レイヤ2の仕組み、理解してみませんか？

2026年2月11日

自己紹介

佐藤　克賢 (さとう　かつたか)

- 所属：
  NTT東日本　先端テクノロジー部
  法人向けサービス用の基盤ネットワーク開発に従事

- JANOGとの関わり：
  JANOG53から毎回現地参加
  JANOG55ではプログラム委員として参加

チュートリアルの構成

1. 導入編
   - なぜレイヤ2について学ぶか？
   - 本セッションのゴール
2. 入門編
   - フレーム構造
   - フレーム転送
   - フレームの種類
   - ブロードキャストストーム
3. 基礎編
   - VLAN
   - STP
   - LAG
4. 応用編
   - トラブル事例
5. 発展編
   - EVPNの概要
   - EVPNのフレーム転送
   - EVPNのマルチホーミング
   - EVPN/(SR)MPLS、EVPN/VXLAN

導入編

なぜレイヤ2について学ぶか？

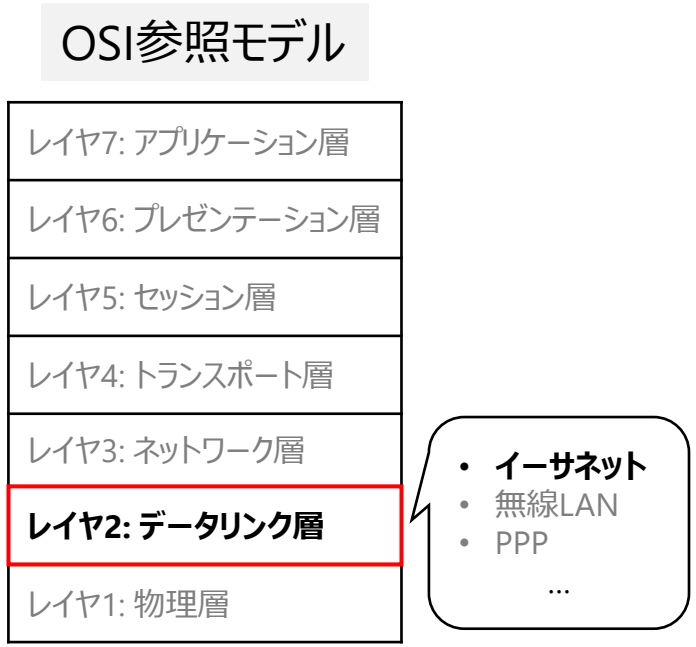
レイヤ2(データリンク層)は、OSI参照モデルの第2層に位置し、
隣接する機器間のデータ転送を担っています。

本セッションでは、データリンク層を代表するプロトコルである**「イーサネット」**を取り上げます。
世界で最も普及しているプロトコルの1つで、TCP/IP通信を支える不可欠な基盤です。

仕組みはシンプルですが、裏には奥深さがあり、
設計や運用を誤ると大規模な障害となり**私たちに"牙をむく"**こともあります。

また、近年のデータセンタの標準となったEVPN/VXLANなどの新技術も、
根幹にあるのは、このレイヤ2をいかに拡張し、制御するかという思想です。

本セッションでは、改めて基礎から紐解き、日々のネットワーク構築・運用、
トラブルシューティングに役立つ情報を共有できれば幸いです。



本セッションのゴール (Take-away)

1. 【入門編】フレーム転送の**「仕組み」**と**「リスク」**を理解する
   - MACアドレス学習とフォワーディングの流れ
   - ユニキャストとBUMの違い
   - ループ(ブロードキャストストーム)の怖さ

2. 【基礎編・応用編】L2ネットワークを支える**「制御技術」**と**「トラブル」**を知る
   - VLAN, STP, LAGなどの目的と手段
   - 設計と運用の落とし穴

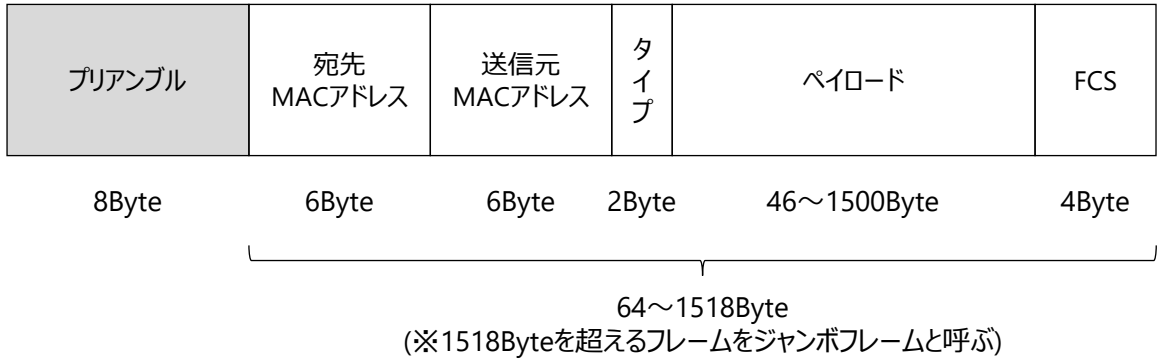
3. 【発展編】基礎から**新技術へ視野を広げる**
   - EVPNの目的と手段
   - (時間があれば)EVPN/(SR)MPLSとEVPN/VXLANの経路学習とフォワーディングの流れ

入門編

フレーム転送の**「仕組み」**と**「リスク」**を理解する

イーサネットのフレーム構造

Ethernet Ⅱフレーム



※802.3の規格もあるがここでは最も使われているEthernet Ⅱを紹介する

■ **プリアンブル**
10101010 × 7Byte + 10101011 × 1Byte
でイーサネットフレームの始まりを示す
※正確にはイーサネットフレームに含まない

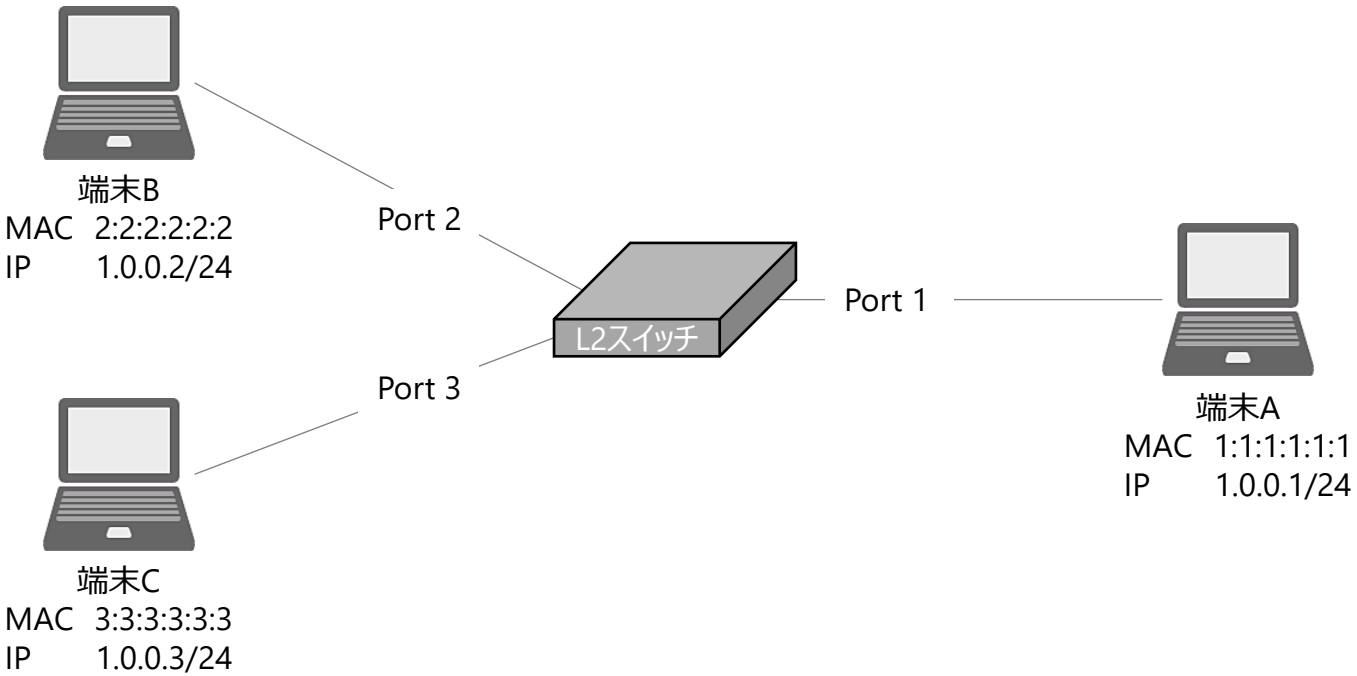
■ **宛先MACアドレス / 送信元MACアドレス**
48ビットの長さを持つ。種別としては、
ユニキャスト / マルチキャスト / ブロードキャストアドレスがある
ユニキャストアドレスは基本的にグローバルで一意

■ **タイプ**
ペイロードのプロトコルを示す。
(例: 0x0800=IPv4, 0x0806=ARP, 0x86DD=IPv6 等)

■ **ペイロード**
データ部分。IPパケット等が格納される

■ **FCS**
受信側でフレームの破損をチェックするためのもの

フレーム転送 (0/8 構成)

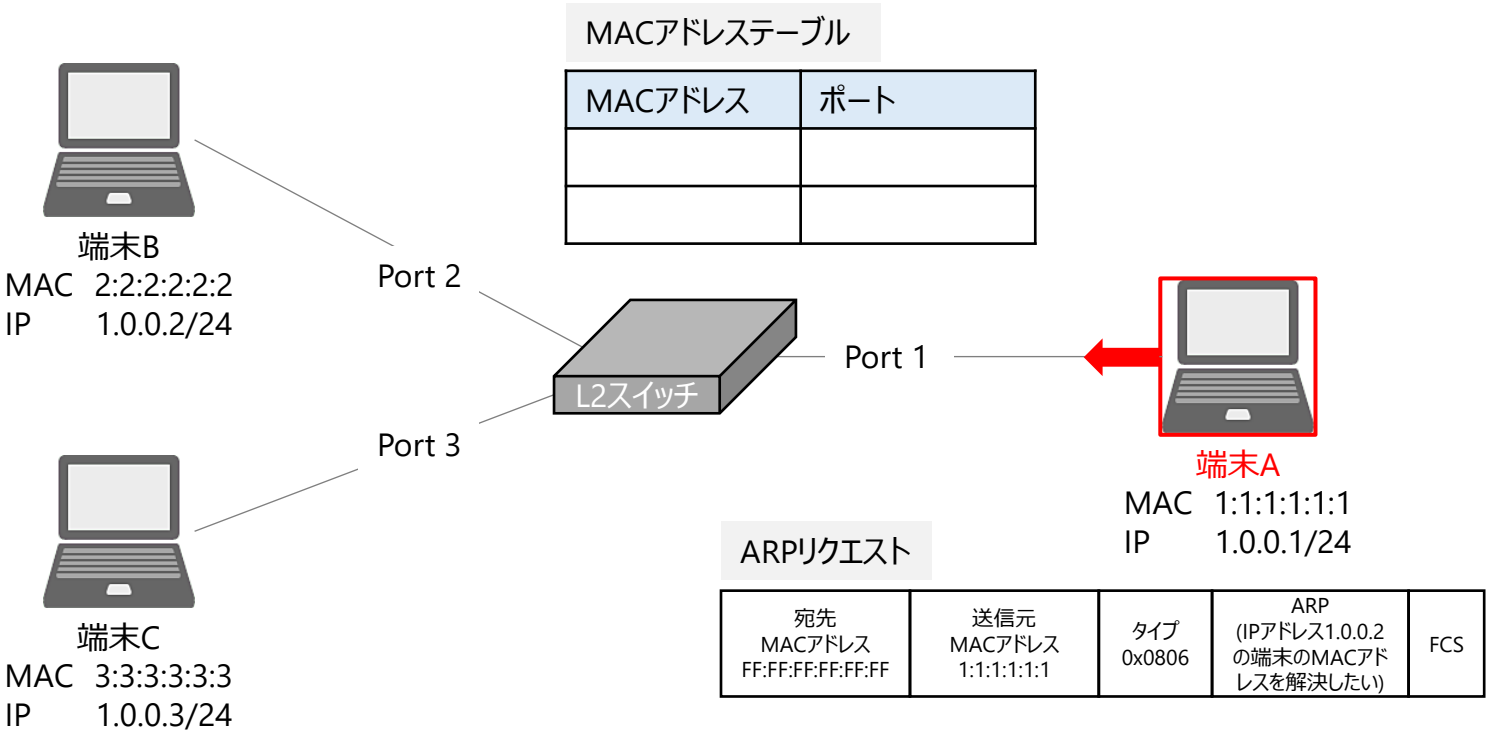


■ **物理構成**
L2スイッチの各ポートに端末を接続
- Port 1ー端末A
- Port 2ー端末B
- Port 3ー端末C

■ **MACアドレス**
各端末は以下のMACアドレスを持つ
- 端末Aー1:1:1:1:1:1
- 端末Bー2:2:2:2:2:2
- 端末Cー3:3:3:3:3:3

■ **IPアドレス**
各端末は以下のIPアドレスを持つ
- 端末Aー1.0.0.1/24
- 端末Bー1.0.0.2/24
- 端末Cー1.0.0.3/24

フレーム転送 (1/8 端末AがARPリクエスト送出)



■ **目的**

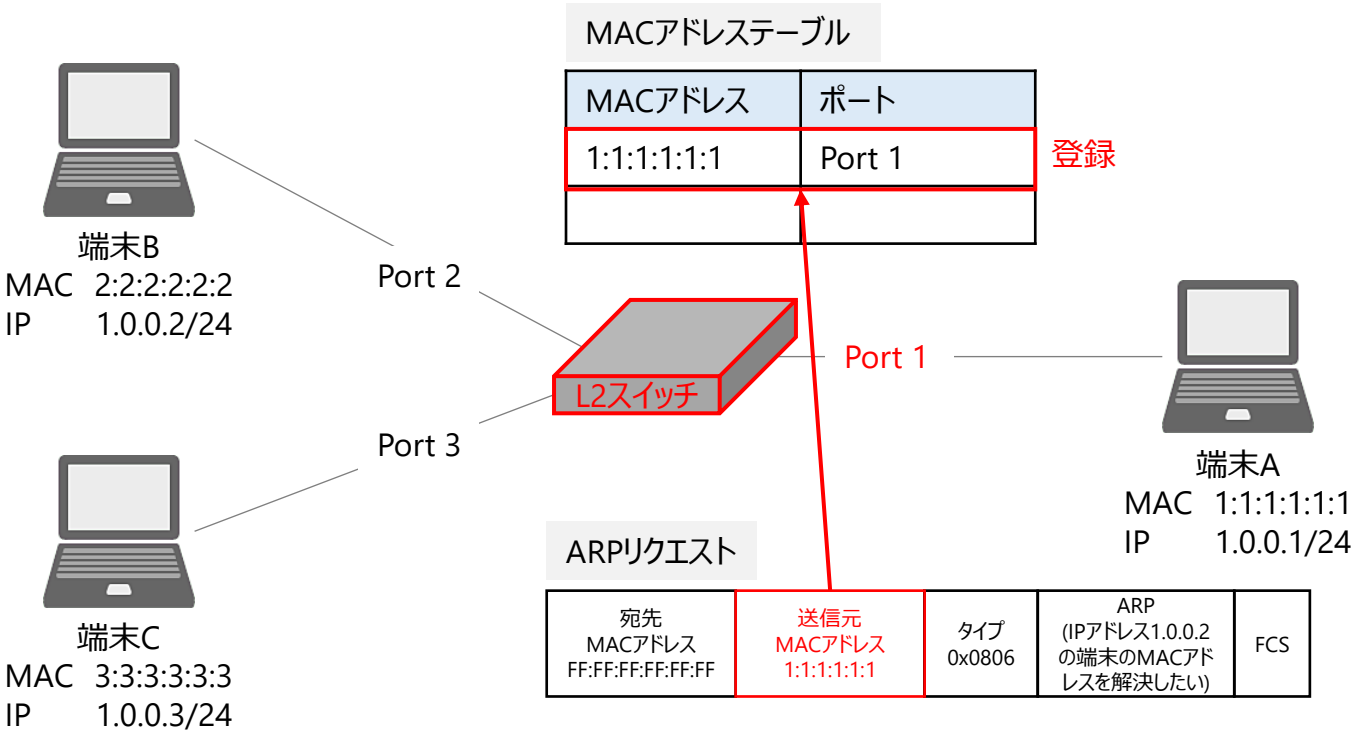
端末Aは端末Bと通信するために、
端末BのMACアドレスを知りたい

■ **方法**

端末Aから端末BのMACアドレスを解決するためのARPリクエストを送出

- 宛先MACアドレス
  FF:FF:FF:FF:FF:FF(Broadcastアドレス)
- 送信元MACアドレス
  1:1:1:1:1:1
- タイプ
  0x0806
- ARP
  ターゲットとなるIPアドレス1.0.0.2の情報を格納

フレーム転送 (2/8 L2スイッチが端末AのMACアドレス学習)

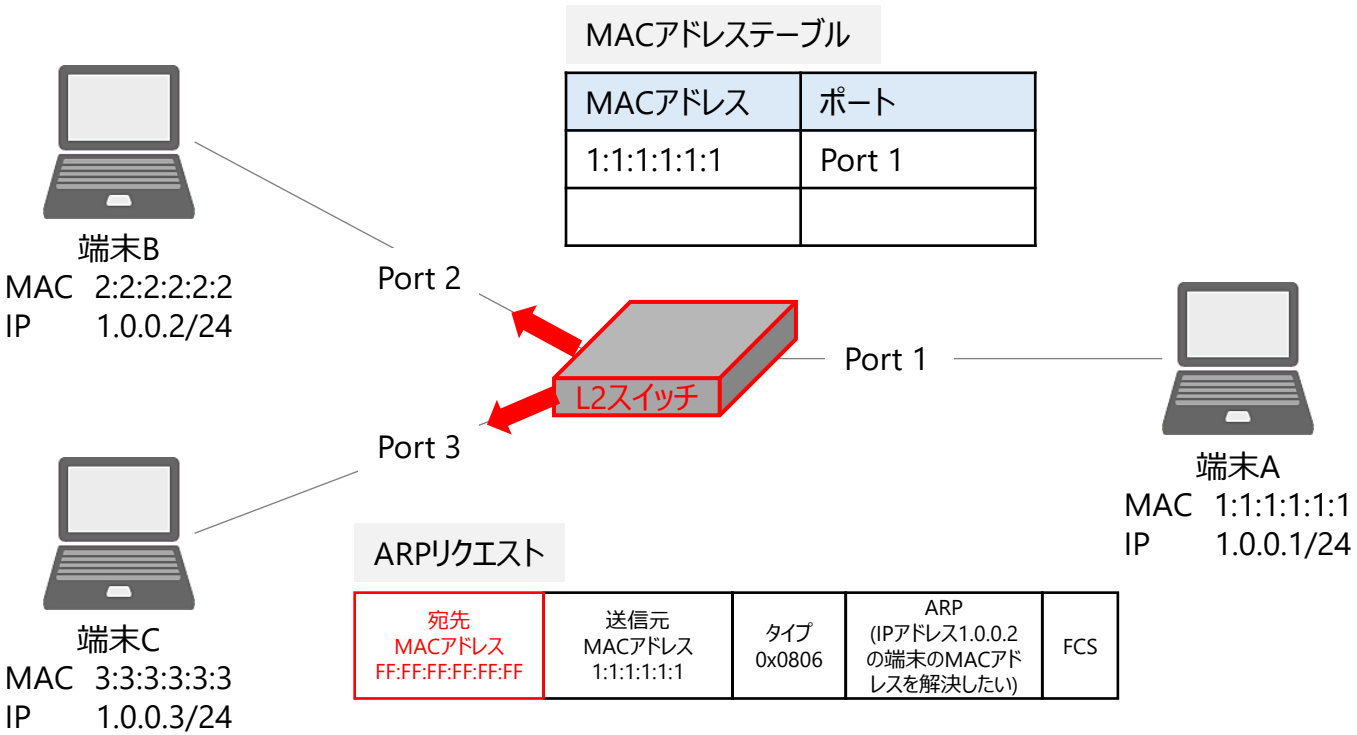


■ **目的**
L2スイッチは宛先MACアドレスをもとに、
フレームを転送できるようになりたい

■ **方法**
受信したフレームの送信元MACアドレスを、
Port情報と共にMACアドレステーブルに登録

(※今後、宛先MAC1:1:1:1:1:1のフレームが来たらPort 1に転送できる)

フレーム転送 (3/8 L2スイッチがARPリクエストをフラッディング)



※端末Cは破棄

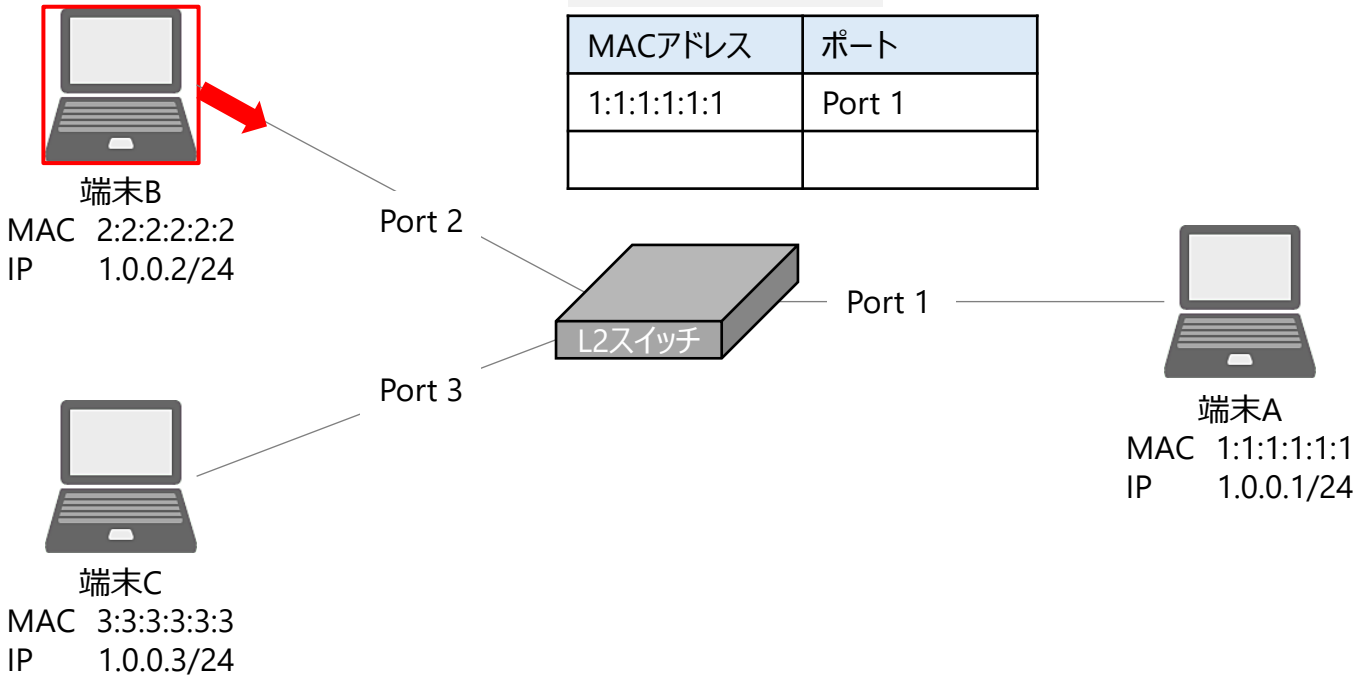
■ **目的**
同一ネットワークの全端末に、IPアドレス1.0.0.2のMACアドレスの聞き込みをしたい

■ **方法**
ARPリクエストの宛先MACアドレスがBroadcastアドレスなので、L2スイッチは受信ポート以外の全てのポートに同一のフレームをコピーして送信する **(フラッディング)**

フレーム転送 (4/8 端末BがARPリプライ送出)

ARPリプライ

宛先 MACアドレス 1:1:1:1:1:1	送信元 MACアドレス 2:2:2:2:2:2	タイプ 0x0806	ARP (端末BのMACアドレス情報)	FCS



