S8 ルーティング関連セッション（Ⅱ）
～ 経路爆発を考える ～ rev1.2

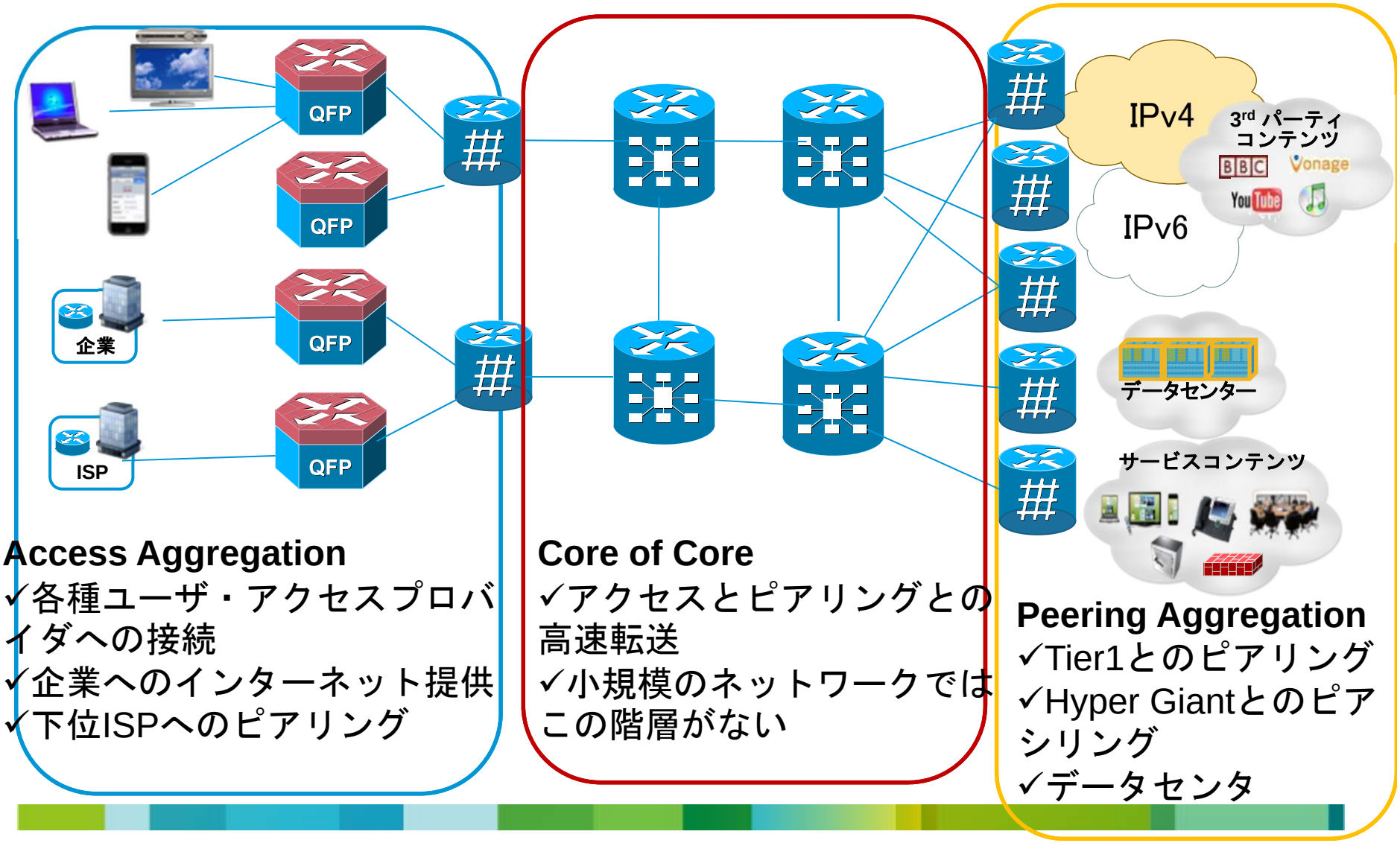
Shishio Tsuchiya

shtsuchi@cisco.com

Part 2. – 経路爆発問題に対する対処は？！

- 通信事業者自身による対応
- 通信機器による対応
- 新しいプロトコルの導入

インターネットサービスプロバイダのネットワーク



Access Aggregation
✓各種ユーザ・アクセスプロバイダへの接続
✓企業へのインターネット提供
✓下位ISPへのピアリング

Core of Core
✓アクセスとピアリングとの高速転送
✓小規模のネットワークではこの階層がない

Peering Aggregation
✓Tier1とのピアリング
✓Hyper Giantとのピアシリング
✓データセンタ

各レイヤーの要件

	Access	Core of Core	Peering
インターフェース種別	様々	100GE/40GE/10GE	10GE/1GE
BGPルート数(広告数)	フルルート	無し	顧客分 サービス分
BGPルート数(受信数)	フルルート	フルルート	フルルート
FIB	大きい	大きい(transitの為)	大きい
デュアルスタック	必要	必要(transitの為)	必要
コスト	$	$$$	$$

アクセスでの要件

	Access	Core of Core	Peering
インターフェース種別	様々	100GE/40GE/10GE	10GE/1GE
BGPルート数(広告数)	フルルート	無し	顧客分 サービス分
BGPルート数(受信数)	フルルート	フルルート	フルルート
FIB	大きい	大きい(transitの為)	大きい
デュアルスタック	必要	必要(transitの為)	必要
コスト	$	$$$	$$

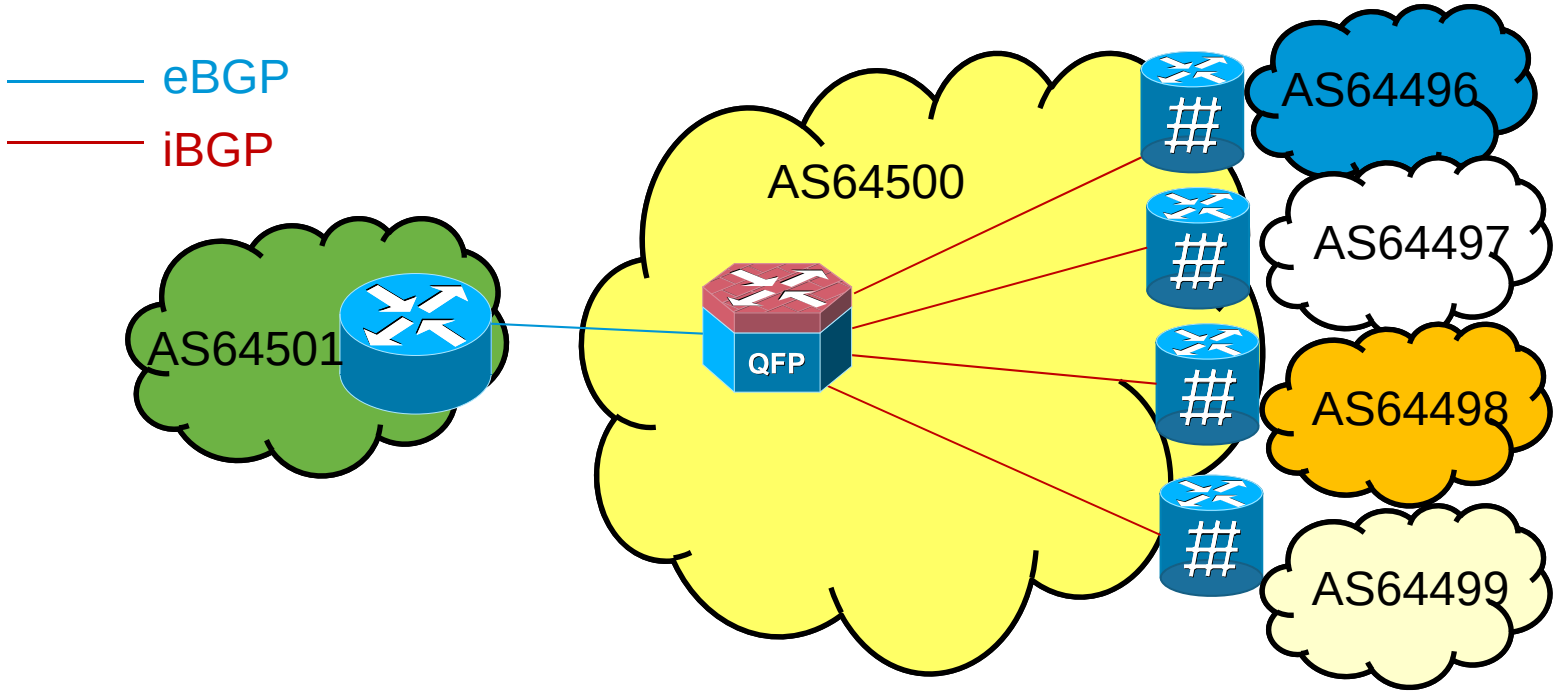
- 下位プロバイダーへのフルルート提供の為に、すべてのルータでフルルートを持つ必要がある。

BGP環境でのFIBの作成の方法

FIB / RIB Table	Data
Active BGP entries (FIB)	396,184
All BGP entries (RIB)	12,561,626
RIB/FIB ratio	31.7065

