10:
Inter and intra AS, RIP, OSPF, BGP, Router Architecture

Last Modified:
3/24/2003 2:39:16 PM

Goals of Routing Protocols

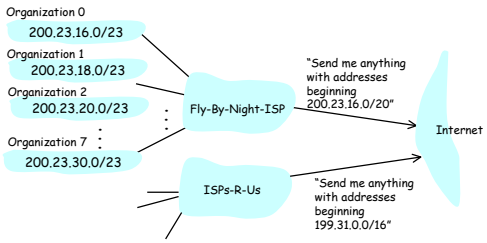
- Find the "optimal route"
- Rapid Convergence
- Robustness
- Configurable to respond to changes in many variables (changes in bandwidth, delay, queue size, policy, etc.)
- Ease of configuration

Real Internet Routing?

- CIDR?
- Dynamic routing protocols running between every router?

Recall CIDR

We already talked about how routing based on hierarchical allocation of IP address space can allows efficient advertisement of routing information:



CIDR? Dynamic Routing?

- CIDR by itself is a nice idea but..
  - Hard to maintain
  - Work around existing IP address space allocations
  - What about redundant paths?
- Dynamic routing protocols?
  - They maintain/update themselves
  - Allow for redundant paths
  - But could every router in the Internet be a node in the graph?

Dynamic Routing Protocols?

Our study of dynamic routing protocols thus far = idealized graph problem

- all routers identical
- network "flat"

… *not* true in practice

scale: with 50 million destinations:

- can't store all destinations in routing tables!
- routing table exchange would swamp links!
- Neither link state nor distance vector could handle the whole Internet!

Routing in the Internet

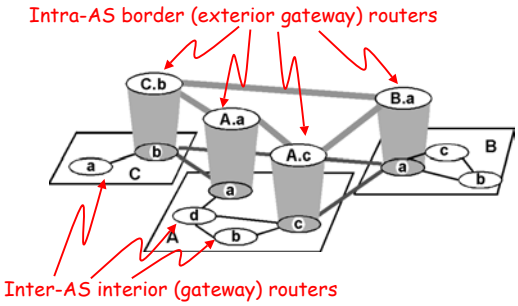
- Administrative Autonomy
  - Internet = network of networks
  - Each network controls routing in its own network
  - Global routing system to route between Autonomous Systems (AS)
- Two-level routing:
  - **Intra-AS**: administrator is responsible for choice
  - **Inter-AS**: unique standard

Hierarchical Routing

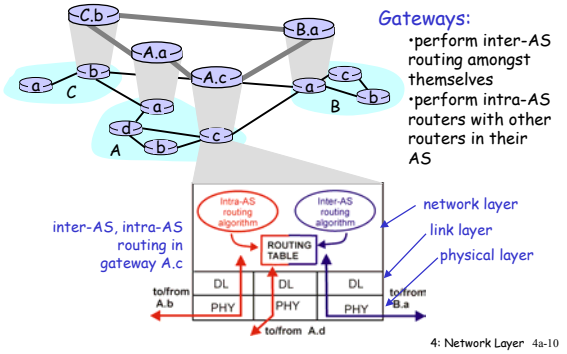
Routers in same AS run routing protocol chosen by administrators of that domain
- "intra-AS" routing protocol
- routers in different AS can run different intra-AS routing protocol

gateway routers
- special routers in AS
- run intra-AS routing protocol with all other routers in AS
- *also* responsible for routing to destinations outside AS
  - run *inter-AS routing* protocol with other gateway routers

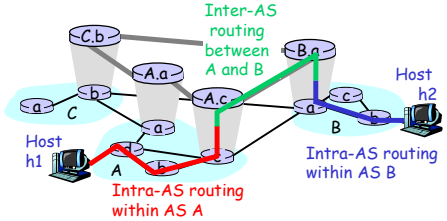
Internet AS Hierarchy



Intra-AS and Inter-AS routing



Intra-AS and Inter-AS routing



- Single datagram is often routed over many hops via routes established by several intra-AS routing protocols and an inter-AS routing protocol

