COMPSCI 314 S1 C

Internet Protocol (IP)

Fragmentation and Reassembly

IP routing

The Internet Protocol (IP)

- Defined in RFC791 (available from http://www.ietf.org/) in 1981
- *Version 4* is now deployed as the predominant Internet protocol – unless otherwise mentioned, this is the version referred to here
- Part of a suite of protocols that has collectively become known as TCP/IP
- Provides connectionless best-effort packet delivery between *hosts* on the Internet
- Based on a 32-bit (4 byte) addressing scheme that assists in routing (RFC790)

IP addresses

- An IP address identifies an *interface* on a host (i.e. on a network node)
- One host can have multiple IP addresses, but one IP address cannot be assigned to more than one interface on one host
- An IP address has two parts: *network* and *host*. A network mask (*netmask*) indicates the network bits
- IP addresses originally came in five flavours: Class A, B, C, D, E
- CIDR (Classless Inter-Domain Routing) makes those ‘classes’ obsolete

Class-based IP Addresses

Class A: First octet designates network, last three octets designate host within the network (netmask 255.0.0.0). Network contains 2^{24} addresses.
First bit in first octet of address is always zero.

Class B: First two octets designate network, last two octets designate host within the network (netmask 255.255.0.0). Contains 65536 addresses. First octect starts with 10.

Class C: First three octets designate network, last octet designates host within the network (netmask 255.255.255.0). Contains 256 addresses.
First octet starts with 110.

Class D: Multicast address (packet gets sent to more than one host). First octet starts with 1110.

Class E: Reserved for future use (unlikely).
First octet starts with 11110.

CIDR addresses

CIDR (‘cider’) = Classless InterDomain Routing

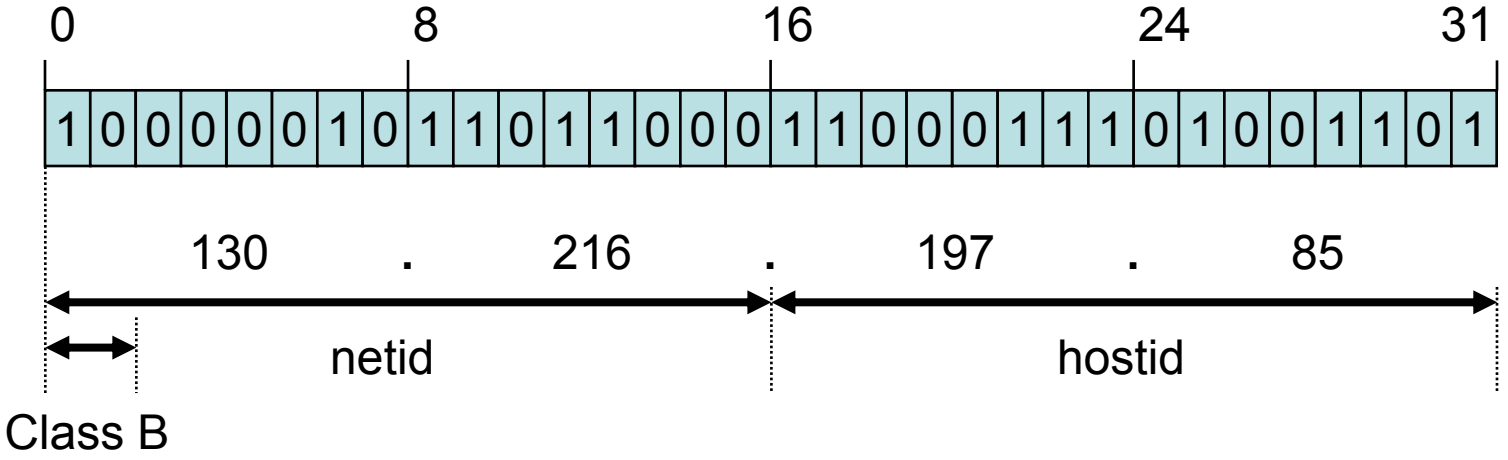
- A newer concept where any number of leading bits can designate the network, with the remaining bits designating the host
- Written with a trailing slash indicating the number of network bits
- e.g.: 130.247.156.87/22 is a host on a network with 2^{32-22} = 1024 addresses

