Routing

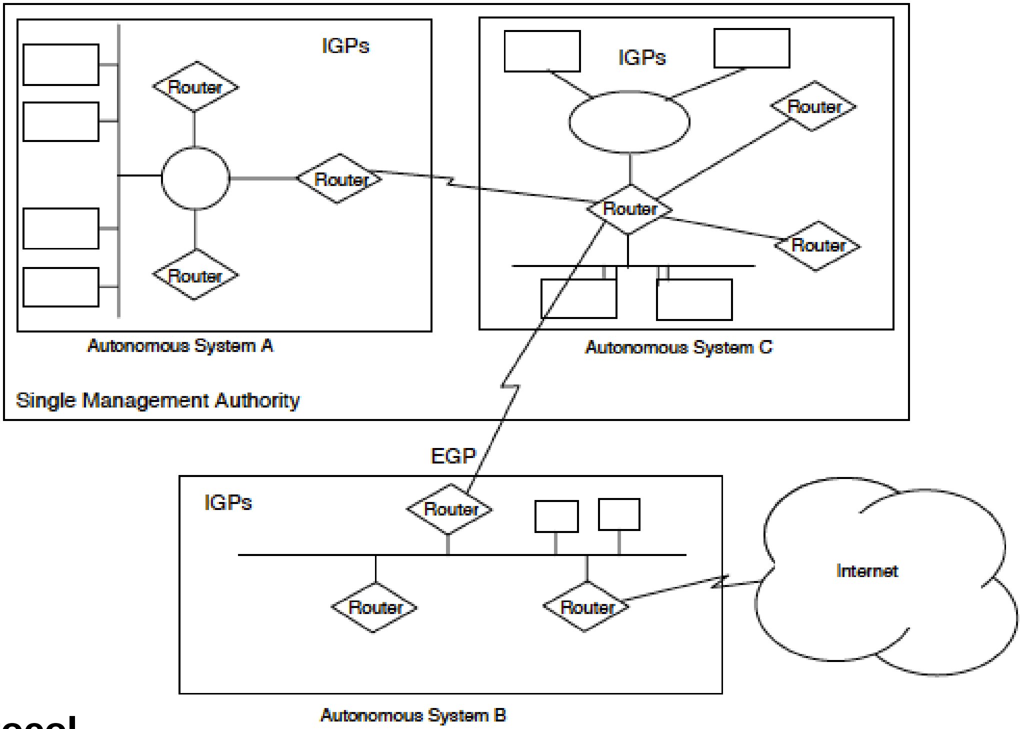
Interdomain – EGP (BGP)

Content

- Introduction to BGP
- BGP terminology, concepts, messages, timers
- Attributes and path selection mechanisms
- Differences between I-BGP and E-BGP
- Challenges
- More on BGP scaling

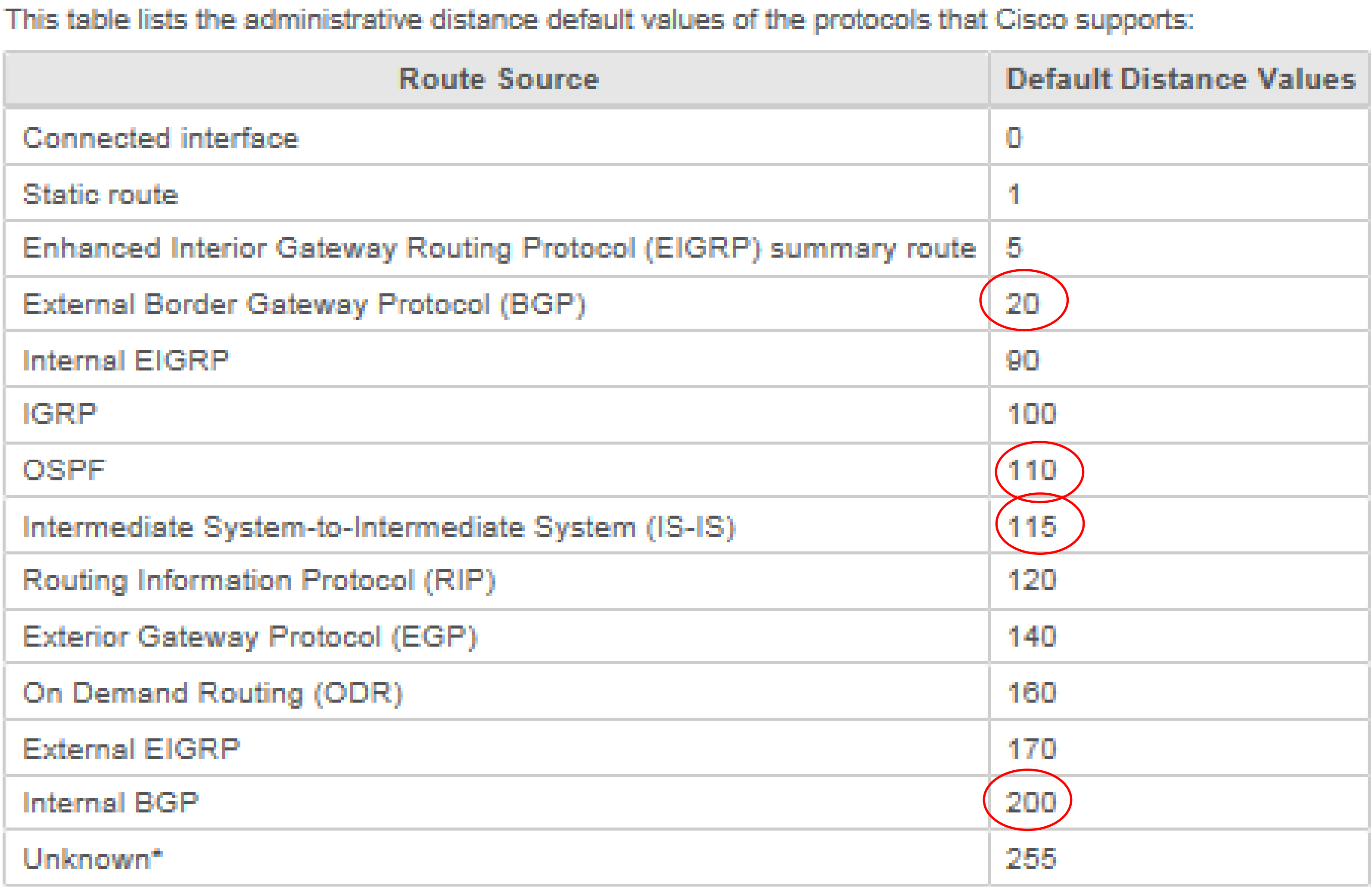
Organization of the Internet – autonomous systems and hierarchical routing

- An **autonomous system (AS)**: a domain (collection of IP networks and routers) under the control of one entity (or sometimes more) that presents a common routing policy to the Internet



- Hierarchical routing
  - **Intra-domain** (interior) routing: within an AS, Interior Gateway Protocol
  - **Inter-domain** (exterior) routing: between AS, Exterior Gateway Protocol
  - Reduces the routing information to be kept within an AS
    - Uses ”standard router” (default router) towards other AS’s
    - Simple for all to know which route to be applied towards the external

Choosing between information provided by different routing protocols

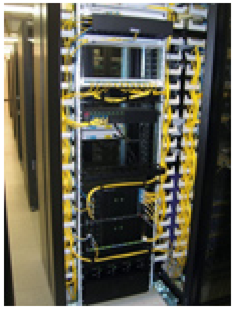


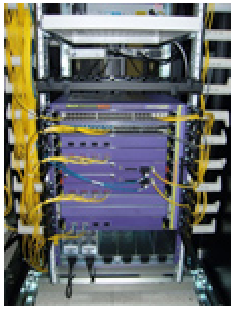
This table lists the administrative distance default values of the protocols that Cisco supports:

Route Source	Default Distance Values
Connected interface	0
Static route	1
Enhanced Interior Gateway Routing Protocol (EIGRP) summary route	5
External Border Gateway Protocol (BGP)	20
Internal EIGRP	90
IGRP	100
OSPF	110
Intermediate System-to-Intermediate System (IS-IS)	115
Routing Information Protocol (RIP)	120
Exterior Gateway Protocol (EGP)	140
On Demand Routing (ODR)	160
External EIGRP	170
Internal BGP	200
Unknown*	255

Peering of networks to interconnect AS's in exchange points or through direct physical peering

- Peering is interconnection of administratively separate internet networks for the purpose of exchanging traffic
  - Peering requires physical interconnection of the networks
    - Internet exchange point is a physical infrastructure that houses servers, routers, switches
  - Exchange of routing information through the BGP routing protocol
  - Peering usually involves a business relationship between the ISPs

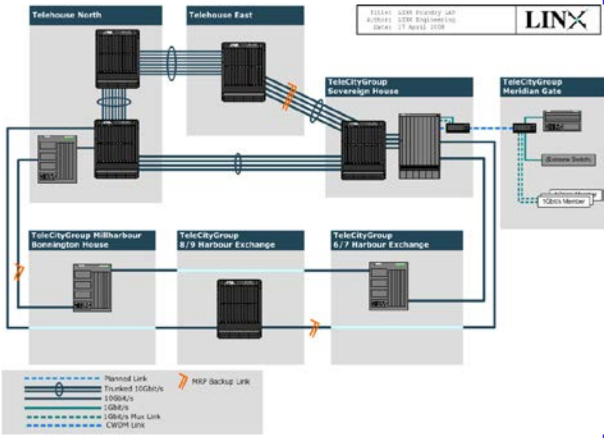




  - List of internet exchanges: http://www.bgp4.as/internet-exchanges
    - AMS-IX: an Amsterdam internet exchange point: http://www.ams-ix.net/
    - LINX: London Internet eXchange https://www.linx.net
    - NIX: Norwegian Internet eXchange http://www.uio.no/nix

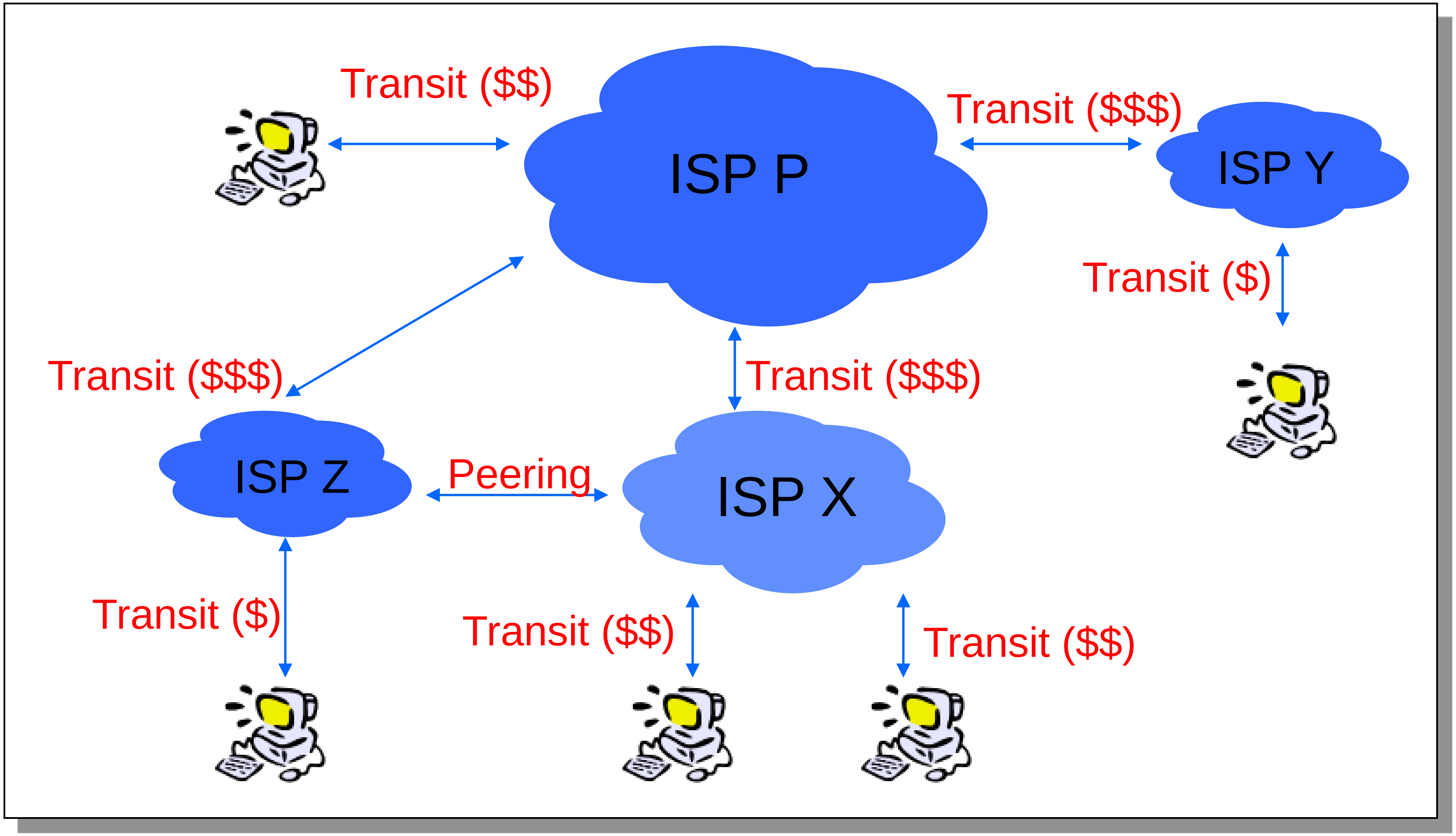
      Info and prices http://www.uio.no/nix/nix-info.html

      Members http://www.uio.no/nix/nix-ops.html


Telehouse North
Telehouse East
LINX
TeleCityGroup Sovereign House
TeleCityGroup Meridian Gate
Extreme Switch
10Gbit/s Member
TeleCityGroup Millharbour Bonnington House
TeleCityGroup 8/9 Harbour Exchange
TeleCityGroup 6/7 Harbour Exchange
Planned Link
Trunked 10Gbit/s
10Gbit/s
1Gbit/s
1Gbit/s Mux Link
CWDM Link
MRP Backup Link

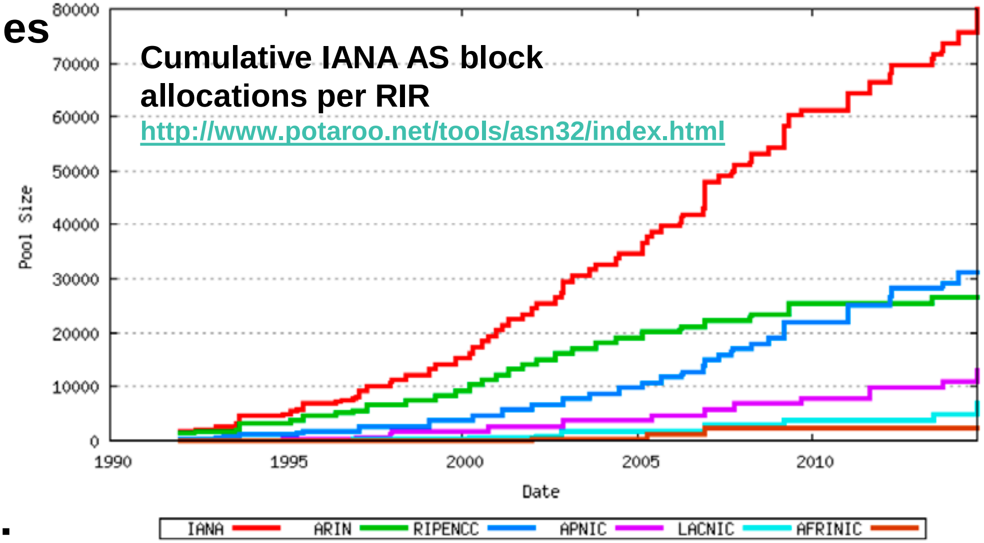
https://www.peeringdb.com/

Not all challenges are technical



Inter-domain route propagation is based on identification of autonomous systems (AS's)

- A unique AS number is allocated to each AS
  - assigned by the same authorities that allocate IP addresses
  - a unique 16/(32)-bit id
  - Eg.: universities (Uninett 224), corporate, backbone-network http://www.ripe.net/perl/whois

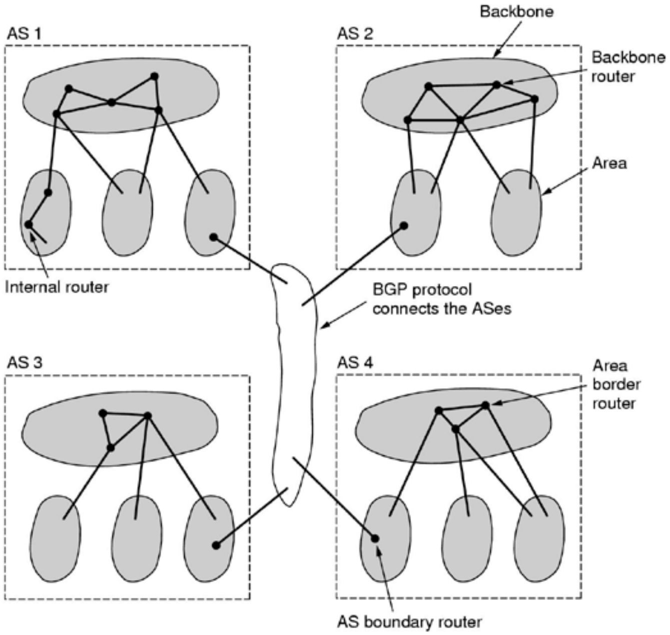


Cumulative IANA AS block allocations per RIR
http://www.potaroo.net/tools/asn32/index.html

