S8 サービスプロバイダ バックボーン設計入門 後編 BGP設計 後半

Norisuke Hirai
SoftBank Corp./BBIX, Inc.

Who am I?

平井 則輔（ひらいのりすけ）

SoftBank (2005～)

- 対外接続の最適化
- IPアドレス設計企画
- 固定BB技術企画
- インターネットサービスプロダクト開発
- BBIX出向

JANOG Committee (2013～)

- 会長(2017/9～)

好きなもの

- ビートルズとヘビメタが大好き！
- 最近はキングダムにハマってます！

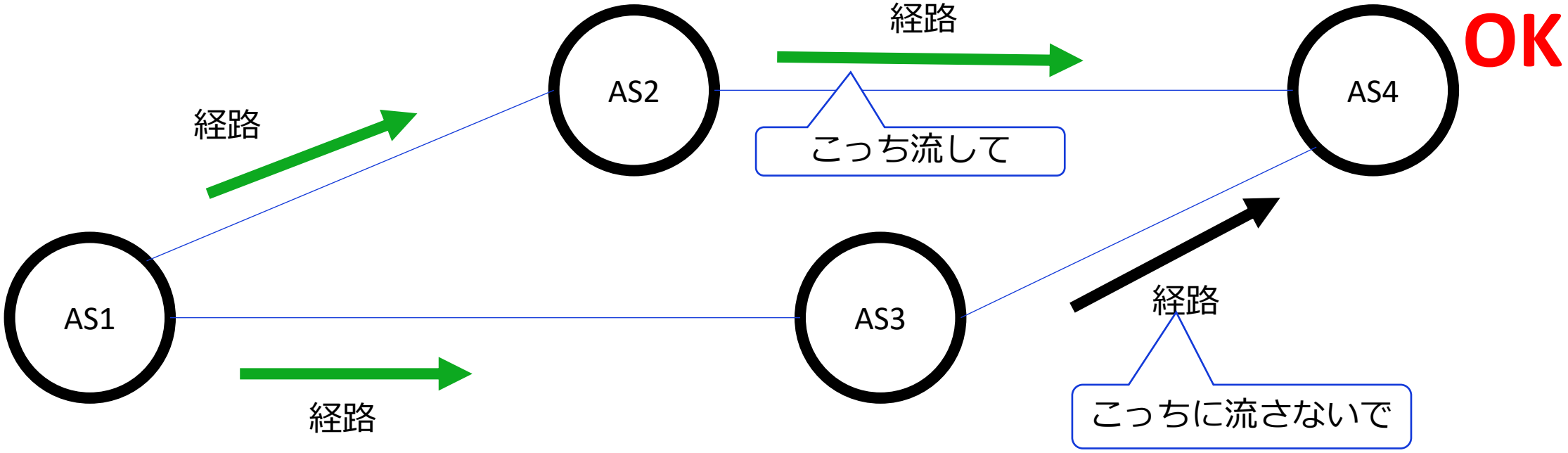
今日のお題

- eBGPの設計に関するお話
  eBGPの設計を決める際に考えないといけないこと

- eBGP設計の応用
- トラフィック制御ケーススタディ
  BGP Sessionの先にいるAS(Customer/Peering Partner/Transit Provider)それぞれに個別の Routing Policy がある中で、どのようにTraffic Controlを実現するか

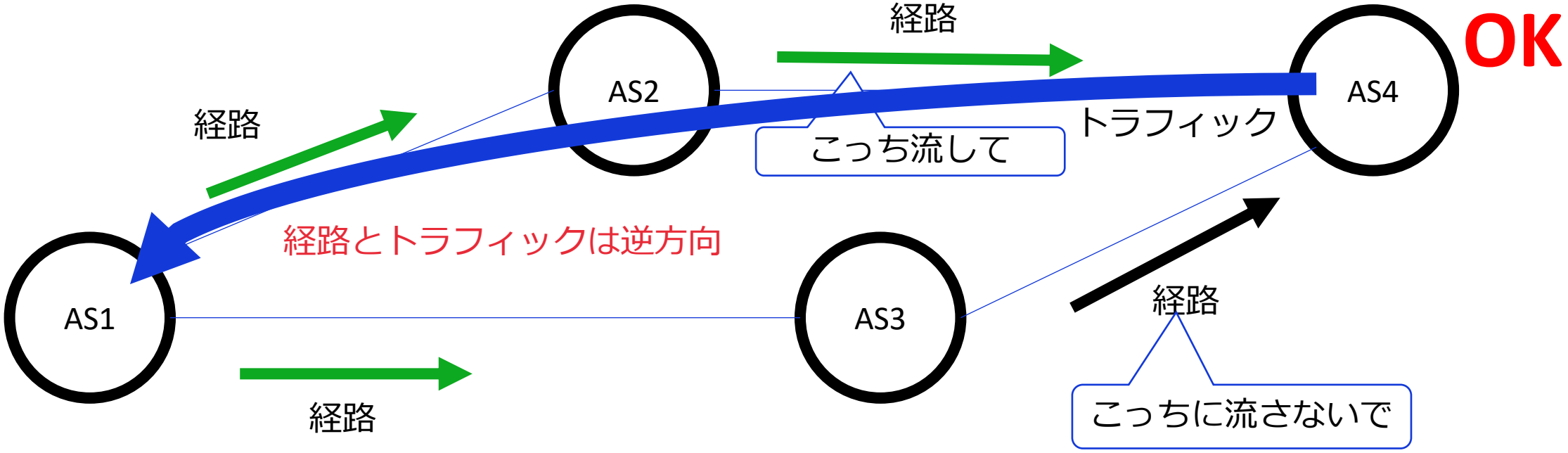
BGPとは

- ASが相互接続するのに使う約束事(プロトコル)EGP(External Gateway Protocol)の一つ。実際にはBGP(Border Gateway Protocol)が唯一のプロトコルとなってる。
- 異なるAS間でrouting情報を交換するプロトコル



BGPとは

- ASが相互接続するのに使う約束事(プロトコル)EGP(External Gateway Protocol)の一つ。実際にはBGP(Border Gateway Protocol)が唯一のプロトコルとなってる。
- 異なるAS間でrouting情報を交換するプロトコル



BGPの経路選択アルゴリズム

1. (最も高い WEIGHT を持つパスが優先されます。一部メーカーのみ)
2. 最も高い LOCAL_PREF を持つパスが優先されます
3. network または aggregate BGP サブコマンドによって, あるいは IGP からの再配布を通じて, ローカルで発信されたパスが優先されます
4. 最短の AS_PATH を持つパスが優先されます
5. 最小のオリジン タイプを持つパスが優先されます
6. 最小の Multi-Exit Discriminator (MED) を持つパスが優先されます

(MED は remote AS が同じ場合のみ評価される)

7. iBGP パスよりも eBGP パスの方が優先されます
8. BGP ネクストホップへの最小の IGP メトリックを持つパスが優先されます
9. 両方のパスが外部のときは, 先に受信したパス (最も古いパス) が優先されます
10. 最小のルータ ID を持つ BGP ルータから送られたルートが優先されます
11. 発信元 ID またはルータ ID が複数のパスで同じ場合は, 最小のクラスタリスト長を持つパスが優先されます
12. 最小の隣接ルータ アドレスから送られたパスが優先されます

BGPの設計を考える

BGPの設計とは？

- BGP Policyを設計すること
  - 経路広告をどのようにするか？
  - 経路受信をどのようにするか？

    を決めること

ところで、、、BGP Policy設計のまえに

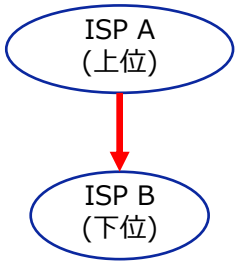
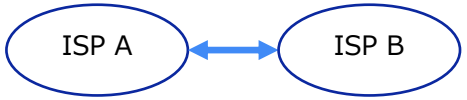
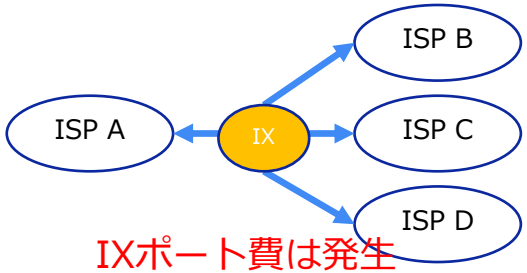
- 自社NetworkのPolicyってなんですか？
  - どんな業態のNetwork(何を実現したいのか?)
  - 可用性
  - 品質
  - 運用性
  - 拡張性
  - セキュリティ
  - コスト

