A novel internal BGP route distribution architecture

Cristel Pelsser Akeo Masuda Kohei Shiomoto
NTT Network Service Systems Laboratories, NTT Corporation

1 Abstract

Route-Reflection and confederations were introduced to alleviate the scalability issue of maintaining a full-mesh of iBGP sessions. However, these techniques may lead to routing, forwarding, route diversity and sub-optimal routing issues. In this paper, we propose a new scalable internal BGP route distribution architecture that is rid of these issues.

We propose an iBGP route distribution architecture relying on Route Servers (RS). Compared to the work of Ceasar et al. [1], there are multiple RSs per AS in our proposal. This ensures scalability and robustness of our new internal BGP route distribution architecture. Each route server is responsible for a subset of the external destinations. For this subset, the RS selects the egress ASBR to be used by each router in the AS.

2 Issues

In [2], Griffin and Wilfong describe a list of iBGP correctness issues. They distinguish routing issues from forwarding issues. There are two causes to the routing issues presented in [2]. The first cause is the sparsity of the iBGP topology when RRs and a confederation of RRs are used. The second cause lies in the advertisement of a single route per prefix on an iBGP session. These two causes result in a partial view of the external routes in the routers of the AS. A node does not know all the BGP routes that are received at the border of the AS when it takes a decision. Thus, routing oscillations may be observed. Or, different routing solutions may be obtained upon different timings of the BGP messages advertised in the AS.

The forwarding issues mentioned in [2] that occur today with RRs[1] are due to the fact that RRs do not consider the location of their clients when they select a BGP route. They select the route that is best for themselves not their clients. Then, they send this route to their iBGP clients. However, this route may not be the one with the shortest IGP cost (among the routes that pass rules 1 to 3 of the BGP decision process). Thus, all the routers on the IGP path to the egress ASBR may not select the same egress ASBR for a given prefix. This leads to deflection and, eventually, forwarding loops in the AS.

Additionally, Uhlig and Tandel [3] have shown that routers lack diverse BGP routes. This leads to connectivity losses of a few tenth of seconds. Finally, due to the route hiding phenomenon in RRs, we observe that (1) the BGP routes used by the routers may be suboptimal with regard to the BGP decision process and (2) path exploration occurs upon a route change.

3 Distributed route servers

We introduce route servers in the AS. Each server is responsible of BGP path selection for a subset of the external prefixes. Each route server is assigned an ID. There is a function that maps each prefix to a key. Route servers' identifiers and prefix's keys belong to the same domain R. Each route server with ID r_i is responsible for prefixes with key k comprised in $r_j < k \leq r_i$, such that $\forall n \in R, n \neq r_i$, if $r_j < n$ then $r_i < n$, where $r_j, r_i \in R$. r_j is the largest ID assigned to a route server that is smaller than r_i.

[1]This problem is also observed in a confederation of ASs.

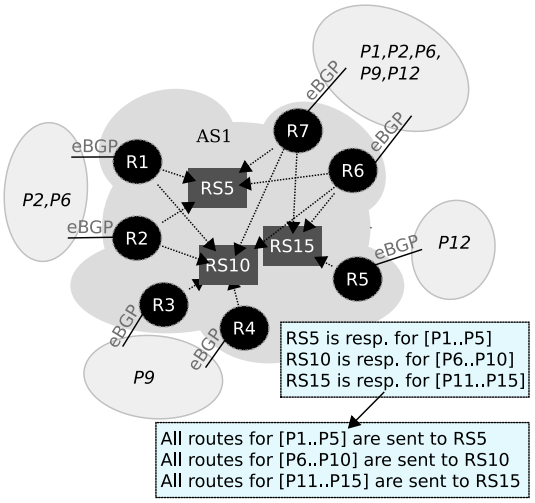

Figure 1: Distributed route servers.

The ASBRs discover the route servers that are present in the AS, as well as their ID, by means of the IGP.

The ASBRs send all the routes they learn on eBGP sessions to the appropriate route servers. In order for the route server to receive all the external routes for a prefix, it has an iBGP session with all the ASBRs. Moreover, "multiple route advertisements" option [4] is activated on these sessions. A route server learns all the routes for the prefixes for which it is responsible. For example, in figure 1, the ID of a route server is the number used in its label. Thus, the ID of $RS5$ is 5. Moreover, the key of a prefix is the index used in its label. For example, $P1$ has key 1. In figure 1, $RS5$ receives all the routes for prefix $P1$ and $P2$. A BGP route for $P1$ is received from ASBRs $R6$ and $R7$. These routers send the advertisements for $P1$ to $RS5$.

Our proposal solves the routing issues presented in [2] because the router server knows all the external routes for a prefix.

After receiving BGP routes, a route server performs the BGP route selection. The current BGP decision process is shown in table 1. In our proposal, the best BGP routes are selected by a route server as follows. First, there are no changes concerning the rules 1 to 3, in table 1. In traditional BGP, these rules lead to the selection of the same set of routes independently of the router that performs the route selection. Thus, the route server will select the same set of routes independently of the router for which the route is destined. However, the 4th and the 5th rules make use of the location of the router in the topology. Therefore, a route server has to take into account the location of the router in the topology, in order to compute a route that is appropriate for the router. In our proposal, the 4th rule of the BGP decision process becomes: "If the NH of the route is directly connected to the router for which the selection is performed, select this route". In the 5th rule, the route server will keep the routes with NHs that are the closest to the considered router. The route server has to know the IGP cost from the router to each possible NH. This is computed from the link costs distributed by the IGP. Finally, there are no changes in the application of the tie-breaking rules.

