IDR working group
Internet Draft
Intended status: Standards Track
Expires: January 2012

Juan Alcaide
Cisco
Fernando Calabria
Cisco

September 2 , 2011

BGP prefix priority

draft-alcaide-calabria-idr-bgp-prefix-priority-00

Status of this Memo

Copyright Notice

Abstract

This document defines a set of extended communities to carry priority information. This information provides a mechanism for assigning a processing preference to the routes that carries it. It also provides a scheme for processing routes with strict priority order during update reception, best-path computation, and update transmission.

Table of Contents

1. Introduction

BGP scale has been growing in the last years, in terms of neighbors and routes. This impacts convergence times after, for example, a BGP re-initialization event. One solution is a continuous upgrade of the hardware used by BGP speakers, by adding faster CPU and additional memory. This approach, however, is expensive and cannot reduce convergence times indefinitely. It is desirable having a software based solution, in which a BGP speaker can prioritize some selected routes. In other words, there is a need for a Qos-like mechanism in the BGP control plane.

Processing of routes with a given priority SHOULD be performed before any lower priority ones. This process SHOULD be performed in a preemptive manner. Thus, the convergence times obtained for high priority routes would be the same as if there were no lower priority routes at all. Implementations are not expected to reach this theoretical limit, but closely approach to it.

Priority information is signaled by adding to the route an extended community hereby named PEC (Priority Extended Community). A PEC is meant to have network wide significance and transparent to speakers that do not understand it. It MAY be set at the origination of the route and propagated across the network, thus greatly reducing management burden, but it can also be set by a policy if required.

Route processing during reception of routes is based on the priority assigned to the received path; while the remaining tasks are based on the priority of the computed best-path. Provisions to prevent that a change in the priorities associated to the path results in miss ordered routes are also covered in the present document.

The design of how a given priority marking is honored is twofold: a given speaker SHOULD process the reception of a path with the priority that the received path has; and it should process any local or transmission task with the priority associated to the best-path of the net. Thus, the design supports different paths being originated with different priority marking; and it deals with the conflict by aggregating these markings during best-path computation and propagating them downstream. Thus, aggregated marking is honored as close to the source of this aggregation as possible.

2. Conventions Used in this Document

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 BCP 14, RFC 2119 [RFC2119]. RFC 2119 defines the use of these key words to help make the intent of standards track documents as clear as possible. While this document uses these keywords, this document is not a standards track document.

3. Definitions of Commonly Used Terms

The set of definitions below are used through this document. Some terms are well-known, some terms are defined to avoid confusion and some (those marked with a “*”) are defined for the purpose of this implementation (and thus referenced by other sections throughout the entire document).

BGP process: internal implementation of a BGP speaker. The router may implement the BGP process as one or more OS processes or threads.

net: BGP prefix, including all the paths received from all the neighbors.

path: BGP prefix received from a particular neighbor. Multiple paths can be associated to a given net.

BGP table: database where all the BGP routes are kept. It’s a set of nets, each of them with their associated paths.

RIB (Routing Information Base): database where all the forwarding information is kept. It’s a set of nets with their associated forwarding paths (more than one if it’s a multipath net). Nets can be learned from different routing protocols, in particular they can have a correspondent entry from the BGP table, and the forwarding path used will be the BGP best-path for that net (plus additional ones if it’s a multipath net).

upstream/downstream directions: When routes flow in a given direction, a BGP speaker receives routes from upstream and advertises them downstream.

receiving peer/sending peer: When routes flow in a given direction between two speakers, the BGP speaker that sends the routes is the sending peer and the BGP speaker that receives them is the receiving peer.

PEC* (Priority Extended Community): extended-community associated to a BGP path that is an indication of the *path-priority* for that path. PEC=*priority* denotes that a given PEC indicates that priority. PEC=NULL indicates that no PEC is actually send in an update message.

strict priority: method of servicing the process of several tasks. Tasks with a given priority are processed before any other task with lower priority. In the context of this document, they SHOULD also preempt the processing of any lower priority task.

route priority*: integer from 0 to 7 associated to a route. It indicates the priority or urgency with which this route is processed. Priority=0 indicates the lowest urgency, and priority=7 indicates the highest urgency. It is a generic term that can actually have a different value based on the specific task a BGP process is performing:

in-message-priority*: priority associated to a received BGP message as it is received from the TCP session. It’s derived by calculating the maximum of all path-priorities in a given update message. It determines the priority for message processing during reception.