- AS is classified in three types:
  - "Stub AS": an AS with only one connection to another AS, such an AS allows only local traffic
  - "Multi-homed AS": an AS with connection to more than one AS, and which does not allow transit traffic
  - "Transit AS": an AS with connection to more than one AS, and which allows both local and transit traffic

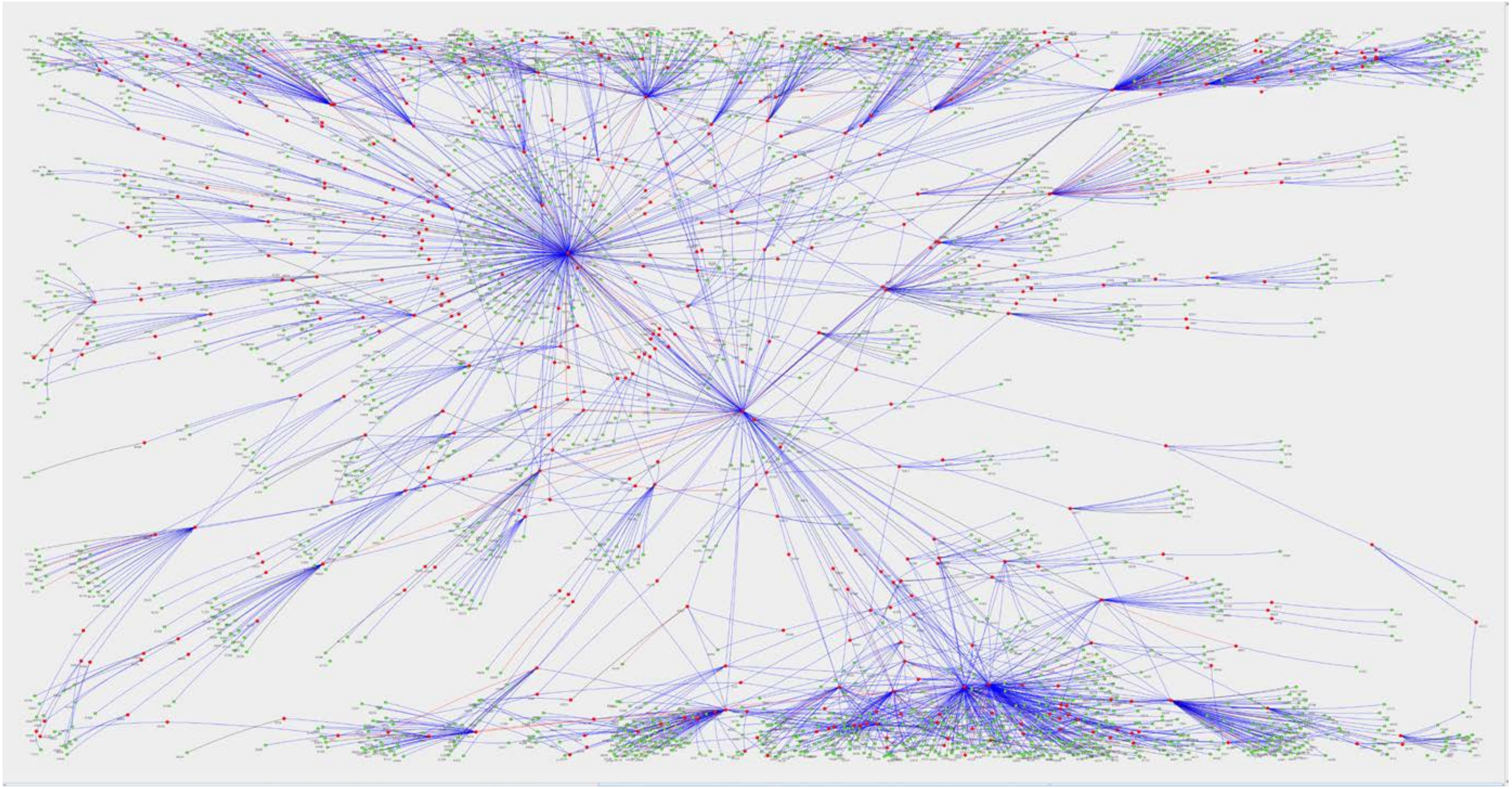
BGP(v4) (Border Gateway Protocol) is THE inter-domain routing protocol

- Primary function: to exchange network reachability information with other BGP systems, http://www.bgp4.as/
  - http://www.itransformers.net/logo/bg_peering.png
- Reachability information = the list of traversed AS's
- Path vector protocol - carries the complete AS path vector (or AS set when want to hide specific routes)
- Constructs a graph of AS connectivity
- Supports CIDR and advertises aggregates of routes as an IP prefix



This is how the “tree of AS’es” looks like

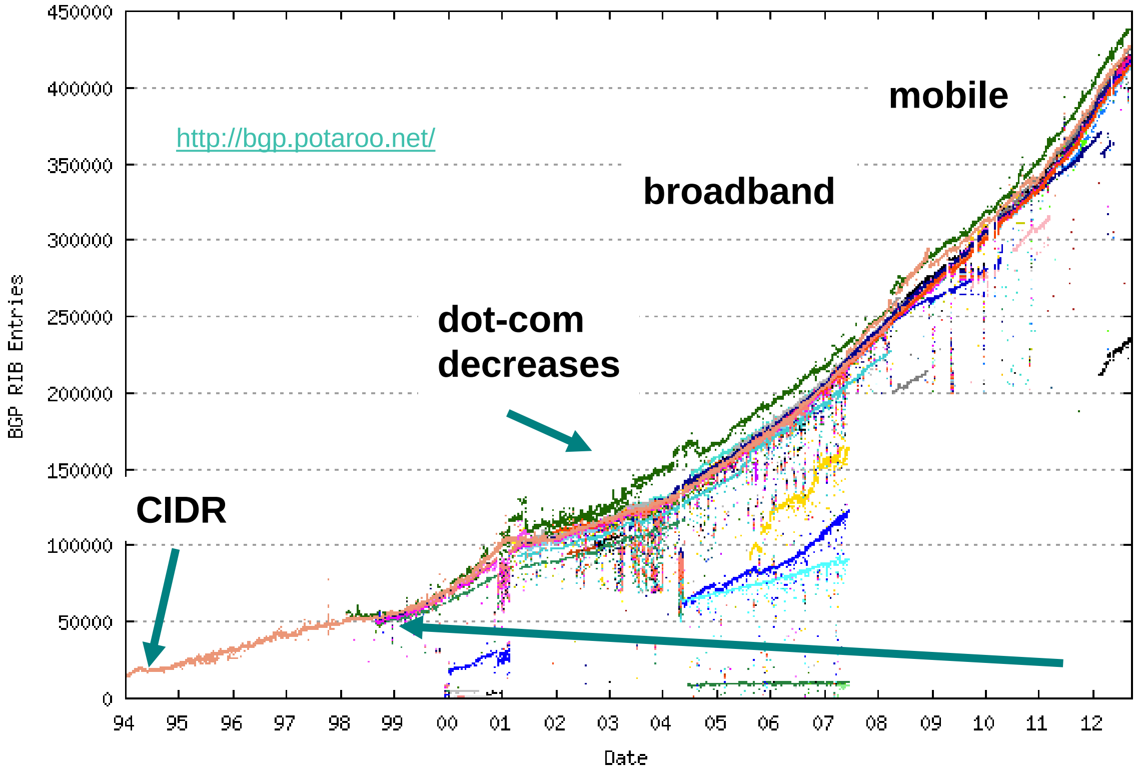
http://www.itransformers.net/logo/bg_peering.png



http://www.ripe.net/perl/whois

BGP routing table size is a metric on the number of distinct routes within Internet

- Growth in routing advertisements >> amount of address space advertised
  - Growth in number of connected devices is behind NAT gateways
  - Discrete exterior routing policy being applied to finer address blocks



http://bgp.potaroo.net/

mobile

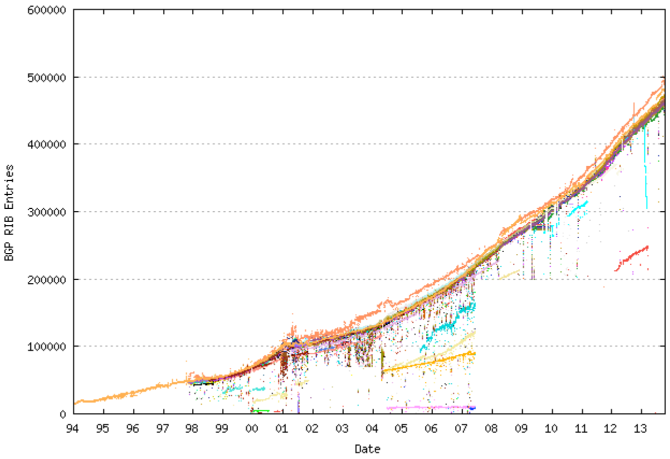
broadband

dot-com decreases

CIDR

CIDR
not able to keep
pace with growth

Routing scalability is essential to a large network and depends on routing protocol and network design

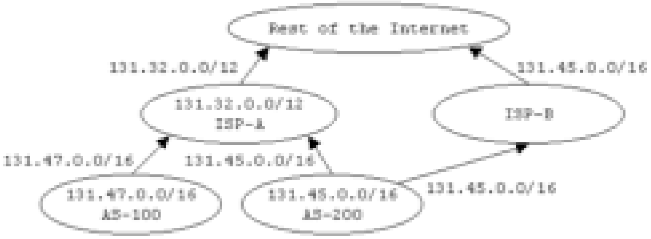


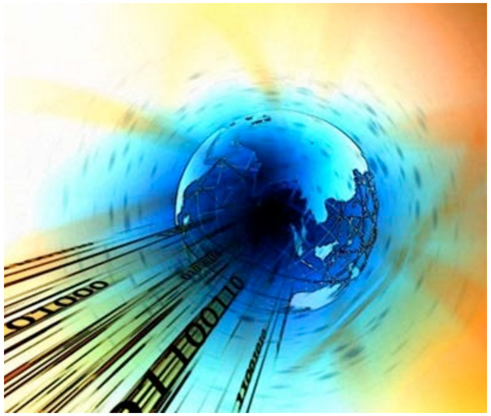
- Today’s large networks: a large number of nodes with rich connectivity
  - Customer aggregation routers connect many customer routers
  - Default-free routers with many routing entries
- Router resource consumption increases
  - Large number of routers and adjacencies
  - Large number of routes with multiple paths
    - Route flapping on link instability
    - Large number of BGP sessions
  - Consequence: Slow routing convergence
- Routing complexity increases
  - Complicated routing policy, prefix-based filtering
  - Consequence: Poor management of network and low service quality

Source: http://bgp.potaroo.net/

The scalability problem

- Provider independent addressing
- Multi-homing for increased reliability
- Multi-homing for traffic engineering
- Countermeasure against prefix hijacking





Scalable routing design principles

- **Build hierarchy**
  - Avoid full mesh overlay topology
  - Core (transit) + region (access)
  - One vs. separate routing domains
- **Compartmentalization**
  - for problem and failure localization
- **Making proper trade-offs**
  - Redundancy vs. scalability
  - Convergence vs. stability
- **Reducing route processing burdens**
  - Routing intelligence placement at edges
  - Reduce routes and routing information: CIDR, default routing, reduce alternative paths
  - Static route at edges
- **Defining scalable routing policies and implementation**

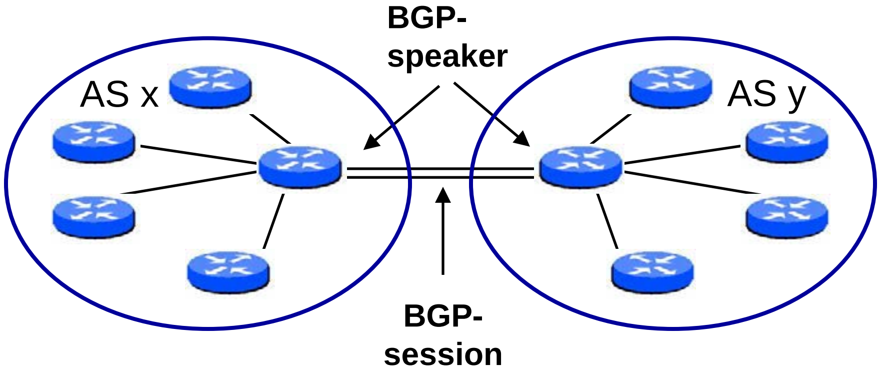
Policy with BGP

- **BGP provides capability for enforcing various policies**
- **Policies are not part of BGP: they are provided to BGP as configuration information**
- **BGP enforces policies by choosing paths from multiple alternatives and controlling advertisement to other AS's**

Content

- Introduction to BGP
- BGP terminology, concepts, messages, timers
- Attributes and path selection mechanisms
- Differences between I-BGP and E-BGP
- Challenges
- More on BGP scaling
-

Each AS is connected to the rest of the Internet through a border router



- **BGP-speaker:**
  announces local networks,
  other reachable networks
  (transit AS) and
  other route information
- **BGP neighbors (peers):**
  pair of BGP speakers exchanging routing information

- **BGP–session: TCP session connecting two BGP neighbors**
  - **Keepalive messages monitors the state of the session**
  - **Need not explicitly conform receipt of BGP message**
  - **No refreshment of the whole routing table, only incremental updates**
  - **Conserve bandwidth and processing power**

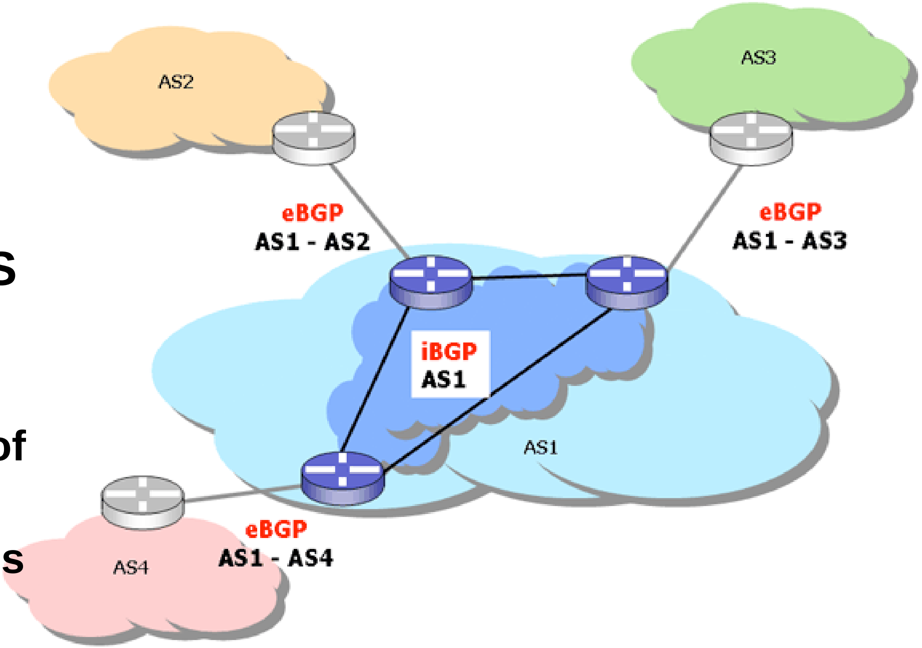
- **Routing policy constrains the flow of data packets through the network**
  - **Decides exchange of routing information between AS’s**
  - **Configured administratively**

