PEMP: Peering Equilibrium MultiPath routing

Stefano Secci[a,b], Jean-Louis Rougier[a], Achille Pattavina[b], Fioravante Patrone[c], Guido Maier[b]
[a]Institut Telecom, Telecom ParisTech, LTCI CNRS, France. E-mail: {secci, rougier}@telecom-paristech.fr
[b]Politecnico di Milano, Italy. E-mail: {secci, pattavina, maier}@elet.polimi.it
[c]Università di Genova, Italy. E-mail: patrone@diptem.unige.it

***Abstract*—It is generally admitted that Inter-domain peering links represent nowadays the main bottleneck of the Internet, particularly because of lack of coordination between providers, which use independent and "selfish" routing policies. We are interested in identifying possible "light" coordination strategies, that would allow carriers to better control their peering links, while preserving their independence and respective interests. We propose a robust multi-path routing coordination framework for peering carriers, which relies on the MED attribute of BGP as signalling medium. Our scheme relies on a game theoretic modelling, with a non-cooperative potential game considering both routing and congestions costs. Peering Equilibrium MultiPath (PEMP) coordination policies can be implemented by selecting Pareto-superior Nash equilibria at each carrier. We compare different PEMP policies to BGP Multipath schemes by emulating a realistic peering scenario. Our results show that the routing cost can be decreased by roughly 10% with PEMP. We also show that the stability of routes can be significantly improved and that congestion can be practically avoided on the peering links[1].**

I. Introduction

Multipath routing has received interest for a long time, as it is considered as a very efficient solution providing more robustness and better load distribution on the network. Intra-domain multipath routing is commonly performed in Interior Gateway Protocol (IGP) networks, by balancing the load over Equal Cost Multiple Paths (ECMP) [1]. In the multi-domain context, multi-path routing is generally not implemented, its introduction raising important scalability and complexity issues (see eg. [2]). Multipath interdomain routing is, to our knowledge, still an open issue (and a target for future internet architectures). However, some limited solutions based on the Border Gateway Protocol (BGP) have been introduced, at least with some vendor's routers (see e.g. [3] [4]). Multipath BGP can then be used to balance load on different routes under specific conditions (detailed in the next section), in particular on several peering links between two adjacent carriers.

Nevertheless, the lack of routing collaboration among neighboring carriers causes BGP Multipath to produce unilateral routing choices that, even if potentially efficient for the upstream carrier w.r.t. load distribution, may lead to an inefficient situation for the downstream carrier. In this paper we propose a framework that allow carriers to select efficient load balancing strategies in a coordinated manner, while preserving their independence and respective interests. Our proposal is based on a game theoretical model, as a natural tool to study possible trade-offs between selfishness and cooperation. Possible coordination policies can be highlighted, from quite selfish to more cooperative ones, with different degrees of Pareto-efficiency.

We propose to re-use the Multiple Exit Discriminator (MED) attribute of BGP as the simple medium to convey coordination costs between carriers. A potential non-cooperative game that arises from load balancing based upon this data is then proposed. Pareto-efficient equilibrium solutions can be coordinatively selected by carriers. We show by simulations that this choice prevents congestion on peering links, decreases the global routing cost while increasing the route stability.

Sect. II presents the inter-carrier routing issues that we tackle. Sect. III present the ClubMED (Coordinated MED) framework for inter-domain multipath routing over peering links. We explain how load balancing shall be implemented over efficient equilibrium strategies. Sect. IV defines the Peering Equilibrium MultiPath (PEMP) routing coordination policies and discuss their possible benefits and technical implementation issues. Sect. V presents results from realistic simulations assessing the PEMP policy performance. We show how our approach can outperform BGP Multipath in terms of routing cost, route stability and peering link congestion. Sect. VI concludes the paper and discusses further work.

II. Inter-carrier routing issues

A. BGP and selfish routing

It is worth briefly reminding how the route selection is performed via BGP [5]. When multiple paths to a destination network are available, a cascade of criteria is employed to compare them. The first is the "local preference" through which local policies with neighbor Autonomous Systems (ASs), mainly guided by economic issues, can be applied: e.g., a peering link (i.e., free transit) is preferred to a transit link (transit fees). The subsequent criteria incorporate purely operational network issues to select the best route: (i) the route with a smaller AS hop count; (ii) if the routes are received by the same neighbor AS, the route with a smaller MED; (iii) the route with the closer egress point ("hot-potato" rule), using as distance metric the IGP path cost; (iv) the more recent route; (v) the AS path learned by the router with the smaller IP ("tie-breaking" rule). Considering these criteria, BGP selects the best route. This best route is then eventually advertised to its peers (if not filtered by local policies).

Two peering ASs have usually many links in several distributed locations and can thus dispose of many routes to the same network through the same AS. By default, these routes have equal local preferences and AS hop counts. Hence, the best route is chosen w.r.t. either the smaller MED or (if the MED is disabled) the smaller IGP path cost. The decision is taken minimizing the routing cost of a single peer: either the upstreaming AS's IGP path cost (hot-potato), or the downstreaming AS's weight (smaller MED). The challenge is thus the definition of methods that consider both the routing costs when taking the peering routing decision.

[1]Work funded by the European ICT FP7 Euro-NF Network of Excellence and the I-GATE project of the Institut Telecom, ICF Networks of the Future.

978-1-4244-4148-8/09/$25.00 ©2009

The Multi-Exit Discriminator (MED): The MED is a metric that an AS can attach to route advertisements toward a potential upstream AS, to suggest an entry point when many exist. In this way, the upstream AS can prefer an entry point toward the advertised network. By default, the MED is set to the corresponding intra-AS IGP path cost (from the downstream border router to the egress router). On transit links, subject to provider/customer agreements, the provider should always follow "MED-icated" routes suggesting preferred entry points because the customers pay for. This is not the case for peering settlements, and this can be considered as the main reason why the MED is often disabled between peers [6].

