

IP Forwarding Table

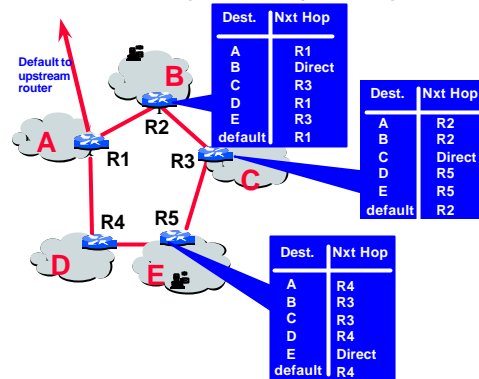
Destination	Next Hop	Interface
Net A	Router 1	INT 7
Net B	Direct	INT 4
Net C	Router 2	INT 3
Default	Router 1	INT 7

A destination is either a network, a host, or a "gateway of last resort"

The next hop is either a directly connected network or a router on a directly connected network

A physical interface

End-to-End Routes are Implemented by Next Hop Routes



How are Forwarding Tables Populated?

Statically

Administrator manually configures table entries

- + More control
- + Not restricted to destination-based forwarding
- Doesn't scale
- Slow to adapt to network failures

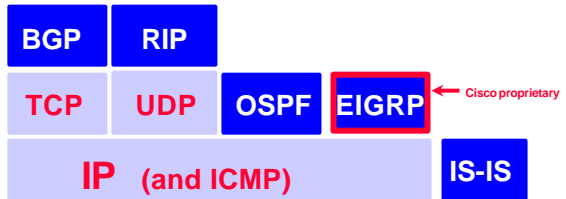
Dynamically

Routers exchange information using **ROUTING PROTOCOLS** that compute "best" routes

- + Can rapidly adapt to changes in network topology
- + Can be made to scale well
- Complex distributed algorithms
- Consume CPU, Bandwidth, Memory
- Debugging can be hell
- Current protocols are destination-based

In practice : a mix of these....

And a few, rather obscure, Routing Protocols

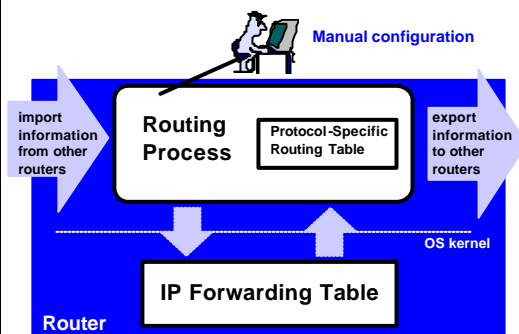


Routing protocols exchange network reachability information between routers.

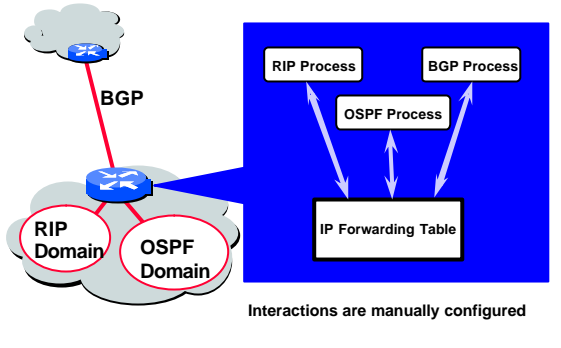
Dynamically route around network congestion? NO!

- IP traffic is very bursty
- Dynamic adjustments in routing typically operate more slowly than fluctuations in traffic load
- Attempt to adapt routing to account for load can lead to wild, unstable oscillations of routing system

What is a Routing Process?

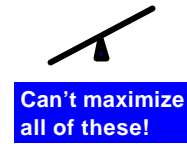


Many routing processes can run on a single router

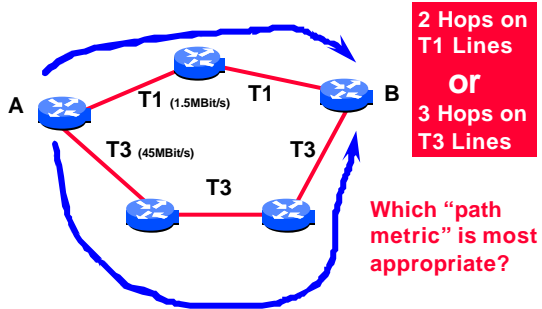


A Good Routing Protocol Should be ...

- Simple
- Stable
- Accurate
- Scalable
- Efficient (in use of CPU, memory, and link bandwidth)



Which route is the best route?



The More Information the better, NOT

The more information transmitted in a dynamic routing protocol the more precise a "best route" decision procedures can be...

But, routing information consumes bandwidth...

Routing decision procedures also consume CPU time and memory...

Potential Path Metrics

- Hop Count
- Bandwidth
- Cost
- Delay
-
- Any combination of the above:

Avoid that other IP -- Integer Programming!

