Automating the iBGP organization in large IP networks

Virginie Van den Schrieck
Université catholique de Louvain (Belgium)
virginie.vandenschrieck@uclouvain.be

1. INTRODUCTION

For years, the Border Gateway Protocol has been used as the interdomain routing protocol inside the Internet [9]. This protocol allows ASes to exchange routes to reachable destinations. A BGP route contains, among other attributes, the list of ASes that form a path to the destination, i.e. an IP prefix. This list is called an AS Path. Thanks to the BGP routes it receives, an AS has all the information needed to forward packets towards their destination by sending them to the best BGP nexthop in the first AS in the AS Path. In practice, there are several BGP routers inside an AS. The routes from a neighbouring AS are received by the routers that have a peering session with this neighboring AS. Such sessions are called eBGP sessions. However, other routers that do not have a peering session with this particular neighbour also need to receive this information. For this purpose, routers inside an AS establish internal BGP sessions, called iBGP sessions.

2. CURRENT IBGP ORGANIZATION

Initially, BGP routers of an AS formed a Full Mesh of iBGP sessions, which is not scalable ($O(n^2)$). Route Reflection has been proposed to address this scalability issue [1]. However, after several years of usage, it appears that this solution has introduced drawbacks, notably in terms of correctness [4] [7] and lack of diversity [10]. We evaluated route diversity inside a Tier-1 ISP of about 100 routers with a hierarchy of Route Reflectors. For this, we generated artificial BGP routes of equal quality, and used the C-BGP simulator [8] to advertise them. We then measured the resulting diversity by counting the percentage of routers having at least two routes with different nexthops to the advertised prefix. Results show that route diversity is very poor compared to the Full Mesh of iBGP sessions. Furthermore, configuring iBGP is not an easy task especially with hierarchies of route reflectors, and can be a source of misconfigurations when done by human operators [6].

After having identified the requirements for iBGP [11], we propose in this paper a new iBGP organization that solves the problem of the manual iBGP session setup while providing on-demand diversity to ensure rapid convergence and facilitate traffic engineering and load balancing.

3. AUTOMATING IBGP SESSIONS ESTABLISHMENT

A group of BGP routers connected to a common AS can be seen as the repository for all the routes announced by this AS. We use this logical grouping as a starting point for a new organization : In order to receive all best routes announced in the AS, a router will automatically establish iBGP sessions with a router in each group of border routers connected to each neighbor AS. We call the group of routers connected to the same neighboring AS the **Contact Group** of this AS. A member of this group is called a **Contact Node**. All routers that do not have eBGP session with a given AS are called **Clients** of the Contact Group of this AS, because they will contact one Contact Node in order to receive the routes from this AS. In the example of figure 1, R1, R2 and R3 belongs to the Contact Group of AS2. R1 is the Contact Node of R4 for AS2.

IBGP sessions between Clients and Contact Nodes are only partial : A Contact Node will only send a route to a Client if it actually plays the role of Contact Node for the AS that advertised that route. We will call all those partial iBGP sessions **liBGP sessions**, for Light iBGP sessions. The liBGP sessions are unidirectional, because routes are only sent from the Contact Node to the Client. In figure 1, R1 only sends the routes from AS3 to its Client R4.

All routers belonging to a Contact Group must agree on the routes that have to be propagated inside the AS : if, for example, there exists a route to a given prefix with a lower MED than the other, all Clients of the Contact Group must learn this route so that they can select it as best. Therefore, there must exist a Full Mesh of liBGP sessions between the members of a Contact Group to allow them to

*CoNEXT'*07, December 10-13, 2007, New York, NY, U.S.A.
Copyright 2007 ACM 978-1-59593-770-4/ 07/ 0012 ...$5.00.

share their best routes. In this particular case, the liBGP sessions are bidirectionnal, as Contact Nodes exchange their Contact Group routes with each other.

LiBGP sessions can also be configured using the Add-Paths feature [12] which allows routers to advertise more than one route to each prefix instead of advertising only the best route. This provides on-demand diversity, as Clients can ask their Contact Node for backup routes.

As a router needs to learn all routes advertised by all neighboring AS, it needs to know which routers are connected to which ASes in order to participate to the Contact Groups of the ASes to which it has an eBGP session and be Client of the others. For this, we propose to use a dedicated router called **Contact Information Server**, or **CIS**, that gathers the information about which eBGP sessions are attached to which border router. This CIS can send or receive this information as BGP routes using the new address family presented in [2]. In practice, a simple Route Reflector can act as a CIS, as it only needs to reflect a number of routes equal to the number of eBGP peerings. There can be more than one CIS in the AS to ensure robustness, and this role can even be distributed among several nodes. This mechanism allows the complete automatization of all liBGP session establishment.