BGP Multipath: If the MEDs and/or the IGP path costs are equal, to avoid tie-breaking the load may be balanced on the equivalent routes. At the time being, such multipath extensions for BGP did not find consensus at the IETF, and for this reason there is no standard specification. However, some suggestions are indicated in [7]. As of our knowledge, the only implemented method carriers can use for multipath inter-domain routing is the "BGP Multipath" mode that some router vendors now provide (e.g., Juniper [3] and Cisco [4]), with some little variations on the routing decision. Therefore, BGP Multipath allows adding multiple paths to the same destination in the routing table. This does not affect the best path selection: a router still designates a single best path and advertises it to its neighbors. More precisely, BGP Multipath can be used when more than one IBGP (Internal BGP) routers have equivalent routes to a destination through many border routers, or when all of the candidates routes are learned via EBGP (External BGP). As stated in [7], other cases, with a combination of routes learned from IBGP and EBGP peers, should be avoided, as they may lead to routing loops for instance.

B. BGP route deflection

The peering routing decision with BGP thus relies on IGP routing costs. Nowadays, the interaction between IGP routing and inter-AS routing represents a major issue because IGP weights are optimized and reconfigured automatically. To react to non-transient network events, a carrier may re-optimize the IGP weights, inducing changes in the BGP routing decision, so that congestions might appear where not expected.

Many works concern BGP route deflection control methods. [8] reformulates the egress routing problem and proposes to replace the hot-potato rule with a more expressive and efficient rule. [9] presents a comprehensive yet hard IGP Weight Optimization (IGP-WO) method aware of possible hot-potato route deflections to bound them (they report that 70% of traffic can be affected in a real network). [10] presents a similar proposition relying on graph expansion tricks. However, while effective, a problem seems to persist with the latter propositions: each time the BGP routes change, the BGP-aware IGP-WO is to be triggered. The scalability may be thus a practical issue: the occurrence of IGP-WOs, normally triggered only for intra-AS issues, would drastically increase. To better assess this issue, we worked at the detection of deflections using TRACETREE radar data [11]. Preliminary results confirm that top-tier AS interconnections suffer from frequent deflections, and some periodic oscillations [12]. The challenge is thus the definition of methods to control the coupling between inter-AS and intra-AS routing, as the authors in [13] conclude after studying these interactions.

C. Peering link congestion

It should also be noted that the incentives for increasing the capacities of peering links are not straightforward. Indeed, peering agreements do not relay on any payment, as opposed to transit agreement. Controlling the load on the peering links is thus essential. However, this is difficult, as it requires to set-up very complex routing policies [2]. Furthermore, the current inability to estimate possible IGP weight variations, and thus to foresee the associated inter-domain route deflections they might cause, prevents carriers from controlling the inter-AS link congestion precisely. Whenever available, Multipath BGP is expected to reduce congestion, by better distributing the load over the different available routes (through the different peering links) with the same IGP costs. However, the choice of routes on which to distribute the load is based on internal costs, which might lead to inefficient traffic distribution for the peer's network. The challenge is thus the definition of scalable peering link control methods, with some collaboration.

III. The ClubMED Framework

We present the ClubMED (Coordinated MED) framework, characterized in detail in [14]. Within it the MED signalling between peering ASs is modeled as a non-cooperative peering game that can allows the peers to coordinate towards rational, efficient and stable multipath routing solutions.

A. The ClubMED peering game

The idea is to re-use the MED as the means to exchange loose routing and link congestion costs between peer networks for a subset of destination prefixes, in order to help carriers to better collaborate in the load sharing decisions. The scheme relies on a game theoritical modeling of the load sharing problem. Each peer is represented as a rational player that can take benefit by routing accordingly to a cost game built upon routing and congestion costs. The principle is to take the peering routing decision following efficient equilibrium strategy profiles of the game - in its one-shot form or repeated form - thus allowing better collaboration between carriers.

We can introduce the game on a simple example, depicted in Fig. 1, with two peers, AS I and AS II. Let us first define a *destination cone* as a set of customers' destination prefixes. On Fig. 1, Community A and Community B represent two critical destination cones that may deserve careful peer routing, e.g. because they produce high bit-rate flow aggregates. The inter-cone flows are supposed to be equivalent, for instance w.r.t. their bandwidth, so that their path cost can be fairly compared and their routing coordinated. We assume that the cones represent direct customers or stub ASs, which would often assure that their entry point in a peer network is unique (this would reinforce the equivalence condition of the two flows, but is not, however, a strict requirement).

We propose that the two ASs coordinate the choice on the egress peering link for each outgoing flow, from Community A to Community B and vice-versa. A "ClubMED peering game" is built at R_a and R_b routers, called *ClubMED nodes*, using the egress IGP path cost, the ingress IGP path cost, the same costs for the peer announced via the MED, and endogenously-set peering link congestion costs. At ClubMED nodes, efficient equilibria can be selected, accordingly to the different policies detailed in the next section, so as to decide the egress route(s) for each inter-community flow.

978-1-4244-4148-8/09/$25.00 ©2009

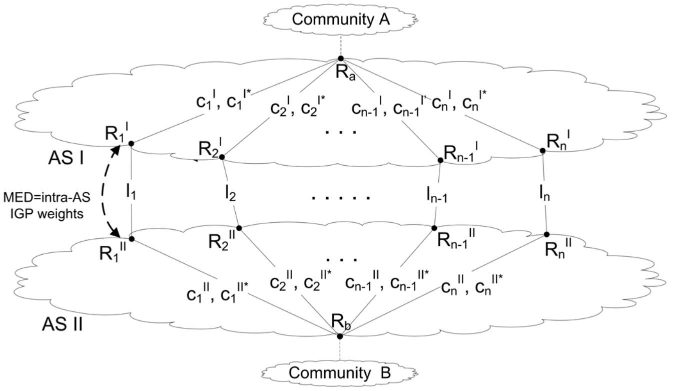


Fig. 1. Single-pair ClubMED interaction example.

