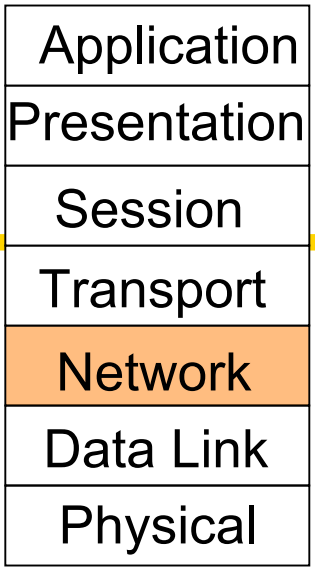
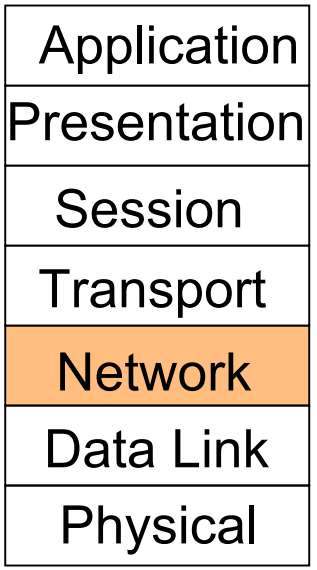
CSE/EE 461
The Network Layer



This Lecture

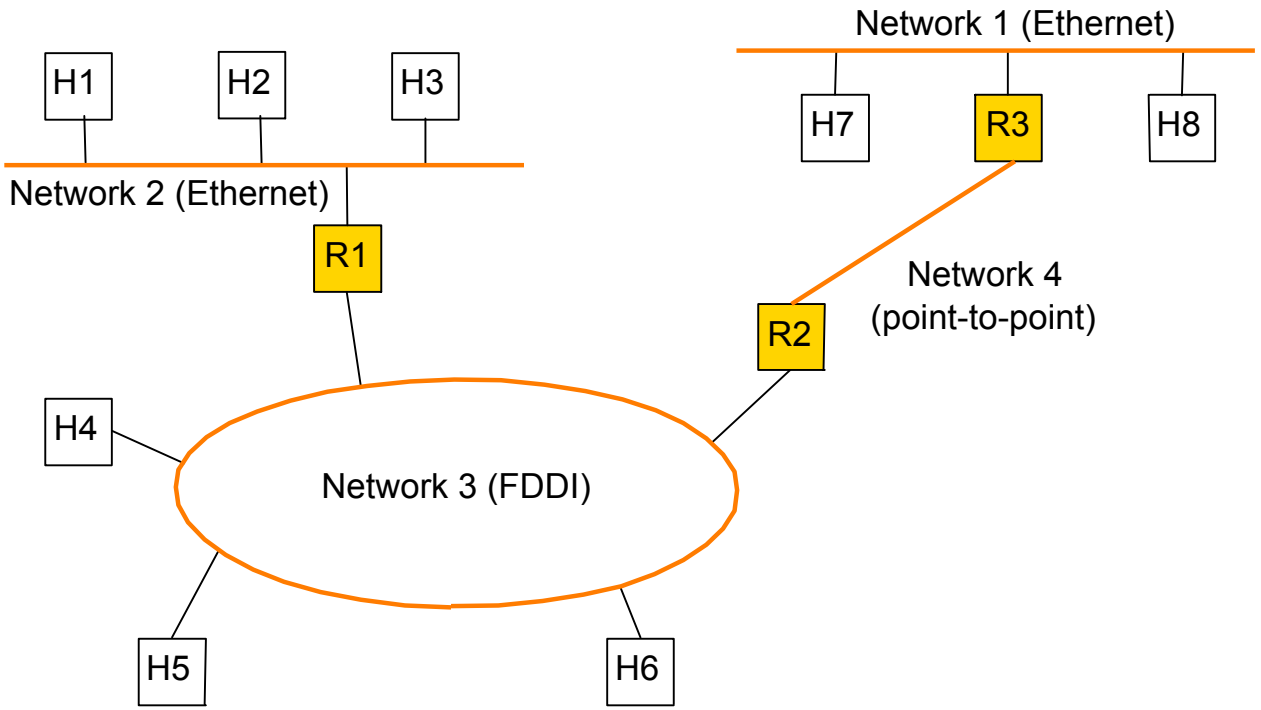
- Focus:
  - What to do when one wire isn't big enough?
    - Point to point link
    - Broadcast link
      - (Ethernet discussion coming up)
  - How do we build large networks?

- Introduction to the Network layer
  - Internetworks
  - Service models
  - IP, ICMP



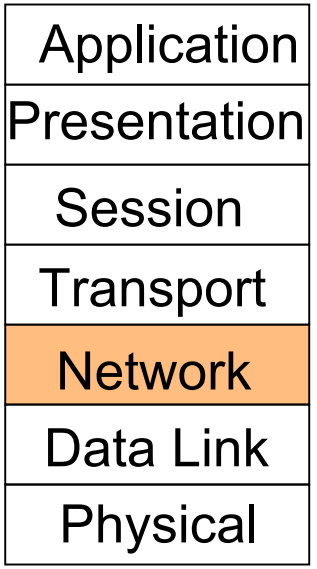
Internetworks

- Set of interconnected networks, e.g., the Internet
  - Scale and heterogeneity



The Network Layer

- Job is to provide end-to-end data delivery between hosts on an internetwork
- Provides a higher layer of addressing



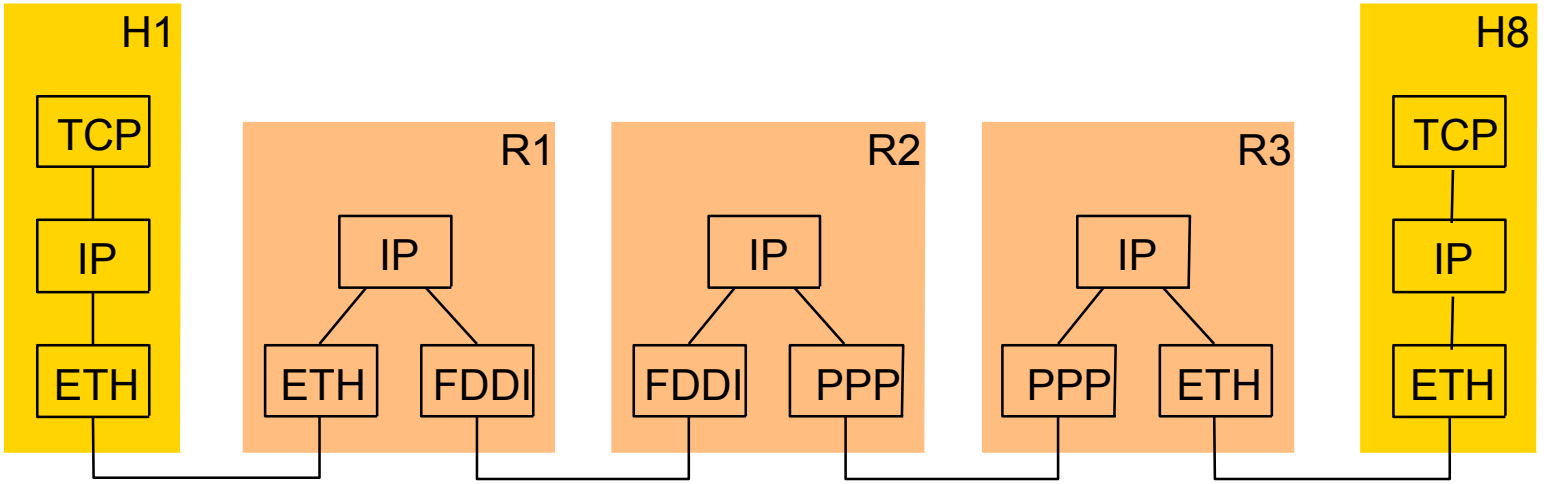
Network Service Models

- Datagram delivery: postal service
  - connectionless, best-effort or unreliable service
  - Network can't guarantee delivery of the packet
  - Each packet from a host is routed independently
  - Example: IP

- Virtual circuit models: telephone
  - connection-oriented service
  - Signaling: connection establishment, data transfer, teardown
  - All packets from a host are routed the same way (router state)
  - Example: ATM, Frame Relay, X.25
- You can build one with the other
  - But one is much easier to build than the other

