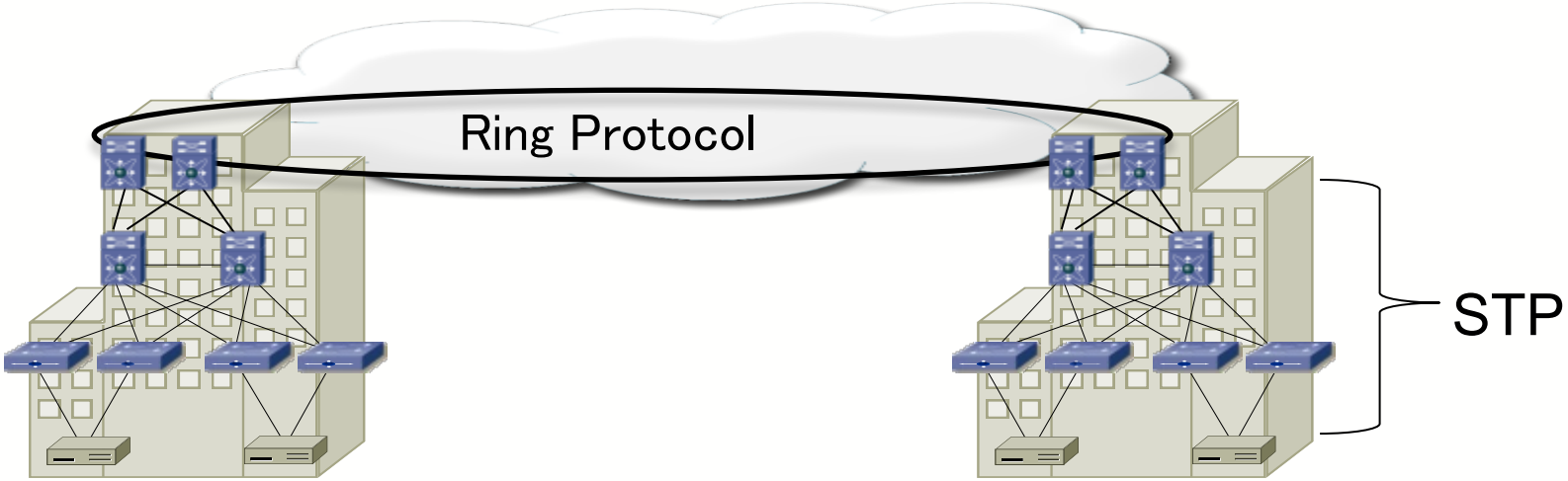
データセンターでのルーティングプロトコル

Shishio Tsuchiya
shtsuchi@arista.com

アジェンダ

- **今のデータセンタールーティングが生まれてきた訳**
- EVPNデータセンター内デザイン
- ダイナミックディスカバリー
- IETF Dcrouting

2011年ごろのデータセンターネットワーク



- フラットなレイヤー2のネットワークデザイン
- データセンター内ではスパニングツリーが使用
- ベンダー独自のリングプロトコル (またはG.8032)がデータセンター間で使用される
- VMライブマイグレーションは必要 (GARPによって移動を通知)
- VLANがユーザ毎にアサインされる

問題点

- データセンター間/データセンター内
  - VLANスケーラビリティ ＞ 4K
  - MACアドレステーブルのスケーラビリティ
  - VMライブマイグレーションや簡単に使うためのブロードキャストドメインの拡張
  - East / Westトラフィック帯域の増加
  - 高速収束
  - 自動化/オーケストレーション
- データセンター間
  - 要求に応じた帯域増強
  - ベンダーロックイン技術からの解放
  - 柔軟性のあるトポロジーデザイン
  - トラフィックエンジニアリング
  - BUM(Broadcast/Unknown unicast/Multicast) トラフィックの最適化
- ゲートウェイ
  - ARP/NDPスケーラビリティ
  - IETF ARMD(Address Resolution for Massive numbers of hosts in the Data center) Groupは一つのinformational RFCを発行 **RFC6820 Address Resolution Problems in Large Data Center Networks**

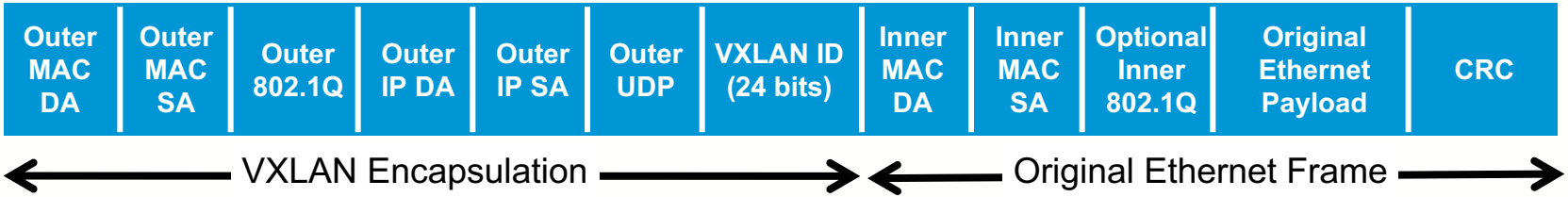
問題に関するソリューション

- 仮想オーバーレイプロトコル
  - VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks
    - draft-mahalingam-dutt-dcops-vxlanは2011年8月に experimental として発行
      - vmware/cisco/arista/broadcom/citrix/Redhat
    - RFC7348は2014年8月に Informationalとして発行
  - NVGRE: Network Virtualization using Generic Routing Encapsulation
    - draft-sridharan-virtualization-nvgreは2011年9月に発行
      - microsoft/arista/Intel/Dell/HPE/Broadcom/Emulux
    - RFC7637は2015年にマイクロソフトによって informationalとして発行
  - A Stateless Transport Tunneling Protocol for Network Virtualization(STT)
    - draft-davie-sttは2012年2月nformationとして発行
    - Nicira Networks
    - 2016年にexpire
  - Overlay Transport Virtualization(OTV)
    - draft-hasmit-otvは2010年4月にスタンダードとして発行
      - Cisco
    - 2013年にexpire

問題に関するソリューション

- ネットワークデザイン
  - Use of BGP for routing in large-scale data centers
    - Microsoft Petr Lapukhov がNANOG55 June 3-6, 2012でプレゼン
    - またdraft-lapukhov-bgp-routing-large-dcを2012年7月に発行 Ariff Premji(Arista)が共同著者
    - RFC7938 は2012年8月にInformationalとして発行
  - Transparent Interconnection of Lots of Links (trill)
    - draft-perlman-trill-rbridge-protocolは2006年に発行
    - Trill WG はRbridges(RFC6325)を次世代 IEEE802.1Dプロトコルとして定義 July 2011
    - RFCの著者は Intel/Huawei/Cisco/Brocade

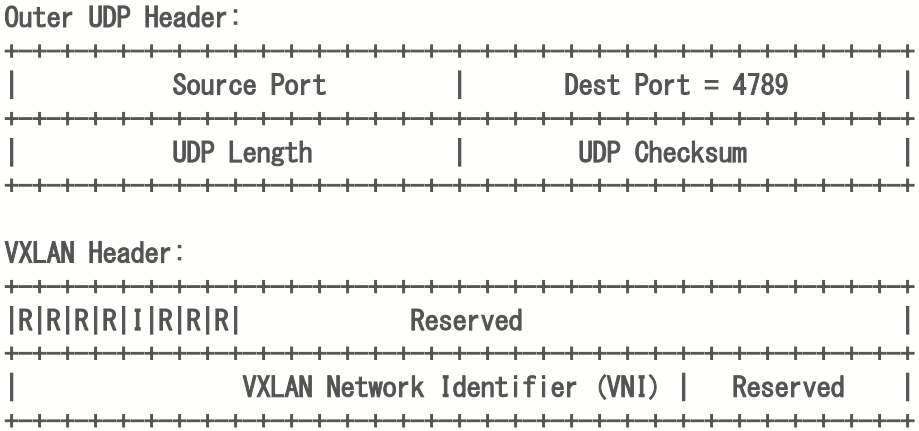
Virtual eXtensible LAN (VXLAN)



