BGP Techniques for Internet Service Providers

Dawit Birhanu (dawit@cisco.com)

Technical Marketing Engineer

Cisco System

Introduction

- Objective
  - BGP basics
  - Scaling techniques
  - Best practices
- Acknowledgment
  - **Philip Smith – original developer of this tutorial**
- Self-practice lab
  - Get your lab POD # at the end of first session
  - Lab setup will be available for one week

Agenda

2:00 PM – 3:30 PM

- BGP Basics
- Scaling BGP
- Using Communities
- Lab logistics

4:30 PM – 5:30 PM

- Deploying BGP in an ISP network

BGP Basics

Agenda – BGP Basics

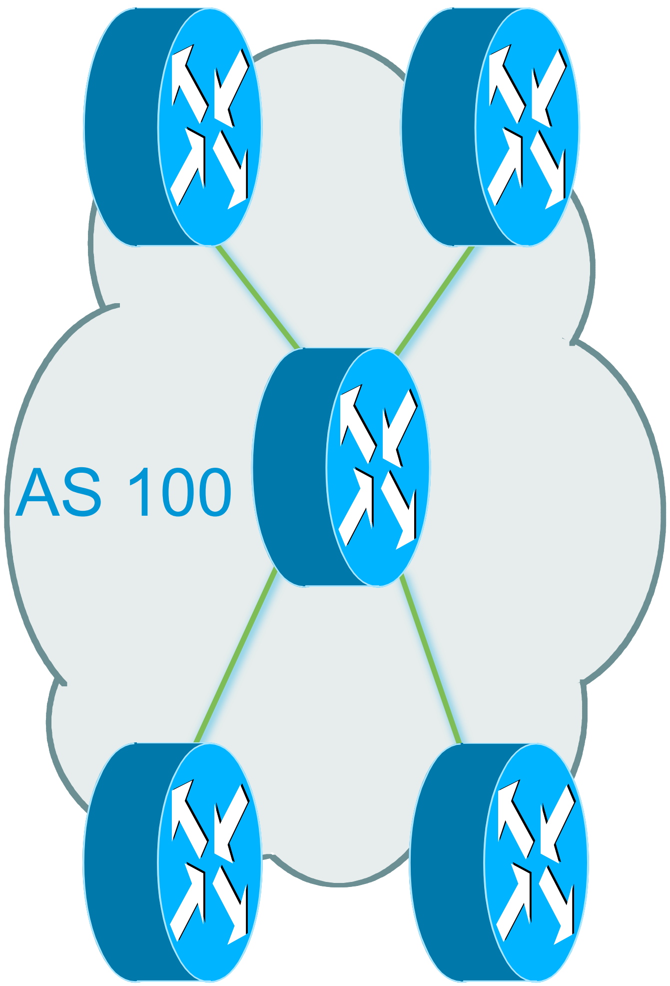
- What is BGP?
- BGP Attributes
- BGP Path Selection Algorithm
- Applying Policy with BGP
- BGP Capabilities

What is BGP?

Border Gateway Protocol

- A Routing Protocol used to exchange routing information between different networks
  - Exterior gateway protocol
- Described in RFC4271
  - RFC4276 gives an implementation report on BGP
  - RFC4277 describes operational experiences using BGP
- IETF Working Groups
  - IDR (Internet-Domain Routing: http://datatracker.ietf.org/wg/idr/)
  - SIDR (Secure IDR: http://datatracker.ietf.org/wg/sidr/)
- The Autonomous System is the cornerstone of BGP
  - It is used to uniquely identify networks with a common routing policy

Autonomous System



- Collection of networks with same routing policy
- Single routing protocol
- Usually under single ownership, trust and administrative control
- Identified by a unique AS number (ASN)
  - 2-octet(16-bit) integer number, or
  - 4-octet (32-bit) integer number
- 4-octet ASN was introduced by RFC4893

Autonomous System Number (ASN)

- Two ranges

0-65535	(original 16-bit range)
65536-4294967295	(32-bit range - RFC4893)

- Usage:

0 and 65535	(reserved)
1-64495	(public Internet)
64496-64511	(documentation - RFC5398)
64512-65534	(private use only)
23456	(represent 32-bit range in 16-bit world)
65536-65551	(documentation - RFC5398)
65552-4294967295	(public Internet)

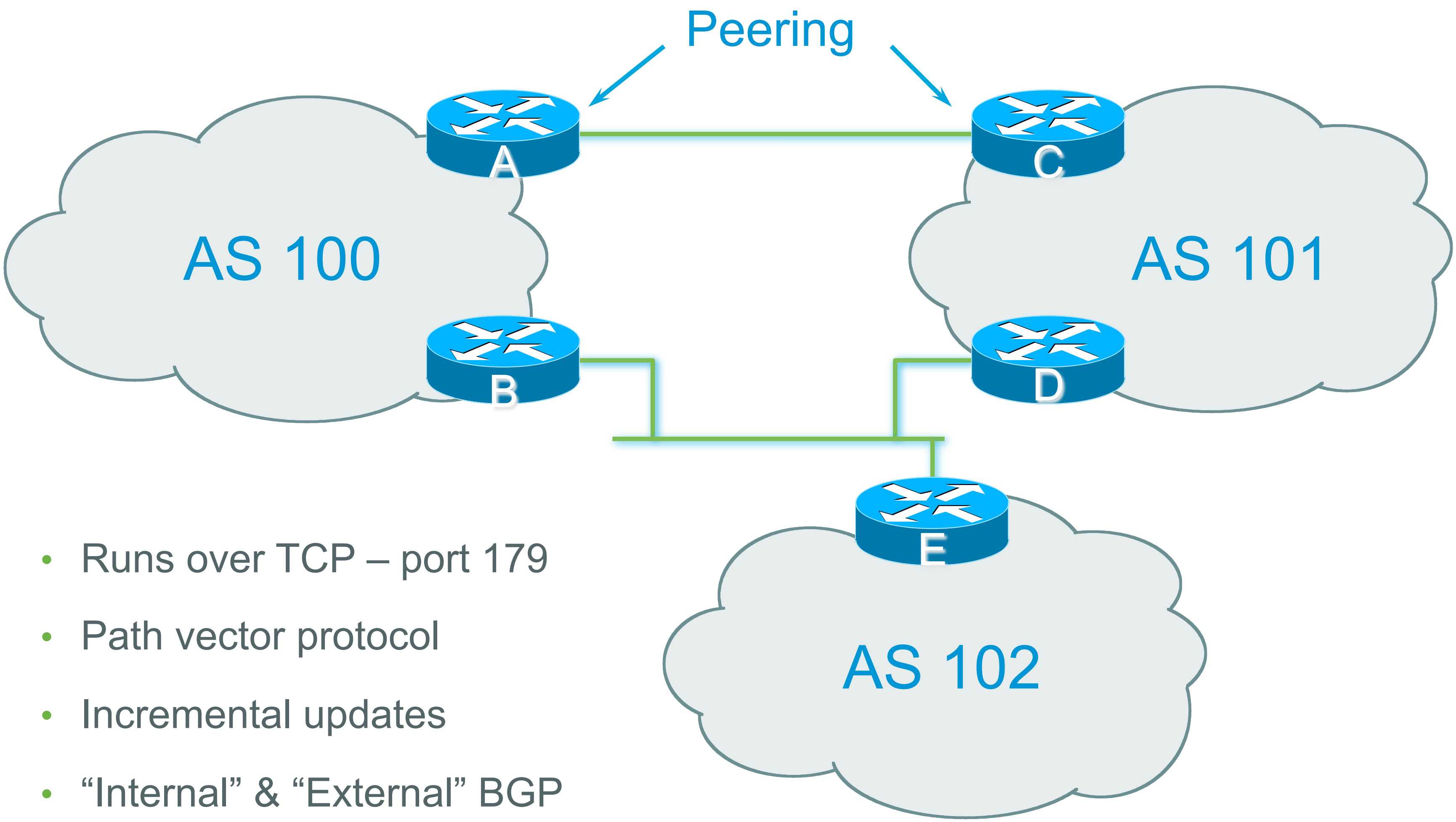
- 32-bit range representation specified in RFC5396

  Defines “asplain” (traditional format) as standard notation

Autonomous System Number (ASN)

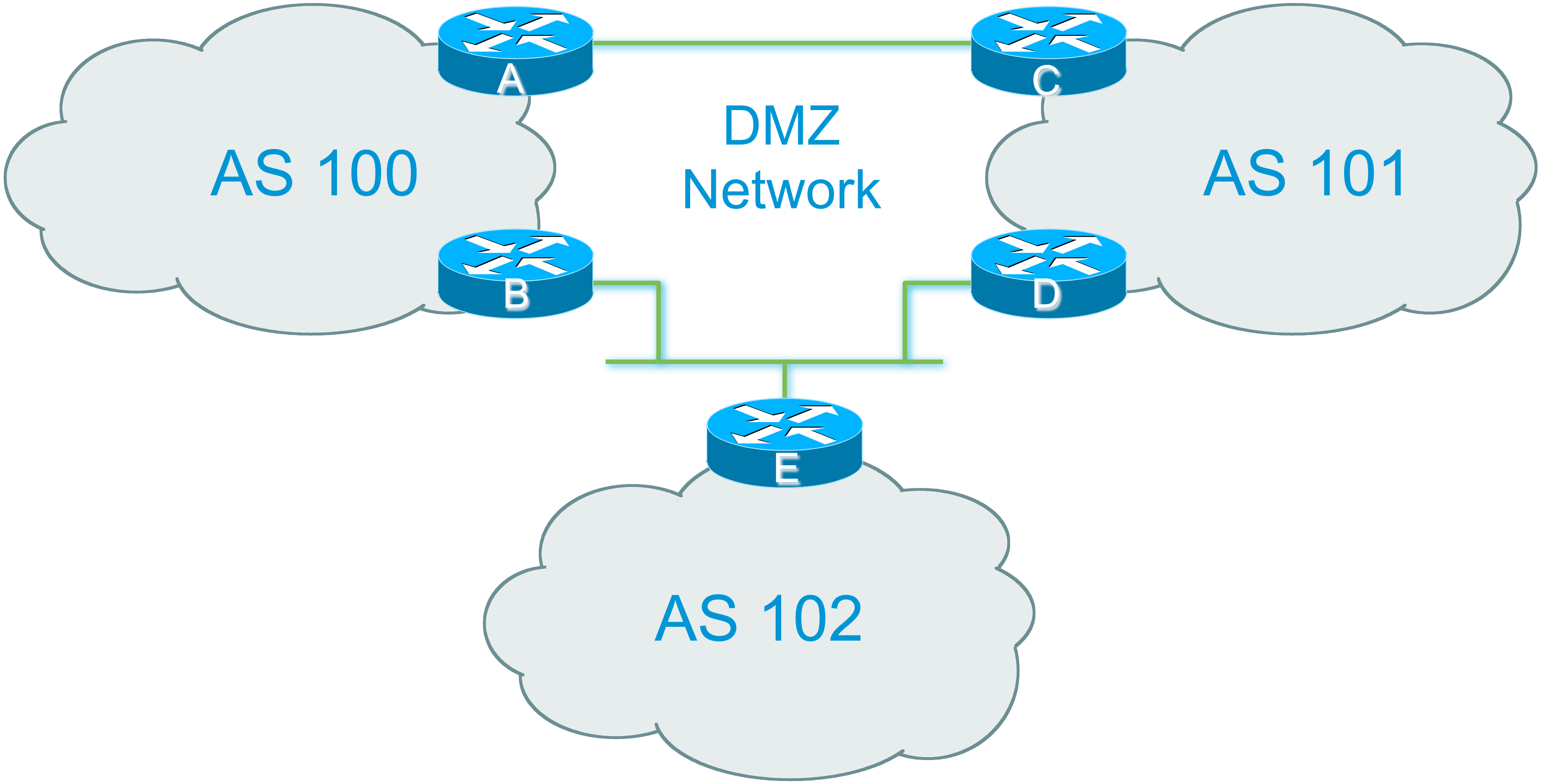
- ASNs are distributed by the Regional Internet Registries
  - They are also available from upstream ISPs who are members of one of the RIRs
- Current 16-bit (2-octet) ASN allocations up to 61438 have been made to the RIRs
  - Around 38860 are visible on the Internet
- Each RIR has also received a block of 32-bit (4-octet) ASNs
  - Out of 8192 assignments, around 3029 (October 2012) are visible on the Internet
- See **www.iana.org/assignments/as-numbers** and http://www.potaroo.net/tools/asn32/

BGP Basics



- Runs over TCP – port 179
- Path vector protocol
- Incremental updates
- "Internal" & "External" BGP

Demarcation Zone (DMZ)



- Shared network between ASes

BGP General Operation

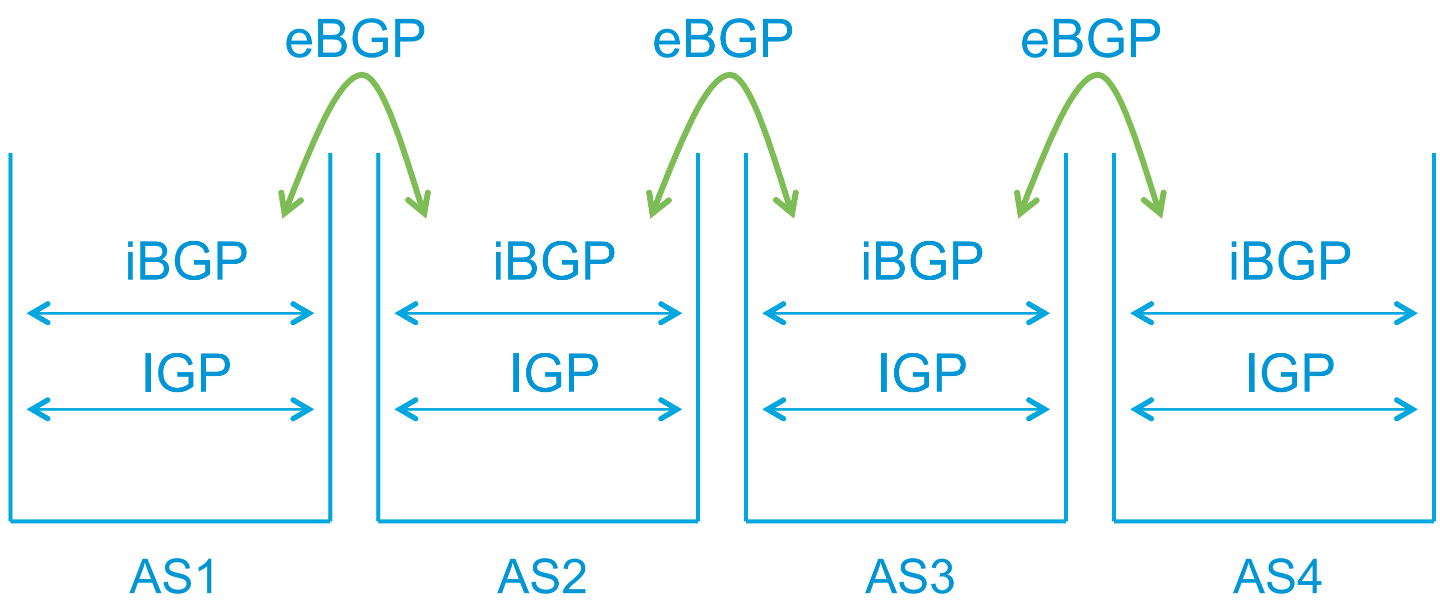
- Learns multiple paths via internal and external BGP speakers
- Picks the best path and installs in the forwarding table
- Best path is sent to external BGP neighbours
- Policies are applied by influencing the best path selection

eBGP & iBGP

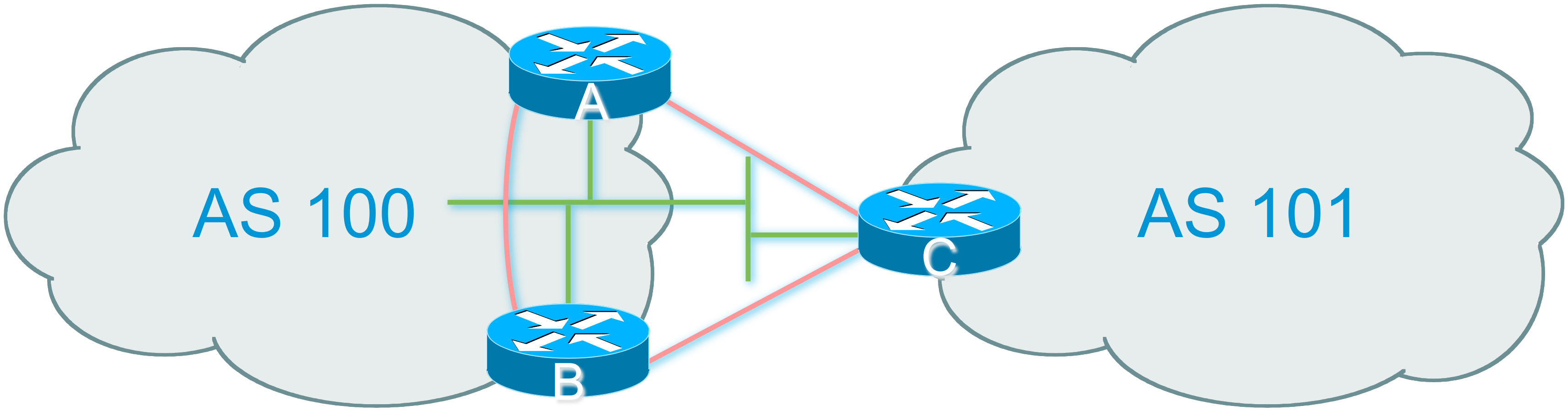
- BGP used internally (iBGP) and externally (eBGP)
- iBGP used to carry
  - Some/all Internet prefixes across ISP backbone
  - ISP's customer prefixes
- eBGP used to
  - Exchange prefixes with other ASes
  - Implement routing policy

BGP/IGP model used in ISP networks

- Model representation



External BGP Peering (eBGP)

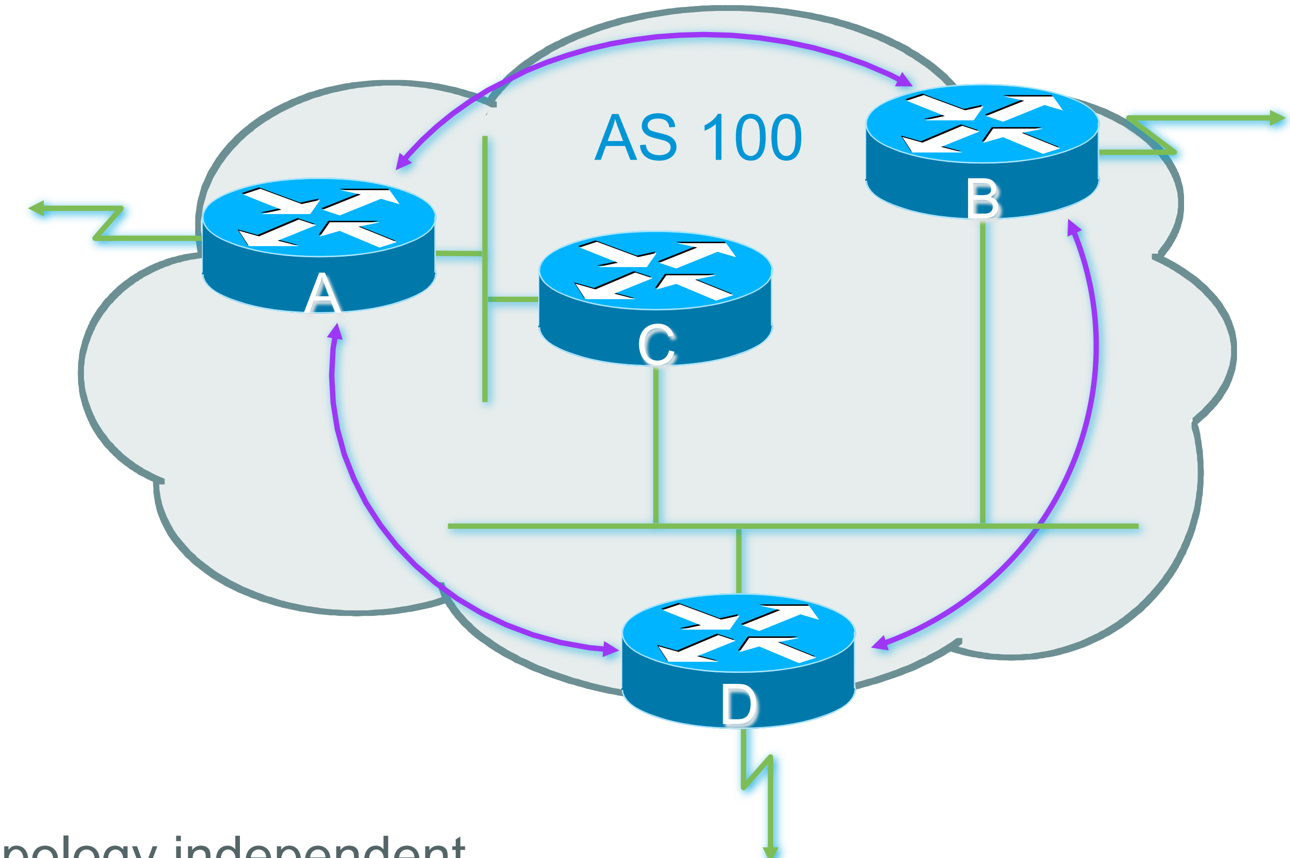


- Between BGP speakers in different AS
- Should be directly connected
- **Never** run an IGP between eBGP peers

Internal BGP (iBGP)

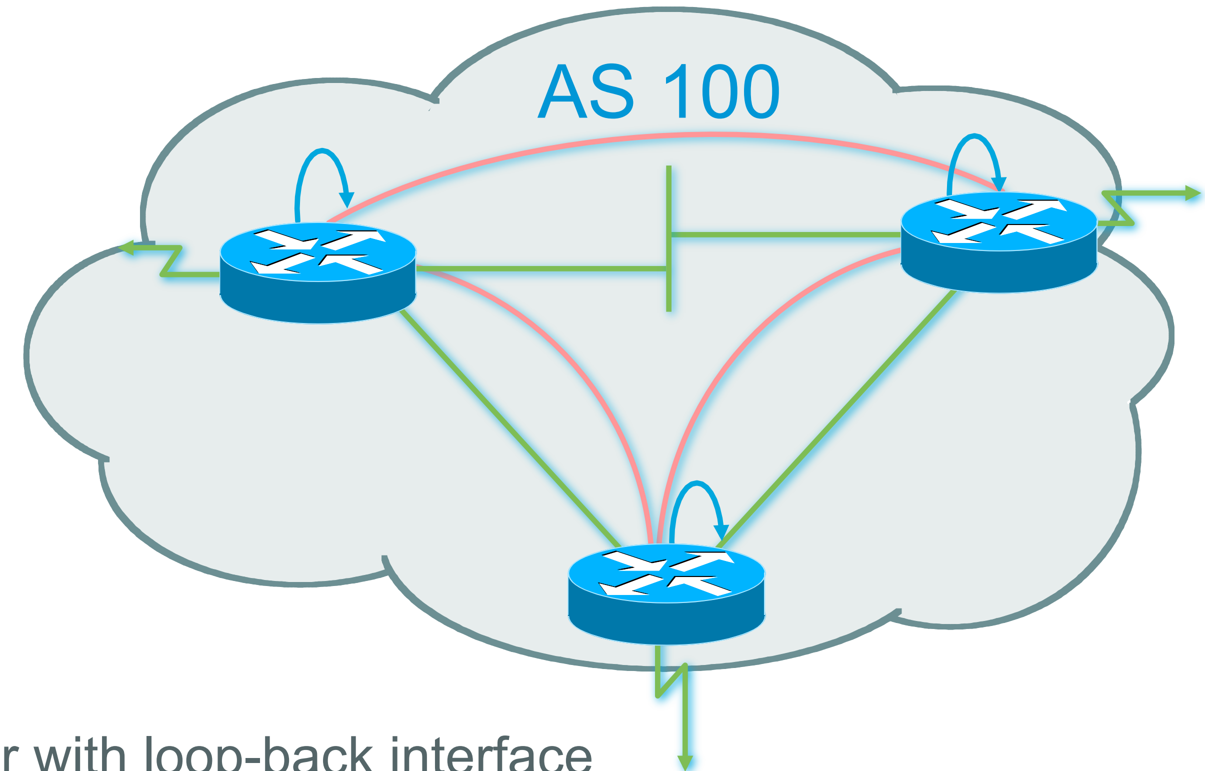
- BGP peer within the same AS
- Not required to be directly connected
  - IGP takes care of inter-BGP speaker connectivity
- iBGP speakers must be fully meshed:
  - They originate connected networks
  - They pass on prefixes learned from outside the ASN
  - They do **not** pass on prefixes learned from other iBGP speakers

Internal BGP Peering (iBGP)



- Topology independent
- Each iBGP speaker must peer with every other iBGP speaker in the AS

