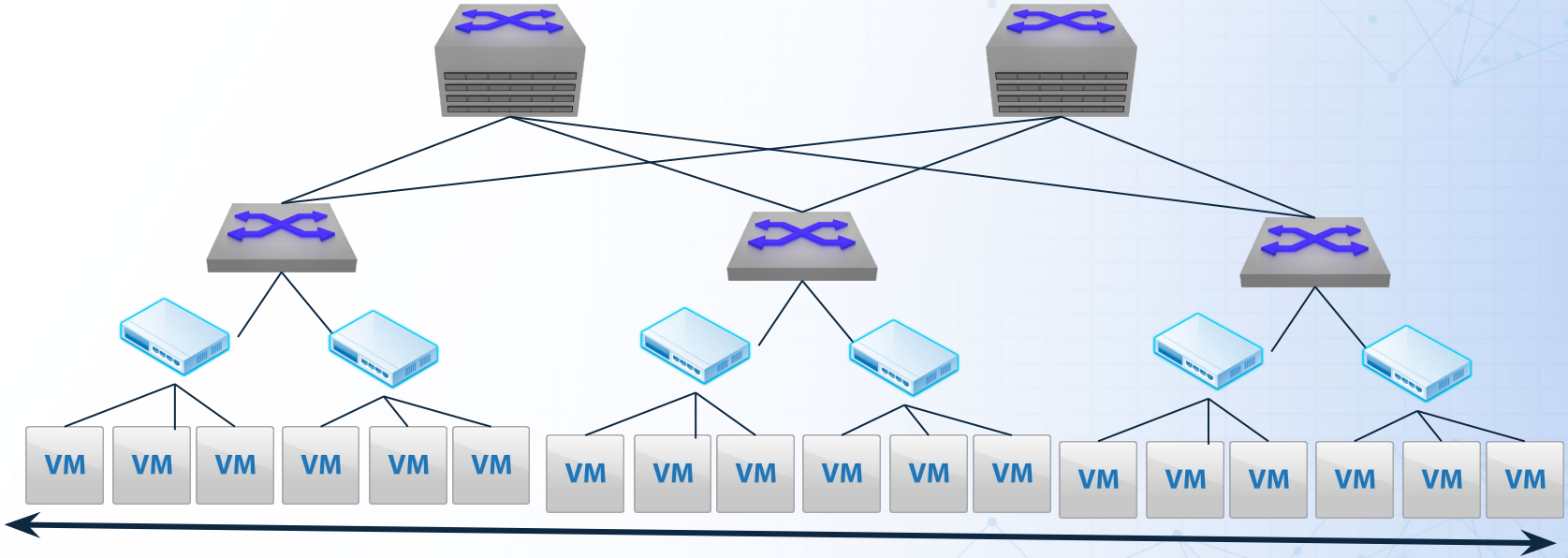
ARISTA

# 爆発するMACアドレスとの戦い

## EVPN VESPA/仮想イーサネットセグメントとProxy ARPの活用

Shishio Tsuchiya
shtsuchi@arista.com

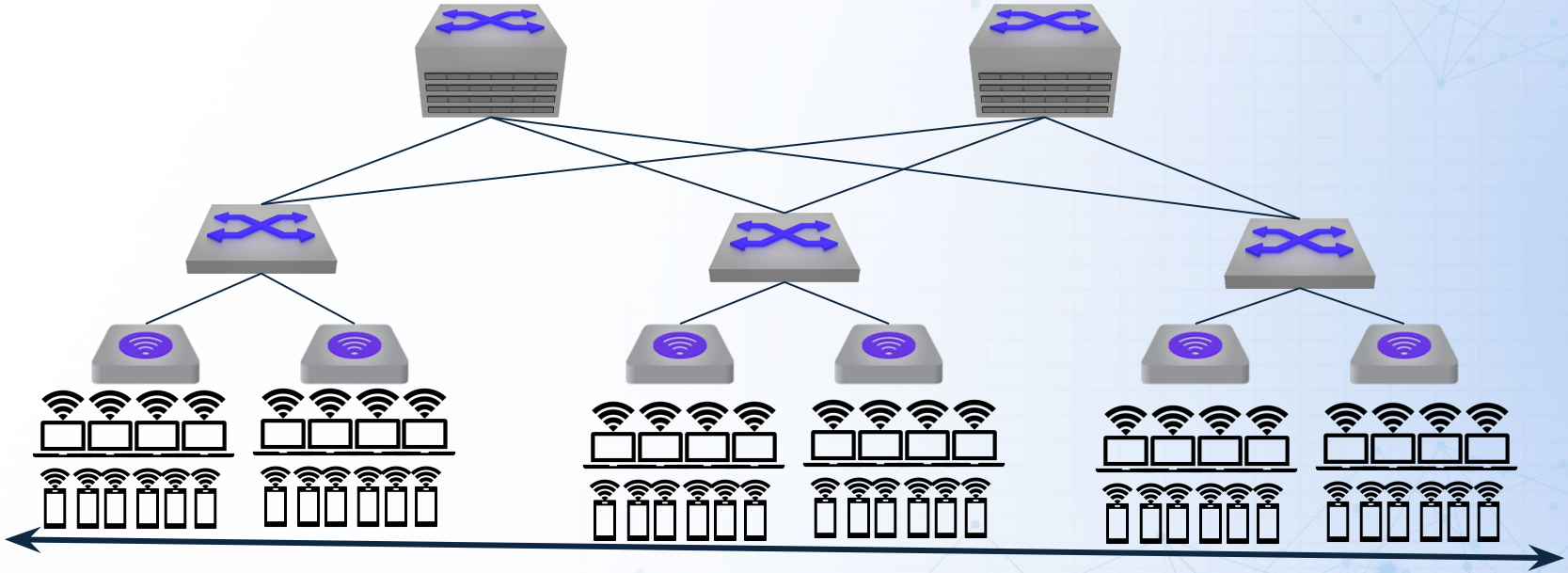
# データセンターでの MACアドレスの増加要因




- ムーアの法則でCPU/メモリのスペックは上がっていく1台のサーバーの集積密度は上がっていく
- 一方仮想マシンのサービス要求は変わらない
- コンテナなど仮想化の技術も増えてきた

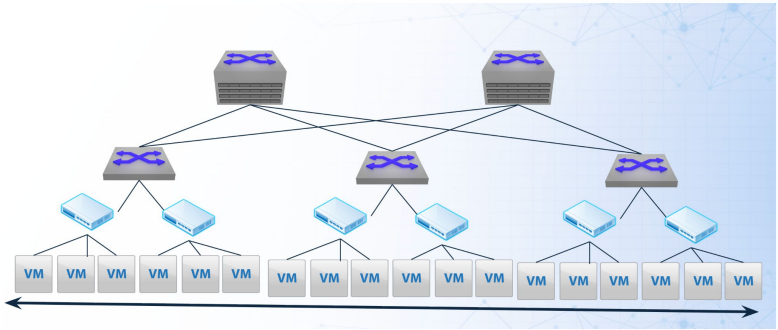
  

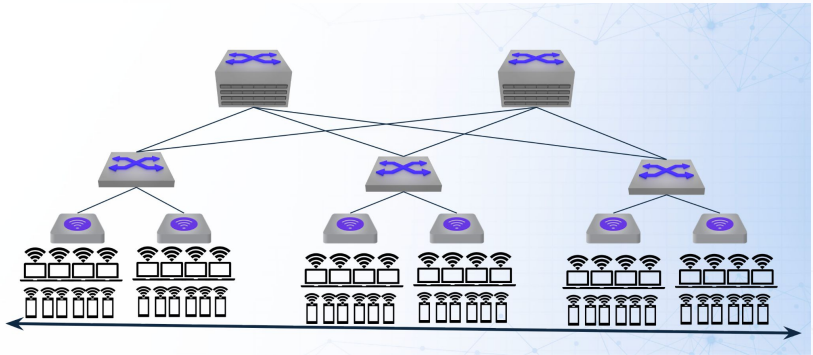
# 企業ネットワークでの MACアドレスの増加要因



- BYOD端末やノートPCのモビリティの確保
- MACアドレスのランダム化も要因の1つに!?