明確に決まってなくても、
考えたり、チームで話し合ったりすることは非常に大事

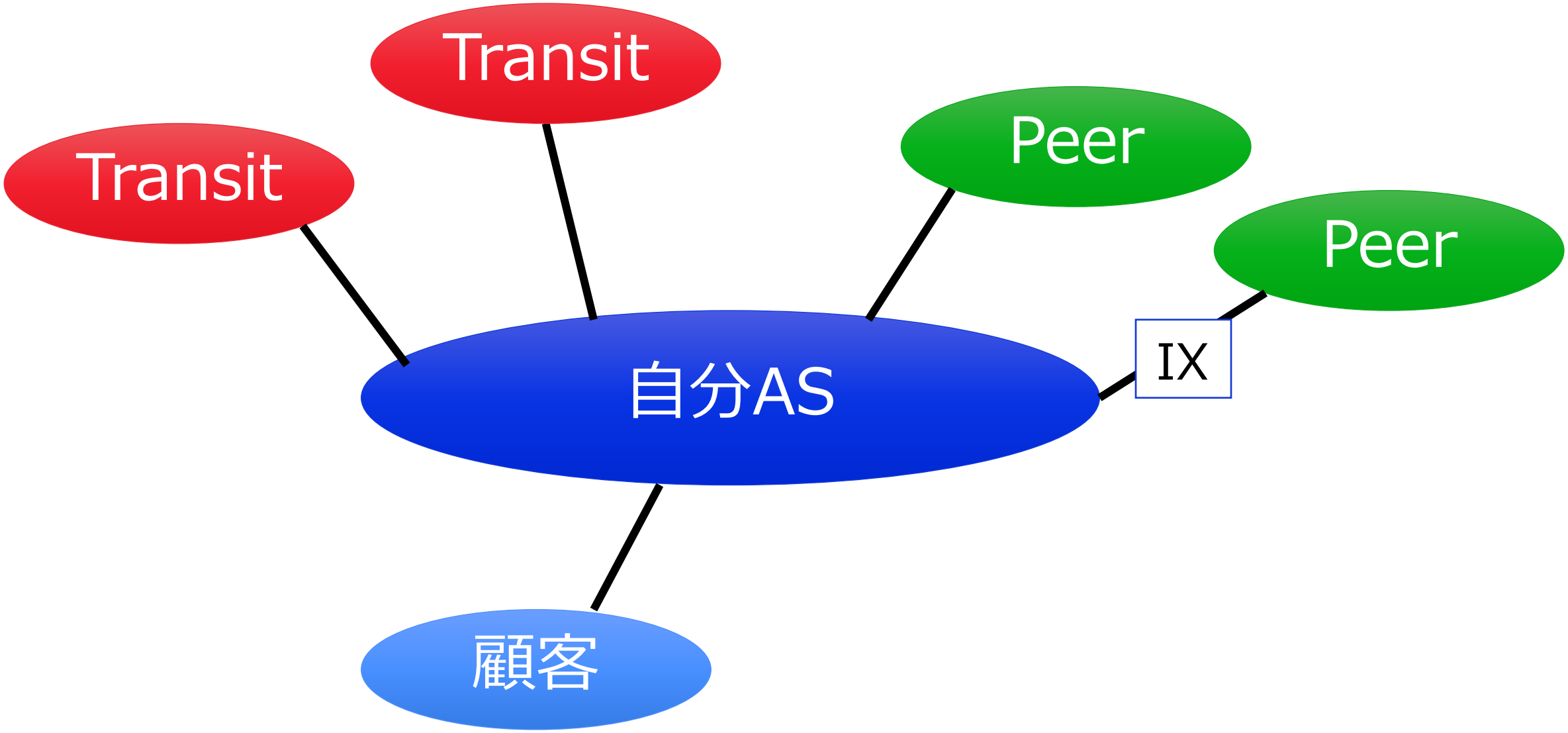
それでは、

- どんな経路を
- どんなASからどのように受信し
- どんなASにどのように広報するか

ISPの接続形態

接続形態		イメージ	関係	課金	経路交換	接続方法
Transit			上位ISPに依存	あり	上位ISPがFull Route（Internet全体へのReachability）を提供	1対1
Peer	Private		対等	なし	相互に自NWの経路と下位ISPの経路を交換	1対1
	Public	 IXポート費は発生	対等	なし	相互に自NWの経路と下位ISPの経路を交換	1 対N

たとえば、こんなTopology



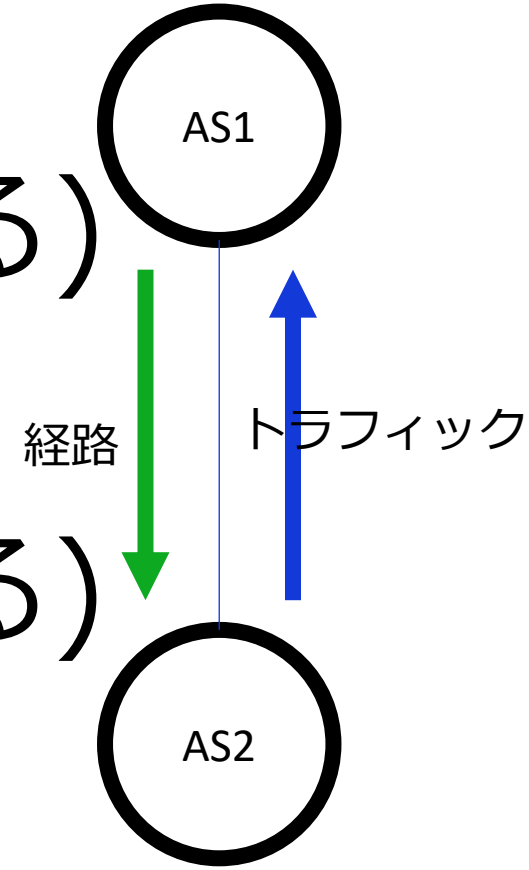
どんなASがいるか？

高

- 顧客AS
- Peer AS
  - Paid Peer(お金もらってる)
  - Private Peer
  - Public Peer (IX)
  - Paid Peer(お金はらってる)