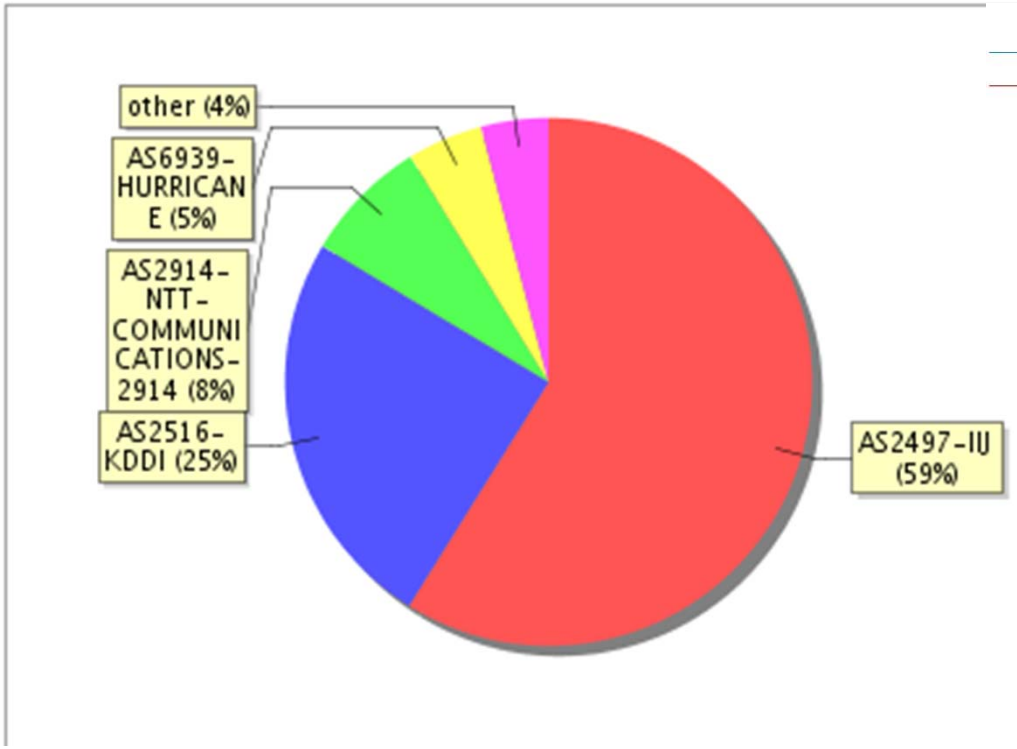
- BGPテーブルからルーティングテーブルを作成
- ルーティングテーブルよりFIBが出来る
- FIBの情報を元にパケットをフォワーディングする
- BGP経路の爆発により、すべてのリソースを消費する

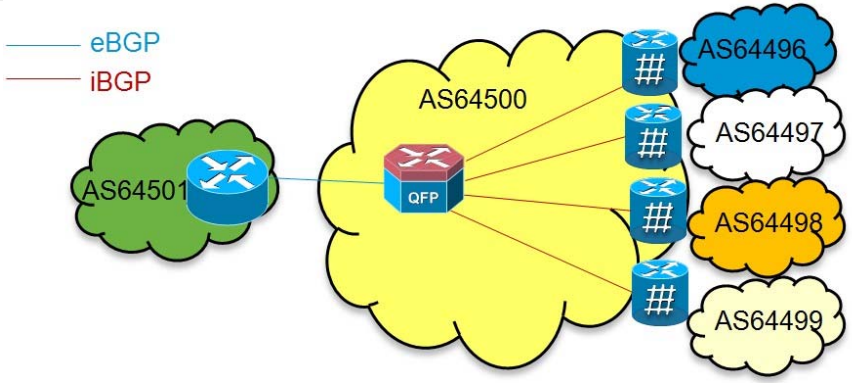
フルメッシュルーティング環境



- Small/Mediumサイズのネットワークではrrは用いず、BGPフルメッシュで構築する。
- 最近では大規模であっても、フルメッシュは多い

フルルートが必要か？





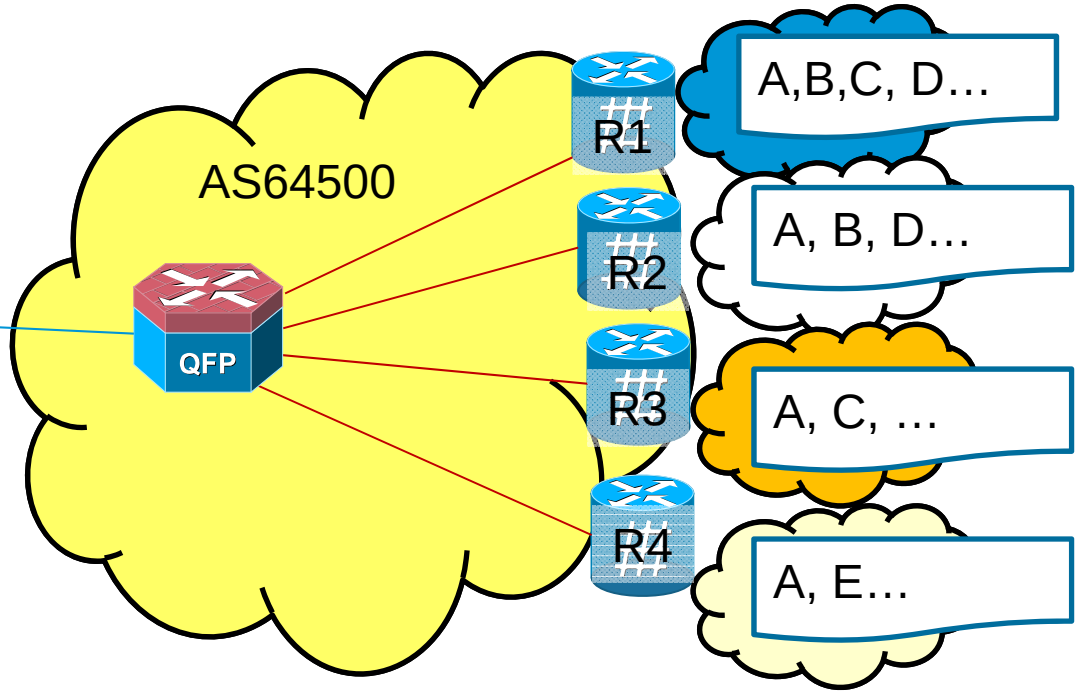
- 下位プロバイダーへのフルルート提供の為には必要
- RIPE AS dashboardのtransits distributionではパス情報から、偏りを見る事が出来る。

Simple Virtual Aggregation(S-VA)

draft-ietf-grow-simple-va

cont'd

BGPテーブル	
Destination	Nexthop
0.0.0.0	*R1
A	*R1 R2 R3 R4
B	*R1 R2
C	R1 *R3
D	*R1 R2
E	*R4



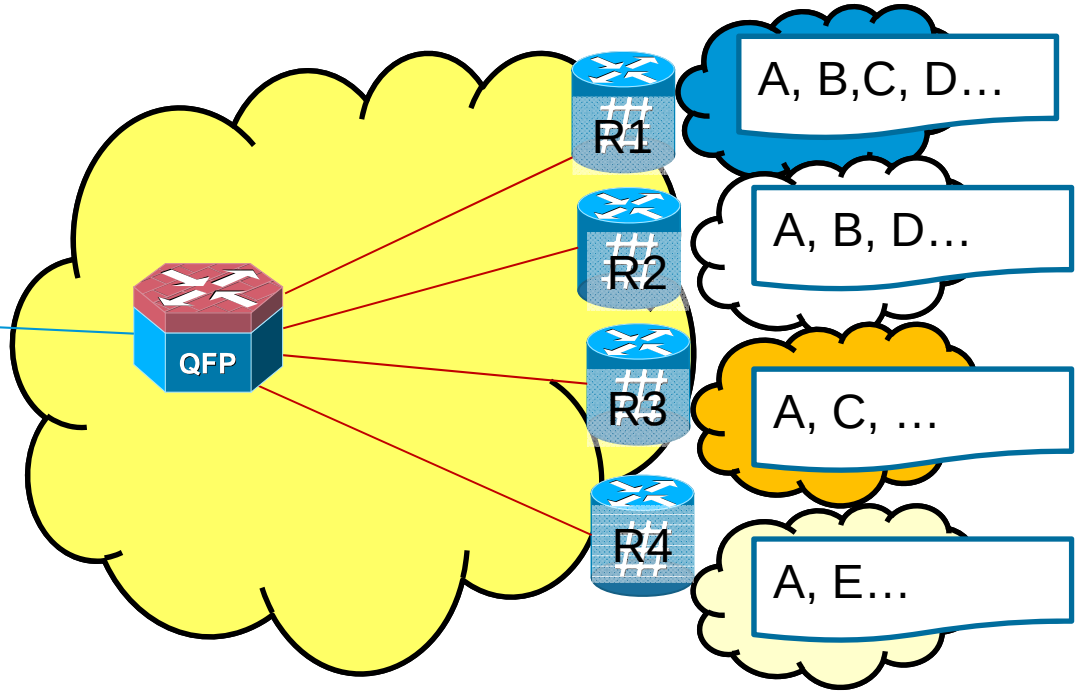
- 通常のBGPテーブル
- 複数のパスを持つ場合はベストパスを選出する

Simple Virtual Aggregation(S-VA)