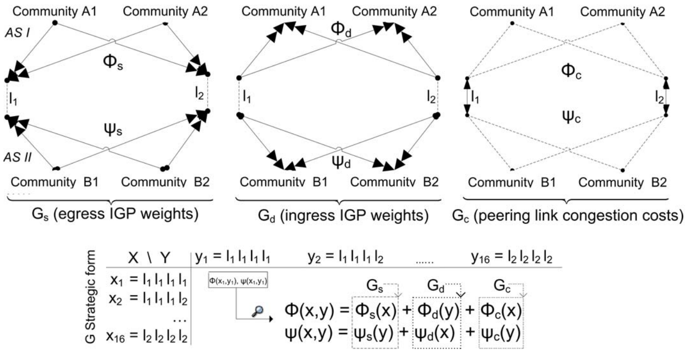


Fig. 2. Multi-pair 2-link ClubMED game composition example.

In order to take broader decisions, many pairs of inter-cone flows shall be considered in a same ClubMED game. In this way, the equivalence condition (e.g., on the bandwidth) can be extended to all the pairs together, not necessarily related to a same couple of ClubMED nodes. Therefore, the final ClubMED game derives from the superposition of many inter-community flows (e.g. in Fig.2 we have 4 pairs and 8 flows).

With multiple pairs of cones, carriers shall control the congestion on inter-peer links. The more egress flows are routed on a peering link, the more loaded the link and the congestion risk, and the higher the routing cost. Hence, we aim at weighting the inter-carrier links with congestion costs when congestion may arise. This could be alternatively done by modeling the inter-peer link in IGP-WO operations (e.g. [10]), but this would violate, however, the requirement of decoupling intra-domain from inter-domain routing [13].

1) Notations: The ClubMED game can be described as $G = G_s + G_d + G_c$, sum of a selfish game, a dummy game and a congestion game, respectively, as depicted in Fig. 2. Let X and Y be the set of strategies available to AS I and AS II (resp.): each strategy indicates the peering link where to route each inter-community flow. Let $(\phi(x, y), \psi(x, y))$ be the strategy cost vector for the the strategy profile (x, y), $x \in X$, $y \in Y$. E.g., in Fig. 2, we have 4 pairs (A1↔B1,A1↔B2,A2↔B1,A2↔B2) and 2 links (l_1,l_2), and X and Y become $\{l_1l_1l_1l_1$, $l_1l_1l_1l_2$, ..., $l_2l_2l_2l_2\}$. For m pairs and n links, the game is the repeated permutation of m single-pair n-link games, thus with $|X|$=$|Y|$=n^m. G_s considers egress IGP weights only, modeling a sort of extended hot-potato rule. G_d considers ingress IGP weights only, impacted by the other peer's routing decision (not taken into account in the legacy BGP decision process). G_c considers peering link congestion costs computed as explained hereafter.

Let c_{ji}^{I} and c_{ji}^{II} be the egress IGP weight from the j^{th} ClubMED node of AS I and AS II to the i^{th} peering link l_i, $i \in E$, $|E|$=n. Let ${c_{ij}^{I}}^*$ and ${c_{ij}^{II}}^*$ be the corresponding ingress weights, from the i^{th} link to the j^{th} ClubMED node.

$G_s = (X, Y ; f_s, g_s)$, is a purely endogenous game, where $f_s, g_s : X \times Y \rightarrow \mathbf{N}$ are the cost functions for AS I and AS II (resp.). In particular, $f_s(x, y) = \phi_s(x)$, where $\phi_s : X \rightarrow \mathbf{N}$, and $g_s(x, y) = \psi_s(y)$, where $\psi_s : Y \rightarrow \mathbf{N}$. E.g., for the topology in Fig. 2, consider the profile (x^+, y^+) with $x^+ = l_1l_2l_1l_1$ and $y^+ = l_1l_1l_1l_2$; we have:
$f_s(x^+, y^+) = \phi_s(x^+) = c_{11}^{I} + c_{12}^{I} + 2c_{21}^{I}$
$g_s(x^+, y^+) = \psi_s(y^+) = 2c_{11}^{II} + c_{21}^{II} + c_{22}^{II}$.

$G_d = (X, Y ; f_d, g_d)$, is a game of pure externality, where $f_d, g_d : X \times Y \rightarrow \mathbf{N}$, $f_d(x, y) = \phi_d(y)$ and $\phi_d : Y \rightarrow \mathbf{N}$, $g_d(x, y) = \psi_d(x)$ and $\psi_d : X \rightarrow \mathbf{N}$. For the above example:
$f_d(x^+, y^+) = \phi_d(y^+) = 2{c_{11}^{I}}^* + {c_{12}^{I}}^* + {c_{22}^{I}}^*$
$g_d(x^+, y^+) = \psi_d(x^+) = {c_{11}^{II}}^* + 2{c_{12}^{II}}^* + {c_{21}^{II}}^*$.

