Juniper NETWORKS

Internet Week 2007
~東京でディープに語る4日間~

# The Internet Operations
# - 次世代ルーティング -

21 November 2007

Miya Kohno, mkohno@juniper.net



## Agenda

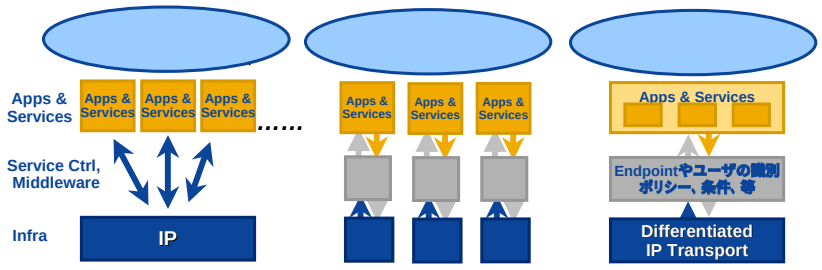
- *はじめに*
  - *メタレヴェル テーマ --- "The Internetはトランジションできるか"*
- 現在のインターネットルーティングが抱える問題
- アーキテクチャ再考?
- ルータ開発の取り組み
- 次世代ルーティング?

## メタレヴェルテーマ
## - インターネットはトランジションできるか

- 類のない、地球規模の分散協調システム
- 何でも取り込む、という包摂力

[ネットワーク構成方法のモデル]




End-to-end
Everything over IP

Optimization
Closed Architecture

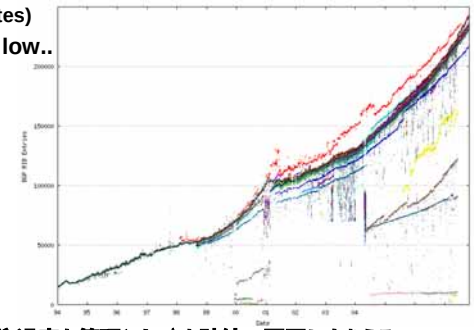
Virtualization
Open Architecture




## メタレヴェルテーマ
## - インターネットはトランジションできるか

- トランジションは必要？
  - 資源問題（環境問題と同じ）
    - 有限なもの
      - IPアドレススペース(4 bytes/16 bytes)
      - AS番号スペース (2 bytes/4bytes)
      - TCP/UDP port番号スペース (2 bytes)
    - Breakthrough, Trade-off, Moore's low..
      - 半導体性能、集積度
      - オプティカル技術、帯域
      - Multi-core、並列処理
      - 消費電力密度
  - 技術＋社会問題
    - セキュリティ脅威、低S/N比
    - ディジタル著作権
    - インフラとしての信頼性

http://bgp.potaroo.net/



→ 思慮なくば破綻する。しかし、思慮あり過ぎ(過度な管理とか?)も破綻の要因になりうる。

## メタレヴェルテーマ
## - インターネットはトランジションできるか

- これまでのトランジション
  - 1980年代初頭
    - /etc/hosts.txt からDNSへの移行
      → host nameの分散管理が実現する
  - 1990年代初頭
    - BGP
      → Policy Routingの実現 (商用用途が出現し、NSFnetでは急速学術用途と商用の Routing Policyを分ける必要が生じた)
    - CIDR
      → Address blockの有効利用
    - OSI (??!!)
      → GOSIP (rfc1169参照) なんてのもあったけれども、これは普及しなかった
- その後?
  - MPLS, ECN(*), diffserv, IP Multicast, Mobile IP…
    → 主に非インターネット用途
    使われたとしてもIntra-domain only

(*) Explicit Congestion Notification




## メタレヴェルテーマ
## - インターネットはトランジションできるか

InternetWeek2007 W4: The Internet Operation (後半) “次世代ルーティング”

湧川 隆次 慶應義塾大学 環境情報学部
Mobile architecture研究、標準化の第一線で活躍中
次世代ルーティングアーキテクチャとして、現固定インターネットのsupersetとなりうる可能性を持つMobile Architecture、および関連activityを紹介する。

河野 美也 Juniper Networks
Routing, Network Architecture担当
現在のインターネットルーティングが抱える問題を概観し、またそれに対する取り組みを考察する。

”Scientific Revolution”/ ”Normal Science”
(「科学革命の構造」 トーマス・クーン 中山 茂 訳 みすず書房)

## Agenda

- はじめに
  - メタレヴェル テーマ --- “The Internetはトランジションできるか”
- *現在のインターネットルーティングが抱える問題*
- *アーキテクチャ再考?*
- ルータ開発の取り組み
- 次世代ルーティング?




