今朝HDDが飛んで資料も消えちゃった。
でも事前に担当PCに送っておいたから
なんとか発表までに復旧できました版

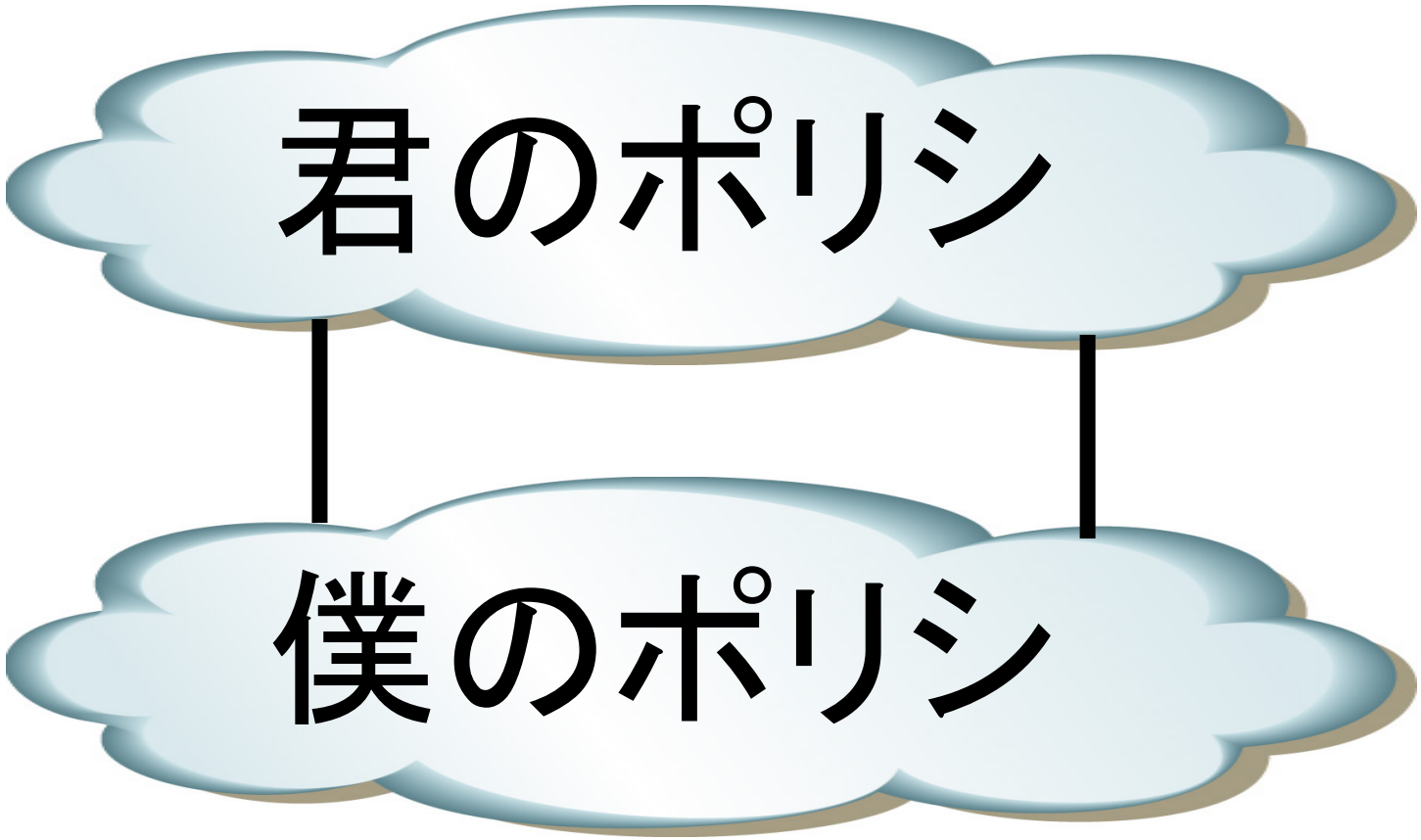
BGP運用
your policy vs my policy

Matsuzaki ‘maz’ Yoshinobu
<maz@iij.ad.jp>

全体のお話のサマリ

- ほころびの始まり
  - その時は、それでもいけると思った
  - 営業が‘無いサービス’売っちゃった
- スパイラル
  - よりディープな世界へ
  - 使える制御を使い切っちゃう
  - 矛盾を解決できなくなる・・・運用でカバー？
- 素直なポリシは身を助ける

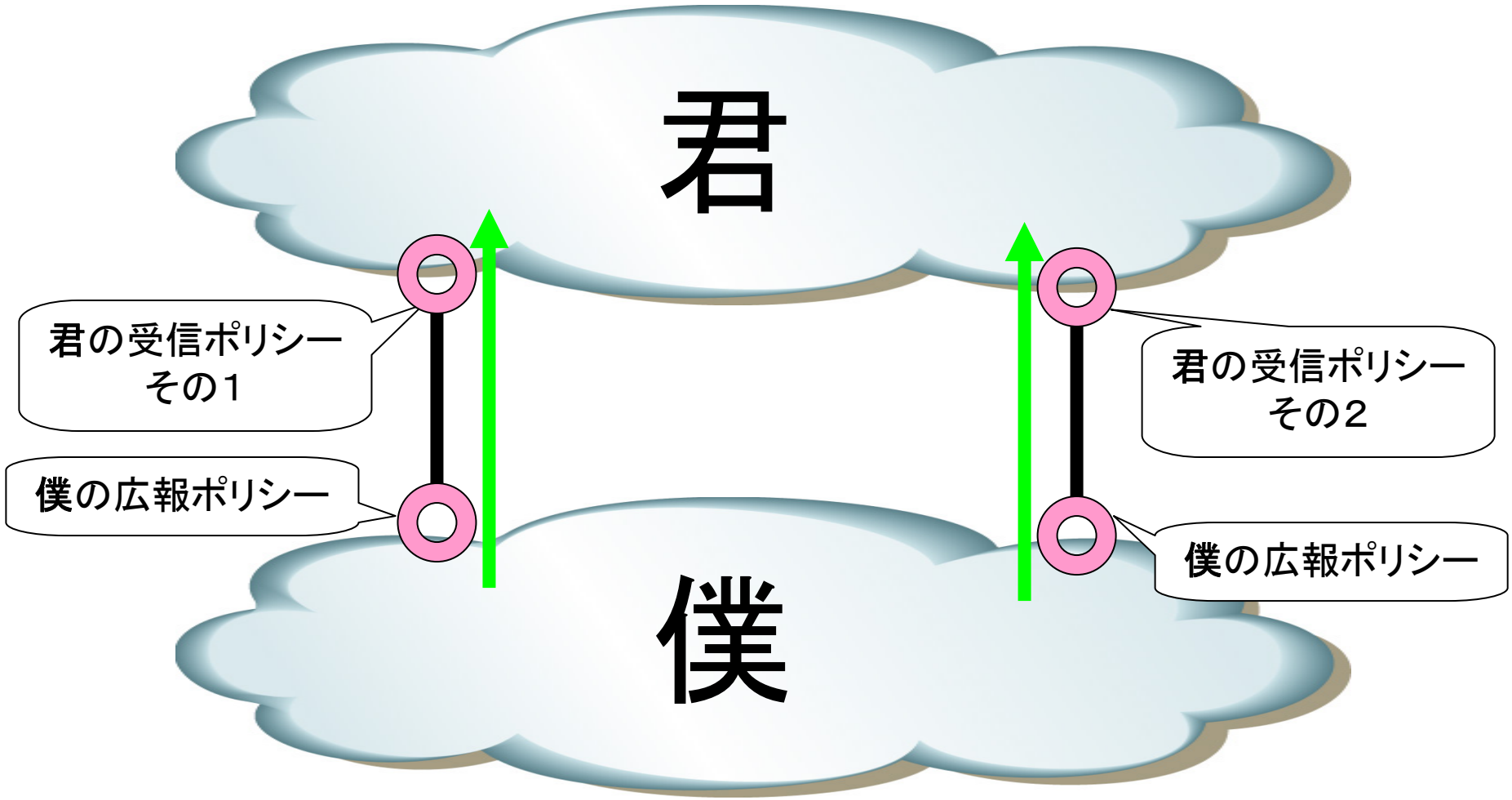
ポリシと相互接続



いろいろな前提

- BGPの仕様
  - 最適なprefixのみを広報
  - 経路広報時、受信時にそれぞれポリシを適用
- 受信した経路の優先度
  - 細かい経路が優先
  - 顧客経路　＞　ピア経路　＞＝　上流経路
- 広報する経路
  - ピア、上流には自身＋顧客経路を広報
  - 顧客にはfull-routeを広報

