      Performance Analysis of Inter-Domain Path Computation Methodologies

Status of This Memo

Copyright Notice

Abstract

   This document presents a performance comparison between the per-
   domain path computation method and the Path Computation Element (PCE)
   Architecture-based Backward Recursive Path Computation (BRPC)
   procedure. Metrics to capture the significant performance aspects
   are identified, and detailed simulations are carried out on realistic
   scenarios. A performance analysis for each of the path computation
   methods is then undertaken.

Table of Contents

1. Introduction

   The IETF has specified two approaches for the computation of inter-
   domain (Generalized) Multi-Protocol Label Switching ((G)MPLS) Traffic
   Engineering (TE) Label Switched Paths (LSP): the per-domain path
   computation approach defined in [RFC5152] and the PCE-based approach
   specified in [RFC4655]. More specifically, we study the PCE-based
   path computation model that makes use of the BRPC method outlined in
   [RFC5441]. In the rest of this document, we will call PD and PCE the
   per-domain path computation approach and the PCE path computation
   approach, respectively.

   In the per-domain path computation approach, each path segment within
   a domain is computed during the signaling process by each entry node
   of the domain up to the next-hop exit node of that same domain.

   In contrast, the PCE-based approach and, in particular, the BRPC
   method defined in [RFC5441] rely on the collaboration between a set
   of PCEs to find to shortest inter-domain path after the computation
   of which the corresponding TE LSP is signaled: path computation is
   undertaken using multiple PCEs in a backward recursive fashion from
   the destination domain to the source domain. The notion of a Virtual
   Shortest Path Tree (VSPT) is introduced. Each link of a VSPT
   represents the shortest path satisfying the set of required
   constraints between the border nodes of a domain and the destination
   LSR. The VSPT of each domain is returned by the corresponding PCE to
   create a new VSPT by PCEs present in other domains. [RFC5441]
   discusses the BRPC procedure in complete detail.

This document presents some simulation results and analysis to
compare the performance of the above two inter-domain path
computation approaches. Two realistic topologies with accompanying
traffic matrices are used to undertake the simulations.

Note that although the simulation results discussed in this document
have used inter-area networks, they also apply to Inter-AS cases.

Disclaimer: although simulations have been made on different and
realistic topologies showing consistent results, the metrics shown
below may vary with the network topology.

Note that this document refers to multiple figures that are only
available in the PDF version.

2. Terminology

Terminology used in this document:

TE LSP: Traffic Engineered Label Switched Path.

CSPF: Constrained Shortest Path First.

PCE: Path Computation Element.

BRPC: Backward Recursive PCE-based Computation.

AS: Autonomous System.

ABR: Routers used to connect two IGP areas (areas in OSPF or levels
in IS-IS).

ASBR: Routers used to connect together ASes of a different or the
same Service Provider via one or more Inter-AS links.

Border LSR: A border LSR is either an ABR in the context of inter-
area TE or an ASBR in the context of Inter-AS TE.

VSPT: Virtual Shortest Path Tree.

LSA: Link State Advertisement.

LSR: Label Switching Router.

IGP: Interior Gateway Protocol.

TED: Traffic Engineering Database.

PD: Per-Domain

3. Evaluation Metrics

   This section discusses the metrics that are used to quantify and
   compare the performance of the two approaches.

   o Path Cost. The maximum and average path costs are observed for
      each TE LSP. The distributions for the maximum and average path
      costs are then compared for the two path computation approaches.

   o Signaling Failures. Signaling failures may occur in various
      circumstances. With PD, the head-end LSR chooses the downstream
      border router (ABR, ASBR) according to some selection criteria
      (IGP shortest path, ...) based on the information in its TED.
      This ABR then selects the next ABR using its TED, continuing the
      process till the destination is reached. At each step, the TED
      information could be out of date, potentially resulting in a
      signaling failure during setup. In the BRPC procedure, the PCEs
      are the ABRs that cooperate to form the VSPT based on the
      information in their respective TEDs. As in the case of the PD
      approach, information in the TED could be out of date, potentially
      resulting in signaling failures during setup. Also, only with the
      PD approach, another situation that leads to a signaling failure
      is when the selected exit ABR does not have any path obeying the
      set of constraints toward a downstream exit node or the TE LSP
      destination. This situation does not occur with the BRPC. The
      signaling failure metric captures the total number of signaling
      failures that occur during initial setup and re-route (on link
      failure) of a TE LSP. The distribution of the number of signaling
      failures encountered for all TE LSPs is then compared for the PD
      and BRPC methods.

