Lecture 11
IP addressing – Router Internals

Peter Steenkiste
Departments of Computer Science and
Electrical and Computer Engineering
Carnegie Mellon University

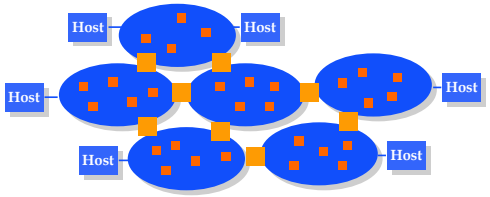
15-441 Networking, Spring 2006
http://www.cs.cmu.edu/~prs/15-441

Outline

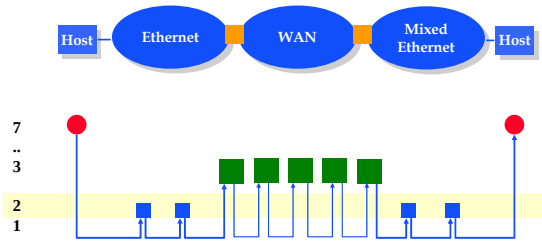
- IP addressing
- CIDR
- ICMP
- Router architecture

Internetworking

- Multiple networks connected by routers.
- Networks share some features
  - IP protocol, addressing, ..
- But differ in many other ways
  - Technology, ownerships, usage policies, scale, ..



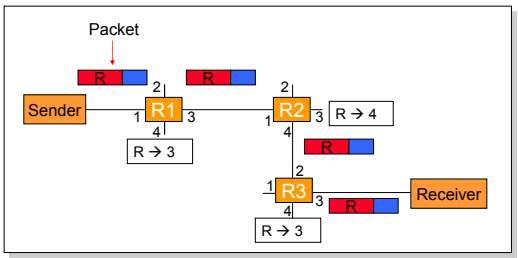
Hop-by-Hop Packet Forwarding in the Internet



IP Packet Forwarding

- Each packet has an IP destination address
- Each router has forwarding table of destination → next hop
  - Similar to Ethernet bridges and switches
  - What is different???
- Forwarding table is created by a routing protocol
  - Manual solution would be error-prone
  - How is this done for Ethernet??

Global Address Example



Router Table Size

- Simple solution: one entry for every host
  - Several 100M entries and growing
  - Sometimes called “flat” addressing
- Solution: use hierarchical addressing
  - Similar to postal addresses, telephone numbers, ...

- One entry for every organization
  - Every host in organization shares prefix
  - Requires careful address allocation
- But how do split up the address bit?

Traditional Address Classes in IP v4

- Each host has an internet address.
  - Example: 128.2.209.19
- Addresses are hierarchical.
  - address contains hint about location
- Original design: 4 classes of networks.

	type	network	host
A	0	7	24
B	10	14	16
C	110	21	8
D	1110	28	

- IP address space was assigned by the Internet Assigned Numbers Authority (IANA).
  - See http://www.iana.org

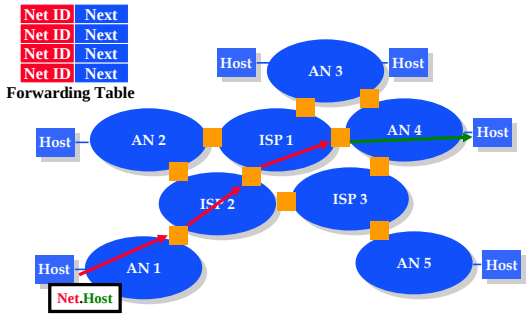
Motivation for Address Structure

- Hierarchical structure gives a hint about the location that can be used to simplify routing.
  - Compare with postal addresses
  - Routing in core based on network identifier only
    - reduces the size of the routing tables
  - But what happens when your machine moves?
- Can assign subnet address spaces that are appropriate for the size of the organizations.
  - Spaces can be managed independently
  - A small number of large spaces; many smaller spaces
  - But what happens when your organization grows?
- Class D corresponds to multicast.
  - Flat address structure

Original IP Route Lookup

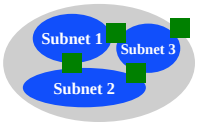
- Address would specify prefix for forwarding table
  - Simple lookup
- www.cmu.edu address 128.2.11.43
  - Class B address – class + network is 128.2
  - Lookup 128.2 in forwarding table
  - Prefix – part of address that really matters for routing
- Forwarding table contains
  - List of class+network entries
  - A few fixed prefix lengths (8/16/24)
- Did we really solve the table size problem?
  - 2 Million class C networks!

