Network Working Group   C. Pelsser
                                                       S. Uhlig
                                                  O. Bonaventure
Internet-Draft   UCL (Belgium)
Expires: April 2005   October 2004

      Limitations induced by BGP on the computation of interdomain LSPs
                   draft-pelsser-bgp-pce-00.txt

Status of this Memo

   This document is an Internet-Draft and is subject to all provisions
   of section 3 of RFC 3667. By submitting this Internet-Draft, each
   author represents that any applicable patent or other IPR claims of
   which he or she is aware have been or will be disclosed, and any of
   which he or she become aware will be disclosed, in accordance with
   RFC 3668.

   Internet-Drafts are working documents of the Internet Engineering
   Task Force (IETF), its areas, and its working groups. Note that
   other groups may also distribute working documents as
   Internet-Drafts.

   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."

   The list of current Internet-Drafts can be accessed at
   http://www.ietf.org/ietf/1id-abstracts.txt.

   The list of Internet-Draft Shadow Directories can be accessed at
   http://www.ietf.org/shadow.html.

   This Internet-Draft will expire on April 1, 2005.

Copyright Notice

Abstract

   Path Computation Elements have been proposed to aid the establishment
   of interdomain Label Switched Paths. We propose to colocate the PCE
   with a route reflector and show that the performance of such a PCE
   depends on the quality of the interdomain routes that it collects.

1. Introduction

Service Providers have been using MPLS for various purposes inside
their networks. Recently, several service providers have expressed
their requirements[INTER-AS] to also use MPLS accross interdomain
boundaries, notably for QoS, traffic engineering and fast restoration
purposes. One of the proposed solutions to aid the establishment of
interdomain LSPs is the utilization of a Path Computation Element
(PCE).

The current PCE architecture document [PCE] mentions several
solutions for the synchronization of the PCE TED. A first solution is
that the PCE will participate in the IGP of the different ASes
involved in the interdomain path. This solution is applicable in
limited environments, for example when two domains belong to the same
company, but we do not expect that it will become widely used,
notably due to the confidentiality requirements expressed in [INTER-
AS]. The second solution proposed in [PCE] is to use an out-of-band
TED synchronization. In this case, each PCE will regularly obtain
topological information from the neighboring domains by using a
mechanism or a protocol that is still to be defined.

When considering the utilization of a PCE to aid in the establishment
of interdomain LSPs, the PCE should collect interdomain in addition
to intradomain routes. A possible solution would be to co-locate a
PCE with a BGP route reflector [VERSATILERR] as a route reflector
naturally collects those interdomain routes.

```
   PE1 ---- R1 ----- R2 ======= R5 ----- R6 --- PE2
             |        |          |         |
             |        |          |         |
            R3 ----- R4 ======= R7 ----- R8 --- PE3


  <------- AS1 ------->        <------- AS2 ------- >
```
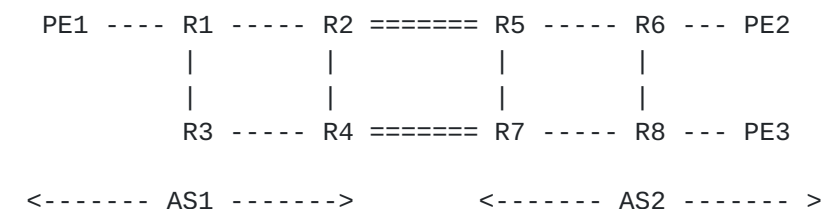Figure 1: Simple interdomain topology

However, simply using a route reflector to collect the interdomain
routes may not be sufficient to allow the PCE to have enough
visibility on the interdomain routes to establish primary and backup
interdomain LSPs. For example, let us consider the simple interdomain
topology shown in figure 1 and assume that router PE1 needs to
establish interdomain LSPs towards PE2 and PE3 and that there are two
peering links between AS1 and AS2 (R2-R5 and R4-R7).

Let us first consider the case of a full iBGP mesh inside AS1 and
that the IP addresses of PE2 and PE3 belong to the same prefix
advertised by AS2.

