ClubMED: Coordinated Multi-Exit Discriminator Strategies for Peering Carriers

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*—Inter-domain peering links represent nowadays the real bottleneck of the Internet. On peering links carriers may coordinate to efficiently balance the load, but the current practice is often based on an uncoordinated selfish routing supported by the peer relationship. We present a novel game theoretical routing framework to efficiently coordinate the routing on peering links while modelling the non-cooperative carrier behaviour. It relies on a collaborative use of the Multi-Exit Discriminator (MED) attribute of BGP, hence it is nicknamed ClubMED (Coordinated MED). The incentives are the minimization of carrier routing costs, the control of peering link congestions and peering route stability. For the ClubMED game, we define the Nash Equilibrium Multi-Path (NEMP) routing policy that shall be implemented upon Nash equilibria and Pareto-efficient profiles. Intra-domain Interior Gateway Protocol (IGP) weight optimizations are consistently integrated into the framework.**

We emulated the peering settlement between the Internet2 and the Geant2 networks, employing real datasets, comparing the ClubMED results to the current BGP practice. The results show that the global routing cost can be reduced of roughly 17%, that the peering link congestion can be avoided and that the stability of the routes can significantly be reinforced[1].

I. Introduction

The Internet backbone is composed of a few Autonomous Systems (ASs). To simplify, one may say it is composed of a few inter-continental carrier providers that provide transit connectivity to those regional providers the most part of customers and stub ASs are connected to. Some of the core ASs are "peers" when they agree on free reciprocal transit of their clients' IP flows. Peering agreements are usually signed when ASs get mutual operational and economical benefits from peering. The Border Gateway Protocol (BGP) v.4 is the current inter-AS IP routing protocol. It includes criteria that allow implementing peering settlements. As a matter of fact, the inefficient way in which these criteria are currently used overstresses inter-peer routing and overloads the peering links. This is mainly due to the unpredictability of the aggregate IP flows and to the fact that the free transit over peering links releases an AS from following the peer's routing preferences [1]. This yields to selfish routing while, instead, coordination schemes may improve the bilateral routing efficiency.

In previous work about inter-carrier connection-oriented services [2], it sorted out that a form of *cooperation* among carriers is needed to overtake privacy, billing and monitoring issues. In this paper we argue how, instead, for connectionless IP services - for which such issues are not present - cooperation is not necessary in that *coordination* is enough. In particular, we concentrate on the coordination issue for peering settlements to reduce congestions, routing cost and route deflections.

In Sect. II we link recent ideas in the area that motivated this work. We rely on the MED BGP attribute as the natural medium to convey coordination data. In Sect. III we define the ClubMED (Coordinated MED) framework, in which efficient strategy profiles can be detected in a non-cooperative game modelling. We define an effective routing policy relying on the concepts of Nash equilibria and Pareto-efficiency. We explain how, within the ClubMED framework, a form of load balancing can be implemented on selected strategy profiles for a subset of the destination networks whose traffic routing can be coordinated. We consistently integrate IGP weight optimization operations and peering link congestion controls, which increases the number of possible Nash equilibria and, thus, the importance of coordination schemes to select the most efficient ones. Sect. IV presents practical implementation aspects. Sect. V reports the results from realistic simulations and comments the significant gains the ClubMED framework can offer with respect to the current practice. Finally, Sect. VI summarizes the paper.

II. Rationales

A. BGP, route deflection and congestion

It is worth briefly reminding how the inter-AS route selection is performed via BGP. When multiple AS paths to a destination network prefix are available, a cascade of criteria is employed to compare them. The first is the "local preference" through which local policies, 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). Marking routes with local preferences, an AS can thus implement peering and transit settlements. The subsequent BGP criteria incorporate purely operational network issues: smaller AS hop count, smaller MED, closer egress point (also called "hot-potato"), more recent route. If not enough, the AS path learned by the router with the smaller IP is selected (rule also called "tie-breaking"). Considering these criteria, BGP selects the best AS path which is the single one advertised to the neighbours (if not filtered by local policies).

[1]Work funded by the INCAS S.JRA of the EU IST Euro-NF Network of Excellence and the ICF I-GATE project of the Institut Télécom, France.

978-1-4244-4245-4/09/$25.00 (c)2009 IEEE

Operationally speaking, carriers desire that AS paths pointing to them have been selected using the highest possible priority rule to obtain good performance, e.g., on the end-to-end delay for the connections along that AS path. The smaller AS hop count is a rude yet simple rule that avoids routing inefficiencies [3]. For a given AS path, if several border routes are available, the MED can be used by the downstream AS to suggest an egress router. However, it is rarely used: it is only for very specific cases or when requested by a client (see Sect. II-C). In the absence of MED settings, IGP weights are compared and the closer egress point is selected (hot-potato).

The interaction between hot-potato routing and intra-AS routing represents a major issue. To react to non-transient network events, a carrier may re-optimize the IGP weights, inducing changes in the egress router selection so that congestions might appear where not expected. [4] reformulates the egress routing problem and proposes to replace hot-potato with a more expressive and efficient rule. [5] presents a comprehensive yet hard IGP Weight Optimization (IGP-WO) method aware of BGP hot-potato routing deflections, opportunely bounding them (they report that in real cases 70% of traffic could be affected). [6] 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 would be thus a practical issue: the occurrence of IGP-WOs, normally triggered only for intra-AS issues, would drastically increase given the frequency of BGP deflections. The reduction of the coupling between inter-AS and intra-AS routing is thus really an open issue [1].