低

- Transit事業者



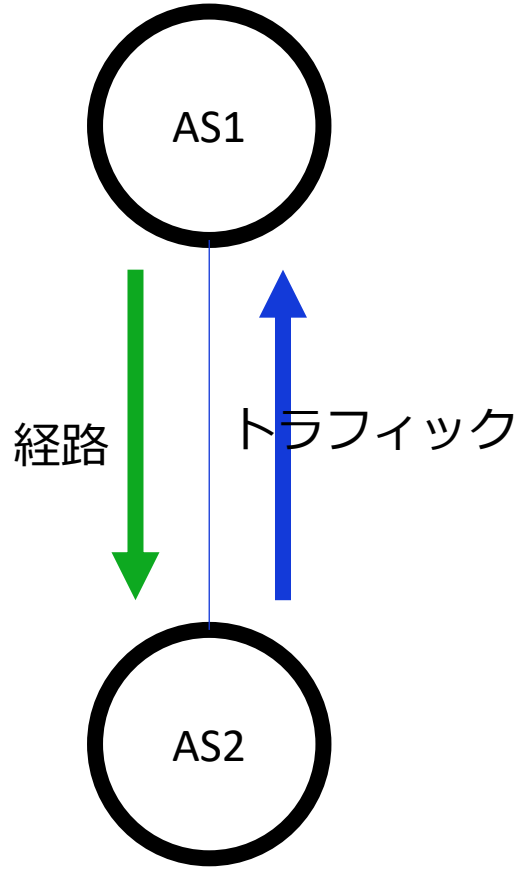
経路の種類

高

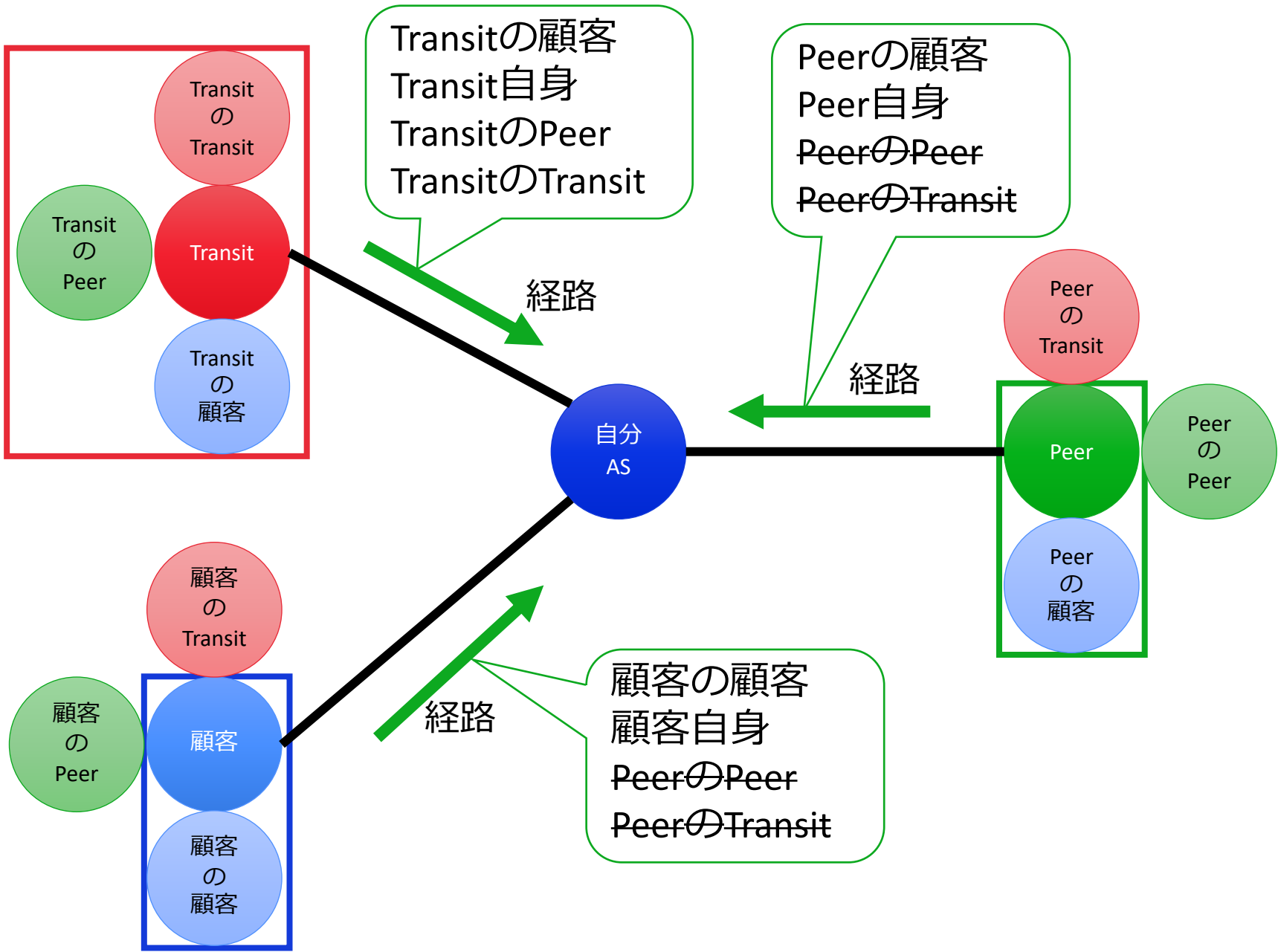
低

- 顧客の経路
- 自分の経路
- Peer先ASの経路
- Transit事業者の経路

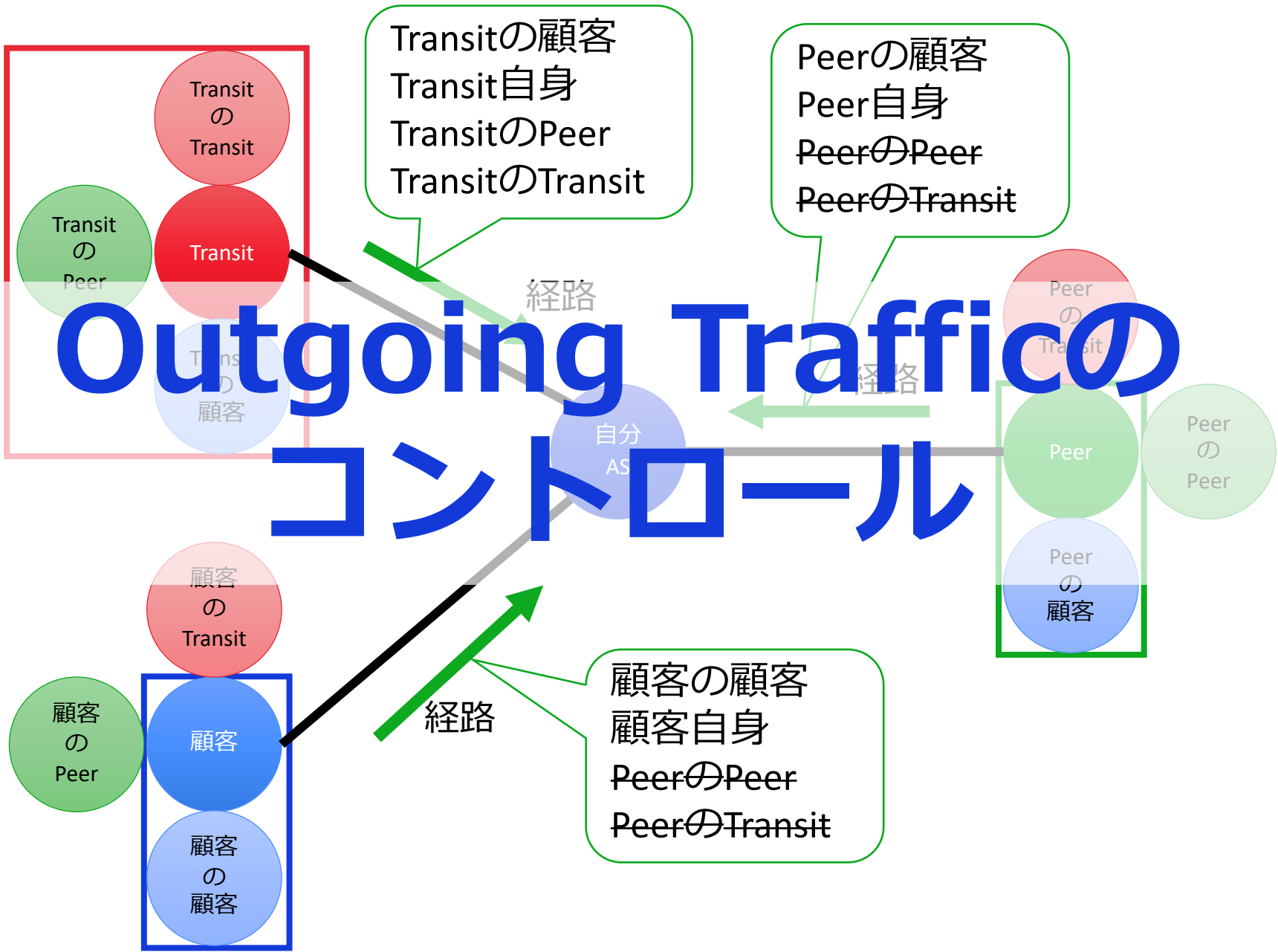
Full Route (Internet上のすべての経路)



どんな経路を受信するか



どんな経路を受信するか



Outgoing Trafficのコントロール

受信経路ポリシー (例)

優先度	経路種別	LP	MED
1	顧客	1200	
2	自AS	1100	-
3	Private Peer AS	1000	複数のBGP Sessionがある場合、100ごと
3	Public Peer AS	1000	複数のBGP Sessionがある場合、100ごと
4	Transit	900	複数のBGP Sessionがある場合、100ごと

経路フィルタ

経路種別	フィルタ	備考
顧客	Prefix Filter + AS-PATH Filter	Prefix Filter はExact Match、IRRベースで運用できるとベスト
Peer AS	Max-Prefix Filter IRRベースのPrefix Filter運用してるASあり	AS-PATH Update メールは嫌われる
Transit	(Max Prefix Filter)	全断の恐れがあるので慎重に