Peering to Loopback Interfaces



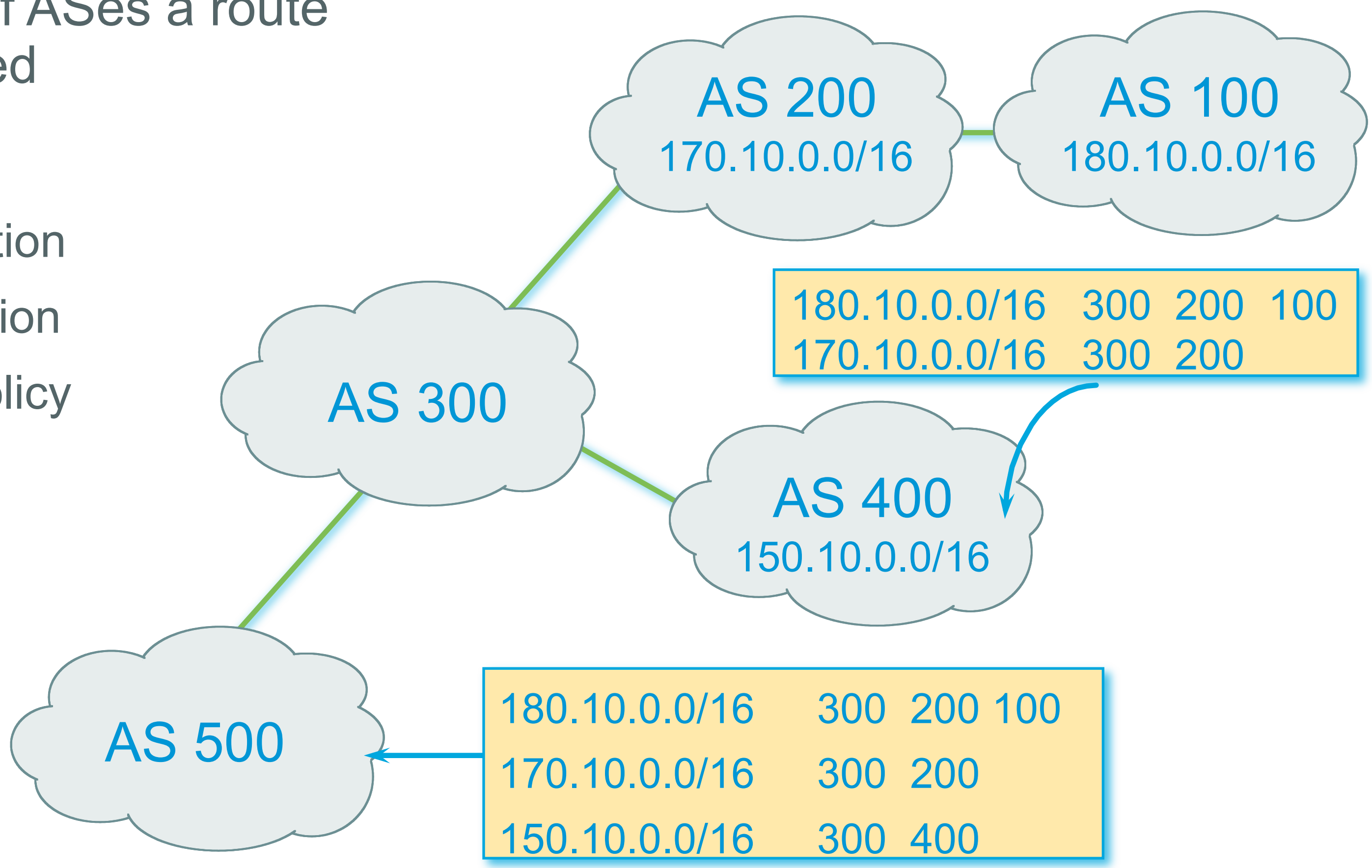
- Peer with loop-back interface
  - Loop-back interface does not go down – ever!
- Do not want iBGP session to depend on state of a single interface or the physical topology

BGP Attributes

Information about BGP

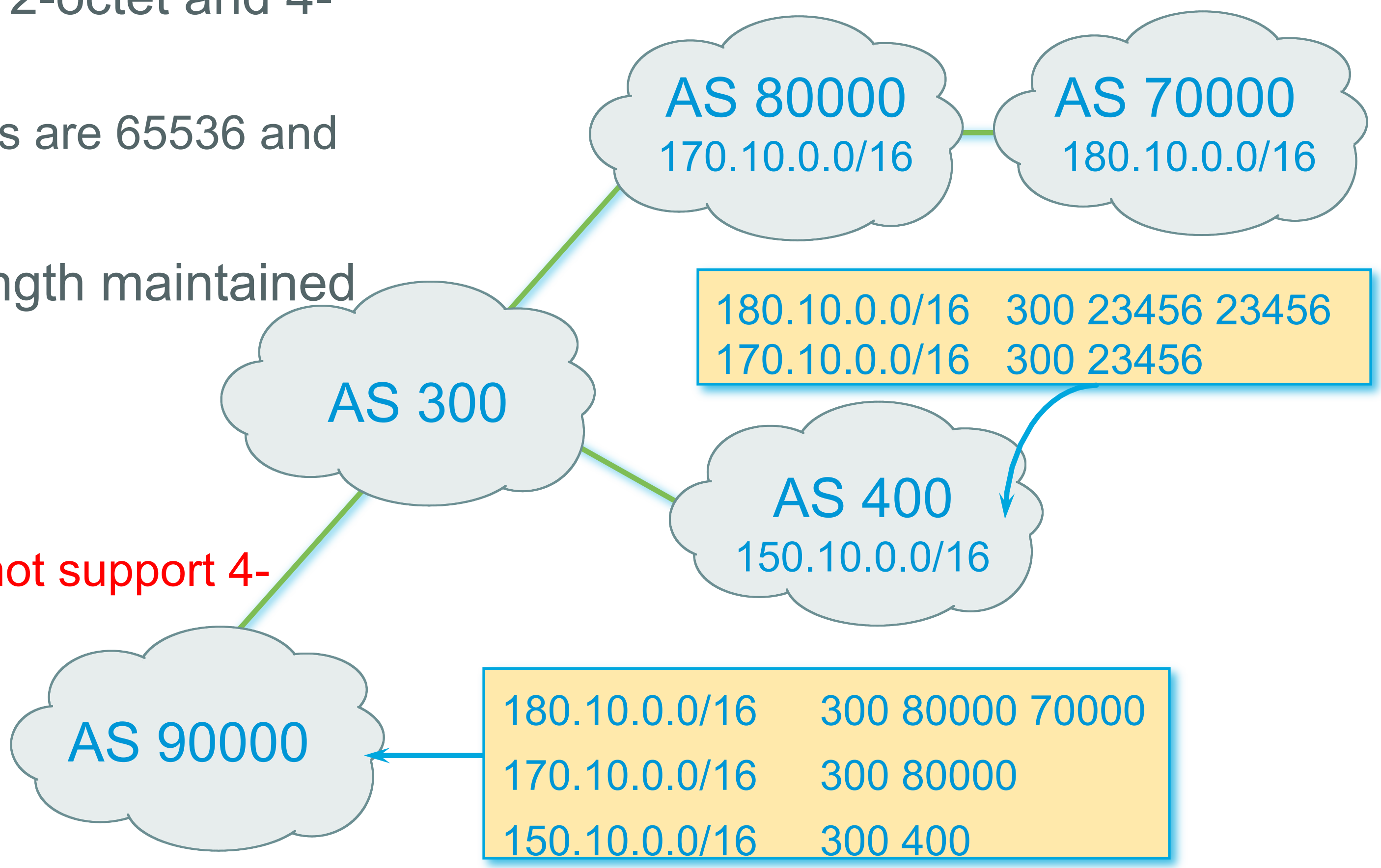
AS-Path

- Sequence of ASes a route has traversed
- Used for:
  - Loop detection
  - Path Selection
  - Applying policy

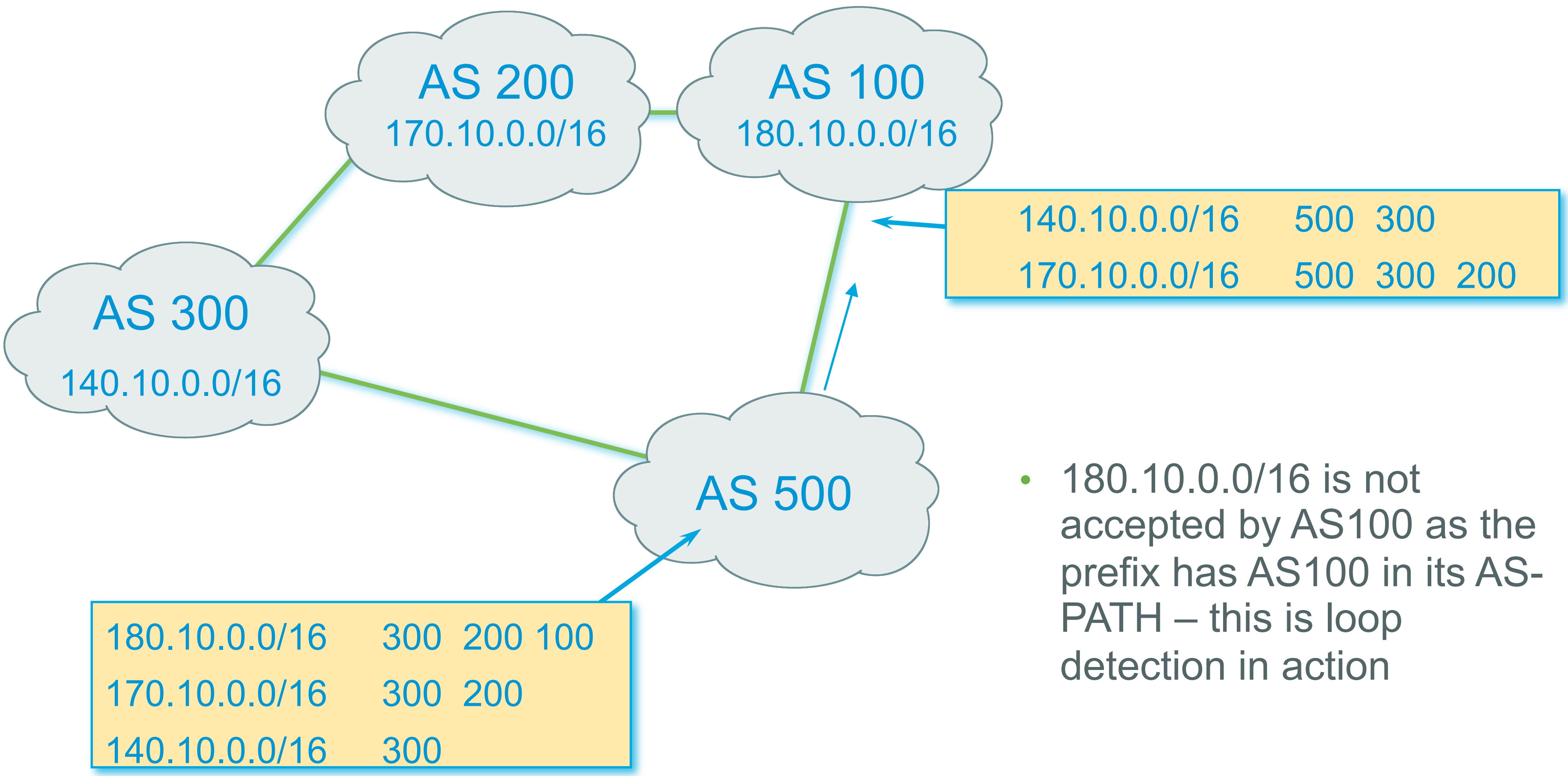


AS-Path (with 2 and 4-octet ASNs)

- Internet with 2-octet and 4-octet ASNs
  - 4-octet ASNs are 65536 and above
- AS-PATH length maintained
- AS400 does not support 4-octet ASN

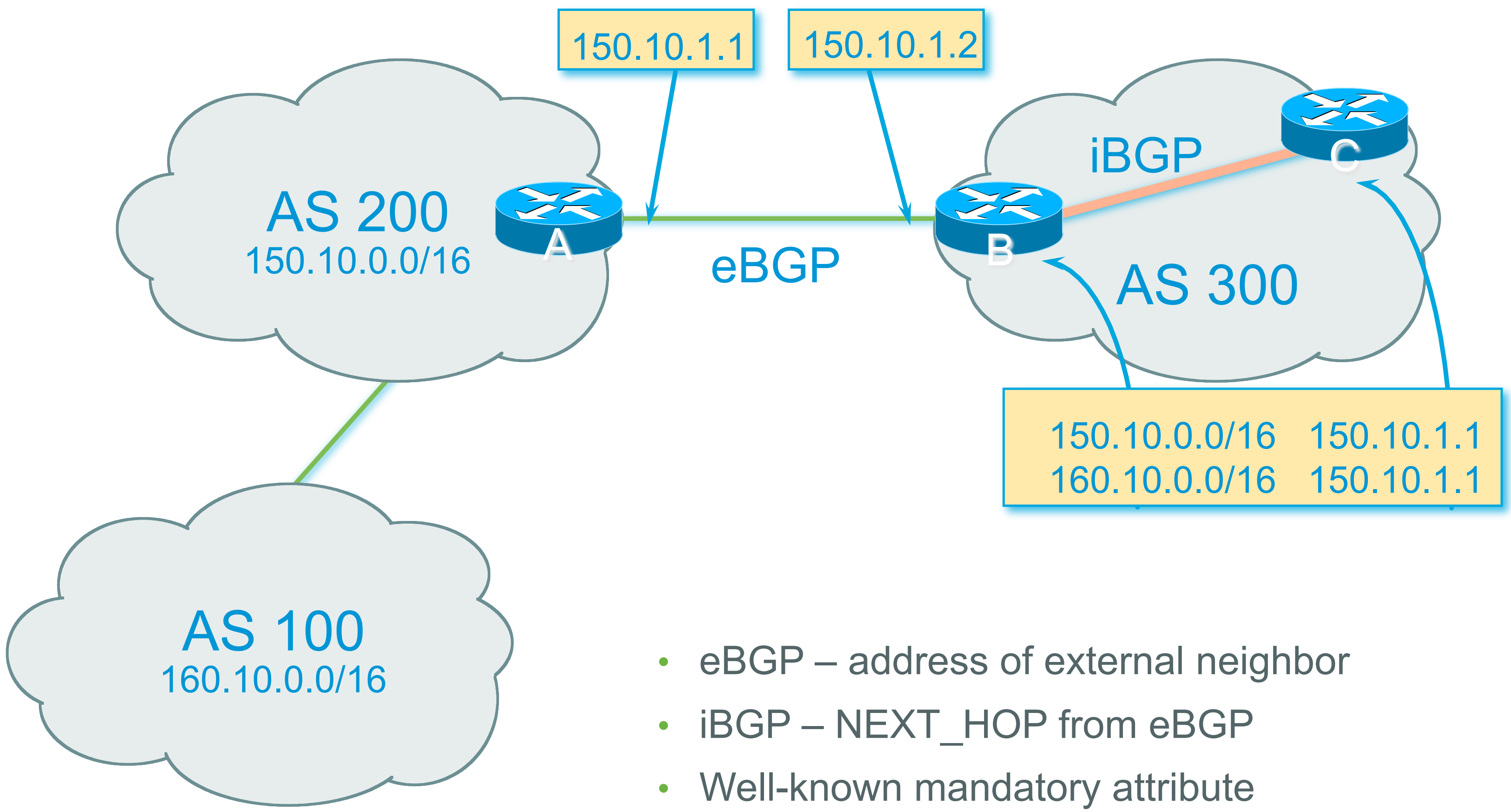


AS-Path loop detection



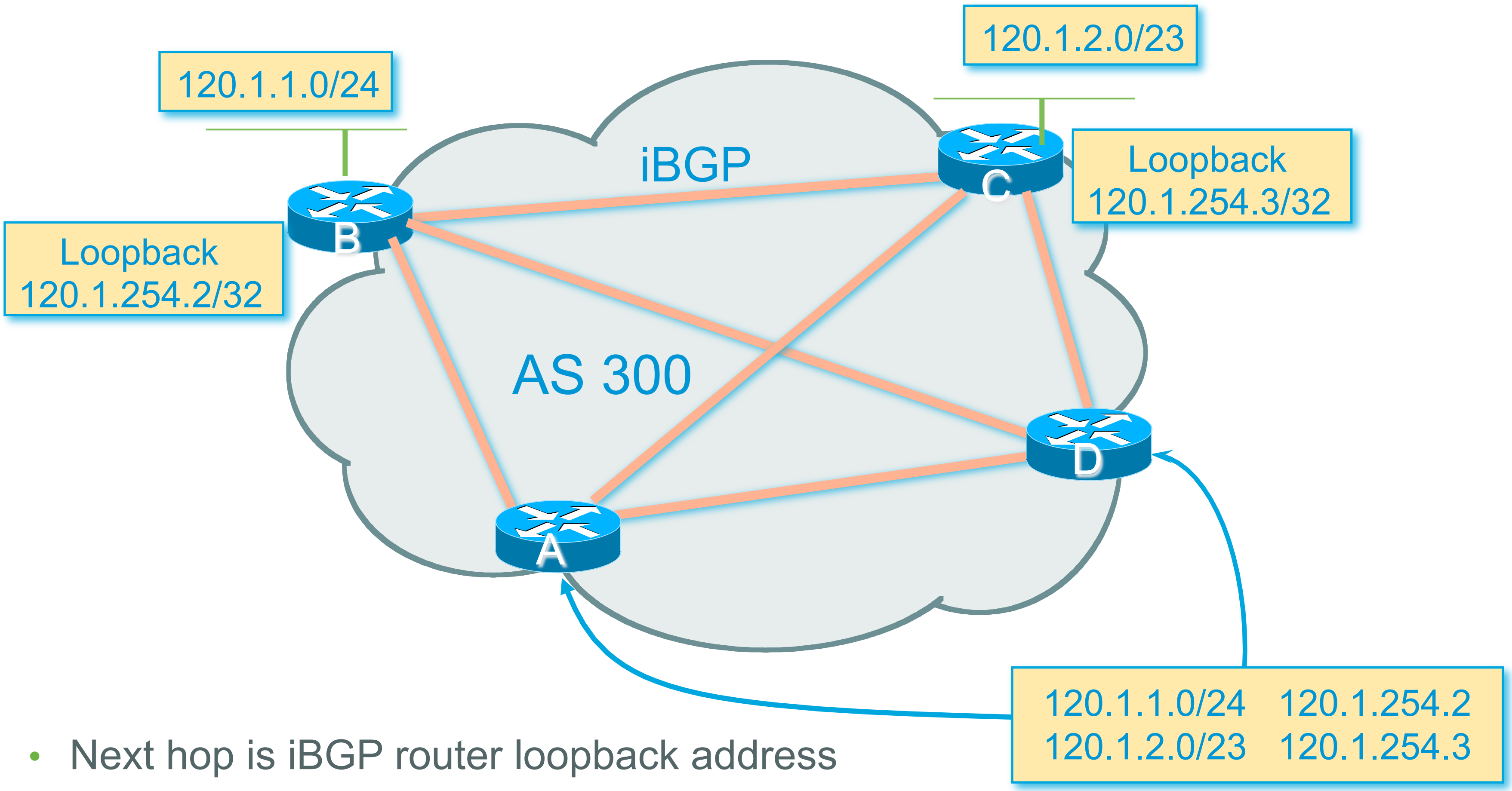
- 180.10.0.0/16 is not accepted by AS100 as the prefix has AS100 in its AS-PATH – this is loop detection in action

Next Hop



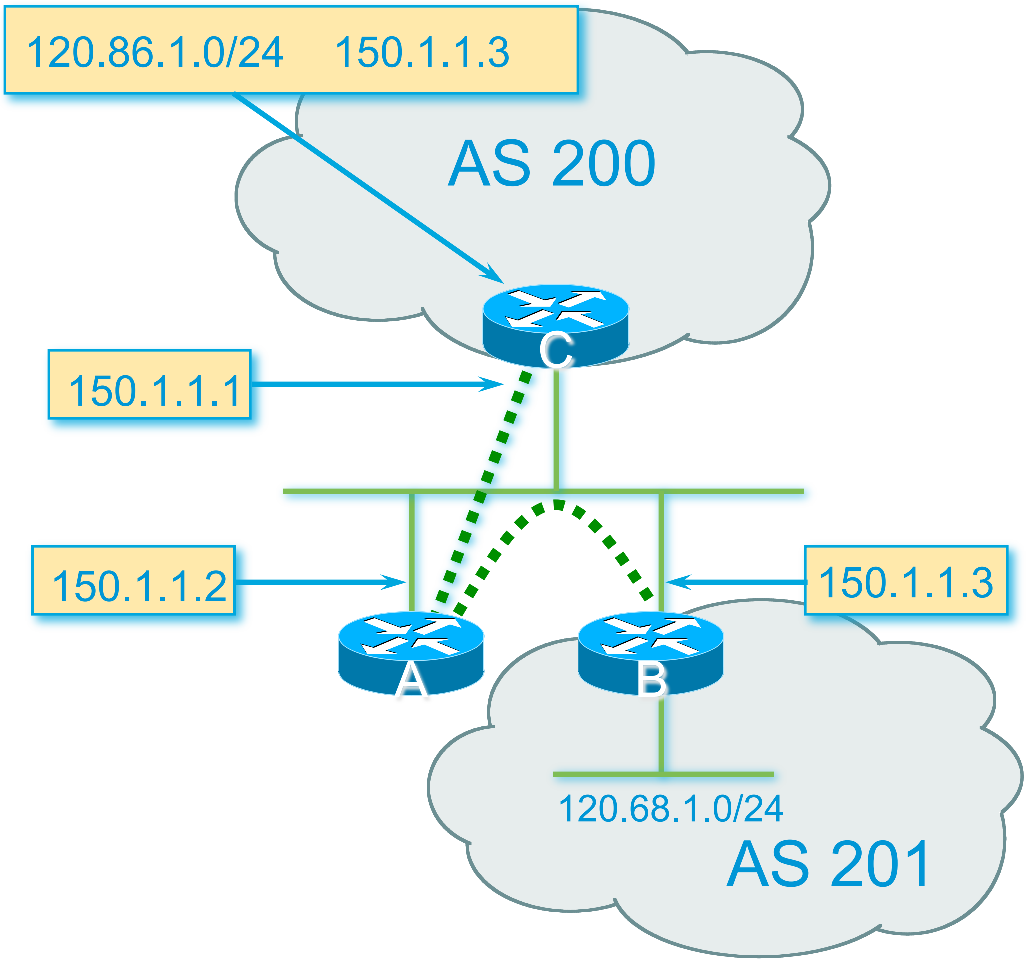
- eBGP – address of external neighbor
- iBGP – NEXT_HOP from eBGP
- Well-known mandatory attribute

iBGP Next Hop



- Next hop is iBGP router loopback address
- Recursive route look-up

Third Party Next Hop



- eBGP between Router A and Router C
- eBGP between RouterA and RouterB
- 120.68.1/24 prefix has next hop address of 150.1.1.3 – this is passed on to RouterC instead of 150.1.1.2
- More efficient
- No extra config needed

Next Hop Best Practice

- BGP default is for external next-hop to be propagated unchanged to iBGP peers
  - This means that IGP has to carry external next-hops
  - Forgetting means external network is invisible
  - With many eBGP peers, it is unnecessary extra load on IGP
- ISP Best Practice is to change external next-hop to be that of the local router

Next Hop (Summary)

- IGP should carry route to next hops
- Recursive route look-up
- Unlinks BGP from actual physical topology
- Change external next hops to that of local router
- Allows IGP to make intelligent forwarding decision

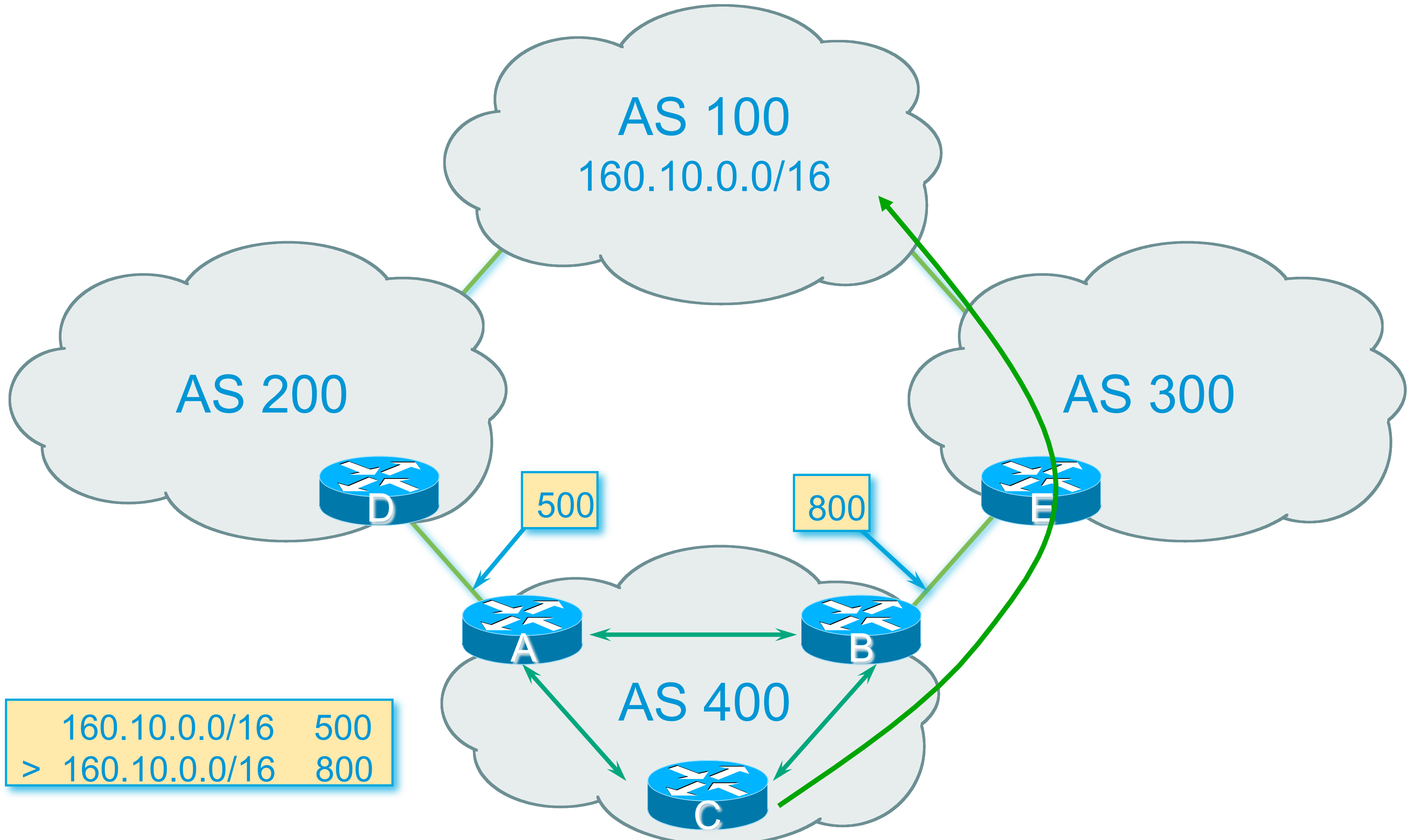
Origin

- Conveys the origin of the prefix
- Historical attribute
  - Used in transition from EGP to BGP
- Transitive and Mandatory Attribute
- Influences best path selection
- Three values: IGP, EGP, incomplete
  - IGP – generated by BGP network statement
  - EGP – generated by EGP
  - incomplete – redistributed from another routing protocol

Aggregator

- Conveys the IP address of the router or BGP speaker generating the aggregate route
- Optional & transitive attribute
- Useful for debugging purposes
- Does not influence best path selection

