Multi-domain G/MPLS recovery paths using PCE

Ricardo Romeral[1], Marcelo Yannuzzi[2], David Larrabeiti[1], Xavier Masip-Bruin[2], Manuel Urueña[1]

[1] Dpt. of Telematics, University Carlos III of Madrid, Leganes, Madrid, Spain
Tel: +34 91 624 8794, Fax: +34 91 624 8749,
E-mail: {rromeral,dlarra,muruenya}@uc3m.es
[2] Dpt. Of Computer Architecture, Technical University of Catalonia, Barcelona, Spain
E-mail: {yannuzzi, xmasip}@ac.upc.edu

The Path Computation Element WG was chartered at IETF in the beginning of 2005. The purpose of this working group is the computation of paths for MPLS-based traffic engineering across Autonomous Systems in the Internet. This paper provides a short overview of the purpose of this architecture and explores some of its possibilities in the provision of inter-domain exchange of traffic at G/MPLS level in a realistic application scenario. The focus is set on the automatic set up of primary and backup paths spanning multiple domains.

1. Introduction

Today, G/MPLS is being deployed as an intra-domain traffic engineering tool giving the operator the flexibility and performance of a connection-oriented technology seamlessly integrated with IP. However, the potential of G/MPLS across domains in the Internet context is almost unexplored partly due to the way Internet traffic exchange is conceived and architectural constraints of the current routing protocol driving the exchange of traffic among domains: the Border Gateway Protocol (BGP). There are a few initiatives trying to break the barriers for inter-domain MPLS and hence for optical exchange across domains, while keeping the requirements to extend MPLS to the interdomain context recently pointed out by Service Providers [1]. The Path Computation Element (PCE) is one of such enterprises. The PCE WG was chartered at IETF in the beginning of 2005. The purpose of this working group is the computation of paths for G/MPLS traffic engineering across a bundle of Autonomous Systems in the Internet. This will imply a number of extensions to current IGPs and to BGP, together with the definition of the entity and protocols required to deliver path computation information. This paper provides a brief overview of this architecture and explores some of its possibilities in the applications identified as realistic in this context: fast-recovery and constraint-based routing.

2. BGP Limitations to Inter-domain MPLS end-to-end disjoint paths

Instead of advertising networks in terms of a destination and the distance to that destination, BGP routers advertise networks as: destination addresses and AS (Autonomous System) path descriptions to reach those destinations. This means that BGP can be classified as a path-vector routing protocol. Since BGP hides most

intradomain routing information, BGP cannot guarantee that it will select the fastest, shortest route to a destination. In general, BGP just aims to minimize the number of traversed ASes, irrespective of parameters such as: the number of internal hops taken within each AS, end-to-end delay or traversed link capacities. Policies can be used to influence route selection to some extent, but the capabilities of these policies are constrained by the limited amount of information (AS path) supplied by BGP peers. However, the most important limitation of BGP to enable global inter-domain Traffic Engineering (TE) is the fact that each AS advertises to its border routers and, through them, to their neighbour ASes, only the route considered the best for a given destination. This means that a source AS has no means to acquire information and compute alternative paths to the destination where to check a set of constraints. Hence, the TE facilities of BGP are only local [2] (e.g. the Multi Exit Discriminator permits to select the preferred ingress router for a destination and the Local_preference attribute permits to select the egress router from several choices).

Finally, the current Internet routing paradigm entirely obscures the availability of intradomain resources (e.g. available bandwidth) in the computation of end-to-end paths across multiple domains. Certainly, this limitation and the lack of multipath routing have a deep impact on the end-to-end QoSR possibilities of the paradigm. The PCE model aims to provide a solid way to cope with both limitations at the same time.

3. PCE Architecture

Constraint-based path computation for MPLS or GMPLS traffic engineering in large multi-domain networks could be a CPU-intensive and highly complex process that might require the cooperation among several entities. The IETF has chartered a new Routing Area Working Group called “Path Computation Element (pce)” in order to specify a distributed architecture and its related protocols to solve this problem.

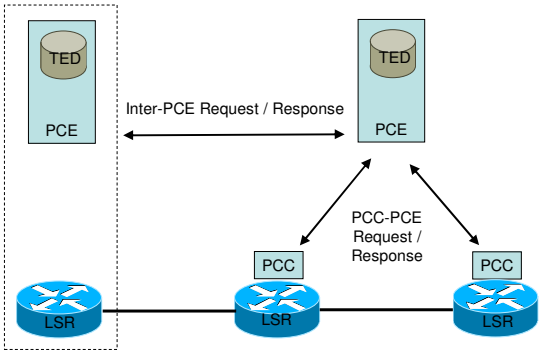

Figure 1: PCE Architecture and PCC-PCE/Inter-PCE Communication.