Routing based on Network Identifier



Subnetting

- Hierarchy can be extended to more than two layers.
- Makes it possible to break up a network in multiple subnets.
  - provides flexibility to manage networks
  - packet forwarding between subnets is also done using routers, i.e. same as in Internet
- Provides autonomy.
  - subnets inside network are not visible outside the network
- Example: CMU has a number of independently managed subnets.
  - E.g. computer science



Subnetting Example

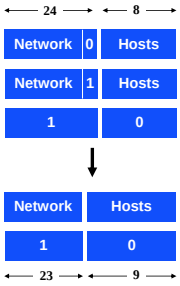
- Assume an organization was assigned address 150.100
- Assume < 100 hosts per subnet
- How many host bits do we need?
  - Seven
- What is the network mask?
  - 11111111 11111111 11111111 10000000
  - 255.255.255.128

IP Addressing: Problems

- IANA was running out of class B addresses.
  - Class-based addressing does not use space efficiently
  - A networks too large; C networks too small
- Routing tables in the network core grow too fast as a result of the success of the Internet.
  - Too many networks!
- Running out of IP addresses: too many hosts.
  - 32-bit address is not sufficient on the long run
- Combination of solutions:
  - Classless Inter-Domain Routing (CIDR)
  - Dynamic address assignment
  - Network Address translation (NAT)
  - IPv6 with 128 bit address

CIDR Addressing: Variable Length Network ID

- Length of network address is variable and specified using a netmask.
  - Generalization of the idea of subnetting
- Can make the address space just large enough
  - Can merge a group of adjacent class C addresses to form a larger network address
  - Can break up a class B address into multiple network addresses
- Must now have the netmask to identify the network id.
  - Address class bits no longer useful

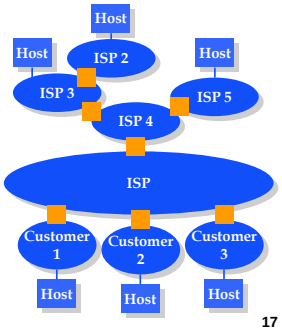


CIDR Example

- ISP is allocated 8 class C chunks, 200.10.0.0 to 200.10.7.255
  - Allocation uses 3 bits of class C space
  - Remaining 21 bits are network number, written as 200.10.0.0/21
- Replaces 8 class C routing entries with 1 combined entry
  - Routing protocols carry prefix with destination network address
  - Longest prefix match for forwarding
- ISPs get block of addresses from Regional Internet Registries (RIRs)
  - ARIN (North America, Southern Africa), APNIC (Asia-Pacific), RIPE (Europe, Northern Africa), LACNIC (South America)

CIDR Addressing: ISP-based Allocation

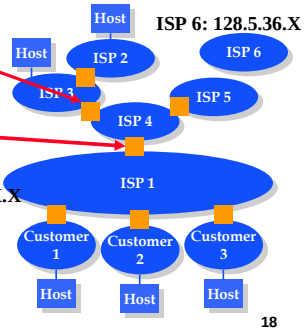
- Service provider hands out network addresses to its customers from a large block assigned to it.
  - To senders, the customers of the ISP look like subnets in the ISP
- Routing entries for the ISPs customers can be merged in many cases.
  - Packets should be forwarded to that ISP, which will then forward them to the customer
  - Only the destination ISP has to distinguish between its customers



CIDR Address Allocation: Example

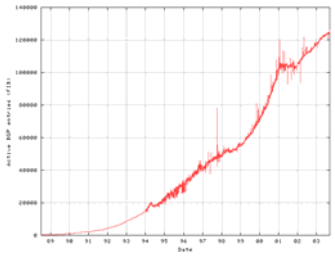
Single route entry:
128.5/16

Separate route entries for each customer:
128.5.010/19
128.5.110/19
128.5.011/19



Customer 1: 128.5.010xxxxx.X
Customer 2: 128.5.110xxxxx.X
Customer 3: 128.5.011xxxxx.X

Size of Complete Routing Table



Source: www.cidr-report.org

Shows that CIDR has kept # table entries in check

- Currently require 124,894 entries for a complete table
- Only required by backbone routers

Shortcomings of CIDR

- CIDR does not help with the large number of addresses that were already assigned before CIDR was introduced.
  - » E.g. 128.2 for CMU
- Many exceptions to CIDR addresses.
  - » Customer receives a block of addresses and then moves to a different ISP
    - – Typically keeps the same addresses
  - » Many customers subscribe with several ISPs for redundancy
    - – Example: 45 Mbs with a primary ISP, and 5 Mbs with two backup ISPs
    - – Can only have one set of addresses
  - » These exceptions require adding that network as a separate route entry in the forwarding tables

Route Lookup with CIDR

- Problem: with CIDR there can be multiple matches when looking up an address.
  - » Can for example happen when a customer switches ISPs but keeps addresses
- Solution: lookup is based on longest prefix match.
  - » If there are multiple matches in the lookup, the match with the most bits (longest netmask) wins
  - » Complicates route lookup!