draft-ietf-grow-simple-va

cont'd

ルーティングテーブル	
Destination	Nexthop
0.0.0.0	*R1
A	*R1
B	*R1
C	*R3
D	*R1
E	*R4



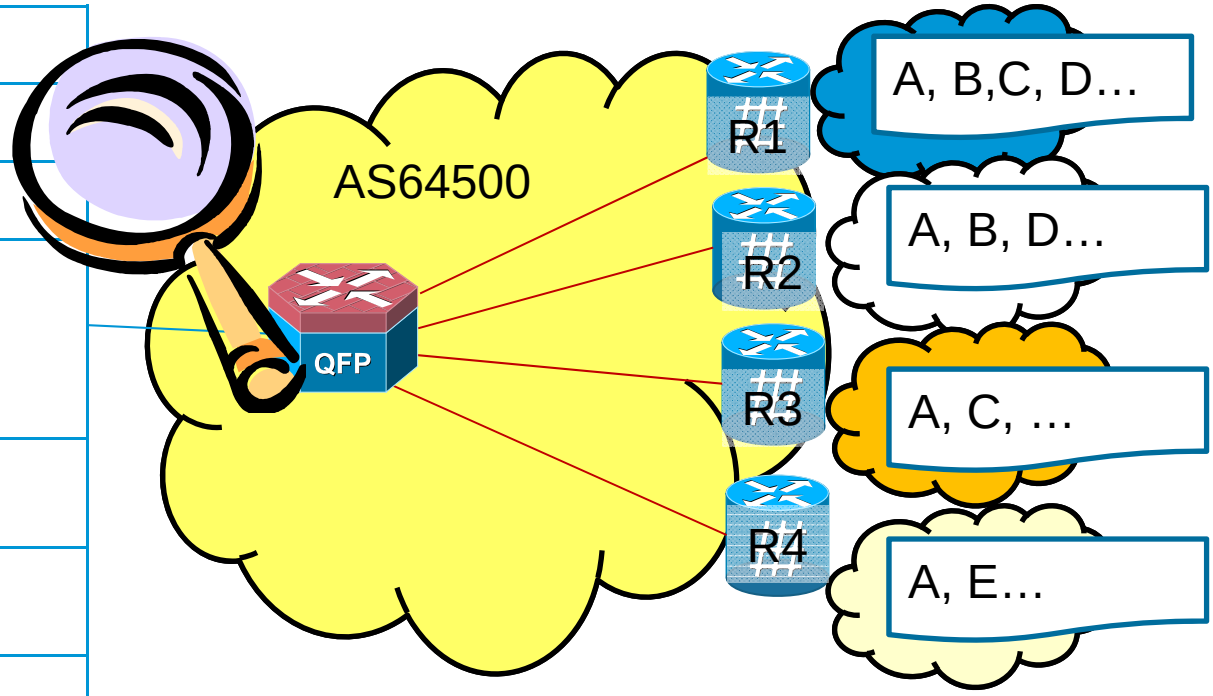
- 通常のルーティングテーブル
- 宛先毎のnexthopを持つ

Simple Virtual Aggregation(S-VA)

draft-ietf-grow-simple-va

cont'd

BGPテーブル	
Destination	Nexthop
0.0.0.0	*R1
A	***R1** R2 R3 R4
B	***R1** R2
C	R1 *R3
D	***R1** R2
E	*R4



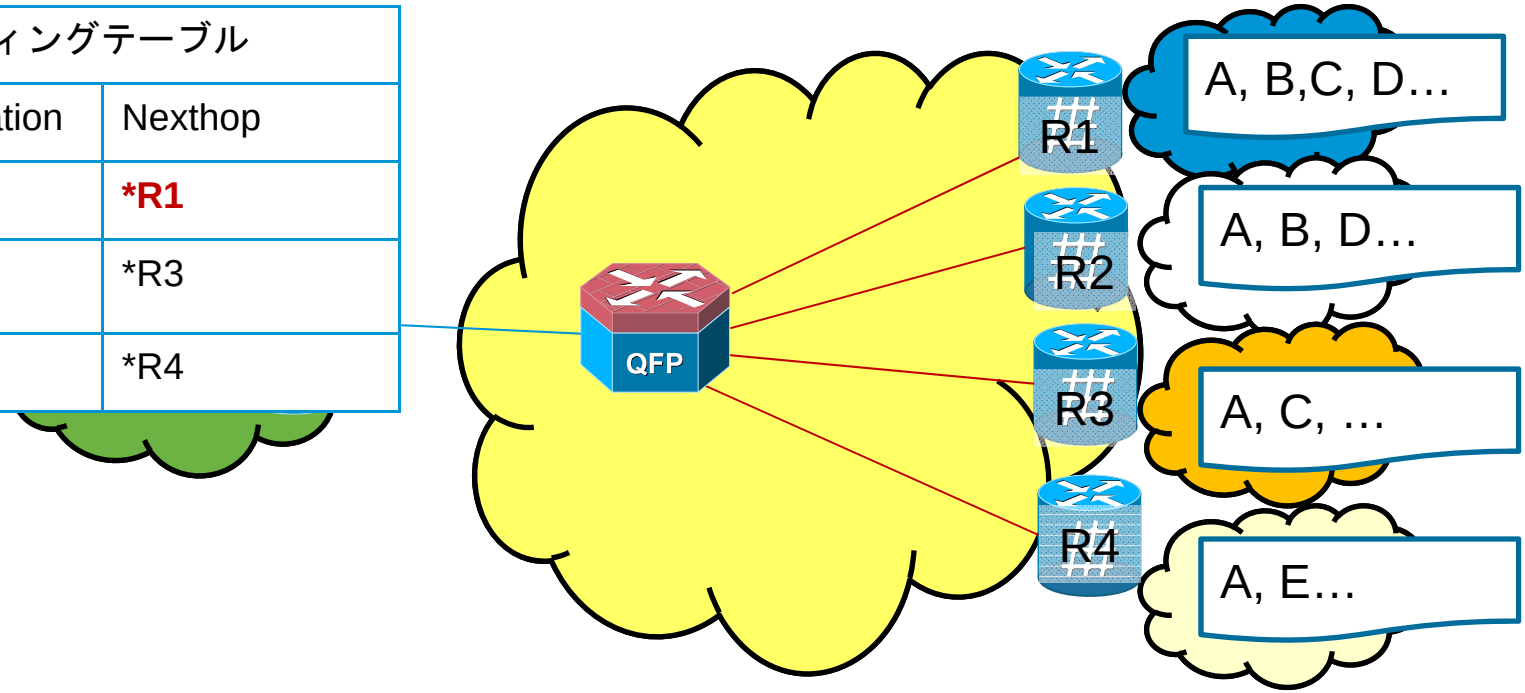
- S-VAではVA Prefix0/0をまず計算する
- VA Prefixと同じnext hop持つルートをSuppress する

Simple Virtual Aggregation(S-VA)

