On the Selection of Optimal Diverse AS-Paths for Inter-Domain IP/(G)MPLS Tunnel Provisioning

Stefano Secci [*#1], Jean-Louis Rougier [*2], Achille Pattavina [#2]

[*] *Départment Informatique et Réseaux, GET/ENST ParisTech*
37/39 rue Dareau, 75012 Paris, France
[1]secci@enst.fr [2]rougier@enst.fr

[#] *Dipartimento di Elettronica e Informazione, Politecnico di Milano*
Piazza Leonardo da Vinci, 32, 20133 Milano, Italia
[1]secci@elet.polimi.it [2]pattavina@elet.polimi.it

Abstract— **This article proposes an architecture and algorithms to select optimal diverse AS paths for end-to-end LSPs computation. The multi-domain architecture relies upon a service plane consisting of a service broker and an AS Selection Agent. Through the broker, every domain advertises transit metrics representing its transit policies (cost, routing policies) and potentially some Traffic Engineering (TE) information. The metrics are assumed to be *directional*, i.e. depending on the incoming and outgoing ASs. The Agent uses them to compute AS paths based on both costs and TE constraints, considering also, if needed, local policies and statistics on past transactions stored by the broker. A set of diverse AS paths can be computed, in order to proactively increase the success rate of tunnel set-up, in the case of imprecision or absence of advertised TE information (each AS path being subsequently tested), or to meet end-to-end protection requirements. If an AS path can be activated, the source router trigger the router-level inter-AS path computation along the AS path, which is accomplished by the PCE-based architecture.**

Within this framework, we formalize the inter-AS diverse route selection problem with directional metrics, and compare a breadth-first search heuristic with limited depth to the optimal approach. Simulations on realistic topologies prove that the heuristic scales with the number of diverse routes, and that it has an optimality gap under the 5% at least once every two times.

I. Introduction

Extensive researches have been devoted for many years to the definition of intra-domain Traffic Engineering (TE) tools and protocols, in order to support advanced services. Nowadays, there is clearly a requirement to extend these services beyond domain boundaries, particularly for critical inter-AS VPNs, TV transport or voice gateways interconnection.

A relevant solution to support inter-domain TE relies on Inter-AS (G)MPLS tunnels (i.e. LSPs) [1], [2]. Inter-AS Path Computation (PC) has to deal with TE visibility limitations since the domains can not diffuse complete TE information: the inter-AS PC method needs to be carefully designed. A Path Computation Element (PCE)-based method [3] can be adopted. Each domain elects at least one PCE. Given that TE information is not shared between domains for scalability and confidentiality reasons [1], a single PCE is unlikely to be able to compute a full inter-domain path, and has to collaborate with the PCEs of the other domains.

A few schemes can be adopted for inter-domain PC with cooperation between PCEs. A Backward Recursive PC (BRPC) approach [4] can be used, in which, starting from the destination, each domain computes an inverse tree of shortest (constrained) single-hop paths from ingress routers towards the destination, until the source domain is reached. Alternatively, PC performed on abstract domain routing information has also been proposed [5], similarly to certain routing techniques for ASON networks [6]. Nevertheless, this last approach could imply frequent cranckbacks that, still acceptable in intra-domain inter-area networks, would stress inter-domain tunnel signalling. Hence PCE-based PC methods seem to better meet the requirements for inter-domain PC. They do not deal, however, with the prior selection of AS-paths, but consider it given. It is worth noting that using BGP is not really appropriate in this case, because of lack of path diversity offered by this protocol [7] (or even its past proposed extensions [8]).

Inter-AS PC should be based on economical constraints and consider individual AS routing policies (whereas works on inter-domain routing tend to consider technical constraints only). The introduction of a "service plane", working on abstract representations of inter-domain relationships, seems attractive in order to capture these features. In this context, the current developments at the IP Sphere Forum (IPSF) [9] are of interest. They are modelling the functional features of a multi-domain service plane supporting, among other things, the advertisement of providers' network service capabilities. This service plane does not carry explicit routing data, but *multi-domain service data* that the current protocols do not handle. This data may include guarantees on the offered transit performance, and policies for the service. We believe that such a service plane is not necessarily meant to be extended to the whole Internet, but could be used by a limited group of neighbouring providers wishing to jointly offer inter-domain services (in the context of an alliance for instance).

Consistently to IPSF requirements, service elements could be published into a "service broker" and used by "AS Selection Agents (ASAs)" for service selection and activation. The service broker offers to the ASAa a common service repository indicating routing policies between domains offering inter-

978-1-4244-1845-9/08/$25.00 © 2008 IEEE

domain transit services, their costs, and potentially inaccurate TE information and statistics on past transactions. The ASAs use this service-layer information to compute constrained inter-AS routes, i.e., point-to-point (AS paths) or multipoint (AS trees) routes based on both cost and QoS constraints. The service plane is also responsible for managing the transactions needed for service acceptance by all the domains of an inter-AS route, and is then interfaced to the underlying PCE-based control plane for end-to-end path computation and signalling. We describe a consistent architecture involving multiple domains interested in QoS connectivity brokering, and analyse the unexplored open issue of selecting diverse *inter-AS routes* that meet both TE and economical requirements.