ポリシと実装



AS間の経路制御で使える手段

- ネットワーク構成
  - 物理構成とか論理構成で制御
  - closest exitとか
- BGP的な手法で制御
  - 経路フィルタ
  - prefix分割 ☹
  - パス属性でいろいろ制御
    - 優先度
    - bgp community

僕のポリシと君達のポリシ

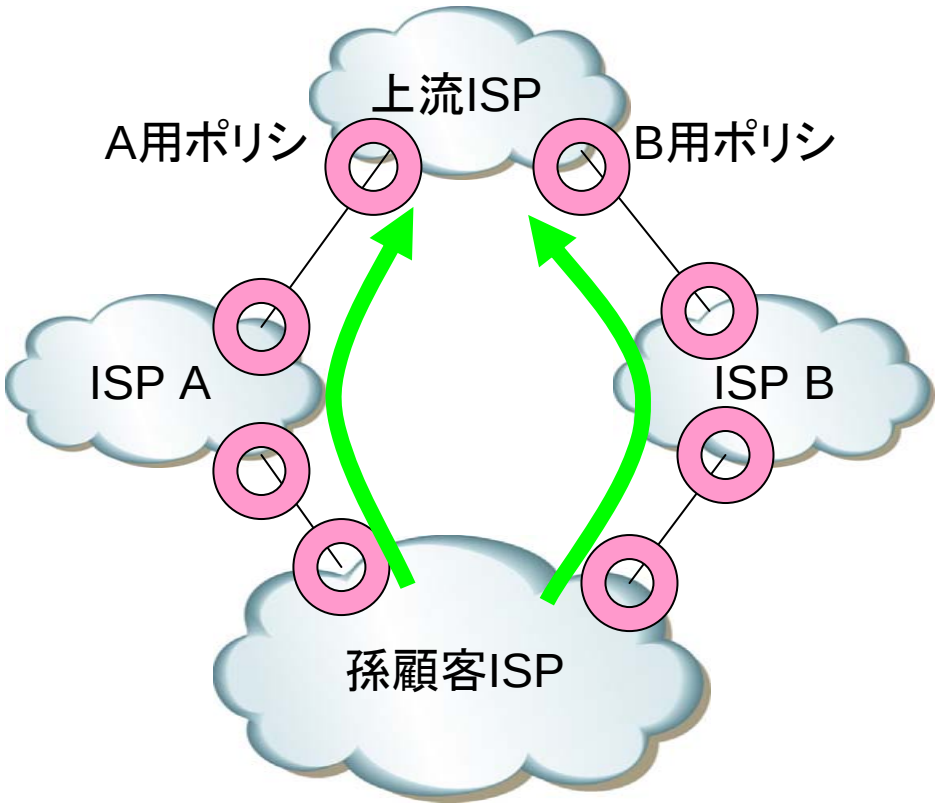
- インターネットはfull-meshじゃない
  - トランジットASの存在
  - 階層構造の存在

- 直接接続していないネットワークへは他のネットワークを経由する
  - ポリシが多段に適用される！

階層構造と経路制御

- ポリシが階層通りに適用される
  - 経路広報時、経路受信時
- 受信経路へのポリシ適用は強力
  - localprefはas path長(prepend)より強い
  - 顧客へのポリシは、孫顧客にも同様に適用
  - ピアへのポリシは、ピアの顧客やその孫顧客にも同様に適用

ポリシの階層構造



- 広報する経路は、いろいろな接続ポリシを経て他のASに届く
- 上流ISPで孫顧客の経路にどのようなポリシが適用されるかは、ISP A/Bがどの様な接続契約を行っているかに依存する

無理なポリシは不幸を招く

- 営業が売っちゃった
  - 実現するには無理なポリシ/設定が必要
  - それでも実装しなきゃいけない
- 最初はそれでいけると思ってた
  - 頭の中ではこれでOKのはず

- ちょっとした不幸から始まるスパイラル

売った側の不幸

- 無理なポリシの実装に苦しむ
  - トラブルの可能性
  - バグ、運用負荷、ミス
- 異常事態の発生
  - そんな挙動になるなんて思ってもみなかった
- 他の顧客やポリシとのすり合わせ
  - 次々に現れる矛盾とその解決
  - 解決できない問題も・・・

買った側の不幸

- インターネットの変動に苦しむ
  - 色々試すが思った通りにならない
  - 各地で記録されて駄目ASとして有名になる
- トラブルに巻き込まれる
  - ちゃんと運用されてると思ってた
  - 特殊事例は緊急時や移行時に忘れられる

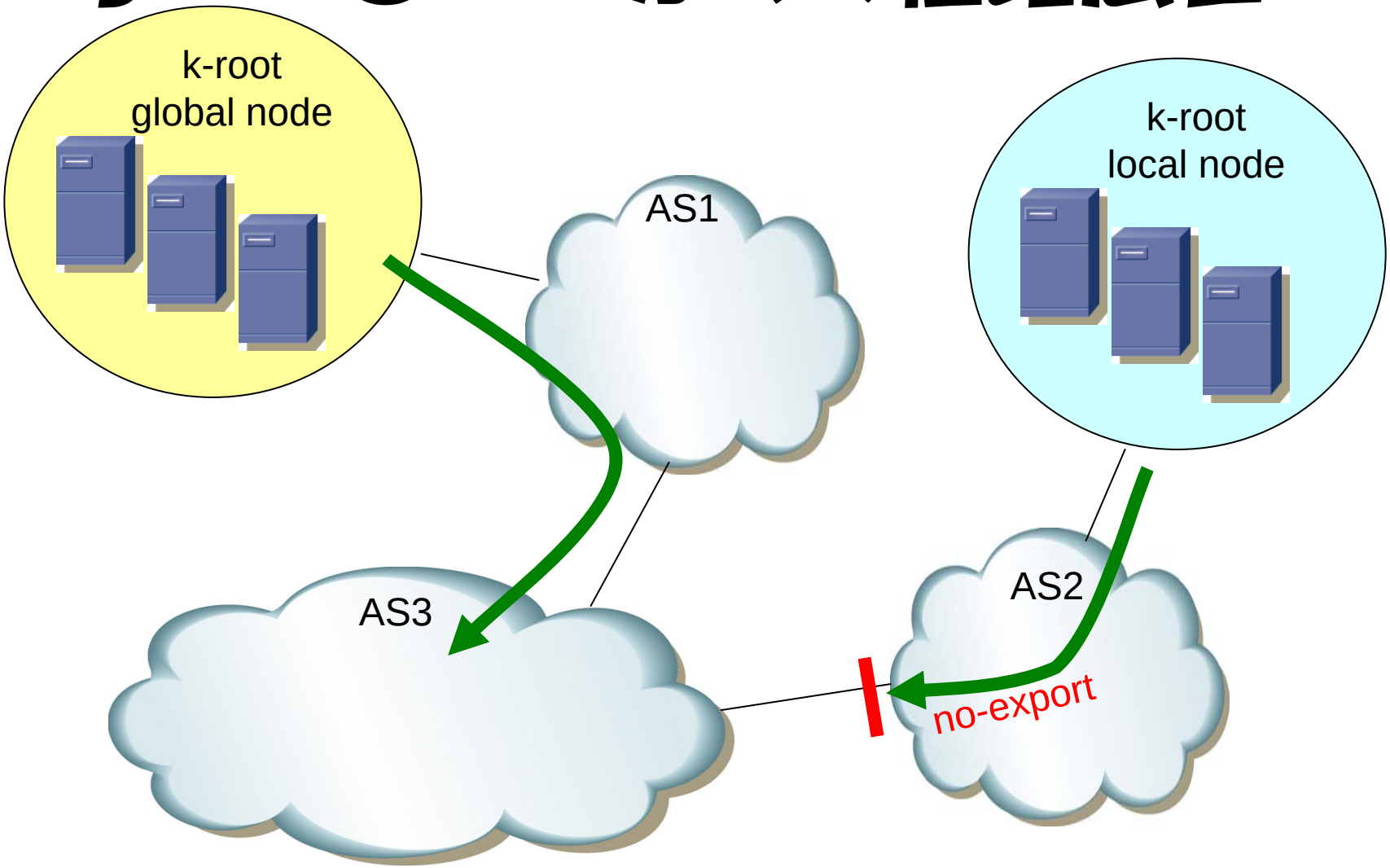
不幸な事例

- k-root
- パーシャルトランジット

例えばk-root

- global nodeとlocal nodeでBGP anycast
- global node
  - 世界から参照される
  - prependして非優先
- local node
  - 限定された地域から参照される
  - prependしないので優先される
  - NO-EXPORTで経路の流通を制限