## 現在のインターネットルーティングが抱える問題

- 資源問題
  - AS番号スペース
  - IPv4アドレススペース
  - RIB/FIB容量
- Stability vs Convergence
- Security

Juniper

# AS番号

**RFC1930 “Guidelines for creation, selection, and registration of an Autonomous System (AS)” 1996年3月**

9. AS Space exhaustion

The AS number space is a finite amount of address space. It is currently defined as a 16 bit integer and hence limited to 65535 unique AS numbers. At the time of writing some 5,100 ASes have been allocated and a little under 600 ASes are actively routed in the global Internet. It is clear that this growth needs to be continually monitored. However, if the criteria applied above are adhered to, then there is no immediate danger of AS space exhaustion. It is expected that IDRP will be deployed before this becomes an issue. IDRP does not have a fixed limit on the size of an RDI. （下線筆者）

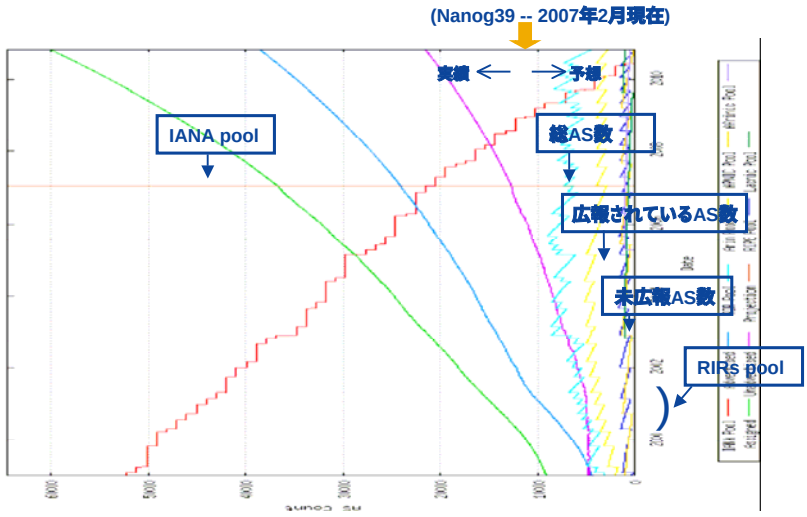
**IDRP ：OSI Inter-Domain Routing Protocol. ISO/IEC 10747(*)にて規定している。IDRPにおけるAS番号に相当するものがRDI (Routing Domain Identifier)であるが、RDIは可変長 (length/valueを記述) である。**

**(*) http://www.sigcomm.org/standards/iso_stds/IDRP/10747.TXT**




Juniper

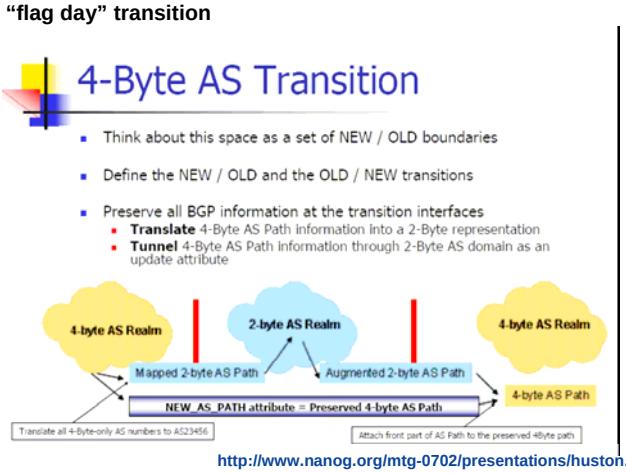
# AS番号消費状況




**http://www.nanog.org/mtg-0702/presentations/huston.pdf**

## AS番号拡張におけるアプローチ

- BGP specへの最小限の侵襲
- Backward Compatibilityの確保
- 混在運用可能 → No “flag day” transition