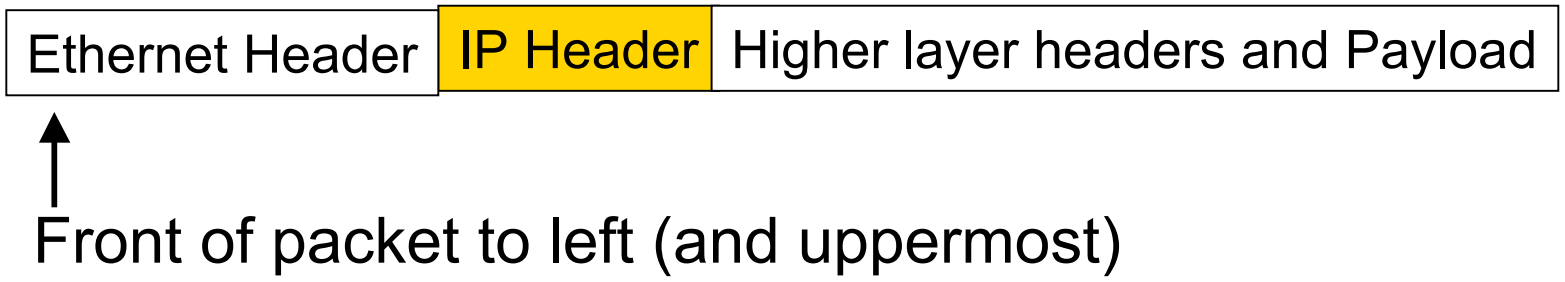
In terms of protocol stacks

- IP is the network layer protocol used in the Internet
- Routers are network level gateways
- Packet is the term for network layer PDUs (protocol data unit)



In terms of packet formats

- View of a packet on the wire on network 1 or 2
- Routers work with IP header, not higher
  - Higher would be a "layer violation"
    - Imagine if the post office read your mail
- Routers strip and add link layer headers

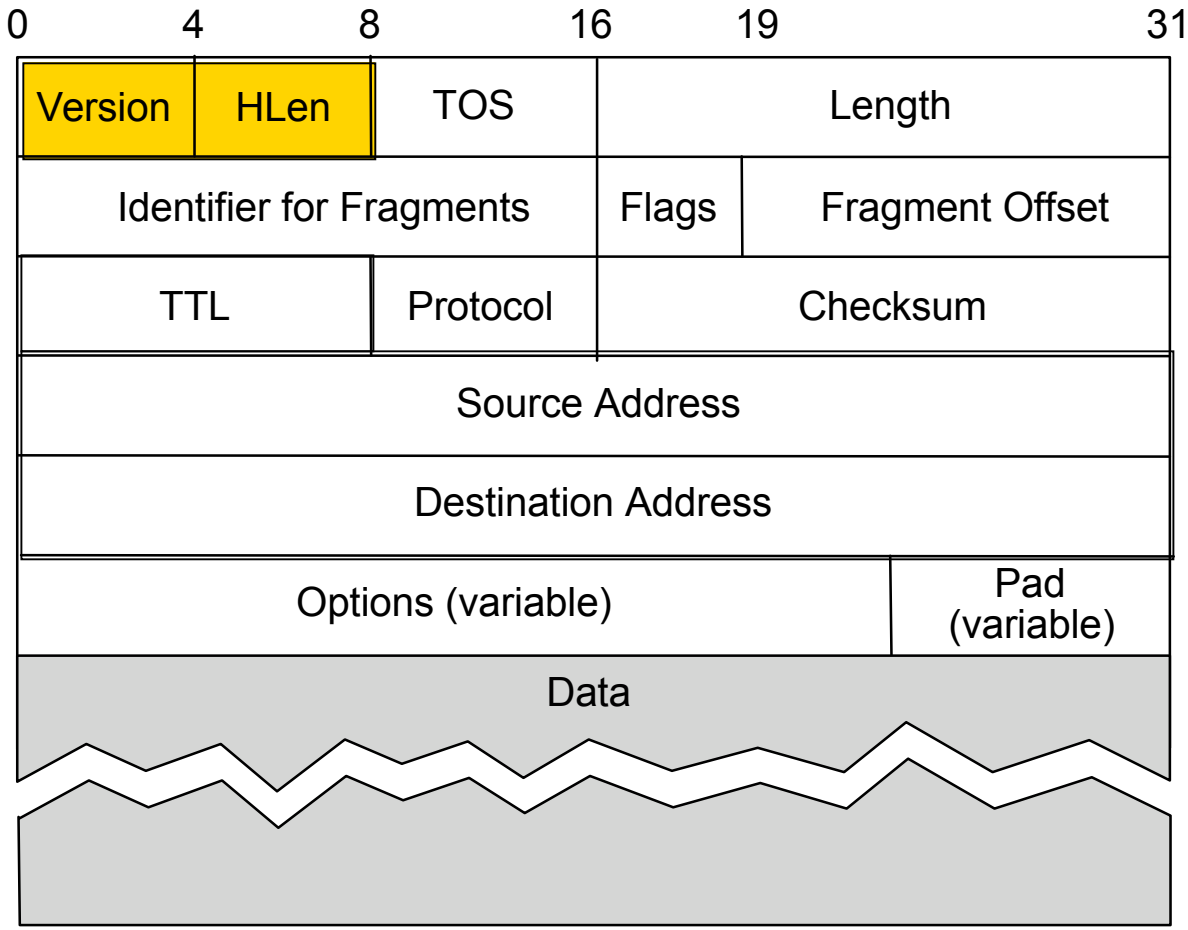


Internet Protocol (IP)

- IP (RFC791) defines a datagram "best effort" service
  - May be loss, reordering, duplication, and errors!
  - Currently IPv4 (IP version 4), IPv6 here/coming way
- Global, hierarchical addresses, not flat addresses
  - 32 bits in IPv4 address; 128 bits in IPv6 address
  - ARP (Address Resolution Protocol) maps IP to MAC addresses
- Routers forward packets using predetermined routes
  - Routing protocols (RIP, OSPF, BGP) run between routers to maintain routes (routing table, forwarding information base)
    - We'll deal with this later