Integrated intra-/interdomain routing with interior and exterior BGP

- **Default route is adopted by intra-domain routing protocol**

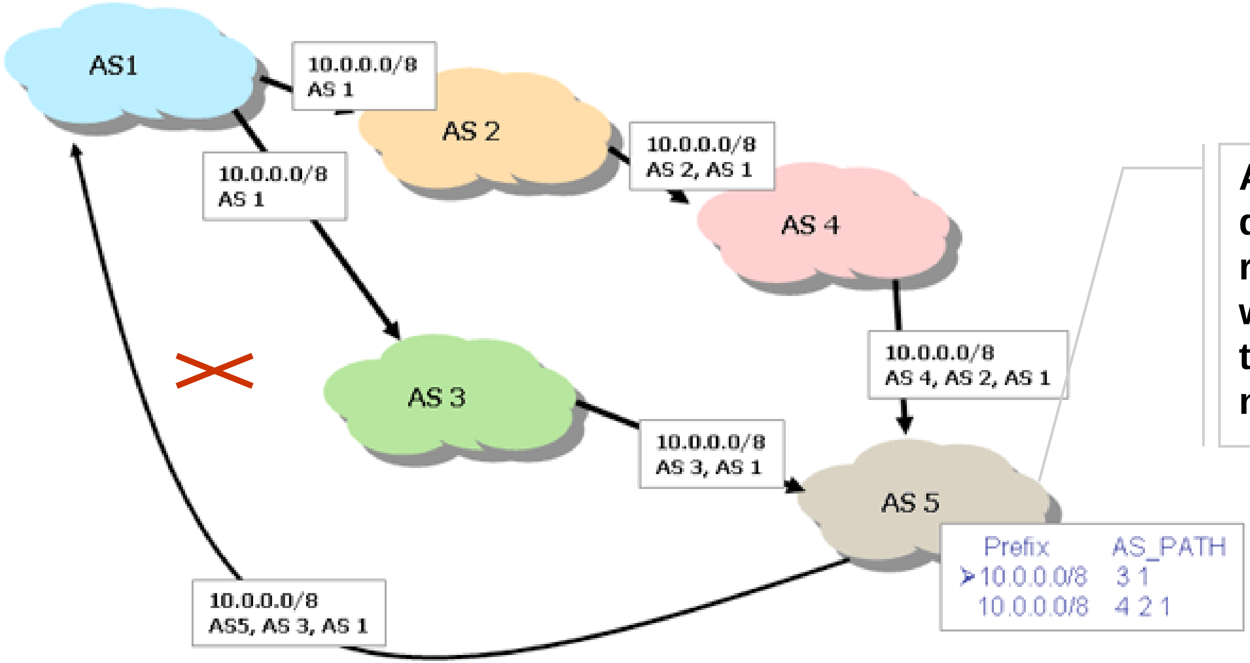
- **”Border” router inject specific routes outside the AS**

- **Interior BGP (IBGP) – between BGP nodes (anywhere) in the AS**
  - **All routes cannot be injected into intra-domain protocol**
  - **IBGP for effective redistribution of BGP information**
  - **Which AS ”border” router reaches specific address**



- **Exterior BGP (EGBP) across AS boundaries, typically directly connected**

Routers speaking BGP exchange AS_PATH vectors



Prefix	AS_PATH
➤10.0.0.0/8	3 1
10.0.0.0/8	4 2 1

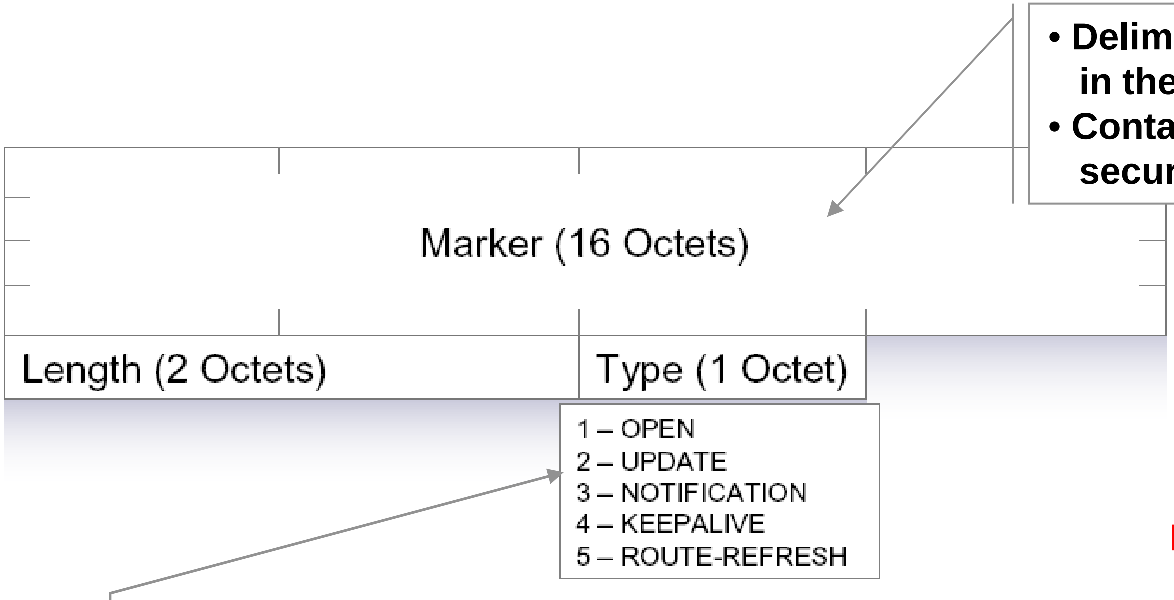
AS5 makes a policy decision as to which route path to accept, which, by default will be the shorter AS path, namely <AS3, AS1>

http://looking-glass.connect.com.au/lg/
http://lg.he.net/

Address prefix reachability information traverses the Internet in the form of individual route objects, this routing information is augmented by the list of AS's that have been traversed thus far, forming the AS_PATH attribute

BGP figures:
http://www.internetsociety.org/publications/isp-column-may-2006-introduction-bgp-%E2%80%93-protocol

BGP common header message format represents 5 message types



Marker (16 Octets)

Length (2 Octets)

Type (1 Octet)

1 – OPEN
2 – UPDATE
3 – NOTIFICATION
4 – KEEPALIVE
5 – ROUTE-REFRESH

- **Delimits BGP messages in the TCP byte stream**
- **Contains all 1’s, unless security option**

BGP uses TCP port 179

- **OPEN:** **start a BGP session**
- **UPDATE:** **exchange reachability information**
- **NOTIFICATION:** **convey a reason code prior to termination of BGP session**
- **KEEPALIVE:** **confirm the continued availability of the BGP peer**
- **ROUTE-REFRESH:** **request messages**

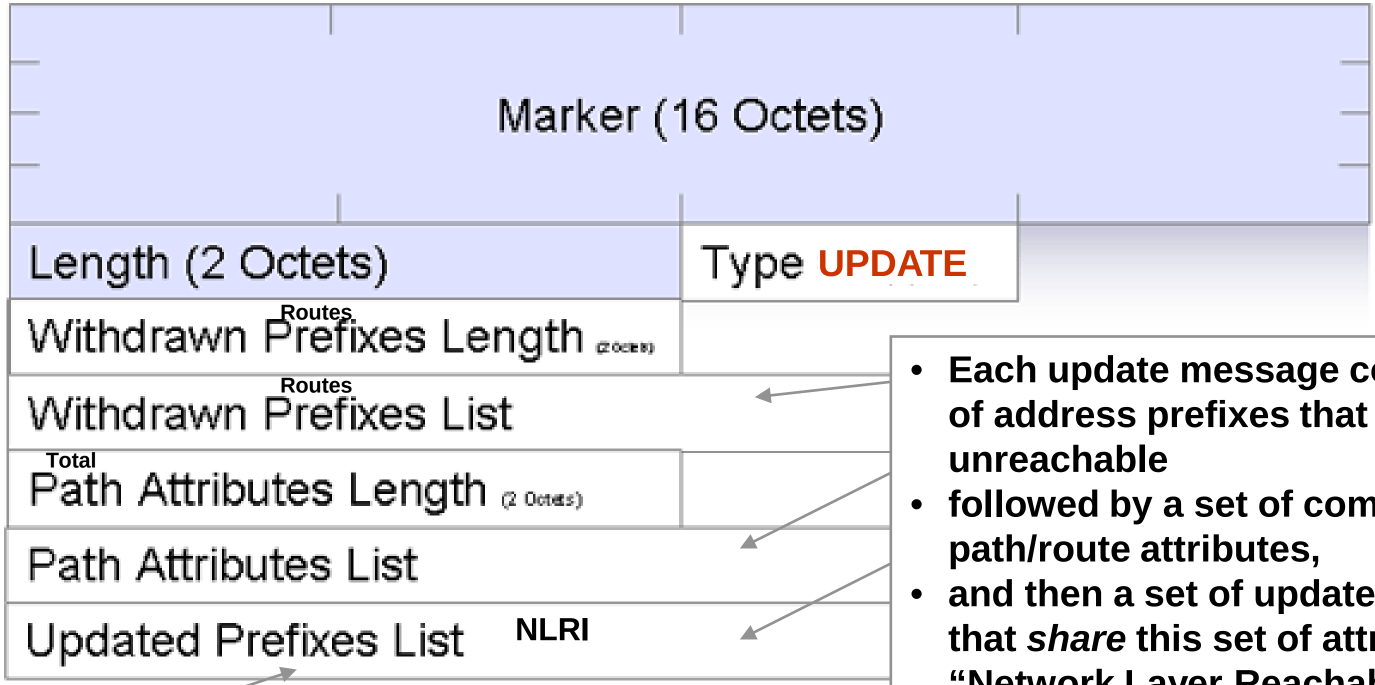
BGP Messages

- Open
  - Announces AS ID
  - Determines hold timer – interval between keep_alive or update messages, zero interval implies no keep_alive
- Keep_alive
  - Sent periodically (but before hold timer expires) to peers to ensure connectivity.
  - Sent in place of an UPDATE message
- Notification
  - Used for error notification
  - TCP connection is closed *immediately* after notification

BGP UPDATE Message

- List of withdrawn routes
- Network layer reachability information
  - List of reachable prefixes
- Path attributes
  - Origin
  - Path
  - Metrics
- All prefixes advertised in message have same path attributes

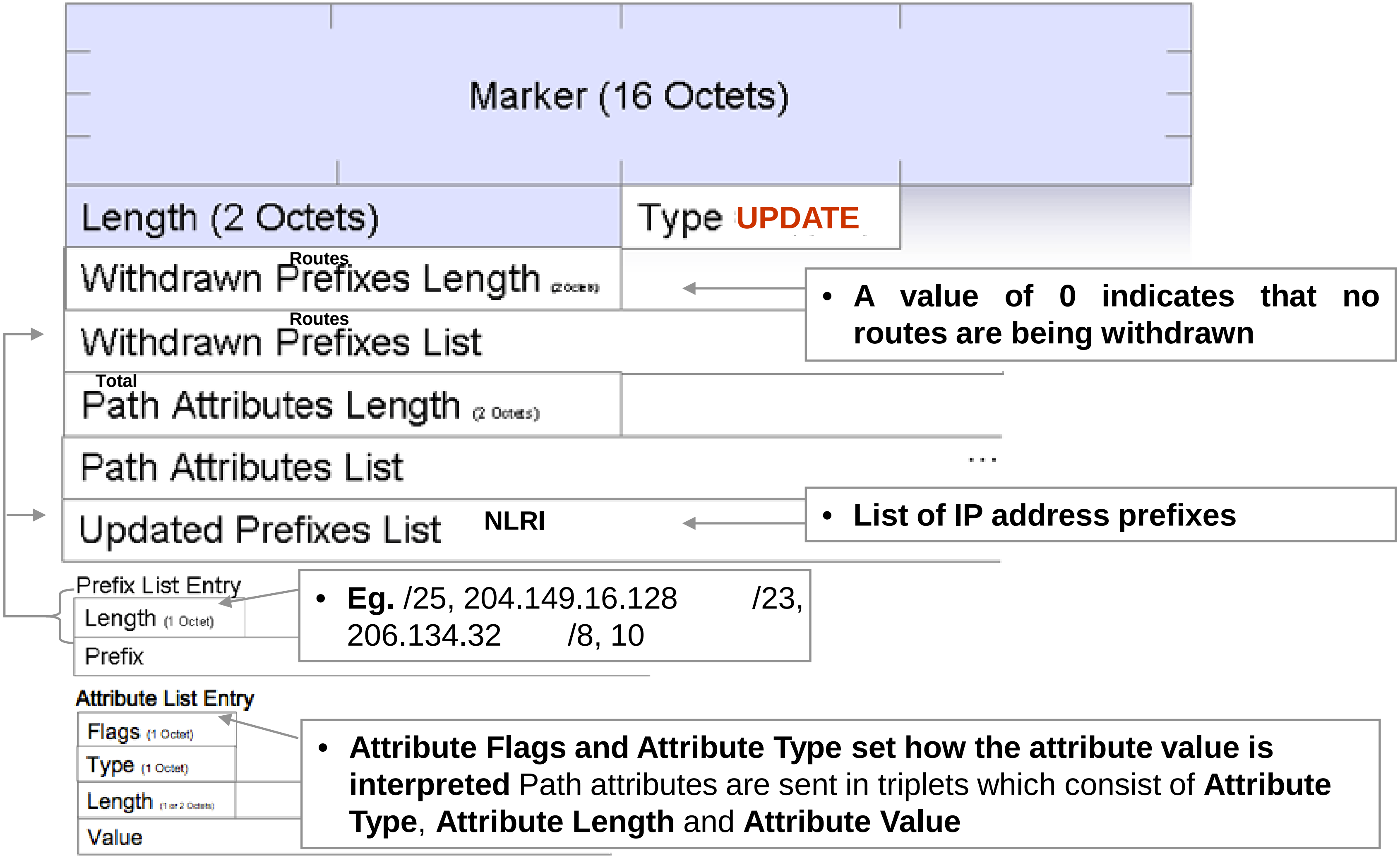
An UPDATE message MAY simultaneously withdraw multiple unfeasible routes and advertise feasible routes



- Each update message contains a set of address prefixes that are unreachable
- followed by a set of common path/route attributes,
- and then a set of updated prefixes that *share* this set of attributes (or “Network Layer Reachability Information” (NLRI) field)

- The updated prefixes are those prefixes where the local BGP instance has an updated view of the reachability of a prefix, or an updated view of the attributes of the locally selected ‘best’ route object for a prefix

An UPDATE message MAY simultaneously withdraw multiple unfeasible routes and advertise one feasible route



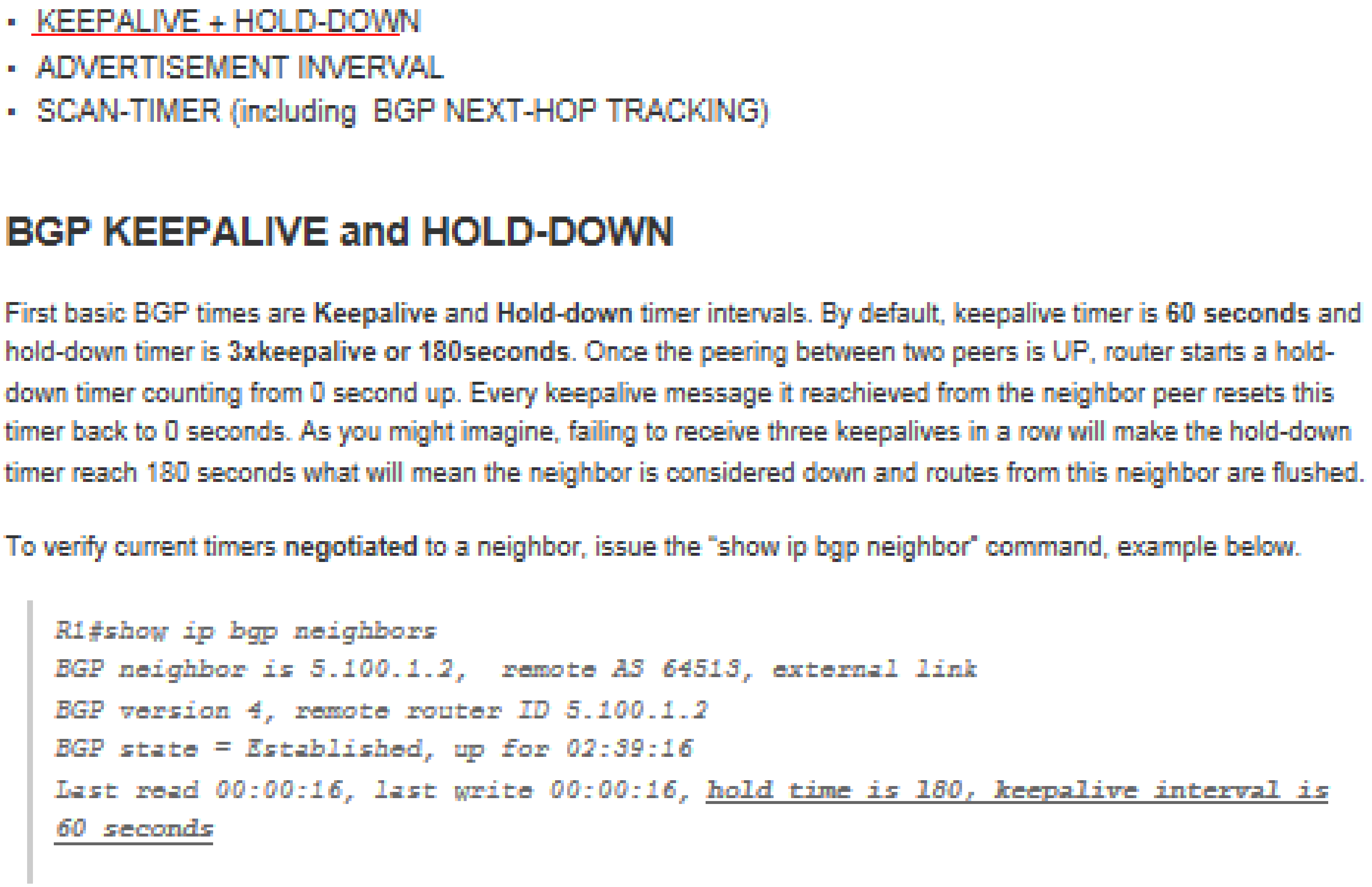
- A value of 0 indicates that no routes are being withdrawn

- List of IP address prefixes

- **Eg.** /25, 204.149.16.128 /23, 206.134.32 /8, 10

- **Attribute Flags and Attribute Type set how the attribute value is interpreted** Path attributes are sent in triplets which consist of **Attribute Type**, **Attribute Length** and **Attribute Value**

Selected BGP timers



- KEEPALIVE + HOLD-DOWN
- ADVERTISEMENT INVERVAL
- SCAN-TIMER (including BGP NEXT-HOP TRACKING)

BGP KEEPALIVE and HOLD-DOWN

First basic BGP times are **Keepalive** and **Hold-down** timer intervals. By default, keepalive timer is **60 seconds** and hold-down timer is **3xkeepalive or 180seconds**. Once the peering between two peers is UP, router starts a hold-down timer counting from 0 second up. Every keepalive message it reachieved from the neighbor peer resets this timer back to 0 seconds. As you might imagine, failing to receive three keepalives in a row will make the hold-down timer reach 180 seconds what will mean the neighbor is considered down and routes from this neighbor are flushed.

