Providing scalable NH-diverse iBGP route redistribution to achieve sub-second switch-over time ⋆

Cristel Pelsser [a], Steve Uhlig [c], Tomonori Takeda [b,∗], Bruno Quoitin [d] and Kohei Shiomoto [b]

[a] *Internet Initiative Japan (IIJ), Japan*

[b] *NTT Network Service Systems Laboratories, Japan*

[c] *TU Berlin / Deutsche Telekom Laboratories, Germany*

[d] *Université de Mons (UMons), Belgium*

Abstract

The role of BGP inside an AS is to disseminate the routes learned from external peers to all routers of the AS. A straightforward, but not scalable, solution, is to resort to a full-mesh of iBGP sessions between the routers of the domain. Achieving scalability in the number of iBGP sessions is possible by using Route Reflectors (RR). Relying on a sparse iBGP graph using RRs however has a negative impact on routers' ability to quickly switch to an alternate route in case of a failure. This stems from the fact that routers do not often know routes towards distinct next-hops, for any given prefix.

In this paper, we propose a solution to build sparse iBGP topologies, where each BGP router learns two routes with distinct next-hops (NH) for each prefix. We qualify such iBGP topologies as *NH-diverse*. We propose to leverage the "best-external" option available on routers. By activating this option, and adding a limited number of iBGP sessions to the existing iBGP topology, we obtain NH-diverse iBGP topologies that scale, both in number of sessions and routing table sizes. We show that NH diversity enables to achieve sub-second switch-over time upon the failure of an ASBR or interdomain link. The scalability of our approach is confirmed by an evaluation on a research and a Service Provider network.

Key words: BGP, iBGP topology design, diversity, fast-recovery

⋆ Extended version of a paper published at ICC'2008. This work was done while Cristel Pelsser was at NTT.

∗ Corresponding author. Present address: NTT Network Service Systems Laboratories, 9-11, Midori-Cho 3-Chrome Musashino-shi, Tokyo 180-8585 Japan.

Email addresses: `cristel@iij.ad.jp` (Cristel Pelsser),

1 Introduction

The Internet is divided in domains, also called Autonomous Systems (AS). Each AS is usually administered by a single entity, such as a company or a university. The protocol currently deployed to distribute routing information between domains is the Border Gateway Protocol (BGP) [1]. In BGP, external BGP (eBGP) sessions are established to exchange routes with neighboring ASs. BGP routes are distributed inside an AS by means of internal BGP (iBGP) sessions.

A BGP route is composed of a prefix, a Next-Hop (NH), and a set of attributes. The attributes are used in the BGP decision process. The NH is the address of a router at the border of the domain. This router is able to farther 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 [1]. This leads to scalability issues in ASs with hundreds of routers. Today, the trend is to use Route-Reflectors (RR) [2] in large ASs. A RR may re-advertise routes learned on some iBGP sessions on some other iBGP sessions. Thus, they enable a reduction of the number of iBGP sessions established in the network and the number of routes maintained in the routers.

A router holds a routing table per BGP session (i.e., per BGP peer/neighbor). It stores the routes received on each session in these tables. A router may receive multiple routes for the same prefix from several neighbors. 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 selection process, called "decision process", is composed of a set of rules applied in sequence. A summary of these rules is provided in Table 1. 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.

1.1 The slow BGP convergence

The slow convergence of BGP has been highlighted in the literature. In [3], Labovitz et al. claim that recovery from a failure affecting inter-domain routes takes three

steve@net.t-labs.tu-berlin.de (Steve Uhlig),
takeda.tomonori@lab.ntt.co.jp (Tomonori Takeda),
bruno.quoitin@umons.ac.be (Bruno Quoitin),
shiomoto.kohei@lab.ntt.co.jp (Kohei Shiomoto).

Table 1
Simplified BGP decision process

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

minutes on average. Moreover, Wang et al. show in [4] that routing changes subsequent to a failure contribute significantly to end-to-end packet loss. Kushman et al. [5] measure the impact of BGP route updates on VoIP traffic. They designate BGP routing changes as being the cause of 50% of the perturbations in the VoIP calls they observed. Several techniques to improve BGP convergence have been proposed [6, 7]. However, as claimed by [8], reducing BGP convergence time is not sufficient, at least for the reliability required by loss and delay sensitive applications.

Solutions have been proposed in order for a domain to receive multiple AS paths to external destinations [8, 9]. These routes are present at the frontier of the domain. However, this diversity is not be redistributed to all routers inside the domain. Uhlig et al. [10] have demonstrated this for a Tier-1 Service Provider network making use of a hierarchy of RRs. Uhlig et al. have shown that 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 reach 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 NH 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. We confirm through measurements the significant gain in switch-over time when diverse NHs are present in the routers.

1.2 The complexity of iBGP design

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 routers might be considered. For example, not all routers are able to support the load incurred to RRs. Moreover, choices have to be made about the implications of the trade-offs on the iBGP topology design. For instance, operators have the choice between approaches requiring few iBGP sessions compared to solutions with lower amount of BGP messages exchanged upon failures. In this paper, we leave the decision of those trade-offs to the operators, and focus on NH diversity.

1.3 Problem statement

We say that *NH diversity* is achieved for prefix p in domain d if and only if, there are at least two BGP routes ρ_1 and ρ_2 with NHs NH_{ρ_1} and NH_{ρ_2}, s.t. $NH_{\rho_1} \neq NH_{\rho_2}$, in the routing tables of each BGP router in domain d. We only consider the BGP routers of an AS [1]. We note that NH diversity can be reached only if routes for the prefix are received at two ASBRs and from two different nodes in neighboring ASs. If there is no physical diversity at the border of the AS, it is not possible to reach NH diversity without negotiating additional external peering links.

An AS is NH diverse if NH diversity is reached in that AS, at every BGP router, for every prefix advertised in BGP. In a NH diverse AS, 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. NH diversity protects against the failure of the NHs and the links directly connected to the NHs, used to reach the NHs.

NH diversity in every AS along the path combined with IGP fast reconvergence, fast recovery or protection techniques in these ASs ensures that the BGP routes through the diverse NHs do not fail simultaneously upon the failure of a single resource or a Shared Risk Link Group (SRLG) along the active path.

We say that diverse NHs are *Shared Risk Link Group (SRLG) diverse* if there exists SRLG diverse paths to the NHs. If the diverse NHs are SRLG diverse, protection against SRLG failure to the active NH is ensured. Taking SRLGs into account ensures fast switch-over time in case of a SRLG failure toward the active NH.

We define diverse NHs as being BGP policy equivalent if their respective routes are policy equivalent. Two routes x and y for the same prefix are *policy equivalent* if and only if the output filters of the AS are configured such that both routes are allowed to be advertised on the same set of eBGP sessions. Guaranteeing diversity of NHs that are policy equivalent ensures that no BGP updates will be sent outside the AS [14] upon the failure of the route through the active NH. This enables failure restoration to be confined in the local AS.

In this paper, we propose an algorithm that leads to NH diversity in the routers by adding iBGP sessions to an initial topology of RRs. The proposed scheme is desired as soon as the administrators of an AS opt for a route-reflection topology.

[1] Often, all the routers of an AS are running BGP. When it is not the case, the routers that are not running BGP either have multiple default routes (primary and backups) or they are in the core of an MPLS network. In an MPLS core, the routers are configured with protection tunnels. Thus, non BGP routers can switch to an alternate route when the primary route fails.

Operators may adopt a route-reflection for multiple reasons among these is scalability of the routing tables size and easiness to introduce new routers in the network. Our algorithm aims at achieving this diversity by adding only a **limited number of iBGP sessions**. In the resulting iBGP configurations, each router learns at least two different NHs to reach every destination.

We limit ourselves to an objective of two diverse NHs per prefix inside each router. Our solution can easily be adapted to reach a larger NH diversity. However, the desire for more than 2 diverse NHs in the routers has to be weighted with the cost of maintaining these routes in the routers.

The alternate NH is solely used upon the failure of the route to the best NH. We do not aim to load balance traffic among the multiple NHs. As the alternate NH is only used temporarily, the quality of the path to the NH does not have to be as good as toward the best NH [2] . When an operator makes use of BGP attributes to indicate the preference of a BGP route, its objective is usually not to render a route unusable. Rather, its objective is to indicate that this route should be used only if better routes are not available. Our solution may enable this route to be used exactly in such a circumstance. Moreover, this does not compromise the correctness of the iBGP topology (see section 5).

In this paper, we present an algorithm that does not take SRLGs and BGP policies into account. However, these two aspects can be easily incorporated in our proposal. NHs that are not SRLG diverse or not BGP policy equivalent can be removed from consideration when selecting the candidate diverse NHs for a prefix.