- レイヤー2フレームをIPでカプセルプロトコルを定義
- レイヤー3インフラストラクチャー上でオーバーレイのネットワークを作成する為に使用
- レイヤー2の接続性をユーザに提供

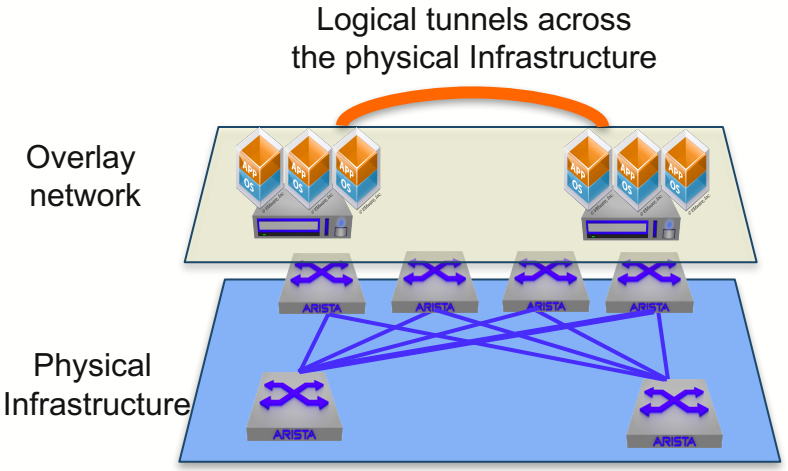
Virtual eXtensible LAN (VXLAN)フレームフォーマット

- フラグ(8 bits):
  - 有効なVXLANのネットワークIDの場合、Iフラグは1をセットしなければならない。他の7ビット(R)は予約フィールドで送信時に0をセットしなければならず、受信時に無視される
- VXLANセグメントID/VXLANネットワーク識別子 (VNI):
  - これは、通信するVMが配置されている個々のVXLANオーバーレイネットワークを指定するために使用される24ビットの値です。異なるVXLANオーバーレイネットワーク内のVMは互いに通信できません。
- 宛先ポート:
  - IANAはVXLANのポートとして4789をアサインした。このポートをデフォルトの宛先ポートとして使う
- 送信元ポート:
  - UDPソースポート番号は、ロードバランスの際のハッシュの計算に使用される。動的にプライベートポート範囲49152-65535である事が推奨される

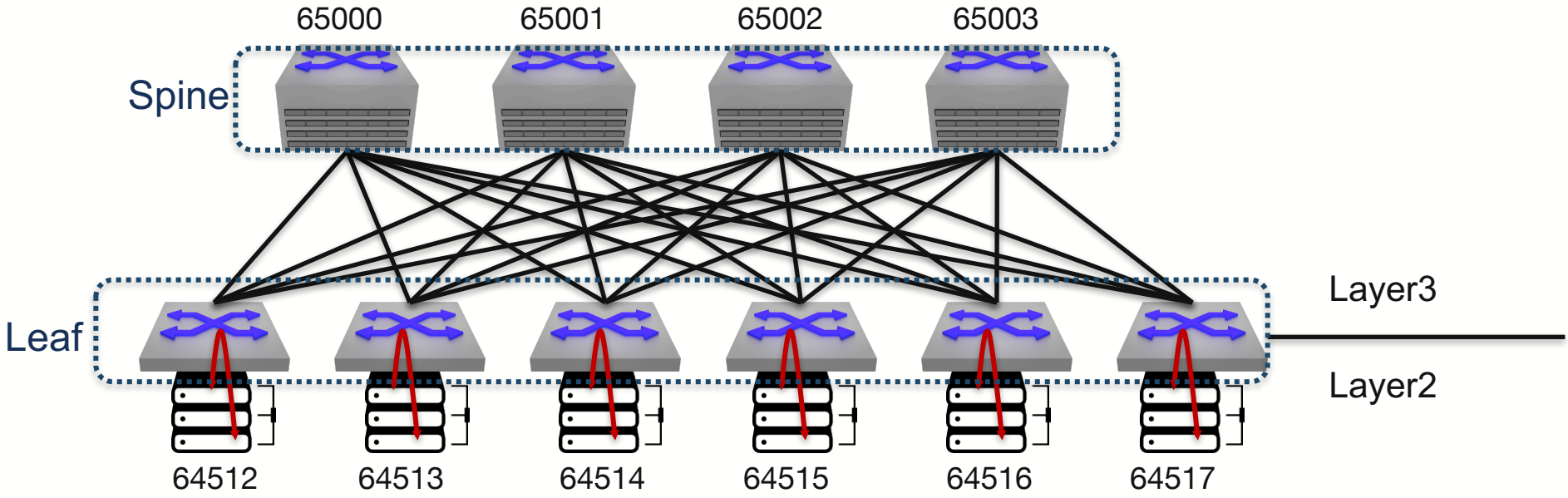


オーバレイとアンダーレイ

- オーバーレイネットワーク（論理インフラ）
  - 物理から切り離された論理トポロジー
  - ルーティングされる物理インフラストラクチャーを通るトンネルにて構成される
  - トンネルは物理と論理エンドポイントの接続性を提供
- アンダーレイ(物理インフラ)
  - オーバーレイ技術に透過的
  - レイヤー3インフラでの構築を許可
  - 物理ネットワークは通信の為の帯域幅とトポロジーのスケールを提供
  - 仮想化デバイスとノードのスケール要件は物理から抽象化される



大規模データセンタールーティングでのBGPの使用

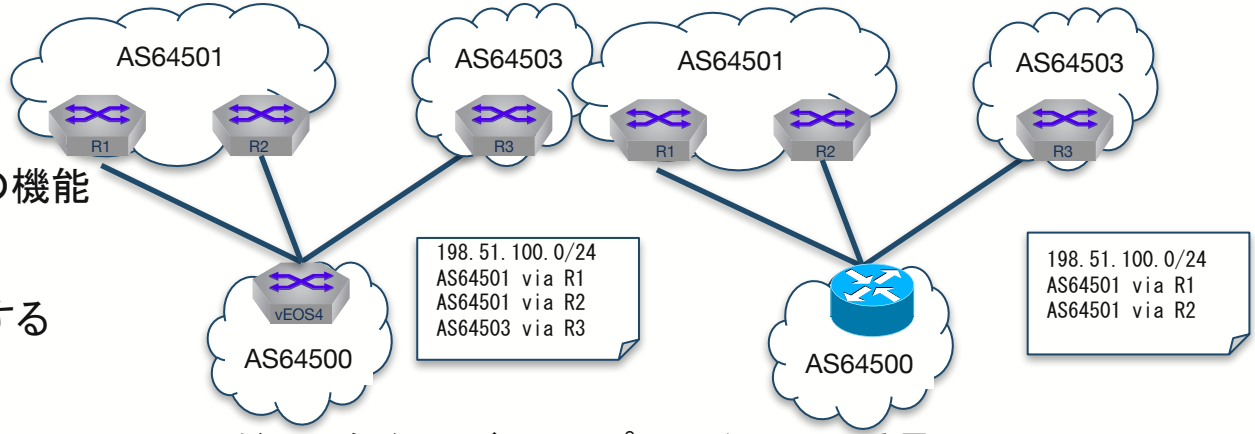


- スケールアウトするClosデザイン
- 安定した標準的なBGPプロトコルをToR/Leafスイッチに使用
- 安定性にフォーカスし、VMモビリティはラック内に留める

