

コンテンツ配信コスト構造の 変化を目指して

吉野純平

自己紹介

- 株式会社ミクシィ
 - システム統括室 たんぽぽグループ
- 主な業務
 - コスト最適化
 - ネットワークの面倒をみる
 - クラウド活用
 - 社内のツール開発

私の正義

おいしいご飯を食べたい

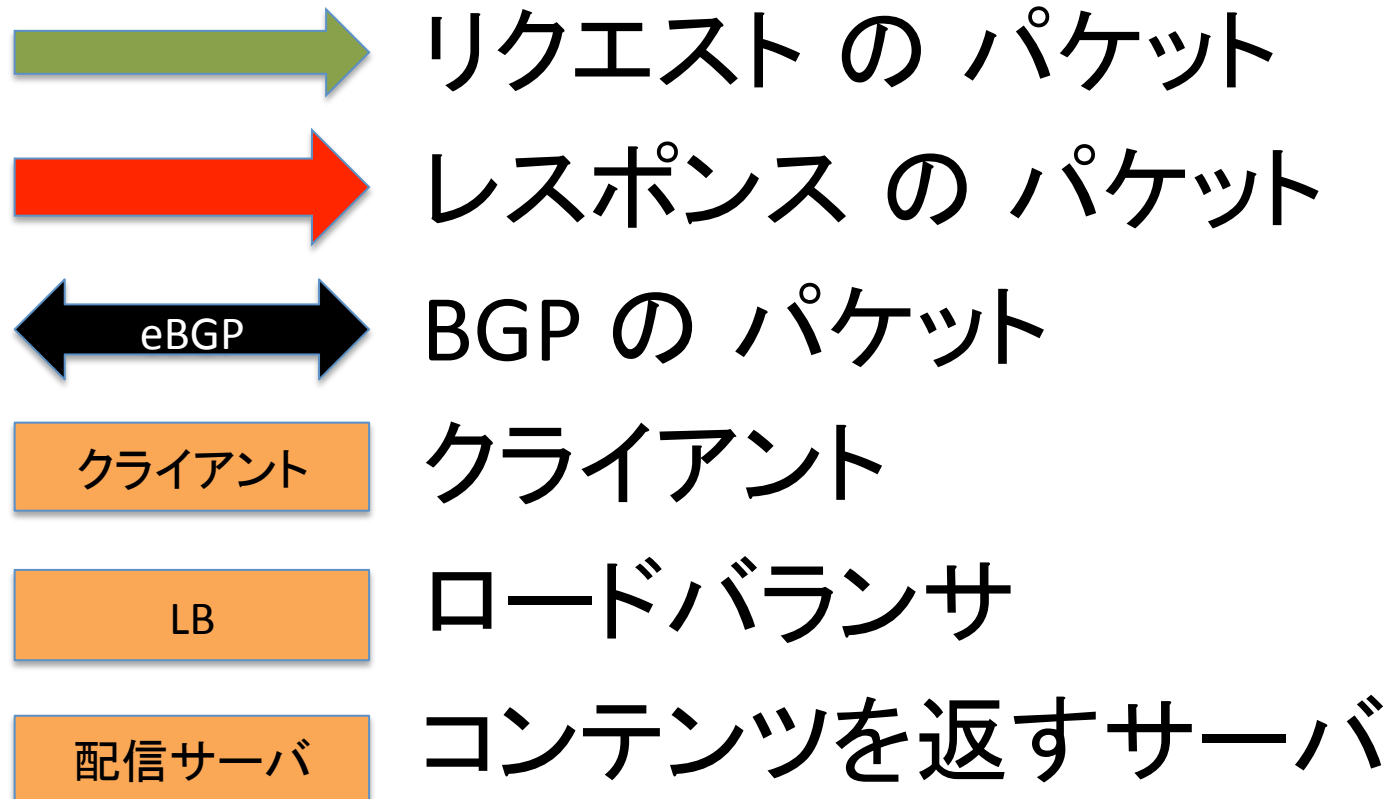
技術で利益を向上させる

今日の流れ

