Network Layer

Arnav Gupta

November 28, 2024

Contents

1 Overview

Network layer protocols transport segments from sending to receiving hosts. The sender encapsulates segments into datagrams and passes them to the link layer. The receiver delivers segments to the transport layer protocol.

Every Internet device uses network layer protocols.

Routers examine header fields in all IP datagrams passing through them and move datagrams from input interfaces to output inferfaces to transfer datagrams along end-to-end paths.

Key network-layer functions:

- **forwarding**: move packets from a router's input link to appropriate router output links (done for all datagrams very fast)
- **routing**: network-wide process that determines route taken by packets from source to destination to fill forwarding tables (done in the background, takes longer), uses routing algorithms

The **data plane** is a local, per-router function and determines how a datagram arriving on router input interface is forwarded to router output interface.

The **control plane** uses network-wide logic to determine how a datagram is routed among routers along end-to-end paths from source host to destination host.

Control plane approaches:

- **traditional routing algorithms**: implemented in routers (both planes implemented monolithically within a router)
  - individual routing algorithm components in every router interact in the control plane
- **software-defined networking**: explicitly separate the two planes by implementing the control plane as a service in remote servers
  - remote controller computes and installs forwarding tables in routers

1.1 Best Effort Service

The Internet runs on a "best effort" service model, so there are no guarantees on:

- successful datagram delivery to the destination
- timing or order of delivery
- bandwidth available for end-to-end flow

Best effort is simple and has allowed Internet to be widely adopted. Successful provisioning of bandwidth allows performance of real-time applications to be good enough for most of the time.

Replicated, application-layer distributed services connect close to client networks which allows services to be provided from multiple locations.

2 Data Plane: The Internet Protocol

Data plane is:

- connectionless (datagram-based)
- best-effort delivery
  - packets can be lost, delivered out of order, or delayed
- a common packet format for IPv4 and new packet format for IPv6
- global addressing for identifying all hosts (ARP)
- sister protocol that performs error reporting and enables signaling between routers: ICMP (v4, v6)

IPv4 datagram has

- IP version number
- header length (bytes)
- type of service
- total datagram length (bytes)
- 16-bit identifier
- flag
- fragmentation/reassembly and offset info
- time to live: remaining max hops
- upper layer protocol (TCP or UDP)

- header checksum
- 32-bit source IP address
- 32-bit destination IP address
- options (timestamp, record route taken, etc)
- payload data

When no options, overhead is 20 bytes. Upper layer protocol can be:

- ICMP: 1
- TCP: 6
- UDP: 17
- IPv4: 4
- IPv6: 41

If header checksum detects an error, datagram is dropped. Must be recomputed at every hop because of TTL and options.

2.1 Fragmentation and Assembly

Link layer protocols have maximum transfer unit size, which is the largest possible data size in a frame.

Large IP datagrams can be divided (fragmented) within a network, and then reassembled at the final destination. IP header bits are used to identify and order related fragments.

A receiver cannot hold fragments forever and fragments can arrive out of order, so loss of fragments can mean loss of entire datagram. The receiver starts a timer when the first fragment of a datagram arrives. If the timer expires before all the fragments are received, those already received are discarded.

Fragmentation complicated routers and end-systems, which is used by attackers.

2.2 IPv4 Addressing

IP address: 32-bit identifier associated with each host or router interface (about 4 billion total)

Interface: connection between host/router and physical link

- routers typically have multiple interfaces
- hosts typically have 1 or 2 interfaces

2.2.1 Subnets

The addressing scheme is **Classless InterDomain Routing (CIDR)**:

- IP addresses have a subnet part and a host part
- the subnet portion of the address has arbitrary length
- the address has format a.b.c.d/x where x is the prefix, the number of bits in the subnet portion

Subnet: set of interfaces that have IP addresses with the same prefix and same subnet portion, and can physically reach each other without passing through an intervening router

A subnet mask is x 1 bits followed by 0 bits, which is bitwise ANDed to the IP address. With masking, for the same subnet, can send a datagram directly to the destination and for a different subnet, send the datagram to the router.

