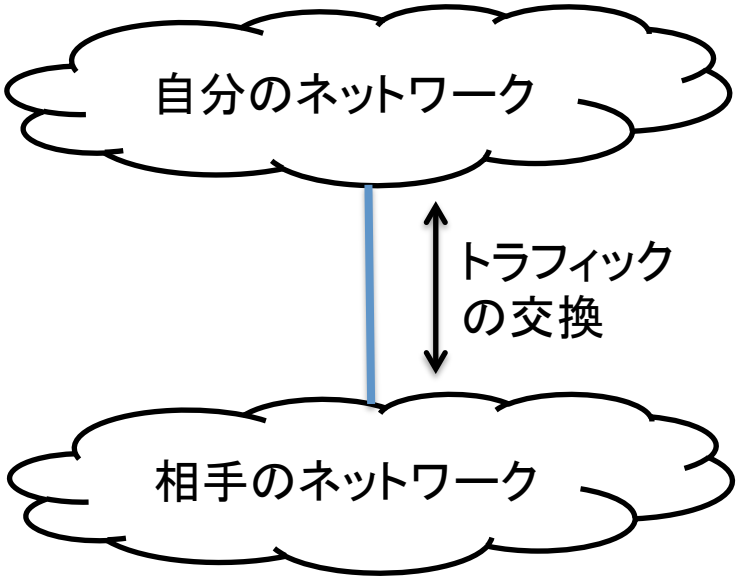
相互接続とルーティングの裏事情入門編

BIGLOBE Inc.
川村　聖一
kawamucho at mesh.ad.jp

相互接続(ピアリング)の目的

他のネットワークと自分のネットワークを「つないで」
自分のネットワークのトラフィックを相手に流し、相手のネットワークのトラフィックを流してもらうこと



用語

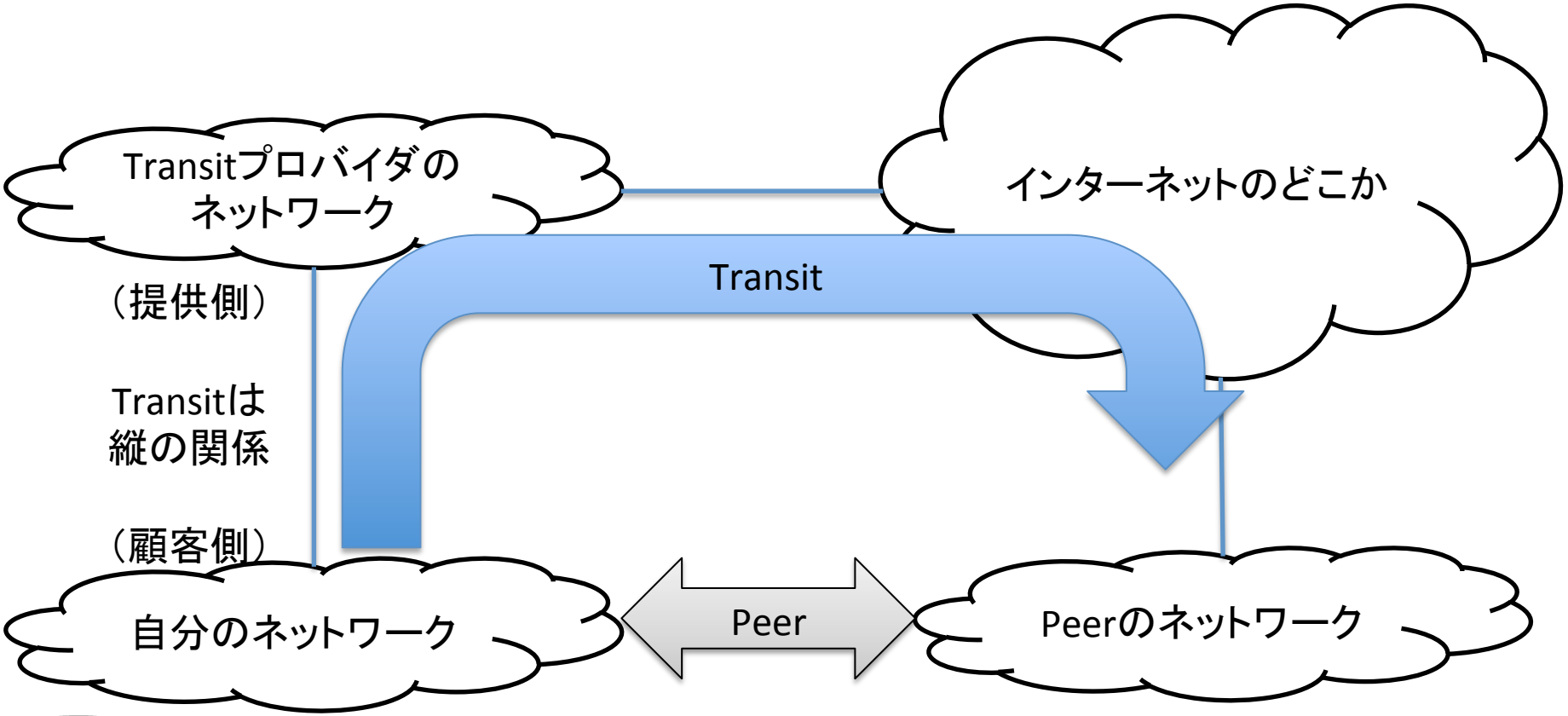
① Peer(ピアする、ピアリング)
- 動詞
- 2つのルータで、BGPを設定してセッションをはること
- 例：さくらインターネットとBIGLOBEがピアした

