Washington University in St. Louis

Engineering

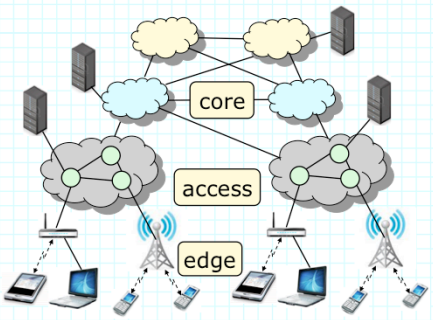
6. Internet Routers and the IPv4 Protocol

- IP Addressing and packet format
- Fragmentation and reassembly
- IP address lookup
- Router datapath components
- Interconnect options and queueing

Jon Turner

Understanding the Internet Context

- The *internet* is a "network of networks"
  - » multiple heterogenous "Layer 2" (L2) networks
  - » multiple owners administrative domains
  - » routers forward packets across networks
  - » routers make limited assumptions about capabilities of lower level networks

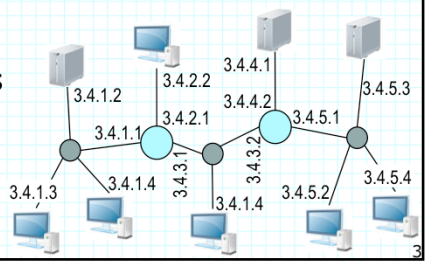


- Relationship between L2 subnets and Internet routers
  - » from L2 network perspective, an internet router is an endpoint
  - » IP addresses are assigned in groups, corresponding to networks
  - » so router with link to several network has separate address for its connection to each one

2

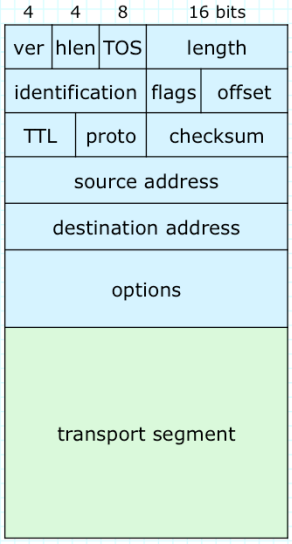
IP Addressing

- Address is a 32 bit number
  - » usually written in "dotted decimal" (e.g. 132.45.21.7) where each part represents eight bits
- Addresses are assigned hierarchically
  - » high order bits specify a network, low order bits specify a host
    - for example, 132.45.0.0/16 represents block of addresses starting with 132.45
    - boundary is flexible, so 132.45.96.0/20 represents block of addresses that start with first 20 bits of 132.45.96.0
    - can divide address blocks into sub-blocks for sub-networks
- Associated with link endpoints
  - » so, routers and hosts may have multiple addresses
- *Broadcast address*: all 1s
- *Multicast*: [224.0.0.0,240.0.0.0)



3

IPv4 Packet Format



- *Version number* specifies the version of the IP protocol and determines packet format
- *Header Length* (hLen) in 32 bit words
- *Type of Service* (TOS) field used to specify how packets are forwarded
- *Identification, flags* and *offset* used for fragmentation/reassembly
- *Time-to-live* (TTL) specifies number of seconds before packet must be discarded
  - » prevents infinite looping of packets
  - » every router decrements by at least 1
- *Protocol* identifies higher level protocol
  - » TCP (6) or UDP (17)
- *Header checksum* for error detection
- *Address* fields specify source, destination
- *Options* – up to 10 words long

Exercises

1. RFC 791 refers frequently to a "gateway". What is meant by this?
2. What is the smallest legitimate value for the header length field? What is the maximum number of octets in the options part of an IP packet? Why do you think the RFC uses the term "octet", instead of "byte"?
3. What is the largest number of times an IP packet can be forwarded?
4. Name three options supported by the IP protocol. What do they do?