2.2.2 DHCP

An IP address can be hard-coded by sysadmin in the config file or using **DHCP (Dynamic Host Configuration Protocol)** which dynamically gets address from a server when the host joins the network.

DHCP allows:

- addresses to be renewed on use
- reused since only holding addresses while host connected/on

For a client-server protocol using UDP, DHCP is a network function implemented as an application protocol:

- host broadcasts DHCP discover message
- DHCP server responds with DHCP offer message
- host requests IP address with a DHCP request message
- DHCP server sends the address with a DHCP ack message

Only last 2 steps are needed if the client remembers and wishes to reuse a previously allocated network address.

A DHCP server is co-located in the router, serving all subnets to which the router is attached.

DHCP also returns:

- address of first hop router for client
- name and IP address of DNS server
- address prefix (indicating network vs host portion of address)

For a network to get the subnet part of the IP address, it gets allocated a portion of its provider ISP's address space.

2.2.3 Hierarchical Addressing

Allows efficient advertisement of routing info.

When the router receives a datagram with some destination address:

1. mask the destination address with the mask for that row of the forwarding table
2. check if the results correspond to the value in the table, if so then remember as candidate for forwarding, and then regardless continue to the next row
   (a) if there are no candidates, datagram sent to otherwise output interface
   (b) if there is one candidate, datagram sent to the corresponding output interface
   (c) if there are multiple candidates (due to multiple routes to some host), the most specific one is taken: **longest prefix rule**

To get a block of addresses, ICANN (Internet Corportation for Assigned Names and Numbers) allocates IP addresses through 5 regional registries (RRs) and manages DNS.

DHCP and NAT help with IPv4 address space exhaustion. IPv6 has 128-bit address space.

2.2.4 Network Address Translation (NAT)

Limiting the number of addresses and re-using IP addresses in a smart way.

Addresses in a private address space are not routable outside it and can be reused as much as desired.

Advantages:

- range of addresses not needed from ISP: just one IP address for all devices
- can change addresses of devices in local network without notifying outside world
- can change ISP without changing addresses of devices in local network
- devices inside local net not explicitly addressable or visible by outside world

Host creates IP datagrams with source and destination IP address that carries 1 transport layer segment. Each segment has source and destination port numbers (Layer 4 entity).

Client knows the port number at the server for the service it needs and OS selects a unique source port number.

Receiving host uses IP addresses and port numbers to direct segment to appropriate process (via sockets).

All devices in local network share just one IPv4 address to the outside world:

- all datagrams leaving local network share same source NAT IP address but different source port numbers
- datagrams with source or destination in local network have subnetted address for source and destination, as usual

The NAT router must transparently:

- for outgoing datagrams, replace source IP address and port number of every outgoing datagram to NAT IP address and new port number
  - remote clients and servers will respond using the NAT IP address and new port number as destination address

- remember in the NAT translation table every pair of source IP address and port number to NAT IP address and new port number translation mapping
- for incoming datagrams, replace the NAT IP address and new port number in the destination fields of every incoming datagram with the corresponding source IP address and port number stored in the NAT table

The 16-bit port number field allows 60000 simultaneous connections with a single public IP address.

The router keeps NAT entries in the translation table for a configurable length of time. For TCP connections, default timeout is 24 hours. Since UDP is not connection based, default timeout is 5 minutes.

NAT is controversial since:

- routers should only process up to layer 3
- address shortage should be solved by IPv6
- violates end-to-end argument since port number is manipulated by network-layer device
- can be tricky if client wants to connect to the server behind NAT

2.2.5 Middleboxes

Any intermediate box performing functions apart from normal, standard functions of an IP router on the data path between a source host and destination host.

Includes NAT, application-specific, firewalls, intrusion detection systems, load balancers, and caches.

The internet has a thin waist, since there is a single network layer protocol: IP that must be implemented by every Internet-connected device (compared to many protocols in other layers).

Middleboxes give love handles that operate inside the network.

2.3 IPv6 Addressing

Possibly not enough 32-bit IPv4 addresses. Also, IPv4 is slow (variable length header). IPv6 allows different network-layer treatment of flows and better mobility management.