Another example

The University of Auckland has a Class B address, 130.216.0.0/16

We run the network as though it were a set of Class C subnets, 130.216.0.0/24

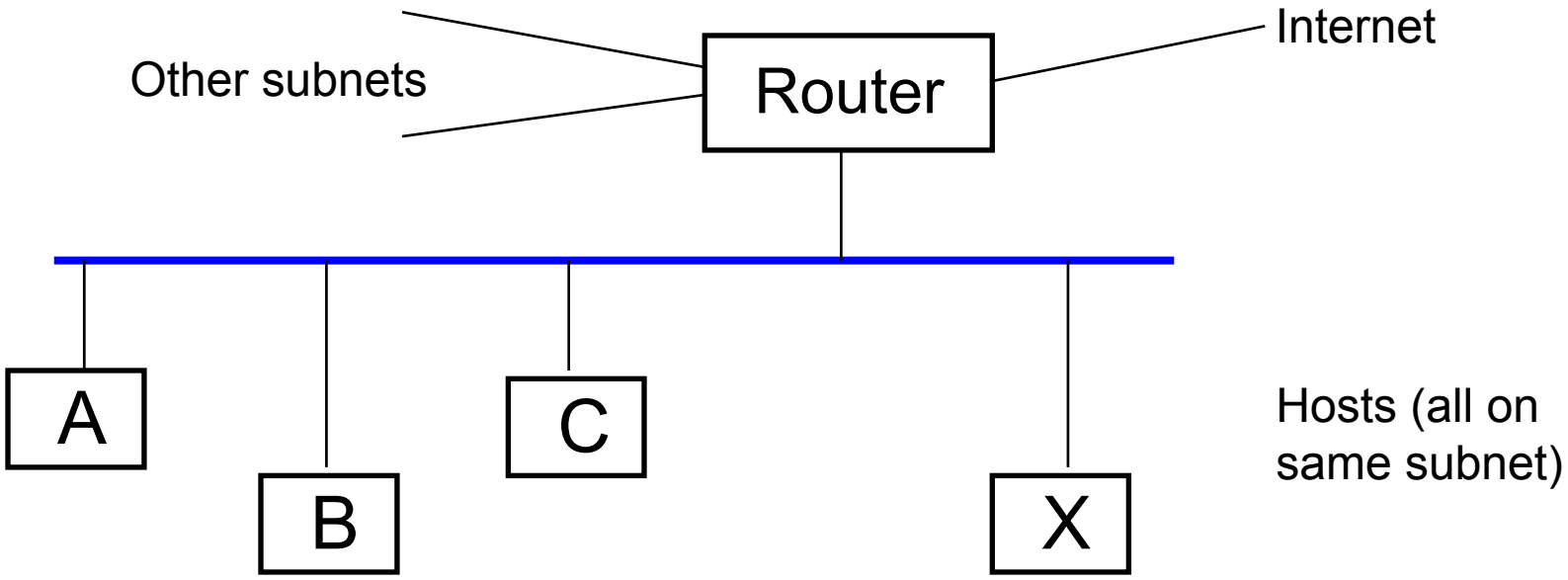
Network addresses like these (with trailing zeros) are usually called *network prefixes*



Special IP addresses

- Network or host bits all set to zero – refers to own host or network
- 255.255.255.255 (all bits set to 1) – broadcast address. Used, e.g., in BOOTP and DHCP for host configuration
- 127.0.0.1 – local loopback address referring to the host itself
- 10.0.0.0 - 10.255.255.255 (10/8 prefix),
  172.16.0.0 - 172.31.255.255 (172.16/12 prefix),
  192.168.0.0 - 192.168.255.255 (192.168/16 prefix)
  These are private addresses reserved for private Internets. Defined in RFC 1918. Very commonly used in conjunction with Network Address Translators (NATs) behind enterprise gateways.
- Firewalls (filtering gateways) are usually configured to drop packets to these addresses as they are not supposed to be used outside the local network

IP host configuration



- Each host knows its *address, netmask* and *network prefix*
- It also knows the IP address of its default router
- As well, it will need to know the IP address of a *nameserver* (more later)

Mapping IP addresses to hardware addresses (MAC addresses)