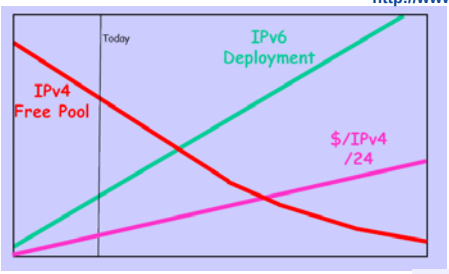

http://www.nanog.org/mtg-0702/presentations/huston.pdf



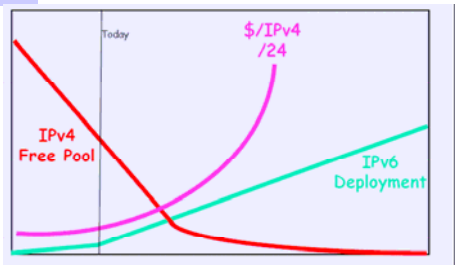

## IPv4アドレススペース

こういう目論見だったのに。

http://www.nanog.org/mtg-0710/presentations/Bush-v6-op-reality.pdf




実際はこんな感じ。



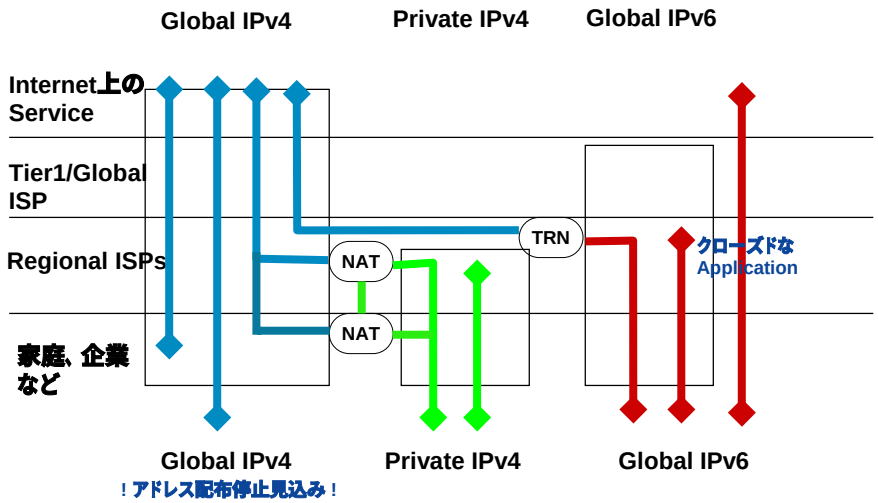
## IPv4 → IPv6 transitionのアプローチ

- Transitionアプローチは大きく分けて以下に分類される
  - Dual Stack
  - Tunneling
  - Translation
- 理想的には、dual stack everywhere !
  - DNSへの依存度増
  - それ以前に、殆どのService/ApplicationそしてHostはIPv4、という現実
- Tunnelingはしかし、v4 islandとv6 islandの疎通を可能にしない
  - 適用領域はあるが限られる
  - アドレススペース問題を解決しない
- Translation (NAT/NAPT(rfc2663), NAT-PT(rfc2766))は問題多い
  - そうはいっても、Translation無しで移行できないのでは




## IPv4 → IPv6 transitionのアプローチ

…かなり苦しい。



## IPv4 → IPv6 transitionのアプローチ

- 外堀を埋める？
- 政治経済的アプローチ?

…

- Transitionよりは、まずはCo-existenceを真剣に考えよう。
  - IPv6の良さを強調し過ぎると、話がそこで終わってしまう可能性がある。
- すべてを一気に移行できない以上、co-existenceなきtransitionはありえない。




## RIB/FIB容量

- IPv6が解決するのは、基本的にはアドレススペースの問題のみ。
- IPv6でもやはりMultihoming, Traffic Engineeringは行いたい。

- ARIN M.Azingerから、v6ops@ops.ietf.orgへの投稿(2006年6月28日)
  draft-ietf-v6ops-routing-guidelines-00はmultihomeに関するguildelineが無い。
  v6でもv4と同様multihomeできるようにすべき。IETFにGuidelineを示すように要請。

(当初の反応)

- More specificジャンクによって、v6アドレススペースの泥沼を作るのはやめよう。
  Let's not create a swamp out of v6 address space with more specific junk. (Pekka Savola)

- RFC 4177 section 5.1　このアプローチ(L3 multihoming)は、マルチホーム方式の全ての目標に合致するが、一つだけ問題がある。ー それはスケーラビリティ。
  "This approach generally meets all the goals for multi-homing approaches with one notable exception: scaleability."

→ draft-baker-v6ops-l3-multihoming-analysis-00

## RIB/FIB容量

**IAB workshop 18-19 Oct.2006**

http://www.iab.org/about/workshops/routingandaddressing/

[問題の定義]

1. Routing Scalability
2. The overloading of IP address semantics
3. Routing Convergence
4. Misaligned Costs and Benefits
5. Others
   1. Mobility
   2. Routing Security

[Workshopからの提言]

1. Scalability of Routing and Addressing System is a concern
2. More discussion is needed with broader audience
3. Solution development should be open and transparent
4. Short/Intermediate term solution is needed concurrently
5. Roadmap to the solution deployment
6. Miscs (to create the mailing list (ram@iag.org) , etc.)