# 解決しなければいけないスケール



- クライアント数(VMやWi-Fi端末)
  - 500,000
- アクセスデバイス(vSwitchやAP)
  - 30,000
- レイヤー2ドメイン
  - 制限では無く自分で決める事が出来る

# こんな事がありました


🦁いうてDCと比べて少なくない？

🗿JANOGの会場ネットワークもMACアドレスがやばい

🦁あーーそういう事

🗿MACアドレスのランダム化って知ってる？

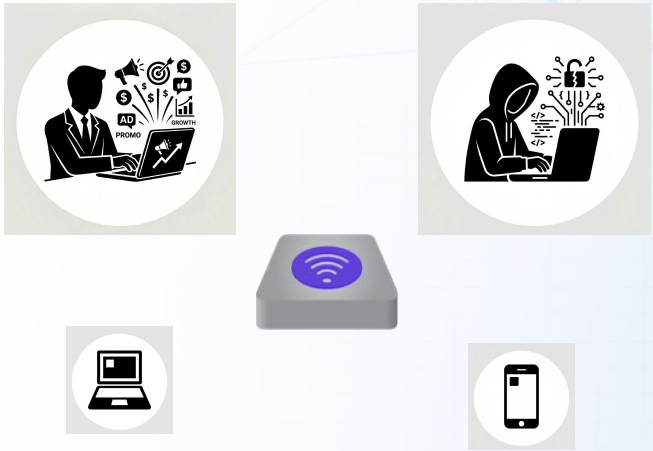
 





ARISTA

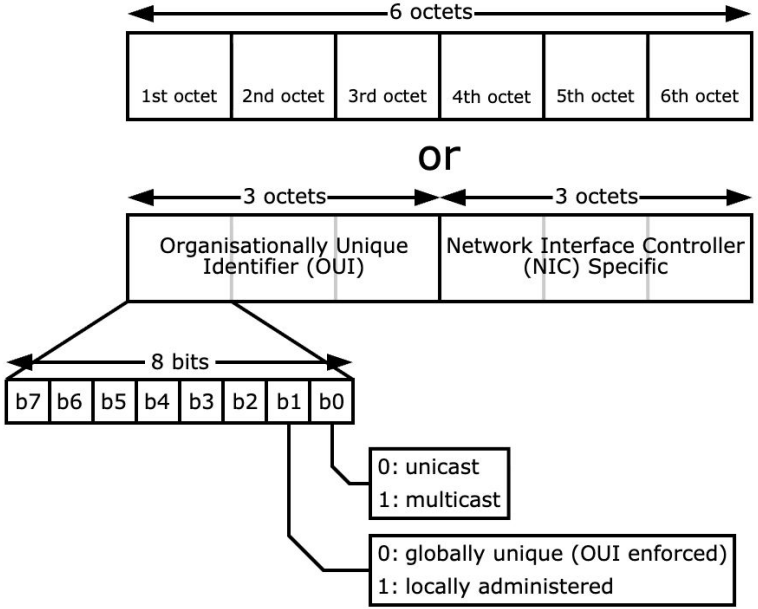
# MACアドレスランダム化の背景




- カフェやショッピングモールなどWi-Fi環境においてMACアドレスは収集され下記の様な目的に使用される
  - 現在の位置情報
  - 滞在時間
  - 滞在時間
- プライバシーの観点から各OSベンダーが積極的に採用
- 接続要求時のみでは無く、接続中もランダム化
- 通常運用にも影響が!?

  Public.

# どのようにランダム化されるか




- 48ビットのMACアドレスの7bit目はグローバルにユニークか、ローカルに管理されているかを示す
- 1が立っているとローカルに管理されたアドレスであることを示す。

# OSベンダー毎の MACアドレスランダム化サポート状況

| | OS毎の挙動 | | |
|---|---|---|---|
| 挙動 | iOS14 | Android 11 | Windows 10 |
| デフォルト | ON/ユーザはデフォルトの動作を変更出来ない | ON/ユーザはデフォルトの動作を変更出来ない | OFF/ユーザーはデフォルト設定変更可能 |
| 初めてのSSIDへの接続 | SSIDに初回接続時、接続用に新しいランダムMACを生成 | SSIDに初回接続時、接続用に新しいランダムMACを生成 | OFF:ハードウェアMACアドレスで接続<br>ON:SSIDに初回接続時、接続用に新しいランダムMACを生成 |
| 既存のSSIDへの接続 | 同じSSIDから切断し再接続する際、接続には同じランダムMACアドレスを使用 | 同じSSIDから切断し再接続する際、接続には同じランダムMACアドレスを使用 | ONの場合<br>同じSSIDから切断し再接続する際、接続には同じランダムMACアドレスを使用 |
| 特定のSSIDに対してのランダムMACの無効化 | ハードウェアMACで接続 | ハードウェアMACで接続 | ハードウェアMACで接続 |
| すべてのSSIDでのMACアドレスランダム化の無効 | N/A | N/A | ハードウェアMACアドレスを使用 |
| SSID プロファイル削除と再接続 | 同じランダムMACアドレスを使用 | 同じランダムMACアドレスを使用 | OFF/ハードウェアMACアドレスを使用<br>ON/新たに生成されたランダムMACが使用される |