path-priority*: priority associated to a BGP path. It’s calculated by looking at the PEC associated to the path. It determines the priority of a path during reception, after it has been parsed from a message. It is also used to calculate the rest of priorities.

max-path-priority*: priority associated to a BGP net. It’s derived by calculating the maximum path-priority for all the paths of a given net. It determines the processing priority for best-path computation.

net-priority*: priority associated to a BGP net. It's derived from by calculating the path-priority of the best-path. It determines the processing priority for any further local processing (after best-path computation) and advertisement of routes.

4. Scope

As mentioned before, this document focuses on the following:

- A scheme that assigns and signals priority values on a prefix basis.

- Proposing a solution for processing prioritized routes during update reception.

- Proposing a solution for processing prioritized routes during best-path computation, and update transmission.

- Proposing a solution for managing prefixes whose priority changed by an administrative task.

- Guidelines to "interact" with speakers that do not (fully or partially) support prefix prioritization.

5. Solution Specification

5.1. Network Wide Prefix Priority

Priority for a prefix is set by the assignment of a BGP extended community attribute, in order to indicate preference of processing. This community is hereby named PEC (Priority Extended Community), and MUST contain priority values from 0 to 7. PECs are defined as a new transitive extended-community of experimental use as defined by [RFC4360] and [RFC3692].

The extended community type is: 0x80FE whose value is encoded as a sequence of 5 zero bytes and the priority value set by the 3 most significant bits of the last byte, resulting in:

Highest priority (7) : 0x80FE:0000000000E0

Lowest priority (0) : 0x80FEA:000000000000

and all the pertinent values in between.

In a trusted environment, PEC is set by the speaker originating the route and has neighbor significance. This approach greatly reduces the management burden of mapping routes to priorities. If PECs are not trusted, they MAY be changed by any other speaker downstream based on its policy.

PECs are propagated on a per path basis. The correlation between paths and nets for a given priority is as follows:

- Path-priority is associated to a BGP path upon receiving it, typically based on PECs.

- Net-priority is assigned to the net, and corresponds to the path-priority of the best-path for that prefix.

- Net-priority is signaled when the route is advertised, typically by PECs.

5.2. Network Wide Prefix Priority in a "Trusted" Environment

In a trusted environment, priority signaling is based on the advertisement of one single PEC by the originator of the route. In particular:

- Path-priority for a path is based on the PEC received to that BGP path.

- If multiple PECs are received for the same prefix, the speaker SHOULD use the PEC that indicates a higher priority.

- If no PEC is received (PEC=NULL), the speaker SHOULD explicitly set path-priority=0.

- When advertising updates, all PECs are removed and one single PEC is advertised, corresponding to the net-priority of the advertised net. In particular, if net-priority=0 an explicit PEC=0 SHOULD be sent.

5.3. Network Wide Prefix Priority in a "on-Trusted" Environment

In a non-trusted environment, it's possible to change the above procedures by local configuration. In particular:

- Path-priority can be overwritten when receiving a route.

- PECs transmitted can be overwritten when advertising a route.

5.4. Prioritizing Reception of Routes

Processing routes during reception involves tasks like reading update messages, parsing the prefixes inside those messages, and installing them in the BGP table as a path belonging to the neighbor associated to the session the message was received from.

These tasks SHOULD be performed in strict priority order based on the path-priority set by a speaker or by local configuration.

Using path-priority to select the priority for inbound processing carries within some challenges, since path-priority is unknown till inbound processing itself is performed. The following solutions to this challenge are presented:

- After reading an update message from the TCP session, inspect the message and calculate an in-message-priority, which corresponds to the highest path-priority of all the prefixes present in the message. Any further processing of the message, like a detailed parsing, it’s performed in strict priority order based on in-message-priority.

- Calculating in-message-priority itself is not a task that can be prioritized, and therefore it should be a light-weight task. For the most common case, where path-priority is determined based on PEC, this consideration does not apply. Assigning statically a path-priority to a given session is a task that requires no processing at all. On the other side of the spectrum, if path-priority is determined by the prefix itself (i.e. prefixes in the same update can have different path-priority), the task becomes non-trivial. Furthermore, some prefixes may get a preferential treatment (if their in-message-priority is higher than their path-priority).