[Report]

- http://www.potaroo.net/ispcol/2006-11/raw.html
- http://www3.ietf.org/proceedings/06nov/slides/RRG-0.pdf
- http://www3.ietf.org/proceedings/06nov/slides/plenaryt-5.pdf

→ rfc4984




## RIB/FIB容量

- **NANOG 39 BOF 5 Feb.2007**

  http://www.nanog.org/mtg-0702/jaeggli.html/

  Joel Jeaggli氏の呼びかけで、各ヴェンダーがFIB memoryアーキテクチャや取り組みについて報告

  (Force 10, Cisco, Foundry, Juniper, Extreme)

- **Apricot “future of routing” workshop 26-27 Feb.2007**

  http://www.apricot2007.net/presentation/apia-future-routing/

  Dave MeyerがChair

  Jari Arkko, John Scudder, Vince FullerがRouting Scalabilityを考察

## アーキテクチャ再考？

"Addresses can follow topology,
or topology can follow addresses,
but you can only pick one. "

--- Yakov Rekhter




## アーキテクチャ再考？

IP addressには、次の2つの意味が内包されている。

- ネットワークトポロジーにおける位置を特定する情報 (Locator)
- エンドシステムを識別する情報 (End Identifier)

→分離する？

[分離のメリット]

- DFZ(Default Free Zone)にはLocatorのみ存在すればよい。
- Locatorは集約可能。
- MultihomeやMobilityの実現可能性。

→どうやって？

(Host based)

- HIP, etc.

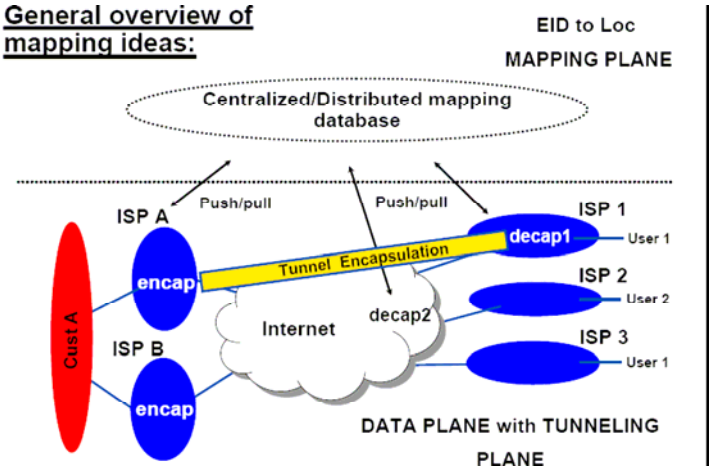
(Network based)

- Rewriting/Translating (8+8/GSE, etc.)
- Map and Encap/Tunneling (LISP, etc.)

## LocatorとIDの分離 (Network based Tunneling)

- コアネットワークでは、”Locator”情報のみを運ぶ
- Edge device間はTunnelされる。

General overview of mapping ideas:




http://www.janog.gr.jp/meeting/janog20/pg-routing.html

## LocatorとIDの分離 (Network based Tunneling)

**Many Open Questions:**

- 次のものが必要
  - “ID”情報と”Locator”をbindするmapping service
  - Mapping情報を伝達する仕組み
    - BGP --- “Push” protocol
    - Caching --- “Pull” model

  → 経路数を削減するのでなく、別の場所に移している。
- 障害時や、Mappingが変わったときのConvergence時間
  → ICMP等に頼る? パケットを送ってみないとわからない?
- Tunnelingによるoverhead, performance, bottleneck
- Traffic Engineering ?
- Deployability, 移行可能性
- コストを凌駕する効用があるのか?
- BGP free core (e.g. by MPLS) は現時点でも可能であるが、deployment事例は少ない。

## アーキテクチャ再考？

"Any problem in computer science can be solved with another layer of indirection."

—David Wheeler

"But that usually will create another problem."

—rest of the quote

- "A problem": RIB/FIB容量
- "Another problem" : 新たなMapping/Tunnelingによるオーヴァヘッド

なお、Mapping/Tunnelingは比較的静的なoverlayであり、実は"Loc/ID分離"とはいえないのでは。

いずれにせよ、キーは、
- 効用がコスト(移行コスト、運用コスト)を凌駕するか。
- 移行できるか。




## RIB/FIB容量

- RIB/FIB容量自体は、現在の商用ルータにおいては、さほど問題ではない。少なくとも、この先10年程度のHW feasibilityは見えている。
- メモリ容量と、転送効率・Convergence・Stabilityをいかに両立させるかが問題。
  (複数のトレードオフ関係)
  - 転送効率 <-> Convergence
  - Convergence <-> Stability
  - Protection <-> メモリ容量

  …

## Convergence vs Stability

- ConvergenceとStabilityは拮抗する要素
- しかし、それ以前に重要なのはConnectivity
  (happy packets © Randy Bush)

Connectivity > Stability

多少網が不安定になっても、トラフィックが流れていればよい。

Connectivity > Convergence

コンヴァージェンスはしていなくても、トラフィックが流れていればよい。

- 従って、重要なことは：
  1. トラフィック断無し、または最低限に留める。
  2. 過負荷による不安定を避ける。




## Convergence vs Stability

- 高速コンヴァージェンスは通常、高速に検知し、高速にreactする必要があるため、Stabilityとは拮抗する。

→ 共存させるために

- Carrier delay / Interface Dampening
  for DOWN event, UP event
- Exponential Back-off
  for IGP spf-delay, LSA generation
- BGP RFD (*), MRAI
- Protection