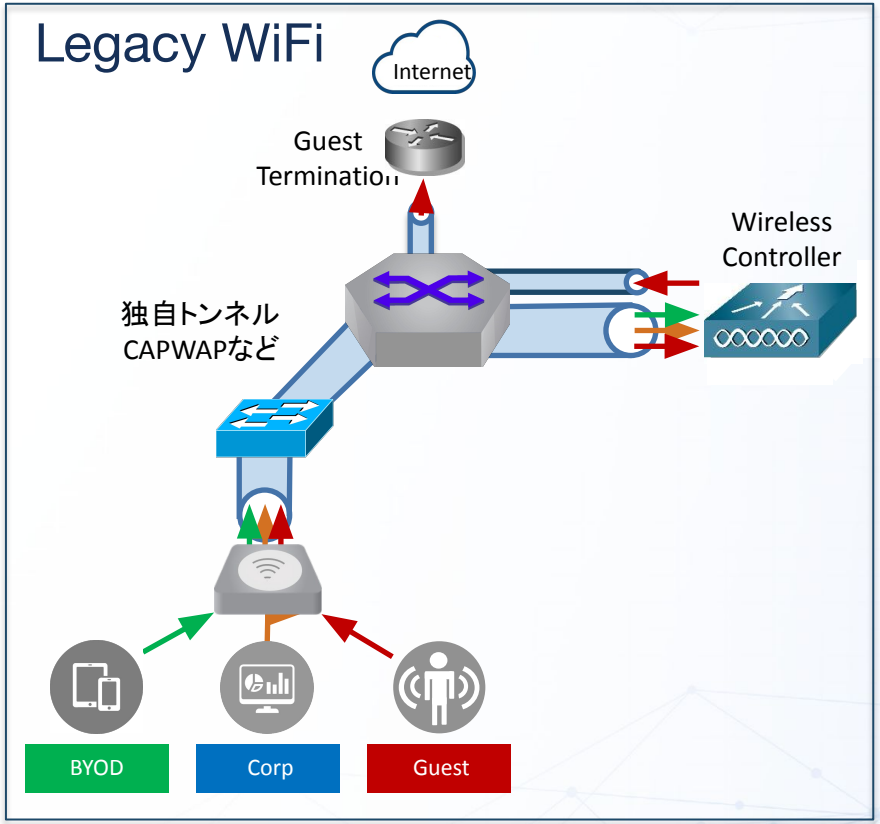
 

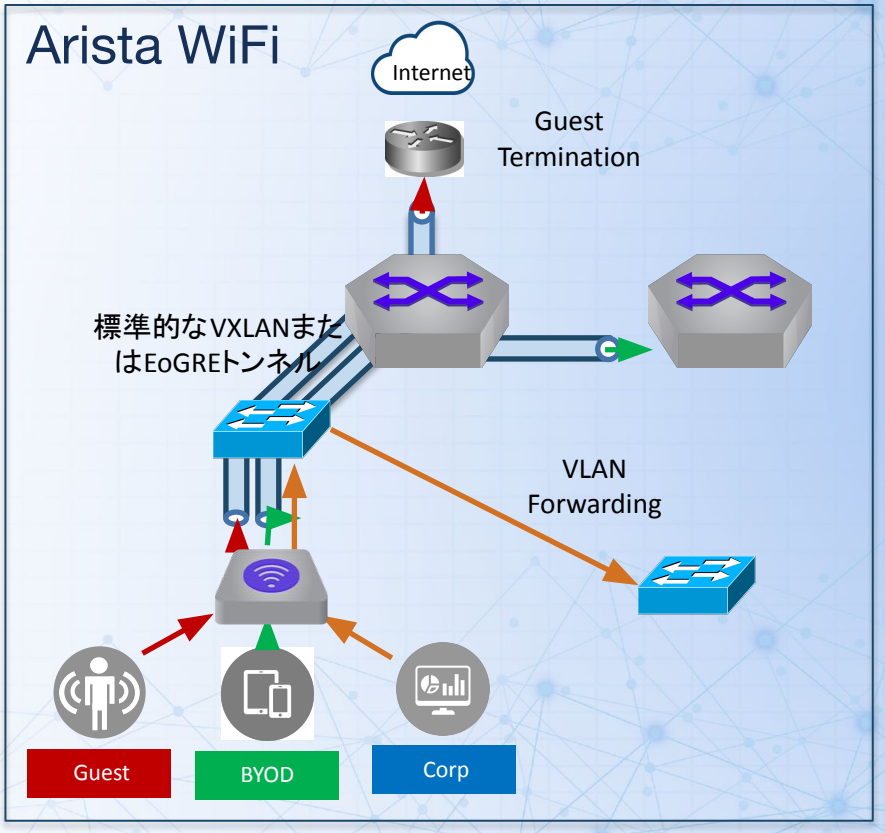
# インパクト

- iOS14にアップデート後に最初にSSIDにログインするとゲストWiFiなどの情報はリセットされる
- 更に最新のOS(Windows11/iOS18/macOS15では接続時間やネットワークの安全性など特定条件において同じSSIDでもMACアドレスのランダム化が行われる(Rotating Wi-Fi Address)
- 下記の様なネットワークの機能に影響がある
  - MACアドレス認証やPortセキュリティやMAC ACLなどのセキュリティ
  - ユーザ情報分析
  - DHCP IP アドレスの消費とサーバー負荷
  - スケーラビリティなどにインパクトが起こり得る

# コントローラーレス WiFiアーキテクチャー



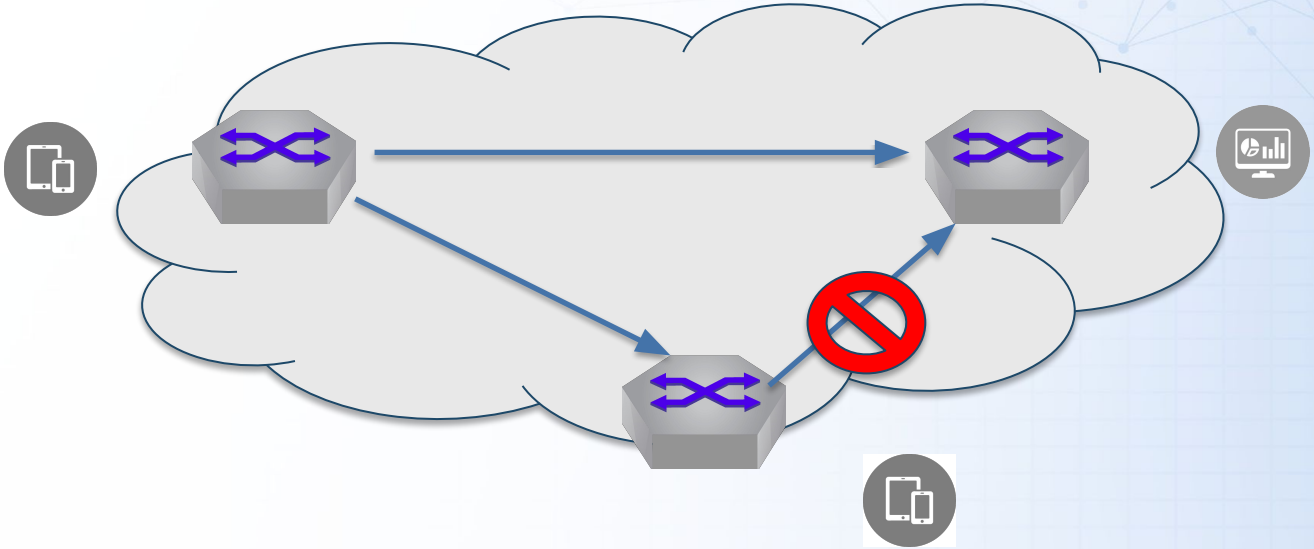





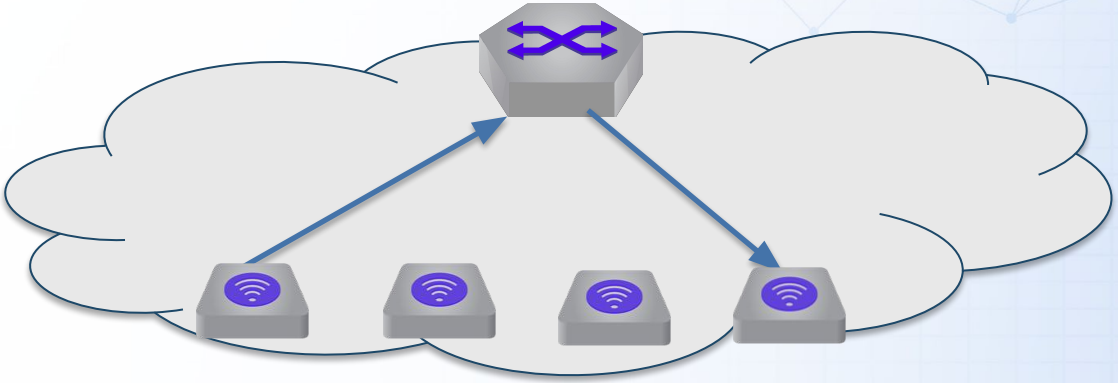
ARISTA

# 通常のvxlanの場合



- すべてのvxlanのエンドポイントはフルメッシュでつながっている
- 他VTEPからフラッドされたトラフィックを別なVTEPにフラッドする事は無い
- これでルーティングループを防ぐ