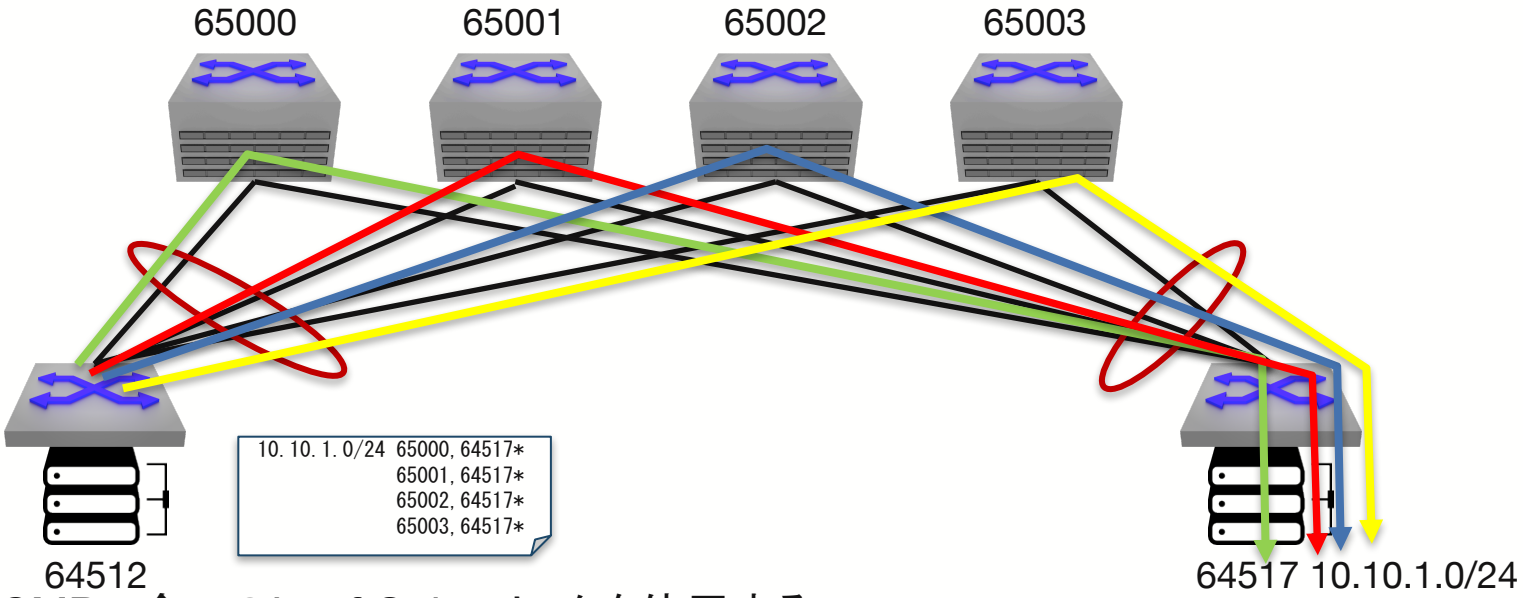
ARISTA

このネットワークデザインにおける特別なBGPの要件

- AS_PATH Multipath Relax
- Allow AS In
  - 重複したAS番号を使用する為の機能
- Fast eBGP Fall-over
  - 高速コンバージェンスを可能にする
- Remove Private AS
  - 2バイトのプライベートAS番号 (64512-65534)だけでなく、4バイトのプライベートAS番号 (4200000000 – 4294967294) もエッジで除去
  - RFC6996 Autonomous System (AS) Reservation for Private Use

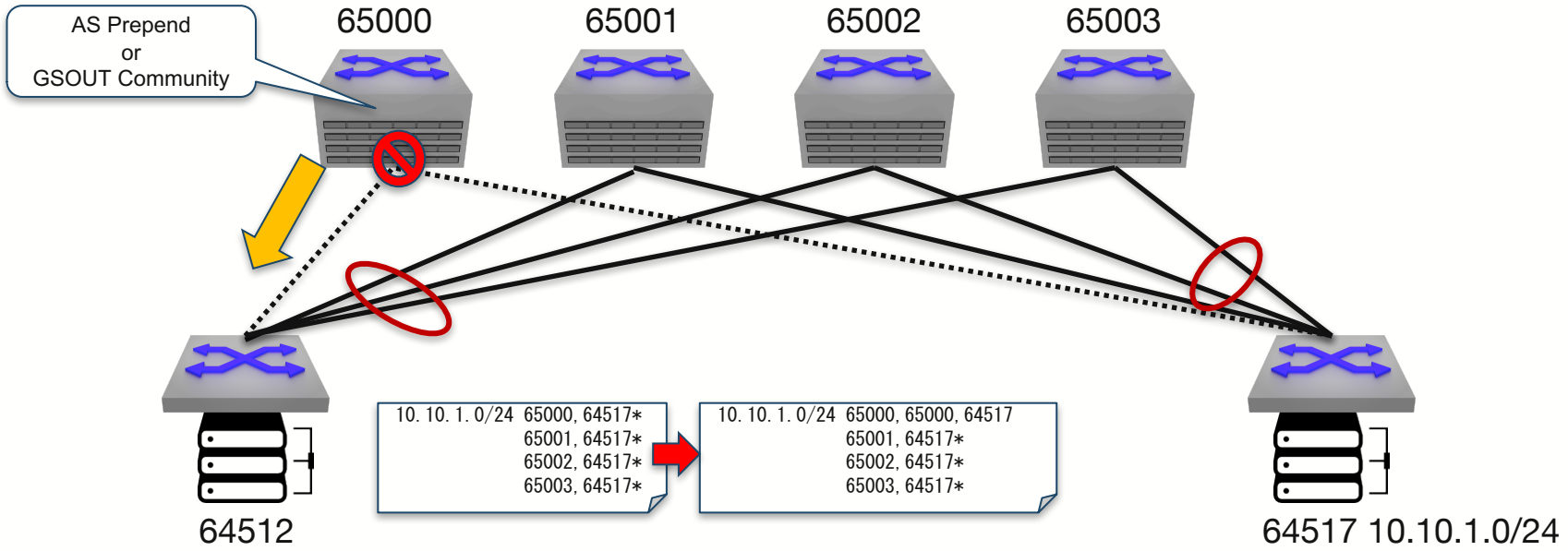


IP ECMPの利点



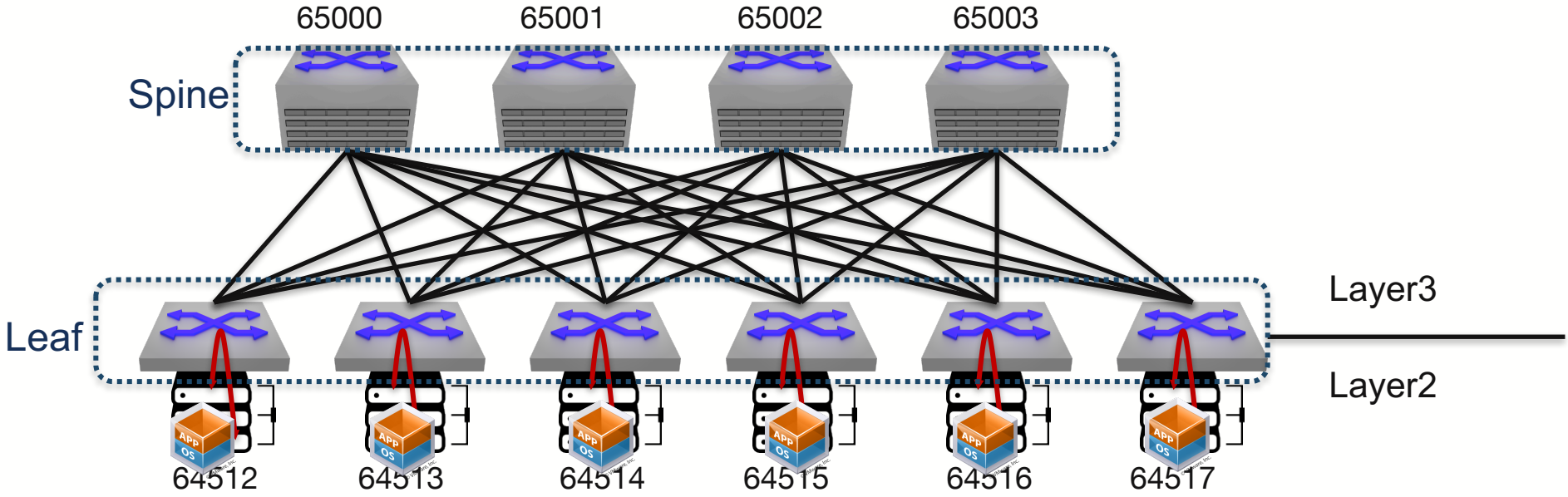
- ECMPで全てのLeaf-Spineリンクを使用する
- それぞれのフローはECMPハッシュにてバランスされる(既に実装されている)
- ルーティングパスはBGPパス属性により可視化される

