Network Working Group G. Cristallo
Internet Draft Alcatel
Document: draft-jacquenet-bgp-qos-00.txt C. Jacquenet
Category: Experimental France Telecom
Expires August 2004 February 2004

The BGP QOS_NLRI Attribute
<draft-jacquenet-bgp-qos-00.txt>

Status of this Memo

NOTE: a PDF version of this document (which includes the figures mentioned in section 7) can be accessed at http://www.mescal.org.

Abstract

This draft specifies an additional BGP4 (Border Gateway Protocol, version 4) attribute, named the "QOS_NLRI" attribute, which aims at propagating QoS (Quality of Service)-related information associated to the NLRI (Network Layer Reachability Information) information conveyed in a BGP UPDATE message.

Table of Contents

1. 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 RFC 2119 [2].

2. Introduction

Providing end-to-end quality of service is one of the most important challenges of the Internet, not only because of the massive development of value-added IP service offerings, but also because of the various QoS policies that are currently enforced within an autonomous system, and which may well differ from one AS (Autonomous System) to another.

For the last decade, value-added IP service offerings have been deployed over the Internet, thus yielding a dramatic development of the specification effort, as far as quality of service in IP networks is concerned. Nevertheless, providing end-to-end quality of service across administrative domains still remains an issue, mainly because:

- QoS policies may dramatically differ from one service provider to another,

- The enforcement of a specific QoS policy may also differ from one domain to another, although the definition of a set of common quality of service indicators may be shared between the service providers.

Activate the BGP4 protocol ([3]) for exchanging reachability information between autonomous systems has been a must for many years. Therefore, disseminating QoS information coupled with reachability information in a given BGP UPDATE message appears to be helpful in enforcing an end-to-end QoS policy.

This draft aims at specifying a new BGP4 attribute, the QOS_NLRI attribute, which will convey QoS-related information associated to the routes described in the corresponding NLRI field of the attribute.

This document is organized according to the following sections:

- Section 3 describes the basic requirements that motivate the approach,

- Section 4 describes the attribute,

- Section 5 elaborates on the mode of operation,

- Section 6 elaborates on the use of the capabilities advertisement feature of the BGP4 protocol,

- Section 7 depicts the results of a simulation work,

- Finally, sections 8 and 9 introduce IANA and some security considerations, respectively.

3. Basic Requirements

The choice of using the BGP4 protocol for exchanging QoS information between domains is not only motivated by the fact BGP is currently the only inter-domain (routing) protocol activated in the Internet, but also because the manipulation of attributes is a powerful means for service providers to disseminate QoS information with the desired level of precision.

The approach presented in this draft has identified the following requirements:

- Keep the approach scalable. The scalability of the approach can be defined in many ways that include the convergence time taken by the BGP peers to reach a consistent view of the network connectivity, the number of route entries that will have to be maintained by a

BGP peer, the dynamics of the route announcement mechanism (e.g., how frequently and under which conditions should an UPDATE message containing QoS information be sent?), etc.

- Keep the BGP4 protocol operation unchanged. The introduction of a new attribute should not affect the way the protocol operates, but the information contained in this attribute may very well influence the BGP route selection process.

- Allow for a smooth migration. The use of a specific BGP attribute to convey QoS information should not constrain network operators to migrate the whole installed base at once, but rather help them in gradually deploying the QoS information processing capability.

4. The QOS_NLRI Attribute (Type Code tbd*)

(*): "tbd" is subject to the IANA considerations section of this draft.

The QOS_NLRI attribute is an optional transitive attribute that can be used:

1. To advertise a QoS route to a peer. A QoS route is a route that meets one or a set of QoS requirement(s) to reach a given (set of) destination prefixes. Such QoS requirements can be expressed in terms of minimum one-way delay ([4]) to reach a destination, the experienced delay variation for IP datagrams that are destined to a given destination prefix ([5]), the loss rate experienced along the path to reach a destination, and/or the identification of the traffic that is expected to use this specific route (identification means for such traffic include DSCP (DiffServ Code Point, [6]) marking). These QoS requirements can be used as an input for the BGP route calculation process,

2. To provide QoS-related information along with the NLRI information in a single BGP UPDATE message. It is assumed that this information will be related to the route (or set of routes) described in the NLRI field of the attribute.