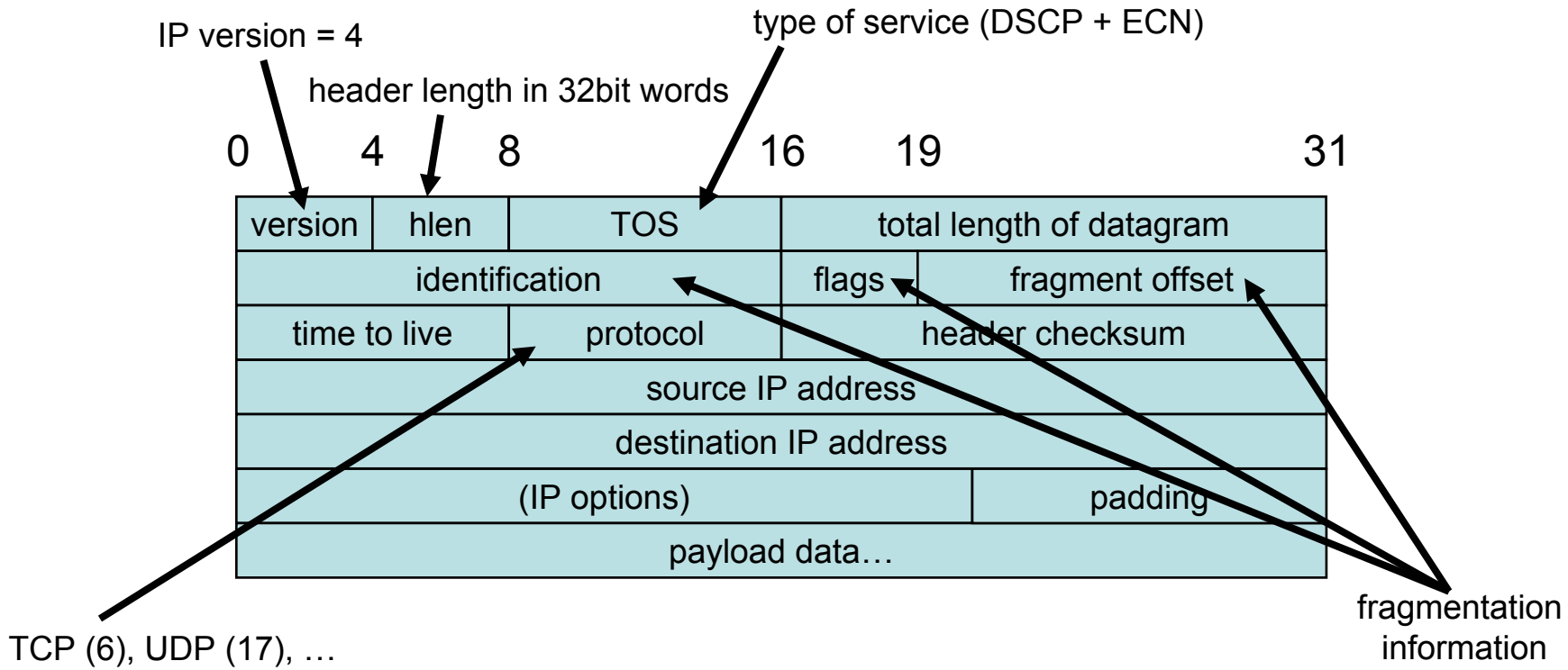
- How do hosts on a shared medium, e.g. Ethernet, recognize that an IP datagram is for them?
- First approach: Could *always* broadcast to all hosts on the network; each host decodes the packet and looks at the IP address: Lots of decoding overhead and what do we do with bridges?
- Second approach: Find out the MAC address of the host first. Need some way of doing this: The Address Resolution Protocol (ARP)

Address Resolution Protocol (ARP)

- Part of the 'TCP/IP family,' Defined in RFC 826
- Gateway or other host that wants to find out a MAC address for an IP number broadcasts an ARP request throughout the (local) network. This broadcast is received by all hosts in the network
- Each host hands this packet to its ARP implementation
- If the IP address in the packet does NOT match the IP address of the host, the host remains silent
- If the IP address matches that of the host, the host replies with an ARP response packet that contains both the IP address and the MAC address of the host
- The requesting host can now cache this mapping for future use

IP packet header

- The IP packet header bytes precede the payload data in the IP datagram (= IP packet)

