2006-2: Micro-allocations for Internal Infrastructure

Heather Skanks (heather.skanks@mci.com)
Jason Schiller (schiller@uu.net)
Chris Morrow (chris@uu.net)

History..

- Previously reviewed in April
- 14 voted for
- 16 voted against
- 29 supported with revisions
- A significant number of folks abstained
- Why?

BGP Re-convergence

- Router /32 loopbacks are in the IGP
- 137.39.0.0/16 is a "pull-up" route in the core
- "Pull-up" routes are re-distributed into BGP

- Customer advertises 192.0.2.0/24 MED of 0 via eBGP across primary link to Edge Router 1

- Edge Router 1 advertises 192.0.2.0/24 MED 0 via iBGP and sets next-hop self (137.39.1.1)

- Customer advertises 192.0.2.0/24 MED of 10 via eBGP across secondary link to Edge Router 2

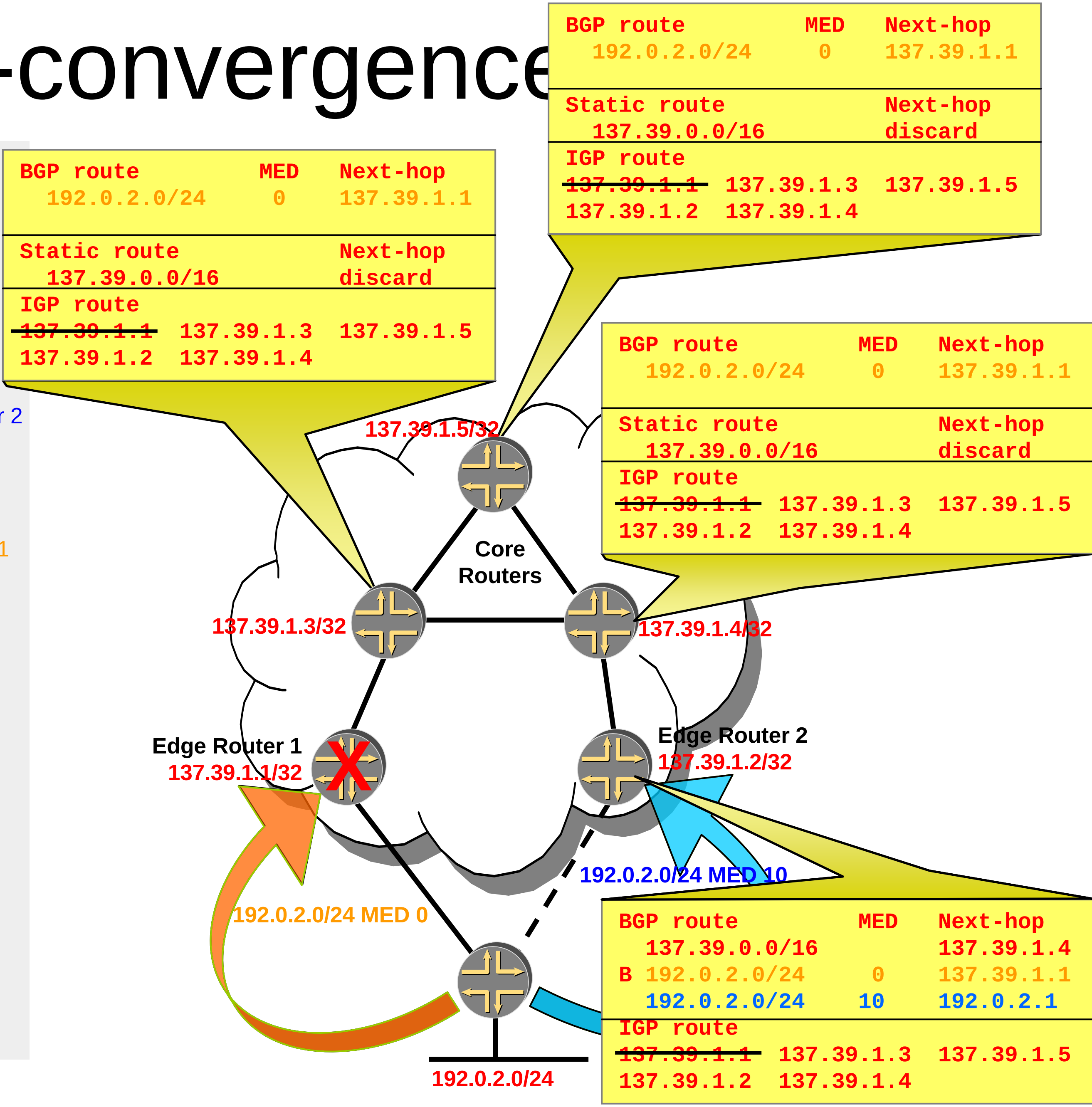
- Edge Router 2 learns 192.0.2.0/24 MED 10 but the best path is 192.0.2.0/24 MED 0 learned via iBGP, so no announcement is made

