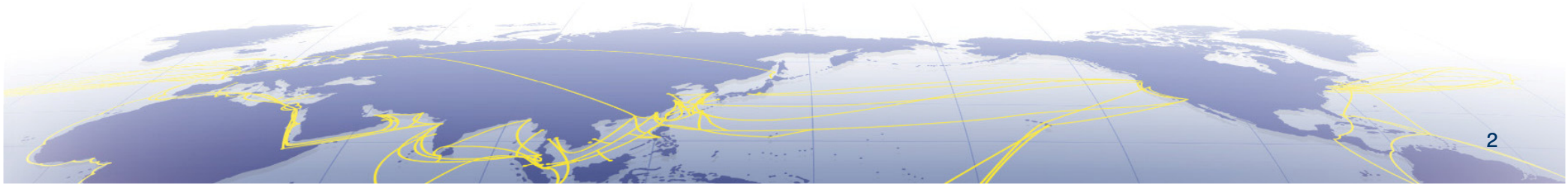
ますます増えるトラフィック！
ISPバックボーン設計の過去、現在、未来

友近 剛史
吉村 知夏
NTT Communications, OCN

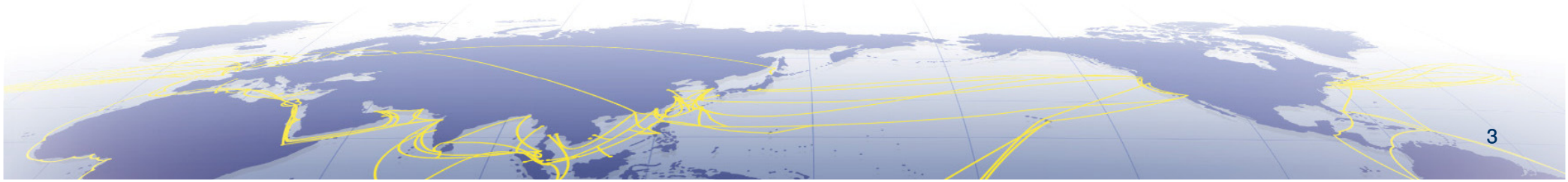
はじめに

- インターネットトラフィックは順調に増加
- 増えゆくトラフィックを輻輳なく、安定して運ぶことがISPの最大のミッション
  - 設計上の努力が欠かせない
- このセッションでは、
  -インターネットトラフィックの現状
  -OCNにおける、過去～現在の具体的な設計事例、課題を共有し、未来の設計についてディスカッションします



今回のフォーカス

- 話します
  - –設計技術
  - –トラフィック状況
- 話しません！！
  - –政治
    - –Peering戦略とか
    - –Peerの構成とか
  - –法律
    - –帯域制御の是非とか



アウトライン

1. 日本のインターネットトラフィックの現状

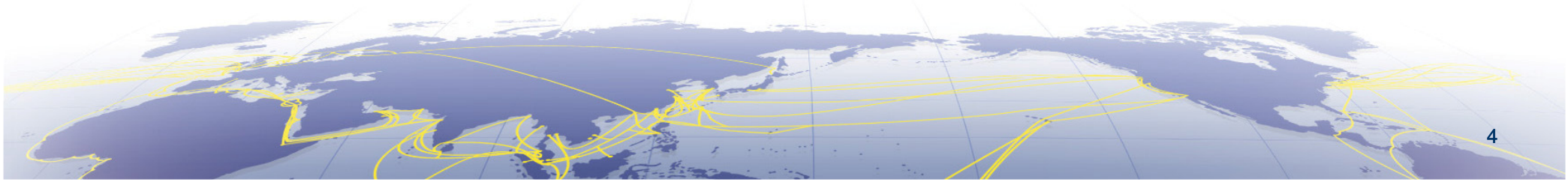
2. OCNの変遷（サマリ）

3. 過去の設計

4. 現在の設計

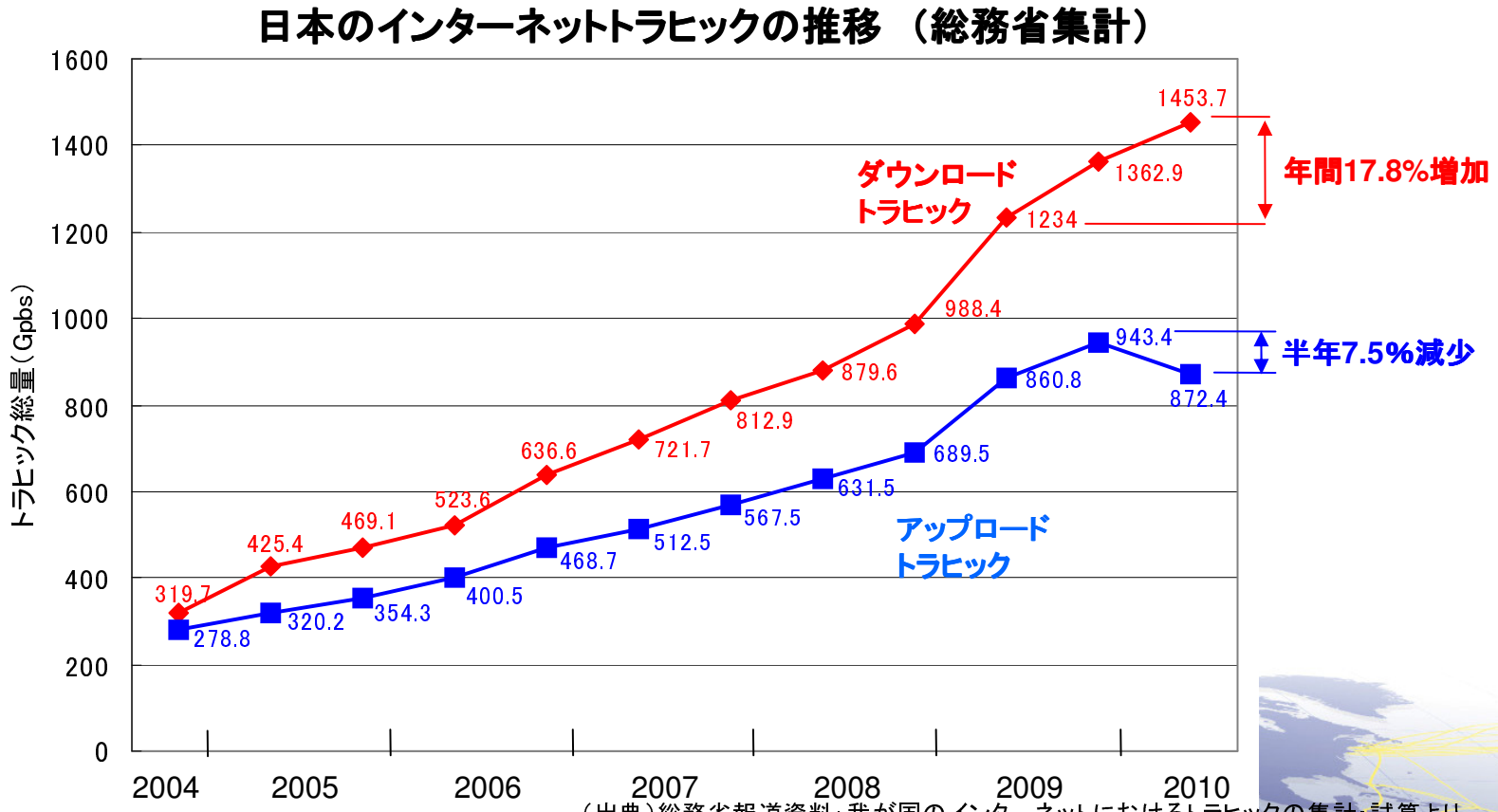
5. 未来の設計、構想

6. まとめ



日本のインターネットトラヒック推移

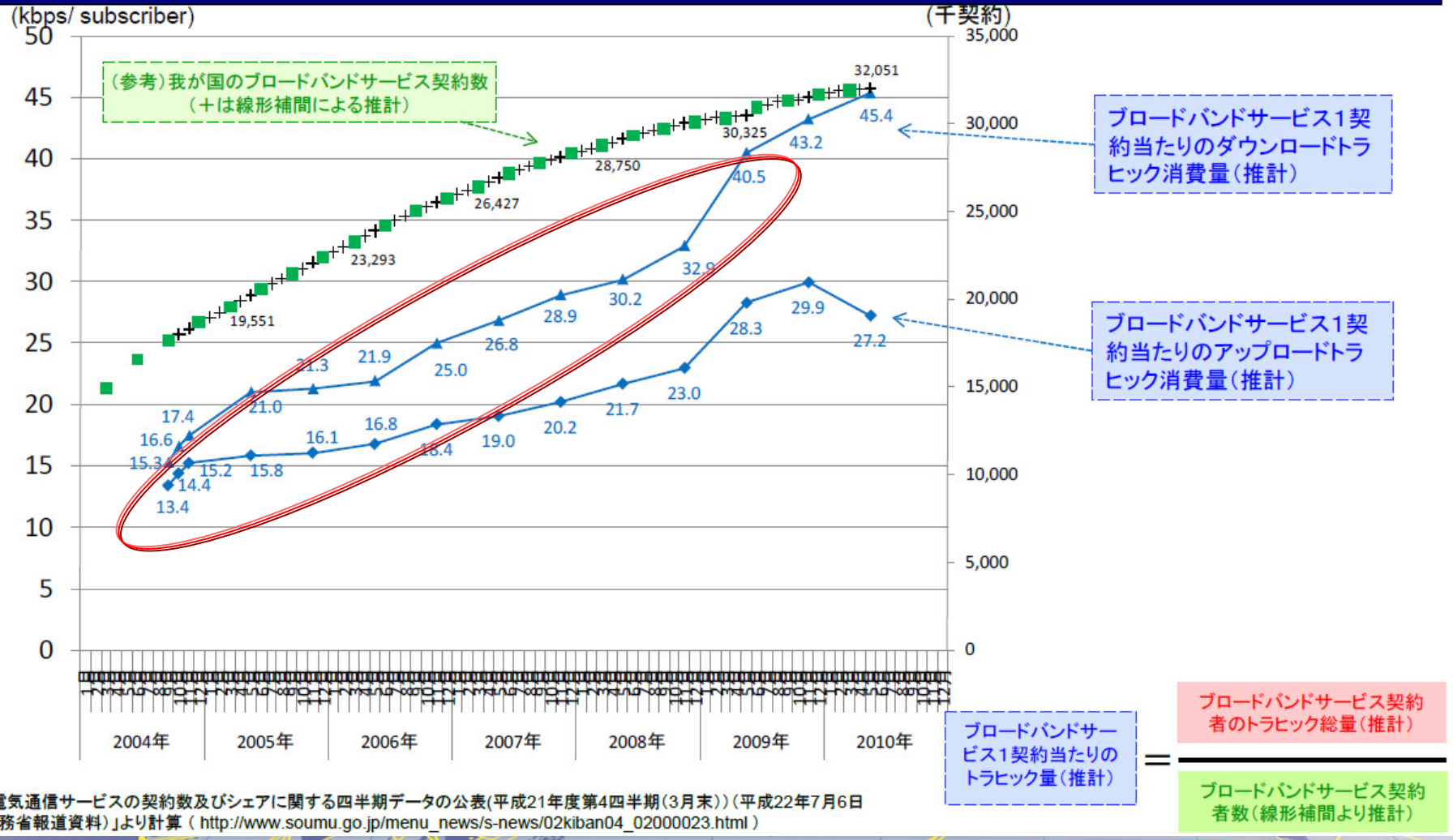
- 日本のブロードバンド契約者のダウンロードトラヒックの総量は1.46T(テラ)bps。1年で1.2倍、5年間では3.4倍の増加(ダウンロードトラヒック)
- アップロードトラヒックにおいては、初めて減少(872Gbps)



(出典)総務省報道資料:我が国のインターネットにおけるトラヒックの集計・試算より
http://www.soumu.go.jp/menu_news/s-news/01kiban04_01000001.html