Similarly, one operator may want NHs that are peering with different ASs. Again the algorithm can easily be modified to reach that goal. NHs that are peering with the same AS as the primary NH are removed from consideration in the set of candidate NHs for a prefix. The diversity one may want to consider depends on the type of failures that will occur. Unfortunately, we do not have a proper model of where failures happen in the Internet. We do not know if all the routes through an AS are more likely to fail simultaneously than the routes that go through different nodes. Thus, we cannot determine if such a consideration is meaningful.

Our algorithm aims at achieving route diversity by adding only a **limited number of iBGP sessions** to a sparse iBGP topology. We show that, for a particular research network, between 1.2% and 1.7% of the total number of sessions contained in a full-mesh are added to conventional iBGP topology designs. For the ISP network, between 0.6% and 1% of the sessions in a full-mesh are added to conventional iBGP topology designs. These, additional sessions bring new routes that need to be stored in the routing tables. We show that the increase in routing table is rather small, especially for the ISP network. For the research network, the initial routing tables

[2] We invite the reader to read [13] for a study on the cost along the backup routes provided by the algorithm in this paper.

are on average 3 times smaller than the tables with a full-mesh. After our design, they become 2 times smaller than the tables with a full-mesh. The tables in the routers of the ISP network are on average 3 to 7 times smaller with the traditional iBGP design techniques than with a full-mesh. They become 3 to 6 times smaller than with a full-mesh after the application of our algorithm.

This paper is structured as follows. First, we introduce the problem of lack of route diversity inside a domain in section 2. In section 3, we describe our methodology and our design algorithm. Then, we quantify the gain in switch-over time with our proposal, in section 4. The correctness of the NH-diverse iBGP topology is proved in section 5. Our proposal is evaluated in section 6. This section also contains a description of conventional iBGP topology designs. Then, we present solutions that have been proposed in the literature to solve related issues in section 7. Finally, we conclude the paper.

2 Lack of route diversity in BGP

In this section, we expose the causes for the lack of NH diversity in the routers of an AS. We show that the lack of NH diversity may occur both in full-mesh and route reflection topologies. The first two causes apply to both full-mesh and route reflection topologies. The last cause is applicable only to sparse iBGP topologies such as route reflection topologies.

When an iBGP full-mesh is used to propagate the external routes inside the AS, all the external routes that are chosen as best by the routers are known to the routers of the AS. An iBGP full-mesh might thus be seen as the ideal case for the visibility of the external routes. However, this apparently “ideal” situation of the full-mesh does not automatically imply that NH diversity is achieved. Three aspects that affect the diversity inside an AS:

- **Location of eBGP peerings**: If an AS Border Router (ASBR) has multiple peerings with neighboring ASs, external routes can be hidden at the ASBR and never be propagated inside the AS unless some failure occurs.
- **eBGP attributes of the routes**: The BGP decision process defines an ordering of the routes. The external routes that have the best ordering for the 3 eBGP attributes (highest `Loc_pref`, shortest `AS-path` length, and lowest `MED`[3]) dominate all the other routes learned by the routers of the AS. The dominated routes will never be selected by any router inside the AS, unless all the dominating routes are withdrawn. After the convergence of BGP, the dominated routes will

[3] If the “always-compare” option is used, the route with the lowest MED dominates the other routes. Without the option, the route with lowest MED dominates the other routes received from the same AS only.

only be present in the routing tables of the ASBRs that received them on an eBGP session.

- **iBGP propagation graph**: A sparse iBGP graph creates dependencies between routers. Routers that do not have multiple eBGP sessions must rely on their iBGP neighbors to achieve NH diversity. Routers without external peering sessions can only achieve NH diversity if their iBGP peers receive, select and advertise routes with different NHs.

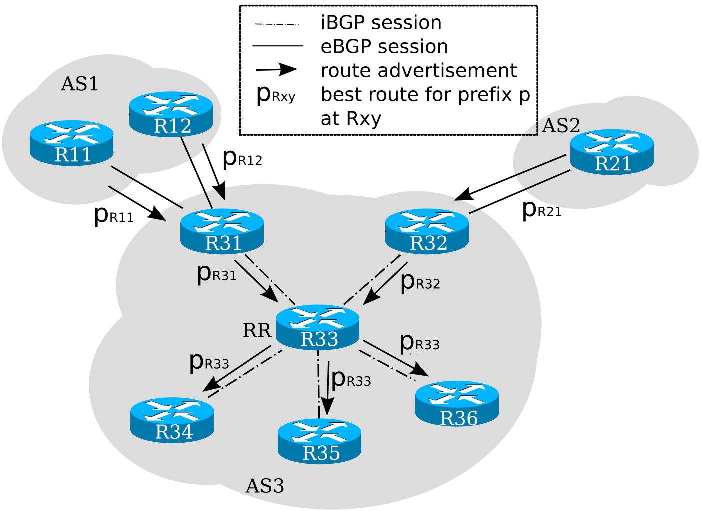


Fig. 1. Example of route diversity loss. Location of eBGP peerings: R31 selects either p_{R11} or p_{R12} as best route. R31 only propagates this route in iBGP. eBGP attributes of the routes: If R32 assigns a high `Loc_pref` to p_{R21}, neither p_{R11} nor p_{R12} will be propagated in iBGP. iBGP propagation graph: R33 selects either p_{R31} or p_{R32} as best route. R33 only sends this route to R34, R35 and R36.

The example of Fig. 1 illustrates the previous three aspects. Note that this example is not meant to represent a realistic iBGP topology, but is designed to illustrate the different reasons for lack of routing diversity in BGP. Such lack of route diversity has been observed in [10] and is illustrated in section 6 for traditional iBGP designs. In Fig. 1, we place ourselves as being the operator of AS3. This AS receives routes towards a prefix p from two neighboring ASs: AS1 and AS2. Two ASBRs of AS3 receive external routes toward p. However, only two of those three external routes can be used, as ASBR R_{31} only chooses and propagates one of the two external routes it receives from AS1. This illustrates the first reason for loss of route diversity: *the location of eBGP peerings*. Such route diversity at a single ASBR is not robust against the failure of the ASBR. In this case, changing the location of one of the eBGP peerings AS3 has with AS1, to some router who does not receive an external route towards p, will prevent one of the external routes from AS1 to be hidden by an ASBR. Note that in practice, changing the location of eBGP peerings is not so simple as it depends on the geographic location of the routers to be interconnected and on the availability of ports on these routers.

Among the three external routes that are received by the ASBRs of AS3, some routes might dominate others. For example, AS3 might prefer the routes from one of its neighboring ASs, AS2, due to routing policies. AS3 may assign a higher value to the `Loc_pref` attribute of the routes received from AS2. In this case, the two routes received from AS1 will not be selected as best by any router inside AS3. Thus, they will only be available at $R31$. The same reasoning applies if routes have the same value of the `Loc_pref` attribute but a shorter `AS-path`, or the same value of `Loc_pref`, the same `AS-path` length, but a lower value of the `MED` attribute. Routing policies can cause a great loss of diversity. They may prevent alternative NHs to be observed in a domain, unless some failure leads to the dominating route(s) to be withdrawn.

Router vendors are proposing an extension to BGP called "best-external" [15]. The use of this extension will help to solve the diversity problem when the lack of diversity is due to the *eBGP attributes of the routes*. With the "best-external" option activated, the ASBRs will advertise their best eBGP route to their iBGP peers. Thus, dominated routes may be propagated in the AS. In Fig. 1, with the "best-external" option, $R31$ would advertise one of its eBGP routes to $R33$ even if its eBGP routes are dominated. However, this route will not be propagated further in the AS. This option does not allow by itself to solve the diversity issue in a sparse iBGP topology. Because a "best-external" route is only propagated one hop in the iBGP topology, this option does not cause issues, such as routing loops, in the iBGP propagation of routes.

Finally, dependencies in the *iBGP graph* may create diversity loss during the propagation of the BGP routes. In Fig. 1, we intentionally built an iBGP topology that leads to a very poor diversity. As routers R_{34}, R_{35} and R_{36} depend exclusively on R_{33}. Those three routers will only see a single external route as being available to reach p; the route advertised by $R33$. We note that, even if these routers had iBGP sessions with several RRs, they may not learn routes with diverse NHs, as observed in section 6.

We have seen in this section that the lack of route diversity does not only concern route-reflector based iBGP topologies. It is a more generic problem. There is no protocol-based solution to counter the first cause. The ASBR is a single point of failure in this case. The establishment of additional eBGP sessions ending at diverse nodes counters this cause. The second cause can be countered by the use of "best-external". The last cause only occurs in sparse topologies such as route-reflection topologies and a confederation of ASs. The proposal in this paper enables to counter this last cause, for route-reflection topologies.

3 Improving diversity

In this section, we present our solution to reach NH diversity at the routers of an AS. By routers, we mean all the routers that rely on BGP routes to reach destinations external to the local AS. Our proposal consists of configuring the “best-external” option at the routers, coupled with an algorithm for the design of iBGP topologies. As a result, NH diversity is achieved at each router in the network, for all prefixes that are learned at different AS Border Routers (ASBR).