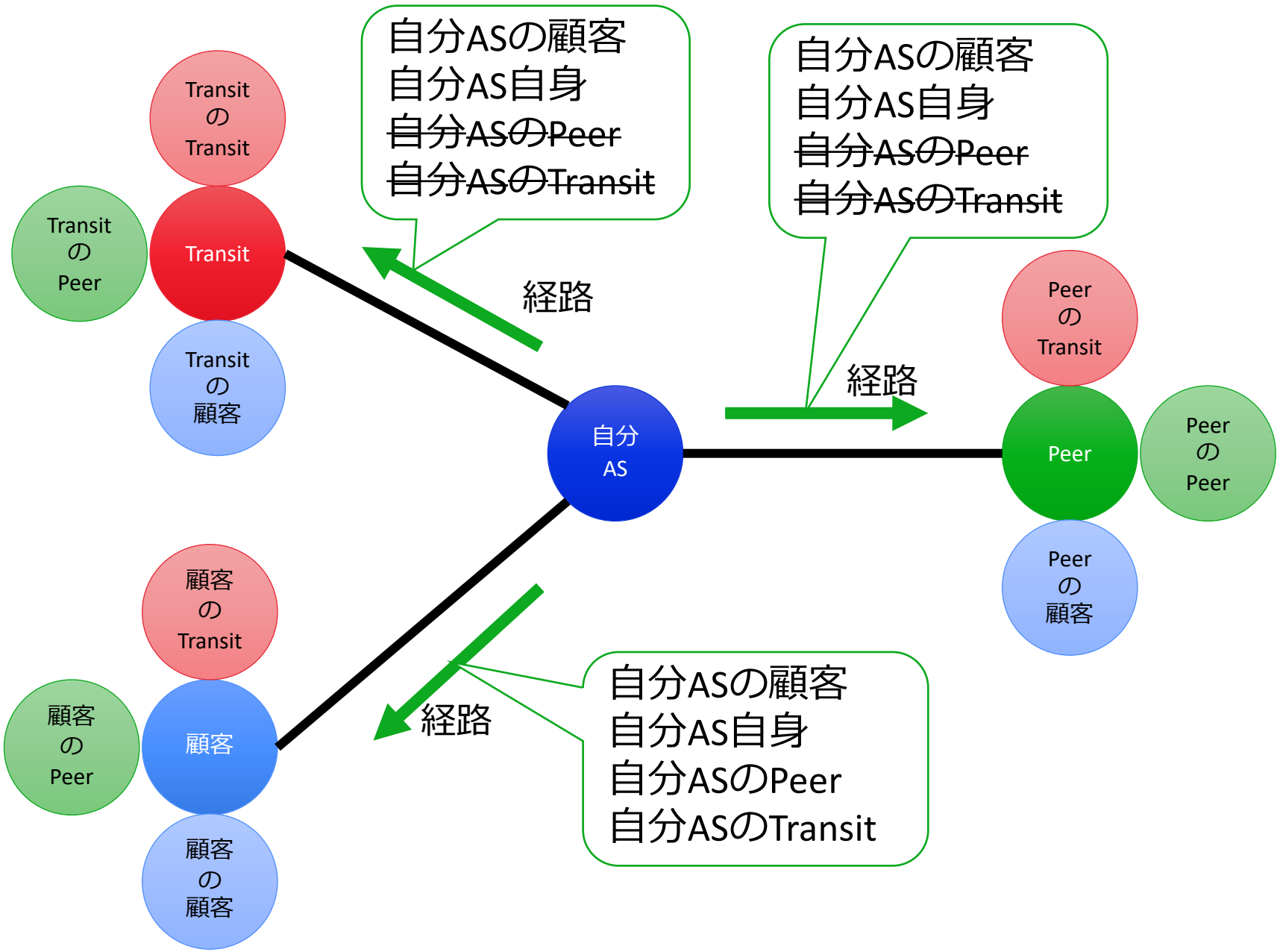
経路フィルタ

- JANOG Comment JC1000～JC1006に良くまとまってるので、ご一読を！！

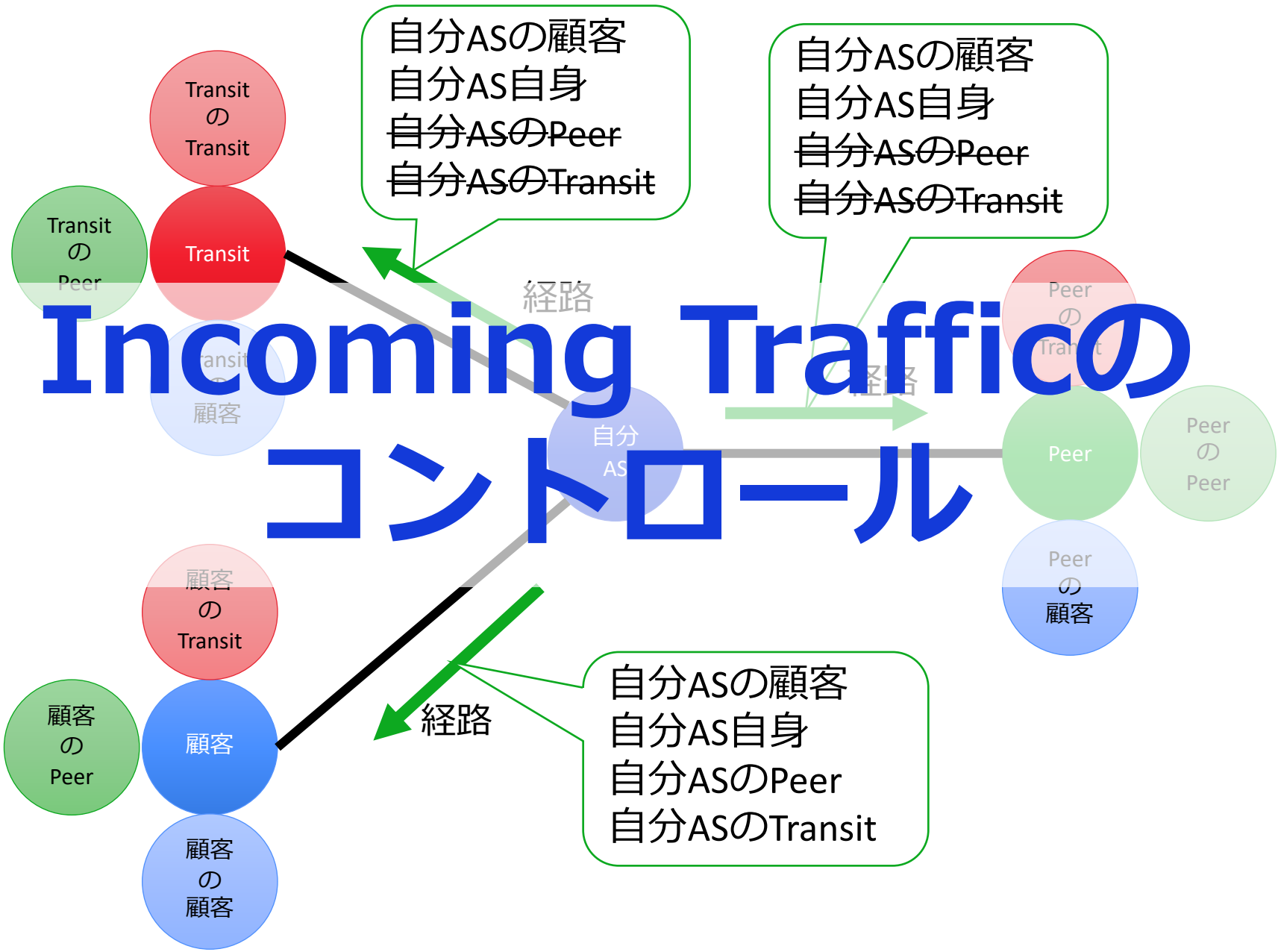
https://www.janog.gr.jp/doc/janog-comment/

JC1000: xSP のルータにおいて設定を推奨するフィルタの項目について
JC1001: ～ トランジット接続部分編
JC1002: ～ ピア接続部分編
JC1003: ～ 顧客接続部分編
JC1004: ～ ルータ自身へのアクセス編
JC1005: IXに接続する際の標準的な設定について
JC1006: xSP のルータにおいて設定を推奨するフィルタの項目について (IPv6版)

