Improving Route Diversity through the Design of iBGP Topologies

Cristel Pelsser, Tomonori Takeda, Eiji Oki and Kohei Shiomoto
NTT Network Service Systems Laboratories, NTT Corporation, Japan

***Abstract*—In a Service Provider (SP) network, routes for external destinations are distributed on iBGP sessions. This traditionally required the establishment of a full-mesh of iBGP sessions in the network. A common practice is now to make use of Route Reflectors (RR). Such a practice is more scalable in the number of iBGP sessions to be configured in a SP network. However, it has been shown that RRs have a negative impact on the diversity of routes available in the network. This is an important issue as routers may not be able to quickly use an alternate route in case of a route failure.**

In this paper we tackle the problem of route diversity in a Service Provider network composed of RRs. We propose an algorithm to design iBGP session topologies with improved route diversity. We rely on an initial route reflection topology. Our algorithm proposes the addition of a few iBGP sessions to some border routers of the domain. These border routers receive a large number of external routes for which routers lack diversity. We show by means of simulations that our algorithm meets its goals. In the resulting topologies, each BGP router knows at least two different ways to reach distant destinations. This is ensured as long as a prefix advertisement is received at different nodes at the border of the AS. Secondly, we observe that the number of iBGP sessions required to achieve this goal is significantly below the number of sessions required in the case of a full-mesh. Finally, the remaining lack of route diversity after the use of our design algorithm indicates that new external peering sessions should be established. In this case, our algorithm shows that diversity cannot be reached for some prefixes independently of the iBGP topology, with the current external peering sessions.

I. Introduction

The Internet is divided in domains, also called Autonomous Systems (AS). Each AS is usually administrated by a single company. The protocol currently deployed to distribute routing information between domains is the Border Gateway Protocol (BGP). In BGP, external BGP (eBGP) sessions are established to exchange routes with neighboring ASs. BGP routes are distributed in an AS by means of internal BGP (iBGP) sessions.

A BGP route[1] is composed of a prefix, a Next-Hop (NH), and a set of attributes. The NH is the address of a router at the border of the domain. This router is able to forward traffic toward the destinations belonging to the prefix.

Initially, routers were only allowed to advertise, on iBGP sessions, routes that were received on eBGP sessions. Thus, redistributing BGP routes to all the routers of an AS required to setup a full-mesh of iBGP sessions in the AS. This leads to scalability issues when the number of routers in the AS becomes large. Today, the trend is to use Route-Reflectors (RR) [1] to redistribute routes in an AS. A RR may readvertise routes learned on some iBGP sessions on other iBGP sessions. This enables to reduce the number of iBGP sessions to be established in the network[2].

A router holds a routing table per BGP session (i.e. per BGP peer). It stores the routes received on each session in these tables. A router may receive multiple routes for the same prefix. In this case, it selects a single of these routes for packet forwarding. **Only this route is redistributed by the router on iBGP sessions**. The selection of a single route for each destination relies on the values of the routes' attributes. The route selection process is composed of a set of rules applied in sequence. These rules are provided in Table I. Each rule eliminates from consideration all the routes that do not have the best value for a given attribute. When a single route remains, it is selected for packet forwarding. First, only the routes that are preferred by the local AS are kept. Then, among the remaining routes, the routes that have crossed a larger number of ASs are removed. The Multi-Exit Discriminator (MED) is used between multi-connected ASs. It indicates the preference of the neighboring AS concerning entry points for the traffic in its domain. Among the routes with the lowest MED, routes learned from eBGP peers are preferred. Then, the cost of the path (the Internal Gateway Protocol cost) to the NH is taken into account. Finally, if multiple routes are still available, tie-breaking rules are applied to obtain a single route.

TABLE I
Simplified BGP decision process

Sequence of rules	
1	Highest Loc_pref
2	Shortest AS-path
3	Lowest MED
4	eBGP over iBGP
5	Lowest IGP cost to NH
6	Tie-break

The slow convergence of BGP has been highlighted in the literature. In [3], Labovitz et al. say that recovery from a failure impacting inter-domain routes takes three minutes in average. Moreover, Wang et al. show in [4] that routing changes

[1]When we use the term "route" in this paper, we refer to the notion of BGP route. Similarly, the term "router" is used to designate a BGP router. That is, a router running BGP.

[2]An alternative solution to improve scalability in the number of required iBGP sessions relies on confederations [2]. We send the reader to section VI for hints on improving route diversity in a confederation of ASs.

subsequent to a failure contribute significantly to end-to-end packet loss. Several techniques to improve BGP convergence have been proposed [5], [6], [7]. However, as claimed by [8], reducing BGP convergence time is not sufficient in itself to ensure the level of reliability required by loss and delay sensitive applications.