The “best-external” option is required because, as we have seen in section 2, in some configurations, it is not possible to achieve NH diversity without this option. In addition, we have shown that this occurs independently of the iBGP topology configured in the AS. When all the routers in a domain prefer the same route (i.e., the same NH) for a prefix, the routes that are received at other ASBRs for this prefix are not propagated in the domain.

3.1 Algorithm

Our algorithm determines a small number of iBGP sessions to add to an existing iBGP route reflection topology. The pseudo-code of our algorithm is provided in Algorithm 1. As input, the algorithm takes the eBGP routes received at the ASBRs, the IGP topology and an iBGP route reflection topology (line 1). Our solution relies on a tool such as [16] to compute the routing tables of the BGP routers in the domain (line 2).

The principle of the algorithm is as follows. First, we remove from consideration all the prefixes for which it is not possible to achieve NH diversity (Algorithm 1, line 3). These are the prefixes for which an external route is received only at one ASBR. For example, in Fig. 2 diversity cannot be reached for prefix $p1$. This is due to the fact that only ASBR $R21$ receives an external route for $p1$. From this step of the algorithm, a set S of prefixes is obtained. NH diversity will not be reached with any iBGP topology for prefixes that do not belong to S, even with a full-mesh. New external peering links need to be negotiated by the operator of the domain, in order to be able to achieve NH diversity for these prefixes. In the example of Fig. 2, the operator of $AS2$ could contact the operator of $AS1$ to schedule the establishment of a new link between $R22$ and $R11$.

After the removal from S of the prefixes for which diversity cannot be achieved, we compute the set of routers lacking NH diversity for a least one prefix in S (Algorithm 1, line 4). This set of routers is noted R.

The core of the algorithm is composed of the set of operations in lines 5 to 17. We call this set of instructions a *step* of the algorithm. These operations are performed

Algorithm 1 Addition of iBGP sessions

```
1:  $self.T$=$LoadTopology$()
2:  $self.RIBIn$=$self.ComputeBGPRoutes$()
3:  $self.S$=$self.RemoveUnsolvablePrefixes$()
4:  $R$=$self.GetLowDivRouterSet$()
5:  while ($|R| > 0$) do
6:      {improve diversity for one router}
7:      $r$=$self.GetMostInterestingRouter(R)$
8:      $P$=$self.GetLowDivPrefixSet(r)$
9:      $C$=$self.GetCandidateIBGPPeersSet(r,P)$
10:     if ($P \neq \emptyset$ and $C \neq \emptyset$) then
11:         {select candidate with maximum number of eBGP prefixes in $P$}
12:         $n$=$self.SelectNewIBGPPeer(r,C,P)$
13:         $self.T$=$self.AddIBGPSession(r,n)$
14:         $self.RIBIn$=$self.ComputeBGPRoutes$()
15:         $R$=$self.GetLowDivRouterSet$()
16:     end if
17: end while
```

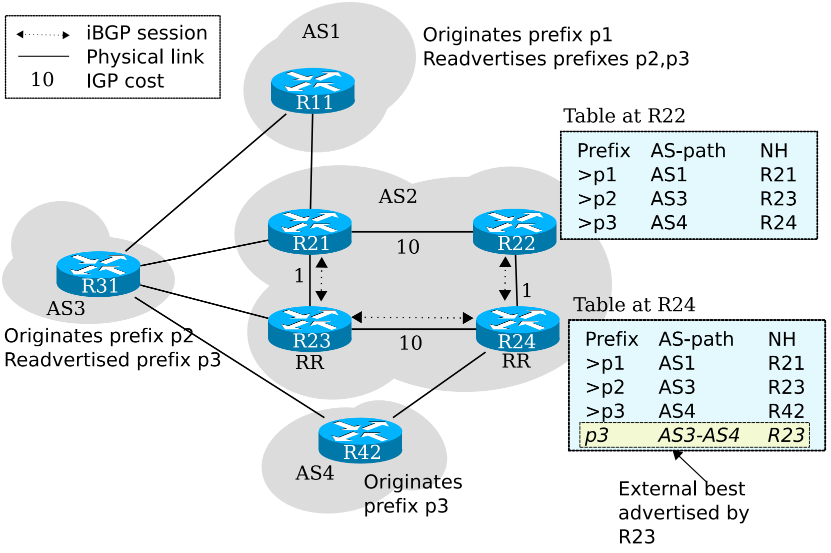


Table at R22

Prefix	AS-path	NH
>p1	AS1	R21
>p2	AS3	R23
>p3	AS4	R24

Table at R24

Prefix	AS-path	NH
>p1	AS1	R21
>p2	AS3	R23
>p3	AS4	R42
p3	*AS3-AS4*	*R23*

Fig. 2. Improving diversity

until NH diversity is reached for all the routers in R. First, we pick a router r from the set of routers lacking diversity, R (line 7). Then, we improve NH diversity for r through the addition of an iBGP session between r and an ASBR. With “best-external”, 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 thus good candidates for becoming new iBGP peers. An ASBR is selected to become a new iBGP peer for r if adding a session to this ASBR increases NH diversity for the largest number of prefixes at r.

Assume that router r is $R22$, in Fig. 2. We see that $R22$ lacks NH diversity for prefixes $p2$ and $p3$ (line 8). Then, we determine the set of ASBRs that are candidates to become new iBGP peers for r (line 9). Routers $R21$, $R23$ and $R24$ are ASBRs in $AS2$. In the example, $R21$ distributes a route with NH $R21$ for prefixes $p2$ and $p3$ to its iBGP peers. $R23$ sends routes for prefixes $p2$ and $p3$ with NH $R23$. And, $R24$ only sends a route for $p3$ with itself as the NH.

Some iBGP sessions do not 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 neighbor, (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. These routers are not candidate ASBRs. In the example, the algorithm will not propose to add an iBGP session between $R22$ and $R24$ because of (1). Only $R21$ and $R23$ belong to the set C of candidate iBGP peers (Algorithm 1, line 9).

The algorithm now chooses between $R21$ and $R23$ as new iBGP peer (Algorithm 1, line 12). For this purpose, it determines the ASBR that will 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, $p2$ and $p3$. An iBGP session with $R23$ will only increase NH diversity for prefix $p3$ because $R23$ is already the NH for $p2$ at $R22$. 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 arbitrarily[4] .

Then, we recompute the BGP routes received at the routers (Algorithm 1, line 14). For this purpose, we first add the new session to the model of the iBGP topology at line 13. The BGP routes are computed after the addition of each iBGP session because when a router receives additional routes, it may select different routes as best. Subsequently, it stops advertising the previous best routes to its peers. This may lead to reduced NH diversity in some routers. We note that this occurs only when the initial iBGP topology is not *fm-optimal*. That is, when routers cannot choose as best route the one towards the closest NH in terms of IGP cost. This concept is defined in [17]. It is desirable that an initial iBGP topology meets this fm-optimality constraint in order to avoid deflection and forwarding loops. In case of fm-optimality of the iBGP topology, this route computation step is not necessary.

3.2 External routes model

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

[4] The tie-beak may be based on the IGP cost to the ASBR or the peering cost at the ASBR. The choice is left to the operator.

diversity in the AS. To avoid having to re-optimize 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. This part of the contract the SP 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 classes of prefixes to expect in a subset. 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. In our model of the routes, a single prefix is used for each class of prefixes. An iBGP session that is added to improve diversity for this prefix improves diversity for all the prefixes in the class. Such a model has already been used in [10, 16, 18]. 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. That is, if a prefix is added or removed from a class, diversity is maintained. The model can also take into account predictions of changes in agreements and of the removal or the addition of external peers. We note that the real eBGP routes can be used instead of building such a model.

3.3 Properties of the solution

Our algorithm adds iBGP sessions to ASBRs that receive many external routes. Thus, an ASBR that receives many routes should be able to support a higher number of iBGP sessions than other ASBRs. This effect is predictable. Therefore, those ASBRs can be correctly dimensioned to support the additional load. Moreover, this aspect can be taken into account when selecting a location for the addition of external peerings. We note that the number of iBGP sessions at an ASBR will never be larger than the number of sessions it would have to support in a full-mesh. We show in section 6.3 that the average and maximum number of iBGP sessions supported by the routers in a ISP network is much lower than the number of sessions to be supported by a router in an iBGP full-mesh.

The addition of iBGP sessions enables us to achieve NH diversity at the cost of a limited increase in the amount of routes to be maintained in the routing tables. We evaluate this cost in section 6. We see in that section that there is a trade-off. The size of the routing tables is kept smaller in the situations requiring the addition of a larger number of iBGP sessions at the ASBRs, such as the research network (see section 6.2). When diversity is easily achieved, because diversity is largely present at the border of the domain, a very small number of additional sessions leads to a larger increase of the routing table sizes. This is observed in the evaluation of our proposal for the ISP network (see section 6.3).