② Peer(ピア)
- 関係性を表す言葉
- BGPで接続していて、お互いに対当の関係
- 例：さくらインターネットとBIGLOBEはピアの関係

③ Peer(ピア)またはSession
- 単にBGPセッションの事
- 例：このさくらとのピア、セッションが安定してないね

④ Transit
- 関係性を表す言葉
- 上流ISPのこと
- 「Transit事業者とのPeer」という表現でのPeerは(3)のPeerのこと

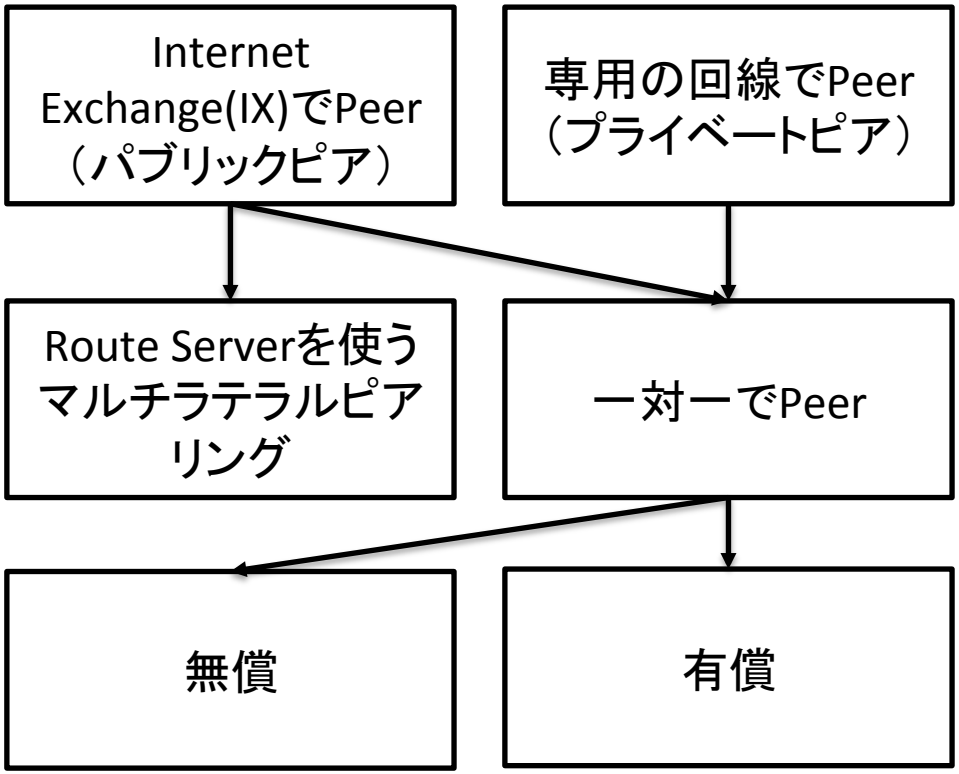
PeerとTransit(関係性の話)