- After path-priority is computed for a route, any further inbound processing of the route can be performed based on path-priority. This may involve tasks like installing the route into the BGP table.

A path MUST be discarded (and not installed in the BGP table) if it has been received before a path for the same prefix and TCP session that already exists in the BGP table. This non-FIFO scenario is possible when receiving the same prefix with different priorities. If the second prefix received has a higher in-message-priority or path-priority, the first prefix could be a candidate to be installed in the BGP table after the second has actually already been installed. Note that with these modifications, the sequence of routes installed in the BGP table could be different than it would be without the use of priorities. This change of behavior is acceptable under BGP protocol rules ([RFC4271]).

Any received BGP messages that are not update messages SHOULD be processed in strict priority order, based on a higher priority than the maximum in-message-priority.

5.5. Prioritizing Local and Outbound Processing of Routes

After a path has been installed in the BGP table, the processing priority of all the tasks that correspond to the associated prefix is not dependent anymore into the priority of the path itself (path-priority), but on that of the net it belongs to, namely net-priority. However, net-priority cannot be known till the best-path is resolved, and to prioritize itself the task that resolves best-path, max-path-priority is used. Max-path-priority is defined as the maximum path-priority of all the paths associated to a given net, including the path-priority of any new path that triggered the best-path computation.

Calculating max-path-priority itself is a task that SHOULD be processed in strict priority order, based on the path-priority of the path that triggers best-path computation.

Best-path processing is a local task that SHOULD be processed in strict priority order, based on max-path-priority.

Further local processing of routes includes tasks like installation of the net in the routing table. Outbound processing includes tasks like formatting nets into update messages and transmitting them through the TCP session. All these tasks SHOULD be performed in strict priority order based on net-priority.

Note that the rules above force that all the prefixes in a given message to have associated the same net-priority (if the transmission of update messages is to be prioritized based on the common net-priority). This is already a constriction if PECs are used to signal priorities to downstream peers.

Any transmitted BGP messages that are not update messages SHOULD be processed in strict priority order, based on a higher priority than the maximum net-priority.

5.6. Change of priority

As previously described, the advertisement of routes is done with a priority based on net-priority (assigned to a given prefix). There are no conflicts as long as, over time, net-priority remains the same for a given prefix. However, net-priority derives from path-priority, and therefore it may change. Without any further mechanisms, the order in which routes are advertised would be incorrect, and inconsistencies across the BGP tables of the sending and receiving peers would appear.

This non-FIFO scenario is possible when advertising the same prefix with different priorities. If the second prefix that needs to be advertised to a given neighbor has a higher net-priority than a first one already scheduled for transmission, the second one could be transmitted actually before the first one is.

When sent through the BGP session, advertisements for a given prefix MUST keep, in all cases, the same order than they would have without route prioritization (i.e., FIFO-like processing), or perform only the last advertisement. In other words, a route computed as best-path MUST NOT be transmitted over a BGP session before a route that was computed previously as best-path. Note that the offending scenarios are only possible when increasing net-priority. If net-priority decreases, the problem does not happen. How an implementation deals with this situation is outside the scope of this document. However, these two general approaches are discussed:

- One obvious option is making sure that any previous low-priority route is not actually advertised (and thus it's discarded). This option has the drawback of complexity (updates already scheduled for transmission may have to be reformatted). Note also that the sequence of routes transmitted could be different than it would be without the use of priorities. This change of behavior is acceptable under BGP protocol rules ([RFC4271]).

- A second option is that, whenever net-priority needs to increase, the BGP speaker simply waits for all the routes with lower net-priority to be transmitted across all sessions. After they are transmitted, net-priority can be safely increased. While net-priority has not transitioned, any task depending on net-priority for that route is processed as usual, considering the old net-priority. Note that this may imply sending two updates upon a transition, if attributes transmitted (like PEC) depend on net-priority. The drawback of this approach is that it introduces a delay in how priority information is propagated across the network (indefinitely in a worst case scenario, if a prefix is constantly flapping at a high rate).

Same considerations apply for any other local processing tasks, if the implementation of these tasks makes them susceptible of miss ordering their execution.