日本のインターネットトラヒック推移

- 一契約あたりのダウンロードトラフィックも増加傾向
  15.3kbps（2004）→ 45.4kbps（2010）

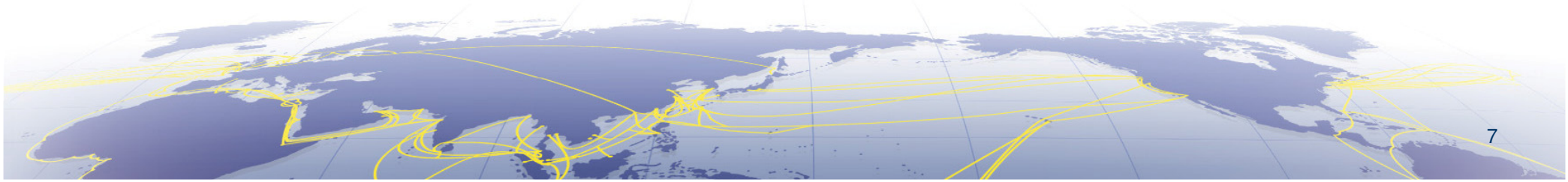


「電気通信サービスの契約数及びシェアに関する四半期データの公表(平成21年度第4四半期(3月末))(平成22年7月6日総務省報道資料)」より計算（ http://www.soumu.go.jp/menu_news/s-news/02kiban04_02000023.html ）

【出典】我が国のインターネットにおけるトラヒックの集計・試算 http://www.soumu.go.jp/menu_news/s-news/01kiban04_01000001.html

トラフィック状況　サマリ

- 2009年11月→2010年5月のアップロードトラフィックの減少
  - 2010年1月　著作権法改正（P2Pトラフィックの減少）
  - アップロードの帯域制御
- 昨今のダウンロードトラフィックの増加
  - HTTPストリーミング（動画視聴）
  - リッチコンテンツの更なる流行（楽曲、映画、アプリケーション）
- 日本のインターネットトラフィックは順調に増加
  - おそらく今後も増え続ける
- ISPはどんな勢いで増速してきたか？
  - OCNの例



