Beyond interdomain reachability

Position paper for the Workshop on Internet Routing Evolution and Design (WIRED), October 2003

Olivier Bonaventure*, Bruno Quotin+, Steve Uhlig*

I. Introduction

The Border Gateway Protocol (BGP) was designed as a successor to the Exterior Gateway Protocol. BGP started as a subset of the IDRP protocol [ISO93] being developed by ISO. During the last ten years, BGP has evolved in an incremental and backward compatible manner. In the early nineties, the main objective of BGP was to make possible the distribution of routes constrained by routeing policies, such as those of the NSFNet. As such, BGP mainly provided reachability information and this was the main concern for most Autonomous Systems since the Internet was being built.

Since then, the Internet has grown tremendously, both in size and economical importance. Today, the Internet is composed of 15804 different domains. Among those domains, only 2109 provide a transit service. Thus, most of the domains on the global Internet are stub domains. Stub domains include enterprise networks or universities, content providers and access providers.

II. Limitations of the BGP decision process

Initially, BGP was only concerned by reachability and the BGP specification [LR89] did not define precisely the decision process to be applied to the received routes. The next versions of the BGP protocol, until BGP-4 [RL95] introduced several important modifications. Besides the support of CIDR prefixes, a first implicit addition was the ability to perform AS-Path prepending. A second addition is the clarification of the BGP decision process. Initially, this decision process did not define precisely the tie-breaking rules (AS-Path length, IGP cost, ...) but the current draft [RLH03] defines this process clearly.

The BGP decision process is now composed of the six main rules shown in figure II. The first rule is often used to enforce economical relationships between domains.

The second rule is considered by some as a way to compare the quality of routes, assuming that the length of the `AS-Path` is a valid indication of the quality of a route [McM99]. The validity of this assumption is not really confirmed by the available measurement studies [HFP+02]. However, a consequence of this second rule is that most Internet routes tend to be rather short (usually 3 to 4 AS hops). An indirect consequence is probably that domains try to be connected directly to tier-1 domains to have shorter routes that are assumed to be of better quality.

1) `Prefer routes with highest local preference`
2) `Prefer routes with the shortest AS-Path;`
3) `Prefer routes with the lowest MED attribute ;`
4) `Prefer eBGP routes to iBGP routes;`
5) `Prefer routes via the nearest IGP neighbour;`
6) `Tie breaking rules`

Fig. 1. The (simplified) BGP decision process

The third step of the BGP decision process is used to implement *cold potato* routing between neighbouring domains. It has several problems since its utilisation may cause routing oscillations.

The fourth and fifth steps of the BGP decision process are used to implement *hot potato* routing. They are typically useful for transit domains.

The last step of the BGP decision process is used to break ties when several routes are equivalent on the basis of the other steps. Some BGP implementations break ties by preferring the routers received from the router with the lowest `routerid` [RLH03] while other prefer the oldest route [EC03]. This step introduces some randomness in interdomain routing.

The current BGP decision process is well adapted to large transit ISPs since it reflects the cost of using the received routes and allows them to prefer the cheapest ones.

However, when considering the 13695 stub domains in today's Internet, the BGP decision process is less useful since only the first, second and sixth steps are be used in practice. Often, the selection of the route is performed by the sixth step and thus a large fraction of the interdomain routes are selected randomly in stub domains.

The BGP decision process should be improved to allow stub domains to better select their interdomain routes. A first solution is to continuously measure the quality of the interdomain routes that carry lots of traffic and dynamically configure BGP to always select the best one [Bar02]. Apparently, some content providers already rely on such solutions today. However, the overall impact of those techniques on the interdomain routing system is as of today unknown.

Instead of placing a measurement probe inside each domain, it would be useful to develop a scalable method to perform and

* Olivier Bonaventure and Steve Uhlig are with Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium. Emails: {Bonaventure,Uhlig}@info.ucl.ac.be, URL: http://www.info.ucl.ac.be/people/OBO/

+ Bruno Quoitin is with the University of Namur (FUNDP), Namur, Belgium. Email: Bruno.Quoitin@info.fundp.ac.be

distribute those measurements. [FJJ+01] could be solution as well as QoS extensions to BGP[CJ01], [XLWN02], although those extensions might be difficult to deploy in practice. We believe that providing more accurate "quality" information about the interdomain routes will be useful in the near future.

