Reconsidering the Internet Routing Architecture

Olivier Bonaventure
Université catholique de Louvain, Belgium

March 12, 2007

1 Introduction

The current growth of the BGP routing tables [15] and Forwarding Information bases (FIB) on core routers worries several operators and vendors [21]. Another issue is the growth of the number of BGP messages that need to be processed by BGP routers [16]. This is not the first time that the Internet is confronted with such problems. The most successful solution has been the introduction of Classless Interdomain Routing (CIDR) [13]. Other solutions such as using the DNS to aid routing [14] were proposed, but not implemented. CIDR allowed to significantly reduce the growth rate of the BGP routing tables until a few years ago. Today, CIDR is not sufficient anymore due to the growth of multihomed stub ASes [1, 15].

This work in progress document is divided in two parts. In section 2, we first discuss several of the implicit assumptions of the current interdomain routing architecture. Then, in section 3, we discuss several alternatives to these assumptions and highlight some design choices.

2 Assumptions

In this section, we discuss several of the implicit assumptions of the current interdomain routing architecture and their implications.

2.1 IANA-based address allocation

For many years, the IP addresses have been manually assigned. IANA or the regional registries maintain a pool of IP addresses and allocate the IP addresses in two flavors :

- Provider Independent (PI) address
- Provider Aggregatable (PA) address

The IANA-based manual address allocation has been extended to the providers receiving PA addresses. Those providers also maintain a registry and manually, or partially manually, allocate sub prefixes in their PA space. When a network receives

a prefix from IANA, a regional registry or its provider, its network operators manually assign addresses from the received prefix to their routers and servers. The only place where the Internet managed to achieve automatic allocation of addresses is the endsystem with the success of DHCP and auto-configuration. Due to the large number of places where IP addresses are manually configured or specified, renumbering the IP addresses of a network is a difficult problem [7, 5]. The available renumbering techniques [3, 4] are complex and cannot be fully automated.

2.2 An IP address is both a locator and an identifier

Since the beginning of the Internet, an IP address has been associated to a single layer-2 interface. Thus, the IP address serves as a locator. Unfortunately, an IP address also serves as an identifier. It is only recently that standardised solutions have been proposed to allow the transport layer to work more independently from the network layer [30]. There is also ongoing work to support multiple locators below the transport layer [22, 24].

2.3 ASes are visible entities in the interdomain routing system

The current interdomain routing system is used to distribute prefixes, but it assumes that each prefix belongs to one Autonomous System. An AS is defined as "a set of routers under a single technical administration ... the administration of an AS appears to other ASes to have a single coherent interior routing plan, and presents a consistent picture of the destinations that are reachable through it" in [28].

Mechanisms have been added to the interdomain routing system to allow an AS to aggregate several prefixes received from its peers before announcing a larger IP prefix [8], but those mechanisms are not aggressively used by ISPs. The large number of multihomed stubs [15] is one of the reasons for the limited usage of aggregation.

The utilisation of the AS-Path as a loop avoidance mechanism in the BGP protocol has increased the reliance on the ASes in the current interdomain routing system. AS-Paths have been used for other purposes such as inferring the interdomain topology [31].

2.4 Interdomain routing convergence

When a link fails, the interdomain routing protocol needs to converge to let the entire Internet know an alternate path to reach the destinations affected by the failure. The convergence of the interdomain routing system has been relatively slow in the past [17]. Recent measurements show that interdomain routing convergence remains relatively slow [34].

2.5 Traffic engineering

The interdomain routing system was designed to support a best-effort service by allowing each participating router to advertise one path towards each destination prefix modulo its routing policies. However, due to congestion, routing policies, traffic storms

and other issues, operators have tweaked the interdomain routing protocol and used it to redirect traffic flows to achieve load balancing and other traffic engineering objectives.

2.6 Security is not a strong concern

When the interdomain routing architecture was designed, few researchers considered security issues or misconfiguration problems. The evolution of the Internet has shown that misconfiguration are common events [18] and attacks have become a stronger concern recently [23, 25].

3 Alternative Interdomain Routing Architecture

In this section, we evaluate several alternatives to the current interdomain routing architecture and discuss their potential benefits. We try to favour simplicity whenever possible with the aim of reducing the size of the routers' FIB and the number of BGP update messages while still allowing the support of added value services such as fast convergence or traffic engineering.

3.1 Separating locators and identifiers

The separation between locator and identifier functions of IP addresses has been proposed to solve several problems of the current Internet architecture.

