Inter-Domain Routing Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 13, 2014

Q. Wu
D. Wang
Huawei

July 12, 2013

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

Abstract

In order to populate network performance information like link latency, latency variation and packet loss 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, and available bandwidth,link utilization, channel throughput).

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 13, 2014.

Copyright Notice

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-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-connecting 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 path computation and selection to establish segment overlay tunnel between overlay nodes and stitch them together to compute end to end path.

In order to populate network performance information like link latency, latency variation and packet loss 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, and available bandwidth,link utilization, channel throughput).

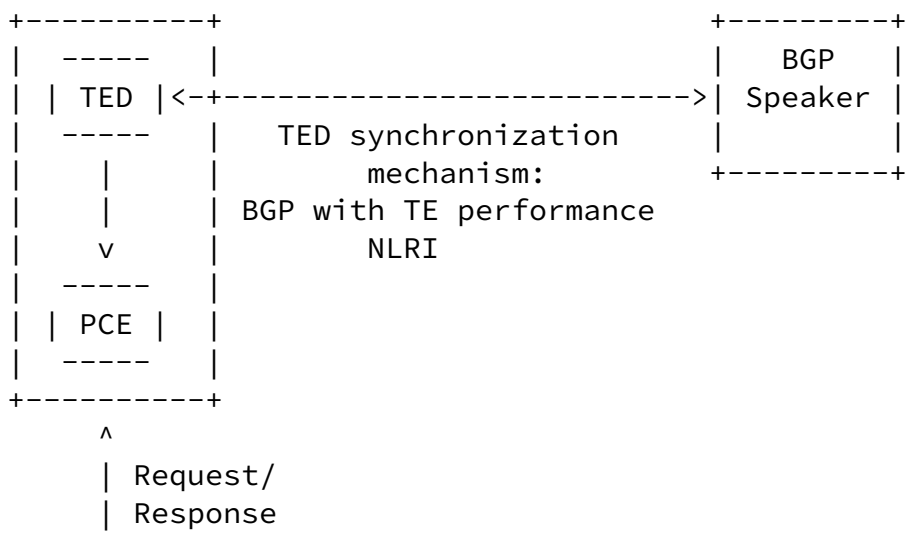
2. Conventions used in this document

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 PCE

The following figure shows how a PCE can get its TE performance information beyond that contained in the LINK_STATE attributes [I.D -ietf-idr- ls-distribution] using the mechanism described in this document.



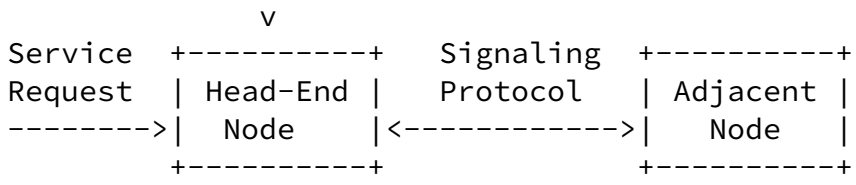


Figure 1: External PCE node using a TED synchronization mechanism

3.2. ALTO Server Network API

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.