The manuscript is organized as follows. Sect.II describes the proposed service plane architecture. Sect.III resumes the state of the art on the subject and places our contribution. In Sect.IV we formulate the domain selection problem and in Sect.V we devise a two-step approach for its resolution. We evaluate the algorithms on realistic topologies, as exposed in Sect.VI. Sect. VII concludes the paper.

II. Architectural Framework

We consider a provider alliance linked by a common service plane. All the following assumptions meet the inter-AS (G)MPLS requirements and can be implemented within the IPSF framework [9]. The service plane is used in particular to exchange information needed to select a provider chain, to activate and maintain the service. An inter-domain tunnel request triggers, at the source domain, the computation of possible provider chains (inter-AS routes) on the basis of the cost and of the service requirements. Then, the best route accepted by all the involved providers is passed to the PCE-based control planes that compute the final (G)MPLS path. By monitoring the performance on the established tunnels, the providers can renegotiate transit policies for future requests.

Three actors are kept separated: (i) the *IP Network Provider (INP)*, an AS offering IP broadband connectivity to its customers; (ii) the *Service Provider (SP)*, a company offering QoS connectivity across different INPs; (iii) the *Service Customer (SC)*, user of a INP or a SP requiring a service with particular QoS requirements and for which a dedicated inter-domain tunnel needs to be established. In some previous models, the SP was non-existent as a stand-alone entity and coincided with the INP [10]. In other propositions, the INP was supposed to take external decisions about transit path [11]. In this paper, we assume a model where SCs, SPs and, finally, INPs cooperate in the inter-domain QoS services provisioning, and profit from this cooperation. The INP should be able to act as SP for its directly-connected customers, but it should not deny to its customers to subscribe to services of external SPs.

Two scenarios are supported. (i) First, a SC subscribes to an inter-domain service offered by its INP; the INP acts as SP for its customer, and collaborates with other INPs to set up the inter-domain tunnel. Second, a SC subscribes to an inter-domain tunnel service offered by a SP; the SP is disjoint with respect to both INP and SC. Hence a SP may build its multi-domain infrastructure acting as Virtual Network Operator.

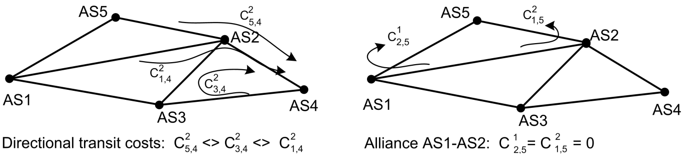

AS5
$C^2_{5,4}$
AS2
$C^2_{1,4}$
$C^2_{3,4}$
AS1
AS3
AS4
Directional transit costs: $C^2_{5,4}$ <> $C^2_{3,4}$ <> $C^2_{1,4}$
AS5
$C^2_{1,5}$
$C^1_{2,5}$
AS2
AS1
AS3
AS4
Alliance AS1-AS2: $C^1_{2,5} = C^2_{1,5} = 0$

Fig. 1. How to play with directional metrics.

A. The service plane

We assume that a provider-independent service broker is responsible of managing the inter-domain tunnel transactions. Actually, network operators would prefer to offer inter-domain QoS tunnels only if ad-hoc bilateral agreements have been signed with adjacent ASs (as assumed in [10]). But, the client would suffer the dependence of an AS from the other adjacent ones for service offerings: it would not be guaranteed that a tunnel could be established after exhaustive negotiations. The use of a shared service repository, where providers advertise their inter-domain QoS capabilities, should guarantee to the user that the achievement of connectivity ensues from competition between ASs, and that a larger path diversity is offered.

Definition 1 (AS Selection Agent): An AS Selection Agent (ASA) is a functional element responsible for inter-domain route (i.e. AS path) computation.

At the service plane, the ASAs receive the requests, query the service repository, perform the required selection, assess if the service can be instantiated. If so, the service is activated and the source ASA triggers the related router-level path computation at the PCE-based control plane.

We call *Network Service Broker (NSB)* a centralized entity that provides for ASAs: (i) the *Partial Internet Topology (PIT)*, where a node is an AS offering tunnel transit, and an arc the unidirectional logical interconnection between two nodes; (ii) the *Transit Capabilities and Costs (TCC)* of every AS-node as function of different tunnel types; (iii) statistics about past connection requests in a *Transactions DataBase (TDB)*.

In order to capture the various AS policies and business models, we introduce the following flexible cost model based on *directional policies*. We assume that the TCC contains a directional transit cost defined for the 3-uple: transit AS-node of the PIT, incoming AS and outgoing AS. In other words, the cost is defined for any transit AS, for a path coming from a given neighbour AS-node and going towards another distinct neighbour AS-node. The cost can be function of the service type and its QoS requirements (e.g. minimum bandwidth). We also assume that a transit AS may declare in the TCC inaccurate TE information corresponding to these directional links (3-uples), for instance some bounds on transit latency and bandwidth availability (potentially for each service type).