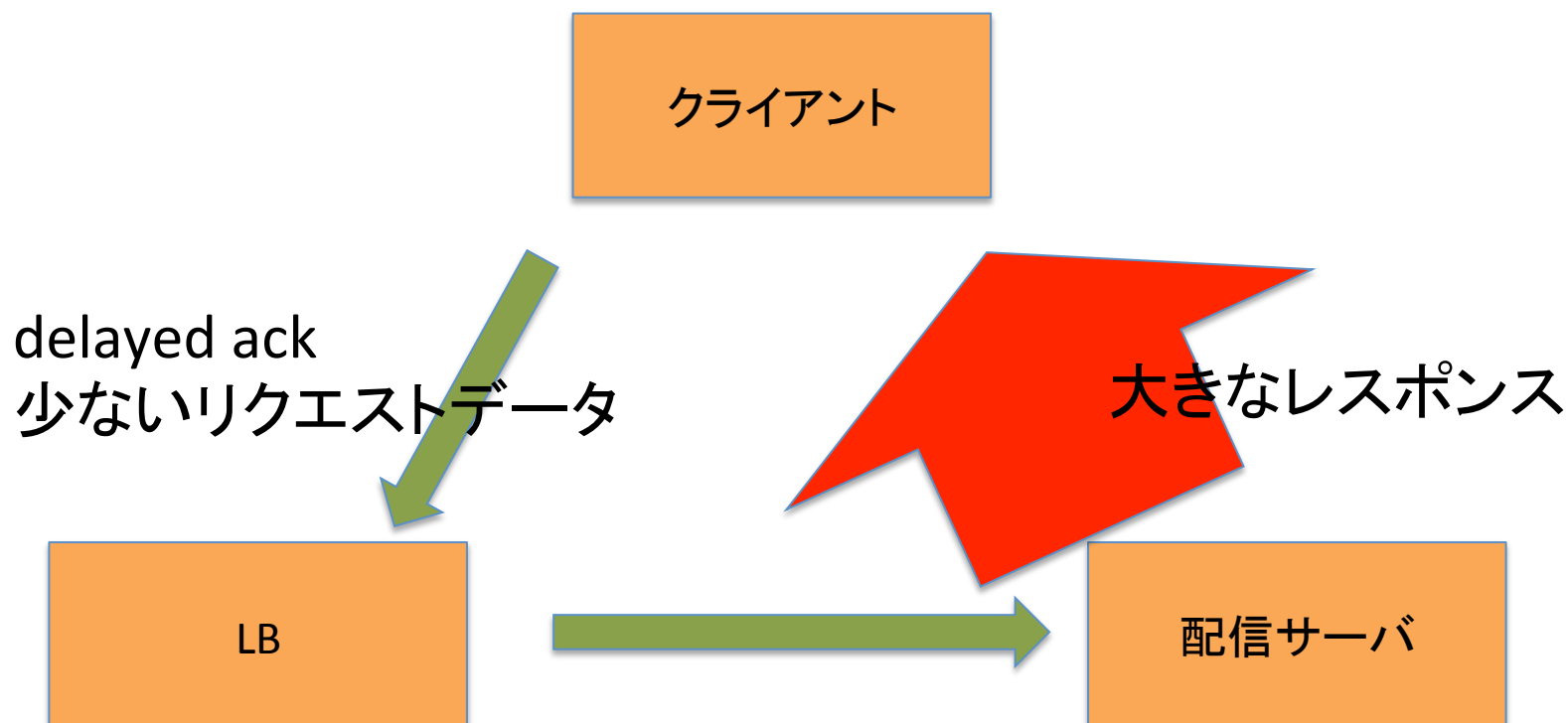
L3DSRで効果100倍！

トラフィックをもっと外へ！

前置き

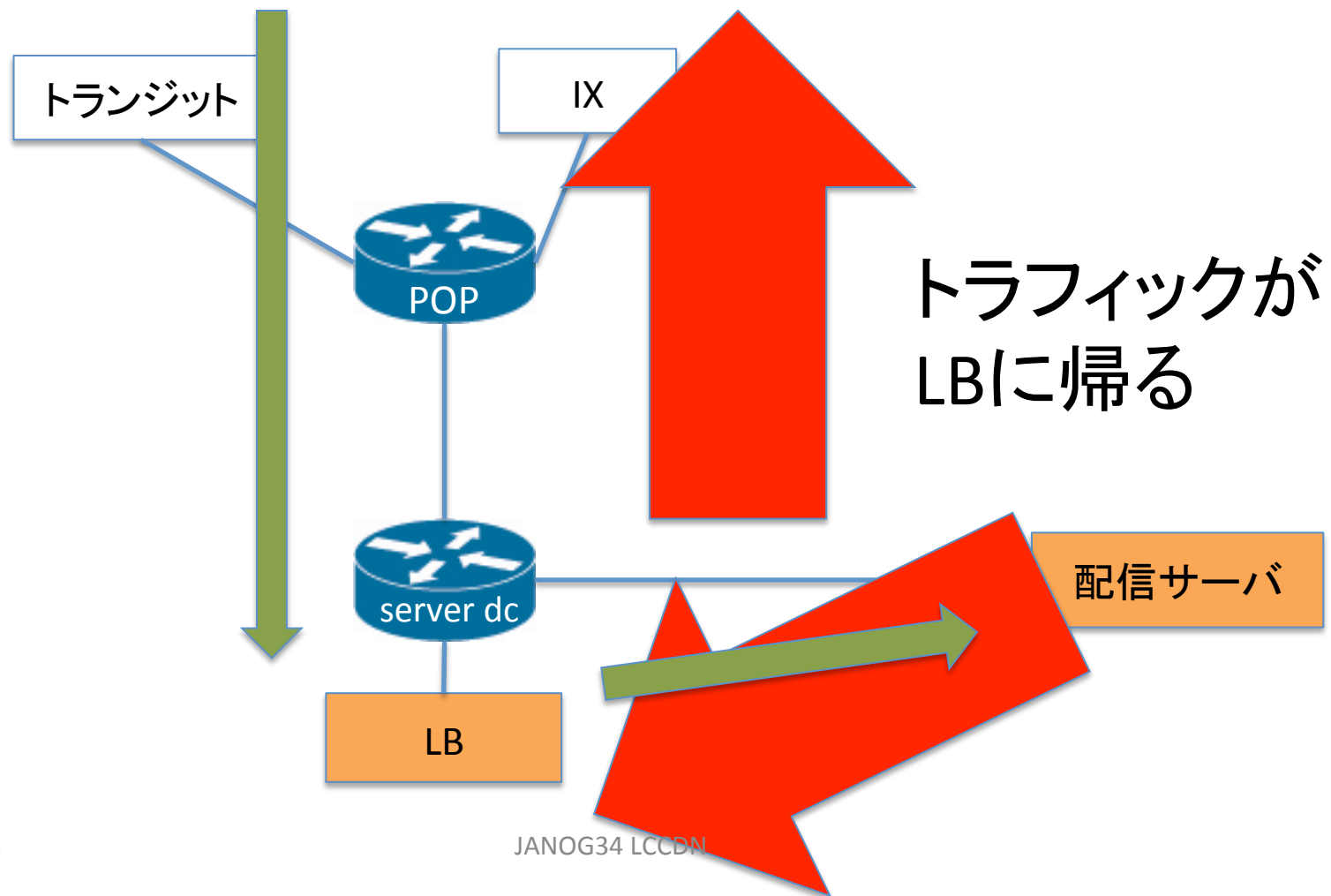


LBにレスポンスを通さないDSR

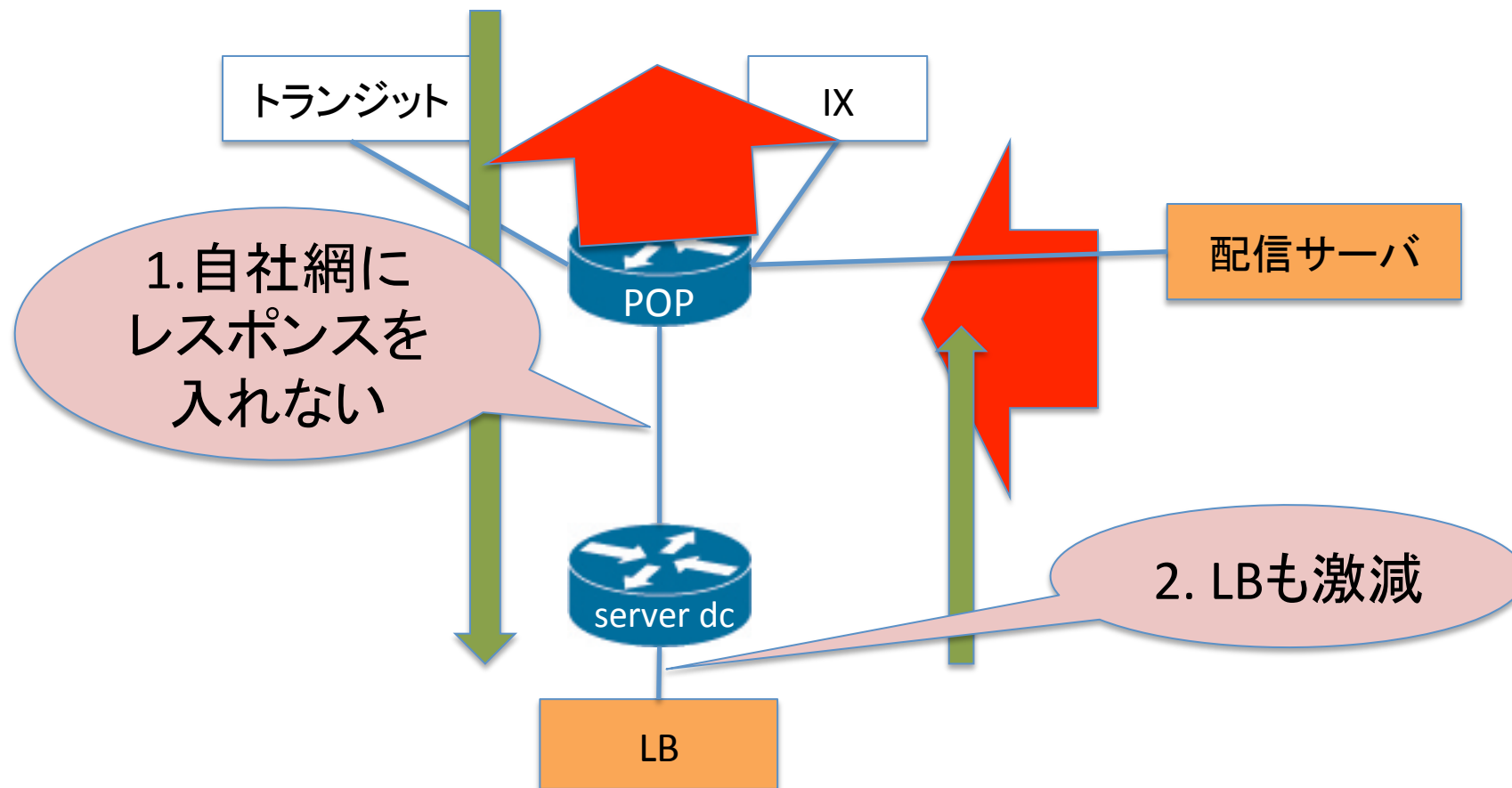


DSRは、JANOG32でも以下の問題を紹介
ヘルスチェックの問題
MTUの問題(トンネルを使った場合)