- All routers select 192.0.2.0/24 MED 0 137.39.1.1 as the best BGP route to the customer

- Edge Router 1 fails

- 137.39.1.1/32 is removed from IGP

- The network still has best BGP route 192.0.2.0/24 MED 0 NH 137.39.1.1
- 137.39.1.1 is reachable through the route to 137.39.0.0/16
- Traffic is drawn to core routers and discarded



BGP route		MED	Next-hop
192.0.2.0/24		0	137.39.1.1
Static route			Next-hop
137.39.0.0/16			discard
IGP route			
~~137.39.1.1~~	137.39.1.3		137.39.1.5
137.39.1.2	137.39.1.4		

BGP route		MED	Next-hop
192.0.2.0/24		0	137.39.1.1
Static route			Next-hop
137.39.0.0/16			discard
IGP route			
~~137.39.1.1~~	137.39.1.3		137.39.1.5
137.39.1.2	137.39.1.4		

BGP route		MED	Next-hop
192.0.2.0/24		0	137.39.1.1
Static route			Next-hop
137.39.0.0/16			discard
IGP route			
~~137.39.1.1~~	137.39.1.3		137.39.1.5
137.39.1.2	137.39.1.4		

BGP route		MED	Next-hop
137.39.0.0/16			137.39.1.4
B 192.0.2.0/24		0	137.39.1.1
192.0.2.0/24		10	192.0.2.1
IGP route			
~~137.39.1.1~~	137.39.1.3		137.39.1.5
137.39.1.2	137.39.1.4		

Policy Revision

- Some confusion resulted from an editorial rewrite of the existing Mirco-allocation policy
- While the rewrite consisted of cut and past, many thought it was an attempt to change the current policy
- Editorial rewrites have been removed from this revision
- Removed the requirement that the micro-allocation for critical infrastructure MUST not be routed

What has changed?

- Proposal type: modify
- Policy term: permanent
- Policy statement:
- 6.10.1 Micro-allocations for Internal Infrastructure
- Organizations that currently hold IPv6 allocations may apply for a micro-allocation for internal infrastructure. Applicant must provide technical justification indicating why a separate non-routed block is required. Justification must include why a sub-allocation of currently held IP space cannot be utilized. Internal infrastructure allocations must be allocated from specific blocks reserved only for this purpose.

Overview

- 2006-2 allows for an additional non-contiguous /48 IPv6 allocation to organizations that already have v6 space, if needed for internal infrastructure
- Alleviate BGP convergence issue
  - 3 min black-holing of sensitive traffic such as VoIP
- Address security considerations

What happens.. Looking at IPv4 for a moment..

