Computer Networks and Distributed Systems

Network Layer

Dr Fidelis Perkonigg

March 1, 2018

Overview

- Routers/Gateways
- Routing Strategies and Algorithms
- Internet Protocol (IP)
- Datagrams (packets)
- IP addressing
- Other protocols (ARP, DHCP, ICMP)

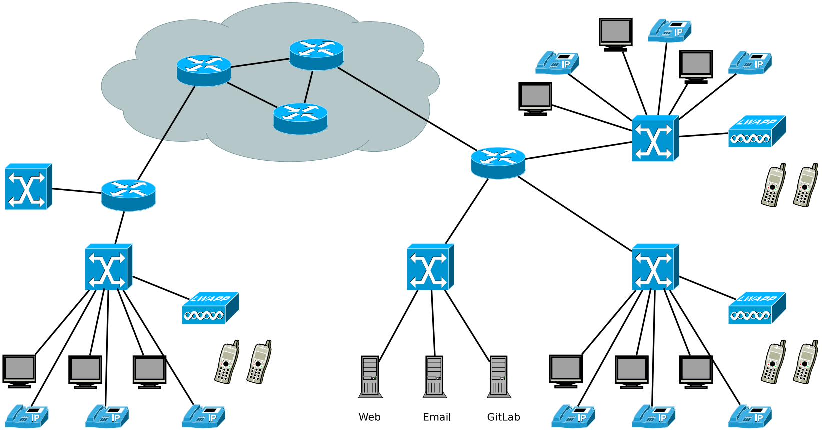
Network Layer Services

- Provides end-to-end transmission of data
- Global addressing and routing
- Hides differences in underlying networks
- Uses data link layer to provide transmission over single hops

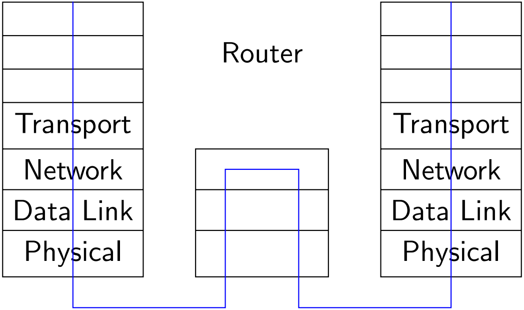
Routing

- Problem: No single network can serve all users
  - Too much traffic, too complex for lower layers, cannot maintain complete network plan
- Solution
  - LANs (subnets) interconnected using routers
  - Routing refers to selecting path from source to destination across multiple subnets

Example Topology



Router/Gateway



- Operates at network layer
- Router forwards packets based on destination networks, unlike bridges, which use hosts
- Lookup in routing table
- Verifies/modifies packets
    - Updates fields affected by routing
    - Checks/recalculates checksum
- Adds transmission, processing delays, and potentially queuing

Router/Gateway

- Typically used for connecting sites
  - Overcome physical and administrative boundaries
  - Greater management and traffic isolation
- Not transparent to end nodes
  - Host needs to know whether/which router to send to

- **Correctness**: Find a route (if it exists)
- **Efficiency**: Routes should provide good performance (should use minimal resources)
- **Robustness**: Route even when links/nodes fail
- **Adaptability**: Routes should reflect network conditions without overreacting
- **Fairness**: Hosts should have equal access to network but Quality of Service (QoS) should be respected
- **Simplicity**: Cheap, predictable and verifiable

Find routes with good properties in terms of:

- Available bandwidth
- Delay
- Hop count
- Price
- Priority for traffic types

- No centralised control
- Knowledge of the whole topology or underlying protocols does not exist
- May use intermediate networks to get to destination

Routing Strategies

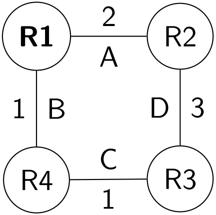
- Static (non-adaptive) routing
  - Compute routes once and load into router
  - Worked for early ARPANET
- Dynamic (adaptive) routing
  - Change routes to reflect changes in topology and load (as seen through congestion)
  - Usually used in packet-switched networks
  - 2 major classes of popular algorithms: Distance Vector Routing and Link State Routing

Static Routing

- Routing using fixed routing tables
- Often used with list of known hosts/networks/links
- All packets for host pair always take same route
- Default link for unknown hosts
- Most workstations use static routing; route most traffic to default gateway/router

- Send packet to all neighbours except source
- Unless packet seen before (add sequence number to remove loops)
- But inefficient and leads to high load on network
- Shortest path and fast discovery
- Extremely robust (data is delivered if there is at least one path)