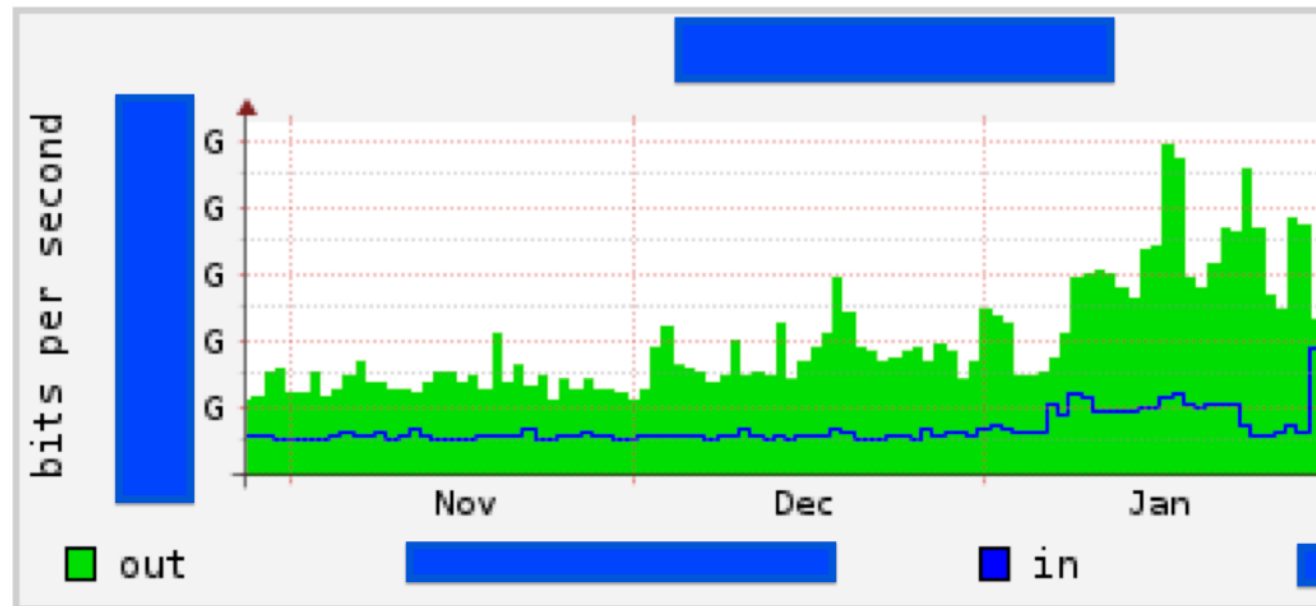
DSRしない例



物理への効率的なマッピング

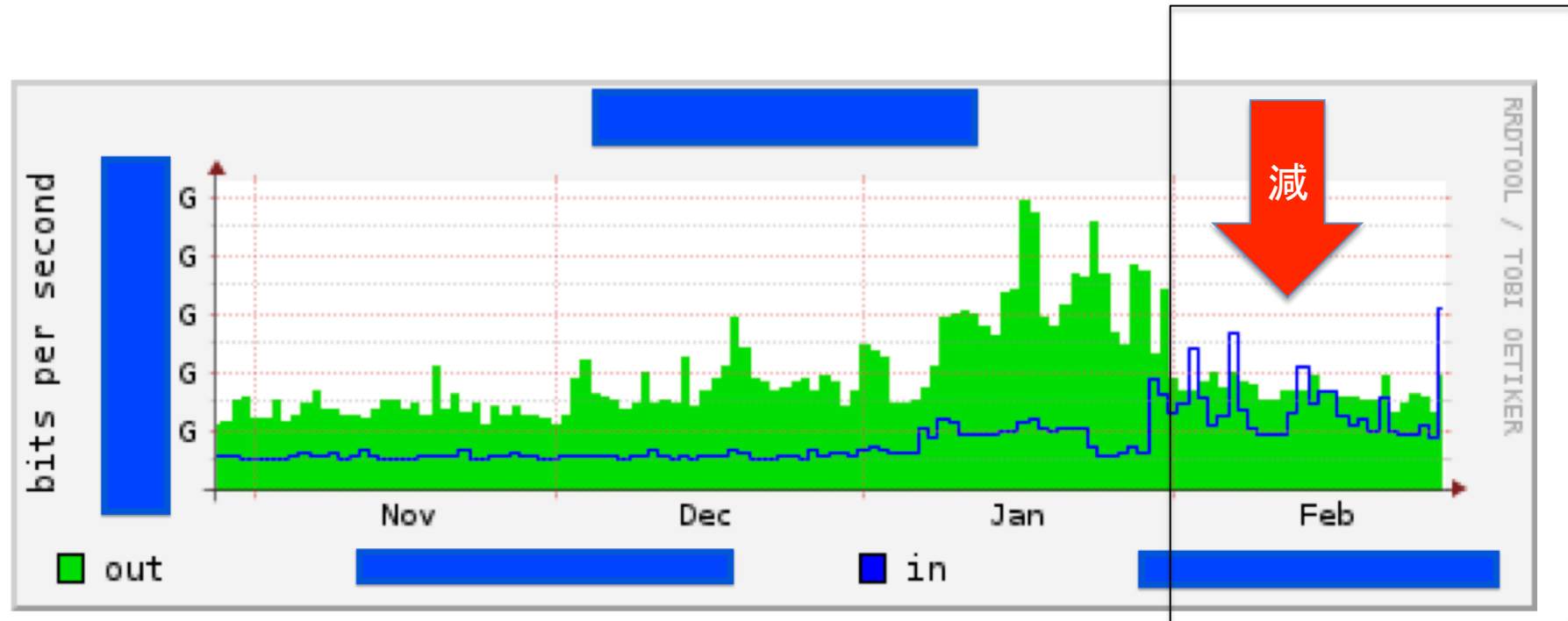


効果はバツグンだ 1



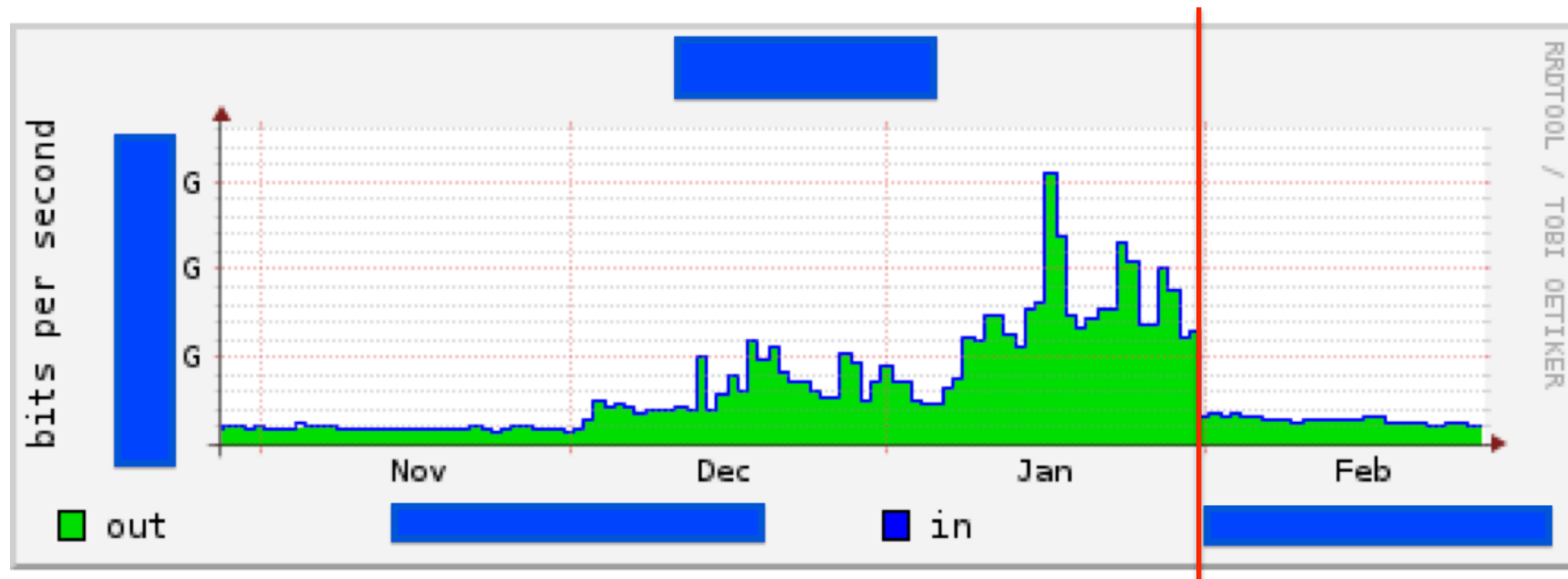
サーバのDCとPOPの間の専用線のトラフィック

効果はバツグンだ 1

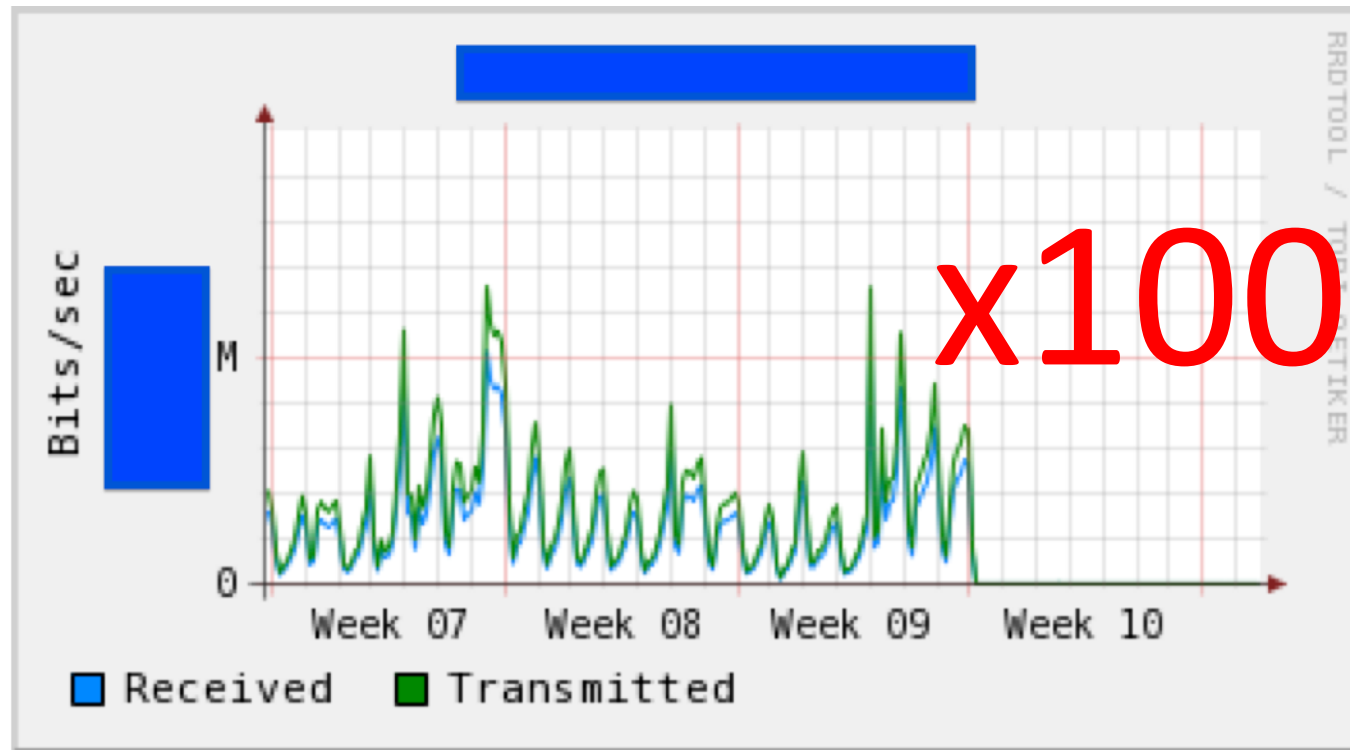


サーバのDCとPOPの間の専用線のトラフィック

効果はバツグンだ 2