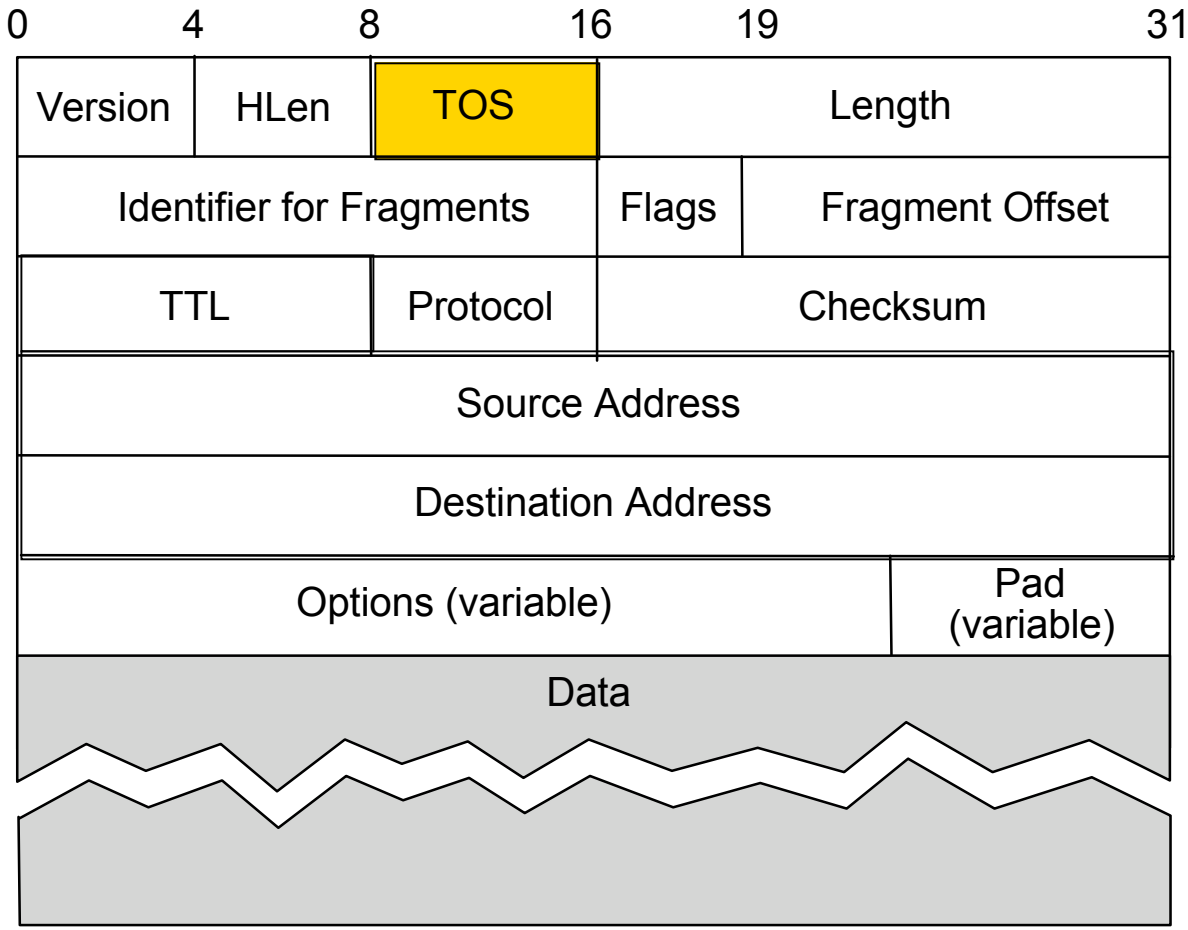
IPv4 Packet Format

- Version is 4
- Header length is number of 32 bit words
- Limits size of options



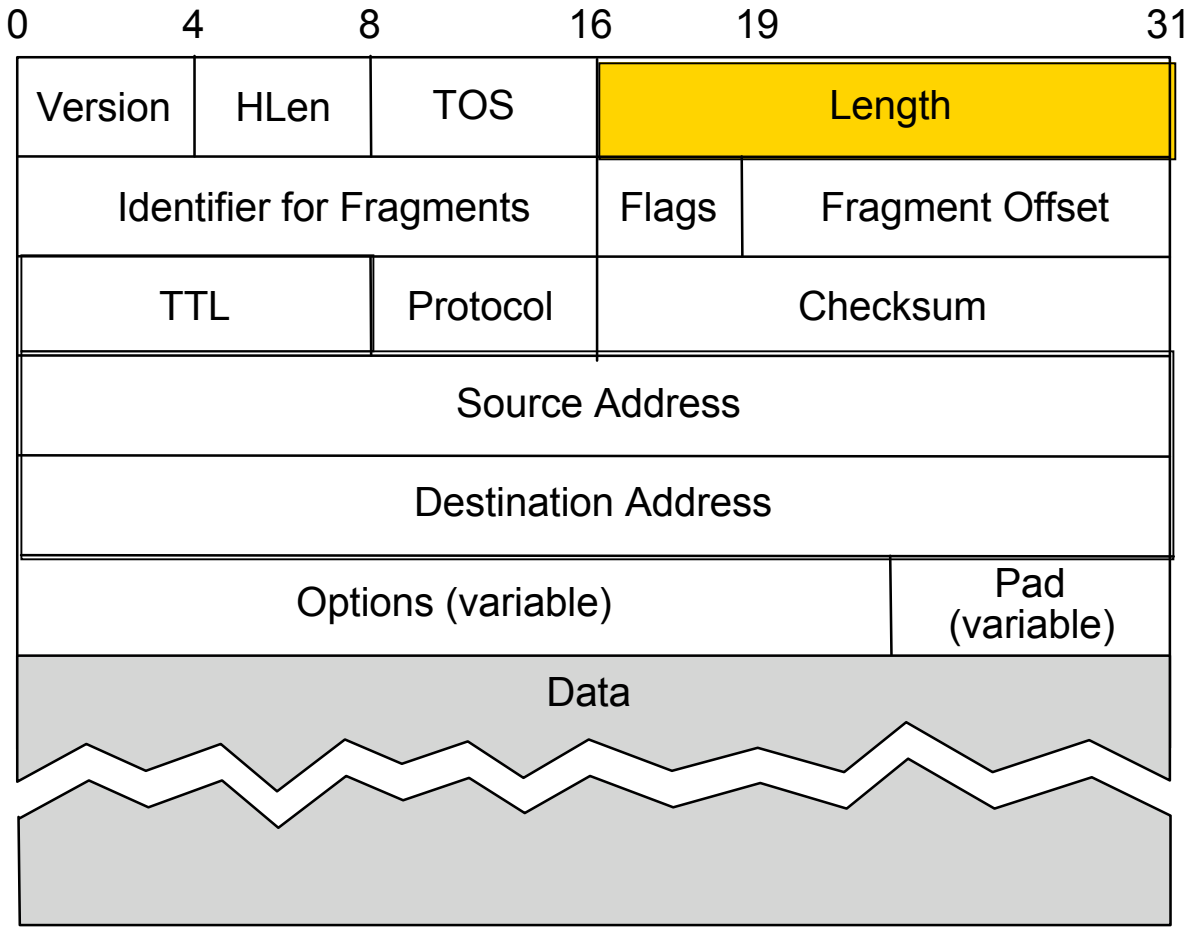
IPv4 Header Fields …

- Type of Service
- Abstract notion, never really worked out
  - Routers ignored
- But now being redefined for Diffserv



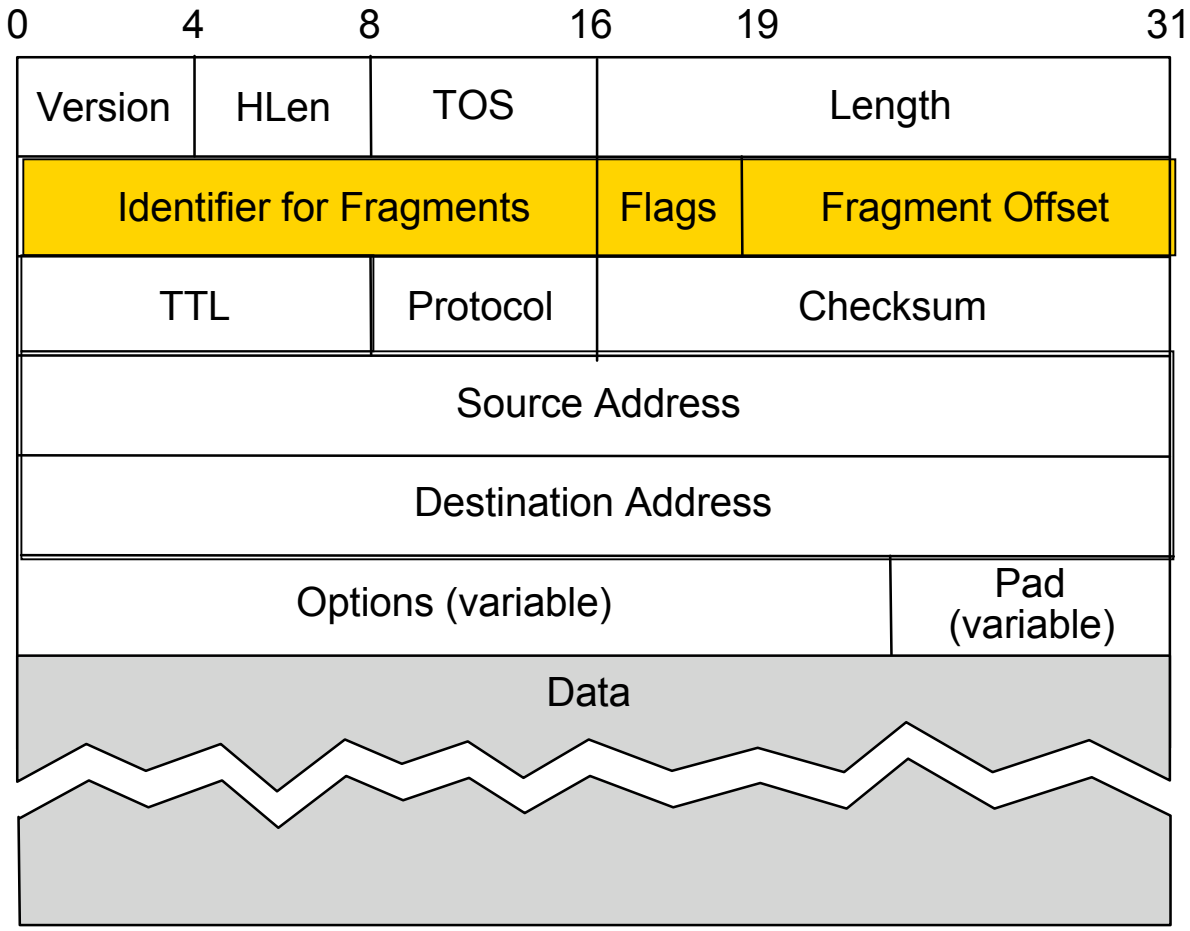
IPv4 Header Fields …

- Length of packet
  - Hdr+data
- Min 20 bytes, max 65K bytes (limit to packet size)



