CS 640: Introduction to Computer Networks

Aditya Akella

Lecture 9 - IP: Packets and Routers

The Road Ahead

• Last lecture
  - How does choice of address impact network architecture and scalability?
  - What do IP addresses look like?
  - How to get an IP address?

• This lecture
  - What do IP packets look like?
  - How to handle differences between LANs?
  - How do routers work?

IP Packets

• Low-level communication model provided by Internet
  - Unit: "Datagram"

• Datagram
  - Each packet self-contained
  - All information needed to get to destination
  - Analogous to letter or telegram

![IPv4 Packet Format]
**IPv4 Header Fields**

- **Version**: IP Version
  - 4 for IPv4
  - 6 for IPv6
- **Header Length**: 32-bit words (typically 5)
- **Type of Service**
  - Priority information
- **Packet Length**: Bytes (including header)
  - Length field limits packets to 65,535 bytes
  - Impose break internetwork packets for network performance considerations
- **Time to Live**
  - Must be decremented at each router
  - Packets with TTL=0 are thrown away
  - Ensure packets exit the network
- **Protocol**
  - Multiplexing to higher layer protocols
  - TCP = 6, ICMP = 1, UDP = 17...
- **Header checksum**
  - Ensures some degree of header integrity
  - Relatively weak - only 16 bits
- **Options**
  - E.g. Source routing, record route, etc.
  - Performance issues at routers
  - Poorly supported or not at all
- **Source Address**
  - 32-bit IP address of sender
- **Destination Address**
  - 32-bit IP address of destination

**Like the addresses on an envelope**

**Globally unique identification of sender & receiver**

- NAT?
IP Delivery Model

- **Best effort service**: Network will do its best to get packet to destination.

- **Does NOT guarantee**: 
  - Any maximum latency or even ultimate success 
  - Sender will be informed if packet doesn’t make it 
  - Packets will arrive in same order sent 
  - Just one copy of packet will arrive 

- **Implications** 
  - Scales very well → simple, dumb network: “plug-n-play” 
  - Higher level protocols must make up for shortcomings 
  - Packets delivering ordered sequence of bytes → TCP 
  - Some services not feasible 
    - Latency or bandwidth guarantee 
    - Hard real-time support

IP Fragmentation

- **Every Network has Own Maximum Transmission Unit (MTU)** 
  - Largest IP datagram it can carry within its own packet frame 
  - E.g., Ethernet is 1500 bytes 
  - Don’t know MTUs of all intermediate networks in advance 

- **IP Solution** 
  - When hit network with small MTU, fragment packets 
  - Might get further fragmentation as passed further

Reassembly

- **Where to do reassembly?** 
  - End nodes or at routers? 

- **End nodes -- better**: 
  - Avoids unnecessary work where large packets are fragmented multiple times 
  - If any fragment missing, delete entire packet 

- **Intermediate nodes -- Dangerous**: 
  - How much buffer space required at routers? 
  - What if routes in network change? 
  - Multiple paths through network 
  - All fragments only required to go through destination
Fragmentation Related Fields

- **Length**
  - Length of IP fragment

- **Identification**
  - To match up with other fragments

- **Fragment offset**
  - Where this fragment lies in entire IP datagram

- **Flags**
  - “More fragments” flag
  - “Don’t fragment” flag

---

**IP Fragmentation Example #1**

---

**IP Fragmentation Example #2**
IP Fragmentation Example #3

IP Reassembly

Fragmentation and Reassembly
Internet Control Message Protocol (ICMP)

- Short messages used to send error & other control information

- Examples
  - Echo request / response
    - Can use to check whether remote host reachable
  - Destination unreachable
    - Indicates how far packet got & why couldn't go further
  - Flow control (source quench)
    - Slow down packet delivery rate
  - Timeout
    - Packet exceeded maximum hop limit
  - Router solicitation / advertisement
    - Helps newly connected host discover local router
  - Redirect
    - Suggest alternate routing path for future messages