BGPの利点



- ASパスを追加するもしくはBGPグレースフルシャットダウンコミュニティ(GSHUT (0xFFFF0000) community)を使って簡単にメンテナンス可能

VXLAN + IP Closデザイン



- データセンターデザインで完全なソリューション
- 標準で安定したBGPプロトコル/広く展開されたVXLANフォーマットを使用
- 特に特別な要求が中間ノードには存在しない

VXLANコントロールプレーンの選択

- VXLANコントロールプレーンはMAC学習とパケットフラッディングに使用される
  - リモートVTEPの背後にいるホストを発見するメカニズム
  - どのようにVTEPとVNIのメンバーを発見するのか?
  - レイヤー2セグメント(VNI)内でブロードキャスト/マルチキャストを転送する為のメカニズム

IPマルチキャストコントロールプレーン

- VTEPはVNIの所属するマルチキャストグループにjoin
- VNI内のUnknownユニキャストはVTEPにマルチキャストで転送
- サードパティVTEPをサポート
- 転送と学習はIPマルチキャストを要求-限られた展開例

HeadEnd Replication (HER)

- BUMトラフィックはVNIの中のリモートVTEPに複製される
- ingress VTEPで複製される
- サードパーティのVTEPをサポート
- MAC学習は転送され学習されるしかし、マルチキャストは必要が無い

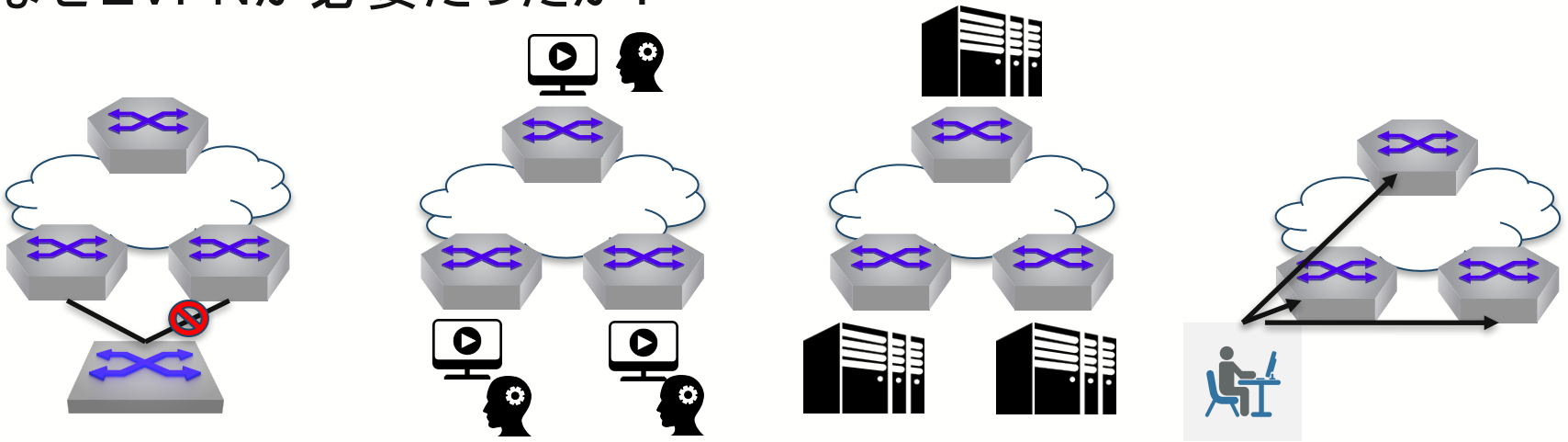
EVPN

- VTEP間のローカルに学習されたMACとIPの紐付けの為にBGPを使用
- ブロードキャストトラフィックはIPマルチキャストやHERによりハンドルされる
- 設定されたBGPによりMACアドレスとVNIはダイナミックに分配される
- サードパーティVTEPをサポート

アジェンダ

- 今のデータセンタールーティングが生まれてきた訳
- **EVPNデータセンター内デザイン**
- ダイナミックディスカバリー
- IETF Dcrouting

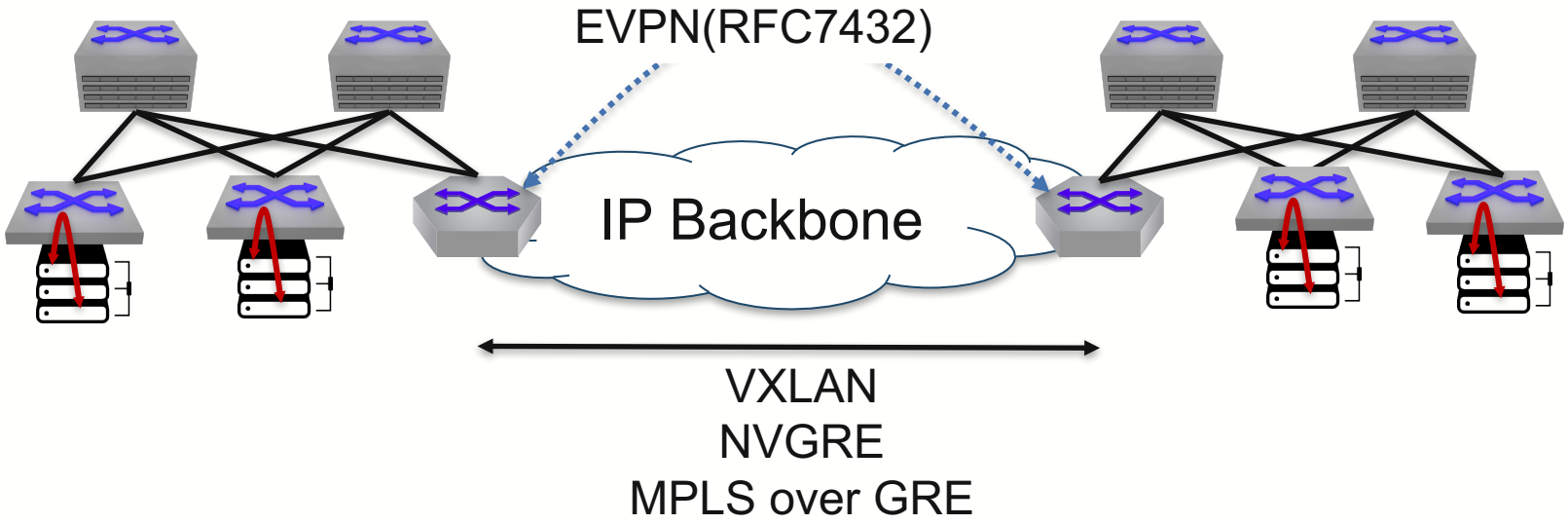
なぜEVPNが必要だったか?