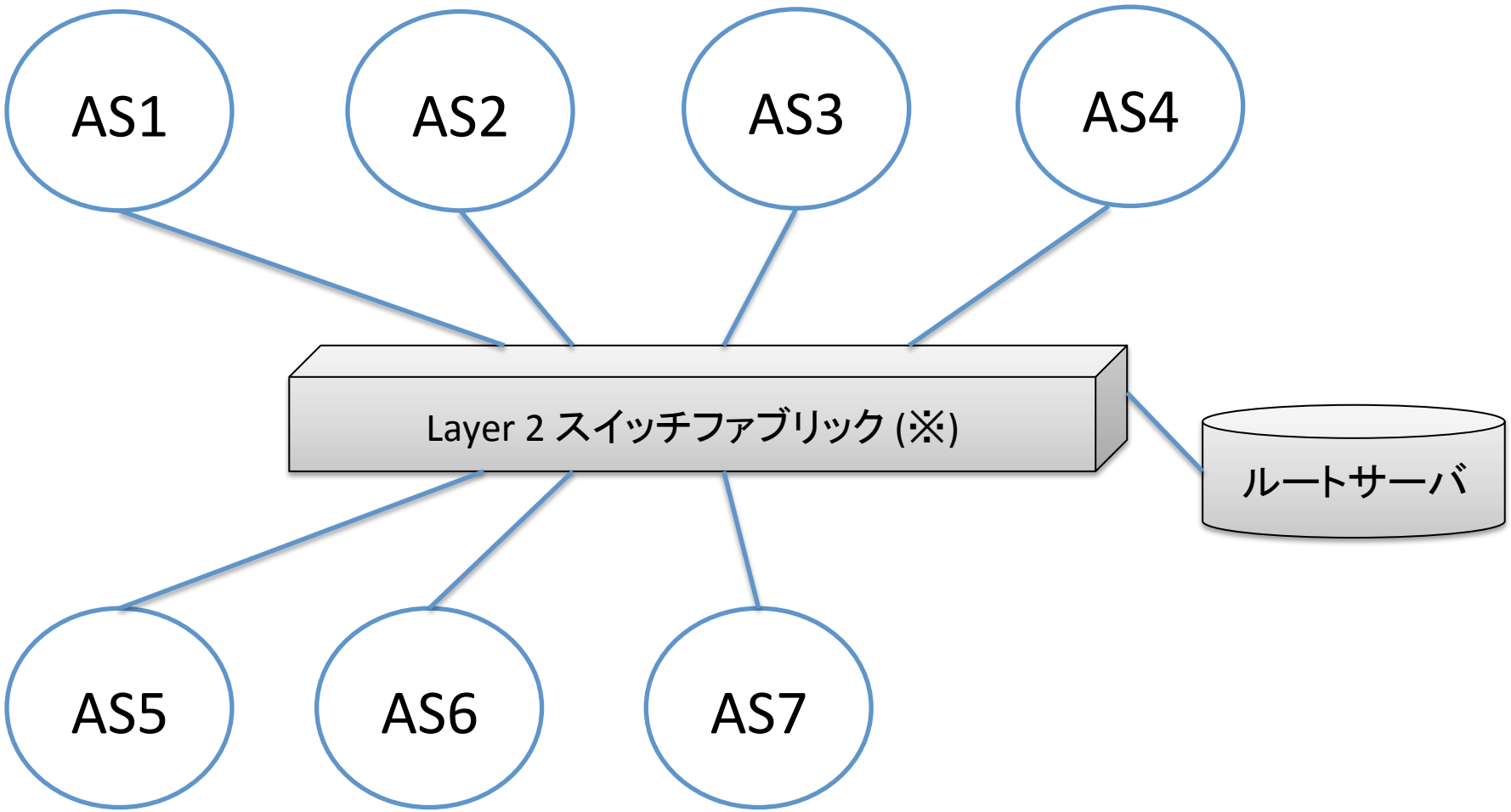
Peerは横の関係

Peerとは

- Peer:
  - 隣接関係（隣のネットワーク）
  - 直接のトラフィック交換が可能
    - どこからトラフィックが飛んでいるか管理できる

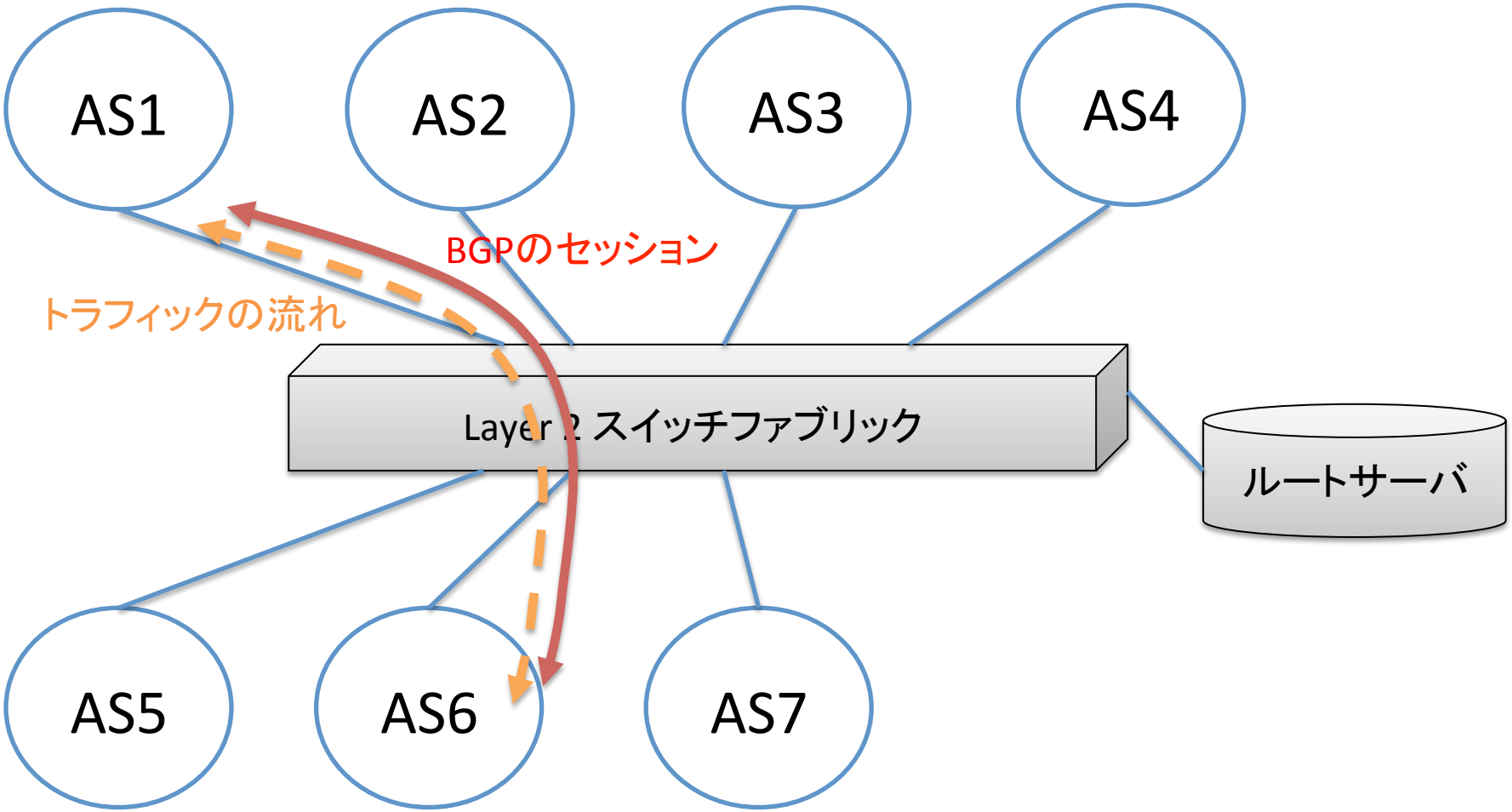


Internet Exchange



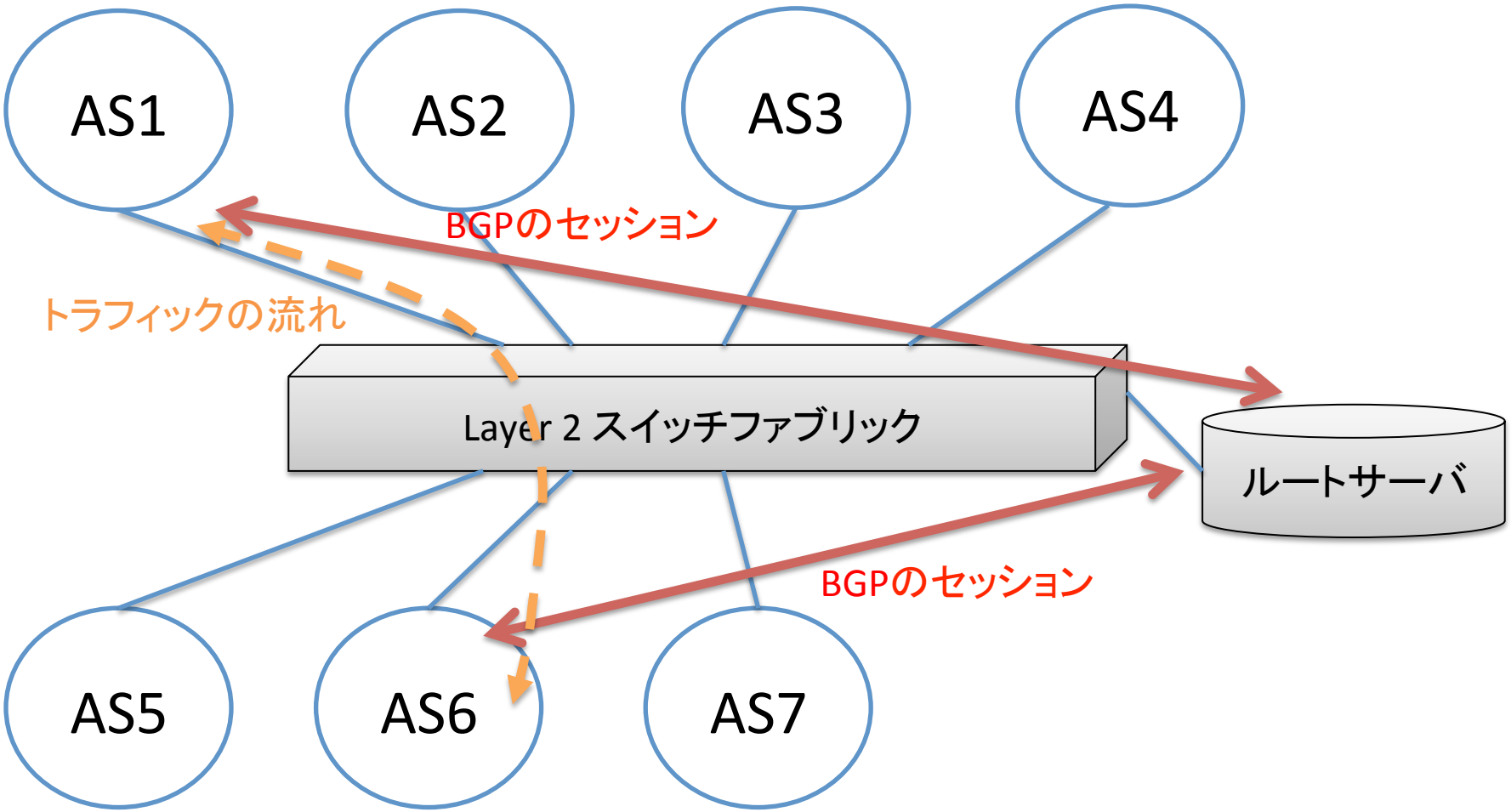
※実際は1台のスイッチでなく、複数台のスイッチでファブリックのようなネットワークになっている。冗長性や、光スイッチによるLayer1切り替えなどIXによって様々

一対一のPeer



AS1とAS6がPeerに合意した場合、それぞれのルータでBGPのセッションを設定してPeerする。IXに接続するルータは、同じSegment(LAN)上にいる

マルチラテラルのPeer



AS1とAS6はそれぞれルートサーバ（ルートサーバもAS番号がついている）とセッションを確率。ルートサーバから経路をもらい、IXのスイッチ上でダイレクトにトラフィックを交換

PeerとTransit

	Peer	Transit
目的	お互いのトラフィック交換	インターネットに接続する
有償無償	無償(settlement free peering)が多いが、たまに有償(paid peering)	原則有償
広告する経路	お互いのASで使っているPrefix、お互いのASのTransit顧客のPrefix	提供側：一般的にフルルート、たまにデフォルトルート 顧客側：自分のASのPrefixおよびTransit顧客のPrefix
契約関係	多くの場合不要。たまに覚え書きなどを締結	サービス提供約款、契約