- ARPANET used this routing protocol
- Implemented as Routing Information Protocol (RIP)
- Router maintains table of distances (vectors)
  - Usually hops/delay/queue length to destination network
  - Periodically exchanges this information with neighbours
  - Re-computes distance and updates its tables



	R1	R2	R3	R4
R1 Vector	0 .	2 A	. .	1 B
R2 Vector	2 A	0 .	3 D	. .
R3 Vector	. .	3 D	0 .	1 C
R4 Vector	1 B	. .	1 C	0 .
R1 Vector	0 .	2 A	2 B	1 B

- Needs time to converge
- What if R3 goes down?
- Count-to-infinity problem

Distance Vector Problems

- Poor efficiency
  - Slow to converge after changes (especially "bad news")
    - Main reason for its demise
  - Count-to-infinity problem
  - Distance vectors increase linearly with network size and may not fit inside packet
- Route finding suboptimal
  - Only considers delay or hop count not bandwidth of links
  - Routing tables do not include paths

- Each router maintains (partial) map of network
  - Consists of more than just neighbours
  - Includes cost metrics (e.g. distance, delay, bandwidth, cost)
- Properties
  - Faster convergence and more reliable
  - Less bandwidth intensive than distance vector routing
  - But more complex and memory/CPU intensive
- Variants used today: IS-IS (Intermediate System to Intermediate System) and OSPF (Open Shortest Path First)

- Each router does the following:
  - Discover identities of all neighbours through HELLO packets
  - Set/measure metric of link to neighbours (automatic or set by administrator)
  - Send this information to all other routers (Link State packet)
  - Also, receive Link State packets from other routers
  - Compute shortest path to every other router using Dijkstra's algorithm
- When link state changes
  - Notification packet flooded throughout network
  - All routers re-compute routes

Link State Packets

A	
SeqNo	
TTL	
B	4
F	5

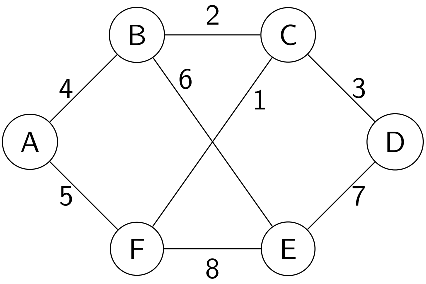
B	
SeqNo	
TTL	
A	4
C	2
E	6

C	
SeqNo	
TTL	
B	2
D	3
F	1

D	
SeqNo	
TTL	
C	3
E	7

E	
SeqNo	
TTL	
B	6
D	7
F	8

F	
SeqNo	
TTL	
A	5
C	1
E	8



Link State Distribution

- Based on flooding algorithm
  - Do not send on incoming link
- ID of source
  - Unique identifier
- Sequence number
  - Routers record received sequence numbers
  - Only newer state packets are forwarded
  - Old or duplicate packets are dropped
  - How to handle corrupt sequence numbers?
  - How to handle router reboots and crashes? (resets sequence number)
- Time-to-live (TTL)
  - Discard packets if TTL reaches 0 (decremented by routers)
  - Discard information after TTL (resets recorded sequence number)
- Different routers have different views of topology
  - E.g. link from A to B not necessarily the same cost as for link from B to A

Hierarchical Routing

- Complete map of topology in every router infeasible
- Instead exploit hierarchy and use regions
  - Router knows local topology in detail
  - Router knows route to other regions but not their internal arrangements
- Regions may map to
  - Geographical area (e.g. London academic network routes between universities)
  - Organisations network (e.g. Imperial has routers in core network, routing between departments and to external links)

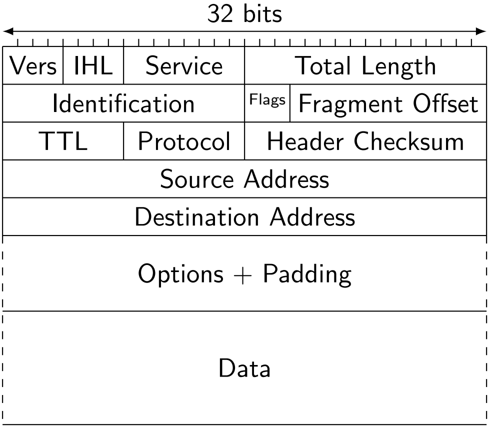
Internet Routing

- Autonomous systems (AS) are regions on the Internet
  - E.g. Internet Service Providers (ISPs) manage regions
- Within ASs
  - Routing protocols: Open Shortest Path First (OSPF) and Intermediate System-Intermediate System (IS-IS)
  - Variant of Link State routing algorithm
- Between ASs
  - Routing protocol: Border Gateway Protocol (BGP)
  - Variant of Distance Vector algorithm

