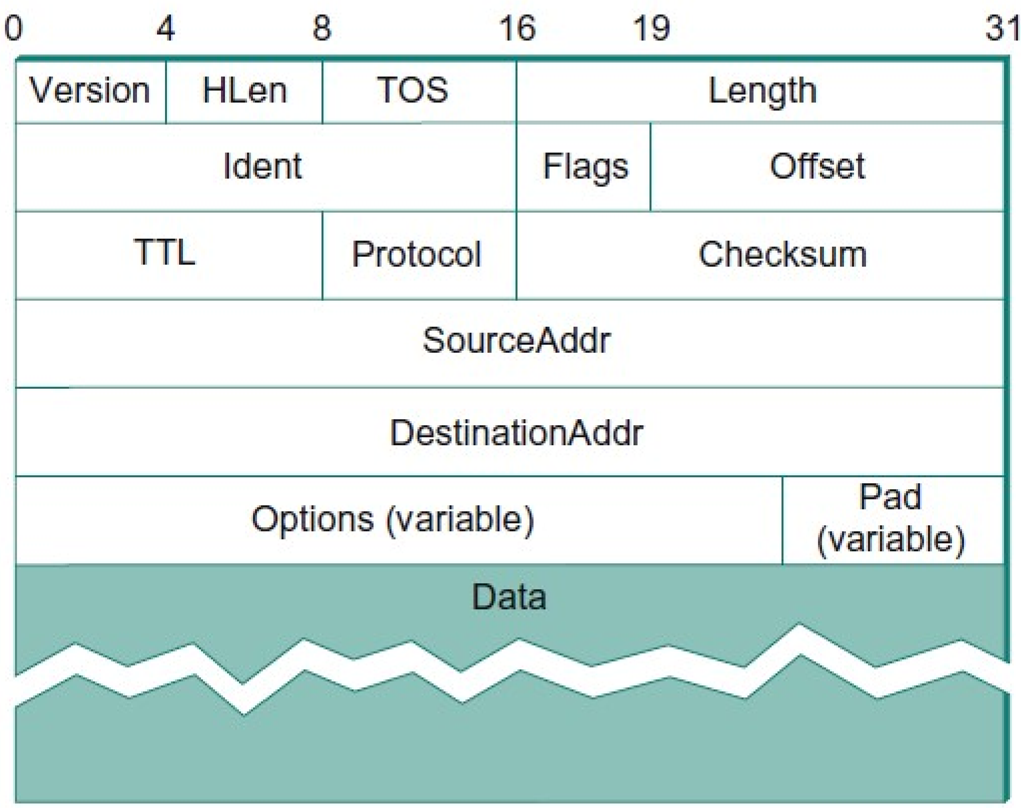
Network Layer

- Network layer vs Link layer
    - Linked layer: moving packets between hosts within a LAN
    - Network layer: moving packets between networks (LANs, or larger networks)
- Two issues at Network layer: addressing + routing.
- Packet format



HLen: header length. Length: Length of packet (header + payload) in bytes. Ident + Flags + Offset: for fragmentation and reassembly. TTL: Time-To-Leave, a hop count to avoid infinite forwarding. The header takes 20 bytes.

- Fragmentation and Reassembly

Networks can have different Maximum Transmission Unit (MTU). *Fragmentation* happens at **router** when it receives a packet that it wants to forward to a destination network with MTU smaller than the packet. Fragments of the same packet has the same `Ident`, a `M` flag set for all fragments except for the last one, and `Offset` being the byte position of the start of the fragment (8-byte aligned). Each fragment is a self-contained IP packet with complete header. This allows fragmentation to be repeated. *Reassembly* happens at the receiving host, instead of at router. If some fragments aren't received, the received fragments are discarded.

Reassembly can be expensive. To avoid fragmentation, *Path MTU discovery* is used to find the minimum MTU on the path prior to sending. The sender sends a packet with "do not fragment" flag. Any router that would have to fragment the packet would drop the packet and sends back a response. Keeps halfing the packet size until finding the suitable one.

- IPv4

  IPv4 address: 32-bit long, organized via "dot notation", separating 8-bit portions. `0.0.0.0` - `255.255.255.255`. An IP address is divided into two parts: *network portion* and *host portion*.

