

BGP ○○編 時代と共に ～10年の歴史を振り返りながら～

Yoshida '**tomo**' Tomoya
<yoshida@ocn.ad.jp>

Matsuzaki '**maz**' Yoshinobu
<maz@iij.ad.jp>

0. 昔ばなし編

- ・OCN
- ・IIJ

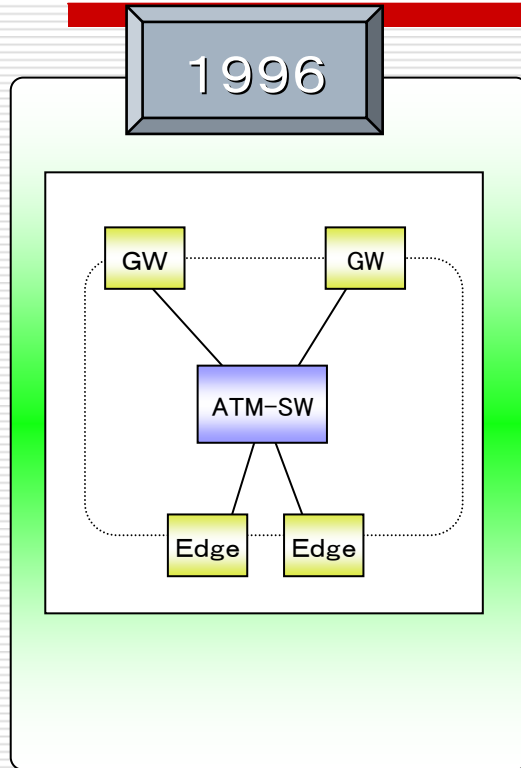
OCNの昔話(おかげさまで10周年)

- OCNエコノミー(1996年12月25日)
 - 完全2重化のネットワーク
 - ロードバランスでバーストラフィックも対処するぞ
 - OSPFに再配信
- 徐々に色々と問題が
 - OSPFの経路(/28, /29)が増大
 - エリア分けを実施した後、1エリアのルータ台数やLSDBが肥大化

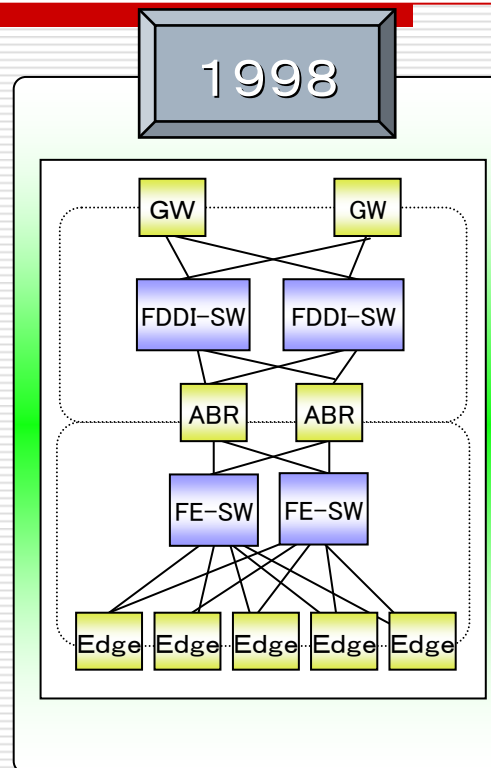
OCNの昔話(続き)

- アドレスブロックのアグリゲーション問題
 - JPNICのおかわりが厳しかった
 - 当然ちゃんと使い切らないともらえない
 - 全国にブロックがまばらに
- Static経路をOSPFに再配信していたのをやめてBGPに乗っける方針に転換
 - 最初はno-exportだけ付与して内部へ伝播
 - 地域とか経路の役割に応じた色づけを徐々に開始し、現在は全てコミュニティで経路制御
- 内部のBGP経路もばかにならない
 - Confedeとか色々検討したが、リフレクタの階層化で対処

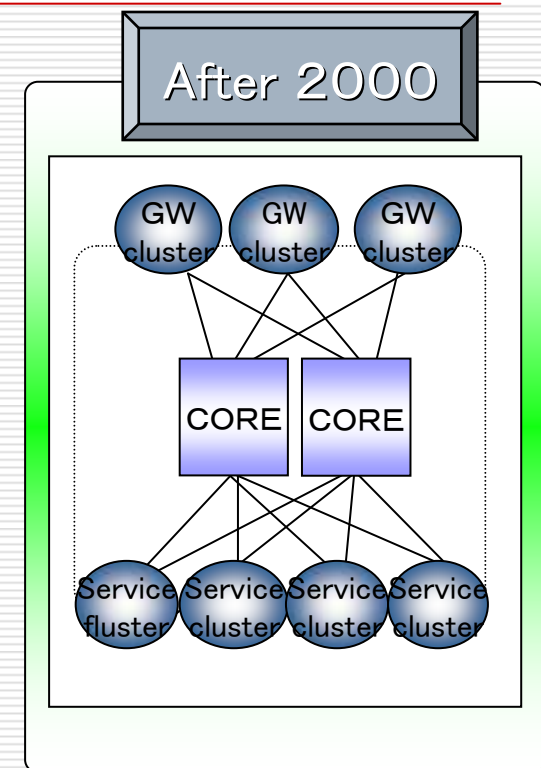
OCNバックボーンの変遷



- ・ ATM-SWを中心としたメッシュ構成



- ・ ABRによるルーティング
負荷分散
- ・ FDDI-SWやFE-SW
にて構成



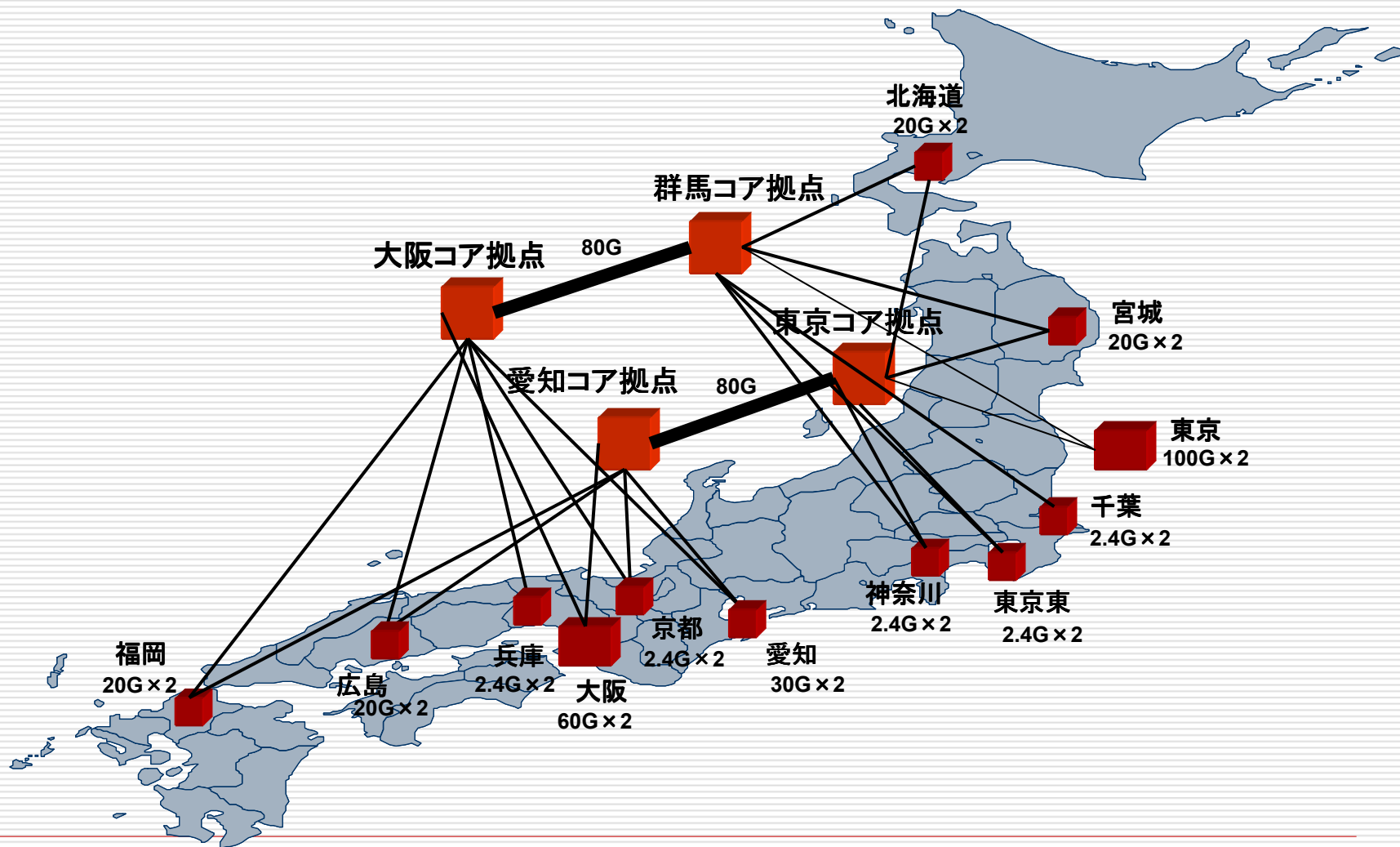
- ・ コアを中心とした構成
- ・ サービス等に応じて
クラスタ化

最近

□ コアを見直しました

- 激甚災害にも耐えられるインフラを目指して
- 100km以上コア拠点を分散

7月末現在のOCNバックボーン(主要部分)



OCNバックボーンの経路制御

- トポロジー情報の管理：OSPF
 - 基本的にはECMPの冗長化運用
 - 一部経路のロードバランス適応時にOSPFを利用
 - V4/V6ともに同様のトポロジー
- 経路情報の管理：BGP
 - 基本はフルルートを送配、配れる範囲で適切に配る
 - V4/V6ともに同様のトポロジー