Essentially, the PCE framework (Figure 1) allows a Path Computation Client (PCC) to request a Path Computation Element (PCE) to calculate one or more network paths between the specified source and destination, which satisfies a set of constraints such as QoS parameters, the required number of disjoint paths, etc. When the requested

paths are computed by the PCEs and returned to the PCCs (e.g. MPLS LSRs), the appropriate Traffic Engineering LSP can be set up employing existing signaling mechanisms such as RSVP-TE.

Computed paths could be either explicit PCE paths that list all the intermediate hops, or strict/loose ones that mix specific and abstract hop identifiers, like a router address, or an Autonomous System (AS) number when no detailed topology information is available for confidentiality reasons [1]. The PCE could perform these path computations based on the network graph and the Traffic Engineering Database (TED). The TED could be built by running an IGP with Traffic Engineering extensions, like OSPF-TE or ISIS-TE, or out-of-band via configuration commands. The TED may also include additional information as LSP routes or traffic statistics.

Currently the PCE architecture is still in an early design stage [3]. Many design alternatives are open. For example, inside a domain (e.g. an IGP area or an AS) the PCE performing the path computation could be at the head-end LSR (composite PCE node), centralized in a remote dedicated server, or even distributed among several PCEs.

Moreover, in the inter-domain case, the global path could be computed using a global “all-seeing” PCE that is aware of the overall topology or, most probably, by the cooperation of the PCEs within each domain, as it is a more scalable solution and enables a Service Provider to hide its internal network topology. However, the distributed PCE model requires additional mechanisms to discover available PCEs and its capabilities, and to reliable synchronize stateful PCEs to cooperatively perform traffic load-balancing among LSPs or build backup paths.

This PCE model allows two types of communication: PCC-PCE when the PCE is not co-located with the head-end LSR and PCE-PCE for distributed path computation. Therefore a new protocol should be specified by the PCE Working Group, although currently only the requirement of a single client-server protocol that covers both types of communication is being defined [4].

In principle, the availability of the PCE service would enable any TE interdomain operation involving several ASes. However, non-technical issues such as confidentiality and administrative hurdles may prevent the deployment of all these TE facilities at full extent. In the following section we propose a practical fast-reroute protection mechanism that may be a first step in the progressive deployment of multidomain TE in Internet.

4. Inter-domain Recovery based on PCE

From the two possibilities offered by the PCE architecture [3] the most suitable for the set up of LSPs spanning multiple domains seems to be *centralized* within domains, and *distributed* (PCE-PCE) on a global basis. If an AS is very large, the domain can be split into regions with communicating PCEs delivering path computation within their regions. In the multi-domain context, it is simpler to establish a PCE per AS in order to make interdomain path computation information consistent, facilitate path computation peering agreements and security configuration. In this model, PCEs just need to interact with adjacent PCEs.

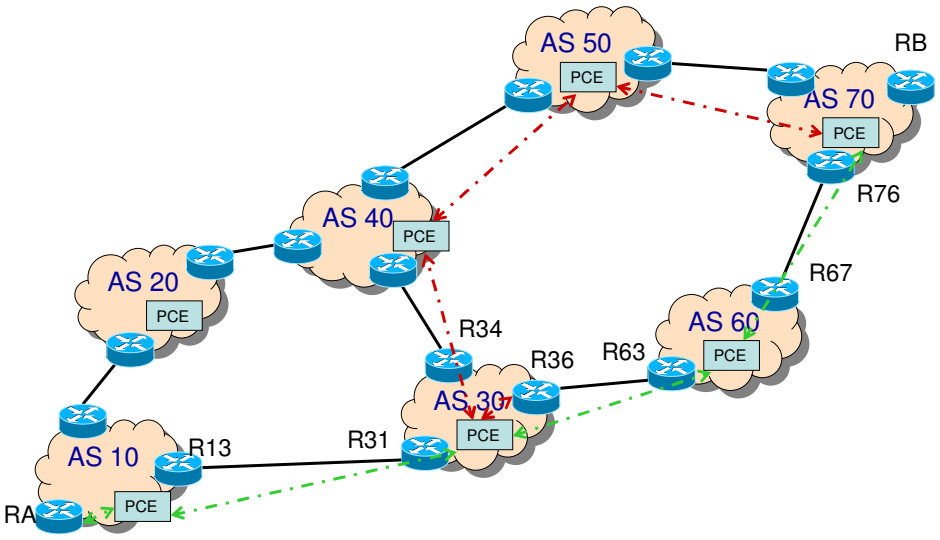

Figure 2: Inter-AS reroute example.

