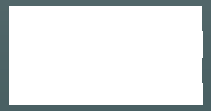
Routing & Protocols

Paul Traina
cisco Engineering



Today's Talk

- Terminology
- Routing
- Static Routes
- Interior Gateway Protocols
- Exterior Gateway Protocols
- Building an ISP network

Terminology

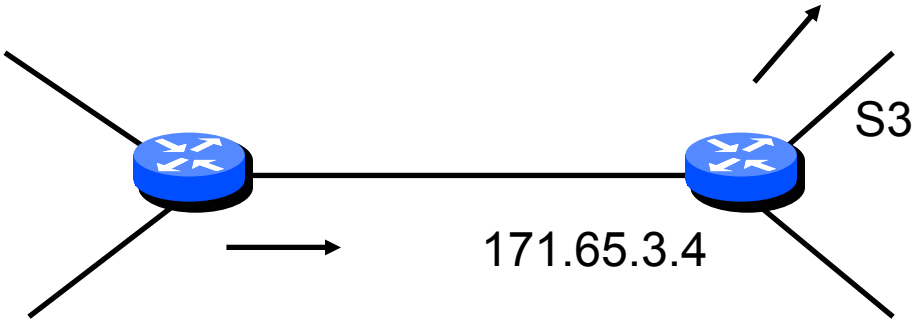
- network number
- prefix
- mask (or length)

Static routes

hand configured routing

- tell the router which way to send packets
- based upon final packet destination

Static routes



- `ip route 10.0.0.0 255.0.0.0 serial 3`
- `ip route 131.108.0.0 255.255.0.0 171.65.3.4`

Terminology

Interior Gateway Protocol (IGP)

- RIP, IGRP, HELLO, OSPF
- Primary goal is optimal connectivity
- Strong distance metrics
- May not have good administrative controls

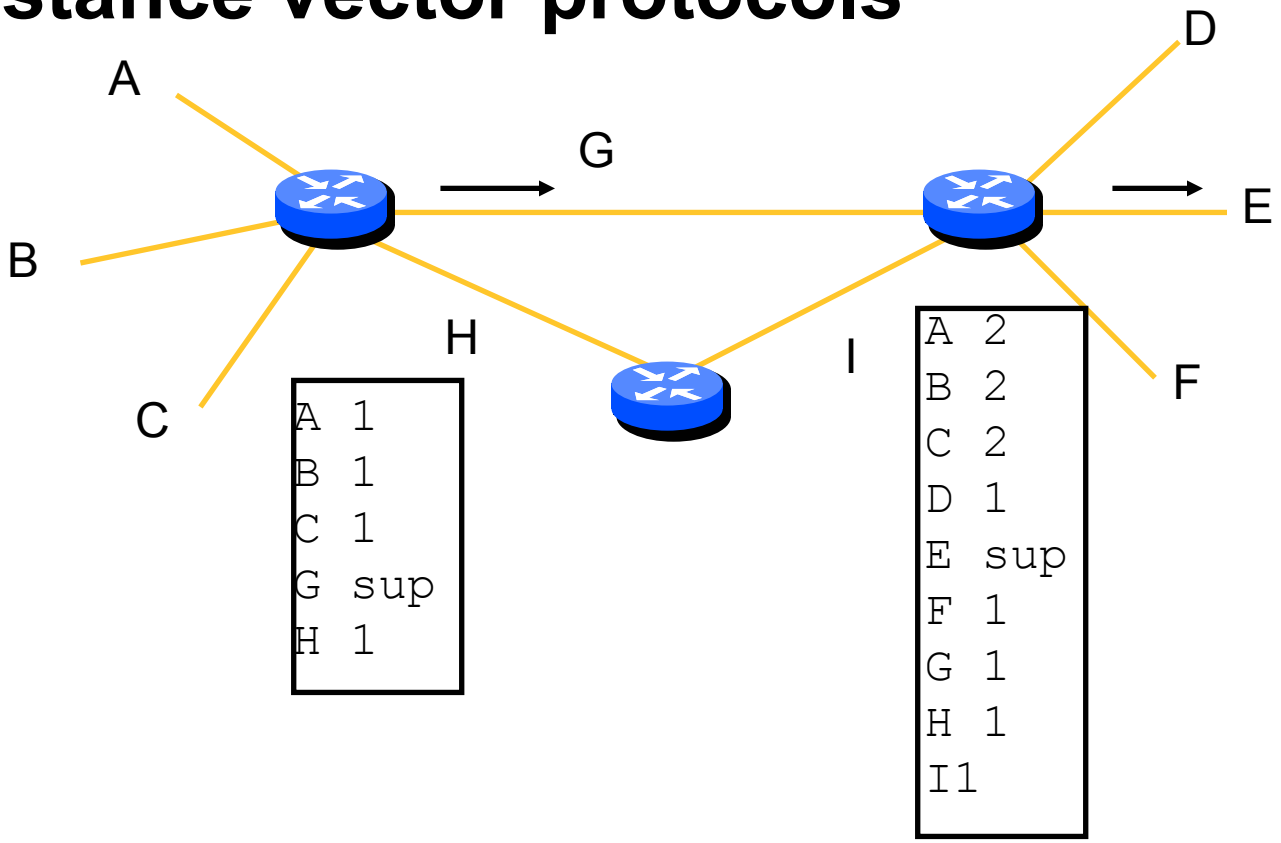
Terminology

Distance vector protocols

- listen to neighboring routers
- install routes in table, lowest distance wins
- advertise all routes in table
- very simple
- very stupid

Terminology

Distance vector protocols



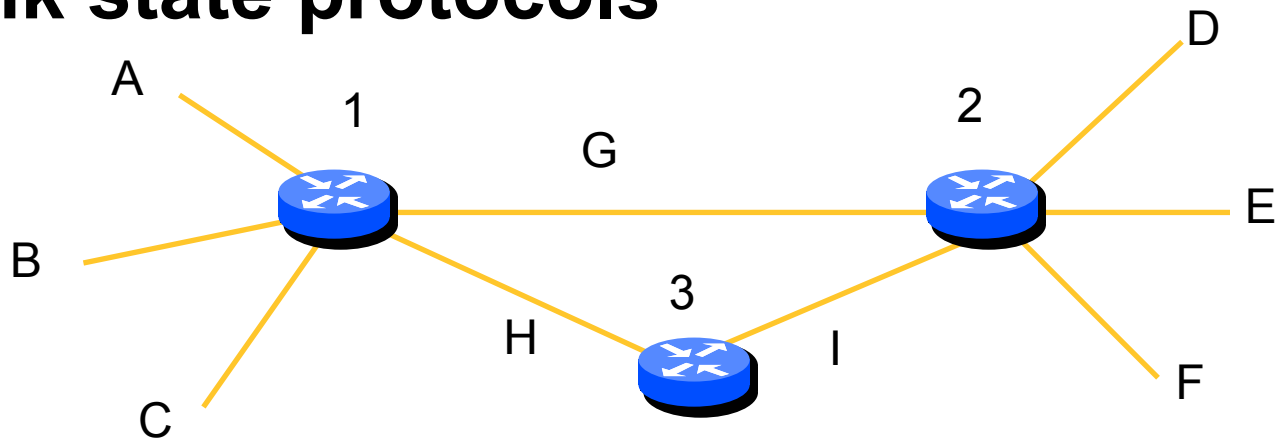
Terminology

Link state protocols

- information about adjacencies sent to all routers
- each router builds a topology database
- a "shortest path" algorithm is used to find best route
- converge as quickly as databases can be updated

Terminology

Link state protocols



router 1
A, B, C, G, H

router 3
H, I

router 2
D, E, F, G, I

A - 1 - G - 2 - D

Interior Gateway Protocols

Routing Information Protocol (RIP)

- IP only
- distance vector protocol
- slow convergence
- does not carry mask information
- reasonably simple design & configuration
- does not scale (maximum 15 hops)
- poor metrics (hop-count)

Interior Gateway Protocols

Interior Gateway Routing Protocol (IGRP)

- IP only
- distance vector protocol
- slow convergence (like RIP)
- does not carry mask information (like RIP)
- very simple design & configuration
  - powerful proprietary metric
  - load sharing across diverse links

Interior Gateway Protocols

The IGRP metric

- always get optimal routing
- metric vector, not single value
  - bandwidth
  - delay
  - hops
  - reliability
  - loading

Interior Gateway Protocols

Enhanced IGRP

- multi-protocol (IP, IPX, Appletalk)
- fast convergence (like OSPF)
- very simple design & configuration (like IGRP)
  - IGRP metric
  - allows load sharing across diverse links

Interior Gateway Protocols

Enhanced IGRP

- distance vector based protocol
- NOT a Bellman-Ford protocol

Uses "dual" algorithm

- alternative to OSPF & I-ISIS
- can be bandwidth intensive on slow links

Interior Gateway Protocols

Integrated IS-IS (I-ISIS)

- multi-protocol (CLNP, IP, IPX, ...)
- link state protocol
- fast convergence
- design and architecture moderately complex
- configuration may be simple

Interior Gateway Protocols

Open Shortest Path First (OSPF)

- IS - IS = 0

Interior Gateway Protocols

Open Shortest Path First (OSPF)

- IP only
- link state protocol
- fast convergence
- design and architecture very complex
- configuration can be simple

Interior Gateway Protocols

Which to use?

- Your interior network is actually VERY simple.
- Your IGP should only carry your routes and your direct customers'

Interior Gateway Protcols

Problems with "classic" protocols

- slow convergence
- count to infinity
- no mask information
  - no CIDR
  - no VLSM
  - no subnet 0

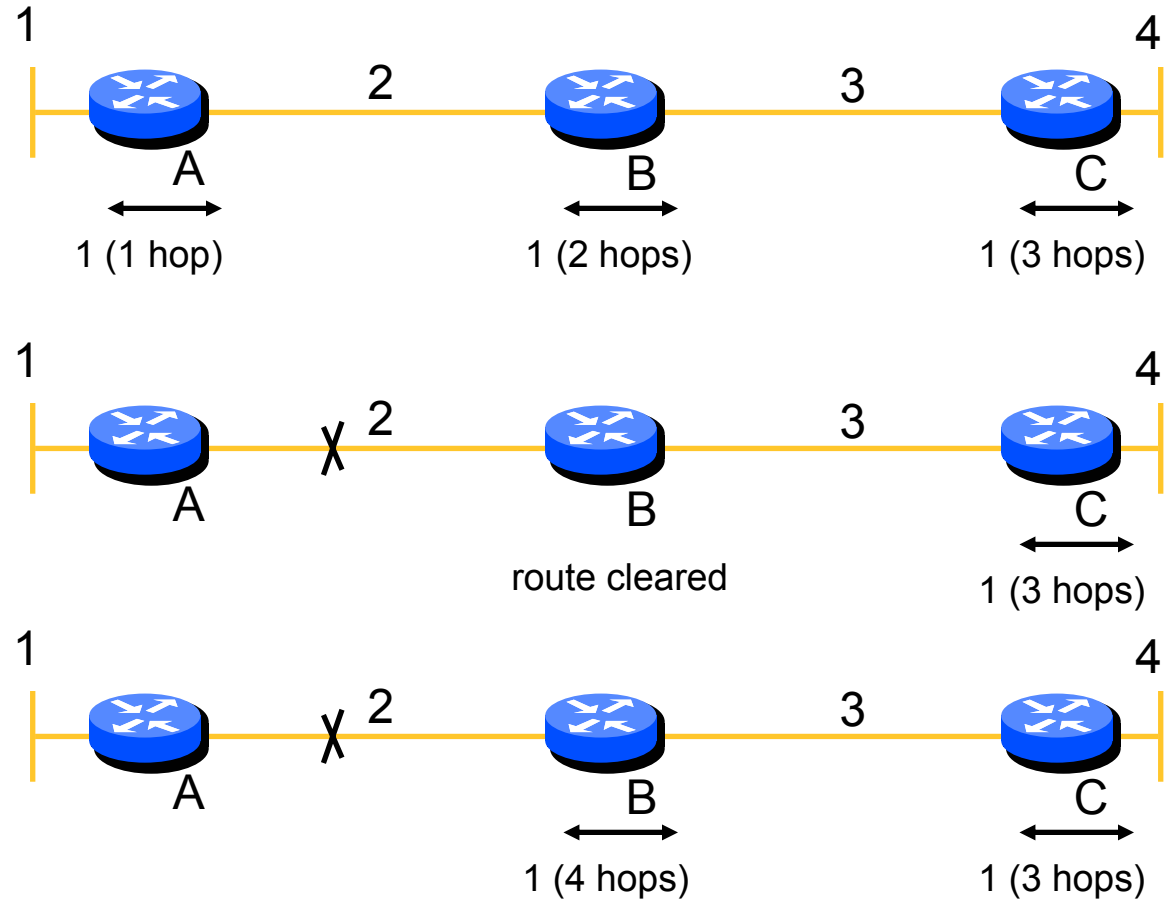
Interior Gateway Protocols

Slow convergence

- advertisement period
  - entire routing table dumped every n seconds
- timeout period
  - usually 3 times advertisement period
- RIP values are normally 30 and 90 seconds!