IIJの昔ばなし

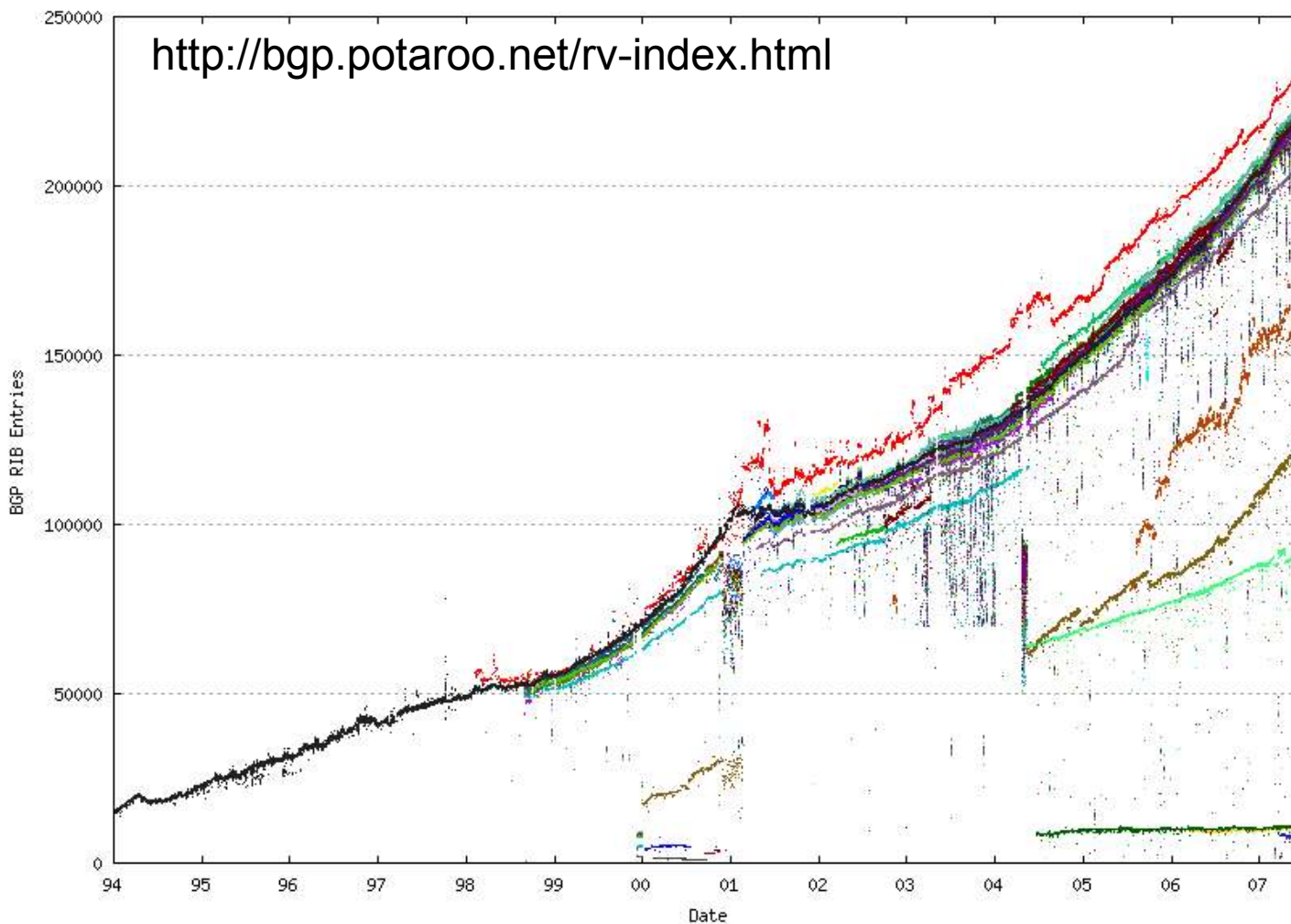
- 網内はstaticをBGP/OSPFに
 - ルータ毎に/24を割り振って、アドレス割り当て
 - /24だけをBGPで網内に広報
 - 細かいのは一部OSPF
- 他ASに経路広報する時にはaccess-listとas_path filterで制御

IIJの昔ばなし(その後)

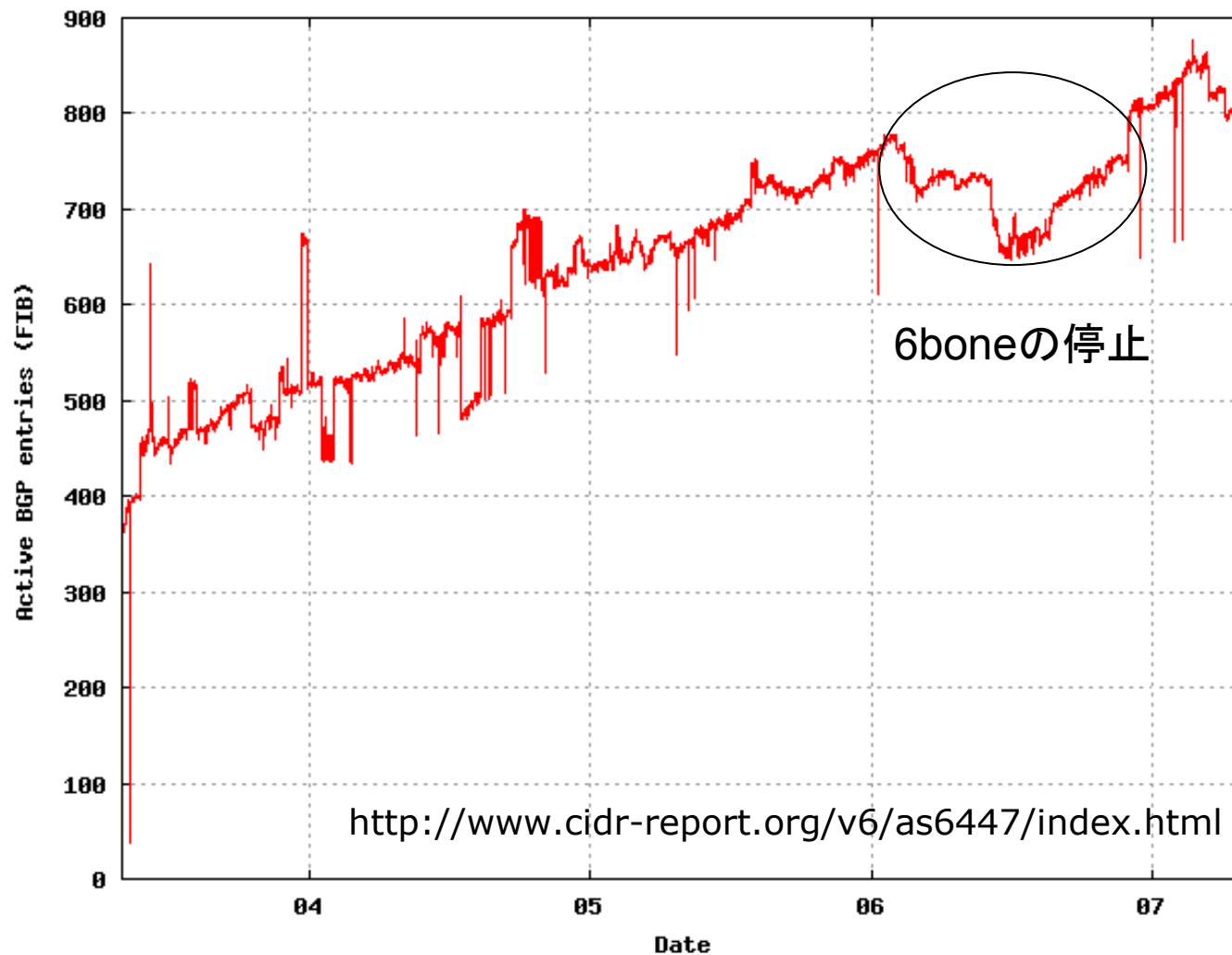
- 集約を止めて、顧客経路は全てBGPで広報
 - 経路数の増加に対して、網内で集約するうれしさ
があまり無かった
- 他ASに経路広報する時にはprefix-listと
bgp communityで制御
 - 経路の受信/生成時にbgp communityを付加
 - 出口で評価して、広報/抑止を制御

1. 経路の推移を見てもよう編

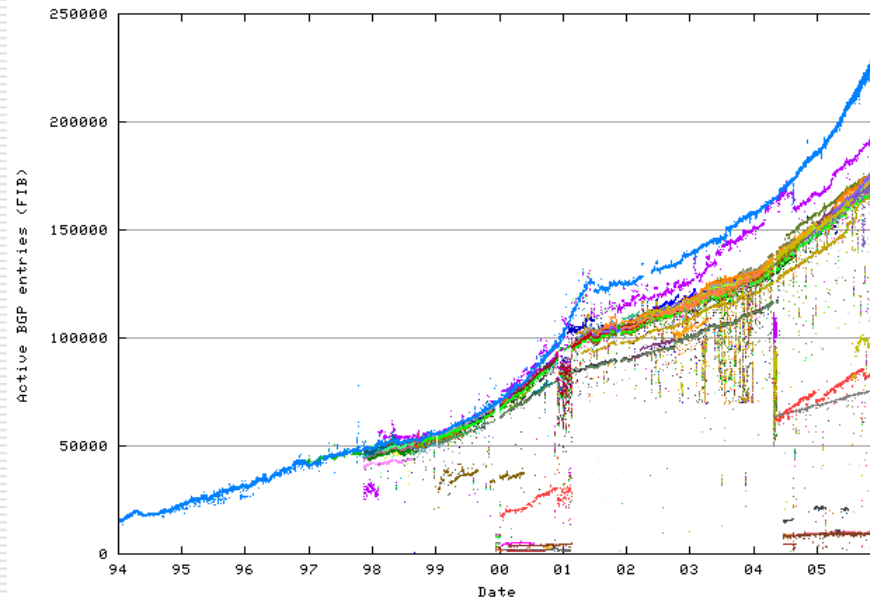
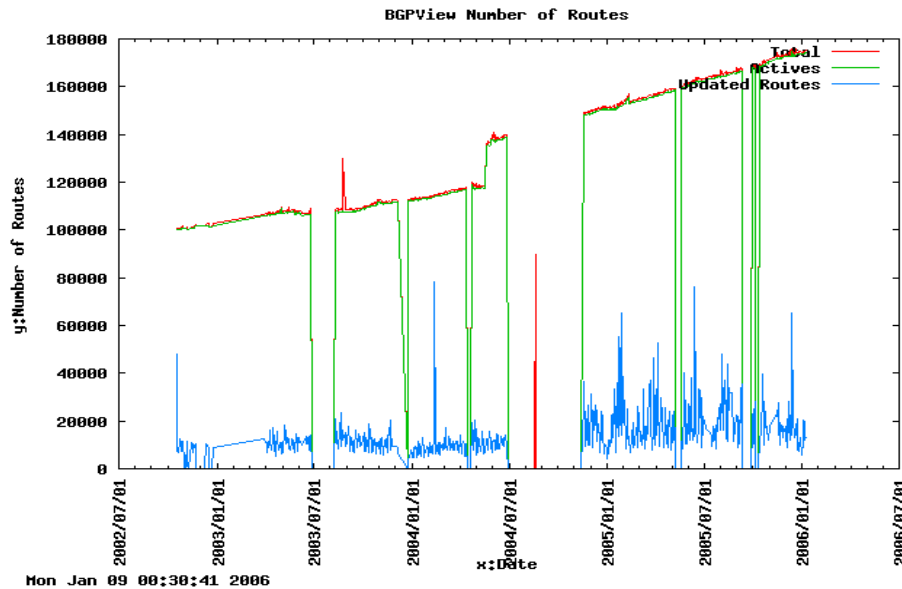
IPv4経路数



