idr
Internet-Draft
Intended status: Informational
Expires: April 21, 2013

W. Kumari
Google
K. Patel
Cisco Systems
J. Scudder
Juniper Networks
October 18, 2012

Automagic peering at IXPs. draft-wkumari-idr-socialite-02

Abstract

This document describes a method for automatically establishing BGP peering sessions at an Internet exchange point. Creation of these peering sessions is facilitated by a host.

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 http://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 April 21, 2013.

Copyright Notice

Table of Contents

1. Introduction

A large amount of Internet traffic is exchanged at Internet Exchange Points (IXP). These are networks that are specifically built and operated as locations for networks to peer and exchange traffic.

Public peering refers to peering across the (IXP provided) switch fabric. In order to avoid having each participant at the IXP having to contact all of the other participants to enter into peering relationships, the IXP often provides a Route Server (RS). The Route Server is a BGP speaker that participants peer with and announce routes to. The Route Server takes these announcements and serves them to all of the other participants who peer with it (so far this is just like any other BGP router!). The Route Server differs from a standard eBGP speaker in that it neither updates the Next Hop, nor prepends its own AS to the AS Path attribute. By not changing the Next Hop attribute, traffic between participants flows directly between those participants (and does not pass through the Route Server), as the traffic doesn't flow though it, it is appropriate that it doesn't appear in the AS Path - this is known as a transparent Route Server (by not showing up in the AS Path, the fact that the peering between the participants occurs over a public peering session is hidden, and participants are not penalized by having longer AS Paths).

This document describes an alternate solution for peering at an IXP. Instead of having a server that re-announces the routes from each participant to all of the others, we introduce a "socialite", a device that is responsible to making introductions between all of the participants and facilitating connections between them. This socialite can be thought of like a host at a dinner party. The guests arrive and the socialite introduces them to each other, and then steps out of the way to allow them to communicate (and peer!) on their own.

This solution is aimed at operators who are currently peering with route-servers (and operators of those route-servers), and it is not expected to be a good alternative to "private peerings".

1.1. Requirements notation

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

2. Terminology

Internet Exchange Point A network for exchanging BGP routing information and traffic.

Route Server A BGP speaker at an IXP that "reflects" routes from one participant to all the other participants. See [I-D.jasinska-ix-bgp-route-server]

Socialite A device running the Introduction protocol, responsible for making introductions between Guests.

Guests Participants of the IXP that speak the Debut protocol with the Socialite, and are introduced by the Socialite to other Guests.

Debut The protocol spoken between the Socialite and the Guests.

3. Overview

The Guests at the IXP form a BGP peering relationship with the Socialite, announcing support for the Debut protocol. The Socialite sends the Guests a set of Debut updates, containing informations about the other participants. The Guests use this information to form direct BGP peerings between themselves. Policy can be configured on the Socialite to only make introductions between subsets of participants if so desired.

4. Protocol Extensions

The BGP protocol extensions introduced in this document include the definition of a new BGP capability, named "'Debut Capability", and the specification of the message subtypes for the Debut messages.

4.1. Debut Capability

The "Debut Capability" is a new BGP capability [RFC5492]. The Capability Code for this capability is specified in the IANA Considerations section of this document. The Capability Length field of this capability is zero. By advertising this capability to a peer, a BGP speaker conveys to the peer that the speaker supports the message subtypes for the Debut protocol and the related procedures described in this document.

5. Packet Formats

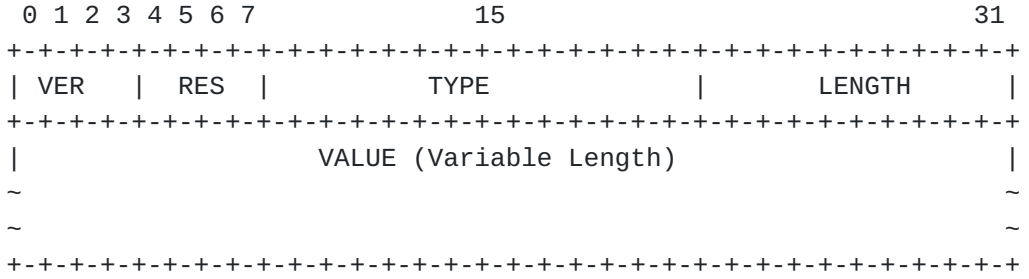
The Debut protocol is implemented using TLV structures, and fields are in network byte order. These TLV records are carried as payload in a standard BGP Message packet (RFC 4271, Section 4.1. Message

Header Format) [RFC4271]

5.1. Message Header

The Debut protocol is implemented using the standard Type-Length-Value paradigm.

