COMP312-09A Communications and Systems Software

Routing 3 – BGP

Richard Nelson

richardn@cs.waikato.ac.nz

Exterior Routing: BGP

- CIDR
- Autonomous Systems
- Interior vs. Exterior Routing
- BGP

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Classless InterDomain Routing

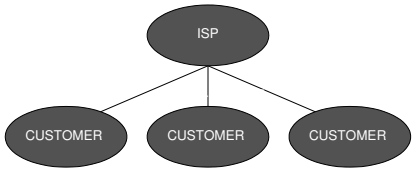
- Appropriate-sized address blocks are allocated to organisations.
- Organisations assign blocks of space to parts of their networks, e.g.
  - 130.217.0.0/25 – 126 host addresses
  - 130.217.2/23 – 510 host addresses
- Organisations advertise only aggregated blocks, e.g. 130.216/16

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Classless InterDomain Routing

- In the same way, ISP's assign “provider aggregatable” addresses from within their allocations to customers and advertise only their aggregate blocks.

e.g.



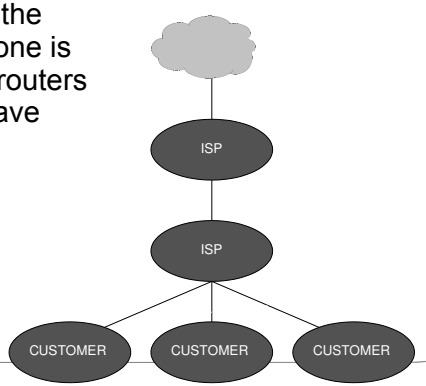
THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Why Aggregate?

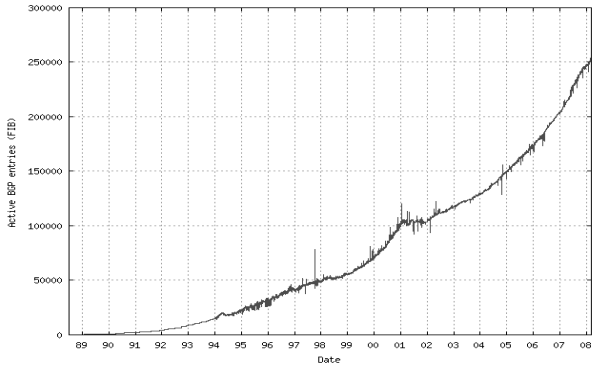
- To keep down the number of routes and volume of updates a default-free router needs to hold.
  - Router memory and CPU
  - Route stability - and so a more stable network
  - Prompt processing of important route updates

The Default-Free Zone

- In the Internet, the Default-Free Zone is made up of all routers which do not have default routes.



"Full" BGP Table Size http://www.cidr-report.org/

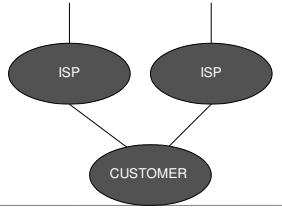


"Full" BGP Table

- Exactly which routes a router sees will vary a little with where in the Internet it is, but a router in the default-free zone today will be working with around 300 000 routes.

Classless InterDomain Routing

- Didn't work as well as hoped:
  - Need to renumber on changing ISP's
  - Multihoming
  - ISP's not as careful as they should be



CIDR (Very Bad) Example

```
Rank AS            AS Name                                    Current  Wthdw  Aggte  Annce Redctn       %
  2 AS9498         BBIL-AP BHARTI BT INTERNET LTD.              1194   1094     22    122   1072  89.78%
Prefix               AS Path                          Aggregation Suggestion
$8.2.236.0/23        12654 7018 9498
$9.144.0.0/15        12654 7018 9498
$9.144.0.0/19        12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.0.0/20        12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.0.0/21        12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.8.0/21        12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.8.0/24        12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.16.0/20       12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.27.0/24       12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.32.0/24       12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.40.0/24       12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.46.0/23       12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.47.0/24       12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.49.0/24       12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.51.0/24       12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.52.0/24       12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.57.0/24       12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.59.0/24       12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.83.0/24       12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
$9.144.84.0/24       12654 7018 9498      - Withdrawn - matching aggregate 59.144.0.0/15 12654 7018 9498
```


Exterior Routing: BGP

- CIDR
- Autonomous Systems
- Interior vs. Exterior Routing
- BGP

Autonomous Systems

An autonomous system is the basic unit in exterior routing.