The strength of our approach is that it is applicable today. No changes are required to the implementation of BGP [5] . Moreover, as we will see in section 6, the iBGP route reflection topologies that are generated by our algorithm are rather small, especially for larger topologies. They require a small average number of sessions and routing entries to be maintained at the routers.

4 Switch-over time

Failure recovery can be divided into three steps: failure detection, failure notification and route switch-over. NH diversity aims at reducing the switch-over time. Once a BGP route is withdrawn or once the router learns that the current NH of a route is no more reachable, it is able to directly switch to the NH-diverse route.

There are multiple ways to speed up failure detection and notification. The IGP [11, 19] and the Bidirectional Forwarding Detection (BFD) protocol [20, 21] enable a router to learn the occurrence of a distant failure within a few hundreds of milliseconds.

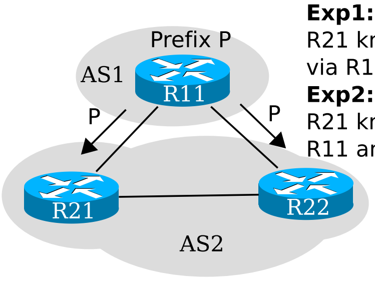
After the few hundreds of milliseconds necessary for a router to detect a failure, the alternate NH, available at the router in a NH-diverse network, is installed in the Forwarding Information Base (FIB). This takes around 100ms, with a hierarchical FIB [11]. With such a FIB architecture, installing a new NH upon a failure does not depend on the number of prefixes impacted by the failure anymore. It only consists of the time that is required to change the value referenced by a pointer.

Thus, **in a domain with NH diversity, fast failure notification and a hierarchical FIB architecture, switch-over can be achieved in much less than a second** [11]. This is a significant improvement compared to the few tens of seconds required today, in ASs without NH diversity.

We perform measurements to determine the gain in switch-over time when diverse NHs are present in a commercial router, without any fast notification mechanisms and without a hierarchical FIB. First, we measure the recovery time at $R21$ for 10000 routes upon the failure of link $R21 - R11$, in the topology illustrated in Fig. 3. Without NH diversity, $R21$ takes 5.85 seconds, on average, to detect the failure, learn the new route and install it in its FIB. With NH diversity, 0.92 seconds are required for $R21$ to detect the failure and install the NH diverse route in its

[5] We expect the “best-external” option to be delivered for most routing equipment very soon. The market leaders are pushing the standardization of this solution at the IETF. Moreover, one implementation is already available.

FIB. Switch-over time is reduced by 4.93 seconds with NH diversity. It is already a significant gain for such a simple topology.



Exp1: No NH diversity
R21 knows only route via R11 to prefix P
Exp2: NH diversity
R21 knows route via R11 and R22

Fig. 3. Switch-over time measurements

Now, we perform similar measurements on a larger topology, the topology illustrated in Fig. 2 in [22]. We measure the recovery time at $PE3$ after the failure of link $PE2 - CE1$. Again, we do not use any fast failure notification mechanisms and $PE3$ does not possess a hierarchical FIB. Without NH diversity, it takes 44 seconds for $PE3$ to learn the alternate routes and update the 10000 route entries. With NH diversity, only 2 seconds are required. Out of these two seconds, it takes already one second for $PE3$ to be notified of the failure. Fast failure notification techniques do not necessarily help to speed up the recovery time when a diverse NH is not known. However, with NH diversity, a fast notification technique and hierarchical FIBs, a sub-second recovery time is at our reach.

5 Correctness

The iBGP sessions added to the initial iBGP topology, in order to reach NH-diversity, do not affect the correctness of the iBGP topology. The sessions that are added are of type “over”, according to the terminology introduced in [23]. A route learned on such an iBGP session is not readvertised on another iBGP session. Thus, the NH-diverse routes will not be re-distributed in iBGP. Moreover, if the initial iBGP topology is correct, it is likely to be “fm-optimal”. Designing iBGP topologies that are fm-optimal even in case of failures is possible. “Fm-optimality” [17] ensures that any router can choose as best route the one that is advertised by the closest egress point, with regard to the IGP cost, as in a full-mesh. In an “fm-optimal” iBGP topology and an iBGP topology “fm-optimal” with regard to failures [24], the routes learned on the additional iBGP sessions do not change the output of the BGP decision process in the routers. The NH-diverse route is never selected as best route instead of the primary route, when there is no failure. After a failure, at the end of the BGP convergence, the best backup routes are learned via the sessions of the initial iBGP topology. If the NH-diverse route is the best, it is also learned via the sessions of the initial iBGP topology. Therefore, the additional sessions do not affect the convergence of BGP inside the AS. They will not be used to forward packets in normal network operation nor be re-distributed. Thus, they will not lead to forwarding loops nor route oscillations. Rather, they allow that an alternative route be used upon a network failure, before the final route is learned. As a conclusion, if the initial iBGP topology is fm-optimal, the iBGP topology we produce is fm-optimal and correct.

Transient forwarding loops may occur during any BGP convergence. The iBGP topologies that we generate do not make an exception. We note that, the state of the art to ensure that no forwarding loops occur during the convergence of BGP, is to encapsulate traffic to the outgoing interface of the ASBR [25].

6 Evaluation

We perform our evaluation on two types of networks: a research network and an ISP network. The ISP network topology has been inferred by the rocketfuel project [27]. For each network, we study the NH diversity achieved with conventional iBGP topology designs. We compare the diversity reached by the conventional iBGP topologies with the one achieved by the iBGP topologies generated by our algorithm [6] . Then, we study the scalability of our proposal. We examine the number of iBGP sessions in conventional iBGP topology designs, in the iBGP topologies that we generate, and in a full-mesh. We also study the number of iBGP sessions and the amount of routing table entries that need to be supported by each router. The “best-external” option is activated for each simulation.

6.1 Settings of the simulations

6.1.1 Model of a research network

We construct the model of the research network used in our evaluation based on public information relative to its topology and external peers. The intra-domain topology of the research network is available on its website (`http://two.wide.ad.jp/`). The research network is composed of 17 nodes. Eight of these nodes are ASBRs. Similarly, a list of its external peers is also available on the website of the research network. The research network has 12 external peers.

We follow the methodology introduced in section 3 to model the external routes received from each peer. First, we determine the roles of the external peers. These roles are deduced from studies on the relationships between ASs, such as [26], and from the service the peers advertise on their websites. We conclude that two of the

[6] As mentioned in section 1.3, we do not consider SRLGs in this paper even though it is rather straight forward to take them into account. The reason is that we don’t know the SRLGs for the networks considered in the evaluation.

peers are well known commercial Internet Service Providers (ISP). Moreover, four peers are research networks. Finally, there are six connections to major Internet eXchange (IX) points. We looked at the IXs' websites to determine the peers connected to the IXs. From this information, we deduced the classes of external routes received from the commercial providers, the research network peers and at the IXs.

6.1.2 Model of an ISP network

In this section, we describe the model of the ISP network that we use in our evaluation. Mahajan et al. [27] have inferred the internal topology of a few Internet Service Provider networks. For each of these ISP networks, they have inferred the link costs and the PoP structure of the network. We use their model of one of the ISP networks, AS1239. Their model is composed of 315 nodes spread across 44 PoPs. We use the AS relationships inferred by Subramanian et al. [26] to determine the external peers of AS1239. According to [26], AS1239 is one of the few tier-1 ASs that are in the core on the Internet. It is connected to 1750 other ASs. Among these ASs, 41 ASs are shared-cost peers. The remaining 1709 ASs are customer ASs of AS1239. The shared-cost peers are ASs of about the same size as AS1239. We assume that these ASs advertise a large number of prefixes, including customer ASs' prefixes. In our model, each shared-cost peer has between 2 and 4 peering links with AS1239. Each peering link ends at a random node in a randomly selected PoP of AS1239.

We build the model of the external routes advertised by each shared-cost peer as follows. First, we assume that the Internet is divided in three major geographic areas, the continents. Furthermore, each of these areas is divided into 30 regions. A region may represent a country. We consider all the 1750 peers connected to AS1239. Each peer is assigned a geographic coverage. Depending on its level in the AS hierarchy inferred by Subramanian et al. [26], it is assumed that a peer covers a wide or a small geographical area. For example, tier-1 peers are assumed to cover all the three major continents. Level-2 and level-3 peers are assumed to cover a single continent and all the countries in this continent. Finally, level-4 and 5 peers cover a single region. The countries are assigned randomly to the level-4 and level-5 peers. Similarly, the continents covered by level-2 and 3 peers are also assigned randomly. Secondly, a prefix is assigned to each region and each continent that is covered by one of the peers of AS1239. Finally, a shared-cost peer advertises to AS1239 the prefixes that are attributed to the regions and areas that it covers. Together, the shared-cost peers of AS1239 cover all geographical areas. Thus, even though the 1709 customer ASs are not directly connected to AS1239, the shared-cost peers advertise their prefixes to AS1239.