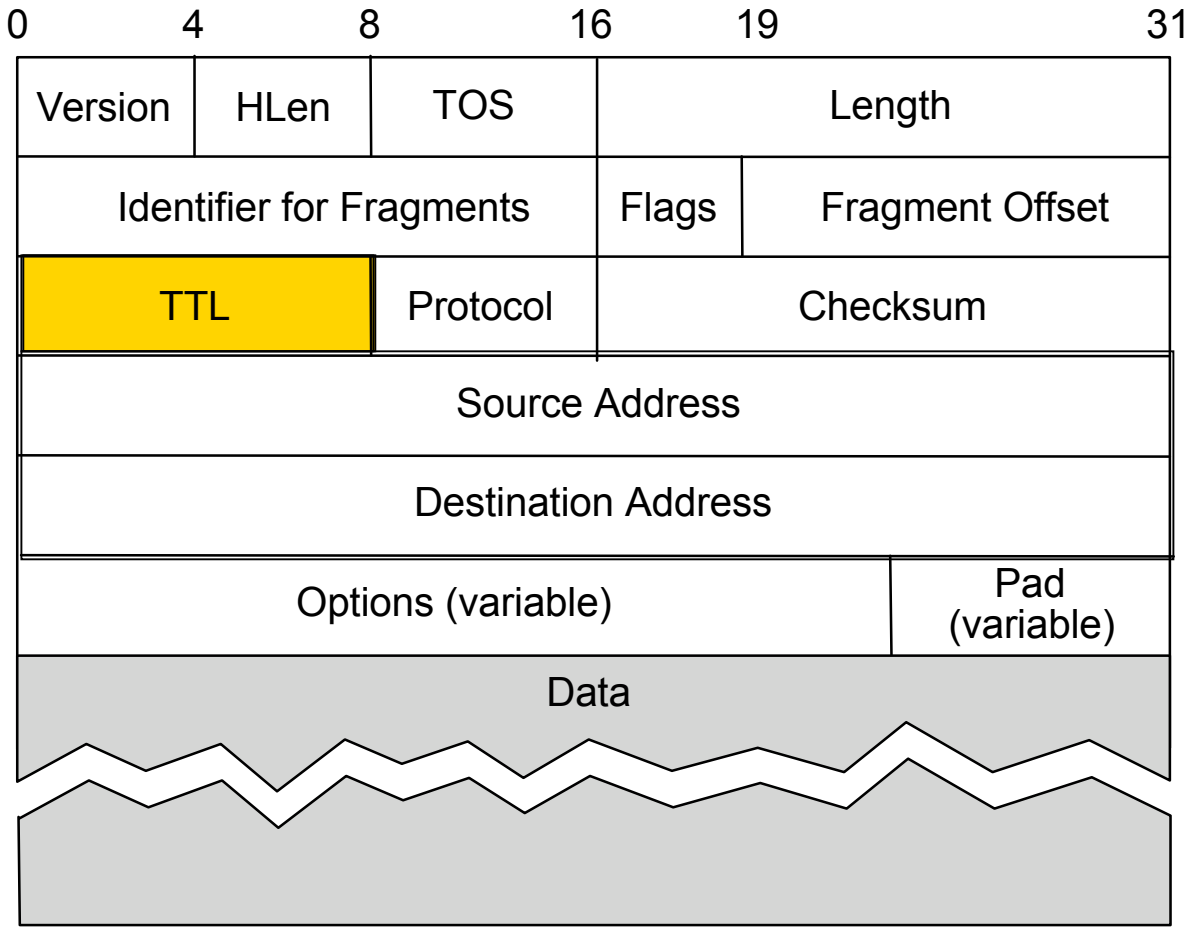
IPv4 Header Fields …

- Fragment fields
- Different links have different frame size limits
- May need to break large packet into smaller fragments



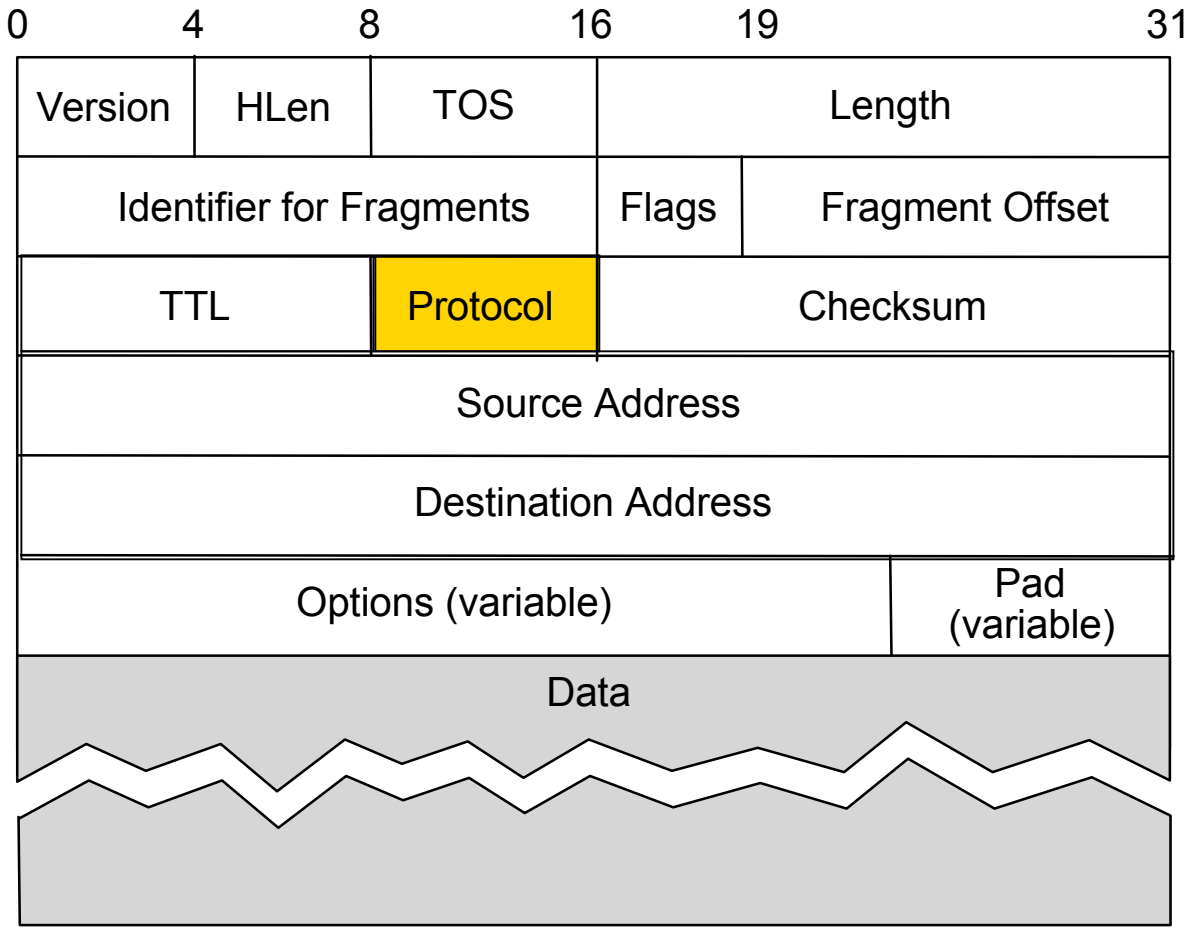
IPv4 Header Fields …

- Time To Live
- Decremented by router and packet discarded if = 0
- Prevents immortal packets