どんな経路を広報するか



どんな経路を広報するか



Incoming Trafficのコントロール

どんな経路を広報するか

Incoming Trafficのコントロール

ISPにとっては、とても重要なのに、
意外と手がない。。。

経路広告の制御選択肢

- Peer/Transit向け
  - AS-PATH Prepend
  - MED 調整
  - BGP Community（一部Transit Provider）
    - Transit AS内でのLPを制御
    - Transit ASから他ASへの経路広告を制御 (広報止める/Prepend)
  - 経路を止める
  - (経路を分割して広報)

経路を分割して広報するならno-export communityをつけるなどのエチケットが必要

eBGP設計の応用(Communityの活用)

BGP Community

- 経路につけるTag
- 2byte：2byte という表記
- あってもなくてもOK、ASまたいで伝搬する
- Well-known Community (定義済みのCommunity)がある
  - NO_EXPORT, NO_ADVERTISE など

たとえば、一般的にはこんな感じ

17676**:2080**

自AS番号
(2byte)

なんでも良いよ
(2byte)

※4byte AS向けにBGP Large CommunityやExtended Communityがある

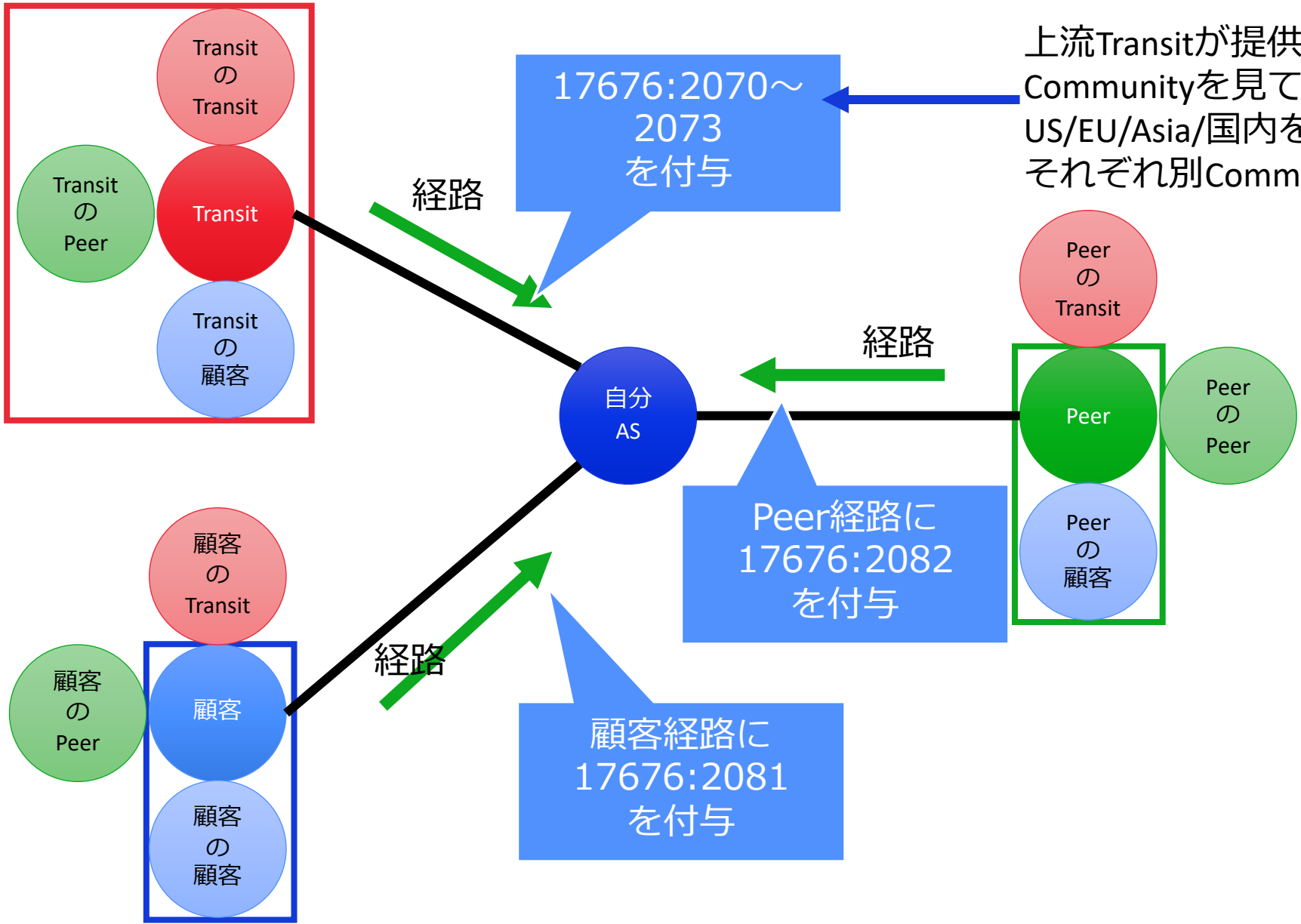
BGP Community

- 経路のカラーリング
  - Transit Provider、Peer、顧客から受信した経路それぞれにタグをつける
- Community付与された経路に対し、制御を行う
  - 特定コミュニティが付与されている経路を受信したら、Local Preference増減させる
  - 特定コミュニティが付与されている経路を受信したら、他AS向けにPrependを付与して広報する
  - 特定コミュニティが付与されている経路を受信したら、その経路向けパケットをBlackholeさせる

など、いろいろ活用されてます。

https://onestep.net/communities/
ここのサイトにTransit ProviderのCommunity制御がまとまってるので参考になります

経路のカラーリング(例)


Transit
の
Transit
Transit
の
Peer
Transit
Transit
の
顧客
経路
17676:2070～
2073
を付与
上流Transitが提供している
Communityを見て
US/EU/Asia/国内を判別して、
それぞれ別Community付与
自分
AS
経路
Peer
の
Transit
Peer
Peer
の
Peer
Peer
の
顧客
Peer経路に
17676:2082
を付与
顧客
の
Transit
顧客
の
Peer
顧客
顧客
の
顧客
経路
顧客経路に
17676:2081
を付与

経路のカラーリングを顧客に提供

経路に付与されたCommunityを参照して、
お客様のトラフィック制御に活用してもらう

コミュニティ	意味
17676:2070	US経路
17676:2071	Europe経路
17676:2072	Asia経路
17676:2073	国内経路
17676:2080	ULTINA Internet(AS17676) Originの経路
17676:2081	ULTINA Internet(AS17676) BGPユーザの経路
17676:2082	ULTINA Internet(AS17676) ピアの経路

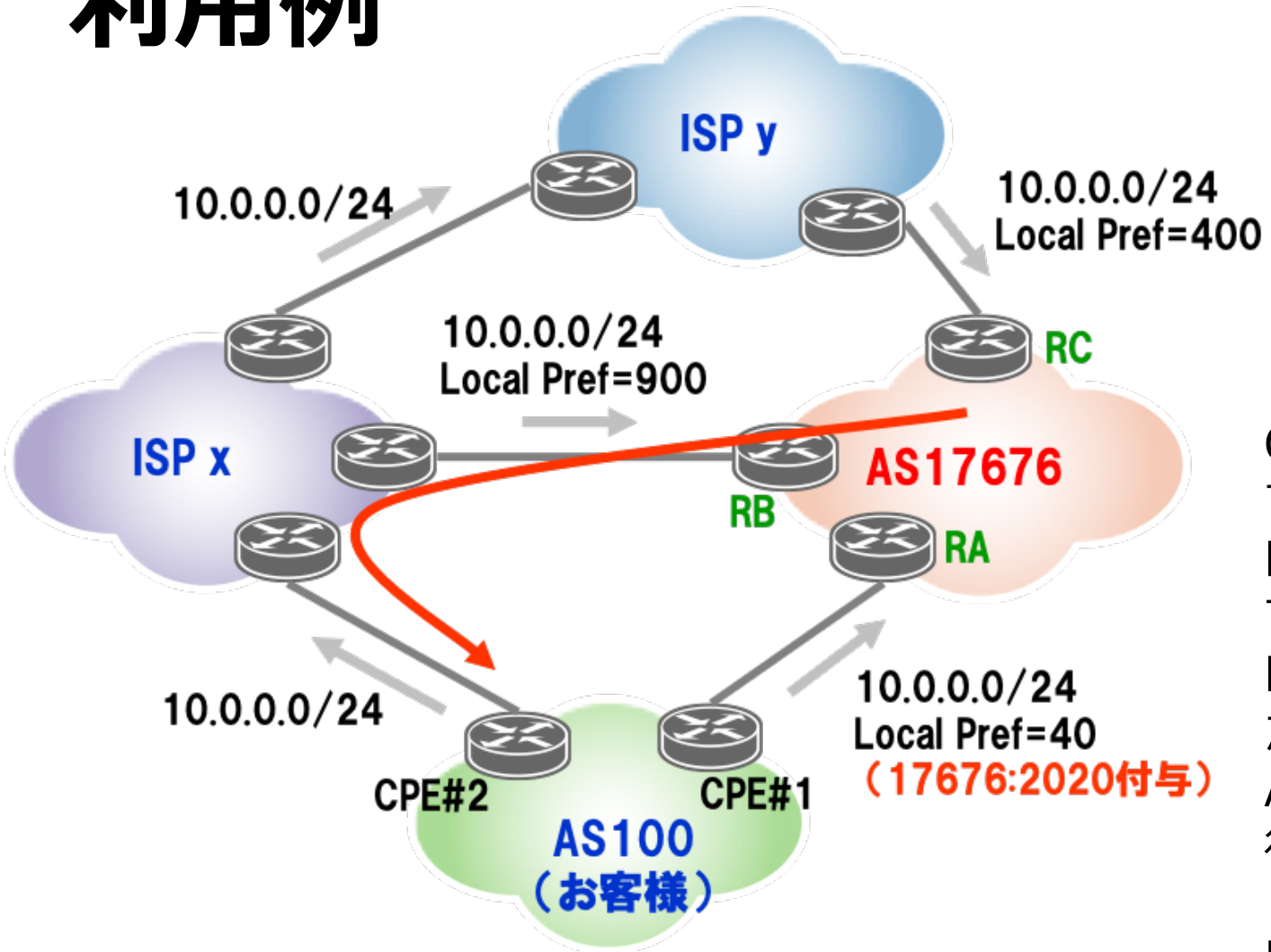
たとえば、US向け経路をLocal Preferenceを上下させて優先制御したりする