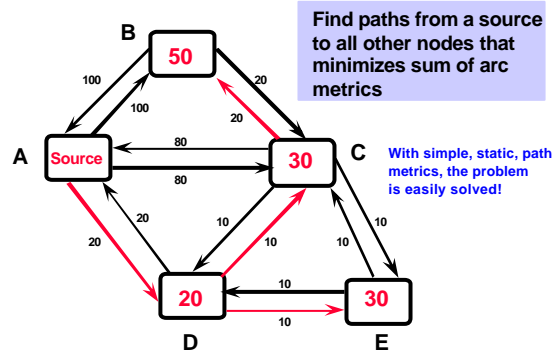
Keep it Simple!
Hop Count is most common

$$\sqrt[3]{\frac{?? \text{Hop Count} - \text{Delay}^?}{4 \text{?} \text{Cost}^?}}$$

??? ?? !!



Finding Shortest Paths



Finding Shortest Paths (cont.)

Directed Graph data structures

Arcs,
Nodes,
Arc Weights

+
Algorithm

Dijkstra's or
Bellman-Ford

=
Solution to shortest path problem

How can this computation be decentralized and performed by many cooperating routers?

Distribute computation.

Keep only local link data.

"Distance Vector" approach

RIP, EIGRP, BGP

Distribute all link data.

Perform path computations locally.

"Link State" approach

OSPF, IS-IS

Debugging can be Hell

Difficult because packets have no history, flows leave no traces. Routes are computed in a distributed manner. Loops and Black Holes can be transient, intermittent.

A few meager tools :

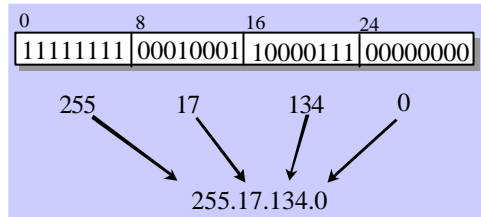
- ping / traceroute
- SNMP
- Logs
- protocol analyzers
- pre-installation lab testing

Outline

- What is an IP Routing Protocol?
- How is Addressing Implemented in IPv4?
- Routing in small to intermediate sized networks.
 - RIP
 - OSPF
- Routing In the Global Internet
 - What does the Internet Look Like?
 - BGP

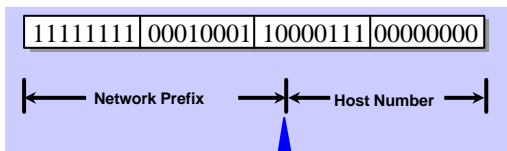
IPv4 Implementation of Addresses

32 Bit Addresses:



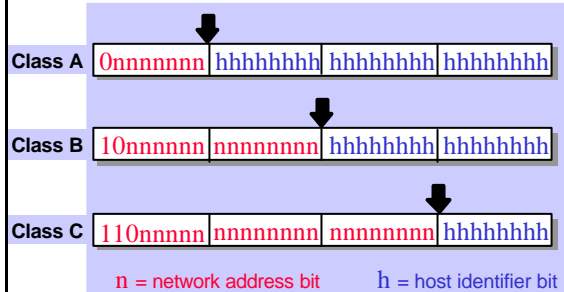
Dotted Quad notation for "human readability"

IP Addresses come in two parts



Where is this dividing line?
Well, that depends

Classful Addresses



1981, RFC 791 (definition of IPv4)

The Classful Address Space

Class	Networks	Hosts	Share of IP address space
A	127	16,777,214	1/2
B	16,384	65,534	1/4
C	2,097,152	254	1/8

Special Addresses :

- 0.0.0.0 : default route
- 127.x.y.z : Loopback addresses
- Host part all 0's : "this network"
- Host part all 1's : "broadcast to this network"

The Internet is DEAD!

Sorry, no more addresses :

If I need an address for a network with 275 hosts, then I need one Class B, and I'll waste over 65,000 addresses!
The address space is rapidly depleted!

Address assignment inflexible :

I need to run to numbering authorities every time a new network is added.

Flat classful address space does not scale:

Core route tables explode.

Long Live the Internet!

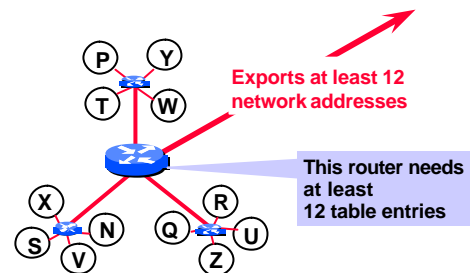
Short term solution: Addressing hierarchy

IPv4 Hierarchical Addressing = Subnetting, CIDR

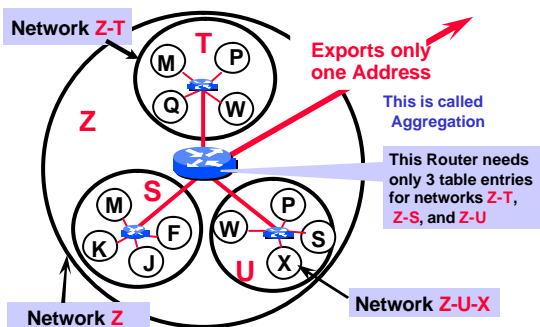
Long term solution: IPv6 with larger address space and improved hierarchical addressing

