広域イーサネット網で本当にあったトラシュー話

- ENOG16 Meeting -

2012年8月31日

株式会社グローバルネットコア
金子 康行
<yasuyuki.kaneko@global-netcore.jp>

概要

- 弊社で運用している広域イーサネット網（というほど広くないけど）で経験したトラブルシュートネタです。
- ちょっと興味深い現象だったので、共有したら面白いかなと思って、あけっぴろげに発表します。
- とりあえずトラシューしたけど、実は根本的に解決できているわけではないのです。
- これどうしたらいいんだろうって未だに思い悩んでいるので、いいアドバイスお待ちしてます ;–）

Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

プロローグ

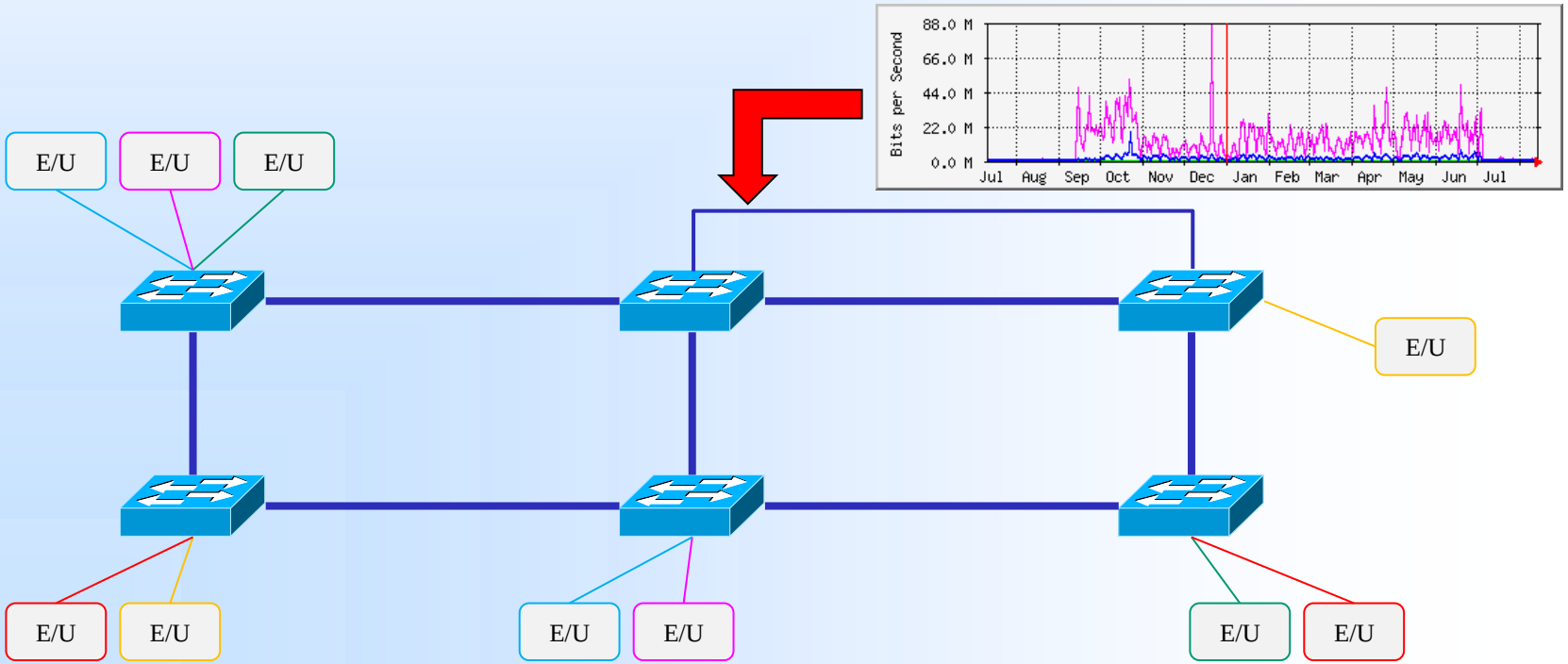
- ある日担当者はふと気がついた・・・
  - とあるスイッチ・とあるポートのトラフィックグラフ
  - そこはとあるスイッチ間のバックアップ用リンク
  - 「このリンクにデータが流れることはないはず」
  - でも、思いっきりトラフィックでてますね・・・

Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

事象を簡単に説明

- 舞台は**広域イーサネット網の基幹線**（基本的にすべてのVLANをトランク）
- **同一スイッチ間に複数リンク**が存在（キャリア冗長目的、異速度）
- 片側のリンクは**完全にバックアップ**の位置づけ（すべてのMST instanceで常時待機系）
- つまり、**このリンクにはほとんどフレームが流れない**はず

にもかかわらず、数十Mbpsのトラフィックを観測・・・なにこれ気持ち悪い



Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

余談・・・

- 実はしばらく放置してましたゴメンナサイ
  - 事象の観測は昨年9月から
  - たしかに気持ち悪いけど、さほど緊急性があるわけでなく・・・
  - まあなんとなく発生源は推測できてたんですけどね・・・
  - いろいろ業務が忙しくて・・・
  - ごめんなさい・・・

- で、夏になる前にようやく本格的にトラシュー着手
  - 今年7月に一旦事象は解消されました
  - 調べてなるほど、挙動がわかってすっきりはしました
  - 興味深い、レアケースな、他でも起こりうる、かなり悩ましい事例
  - というわけで、詳細は、次のページから

Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

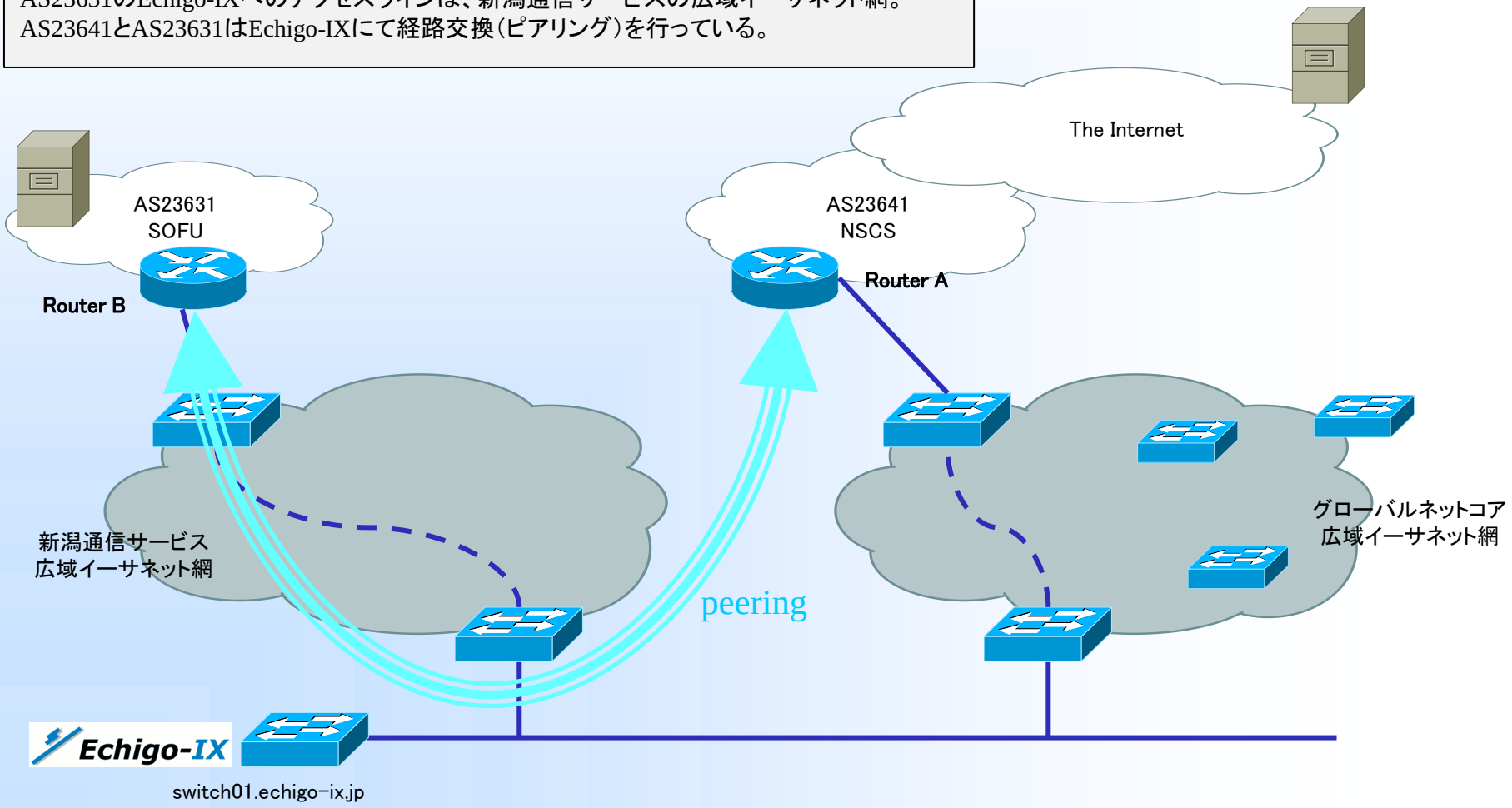
解説しよう！

- 舞台は当社の広域イーサネット網です。
- この事象の原因は、当社がNS・コンピュータサービス様に提供している専用線での通信に起因していました。
- この事象には、NS・コンピュータサービス様だけでなく、創風システム様、そしてEchigo-IXが絡んでいます。
- 誰も悪くないです。誰かのミスとかではないです。

Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

解説しよう！

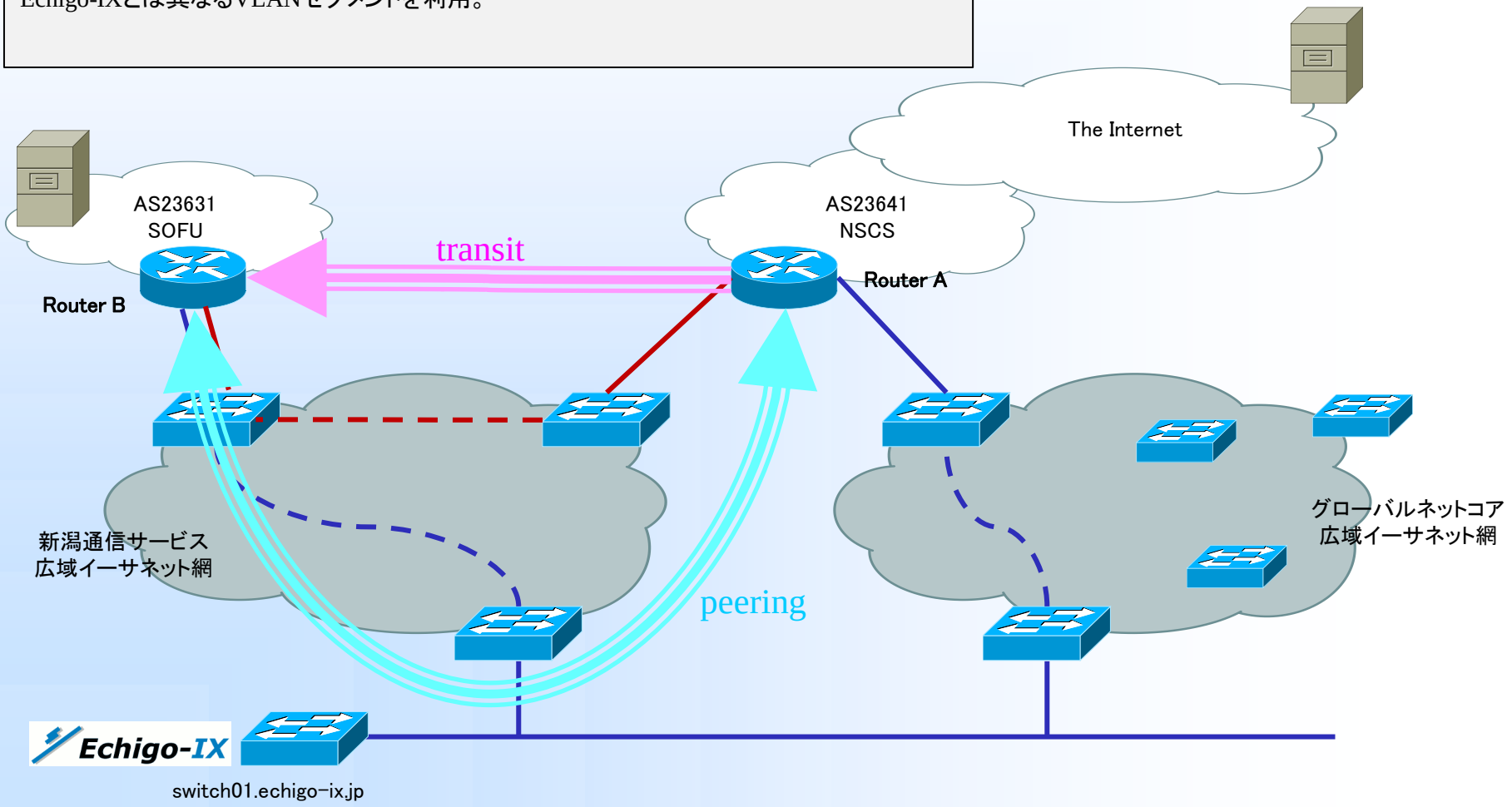
AS23641とAS23631はEchigo-IXに接続している。
AS23641のEchigo-IXへのアクセスラインは、グローバルネットコアの広域イーサネット網。
AS23631のEchigo-IXへのアクセスラインは、新潟通信サービスの広域イーサネット網。
AS23641とAS23631はEchigo-IXにて経路交換（ピアリング）を行っている。



Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

解説しよう！

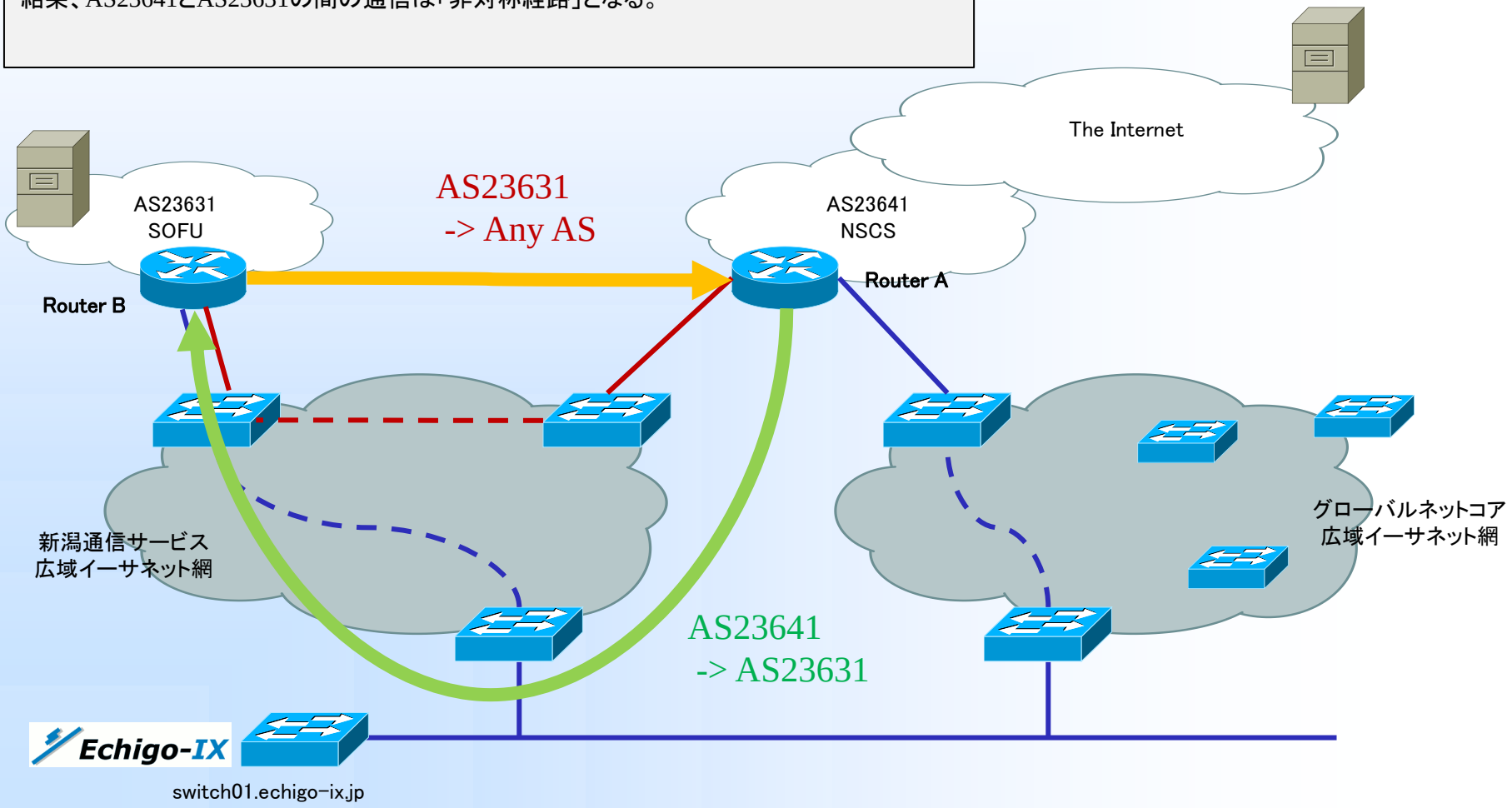
AS23641はAS23631に対して、トランジットを提供している。
トランジットのアクセスラインは、新潟通信サービスの広域イーサネット網。
Echigo-IXとは異なるVLANセグメントを利用。



Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

解説しよう！

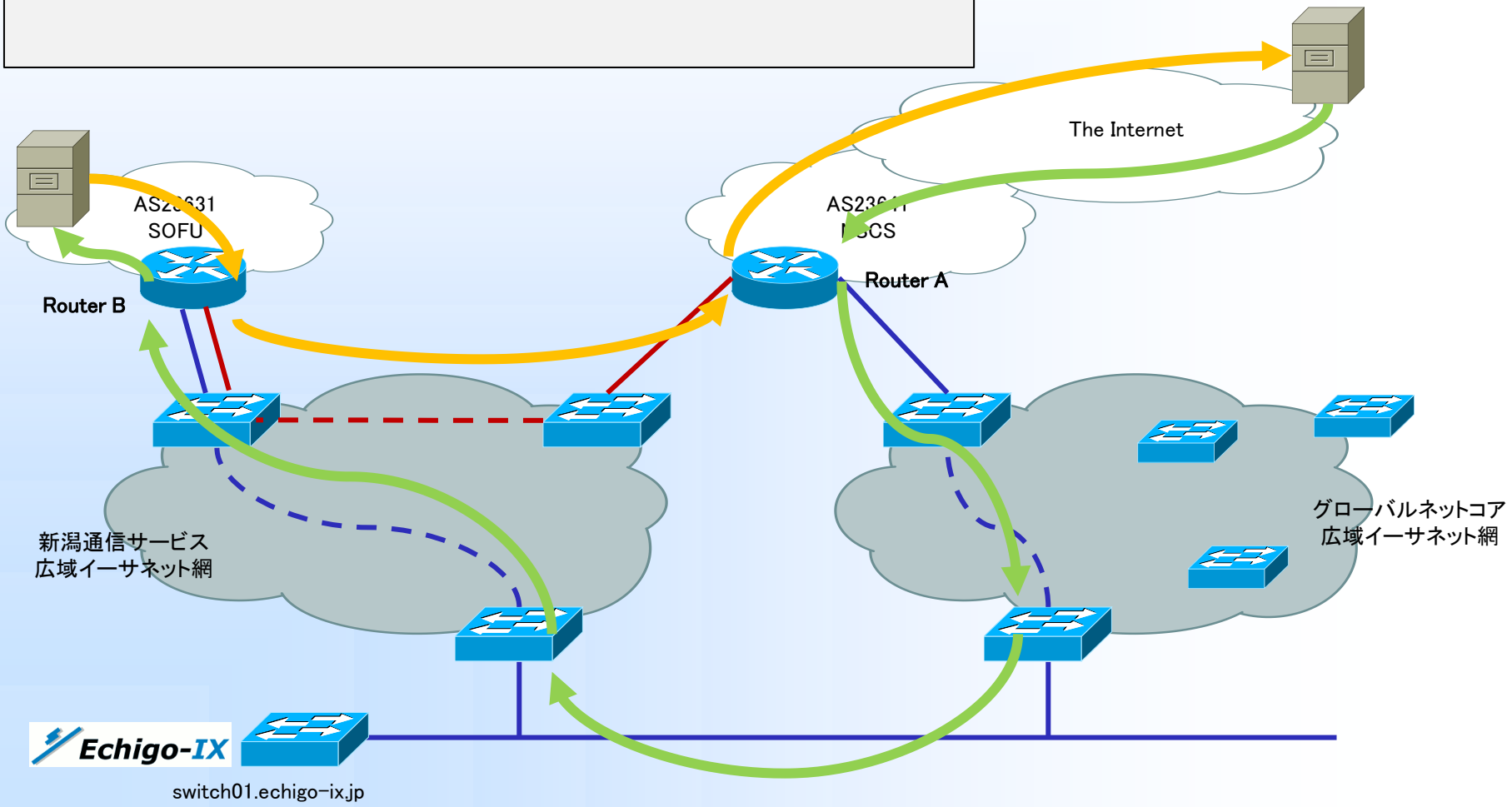
AS23641からAS23631への経路制御では、ピアリング経路が優先されている。
AS23631からAS23641を含むインターネットへの経路は、トランジット経路が優先されている。
結果、AS23641とAS23631の間の通信は「非対称経路」となる。



Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

解説しよう！

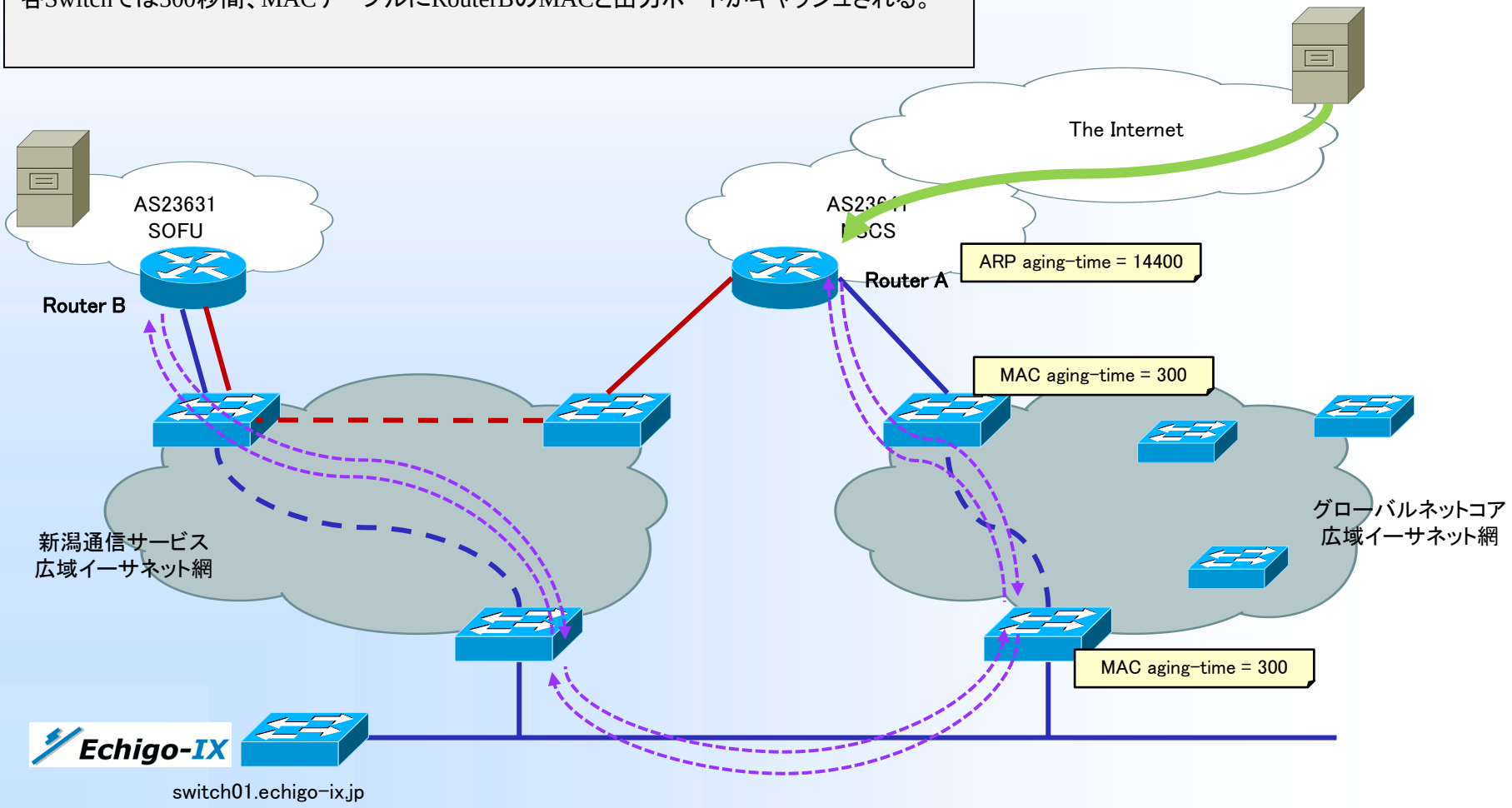
つまり、インターネットからAS23631への通信は、図のような経路を通ることとなる。



Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

解説しよう！

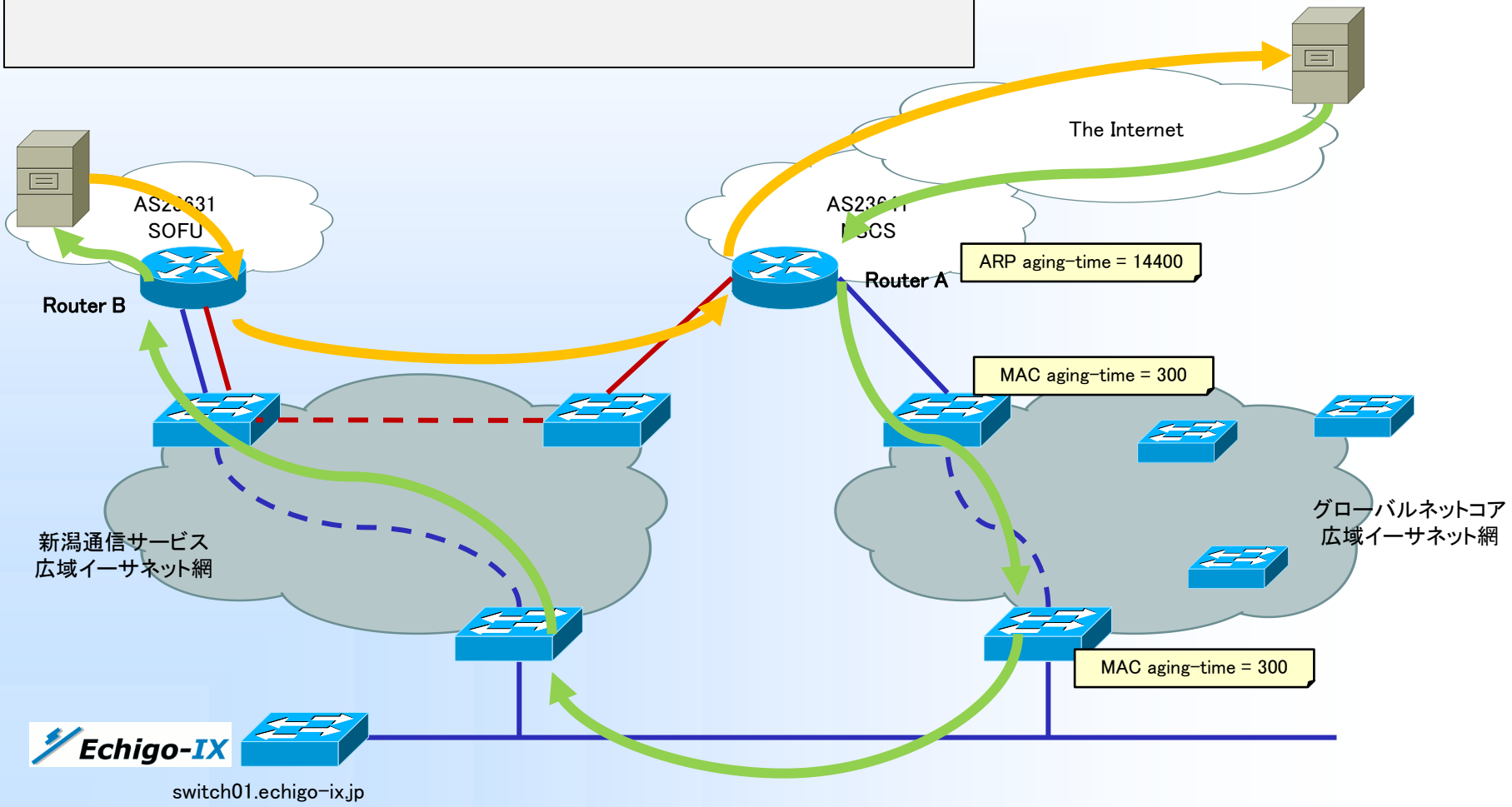
AS23641からAS23631への通信発生時、Router AはRouter BのMACをARPで取得する。
Router Aでは14400秒間、ARPテーブルにRouter BのIPとMACがキャッシュされる。
各Switchでは300秒間、MACテーブルにRouterBのMACと出力ポートがキャッシュされる。



Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

解説しよう！

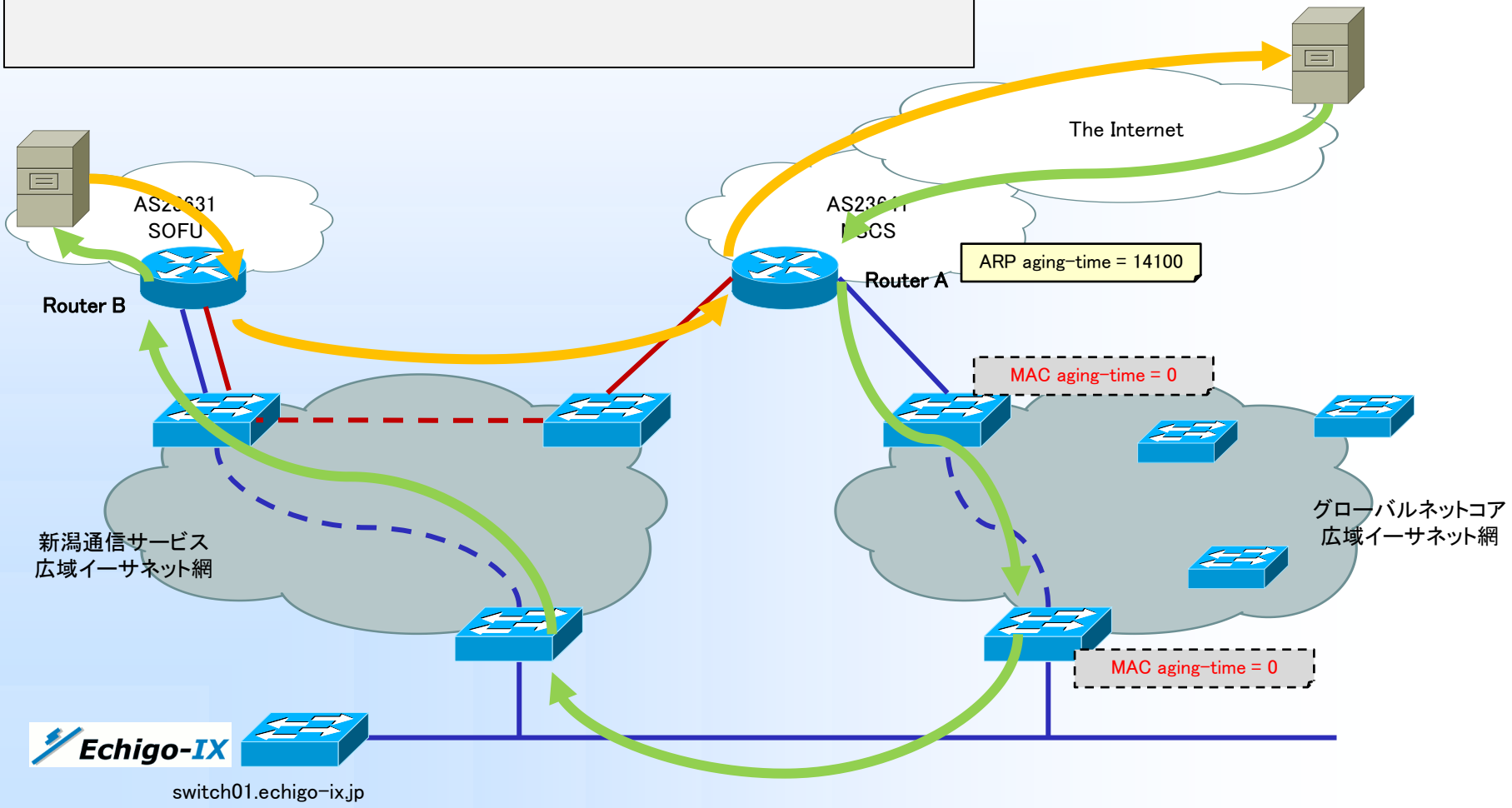
ARPから300秒間は図の通りの通信となる。
しかし・・・



Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

解説しよう！

ARPから300秒後、各スイッチのMACテーブルからRouter Bの情報が消滅する。
しかし、Router AのARPテーブルには、依然としてRouter Bの情報は保持されている。