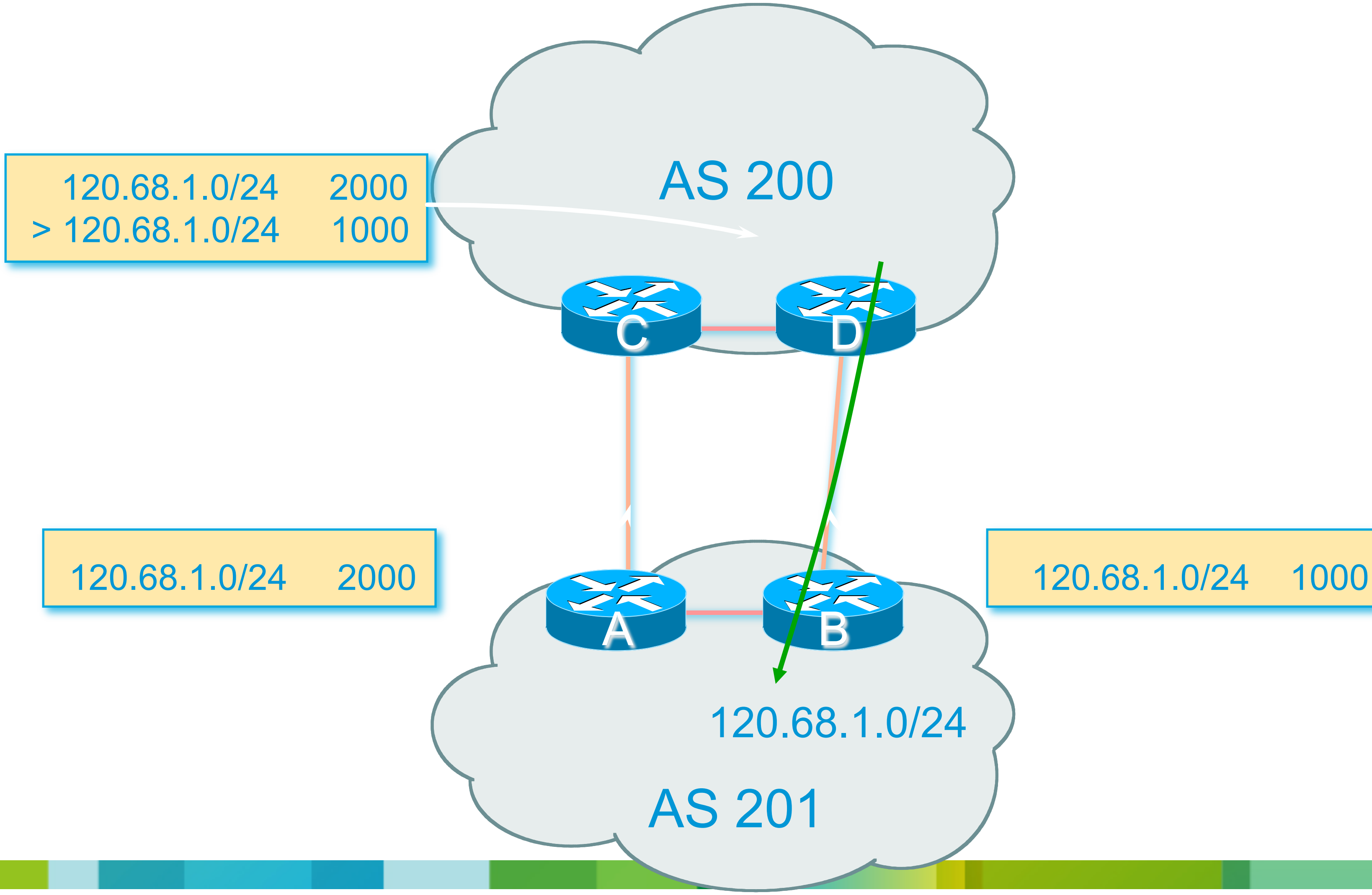
Local Preference



Local Preference

- Non-transitive and optional attribute
- Local to an AS – non-transitive
  - Default local preference is 100 (Cisco IOS)
- Used to influence BGP path selection
  - determines best path for *outbound* traffic
- Path with highest local preference wins

Multi-Exit Discriminator (MED)



Multi-Exit Discriminator

- Inter-AS – non-transitive & optional attribute
- Used to convey the relative preference of entry points
  - determines best path for inbound traffic
- Comparable if paths are from same AS
  - Implementations have a knob to allow comparisons of MEDs from different ASes
- Path with lowest MED wins
- Absence of MED attribute implies MED value of zero (RFC4271)

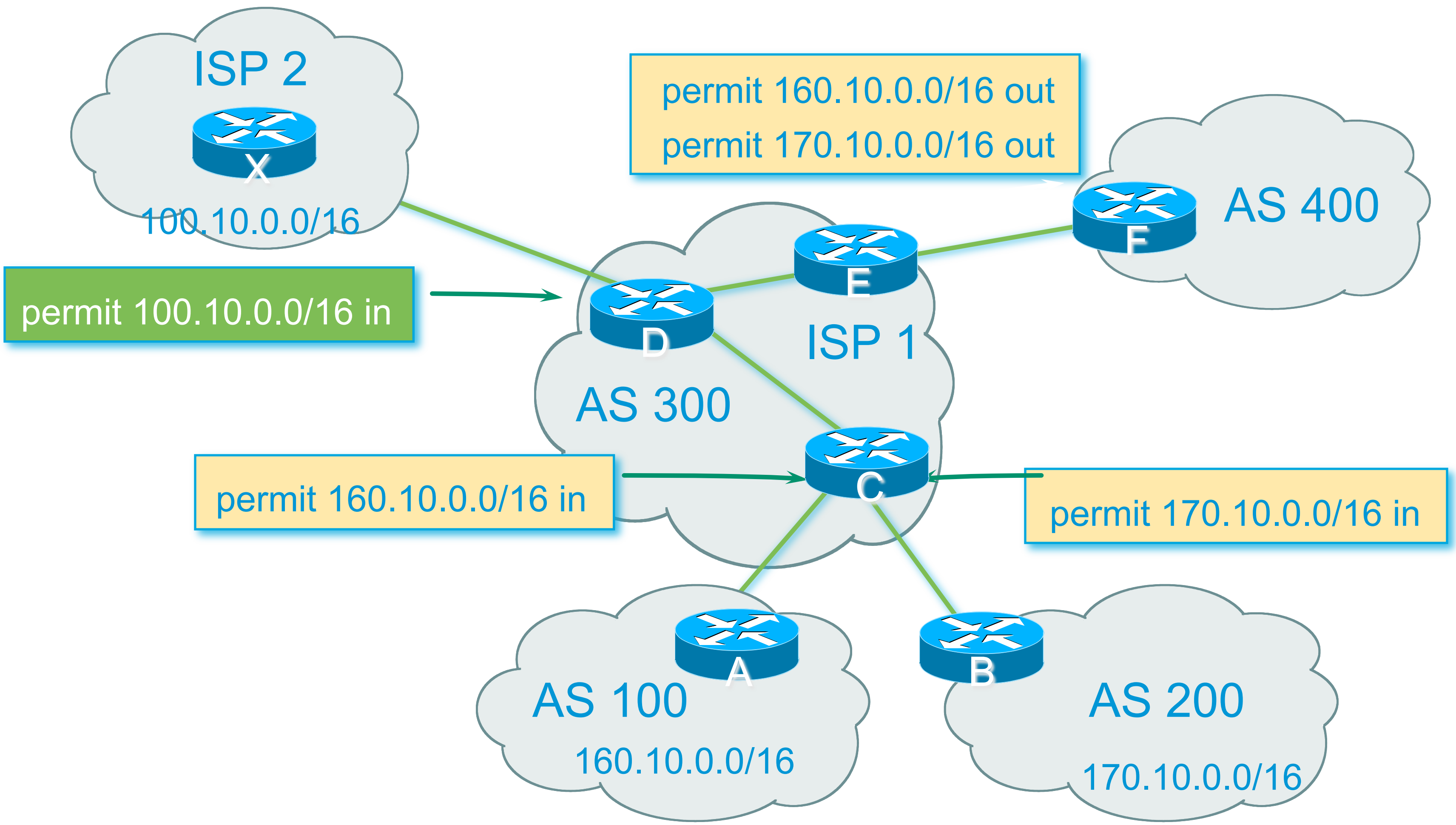
Multi-Exit Discriminator “metric confusion”

- MED is non-transitive and optional attribute
  - Some implementations send learned MEDs to iBGP peers by default, others do not
  - Some implementations send MEDs to eBGP peers by default, others do not
- Default metric varies according to vendor implementation
  - Original BGP spec (RFC1771) made no recommendation
  - Some implementations handled absence of metric as meaning a metric of 0
  - Other implementations handled the absence of metric as meaning a metric of 2^{32}-1 (highest possible) or 2^{32}-2
  - Potential for “metric confusion”

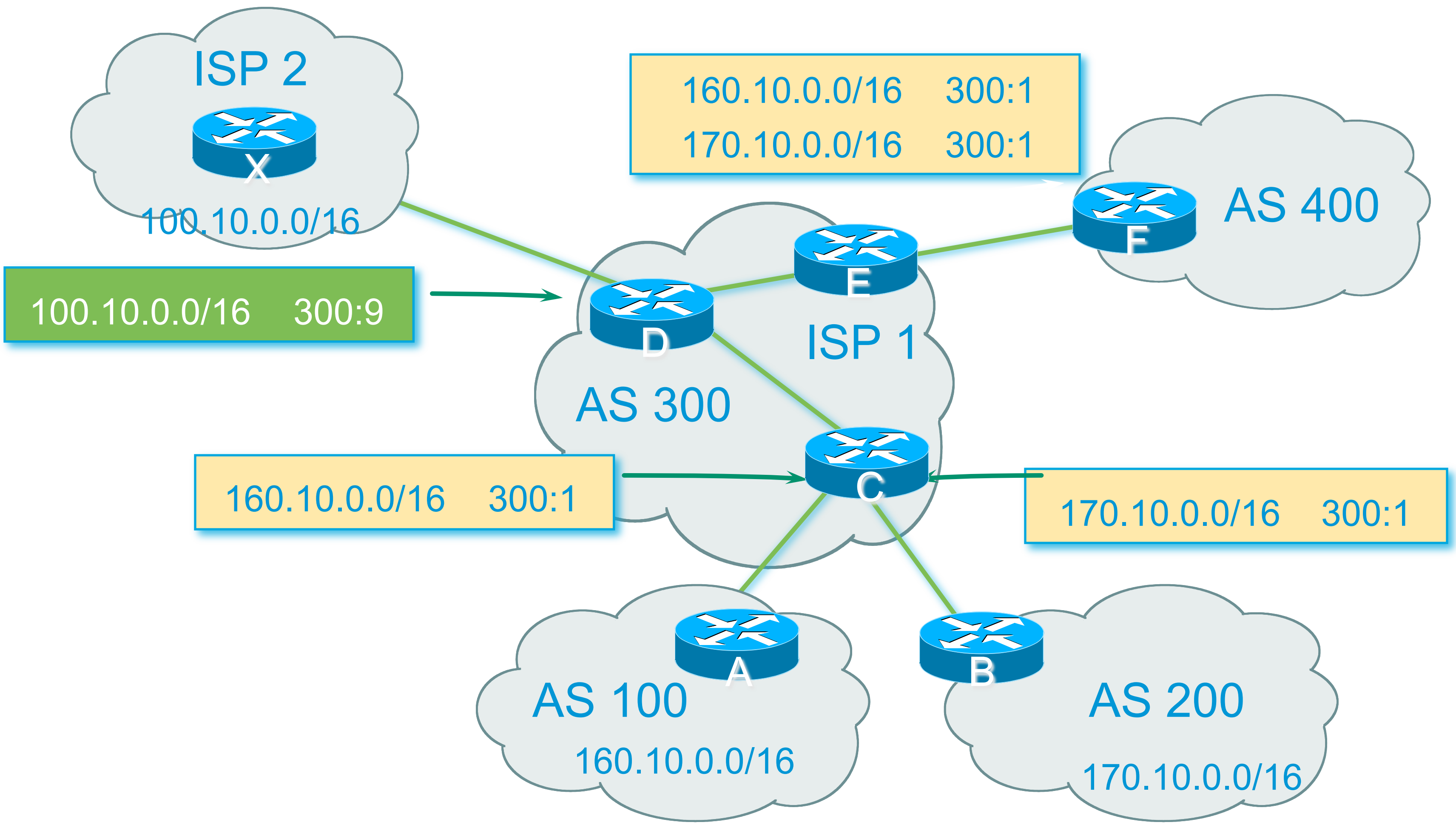
Community

- Communities are described in RFC1997
  - Transitive and Optional Attribute
- 32 bit integer
  - Represented as two 16 bit integers (RFC1998)
  - Common format is <local-ASN>:xx
  - 0:0 to 0:65535 and 65535:0 to 65535:65535 are reserved
- Used to group destinations
  - Each destination could be member of multiple communities
- Very useful in applying policies within and between ASes

Community Example (before)



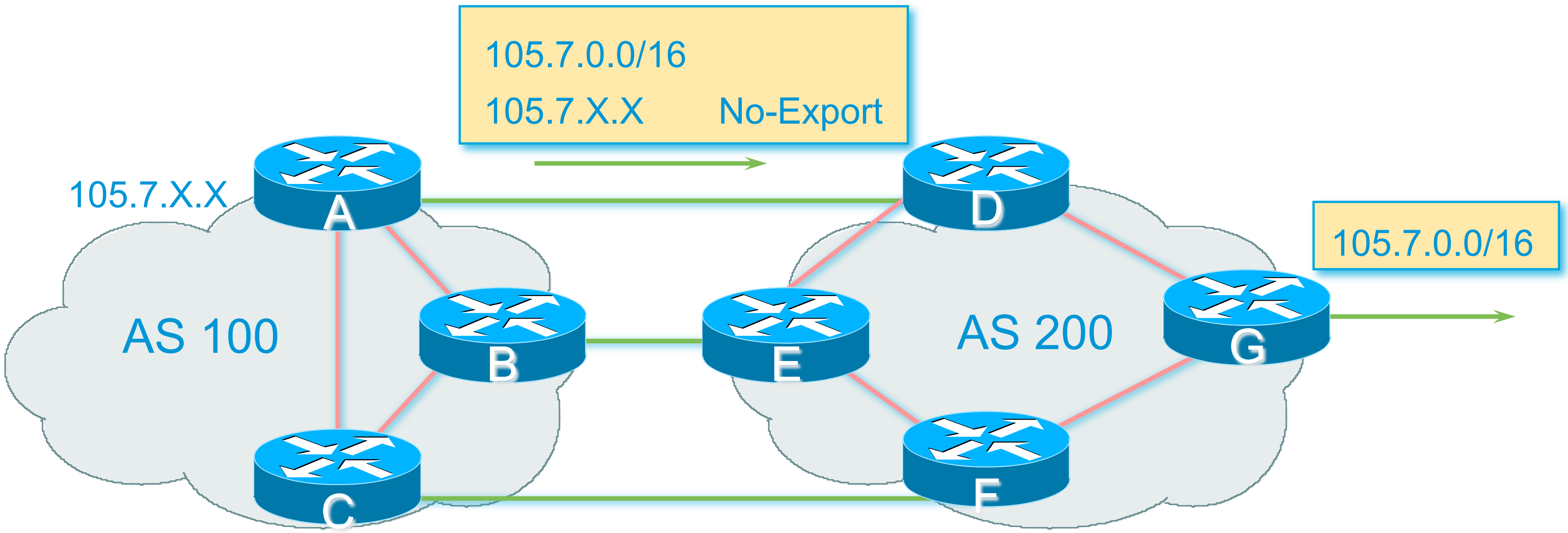
Community Example (after)



Well-Known Communities

- Several well known communities
  www.iana.org/assignments/bgp-well-known-communities
- no-export 65535:65281
  do not advertise to any eBGP peers
- no-advertise 65535:65282
  do not advertise to any BGP peer
- no-export-subconfed 65535:65283
  do not advertise outside local AS (only used with confederations)
- no-peer 65535:65284
  do not advertise to bi-lateral peers (RFC3765)

No-Export Community

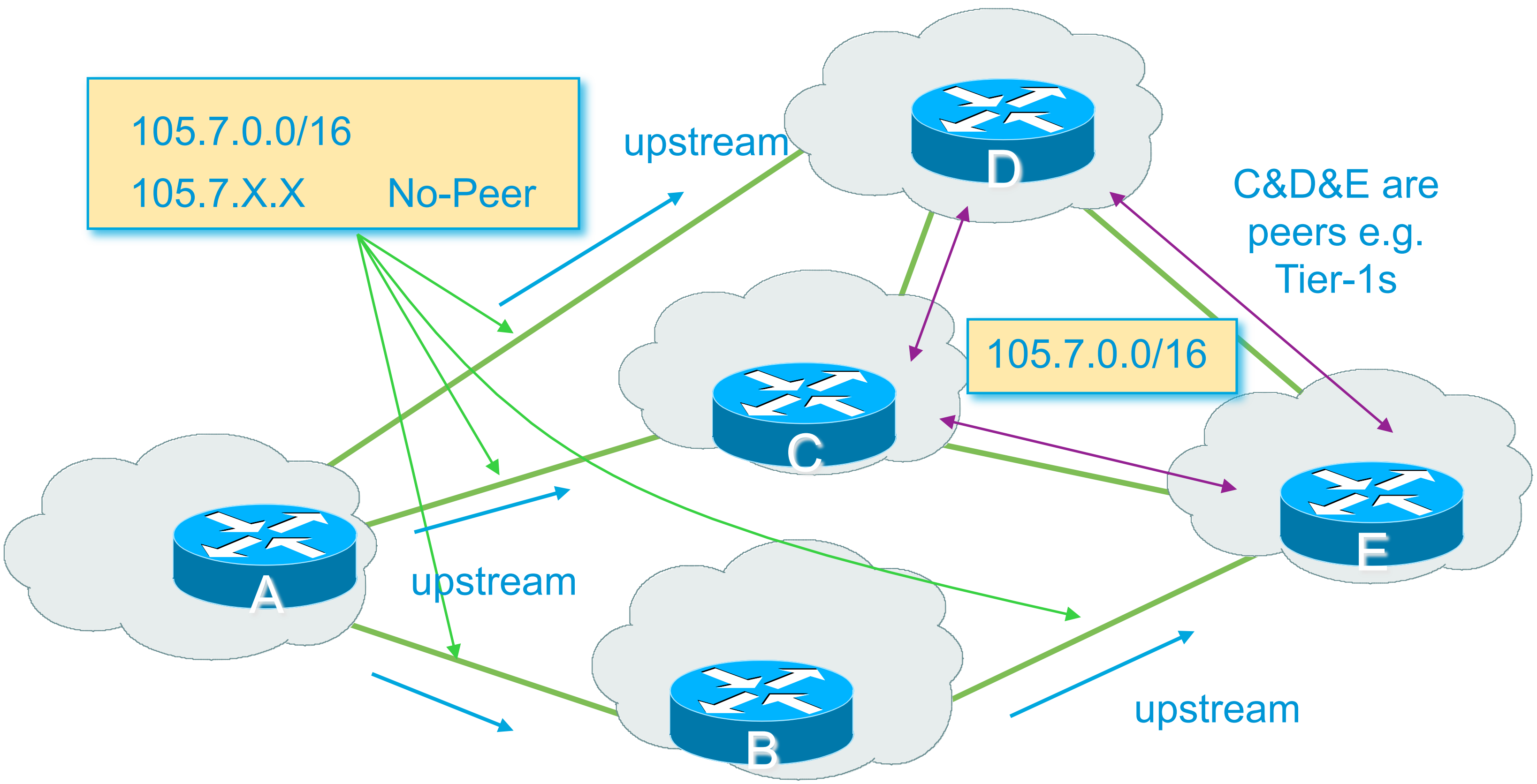


- AS100 announces aggregate and subprefixes

  Intention is to improve loadsharing by leaking subprefixes

- Subprefixes marked with no-export community
- Router G in AS200 does not announce prefixes with no-export community set

No-Peer Community



- Sub-prefixes marked with no-peer community are not sent to bi-lateral peers

  They are only sent to upstream providers

What about 4-byte ASNs?

- Communities are widely used for encoding ISP routing policy
  - 32 bit attribute
- RFC1998 format is now "standard" practice
  - ***ASN:number***
- Fine for 2-byte ASNs, but 4-byte ASNs cannot be encoded
- Solutions:
  - Use "private ASN" for the first 16 bits
  - Use AS_TRANS (23456) for the first 16 bits
  - Wait for http://tools.ietf.org/id/draft-ietf-idr-as4octet-extcomm-generic-subtype-05.txt to be implemented

Community Implementation details

- Community is an optional attribute
  - Some implementations send communities to iBGP peers by default, some do not
  - Some implementations send communities to eBGP peers by default, some do not
- Being careless can lead to community "confusion"
  - ISPs need consistent community policy within their own networks
  - And they need to inform peers, upstreams and customers about their community expectations

BGP Path Selection Algorithm

BGP Path Selection Algorithm Part One

- Do not consider path if no route to next hop
- Do not consider a path that has the maximum possible MED (2^{32}-1)
- Highest weight (local to router)
- Highest local preference (global within AS)
- Prefer locally originated route
- Shortest AS path
  - Skipped if `bgp bestpath as-path ignore` configured

BGP Path Selection Algorithm Part Two

- Lowest origin code

  IGP < EGP < incomplete

- Lowest Multi-Exit Discriminator (MED)

  Order the paths before comparing

  (BGP spec does not specify in which order the paths should be compared. This means best path depends on order in which the paths are compared.)

  If bgp always-compare-med, then compare for all paths

  otherwise MED only considered if paths are received from the same AS (default)

BGP Path Selection Algorithm Part Three

- Prefer eBGP path over iBGP path
- Path with lowest IGP metric to next-hop
- Lowest router-id (originator-id for reflected routes)
- Shortest Cluster-List

  Client **must** be aware of Route Reflector attributes!

- Lowest neighbor IP address

BGP Path Selection Algorithm

- In multi-vendor environments:
  - Make sure the path selection processes are understood for each brand of equipment
  - Each vendor has slightly different implementations, extra steps, extra features, etc.
  - Watch out for possible MED confusion

Applying Policy with BGP

Controlling Traffic Flow and Traffic Engineering

Applying Policy in BGP: Why?

- Network operators rarely “plug in routers and go”
- External relationships:
  - Control who they peer with
  - Control who they give transit to
  - Control who they get transit from
- Traffic flow control:
  - Efficiently use the scarce infrastructure resources (external link load balancing)
  - Congestion avoidance
  - Terminology: Traffic Engineering

Applying Policy in BGP: How?

- Policies are applied by:
  - Setting BGP attributes (local-pref, MED, AS-PATH, community), thereby influencing the path selection process
  - Advertising or Filtering prefixes
  - Advertising or Filtering prefixes according to ASN and AS-PATHs
  - Advertising or Filtering prefixes according to Community membership

Applying Policy with BGP: Tools

- Most implementations have tools to apply policies to BGP:
  - Prefix manipulation/filtering
  - AS-PATH manipulation/filtering
  - Community Attribute setting and matching
- Implementations also have policy language which can do various match/ set constructs on the attributes of chosen BGP routes

BGP Capabilities

Extending BGP

BGP Capabilities

- Documented in RFC2842
- Capabilities parameters passed in BGP open message
- Unknown or unsupported capabilities will result in NOTIFICATION message
- Codes:

  0 to 63 are assigned by IANA by IETF consensus

  64 to 127 are assigned by IANA “first come first served”

  128 to 255 are vendor specific

BGP Capabilities

Current capabilities are:

0	Reserved	[RFC3392]
1	Multiprotocol Extensions for BGP-4	[RFC4760]
2	Route Refresh Capability for BGP-4	[RFC2918]
3	Outbound Route Filtering Capability	[RFC5291]
4	Multiple routes to a destination capability	[RFC3107]
5	Extended Next Hop Encoding	[RFC5549]
64	Graceful Restart Capability	[RFC4724]
65	Support for 4 octet ASNs	[RFC4893]
66	Deprecated	
67	Support for Dynamic Capability	[ID]
68	Multisession BGP	[ID]
69	Add Path Capability	[ID]
70	Enhanced Route Refresh Capability	[ID]

See www.iana.org/assignments/capability-codes

BGP Capabilities

- Multiprotocol extensions
  - This is a whole different world, allowing BGP to support more than IPv4 unicast routes
  - Examples include: v4 multicast, IPv6, v6 multicast, VPNs
  - Another tutorial (or many!)
- Route refresh is a well known scaling technique – covered shortly
- 32-bit ASNs have recently arrived
- The other capabilities are still in development or not widely implemented or deployed yet

Scaling BGP

Agenda – Scaling BGP

- BGP Scaling Techniques
- Dynamic Reconfigurations
- Route Reflectors
- BGP Confederations
- Deploying 4-octect ASNs

BGP Scaling Techniques

BGP Scaling Techniques

- Original BGP specification and implementation was fine for the Internet of the early 1990s
  - But didn't scale
- Issues as the Internet grew included:
  - Scaling the iBGP mesh beyond a few peers?
  - Implement new policy without causing flaps and route churning?
  - Keep the network stable, scalable, as well as simple?

BGP Scaling Techniques

- Current Best Practice Scaling Techniques
  - Route Refresh
  - Configuration Templates
  - Update Groups
  - Route Reflectors (and Confederations)
  - Route Aggregation
- Deploying 4-octect ASNs
- Deprecated Scaling Techniques
  - Route Flap Damping

Dynamic Reconfiguration

Route Refresh

Route Refresh

- BGP peer reset required after every policy change
  - Because the router does not store prefixes which are rejected by policy
- Hard BGP peer reset:
  - Terminates BGP peering & Consumes CPU
  - Severely disrupts connectivity for all networks
- Soft BGP peer reset without Route Refresh capability
  - BGP peering remains active
  - Router needs to keep full update received from each peer (memory resource intensive)
- Soft BGP peer reset (or Route Refresh):
  - BGP peering remains active
  - Impacts only those prefixes affected by policy change

Route Refresh Capability

- Facilitates non-disruptive policy changes
- For most implementations, no configuration is needed
  - Automatically negotiated at peer establishment
- No additional memory is used
- Requires peering routers to support “route refresh capability” – RFC2918

Dynamic Reconfiguration

- Use Route Refresh capability if supported
  - find out from the BGP neighbour status display
  - Non-disruptive, "Good For the Internet"
- If not supported, see if implementation has a workaround
- Only hard-reset a BGP peering as a last resort
  - Consider the impact to be equivalent to a router reload

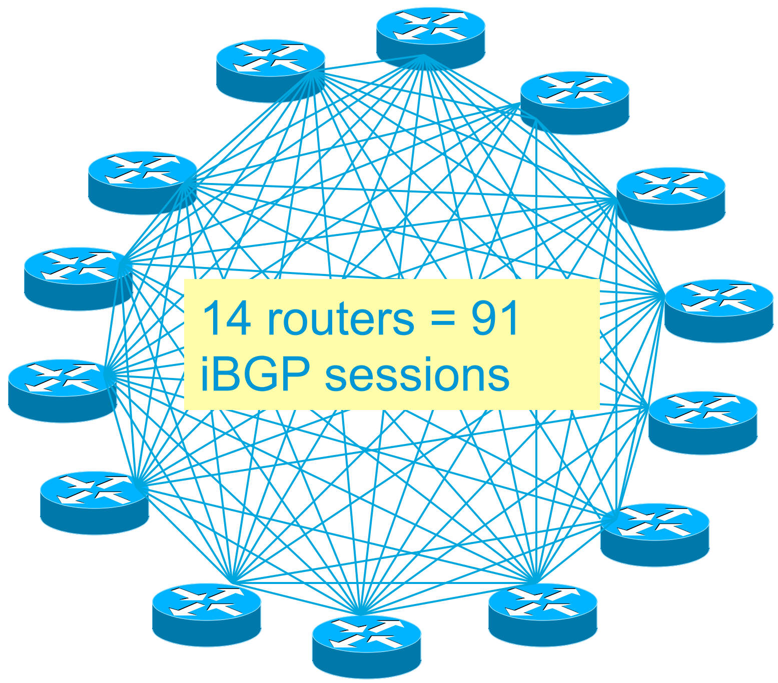
Route Reflectors

Scaling the iBGP mesh

Scaling iBGP mesh

- Avoid n(n-1)/2 iBGP mesh

n=1000 ⇒ nearly half a million ibgp sessions!



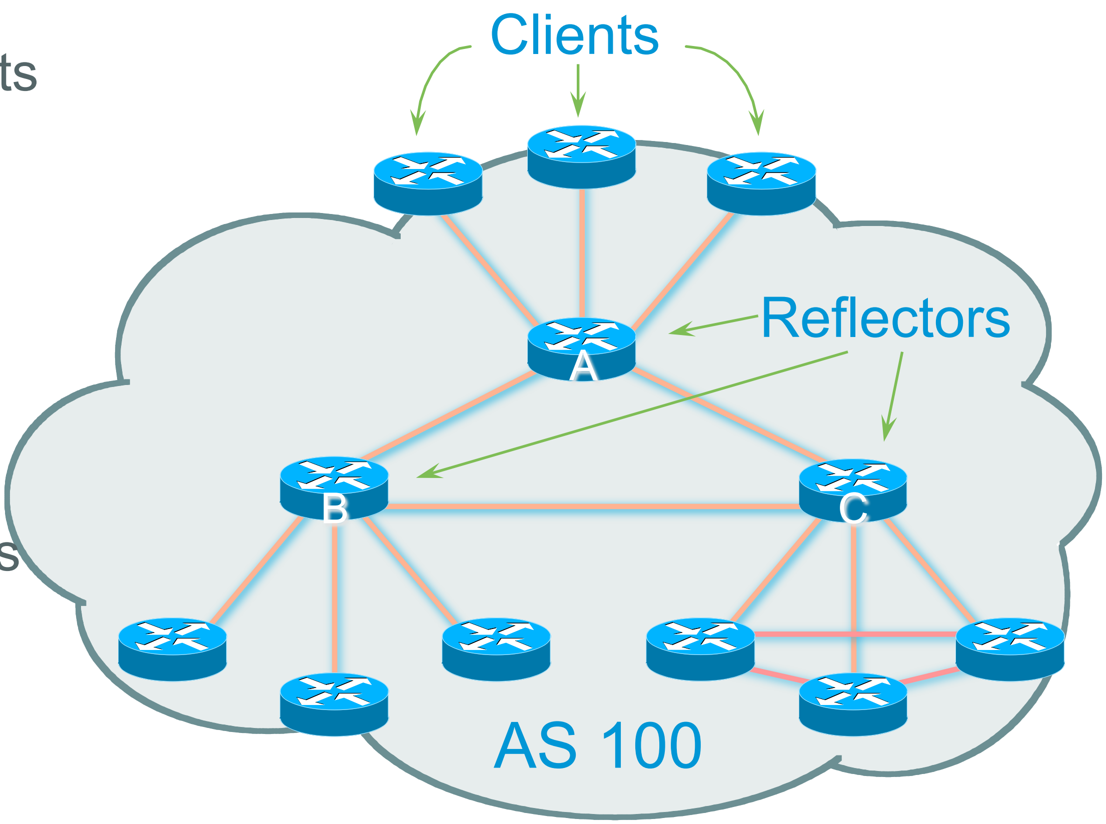
- Two solutions

  Route reflector – simpler to deploy and run

  Confederation – more complex, has corner case advantages

Route Reflector: Principles

- 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 RFC4456



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: Loop Avoidance

- Originator_ID attribute

  Carries the RID of the originator of the route in the local AS (created by the RR)

- Cluster_list attribute

  The local cluster-id is added when the update is sent by the RR

  Best to set cluster-id is from router-id (address of loopback)

  (Some ISPs use their own cluster-id assignment strategy – but needs to be well documented!)

Route Reflector: Redundancy

- Multiple RRs can be configured in the same cluster – not advised!

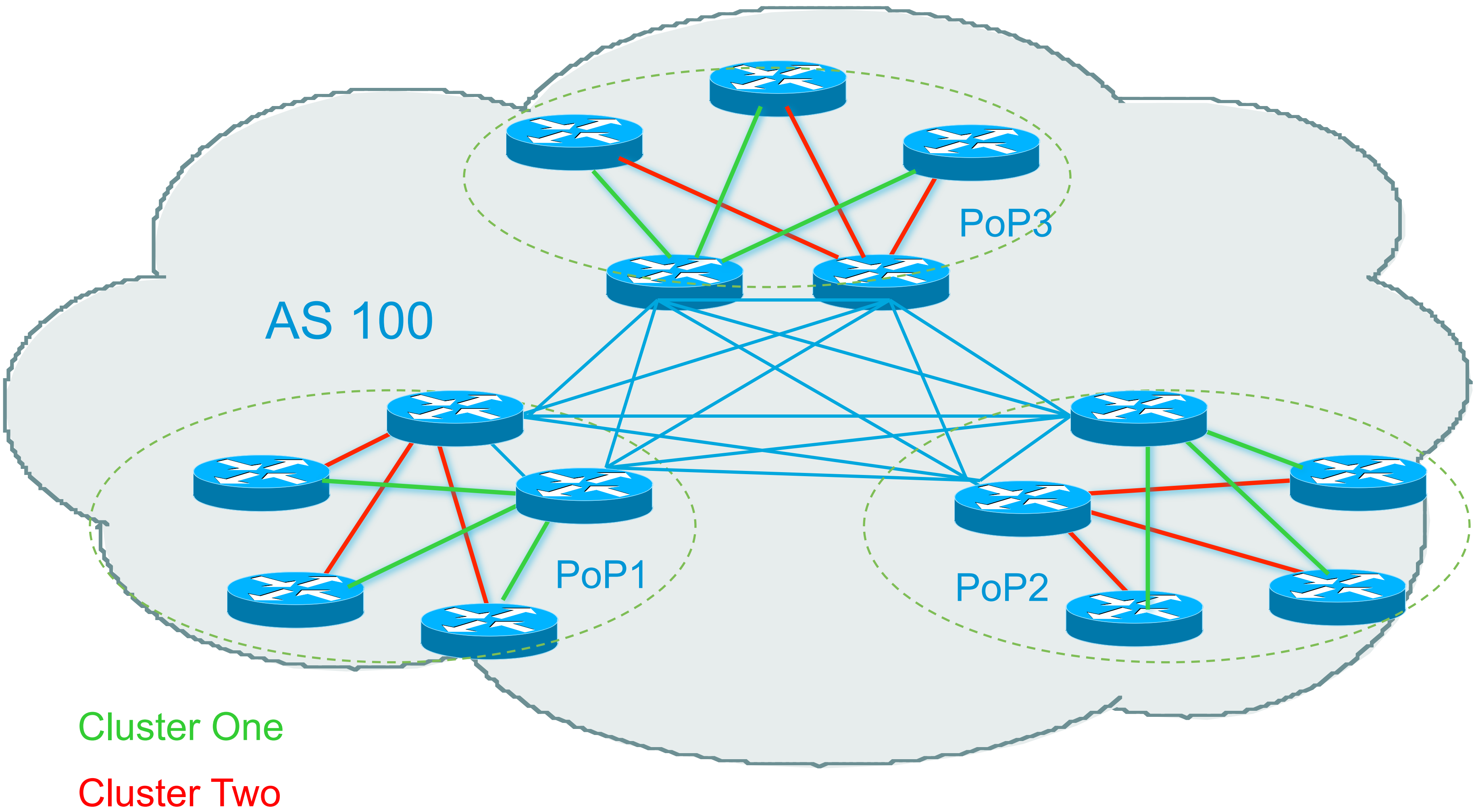
  All RRs in the cluster must have the same cluster-id (otherwise it is a different cluster)

- A router may be a client of RRs in different clusters

  Common today in ISP networks to overlay two clusters – redundancy achieved that way

  → Each client has two RRs = redundancy

Route Reflectors: Redundancy



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

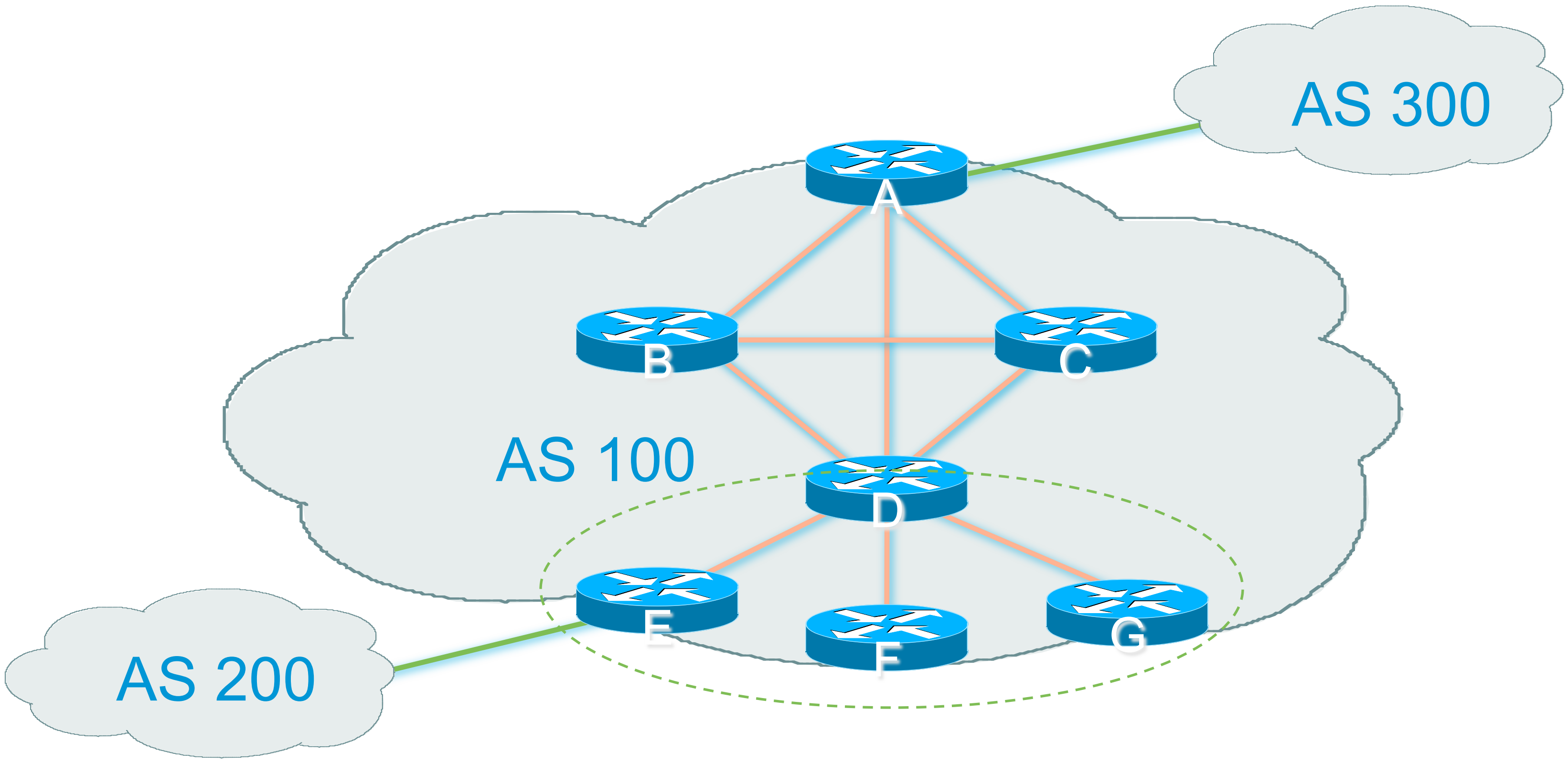
Route Reflector: Deployment

- Where to place the route reflectors?
  - Always follow the physical topology!
  - This will guarantee that the packet forwarding won't be affected
- Typical ISP network:
  - PoP has two core routers
  - Core routers are RR for the PoP
  - Two overlaid clusters

Route Reflector: Migration

- Typical ISP network:
  - Core routers have fully meshed iBGP
  - Create further hierarchy if core mesh too big
    - Split backbone into regions
- Configure one cluster pair at a time
  - Eliminate redundant iBGP sessions
  - Place maximum one RR per cluster
  - Easy migration, multiple levels

Route Reflector: Migration



- Migrate small parts of the network, one part at a time

BGP Confederations

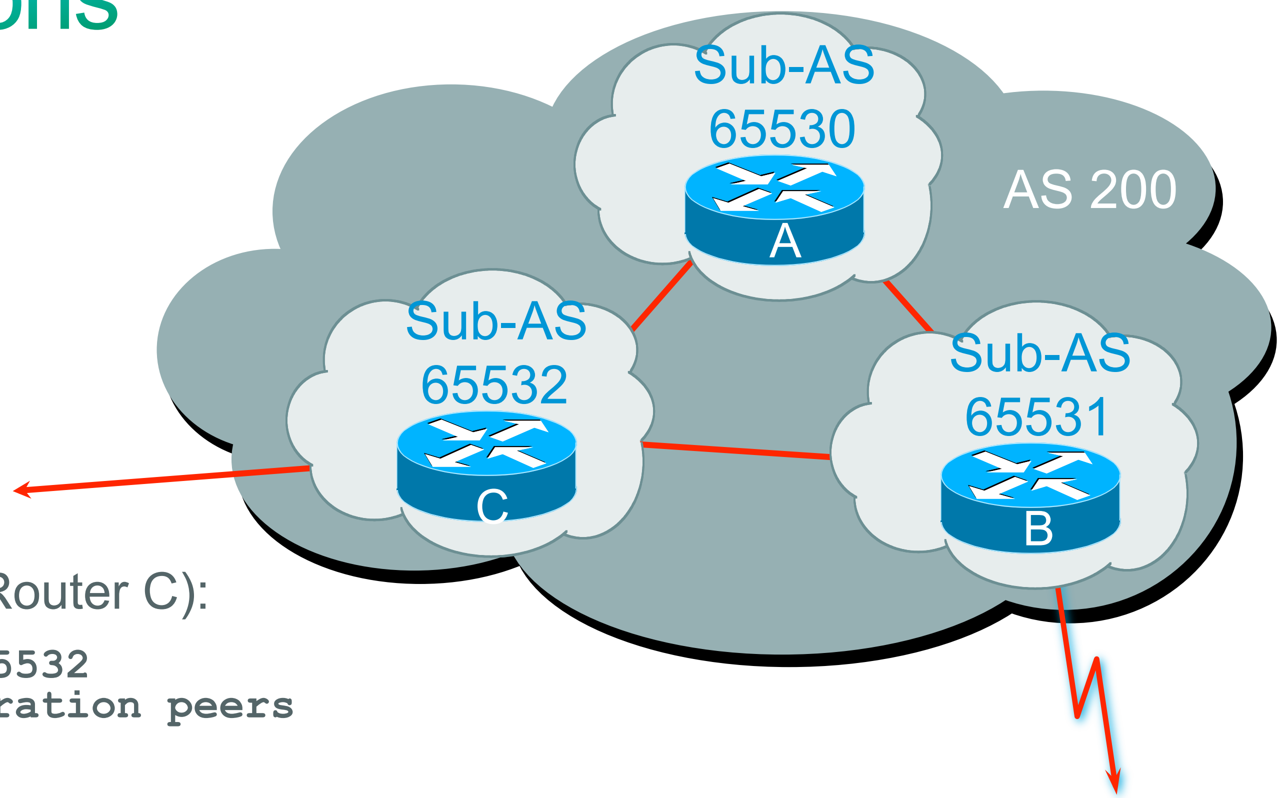
Confederations

- Divide the AS into sub-AS
  - eBGP between sub-AS, but some iBGP information is kept
    - Preserve NEXT_HOP across the sub-AS (IGP carries this information)
    - Preserve LOCAL_PREF and MED
- Usually a single IGP
- Described in RFC5065

Confederations (Cont.)

- Visible to outside world as single AS – "Confederation Identifier"
  - Each sub-AS uses a number from the private AS range (64512-65534)
- iBGP speakers in each sub-AS are fully meshed
  - The total number of neighbours is reduced by limiting the full mesh requirement to only the peers in the sub-AS
  - Can also use Route-Reflector within sub-AS

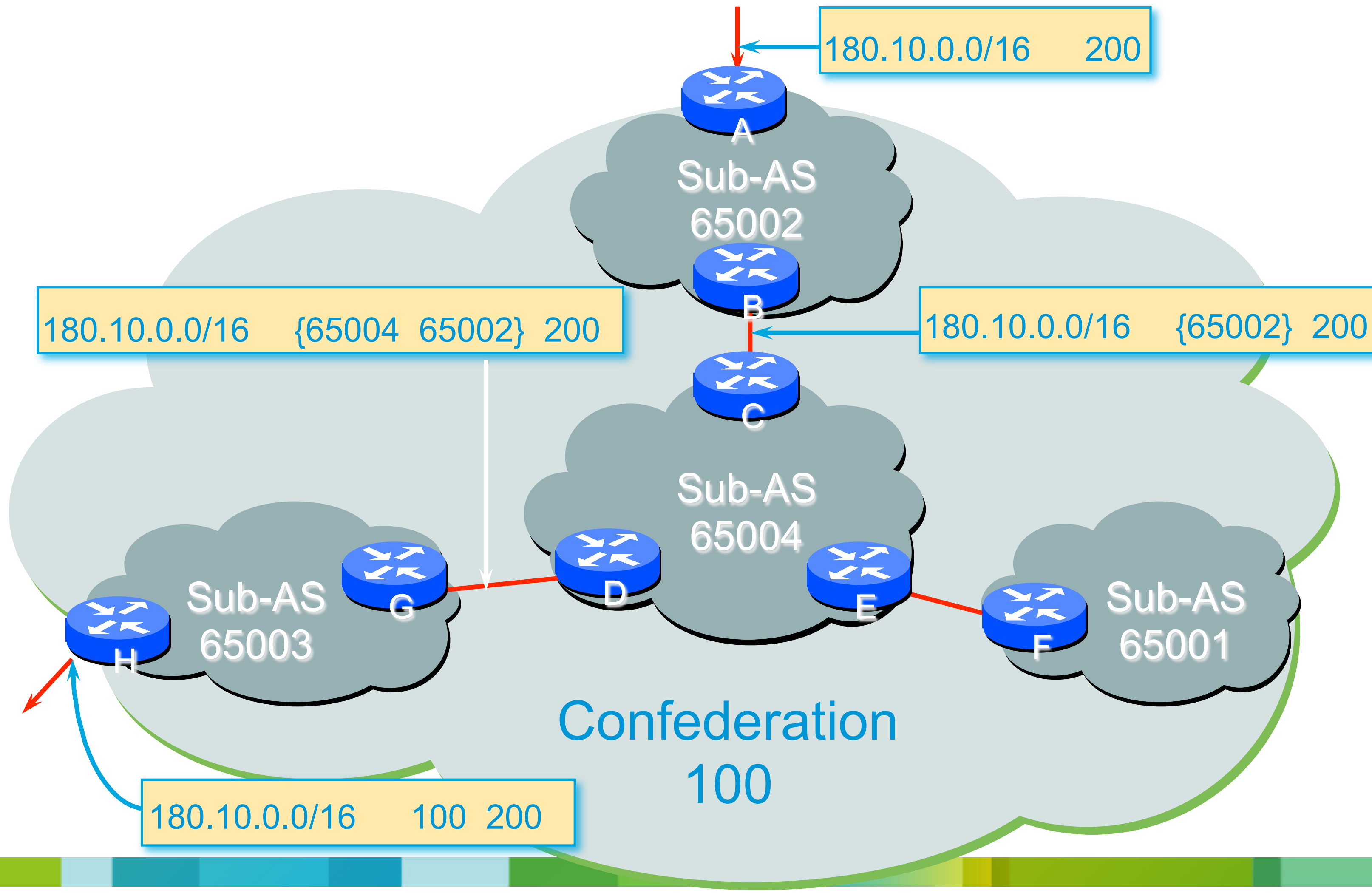
Confederations



- Configuration (Router C):

```
router bgp 65532
 bgp confederation peers
  65530
  65531
 !
 bgp confederation identifier 200
 neighbor 10.10.1.1
  remote-as 65530
 !
 neighbor 10.10.2.2
  remote-as 65531
 !
!
```

Confederations: AS-Sequence



Route Propagation Decisions

- Same as with “normal” BGP:
  - From peer in same sub-AS → only to external peers
  - From external peers → to all neighbors
- “External peers” refers to
  - Peers outside the confederation
  - Peers in a different sub-AS
    - Preserve LOCAL_PREF, MED and NEXT_HOP

RRs or Confederations

	Internet Connectivity	Multi-Level Hierarchy	Policy Control	Scalability	Migration Complexity
Confed-erations	Anywhere In the Network	Yes	Yes	Medium	Medium to High
Route Reflectors	Anywhere In the Network	Yes	Yes	Very High	Very Low

Most new service provider networks now deploy Route Reflectors from Day One

More points about Confederations

- Can ease "absorbing" other ISPs into you ISP – e.g., if one ISP buys another
  - Or can use AS masquerading feature available in some implementations to do a similar thing
- Can use route-reflectors with confederation sub-AS to reduce the sub-AS iBGP mesh

Deploying 4-octet ASN

How to support customers using the extended ASN range

4-Octet ASNs

- Standards documents
  - Description of 4-octet ASNs
    - www.rfc-editor.org/rfc/rfc4893.txt
  - Textual representation
    - www.rfc-editor.org/rfc/rfc5396.txt
  - New extended community
    - www.rfc-editor.org/rfc/rfc5668.txt
- AS 23456 is reserved as interface between 2-octet and 4-octet ASN world

4-octet ASNs – terminology

- 2-octet ASNs

  Refers to the range 0 to 65535

- 4-octet ASNs

  Refers to the range 65536 to 4294967295

  (or the extended range)

- 4-octet ASN pool

  Refers to the range 0 to 4294967295

Getting a 4-octet ASN

- Sample RIR policy
  - www.apnic.net/docs/policy/asn-policy.html
- From 1st January 2007
  - 4-octet ASNs were available on request
- From 1st January 2009
  - 4-octet ASNs were assigned by default
  - 2-octet ASNs were only available on request
- From 1st January 2010
  - No distinction – ASNs assigned from the 4-octet pool

Representation

- Representation of 0-4294967295 ASN range
  - Most operators favour traditional format (asplain)
  - A few prefer dot notation (X.Y):
    - asdot for 65536-4294967295, e.g 2.4
    - asdot+ for 0-4294967295, e.g 0.64513
  - **But regular expressions will have to be completely rewritten for asdot and asdot+ !!!**
- For example:
  - ^[0-9]+$ matches any ASN (16-bit and asplain)
  - This and equivalents extensively used in BGP multihoming configurations for traffic engineering
- Equivalent regexp for asdot is:^([0-9]+)|([0-9]+\.[0-9]+)$
- Equivalent regexp for asdot+ is: ^[0-9]+\.[0-9]+$

Changes

- 4-octet ASNs are backward compatible with 2-octet ASNs
- **There is no flag day**
- You do NOT need to:

  Throw out your old routers

  Replace your 2-octet ASN with a 4-octet ASN
- You do need to be aware that:

  Your customers will come with 4-octet ASNs

  ASN 23456 is not a bogon!

  You will need a router supporting 4-octet ASNs to use a 4-octet ASN locally
- If you have a proper BGP implementation, 4-octet ASNs will be transported silently across your network

How does it work?

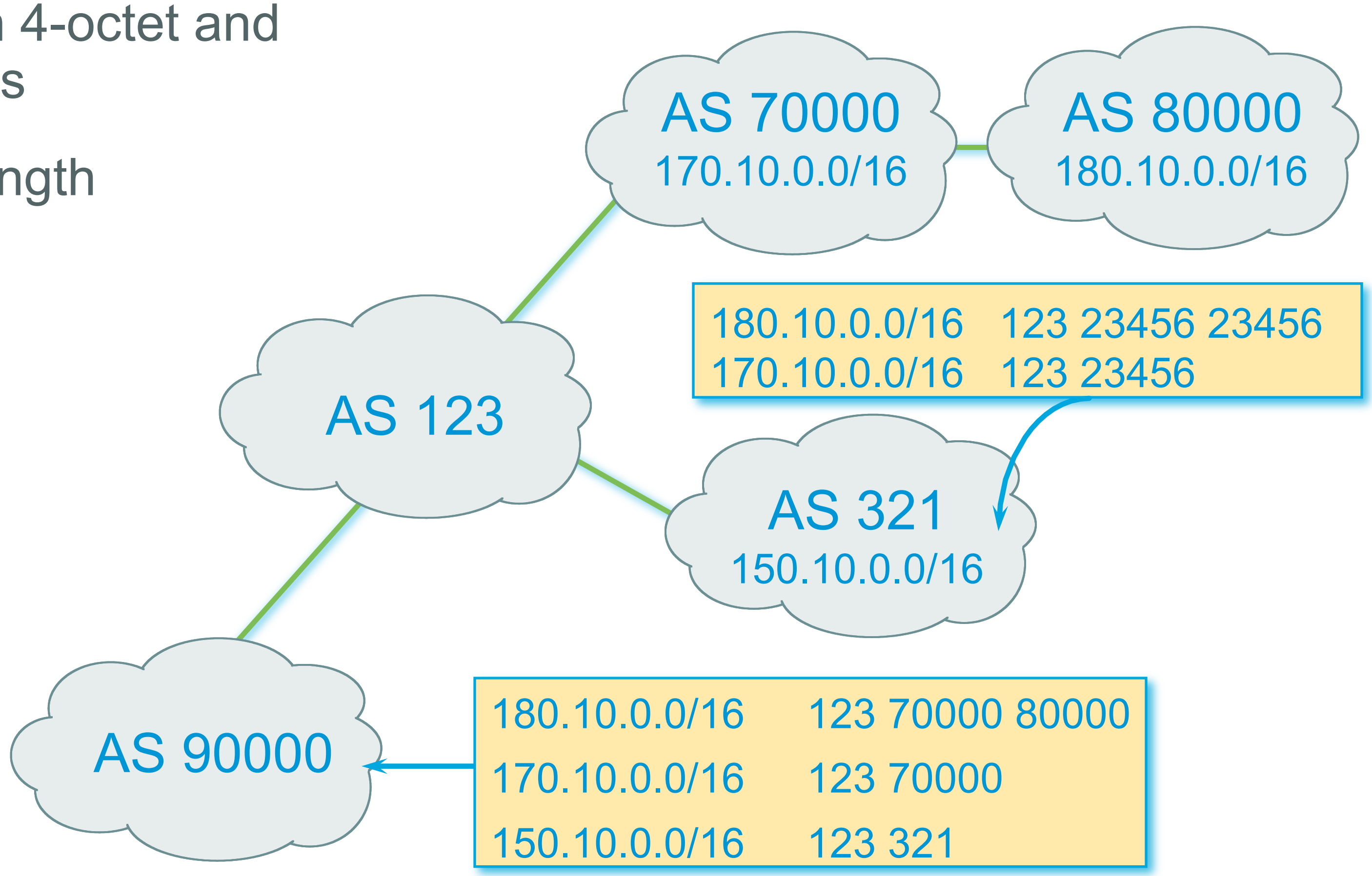
- If local router and remote router supports configuration of 4-octet ASNs
  - BGP peering is configured as normal using the 4-octet ASN
- If local router and remote router does not support configuration of 4-octet ASNs
  - BGP peering can only use a 2-octet ASN
- If local router only supports 2-octet ASN and remote router/network has a 4-octet ASN
  - Compatibility mode is initiated…

Compatibility Mode:

- Local router only supports 2-octet ASN and remote router uses 4-octet ASN
- BGP peering initiated:
  - Remote asks local if 4-octet supported (BGP capability negotiation)
  - When local says "no", remote then presents AS23456
  - Local needs to be configured to peer with remote using AS23456
- BGP peering initiated (cont):
  - BGP session established using AS23456
  - 4-octet ASN included in a new BGP attribute called AS4_PATH
    (as opposed to AS_PATH for 2-octet ASNs)
- Result:
  - 2-octet ASN world sees 2-octet ASNs and 23456 standing in for 4-octet ASNs
  - 4-octet ASN world sees 2 and 4-octet ASNs

Example:

- Internet with 4-octet and 2-octet ASNs
- AS-PATH length maintained



What has changed?

