経路爆発を考える

Internet Week 2011

NEC BIGLOBE, Ltd.
Seiichi Kawamura
kawamucho at mesh.ad.jp

経路って何だ？

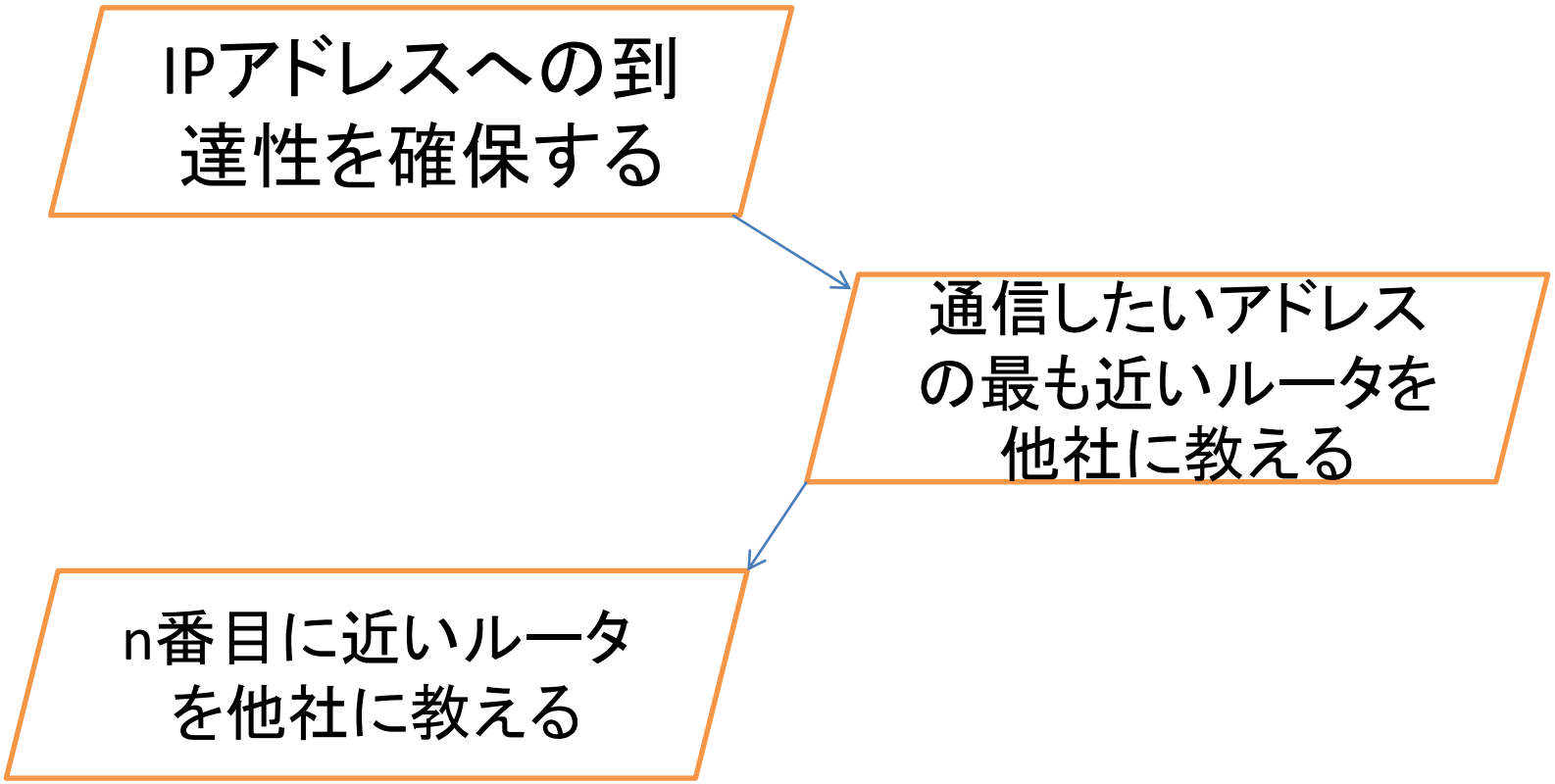
IPアドレスへの到達性を確保する

経路って何だ？

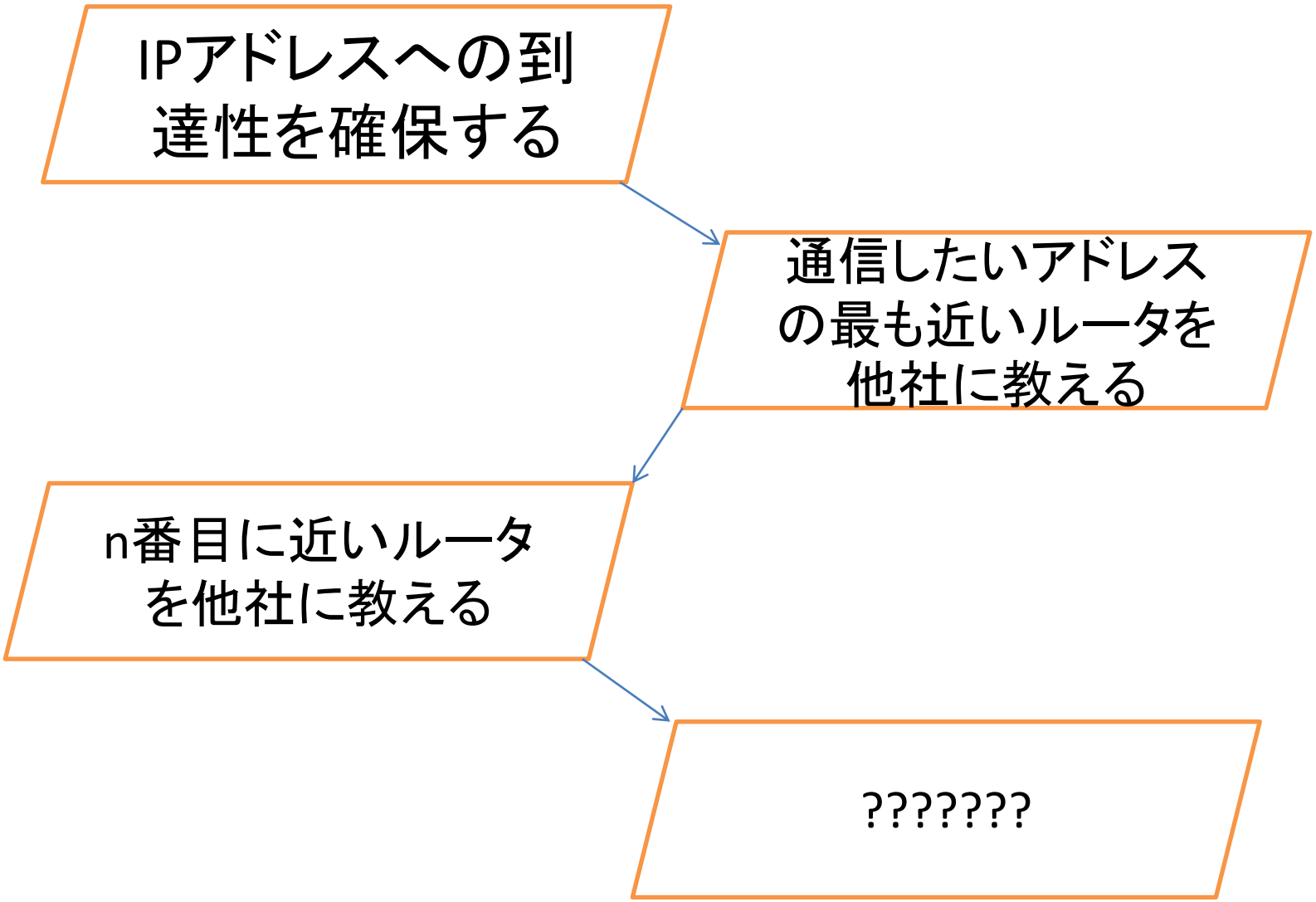
IPアドレスへの到達性を確保する

通信したいアドレスの最も近いルータを他社に教える

経路って何だ？


IPアドレスへの到達性を確保する
通信したいアドレスの最も近いルータを他社に教える
n番目に近いルータを他社に教える

経路って何だ？


IPアドレスへの到達性を確保する
通信したいアドレスの最も近いルータを他社に教える
n番目に近いルータを他社に教える
???????

IPアドレスと経路のおさらい

- IPアドレスの流れ
  - RIR(またはNIR)→LIR(ISPなど)→顧客
- ルーティングする事業者と、IPアドレスを取得する事業者は、一致していなくてもよい
  - ネットワーク運用をアウトソースしているなど
- ルーティング運用とIPアドレス運用は管理が独立している
  - IPアドレスは階層型
  - ルーティングは自立分散型

権威をもった存在というものは「仕組み的には」存在しない

経路とASの関係