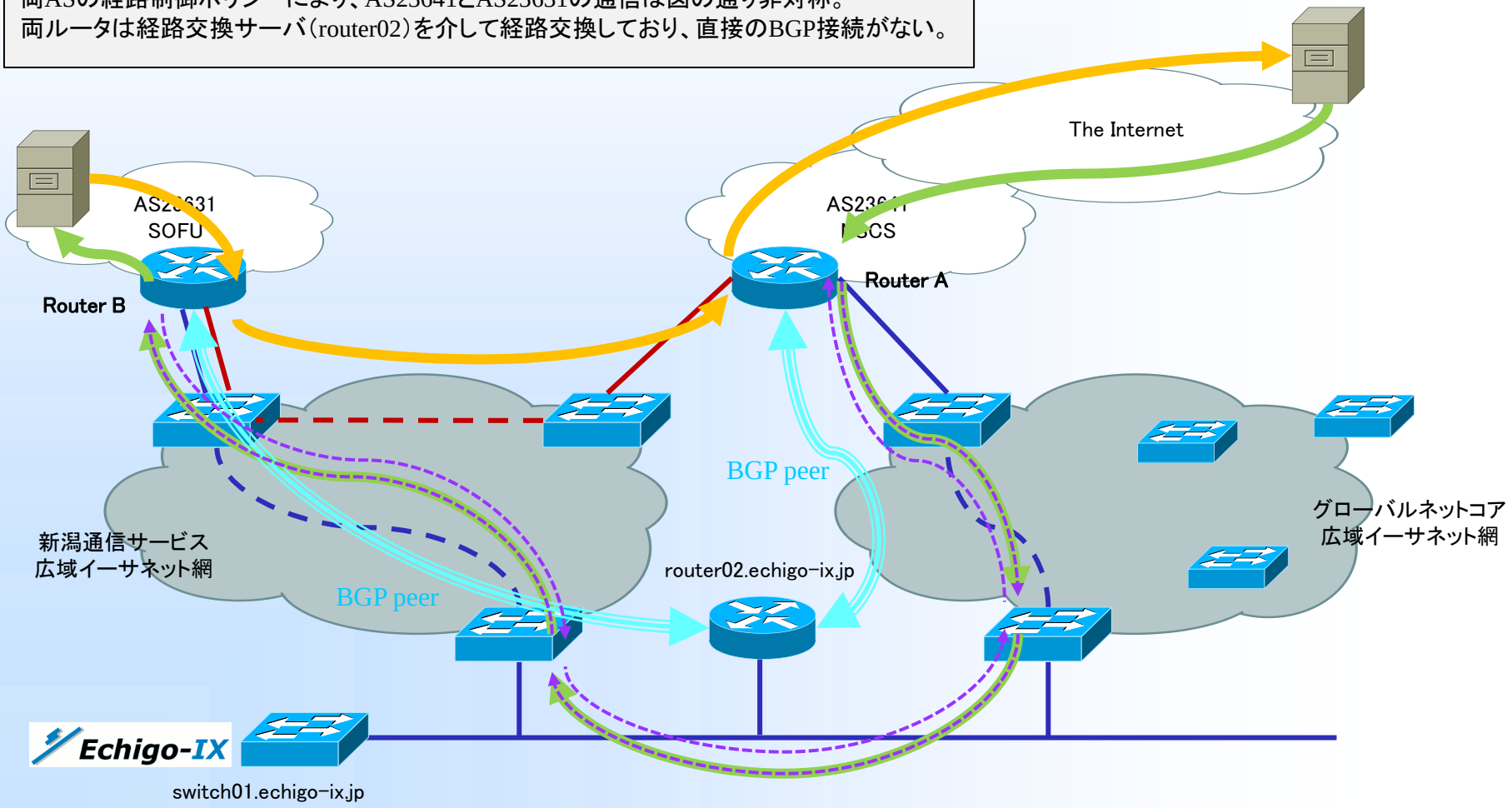


Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

解説しよう！

Echigo-IXでRouter BからRouterAへ向かう通信はARP応答のみ、他には一切発生しない。

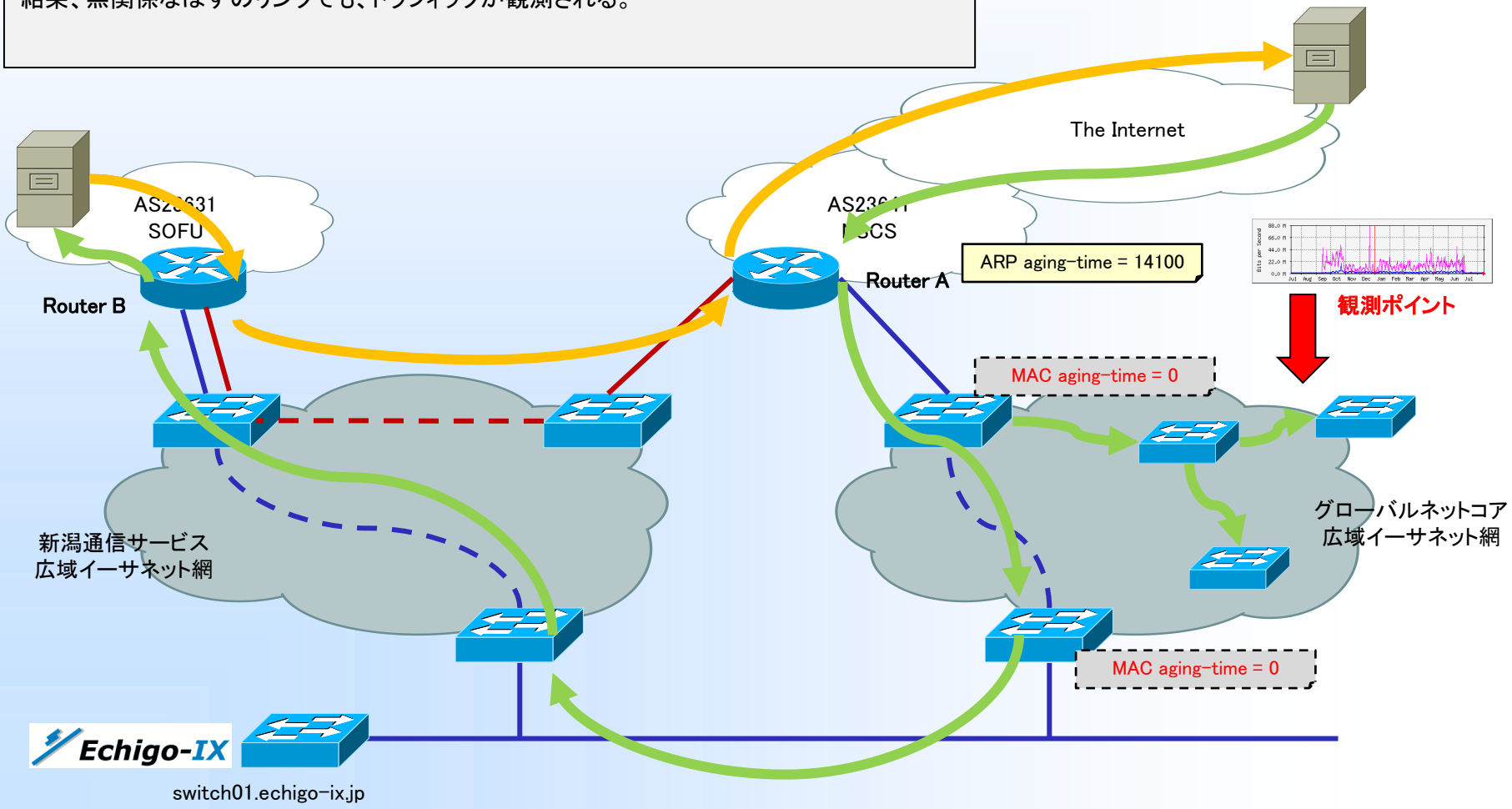
両ASの経路制御ポリシーにより、AS23641とAS23631の通信は図の通り非対称。
両ルータは経路交換サーバ（router02）を介して経路交換しており、直接のBGP接続がない。



Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

解説しよう！

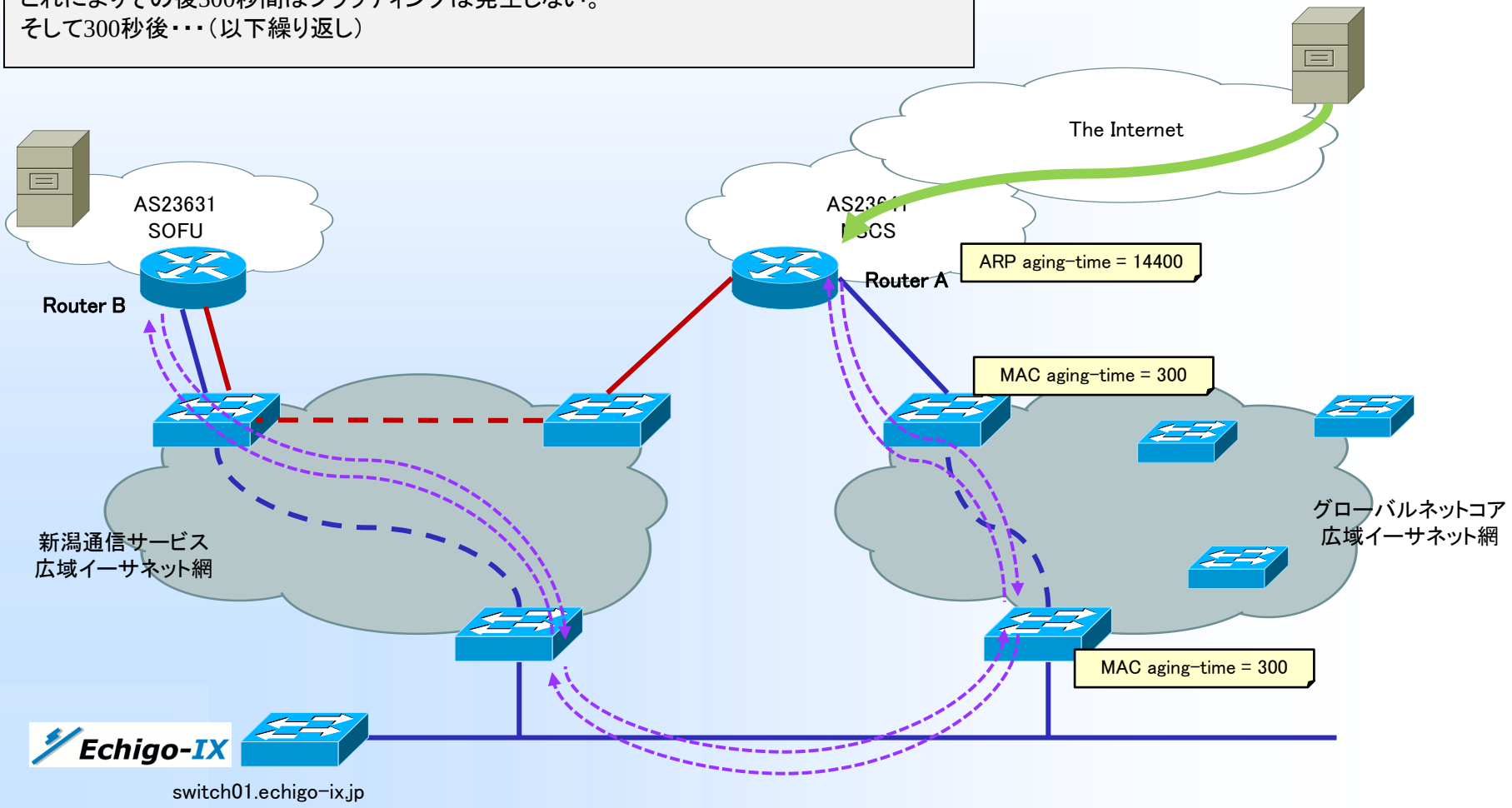
AS23641からAS23631への通信時、途中のSwitchではRouter Bの出力ポートが不明。
このため、Router Bへ向かうフレームは、各スイッチがフラッディングする。
結果、無関係なはずのリンクでも、トラフィックが観測される。



Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

解説しよう！

ARPから14400秒後（フラッディングが始まってから14100秒後）に再度ARPが発生。
Router BからのARP 応答により、各スイッチは再度MACテーブルに学習結果をキャッシュ。
これによりその後300秒間はフラッディングは発生しない。
そして300秒後・・・（以下繰り返し）



Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

結論

- 謎のトラフィックは、MACアドレス学習がうまくいかないことに起因して、網内のスイッチが継続的にフラッディングしていたフレームがその正体でした。

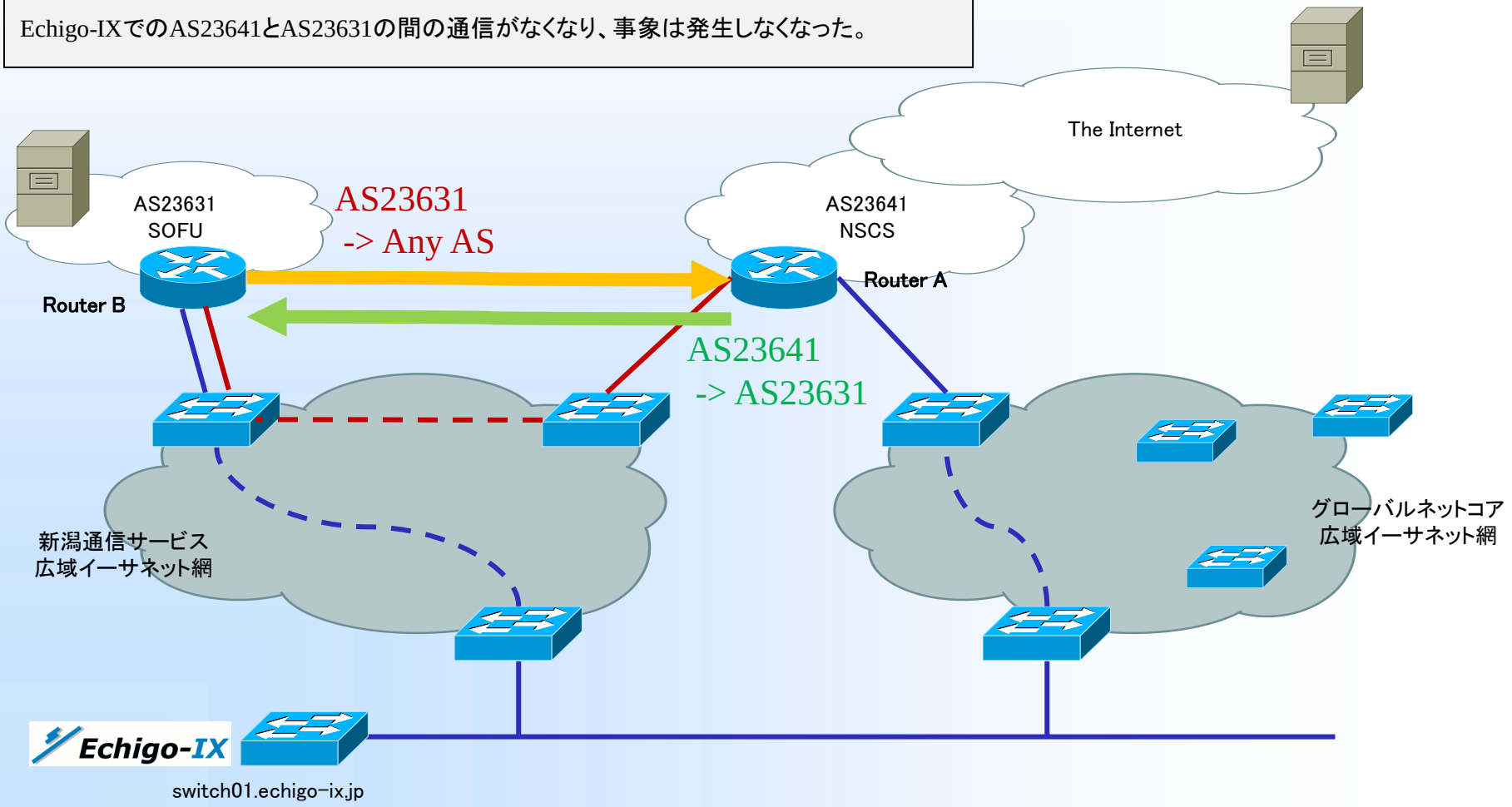
- もちろん、観測ポイント以外にも同じフレームが流れていたわけですが、いろんなフレームが継続的に流れているリンクではそのノイズは埋もれてしまうので発見できず、通常フレームが流れないリンクだったからこそ観測・発見できた、ということでした。

Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

現在の状態（トラシュー後）

AS23641からAS23631への経路制御で、トランジット経路を優先させた。
これにより、AS23641とAS23631の間の通信は「対称経路」となった。

Echigo-IXでのAS23641とAS23631の間の通信がなくなり、事象は発生しなくなった。



Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

考察＆ご意見頂戴

- ルーティングポリシーに関して
  - 私の中での一般的なルーティングポリシー
    - トランジットとIXピアの両方の接続がある場合、トランジッターは通常、IXより自らの顧客経路を優先する
    - 同様に、顧客はIX経路を優先する
  - でも実際のポリシーは違っていました（私の想定と逆だった）
    - AS23641はIX経路を優先していた
    - AS23631はトランジット経路を優先していた
    - 後学のために、ぜひともその意図が知りたいです・・・

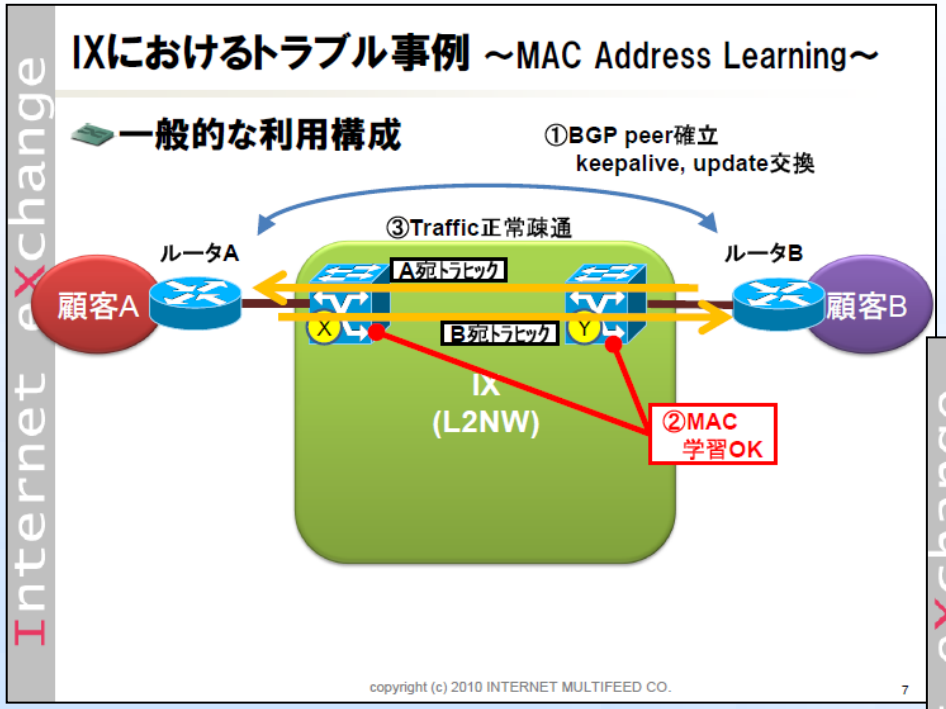
Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

