November 29, 2016

BGP COMMUNITYの世界動向

吉村 知夏 <c.yoshimura@ntta.com>
NTT America

Agenda

- 自己紹介
- BGPコミュニティについて
- BGPコミュニティを利用した経路制御
- BGPコミュニティをめぐるトレンド
- まとめ

自己紹介

- NTTアメリカ Solution Development Engineer (2016年11月より)
  - ネットワークプロダクトマネジメント全般
  - SD-WANの顧客導入案件など
- 2012年9月～ NTTアメリカ
  - GIN (AS2914) バックボーンNW運用開発
- 2003年4月～ NTTコミュニケーションズ
  - OCN (AS4713) サーバ運用
  - OCN (AS4713) バックボーンNW開発
- JANOG会長 (2015年9月より)

BGPコミュニティ？？？

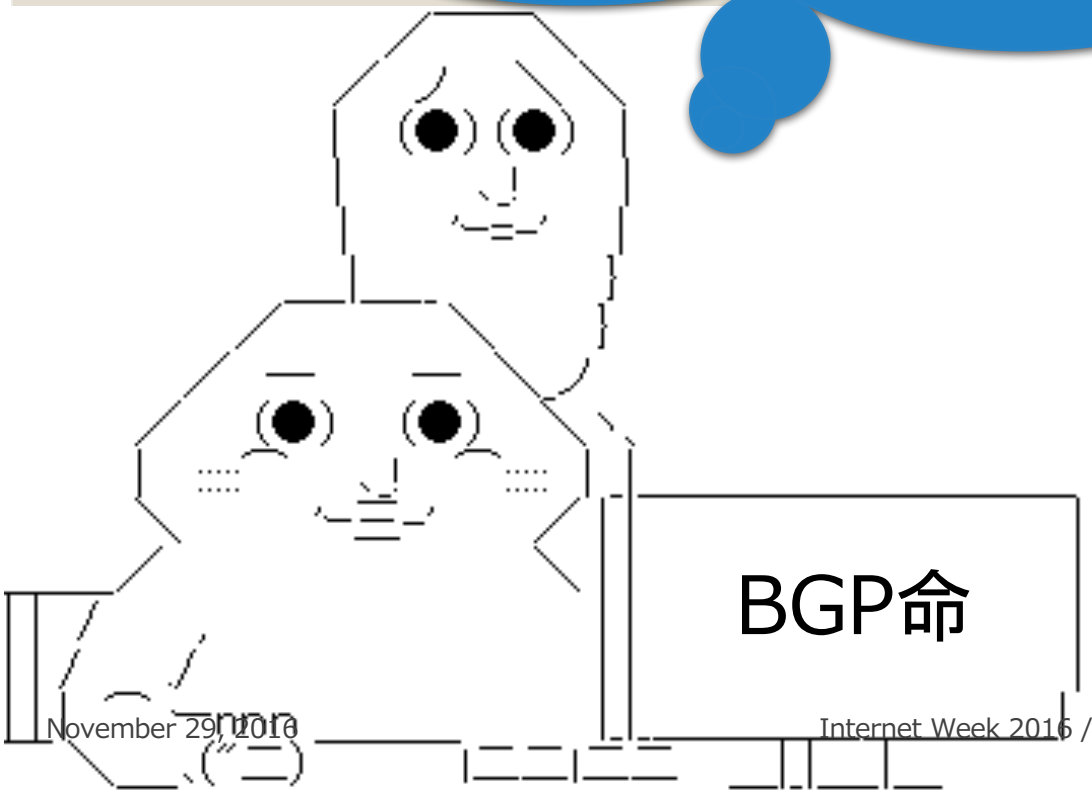
```
BGP routing table entry for 180.0.0.0/10
Versions:
  Process              bRIB/RIB  SendTblVer
  Speaker            141186469   141186469
Last Modified: Nov 14 22:23:20.039 for 1d06h
Paths: (2 available, best #1)
  Advertised to update-groups (with more than one peer):
    0.1 0.3 0.14 0.30 0.33 0.34 0.36 0.40
    129.250.31.182  195.22.206.36   129.250.193.162 64.125.14.69
    129.250.8.226   206.126.236.1   206.126.237.220
  Path #1: Received by speaker 0
  Advertised to update-groups (wi   more than one peer):
    0.1 0.3 0.14 0.30 0.33 0.34 0     0.40
  Advertised to peers (in unique    ate groups):
    202.97.32.87    4.68.71.237     129.250.66.50   193.251.248.137
    128.241.219.82  63.146.27.245   129.250.8.122   206.126.236.137
    206.126.237.42  206.126.236.4   206.126.239.250 206.126.237.141
    129.250.31.182  195.22.206.36   129.250.193.162 64.125.14.69
    129.250.8.226   206.126.23      .126.237.220
  4713
    117.103.176.22 (metric 16325    m 129.250.0.125 (129.250.0.125)
      Origin IGP, metric 100, loca   ref 120, valid, confed-internal, best,
      Received Path ID 0, Local Pa   ID 1, version 141186469
      Community: 2914:370 2914:490 2914:1403 2914:2401 2914:3400
```