To verify current timers **negotiated** to a neighbor, issue the "show ip bgp neighbor" command, example below.

```
R1#show ip bgp neighbors
BGP neighbor is 5.100.1.2,  remote AS 64513, external link
BGP version 4, remote router ID 5.100.1.2
BGP state = Established, up for 02:39:16
Last read 00:00:16, last write 00:00:16, hold time is 180, keepalive interval is
60 seconds
```

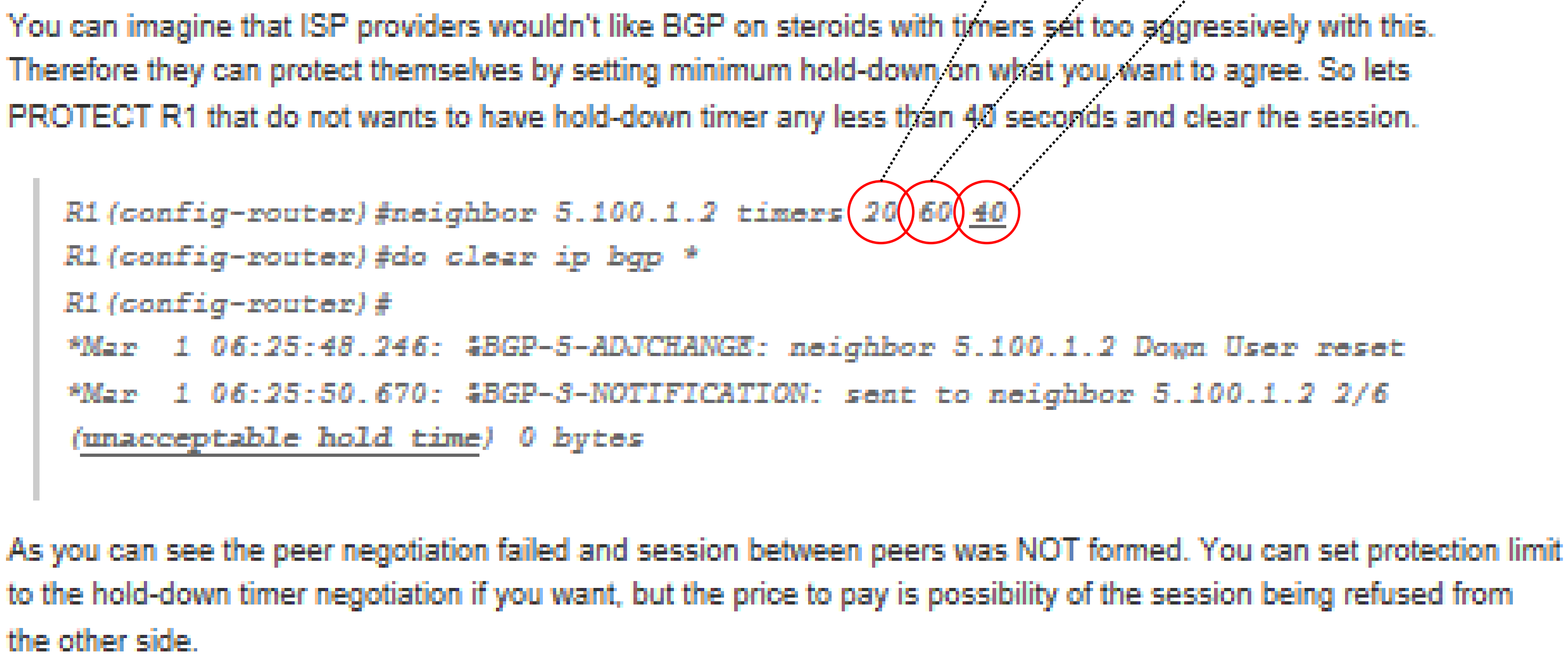
Selected BGP timers

- KEEPALIVE + HOLD-DOWN
- ADVERTISEMENT INVERVAL
- SCAN-TIMER (including BGP NEXT-HOP TRACKING)

keepalive

holddown

Minimum holddown



You can imagine that ISP providers wouldn't like BGP on steroids with timers set too aggressively with this. Therefore they can protect themselves by setting minimum hold-down on what you want to agree. So lets PROTECT R1 that do not wants to have hold-down timer any less than 40 seconds and clear the session.

```
R1(config-router)#neighbor 5.100.1.2 timers 20 60 40
R1(config-router)#do clear ip bgp *
R1(config-router)#
*Mar  1 06:25:48.246: %BGP-5-ADJCHANGE: neighbor 5.100.1.2 Down User reset
*Mar  1 06:25:50.670: %BGP-3-NOTIFICATION: sent to neighbor 5.100.1.2 2/6
(unacceptable hold time) 0 bytes
```

As you can see the peer negotiation failed and session between peers was NOT formed. You can set protection limit to the hold-down timer negotiation if you want, but the price to pay is possibility of the session being refused from the other side.

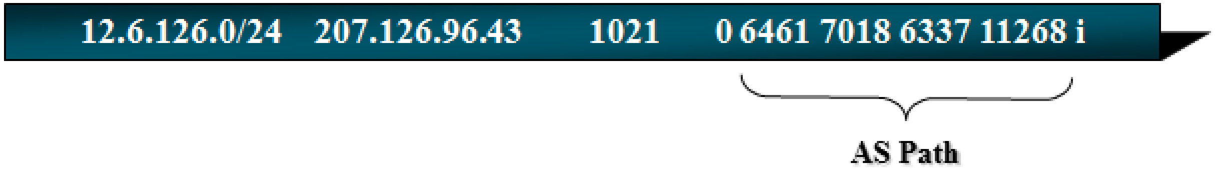
Content

- Introduction to BGP
- BGP terminology, concepts, messages, timers
- Attributes and path selection mechanisms
- Differences between I-BGP and E-BGP
- Challenges
- More on BGP scaling

What Is an Attribute?

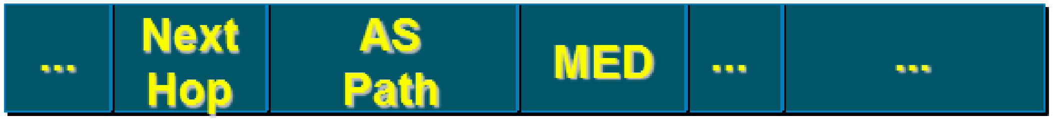
- **BGP is classified as a *path vector* routing protocol** (see RFC 1322)

  A path vector protocol defines a route as a pairing between a destination and the attributes of the path to that destination.



12.6.126.0/24 207.126.96.43 1021 0 6461 7018 6337 11268 i

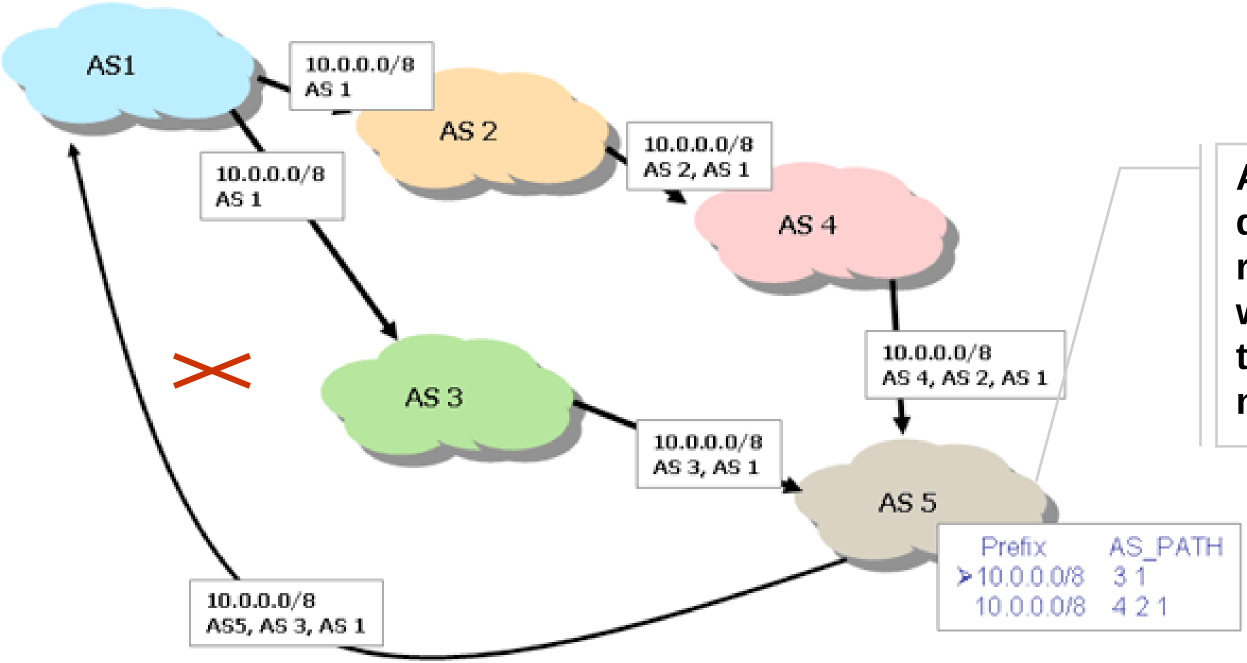
AS Path



... | Next Hop | AS Path | MED | ... | ...

- **Describes the characteristics of prefix**
- **Transitive or non-transitive**
- **Some are mandatory**

Routers speaking BGP exchange AS_PATH vectors



AS5 makes a policy decision as to which route path to accept, which, by default will be the shorter AS path, namely <AS3, AS1>

http://looking-glass.connect.com.au/lg/
http://lg.he.net/

Address prefix reachability information traverses the Internet in the form of individual route objects, this routing information is augmented by the list of AS's that have been traversed thus far, forming the AS_PATH attribute

BGP figures:
http://www.internetsociety.org/publications/isp-column-may-2006-introduction-bgp-%E2%80%93-protocol

BGP defines attribute types to describe an advertised route

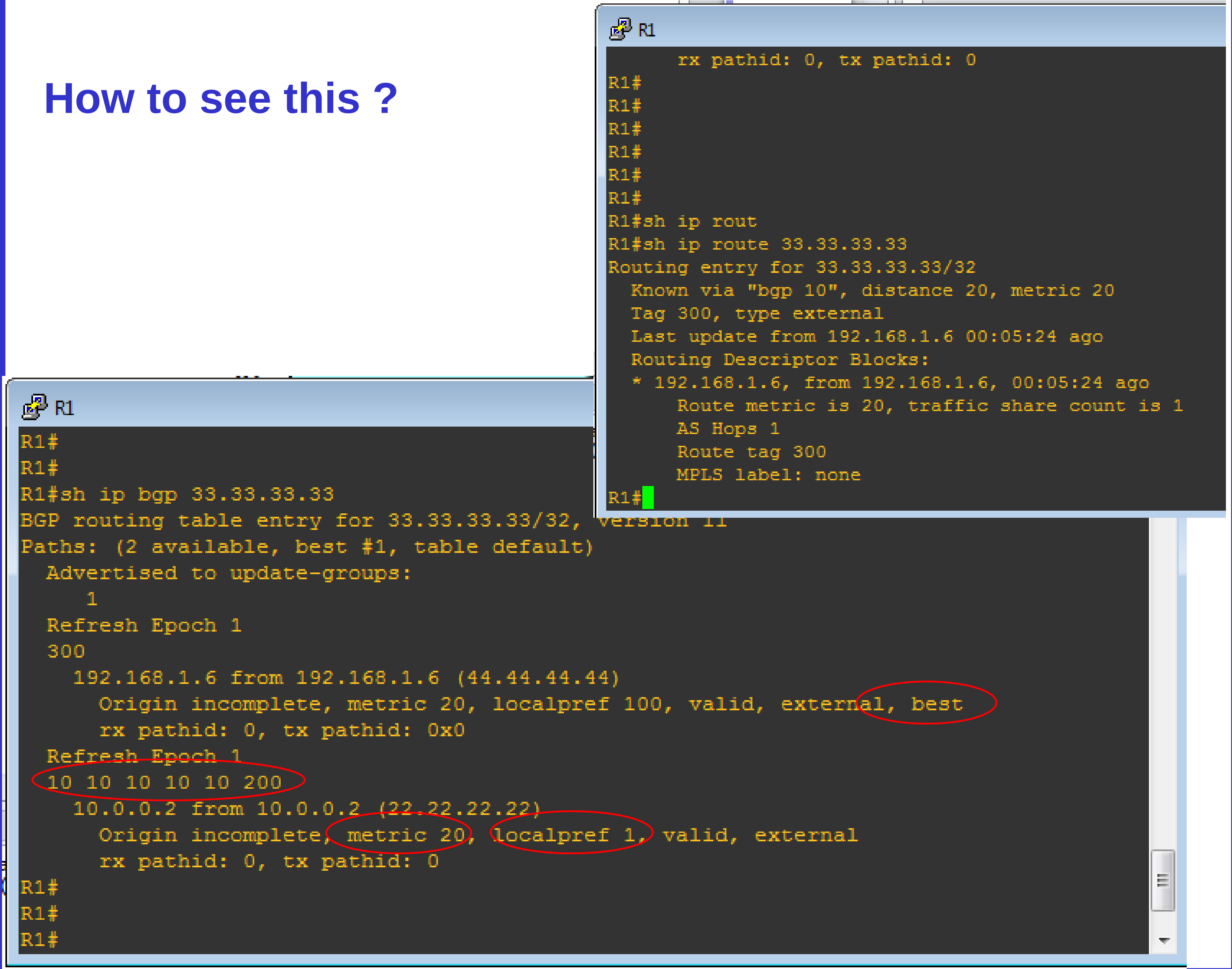
Well-known attributes

1. ORIGIN: Indicates origin of the path information: IGP, EGP or incomplete
2. AS_PATH: Describes the AS path vector associated with the prefix/destination
3. NEXT_HOP: IP address of border router to be used as next hop to reach the destination
4. LOCAL_PREF: Local preference is used to inform other BGP routers in the local AS of the originating BGP session's degree of preference for an advertised route
5. ATOMIC_AGGREGATE: Informs other BGP routers that the local system selected a less specific route without selecting a more specific route which is included in it

Optional

6. MULTI_EXIT_DISC: The Multiple Exit Discriminator (MED) discriminate among multiple exit points to a neighboring AS. If this information is received from an EBGP peer, it is propagated to each IBGP peer.
7. AGGREGATOR: Indicates the last AS number that formed the aggregate route and the IP address of the BGP router within the AS

How to see this ?



```
R1#
R1#
R1#sh ip bgp 33.33.33.33
BGP routing table entry for 33.33.33.33/32, version 11
Paths: (2 available, best #1, table default)
  Advertised to update-groups:
     1
  Refresh Epoch 1
  300
    192.168.1.6 from 192.168.1.6 (44.44.44.44)
      Origin incomplete, metric 20, localpref 100, valid, external, best
      rx pathid: 0, tx pathid: 0x0
  Refresh Epoch 1
  10 10 10 10 10 200
    10.0.0.2 from 10.0.0.2 (22.22.22.22)
      Origin incomplete, metric 20, localpref 1, valid, external
      rx pathid: 0, tx pathid: 0
R1#
R1#
R1#
```