- IP address allocation

  ```
  Class A:   0 + 7-bit  network + 24-bit host
  Class B:  10 + 14-bit network + 16-bit host
  Class C: 110 + 21-bit network + 8-bit  host
  ```

  The network layer only cares about the network portion of the IP address, to move the packet from source network to destination network. Inefficiency: (1) each physical network requires a unique network number; (2) big gaps between class sizes cause low utilization of IP addresses.

  - Subnetting

    A logical network contains multiple physical networks. Instead of allocating a network number to each physical network, allocates a single IP network number to the logical network and allocates portions to the underlying physical networks, called *subnet*s.

    Each subnet has a *subnet mask*. The subnet number is computed by ANDing the subnet mask with the IP address. When a router wants to forward a packet to a certain IP address, it firstly compute the subnet number of the destination. If within the same subnet, send the packet via link-layer protocol; otherwise, send the packet to a rounter to be forwarded to another subnet. The forwarding table is slightly changed to contain `<Subnet number, subnet mask, next hop>`, instead of `<network number, next hop>`.

    In addition to saving network numbers, subnetting also implicitly solves a problem. Consider a geographically large network. LANs have geographical range restriction and cannot be too large. Subnetting enables us using IP protocol to deliver packets to each subnet, probably physically small.

  - Supernetting/Classless Inter-Domain Routing (CIDR)

    Difference between subnetting and supernetting: subnetting relaxes contraint under classful allocation, while supernetting completely abandons the classful allocation scheme.

    Supernetting assign networks on power of 2 and uses `/x` to represent network numbers.

- Network Address Translation (NAT)

  A "private IP address" that's no globally unique, but unique within some limited scope. This's sufficient if the host communicates only with other hosts within tthe same scope. To communicate with hosts outside the scope, it indirectly does so via a NAT box. The NAT box can translate a private address to a globally unique address assigned to the NAT box. This mechanism allows less globally unique IP address to be allocated.

- IP address allocation

  Statically, or dynamically via Dynamic Host Control Protocol (DHCP). If a host doesn't have an IP address assigned, it sends a request to `255.255.255.255`. The DHCP server would respond with an IP address **via link-layer**.

- IP forwarding

  Each IP packet contains destination IP address. The network portion uniquely identifies a *physical network*. Routers and hosts sharing the same network number is in the same LAN and can communicate via link-layer protocol. The goal is to move the packet from the **source network** to the **destination network** (instead of source host to destination host).

  Suppose we already have the forwarding table established by the routing process, containing `<Network num, Next Hop>` entries. When a router receives a packet, first examine whether the dst is in the same physical network (by checking network number with the network number of each of its interfaces). If yes, send via link-layer protocol using ARP for address translation (see below). Otherwise, sends on the `next hop` interface. Normally there's also a default router in case no match found.

- Address Resolution Protocol (ARP)

  When the destination host is in the same network as the router, link layer protocol is used to forward the packed to the destination host, and thus MAC address is needed (the packet doesn't contains MAC address since the sender might not know the MAC address of the receiver).

  ARP enables **each host** to build a table (called *ARP cache*) of `<IP address, MAC address>` mapping. When trying to forward a IP packet, if no mapping is found, the host broadcasts an ARP query to the network. The query contains: sender's IP address, sender's MAC address, and target IP address. The receiver adds the `<sender IP, sender MAC>` mapping to its own ARP cache. If the receiver's IP address matches the query, it responses its MAC address via the link-layer protocol. The sender then adds the mapping to the ARP cache.

- Three kinds of forwarding tables

  - switch: `<IP, MAC address>`
  - intra-network router (with subnetting): `<Subnet number, subnet mask, interface>`
  - inter-network router: `<Network number, interface>`

- Internet Control Message Control (ICMP)

  Designed to provide error reporting, usually enabled on routers.

- Routing

  Process to establish forwarding table at routers (not switches). A *domain* is an internetwork in which all routers are under the same administrative control.

  - Intra-domain routing

    Goal: shortest path between nodes.

    Protocols:

- Routing Information Protocol (RIP), based on *distance vector routing* via Bellman-Ford.

  Each node constructs a distance-vector containing distances to all other nodes and distributes the vector to its immediate neighbors, assuming each node knows the cost to its immediate neighbors.