Community付与された経路に制御を行う

特定コミュニティが付与されている経路を受信したら、
Local Preference増減させる

コミュニティ	意味
なし	お客様が広報した経路に関してAS17676網内でのLocal_Prf値をデフォルト値とする
17676:2010	お客様が広報した経路に関してAS17676網内でのLocal_Prf値をAS17676 BGPユーザ・ピアよりは弱いがトランジットより強い値とする
17676:2020	お客様が広報した経路に関してAS17676網内でのLocal_Prf値をAS17676 BGPユーザ・ピア・トランジットより弱い値とする

利用例



CPE#1からRAに対してコミュニティ付与して広報した10.0.0.0/24に対し、LP(Local Preference)の値は40と設定されます。RBでは同じ経路に対しLP=900、RCではLP=400としているため、AS17676からみた10.0.0.0/24宛のベストパスはRBとなり、AS17676→ISP x経由でパケットの送信が行われることになります。
（※LocalPreferenceより優先度の高いBGP属性は同一と仮定しています）

Community付与された経路に制御を行う

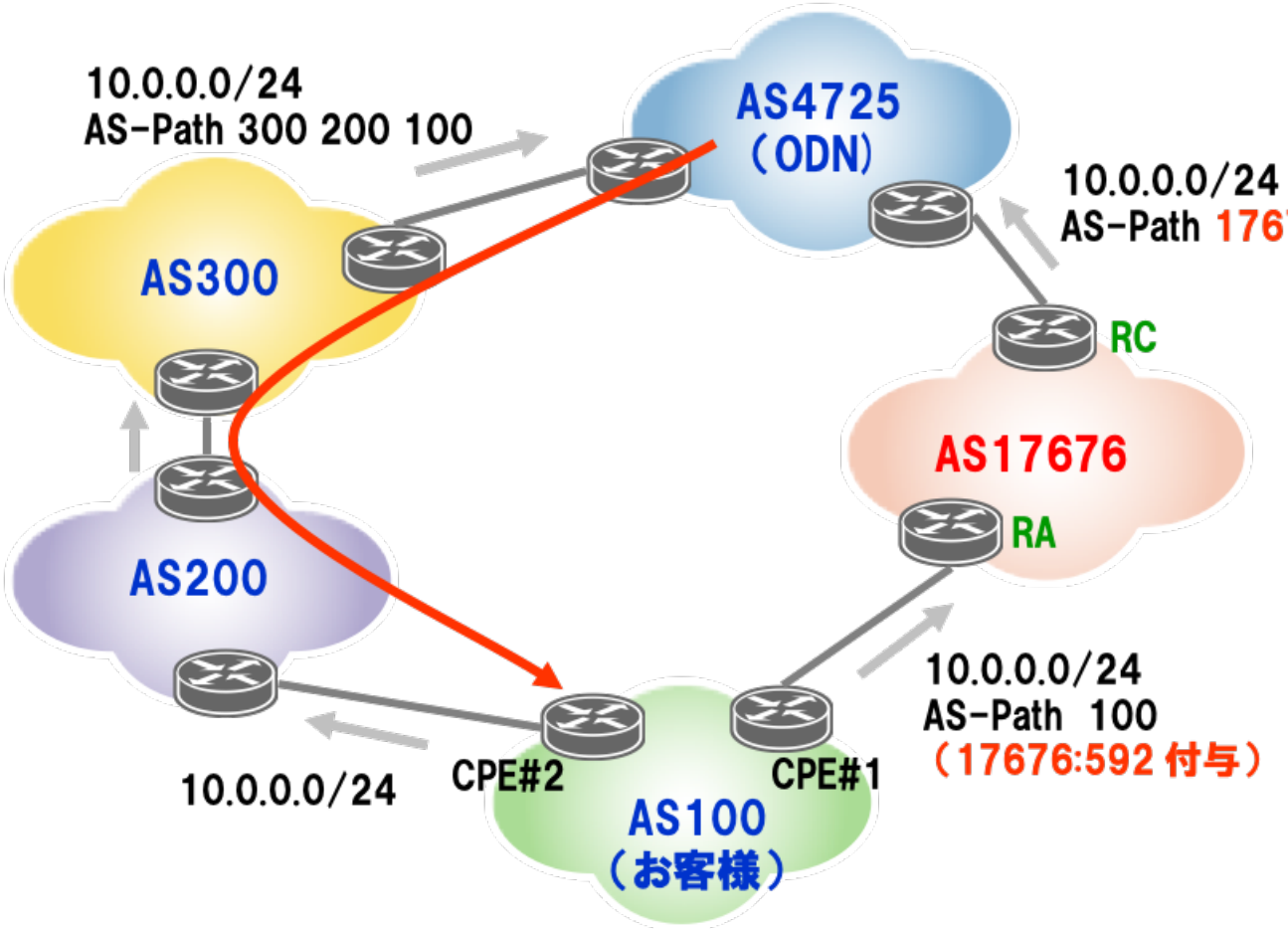
特定コミュニティが付与されている経路を受信したら、
他AS向けにPrependを付与して広報する

コミュニティ	意味
17676:XX0	特定ISPに対し広報禁止
17676:XX1	特定ISPに対しprependを1つ施して広報
17676:XX2	特定ISPに対しprependを2つ施して広報
17676:XX3	特定ISPに対しprependを3つ施して広報

■　XXと特定ISPの対応表
51　・・・　全てのPeer (Transit/Custは含まれておりません)
55　・・・　ntt.net　<AS2914>
57　・・・　KDDI　<AS2516>
59　・・・　ODN　<AS4725>
65　・・・　OCN　<AS4713>
66　・・・　IIJ　　<AS2497>