6.1.3 Conventional iBGP topologies

In this section we describe the iBGP route-reflection topologies that are used as input to our algorithm for our evaluation. They are also used as a reference point for assessing the NH diversity and iBGP sessions scalability in real networks. These topologies are the result of conventional iBGP design methodologies. Those conventional topologies should be similar to iBGP topologies used by ISPs.

6.1.3.1 Bates recommendation In [2], Bates et al. state some recommendations for iBGP route reflection topologies. They advise to configure one or multiple RRs per Point of Presence (PoP) in the network. All the routers in a PoP are clients of the RR(s) in this PoP. In addition, the authors require a full-mesh of iBGP sessions between the RRs. Moreover, they also recommend the configuration of a full-mesh of iBGP sessions between all the routers in a PoP.

6.1.3.2 "Bates1" iBGP design In our first initial iBGP topology, we implement this recommendation as follows. The most connected router in each PoP is selected to be the RR. Each router is a client of the RR in its PoP. A full-mesh of iBGP sessions is established between the RRs. Finally, there is a full-mesh of iBGP sessions between all the routers in a PoP. In the remaining of this paper, we call this iBGP topology *"Bates1"*. An overview of the properties of this topology is given in the first line of Table 2.

6.1.3.3 "Bates2" iBGP design Our second initial iBGP route-reflection topology is built as follows. Two RRs are selected in each PoP for redundancy purposes. These two RRs are the two most connected routers in the PoP. All the routers in a PoP are iBGP clients of the two RRs in the PoP. Moreover, a full-mesh of iBGP sessions is configured between the RRs. This topology also follows the recommendations of Bates et al. [2], expressed earlier. It is called *"Bates2"* in the following sections. A short description of this topology is provided in Table 2.

6.1.3.4 "Zhang" iBGP design Large Service Provider networks may make use of a hierarchical route-reflection topology [10]. Such a topology is characterized by multiple levels of RRs. Routers that are clients of RRs at the top-level may on the other hand be RRs for routers at lower levels. In [1], Zhang and Bartell provide recommendations for the design of such hierarchical iBGP topologies. They say that the RRs at the top-level must be fully meshed. On the contrary, this is not required for RRs at lower levels.

Our third initial iBGP topology verifies the recommendations in [1]. It is built as follows. There are two levels of RRs. At the lowest level, the routers of a PoP are clients of two RRs in the PoP. These two RRs are the most connected routers of the PoP. In turn, these RRs are clients of two RRs at the top-level. RRs are at the top-level of the hierarchy if the number of nodes in their PoP is above a critical number. This number is set to 10 in our topology. However different values can be envisaged. Low-level RRs of a PoP are connected to the two top-level RRs of the closest PoP. The closest PoP is determined based on the IGP cost of the links. Finally, a full-mesh of iBGP sessions is configured between the top-level RRs. Such a configuration is illustrated in Fig. 7-11, page 265 of [1]. We call this topology: *"Zhang"*. Table 2 provides a brief description of this iBGP route-reflection topology.

Table 2
Conventional iBGP topologies

Name	hierarchy	top-level full-mesh	PoP full-mesh	RR redundancy
Bates1	no	yes	yes	no
Bates2	no	yes	no	yes
Zhang	yes	yes	no	yes

Since the number of nodes in the research network model is rather small, it is not relevant to make use of a hierarchy of RRs. In order to obtain a model with a larger number of nodes for the research network, we proceed as illustrated in Fig. 4. This enlarged model of the research network will only be used with the conventional hierarchical iBGP topology, "Zhang".

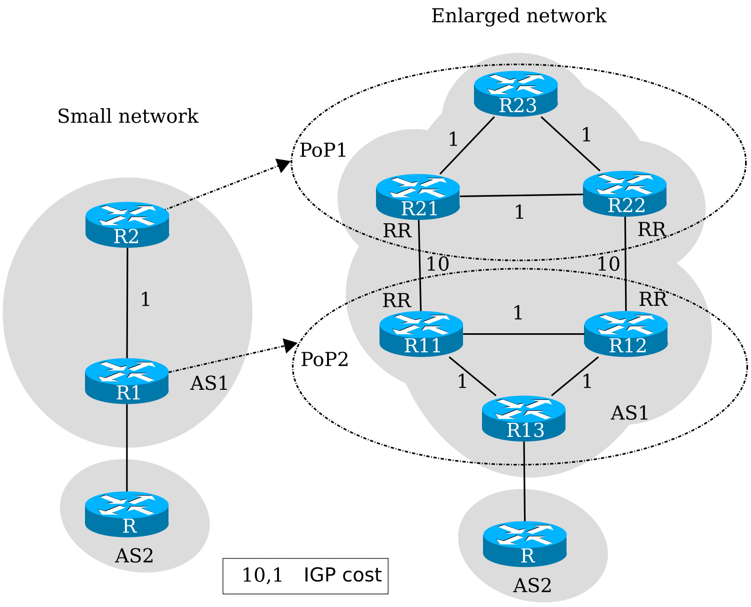


Fig. 4. Enlarging a network

6.2 Evaluation for the research network

In this section, we evaluate the effectiveness of our design algorithm for the research network presented earlier. The "best-external" option is activated for all the simulations. First, we study the NH diversity present in the routers with the three conventional iBGP topologies introduced in section 6.1.3. We observe that this diversity is poor. Then, we show that with the iBGP topologies resulting from our algorithm, we achieve the same NH diversity as with a full-mesh of iBGP sessions. From the distribution of the number of iBGP sessions at the routers, we show that the total number of iBGP sessions present in the topologies generated by our algorithm is low compared to the number of sessions in an iBGP full-mesh. Finally, we draw conclusions on the size of the routing tables.

Table 3 shows the average percentage of prefixes for which NH diversity is observed in the routers of the research network. This value is equal to

$$\frac{100}{p} * \frac{\sum_{i=0}^{n} d_i}{n}, \tag{1}$$

where n is the number of routers, p the total number of prefixes and d_i the number of prefixes with NH diversity at router i. Each line in Table 3 relates to a different initial iBGP topology.

Table 3
Research network: NH diversity

	iBGP topologies		
Name	initial	proposed	full-mesh
"Bates1"	39%	90%	90%
"Bates2"	26%	90%	90%
"Zhang"	13%	90%	90%

We observe in the second column of Table 3 that with the "Bates1" iBGP topology, diversity is achieved for 39% of the prefixes, on average over all the routers in the network. With the "Bates2" iBGP topology, the average NH diversity in the routers is lower, with 26% of the prefixes. Finally, with the "Zhang" iBGP topology, there is NH diversity for only 13% of the prefixes in the routers on average. The differences in diversity observed with these three types of iBGP topologies come from two types of aspects. First, the iBGP designs used lead to different number of iBGP peerings. The "Bates1" iBGP topology has most iBGP sessions, then "Bates2" comes in the middle, and finally the "Zhang" topology has the lowest number of iBGP sessions. Less iBGP sessions reduce the possibility of learning routes with alternative NHs. Furthermore, relying on a full-mesh within a PoP increases the chances of learning more than a single NH, hence improving diversity.

One NH may be learned from each RR and other NHs may be learned from the routers in the PoP.

Now let us look at the NH diversity achieved with our solution, in the third column of Table 3. We observe that NH diversity is achieved at all routers for 90% of the prefixes. The other prefixes are advertised by a single eBGP peer in our model. Moreover, we note that the diversity obtained with our iBGP topology is the same as the diversity observed in a topology with a full-mesh of iBGP sessions (fourth column in Table 3). 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 route reflection topology generated by our algorithm is very important. It enables us to detect situations where the only solution to achieve diversity requires the establishment of new external peerings. Here, we deduce from Fig. 3 that new external peerings session should be negotiated to reach diversity for 10% of the prefixes.

Fig. 5 provides statistics on the number of iBGP sessions configured at the nodes, in different iBGP topologies. On the y-axis, we have the number of iBGP sessions at a router, normalized by the number of iBGP sessions at a router in an iBGP full-mesh. The number of iBGP sessions observed in a full-mesh is our reference point. A router in a full-mesh supports the maximum possible number of sessions, or 100%. On the x-axis, we have the different iBGP topologies. For each iBGP topology, we show the minimum, average and maximum number of iBGP peers that are observed at the nodes in this topology. The conventional iBGP topologies are labeled “init”. The topologies generated with our algorithm are labeled “prop”. Full-mesh iBGP topologies are labeled “f-m”.

We observe in Fig. 5 that the average number of iBGP sessions to be supported at a router, with our proposal, is much lower than in a full-mesh. With the iBGP topologies we generated from “Bates1” and “Bates2” iBGP topologies, only two routers have has many iBGP sessions as in an iBGP full-mesh. These routers are RRs that learn many prefixes on their eBGP sessions. In the topology generated from the “Zhang” input iBGP topology, the maximum number of sessions a router supports is almost two times smaller than the number of sessions routers must keep under a full-mesh. Again, the routers with the largest number of sessions are ASBRs receiving many external routes. As mentioned in section 3, such an effect is predictable. Appropriate dimensioning of these routers and proper selection of external peering locations is thus possible. Moreover, such ASBRs are typically high-end routers that can easily sustain the stress from many sessions.