$G_c = (X, Y ; f_c, g_c)$ is an endogenous game too, where $f_c, g_c : X \times Y \rightarrow \mathbf{N}$. $f_c(x, y) = \phi_c(x)$ and $g_c(x, y) = \psi_c(y)$. In order to build the congestion game, the flow bit-rates have to be known. Let H be the set of inter-peer flow pairs, ρ_h the outgoing flow bitrate of the pair $h \in H$, and C_i the egress available capacity of l_i. With multipath, ρ_h can be partioned, and ρ_h^i is the fraction routed towards l_i. G_c should not count when $\sum_{h \in H} \rho_h \ll \min_{i \in E}\{C_i\}$, otherwise it would affect the G equilibrium selection. The congestion cost is to be monotonically increasing with the number of flows routed on a link [18]; one can use (idem for $\psi_c(y)$):

$$\phi_c(x) = \sum_{i \in E | l_i \in x} \left\lceil K_i \frac{1}{C_i - \sum_{h \in H} \rho_h^i} \right\rceil \qquad (1)$$

If $C_i < \sum_{h \in H} \rho_h^i$, $K_i = \infty$. Otherwise, K_i are constants to be scaled to make the cost comparable to IGP costs, e.g., such that it is 1 when the idle capacity is maximum, i.e., $K_i = C_i$.

2) Peering Nash equilibrium: $G_s + G_c$ is a cardinal potential game [17], i.e., the incentive to change players' strategy can be expressed in one potential function, and the difference in individual costs by an individual strategy move has the same value as the potential difference. G_d can be seen as a potential game too, but with null potential. Hence, the G potential $P : X \times Y \rightarrow \mathbf{N}$ depends on G_s and G_c only. As property of potential games [17], the P minimum corresponds to a Nash equilibrium and always exists. The inverse is not necessarily true, but it can be easily proven that for G it is thanks to the endogenous nature of G_s and G_c. The ClubMED peering Nash equilibrium is thus guided by the egress IGP weights and the congestion costs, and may be not unique when their sum is equal over different strategies.

The opportunity of minimizing of the potential function to catch all the peering Nash equilibria represents a key advantage. It decreases the equilibrium computation complexity, which would have been very high for instances with many links and pairs. When there are multiple equilibria, G_d can help in avoiding inefficient solutions (e.g. due to tie-breaking) by the selection of an efficient equilibrium in the Pareto-sense.

3) Pareto-efficiency: A strategy profile p is *Pareto-superior* to another profile p' if a player's cost can be decreased from p to p' without increasing the other players' costs. The *Pareto-frontier* contains the *Pareto-efficient* profiles, i.e. those not Pareto-inferior to any other. In the ClubMED game, ingress costs affect the Pareto-efficiency (because of the G_d pure externality). In particular, given many Nash equilibria, the Pareto-superiority strictly depends on G_d. E.g., Fig. 3 depicts two cases with 3 links and their strategic forms (G_c is not considered). The exponent indicates the corresponding potential value. Egress costs are close to the egress points, while ingress costs to the communities. For the upper case, there is a single equilibrium, (l_2, l_2). For the lower one, there are four equilibria, and (l_3, l_1) is the single Pareto-superior one; however, it is not Pareto-efficient, but Pareto-inferior to (l_1,l_3) that is not an equilibrium because AS I will always prefer l_2 or l_3 to l_1 ($11 < 13$). This is due to the external effect of G_d. Indeed, it is possible that, after an iterated reduction of strategies, G assumes the form of a Prisoner-dilemma game, in which equilibria are Pareto-inferior to other profiles.
Note 1: To explicate P in calculus, we use a form in which we set to 0 the minimum of ϕ_s and ψ_s, i.e., $P_s(x_0, y_0) = 0$ where: $\phi_s(x_0) \leq \phi_s(x)\ \forall x \in X$, and $\psi_s(y_0) \leq \psi_s(y)\ \forall y \in Y$.
Note 2: In the simple example of Fig. 3, all the Nash equilibria have a null potential value, but this is not the case in general.

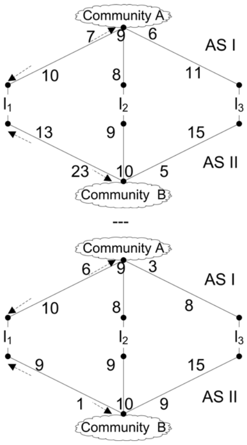
B. Modeling of IGP-WO operations

Nowadays, IGP weights are frequently optimized and automatically updated rather than being manually configured. In this sense, we should assume that the ClubMED costs are subject to changes when the ingress/egress flow directions changes. In the following we explain how, in the ClubMED framework, the coupling among IGP and BGP routing can be modeled to anticipate the route deflection issue presented in Sect. II-B. We aim at selecting a robust peering equilibrium with an approach that is vaguely related somehow to the idea presented in [15] to stabilize intra-domain routing w.r.t. traffic pattern variations.

At a given ClubMED node i of AS I, let $\delta_s^{i,j,I}$ and $\delta_s^{j,i,I*}$ be the (i, j) path cost variations in the egress and ingress directions (resp.) when passing from the current routing to the routing profile $s \in X$ (idem $\delta_s^{i,j,II}$ and $\delta_s^{j,i,II*}$ for AS II). δ variations could be used to extend the G Nash set and Pareto-frontier. However, the δ should not be announced via the MED to avoid a large overhead and an excessive insight in a carrier's operations. Each peer can announce just a directional path cost error. Let ϵ^I and ϵ^{II} be these egress cost errors for AS I and AS II (resp.). Being aware that IGP weights may significantly increase, an optimistic min-max computation can be:

$$\epsilon^I = \min_{(i,j)} \left\{ \max_{s \in X} \left\{ \delta_s^{i,j,I} \right\} / c_{i,j}^I \right\} \tag{2}$$

Similarly for ϵ^{II}, ϵ^{I^*} and ϵ^{II^*}. The ϵ cost errors represent a good trade-offs between network information hiding and coordination requirement: not announcing per-link errors avoid revealing the δ variations; announcing directed errors (ingress and egress) allows reflecting the fact that upstream and downstream availability is likely to be unbalanced because of the bottleneck asymmetry in inter-AS links.