“The classic definition of an Autonomous System is a set of routers under a single technical administration, using an interior gateway protocol and common metrics to route packets within the AS, and using an exterior gateway protocol to route packets to other ASes.

Since this classic definition was developed, it has become common for a single AS to use several interior gateway protocols and sometimes several sets of metrics within an AS. The use of the term Autonomous System here stresses the fact that, even when multiple IGPs and metrics are used, the administration of an AS appears to other ASes to have a single coherent interior routing plan and presents a consistent picture of what networks are reachable through it.

To rephrase succinctly:

An AS is a connected group of one or more IP prefixes run by one or more network operators which has a SINGLE and CLEARLY DEFINED routing policy.”

- RFC1930

AS Numbers

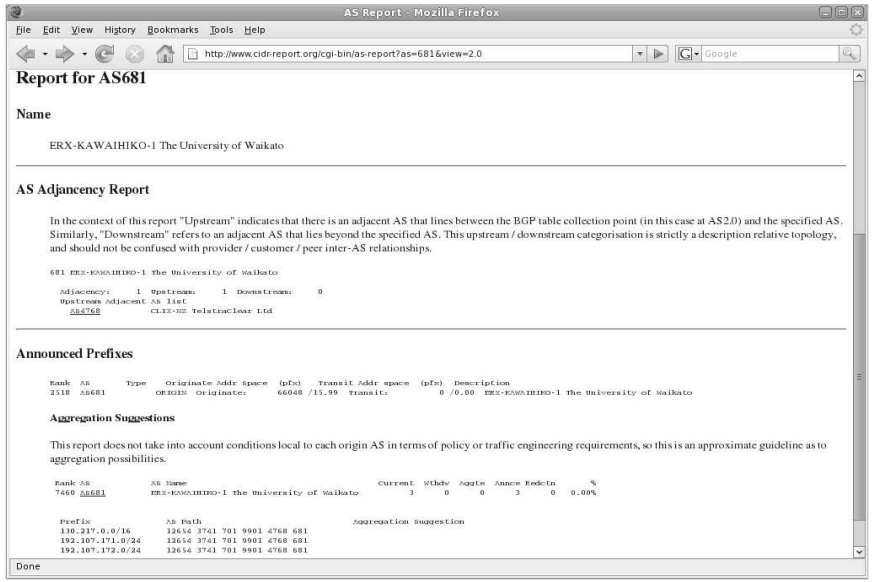
An Autonomous System is identified by an AS number.

- Issued by the same RIR's who issue IP addresses (see BGP pp. 70-71)
- Originally two bytes, now four-byte numbers are available.
- Numbers 64512 to 65535 are “private” (Just like “private” address space)
- AS numbers are used by the BGP protocol, and are expected to be used by other exterior protocols.

AS Numbers

e.g.

- 681 The University of Waikato
- 4763 TelstraClear
- 4684 Telecom NZ Netgate
- 4771 Telecom NZ
- 9325 Xtra
- 9431 Auckland University
- 9439 Wellington Internet Exchange (WIX)
- 9560 Auckland Peering Exchange (APE)