draft-ietf-grow-simple-va

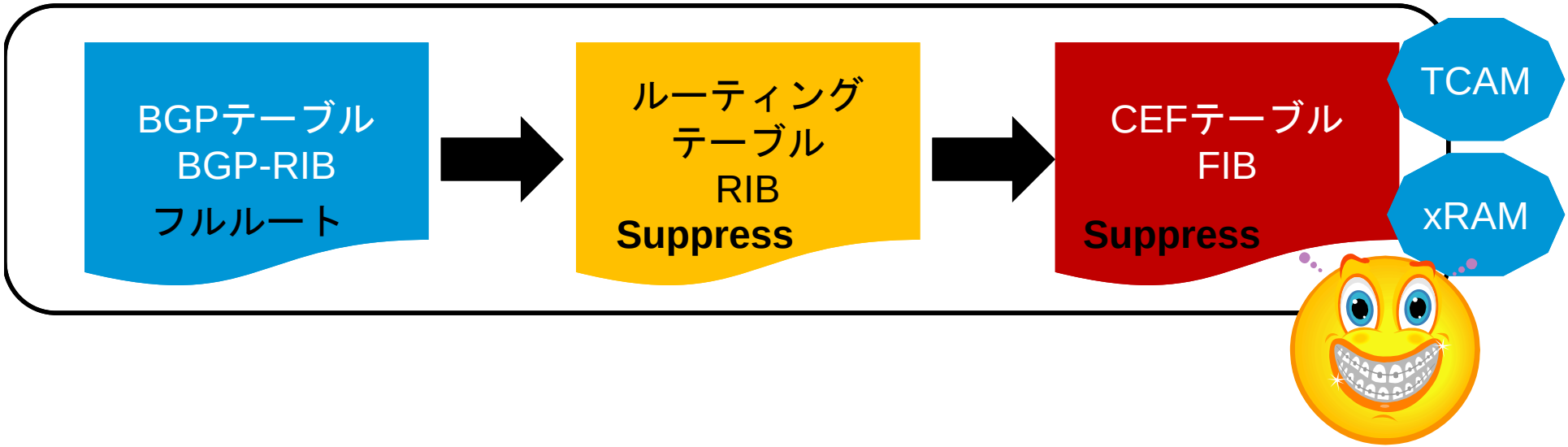
cont'd

ルーティングテーブル	
Destination	Nexthop
0.0.0.0	***R1**
C	*R3
E	*R4



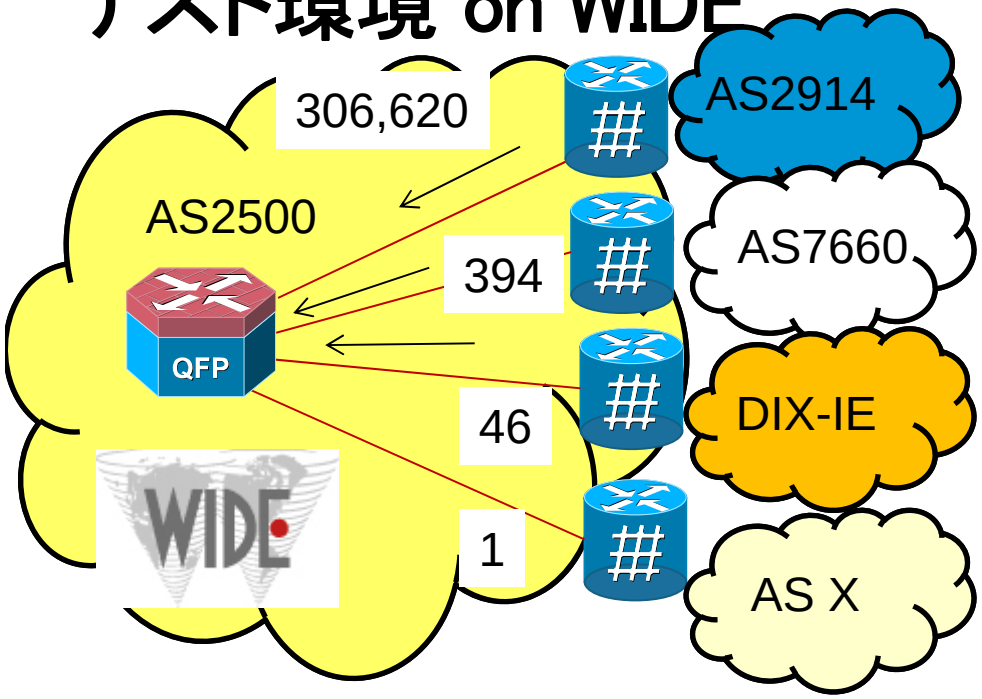
- BGPテーブルは通常通りだが、ルーティングテーブル/FIBが削減される

S-VAでのFIB作成方法



- RIB/FIBのメモリ容量が削減される
- TCAM/xRAMでのエントリーも少なくなり、他のリソースに影響を与えない。ルックアップも簡略化される。

テスト環境 on WIDE



	BGP RIB	RIB	BGP メモリ	RIB メモリ
通常	307,061	307,041	61MB	169 MB
S-VA	307,061	429	61MB	12MB

- WIDE(AS2500)での試験結果
- ルーティングテーブルは0.14%に縮小(300K->400)
- メモリ使用量は92%圧縮された(169MB->12MB)

Simple Virtual Aggregation(S-VA) まとめ

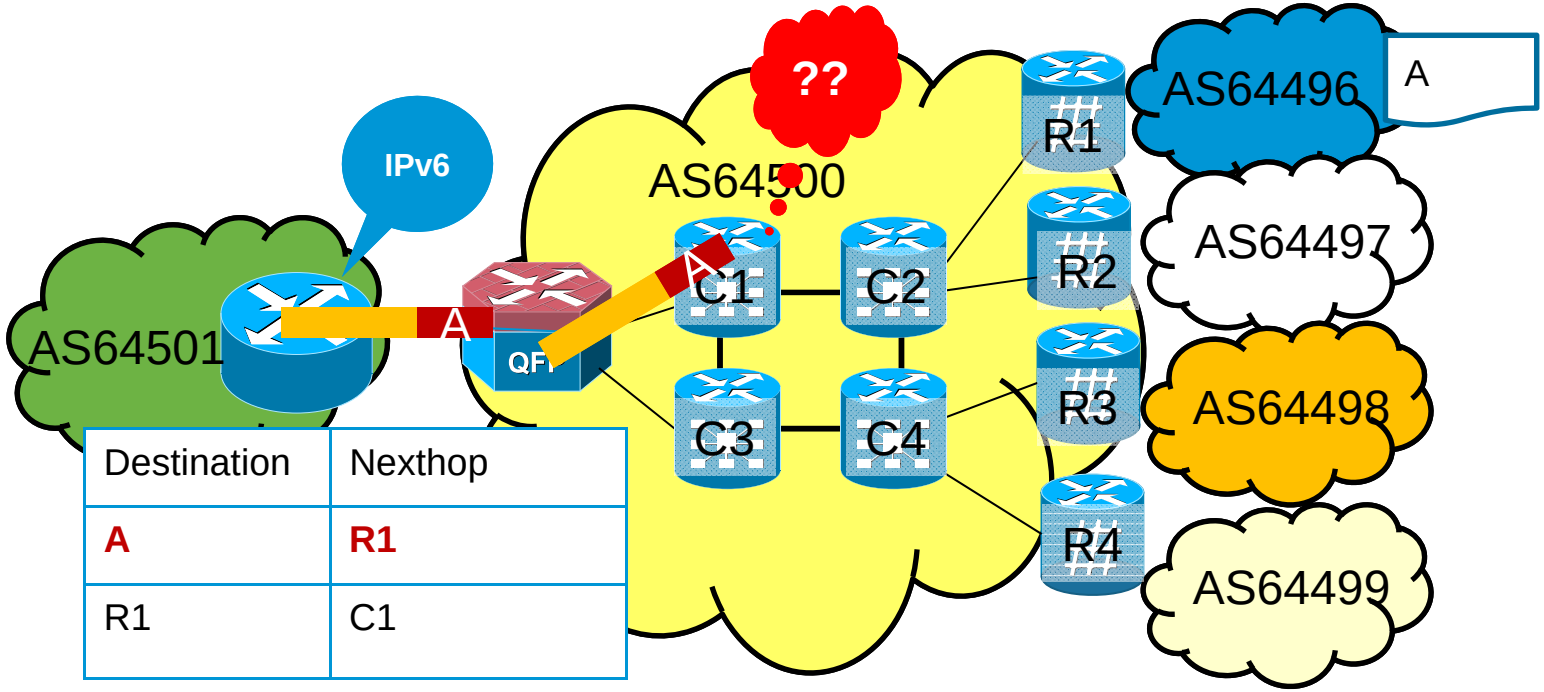
- S-VAは現在の多くのネットワーク(BGPフルメッシュ環境でFIBを節約する為のテクニックである
- プロトコルの拡張はせずに、エッジルータ(FSR)の機能拡張のみで、実現が可能である

Core of Coreの要件

	Access	Core of Core	Peering
インターフェース種別	様々	100GE/40GE/10GE	10GE/1GE
BGPルート数(広告数)	フルルート	無し	顧客分 サービス分
BGPルート数(受信数)	フルルート	フルルート(transitの為)	フルルート
FIB	S-VAで削減	大きい(transitの為)	大きい
デュアルスタック	必要	必要(transitの為)	必要
コスト	$	$$$	$$

- Core of Coreまたはスーパコアでは安定性・大容量・高速処理が求められる
- IP環境ではルーティング処理の為にCore of Coreでもフルルートが必要となる