In Fig. 5, we see that on average, the nodes have two times less iBGP peers in the iBGP topologies that we generate based on “Bates1” and “Bates2” iBGP topologies than in a full-mesh. Moreover, the average number of sessions at a router is very small, in the iBGP topology generated from the “Zhang” initial topology. It is 4

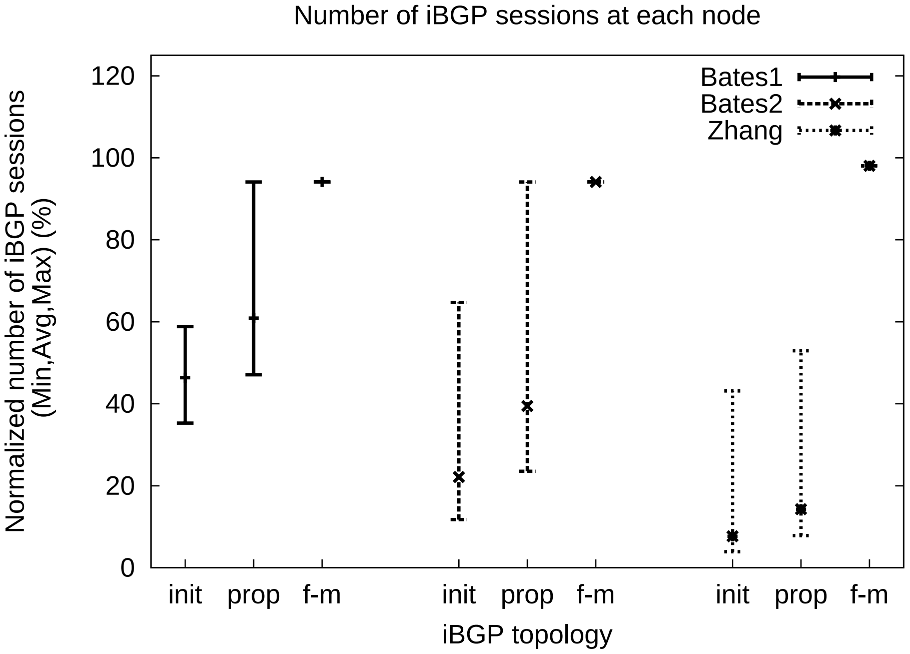


Fig. 5. Research network: Distribution of the iBGP sessions on the routers

times lower than the number of sessions that need to be supported by the nodes in a full-mesh of iBGP sessions.

In Fig. 5, the middle point provided for each iBGP topology also indicates the percentage of iBGP sessions in the topology. When looking at these average values, we observe that our algorithm generates iBGP topologies with far less iBGP sessions than in a full-mesh. Moreover, we see that the number of iBGP sessions in the topologies generated by our algorithm varies based on the initial iBGP topology that is provided as input to our algorithm. **The sparser the original iBGP topology is, the lower the number of total iBGP sessions required in order to reach the target NH diversity with our approach.** The initial "Zhang" iBGP topology is the sparsest iBGP topology. It contains only 8% of the sessions in a full-mesh. The iBGP topology generated from the "Zhang" topology with our algorithm is only composed of 15% of the sessions in a full-mesh. As illustrated in Table 3, while the initial topology provides very poor NH diversity, the same diversity as in a full-mesh is achieved with the resulting iBGP topology. Our approach hence does not require that the original iBGP topology be particularly well designed to work well.

With our proposal, diversity is achieved because additional routes are exchanged compared to the initial iBGP topology. However, this process increases the size of the routing tables. In the conventional iBGP topologies, routers store between 25 and 27 routes on average. In the iBGP topologies generated with our algorithm, the routers store an average of 39 to 40 routes. The average number of routes in the router is reasonably low compared to a full-mesh. With a full-mesh, the routers receive 79 routes on average.

Our study has shown that the total number of iBGP sessions is kept low compared to the number of sessions in a full-mesh. The average number of iBGP sessions and routes to be supported by the routers is also kept low compared to a full-mesh. The number of sessions and routes to be supported is higher at ASBRs receiving a large number of external routes. This is unlikely to be a problem as those large ASBRs will be dimensioned to support a large number of sessions, because they are located at important peering points.

6.3 Evaluation for an ISP network

In this section, we evaluate the efficiency of our iBGP topology design algorithm when applied to a large ISP network. For this purpose, we use the network model presented in section 6.1.2. A different set of external peerings and external routes is used for each initial iBGP topology.

First, we examine, in the second column of Table 4, the average NH diversity achieved in the routers, with the three conventional iBGP design techniques presented in section 6.1.3. We observe that, with the conventional iBGP topologies, a router on average has NH diversity for 14% and 15% of the prefixes. Even though the three conventional iBGP design techniques lead to topologies with different numbers of iBGP sessions, their NH diversity is similar. This confirms the findings of Uhlig et al. [10].

Table 4
ISP network: NH diversity

	iBGP topologies		
Name	initial	proposed	full-mesh
“Bates1”	14%	100%	100%
“Bates2”	15%	100%	100%
“Zhang”	14%	100%	100%

When we look at the NH diversity in routers with the iBGP topologies designed by our algorithm, the column labeled “proposed” in Table 4, we see that diversity is achieved for all the prefixes in all the routers, as with a full-mesh.

We see in Fig. 6 that the number of sessions to be supported by the routers with the conventional iBGP designs as well as with our proposal is low compared to a full-mesh. With the conventional iBGP topology designs, the routers support on average 6.8% (“Bates1), 7.8% (“Bates2”) and 1.7% (“Zhang”) of the iBGP sessions they would support in a full-mesh. In the topologies generated by our algorithm, these numbers become 7.8%, 8.4% and 2.5%, for the topologies based on the “Bates1, “Bates2” and “Zhang” initial topologies, respectively. Thus, on average, the routers

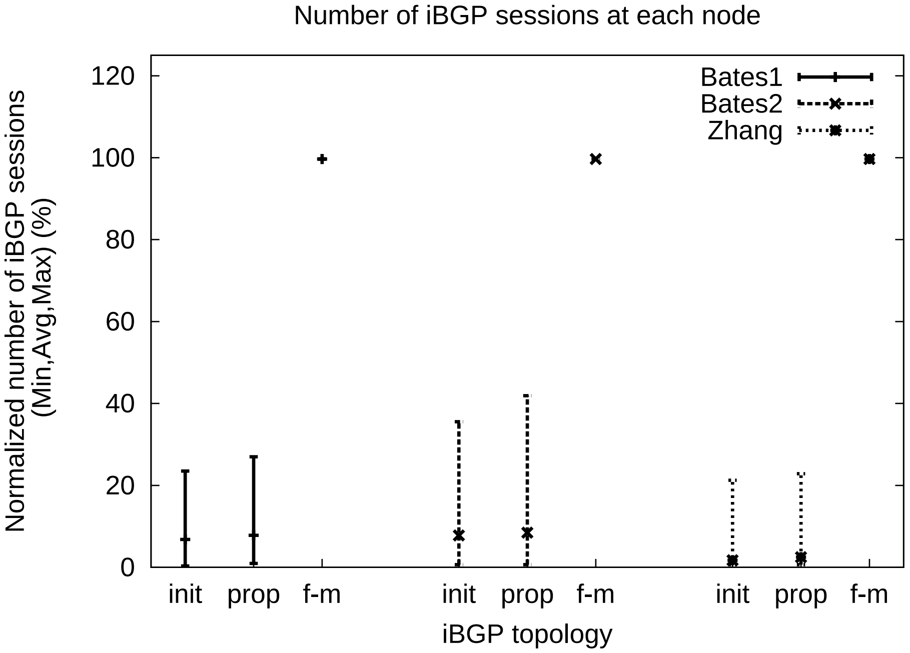


Fig. 6. ISP network: Distribution of the iBGP sessions on the routers

support a similar number of sessions with the conventional topologies and their resulting NH-diverse iBGP topologies. Moreover, there is only a small increase in the maximum number of sessions supported by the routers with the NH-diverse iBGP topologies, compared to their initial conventional topologies.

With our algorithm, 487 iBGP sessions are added to the "Bates1" iBGP topology in order to reach 100% NH diversity. This corresponds to an increase of 1% of sessions. Similarly, our algorithm proposes 331 additional iBGP sessions (0.6% of sessions) to "Bates2" iBGP topology for NH diversity to be achieved.