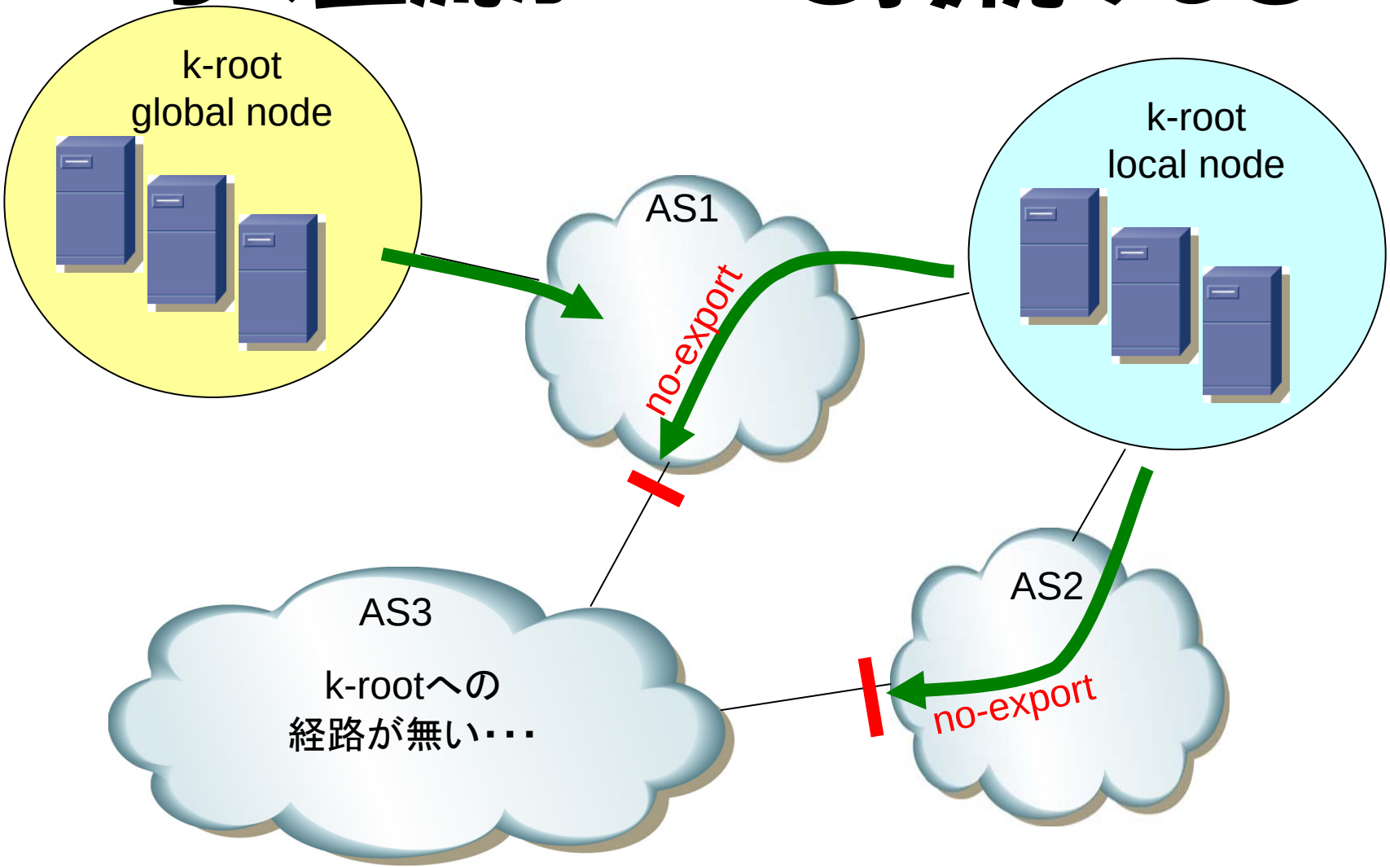
globalとlocalからの経路広告



そして問題発生

- k-rootに到達できない
- Randy Bush – **oh k can you see**
  - http://www.merit.edu/mail.archives/nanog/2005-10/msg01226.html

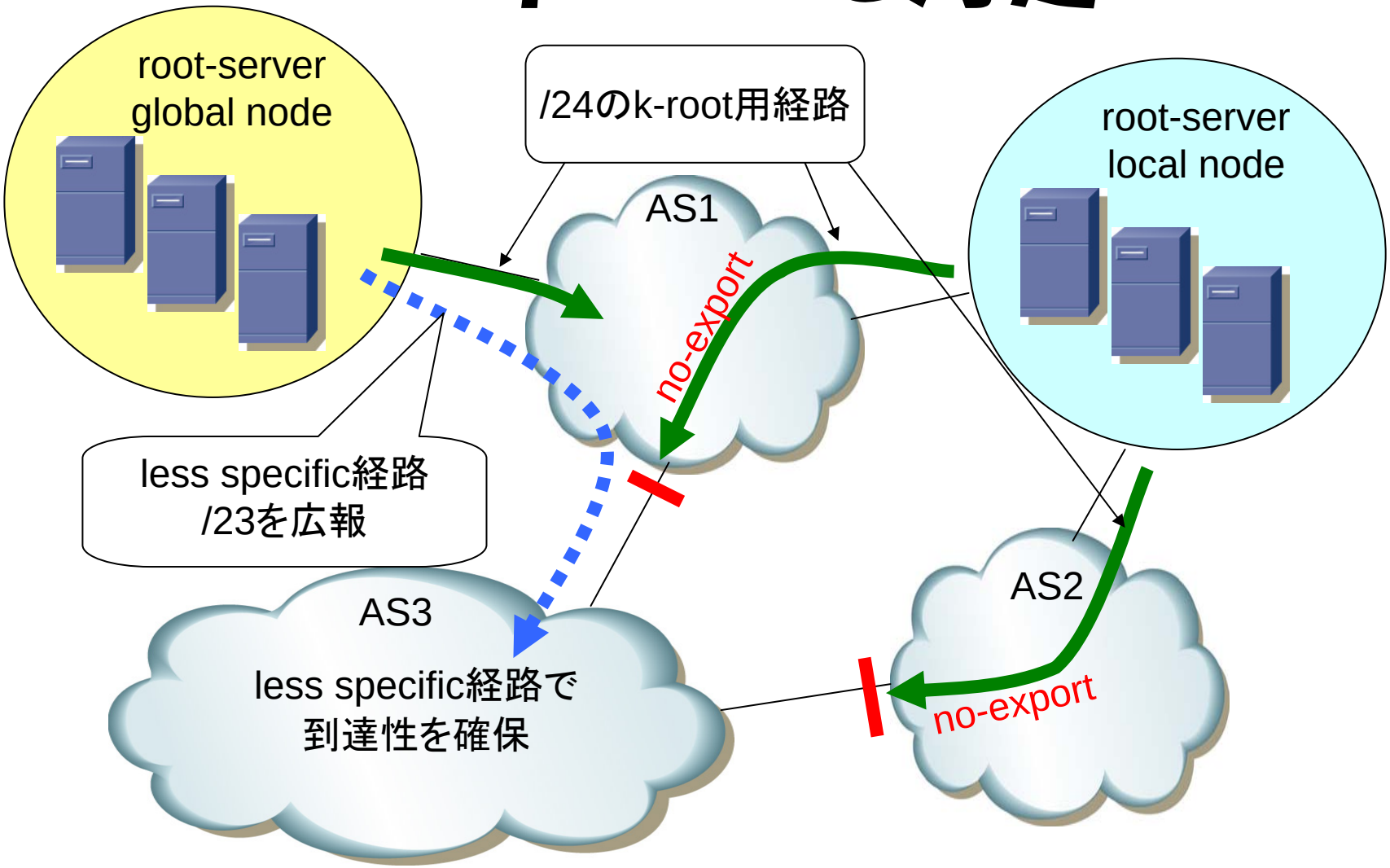
AS3の上流がlocalと接続すると



k-rootの解決策

- globalなless specific経路を広報
  - NO-EXPORTを付けない
  - 途中でlocal nodeがあれば、more specific経路が見えるはずなので、そちらに吸われるはず
- RIPE393 – **Evaluating the Effects of Anycast on DNS Root Nameservers**
  - http://www.ripe.net/ripe/docs/ripe-393.html#conn-loss

less specificで対処



と書いたものの

- ふと思って確認してみると、無い
  - less specificなcovering prefix
  - 昨年withdrawされてそのまま
  - 他のASも持ってない。あれれ？