From a service provider's perspective, the choice of defining the QOS_NLRI attribute as an optional transitive attribute is motivated by the fact that this kind of attribute allows for gradual deployment of the dissemination of QoS-related information by BGP4: not all the

BGP peers are supposed to be updated accordingly, while partial deployment of such QoS extensions can already provide an added value, e.g. in the case where a service provider manages multiple domains, and/or has deployed a BGP confederation ([7]).

This draft makes no specific assumption about the means to actually value this attribute, since this is mostly a matter of implementation, but the reader is suggested to have a look on document [8], as an example of a means to feed the BGP peer with the appropriate information. The QOS_NLRI attribute is encoded as follows:

```
+---------------------------------------------------------+
| QoS Information Code (1 octet)                          |
+---------------------------------------------------------+
| QoS Information Sub-code (1 octet)                      |
+---------------------------------------------------------+
| QoS Information Value (2 octets)                        |
+---------------------------------------------------------+
| QoS Information Origin (1 octet)                        |
+---------------------------------------------------------+
| Address Family Identifier (2 octets)                    |
+---------------------------------------------------------+
| Subsequent Address Family Identifier (1 octet)          |
+---------------------------------------------------------+
| Network Address of Next Hop (4 octets)                  |
+---------------------------------------------------------+
| Flags (1 octet)                                         |
```

```
+---------------------------------------------------------+
| Identifier (2 octets)                                   |
+---------------------------------------------------------+
| Length (1 octet)                                        |
+---------------------------------------------------------+
| Prefix (variable)                                       |
+---------------------------------------------------------+
```

The use and meaning of the fields of the QOS_NLRI attribute are defined as follows:

- QoS Information Code:

  This field carries the type of the QOS information. The following types have been identified so far:

(0) Reserved
(1) Packet rate, i.e. the number of IP datagrams that can be transmitted (or have been lost) per unit of time, this number being characterized by the elaboration provided in the QoS Information Sub-code (see below)
(2) One-way delay metric
(3) Inter-packet delay variation
(4) PHB Identifier

- QoS Information Sub-Code:

  This field carries the sub-type of the QoS information. The following sub-types have been identified so far:

(0) None (i.e. no sub-type, or sub-type unavailable, or unknown sub-type)
(1) Reserved rate
(2) Available rate
(3) Loss rate
(4) Minimum one-way delay
(5) Maximum one-way delay
(6) Average one-way delay

The instantiation of this sub-code field MUST be compatible with the value conveyed in the QoS Information code field, as stated in the following table (the rows represent the QoS Information Code possible values, the columns represent the QoS Information Sub-code values identified so far, while the "X" sign indicates incompatibility).

	0	1	2	3	4	5	6
0							
1					X	X	X

2		X	X	X			
3		X	X	X	X	X	X
4		X	X	X	X	X	X

- QoS Information Value:

  This field indicates the value of the QoS information. The corresponding units obviously depend on the instantiation of the QoS Information Code. Namely, if:

(a) QoS Information Code field is "0", no unit specified,
(b) QoS Information Code field is "1", unit is kilobits per second (kbps), and the rate encoding rule is composed of a 3-bit exponent (with an assumed base of 8) followed by a 13-bit mantissa, as depicted in the figure below:

```
0      8       16
|       |       |
-----------------
|Exp| Mantissa  |
-----------------
```

This encoding scheme advertises a numeric value that is (2^16 -1 - exponential encoding of the considered rate), as depicted in [9].
(c) QoS Information Code field is "2", unit is milliseconds,
(d) QoS Information Code field is "3", unit is milliseconds,
(e) QoS Information Code field is "4", no unit specified.

- QoS Information Origin:

  This field provides indication on the origin of the path information, as defined in section 4.3.of [3].

- Address Family Identifier (AFI):

  This field carries the identity of the Network Layer protocol associated with the Network Address that follows. Currently defined values for this field are specified in [10] (see the Address Family Numbers section of this reference document).

- Subsequent Address Family Identifier (SAFI):

  This field provides additional information about the type of the prefix carried in the QOS_NLRI attribute.

