Interdomain Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 4, 2015

S. Litkowski
Orange Business Service
K. Patel
Cisco Systems
J. Haas
Juniper Networks

July 3, 2014

Timestamp support for BGP paths
draft-litkowski-idr-bgp-timestamp-00

Abstract

BGP is more and more used to transport routing information for critical services. Some BGP updates may be critical to be received as fast as possible : for example, in a layer 3 VPN scenario where a dual-attached site is loosing primary connection, the BGP withdraw message should be propagated as fast as possible to restore the service. The same criticity exists for other address-families like multicast VPNs where "join" messages should also be propagated very fast.

Experience of service providers shows that BGP path propagation time may vary depending on network conditions (especially load of BGP speaker on the path) and too long propagation time are affecting customer service.

It is important for service providers to keep track of BGP updates propagation time to monitor quality of service for the customers. It is also important to be able to identify BGP Speakers that are slowing down the propagation.

This document presents a solution to transport timestamps of a BGP path.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

This Internet-Draft will expire on January 4, 2015.

Table of Contents

1. Problem statement

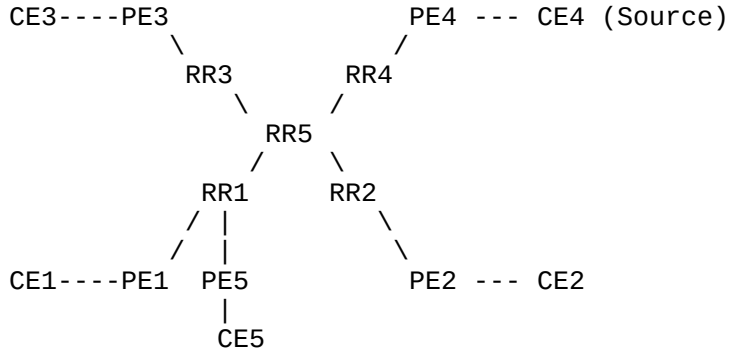


Figure 1

The figure 1 describes a typical hierarchical RR design where PEs are meshed to local RRs and local RRs are meshed to more centric RRs. We consider a single multicast VPN between all CEs. CE4 is the source, all others may be receivers. The BGP controlplane also supports some other BGP service like L3VPN service.

We consider an event in L3VPN service leading to RR1 being temporarly overloaded (for example, RR1 is processing massive updates due to a router failure or formatting updates for a route-refresh). In the same timeframe, CE1 wants to join the multicast flow from CE4. PE1 propagates the C-multicast route to RR1, but RR1 fails to propagate the route to RR5 because it is busy processing L3VPN. When RR1 finishes the L3VPN job, it would send the C-multicast route to RR5 and updates would be imported by PE4. The long time to join the flow may cause CE4 to miss part of the multicast flow.

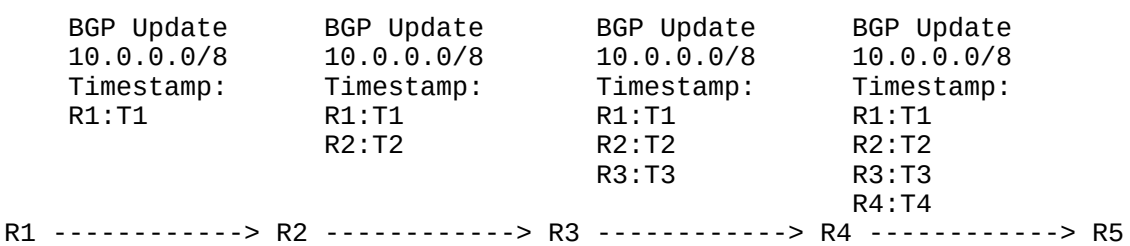
All BGP implementations are different in term of internal processing within an address family or between address family. The issue described above is just given as an example, and the document does not presume that all implementations are suffering from this exact issue. But whatever the implementation, their always be cases where BGP update processing could be delayed.

Service providers currently lack of performant solution to keep track of BGP update propagation time as well as solution to identify the BGP speakers causing issues.

BMP (BGP Monitoring Protocol) may be a solution but as several drawbacks (see Section 6).

2. Proposal

Our proposal is based on the path vector property of BGP. Each hop within the path would add a tuple (ID,timestamp) information in the BGP path. An ordered list of timestamps would so be built along the path.