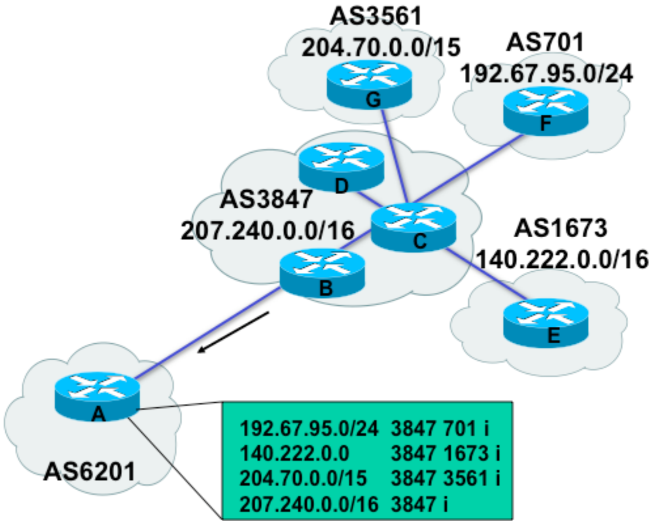
```
      rx pathid: 0, tx pathid: 0
R1#
R1#
R1#
R1#
R1#
R1#
R1#sh ip rout
R1#sh ip route 33.33.33.33
Routing entry for 33.33.33.33/32
  Known via "bgp 10", distance 20, metric 20
  Tag 300, type external
  Last update from 192.168.1.6 00:05:24 ago
  Routing Descriptor Blocks:
  * 192.168.1.6, from 192.168.1.6, 00:05:24 ago
      Route metric is 20, traffic share count is 1
      AS Hops 1
      Route tag 300
      MPLS label: none
R1#
```

BGP defines attribute types to describe an advertised route

Well-known attributes

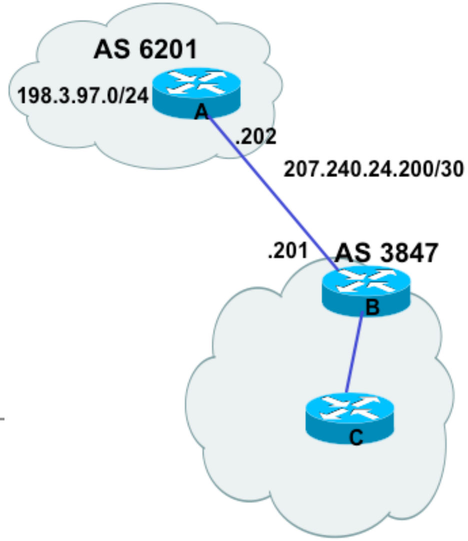
1. **ORIGIN**: Indicates origin of the path information: IGP, EGP or incomplete

2. **AS_PATH**: Describes the AS path vector associated with the prefix/destination



192.67.95.0/24	3847 701 i
140.222.0.0	3847 1673 i
204.70.0.0/15	3847 3561 i
207.240.0.0/16	3847 i

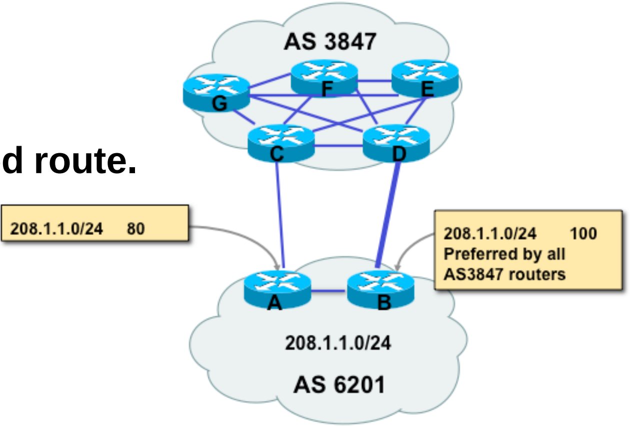
3. **NEXT_HOP**: IP address of border router to be used as next hop to reach the destination



BGP defines attribute types to describe an advertised route

Well-known attributes

4. **LOCAL_PREF**: **Local preference is used to inform other BGP routers in the local AS of the originating BGP session's degree of preference for an advertised route. If all other attributes are equal, the route with the higher degree of preference is preferred.**

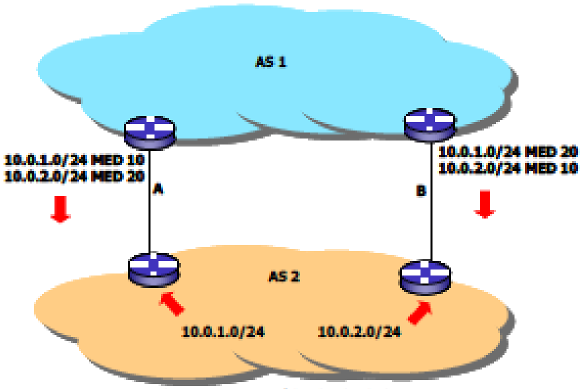


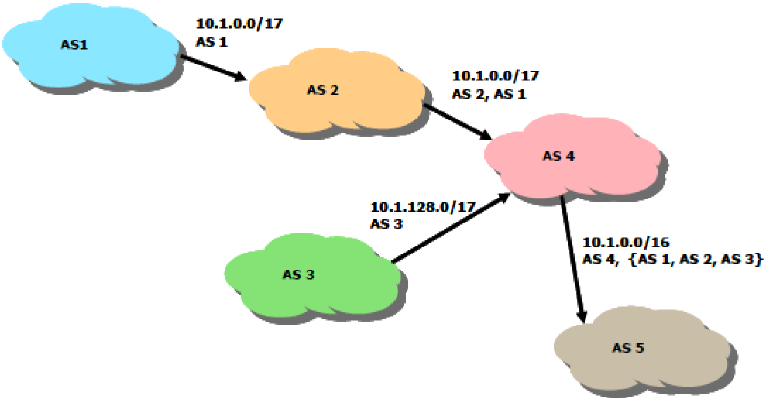
5. **ATOMIC_AGGREGATE**: **Used by BGP speakers to inform other BGP routers that the local system selected a less specific route without selecting a more specific route which is included in it**
A route with this attribute included may actually traverse autonomous systems not listed in the AS_PATH.

BGP defines attribute types to describe an advertised route

Optional

6. MULTI_EXIT_DISC: The Multiple Exit Discriminator (MED) discriminate among multiple exit points to a neighboring AS. If this information is received from an EBGP peer, it is propagated to each IBGP peer. If all other attributes are equal, the exit point with the lowest MED value is preferred.





7. AGGREGATOR: Indicates the last AS number that formed the aggregate route and the IP address of the BGP router that formed the aggregate route.

Cisco-defined BGP attributes (similar for other vendors)

Optional

- **COMMUNITY: allows BGP communities to be set up and provides a way of grouping destinations according to common BGP attributes, filters and policies. Four well-known communities:**
  - **INTERNET - by default all routes belong to this community and are advertised**
  - **NO_EXPORT - routes with this attribute cannot be advertised to EBGP peers i.e. outside of their AS with the exception of internal ASs within a Confederation.**
  - **NO_ADVERTISE - routes with this attribute cannot be advertised to either EBGP or IBGP peers**
  - **LOCAL_AS - acts in the same way as NO_EXPORT except that routes with this attribute cannot even be advertised between EBGP peers in private ASs within a Conferation**

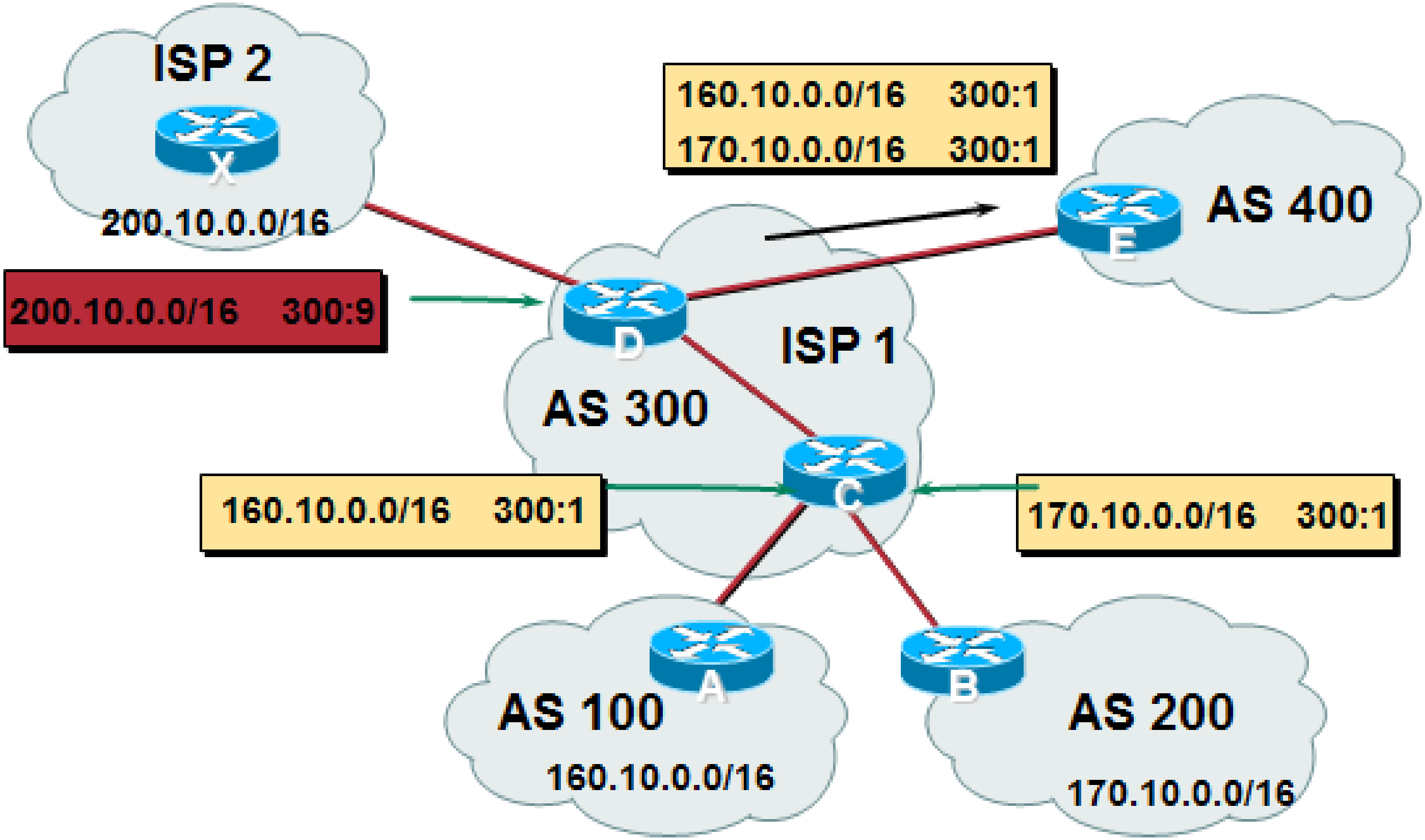
- **ORIGINATOR_ID: used by a Route Reflector as a Reflector ID (RID) to ensure that no loops occur in an AS using Route Reflectors**

- **CLUSTER_LIST: lists the route reflector cluster IDs that the route has passed through so that if a route reflector sees it's own cluster ID it drops the route to stop loops**

BGP Communities

- BGP attribute
- Used to group destinations
- Represented as two 16bit integers
- Each destination could be member of multiple communities
- Community attribute carried across AS's
- Useful in applying policies

BGP Communities



The default BGP route selection process is to prefer a path with the longest prefix match

- **When comparing two route objects that refer to the same prefix, then there are a sequence of comparisons to determine which route object is selected by the local BGP speaker**

1. **Select the route object with the highest value for LOCAL_PREF**
2. **Select the route object shortest AS_PATH**
3. **Select the lowest MULTI_EXIT_DISCRIMINATOR**
4. **Select the minimum IGP cost to the NEXT_HOP address**
5. **Select eBGP over iBGP-learned routes**
6. **If iBGP select the lowest BGP Identifier value**

- **Shall we try it ? On GNS3 ?**

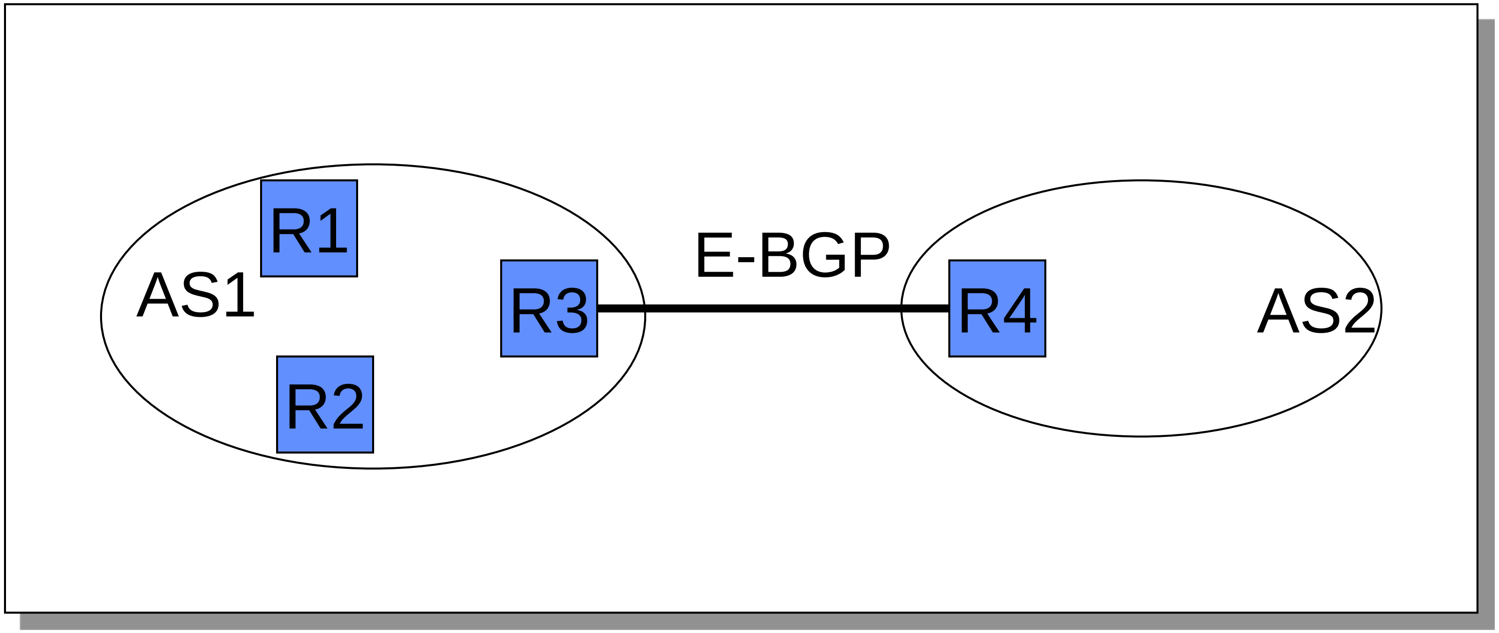
- **http://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/13753-25.html**

Content

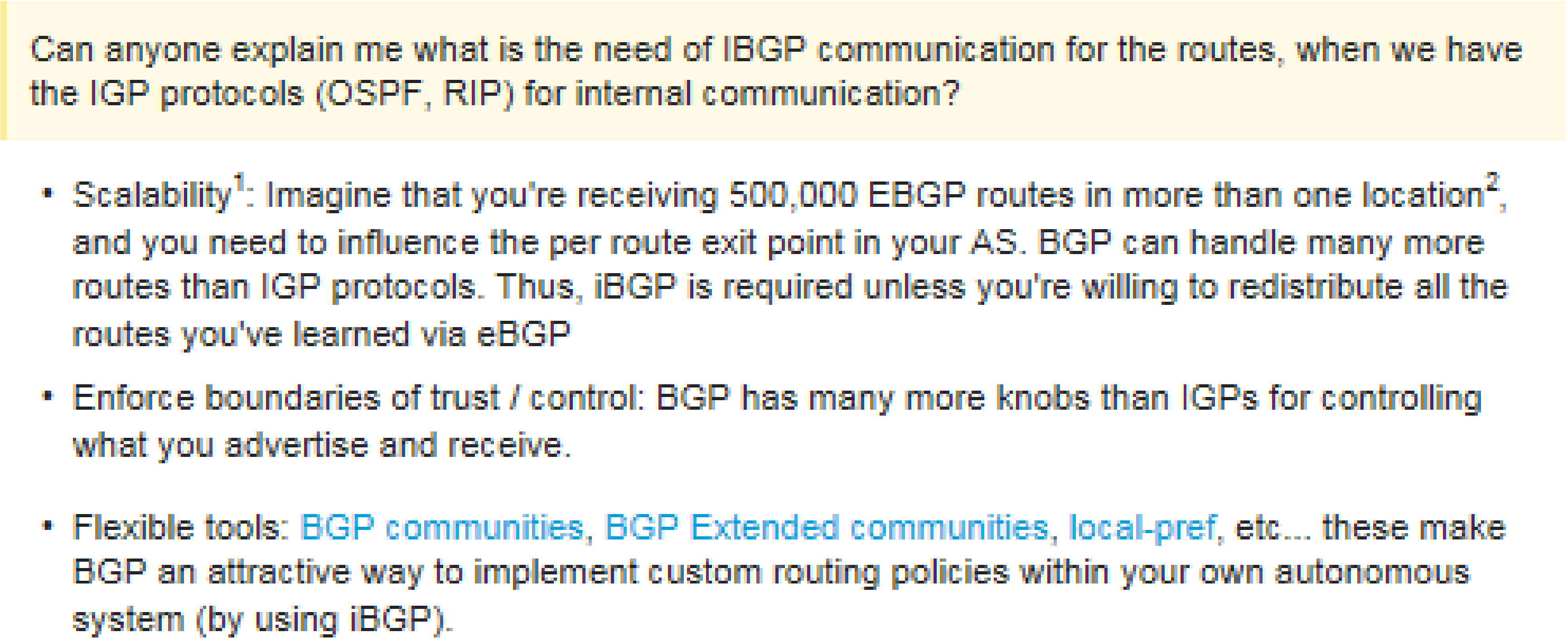
- Introduction to BGP
- BGP terminology, concepts, messages, timers
- Attributes and path selection mechanisms
- Differences between I-BGP and E-BGP
- Challenges
- More on BGP scaling

Internal vs. External BGP

- BGP can be used by R3 and R4 to learn routes
- How do R1 and R2 learn routes?
  - Yes, IGP is an alternative – but what if you need more information than what IGP can carry ?
  - What type of information could that be ?
  - What purpose ?