B. From selfish to coordinated inter-carrier traffic engineering

With a more far-sighted standpoint, in [7] it is proven that, if part of the profits due to inter-carrier services were shared, the Internet carriers would behave less selfishly, yielding better global routing with lower routing cost than under the current practice. Using the Shapley value concept from *cooperative* games, they argue that profits and costs may be fairly imputed considering the importance of each AS in the interconnected "coalition" composed of ASs routing "common" inter-AS flows [8]. In this way, they prove that ASs have incentive to better route yielding to a common inter-domain routing cost lower than with BGP routing.

More pragmatically, the authors in [9] show how much the hot-potato routing is far from being the ideal desirable solution. They compare it to a cooperative routing resulting from the maximization of a common utility for the two network configurations, i.e., the *bargaining problem* of the common utility, the (Nash) product of ASs' utilities, each one estimated somehow from the current intra-AS routing status (somehow withstanding possibly also a congestion risk, ignored in [7]). Then, the maximization of the common utility is solved by decomposition. However, to cope with multiple AS cooperative scenarios, their method should be at least re-designed given that the Nash product maximization solution - which can formally be extended[2] to the case of n/player games - does not take into account the role of subcoalitions [11] [12]. They also show how their method outperforms a generic best-reply "Nash equilibrium" method from non-cooperative games, however not detailed, thus preventing the possibility of making a comparison.

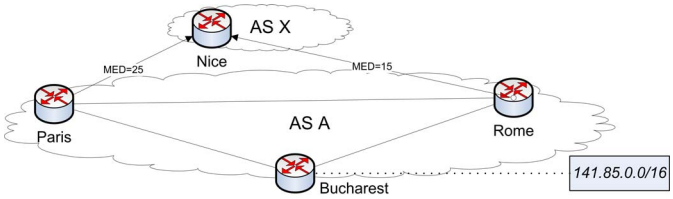


Fig. 1. Multi-Exit Discriminator signalling.

Modelling the inter-AS routing as a cooperative bargaining problem of a common utility may be not necessary. Besides appearing not enough expressive (hardly acceptable for operation engineers), the utility maximization result seems relying on an excessive abstraction of the real network status at the risk of loosing the real routing optimality. In other words, the way in which the utility is computed may not be enough representative of the real operational status of the network. We propose, instead, a non-cooperative approach since it allows more straightforward solutions to implement w.r.t. the current practice (both technically and economically). In the next sections we explain how using MED signalling it is possible to implement strategies that are, in game theory parlance, *non-cooperative but coordinated*, i.e., that solve the inter-peers routing problem without binding agreements between peers.

C. The Multi-Exit Discriminator (MED)

The MED is an integer 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 has the choice on the entry point toward the advertised network. In Fig. 1, the upstream AS X selects a route for the network 141.85.0.0/16. It has two route alternatives through AS A: by the Paris router or the Rome router. MEDs are attached to the routes announced by AS A's Paris and Rome routers. If accepting MEDs, the AS X router will then select the route with smaller MED, hence the route passing by the Rome router. The default MED value is equal to the IGP cost of the corresponding intra-AS path.

Nowadays, the MED is often disabled. Even if a downstream AS uses it to suggest preferred entry points, the neighbour can discard its announcements. The MED can be used on transit or peering links. On transit links, subject to provider/customer agreements, the provider should always follow "MED-icated" routes suggesting the preferred entry points because the customers pay for. This is not the case for peering agreement, and this is the main reason why the MED is usually not employed on peering links [13].

[2]Extensions to cope with the role of subcoalitions have been devised by Harsanyi and Shapley [10], but they are not easily manageable.

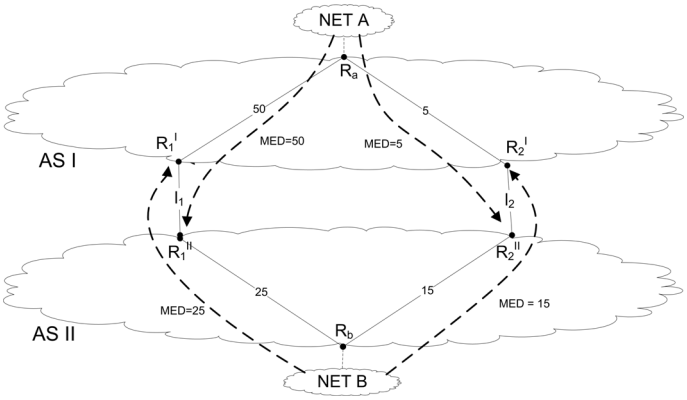


Fig. 2. Peering MED interaction example

TABLE I
A DUMMY GAME.

I \ II	l_1	l_2
l_1	**(50,25)**	**(5,25)**
l_2	**(50,15)**	**(5,15)**

TABLE II
A CLUBMED GAME.

I \ II	l_1	l_2
l_1	(100,50)	(55,40)
l_2	(55,40)	**(10,30)**

III. THE CLUBMED FRAMEWORK

We model the MED signalling between peering ASs as a non-cooperative game wherein peering ASs can implicitly coordinate their routing strategies. As of our knowledge, this is the first attempt in this direction. We nickname it the ClubMED (Coordinated MED) framework. For the sake of clarity, we first start with a simple but unrealistic model with 2 peering links and bidirectional routing costs. Then, we generalize it to the complete realistic generic form, integrating IGP-WO operations and peering link congestion controls.