アウトライン

1. 日本のインターネットトラフィックの現状

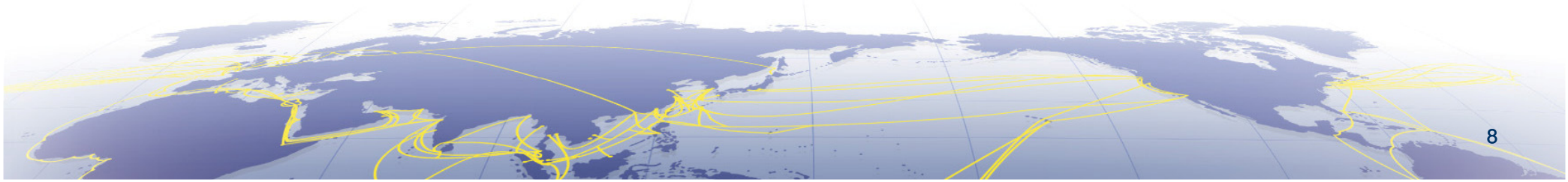
2. OCNの変遷（サマリ）

3. 過去の設計

4. 現在の設計

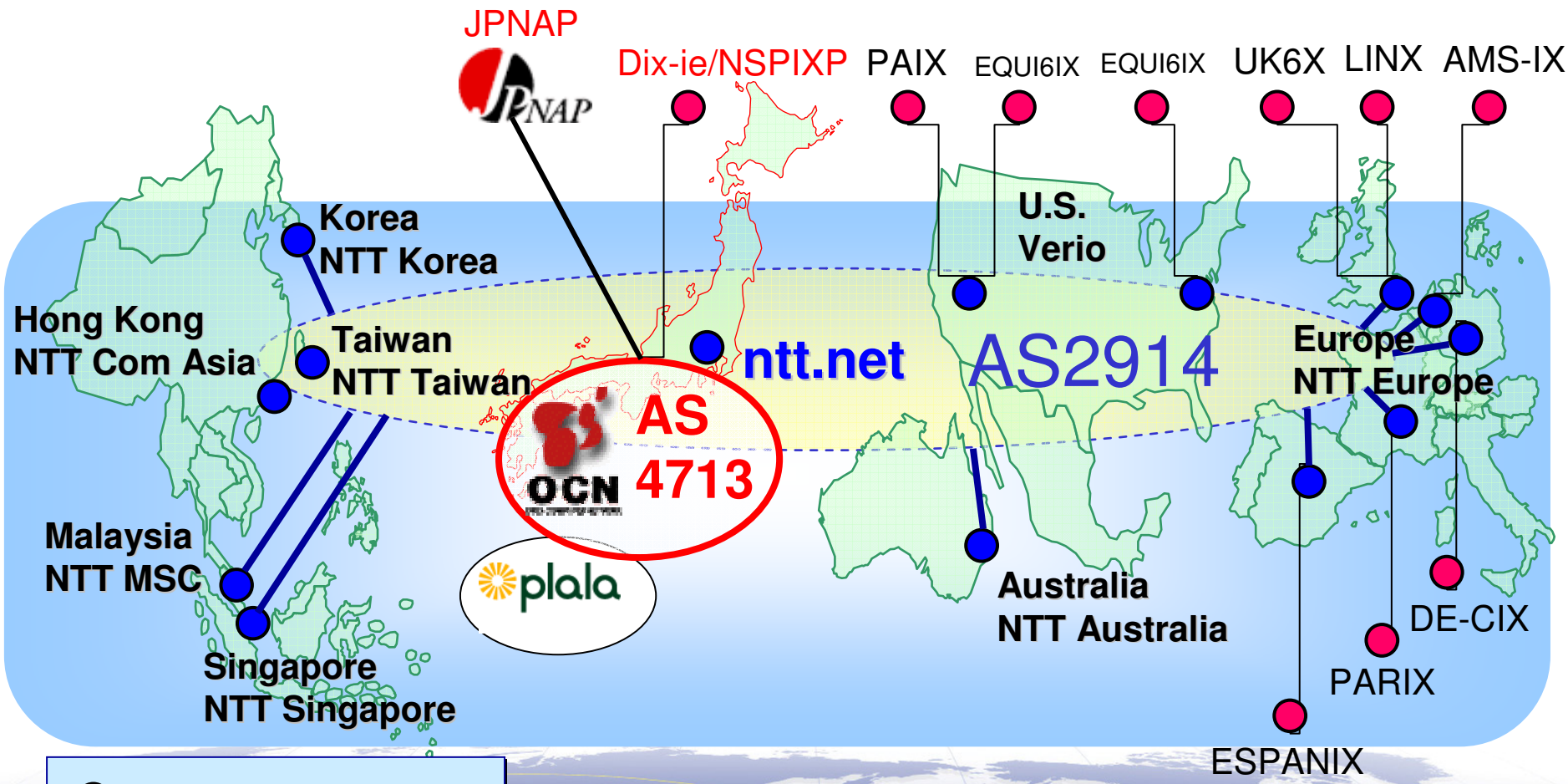
5. 未来の設計、構想

6. まとめ



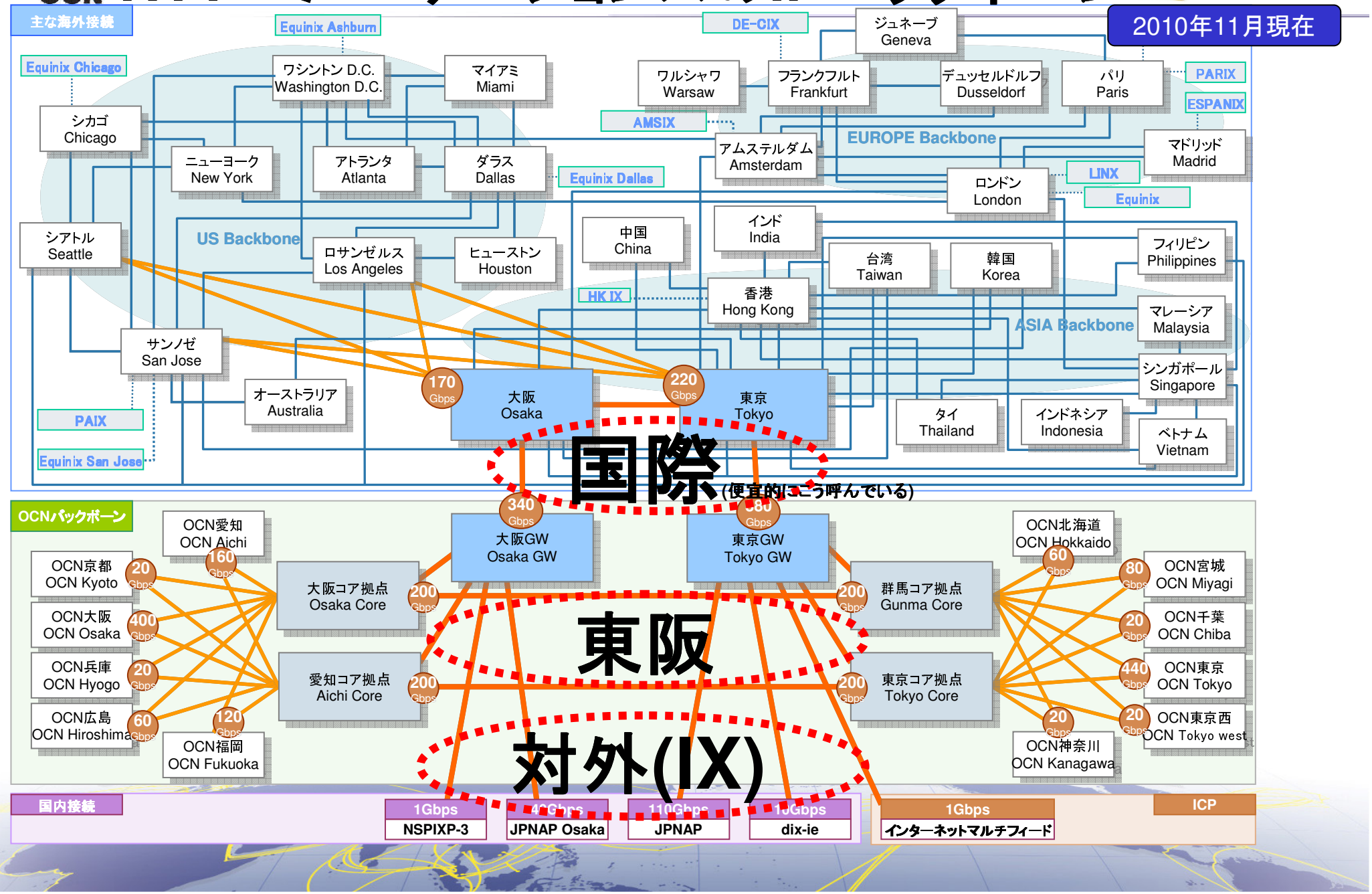
ntt.net と OCN

- 今回はOCN（AS4713）にフォーカス

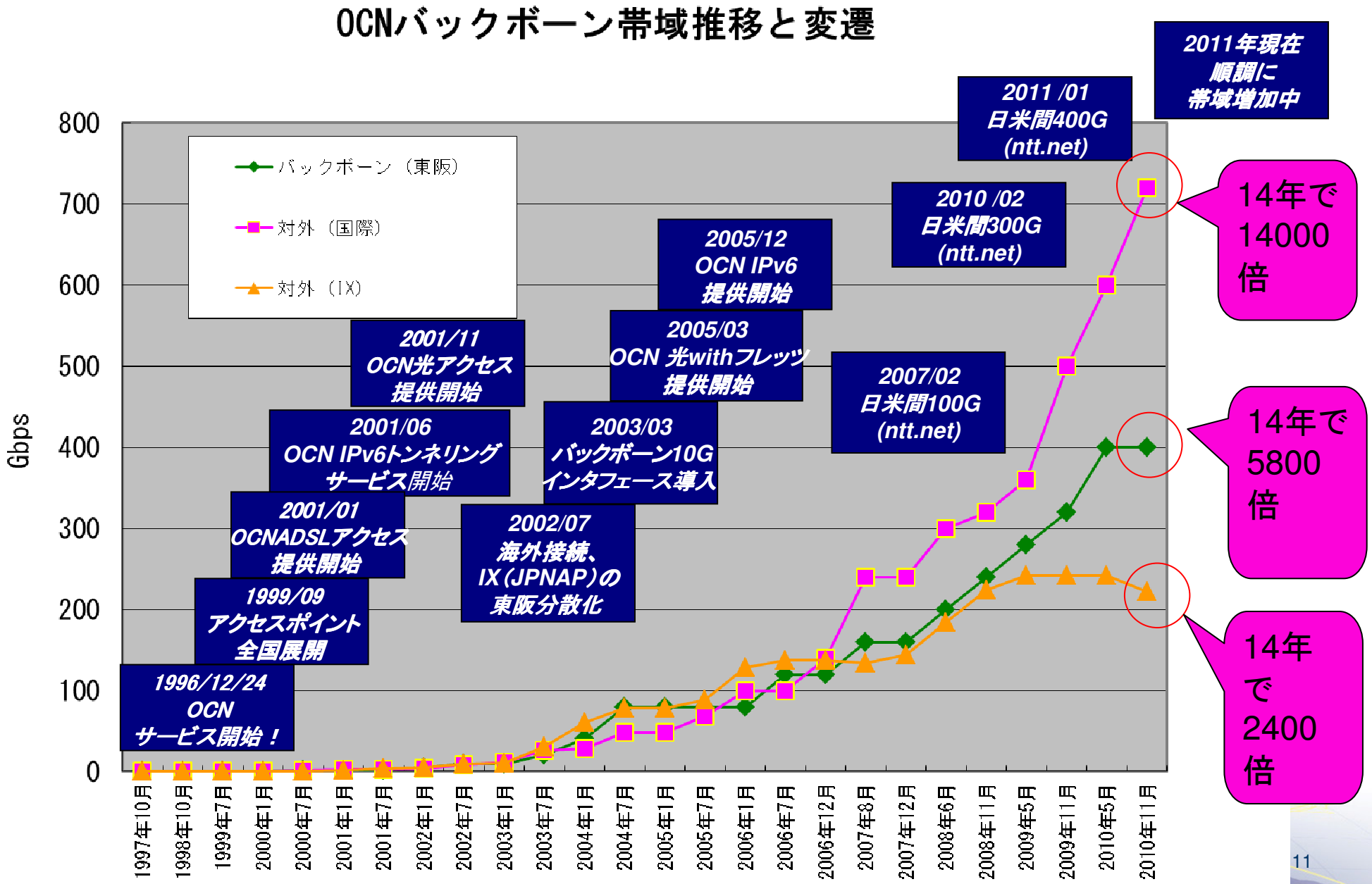


NTTコミュニケーションズのIPバックボーン

2010年11月現在

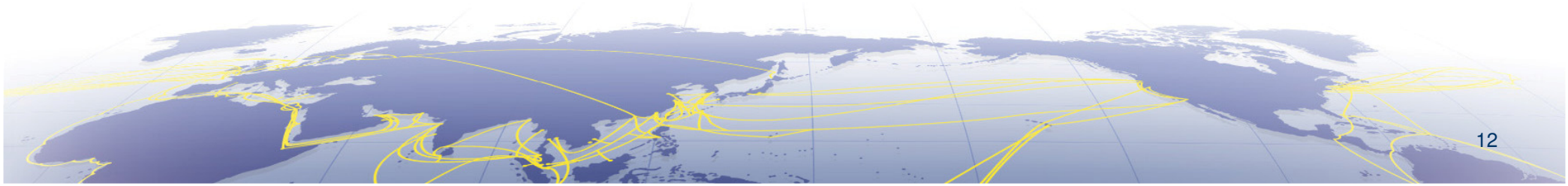


OCNの変遷（サマリ）


OCNバックボーン帯域推移と変遷
Gbps
800
700
600
500
400
300
200
100
0
バックボーン（東阪）
対外（国際）
対外（IX）
1996/12/24 OCN サービス開始！
1999/09 アクセスポイント 全国展開
2001/01 OCNADSLアクセス 提供開始
2001/06 OCN IPv6トンネリング サービス開始
2001/11 OCN光アクセス 提供開始
2002/07 海外接続、IX（JPNAP）の 東阪分散化
2003/03 バックボーン10G インタフェース導入
2005/03 OCN 光withフレッツ 提供開始
2005/12 OCN IPv6 提供開始
2007/02 日米間100G (ntt.net)
2010 /02 日米間300G (ntt.net)
2011 /01 日米間400G (ntt.net)
2011年現在 順調に 帯域増加中
14年で 14000 倍
14年で 5800 倍
14年 で 2400 倍
1997年10月
1998年10月
1999年7月
2000年1月
2000年7月
2001年1月
2001年7月
2002年1月
2002年7月
2003年1月
2003年7月
2004年1月
2004年7月
2005年1月
2005年7月
2006年1月
2006年7月
2006年12月
2007年8月
2007年12月
2008年6月
2008年11月
2009年5月
2009年11月
2010年5月
2010年11月

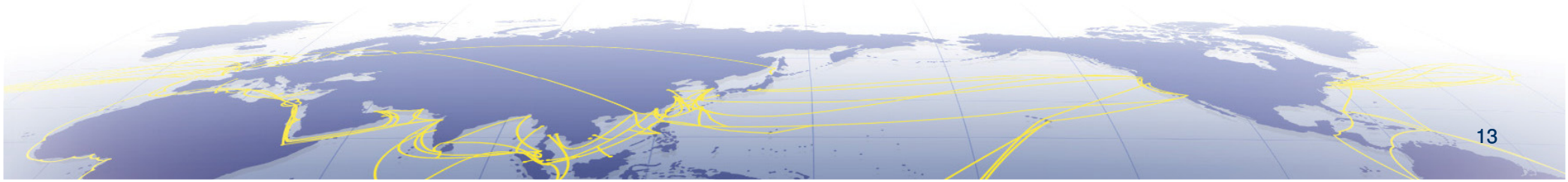
設計上の課題、苦労

- トラフィックの増加に伴ってバックボーンを拡張している
- バックボーンの設計にあたって、
  - どんな課題にぶつかったか？
  - どんな工夫をしてきたか？
- 過去から最近まで、ざっと見てみましょう！



アウトライン

1. 日本のインターネットトラフィックの現状
2. OCNの変遷（サマリ）
3. 過去の設計
4. 現在の設計
5. 未来の設計、構想
6. まとめ



過去の設計上の課題、工夫

1. メディアの変遷
   - ATM、FDDI、Fast Ethernet

2. トラフィック増
   - 増設、新装置

3. 論理的な課題
   - OSPF経路の増大
     → static to BGP

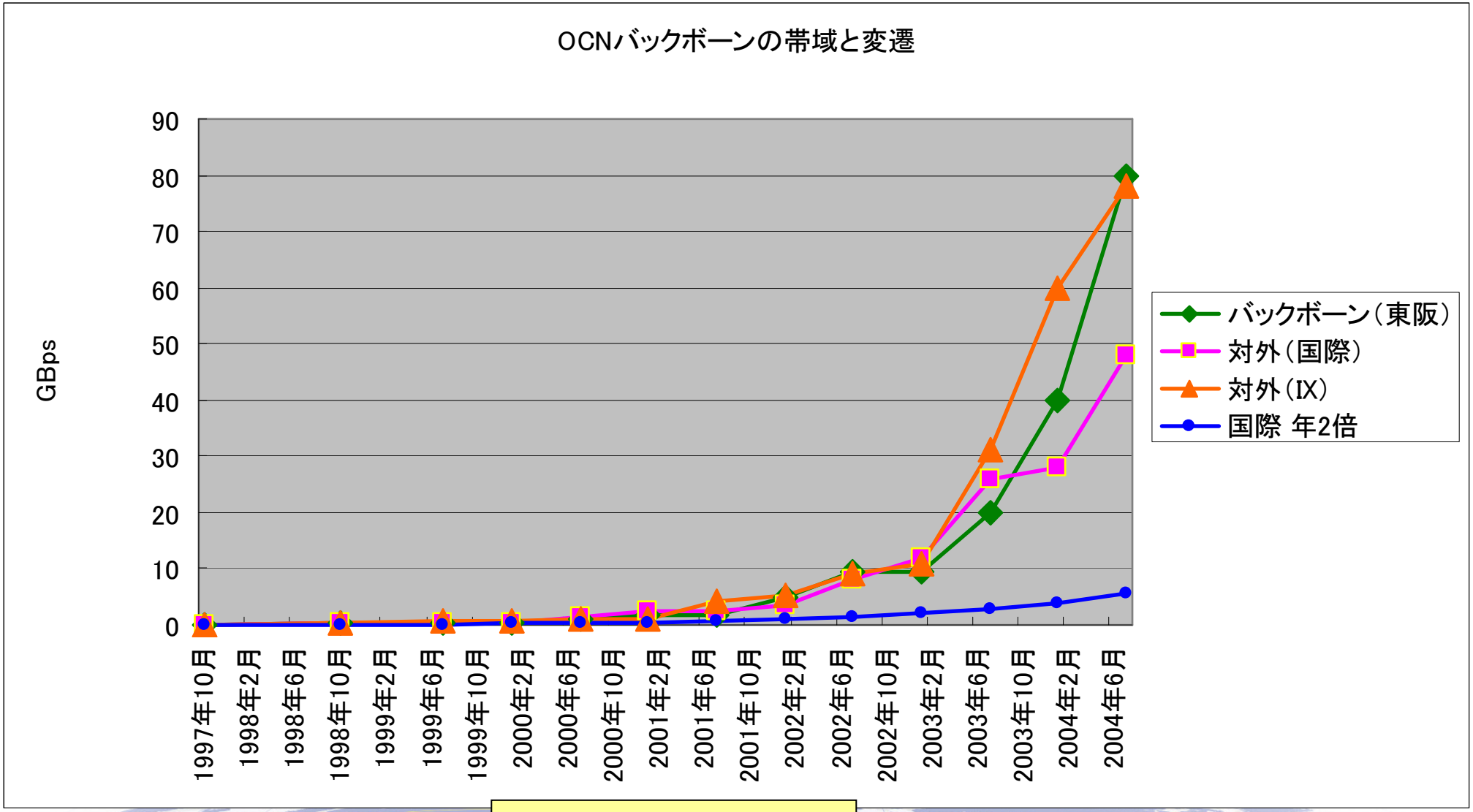
4. 信頼性向上
   - コアPOPスクエア化
   - 激甚災害耐性化

2. トラフィック増：OCNバックボーン帯域の変革

TIME	OCNバックボーン（東阪）	対外接続（国際）	対外接続（国内IX）
2004.07	80. 0G	48. 0G	78. 0G
2004.01	40. 0G	28. 0G	60. 0G
2003.07	20. 0G	26. 0G	31. 0G
2003.01	9. 6G	12. 0G	11. 0G
2002.07	9. 6G	8. 0G	9. 0G
2002.01	4. 8G	3. 6G	5. 2G
2001.07	1. 8G	2. 5G	4. 3G
2001.01	1. 8G	2. 5G	1. 2G
2000.07	1. 2G	1. 3G	1. 0G
2000.01	500M	500M	600M
1999.07	375M	400M	600M
1998.10	270M	300M	500M
1997.10	69M	51M	91. 5M

