                  Calculating Interior Gateway Protocol (IGP) Routes
                        Over Traffic Engineering Tunnels

Status of this Memo

Copyright Notice

Abstract

   This document describes how conventional hop-by-hop link-state
   routing protocols interact with new Traffic Engineering capabilities
   to create Interior Gateway Protocol (IGP) shortcuts. In particular,
   this document describes how Dijkstra's Shortest Path First (SPF)
   algorithm can be adapted so that link-state IGPs will calculate IP
   routes to forward traffic over tunnels that are set up by Traffic
   Engineering.

1. Introduction

   Link-state protocols like integrated Intermediate System to
   Intermediate System (IS-IS) [1] and OSPF [2] use Dijkstra's SPF
   algorithm to compute a shortest path tree to all nodes in the
   network. Routing tables are derived from this shortest path tree.
   The routing tables contain tuples of destination and first-hop
   information. If a router does normal hop-by-hop routing, the first-
   hop will be a physical interface attached to the router. New traffic
   engineering algorithms calculate explicit routes to one or more nodes
   in the network. At the router that originates explicit routes, such
   routes can be viewed as logical interfaces which supply Label
   Switched Paths through the network. In the context of this document,
   we refer to these Label Switched Paths as Traffic Engineering tunnels
   (TE-tunnels). Such capabilities are specified in [3] and [4].

   The existence of TE-tunnels in the network and how the traffic in the
   network is switched over those tunnels are orthogonal issues. A node
   may define static routes pointing to the TE-tunnels, it may match the

recursive route next-hop with the TE-tunnel end-point address, or it
may define local policy such as affinity based tunnel selection for
switching certain traffic. This document describes a mechanism
utilizing link-state IGPs to dynamically install IGP routes over
those TE-tunnels.

The tunnels under consideration are tunnels created explicitly by the
node performing the calculation, and with an end-point address known
to this node. For use in algorithms such as the one described in
this document, it does not matter whether the tunnel itself is
strictly or loosely routed. A simple constraint can ensure that the
mechanism be loop free. When a router chooses to inject a packet
addressed to a destination D, the router may inject the packet into a
tunnel where the end-point is closer (according to link-state IGP
topology) to the destination D than is the injecting router. In
other words, the tail-end of the tunnel has to be a downstream IGP
node for the destination D. The algorithms that follow are one way
that a router may obey this rule and dynamically make intelligent
choices about when to use TE-tunnels for traffic. This algorithm may
be used in conjunction with other mechanisms such as statically
defined routes over TE-tunnels or traffic flow and QoS based TE-
tunnel selection.

This IGP shortcut mechanism assumes the TE-tunnels have already been
setup. The TE-tunnels in the network may be used for QoS, bandwidth,
redundancy, or fastreroute reasons. When an IGP shortcut mechanism
is applied on those tunnels, or other mechanisms are used in
conjunction with an IGP shortcut, the physical traffic switching
through those tunnels may not match the initial traffic engineering
setup goal. Also the traffic pattern in the network may change with
time. Some forwarding plane measurement and feedback into the
adjustment of TE-tunnel attributes need to be there to ensure that
the network is being traffic engineered efficiently [6].

2. Enhancement to the Shortest Path First Computation

During each step of the SPF computation, a router discovers the path
to one node in the network. If that node is directly connected to
the calculating router, the first-hop information is derived from the
adjacency database. If a node is not directly connected to the
calculating router, it inherits the first-hop information from the
parent(s) of that node. Each node has one or more parents. Each
node is the parent of zero or more down-stream nodes.

For traffic engineering purposes, each router maintains a list of all
TE-tunnels that originate at this router. For each of those TE-
tunnels, the router at the tail-end is known.

During SPF, when a router finds the path to a new node (in other
words, this new node is moved from the TENTative list to the PATHS
list), the router must determine the first-hop information. There
are three possible ways to do this:

   - Examine the list of tail-end routers directly reachable via a
      TE-tunnel. If there is a TE-tunnel to this node, we use the
      TE-tunnel as the first-hop.

   - If there is no TE-tunnel, and the node is directly connected,
      we use the first-hop information from the adjacency database.

   - If the node is not directly connected, and is not directly
      reachable via a TE-tunnel, we copy the first-hop information
      from the parent node(s) to the new node.