A. MED-based coordination

In Fig. 2, AS I and AS II are two peers. NET A and NET B are two destination networks whose flows are supposed to be equivalent (e.g., w.r.t. the bandwidth), so that their path cost can be fairly compared and their routing coordinated. Each peer would desire to minimize its routing cost for the incoming flow. The routing costs are indicated in Fig. 2. AS I and AS II announce NET A and NET B with the MED attribute set to the routing cost by the corresponding egress router. The peering interaction can be described with the strategic form in Table I. The cost of each player is the MED of the route it announced, then selected by the peer. Each AS has the choice if routing the outgoing flow on link 1 (l_1) or on link 2 (l_2).

In non-cooperative games, a Nash equilibrium is to be selected by rational players because it yields stability to the strategy profile, the players not being motivated in deviating from it [11]. In Table I every profile is a Nash Equilbrium. We have a dummy game: whatever the other player's strategy is, there is no gain in changing its strategy. This somehow shows that a simple MED usage is dummy for such a case. We should enrich the dummy game considering the egress cost of the flow in the opposite direction, thus summing the routing costs of both the flows in opposite directions for each AS. However, in this way we would assume that both the NET A $\leftrightarrow$ NET B flows pass through the peering AS I-AS II, which would not be realistic (BGP policies can induce asymmetric routing). Moreover, traffic flows to care of are typically between content and "eyeball" providers (with a lot of clients) [8], which would not make the A $\leftrightarrow$ B flows equivalent. Instead of single prefix network, we should consider *destination cones* (i.e., groups of network prefixes). The cone prefixes shall belong to direct customers or stub ASs, whose entry point in a peer network is likely to be unique (even if they are multi-homed, they should have chosen backbone-disjoint providers, referring to disjoint core carriers; see Sect. IV for more practical aspects).

TABLE III
2-LINK CLUBMED GAME, SUM OF TWO GAMES WITH POTENTIAL.

I \ II	l_1	l_2
l_1	(c_1^I, c_1^{II})	(c_1^I, c_2^{II})
l_2	(c_2^I, c_1^{II})	(c_2^I, c_2^{II})

+

I \ II	l_1	l_2
l_1	(c_1^I, c_1^{II})	(c_2^I, c_1^{II})
l_2	(c_1^I, c_2^{II})	(c_2^I, c_2^{II})

$$\begin{pmatrix} 0 & c_1^{II} - c_2^{II} \\ c_1^I - c_2^I & c_1^{II} - c_2^{II} + c_1^I - c_2^I \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Therefore, in the complete strategic form in Table II, each AS sums the costs due to the two community A $\leftrightarrow$ community B flows. (l_2, l_2) is the unique Nash equilibrium. Hence, rational ASs would implicitly coordinate as suggested by (l_2, l_2), which in this case corresponds to accept the suggestion to routing the flow toward the neighbor's preferred egress router, and moreover to route alike hot-potato routing. Swapping e.g. the R_a-R_1^I and R_b-R_2^I IGP path costs (in Fig. 2) it is easy to verify that the Nash equilibrium is (l_1, l_2) with costs (55,40). In this case (l_2, l_1) has costs equal to (l_1, l_2), but it is not an equilibrium; ClubMED still behaves as hot-potato routing, but in this case the MEDs of AS I are not respected by AS II.

Let c_i^I and c_i^{II} be the IGP costs between R_a and R_b (resp.) and l_i, $i \in E$. For the generic case of two inter-AS links, the cost vector for the strategy profile (l_i, l_j), $i, j \in \{1, 2\}$, is thus $(c_i^I + c_j^I,\ c_i^{II} + c_j^{II})$. The resulting ClubMED game (Table III) can be described as $G = G_s + G_d$, sum of two games. $G_s = (X, Y, f_s, g_s)$, a selfish game, purely endogenous, where X and Y are the set of strategies and $f_s, g_s : X \times Y \to \mathbf{N}$ the cost functions, for AS I and AS II (resp.). In particular, $f_s(x, y) = \phi_s(x)$, where $\phi_s : X \to \mathbf{N}$, and $g_s(x, y) = \psi_s(y)$, where $\psi_s : Y \to \mathbf{N}$. $G_d = (X, Y, f_d, g_d)$, a dummy game, of pure externality, where $f_d, g_d : X \times Y \to \mathbf{N}$ are the cost functions for AS I and AS II (resp.). In particular, $f_d(x, y) = \phi_d(y)$, where $\phi_d : Y \to \mathbf{N}$, and $g_d(x, y) = \psi_d(x)$, where $\psi_d : X \to \mathbf{N}$. G_s is a cardinal potential game [14], i.e., the incentive to change players' strategy can be expressed in one global function, a potential function (P_s), 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, with null potential (P_d). G has thus a potential $P = P_s + P_d = P_s$. In the bottom of Table III we report P_d

and P_s. To explicate P_s (thus P) 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 $\psi_s(y_0) \leq \psi_s(y)\ \forall y \in Y$. 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 G.

Theorem III.1. *A ClubMED Nash equilibrium corresponds to the strategy profile with minimum potential.*

If (x^*, y^*) is an equilibrium, $P(x^*, y^*) \leq P(x, y^*), \forall x{\in}X$. But: $P(x^*, y^*) = \phi_s(x^*) - \phi_s(x_0)$ 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^*. So $P(x^*, y^*) = 0$, that is a minimum of P. ■

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

Corollary III.2. *G always possesses a Nash equilibrium.*

Indeed, authors in [14] prove that finite potential games always possess a (pure-strategy) equilibrium. The opportunity of using the minimization of the potential function represents a key advantage of the ClubMED solution. It allows decreasing the complexity of the Nash equilibrium computation, which may be very high for large instances (especially for the generalized framework presented in the following). Therefore, for this base ClubMED modelling, if the equilibrium is unique it corresponds to hot-potato routing because G_s considers egress costs only, which somehow validates the current practice (however, we will explain how this differs in the generalized framework). When there are multiple equilibria, G_d can help in avoiding tie-breaking routing by the selection of an efficient equilibrium in the Pareto-sense (as detailed below).