7年弱で、東阪 1,159倍、国際 941倍、国内IX 852倍

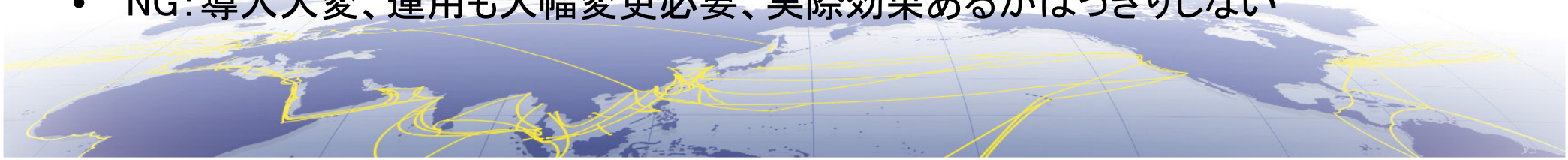
OCNの変遷（サマリ）



- 増設
- 新装置

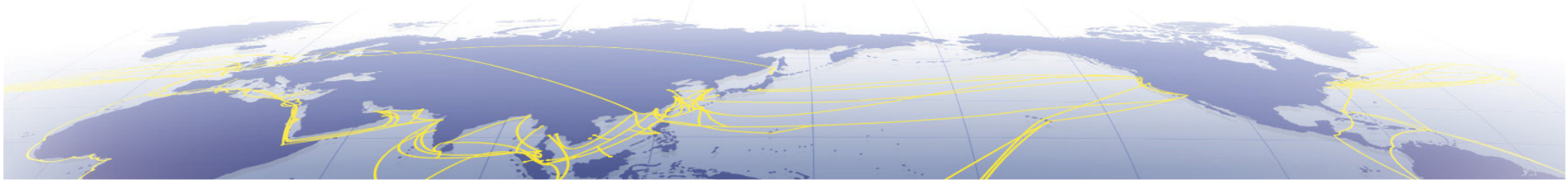
3. 論理的な課題
OSPF経路の増大と対策 1/2

- **ユーザの伸びが激しかった**
- OSFP経路数の増大
  - OCNでは経路数が非常に増大していた
    - 1998年8月21日：16557（そのうちExternalが15546(94%)）
    - 1998年9月21日：17346（そのうちExternalが16275(94%)）
- Static経路をOSPFにRedistributeしていた
  - Confederation検討
- OSPF分割
- NG：運用負荷増大
  - IS-IS検討
- JANOG IS-IS WGとかも
- NG：導入大変、運用も大幅変更必要、実際効果あるかはっきりしない



3. 論理的な課題
OSPF経路の増大と対策 2/2

- Static経路をOSPFでなく、iBGPへ直接Redistributeした
  - external経路はBGP、トポロジはOSPFで管理
  - static設定しているルータでBGPを話した
  - その他
    - no export communityを利用
    - Route Reflectorの階層化
    - Full routeが不要なところではfilter
- 結果：内部ルーティングの安定性の向上が見られた
  - 運用の変更もほとんど無し
- JANOG4で発表

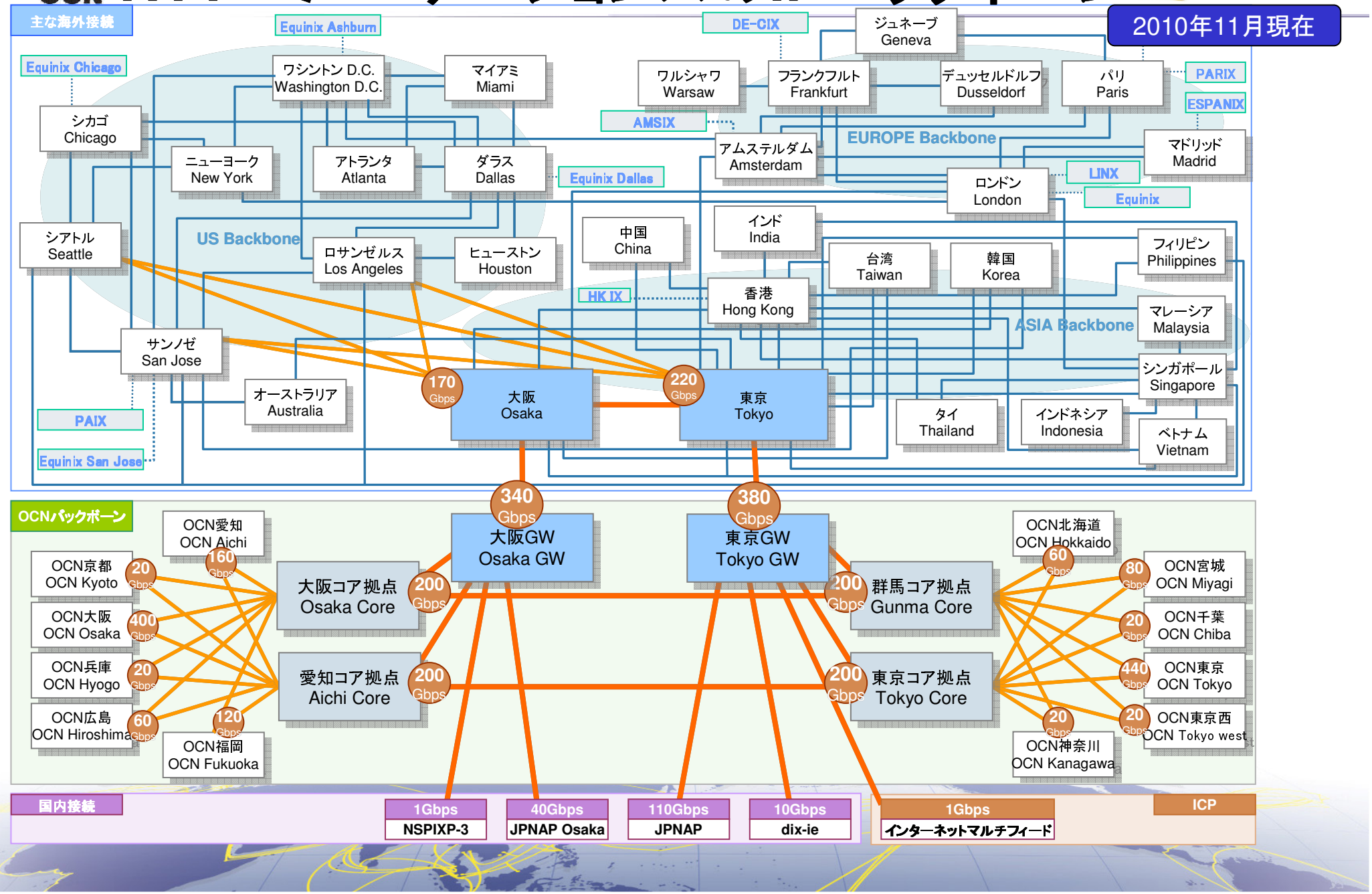


4. 信頼性向上

- インターネットの普及が進み、信頼性向上がますます必要になってきた
- 全経路を完全に二重化
  - ✓ 設備の二重化、ビル分散、ルート分散
- 障害時にも100%救済
  - 回線容量は、障害時にも100%救済が可能となるよう冗長回線を用意
- 激甚耐性
  - コア拠点間を100km以上離すことで、大規模災害が主要都市で発生しても影響範囲を最小限にくいとめる構成
    - 電気通信網基本計画解説書第1版(昭和63年9月)
      - 「震度Ⅶ級の地震が発生しても，約40kmビル間を離すことで，どちらかのビルは震度Ⅵ以下」
    - 東日本：東京と群馬
      - 仮に東京エリアで大規模災害が発生しても、北海道や東北エリアでOCNに接続されるユーザのインターネット通信が救済
    - 西日本：大阪と愛知

NTTコミュニケーションズのIPバックボーン

2010年11月現在



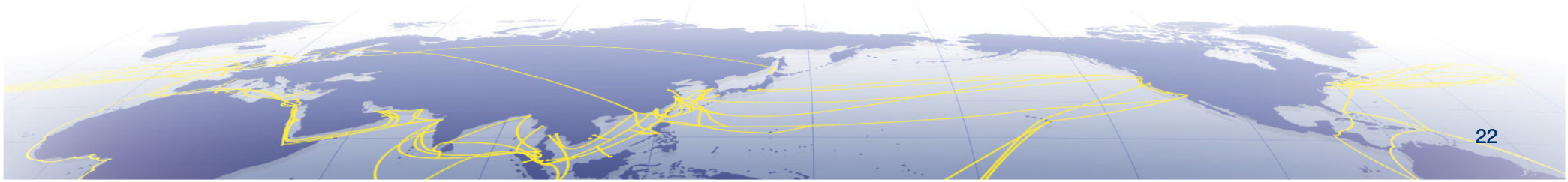
過去のまとめ

1. メディアの変遷
2. トラフィック増
   - 増設、新装置
3. 論理的な課題
   - OSPF経路の増大
4. 信頼性向上
   - ビル分散、ルート分散
   - 激甚災害耐性化