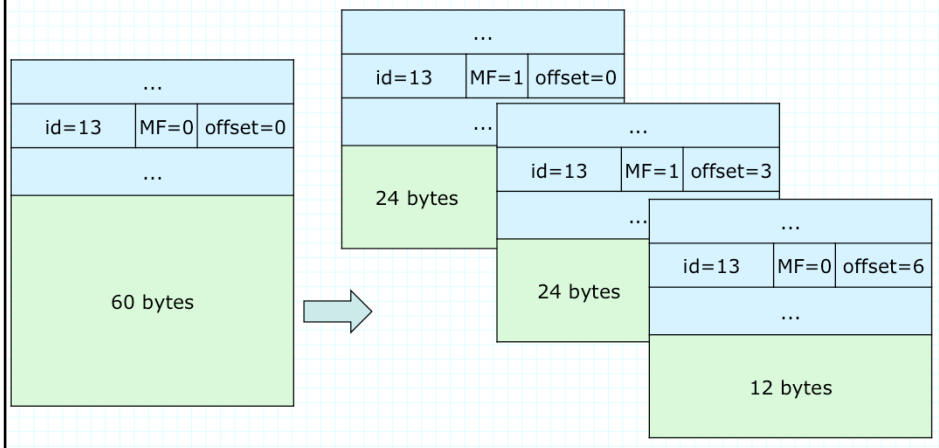
Fragmentation and Reassembly

- Link layer networks can have different max packet sizes
  - » common values are 576 and 1500 bytes
- If a packet is too long to be forwarded through a given network, it must be "fragmented"
  - » role of *identification* field in fragmentation
    - packets labeled by sending host with distinct *id* values
    - if router fragments packet, all fragments have same id field
  - » role of *offset field*
    - specifies location of "this fragment" relative to start of original packet *payload* – specified in units of eight bytes
  - » role of *flags*
    - the *More Fragments* flag is 1 for all fragments but the last
    - the *Don't Fragment* flag can be set by a source to block fragmentation (may result in packet being discarded)

Washington University in St. Louis | Engineering

Example of Fragmentation

- Original packet with 20B header and 60B payload
- Must send through network with max packet size of 50B



More on Packet Fragmentation

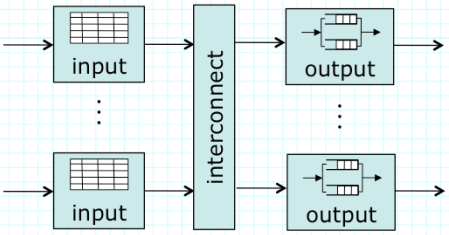
- Fragments are generally not reassembled by routers
  - » destination host must perform reassembly
  - » adds to destination host's packet processing overhead
  - » more efficient if sender uses appropriately sized packets in the first place
- Sender can "discover" the largest safe packet size by doing some experiments
  - » send packets with Don't Fragment flag set
    - packets that are too long to get through without fragmenting will be discarded and ICMP message with "correct" MTU are returned
    - hosts can detect changes by continuing to send packets with don't fragment set and periodically probing to detect increases in PMTU
  - » this procedure is called Path MTU Discovery (MTU stands for Maximum Transfer Unit)
    - IPv6 drops support for fragmentation, making PMTU discovery required

Exercises

1. What happens to a packet for which the "do not fragment" flag is set, if an intermediate router cannot deliver it without fragmenting it?
2. Suppose a packet with no IP options and a payload of 1200 bytes must be sent through a network with a maximum packet size of 150 bytes. What is the minimum number of fragments needed to represent this packet?
3. What is the role of the "copy bit" in the IP options?
4. What is the format of the *record route* option? What does it do? Should the copy bit be set for this option?
5. What is the format of the *strict source routing* option? What does this option do? How does it differ from the *loose source routing* option? In the modern internet, do you think these options make sense? Should the copy bit be set for these options?

