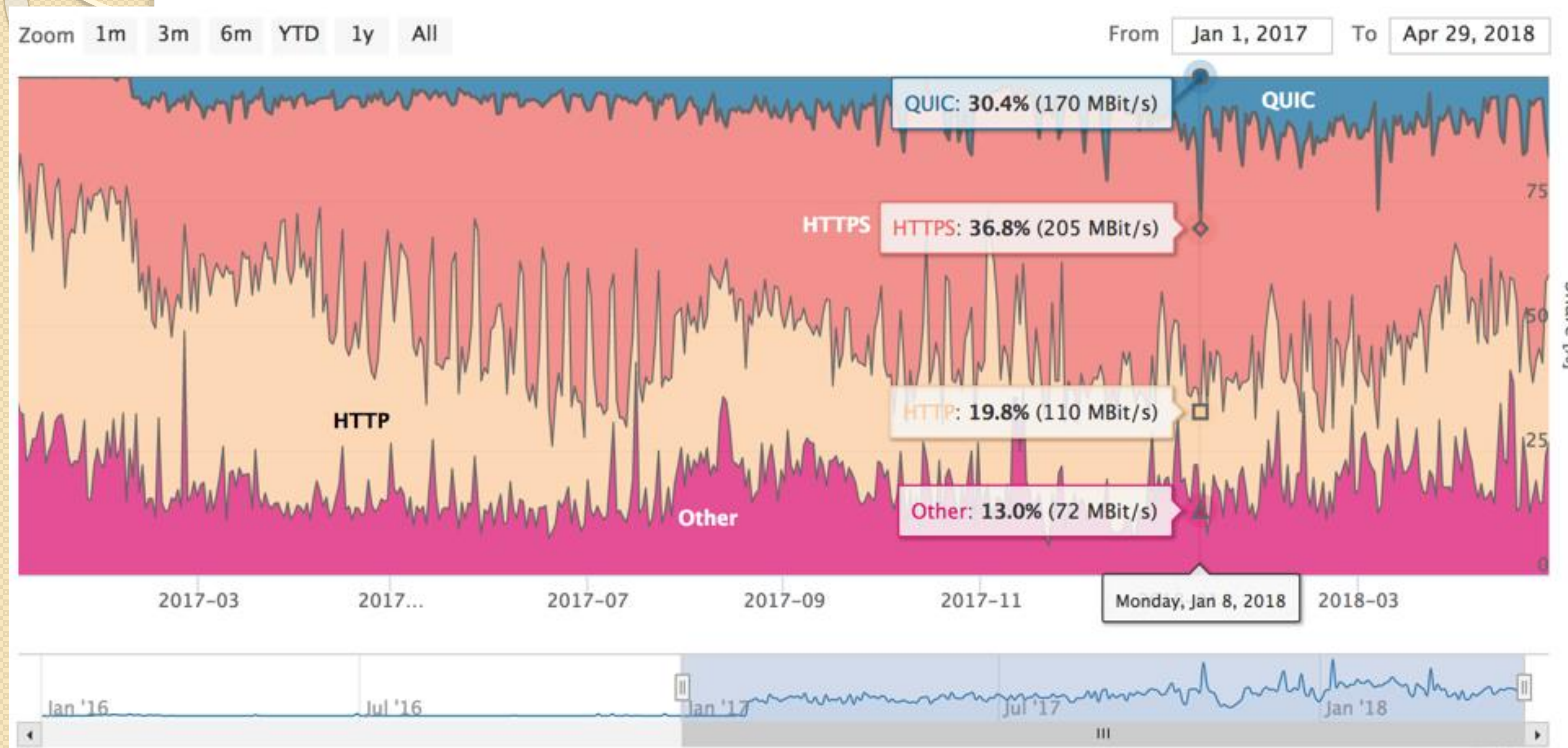


- QUIC connection success rate
  - 92% works
  - 7% doesn’t work
  - 1% is rate limited
- Google disables QUIC to specific ASNs