   o Crankback Signaling. In this document, we made the assumption
      that in the case of PD, when an entry border node fails to find a
      route in the corresponding domain, boundary re-routing crankback
      [RFC4920] signaling was used. A crankback signaling message
      propagates to the entry border node of the domain and a new exit
      border node is chosen. After this, path computation takes place
      to find a path segment to a new entry border node of the next
      domain. This causes an additional delay in setup time. This
      metric captures the distribution of the number of crankback
      signals and the corresponding delay in setup time for a TE LSP
      when using PD. The total delay arising from the crankback
      signaling is proportional to the costs of the links over which the
      signal travels, i.e., the path that is setup from the entry border
      node of a domain to its exit border node (the assumption was made

   that link metrics reflect propagation delays). Similar to the
   above metrics, the distribution of total crankback signaling and
   corresponding proportional delay across all TE LSPs is compared.

   o TE LSPs/Bandwidth Setup Capacity. Due to the different path
      computation techniques, there is a significant difference in the
      amount of TE LSPs/bandwidth that can be set up. This metric
      captures the difference in the number of TE LSPs and corresponding
      bandwidth that can be set up using the two path computation
      techniques. The traffic matrix is continuously scaled and stopped
      when the first TE LSP cannot be set up for both the methods. The
      difference in the scaling factor gives the extra bandwidth that
      can be set up using the corresponding path computation technique.

   o Failed TE LSPs/Bandwidth on Link Failure. Link failures are
      induced in the network during the course of the simulations
      conducted. This metric captures the number of TE LSPs and the
      corresponding bandwidth that failed to find a route when one or
      more links lying on its path failed.

4. Simulation Setup

   A very detailed simulator has been developed to replicate a real-life
   network scenario accurately. Following is the set of entities used
   in the simulation with a brief description of their behavior.

Domain Name	# of nodes	# of links	OC48 links	OC192 links	[0,20) Mbps	[20,100] Mbps
D1	17	24	18	6	125	368
D2	14	17	12	5	76	186
D3	19	26	20	6	14	20
D4	9	12	9	3	7	18
MESH-CORE (backbone)	83	167	132	35	0	0
SYM-CORE (backbone)	29	377	26	11	0	0

            Table 1. Domain Details and TE LSP Size Distribution

   o Topology Description. To obtain meaningful results applicable to
      present-day Service Provider topologies, simulations have been run
      on two representative topologies. They consists of a large
      backbone area to which four smaller areas are connected. For the
      first topology named MESH-CORE, a densely connected backbone was
      obtained from RocketFuel [ROCKETFUEL]. The second topology has a

symmetrical backbone and is called SYM-CORE. The four connected
smaller areas are obtained from [DEF-DES]. Details of the
topologies are shown in Table 1 along with their layout in Figure
1. All TE LSPs set up on this network have their source and
destinations in different areas and all of them need to traverse
the backbone network. Table 1 also shows the number of TE LSPs
that have their sources in the corresponding areas along with
their size distribution.

o Node Behavior. Every node in the topology represents a router
   that maintains states for all the TE LSPs passing through it.
   Each node in a domain is a source for TE LSPs to all the other
   nodes in the other domains. As in a real-life scenario, where
   routers boot up at random points in time, the nodes in the
   topologies also start sending traffic on the TE LSPs originating
   from them at a random start time (to take into account the
   different boot-up times). All nodes are up within an hour of the
   start of simulation. All nodes maintain a TED that is updated
   using LSA updates as outlined in [RFC3630]. The flooding scope of
   the Traffic Engineering IGP updates are restricted only to the
   domain in which they originate in compliance with [RFC3630] and
   [RFC5305].

o TE LSP Setup. When a node boots up, it sets up all TE LSPs that
   originate from it in descending order of size. The network is
   dimensioned such that all TE LSPs can find a path. Once set up,
   all TE LSPs stay in the network for the complete duration of the
   simulation unless they fail due to a link failure. Even though
   the TE LSPs are set up in descending order of size from a head-end
   router, from the network perspective, TE LSPs are set up in random
   fashion as the routers boot up at random times.

o Inducing Failures. For thorough performance analysis and
   comparison, link failures are induced in all the areas. Each link
   in a domain can fail independently with a mean failure time of 24
   hours and be restored with a mean restore time of 15 minutes.
   Both inter-failure and inter-restore times are uniformly
   distributed. No attempt to re-optimize the path of a TE LSP is
   made when a link is restored. The links that join two domains
   never fail. This step has been taken to concentrate only on how
   link failures within domains affect the performance.