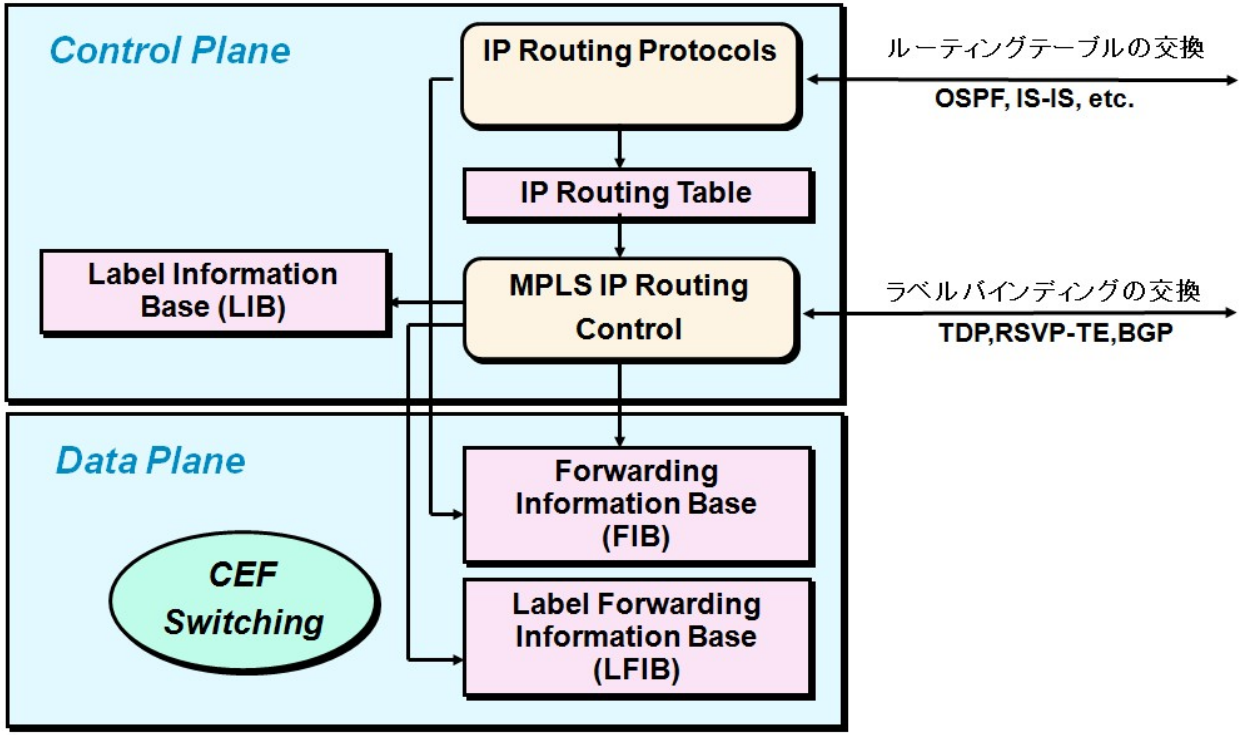
IP環境でのCore of Core



Destination	Nexthop
A	**R1**
R1	C1

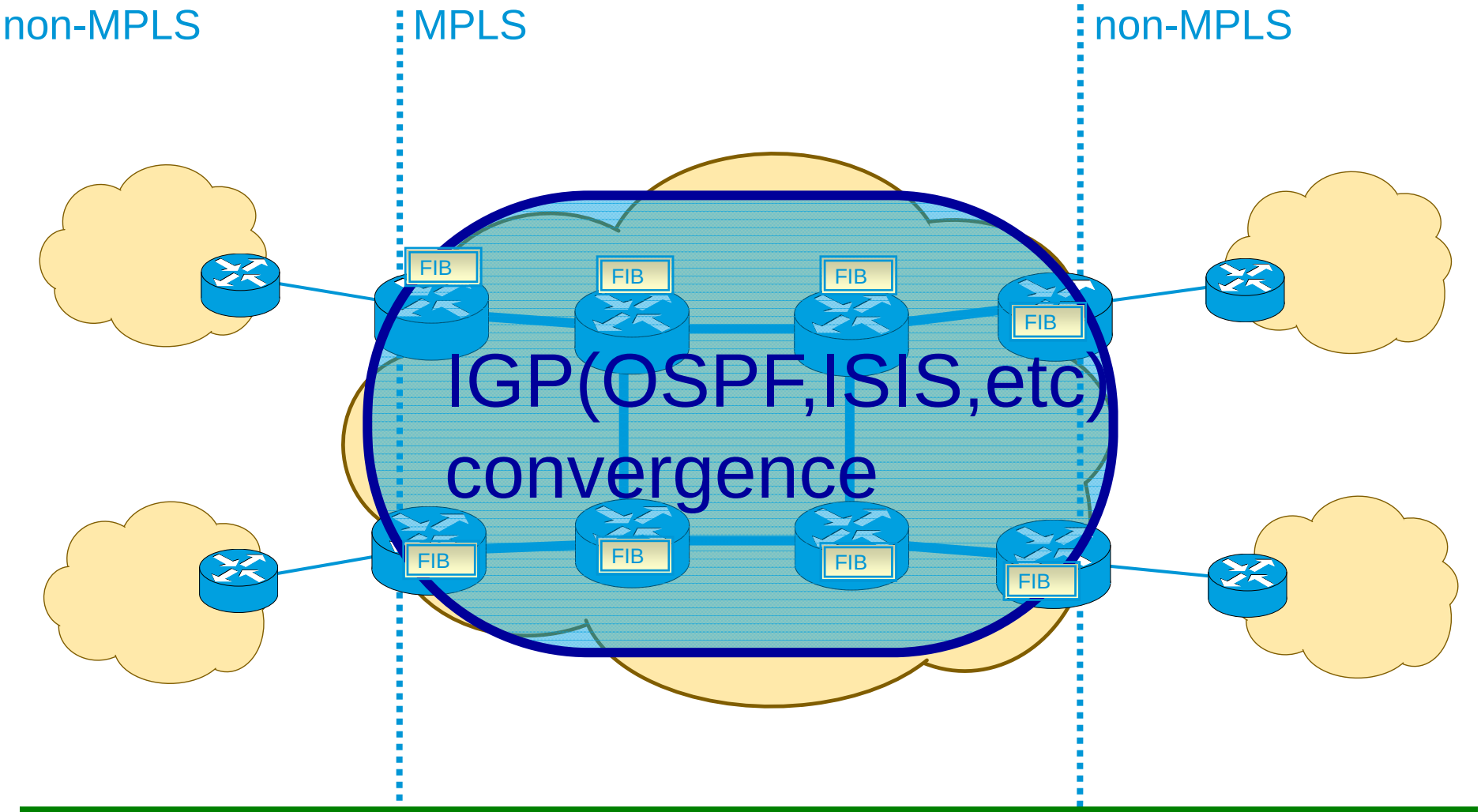
- CoreノードではASBR宛のパケットを転送する
- ルーティングテーブルの解決の為に同等なFIBが必要
- 顧客がIPv6を要求した時にはコアもIPv6にする必要がある
- 単純なアクションにも関わらず、要求に依存する

コアノードのMPLS化



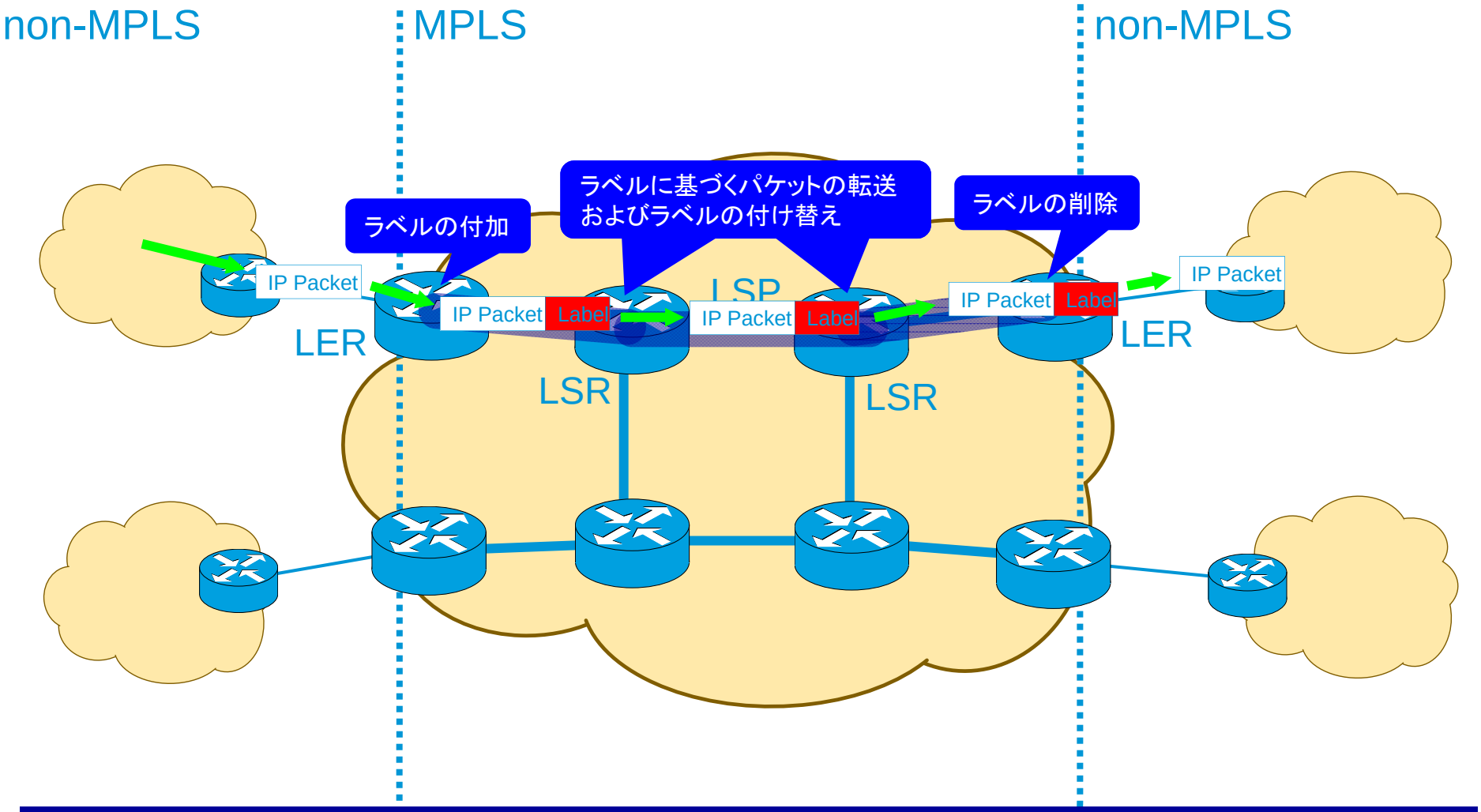
- MPLS(Multiprotocol Label Switching)ではアプリケーション(VPN,IPv6,Ethernetなど)に依存せず、共有のコアを使用可能
- コアではLFIBを元にパケットをフォワーディングしていく