- You go to ARIN and get a netblock

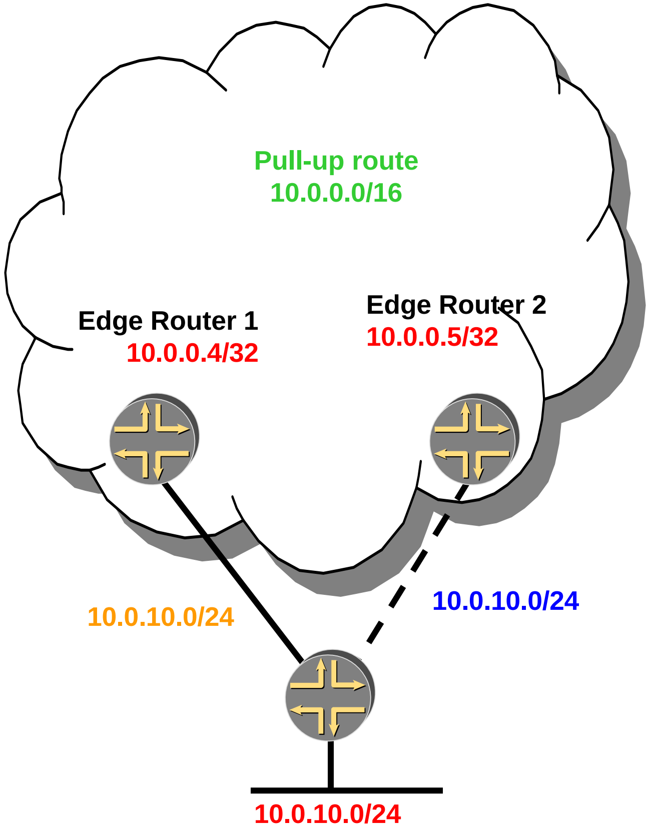
  Example: 10.0.0.0/16
- You pullup or advertise the whole netblock to the world in order to draw traffic to you.
- You use this for all your addressing needs. Giving some for your infrastructure and some to your customers

Example: You give your customer Acme VoIP 10.0.10.0/24

The problem..

- Acme VoIP wants to buy 2 connections to you, in different places. One as a primary and the other as a fail over.
  - They announce the netblock you gave them, out of both connections. The edge router tells all it's neighbors, Acme VoIP's /24 (10.0.10.0/24) is reachable through it's loopback, example: 10.0.0.4/32
  - You expect when the primary connection goes down, the other connection should pick up.
  - But.. since you numbered the loopback for the edge router (10.0.0.4/32), out of the same aggregate you got (10.0.0.0/16) when the primary edge router goes down, its loopback still looks reachable through the pull up route you put in place to draw traffic to your /16.

If I had a whiteboard..



So…? What happens?

- When your connection is down, instead of going to your failover connection, traffic goes to the pullup until the iBGP session times out, and tells everyone that the primary connection is down.
- It can take up to 3 minutes for that session to time out.
- Applications such as VoIP and payment processing are sensitive to this 3 minute downtime.
- If your routing table has a less specific route (like a pullup) for your BGP next-hops then you have this problem

Solution Considerations and IPv6

- Solution required netblocks for loopbacks that would not be part of the aggregate pullup that goes into BGP
- Not a problem for IPv4 as loopbacks can easily be in a separate block which is not pulled-up, since many folks have multiple IPv4 netblocks
- ARIN policy provides for only a single IPv6 block, and is set up to give you contiguous address space should you need it
- Breaking up the IPv6 netblock you get from ARIN, into smaller pieces adds to the internet routing table, and does not uphold the principle that IPv6 aggregation is important for IPv6 stewardship

Private Address Considerations

- Global uniqueness is not guaranteed
- Private addresses in traceroutes across the public Internet may create confusion
- If routers source ICMP messages with private addresses, and there is wide spread packet filtering of private addresses, then confusion and additional problems troubleshooting may result
- For reverse DNS to work for private addresses requires split plane DNS and hijacking of IANA’s authority of the reverse zones

Staff Concerns

- Applicant must provide technical justification indicating why a separate non-routed block is required
- Justification must include why a sub-allocation of currently held IP space cannot be utilized.
- Staff may want to consider using the term unique routing policy which requires a separate block

Technical Justification

- ARIN staff believes the policy is not entirely clear. Specifically, the term 'justification' is broad and requires further definition.
- AC thought the example of justification was very narrow
- Where there is a need to solve the convergence issue
- Where there is need to number internal infrastructure and need to have DNS
- If the space is used to solve this convergence problem, it should not be seen in the internet routing table, as that would rather defeat the purpose and not solve the problem!