Community: 2914:370 2914:490…


謎の数字と
コロンが付い
ている

BGPオペレーターがこれを見ると、

「NTTの顧客経路で、
アジアの日本の大阪の経路。
NTT内部でのLocal
Preferenceは120（略」



みたいなことが
0.1秒で分かる

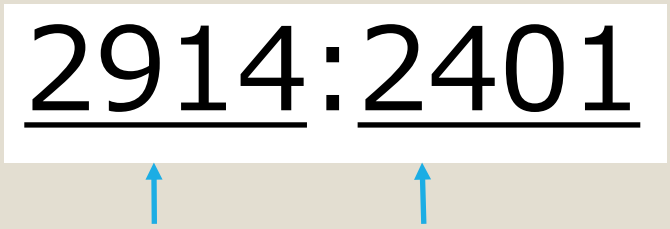
…のは一部の
BGPオタクだけ

今日は一緒にBGP
オタクになりましょう！

ようこそ
BGPコミュニティの世界へ

BGPコミュニティ (1)

- BGPパス属性の一つ
- BGP経路に任意情報を付与するためのパス属性
- RFC1997 “BGP Communities Attribute”
  - 32 byte attribute (16-bit ASN : 16-bit information)
- RFC4360 “BGP Extended Communities Attribute”
  - 64 byte attribute (16-bit ASN : 48-bit information)
- 某ASの例



BGPコミュニティ (2)

Asian country origins (2914:24--)

2914:2401	jp (Japan)
2914:2402	au (Australia)
2914:2403	hk (Hong Kong)
2914:2404	tw (Taiwan)
2914:2405	kr (Korea)
2914:2406	sg (Singapore)
2914:2407	my (Malaysia)
2914:2408	id (Indonesia)
2914:2409	bn (Brunei)
2914:2410	th (Thailand)

ASIA MSA origins (2914:14--)

2914:1401	Kuala Lumpur, Malaysia
2914:1402	New Territories, Hong Kong
2914:1403	Osaka, Japan
2914:1404	Seoul, South Korea
2914:1405	Singapore
2914:1406	Sydney, Australia
2914:1407	Taipei, Taiwan
2914:1408	Tokyo, Japan
2914:1409	Tseung Kwan O, Hong Kong
2914:1410	Jakarta, Indonesia
2914:1411	Bandar Seri Begawan, Brunei
2914:1412	Bangkok, Thailand

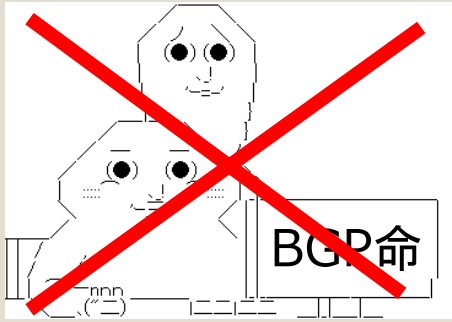
経路に付ける情報タグみたいなもの
タグの中身はAS毎に好きに決めて良い

BGPコミュニティ (3)