- 「AS番号は、自律ネットワークを運営する組織がインターネットにおける外部経路制御を行うことを目的とし利用するために、自律ネットワークに付与されます。」
  - http://www.nic.ad.jp/doc/jpnic-01089.html

経路とASの関係

- 「AS番号は、自律ネットワークを運営する組織がインターネットにおける外部経路制御を行うことを目的とし利用するために、自律ネットワークに付与されます。」
  - http://www.nic.ad.jp/doc/jpnic-01089.html
- ASを運用する、という事はインターネットの経路を受け取り、そしてインターネットに経路を広報する「呼吸活動」の事を意味します

経路とASの関係・・・つまり

- 呼吸が乱れる、という事はASの運用にとって致命的であると言えます

正しく息を吸って
正しく息をはく

- これが経路運用の大前提と言えます
- なぜ「大前提」なのか？

インターネットの安定と自社の事業の関係

- インターネットが安定した運用を続けられる事が、自社のお客様へ安定した良いサービスを提供するための前提だからです

インターネットの安定と自社の事業の関係

- インターネットが安定した運用を続けられる事が、自社のお客様へ安定した良いサービスを提供できる大きな前提だからです
- それができない/気にしたくない人は
- 経路運用をしなくてもよいのです
  - AS番号を持たなくてもいい
  - ネットワークはアウトソースしてもいい