As observed for the research network, we can also see the benefit of using a hierarchy of RRs for the ISP network. We observe, in Fig. 6, that the "Zhang" iBGP topology is composed of only 1.7% of the sessions contained in a full-mesh. With our solution, 368 sessions are added to the "Zhang" initial iBGP topology. This results in a topology composed of only 2.5% of the sessions that would be established in the case of a full-mesh. It is hence possible to achieve NH diversity with a scalable number of iBGP sessions.

Vutukuru et al. [28] obtained an iBGP topology with 8 levels of RRs and 26% of the sessions of an iBGP full-mesh, for the same ISP network with the same intra-domain topology as in our model [27]. Their design objectives are very different from our NH diversity objective as they provide reliability to IGP failures. By comparison, with a two level iBGP topology, we achieve NH-diversity with only 2% of the sessions present in a full-mesh.

When looking at the size of the routing tables, we observed that the average routing table sizes are 3 to 7 times smaller with the conventional iBGP topologies than with the full-mesh. Additionally, on average, the routing tables are 3 to 6 times smaller with the iBGP topologies resulting from our algorithm compared to the sizes obtained with the full-mesh. We note that, for the three initial designs, at most 100 additional routes are maintained in average at the routers with our proposal compared to the original conventional iBGP design. We believe that this is a reasonable increase. Similar to the small increase in iBGP sessions, the increase in routing table size with our proposed iBGP topologies is small compared to conventional iBGP topologies.

6.4 Computation time

In this section, we first discuss the theoretical complexity of our algorithm. Then, we look at the time required by our algorithm for the computation of iBGP topologies with NH diversity. We relate this information to the number of times the instructions in the loop of our algorithm are executed. We call this number, the *number of steps*.

Let n be the number of nodes in a network. With our algorithm, at most $m - \frac{n.(n-1)}{2}$ iBGP sessions may be added to an initial iBGP topology. Where m is the number of sessions in the initial topology. The BGP routes are computed after the addition of each iBGP session. This results in at most $m - \frac{n.(n-1)}{2}$ BGP route computations. The generation of each of these iBGP topologies has a computational time complexity of $\mathcal{O}(n^2.p)$, where p is the number of external prefixes distributed by BGP.

Providing the theoretical time complexity for the computation of BGP routes remains an issue [29]. If the initial iBGP topology is fm-optimal, the computation of the BGP routes is not necessary to determine the impact of each additional iBGP sessions on the NH diversity in the network. Such an impact is easily predictable. Thus, BGP routes only need to be computed once, at the beginning of the algorithm, to determine the initial diversity. Alternatively, BGP routes may also be retrieved directly from the routers in the operational network. In the latter case, there is no BGP route computation. The complexity of our algorithm becomes $\mathcal{O}(n^4.p)$.

To have an idea of the practical execution time of our algorithm for the research network, we measured, for each conventional iBGP topology, the time required to generate ten different NH-diverse solutions. For this purpose, we replace the $GetMostInterestingRouter$ function (Algorithm 1, line 7) and the tie-breaking function used as the final decision for the selection of a new iBGP peer, inside the $SelectNewIBGPPeer$ function (Algorithm 1, line 12), by a random selection. We observed that all simulations with the “Bates1” initial topology have similar execution times: 2.41 seconds on average. Moreover, they all require 21 steps. As each step of the algorithm adds an iBGP session to the iBGP topology, 21 sessions are added in all ten executions of the algorithm. The same observation holds for the simulations with “Bates2” as initial iBGP topology. With the “Bates2” input iBGP topology, our algorithm always suggests the addition of 25 sessions. This design takes 2.53 seconds, on average.

Concerning the simulations using the “Zhang” iBGP topology as input, we note very little variability in the execution times and in the number of steps carried out. Execution times are comprised between 25.07 and 26.81 seconds. The number of steps and the number of additional sessions is between 86 and 90. This represents a variation of only 0.3% of the sessions contained in a full-mesh. Thus, the choice of a particular objective for the $GetMostInterestingRouter$ function and the tie-break function does not have a significant impact on the resulting iBGP topologies.

For the ISP network model, the design of the iBGP topologies was achieved in 227, 262 and 106 minutes, from “Bates1”, “Bates2” and “Zhang” initial iBGP topologies, respectively. This network is composed of 315 nodes. It is larger than the research network. Moreover, more steps are required to reach a solution compared to the research network. A NH-diverse solution is reached in 486, 331 and 368 steps for the “Bates1”, “Bates2” and “Zhang” initial iBGP topologies, respectively. BGP routes are computed 486, 331 and 368 times.

The same amount of steps would be required when considering the complete routing information instead of classes of prefixes. Grouping the prefixes in classes aims to provide scalability for BGP route computation.

Since most transit networks rely on hot-potato routing [30], we believe that the initial iBGP topology of many ISP networks is fm-optimal. For these networks, BGP route computations are not required. The time to generate a NH-diverse iBGP topology is thus much shorter as in our evaluation for the ISP network. A new NH-diverse iBGP topology can be computed, upon a change in the set of prefixes received from the external peers. We do not expect this to occur frequently. The set of routes received from peers currently depends on the details of the peering agreements that are negotiated between the concerned ASs. In this section, we have shown that the execution time of our algorithm is a function of the number of steps required to find a solution, and thus, the number of sessions added to the initial topology. This number is bounded by the maximum number of sessions that can be present in an iBGP topology, i.e the number of sessions in a full-mesh. However, this bound is far from being reached. Moreover, we have shown that for the research network, the choice of a particular objective for the $GetMostInterestingRouter$ function and the tie-break function does not have a significant impact on the resulting iBGP topologies.

7 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 to the ones considered in this paper, has drawn attention. Here, we present an overview of this work.

[8] and [9] aim to provide route diversity at the frontier of a domain. Inside an AS, several aspects of route resiliency toward distant destinations have been considered. Bonaventure et al. [25] propose a technique for the protection of external peering links by means of tunnels. Their technique requires changes to the various BGP implementations and their deployment, to support a new type of route in BGP. Routes of this new type are called protection routes. They convey information about a backup NH and parameters for tunnel establishment to the NH. Protection routes are advertised on iBGP sessions, inside the AS. Our solution provides this type of protection without requiring any modifications to BGP implementations. Moreover, our approach enables the protection of the ASBRs. Another approach is to obtain higher NH diversity in the routers through an extension to BGP allowing multiple route advertisements for a single prefix [31]. However, Van den Schrieck et al. [32] have shown that such an extension may lead to BGP route oscillations. Lastly, 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 [25] or with our work to achieve the results expected by the author.

The design of iBGP route reflection topologies is considered in [24, 28] and [12]. Buob et al. [24] provide a method to generate iBGP topologies where each router selects the same route it would have selected in the case of a full-mesh of iBGP sessions. Vutukuru et al. [28] rely on a hierarchy of RRs to design iBGP topologies that are robust to IGP failures.

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.

We recommend to the reader interested in the BGP convergence problematic to take look at the work of Flavel et al. [33]. The authors propose a modification to the BGP decision process in order to solve the iBGP routing oscillation issues. They do not tackle the NH diversity problem. Their proposal can be used in combination with our solution.

Finally, Caesar et al. [34] propose an architecture for route distribution inside an

AS. This architecture is an application to BGP of the “4D” concept, proposed in [35]. Inside a domain, a central server redistributes external routes to all the routers in the domain. Such an architecture removes the burden of designing iBGP topologies. However, it is a drastic evolution from the distributed approach that makes the success of the current Internet. It requires that the central entity manages the BGP routing information and controls the routers of the entire AS. Moreover, in its current implementation, the remote control server distributes a single BGP route per destination to each router in a domain.

8 Conclusion

In this paper, we have addressed the problem of NH diversity in an AS, i.e., ensuring that each router learns two routes with different NH towards each prefix. We have shown that the presence of NH diverse routes enables to significantly reduce the switch-over time upon the failure of an ASBR or inter-domain link. Sub-second switch-over time can be achieved.

We propose an algorithm that relies on an initial iBGP route reflection topology. Our algorithm adds a few iBGP sessions to some border routers of the domain, without compromising the correctness of the iBGP topology. Finally, we achieve our NH-diverse goal by leveraging the “best-external” option available on routers.

We evaluated our approach on two different networks, a research and an ISP network, and compared it to conventional iBGP topology designs. In the ISP network, between 0.6% and 1% of the total number of sessions contained in a full-mesh are added to conventional iBGP topology designs. Moreover, the number of routes that have to be stored on average by the routers with our approach increases marginally compared with the conventional iBGP topologies. On average, it is far lower than would be the case under a full-mesh. Our work shows that providing NH-diversity from design lead to a scalable solution, hence should be considered by ISPs today.

We believe that in the long term a new mechanism for the redistribution of the BGP routes in the AS will be developed and adopted by operators. Such a mechanism would ensure NH diversity and correctness of the redistribution without requiring careful design and configuration tasks from the operator.

Acknowledgments

We thank Olivier Bonaventure, Eiji Oki and the anonymous reviewers for their useful comments.