IPv6 addresses are 128 bits

Flat Network Addressing



Hierarchical Network Addressing

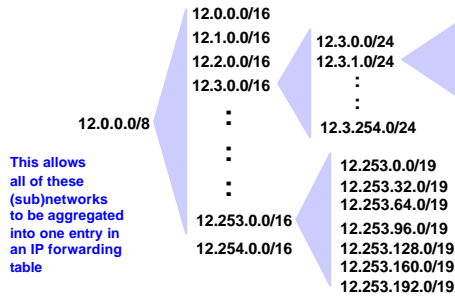


Slash Notation

Network : 12.3.0.0
Mask: 255.255.0.0 → 12.3.0.0/16

10.0.0.0/8
135.250.0.0/16
192.17.99.0/24 → Classful prefixes with "natural masks"

Variable Length Subnet Mask (VLSM)



1987, RFC 1009

Classless Inter-Domain Routing (CIDR)

Forget about Classes!
32.0.0.0/3 and 192.0.0.0/8
are legal CIDR
address prefixes

Assign Addresses
in topologically
significant manner

Now if I need an address for a network
with 275 hosts, then one "/23" address block
will provide for 512 hosts (9 bits). Great!
But I may be forced to get it from my ISP...

1993, RFCs 1517, 1518, 1519, 1520

Outline

- What is an IP Routing Protocol?
- How is Addressing Implemented in IPv4?
- Routing in small to intermediate sized networks.
 - RIP
 - OSPF
- Routing in the Global Internet
 - What does the Internet Look Like?
 - BGP

RIP

- RIP = Routing Information Protocol
- Does not scale well, designed for small LANs
- Is a "distance vector protocol"
- Very simple, easy to configure, easy to implement
- (Obsolete)

RIP Routing Table

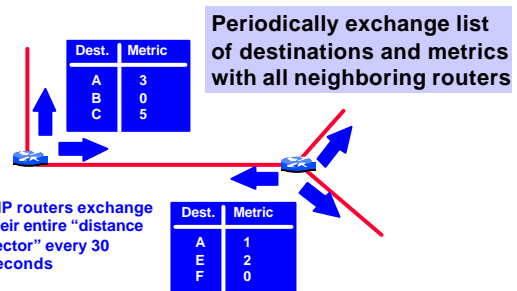
Destination	Next Hop	Metric
Net A	Router 1	3
Net B	Direct	0
Net C	Router 2	5
Default	Router 1	0

A destination is either a network, a host, or a "gateway of last resort"

The next hop is either a directly connected network or a directly connected router

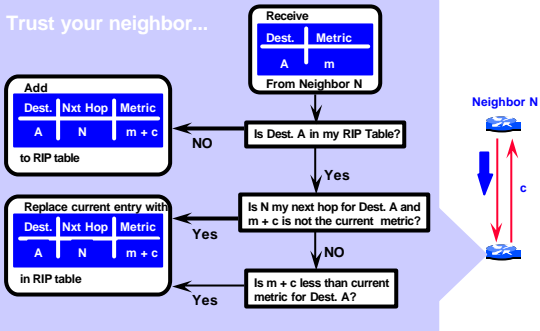
Measures how many "hops away" is the destination

Basic RIP Protocol



Basic RIP Protocol (cont.)

Trust your neighbor...



Basic RIP Protocol (cont.)

Destination	Next Hop	Metric
Network A	Router 1	infinity

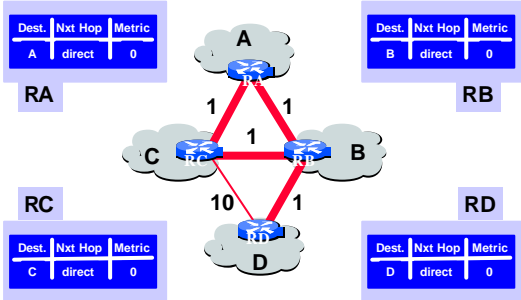
If any route is not refreshed within 180 seconds, declare it unreachable

Unreachable!

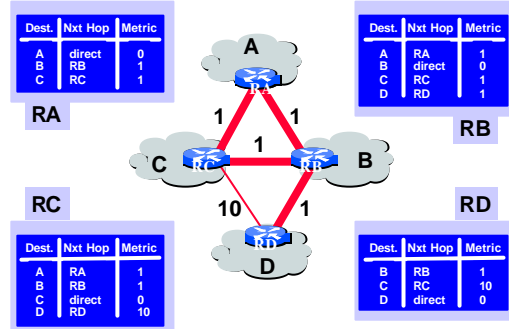
OR

If link to neighbor N goes down, declare all routes with Next Hop = N to be unreachable

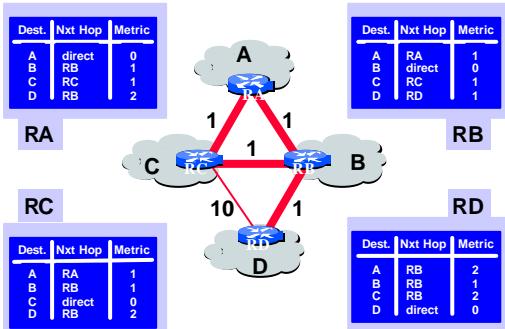
Example : Cold Start



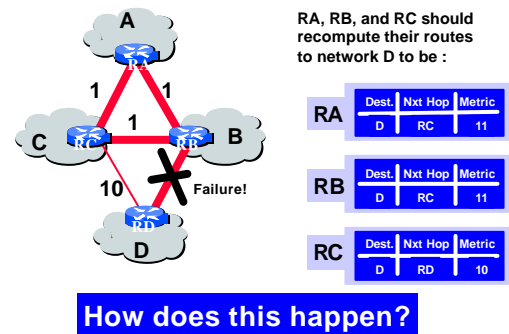
RIP Tables after one exchange



Converged RIP Tables



Topological Change



How does this happen?