- Two new BGP attributes:
  - AS4_PATH
    - Carries 4-octet ASN path info
  - AS4_AGGREGATOR
    - Carries 4-octet ASN aggregator info
  - Well-behaved BGP implementations will simply pass these along if they don't understand them
- AS23456 (AS_TRANS)

What do they look like?

- IPv4 prefix originated by AS196613

```
As4-crs#sh bgp 145.125.0.0/20
BGP routing table entry for 145.125.0.0/20, version 58734
Paths: (1 available, best #1, table default)
  131072 12654 196613
    204.69.200.25 from 204.69.200.25 (204.69.200.25)
      Origin IGP, localpref 100, valid, internal, best
```

asplain format

- IPv4 prefix originated by AS3.5

```
As4-crs#sh bgp 145.125.0.0/20
BGP routing table entry for 145.125.0.0/20, version 58734
Paths: (1 available, best #1, table default)
  2.0 12654 3.5
    204.69.200.25 from 204.69.200.25 (204.69.200.25)
      Origin IGP, localpref 100, valid, internal, best
```

asdot format

What do they look like?

- IPv4 prefix originated by AS196613

  But 2-octet AS world view:

```
BGP-RTR# sh bgp 145.125.0.0/20
BGP routing table entry for 145.125.0.0/20, version 113382
Paths: (1 available, best #1, table Default-IP-Routing-Table)
23456 12654 23456
    204.69.200.25 from 204.69.200.25 (204.69.200.25)
      Origin IGP, localpref 100, valid, external, best
```

Transition AS

If 4-octet ASN not supported:

- Inability to distinguish between peer ASes using 4-octet ASNs
  - They will all be represented by AS23456
  - Could be problematic for transit provider’s policy
- Inability to distinguish prefix’s origin AS
  - How to tell whether origin is real or fake?
  - The real and fake both represented by AS23456
  - (There should be a better solution here!)
- Incorrect NetFlow summaries:
  - Prefixes from 4-octet ASNs will all be summarised under AS23456
  - Traffic statistics need to be measured per prefix and aggregated
  - Makes it hard to determine peerability of a neighbouring network

Implementations (example)

- Cisco IOS-XR 3.4 onwards
- Cisco IOS-XE 2.3 onwards
- Cisco IOS 12.0(32)S12, 12.4(24)T, 12.2SRE, 12.2(33)SXI1 onwards
- Cisco NX-OS 4.0(1) onwards
- Quagga 0.99.10 (patches for 0.99.6)
- OpenBGPd 4.2 (patches for 3.9 & 4.0)
- Juniper JunOSe 4.1.0 & JunOS 9.1 onwards
- Redback SEOS
- Force10 FTOS7.7.1 onwards

http://as4.cluepon.net/index.php/Software_Support for a complete list

Service Provider use of Communities

Some examples of how ISPs make life easier for themselves

BGP Communities

- Another ISP “scaling technique”
- Prefixes are grouped into different “classes” or communities within the ISP network
- Each community means a different thing, has a different result in the ISP network

BGP Communities

- Communities are generally set at the edge of the ISP network
  - Customer edge: customer prefixes belong to different communities depending on the services they have purchased
  - Internet edge: transit provider prefixes belong to difference communities, depending on the loadsharing or traffic engineering requirements of the local ISP, or what the demands from its BGP customers might be
- Two simple examples follow to explain the concept

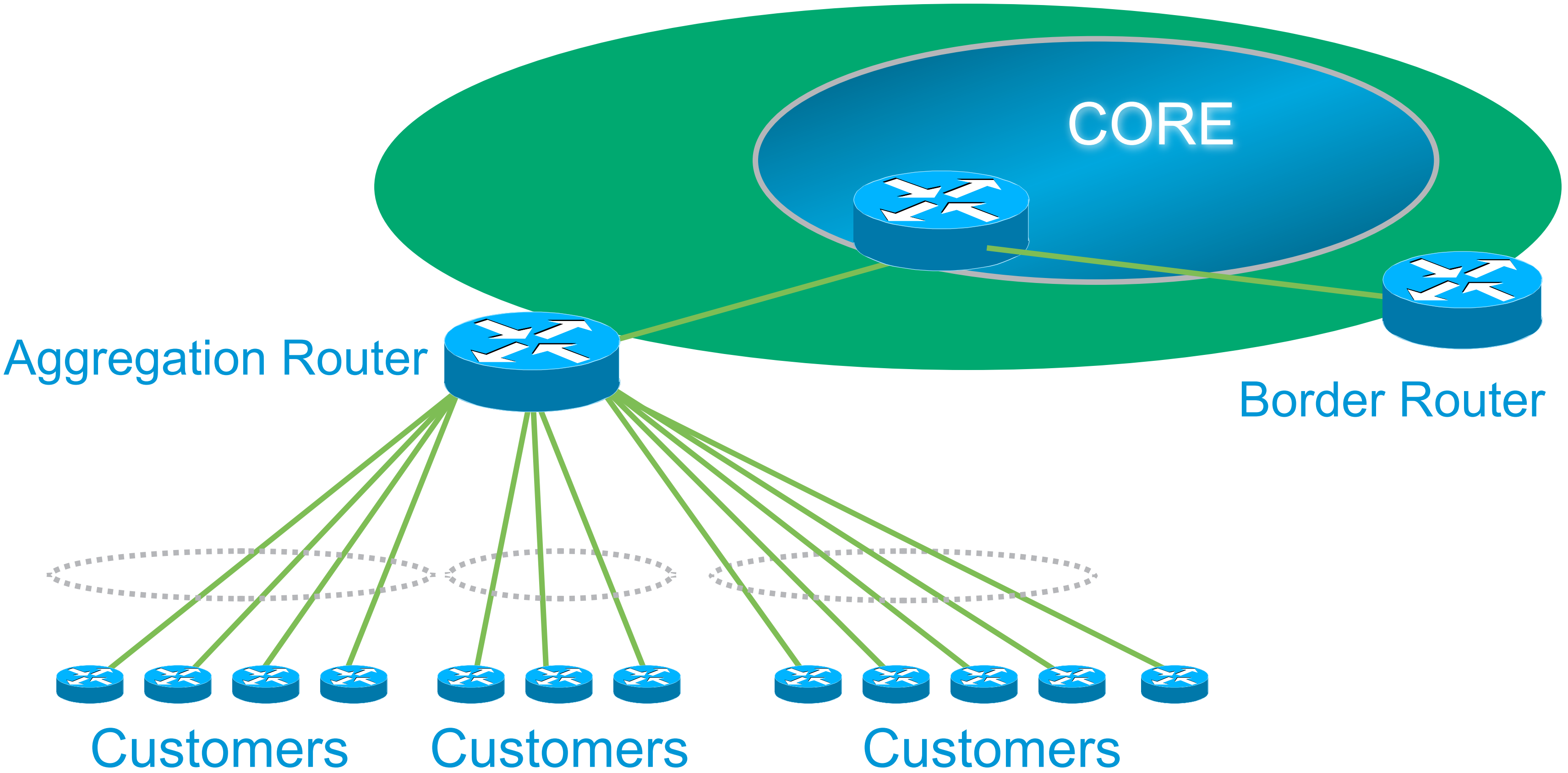
Community Example: Customer Edge

- This demonstrates how communities might be used at the customer edge of an ISP network
- ISP has three connections to the Internet:
  - IXP connection, for local peers
  - Private peering with a competing ISP in the region
  - Transit provider, who provides visibility to the entire Internet
- Customers have the option of purchasing combinations of the above connections

Community Example: Customer Edge

- Community assignments:
  - IXP connection: community 100:2100
  - Private peer: community 100:2200
- Customer who buys local connectivity (via IXP) is put in community 100:2100
- Customer who buys peer connectivity is put in community 100:2200
- Customer who wants both IXP and peer connectivity is put in 100:2100 and 100:2200
- Customer who wants "the Internet" has no community set
  - We are going to announce his prefix everywhere

Community Example: Customer Edge



- Communities set at the aggregation router where the prefix is injected into the ISP's iBGP

Community Example: Customer Edge

- No need to alter filters at the network border when adding a new customer
- New customer simply is added to the appropriate community
  - Border filters already in place take care of announcements
  - ⇒ Ease of operation!

Community Example: Internet Edge

- This demonstrates how communities might be used at the peering edge of an ISP network
- ISP has four types of BGP peers:
  - Customer
  - IXP peer
  - Private peer
  - Transit provider
- The prefixes received from each can be classified using communities
- Customers can opt to receive any or all of the above

Community Example: Internet Edge

- Community assignments:

Customer prefix:	community 100:3000
IXP prefix:	community 100:3100
Private peer prefix:	community 100:3200

- BGP customer who buys local connectivity gets 100:3000
- BGP customer who buys local and IXP connectivity receives community 100:3000 and 100:3100
- BGP customer who buys full peer connectivity receives community 100:3000, 100:3100, and 100:3200
- Customer who wants "the Internet" gets everything

  Gets default route originated by aggregation router

  Or pays money to get all 420k+ prefixes

Community Example: Internet Edge

- No need to create customised filters when adding customers
  - Border router already sets communities
  - Installation engineers pick the appropriate community set when establishing the customer BGP session
  - ⇒ Ease of operation!

Community Example – Summary

- Two examples of customer edge and Internet edge can be combined to form a simple community solution for ISP prefix policy control
- More experienced operators tend to have more sophisticated options available
    - Advice is to start with the easy examples given, and then proceed onwards as experience is gained

ISP BGP Communities

- There are no recommended ISP BGP communities apart from
  - RFC1998
  - The five standard communities
    - www.iana.org/assignments/bgp-well-known-communities
- Efforts have been made to document from time to time
  - totem.info.ucl.ac.be/publications/papers-elec-versions/draft-quoitin-bgp-comm-survey-00.pdf
  - But so far… nothing more… ☹
  - Collection of ISP communities at www.onesc.net/communities
  - NANOG Tutorial: www.nanog.org/meetings/nanog40/presentations/BGPcommunities.pdf
- ISP policy is usually published
  - On the ISP’s website
  - Referenced in the AS Object in the IRR

ISP Examples: Sprint

IP/MPLS Products from Sprint

https://www.sprint.net/index.php?p=policy_bgp

Radio Philip Networking Cisco Miscellaneous Smart Bookmarks TinyURL!

within 3 business days of receipt of the request.

WHAT YOU CAN CONTROL

AS-PATH PREPENDS

Sprint allows customers to use AS-path prepending to adjust route preference on the network. Such prepending will be received and passed on properly without notifiying Sprint of your change in announcments.

Additionally, Sprint will prepend AS1239 to eBGP sessions with certain autonomous systems depending on a received community. Currently, the following ASes are supported: 1668, 209, 2914, 3300, 3356, 3549, 3561, 4635, 701, 7018, 702 and 8220.

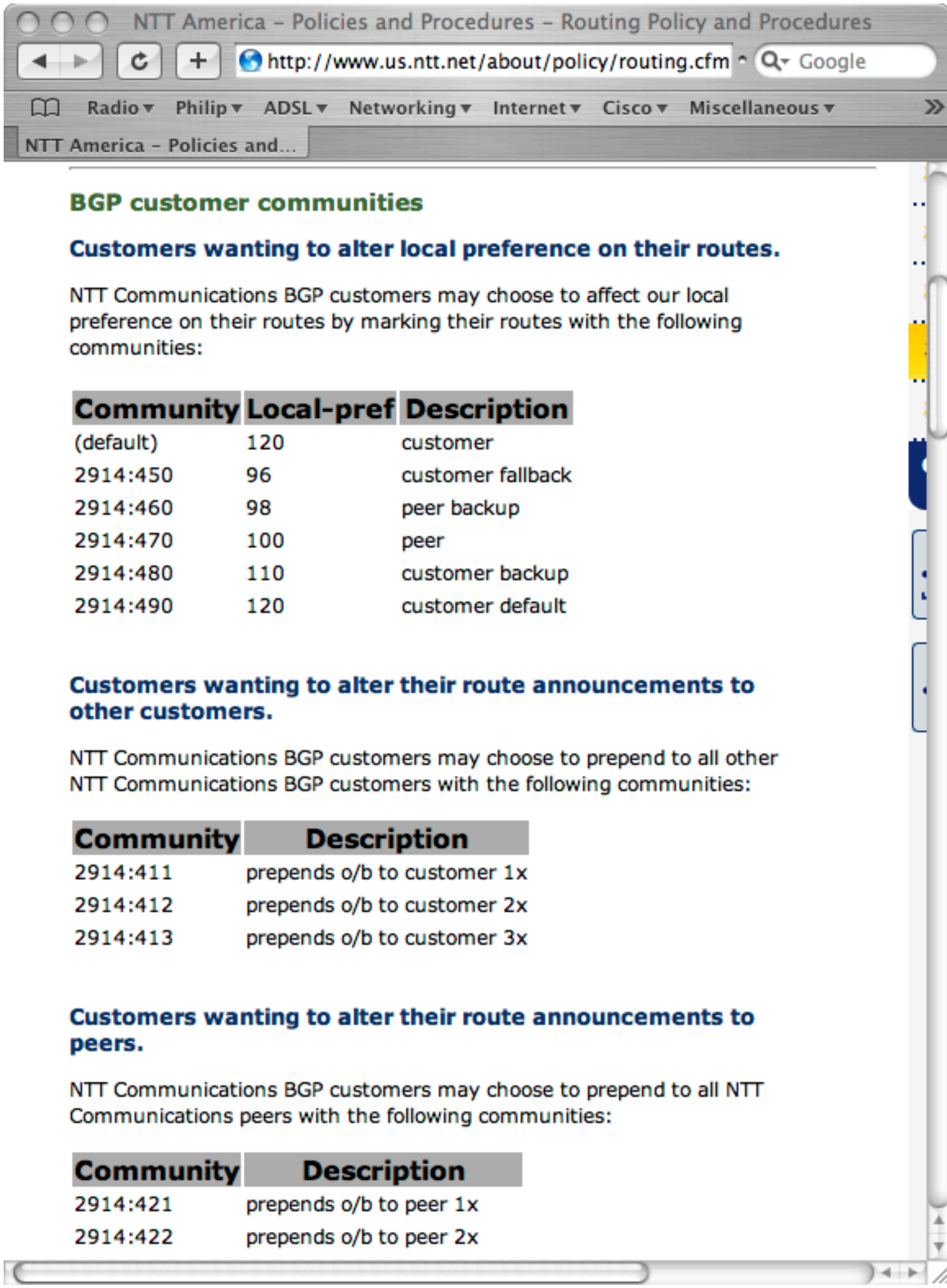
String	Resulting AS Path to ASXXX
65000:XXX	Do not advertise to ASXXX
65001:XXX	1239 (default) ...
65002:XXX	1239 1239 ...
65003:XXX	1239 1239 1239 ...
65004:XXX	1239 1239 1239 1239 ...

String	Resulting AS Path to ASXXX in Asia
65070:XXX	Do not advertise to ASXXX
65071:XXX	1239 (default) ...
65072:XXX	1239 1239 ...
65073:XXX	1239 1239 1239 ...
65074:XXX	1239 1239 1239 1239 ...

String	Resulting AS Path to ASXXX in Europe
65050:XXX	Do not advertise to ASXXX
65051:XXX	1239 (default) ...
65052:XXX	1239 1239 ...
65053:XXX	1239 1239 1239 ...
65054:XXX	1239 1239 1239 1239 ...

More info at https://www.sprint.net/index.php?p=policy_bgp

Some ISP Examples: NTT



NTT America – Policies and Procedures – Routing Policy and Procedures

http://www.us.ntt.net/about/policy/routing.cfm

Radio Philip ADSL Networking Internet Cisco Miscellaneous

NTT America – Policies and...

BGP customer communities

Customers wanting to alter local preference on their routes.

NTT Communications BGP customers may choose to affect our local preference on their routes by marking their routes with the following communities:

Community	Local-pref	Description
(default)	120	customer
2914:450	96	customer fallback
2914:460	98	peer backup
2914:470	100	peer
2914:480	110	customer backup
2914:490	120	customer default

Customers wanting to alter their route announcements to other customers.

NTT Communications BGP customers may choose to prepend to all other NTT Communications BGP customers with the following communities:

Community	Description
2914:411	prepends o/b to customer 1x
2914:412	prepends o/b to customer 2x
2914:413	prepends o/b to customer 3x

Customers wanting to alter their route announcements to peers.

NTT Communications BGP customers may choose to prepend to all NTT Communications peers with the following communities:

Community	Description
2914:421	prepends o/b to peer 1x
2914:422	prepends o/b to peer 2x

More info at www.us.ntt.net/about/policy/routing.cfm

ISP Examples: Verizon Business Europe

```
aut-num: AS702
descr:   Verizon Business EMEA - Commercial IP service provider in Eur
remarks: VzBi uses the following communities with its customers:
         702:80     Set Local Pref 80 within AS702
         702:120    Set Local Pref 120 within AS702
         702:20     Announce only to VzBi AS'es and VzBi customers
         702:30     Keep within Europe, don't announce to other VzBi AS
         702:1      Prepend AS702 once at edges of VzBi to Peers
         702:2      Prepend AS702 twice at edges of VzBi to Peers
         702:3      Prepend AS702 thrice at edges of VzBi to Peers
         Advanced communities for customers
         702:7020   Do not announce to AS702 peers with a scope of
                    National but advertise to Global Peers, European
                    Peers and VzBi customers.
         702:7001   Prepend AS702 once at edges of VzBi to AS702
                    peers with a scope of National.
         702:7002   Prepend AS702 twice at edges of VzBi to AS702
                    peers with a scope of  National.
(more)
```

ISP Examples: Verizon Business Europe

```
(more)
          702:7003 Prepend AS702 thrice at edges of VzBi to AS702
                   peers with a scope  of National.
          702:8020 Do not announce to AS702 peers with a scope of
                   European but advertise to Global Peers, National
                   Peers and VzBi  customers.
          702:8001 Prepend AS702 once at edges of VzBi to AS702
                   peers with a scope of European.
          702:8002 Prepend AS702 twice at edges of VzBi to AS702
                   peers with a scope of  European.
          702:8003 Prepend AS702 thrice at edges of VzBi to AS702
                   peers with a scope  of European.
          --------------------------------------------------------------
          Additional details of the VzBi communities are located at:
          http://www.verizonbusiness.com/uk/customer/bgp/
          --------------------------------------------------------------
mnt-by:   WCOM-EMEA-RICE-MNT
source:   RIPE
```

Some ISP Examples
BT Ignite

```
aut-num:        AS5400
descr:          BT Ignite European Backbone
remarks:
remarks:        Community to                               Community to
remarks:        Not announce      To peer:                 AS prepend 5400
remarks:
remarks:        5400:1000 All peers & Transits             5400:2000
remarks:
remarks:        5400:1500 All Transits                     5400:2500
remarks:        5400:1501 Sprint Transit (AS1239)          5400:2501
remarks:        5400:1502 SAVVIS Transit (AS3561)          5400:2502
remarks:        5400:1503 Level 3 Transit (AS3356)         5400:2503
remarks:        5400:1504 AT&T Transit (AS7018)            5400:2504
remarks:        5400:1506 GlobalCrossing Trans(AS3549)     5400:2506
remarks:
remarks:        5400:1001 Nexica (AS24592)                 5400:2001
remarks:        5400:1002 Fujitsu (AS3324)                 5400:2002
remarks:        5400:1004 C&W EU (1273)                    5400:2004
<snip>
notify:         notify@eu.bt.net
mnt-by:         CIP-MNT
source:         RIPE
```

And many many more!

Some ISP Examples
Level 3

```
aut-num:       AS3356
descr:         Level 3 Communications
<snip>
remarks:       -------------------------------------------------------
remarks:       customer traffic engineering communities - Suppression
remarks:       -------------------------------------------------------
remarks:       64960:XXX - announce to AS XXX if 65000:0
remarks:       65000:0   - announce to customers but not to peers
remarks:       65000:XXX - do not announce at peerings to AS XXX
remarks:       -------------------------------------------------------
remarks:       customer traffic engineering communities - Prepending
remarks:       -------------------------------------------------------
remarks:       65001:0   - prepend once  to all peers
remarks:       65001:XXX - prepend once  at peerings to AS XXX
<snip>
remarks:       3356:70   - set local preference to 70
remarks:       3356:80   - set local preference to 80
remarks:       3356:90   - set local preference to 90
remarks:       3356:9999 - blackhole (discard) traffic
<snip>
mnt-by:        LEVEL3-MNT
source:        RIPE
```

And many many more!

Deploying BGP in an ISP Network

Okay, so we have learned all about BGP.
How do we use it on our network??

Agenda – Deploying BGP

- The role of IGPs and iBGP
- Aggregation
- Receiving Prefixes
- BGP Origin-AS Validation
- Preparing the Network
- Configuration Tips

The role of IGPs and iBGP

BGP versus OSPF/ISIS

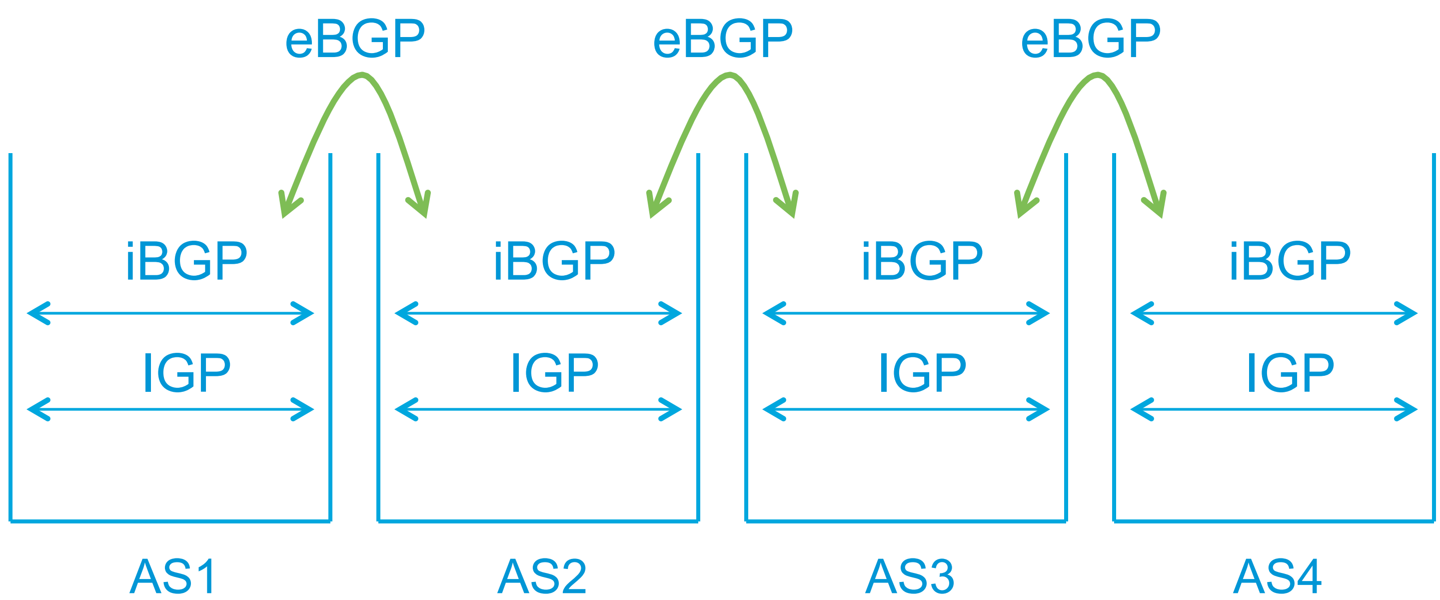
- Internal Routing Protocols (IGPs)
  - examples are ISIS and OSPF
  - used for carrying infrastructure addresses
  - **NOT** used for carrying Internet prefixes or customer prefixes
  - design goal is to minimize number of prefixes in IGP to aid scalability and rapid convergence

BGP versus OSPF/ISIS

- BGP used internally (iBGP) and externally (eBGP)
- iBGP used to carry
  - some/all Internet prefixes across backbone
  - customer prefixes
- eBGP used to
  - exchange prefixes with other ASes
  - implement routing policy

BGP/IGP model used in ISP networks

- Model representation



BGP versus OSPF/ISIS

- DO NOT:

  distribute BGP prefixes into an IGP

  distribute IGP routes into BGP

  use an IGP to carry customer prefixes

- YOUR NETWORK WILL NOT SCALE

Injecting prefixes into iBGP

- Use iBGP to carry customer prefixes
  - Don't ever use IGP
- Point static route to customer interface if customer is single-homed
  - Enter network into BGP process
  - Ensure that implementation options are used so that the prefix always remains in iBGP, regardless of state of interface
  - i.e. avoid iBGP flaps caused by interface flaps
- Consider eBGP with customer only if:
  - Customer is multi-homed to your network or to other provider, and
  - Customer has its own ASN from one of the RIRs

Aggregation

Quality or Quantity?

Aggregation

- Aggregation means announcing the address block received from the RIR to the other ASes connected to your network
- Subprefixes of this aggregate *may* be:
  - Used internally in the ISP network
  - Announced to other ASes to aid with multihoming
- Unfortunately too many people are still thinking about class Cs, resulting in a proliferation of /24s in the Internet routing table

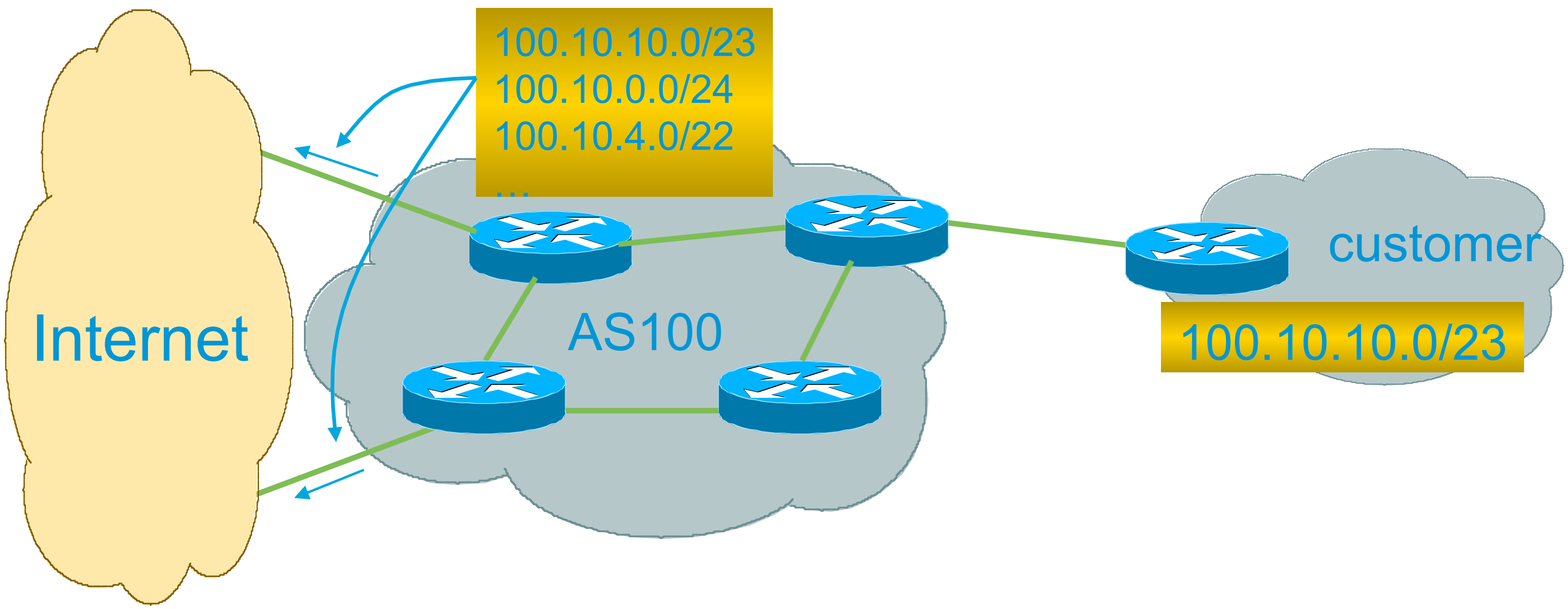
Aggregation

- Address block should be announced to the Internet as an aggregate
- Subprefixes of address block should NOT be announced to Internet unless for traffic engineering purposes
- Aggregate should be generated internally
  - Not on the network borders!

Announcing an Aggregate

- ISPs who don't and won't aggregate are held in poor regard by community
- Registries publish their minimum allocation size
  - Anything from a /20 to a /22 depending on RIR
  - Different sizes for different address blocks
- No real reason to see anything longer than a /22 prefix in the Internet
  - BUT there are currently >185000 /24s!
- But: APNIC changed (Oct 2010) its minimum allocation size on all blocks to /24
  - IPv4 run-out is starting to have an impact

Aggregation – Bad Example

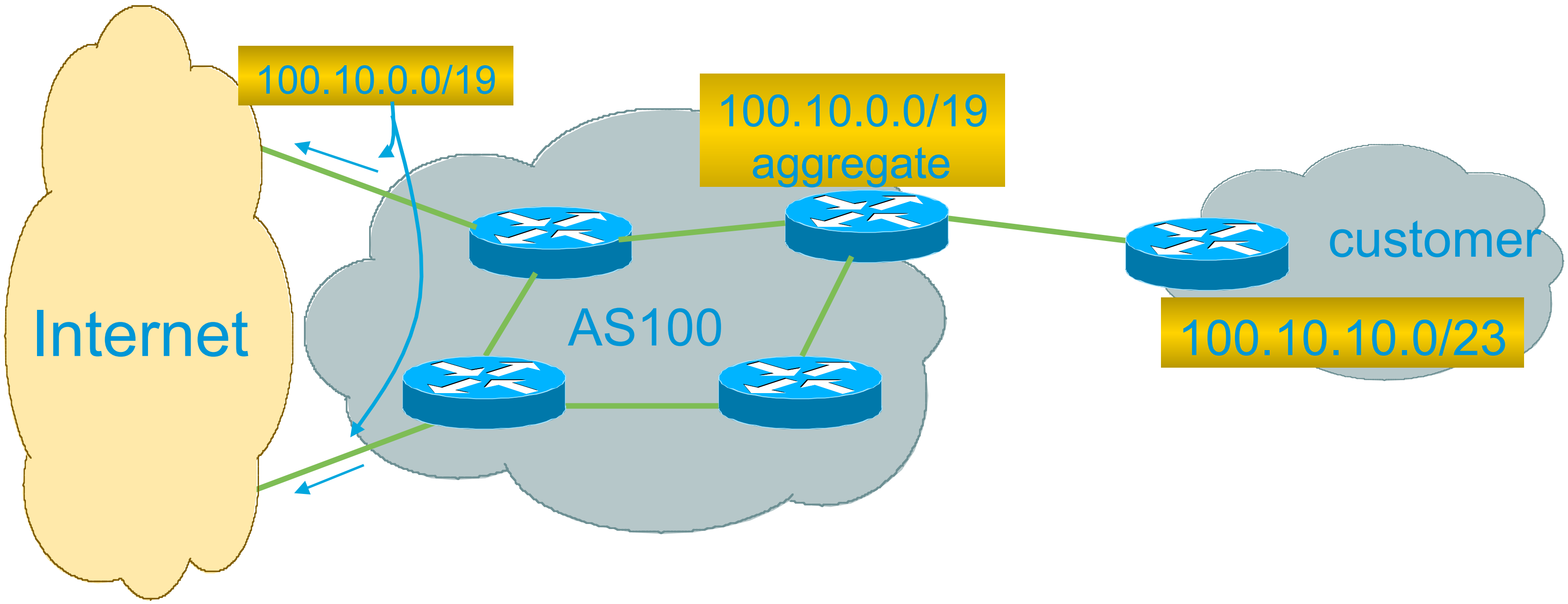


- Customer has /23 network assigned from AS100's /19 address block
- AS100 announces customers' individual networks to the Internet

Aggregation – Bad Example

- Customer link goes down
  - Their /23 network becomes unreachable
  - /23 is withdrawn from AS100's iBGP
- Their ISP doesn't aggregate its /19 network block
  - /23 network withdrawal announced to peers
  - starts rippling through the Internet
  - added load on all Internet backbone routers as network is removed from routing table
- Customer link returns
  - Their /23 network is now visible to their ISP
  - Their /23 network is re-advertised to peers
  - Starts rippling through Internet
  - Load on Internet backbone routers as network is reinserted into routing table
  - Some ISP's suppress the flaps
  - Internet may take 10-20 min or longer to be visible
  - Where is the Quality of Service???

Aggregation – Good Example



- Customer has /23 network assigned from AS100's /19 address block
- AS100 announced /19 aggregate to the Internet

Aggregation – Good Example

- Customer link goes down
  - their /23 network becomes unreachable
  - /23 is withdrawn from AS100's iBGP
- /19 aggregate is still being announced
  - no BGP hold down problems
  - no BGP propagation delays
  - no damping by other ISPs

- Customer link returns
- Their /23 network is visible again
  - The /23 is re-injected into AS100's iBGP
- The whole Internet becomes visible immediately
- Customer has Quality of Service perception

Aggregation – Summary

- Good example is what everyone should do!
  - Adds to Internet stability
  - Reduces size of routing table
  - Reduces routing churn
  - Improves Internet QoS for everyone
- Bad example is what too many still do!
  - Why? Lack of knowledge?
  - Laziness?

Separation of iBGP and eBGP

- Many ISPs do not understand the importance of separating iBGP and eBGP
  - iBGP is where all customer prefixes are carried
  - eBGP is used for announcing aggregate to Internet and for Traffic Engineering
- Do NOT do traffic engineering with customer originated iBGP prefixes
  - Leads to instability similar to that mentioned in the earlier bad example
  - Even though aggregate is announced, a flapping subprefix will lead to instability for the customer concerned
- **Generate traffic engineering prefixes on the Border Router**

The Internet Today (July 2012)

- Current Internet Routing Table Statistics

BGP Routing Table Entries	420845
*CIDR Aggregated	243337
Prefixes after maximum aggregation	181133
*Unique prefixes in Internet	178173
*Prefixes smaller than registry alloc	149545
/24s announced	224148
ASes in use	41910

"The New Swamp"

- Swamp space is name used for areas of poor aggregation
  - The original swamp was 192.0.0.0/8 from the former class C block
    - Name given just after the deployment of CIDR
  - The new swamp is creeping across all parts of the Internet
    - Not just RIR space, but "legacy" space too

"The New Swamp"
RIR Space – February 1999

RIR blocks contribute 88% of the Internet Routing Table

Block	Networks	Block	Networks	Block	Networks	Block	Networks
24/8	165	79/8	0	118/8	0	201/8	0
41/8	0	80/8	0	119/8	0	202/8	2276
58/8	0	81/8	0	120/8	0	203/8	3622
59/8	0	82/8	0	121/8	0	204/8	3792
60/8	0	83/8	0	122/8	0	205/8	2584
61/8	3	84/8	0	123/8	0	206/8	3127
62/8	87	85/8	0	124/8	0	207/8	2723
63/8	20	86/8	0	125/8	0	208/8	2817
64/8	0	87/8	0	126/8	0	209/8	2574
65/8	0	88/8	0	173/8	0	210/8	617
66/8	0	89/8	0	174/8	0	211/8	0
67/8	0	90/8	0	186/8	0	212/8	717
68/8	0	91/8	0	187/8	0	213/8	1
69/8	0	96/8	0	189/8	0	216/8	943
70/8	0	97/8	0	190/8	0	217/8	0
71/8	0	98/8	0	**192/8**	**6275**	218/8	0
72/8	0	99/8	0	193/8	2390	219/8	0
73/8	0	112/8	0	194/8	2932	220/8	0
74/8	0	113/8	0	195/8	1338	221/8	0
75/8	0	114/8	0	196/8	513	222/8	0
76/8	0	115/8	0	198/8	4034		

"The New Swamp"
RIR Space – February 2010

RIR blocks contribute about 87% of the Internet Routing Table

Block	Networks	Block	Networks	Block	Networks	Block	Networks
24/8	3328	79/8	1119	118/8	1349	201/8	4136
41/8	3448	80/8	2335	119/8	1694	202/8	11354
58/8	1675	81/8	1709	120/8	531	203/8	11677
59/8	1575	82/8	1358	121/8	1756	204/8	5744
60/8	888	83/8	1357	122/8	2687	205/8	3037
61/8	2890	84/8	1341	123/8	2400	206/8	3951
62/8	2418	85/8	2492	124/8	2259	207/8	4635
63/8	3114	86/8	780	125/8	2514	208/8	6498
64/8	*6601*	87/8	1466	126/8	106	209/8	5536
65/8	3966	88/8	1068	173/8	1994	210/8	4977
66/8	*7782*	89/8	3168	174/8	1089	211/8	3130
67/8	3771	90/8	377	186/8	1223	212/8	3550
68/8	3221	91/8	4555	187/8	1501	213/8	3442
69/8	5280	96/8	778	189/8	3063	216/8	7645
70/8	2008	97/8	725	190/8	6945	217/8	3136
71/8	1327	98/8	1312	192/8	6952	218/8	1512
72/8	4050	99/8	288	193/8	6820	219/8	1303
73/8	4	112/8	883	194/8	5177	220/8	2108
74/8	*5074*	113/8	890	195/8	5325	221/8	980
75/8	1164	114/8	996	196/8	1857	222/8	1058
76/8	1034	115/8	1616	198/8	4504		
77/8	1964	116/8	1755	199/8	4372		
78/8	1397	117/8	1611	200/8	8884		

“The New Swamp” Summary

- RIR space shows creeping deaggregation

  It seems that an RIR /8 block averages around 5000 prefixes (and upwards) once fully allocated

- Food for thought:

  The 120 RIR /8s combined will cause:

  635000 prefixes with 5000 prefixes per /8 density

  762000 prefixes with 6000 prefixes per /8 density

  Plus 12% due to “non RIR space deaggregation”

  → Routing Table size of 853440 prefixes

"The New Swamp" Summary

- Rest of address space is showing similar deaggregation too ☹
- What are the reasons?
  - Main justification is traffic engineering
- Real reasons are:
  - Lack of knowledge
  - Laziness
  - Deliberate & knowing actions

Efforts to improve aggregation

- The CIDR Report
  - Initiated and operated for many years by Tony Bates and revised by Philip Smith
  - Now combined with Geoff Huston's routing analysis
    - **www.cidr-report.org**
  - Results e-mailed on a weekly basis to most operations lists around the world
  - Lists the top 30 service providers who could do better at aggregating
- RIPE Routing WG aggregation recommendation
  - **RIPE-399 — http://www.ripe.net/ripe/docs/ripe-399.html**

Efforts to Improve Aggregation The CIDR Report

- Also computes the size of the routing table assuming ISPs performed optimal aggregation
- Website allows searches and computations of aggregation to be made on a per AS basis
  - Flexible and powerful tool to aid ISPs
  - Intended to show how greater efficiency in terms of BGP table size can be obtained without loss of routing and policy information
  - Shows what forms of origin AS aggregation could be performed and the potential benefit of such actions to the total table size
  - Very effectively challenges the traffic engineering excuse

CIDR Report

http://www.cidr-report.org/as2.0/

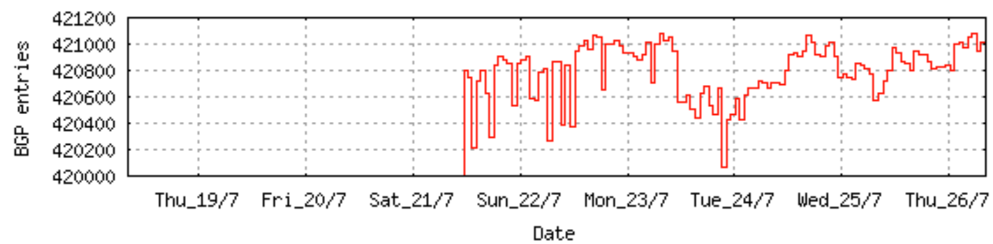
Cisco.com CEC WebEx Apple Wikipedia News Popular

A list of advertisements of address blocks and Autonomous System numbers where there is no matching allocation data.

Status Summary

Table History

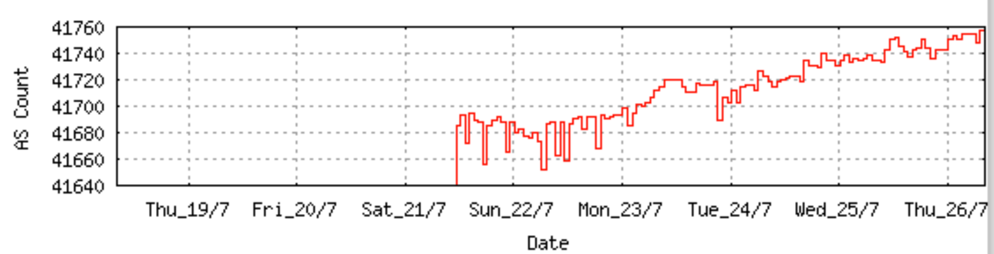
Date	Prefixes	CIDR Aggregated
19-07-12	419152	241935
20-07-12	420802	241935
21-07-12	420802	243450
22-07-12	420851	242316
23-07-12	420929	242400
24-07-12	420469	242764
25-07-12	420742	242807
26-07-12	420845	243337



Plot: BGP Table Size

AS Summary

41753	Number of ASes in routing system
17454	Number of ASes announcing only one prefix
3414	Largest number of prefixes announced by an AS AS7029: WINDSTREAM - Windstream Communications Inc
114212832	Largest address span announced by an AS (/32s) AS4134: CHINANET-BACKBONE



CIDR Report

http://www.cidr-report.org/as2.0/

Cisco.com CEC WebEx Apple Wikipedia News Popular

Aggregation Summary

The algorithm used in this report proposes aggregation only when there is a precise match using AS path so as to preserve traffic transit policies. Aggregation is also proposed across non-advertised address space ('holes').

--- 26Jul12 ---

ASnum	NetsNow	NetsAggr	NetGain	% Gain	Description
Table	421009	243345	177664	42.2%	All ASes
AS6389	3384	189	3195	94.4%	BELLSOUTH-NET-BLK - BellSouth.net Inc.
AS17974	2267	456	1811	79.9%	TELKOMNET-AS2-AP PT Telekomunikasi Indonesia
AS7029	3414	1739	1675	49.1%	WINDSTREAM - Windstream Communications Inc
AS18566	2088	417	1671	80.0%	COVAD - Covad Communications Co.
AS28573	2037	468	1569	77.0%	NET Servicos de Comunicao S.A.
AS4766	2761	1294	1467	53.1%	KIXS-AS-KR Korea Telecom
AS10620	2027	603	1424	70.3%	Telmex Colombia S.A.
AS4323	1578	387	1191	75.5%	TWTC - tw telecom holdings, inc.
AS22773	1694	566	1128	66.6%	ASN-CXA-ALL-CCI-22773-RDC - Cox Communications Inc.
AS1785	1941	817	1124	57.9%	AS-PAETEC-NET - PaeTec Communications, Inc.
AS4755	1617	577	1040	64.3%	TATACOMM-AS TATA Communications formerly VSNL is Leading ISP
AS7303	1457	450	1007	69.1%	Telecom Argentina S.A.
AS7552	1128	231	897	79.5%	VIETEL-AS-AP Vietel Corporation
AS6458	881	45	836	94.9%	Telgua
AS8151	1477	670	807	54.6%	Uninet S.A. de C.V.
AS18101	942	157	785	83.3%	RELIANCE-COMMUNICATIONS-IN Reliance Communications Ltd.DAKC MUMBAI
AS17908	828	60	768	92.8%	TCISL Tata Communications
AS4808	1118	351	767	68.6%	CHINA169-BJ CNCGROUP IP network China169 Beijing Province Network
AS9394	908	166	742	81.7%	CRNET CHINA RAILWAY Internet(CRNET)

CIDR Report

http://www.cidr-report.org/as2.0/

Cisco.com CEC WebEx Apple Wikipedia News Popular

Top 20 Route Count per Originating AS

Prefixes	ASnum	AS Description
3414	AS7029	WINDSTREAM - Windstream Communications Inc
3384	AS6389	BELLSOUTH-NET-BLK - BellSouth.net Inc.
2761	AS4766	KIXS-AS-KR Korea Telecom
2267	AS17974	TELKOMNET-AS2-AP PT Telekomunikasi Indonesia
2088	AS18566	COVAD - Covad Communications Co.
2037	AS28573	NET Servicos de Comunicao S.A.
2027	AS10620	Telmex Colombia S.A.
1941	AS1785	AS-PAETEC-NET - PaeTec Communications, Inc.
1705	AS7545	TPG-INTERNET-AP TPG Internet Pty Ltd
1694	AS22773	ASN-CXA-ALL-CCI-22773-RDC - Cox Communications Inc.
1647	AS20115	CHARTER-NET-HKY-NC - Charter Communications
1617	AS4755	TATACOMM-AS TATA Communications formerly VSNL is Leading ISP
1578	AS4323	TWTC - tw telecom holdings, inc.
1525	AS6503	Axtel, S.A.B. de C.V.
1495	AS8402	CORBINA-AS OJSC "Vimpelcom"
1477	AS8151	Uninet S.A. de C.V.
1457	AS7303	Telecom Argentina S.A.
1401	AS30036	MEDIACOM-ENTERPRISE-BUSINESS - Mediacom Communications Corp
1305	AS9829	BSNL-NIB National Internet Backbone
1257	AS7018	ATT-INTERNET4 - AT&T Services, Inc.

Last Week's Changes

This a daily snapshot of changes in routes being withdrawn and added. The deltas are calculated over a rolling 7 day period. Please bear in mind this is purely a "snapshot" and a large flucuation could be caused by a connectivity problem for example.

CIDR Report

http://www.cidr-report.org/as2.0/

Cisco.com CEC WebEx Apple Wikipedia News Popular

More Specifics

A list of route advertisements that appear to be more specfic than the original Class-based prefix mask, or more specific than the registry allocation size.

Top 20 ASes advertising more specific prefixes

More Specifics	Total Prefixes	ASnum	AS Description
3321	3384	AS6389	BELLSOUTH-NET-BLK - BellSouth.net Inc.
3272	3414	AS7029	WINDSTREAM - Windstream Communications Inc
2684	2761	AS4766	KIXS-AS-KR Korea Telecom
2294	2334	AS4	ISI-AS - University of Southern California
2249	2267	AS17974	TELKOMNET-AS2-AP PT Telekomunikasi Indonesia
2067	2088	AS18566	COVAD - Covad Communications Co.
2037	2037	AS28573	NET Servicos de Comunicao S.A.
2025	2027	AS10620	Telmex Colombia S.A.
1851	1941	AS1785	AS-PAETEC-NET - PaeTec Communications, Inc.
1714	3420	AS3	MIT-GATEWAYS - Massachusetts Institute of Technology
1650	1705	AS7545	TPG-INTERNET-AP TPG Internet Pty Ltd
1638	1694	AS22773	ASN-CXA-ALL-CCI-22773-RDC - Cox Communications Inc.
1607	1617	AS4755	TATACOMM-AS TATA Communications formerly VSNL is Leading ISP
1596	1647	AS20115	CHARTER-NET-HKY-NC - Charter Communications
1481	1495	AS8402	CORBINA-AS OJSC "Vimpelcom"
1455	1525	AS6503	Axtel, S.A.B. de C.V.
1450	1457	AS7303	Telecom Argentina S.A.
1399	1401	AS30036	MEDIACOM-ENTERPRISE-BUSINESS - Mediacom Communications Corp
1396	1477	AS8151	Uninet S.A. de C.V.
1381	1578	AS4323	TWTC - tw telecom holdings, inc.

Report: ASes ordered by number of more specific prefixes
Report: More Specific prefix list (by AS)
Report: More Specific prefix list (ordered by prefix)

AS Report

http://www.cidr-report.org/cgi-bin/as-report?as=AS4755&view=2.0

Cisco.com CEC WebEx Apple Wikipedia News Popular

Announced Prefixes

Rank	AS	Type	Originate Addr Space (pfx)	Transit Addr space (pfx)	Description
183	AS4755	ORG+TRN	Originate: 2760448 /10.60	Transit: 14170624 /8.24	TATACOMM-AS TATA Communications formerly VS

Aggregation Suggestions

This report does not take into account conditions local to each origin AS in terms of policy or traffic engineering requirements, so this is an approximate guideline as to aggregation possibilities.