MACアドレステーブル

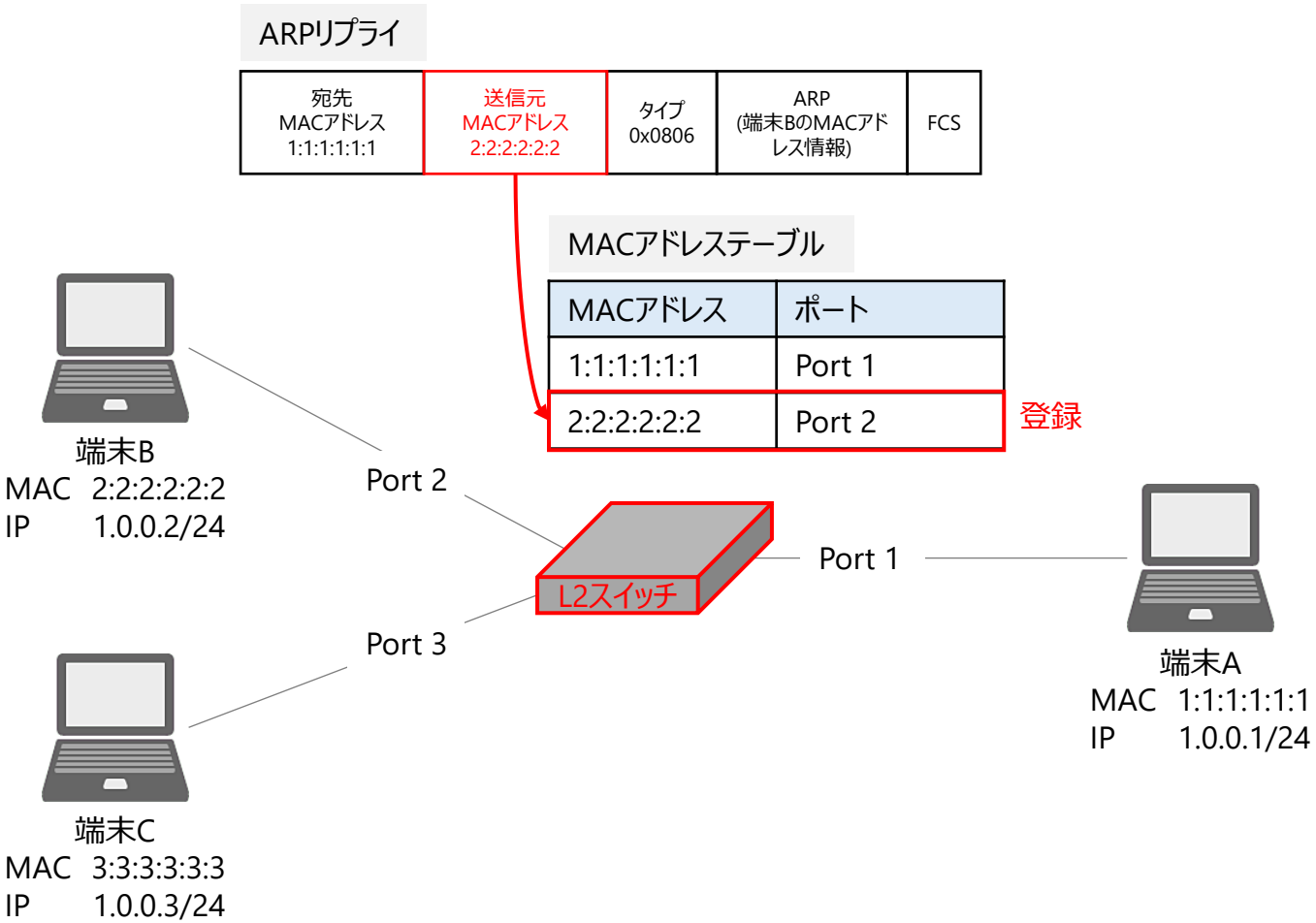
MACアドレス	ポート
1:1:1:1:1:1	Port 1

※端末Cは破棄

■ **目的**
端末Bは端末AへMACアドレス情報を回答したい

■ **方法**
端末BからMACアドレス情報を回答するためのARPリプライを送出
- 宛先MACアドレス
  1:1:1:1:1:1
- 送信元MACアドレス
  2:2:2:2:2:2
- タイプ
  0x0806
- ARP
  端末BのMACアドレス情報を格納

フレーム転送 (5/8 L2スイッチが端末BのMACアドレス学習)

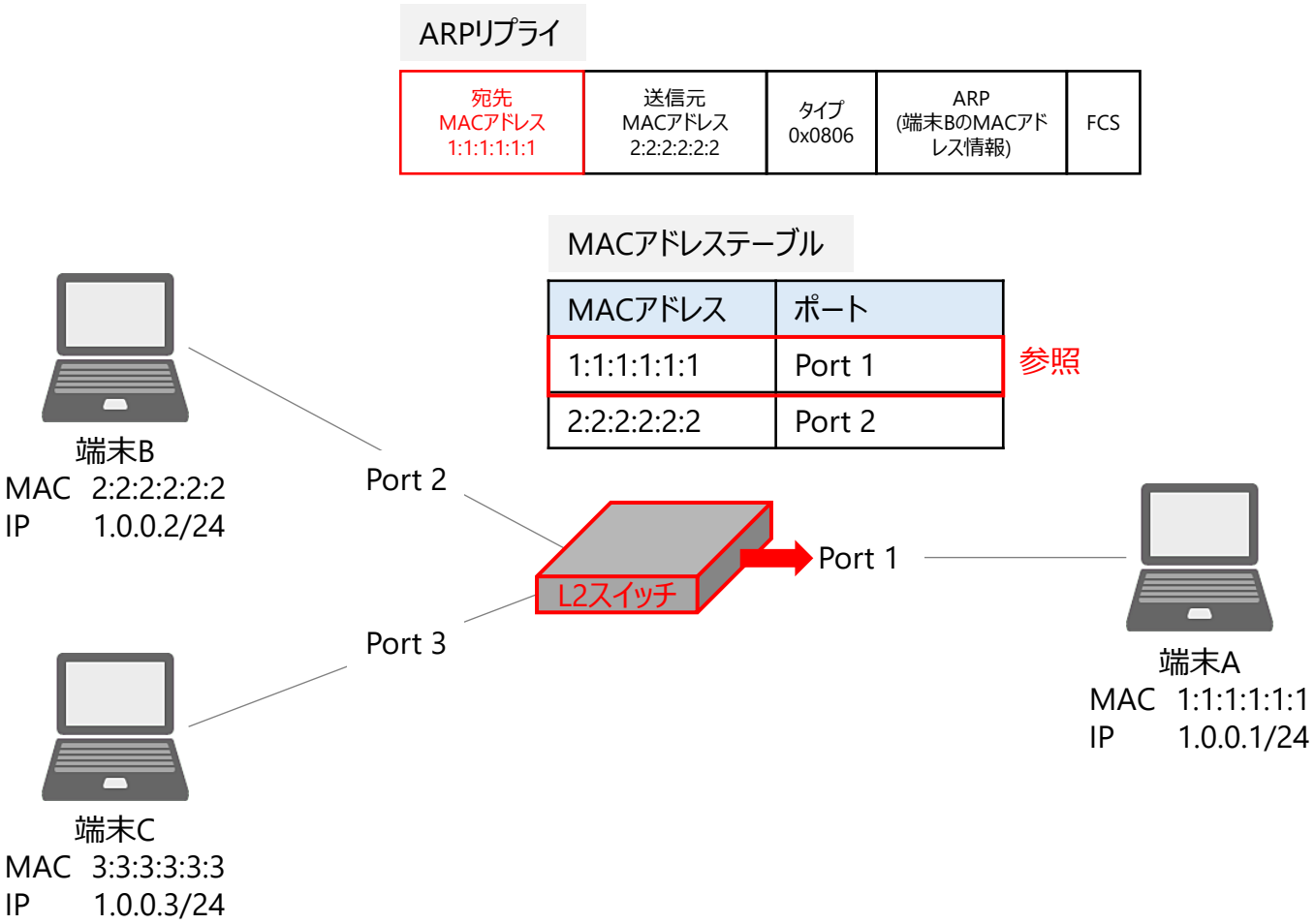


■ **目的**
L2スイッチは宛先MACアドレスをもとに、
フレームを転送できるようになりたい

■ **方法**
受信したフレームの送信元MACアドレスを、
Port情報と共にMACアドレステーブルに登録

(※今後、宛先MAC2:2:2:2:2:2のフレームが来たらPort 2に転送できる)

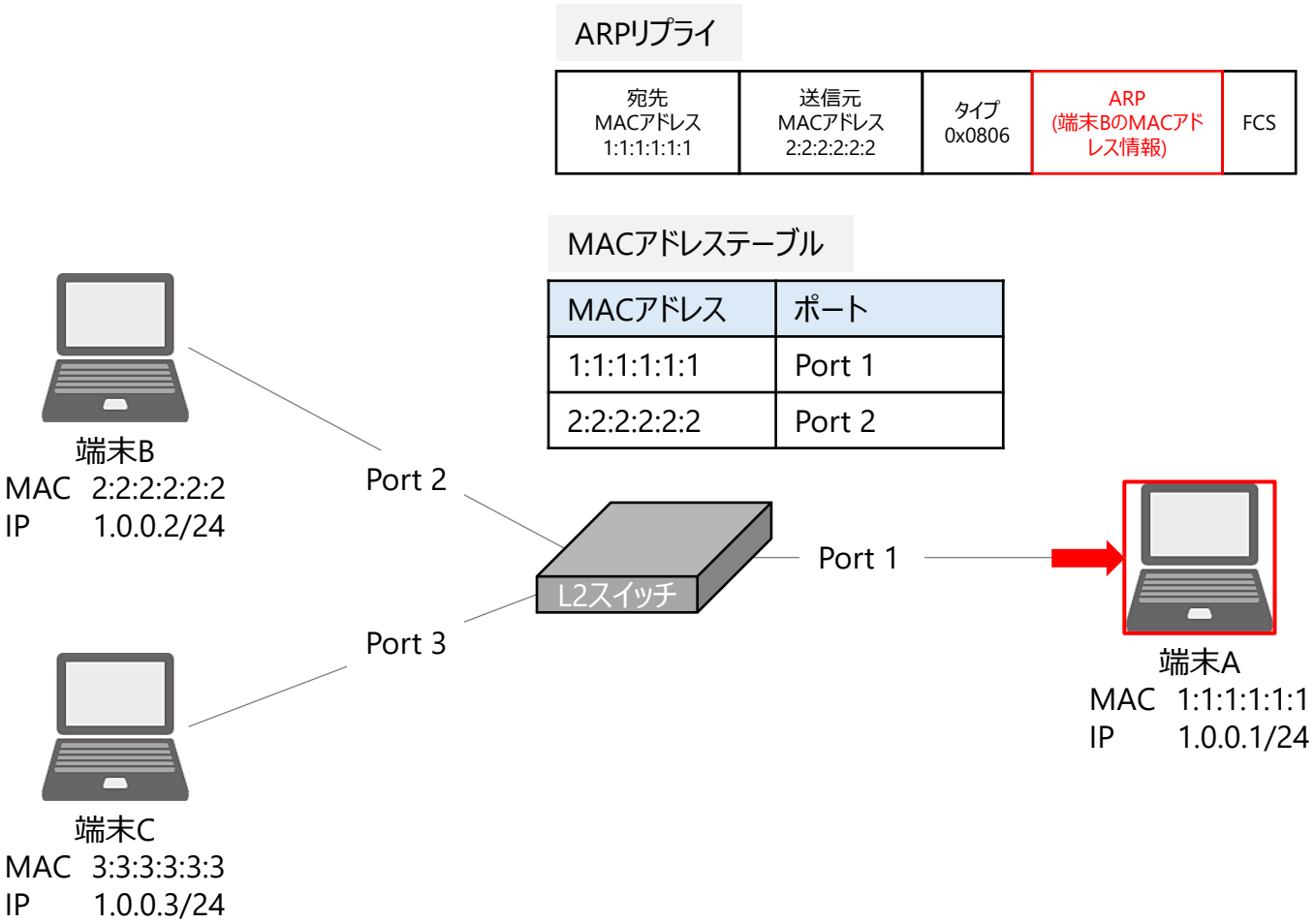
フレーム転送 (6/8 L2スイッチがARPリプライを端末Aに転送)



■ **目的**
L2スイッチは宛先MACアドレスをもとに、ARPリプライを転送したい

■ **方法**
受信したフレームの宛先MACアドレスとMACアドレステーブルを参照して適切なポートに転送する

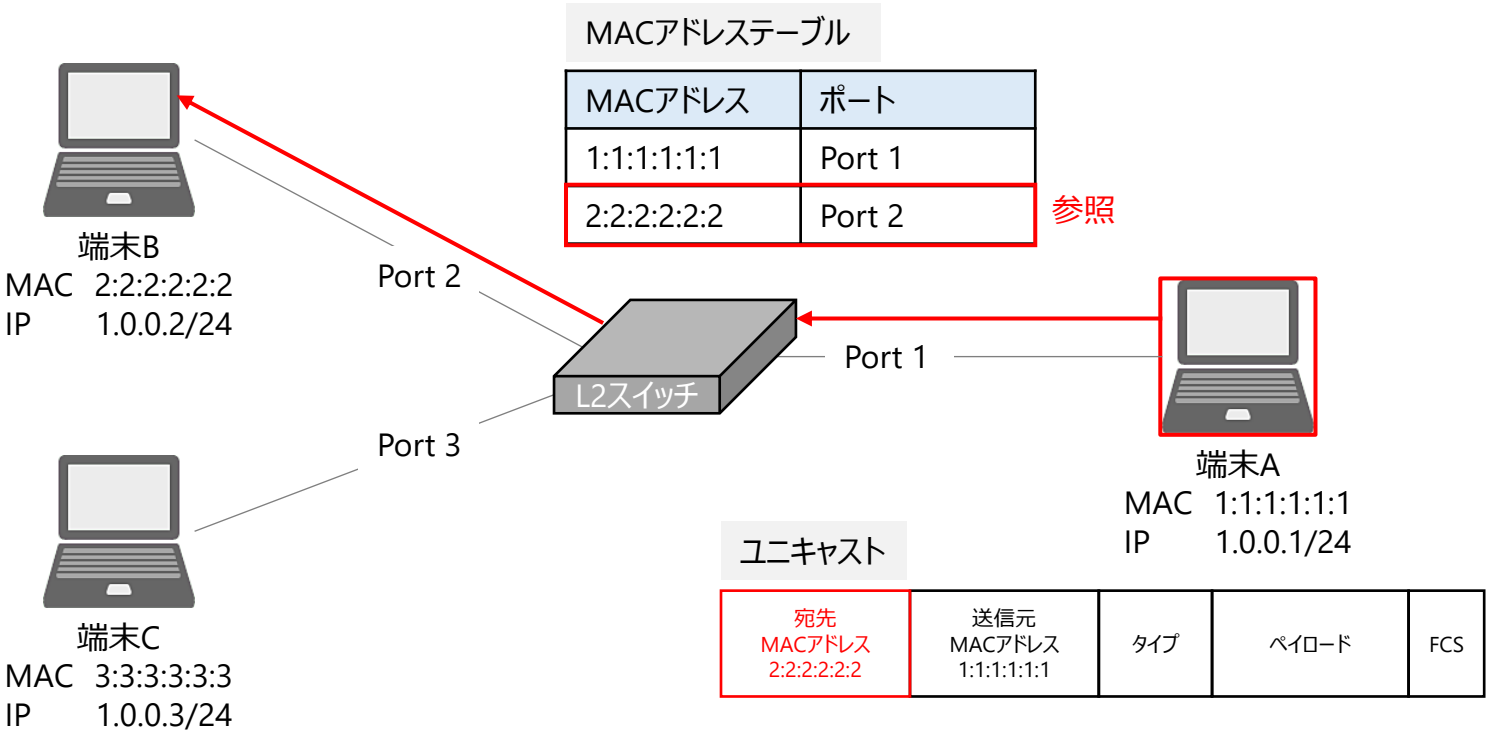
フレーム転送 (7/8 端末Aが端末BのMACアドレスを学習)



■ **目的**
端末Aは端末Bと通信するために、
端末BのMACアドレスを知りたい

■ **方法**
受信したARPリプライから端末Bの
MACアドレス情報を学習

フレーム転送 (8/8 端末Aから端末Bへ通信)



- **目的**
  端末Aは端末Bと通信したい

- **方法**
  学習した端末BのMACアドレスを宛先MACアドレスにしてフレームを送信 **(Unicast)**

  L2スイッチはMACアドレステーブルの情報に基づきフレームを端末Bに転送

フレームの種類

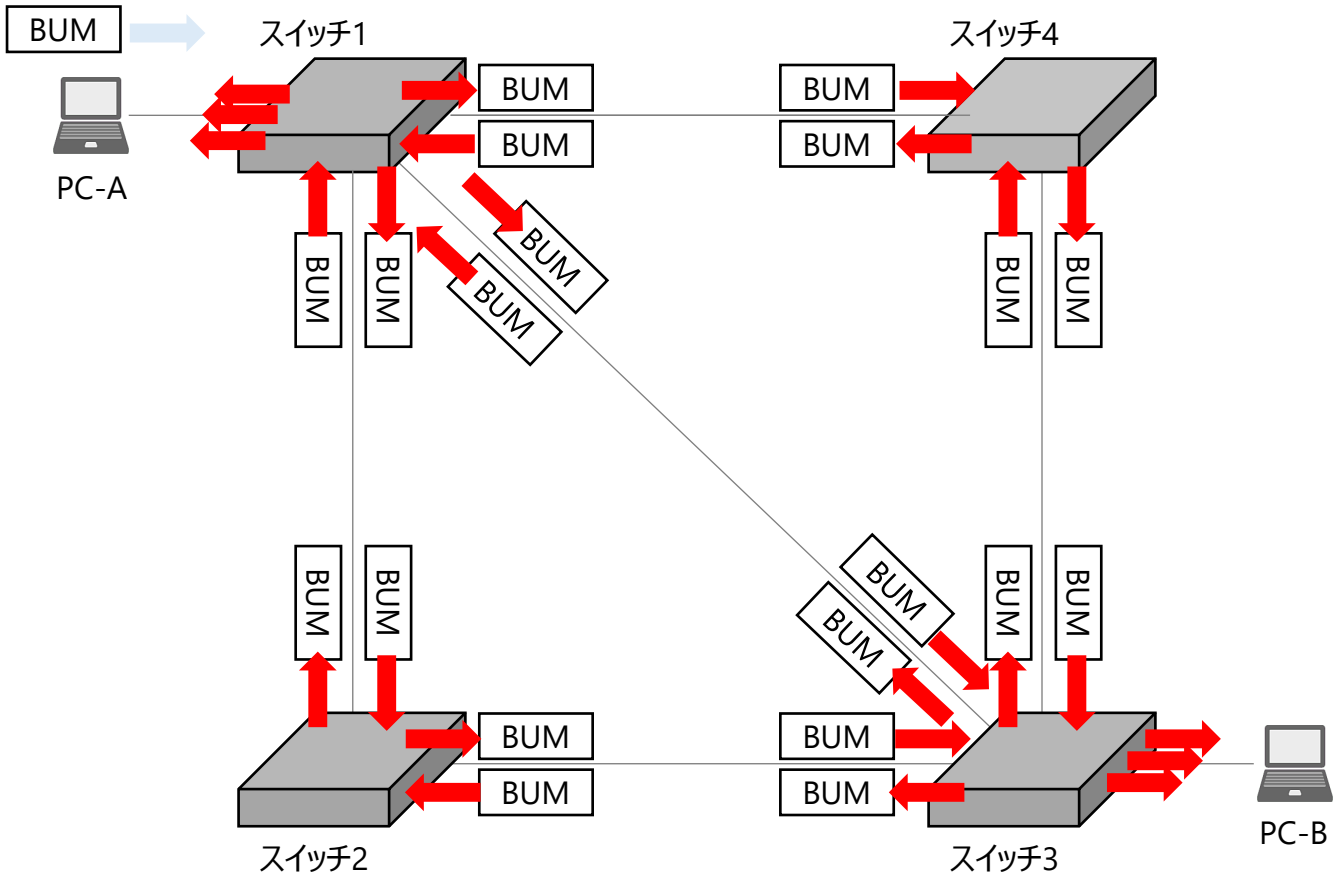
フレームの種類	宛先	L2スイッチの動作	用途
Unicast	特定の1台	特定のポートへ転送	通常の通信
Broadcast	同一ネットワークの全台	フラッディング	ARP等
Unknown Unicast	特定の1台	フラッディング	(エージングにより発生)
Multicast	特定のグループ	フラッディング (※IGMP Snoopingで効率化可能)	OSPF等

まとめてBUMと呼ぶ

■ **フレームの種類**
- Unicast
- Broadcast
- Unknown Unicast
- Multicast

■ **BUM**
Broadcast, Unknown Unicast, Multicast はL2スイッチでフラッディングされる特徴を持ち、頭文字を取ってBUMと呼ばれる

ブロードキャストストーム



■ **ブロードキャストストームの概要**

L2ネットワークでループ構成を作ってしまうと、
BUMが永久に増殖・転送され続けて、
ネットワークが使えなくなる

■ **なぜ問題が起きるのか**

- BUMのフラッディング
  各スイッチがBUMを受信する度に
  受信ポート以外の全ポートにコピーし、
  指数関数的に増殖する
- フレームに寿命が無いこと
  一度ループすると永久に回り続ける

■ **具体的な影響**

- BUMによる帯域の占有
- スイッチに接続されている端末への負荷
- MACアドレステーブルの激しい書き換わりによるスイッチへの負荷

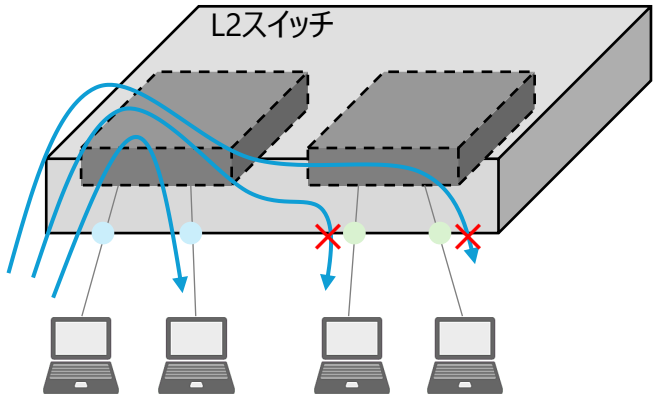
基礎編

L2ネットワークを支える**「制御技術」**と「トラブル」を知る

VLAN (概要)

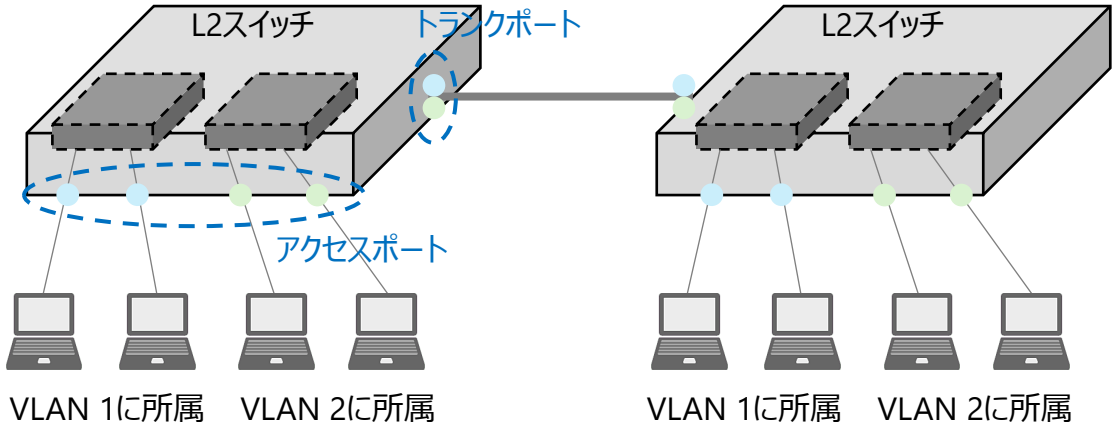
※主にどのネットワークに関連するか示すラベル → Enterprise Service Provider Data Center

VLANのイメージ



ブロードキャストドメインを論理的に分割できる

アクセスポートとトランクポート



■ 目的

1. BUMフラッディング範囲(**ブロードキャストドメイン**)分割
   - セキュリティの向上
     部署間や用途で通信範囲を分離
     盗聴や不正アクセスを防ぐ
   - L2スイッチと端末のパフォーマンス最適化
     L2スイッチ：不要なポートにBUMコピーしない
     端末：不要なBUMの受信・CPU処理しない
2. 論理的な分割による設備利用の効率化
   - 複数L2スイッチの用意、物理配線の変更をせず論理的な設定だけで分割

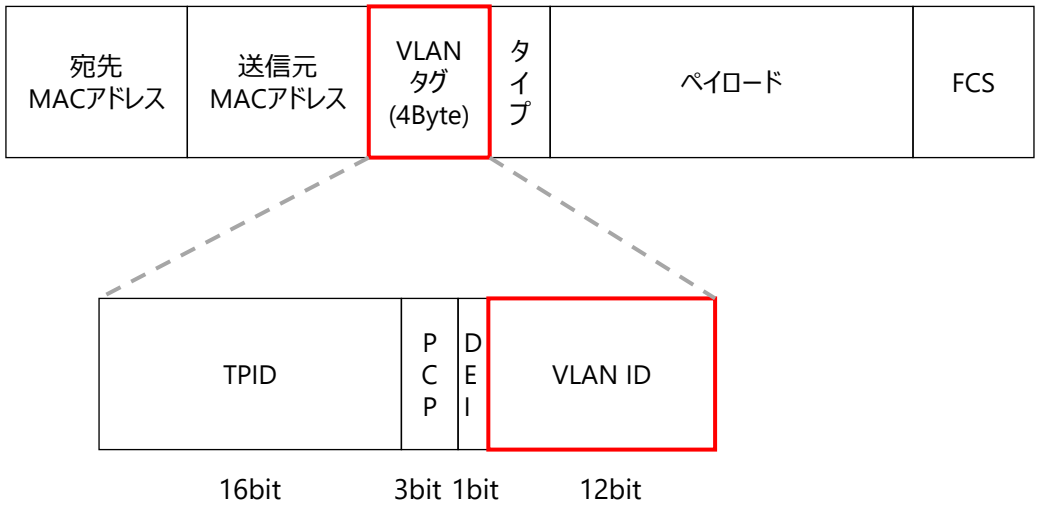
■ 方法

- L2スイッチの各ポートがどのVLANに所属するか定めてブロードキャストドメインを論理的に分割する
- ポートは2種類ある
  ① **アクセスポート**：L2スイッチと端末等を接続
  1つのVLANに所属
  ② **トランクポート**：L2スイッチとL2スイッチ等を接続
  複数のVLANに所属