I\II	l_1	l_2	l_3
l_1	(17,36)^6	(19,32)^2	(16,38)^8
l_2	(15,23)^4	(**17,19**)^0	(14,25)^6
l_3	(18,18)^7	(20,14)^3	(17,20)^9

I\II	l_1	l_2	l_3
l_1	(16,10)^2	(19,10)^2	(13,18)^8
l_2	(**14,19**)^0	(**17,19**)^0	(11,25)^6
l_3	(***14,18***)^0	(**17,18**)^0	(11,24)^6

Fig. 3. 3-link examples.

The ϵ errors induce a larger number of equilibria for the multipath routing solution. The game can be easily extended to take into account these error margins. They define a *potential threshold* under which a profile becomes an equilibrium. More precisely, the minimum potential strategies are found, then the other profiles that have a potential within the minimum plus the threshold (T_P) are considered as equilibria too. Each potential difference ΔP from (x_1, y_1) to (x_2, y_2) can be increased of $a_I(x_1, x_2) + a_{II}(y_1, y_2)$, where $a_I(x_1, x_2) = \epsilon^I(\phi_s(x_1) + \phi_s(x_2))$ and $a_{II}(y_1, y_2) = \epsilon^{II}(\psi_s(y_1) + \psi_s(y_2))$. An optimistic threshold can be:

$$T_P = \min_{x_1, x_2 \in X} \{a(x_1, x_2)\} + \min_{y_1, y_2 \in Y} \{a(y_1, y_2)\} \tag{3}$$

Indicating with $P(x_0, y_0)$ the potential minimum, all strategy profiles (x, y) such that $P(x, y) \leq P(x_0, y_0) + T_P$ will be considered as equilibria. This operation can also allow escaping selfish (endogenous) solutions mainly guided by G_s + G_c, introducing Pareto-superior profiles in the Nash set.

IV. Peering Equilibrium MultiPath (PEMP) policies

Peers would route accordingly to an equilibrium because it grants a rational stability to the routing decision. The Nash set and the Pareto-frontier may be quite broad, especially considering IGP path cost errors. This leads to different possible Peering Equilibrium MultiPath (PEMP) load balancing policies (upon these profile sets), which are presented below.

A. Nash Equilibrium MultiPath (NEMP) implicit coordination

Assuming thus that ClubMED remains a fully non-cooperative framework, its implicit solution strategy to which to coordinate without any signaling message is: *play the equilibria of the Nash set*. Hence, it is feasible to natively implement a Nash Equilibrium MultiPath (NEMP) routing policy. E.g, in the bottom of Fig. 3 AS I may balance the load on l_2 and l_3, being aware that AS II may balance its load on l_1 and l_2. However, the set of equilibria can be shrinked to the Pareto-superior ones; but many Pareto-superior equilibria can exist, so the NEMP policy is to be used in this case too. Please note that there may not exist Pareto-superior equilibria: in this case, NEMP is performed over all the equilibria.

B. Repeated coordination

Given that the the G Pareto-frontier may not contain equilibria, in a repeated ClubMED context, an explicit coordination strategy is: *play the profiles of the Pareto-frontier*. The ClubMED game would be repeated an indefinite number of times, indeed. From "folk-theorem"-like results [16], this strategy is an equilibrium of the repeated game and grants a maximum gain for the players in the long-run. Nevertheless, the unilateral trust for such a strategy could decrease whether in a short period of analysis the gains reveal to be unbalanced and in favor of a single peer. The reciprocal trust among peers can thus affect the reliability of such a Pareto coordination.

Unself-Jump: Another strategy is conceivable to guarantee balancedness in gains in the short term, and thus helping to keep a high level of reciprocal trust. After shrinking the Nash set w.r.t. the Pareto-efficiency, for each equilibrium the ASs might agree to make both a further step towards the best available strategy profile (x^j, y^j) such that:

$$\psi(x^j, y^j) - \psi(x_0, y_0) + \phi(x^j, y^j) - \phi(x_0, y_0) < 0 \qquad (4)$$

where (x_0,y_0) is the starting equilibrium. One AS may unselfishly sacrifice for a better bilateral solution: the loss that one may have moving from the selected equilibrium is compensated by the improvement upon the other AS. This strategy makes sense only if the other AS is compensated with a bigger improvement, and returns the favor the next times.

Pareto-Jump: Instead, with the addition of the constraint:

$$\psi(x^j, y^j) - \psi(x_0, y_0) \leq 0 \ \wedge \ \phi(x^j, y^j) - \phi(x_0, y_0) \leq 0 \quad (5)$$

we select a Pareto-superior profile (not necessarily in the Pareto-frontier), without unselfishly sacrifices. If at least one (x^j, y^j) is found we obtain a new profile set that is to be shrinked w.r.t. the Pareto-superiority for the final solution.

E.g., in the bottom example of Fig. 3, we would jump from the Pareto-superior Nash equilibrium (l_3, l_1) to the Pareto-superior profile (l_1, l_3). We would not have this jump for the Unself-Jump policy, that would prefer instead (l_1, l_1) with a global gain of 6 instead of "just" 3 with (l_1, l_3).

Finally, note the last two policies are not binding: it would be enough to associate the policy with the menace to pass to one of the more selfish choices. Also note that MEDs from different ASs should be normalized to the same IGP weight scale in order to be comparable.

V. Performance Evaluation

We evaluated the performance of the three PEMP routing policies with realistic simulations. We created a virtual interconnection scenario among the Geant2 and the Internet2 ASs, depicted in Fig. 6, emulating their existing peering with three cross-atlantic links. We considered six pairs of inter-cone flows among the routers depicted with crossed circles. The TOTEM toolbox [19] was used to run a IGP-WO heuristic, with a maximum IGP weigth of 50 for both ASs. We used 252 successive traffic samples, oversampling the datasets from [20] for Geant2 and from [21] for Internet2 on a 8h basis (to cover all the day times). The original link capacity was scaled by 10 to create an intra-domain congestion risk. The inter-cone routing generates additional traffic for the traffic matrices. We used a random inter-cone traffic matrix such that flows are balanced with 200 Mb/s per direction, which corresponds to $2/3$ of the total available peering capacity. To evaluate the effectiveness of the congestion game we considered peering links with 100 Mb/s available per direction.