A common configuration in the case of backup links is to use a low
local-pref value for the routes learned via the backup link and a
normal local-pref for the routes learned via the other link. If link
R2-R5 is the primary link and link R4-R7 a backup link, R2 will
advertise the routes learned from R5 in the iBGP mesh with a high
local-pref attribute. For each destination reachable via the R4-R7
link, R4 will receive a better route via iBGP from R2. Thus, R4 will
not advertise any route learned from AS2 in the iBGP mesh and PE1
will not be aware that the R4-R7 link can be used to establish a
primary or a backup interdomain LSP.

If the same local-pref values are used for both links, a similar
problem would occur if AS2 attaches different MED values to the
interdomain routes advertised by R5 and R7. In this case, only one
route will be advertised in the iBGP full-mesh.

If PE1 needs to send VoIP packets towards PE2 and PE3, it would be
interesting to allow PE1 to use the shortest path towards PE2 and
PE3. In a pure IP network, operators might want to use the MED
attribute for this purpose. Assume that AS2 advertises two /32
addresses for PE2 and PE3. R5 advertises PE2/32 with a MED of 2 and
PE3/32 with a MED of 3 and similarly for R7. The result of this
utilization of the MED attribute will be that PE1 will only receive a
single route to reach PE2/32 (via R5) and a single route to reach
PE3/32 (via R7). With those routes, it will be difficult for PE1 to
establish both a primary and a link-disjoint backup LSP.

In large ASes, the iBGP full mesh is replaced by Confederations
[CONF] or Route Reflectors [RR]. Let us consider that R1, R2, R3 and
R4 are Route Reflectors and that PE1 is a client of R1. In this case,
R1 will only advertise its own best route towards each destination to
PE1. Thus, PE1 will never learn two distinct routes towards a
destination. If PE1 uses a PCE, for example implemented on R3, to
compute the interdomain paths, then the PCE can benefit from all the
routes that it learned via iBGP. If local-pref and MED are not used,
then the PCE will learn two routes towards PE2 and PE3. Thus, the PCE
will be able to compute disjoint paths to reach PE2 and PE3 on behalf
of PE1. Unfortunately, this solution is not sufficient if different
local-pref values are used for the peering links with AS2. In this
case, the PCE will suffer from the same problem of limited visibility
as discussed above. The same applies if the MED attribute is used.

2. **Interdomain path computation techniques**

For the purpose of illustrating the issues in interdomain constrained
LSPs computation, we consider two alternative techniques. The first
technique, CSPF, relies on the availability of the complete topology
at a PCE in the network. This technique is only applicable when the
administrators running the different ASes are willing to share
topology information. It is an ideal situation that may not occur in
practice except eventually between 2 ASes that belong to the same
company. We use this technique as a benchmark. It consists of a
centralized approach where the PCE possessing the intradomain
topology of all the ASes is responsible for the computation of
interdomain paths. The second approach is applicable in a more
general framework. It is a decentralized technique where each node on
the path of the LSP completes the path computation toward the
destination based on local routing information. We call this
technique the Distributed Path Computation (DPC) technique. This
technique is applicable for the establishment of LSPs crossing any
number of ASes.

The LSPs considered in this paper are subject to end-to-end delay
guarantees as well as link and node disjointness constraints.

In the centralized path computation technique, a Path Computation
Element (PCE) [PCE] centralizes the topology information of both ASes
and uses this information to compute the path of the inter-AS LSPs.
The PCE collects the link state packets advertised by the IGP in both
ASes and thus possesses the complete topology of the two ASes with
the TE information, if either IS-IS TE or OSPF-TE is used. For the
purpose of this draft we assume that the delays of the links are
distributed by the IGP. Based on this information, the PCE runs a
CSPF algorithm. It runs Dijkstra algorithm with costs set to the
delay of the links and sends the computed path to the source of the
LSP, if the path respects the delay constraint. For the disjoint
path computation, the PCE first prunes from the topology the links
and nodes that are on the primary path. Then, it runs the computation
as for the primary path.

The Distributed Path Computation technique relies on the routing
information distributed by BGP. Each router uses a single best BGP
route to forward IP packets toward each distant destination prefix.
These routes are stored in its Local Routing Information Base (Loc-
RIB). However, a router may receive one route toward each prefix from
each of its peers. If they pass the import filters, these routes are
stored in its Adj-RIB-Ins. We use these routes to compute our
constrained paths. As a consequence, the computed paths respect the
BGP policies of the ASes that are enforced by the import and export
filters inside the BGP routers.