# VXLANをHub & Spokeとして動作させる



- APとの組み合わせでは従来のVXLANとは違う動作が求められる
- “vxlan bridging vtep-to-vtep” をVTEPに設定する

# AP側の設定

https://arista.my.site.com/AristaCommunity/s/article/VXLAN-Configuration

- SSIDとトンネルが紐づく
- StaticにVTEPを設定



# 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

### HER with CloudVision eXchange (CVX)

- ローカルに学習された MACとVNIバインドはCVXに公開される
- CVXはダイナミックにリモート VTEPにステートを分配する
- サードパーティ VTEPをサポート
- MAC学習とfloodlistの設定は自動にされる
- 信頼性の為に HAクラスタサポート

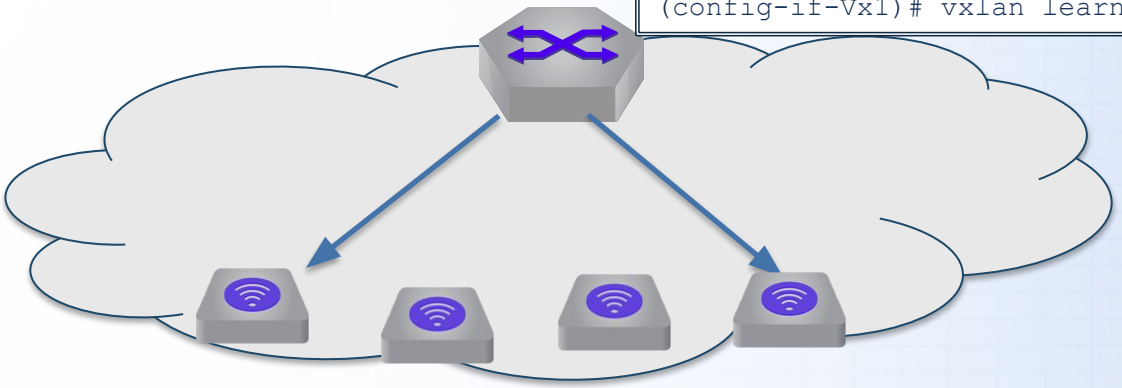
### EVPNモデル Model

- 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

# 従来の方法の Flood-listではAPは管理出来ない

https://www.arista.com/en/support/toi/eos-4-23-0f/14368-vxlan-auto-flood-list-construction

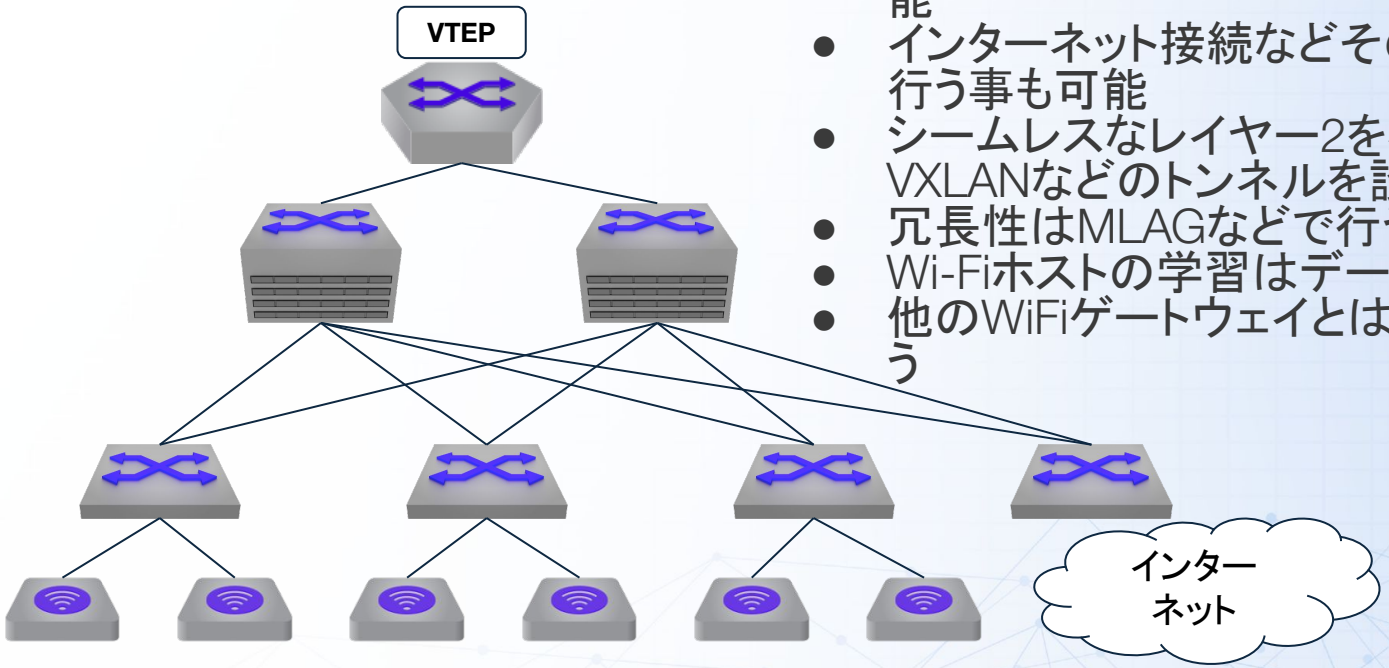
```
(config)# interface Vxlan 1
(config-if-Vx1)# vxlan flood vtep learned data-plane
(config-if-Vx1)# vxlan learn-restrict vtep 10.1.1.0/24
```



- 実際にパケットを送ってきた(データプレーン)VTEPをflood-listに登録する
- アドレス範囲を制限する事も可能

# 現在のコントローラーレス WiFiアーキテクチャー




- 各SSIDでそれぞれのデータプレーンを選択可能
- インターネット接続などそのままルーティングで行う事も可能
- シームレスなレイヤー2を構築する為にはVXLANなどのトンネルを設定する
- 冗長性はMLAGなどで行う
- Wi-Fiホストの学習はデータプレーンで行う
- 他のWiFiゲートウェイとはEVPNでやり取りを行う