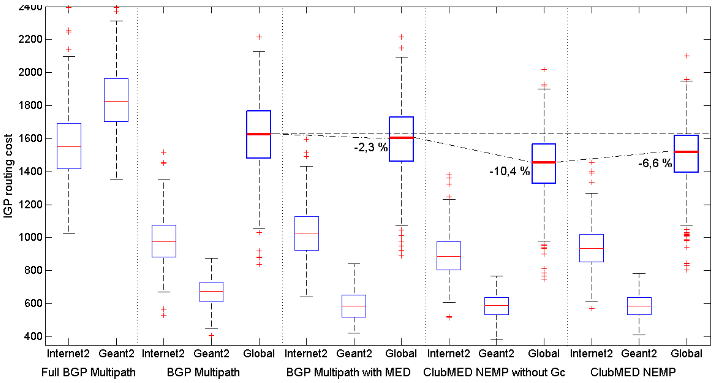


Fig. 4. IGP routing cost Boxplot statistics: NEMP vs BGP Multipath.

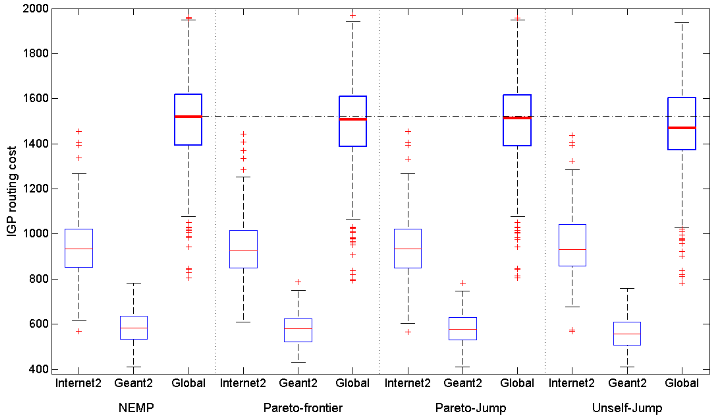


Fig. 5. IGP routing cost Boxplot statistics: PEMP strategies.

We compare the PEMP routing policies ('NEMP', 'Pareto-Frontier', 'Pareto-Jump', 'Unself-Jump') to the 'BGP Multipath' solution without and with ('...+MED') classical MED signalling enabled at both sides, and to a 'Full BGP Multipath' solution in which all the peering links (i.e., the available routes) are used for the multipath solution.

A. Routing cost

Fig. 4 reports the IGP routing costs statistics in BoxPlot format (minimum; box with lower quartile, median, upper quartile; maximum; outliers). We show four solutions: Full BGP Multipath, BGP Multipath, the NEMP policy without and with the congestion game G_c. For each method, we display the Internet2, the Geant2 and the global routing costs. We considered two ClubMED solutions, with and without the congestion game G_c (for the first two figures only).

The full BGP multipath solution obviously guarantees an even load on all the peering links. However, its routing cost almost doubles than with normal BGP multipath, which balances the load only on equal cost paths (egress IGPs or MEDs). Simple MED usage decreases the cost of the BGP case without MED, due to one network that is more loaded (hence, higher IGP weights), and to the fact that with the MED the chance of ECMP is higher (not only on equal IGP path cost routes, but also on equal MED routes). The ClubMED solution, instead, outperforms BGP with a median cost lower by 10% without G_c, and by 6,6% in its complete form.

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.

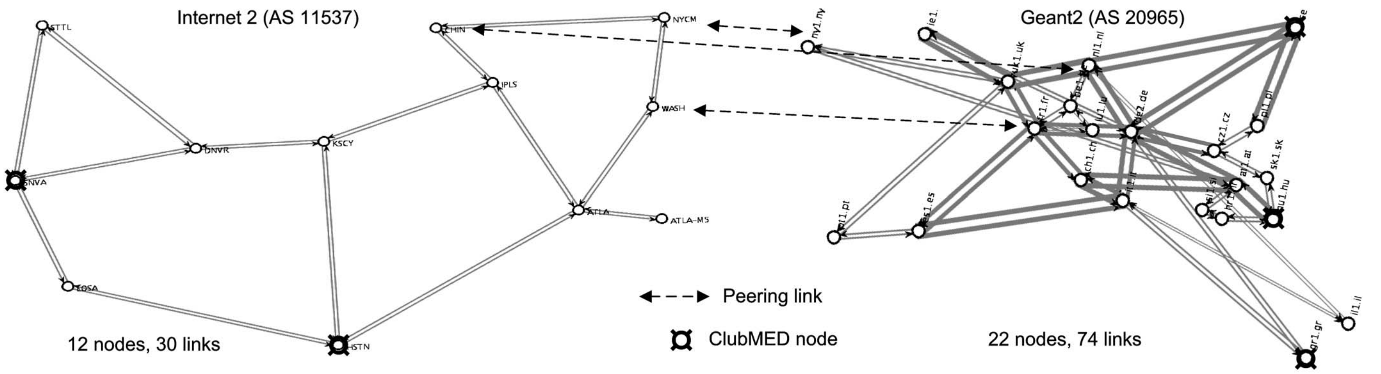


Fig. 6. Internet2 - Geant2 peering scenario with 3 peering links.