- RFC7209 Requirements for Ethernet VPN (EVPN) は2つのVPLSモード(RFC4761 と RFC4762) 問題点を解決するために定義された
  - 冗長性の要求事項 Active/Activeリダンダンシーのサポート
  - マルチキャスト最適化要求: P2MPだけでは無く、MP2MPのサポート
  - データセンター間接続
  - プロビジョニングの簡素化
  - RFC7432 BGP MPLS-Based Ethernet VPN はスタンダードプロトコルとして発行された

A Network Virtualization Overlay Solution using EVPN

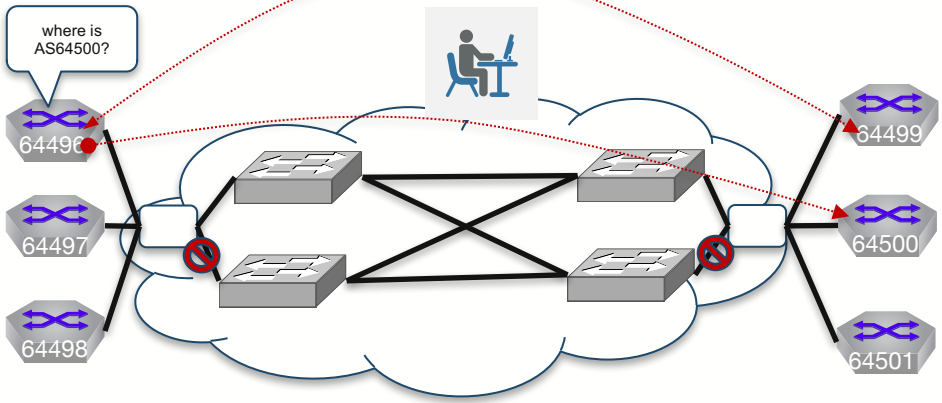
https://tools.ietf.org/html/draft-ietf-bess-evpn-overlay

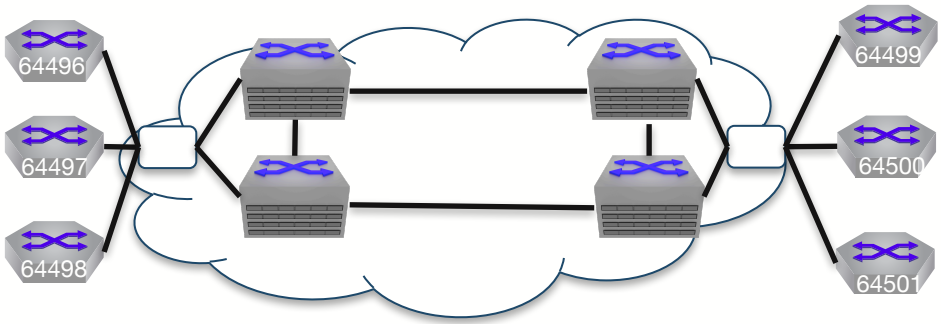


- IETF NVO3(Network Virtualization over Layer 3) WGは現在のマルチテナントネットワークの問題点を定義した RFC7364 Problem Statement: Overlays for Network Virtualization .
- draft-ietf-bess-evpn-overlayはEVPNの技術を使って、どの様にNVO3の要件を提供するのかのソリューションドラフト
- 現在のIESGステータスは Publication Requested, proposal standardとして発行予定

Internet EXchange(IXP) Use case

https://tools.ietf.org/html/draft-ietf-bess-evpn-proxy-arp-nd

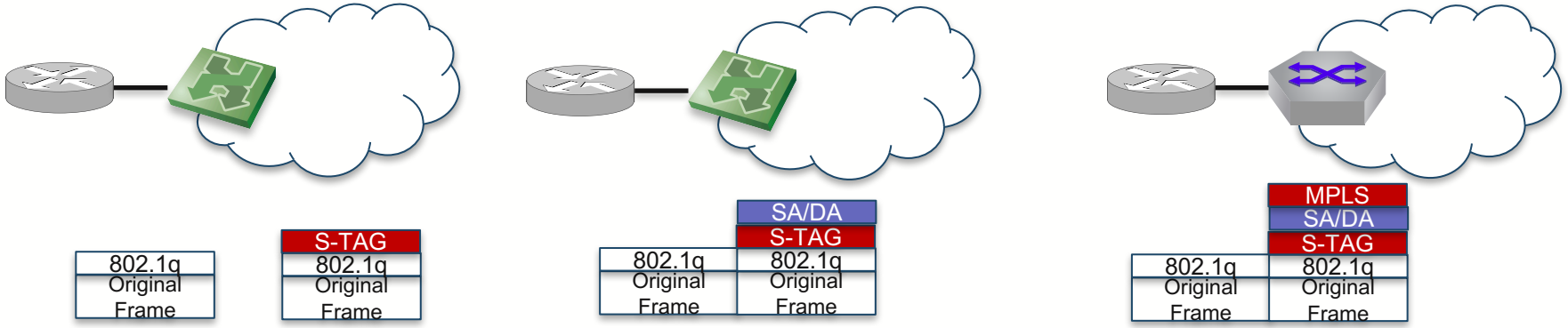




- IXPはBGPピアリングの為にレイヤー2の到達性を提供する
- レイヤー2技術と光スイッチを使用し、トポロジーは限られている
- BGPピアリングは IXPの顧客により行われる。もし片方のAS (AS64500)がIPアドレスを変えたり、そのサイトから切断された場合は他方のAS(AS64496)はBGPの為にARPやNDPを送り続ける。結果としてBUMトラフィックがネットワーク全体に波及する
- EVPNはActive-Acitiveマルチホーミングやダイナミックプロビジョニング/プロキシーARP/NDPによりこれらの問題を解決する
- draft-ietf-bess-evpn-proxy-arp-nd はIXPを含むProxy ARP/NDP使用例を記載している

ARISTA

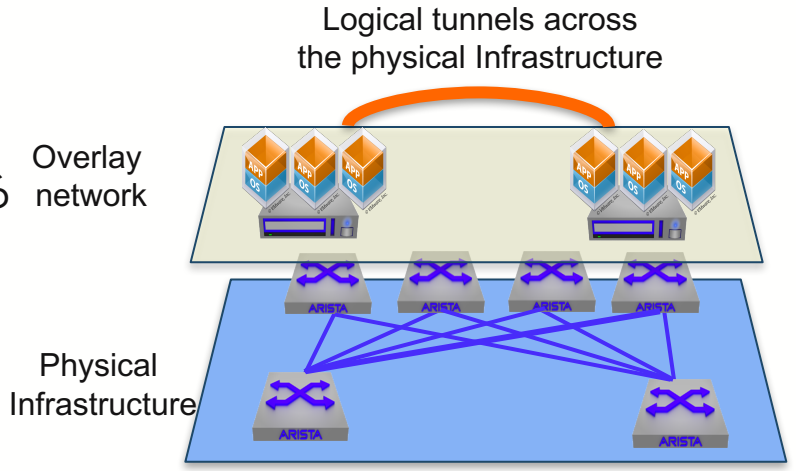
Business L2VPN services



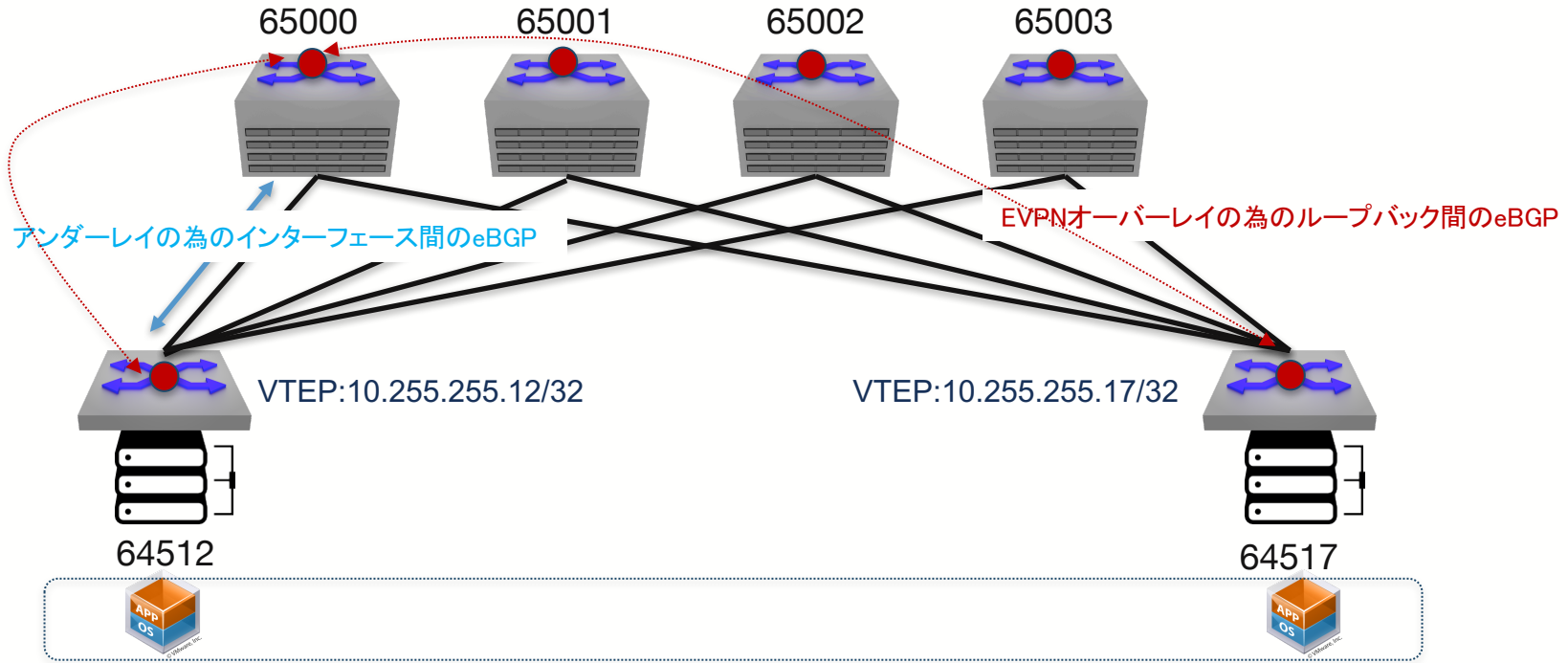
- 日本ではレイヤー2のイーサネットVPNサービスは2000年から続けている
- サービスプロバイダーはvlanヘッダーをスタックするExtreame vMAN(Ether type:0x9100) や Cisco QinQ(0x8100)を使用
- IEEEはその技術を 802.1ad Provider Bridges(0x88A8)として2006年に定義
- vMAN/QinQはvlan id(>4096)の制限を緩和させたが、安定はしていなかった
- 日立電線がEoE(Ether Over Ethernet)と呼ばれる新しいフレームフォーマットを開発した
- IEEEはその後2008年にこの技術を802.1ah Provider Backbone Bridgesとして定義する
- RFC7623 Provider Backbone Bridging はEthernet VPN (PBB-EVPN)でPBBフレームをMPLS上で転送する為の技術
- いくつかのサービスプロバイダーはここでVPLSを使っている