- ポリシの変更かな？
  - ripe-393の発行は10月
  - withdrawはそれよりも前・・・おかしいね

RIPE NCCに聞いてみた

maz: ねえねえ、何か変えた？
ripe ncc: ええ！？変えてないはずだよ？

というわけで、チケット発行され現在調査中
NCC#2007012296

ripe ncc: なんでこんなことになってるのか遡って調べます・・・

ASとインターネット

- 上流からfull-routeを受信する
  → インターネットに接続する全ネットワークが受信できるはず
- 上流に自身＋顧客経路を広報する
  → 上流が自分の代わりに他のネットワークへ経路を広報してくれるはず

- もっと違う経路制御の要望が・・・

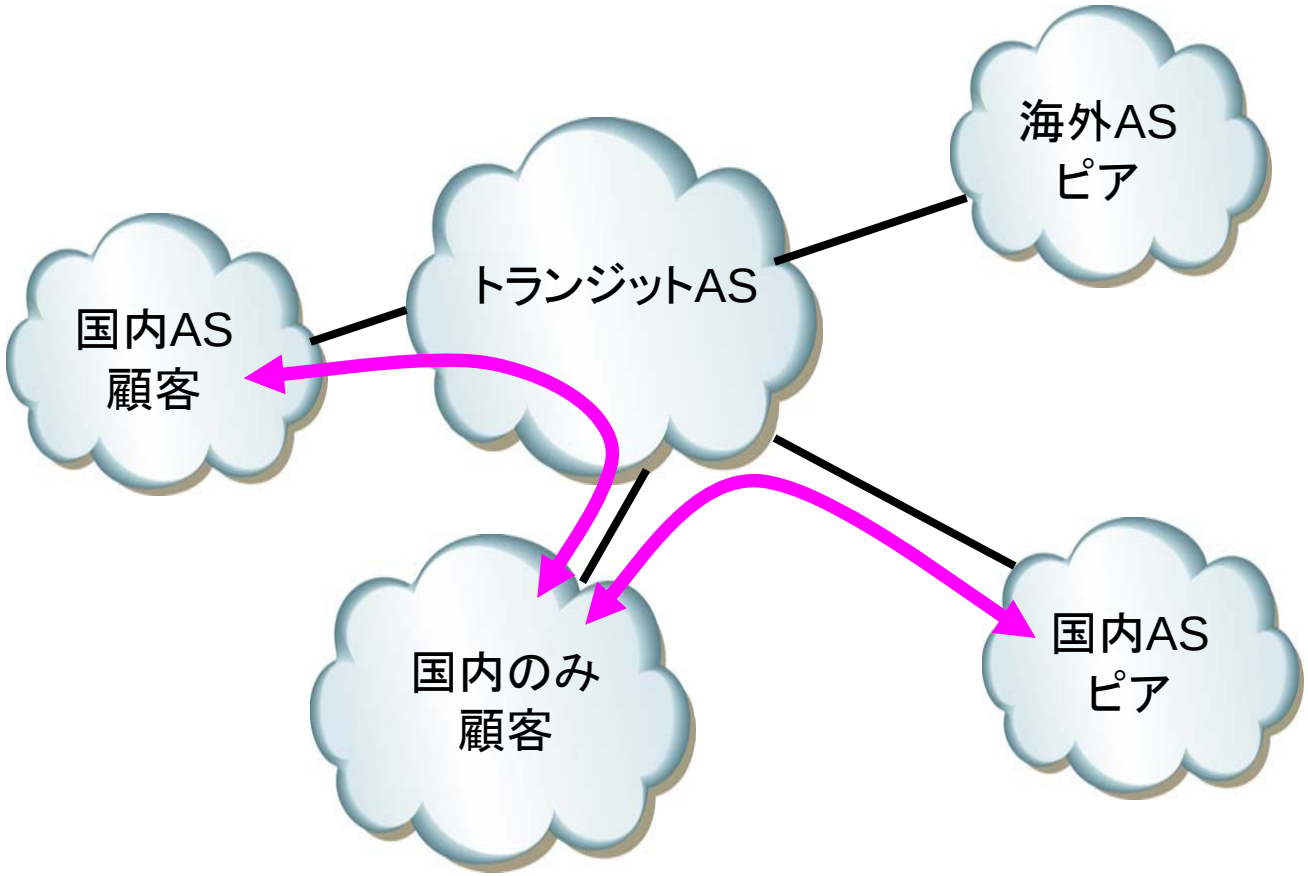
パーシャルトランジット

- full-transitにしたくない
  - トラヒック制御
  - お金の問題
- 目的に応じたいくつかの種類
  - 国際のみトランジット
  - 国内のみトランジット
  - 特定AS向けのみトランジット