同じeBGPでも大きく違う

PeerとTransitの数

- Transitは一般的に2社以上契約する事が多い
  - （シングルホーム）1社しか契約しない人もいる
  - 3-4社程度契約しているケースが多い
- Peerの数は事業者の規模やネットワーク構成によって大きく異なる
  - BIGLOBEは現在約150ASとPeerしています

Peerの設定例(JUNOSの場合)

当日ご紹介します

Peerの設定例(参考ドキュメント)

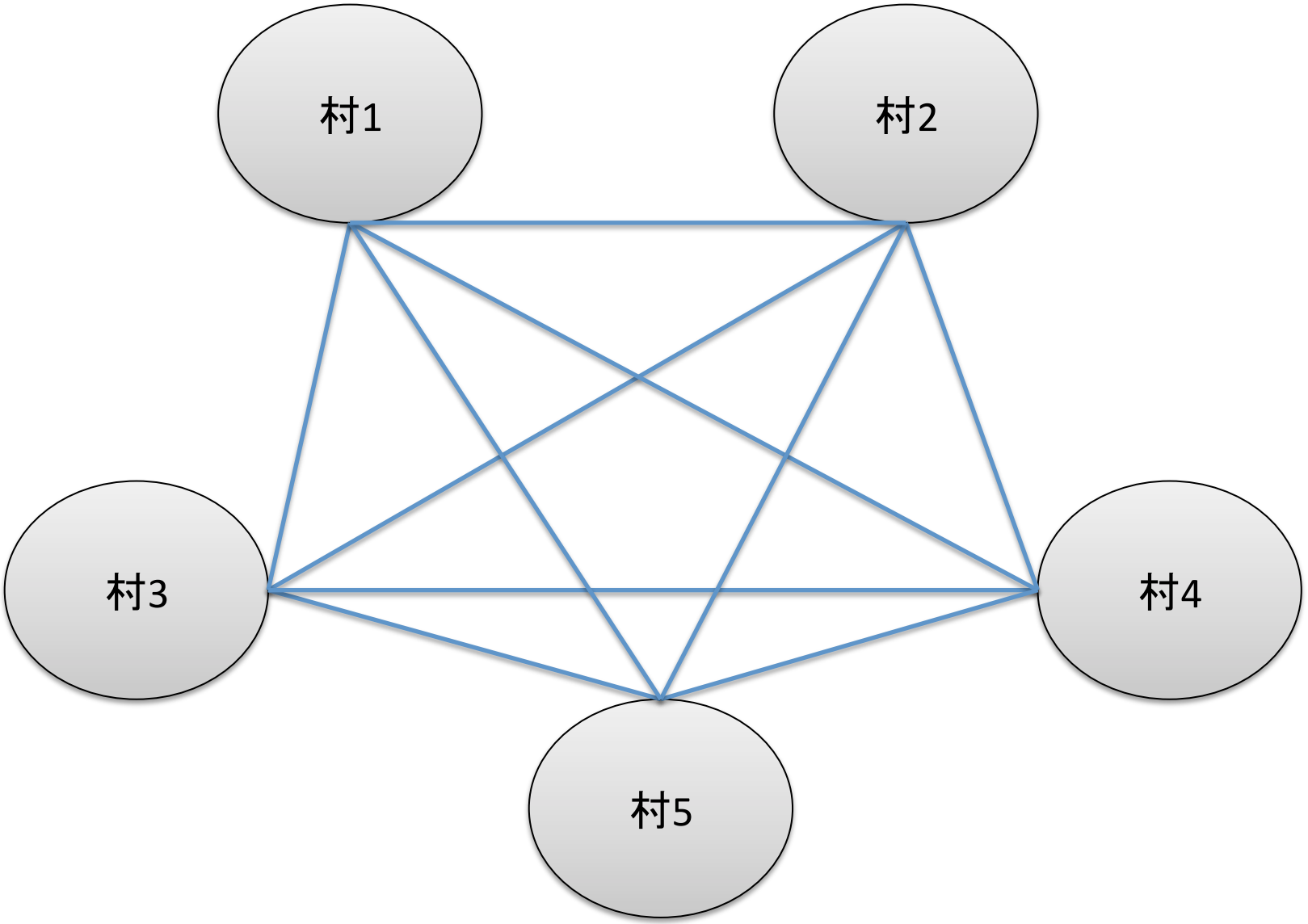
- http://bcop.nanog.org/index.php/EBGP_Configuration_BCOP_v0.1
- http://www.janog.gr.jp/doc/janog-comment/jc1002.txt
- http://www.janog.gr.jp/doc/janog-comment/bcop-ebgp.txt
  - リンクは移動するかもしれません

インターネットが5つの村だったら

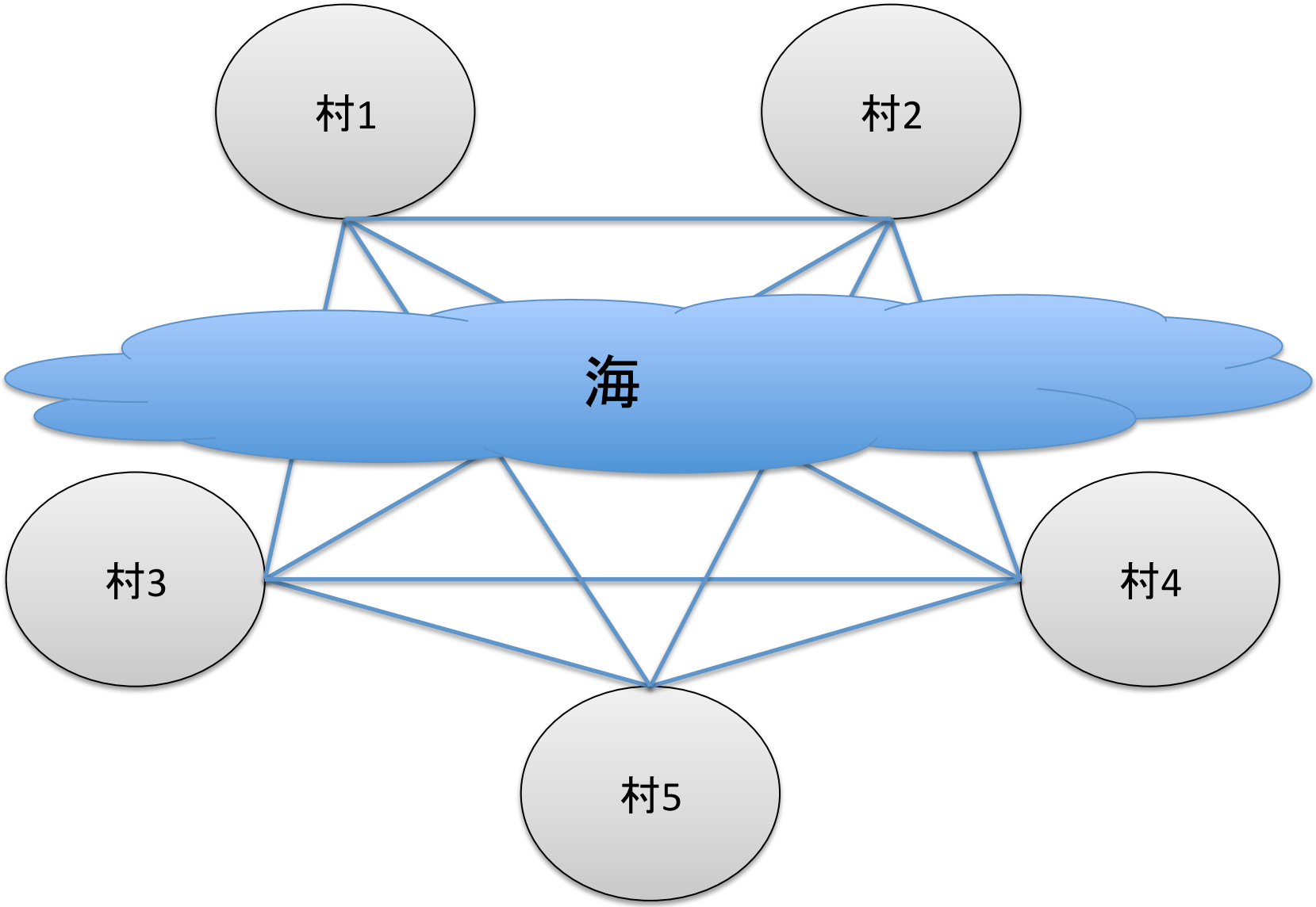
- 技術的にはシンプルなはずのeBGPを使ったPeering
- これが複雑になる背景とは？