  When a node receives a distance vector, it recomputes and possibly update its own distance vector. In two cases a node sends distance vector to its neighbors: (1) periodically, (2) after it updates its distance vector or when it notices a link is down.

  The problem with RIP is the *count to infinity* problem: when a link is down, it can take long for the system to stablize. Possible solution is to use smaller value as approximation of infinity, a typical value is 16. This limits RIP to small/medium-sized network.

- Open Shortest Path First (OSPF), based on *link state routing* via Dijkstra's algorithm.

  The most widely used intra-domain routing protocol.

  Each node knows the cost to reach its direct neighbors, and tries to disseminate this information throughout the network in *link-state packet*, with *reliable flooding*.

  Each link-state packet (LSP) contains: (1) ID of the node that created the LSP, (2) a list of directly connected neighbors of that node, with the cost of the link to each one, (3) a sequence number, and (4) a TTL (time to leave) for the packet. Reliable LSP transmission between adjacent routers is made using acknowledgements and retransmissions.

  When a node X receives a copy of an LSP **originated** (but might be forwarded by some other node Z) from node Y, X checks if it has already stored a copy of an LSP from Y. If no, it stores the LSP. Otherwise, it compares the sequence numbers; If the new LSP has a larger sequence number, i.e., more recent, X replaces the LSP. If the new LSP has a smaller sequence number, it gets discarded. If the received LSP is the newer one, X sends a copy of the LSP to all its neighbors except the neighbor from which the LSP was received, this helps to bring an end to the flooding of LSP. Eventually, the most recent copy of the LSP would reaches all nodes in the network.

  Each node generates LSPs either **periodically**, or if one of its **immediate neighbors goes down**.

  ```
  N = the set of nodes in the graph
  M = set of nodes incorporated
  l(i, j) = cost with edge between nodes i, j, and l(i, j) =
  infinity if no edge
  s = source node
  C(n) = cost of path from s to n

  Algorithm:
      M = {s}
      for each n in N - {s}
          C(n) = l(s,n)

      while (N != M)
  ```

```
        M = M + {w} where C(w) is the min for all w in (N-M)
        for each n in (N-M):
            C(n) = MIN(C(n), C(w)+l(w,n))
```

In distributed manner, the algorithm above works as follows. Each router maintains two lists `Tentative` and `Confirmed`. Each list contains entries of the form `(Destination, Cost, NextHop)`. Once a given node has a copy of the LSP **from every other node**, it's able to compute the distance to every other node:

1. Initialize the `Confirmed` list, `M`, with an entry for myself; this entry has a cost of 0.
2. For the node just added to the `Confirmed` list, call it node `Next`, `w`, and select its LSP.
3. For each neighbor (`Neighbor`) of `Next`, calculate the cost (`Cost`) to reach this `Neighbor` as the sum of the cost from myself to `Next` and from `Next` to `Neighbor`, `C(w)+l(w,n)`.
    1. If `Neighbor` is currently on neither the `Confirmed` nor the `Tentative` list, then add `(Neighbor, Cost, NextHop)` to the `Tentative` list, where `NextHop` is the direction I go to reach `Next`.
    2. If `Neighbor` is currently on the `Tentative` list, and the `Cost` is less than the currently listed cost for `Neighbor`, then replace the current entry with `(Neighbor, Cost, NextHop)`, where `NextHop` is the direction I go to reach `Next`, `C(n) = MIN(C(n), C(w)+l(w,n))`.
4. If the `Tentative` List is empty, stop. Otherwise, move the entry with the minimum cost in `Tentative` to `Confirmed`, and go to step 2.

Link-state routing algorithm stabilizes quickly, generates little traffic, has **no problem of count to infinity** and responds rapidly to topology changes. The con is that the amount of information stored at each node can be quite large. Why OSPF is preferred over RIP? **Faster, loop-free convergence**.

One practial detail about intra-domain routing: the cost communicated and computed are cost to *subnets*, instead of routers. The responsibility for intra-domain routing is to move packet to the **destination subnet**.