- Optional Transitive
  - 他ASにも伝播する
  - 消去や上書きが可能
  - 経路に付与したコミュニティが対向ASに伝わるかは分からない (消去や上書きされる場合があるため)
  - ASごとに確認が必要
- Well-known community
  - no-export (AS外に経路広告しない)
  - no-advertise (他のBGPピアに経路広告しない)
  - blackhole <new!> (後述)

BGPコミュニティでできることとその目的

- BGPコミュニティは、**経路への情報付加のみを行う**
- ~~付加された情報を見てニヤニヤするのが目的~~



- 付加情報に応じた経路制御を別途行い、**トラフィックの流れをできるだけ思い通りに操作することが最終目的**
  - 多数の経路に一括して情報を付与することで、多くの経路を効率的に制御できるメリットがある
  - AS間で連携を取り、情報に応じたきめ細やかな経路制御を行う

制御＝BGPフィルターを書く

- **経路制御をするためにはBGPフィルターを別途書く必要がある**

1. コミュニティを設計する@机上
   - (例) 日本の経路には2914:2401を付与する
2. コミュニティに応じた任意のアクションを設計する@机上
   - (例) 2914:2401が付与されている経路は、全顧客へ広報する
3. アクションを実現するためのBGPフィルターを書く@ルータ
4. コミュニティを経路に付与し広報する@ルータ
5. 任意のアクションが行われる@ルータ

BGPコミュニティを利用した制御例

1. 広告経路を制御する
   - 広告先ASでの扱い方を指定 (広告範囲、優先度など)
   - トラフィックの流入元を管理できる
2. 受信経路を制御する
   - 自網ASでの扱い方を指定
   - トラフィックの流出先を管理できる
3. 内部管理用
   - 経路数をカウントする
   - 経路宛のパケット数をカウントする など

可能性は無限大！BGPオタク歓喜！

主に2つの制御がメイン

1.広告経路を制御する
2.受信経路を制御する

これから
話します

広告経路を制御する

- **自分が相手に広告した経路を、相手にどう扱ってもらうか指定すること**

1. 広告範囲を指定
   - 特定{地域, 国, ピア, 顧客}にのみ広報
   - 特定{地域, 国, ピア, 顧客}に広報しない
   - **Blackholeする (ここ数年のトレンド)**
2. 任意の優先度を指定
   - Local Preferenceを任意の値にする
3. Prepend付与 (あんまり人気ない)
   - 特定{地域, 国, ピア, 顧客}に[1-3]個Prepend

こんな構成があったとき

海外からDDoSが来たら…

大きく2つの打ち手がある

1. トラフィックをフィルターする
2. 経路を操作する
   - 経路広告を停止する
   - 経路の優先度を下げる
   - 経路広告を維持しつつ、next hop を null にする (blackhole)