全員がPeerの関係

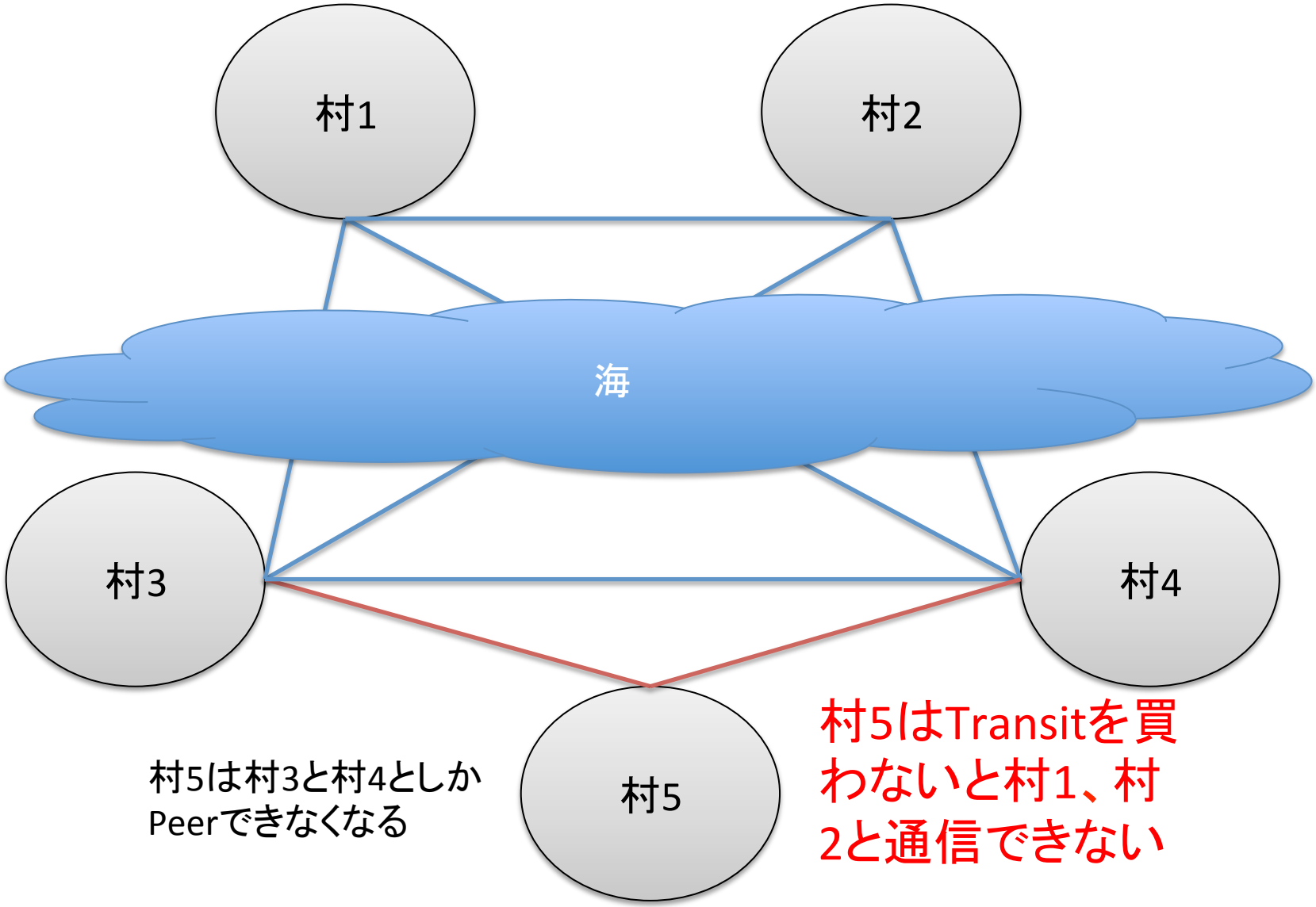
注：青線がPeer、赤線がTransit



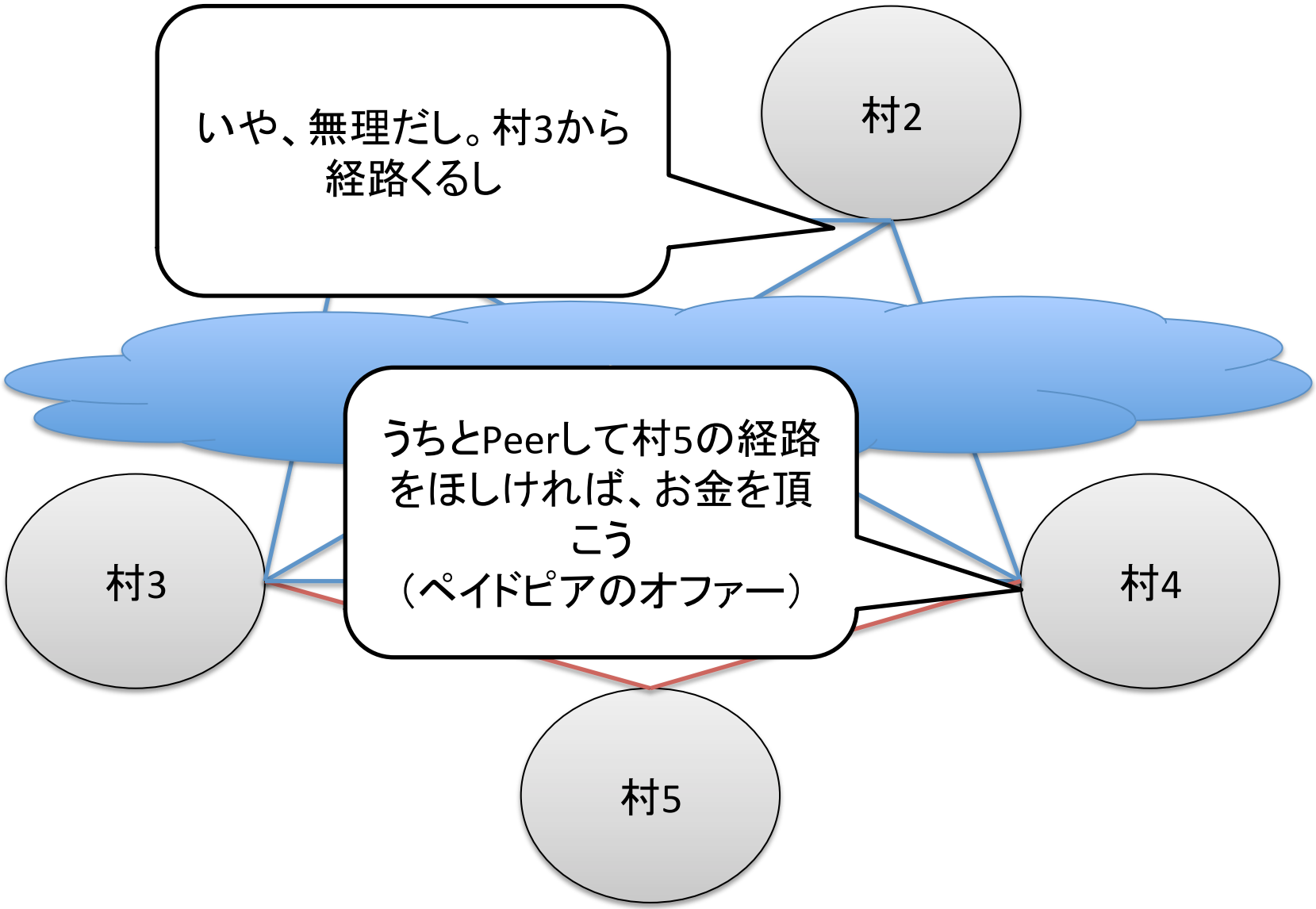
もし間に海があったら・・・



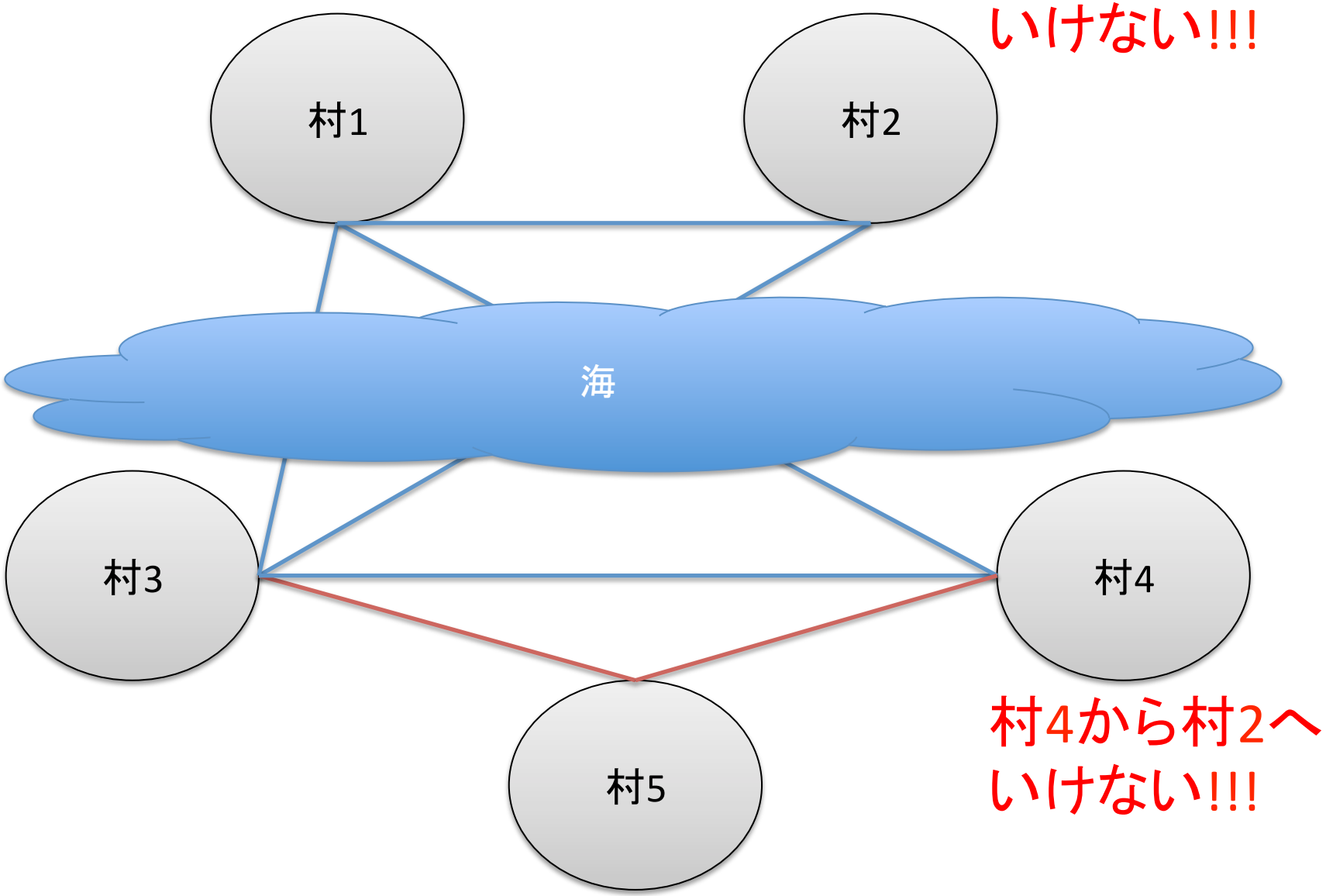
なんと村5は海に接していませんでした



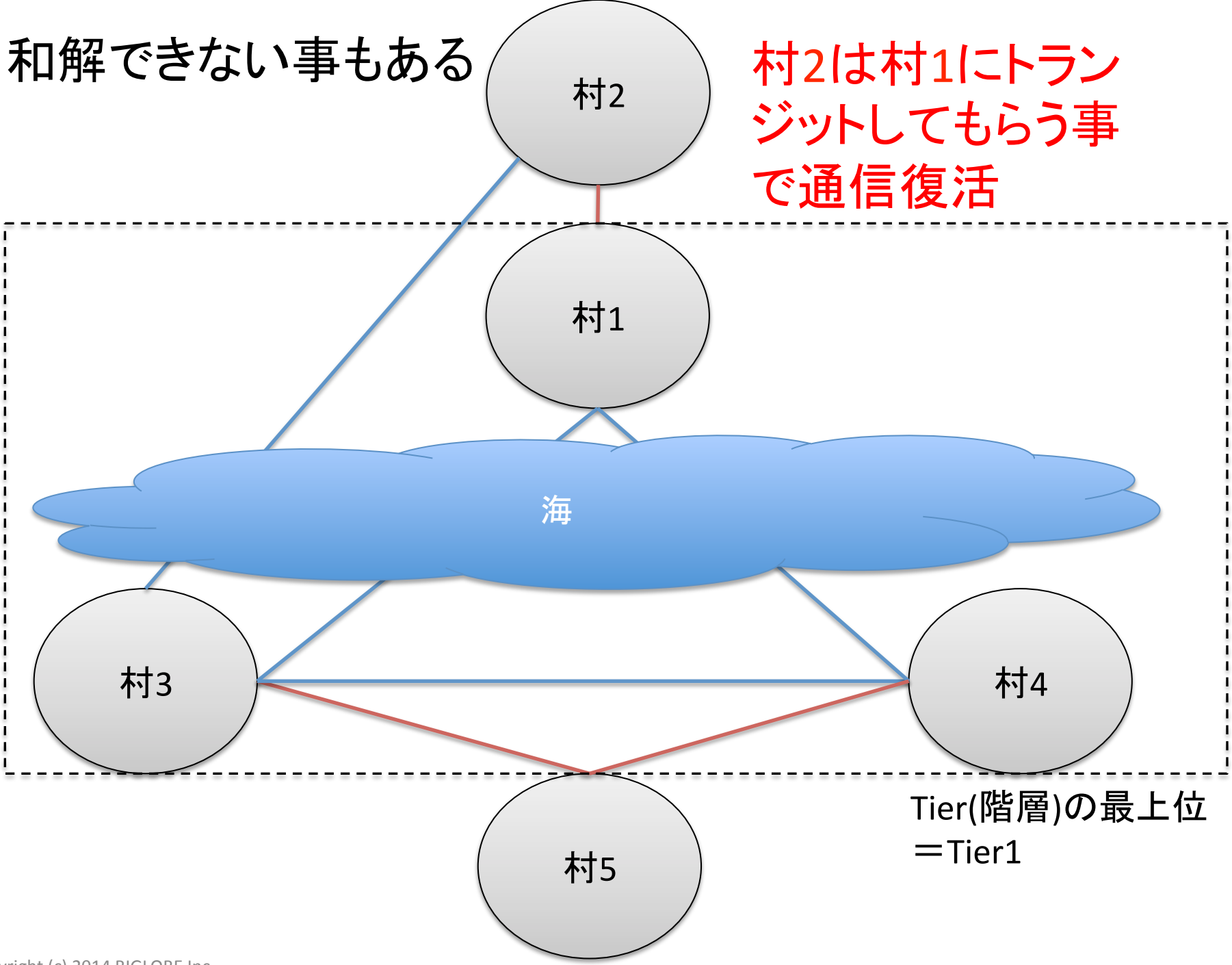
村4が強気に出ました・・・



デピア（depeer）



和解できない事もある



村2は村1にトランジットしてもらう事で通信復活

Tier(階層)の最上位
＝Tier1

実際におきます

- http://slashdot.jp/story/05/10/08/0239233/米国ISP間でネットワークの断絶が発生
- 米国のTier1事業者であるLevel3とCogentの間で条件が折り合わず、Level3側がPeerを切断
- 大規模な通信断が発生
- ただしこれはTier1事業者間で起きると大問題になるが、下の階層の事業者ではdepeerで大規模通信断になる事は稀
  - 通信品質劣化は起きる事があります
  - 逆に、Peerする事で品質劣化になる場合もあります（Peerすると必ずしも良い事ばかりではありません）

ピアリングとポリシー

- ピアリング（相互接続）には何かしらのポリシー（判断基準）が介在します
- ポリシーは、そのネットワークが置かれた状況に依存します
  - 地理的状況
  - 政策・経済的状況
  - コスト構造
  - 提供するサービス内容
  - ネットワークの設計方針

おおまかなポリシーの種類

Restrictive

特殊の状況をのぞき、新規にはほとんどPeerしない。
Tier1、超大手国際キャリア、米国巨大ISPなど歴史の長い事業者の場合が多い。また買収の結果Restrictiveになってしまうケースもある

Selective

条件を満たす場合にはPeerする。条件のキツさは事業者によってバラバラ。ローカルISP、巨大な国際コンテンツ/CDNサービスプロバイダ、データセンタ事業者、Transitサービスを提供している事業者に多い。

Open

原則どの事業者ともPeerする。コンテンツ事業者やローカルISPで良く設定されるポリシー

- 過去にOpenだった事業者がSelective/Restrictiveになる事があるが、逆はあまり見ない
- タイミングがよければPeerできる事もある
  - 担当者変更など

ピアリングの力学

- 調達量が多い場合、1Mbpsあたりの価格が低くなる
- ネットワークの「面」が大きい程かかえる問題は複雑になりがち