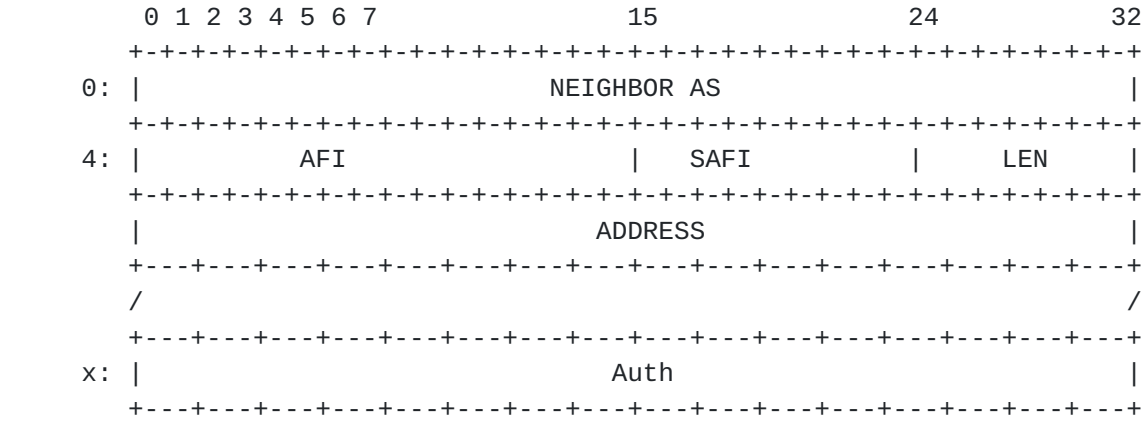
All Debut messages are carried as payload in a standard BGP Message of Type [TBD_BGP] and are preceded by a standard header that specific the type and length of the message.



- VER (4 bits): The VER (version) field specifies the version of the Debut protocol. For the initial (this) version of the protocol it will be set to 0.
- RES (4 bits): The RES (reserved) field is reserved in the initial (this) version of the protocol. It SHOULD be initialized to 0 on transmit and should be ignored on reception.
- TYPE (2 octets): The TYPE field species the TYPE of the TLV record, and allows an implementation to determine what type of information is carried in the record. If the highest bit of the TYPE field is set (the TYPE value is >= 32786), understanding / implementation of the TYPE is optional - if an implementation does not implement this type it may ignore this message (this capability is included to allow for possible future logging, diagnostics, etc). If the highest bit is not set, and the implementation receives a TYPE that it does not implement, it should send a BGP NOTIFY and tear down the session. The TYPE codes are defined in the
- LENGTH (2 octets): The number of octets in the VALUE field of the TLV record. The total length of the TLV record in octets can be calculated by adding 4 (the number of octets in the TYPE and LENGTH fields) to the value of this field. This allows implementations to skip over TLV records that it cannot handle.
- VALUE (Variable length): The actual data. The meaning of this data is given by the TYPE filed, and the length by the LENGTH field. Parsing of the data field is performed according to the value of the TYPE field.

5.2. INTRODUCTION Record

INTRODUCTION (TYPE 0) TLV records are used to "make introductions" between the Guests speaking the Debut protocol. They carry the information needed by Guests to contact the other Guests and establish a BGP peering session.



- NEIGHBOR AS (4 octets): This specifies the Autonomous System of the Guest being introduced. In "Four-octet AS Number" format as specified in RFC4893 [RFC4893]
- AFI (2 octets): Address Family Identifier [RFC4760] [RFC4760]
- SAFI (1 octet): Subsequent Address Family Identifier [RFC4760] [RFC4760]
- LEN (1 octet): The length of the address in the ADDRESS field (32 for IPv4, 128 for IPv6).
- ADDRESS (Variable length) This contains the IP Address of the Guest being introduced.
- Auth (optional, variable length): The existence of this field is determined from LENGTH of the TLV. If the LENGTH is greater than the length of NEIGHBOR AS, FAMILY and ADDRESS, there is Auth data).