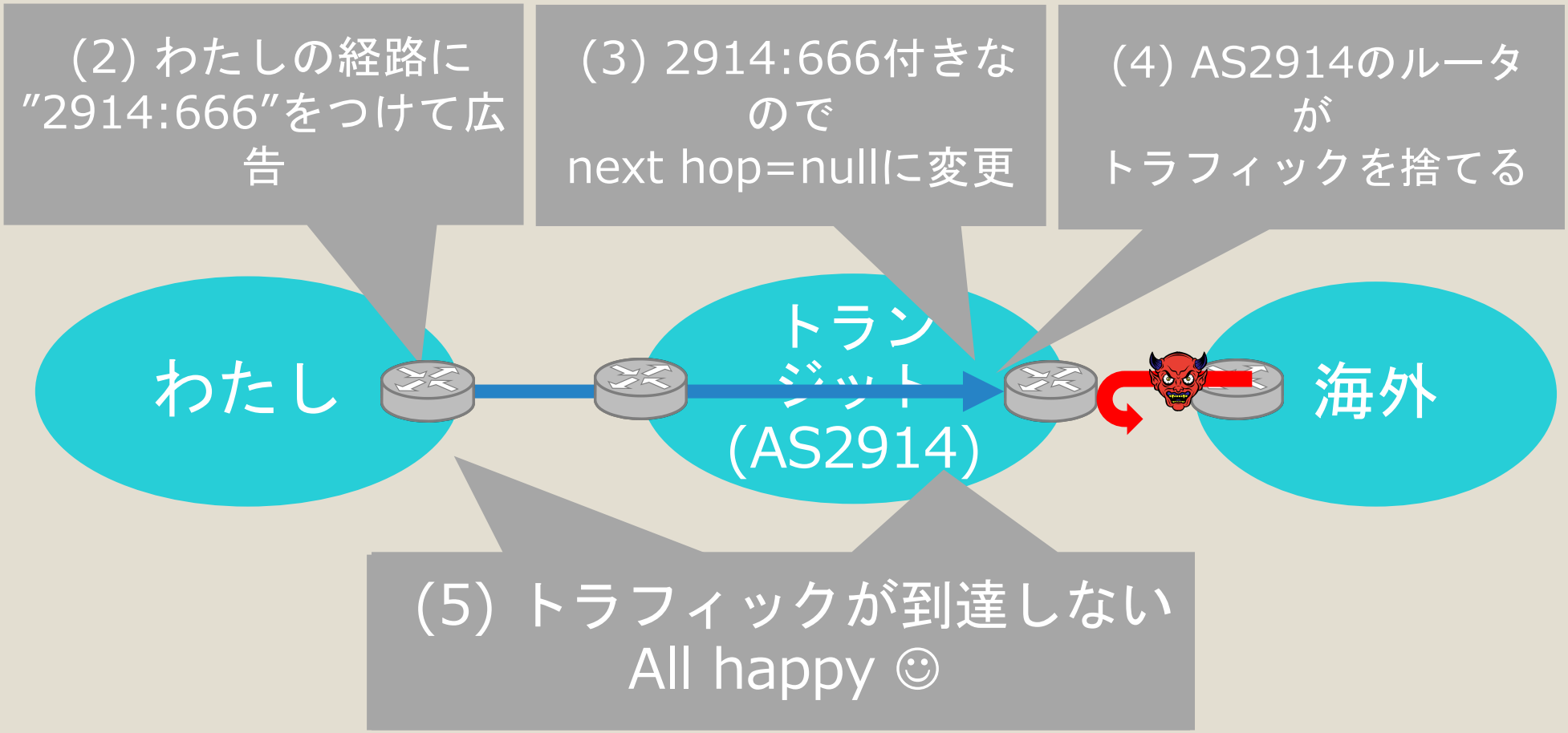
トランジット事業者(上流ISP)のBGPコミュニティを利用すれば、自分で実施できる

- DDoS対応サービスを買うという選択肢もあるがここでは割愛

BGPコミュニティを使って経路広告を自分で操作する(blackholeを例に)

(1) トランジット側：あらかじめBGPフィルターを設定しておく
「2914:666付きの経路は**next hop = nullとし、トラフィックをdiscard**する」

BGPコミュニティを使って経路広告を自分で操作する(blackholeを例に)



よりきめ細やかな操作

- 実際には、よりきめ細やかな経路操作ができることが多い (地域、国、ASごと、顧客ごと、ピアごと)

操作内容	操作できる範囲
Blackhole	任意の{地域, 国}のみnullにする
経路広告停止	任意の{地域, 国, AS} のみ経路を広告しない
優先度変更	任意の地域のみLPを下げる
Prepend	特定の{地域, ピア, 顧客}のみPrependする

- トランジット事業者は、世の中のニーズを見ながらコミュニティを設計し、ユーザに利用開放している

トランジット事業者の裏事情

- **膨大なBGPフィルターを全ルータに設定**している
- **矛盾するコミュニティの扱いについても検討**が必要
  - 例えば、「アジアに経路を出さないコミュニティ ＋ 日本に経路広告するコミュニティ」は両立できるか？
  - とても悩みどころ (設計している側も度々わからなくなる…)
- BGPコミュニティをユーザに利用解放しているのは一部事業者
  - 国際トランジット事業者 (Tier1、グローバル展開をしている事業者) や北米事業者に多い
  - 日本のトランジット事業者 (Tier2) における解放は今のところ一般的ではない
    - コストとニーズのバランス
    - 日本国内のみで経路を操作するニーズがそれほどない…?