Staff Concerns

- ARIN cannot verify the blocks allocated under this policy are not routed.
- Existing IPv4 policy allowing non-connected networks to use globally unique space
- Organizations that filter ping
- Applicants should provide documentation on how the internal infrastructure block is being used – including internal routing information

Staff Concerns

- In the event that these blocks are routed, staff proposes revoking the number resources.
- The authors are acceptable to this suggestion, however it may be outside the charter of ARIN as ARIN does not set routing policy – but may already be covered
- Similar text was proposed recently on PPML and last April
  - Both times the members suggested this text should not be part of the policy
  - The authors accepted this rewrite and removed the text
- 4.1.3 ARIN may invalidate any IP allocation if it determines that the requirement for the address space no longer exists.

Other Comments on PPML

- Will have zero impact on the Internet routing table
- Policy is not very wasteful
- Members are opposed to a sunset clause
- Members see value in global uniqueness even for networks that do not connect to the Internet

Questions?

Its already fixed in the protocol

- BGP re-convergence rational is outdated
  - Juniper has support for route resolution policies
- Not all vendors support this functionality is stable code
- In the process of drafting an RFC with Cisco to make this functionality standard
  - But that will take some time to pass the standards process, and get implemented
- Policy can solve this problem now

Crazy Detailed Backup Slides

From the previous presentation in case anyone is actually interested in all the details.

BGP Re-convergence Problem

- If a route to a destination has a protocol next-hop that is reachable through a pull-up or less specific route, then the route to that destination will never be invalidated due to next-hop unreachability
- Must wait for the iBGP sessions with the failed edge device to time out (up to 3 min hold timer)
- If your routing table has a less specific route for your BGP protocol Next-hops then you have this problem

BGP Re-convergence Problem

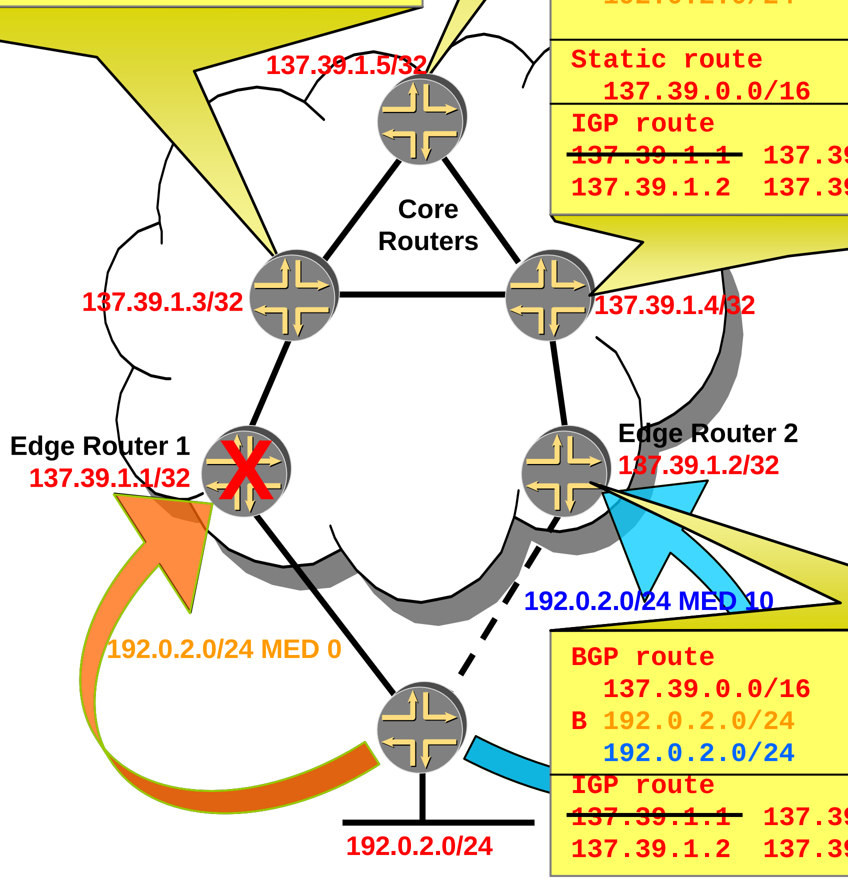
- Take a multi-homed customer with prefix 192.0.2.0/24 connected to two different ISP edge routers (edge router 1 and edge router 2)
- Assume the connection to edge router 1 is a primary link with an eBGP announcement of 192.0.2.0/24 with a MED of 0
- Assume the connection to edge router 2 is a secondary link with an eBGP announcement of 192.0.2.0/24 with a MED of 10
- Assume both edge routers set next-hop self
- Assume that there is a “pull-up” or aggregate route that is less specific than the edge routers’ loopback IP address