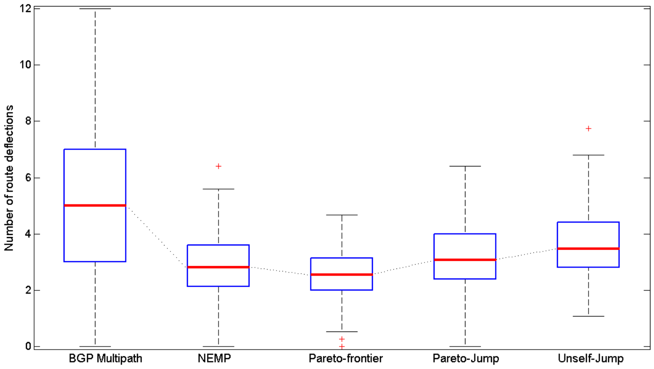


Fig. 7. Number of route deflections.

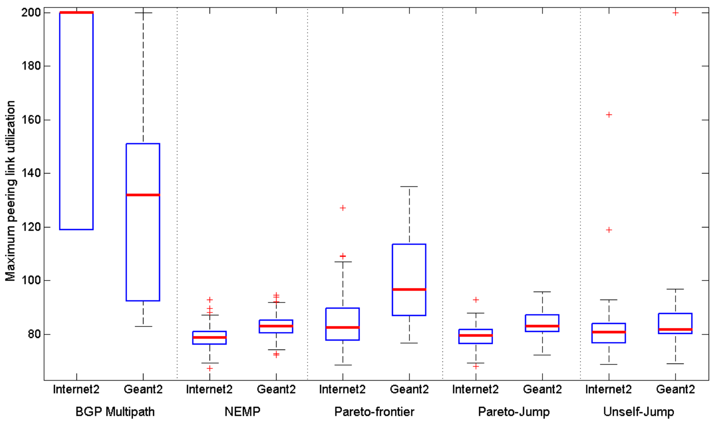


Fig. 8. Maximum peering link utilization boxplot statistics.

Fig. 5 compares the four PEMP policies. With respect to NEMP, the Pareto policies give statistically very close results. This may sound disappointing: one may expect more from the Pareto-frontier and the Pareto-Jump policies. By analyzing the results in detail, we verified that the reason of this poor performance is that the Pareto-frontier often contains strategy profiles with the least cost for a peer and very high cost for the other peer. Such strategy profiles are not marked as Pareto-inferior because of the single peer's least cost and thus belong to the Pareto-frontier. Such situations are likely to be frequent since an uncongested intra-domain link may produce a IGP weight much lower than the others thus affecting the G profile cost components. This risk is augmented in the Pareto-Jump policy since the new selected profiles can "just" be Pareto-superior: they do not necessarily belong to the Pareto-frontier. However, for the Pareto-jump policy the median, the minimum and the upper and lower quartiles outperform the NEMP result; in fact, the starting Nash set for its Pareto-improvement is the NEMP one (see Sect. IV-B). Finally, the Unself-Jump policy should outperform or equalize the Pareto-Jump one w.r.t. the routing cost since, without (5), it can be see as its relaxation. Indeed, as reported in Fig. 5, the Unself-Jump gives a median cost roughly 3% inferior than the NEMP cost.

B. Route deflections

Fig. 7 reports the statistics of routing changes with respect to the previous round (with an upper bound equal to the total number of flows). The PEMP policies behave significantly better than BGP Multipath: they have a median of around 3 route deflections against 5, and the upper quartile and the maximum much lower. Interestingly, among the PEMP policies, the Pareto-frontier one statistically behaves better than the other policies for all the criteria but for the minimum. The reason may be that the Pareto-superiority condition - applied on a very large set of candidate profiles (in fact, $n^{2m} = 531441$) - offers a finer selection than the approximate potential threshold one. Finally, the Jump policies present a lower route stability w.r.t. all the statistical criteria. This is reasonably due to the fact that the jump from the Nash set, i.e., the unself and Pareto-superior conditions, are computed in the simulations without considering the cost errors.

C. Peering link congestion

Fig. 8 reports the Boxplot statistics maximum link utilization as seen by each peer, with all the methods. All the PEMP strategies but the Pareto-frontier one never caused congestion on peering links (utilization above 100%). The enabling of the Multipath mode in BGP does not have a significant effect on the peering link congestion. With ClubMED, instead, the multipath routing choice is carefully guided toward efficient solutions. The NEMP, Pareto-Jump and Unself-Jump policies show the median, the upper and lower quartiles always above 85%, remembering that with full BGP Multipath one would have a $200/300 = 66,7\%$ utilization. The Pareto-frontier strategy does not guarantee, however, a congestion-free solution, with a median close to 100% utilization. The reason

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.

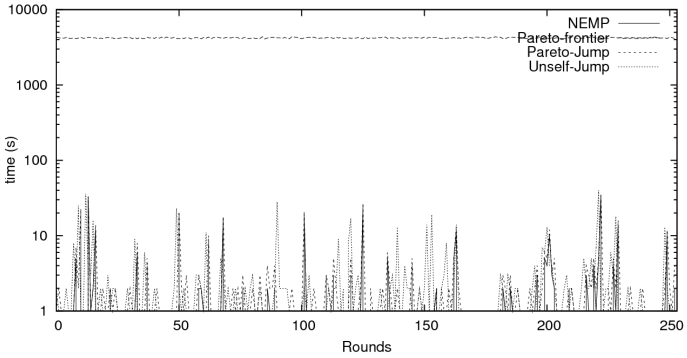


Fig. 9. PEMP strategies execution time.