QUIC Performance

- 5% latency reduction on average
- 30% reduction in rebuffers (video pauses) on YouTube
- 1 second faster at the 99th percentile for Google web search
- Helps more for higher latency networks

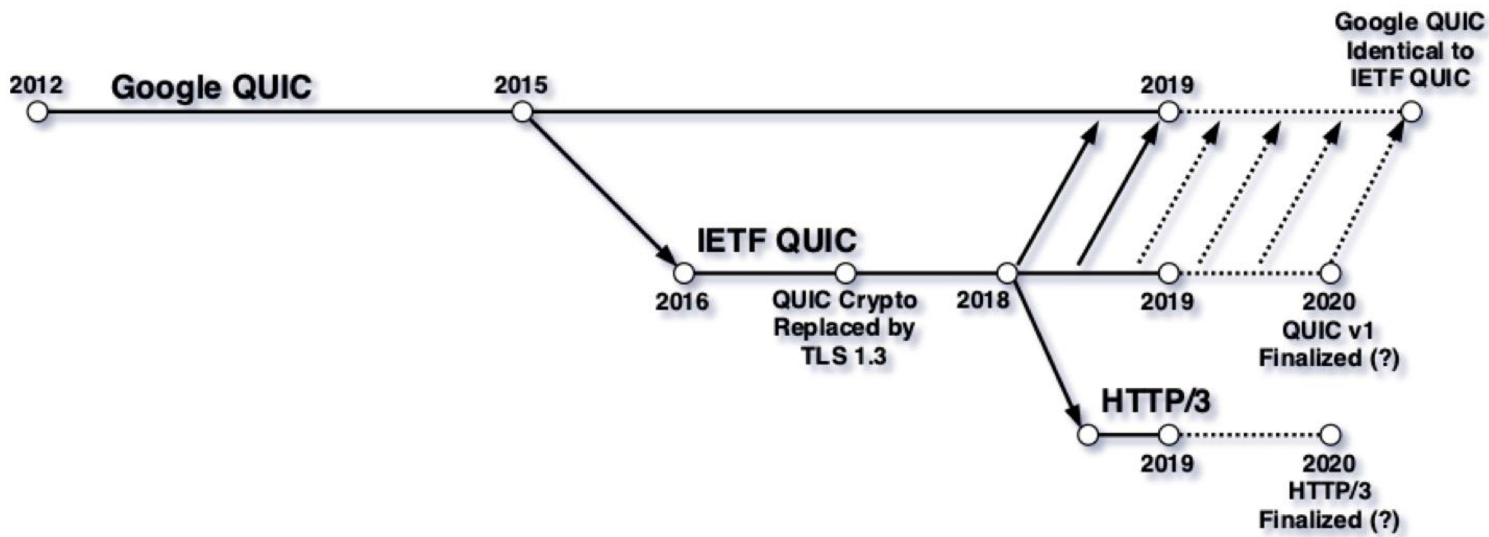
Protocol Usage on the Internet



QUIC Support

- Google QUIC
  - Chrome (Desktop/Mobiles)
- IETF QUIC
  - Chrome Canary
  - Firefox Nightly
  - Safari 14
- Early Trials
  - https://quic.nginx.org
  - https://ietf.akaquic.com

QUIC Lineage



Differences

- HTTP/1 and HTTP/2 on TCP
  - Until Now
  - TLS used for Security
  - Share a Single TCP Connection (HTTP/2)
- HTTP/3 on UDP
  - Debut 2020
  - QUIC to Replace TCP+TLS
  - Low Latency (Reduced Overhead)
  - Higher Throughput (Multiple Connections)

Thoughts on QUIC

- Challenges
  - UDP blocking by Service Providers
  - Complexity Associated with Multiple Threadedness/Connections/Streams
  - Higher CPU Utilization
- Possible Solutions
  - OS Bypassing Drivers
  - Hardware Offload Engines
- Why Not Replace Other TCP packets
  - Likely will follow after HTTP/3
  - What would it take?