Solutions have been proposed in order for a domain to receive multiple paths to external destinations [8], [9]. These routes are present at the frontier of the domain. However, this diversity may not be redistributed to all the routers inside the domain. Uhlig et al. [10] have shown that, in a network with RRs, most routers do not possess multiple routes with alternate NHs for most of the destinations. Thus, if a route fails, the routers lose reachability to the destination of the route. They have to wait for BGP to converge inside the AS before being able to join the destination again. Depending on the value of BGP timers and on the number of routes that fail, BGP convergence may take a few tens of seconds. If routers had diverse routes, network resilience would be improved. The switch-over to an alternate route would take much less than a second [11]. The objective of this paper is to achieve such NH diversity in the routers of a domain. For this purpose, we focus on the design of the iBGP topology of a domain. To our knowledge, it is the first time such an approach is considered.

The design of iBGP route reflection topologies is a NP-hard problem [12]. The solution space is wide and many factors, such as CPU and memory capacity of the nodes, need to be considered. In this paper, we rely on an initial iBGP route reflection topology. **We propose a simple algorithm that leads to NH diversity in the routers by adding iBGP sessions to an initial topology of RRs.** In the resulting iBGP configurations, each router learns at least two different NHs to reach every destination. This way, when the route through one of the NHs fails another route may still be available. Such a route may then be used before new routes are learned through BGP convergence. Our algorithm aims at achieving route diversity by adding only a **small number of iBGP sessions**.

This paper is structured as follows. First, we introduce the problem of insufficient inter-domain route diversity inside a domain, in section II. Secondly, we present solutions that have been proposed in the literature to solve related issues, in section III. In section IV, we describe our methodology and our design algorithm. An evaluation of our proposal is presented in section V. Then, we give several directions for future work in the design of iBGP topologies with constraint on resiliency, in section VI. Finally, we conclude the paper.

II. Lack of route diversity in BGP

In this paper, we say that **NH diversity is achieved if and only if, for any given prefix, there are at least two BGP routes with different NHs in the routing tables of each router in a domain.**

Let us consider the case of $AS2$ in Figure 1. In this example, a prefix, P, is advertised by routers $R11$ and $R12$ in $AS1$. Inside $AS2$, $R23$ and $R24$ are RRs. $R21$ is a client of RR $R23$ and $R22$ is a client of $R24$. The boxes represent the routing tables of the routers in $AS2$. We see that $R21$ and $R22$ each learn a single BGP route for prefix P. However, the RRs know two routes for P. We observe that the two routes for prefix P have a different NH, in the routing tables of the RRs. Thus, we say that route diversity is achieved for prefix P at routers $R23$ and $R24$. This is not true for their clients. There is no route diversity for prefix P at routers $R21$ and $R22$.

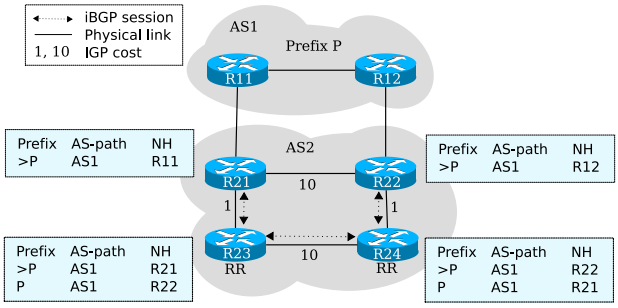

Fig. 1. Lack of route diversity

The lack of NH diversity for a prefix may lead to packet losses in case of failure of the NH, one of the NH's inter-AS links or of the intra-domain path to the NH. In the example of Figure 1, if link $R21-R11$ fails, $R21$ looses its route toward prefix P. Thus, $R21$ will drop all packets destined to P until it learns the route to P via $R22$ from its RR.

Different factors have an impact on route diversity. First, the presence of RRs may reduce route diversity. A route reflector chooses routes on behalf of its clients. It may receive multiple routes for a destination but it only advertises one of these routes to its clients. In Figure 1, RR $R23$ learn one route from RR $R24$ with NH $R22$. It also learns one route from its client $R21$. $R23$ selects the route from $R21$ as best route because the cost to NH $R21$ is lower that the cost to $R22$. The intra-domain path to $R21$ has a cost of 1 while the cost to $R22$ is equal to 11. Since the best route at RR $R23$ is received from $R21$, $R23$ does not advertise any route for P to $R21$. Consequently, $R21$ does not know that $R22$ is able to forward packets destined to P. If a full-mesh of iBGP sessions was used, $R21$ would learn the route through NH $R22$ directly from $R22$. Thus, $R21$ would know diverse NHs for P.