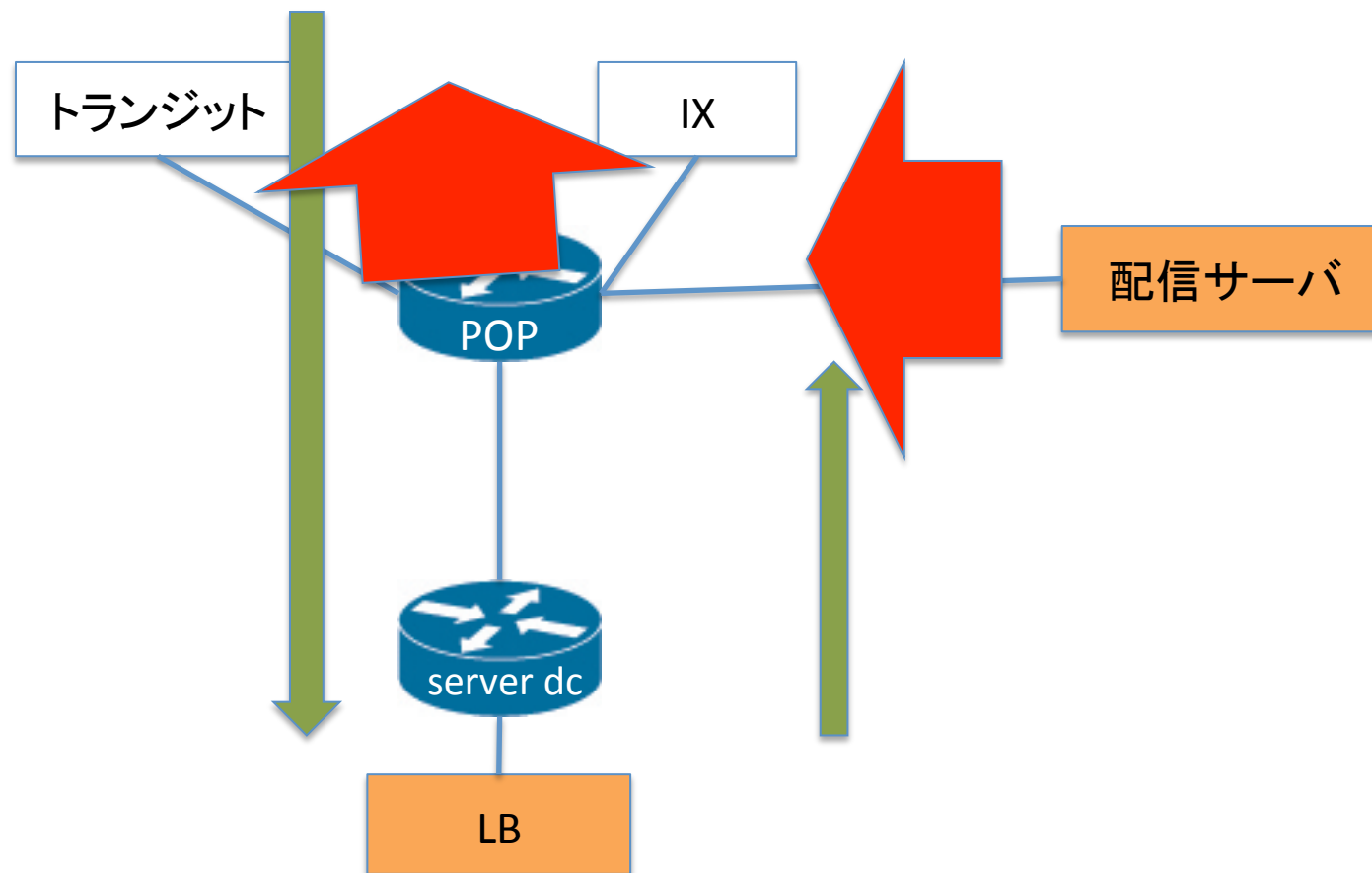
効果はバツグンだ 3



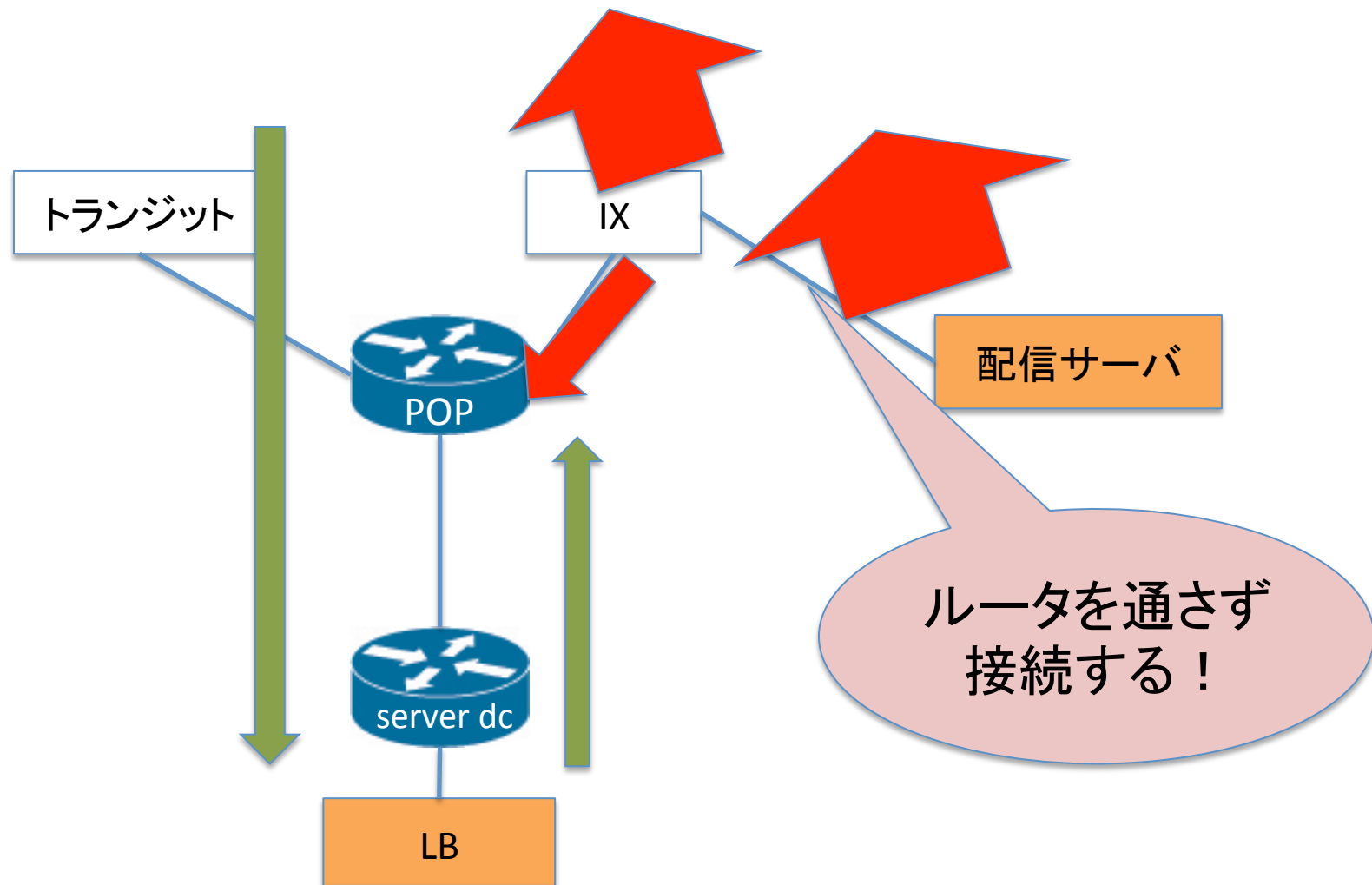
- 10GbpsオーバーでもLBには100Mbps
ーリクエストパケットだけ

「前振り」はこのくらいに
以上、100倍の話
まだまだ満足できない

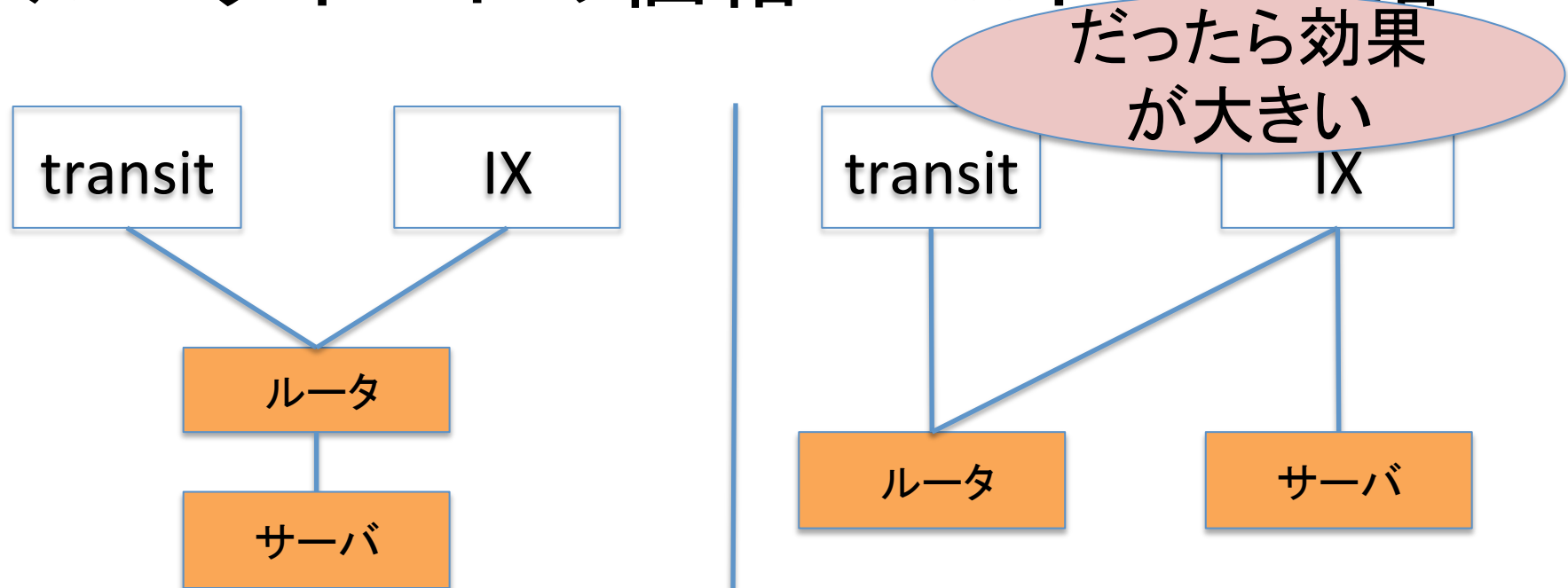
物理への効率的なマッピング(再掲)



物理への効率的なマッピング



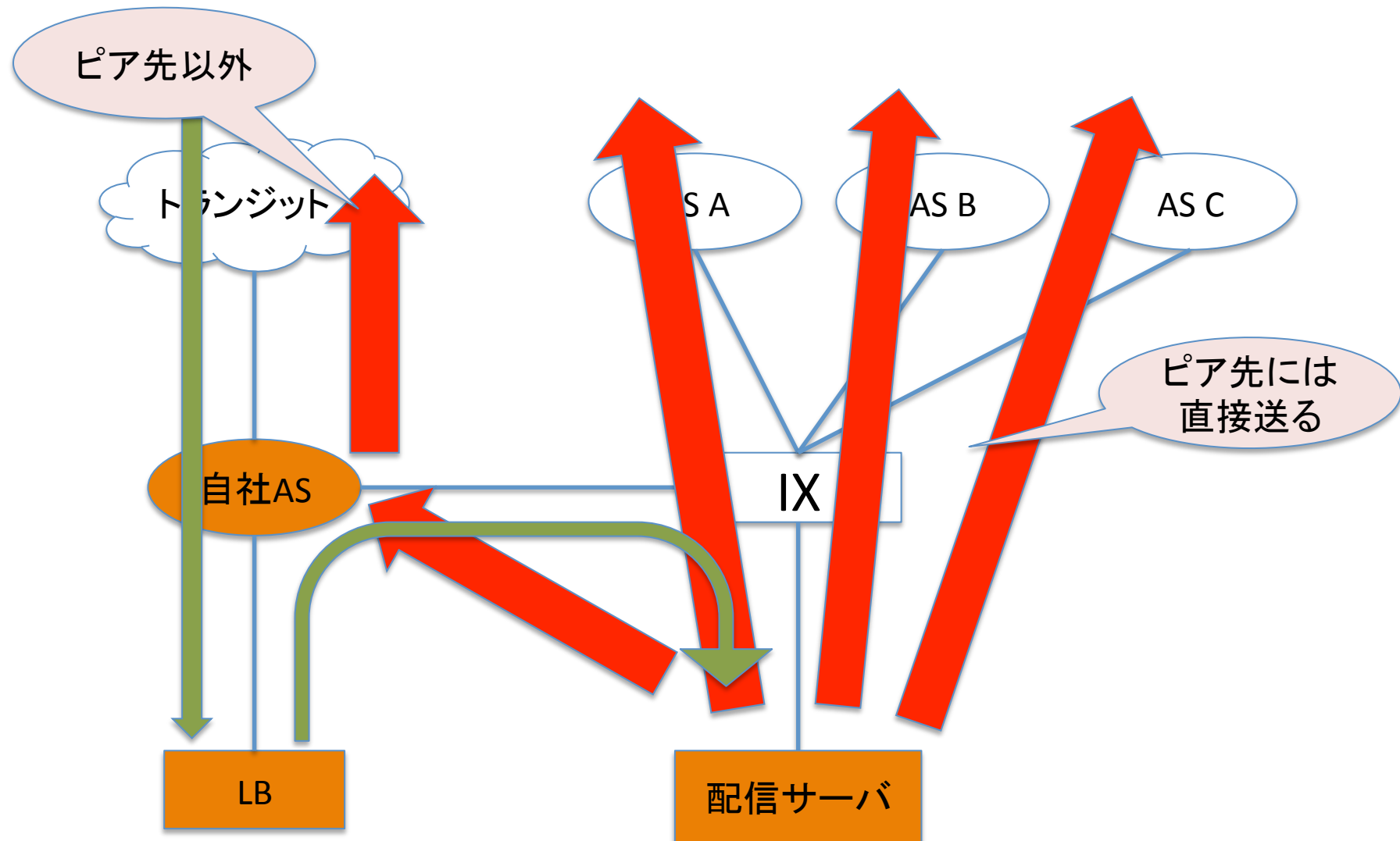