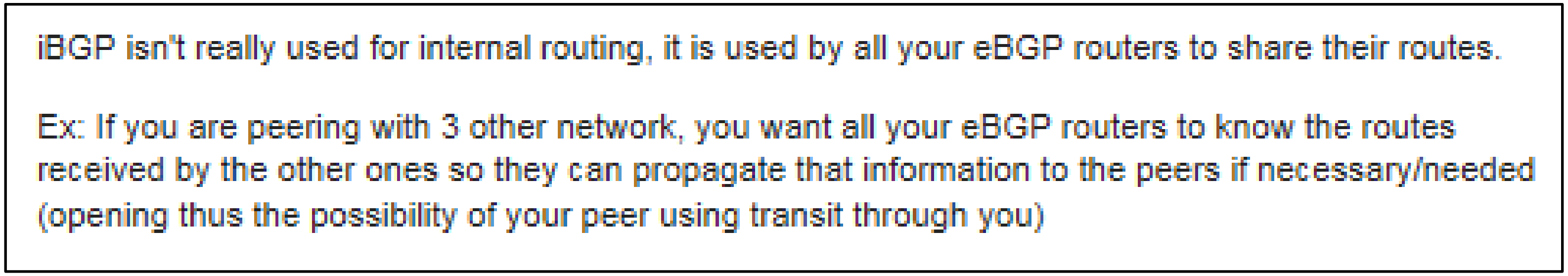


Reasons for using i-BGP



> Can anyone explain me what is the need of IBGP communication for the routes, when we have the IGP protocols (OSPF, RIP) for internal communication?

- Scalability[1]: Imagine that you're receiving 500,000 EBGP routes in more than one location[2], and you need to influence the per route exit point in your AS. BGP can handle many more routes than IGP protocols. Thus, iBGP is required unless you're willing to redistribute all the routes you've learned via eBGP
- Enforce boundaries of trust / control: BGP has many more knobs than IGPs for controlling what you advertise and receive.
- Flexible tools: BGP communities, BGP Extended communities, local-pref, etc... these make BGP an attractive way to implement custom routing policies within your own autonomous system (by using iBGP).



iBGP isn't really used for internal routing, it is used by all your eBGP routers to share their routes.

Ex: If you are peering with 3 other network, you want all your eBGP routers to know the routes received by the other ones so they can propagate that information to the peers if necessary/needed (opening thus the possibility of your peer using transit through you)

Internal BGP (I-BGP)

- **Same messages as E-BGP**
- **Different rules about re-advertising prefixes:**
  - **Prefix learned from E-BGP can be advertised to I-BGP neighbor and vice-versa, but**
  - **Prefix learned from one I-BGP neighbor cannot be advertised to another I-BGP neighbor**
  - **Reason: no AS PATH within the same AS and thus danger of looping.**

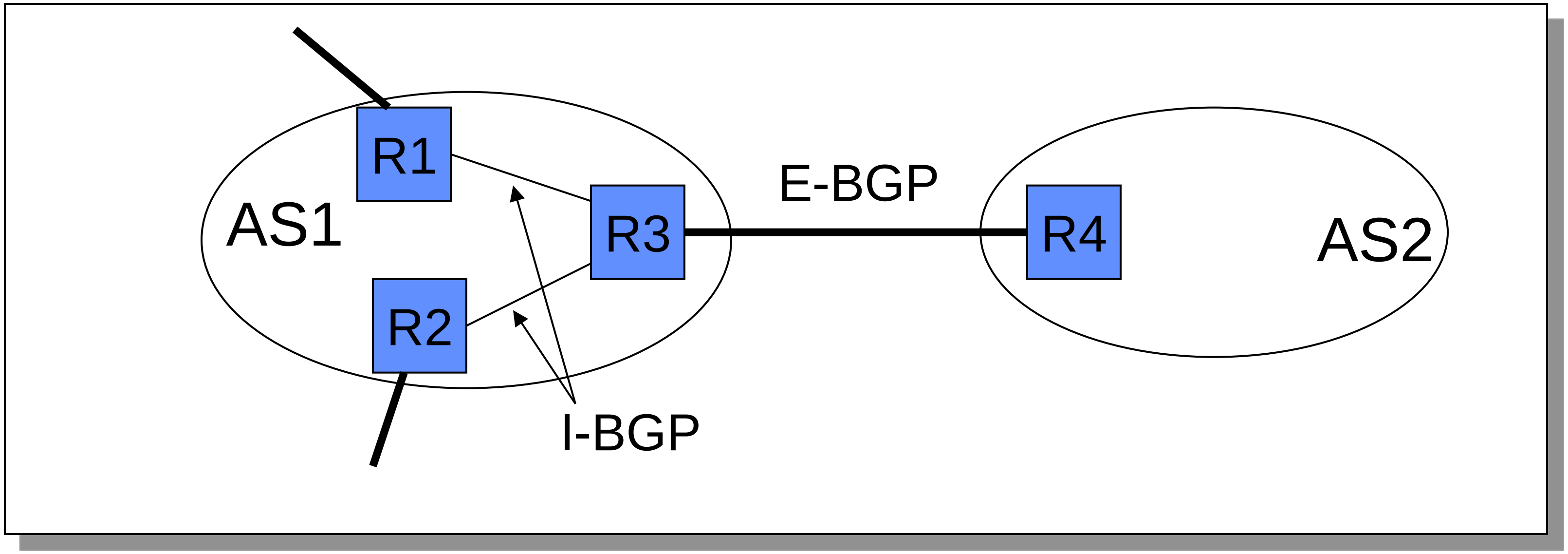
Internal BGP (I-BGP)

- R3 can tell R1 and R2 prefixes from R4
- R3 can tell R4 prefixes from R1 and R2
- R3 cannot tell R2 prefixes from R1

R2 can only find these prefixes through a *direct connection* to R1

Result: I-BGP routers must be fully connected (via TCP)!

- contrast with E-BGP sessions that map to physical links

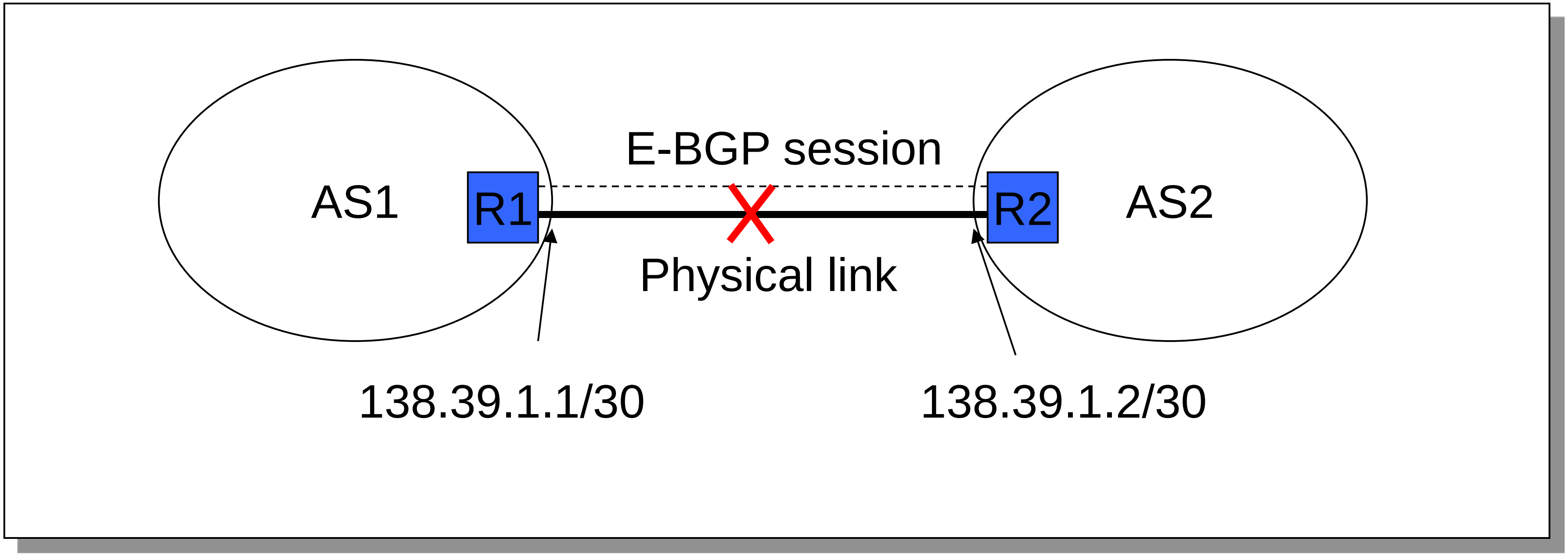


Link Failures

- **Two types of link failures:**
  - **Failure on an E-BGP link**
  - **Failure on an I-BGP Link**
- **These failures are treated completely different in BGP**
- **Why?**

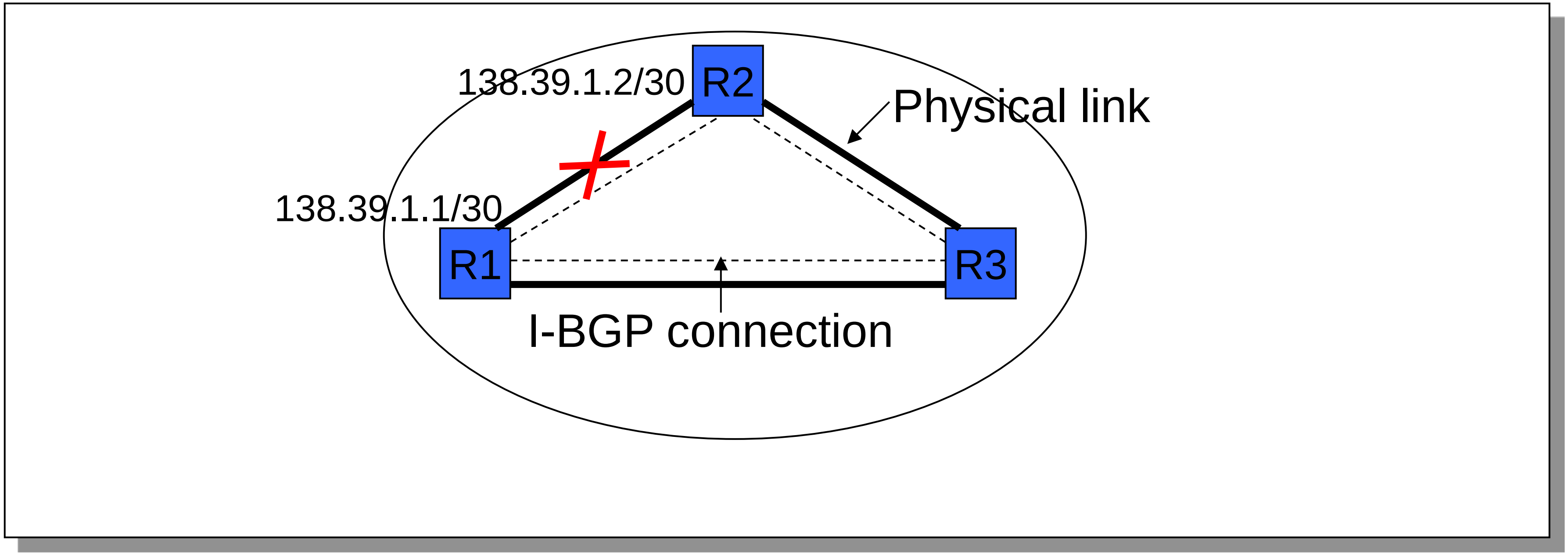
Failure on an E-BGP Link

- If the link R1-R2 goes down
  - The TCP connection breaks
  - BGP routes are removed
- This is the *desired* behavior



Failure on an I-BGP Link

- If link R1-R2 goes down, R1 and R2 should still be able to exchange traffic
- The indirect path through R3 must be used
- Thus, E-BGP and I-BGP must use *different conventions* with respect to TCP endpoints



Content

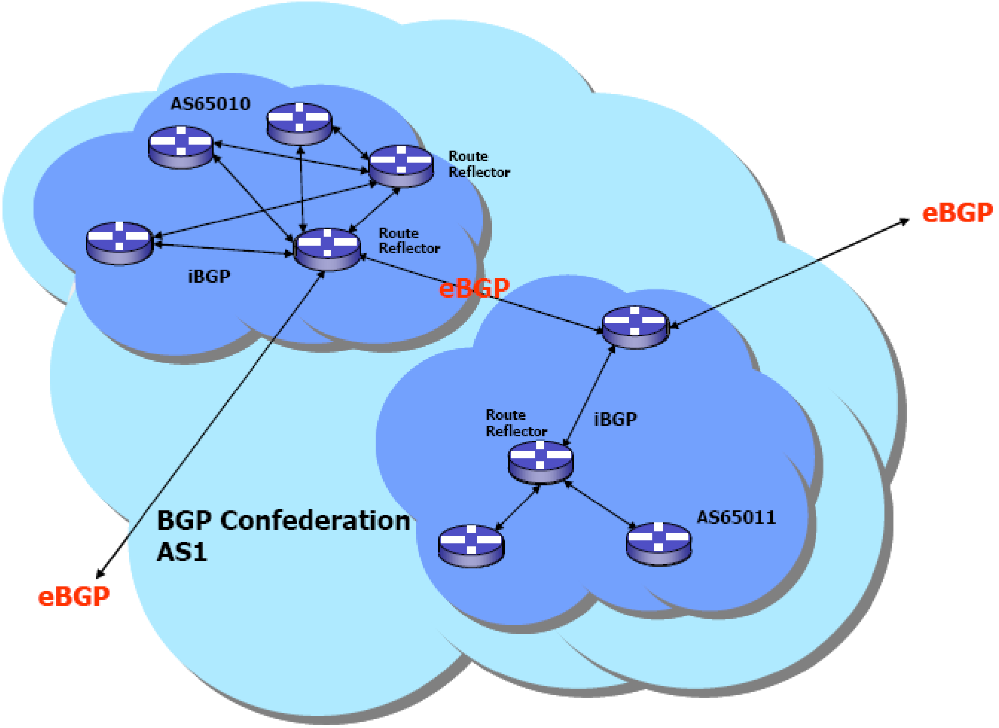
- Introduction to BGP
- BGP terminology, concepts, messages, timers
- Attributes and path selection mechanisms
- Differences between I-BGP and E-BGP
- Challenges
- More on BGP scaling

General costs of BGP

- CPU consumption in
  - BGP session establishment
  - route selection
  - routing information processing
  - handling of routing updates
- Router memory
  - to install routes and multiple paths associated with the routes
- System complexity
  - In an AS with N routers, each router will have to establish I-BGP sessions with N-1 routers (if I-BGP required and used)
- The large number of iBGP sessions and routes consumes tremendous resources from each router, especially during BGP session establishment and during periods of heavy route flapping (path going up and down)