In the DPC technique, the ingress ASBRs (or the head-end of the LSP)
select the egress ASBR based on the routes in their local Adj-RIB-In.
Ingress ASBRs select the route with the Next-Hop (NH) that is
reachable through a path with the smallest delay and respecting the
disjointness constraints. This consists in performing a CSPF inside
the AS toward all the NHs advertised with the destination prefix,
with the delay as metric. Once the NH is selected, the LSP is
established toward this NH using RSVP-TE with an ERO containing the
computed constrained path segment. We assume that egress ASBRs know
the delays toward directly connected eBGP peers. Egress ASBRs then,
select a NH in the neighboring AS from the NHs of the routes toward
the destination of the LSP, in the local Adj-RIB-Ins. In the
destination AS, the ingress ASBR performs a CSPF toward the tail-end
of the LSP.

We note that if a node needs to complete the path computation but
does not have routes in its Adj-RIB-Ins, with NHs that can be joined
by a path segment respecting the constraints, cranckback takes place.
A RSVP Path Error message is sent back to the source. An upstream
node on the path, the previous ASBR in our case, computes an
alternative path toward the destination, based on interdomain route
advertisement toward the PE destination prefix that have not been
tried.

[3](#). Simulation results

Our simulations were performed over the C-BGP simulator [CBGP]. We
consider interdomain topologies composed of two interconnected ASes,
the simplest case for interdomain LSPs. Each AS contains several
interconnected routers. Furthermore, the routers in each AS are
grouped in POPs as in most networks. A small POP may contain a single
router while a large POP may be composed of a few tens of routers.
The ASes are interconnected with one peering link in each city where
both ASes have a POP. To establish interdomain LSPs, we consider the
case of inter-AS VPNs where each AS may offer VPNs services toward
the POPs of the other AS. For this reason, we attach a Provider Edge
(PE) router to each POP containing more than one router. This PE
router is connected to two different routers inside the POP for
redundancy reasons. We establish a full mesh of traffic engineered
LSPs between those PE routers.

The AS topologies, with link delays and routers grouped in POPs, used
for this purpose, have been collected by the rocketfuel project
[ROCKETFUEL].

We assigned a bandwidth of 10 Gbps to each link. Moreover, each link

connecting a PE router to other routers has a delay set to 1 ms. The
same delay of 1 ms is assigned to the inter-AS links that we added to
interconnect the ASes two by two. A router in each POP is configured
as a route reflector, all the routers inside the POP are fully meshed
from an iBGP viewpoint, for optimal intra-POP routing, and the route
reflectors themselves are fully-meshed as recommended by [RR].

```
 ---------------------------------------------------------------
| Topo |      ASes      | Nodes |          Links         | LSPs |
 ---------------------------------------------------------------
|      | ASN1 | ASN2 |         | intra | inter | total |      |
 ---------------------------------------------------------------
|topo0 | 3257 | 3967 |   281 |   557 |     3 |   560 |  828 |
|topo3 | 1755 | 3257 |   291 |   575 |    14 |   589 |  920 |
|topo7 | 1239 | 6461 |   495 |  1428 |     8 |  1436 |  682 |
 ---------------------------------------------------------------
```
Table 1: Properties of the topologies

To illustrate the techniques described earlier, we compute primary
and backup paths with a 100ms delay constraint, with or without
100Mbps bandwidth reservations. That is, for each primary path, we
compute an end-to-end link and node disjoint path with the same
constraints as for the primary path, for protection purposes. The
existence of backup paths is used as an indication of the diversity
of the paths available to the centralized and the distributed
techniques.

```
 --------------------------
| Topo | primary | backup |
|--------------------------
|topo0 | 100     | 100    |
|topo3 | 100     | 100    |
|topo7 | 100     | 100    |
 --------------------------
```
Table 2: Percentage of interdomain LSPs established when a centralised
PCE computes the paths for the interdomain LSPs