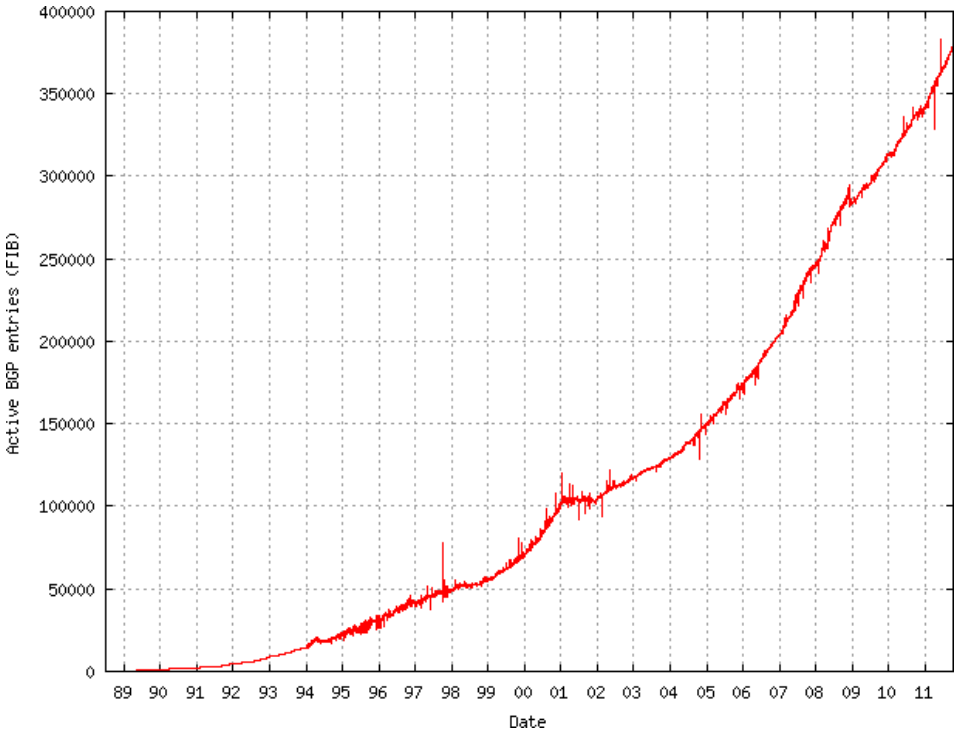
本日のお話は

- 一生懸命インターネットに接続するASを運用されている方
- 内部経路、外部経路の呼吸活動を一生懸命がんばっている方
- これから頑張ろうとしている方
- そんなみなさまに、経路の驚異的な増大が何故起きるのか、今後どういう事を考えながら運用することが望ましいのか、についていっしょに考察してみたいと思います
- 今後のみなさまの運用業務、設備投資計画、技術設計計画のお役に立てれば幸いです

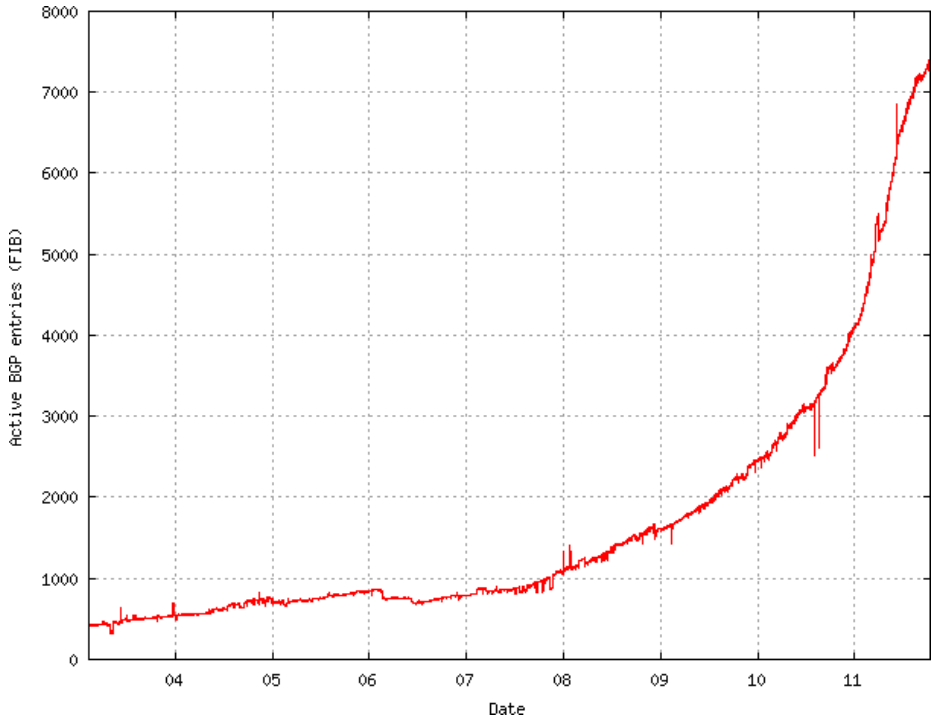
自己紹介

- NECビッグローブ株式会社勤務
  - 2001年 NEC入社
    - 2004年まで営業SE（インターネット関係なし）
  - 2004年からBIGLOBEに移籍
    - 2008年まではIPv6/VPN担当
    - 2008年ごろからIPv6/コアネットワーク運用担当
    - 2011年から運用現場を離れIPv6、NW設計、Peering担当
- JANOG（日本ネットワークオペレーターズグループ　運営委員）
- 標準化（IETF）活動、APNICなどに出没します

問題提起



IPv4フルルート増加



IPv6フルルート増加

出展 : http://www.cidr-report.org/as2.0/ by Geoff Houston

将来予測

- 2011/11月時点で
  - IPv4: 383K
  - IPv6: 7.5K
- 現在のペースで増えた場合
  - 2013: 486K(IPv4)/19K(IPv6)
  - 2015: 616K(IPv4)/49K(IPv6)

不確定要素

1. IPv6
   - ルーティングポリシーは絶えず変化している
     - 今は/48よりも長いPrefixはルーティングしないのが一般的
     - トラフィックが増加するとともに、ポリシーは変化する
   - 割り振りポリシーも絶えず変化している
     - 簡単に複数Prefixを保持できる時代がくるかもしれない
2. IPv4
   - ルーティングポリシーはIPv6同様変化する
     - 今は/24よりも長いPrefixはルーティングしないのが一般的
   - IPアドレス移転制度がもたらす影響は未知数
3. 先進国の経済成長低下
4. 成長国の経済成長スピード