Intra vs Inter AS Routing protcols

- For Intra AS routing protocols: many choices; For Inter AS routing protocols: standard
  - Why does this make sense?
- Intra AS routing protocols focus on performance optimization; Inter AS routing protocols focus on administrative issues
  - Why does this make sense?
- Choice in Intra-AS
  - Intra-AS often static routing based on CIDR, can also be dynamic (usually RIP or OSPF)
- Standard Inter-AS BGP is dynamic

Intra-AS Routing

- Also known as Interior Gateway Protocols (IGP)
- Most common IGPs:
  - RIP: Routing Information Protocol
  - OSPF: Open Shortest Path First
  - IGRP: Interior Gateway Routing Protocol (Cisco proprietary)
  - Can also be static (via CIDR) but that is not called an IGP

RIP (Routing Information Protocol)

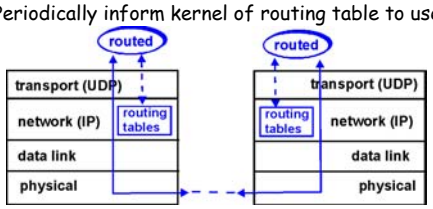
- Distance vector algorithm
- Included in BSD-UNIX Distribution in 1982
- Single Distance metric: # of hops (max = 15 hops)
  - *Can you guess why?*
  - *Count to infinity less painful if infinity = 16 ☺*
  - *But limits RIP to networks with a diameter of 15 hops*
- Distance vectors: exchanged every 30 sec via Response Message (also called **advertisement**)
- Each advertisement: route to up to 25 destination nets

RIP: Link Failure and Recovery

If no advertisement heard after 180 sec --> neighbor/link declared dead
- routes via neighbor invalidated
- new advertisements sent to neighbors
- neighbors in turn send out new advertisements (if tables changed)
- link failure info quickly propagates to entire net
- poison reverse used to prevent small loops
- infinite distance = 16 hops to make make problem with larger loops less painful

RIP Table processing

- RIP routing tables managed by **application-level** process called route-d (daemon)
- advertisements sent in UDP packets, periodically repeated
- Periodically inform kernel of routing table to use



RIP Table example: netstat -rn

Destination	Gateway	Flags	Ref	Use	Interface
127.0.0.1	127.0.0.1	UH	0	26492	lo0
192.168.2.	192.168.2.5	U	2	13	fa0
193.55.114.	193.55.114.6	U	3	58503	le0
192.168.3.	192.168.3.5	U	2	25	qaa0
224.0.0.0	193.55.114.6	U	3	0	le0
default	193.55.114.129	UG	0	143454	

- Three attached class C networks (LANs)
- Router only knows routes to attached LANs
- Default router used to "go up"
- Route multicast address: 224.0.0.0
- Loopback interface (for debugging)

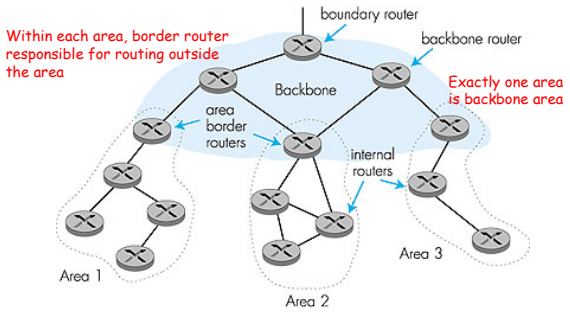
OSPF (Open Shortest Path First)

- "open": publicly available
- Uses Link State algorithm
  - LS packet dissemination
  - Topology map at each node
  - Route computation using Dijkstra's algorithm
- OSPF advertisement carries one entry per neighbor router (i.e. cost to each neighbor)
- Advertisements disseminated to entire AS (via flooding)

OSPF "advanced" features (not in RIP)