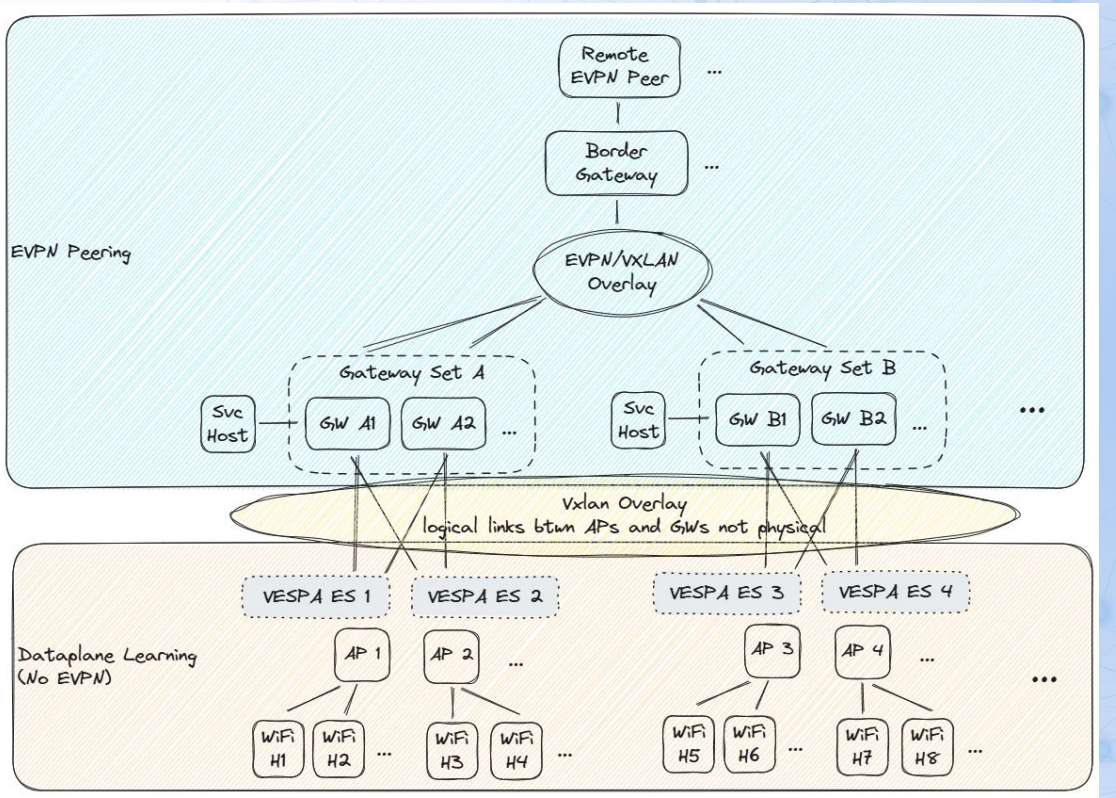
16

# 既存の構成の問題点

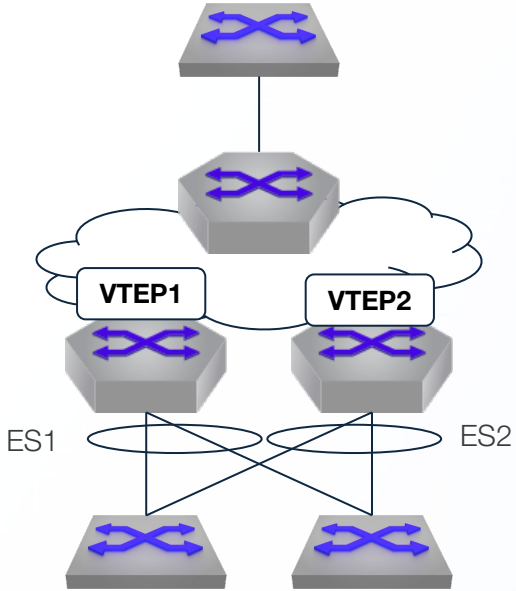
- 冗長性の欠如
  - WiFiゲートウェイが単一障害ポイントとなり得る
  - プライマリーとセカンダリーの設定
    - ICMPを定期的にポーリングし、プライマリーが切れた時はセカンダリーに切り替わる。
- 解決策
  - MLAG
    - 独自プロトコルとなる
    - 地理的制限がある
    - 2台のみの冗長となる
  - EVPN All Active Multihoming
    - EVPN マルチホーミングのイーサネットセグメントはStaticに設定される。APの数が多いほど、Static設定の量も増える
    - APはゲートウェイに直接接続されておらず、アンダーレイネットワークを経由したVXLANトンネルを介して接続されている

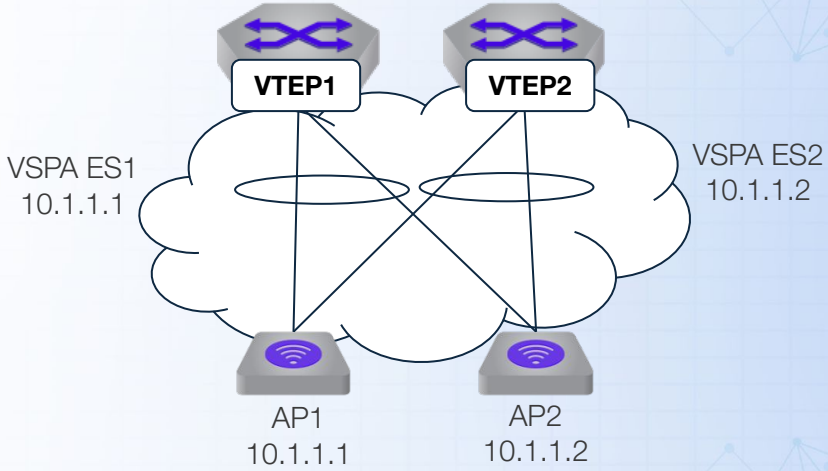
# VESPA(Virtual Ethernet Segment with Proxy Arp)

- VESPAは、本質的に通常のEVPNマルチホーミングの拡張機能
- ローカルインターフェースに加えてトンネルにも対応し、イーサネットセグメントの動的な構築を可能に
- MACアドレスのスケーラビリティはAPがL2 Proxyとして動作する事により実現する



# EVPN VESPA(Virtual Ethernet Segment with Proxy Arp)

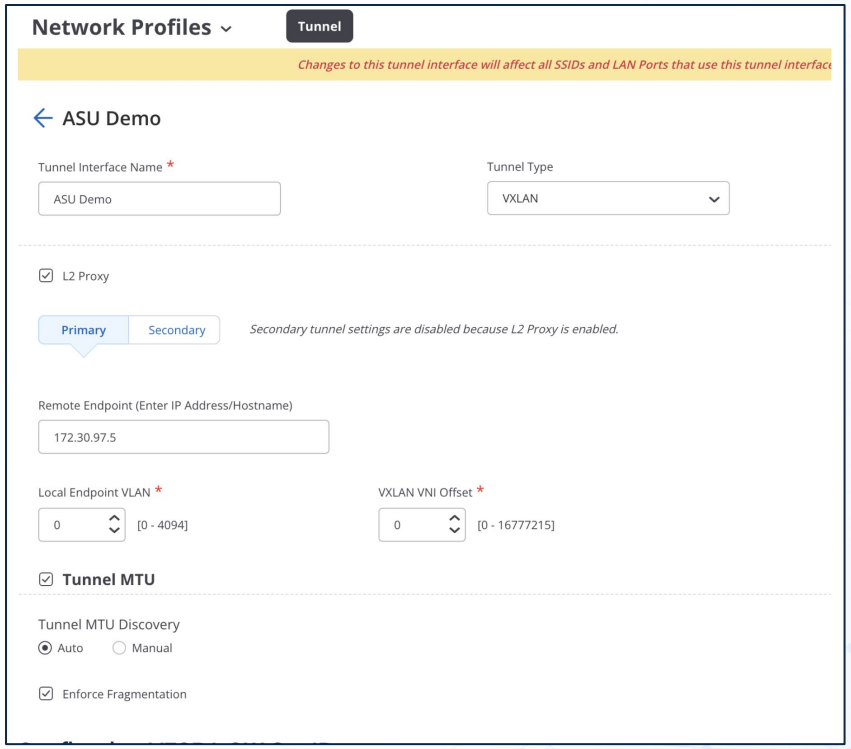


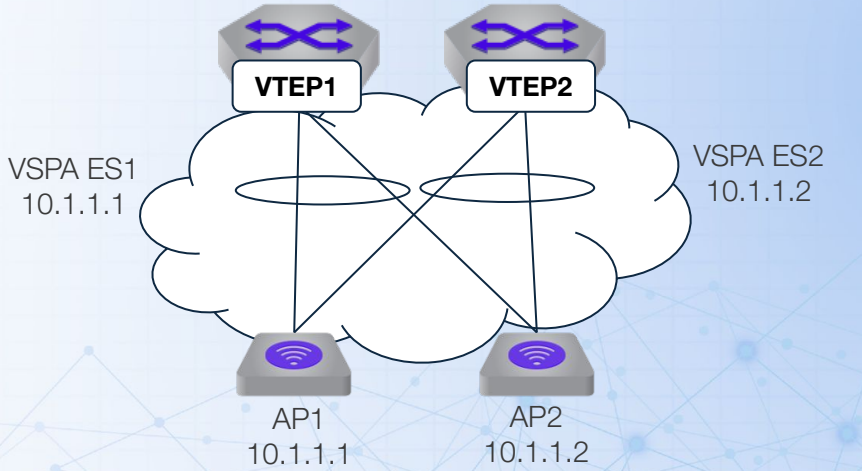

- 通常のEVPN Ethernet Segment
  - セグメントは物理インターフェースに紐づく
- EVPN VESPA(Virtual Ethernet Segment with Proxy Arp)
  - セグメントはトンネルに紐づく(VXLAN VTEP)