心構え

- 予測値の1.2~1.3倍を見ておく事が望ましい
- 状況変化には常に気をつける必要がある
- 重要な要素として位置付け、行動する

具体的にどうするか、はPart2でお話します

経路が増加する事で、何がおきるか？

- ルータの処理能力を超える:

  – コストインパクト
- 経路収束が遅くなる:

  – 品質インパクト
- 運用が複雑になる:

  – コスト、品質インパクト

経路増大についての様々な「意見」

- 前項の影響があるため、みんなが苦労する。みんなもっと責任もって運用するべき
- Moore's Lawに従えばハードウェア面は問題ではない。むしろ高級な装置を必要とする事で参入障壁になるから競争に有利になる
- 人間が運用している以上、経路が増えるのは仕方ない。発生する問題を軽減する方向を技術で解決する方法を探すべきだ

※NANOG、JANOGなどで意見交換して聞いた生の声です

正しい答えはありません

でも、Internetなんです

- 「ネットワーク」は相手があってこそ、上手く通信ができ、利益を上げることができ、そしてユーザが満足することができる
  - 経路増大は全てのASに影響します
- 同時に、慈善事業ではなく、市場原理も適切に働かなければいけない
  - コスト的に説明できる事をやらないといけない
- 少なくとも、経路増大で「良い」事はあまり無さそうです。
  - たくさんの経路を処理できる事は新規参入障壁にはならなそうです

経路増大については、全てのASが
一人称で考える必要がありそうです

経路増大はどうしておきるのか

1. 運用目的のDeaggregation（経路の分割広告）
2. 設定ミス/設計ミス
   - iBGP流出
3. マルチホーム
   - 検討不足のマルチホーム
   - トラフィックエンジニアリング

参考：

- **RIPE Routing Working Group Recommendations on Route Aggregation**
  - http://www.ripe.net/ripe/docs/ripe-399
  - 日本語訳：http://www.janog.gr.jp/doc/ripe399.pdf

CIDR前の時代の経路が分割広告されている事がありますが、「増大」はしていないので割愛しています

運用目的のDeaggregation

- レジストリから割り振られたPrefixや上位ISPから割り当てられたPrefixを、細かく分割してInternetに向けて広告する行為
- 良く耳にする理由
  - 経路ハイジャック対策
    - たとえば/24で経路を広告しておけば、誰かが間違って同じ/24を広告しても、最悪同点。/23までなら勝てるから影響がない
    - 分割する事で影響範囲が限定できる

運用目的のDeaggregation

- 良く耳にする理由・・・続き
  - ノイズ対策
    - たとえば、10.1.0.0/16をRIRから割り振られた場合
    - 10.1.1.0/24と10.1.2.0/24しか利用していない状況で
    - 10.1.0.0/16を広告すると、10.1.3.0-10.1.255.0の間は未使用だが、ProbeやScanなどのトラフィックがネットワークに入ってくる
    - ルータの処理負荷を下げるために分割広告している