Interior Gateway Protocols

Count to infinty problem



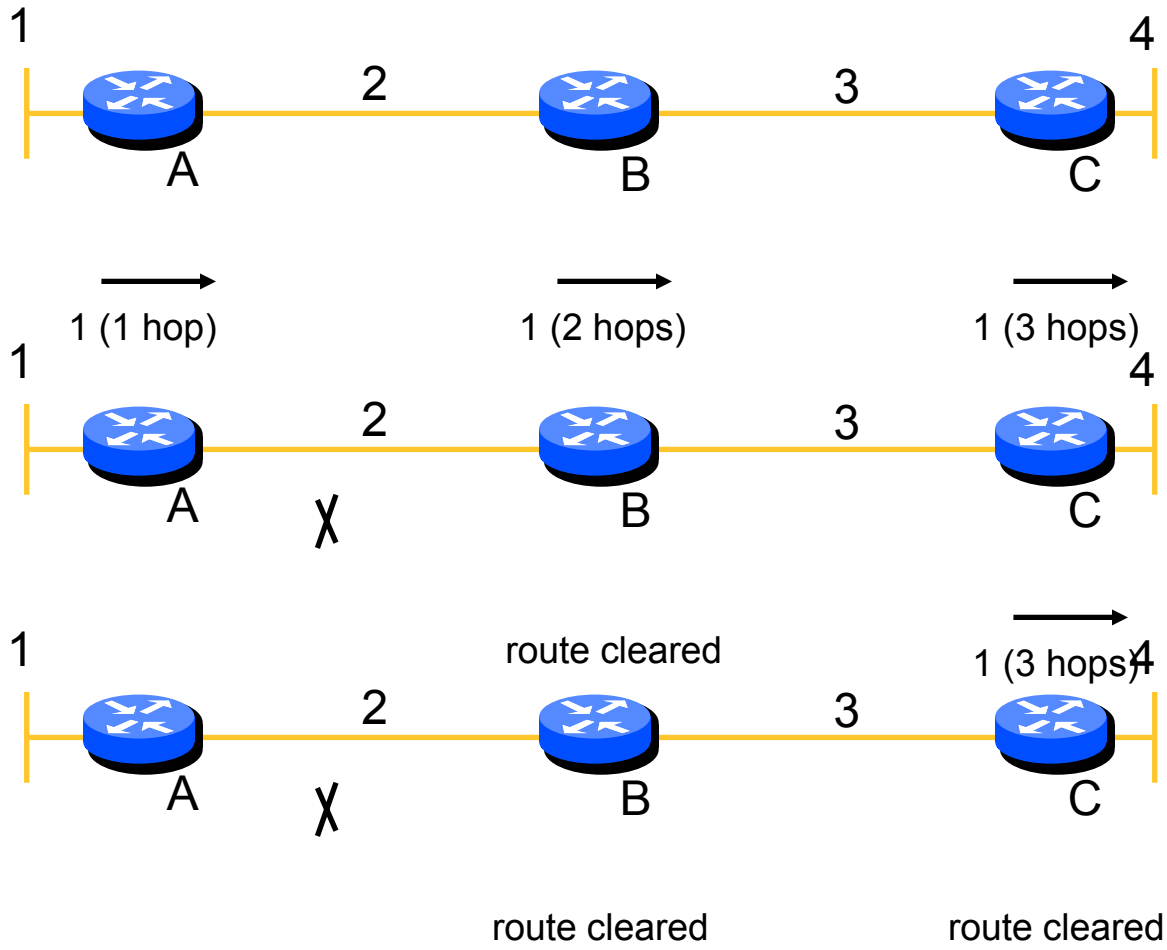
Interior Gateway Protocols

Count to infinity: split-horizon

- Don't feed selected route back to source
  - no feedback on source interface
  - no feedback to source neighbor

Interior Gateway Protocols

Count to infinity: split-horizon



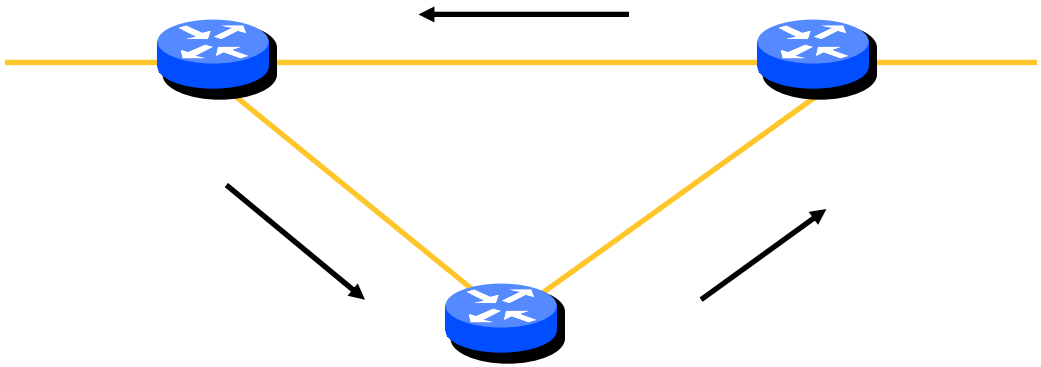
Interior Gateway Protocols

Count to infinity: hold-down

- Split horizon not sufficient!
- Holddown period
  - interval during which "less attractive" updates are ignored

Interior Gateway Protocols

Count to infinity: hold-down



Interior Gateway Protocols

The universal rule

- You will always trade bandwidth for speed of convergence

Interior Gateway Protocols

OSPF configuration

- myth
  - OSPF is hard to use
- reality:
  -
```
router ospf 1
network 192.111.107.0 0.0.0.255 area 0
```

Interior Gateway Protocols

OSPF operation

- every OSPF router sends out 'hello' packets
- hello packets used to determine if neighbor is up
- hello packets are small easy to process packets
- hello packets are sent periodically (usually short interval)

Interior Gateway Protocols

OSPF operation

- once an adjacency is established, trade information with your neighbor
- topology information is packaged in a "link state announcement"
- announcements are sent ONCE, and only updated if there's a change
  - (or every 45mins...)

Interior Gateway Protocols

OSPF operation

- change occurs
- broadcast change
- run SPF algorithm
- install output into forwarding table

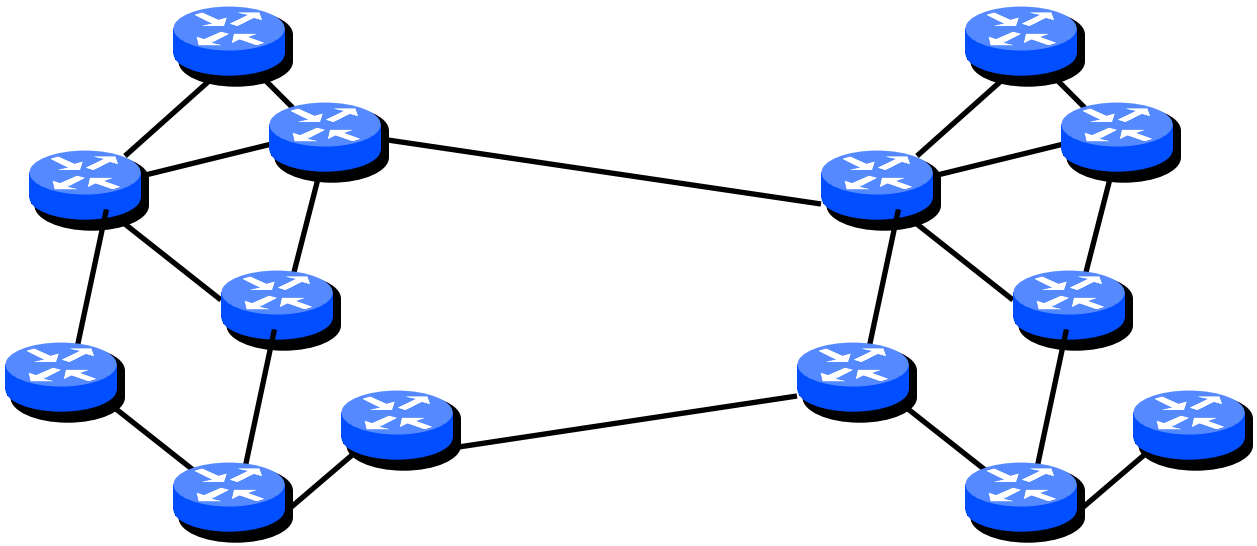
Interior Gateway Protocols

making OSPF scale

- each link transition causes a broadcast and SPF run
- OSPF can group routers to appear as one single router
- OSPF areas

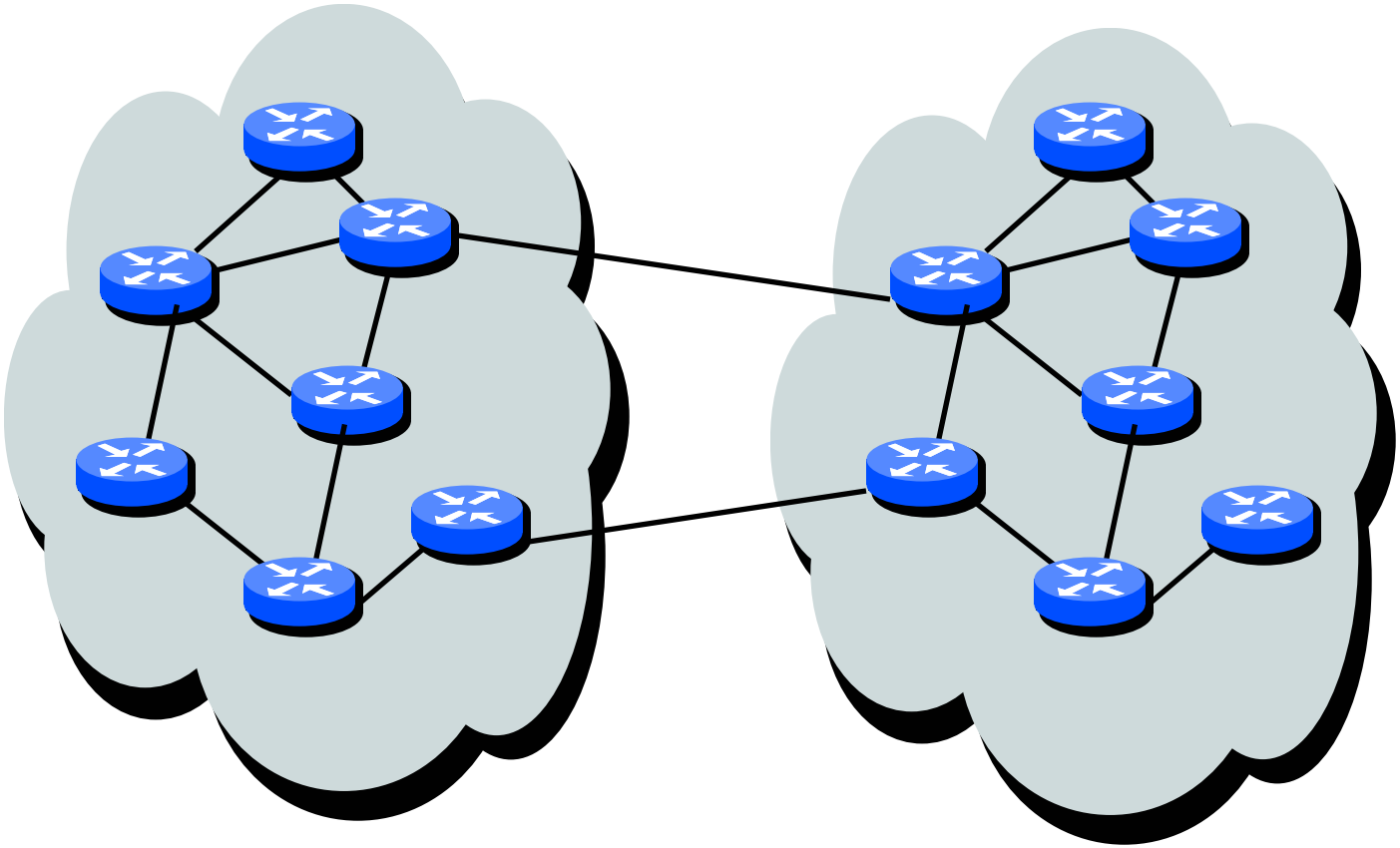
Interior Gateway Protocols

OSPF areas (before)



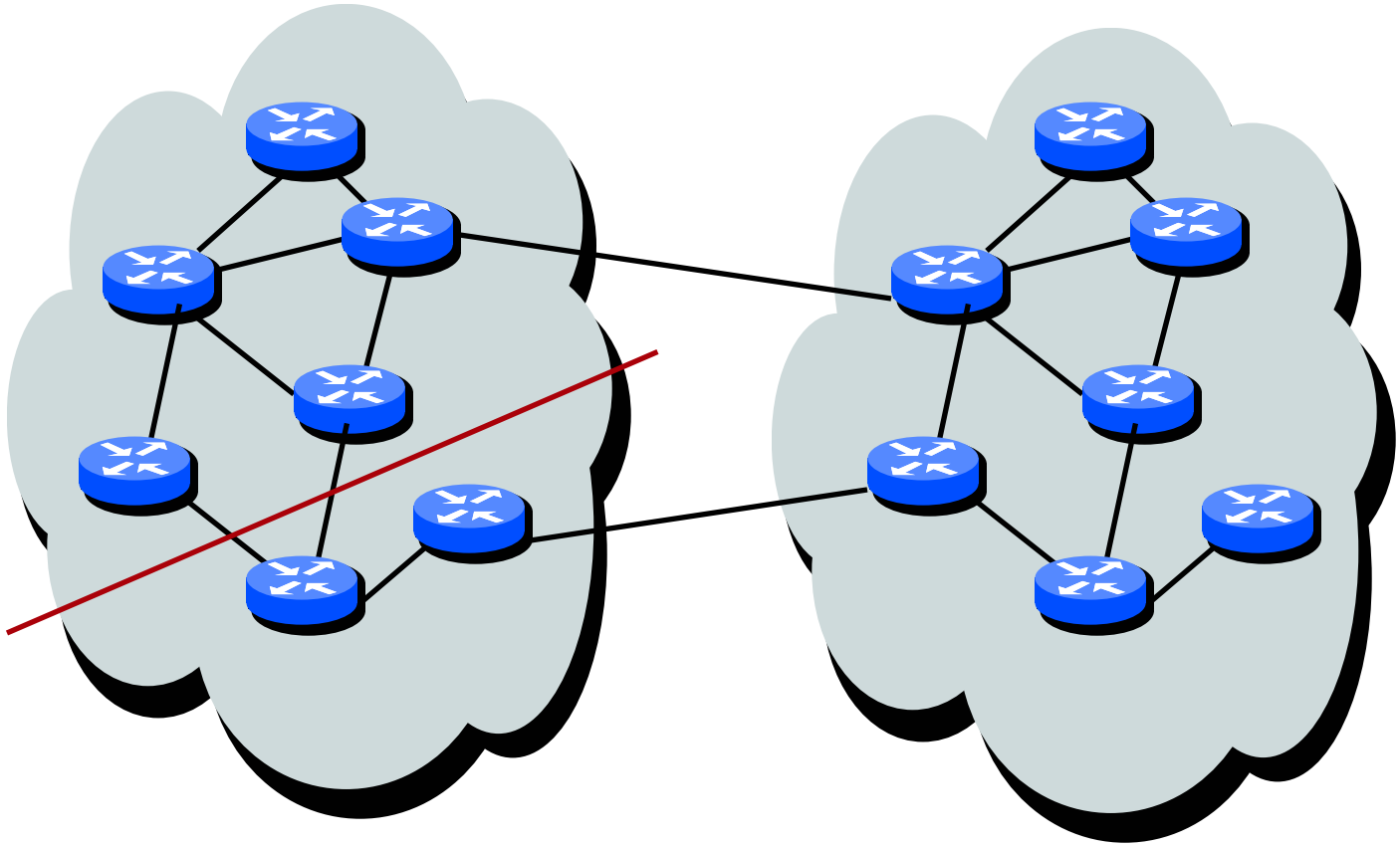
Interior Gateway Protocols

OSPF areas (after)