# L2 Proxy(MRO:MAC Rewrite Offload)を有効にする



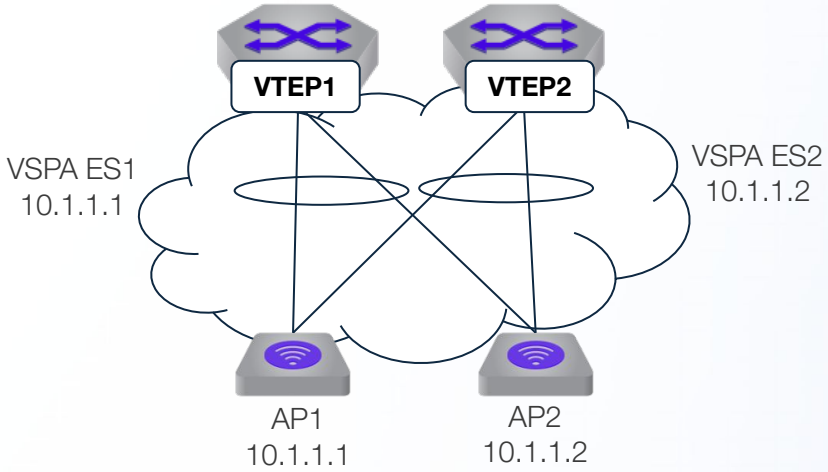

- Network Profiles → VxLAN Tunnel TypeでL2 Proxyにチェックする






20

# VESPA GW Set IDの設定

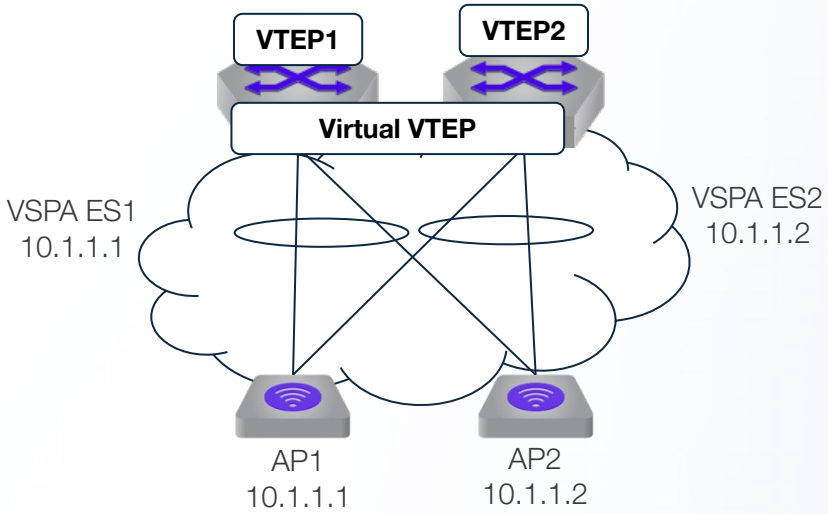



- VESPAゲートウェイにはSet IDを設定する必要がある。
- 接続されているAPに関して、互いに同じイーサネットセグメントに属する必要があるすべてのVESPAゲートウェイには、同じセットIDを設定する必要がある

```
router bgp 65000
   address-family evpn
      evpn ethernet-segment dynamic-tunnel mh-set-id
10
```

# VESPA GW でのVirtsual VTEP の設定

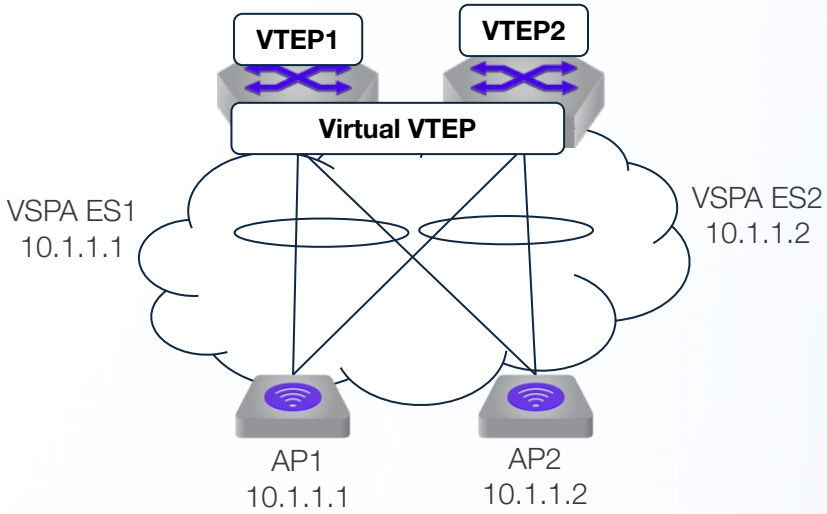



```
VESPA-GW1(config)#interface loopback 1
VESPA-GW1(config-if-Lo1)#    description VXLAN VTEP
VESPA-GW1(config-if-Lo1)#    ip address 10.255.255.1/32
VESPA-GW1(config-if-Lo1)#interface loopback 2
VESPA-GW1(config-if-Lo2)#  description VXLAN VIRTUAL VTEP
VESPA-GW1(config-if-Lo2)#    ip address 10.255.255.250/32
VESPA-GW1(config-if-Lo2)#interface vxlan 1
VESPA-GW1(config-if-Vx1)#    vxlan source-interface Loopback1
VESPA-GW1(config-if-Vx1)#    vxlan virtual-vtep local-interface
Loopback2
```

```
VESPA-GW2(config)#interface loopback 1
VESPA-GW2(config-if-Lo1)#    description VXLAN VTEP
VESPA-GW2(config-if-Lo1)#    ip address10.255.255.2/32
VESPA-GW2(config-if-Lo1)#interface loopback 2
VESPA-GW2(config-if-Lo2)#  description VXLAN VIRTUAL VTEP
VESPA-GW2(config-if-Lo2)#    ip address 10.255.255.250/32
VESPA-GW2(config-if-Lo2)#interface vxlan 1
VESPA-GW2(config-if-Vx1)#    vxlan source-interface Loopback1
VESPA-GW2(config-if-Vx1)#    vxlan virtual-vtep local-interface
Loopback2
```

- GWでは2つのVTEP IPを設定する必要がある。
- VXLANの送信元インターフェースとして使用するアドレス(loopback 1)
- virtual-vtep interface アクセスポイントがトンネルエンドポイントとして設定するVESPA Anycast GWアドレス(loopback 2)

# SVIの設定



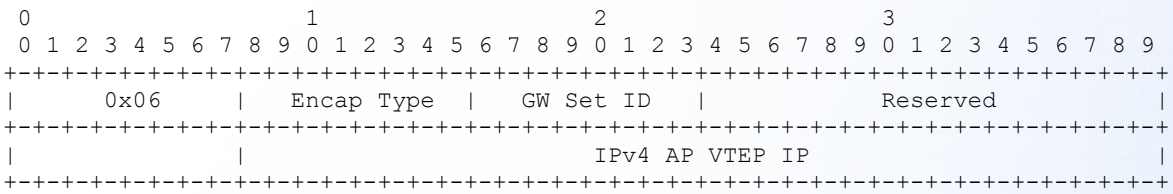

```
interface Vlan3525
   description USER_VRF_A
   mtu 9090
   no autostate
   vrf USER-3520
   arp aging timeout 3600
   arp refresh retries unicast 5 broadcast
0
   ip helper-address 192.168.21.2
   ip address virtual 10.1.1.250/16
```