IPv6 datagram has:

- IP version
- priority among datagrams in flow
- flow label: identify datagrams in same flow
- payload length
- next header
- hop limit
- 128-bit source address
- 128-bit destination address
- data payload

Compared to IPv4, has no checksum, fragmentation/reassembly, and options.

Not all routers can be upgraded simultaneously so must operate with mixed IPv4 and IPv6.

Tunneling: IPv6 datagram is carried as payload in IPv4 datagram among IPv4 routers (packet within a packet)

44.5% of clients access services via IPv6, so takes time to deploy.

3 What's Inside a Router

High-level view of generic router architecture has **routing processor** (control plane) and **high-speed switching fabric** (data plane).

Input ports have a physical layer, link layer, and decentralized switching (using header field values, lookup output port using forwarding table in input port memory).

Destination-based forwarding: forward based only on destination IP address

Generalized forwarding: forward based on any set of header field values

3.1 Switching Fabrics

Transfer packets from input link to appropriate output link.

Switching rate: rate at which packets can be transferred from inputs to outputs, measured as multiple of input/output line rate

For N inputs, switching rate of N times the line rate is desirable.

Major types of switching fabrics are:

- memory
- bus
- interconnection network

3.1.1 Switching via Memory

Used traditionally, with switching under direct control of CPU.

The packet is copied into system memory and speed is limited by memory bandwidth (2 bus crossings per datagram).

3.1.2 Switching via Bus

Datagram from input port memory to output port memory via shared bus.

Bus contention: switching speed is limited by bus bandwidth

3.1.3 Interconnection Network

Initially developed to connect processors in multiprocessor.

Multistage switch: $n \times n$ switch from multiple stages of smaller switches

With parallelism:

- fragment datagram into fixed length cells on entry
- switch cells through the fabric and reassemble datagram at exit

Can scale by using multiple switching planes in parallel.

Cisco CRS router:

- basic unit has 8 switching planes
- each plane has a 3 stage interconnection network

- up to 100s of Tbps switching capacity

3.2 Port Queuing

3.2.1 Input Port Queuing

If switch fabric slower than input ports combined, queuing may occur at input queues. This can lead to queuing delay and loss due to input buffer overflow.

Head of the Line (HOL) blocking: queued datagram at front of queue prevents others in queue from moving forward

3.2.2 Output Port Queuing

Buffering occurs when arrival rate via switch exceeds output line speed.

Queuing delay (and loss) due to output port buffer congestion (overflow).

Buffering is required when datagrams arrive from fabric faster than the link transmission rate. Must have **drop policy** to decide which datagrams to drop if no free buffers.

Scheduling discipline chooses among queued datagrams for transmission. With priority scheduling, this decides who gets best performance.

4 Control Plane: ICMP, Routing

Even though best effort, IP attempts to avoid errors and report problems when they occur.

IP does introduce errors or ignore all errors.

Errors detected can be:

- corrupted header bits: header checksum
- illegal addresses: routing tables
- routing loop: TTL field
- fragment loss: timeout

4.1 ICMP

Internet Control Message Protocol is a separate protocol for errors reporting and information. Required part of IP (just above IP, layer 3.5) and sends error message to original source.

Used by hosts and routers to communicate network-level info like error reporting (unreachable host, network, port, protocol, etc.) and uses echo request/reply (used by ping).

Network-layer above IP so ICMP messages carried in IP datagrams.

ICMP message has a type, code and first 8 bytes of IP datagram causing error.

IP datagram header contains a bit to specify no fragmentation allowed, which can be bit 0 $\rightarrow$ must be zero, bit 1 $\rightarrow$ don't fragment, bit 2 $\rightarrow$ more fragments.

ICMP sends an error message when fragmentation required but not permitted. This is done by probing to find the largest MTU that does not generate an error message. This MTU is not guaranteed if routes change.

For traceroute (provides delay measurement from source to router):

- source sends sets of UDP segments to destination with an unknown port number where each set has an increasing TTL starting from 1
- datagram in set n arrives to router n where router discards datagram and sends source ICMP message which possibly includes name of router and IP address
- when ICMP message arrives at source, it records RTTs

The stopping criteria for traceroute is that the UDP segment eventually arrives at destination host. The destination returns ICMP "port unreachable" message and so the source stops.