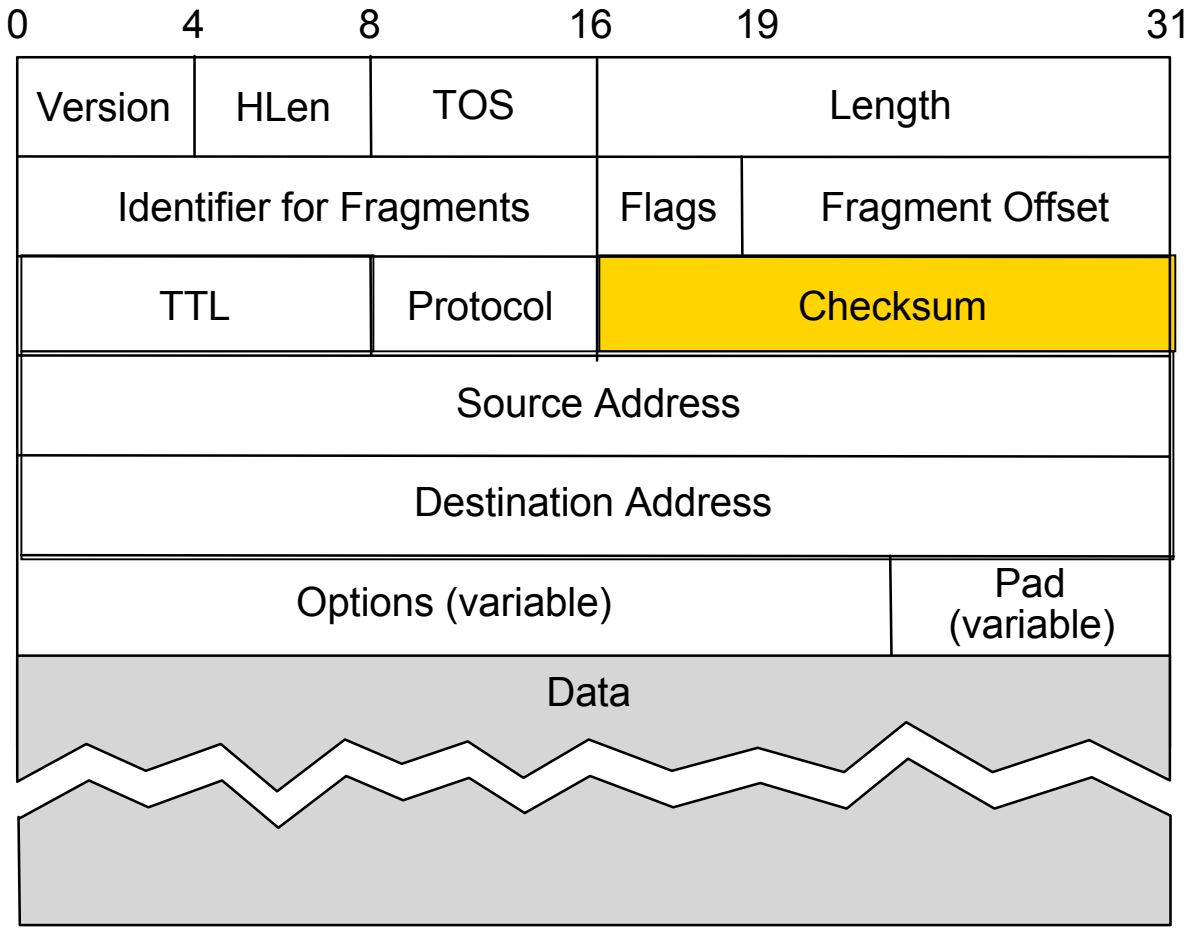
IPv4 Header Fields …

- Identifies higher layer protocol
  - E.g., TCP, UDP



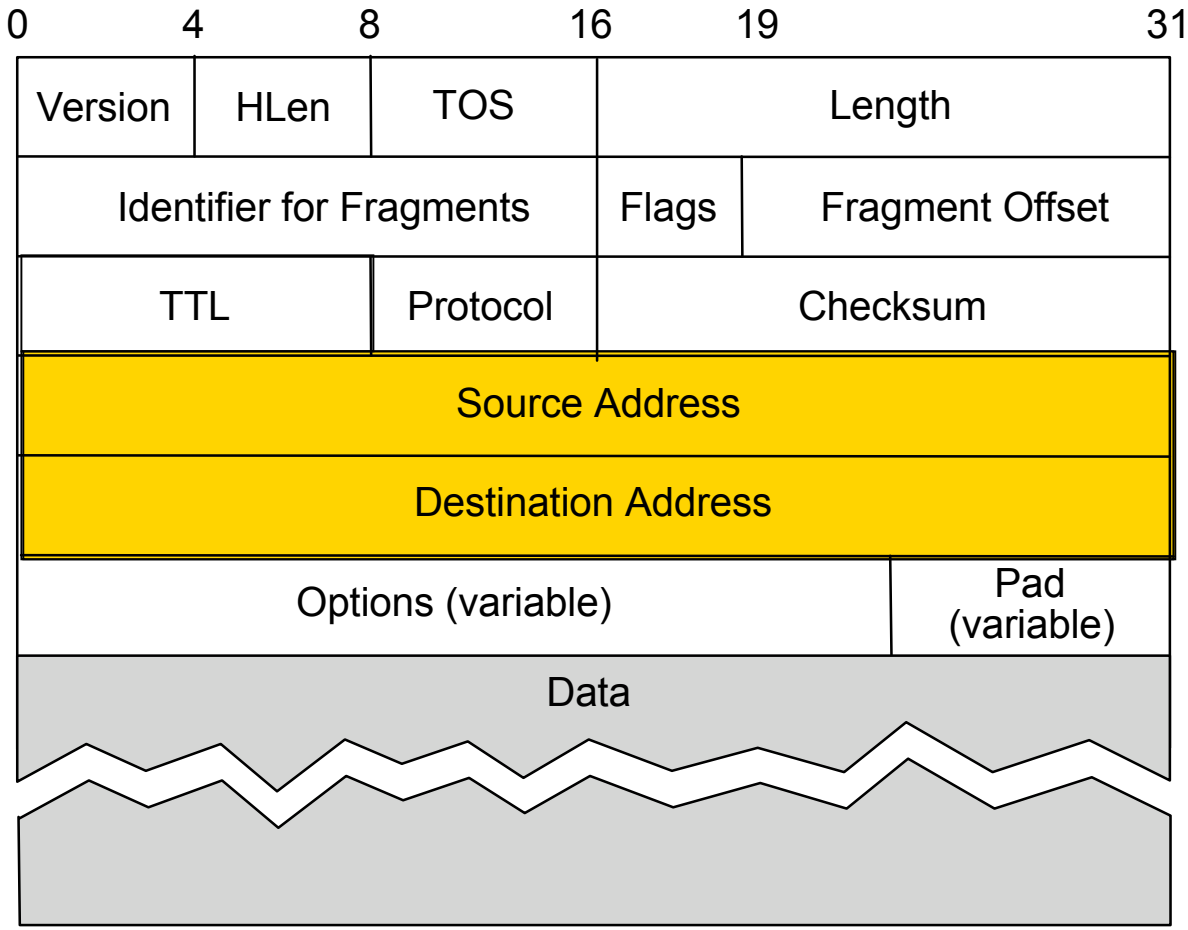
IPv4 Header Fields …

- Header checksum
- Recalculated by routers (TTL drops)
- Doesn't cover data
  - Why not?
- Disappears for IPv6