The following behaviours are thus possible. (i) An AS applies different transit costs as function of the direction of the route, based on both the arrival and departure ASs. For instance, in Fig.1, AS_2 may apply different transit costs towards AS_4 whether the route comes from AS_1 or from AS_3. Or, AS_2 may set an opportunely high transit cost from AS_5 towards AS_4, hopefully less than the cost of concurrent routes. (ii) Adjacent ASs may apply agreed transit policies, e.g. modelling free of charge peering-per-destination with a null reciprocal transit cost (i.e. policing routing is supported). For instance, in Fig.1, AS_1 and AS_2 may set null reciprocal transit costs towards AS_5. (iii) The directional transit cost may vary as function of a service type (e.g. to model protection or QoS level of the tunnel), and so an AS can deny a service type on a specific direction by setting infinite transit cost.

IPSF standards being developed would help in supporting such an architecture in a standard manner.

B. Functional Architecture

The functional flow at the service, management and network layers is the following. *1.* An ASA receives an inter-domain tunnel request (from a SC or a SP) characterized by the service type, the bandwidth, and upper bounds for the QoS additive constraints. *2.* The ASA computes the selection of an inter-AS route employing local policies and the data supplied by the NSB. *3.* The ASA sends and *instantiation message* with the selected inter-AS route and the request details, to the NSB. This message also contains an identifier called *Service Id (SID)*. *4.* The NSB forwards the message to the ASAs of the ASs along the selected route. *5.* The ASAs communicate to the NSB if they have instantiated the service, depending on their admission control policies. If yes, they can update the SLS if needed within the same message. *6.* The NSB forwards the responses to the source ASA. *7.* The source ASA decides if it will finally set up the inter-AS tunnel communicating back this information through an *activation/deactivation message*, with the SID. *8.* The NSB forwards the message. *9.* If the response is positive the ASAs transfer the request information and the SID to their local policy managers. *10.* Via the NMS the source ASA commands a tunnel to the top router, passing the SID *11.* The router queries the local PCE triggering the inter-AS path computation, e.g. using the BRPC procedure. *12.* The PCEs communicate via the PCEP protocol, which can be extended to transport the SID [12]: using it, the PCE along the AS path can filter requests querying the local policy manager. *13.* The tunnel is signaled across the inter-AS path via the inter-AS RSVP-TE protocol [2], extended to transport the SID so as to filter the inter-AS messages at the policy managers.

A major reason for using the NSB as a gateway is to store the transaction statistics into the TDB for every request and for a limited amount of time, e.g. success rate, guaranteed QoS over configured tunnels, etc., all those information useful for an ASA to prune and weight the topology conveniently before selecting a route. An ASA may employ this local information to improve the success rate of the selected routes. This paper does not specify how such optimizations may be used.

In order to increase the possibility of route acceptance, an ASA should evaluate disjoint *route alternatives*. Computing disjoint alternatives may also be desirable to meet protection requirements of some customers. The basic functional flow slightly changes: once the ASA has computed several alternatives, it orders them on the basis of their cost in a priority list, which is sent to the NSB. Then, on the strength of the route acceptance (one by one, points *4.-6.* above), the first accepted inter-AS route is chosen, and the resources of the others, if any, are released; if none is accepted, the computation is repeated considering the past rejections (stateful computation). The alternatives should have an appropriate disjointness degree in order to improve the success probability.

Accounting and Billing: The service broker has tracked all the successful transactions; likewise, every INP/SP has tracked all the transactions in which it was involved. Two cases seem possible for *provider-provider* billings: (i) a clearing house resolves paybacks and determinates the reciprocal dues, similarly to what happens for mobile phones roaming or airlines alliances; (ii) every service provider arranges directly the cash flows with the others, so that bilateral agreements hidden in the common repository may be signed. The second case seems the most promising in the short term: a SP could hide bilateral agreements by advertising official transit costs, and by billing different ones toward certain INPs.

Service types: Several service types may potentially be considered. The service type may characterize QoS flow constraints (resiliency, bit-rate parameters, premium classes, etc.) and/or a required switching capability (switching technology, point-to-point or point-to-multipoint LSP signalling features, etc). In the following, we will consider one single service type: a point-to-point tunnel with a delay upper bound and a given bit-rate. The extension of the modelling and algorithmic parameters to multi-service cases is straightforward.

III. Related work and Our Contribution

Apart the IETF activities already resumed, previous relevant works on inter-domain TE have been achieved, particularly in the context of Agave, Dragon, EuQoS and Mescal projects.

Within the NSF Dragon project [13] an experimental PCE-based framework for multi-domain provisioning of TE paths has been implemented. A distributed control plane across heterogeneous networks, with different switching technologies and granularities, has been tested, including mechanisms for authentication, authorization, accounting (AAA), and scheduling. This work seems particularly useful for grid networks where economical constraints are absent and QoS constraints are often limited to availability and survivability.