VLAN (IEEE802.1Q)

Enterprise　Service Provider　Data Center

VLANタグ付きフレーム



■ **目的**

- トランクポートにおけるVLANの識別
  L2スイッチはトランクポートで受信したフレームが、
  どのVLANに所属するのか識別する必要がある
  (※通常のイーサネットフレームではそれを示す情報は
  格納されていない)

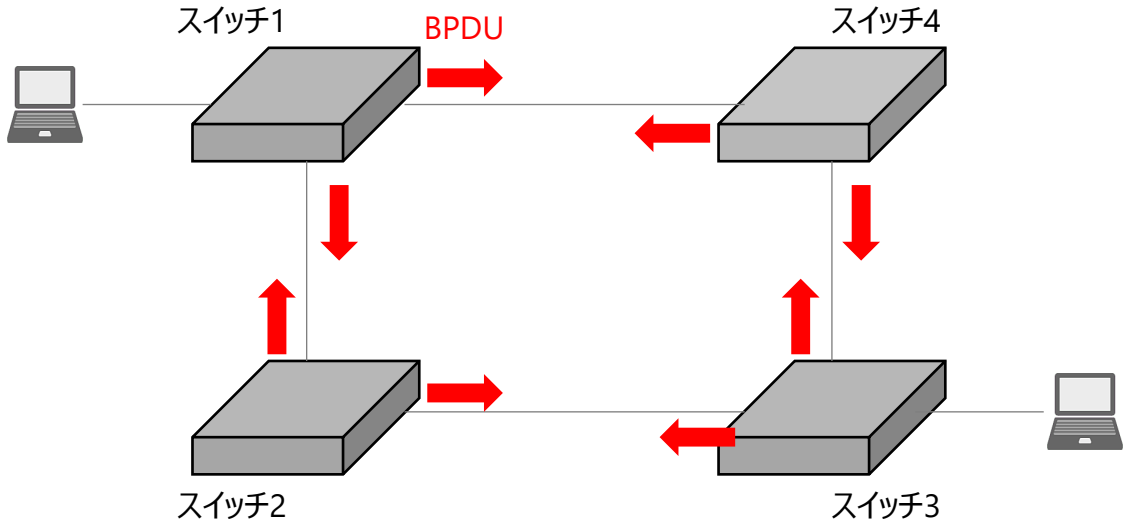
■ **方法**

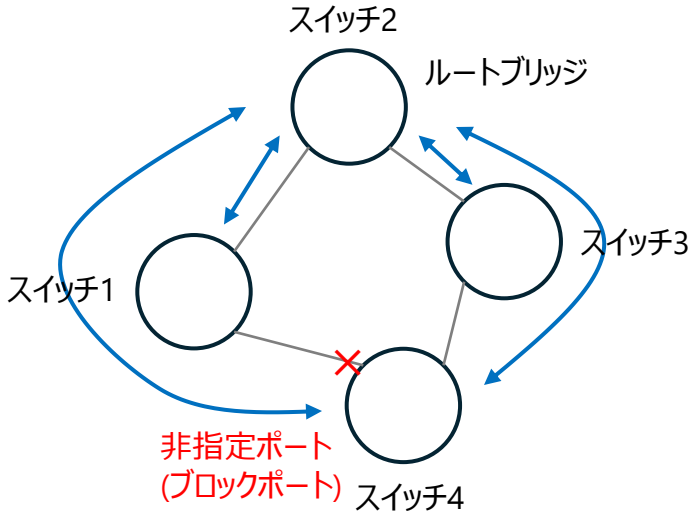
- IEEE802.1Qで定めた**VLANタグ**の**VLAN ID**で識別
- VLANタグは以下の要素で構成される

  TPID
  タグのプロトコル種別を示す。
  IEEE802.1Qは0x8100。
  PCP
  0～7の8段階で優先度を示す。CoS。
  DEI
  ネットワーク輻輳時のフレーム破棄の可否を示す。
  VLAN ID
  VLANの識別子(1～4094)

STP

Enterprise





■ 目的

- ネットワークの耐障害性の向上
  ブロードキャストストームを回避しつつ、
  ネットワークを冗長化したい

■ 方法

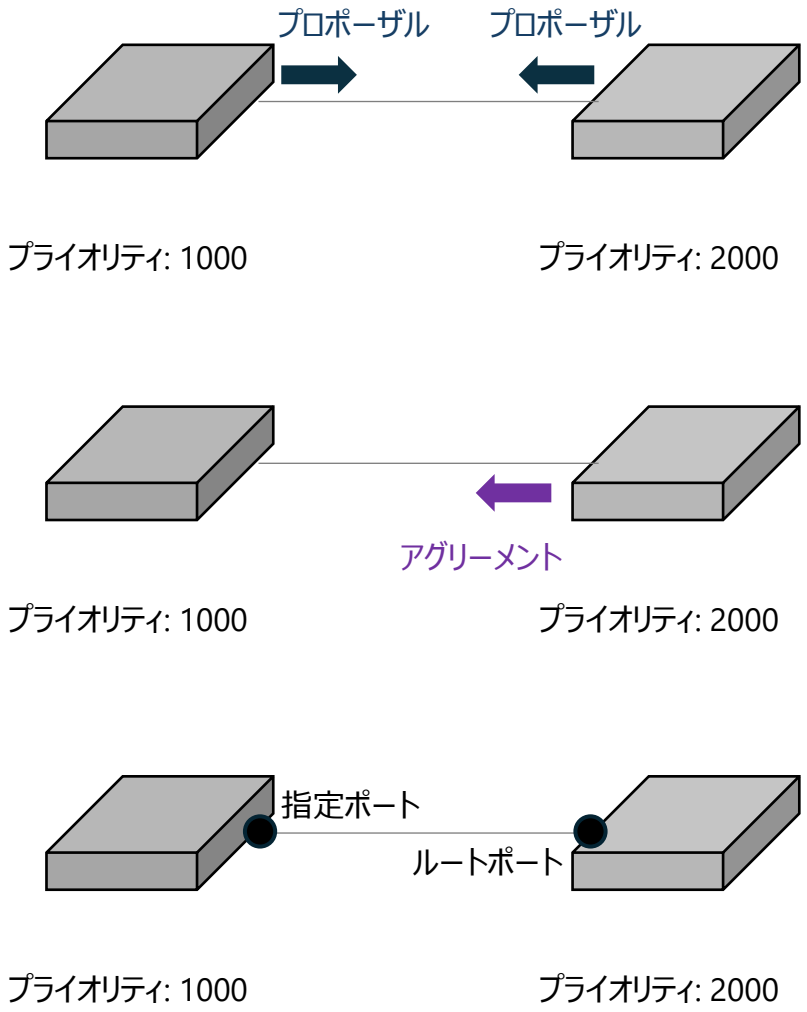
- IEEE802.1Dで定めた**STP** (Spanning Tree Protocol)
  ループが発生しないように、意図的に特定のポート
  (非指定ポート)をブロック

  制御用フレームBPDU(Bridge Protocol Data Unit)を
  スイッチ間でやり取りし、ブロックするポートを決める
  ① ルートブリッジとなるスイッチの決定
  ② ルートポート、指定ポートの決定
  ③ ルートポート、指定ポートに選ばれなかった
  非指定ポートをブロック

  ※ブロードキャストストームの発生を防ぐために、非指
  定ポートが決定するまでトラフィックの転送を行わない
  ように状態を遷移させるため、収束まで約50秒かかる

RSTP

Enterprise



■ 目的

- ネットワークの耐障害性の向上と収束の高速化
  ブロードキャストストームを回避しつつ、
  ネットワークを冗長化し、収束も高速化したい

■ 方法

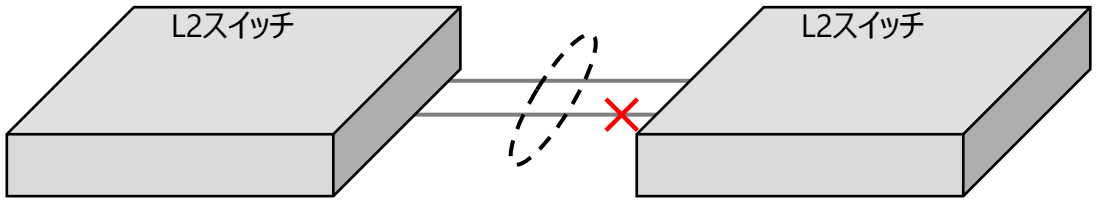
- IEEE802.1wで定めた**RSTP** (Rapid STP)
  スイッチ間で以下のBPDUを使用し、
  タイマーを待たずポート役割を即決
  - **プロポーザル**
    「私が指定ポートとして動作します」という提案
  - **アグリーメント**
    「承知、私はルートポートになります」という同意

  端末が繋がるポートとスイッチが繋がるポートを区別
  - **エッジポート**(端末等が繋がるポート)
    ループの危険が無いので即座にフォワーディング状態にする
  - **非エッジポート**(スイッチが繋がるポート)
    BPDUのやり取りをする

LAG (Link Aggregation Group)

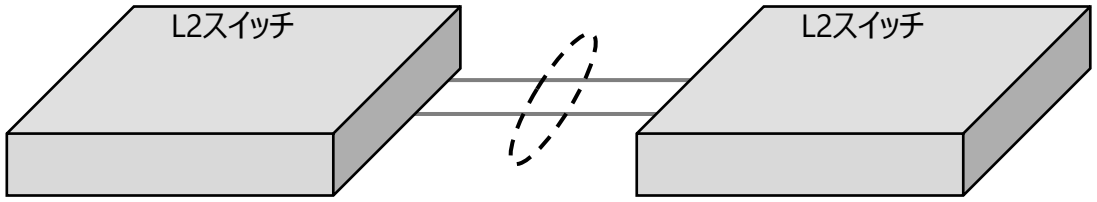
Enterprise　Service Provider　Data Center

耐障害性の向上



ある物理リンクに障害が発生した場合でも通信を継続できる

帯域幅の柔軟な増加



(例)1Gの物理ポートを2つ束ねて最大2Gの通信ができるようになる

■ 目的

- 耐障害性の向上
  物理リンクが切れても別の物理リンクで転送し続けたい
- 帯域幅の柔軟な増加
  複数の物理リンクの帯域を足し合わせて使いたい

■ 方法

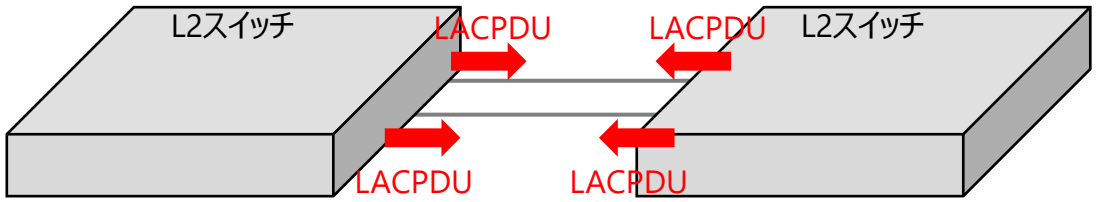
- IEEE802.1AXの**LAG (Link Aggregation Group)**
  複数の物理リンクを1本の論理的なリンクとして扱う
  - ➢ LAG内のある物理リンクに障害が発生しても通信を継続できる
  - ➢ 最大帯域は物理リンクの帯域合計
    (例)1Gの物理ポート2つで最大2Gの通信が可能

- LAGを形成する方法としては以下の2つがある
  ① 手動設定による静的な方法
  ② 動的な方法(LACPが主流)

LAG (LACP)

Enterprise　Service Provider　Data Center

LACPDU交換イメージ



LACPDUのフレームフォーマット

DA	SA	Type	Sub Type	LACP ver.	Actor Info.	Partner Info.	Collector Info.	Terminator	FCS
0180:c200:0002		0x8809	0x01	0x01					

Actor Information Length
Actor_System_Priority
Actor_System (MACアドレス)
Actor_Key
Actor_Port_Priority
Actor_Port
Actor_State
Reserved

Partner Information Length
Partner_System_Priority
Partner_System (MACアドレス)
Partner_Key
Partner_Port_Priority
Partner_Port
Partner_State
Reserved

■ 目的

- LAGを動的に形成する

■ 方法

- LACP (Link Aggregation Control Protocol)
  物理リンクごとにLACPDU(Data Unit)を送受信

➢ **識別情報の交換**
LACPDU内の「Actor(自装置)」と「Partner(対向装置)」の情報を交換

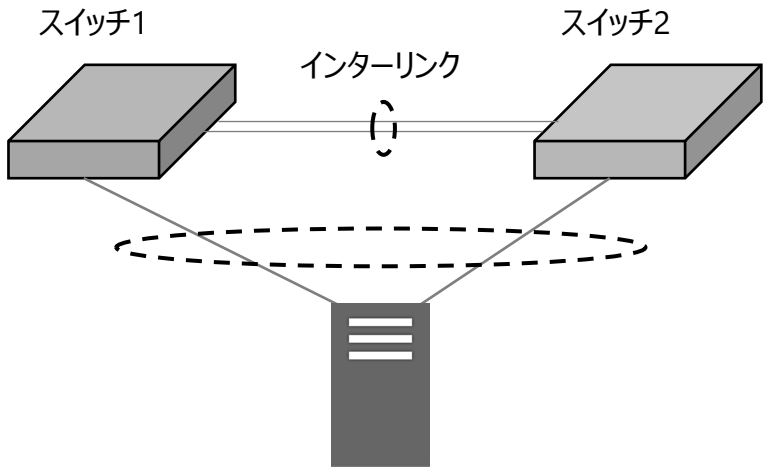
➢ **初期状態**
対向装置の情報が分からないので、Partner情報は0の状態で開始

➢ **LAG確立**
お互いの情報を学習し、LACPDUのPartner情報が正しく記載されたことを確認しLAG確立

LAG (複数筐体)

Enterprise　Service Provider　Data Center

複数筐体によるLAG形成のイメージ



■ **目的**

- 耐障害性の更なる向上
  リンク観点の冗長だけでなく、筐体観点の冗長

■ **方法**

- 複数筐体に跨ったLAG形成
- 注意点
  - ベンダ独自プロトコルの使用
  - 状態同期のためのインターリンクが必要

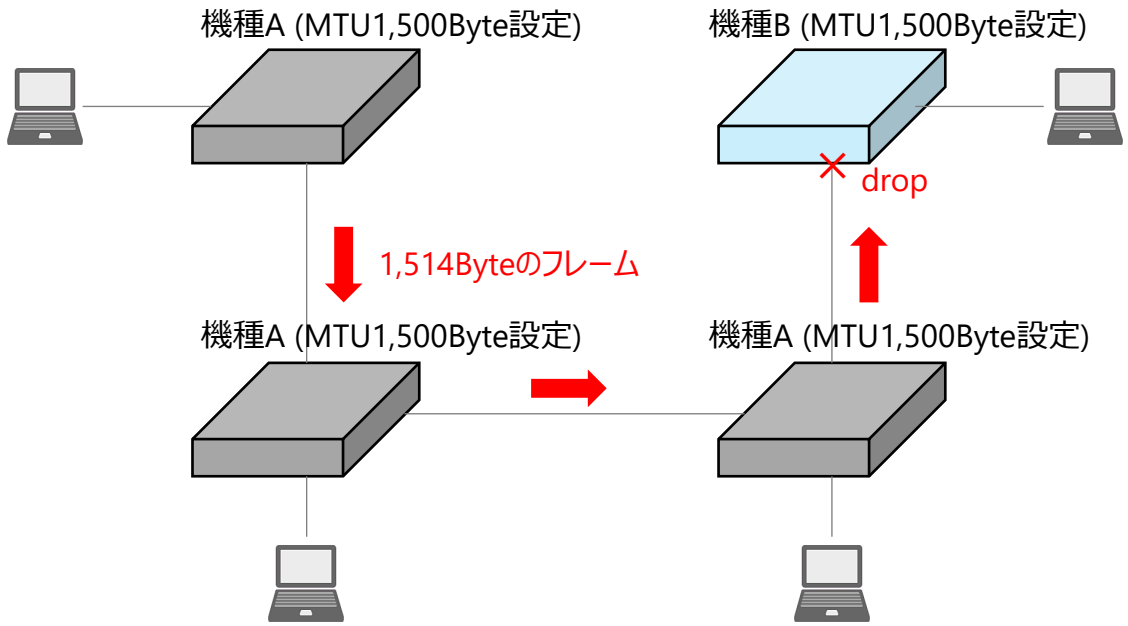
- 標準技術且つインターリンク不要のEVPN
  マルチホーミングが冗長化ソリューションとして注目

応用編

L2ネットワークを支える「制御技術」と**「トラブル」**を知る

トラブル事例 (MTU設定ミス)

Enterprise　Service Provider　Data Center



■ **トラブル概要**

- MTU設定ミスによるフレームdrop
  L2スイッチが1回の転送で送れるフレームサイズを設定を誤ると大きなフレームが落ちることがある

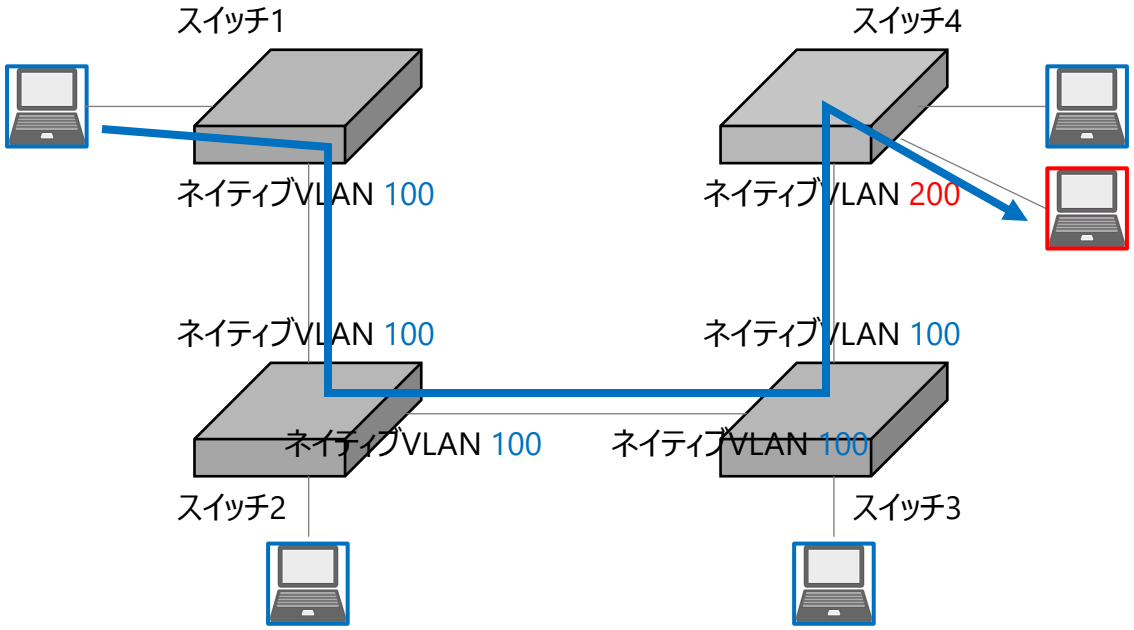
■ **具体例**

- MTUの計算の仕方が機種・ベンダ毎で異なるケース
  - 機種AはMTUに宛先MACアドレス、送信元MACアドレス、タイプ (計14 Byte) を含まない
  - 機種BはMTUに宛先MACアドレス、送信元MACアドレス、タイプ (計14 Byte) を含む
  - 最大1,514 Byte(※FCS除く)を通したい場合は、機種Aは1,500 Byte、機種Bは1,514 Byteの設定が適切
  - 機種BでもMTUを1,500 Byteに設定すると、FCS除くフレームサイズが1,500 Byteを超えるフレームを転送しようとするとdropしてしまう
  - (小さなサイズのフレームは通るのでPingは通るのに…ということが起こり得る)

トラブル事例 (ネイティブVLAN設定ミス)

Enterprise

Data Center



■ **トラブル概要**

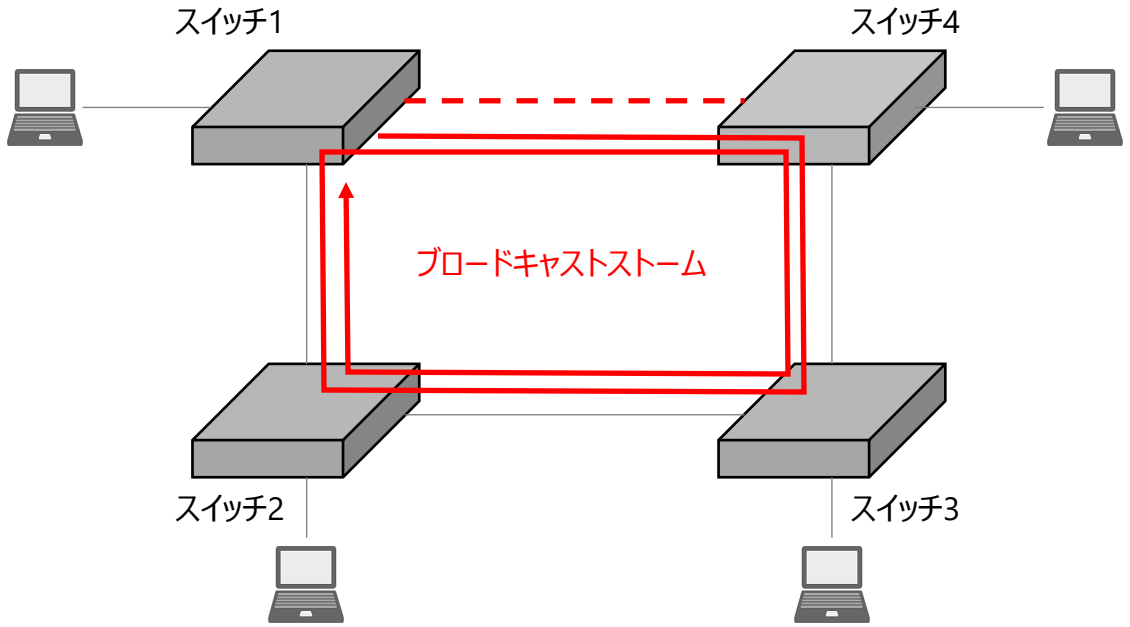
- ネイティブVLAN設定ミスによるVLANリーク
  VLANタグ無しフレームを特定のVLANとして扱う
  ネイティブVLANの設定を誤ると意図しないVLAN間通信が発生することがある

■ **具体例**

- ネイティブVLANを以下のように誤って設定
  - スイッチ1, 2, 3ではネイティブVLAN100
  - スイッチ4ではネイティブVLAN200
- VLAN100として送られてきたフレームがスイッチ4でVLAN200として扱われ不適切な端末に転送される

トラブル事例 (ループ構成①)

Enterprise



■ **トラブル概要**

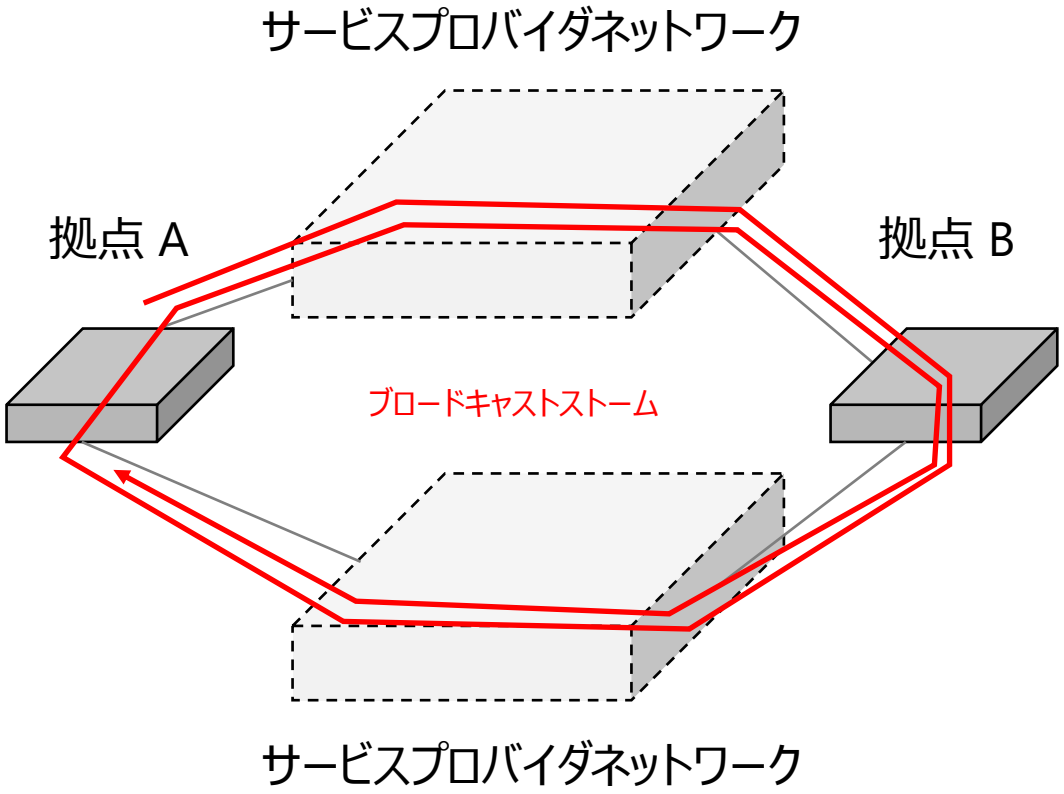
- 物理配線の誤りによるブロードキャストストーム
  STPを無効にした環境で物理配線の誤りにより、
  ループが形成され、ブロードキャストストームが発生