ユーザの増加、トラフィックの増加、
インターネットの重要性増大
に付随した課題に対応してきた

アウトライン

1. 日本のインターネットトラフィックの現状
2. OCNの変遷（サマリ）
3. 過去の設計
4. **現在の設計**
5. 未来の設計、構想
6. まとめ



最近の設計上の課題、苦労

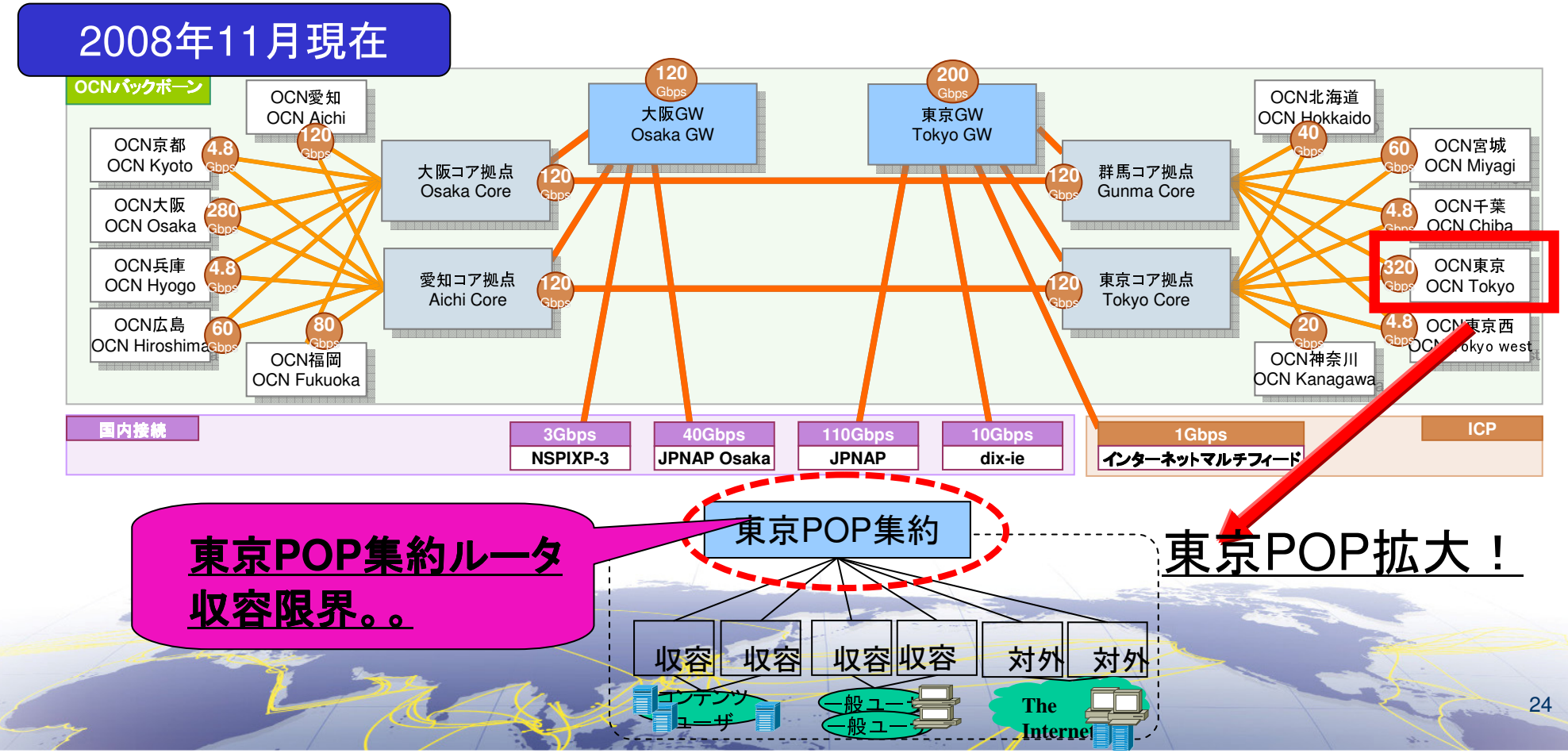
1. 東京POPの肥大化
   - トラフィックの集中

2. フォワーディングスケーラビリティの対策
   - FIBの増大

3. Link Aggregationの限界
   - 限界本数への到達
   - トラフィック偏り
   - 運用上の課題

東京POPのトラフィック増加対策

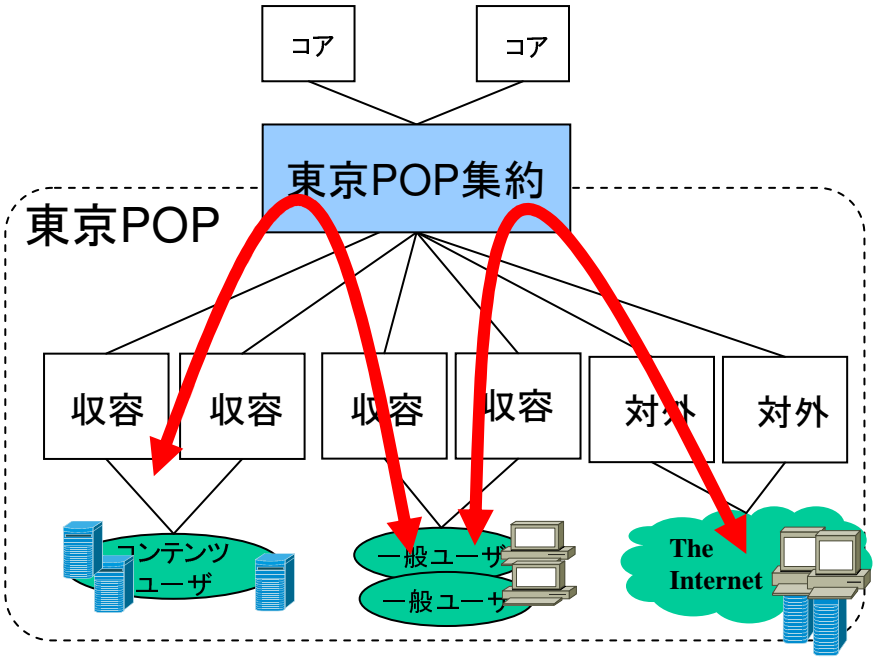
- 2007～2008年ごろから東京POPの肥大化が課題に
- 現状トポロジのまま増強した場合、2009年中に東京POP 集約ルータの収容限界を迎えそう



東京POPのトラフィック増加対策

- 増加の原因
  1. 東京POPにユーザが多い
  2. 対外トラフィックの増加
     → 対外接続拠点は東京に集中

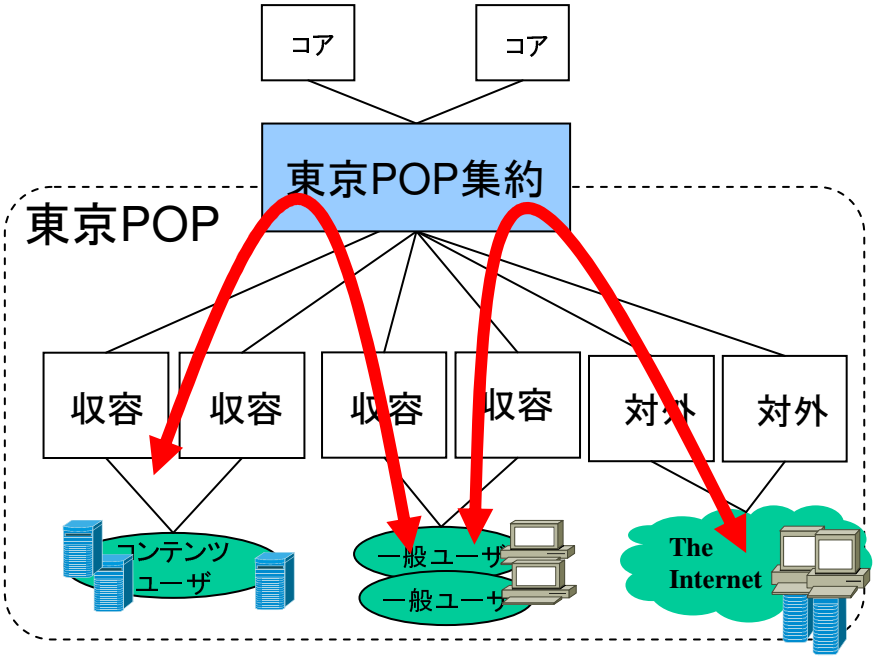
- よくよく分析してみると、、
  - どうやら東京内に吸い込まれるトラフィックが多い
  1. 東京のコンテンツユーザ等が出すトラフィック
  2. 東京の対外接続トラフィック



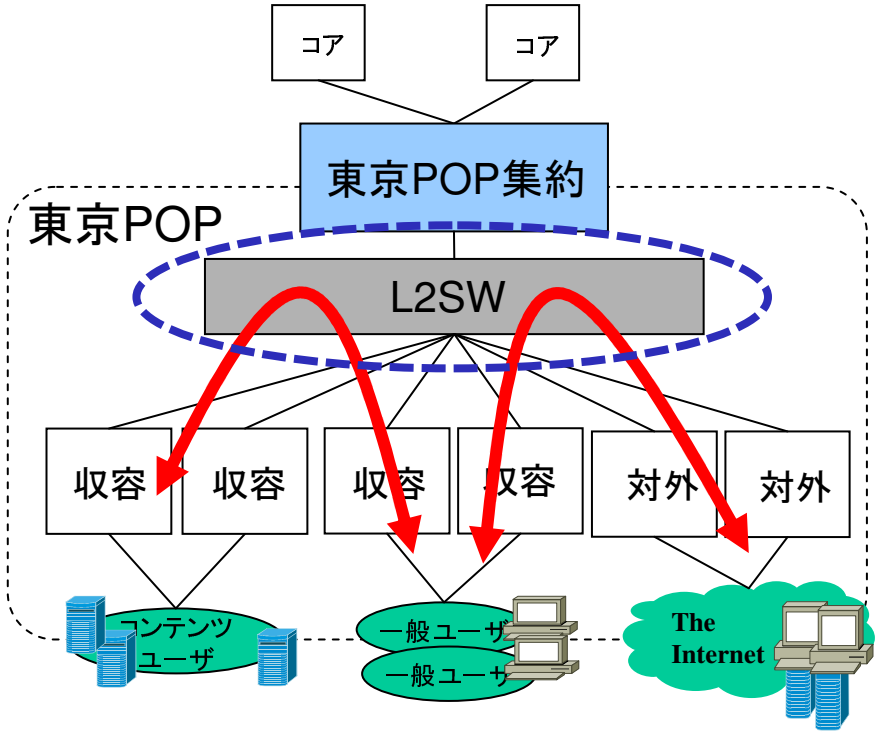
東京POPのトラフィック増加対策

- 別にルータ機能は必要じゃなかった
- トラフィック集約用L2スイッチを導入

L2SW導入前の構成



L2SW導入後の構成



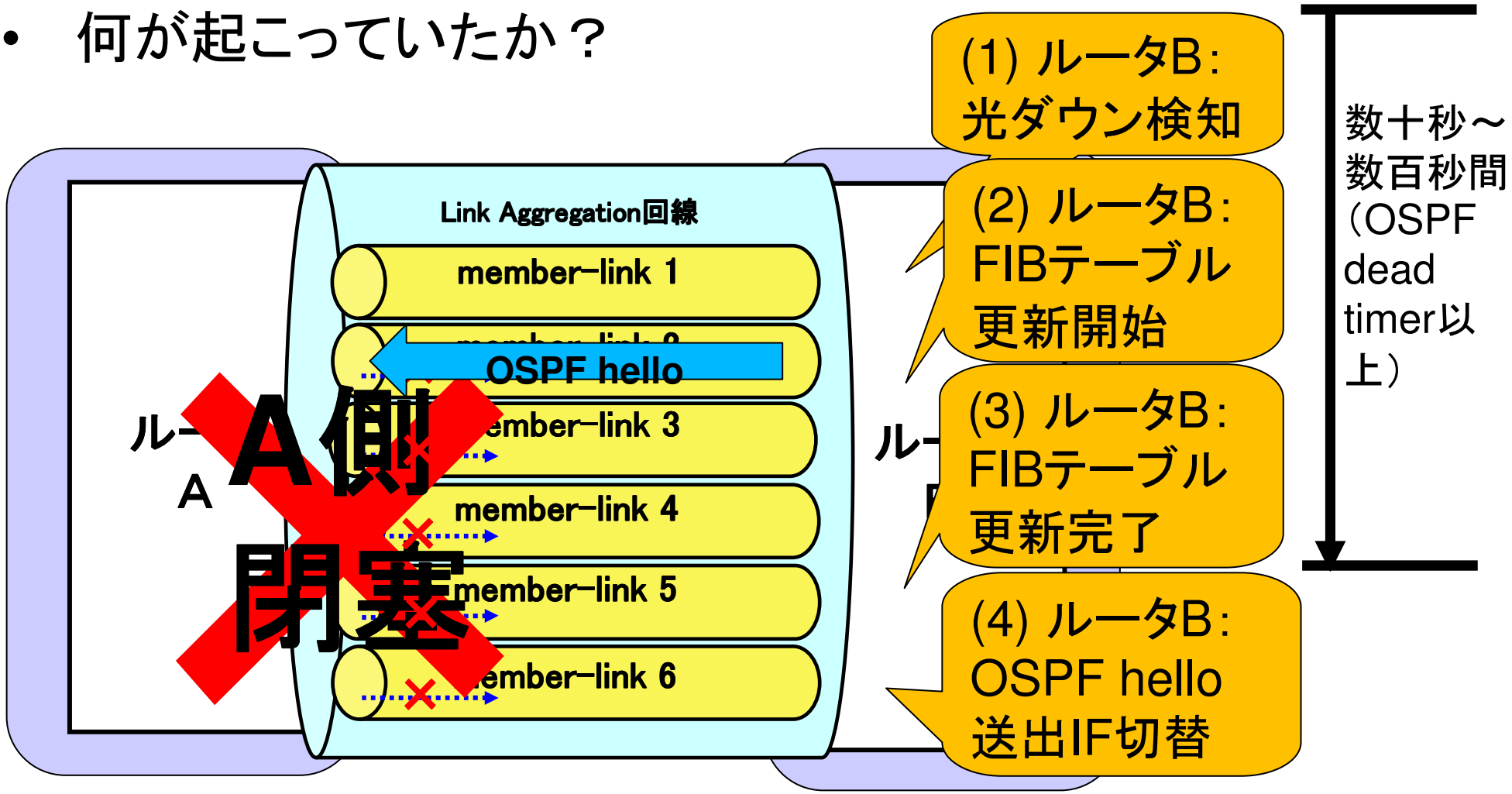
- トラフィック特性をよく見る
- コストとバックボーン拡張のバランスを取る

フォワーディングスケーラビリティの対策

- ルータAとB
  - 複数のmember-linkでLink Aggregationを組んでいる(min-links=1)
  - OSPF neighborな関係
- ルータA側で、メンバリンク1本を残して閉塞
  → 残1本でOSPF Helloをやりとりするはず
  → なぜかOSPFダウン。。