MPLS動作概要



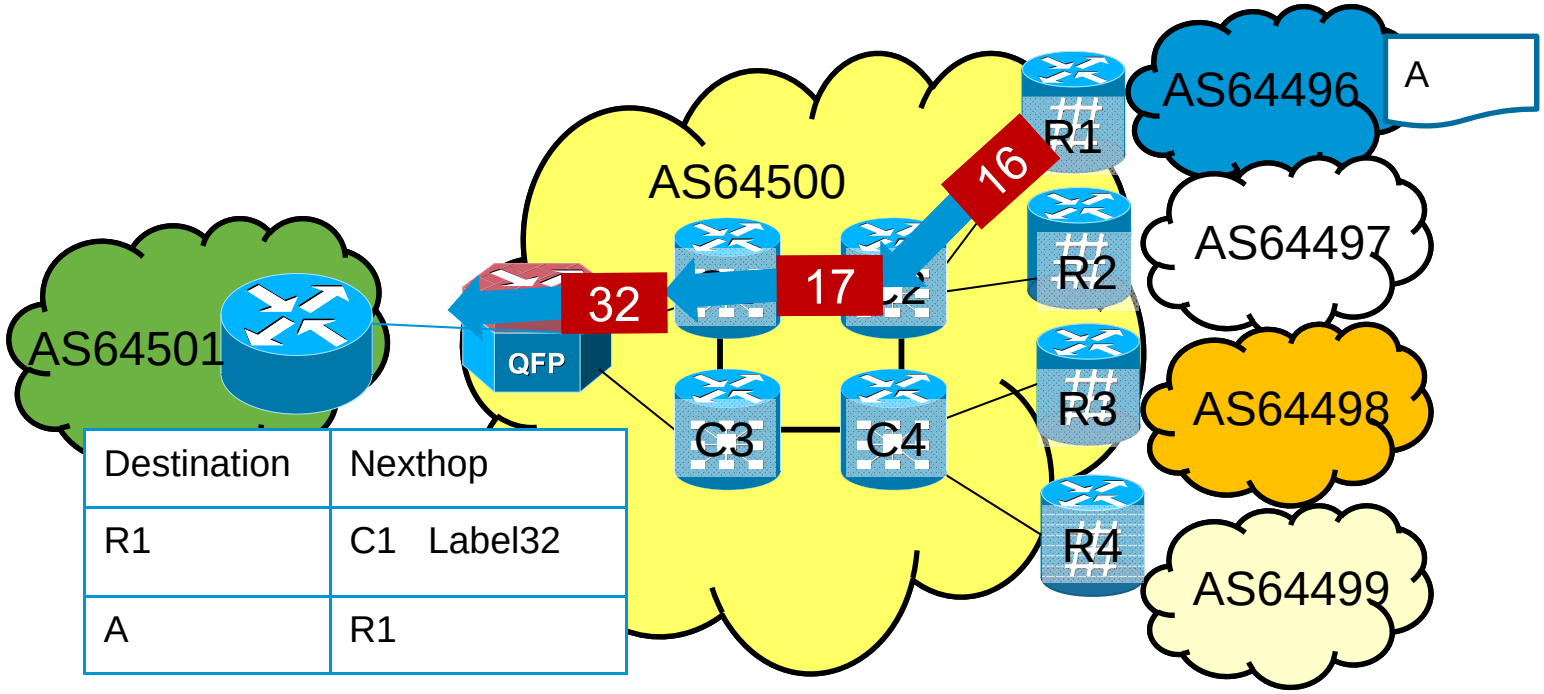
Control Plane : Routing Protocolsの収束、FIBの作成

MPLS動作概要



Data Plane : Label値に基づくパケットフォワーディング

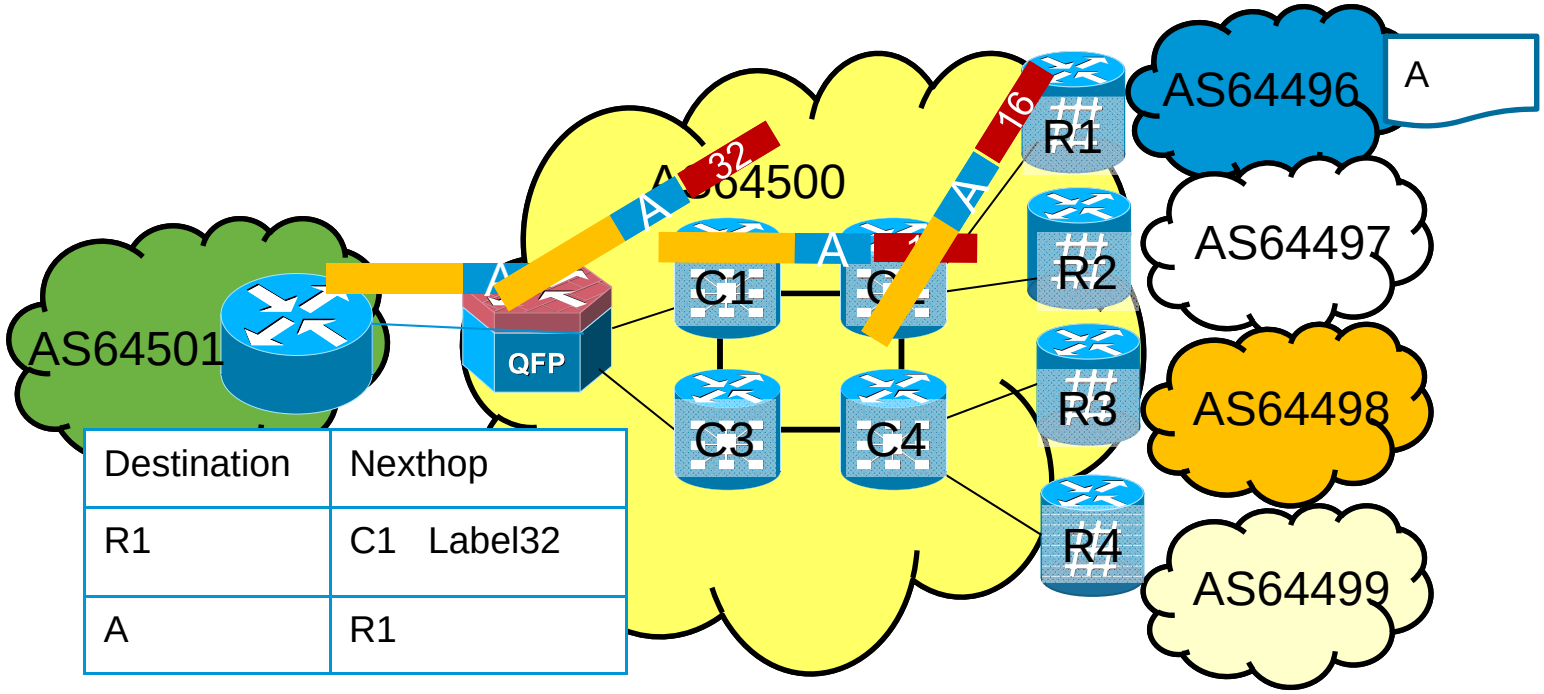
Core of CoreのMPLS化



Destination	Nexthop
R1	C1　Label32
A	R1

- AS内のloopbackアドレスをIGPで学習する
- LDPもしくはRSVP-TEを有効にし、loopbackへのLSPを作成
- EdgeでiBGPを設定する

Core of CoreのMPLS化



Destination	Nexthop
R1	C1 Label32
A	R1

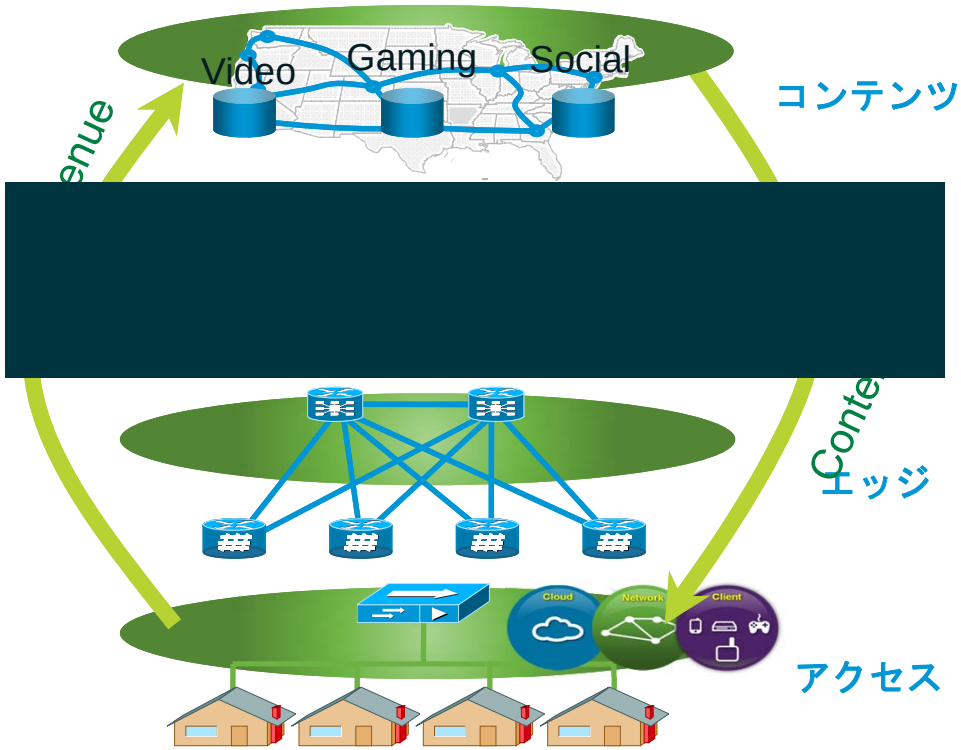
- Edgeではルーティングテーブルのルックアップを行う
- BGPネクストホップのラベル値を取り付け転送する
- コアではLFIBに基づきラベル値だけを見て転送する