Second, local policies may also hinder the distribution of diverse routes in an AS. Local policies are usually enforced by assigning local preference values to the routes. The Loc_pref value is considered in the first step of the BGP decision process (see Table I). Only the routes with the highest preference will be selected by the border routers and redistributed in the domain. If there is a single route with highest preference for a prefix, there will be a lack of diversity in the routers for this prefix. Let's consider the topology in Figure 1. Local preferences are often used between multiply connected ASs, such as $AS1$ and $AS2$, to configure one link as the primary and the other as the backup link. Assume that the operator of $AS2$ wants link $R21-R11$ to carry primary traffic and

link $R22 - R12$ only to be used in case of failure of the primary link. The operator will configure $R21$ to assign a high Loc_pref to the routes received from $R11$. The Loc_pref of the routes received from $R12$ at $R22$ will be set to a low value. All the routers in $AS2$ will prefer and redistribute the route with $R21$ as NH. After the convergence of BGP, only router $R22$ will know that there are two possible routes for prefix P. If link $R22 - R12$ fails, the nodes in $AS2$ have to wait for the convergence of BGP before being able to use the backup link.

Similarly to the use of the Loc_pref attribute, the values of the other attributes considered by BGP route's selection process (see Table I) also have an impact on NH diversity.

Several researchers have highlighted the lack of NH diversity in BGP routers. Among them, Uhlig et al. [10] have studied route diversity in a Tier-1 ISP with a hierarchy of RRs. They observed that most routers do not learn diverse NHs for most of the prefixes.

Router vendors have implemented an extension to BGP called "external best" [13]. This extension is useful in case a router prefers a route learned on an iBGP session from the routes received on eBGP sessions. When the "external best" option is activated, the router advertises its best eBGP route to its iBGP peers. The use of this extension may increase NH diversity in some routers of the domain. However, it does not completely solve the diversity problem. Let us consider the topology in Figure 1. Again assume that link $R22 - R12$ is configured as a backup link through the setting of a low Loc_pref. We have shown earlier that, in this situation, the external route learned for prefix P by $R22$ is not advertised by $R22$ to the other routers in $AS2$. Thus, $R21$, $R23$ and $R24$ do not have NH diversity for P. We note that this is also true in any other iBGP session topology. In the example of Figure 1, with the "external best" option, $R22$ advertises its eBGP route to its RR $R24$. Now, $R24$ has route diversity for P. However, the best route at $R24$ is still the route via $R21$, making use of the primary link. Thus, it will not advertise the route with NH $R22$ to the other routers in the AS. For this router, the route with NH $R22$ is not an external route. Consequently, the "external best" option does not enable $R24$ to advertise this route in the AS. Even with the "external best" option, $R23$ and $R21$ will only know a single NH for prefix P.

III. Related work

Several aspects of resilience toward prefixes distributed by BGP have been studied in the literature. Moreover, the design of iBGP topologies, meeting different objectives as the ones considered in this paper, has drawn attention. Here, we present an overview of this work.

Several authors have proposed mechanisms to ensure that an AS knows multiple routes for destinations outside its AS. These techniques aim to provide route diversity at the frontier of a domain. They do not ensure the distribution of these diverse routes to all the BGP routers inside a domain. In R-BGP [8], an AS learns failover paths from its neighboring ASs. These paths are not distributed with the classical implementation of BGP. The failover paths are used only if the usual routes advertised by BGP fail. In [9], pairs of domains negotiate the use of paths that are not distributed by BGP. Tunnels are established to carry data traffic along these paths.

Inside an AS, the following aspects of route resiliency toward distant destinations have been considered. Bonaventure et al. [14] propose a technique for the protection of external peering links by means of tunnels. Their technique requires the support of a new type of routes in BGP. Protection routes are advertised on iBGP sessions, inside the AS. Our solution provides this type of protection without requiring any modifications to BGP implementation. Another approach is to obtain higher NH diversity in the routers through an extension to BGP allowing multiple route advertisements for a single prefix [15]. However, Van den Schrieck et al. [16] have shown that such an extension may lead to BGP route oscillations. Filsfils [11] has proposed BGP Prefix Independent Convergence (BGP PIC). It is a routing table architecture that relies on the knowledge of backup NHs to reduce BGP convergence time. This architecture has to be used in combination with [14] or with this work to achieve the results expected by the author.

The design of iBGP route reflection topologies has been considered in [17], [18] and [12]. Buob et al. [18] provide a method to check that hot-potato routing is enforced in a given iBGP topology. They check that each router selects the same route it would have selected in the case of a full-mesh of iBGP sessions. Vutukuru at al. [17] propose an algorithm for the construction of iBGP topologies ensuring hot-potato routing. The objective of these two papers is to ensure that deflection and, thus, forwarding loops, do not occur in an AS.