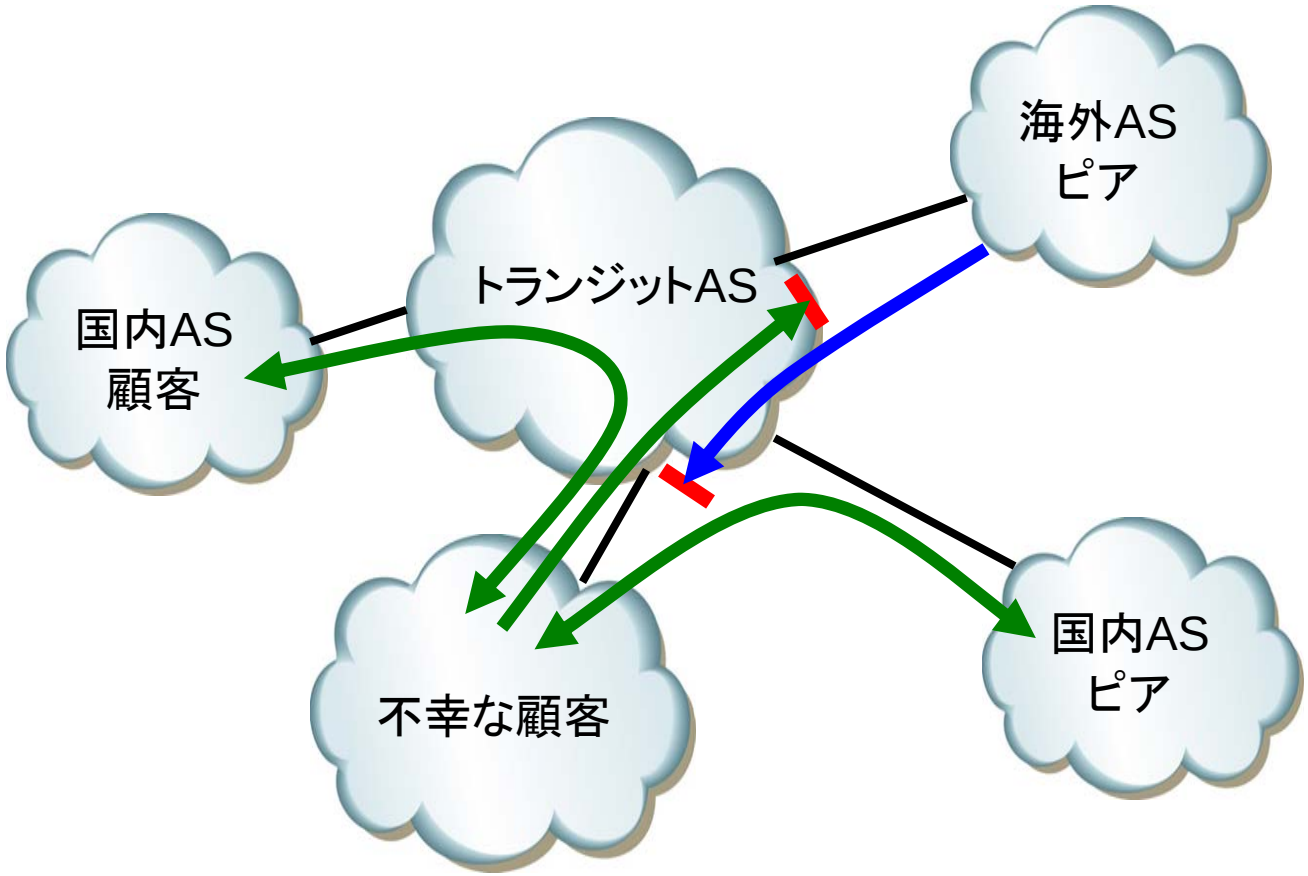
パーシャルトランジットの制御

- 経路広報時に、一部あて先への広報を抑制
  - bgp community
  - AS Path
  - さらにコンフェデレーションとの組み合わせ

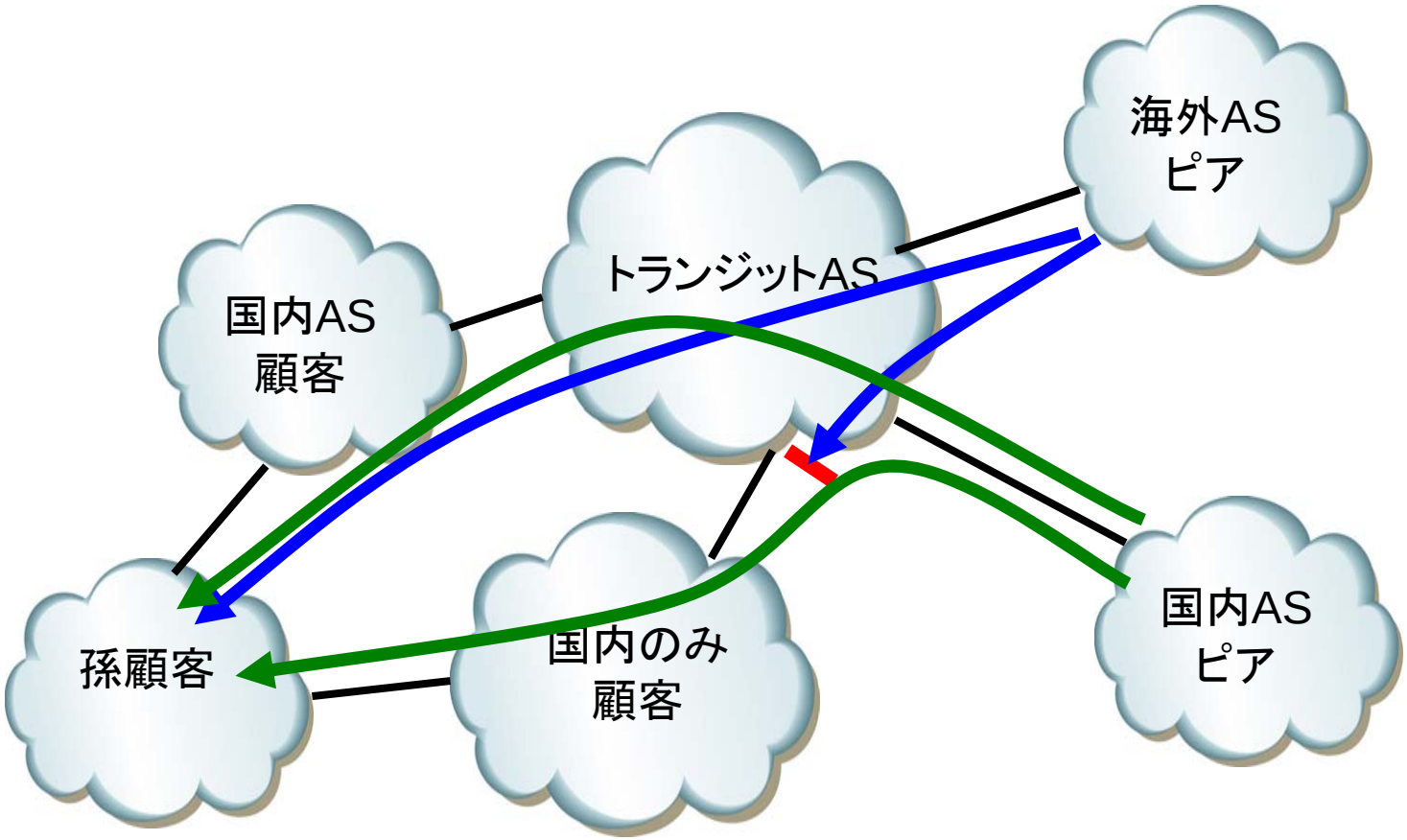
国内トランジットで期待すること



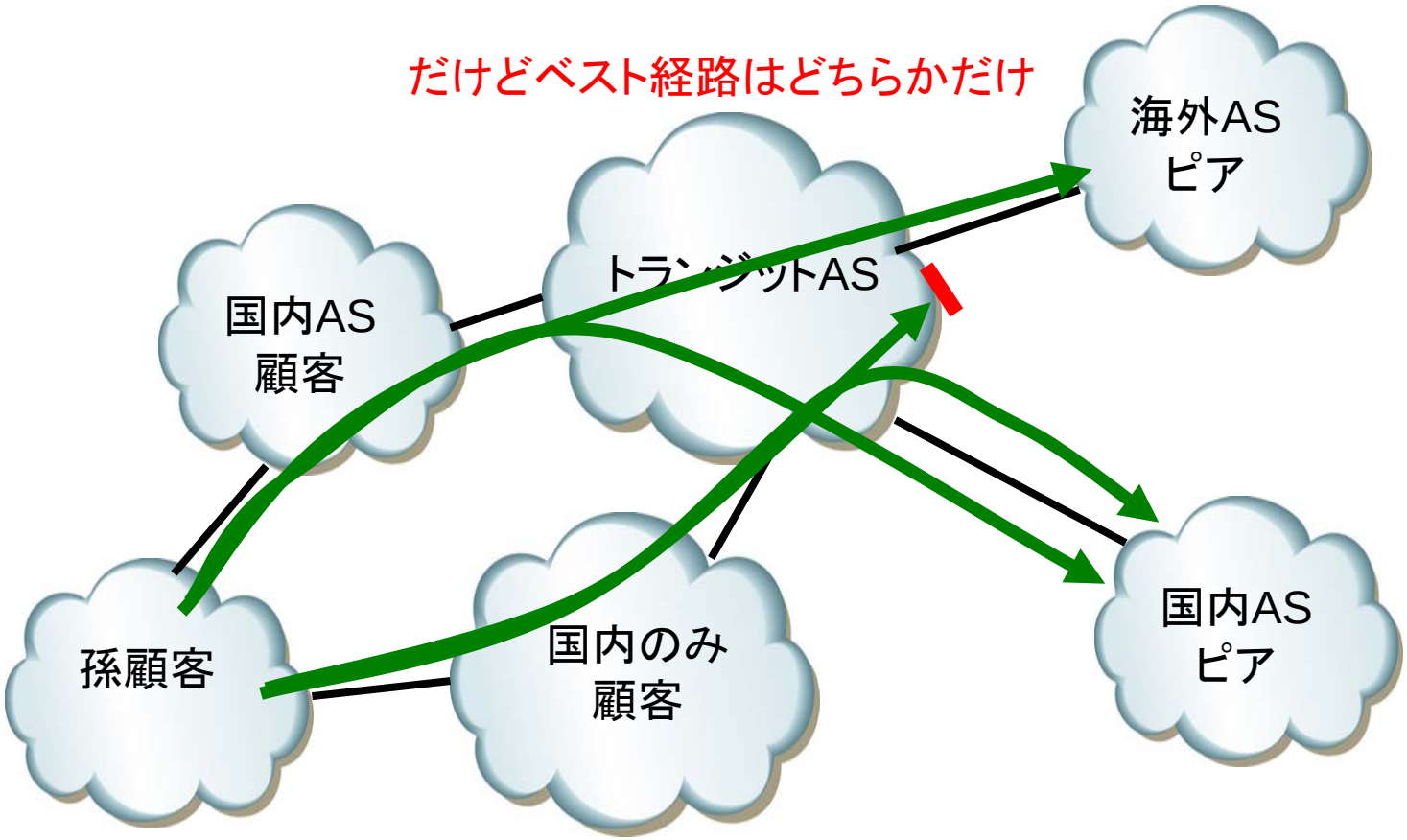
国内トランジットの経路制御例