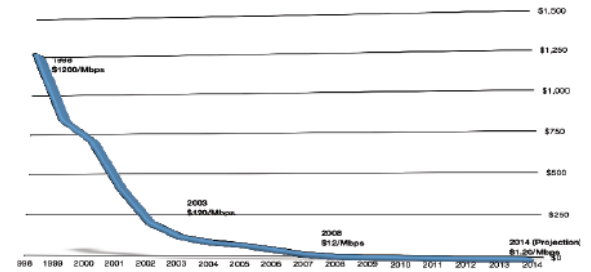
MPLSなら十分か!?

	Access	Core of Core	Peering
インターフェース種別	様々	100GE/40GE/10GE	10GE/1GE
BGPルート数(広告数)	フルルート	無し	顧客分 サービス分
BGPルート数(受信数)	フルルート	無し(MPLS化)	フルルート
FIB	S-VAで削減	小さい(LFIB化)	大きい
デュアルスタック	必要	不要(6PEの活用)	必要
コスト	$	$$$	$$

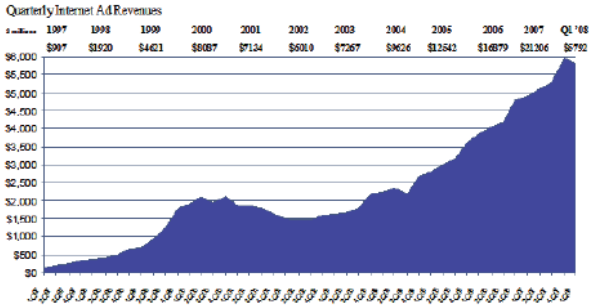
- コアを通過するトラフィックは年々成長している
- 帯域は増加し、大容量なインターフェースが求められる

ハイパージャイアントの出現による収益モデルの変化





Revenue from
Internet Transit
Source: Dr. Peering, Bill Norton



Revenue from
Internet Advertisement
Source: Interactive Advertising Bureau

- 顧客はコンテンツをASPから直接購入する。
- 帯域と機器コストは利益を持たないSPが受け持つ

サービスプロバイダーの憂鬱

- トラフィックはどんどん成長していく
  100GE/40GE？10GEを複数束ねる?
  WDMとの接続？
- ハイパージャイアントの出現
  利益を生まないトラフィックが増え続ける
- コストがどんどん高くなる。。。

Lean Core

Lean（リーン）:「贅肉のとれた」なコアに特化したテクノロジー

贅沢な機能を削り、無駄なコストを省いたプロダクトの提供

Lean Core Implementation Example
CRS-LSP

Cisco CRS
Single, MC, B2B

MSC140
Full IP/MPLS

FP140
Peering/ Aggregation

CRS-LSP
LSR Core

New

Flex Packet Transport/LSRとは何か？

- ピュアMPLSラベル・トランスポート・デバイス
- コアに適したスケーラビリティ/トランスポート/QoS
- 制限されたIP FIBテーブル
- 広帯域
- 低遅延
- MPLS TEミッドポイントとしてのスケーラビリティ
- ファーストリルート 50msecコンバージェンス
- OAM
- IPoDWDM

Service Card Feature比較

	LSP	FP140	MSC140
Bandwidth Full Duplex	140Gbps	140Gbps	140Gbps
Packets Per Second	125Mpps	125Mpps	125Mpps
CRS-3 PLIM Support	All	All	All
IP FIB Size (shared between v4, v6, Mcast)	16K	1M-4M*	4M
ACLs	Limited Support	Yes	Yes
Policiers	Limited Support	Yes	Yes
Multichassis	License *	License	Yes
QoS Queues	8/Port	8/Port	64K Total
MPLS TE midpoints	100K **	License	100K
Enhanced Netflow	License	License	Yes
VPLS	Not Supported	Yes	Yes
L3VPN,L2VPN	Not Supported	License	Yes
VIDMON	Not Supported	Yes	Yes
PSE+PLIM価格感	35%	62.5%	100%

Lean Coreまとめ

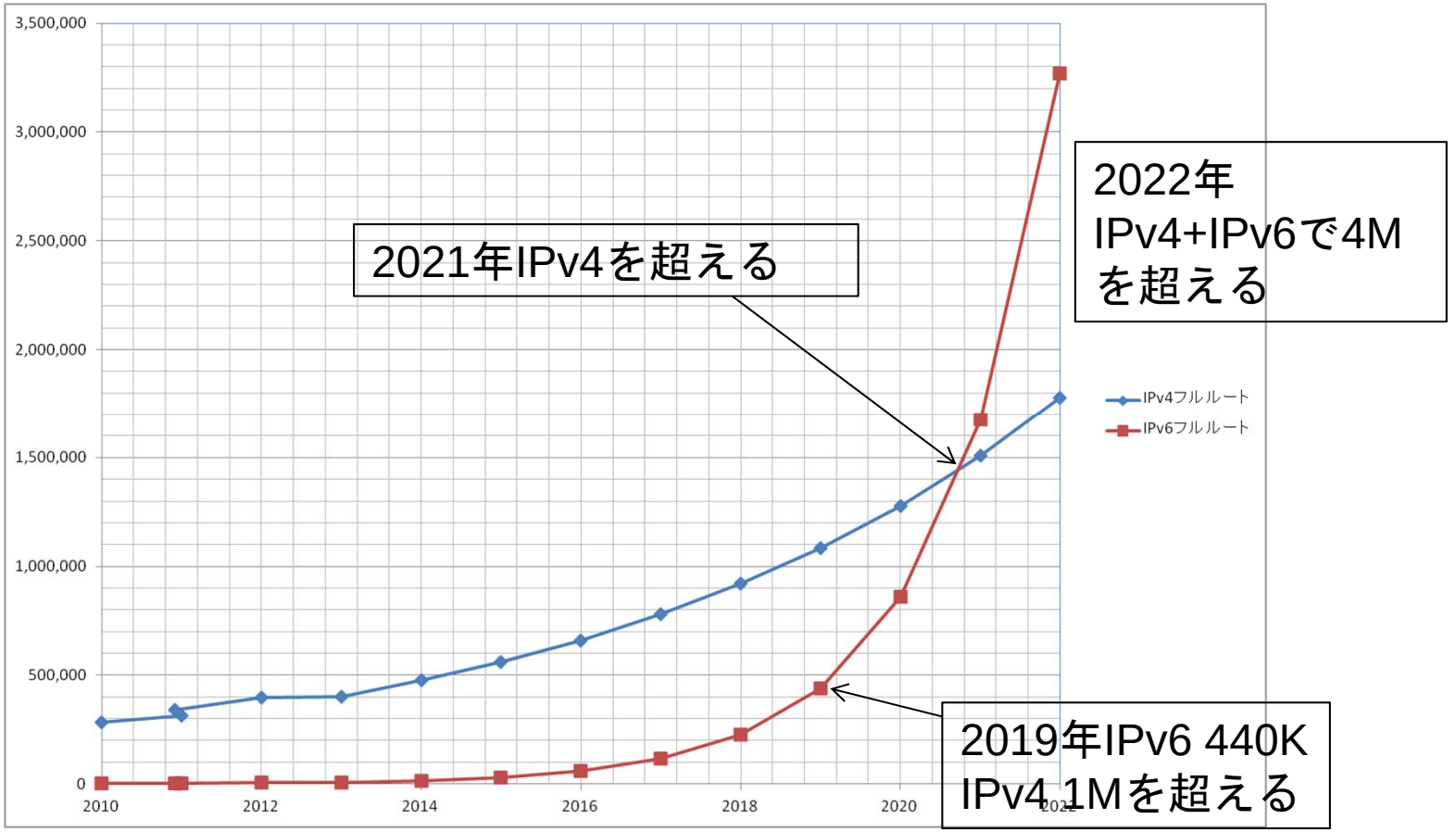
- FIB Suppression with Virtual Aggregation にてコアでのMPLSを含めたFIBの抑制方法に関して述べられている
- Auto-Configuration in Virtual AggregationではVP-Listの設定をBGP Extended Communityを行う
- コアのMLS化はBGPフリーコアと呼ばれ多くのキャリアで実績がある
- CRS-LSPはLean Coreの実装例であり、Lean Coreは安定的なコアを経済的に提供する。

