2019/07/26 JANOG44

目指せ！Goodbye IPv4 on L3 ToR

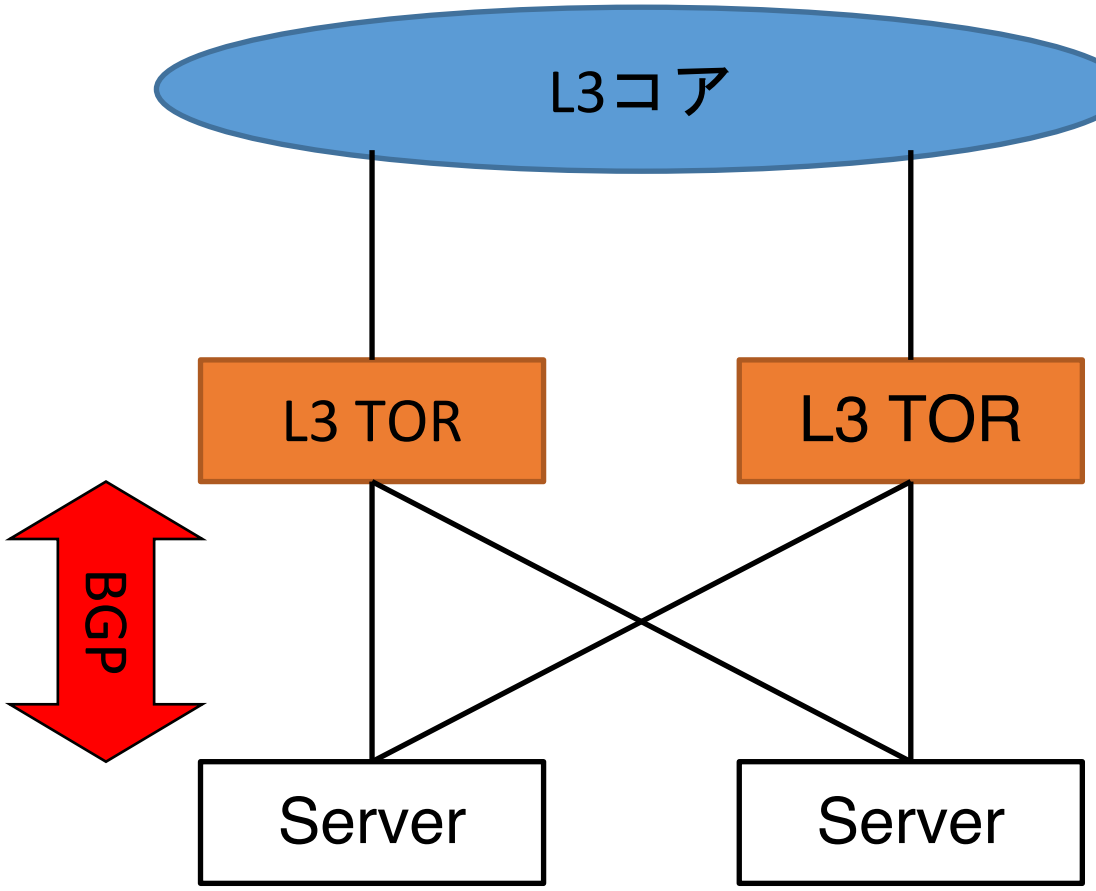
ルータIDとLoopbackを除く！

XFLAG
開発本部 インフラ室 上竹嘉史

私たちの思い

自動化よりも
管理対象削減

目指している最終形



- IP CLOS Network
  - OSPFでToRまでL3にしてみた(J40)
- ToRとサーバの接続 L2 => L3
  - フルL3化
- ルーティングプロトコルにBGPを選択
- ToR のメンテナンス性の大幅な向上

- ただし設定に関して楽をしたい
  - 管理が必要なパラメーターを最小限に！

自動インストール復習(PXEブート)