In [12], the authors consider the design of robust iBGP topologies. They aim to minimize the probability of failure of iBGP sessions and the number of iBGP sessions that may fail. This approach does not ensure NH diversity in the routers. When maintenance of routers is performed, some iBGP sessions may still be taken down. This may lead to packet loss since diverse NHs are not necessarily available at the routers.

Finally, Caesar et al. [19] propose a novel architecture for route distribution inside an AS. Inside a domain, a server distributes external routes to all the routers in the domain. Such an architecture removes the burden of designing iBGP topologies. In its current implementation, the server distributes a single BGP route per destination to each router in a domain. Such an architecture may be promising in the future for diverse route distribution.

IV. Improving diversity

In this section, we present our solution for improving NH diversity at the routers of an AS. We propose an algorithm to be used offline, in the design phase of iBGP topologies. This algorithm determines a small number of iBGP sessions to add to an existing iBGP route reflection topology. With the resulting iBGP topology, NH diversity is achieved for all prefixes at each router in the operating network. Routers can

thus directly switch to an alternate route upon a failure of a BGP route.

As input, the algorithm takes the eBGP routes received at the AS Border Routers (ASBR), the IGP topology and an iBGP route reflection topology. Our solution relies on a tool such as [20] to compute the routing tables of the BGP routers in the domain. Alternatively, routing table dumps may be used if they are available.

The algorithm relies on the assumption that the "external best" option [13] is activated in the routers. This option enables to improve NH diversity in a domain. Moreover, in some configurations it is not possible to achieve NH diversity without this option. When all the routers in a domain prefer the same route (i.e. the same NH) for a prefix, the routes that may be received at other ASBRs for this prefix, are not propagated in the domain.

The principle of the algorithm is as follows. We consider, in sequence, the routers lacking diversity for a set of prefixes. We improve NH diversity for a router through the addition of iBGP sessions with ASBRs[3]. An ASBR is selected to become a new iBGP peer if adding a session to this ASBR most contributes to increase NH diversity at the router under consideration.

Assume that we want to improve NH diversity for router $R22$, in Figure 2. We see that $R22$ lacks NH diversity for prefixes P, Q and R. Routers $R21$, $R23$ and $R24$ are ASBRs in $AS2$. **With "external best", we are sure that an ASBR distributes, to its iBGP peers, one route with itself as NH, for each prefix it learns on an eBGP session.** ASBRs are good candidates for becoming an iBGP peer. In the example, $R21$ distributes a route with NH $R21$ for prefixes P, Q and R to its iBGP peers. $R23$ sends routes for prefixes Q and R with NH $R23$. And, $R24$ only sends a route for R with itself as the NH.

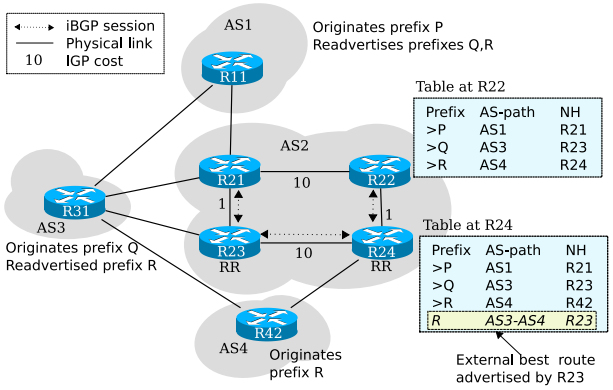

Fig. 2. Improving diversity

Some iBGP sessions do not contribute to increase the NH diversity at the considered router ($R22$, in the example). These are sessions with ASBRs such as: (1) an ASBR that is already an iBGP peer, (2) an ASBR that is already the NH for all the prefixes lacking diversity, (3) an ASBR that does not advertise any of the prefixes lacking diversity to its iBGP peers. Therefore, the algorithm does not consider to add an iBGP session with such routers. In the example, the algorithm will not propose to add an iBGP session between $R22$ and $R24$ because of (1). Thus, the algorithm needs now to choose between $R21$ and $R23$ as candidate iBGP peers. For this purpose, it needs to determine the ASBR that will contribute to increase the diversity for **most** of the prefixes lacking diversity. In our example, an iBGP session with $R21$ will increase the NH diversity for two prefixes, Q and R. Diversity will not be increased for P by adding a session to $R21$ as $R21$ is already the NH for P in $R22$'s routing table. On the other hand, an iBGP session with $R23$ will only increase NH diversity for prefix R. $R23$ is already the NH for Q at $R22$ and $R23$ does not advertise P to its iBGP peers. Thus, $R21$ is selected as new iBGP peer. If multiple ASBRs contribute to increase diversity for the same number of prefixes, our algorithm selects one of them in an arbitrary fashion.