Security Considerations

- Non routed internal only addresses can be used for internal only services
  - iBGP
  - SNMP
  - Radius / TACACS
  - OOB management
- Two tiered approach to network security
  - Can reduce many attacks to internal infrastructure in control plane by not routing the internal address
  - Additional forwarding filters can be easily constructed by the uniqueness of internal only address block

BGP Re-convergence

- Router /32 loopbacks are in the IGP
- 137.39.0.0/16 is a "pull-up" route in the core
- "Pull-up" routes are re-distributed into BGP
- Customer advertises 192.0.2.0/24 MED of 0 via eBGP across primary link to Edge Router 1
- Edge Router 1 advertises 192.0.2.0/24 MED 0 via iBGP and sets next-hop self (137.39.1.1)
- Customer advertises 192.0.2.0/24 MED of 10 via eBGP across secondary link to Edge Router 2
- Edge Router 2 learns 192.0.2.0/24 MED 10 but the best path is 192.0.2.0/24 MED 0 learned via iBGP, so no announcement is made
- All routers select 192.0.2.0/24 MED 0 137.39.1.1 as the best BGP route to the customer
- Edge Router 1 fails
- 137.39.1.1/32 is removed from IGP
- The network still has best BGP route 192.0.2.0/24 MED 0 NH 137.39.1.1
- 137.39.1.1 is reachable through the route to 137.39.0.0/16
- Traffic is drawn to core routers and discarded



BGP route	MED	Next-hop
192.0.2.0/24	0	137.39.1.1
Static route		Next-hop
137.39.0.0/16		discard
IGP route		
~~137.39.1.1~~	137.39.1.3	137.39.1.5
137.39.1.2	137.39.1.4	

