Peering Equilibrium Multipath Routing: A Game Theory Framework for Internet Peering Settlements

Stefano Secci, *Member, IEEE*, Jean-Louis Rougier, Achille Pattavina, *Senior Member, IEEE*, Fioravante Patrone, and Guido Maier, *Senior Member, IEEE*

***Abstract*—It is generally admitted that interdomain 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 multipath routing coordination framework for peering carriers, which relies on the multiple-exit discriminator (MED) attribute of Border Gateway Protocol (BGP) as signaling medium. Our scheme relies on a game theory modeling, 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. Finally, we discuss practical implementation aspects and extend the model to multiple players highlighting the possible incentives for the resulting extended peering framework.**

***Index Terms*—Border Gateway Protocol (BGP), game theory, interdomain routing, multiple-exit discriminator (MED), multipath, peering.**

I. Introduction

MULTIPATH routing has received interest for a long time, as it is considered to be a very efficient solution providing more robustness and better load distribution on the network. Intradomain multipath routing is commonly performed in Interior Gateway Protocol (IGP) networks by balancing the load over equal-cost multiple paths (ECMPs) [2]. In the multidomain context, multipath routing is generally not implemented, its introduction raising important scalability and complexity issues (see, e.g., [3]). 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., [4] and [5]). 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 with respect to load distribution, may lead to an inefficient situation for the downstream carrier. In this paper, we propose a framework that allows 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 tradeoffs 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 reuse 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 selected by carriers in coordination with each other. We show by simulations that this choice prevents congestion on peering links and decreases the global routing cost while increasing the route stability.

Section II presents the intercarrier routing issues that we tackle. Section III presents the ClubMED (Coordinated MED) framework for multipath routing across peering links. We explain how load balancing shall be implemented over efficient equilibrium strategies. Section IV defines the peering equilibrium multipath (PEMP) routing coordination policies and discusses their possible benefits and implementation issues. Section 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. In Section VI, we discuss practical implementation aspects. In Section VII, we discuss how the two-player model can be extended to an arbitrary number of players, highlighting the possible incentives for the resulting extended peering framework. Section VIII concludes the paper.

Manuscript received October 08, 2009; revised May 18, 2010; accepted July 22, 2010; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor V. Misra. Date of publication August 23, 2010; date of current version April 15, 2011.This work was supported in part by the European ICT FP7 Euro-NF Network of Excellence, the I-GATE project of the Institut Telecom, and the FP7 IP project ETICS (Economics and Technologies for Inter-Carrier Services). A preliminary version of this paper appears in the proceedings of the IEEE Global Communications Conference (GLOBECOM 2009), Honolulu, HI, November 30–December 4, 2009.

S. Secci is with the LIP6, Université Pierre et Marie Curie, Paris 75005, France (e-mail: stefano.secci@ieee.org).

J.-L. Rougier is with the Institut Telecom, Telecom ParisTech, LTCI CNRS, Paris 75013, France (e-mail: rougier@telecom-paristech.fr).

F. Patrone is with the Università di Genova, Genova 16145, Italy (e-mail: patrone@diptem.unige.it).

A. Pattavina and G. Maier are with the Politecnico di Milano, Milan 20133, Italy (e-mail: pattavina@elet.polimi.it; maier@elet.polimi.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNET.2010.2062535

II. Intercarrier Routing Issues

A. BGP and Selfish Routing

It is worth briefly recalling how the route selection is performed via BGP [6]. When multiple paths to a destination network are available, a cascade of criteria is employed to compare them, eliminating routes until only one (the best) remains. 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: 1) the route with a smaller AS hop count; 2) if the routes are received by the same neighbor AS, the route with a smaller MED; 3) the route with the closer egress point ("hot-potato" rule), using as distance metric the IGP path cost; 4) the more recent route; 5) 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 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 with respect to 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 routing costs when taking the peering routing decision.

1) 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 them. 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 [7].

2) 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. For the time being, such multipath extensions for BGP have not found consensus at the IETF, and for this reason there is no standard specification. However, some suggestions are indicated in [8]. As of our knowledge, the only implemented method carriers can use for multipath interdomain routing is the "BGP Multipath" mode that some router vendors now provide (e.g., Juniper [4] and Cisco [5]), 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 internal BGP (IBGP) router has equivalent routes to a destination through many border routers, or when all of the candidates routes are learned via external BGP (EBGP). As stated in [8], other cases, with a combination of IBGP and EBGP routes, should be avoided, as they may lead to routing loops for instance.

B. BGP Route Deviation

The peering routing decision with BGP thus relies on IGP routing costs (least MED or hot-potato criteria). 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 nontransient network events (which persist for a long period), a carrier may reoptimize the IGP weights, inducing changes in the BGP routing decision, so that congestions might appear where not expected.

Many works concern BGP route deviation control methods. Reference [9] reformulates the egress routing problem and proposes to replace the hot-potato rule with a more expressive and efficient rule. Reference [10] presents a comprehensive yet hard IGP weight optimization (IGP-WO) method aware of possible hot-potato route deviations to bound them (they report that 70% of traffic can be affected in a real network). Reference [11] 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 thus be 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 deviations using TRACETREE radar data [12] and Paris-traceroute data. Preliminary results confirm that top-tier AS interconnections suffer from frequent deviations and some periodic oscillations [13]. The challenge is thus the definition of methods to control the coupling between inter-AS and intra-AS routing, as the authors in [14] 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 rely 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 setting very complex routing policies [3].

Furthermore, the current inability to estimate possible IGP weight variations, and thus to foresee the associated interdomain route deviations 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 with the same IGP costs (and through different peering links). 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. ClubMED Framework

We present the ClubMED (Coordinated MED) framework. Within it, the MED signaling between peering ASs is modeled

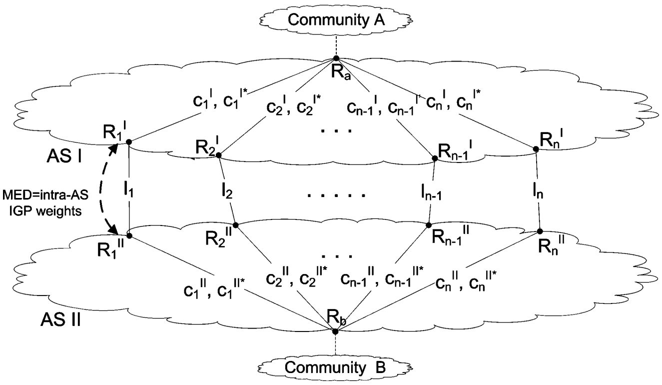


Fig. 1. Single-pair ClubMED interaction example.

as a non-cooperative peering game that can allow the peers to coordinate toward rational, efficient, and stable multipath routing solutions.

A. ClubMED Peering Game

We propose to reuse the MED attribute as the means to exchange loose routing costs and peering link congestion costs between peer networks: a coordinated MED signaling can help carriers to better collaborate in the load-sharing decisions. For the sake of implementation, we consider IGP weights for the game-routing cost components; such costs are classically optimized so that the lower they are, the more they attract shortest paths and traffic, and the higher the link congestion risk is.

Our scheme relies on a game-theoretic modeling of the load-sharing problem. Each peer is represented as a rational player that can take benefit by routing according to a cost game built upon routing and congestion costs. The basic idea 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 higher collaboration.

We introduce the game in 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. In 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 with respect to their bandwidth, so that their path cost can be fairly compared and their routing coordinated. We also assume that these cones represent networks that belong to direct customer ASs or stub ASs. This would often assure that their entry point in a peer network is unique. This condition 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 of 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.

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 four pairs and eight flows). With multiple pairs of cones, carriers shall control the congestion on interpeer 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 due to the interpeer flow routing.[1]

1) Notations: The ClubMED game can be described as $G = G_{\rm s}+G_{\rm d}+G_{\rm c}$, 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, respectively: Each strategy indicates the peering link where each inter-community flow is then routed. Let $(\phi(x,y),\psi(x,y))$ be the strategy cost vector for the strategy profile (x,y), $x \in X$, $y \in Y$. In Fig. 2, e.g., we have four pairs $(\mathrm{A1} \leftrightarrow \mathrm{B1}, \mathrm{A1} \leftrightarrow \mathrm{B2}, \mathrm{A2} \leftrightarrow \mathrm{B1}, \mathrm{A2} \leftrightarrow \mathrm{B2})$ and two links (l_1, l_2), and X and Y become $\{l_1l_1l_1l_1, l_1l_1l_1l_2, \ldots, l_2l_2l_2l_2\}$. For m pairs and n links, the game is the repeated permutation of m single-pair n-link games, thus with $|X| = |Y| = n^m$. $G_{\rm s}$ considers egress IGP weights only, modeling a sort of extended hot-potato rule (i.e., extended to many destinations for a same decision). $G_{\rm d}$ considers ingress IGP weights only, impacted by the other peer's routing decision (not taken into account in the legacy BGP decision process). $G_{\rm c}$ considers peering link congestion costs as explained hereafter.