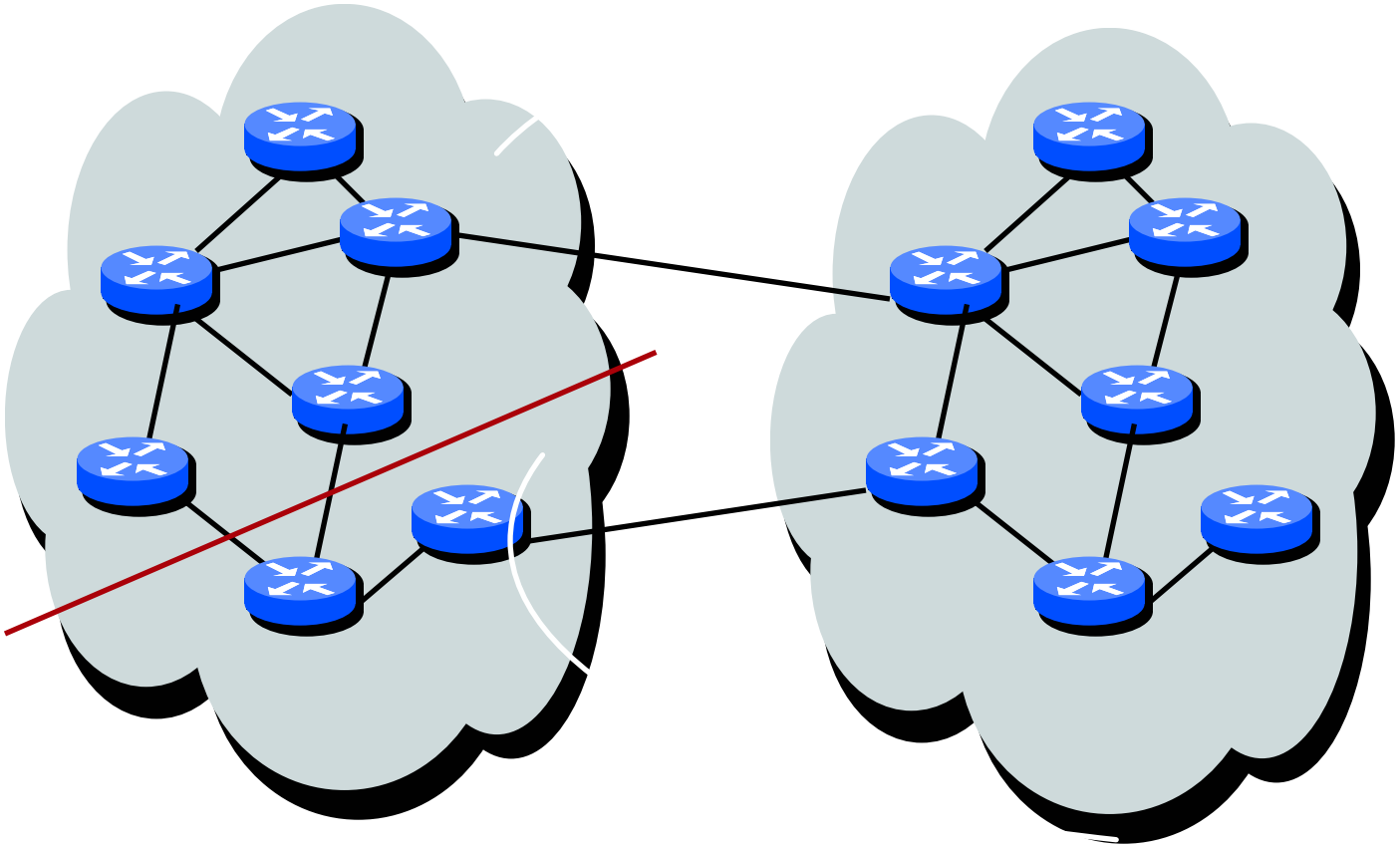
Interior Gateway Protocols

OSPF areas - partitioning



Interior Gateway Protocols

OSPF areas - partition repair



Interior Gateway Protocols

OSPF areas

- rule of thumb:

no more than 150 routers/area

- reality:

no more than 500 routers/area

- backbone "area" is an area
- proper use of areas reduce bandwidth & CPU utilization

Interior Gateway Protocols

EIGRP operation

- design goals were
  - make it as fast as OSPF & IS-IS
  - make it trivial to configure
  - easy migration from IGRP

Interior Gateway Protocols

EIGRP operation

```
router eigrp 1
network 192.108.0.0 mask 255.255.0.0
```

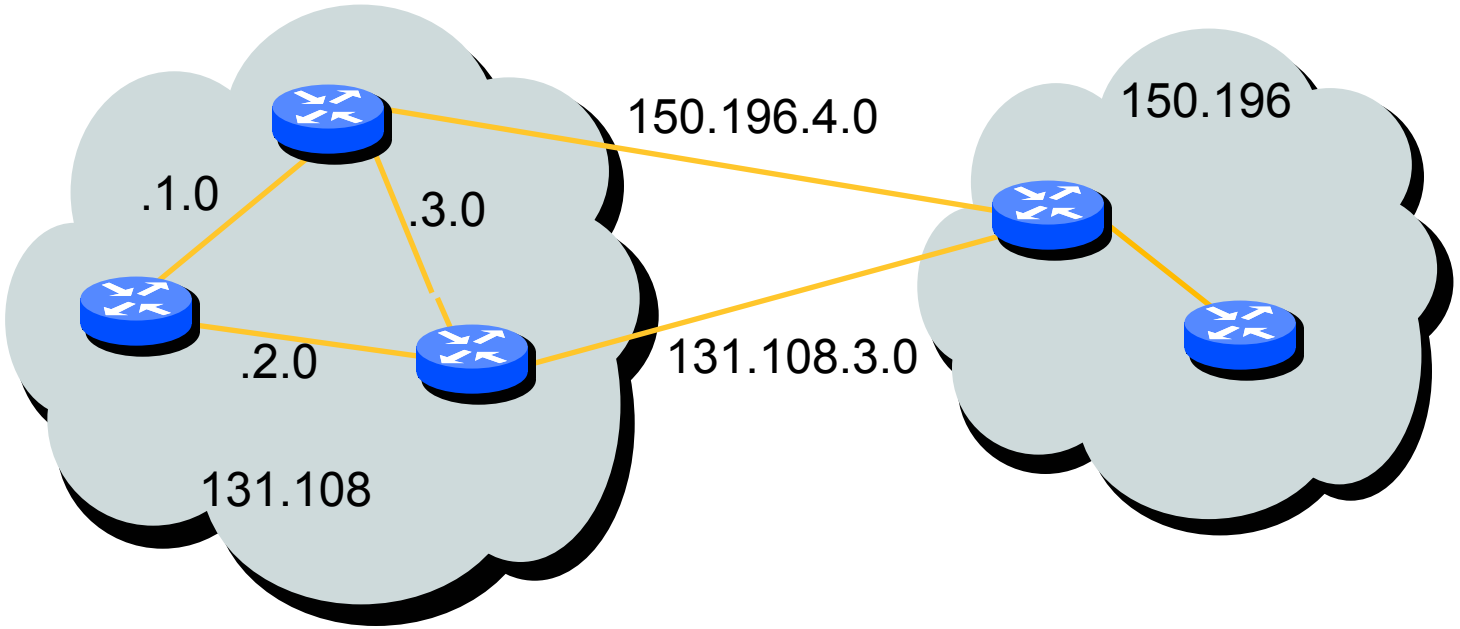
Interior Gateway Protocols

EIGRP operation - caveats

- nothing is for free
- EIGRP works best on high speed links
- EIGRP doesn't scale well in high-meshed frame-relay networks
  - star networks OK

Interior Gateway Protocols

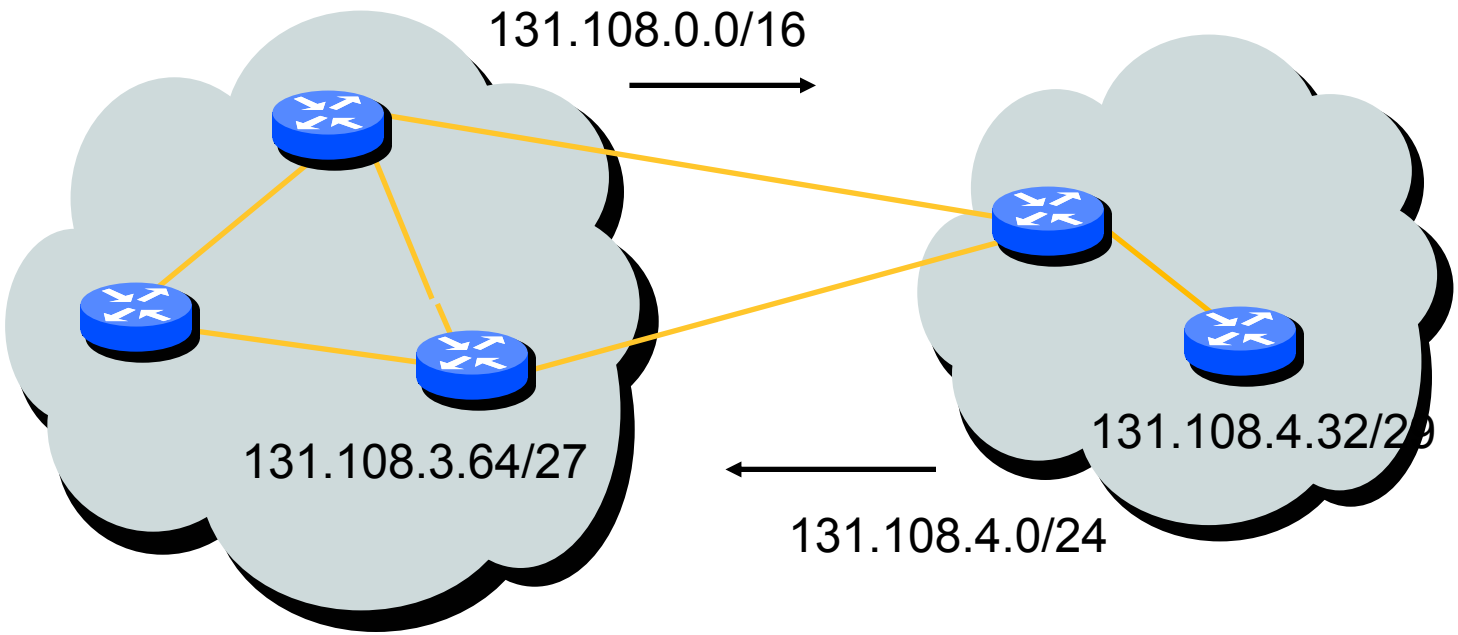
summarization



- classful routing protocols naturally summarize to network numbers at boundaries