- Inter-domain routing overview

While intra-domain routing moves packet to the destination subnet, inter-domain routing moves packet to the destination network (which contains subnets). The problem with intra-domain routing schemes, like RIP or OSPF, is that it requires a router to know all subnets in the network, and all subnet numbers are exchanged in the routing protocol. This doesn't scale to the hugh Internet.

Autonomous system (AS): an administratively independent network. We consider the Internet as a connected ASes. The other name for AS is routing domain. So routing within an AS is called intra-domain routing, and routing between ASes is called inter-domain routing. The basic idea behind AS is to provide an additional way to hierarchically aggregate routing information in a large internet. The AS model also decouples the intra-domain routing from inter-domain routing, so each AS can run its own intra-domain routing protocol. Under the setting of AS, each network is identified by its IP network number, and the AS Number (ASN), a 32-bit identifier.

Routing policy can be complicated when commercial contracts are involved.

Goals for inter-domain routing: (1) minimize number of network number exchanged/stored; (2) Loop-free path to destination network; (3) The path should be compilant with possibly complicated policies.

In contrast to intra-domain routing with tries to find shortest path, inter-domain routing tries to find loop-free, policy-compliant path.

Traffic classification: local traffic vs transit traffic.

AS classification: *stub AS* has single connection to one other AS, carries only local traffic. *Multi-homed AS* has multiple connections to other ASes, carrying only local traffic. *Transit AS* has multiple connections to other ASes, carrying both local and transit traffic.

Relationship between ASes can be: customer-provider, peer. The relationship bring hierarchy into the Internet: at the bottom are stub ASes, higher in the hierarchy are ASes being providers and customers are the same time, at the top are pure providers.

- Border Gateway Protocol (BGP)

  Each AS has >= 1 *border routers* through which **packets enter and leave** the AS (the only entrance & exit of the AS), aka *gateways*. A border is just an IP router forwarding packets between AS's.

  Each AS must also have >= 1 *BGP speaker*, **a router** that "speaks" BGP to other BGP speakers in other AS's. It's common to have border routers as BGP speakers. Keep in mind that **a BGP speaker is a router**.

  BGP advertises complete paths as a vector of ASes to a particular network, to support policies and detect routing loops. Each AS needs an unique identifier.

  Example: speaker in AS X sends speaker in AS Y the path to AS Z: `path(X, Z) = X, H1, H2, ..., Z`. If Y selects the path (Y, Z) from X, it will send `path(Y, Z) = Y, path(X, Z)`.

  When a router receives a path vector, if it finds itself in the path, it doesn't propagate it. A given AS only advertises routes that it considers good enough. If a BGP speaker has several choices of several different routes to a destination, it'll choose the best one according to its policies and advertise it. A BGP speaker is not obliged to advertise any route to a destination, even if it has one. This's how an AS can be implemented not to provide transit service.

  As link fails and policies changes, BGP speakers need to cancel previously advertised paths. This is done with a form of negative advertisement called *withdrawn route*. So each BGP update message contains both withdrawn routes and reachable routes.

  BGP runs on TCP's reliable transmission.

- Why BGP helps to build scalable network?

    1. The number of nodes participating in BGP is on the order of number of AS's, instead of the number of networks. Usually, each AS contains many networks.

    2. Finding a good inter-domain route is only about finding a path to the border router of the destination AS, of which there are only a few per AS.

    In this way, the complexity of inter-domain routing is on the order of number of AS's, and the complexity of intra-domain routing is on the order of the number of networks in a single AS.

  - Integrating intra-domain and inter-domain

    In the case of a stub AS that only connects to other AS at a single point, the border router is the only choice for all routes that are outside the stub AS. Thus, the border router can inject a *default route* into the intra-domain routing protocol, stating that any network that has not been explicitly advertised in the intra-domain protocol is reachable through the border router. The default entry is the last one in the forwarding table and matches anything that failed to match a specific entry.

    The next step is to have the border routers inject specific routes they learned. Consider, for example, the border router of a provider AS that connects to a customer AS. That router could learn that the network prefix 192.4.54/24 is located inside the customer AS. It could inject a route to that prefix into the routing protocol running **inside the provider AS**. This would cause other routers in the provider AS to learn that this border router is the place to send packets destined for that prefix.

Multicast, IPv6, Mobile IP are skipped for now. Don't seem like popular interview topics.