BGP Link Bandwidth Extended Communities

https://datatracker.ietf.org/doc/html/draft-ietf-idr-link-bandwidth-07

IETF 122 IDR Meeting

Mar 17, 2025

Presenters:

Reshma Das (dreshma@juniper.net)

On behalf of authors

Agenda

- Goal
- Solution
- Error Handling
- Current Status
- Next Steps

Goal

- Revive and retain existing draft (draft-ietf-idr-link-bandwidth)
- Achieve interop of existing LBW extended community with minimum changes to procedures.
- Keep the changes to procedure simple.
  - Assumption: Receiver runs upgraded code to be able to interop.
  - Keep RR procedures simple; RR is transparent, doesn't modify LBW when reflecting
- Cover all applicable use cases. (IntraAS, InterAS)
- Define Error handling

Solution (1/2)

A. Sender (originating link bandwidth community) :

An originator of the link bandwidth community SHOULD be able to originate either a transitive or a non-transitive link bandwidth extended community. Implementation SHOULD provide **configuration** to set the transitivity type of the link bandwidth community. No more than one link bandwidth extended community SHALL be attached to a route.

B. Receiver (receiving link bandwidth community) :

A BGP receiver MUST be able to process link bandwidth community of either transitive or non-transitive type. The receiver SHOULD NOT flap or treat the route as malformed based on the transitivity of the link bandwidth community.

Solution (2/2)

C. Re-advertisement with Next hop Self :

When a BGP speaker re-advertises a route with the link bandwidth extended community and sets the next hop to itself, it SHOULD follow the same procedures as outlined as sender.

In the absence of any import or export policies that alter the Link Bandwidth Extended Community, any received Link Bandwidth extended community on the route will be re-advertised unchanged, in accordance with standard BGP procedures.

D. Re-advertisement with Next hop unchanged :

A BGP speaker that receives a route with link bandwidth community and re-advertises or reflects the same without changing its next hop SHOULD NOT change the link bandwidth extended community in any way.

Error Handling (Discussion)

- **Current Draft Text**

If a receiver receives a route with more than one Link Bandwidth Extended Community, it SHOULD:

1. Prefer the lowest value of the attached link bandwidth community, irrespective of the extended community's transitivity.

2. Prefer the transitive link bandwidth extended community when choosing between transitive and non-transitive types that have the same value.

Implementations MAY provide configuration to change the above preferences.

- **WG Discussion on this text:**

The current text lacks clarity about where these rules apply, causing ambiguity for RR behavior. Clarification is needed to say that this error handling applies to the WECMP computation layer in a node and not at BGP send/receive layer. Thus, it does not affect RR behavior.

- **Suggested text to be added:**

If a BGP speaker receives a route with more than one Link Bandwidth Extended Community and uses the route to compute WECMP, it SHOULD prefer the lowest value of the attached link bandwidth community, irrespective of the extended community's transitivity. Implementations MAY provide configuration to change the above preference.

Between transitive and non-transitive types of LBW extended communities that have the same value, the transitivity doesn't matter for purpose of computing WECMP or programming to FIB.

Note that these procedures mean, a BGP speaker reflecting a route with next hop unchanged (e.g. RR) will re-advertises the LBW communities received on the route as-is without any modification, while following the extended community transitivity rules.

Current Status:

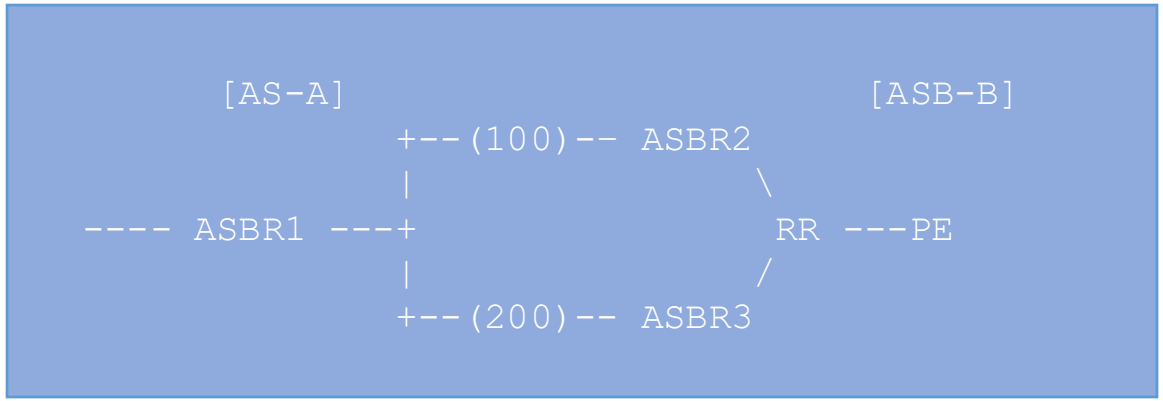
- Customers are requesting both transitive and non-transitive versions to be supported by all vendors, to accommodate their existing deployments
- Juniper implementation now supports sending/receiving both transitive and non-transitive form of LBW extended community
- Incremental support from a few other vendor implementations for both transitive and non-transitive versions are seen.
- Thank you to all vendors participating in the ongoing discussions. [Arista, Cisco, Google, Juniper, Zscaler] .

Next Steps

- Close discussion on Error handling and clarify text
- Standardize the solution
- Follow on work: 4360-bis

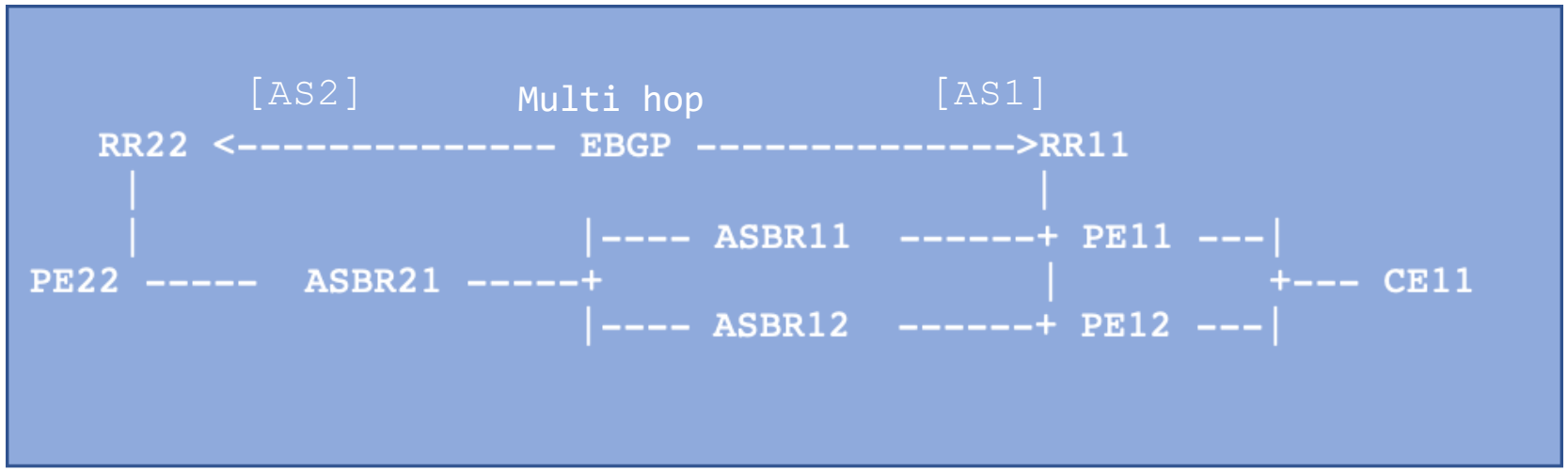
Thank you.

Example Topology from IDR mailing list



- ASBR2,RR,ASBR3,PE initially supports only non-transitive.
- RR is upgraded to understand both transitive and non-transitive.
- ASBR1 (AS_A) supports only transitive and leaked it to ASBR2/ASBR3.
- ASBR2/ASBR3 source non-transitive based on interface lbw 100 and 200 for PE to do UCMP
- Problem does not exist as RR reflects all received LBW community

Example Topology for Transitive LBW Community



- RR22 and RR11 are service RRs connecting PE22 and PE11/PE12.
- RR22 and RR11 is connected via multi-hop EBGP session.
- The load balancing point for service routes in this topology is PE22towards PE11 and PE12.
- The load balancing point for transport routes in this topology is ASBR21 towards ASBR11 and ASBR12.