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

$G_{\rm s} = (X,Y; f_{\rm s}, g_{\rm s})$ is a purely endogenous game, where $f_{\rm s}, g_{\rm s} : X \times Y \rightarrow \mathbf{N}$ are the cost functions for AS I and AS II, respectively. In particular, $f_{\rm s}(x,y) = \phi_{\rm s}(x)$, where $\phi_{\rm s} : X \rightarrow \mathbf{N}$, and $g_{\rm s}(x,y) = \psi_{\rm s}(y)$, where $\psi_{\rm s} : Y \rightarrow \mathbf{N}$. For the topology in Fig. 2, e.g., consider the profile $(\tilde{x},\tilde{y})$ with $\tilde{x} = l_1l_2l_1l_1$ and $\tilde{y} = l_1l_1l_1l_2$. We have

$$f_{\rm s}(\tilde{x},\tilde{y}) = \phi_{\rm s}(\tilde{x}) = c_{11}^{\rm I} + c_{12}^{\rm I} + 2c_{21}^{\rm I}$$
$$g_{\rm s}(\tilde{x},\tilde{y}) = \psi_{\rm s}(\tilde{y}) = 2c_{11}^{\rm II} + c_{21}^{\rm II} + c_{22}^{\rm II}.$$

$G_{\rm d} = (X,Y; f_{\rm d}, g_{\rm d})$ is a game of pure externality, where $f_{\rm d},\ g_{\rm d} : X \times Y \rightarrow \mathbf{N}, f_{\rm d}(x,y) = \phi_{\rm d}(y)$ and $\phi_{\rm d} : Y \rightarrow \mathbf{N}, g_{\rm d}(x,y) = \psi_{\rm d}(x)$ and $\psi_{\rm d} : X \rightarrow \mathbf{N}$. For the above example

$$f_{\rm d}(\tilde{x},\tilde{y}) = \phi_{\rm d}(\tilde{y}) = 2{c_{11}^{\rm I}}^* + {c_{12}^{\rm I}}^* + {c_{22}^{\rm I}}^*$$
$$g_{\rm d}(\tilde{x},\tilde{y}) = \psi_{\rm d}(\tilde{x}) = 2{c_{11}^{\rm II}}^* + {c_{12}^{\rm II}}^* + {c_{21}^{\rm II}}^*.$$

$G_{\rm c} = (X,Y; f_{\rm c}, g_{\rm c})$ is an endogenous game too, where $f_{\rm c},\ g_{\rm c} : X \times Y \rightarrow \mathbf{N}$. $f_{\rm c}(x,y) = \phi_{\rm c}(x)$ and $g_{\rm c}(x,y) = \psi_{\rm c}(y)$. In order to build the congestion game, the flow bit rates have to be known. Let H be the set of interpeer flow pairs, ρ_h the

[1]It is worth noting that inter-community flows may represent very high bit rates that dangerously approach the link capacity scale. In these cases, a community pair may be decomposed in many functional community pairs among the same pair of ClubMED nodes.

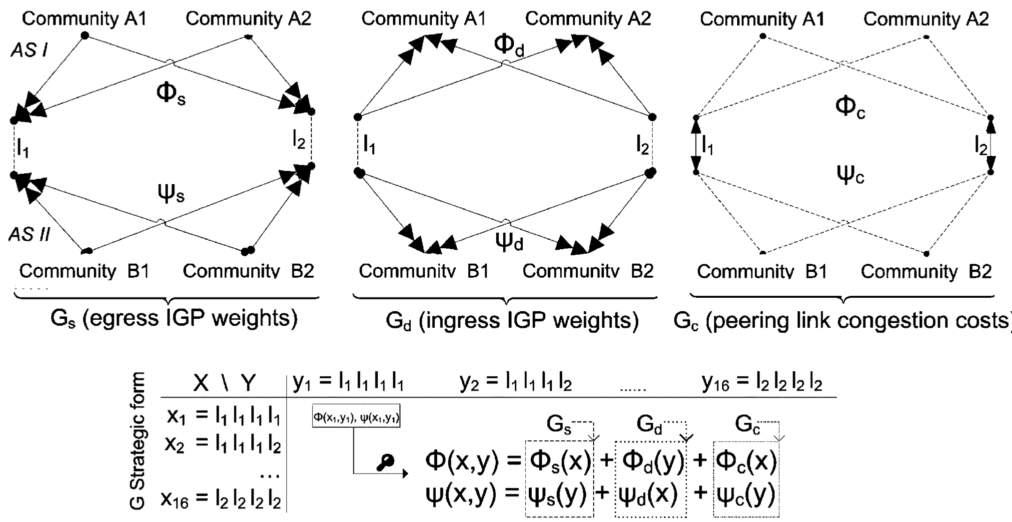


Fig. 2. Multipair two-link ClubMED game composition example.

outgoing flow bit rate of the pair $h \in H$, and C_i the egress available capacity of l_i. With multipath, ρ_h can be portioned, and ρ_h^i is the fraction routed toward l_i. $G_{\rm c}$ 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 function should be monotone increasing with the number of flows routed on a link [18]; one can use [idem for $\psi_{\rm c}(y)$]

$$\phi_{\rm c}(x) = \sum_{i \in E \,|\, l_i \in x} \left\lceil K_i \frac{1}{C_i - \sum_{h \in H} \rho_h^i} \right\rceil. \quad (1)$$

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., $K_i = C_i$.

It is worth noting that when the peering links rely on Internet exchange points (IXPs) and links are full-duplex (e.g., for public IXPs where small-size carriers often peer), and when the incoming flow bit rates can be collected, the congestion cost functions may be designed considering both the incoming and the outgoing flow bit rates. However, in this case the game may no longer be a potential game.

2) Peering Nash Equilibrium: $G_{\rm s} + G_{\rm c}$ 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. $G_{\rm d}$ can be seen as a potential game, too, but with null potential. This decomposition is characterized for the general case in Appendix A. Hence, the G potential $P : X \times Y \to \mathbf{N}$ depends on $G_{\rm s}$ and $G_{\rm c}$ 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 is easy to prove that it is true for G thanks to the endogenous nature of $G_{\rm s}$ and $G_{\rm c}$ (the proof is given in Appendix B).

The ClubMED peering Nash equilibrium is thus guided by the egress IGP weights and the congestion costs and may not be unique when their sum is equal over different strategies.

The opportunity of using the minimization of the potential function to catch all the peering Nash equilibria represents a key advantage. It decreases the Nash equilibrium computation complexity, which would have been very high for instances with many links and pairs. When there are multiple equilibria (which happens in fact quite often), $G_{\rm d}$ can help in avoiding tie-breaking routing by the selection of an efficient equilibrium in the Pareto sense.

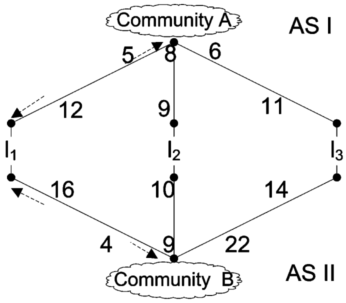
3) Pareto Efficiency: A strategy profile p is *Pareto-superior* to another profile p' if a player's cost can be decreased from p to p' without increasing the other players' costs. The *Pareto frontier* contains the *Pareto-efficient* profiles, i.e., those not Pareto-inferior to any other. In the ClubMED game, ingress costs affect the Pareto efficiency (because of the pure externality of $G_{\rm d}$). In particular, given many Nash equilibria, the Pareto superiority strictly depends on $G_{\rm d}$. Figs. 3 and 4, e.g., depict example cases with three links and their strategic forms ($G_{\rm c}$ is not considered). The exponent indicates the corresponding potential value. Egress costs are close to the egress points, and ingress costs are close to the communities. For Fig. 3, there is a single equilibrium, (l_2, l_2). For Fig. 4, 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), which 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 $G_{\rm d}$. 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 $\phi_{\rm s}$ and $\psi_{\rm s}$, i.e., $P_{\rm s}(x_0, y_0) = 0$ where: $\phi_{\rm s}(x_0) \leq \phi_{\rm s}(x)\ \forall x \in X$, and $\psi_{\rm s}(y_0) \leq \psi_{\rm s}(y)\ \forall y \in Y$.