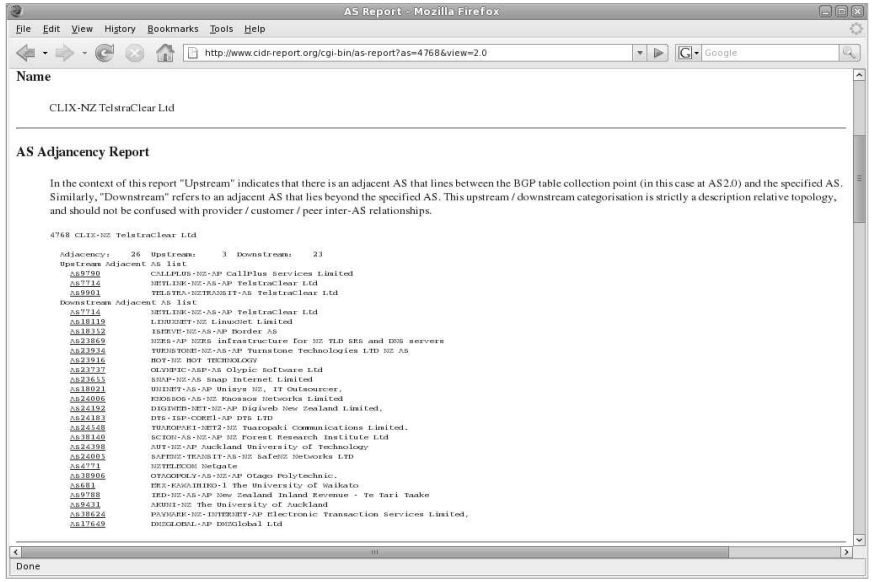


Aggregation Reminder

Why is The University of Waikato advertising all of

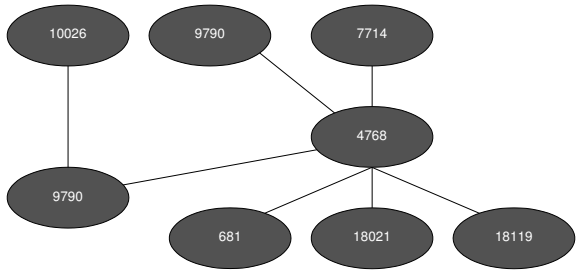
- 130.217.0.0/16
- 192.107.171.0/24
- 192.107.172.0/24

?



Autonomous Systems

An autonomous system is the basic unit in exterior routing.



THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Autonomous Systems

An AS may be a

- Stub AS – communicates with only one other AS.
- Multihomed AS – communicates with more than one other AS, but only passes its own traffic.
- Transit AS – communicates with more than one other AS and will pass through traffic from one outside AS to another.

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Exterior Routing: BGP

- CIDR
- Autonomous Systems
- Interior vs. Exterior Routing
- BGP

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Interior vs. Exterior Routing

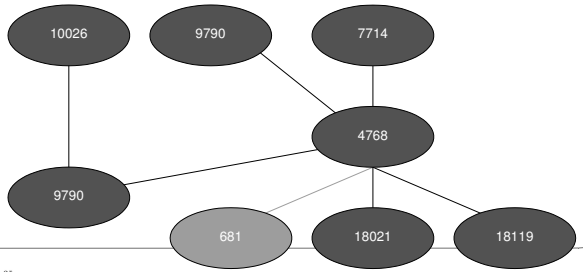
- In any routing system, there's a tradeoff between stability and low cost on one hand and amount of information on the other. An Exterior Gateway Protocol (EGP) hides most of what's inside an AS.
- An EGP is used to dictate how traffic flows between companies. That means money, so must allow more administrative control, so that business considerations can be made to drive network behaviour.
- The Internet's EGP must cope with BIG numbers of routes.

Interior vs. Exterior Routing

- No auto-discovery of neighbours – the administrators decide who we'll exchange routes with.
- Associated with each route are a series of attributes which allow (possibly quite complex) routing policies to be applied.
  - e.g. Favour routes advertised by “peers” over those provided by “transit” provider(s).

Exterior Routing

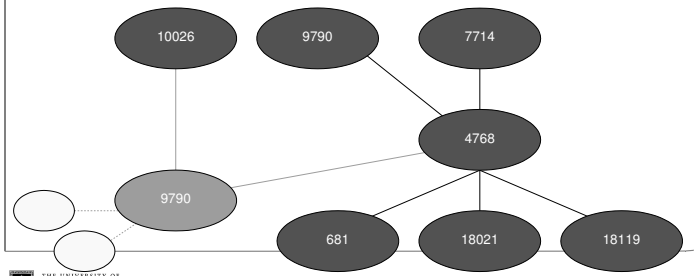
A stub AS doesn't need an EGP. An organisation which connects to only one other organisation (such as its telco) could use static routing or an IGP.



Exterior Routing

An AS does need an EGP if it

- Uses multihoming
- Carries routing information to and from its customers.



Exterior Routing: BGP

- CIDR
- Autonomous Systems
- Interior vs. Exterior Routing
- BGP

BGP

- BGP version 4 is the only EGP in use on the Internet.
  - Everybody has to speak to everybody else
- Standard protocol – RFC1771 and extensions.
- Multi-protocol – uses AS's for routing and can provide a path for any protocol (Ipv4-IPv6)
- Primary metric is AS hop count.
- Many additional attributes of paths to affect route choosing.