Topological Change (cont.)

RA			RB			RC		
Dest.	Nxt Hop	Metric	Dest.	Nxt Hop	Metric	Dest.	Nxt Hop	Metric
D	RB	2	D	RD	1	D	RB	2
D	RC	3	D	RD	infinity	D	RB	2
D	RC	3	D	RC	3	D	RA	3
D	RC	4	D	RC	4	D	RA	4
⋮			⋮			⋮		

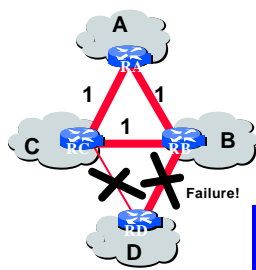
Topological Change (cont.)

RA			RB			RC		
Dest.	Nxt Hop	Metric	Dest.	Nxt Hop	Metric	Dest.	Nxt Hop	Metric
D	RC	10	D	RC	10	D	RA	10
D	RC	11	D	RC	11	D	RD	10

Note the transient routing loop between RA and RC!

Finally Converge!

Counting to Infinity



If both Links to Network D go down this will result in a "count to infinity."

RIP solution : infinity = 16

Increasing infinity allows for larger networks but also increases convergence time.

Speeding Convergence Time

- **Split Horizon**
 - **simple** : don't send neighbor N routes with next hop N
 - **with poisoned reverse** : fib to your neighbor N and say all routes with next hop N are unreachable
 - **Triggered Updates**
 - Immediately send changes upon link failure
- With these tricks, convergence is normally faster but counting to infinity and transient loops can still occur**

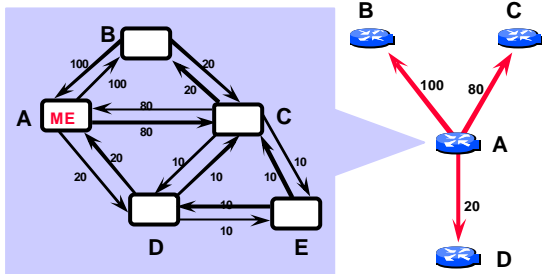
EIGRP

- EIGRP = Enhanced Internet Gateway Routing Protocol
- Cisco's proprietary distance vector protocol – works only in an all-Cisco network
- Uses complex "Distributed Update Algorithm (DUAL)" to avoid loops and counting to infinity
- Uses six element vector of metric information :
 - Delay, bandwidth, error rate, hop count, MTU, load
- Composite link metric

OSPF

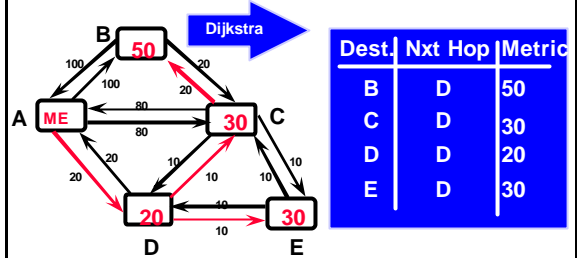
- OSPF = **O**pen **S**hortest **P**ath **F**irst
- Developed to address shortcomings of RIP
 - has rapid, loop-free convergence
 - does not count to infinity
- Link metrics between 0 and 65,535, no limit on path metric
- Is a "link state protocol"
- Has reputation for being complex
- Scales well
- Defined in RFCs 1247 (1991), 1583 (1994), 2178 (1997), 2328 (1998).

Link State Database



Each Router has a database representing the entire network that is constructed from the local knowledge at each router

Building OSPF Routing Table



Dest.	Nxt Hop	Metric
B	D	50
C	D	30
D	D	20
E	D	30

Compute locally using Link State Database!

That's Easy!

Not so fast!

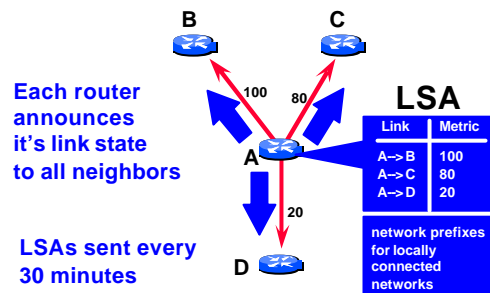
RIP RFC 1058 : 33 pages

OSPF RFC 2328 : 244 pages

Much of this complexity is related to the synchronization of the distributed, replicated link state database. Plus network modeling

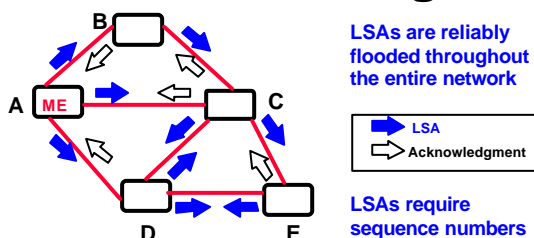
Distance Vector vs. Link State....