- アクセスポイントは、そこに接続されているすべてのワイヤレスクライアントに対して、レイヤ2-MACプロキシとして機能
- VESPA GWはこれらすべてのワイヤレスクライアントのARPエントリを維持する必要がある
- ワイヤレスクライアントのARPエントリは、APが能動的にGARPパケットを送信することに基づいて作成
- GWは、エージェントベースの定期的なユニキャストリフレッシュを使用して、これらのARPエントリを維持する



23

# ESIタイプフォーマット




- RFC7432では5つのESIフォーマットが定義
- VESPAではType6を使用

| Type | 名前 | フォーマット | モード |
|---|---|---|---|
| 0 | User | 手動設定 | 手動設定 |
| 1 | LACP | LACP System MAC + port key + 00 | 自動生成 |
| 2 | Bridged | Root bridge MAC + priority key + 00 | 自動生成 |
| 3 | MAC | System MAC + Local Discriminator | 手動設定/自動生成 |
| 4 | Router-ID | RouterID + Local Discriminator + 00 | 手動設定/自動生成 |
| 5 | ASN | AS番号 + Local Discriminator + 00 | 手動設定/自動生成 |

# まとめ

- 大規模のレイヤー2ネットワークをサポートするEVPNを使った VESPA(Virtual Ethernet Segment with Proxy Arp)を紹介
- 大規模なMACアドレス/不安定なMACアドレスに対して安定したネットワークを提供



25

# 聞きたい事

- MACアドレス爆発してます？
- それはWiFiネットワーク？データセンター？
- MACアドレスのランダム化で困った事ある？
- L2 Proxy議論された事もありましたが、やってたりする人います？
  OpenFlow的ななにかで？
- 今どうですかね？



26

# 参考

- Mega Scale Campus Mobility Powered by Arista VESPA
  - https://www.youtube.com/watch?v=FUriyTCSmUU
- Arista VESPA User Guide
  - https://www.arista.com/en/support/toi/eos-4-35-2f/23763-arista-vespa-user-guide
- RFC7432 BGP MPLS-Based Ethernet VPN 5.  Ethernet Segment
  - https://datatracker.ietf.org/doc/html/rfc7432#section-5
- The Scalable Address Resolution Protocol (SARP) for Large Data Centers
  - https://datatracker.ietf.org/doc/html/rfc7586
- Address Resolution Problems in Large Data Center Networks
  - https://datatracker.ietf.org/doc/html/rfc6820
- MAC Randomization: Behavior and Impact
  - https://www.arista.com/assets/data/pdf/Whitepapers/MAC-Randomization-Behavior-and-Impact.pdf