- Many have nothing to do with link-state vs distance vector!!
- Security: all OSPF messages authenticated (to prevent malicious intrusion); TCP connections used
- Multiple same-cost paths can be used at once (single path need not be chosen as in RIP)
- For each link, multiple cost metrics for different TOS (eg, high BW, high delay satellite link cost may set "low" for best effort; high for real time)
- Integrated uni- and multicast support:
  - Multicast OSPF (MOSPF) uses same topology data base as OSPF
- Hierarchical OSPF in large domains
  - Full broadcast in each sub domain only

Hierarchical OSPF: Mini Internet



Backbone area contains all area border routers and possibly others

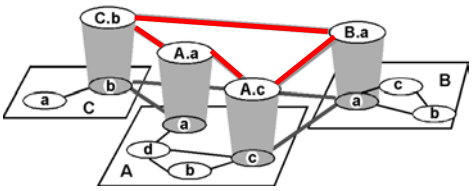
Hierarchical OSPF

- Two-level hierarchy: local area, backbone.
  - Link-state advertisements only in area
  - each nodes has detailed area topology; only know direction (shortest path) to nets in other areas.
- **Area border routers:** "summarize" distances to nets in own area, advertise to other Area Border routers.
- **Backbone routers:** run OSPF routing limited to backbone.
- **Boundary routers:** connect to other ASs.

IGRP (Interior Gateway Routing Protocol)

- CISCO proprietary; successor of RIP (mid 80s)
- Distance Vector, like RIP but with advanced features like OSPF
- several cost metrics (delay, bandwidth, reliability, load etc); administer decides which cost metrics to use
- uses TCP to exchange routing updates
- Loop-free routing via Distributed Updating Alg. (DUAL) based on *diffused computation*

Now on to Inter-AS routing



Autonomous systems

- The Global Internet consists of Autonomous Systems (AS) interconnected with each other:
  - **Stub AS**: small corporation
  - **Multihomed AS**: large corporation (no transit traffic)
  - **Transit AS**: provider (carries transit traffic)
- Major goal of Inter-AS routing protocol is to reduce transit traffic

Internet inter-AS routing: BGP

- BGP (Border Gateway Protocol): *the* de facto standard
- **Path Vector** protocol:
  - similar to Distance Vector protocol
  - Avoids count-to-infinity problem by identifying yourself in a path advertised to you
  - each Border Gateway broadcast to neighbors (peers) *entire path* (I.e, sequence of ASs) to destination
  - E.g., Gateway X may send its path to dest. Z:

Path (X,Z) = X,Y1,Y2,Y3,…,Z

Internet inter-AS routing: BGP

Suppose: gateway X send its path to peer gateway W

- W may or may not select path offered by X
  - cost, policy (don't route via competitors AS!), loop prevention reasons.
- If W selects path advertised by X, then:

Path (W,Z) = w, Path (X,Z)

- Note: X can control incoming traffic by controlling its route advertisements to peers:
  - e.g., don't want to route traffic to Z -> don't advertise any routes to Z

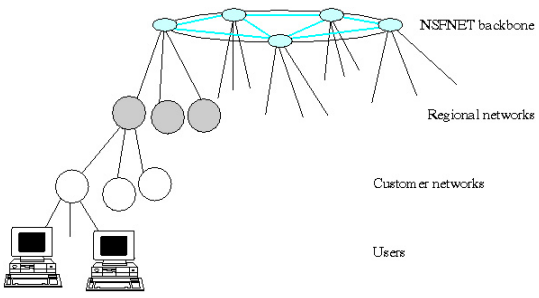
Internet inter-AS routing: BGP

- BGP messages exchanged using TCP.
- BGP messages:
  - OPEN: opens TCP connection to peer and authenticates sender
  - UPDATE: advertises new path (or withdraws old)
  - KEEPALIVE keeps connection alive in absence of UPDATES; also ACKs OPEN request
  - NOTIFICATION: reports errors in previous msg; also used to close connec tion

