Network Working Group
Internet Draft
Document: draft-jacquenet-qos-nlri-03.txt
Category: Experimental
Expires January 2002

G. Cristallo
Alcatel
C. Jacquenet
France Telecom R&D
July 2001

Providing Quality of Service Indication by the BGP-4 Protocol: the QOS_NLRI attribute
<draft-jacquenet-qos-nlri-03.txt>

Status of this Memo

Abstract

This draft specifies an additional BGP4 (Border Gateway Protocol, version 4, [2]) attribute, named the "QOS_NLRI" attribute, which aims at providing QoS (Quality of Service)-related information associated to the NLRI (Network Layer Reachability Information) information conveyed in a BGP UPDATE message.

1. Introduction

Providing end-to-end quality of service is probably 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 deployed and enforced within an autonomous system, and which may well differ from one AS (Autonomous System) to another.

For almost 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 by crossing 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 basic and common quality of service indicators may be shared between the service providers.

Activate the BGP4 protocol for exchanging reachability information between autonomous systems has been a must for many years, and, from this standpoint, the BGP4 protocol is one of the key components for the enforcement of end-to-end QoS policies.

Therefore, exchanging QoS-related information as well as 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 into the following sections:

- Section 3 identifies the changes that have been made in the document since the last version,

- Section 4 describes the attribute and its mode of operation,

- Section 5 elaborates on the use of the capabilities advertisement feature of the BGP4 protocol,

- Section 6 introduces the first results of an ongoing simulation work,

- Finally, sections 7 and 8 introduce IANA and some security considerations, respectively.

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 RFC 2119 [3].

3. Changes since the last version of this draft

The current version of this draft reflects the following changes:

- Slight re-wording of the Introduction section (section 1),

- Added a section on the first simulation results (section 6),

- Authors' list has been updated,

- Correction of remaining typos.

4. The QOS_NLRI attribute (Type Code XY*)

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

The QOS_NLRI attribute is an optional transitive attribute that can be used for the following purposes:

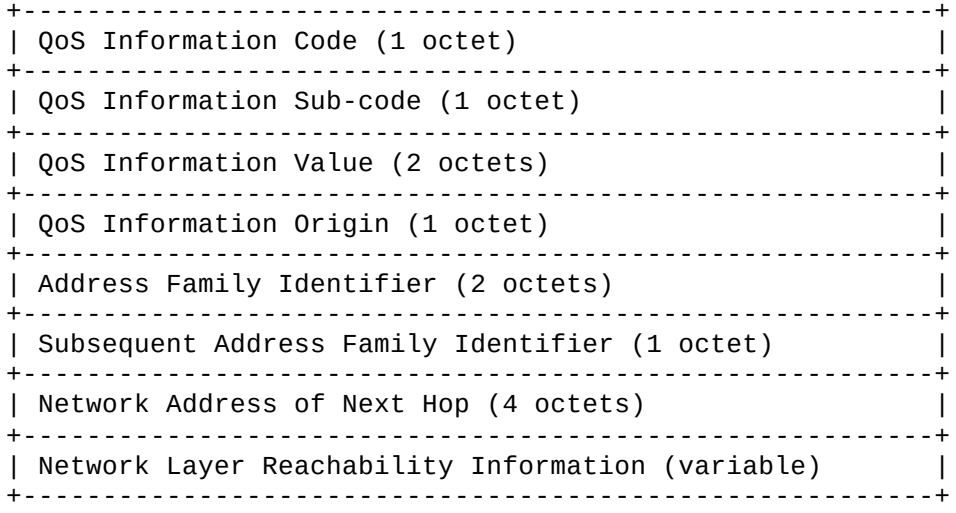
(a) To advertise a QoS route to a peer. A QoS route is a route that meets one or a set of QoS requirement(s) to reach a given (set of) destination prefixes (see [4], for example). Such QoS requirements can be expressed in terms of minimum one-way delay ([5]) to reach a destination, the experienced delay variation for IP datagrams that are destined to a given destination prefix ([6]), 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, [7]) marking). These QoS requirements can be used as an input for the route calculation process embedded in the BGP peers, e.g. thanks to the activation of a signaling protocol, such as RSVP (Resource ReSerVation Protocol, [8]),

(b) To provide QoS information along with the NLRI information in a single BGP UPDATE message. It is assumed that this QoS 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 basically motivated by the fact that this kind of attribute allows for gradual deployment of QoS extensions to 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.

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 kindly suggested to have a look on document [9], as an example of a means to feed the BGP peer with the appropriate information.

The QOS_NLRI attribute is encoded as follows:



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, as defined in [5]
(3) Inter-packet delay variation, as defined in [6]
(4) PHB Identifier, as defined in [10]

- 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

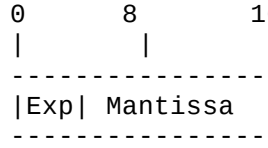
The instantiation of this sub-code field MUST be compatible with the value conveyed in the QoS Information code field, as stated in the following table (the rows represent the QoS Information Code possible values, the columns represent the QoS Information Sub-code values identified so far, while the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



  This encoding scheme advertises a numeric value that is (2^16 -1 - exponential encoding of the considered rate), as depicted in [11].
(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 [2].

- Address Family Identifier (AFI):

  This field carries the identity of the Network Layer protocol associated with the Network Address that follows. Presently defined values for this field are specified in [12] (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 NLRI 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.

- Network Layer Reachability Information:

  This variable length field lists the NLRI information for the feasible routes that are being advertised by this attribute. The next hop information carried in the QOS_NLRI path attribute defines the Network Layer address of the border router that should be used as the next hop to the destinations listed in the QOS_NLRI attribute in the UPDATE message.

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 the route is 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 can advertise any external peer router in the next hop component, provided that the Network Layer address of this border router was learned from an external peer, and the external peer to which the route is being advertised shares a common subnet with the next hop address. This is a second form of "third party" next hop information.

Normally the next hop information is chosen so that the shortest available path will be taken. A BGP speaker must be able to support disabling advertisement of third party next hop information to handle imperfectly bridged media or for reasons of policy.

A BGP speaker must never advertise an address of a peer to that peer as a next hop, for a route that the speaker is originating. A BGP speaker must never install a route with itself as the next hop.

When a BGP speaker advertises the route to an internal peer, the advertising speaker should not modify the next hop information associated with the route. When a BGP speaker receives the route via an internal link, it may forward packets to the next hop address if the address contained in the attribute is on a common subnet with the local and remote BGP speakers.

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.

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.

5. 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 [13], 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 [14]), 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 4 of the present draft.

6. First simulation results

6.1. A step-by-step approach

The simulation work that has begun within the context of the TEQUILA project (see [4]) basically aims at qualifying the scalability of the usage of the QOS_NLRI attribute for propagating QoS-related information between domains. This work also aims at quantifying the

added value provided by the QoS extensions to BGP4, as a function of the percentage of the accordingly updated BGP peers.

This effort has also been launched to focus on the impact on the stability of the BGP routes, by defining a set of basic engineering rules for the introduction of new QoS information, as well as design considerations for the calculation 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. Each of the autonomous systems is composed of BGP peers that have established iBGP connections between each other. 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 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: these routers 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 corresponding UPDATE messages, meaning that the information conveyed in the message is incomplete. This is the typical behavior that is expected when a BGP peer has to deal with an optional transitive attribute. This approach allows to elaborate on the added value introduced by a gradual deployment of the QoS extensions to BGP4,

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 installation) of distinct routes towards a same destination prefix, depending on the QoS-related information conveyed in the QOS_NLRI attribute. From this implementation perspective, the BGP routing tables have been modeled so that they contain a "sub-section" where QOS_NLRI-capable peers will store the information conveyed in the 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 will also take into account the value of the Partial bit in the UPDATE message, in the case where the QoS-related information contained in the QOS_NLRI attribute happens to be incomplete, because it's been relayed by a non-QOS_NLRI aware BGP peer.

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 incomplete (partial bit valued to "1" in the UPDATE message), or if both routes are said to be complete, the best route is the route with the lowest value of the QoS Information Sub-code field of the QOS_NLRI attribute,

- Otherwise, a complete QoS-related information is always preferred over an incomplete one, even if the complete route has a QoS Information Sub-code field with a better value.