Definition III.3. A strategy profile s is *Pareto-superior* to another profile s' if a player's cost can be decreased from s to s' without increasing the other players' costs.
Remark: And s' is *Pareto-inferior* to s.

Definition III.4. A strategy profile is *Pareto-efficient* if it is not Pareto-inferior to any strategy profle.
Remark: Pareto-efficient profiles form the *Pareto-frontier*.

In Table II, (l_2,l_2) is Pareto-inferior to (l_1,l_2) because $2c_2^{II} < c_1^{II} + c_2^{II}$, and ($l_1$,$l_2$) forms a singleton Pareto-frontier.

It is worth noting that the MEDs of AS I and AS II are never compared, never summed together, hence they can be calculated over different integer scales. What is important to sort strategy profiles is the ordering of individual AS costs[3].

B. Generalization to directed metrics and multiple links

So far, we assumed that the cost metric announced through the MED stands for the two directions, the incoming and the outgoing ones. Normally, it corresponds only to the incoming one (IGPs can manage directed costs). The modelling of directed costs intuitively does not change the ClubMED equilibrium and the Pareto-efficiency[4]. It decouples the (egress) costs used to form G_s from which the potential function is build from the (ingress) ones used to form G_d.

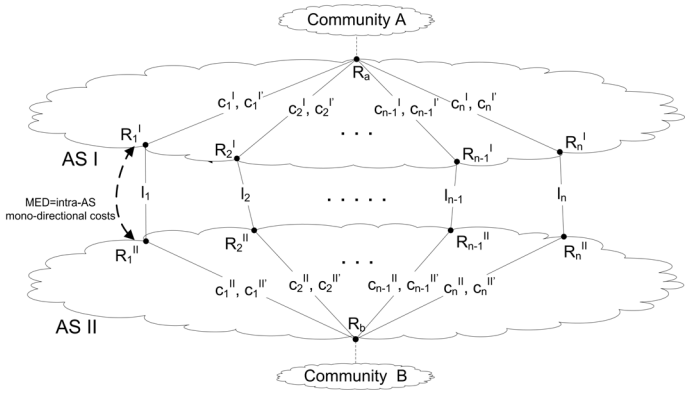


Fig. 3. ClubMED interaction example with multiple inter-AS links.

Further generalizing, multiple peering links are to be considered, as depicted in Fig. 3. Let $|E|{=}n$ be the link number, l_i and l_j, $i, j \in E$, the links chosen by AS I and AS II (resp.). Let c_i^I and $c_i^{I'}$ be the ingress and egress costs at link i for AS I, and idem c_i^{II} and $c_i^{II'}$ for AS II. The costs corresponding to the strategy profile (l_i,l_j) are ($c_i^I + c_j^{I'}$, $c_i^{II'} + c_j^{II}$). With multiple links, the occurrence of multiple equilibria increases. This happens under the necessary conditions:

$$\exists\ i, k \in E \mid i \neq k\ \wedge\ c_i^{I'} = c_k^{I'} \tag{1}$$

$$\exists\ i, k \in E \mid i \neq k\ \wedge\ c_i^{II'} = c_k^{II'} \tag{2}$$

for AS I and AS II (resp.). Given a link l_k, thus the ingress cost c_k^I, the strategy profiles (l_k, l_i) or (l_i, l_k) - many $i \in E$ may satisfy (1) or (2) - are eligible equilibria if both (1) and (2) are satisfied. Fig. 4 depicts two examples with 3 inter-peer links and their strategic forms. The exponent indicates the corresponding potential value. Egress costs are given close to the egress points, while ingress costs are close to the destination communities. For the upper case, (1) and (2) are not satisfied: there is a single equilibrium (l_2, l_2). For the lower case, (1) and (2) hold: there are four equilibria. Which should be selected? The peers could easily coordinate in playing the Pareto-superior one, (l_3, l_1). In ClubMED, a strategy profile (l_r, l_s) is Pareto-superior to a strategy profile (l_i, l_j) if:

$$\begin{aligned}&(c_r^I + c_s^{I'} < c_i^I + c_j^{I'} \quad \wedge \quad c_r^{II'} + c_s^{II} \leq c_i^{II'} + c_j^{II}) \quad \vee \\ &(c_r^I + c_s^{I'} \leq c_i^I + c_j^{I'} \quad \wedge \quad c_r^{II'} + c_s^{II} < c_i^{II'} + c_j^{II})\end{aligned} \tag{3}$$

For the lower case, note that the Pareto-superior equilibrium is not Pareto-efficient, it is Pareto-inferior to (l_1,l_3) that is the single element of the Pareto-frontier - (l_1,l_3) 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 assume the form

[3]Additional mathematical properties not found here are detailed in [12]

[4]We assume that composite MED attributes can be easily coded to transport both the ingress and the egress costs (and other costs mentioned hereafter).

of a Prisoner-dilemma game, in which equilibria are Pareto-inferior to non-equilibrium strategy profiles.