IPv6経路数の推移



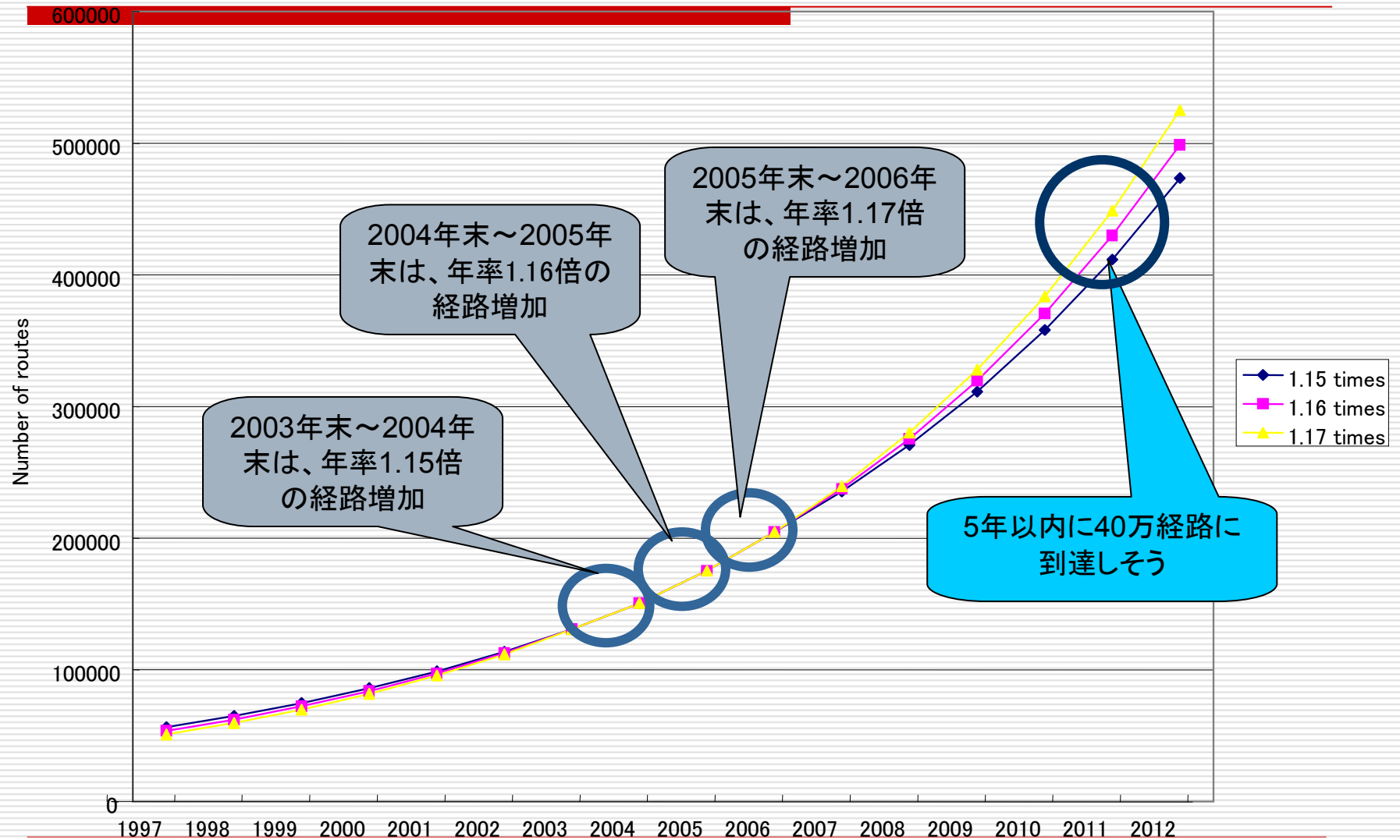
IPv4経路数の推移 ～予測と過去の実測値～



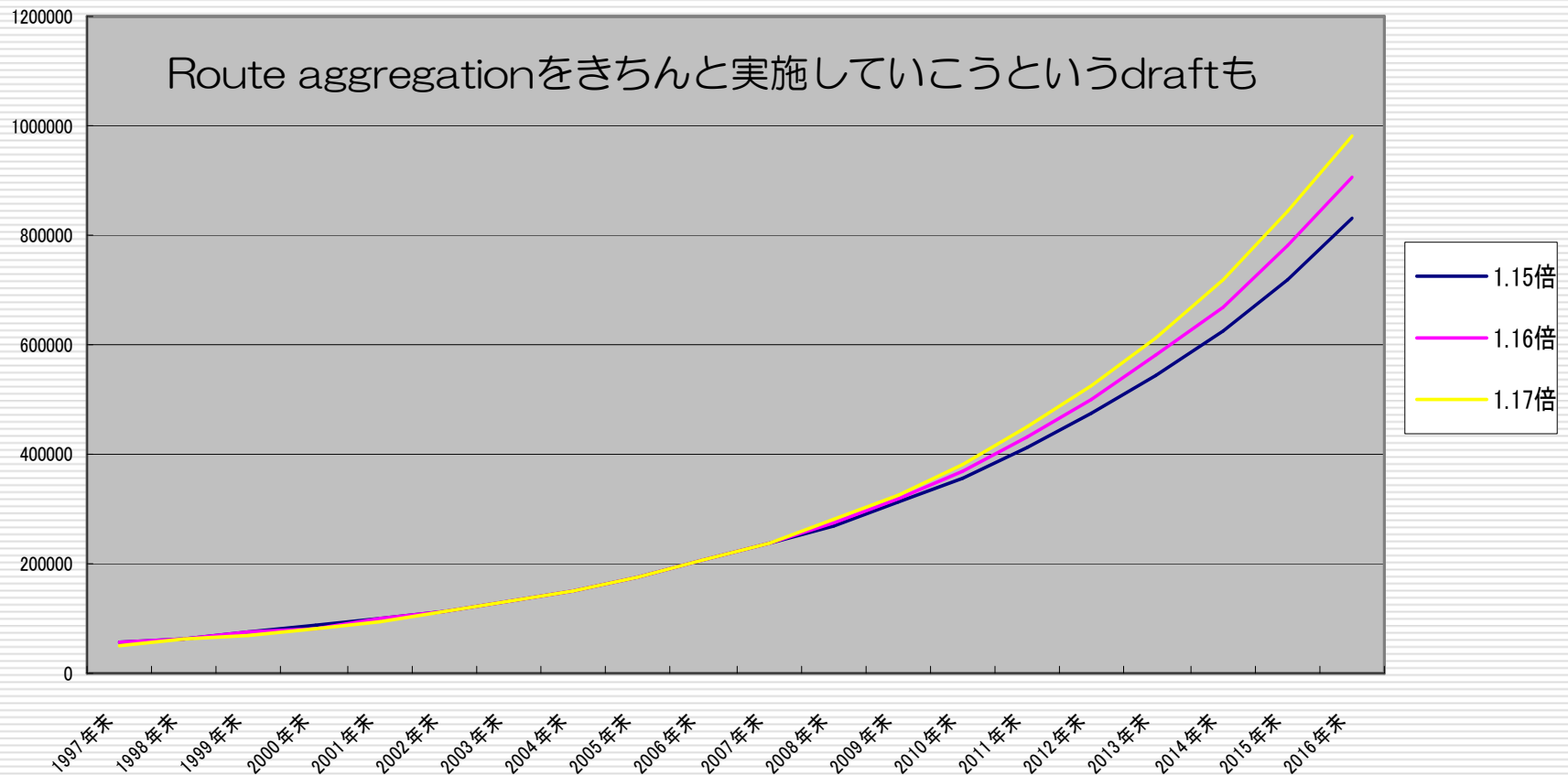
	推移1		推移2	
1997年末	56580	-1.15倍	53715	-1.16倍
1998年末	65067	-1.15倍	62310	-1.16倍
1999年末	74827	-1.15倍	72279	-1.16倍
2000年末	86051	-1.15倍	83844	-1.16倍
2001年末	98958	-1.15倍	97259	-1.16倍
2002年末	113802	-1.15倍	112821	-1.16倍
2003年末	130873	実測	130873	実測
2004年末	150712	1.15倍:実測	150712	1.15倍:実測
2005年末	175261	1.16倍:実測	175261	1.16倍:実測
2006年末	201550	1.15倍	203302	1.16倍
2007年末	231782	1.15倍	235831	1.16倍
2008年末	266550	1.15倍	273564	1.16倍
2009年末	306532	1.15倍	317334	1.16倍
2010年末	352512	1.15倍	368107	1.16倍
2011年末	405389	1.15倍	427005	1.16倍
2012年末	466197	1.15倍	495326	1.16倍
2013年末	536127	1.15倍	574578	1.16倍
2014年末	616546	1.15倍	666510	1.16倍

2003年時の実測値と過去の推移に基づいた予測
→ かなり予想通りで推移している

IPv4経路数の推移予測

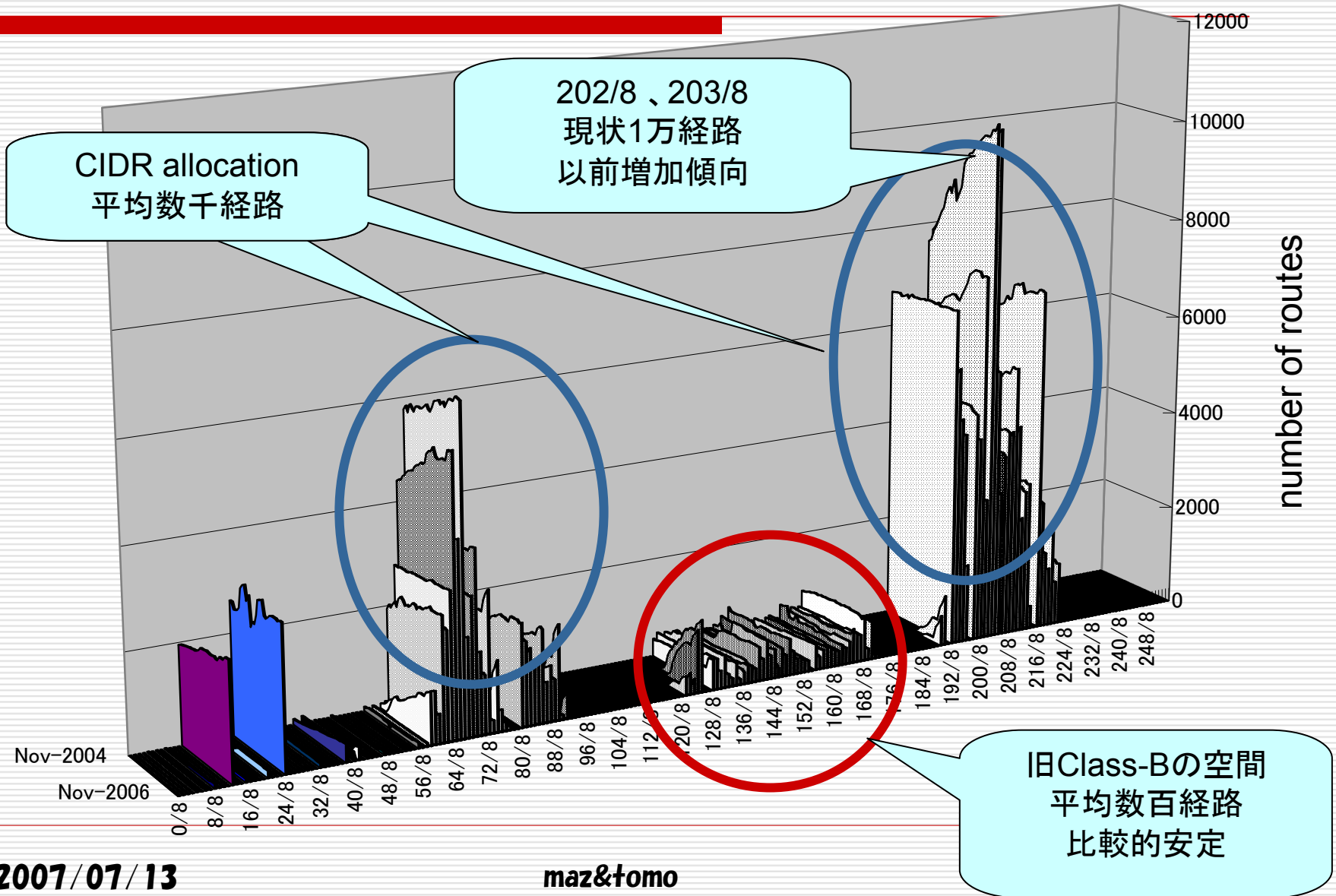


IPv4経路数の推移予測(続き)