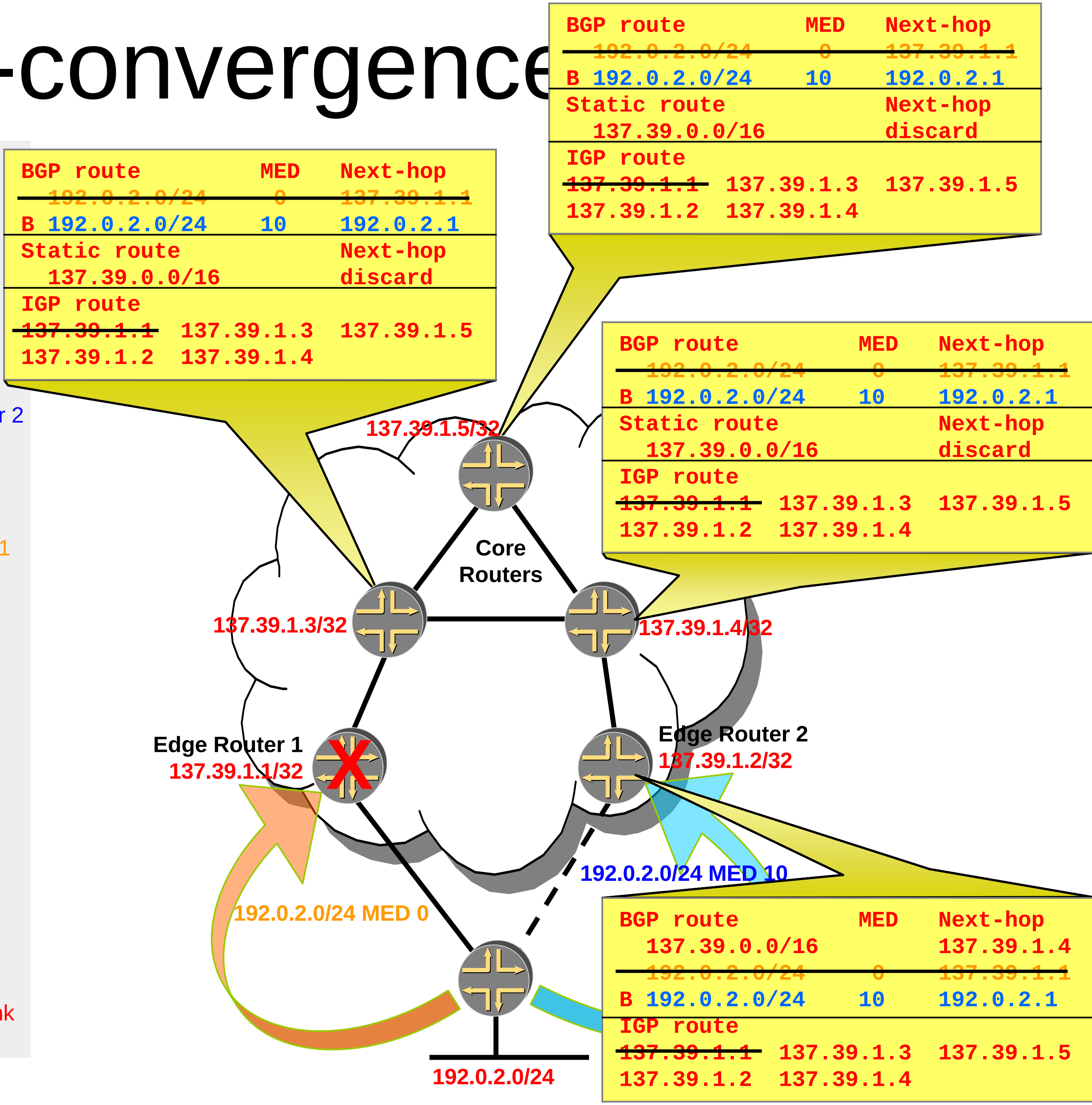
BGP route	MED	Next-hop
192.0.2.0/24	0	137.39.1.1
Static route		Next-hop
137.39.0.0/16		discard
IGP route		
~~137.39.1.1~~	137.39.1.3	137.39.1.5
137.39.1.2	137.39.1.4	

BGP route	MED	Next-hop
192.0.2.0/24	0	137.39.1.1
Static route		Next-hop
137.39.0.0/16		discard
IGP route		
~~137.39.1.1~~	137.39.1.3	137.39.1.5
137.39.1.2	137.39.1.4	

BGP route	MED	Next-hop
137.39.0.0/16		137.39.1.4
B 192.0.2.0/24	0	137.39.1.1
192.0.2.0/24	10	192.0.2.1
IGP route		
~~137.39.1.1~~	137.39.1.3	137.39.1.5
137.39.1.2	137.39.1.4	

BGP Re-convergence

- Router /32 loopbacks are in the IGP
- 137.39.0.0/16 is a “pull-up” route in the core
- “Pull-up” routes are re-distributed into BGP
  - Customer advertises 192.0.2.0/24 MED of 0 via eBGP across primary link to Edge Router 1
  - Edge Router 1 advertises 192.0.2.0/24 MED 0 via iBGP and sets next-hop self (137.39.1.1)
  - Customer advertises 192.0.2.0/24 MED of 10 via eBGP across secondary link to Edge Router 2
  - Edge Router 2 learns 192.0.2.0/24 MED 10 but the best path is 192.0.2.0/24 MED 0 learned via iBGP, so no announcement is made
  - All routers select 192.0.2.0/24 MED 0 137.39.1.1 as the best BGP route to the customer
  - Edge Router 1 fails
  - 137.39.1.1/32 is removed from IGP
  - The network still has best BGP route 192.0.2.0/24 MED 0 NH 137.39.1.1
  - 137.39.1.1 is reachable through the route to 137.39.0.0/16
  - Traffic is drawn to core routers and discarded
  - After 3 mins the iBGP sessions with Edge Router 1 time out
  - Route for 192.0.2.0/24 MED 0 is retracted
  - Edge Router 2 route for 192.0.2.0/24 MED 10 is now best. It advertises 192.0.2.0/24 MED 10 via iBGP and sets next-hop self (137.39.1.2)
  - Traffic is forwarded to CPE across secondary link