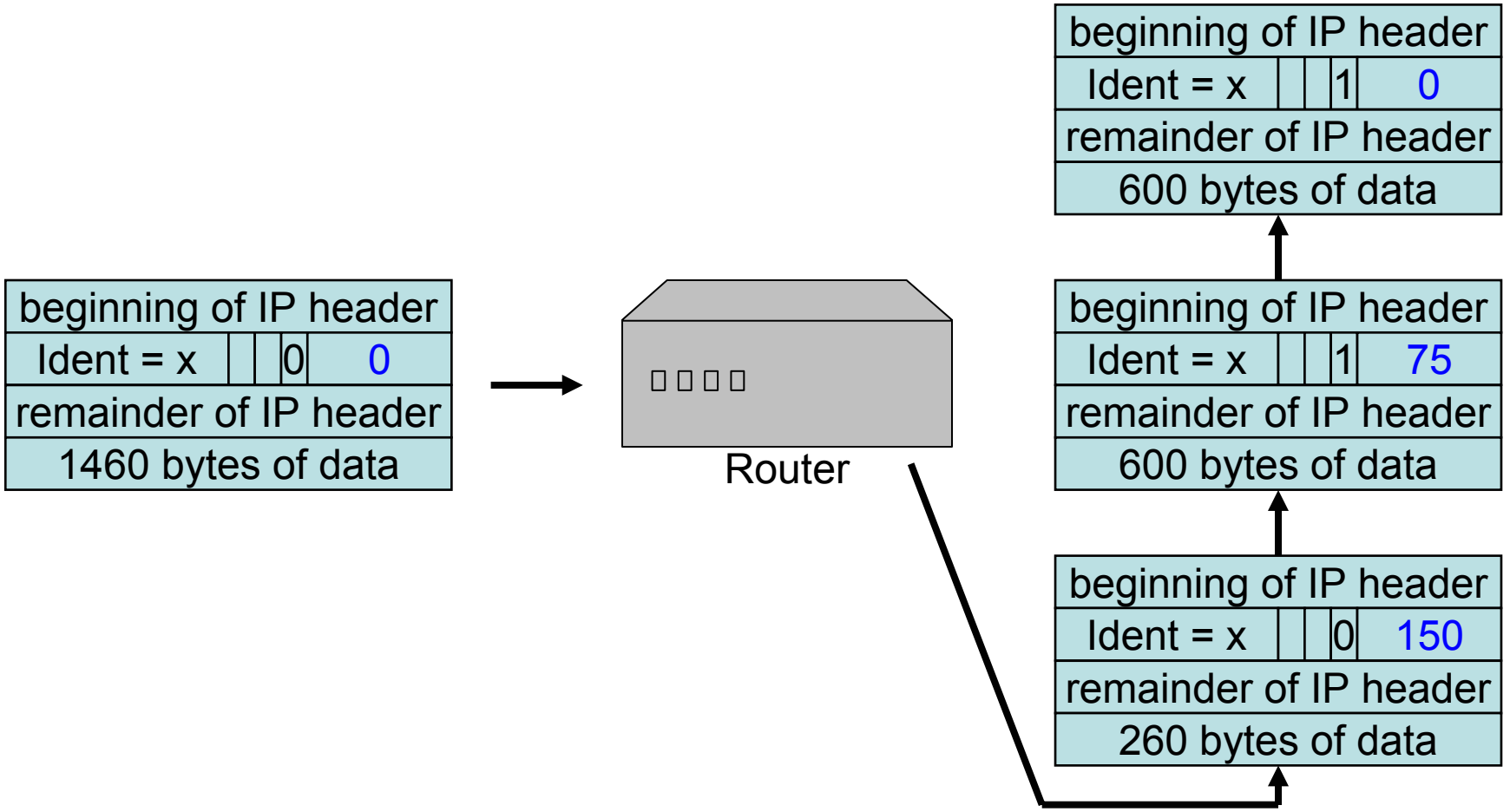


IP fragmentation and reassembly

- Any router along the path may split an IPv4 datagram into two or more fragments
- Each fragment becomes its own IP datagram with its own header
- All fragments are given a common *ident* field – in fact the same ident as the original datagram
- Third bit ('M' bit as in 'more to follow') in the flags field is set in all but the last fragment
- The fragment offset field contains the offset of the fragment's payload within the original payload, in units of 8 bytes.
- All fragments travel separately to the final destination (i.e., they do not get reassembled by the next router with larger MTU)

Example: IP fragmentation

Example: MTU of onward link frame only permits 600 bytes of payload after IP header in frame