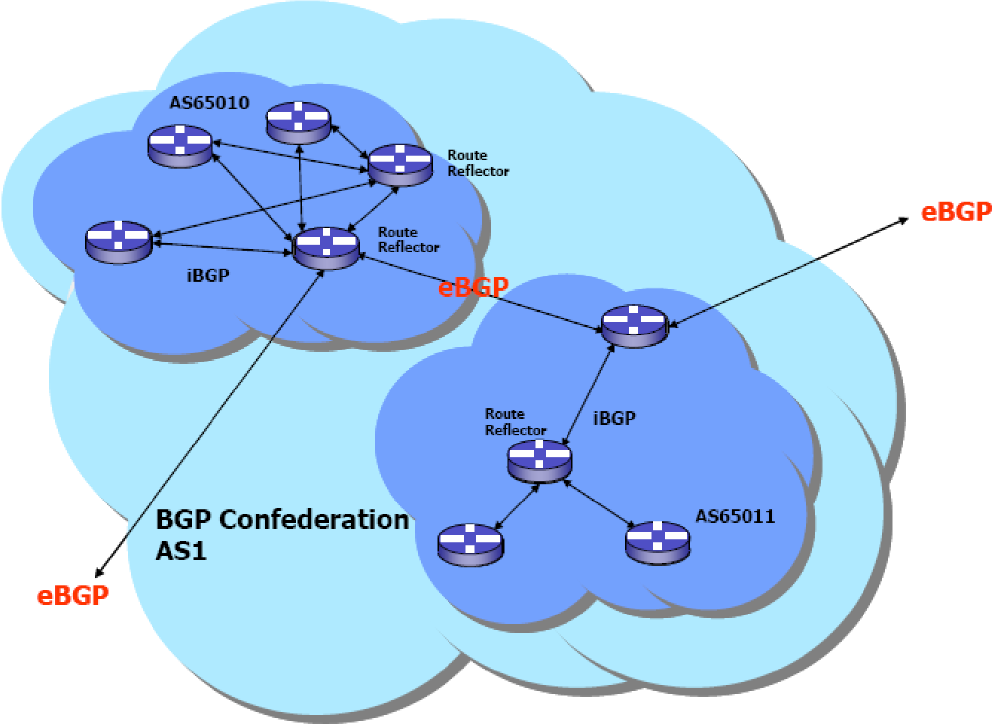
Route reflectors for scalability to avoid full mesh iBGP

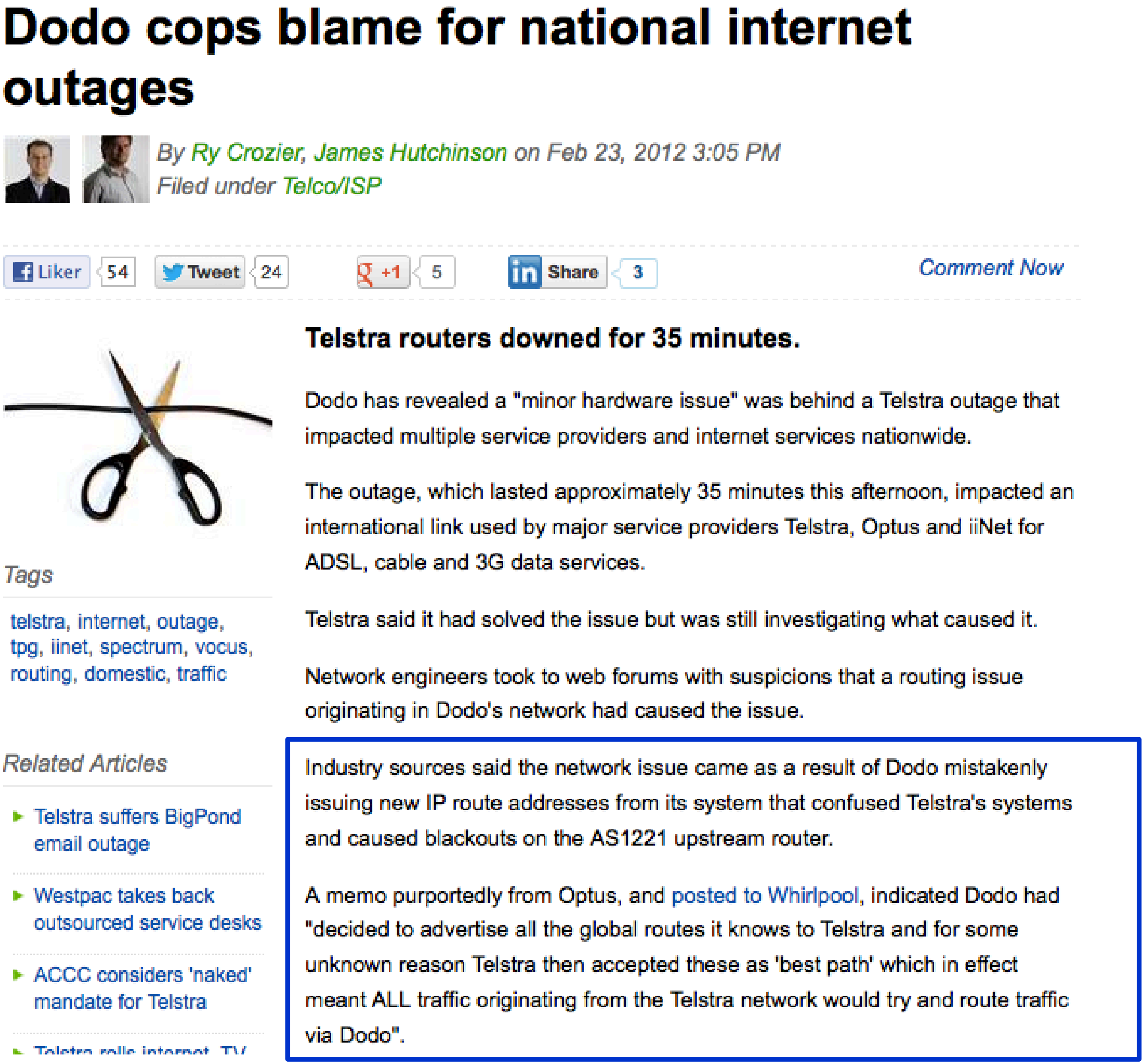
- **Route reflectors** reduce the number of connections required in an AS.
- A single router (or two for redundancy) can be made a route reflector: other routers in the AS need only be configured as peer to them



Confederations for scalability to avoid full mesh iBGP

- Confederations are sets of autonomous systems
- A confederation of autonomous systems is represented as a single autonomous system to BGP peers external to the confederation, thereby removing the "full mesh" requirement
- Confederations can be used in conjunction with route reflectors





Dodo cops blame for national internet outages

By Ry Crozier, James Hutchinson on Feb 23, 2012 3:05 PM
Filed under Telco/ISP

Telstra routers downed for 35 minutes.

Dodo has revealed a "minor hardware issue" was behind a Telstra outage that impacted multiple service providers and internet services nationwide.

The outage, which lasted approximately 35 minutes this afternoon, impacted an international link used by major service providers Telstra, Optus and iiNet for ADSL, cable and 3G data services.

Telstra said it had solved the issue but was still investigating what caused it.

Network engineers took to web forums with suspicions that a routing issue originating in Dodo's network had caused the issue.

Industry sources said the network issue came as a result of Dodo mistakenly issuing new IP route addresses from its system that confused Telstra's systems and caused blackouts on the AS1221 upstream router.

A memo purportedly from Optus, and posted to Whirlpool, indicated Dodo had "decided to advertise all the global routes it knows to Telstra and for some unknown reason Telstra then accepted these as 'best path' which in effect meant ALL traffic originating from the Telstra network would try and route traffic via Dodo".

Tags

telstra, internet, outage, tpg, iinet, spectrum, vocus, routing, domestic, traffic

Related Articles

- Telstra suffers BigPond email outage
- Westpac takes back outsourced service desks
- ACCC considers 'naked' mandate for Telstra
- Telstra rolls internet, TV

http://www.itnews.com.au/News/291364,dodo-cops-blame-for-national-internet-outages.aspx

How China swallowed 15% of Net traffic for 18 minutes

- For about 18 minutes on April 8, 2010, China Telecom advertised erroneous network traffic routes that instructed US and other foreign Internet traffic to travel through Chinese servers. Other servers around the world quickly adopted these paths, routing all traffic to about 15 percent of the Internet’s destinations through servers located in China. This incident affected traffic to and from US government (‘‘.gov’’) and military (‘‘.mil’’) sites, including those for the Senate, the army, the navy, the marine corps, the air force, the office of secretary of Defense, the National Aeronautics and Space Administration, the Department of Commerce, the National Oceanic and Atmospheric Administration, and many others. Certain commercial websites were also affected, such as those for Dell, Yahoo!, Microsoft, and IBM.
- The culprit here was "IP hijacking," a well-known routing problem in a worldwide system based largely on trust. Routers rely on the Border Gateway Protocol (BGP) to puzzle out the best route between two IP addresses; when one party advertises incorrect routing information, routers across the globe can be convinced to send traffic on geographically absurd paths.
- This happened famously in 2008, when Pakistan blocked YouTube. The block was meant only for internal use, and it relied on new routing information that would send YouTube requests not to the company's servers but into a "black hole.”
- As we described the situation at the time, "this routing information escaped from Pakistan Telecom to its ISP PCCW in Hong Kong, which propagated the route to the rest of the world. Soany packets for YouTube would end up in Pakistan Telecom's black hole instead." The mistake broke YouTube access from across much of the Internet.

http://arstechnica.com/security/news/2010/11/how-china-swallowed-15-of-net-traffic-for-18-minutes.ars

Black-holing of Youtube by Pakistan Telecom, an intentional hijacking designed to make a political statement?

- PieNet configured their Internet routers to appear to have a better route to www.youtube.com than other routers.
  - started advertising a route for 208.65.153.0/24 to its provider, PCCW (AS 3491)
  - a more specific route than the ones used by YouTube (208.65.152.0/22), and therefore most routers would choose to send traffic to Pakistan Telecom for this slice of YouTube's network.
- Soon their routers shared this new and better route with other routers who in turn shared it with even more routers and soon the news was spreading like wildfire. The problem was, PieNet made sure that this hot new route to www.youtube.com actually led to PieNet's own servers and not YouTube at all.
- "Whether accidental or not, the black-holing of Youtube by Pakistan Telecom demonstrates a serious weakness in the 'longest prefix wins' rule: there is no concept of trust contained in it," Tomas Byrnes wrote on the NANOG list. "Trust, whether implicit or explicit, is inherent in all human interactions, yet expressing it in cyberspace has continued to be troublesome. In routing decisions, once you are beyond a connected (either directly or multi-hop) peer, it becomes much more difficult."

- http://billso.com/2008/02/24/pakistan-blocks-youtube-breaks-trust/
- http://www.datacenterknowledge.com/archives/2008/02/25/how-to-avoid-another-major-ip-hijacking/

Other serious considerations

- For public peering: filter EBGP routes inbound and outbound
  - Block your own address space inbound
  - Block RFC 1918 space (inbound and outbound)
  - Block DSUA space (inbound and outbound):
  - http://www.ietf.org/internet-drafts/draft-manning-dsua-08.txt
- Use prefix-lists for route-filtering when possible (easier to read than ACLs)

Content

- Introduction to BGP
- BGP terminology, concepts, messages, timers
- Attributes and path selection mechanisms
- Differences between I-BGP and E-BGP
- Challenges
- More on BGP scaling

BGP Scaling Techniques

- How to scale iBGP mesh beyond a few peers?
- How to implement new policy without causing flaps and route churning?
- How to reduce the overhead on the routers?

BGP Scaling Techniques

- Dynamic reconfiguration
- Peer groups
- Route flap damping
- Route reflectors

Soft Reconfiguration

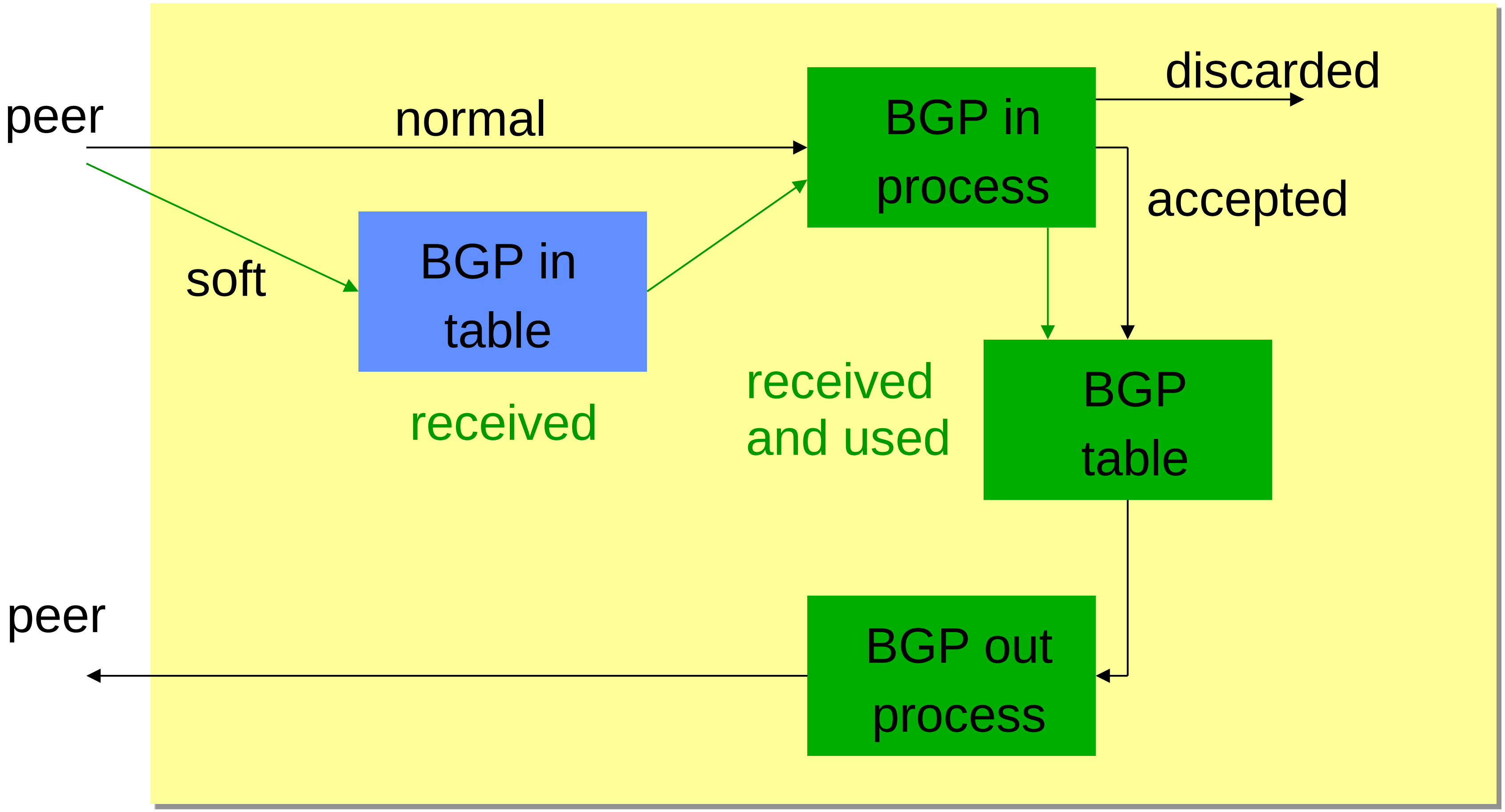
Problem:

- Hard BGP peer clear required after every policy change because the router does not store prefixes that are denied by a filter
- Hard BGP peer clearing consumes CPU and affects connectivity for all networks

Solution:

- Soft-reconfiguration

Soft Reconfiguration



Soft Reconfiguration

- New policy is activated without tearing down and restarting the peering session
- Per-neighbour basis
- Use more memory to keep prefixes whose attributes have been changed or have not been accepted

Configuring Soft reconfiguration

```
router bgp 100
 neighbor 1.1.1.1 remote-as 101
 neighbor 1.1.1.1 route-map infilter in
 neighbor 1.1.1.1 soft-reconfiguration inbound
```

! *Outbound does not need to be configured* !

Then when we change the policy, we issue an exec command

```
clear ip bgp 1.1.1.1 soft [in | out]
```

Managing Policy Changes

- `clear ip bgp <addr> [soft] [in|out]`
  - <addr> may be any of the following

x.x.x.x	IP address of a peer
*	all peers
ASN	all peers in an AS
external	all external peers
peer-group <name>	all peers in a peer-group

Route Refresh Capability

- **Facilitates non-disruptive policy changes**
- **No configuration is needed**
- **No additional memory is used**
- **Requires peering routers to support “route refresh capability” – RFC2918**
- **clear ip bgp x.x.x.x in tells peer to resend full BGP announcement**

Soft Reconfiguration vs Route Refresh

- Use Route Refresh capability if supported
  - find out from “show ip bgp neighbor”
  - uses much less memory
- Otherwise use Soft Reconfiguration

Peer Groups

Without peer groups

- iBGP neighbours receive same update
- Large iBGP mesh slow to build
- Router CPU wasted on repeat calculations

Solution – peer groups!

- Group peers with same outbound policy
- Updates are generated once per group

Peer Groups - Advantages

- Makes configuration easier
- Makes configuration less prone to error
- Makes configuration more readable
- Lower router CPU load
- iBGP mesh builds more quickly
- Members can have different inbound policy
- Can be used for eBGP neighbours too!

Configuring Peer Group

```
router bgp 100
 neighbor ibgp-peer peer-group
 neighbor ibgp-peer remote-as 100
 neighbor ibgp-peer update-source loopback 0
 neighbor ibgp-peer send-community
 neighbor ibgp-peer route-map outfilter out
 neighbor 1.1.1.1 peer-group ibgp-peer
 neighbor 2.2.2.2 peer-group ibgp-peer
 neighbor 2.2.2.2 route-map  infilter in
 neighbor 3.3.3.3 peer-group ibgp-peer
```

! *note how 2.2.2.2 has different inbound filter from peer-group* !