BGP Algorithm

- What routing algorithm do you use for routing between ten thousand autonomous systems (and more)?
- Distance vector has convergence problems
- Link state scales to (maybe) 1000 nodes, but not 10000
- No clear hierarchy in the Internet
- Need to know full paths for applying policy

Path Vector Routing

- Each route has an AS path attached. This is empty if the route originated inside this AS.
- On advertising a route to another AS, this AS's number is prepended to the path. e.g. A route that comes into AS 9901 with the path “4768 681” will be readvertised with the path “9901 4768 681”.
- A route whose path already contains this AS number will be discarded. This is how BGP prevents routing loops.
- BGP routers readvertise only routes they have selected to install in their routing table.

Path Vector Routing

- Each route is a network prefix with an AS path attached, e.g.

Prefix	AS Path
130.217.0.0/16	12654 3741 701 9901 4768 681
192.107.171.0/24	12654 3741 701 9901 4768 681
192.107.172.0/24	12654 3741 701 9901 4768 681

- Many other attributes are carried so that route selection policy can be applied.
- Any attribute of a route (including the AS path) can be rewritten by policy.

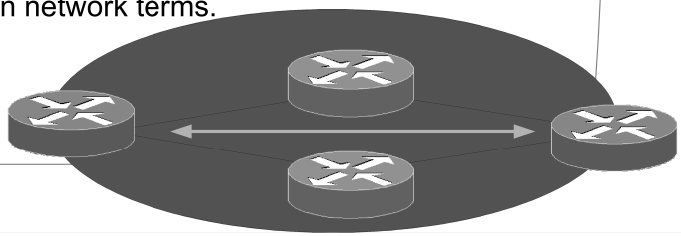
BGP Peers

- BGP performs no neighbour discovery. Relationships are formed only with configured neighbours.
- First a TCP connection is opened using well-known port 179.
- Then a BGP Open packet is sent in each direction. Open packets are used to negotiate BGP session parameters, like how often a Keepalive packet is to be sent.
- Once a BGP session is “established”, routing information is exchanged.

BGP

Use of TCP means

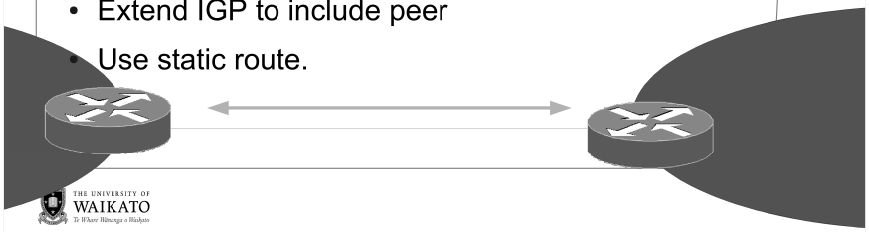
- TCP provides reliable transport, so retransmission not built into BGP
- BGP peers need to be able to exchange TCP packets (which often means need a working IGP between them) but do not need to be near to each other in network terms.



BGP

BGP peers need to be able to exchange TCP packets so they need a route to their peer's address even when it's outside their AS.

- On a broadcast-capable numbered interface (i.e. ethernet), rely on route created by interface configuration (address/netmask).
- Extend IGP to include peer
- Use static route.



BGP Packets

As originally defined, BGP had only four packet types:

- Open – Used in creating sessions. Carries information about the capabilities of the sender.
- Update – Carries routing information.
- Notification – Sent to tear down a BGP session.
- Keepalive – Sent periodically to confirm that this BGP process is still alive. If these are not received by a peer for a long enough period, the peer will send a Notification packet and shut down the session.(Then try periodically to re-establish it.)

BGP Route Attributes

Update packet contains one or more prefixes, then a set of attributes which apply to all of those prefixes. To carry a new kind of routing information, define a new attribute. No need for multiple packet types to carry routes (cf. OSPF).

BGP route attributes include:

- Local Preference – a number. There to allow policy to change how preferred a route is. Not advertised outside this AS.
- AS Path – list of AS's through which this route has passed to reach here.
- Next Hop – IP address to which traffic for addresses in this prefix is to be sent.