Internet Map

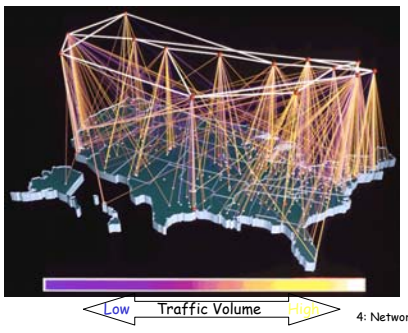
- Now that we know about autonomous systems and intra and inter AS routing protocols
- What does the Internet really look like?
  - That is a actually a hard question to answer
  - Internet Atlas Project
    - http://www.caida.org/projects/internetatlas/
    - Techniques, software, and protocols for mapping the Internet, focusing on Internet topology, performance, workload, and routing data

The Internet around 1990



CAIDA: NSFNET growth until 1995

Backbone nodes elevated



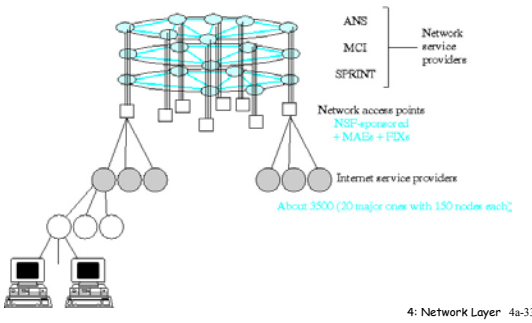
NSF Networking Architecture of Late 1990s

- NSFNET Backbone Project successfully transitioned to a new networking architecture in 1995.
  - vBNS (very high speed Backbone Network Services) - NSF funded, provided by MCI
  - 4 original Network Access Points (NSF awarded)
  - NSF funded Routing Arbiter project
  - Network Service Providers (not NSF funded)

Network Access Point

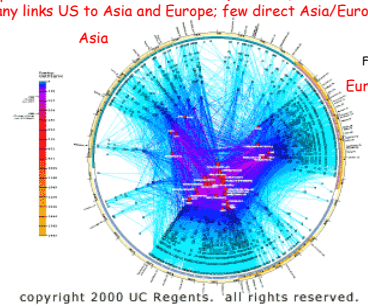
- Allows Internet Service Providers (ISPs), government, research, and educational organizations to interconnect and exchange information
- ISPs connect their networks to the NAP for the purpose of exchanging traffic with other ISPs
- Such exchange of Internet traffic is often referred to as "peering"

The Internet in 1997



CAIDA's skitter plot

Top 15 ASes are in North America (14 in US, 1 in Canada)
Many links US to Asia and Europe; few direct Asia/Europe Links



Skitter data

16 monitors probing approximately 400,000 destinations

626,773 IP addresses

1,007.723 IP links

48,302 (52%) of globally routable network prefixes

Economics of Internet Connectivity

- Upstream ISPs charge downstream ISPs for connectivity (transit traffic)
- Downstream ISPs change customers
- Upper level ISPs exchange traffic at NAPs for mutual convenience

Roadmap

- Mechanics of Routing
  - Sending datagram to destination on same network
  - Sending datagram to destination on a different network
- Router Architecture
- Router Configuration Demo

Getting a datagram from source to dest.

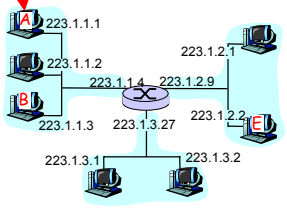
IP datagram:

misc fields	source IP addr	dest IP addr	data

- datagram remains unchanged, as it travels source to destination
- addr fields of interest here

routing table in A

Dest. Net.	next router	Nhops
223.1.1.1		1
223.1.1.2	223.1.1.4	2
223.1.1.3	223.1.1.4	2



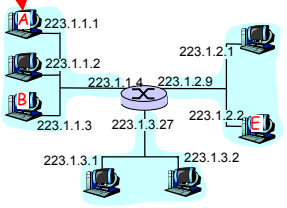
Destination on same network as source

misc fields	223.1.1.1	223.1.1.3	data

Starting at A, given IP datagram addressed to B:

- look up net. address of B
- find B is on same net. as A
- link layer will send datagram directly to B inside link-layer frame
  - B and A are directly connected

Dest. Net.	next router	Nhops
223.1.1.1		1
223.1.1.2	223.1.1.4	2
223.1.1.3	223.1.1.4	2



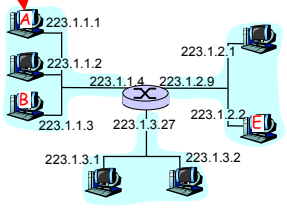
Destination on different network than source, Step 1

misc fields	223.1.1.1	223.1.2.3	data

Starting at A, dest. E:

- look up network address of E
- E on *different* network
  - A, E not directly attached
- routing table: next hop router to E is 223.1.1.4
- link layer sends datagram to router 223.1.1.4 inside link-layer frame
- datagram arrives at 223.1.1.4
- continued.....

Dest. Net.	next router	Nhops
223.1.1.1		1
223.1.1.2	223.1.1.4	2
223.1.1.3	223.1.1.4	2



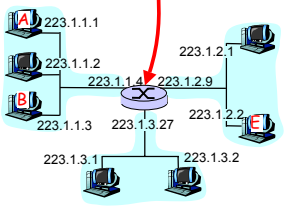
Destination on different network than source, Step 2

misc fields	223.1.1.1	223.1.2.3	data

Arriving at 223.1.4, destined for 223.1.2.2

- look up network address of E
- E on *same* network as router's interface 223.1.2.9
  - router, E directly attached
- link layer sends datagram to 223.1.2.2 inside link-layer frame via interface 223.1.2.9
- datagram arrives at 223.1.2.2!!! (hooray!)

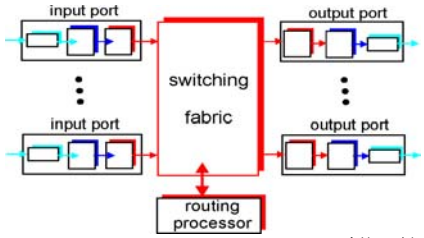
Dest. network	next router	Nhops	interface
223.1.1.1	-	1	223.1.1.4
223.1.1.2	-	1	223.1.2.9
223.1.1.3	-	1	223.1.3.27



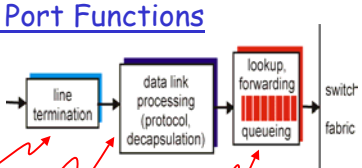
Router Architecture Overview

Two key router functions:

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



Input Port Functions



Physical layer:
bit-level reception

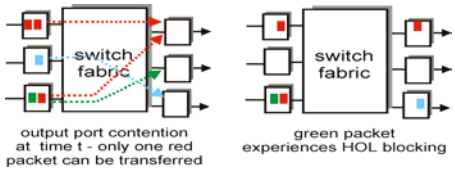
Data link layer:
e.g., Ethernet

Decentralized switching:

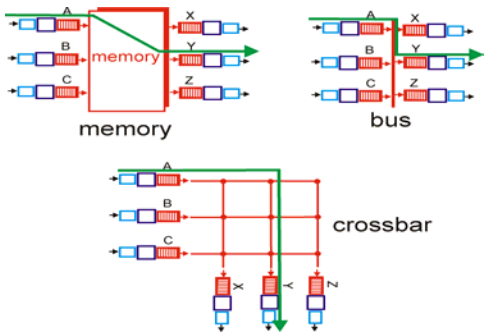
- given datagram dest., lookup output port using routing table in input port memory
- goal: complete input port processing at 'line speed'
- queuing: if datagrams arrive faster than forwarding rate into switch fabric

Input Port Queuing

- Fabric slower that input ports combined -> queueing may occur at input queues
- Head-of-the-Line (HOL) blocking: queued datagram at front of queue prevents others in queue from moving forward
- *queueing delay and loss due to input buffer overflow!*



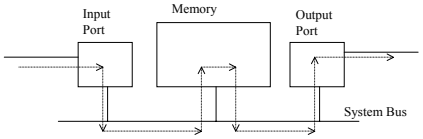
Three types of switching fabrics



Switching Via Memory

First generation routers:

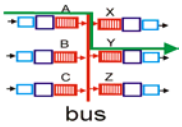
- packet copied by system's (single) CPU
- speed limited by memory bandwidth (2 bus crossings per datagram)



Modern routers:

- input port processor performs lookup, copy into memory
- Example: Cisco Catalyst 8500

Switching Via Bus

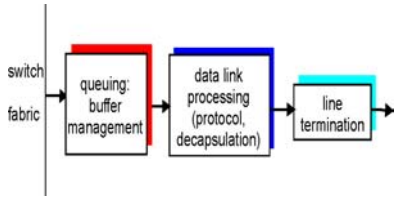


- datagram from input port memory to output port memory via a shared bus
- bus contention: switching speed limited by bus bandwidth
- 1 Gbps bus (Example: Cisco 1900): sufficient speed for access and enterprise routers (not regional or backbone)

Switching Via An Interconnection Network

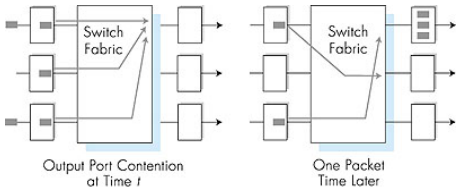
- overcome bus bandwidth limitations
- Banyan networks, other interconnection nets initially developed to connect processors in multiprocessor
  - Consider things like cross sectional BW
- Used as interconnection network in the router instead of simple crossbar
- Advanced design: fragmenting datagram into fixed length cells, switch cells through the fabric.
- Example: Cisco 12000 switches Gbps through the interconnection network

Output Ports



- *Buffering* required when datagrams arrive from fabric faster than the transmission rate
- *Scheduling discipline* chooses among queued datagrams for transmission

Output port queueing



- buffering when arrival rate via switch exceeds ouput line speed
- *queueing (delay) and loss due to output port buffer overflow!*

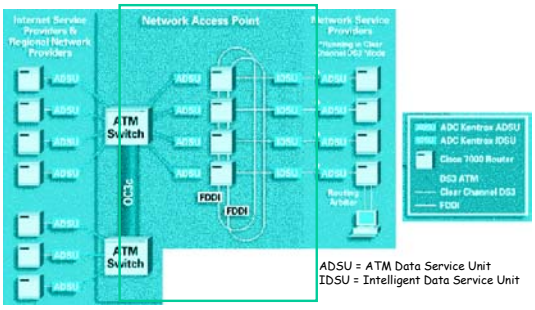
Router Hardware

Router Configuration

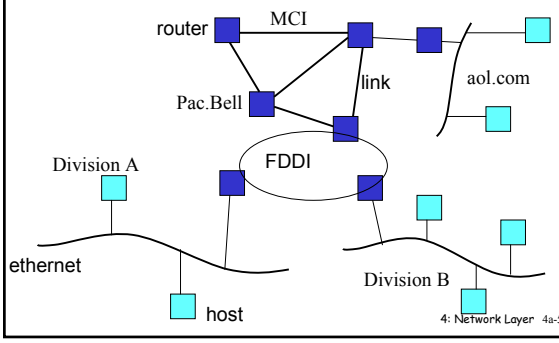
- Router Software: operating system with built in applications (command line interpreters, web servers)
- Configure Each Interface
- Configure Routing Protocol

Outtakes

A typical Network Access Point (NAP)



A small Internet



Why different Intra- and Inter-AS routing ?

Policy:

- Inter-AS: admin wants control over how its traffic routed, who routes through its net.
- Intra-AS: single admin, so no policy decisions needed

Scale:

- hierarchical routing saves table size, reduced update traffic

Performance:

- Intra-AS: can focus on performance
- Inter-AS: policy may dominate over performance

CAIDA: Layout showing Major ISPs