…

(*) しかし、arbitraryなdampeningは却って害がある。
Is RFD harmful ?!
http://www.ripe.net/docs/routeflap-damping.html

## RFD見直し

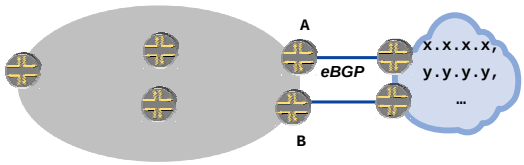
draft-li-bgp-stability-01
http://www.potaroo.net/ispcol/2007-06/dampbgp.html

[Goal]
- Flap damping
- Rapid Convergence
- Overhead削減
  - Path hunting
  - 屈折(Refraction)を避ける

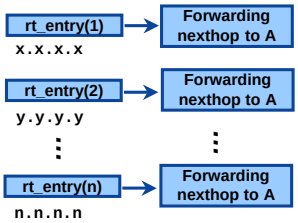
[仮説]
- 止める
- Band-stop filtering
- Path length damping
- 最適パスヒステリシス
- path selectionを遅らせる
- MRAIの廃止
- これらの組み合わせ
- その他
  - Aggregate withdraw




## FIBの階層性

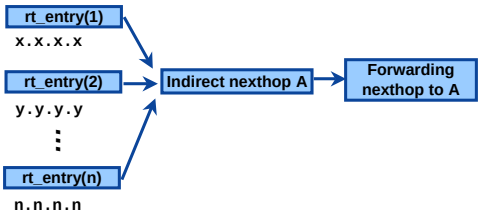



1) Flattened FIB




→ フォワーディング効率に有利

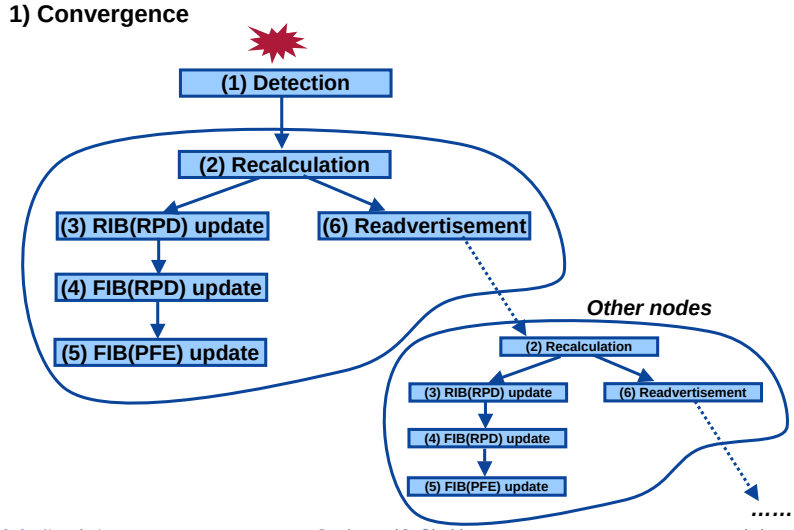
2) Hierarchical FIB




→ コンヴァージェンスに有利
→ Protectionのための土台となる

## Convergence vs Protection

1) Convergence

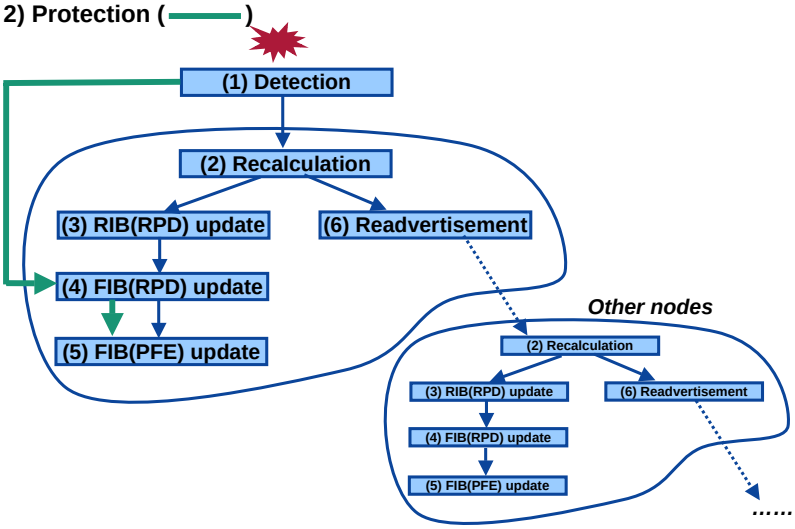






## Convergence vs Protection

2) Protection ( ——— )