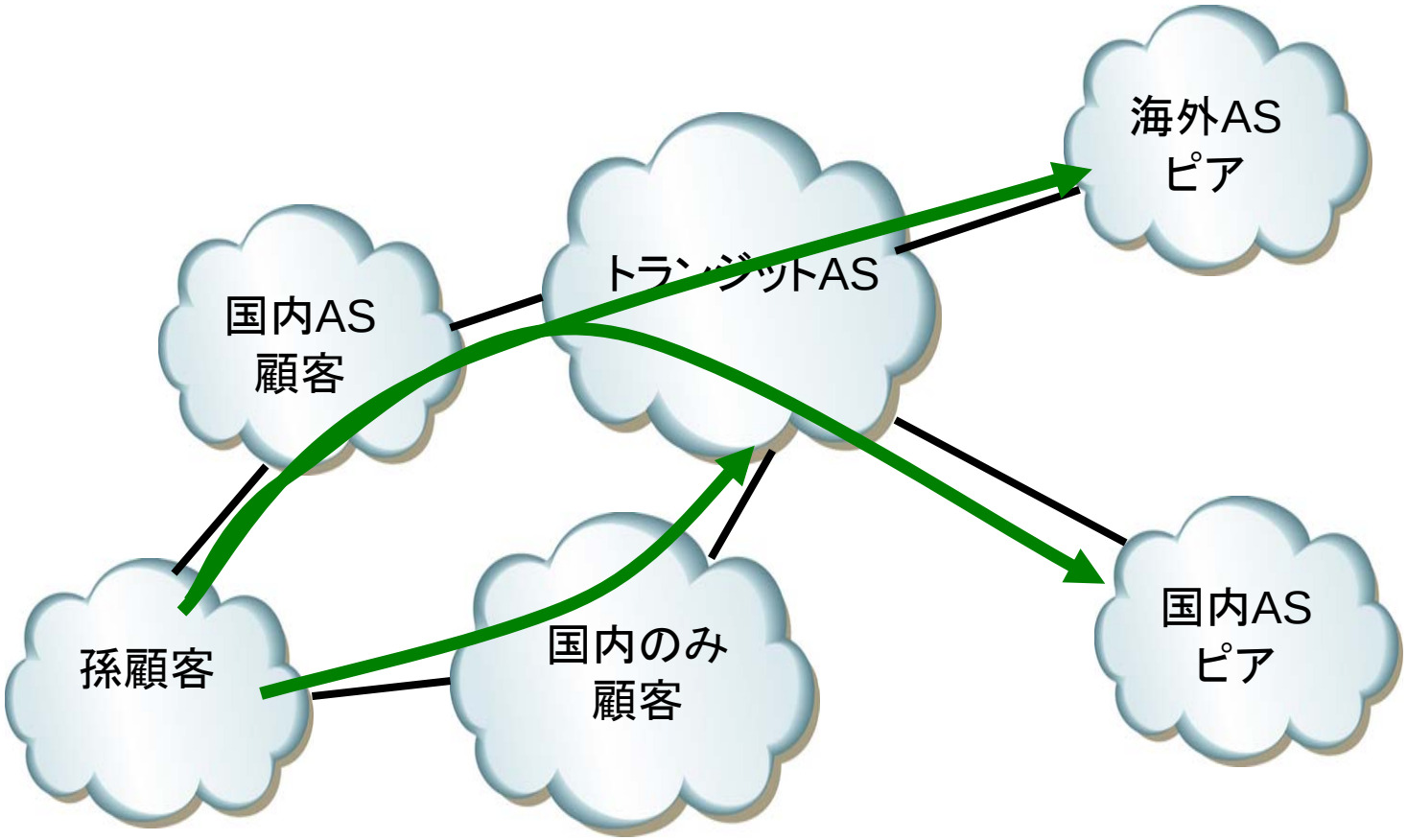
孫顧客が見る経路



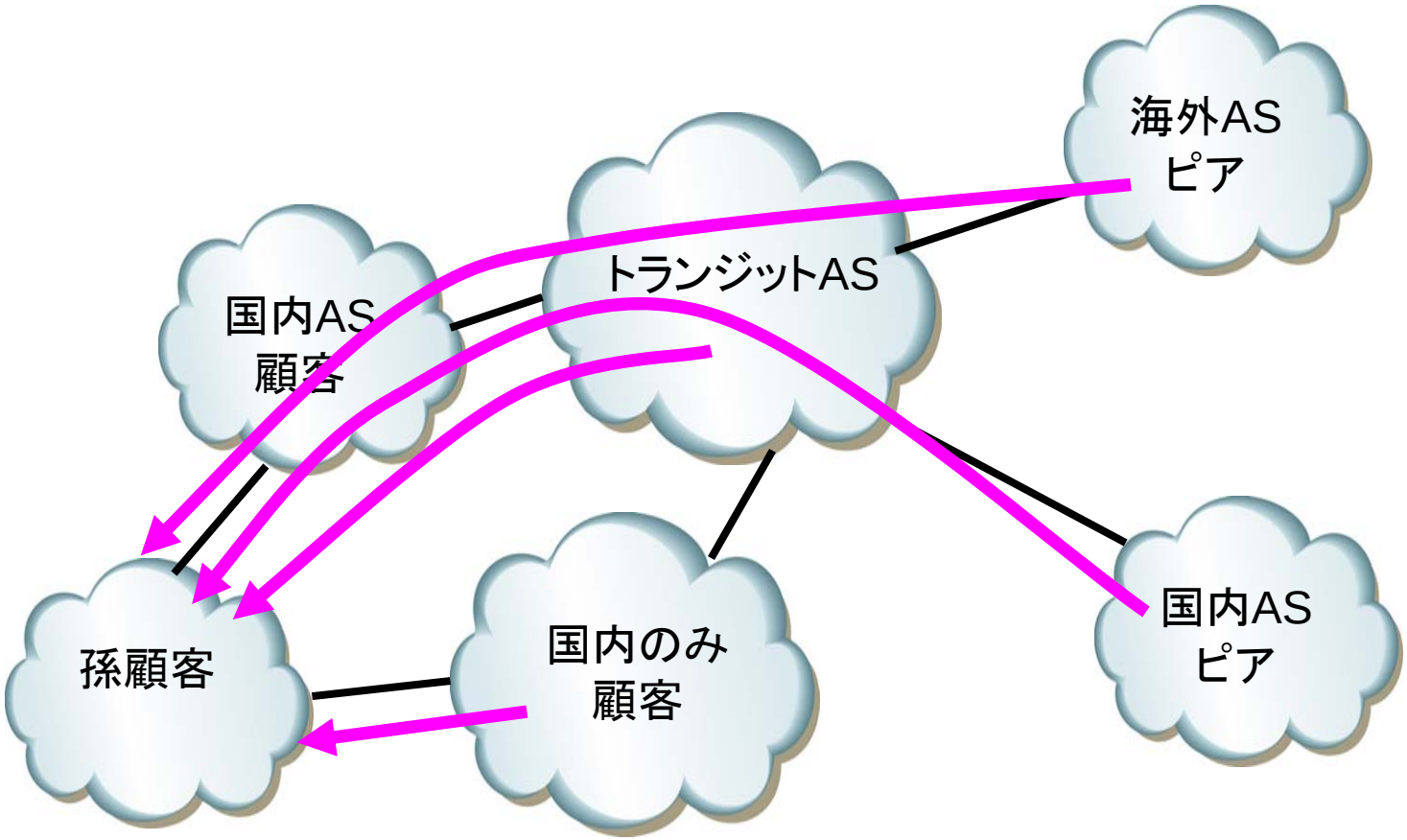
孫顧客が広報する経路



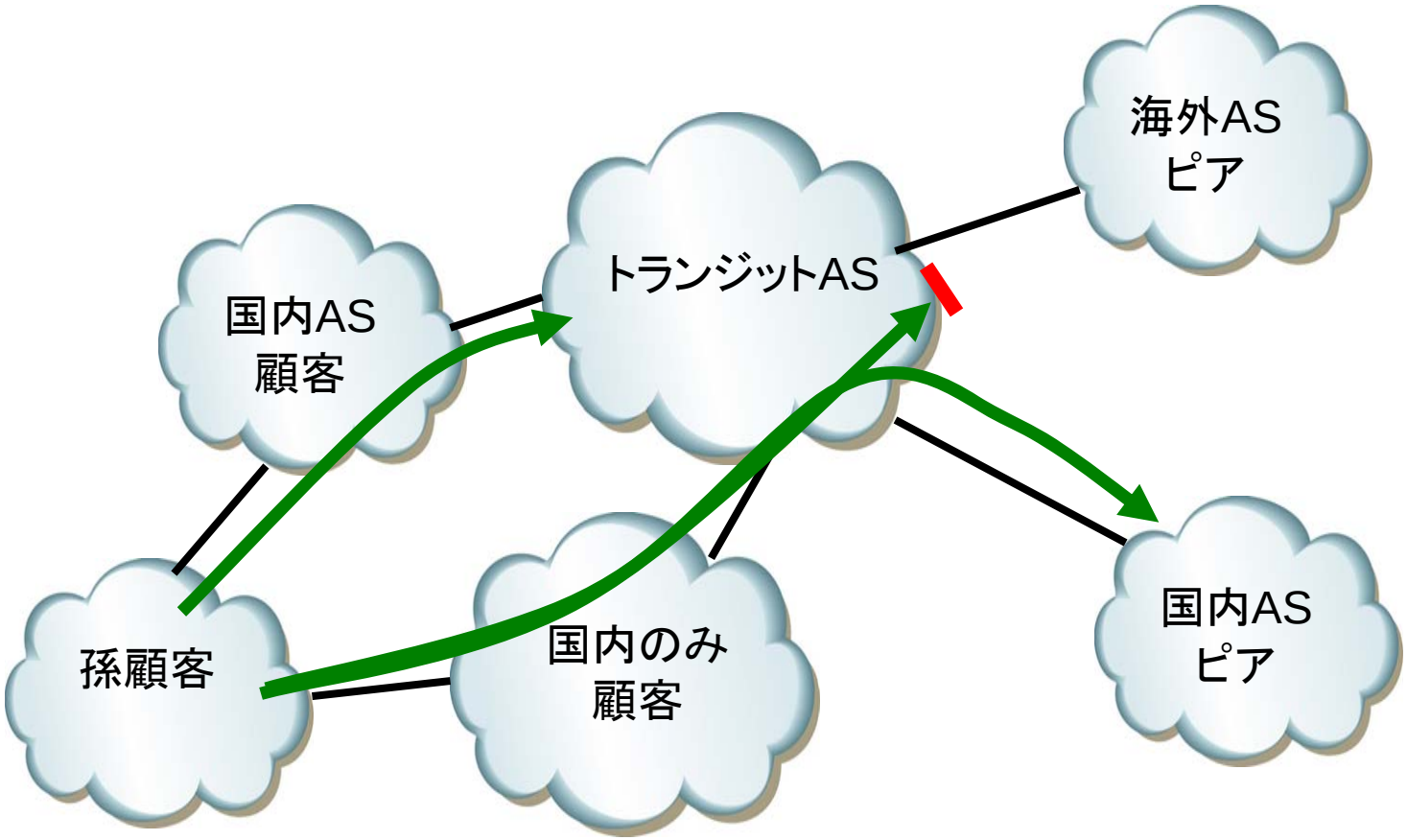
孫顧客　フルトランジット側優先時



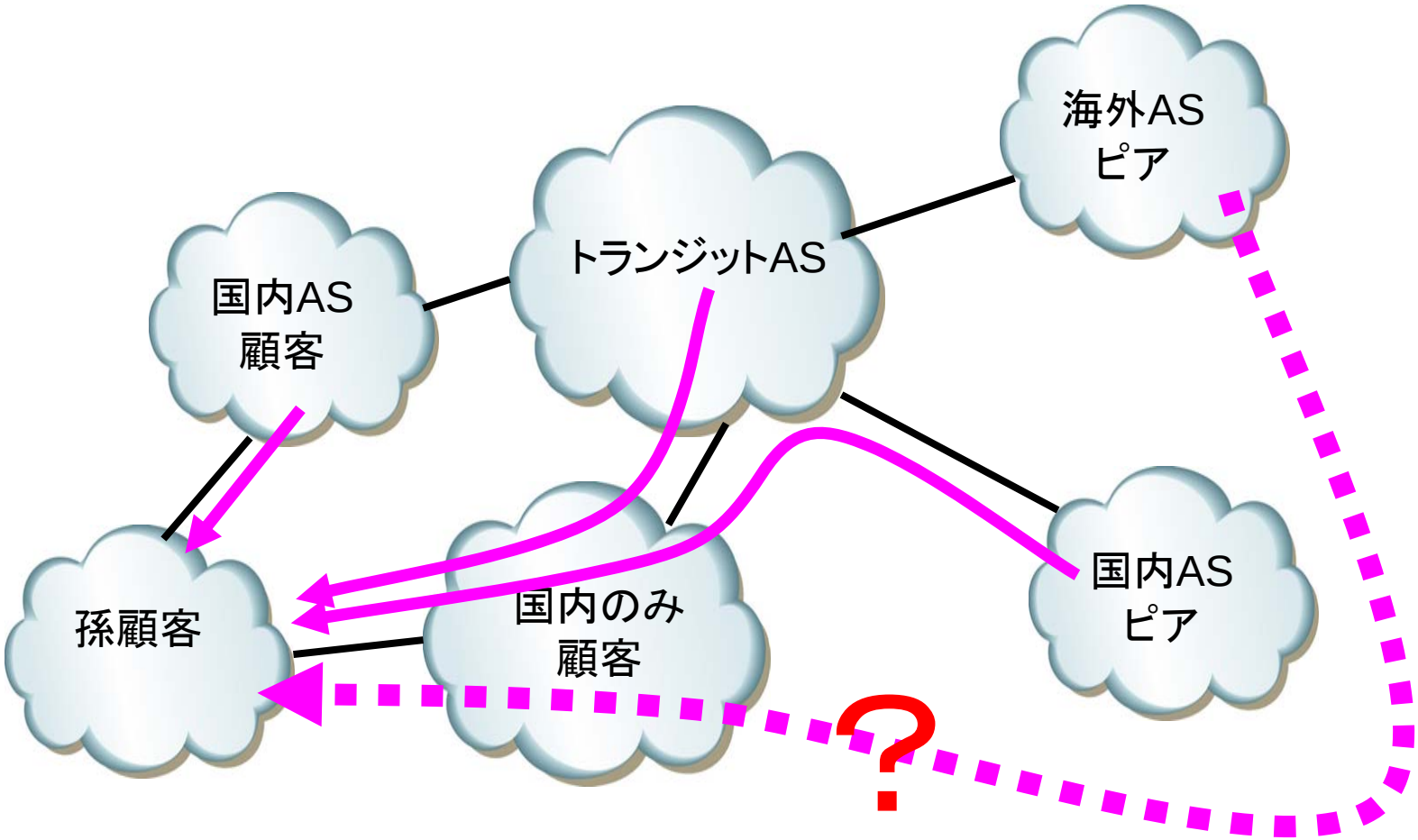
孫顧客　フルトランジット側優先時



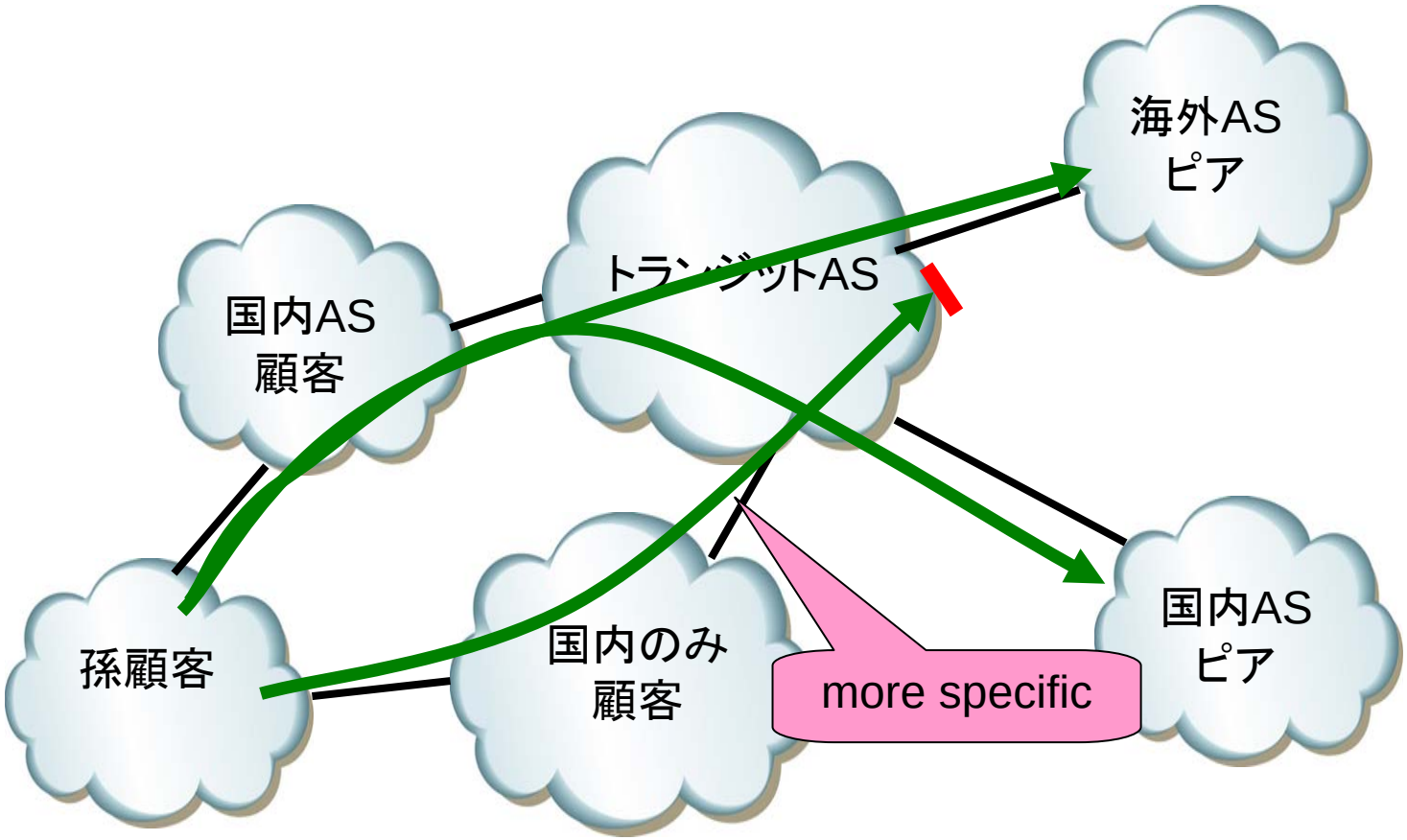
孫顧客　国内のみ優先時