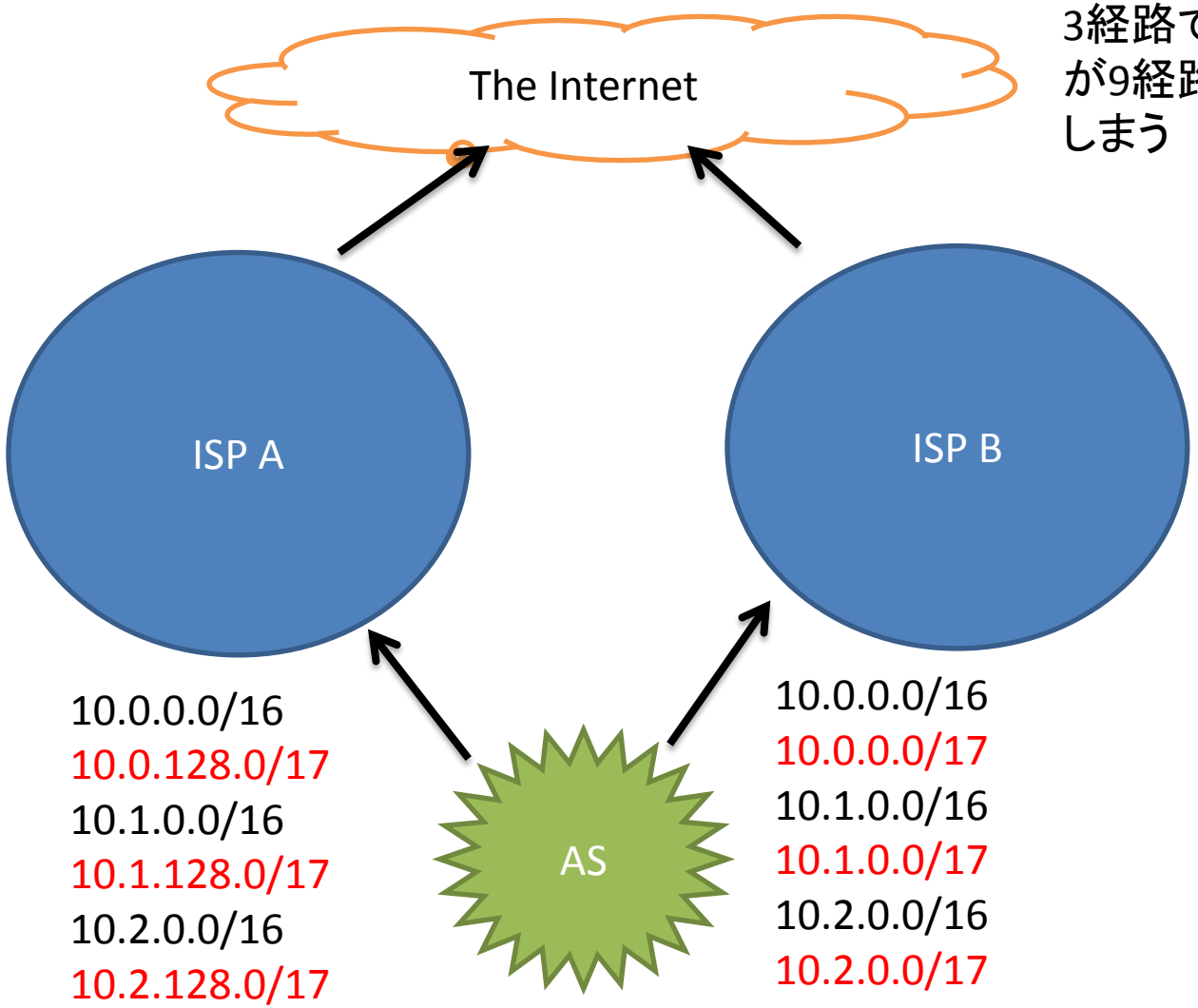
設定ミス/設計ミス

- 内部経路(一般的にはiBGPで使っているもの）がそのままeBGPにのってインターネットに出て行ってしまう
- ミスしても、到達性は確保されるため実影響が出ないので気にしない人がいる
- 本来数経路で済んだものが何十経路にもなる

マルチホームと経路増大

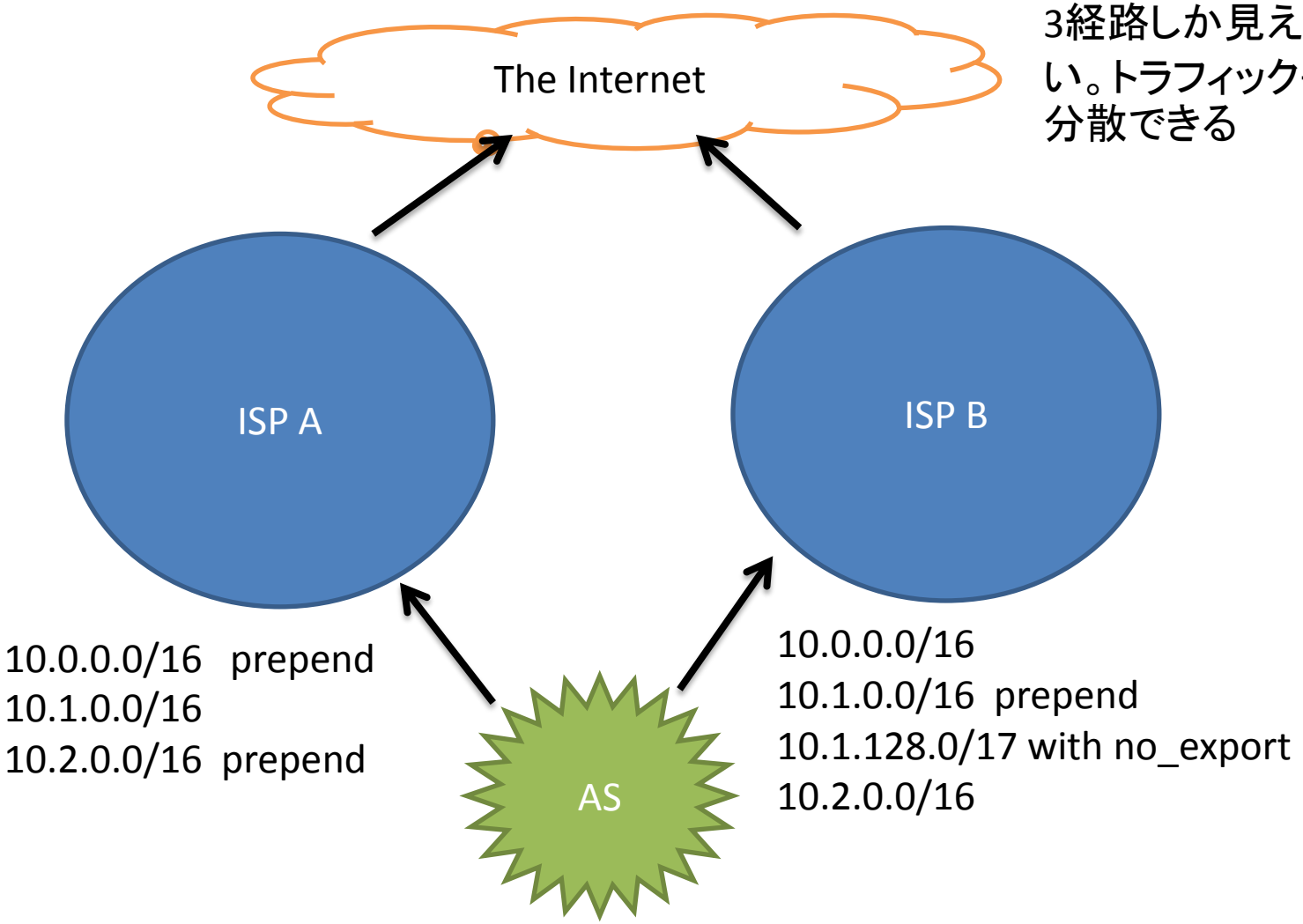
- マルチホームには様々な手法がありますが、ここでは経路増大に関連するものだけ紹介します
- 上流ISPを分割して、冗長性を確保する場合、もしくはトラフィックを分散させて支払う金額をコントロールしたい場合に、経路を分割して上流に広告する場合があります