年に1.15倍~1.17倍の増加（最近は倍率が増加傾向）
割り振りが停止しても一定期間は広告経路の増加は継続される（予想）

/ 8毎に見ると



プロフィックス長毎に見てみると

	2003末	2004末	2005末	2006末	増減 2003- 4	増減 2004- 5	増減 2005- 6
/1	0	0	0	0	0	0	0
/2	0	0	0	0	0	0	0
/3	0	0	0	0	0	0	0
/4	0	0	0	0	0	0	0
/5	0	0	0	0	0	0	0
/6	0	0	0	0	0	0	0
/7	0	0	0	0	0	0	0
/8	19	19	18	19	0	-1	1
/9	4	3	5	10	-1	2	5
/10	6	7	8	13	1	1	5
/11	14	15	17	30	1	2	13
/12	57	61	81	111	4	20	30
/13	100	138	187	222	38	49	35
/14	277	314	340	397	37	26	57
/15	483	553	666	794	70	113	128
/16	7506	8113	8597	9077	607	484	480

/8~/16は
微小ながら
増加している

2003末 : 130873
2004末 : 150712
2005末 : 175261
2006末 : 204725

半分が/24

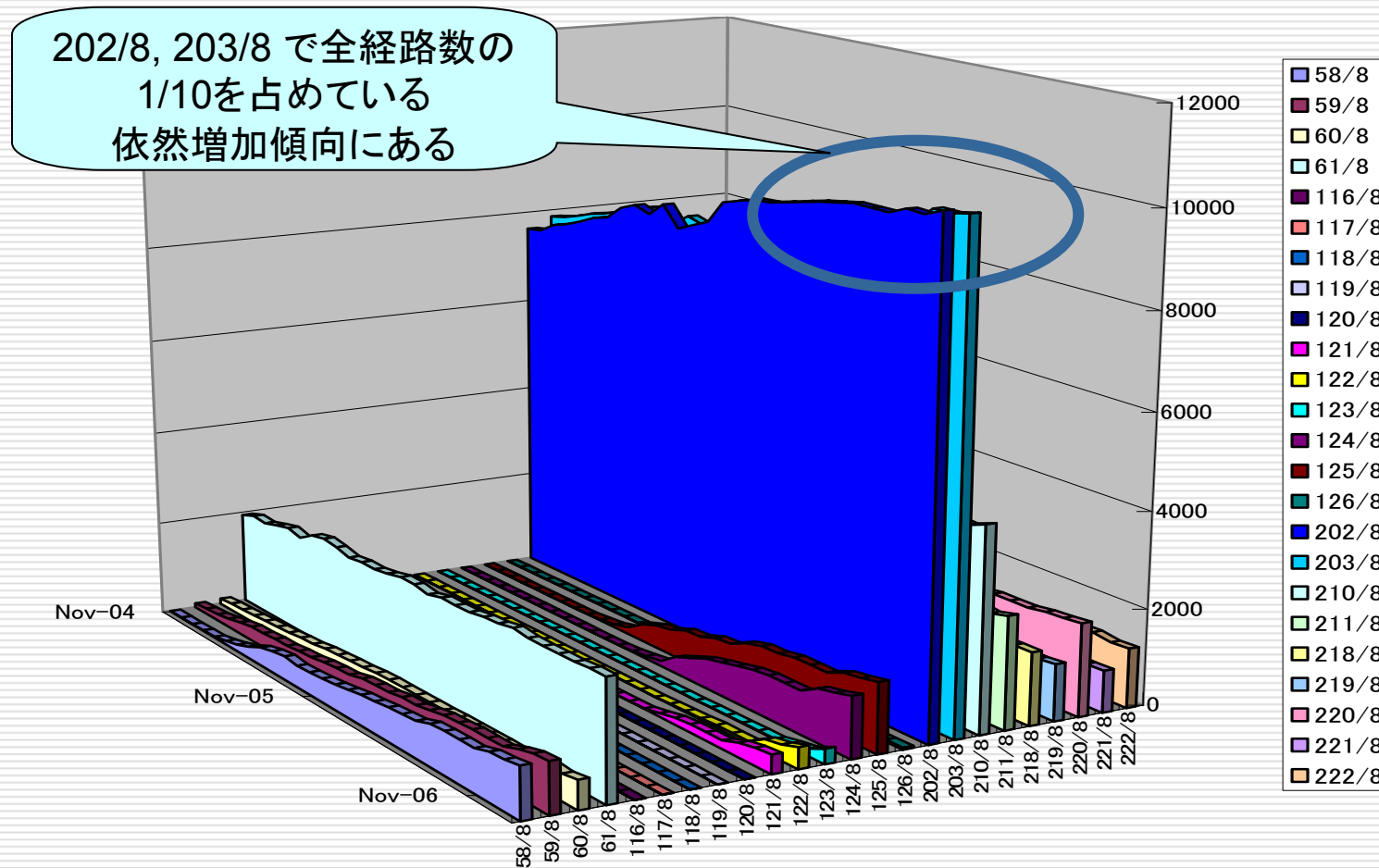
	2003末	2004末	2005末	2006末	増減 2003-4	増減 2004-5	増減 2005-6
/17	1829	2270	2880	3625	441	610	745
/18	3334	3933	4871	5826	599	938	955
/19	8716	9818	11026	12664	1102	1208	1638
/20	9249	10402	12142	14281	1153	1740	2139
/21	6656	8007	10194	12838	1351	2187	2644
/22	9386	11066	13440	16203	1680	2374	2763
/23	10943	12707	14626	17682	1764	1919	3056
/24	71541	82382	95225	109219	10841	12843	13994
/25	182	252	345	658	70	93	313
/26	233	239	292	468	6	53	176
/27	156	130	194	364	-26	64	170
/28	70	69	26	69	-1	-43	43
/29	21	54	12	44	33	-42	32
/30	50	120	36	80	70	-84	44
/31	0	0	3	0	0	3	-3
/32	41	40	30	31	-1	-10	1