ルータポートの価格 >> IXポート価格



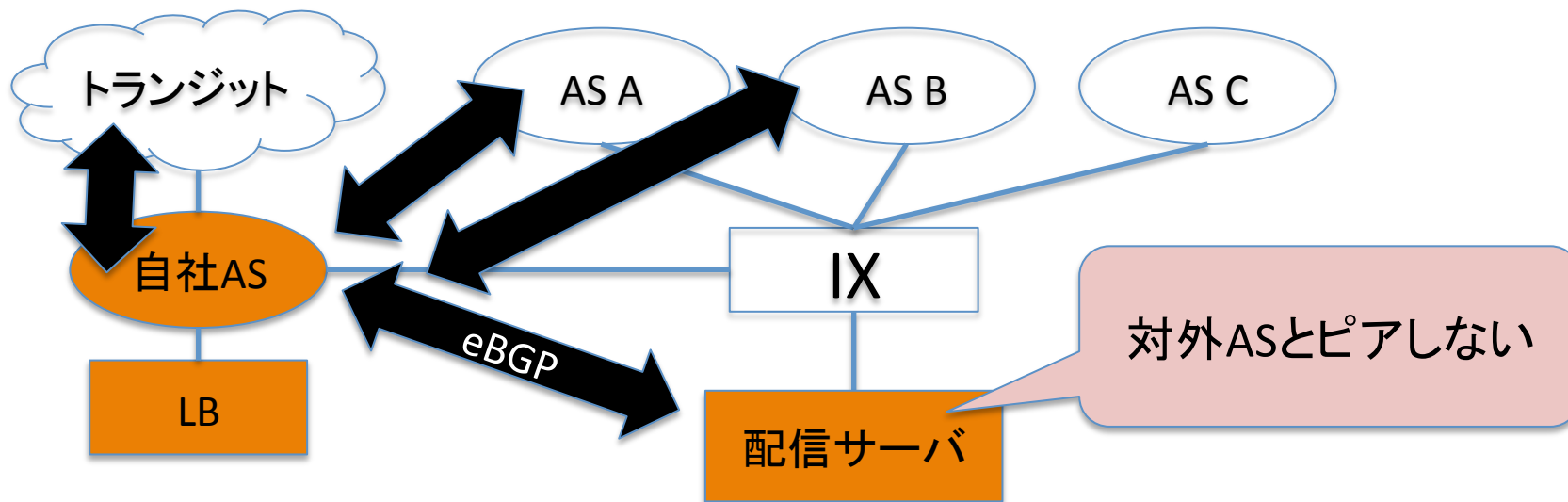
- ルータのポート1つを減らして
- IXのポートを1つ増やす

※ラインカードは多ポートのことが多い

トラフィックの流れ方



ルーティングの組み方



サーバは自社ASとのみ eBGPピア
ASN サーバにはprivate asを設定
受信 ixセグメントがnexthopの経路、default経路
広報 配信のために使うIP帯
(LBからのバランシングやオリジン取得に使う)