フォワーディングスケーラビリティの対策

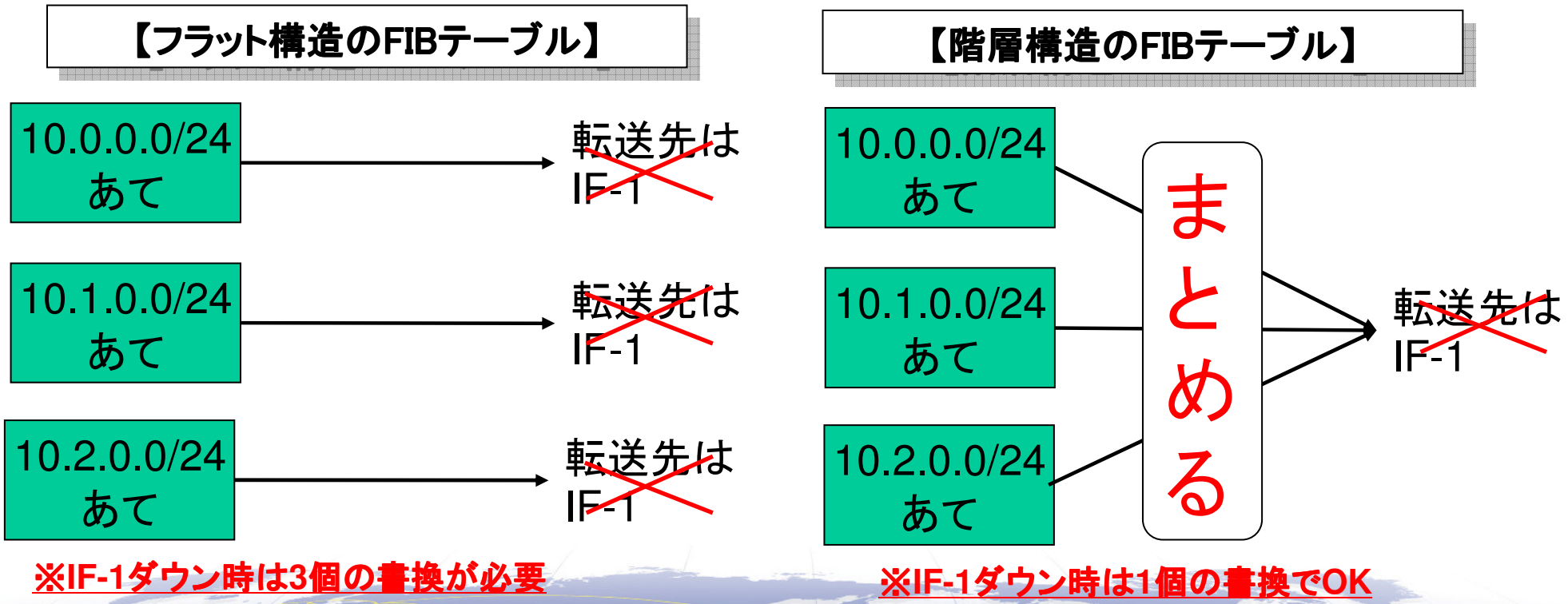
- 何が起こっていたか？



光ダウン検知～FIBテーブルの更新完了までは、OSPF helloを送信できていなかった

フォワーディングスケーラビリティの対策（FIB）

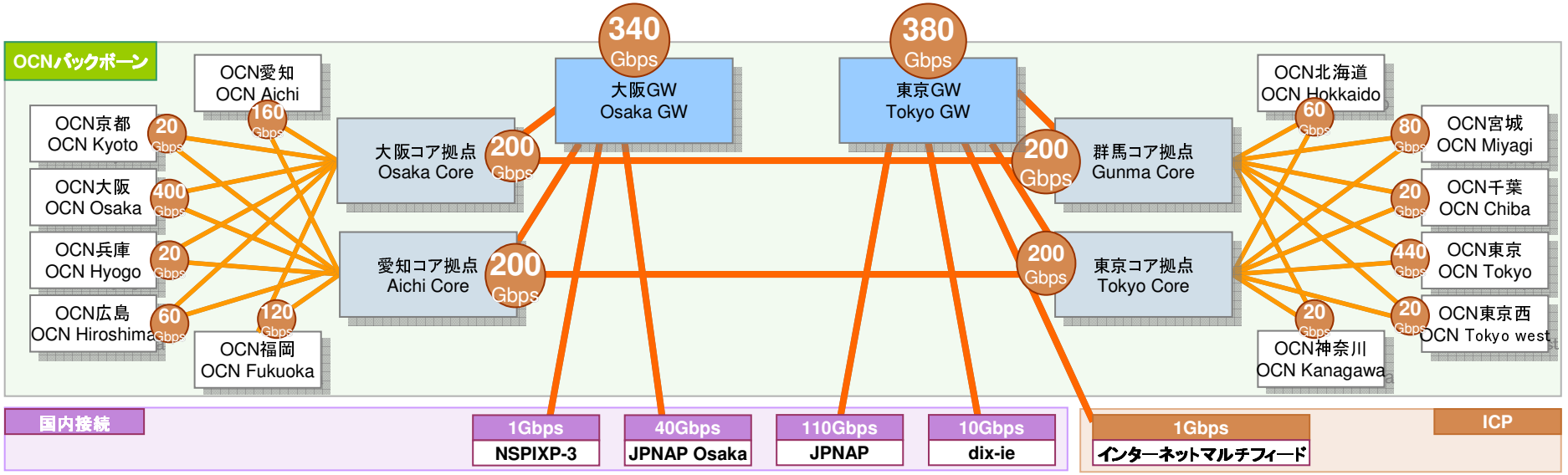
- フラットFIB構造をやめ、階層化FIBを導入
  - Cisco: BGP Prefix Independent Convergence(PIC)
  - Juniper: indirect-nexthop
  - BGP経路との紐付けではなく、ポインタ毎にFIBテーブルを管理



FIBテーブルの集約が可能

Link Aggregationの限界

- 2011年現在、Ling Aggregationを複数回線で利用
- 課題が多い（バンドル上限、トラフィック偏り、運用課題など）

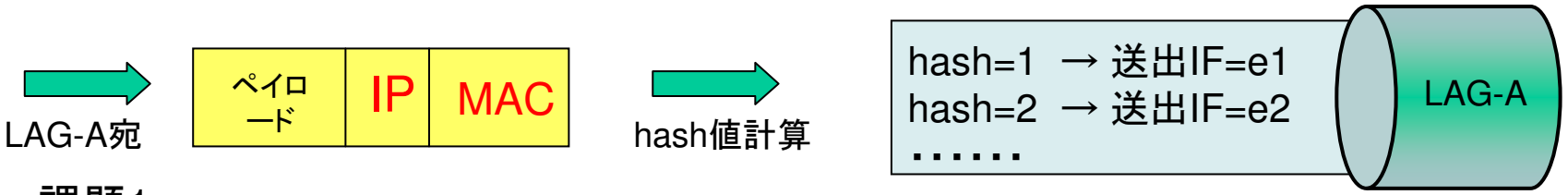


LAGの課題 ～トラヒック偏り1/3～

LAG内におけるトラヒック偏り（1）

前提：単純なラウンド・ロビン方式（per packet）は、パケットの順序逆転が起こるため、採用できない →per flowへ（ラウンド・ロビン以外）

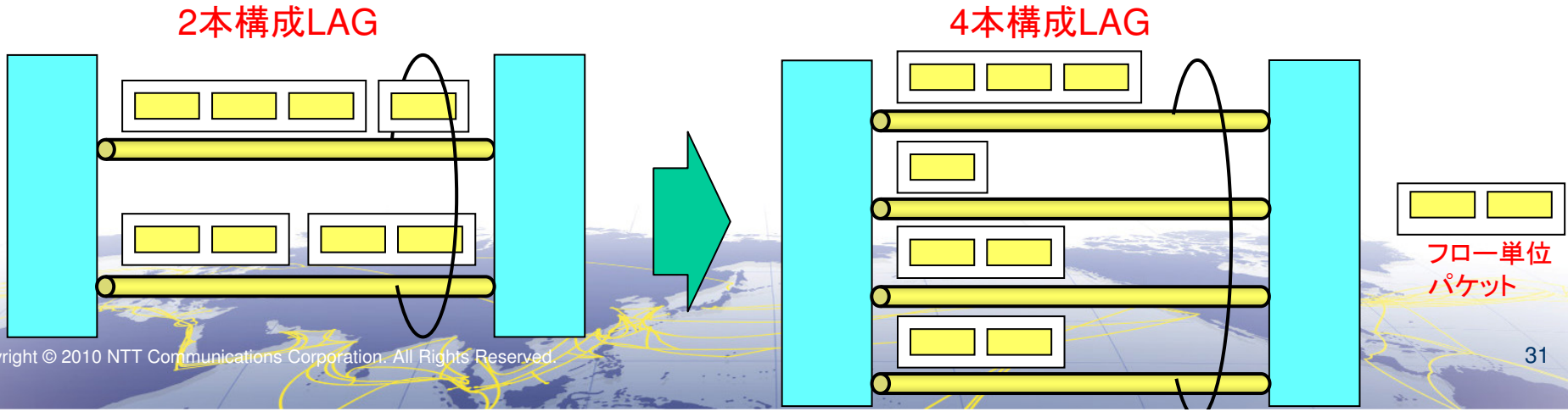
バランス方式：　IPアドレスやMACアドレス等パケットの情報を元に**hash算出**して送出IFを決めるのが一般的



課題1：

hash値毎のトラヒックはフロー単位で分散

物理回線数が多くなると、フロー毎のトラフィック偏りの影響を受けやすくなる



LAGの課題 ～トラヒック偏り2/3～

LAG内におけるトラヒック偏り（2）

課題2：hash値の要素数とトラヒック偏り

hash値の要素数が少ない場合、トラヒック偏りが発生する可能性が高い
結果として帯域を有効活用出来ない

例： hash値が8個の場合のLAG内トラヒック分散

5本LAG	4本LAG	3本LAG
IF#1　H1、H6 IF#2　H2、H7 IF#3　H3、H8 IF#4　H4 IF#5　H5	IF#1　H1、H5 IF#2　H2、H6 IF#3　H3、H7 IF#4　H4、H8	IF#1　H1、H4、H7 IF#2　H2、H5、H8 IF#3　H3、H6
2:2:2:1:1 10+10+10+10*1/2+10*1/2=40	2:2:2:2 10+10+10+10=40	3:3:2 10+10+10*2/3=26.7

hash値の数と物理IFの数によって均等分散できないケースがある

←分散比率
←LAGの実効帯域

5本LAGでも40Gしか流せない

3本LAGでも26Gしか流せない

LAGの課題 ～トラヒック偏り2/3～

参考： hash値の要素数による分散の違い

例1： hash値が8個の場合のLAG内トラヒック分散

5本LAG	4本LAG	3本LAG
IF#1 H1, H6 IF#2 H2, H7 IF#3 H3, H8 IF#4 H4 IF#5 H5	IF#1 H1, H5 IF#2 H2, H6 IF#3 H3, H7 IF#4 H4, H8	IF#1 H1, H4, H7 IF#2 H2, H5, H8 IF#3 H3, H6
40	40	**26.7**

hash値の要素数が多いほどマシになる

例2： hash値が32個の場合のLAG内トラヒック分散

5本LAG	4本LAG	3本LAG
IF#1 H1, H6, ･･･H26, H31 IF#2 H2, H7, ･･･H27, H32 IF#3 H3, H8, ･･･H28 IF#4 H4, H9, ･･･H29 IF#5 H5, H10, ･･･H30	IF#1 H1, H5, ･･･H29 IF#2 H2, H6, ･･･H30 IF#3 H3, H7, ･･･H31 IF#4 H4, H8, ･･･H32	IF#1 H1, H4, ･･･H28, H31 IF#2 H2, H5, ･･･H29, H32 IF#3 H3, H6, ･･･H30
7:7:6:6:6 10+10+10*6/7+10*6/7＋10*6/7=**45.7**	8:8:8:8 10+10+10+10=40	11:11:10 10+10+10*10/11=**29.1**