There is debate currently on whether the locators should identify routers/middleboxes (e.g. LISP, proxy-shim6, ...) or endsystems (e.g. shim6, HIP). We believe that it is too early to choose. Both have advantages and drawbacks when considering both long term and short term issues. **A new interdomain routing architecture should not dictate the exact placement of locators.** It should allow the architecture to evolve from host-attached to middlebox attached locators and the opposite.

In some cases, e.g. a single endsystem attached to two different providers, a host-based solution would be natural. However, in enterprise networks, requiring each endsystem to implement a complex protocol may not be the best solution, especially since the network managers often want to control the flow of the packets for traffic engineering purposes.

3.2 Automatic allocation of locators

When CIDR was proposed, it was successful in reducing the growth rate of the BGP routing tables by allowing ISPs to better aggregate the prefixes that they advertise [15]. Unfortunately, this success did not continue, mainly due to two factors :

1. the growth of multihoming forced ISPs to advertise more and more more specific prefixes, which increased the size of the FIBs and the number of BGP Update messages [16]

2. there is a pressure from enterprise networks to use PI address space because IP address renumbering is considered too costly

Despite all the work done on renumbering IP addresses [7, 9, 3], enterprise networks are still reluctant to renumber their IP addresses when their provider changes.

The renumbering problem, given all the existing manual configurations of addresses will always remain a difficult problem. If the Internet moves towards the utilisation of locators, **we should not redo the same error and start immediately to work on the development of a protocol/mechanism to automatically allocate locators**. If there is a mechanism in place to automatically allocate locators, then it will be easier for a network to change them when required.

Let us briefly describe a possible way to solve this problem. When a client network is attached to a provider network, a protocol should be used by the provider network to announce to the client network the prefix that the client network should use. This protocol could be a new BGP AFI or could be a new protocol relying on certificates such as the certificates being defined by the SIDR working group. With such certificates, when the customer signs a contracts with its provider, its receives a certificate signed by the provider and valid for the duration of the contract. When the client presents a valid certificate to its provider over a BGP peering, the provider replies with a certificate containing the locator prefix allocated to the client network. Then, a mechanism should be used inside the client network to distribute the received prefix. This mechanism should also allow the network to allocate sub-prefixes to its own customers. It should be possible for a network to indicate to its customers a list of allocated sub-prefixes with orders of preference. It should also be possible for an ISP to remove (or mark as unusable) a previously allocated locator prefix.

Of course, automatic allocation of locators is not sufficient to maintain low FIB sizes. The key issue to reduce the size of the FIBs is to be able to efficiently aggregate locators. To achieve this aggregation, multiple locator prefixes will be allocated to each customer network. Using multiple locators inside a single customer network provide several benefits in terms of path diversity [11]. Provider Aggregatable Locators (PAL) should be the default for stub ASes. Provider Independent Locators (PIL) would be the default for Tier-1 ISPs. For smaller ISPs, either PA or PI locators could be used. We believe that in the long term the benefits of PA locators will encourage small transit ISPs to also utilise such locators.

3.3 Removing ASes from the interdomain routing system

The main objective of the interdomain routing system is the distribution of locators. When redesigning the interdomain routing system, **we should not assume that ASes necessarily need to be visible entities exposed by the interdomain routing protocol**. The visibility of the ASes is an artifact of today's path-vector based interdomain routing. Path vectors have been introduced in BGP to avoid loops, but they cause path exploration and are responsible for a fraction of the exchanged BGP messages [26]. Alternative interdomain routing protocols have been proposed and should be explored [32].

3.4 Interdomain routing convergence

Link failures are common events that affect both internal links inside ASes and peering links between ASes [6]. Studies published in the literature show that many link failures last for a short period of time [20, 6]. Inside ASes, the success of MPLS-based restoration local techniques [33] has shown that it is better to react locally instead of reacting globally. Solutions have been proposed to protect peering links [6] in transit and stub ASes. and other IP-based solutions are being developed by IETF to protect intradomain links [29].

With the current BGP interdomain routing protocol, two timer-based techniques have been introduced to reduce the amount of BGP messages exchanged after a topology change. The MRAI timer allows to wait some time for an iBGP convergence before advertising update messages over eBGP sessions. The BGP dampening process was a reaction to link flaps, but it is not anymore favoured by operators since it has a negative impact on the BGP convergence time [19].