Note 2: In the simple example of Figs. 3 and 4, all the Nash equilibria have a null potential value, but this is not the case in general.

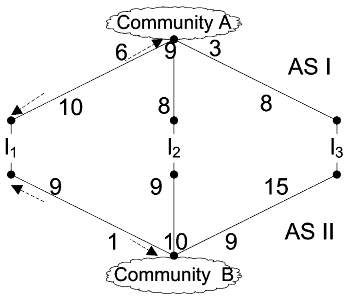
B. Modeling of IGP-WO Operations

Nowadays, IGP weights are frequently optimized, and these operations are often scheduled and automated. In this sense, we should assume that the ClubMED costs are subject to changes when the ingress/egress path changes. In the following, we explain how, in the ClubMED framework, the coupling among



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

Fig. 3. Three-link example with one equilibrium.



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

Fig. 4. Three-link example with multiple equilibria.

IGP and BGP routing can be modeled to anticipate route deviations.

At a given ClubMED node i of AS I, let $\delta_{\mathrm{s}}^{i,j,\mathrm{I}}$ and $\delta_{\mathrm{s}}^{j,i,\mathrm{I}^*}$ be the (i,j) path cost variations in the egress and ingress directions, respectively, when passing from the current routing to the routing profile $\mathrm{s} \in X$ (idem $\delta_{\mathrm{s}}^{i,j,\mathrm{II}}$ and $\delta_{\mathrm{s}}^{j,i,\mathrm{II}^*}$ 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 excessive insight in a carrier's operations. Each peer can just announce a directional path cost error. Let ϵ^{I} and ϵ^{II} be these egress cost errors for AS I and AS II, respectively. Being aware that IGP weights may significantly increase, an optimistic min–max computation can be

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

Similarly for ϵ^{II}, ϵ^{I^*} and ϵ^{II^*}. The ϵ cost errors represent good tradeoffs between network information hiding and coordination requirement: Not announcing per-link errors avoids 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.

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 *potential thresholds* 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 by $a_{\mathrm{I}}(x_1, x_2) + a_{\mathrm{II}}(y_1, y_2)$, where $a_{\mathrm{I}}(x_1, x_2) = \epsilon^{\mathrm{I}}(\phi_{\mathrm{s}}(x_1) + \phi_{\mathrm{s}}(x_2))$ and $a_{\mathrm{II}}(y_1, y_2) = \epsilon^{\mathrm{II}}(\psi_{\mathrm{s}}(y_1) + \psi_{\mathrm{s}}(y_2))$. 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 escape selfish (endogeneous) solutions mainly guided by $G_{\mathrm{s}} + G_{\mathrm{c}}$, introducing Pareto-superior profiles in the Nash set.

IV. Peering Equilibrium Multipath (PEMP)

Within the ClubMED framework, 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 the IGP path cost errors. This leads to different possible PEMP load-balancing policies (upon these sets of profiles).

A. Implicit Coordination

Assuming thus that ClubMED remains a fully non-cooperative framework, its implicit solution policy to which to coordinate without any signaling message is: *play the equilibria of the Nash set, and only the Pareto-superior ones if there is at least one*. Hence, it is feasible to natively implement a Nash equilibrium multipath (NEMP) routing policy. When no Pareto-superior equilibria exist in the Nash set (as already mentioned, this can happen), NEMP is performed over all the equilibria. For example, in Fig. 4, 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.

B. Repeated Coordination

Given that the G Pareto frontier may not contain equilibria, in a repeated ClubMED context, an explicit coordination policy 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 policy 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.

Unselfish-Jump: Another policy is conceivable to guarantee a state of balance in gains in the short term, and thus helping to keep a high level of reciprocal trust. After shrinking the Nash

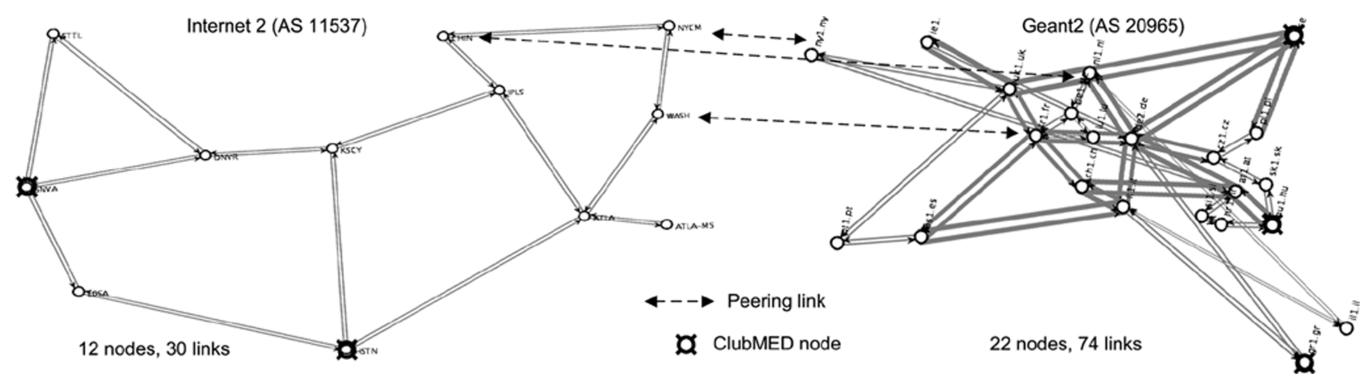


Fig. 5. Internet2–Geant2 peering scenario with three peering links.

set with respect to the Pareto efficiency, for each equilibrium the ASs might agree to make both a further step toward 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 policy makes sense only if the other AS is compensated with a bigger improvement and returns the favor in the future.

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 unselfish sacrifices. If at least one (x^j, y^j) is found, we obtain a new profile set that is to be shrunken with respect to the Pareto superiority for the final solution. In Fig. 4, e.g., 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 Unselfish-Jump policy, which would instead prefer (l_1, l_1) with a global gain of 6 instead of "just" 3 with (l_1, l_3).

It is worth noting that the two Jump policies are not binding: the implicit threat to change to one of the more selfish choices is enough. Moreover, with the Jump policies, we assume that MEDs from different ASs are normalized to the same IGP weight scale in order to be comparable. Finally, also note that we have a decreasing level of collaboration (thus trust), starting with a high level for "Pareto-Frontier," lower for "Unselfish-Jump," still lower for "Pareto-Jump" and basic coordination with "NEMP."

V. Performance Evaluation

We evaluated the performance of the PEMP routing policies with realistic simulations. We created a virtual interconnection scenario among the Geant2 and the Internet2 ASs, depicted in Fig. 5, emulating their existing peering with $n = 3$ cross-Atlantic links. We considered $m = 6$ 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 data sets from [20] for Geant2 and from [21] for Internet2 on an 8-h basis (to cover all the day times). The original link capacity was scaled by 10 to create an intra-AS congestion risk. The inter-cone routing generates additional volume 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.

We compare the PEMP routing policies ("NEMP," "Pareto-Frontier," "Pareto-Jump," "Unselfish-Jump") to the "BGP Multipath" solution without and with ("$\cdots+\mathrm{MED}$") classical MED signaling 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. 6 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 as described in Section II-A, without and with MED signaling enabled; the first PEMP policy, NEMP, without and with the congestion game G_c. For each method, we display the Internet2, the Geant2, and the global IGP 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 compared to normal BGP multipath, which balances the load only on equal cost paths (egress IGPs or MEDs). Curiously, the simple usage of the MED would decrease by 2% the cost of the BGP case without MED. This is probably due to the fact that for the most utilized network (Internet2), the ingress paths are more loaded than the egress one (hence, with higher ingress IGP weights), which leads to a lower global IGP cost. Another reason may be that the chance of doing 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 10% lower without G_c, and 6.6% in its complete form.