Even though the specification of PCE is still in the requirements phase, we shall try to sketch a possible application scenario. Figure 2 shows a set of ASes that have established an agreement to provide mutual protection in such a way that traversal of any AS is protected by a disjoint AS path. For example, an LSP AS30-AS40-AS50-AS70 might be used to protect the traffic from failures in transit traffic within AS60. In this case the information required to establish such a backup path (at least this information for RSVP-TE is the specification of the AS path) is not available from BGP routing information because BGP in AS30 decides that the route received via R36 is the best one to AS70 and the route over AS path AS40-AS50-AS70 is filtered out. If a PCE is deployed in each AS, alternative paths information would be available at the TED (Traffic Engineering Database), included the target backup path, and after a query to the PCE, R36 can issue a RSVP-TE LSP setup request (by means of a PATH message with an Explicit Route object -typically indicating a succession of AS numbers- and a Label Request object).

In this scenario, if an interdomain link group or a full domain fails, BGP would slowly recover full connectivity [5]. The availability of the protection LSP makes it possible to fast reroute all IP packets and LSPs that transit through the failing domain over this LSP, irrespectively of the stability of IP routing tables.

Another more complex alternative left open by the current conception of PCE is a more explicit hop by hop path computation including traversed nodes inside ASes. Moreover, PCEs can be used as path brokers enabling the application of any admission policy, and PCEs can relay queries to all involved ASes. In this case, it is also possible to look for a backup path dynamically upon failure because the path computation request would travel hop by hop, in a similar way to routing on-demand protocols work (e.g. RFC3561). As well known, this approach is slower than the pre-configured backup solution described above.

In the example of Figure 2, R36 would query PCE_{30} for a backup path to AS70 that excludes AS60. Firstly, PCE_{30} determines that the optimum alternative path would be AS30-AS40-AS50-AS70, identifies an internal path to R34 and decides to query PCE_{40} for an LSP suitable to temporarily carry the amount of traffic currently traversing AS60 to AS70. PCE_{40} would then compute a suitable internal path toward the next-hop AS50 and would forward the request to PCE_{50}, etc to PCE_{70}. The response would follow the reverse path and would provide PCE_{30} with a strict path R36-R76.

Usually, domain administrators are reluctant to disclose internal routing information to other parties [1]; that is why loose path specifications indicating border routers IP addresses may be more frequent. However, as stated in [3], this may not guarantee that a suitable (shortest path or with available resources) will be found by the subsequent LSP set up request. In this case we propose to include an opaque token with a given expiry time associated to the internal path that must be delivered by the PCE to the edge router, and that must be mapped to a strict Explicit Route (ER) object. For this purpose the RSVP-TE ER object should be enhanced to specify combined sequences of tokens and IP addresses.

Notice that this sort of solution does not scale to the full Internet. As a matter of fact, in general, global end-to-end inter-AS LSPs -in other words, fully optical end to end circuits- do not scale, even with statistical merging of lambdas provided by optical burst switching multiplexing [6].

5. Getting Further: Optical QoS Routing with PCE

One of the main components of a TE system is the ability to compute and solve the problem of finding primary and backup paths, wherein each of these paths simultaneously satisfies a set of independent QoS constraints. This Multi-Constrained Path (MCP) problem is an extremely complex problem (typically NP-hard) and is the focus of what is known as Quality of Service Routing (QoSR).

In the last few years, QoSR has been recognized as a strong need both at the intradomain and the interdomain level [7], and it is certainly expected that this need will also be present in the future optical-based Internet. The PCE represents a highly appealing and flexible approach to address the issue of QoSR within the context of GMPLS optical networks for a number of reasons. In the first place, because it allows to entirely decouple the complexity and CPU demanding operations of solving the MCP problem from optical LSRs or PCCs. A PCE will gather information about the current state of some QoS metrics from the TED [3], and based on this it will typically find a sub-optimal path (or part of an end-to-end path if loose hop computation between PCEs is in use) by means of a set of heuristics especially designed to tackle the MCP problem in polynomial time.

Among the QoS information that any QoSR-capable device may need to gather from the network are quite diverse metrics like the blocking probability, the restoration capabilities or level of robustness, as well as physical parameters such as the Bit Error Rate (BER) and attenuation, just to name a few. Whereas some of these metrics may be used during the computation of an interdomain segment of an optical path, others (not necessarily disjoint) could be used while performing computations inside a domain. This is mainly because the aggregation of QoS information derived by non-disclosing policies among distinct domains will typically yield a reduced set of QoS metrics to be exchanged at the interdomain level (i.e. domains will only have a partial view of overall QoS state). Given that it is quite likely that the future GMPLS optical Internet will be

deployed based on these premises, the second reason supporting the utilization of PCEs is that they provide a suitable and viable model for aggregating and signaling QoS information, not only among PCEs belonging to different domains, but also between PCEs within a same large transit domain.

