Scalable De-Aggregation for Overlays Using the Border Gateway Protocol (BGP)
draft-templin-rtgwg-scalable-bgp-00.txt

Abstract

The Border Gateway Protocol (BGP) has well-known limitations in terms of the numbers of routes that can be carried and stability of the routing system. This is especially true when mobile nodes frequently change their network attachment points, which in the past has resulted in excessive announcements and withdrawals of de-aggregated prefixes. This document discusses a means of accommodating scalable de-aggregation of IPv6 prefixes for overlay networks using BGP.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on July 27, 2019.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must
1. Introduction

The Border Gateway Protocol (BGP) [RFC4271] has well-known limitations in terms of the numbers of routes that can be carried and the stability of the routing system. This is especially true for routing systems that include mobile nodes that frequently change their network attachment points, which in the past have resulted in excessive announcements and withdrawals of de-aggregated prefixes. This document discusses a means of accommodating scalable de-aggregation of IPv6 prefixes [RFC8200] for overlay networks using BGP.

2. Overview and Analysis

As discussed in [I-D.ietf-rtgwg-atn-bgp] and [I-D.templin-intarea-6706bis], the method for accommodating scalable de-aggregation is to institute an overlay network instance of BGP that is separate and independent from the global Internet BGP routing system. The overlay is presented to the global Internet as a small number of aggregated IPv6 prefixes (also known as Mobility Service Prefixes (MSPs)) that never change. In this way, the Internet BGP routing system sees only stable aggregated MSPs (e.g., 2001:db8::/32) and is completely unaware of any de-aggregation or mobility-related churn that may be occurring within the overlay.

The overlay consists of a core Autonomous System (AS) with core AS Border Routers (c-ASBRs) that connect to stub ASes with stub ASBRs (s-ASBRs) in a hub-and-spokes fashion. Mobile nodes associate with nearby (i.e., regional) s-ASBRs for extended timeframes, and change to new s-ASBRs only after significant topological or geographic
movements. Mobility-related changes between stub ASes are therefore normally on a long-duration timescale.

The s-ASBRs use eBGP to announce de-aggregated Mobile Network Prefixes (MNP) of mobile nodes (e.g., 2001:db8:1:2::/64) to their neighboring c-ASBRs, but do not announce fine-grained mobility events such as a mobile node moving to a new network attachment point. Instead, mobile nodes coordinate with s-ASBRs using mobility protocols such as MIPv6, LISP, AERO, etc. and s-ASBRs accommodate these localized mobility events without disturbing the c-ASBRs.

The c-ASBRs originate "default" to their neighboring s-ASBRs but do not announce any MNP routes. In this way, MNP announcements and withdrawals are unidirectional from s-ASBRs to c-ASBRs only, thereby suppressing BGP updates on the reverse path. The c-ASBRs in turn use iBGP to maintain a consistent view of the full topology.

We expect that each c-ASBR should be able to carry at least as many routes as can be carried by a typical core router in the global public Internet BGP routing system. Since the number of active routes in the Internet is quickly approaching 1 million (1M), we therefore assume that each set of c-ASBRs can carry at least 1M MNP routes which has been proven even for BGP running on lightweight virtual machines. The method for increasing scaling therefore is to divide the MSP into longer sub-MSPs, and to assign a different set of c-ASBRs for each sub-MSP.

For example, the MSP 2001:db8::/32 could be sub-divided into sub-MSPs such as 2001:db8:0010::/44, 2001:db8:0020::/44, 2001:db8:0030::/44, etc. with each sub-MSP assigned to a different set of c-ASBRs. Each s-ASBR peers with at least one member of each c-ASBR set and uses route filters such that BGP updates are only sent to the c-ASBR(s) that aggregate the specific sub-MSP. Then, assuming 1000 or more sub-MSPs (each with its own set of c-ASBRs) the entire BGP overlay routing system should be able to service 1 billion (1B) MSPs or more.

3. Opportunities and Limitations

Since a lightweight virtual machine (e.g., a Ubuntu linux image running Quagga in the cloud) can service up to 1M MNPs using BGP, it is conceivable that dedicated high-performance router hardware could support even more - perhaps by a factor of 10 or more. With such dedicated high-performance hardware, the numbers of MNPs that can be serviced could be increased further.

The deployed numbers of s-ASBRs even for very large overlays should not exceed the c-ASBR’s capacity for BGP peering sessions. For example, c-ASBRs should be capable of supporting a few thousands to a
few tens of thousands of BGP peering sessions but it is not known whether more could be supported.

By the same token, the maximum number of c-ASBR sets should be based on the number of BGP peering sessions each s-ASBR can comfortably accommodate, since each s-ASBR must peer with each c-ASBR set.

Packets sent between end systems that associate with different s-ASBRs would initially need to be forwarded through the core AS, which presents a forwarding bottleneck. For this reason, some form of route optimization is needed to significantly reduce congestion in the core and preferably to also allow for direct end system to end system communications without involving s-ASBRs. Since c-ASBRs should be standard commercial off-the-shelf (COTS) dedicated high-performance IPv6 routers, however, they should not be required to participate in any ancillary route optimization signaling. The AERO route optimization function honors this design consideration.

Further opportunities and limitations are discussed in more detail in the references [I-D.ietf-rtgwg-atn-bgp][I-D.templin-intarea-6706bis].

4. IANA Considerations

This document does not introduce any IANA considerations.

5. Security Considerations

Security considerations are discussed in the references.

6. Acknowledgements

This work is aligned with the FAA as per the SE2025 contract number DTFWA-15-D-00030.

This work is aligned with the NASA Safe Autonomous Systems Operation (SASO) program under NASA contract number NNA16BD84C.

This work is aligned with the Boeing Information Technology (BIT) MobileNet program.

7. References

7.1. Normative References

7.2. Informative References

[I-D.ietf-rtgwg-atn-bgp]

[I-D.templin-intarea-6706bis]
Templin, F., "Asymmetric Extended Route Optimization (AERO)", draft-templin-intarea-6706bis-03 (work in progress), December 2018.

Appendix A. Change Log

<< RFC Editor - remove prior to publication >>

Status as of 01/23/2018:

-00 draft published

Author’s Address

Fred L. Templin (editor)
Boeing Research & Technology
P.O. Box 3707
Seattle, WA 98124
USA

Email: fltemplin@acm.org