アンダーレイとオーバーレイのプロトコルの選択

- アンダーレイ:
  - 前述の通り、安定性とオープンなプロトコルを重視して、BGP
- オーバーレイ:
  - EVPNはコントロールプレーンとしてBGPを使用する
- BGP over BGP?



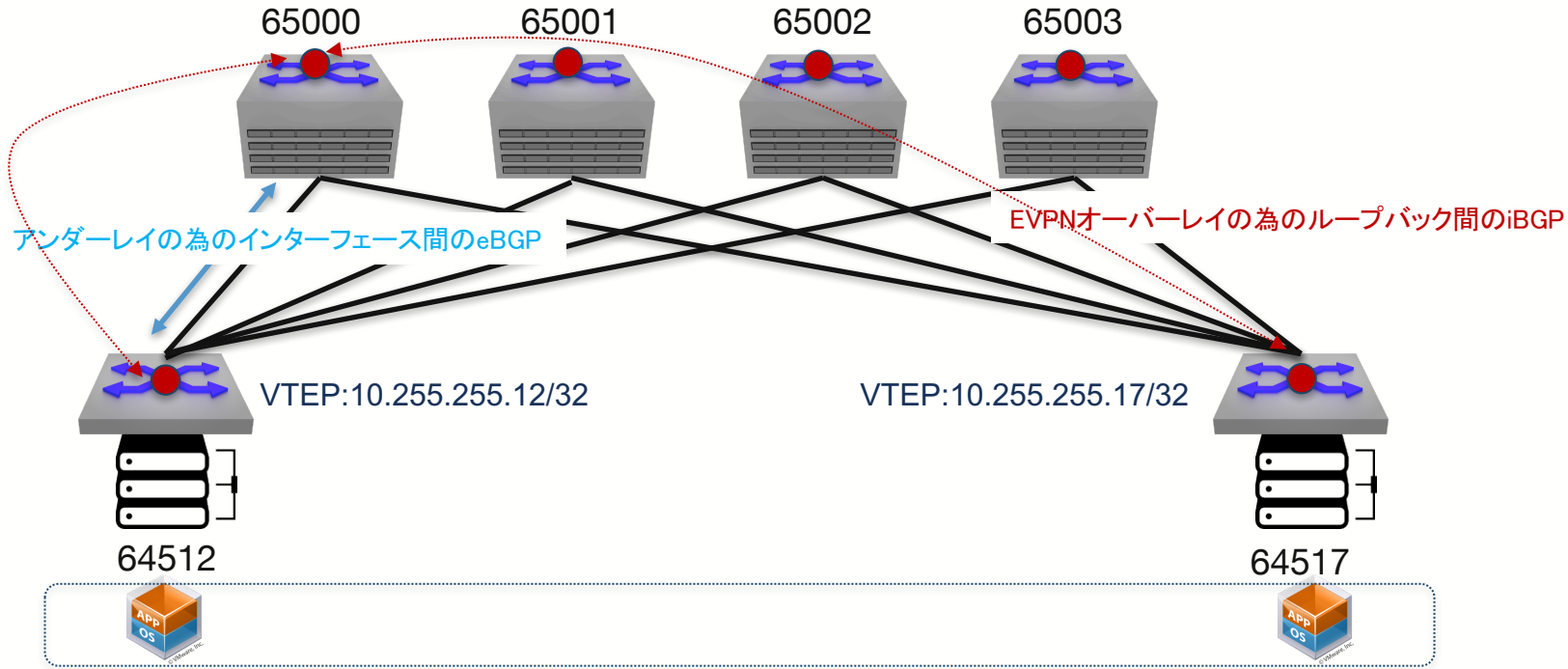
eBGP with Route Server(nexthop unchange)



- eBGPセッションをVTEPとSpineの間でEVPNの為に設定
- それぞれのSpineはEVPNコミュニティやnexthopアトリビュートを透過しないといけない
- つまりルートサーバーの機能が必要

ARISTA

iBGP with Route Reflector



- iBGPセッションを VTEPとSpine のloopbackでEVPNの為に設定
- SpineをRRとして設定し、local-asを使う

ARISTA

アジェンダ

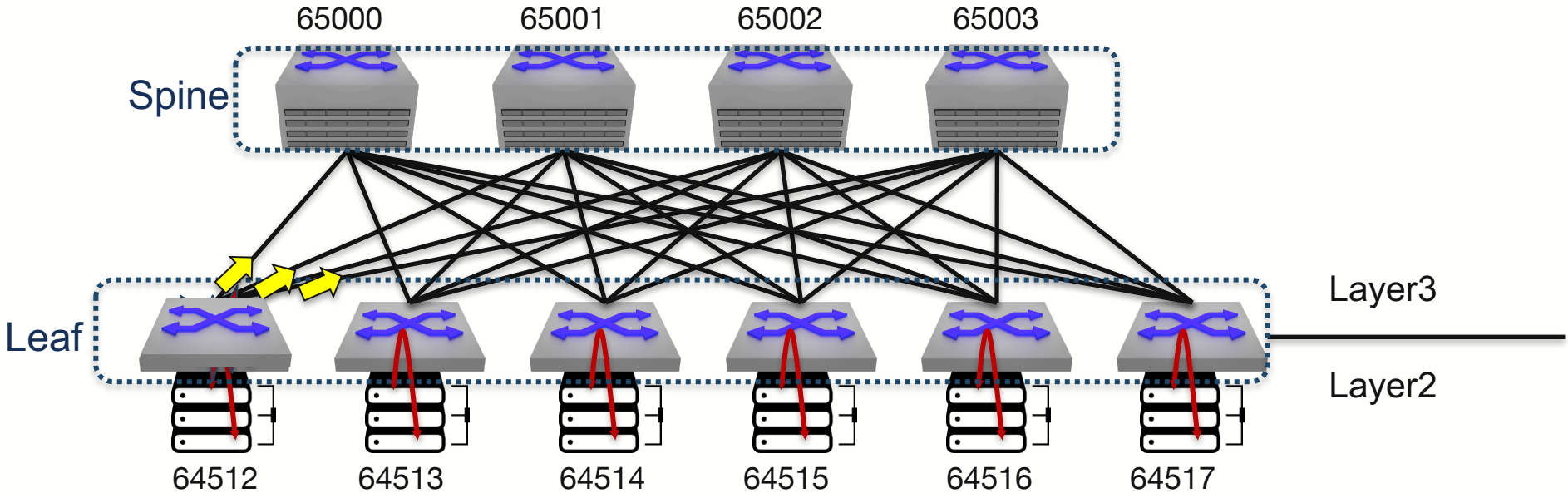
- 今のデータセンタールーティングが生まれてきた訳
- EVPNデータセンター内デザイン
- **ダイナミックディスカバリー**
- IETF Dcrouting