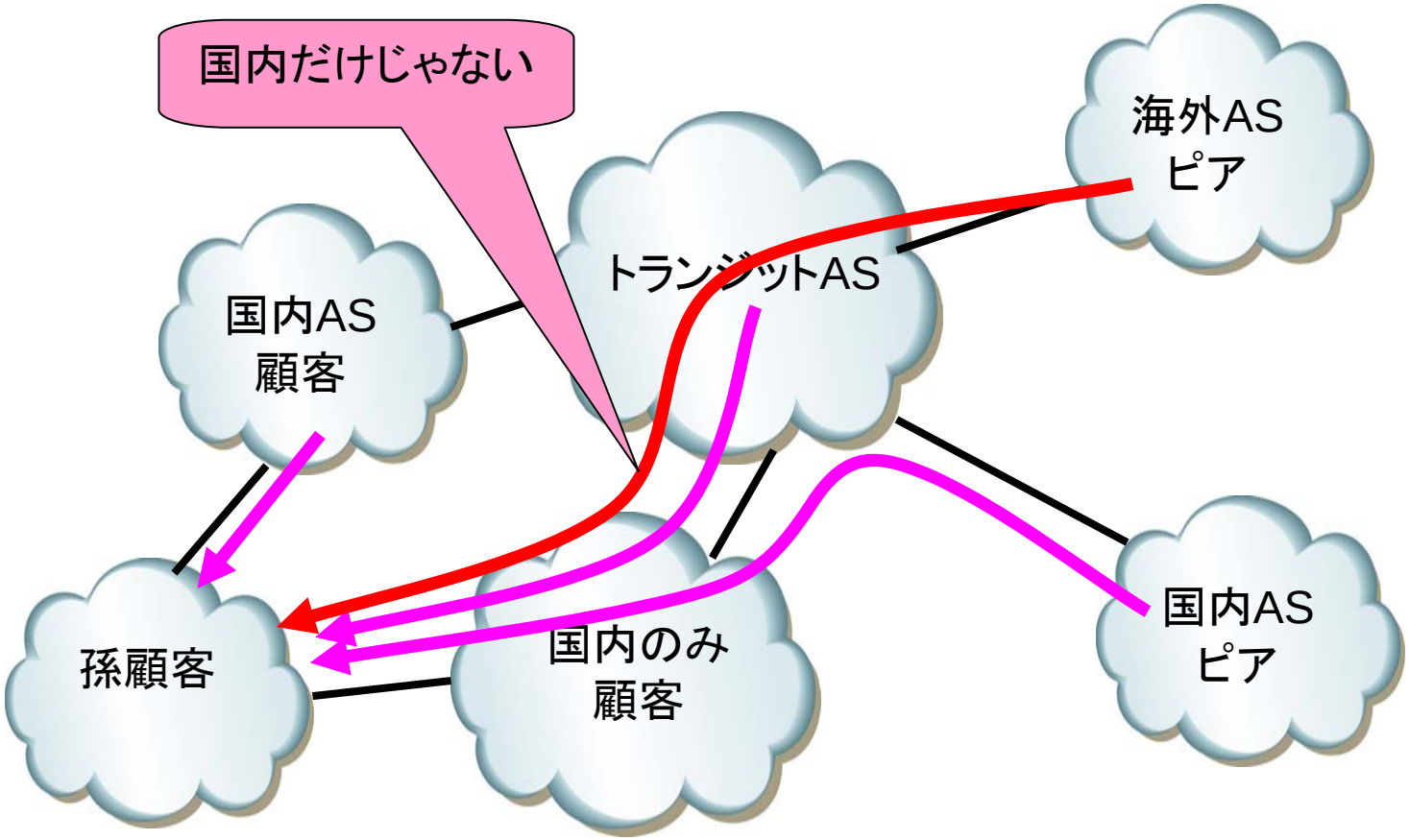
孫顧客　国内のみ側優先時



孫顧客　ただ乗り時



孫顧客　ただ乗り時



素直なポリシへの移行

- ポリシと実装の見直し
  - 関係するASの話し合いが必要
  - 出来るだけシンプルな方が安心

- 実態は・・・
  - 解約されるまで待つ
  - 構成変更のときに、新構成を提案

つまり

- 複雑なポリシは身を滅ぼす
  - みんな不幸
  - お金の問題なら、営業頑張れ
- 素直で汎用的な構成を保つ
  - スケールする
  - いつ他社の機器に置き換えても実装できる
  - 新機能の実装、構成変更の時も安心

さてさて

- パーシャルトランジット
  - 買っちゃった人、提供してる人
  - どうよ？
- その他特殊事例
  - 苦労話
- 解決事例
  - 交渉に成功しました！とかとか

そして将来

- 直近は素直なポリシで延命
  - でもいろいろ破綻要素が
  - 経路爆発とかとかとかとかとかとか
- 将来はもっと根本的にきれいに
  - 今の経路制御の問題点を考える
    - RAM (Routing & Addressing Mailinglist)
  - とかとかとかとか