The IST Mescal and Agave projects mainly recommend a provider-centric approach for connection-less services based on a cascaded model [10]: in this framework an AS can discover transit QoS capabilities only of adjacent ASs, and only towards specific destination networks. This limits the path diversity. Within these projects, and similarly in the EuQoS project, extensions to BGP have been proposed to advertise TE information [8] [14]. Other studies propose the combination of distributed overlay architectures and BGP extensions (e.g. [15]). Such approaches require changing BGP, which is problematic, given the number of existing routers deployed currently. The exchanging of QoS information on

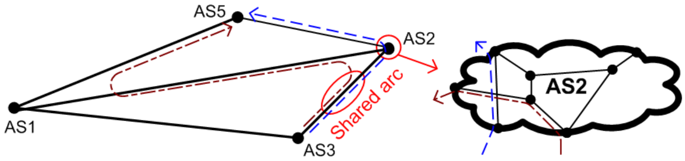


Fig. 2. Example of two diverse (directionally disjoint) inter-AS routes: different inter-domain directions imply different intra-domain resources availability

BGP is also questionable even if the proposers claim that it scales [8]. Moreover, these choices do not meet the route diversity requirement [7]: assuming a cascaded model, proposing extensions to BGP, not enough TE inter-AS routes per destination can be taken into account.

In Sect.II, instead, we propose a user-centric model based on a service plane restricted to some providers willing to collaborate for connection-oriented inter-AS services. In the rest of the paper we tackle the related problem of selecting diverse inter-AS routes in a graph with directional metrics.

IV. The Inter-AS Diverse Route Selection problem

As previously mentioned in Sect.II-B, a set of route alternatives should be selected to offer enough diversity for a successful route selection, or to set-up disjoint tunnels for protection purposes. We introduce the following definitions:

Definition 2 (Directional arc): It denotes a succession of two inter-AS logical arcs linking three AS-nodes.

Definition 3 (Diverse routes): Two inter-AS routes are diverse if they do not share any directional arc.

An inter-domain logical arc may correspond to several ingress and egress inter-AS links at every transit AS: forcing a directional disjointness, two route alternatives may involve the same AS-node, but follow different directions. As depicted in Fig.2, diverse alternatives do not share the same ingress/egress links pair at an AS. Note that only one route may be possible because of intra-domain resource availability.

We consider local directional disjointness, and not end-to-end disjointness, because it allows benefiting from the scale-free nature of the AS graph, which currently presents a few transit hubs interconnecting lots of ASs, and the most part of ASs with a few adjacencies. End-to-end disjointness at the AS-level would be, instead, very hard to achieve. When two ASs are connected with a single inter-AS link, the end-to-end disjointness may not be guaranteed: this would be the case for the most part of the AS-node in the Internet graph. Moreover, as already mentioned, imposing directional disjointness for a given transit AS it is possible to have unavailability over a directional arc and availability over another one, improving the acceptance ratio of the selected routes.

Definition 4 (Inter-AS Diverse Route Selection problem): It consists in selecting the less costly set of diverse inter-AS routes satisfying a given connection request.

An inter-domain connection request is characterized by a source AS-node, a destination AS-node, a bandwidth β and an end-to-end delay bound D. Many routes can satisfy a request. The selected routes have to be diverse in order to increase the probability that at least one of them is accepted while minimizing the signalling required at the service plane to "test" these paths. Before computing the selection of paths the PIT is pruned removing those AS-nodes and (direct and directional) arcs that are incapable of routing the request. The pruned arcs are those with an advertised delay bigger than D, those that do not dispose of enough bandwidth (if these information are available) or those in a taboo list. The pruned nodes are those remained isolated after arcs pruning, or those that are in local taboo lists. Arcs and node taboo lists can be specified based on some local policies (ASs competing in certain geographical regions may exclude each-other for some specific destinations, while accepting to collaborate for others), or after a TDB analysis that may indicate some nodes and arcs temporarily not ready for tunnel configuration. Finally, note that an ASA besides pruning the PIT may also expand it with non-published arcs, costs and constraints, applying private agreements with other domains or members of some alliance.

Over reduced PIT the route selection is computed. Let $G(N,E)$ be the given graph, where N is the set of AS-nodes and E the set of inter-AS logical connections. Using the cost model introduced in Sect.II-A, the ASA of the source AS associates to every directional arc $(i,k,j) \subset E \times E \times E$ of the reduced PIT, the directional cost function $c_{i,k,j}(\beta)$, and the directional transit delay $d_{i,k,j}$ gathered from the TCC.