2003末 : 130873
 2004末 : 150712
 2005末 : 175261
 2006末 : 204725

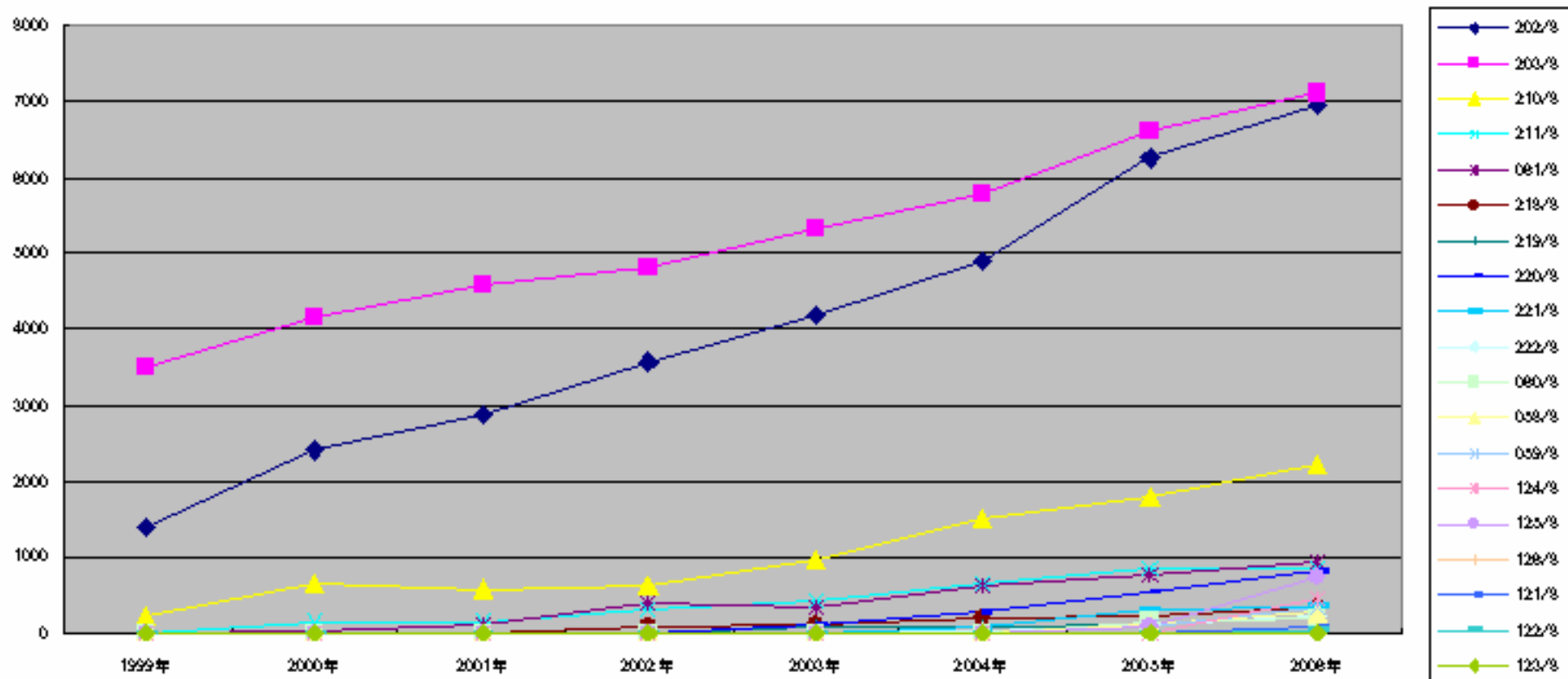
何が伸びてるのか

- 新しい経路はもちろん増加
- 昔の経路も実は延び続けている

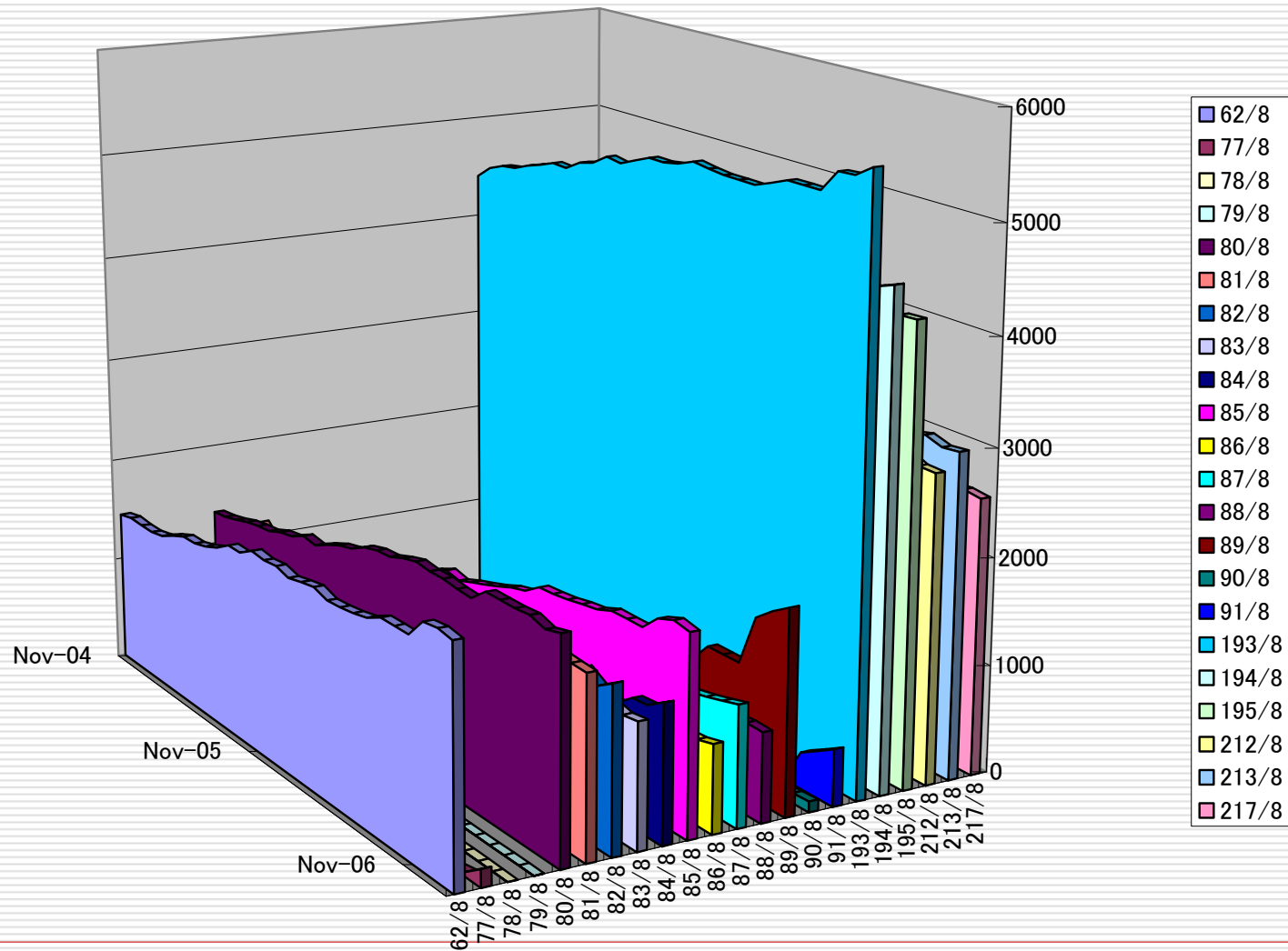
APNIC地域の / 8毎のトレンド



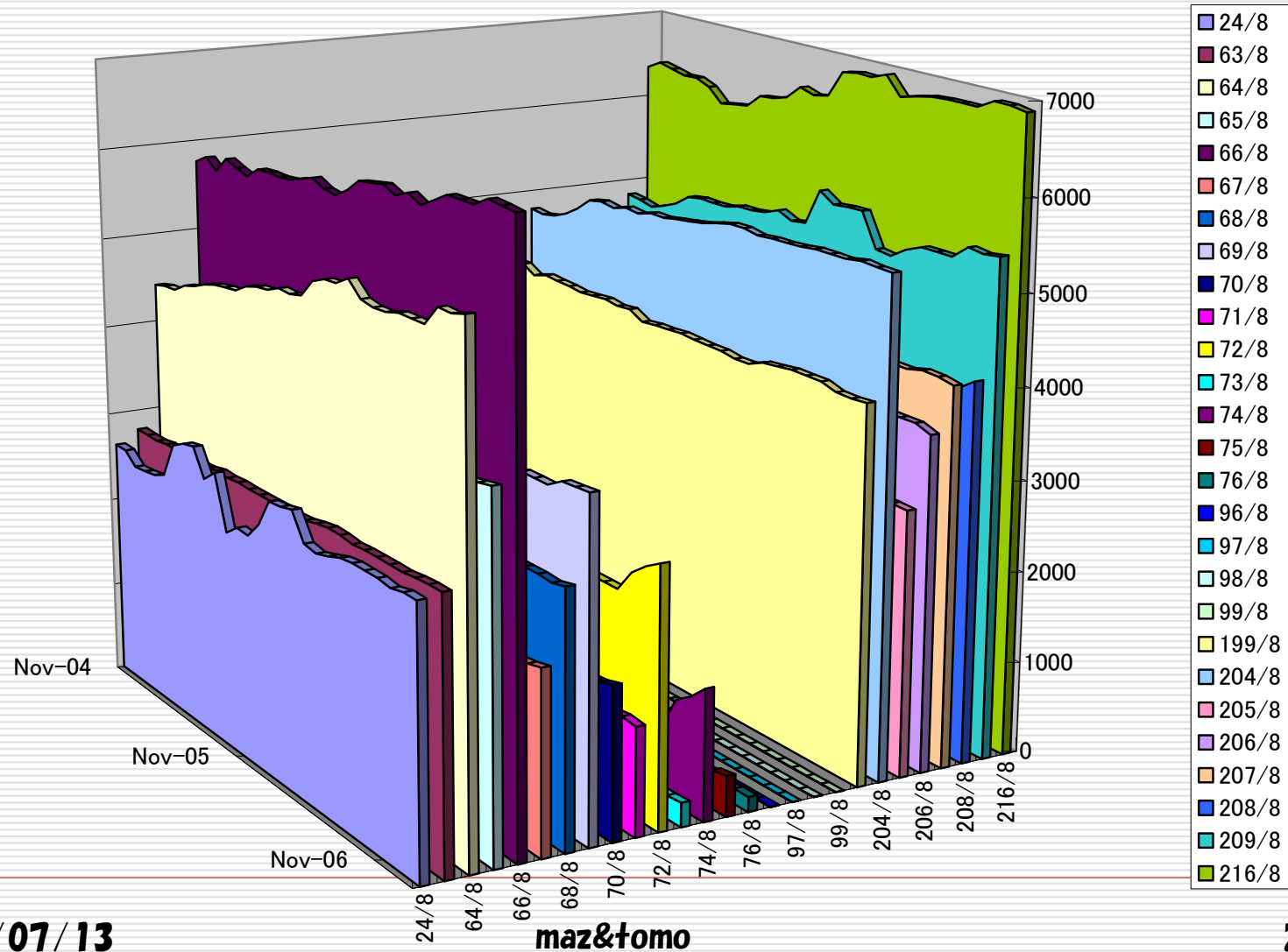
APNIC地域を細かく見ると



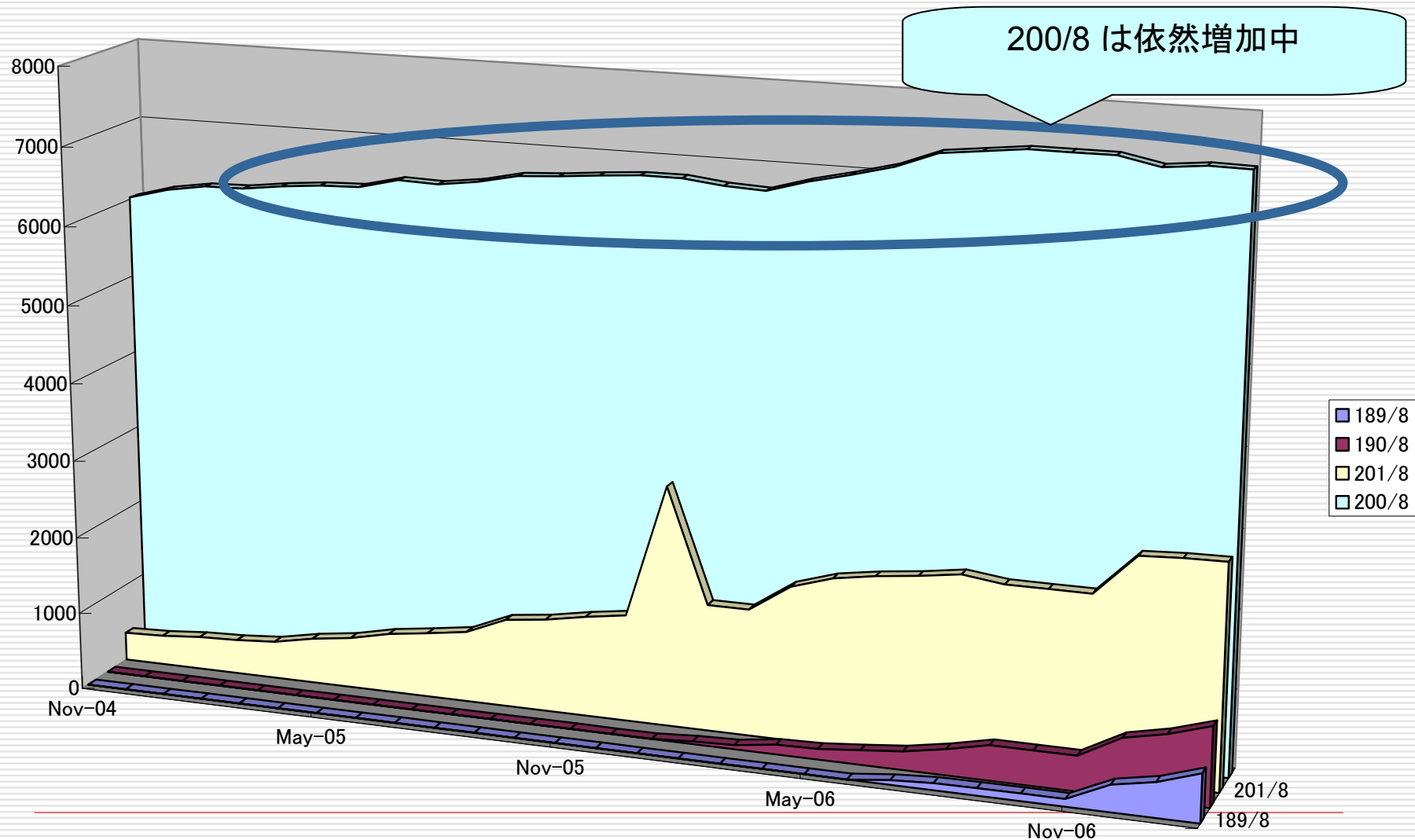
RIPE地域の / 8毎のトレンド



ARIN地域の / 8毎のトレンド

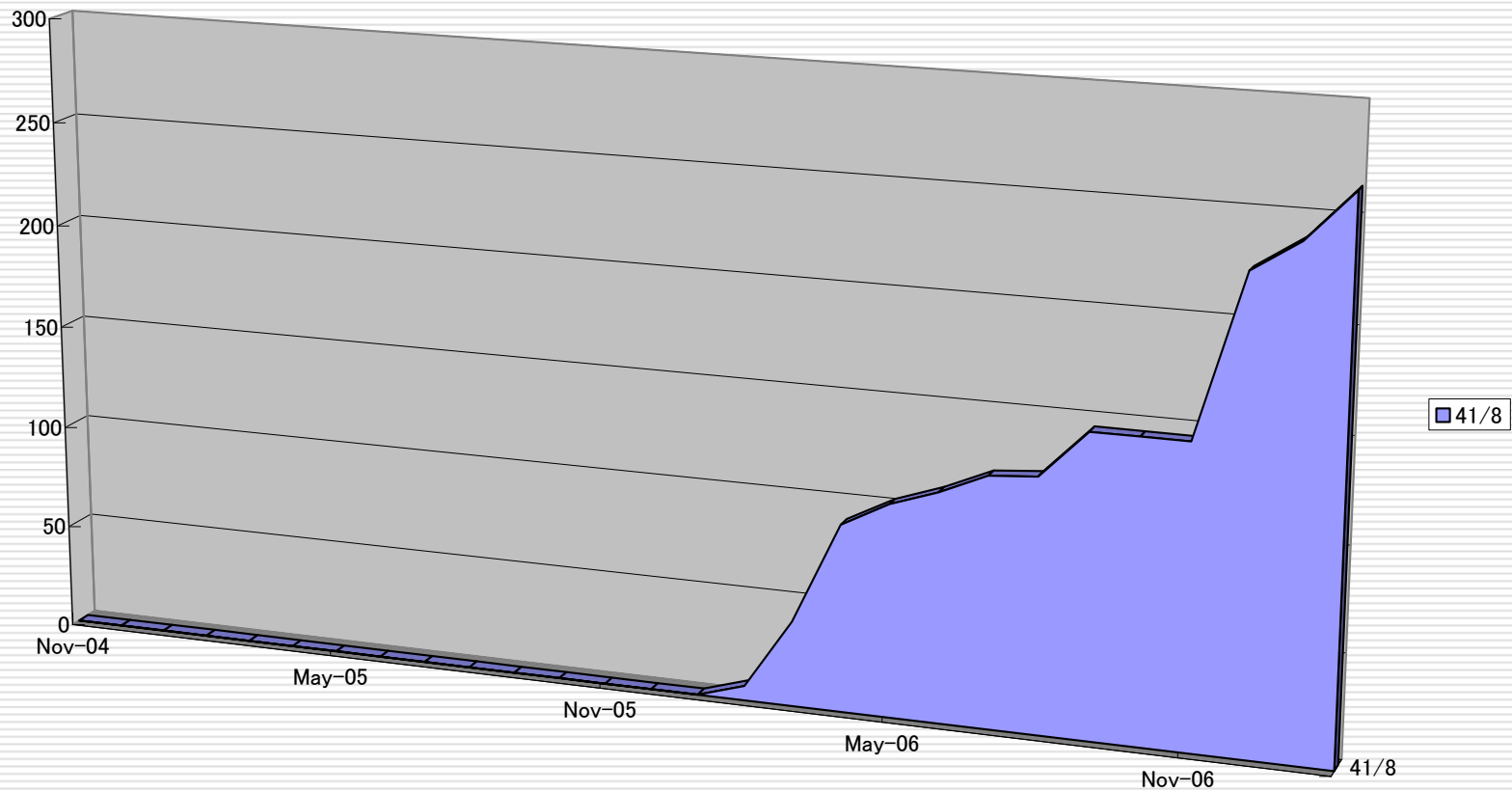


LACNIC地域の / 8毎のトレンド



AfriNIC地域の / 8毎のトレンド

41/8



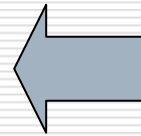
2. BGPでトラフィック制御編

トラフィック制御の手法例

□ Path Attribute

- Local Preference