Rank	AS	AS Name	Current	Wthdw	Aggte	Annce	Redctn	%
12	AS4755	TATACOMM-AS TATA Communications formerly VSNL	1617	1139	99	577	1040	64.32%

Prefix	AS Path	Aggregation Suggestion
14.140.0.0/14	4777 2516 6453 4755	
14.140.0.0/21	4608 1221 4637 6453 4755	+ Announce - aggregate of 14.140.0.0/22 (4608 1221 4637 6453 4755) and 14.140.4.0
14.140.0.0/22	4608 1221 4637 6453 4755	- Withdrawn - aggregated with 14.140.4.0/22 (4608 1221 4637 6453 4755)
14.140.4.0/23	4608 1221 4637 6453 4755	- Withdrawn - aggregated with 14.140.6.0/23 (4608 1221 4637 6453 4755)
14.140.6.0/23	4608 1221 4637 6453 4755	- Withdrawn - aggregated with 14.140.4.0/23 (4608 1221 4637 6453 4755)
14.140.16.0/21	4608 1221 4637 6453 4755	+ Announce - aggregate of 14.140.16.0/22 (4608 1221 4637 6453 4755) and 14.140.20
14.140.16.0/22	4608 1221 4637 6453 4755	- Withdrawn - aggregated with 14.140.20.0/22 (4608 1221 4637 6453 4755)
14.140.18.0/24	4608 1221 4637 6453 4755	- Withdrawn - matching aggregate 14.140.16.0/22 4608 1221 4637 6453 4755
14.140.20.0/22	4608 1221 4637 6453 4755	- Withdrawn - aggregated with 14.140.16.0/22 (4608 1221 4637 6453 4755)
14.140.24.0/22	4608 1221 4637 6453 4755	
14.140.32.0/23	4608 1221 4637 6453 4755	
14.140.40.0/21	4608 1221 4637 6453 4755	
14.140.48.0/20	4608 1221 4637 6453 4755	+ Announce - aggregate of 14.140.48.0/21 (4608 1221 4637 6453 4755) and 14.140.56
14.140.48.0/21	4608 1221 4637 6453 4755	- Withdrawn - aggregated with 14.140.56.0/21 (4608 1221 4637 6453 4755)
14.140.56.0/21	4608 1221 4637 6453 4755	- Withdrawn - aggregated with 14.140.48.0/21 (4608 1221 4637 6453 4755)
14.140.64.0/21	4608 1221 4637 6453 4755	
14.140.72.0/22	4608 1221 4637 6453 4755	
14.140.80.0/20	4608 1221 4637 6453 4755	+ Announce - aggregate of 14.140.80.0/21 (4608 1221 4637 6453 4755) and 14.140.88
14.140.80.0/23	4608 1221 4637 6453 4755	- Withdrawn - aggregated with 14.140.82.0/23 (4608 1221 4637 6453 4755)
14.140.82.0/23	4608 1221 4637 6453 4755	- Withdrawn - aggregated with 14.140.80.0/23 (4608 1221 4637 6453 4755)
14.140.84.0/22	4608 1221 4637 6453 4755	- Withdrawn - aggregated with 14.140.80.0/22 (4608 1221 4637 6453 4755)
14.140.88.0/21	4608 1221 4637 6453 4755	- Withdrawn - aggregated with 14.140.80.0/21 (4608 1221 4637 6453 4755)
14.140.96.0/22	4608 1221 4637 6453 4755	
14.140.104.0/21	4608 1221 4637 6453 4755	
14.140.112.0/20	4608 1221 4637 6453 4755	+ Announce - aggregate of 14.140.112.0/21 (4608 1221 4637 6453 4755) and 14.140.1
14.140.112.0/22	4608 1221 4637 6453 4755	- Withdrawn - aggregated with 14.140.116.0/22 (4608 1221 4637 6453 4755)
14.140.116.0/23	4608 1221 4637 6453 4755	- Withdrawn - aggregated with 14.140.118.0/23 (4608 1221 4637 6453 4755)
14.140.118.0/23	4608 1221 4637 6453 4755	- Withdrawn - aggregated with 14.140.116.0/23 (4608 1221 4637 6453 4755)
14.140.120.0/21	4608 1221 4637 6453 4755	- Withdrawn - aggregated with 14.140.112.0/21 (4608 1221 4637 6453 4755)

Importance of Aggregation

- Size of routing table
  - Router Memory is not so much of a problem as it was in the 1990s
  - Routers can be specified to carry 1 million+ prefixes
- Convergence of the Routing System
  - This is a problem
  - Bigger table takes longer for CPU to process
  - BGP updates take longer to deal with
  - BGP Instability Report tracks routing system update activity
  - **http://bgpupdates.potaroo.net/instability/bgpupd.html**

The BGP Instability Report

http://bgpupdates.potaroo.net/instability/bgpupd.html

The BGP Instability Report

The BGP Instability Report is updated daily. This report was generated on 26 July 2012 06:19 (UTC+1000)

50 Most active ASes for the past 7 days

RANK	ASN	UPDs	%	Prefixes	UPDs/Prefix	AS NAME
1	8402	31474	1.38%	1766	17.82	CORBINA-AS OJSC "Vimpelcom"
2	1637	30729	1.35%	108	284.53	DNIC-AS-01637 - Headquarters, USAISC
3	17813	29341	1.28%	136	215.74	MTNL-AP Mahanagar Telephone Nigam Ltd.
4	47931	25100	1.10%	123	204.07	ALENETWORK A.L.E. COM NETWORK S.R.L
5	9829	21569	0.94%	1305	16.53	BSNL-NIB National Internet Backbone
6	24560	19759	0.86%	1037	19.05	AIRTELBROADBAND-AS-AP Bharti Airtel Ltd., Telemedia Services
7	7029	15412	0.67%	3508	4.39	WINDSTREAM - Windstream Communications Inc
8	7552	13226	0.58%	1131	11.69	VIETEL-AS-AP Vietel Corporation
9	13118	11776	0.52%	48	245.33	ASN-YARTELECOM OJSC Rostelecom
10	6458	11752	0.51%	882	13.32	Telgua
11	27738	11509	0.50%	557	20.66	Ecuadortelecom S.A.
12	48277	11271	0.49%	56	201.27	SOREX SOREX MEDIA S.R.L.
13	49074	10768	0.47%	49	219.76	TECHNOLOGICAL SC TECHNOLOGICAL SRL
14	6389	10345	0.45%	3387	3.05	BELLSOUTH-NET-BLK - BellSouth.net Inc.
15	28573	9562	0.42%	2054	4.66	NET Servicos de Comunicao S.A.
16	10620	9514	0.42%	2027	4.69	Telmex Colombia S.A.
17	5800	8667	0.38%	258	33.59	DNIC-ASBLK-05800-06055 - DoD Network Information Center
18	4766	8347	0.37%	2764	3.02	KIXS-AS-KR Korea Telecom
19	8151	8307	0.36%	1492	5.57	Uninet S.A. de C.V.
20	43875	8261	0.36%	40	206.53	DATAINFO-ASN SC Data Media Info SRL
21	28885	8126	0.36%	137	59.31	OMANTEL-NAP-AS OmanTel NAP

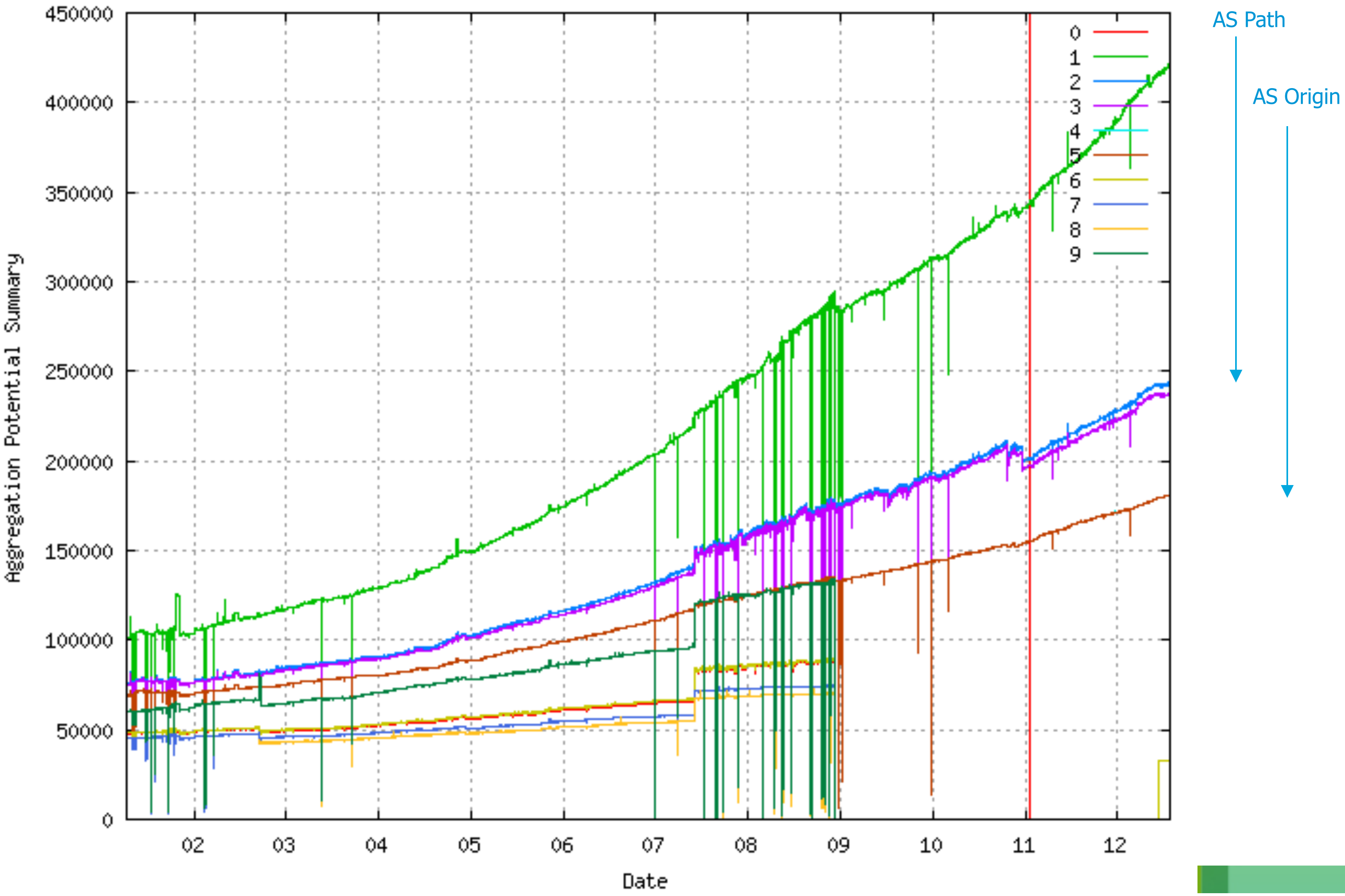
The BGP Instability Report

http://bgpupdates.potaroo.net/instability/bgpupd.html

50 Most active ASes for the past 7 days

RANK	ASN	UPDs/Prefix	%	Prefixes	UPDs	AS NAME
1	16535	1121.3	0.15%	3	3364	ECHOS-3 - Echostar Holding Purchasing Corporation
2	44410	884.7	0.12%	3	2654	ENTEKHAB-AS ENTEKHAB INDUSTRIAL GROUP
3	43348	876.0	0.08%	2	1752	TATARINOVA-AS PE Tatarinova Alla Ivanovna
4	49072	837.0	0.04%	1	837	APSUARA-AS TCA Apsuara Ltd.
5	54037	770.0	0.03%	1	770	CAREER-GROUP-INC - CAREER GROUP INC
6	14452	701.3	0.28%	9	6312	IOS-ASN - INTERNET OF THE SANDHILLS
7	26184	645.0	0.03%	1	645	ASA-HQAS - American Society of Anesthesiologists
8	58655	580.0	0.05%	2	1160	SKYTEL6-BD SkyTel Communications Limited
9	51250	552.0	0.02%	1	552	ITE-PROTON-AS "Information technologies enterprise "Proton" LTD
10	3	440.0	0.02%	1	440	MIT-GATEWAYS - Massachusetts Institute of Technology
11	42806	411.0	0.02%	1	411	TELECOM-AS Telecom Georgia
12	38857	387.5	0.03%	2	775	ESOFT-TRANSIT-AS-AP e.Soft Technologies Ltd.
13	23007	296.0	0.04%	3	888	Universidad de Los Andes
14	4	296.0	0.01%	1	296	ISI-AS - University of Southern California
15	27890	288.0	0.03%	2	576	Universidad de Oriente
16	1637	284.5	1.35%	108	30729	DNIC-AS-01637 - Headquarters, USAISC
17	23237	279.2	0.05%	4	1117	MCMASTER - McMaster University
18	29398	277.0	0.01%	1	277	PETROBALTIC "Petrobaltic" S.A.
19	34744	247.3	0.24%	22	5440	GVM S.C. GVM SISTEM 2003 S.R.L.
20	50704	246.1	0.08%	7	1723	BENEFIC-INTERNET Benefic Consult SRL
21	13118	245.3	0.52%	48	11776	ASN-YARTELECOM OJSC Rostelecom
22	3388	243.3	0.12%	11	2676	UNM-AS - University of New Mexico
23	15478	240.4	0.12%	11	2644	W-MEDIA White Market Media SRL
24	57201	232.0	0.01%	1	232	EDF-AS Estonian Defence Forces
25	47147	226.4	0.08%	8	1811	VISNET-AS VisNetwork Media SRL
26	19406	223.4	0.11%	11	2457	TWRS-MA - Towerstream I, Inc.

Aggregation Potential (source: bgp.potaroo.net/as2.0/)



Aggregation Summary

- Aggregation on the Internet could be MUCH better
  - 35% saving on Internet routing table size is quite feasible
  - Tools are available
    - Commands on the routers are not hard
    - CIDR-Report webpage

Receiving Prefixes

Receiving Prefixes

- There are three scenarios for receiving prefixes from other ASNs
  - Customer talking BGP
  - Peer talking BGP
  - Upstream/Transit talking BGP
- Each has different filtering requirements and need to be considered separately

Receiving Prefixes: From Customers

- ISPs should only accept prefixes which have been assigned or allocated to their downstream customer
- If ISP has assigned address space to its customer, then the customer IS entitled to announce it back to his ISP
- If the ISP has NOT assigned address space to its customer, then:

  Check the five RIR databases to see if this address space really has been assigned to the customer

  The tool: whois

Receiving Prefixes: From Customers

- Example use of whois to check if customer is entitled to announce address space:

```
$ whois -h whois.apnic.net 202.12.29.0
inetnum:        202.12.28.0 - 202.12.29.255
netname:        APNIC-AP
descr:          Asia Pacific Network Information Centre
descr:          Regional Internet Registry for the Asia-Pacific
descr:          6 Cordelia Street
descr:          South Brisbane, QLD 4101
descr:          Australia
country:        AU
admin-c:        AIC1-AP
tech-c:         NO4-AP
mnt-by:         APNIC-HM
mnt-irt:        IRT-APNIC-AP
changed:        hm-changed@apnic.net
status:         ASSIGNED PORTABLE
changed:        hm-changed@apnic.net 20110309
source:         APNIC
```

Portable – means its an assignment to the customer, the customer can announce it to you

Receiving Prefixes: From Customers

- Example use of whois to check if customer is entitled to announce address space:

```
$ whois -h whois.ripe.net 193.128.0.0
inetnum:        193.128.0.0 - 193.133.255.255
netname:        UK-PIPEX-193-128-133
descr:          Verizon UK Limited
country:        GB
org:            ORG-UA24-RIPE
admin-c:        WERT1-RIPE
tech-c:         UPHM1-RIPE
status:         ALLOCATED UNSPECIFIED
remarks:        Please send abuse notification to abuse@uk.uu.net
mnt-by:         RIPE-NCC-HM-MNT
mnt-lower:      AS1849-MNT
mnt-routes:     AS1849-MNT
mnt-routes:     WCOM-EMEA-RICE-MNT
mnt-irt:        IRT-MCI-GB
source:         RIPE # Filtered
```

ALLOCATED – means that this is Provider Aggregatable address space and can only be announced by the ISP holding the allocation (in this case Verizon UK)

Receiving Prefixes: From Peers

- A peer is an ISP with whom you agree to exchange prefixes you originate into the Internet routing table
  - Prefixes you accept from a peer are only those they have indicated they will announce
  - Prefixes you announce to your peer are only those you have indicated you will announce

Receiving Prefixes: From Peers

- Agreeing what each will announce to the other:
  - Exchange of e-mail documentation as part of the peering agreement, and then ongoing updates

    OR

  - Use of the Internet Routing Registry and configuration tools such as the IRRToolSet

    www.isc.org/sw/IRRToolSet/
- Alternatively, you can use origin-AS validation
  - Recommended if (or when) your routers support it
  - Enables you to automatically validate that the origin AS in the AS path is valid using RIRs registries
  - Discussed in the next section

Receiving Prefixes: From Upstream/Transit Provider

- Upstream/Transit Provider is an ISP who you pay to give you transit to the WHOLE Internet
- Receiving prefixes from them is not desirable unless really necessary
  - Traffic Engineering – see BGP Multihoming Tutorial
- Ask upstream/transit provider to either:
  - originate a default-route

    OR
  - announce one prefix you can use as default

Receiving Prefixes: From Upstream/Transit Provider

- If necessary to receive prefixes from any provider, care is required.
  - Don't accept default (unless you need it)
  - Don't accept your own prefixes
- For IPv4:
  - Don't accept private (RFC1918) and certain special use prefixes:
    - **http://www.rfc-editor.org/rfc/rfc5735.txt**
  - Don't accept prefixes longer than /24 (?)
- For IPv6:
  - Don't accept certain special use prefixes:
    - **http://www.rfc-editor.org/rfc/rfc5156.txt**
  - Don't accept prefixes longer than /48 (?)

Receiving Prefixes: From Upstream/Transit Provider

- Check Team Cymru's list of "bogons"
  - Cymru is pronounced kum-ree
  - www.team-cymru.org/Services/Bogons/http.html
- For IPv6 also consult:
  - www.space.net/~gert/RIPE/ipv6-filters.html
- Bogon Route Server:
  - www.team-cymru.org/Services/Bogons/routeserver.html
  - Supplies a BGP feed (IPv4 and/or IPv6) of address blocks which should not appear in the BGP table

Receiving Prefixes

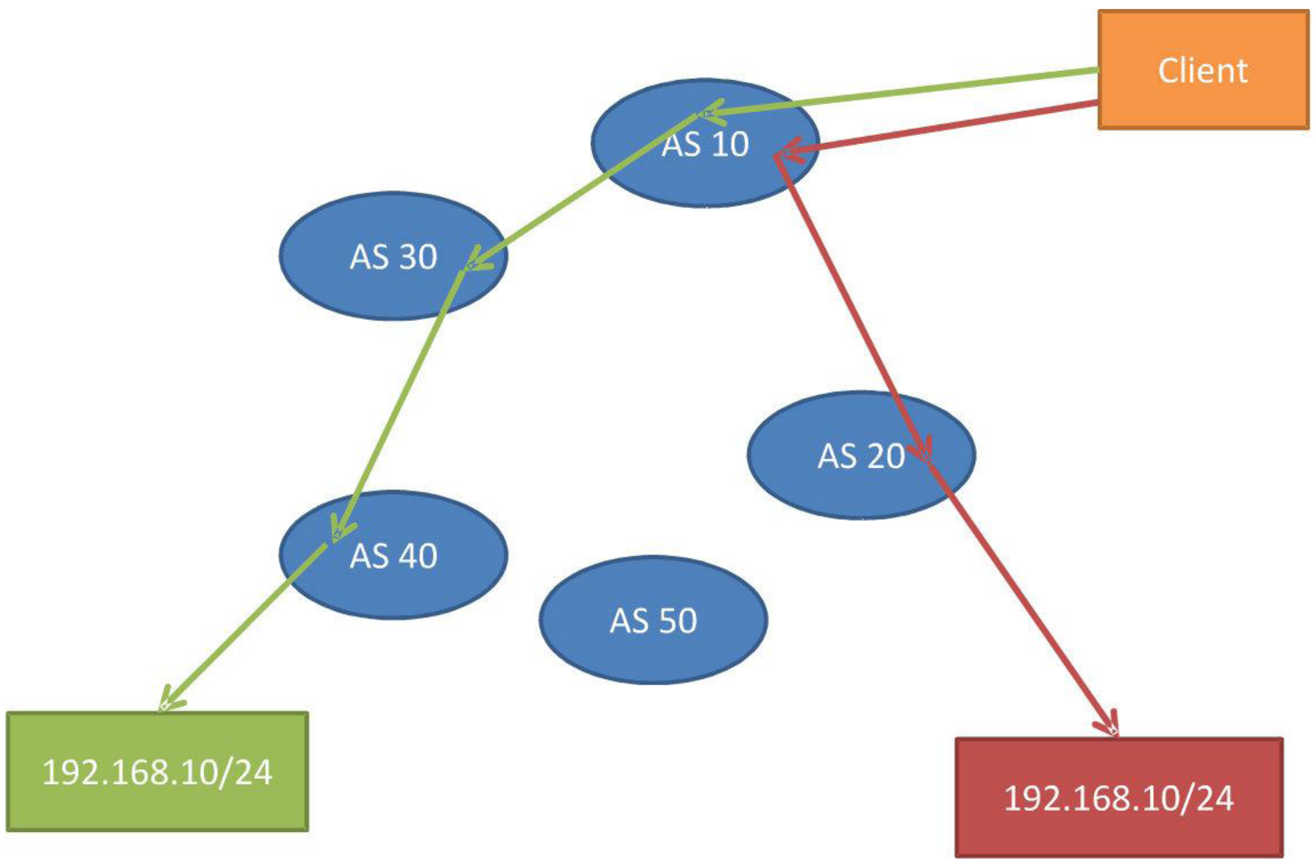
- Paying attention to prefixes received from customers, peers and transit providers assists with:
  - The integrity of the local network
  - The integrity of the Internet
- Responsibility of all ISPs to be good Internet citizens

BGP Origin-AS Validation

Security issue for BGP route distribution

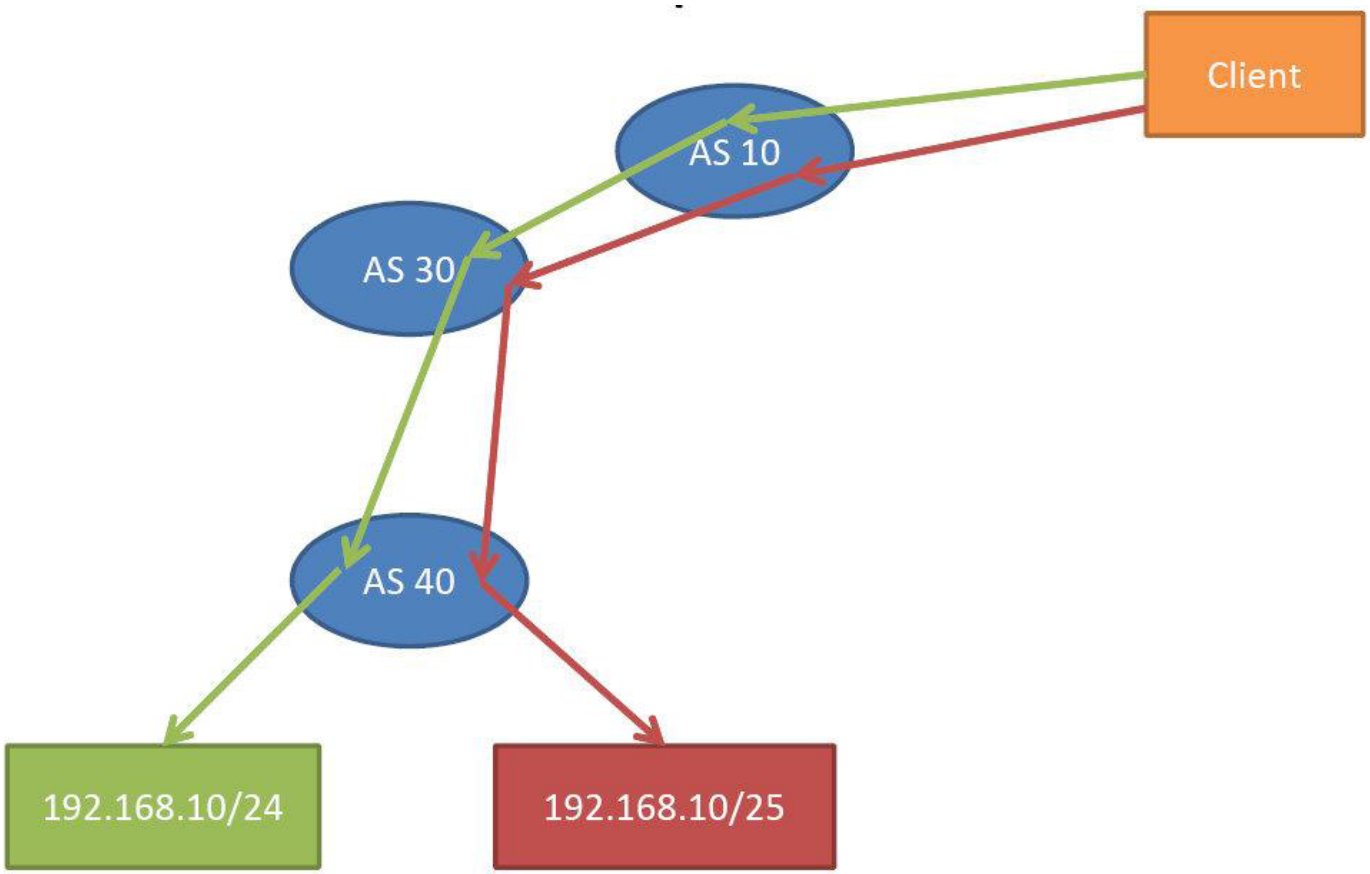
- Any AS can inject any prefixes in BGP, leading to prefix hijacking
  - Inadvertently or maliciously
- The manifestation of prefix hijacking are
  - an AS announcing someone else's prefix
  - as AS announcing a more specific of someone else's prefix
- The actual incidents are:
  - http://www.networkworld.com/news/2009/011509-bgp-attacks.html
- Need a mechanism to differentiate between invalid and legitimate routes for a BGP destination