Furthermore, end-to-end primary and backup multi-constrained optical paths may be computed using a distributed loose hop computation approach, in which each PCE along a path could be fed by and use both intradomain and aggregated interdomain QoS information to compute its corresponding portion of the path [8].

For all of the aforementioned reasons, the PCE arises as a good candidate for supporting the QoSR building block of a GMPLS-TE framework, and for empowering the end-to-end recovery capabilities of the future optical-based Internet. Despite these irrefutable advantages, issues such as handling the intricate interactions and dependencies between domains within the PCE-PCE protocol require further analysis. This is especially important given that routing between domains could be supported by potentially conflicting routing policies which may represent a strong limitation for QoSR.

6. Conclusions and Further Work

Today, most AS-AS interconnection is shielded by link protection mechanisms, usually involving two routers from each AS that have primary and backup links to the neighbouring AS. This protection can be easily achieved both at link or MPLS layer properly adapted to optical transmission technology, and complemented at layer 3 with multiple eBGP sessions. However, fast recovery from more complex failures involving both primary and protection links/LSPs or even sets of ASes, which obviously take longer for the routing protocol to solve, implies the establishment of Interdomain LSPs. This is not feasible today due to the limitations of routing information conveyed by BGP and the confidentiality requirements imposed by Service Providers.

This paper has reviewed PCE as a powerful tool to overcome these limitations that actually can enable a wide range of MPLS-based Traffic Engineering facilities across multiple domains. This is the most realistic application of label switching in the inter-domain context, since, letting alone scalability, a hard constraint towards an all-optical Internet is that label-based inter-domain communication prevents the application of IP-layer ingress filtering policies across domains.

In this context, the authors propose a practical framework for the usage of PCE-based protection LSPs to be used for multi-domain fast recovery as a temporary emergency measure. The application scenario assumes that a set of ASes establish mutual agreements to protect each other. Under normal working conditions traffic exchange occurs at IP layer at the network boundaries and fully-optical label-switched within the AS. In the case of a failure that is not solvable by local-repair mechanisms, pre-established LSPs are used until the routing protocol converges. Getting further, those LSPs must be subject to QoS constraints and admission control policies consistent with Diffserv networks. Thus, PCE design is facing a big challenge: support QoS-aware path computation with the minimum possible information as imposed by scalability and confidentiality.

Acknowledgements

This work is supported by FP6 IST e-Photon/ONE NoE and Spanish CAPITAL project (MEC, TEC2004-05622-C04-03/TCM). The position expressed in this article regarding

interdomain recovery corresponds to the authors' view on the topic, not necessarily to the referred projects viewpoints.

References

[1] R. Zhang, JP Vasseur. "MPLS Inter-AS Traffic Engineering requirements". IETF Internet Draft draft-ietf-tewg-interas-mpls-te-req-09.txt, September 2004.

[2] Quoitin, B.; Pelsser, C.; Swinnen, L.; Bonaventure, O.; Uhlig, S. Interdomain traffic engineering with BGP.; Communications Magazine, IEEE , Volume: 41 , Issue: 5 , May 2003.

[3] A. Farrel, J. P. Vasseur, J. Ash, "Path Computation Element (PCE) Architecture," Internet draft, draft-ietf-pce-architecture-00.txt, work in progress, March 2005.

[4] J. Ash and J. L. Le Roux, "PCE Communication Protocol Generic Requirements", Internet draft, draft-ietf-pce-comm-protocol-gen-reqs-00.txt, work in progress, May 2005.

[5] Changcheng Huang and Donald Messier . A Fast and Scalable Inter-Domain MPLS Protection Mechanism. Journal of Communications and Networks, Vol. 6, No. 1, March 2004.

[6] Ricardo Romeral, David Larrabeiti, Miguel Couto, Macelo Bagnulo, Alberto García, "MPLS-supported interdomain recovery in the public Internet", in Proceedings of the IV Workshop in MPLS/GMPLS networks, 21-22 April 2005, Girona, Spain.

[7] E. Crawley, R. Nair, B. Rajagopalan, H. Sandick, "A Framework for QoS-based Routing in the Internet," Internet Engineering Task Force, Request for Comments 2386, August 1998.

[8] M. Yannuzzi, S. Sánchez, X. Masip, J. Solé, J. Domingo, "A combined intra-domain and inter-domain QoS routing model for optical networks," in Proceedings of the 9th Conference on Optical Network Design and Modelling (ONDM 2005), IFIP/IEEE, Milan, Italy, February 2005.