BGP route	MED	Next-hop
~~192.0.2.0/24~~	~~0~~	~~137.39.1.1~~
B 192.0.2.0/24	10	192.0.2.1
Static route		Next-hop
137.39.0.0/16		discard
IGP route		
~~137.39.1.1~~	137.39.1.3	137.39.1.5
137.39.1.2	137.39.1.4	

BGP route	MED	Next-hop
~~192.0.2.0/24~~	~~0~~	~~137.39.1.1~~
B 192.0.2.0/24	10	192.0.2.1
Static route		Next-hop
137.39.0.0/16		discard
IGP route		
~~137.39.1.1~~	137.39.1.3	137.39.1.5
137.39.1.2	137.39.1.4	

BGP route	MED	Next-hop
~~192.0.2.0/24~~	~~0~~	~~137.39.1.1~~
B 192.0.2.0/24	10	192.0.2.1
Static route		Next-hop
137.39.0.0/16		discard
IGP route		
~~137.39.1.1~~	137.39.1.3	137.39.1.5
137.39.1.2	137.39.1.4	

BGP route	MED	Next-hop
137.39.0.0/16		137.39.1.4
~~192.0.2.0/24~~	~~0~~	~~137.39.1.1~~
B 192.0.2.0/24	10	192.0.2.1
IGP route		
~~137.39.1.1~~	137.39.1.3	137.39.1.5
137.39.1.2	137.39.1.4	

BGP Re-convergence

- BGP next-hops are not aggregated
- The aggregate of the BGP next-hops are not announced to the Internet

- Router /32 loopbacks are in the IGP
- 137.39.0.0/16 is a “pull-up” route in the core
- “Pull-up” routes are re-distributed into BGP

- Customer advertises 192.0.2.0/24 MED of 0 via eBGP across primary link to Edge Router 1

- Edge Router 1 advertises 192.0.2.0/24 MED 0 via iBGP and sets next-hop self (157.130.1.1)

- Customer advertises 192.0.2.0/24 MED of 10 via eBGP across secondary link to Edge Router 2