5. Results and Analysis

   Simulations were carried out on the two topologies previously
   described. The results are presented and discussed in this section.
   All figures are from the PDF version of this document. In the
   figures, "PD-Setup" and "PCE-Setup" represent results corresponding

to the initial setting up of TE LSPs on an empty network using the
per-domain and the PCE approach, respectively. Similarly, "PD-
Failure" and "PCE-Failure" denote the results under the link failure
scenario. A period of one week was simulated and results were
collected after the transient period. Figure 2 and Figure 3
illustrate the behavior of the metrics for topologies MESH-CORE and
SYM-CORE, respectively.

5.1. Path Cost

Figures 2a and 3a show the distribution of the average path cost of
the TE LSPs for MESH-CORE and SYM-CORE, respectively. During the
initial setup, roughly 40% of TE LSPs for MESH-CORE and 70% of TE
LSPs for SYM-CORE have path costs greater with PD (PD-Setup) than
with the PCE approach (PCE-Setup). This is due to the ability of the
BRPC procedure to select the inter-domain shortest constrained paths
that satisfy the constraints. Since the per-domain approach to path
computation is undertaken in stages where every entry border router
to a domain computes the path in the corresponding domain, the most
optimal (shortest constrained inter-domain) route is not always
found. When failures start to take place in the network, TE LSPs are
re-routed over different paths resulting in path costs that are
different from the initial costs. PD-Failure and PCE-Failure in
Figures 2a and 3a show the distribution of the average path costs
that the TE LSPs have over the duration of the simulation with link
failures occurring. Similarly, the average path costs with the PD
approach are much higher than the PCE approach when link failures
occur. Figures 2b and 3b show similar trends and present the maximum
path costs for a TE LSP for the two topologies, respectively. It can
be seen that with per-domain path computation, the maximum path costs
are larger for 30% and 100% of the TE LSPs for MESH-CORE and SYM-
CORE, respectively.

5.2. Crankback/Setup Delay

Due to crankbacks that take place in the per-domain approach of path
computation, TE LSP setup time is significantly increased. This
could lead to Quality-of-Service (QoS) requirements not being met,
especially during failures when re-routing needs to be quick in order
to keep traffic disruption to a minimum (for example in the absence
of local repair mechanisms such as defined in [RFC4090]). Since
crankbacks do not take place during path computation with a PCE,
setup delays are significantly reduced. Figures 2c and 3c show the
distributions of the number of crankbacks that took place during the
setup of the corresponding TE LSPs for MESH-CORE and SYM-CORE,
respectively. It can be seen that all crankbacks occurred when
failures were taking place in the networks. Figures 2d and 3d
illustrate the "proportional" setup delays experienced by the TE LSPs

due to crankbacks for the two topologies. It can be observed that
for a large proportion of the TE LSPs, the setup delays arising out
of crankbacks are very large, possibly proving to be very detrimental
to QoS requirements. The large delays arise out of the crankback
signaling that needs to propagate back and forth from the exit border
router of a domain to its entry border router. More crankbacks occur
for SYM-CORE as compared to MESH-CORE as it is a very "restricted"
and "constrained" network in terms of connectivity. This causes a
lack of routes and often several cycles of crankback signaling are
required to find a constrained path.

5.3. Signaling Failures

   As discussed in the previous sections, signaling failures occur
   either due to an outdated TED or when a path cannot be found from the
   selected entry border router. Figures 2e and 3e show the
   distribution of the total number of signaling failures experienced by
   the TE LSPs during setup. About 38% and 55% of TE LSPs for MESH-CORE
   and SYM-CORE, respectively, experience a signaling failures with per-
   domain path computation when link failures take place in the network.
   In contrast, only about 3% of the TE LSPs experience signaling
   failures with the PCE method. It should be noted that the signaling
   failures experienced with the PCE correspond only to the TEDs being
   out of date.