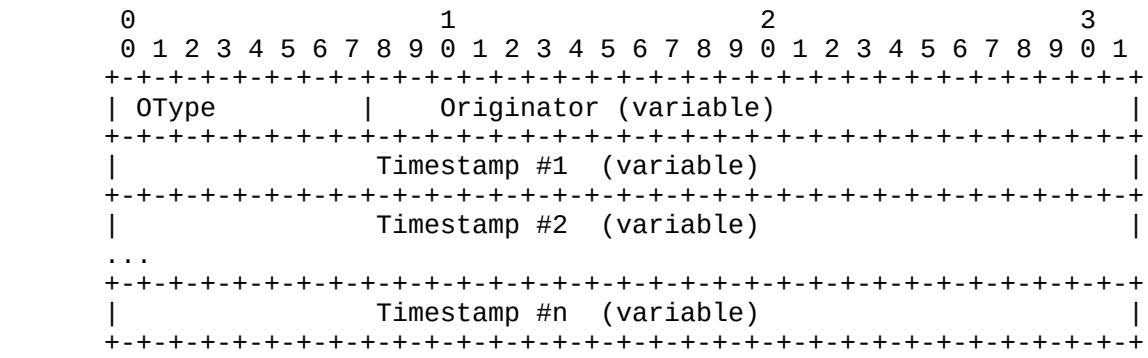
Using this mechanism, we can easily identify if a hop within a path is slowing down the propagation.

We propose to use a new BGP attribute, BGP timestamp attribute to encode timestamps information.

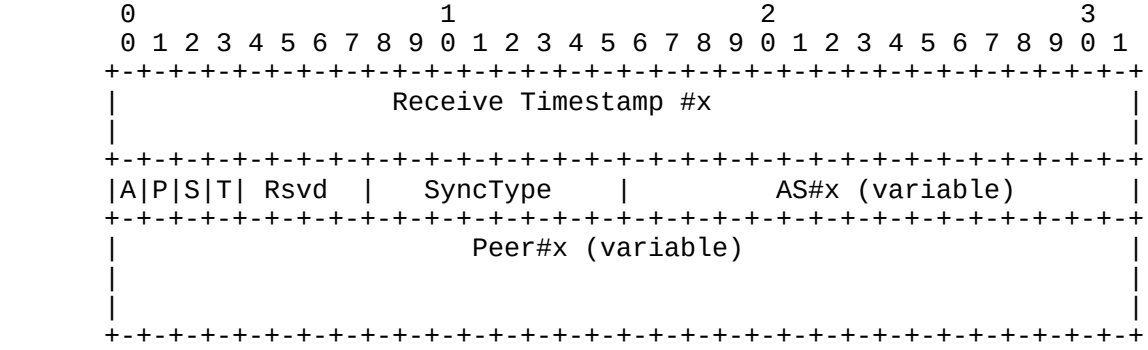
3. BGP timestamp attribute

The BGP timestamp (BGP-TS) Attribute is an optional transitive BGP Path Attribute. The attribute type code is TBD.

The value field of the BGP timestamp attribute is defined here :



- OType : A single octet encoding the originator type
  - Type 1 : 2 bytes ASN.
  - Type 2 : 4 bytes ASN.
  - Type 3 : IPv4 address.
  - Type 4 : IPv6 address.
- Originator : IP address or AS number identifying the first router or AS that added the BGP timestamp attribute.
- Timestamp : ordered list of timestamps, the first timestamp is the first router information. The timestamps are encoded as follows :



  - Receive timestamp : the time at which the BGP path was received. When originating a path in BGP, the timestamp is the originating time. The format of the timestamp is the same as in [RFC5905] and is as follows: the first 32 bits represent the unsigned integer number of seconds elapsed since 0h on 1

January 1900; the next 32 bits represent the fractional part of a second that has elapsed since then.

* Flags :
  + A : AS type, if unset the AS field is a 2 bytes ASN, otherwise the AS field is a 4 bytes ASN.
  + P : Peer type, if unset the peer field is an IPv4 address, otherwise the peer field is an IPv6 address.
  + S : Summary, if set, the timestamp is a summary entry and does not contain a peer field. If set, the P bit MUST be set to zero.
  + T : Synchronized, if set, the BGP speaker clock is synchronized to an external system.
* SyncType : defines the stratum as defined in [RFC5905].
* AS : the local AS of the BGP Speaker.
* Peer : the routerID of the BGP Speaker.