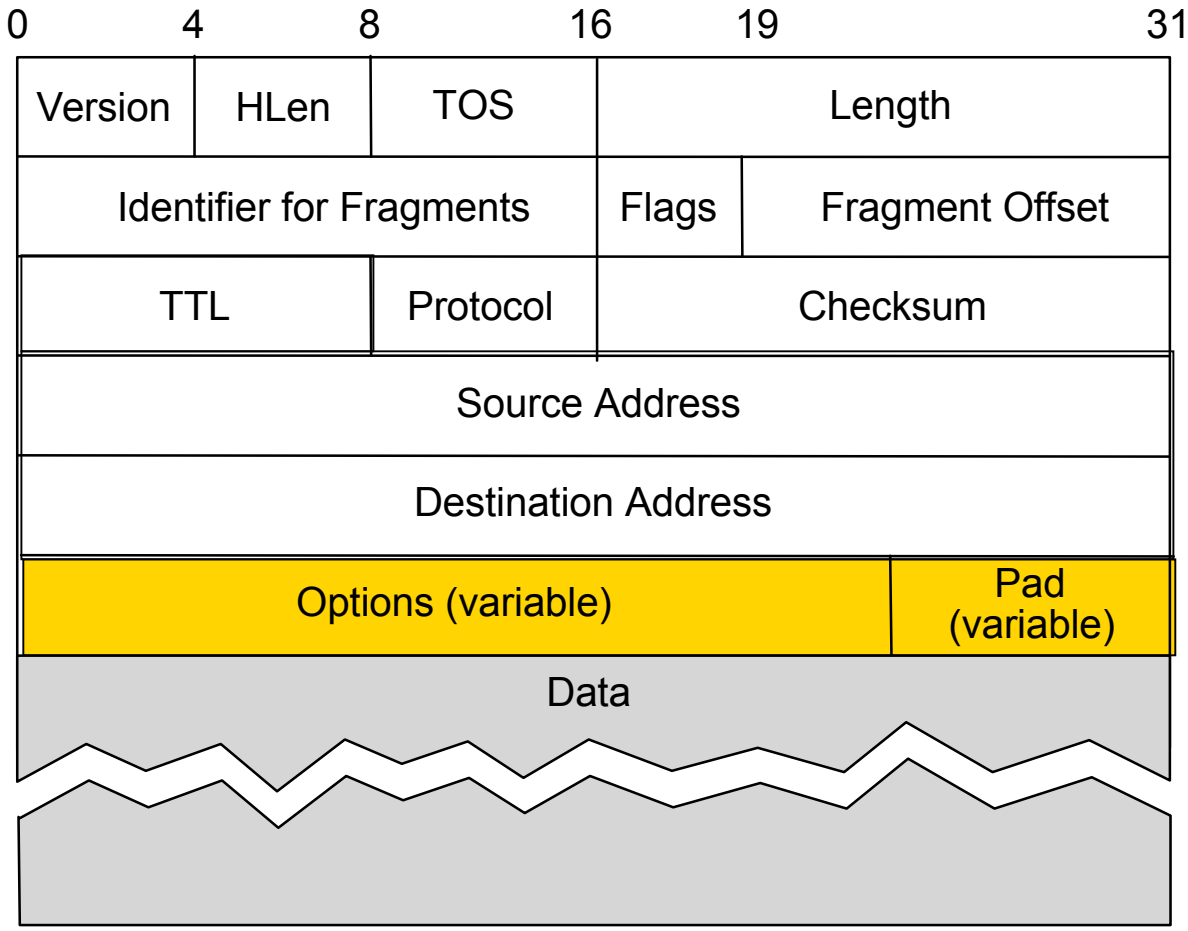
IPv4 Header Fields …

- Source/destination addresses
  - Not Ethernet
- Unchanged by routers
- Not authenticated



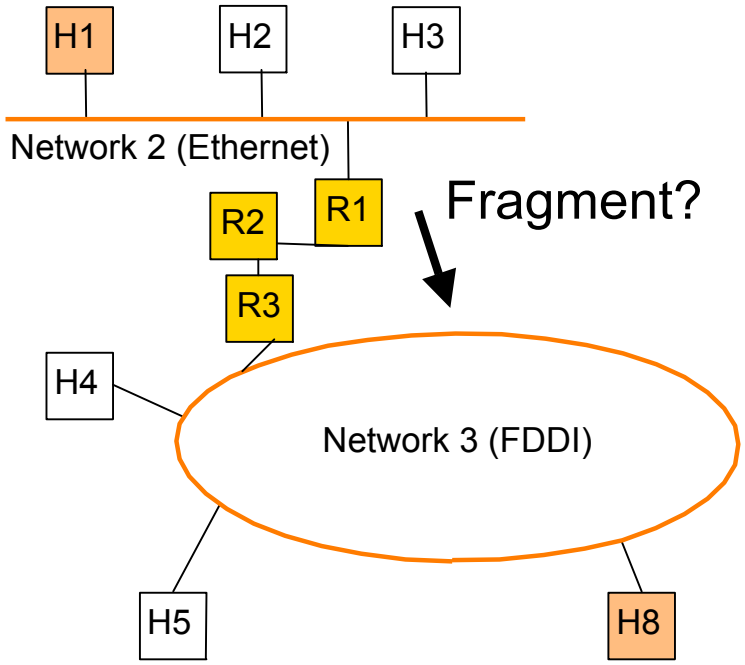
IPv4 Header Fields …

- IP options indicate special handling
  - Timestamps
  - "Source" routes
- Rarely used …



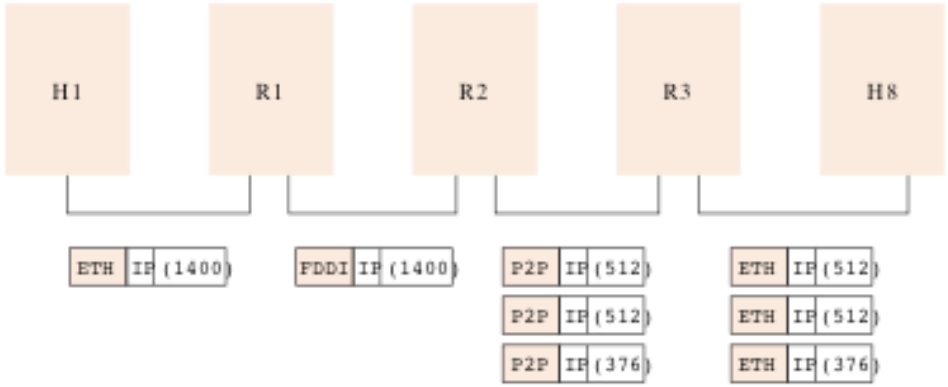
Fragmentation Issue

- Different networks may have different frame limits (MTUs)
  - Ethernet 1.5K, FDDI 4.5K
- Don't know if packet will be too big for path beforehand
  - IPv4: fragment on demand and reassemble at destination
  - IPv6: network returns error message so host can learn limit