Peering Aggregationの要件

	Access	Core of Core	Peering
インターフェース種別	様々	100GE/40GE/10GE	10GE/1GE
BGPルート数(広告数)	フルルート	無し	顧客分 サービス分
BGPルート数(受信数)	フルルート	無し(MPLS化)	フルルート
FIB	S-VAで削減 VP-list	小さい(LFIB化)	大きい
デュアルスタック	必要	不要(6PEの活用)	必要
コスト	$	Lean Core	$$

- インターネットルーティングテーブルの成長率の影響をうける

インターネットルーティングテーブルの成長率



- 2011年の11月の成長率を計算し、ルート増加数を計算したグラフ

ルーティングテーブルを圧迫する理由
draft-narten-radir-problem-statement

- トラフィックエンジニアリング
  異なるパスをトラフィックが通過させる為にルートを追加する
- マルチホーミング
  信頼性を上げるために複数のサイトに接続し、ルートを追加する
- リナンバリング
  ISPを変更してもアドレス変更をしたくないサイトは自分のprefixを持つ
- 買収、合併
  会社が統合されてもリナンバリングはしたくない
- RIRポリシー・IPv6化・内部ルート増加・IPv4枯渇など多くの理由があげられる。

IRTF RRG Recommendation

Recommendation for a Routing Architecture

- LISP
- RANGI
- Ivip
- hIPv4
- NOL
- Evolution
- Name-based sockets
- LMS
- 2-phased mapping
- GLI-Split
- TIDR
- ILNP
- EEMDP
- IRON-RANGER

- IRTFではこれらの問題に対し、2007年から2010年まで議論した。
- Chairが3つのテクノロジーの推薦を行なった

RRG Chair Recommendation

- Evolution

  VAの概念をインターASに広げる

- Work on renumbering

  IPv6でも少しづつ問題が残っているので検討をする

- Identifier/Locator Network Protocol (ILNP)

Name

- Nameの定義

  指定するオブジェクトの中身を意味する

- 例：

  プロトコルname http

  ポート番号　80

  fully qualified domain name(FQDN)

  shtsuchi.apac.cisco.com

  IPアドレス　10.10.10.1

Application Layer Protocol

- URLs

  https://shtsuchi.apac.cisco.com

- IPアドレスも使用可能

  https://10.10.10.1

- DNSネームとIPアドレスが共に使われている-IPアドレスとFQDNが同義語
- IPアドレスは過負荷です。
- Application Layerでセッション**ID**として使われている

トランスポートプロトコル

- TCPではセッションを識別する為に、
  - ローカルIPアドレス
  - ローカルポート番号
  - リモートIPアドレス
  - リモートポート番号
- TCPステートはすべてローカル/リモートIPアドレスに紐づいている
- IPアドレスはIdentifier(識別子)として使われている

ネットワークレイヤー

- IPアドレスはルーティングで使用される
  10.10.10.1/24は10.10.10がルーティングに使用される
- アドレスのホスト部はサブネットに使用されてもよい。
  10.10.10.1/25 アドレスのホスト部分の1ビットが使用されている

- IPアドレスはLocatorに使われている

IPのBug

プロトコルレイヤー	IP
アプリケーション	FQDN or IPアドレス
トランスポート	IPアドレス (+ポート番号)
ネットワーク	IPアドレス
(インターフェース)	IPアドレス

- すべてのレイヤーにIPアドレスが複雑にからみ合ってしまっている

ILNPv6

- IPv6拡張に見える
  - ネットワークコアではIPv6と同じパケットフォーマット
  - IPv6コアルーターは変更を必要としない
  - IPv6コアで大規模展開が可能
  - IPv6へ下位互換性がある
- 128ビットのIPv6アドレスをSplitする
  - ✓64-bit Locator (L) Network Name
    - ルーティング・フォワーディングのみに使用される
    - 変更出来る・複数同時に持つことが出来る
  - ✓64-bit Identifier (I)　node name
    - ノードの名前（インターフェースの名前では無い）
    - トランスポートのライフタイムの間は保持する。複数同時に持てるが同じセッションでは出来無い
- 上位レイヤーはIdentifierのみをバインドする

IP vs ILNPv6

プロトコルレイヤー	IP	ILNPv6
アプリケーション	FQDN or IPアドレス	FQDN (RFC1958)
トランスポート	IPアドレス (+ポート番号)	Identifier (+ ポート番号)
ネットワーク	IPアドレス	Locator
(インターフェース)	IPアドレス	Dynamic Mapping

- ILNPではすべてのレイヤーで分離する

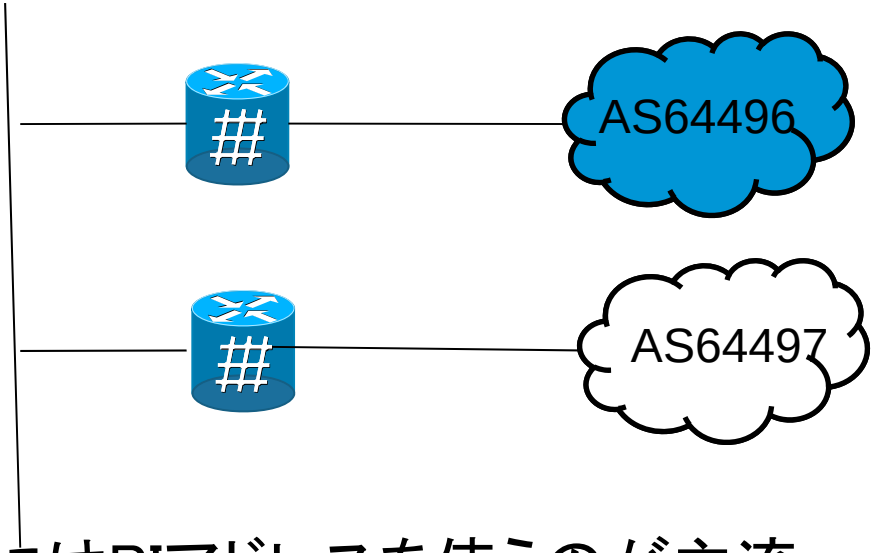
ILNP operation

- ホストはlocaterとIdentifier をDNSを通じて知る
- 通常は一つのIdentifierで、複数のlocatorを持つ

```
ahost.foo.com IN   AAAA        2001:db8::53
              IN   ID          10 00:00:d4:9a:20:0f:08:e0
              IN   L64         11 2001:0db8:0102:0304
              IN   L64         12 2001:0db8:0506:0708
```

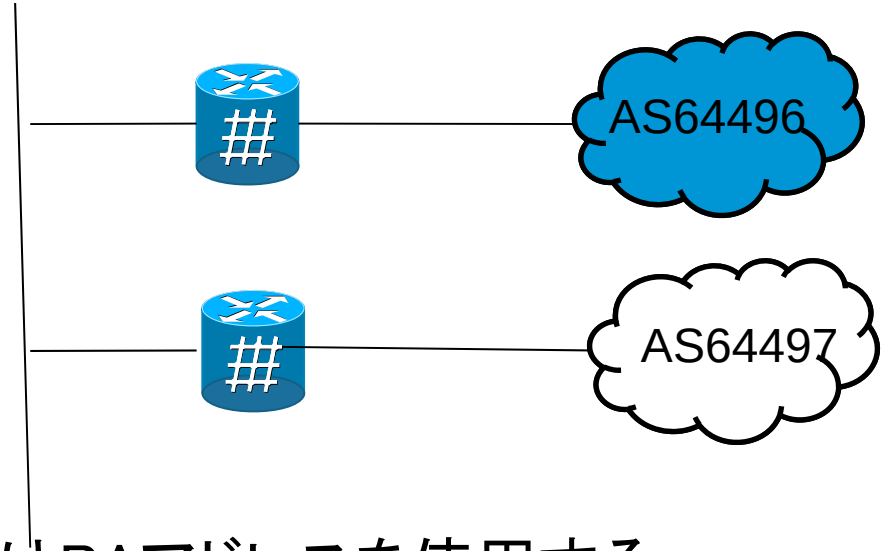
- Locatorの変更はICMP Locator Updateメッセージで行う。セキュルティの為にnonceを含む

マルチホーミング



- マルチホーミングにはPIアドレスを使うのが主流
- Prefix=Identity
- リナンバリング出来無い
- アグリゲート出来無いアドレスがコアに流入する
- サイト毎にリンクの数、Prefixの数でRIBが増加する

ILNPでのマルチホーミング



- ILNPではLocatorはPAアドレスを使用する。
- IdentityはLocatorには関係無い
- リナンバリングは通常のIPv6だけで行われ、ICMP Locator Updateが送られる
- 余計なPrefixは何も送られない。

ILNPのまとめ

- ILNPは既存のIPの問題点をシームレスに解決する為のホストベースのプロトコルである
- ILNPが導入されれば、現在のインターネットの経路爆発の原因となっているマルチホーミングやリナンバリングを解決しうる事が出来る。
- またモビリティやVM Migrationなど多くの問題も解決可能

Key Takeaway

- FIB/RIBの違い
- TCAM/xDRAMの違い
- S-VAの動作
- Lean Coreの概要
- ILNPの概要

Thank you.