BGPコミュニティを使う側

- **予備知識がモノを言うことが多い**
- 利用前の確認事項
  1. トランジットISPがBGPコミュニティを解放しているか
  2. 解放している場合、コミュニティ一覧と動きを把握する
     - 矛盾するコミュニティを付与した時の動きも把握
  3. 経路操作＝トラフィック流入元が変化することを念頭に置き、どのように変化するか机上検討する
- 必要な時にさっとBGPコミュニティを付与できるよう、あらかじめBGPフィルターを準備しておく
  - Blackholeなど緊急性を要するものは特に

トランジット事業者のBGPコミュニティ確認方法

- 事業者にヒアリングする
  - 「BGP community を使った経路操作をサポートしてますか？」
- Webサイトを見る
  - AS2914 https://www.us.ntt.net/support/policy/routing.cfm
  - https://onestep.net/communities/

Community	Effect
7922:999	Prefixes are not sent to anyone. They are contained entirely within AS7922 only
7922:888	Prefixes are not sent to ALL peers. Prefixes still sent to customers
65100:XXX	Do not announce prefixes to AS XXX
65200:XXX	Announce to AS XXX if 7922:888 is also set

Comcast (AS7922)

```
6461:5000       suppress announcement to all peers
6461:5001       prepend once to all peers
6461:5002       prepend twice to all peers
6461:5003       prepend three times to all peers

6461:5010       suppress announcement to all EU peers
6461:5011       prepend once to all EU peers
6461:5012       prepend twice to all EU peers
6461:5013       prepend three times to all EU peers
```

Zayo Communications (AS6461)

8220:63099 Do not advertise to any peers
8220:63999 Blackhole all traffic
8220:63800 No Export
8220:63700 Set local preference to 120
8220:63900 Set local preference to 40

COLT (AS8220)

Local-pref	Community	Description
70	2828:1507	Lower (less preferred) than all other routes on network, including all public & private peer routes.
80	2828:1508	Same as public & private peer routes.
90	2828:1509	
100	none	Default for all customer routes, BGP and static.
100	2828:1510	Explicitly set customer BGP announcements to 100.
110	2828:1511	Higher (more preferred) than all default customer BGP and static routes.
120	2828:1512	Highest (most preferred) local-pref that a customer can specify.

XO Communications (AS2828)

(おまけ) 受信経路を制御する

- **相手からもらった経路を、自分がどう扱うか指定すること**

1. 広告範囲を指定
   - 特定{地域, 国, ピア, 顧客}にのみ広報
   - 特定{地域, 国, ピア, 顧客}に広報しない

2. 任意の優先度を指定
   - Local Preference
   - MED
   - 地域、国、ピア、顧客ごとに優先度を変える

BGP Communityをめぐるトレンド

1. DDoSトラフィックをどうにかしたい！
   - 先のページで書いたような、経路のblackholeを実現するためのBGPコミュニティが、大規模事業者を中心に提供されつつある
   - “Blackhole Community” (*:666) がIETFでRFC化

2. 4-byte AS対応が必要なんじゃないか？
   - “BGP Large Community” のInternet DraftがIETFに提出

Blackhole Community

Internet Engineering Task Force (IETF)
Request for Comments: 7999
Category: Informational
ISSN: 2070-1721

T. King
C. Dietzel
DE-CIX
J. Snijders
NTT
G. Doering
SpaceNet AG
G. Hankins
Nokia

October 2016

BLACKHOLE Community

Abstract

This document describes the use of a well-known Border Gateway Protocol (BGP) community for destination-based blackholing in IP networks. This well-known advisory transitive BGP community named "BLACKHOLE" allows an origin Autonomous System (AS) to specify that a neighboring network should discard any traffic destined towards the tagged IP prefix.