Reassembly after fragmentation

- Reassembly **always** happens at the final destination, even if MTU size increases again
- Fragments may have taken different paths and may arrive in different order!
- Further fragmentation of already fragmented datagrams is possible – in this case, if the M bit is already 1, the M bits in the flags of all additional fragments are set to 1
- May need to buffer fragments for a while until the missing parts arrive

Shortcomings of IP

- Severely restricted address space in IPv4 (solved in IPv6)
- IP address reflects physical network topology. If a node moves from one part of the network to another, the IP address cannot stay the same. Bad news for mobile routing. (Current work on *Host Identity Protocol* could solve this)

IP routing

- IP datagrams outside the local network (as determined by the host's netmask) are sent to the local gateway router
- The gateway is another essential part of a host's IP configuration and has an address within the host's netmask
- The gateway routes the datagram by network. Note that IP contains no rule about how that route is obtained, or what 'cost' means, or if it indeed exists at all
- Same applies to all gateways/routers between there and the destination
- IP routing is 'best effort,' i.e., datagrams may get to the destination or they may not
- No notification if delivery fails – any recovery from lost data is left to the end points of the communication
- ICMP (IP Control Message Protocol) can return information, e.g., 'no route to host'

IP Routing Protocols

- Distance Vector (Bellman Ford):
  RIP and RIP2
- Link State (Dijkstra):
  OSPF, IS-IS
- Hierarchical
  BGP

Routing Information Protocol: RIP

- Popular IGP (Interior Gateway Protocol) for IP networks. Based on Bellman-Ford algorithm
- Bundled with BSD Unix as a program called **routed** – pronounced ‘route-d’ (‘d’ as in ‘daemon’)
- Uses routing information broadcasts rather than peer-to-peer updates (i.e., tries to take advantage of shared media such as Ethernet)
- *Broadcasts take place at 30-second intervals*
- Not all nodes broadcast routes, only active nodes do – usually the gateways
- Uses *hop count* metric
- No routing loop detection, not suitable for large networks!

More on RIP: RIP2

- Successor protocol to RIP
- RIP can only handle Class-Based network addresses.
- RIP only sends network addresses (with trailing zeroes) to other nodes. It uses the first few bits to determine each network's class
- RIP2 is CIDR-based, it sends network prefixes
- Simplest routing protocol to use for small- to medium-sized networks

Link State Routing protocols

- Nodes (i.e. routers) send out announcements whenever they, or the links connected to them, change state. Announcements are sent using *reliable flooding*
- When a router receives a state change announcement, it updates its network topology graph, then runs a shortest-path-first algorithm to compute its new routing table
- The announcements are usually called *Link State Packets (LSPs)*

Reliable Flooding Challenges

- Need to remove old data when link fails … how?
  - LSPs carry sequence numbers to distinguish new from old
  - Only accept (and forward) the 'newest' LSP seen from a node
  - Send a new LSP with cost infinity to signal a link is down
- What happens when a node (router) fails and restarts?
  - What sequence number should it use? Don't want data ignored
  - Aging
    - Put a TTL in the LSP, periodically decremented by each router
    - When TTL=0, purge the LSP and flood that LSP to tell everyone else to do the same
  - Or : when receiving an 'old' LSP from a node, tell that node what the current sequence number is, rather than just dropping the LSP

Open SPF (OSPF)

- Open standard proposed by IETF
- Implements SPF/link state algorithm with Dijkstra at core
- Link State Packet (LSP) flooding communicates link states to all nodes
- Supports host-specific routes and network-specific routes
- Supports type-of-service routing and load balancing
- Addresses security issues between gateways (authentication)
- Allows for network partitioning etc.