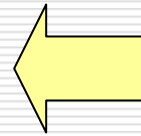
- MED



優先度による制御

□ IGP cost

- closest exit



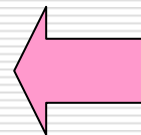
トポロジに応じた最適化

□ eBGP multihop

□ multipath

- iBGP

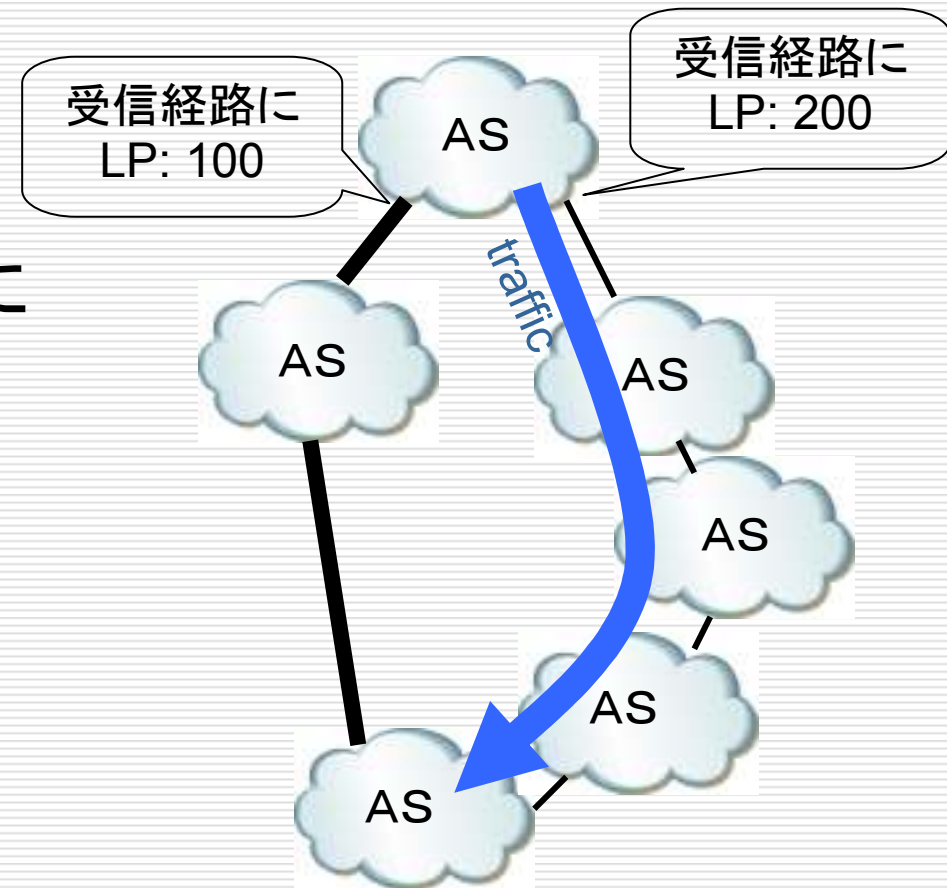
- eBGP



負荷分散

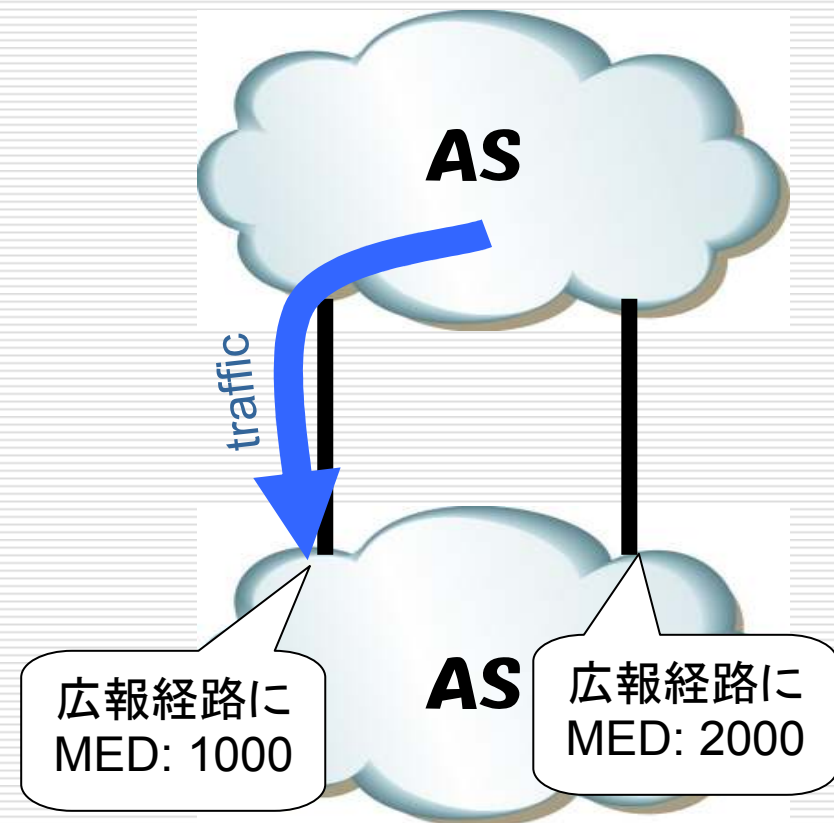
Local Preference

- 経路受信時に設定
 - default: 100
- 主に優先度を明確にしたい箇所
 - 顧客、ピア、上流
- 強すぎるので注意
 - AS Pathより強い



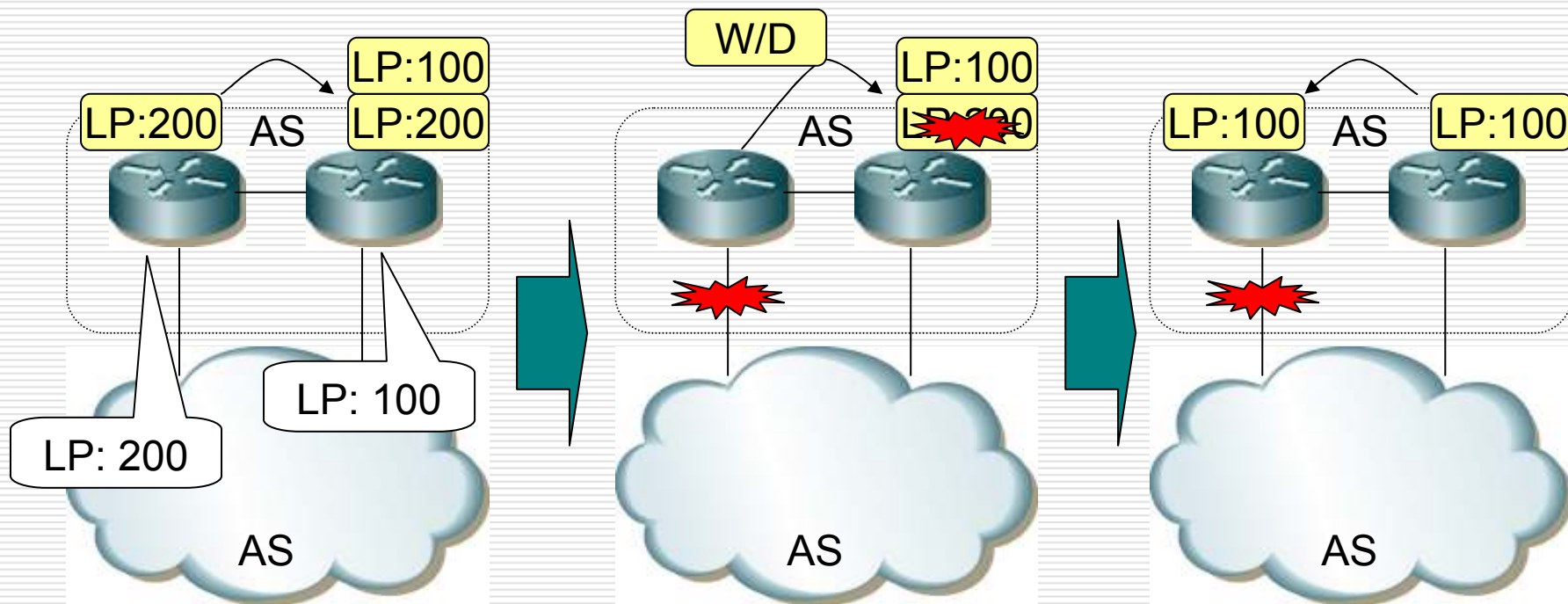
MED

- 付けても良い
 - 無ければ0と見なすか
設定によってはmax値
- 程よく弱い
 - AS Pathより弱く、IGP Costよりは強い
- 網内で一貫した扱いを
 - always-compare?
 - deterministic?