- Network Address of Next Hop:

  This field contains the IPv4 Network Address of the next router on the path to the destination prefix, (reasonably) assuming that such routers can at least be addressed according to the IPv4 formalism.

- Flags, Identifier, Length and Prefix fields:

  These four fields actually compose the NLRI field of the QOS_NLRI attribute, and their respective meanings are as defined in section 2.2.2 of [11].

5. Operation

When advertising a QOS_NLRI attribute to an external peer, a router may use one of its own interface addresses in the next hop component of the attribute, given the external peer to which one or several route(s) is (are) being advertised shares a common subnet with the next hop address. This is known as a "first party" next hop information.

A BGP speaker can advertise to an external peer an interface of any internal peer router in the next hop component, provided the external peer to which the route is being advertised shares a common subnet with the next hop address. This is known as a "third party" next hop information.

A BGP speaker that sends an UPDATE message with the QOS_NLRI attribute has the ability to advertise multiple QoS routes, since the Identifier field of the attribute is part of the NLRI description. In particular, the same prefix can be advertised more than once without subsequent advertisements that would replace previous announcements.

Since the resolution of the NEXT_HOP address that is always conveyed in a BGP UPDATE message is left to the responsibility of the IGP that has been activated within the domain, the best path according to the BGP route selection process depicted in [3] SHOULD also be advertised. As long as the routers on the path towards the address depicted in the NEXT_HOP attribute of the message have the additional paths depicted in the QOS_NLRI attribute, the propagation of QoS routes within a domain where all the routers are QOS_NLRI aware should not yield inconsistent routing.

A BGP UPDATE message that carries the QOS_NLRI MUST also carry the ORIGIN and the AS_PATH attributes (both in eBGP and in iBGP exchanges). Moreover, in iBGP exchanges such a message MUST also

carry the LOCAL_PREF attribute. If such a message is received from an external peer, the local system shall check whether the leftmost AS in the AS_PATH attribute is equal to the autonomous system number of the peer than sent the message. If that is not the case, the local system shall send the NOTIFICATION message with Error Code UPDATE Message Error, and the Error Sub-code set to Malformed AS_PATH.

Finally, an UPDATE message that carries no NLRI, other than the one encoded in the QOS_NLRI attribute, should not carry the NEXT_HOP attribute. If such a message contains the NEXT_HOP attribute, the BGP speaker that receives the message should ignore this attribute.

6. Use of Capabilities Advertisement with BGP-4

A BGP speaker that uses the QOS_NLRI attribute SHOULD use the Capabilities Advertisement procedures, as defined in [12], so that it might be able to determine if it can use such an attribute with a particular peer.

The fields in the Capabilities Optional Parameter are defined as follows:

- The Capability Code field is set to N (127 < N < 256, when considering the "Private Use" range, as specified in [13]), while the Capability Length field is set to "1".

- The Capability Value field is a one-octet field, which contains the Type Code of the QOS_NLRI attribute, as defined in the introduction of section 5 of the present draft.

In addition, the multiple path advertisement capability MUST be supported, as defined in section 2.1 of [4].

7. Simulation Results

7.1. A Phased Approach

The simulation work basically aims at qualifying the scalability of the usage of the QOS_NLRI attribute for propagating QoS-related information across domains.

This effort also focused on the impact on the stability of the BGP routes, by defining a set of basic engineering rules for the introduction of additional QoS information, as well as design considerations for the computation and the selection of "QoS routes".

This ongoing development effort is organized into a step-by-step approach, which consists in the following phases:

1. Model an IP network composed of several autonomous systems. Since this simulation effort is primarily focused on the

qualification of the scalability related to the use of the QOS_NLRI attribute for exchanging QoS-related information between domains, it has been decided that the internal architecture of such domains should be kept very simple, i.e. without any specific IGP interaction,

2. Within this IP network, there are BGP peers that are QOS_NLRI aware, i.e. they have the ability to process the information conveyed in the attribute, while the other routers are not: the latter do not recognize the QOS_NLRI attribute by definition, and they will forward the information to other peers, by setting the Partial bit in the attribute, meaning that the information conveyed in the message is incomplete. This approach contributes to the qualification of a progressive deployment of QOS_NLRI-aware BGP peers,

3. As far as QOS_NLRI aware BGP peers are concerned, they will process the information contained in the QOS_NLRI attribute to possibly influence the route decision process, thus yielding the selection (and the announcement) of distinct routes towards a same destination prefix, depending on the QoS-related information conveyed in the QOS_NLRI attribute,

4. Modify the BGP route decision process: at this stage of the simulation, the modified decision process relies upon the one-way delay information (which corresponds to the QoS Information Code field of the attribute valued at "2"), and it also takes into account the value of the Partial bit of the attribute.