We note that diversity cannot be increased for prefix P in our example. This is due to the fact that a single ASBR receives an external route for P. The failure of our algorithm to increase NH diversity for some prefixes indicates that NH diversity will not be reached in any iBGP topology for these prefixes. New external peering links need to be negotiated by the operator of the domain. In our example, the operator of $AS2$ could contact the operator of $AS1$ to schedule the establishment of a new link between $R22$ and $R11$.

Our algorithm relies on the eBGP routes received at the ASBRs. A change in the prefixes that are received from the external peers may have an impact on the NH diversity in the AS. To avoid having to reoptimize the iBGP route reflection topology every time a change in the external routes is observed, we suggest to build a model of the eBGP routes. We suggest to use classes of prefixes in this model. A Service Provider (SP) knows the type of connectivity that is provided by each of its external peers based on the contract it has negotiated with its peer. Thus, the SP knows if it will receive all the Internet routes from the peer or a subset of the routes. In the case of a subset of prefixes, the administrator knows the prefixes to expect. The prefixes that are always advertised together with the same BGP attributes belong to a class. For example, a class may contain all the prefixes assigned to European universities. Another class may be all the prefixes assigned to the American customers of the peer. Instead of trying to improve NH diversity for single prefixes, diversity is considered on a per class basis. A single prefix is used to represent a class in the model. An iBGP session that is added to improve diversity for this prefix improves diversity for all the prefixes in the class. Such a modeling is common [20], [10]. An iBGP topology computed based on such a model is likely to be robust to changes in eBGP routes, if the current peering agreements are respected. The model can also take into account predictions for changes in agreements and for the removal or the addition of external peers. We note that the

[3]We do not consider the addition of client sessions to RRs because adding such sessions is likely to prevent BGP convergence. We refer the reader to [21] for guidelines on building iBGP route reflection topologies that ensure the convergence of BGP.

real eBGP routes can be used instead of building such a model. Identifying the classes of prefixes may be computationally intensive. However, this is counter-balanced by the drastic time reduction achieved in the computation of the BGP routes.

We note that our algorithm adds iBGP sessions to ASBRs that receive a lot of external routes. Thus, an ASBR that receives many routes, that has many external peers, should be able to support a higher number of iBGP sessions than the other ASBRs. This effect is predictable. Therefore, these ASBRs can be correctly dimensioned to support the additional load. We note that the number of iBGP sessions at an ASBR will never be over the number of sessions it would have to support in a full-mesh.

The strength of our approach is that is it applicable today. No changes are required to the implementation of BGP. Moreover, as we will see in section V, the iBGP route reflection topologies that are generated by our algorithm contain far less iBGP sessions that a full-mesh of sessions.

In a topology with NH diversity, the routers can directly switch to an alternate route upon the detection of a route failure. Failure detection and rerouting to a locally available route can be done in less than a second with router architectures such as [11]. This is a significant gain compared to the few tens of seconds required today.

V. Evaluation

In this section, we present an evaluation of our algorithm on a research network. We study the NH diversity achieved with an initial topology and the iBGP topology generated by our algorithm. Then, we look at the number of iBGP session required with our algorithm to achieve NH diversity for all prefixes and in all the routers of the network.

We construct the model of the network used in our evaluation based on public information relative to its topology and external peers. The research network is composed of 17 nodes. Eight of these nodes are ASBRs. It has 12 external peers. We build a model of the external routes from our knowledge of the roles of the different peers. We follow the methodology introduced in section IV. Two of the peers are well known commercial Internet Service Providers (ISP). Four peers are research networks in the same continent as the considered network. Finally, there are three connections to major Internet eXchange (IX) points in the same continent and three to IXs in another continent. From this information, we assume that the classes of networks in Table II are advertised at the different peering points. Each line of the table represents an external peer. The characterization of the peer is provided in the first column. The second column contains the classes of prefixes advertised by the peer. We retain one prefix per class (see section IV). We observe from this table that there is redundancy in the prefixes received from the peers. There is an exception for prefix "research net4" that is only learned from one peer[4].

[4]This prefix may be advertised to the real network at multiple peering points. However, it is not captured in our model.

TABLE II
External prefixes

Peerings	Prefixes
1 commercial peer in continent 1	commercial continent1 commercial global
1 commercial peer in continent 2	commercial continent2 commercial global
1 research peer in continent 1	research net1
1 research peer in continent 1	research net2
1 research peer in continent 1	research net3
1 research peer in continent 1	research net4
3 IXs in continent 1	research net1 research net2 research net3 research continent1 research global commercial continent1 commercial global
3 IXs in continent 2	research continent2 research global commercial continent2 commercial global

We build the initial iBGP route reflection topology[5] as follows. There are two Points of Presence (PoP) in this network. We selected the mostly connected router of each PoP as the RR. We established an iBGP session between the RRs. All the routers in a PoP are clients of the RR.

Figure 3 illustrates the NH diversity at each router of the studied network. There are two bars for each router. The first bar shows the number of prefixes for which the router knows a single NH (label "no", on the x-axis). The second bar gives the number of prefixes for which at least two NHs are learned (label "yes" on the x-axis) at the router.

$Router1$ and $router2$ are the route reflectors. We observe in Figure 3 that these routers learn diverse NHs for the largest number of prefixes. There is NH diversity for 70% and 90% of the prefixes at $router1$ and $router2$, respectively. In average, there is diversity for only 8.7% of the prefixes at the other routers.

Let us look more closely at the reasons for NH diversity in some of the routers, with the initial iBGP topology. We observe that these routers ($router3$, $router12$, $router15$, $router16$ and $router17$) are ASBRs. For some prefixes, there is diversity at these routers because they receive one route from their eBGP peer(s) and the other route from their RR. We note that diversity is not present in all ASBRs and neither for all prefixes.

Figure 4 shows the diversity achieved in the routers with the iBGP topology computed by our algorithm. We observe that NH diversity is achieved at all routers for all the prefixes except "research net4". As mentioned earlier, this was expected since this prefix is received by a single ASBR in our model. The diversity obtained with our iBGP topology is the same as the diversity observed in a topology with a full-mesh of iBGP sessions. Our algorithm generates topologies where diversity is ensured for all prefixes that are received at different ASBRs.

In addition, we see that studying NH diversity for the iBGP

[5]The iBGP topology of this network is not available to the public.

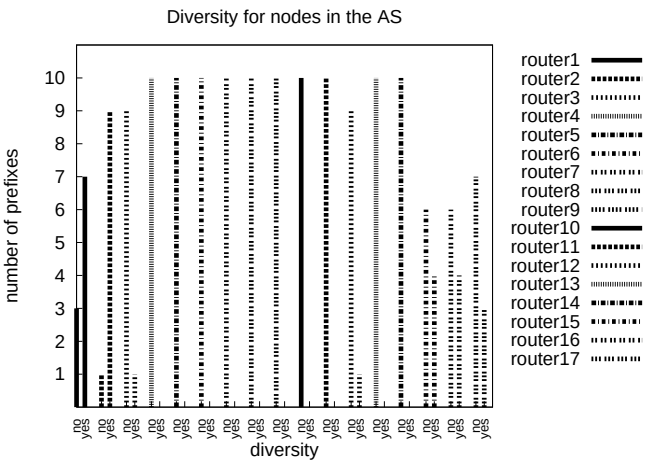

Fig. 3. Initial diversity

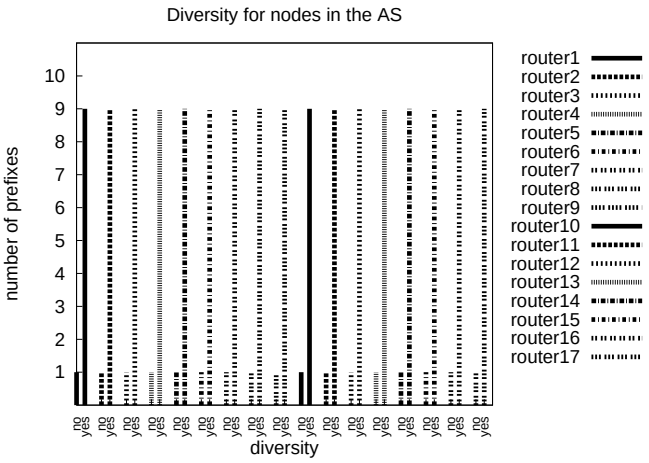

Fig. 4. Improved diversity

route reflection topology generated by our algorithm is very important. It enables us to detect situations when the only solution to achieve diversity requires the establishment of new external peerings. Here, we deduce from Figure 4 that a new external peering session should be negotiated to reach diversity for prefix "research net4".

By looking at the number of iBGP sessions, we assert that our algorithm generates iBGP topologies with far less session than in a full-mesh. The initial iBGP route reflection topology is composed of 16 iBGP sessions. To achieve the NH diversity of Figure 4, our algorithm suggests the addition of 28 iBGP sessions. This leads to a topology with a total of 44 iBGP sessions. A full-mesh of iBGP session in this domain would require 136 iBGP sessions. Thus, we observe that there are 68% less iBGP sessions in the generated topology than in a full-mesh of iBGP sessions. Our algorithm leads to topologies with a scalable number of iBGP sessions.

From the distribution of the iBGP sessions at the routers, we note that the algorithm suggests the addition of sessions mostly to the RRs and $router15$. These are the routers that learn a large number of prefixes on eBGP sessions. As mentioned in section IV, this is a predictable effect of our algorithm. These routers should be dimensioned accordingly.

In [1], Bates et al. state some recommendations for iBGP route reflection topologies. They recommend to add a full-mesh of iBGP sessions between all the routers in a PoP to the initial iBGP session topology that we used above. We performed simulations with a configuration meeting the recommendations expressed in [1]. We made the following observations. First, the initial iBGP topology is now composed of 67 sessions instead of 16 sessions. This initial topology has more iBGP sessions than the topology providing the diversity of Figure 4, generated by our algorithm. However, the diversity of Figure 4 is not reached. That is, NH diversity is not ensured for all the prefixes for which diversity is possible. With our algorithm, 21 sessions are added to this new initial topology in order to achieve the same NH diversity as in Figure 4. We note that this still gives a topology with far less iBGP sessions than in a full-mesh. The number of iBGP session is reduced by 35% compared to an iBGP full-mesh.

To conclude, we have shown in this section that with the iBGP session topologies generated by our algorithm, each BGP router in the network knows at least two different NHs for each distant destinations. This is ensured for all prefixes that are received at different ASBRs. We observed that the number of iBGP sessions required to achieve this goal is far below the number of sessions in a full-mesh. Finally, we have shown that our algorithm enables us to determine the set of prefixes that will always lack NH diversity with any iBGP session topology. We are thus able to detect that new external peering sessions should be established to achieve diversity for these prefixes.

VI. Further work

We have shown in this paper that our algorithm generates scalable iBGP route reflection topologies meeting our NH diversity goal. However, this algorithm is a first step toward the design of scalable iBGP topologies achieving this objective of NH diversity.

Several improvements can be brought to the algorithm to further reduce the number of iBGP sessions added to the initial route reflection topology. First, one should determine an intelligent ordering the routers considered by our algorithm. This ordering may have an impact on the sessions that are added. Adding certain iBGP sessions at a router may avoid adding other sessions to routers that are considered later by the algorithm. Second, we should determine an appropriate tie-breaking function, to select a single iBGP peer from multiple candidate iBGP peers that improve diversity by the same amount for the considered router. Such tie-break could take into account the current distribution of the iBGP sessions on the ASBRs. Third, in some networks some iBGP sessions may largely contribute to increase route diversity in a router while the contribution of other sessions may be small. One

can envisage to only add sessions with a large contribution to NH diversity.

Our algorithm focusses on the addition of non-client sessions to ASBRs. Instead, we could envisage to add client sessions to RRs. Such an approach is likely to require even less sessions than our proposal. However, one has to be careful when adding such sessions. One should pay attention to respect the guidelines expressed in [21] to ensure the convergence of BGP. Yet, these guidelines may lead to the addition of many unnecessary iBGP sessions with regard to diversity. Another approach would be to ignore these guidelines and check the correctness of the BGP convergence a posteriori. However, such a problem has been proven to be NP-hard [22].

Some domains are composed of a confederation of sub-ASs with eBGP sessions between routers of different sub-ASs. In these networks, NH diversity may not be achieved in the routers due to the internal iBGP session topology of a sub-AS or, due to the lack of some peering sessions between sub-ASs. We propose to solve this problem by running our algorithm for the iBGP topology of each sub-AS. Then, if diversity is not achieved, eBGP sessions have to be added between sub-ASs.

Finally, if the sets of prefixes received from the external peers of the domain change often, a dynamic solution for constructing the iBGP topology should be envisaged. In such a situation, the idea of Van den Schrieck et al. [23] that consists of going toward an automatic configuration of the iBGP topology should be considered.

VII. Conclusion

In this paper, we have illustrated the problem of NH diversity in an AS. We have shown that solutions have been proposed to solve the problem of learning multiple routes at the frontier of a domain, However, the problem of distributing these multiple routes in a domain remained to be solved.

To address this problem, we considered the design of iBGP topologies. We proposed an algorithm for the design of iBGP topologies in which it is ensured that each BGP router learns multiple Next-Hops for each destination. For this purpose, our algorithm relies on an initial iBGP route reflection topology. It proposes the addition of a few iBGP sessions to some border routers of the domain. These border routers receive a large number of external routes for which the BGP routers in the AS lack diversity.

We have shown by means of simulations that our algorithm leads to iBGP topologies where NH diversity is achieved in all the routers of the domain. This is true for all prefixes that are advertised at different external peering points. Moreover, our algorithm highlights the need to negotiate new external peerings to reach NH diversity for the remaining prefixes. For these prefixes, NH diversity cannot be achieved with any iBGP topology given the current external peering sessions.