Internet Protocol (IP)

- Basic protocol for the Internet (Network Layer)
  - Defined in RFC 791 (updated in 1349, 2474, 6864)
- Datagram (packet) oriented
  - Variable sized data payload
  - Unreliable delivery
  - No checksum on data payload, just on header
- Global addressing; packets contain complete addressing information
- Fragmentation
  - May split packets if underlying network requires it
- Priorities through Type of Service (ToS) information
  - Requires routers on path to read and treat packets differently

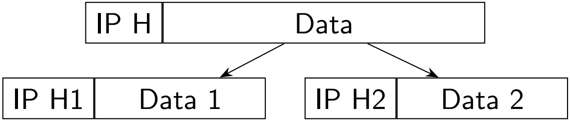
IP Datagram Format



- Version: IP version (usually 4 or 6)
- Internet Header Length
  - In 4-byte multiples (from 5 to 15)
  - Options increase this
  - Gives data offset
- Type of Service
  - Prioritize data (e.g. VoIP)
  - ToS vs DSCP (Differentiated Services Code Point)
- Total Length
  - Entire packet size in bytes (max 64KB for IPv4)

Fragmentation

- Networks have maximum transfer unit (MTU)
- E.g. Ethernet frames are limited to $\leq$ 1500 bytes
- Too large IP datagrams broken up into smaller datagrams
- Smaller datagrams need their own complete IP header



- Identifier, Flags and Fragment Offset aid reassembling fragmented datagrams
- Final destination can reassemble original datagram
  - Missing fragments are waited for
  - Whole datagram discarded if any fragment is lost

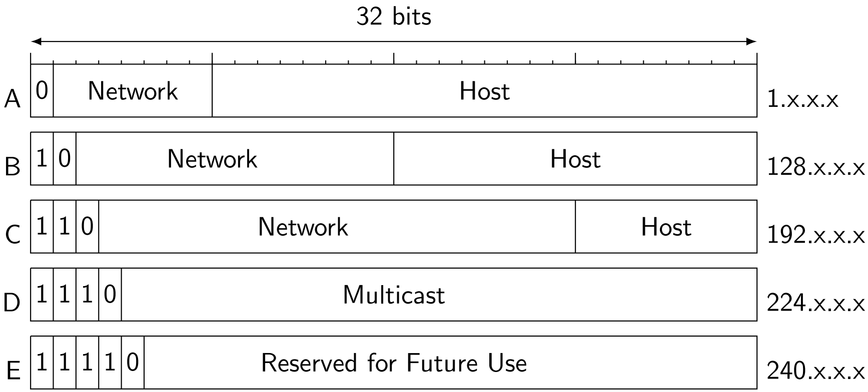
IP Datagram Format

- Time to Live (TTL)
  - Handles routing loops
  - Decremented at each hop (router)
  - Datagram dropped when 0
- Protocol
  - 0 = reserved, 1 = ICMP, 6 = TCP, 17 = UDP
  - Similar to Ethernet 'protocol type' field
- Header checksum
  - 1s complement sum of header (not data)
  - Sum of header and checksum should be 0
- Source and destination addresses
- Options
  - Security, loose/strict source routing, record route, stream ID, timestamp, ...
  - Padded to multiples of 32 bits

IP Addressing

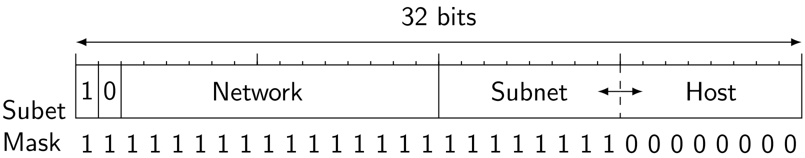
- Ethernet addresses are 48 bits and written as hex pairs
- IP addresses are 32 bits and written as dotted decimal
  - E.g. 146.169.7.41
  - No direct mapping of IP addresses to Ethernet addresses
  - IP address identifies `network` and `host` on that network
  - Device on *n* networks has *n* IP addresses - one for each
- Address space administered by ICANN (Internet Corporation for Assigned Names and Numbers)