Once the creation of these components of the IP network has been completed (together with the modification of the BGP route selection process), the behavior of a QOS_NLRI-capable BGP peer is as follows.

Upon receipt of a BGP UPDATE message that contains the QOS_NLRI attribute, the router will first check if the corresponding route is already stored in its local RIB, according to the value of the one-way delay information contained in both QoS Information Code and Sub-code fields of the attribute.

If not, the BGP peer will install the route in its local RIB. Otherwise (i.e. an equivalent route already exists in its database), the BGP peer will select the best of both routes according to the following criteria:

- If both routes are said to be either incomplete (Partial bit has been set) or complete (Partial bit is unset), the route with the lowest delay will be selected,

- Otherwise, a route with the Partial bit unset is always preferred over any other route, even if this route reflects a higher transit delay.

If ever both Partial bit and transit delay information are not sufficient to make a decision, the standard BGP decision process (according to the breaking ties mechanism depicted in [3]) is performed.

7.2. A Case Study

REMINDER: a PDF version of this document (which includes the figures mentioned in this section) can be accessed at http://www.mescal.org.

As stated in the previous section 7.1, the current status of the simulation work basically relies upon the one-way transit delay information only, as well as the complete/incomplete indication of the Partial bit conveyed in the QOS_NLRI attribute.

The following figures depict the actual processing of the QoS-related information conveyed in the QOS_NLRI attribute, depending on whether the peer is QOS_NRLI-aware or not.

[Fig. 1: A Case Study.]

Figure 1 depicts the IP network that has been modelled, while figure 2 depicts the propagation of a BGP UPDATE message that contains the QOS_NLRI attribute, in the case where the contents of the attribute are changed, because of complete/incomplete conditions depicted by the Partial bit of the QOS_NLRI attribute.

[Fig. 2: Propagation of One-way Delay Information via BGP4.]

Router S in figure 2 is a QOS_NRLI-capable speaker. It takes 20 milliseconds for node S to reach network 192.0.20.0: this information will be conveyed in a QOS_NLRI attribute that will be sent by node S in a BGP UPDATE message with the Partial bit of the QOS_NLRI attribute unset.

Router A is another QOS_NLRI BGP peer, and it takes 3 milliseconds for A to reach router S. Node A will update the QoS-related information of a QOS_NLRI attribute, indicating that, to reach network 192.0.20.0, it takes 23 milliseconds. Router A will install this new route in its database, and will propagate the corresponding UPDATE message to its peers.

On the other hand, router B is not capable of processing the information conveyed in the QOS_NLRI attribute, and it will therefore set the Partial bit of the QOS_NLRI attribute in the corresponding UPDATE message, leaving the one-way delay information detailed in both QoS Information Code and Sub-code unchanged.

Upon receipt of the UPDATE message sent by router A, router E will update the one-way delay information since it is a QOS_NRLI-capable peer. Finally, router D receives the UPDATE message, and selects a

route with a 40 milliseconds one-way delay to reach network 192.0.20.0, as depicted in figure 3.

[Fig. 3: Selecting QoS Routes Across Domains.]

This simulation result shows that the selection of a delay-inferred route over a BGP route may not yield an optimal decision. In the above example, the 40 ms-route goes through routers D-E-A-S, while a "truly optimal" BGP route would be through routers D-E-F-A-S, hence a 38 ms-route. This is because of a BGP4 rule that does not allow router F to send an UPDATE message towards router E, because router F received the UPDATE message from router A thanks to the iBGP connection it has established with A.

7.3. Additional Results

The following table reflects the results obtained from a simulation network composed of 9 autonomous systems and 20 BGP peers. The numbers contained in the columns reflect the percentage of serviced requirements, where the requirements are expressed in terms of delay.