We model the problem of finding a feasible diverse inter-AS routes by ILP. In the following, $x^a_{i,j}$ is a binary variable equal to 1 if the a^{th} route alternative passes over the arc (i,j), and $f^a_{i,k,j}$ is equal to 1 if the a^{th} route alternative passes over the directional arc (i,k,j) (0 otherwise).

$$\min \phi(f) = \sum_{a} \sum_{(i,k,j)} c_{i,k,j}(\beta)\, f^a_{i,k,j} \quad (1)$$

$$s.t. \sum_{j \in N} x^a_{i,j} - \sum_{j \in N} x^a_{j,i} = \begin{cases} 1 & \text{if } i = \text{origin AS} \\ -1 & \text{if } i = \text{tail AS} \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in N, \forall a \quad (2)$$

$$f^a_{i,k,j} \geq x^a_{i,k} + x^a_{k,j} - 1 \qquad \forall (i,k,j),\ \forall a \quad (3)$$

$$\sum_{(i,k,j)} f^a_{i,k,j} d_{i,k,j} \leq D,\ \forall a \quad (4)$$

$$\sum_{(i,k,j)} f^a_{i,k,j} \leq H_m \ \forall a \quad (5)$$

$$\sum_{a} f^a_{i,k,j} \leq 1 \qquad \forall (i,k,j) \quad (6)$$

$$f^a_{i,k,j} \in \{0,1\},\ x^a_{i,j} \in \{0,1\} \quad (7)$$

The objective (1) is to minimize the total route cost. (2) sets the flow conservation. (3) sets the enabling of directional arc variables. (4) enforces the delay upper bound. (5) imposes a maximum hop bound. (6) enforces the directional disjointness. (7) imposes the binary domain to variables.

V. Route Collection and Clique Selection (RECS)

The (1)-(7) formulation is a formal optimal approach to solve the Diverse Inter-AS Route Selection problem. However, with a number of variables and constraints $\sim |N|^3$ for the worst case, its optimization can not always be solved by general purpose solvers in a reasonable time. To solve it we propose a two-step approach called Route Collection and Clique Selection (RECS). Firstly, we collect some feasible routes to reach the destination. Then, we look for the least cost clique of a diverse route alternatives among the collected ones. To clarify the taxonomy used in the description of the RECS algorithm, we introduce the following definitions:

Definition 5 (Feasible route): A feasible route satisfies the QoS constraints and has a cost not too high with respect to the other routes already collected, i.e., sufficiently low to be a good candidate for the final route clique.

Definition 6 (Route clique): A route clique of a set of routes is a subset of diverse routes.

A. Route Collection

For route collection we devise an ad-hoc breadth-first search algorithm with limited depth, which begins at the root and explores all the neighbouring nodes. Then, for each nearest node, it explores its unexplored neighbour nodes, and so on, until no further improvement is reached, where an improvement is the selection of a new route for the target destination. It stops at a given number of hops from the root and during the search it prunes branches on the basis of metric bounds.

We chose a search algorithm for its ability to reach all the feasible paths. The algorithm we propose is inspired by the A*prune algorithm [21], used to solve the constrained k-shortest paths problem. Our approach differs from it in that : (i) given that the final objective is the selection of the optimal route clique, a further pruning (besides that on the additive constraints) on the basis of the route cost is performed, giving priority to least hop routes; (ii) given that there is no need to sort the candidate routes (as best-first-search approaches, such as A*prune, do when choosing the next path to expand during the graph exploration), the number k of shortest routes is not fixed and all the experienced feasible routes are collected.

Collection algorithm: The reduced PIT (after arcs and nodes pruning) is explored for route collection starting form the source. During the exploration, if the delay bound is not respected, or if the cost is bigger than a threshold cost updated so far, a path in the graph is no longer considered.

Let v be the threshold cost. It is re-calculated at each new route collection if at least F routes have been already collected, where F is a start threshold number to be chosen conveniently (we use $F{=}\sqrt[3]{|N|}$). Only afterwards the constraint on the route cost is checked; thus, the first F routes are collected without checking their cost. v is calculated as the average cost of those routes with a variance on the average cost less than the average of this variance: simply, within the first F routes, those with a very high cost with respect to the others are not taken into account. In this way, v has a decreasing trend, with a starting value not excessively high. The least hop routes are thus privileged because v is higher in the first hops. Favouring routes of few hops is a suitable approach for our specific problem, since long routes crossing several ASs may only have a small number of arcs in common with those previously selected, which tends to increase the cost. In this way we try to cut a lot of branches that would have been considered by general purpose solvers for the (1)-(7).

The pseudo-code is shown in Alg.V.1. The search starts (main body) looking for feasible routes at 2 hops, then 3, and so on. To populate a list of feasible routes we proceed with an AS graph exploration by evaluating for feasibility, at each iteration, only the routes of equal hops H, up to a given hop bound H_m. At the H^{th} hop, a feasible route is collected in the set ζ_{sel} if its tail is the destination, otherwise it is collected as feasible sub-routes in ζ_{cand} for further expanding.