Interior Gateway Protocols

summarization



- classless routing protocols summarize at arbitrary bit boundaries

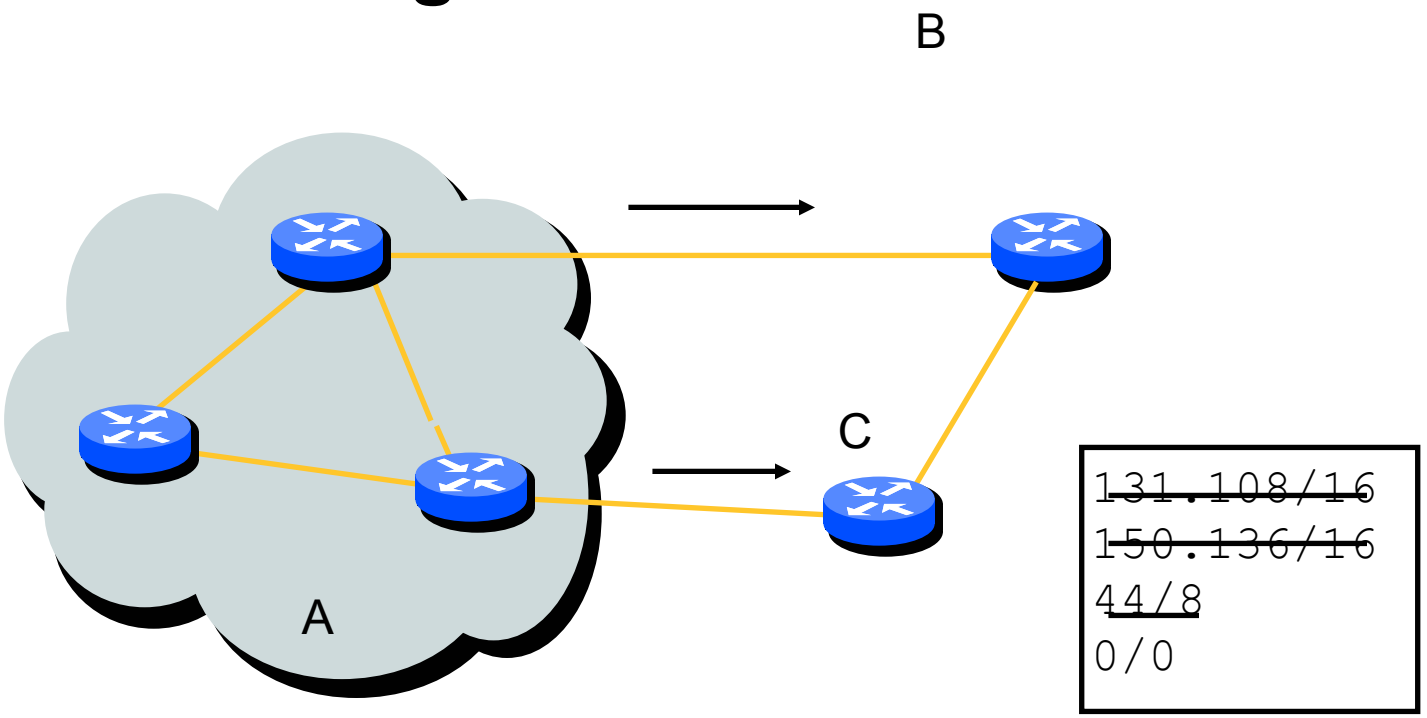
Interior Gateway Protocols

route filtering

- pseudo-security (bad idea!)
- low bandwidth links
- eliminate unnecessary information

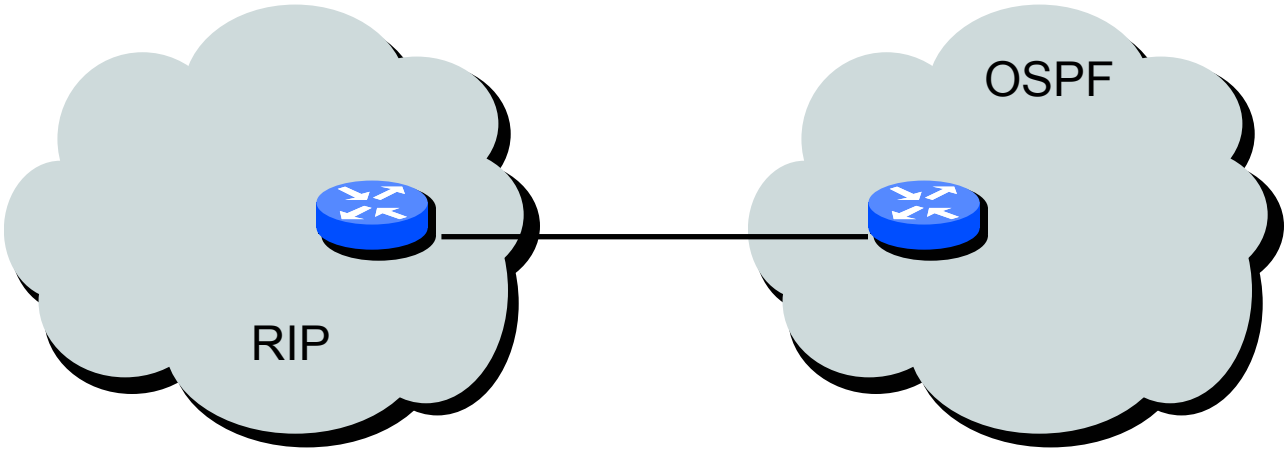
Interior Gateway Protocols

route filtering



Interior Gateway Protocols

redistribution



- you run OSPF
- your neighbor runs RIP

Interior Gateway Protocols

redistribution

- run RIP on their interface
- router rip

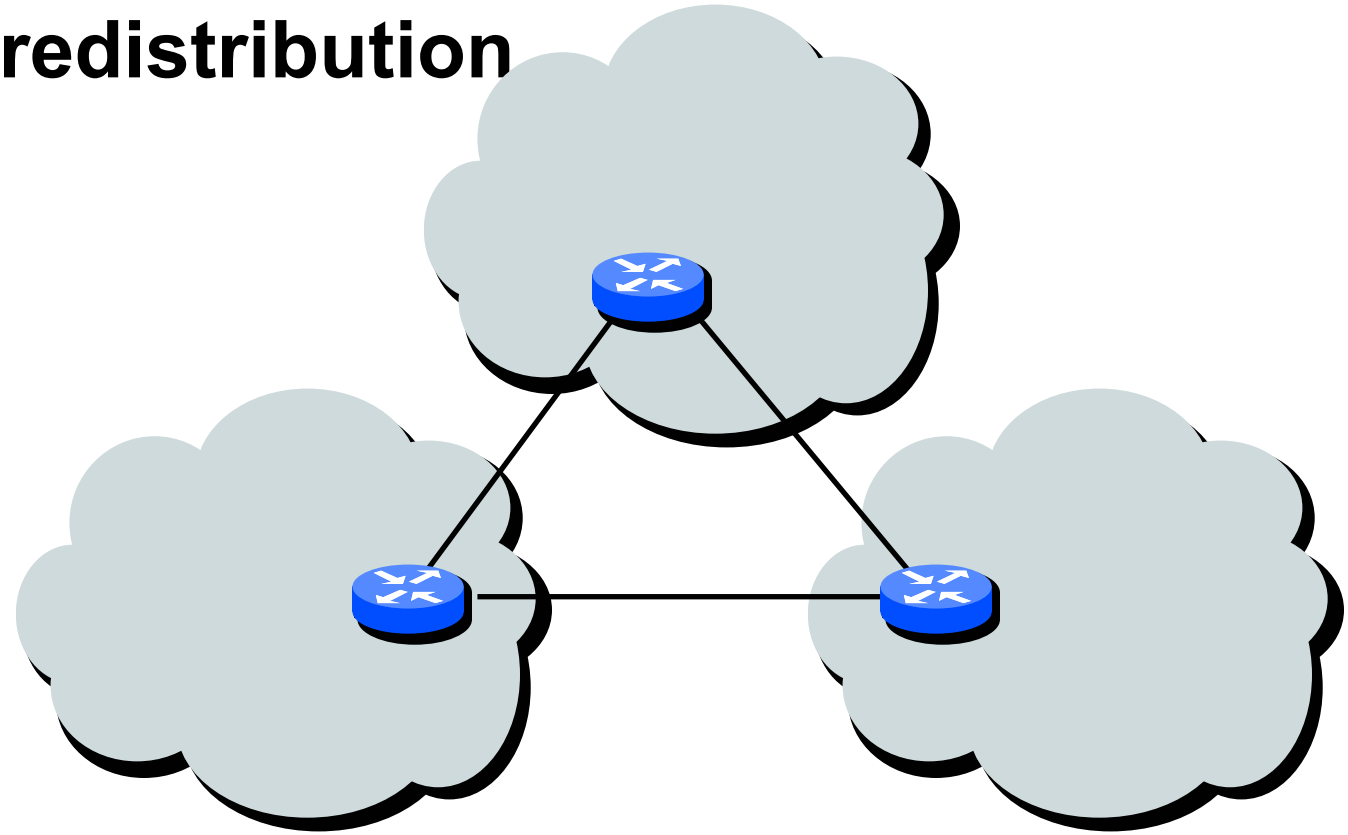
```
network 192.111.107.0
```

- configure OSPF to redistribute RIP
- router ospf 1

```
network 135.111.104.0
0.0.0.255 area 0
redistribute rip metric 10
```

Interior Gateway Protocols

redistribution



- bi-directional redistribution MUST be filtered!

Interior Gateway Protocols

redistribution

- router rip

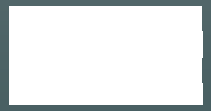
```
network 192.111.107.0
```

- router ospf 1

```
network 135.111.104.0 0.0.0.255
area 0
redistribute rip metric 10
distribute-list 1 out rip
```

- access-list 1 permit 192.111.107.0 0.0.0.255

Exterior routing



Exterior routing

- Terminology
- What is exterior routing?
- Routing protocols
- Overview of BGP
- Putting it all together
- Further information

Terminology

Autonomous System

- A set of networks sharing the same routing policy.
- Internal connectivity
- One contiguious unit
- Identified by "AS number"
- Examples
  - service provider
  - multi-homed customer
  - anyone needing policy discrimination

Terminology

Exterior routes

- Routes learned from other autonomous systems

Terminology

Exterior Gateway Protocol

- egp vs EGP
- EGP, BGP, IDRP
- Primary goal is to provide reachability information outside administrative domain
- Secondary goal is administrative control
- Metrics may be arbitrary or weak

Terminology

Natural network mask

- Classful mask
  - Class A = 8 bits
    - networks 1...127
  - Class B = 16 bits
    - networks 128.0...191.255
  - Class C = 24 bits
    - networks 192.0.0...223.255.255

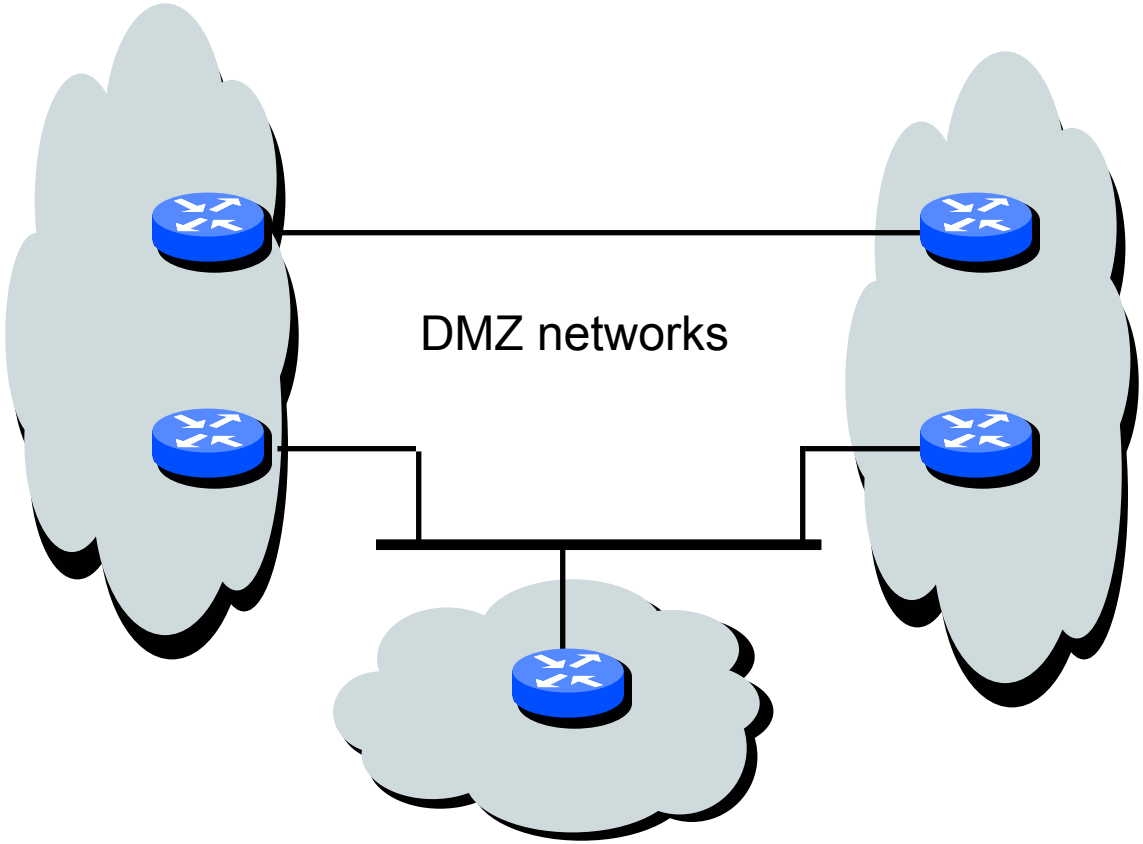
Terminology

DMZ network

- de-militarised zone
- area between North and South Korea
- shared network between ASs
  - before, neither AS carried it in IGP
  - now, both carry it in IGP

Terminology

DMZ network



Why do we need exterior routing?

Why not make entire internet a single cloud?

- separate policy control
- filtering on networks doesn't scale well
- service provider selection given multiple choices
- everything must scale to hundreds of thousands of routes

Exterior Routing

- static routes
- multiple IGP instances
- OSPF inter-domain routing
- EGP
- IDRP
- BGP version 4

Exterior Routing

Static routes

- no path information
- very versatile
- low protocol overhead
- high maintenance overhead
- very very very bad convergence time
  - requires manual configuration

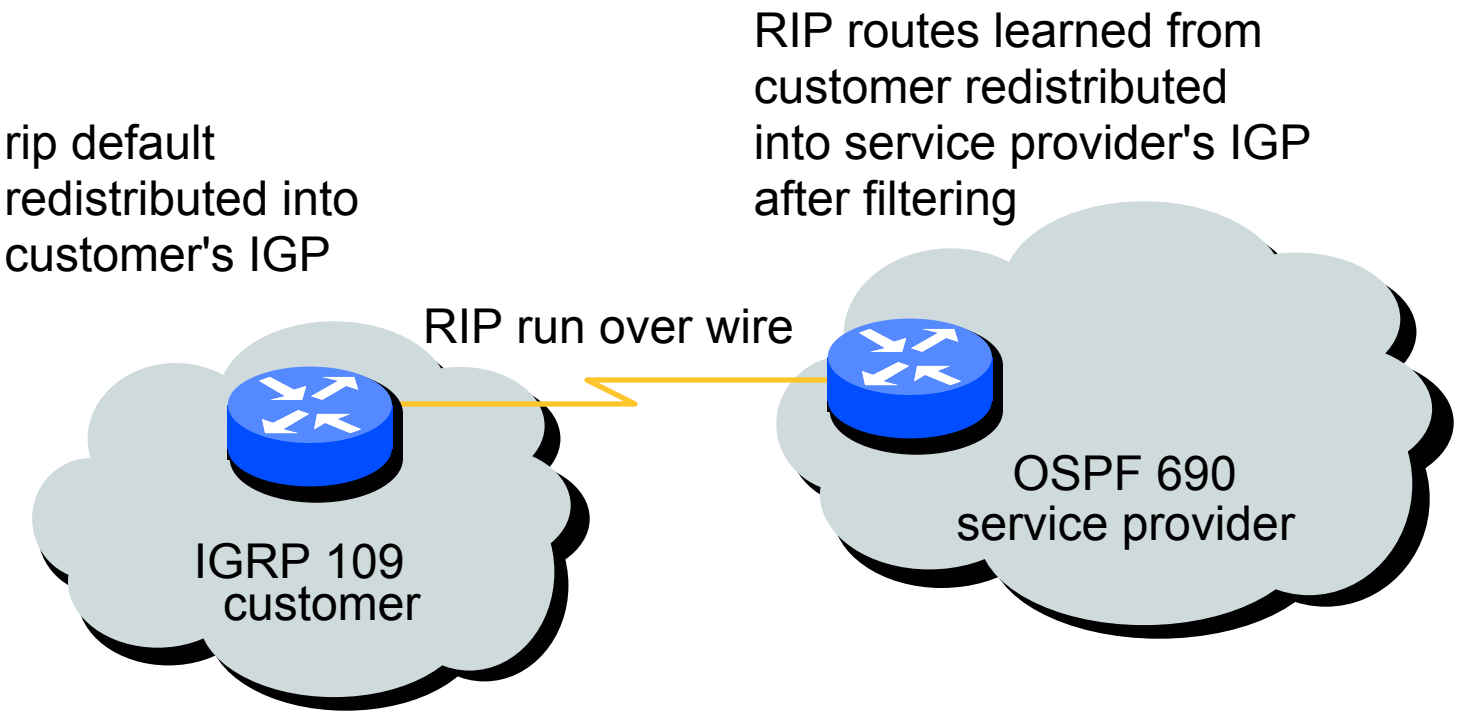
Exterior Routing

Multiple IGPs with route leaking

- Run an instance of an IGP at each site for local routing
- Run a backbone IGP at each border router
- redistribute local IGP into backbone IGP
- redistribute backbone IGP into local IGP (or default)
- backbone routers share common administration

Exterior Routing

Multiple IGPs with route leaking



Exterior Routing

Multiple IGPs with route leaking

- backbone IGP
  - router ospf 690

```
network 129.119.0.0 0.0.255.255

area 0

redistribute rip metric 5

distribute-list 1 rip out
```

- local IGP
  - router igrp 109

```
network 131.108.0.0
ip default-network 140.222.0.0
```

Exterior Routing

OSPF inter-domain routing

- Route leaking formalised for one protocol
- OSPF tag carries originating AS
  - limited policy control
    - only have 32 bit OSPF tag
    - OSPF tag contains originating AS

Exterior Routing

Exterior Gateway Protocol

- historical protocol
- obsolete
- assumes a central core
- no transit service except via core

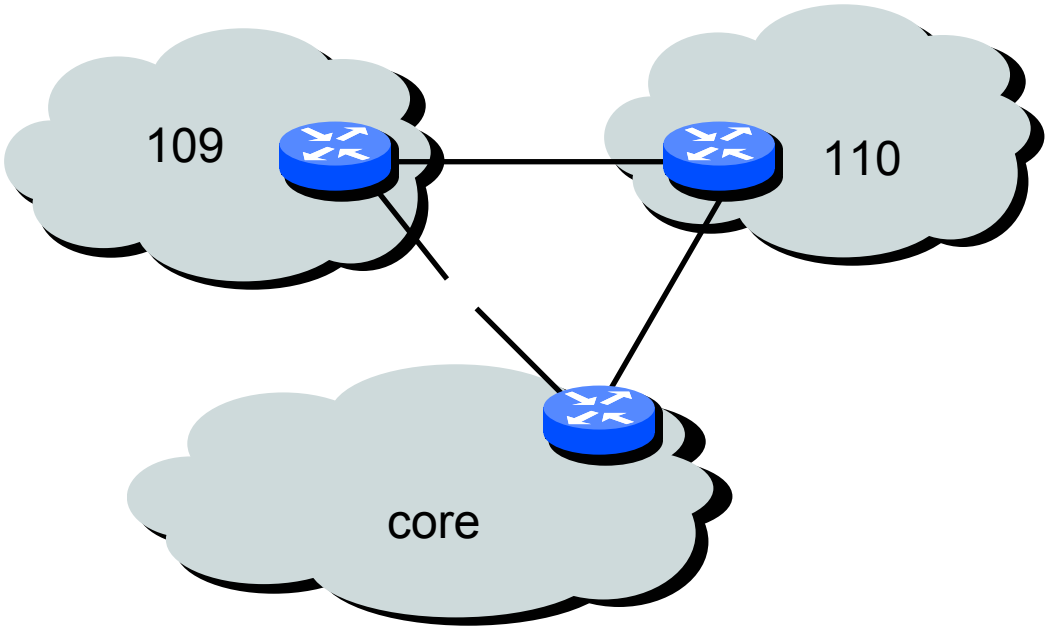
Exterior Routing

Exterior Gateway Protocol (historical)

- RIP by any other name
- fancy "hello dance"
- periodic update protocol
- entire routing table sent with each update
- no metric
  - everything is one hop from core

Exterior Routing

Exterior Gateway Protocol



- AS 110 may not advertise AS 109 to core

IDRP (future expansion path)

Inter-domain routing protocol

- IDRP is an almost identical clone of BGP-4
- IDRP is multi-protocol
  - IP
  - CLNP
  - IPX
- For purposes of this talk:

```
g/BGP-4/s//IDRP/g
```

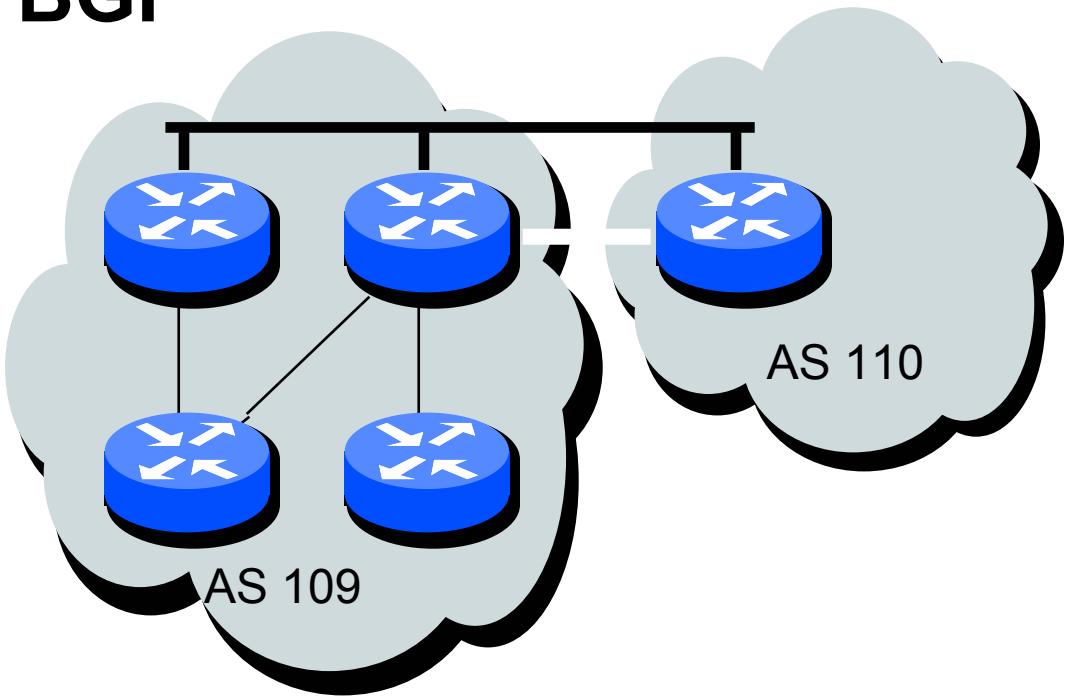
BGP-4

Border Gateway Protocol version 4

- carries external routes only
- uses reliable transport mechanism (TCP)
- not a periodic routing protocol
- allows limited policy selection
- AS path insures loop free routing
- "best path" determined at AS granularity

BGP peer relationships

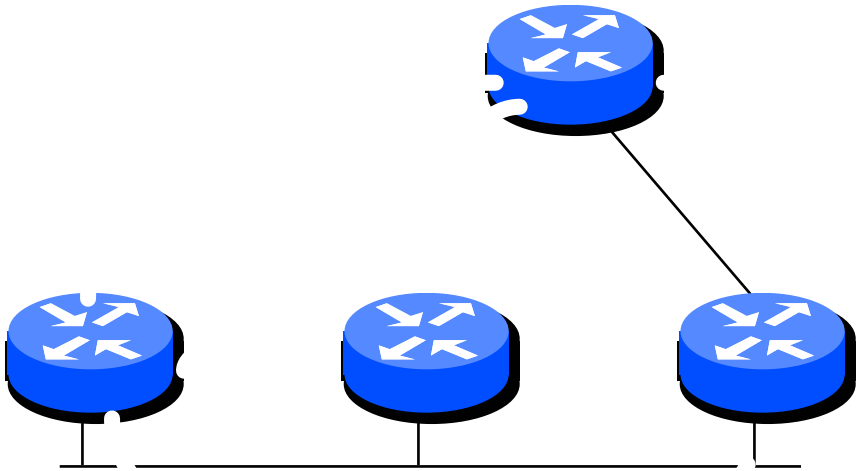
External BGP



- neighbor is in a different AS
- neighbors share a common network

BGP peer relationships

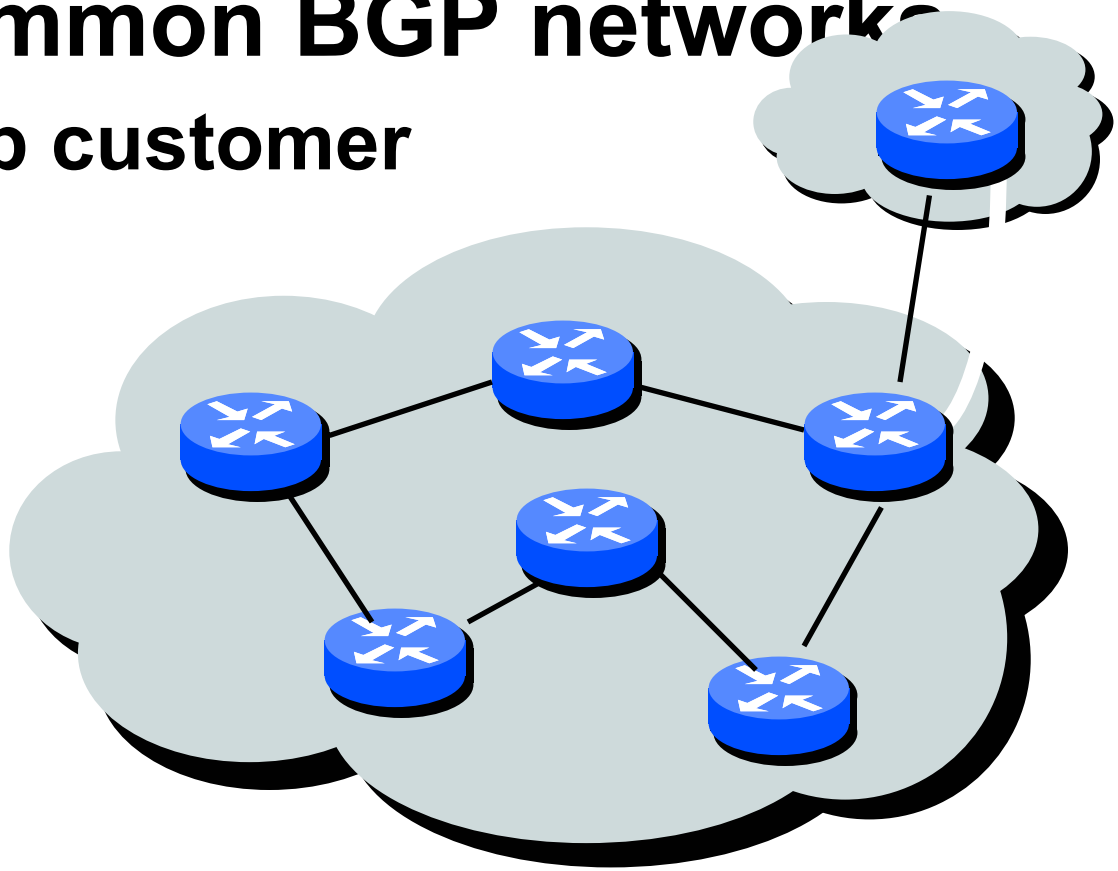
Internal BGP



- neighbor in same AS
- may be several hops away
- full neighbor mesh required

Common BGP networks

Stub customer



- BGP only at border
- default to border

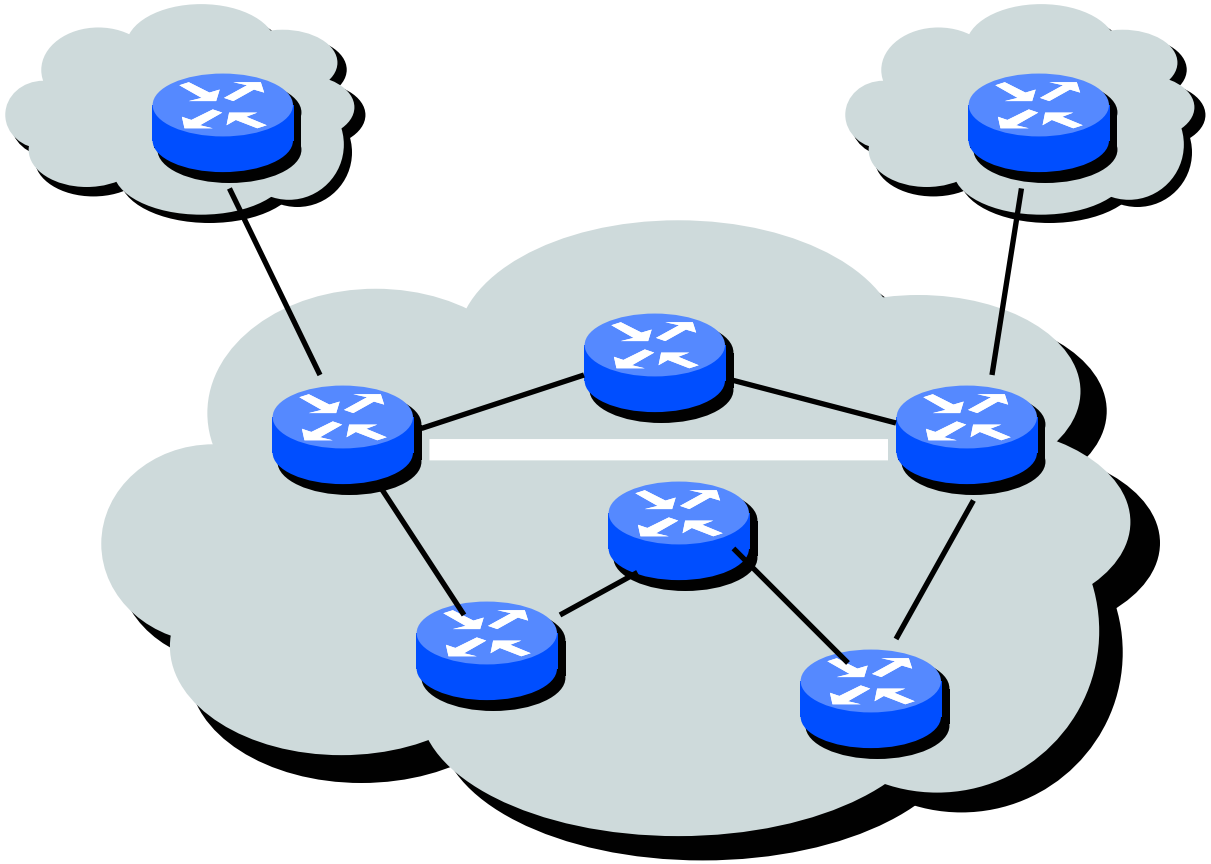
Common BGP networks

Multi-homed customer

- Internal BGP used with IGP
- IBGP only between border gateways
- Only border gateways speak BGP
- Synchronization with IGP required
- May use one IGP for exterior routes, and another for internal nodes
  - exterior routes must be redistributed into IGP

Common BGP networks

Multi-homed customer



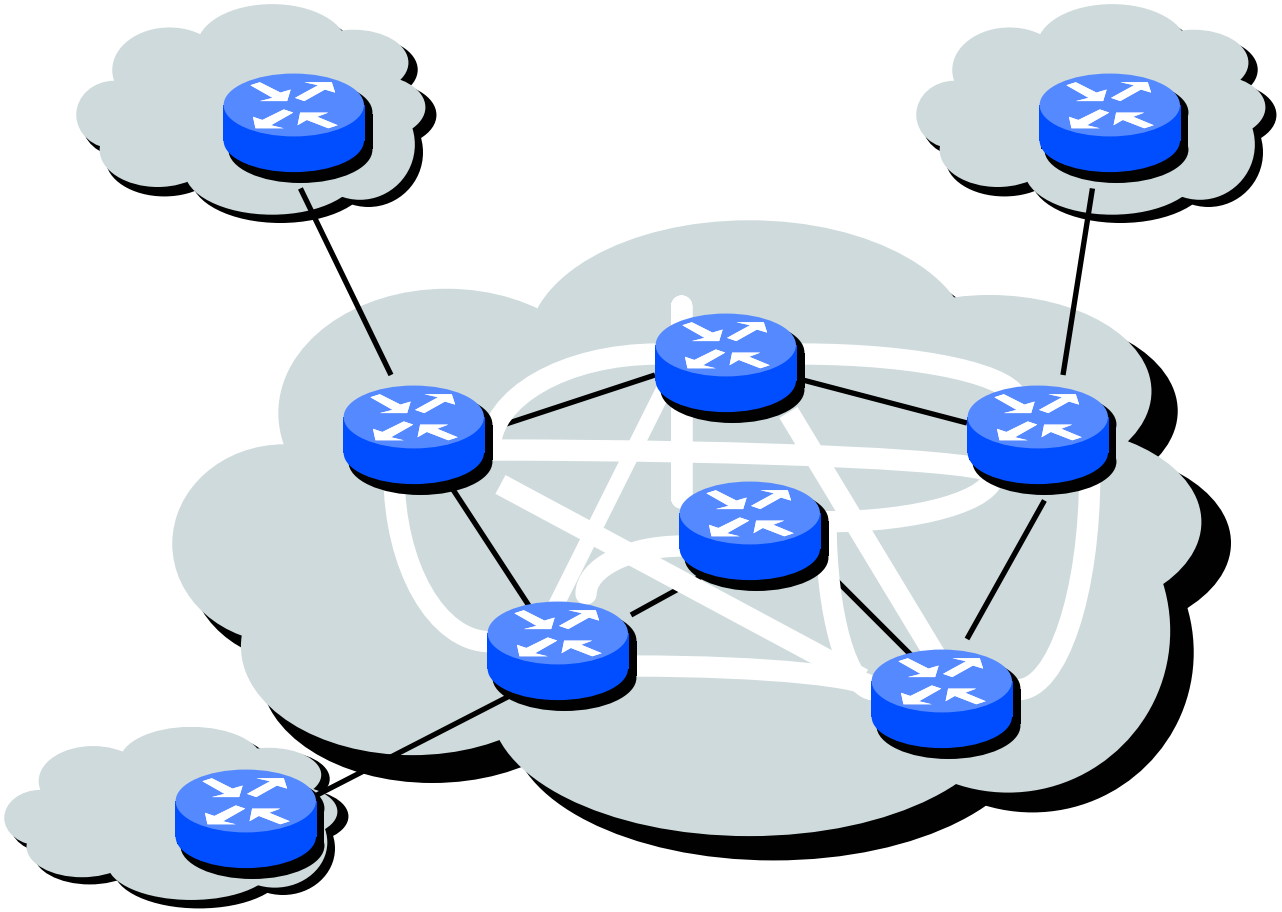
Common BGP networks

Service provider

- Internal BGP used to carry exterior routes
  - IGP carries local information only
  - Full mesh required if no IGP synchronization

Common BGP networks

Service provider



Common BGP networks

Service provider confederation

- A group of service providers
- Multiple connectivity points
  - multi-exit discriminator useful
- Not a special case

The BGP protocol

Update messages

- withdrawn routes
- attributes
- advertised routes

Update messages

Network reachability information

- prefix length
  - number of significant bits
- network prefix
  - 0 to 4 bytes
- Example:
  - 131.108/16
    - 131.108.0.0 255.255.0.0
  - 193/8
    - 193.0.0.0 255.0.0.0

Update messages

Attributes

- AS path
- next hop
- origin
- local preference
- multi-exit discriminator
- atomic aggregate
- aggregator

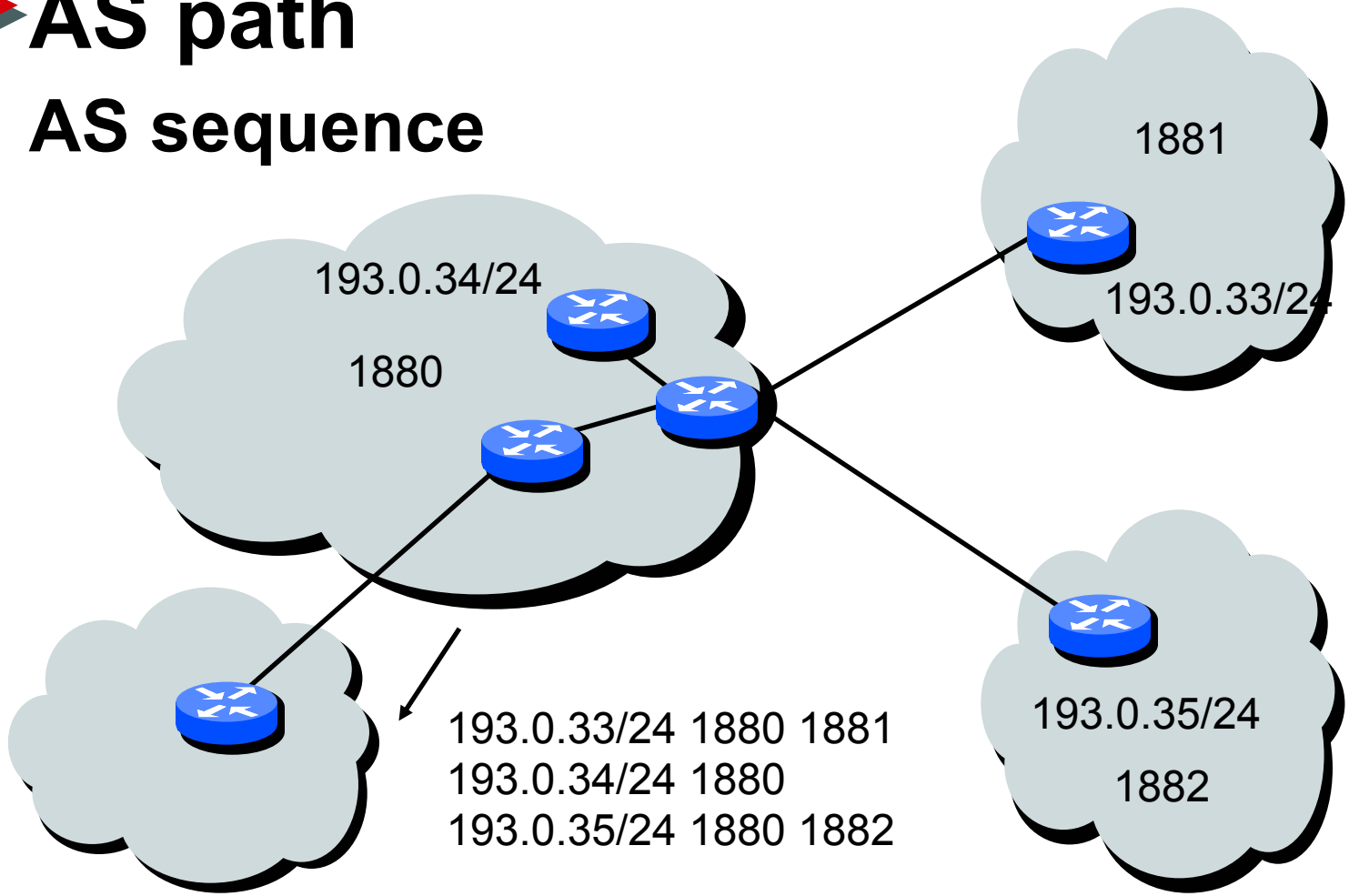
AS path

AS sequence

- a list of AS's that a route has traversed
  - `109 200 690 1755 1883`

AS path

AS sequence



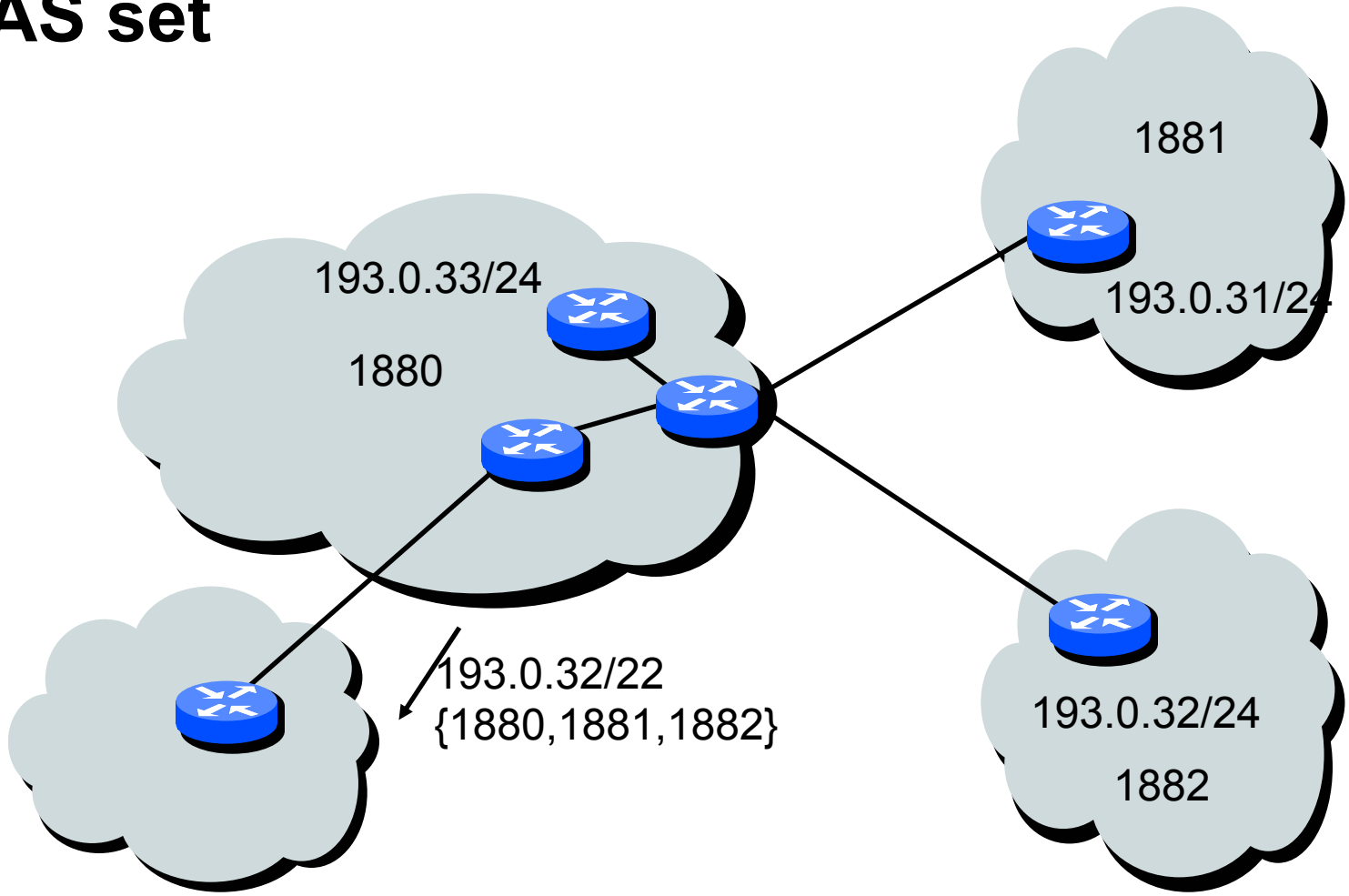
AS path

AS set

- path traversed one or more members of a set
  - `{1880,1881,1882}`

AS path

AS set



AS path

Sets and sequences combined

- local aggregation
  - `109 200 690 1755 {1881,1882,1883}`
- regional aggregation
  - `109 200 690 {1755,1881,1882,1883,...}`

BGP path selection

BGP maintains multiple "feasable" paths to a destination

- fast convergence
- routing based upon preferences
- Example:
  - `131.108/16` may be reached via AS path `690 200 109` or via AS path `690 1340 109`

BGP path selection algorithm

Initial route determination

- do not consider path if no next hop route
- largest weight
  - local to router
- highest local preference
  - global within AS
- shortest AS path

BGP path selection

Tie breaking

- multi-exit discriminator
  - only considered if AS paths identical
- external routes
- best IGP metric to next hop
- highest IP address

Policy Control

- distribute list
  - filter individual networks
- filter list
  - filter by AS path
- route maps
  - general policy control and tuning

More information

Technical information on BGP

- RFC-1772
  - application of the Border Gateway Protocol
- RFC-1771
  - BGP-4 protocol reference document
- RFC-1745
  - BGP <-> OSPF interaction

Building an Internet

Putting it all together

General philosophy

- Your network is going to grow at an exponential rate!
- Design to scale...but be prepared to reorganize from scratch
- Don't be afraid of change!
  - Most network redesigns are only configuration changes

Putting it all together

- Requirements for IGPs for backbones
- IGP connects your backbone together, not your client's routes
- Must
  - converge quickly
- Should
  - carry netmask information

Putting it all together

connecting to a customer

- static routes
  - you control directly
  - no route flaps
  - no packets to be charged
- shared routing protocol or leaking...
  - you MUST filter your customers info
  - route flaps
- BGP for multi homed customers

Putting it all together

building your backbone

- keep it simple
- redundancy is good, but expensive
- use an IGP that carrys mask information
- use an IGP that converges quickly
- use OSPF, ISIS, or EIGRP

Putting it all together

connecting to other ISPs

- Use BGP-4
- advertise only what you serve
- take back as little as you can

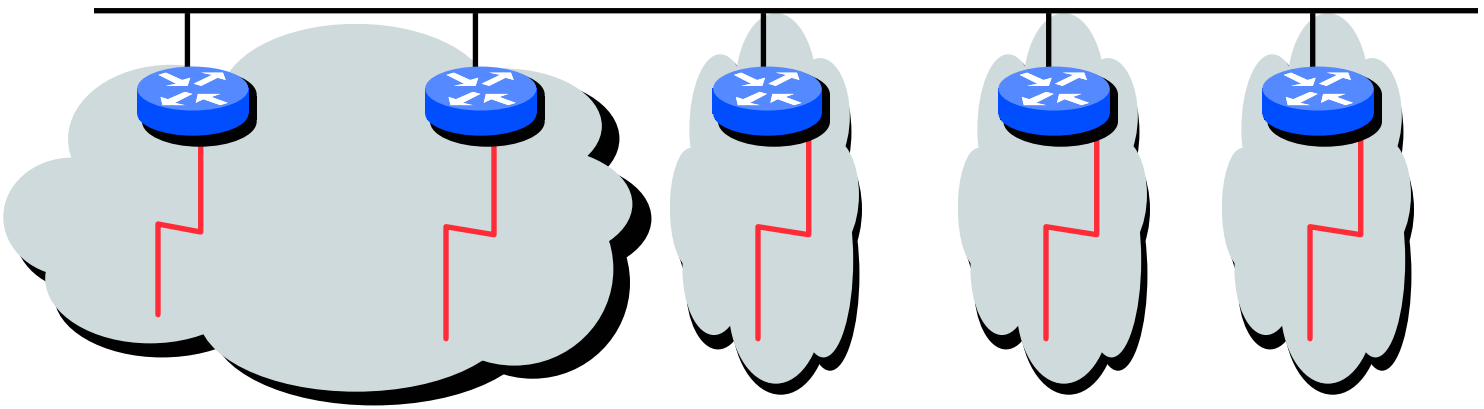
Putting it all together

the internet exchange

- long distance connectivity is expensive
- connect to several providers at a single point

Internet exchanges - FIX

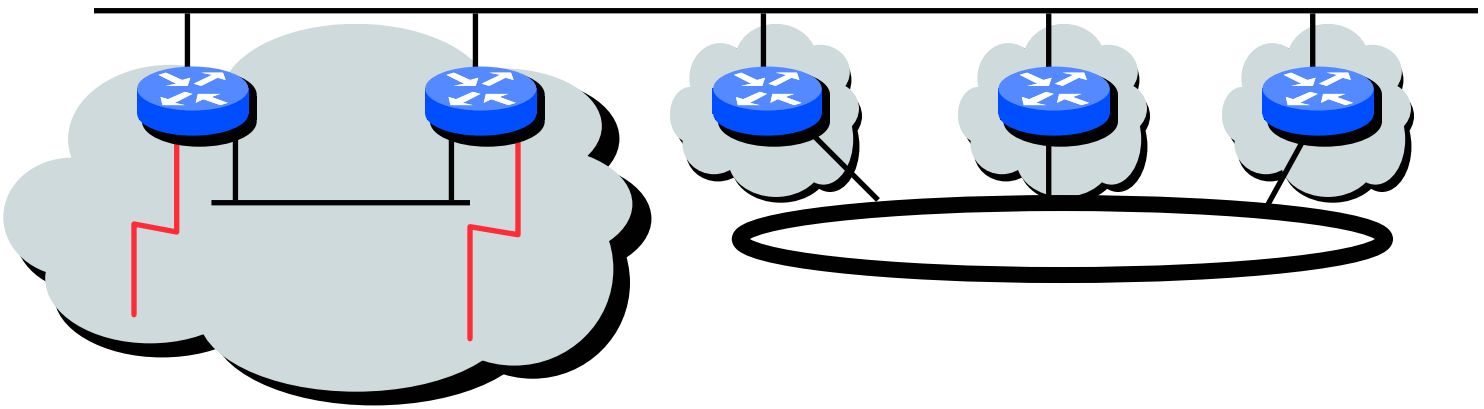
Federal internet exchange (historical)



- dumb ethernet connecting a group of service providers

Internet exchanges - FIX

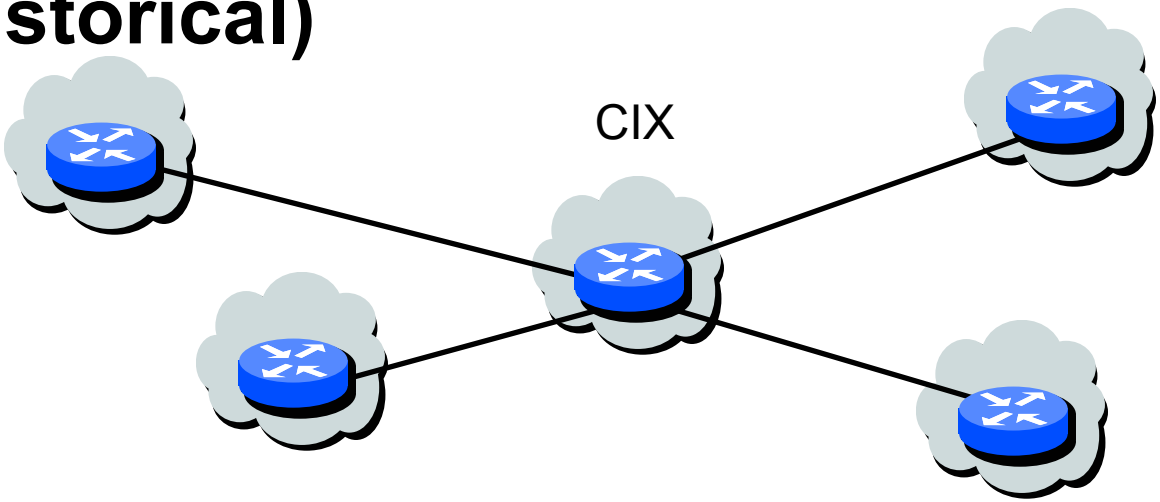
Federal internet exchange



- single primary media all systems share
- secondary media may be shared by a subset of systems to reduce load on primary media

Non-Internet exchange - CIX

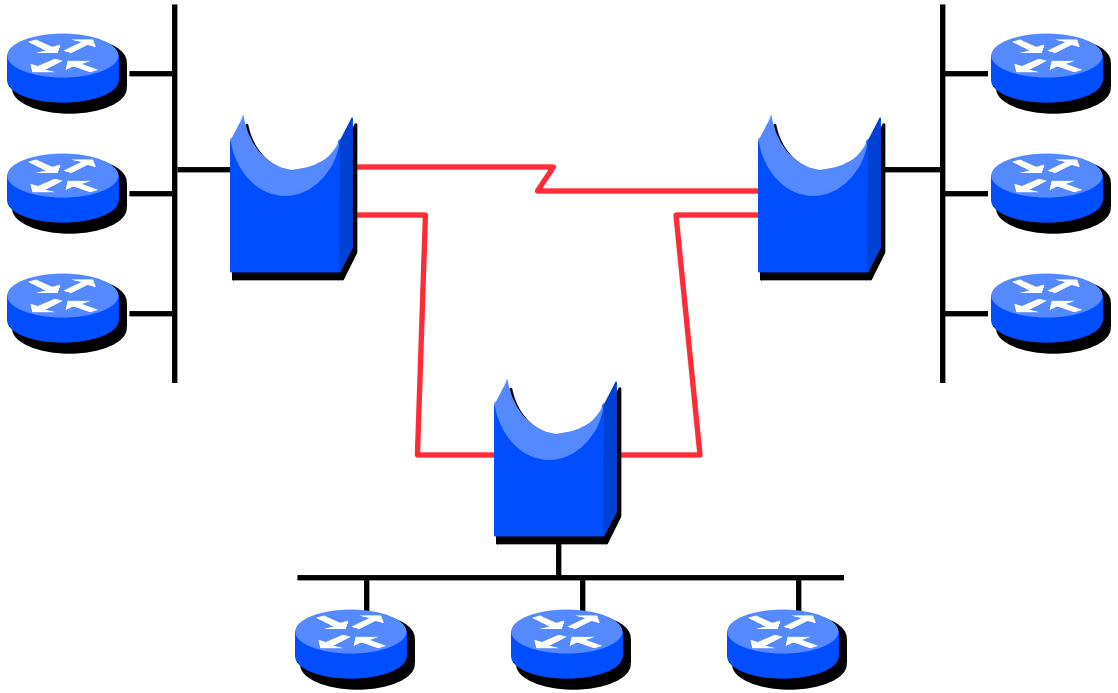
Commercial internet exchange (historical)



- actually a one-router transit AS
- CIX clients only receive best path as determined by CIX router

Internet exchanges - d-GIX

Distributed global internet exchange



- emulates a single ethernet

Internet exchanges - d-GIX

Distributed global internet exchange

- share the cost of high speed lines
- single virtual level-2 media
  - bridges, not routers, connect the link access points
    - bridge table entries are static
    - don't need spanning tree
    - mac address filtering used

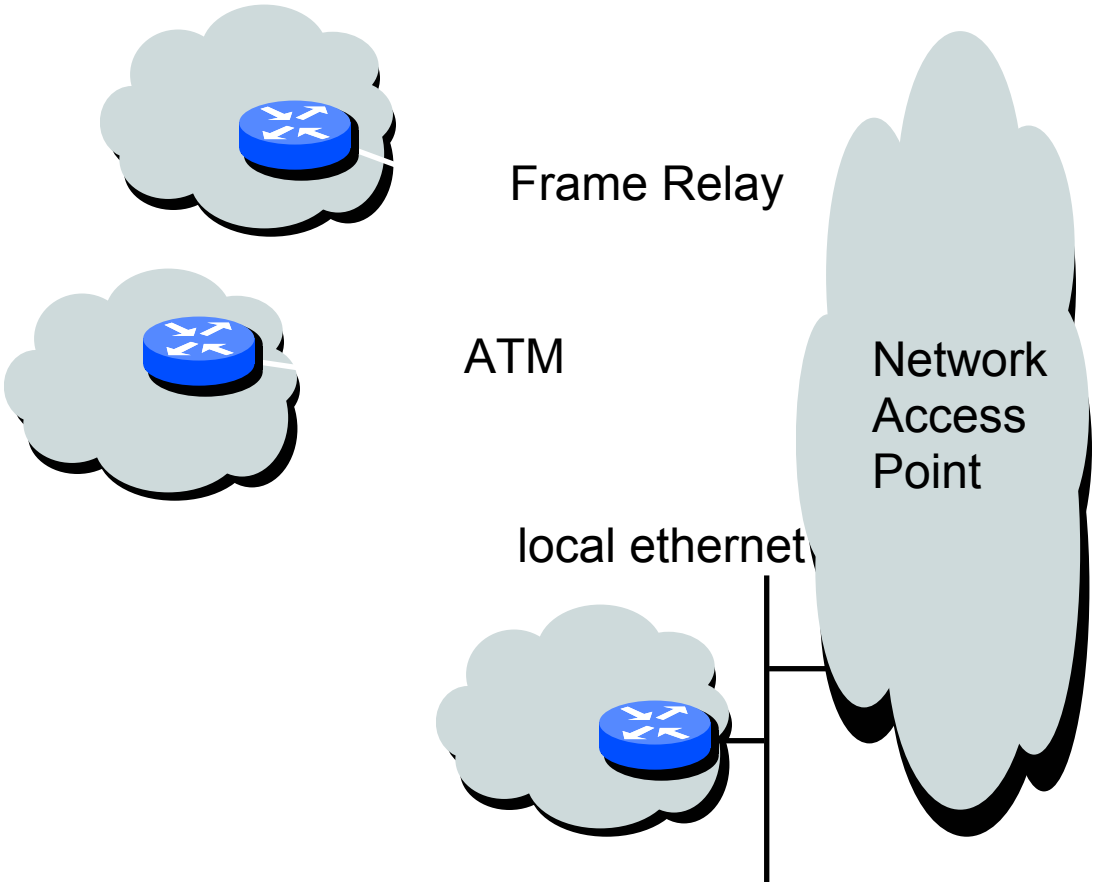
Internet exchanges - d-GIX

Distributed global internet exchange

- the GIX itself still has no routing policy
- in that case, how do you pay for it?
- the GIX does have connectivity policy
  - charge for MAC address filters (source/destination filtering)

Internet exchanges - multi-NAP

Multiple-media network access point



Internet exchanges - multi-NAP

Multiple-media network access point

- Problem:
  - How do you allow one NAP client to connect via Frame Relay and another customer connect via ATM?
- Answer:
  - Don't do this! Extend the NAP and keep it policy free.

Interenet exchanges - multi-NAP

Multiple-media network access point

- **NAPs and IXs need to be policy free**
- **Routers implicity have an 'advertise only what you use' policy.**
- **If routers are used, NAP becomes a transit AS, not an "IX," and clients of the NAP are limited by the NAP's route selection policy.**

More information

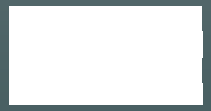
- Original GIX proposal
  - ftp://ftp.ripe.net/ripe/docs/ripe-082.ps
  - ftp://ftp.ripe.net/ripe/drafts/gix15jun.txt
- d-GIX - distributed global internet exchange
  - ftp://ftp.ripe.net/ripe/drafts/d-gix-proposal.ps

Routing registries

What are they?

- database containing
  - route prefix/ origin autonomous system
  - autonomous system/ connectivity policy
- RIPE-181 aka RC-1786

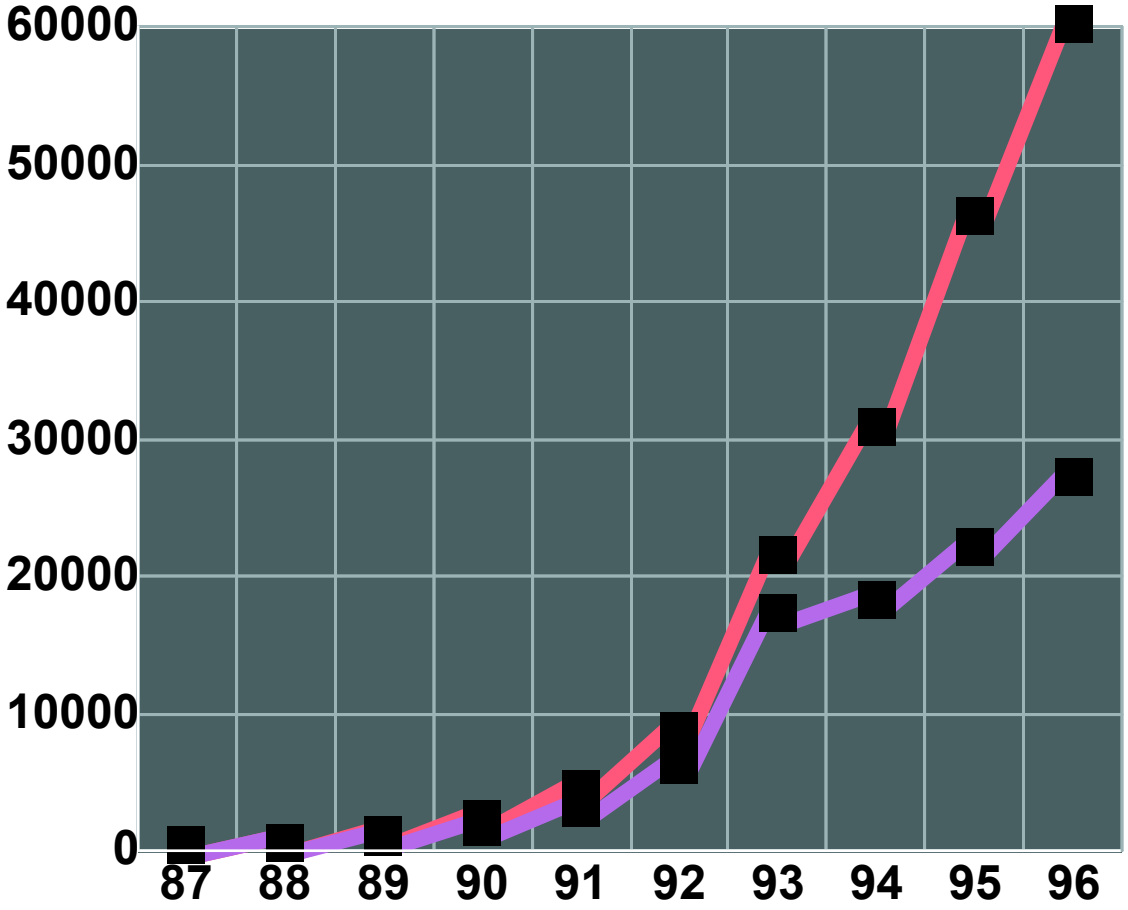
Classless routing



Why CIDR?

- IP route advertisements have been growing exponentially.
- Class A networks are too big
- Class C networks are too small
- Only 65534 class B networks available

Routing Table Growth



Why CIDR?

Classful networks mis-sized

- Class A networks are too big
  - not desirable because of connectivity constraints
- Class B address space is depleted
- Class C networks are useful only for small customers
  - large gap between "C" customer and "B" customer

Classless routing

CIDR at the service provider level

- Service provider given CIDR blocks by numbering authority
- Example:
  - 198.24/15 == 512 class "C" nets
- Service provider advertises only a summary route for CIDR block to neighboring providers, not 512 separate class "C" routes.

Classless routing

The client interface

- Partition local CIDR block and assign to customers
- Example:
  - 198.24.62/23 == 2 "C" nets
  - 198.24.192/18 == 64 "C" nets
  - 198.24.61/24 == 1 "C" net

Classless routing

Do's and don'ts

- Don't assign blocks smaller than class "C" sized networks without prior agreement from customers
  - most hosts & routing protocols are not classless
- Do help customers use their address space wisely!

Classless routing

Do's and don'ts

- Do give customers enough address space for what they need
- Do parition your CIDR block to provide for customer growth
  - get the tree program
  - understand RFCs 1519 and 1219

Classless routing

Do's and don'ts

- Don't be afraid of "holes" when aggregating
- Longest match routing means "he who has the longest prefix wins"

Classless routing

Getting the most out of your allocation

- It's natural, but inefficent to subnet on 8 bit boundaries
  - 131.108.1 = subnet 1
  - 131.108.2 = subnet 2
  - 131.108.3 = subnet 3
- 254 subnets with up to 254 hosts per subnet out of a 16 bit address allocation

Classless routing

There are NO NETWORK NUMBERS!!!

- ...just address space prefixes
  - 131/8
  - 131.0/12
  - 131.108/16
  - 131.108.5/24
  - 131.108.5.32/29
  - 131.108.5.33/32

Classless routing

There are NO SUBNET MASKS!!!

- It's no longer a mask, just a prefix length
- There can be no '0' holes in the mask
- /16 = 255.255.0.0
- /32 = 255.255.255.255
- /14 = 255.252.0.0
- /0 = default = 0.0.0.0

Classless routing

Getting the most out of your allocation

- Unnumbered serial links
- Variable length subnet masks
- Small ethernet
  - 28 bit mask = 14 hosts
- Larger ethernet
  - 26 bit mask = 62 hosts
- VLSM allocation rules are the same as CIDR allocation

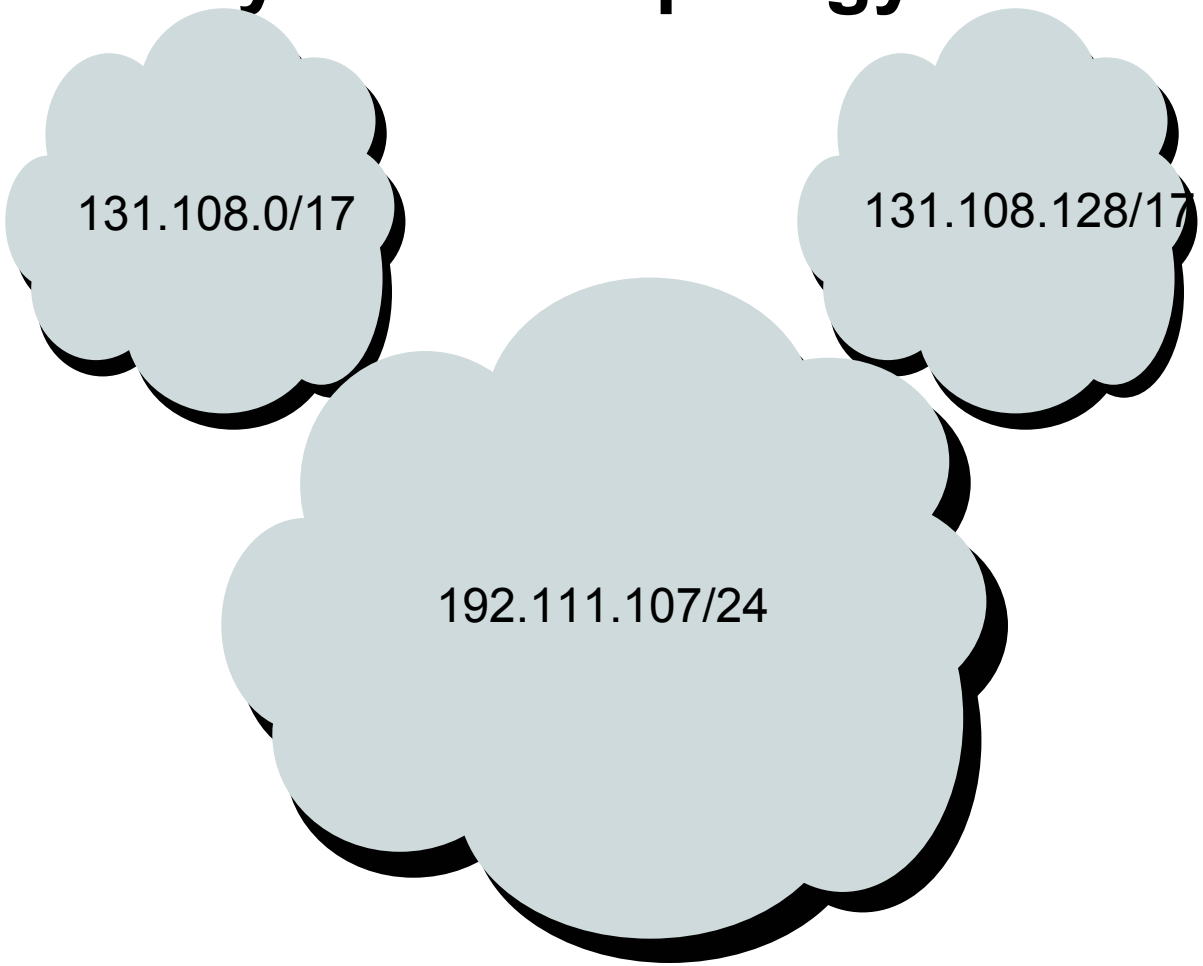
Classless routing

restrictions removed

- no such thing as a "subnet" anymore
  - subnet 0 is no longer special
  - all 1's subnet is no longer special
  - no such thing as a disconnected subnet

Classless routing

Mickey Mouse topology is OK



Classless routing

Plan for entropy

- What is your policy when customers move to a different service provider?
  - do you own the numbers in the CIDR block?
  - will new service provider supply more specific routing information?

Classless routing

Allocate addresses efficiently!

- you don't get very many
- what happens as organizations grow?
- what happens when your customers lie to you?

More information

Technical information on classless routing

- RFCs 1517, 1518, and 1519
  - address assignment and aggregation strategy
- RFC1219
  - assignment of subnet numbers
- `ftp://ftp.sesqui.net/pub/tools/tree.tar`
  - program to help calculate address assignment

More information

Technical information on address allocation

- RIPE NCC address allocation guidelines