We believe that the interdomain routing system should follow the same approach to handle link failures as for links inside ISP networks. **Instead of potentially advertising globally each link failure, the interdomain routing system should always favour a local reaction** :

- local restoration provides a faster recovery upon peering link failure without requiring a complete convergence [6]
- as many ASes are multiconnected, there is often a backup link between the two concerned ASes that allows to maintain the reachability of the affected locators
- if a failure lasts for a long period, then instead of advertising the failure in the interdomain routing system, it should be possible to deprecate the locator allocated over the failed link

The last bullet implies that the locator allocation mechanism should support the removal or the deprecation of a previously assigned set of locators. When there is a single unprotected link between a provider and a customer network, the failure of this link implies that the locators assigned to the customer network are no longer reachable. Upon detection of the failure, the router attached to the peering link in the provider network will send ICMP locator unreachable messages upon reception of a packet destined to the customer network. Upon reception of such messages the correspondent ASes should consult the mapping mechanism to determine a new locator to reach the concerned identifiers. The customer network should handle the failure by updating its id-locator mapping to indicate that the locators assigned by the provider attached to the failed peering link are not reachable anymore. This will ensure that new flows established by remote networks will not use the failed locator anymore.

3.5 Traffic engineering

Traffic engineering is often a requirement for both transit and stub ASes [2]. In the current interdomain routing system, traffic engineering is often done by tweaking the

BGP advertisements [27]. However, we would like to point out that there are different types of traffic engineering activities, such as :

- selecting the path with the lowest delay to reach a given destination
- selecting the path with the highest bandwidth to reach a given destination
- forcing outgoing packets to use a given peering link
- forcing incoming packets to use a given peering link

With a locator-id split, the traffic engineering problem can be solved at different levels. The first level is the mapping mechanism that will be used to map identifiers to locators. This mechanism should allow to map an identifier to a set of locators, possibly with a preference ordering. The entity that controls such a mapping mechanism, can easily use it for traffic engineering purposes. This solution can work as demonstrated by the success and deployment of Content Distribution Networks today. We believe that **tuning the mapping mechanism is the most efficient and the most scalable way to allow networks to engineer their incoming packet flows**. When an entity selects a locator to reach a given identifier among the various locators proposed by the mapping mechanism, it can also engineer its outgoing traffic by performing this selection. For example, in a shim6 context, we have shown how to perform such traffic engineering to achieve load-balancing [10] or favour low-delay paths [12].

The mapping mechanism offers much better flexibility for the traffic engineering decisions than the interdomain routing protocol. Thus, the mapping mechanism should be the preferred way to perform traffic engineering actions and the interdomain routing should remain as static as possible.

4 Conclusion

In this work-in-progress document, we have discussed several requirements for a new interdomain routing architecture relying on the locator-id split. Our key findings can be summarised as follows :

- The new interdomain architecture should not dictate the exact placement of locators. It should allow both host- and router-attached locators.
- The new interdomain architecture must support the automatic allocation of locators.
- ASes should not necessarily be visible entities in the interdomain routing system.
- Fast convergence should not be a requirement for the interdomain routing protocol of the new architecture. Local restoration mechanisms should be favoured to handle link failures.
- The identifier-locator mapping mechanism should be able to support traffic engineering without any impact on the interdomain routing protocol.

Acknowledgements

We would like to thank Pierre François, Bruno Quoitin, Luigi Iannone and Clarence Filsfils for many fruitful discussions on this topic.

Olivier Bonaventure is partly funded by AGAVE (`http://www.ist-agave.org`), a research project supported by the European Commission under its Sixth Framework Program. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the AGAVE project or the European Commission.

References

[1] S. Agarwal, C. Chuah, and R. Katz. OPCA: Robust Interdomain Policy Rrouting and Traffic Control. In *Proceedings of the 6th International Conference on Open Architecture and Network Programming, IEEE OpenArch*, April 2003.

[2] D. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao. Overview and Principles of Internet Traffic Engineering. RFC 3272 (Informational), May 2002.

[3] F. Baker, E. Lear, and R. Droms. Procedures for Renumbering an IPv6 Network without a Flag Day. RFC 4192 (Informational), September 2005.

[4] H. Berkowitz. Router Renumbering Guide. RFC 2072 (Informational), January 1997. Updated by RFC 4192.

[5] H. Berkowitz, P. Ferguson, W. Leland, and P. Nesser. Enterprise Renumbering: Experience and Information Solicitation. RFC 1916 (Informational), February 1996.

[6] 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.

[7] B. Carpenter and Y. Rekhter. Renumbering Needs Work. RFC 1900 (Informational), February 1996.

[8] E. Chen and J. Stewart. A Framework for Inter-Domain Route Aggregation. RFC 2519 (Informational), February 1999.

[9] M. Crawford. Router Renumbering for IPv6. RFC 2894 (Proposed Standard), August 2000.