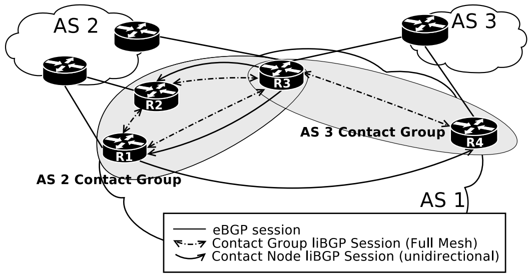
AS 2
AS 3
R3
R2
R1
R4
AS 3 Contact Group
AS 2 Contact Group
AS 1
eBGP session
Contact Group liBGP Session (Full Mesh)
Contact Node liBGP Session (unidirectional)

Figure 1: liBGP architecture

4. EVALUATION OF THE PROPOSED ORGANIZATION

Route diversity : We mentionned in the introduction that ISP networks using route reflection had a bad diversity repartition. It is then logical to raise the same issue with our proposed iBGP organization. We applied the same methodology as explained in the introduction on the same Tier-1 ISP, but with the iBGP configured with the proposed liBGP organization. When using one Contact Node per AS without Add-Paths, diversity is very bad (< 10%) when routes are advertise by single ASes. This is logical, as Clients receive only one route from their Contact Nodes. As explained earlier, using Add-Paths naturally solves this problem, as it allows Clients to receive two different routes for each destination and then have all the diversity they need.

Load on the routers : The number of sessions with this new iBGP organization is high, but this does not significantly rise memory usage on the routers. Indeed, the liBGP sessions are only partial sessions, and only a small subset of the routes are exchanged on those sessions. The amount of computation performed by iBGP routers can be reduced by using Peer Groups [3][5] on liBGP sessions sharing common characteristics. Furthermore, the size of the Adj-Rib-Ins with this organization is much lower than with a Full-Mesh or with Route Reflection when Add-Paths is not used. With Add-Paths, memory usage is comparable to the one obtained with redundant Route Reflection. But with redundant Route Reflection, routers often have several versions of the same route in their Adj-Rib-Ins, which is not useful in case of failure of the primary path. With our proposal, when several routes to the same prefix exist in the Adj-Rib-In, they have different nexthops and the diversity is exploitable.

5. CONCLUSION

We propose a new method to organize iBGP in large ISP networks. For this, we rely on lightweight iBGP sessions. On a liBGP session, a router will only advertise the routes learned from a given neighbor. Compared to existing solutions such as Full Mesh or Route Reflection, our solution offers several advantages. First, it can be completely automated, i.e. iBGP sessions do not need anymore to be manually configured on the routers. This is key to reduce the misconfiguration errors. Second, it reduces memory usage on routers, compared to the Full-Mesh. Finally, our organization also allows routers to learn two distinct paths by using Add-Paths. This provides on-demand diversity and is key to allow routers to quickly reroute after the failure of the primary path.

Further work on this topic includes the integration of this proposal in the XORP software in order to deploy it in a lab and study its dynamic behavior. We are also planning to study how BGP-MPLS VPN deployment can be facilitated with this iBGP organization.

6. REFERENCES

[1] T. Bates, R. Chandra, and E. Chen. BGP route reflection - an alternative to full mesh iBGP. Internet RFC 2796, April 2000.

[2] O. Bonaventure, C. Filsfils, and P. Francois. Achieving sub-50 milliseconds recovery upon BGP peering link failures. In *Co-Next 2005*, Toulouse, France, October 2005.

[3] Cisco Systems. IOS 12.4. `http://www.cisco.com`, May 2007.

[4] T. Griffin and G. Wilfong. Analysis of the MED oscillation problem in BGP. In *ICNP2002*, 2002.

[5] Juniper. Junos 7.6. `http://www.juniper.net/techpubs/software/junos/junos76/index.html`, May 2006.

[6] R. Mahajan, D. Wetherall, and T. Anderson. Understanding BGP misconfigurations. In *ACM SIGCOMM 2002*, August 2002.

[7] D. McPherson, V. Gill, D. Walton, and A. Retana. Border Gateway Protocol (BGP) Persistent Route Oscillation Condition. RFC 3345 (Informational), Aug. 2002.

[8] B. Quoitin and S. Uhlig. Modeling the routing of an Autonomous System with C-BGP. *IEEE Network*, 19(6), November 2005.

[9] Y. Rekhter, T. Li, and S. Hares. A Border Gateway Protocol 4 (BGP-4). RFC 4271 (Draft Standard), Jan. 2006.

[10] S. Uhlig and S. Tandel. Quantifying the impact of route-reflection on bgp routes diversity inside a tier-1 network. In *IFIP Networking 2006*, Coimbra, Portugal, May 2006.

[11] V. Van den Schrieck, P. Francois, S. Tandel, and O. Bonaventure. Let BGP speakers configure their ibgp sessions on their own. Position Paper, Wired2006 Workshop, Atlanta, October 2006.

[12] D. Walton, D. Cook, A. Retana, and J. Scudder. Advertisement of Multiple Paths in BGP. Internet draft, November 2002.