Same Prefix: Shorter AS_PATH length wins



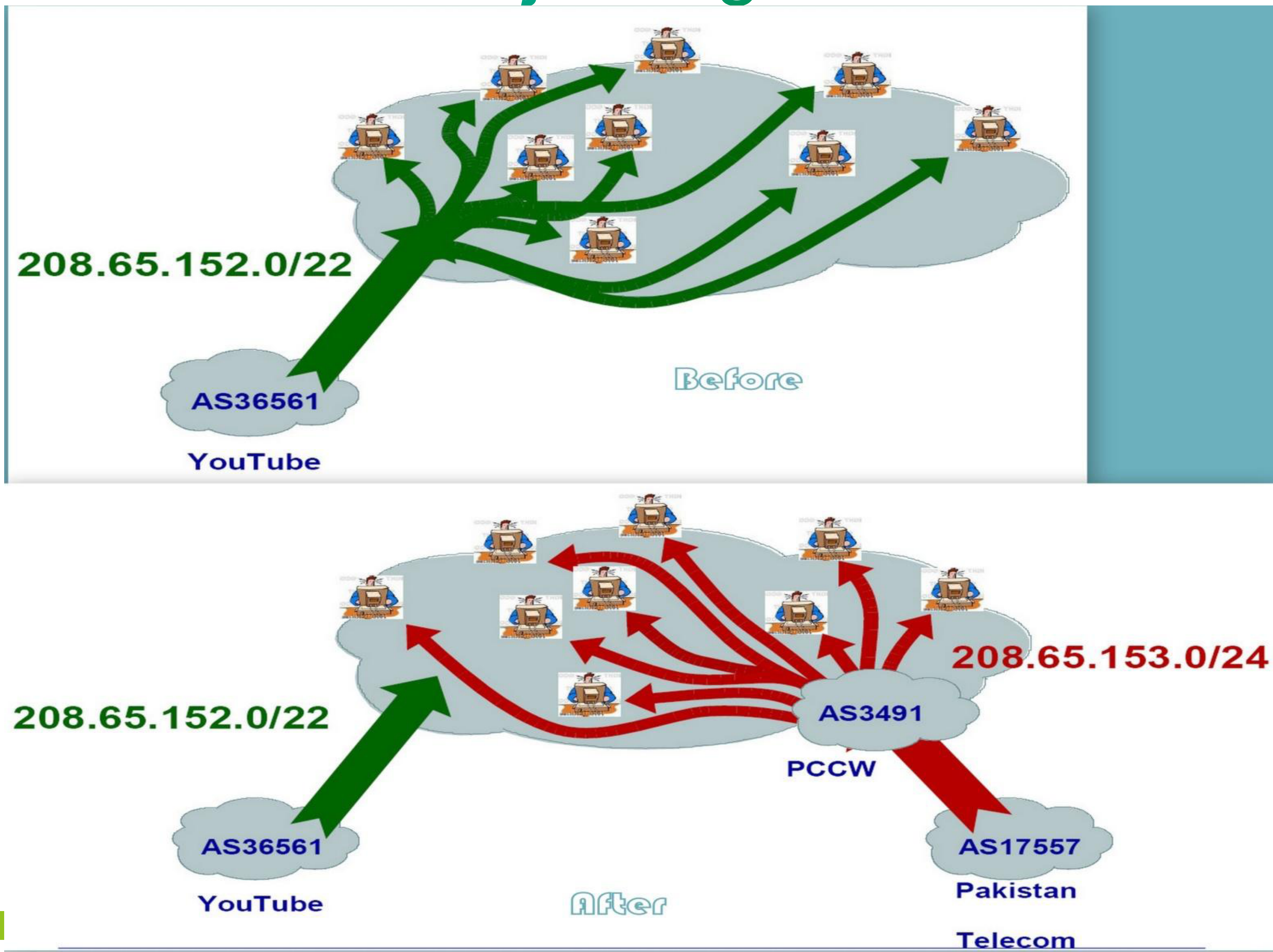
Source: nanog 46 preso

Same Prefix: More specific wins



Source: nanog 46 preso

Youtube Prefix Hijacking



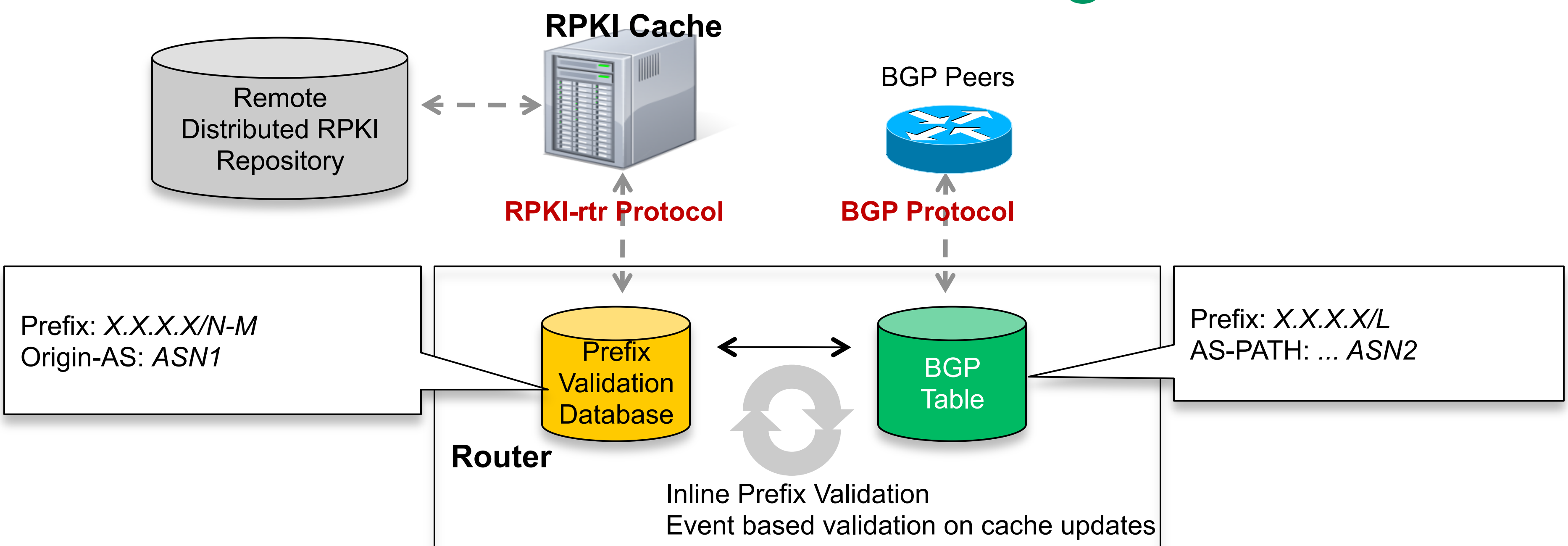
Standardization: IETF

- IETF Security Inter Domain Routing WG
  - Focus on Inter Provider Internet Security
- Origin-AS Validation
  - http://datatracker.ietf.org/wg/sidr/
  - draft-ietf-sidr-pfx-validate-10.txt
  - draft-ietf-sidr-rpki-rtr-26.txt
  - RFC6483

RPKI (Resource Public Key Infrastructure)

- RPKI is a globally distributed database containing, among other things, information mapping BGP (Internet) prefixes to their authorized origin-AS numbers
- Routers running BGP can connect to the RPKI to validate the origin-AS of BGP paths

RPKI Database and BGP Design



ROA: Route Origin Authorization

- Input for the RPKI database for a BGP path:
  - BGP prefix/mask-length (X.X.X.X/N or X:X::X/N)
  - Origin-AS
- If the BGP prefix/mask-length has no covering ROAs in the RPKI database, the validity of path is “unknown”
- If the BGP prefix is covered by one or more ROAs in the RPKI database,
  - If any of the covering ROAs maps to the input origin-AS, the validity of the BGP route is “valid”
  - If none of the covering ROAs map to the input origin-AS, the validity of the BGP route is “invalid”

Origin-AS Validity Check Example

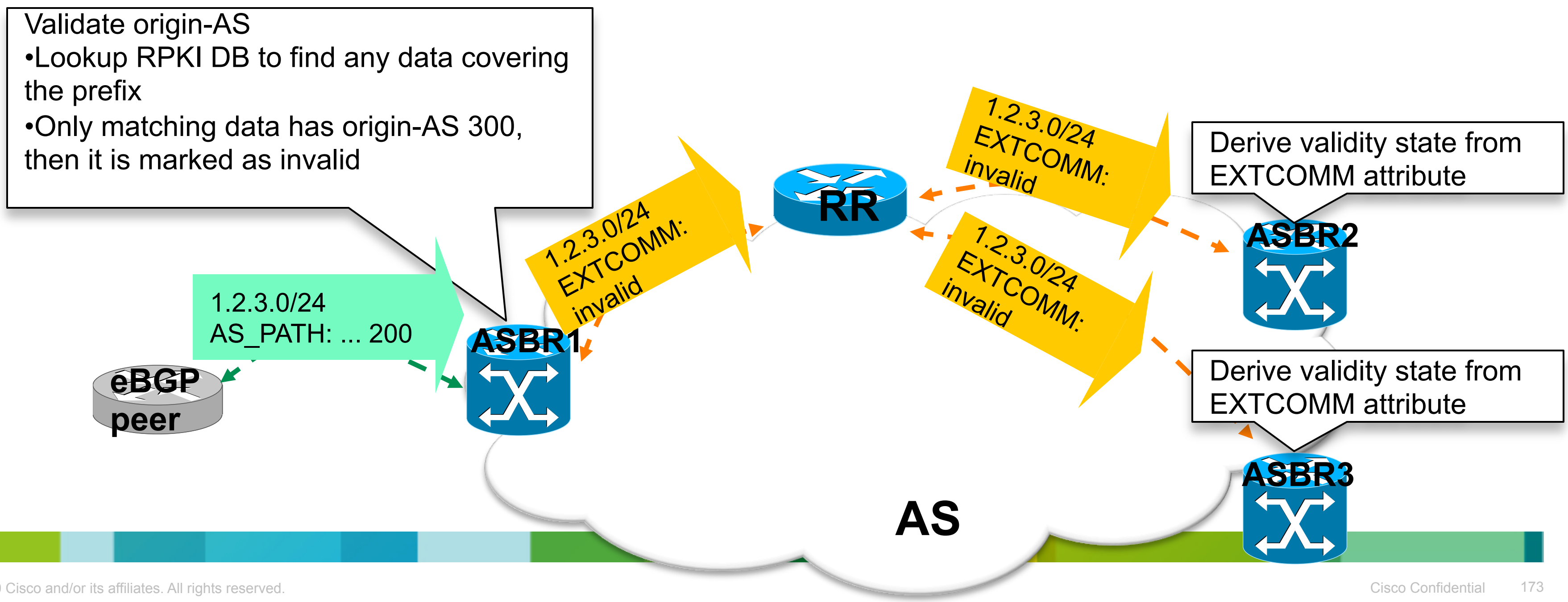
BGP Prefix / Origin-AS		RPKI Database ROAs		
10.0.1/24	AS 300	10/8-20	AS 100	Does not cover BGP prefix
valid		10.0/16-24	AS 200	Cover BGP prefix
		10.0/16-32	AS 300	Cover BGP prefix / **Origin AS matches**

BGP Prefix / Origin-AS		RPKI Database ROAs		
10.0.1/24	AS 400	10/8-20	AS 100	Does not cover BGP prefix
invalid		10.0/16-24	AS 200	Cover BGP prefix
		10.0/16-32	AS 300	Cover BGP prefix

BGP Prefix / Origin-AS		RPKI Database ROAs		
20.0.1/24	AS 500	10/8-20	AS 100	Does not cover BGP prefix
unknown		10.0/16-24	AS 200	Does not cover BGP prefix

iBGP Signaling of Origin-AS Validity State

- When a BGP route is received from outside AS, ASBRs should check this received path for origin-AS validity
- ASBRs that validates the origin-AS should signal the validity state of the route to its iBGP peers through a non-transitive BGP extended community attribute
- Upon receiving validity state information via extended community, iBGP peers can derive the validity state without having to lookup RPKI database



Preparing the Network

Before we begin …

Preparing the Network

- We will deploy BGP across the network before we try and multihome
- BGP will be used; therefore an ASN is required
- If multihoming to different ISPs, public ASN needed:
  - Either go to upstream ISP who is a registry member, or
  - Apply to the RIR yourself for a one off assignment, or
  - Ask an ISP who is a registry member, or
  - Join the RIR and get your own IP address allocation too
    - (this option strongly recommended)!

Preparing the Network
Initial Assumptions

- The network is not running any BGP at the moment
  - single statically routed connection to upstream ISP
- The network is not running any IGP at all
  - Static default and routes through the network to do "routing"

Preparing the Network
First Step: IGP

- Decide on an IGP: OSPF or ISIS ☺
- Assign loopback interfaces and /32 address to each router which will run the IGP
  - Loopback is used for OSPF and BGP router id anchor
  - Used for iBGP and route origination
- Deploy IGP (e.g. OSPF)
  - IGP can be deployed with NO IMPACT on the existing static routing
  - e.g. OSPF distance might be 110m static distance is 1
  - Smallest distance wins

Preparing the Network
IGP (cont)

- Be prudent deploying IGP – keep the Link State Database Lean!
    - Router loopbacks go in IGP
    - WAN point to point links go in IGP
    - (In fact, any link where IGP dynamic routing will be run should go into IGP)
    - Summarise on area/level boundaries (if possible) – i.e. think about your IGP address plan

Preparing the Network
IGP (cont)

- Routes which don't go into the IGP include:
  - Dynamic assignment pools (DSL/Cable/Dial)
  - Customer point to point link addressing
    - (using next-hop-self in iBGP ensures that these do NOT need to be in IGP)
  - Static/Hosting LANs
  - Customer assigned address space
  - Anything else not listed in the previous slide

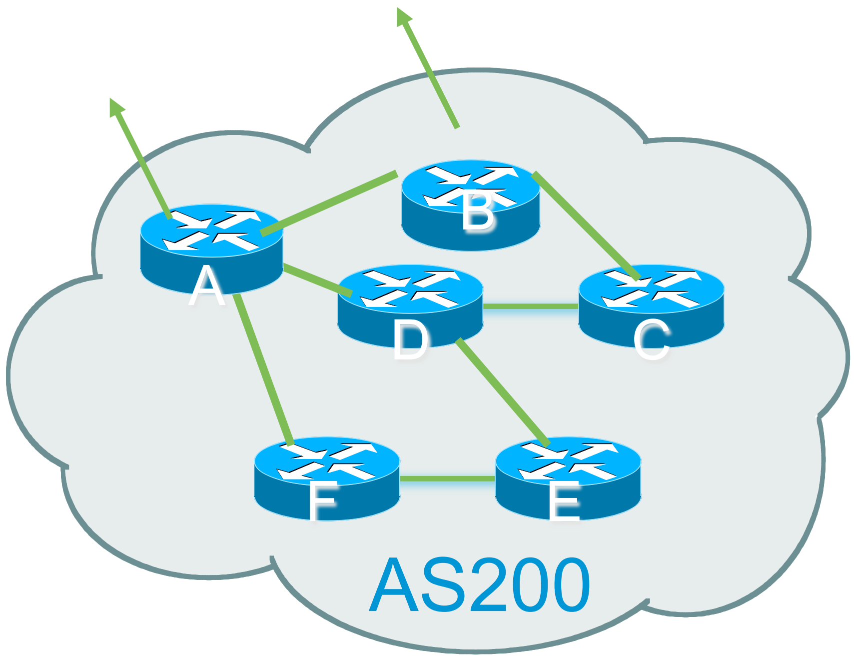
Preparing the Network
Second Step: iBGP

- Second step is to configure the local network to use iBGP
- iBGP can run on

  all routers, or

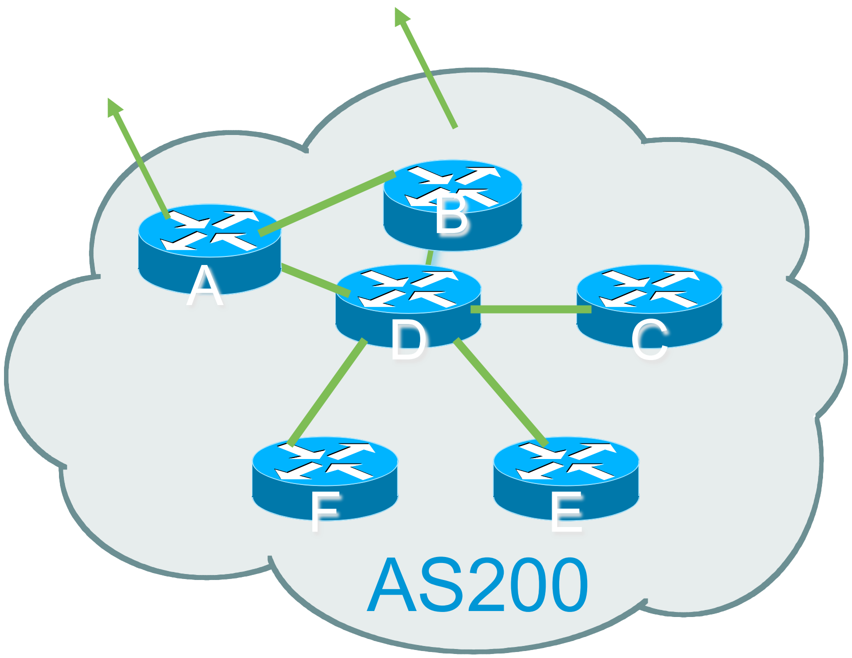
  a subset of routers, or

  just on the upstream edge
- *iBGP must run on all routers which are in the transit path between external connections*



Preparing the Network
Second Step: iBGP (Transit Path)

- *iBGP must run on all routers which are in the transit path between external connections*
- Routers C, E and F are not in the transit path
  - Static routes or IGP will suffice
- Router D is in the transit path
  - Will need to be in iBGP mesh, otherwise routing loops will result



Preparing the Network Layers

- Typical SP networks have three layers:
  - Core – the backbone, usually the transit path
  - Distribution – the middle, PoP aggregation layer
  - Aggregation – the edge, the devices connecting customers

Preparing the Network Aggregation Layer

- iBGP is optional
  - Many ISPs run iBGP here, either partial routing (more common) or full routing (less common)
  - Full routing is not needed unless customers want full table
  - Partial routing is cheaper/easier, might usually consist of internal prefixes and, optionally, external prefixes to aid external load balancing
    - Communities and peer-groups make this administratively easy
- Many aggregation devices can't run iBGP
  - Static routes from distribution devices for address pools
  - IGP for best exit

Preparing the Network Distribution Layer

- Usually runs iBGP
  - Partial or full routing (as with aggregation layer)
- But does not have to run iBGP
  - IGP is then used to carry customer prefixes (does not scale)
  - IGP is used to determine nearest exit
- Networks which plan to grow large should deploy iBGP from day one
  - Migration at a later date is extra work
  - No extra overhead in deploying iBGP.
  - Indeed IGP benefits

Preparing the Network
Core Layer

- Core of network is usually the transit path
- iBGP necessary between core devices
  - Full routes or partial routes:
    - Transit ISPs carry full routes in core
    - Edge ISPs carry partial routes only
- Core layer includes AS border routers

Preparing the Network
iBGP Implementation

Decide on:

- Best iBGP policy
  - Will it be full routes everywhere, or partial, or some mix?
- iBGP scaling technique
  - Community policy?
  - Route-reflectors?
  - Configuration templates such as neighbor groups, sessions groups?

Preparing the Network
iBGP Implementation

- Then deploy iBGP:

  Step 1: Introduce iBGP mesh on chosen routers

  make sure that iBGP distance is greater than IGP distance (it usually is)

  Step 2: Install "customer" prefixes into iBGP

  Check! Does the network still work?

  Step 3: Carefully remove the static routing for the prefixes now in IGP and iBGP

  Check! Does the network still work?

  Step 4: Deployment of eBGP follows

Preparing the Network
iBGP Implementation

Install “customer” prefixes into iBGP?

- Customer assigned address space

  Network statement/static route combination

  Use unique community to identify customer assignments

- Customer facing point-to-point links

  Redistribute connected through filters which only permit point-to-point link addresses to enter iBGP

  Use a unique community to identify point-to-point link addresses (these are only required for your monitoring system)

- Dynamic assignment pools & local LANs

  Simple network statement will do this

  Use unique community to identify these networks

Preparing the Network
iBGP Implementation

Carefully remove static routes?

- Work on one router at a time:

  Check that static route for a particular destination is also learned by the iBGP

  If so, remove it

  If not, establish why and fix the problem

  (Remember to look in the RIB, not the FIB!)

- Then the next router, until the whole PoP is done
- Then the next PoP, and so on until the network is now dependent on the IGP and iBGP you have deployed

Preparing the Network Completion

- Previous steps are NOT flag day steps
  - Each can be carried out during different maintenance periods, for example:
  - Step One on Week One
  - Step Two on Week Two
  - Step Three on Week Three
  - And so on
  - And with proper planning will have NO customer visible impact at all

Preparing the Network Example Two

- The network is not running any BGP at the moment

  single statically routed connection to upstream ISP

- The network is running an IGP though

  All internal routing information is in the IGP

  By IGP, OSPF or ISIS is assumed

Preparing the Network
IGP

- If not already done, assign loopback interfaces and /32 addresses to each router which is running the IGP
  - Loopback is used for OSPF and BGP router id anchor
  - Used for iBGP and route origination
- Ensure that the loopback /32s are appearing in the IGP

Preparing the Network
iBGP

- Go through the iBGP decision process as in Example One
- Decide full or partial, and the extent of the iBGP reach in the network

Preparing the Network
iBGP Implementation

- Then deploy iBGP:

  Step 1: Introduce iBGP mesh on chosen routers

  make sure that iBGP distance is greater than IGP distance (it usually is)

  Step 2: Install "customer" prefixes into iBGP

  Check! Does the network still work?

  Step 3: Reduce BGP distance to be less than the IGP

  (so that iBGP routes take priority)

  Step 4: Carefully remove the "customer" prefixes from the IGP

  Check! Does the network still work?

  Step 5: Restore BGP distance to less than IGP

  Step 6: Deployment of eBGP follows

Preparing the Network
iBGP implementation

Install "customer" prefixes into iBGP?

- Customer assigned address space
  - Network statement/static route combination
  - Use unique community to identify customer assignments
- Customer facing point-to-point links
  - Redistribute connected through filters which only permit point-to-point link addresses to enter iBGP
  - Use a unique community to identify point-to-point link addresses (these are only required for your monitoring system)
- Dynamic assignment pools & local LANs
  - Simple network statement will do this
  - Use unique community to identify these networks

Preparing the Network
iBGP implementation

Carefully remove "customer" routes from IGP?

- Work on one router at a time:
  - Check that IGP route for a particular destination is also learned by iBGP
  - If so, remove it from the IGP
  - If not, establish why and fix the problem
  - (Remember to look in the RIB, not the FIB!)
- Then the next router, until the whole PoP is done
- Then the next PoP, and so on until the network is now dependent on the iBGP you have deployed

Preparing the Network Completion

- Previous steps are NOT flag day steps

  Each can be carried out during different maintenance periods, for example:

  Step One on Week One

  Step Two on Week Two

  Step Three on Week Three

  And so on

  And with proper planning will have NO customer visible impact at all

Preparing the Network Configuration Summary

- IGP essential networks are in IGP
- Customer networks are now in iBGP
  - iBGP deployed over the backbone
  - Full or Partial or Upstream Edge only
- iBGP distance is greater than any IGP
- Now ready to deploy eBGP

Configuration Tips

iBGP and IGPs Reminder!

- Make sure loopback is configured on router
  - iBGP between loopbacks, NOT real interfaces
- Make sure IGP carries loopback /32 address
- Consider the DMZ nets:
  - Use unnumbered interfaces?
  - Use next-hop-self on iBGP neighbours
  - Or carry the DMZ /30s in the iBGP
  - Basically keep the DMZ nets out of the IGP!

iBGP: Next-hop-self

- BGP speaker announces external network to iBGP peers using router’s local address (loopback) as next-hop
- Used by many ISPs on edge routers
  - Preferable to carrying DMZ /30 addresses in the IGP
  - Reduces size of IGP to just core infrastructure
  - Alternative to using unnumbered interfaces
  - Helps scale network
  - Many ISPs consider this “best practice”

Limiting AS Path Length

- Some BGP implementations have problems with long AS_PATHS
  - Memory corruption
  - Memory fragmentation
- Even using AS_PATH prepends, it is not normal to see more than 20 ASes in a typical AS_PATH in the Internet today
- July 26, 2012 Internet AS path report for AS6447 (http://bgp.potaroo.net/as6447/) shows that
  - Average AS path length is 3.8
  - Maximum AS path length is 13
  - Maximum prepended AS path length is 34

Limiting AS Path Length

- Some announcements have ridiculous lengths of AS-paths:

```
*> 3FFE:1600::/24       22 11537 145 12199 10318
10566 13193 1930 2200 3425 293 5609 5430 13285 6939
14277 1849 33 15589 25336 6830 8002 2042 7610 i
```

  This example is an error in one IPv6 implementation

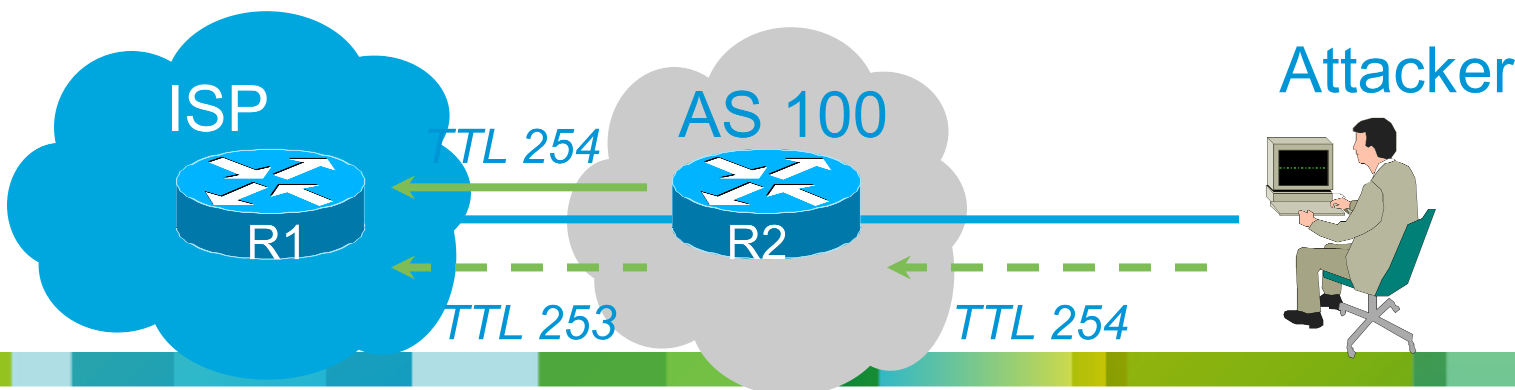
```
*>  96.27.246.0/24       2497 1239 12026 12026 12026
12026 12026 12026 12026 12026 12026 12026 12026 12026
12026 12026 12026 12026 12026 12026 12026 12026 12026
12026 i
```

  This example shows 21 prepends (for no obvious reason)

- If your implementation supports it, consider limiting the maximum AS-path length you will accept

Generalized TTL Security Mechanism (GTSM)

- Also known as BGP TTL Security "Hack" (BTSH)
- Implement RFC5082 on BGP peerings
  - Neighbour sets TTL to 255
  - Local router expects TTL of incoming BGP packets to be 254
  - No one apart from directly attached devices can send BGP packets which arrive with TTL of 254, so any possible attack by a remote miscreant is dropped due to TTL mismatch
  - Some implementations drop it in HW without any CPU impact



Generalized TTL Security Mechanism

- GTSM:
  - Both neighbors must agree to use the feature
  - TTL check is much easier to perform than MD5
- Provides "security" for BGP sessions
  - In addition to packet filters of course
  - MD5 should still be used for messages which slip through the TTL hack
  - See www.nanog.org/mtg-0302/hack.html for more details

Templates

- Good practice to configure templates for everything
  - Vendor defaults tend not to be optimal or even very useful for ISPs
  - ISPs create their own defaults by using configuration templates
- eBGP and iBGP examples follow
  - Also see Team Cymru's BGP templates
    - http://www.team-cymru.org/ReadingRoom/Documents/

iBGP Template Example

- iBGP between loopbacks!
- Next-hop-self
  - Keep DMZ and external point-to-point out of IGP
- Always send communities in iBGP
  - Otherwise accidents will happen
- Hardwire BGP to version 4, if there is a version configuration option
  - Yes, this is being paranoid!

iBGP Template
Example continued

- Use passwords on iBGP session
  - Not being paranoid, VERY necessary
  - It's a secret shared between you and your peer
  - If arriving packets don't have the correct MD5 hash, they are ignored
  - Helps defeat miscreants who wish to attack BGP sessions – particularly, from man-in-the-middle type of attack
- Powerful preventative tool, especially when combined with filters and the TTL "hack"

eBGP Template Example

- Remove private ASes from announcements
  - Common omission today
- Use extensive filters, with “backup”
  - Use as-path filters to backup prefix filters
  - Keep policy language for implementing policy, rather than basic filtering
- Use password agreed between you and peer on eBGP session
- Use TTL security (GTSM) if both peers support it

eBGP Template
Example continued

- Use maximum-prefix tracking

  Router will warn you if there are sudden increases in BGP table size, bringing down eBGP if desired

- Limit maximum as-path length inbound
- Log changes of neighbor state

  …and monitor those logs!

- Either make BGP admin distance higher than that of any IGP, or make sure to block your own prefixes inbound,

  Otherwise prefixes heard from outside your network could override your IGP!!

Summary

- Use configuration templates
- Standardise the configuration
- Be aware of standard “tricks” to avoid compromise of the BGP session
- Anything to make your life easier, network less prone to errors, network more likely to scale
- It’s all about scaling – if your network won’t scale, then it won’t be successful

Thank you.

CISCO