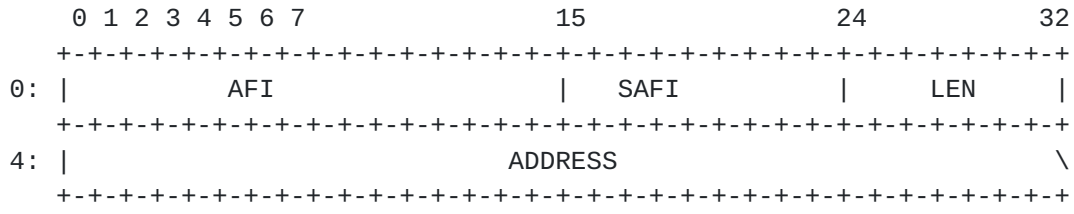
The AFI and SAFI are included in the INTRODUCTION message to allow the Socialite to introduce Guests with multiple address families.

On reception of an INTRODUCTION message a Guest should store the information and then consult local policy (if any) to determine if it is willing to peer with the newly introduced Guest. If so, it should proceed as though this were a manually configured peer. This peering SHOULD be annotated to note that this is a Socialite created peering. It is recommended that the peering show up in the configuration, but not persist across reboots -- this is to allow operators to more easily see all neighbors while looking through the config.

5.3. WITHDRAW Record

WITHDRAW (TYPE 1) TLV records are used to inform a Guest that another previously introduced Guest is no longer participating. A Guest can use this information to abort in progress connection attempts, invalidate information from a cache, for informational logging or in any other way is sees fit, but it SHOULD NOT use this information to tear down peering sessions to other Guests in ESTABLISHED state. Debut is intended to make initial introductions between participants and does not provide any mechanisms to invalidate / abort sessions once the introductions have been made.

If a Socialite attempts to unintroduce an unknown Guest, this information should be logged and then ignored.



- AFI (2 octets): Address Family Identifier [RFC4760] [RFC4760]
- SAFI (1 octet): Subsequent Address Family Identifier [RFC4760] [RFC4760]
- LEN (1 octet): The length of the address in the ADDRESS field (32 for IPv4, 128 for IPv6).
- ADDRESS (Variable length): This contains the IP address of the Guest that is no longer participating.

The AFI and SAFI are included so that the Socialite can inform Guests that only one of the AFI / SAFIs is being removed.

6. Protocol operation

Debut BGP sessions behave just like any other BGP sessions, just the information carried is different - Guests should use the standard BGP peering process to contact Socialites (or Socialites, Guests). Once the peering is ESTABLISHED, Guests will begin receiving INTRODUCTION messages in UPDATES, and will store them in something resembling Adj-RIB-IN. Standard BGP logic applied for things like error handling, invalidation of previously received information, etc.

As Debut is only intended to make initial introductions between Guests (and not to manage sessions between those Guests), if the BGP session between Guest and Socialite goes down, established BGP peerings between Guests will continue to remain active.

7. Operational overview / implications

There are many reasons why participants peer with route-servers at IXPs (see [I-D.jasinska-ix-bgp-route-server]) including

- reducing the administrative burden of arranging and configuring BGP sessions with all the other participants,
- not wanting (or being able) to carry views from all the participants,
- relying on the IXP operator to implement routing policy decisions (see [I-D.jasinska-ix-bgp-route-server], section 2.3)

This solution only attempts to address the first reason for using a route-server, and the implications of deploying this are described below.

7.1. Additional eBGP sessions.

Debut is used to make introductions between all (or a subset of) participants at an IXP, and then the participants peer over "regular" BGP peerings. This means that each participating router will build a separate BGP peering session with every other participating router. As participants at IXPs (usually) only advertise a small subset of the full Internet routing table (such as internal or customer routes) and there is (usually) not a huge overlap of this routing information, the additional memory requirements are expected to not be too onerous (especially with the capacity of modern routers). As with all operational matters though, "Your network, your rules" applies -- it is up to each operator to determine the applicability / utility of the solution and how it fits (or doesn't) into their network (see what I did there? If this ends poorly, it's your own fault!)