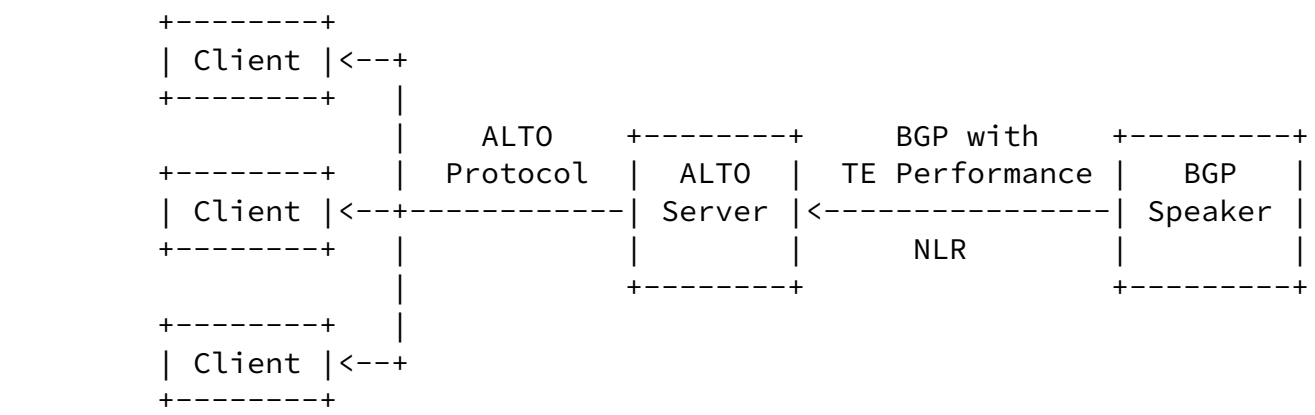


Figure 2: ALTO Server using network performance information

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 NLRI (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 NLRI defined in [I.D-ietf-idr-ls-distribution] has nested TLVs which allow the BGP-LS NLRI to be readily extended. This document

proposes several additional TLVs as its attributes:

Type	Value
TBD1	Unidirectional Link Delay
TBD2	Unidirectional Delay Variation
TBD3	Unidirectional Packet Loss
TBD4	Unidirectional Residual Bandwidth
TBD5	Unidirectional Available Bandwidth
TBD6	Link Utilization
TBD7	Channel Throughput

As can be seen in the list above, the TLVs described in this document carry different types of network performance information. Many (but not all) of the TLVs include a bit called the Anomalous (or "A") bit. 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 falls below configurable link-local thresholds a TLV with the A bit set is advertised. These TLVs could be used by the receiving node 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 exceeds a configurable minimum value (i.e.,threshold), that TLV can be re-advertised with the Anomalous bit cleared. In this case, a receiving node 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 sub-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 [RFC5305][ISIS-TE-METRIC], the bandwidth advertisements defined in this document MUST be encoded as IEEE floating point values. The delay and delay variation advertisements defined in this draft MUST be encoded as integer values. Delay values MUST be quantified in units of microseconds, packet loss MUST be quantified as a percentage of packets sent, and bandwidth MUST be sent as bytes per second. All values (except residual bandwidth) MUST be calculated as rolling averages where the averaging period MUST be a configurable period of time.

5. Attribute TLV Details

Link attribute TLVs 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 some 'Link Attribute' TLVs correspond to the format and semantics of value fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305] and . 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 are valid in the LINK_STATE attribute:

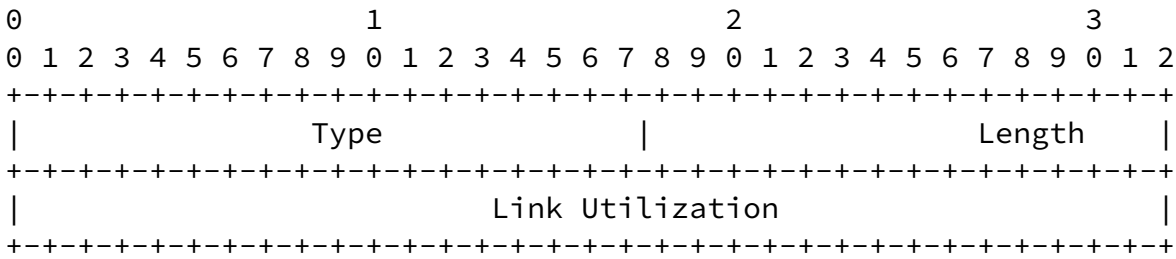
TLV Code Point	Description	IS-IS TLV/Sub-TLV	Defined in:
xxxx	Unidirectional Link Delay	22/xx	[ISIS-TE-METRIC]/4.1
xxxx	Min/Max Unidirection Link Delay	22/xx	[ISIS-TE-METRIC]/4.2
xxxx	Unidirectional Delay Variation	22/xx	[ISIS-TE-METRIC]/4.3
xxxx	Unidirectional Link Loss	22/xx	[ISIS-TE-METRIC]/4.4

xxxx	Unidirectional Residual Bandwidth	22/xx	[ISIS-TE-METRIC]/4.5
xxxx	Unidirectional Available Bandwidth	22/xx	[ISIS-TE-METRIC]/4.6
xxxx	Link Utilization	----	section 5.1
xxxx	Channel Throughput	----	section 5.2