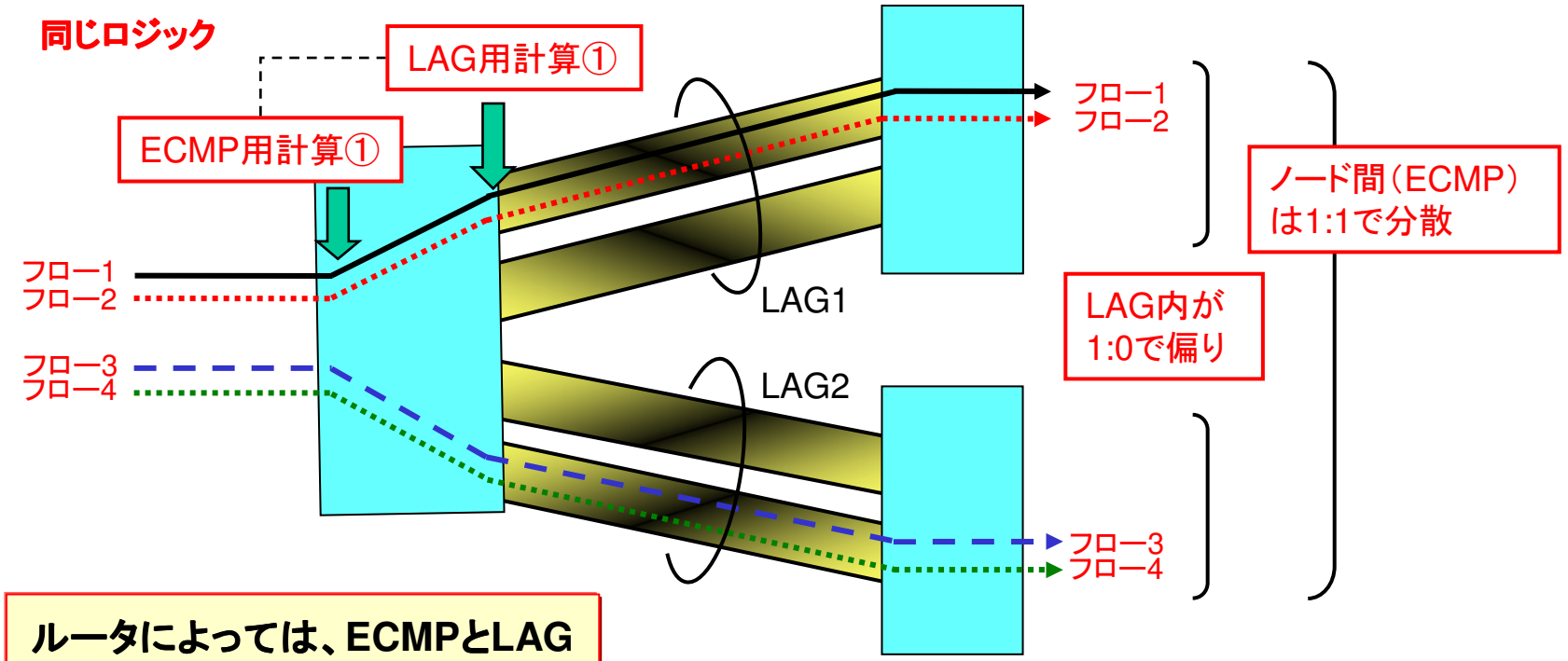
←分散比率
←LAGの実効帯域

33

LAGの課題 ～トラヒック偏り3/3～

課題3：ECMPバランスとLAG内バランス　Case1

トラヒックの分散ロジックの関係で、ECMPバランスの後にLAG内のバランスを行うとLAG内の物理回線で偏りが生じる場合がある

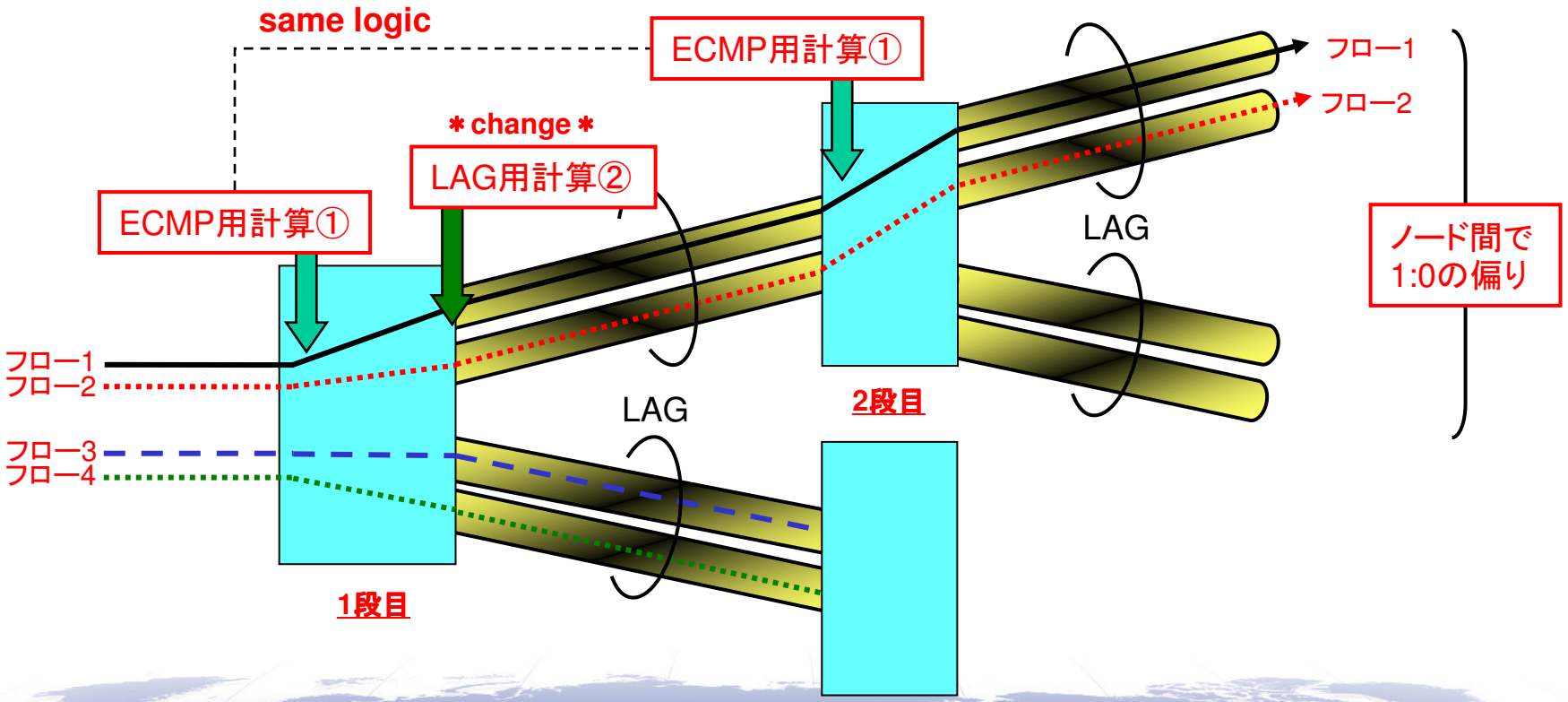


ルータによっては、ECMPとLAG内のバランス計算がデフォルトで同じ場合がある

LAGの課題 ～トラヒック偏り3/3～

課題3：ECMPバランスとLAG内バランス　Case2

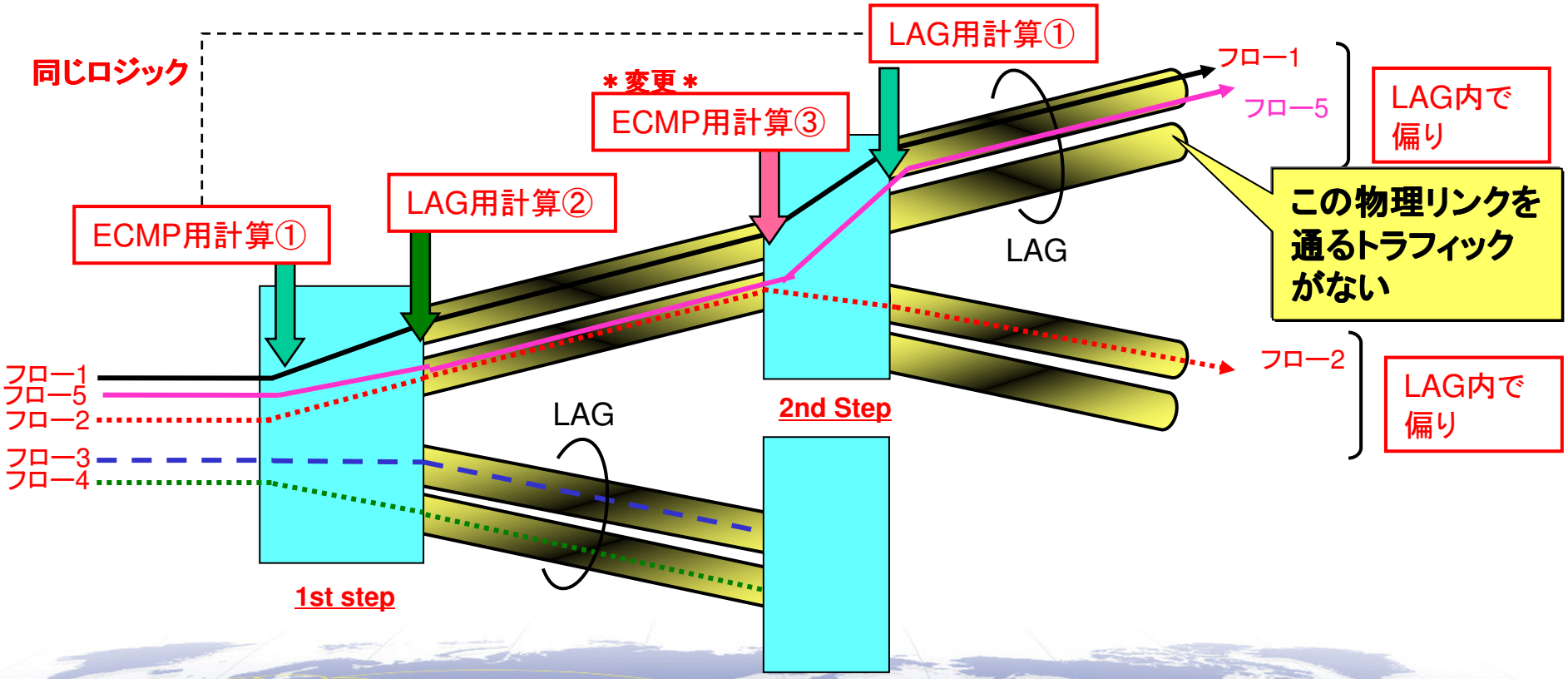
同一機種のルータで構成するNWにおいては、ECMPバランスにおいて同じ計算ロジックを使わないように注意する必要がある



LAGの課題 ～トラヒック偏り3/3～

課題3：ECMPバランスとLAG内バランス　Case3

Case2の対処としてECMPバランスのロジックを変えた場合偏りは改善するが、
IF故障等によってトラヒックが寄る場合、2段目ルータのLAG内で偏りが発生
※最近ではCase2,3を避けるためにrouter-IDを含むことができるものもある