5.4. Failed TE-LSPs/Bandwidth on Link Failures

   Figures 2f and 3f show the number of TE LSPs and the associated
   required bandwidth that fail to find a route when link failures are
   taking place in the topologies. For MESH-CORE, with the per-domain
   approach, 395 TE LSPs failed to find a path corresponding to 1612
   Mbps of bandwidth. For PCE, this number is lesser at 374
   corresponding to 1546 Mbps of bandwidth. For SYM-CORE, with the per-
   domain approach, 434 TE LSPs fail to find a route corresponding to
   1893 Mbps of bandwidth. With the PCE approach, only 192 TE LSPs fail
   to find a route, corresponding to 895 Mbps of bandwidth. It is
   clearly visible that the PCE allows more TE LSPs to find a route thus
   leading to better performance during link failures.

5.5. TE LSP/Bandwidth Setup Capacity

   Since PCE and the per-domain path computation approach differ in how
   path computation takes place, more bandwidth can be set up with PCE.
   This is primarily due to the way in which BRPC functions. To observe
   the extra bandwidth that can fit into the network, the traffic matrix
   was scaled. Scaling was stopped when the first TE LSP failed to set
   up with PCE. This metric, like all the others discussed above, is
   topology dependent (therefore, the choice of two topologies for this

study). This metric highlights the ability of PCE to fit more
bandwidth in the network. For MESH-CORE, on scaling, 1556 Mbps more
could be set up with PCE. In comparison, for SYM-CORE, this value is
986 Mbps. The amount of extra bandwidth that can be set up on SYM-
CORE is lesser due to its restricted nature and limited capacity.

6. Security Considerations

This document does not raise any security issues.

7. Acknowledgment

The authors would like to acknowledge Dimitri Papadimitriou for his
helpful comments to clarify the text.

8. Informative References

   [DEF-DES]   J. Guichard, F. Le Faucheur, and J.-P. Vasseur,
                "Definitive MPLS Network Designs", Cisco Press, 2005.

   [RFC5152]   Vasseur, JP., Ed., Ayyangar, A., Ed., and R. Zhang, "A
                Per-Domain Path Computation Method for Establishing
                Inter-Domain Traffic Engineering (TE) Label Switched
                Paths (LSPs)", RFC 5152, February 2008.

   [RFC5441]   Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A
                Backward Recursive PCE-Based Computation (BRPC)
                Procedure to Compute Shortest Constrained Inter-Domain
                Traffic Engineering Label Switched Paths", RFC 5441,
                April 2009.

   [RFC3630]   Katz, D., Kompella, K., and D. Yeung, "Traffic
                Engineering (TE) Extensions to OSPF Version 2", RFC
                3630, September 2003.

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

   [RFC4090]   Pan, P., Ed., Swallow, G., Ed., and A. Atlas, Ed., "Fast
                Reroute Extensions to RSVP-TE for LSP Tunnels", RFC
                4090, May 2005.

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

   [RFC4920]   Farrel, A., Ed., Satyanarayana, A., Iwata, A., Fujita,
                  N., and G. Ash, "Crankback Signaling Extensions for MPLS
                  and GMPLS RSVP-TE", RFC 4920, July 2007.

   [ROCKETFUEL] N. Spring, R. Mahajan, and D. Wehterall, "Measuring ISP
                  Topologies with Rocketfuel", Proceedings of ACM SIGCOMM,
                  2002.

Authors' Addresses

   Sukrit Dasgupta
   Drexel University
   Dept of ECE, 3141 Chestnut Street
   Philadelphia, PA 19104
   USA

   Phone: 215-895-1862
   EMail: sukrit@ece.drexel.edu
   URI: www.pages.drexel.edu/~sd88

   Jaudelice C. de Oliveira
   Drexel University
   Dept. of ECE, 3141 Chestnut Street
   Philadelphia, PA 19104
   USA

   Phone: 215-895-2248
   EMail: jau@ece.drexel.edu
   URI: www.ece.drexel.edu/faculty/deoliveira

   JP Vasseur
   Cisco Systems
   1414 Massachussetts Avenue
   Boxborough, MA 01719
   USA

   EMail: jpv@cisco.com