IP Address Classes



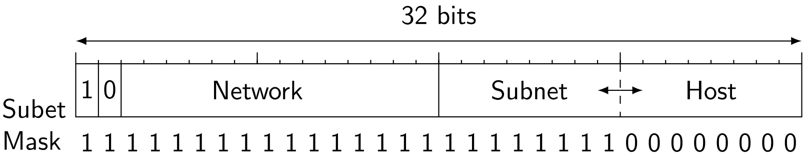
Subnets

- As organisations grow, they need finer control over network sizes
- Single class A/B/C network not good enough
- Create subnet within assigned IP network



- Subnets can be any size within host field
- Internally subnet addresses may be used for routing and creating administrative boundaries

Subnets



- Use high-order bits from host field to create subnets within network class
- Subnet mask `AND` address = network portion

Number of subnets and hosts:

- 2^{subnet_bits} = number of subnets per network
  - Usage of all 0s and 1s is not RFC-compliant
- $2^{32-network\_bits-subnet\_bits} - 2$ = number of hosts per subnet
  - All 0s and all 1s are not valid addresses; cannot be assigend to hosts

Special IP Addresses

- 127.0.0.0/8 loopback address
- Loopback is for local inter-process communication (IPC) and should never exist on the network
- 0.0.0.0 local host
- Network part + all 0s for host part: Network ID
- Network part + all 1s for host part: Broadcast address
- A broadcast addresse is never a valid source addresse
- Addresses with all bits 0 or 1 are not assigned to hosts for routing purposes

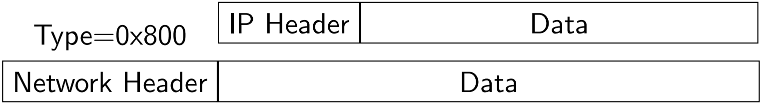
Subnet Example

- In DoC, we have a class B network
- How would you have found out?
- What about subnets?
  - Subnet mask 255.255.254.0
  - 7 bits for subnets and 9 bits for hosts
  - In a class B network this means: 128 subnets, each with 510 hosts
  - What is the network ID of this subnet?
  - What is the range of IP addresses for this subnet?
  - What is the broadcast address?
  - What is the network ID of the next subnet?

Private Internet Address Ranges

- Address ranges for internal use
- Addresses never routed on public Internet
  - Not all devices need to be globally visible
  - Used for testing and NAT (see later)
- 10.0.0.0/8
- 172.16.0.0/12
- 192.168.0.0/16

IP Datagrams and Ethernet Frames



- IP destination address is always final destination address
- Physical destination address (MAC address) in frame is changed at each hop
- Along the path each router
  - Removes packet from the frame
  - Determines next router or local link
  - Re-encapsulates in appropriate frame for next hop

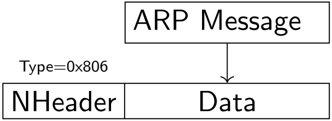
- Network layer: hosts send packets using IP addresses
- Data link layer: frames between devices use MAC addresses
- Need to translate between addresses
- Static mapping
  - May be sufficient for small isolated network
  - IP addresses are virtual (no relation to hardware, maintained in software)

- Hosts maintain lookup tables (e.g. hash tables) of IP/data link address mappings for LAN
- If host A has no entry for host B
  - A broadcasts ARP Request requesting data link address for B's IP address
  - B recognises its IP address and returns ARP response with its data link address
- ARP is network layer protocol, not visible to the user
- Usual method on IP networks that use Ethernet
- Optimisations
  - Hosts cache these tables
  - B also adds A's mapping to its own table
    - Likely to need it in future exchanges
- Defined in RFC 826

ARP Message Format

32 bits

<table>
<tr><td colspan="2">Hardware Address Type</td><td colspan="2">Protocol Address Type</td></tr>
<tr><td>HA Length</td><td>PA Length</td><td colspan="2">Operation</td></tr>
<tr><td colspan="4">Sender HA (first 4 bytes)</td></tr>
<tr><td colspan="2">Sender HA (last 2 bytes)</td><td colspan="2">Sender PA (first 2 bytes)</td></tr>
<tr><td colspan="2">Sender PA (last 2 bytes)</td><td colspan="2">Target HA (first 2 bytes)</td></tr>
<tr><td colspan="4">Target HA (last 4 bytes)</td></tr>
<tr><td colspan="4">Target PA (all 4 bytes)</td></tr>
</table>



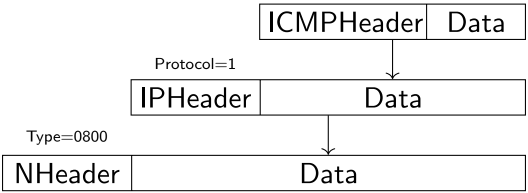
- HW Address Type: 1 for Ethernet
- Protocol Address Type: 0x800 for IP
- HW Address Length: 6 bytes
- Protocol Address Length: 4 bytes
- Operation: 1 for request, 2 for response
- Target HW Address: undefined on request
- Target machine swaps target and sender in response