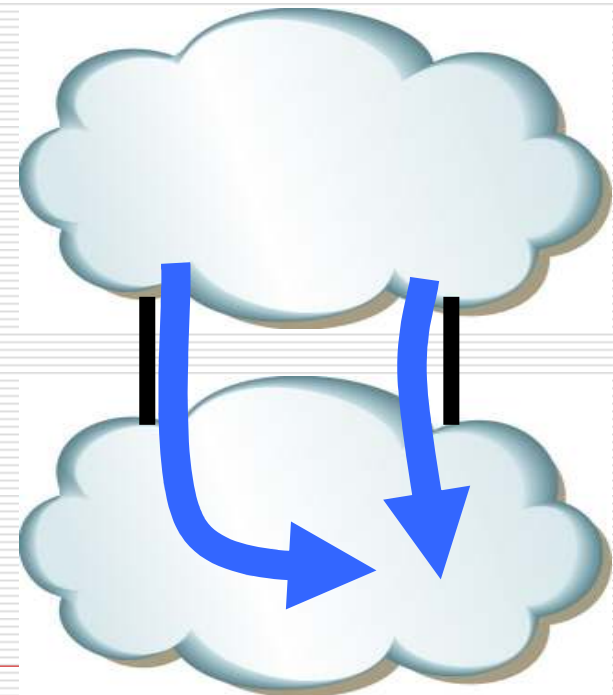
優先制御の憂鬱

- BGPの特性上、収束にちょっと時間がかかる
 - 優先度の高い経路が消えてから、次が見える



closest exit

- 大好き 😊
- 同じ規模、同じ傾向のAS/ネットワークの相互接続であれば、これが一番
- シンプルなので障害時の切り分けも楽だし、困ったことにならない
- 網内に複数Pathが伝播する



closest exitの維持

□ シンプル一番

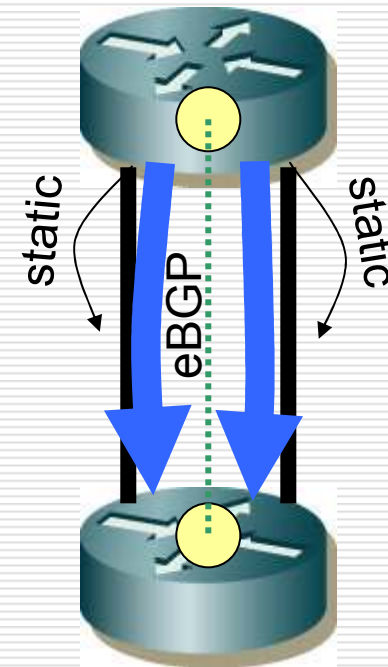
- 細かい制御はしないと腹をくくる
- とにかく一貫したLocalPrefとMEDの管理
- 網のIGP設計のみで維持

□ 地域分割

- confederationで地域分割
- 地域ごとのroute reflectorが経路交換するときにMEDを+2000とかしちゃう
- 網の拡張の際に注意が必要

eBGP multihopで負荷分散

- 古の負荷分散手法
 - 最近流行らない
- 設定が煩雑
 - お互いのloopbackをstaticで向け合う
 - eBGPのIP TTLを増やす
 - loopback同士でeBGPを張る
- 障害時は怖かった
 - 回線が切れた時の状態遷移にいろいろと...

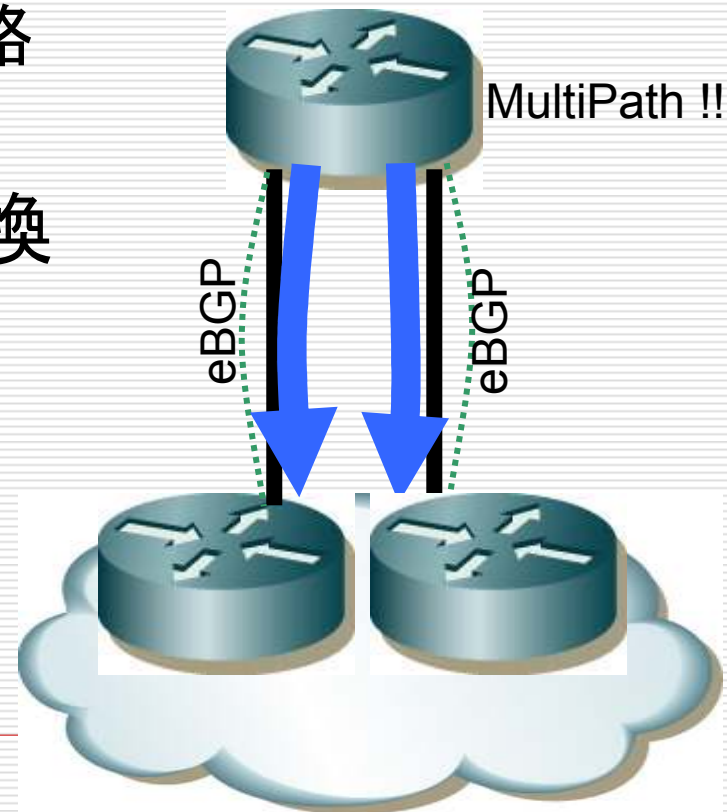


BGP multipath

- Best Pathが2つあってもいいじゃない！
 - BGPのBest Path selectionを少しリラックス
 - LocalPreference, AS Path長、MEDは同じだけど next_hopの異なるBGP Pathを同時に採用
 - どこまでをmultipathに含めるかは実装依存
 - iBGPに広報するときは、通常通り1つのBest Pathのみ
 - 局所的に複数Pathを利用できる
 - eBGP multipathだとnext_hop attributeを書き換えてiBGPに広報する人もいる

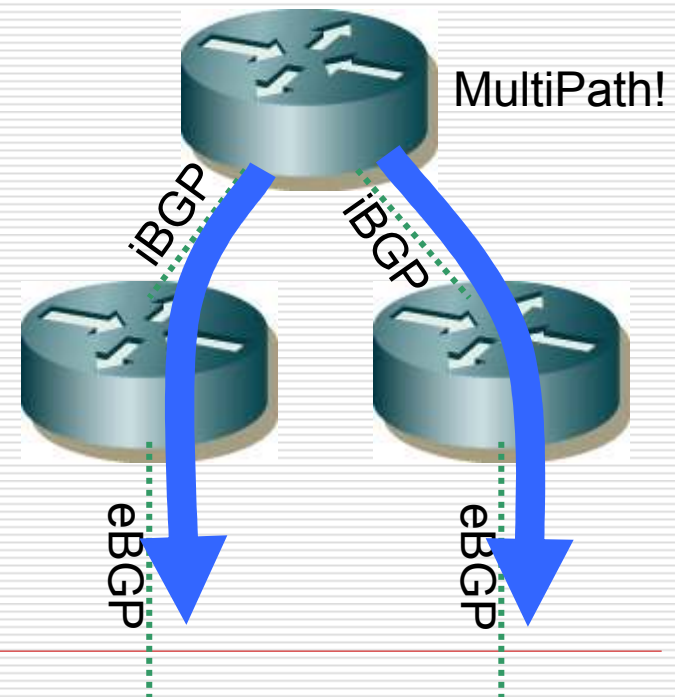
eBGP multipath

- eBGPから聞いた複数Pathを同時に採用
 - どちらを使ってもパケットが届くなら、両方採用！
 - Multipath対象になる経路は実装依存
- eBGP multihopの置き換えに便利
 - 複雑な設定がいらぬ
 - 片方向だけでも使える



iBGP multipath

- iBGPから聞いた複数Pathを同時に採用
 - どちらを使ってもパケットが届くなら、両方採用！
 - Multipath対象になる経路は実装依存
- POP内で複数のゲートウェイを利用する場合に便利



経路制御の苦悩

□ closest exitは正義

- 簡単なポリシーで最適な経路制御になりうる
- 細かい経路制御のあきらめも肝心

□ でも、帯域は限られてる

- 思い出されるIX接続帯域の移り変わり
 - FE → GbE
 - GbE → 10GE
 - 10GE → ????

□ ポリシを維持しつつ、足りない所はやりくり

ISP間でバックアップの難しさ

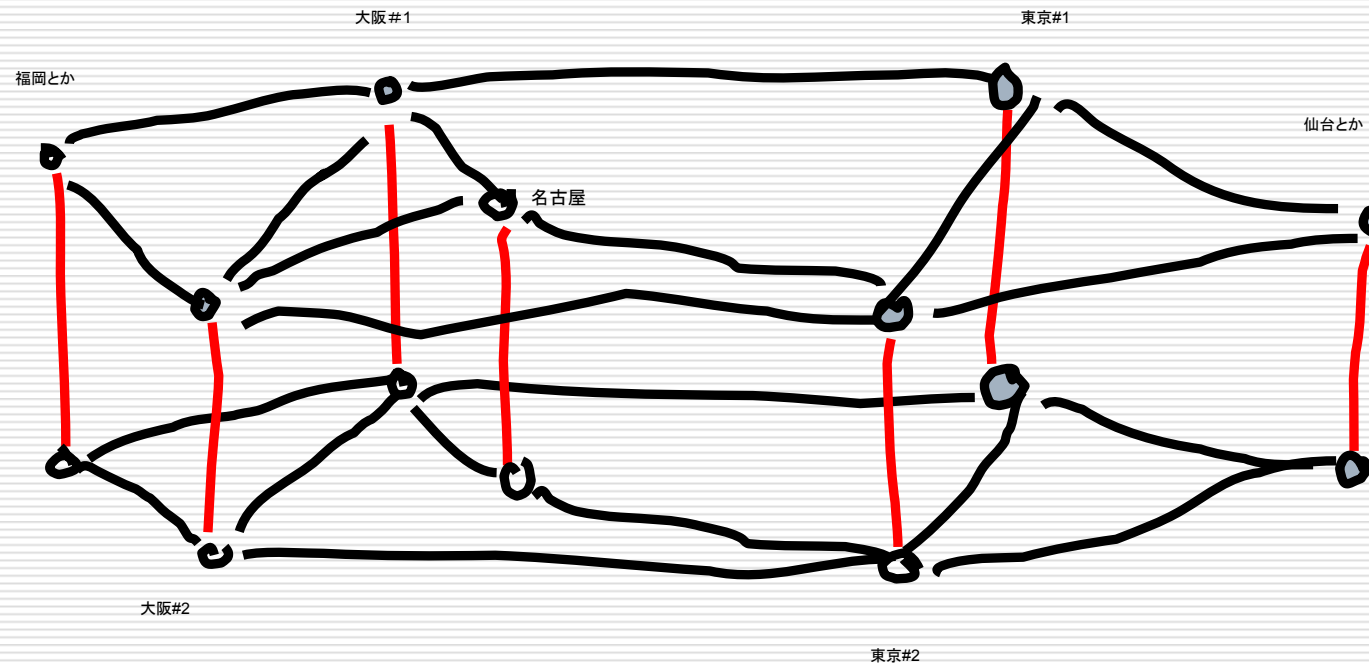
- 主に設備/お金問題
 - インターフェース帯域、ポート数
 - お互いの設備投資の都合やタイミング
 - 相互接続の回線費用
- 基本ポリシーは日本的
 - 東は東でバックアップ
 - 西は西or東でバックアップ