マルチホームの例



3経路ですむはずが9経路も見えてしまう

マルチホームの例：改善してみる



健全な経済活動による経路増大

- 内部経路の増加
  - 顧客（PAアドレス配布）が増えると経路は増える
  - 大きなプロバイダでは50k～150k経路かかえる場合がある
    - http://tools.ietf.org/html/draft-narten-radir-problem-statement-05
- 事業売却など
  - 稼働中のシステム/サービスを売却すると、アドレス毎譲渡する事になる場合がある
  - 経路はコマ切れになって、譲渡先に渡されることも

時事的な要素による経路増大

- IPv4アドレス枯渇の影響
  - 現在APNICにIPアドレスを申請するとどの事業者も最大/22が1つもらえる
  - APNIC地域だけで数万経路になる
  - IPアドレス移転制度
    - IPアドレスを部分的に他社に移転する事ができる
    - IPアドレス枯渇対策としてとても有効な策だが、ルーティングにとっては不都合な仕組み
  - http://www.nic.ad.jp/ja/ip/ipv4transfer-log.html

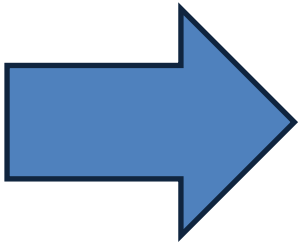
経路増大の影響はどれだけ深刻か

- 運用しているルータの処理能力
  - ラインカード（FIB）に512K経路しか保持できないものが大量に出回っている
  - 新しいものでも、IPv6/IPv4デュアルスタックで運用すると1M経路持てるものが512Kしか持てなくなる場合がある
- 経路収束
  - フルルートを持つルータが突然落ちた場合、経路の収束までに1分以上かかるケースがある
  - たとえばNTT（AS2914）とKDDI（AS2516）にマルチホームしていて、AS2914がメンテナンスに入った場合、2516側に切り替わるまでに1分以上通信が断になる場合もある、ということ
  - 単純な計算式はない

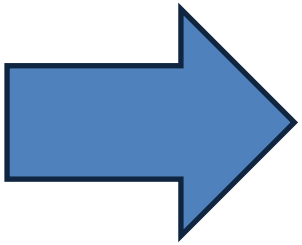
運用現場にとっては、見えない時限爆弾

経路増大の影響はどれだけ深刻か

- まだ生まれてきていないルータたちの未来
  - 経路情報はDRAMにストアされる
  - DRAMのlatency性能向上は年7%程度*



IPv6の伸び率、IPv4の伸び率は、7%を大幅に上回っている！！！**



将来のルータ性能は、経路爆発のスピードにおいつけなくなる！！！

* Tony Li氏 Future of the Internet 講演より
** Tony Li氏によるとIPv4は年12.6%, IPv6は60.2%

無策ではありません

- 適切な知識、適切な対応、計画的な見通しを持つ事により、影響は最小限におさえる事ができます
- 何よりも、AS運用者として正しいマナーを守ることは、インターネットの運営に関わる企業としてはとても大事な原則です
  - *Robustness Principle: Be liberal in what you accept, and conservative in what you send.*

対応は、運用者とメーカーがいっしょになって考えていく必要があります

＜参考＞CIDR Reportでの情報

NANOGなどに、週1度レポートメールが投稿されている
以下、deaggregation factor部分のTop10を切りだしたもの

ASnum	NetsNow	NetsAggr	NetGain	% Gain	Description
Table	379447	222433	157014	41.4%	All ASes
AS6389	3549	224	3325	93.7%	BELLSOUTH-NET-BLK - BellSouth.net Inc.
AS18566	1916	383	1533	80.0%	COVAD - Covad Communications Co.
AS4766	2498	978	1520	60.8%	KIXS-AS-KR Korea Telecom
AS22773	1460	110	1350	92.5%	ASN-CXA-ALL-CCI-22773-RDC - Cox Communications Inc.
AS4755	1543	243	1300	84.3%	TATACOMM-AS TATA Communications
AS4323	1629	395	1234	75.8%	TWTC - tw telecom holdings, inc.
AS1785	1839	784	1055	57.4%	AS-PAETEC-NET - PaeTec Communications, Inc.
AS28573	1378	323	1055	76.6%	NET Servicos de Comunicao S.A.
AS19262	1394	400	994	71.3%	VZGNI-TRANSIT - Verizon Online LLC
AS7552	1393	418	975	70.0%	VIETEL-AS-AP Vietel Corporation

http://www.cidr-report.org/as2.0/ を参考に編集

Part II
運用視点での対応

運用者がとれる対応

- まずは「見る」こと
- しかし、日常業務でネットワーク運用をやっていても、「ルーティングは今どういう状態か」という事を気にするケースはかなり少ない
- 通常大きな問題はそんなに起きない
  - 年に1度くらい大きな経路ハイジャックなどがあるくらい