[10] C. de Launois, O. Bonaventure, and M. Lobelle. The NAROS approach for IPv6 multihoming with traffic engineering. In *4th COST 263 International Workshop on Quality of Future Internet Services (QoFIS 2003)*, volume LNCS 2811, pages 112–121, Stockholm, Sweden, October 1-3rd 2003. Springer-Verlag.

[11] C. de Launois, B. Quoitin, and O. Bonaventure. Leveraging network performances with IPv6 multihoming and multiple provider-dependent aggregatable prefixes. In *3rd International Workshop on QoS in Multiservice IP Networks (QoSIP 2005)*, Catania, Italy, February 2-4th 2005.

[12] C. de Launois, S. Uhlig, and O. Bonaventure. Scalable route selection for IPv6 multihomed sites. In *Proceedings of Networking 2005*, Waterloo, Ontario, Canada, May 2-6th 2005.

[13] V. Fuller, T. Li, J. Yu, and K. Varadhan. Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy. RFC 1519 (Proposed Standard), September 1993. Obsoleted by RFC 4632.

[14] C. Huitema. An Experiment in DNS Based IP Routing. RFC 1383 (Experimental), December 1992.

[15] G. Huston. Analyzing the internet's BGP routing table. *Internet Protocol Journal*, 4(1), 2001.

[16] G. Huston and G. Armitage. "projecting future ipv4 router requirements from trends in dynamic bgp behaviour". In *Australian Telecommunication Networks and Applications Conference (ATNAC)*, 2006.

[17] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian. An experimental study of internet routing convergence. In *SIGCOMM 2000*, August 2000.

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

[19] Z. M. Mao, R. Govindan, G. Varghese, and R. Katz. Route flap damping exacerbates internet routing convergence. In *ACM SIGCOMM'2002*, 2002.

[20] A. Markopoulou, G. Iannaccone, S. Bhattacharyya, C. Chuah, and C. Diot. Characterization of failures in an IP backbone. In *IEEE Infocom2004*, Hong Kong, March 2004.

[21] D. Meyer, L. Zhang, and K. Fall. "report from the iab workshop on routing and addressing". Internet draft, draft-iab-raws-report-01.txt, work in progress, February 2007.

[22] R. Moskowitz and P. Nikander. Host Identity Protocol (HIP) Architecture. RFC 4423 (Informational), May 2006.

[23] S. Murphy. BGP Security Vulnerabilities Analysis. RFC 4272 (Informational), January 2006.

[24] E. Nordmark and M. Bagnulo. Level 3 multihoming shim protocol. Internet draft draft-ietf-shim6-proto-07.txt, work in progress, Nov 2006.

[25] Ola Nordstrom and Constantinos Dovrolis. Beware of bgp attacks. *SIGCOMM Comput. Commun. Rev.*, 34(2):1–8, 2004.

[26] R. Oliveira, B. Zhanf, D. Pei, R. Izhak-Ratzin, and L. Zhang. Quantifying path exploration in the internet. In *Internet Measurement Conference*, Rio de Janeiro, Brazil, October 2006.

[27] B. Quoitin, S. Uhlig, C. Pelsser, L. Swinnen, and O. Bonaventure. Interdomain traffic engineering with BGP. *IEEE Communications Magazine*, May 2003.

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

[29] M. Shand and S. Bryant. IP Fast Reroute Framework. Internet draft, draft-ietf-rtgwg-ipfrr-framework-06.txt, work in progress, October 2006.

[30] R. Stewart, Q. Xie, K. Morneault, C. Sharp, H. Schwarzbauer, T. Taylor, I. Rytina, M. Kalla, L. Zhang, and V. Paxson. Stream Control Transmission Protocol. RFC 2960 (Proposed Standard), October 2000. Updated by RFC 3309.

[31] L. Subramanian, S. Agarwal, J. Rexford, and R. Katz. Characterizing the Internet Hierarchy from Multiple Vantage Points. In *INFOCOM 2002*, June 2002.

[32] L. Subramanian, M. Caesar, C. T. Ee, M. Handley, M. Mao, S. Shenker, and I. Stoica. Hlp: a next generation inter-domain routing protocol. *SIGCOMM Comput. Commun. Rev.*, 35(4):13–24, 2005.

[33] J.-P. Vasseur, M. Pickavet, and P. Demeester. *Network Recovery: Protection and Restoration of Optical, SONET-SDH, and MPLS*. Morgan Kaufmann, 2004.

[34] F. Wang, Z. 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 *ACM SIGCOMM*, pages 375–387, Pisa, Italy, September 2006.