C. Nash Equilibrium Multi-Path (NEMP) coordination policy

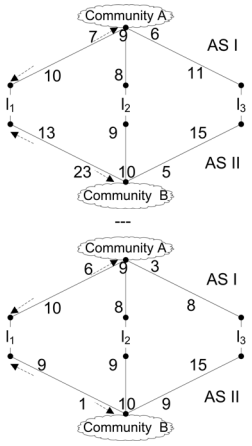
Within the described framework two peers would rationally route accordingly to a ClubMED equilibrium profile because it grants a rational stability to the routing decision. With many links, directed metrics and the needed practical extensions defined hereafter (see sections III-D and III-E), the number of equilibria in the Nash set drastically increases. The issue is thus to highlight the possible coordination policies on these equilibria.

In a fully non-cooperative framework, the ClubMED implicit policy to which to coordinate is: *to play the equilibria in the Nash set*. Hence, it is feasible to natively implement a *Nash Equilibrium Multi-Path (NEMP)* inter-peer routing policy. No coordination signalling message is needed: NEMP can be applied at only one side, under the assumption that a rational agent, as a carrier shall be, would route accordingly to a Nash equilibrium. E.g, in the bottom of Fig. 4 AS I may balance the load on l_2 and l_3, being aware that AS II may or may not balance its load on l_1 and l_2. Especially top-tier carriers, interconnected at numerous Points of Presence worldwide, would benefit from NEMP avoiding so sudden bottlenecks at inter-carrier links. In fact, the NEMP policy implies multipath routing on peering link when more than one equilibrium is selected. However, as above mentioned, the set of equilibria can be restricted to the Pareto-superior ones; but many Pareto-superior equilibria can exist, so the NEMP policy is to be used on the Pareto-superior profiles of the Nash set. Please note that there may not exist Pareto-superior equilibria: in this case, NEMP is performed on all the equilibria.

Finally, it is worth remarking that, from a computational standpoint, the NEMP policy is very efficient in that it simply requires the minimization of the potential value and a trivial Pareto-restriction of the Nash set, even if this contains "simple" equilibria of the one-shot game. Indeed, the equilibrium policy of the repeated game, from "folk-theorem"-like results [11], would be to play the profiles in the Pareto-frontier, even if they do not necessarily represent an equilibrium of the one-shot game. However, such a policy would be prohibitive because very time consuming: it needs the enumeration of all the strategy profiles, whose number grows exponentially when many links and multiple pairs are considered (see Sect. III-E). This is the reason why we evaluated, in our experimentation, only the NEMP policy, obtaining solutions in the order of the ms with nowadays processors (instead of hours).

D. Dealing with incomplete cost information

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. E.g., the costs in Fig. 4 may be computed for the starting profile (l_1, l_1). A change of the AS II-head flow via l_2 or l_3 may cause a decrease of the ingress cost by l_1, because



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. 4. 3-link examples.

of available bandwidth decrease on the corresponding links, and/or an increase of the ingress cost at l_2 or l_3 for the opposite reason. It may also happen that the ingress/egress cost from an AS to another changes when an egress/ingress flow direction changes because, for large topologies, flows with inverse inter-domain directions may use core links in the same direction.

As currently formulated, with IGP-WO operations the ClubMED would converge to a stable configuration after some repetitions. The ClubMED decision shall be kept stable as long as needed to avoid too many route oscillations while assuring a good solution. In order to reach this purpose, we integrate IGP-WO operations as follows. Let $\delta_{i,j}^{k,I}$ and $\delta_{i,j}^{k,I'}$ be the variations of the egress and the ingress cost (resp.), at AS I for link k, passing from the current strategy profile to the profile (l_i, l_j). Similarly, $\delta_{i,j}^{k,II}$ and $\delta_{i,j}^{k,II'}$ for AS II. Once pre-computed, the δ may be conveyed within the MED attribute to refine the strategic form. However, while announcing loose intra-AS costs is not so critic because the intra-domain topology is fully abstracted, exchanging the δ may represent an excessive insight in a carrier's operations (that might allow a peer to partially infer the peer's network status). Hence, the δ should be abstracted. Using the current cost and its δ variations, each peer can announce just an error to give an interval of equivalence for the computation of the equilibrium. 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, we opt for an optimistic min-max computation:

$$\epsilon^I = \min_{k \in E} \left\{ \max_{i,j \in E} \left\{ \delta_{i,j}^{k,I} \right\} / c_k^I \right\} \tag{4}$$

Similarly for ϵ^{II}, and the ingress cost errors of AS I and AS II, i.e. $\epsilon^{I'}$ and $\epsilon^{II'}$ (resp.). 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 asymmetric bottlenecks around inter-AS links. The important effect of the errors is a larger number of equilibria. Indeed, they arise a *potential threshold* under which a profile becomes an equilibrium. That is, first the minimum potential strategy ($P(x^*, y^*)$) is found, then the other profiles that have a potential within the minimum plus the threshold (T_P) are considered as equilbria too. More precisely, in the worst case, each potential difference ΔP from strategy i to j can be increased of the amount (for AS I) $\epsilon^I(c_i + c_j)$. Generically, in the worst case the 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:

$$a_I(x_1, x_2) = \epsilon^I \left[\phi_s(x_1) + \phi_s(x_2)\right]$$
$$a_{II}(y_1, y_2) = \epsilon^{II} \left[\psi_s(y_1) + \psi_s(y_2)\right] \quad (5)$$

It is reasonable to opt for the following optimistic threshold:

$$T_P = \min_{x_1, x_2 \in X} \{a(x_1, x_2)\} + \min_{y_1, y_2 \in Y} \{a(y_1, y_2)\} \quad (6)$$

All strategy profiles (x, y) such that $P(x, y) \leq P(x^*, y^*) + T_P$ will be considered as equilibria. More straightforwardly, the Pareto-superiority condition (3) can be easily extended considering $\epsilon^{I'}$ and $\epsilon^{II'}$ [12]. E.g, for the upper case in Fig. 4, for simplicity let all the ϵ cost errors be equal to 12%. Besides the existing equilibrium (l_2,l_2), one new equilibria is added: (l_1,l_2). Besides anticipating possible future routing deflections, the potential threshold may also allow escaping selfish solutions mainly guided by G_s: Pareto-superior profiles may be introduced in the Nash set and then selected.