  - 国際通信、たくさんのASをトランジットしていると様々な国/事業者の運用の違いを吸収しないといけない
- 大きな事業者と小さな事業者ではコスト構造や体制が異なるため、Peerに対しての考え方も違う
- トラフィックを出す側（コンテンツ）とトラフィックを吸う側（プロバイダ）でのバックボーン運用維持コストの違い
- などなど・・・

ピアリングポリシーの例

- http://www.biglobe.co.jp/pdf/PeeringPolicy2013_jp_03.pdf
- まずは下記条件のどれかを満たした上で
  - (ア) 公開された IRR(RADB や JPIRR など)に as-set object および route object を適切に登録している
  - (イ) PeeringDB (http://www.peeringdb.com/)に登録している
  - (ウ) Peeering の情報を公開している Web を保持している
  - (エ) 上記(ア)から(ウ)以外だが、AS 情報を適切に連絡できる手段がある
- 事業者毎の基準がある
  - AS を Transit する事業者さまの場合
    - 15 以上の AS を Transit している
  - FTTH、ADSL、モバイルなどエンドユーザにアクセスサービスを提供している国内事 業者、またはホスティングサービス(クラウド、VPS を含む)事業者さまの場合
    - BIGLOBE と Peer する AS の Origin トラフィック(incoming もしくは outgoing) が最大値で定常的に 180Mbps をこえている
  - モバイル向けアプリ・コンテンツサービスを提供している事業者さまの場合
    - 変化の激しい市場のため、可否について個別に検討させて頂きますのでお問い合わせください。なるべくご期待に沿えるよう努力はします
  - Root DNS などのクリティカルインフラ、The Internet の運営にとって重要な非営利組織、The Internet の安定運用に貢献しているプロジェクト、などの事業者さまの場合
    - 詳細を教えて頂き、懸念事項がなければ Peer します

つまりピアリングとは

- ビジネスデシジョンである
  - 自分のネットワークにとって損か特か、相手にとって損か特か
- ビジネスデシジョンをするためには、仕組み（技術的な背景）の理解が必要
  - これをやる人は「コーディネーター・アーキテクト・ストラテジスト」のような肩書きを持っていたりする
- デシジョンをインプリ（設定・導入）し、問題が発生しないように設定調整するのがオペレーション（運用・技術）
- 「ネットの品質」はビジネス取引外であっても自分に影響するため、ビジネスを超えた協業・協力は重要