The result of this algorithm is that traffic to nodes that are the
tail-end of TE-tunnels, will flow over those TE-tunnels. Traffic to
nodes that are downstream of the tail-end nodes will also flow over
those TE-tunnels. If there are multiple TE-tunnels to different
intermediate nodes on the path to destination node X, traffic will
flow over the TE-tunnel whose tail-end node is closest to node X. In
certain applications, there is a need to carry both the native
adjacency and the TE-tunnel next-hop information for the TE-tunnel
tail-end and its downstream nodes. The head-end node may
conditionally switch the data traffic onto TE-tunnels based on user
defined criteria or events; the head-end node may also split flow of
traffic towards either types of the next-hops; the head-end node may
install the routes with two different types of next-hops into two
separate RIBs. Multicast protocols running over physical links may
have to perform RPF checks using the native adjacency next-hops
rather than the TE-tunnel next-hops.

3. Special Cases and Exceptions

The Shortest Path First algorithm will find equal-cost parallel paths
to destinations. The enhancement described in this document does not
change this. Traffic can be forwarded over one or more native IP
paths, over one or more TE-tunnels, or over a combination of native
IP paths and TE-tunnels.

A special situation occurs in the following topology:

    rtrA -- rtrB -- rtrC
              | |
            rtrD -- rtrE

Assume all links have the same cost. Assume a TE-tunnel is set up
from rtrA to rtrD. When the SPF calculation puts rtrC on the
TENTative list, it will realize that rtrC is not directly connected,
and thus it will use the first-hop information from the parent, which
is rtrB. When the SPF calculation on rtrA moves rtrD from the
TENTative list to the PATHS list, it realizes that rtrD is the tail-
end of a TE-tunnel. Thus rtrA will install a route to rtrD via the
TE-tunnel, and not via rtrB.

When rtrA puts rtrE on the TENTative list, it realizes that rtrE is
not directly connected, and that rtrE is not the tail-end of a TE-
tunnel. Therefore, rtrA will copy the first-hop information from the
parents (rtrC and rtrD) to the first-hop information of rtrE.
Traffic to rtrE will now load-balance over the native IP path via
rtrA->rtrB->rtrC, and the TE-tunnel rtrA->rtrD.

In the case where both parallel native IP paths and paths over TE-
tunnels are available, implementations can allow the network
administrator to force traffic to flow over only TE-tunnels (or only
over native IP paths) or both to be used for load sharing.

4. Metric Adjustment of IP Routes over TE-tunnels

When an IGP route is installed in the routing table with a TE-tunnel
as the next hop, an interesting question is what should be the cost
or metric of this route? The most obvious answer is to assign a
metric that is the same as the IGP metric of the native IP path as if
the TE-tunnels did not exist. For example, rtrA can reach rtrC over
a path with a cost of 20. X is an IP prefix advertised by rtrC. We
install the route to X in rtrA's routing table with a cost of 20.
When a TE-tunnel from rtrA to rtrC comes up, by default the route is
still installed with metric of 20, only the next-hop information for
X is changed.

While this scheme works well, in some networks it might be useful to
change the cost of the path over a TE-tunnel, to make the route over
the TE-tunnel less or more preferred than other routes.

For instance, when equal cost paths exist over a TE-tunnel and over a
native IP path, by adjusting the cost of the path over the TE-tunnel,
we can force traffic to prefer the path via the TE-tunnel, to prefer
the native IP path, or to load-balance among them. Another example
is when multiple TE-tunnels go to the same or different destinations.
Adjusting TE-tunnel metrics can force the traffic to prefer some TE-
tunnels over others regardless of underlining IGP cost to those
destinations.

Setting a manual metric on a TE-tunnel does not impact the SPF
algorithm itself. It only affects the comparison of the new route
with existing routes in the routing table. Existing routes can be
either IP routes to another router that advertises the same IP
prefix, or it can be a path to the same router, but via a different
outgoing interface or different TE-tunnel. All routes to IP prefixes
advertised by the tail-end router will be affected by the TE-tunnel
metric. Also, the metrics of paths to routers that are downstream of
the tail-end router will be influenced by the manual TE-tunnel
metric.

This mechanism is loop free since the TE-tunnels are source-routed
and the tunnel egress is a downstream node to reach the computed
destinations. The end result of TE-tunnel metric adjustment is more
control over traffic loadsharing. If there is only one way to reach
a particular IP prefix through a single TE-tunnel, then no matter
what metric is assigned, the traffic has only one path to go.

The routing table described in this section can be viewed as the
private RIB for the IGP. The metric is an important attribute to the
routes in the routing table. A path or paths with lower metric will
be selected over other paths for the same route in the routing table.