**Case 1からCase3を通して、計算ロジックの重なりや接続構成を踏まえて
トラヒックバランスを考慮した設計・運用が必要**

LAGの課題 ～運用面1/3～

LAGの運用（1）

回線故障時のトラヒックあふれ

中継SW、伝送路等においてIF断とならない故障（サイレント故障）が発生した際に、該当物理回線を流れるトラヒックはロスとなる

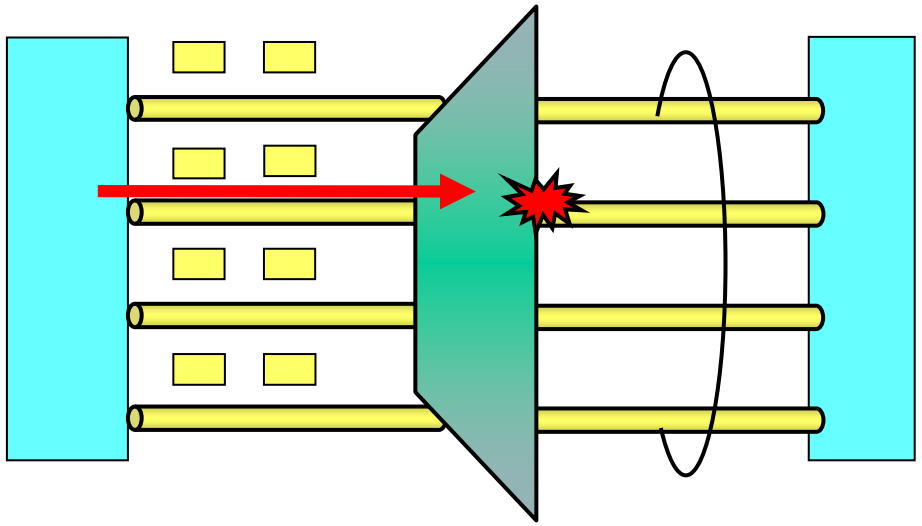
回避策：

LACP（Link Aggregation Control Protocol）

・物理回線間で制御フレームをやり取りする

・ただし、機種によって実装が違う点もあるため、相互接続による運用は注意

BFD Per Member Link



LAGの課題 ～運用面3/3～

LAGの運用(3)

LAG回線の新規開通、工事時の確認、故障切り分け

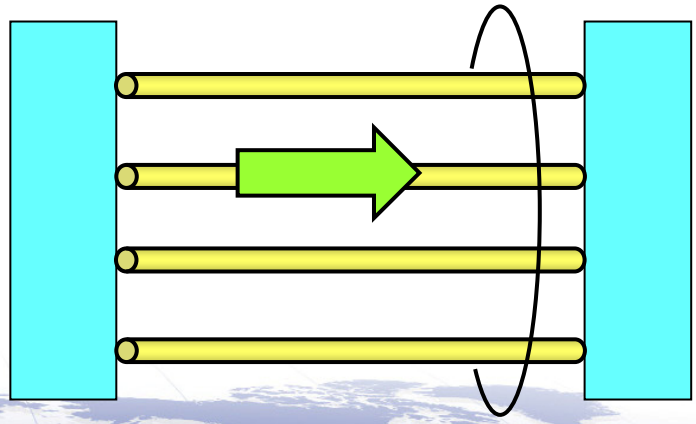
確認用ping:

hashに基づきLAGのうちどこか1ポートを通るため、他のIFをpingで狙い撃ちできない

- LAGの物理回線試験を行う場合には、生存IFを1本ずつにして、各々のping確認する必要がある
- minimum-linkを設定している場合には、回線試験時/戻し時に設定変更が必要

→ 今後、Ethernet OAM等に期待

- ITU-T Y.1731／IEEE 802.1ag
  LAGへのOAMは論理アドレス
- IEEE 802.3ah
  LAGの物理単位にOAM送出可



LAGの課題 ～その他～

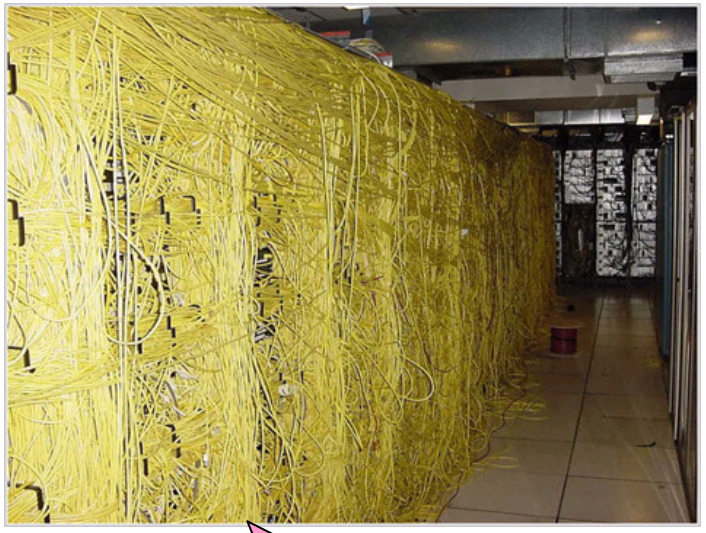
その他

物理配線の問題

ルータ、SWの大容量化が進むにつれ、
物理回線数が増える

- ケーブル配線・交換等の作業が煩雑に
- 設定等大変

異速度LAG、異メディアLAG



参考資料なので、
コム（OCN）設備では
ありません

LAG構築時の考慮

物理回線をラインカード（slot)にどう割り当てるか？

それぞれのMTBF

コスト

例1：出来るだけLAG回線を維持するポリシー

ひとつずつの物理リンクをそれぞれラインカードに割り当て、minimum-link = 1

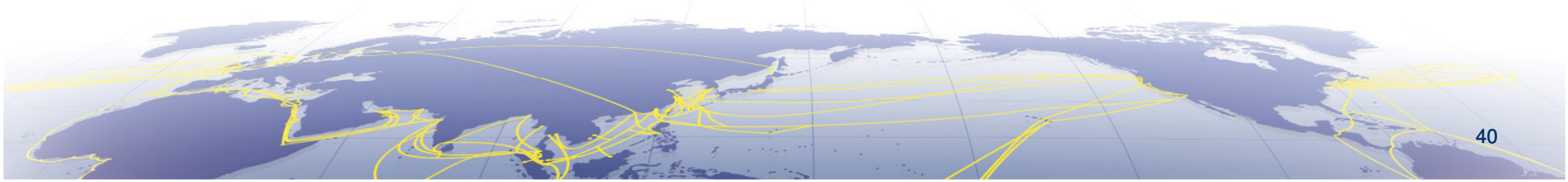
例2：LAGはとっとダウンさせ、他へトラヒックを切り替えるポリシー

全ての物理リンクをひとつのラインカードに割り当て、minimum-link = 物理回線数

多くの回線を束ねることはいろいろ大変
100GEに期待

アウトライン

1. 日本のインターネットトラフィックの現状
2. OCNの変遷（サマリ）
3. 過去の設計
4. 現在の設計
5. **未来の設計、構想**
6. まとめ



トラフィック増加向けた技術的対策

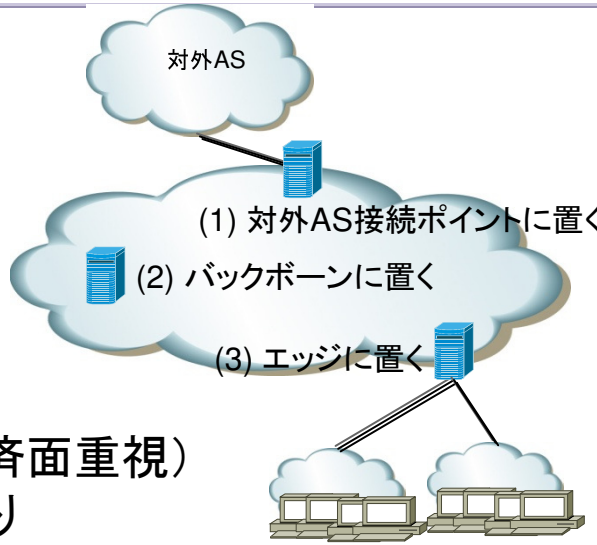
(1)NW側の構成変更	• ルータ、インタフェースの広帯域化（100GE化等） • 新装置導入、L2SW導入等
(2) 流れるトラフィックの効率化	• キャッシュの導入 • CDN

- トラフィック増加にどうせ対応するなら
  - 今よりさらによくしたい（今の課題のカイゼン）
  - トラフィック増加が収入増につながれば

100GEへの要望

- ぜひほしい
  - 1Tbpを超える帯域
  - 10GE、LAG、多数回線の運用限界
- 要望
  - もっと低価格に
    - CFP（光トランシーバモジュール）も高い
    - LR4、SR10だけでなくER4も
    - LR10とかあるみたいだけど
  - ポート密度を高めてほしい
    - 10Gを100Gに変えると、筐体当たりの帯域が減ったりして欲しくない
  - 異速度LAG：100Gと10G
  - 相互接続性、運用性、Ether OAM、100G LAG
    - LAG問題は続く
  - さらなる先へ：400G、1T Ether

コンテンツキャッシュの導入

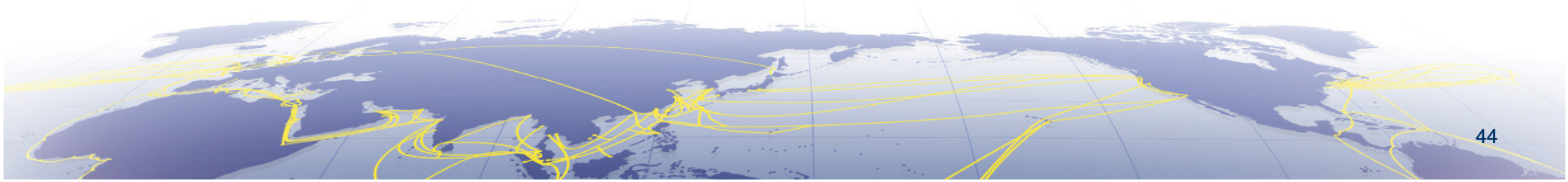


- 2010年1月に法律改正
  - 日本でも設置可能となった
- 必要に応じて動画等のコンテンツをキャッシュにためこんでおく
  - これまでは、Peer/Transit経由で運んでいた
    - → 遠くても安いネットワークで運びがち（速さよりも経済面重視）
    - → 回線帯域、Transit費用、遅延、、、いろいろ課題あり
  - これからは、ISP内にキャッシュサーバを置いてさくっと運ぶ
    - → 帯域節約、Transit費用節約、遅延解消、、、 いろいろいいことありそう
- 実際の設置にあたっては課題もあり

(1) 装置課題	キャッシュヒット率、キャッシュの保持時間（古いキャッシュはいつ捨てる？） 装置の帯域不足
(2) 設置箇所	どこに設置？
(3) 故障時の対応	バックアップ回線を用意？

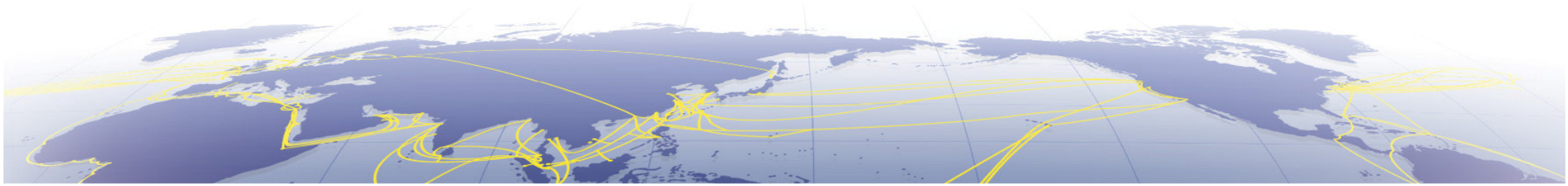
アウトライン

1. 日本のインターネットトラフィックの現状
2. OCNの変遷（サマリ）
3. 過去の設計
4. 現在の設計
5. 未来の設計、構想
6. まとめ



まとめ

- トラフィックは今後も伸び続ける
- 過去様々な設計を行ってきた
  - トラフィック対応
    - メディア更改、ルータ更改、トポロジ見直し、回線増強
  - フォワーディングスケーラビリティ対応
  - etc…
- 我々のミッションはトラフィックを安定して運ぶこと
  - Hyper Giantsの時代
  - トラフィックが増加したって、技術を駆使して前向きに設計！
- 未来の設計は？



OCN
OPEN COMPUTER NETWORK

Thank you!

NTT Communications