最近のISP間トラフィック制御

- 基本はclosest exit
- 複数本で接続して負荷分散
 - 構成的に可能ならばmultipath等で分散
 - 他に手段が無ければ、MEDなどを駆使してトラフィックを制御するが、それでも地域的なclosest exitは維持
 - 10GEのリンクアグリゲーションがそろそろ普通な時代に
 - さらにそこでmultipath...
- 通常時はプライベートな接続を利用し、バックアップ用としてIXでも接続しておく

ゆくゆくはこんな接続を希望

- 全地域で相互接続...できるといいなあ



3. BGPタイマー編

Keepalive

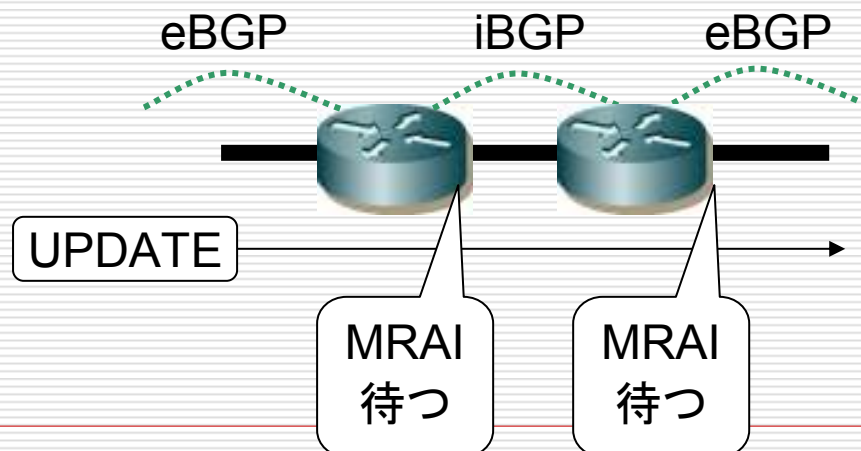
- UPDATEが無いときに定期的を送信
 - ホールドタイム(通常keepaliveの3倍)の期間、何もメッセージが届かなければ、ピアダウン
- keepalive/holdtimeにはいくつかの流儀が
 - 30秒/90秒派(Juniper?)
 - 60秒/180秒派(Cisco?)
 - もっとチューニング派
 - ちなみにIIJでは10秒/40秒で元気に運用中

最近のkeepalive

- 短くしても大きな問題は無し
- でも、これ以上頑張る？
 - 最近では短くしすぎるのを禁止する箱もある
 - BFDとかの方がルータに優しそう
 - たぶん、keepaliveはそこそこで、頑張るところはBFDを実装するのが生きる道
 - OCNではInterop接続時にBFD for BGP
 - 1000msec間隔で3発
 - 特に問題は無し

MRAI timer

- UPDATEの送信間隔を決めている
 - Ciscoの(古い)実装ではかなり長め
 - iBGPに5秒
 - eBGPに30秒
- RR階層とかあると、経路の伝播が遅い...



最近のMRAI

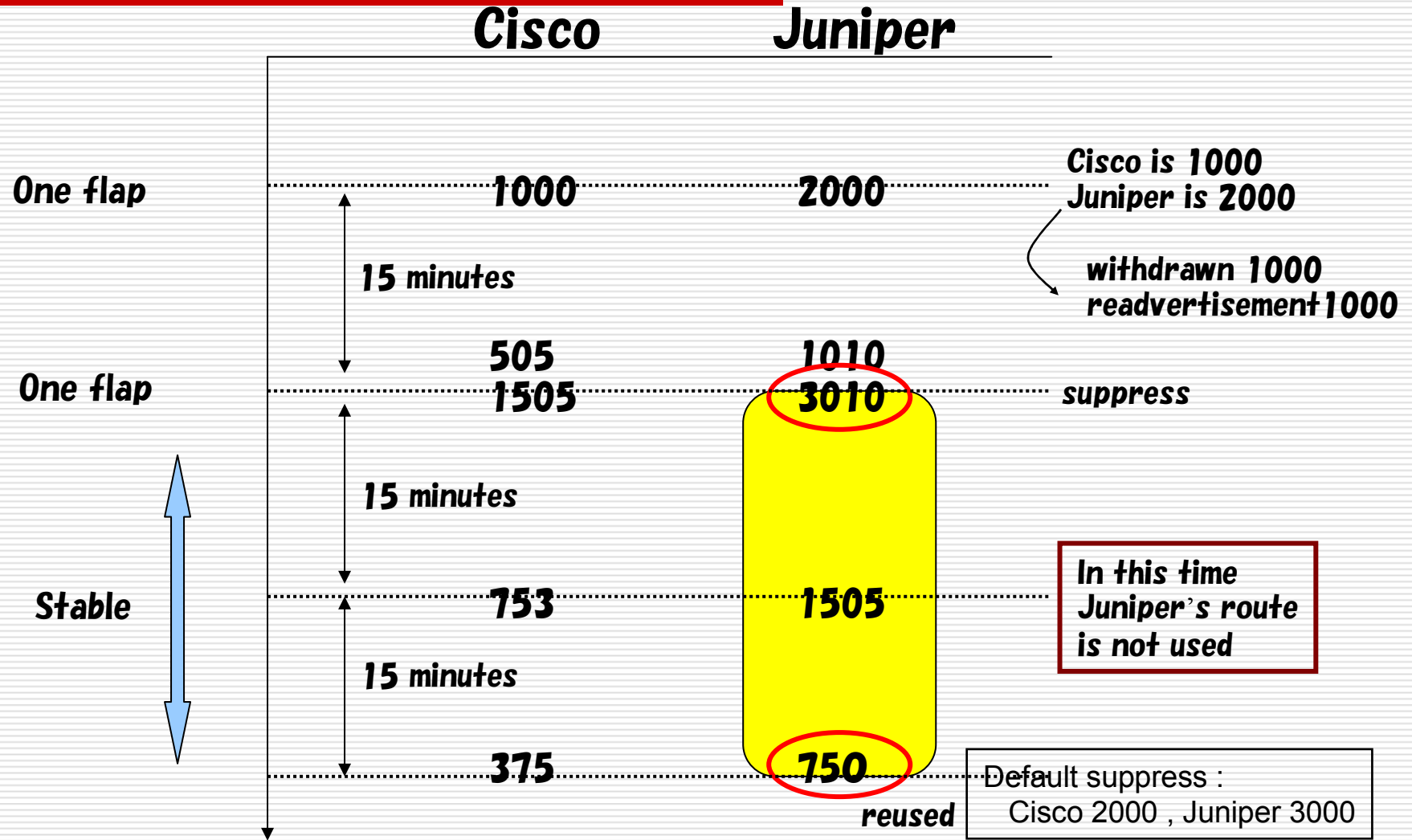
- 0sec for iBGP
- 30sec for eBGP

- Cisco
 - 12.0(32)S or R3.3以降使えば自動的にこの値
 - 設定不可
- Juniper
 - MRAIは標準で全部0sec

bgp damping

- ばたつく経路を無効にして、経路計算の負荷を控えよう
 - タイマーの設定がベンダによって異なる
 - ベンダによって無効になるタイミングが違う
- 実装が使いにくい

bgp damping(続き)



最近のbgp damping

- 使ったり使わなかったり
 - そもそもIIJでは実装しなかった
 - OCNでは場所に応じてリソース消費を考慮して弱めに実装(顧客接続部分は実装していない)
- ばたつく経路があるとMRAIが効いてくるので、迂回性能が悪くなる可能性あり
 - 要検討
 - ばたつく経路には、もっとグローバルなアプローチが欲しいかも

他

□ MD5がほぼ標準になった

■ 2004年のTCP Vulnerability以降

□ 日本のISPに集まってもらって問題解決

□ http://www.bugest.net/irs/docs_20040707/Tcp_Vulnerability_yoshida.pdf

■ eもiも頑張った

□ Cluster-ID

■ 昔の教科書では、同一階層化の冗長ルータで同一

■ 今の基本はloopback(ルータID)

4. これから編

経路数の増大

- 線形以上に増大中
- 現在、約22万経路
 - 半分ぐらいは/24
 - 綺麗に集約すると22万の半分ぐらいになる
 - <http://www.cidr-report.org/>
 - Route Aggregation
 - <http://www.ripe.net/docs/ripe-399.html>

細かい経路を減らす努力

- /25より細かい経路を受け取らない(/25含む)
 - 細かい経路を受け取る必要がなかった
 - IIJ/OCN とともに数年前より実装済み
- PAは全て割り振られたサイズで広報
 - 細かい経路をインターネットに流さない
 - IGPで分割されている経路もaggregateして広告

AP地域、頑張ろう

□ APNIC地域

- de-aggregation routes が多い
- 1つのprefixのみを広告するASの率は低い

	TOTAL	APNIC	ARIN	RIPE	AfriNIC	LACNIC
BGP routing table entries	213110	48310	104128	44039	2527	14058
after maximum aggregation	114382	19415	61213	28738	964	4050
Deaggregation factor	1.86	2.49	1.7	1.53	2.67	3.47
Total ASes present	24455	2877	11364	9195	180	760
Origin ASes announcing only one prefix	10282	782	4363	4831	59	246
	42.0%	27.2%	38.4%	52.5%	32.8%	32.4%

* BGP Routing Table Analysis Reports *

Routing Table Report 04:00 +10GMT Mon 26 Feb, 2007

BGP経路情報の信頼性向上

- Secure Inter-domain Routing WG
 - Inter-domain Routingにおけるセキュリティフレームワークアーキテクチャの検討
 - 拡張可能なインタードメインにおけるセキュアルーティングアーキテクチャの文書化
 - セキュアルーティングアーキテクチャに含まれる証明書の利用方法に関する文書化
 - RPSEC WGにより決定された安全なルーティングへの要求に焦点をおいた、このアーキテクチャが提供するルーティング機能の構成要素の文書化
 - 2006年末よりWG化
 - <http://www.ietf.org/html.charters/sidr-charter.html>

もろもろの課題

- 帯域
 - 10G Link-aggregation → 100GE → 1TE?
 - 100Gが出るころにはもうおなかいっぱいかも
- IX(eBGP)上での高速切断検知技術の追及
 - BFD for BGPをそろそろ実験
- 4-octet ASNのDeployment
 - AS23456(AS_Trans)は単なる移行措置だろう(と僕らは思っている)
 - 3年以内ぐらいに全て4-octet ASN対応になる(と僕らは信じている)
- IPv6 NW へのマイグレーション

もろもろの課題2

□ 経路の切り替わり

- 20万経路の切り替わり。。。
- 40万経路の切り替わり。。。
- IPv4+IPv6
- やっぱりaggregationしようよ運動
- Intraも大事だけど、Inter-ASの切り替わりももっと頑張ろう

**日本のすごさをもっと国際的にアピール
できるように、今後とも頑張って参ります**

おしまい