Table 2 shows that, as expected, when a centralised PCE with complete
knowledge of the IGP routing tables of both ASes is used, then all
interdomain LSPs, both primary and link-disjoint backup LSPs can be
established. However, such a PCE design does not meet the
confidentiality requirements of [INTER-AS].

```
 --------------------------
| Topo | primary | backup |
|--------------------------
|topo0 | 100     | 0      |
```

```
|topo3 | 100     | 5      |
|topo7 | 100     | 3      |
 --------------------------
```
Table 3: Percentage of interdomain LSPs established when interdomain
paths are computed by head-end LSRs

Table 3 shows that when no PCE is used, the head-end LSRs have
difficulties in finding the appropriate paths for the interdomain
LSPs. The main reason why the head-end LSR is unable to establish
the backup interdomain LSPs is that it did not receive enough
interdomain routes from its route reflector. Table 4, shows that the
utilization of a PCE co-located with a Route Reflector significantly
improves the percentage of backup LSPs that are established in the
considered topologies. The main reason why it is not possible to
establish all backup LSPs is that the Route Reflectors already
summarize interdomain routes and not all interdomain routes appear in
the iBGP mesh between the route reflectors.

```
 --------------------------
| Topo | primary | backup |
|--------------------------
|topo0 | 100     | 64     |
|topo3 | 100     | 78     |
|topo7 | 100     | 74     |
 --------------------------
```
Table 4: Percentage of interdomain LSPs established when interdomain
paths are computed by a PCE colocated with a Route Reflector

4. Conclusion

PCE can play an important role to aid head-end LSRs in the
establishment of interdomain LSPs. However, to meet the
confidentiality requirements of [INTER-AS], it must be noted that the
performance of the PCE will depend on the interdomain routes that it
receives. One solution to ensure that a PCE receives enough
interdomain routes would be to colocate it with a Route Reflector.
However, even in this case, the PCE may not know all the available
routes and may have difficulties to find appropriate interdomain
paths.

The development of the Path Computation Elements should take into
account the need to collect enough interdomain routes on the PCE. A
possible approach would be to allow border routers to advertise
multiple paths (their best and possibly several non-best paths) to
the PCE, at least for the interdomain routes towards taild-end LSRs.

BGP extensions allowing to advertise multiple BGP routes have already
been proposed earlier [MULTIPLE].

5. Security Considerations

The security consideration of [PCE] are applicable for this document.

References

[INTER-AS] Zhang, R., Vasseur, JP., et. al., "MPLS Inter-AS Traffic
Engineering requirements", draft-ietf-tewg-interas-mpls-te-
req-06.txt, January 2004 (work in progress).

[VERSATILE-RR] O. Bonaventure, S. Uhlig, and B. Quoitin. The case for
more versatile BGP Route Reflectors, July 2004. Work in progress,
draft-bonaventure-bgp-route-reflectors-00.txt.

[PCE] A. Farrel, J.P. Vasseur, J. Ash, Path Computation Element (PCE)
Architecture, Internet draft, draft-ash-pce-architecture-00.txt, work
in progress, September 2004

[ROCKETFUEL] R. Mahajan, N. Spring, D. Wetherall, T. Anderson,
Inferring link weights using end-to-end measurements, 2nd Internet
Measurement Workshop (IMW2002), Marseille, France, Nov. 2002

[CBGP] B. Quoitin, C-BGP, an efficient BGP simulator,
http://cbgp.info.ucl.ac.be, March 2004

[CONF] P. Traina, Autonomous System confederations, RFC1965, June
1996

[RR] T. Bates, R. Chandra, E. Chen, BGP Route Reflection - an
alternative to full mesh iBGP, April 2000, RFC 2796

[MULTIPLE] D. Walton and D. Cook and A. Retana and J. Scudder,
Advertisement of multiple paths in BGP, Internet draft, draft-walton-
bgp-add-paths-01.txt, work in progress, Nov. 2002

[IPOM] C. Pelsser, S. Uhlig, O. Bonaventure, On the difficulty of
establishing interdomain LSPs, Proceedings IPOM'2004, Beijing China,
October 2004

Acknowledgment

TOTEM project (http://totem.info.ucl.ac.be).

Authors' Addresses

     Cristel Pelsser
     Steve Uhlig
     Olivier Bonaventure
     Dept. CSE, Universite catholique de Louvain
     EMail: {name}@info.ucl.ac.be