We have also shown that NH diversity can be achieved with the addition of a low number of iBGP sessions. The total number of iBGP sessions in the AS remains very low compared to the number of sessions required by a full-mesh. Thus, we have shown that NH diversity can be achieved in a scalable way.

Acknowledgements

We thank Bruno Quoitin for making the C-BGP tool available to the research community. We also thank him for the excellent user support he provides.

References

[1] T. Bates, E. Chen, and R. Chandra, "BGP route reflection - an alternative to full mesh internal BGP (IBGP)," April 2006, RFC 4456.
[2] P. Traina, D. McPherson, and J. Scudder, "Autonomous system confederations for BGP," August 2007, RFC 5065.
[3] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed internet routing convergence," in *ACM SIGCOMM 2000*, August 2000.
[4] F. Wang, Z. M. Mao, J.Wang, L. Gao, and R. Bush, "A measurement study on the impact of routing events on end-to-end internet path performance," in *ACM SIGCOMM 2006*, September 2006.
[5] A. Bremler-Barr, Y. Afek, and S. Schwarz, "Improved BGP convergence via ghost flushing," in *IEEE INFOCOM*, March 2003.
[6] D. Pei, X. Zhao, L. Wang, D. Massey, A. Mankin, S. F. Wu, and L. Zhang, "Improving BGP convergence through consistency assertions," in *IEEE INFOCOM*, June 2002.
[7] J. Chandrashekar, Z. Duan, Z.-L. Zhang, and J. Krasky, "Limiting path exploration in BGP," in *IEEE INFOCOM*, March 2005.
[8] N. Kushman, S. Kandula, D. Katabi, and B. M. Maggs, "R-BGP: Staying connected in a connected world," in *4th USENIX Symposium on Networked Systems Design & Implementation (NSDI'07)*, April 11-13th 2007.
[9] W. Xu and J. Rexford, "Miro: multi-path interdomain routing," in *ACM SIGCOMM 2006*, September 2006.
[10] S. Uhlig and S. Tandel, "Quantifying the BGP routes diversity inside a tier-1 network," in *Proceedings of Networking 2006*, Coimbra, Portugal, May 15-19th 2006.
[11] C. Filsfils, "BGP convergence in much less than a second," June 2007, presentation at NANOG 40.
[12] L. Xiao, J. Wang, and K. Nahrstedt, "Reliability-aware iBGP route reflection topology design," in *11th IEEE International Conference on Network Protocols (ICNP)*, November 2003.
[13] Juniper Networks, Inc., "Configuring BGP routing - advertising routes: bgp advertise-best-external-to-internal," 2007, http://www.juniper.net/techpubs/software/erx/junose71/swconfig-bgp-mpls/html/bgp-config10.html.
[14] O. Bonaventure, C. Filsfils, and P. Francois, "Achieving sub-50 milliseconds recovery upon bgp peering link failures," in *Proceedings of the 2005 ACM CoNext*, 2005, pp. 31–42.
[15] M. Bhatia, "Advertising multiple nexthop routes in BGP," August 2006, internet draft, draft-bhatia-bgp-multiple-next-hops-01, work in progress.
[16] V. Van den Scrieck and O. Bonaventure, "Routing oscillations using BGP multiple paths advertisement," June 2007, internet draft, draft-vandenschrieck-bgp-add-paths-oscillations-00.txt, work in progress.
[17] M. Vutukuru, P. Valiant, S. Kopparty, and H. Balakrishnan, "How to construct a correct and scalable iBGP configuration," in *IEEE INFOCOM*, April 2006.
[18] M.-O. Buob, M. Meulle, and S. Uhlig, "Checking for optimal egress points in iBGP routing of a tier-1 AS," in *Proc. of the 6th International Workshop on Design and Reliable Communication Networks - DRCN'2007*, October 2007.
[19] M. Caesar, D. Caldwell, N. Feamster, J. Rexford, A. Shaikh, and J. van der Merwe, "Design and implementation of a routing control platform," in *Networked Systems Design and Implementation (NSDI)*, May 2005.
[20] B. Quoitin and S. Uhlig, "Modeling the routing of an Autonomous System with C-BGP," *IEEE Network*, vol. 19, no. 6, November 2005.
[21] N. Feamster and H. Balakrishnan, "Detecting BGP configuration faults with static analysis," in *Networked Systems Design and Implementation (NSDI)*, May 2005.
[22] T. Griffin and G. Wilfong, "On the correctness of iBGP configuration," in *ACM SIGCOMM 2002*, August 2002.
[23] V. V. den Schrieck, "Automating iBGP organization in large IP networks," in *Proc. ACM CoNEXT Student Workshop*, december 2007.