Algorithm V.1: ROUTE COLLECTION(G)

```
procedure Pop($c, d, h, \pi$)
  – $f$ : counter of found routes so far
  – $a$, $d_a$, $c_a$ : next directional arc, delay and cost of $a$
  if $h = H$
    then
      if $\pi[h]$is the destination
        then
          if $(f \geq F$ and $c < v)$ or $(f < F)$
            then
              Add $\pi$to $\zeta_{sel}$
              $f \leftarrow f + 1$
              Update $v$
              if $f = F$
                then Calculate the starting $v$
      else if $(c + SPC[\pi[h]][d] < v)$ or $(f < F)$
      then Add $\pi$ to $\zeta_{cand}$
    else
      for $i \leftarrow 1$ to $N$
        do
          if AS$_i$ connected to AS$_{\pi[h]}$, and AS$_i \notin \pi$
            then
              $\pi[h+1] \Leftarrow i$
              $a \leftarrow (\pi[h-1], \pi[h], \pi[h+1])$
              if $h = 0$
                then Pop($c, d, h+1, \pi$)
                else if $d + d_a < D$
                then Pop($c + c_a, d + d_a, h+1, \pi$)

main
  $H \leftarrow 1$
  Pop($0, 0, H-1, \pi$)
  while $\zeta_{cand} \neq \emptyset$ or $H < H_m$
    do
      $H \leftarrow H + 1$
      take a $\pi \in \zeta_{cand}$
      $\zeta_{cand} = \zeta_{cand} - \{\pi\}$
      Pop($cost(\pi), delay(\pi), H-1, \pi$)
```

Definition 7 (Projected cost): The projected cost of a sub-route is given by the sum of the current sub-route cost and the cost of the shortest path from the tail toward the destination.

A simplified version of the Floyd-Warshall algorithm [22] is employed to calculate the cost of the shortest paths from any node to any node (A2ASP) (SPC matrix in Alg.V.1).

At every iteration, the sub-routes in ζ_{cand} are the starting point of the search. At every call of Pop(), c and d are the cumulative cost and delay of the route handled by the current route vector π with h hops number. At the very first iteration ζ_{cand} is empty, π has only the source node, and the delay constraint is not verified. Afterwards, the function recursively visit every neighbour of the sub-route tail node, updating π, and evaluating the route feasibility on the cumulative delay. When the H^{th} hop is reached, the route is collected in ζ_{sel} if the destination is attained, if its cost is minor than v, and if the delay bound is respected; otherwise it is added to ζ_{cand} only if the delay bound is respected and its projected cost is equal to or less than the v.

Complexity: Including the A2ASP pre-computation the time complexity of the collection algorithm is O(n^4). Without it is

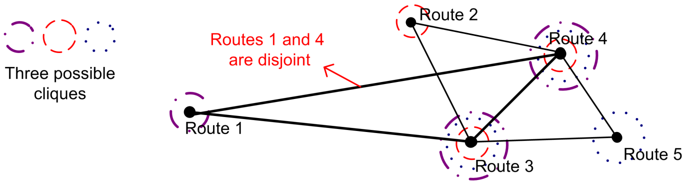


Fig. 3. Three possible cliques of 3 diverse routes in a 4-route graph

O($n^{\frac{1}{3}H_m}$) (details not included because of page limit).

B. Optimal Clique Selection

The next step consists in extracting the least cost clique of a diverse (collected) routes. Every route-element of ζ_{sel} has a cost and can be included in the final clique as by Definition 6. This problem is linked to the Generalized Minimum Clique Problem (GMCP), with a fixed clique size. The routes of ζ_{sel} are considered as vertices, which are connected only if diverse. The GMPC considers weighted vertex and links, and is NP-hard [20]. In our problem only vertex have a cost, so it becomes a node-weighted minimum clique problem, which is still less complex. The optimal clique selection sub-problem can be solved by ILP (formulation not included because of page limit). In Fig.3, e.g., we have a 5-route graph from which only 3 cliques of 3 vertices can be extracted. Every route-vertex has a cost, and the least cost clique is the solution.

VI. Performance Analysis

1) PIT building: We chose to build realistic topologies: we dumped the AS whois database containing interconnection data available at [17]. As stated before, our architecture is not meant to be used at Internet-wide scale (even the PCE-based one is not meant to be) but on a set of ASs collaborating to a common service plane. We then generate topologies with a limited number of ASs (a few hundreds to thousands), but use Internet topology estimations in order to be as realistic as possible. The selection method is described below: among all the ASs, we select only those with at least 4 adjacencies, focusing so on possible INPs with several transit directions; then, only those ASs with more than 2 adjacencies within the selection are kept in. With these parameters, the final graph has 1716 sparsely interconnected ASs.

2) Capacities and costs: For capacities and costs assignment, we classify as Tier 3 (T3) an AS with a number of interconnections less than the average, Tier 1 (T1) one with a number of interconnections with non-T3 ASs over the average, and Tier 2 (T2) the remaining ones. This deviates from the conventional terminology of T1, in that, an AS would be classified as a T1 if it is not an explicit customer of any other AS; this does not apply to our framework, since we overtake the BGP-policy-based peering and customer-provider relationships. Moreover, we prefer a degree-based instead of a betwenness-based ranking because this last could not apply since we are not aware of BGP routes.