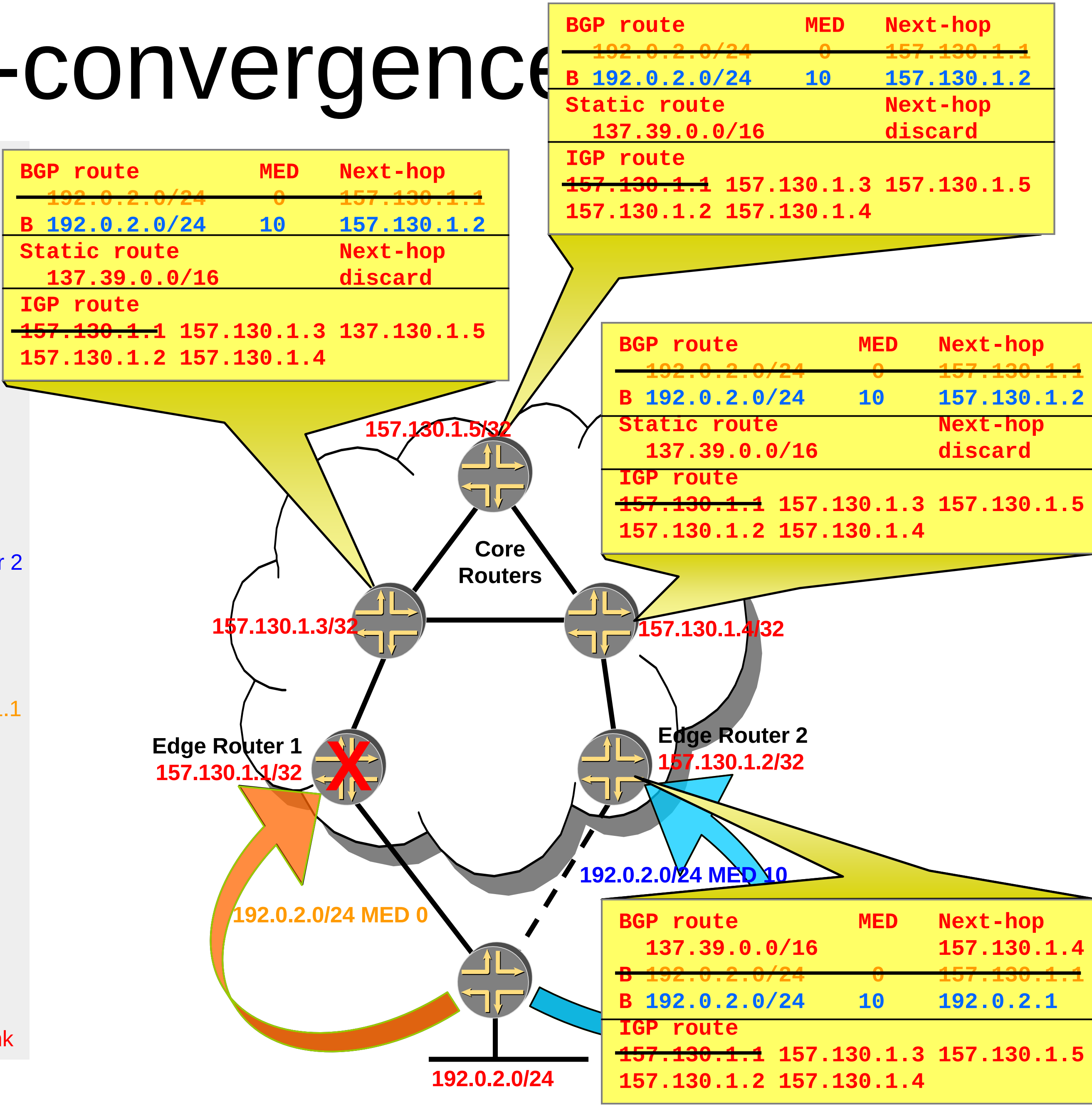
- Edge Router 2 learns 192.0.2.0/24 MED 10 but the best path is 192.0.2.0/24 MED 0 learned via iBGP, so no announcement is made

- All routers select 192.0.2.0/24 MED 0 157.130.1.1 as the best BGP route to the customer

- Edge Router 1 fails

- 157.130.1.1/32 is removed from IGP
- The best BGP route 192.0.2.0/24 MED 0 has an unreachable next-hop (157.130.1.1) and is invalidated

- Edge Router 2 route for 192.0.2.0/24 MED 10 is now best. It advertises 192.0.2.0/24 MED 10 via iBGP and sets next-hop self (157.130.1.2)
- Traffic is forwarded to CPE across secondary link



BGP route	MED	Next-hop
~~192.0.2.0/24~~	~~0~~	~~157.130.1.1~~
B 192.0.2.0/24	10	157.130.1.2
Static route		Next-hop
137.39.0.0/16		discard
IGP route		
~~157.130.1.1~~ 157.130.1.3		157.130.1.5
157.130.1.2 157.130.1.4		

BGP route	MED	Next-hop
~~192.0.2.0/24~~	~~0~~	~~157.130.1.1~~
B 192.0.2.0/24	10	157.130.1.2
Static route		Next-hop
137.39.0.0/16		discard
IGP route		
~~157.130.1.1~~ 157.130.1.3		137.130.1.5
157.130.1.2 157.130.1.4		

BGP route	MED	Next-hop
~~192.0.2.0/24~~	~~0~~	~~157.130.1.1~~
B 192.0.2.0/24	10	157.130.1.2
Static route		Next-hop
137.39.0.0/16		discard
IGP route		
~~157.130.1.1~~ 157.130.1.3		157.130.1.5
157.130.1.2 157.130.1.4		

BGP route	MED	Next-hop
137.39.0.0/16		157.130.1.4
~~B 192.0.2.0/24~~	~~0~~	~~157.130.1.1~~
B 192.0.2.0/24	10	192.0.2.1
IGP route		
~~157.130.1.1~~ 157.130.1.3		157.130.1.5
157.130.1.2 157.130.1.4		