■ **具体例**

- 誤ったスイッチ間のケーブル追加
  スイッチ1とスイッチ4の間に誤ってケーブルを追加して
  ブロードキャストストームが発生
  (※図のような単純な構成ではミスは起こりにくいが、
  トポロジーが複雑化し、ラックを跨ぐ配線が増えるほど、
  危険性が増す)

トラブル事例 (ループ構成②)

Enterprise　Service Provider



■ **トラブル概要**

- 拠点をまたがったループ形成でブロードキャストストーム
  サービスプロバイダネットワークを利用して離れた拠点のL2ネットワークを延伸する際、拠点をまたがった大きなループが形成されてブロードキャストストームが発生

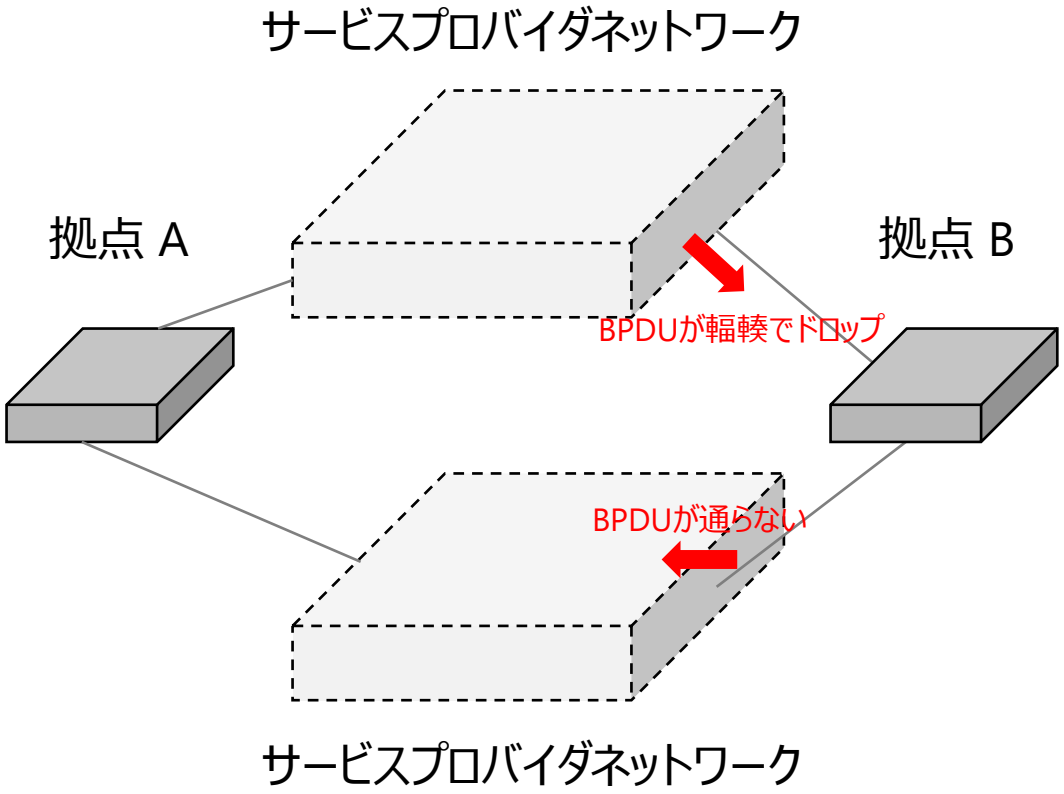
■ **具体例**

- 2拠点を2回線で冗長してL2ネットワーク延伸
  (※物理的に離れた所でループが形成されてしまうので、発見しづらい。)

  サービスプロバイダネットワークにも大きな負荷がかかる
  対策としては流入するBUMの最大量を制限する
  Storm Controlがある

トラブル事例 (ループ構成③)

Enterprise　Service Provider



■ **トラブル概要**

- 拠点をまたがったループ形成でブロードキャストストーム
  サービスプロバイダネットワークを利用して離れた拠点のL2ネットワークを延伸する際、拠点をまたがった大きなループが形成されてブロードキャストストームが発生

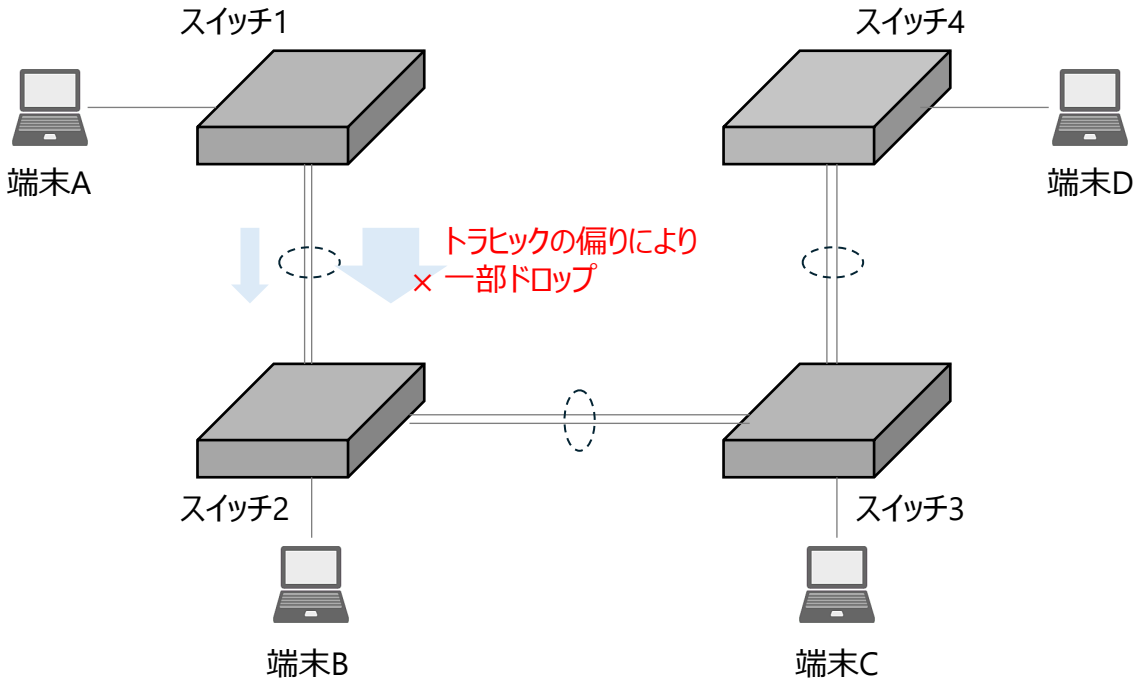
■ **具体例**

- STPのBPDUが対向に届かない
  拠点をまたがった大きなループについても、STPでブロードキャストストームを防止可能だが…

  以下に注意
  - ➢ サービスプロバイダネットワークの仕様で、BPDUが通らない
  - ➢ 輻輳が起きてBPDUがdrop

トラブル事例 (LAGのトラフィック偏り)

Enterprise　Service Provider　Data Center



■ **トラブル概要**

- LAGのトラフィック偏りによるdrop
  LAGに所属するどの物理リンクにフレームを投げるか
  L2～L4ヘッダを参照するルールが存在し、
  同じフレーム(フロー)は同じ物理リンクを通る

  フローの数が少ない場合はフレームの振り分けに
  偏りが生じて、期待した速度が出ずにドロップが発生

■ **具体例**

- 特定のフローのトラフィック量が膨大
  - スイッチ1とスイッチ2間は1G×2でLAG形成
  - 端末Aから端末BへL2～L4のヘッダが全く同じフレームを1.5G送信
  - 1.5Gのフレームは全て同一の物理リンクを通り、dropが発生

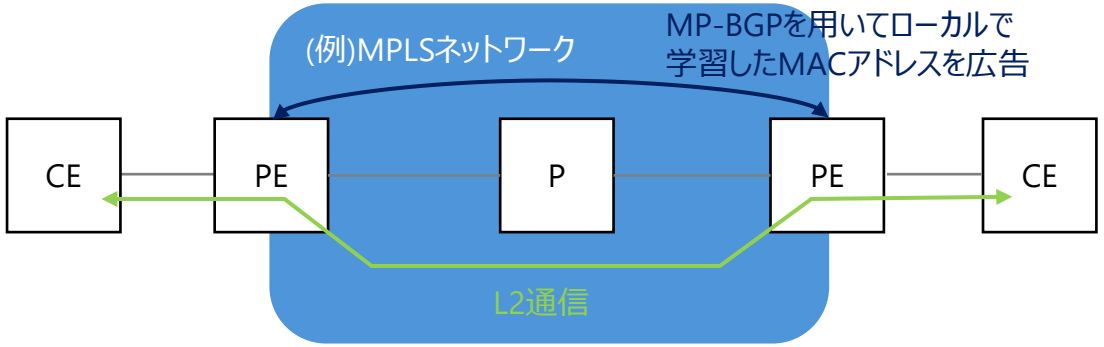
発展編

基礎から**新技術へ視野を広げる**

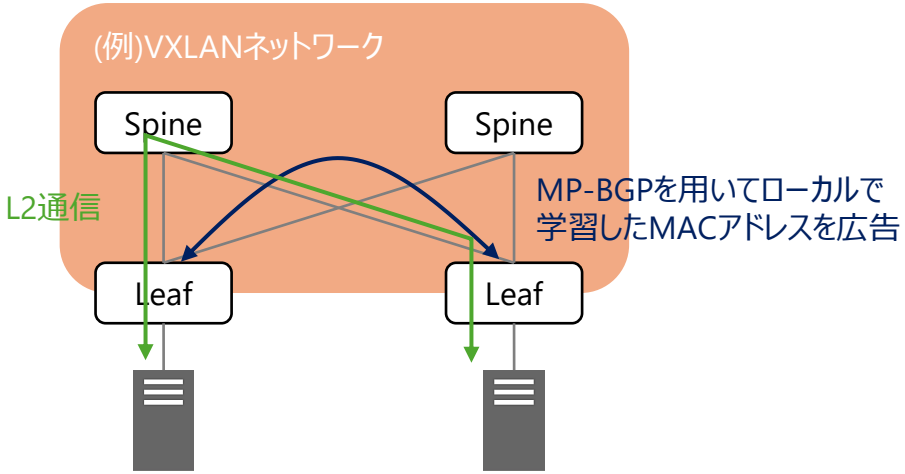
EVPN (概要)

Enterprise　Service Provider　Data Center

サービスプロバイダネットワークのイメージ図



データセンタネットワークのイメージ図



■ 目的

- 物理的に離れた拠点をL2で接続

■ 方法

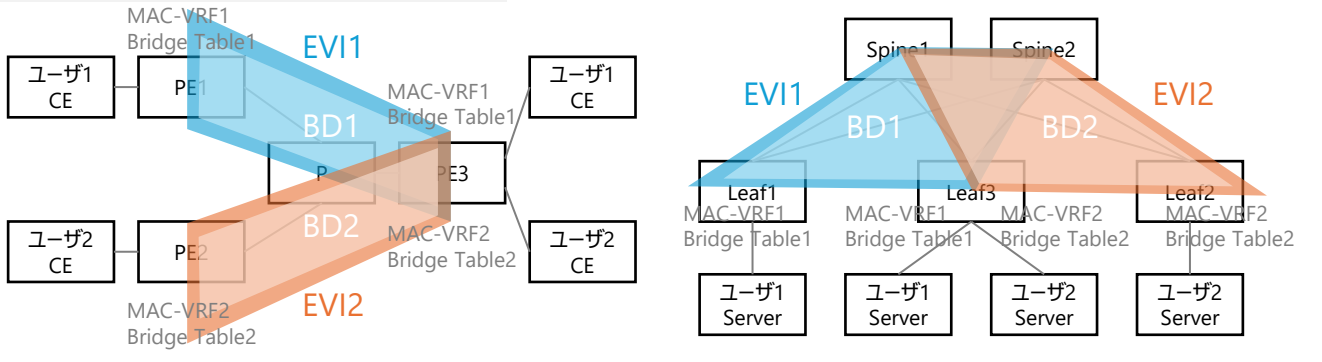
- RFC7432の**EVPN** (Ethernet VPN)
  サービスプロバイダ・データセンタのネットワーク基盤として広く採用
- 「MACアドレスの学習方法」が特徴的
  MP-BGP(Multi Protocol-BGP)を用いたコントロールプレーンでの学習
  (※従来はデータプレーンでの学習)

■ 用語

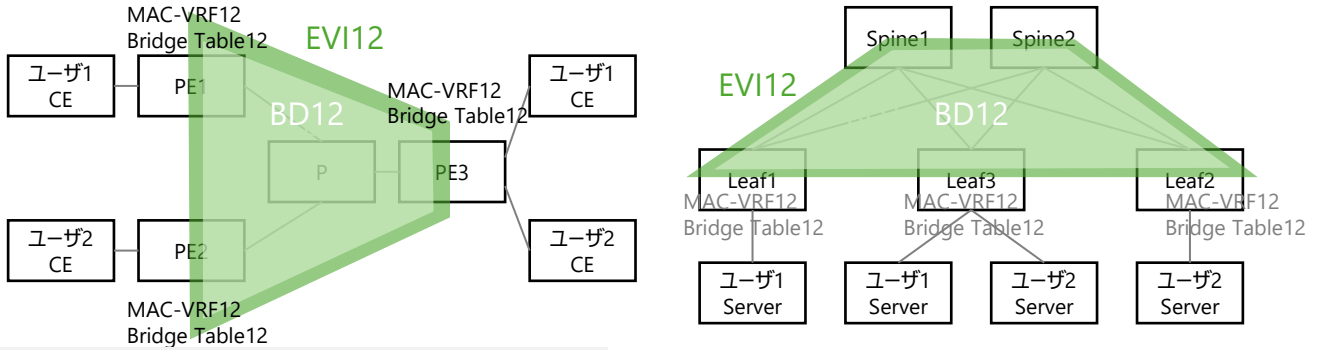
- **CE (Customer Edge)**
  お客様装置
- **PEルータ (Provider Edge)**
  CEと繋がるサービスプロバイダ装置 (Leaf相当)
- **Pルータ**
  PEルータを繋げるサービスプロバイダ装置 (Spine相当)
- **Leaf**
  Serverと繋がるデータセンタ装置 (PEルータ相当)
- **Spine**
  Leafを繋げるデータセンタ装置 (Pルータ相当)

EVPN (サービスインターフェイスとEVI)

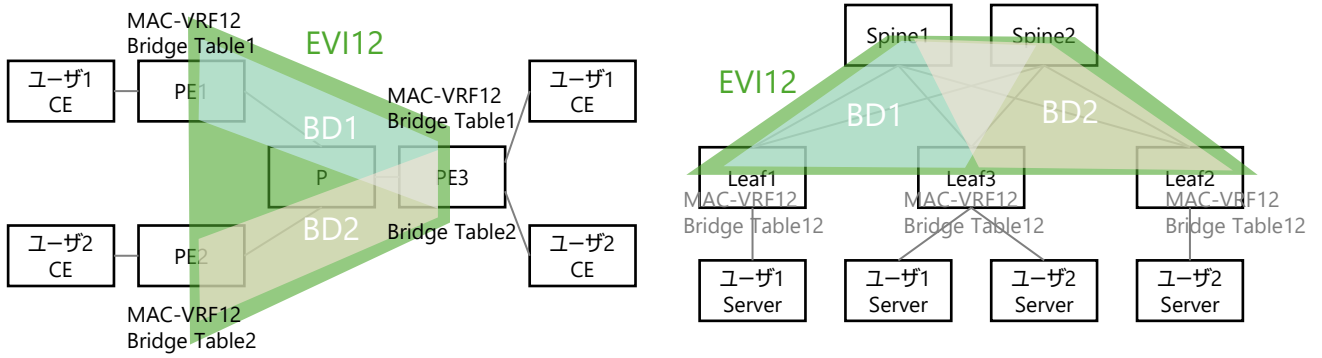
VLAN-Basedサービスインターフェイス



VLAN Bundleサービスインターフェイス



VLAN-Aware Bundleサービスインターフェイス



■ 目的

- 仮想的なL2スイッチの概念を導入する

■ 方法

仮想的なL2スイッチの作り方は3つある

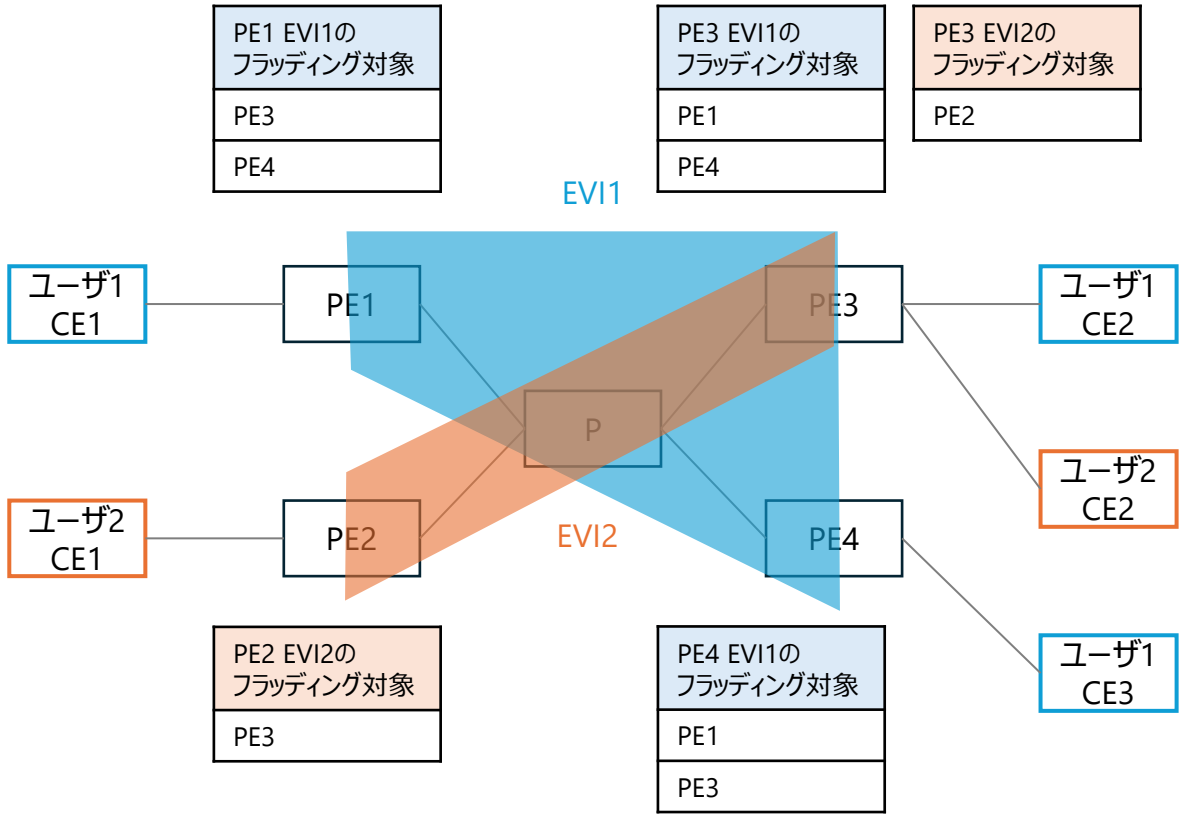
1. **VLAN-Basedサービスインターフェイス**
   ユーザごとに仮想的な物理L2スイッチ(EVI)を作成
2. **VLAN Bundleサービスインターフェイス**
   複数ユーザで仮想的な物理L2スイッチ(EVI)を共有
3. **VLAN-Aware Bundleサービスインターフェイス**
   複数ユーザで仮想的な物理L2スイッチ(EVI)を共有
   ただしVLANでブロードキャストドメインを分割

※以降のページでは一番シンプルなVLAN-Based
サービスインターフェイスで説明

■ 用語

- **EVI (EVPN Instance)**
  仮想的な物理L2スイッチの筐体そのもの
- MAC-VRF
  仮想的な物理L2スイッチ(EVI)ごとのMACアドレステーブル (L2のRIB)
- Bridge Domain
  仮想的な物理L2スイッチ(EVI)におけるブロードキャストドメイン
- Bridge Table
  Bridge DomainごとのMACアドレステーブル (L2のFIB)

EVPN (フラッディング対象のリスト化)



■ **目的**
- BUMフラッディングの最適化

■ **方法**

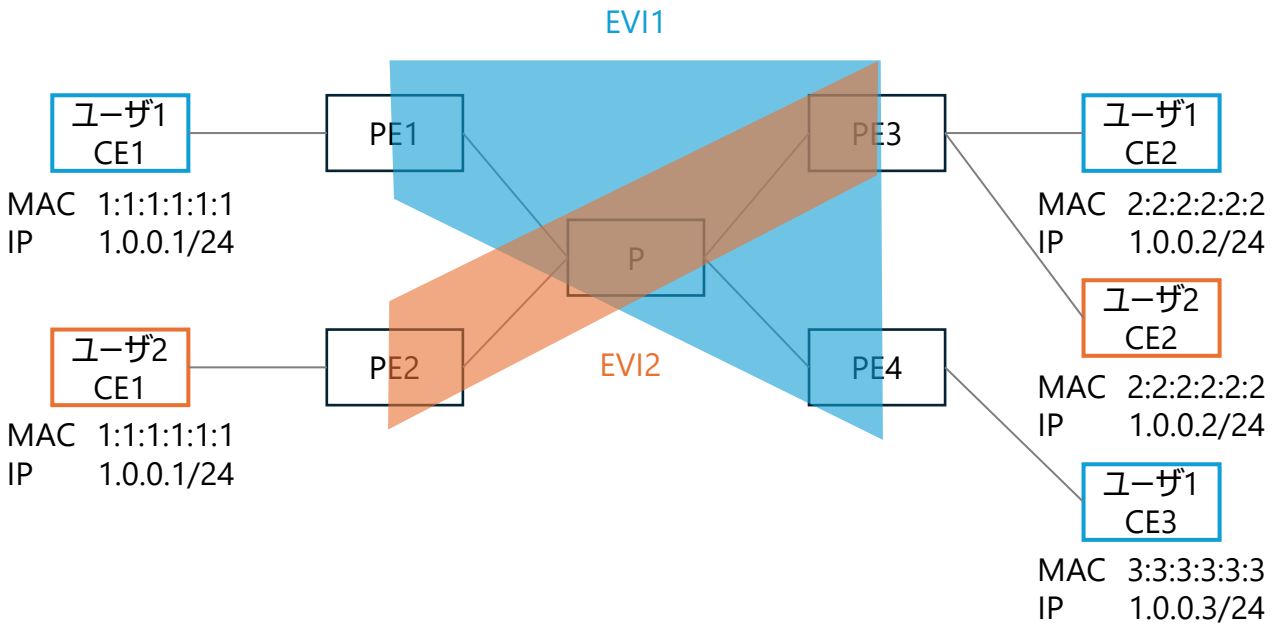
同一のEVIを構成するPE (or Leaf) で情報交換し、
BUMをどのPE (or Leaf) にフラッディングすればよいか学習

関係の無いPE (or Leaf) へのフラッディングを無くせる

■ 具体例
- PE1, PE3, PE4でEVI1を構成
  PE1はPE3とPE4をフラッディング対象として学習
  PE3はPE1とPE4をフラッディング対象として学習
  PE4はPE1とPE3をフラッディング対象として学習

- PE2, PE3でEVI2を構成
  PE2はPE3をフラッディング対象として学習
  PE3はPE2をフラッディング対象として学習

EVPNのフレーム転送 (0/10 構成)



■ **物理構成**

各PEにCEを接続

- PE1ーユーザ1 CE1
- PE2ーユーザ2 CE1
- PE3ーユーザ1 CE2、ユーザ2 CE2
- PE4ーユーザ1 CE3

■ **MACアドレス**

各端末は以下のMACアドレスを持つ

- ユーザ1 CE1ー1:1:1:1:1:1
- ユーザ1 CE2ー2:2:2:2:2:2
- ユーザ1 CE3ー3:3:3:3:3:3
- ユーザ2 CE1ー1:1:1:1:1:1
- ユーザ2 CE2ー2:2:2:2:2:2

■ **IPアドレス**

各端末は以下のIPアドレスを持つ

- ユーザ1 CE1ー1.0.0.1/24
- ユーザ1 CE2ー1.0.0.2/24
- ユーザ1 CE3ー1.0.0.3/24
- ユーザ2 CE1ー1.0.0.1/24
- ユーザ2 CE2ー1.0.0.2/24

EVPNのフレーム転送 (1/11 ユーザ1 CE1がARPリクエスト送出)