References

[1] R. Zhang, M. Bartell, BGP Design and Implementation, 1st Edition, Cisco Press, 2003.

[2] T. Bates, E. Chen, R. Chandra, BGP route reflection - an alternative to full mesh internal BGP (IBGP), rFC 4456 (April 2006).

[3] C. Labovitz, A. Ahuja, A. Bose, F. Jahanian, Delayed internet routing convergence, in: ACM SIGCOMM 2000, 2000.

[4] F. Wang, Z. M. Mao, J.Wang, L. Gao, R. Bush, A measurement study on the impact of routing events on end-to-end internet path performance, in: ACM SIGCOMM 2006, 2006.

[5] N. Kushman, S. Kandula, D. Katabi, Can you hear me now?!: it must be BGP, SIGCOMM Comput. Commun. Rev. 37 (2).

[6] D. Pei, M. Azuma, D. Massey, L. Zhang, BGP-RCN: Improving BGP convergence through root cause notification, Comput. Netw. ISDN Syst. 48 (2).

[7] J. Chandrashekar, Z. Duan, Z.-L. Zhang, J. Krasky, Limiting path exploration in BGP, in: IEEE INFOCOM, 2005.

[8] N. Kushman, S. Kandula, D. Katabi, B. M. Maggs, R-BGP: Staying connected in a connected world, in: 4th USENIX Symposium on Networked Systems Design & Implementation (NSDI'07), 2007.

[9] W. Xu, J. Rexford, MIRO: multi-path interdomain routing, in: ACM SIGCOMM 2006, 2006.

[10] S. Uhlig, S. Tandel, Quantifying the BGP routes diversity inside a tier-1 network, in: Proceedings of Networking 2006, Coimbra, Portugal, 2006.

[11] C. Filsfils, BGP convergence in much less than a second, presentation at NANOG 40 (June 2007).

[12] L. Xiao, J. Wang, K. Nahrstedt, Reliability-aware iBGP route reflection topology design, in: 11th IEEE International Conference on Network Protocols (ICNP), 2003.

[13] C. Pelsser, T. Takeda, Metrics to evaluate the cost of maintaining diverse BGP routes, in: IEICE General Symposium, 2008.

[14] V. V. den Schrieck, P. Francois, C. Pelsser, O. Bonaventure, Preventing the unnecessary propagation of BGP withdraws, in: Proceedings of IFIP Networking, 2009.

[15] P. Marques, R. Fernando, E. Chen, P. Mohapatra, Advertisement of the best-external route to iBGP, internet draft, draft-ietf-idr-best-external-00.txt (May 2009).

[16] B. Quoitin, S. Uhlig, Modeling the routing of an Autonomous System with C-BGP, IEEE Network 19 (6).

[17] M.-O. Buob, M. Meulle, S. Uhlig, Checking for optimal egress points in iBGP routing of a tier-1 AS, in: The 6th International Workshop on Design and Reliable Communication Networks - DRCN'2007, 2007.

[18] N. Feamster, H. Balakrishnan, Detecting BGP configuration faults with static analysis, in: Networked Systems Design and Implementation (NSDI), 2005.

[19] P. Francois, C. Filsfils, J. Evans, O. Bonaventure, Achieving sub-second IGP convergence in large IP networks, SIGCOMM Comput. Commun. Rev. 35 (3).

[20] D. Katz, D. Ward, Bidirectional Forwarding Detection, internet draft, draft-ietf-bdf-base-09.txt, work in progress (February 2009).

[21] R. Aggarwal, Applications of Bidirectional Forwarding Detection (BDF), presentation at RIPE 48 meeting (May 2004).

[22] D. Pei, J. Van der Merwe, BGP convergence in Virtual Private Networks, in: Internet Measurement Conference (IMC 2006), 2006.

[23] T. Griffin, G. Wilfong, On the correctness of iBGP configuration, in: ACM SIGCOMM 2002, 2002.

[24] M.-O. Buob, S. Uhlig, M. Meulle, Designing optimal iBGP Route-Reflection topologies, in: IFIP Networking 2008, 2008.

[25] O. Bonaventure, C. Filsfils, P. Francois, Achieving sub-50 milliseconds recovery upon bgp peering link failures, IEEE/ACM Transactions on Networking 15 (5).

[26] J. R. L. Subramanian, S. Agarwal, R. H. Katz, Characterizing the internet hierarchy from multiple vantage points, in: IEEE INFOCOM, 2002.

[27] R. Mahajan, N. Spring, D. Wetherall, T. Anderson, Inferring link weights using end-to-end measurements, in: ACM SIGCOMM Internet Measurement Workshop 2002, 2002.

[28] M. Vutukuru, P. Valiant, S. Kopparty, H. Balakrishnan, How to construct a correct and scalable iBGP configuration, in: IEEE INFOCOM, 2006.

[29] D. Pei, B. Zhang, D. Massey, L. Zhang, An analysis of path vector convergence algorithms, Computer Networks 50 (3).

[30] L. Subramanian, V. N. Padmanabhan, R. H. Katz, Geographic properties of Internet routing, in: Proceedings of the 2002 USENIX Annual Technical Conference, 2002.

[31] D. Walton, A. Retana, E. Chen, J. Scudder, Advertisement of multiple paths in BGP, internet draft, draft-ietf-idr-add-paths-02, work in progress (August 2009).

[32] V. Van den Schrieck, O. Bonaventure, Routing oscillations using BGP multiple paths advertisement, internet draft, draft-vandenschrieck-bgp-add-paths-oscillations-00.txt, work in progress (June 2007).

[33] A. Flavel, M. Roughan, Stable and flexible iBGP, in: ACM SIGCOMM, 2009.

[34] M. Caesar, D. Caldwell, N. Feamster, J. Rexford, A. Shaikh, J. van der Merwe, Design and implementation of a routing control platform, in: Networked Systems Design and Implementation (NSDI), 2005.

[35] A. Greenberg, G. Hjalmtysson, D. A. Maltz, A. Myers, J. Rexford, G. Xie, H. Yan, J. Zhan, H. Zhang, A clean slate 4D approach to network control and management, SIGCOMM Comput. Commun. Rev. 35 (5).

Vitae

Cristel Pelsser received her Master degree in computer science from the FUNDP in Belgium in 2001. She then obtained her PhD in applied sciences from the UCL, in Belgium, in 2006. From 2007 to 2009, she held a post-doctorate position at NTT Network Service Systems Laboratories in Japan. She is now a researcher at Internet Initiative Japan (IIJ). Her current research interests are in inter and intra domain routing. She is more specifically interested in interdomain traffic engineering (load balancing, QoS, resilience), BGP routing scalability and IP Fast Rerouting.

Steve Uhlig obtained a PhD in Applied Sciences in March 2004 from the University of Louvain, Louvain-la-neuve, Belgium. He was a Postdoctoral Fellow of the Belgian National Fund for Scientific Research between 2004 and 2006. Between 2006 and 2008, he was assistant professor at Delft University of Technology. He is currently with T-labs/TU Berlin. His main research interests are focused on the behavior of routing and traffic in the Internet, and their interactions on the network topology.

Tomonori Takeda received the M.E. degree in electronics, information and communication engineering from Waseda University, Japan, in 2001. He joined NTT in 2001 and has been engaged in R&D on the IP optical network architecture and related protocols. He has been involved in standardization activities, and co-chaired the Layer 1 Virtual Private Network (L1VPN) working group in the IETF (2005-2009).

Bruno Quoitin is an assistant professor at UMons in Belgium. He obtained his B.S. degree in Mathematics and M.S. degree in computer science from University of Namur in 1996 and 1999 respectively. He then worked in the industry in the field of industrial control networks until 2002. He obtained his PhD degree in applied sciences from UCL in 2006. His main research interests are interdomain routing, network management, interdomain traffic engineering, network modeling and simulation.

Kohei Shiomoto is a Senior Research Engineer, Supervisor, Group Leader at NTT Network Service Systems Laboratories, Tokyo, Japan. He joined the Nippon Telegraph and Telephone Corporation (NTT), Tokyo, Japan in April 1989. Since then he had been engaged in research and development of ATM networks in NTT Laboratories. From August 1996 to September 1997 he was engaged in research in high-speed networking as a Visiting Scholar at Washington University in St. Louis, MO, USA. Since September 1997, he had been engaged in research and development in the areas of IP/GMPLS networking, IP and optical networking at NTT Network Innovation Laboratories and NTT Network Service Systems Laboratories. Since April 2006, he has been leading the IP Optical Networking Research Group in NTT Network Service Systems Laboratories. He received the B.E., M.E., and Ph.D degrees in information and computer sciences from Osaka University, Osaka in 1987 1989, and 1998, respectively. He is a Fellow of IEICE, a member of IEEE, and ACM. He received the Young Engineer Award from the IEICE in 1995. He received the Switching System Research Award from the IEICE in 1995 and 2001.