7.2. Simplified debugging.

As "normal" eBGP peering sessions are setup between the participants (and there is no third party performing route-selection, etc), operators have more visibility into the system and can more easily leverage their existing troubleshooting / debugging skills to debug issues. Debut also more closely aligns the data and control plane, etc.

7.3. BGP PATH / SIDR implications

Part of the justification for this work is to simplify the design and implementation of path security in SIDR. As a Route-Server passes routing information between peers, but does not show up in the BGP AS-PATH it is indistinguishable from a BGP path shortening attack. By using Debut, eBGP speakers peer directly with each other and this

problem is avoided.

8. IANA Considerations

IANA is requested to assign an AFI and SAFI for the Debut protocol. The text TBD1 should be replaced with the allocated AFI and the text TBD2 should be replaced with the allocated SAFI (and then this sentence should be removed).

The IANA is requested to assign a value from the "BGP Message Types" registry and replace the text [TBD_BGP] with this value. The definition should be "Debut protocol".

8.1. Debut TYPE registry

This document creates a new registry, "Debut Message Types".

The registry policy is ""Specification Required".

The initial entries in the registry are:

Value	Short description	Reference
0	INTRODUCTION	[This]
1	WITHDRAW	[This]
2-3200	Unassigned	
3200-32767	Private Use	
32768-65480	Unassigned	
65481-65535	Private Use	

Applications to the registry can request specific values that have yet to be assigned.

9. Security considerations

This protocol is designed to facilitate direct BGP peerings between participants at an IXP, which eliminates the need for transparent route servers (which do not show up in the AS_PATH). This will facilitate the deployment of SIDR.

As participants peer with each other directly (and not through a third party) there is less opportunity for malicious tampering with the control plane (for example, by the IXP).

Debut currently does not provide a means to securely distribute Authentication information (there is a field, but it's not really

defined). Depending on if needed this may be addressed.

An attacker who manages to subvert the Socialite (or inject UPDATES that into the Socialite to Guest communication) will be able to make Guests peer with a device under his control -- the impact of this seems to be no worse than in the routeserver model.

Currently routeserver operators perform some base level checking / sanitization of routing information (such as enforcing max-paths) - in the socialte model each operator is expected to perform thier own checks.

9.1. Privacy

By having participants peer directly (as opposed to having their routing information pass through a route-server) the routing information is hidden from the IXP / route-server operator. Please note that this doesn't protect the data-plane, and the routing information could still be sniffed off the wire.

The biggest concern with regards to privacy on a route server is towards propagating your policy to a third party, rather than propagating your routing information.

10. Acknowledgements

The authors wish to thank Elisa Jasinska, Masataka MAWATARI, Robert Raszuk, Martin Hannigan, Simon Leinen.

11. Author Notes

[RFC Editor -- Please remove this section before publication!]

1. Choose a better name than "Debut"

11.1. Changelog.

- Changed the name of the protocol from 'elo-'elo to Debut - this is still not great, but "Introduction" is worse.
- Added Operational section, incorporated notes from John Scudder, Keyur.

11.2. Changes from -00 to -01

- Incorporated some comments from Elisa Jasinska
- Mainly version bump tp prevent expire!

11.3. Changes from -01 to -02

- Incorporated some long lingering nits / suggestions.
- 9 or 10 folk have expressed interest and asked us to revive this. I (Warren) have done a really poor job of taking notes and incorporating them. Appologies, if you mentioned issues to me in person I have probably forgotten to incorporate them, *please* send them in email and I'll get to them.
- Clarified the audience slightly, improved some security bits.

12. References

12.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

[RFC4893] Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space", RFC 4893, May 2007.

12.2. Informative References

[I-D.jasinska-ix-bgp-route-server] Jasinska, E., Hilliard, N., Raszuk, R., and N. Bakker, "Internet Exchange Route Server", draft-jasinska-ix-bgp-route-server-03 (work in progress), October 2011.

Authors' Addresses

Warren Kumari
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
US

Email: warren@kumari.net

Keyur Patel
Cisco Systems

Phone:
Fax:
Email: keyupate@cisco.com
URI:

John Scudder
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA
USA

Phone:
Fax:
Email: jgs@juniper.net
URI: