IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: July 14, 2014

Q. Wu
D. Wang
Huawei
S. Previdi
Cisco
H. Gredler
Juniper
S. Ray
Cisco

January 10, 2014

BGP attribute for North-Bound Distribution of Traffic Engineering (TE) performance Metrics
draft-ietf-idr-te-pm-bgp-00

Abstract

In order to populate network performance information like link latency, latency variation, packet loss and bandwidth into Traffic Engineering Database(TED) and ALTO server, this document describes extensions to BGP protocol, that can be used to distribute network performance information (such as link delay, delay variation, packet loss, residual bandwidth, available bandwidth and utilized bandwidth).

Table of Contents

1. Introduction

As specified in [RFC4655],a Path Computation Element (PCE) is an entity that is capable of computing a network path or route based on a network graph, and of applying computational constraints during the computation. In order to compute an end to end path, the PCE needs to have a unified view of the overall topology[I-D.ietf-pce-pcep-service-aware]. [I.D-ietf-idr-ls-distribution] describes a mechanism by which links state and traffic engineering information can be collected from networks and shared with external components using the BGP routing protocol. This mechanism can be used by both PCE and ALTO server to gather information about the topologies and capabilities of the network.

With the growth of network virtualization technology, the needs for inter-connection between various overlay technologies (e.g. Enterprise BGP/MPLS IP VPNs) in the Wide Area Network (WAN) become important. The Network performance or QoS requirements such as latency, limited bandwidth, packet loss, and jitter, are all critical factors that must be taken into account in the end to end path computation ([I-D.ietf-pce-pcep-service-aware]) and selection which enable establishing segment overlay tunnel between overlay nodes and stitching them together to compute end to end path.

In order to populate network performance information like link latency, latency variation, packet loss and bandwidth into TED and ALTO server, this document describes extensions to BGP protocol, that can be used to distribute network performance information (such as link delay, delay variation, packet loss, residual bandwidth, available bandwidth, and utilized bandwidth). The network performance information can be distributed in the same way as link state information distribution,i.e., either directly or via a peer BGP speaker (see figure 1 of [I.D-ietf-idr-ls-distribution]).

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 RFC2119 [RFC2119].

3. Use Cases

3.1. MPLS-TE with H-PCE

For inter-AS path computation the Hierarchical PCE (H-PCE) [RFC6805] may be used to compute the optimal sequence of domains. Within the H-PCE architecture, the child PCE communicates domain connectivity information to the parent PCE, and the parent PCE will use this information to compute a multi-domain path based on the optimal TE links between domains [I.D-ietf-pce-hierarchy-extensions] for the end-to-end path.

The following figure demonstrates how a parent PCE may obtain TE performance information beyond that contained in the LINK_STATE attributes [I.D-ietf-idr-ls-distribution] using the mechanism described in this document.

```
                  +----------+                           +---------+
                  |  -----   |                           |   BGP   |
                  | | TED |<-+-------------------------->| Speaker |
                  |  -----   |   TED synchronization     |         |
                  |    |     |        mechanism:         +---------+
                  |    |     | BGP with TE performance
                  |    v     |        NLRI
                  |  -----   |
                  | | PCE |  |
                  |  -----   |
                  +----------+
                       ^
                       | Request/
                       | Response
                       v
         Service  +----------+   Signaling  +----------+
         Request  | Head-End |   Protocol   | Adjacent |
         -------->|  Node    |<------------>|   Node   |
                  +----------+              +----------+
```

Figure 1: External PCE node using a TED synchronization mechanism

3.2. ALTO Server Network API

The ALTO Server can aggregate information from multiple systems to provide an abstract and unified view that can be more useful to applications.

The following figure shows how an ALTO Server can get TE performance information from the underlying network beyond that contained in the LINK_STATE attributes [I.D-ietf-idr-ls-distribution] using the mechanism described in this document.

```
+--------+
| Client |<--+
+--------+   |
             |    ALTO    +--------+     BGP with    +---------+
+--------+   |  Protocol  |  ALTO  |  TE Performance |   BGP   |
| Client |<--+------------| Server |<----------------| Speaker |
+--------+   |            |        |      NLR        |         |
             |            +--------+                 +---------+
+--------+   |
| Client |<--+
+--------+
```

Figure 2: ALTO Server using network performance information

4. Carrying TE Performance information in BGP

This document proposes new BGP TE performance TLVs that can be announced as attribute in the BGP-LS attribute (defined in [I.D-ietf-idr-ls-distribution]) to distribute network performance information. The extensions in this document build on the ones provided in BGP-LS [I.D-ietf-idr-ls-distribution] and BGP-4 [RFC4271].

BGP-LS attribute defined in [I.D-ietf-idr-ls-distribution] has nested TLVs which allow the BGP-LS attribute to be readily extended. This document proposes seven additional TLVs as its attributes:

Type	Value
TBD1	Unidirectional Link Delay
TBD2	Min/Max Unidirectional Link Delay
TBD3	Unidirectional Delay Variation
TBD4	Unidirectional Packet Loss
TBD5	Unidirectional Residual Bandwidth
TBD6	Unidirectional Available Bandwidth
TBD7	Unidirectional Utilized Bandwidth

As can be seen in the list above, the TLVs described in this document carry different types of network performance information. These TLVs include a bit called the Anomalous (or "A") bit at the left-most bit after length field of each TLV defined in figure 4 of [[I.D-ietf-idr-ls-distribution]]. The other bits in the first octets after length field of each TLV is reserved for future use. When the A bit is clear (or when the TLV does not include an A bit), the TLV describes steady state link performance. This information could conceivably be used to construct a steady state performance topology for initial tunnel path computation, or to verify alternative failover paths.

When network performance downgrades and exceeds configurable maximum thresholds, a TLV with the A bit set is advertised. These TLVs could be used by the receiving BGP peer to determine whether to redirect failing traffic to a backup path, or whether to calculate an entirely new path. If link performance improves later and falls below a configurable value, that TLV can be re- advertised with the Anomalous bit cleared. In this case, a receiving BGP peer can conceivably do whatever re-optimization (or failback) it wishes to do (including nothing).

Note that when a TLV does not include the A bit, that TLV cannot be used for failover purposes. The A bit was intentionally omitted from some TLVs to help mitigate oscillations.

Consistent with existing ISIS TE specifications [ISIS-TE-METRIC], the bandwidth advertisements, the delay and delay variation advertisements, packet loss defined in this document MUST be encoded in the same unit as one defined in IS-IS Extended IS Reachability sub-TLVs [ISIS-TE-METRIC]. All values (except residual bandwidth) MUST be obtained by a filter that is reasonably representative of an average or calculated as rolling averages where the averaging period MUST be a configurable period of time. The measurement interval, any filter coefficients, and any advertisement intervals MUST be configurable per sub-TLV in the same way as ones defined in section 5 of [ISIS-TE-METRIC].

5. Attribute TLV Details

Link attribute TLVs defined in section 3.2.2 of [I-D.ietf-idr-ls-distribution]are TLVs that may be encoded in the BGP-LS attribute with a link NLRI. Each 'Link Attribute' is a Type/Length/ Value (TLV) triplet formatted as defined in Section 3.1 of [I-D.ietf-idr-ls-distribution]. The format and semantics of the 'value' fields in 'Link Attribute' TLVs correspond to the format and semantics of value fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305]. Although the encodings for 'Link Attribute' TLVs were originally defined for IS-IS, the TLVs can carry data sourced either by IS-IS or OSPF.

The following 'Link Attribute' TLVs are valid in the LINK_STATE attribute:

TLV Code Point	Description	IS-IS TLV/Sub-TLV	Defined in:
xxxx	Unidirectional Link Delay	22/xx	[ISIS-TE-METRIC]/4.1
xxxx	Min/Max Unidirection Link Delay	22/xx	[ISIS-TE-METRIC]/4.2
xxxx	Unidirectional Delay Variation	22/xx	[ISIS-TE-METRIC]/4.3
xxxx	Unidirectional Link Loss	22/xx	[ISIS-TE-METRIC]/4.4
xxxx	Unidirectional Residual Bandwidth	22/xx	[ISIS-TE-METRIC]/4.5
xxxx	Unidirectional Available Bandwidth	22/xx	[ISIS-TE-METRIC]/4.6
xxxx	Unidirectional Utilized Bandwidth	22/xx	[ISIS-TE-METRIC]/4.7

Table 1: Link Attribute TLVs

6. Security Considerations

This document does not introduce security issues beyond those discussed in [I.D-ietf-idr-ls-distribution] and [RFC4271].

7. IANA Considerations

IANA maintains the registry for the TLVs. BGP TE Performance TLV will require one new type code per TLV defined in this document.

8. References

8.1. Normative References

[I-D.ietf-idr-ls-distribution]
Gredler, H., "North-Bound Distribution of Link-State and TE Information using BGP", ID draft-ietf-idr-ls-distribution-03, May 2013.

[I-D.ietf-pce-pcep-service-aware]
Dhruv, D., "Extensions to the Path Computation Element Communication Protocol (PCEP) to compute service aware Label Switched Path (LSP)", ID draft-ietf-pce-pcep-service-aware-01, July 2013.

[ISIS-TE-METRIC]
Giacalone, S., "ISIS Traffic Engineering (TE) Metric Extensions", ID draft-ietf-isis-te-metric-extensions-00, June 2013.

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

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

[RFC5305] Li, T., "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.

8.2. Informative References

[ALTO] Yang, Y., "ALTO Protocol", ID http://tools.ietf.org/html/draft-ietf-alto-protocol-16, May 2013.

[I.D-ietf-pce-hierarchy-extensions]
Zhang, F., Zhao, Q., Gonzalez de Dios, O., Casellas, R., and D. King, "Extensions to Path Computation Element Communication Protocol (PCEP) for Hierarchical Path Computation Elements (PCE)", ID draft-ietf-pce-hierarchy-extensions-00, August 2013.

[RFC4655] Farrel, A., "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.

Appendix A. Contributor Addresses

Jeff Tantsura
Ericsson
300 Holger Way
San Jose, CA 95134
US

Email: Jeff.Tantsura@ericsson.com

Appendix B. Change Log

Note to the RFC-Editor: please remove this section prior to publication as an RFC.

B.1. draft-ietf-idr-te-pm-bgp-00

The following are the major changes compared to previous version draft-wu-idr-te-pm-bgp-03:

- o Update PCE case in section 3.1.
- o Add some texts in section 1 and section 4 to clarify from where to distribute pm info and measurement interval and method.

Authors' Addresses

Qin Wu
Huawei
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: bill.wu@huawei.com

Danhua Wang
Huawei
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: wangdanhua@huawei.com

Stefano Previdi
Cisco Systems, Inc.
Via Del Serafico 200
Rome 00191
Italy

Email: sprevidi@cisco.com

Hannes Gredler
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: hannes@juniper.net

Saikat Ray
Cisco Systems, Inc.
170, West Tasman Drive
San Jose, CA 95134
US

Email: sairay@cisco.com