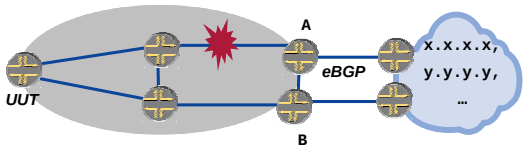


## Convergence vs Protection

| Convergence | Protection |
|---|---|
| Globally consistent | Local |
| Stabilityとのtrade-off可能性 | FIB sizeとのtrade-off可能性 |
| 恒久的 | 一時的 |
| トポロジーは自由 | トポロジーに依存する |
| 技術例:<br>- Convergence (IGP/BGP) | 技術例:<br>- FRR (MPLS, IP)<br>- ECMP |



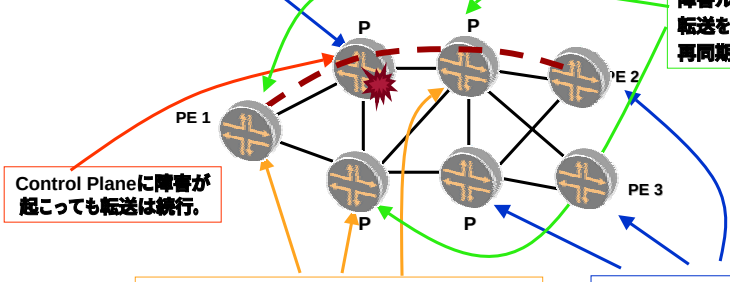

## Convergence vs Protection




- まずはProtectionで救い、後からゆっくりConvergenceさせる。
  - トラフィック断時間は最少
  - Stabilityも維持可能
- ただし、条件と場合によってはループとなる可能性があるので要注意。

## Control Plane障害であれば…
## Graceful Restart / NSR

Control PlaneとData Planeは独立
- RE (Routing Engine)
- FE (Forwarding Engine)




Graceful Restartの場合、障害ルータが復帰すると転送を続行しながら、状態の再同期を取る。

Control Planeに障害が起こっても転送は続行。

Graceful Restartの場合、Neighborはhelperとして動作する。 但し、他のノードに障害が起こったことを伝播しない。

NSRの場合、Neighborも障害を関知しない。

他のノードは、障害を関知しない。

→ 障害は局所に留まる

→ ネットワークは安定

ただし、NSRはoverheadに注意。




## Agenda

- はじめに
  - メタレヴェル テーマ --- “The Internetはトランジションできるか”
- 現在のインターネットルーティングが抱える問題
- アーキテクチャ再考?
- ***ルータ開発の取り組み***
- 次世代ルーティング?