5.7. Interaction with Neighbors not Supporting Route Prioritization

When all the BGP speakers involved in the propagation of a network event do not support route prioritization, priority routes will not be treated with the preference they would have otherwise. It is possible, however, to minimize the effects of this scenario based on the following considerations:

- Priority management is transparent across speakers and domains not supporting route prioritization. This is because PEC is defined as a transitive extended-community.

- If priority of received paths is not marked with a PEC, the same effect can be achieved by local configuration.

- Reception of routes from a neighbor not supporting route priority does not change. The routes are received with the preference that in-message-priority indicates.

- Advertisement of nets towards a neighbor not supporting route priority does not change. The routes are advertised with the preference that net-priority indicates.

Note that if routes are advertised with the order determined by its own net-priority to a downstream speaker not supporting route prioritization, there is a high probability that that this speaker will process those routes with the same (or approximate) order that it received them, since most likely it will treat them in a FIFO or quasi-FIFO fashion. Thus, introducing a single speaker supporting route prioritization upstream in the network can significantly increase the overall prioritization across the entire route propagation path.

6. Rationale behind network wide priorities

This proposal develops a comprehensive use of a network wide priority as a method to give preferential treatment to some routes. Out of all the possible design alternatives, the choices were based in flexibility, performance and stability. Amongst them, the following ones can be pointed out:

- PECs can be used to signal path-priorities for unreachable NLRIs (aka withdraws). In an implementation without priorities, any attributes are meaningless when associated to unreachable NLRIs, but there is nothing in the BGP protocol rules ([RFC4271]) to prevent its use. Note that implementations could use other attributes (besides PECs) associated to unreachable NLRIs.

- An implementation SHOULD send one and one only PEC, but it SHOULD also accept multiple PECs or no PECs at all. With only "good behaved" implementations and configurations, this precaution is not necessary; but the proposal’s designs provisions for it under the philosophy “be liberal with what you receive, be conservative with what you send”.

- When a net with a net-priority=0 is sent, the options are to set PEC explicitly (PEC=0) or implicitly (PEC=NULL). Both options are equally valid and there is not a chance for confusion. Consider, however, the case where the nets coming from two speakers, one supporting route priority and one not supporting it. They traverse a transparent speaker (i.e. one that just forwards nets with the

PECs it received). In this case, confusion is possible: a router downstream using route prioritization won't be able to distinguish the two set of routes (and it's possible that its requirements dictate to differentiate both cases). The drawbacks of using an explicit PEC=0 is that some extra bytes need to be added to the update messages of the lowest net-priority routes, and that more update messages might be transmitted (consider the case above, where a transparent speaker sends routes with both PEC=0 and PEC=NULL: these routes cannot be packed in the same message).

- It's desirable for a given prefix to have the same priority across the network. Propagating the priority of the best-path maximizes the chances of this happening. There is no absolute guarantee, however, since not all the speakers have to select the same best-path, according to BGP propagation and best-path selection rules ([RFC4271]).

- When path-priorities are different for a given net, a different approach could have been chosen to determine net-priority (other than using the path-priority of a best-path). An alternative method, however, could potentially create a chicken-and-egg situation. Consider, for instance, a proposal that chooses as net-priority the higher path-priority of all the paths. Consider also the case of two speakers back to back, mutually advertising routes for a given prefix between them, none of them using the other's route a best-path. The mutually advertised routes could have a higher priority than the best-paths. This would be a self-sustained state that would remain no matter what other PECs are received from other peers.

7. Security Considerations

This document introduces no new security concerns to BGP or other specifications referenced in this document.

8. IANA Considerations

N/A

9. References

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

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

[RFC4360] Sangli, et all "BGP Extended Communities Attribute, RFC 4360, February 2006

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

[RFC2234] Crocker, D. and Overell, P.(Editors), "Augmented BNF for Syntax Specifications: ABNF", RFC 2234, Internet Mail Consortium and Demon Internet Ltd., November 1997.

10. Authors’ Addresses

Juan Alcaide
Cisco
7025 Kit Creek Rd RTP-NC 27709
jalcaide@cisco.com
USA

Fernando Calabria
Cisco
7025 Kit Creek Rd RTP-NC 27709

fcalabri@cisco.com
USA