Let BGP speakers configure their iBGP sessions on their own

Virginie Van den Schrieck, Pierre Francois, Sébastien Tandel, Olivier Bonaventure
Dept CSE
Université catholique de Louvain
Belgium
{vvandens, francois, standel, bonaventure}@info.ucl.ac.be

I. Introduction

Current iBGP is a source of many problems for ISPs today. First, operating iBGP is costly as it requires human configuration and supervision of all iBGP sessions. Such a manual configuration is error prone and hard to troubleshoot. Secondly, the introduction of Route Reflectors to solve scalibility issues has led to a reduced path diversity within the routers of the network. Also, it has been shown that losses of connectivity are frequently caused by routing failures [1], and iBGP is often responsible of them.

This low performance system does not fit for the current Service Level that is required by ISP customers. iBGP needs to be improved or even redefined in such a way that it will allow more robust and stable operation. Prior to the design of a new solution, we propose a list of requirements that any new proposal for iBGP should try to fulfill.

II. Requirements for iBGP

The requirements presented in this section are partially fulfilled by the state of the art, but we hope for new solution to go further in this fulfillment, even if a complete respect of those requirements is probably unfeasable.

They are mostly motivated by informal discussions with ISPs as weel as by panels presented by providers and router vendors.

A. Automatic configuration

The first requirement is motivated by cost reduction plans of ISPs [2] as well as by the fact that iBGP topologies are known to lead to configuration errors due to human operations [3].

It follows from those reflexions that the iBGP system should configure itself on its own, with as few human effort as possible. It should adapt to internal topology changes and automatically reconfigure itself, without human intervention. Such internal topology changes are typically UP and DOWN events in the case of BGP speakers.

Ideally, the only thing that an operator should configure is its inter-domain policies. Note that the solution should allow an operator to define its policies with something less error prone than input and output filters languages as they are featured by current routers operating systems. Also, the langage used by the operator for this purpose should be independent of the hardware and software that is deployed within its network.

The IETF inter-domain routing working group is currently working on ways to let routers automatically detect BGP speakers within an AS [4] and establish iBGP sessions with them. However, the current solution only allows to establish iBGP Full Meshes. It should be possible to automatically establish more scalable iBGP topologies.

B. Scalability

The iBGP system should be scalable, i.e it should be able to support a large number of speakers, with no memory overloading. Also, it should not lead to long periods of high CPU usage during convergence.

C. Path Diversity

Path diversity is the availability of multiple routes to one given prefix in the Adj-Rib-Ins of the routers when the network is stable. This diversity is desirable for fast recovery in case of route withdrawals, as it provides alternative paths that can be used as replacement of the withdrawn route.

Another application of path diversity is load balancing : load balancing traffic among redundant peering links has been proposed at the inter-domain routing working group of the IETF, to satisfy requirements of operators to enable finer Traffic Engineering over their peering links, as requested in [2]. A showstopper for the deployement of such solutions is the low path diversity that has been found in ISPs networks.That is, routers often have only one path in their Adj-Rib-In for a given prefix, which would prevent them from doing load balancing over multiple egress points, even if the routing protocol allows it.

Obviously, path diversity is desirable. Any iBGP system should then provide as much path diversity as possible.

D. Stability

The stability objective is two-fold. First, in the context of iBGP topologies configured by the routers themselves, according to various criteria, the set of iBGP sessions established between BGP speakers must not be continuously changed by the BGP system. Secondly, routing within the system must be stable, so that the system does not leave room for route oscillation.

E. Correct support of all intra-domain forwarding modes

Any iBGP system should correctly support all possible intra-domain forwarding nodes.

One first forwarding mode is Pervasive BGP, with is known to induce deflection and forwarding loops with some iBGP topologies [5]. Any iBGP solution should then configure itself by ensuring that forwarding deflection does not occur.

Other forwarding modes, using tunnels, should also be supported. Those forwarding modes are :

- Using an MPLS tunnel from the Ingress Node to the Egress Node, with forwarding performed by the Egress based on a lookup in its BGP table, for the destination of the encapsulated packet
- Using an MPLS tunnel from the Ingress Node to the Egress Node, with forwarding performed by the Egress based on the MPLS Label of the received packet
- Using an MPLS tunnel from the Infress Node to the Egress Node, with two levels of encapsulation. The outer label is used to forward the packet to the Egress Node. The inner label is used by the Egress Node to select the outgoing peering link.
- A mix of these modes.

F. Robustness

The robustness requirement is the capacity of any iBGP system to support removal or failure of iBGP nodes.

More specifically, the reachability of external destinations must not be compromised when k iBGP sessions or BGP speakers are removed from the iBGP topology. k should be configurable by the operator.

Also, if centralized systems like Route Servers are used, the reachability of external destinations must not be compromised when some, or even all of them fail at the same time, and this even if service redundancy is supposed to be provided.

G. Support for maintenance operations

The convergence phase following a change in the topology due to a maintenance operation should not lead to packet loss. This requirement should cover operations that affect the forwarding plane of routers. In other words, Graceful Restart mechanisms are not sufficient as they only cover reboots of the control plane of routers. It should be possible to establish and shut down iBGP and eBGP sessions without loosing packets, when an alternate path exists in the network.

H. Tunability

Even in the case of an iBGP system able to configure itself, operators can still be willing to introduce small changes in the configuration. iBGP solutions should provide operators the ability to manually set up part of their configuration.

It should then be possible for the operator to specify mandatory iBGP sessions, i.e. sessions that will be established if both ends of the specified session are up. It should also be possible for the operator to specify forbidden sessions, i.e. sessions that will never be established between two BGP speakers.

For example, if Route Reflectors are part of the solution, it should be possible for the operator to specify which routers can or cannot act as a Route Reflector. The operator should also be able to specify a maximum depth for the Route Reflector Hierarchy. memory and CPU establishment of the iBGP

I. Feedback

In order to allow supervision of what happens in the network, the operator should have easy access to the state of the BGP system at any time. Furthermore, any routing error or anomaly should be automatically reported to the operator in order to allow quick reaction or intervention if needed.

In other words, Constant feedback should be provided to the operator by the iBGP system. captured by the operator, and routing errors should be automatically reported.

J. Performance

Recovery and convergence time should be optimized by the system, as well as bandwith utilization on peering links.

K. Support for legacy systems

Legacy BGP speakers should be integrable within the system without harming too much the fulfillement of the other requirements. As few changes as possible should be made to the BGP protocol itself.

REFERENCES

[1] F. Wang, Z. M. Mao, J. Wang, L. Gao, and R. Bush, "A Measurement Study on the Impact of Routing Events on End-to-End Internet Path Performance," in *Proc. of ACM SIGCOMM*, September 2006.

[2] V. Jill, "Panel on BGP," April 2006, presented at Infocom 2006, http://www.ieee-infocom.org/2006/panelist/infocom-panel2-vijay.pdf.

[3] R. Mahajan, D. Wetherall, and T. Anderson, "Understanding BGP misconfigurations," in *ACM SIGCOMM 2002*, August 2002.

[4] R. Raszuk, "IBGP Auto Mesh," June 2003, internet draft, draft-raszuk-idr-ibgp-auto-mesh-00.txt, work in progress.

[5] T. G. Griffin and G. Wilfong, "On the correctness of ibgp configuration," in *SIGCOMM '02: Proceedings of the 2002 conference on Applications, technologies, architectures, and protocols for computer communications*. New York, NY, USA: ACM Press, 2002, pp. 17–29.