E. Dealing with multiple flows and peering link congestion

In order to take broader decisions, it would result more useful to consider many pairs of inter-cone flows in a same ClubMED game. In this way the equivalence condition can be extended to all the pairs together, not necessarily related to a same couple of ClubMED routers. For 2 pairs and 2 links, the set of strategies X^2 or Y^2 becomes $\{l_1l_1, l_1l_2, l_2l_1, l_2l_2\}$. For m pairs and n links, the multi-pair game is the repeated permutation of m single-pair n-link games: $|X^m|$=$|Y^m|$=n^m. G = (X^m; Y^m; f_s,f_d,g_s,g_d: $X^m \times Y^m \rightarrow \mathbf{N}$).

In a multi-pair ClubMED framework, carriers shall control the congestion on inter-peer links. The more egress flows are routed on a peering link, the more congested the link, and the higher the routing cost. We aim at weighting thus the inter-carrier links when congestion may arise due to the inter-peer flow routing. This may be done by modelling the inter-peer link in IGP-WO operations (e.g. [6]), but the requirement of separating intra-domain from inter-domain routing would not be met [1]. We add an endogenous congestion game G_c = (X^m; Y^m; $f_c, g_c : X^m \times Y^m \rightarrow \mathbf{N}$) to G, where $f_c(x^m, y^m) = \phi_c(x^m)$ and $g_c(x^m, y^m) = \psi_c(y^m)$. Let H be the set of inter-peer flow pairs, ρ_h the outgoing bitrate of the pair $h \in H$, and C_i the egress available capacity of l_i. G_c should not count when $\sum_{h \in H} \rho_h \ll \min_{i \in E}\{C_i\}$, otherwise it should do affecting the G equilibrium selection. The G_c costs are to be monotonically increasing with the number of flows routed on a same link. An effective and commonly agreed congestion cost convex function is $1/(C - \rho)$, where $(C - \rho)$ is the idle capacity [15]. We shall use (idem for $\psi_c(y^m)$):

$$\phi_c(x^m) = \sum_{i \in E | l_i \in x^m} \left\lceil K_i \frac{1}{C_i - \sum_{h \in H} \rho_h^i} \right\rceil \quad (7)$$

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 weights, e.g., such that it is 1 when the idle capacity is maximum, i.e., $K_i = C_i$. ρ_h^i is the fraction of the pair h flow that is routed on l_i.

IV. Implementation Aspects

The framework does not require new protocols. It just exploits the MED and refinements to the BGP decision process.

A. Building destination communities

The destination communities building method has to be carefully designed. Each destination cone groups prefixes announced to the same border router or group of border routers. As already mentioned, the cone prefixes shall belong to stub ASs or direct clients whose traffic enters at a single border router, and a cone pair couples two equivalent peers' destination cones. Ingress BGP filters can be configured to mark routes announcing a set of prefixes and learned by a subset of neighbour ASs with a same BGP *community id*, to apply per-cone policies. We assume that each community created in this way is attainable by the same gateway router. Moreover, only network prefixes not announced by another peer shall be part of a peer's destination cone (this condition may be relaxed under a ClubMED-based extended peering, with more than 2 ASs, to be defined in further work).

B. BGP filtering and decision process

Practically, ClubMED decisions are needed only at a few non-adjacent edges of the peering networks. It is required that the decision process of ClubMED nodes (e.g., R_a and R_b in Fig. 3) indicates the egress peering router corresponding to the ClubMED solution. In order to do so, first ClubMED nodes normally receive the MEDs announced by the peer. Then, the discrete potential function is updated accordingly, and so the Nash set found minimizing it, whose equilbria are then shrinked w.r.t. to the Pareto-superiority. A ClubMED process responds with the corresponding router IP(s) when queried by the BGP decision process. The BGP process queries it once verified that many paths with the same AS hop count exist, and once verified that the prefix's community id corresponds to one of those preconfigured. For the multi-pair framework, in order to dispose of the G strategic form at these ClubMED nodes they may synchronize via ad-hoc sessions. Alternatively, they might demand the ClubMED decision to a sort of route server (easily integrating route server architectures recently proposed like [16]) or Path Computation Elements. The data to be shared among ClubMED nodes of a same peer are just: the peers' potential values of the single-pair game, so that the global potential values can be built with a simple permutation of the single-pair functions; the potential threshold contributions (i.e., the $\min\{a(.)\}$, see Sect. III-D); the flow bit-rates to calculate the congestion costs for the congestion game.

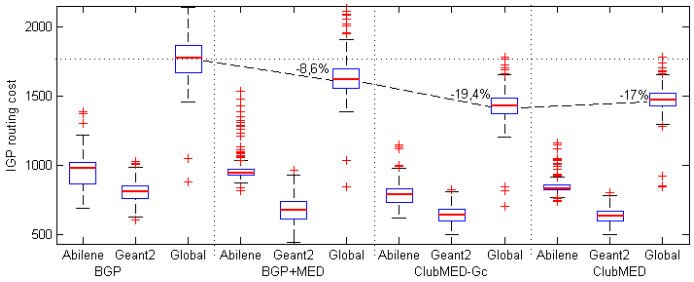


Fig. 5. Global routing cost Boxplot statistics.