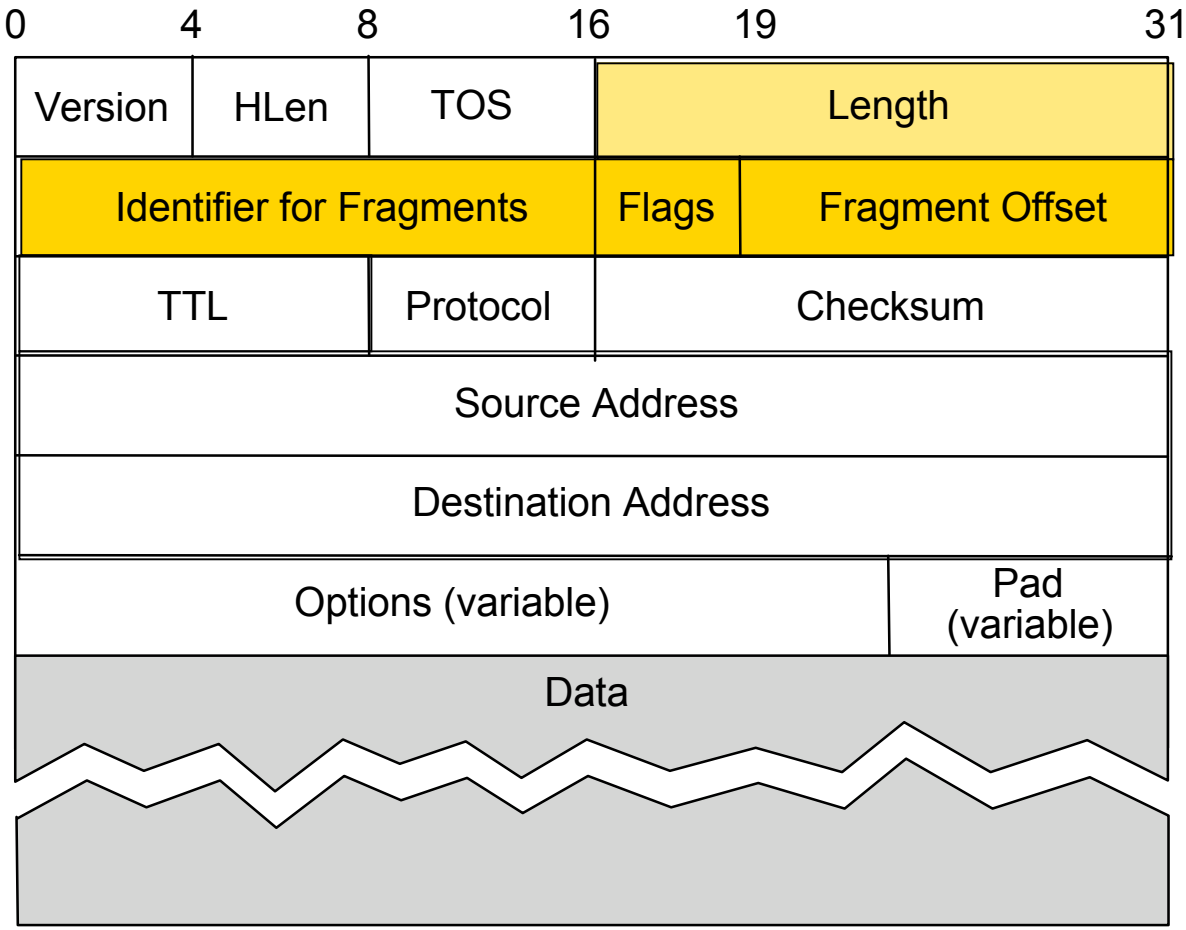
Fragmentation and Reassembly

- Each network has some MTU
- Strategy
  - fragment when necessary (MTU < Datagram)
  - try to avoid fragmentation at source host
  - refragmentation is possible
  - fragments are self-contained datagrams
  - use CS-PDU (not cells) for ATM
  - delay reassembly until destination host
  - do not recover from lost fragments
- Example

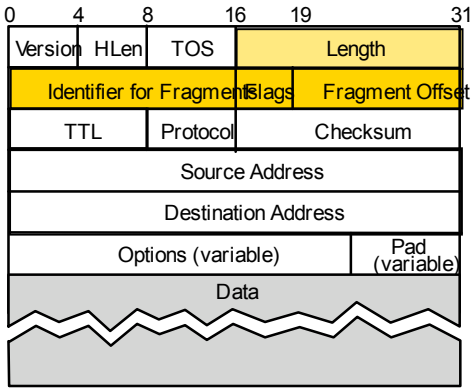


Fragment Fields

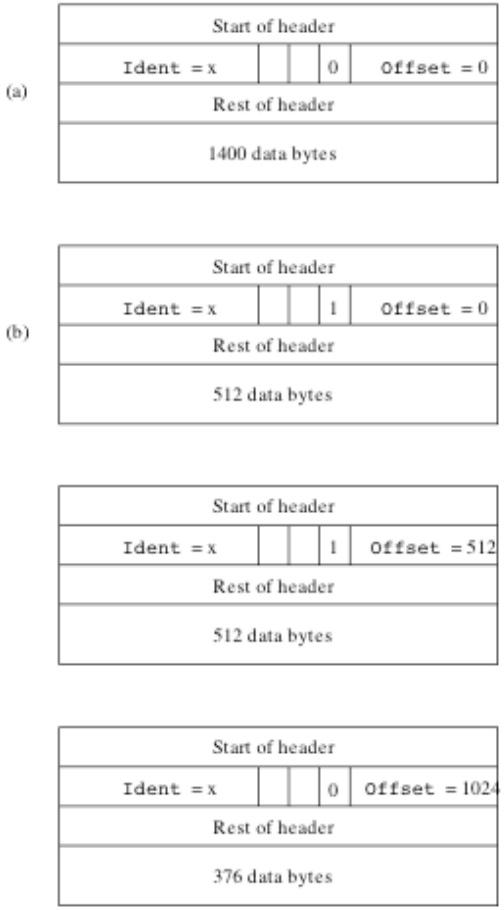
- Fragments of one packet identified by (source, dest, frag id) triple
  - Uniquifies
- Offset gives start, length changed
- Flags are More Fragments (MF) Don't Fragment (DF)



Fragmenting a Packet



Packet Format



Fragment Considerations

- Relating fragments to original datagram provides:
  - Tolerance of loss, reordering and duplication
  - Ability to fragment fragments
- Reassembly done at the endpoint
  - Puts pressure on the receiver
- Consequences of fragmentation:
  - Loss of any fragments causes loss of entire packet
    - The packet train and buffer overflow

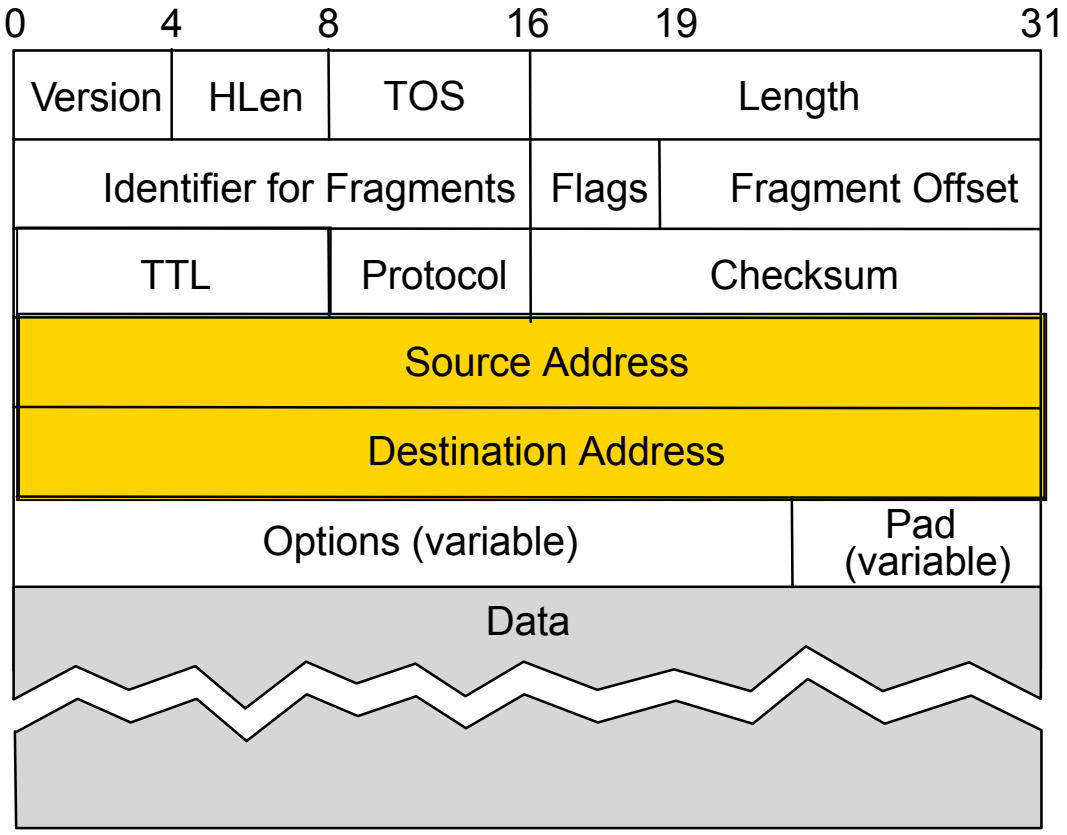
Fragmentation Issues Summary

- Causes inefficient use of resources within the network
  - BW, CPU
  - Eg, App sends 1024 bytes across ARPANET (1007 MTU)
    - 1024 + 40 for TCP/IP header
    - Frag 1 == 1000, Frag 2 == 84
    - Should have sent 1006 bytes!
- Higher level protocols must rexmit entire datagram
  - Really hard with "guaranteed packet loss"
- Efficient reassembly is hard
  - Lots of special cases
  - (think linked lists)

Avoiding Fragmentation

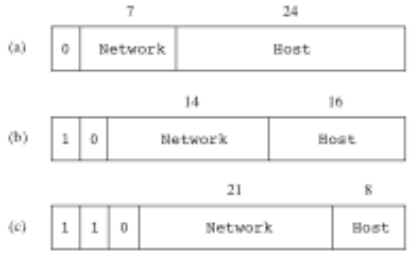
- Always send small datagrams
  - Might be too small
- "Guess" MTU of path
  - Use DF flag. May have large startup time
- Discover actual MTU of path
  - One RT delay w/help, much more w/o.
  - "Help" requires router support
- Guess or discover, but be willing to accept your mistakes

What is an Internet Address?