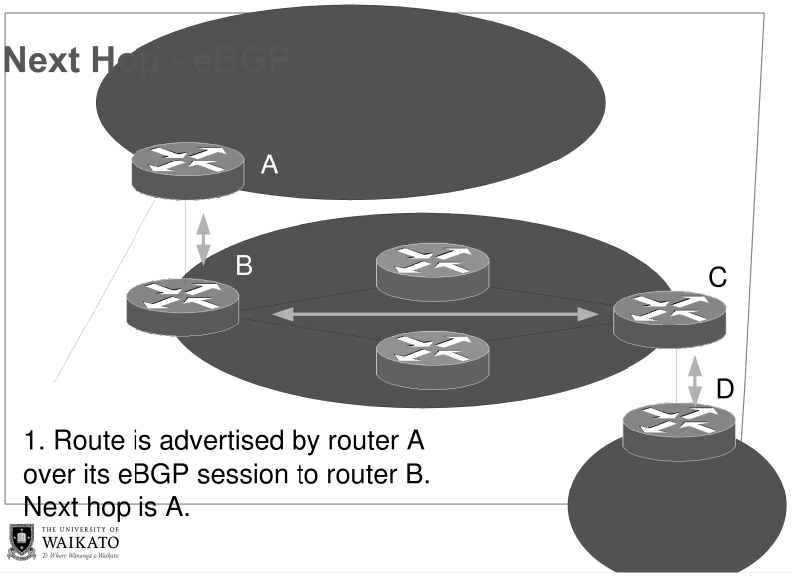
BGP Route Attributes

- Multi Exit Discriminator – A number. Attached to routes sent to neighbouring AS's to tell them which of multiple connections to your AS you prefer them to use. Not re-advertised by neighbouring AS's.
- Community – two or four bytes which could contain anything. Numbers of communities may be added to any route to tag it with information. For example, different communities may be assigned to routes learned at different sites. A common policy is to strip off all communities on routes as they are learned from neighbours.

IBGP vs. eBGP

- IBGP – Internal BGP session, between routers in an AS
- EBGP – External BGP session, between routers in different AS's
- Slightly different rules
  - Add my AS to route's AS path when advertising the route by EBGP.
  - Change the Next Hop to the address I'm using to advertise this route on EBGP.