IS-IS

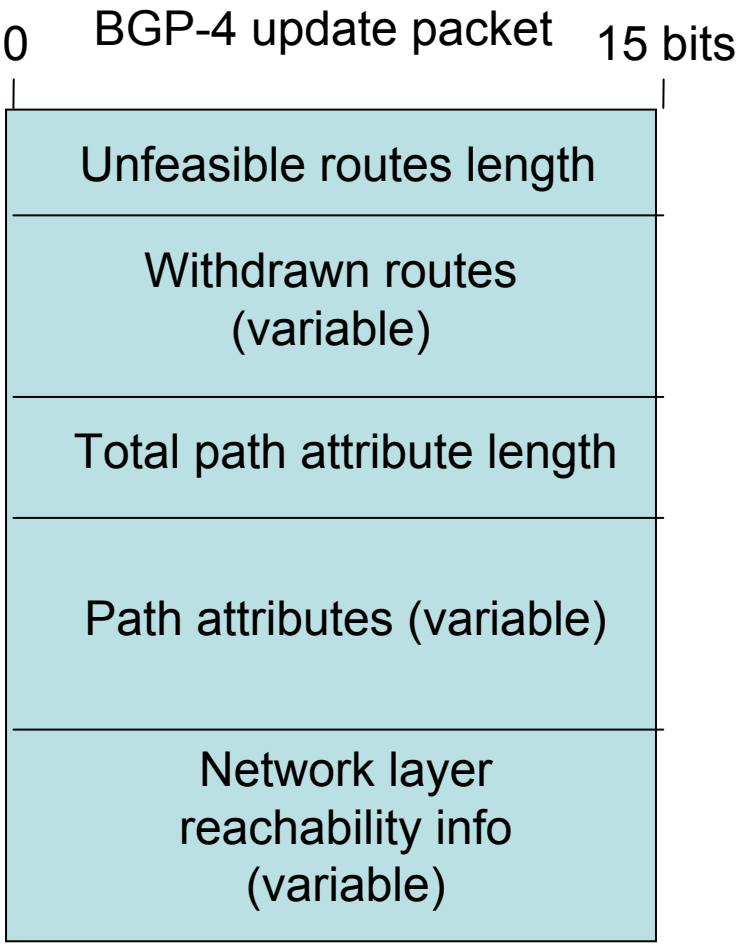
- Open standard proposed by OSI
- Very similar to OSPF, but with fewer ‘added features’
- Widely used as an IGP in large provider networks, e.g. Sprint and AT&T
- Providers often adjust link weights as a way to tune their network, e.g. to move traffic away from overloaded links

Hierarchical Routing: BGP

- We can view the global Internet as a graph linking provider networks together, forming a hierarchy – ‘tier 1,’ tier 2,’ .. down to ‘customer’ networks
- Each network in the graph is described as an Autonomous System (AS), and referred to by an AS number (ASN)
- Autonomous Systems use the Border Gateway Protocol (BGP) as their Exterior Gateway protocol, so as to provide global routing in the Internet
- BGP uses *paths* as its primary ‘cost’ measure
- A BGP *path* is a list of the ASNs indicating the path a packet may traverse to reach a specified *network prefix*

Border Gateway Protocol (BGP)

- Uses a Bellman-Ford (distance vector) type algorithm to route between ASs
- 'Cost' becomes a fuzzy concept – other factors ('policy') can also influence routing decisions
- Swaps 'route information' rather than cost
- 'Getting there' is main concern
- Routing loop suppression
- BGP-4 is most common version