ARISTA

# Thank You

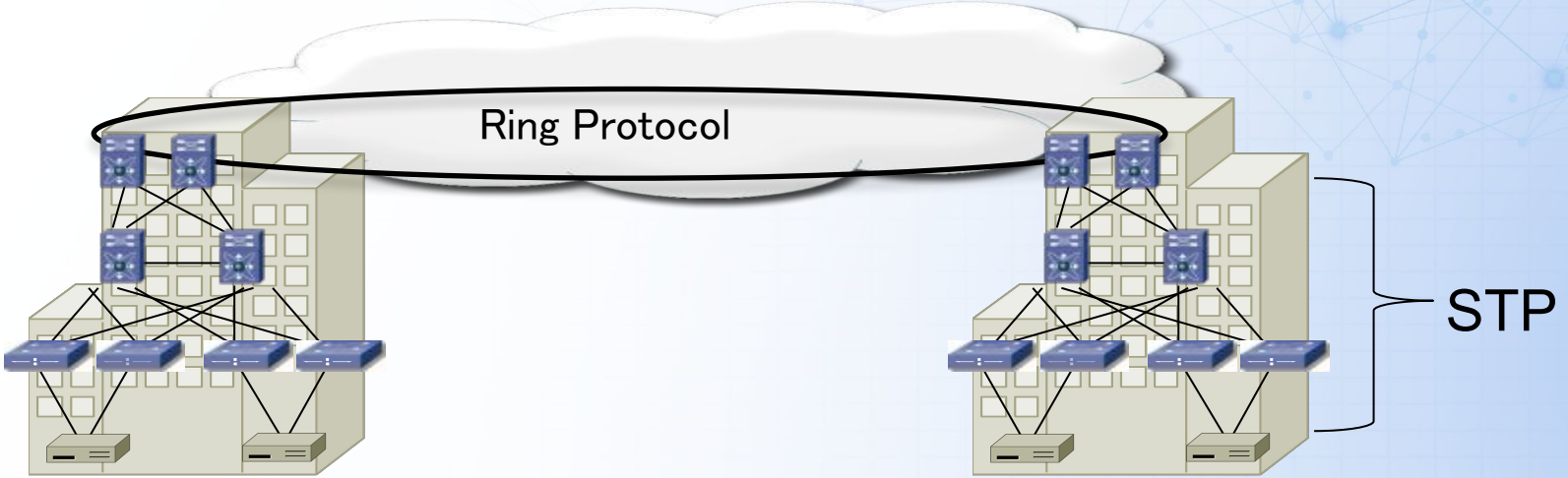
www.arista.com

# 技術の成熟と次代への継承

MACアドレス爆発解決策の光と影

ARISTA

# 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

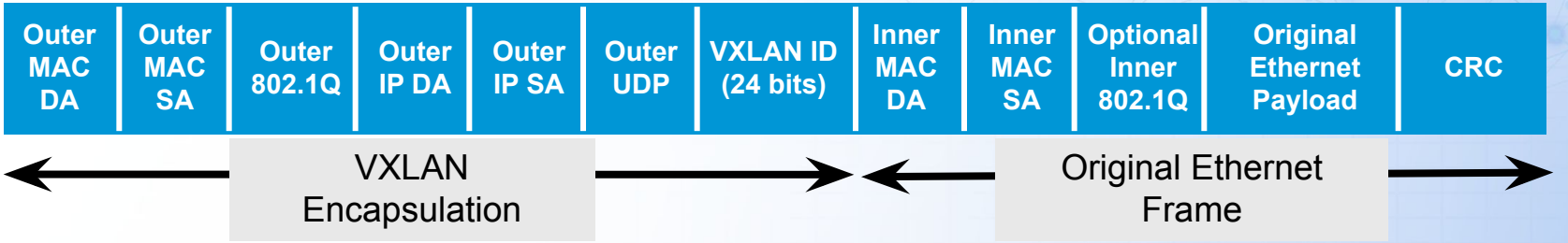
# 問題に関するソリューション

- 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



33

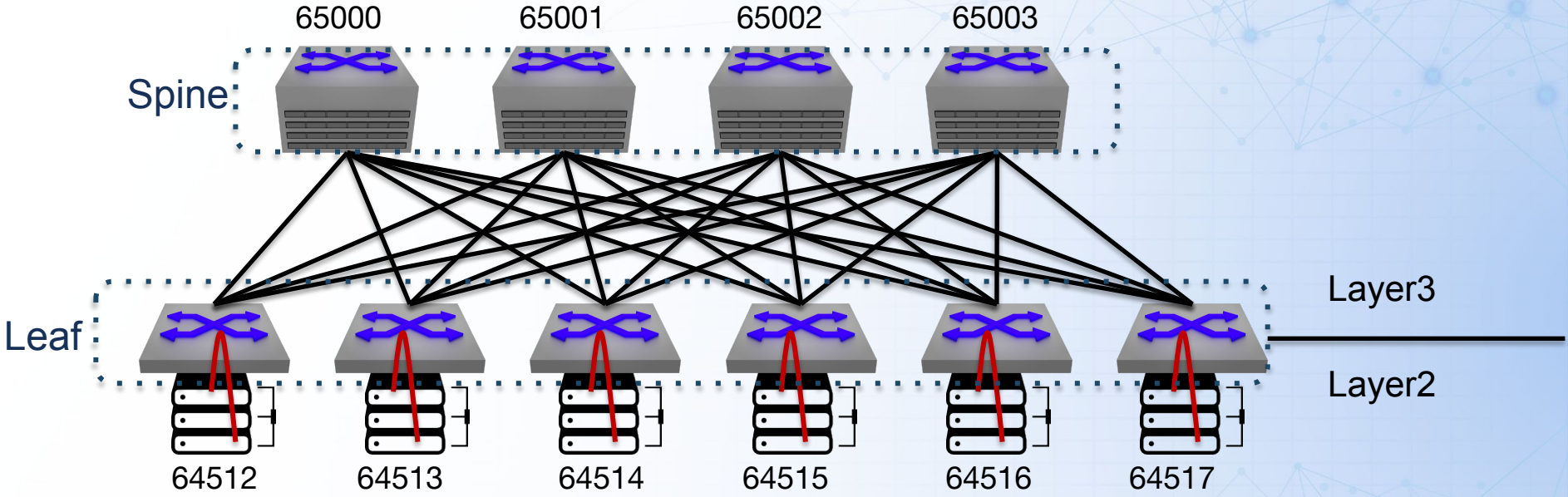
# Virtual eXtensible LAN (VXLAN)




- 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

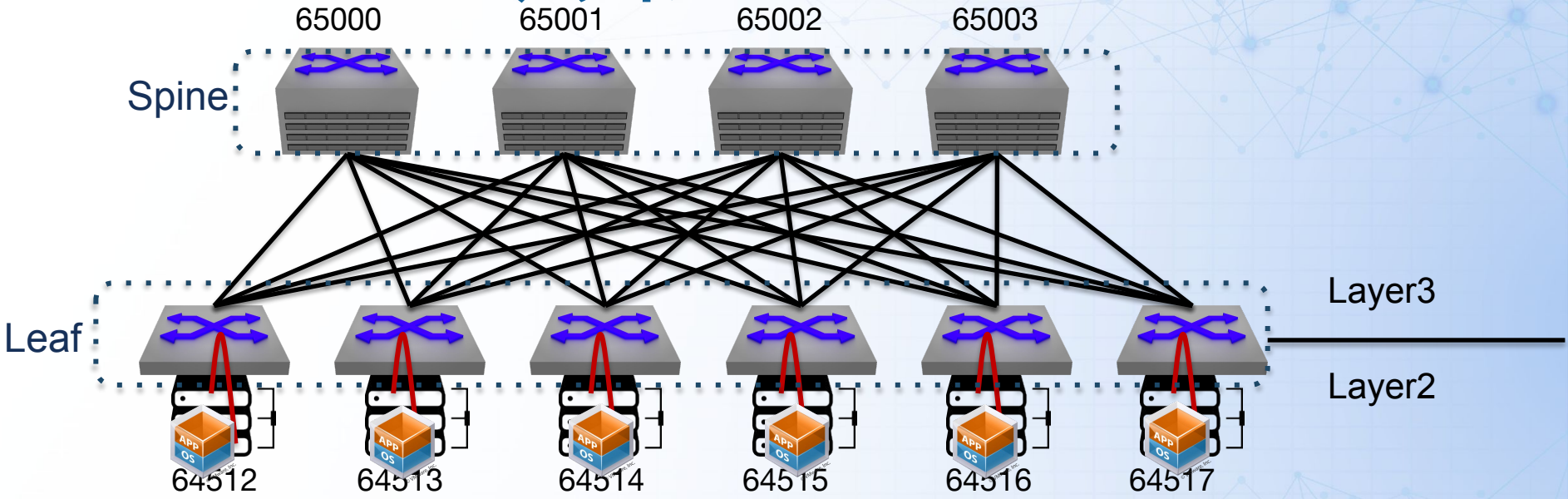
# 大規模データセンタールーティングでの BGPの使用




- 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

  Public. 

# VXLAN + IP Closデザイン

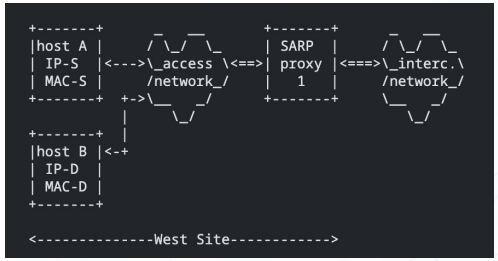


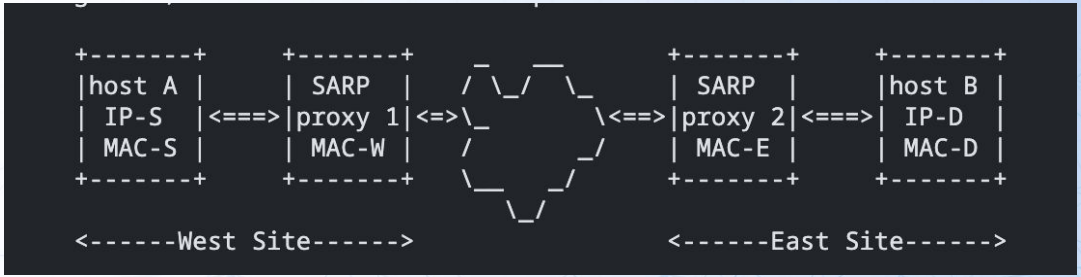

- データセンターデザインで完全なソリューション
- 標準で安定したBGPプロトコル/広く展開されたVXLANフォーマットを使用
- 特に特別な要求が中間ノードには存在しない

  Public. 

# 2011年のMACアドレス爆発

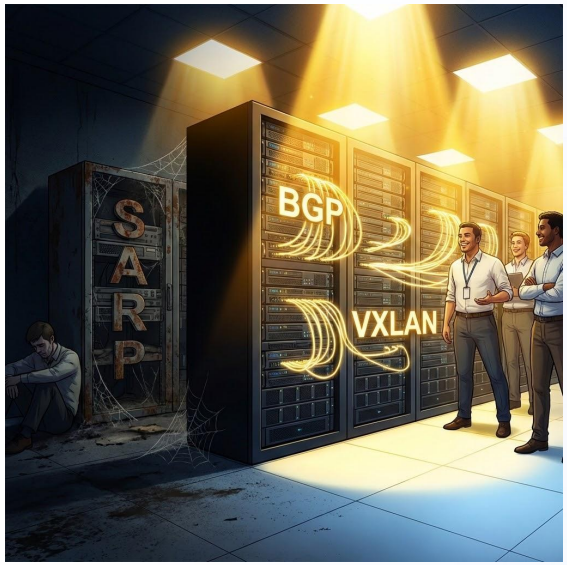
- BGPのデータセンターネットワークでの活用デザイン+プロトコル標準では無いVXLAN(業界標準)のカプセル化での解決
- つまり標準化IETFは課題定義のみを行いなにも出来なかった
- しかし1つのExperimental RFCを発行していた
- The Scalable Address Resolution Protocol (SARP) for Large Data Centers
  - https://datatracker.ietf.org/doc/html/rfc7586



# 2026年のMACアドレス爆発

## VXLAN技術の成熟とSARPの継承





- 2011-12年に脚光を浴びたVXLANとBGP
- BGPの拡張EVPNをコントロールプレーンに引っ提げて
- あの時光を浴びなかったL2 Proxyの技術にも注目そして実装