III. Interdomain traffic engineering

In the first version of BGP [LR89], interdomain traffic engineering was not required. The original specification considered as invalid the routes received with the same AS number several times in the AS path. However, this limitation was unnecessary in practice and it was removed in a later version of BGP. Since then, several interdomain traffic engineering techniques have been used by ISPs. Some allow an ISP to control the flow of its outgoing traffic (setting up local-pref, configuration of import filters). When used by stub ISPs, those techniques do not affect the global interdomain routing since the changes are limited to the stub ISP. However, a dynamic technique used by a large number of stubs could affect the distribution of the interdomain traffic. If a large transit ISP was to start to use such techniques [1] dynamically, it could have a strong impact on the global interdomain routing system.

Others techniques allow ISPs to control the flow of their outgoing traffic (AS-Path prepending, utilisation of the MED attribute, advertisement of more specific prefixes, the various community-based hacks [BQ03]). All these techniques suffer from severe limitations, but it seems that they are required, notably by access providers. When used by stub or transit ISPs, they have an impact on the global interdomain routing system.

If interdomain traffic engineering remains a requirement for the next few years, the Internet will need a set of engineering techniques that go beyond the current "hacks". Those techniques will probably require some form of negotiation between the ISPs that send, transit and receive traffic. As usual when considering interdomain routing, scalability will be a key concern. To better support the interdomain traffic engineering needs of transit and stub domains, we should design techniques that allow to control, possibly dynamically, the flow of interdomain traffic (incoming and outgoing) without any changes to the BGP advertisements. Decoupling BGP routing from interdomain traffic engineering should probably be a medium term goal.

It remains to be seen whether the best solution would be changing the interdomain routing architecture [ACK03], building overlay networks [ABKM01] or something else..

References

[ABKM01] D. Andersen, H. Balakrishnan, M. Kaashoek, and R. Morris. Resilient overlay networks. In *SOSP 2001*, 2001.

[ACK03] S. AGARWAL, C. CHUAH, and R. KATZ. Opca: Robust interdomain policy routing and traffi c control. In *IEEE Openarch (New York, NY, April 2003)*., 2003.

[Bar02] J. Bartlett. Optimizing multi-homed connections. *Business Communications Review*, 32(1):22–27, January 2002.

[BQ03] O. Bonaventure and B. Quoitin. Common utilizations of the BGP community attribute. Internet draft, draft-bonaventure-bgp-communities-00.txt, work in progress, June 2003.

[CJ01] G. Cristallo and C. Jacquenet. Providing quality of service indication by the BGP-4 protocol : the QoS_NLRI attribute. Internet draft, draft-jacquenet-qos-nlri-03.txt, work in progress, July 2001.

[EC03] S. Sangli E. Chen. Avoid bgp best path transition from one external to another. Internet draft, draft-chen-bgp-avoid-transition-00.txt, work in progress, 2003.

[FJJ+01] P. Francis, S. Jamin, C. Jin, Y. Jin, D. Raz, Y. Shavitt, , and L. Zhan. Idmaps: A global internet host distance estimation service. *IEEE/ACM Transactions on Networking*, 2001.

[HFP+02] Bradley Huffaker, Marina Fomenkov, Daniel J. Plummer, David Moore, and k claffy. Distance metrics in the internet. In *IEEE International Telecommunications Symposium*, 2002.

[ISO93] ISO/IEC. Information processing systems - telecommunications and information exchange between systems - protocol for exchange of inter-domain routeing information among intermediate systems to support forwarding of iso 8473 pdus. ISO/IEC 10747:1993, 1993.

[LR89] K. Lougheed and Y. Rekhter. Border gateway protocol (BGP). Request for Comments 1105, Internet Engineering Task Force, June 1989.

[McM99] P. McManus. A passive system for server selection within mirrored resource environments using as path length heuristics. Available from `http://www.gweep.net/~mcmanus/proximate.pdf`, April 1999.

[RL95] Y. Rekhter and T. Li. A border gateway protocol 4 (BGP-4). Request for Comments 1771, Internet Engineering Task Force, March 1995.

[RLH03] Y. Rekhter, T. Li, and S. Hares. A Border Gateway Protocol 4 (BGP-4). Internet draft, draft-ietf-idr-bgp4-20.txt, work in progress, April 2003.

[XLWN02] L. Xiao, K. Lui, J. Wang, and K. Nahrstedt. QoS extensions to BGP. In *ICNP 2002*, Paris, France, November 2002.

[1] Note that for a large transit ISP, the utilisation of intradomain traffi c engineering techniques could affect the fbw of interdomain traffi c and the routing advertisements sent by this ISP.