By taking a decision per router, we are able to avoid

Table 1: Simplified BGP decision process (DP)

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

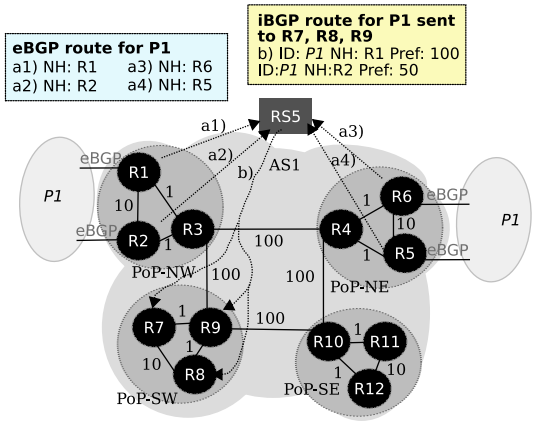

Figure 2: Route selection.

the forwarding issues mentioned in [2].

Figure 2 illustrates the route selection at $RS5$ on behalf of routers $R7$, $R8$ and $R9$. The routes learned for $P1$ at $R1$, $R2$, $R5$ and $R6$ have the same local preference (Loc_pref), the same AS-path length and no MED value assigned. Moreover, these routes are not learned on eBGP sessions at the considered routers ($R7$, $R8$ and $R9$). When $RS5$ makes a selection for the routers inside PoP-SW, it selects R1 and R2 as the best next-hops, based on the IGP cost, the 5th rule in table 1. Then, based on the router ID, $RS5$ selects R1 as the best route for all three routers.

[4] enables a route server to advertise multiple routes to a router. In figure 2, $RS5$ may send the route via $R2$ to the routers inside PoP-SW, in addition to the best route. This route will be used as a backup route, if $R1$ is no more reachable. Thus, with our proposal, we are able to provide route diversity and consequently reduce the duration of connectivity losses upon a failure of an interdomain resource.

$RS5$ also selects the route via $R1$ as best route for $R3$, in figure 2. $R3$ receives the route via $R2$ as backup route. Finally base on the 4th rule of the BGP decision process, $R1$ receives the external route as best route from the route server. The route via $R2$ is its backup route in case $R1$'s interdomain link fails. Similarly, the best route at $R2$ is the external route. The backup route at $R2$ is the route via $R1$. We observe that such routing tables do not lead to forwarding loops.

4 Qualitative evaluation

Our proposal presents a set of advantages compared to the sparse iBGP techniques currently in use. These advantages are summarized in table 2.

Table 2: Qualitative evaluation.

Advantages compared to sparse iBGP topologies	
1	no routing oscillations
2	single routing solution
3	no forwarding loops
4	no deflection
5	route diversity
6	optimal paths according to the DP

In table 3, we compare the scalability of our proposal to the scalability of a full-mesh and sparse iBGP topologies. Therefore, we consider the criterion listed in the first column of the table. We define the following variables: p is the number of external prefixes learned by the AS, n is the number of nodes in the AS, q is the average number of iBGP peers of a node in a sparse iBGP topology. And, s is the number of RS in the AS, with our proposal. It is assumed that $q < s$ and $n > s$.

Table 3: Scalability.

Criterion	iBGP route distribution techniques		
	full-mesh	sparse	proposal
Nb of sessions	large $(n(n-1))$	low $(q*n)$	fair $(s*n)$
Routes/sessions	large (p)	large (p)	low $(2*p/s)$
Table sizes	large $(p(n-1))$	fair $(p*q)$	low $(2*p)$
Nb of messages	large $(p(n-1))$	undef	low $(2*p)$

We can see in table 3 that overall, our proposal is much more scalable than a full-mesh of iBGP sessions and sparse iBGP topologies. Moreover, we note that there is a trade-off between the number of sessions and the number of routes transmitted on a session in our proposal. The number of messages exchanged in a sparse iBGP topology with RRs or a confederation of ASs is not predictable. It highly depends on the iBGP topology. This number may be infinitely high if BGP never converges. Our proposal may require more sessions than a sparse iBGP topology. However, our solution provides route diversity which is highly desirable for the provision of critical services.

It is easy to determine the number of RS required in an AS, s. Given the number of nodes n, the number of external prefixes p a bound on the number of sessions and the number of routes per session, the equations in the last column of table 3, lines 1 and 2, enable to compute s.

5 Conclusion

This paper provides a solution to the routing and forwarding issues highlighted in [2]. With our proposal, routing oscillations due to particular MED or IGP configurations are wiped out. Moreover, the routing solution is always unique. Additionally, with our proposal, no deflection and forwarding loops occur in the converged network.

Our proposal enables each router to know a backup egress ASBR for each prefix. This enables fast recovery upon a failure of the primary route. Lastly, with our proposal less BGP messages are exchanged upon the failure of an external route than with route-reflection topologies and a confederation of ASs.

As next steps, we plan to specify the function that assigns a prefix to a given route server. The focus will be on a balanced distribution of the prefixes on the route servers. We will constrain ourselves to low complexity functions as such a function has to be run each time an ASBR receives a new external route.

References

[1] M. Caesar, D. Caldwell, N. Feamster, J. Rexford, A. Shaikh, and J. van der Merwe. Design and implementation of a routing control platform. In *NSDI 2005*, May 2005.

[2] T. Griffin and G. Wilfong. On the correctness of iBGP configuration. In *ACM SIGCOMM 2002*, 2002.

[3] S. Uhlig and S. Tandel, "Quantifying the BGP routes diversity inside a tier-1 network," In *Networking 2006*, May 2006.

[4] D. Walton, A. Retana, E. Chen, and J. Scudder. Advertisement of multiple paths in BGP, July 2008. Internet draft, draft-walton-bgp-add-paths-06.txt, work in progress.