JUNOSだったら

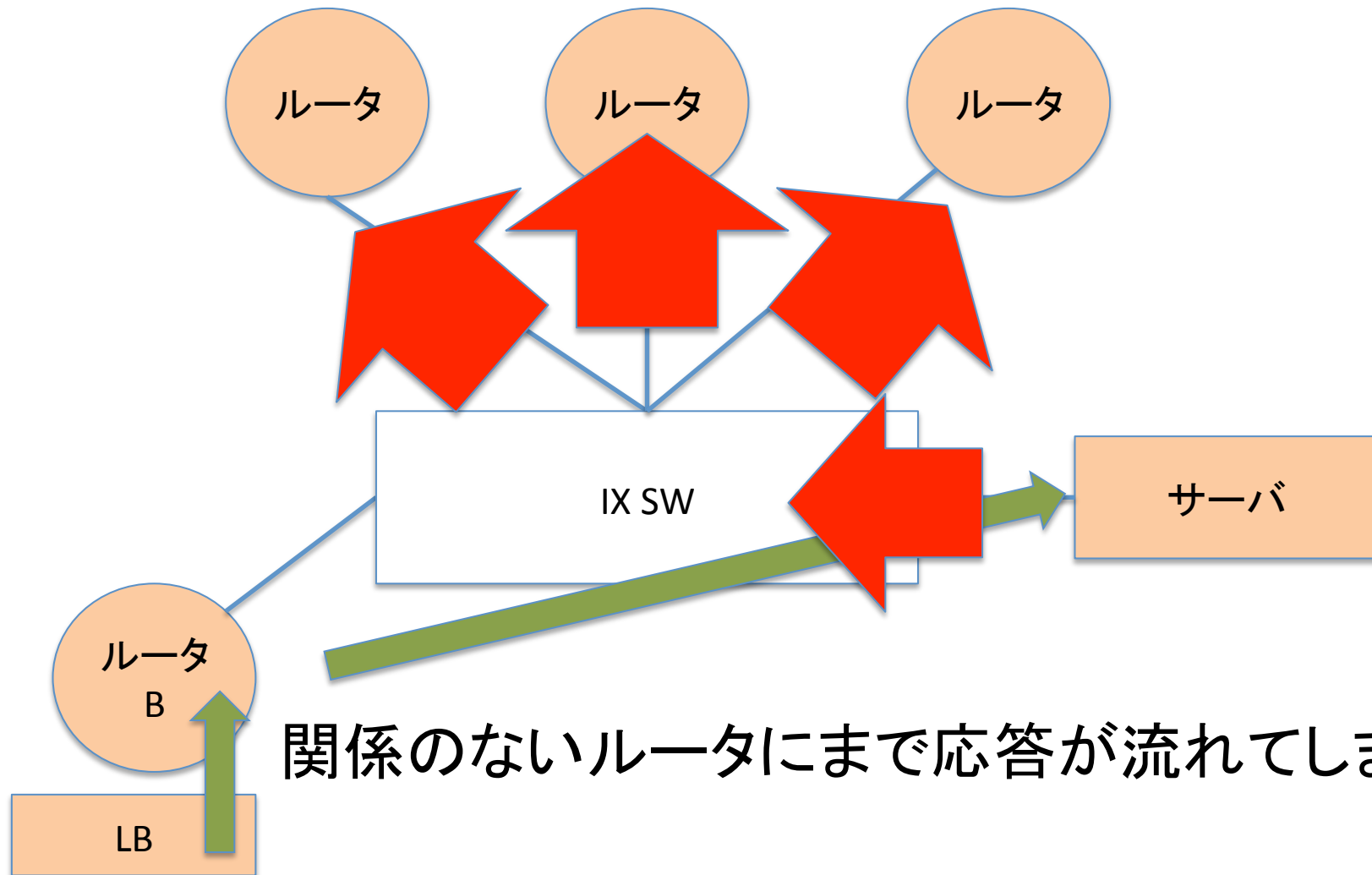
- term export-if-nexthop-is-ix-segment
 - from next-hop x.x.x.0/24
 - then accept
- term export-default
 - from 0.0.0.0/0
 - then accept
- term discard
 - then discard

この構成の懸念点

- CloudIX研究会に提案し、議論してきました
 - CloudIX研究会 <http://cloudix.jp/>
 - AS間でのbfd活用等の活動を行ってきました
- 挙げた懸念点: Unknown Unicast Flooding
 - dest macが学習されずfloodingされる状況
 - IXで発生させると重大な問題を発生させる



Unknown Unicast Floodingの恐怖

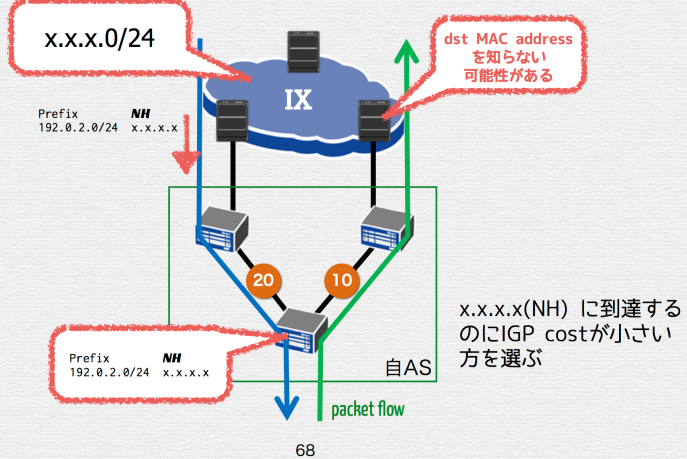


IXではnext-hop selfがセオリー

next-hop self したほうがいい場合

IX経由でもらった経路をiBGPに流すとき

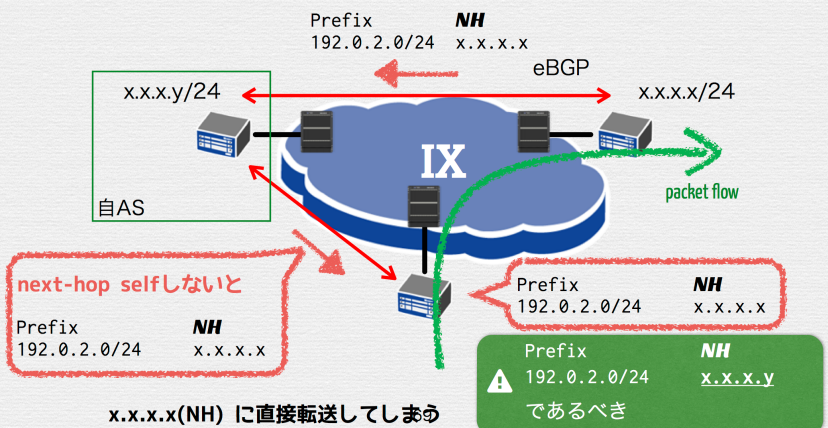
<http://www.janog.gr.jp/doc/janog-comment/jc1005.txt>