※　Prepend先のISPは将来変更になる可能性があります

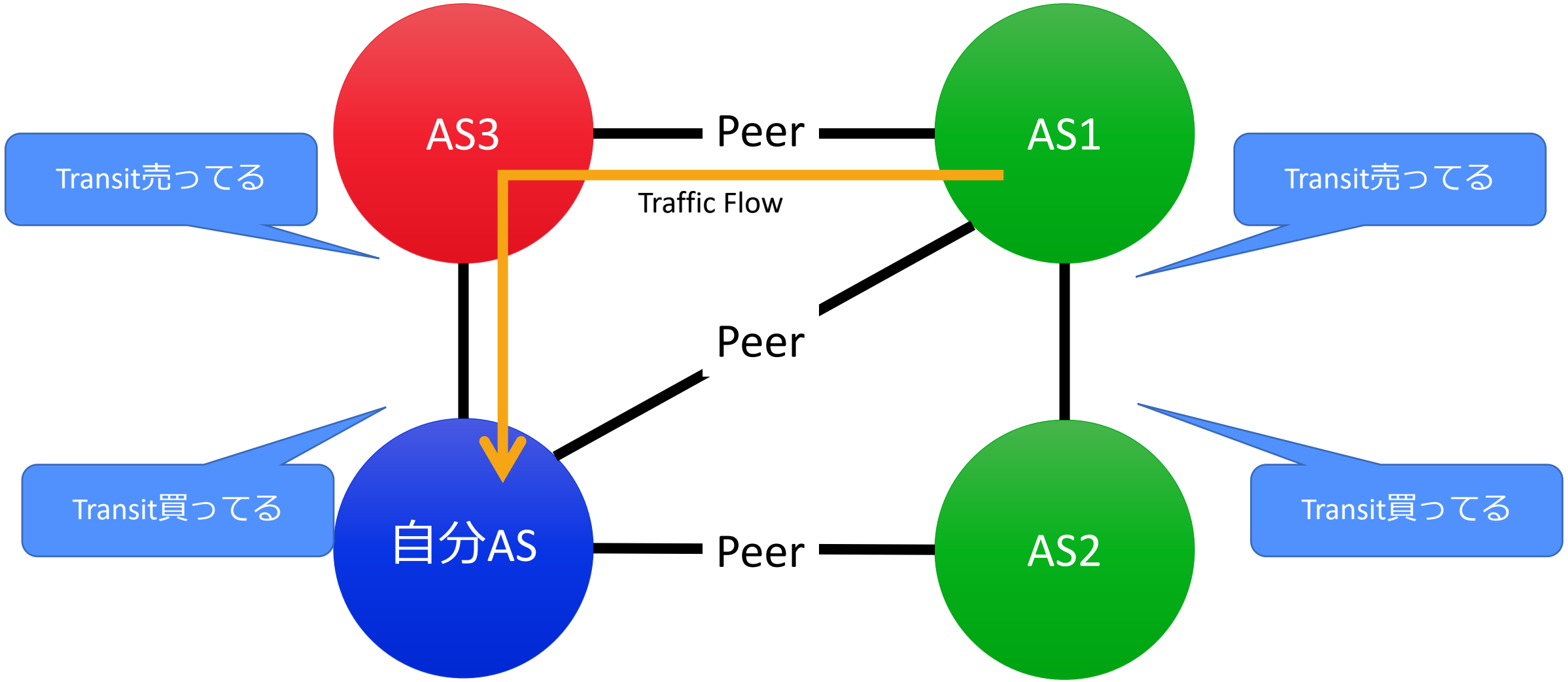
Community付与された経路に制御を行う



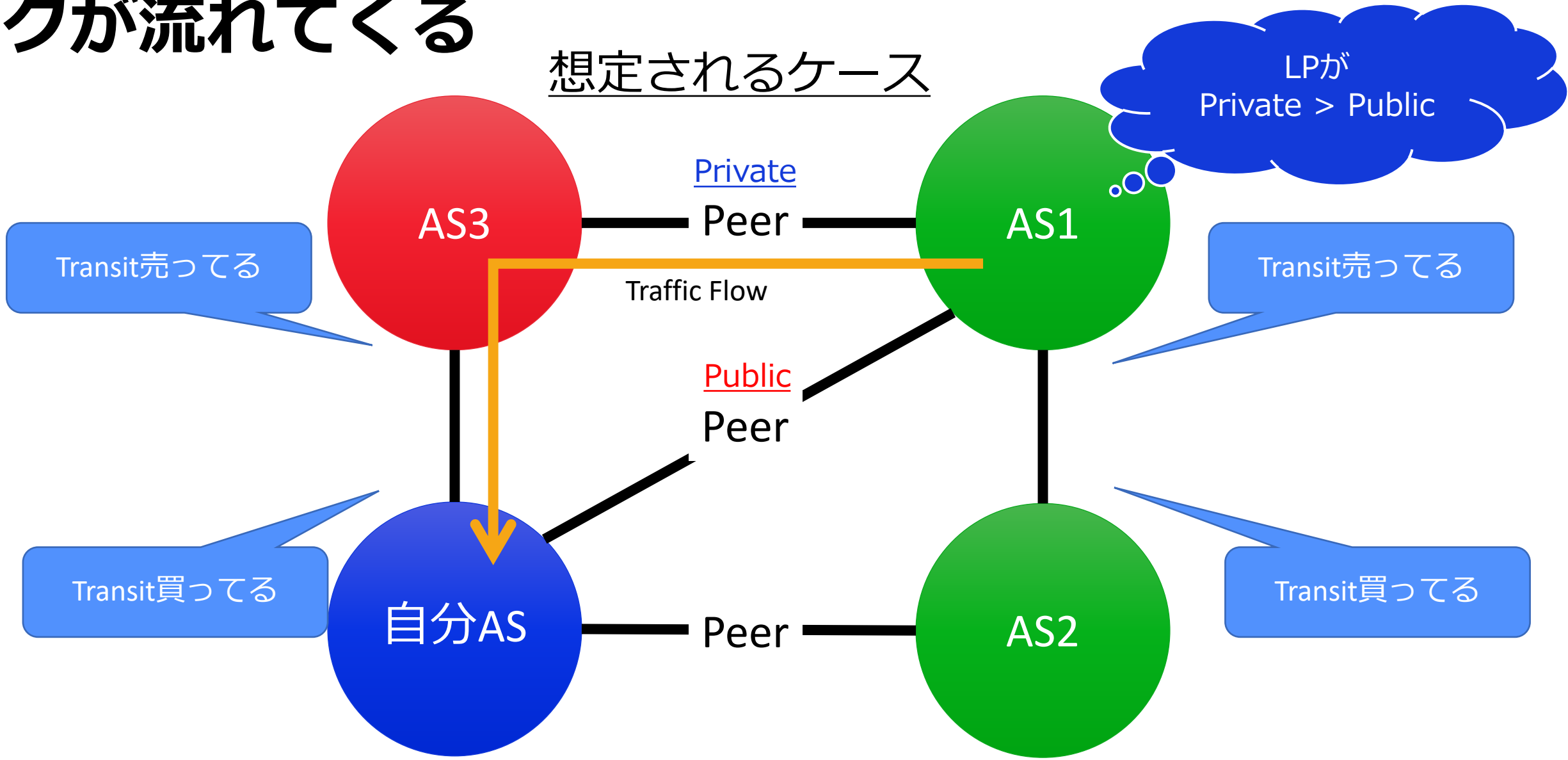
CPE#1からRAにコミュニティ付与して広報した10.0.0.0/24に対し、特定ISP（AS4725）へ広報する際にAS17676を2つ付加します。
AS4725からみた10.0.0.0/24宛のベストパスは経由するASの数が少ないAS300→AS200経由となります。
（※AS-Pathより優先度の高いBGP属性は同一と仮定しています）