4.1. Absolute and Relative Metrics

It is possible to represent the TE-tunnel metric in two different
ways: an absolute (or fixed) metric or a relative metric, which is
merely an adjustment of the dynamic IGP metric as calculated by the
SPF computation. When using an absolute metric on a TE-tunnel, the
cost of the IP routes in the routing table does not depend on the
topology of the network. Note that this fixed metric is not only
used to compute the cost of IP routes advertised by the router that
is the tail-end of the TE-tunnel, but also for all the routes that
are downstream of this tail-end router. For example, if we have TE-
tunnels to two core routers in a remote POP, and one of them is
assigned with an absolute metric of 1, then all the traffic going to
that POP will traverse this low-metric TE-tunnel.

By setting a relative metric, the cost of IP routes in the routing
table is based on the IGP metric as calculated by the SPF
computation. This relative metric can be a positive or a negative
number. Not configuring a metric on a TE-tunnel is a special case of
the relative metric scheme. No metric is the same as a relative
metric of 0. The relative metric is bounded by minimum and maximum
allowed metric values while the positive metric disables the TE-
tunnel in the SPF calculation.

4.2. Examples of Metric Adjustment

   Assume the following topology. X, Y, and Z are IP prefixes
   advertised by rtrC, rtrD, and rtrE respectively. T1 is a TE-tunnel
   from rtrA to rtrC. Each link in the network has an IGP metric of 10.

```
        ===== T1 =====>
      rtrA -- rtrB -- rtrC -- rtrD -- rtrE
          10       10  |  10  |  10  |
                       X      Y      Z
```

   Without TE-tunnel T1, rtrA will install IP routes X, Y, and Z in the
   routing table with metrics 20, 30, and 40 respectively. When rtrA
   has brought up TE-tunnel T1 to rtrC, and if rtrA is configured with
   the relative metric of -5 on tunnel T1, then the routes X, Y, and Z
   will be installed in the routing table with metrics 15, 25, and 35.
   If an absolute metric of 5 is configured on tunnel T1, then rtrA will
   install routes X, Y, and Z all with metrics 5, 15, and 25
   respectively.

5. Security Considerations

   This document does not change the security aspects of IS-IS or OSPF.
   Security considerations specific to each protocol still apply. For
   more information see [5] and [2].

6. Acknowledgments

   The authors would like to thank Joel Halpern and Christian Hopps for
   their comments on this document.

7. Informative References

   [1] ISO. Information Technology - Telecommunications and Information
       Exchange between Systems - Intermediate System to Intermediate
       System Routing Exchange Protocol for Use in Conjunction with the
       Protocol for Providing the Connectionless-Mode Network Service.
       ISO, 1990.

   [2] Moy, J., "OSPF Version 2", RFC 2328, April 1998.

   [3] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J.
       McManus, "Requirements for Traffic Engineering Over MPLS", RFC
       2702, September 1999.

   [4] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G.
       Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209,
       December 2001.

   [5] Li, T. and R. Atkinson, "Intermediate System to Intermediate
       System (IS-IS) Cryptographic Authentication", RFC 3567, July
       2003.

   [6] Awduche, D., Chiu, A., Elwalid, A., Widjaja, I., and X. Xiao,
       "Overview and Principles of Internet Traffic Engineering", RFC
       3272, May 2002.

8. Authors' Addresses

   Naiming Shen
   Redback Networks, Inc.
   300 Holger Way
   San Jose, CA 95134

   EMail: naiming@redback.com

   Henk Smit

   EMail: hhwsmit@xs4all.nl

9. Full Copyright Statement

   Copyright (C) The Internet Society (2004).

   This document is subject to the rights, licenses and restrictions
   contained in BCP 78, and except as set forth therein, the authors
   retain all their rights.

   This document and the information contained herein are provided on an
   "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS
   OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY AND THE INTERNET
   ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED,
   INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE
   INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED
   WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Intellectual Property

   The IETF takes no position regarding the validity or scope of any
   Intellectual Property Rights or other rights that might be claimed to
   pertain to the implementation or use of the technology described in
   this document or the extent to which any license under such rights
   might or might not be available; nor does it represent that it has
   made any independent effort to identify any such rights. Information
   on the IETF's procedures with respect to rights in IETF Documents can
   be found in BCP 78 and BCP 79.

   Copies of IPR disclosures made to the IETF Secretariat and any
   assurances of licenses to be made available, or the result of an
   attempt made to obtain a general license or permission for the use of
   such proprietary rights by implementers or users of this
   specification can be obtained from the IETF on-line IPR repository at
   http://www.ietf.org/ipr.

   The IETF invites any interested party to bring to its attention any
   copyrights, patents or patent applications, or other proprietary
   rights that may cover technology that may be required to implement
   this standard. Please address the information to the IETF at ietf-
   ipr@ietf.org.

Acknowledgement

   Funding for the RFC Editor function is currently provided by the
   Internet Society.