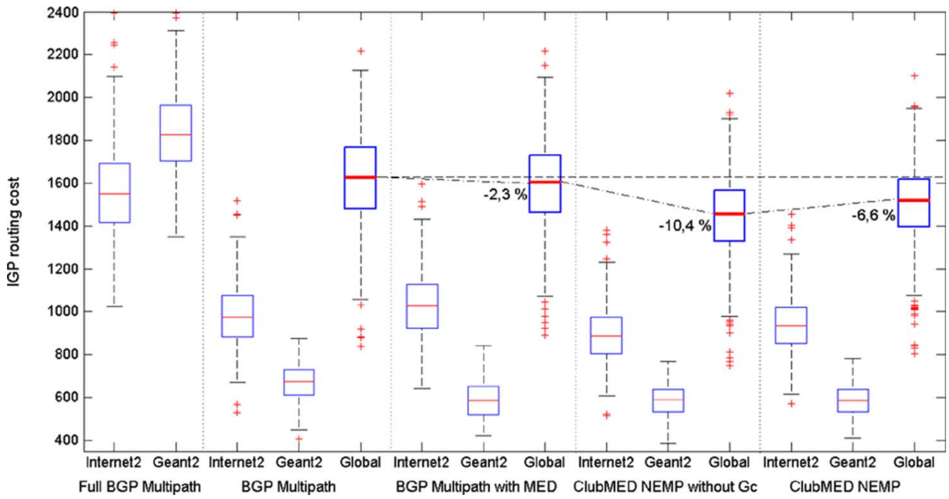


Fig. 6. IGP routing cost Boxplot statistics: NEMP versus BGP Multipath.

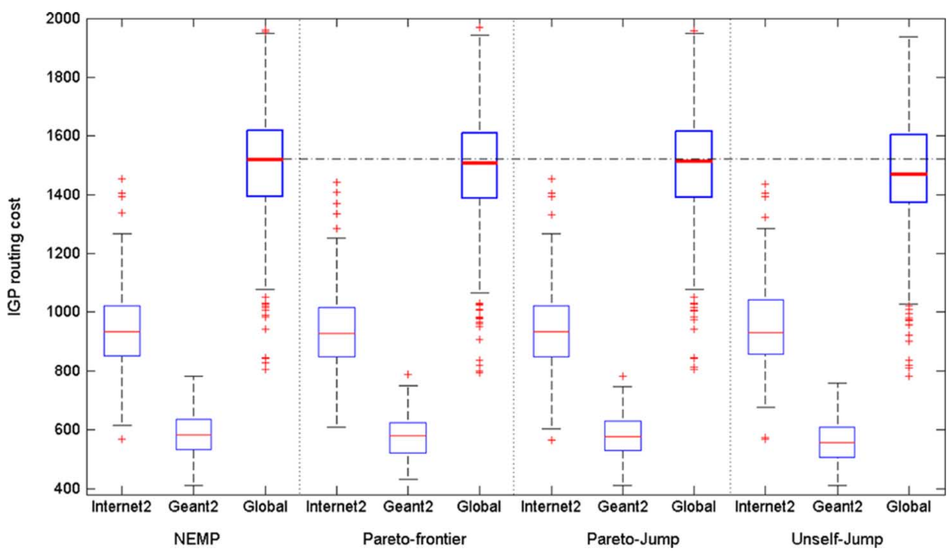


Fig. 7. IGP routing cost Boxplot statistics: PEMP strategies.

Fig. 7 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 for this poor performance is that the Pareto frontier often contains strategy profiles with the least cost for one 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-AS link may produce an IGP weight much lower than the others, thus affecting the G profile cost components. This risk is augmented in the Pareto-Jump policies 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 Section IV-B). Moreover, the Unselfish-Jump one is expected to outperform or equalize the Pareto-Jump strategy with respect to the routing cost since, without (5), it can be seen as its relaxation. Indeed, as reported in Fig. 7, the Unselfish-Jump gives a median cost roughly 3% inferior to the NEMP cost.

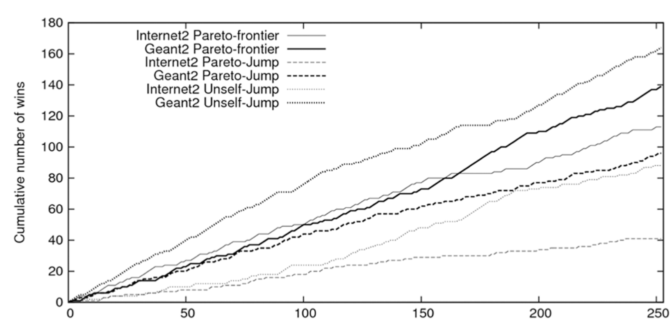


Fig. 8. Dynamics of the cumulative number of wins.

Fig. 8 further compares the Pareto-Frontier and the Jump policies in terms of fairness in routing cost in the long run. The horizontal axis is the round, i.e., a repetition of the ClubMED game

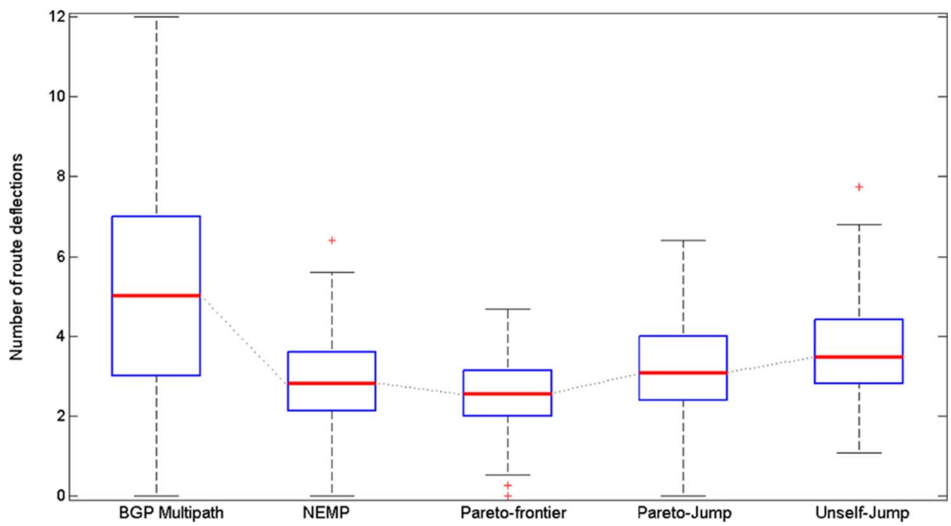


Fig. 9. Number of route deviations.

with a new traffic matrix. The vertical axis displays the cumulative number of times in which a peer obtained a percentage gain with respect to the NEMP solution [e.g., for the first peer, $(\phi_{\mathrm{NEMP}}(x,y) - \phi_{\mathrm{PEMP-str}}(x,y))/\phi_{\mathrm{NEMP}}(x,y)$] bigger than that of the other peer. In this way, we can assess how fair is the solution of the repeated game in the long run. Even if Geant2 is the final winner always, as expected, the Pareto-Frontier policy shows itself the most fair one. This can be measured by the difference between the Internet2 and the Geant2 lines: the lowest with Pareto-Frontier, the highest with Unselfish-Jump.

B. Route Deviations

Fig. 9 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 three route deviations against five, and the upper quartile and the maximum are 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, $n^{2m} = 531\,441$) offers a finer selection than the approximate potential threshold one. Finally, the Jump policies present a lower route stability with respect to all the statistical criteria. This is probably due to the fact that the jump from the Nash set is done without considering the cost errors.

As mentioned, the original link capacity of both networks was scaled by 10 to create an intra-AS congestion risk. It is interesting to observe how much the ClubMED framework can improve the route stability under "normal conditions" in which an operator's network is largely overdimensioned.

In Fig. 10, we report the route deviations' number dynamics, together with the corresponding Boxplot statistics, obtained rerunning the simulations with original intra-AS link capacities, comparing the BGP solution to the ClubMED NEMP solution. The median of route deviations with NEMP falls to 0. The

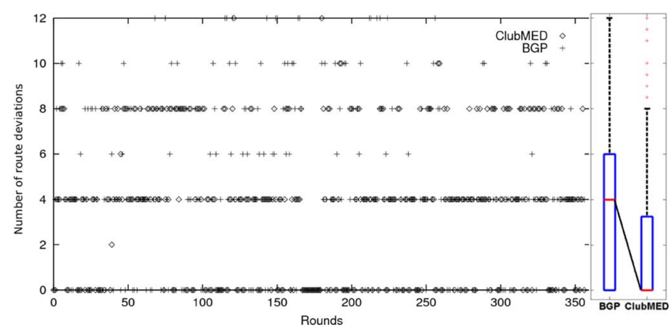


Fig. 10. Number of route deviations with original link capacities.

reason for this very good performance is related to the IGP-WO algorithm used to set the IGP weights. The IGP-WO cost function (such as the one implemented in TOTEM) assigns weights as function of the expected load, so that with loads below 50%, the variation in weight assignment is very low, while it increases more than exponentially as the load approaches 100%. Therefore, we verified that ClubMED works even better with high available networks whose link IGP weight variations are contained.