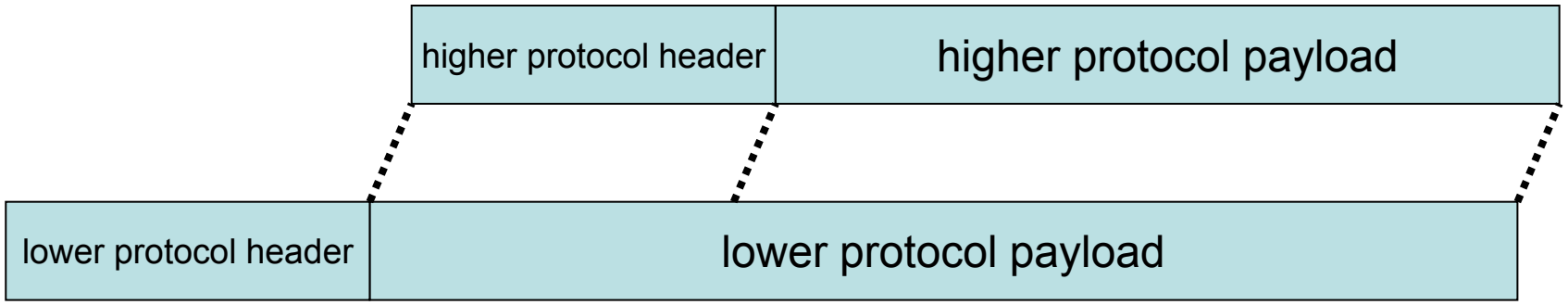


Routing summary

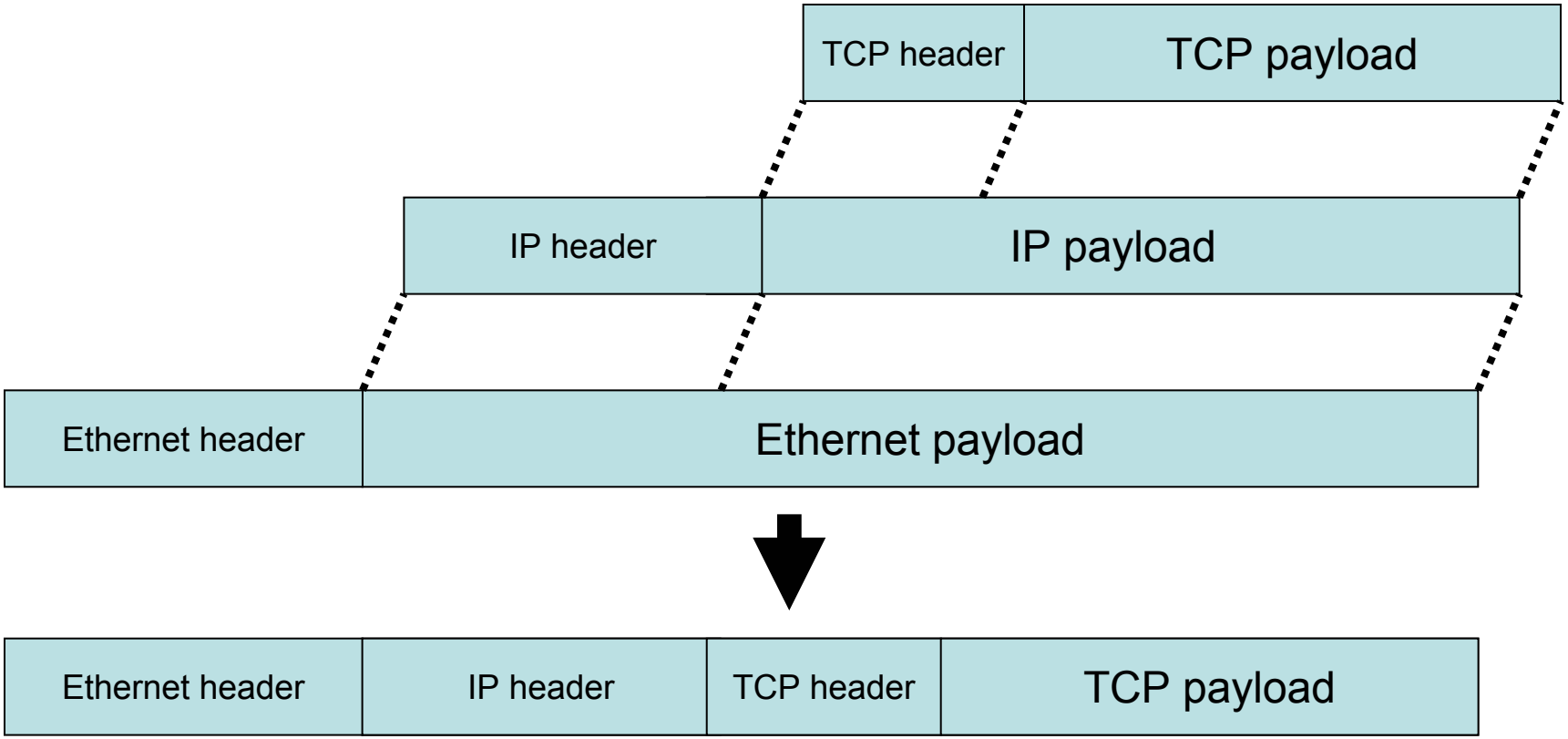
- Scalability is an important aspect
- Reality demands a distributed approach
- Network hierarchies and routing by network reduce complexity – route between gateways at higher levels
- Bellman Ford and SPF make good IGPs
- At top level, ‘cost’ becomes a fuzzy issue
- Can manage routing in very large networks if we compromise ‘cost’ for ‘getting there’

Encapsulation: review

- It is often necessary to transfer data across different network formats
- For example, IP over ATM, TCP over IP, IP over Ethernet, IPv4 over IPv6, IPv6 over IPv4, etc.
- Encapsulation puts one protocol's packet into the payload field(s) of another protocol's packet(s)
- Very widely practised – get used to the idea!



Example: TCP over IP over Ethernet



… anyway, we have yet to look at the details of TCP