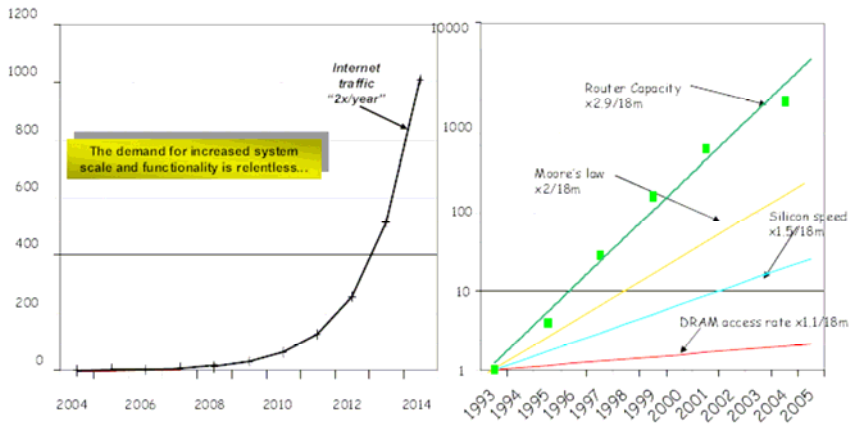
## ルータ開発の取り組み

- Do more with less
- Virtualization
- Automation

…

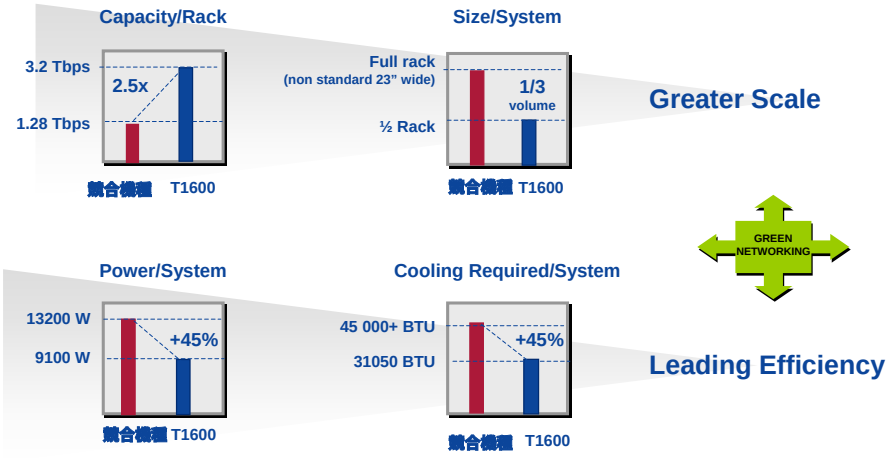



## ルータキャパシティの成長



# Do more with less

**最近のachievement**







# Do more with less

**Micro Processorの性能向上アプローチ**

- **マルチコア化**

  - **シングルコアに対して20% のオーバー・クロック (動作周波数を上げる) を行った場合、性能は13% 向上するが、消費電力は73 % も上昇。**

  - **20%のアンダー・クロック（動作周波数を下げる）では、性能が13%下がるものの、それ以上に消費電力が49%低減。**

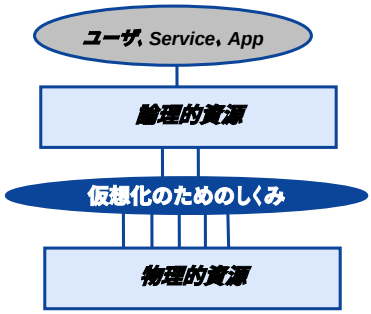
  **→20%のアンダー・クロックを行ったコアを 2 つ使うことで、もともとのシングルコアと同水準の消費電力でありながら、性能は73%向上。**

**http://www.intel.co.jp/jp/business/technologies/focused-tech/process-rule/index02.htm**

# Virtualization

**仮想化とは**

- 物理的資源は、仮想化のためのしくみにより、論理的資源に、
（抽象化｜隠蔽｜分割）される。
- ユーザや、サーヴィス・アプリケーションは、論理的資源を利用する。




→ 分割された資源を統合的に使用する。

→ 冗長/Fault Tolerant性を実現する手法、
という側面も持つ。




# Virtualization

**仮想化の方法**

**1) Aggregation**

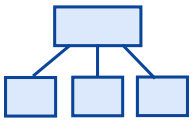
- Link bundling
- RE redundancy
- 階層化
  CsC, RR, LSP hierarchy…
- …

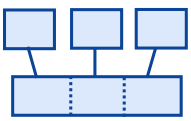
**2) Partitioning**

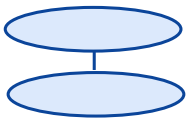
- Logical Router
- vlan
- Modularity
- QoS
- Multi-topology
- Confederation
- …

**3) Emulation**

- VPN
- VPLS
- Pseudo Wire
- …







## Virtualizationにおける課題

「統合」と「分離」という、内在する二律背反をどうするか

Control Plane
- 共通コントロールプレーン
- fate-sharing、リソース競合を避ける

Forwarding Plane
- 共通インフラ、統合トランスポート
- QoS分離、トポロジー分離

Management Plane
- 統合マネジメントシステム、一元管理、集中管理
- 組織主体、サーヴィスによる分離(provisioning, syslog, statistics..)

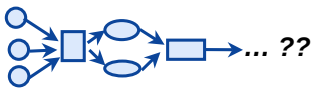
その他
- セキュリティ、責任分解点のためのBoundary
- リソース保護
- Disaster Recovery, Risk Management

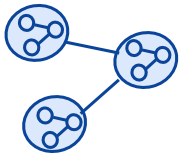



## Virtualization ー 設計指針

複雑性、オーヴァヘッドを最低限に留める

- 統合・分離を、必要以上には繰り返さない
  - Link Bundling (統合) + QoS (分離) ?!!
  - Logical Router (分離) + RE redundancy (統合)
- Self-similarityが保てればOK (の場合が多い)
  - Route Reflector (統合) + Confederation (分離)



拮抗する要素を調整する

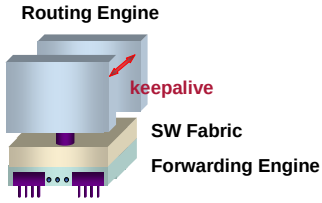
- 統合 vs 分離
- 安定性 vs 高速性 (高速検出・高速処理)
  - GR/NSR vs Fast Convergence, Route Flap Dampening
- 拡張性 vs きめ細かさ
  - 制御単位の粒度
- 運用性 (シンプルさ) vs 多機能