4. Processing the BGP timestamp attribute

4.1. Inspection list

A BGP Speaker supporting the BGP-TS can decide to timestamp only some specific BGP paths. An inspection list may be configured by the user (filter) to apply timestamping on a specific set of BGP prefixes or paths. By default, we suggest that a BGP Speaker supporting BGP-TS SHOULD NOT timestamp any BGP paths.

4.2. Originating a timestamped route in BGP

When a BGP Speaker supporting BGP-TS originates a new path in BGP that matches the inspection list, it MUST add the BGP-TS attribute to the BGP path and MUST set the receive timestamp field to the time the path was originated in BGP. If the BGP Speaker is synchronized to an external system when originating the route, the S-bit MUST be set in the attribute and the SyncType MUST be set to the current stratum.

4.3. Receiving a timestamped route in BGP

When a BGP Speaker supporting BGP-TS receives a BGP path that matches the inspection list and does not contains a BGP-TS attribute, it MUST add a BGP-TS attribute containing :

- The originator type and originator field are set according to local BGP Speaker informations.
- The timestamp entry contains information related to the local BGP Speaker.
- If the BGP Speaker is synchronized to an external system when receiving the route, the S-bit MUST be set in the attribute and the SyncType MUST be set to the current stratum.

When a BGP Speaker supporting BGP-TS receives a BGP path that matches the inspection list and contains a BGP-TS attribute, it MUST append its own timestamp entry in the existing attribute. If the BGP Speaker is synchronized to an external system when receiving the route, the S-bit MUST be set in the attribute and the SyncType MUST be set to the current stratum.

When a BGP Speaker supporting BGP-TS receives a BGP path that does not the inspection list and contains a BGP-TS attribute, it MUST NOT change the existing attribute.

When a BGP Speaker not supporting BGP-TS receives a BGP path that contains a BGP-TS attribute, it MUST follow the standard BGP procedures described in [RFC4271].

4.4. Sending a timestamped route in BGP

For a manageability/security purpose, the authors suggest that BGP timestamp attribute MAY NOT be sent to a peer unless it was explicitly configured for. This would prevent timestamp and internal address informations to be propagated to some external peers for example. See Section 4.5 for more information.

If a BGP path containing a BGP-TS attribute must be sent to be peer not configured with BGP timestamp option, the BGP-TS attribute should be dropped when the update message is sent to the peer.

4.5. Inter-AS considerations

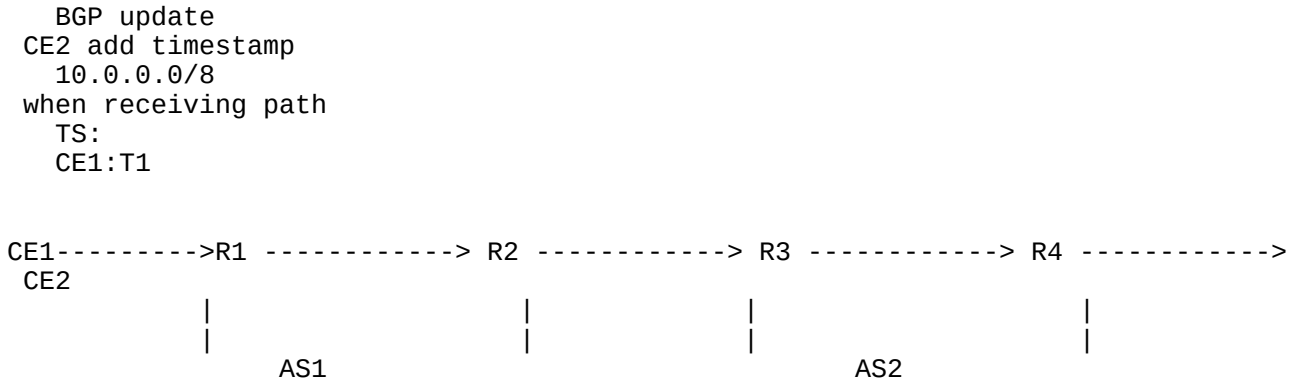


Figure 2

In the figure above, we consider that customer wants to monitor BGP updates propagation time between its two sites.