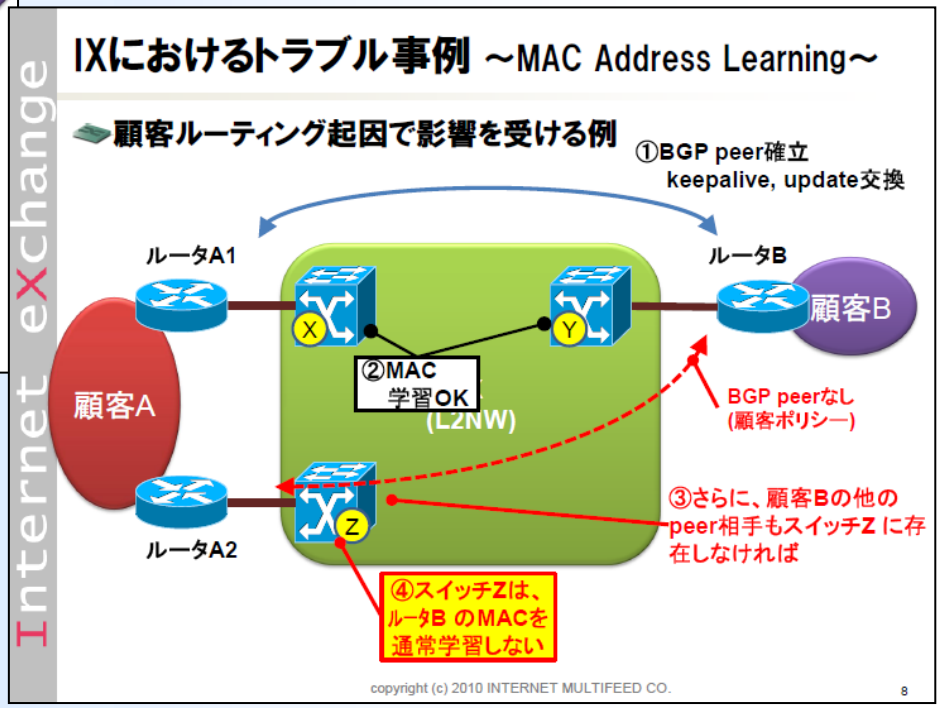
考察＆ご意見頂戴

- IXの運用に関して
  - IXの品質維持管理はこれでいいのか？
    - Echigo-IXの運用するスイッチはswitch01のみ
    - しかし、IXセグメントは広域イーサネット網に乗って各社まで広がっている！（そこにはスイッチがたくさんある）
    - IXの責任範囲はどこまで？L2セグメント全体を管理しなければ、本当の品質管理できないのでは？
  - ちょっと似たようなトラブルと運用の話が、JANOG26に！
    - IXPセッション ～頑張ってトラフィック支えてます～
    - http://www.janog.gr.jp/meeting/janog26/program/ixp.html
    - pingサーバという解決策

Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

JANOG26で発表されていた類似事例

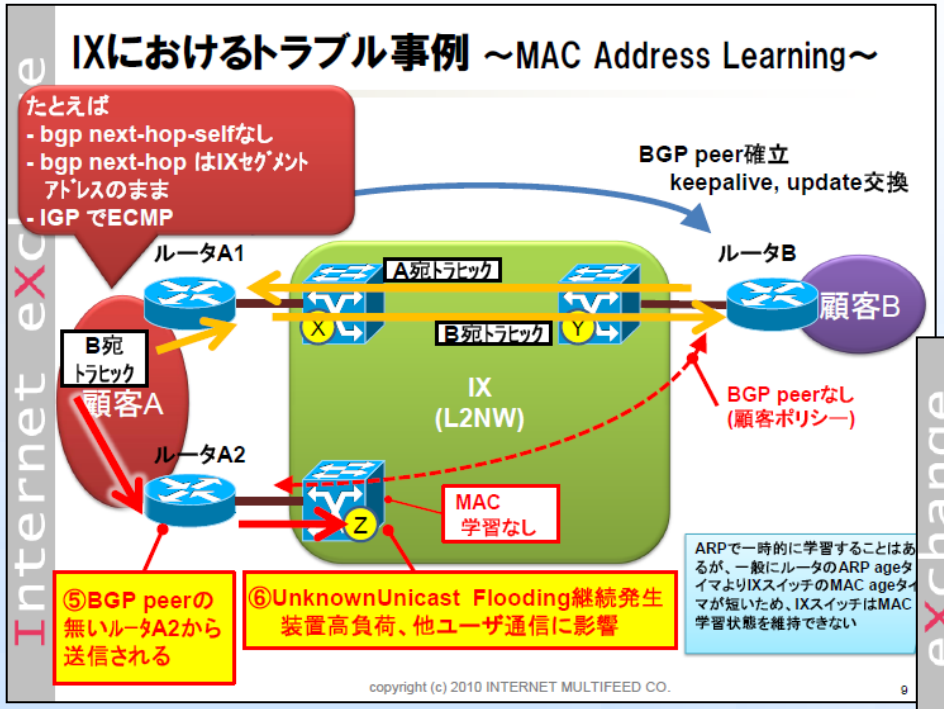


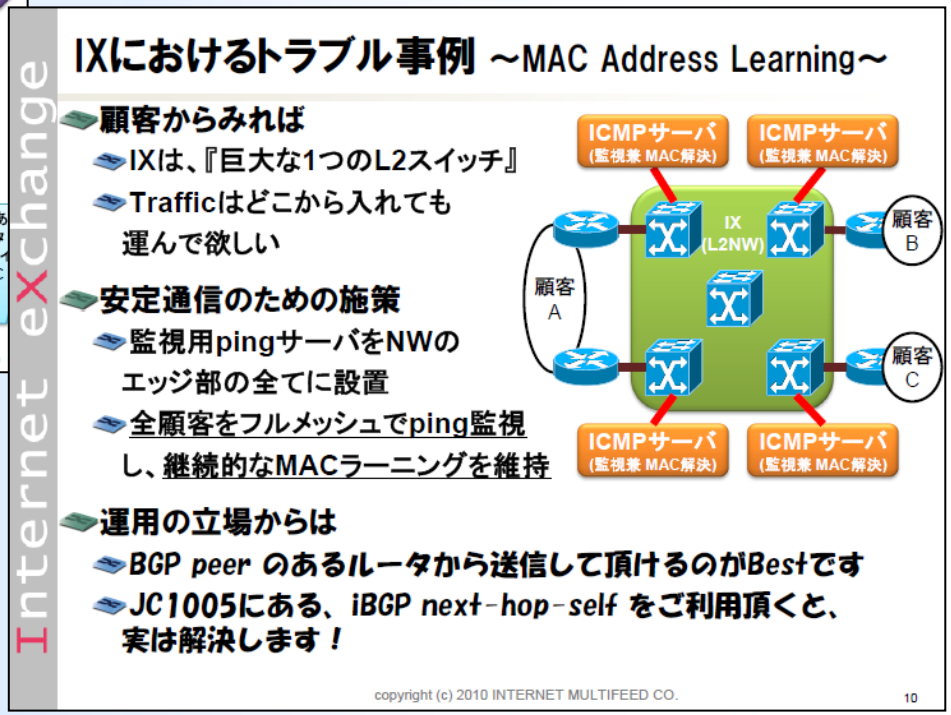


JANOG26 Meeting
IXPセッション ～頑張ってトラフィック支えてます～
任田大介氏(インターネットマルチフィード株式会社）資料より抜粋
http://www.janog.gr.jp/meeting/janog26/program/ixp.html

Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

JANOG26で発表されていた類似事例





JANOG26 Meeting
IXPセッション ～頑張ってトラフィック支えてます～
任田大介氏(インターネットマルチフィード株式会社)資料より抜粋
http://www.janog.gr.jp/meeting/janog26/program/ixp.html

Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

考察＆ご意見頂戴

- 広域イーサネット網の運用に関して
  - これって、めったにないけど、ありえる話
    - 今回はたまたま関係者だったのでご相談できたけど・・・
    - 本来は専用線を提供しているだけの立場なので、お客様の利用形態や構成に口を挟むことはできない
    - IXにおけるpingサーバのように、イーサネット網事業者側でなにか対策はできないものだろうか？

Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.

最後に

- NS・コンピュータサービス様、創風システム様、本件トラブルシュートにご協力いただき、ありがとうございました。
- また、今回ENOGの場で本事例紹介を行うことをご承諾いただき、重ねてありがとうございました。
- あくまでもこういった状況でこのような事象が発生する、という話で、広域イーサネットを運用する事業者としての当社の課題を発表したつもりです。上記2社様に落ち度があったわけではありませんので、あしからず。

Written by Yasuyuki Kaneko / Copyright (c) 2012 Global Network Core Co.,Ltd.