Ex-ISP 10110110 -> ISP 1

My Entry 10110110 010 -> ISP 2

10110110 010 0100011

Host Routing Table Example

Destination	Gateway	Genmask	Iface
128.2.209.100	0.0.0.0	255.255.255.255	eth0
128.2.0.0	0.0.0.0	255.255.0.0	eth0
127.0.0.0	0.0.0.0	255.0.0.0	lo
0.0.0.0	128.2.254.36	0.0.0.0	eth0

- Host 128.2.209.100 when plugged into CS ethernet
- Dest 128.2.209.100 → routing to same machine
- Dest 128.2.0.0 → other hosts on same ethernet
- Dest 127.0.0.0 → special loopback address
- Dest 0.0.0.0 → default route to rest of Internet
  - » Main CS router: gigrouter.net.cs.cmu.edu (128.2.254.36)

Short-term Solutions to Address Space Crunch

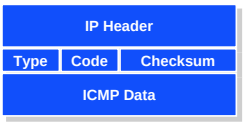
- Dynamic assignment of IP addresses.
  - » Assign an IP address for duration that it is needed only
  - » Example: dial up connections
  - » Is supported by the Dynamic Host Configuration Protocol (DHCP)
- Use of private address spaces.
  - » Many large organizations have hosts that do not need direct access to the network
    - – Only talk to other host in the organization, or
    - – Can use application level gateways to talk to hosts outside of the organization
  - » IANA set aside blocks of address space for use inside such organizations
  - » Uses Network Address Translation (NAT) boxes (later)

Outline

- IP addressing
- CIDR
- ICMP
- Router architecture

ICMP

- Internet Control Message Protocol.
- Used to report errors to sender or for diagnotics.
  - » redirect
  - » expired packets
  - » fragmentation problems
  - » tracing tools, ..
- Type-code identify the purpose of the IP packet.
  - » Related operations can be group under the same type, e.g. type 3 for unknown header fields



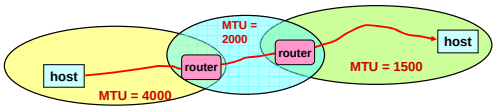
Common ICMP Types

- Echo request and reply (8-0, 0-0).
  - » Check whether a host is alive
- Entity listed in header is unreachable or unknown (type 3).
  - » Host, network, protocol, ...
- Source quench (4-0).
  - » Originally intended for congestion control
- Routing related types.
- TTL expired (11-0).
- Bad IP header (12-0).

Some Tools

- Ping: check whether a host exists and measure the roundtrip time.
  - » Uses an ICMP echo request and reply
  - » Floodping: send as fast as you can - not very friendly
- Traceroute: find the path to a host.
  - » Implemented by sending a sequence of packets with increasing TTL, I.e. 1, 2, ..
  - » Routers on the path will send ICMP error messages when the TTL reaches 0
  - » Destination host sends ICMP error port unknown
- Example: ping and tracert on Windows.

IP MTU Discovery with ICMP



- Typically send series of packets from one host to another
- Typically, all will follow same route
  - » Routes remain stable for minutes at a time
- Makes sense to determine path MTU before sending real packets
- Operation
  - » Send max-sized packet with "do not fragment" flag set
  - » If encounters problem, ICMP message will be returned
    - – "Destination unreachable: Fragmentation needed"
    - – Usually indicates MTU encountered

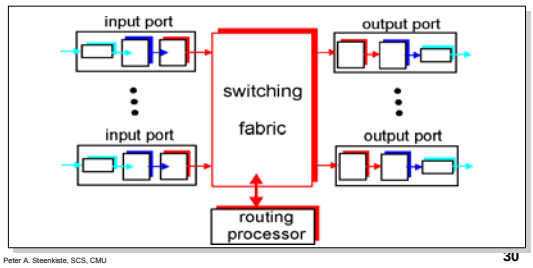
Outline

- IP addressing
- CIDR
- ICMP
- Router architecture

Router Architecture Overview

Two key router functions:

- Run routing algorithms/protocol (RIP, OSPF, BGP)
- *Switching* datagrams from incoming to outgoing link



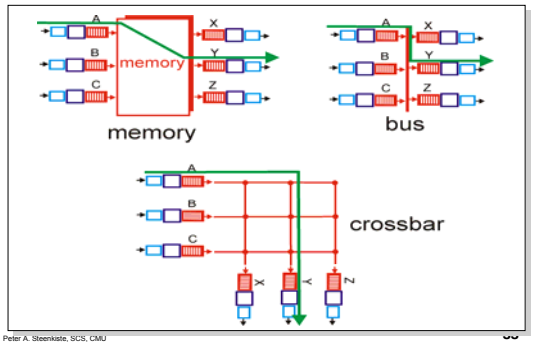
What Does a Router Look Like?