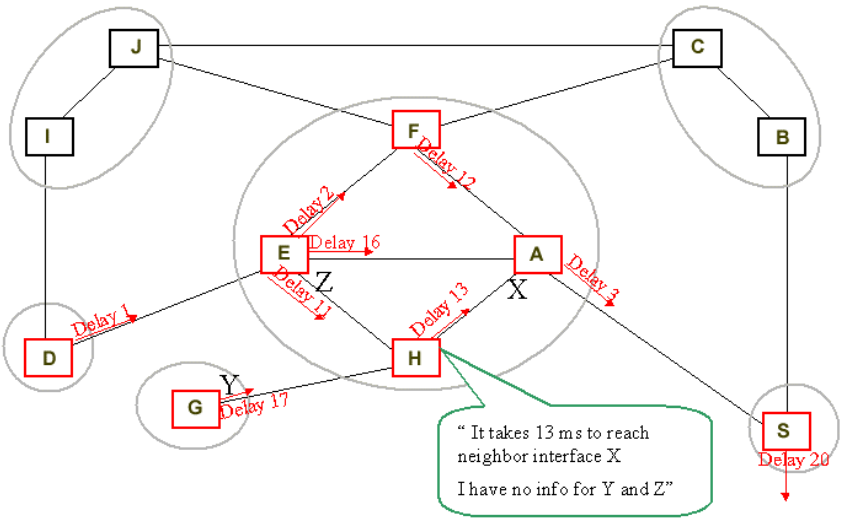
If the BGP route selection process cannot make a decision based upon the QoS Information Code and Sub-code fields (and possibly the complete/incomplete indication of the partial bit), then the BGP route selection process is basically based upon the recommendations stated in [2].

6.2. Current status of the simulation work

As stated in the previous section 6.1, 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 corresponding BGP UPDATE messages.

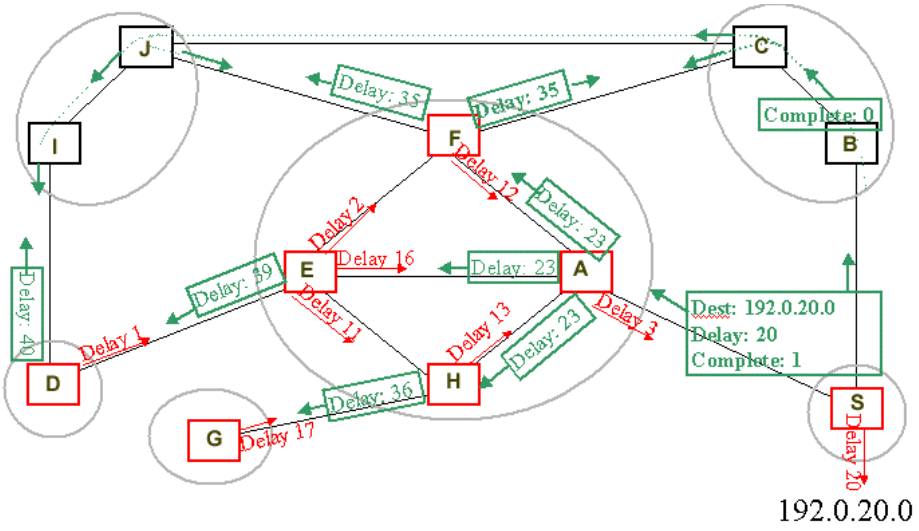
The IP network has been modeled so that it is composed of 6 autonomous systems and 11 BGP peers. Future scenarios will be composed of more ASs. The following figures depict the actual processing of the QoS-related information conveyed in the QOS_NLRI attribute, depending on whether the peer is QOS_NRLI-aware or not.

NOTE: the text version of this draft does not contain the above-mentioned figures, but a PDF version of this document can be accessed at the following link: http://www.ietf.org/ietf/1id-abstracts.txt.



- Figure 1: the modeled IP network. -

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 of the UPDATE message propagation.