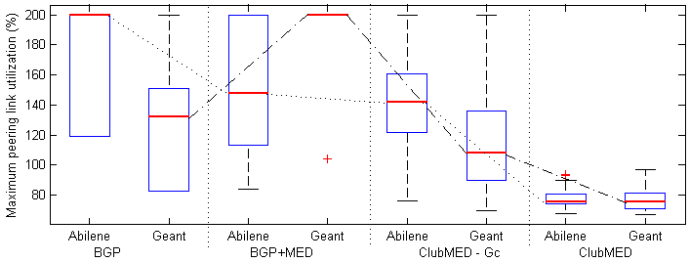


Fig. 6. Boxplot statistics of the maximum link utilization (%)

C. Composite MED coding

As currently defined, the MED transports the ingress cost only. In the ClubMED framework we suppose that a *composite MED* is used to code also the egress and congestion costs, and the cost errors. Thus, a MED coding should be agreed ex-ante.

V. Experimentation results

We created a virtual interconnection scenario among the Geant2 and the Internet2 ASs, depicted in Fig. 9, 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 [17] was used to run a IGP-WO heuristic, with a maximum link weigth of 50 for both ASs. We used 360 successive traffic samples, oversampling the datasets from [18] for Geant2 and from [19] for Internet2 on a 8h basis (to cover all the day times). The original intra-AS link capacities have been considered. The inter-cone routing generates additional traffic for the traffic matrices. We used a random inter-cone traffic matrix such that flows are equivalent 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 ClubMED to the BGP solution, without and with ('BGP+MED' in the figures) MED signalling at both sides. Fig. 5 reports the routing costs statistics in BoxPlot format (minimum; box with lower quartile, median, upper quartile; maximum; outliers). For each method, we display the Internet2, the Geant2 and the global routing cost. For the first two figures only, we considered two ClubMED solutions, without and with the congestion game G_c. Fig. 6 reports the Boxplot statistics maximum link utilization as seen by each peer, with the four above mentioned methods. Fig. 7 reports the number of ClubMED Nash equilibria and those Pareto-superior in a log-scale for all the rounds. When no Pareto-superior equilibria were found, NEMP was applied to all the Nash equilibria. Fig. 8 reports the number of routing changes with respect to the previous round (with an upper bound equal to the total number of flows), together with the Boxplot statistics.

All in all, we can synthetically assess that:

- the median routing cost is reduced of roughly 17% (simple uncoordinated MED signalling already improves it by 8%, but ClubMED further improves it);
- the addition of the congestion game G_c slightly augments it, but allows nullifying the congestion on peering links (that appear over-congested with a median between 130% and 200% with BGP, and a few congested with ClubMED without G_c);
- comparing BGP with BGP+MED, the latter seems improving the performance of Geant2 and Internet2 in terms of routing cost and maximum peering link utilization respectively, and conversely; this is probably due to the higher global path cost for Internet2 and to the higher number of intra-AS connection requests for Geant2;
- 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);
- the routing stability is significantly improved thanks to the consideration of the loose cost errors in G and thus to the arise of the potential minimum threshold: we pass from a *median* of 4 routing changes per round on 12 possible ones with BGP, to a median of 0 with ClubMED; ClubMED can significantly increase routing stability;
- a better behaviour in terms of routing stability seems corresponding to larger Nash sets (c.f. Fig. 7 and Fig. 8).

VI. Summary

We proceeded with a deductive analysis of the coordinated inter-peer routing interaction via the MED attribute of BGP. We modelled it as a game, named ClubMED, composition of a potential game guiding the Nash equilibrium selection and of a dummy game affecting the Pareto-efficiency. The ClubMED game routes equivalent aggregates of inter-carrier