しかし、今後適切な事業継続のためには同じような運用では思わぬ障害、品質低下、コスト増を招いてしまう

運用と管理～健康診断～

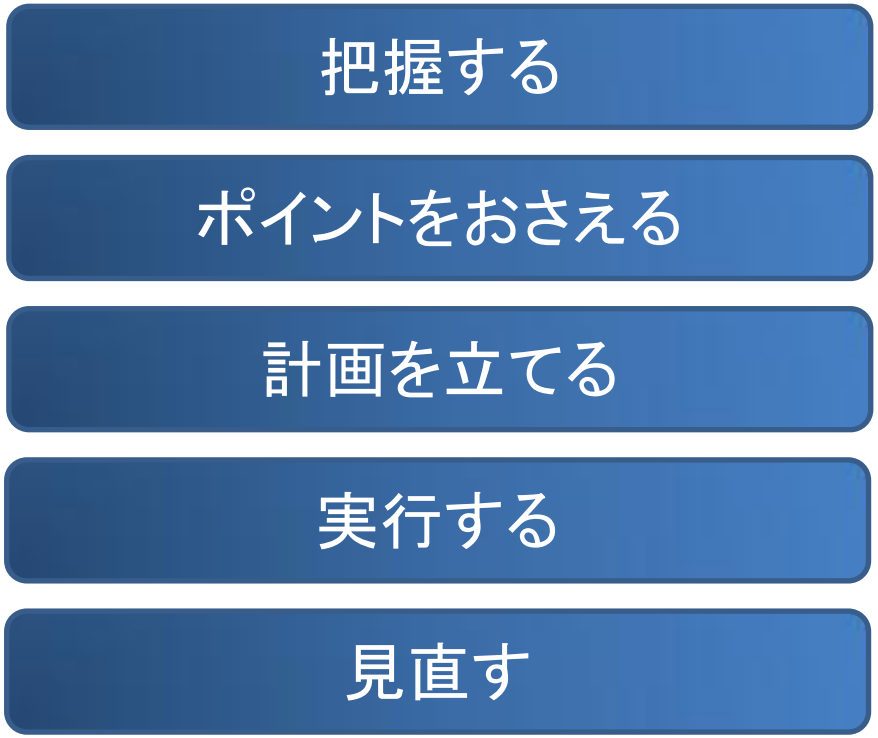
- 自身のネットワーク内のフルルートは何経路だろうか？
  - フルルート＝Internet Full route + 内部経路
- 内部経路の増加状況は把握できているか？
- Internet Full routeの変遷はすぐにわかるか？
- 運用中各ルータのFIBの大きさ（＝持てる経路数）は把握しているか？
- 実際にルータのFIBの仕様率を把握しているか？

具体的な取り組みポイント

1. 運用と管理～平和な日常の中で～
   - 稼働中ルータの「状態」
   - ルータの選定
2. 経路広告～conservative in what you send～
   - 経路設計からルータのコンフィグに落とし込む
3. 経路受領～liberal in what you accept？～
   - 到達性確保のための設計
   - 「状態」を見てポリシーを設計

経路増大問題に取り組もう！

- PDCAのようなサイクルを立てましょう



運用と管理：稼働中ルータの状態

重要

- 稼働中ルータのラインカードは、何経路保持する事ができるか把握しよう
  - ベンダーさんに確認しよう！
    - CiscoのEngine4+はv4 onlyで500K経路
    - 500KものをDual Stackで使ってると来年あたりキケン
    - IPv6の経路数*2をIPv4に足すと感覚としては近い（ルータ毎に制限は異なるが一般的な感覚として）
  - 設定として変更可能か確認しよう！
    - ASR9000は標準で500K経路しか持てない
    - 設定を投入する事で増やす事ができる

運用と管理：稼働中ルータの状態

best effort

- 実際何経路もっているのか、定期的に確認しよう
  - 増減トレンドを把握する
  - http://www.cidr-report.org/as2.0/やroute-viewsで確認できる経路数と、自分の持っている経路数はズレている
    - BIGLOBEでも、IPv4で5%、IPv6で10%ずれていることもある(どこと比較するかにも依存）
  - やり方は次のページで簡単に紹介

ポイント！

経路の管理

- 定期的に目で見る
  - BGP Route Reflectorに全ての経路が集約される構成なら、Route Reflectorの持っている経路数をみる
    - 目で見るのに慣れると、障害対応の時の「感」を育てる事ができる
- システムで管理する
  - SNMP（BGP4-MIBv2はまだ標準化未完了。Juniperは一部実装済み）
  - リモートから定期的にshow ip bgp/ show bgp の結果をとりに行き保存する（ツールを自作する事になる）
  - Route ReflectorまたはRoute Reflectorからフルルートを吐き出すクライアントにソフトウェアルータを導入し(ZebOS, vyattaなど）内部HDDに経路数を保存する
    - ZebOSにはそういう機能が搭載されている
    - Quaggaのコンフィグ例

```
dump bgp updates /home/kawamura/route/update/update.%Y.%m.%d.%H%M 15m
dump bgp routes-mrt /home/kawamura/route/dump%Y.%m.%d.%H%M 60m
```

運用と管理：ルータの選定

普通

- ルータ選定の際は、経路保持の仕組みについて確認しよう
  - 最大経路数の確認は重要（設備の実質の償却を決めてしまう）
  - 経路収束を早める機能の搭載を確認
    - BGP PIC/indirect next-hop
    - デフォルトでONになっている場合もある
    - 将来的に助けてくれる機能の一つ
  - フルルート受信にかかる速度を評価する
    - 経路切り替えの時の断時間に影響を与える

ポイント！

フルルート受信の切り替え

- external側のルータで、フルルートを受けているPeerを落とす（落ちる）と、Route Reflectorから大量に経路を受信する必要がある場合がある
- この間、FIBに経路が存在しない事になる

最も避けたい状態

経路が増えるほど遅くなる

- できるだけ早く受信してFIB（ラインカード）に反映できる事が望ましい
- BGPの経路送信は、若番から送っていくケースが多いようで、22x.x.x.xの経路は復旧まで比較的時間がかかる