アンダーレイとしてのBGP

- コントロールのしやすさ・標準的なプロトコルでロックインされない
- 自動検出はZTPやSDNでカバー

- 最近は更にRFC5549機能拡張など

ZTP(Zero Touch Provisioning)



- 出荷状態において、全てのポートでDHCPクライアントとして動作
- LLDPの隣接情報などでそのまま正規設定をダウンロード
- インベントリシステムに登録後に正規設定をダウンロード

RFC5549 Capability Negotiation

```
+-------------------------------+
| Capability Code (1 octet:5 )  |
+-------------------------------+
| Capability Length (1 octet)   |
+-------------------------------+
| Capability Value (variable)   |
+-------------------------------+
```

```
+-----------------------------------------------------+
| NLRI AFI (2 octets)    1:IPv4                       |
+-----------------------------------------------------+
| NLRI SAFI(2 octets)    1:Unicast 2:Multicast        |
|                        4:Label 128:MPLS-VPN         |
+-----------------------------------------------------+
| Nexthop AFI(2 octets)  2:IPv6                       |
+-----------------------------------------------------+
```

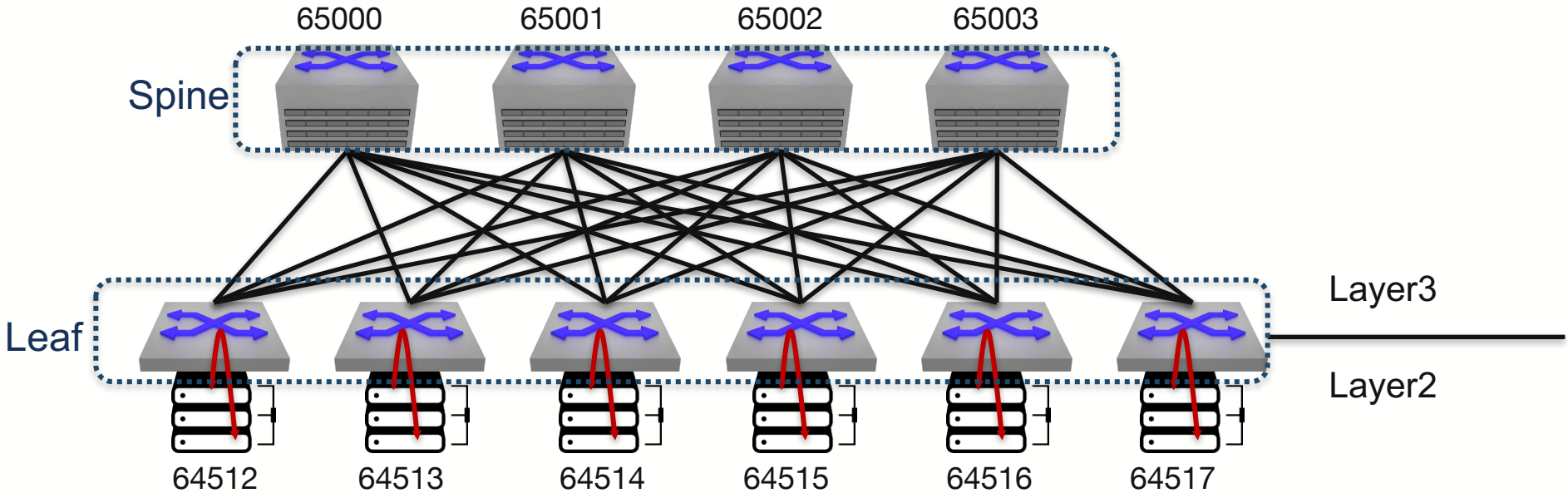
- RFC5492 のCapability Code 5 Extended Next Hop Encodingでネゴシエーション
- Valueにてサポート可能なAFI/SFIおよび Nexthop AFIを通知(IPv6:2)

RFC5549 UPDATE/WITHDRAWN

```
+---------------------------------------------------------+
| Address Family Identifier (2 octets)                    |
+---------------------------------------------------------+
| Subsequent Address Family Identifier (1 octet)          |
+---------------------------------------------------------+
| Length of Next Hop Network Address (1 octet)            |
+---------------------------------------------------------+
| Network Address of Next Hop (variable)                  |
+---------------------------------------------------------+
| Reserved (1 octet)                                      |
+---------------------------------------------------------+
| Network Layer Reachability Information (variable)       |
+---------------------------------------------------------+
```

- RFC4760 のMPBGPのMP_REACH_NLRIとMP_UNREACH_NLRIをそのまま使用可能

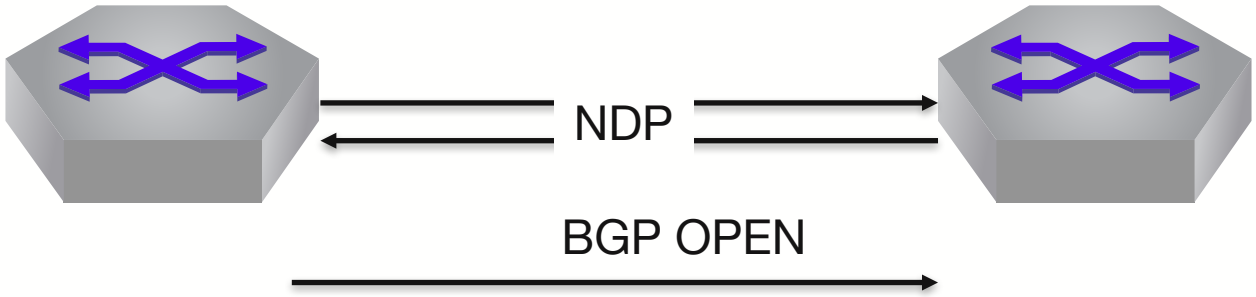
データセンターだと



- お？リンクローカル使いたくならない？？
- fe80::AS番号
- BGP4+ Peering Using IPv6 Link-local Address
  - https://tools.ietf.org/html/draft-kato-bgp-ipv6-link-local

更に拡張: BGP Link-Local Next Hop Capability

https://tools.ietf.org/html/draft-kumar-idr-link-local-nexthop



- RFC5492 のCapability Code 5 Extended Next Hop Encodingでネゴシエーション
- ＋LINK-LOCAL-ONLY-NEXT-HOP (IANA未定義)でBGP OPENメッセージ
- 繋げただけで自動設定が目標

アジェンダ

- 今のデータセンタールーティングが生まれてきた訳
- EVPNデータセンター内デザイン
- ダイナミックディスカバリー
- **IETF Dcrouting**

IETF100 DCROUTING

https://datatracker.ietf.org/meeting/100/session/dcrouting

- IETF100シンガポールにてDCROUTING(データセンタールーティングに関する)BOFを開催
  - 要求事項と問題定義
  - 新しいルーティングテクノロジーでの対応
  - 既存のルーティングテクノロジーでの拡張