C. Peering Link Congestion

Fig. 11 reports the Boxplot statistics maximum link utilization as seen by each peer with the five above-mentioned methods. The PEMP policies except 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 Unselfish-Jump policies show the median and the upper and lower quartiles always above 85%, remembering that with full BGP Multipath, one would have the best $200/300 = 66.7\%$ utilization. The Pareto-Frontier policy does not guarantee, however, a congestion-free solution, with a median close to 100% utilization. The reason for this behavior is again the

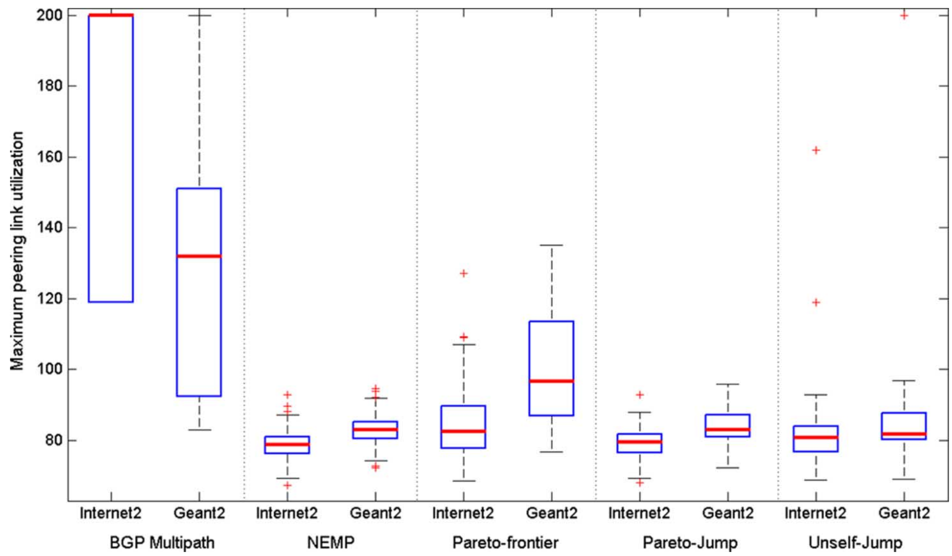


Fig. 11. Maximum peering link utilization boxplot statistics.

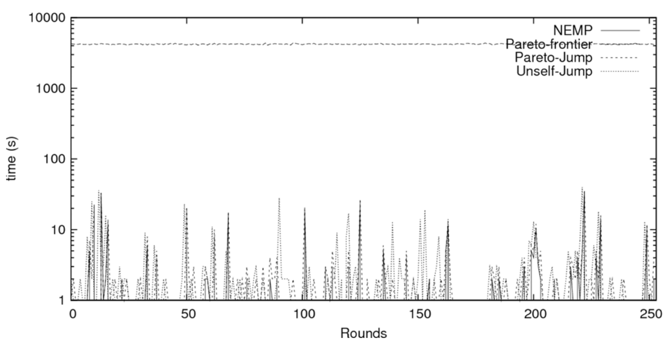


Fig. 12. PEMP strategies execution time.

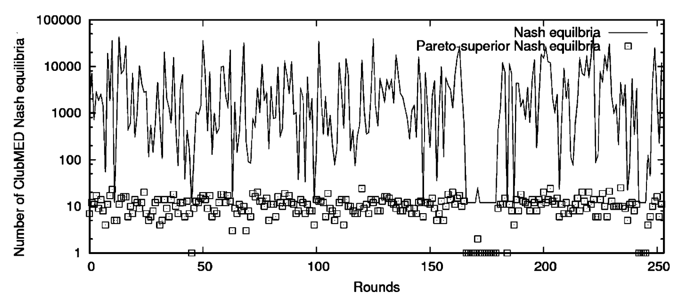


Fig. 13. Nash set dynamics.

Pareto-superiority condition that may introduce highly asymmetric cost profiles in the multipath routing solution.

D. Time Complexity

Fig. 12 reports the execution time for the PEMP policies. As expected, the Pareto-Frontier computation is excessively complex, with a $O(n^{2m})$ time complexity (n is the number of peering links, m the number of community pairs). The other policies have, instead, a polynomial complexity since they asymptotically depend on the minimization of a (monodimension) potential function to populate the Nash set. In fact, the other policies have an average computation time below 2 s (however, rare peaks of a few more seconds appear, probably due to the cases with very large Nash set, as can be seen cross-checking with Fig. 13). Hence, only the NEMP, Pareto-Jump, and Unselfish-Jump policies shall be considered for a practical implementation (we have however introduced the Pareto-Frontier case for a thorough comparison). Their execution times are acceptable in so much as the routing policies are computed after each IGP-WO, which can take much more time.

E. Nash Equilibrium Dynamics

Fig. 13 reports the number of ClubMED Nash equilibria and those Pareto-superior in a log-scale for all the rounds. The Pareto-superiority condition picks a few efficient Nash equilibria over broad sets, whose dimension varies significantly in time. This reveals a high sensitivity to the routing costs, probably due to the endogenous effect of G_c with high congestion costs; in fact, G_c cost components are not taken into account by the IGP ϵ errors.

VI. Implementation Aspects

The proposed framework does not require new protocol definitions or invasive extensions of existing ones. As partially already mentioned, there are important assumptions, possible intra-AS routing issues, and ClubMED operations aspects that we discuss hereafter.

A. Technical Assumptions

An assumption is that at each border a network management system is present to estimate traffic matrix, run IGP-WO, and update IGP weights, as it happens nowadays for large commercial Internet carriers. Nevertheless, from an algorithmic standpoint, the management operations or the IGP-WO algorithm behind IGP weight reconfiguration remain arbitrary and unilateral choices. The requirement is that the ASs exchange IGP path cost variation information via the ϵ errors.

Moreover, the decision on the destination cones to include in the ClubMED communities should rely on an initial setting agreement between the peers. Such an agreement should also contain the scaling rules for the IGP weights (needed, however, only for the repeated coordination PEMP policies), which is particularly important especially for large providers that want to apply ClubMED with a large number of peers.

B. Routing and Signaling

There are some routing and signaling aspects that relate to the following.

1) At the peering nodes:
   a) the coding of multiple subattributes (ingress and egress IGP path costs, cost errors) into the MED attribute for the networks belonging to the destination cones (i.e., the prefix belonging to the ClubMED destination communities)—the new MED subattributes shall pass opaquely across intra-AS routers;
   b) the usage of the MED may be adapted on a per-community identifier fashion rather than on a per-prefix fashion, so as to aggregate the MED information; the community identifiers can in fact pass the AS frontier (i.e., no community strip operations on the prefix belonging to the ClubMED destination cones).
2) At the ClubMED nodes:
   a) the modification of the BGP decision process at the "least MED" stage to select the multipath PEMP solution;
   b) the collection of the interpeer flow bit-rate information for the congestion game (we assume that some metrology infrastructure, e.g., Netflow, is available).
3) With an IBGP AS core, there is no guarantee that at least one MED-icated route for each peering link will be visible at the ClubMED nodes, and (vice versa) that at least one route per ClubMED node will be visible at the peering nodes; let us call both kinds of routes "ClubMED routes." This can happen in some corner cases, in particular when some internal router compared ClubMED routes and announced only the best (with shortest IGP path cost) one. It is worth remarking, however, that the same issues would be present with BGP Multipath, and that in our simulations these route limitation cases were not considered (which actually yielded better than real solutions for BGP Multipath).