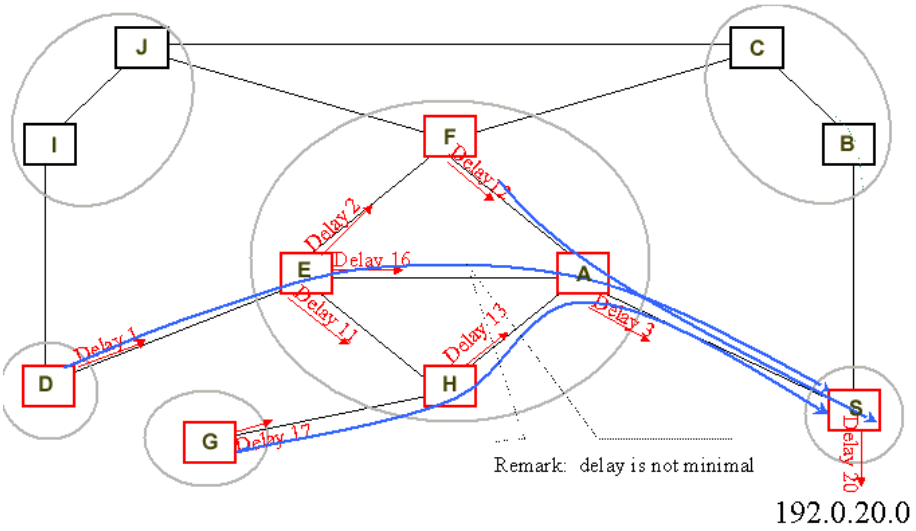
- Figure 2: propagation of a one-way delay information via BGP4. -

Router S in figure 2 is a QOS_NRLI-capable speaker. It takes 20 milliseconds to 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 a partial bit 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 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.



- Figure 3: installing QoS routes between domains. -

This simulation result shows that the selection of a delay-based route over a BGP route (as depicted in [2]) may not yield an optimum 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.

These basic observations confirm that the enforcement of a QoS policy between domains by using the BGP4 protocol is obviously conditioned by the BGP4 routing policies that are enforced within each domain.

6.3. Next steps

The above-mentioned simulation effort will be pursued in order to qualify the interest of using the BGP4 protocol to convey QoS-related information between domains, from a scalability perspective, i.e. the

increase 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.

It is therefore expected that the upcoming versions of this draft will reflect the progress of this simulation work, which will take into account additional autonomous systems, among other tracks of evolution.

7. 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.

8. Security Considerations

This additional BGP-4 attribute specification does not change the underlying security issues inherent in the existing BGP-4 protocol specification [15].

9. References

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

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

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

[4] Goderis, D., T'Joens, Y., Jacquenet, C., Memenios, G., Pavlou, G., Egan, R., Griffin, D., Georgatsos, P., Georgiadis, L., "Specification of a Service Level Specification (SLS) Template", draft-tequila-sls-00.txt, Work in Progress, November 2000. Check http://www.ist-tequila.org for additional information.

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

[6] Demichelis, C., Chimento, P., "IP Packet Delay Variation Metric for IPPM", draft-ietf-ippm-ipdv-07.txt, Work in Progress, February 2001.

[7] 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.

[8] Braden, R., et al., "Resource ReSerVation Protocol (RSVP)- Version 1 Functional Specification", RFC 2205, September 1997.

[9] Jacquenet, C., "A COPS client-type for IP traffic engineering", draft-jacquenet-ip-te-cops-02.txt, Work in Progress, June 2001.

[10] Black, D., Brim, S., Carpenter, B., Le Faucheur, F., "Per Hop Behavior Identification Codes", draft-ietf-diffserv-2839bis-02.txt, Work in Progress, May 2001.

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

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

[13] Chandra, R., Scudder, J., "Capabilities Advertisement with BGP-4", RFC 2842, May 2000.

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

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

10. Acknowledgments

Part of this work is funded by the European Commission, within the context of the TEQUILA (Traffic Engineering for Quality of Service in the Internet At Large Scale, [4]) 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 TEQUILA 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.

11. 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 R & D

DMI/SIR
42, rue des Coutures
BP 6243
14066 Caen Cedex 4
France
Phone: +33 2 31 75 94 28
E-Mail: christian.jacquenet@francetelecom.com

12. Full Copyright Statement