If AS1 and AS2 BGP Speakers does not support BGP-TS, the attribute will be transported transparently accross AS1 without any processing. CE2 will so receive the BGP path with only a single timestamp entry from CE1.

If AS1 and AS2 BGP Speakers does support BGP-TS, three different options are offered : drop, summarize, propagate.

4.5.1. Drop option

If AS1 and/or AS2 BGP Speakers support BGP-TS, they may not want to expose their timestamps or internal BGP topology to other ASes. If a service does not want to propagate timestamp information to external peers, it can decide to not activate the "timestamp" option on the peer configuration , as explained in Section 4.4.

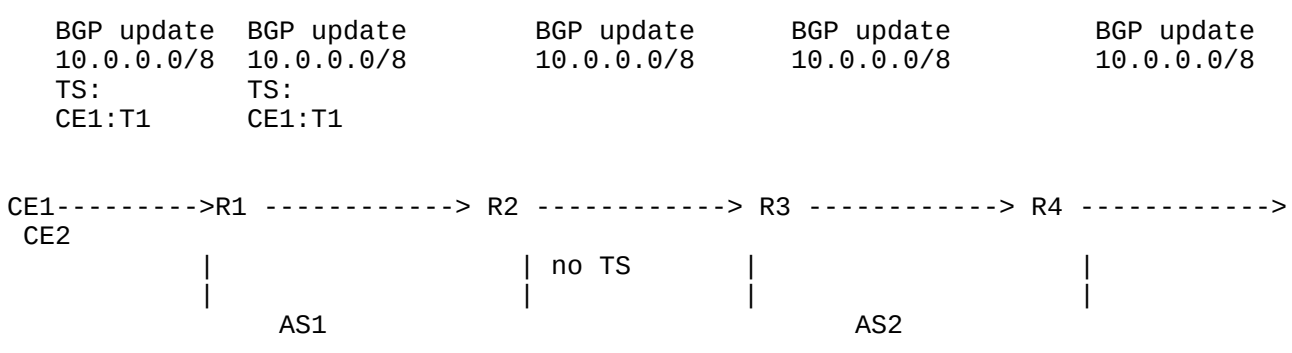


Figure 3

4.5.2. Summary option

If AS1 and/or AS2 BGP Speakers support BGP-TS, they may want to offer timestamp service to their customers but they want to hide their internal topology. In order to achieve the expected behavior, AS1/ AS2 can activate a timestamp summary option on the external peer.

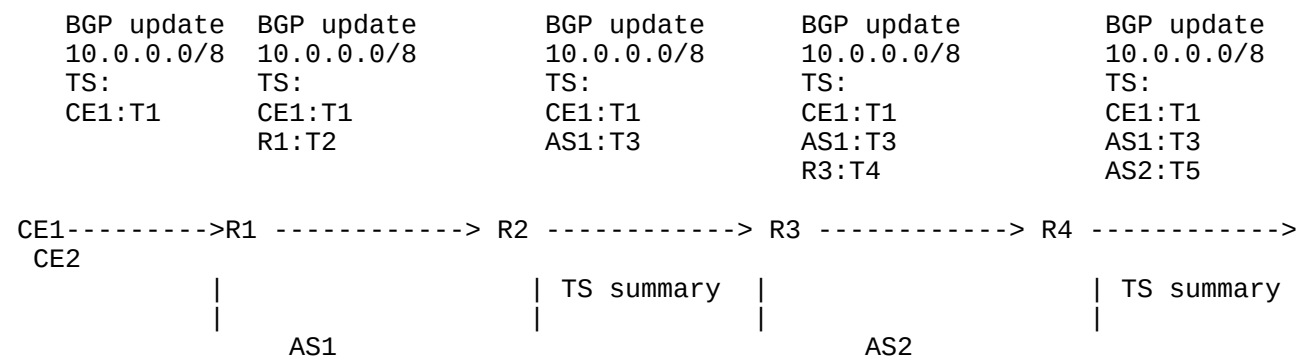


Figure 4

When using summary option, the BGP-TS attribute is modified as follows when exporting the route :

- All timestamp entries containing the local AS in AS field are removed.
- A new timestamp entry is created and inserted in place of removed entries (n entries replaced by 1).
- The new timestamp entry MUST have the S bit set.
- The new timestamp entry MUST NOT have a peer field.
- The timestamp of the new timestamp entry is the receiving timestamp of the last timestamp entry that has been removed.

4.5.3. Propagate option

If AS1 and/or AS2 BGP Speakers support BGP-TS, they may want to offer timestamp service to their customers with a full view. The behavior is the default intraAS behavior.

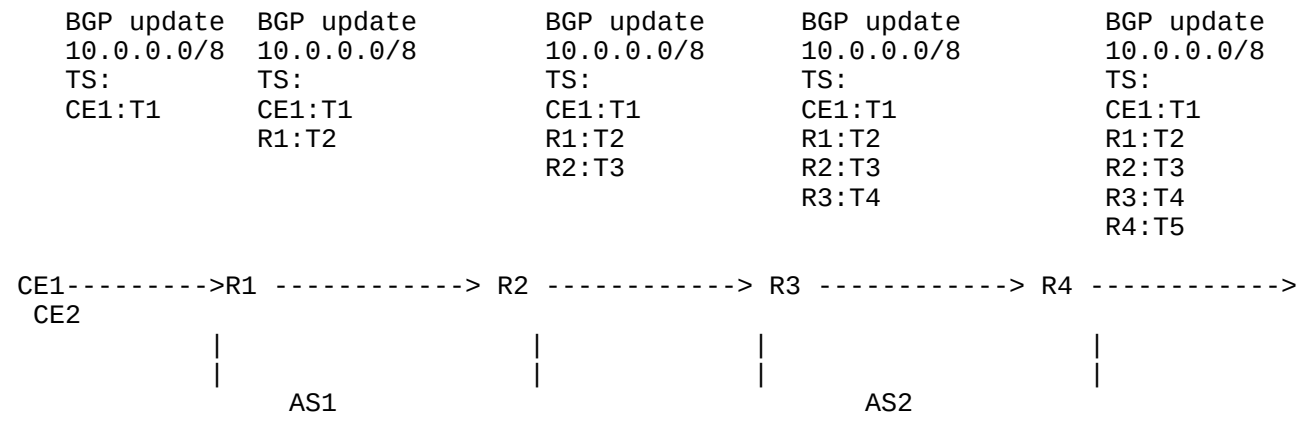


Figure 5

4.6. Handling malformed attribute

When receiving a BGP Update message containing a malformed BGP-TS attribute, an "attribute-discard" action MUST be applied as defined in .

5. Monitoring BGP Update propagation time

5.1. An architecture to measure BGP Update propagation time

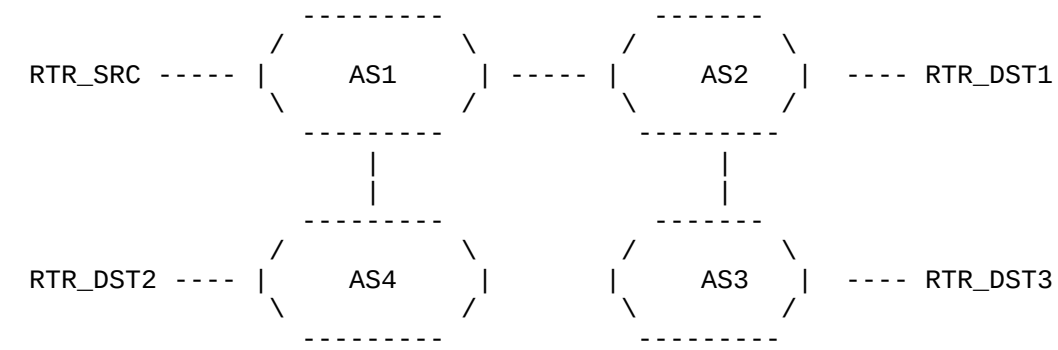


Figure 6

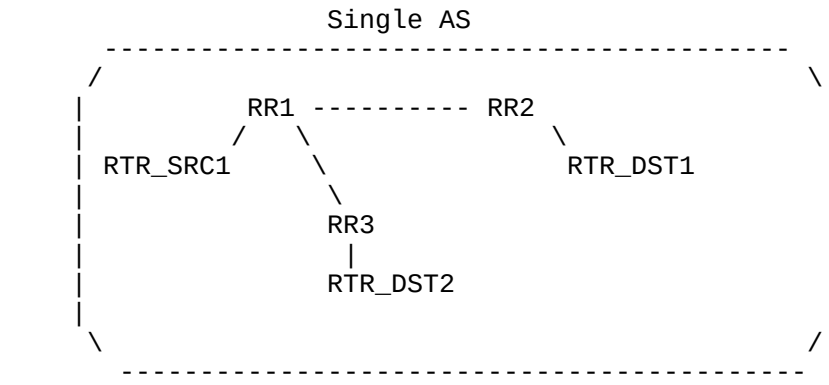


Figure 7

Figure 6 and Figure 7 describes an interAS and a single AS scenario where a service provider wants to monitor BGP Update propagation time from a router to multiple routers. In Figure 6, multiple probing routers are attached to multiple ASes. In Figure 7, all probing routers are in the same AS.

An external tool should command RTR_SRC to originate a probing BGP path. Each probing router is configured to match the path in its inspection list. The BGP path would propagate across ASes whatever they are supporting BGP TS or not. Each probing router would receive the BGP path and add timestamp information. Authors suggest to implementors to use a local wrapping buffer on each node and record entries in the buffer each time a BGP path is timestamped. An external tool should then retrieve timestamps information from RTR_DSTx. How the information is retrieved is out of scope of the document but we can imagine using :

- BMP from the external tool to each RTR_DSTx.
- NetConf call to retrieve wrapping buffer information.
- SNMP call to retrieve wrapping buffer information.
- CLI command to retrieve wrapping buffer information.

5.2. Measurement accuracy

For the solution to be accurate, it is mandatory for BGP Speaker to be synchronized. This could be achieved easily within a single AS but in a inter domain scenario, it is hard to ensure that all Speakers are synchronized to a good clock source.

The S bit and SyncType fields are set to help operators to understand the accuracy of the timestamp measurements and being able to compare timestamps between them.

5.3. Dealing with stale information

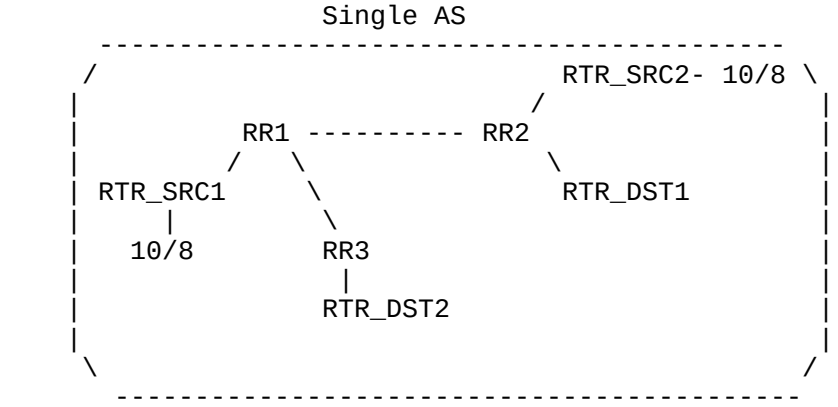


Figure 8

In the figure above, consider that the service provider is keep tracking of propagation time for real NLRIs (corresponding to customer routes). All the BGP Speakers in our figure are configured to inspect the NLRI 10/8 which is multihomed. We consider that the network is starting and the NLRI has not been propagated yet.

RTR_SRC1 starts to propagate 10/8 within the BGP controlplane. All BGP Speakers considers the path as best and this path will be propagated within the whole controlplane. Each BGP Speaker would add its timestamp information and RTR_DST1 and RTR_DST2 would be able to record the timestamp vector. In this case, the timestamp vector is quite accurate because it represents an end to end propagation.

Now RTR_SRC2 starts to propagate its own path. RR2 has two paths for 10/8 and will choose the best one, let’s consider that RTR_SRC2 path is the best one, RTR_SRC2 path will so be propagated and timestamp vector will be updated. RR1 will also have two paths, and we consider that RR1 prefers RTR_SRC1 path, so RTR_SRC2 path will not be propagated by RR1. In this situation, RTR_DST1 will receive the path from RR2 with accurate timestamp (end to end propagation) but RTR_DST2 will never receive it.

We could also consider a stable network situation, where both paths have been advertised for a long time. A network event may occur (e.g. IGP metric change) that would cause a BGP Speaker within a path vector to change its best path. In Figure 8, an IGP event, may

cause RR1 to change its decision and prefers the path originated by RTR_SRC2 as best, the path will be propagated with previous received timestamp information that are no more accurate. RTR_DST2 will receive a BGP timestamp vector containing stale timestamp informations as well as new ones.

The case of sending stale timestamp information can also appear with a single originator as soon as some redundancy in the BGP design is involved (multiple RRs, multiple ASBRs ...).

An external tool that monitors BGP timestamp should take care about analysing only end to end propagation scenarios.

6. Compared to BMP

BMP (BGP Monitoring Protocol) [I-D.ietf-grow-bmp] is a solution to monitor BGP sessions and provides a convenient interface for obtaining route views. BMP is a complete suite of messages to exchange informations regarding a BGP session.

We can imagine to use BMP as a solution to monitor BGP update propagation time but there is multiple drawbacks associated with such solution :

- BMP provides dump of all received BGP update (per peer). If we are interested only in probing BGP routes, a strong filtering of information may be needed in BMP messages.

- BMP does not mandate timestamping of messages (as per [I-D.ietf-grow-bmp] Section 5) : "If the implementation is able to provide information about when routes were received, it MAY provide such information in the BMP timestamp field. Otherwise, the BMP timestamp field MUST be set to zero, indicating that time is not available."

- BMP may provide (if implementation available) timestamps information only for a single router point of view. If we want to retrieve timestamps of all BGP Speakers on a path, a BMP session is required to all BGP speakers. Correlation (based on known design) is also required at the external tool to order timestamps from each BMP session.

- If BMP provides timestamp information, it does not provide information on how the router clock is synchronized (free run, NTP, GPS ...).

Using BMP to monitor BGP update propagation may complexify the design of the monitor solution.

7. Deployment considerations

This solution is not intended to perform timestamp imposition on all BGP updates.

Service provider implementing the BGP timestamp attribute must be aware of the propagation rules of the NLRIs to be inspected. If we consider an implementation scenario, where a path for NLRI is already propagated, a new path may appear and starts to be propagated, propagation of this new path may stop at a certain point because a BGP Speaker may consider the old path as the best one. Another scenario, could be that the two paths are installed, and for a BGP Speaker within the path vector, the best path is changing because of an IGP metric change, this BGP Speaker will send a new BGP update and timestamp information of the path will be updated but will have no more sense : origin timestamp will be quite old, but timestamps recorded after this BGP Speaker will be recent. This kind of scenario is complex to understand.

The deployment scenario we are targeting is really to inspect some specific NLRIs identified by the service provider where the propagation rules are well known (see Section 5 as an example). Service provider may rely on existing NLRIs (real routes), or ephemeral NLRIs (dedicated NLRIs for beaconing). Whatever the NLRI used, the tool used by the service provider to collect and interpret the timestamp must be aware of the propagation rules and must record events only if propagation is end to end (from originator to listener).

The inspection list should be kept as small as possible in order to not introduce processing overhead and as a consequence slow down propagation. Implementors should take care about reducing as much as possible the processing overhead introduced by the inspection list and timestamp imposition.

8. Security considerations

Depending of the implementation and router capacity, adding timestamps to BGP path may consume some router ressources. As proposed in Section 4.1, by default a BGP Speaker will not timestamp any path and inspection list should be configured to activate timestamping on a subset of paths. Using this approach, we consider that overhead that may be introduced by timestamping BGP paths is well controlled by operators. An external router cannot force an internal router to timestamp.

Providing detailled timestamps information to other ASes may introduce security issues by exposing internal datas (part of BGP

topology, IP addresses, internal performance) to external entities. The proposal we make in Section 4.5 solves this security issue by giving flexibility to operators on the level of information he wants to expose to external peers.

9. Acknowledgements

10. IANA Considerations

IANA shall assign a codepoint for the BGP Timestamp attribute. This codepoint will come from the "BGP Path Attributes" registry.

11. Normative References

[I-D.ietf-grow-bmp] Scudder, J., Fernando, R., and S. Stuart, "BGP Monitoring Protocol", draft-ietf-grow-bmp-07 (work in progress), October 2012.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC5905] Mills, D., Martin, J., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, June 2010.

Authors' Addresses

Stephane Litkowski
Orange Business Service

Email: stephane.litkowski@orange.com

Keyur Patel
Cisco Systems

Email: keyupate@cisco.com

Jeff Haas
Juniper Networks

Email: jhaas@juniper.net