Route Flap Damping

- Route flap
  - Going up and down of path or change in attribute
    - BGP WITHDRAW followed by UPDATE = 1 flap
    - eBGP neighbour going down/up is NOT a flap
  - Ripples through the entire Internet
  - Wastes CPU
- Damping aims to reduce scope of route flap propagation

Route Flap Damping (Continued)

- Requirements
  - Fast convergence for normal route changes
  - History predicts future behaviour
  - Suppress oscillating routes
  - Advertise stable routes
- Implementation described in RFC2439

Operation

- Add penalty (1000) for each flap
  - Change in attribute gets penalty of 500
- Exponentially decay penalty

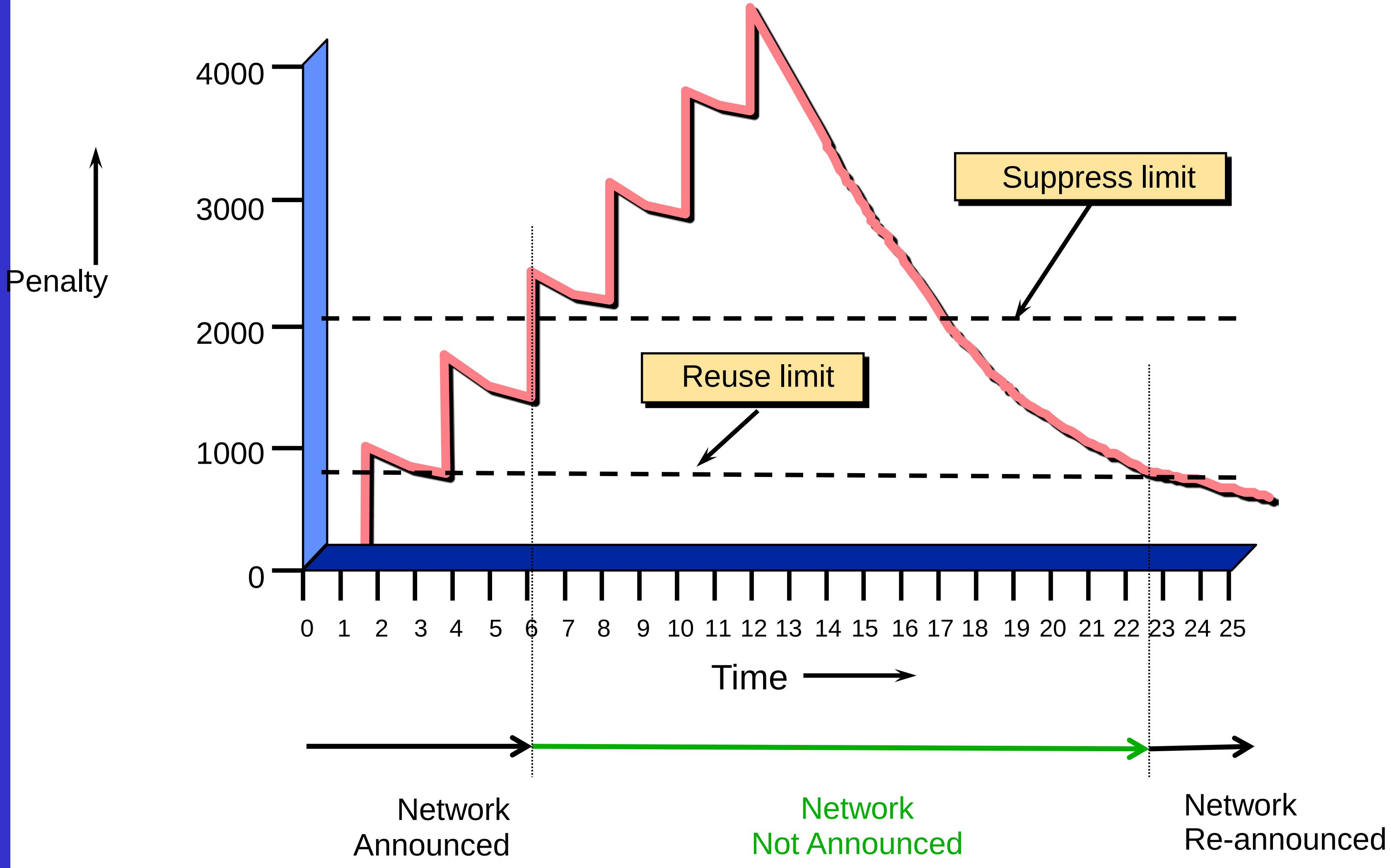
  half life determines decay rate
- Penalty above suppress-limit

  do not advertise route to BGP peers
- Penalty decayed below reuse-limit

  re-advertise route to BGP peers

  penalty reset to zero when it is half of reuse-limit

Operation


4000
3000
2000
1000
0
Penalty
Suppress limit
Reuse limit
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Time
Network Announced
Network Not Announced
Network Re-announced

Operation

- Only applied to inbound announcements from eBGP peers
- Alternate paths still usable
- Controlled by:
  - Half-life (default 15 minutes)
  - reuse-limit (default 750)
  - suppress-limit (default 2000)
  - maximum suppress time (default 60 minutes)

Configuration

Fixed damping

```
router bgp 100
 bgp dampening [<half-life> <reuse-value> <suppress-penalty> <maximum suppress time>]
```

Selective and variable damping

```
 bgp dampening [route-map <name>]
  route-map <name> permit 10
   match ip address prefix-list FLAP-LIST
   set dampening [<half-life> <reuse-value> <suppress-penalty> <maximum suppress time>]
 ip prefix-list FLAP-LIST permit 192.0.2.0/24 le 32
```

Operation

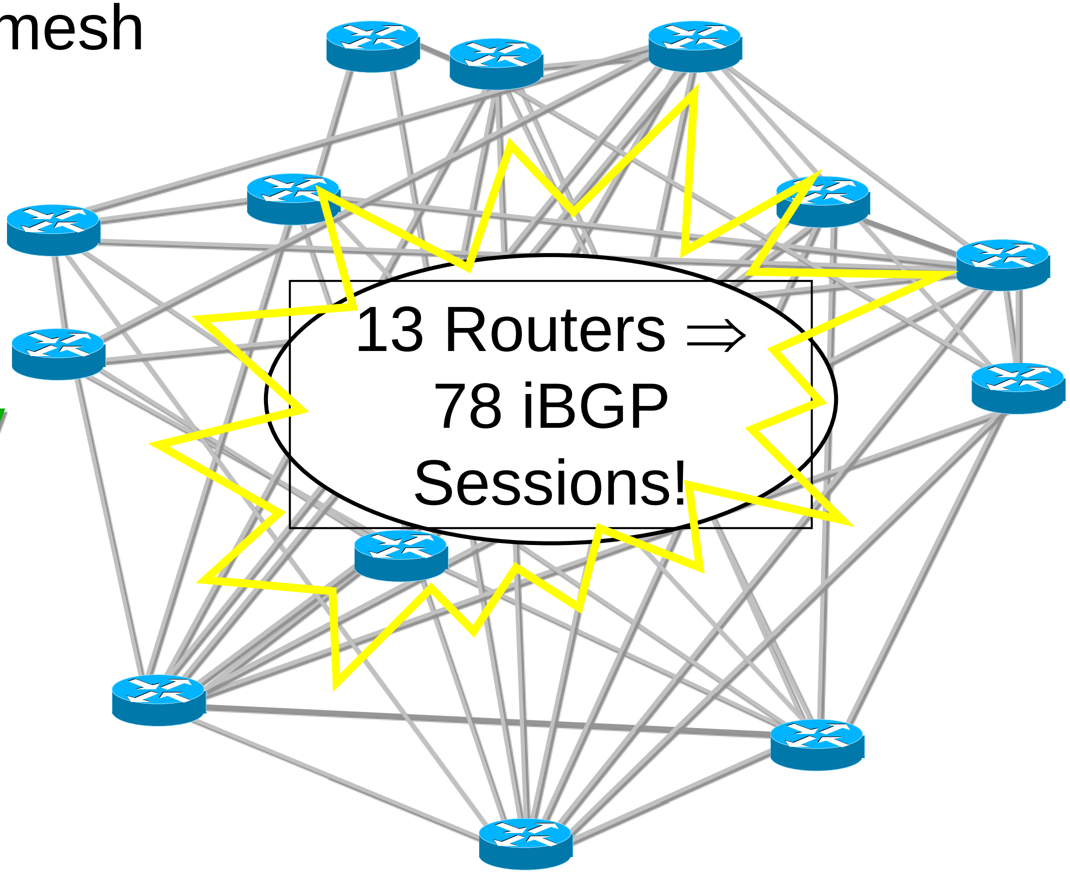
- **Care required when setting parameters**
- **Penalty must be less than reuse-limit at the maximum suppress time**
- **Maximum suppress time and half life must allow penalty to be larger than suppress limit**

Enhancements

- Selective damping based on
  - AS-path, Community, Prefix
- Variable damping
  - recommendations for ISPs
  - http://www.ripe.net/docs/ripe-229.html
- Flap statistics
  - `show ip bgp neighbor <x.x.x.x> [dampened-routes | flap-statistics]`

Scaling iBGP mesh

Avoid n(n-1)/2 iBGP mesh



13 Routers ⇒ 78 iBGP Sessions!

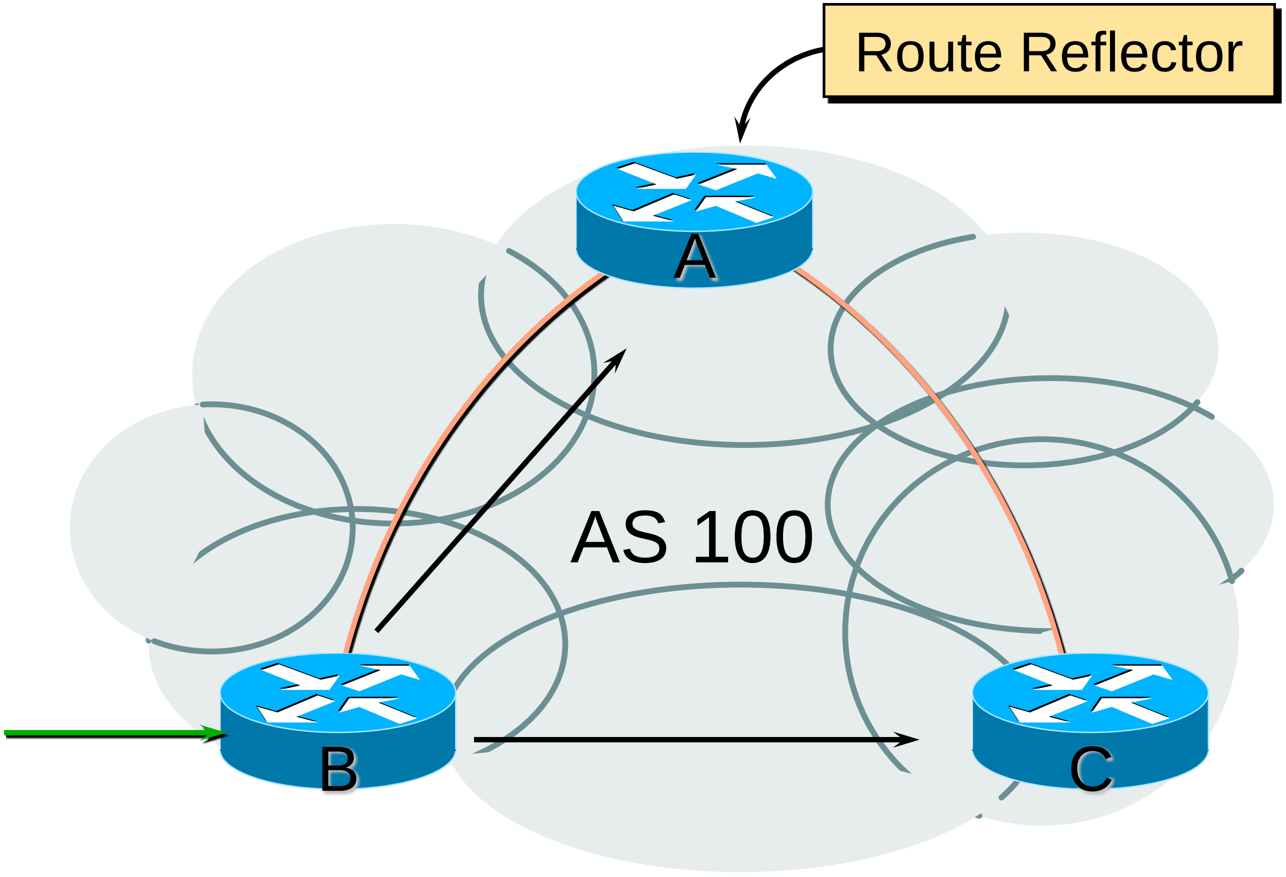
n=1000 ⇒ nearly half a million ibgp sessions!

Two solutions

Route reflector – simpler to deploy and run

Confederation – more complex, corner case benefits

Route Reflector: Principle



Route Reflector

Reflector receives path from clients and non-clients

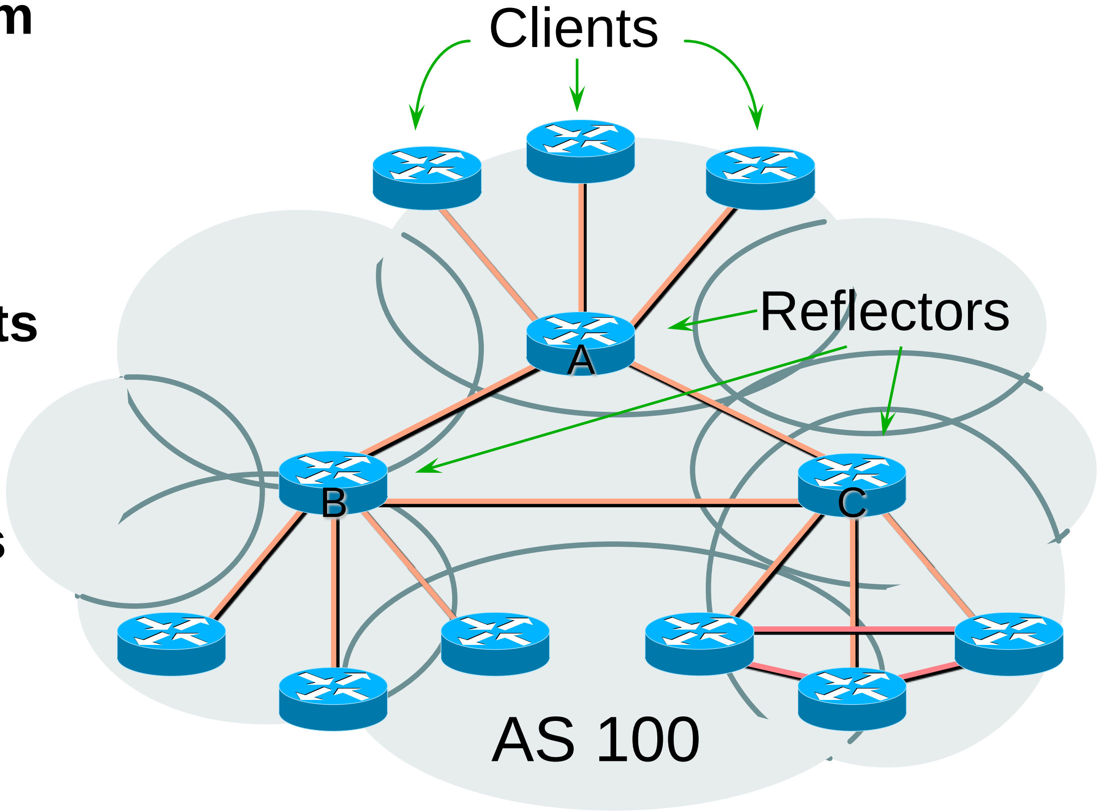
Selects best path

If best path is from client, reflect to other clients and non-clients

If best path is from non-client, reflect to clients only

Non-meshed clients

Described in RFC2796



Route Reflector Topology

- **Divide the backbone into multiple clusters**
- **At least one route reflector and few clients per cluster**
- **Route reflectors are fully meshed**
- **Clients in a cluster could be fully meshed**
- **Single IGP to carry next hop and local routes**

Route Reflector: Benefits

- Solves iBGP mesh problem
- Packet forwarding is not affected
- Normal BGP speakers co-exist
- Multiple reflectors for redundancy
- Easy migration
- Multiple levels of route reflectors

Configuring a Route Reflector

```
router bgp 100
 neighbor 1.1.1.1 remote-as 100
 neighbor 1.1.1.1 route-reflector-client
 neighbor 2.2.2.2 remote-as 100
 neighbor 2.2.2.2 route-reflector-client
 neighbor 3.3.3.3 remote-as 100
 neighbor 3.3.3.3 route-reflector-client
```