   Nonetheless, to deal with such corner cases, a BGP-friendly approach would be to limit the strategy set of a player to the "visible" peering links at each ClubMED node; however, in the absence of specific signaling among each peer's ClubMED nodes announcing which peering links each peer considers in the strategy set, we would have a game with incomplete information in which the strategy sets considered by the peers are not completely known. The ClubMED game with incomplete information, even if respectful of BGP, may no longer be as effective as with complete information since a probability distribution shall be used by each peer over the different types of players (number and type of strategies) it could experience (see [16]).
4) In the case of configuration of BGP route reflectors (RRs), the visibility issue described previously could be even more important. Moreover, ClubMED nodes should not behave as normal RR clients for the networks belonging to the ClubMED destination cones.

Let us further discuss the implementation aspects 3) and 4), pointing out the correlated signaling issues that should be tackled to avoid the incompleteness of the game information and to deal with RRs.

i) Only the ClubMED nodes play the game, not intermediate intra-AS routers (those in between ClubMED nodes and peering routers).
ii) The ClubMED nodes should learn all the different peering routes in order to play the ClubMED game (avoid having only best paths).
iii) Intermediate nodes should forward packets to the proper egress router (without playing the game).

With respect to i), there is no scalability issue in that only a few AS border routers are likely to be elected ClubMED nodes even in large networks.

With respect to ii) and iii), ClubMED nodes could just have configured an IBGP direct session with the peering routers. With a BGP-free-core, i.e., direct BGP sessions only among AS border routers and an MPLS-managed AS core, the game could be played in its complete form. If RRs are configured, since their normal setting contrasts with i) and ii), they should announce the several routes with the same AS path to ClubMED nodes, at least for the routes whose prefix belongs to the ClubMED destination cones.

C. Execution Policy

It is worth stressing that the ClubMED game does not require an execution of an IGP-WO for the computation of each PEMP solution. The G_c components do not depend on IGP metrics and can be updated when a peering link fail or when the intercommunity flow bit-rate (μ_h) change.

There is no need to compute the PEMP solution after each IGP-WO or after each inter-community flow bit-rate variation. An appropriate execution policy, to be defined in a further work, should be able to assess the opportunity to rerun the PEMP computation at each side with respect to IGP weights and inter-community flow bit-rate variations.

D. Dealing With Cheating Behaviors

It is possible to configure the peering border routers to send not the real IGP path costs, but artificial costs, but this should pass via an OS reconfiguration. Announcing false MEDs could allow attracting more convenient equilibria for the cheating peer. However, since the two ASs are not obliged to peer, the result of such a malicious behavior could be a de-peering, which would be bad for the cheating ASs. In fact, non-cooperative interactions with cheating normally make sense only when the two payers have to play. In our case, there remains the threat to stop playing if such a malicious behavior is detected.

VII. Conception of Internet Extended Peering

Within the ClubMED framework, it is thus possible to efficiently control the route deviations by fine-tuning the routing strategy. The major practical benefits from the implementation of the PEMP policies would be the trust-reinforcement of an existing peering agreement and the improvement of the provided QoS related to the lack of congestions and frequent deviations. Let us try to see what happens if we generalize the two-player game to an n-player game to then discuss the practical incentives for such a novel model.

The extension of the ClubMED framework to more than two players could allow the definition of a sort of "extended peering" in which the border one provider has with the other neighbors (peers or candidate peers) is modeled as a single border. Please note that this differs from many sibling settlements (any flow is routed across sibling borders). In an extended peering, only the peers' client traffic would be routed across the peering borders. In order to treat multipeer borders as a single equivalent peering border, (in the extended framework) transit costs at each peer—from each neighbor to every other neighbor—shall be considered in the game modeling.

Referring to the three-peer scenario in Fig. 14, e.g., AS I announces the destination prefixes of community A to AS II with the MED set to the intra-AS I routing cost c_1^{I}. AS II in turn announces the same prefixes to AS III. In such announcements, a composite MED is to be coded including the individual routing costs that the selection of the link l_3 by AS III would cause to the ASs in the extended peering chain, thus in this case to AS II ($c_{3,1}^{\mathrm{II}}$) and to AS I (c_1^{I}).[2] Instead, the routing cost toward local communities, e.g., c_3^{II} for AS II, is sent via a normally MED-icated announcement. AS III disposes of two routes toward the community A, one through the direct link l_2 with AS I, one crossing AS II. Being aware of the costs that its routing decision causes to the other peers (given by the composite MEDs), the router R_{c} of AS III decides consistently with the extended peering game strategy profiles. R_{c} decides toward what peering link to route the aggregate $\mathrm{C} \rightarrow \mathrm{A} + \mathrm{B}$ flow, aware of the routing costs it implies for AS II (transit cost $c_{3,1}^{\mathrm{II}}$, from l_3 to l_1 for $\mathrm{C} \rightarrow \mathrm{A}$, and c_3^{II} for $\mathrm{C} \rightarrow \mathrm{B}$) and for AS I ($c_1^{\mathrm{I}}$).

In the general case, many peering links can connect two peers. Moreover, many ASs can transit traffic toward the same destination community, and the AS chain lengths within the extended peering vary. While inter-community routing is distributed at the edge routers (e.g., R_a, R_b, and R_c) following the extended peering game (thus bypassing BGP), transit routing decisions are instead taken at the peering routers (e.g., R_1^{II} and R_3^{II}) following the normal BGP routing policy for the ingress peering flows (without specific route filtering). The peer routing costs, which depend on the peering router's decisions, are to be coded in the composite MED sent to the neighbors. For those MEDs that are composite, the smaller MED rule shall be applied to the sum of all the MED parts. Finally, it is possible that many "MED-icated" routes from different ASs have the same AS hop count. In such a case, MEDs from different ASs shall be normalized over a same IGP weight scale.

[2]In the composite MED, the reference to the selected peering links over the interpeer path is lost. Only the routing costs impacted to the peers matter.

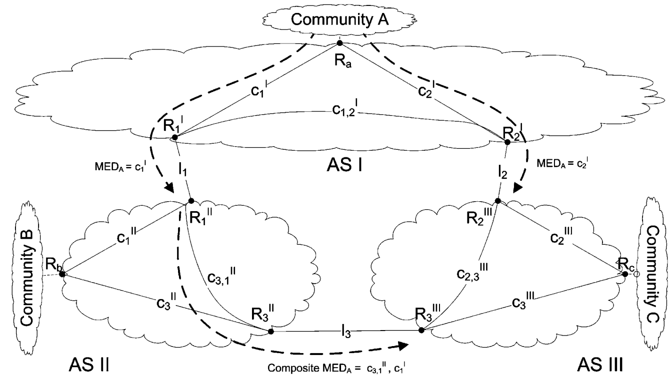


Fig. 14. Extended peering scenario with three peers (for simplicity, only AS I MED signaling and simple bidirectional costs are depicted).

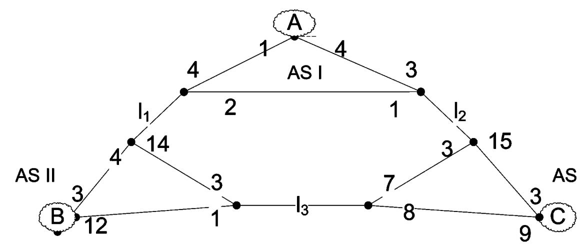


Fig. 15. Extended peering game example.

A. Extended Peering Game

The extended peering game is a straightforward extension of the two-player game.

- The number of strategies increases due to the enlarged interconnection.
- G_{s} and G_{c} maintain the same structure.
- G_{d} includes also the exogenous transit costs toward the external destination communities; a transit cost is simply summed to the ingress cost for the internal destination community.

In Fig. 15, there is a so-built extended peering game example with three carriers, with just one link connecting two peers, and without G_{c}; the corresponding strategic form is in Table I. The decision of routing on a link impacts an egress cost for the deciding AS, an ingress cost and a transit cost for the next AS, and an ingress cost for the last AS; three routing decisions must be taken at community edge routers, one for $\mathrm{A} \rightarrow \mathrm{B} + \mathrm{C}$ flows by AS I, one for $\mathrm{B} \rightarrow \mathrm{A}+\mathrm{C}$ flows by AS II, and one for $\mathrm{C} \rightarrow \mathrm{A}+\mathrm{B}$ flows by AS III. In Table I, there is the strategic form of the game; each cell corresponds to a strategy profile and indicates between brackets the routing cost for AS I, AS II, and AS III, in the listed order. Each peer strategy corresponds to a possible link where to route its own egress flow, thus, e.g., l_1 and l_2 for AS I, l_1 and l_3 for AS II, and l_2 and l_3 for AS III. We have a Nash equilibrium in (l_2, l_3, l_3). Similarly to two-player games, it is possible to have profiles Pareto-superior to the equilibrium (or equilibria), such as (l_1, l_1, l_3), that grants a lower cost for all ASs.