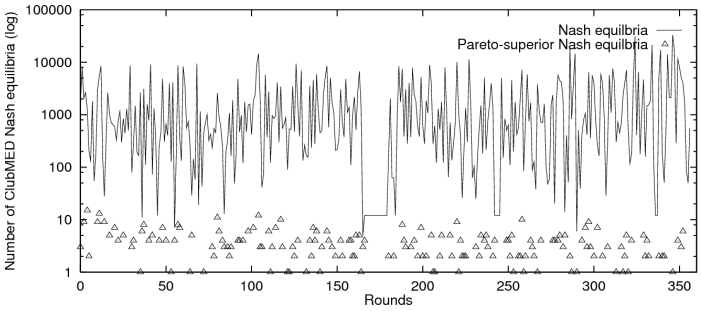


Fig. 7. Dynamics of the number of found Nash equilibria

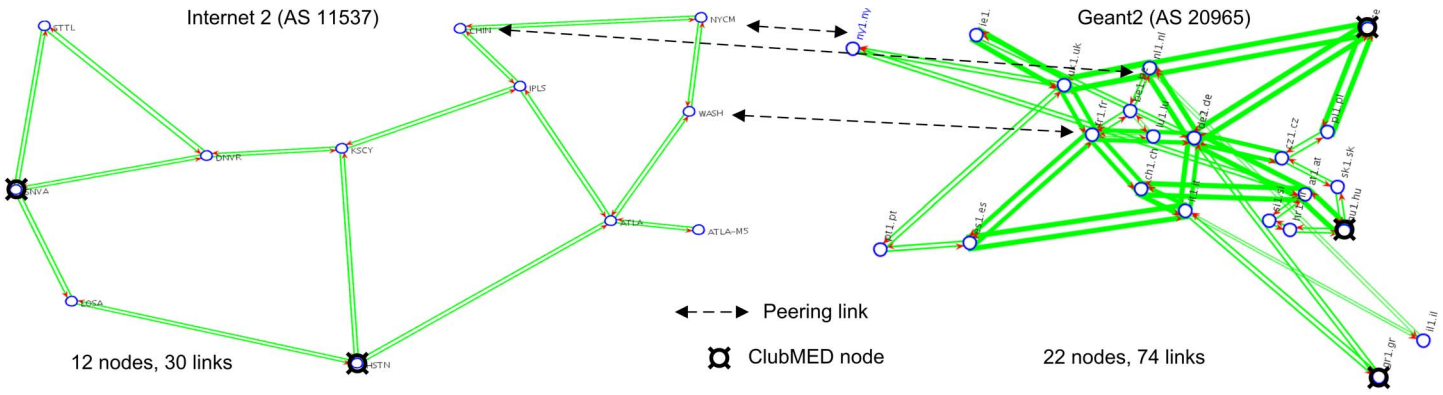


Fig. 9. Internet2 - Geant2 peering scenario with 3 peering links.

flows over multiple links, includes cost errors due to IGP-WO operations and is able to control the congestion of peering links. The frequent occurrence of multiple ClubMED Nash equilibria arises a need for a coordination policy. We presented the NEMP routing policy that shall be implemented on the Nash equilibria and Pareto-superior profiles. We discussed implementation aspects related to a practical deployment in BGP/IP networks. Finally, we validated the ClubMED framework emulating the peering between the European and North American research networks using real datasets.

The results show that the global routing cost can be reduced of roughly 17%, that the peering link congestion can be avoided with the addition of an endogenous congestion game, and that the inter-peer routing can be stabilized. Besides this promising performance, in the ClubMED framework the carriers' selfish and non-cooperative behaviour is respected as an imperative requirement. The ClubMED framework emerges as a pragmatic and effective solution between the current uncoordinated practice and other ideal yet unwise cooperative solutions. It may be implemented to overlay a special peering for critical inter-carrier flows above co-existing settlements concerning flows whose routing can not be coordinated. It may allow a still finer policy routing by the artificial addition of endogenous costs. It may freeze the wild bargaining that nowadays characterizes the peering settlements [20], yielding to long-term and effective inter-carrier agreements for the future Internet.

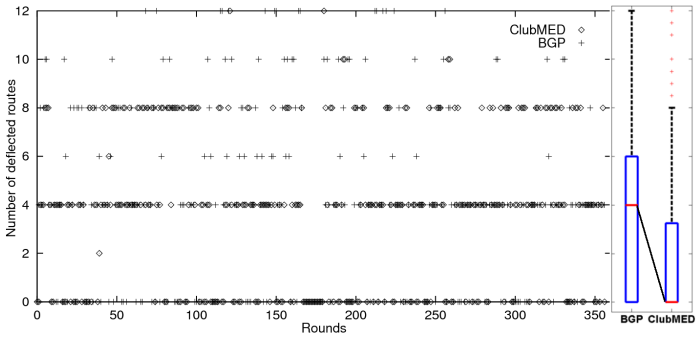


Fig. 8. Dynamics and Boxplot statistics of the routing deflections

Further work is needed to study coordination policies other than the presented NEMP one. The idea is to define suboptimal yet effective policies for the repeated ClubMED game. Moreover, we are working to the definition of an extended peering framework relying on a similar game theoretic modelling, wherein the borders with multiple ASs are modelled as a classical bilateral peering border.

REFERENCES

[1] R. Teixeira et al., "Impact of hot-potato routing changes in IP networks", *IEEE/ACM Trans. on Networking*, Vol. 16, No. 6, Dec. 2008

[2] R. Douville, J.-L. Le Roux, J.-L. Rougier, S. Secci, "A Service Plane over the PCE Architecture for Automatic Multi-Domain Connection-Oriented Services", *IEEE Comm. Magazine*, Vol. 46, No. 6, June 2008.

[3] B. Huffaker et al., "Distance Metrics in the Internet", in *Proc. of IEEE International Telecommunications Symposium (ITS) 2002*.

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

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

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

[7] R.T.B. Ma et al., "Internet Economics: The use of Shapley value for ISP settlement", in *Proc. of CoNEXT 2007*.

[8] R. Ma et al., "Interconnecting eyeballs to content: a Shapley value perspective on ISP peering and settlement", *Proc. of SIGCOMM 2008*.

[9] G. Shrimali et al., "Cooperative Inter-Domain Traffic Engineering Using Nash Bargaining and Decomposition", in *Proc. of INFOCOM 2007*.

[10] A.E. Roth, *The Shapley value, essays in honor of Lloyd S. Shapley*, Cambridge Univ. Press (1988).

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

[12] S. Secci, *Game Theory for Internetworking*, minor Ph.D. thesis, POLIMI.

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

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

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

[16] Y. Wang, I. Avramopoulos, J. Rexford, "Design for configurability: rethinking interdomain routing policies from the ground up", *IEEE J. on Selected Areas in Communications*, Vol. 27, No. 3, April 2009.

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

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

[19] By courtesy of Y. Zhang. Internet2/Abilene topology and traffic dataset. http://www.cs.utexas.edu/yzhang/research/AbileneTM.

[20] P. Faratin et al., "Complexity of Internet Interconnections: Technology, Incentives and Implications for Policy", in *Proc. of TPRC 2007*.