Considering a T3 not able to offer as much connectivity as T2s and T1s do, and the same for T2s versus T1s, we assign capacities to inter-AS links normally with different averages and deviations as indicated in [19]. Moreover, since the bottleneck is not at the intra-domain but at the inter-domain links, and since lower transit costs come with higher availability, we approximate the directional transit cost equal to $K\frac{log[\beta min(C_{i,k},C_{k,j})]}{\beta min(C_{i,k},C_{k,j})}$, $K = 10^5$, for a directional arc (i,k,j) with links capacities of $C_{i,k}$ and $C_{k,j}$; it decreases more than linearly as function of the product between the requested bandwidth and the minimal inter-AS capacity. We halve the cost when the transit involves two AS of the same company, and set it to zero when the threes of them do, so as to try to be more realistic (ASs of the same ISP can be identified approximately exploiting certain whois tags [19]).

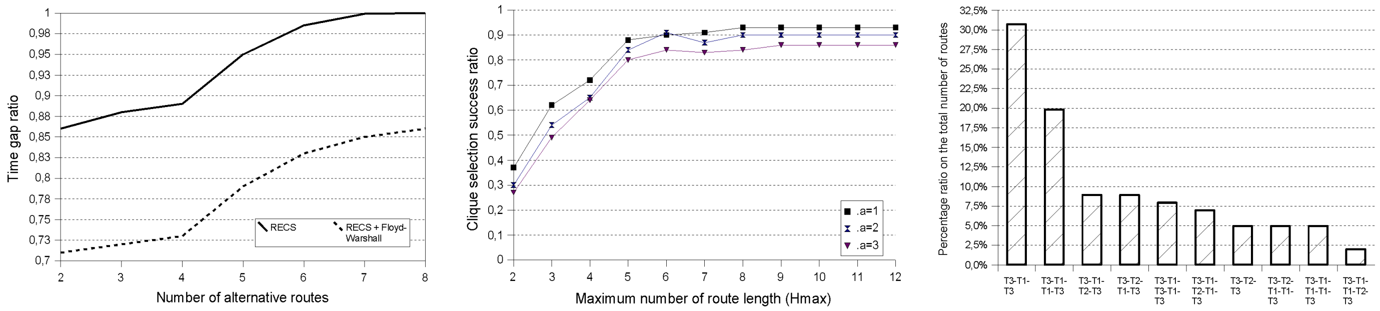
3) Delay bounds: The significant factor affecting the end-to-end delay is the propagation delay [18]. According to the whois tags, we assign ASs to a country. Since carriers can operate in more continents, we calculate the directional transit delay bounds independently of the geographical position of the transit nodes, but as a function of the position of source and destination nodes, following a normal distribution with averages and deviations chosen on the basis of experimental round trip times (see [19]).

4) Simulation results: The algorithms were implemented in C++. CPLEX was employed as ILP solver. Sources and destinations are chosen randomly. Bandwidth and delay bound are chosen randomly between 1 and 10, and between 500ms and 3s, respectively. We display three significant aspects of the simulations results for model evaluation.

Fig.4a displays the average execution time gap ratio ($1 - t_{RECS}/t_{ILP}$, where t_{RECS} and t_{ILP} are the execution times under the two approaches) as function of a. 50 successful simulations are considered. We display two curves: the dotted one consider the A2ASP computation time in t_{RECS}, while the continuous one does not. Indeed, the ASAs should compute the ASASPs off-line prior to inter-AS route request. The higher the number of alternatives is, the harder the optimal approach: the RECS approach scales with the number of alternatives. Indeed, given that the number of collected routes remains always under 1000, the clique selection requires only a few solution searches. Obviously with the A2ASP computation time we have just a shift.

Fig.4b shows the success ratio in selecting a route clique for three clique sizes ($a = 1, 2, 3$), as function of the upper hop bound, for 50 new simulations per case. We can affirm that: (i) the most part of ASs is attainable within 5 hops; (ii) the exploration of the graph for more than 8 hops is not useful; (iii) even for single-element degenerate cliques, a 100% success ratio was never reached because the bandwidth and the delay constraints limit the number of collected routes.

Fig.4c reports the 10 most selected hierarchical routes, for 100 new successful simulations with a hop bound of 8. We can affirm that: (i) the routes have as source and destination T3s, since they form the most part; (ii) more than 80% of routes count less than 5 hops; (iii) a significant part has only T1s transit nodes, while the others use at least one Tier 1. Thus, less than 0.1% of ASs (the T1s) attract the most of the



(a) Time gap ratio ILP - RECS

(b) Success ratio as function of the hop bound

(c) Hierarchical routes ratio

Fig. 4. Simulation results

TABLE I
RECS OPTIMALITY EVALUATION.

	$< 5\%$	$<50\%$	$< 100\%$
$a = 2$	80%	93%	99%
$a = 4$	75%	80%	99%
$a = 16$	69%	77%	96%

traffic. Such results prove that assuming, as we did, a carriers hierarchy where T1s dispose of more resources and can apply lower prices, the economically feasible routes are attracted by top tiers. This does not preclude, however, a lower-tier AS to attract more routes if it tunes prices efficaciously.