Blackhole Community

- 2016年10月 RFC化 (RFC7999)
- “*:666”が付く経路宛のトラフィックは、各ASでdiscardして良いことにするという**取り決め**
  - 実際にdiscardするかは各ASに任されている
  - 実際にdiscardするためには、各ASで別途BGPフィルターを書く必要がある
- 実運用の世界ではこんな感じです
  - AS2914では、お客様から2914:666のつく経路を受信した場合、next hopをnullに変更することでトラフィックをdiscard
  - ASによっては、地域や国別のblackholeができることもある (selective blackhole)

BGP Large Community

IDR
Internet-Draft
Intended status: Standards Track
Expires: May 18, 2017

J. Heitz, Ed.
Cisco
J. Snijders, Ed.
NTT
K. Patel
Arrcus
I. Bagdonas
Equinix
N. Hilliard
INEX
November 14, 2016

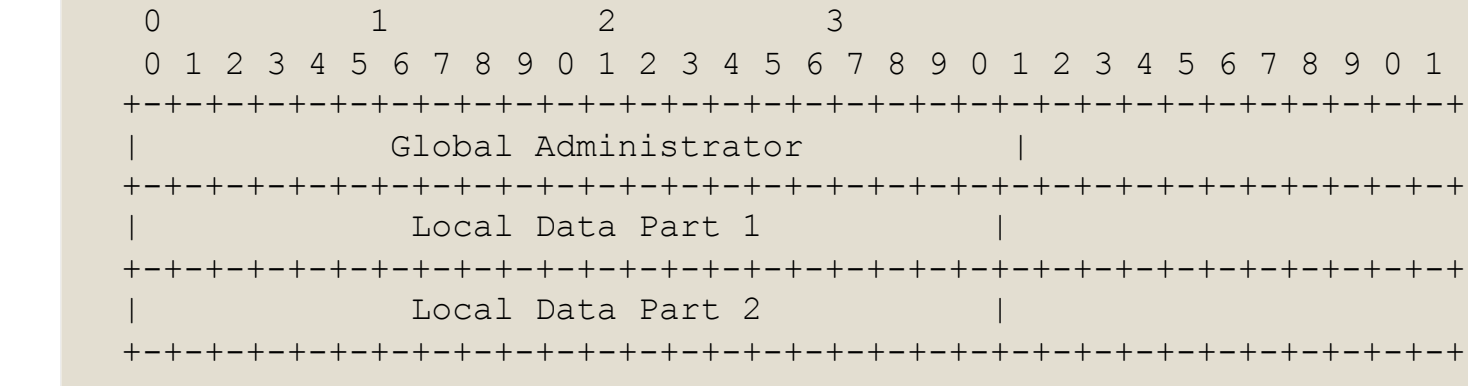
BGP Large Communities
draft-ietf-idr-large-community-08

Abstract

This document describes the BGP Large Communities attribute, an extension to BGP-4. This attribute provides a mechanism to signal opaque information within separate namespaces to aid in routing management. The attribute is suitable for use with four-octet Autonomous System Numbers.

BGP Large Community

- http://largebgpcommunities.net/
- “大きな” BGPコミュニティ
  - 96 byte attribute (32-bit ASN:32-bit value:32-bit value)
  - これまでは 32 or 64 bytes
  - RFC1998を踏襲したスタイル
- IETFにInternet Draft 提出済
  - 11/14 現在第8版



標準化のモチベーション

- **既存のBGP Communityは16bit AS対応を前提に設計されている**

  16-bit-ASN:value

- **32bit AS番号でも使えるBGPコミュニティを作りたい**

  32-bit-ASN:value:value

- 2001年- 32-bit ASNの標準化活動スタート
  - 2007年5月 RFC 4893
  - 2012年12月　RFC 6793
- 2007年- 32-bit ASNがRIRから払い出され始める
- 32bit ASNはグローバルに利用されている
- 32bit ASNのオペレーターは、16bitのプライベートASNを利用したりしながら現状をしのいでいる (らしい)