Link State Announcements (LSAs)

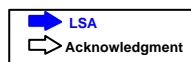


LSAs sent every 30 minutes

LSA Flooding



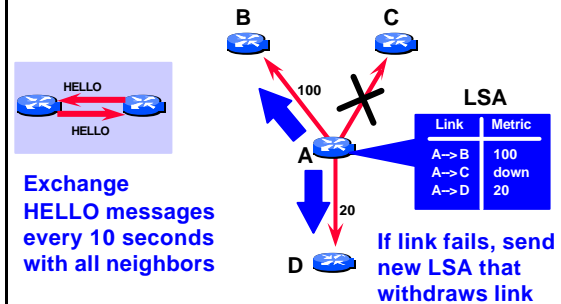
LSAs are reliably flooded throughout the entire network



LSAs require sequence numbers to distinguish old and new data

OSPF runs directly over IP (not TCP), so it must provide its own mechanism for reliable messaging

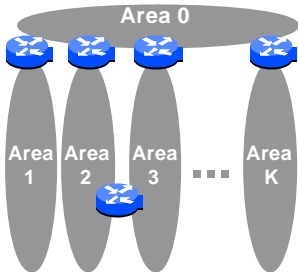
OSPF's Hello Protocol



Exchange HELLO messages every 10 seconds with all neighbors

Scalability: OSPF Areas

LS database unique within an area



- Decentralize administration
- Reduce memory usage per router
- Reduce bandwidth used by flooding

Special OSPF protocol to exchange routes between areas. This is a "distance vector" protocol!

Best IGP?

OSPF

RIP

- | | | |
|--|---|--|
| <ul style="list-style-type: none"> • Rapid, loopless convergence • Wide metric range • Supports multiple metrics • Scales well • Can support multiple best routes • Supports a wide variety of network types | + | <ul style="list-style-type: none"> • Simple to implement • Uses little memory • Works well in small, homogeneous networks |
|--|---|--|

- | | | |
|---|---|--|
| <ul style="list-style-type: none"> • Is complex • Uses a lot of memory • Requires more configuration | - | <ul style="list-style-type: none"> • Obsolete • Uses a lot of bandwidth • Slow convergence time |
|---|---|--|

Don't forget static routing, IS-IS...

Integrated IS-IS

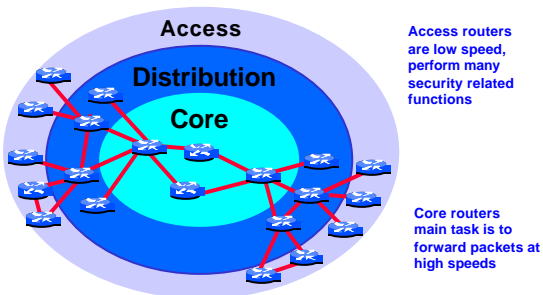
- IS-IS = Intermediate System-to-Intermediate System (ISO standard)
- Integrated IS-IS is version adapted for IP
- Is a "link state protocol" similar to OSPF
- Sometimes called DUAL IS-IS
- Designed for LAN networking
- Is popular with some large ISPs
- Defined in ISO 10589 and RFC 1195

See <http://www.ietf.org/html.charters/isis-charter.html>

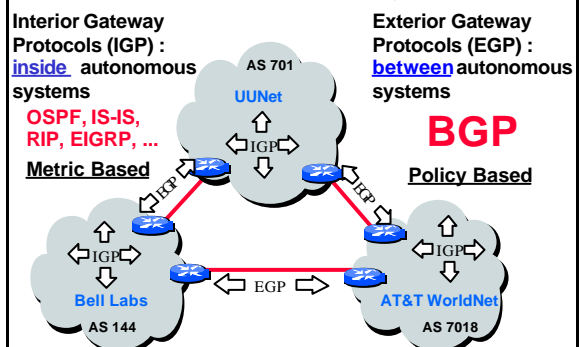
Outline

- What is an IP Routing Protocol?
- How is Addressing Implemented in IPv4?
- Routing in small to intermediate sized networks.
 - RIP
 - OSPF
- Routing In the Global Internet
 - What does the Internet Look Like?
 - BGP

Tradeoffs may depend on where you are in the Internet

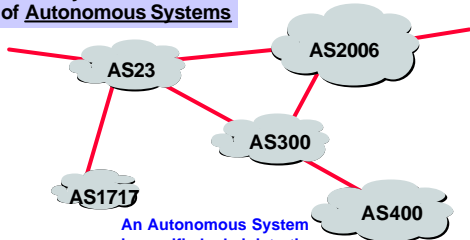


Interior vs. Exterior Routing Protocols



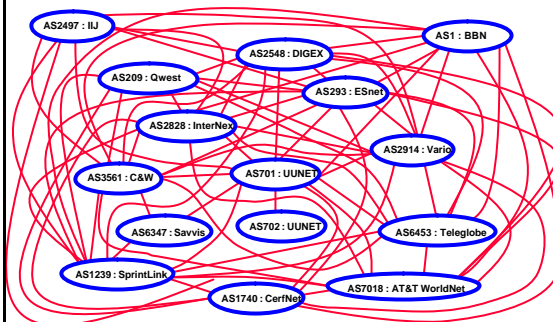
Current Internet Architecture

Arbitrary Internetwork of Autonomous Systems

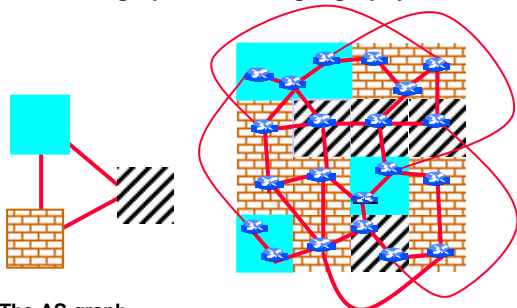


An Autonomous System is a unified administrative domain with a consistent routing policy

A small example



AS graphs obscure geography!



The AS graph may look like this.

Reality may be closer to this...

Interior vs. Exterior Routing Protocols

Interior Gateway Protocols (IGP):

inside autonomous systems

OSPF, IS-IS, RIP, EIGRP, ...

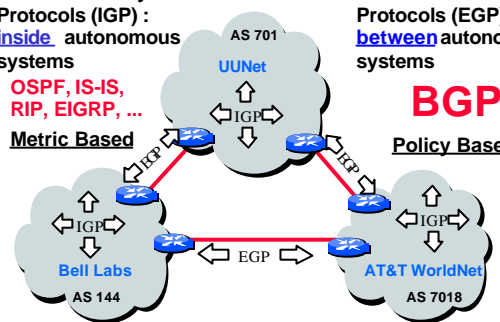
Metric Based

Exterior Gateway Protocols (EGP):

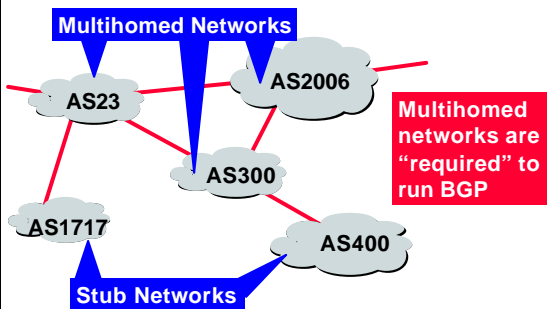
between autonomous systems

BGP

Policy Based



Stub vs. Multihomed Networks



Why are EGPs Needed?

Scalability

Even with highly aggregated addresses, there are currently about 120,000 routes in the core route tables.

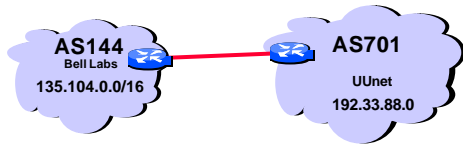
Routing Policies

Connectivity is based on bilateral commercial agreements, constrained by legal, political, and monetary considerations.

- Transit policies
- Address aggregation policies
- Traffic management policies

Policy : Some Terminology

AS144 originates the route 135.104.0.0/16

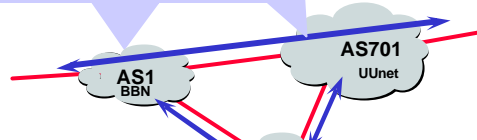


AS144 exports 135.104.0.0/16 to AS701 for inbound traffic

AS144 imports 192.33.88.0/24 from AS701 for outbound traffic

Policy : Transit vs. Nontransit

A transit AS allows traffic with neither source nor destination within AS to flow across the network

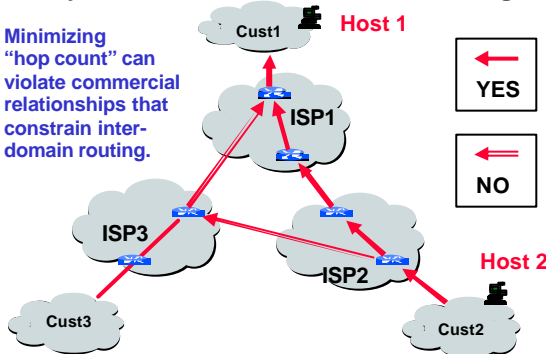


A nontransit AS allows only traffic originating from AS or traffic with destination within AS

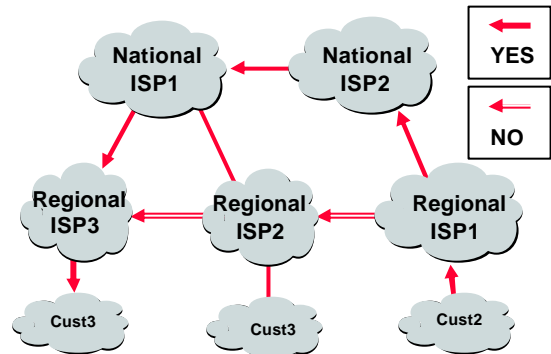
IP traffic

Policy-Based vs. Distance-Based Routing?

Minimizing "hop count" can violate commercial relationships that constrain inter-domain routing.

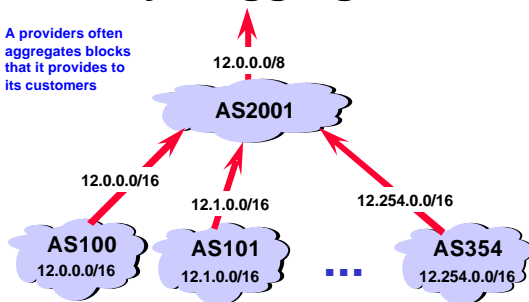


Why not minimize "AS hop count"?



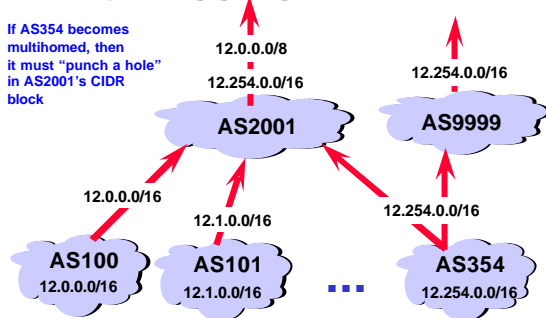
Policy : Aggregation

A providers often aggregates blocks that it provides to its customers

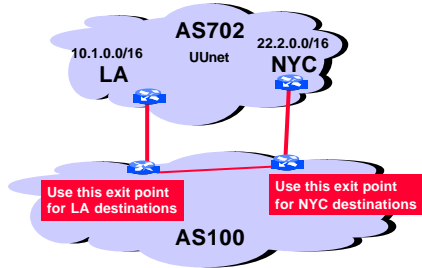


Policy : Aggregation (cont.)

If AS354 becomes multihomed, then it must "punch a hole" in AS2001's CIDR block



Policy : Traffic Management



The No Meltdown "Policy"

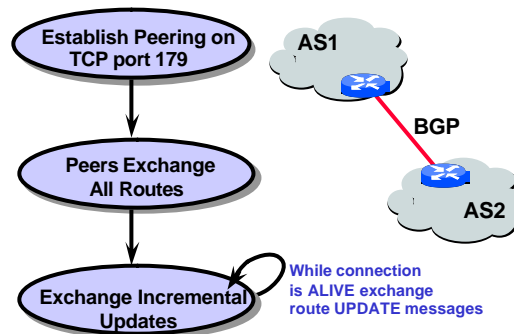
For any given network prefix p , a BGP router can receive at most N announcements for p , where N is the number of BGP peers of the router. All accepted routes must be kept in the BGP table, because if a best route is withdrawn, the router must install a new best route without querying its peers. Therefore, an AS may filter routes simply to ensure that its routers do not run out of memory.



BGP 4

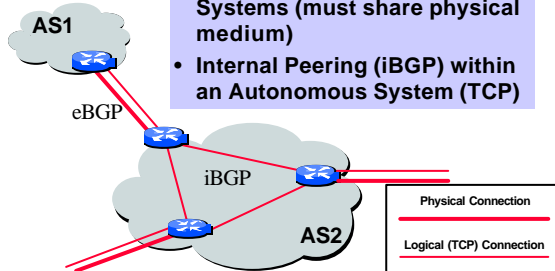
- BGP = **B**order **G**ateway **P**rotocol
- Is an exterior routing protocol (EGP)
- Is a **Policy-Based** routing protocol
- Is the **de facto EGP** of today's global Internet
- Has a reputation for being complex
- Supports hierarchical routing
- Is a distance vector protocol

BGP Operations Simplified



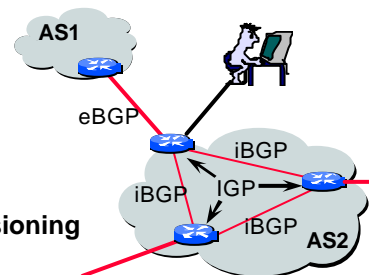
Two Types of BGP Peering Relationships

- External Peering (eBGP) between Autonomous Systems (must share physical medium)
- Internal Peering (iBGP) within an Autonomous System (TCP)



Sources of BGP Routes

- eBGP peers
- iBGP peers
- IGP
- Static provisioning



Four Types of BGP Messages

- **Open** : Establish a peering session.
- **Keep Alive** : Handshake at regular intervals.
- **Notification** : Error messages that terminate a peering session and result in all of peer's routes being invalidated.
- **Update** : Announcing new routes or withdrawing previously announced routes.

announcement = network prefix + attributes

BGP Attributes Types

- Well-known vs. Optional : Well-known attributes must understood by all BGP implementations.
- Mandatory vs. Discretionary : Mandatory attributes must be included in every route description.
- Transitive vs. nontransitive : Transitive attributes can be passed along unmodified.

