Rethinking iBGP Routing

Iuniana M. Oprescu [‡] [†], Mickael J. F. Meulle [‡], Steve Uhlig [*],
Cristel Pelsser [◇], Olaf Maennel [⋆], Philippe Owezarski [†]

[‡] Orange Labs; {mihaela.oprescu, michael.meulle}@orange-ftgroup.com
[*] Deutsche Telekom Laboratories/TU Berlin; steve@net.t-labs.tu-berlin.de
[◇] Internet Initiative Japan; cristel@iij.ad.jp
[⋆] University of Loughborough; olaf@maennel.net
[†] CNRS; LAAS; owe@laas.fr

ABSTRACT

The Internet is organized as a collection of administrative domains, known as Autonomous Systems (ASes). These ASes interact through the Border Gateway Protocol (BGP) that allows them to share reachability information. Adjacent routers in distinct ASes use external BGP (eBGP), whereas in a given AS routes are propagated over internal BGP (iBGP) sessions between any pair of routers. In large ASes where a logical full-mesh is not possible, confederations or route reflectors (RRs) are used. However, these somewhat scalable alternatives have introduced their own set of unpredictable effects (persistent routing oscillations and forwarding loops causing an increase of the convergence time) addressed in the literature [1].

The solution we propose to these issues consists of a structured routing overlay holding a comprehensive view of the routes. We describe the design of a distributed entity that performs BGP route pre-computation for its clients inside a large backbone network and propagates the paths to the routers. Compared to the current iBGP routing, the advantage of the overlay approach is the separation between the responsibility of the control plane (route storage and best path computation) and the forwarding of the packets.

One of the major improvements we bring is the divided routing table tackling the scalability concerns and allowing for parallel computation of paths.

Categories and Subject Descriptors:
C.2.3 [Network Operations]: Network Management

General Terms: Design, Management

1. TODAY'S ROUTING PARADIGM

In this work we focus on improving the scalability of BGP intra-domain routing, keeping eBGP untouched. The current iBGP architecture suffers from multiple drawbacks:

- Scalability - quantified by the number of sessions a router is required to handle and the number of prefixes in the BGP table. During the last decade, the Internet BGP routing table has expanded, growing from $100,000$ prefixes in 2001 to approximatively $320,000$ in 2010.
- Incomplete visibility of routing information - a border router of an AS has several possible paths to a given IP prefix. After the decision process, it selects and advertises only its single best route among all the possible ones, reducing the knowledge of the routers inside the network [2]. This incomplete view of eBGP-learned routes may lead to suboptimal egress point selection, non-deterministic protocol behavior, path exploration, delayed convergence [3].
- Complex routing policy management - the routing policy of an AS is subject to misconfigurations and inconsistencies given the heterogenous nature of routing equipment in an AS. It is difficult to build and maintain automatic tools to consistently manage the entire collection of routes in BGP networks.

SIGCOMM'10, August 30–September 3, 2010, New Delhi, India.
ACM 978-1-4503-0201-2/10/08.

2. SEPARATION OF IBGP PLANE

In today's IP networks, routing is done in a distributed manner, on every router in the network. We propose to detach the two processes of building the routing table (routing plane) and the forwarding of traffic (data plane) on two distant hardwares. This offloading of the control plane from the actual routers to dedicated nodes may be regarded as a natural evolution in routing, driven by the increasing table size.

We reconsider the current design with a separate iBGP routing plane concentrating the full knowledge of routing data. A detached plane would help to solve issues about visibility and diversity of routing information and also permit a more consistent management of the routing policy. Such a separation relieves routers from the overhead of the BGP routing process as it allows for distant storage of the routing information and further customized processing according to rules specific to the client routers. Our proposal is based on a logical overlay of routing processes (or nodes) that are jointly responsible for the following:

- collect, split and store the complete set of eBGP-received routes and the internally originated routes,
- store the routing policies and the configurations of all the routers within the AS,
- compute BGP routes for each router,
- redistribute the computed paths to the routers.

One of the most challenging issues for an iBGP architecture is its ability to scale, or in other terms to support increasing routing table size and messages over time. To design a scalable solution, the iBGP overlay is split into n sub-planes, where nodes of the same sub-plane handle the same routing information for only a fraction of the entire set of prefixes.

Upon reception of a reachable prefix from eBGP (step 1), a router forwards this information to the overlay (step 2) and the corresponding sub-plane is determined through a simple IP lookup in a table containing the previously mentioned set of prefixes (step 3).