- Line cards
  - Network interface cards
- Forwarding engine
  - Fast path routing (hardware vs. software)
  - Usually on line card
- Backplane
  - Switch or bus interconnect
- Control processor
  - Handles routing protocols, error conditions, management support, …

Control 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
  - List is somewhat router specific
- Also runs management functions
  - Monitoring
  - Router configuration

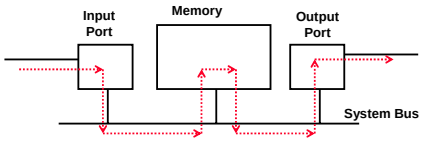
Three Types of Switching Fabrics



Switching Via Memory

First generation routers:

- Packet copied by system's (single) CPU
- Speed limited by memory bandwidth (2 bus crossings per datagram)

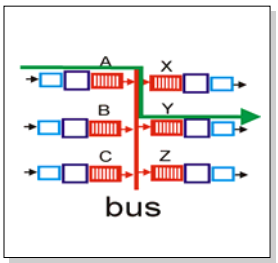


Modern routers:

- Input port processor performs lookup, copy into memory
- Cisco Catalyst 8500

Switching Via 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)



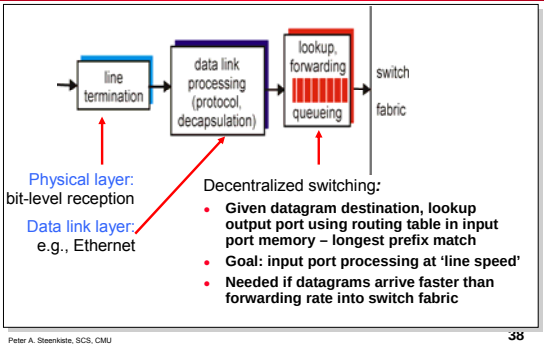
Switching Via An Interconnection Network

- Overcome bus bandwidth limitations
- Crossbar provides full NxN interconnect
  - Expensive
- Banyan networks, other interconnection nets initially developed to connect processors in multiprocessor
  - Typically less capable than complete crossbar
- Cisco 12000: switches Gbps through the interconnection network

Switch Design Issues

- Suppose we have N inputs and M outputs
  - » Multiple packets for same output – output contention
  - » Switch contention – switching fabric cannot support arbitrary set of transfers
    - – I.e, not a full crossbar
- Solution – buffer packets when/where needed
- What happens when these buffers fill up?
  - » Packets are THROWN AWAY!! This is where packet loss comes from

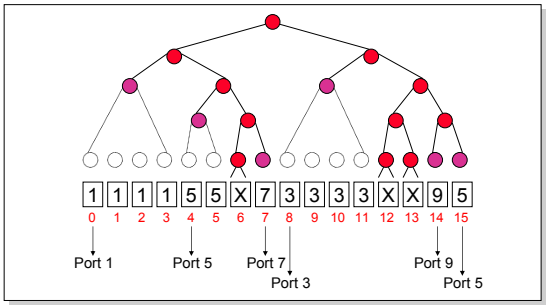
Input Port Functions



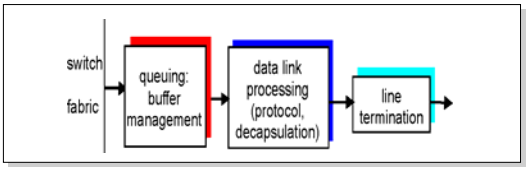
Decentralized switching:

- Given datagram destination, lookup output port using routing table in input port memory – longest prefix match
- Goal: input port processing at 'line speed'
- Needed if datagrams arrive faster than forwarding rate into switch fabric

Prefix Tree



Output Ports



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

Important Dates

- Wednesday March 1: Project 2 Checkpoint
  - » You should be actively working on Project 2
- Monday March 6: Midterm
  - » Use homeworks to help prepare
- Homework 1 due on Friday.
- Homework 2 will be handed out on Friday – due one week later.
  - » No late days allowed so we can post solutions