ARPリクエスト

宛先 MACアドレス FF:FF:FF:FF:FF:FF	送信元 MACアドレス 1:1:1:1:1:1	タイプ 0x0806	ARP (IPアドレス1.0.0.2 の端末のMACアド レスを解決したい)	FCS

PE1 EVI1のMACアドレステーブル

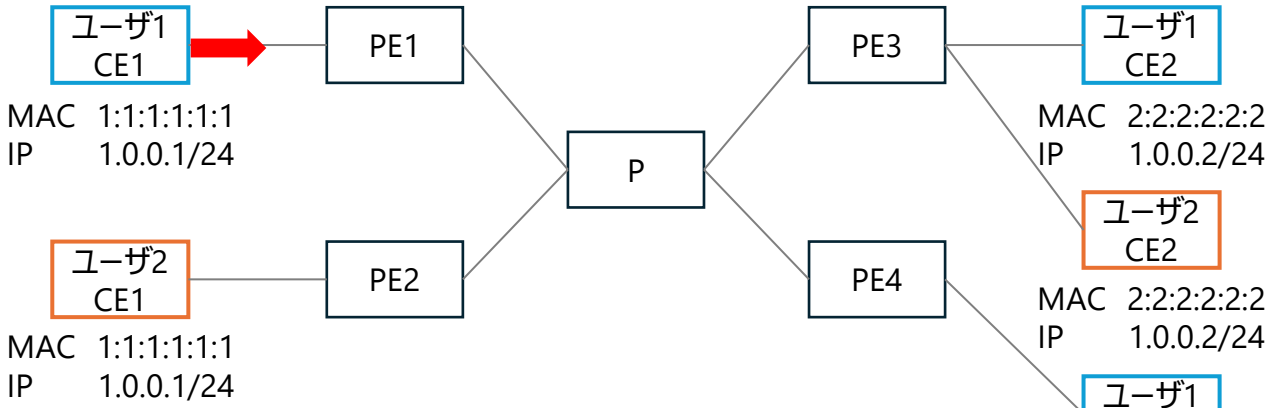
MACアドレス	ネクストホップ

PE3 EVI1のMACアドレステーブル

MACアドレス	ネクストホップ

PE3 EVI2のMACアドレステーブル

MACアドレス	ネクストホップ



PE2 EVI2のMACアドレステーブル

MACアドレス	ネクストホップ

PE4 EVI1のMACアドレステーブル

MACアドレス	ネクストホップ

■ **目的**

ユーザ1 CE1はユーザ1 CE2と通信するために、ユーザ1 CE2のMACアドレスを知りたい

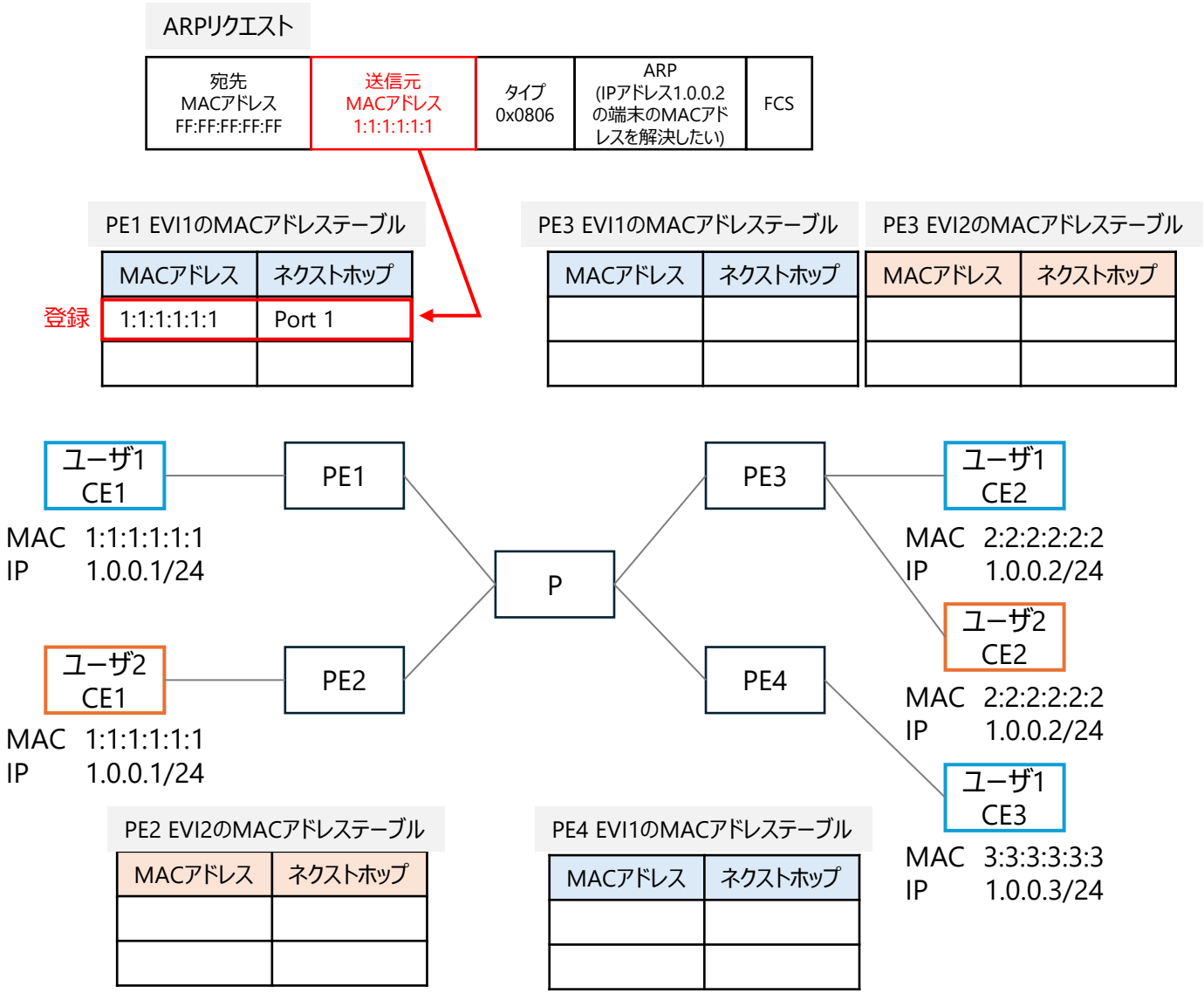
■ **方法**

ユーザ1 CE1からユーザ1 CE2のMACアドレスを解決するためのARPリクエストを送出

- 宛先MACアドレス
  FF:FF:FF:FF:FF:FF(Broadcastアドレス)
- 送信元MACアドレス
  1:1:1:1:1:1
- タイプ
  0x0806
- ARP
  ターゲットとなるIPアドレス1.0.0.2の情報を格納

(※入門編のフレーム転送と同じ)

EVPNのフレーム転送 (2/11 PE1がユーザ1 CE1のMACアドレス学習)



■ **目的**
PE1は宛先MACアドレスをもとに、
フレームを転送できるようになりたい

■ **方法**
受信したフレームの送信元MACアドレスを、
Port情報と共にMACアドレステーブルに登録

(※今後、PE1に宛先MAC1:1:1:1:1:1の
フレームが来たらPort 1に転送できる)

(※入門編のフレーム転送と同じ)

EVPNのフレーム転送 (3/11 PE1がMP-BGPで経路広告)

ARPリクエスト

宛先 MACアドレス FF:FF:FF:FF:FF	送信元 MACアドレス 1:1:1:1:1:1	タイプ 0x0806	ARP (IPアドレス1.0.0.2 の端末のMACアド レスを解決したい)	FCS

PE1 EVI1のMACアドレステーブル

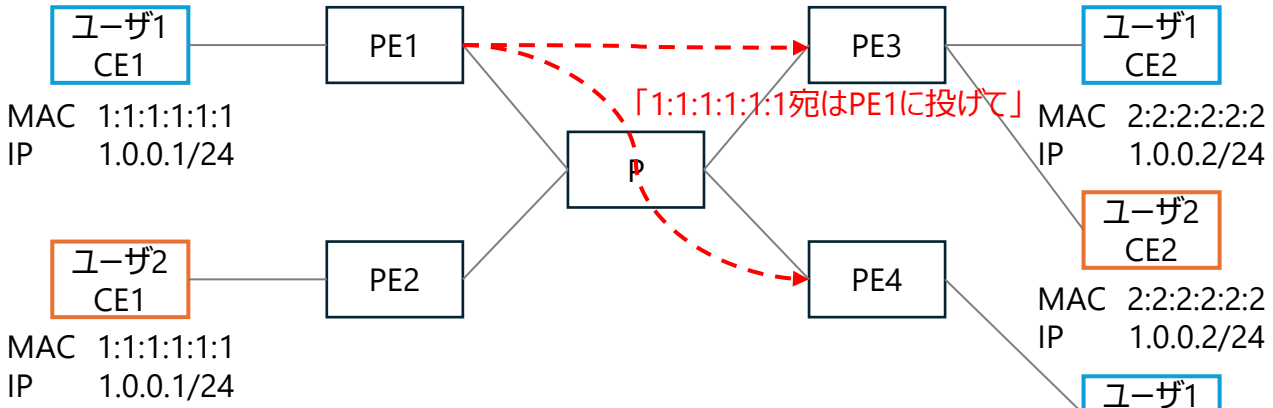
MACアドレス	ネクストホップ
1:1:1:1:1:1	Port 1

PE3 EVI1のMACアドレステーブル

	MACアドレス	ネクストホップ
登録	1:1:1:1:1:1	PE1

PE3 EVI2のMACアドレステーブル

MACアドレス	ネクストホップ



PE2 EVI2のMACアドレステーブル

MACアドレス	ネクストホップ

PE4 EVI1のMACアドレステーブル

	MACアドレス	ネクストホップ
登録	1:1:1:1:1:1	PE1

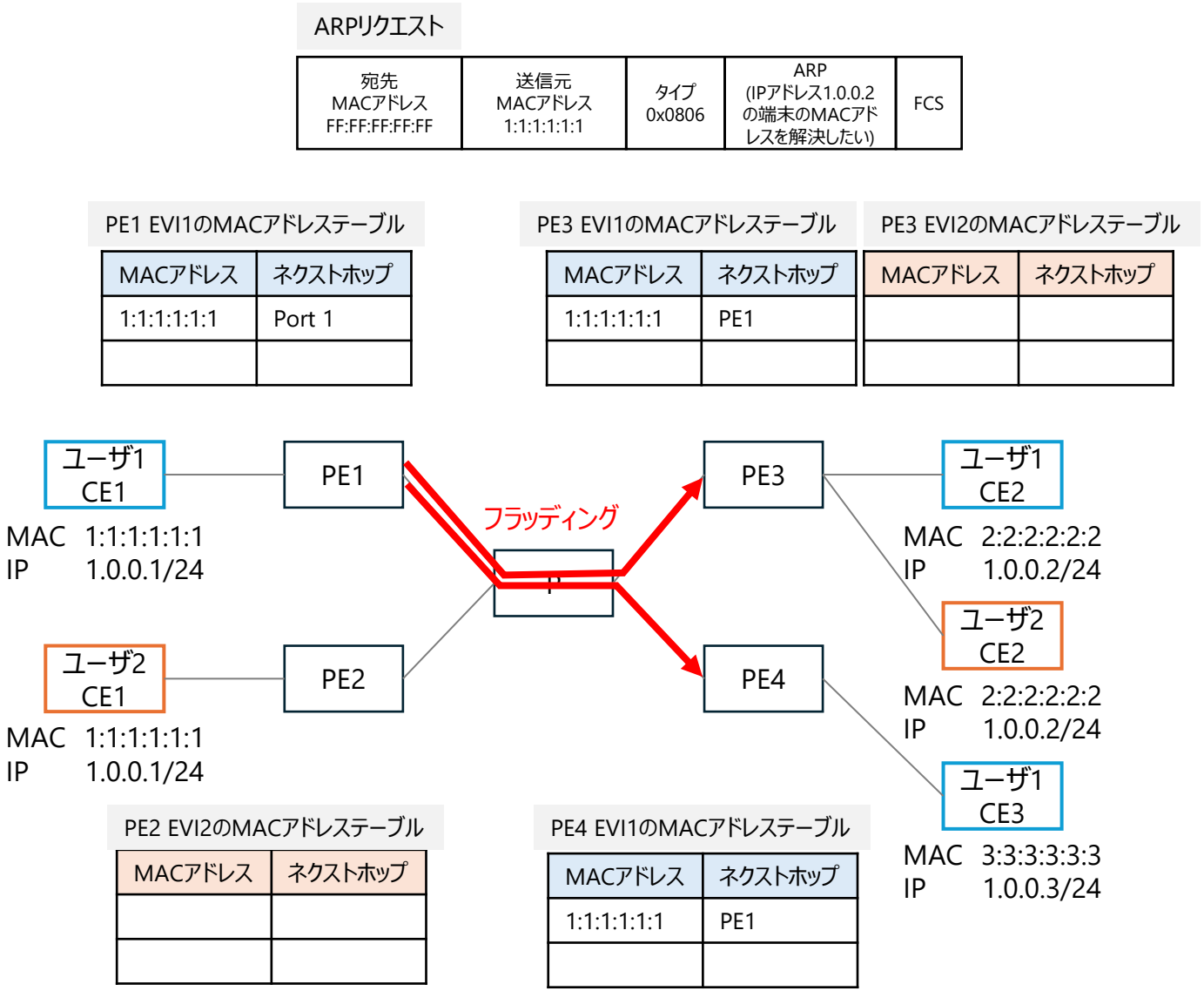
■ **目的**
EVIを構成する他PEでも宛先MACアドレスをもとに、フレームを転送できるようにしたい

■ **方法**
学習したMACアドレスをMP-BGPで経路広告
「1:1:1:1:1:1宛はPE1に投げて」

(※今後、PE3、PE4で宛先MAC1:1:1:1:1:1のフレームを受信するとPE1に転送できる)

(※入門編のフレーム転送と異なる)

EVPNのフレーム転送 (4/11 PE1がPE3, PE4にフラッディング)

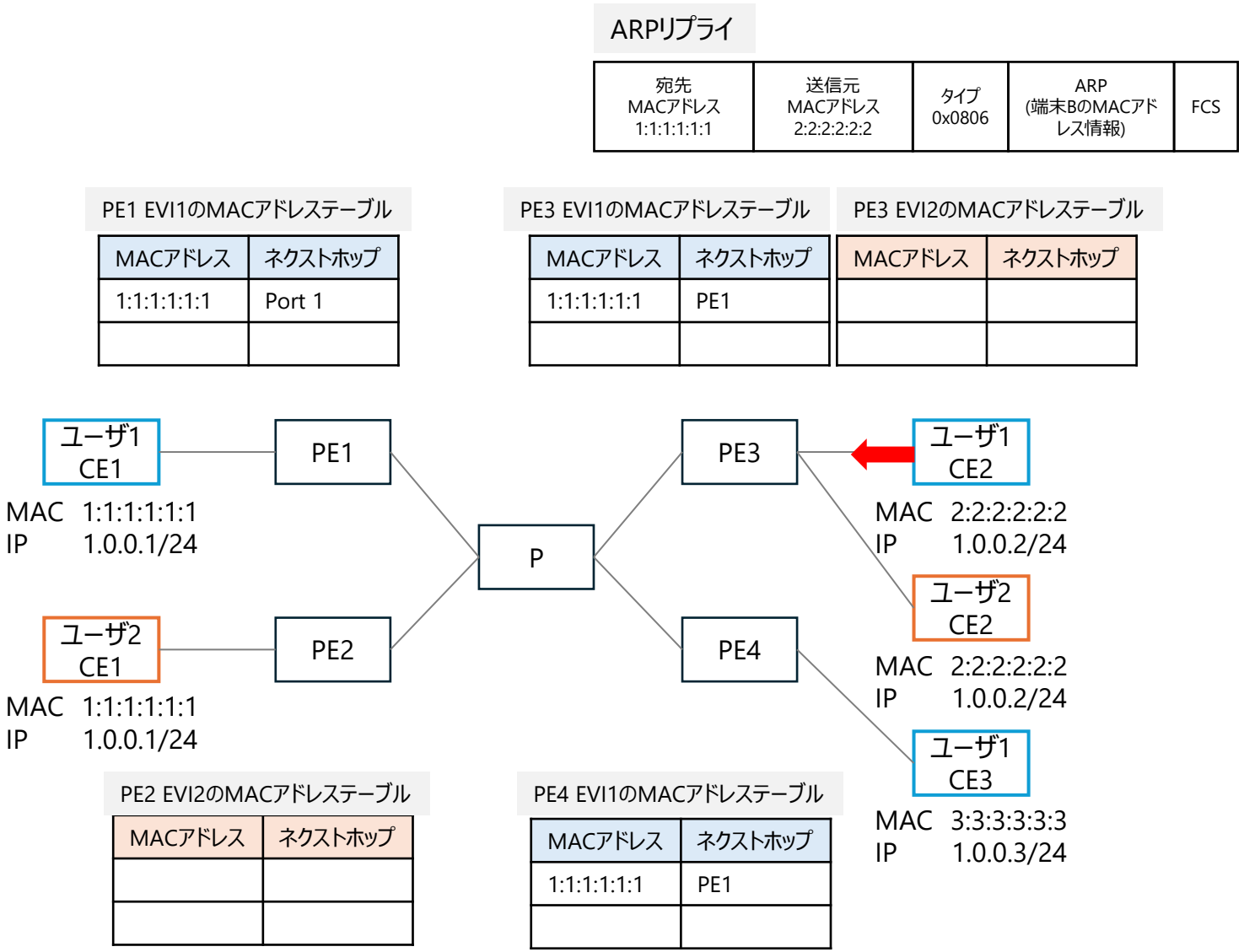


■ **目的**
同一ネットワークの全端末に、IPアドレス1.0.0.2のMACアドレスの聞き込みをしたい

■ **方法**
受信したBUM(ARPリクエスト)を、
フラッディング対象の全てのPEにフラッディング
(※PE2にはフラッディングしない)

(※入門編のフレーム転送と異なる)

EVPNのフレーム転送 (5/11 ユーザ1 CE2がARPリプライ送出)



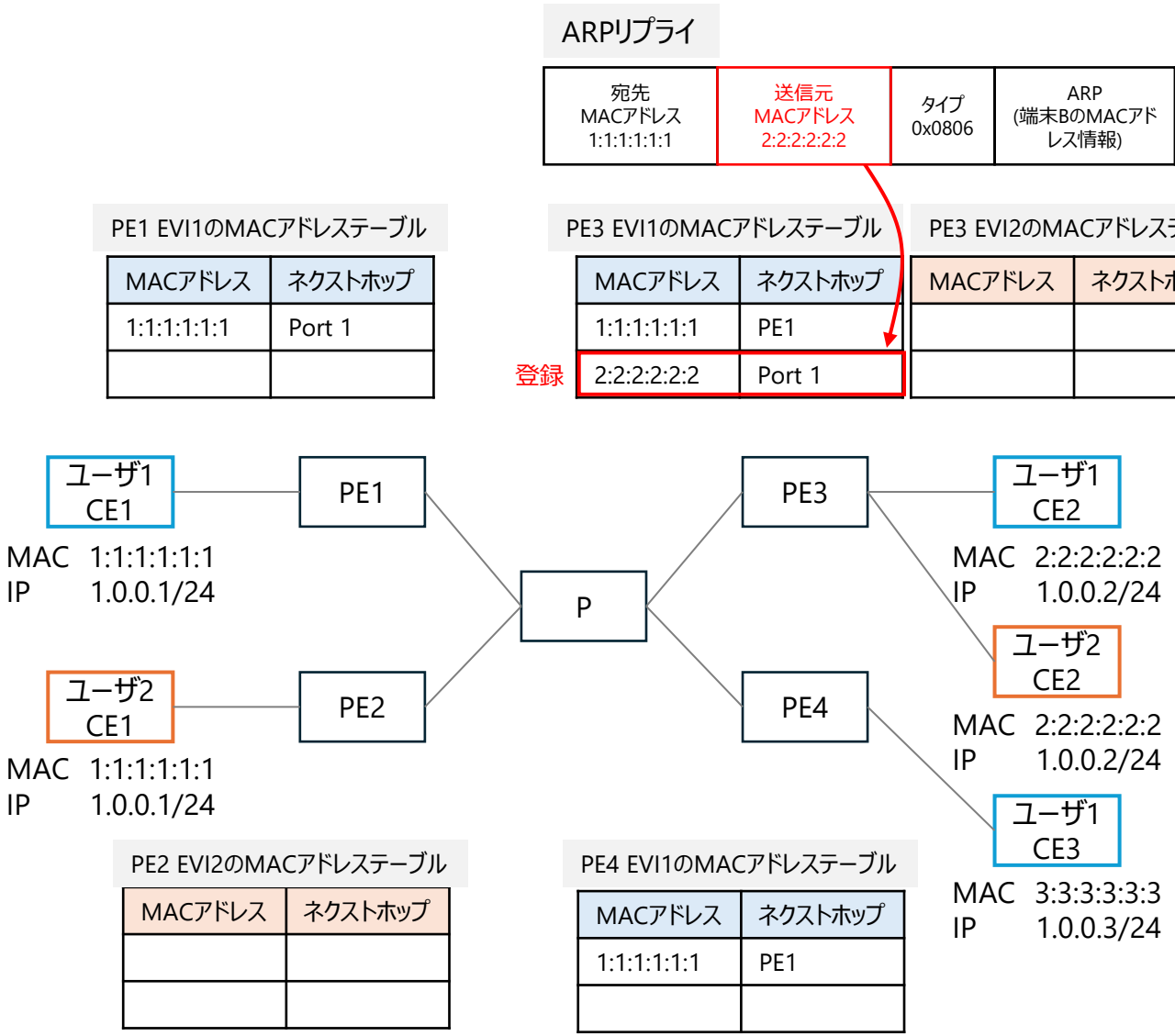
■ **目的**
ユーザ1 CE2はユーザ1 CE1へMACアドレス情報を回答したい

■ **方法**
ユーザ1 CE2からMACアドレス情報を回答するためのARPリプライを送出

- 宛先MACアドレス
  1:1:1:1:1:1
- 送信元MACアドレス
  2:2:2:2:2:2
- タイプ
  0x0806
- ARP
  ユーザ1 CE2のMACアドレス情報を格納

(※入門編のフレーム転送と同じ)

EVPNのフレーム転送 (6/11 PE3がユーザ1 CE1のMACアドレス学習)



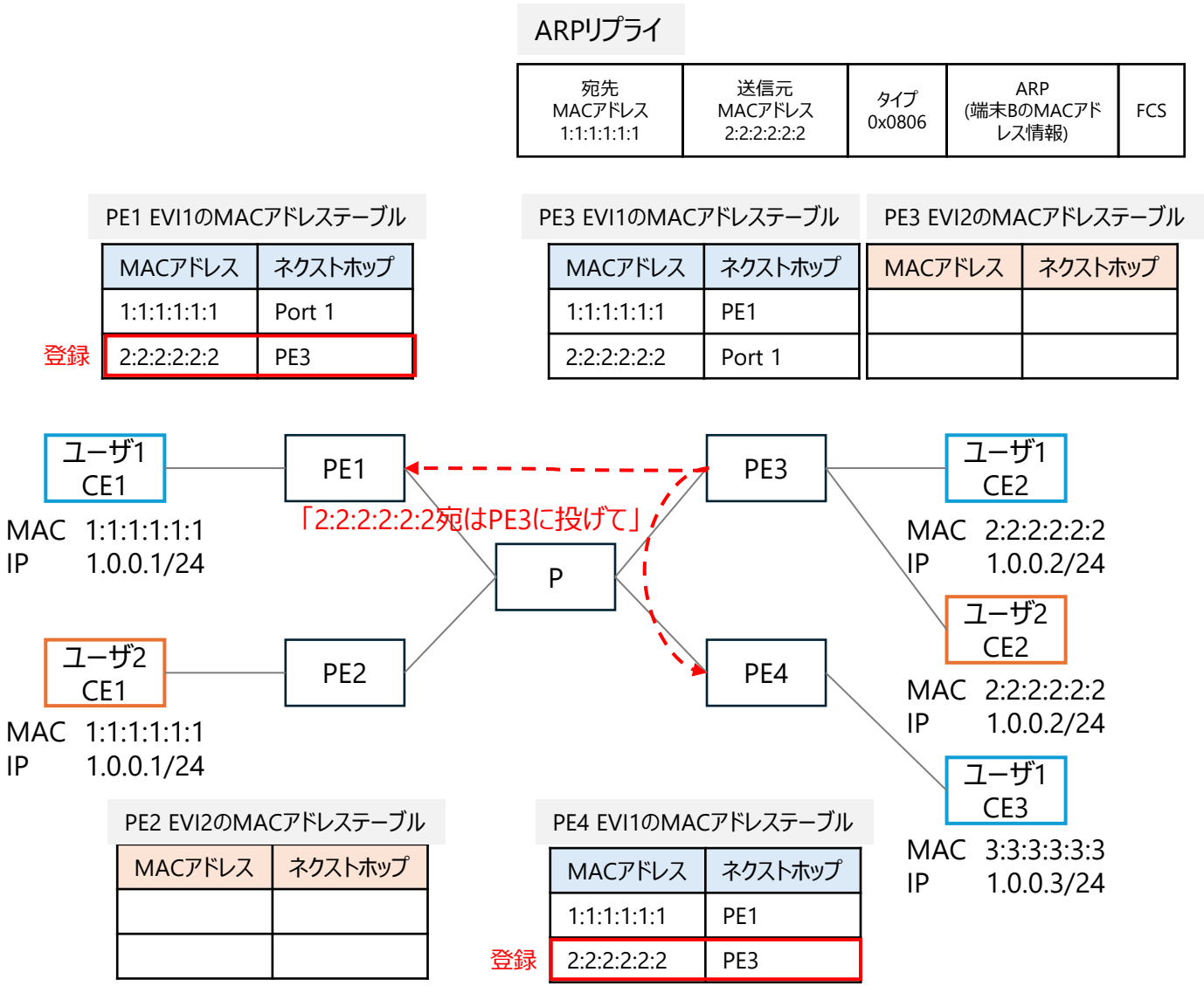
■ **目的**
PE3は宛先MACアドレスをもとに、
フレームを転送できるようになりたい