The subsets of prefixes assigned to sub-planes are decided dynamically. This allows for a flexible load-sharing of routing data

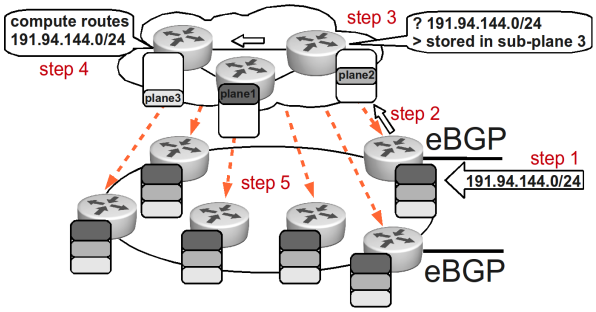

Figure 1: The $n = 3$ nodes in the overlay each manage a subset of the prefixes found in the total routing table

over the sub-planes, that takes into account the capacities of nodes. To ensure that each client router receives routing data from the overlay (step 5) for the full set of prefixes, each client router connects to at least one node of each sub-plane.

2.1 Design guidelines

Various assumptions have been made when building this new architecture. Some important construction aspects are derived from present needs appearing in Internet routing:

Dynamic load sharing of routing data among planes allows for splitting, organizing and re-organizing of the various prefixes handled by the overlay nodes. A change in the network topology graph or an unbalanced distribution of the prefixes over the sub-planes may trigger a redistribution of prefixes. The overlay automatically adjusts the load on the nodes over time.

Distributed computing of the BGP decision process can improve convergence time. We believe that parallel processing of the large amount of control plane data would alleviate today's limited routing engines that are handling great loads due to the BGP activity.

Correlation between best path selection and route propagation is one of the causes leading to opacity in a network. This is partially due to the fact that each router advertises its best route and the propagation inside the AS depends on the sparse router-level graph. Bringing complete visibility to the decision entity enables better route selection, while ensuring a reasonable number of exchanged routing messages. The overlay nodes have knowledge of the entire IGP topology and the routing policy, thus they are able to provide diverse paths towards the same prefix, depending on the position of the different clients within the IGP topology.

3. CONTRIBUTION

Unlike in [4], our solution achieves a division of the routing table in the overlay, therefore decreasing the charge of protocol data on the routers and allowing for parallel computation of routes located on disjoint nodes. If redundancy is sufficient within the overlay, failures inside the network will have minimal impact on reachability inside the AS.

Other studies on centralized routing schemes [5] have shown encouraging results about faster network convergence. Having a comparable convergence time, a well designed routing architecture could be an alternative to link-state protocols.

4. PERSPECTIVES AND CONCLUSION

The splitting of the routing table sounds appealing, but it comes at a cost. We want to determine a threshold beyond which the overhead of computing paths inside the overlay will prevail on the advantages of such a distributed design. Future work includes an implementation of the routing table splitting algorithm and its optimization for handling gracefully the dynamic re-organization of the prefixes in the nodes of the overlay. Further, we will concentrate on an evaluation of the proposed architecture and examine which are the relevant parameters to be studied, e.g. how to quantify convergence time, scalability and compliance to the routing policy.

We have shown a different approach to intra-domain BGP routing, bringing architectural advantages. The proposed design offers better scalability, improved diversity and comparable network convergence time. We estimate that it provides ground for implementations of extra features, such as multiple BGP paths.

5. REFERENCES

[1] A. Rawat and M. A. Shayman, "Preventing persistent oscillations and loops in ibgp configuration with route reflection," *Comput. Netw.*, vol. 50, no. 18, pp. 3642–3665, 2006.

[2] S. Uhlig and S. Tandel, "Quantifying the bgp routes diversity inside a tier-1 network," in *Networking*, pp. 1002–1013, 2006.

[3] T. G. Griffin and G. Wilfong, "On the correctness of iBGP configuration," *SIGCOMM Computer Comm. Review*, vol. 32, no. 4, pp. 17–29, 2002.

[4] M. Caesar, D. Caldwell, N. Feamster, J. Rexford, A. Shaikh, and J. van der Merwe, "Design and implementation of a routing control platform," in *Proc. of the 2nd NSDI*, (Berkeley, CA, USA), 2005.

[5] J. Fu, P. Sjödin, and G. Karlsson, "Intra-domain routing convergence with centralized control," *Comput. Netw.*, vol. 53, no. 18, pp. 2985–2996, 2009.