ポイント！

運用としての対応

- Peerを落とす時には、まずBest Pathを切り替える、という作業を実施する
  - Local Preferenceを落とす
  - MEDを上げる

優しさあふれる運用

- ベストパスで無くなるので、FIBから経路が無くならずに多くの経路を切り替える事ができる（全部は無理）
- 経路が多い状態でも、断時間を短くできる
- ただし、メンテナンスしやすい設計にしないといけない
  - 例：Juniperでは1つのpeer groupでまとめる複数のPeerに対して異なるexport policyを設定していると、export policyを変更するためにはPeerを落とさなければならない
  - フルルートを受けるPeerはpeer groupから独立させる

経路広告:ポリシーの決定

重要

- 経路広告ポリシーを決めよう
  - フルルート顧客向け、Internet向けそれぞれにポリシーを策定
    - Internet向けには極力集約した経路を設定する
    - IGPを間違って出さないように、/24よりも長い経路はフィルタする(IPv6は/48くらい)
    - 集約経路を生成する
    - RouteViewsなどで、自分の経路がどう見えているか確認する

(IPv4)/24より長い経路について
現時点では到達性が無いと思ってよい。しかし、IPアドレス枯渇で/24より長い経路が到達する必要性がでてくるかもしれない。

ポイント！

誤った経路広報の末路

- NANOG53でのこと
  - 「Bell Southの運用者はただちにマイクに並んで、現状のルーティング状態を説明しなさい」
  - 500人の前でこのような事を言われてしまう

- 他者に悪影響を及ぼすような行為は信頼を損ね、事業に影響を及ぼす場合もある

受信ポリシーは、経路爆発に対しての効力は実はあまり無いが、AS運用のためには重要な要素

経路受信：ポリシーの決定

重要

- 経路受信ポリシーを決めよう
- いくつかのタイプがある
  - ① 到達性を確保することに重点を置く
  - ② 本来不要と思われる経路はすべてフィルタをしてインターネットのあるべき姿に重点を置く
  - ③ 明らかに不要と思われる経路はフィルタするが到達性確保には重点を置く
- 顧客向け、Transit向け、Peer向けそれぞれでポリシーが必要

到達性を確保する手段:デフォルト

- デフォルトルートが一番到達性を確保できる
- http://www.janog.gr.jp/meeting/janog24/doc/ST3.pdf
  - 7割近くのASはデフォルトルートを持っている
  - デフォルトルートを持ちつつ、インターネットルートも持つという運用をする人もいる
- デフォルトルート一本だと、経路爆発は無関係！
  - ただし、トラフィックエンジニアリングには無理が生じる
  - 世の中の動きがよくわからなくなる
  - Net flowデータ解析で経路連動の分析ができなくなる
- デフォルトルート無しで運用できる事が本来はベスト
  - そもそもASを運用する必要がない

受信経路フィルタ

厳しくフィルタ

- 基準を明確にしてフィルタする
  - IRRに登録されてる経路のみ許可
  - RIRから割り振られたサイズのみ許可
- インターネットのあるべき姿に近いが、到達性に問題が出てくる可能性が高く運用が難しい
- 保持する経路は少なくなる

バランス型

- /24より長い経路をフィルタ、Martian（private空間やTest-Netなど特殊用途アドレス）経路をフィルタする程度にとどめる
- ほとんどセキュリティ目的で、経路数とは無関係
- 日本ではポピュラーな運用スタイル

フィルタしない

- 顧客からの経路をフィルタしない場合、不要な経路をインターネットにばらまくことになる可能性がある
- あまり責任を持った運用とは言えないが、ほとんどフィルタしていない事業者も海外では結構見かける

ポイント！

顧客からの経路受信

- マルチホームの例でみたように、BGP顧客からは分割された経路を受ける事がある
- Communityを用意しておく事で、顧客のDeaggregationがインターネットに伝搬してしまうのを最小限に抑える事ができる
- トランジット提供者の適切なコンサルティングで経路爆発問題は軽減する可能性がある

おまけ：内部経路設計

- 大きな事業者になると、内部経路だけで数十万を超えて無視できなくなる大きさになる
- このような事業者はごく少数
- 内部経路は、「事情」と「案件」によって増大していくため、初期設計をしっかり立てていても無茶な経路増加を回避できない事はある
- 大きな事業者へと成長する過程での経路再設計は難しいが、OSPFのエリア分け、集約経路の生成など努力できるポイントはある
- ポイントは、サービス毎のルーティングドメイン/インスタンス分けの検討、集約性の考慮、見直しするタイミングを設けるというところにある

まとめ

- 「なるべく経路を受けない」ようにする事は難しい（到達性の問題に遭遇する）
- 一方で、状態を把握し、設備投資や評価/設計に反映させる事は万が一を防ぐ事はとても重要
- 自分が加害者にならないようにする事はできる
  - 送信経路の適切な設計は、AS運用者としての重要な責任の一つ

まとめ

- しかし、経路が増大し、ハードウェアの性能が追い付かなくなる可能性がある状態で、何年も先に品質が落ちていくというインターネットそのものの設計問題は解決されていない
- 今できる事は「運用でカバー」
- 影響要素の動向についてしっかりとフォローする
- 運用でカバーしつつ、装置メーカーのイノベーション、インターネット設計そのもののイノベーションに取り組む、という事が重要になってくる

まとめ

- 今我々が直面している課題は、インターネットがRobustである事に起因している
  - 甘い設計でもつながってしまう
  - 多少問題があってもつながってしまう
- しかしBGPの接続性をStrictにしていくとインターネットの強みは失われ、やがてそれは使えないものになっていく
- 経路爆発問題は、インターネット全体にとっての問題
- まず高い関心を持つだけでも一歩前進！

ありがとうございました