Host Configuration

- How to configure the IP addresses of the hosts?
  - Manually vs dynamically
- Dynamic Host Configuration Protocol (DHCP)
  1. Host broadcasts request for IP address (DHCP DISCOVER packet)
  2. DHCP server allocates free IP address and sends it to host
     - IP addresses are chosen from pool
     - DHCP OFFER packet
     - uses Ethernet address of the DHCP DISCOVER packet
- IP addresses are only `leased` for a fixed period of time
- Host must renew the lease before it expires or IP is returned to the pool
- Other information DHCP is used for
  - IP address of the default gateway including network mask
  - IP addresses of name servers and time servers
  - Can be extended by custom fields
- Described in RFCs 2131 and 2132

Internet Control Message Protocol (ICMP)

- Used to investigate communication problems
    - IP unreliable and no guarantees of delivery
- Allows hosts to send control/error messages to other hosts
- Popular clients: Ping and Traceroute



- Behaves as if higher level protocol, but integral to IP

ICMP Message Format

- Type (8bit) and code (8bit), which gives subtype
- 1s compliment checksum (16bit) of type and code
- Rest of header based on type and code
- Type 3 codes
  - 0: Net unreachable
  - 1: Host unreachable
  - 2: Protocol unreachable
  - 3: Port unreachable
  - 4: Fragmentation needed and DF set
  - 5: Source route failed

- Other types include
  - 0: Echo reply
  - 5: Redirect
  - 8: Echo request (ping)
  - 11: Time exceeded
  - 12: Parameter problem
  - 13: Timestamp
  - 14: Timestamp reply
  - 15: Information request
  - 16: Information reply
  - 17: Address mask request
  - 18: Address mask reply

- Ping
  - Verify connection to hosts
  - Quality of connection (round trip time, dropped packages)
  - Sends echos and receives echo replies
- Traceroute
  - Find out about intermediate hosts (hops)
  - Routers send ICMP error messages back for every received packet with TTL=0 (type 11)
  - Send packets with increasing TTL for each subsequent packet

- Support for mobility (laptops, phones,. . .)
  - Moving between different networks
  - Challenges with transistion as routing depends on address used
- Expansion of networks
  - Renumbering/adding new number ranges hard

- Shortage of unallocated addresses
    - Practical address space in IPv4 is 100 million hosts
    - IP is more popular than its designers expected
- Some address classes are unnecessarily large
    - Some organisations have more than they need
    - Classes A and B are bigger than most people use

Better utilisation:

- Stricter access to allocation
  - Applicant needs to show that addresses will be utilised
- Make address allocation more flexible
  - Allocate networks with subnet masks ("Subnetting for the Internet")
  - Classless Inter-Domain Routing (CIDR)

Increase address space:

- Reuse addresses in different parts of network
  - Network Address Translation (NAT)
- Add more address bits
  - IPv6

Network Address Translation (NAT)

- Hide network in smaller address range
  - External address: routable, allocated IP address
  - Internal address: from private addresses range
  - Gateway box with external and internal address
  - Translates internal address to external address for traffic leaving internal network
  - Box maintains this mapping for incoming traffic (replies)
  - Port numbers are part of the mapping (see Transport Layer)
- Port mappings on gateways can make services on private hosts available to public hosts (see Transport Layer)

IPv6

- 128 bit addresses (vs. 32 bit in IPv4)
  - About $6.5 * 10^{23}$ addresses for every square meter of the Earth's surface
  - Large address space easier to be subdivided into hierarchical domains
- Simplified 7 field header (vs. 13 fields in IPv4)
  - Faster processing in routers possible
  - More options through extension headers
  - Support for authentication, privacy, service types, mobility, ...
- No interoperability with IPv4 but independent network
  - Can exist in parallel
  - Translator gateways to exchange traffic between IPv6 and IPv4 networks

- Transition from IPv4 to IPv6 hard and slow
  - Router/switch manufacturers not pushing it
  - ISPs and network providers not demanding it
  - Many of the benefits lost in gateways to IPv4
- Currently useful within organisation
- Not many hosts to talk IPv6 to
- Mobile phones may push adoption
- Google users using IPv6: 8-13% in 2016; 13-20% in 2017[1]

[1] https://www.google.com/intl/en/ipv6/statistics.html