IP MTU Discovery with ICMP

- Operation
  - Send maximized packet with "do not fragment" flag set
    - If encounters problem, ICMP message will be returned
      - "Destination unreachable: Fragmentation needed"
  - Usually indicates MTU encountered
  - Typically sends series of packets from one host to another
    - Additional discovery cost
  - Typically, all will follow same route
    - Routes remain stable for minutes at a time
    - Makes sense to do MTU discovery
**IP MTU Discovery with ICMP**

- **MTU = 4000**
- **MTU = 1500**

- When successful, no reply at IP level
  - “No news is good news”

- Higher level protocol might have some form of acknowledgement

---

**Router Architecture Overview**

Two key router functions:
- Run routing algorithms/protocol (RIP, OSPF, BGP)
- Switching datagrams from incoming to outgoing link
Line Card: Input Port

Physical layer: bit-level reception
Data link layer: e.g., Ethernet

Decentralized switching:
- Process common case ("fast-path") packets
- Increment TTL, update checksum, forward packet
- Given datagram dest., lookup output port using routing table in input port memory
- Queue needed if datagrams arrive faster than forwarding rate into switch fabric

Line Card: Output Port

- Queuing required when datagrams arrive from fabric faster than the line transmission rate

Three Types of Switching Fabrics
Switching Via a Memory

First generation routers looked like PCs
• Packet copied by system's (single) CPU
• Speed limited by memory bandwidth (2 bus crossings per datagram)

Most modern routers switch via memory, but...
• Input port processor performs lookup, copy into memory
• Cisco Catalyst 8500

Switching Via a Bus

• Datagram from input port memory to output port memory via a shared bus
• Bus contention: switching speed limited by bus bandwidth
• 1 Gbps bus, Cisco 1900: sufficient speed for access and enterprise routers (not regional or backbone)

Switching Via an Interconnection Network

• Overcome bus and memory bandwidth limitations
• Crossbar provides full N×N interconnect
  - Expensive
  - Uses 2N buses
• Banyan networks & other interconnection nets initially developed to connect processors in multiprocessor
  - Typically less scalable than complete crossbar
• Cisco 12000: switches 6 Gbps through the interconnection network
Network Processor

- Runs routing protocol and downloads forwarding table to forwarding engines
- Performs "slow" path processing
  - ICMP error messages
  - IP option processing
  - Fragmentation
  - Packets destined to router

A Note on Buffering

- 3 types of switch buffering
  - Input buffering
    - Fabric slower than input ports combined → queuing may occur at input queues
    - Can avoid any input queuing by making switch speed = N x link speed
  - Output buffering
    - Buffering when arrival rate via switch exceeds output line speed
  - Internal buffering
    - Can have buffering inside switch fabric to deal with limitations of fabric
- What happens when these buffers fill up?
  - Packets are THROWN AWAY!! This is where (most) packet loss comes from

Input Port Queuing

- Which inputs are processed each slot – schedule?
- Head-of-the-Line (HOL) blocking: datagram at front of queue prevents others in queue from moving forward
Output Port Queuing

- Scheduling discipline chooses among queued datagrams for transmission
  - Can be simple (e.g., first-come first-serve) or more clever (e.g., weighted round robin)

Forwarding: Longest Prefix Match

- Traditional method - Patricia Tree
  - Arrange route entries into a series of bit tests
  - Worst case = 32 bit tests
  - Problem: memory speed is a bottleneck

Speeding up Prefix Match - Some Alternatives

- Route caches
  - Packet trains → group of packets belonging to same flow
  - Temporal locality
  - Many packets to same destination
  - Size of the cache is an issue

- Other algorithms
  - Routing with a Clue [Bremer-Barr - Sigcomm 99]
    - Clue = prefix length matched at previous hop
    - Why is this useful?
Next Lecture

• How do forwarding tables get built?

• Routing protocols
  - Distance vector routing
  - Link state routing