Under the assumption that the IGP costs of all the carriers involved in the extended peering are normalized to the same scale, the PEMP polices should be used to coordinate the extended peering routing strategy. Given that the extended peering game

TABLE I
STRATEGIC FORM OF FIG. 15 EXAMPLE

III	I \ II	l_1	l_3
l_2	l_1	(12,13,27)	(14,10,36)
	l_2	(11,19,28)	(13,16,37)
l_3	l_1	(7,36,20)	(9,33,29)
	l_2	(6,42,21)	**(8,39,30)**

is a straightforward extension of the two-player one, we expect similar results and benefits for this framework (no simulation results will be provided herein).

B. *Incentives for an Extended Peering*

The incentives for implementing an extended peering coordination framework are not straightforward. The alternative closer solution would be a full mesh of classical bilateral peering agreements. With respect to the best case with a full mesh of ClubMED-based bilateral peering agreements, an extended peering framework would have the following key advantages.

- *Extended balance*: It may be easier to agree on a peering among many carriers rather than among only two carriers since, e.g., the traffic balance condition may be reached more easily by considering a larger set of flows.
- *Higher Internet reliability*: Congestions or outages at one peering border or de-peerings can be surrounded by an automatic rerouting of the traffic elsewhere within the extended peering settlement without losing visibility toward a piece of the Internet.
- *Larger path diversity*: The resulting increased path diversity can further improve the efficiency of the peering routing solution (with respect to routing costs, congestions, and deviations).

Nevertheless, these incentives may be too weak because of not appealing enough especially for those top-tier providers for which the existing peering settlements are well balanced, for which the reliability is not a relevant issue (with a number of peerings with high cone overlapping), and which would see the extended peering management too cumbersome already with a discrete number of communities.

The "killer incentive" for a generic form of extended peering might be additional revenues related to novel added-value interprovider services. The framework may indeed be used to differentiate the treatment of added-value services overlapping best-effort routing.

VIII. SUMMARY

We modeled the routing on peering links as a non-cooperative game with the aim of allowing carriers to fine-select 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 reoptimizations.

We proposed a low-computational way to compute the Nash equilibria as well as the four possible peering equilibrium multipath (PEMP) routing coordination policies. The first two 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 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 outperform 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 others present some tradeoffs, but all represent promising solutions to perform an efficient and rational routing across peering links. In particular, the Unselfish-Jump policy represents the best tradeoff between peering trust insurance, routing cost, congestion control, routing stability, and execution time.

Finally, we discussed practical implementation aspects. Moreover, we showed that the extension of the framework to an arbitrary number of provider-players can be done straightforwardly by modeling additional transit routing cost in the dummy game. Such an extension might allow the definition of extended peering models that could increase the Internet reliability at the expense of some additional complexity at the border routers.

Our work represents a step toward the definition of peering management frameworks to improve the routing where the real Internet bottleneck is located. The critical situation of peering interconnections nowadays manifests with new forms of peering called "paid peering," in which two peering carriers agree for monetary compensations in case the traffic become excessively unbalanced. Isolating critical flows and managing them in a dedicated framework, as the PEMP one, might allow escaping these astray agreements and reaching acceptable and viable peering situations.

APPENDIX A
PRISONER DILEMMA AND POTENTIAL GAMES

We provide in this Appendix a brief "tutorial" on how to decompose a prisoner's dilemma game as the sum of two interesting types of games (extracted from [22]). Consider the generic symmetric game in Table II, where $a, b, c, d \in \mathbb{R}$. We have a prisoner dilemma *cost* game if $a > b > c > d$, with (B, R) as Nash equilibrium, inefficient since both would prefer (T, L), which is however a dominated strategy profile. Indeed, this is the rationality dilemma offered by such games.

The game can be decomposed as the sum of the two games shown in Table III. For the first game, the cost components for the two players are equal for every profile. For the second game,

TABLE II
GENERIC TWO-PLAYER SYMMETRIC GAME

I\II	L	R
T	(c,c)	(a,d)
B	(d,a)	(b,b)

TABLE III
DECOMPOSITION OF A TWO-PLAYER SYMMETRIC GAME

I\II	L	R
T	$(0,0)$	$(d-c,d-c)$
B	$(d-c,d-c)$	$(d-c+b-a,d-c+b-a)$

I\II	L	R
T	(c,c)	$(a-d+c,c)$
B	$(c,a-d+c)$	$(a-d+c,a-d+c)$

TABLE IV
DECOMPOSITION OF A TWO-PLAYER SYMMETRIC GAME

I\II	L	R
T	$(0,0)$	$(-1,-1)$
B	$(-1,-1)$	$(-2,-2)$

I\II	L	R
T	$(2,2)$	$(5,2)$
B	$(2,5)$	$(5,5)$

the cost components of a player do not depend on its choice, but they depend on the other player's choice. The second game can be called "dummy game" since, for a player, there is no possible discrimination in choosing one strategy instead of the other. It can also be called "game of pure externality," meaning that its action has an effect *only* on the other player. This type of decomposition allows to clearly see the externality effect in the prisoner dilemma game.

With the setting $a=4, b=3, c=2, d=1$, we obtain the game decomposition in Table IV. The choice of *B* allows to decrease the cost of I by 1, independently of the choice of II. At the same time, this choice increases by 3 the cost of II in the second game. It is worth noting that, in the first game, the costs are equal for the two players and that the choice of *B* has a positive externality effect for II: It also decreases by 1 its cost. Clearly, inefficiency stems from the fact that externalities prevail upon selfish improvements.

With a broader perspective, one can note that such a decomposition is a general property of the so-called potential games [17]. For a game in strategic form $G=(X,Y,f,g)$, where X and Y are the strategy sets for the two players, and f and g are real functions, G admits a potential if it exists a function $P: X \times Y \to \mathbb{R}$ such that $\forall x', x'', x \in X, \forall y', y'', y \in Y$

$$P(x',y)-P(x'',y)=f(x',y)-f(x'',y)$$
$$P(x,y')-P(x,y'')=f(x,y')-f(x,y''). \qquad (6)$$

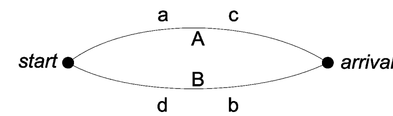


Fig. 16. Representation of a two-player symmetric game.

P is called *potential function*. The analogy with physics relates, e.g., to the ability to substitute a "vector field" (the two payoff functions) with a single scalar valued function or to the condition of being an irrotational field. Minima of the potential function are Nash equilibria for the game, which guarantees that finite potential games have equilibria in *pure* strategies [17].

Potential games emerge from congestion problems [23]. Indeed, we can represent the game of Table II with Fig. 16. Both players have to go from *start* to *arrival* taking either path A or path B (strategy A corresponds to T for I and to L for II, B corresponds to B for I and to R for II). The lowercase letters on each path in Fig. 16 indicate the transit cost for the players in case they walk alone (on the left) or together (on the right). If they travel together on the same path, the path is more congested than if they traveled alone along different paths, i.e., the cost is higher for both.

APPENDIX B
ON THE CLUBMED NASH EQUILIBRIUM

In potential games, the potential function minimum corresponds to a Nash equilibrium, but the inverse is not necessarily true. The next theorem proves that the inverse is also true for the ClubMED game G defined in Section III-A.

Theorem B.1: A ClubMED Nash equilibrium corresponds to the strategy profile with minimum potential.

Proof: If (x^*,y^*) is an equilibrium, $P(x^*,y^*) \leq P(x,y^*), \forall x \in X$. However, $P(x^*,y^*)=\phi_s(x^*)-\phi_s(x_0)$ and $P(x,y^*)=\phi_s(x)-\phi_s(x_0), \forall x \in X$. Thus, $P(x^*,y^*) \leq P(x,y^*), \forall x \in X$, is equivalent to $\phi_s(x^*)-\phi_s(x_0) \leq \phi_s(x)-\phi_s(x_0), \forall x \in X$, that is $\phi_s(x^*) \leq \phi_s(x), \forall x \in X$. Hence, x^* is a minimum for ϕ_s. Idem for y^*. Therefore, $P(x^*,y^*)=0$, which is a minimum of P. ■

Given that $P=P_s, G_s$ fully guides the G Nash equilibrium.