Table 1: Link Attribute TLVs

5.1. Link Utilization TLV

This TLV advertises the average link utilization between two directly connected IS-IS neighbors. The link utilization advertised by this sub-TLV MUST be the utilization percentage per interval from the local neighbor to the remote one. The format of this sub-TLV is shown in the following diagram:



where:

Type: TBA

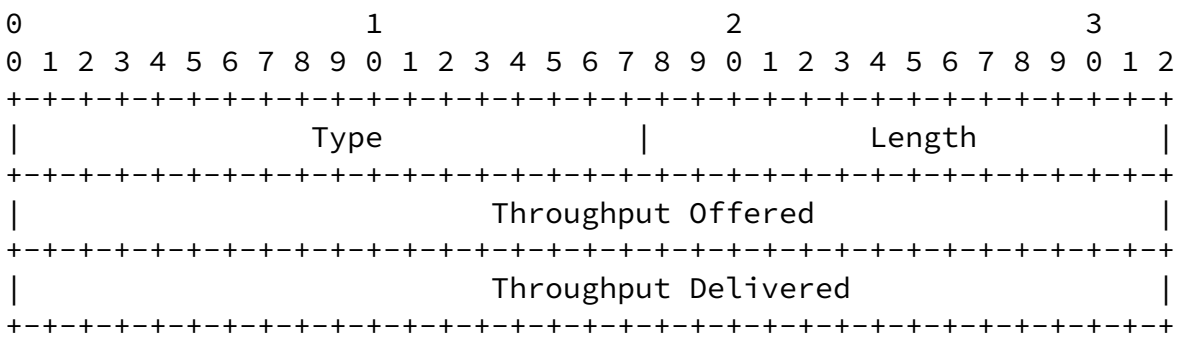
Length: 4

Link Utilization. This 24-bit field carries the average link untilization over a configurable interval. A commonly used time interval is 5 minutes, and this interval has been sufficient to support network operations and design for some

time. link utilization can be calculated by counting the IP-layer (or other layer) octets received over a time interval and dividing by the theoretical maximum number of octets that could have been delivered in the same interval(see section6.4 of [RFC6703]). If there is no value to send (unmeasured and not statically specified), then the sub-TLV should not be sent or be withdrawn.

5.2. Channel Throughput TLV

This TLV advertises the average Channel Throughput between two directly connected IS-IS neighbors. The channel throughput advertised by this sub-TLV MUST be the throughput between the local neighbor and the remote one. The format of this sub-TLV is shown in the following diagram:



where:

Type: TBA

Length: 8

Throughput offered: This 24-bit field carries the average throughput offered over a configurable interval. Throughput offered can be calculated by counting the number of units successfully transmitted in the interval (See section 2.3 of [RFC6374]). If there is no value to send (unmeasured and not statically specified), then

the sub-TLV should not be sent or be withdrawn.

Throughput delivered: This 24-bit field carries the average throughput delivered over a configurable interval. Throughput delivered can be calculated by counting the number of units successfully received in the interval (See section 2.3 of [RFC6374]). If there is no value to send (unmeasured and not statically specified), then the sub-TLV should not be sent or be withdrawn.

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.

[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.

[RFC6374] Frost, D., "Packet Loss and Delay Measurement for MPLS Networks ", RFC 6374, September 2011.

[RFC6703] Morton, A., "Reporting IP Network Performance Metrics: Different Points of View ", RFC 6703, August 2012.

8.2. Informative References

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

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

Authors' Addresses

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

Email: sunseawq@huawei.com

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

Email: wangdanhua@huawei.com