■ **方法**
受信したフレームの送信元MACアドレスを、
Port情報と共にMACアドレステーブルに登録

(※今後、PE3に宛先MAC2:2:2:2:2:2の
フレームが来たらPort 1に転送できる)

(※入門編のフレーム転送と同じ)

EVPNのフレーム転送 (7/11 PE3がMP-BGPで経路広告)



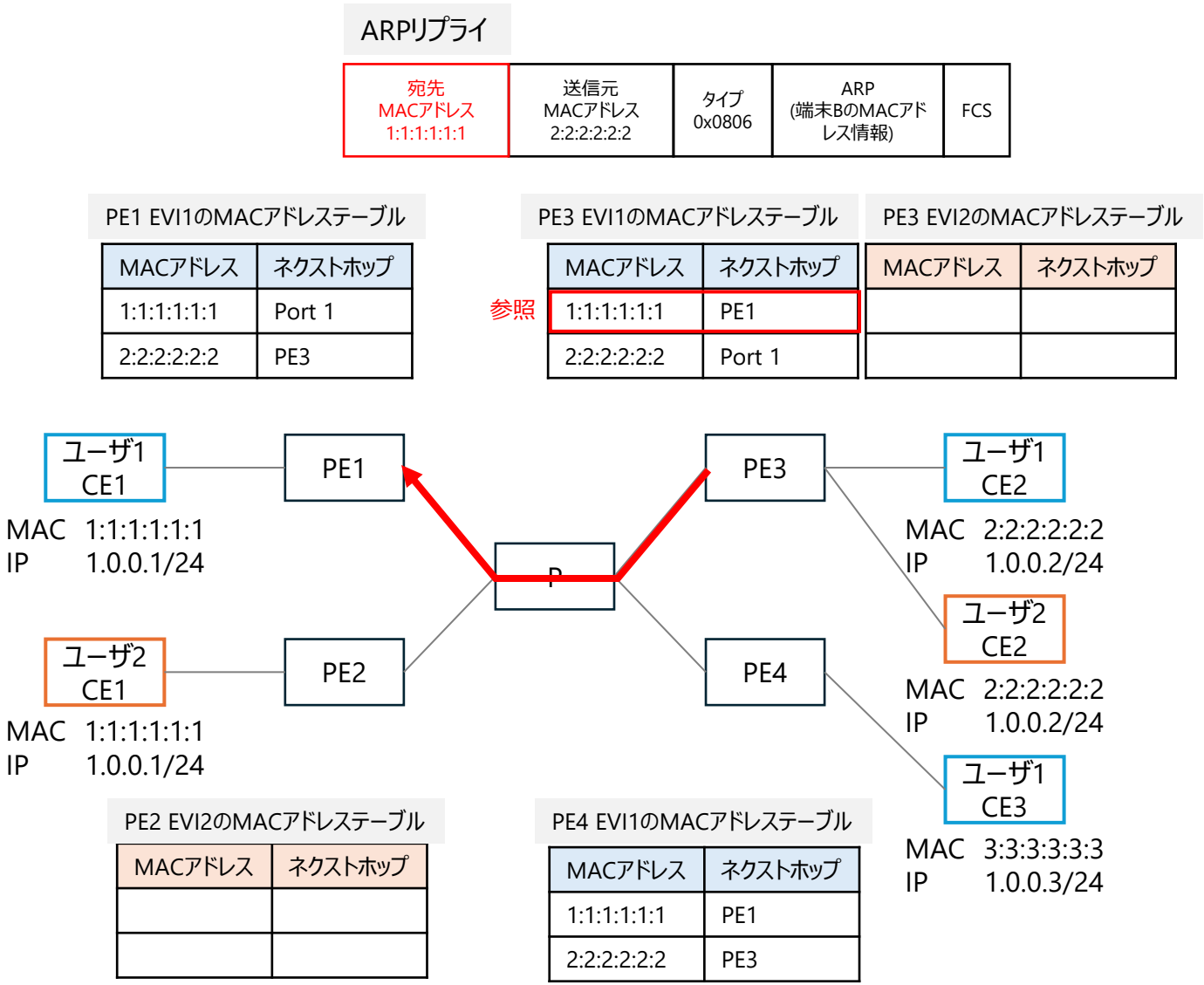
■ **目的**
EVIを構成する他PEでも宛先MACアドレスをもとに、フレームを転送できるようにしたい

■ **方法**
学習したMACアドレスをMP-BGPで経路広告
「2:2:2:2:2:2宛はPE3に投げて」

(※今後、PE1、PE4で宛先MAC2:2:2:2:2:2のフレームを受信するとPE3に転送できる)

(※入門編のフレーム転送と異なる)

EVPNのフレーム転送 (8/11 PE3がARPリプライをPE1に転送)

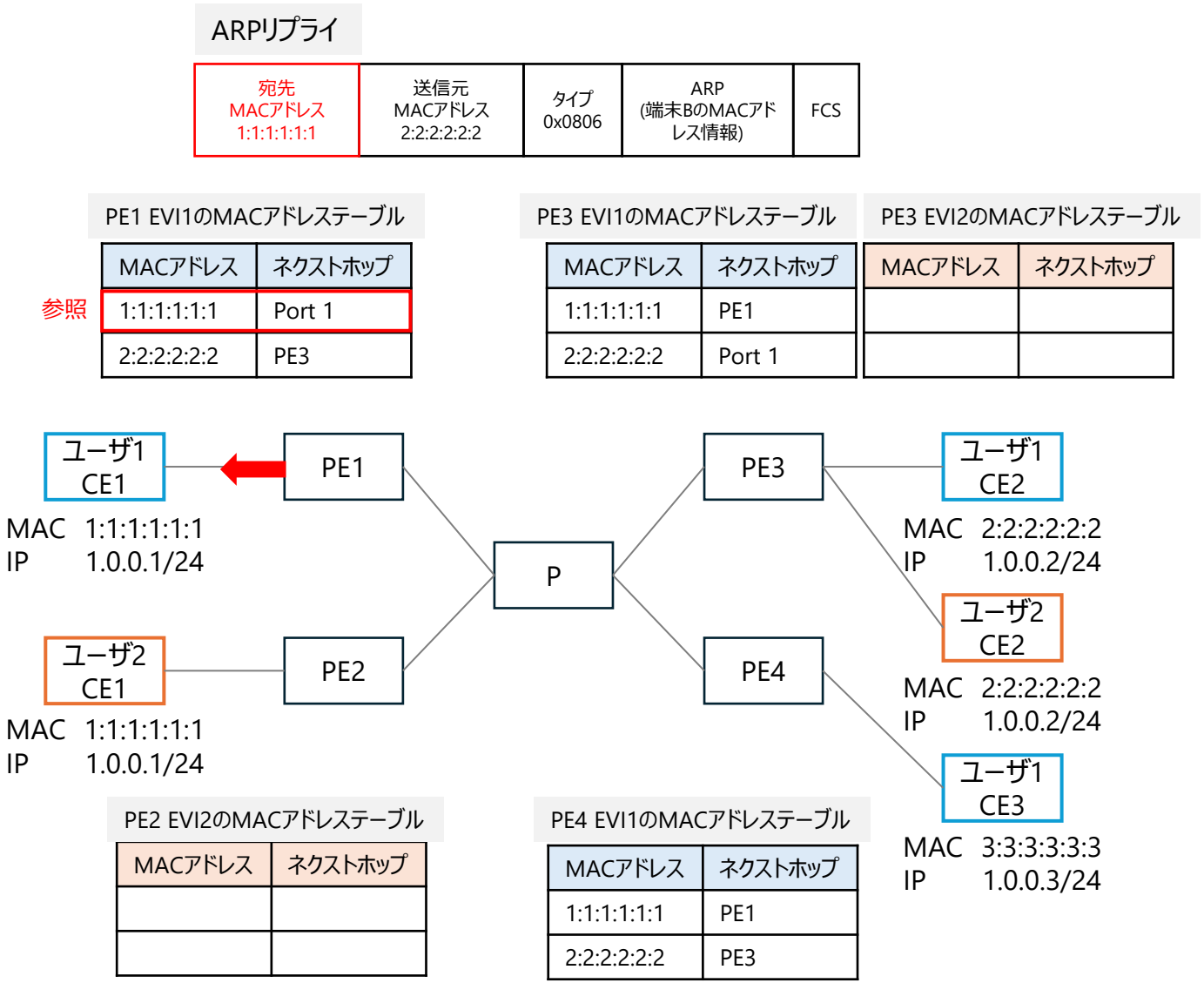


■ **目的**
PE3は宛先MACアドレスをもとに、
ARPリプライを転送したい

■ **方法**
受信したフレームの宛先MACアドレスと
MACアドレステーブルを参照して、
適切なPEに転送する

(※入門編のフレーム転送と同じ)

EVPNのフレーム転送 (9/11 PE1がARPリプライをユーザ1 CE1に転送)



■ **目的**
PE1は宛先MACアドレスをもとに、
ARPリプライを転送したい

■ **方法**
受信したフレームの宛先MACアドレスと
MACアドレステーブルを参照して、
適切なPEに転送する

(※入門編のフレーム転送と同じ)

EVPNのフレーム転送 (10/11 ユーザ1 CE1がユーザ1 CE2のMACアドレスを学習)

ARPリプライ

宛先 MACアドレス 1:1:1:1:1:1	送信元 MACアドレス 2:2:2:2:2:2	タイプ 0x0806	ARP (端末BのMACアドレス情報)	FCS

PE1 EVI1のMACアドレステーブル

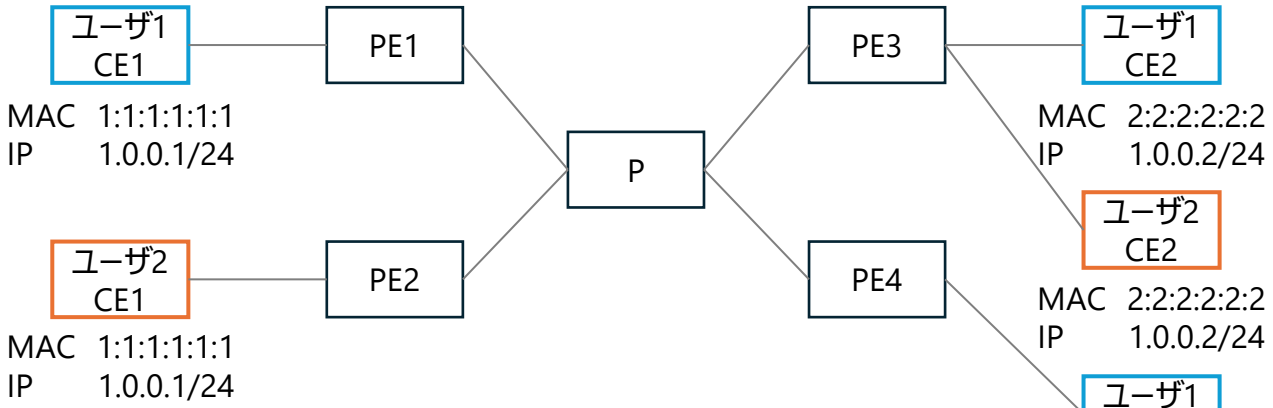
MACアドレス	ネクストホップ
1:1:1:1:1:1	Port 1
2:2:2:2:2:2	PE3

PE3 EVI1のMACアドレステーブル

MACアドレス	ネクストホップ
1:1:1:1:1:1	PE1
2:2:2:2:2:2	Port 1

PE3 EVI2のMACアドレステーブル

MACアドレス	ネクストホップ



PE2 EVI2のMACアドレステーブル

MACアドレス	ネクストホップ

PE4 EVI1のMACアドレステーブル

MACアドレス	ネクストホップ
1:1:1:1:1:1	PE1
2:2:2:2:2:2	PE3

■ **目的**
ユーザ1 CE1はユーザ1 CE2と通信するために、ユーザ1 CE2のMACアドレスを知りたい

■ **方法**
受信したARPリプライからユーザ1 CE2のMACアドレス情報を学習

(※入門編のフレーム転送と同じ)

EVPNのフレーム転送 (11/11 ユーザ1 CE1からユーザ1 CE2へ通信)

PE1 EVI1のMACアドレステーブル

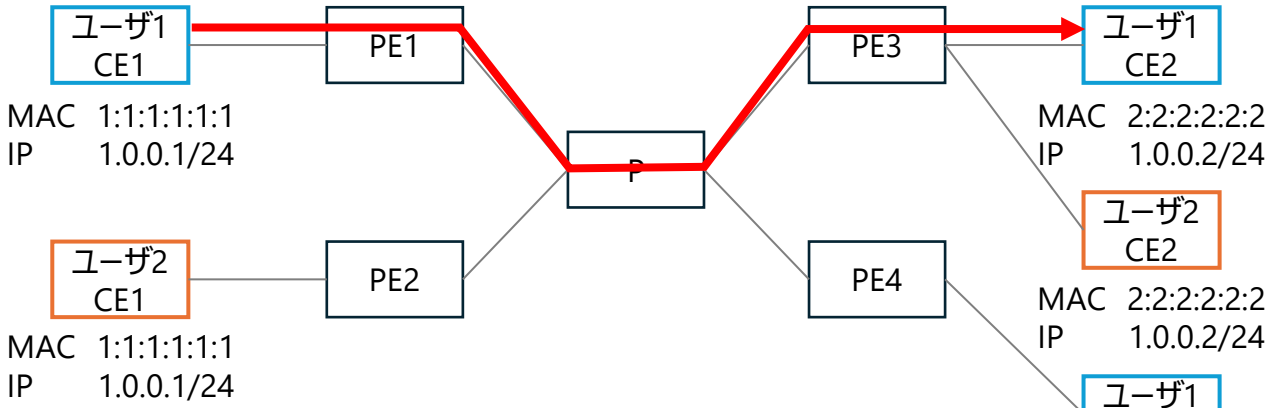
MACアドレス	ネクストホップ
1:1:1:1:1:1	Port 1
2:2:2:2:2:2	PE3

PE3 EVI1のMACアドレステーブル

MACアドレス	ネクストホップ
1:1:1:1:1:1	PE1
2:2:2:2:2:2	Port 1

PE3 EVI2のMACアドレステーブル

MACアドレス	ネクストホップ



PE2 EVI2のMACアドレステーブル

MACアドレス	ネクストホップ

PE4 EVI1のMACアドレステーブル

MACアドレス	ネクストホップ
1:1:1:1:1:1	PE1
2:2:2:2:2:2	PE3

■ **目的**
ユーザ1 CE1はユーザ1 CE2と通信したい

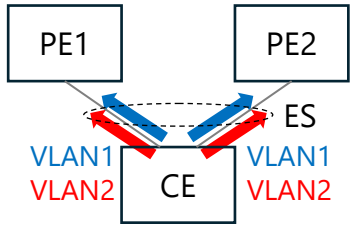
■ **方法**
学習したユーザ1 CE2のMACアドレスを
宛先MACアドレスにしてフレームを送信
(Unicast)

各PEはMACアドレステーブルの情報に基づき
フレームをユーザ1 CE2に転送

(※入門編のフレーム転送と同じ)

EVPN (マルチホーミング)

All-Active

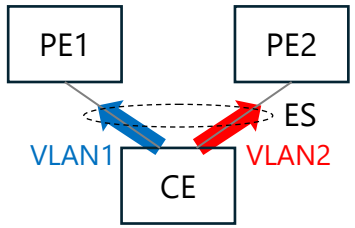


フロー単位のトラフィック分散

	CE→PE1		CE→PE2		PE1→CE		PE2→CE	
	Uni	BUM	Uni	BUM	Uni	BUM	Uni	BUM
VLAN1	○	○	○	○	○	○※	○	×※
VLAN2	○	○	○	○	○	×※	○	○※

※DFとスプリットホライズンに関連

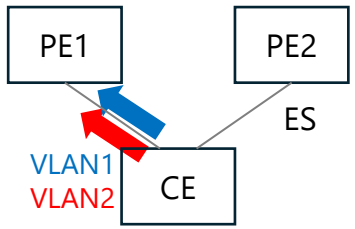
Single-Active



VLAN単位のトラフィック分散

	CE→PE1		CE→PE2		PE1→CE		PE2→CE	
	Uni	BUM	Uni	BUM	Uni	BUM	Uni	BUM
VLAN1	○	○	×	×	○	○	×	×
VLAN2	×	×	○	○	×	×	○	○

Port-Active



片側だけアップ

	CE→PE1		CE→PE2		PE1→CE		PE2→CE	
	Uni	BUM	Uni	BUM	Uni	BUM	Uni	BUM
VLAN1	○	○	×	×	○	○	×	×
VLAN2	○	○	×	×	○	○	×	×

■ 目的

- 耐障害性の更なる向上
  PE (or Leaf)に関する筐体観点の冗長
  ※基礎編のLAG(複数筐体)と同じ

■ 方法

マルチホーミング (以下のようなやり方がある)

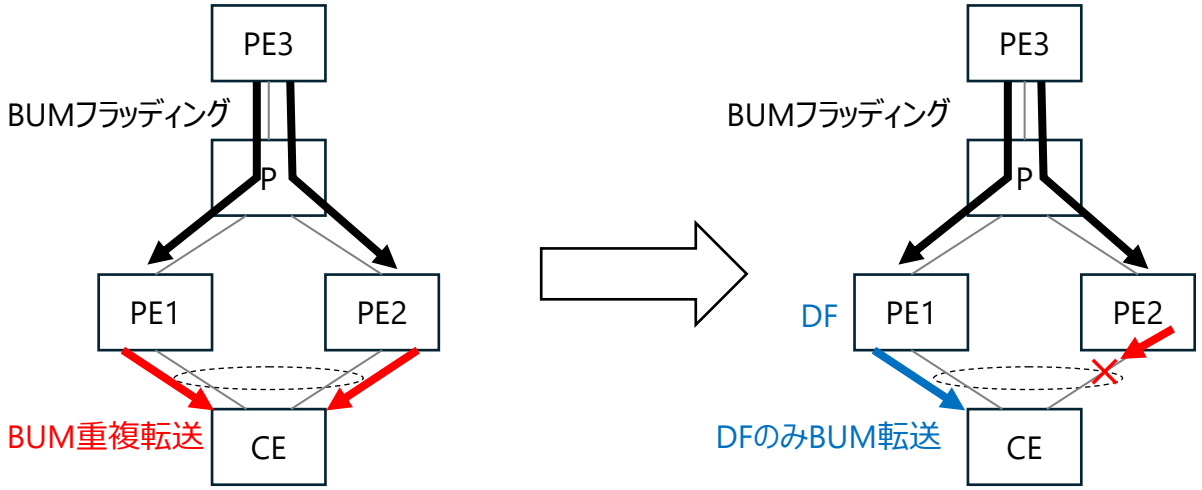
1. **All-Active**
   PE-CE間 (or Leaf-Server間)でLACPによるLAG形成
2. **Single-Active**
   PE-CE間 (or Leaf-Server間)でLACPによるLAG形成
   ただしトラフィックが流れるのは1本の物理リンクのみ
3. **Port-Active**
   1本の物理リンクのみアップ

■ 用語

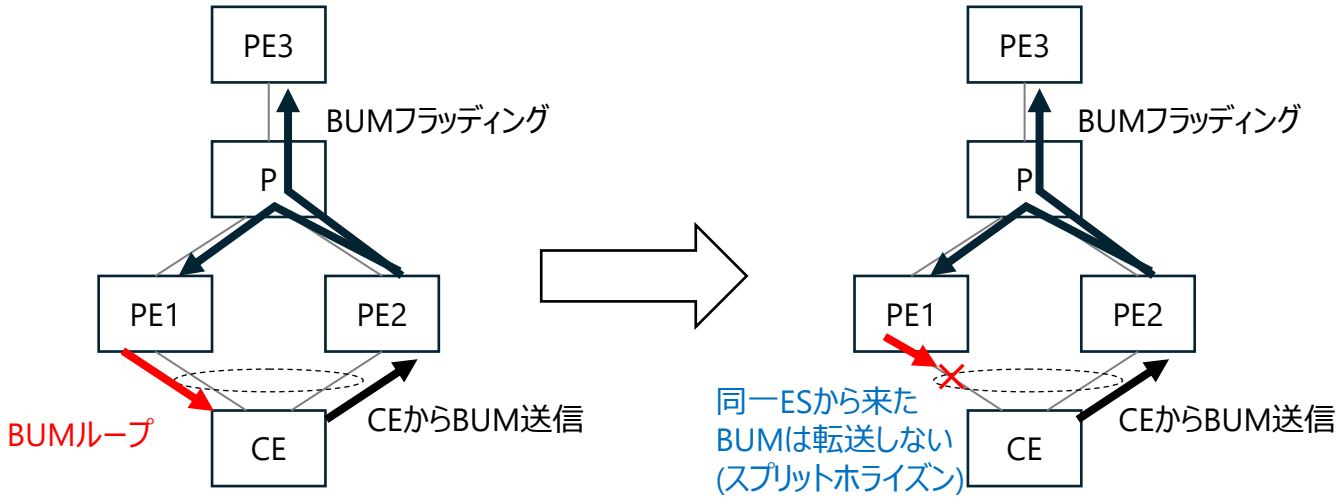
- **ES (Ethernet Segment)**
  PE-CE間 (or Leaf-Server間)の1本以上の物理リンクの集合
  All-Active, Single-ActiveについてはCEから見るとLAGそのもの
- **ESI (Ethernet Segment Identifier)**
  ESを区別するためのID

EVPN (DF選出とスプリットホライズン)

DF選出(BUMの重複転送防止)



スプリットホライズン(BUMのループ防止)



■ **目的**

- All-Activeのマルチホーミングにおいて、BUMの重複転送とループを防ぐ

■ **方法**

- BUM転送を行う代表である**DF (Designated Forwarder)**の選出によるBUMの重複転送防止
- 同一ESから転送されたBUMをCEに送り返さない**スプリットホライズン**によるループ防止

ここから先は時間があればご説明

EVPN (BGP EVPN Route)

Route Type 1

RD (8 octets)
ESI (10 octets)
Ethernet Tag ID (4 octets)
MPLS Label (3 octets)

Route Type 3

RD (8 octets)
ESI (10 octets)
IP Address Length (1 octets)
Originating Router's IP Address (4 or 16 octets)

Route Type 2

RD (8 octets)
ESI (10 octets)
Ethernet Tag ID (4 octets)
MAC Address Length (1 octets)
MAC Address (6 octets)
IP Address Length (1 octet)
IP Address (0, 4, or 16 octets)
MPLS Label 1 (3 octets)
MPLS Label 2 (0 or 3 octets)

Route Type 4

RD (8 octets)
Ethernet Tag ID (4 octets)
IP Address Length (1 octets)
Originating Router's IP Address (4 or 16 octets)