PeeringDB

- https://www.peeringdb.com
  - guest/guest で参照ができる
  - 特定のASの情報を見たければ
    - https://as2518.peeringdb.com/
  - whois -h www.peeringdb.com as2518
  - SQLでも取って来れる
- 米国の有志によって2004年に設立
  - Peerに必要なASの情報、IXの情報、Peerできるデータセンター情報が掲載されている（登録申請し、有志がチェックした上で登録されている）
  - 掲載情報をPeer目的以外で使う事は御法度！

Navigation
- Home Page
- Logout

Your Records
- Peering Record
- User Account

Search Records
- Networks
- Exchange Points
- Facilities
- Common Points

Suggestions
- Comments
- New Exchange
- New Facility

Help
- FAQ
- Statistics

Company Information

Company Name	BIGLOBE Inc.
Also Known As	FullRoute
Company Website	http://www.biglobe.co.jp/en/
Primary ASN	2518
IRR Record	AS-MESH
Network Type	NSP
Approx Prefixes	500
Traffic Levels	200-300 Gbps
Traffic Ratios	Balanced
Geographic Scope	Asia Pacific
Looking Glass URL	http://lg.fullroute.net/lg/
Route Server URL	
Notes	We are connected to BBIX Singapore but not Hong Kong. We will only peer with Singapore switch participants. You can ping and trace to ping.mesh.ad.jp(both IPv4 and IPv6). IRR Records should be viewed with source set to RADB or JPIRR. There are colliding records with the same AS-SET name, but from a different AS on RIPE. We are currently not accepting new sessions at JPNAP. See here for Japanese version of the peering policy http://www.biglobe.co.jp/pdf/PeeringPolicy2013_jp_03.pdf
Protocols Supported	Unicast IPv4 ☑ Multicast ☐ IPv6 ☑
Date Last Updated	2014-11-06 23:33:17 UTC

Peering Policy Information

Peering Policy URL	http://www.biglobe.co.jp/en/peering_policy.pdf
General Policy	Selective
Multiple Locations	Preferred
Ratio Requirement	No

Public Peering Exchange Points

Exchange Point Name	ASN	IP Address	Mbit/sec
BBIX Hong Kong / Singapore	2518	2001:df5:b800:bb00::2518:1	10000
BBIX Hong Kong / Singapore	2518	103.231.152.25	10000
BBIX Tokyo	2518	2001:de8:c::2518:1	10000
BBIX Tokyo	2518	218.100.6.73	10000
BBIX Tokyo	2518	218.100.6.47	10000
BBIX Tokyo	2518	2001:de8:c::2518:2	10000
CoreSite - Any2 California	2518	206.72.210.154	10000
CoreSite - Any2 California	2518	2001:504:13::210:154	10000
Equinix San Jose	2518	2001:504:0:1::2518:1	10000
Equinix San Jose	2518	206.223.116.156	10000
Equinix Singapore	2518	202.79.197.180	10000
Equinix Singapore	2518	2001:de8:4::2518:1	10000

1 2 of 2 Next > Last >>

Private Peering Facilities

Facility Name	ASN	City	Country	SONET	Ethr	ATM
1-Net Singapore	2518	Singapore	SG	☐	☑	☐
Equinix Singapore	2518	Singapore	SG	☐	☑	☐
Equinix Tokyo (TY2)	2518	Tokyo	JP	☐	☑	☐
MEGA iAdvantage Hong Kong	2518	Hong Kong	HK	☐	☑	☐
Telehouse Tokyo	2518	Tokyo	JP	☐	☑	☐

List of Public Exchange Points

Exchange Name	Long Name	City/Region	Country	Continental Region	Media Type	Participants
ASSOCIO	mpls ASSOCIO	Tokyo, Nation wide	JP	Asia Pacific	Multiple	0
BBIX Fukuoka	BroadBand Internet eXchange Fukuoka	Fukuoka	JP	Asia Pacific	Ethernet	0
BBIX Nagoya	BroadBand Internet eXchange Nagoya	Nagoya	JP	Asia Pacific	Ethernet	1
BBIX Osaka	BroadBand Internet eXchange Osaka	Osaka	JP	Asia Pacific	Ethernet	4
BBIX Tokyo	BroadBand Internet eXchange Tokyo	Tokyo	JP	Asia Pacific	Ethernet	77
DIX-IE	Distributed IX in EDO (former NSPIXP2)	Tokyo	JP	Asia Pacific	Ethernet	36
Echigo-IX	Echigo Internet Exchange	Niigata	JP	Asia Pacific	Ethernet	2
Equinix Osaka	Equinix Osaka	Osaka	JP	Asia Pacific	Ethernet	1
Equinix Tokyo	Equinix Tokyo	Tokyo	JP	Asia Pacific	Ethernet	109
JPIX	Japan Internet Exchange	Tokyo	JP	Asia Pacific	Ethernet	168
JPIX OSAKA	Japan Internet Exchange Osaka	Osaka	JP	Asia Pacific	Ethernet	21
JPNAP Osaka	Internet Multifeed JPNAP Osaka	Osaka	JP	Asia Pacific	Ethernet	32
JPNAP Tokyo	Internet Multifeed Company	Tokyo	JP	Asia Pacific	Ethernet	143
JPNAP Tokyo 2	JPNAP Tokyo II Service	Tokyo	JP	Asia Pacific	Ethernet	5
NSPIXP3	Wide NSPIXP3	Osaka	JP	Asia Pacific	Ethernet	8

NOTE: Sending Unsolicited Commercial Emails to contacts mined from PeeringDB will result in a ban and public embarrassment.

Peeringを担当する事になったらやる事 ―情報収集―

- JANOG
  - https://www.janog.gr.jp
  - 年2回のミーティング、メーリングリスト
- Peering in Japan
  - 不定期にBoFの開催
  - AS運用者が集まるメーリングリスト（IX接続している方限定）
  - https://groups.google.com/forum/#!forum/peering-jp
- 英語が得意ならnanog、outagesメーリングリスト
- IX主催のユーザ会に参加
  - ユーザでなくても参加できるものもあります
- 新聞を読む
  - 企業の統廃合、サービスの変化、政策変化などはポリシーに影響する
  - 為替は通信コストを左右する

Peeringを担当する事になったらやる事 ーツールの活用ー

- http://bgp.he.net
- http://www.bgp4.as/looking-glasses
- http://www.routeviews.org
- http://www.ripe.net/data-tools/stats/ris
- IRR
  - http://www.radb.net
  - https://www.nic.ad.jp/ja/ip/irr/index.html
- http://www.bgpmon.net
- http://www.submarinecablemap.com

Peeringを担当する事になったらやる事 ―経路的直感を磨く―

- 飛行機1時間＝10msecの遅延
- Traceroute慣れする
  - traceroute –I –w 1 –A whois.radb.net
  - 逆引きから何となく国や街がわかる
- IPアドレス帯の直感
  - 1.0.0.0, 27.0.0.0は枯渇直前に割り振られたアドレス帯、103.0.0.0は枯渇後のアドレス帯で事件が多い
- ルーティングテーブルと友達になる
  - Routeviewで遊ぶ
  - 手元にfull routeもっておく

本内容は「入門」としているため、内容を一部割愛して説明している部分があります。また、全ての取引関係を網羅する事はできないため、本内容とは異なる見方や解釈をする事もできます。