REFERENCES

[1] S. Secci, J.-L. Rougier, A. Pattavina, F. Patrone, and G. Maier, "PEMP: Peering equilibrium multipath routing," in *Proc. 2009 IEEE GLOBECOM*, Honolulu, HI, Nov. 30–Dec. 4, 2009, pp. 1–7.
[2] R. Teixera, K. Marzullo, S. Savage, and G. M. Voelker, "In search of path diversity in ISP networks," in *Proc. ACM Internet Meas. Conf.*, Oct. 2003, pp. 313–318.
[3] J. He and J. Rexford, "Toward Internet-wide multipath routing," *IEEE Netw.*, vol. 22, no. 2, pp. 16–21, Mar. 2008.
[4] JUNOS, "Configuring BGP to select multiple BGP paths," Sunnyvale, CA, JUNOS Internet Software Configuration Guide, 2005 [Online]. Available: http://www.juniper.net/techpubs/software/junos/swconfig53-ipv6/html/ipv6-bgp-config29.html
[5] Cisco Systems, "BGP best path selection algorithm," San Jose, CA, Doc. ID 13753, 2006.
[6] Y. Rekhter and T. Li, "A Border Gateway Protocol 4 (BGP-4)," RFC 1771, 1995.
[7] D. McPherson and V. Gill, "BGP MED considerations," RFC 4451, 2006.
[8] A. Lange, "Issues in revising BGP-4," Draft-Ietf-Idr-Bgp-Issues, 2003.

[9] R. Teixeira, T. G. Griffin, M. G. C. Resende, and J. Rexford, "TIE breaking: Tunable interdomain egress selection," in *Proc. CoNEXT*, 2005, pp. 93–104.
[10] S. Agarwal, A. Nucci, and S. Bhattacharyya, "Controlling hot potatoes in intradomain traffic engineering," SPRINT RR04-ATL-070677, 2004.
[11] S. Balon and G. Leduc, "Combined intra- and inter-domain traffic engineering using hot-potato aware link weights optimization," in *Proc. ACM SIGMETRICS*, 2008, pp. 441–442.
[12] M. Latapy, C. Magnien, and F. Ouédraogo, "A radar for the Internet," in *Proc. IEEE ICDMW*, 2008, pp. 901–908.
[13] E. Elena, J.-L. Rougier, and S. Secci, "Characterisation of AS-level path deviations and multipath in Internet routing," in *Proc. NGI*, 2010, pp. 1–7.
[14] R. Teixeira, A. Shaikh, T. G. Griffin, and J. Rexford, "Impact of hot-potato routing changes in IP networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 6, pp. 1295–1307, Dec. 2008.
[15] P. Casas, L. Fillatre, and S. Vaton, "Multi hour robust routing and fast load change detection for traffic engineering," in *Proc. IEEE ICC*, 2008, pp. 5777–5782.
[16] R. B. Myerson*, Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard Univ. Press, 1997.
[17] D. Monderer and L. S. Shapley, "Potential games," *Games Econ. Behav.*, vol. 14, no. 1, pp. 124–143, May 1996.
[18] F. Larroca and J.-L. Rougier, "Routing games for traffic engineering," in *Proc. IEEE ICC*, 2009, pp. 1–6.
[19] S. Balon, J. Lepropre, O. Delcourt, F. Skivee, and G. Leduc, "Traffic engineering an operational network with the TOTEM toolbox," *IEEE Trans. Netw. Service Manag.*, vol. 4, no. 1, pp. 51–61, Jun. 2007.
[20] S. Uhlig, B. Quoitin, J. Lepropre, and S. Balon, "Providing public intradomain traffic matrices to the research community," *Comput. Commun. Rev.*, vol. 36, no. 1, pp. 83–86, 2006.
[21] Y. Zhang, "Abilene topology and traffic dataset," 2004 [Online]. Available: http://userweb.cs.utexas.edu/~yzhang/research/AbileneTM
[22] F. Patrone, "Giochi con potenziale," *(Potential Games) Lettera Matematica (Pristem)*, vol. 69, pp. 17–19, 2009.
[23] R. W. Rosenthal, "A class of games possessing pure-strategy Nash equilibria," *Int. J. Game Theory*, vol. 2, pp. 65–67, 1973.

Stefano Secci (S'05–M'10) received the M.Sc. degree in communications engineering from Politecnico di Milano, Milan, Italy, in 2005, and Ph.D. degrees in networks and computer science from Politecnico di Milano and Telecom ParisTech, Paris, France, in 2009.

He is an Associate Professor at the LIP6, Université Pierre et Marie Curie (UPMC–Paris VI–Sorbonne Universités), Paris, France. He worked as Post-Doctoral Research Associate with Telecom ParisTech, then NTNU, Trondheim, Norway, and George Mason University, Fairfax, VA. Before receiving the Ph.D. degrees, he worked as a Network Engineer with Fastweb Italia, Milan, Italy, and as a Research Associate with Ecole Polytechnique de Montréal, Montréal, QC, Canada, and with Politecnico di Milano. He actively participated in the ANR ACTRICE French project, the EU FP7 EuroNF and ETICS projects, the EU CELTIC TIGER2 project, and a US ONR project on future Internet. He is the author of more than 20 articles in major international journals and conferences. His works space from optical network planning and switching to IP routing optimization and traffic engineering. His current research interests are about future Internet routing resiliency, mobility, and policy.

Dr. Secci is the recipient of the NGI 2009 Best Paper Award. He has been a member of the Technical Program Committee of IEEE ICC 2011, IEEE GLOBECOM 2010, NGI 2010, and CFIP 2009, and has been a reviewer on a regular basis for major international journals and conferences in networking.

Jean-Louis Rougier received the Engineering diploma the Ph.D. degree in networks and computer science from Télécom ParisTech (formerly called École Nationale Supérieure des Télécommunications), Paris, France, in 1996 and 1999, respectively.

He joined the Computer Science and Networks Department of Télécom ParisTech in 2000 as an Associate Professor. His research interests are performance evaluation and algorithms, traffic engineering, and routing in networks.

Achille Pattavina (M'85–SM'93) received the Dr.Eng. degree in electronic engineering from the University La Sapienza of Rome, Rome, Italy, in 1977.

He was with the University La Sapienza of Rome until 1991, when he moved to the Politecnico di Milano, Milan, Italy, where he is now a Full Professor. He has been the author of more than 100 papers in the area of communications networks published in leading international journals and conference proceedings. He has authored two books, *Switching Theory, Architectures and Performance in Broadband ATM Networks* (Wiley, 1998) and *Communication Networks* (McGraw-Hill, 1st ed. 2002, 2nd ed. 2007, in Italian). He has been engaged in many research activities, including European Union-funded projects. His current research interests are in the areas of optical switching and networking, traffic modeling, and multilayer network design.

Dr. Pattavina has been Guest or Co-Guest Editor of special issues on switching architectures in IEEE and non-IEEE journals. He has been an Editor for switching architecture performance for the IEEE TRANSACTIONS ON COMMUNICATIONS since 1994, and Editor-in-Chief of the *European Transactions on Telecommunications* since 2001.

Fioravante Patrone received the M.Sc. degree in mathematics from the University of Genova, Genova, Italy, in 1974.

Formerly a Full Professor of mathematical analysis, he is currently a Professor of game theory with the Faculty of Engineering, University of Genova. He has been the promoter of the "Game Practice" meetings that started in 1998 and has served as Director of the Interuniversity Centre for Game Theory and Applications, University of Genova, for many years. He is the author of more than 50 papers and is a member of the Editorial Board of international journals. He has recently published (in Italian) an introductory book for game theory. His recent research interests focus on applications of game theory to diverse fields: telecommunication, molecular biology, health care, natural resources.

Guido Maier (S'97–M'99–SM'08) received the Laurea degree in electronic engineering and the Ph.D. degree in telecommunications from the Politecnico di Milano, Milan, Italy in 1995 and 2000, respectively.

Until February 2006, he was a Researcher and head of the Optical Networking Laboratory at CoreCom, Milan, Italy. In March 2006, he joined Politecnico di Milano as an Assistant Professor. He is the author of more than 70 papers in international journals and conferences in the area of optical networks. He has been or is involved in several research projects, including BONE, MUPBED, and NOBEL2. His main interests are optical network optimization, multidomain ASON/GMPLS, and photonic switching systems.