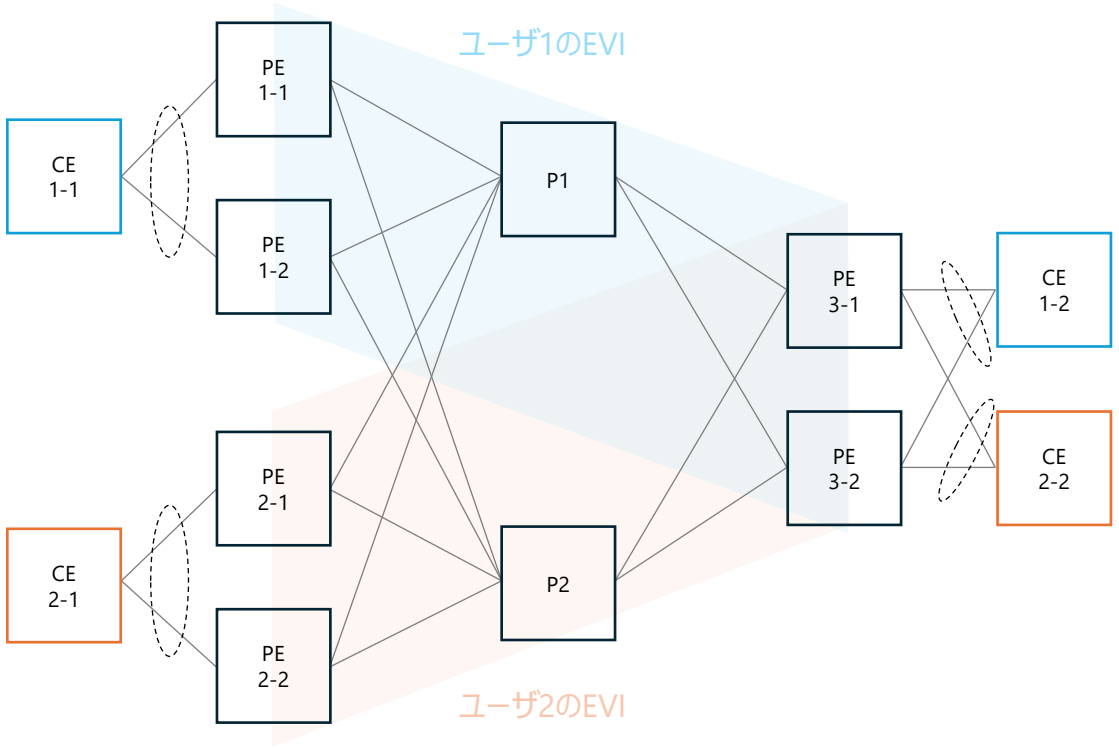
■ **目的**

- MP-BGPを用いて、"誰が" "どこに" "どのように"
  接続されているかPE間 (or Leaf間)で情報交換する

■ **方法**

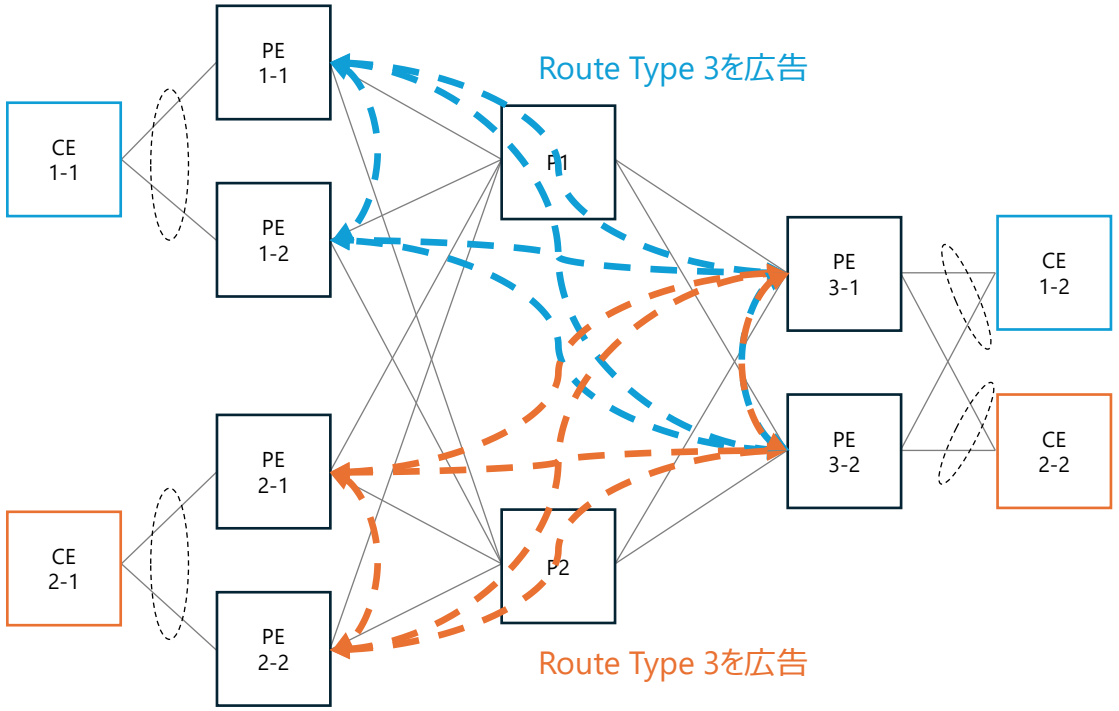
- RFC7432で定義されたBGP EVPN Routeを使う
  - Route Type 1
    Ethernet Auto-discovery Route
    高速切替(per ES)、負荷分散に使用(per EVI)
  - Route Type 2
    MAC/IP Advertisement Route
    MACアドレスの広告に使用
  - Route Type 3
    Inclusive Multicast Ethernet Tag Route
    フラッディング対象のリスト化に使用
  - Route Type 4
    Ethernet Segment Route
    DFの選出に使用

EVPN/(SR)MPLS (0/7 構成)



- 物理構成
  - PE1-1 & PE1-2ーCE1-1
  - PE2-1 & PE2-2ーCE2-1
  - PE3-1 & PE3-2ーCE1-2、CE2-2
- トランスポートネットワーク
  - (SR)MPLS
- サービスインターフェイス
  - VLAN-Based
- マルチホーミング種別
  - All-Active

EVPN/(SR)MPLS (1/7 フラッディング対象のリスト化)



■ 目的
- ユーザ1CEを収容するPE、ユーザ2CEを収容するPE、それぞれでフラッディングの対象をリスト化

■ 方法
- Route Type 3の広告
  - EVI1：PE1-1, PE1-2, PE3-1, PE3-2
  - EVI2：PE2-1, PE2-2, PE3-1, PE3-2
- RT (Route Target) による識別
  EVI毎の識別子RTにより、受信したRoute Type 3がどのEVIに対応するか識別可能
- Originating Router's IP Addressの参照
  Originating Router's IP Addressフィールドを参照し、どのPEにフラッディングすれば良いか学習

Route Type 3のフォーマット

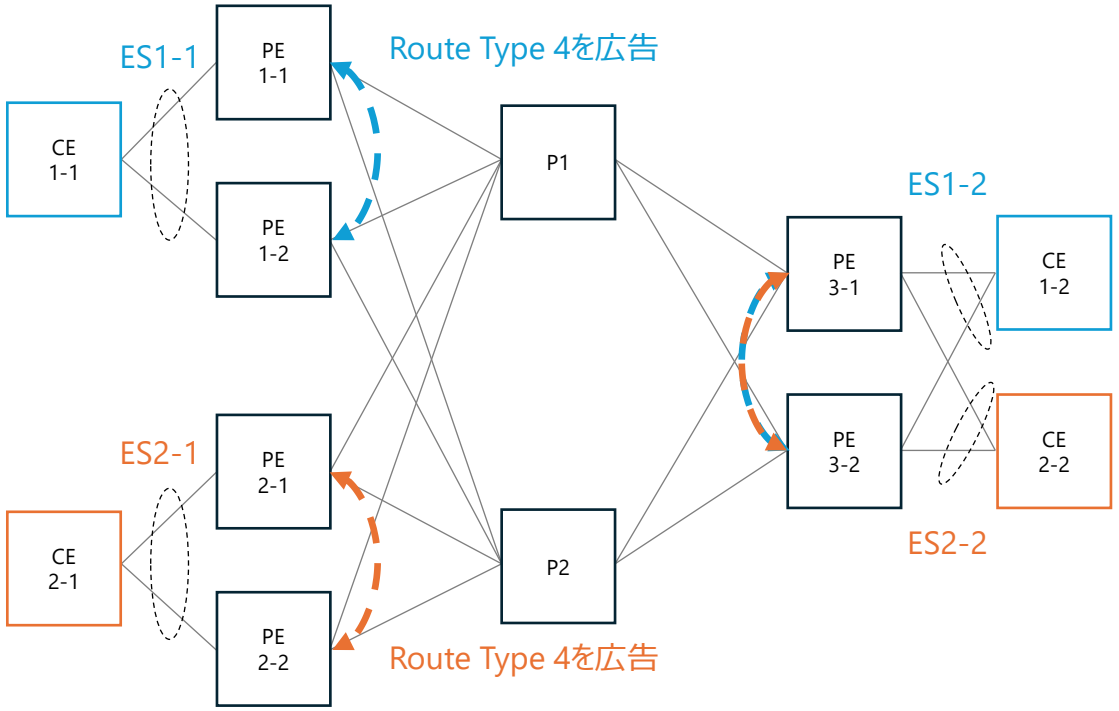
RD (8 octets)
Ethernet Tag ID (4 octets)
IP Address Length (1 octets)
Originating Router's IP Address (4 or 16 octets)

例)PE1-1から広告される
ユーザ1に関するRoute Type 3

1.1.1.1:1 (PE1-1のユーザ1のRD)
0 (VLAN-Basedの場合)
32 (IPv4の場合)
1.1.1.1 (PE1-1のアドレス)

※PMSI Tunnel AttributeでBUM用のラベル情報を広告

EVPN/(SR)MPLS (2/7 DFの選出)



Route Type 4のフォーマット

RD (8 octets)
ESI (10 octets)
IP Address Length (1 octets)
Originating Router's IP Address (4 or 16 octets)

例)PE1-1から広告される Route Type 4

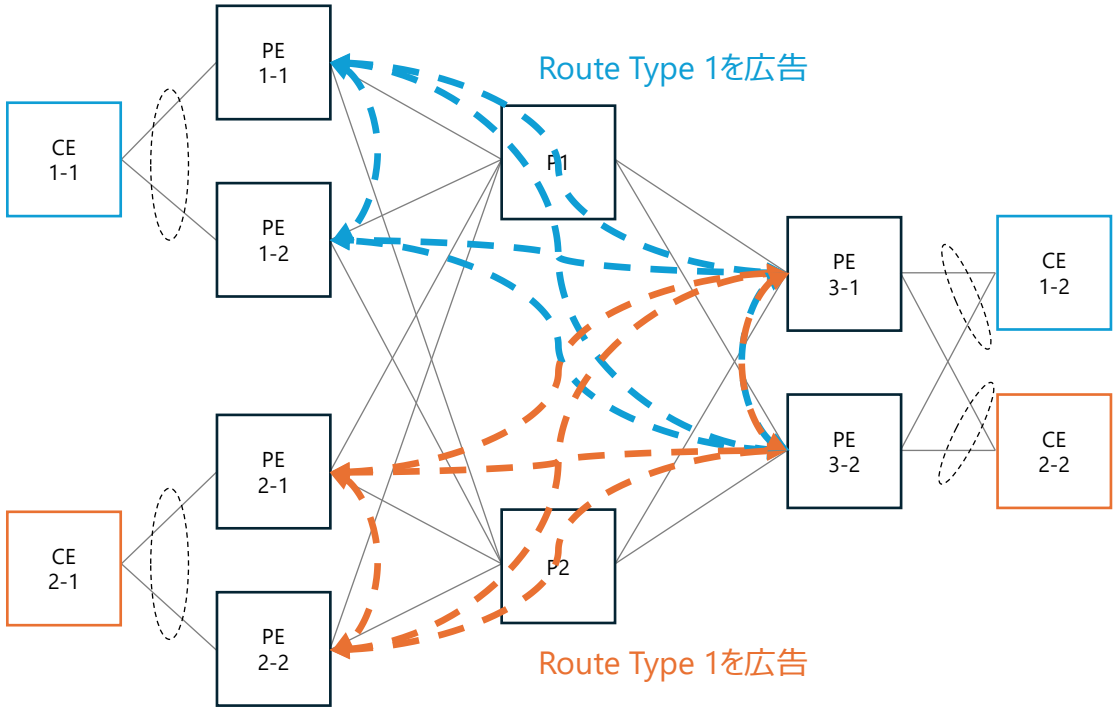
1.1.1.1:1 (PE1-1のユーザ1のRD)
0:1:1:1:1:1:1:1:1:1 (Type 0)
32 (IPv4の場合)
1.1.1.1 (PE1-1のアドレス)

■ 目的
- 各ESにおけるDFの選出

■ 方法
- Route Type 4の広告
  各PEは自身が接続しているESのESIを検出すると、DF選出のためRoute Type 4を広告
- ES-Import Route Targetによる識別
  EVI毎の識別子のRTではなく、
  ES毎の識別子のES-Import Route Targetで識別
  (※Route Type 4はマルチホーミングペアだけで必要な情報のため)
- DFの具体的な決め方
  - 該当VLAN-IDとマルチホーミングPEペア数でmodを取る方法
  - オペレータで固定する方法

  etc

EVPN/(SR)MPLS (3/7 ES情報(物理)の広告)



■ 目的

- CEとPEの物理リンクに障害が発生した際に、Remote-PE側で高速に切り替えを行う

■ 方法

- Route Type 1 (per ES)の広告と撤回
  「あるPEのESが落ちた」という情報を伝えることで、落ちたESに紐づくMAC経路を一気に削除可能
  Mass WithdrawによるFast Convergence

※通常、PEとCEの物理リンクで障害が発生した場合、MAC経路の削除を一つずつ行う必要がある

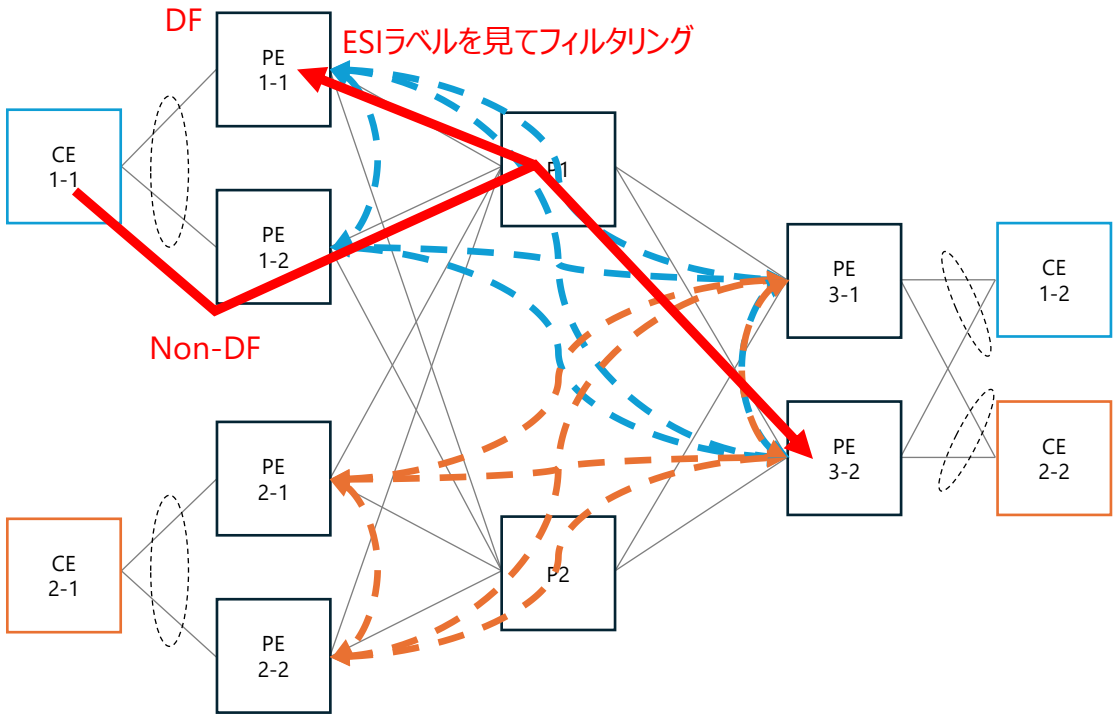
Route Type 1のフォーマット

RD (8 octets)
ESI (10 octets)
Ethernet Tag ID (4 octets)
MPLS Label (3 octets)

例)PE1-1から広告される
Route Type 1 (/ES)

1.1.1.1:1000 (ES固有の値)
0:1:1:1:1:1:1:1:1 (Type 0)
0 (VLAN-Basedの場合)
0 (/ESの場合)

EVPN/(SR)MPLS (4/7 ES情報(物理)の広告 Cont'd)



■ 目的
- 各ESにおけるスプリットホライズン

■ 方法
- Route Type 1 (per ES)と共にESI Label Extended CommunityでESIラベルを広告

  BUMフラッディングの際にESIラベルも付与することで、同一ESに所属するPEでBUMをフィルタリング可能

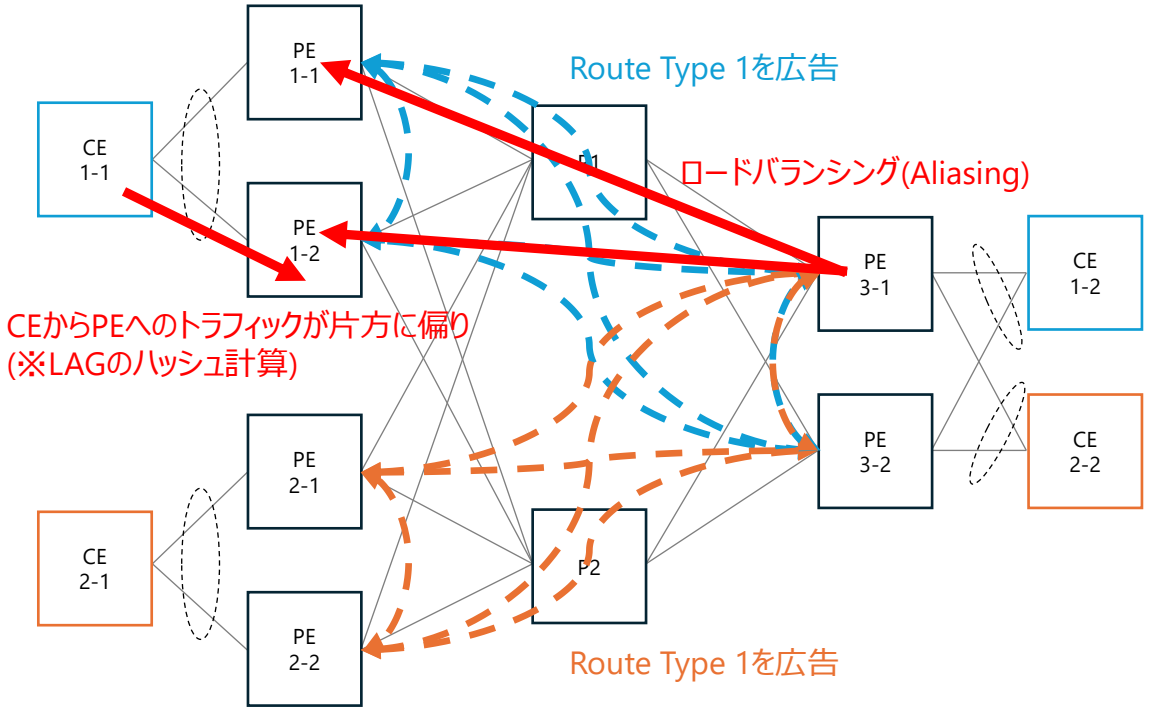
Route Type 1のフォーマット

RD (8 octets)
ESI (10 octets)
Ethernet Tag ID (4 octets)
MPLS Label (3 octets)

例)PE1-1から広告される
Route Type 1 (/ES)

1.1.1.1:1000 (ES固有の値)
0:1:1:1:1:1:1:1:1:1 (Type 0)
0 (VLAN-Basedの場合)
0 (/ESの場合)

EVPN/(SR)MPLS (5/7 ES情報(EVI)の広告)



Route Type 1のフォーマット

RD (8 octets)
ESI (10 octets)
Ethernet Tag ID (4 octets)
MPLS Label (3 octets)

例)PE1-1から広告される
ユーザ1に関するRoute Type 1 (/EVI)

1.1.1.1:1 (PE1-1のユーザ1のRD)
0:1:1:1:1:1:1:1:1:1 (Type 0)
0 (VLAN-Basedの場合)
111111

■ 目的
- ロードバランシング (Remote-PE→マルチホーミングPE)

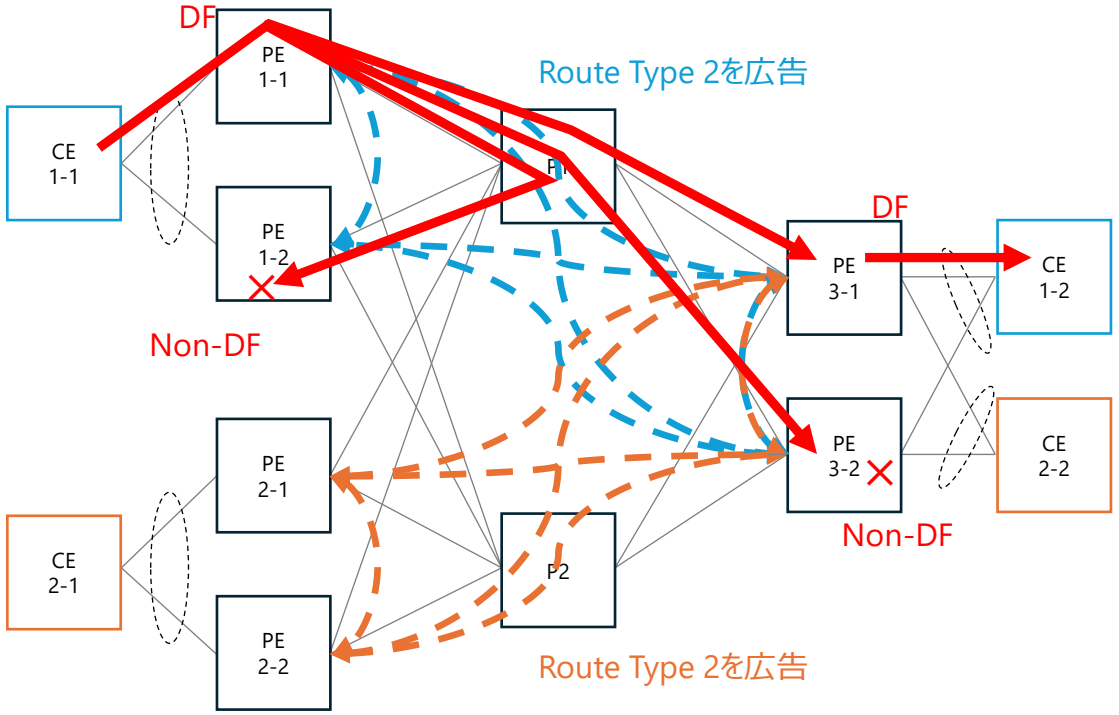
■ 方法
- Route Type 1 (per EVI)の広告

CE→マルチホーミングPEのLAGハッシュ計算によっては片方のPEだけにしかトラフィックが送られず、もう片方のPEではMAC学習できないケースがある

Remote-PEには片方のPEからのみMAC経路が広告され、何も工夫しないと片方のPEに対してのみ転送

Route Type 1 (per EVI)によって、
「あるEVIのある経路について特定のPEからしか学習していなけれど、その特定のPEと同じESに属する別のPEが存在していて、指定されたMPLSラベルを付ければ転送してくれるらしい」
ということをRemote-PEは知ることができる**(Aliasing)**

EVPN/(SR)MPLS (6/7 トラフィック転送)



(例)PE1-1からPE1-2に送るフレーム　PEで付与

DA	SA	Type	転送 Label	ESI Label	BUM Label	DA	SA	Type	ARP	FCS
P1の MAC	PE1-1の MAC		PE1-2の Node-SID	PE1-2の ESI Label	PE1-2の BUM Label	FF:FF:FF:FF:FF:FF	1:1:1:1:1:1			

(例)PE1-1からPE3-1に送るフレーム　PEで付与

DA	SA	Type	転送 Label	BUM Label	DA	SA	Type	ARP	FCS
P1の MAC	PE1-1の MAC		PE3-1の Node-SID	PE3-1の BUM Label	FF:FF:FF:FF:FF:FF	1:1:1:1:1:1			

CE1-1からPE1-1にARPが送られてきたとき、
PE1-1は予め学習した全てのフラッディング先にコピーして転送

送る際は、もともとのARPのフレームに、新たなEtherヘッダと、
MPLSラベル(転送Label, ESI Label, BUM Label)を付与

フラッディングされてきたARP(BUM)について、
Non-DFであるPEはCEに対して転送しない

並行してARPの送信元MACアドレスを学習して、
Route Type 2を他のPEに広告

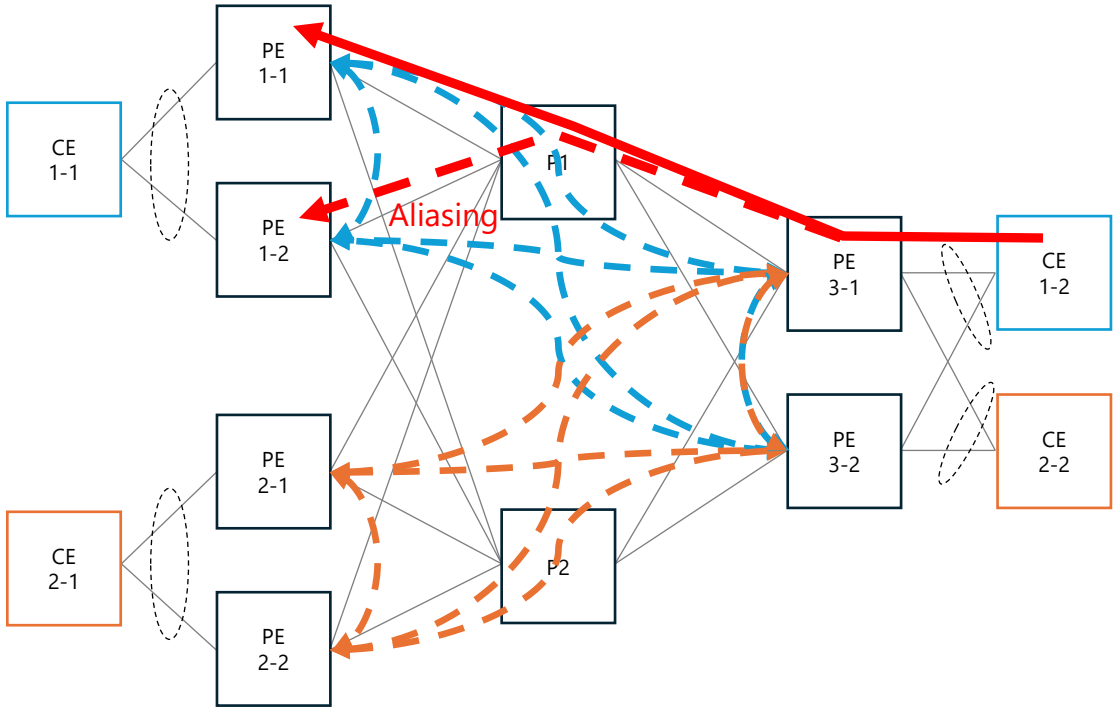
Route Type 2のフォーマット

RD (8 octets)
ESI (10 octets)
Ethernet Tag ID (4 octets)
MAC Address Length (1 octets)
MAC Address (6 octets)
IP Address Length (1 octet)
IP Address (0, 4, or 16 octets)
MPLS Label 1 (3 octets)
MPLS Label 2 (0 or 3 octets)