next-hop self したほうがいい場合

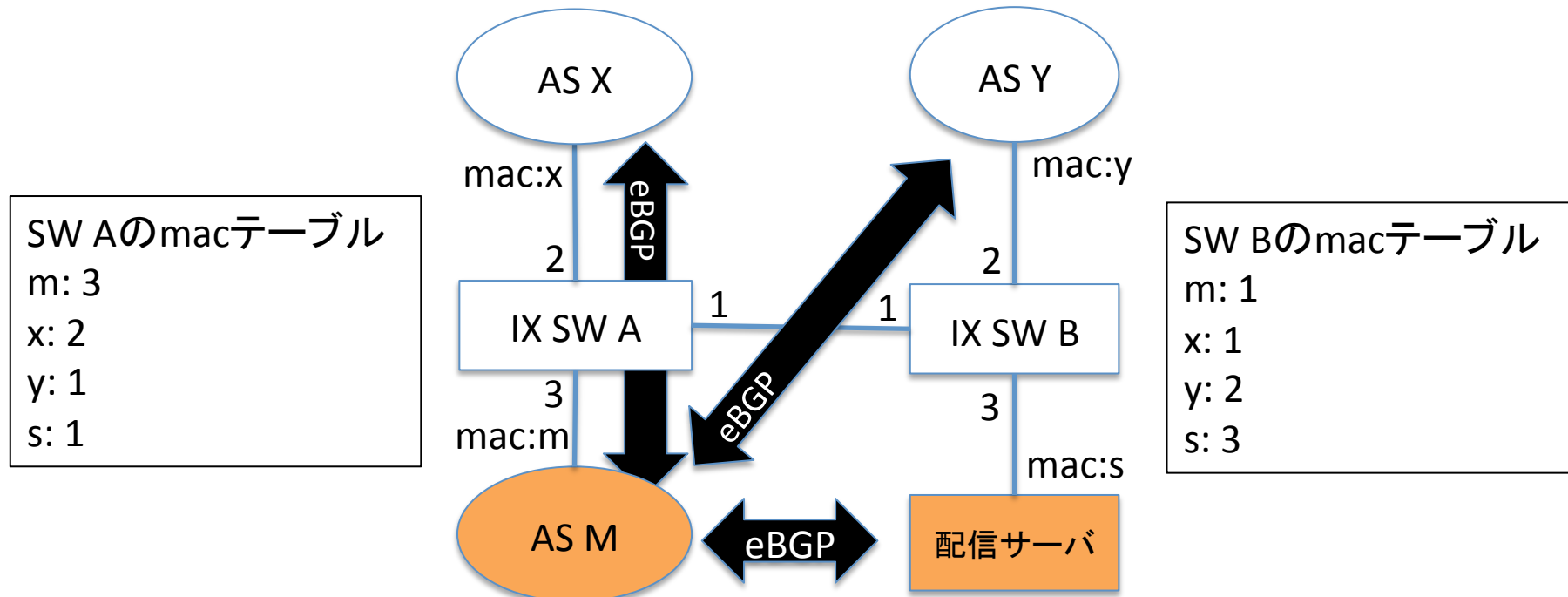
IX経由でもらった経路を同じIX上の別のeBGPの流すとき

<http://www.janog.gr.jp/doc/janog-comment/jc1005.txt>



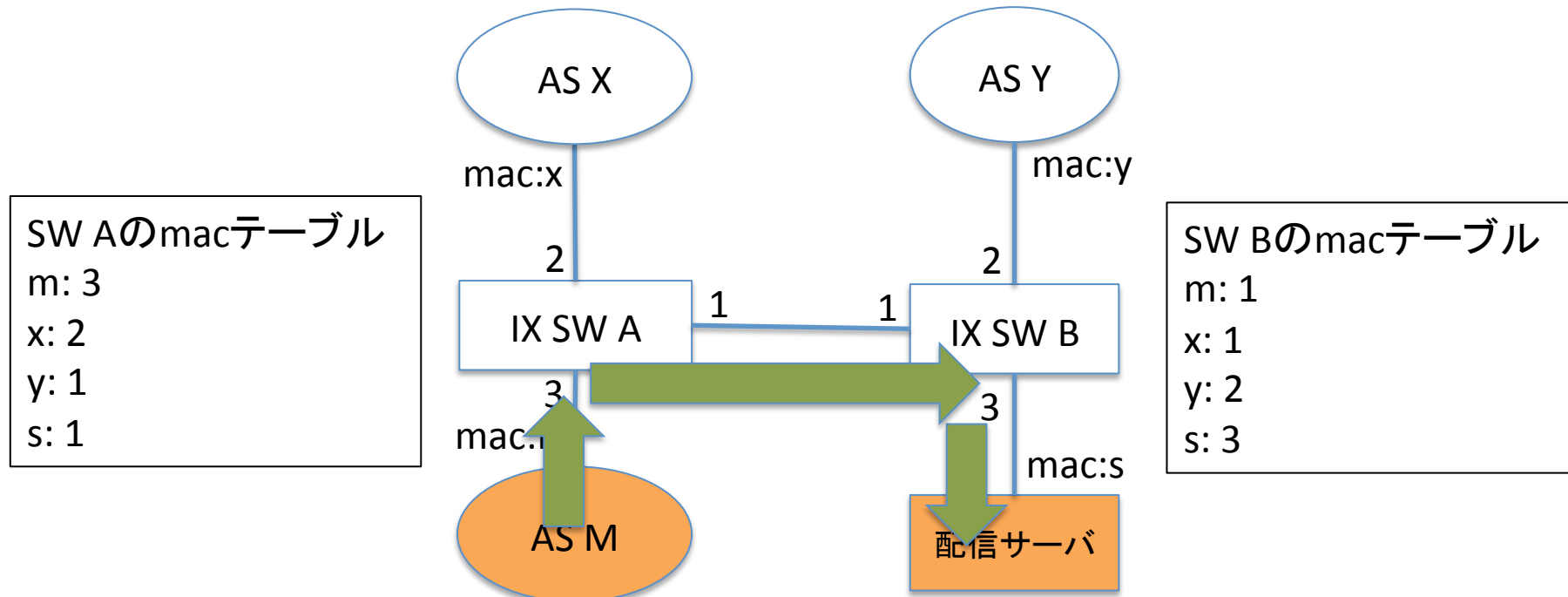
- JANOG33のチュートリアルを引用
- コントロール用のパケット(BGPのkeepaliveやbfd)でmac tableを維持できる

問題になる例(1/3)



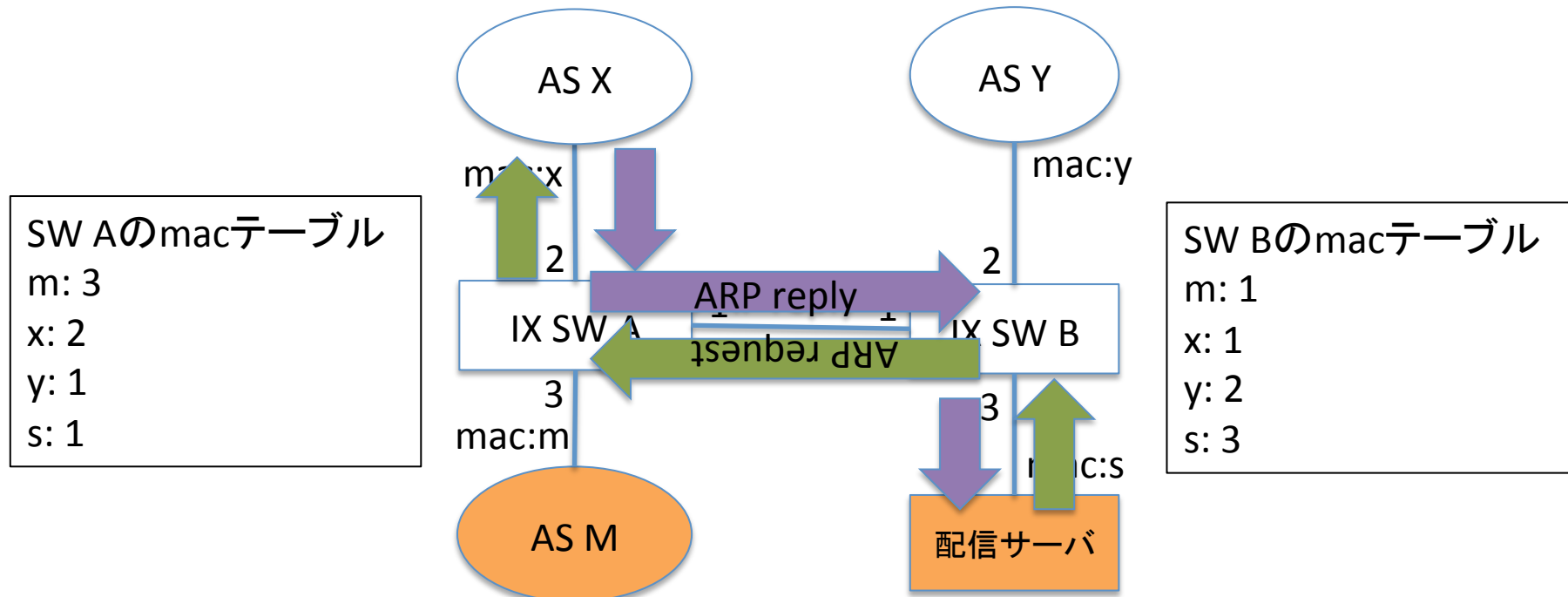
- MはServer, X, Yとピア
- ServerはXのARP requestを投げる
 - ARP ReplyがXからの唯一のパケット
 - mac aging timerで消える可能性がある

問題になる例(1/3)



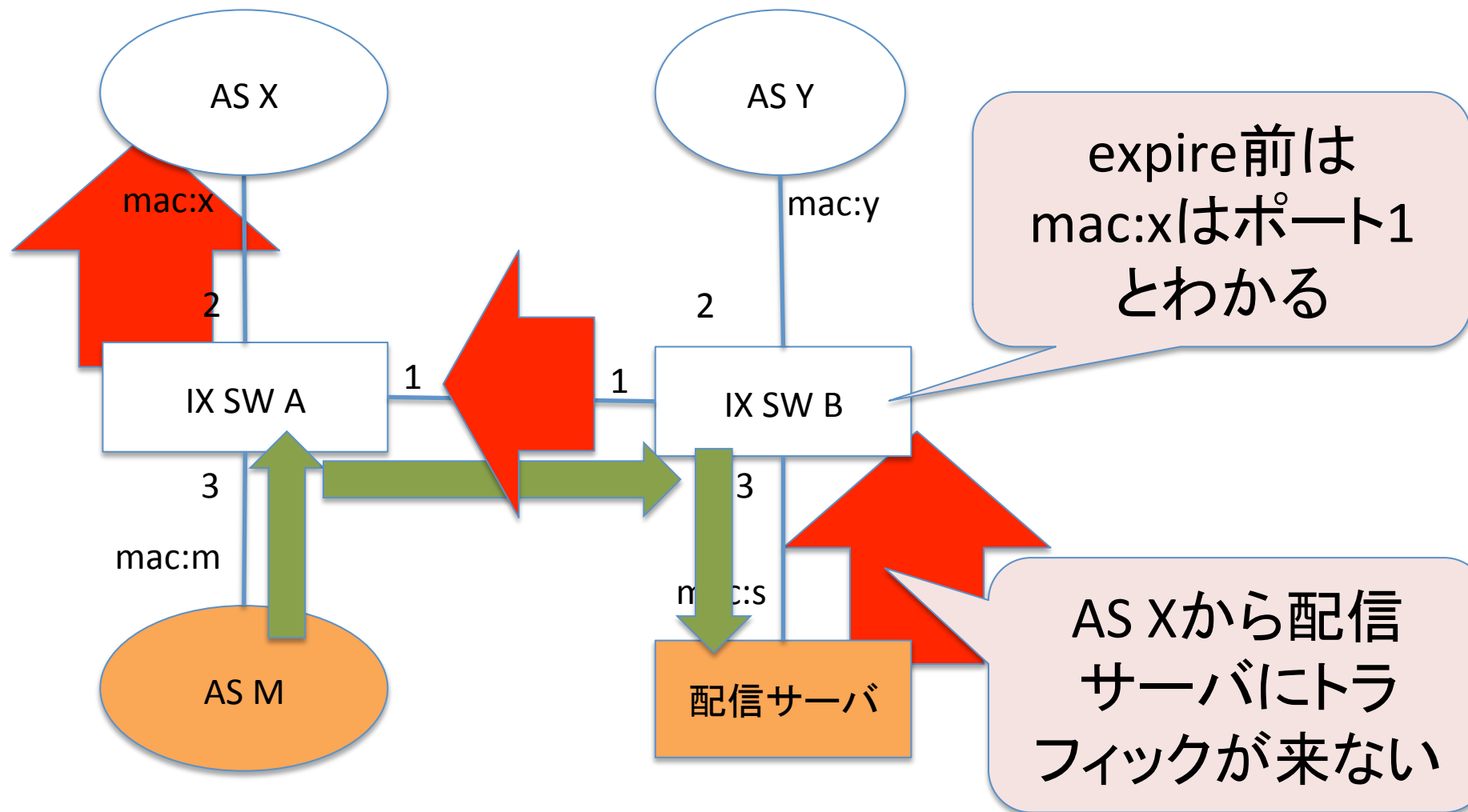
- MはServer, X, Yとピア
- ServerはXのARP requestを投げる
 - ARP ReplyがXからの唯一のパケット
 - mac aging timerで消える可能性がある