Basic Router Datapath

- Input side processing
  - » check packet header
  - » form *lookup key*
  - » retrieve forwarding information
- Interconnect
  - » transfer packets from inputs to specified output(s)
  - » design for *scalability* and *nonblocking performance*
- Output-side processing
  - » one or more queues to hold packets waiting to go out
  - » multiple queues enable preferential treatment
    - may be based on per packet priority or per *flow* shares
- Typically include one or more general purpose processors for control (configuring forwarding tables)



10

IP Address Lookup

- Routing tables contain (*prefix*, *next hop*) pairs
- Address in packet is compared to stored prefixes, starting at left
  - » prefix that matches largest number of address bits is desired match
- Packet is forwarded to the specified next hop
  - » if next-hop includes an address, packet is sent to router at that address
  - » otherwise, next-hop is destination
- Large routers have >10^5 prefixes
  - » need data structures that support efficient lookup

forwarding table

prefix	next hop out link, adr
10*	7, 1.2.3.4
01*	5, -
110*	3, -
1011*	5, 2.3.4.5
0001*	1, 5.4.3.2
0101 1*	7, -
0001 0*	1, -
0011 00*	2, 3.4.5.6
1011 001*	3, 4.5.6.7
1011 010*	5, -
0100 110*	6, -
0100 1100*	4, 8.7.6.5
~~1011 0111*~~	8, -
1001 1000*	10, 6.5.5.5
0101 1001*	9, -

address: 1011 0010 1000

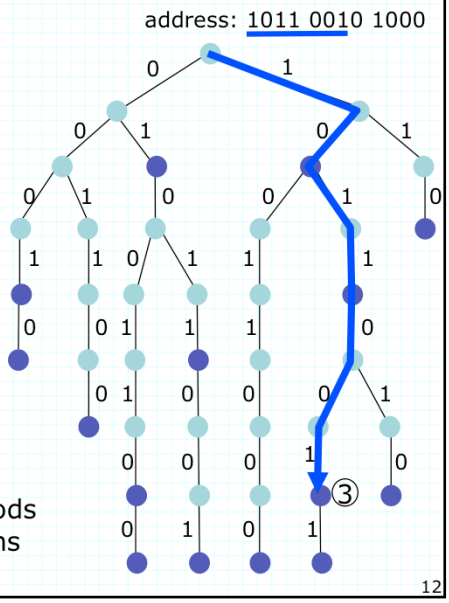
11

Washington University in St. Louis

Engineering

Address Lookup Using Tries

- Prefixes stored in trie
- Prefixes "spelled" out by following path from top
  - » dark nodes mark prefix endings
  - » to find best prefix, spell out address in tree
  - » last dark node marks longest matching prefix
- Straightforward version can be too slow for routers
  - » variations on the basic trie can speed up processing
  - » more sophisticated lookup methods used in high performance systems



12

Packet Classification

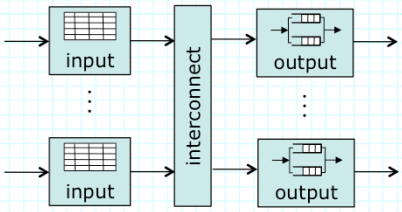
- General router mechanism with several uses
  - » security firewalls, intrusion detection systems
  - » network address translation
  - » load balancing for large web sites
  - » special handling of selected flows or groups of flows
- Common form of *packet filter* based on IP 5-tuple
  - » source address, destination address prefixes
  - » protocol field - may be specified as wildcard
  - » src port number, dest port number range (for TCP, UDP)
- No ideal solution for general case
  - » sequential search - limited to slow links, few filters
  - » ternary content-addressable memory (TCAM)
    - only effective choice in practice, for high speed links
    - drawbacks include cost, power, port range handling
  - » for special case of "exact-match filters", can use hash lookup

Exercises

1. Consider a packet with destination address 90.231.102.61 arrives at a router using the forwarding table on slide 11. On what output port is this packet forwarded? What is the IP address of the "next-hop".
2. In a router using a binary trie in its forwarding table, how many memory accesses are needed to lookup a destination address? If each memory access takes 30 ns, what is the maximum number of lookup operations that can be performed in one second? For this lookup rate, how many bits per second can be forwarded, assuming all packets have the minimum possible length? How does this change if packet payloads are 40 bytes long? What if they are 200 bytes long?
3. What is the minimum number of bits needed to specify an address prefix in a forwarding table? How many bits are needed to specify a packet filter? How many are needed to represent an exact match filter?