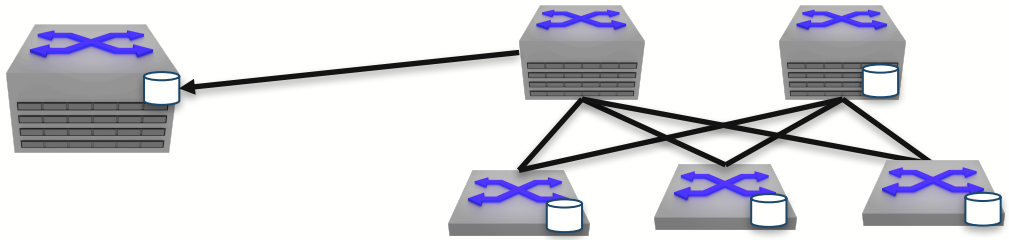
Requirements for the DataCenter Routing

https://tools.ietf.org/html/draft-dt-rtgwg-dcrouting-requirements

- 基本的な接続性と1つ以上のオーバーレイプロトコルを運ばなければならない
- Spine-Leafのトポロジーをサポートしなければならない
- 収束時間の仮のKPIは下記の通り
  - 5,000ノード/250Kルートで250msec以下
  - 7,500ノード/500Kルートで500msec以下
  - 10,000ノード/1Mルートで1秒以下
- ECMP/UCMPを使ったロードバランスをサポートしなければならない
- 様々のアドレスファミリーを運ばなければならない、また到達性以外のデータも運べる様に拡張可能でなれかbならない
- inbandのみでは無く、out band管理も出来なければならない
- ZTPをサポートしなければならない
- BFDをサポート出来なければいけない
- ステート情報をTelemetryで運べなければならない

Shortest Path Routing Extensions for BGP Protocol

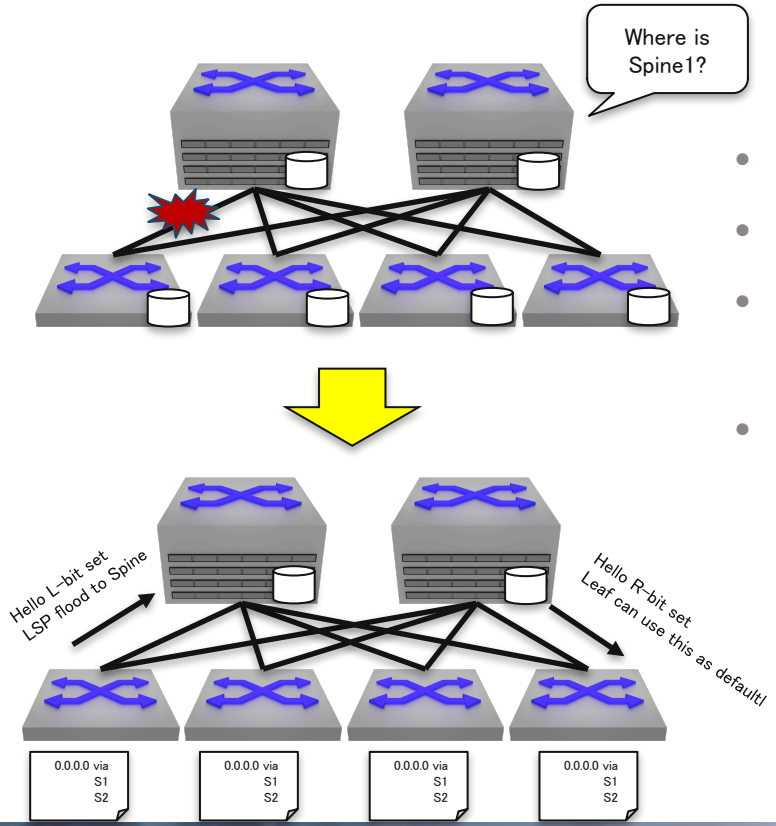
https://tools.ietf.org/html/draft-keyupate-idr-bgp-spf



- SRLGやLFAなどを管理するにはトポロジー全体が必要(LSDB)
- リンクステート情報をBGPで伝播する仕組みは既にある=>BGP-LS
- これを使ってBGPとDijkstraアルゴリズムを組み合わせ、パス選択アルゴリズムを変更
  1. NLRIを発行したものかどうかを確認。発行元のNLRIを優先
  2. 次にBGP-LSのメトリック情報を元にSPFを実施
  3. Router IDの最も高い値

IS-IS Routing for Spine-Leaf Topology

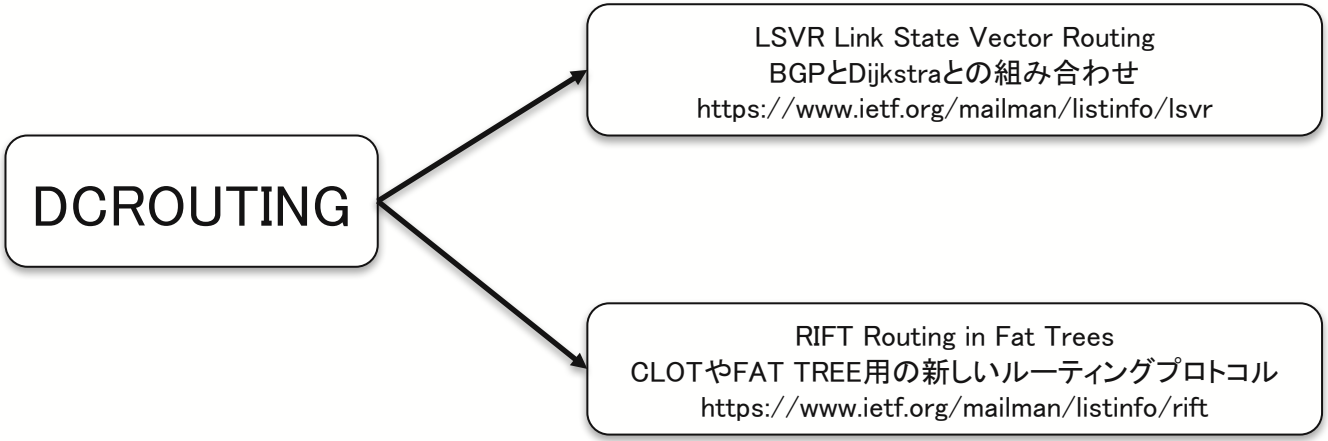
https://tools.ietf.org/html/draft-shen-isis-spine-leaf-ext



- 既存のISISのメカニズムをそのまま使用
- Spine-Leaf TLVを持たせる
- Spine間ではフルデータベースを持ち、Leafでは Spineでのデフォルトルートのみを持つ
- Leafでのスケーラビリティを持たし、障害時の影響範囲を狭める

ARISTA

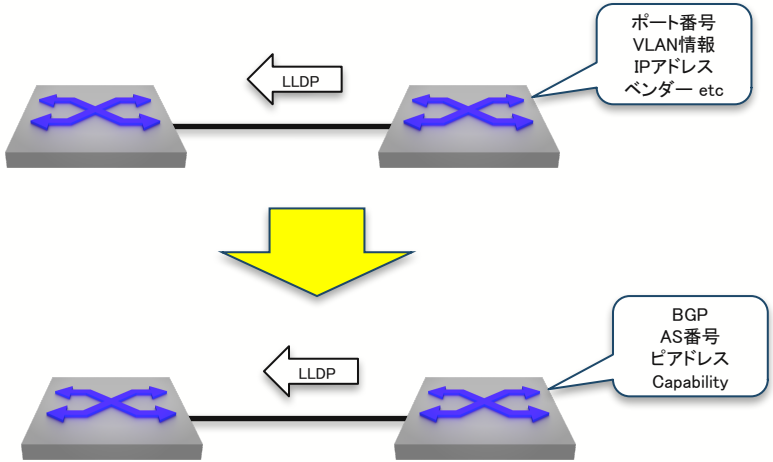
今後のデータセンタールーティング標準化



- 既存のパスベクタープロトコルにDijkstraアルゴリズムを組み合わせたLSVR
- データセンタートポロジー用に新しいルーティングプロトコルを使うRIFT
- それぞれがWGとして標準化および議論を進めていく

おまけ BGP LLDP Peer Discovery

https://tools.ietf.org/html/draft-acee-idr-lldp-peer-discovery



- LLDPでBGP設定に必要な情報も追加
- データセンターの様な1ホップのBGPピアに対して活用

まとめ

- データセンターのスケールによって選択肢は色々
- BGPやISIS(OSPFも)データセンター用のものに少しずつ変わるかも
  - 要求を提案するなら今です。
- EVPNのデザインはインターデータセンターが中心で合ったため、データセンター内ではちょっと癖があるかも

ARISTA