Global Addresses

- Properties
  - globally unique
  - hierarchical: network + host
- Format



- Dot notation
  - 10.3.2.4
  - 128.96.33.81
  - 192.12.69.77

1. Small number of large networks
2. Modest # of medium sized networks
3. Many small networks

CLASS	*SIZE*	*NUMBER*
A	2G	126
B		
C	254	2M

Original Rationale: Beware the Routing Tables

1. You don't care about most networks.
2. The few networks you do care about, you care about them a lot.
3. Not many routing table entries get you "closer" to a lot of the hosts

Names, Forwarding, and Routes

- A Name is where (IP) you want to get.
  - Eg, DEST = 207.171.178.146
- A Forward Hop is the name of the next closes node
- A Route is a series forward hops
  - EG, SRC = 128.208.200.111
    - DEST = 128.30.2.82 (www.lcs.mit.edu)

```
[lt1:/net/share/vmware] bershad% traceroute www.lcs.mit.edu
traceroute to dib.csail.mit.edu (128.30.2.82), 64 hops max, 40 byte packets
 1  filterqueen-ge1-4.cac.washington.edu (128.208.200.100)  2.588 ms  1.920 ms  1.892 ms
 2  uwbr-ads-01-vl1998.cac.washington.edu (140.142.155.23)  2.213 ms  5.020 ms  2.485 ms
 3  hnsp2-wes-ge-0-0-0-0.pnw-gigapop.net (209.124.176.12)  2.291 ms  6.207 ms  2.491 ms
 4  abilene-pnw.pnw-gigapop.net (209.124.179.2)  3.595 ms  17.458 ms  2.574 ms
 5  dnvrng-sttlng.abilene.ucaid.edu (198.32.8.50)  53.462 ms  28.084 ms  28.649 ms
 6  kscyng-dnvrng.abilene.ucaid.edu (198.32.8.14)  39.770 ms  38.647 ms  42.628 ms
 7  iplsng-kscyng.abilene.ucaid.edu (198.32.8.80)  276.104 ms  276.761 ms *
 8  chinng-iplsng.abilene.ucaid.edu (198.32.8.76)  57.336 ms  52.068 ms  53.371 ms
 9  nycmng-chinng.abilene.ucaid.edu (198.32.8.83)  86.404 ms  142.882 ms  86.026 ms
10  nox230gw1-po-9-1-nox-nox.nox.org (192.5.89.9)  76.934 ms  77.059 ms  76.957 ms
11  nox230gw1-peer-nox-mit-192-5-89-90.nox.org (192.5.89.90)  80.865 ms  85.553 ms  87.329 ms
12  b24-rtr-2-backbone.mit.edu (18.168.0.23)  92.911 ms  77.472 ms  77.152 ms
13  mitnet.trantor.csail.mit.edu (18.4.7.65)  77.064 ms  77.944 ms  77.197 ms
14  trantor.kalgan.csail.mit.edu (128.30.0.246)  77.445 ms  77.347 ms  77.533 ms
15  dib.csail.mit.edu (128.30.2.82)  78.776 ms  77.456 ms  78.357 ms
[lt1:/net/share/vmware] bershad%
```

How can it work?

- Every IP DG contains the ultimate destination
  - DEST field
- Network part of an IP address uniquely identifies a single physical network in the internet
  - Eg, DEST = 128.30.2.82
    - NETWORK = 128.30.2
- All IP end points (hosts and routers) that have the same network address (eg, 128.30.2.[0--255]) are connected to the same physical network*
- Every physical network (A) on the internet is connected to at least one "forwarding host" (router) that is also connected to at least one other physical network (B).
  - A can talk to B, B can talk to A.

Datagram Forwarding

- Strategy
  - every datagram contains destination's address
  - if directly connected to destination network, then forward to host
  - if not directly connected to destination network, then forward to some router
  - forwarding table maps network number into next hop
  - each host has a default router
  - each router maintains a forwarding table
- Example (router R2)

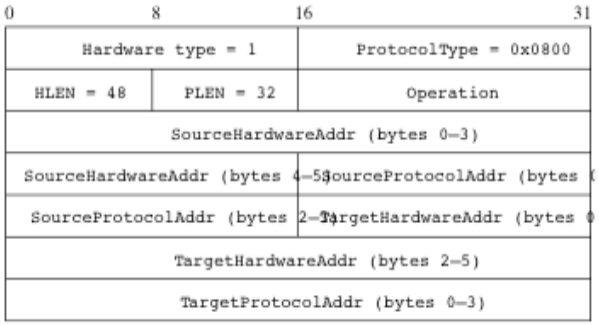
Network Number	Next Hop
1	R3
2	R1
3	interface 1
4	interface 0

May be recursive, eg, how to get to “R3”

Address Translation

- Map IP addresses into physical addresses
  - destination host
  - next hop router
- Techniques
  - encode physical address in host part of IP address
  - table-based
- ARP
  - table of IP to physical address bindings
  - broadcast request if IP address not in table
  - target machine responds with its physical address
  - table entries are discarded if not refreshed

ARP Packets



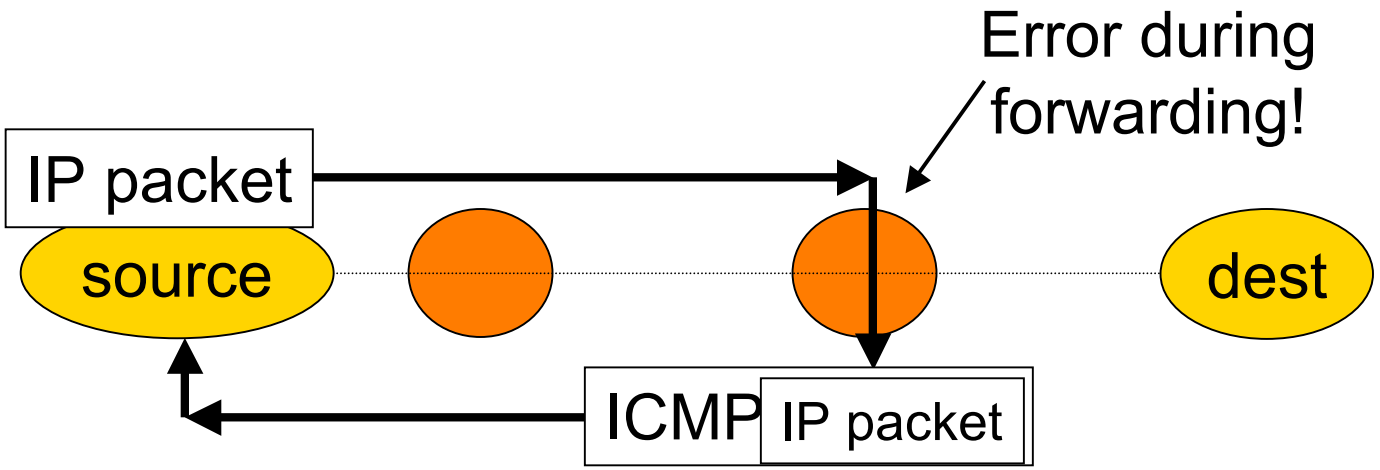
- HardwareType: type of physical network (e.g., Ethernet)
- ProtocolType: type of higher layer protocol (e.g., IP)
- HLEN & PLEN: length of physical and protocol addresses
- Operation: request or response
- Source/Target Physical/Protocol addresses

- Notes
  - table entries timeout in about 10 minutes
  - update table with source when you are the target
  - update table table if already have an entry
  - do not refresh table entries upon reference

ICMP

- What happens when things go wrong?
  - Need a way to test/debug a large, widely distributed system
- ICMP = Internet Control Message Protocol (RFC792)
  - Companion to IP – required functionality
- Used for error and information reporting:
  - Errors that occur during IP forwarding
  - Queries about the status of the network

ICMP Generation



Common ICMP Messages

- Destination unreachable
  - "Destination" can be host, network, port or protocol
- Packet needs fragmenting but DF is set
- Redirect
  - To shortcut circuitous routing
- TTL Expired
  - Used by the "traceroute" program
- Echo request/reply
  - Used by the "ping" program
- Cannot Fragment
- Busted Checksum

- ICMP messages include portion of IP packet that triggered the error (if applicable) in their payload

ICMP Restrictions

- The generation of error messages is limited to avoid cascades … error causes error that causes error!
  - 1->n
- Don't generate new (ICMP) error in response to:
  - An ICMP error
  - Broadcast/multicast messages (link or IP level)
  - IP header that is corrupt or has bogus source address
  - Fragments, except the first
- ICMP messages are often rate-limited too.

Key Concepts

- Network layer provides end-to-end data delivery across an internetwork, not just a LAN
  - Datagram and virtual circuit service models
  - IP/ICMP is the network layer protocol of the Internet