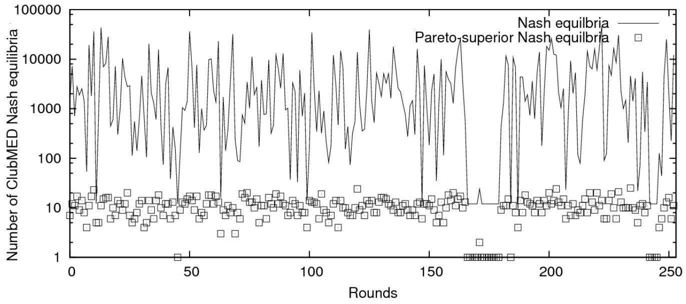


Fig. 10. Nash set dynamics.

for this behavior are still the highly asymmetric cost profiles introduced by the Pareto-superiority condition in the solution.

D. Time complexity

Fig. 9 reports the PEMP execution time. We know that the Pareto-frontier computation is cumbersome, with a $O(n^{2m})$ time complexity, while the other policies have a polynomial complexity that asymptotically depends on the minimization of a (mono-dimension) potential function to populate the Nash set. In fact, the other policies have an average computation time below the 2 seconds (however, rare peaks of a few more seconds appear, probably due to the cases with very large Nash set, as it can be see cross-checking with Fig. (10). Hence, only the NEMP, Pareto-Jump and Unself-Jump policies shall be considered for a practical implementation. We have, however, introduced the Pareto-frontier case for a thorough comparison.

E. Nash equilibrium dynamics

Fig. 10 reports the number of equilibria and those Pareto-superior in a log-scale for all the rounds. The Pareto-superiority condition permits to pick a few efficient Nash equilibria over broad sets, whose dimension varies significantly in time. This reveals a high sensibility to the routing costs due to the endogenous effect of G_c with high congestion costs.

VI. Summary

We modeled the routing on peering links as a non-cooperative game with the aim to allow carriers fine-selecting routes for critical flows by following efficient equilibrium multipath solutions. We presented the mathematical model of the game, composed of a selfish game (with egress IGP costs), a dummy game (with ingress IGP costs) and a congestion game. The game components can be adapted to consider IGP cost variations due to IGP-WO re-optimizations.

We proposed a low-computational way to compute the Nash equilibria, and four possible Peering Equilibrium MultiPath (PEMP) routing coordination policies. The first twos corresponds to balance the load on the Pareto-superior Nash equilibria of the one-shot game, and on the Pareto-frontier (equilibrium of the repeated game), respectively. The latter two policies correspond to improve the first strategy moving from the Pareto-superior Nash set refinement toward exterior Pareto-superior and unselfish routing profiles, respectively.

We simulated the PEMP policies with a realistic emulation, comparing them to BGP Multipath. The results show they outperforms BGP Multipath in terms of routing cost, route stability and peering link congestion. In particular, the route stability is significantly improved and the peering link congestion can be practically avoided. Some differences exist between the PEMP policies. Namely, the Pareto-frontier one is extremely complex and shall not be implemented. The other ones present some trade-offs but represent all promising solutions to perform an efficient and rational routing across peering links. In particular, the Unself-Jump policy represents the best trade-off between peering trust insurance, routing cost, congestion control, routing stability and execution time.

We are currently working on the definition of an extended peering framework, modeling the border with multiple ASs as the single border of a classical peering.

References

[1] R. Teixera et al., "In Search of Path Diversity in ISP Networks", in *Proc. of ACM Internet Measurement Conference*, Oct. 2003.

[2] Jiayue He, J. Rexford, "Towards Internet-wide multipath routing" in *IEEE Network magazine*, March 2008.

[3] "Configuring BGP to Select Multiple BGP Paths", JUNOS document.

[4] "BGP Best Path Selection Algorithm", Cisco documentation.

[5] Y. Rekhter, T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771.

[6] D. McPherson, V. Gill, "BGP MED considerations", RFC 4451.

[7] A. Lange, "Issues in Revising BGP-4", draft-ietf-idr-bgp-issues (2003)

[8] R. Teixeira et al., "TIE Breaking: Tunable Interdomain Egress Selection", in *Proc. of CoNEXT 2005*.

[9] S. Agarwal, A. Nucci, S. Bhattacharyya, "Controlling Hot Potatoes in Intradomain Traffic Engineering", SPRINT RR04-ATL-070677, 2004.

[10] S. Balon, G. Leduc, "Combined Intra- and inter-domain traffic engineering using hot-potato aware link weights optimization", arXiv:0803.2824.

[11] M. Latapy, C. Magnien, F. Ouédraogo, "A Radar for the Internet", in *Proc. of ADN 2008*.

[12] S. Secci et al., "Detection of route deflections across top-tier interconnections", working paper, http://perso.enst.fr/secci/papers/radar.pdf

[13] R. Teixeira et al., "Impact of Hot-Potato Routing Changes in IP Networks", *IEEE/ACM Trans. on Networking*, Vol. 16, Dec. 2008.

[14] S. Secci et al., "ClubMED: Coordinated Multi-Exit Discriminator Strategies for Peering Carriers", in *Proc. of NGI 2009*.

[15] P. Casas, L. Fillatre, S. Vaton, "Multi hour robust routing and fast load change detection for traffic engineering", in *Proc. of IEEE ICC 2008*.

[16] R.B. Myerson, *Game Theory: Analysis of Conflict*, Harvard Univ. Press.

[17] D. Monderer, L.S. Shapley, "Potential Games", *Games and Economic Behavior*, Vol. 14, No. 1, May 1996, Pp: 124-143.

[18] F. Larroca, J.-L. Rougier, "Routing Games for Traffic Engineering", in *Proc. of IEEE ICC 2009*.

[19] J. Lepropre, S. Balon, G. Leduc, "Totem: a toolbox for traffic engineering methods", in *Proc. of INFOCOM 2006*.

[20] S. Uhlig et al., "Providing public intradomain traffic matrices to the research community", *Computer Communication Review* 36 (1) (2006).

[21] By courtesy of Y. Zhang. Abilene topology and traffic dataset. http://www.cs.utexas.edu/yzhang/research/AbileneTM.