例)PE1-1から広告される
ユーザ1に関するRoute Type 2

1.1.1.1:1 (PE1-1のユーザ1のRD)
0:1:1:1:1:1:1:1:1:1 (Type 0)
0 (VLAN-Basedの場合)
48
1:1:1:1:1:1
0 (IPアドレスに関する情報無し)
-
16001
-

EVPN/(SR)MPLS (7/7 トラフィック転送)



(例)PE3-1からPE1-1に送るフレーム

DA	SA	Type	転送 Label	Service Label	DA	SA	Type	ARP	FCS
P1の MAC	PE3-1の MAC		PE1-1の Node-SID	PE1-1の Service Label	1:1:1:1:1:1	2:2:2:2:2:2			

(例)PE3-1からPE1-2に送るフレーム　PEで付与

DA	SA	Type	転送 Label	Service Label	DA	SA	Type	ARP	FCS
P1の MAC	PE3-1の MAC		PE1-2の Node-SID	PE1-2の Aliasing Label	1:1:1:1:1:1	2:2:2:2:2:2			

CE1-2からCE1-1にユニキャストの通信を行うとき、PE3-1はPE1-1からRoute Type 2でMACアドレスを学習しているので、フラッディングの必要が無く、PE1-1を狙って転送

またPE1-2に対してもRoute Type 1(per EVI)で転送できることが分かっているので、Aliasing用のLabelを付けて転送可能

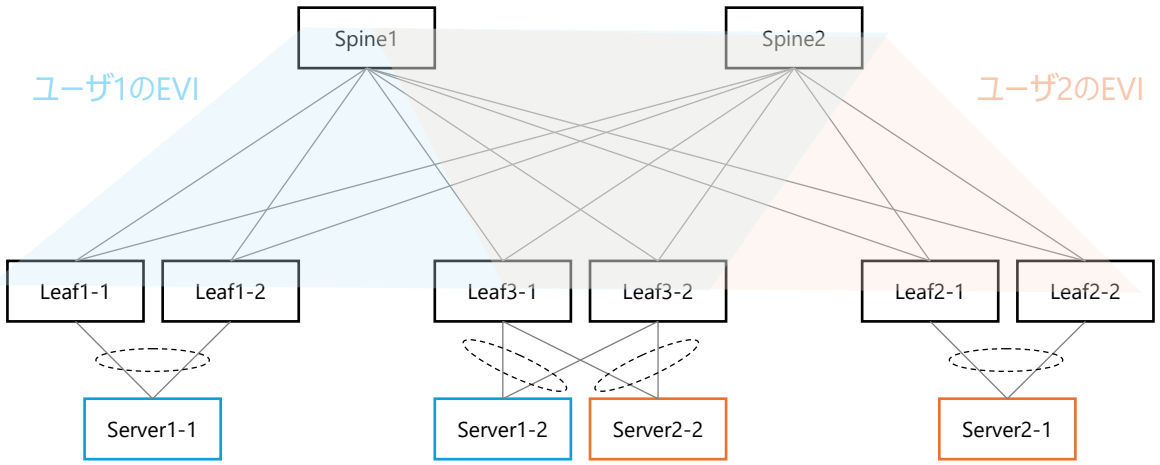
Route Type 2のフォーマット

RD (8 octets)
ESI (10 octets)
Ethernet Tag ID (4 octets)
MAC Address Length (1 octets)
MAC Address (6 octets)
IP Address Length (1 octet)
IP Address (0, 4, or 16 octets)
MPLS Label 1 (3 octets)
MPLS Label 2 (0 or 3 octets)

例)PE1-1から広告されるユーザ1に関するRoute Type 2

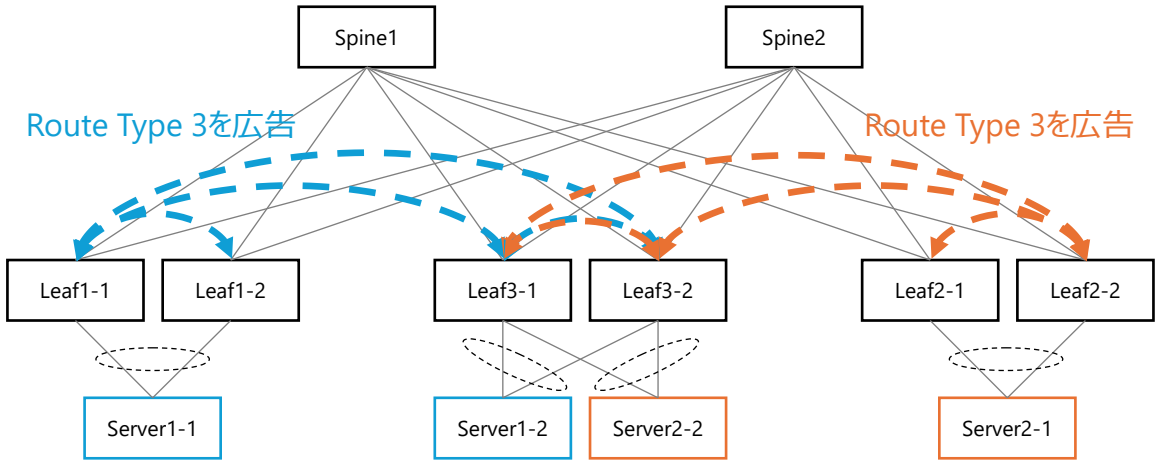
1.1.1.1:1 (PE1-1のユーザ1のRD)
0:1:1:1:1:1:1:1:1:1 (Type 0)
0 (VLAN-Basedの場合)
48
1:1:1:1:1:1
0 (IPアドレスに関する情報無し)
-
16001
-

EVPN/VXLAN (0/6 構成)



- 物理構成
  - Leaf1-1 & Leaf1-2ーServer1-1
  - Leaf2-1 & Leaf2-2ーServer2-1
  - Leaf3-1 & Leaf3-2ーServer1-2、Server2-2
- トランスポートネットワーク
  - VXLAN
- サービスインターフェイス
  - VLAN-Based
- マルチホーミング種別
  - All-Active

EVPN/VXLAN (1/6 フラッディング対象の学習)



■ 目的

- ユーザ1Serverを収容するLeaf、
  ユーザ2Serverを収容するLeaf、
  それぞれでフラッディングの対象をリスト化

■ 方法

- Route Type 3の広告
  - EVI1：Leaf1-1, Leaf1-2, Leaf3-1, Leaf3-2
  - EVI2：Leaf2-1, Leaf2-2, Leaf3-1, Leaf3-2
- RT (Route Target) による識別
  EVI毎の識別子RTにより、受信したRoute Type 3が
  どのEVIに対応するか識別可能
- Originating Router's IP Addressの参照
  Originating Router's IP Addressフィールドを参照し、
  どのPEにフラッディングすれば良いか学習

Route Type 3のフォーマット

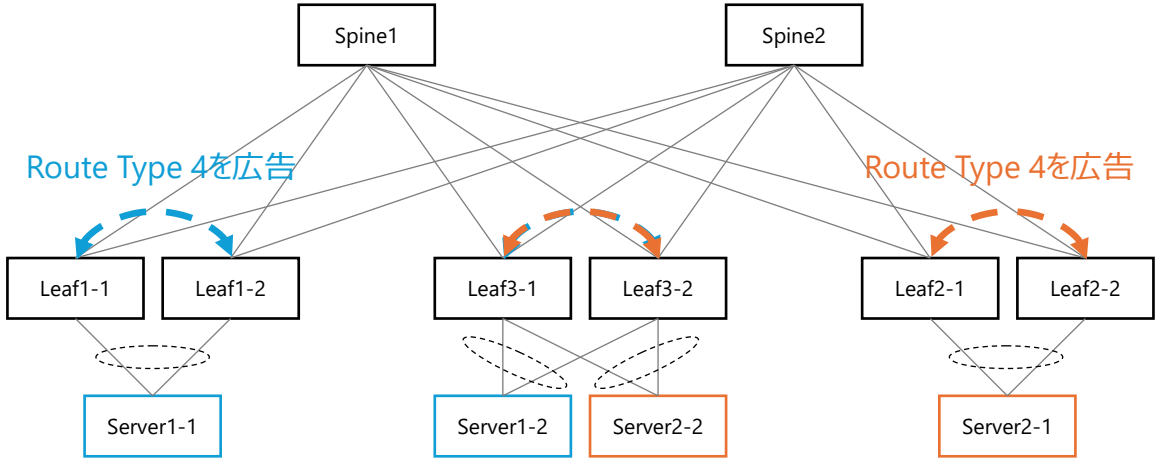
RD (8 octets)
Ethernet Tag ID (4 octets)
IP Address Length (1 octets)
Originating Router's IP Address (4 or 16 octets)

例)Leaf1-1から広告される
ユーザ1に関するRoute Type 3

1.1.1.1:1 (Leaf1-1のユーザ1のRD)
0 (VLAN-Basedの場合)
32 (IPv4の場合)
1.1.1.1 (Leaf1-1のアドレス)

※PMSI Tunnel Attributeで
VNI情報を広告

EVPN/VXLAN (2/6 DFの選出)



Route Type 4のフォーマット

RD (8 octets)
ESI (10 octets)
IP Address Length (1 octets)
Originating Router's IP Address (4 or 16 octets)

例)Leaf1-1から広告される Route Type 4

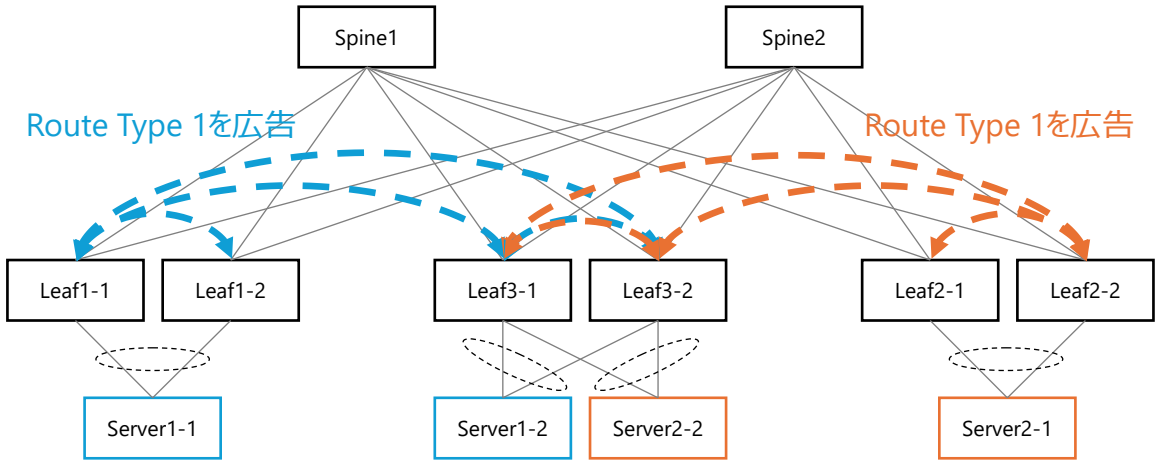
1.1.1.1:1 (Leaf1-1のユーザ1のRD)
0:1:1:1:1:1:1:1:1:1 (Type 0)
32 (IPv4の場合)
1.1.1.1 (Leaf1-1のアドレス)

■ 目的
- 各ESにおけるDFの選出

■ 方法
- Route Type 4の広告
  各Leafは自身が接続しているESのESIを検出すると、DF選出のためRoute Type 4を広告
- ES-Import Route Targetによる識別
  EVI毎の識別子のRTではなく、
  ES毎の識別子のES-Import Route Targetで識別
  (※Route Type 4はマルチホーミングペアだけで必要な情報のため)
- DFの具体的な決め方
  - 該当VLAN-IDとマルチホーミングLeafペア数でmodを取る方法
  - オペレータで固定する方法

  etc

EVPN/VXLAN (3/6 ES情報(物理)の広告)



Route Type 1のフォーマット

RD (8 octets)
ESI (10 octets)
Ethernet Tag ID (4 octets)
VNI field (3 octets)

例)Leaf1-1から広告される
Route Type 1 (/ES)

1.1.1.1:1000 (ES固有の値)
0:1:1:1:1:1:1:1:1:1 (Type 0)
0 (VLAN-Basedの場合)
0 (/ESの場合)

■ 目的

- ServerとLeafの物理リンクに障害が発生した際に、Remote-Leaf側で高速に切り替えを行う

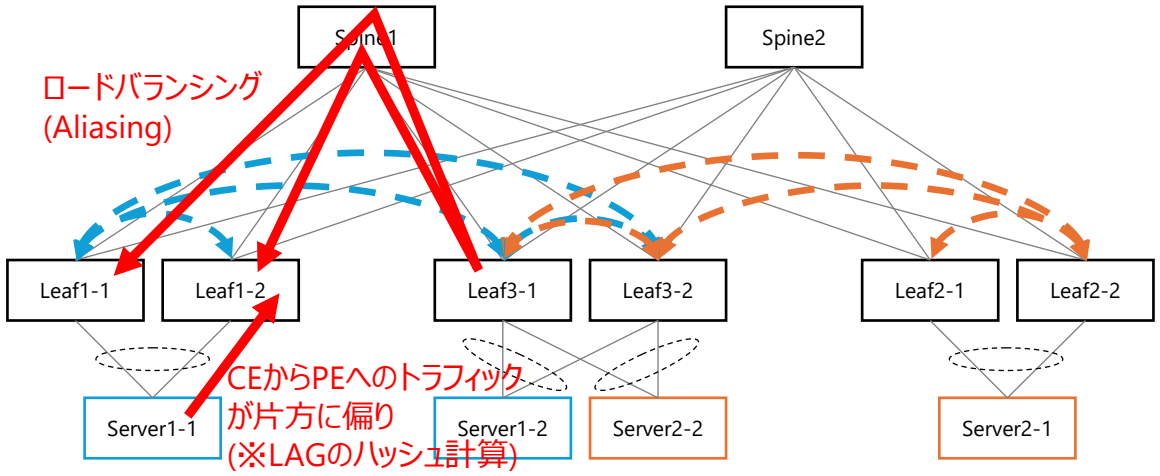
■ 方法

- Route Type 1 (per ES)の広告と撤回
  「あるLeafのESが落ちた」という情報を伝えることで、落ちたESに紐づくMAC経路を一気に削除可能
  Mass WithdrawによるFast Convergence

※通常、LeafとServerの物理リンクで障害が発生した場合、MAC経路の削除を一つずつ行う必要がある

※(SR)MPLSではESIラベルでスプリットホライズンをしていたが、VXLANでは送信元IPアドレスに基づいたフィルタリングをする

EVPN/VXLAN (4/6 ES情報(EVI)の広告)



Route Type 1のフォーマット

RD (8 octets)
ESI (10 octets)
Ethernet Tag ID (4 octets)
VNI field (3 octets)

例)Leaf1-1から広告される
ユーザ1に関するRoute Type 1 (/EVI)

1.1.1.1:1 (Leaf1-1のユーザ1のRD)
0:1:1:1:1:1:1:1:1:1 (Type 0)
0 (VLAN-Basedの場合)
10001

■ 目的
- ロードバランシング
(Remote-Leaf→マルチホーミングLeaf)

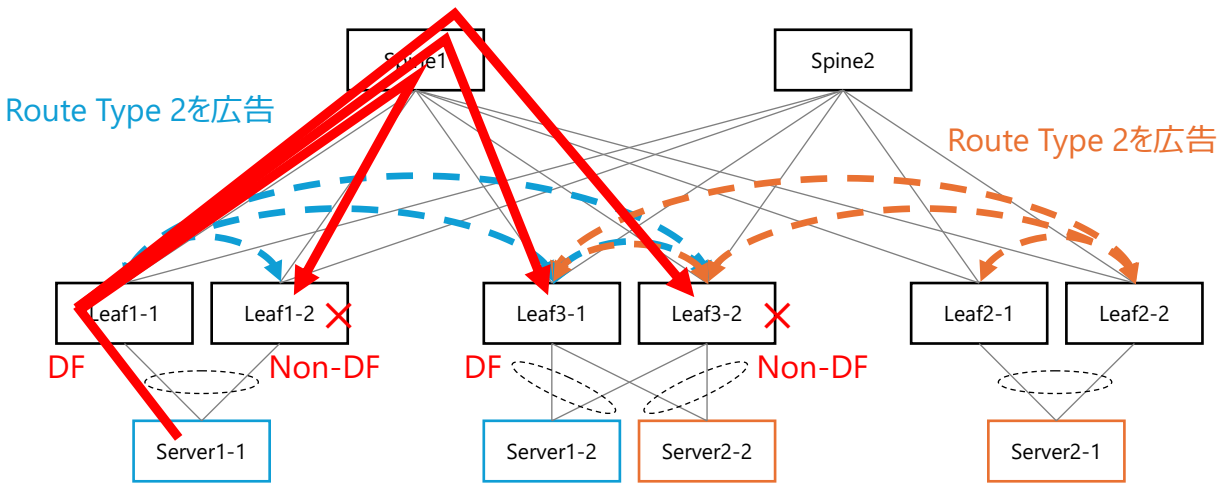
■ 方法
- Route Type 1 (per EVI)の広告

Server→マルチホーミングLeafのLAGハッシュ計算によっては片方のLeafだけにしかトラフィックが送られず、もう片方のLeafではMAC学習できないケースがある

Remote-Leafには片方のPEからのみMAC経路が広告され、工夫しないと片方のLeafに対してのみ転送

Route Type 1 (per EVI)によって、
「あるEVIのある経路について特定のLeafからしか学習していなけれど、その特定のLeafと同じESに属する別のLeafが存在していて、指定されたMPLSラベルを付ければ転送してくれるらしい」
ということをRemote-Leafは知ることができる
(Aliasing)

EVPN/VXLAN (5/6 トラフィック転送)



(例)Leaf1-1からLeaf1-2に送るフレーム　Leafで付与

DA	SA	Type	DIP	SIP	UDP Header	VXLAN Header	DA	SA	Type	ARP	FCS
Spine1のMAC	Leaf1-1のMAC		Leaf1-2のIP	Leaf1-1のIP		L2VNI	FF:FF:FF:FF:FF:FF	1:1:1:1:1:1			

(例)PE1-1からPE3-1に送るフレーム　Leafで付与

DA	SA	Type	DIP	SIP	UDP Header	VXLAN Header	DA	SA	Type	ARP	FCS
Spine1のMAC	Leaf1-1のMAC		Leaf3-1のIP	Leaf1-1のIP		L2VNI	FF:FF:FF:FF:FF:FF	1:1:1:1:1:1			

Server1-1からLeaf1-1にARPが送られてきたとき、
Leaf1-1は予め学習した全てのフラッディング先にコピーして転送

送る際は、もともとのARPのフレームに対してVXLANカプセル化

フラッディングされてきたARP(BUM)について、Non-DFである
LeafはServerに対して転送しない

並行してARPの送信元MACアドレスを学習して、
Route Type 2を他のLeafに広告

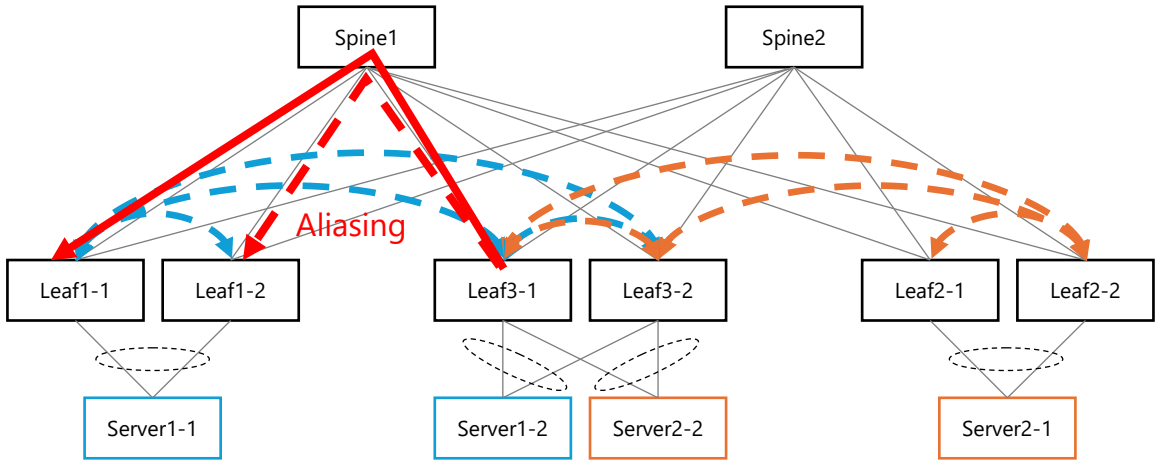
Route Type 2のフォーマット

RD (8 octets)
ESI (10 octets)
Ethernet Tag ID (4 octets)
MAC Address Length (1 octets)
MAC Address (6 octets)
IP Address Length (1 octet)
IP Address (0, 4, or 16 octets)
VNI field 1 (3 octets)
VNI field 2 (0 or 3 octets)

例)Leaf1-1から広告される
ユーザ1に関するRoute Type 2

1.1.1.1:1 (PE1-1のユーザ1のRD)
0:1:1:1:1:1:1:1:1:1 (Type 0)
0 (VLAN-Basedの場合)
48
1:1:1:1:1:1
0 (IPアドレスに関する情報無し)
-
10001
-

EVPN/VXLAN (6/6 トラフィック転送)



(例)Leaf3-1からLeaf1-1に送るフレーム　Leafで付与

DA	SA	Type	DIP	SIP	UDP Header	VXLAN Header	DA	SA	Type	ARP	FCS
Spine1のMAC	Leaf3-1のMAC		Leaf1-1のIP	Leaf3-1のIP		L2VNI	1:1:1:1:1:1	3:3:3:3:3:3			

(例)Leaf3-1からLeaf1-2に送るフレーム　Leafで付与

DA	SA	Type	DIP	SIP	UDP Header	VXLAN Header	DA	SA	Type	ARP	FCS
Spine1のMAC	Leaf3-1のMAC		Leaf1-2のIP	Leaf3-1のIP		L2VNI	1:1:1:1:1:1	3:3:3:3:3:3			

Server1-2からServer1-1にユニキャストの通信を行うとき、Leaf3-1はLeaf1-1からRoute Type 2でMACアドレスを学習しているので、フラッディングの必要が無く、Leaf1-1を狙って転送

またLeaf1-2に対してもRoute Type 1(per EVI)で転送できることが分かっているので転送可能

Route Type 2のフォーマット

RD (8 octets)
ESI (10 octets)
Ethernet Tag ID (4 octets)
MAC Address Length (1 octets)
MAC Address (6 octets)
IP Address Length (1 octet)
IP Address (0, 4, or 16 octets)
VNI field 1 (3 octets)
VNI field 2 (0 or 3 octets)

例)Leaf1-1から広告されるユーザ1に関するRoute Type 2

1.1.1.1:1 (PE1-1のユーザ1のRD)
0:1:1:1:1:1:1:1:1:1 (Type 0)
0 (VLAN-Basedの場合)
48
1:1:1:1:1:1
0 (IPアドレスに関する情報無し)
-
10001
-

本セッションのゴール (Take-away)　※再掲

1. 【入門編】フレーム転送の**「仕組み」**と**「リスク」**を理解する
   - MACアドレス学習とフォワーディングの流れ
   - ユニキャストとBUMの違い
   - ループ(ブロードキャストストーム)の怖さ

2. 【基礎編・応用編】L2ネットワークを支える**「制御技術」**と**「トラブル」**を知る
   - VLAN, STP, LAGなどの目的と手段
   - 設計と運用の落とし穴

3. 【発展編】基礎から**新技術へ視野を広げる**
   - EVPNの目的と手段
   - (時間があれば)EVPN/(SR)MPLSとEVPN/VXLANの経路学習とフォワーディングの流れ

まとめ

本日は、フレーム転送の仕組みから始まり、
制御技術、トラブル事例、そしてEVPNまで幅広くお話させていただきました。

「MACアドレスの学習」と「BUMのフラッディング」というシンプルな原則の中で、
いかにリソースを効率化し、ループを排除して耐障害性を向上させるか。
このシンプルさと複雑さのせめぎ合いがポイントとなっていました。

改めて本セッションの内容が、皆様の日々のネットワーク構築・運用、
トラブルシューティングに役立つ情報を提供できていましたら幸いです。