5) RECS optimality: We compare the average deviation of the selected clique cost using the RECS approach to that given by (1)-(7). Each entry of Table I indicates how many of the performed simulations per case produced a solution with an optimality gap within 5%, 15% or 100%. Three cases are considered for 2, 4 and 16 route alternatives in the clique, with 50 simulations per case. For each case we show how often (in percentage) RECS solutions had an optimality gap that falls in the three intervals. We can affirm that: (i) RECS gave a less than 5% optimal solution more than once every two times; (ii) it can guarantee a solution within the double of the optimal solution for practically all the requests; (iii) an increase on the number of route alternatives slightly worsens the optimality, while allowing at least 7 times on 10 the attainment of a solution within one and a half the optimal one.

VII. CONCLUSIONS

The goal of this paper was to propose solutions in order to provide inter-domain IP/(G)MPLS tunnels. We presented a service plane-oriented architecture based on current state-of-the-art that allows collaboration between domains in order to exchange inter-domain services, while preserving the required level of isolation and privacy. This architecture is based on a network service broker and on an AS selection agent, which is needed for a consistent AS paths selection. A flexible and realistic cost model is provided in order to meet various provider strategies and business policies.

We then studied the problem of selecting diverse inter-domain routes in such a context and provided an optimal approach and a two-step heuristic to solve it. We tested them on realistic topologies extracted from the Internet. We demonstrated that, because of the limited number of economically feasible routes, the heuristic can solve the diverse routing problem even on large topologies. Simulations revealed that it is computationally competitive, and that more than once every two times it can give a solution less within the 5% optimality.

We are currently studying refinements to the multi-domain routing problem for point-to-multipoint IP/(G)MPLS tunnels.

REFERENCES

[1] R. Zhang, J. Vasseur, "MPLS Inter-Autonomous System (AS) Traffic Engineering (TE) Requirements", RFC 4216, Nov. 2005.

[2] A. Farrel et al., "Inter domain MPLS and GMPLS TE - RSVP-TE extensions", draft-ietf-ccamp-inter-domain-rsvp-te-07, Sep. 07.

[3] A. Farrel, JP. Vasseur and J. Ash, "A Path Computation Element (PCE)-based architecture", RFC 4655, Aug. 2006.

[4] JP. Vasseur et al., "A Backward Recursive PCE-based Computation (BRPC) procedure to compute optimal inter-domain Traffic Engineering Label Switched Paths", draft-vasseur-pce-brpc-06.txt, Sep. 2007

[5] A. Sprintson et al., "Reliable Routing with QoS Guarantees for Multi-Domain IP/MPLS Networks", in Proc. of INFOCOM 2007

[6] G. Maier, F. Mizzotti, A. Pattavina, "Multi-domain routing techniques in ASON networks", in Proc. of ECOC 2007

[7] M. Yannuzzi et al., "On the challenges of establishing disjoint QoS IP/MPLS paths across multiple domains", Comm. Magazine 44 no.12

[8] D. Griffin et al., "Interdomain routing through QoS-class planes", Comm. Magazine 45, no.2, pp: 88-95, Feb.2007

[9] IP Sphere Framework Technical Specification, June 2007

[10] M.P. Howarth et al., "Provisioning for inter-domain quality of service: the MESCAL approach", Comm. Magazine 43, no.6, 129-137, Jun.2005

[11] D. Di Sorte, G. Reali, "Minimum price inter-domain routing algorithm", IEEE Communications Letters 6, no. 4, pp. 165-167, Apr. 2002

[12] J.-L. Le Roux et al., "Carrying a Contract Id in the PCE communication Protocol (PCEP)" draft-leroux-pce-contract-id-01.txt. Mar.2007

[13] T. Lehman et al., "DRAGON: a framework for service provisioning in heterogeneous grid networks", Comm. Magazine 44 no.3

[14] X. Masip-Bruin et al., "The EuQoS system: a solution for QoS routing in heterogeneous networks", Comm. Magazine 45, no.2 Feb.2007

[15] M. Yannuzzi et al., "A proposal for inter-domain QoS Routing based on distributed overlay entities and QBGP", in Proc. of WQoSR (2004)

[16] C. Pelsser, O. Bonaventure, "Path selection techniques to establish constrained interdomain MPLS LSPs", in Proc. of IFIP Networking 2006.

[17] The CIDR report, http://www.cidr-report.org.

[18] B. Choi et al., "Analysis of Point-To-Point Packet Delay in an Operational Network", in Proc. of IEEE INFOCOM 2004.

[19] S. Secci, JL. Rougier, "A constrained Internet graph", ENST RR http://www.tsi.enst.fr/publications/enst/article-2006-6628.pdf

[20] A. Koster et al., "The Partial Constraint Satisfaction Problem: Facets and Lifting Theorems", Operations Research Letteres 23, pp. 89-97 (1998).

[21] G Liu, KG Ramakrishnan, "A*Prune: an algorithm for finding K shortest paths subject to multiple constraints", INFOCOM 2001

[22] R. W. Floyd, "Algorithm 97: Shortest Path", Communications of the ACM 5 (6): 345, (June 1962)