Router Interconnect Options

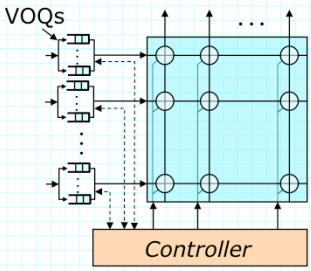
- Bus interconnects
  - » ideal performance, simple operation, but limited scalability
  - » need high bandwidth to lookup and queueing components



- Crossbars
  - » more scalable, but more complex to control
  - » need speedup, complex scheduling for best performance
- Multistage networks
  - » unbounded scalability enabling very large systems
  - » additional design issues
    - load balancing, traffic regulation, maintaining packet order
  - » wide range of specific approaches

15

Crossbar with Virtual Output Queues



- Separate *virtual output queues* at each input
  - »queues implemented using linked lists in common memory
- Controller matches inputs to outputs
  - »requires synchronous operation
    - requires that packets be divided into fixed-length "cells"
  - »objective – keep outputs busy and/or emulate ideal switch
  - »best scheduling methods work well for arbitrary input traffic
    - requires a 2x *speedup*, relative to external links
- Does not extend readily to multicast
  - »cannot associate packet with single VOQ
  - »can copy cells to multiple VOQs, but need extra speedup

Queueing

- Where does queueing occur?
  - » in systems with nonblocking bus interconnects, all queueing happens at output
  - » in crossbars, queueing occurs at both input and outputs
    - in well-designed systems, most queueing happens at the outputs
- FIFO output queuing with tail discard
  - » simplest and most common approach
  - » allows greedy "flows" to hog unfair share of link capacity
- Random mapping of packets to multiple output queues
  - » packets mapped to queues using hash on (src, dst) address
    - so all packets in same flow use same queue and stay in order
  - » queues get equal share of outgoing link capacity
  - » in system with *N* backlogged queues, a greedy flow can get only $1/N^{th}$ of link capacity

Buffer Capacity in Routers

- For random packet arrivals and average link loads of <95%, modest buffer capacities can suffice
- Unfortunately, internet traffic can cause routine overloads on links
  - » primarily cause is TCP's congestion control mechanism
  - » consequently, to ensure efficient operation a link of rate R needs buffer capable of holding at least $R \times RTT$ bits
    - for 10 Gb/s link and RTT=250 ms, that's about 300 MB
  - » smaller buffers can work well under typical conditions but may perform poorly in the worst-case
    - for links carrying traffic from N TCP flows, buffer sizes of $R \times RTT/N^{1/2}$ can work well
    - competitive forces make it hard for router vendors to deviate much from buffer sizes designed for worst-case

Exercises

For the following questions, assume an Internet router with 16 ports and a link speed of 1 Gb/s per port.

1. How many packets can arrive for output 3 in 1 μs? If the router's interconnection network is a bus, what is the maximum increase in the queue length at output port 3 during a 1 millisecond interval? How does this answer change if the interconnection network is a crossbar with a 2x speedup?
2. Suppose a router is forwarding streams of packets to 50 different hosts concurrently, where the arrival rate for each stream is 10 Mb/s. In addition, there is one stream with a rate of 1 Gb/s. If all streams are going to the same router output, and the router has a single queue at that output, at what bandwidth will the low rate streams be forwarded? What about the high rate stream? How does this change if the router distributes the packets evenly across 10 outgoing queues (still, at a single output)?