トラフィック制御を考える

問題1: PeerしてるのにTransitからトラフィックが流れてくる



問題1: Peerしてるのに Transitからトラフィックが流れてくる

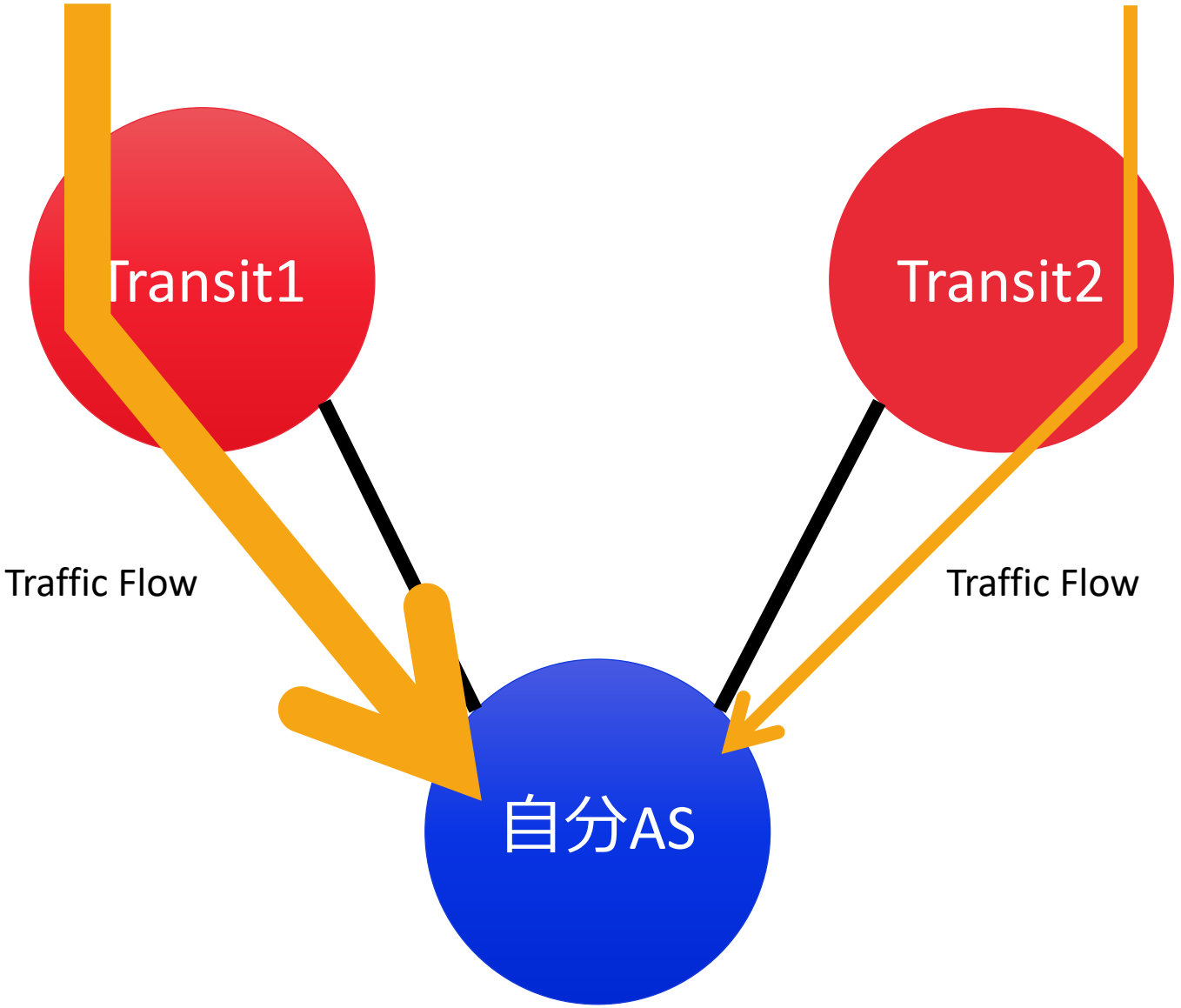


問題1: PeerしてるのにTransitからトラフィックが流れてくる

- 解決案
  - AS1にメールする
  - AS3にメールする
  - AS3(Transit)の提供するCommunityを使って、AS3からAS1向けに経路広報を止める(あれば)
  - IX事業者経由でAS1にメールする
  - AS1向けに経路を分割して広報(no-exportつける)

※ Trafficコントロールポリシーが公開されているPeerもいるので、事情は理解する
※ 直接Peerしてるのに、IX事業者/Transit経由で問い合わせされると、好感はもたれない
(どことPeerしてるのかは、あまり大っぴらに話されると困ることもある)

問題2: Transit間でTrafficを動かしたい

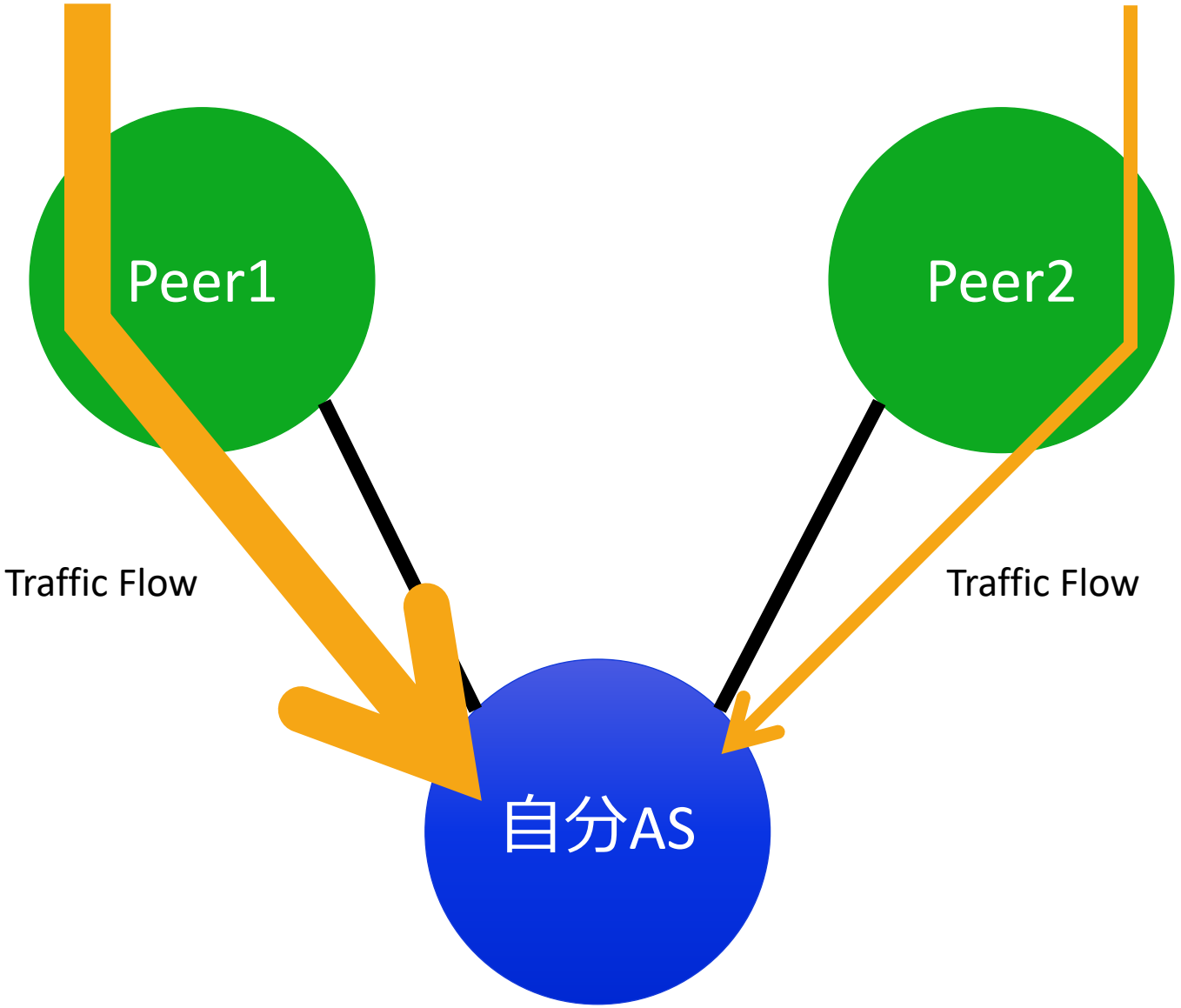


問題2: Transit間でTrafficを動かしたい

- 解決案
  - Trafficの偏りの原因となってるASとPeerする!!!
  - Transit2に偏りの原因となってるASとPeerしてもらう(時間はかかるかもしれない)
  - Transit1へのAS-PATH Prepend実施
    - AS-PATH Prepend+3してだめだったら、それ以上やっても無駄(経験的に)
  - Transit1の提供するCommunityを使って、偏りの原因となってるAS向けに経路広報を止める(あれば)
  - Transit2に経路分割して広報(まあまあ乱暴)

優先順位

問題3: Peer間でTrafficを動かしたい



問題3: Peer間でTrafficを動かしたい

- 解決案

優先順位

- Trafficの偏りの原因となってるASとPeerする!!!
- 偏りの原因となってるASにPeer2向けに流してもらうようメール/話をする
- Peer1へのAS-PATH Prepend実施
  - AS-PATH Prepend+3してだめだったら、それ以上やっても無駄(経験的に)
- Peer2に経路分割して広報(まあまあ乱暴)

本日のまとめ

- 自分たちのNetwork Policyを考えて、意識しましょう
  (そのために、Policyはシンプルに!!)

- 困ってること/疑問は直接Peer先に聞いてみましょう

- ということで、生野さんPeering解説よろしく！