Three parameters have been taken into account:

- The percentage of BGP peers that have the ability to process the information conveyed in the QOS_NLRI attribute (denoted as "x% Q-BGP" in the following table),

- The transit delays "observed" (and artificially simulated) on each transmission link: the higher the delays, the lower the percentage of serviced QoS requirements,

- The QoS requirements themselves, expressed in terms of delay: as such, the strongest requirements (i.e. the lowest delays) have less chance to be satisfied.

Delay	0% Q-BGP	50% Q-BGP	100% Q-BGP
3	11	8,3	11
5	30,5	30,5	36,1
6	40	47,2	55,5
7	47	59,7	72,2
8	62,5	79	91,6
9	63	84,7	97,2
10	70,8	90,2	98,6
11	76,3	93	98,6

12	86,1	97,2	100
13	88,8	98,6	100
14	94,4	100	100
15	94,4	100	100
16	94,4	100	100
17	97,2	100	100
18	98,6	100	100
19	98,6	100	100
20	98,6	100	100
21	98,6	100	100
22	100	100	100

This table clearly demonstrates the technical feasibility of the approach, and how the use of the QOS_NLRI attribute can improve the percentage of serviced QoS requirements.

7.4. Next Steps

This simulation effort is currently pursued in order to better qualify the interest of using the BGP4 protocol to convey QoS-related information between domains, from a scalability perspective, i.e. the growth of BGP traffic vs. the stability of the network.

The stability of the IP network is probably one of the most important aspects, since QoS-related information is subject to very dynamic changes, thus yielding non-negligible risks of flapping.

8. IANA Considerations

Section 4 of this draft documents an optional transitive BGP-4 attribute named "QOS_NLRI" whose type value will be assigned by IANA. Section 5 of this draft also documents a Capability Code whose value should be assigned by IANA as well.

9. Security Considerations

This additional BGP-4 attribute specification does not change the underlying security issues inherent in the existing BGP-4 protocol specification [14].

10. References

[1] Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.

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

[3] Rekhter, Y., Li T., "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.

[4] Almes, G., Kalidindi, S., "A One-Way-Delay Metric for IPPM", RFC 2679, September 1999.

[5] Demichelis, C., Chimento, P., "IP Packet Delay Variation Metric for IP Performance Metrics (IPPM)", RFC 3393, November 2002.

[6] Nichols, K., Blake, S., Baker, F., Black, D., "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

[7] Traina, P., McPherson, D., Scudder, J., "Autonomous System Confederations for BGP", RFC 3065, February 2001.

[8] Jacquenet, C., "A COPS Client-Type for Traffic Engineering", draft-jacquenet-cops-te-00.txt, Work in Progress, February 2004.

[9] Apostolopoulos, G. et al, "QoS Routing Mechanisms and OSPF Extensions", RFC 2676, August 1999.

[10] Reynolds, J., Postel, J., "ASSIGNED NUMBERS", RFC 1700, October 1994.

[11] Walton, D., et al., "Advertisement of Multiple Paths in BGP", draft-walton-bgp-add-paths-01.txt, Work in Progress, November 2002.

[12] Chandra, R., Scudder, J., "Capabilities Advertisement with BGP-4", RFC 3392, November 2002.

[13] Narten, T., Alvestrand, H., "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 2434, October 1998.

[14] Heffernan, A., "Protection of BGP sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.

11. Acknowledgments

Part of this work is funded by the European Commission, within the context of the MESCAL (Management of End-to-End Quality of Service Across the Internet At Large, http://www.mescal.org) project, which is itself part of the IST (Information Society Technologies) research program.

The author would also like to thank all the partners of the MESCAL project for the fruitful discussions that have been conducted within the context of the traffic engineering specification effort of the project, as well as O. Bonaventure and B. Carpenter for their valuable input.

12. Authors' Addresses

Geoffrey Cristallo
Alcatel
Francis Wellesplein, 1
2018 Antwerp
Belgium
Phone: +32 (0)3 240 7890
E-Mail: geoffrey.cristallo@alcatel.be

Christian Jacquenet
France Telecom
3, avenue François Château
CS 36901
35069 Rennes Cedex
France
Phone: +33 2 99 87 63 31
Email: christian.jacquenet@francetelecom.com

13. Full Copyright Statement