要素技術を充分吟味し選択する

## コントロールプレーン冗長

コントロールプレーン冗長の必要性

- コントロールプレーン断は起こりえる
  - 計画停止
    ソフトウェアのアップグレード、保守
  - 無計画停止
    異常事象、バグ




- 起こった場合の影響が予想以上に大きくなる可能性がある
  - ルーティング収束
  - Oscillation
  - Cascade Failure



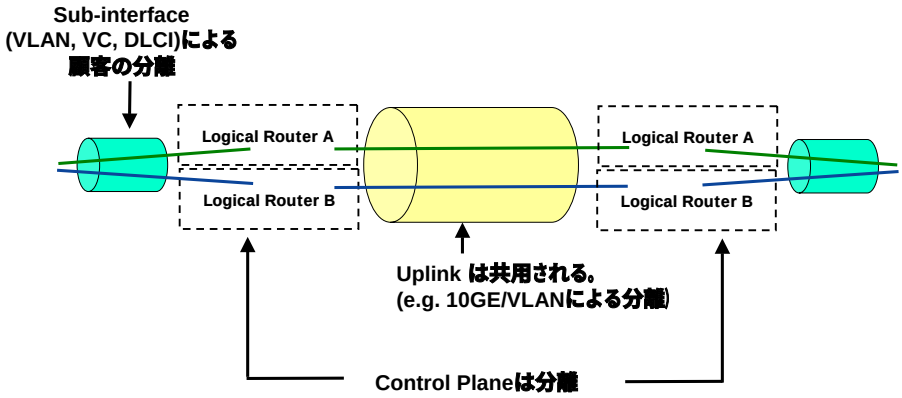

## コントロールプレーン分離

Logical Router (LR)によるコントロールプレーン分離

- 分離
  - Management Plane
  - Control Plane
  - 各LRで動作させるSoftware Instance
- 共用
  - Master Management Plane
  - ハードウェア資源 (Software Logical Routerの場合)
    - Uplink回線
    - CPU, memory
      … with some partitions

→ *Service Separation*
→ *セキュリティ境界*
→ *fate sharingの回避*

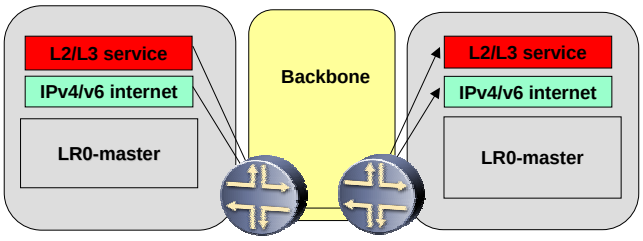
## Logical RouterによるUplinkの共用



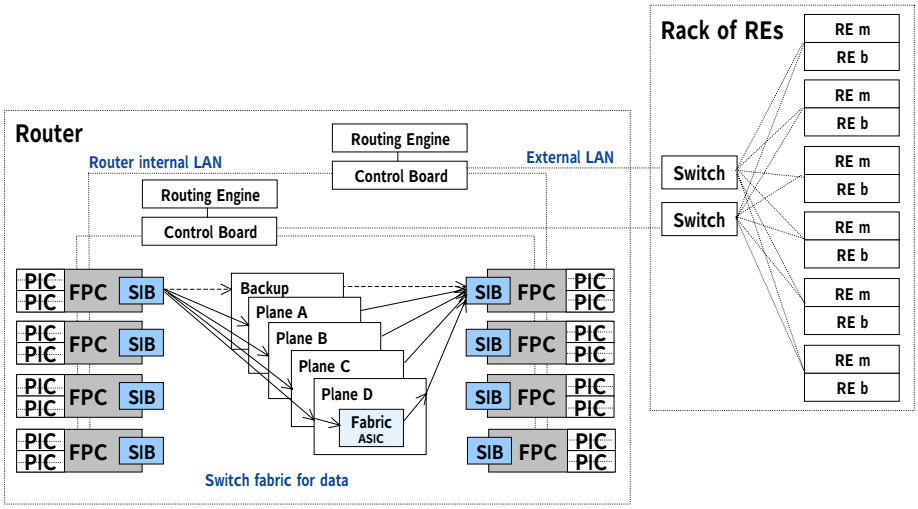




## Logical Routerによるサーヴィス分離例



## More Virtualization







## Agenda

- はじめに
  - メタレヴェル テーマ --- “The Internetはトランジションできるか”
- 現在のインターネットルーティングが抱える問題
- アーキテクチャ再考?
- ルータ開発の取り組み
- *次世代ルーティング?*

Juniper

# 次世代ルーティング？

- 世代交代は、意図した通りには起こらない。
- Transitionは、次の条件が揃ったときに起こりうる。
  - 危機的状況が逼迫する
  - Transitionの効用がコストを上回ることが見込める
  - 段階的移行が可能 (incrementally deployable)
- 大切なのは
  - 現状理解
  - 日々の改善、向上
  - 全ての関係者の高い意識と将来を洞察する力