問題になる例(1/3)

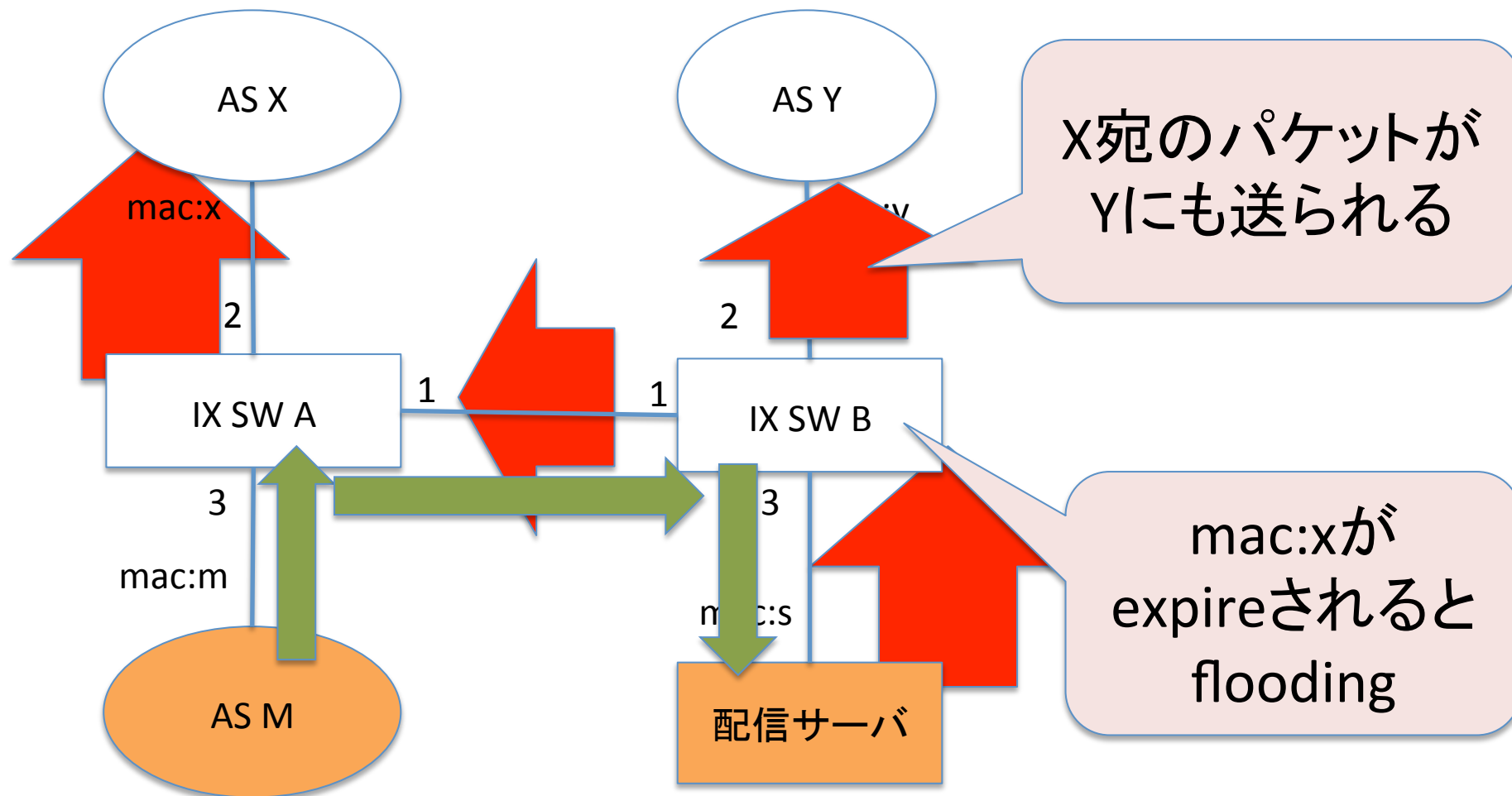


- MはServer, X, Yとピア
- ServerはXのARP requestを投げる
 - ARP ReplyがXからの唯一のパケット
 - mac aging timerで消える可能性がある

問題になる例(2/3)



問題になる例(3/3)

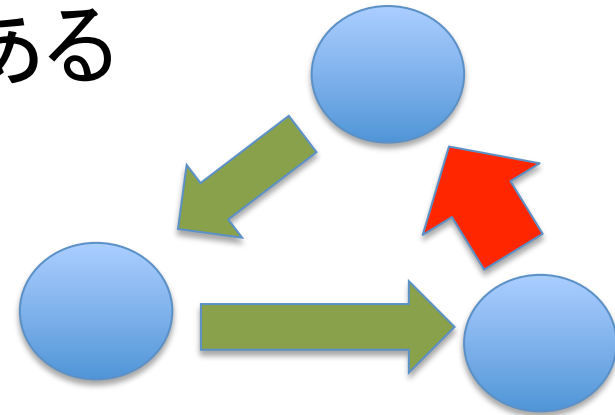


よくあるdefault値

- よくあるL2スイッチのmac aging timer
 - 300sec
- L3機器のarp aging timer
 - 某J社 1200 sec
 - 某C社 14400 sec
 - 某サーバOS 今回の用途だと消えない
 - (ARPの再学習はunicastで行われる！)

今回の問題点の整理

- L3DSRでトラフィックが非対称
 - リクエスト: 自社ASから配信サーバへ
 - レスポンス: IXでつながったどこかのASへ
 - ※対外ASから配信サーバへのパケットは無い
- 直接ピアをしてないIX上のASへの送信
- mac tableに無い可能性がある



解決案

- 全てのIXスイッチでmac tableが完璧なら良い



- 利用者ががんばる
 - next-hopへpingを打つデーモン？
 - arp aging timerをmac aging timerより短くする？
- IXでがんばる
 - 全スイッチから全IPにping打てませんか？

やっていいですか？

- ピア先との調整や追加のピア依頼が不要！
 - 相手から見ると送信元macアドレスが変わるだけ
- 運用負荷はそんなに上がらない
 - 自分のASのポリシーの一部を配信機器に流すだけ
- IXとの調整はした方がいいけど、ばれない？
 - BGPしか使ってない

特別な技術を使ってないから誰でも出来る！
ある日、あなたのルータをfloodingが襲うかも！

まとめ

- L3DSRでPOPまでのトラフィックを減らしました
- IXを使ってASの外から送る案を紹介しました
 - mac tableが維持できないと迷惑をかけうる構成
 - こういうことをやることができると知ること、トラブル対応が早まるかもしれません
- 議論で各方面の方から意見を伺いたいです

議論の流れ

- 時間割

- IX事業者様

9min

- だれでもOK

11min(調整)

- NGな内容

- お金の話。コスト構造の話にしてもらえると。

- 弊社のサービスの話、流れているものの話

- 応用したビジネスの話

IX事業者様

- mac aging timerチューニングしていますか？
 - 長いと切り替わらない
 - 短いとフラッディングが増える
- Unknown Unicast Floodingへの対策は？
 - 実装されているのであれば発動条件は？
 - fail-safeなしのノーケア？
- サーバ直結について
 - ルートサーバ等の経験から気をつけた方がいいことあるでしょうか？

IX利用者様・メーカー様

- ルータから応答があるパケットを投げたい
 - 負荷かからないオススメの方法ありますか？
 - arp aging timerを短くする場合の懸念は？
- スイッチでmac tableを維持できるような機能
- プロトコルは？
 - ICMP? ARP? unicast? I2 multicast?
- レートは？
 - 1sec ? 1min ? 2min?

最後のまとめ

- IXごとに事情は異なりそうですね
 - 事業者様に事前に相談しよう！
 - もしやるならユーザ会とかでシェアしよう