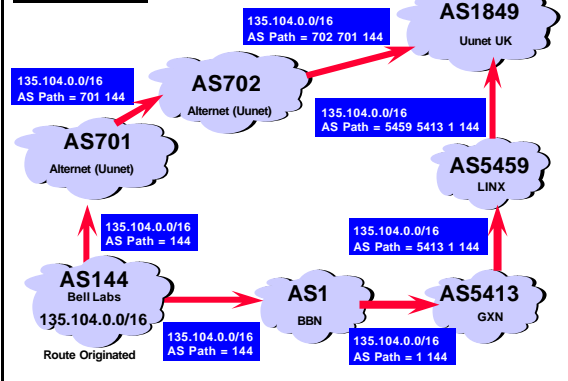
How about these too?

- External vs. Internal
- Filter
- Decision
- Cisco vs. Non-Cisco....

BGP Attributes

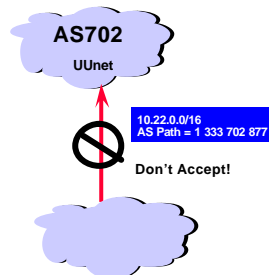
- **AS path** (well-known, mandatory)
- **Next Hop** (well-known, mandatory)
- **Origin** (well-known, mandatory)
- **Mult Exit Discriminator** (Optional, nontrans, eBGP)
- **Local Preference** (well-known, discretionary, iBGP)
- **Community** (Optional, trans)
- **Atomic Aggregate** (well-known, discretionary)
- **Aggregator** (Optional, trans)
- **Originator ID** (Optional, nontrans, Cisco)
- Other vendor-specific optional attributes ...

AS Path Attribute

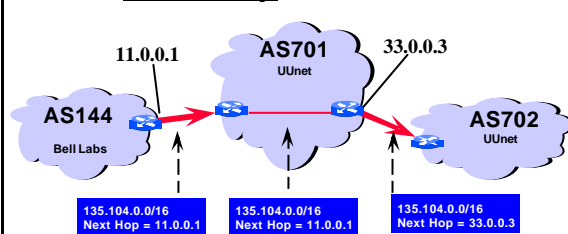


AS Path Attribute (cont.)

BGP at AS YYY will never accept a route whose **AS Path** contains YYY. This helps to minimize interdomain routing loops.

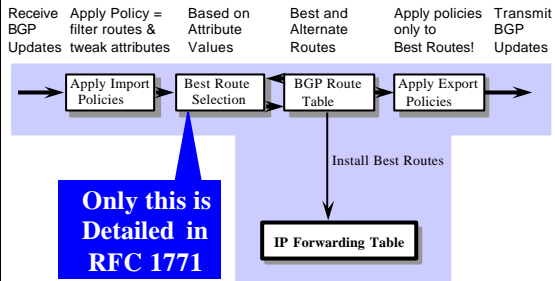


Next Hop Attribute



Every time a route announcement crosses an AS boundary, the Next Hop attribute is changed to the IP address of the border router that announced the route.

BGP Route Processing



BGP Best Route Selection Process

When two routes with the same network prefix ...

- 1) Pick route with highest LOCAL PREFERENCE ▶ Allows policy to override distance metric
- 2) Pick route with shortest AS PATH
- 3) Pick route with lowest MED
- 4) Pick route with closest NEXT HOP (via IGP)
- 5) Pick route from router with lowest IP address (break tie)

What does a BGP Routing Table Look Like ?

destination	Next-hop	AS-path
> 135.104.0.0	198.32.146.20	1740 701 144
+	198.32.136.5	1740 1 144
+	198.32.136.55	1740 1 144
+	198.32.176.25	1740 1 144
+	192.157.69.5	1740 1 144
> 135.180.0.0	198.32.146.20	1740 701 144
+	198.32.136.5	1740 1 144
+	198.32.136.55	1740 1 144
+	198.32.176.25	1740 1 144
+	192.157.69.5	1740 1 144
> 192.20.110.0	198.32.146.20	1740 701 144
+	198.32.136.5	1740 1 144
+	198.32.136.55	1740 701 144
+	198.32.176.25	1740 701 144
+	192.157.69.5	1740 1 144
> 192.20.115.0	198.32.146.20	1740 701 144
+	198.32.136.5	1740 1 144
+	198.32.136.55	1740 701 144
+	198.32.176.25	1740 701 144
+	192.157.69.5	1740 1 144
> 192.20.120.0	198.32.146.20	1740 701 144
+	198.32.136.5	1740 1 144
+	198.32.136.55	1740 701 144
+	198.32.176.25	1740 701 144
+	192.157.69.5	1740 1 144
> 204.178.0.0/19	198.32.96.25	1740 1 144
>	198.32.146.20	1740 1 144
>	198.32.176.25	1740 1 144
+	192.157.69.5	1740 1 144
> 207.140.138.0	198.32.146.20	1740 7018 144
+	198.32.136.25	1740 7018 144
+	198.32.176.25	1740 7018 144
+	192.157.69.5	1740 7018 144

from route-server.cerf.net
(in AS 1847)

This table fragment
shows all routes to
Bell Labs (AS 144)

Full table contained about
120k different prefixes and
about ~600k entries.

1 = BBN (GTE)
144 = Bell Labs
701 = UUNET
1740 = CerfNet
7018 = AT&T