32-bit ASNs in a 16-bit Field

Large BGP Community の例

RFC 1997 (Current)	Large BGP Communities	Action
65400:peer-as	2914:65400:peer-as	Do not Advertise to peer-as in North America (NTT)
0:peer-as	6667:0:peer-as	Do not Announce to Route Server peer-as (AMS-IX)
65520:nnn	2914:65520:nnn	Lower Local Preference in Country nnn (NTT)
2914:410	2914:400:10	Route Received From a Peering Partner (NTT)
2914:420	2914:400:20	Route Received From a Customer (NTT)

32-bit (私が) : 32-bit (どういうアクションを) : 32-bit (誰/どこなど、何をターゲットに行う) という書き方をする

デザインのゴール

- できるだけシンプルに
  - 複雑な機能追加はしない
  - RFC1997をベースに拡張
  - 自分のASNや、target情報の欠落なしにアクションを記載できるように
- できるだけ実装しやすく
  - Transitiveにする
- 柔軟な設計に耐えられるように
  - ネットワークオペレーターがコミュニティを設計しやすいように
- すべての 16-bit and 32-bit ASNsが使えるように
  - プライベートASNを使わなくても良いように
  - ASNの重複が起こらないように
- 覚えやすく、電話でもうまく伝わるくらい簡単に
  - 私：アクション：ターゲット という表記
  - 異言語環境でもうまく話が通じるように

やらないこと

- well-known communityは設定しない
  - RFC1997のwell-known communityは問題なく使えるため
- TLV や headerは利用しない
  - BGP Path Attributes code 30 (0x1E) を利用

実装状況 (11/20現在)

BGP Speakers

Vendor	Software	Status	Details
Arista	EOS	Planned	Feature Requested BUG169446
Cisco	IOS XR	✔ Done!	Engineering Release
cz.nic	BIRD	✔ Done!	BIRD 1.6.3 (commit)
ExaBGP	ExaBGP	✔ Done!	PR482
Juniper	Junos OS	Planned	Second Half 2017
MikroTik	RouterOS	Won't Implement Until RFC	Feature Requested 2016090522001073
Nokia	SR OS	Planned	Third Quarter 2017
OpenBSD	OpenBGPD	✔ Done!	OpenBSD 6.1 (commit)
OSRG	GoBGP	✔ Done!	PR1094
rtbrick	Fullstack	Planned	December 2016
Quagga	Quagga	✔ Done!	Patch Provided for 1.1.0 875
VyOS	VyOS	Requested	Feature Requested T143

Tools / Ecosystem

Vendor	Software	Status	Details
DE-CIX	pbgpp	✔ Done!	PR16
FreeBSD	tcpdump	✔ Done!	PR213423
Marco d'Itri	zebra-dump-parser	✔ Done!	PR3
OpenBSD	tcpdump	✔ Done!	OpenBSD 6.1 (patch)
pmacct.net	pmacct	✔ Done!	PR61
RIPE NCC	bgpdump	Planned	issue 41
tcpdump.org	tcpdump	✔ Done!	PR543 (commit)
Yoshiyuki Yamauchi	mrtparse	✔ Done!	PR13
Wireshark	Dissector	✔ Done!	18172 (patch)

まとめ

- 世の中にはBGPコミュニティというものがある
- BGPコミュニティは経路制御に利用できる
  - 自分の経路を相手にどう見せるか
  - 相手から受信した経路を自分でどう扱うか
  - 最終的には、トラフィック制御を行うことが目的
- ユーザやピア向けに、BGPコミュニティを利用解放している事業者もある
  - 一部の国際・北米大規模事業者がメイン
  - 事業者側はニーズを見ながら設計・利用解放
  - 利用解放有無については各トランジット事業者に問い合わせを
- BGPコミュニティをめぐって2つの動きがある
  - Blackhole Community のRFC化および運用推進 (DDoS対応)
  - BGP Large Community のInternet Draft化 (32bit ASN対応)

November 29, 2016

QUESTIONS?