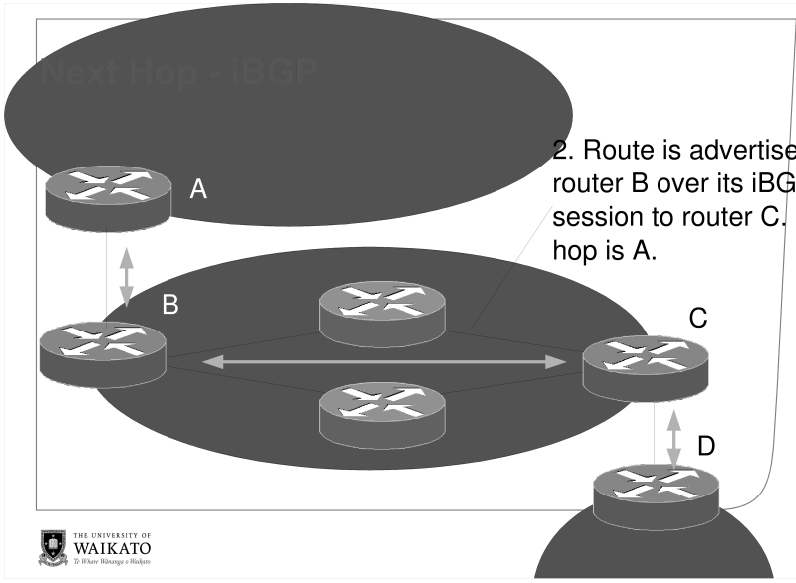
Next Hop - eBGP



1. Route is advertised by router A over its eBGP session to router B. Next hop is A.

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

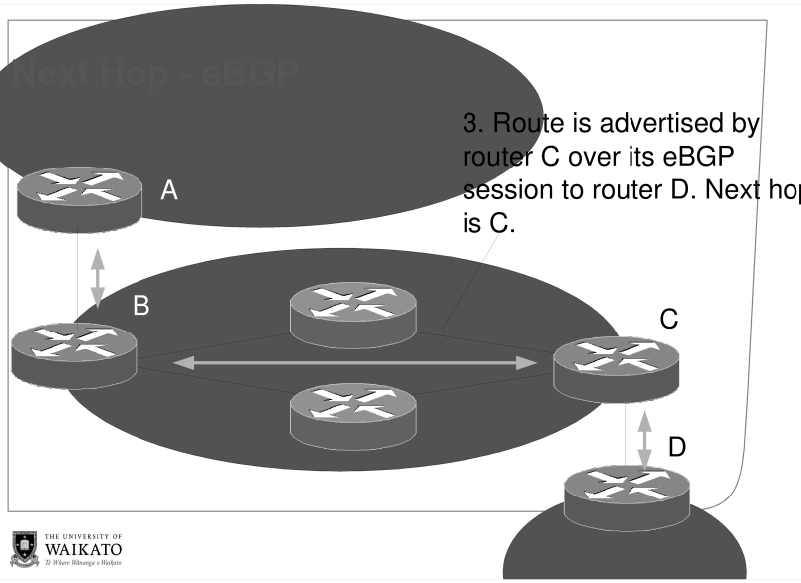
Next Hop - iBGP



2. Route is advertised by router B over its iBGP session to router C. Next hop is A.

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Next Hop - eBGP



3. Route is advertised by router C over its eBGP session to router D. Next hop is C.

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

BGP Route Selection

- Apply any defined policies to change route attributes.
- Reject any BGP route whose Next Hop is not reachable.
- Select the route with the longest prefix match. (Not specific to BGP.)
- Select the route with the highest Local Preference.
- Select the route with the shortest AS Path.
- Among routes received from the same neighbour AS, select the route with the lowest MED.
- Apply the other rules, all the way down to “lowest peer IP address”. No random element, so a route recalculation will not change the route selected without a change of policy or incoming routes.

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Multiprotocol BGP

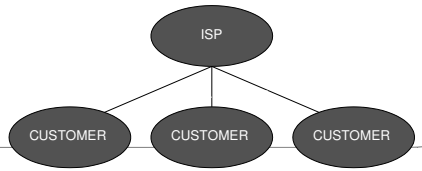
Attributes for a route can include Address Family Identifier and Subsequent Address Family Identifier. (extension in RFC4760) e.g.

- AFI 1, SAFI 1 = IPv4 unicast
- AFI 1, SAFI 2 = IPv4 multicast
- AFI 2, SAFI 1 = IPv6 unicast
- AFI 2, SAFI 2 = IPv6 multicast

Nearly all route attributes are unchanged if the address family used changes. So BGP can carry routing information for multiple protocols at once without change to BGP.

Route Aggregation

Usual practice is to configure a BGP-speaking router to advertise all of the address ranges you or your "downstream" customers have in use. These are or should be advertised in aggregated form, for example the ISP in this example would advertise to its upstream peer 60.234/16, and would not readvertise ranges inside that prefix advertised to the ISP by its customers.



Route Aggregation

- It is usual for a BGP-speaking router to have statically configured on it which routes it is to advertise. This is commonly done by setting the router so that it will advertise a configured range so long as it has routes for some parts of the range. These may be received by IGP.
- A large transit AS may filter the advertisements it accepts, perhaps not accepting any prefix longer than /20.

Routing Incoming Packets

A BGP-speaking router has aggregate route it advertises by BGP, say 130.217/16. Next Hop is to drop the packet.

It also has in its routing table IGP routes, including one for 130.217.4.0/25, which specify next hops inside the AS

A packet comes in with destination 130.217.4.6.

What does the router do?

Choosing Among Protocols

1. Longest match first

2. Given more than one route for the same prefix learned from different protocols, use protocol weighting. (Protocol weightings can be changed.)

Choosing Among Protocols

Cisco Administrative Distance	
Connected Interface	0
Static Route	1
External BGP	20
Internal EIGRP	90
RIP	100
OSPF	110
IS-IS	115
External EIGRP	170
Internal BGP	200

BEST

Juniper Protocol Preference	
Connected Interface	0
Static Route	5
OSPF Internal	10
IS-IS Internal	15,18
RIP	120
OSPF External	150
IS-IS External	160,165
BGP	170

BGP Protocol Summary

- BGP is the only EGP in use in the Internet.
- Static configuration of peer relationships.
- Information carried over TCP sessions.
- Routing governed by policy to reflect business relationships.
- Multiple metrics available in choosing BGP routes.
- “Path vector” routing based in large part on AS paths.
- Over 250 000 routes in a full routing table now, so scaling matters.
- Can carry routing information for multiple address types.

Reading

http://www.cidr-report.org/

Van Beijnum, *BGP*, O'Reilly, 2002 pp. 70-71

Halabi, *Internet Routing Architectures*, Cisco Press, 2000, Chapter 4

Van Beijnum, *BGP*, O'Reilly, 2002 pp. 23-27