- DHCPサーバから情報取得 (untagでdhcp request)
  - IPアドレス
  - tftpサーバアドレス
  - ブートローダのファイル名
- tftpサーバにアクセス
  - ブートイメージ(pxelinux.0)
- インストーラ起動
  - パッケージ取得
- Post Script実行
- 再起動

BGPと自動インストールの相性

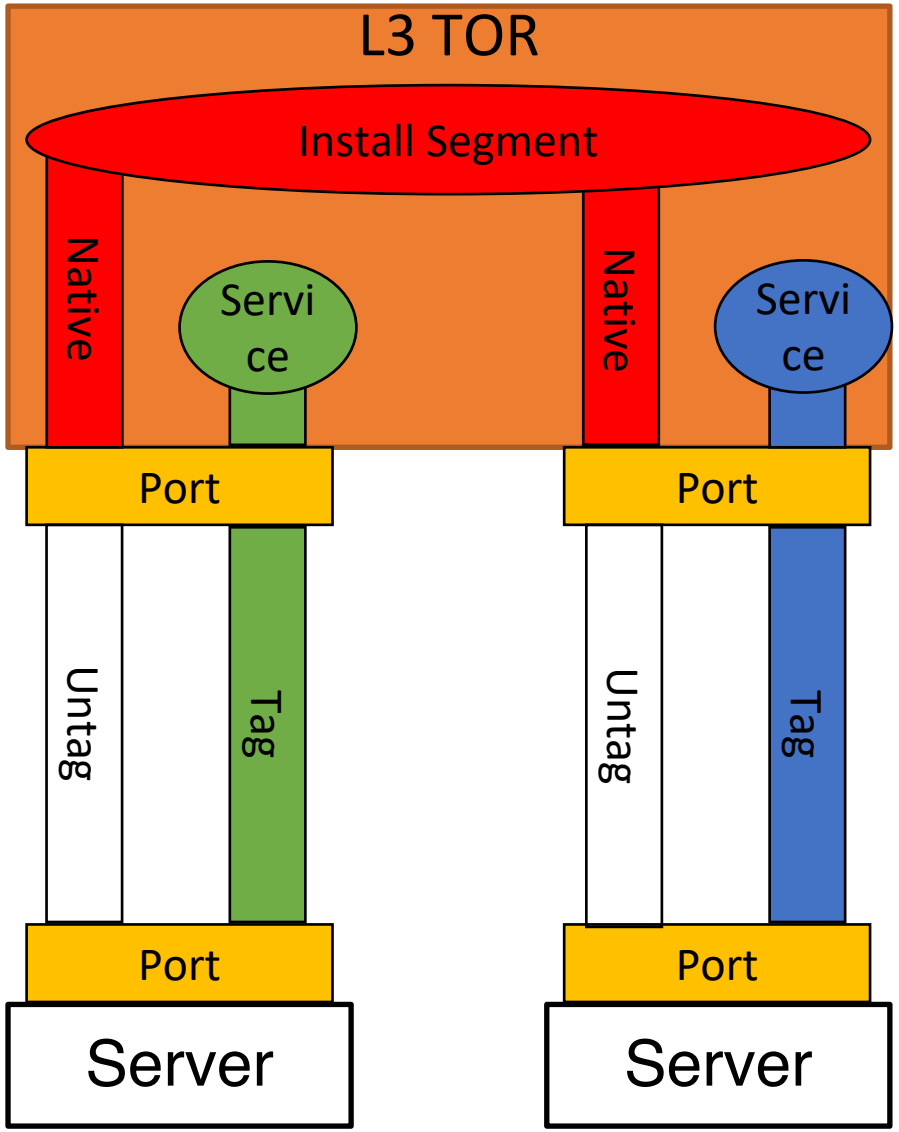
1. ブートイメージの取得 (BGPなし)
2. パッケージの取得 (BGPなし)
3. サーバ通常利用時 (BGPあり)

自動インストーラにはBGPデーモンがない

検討軸

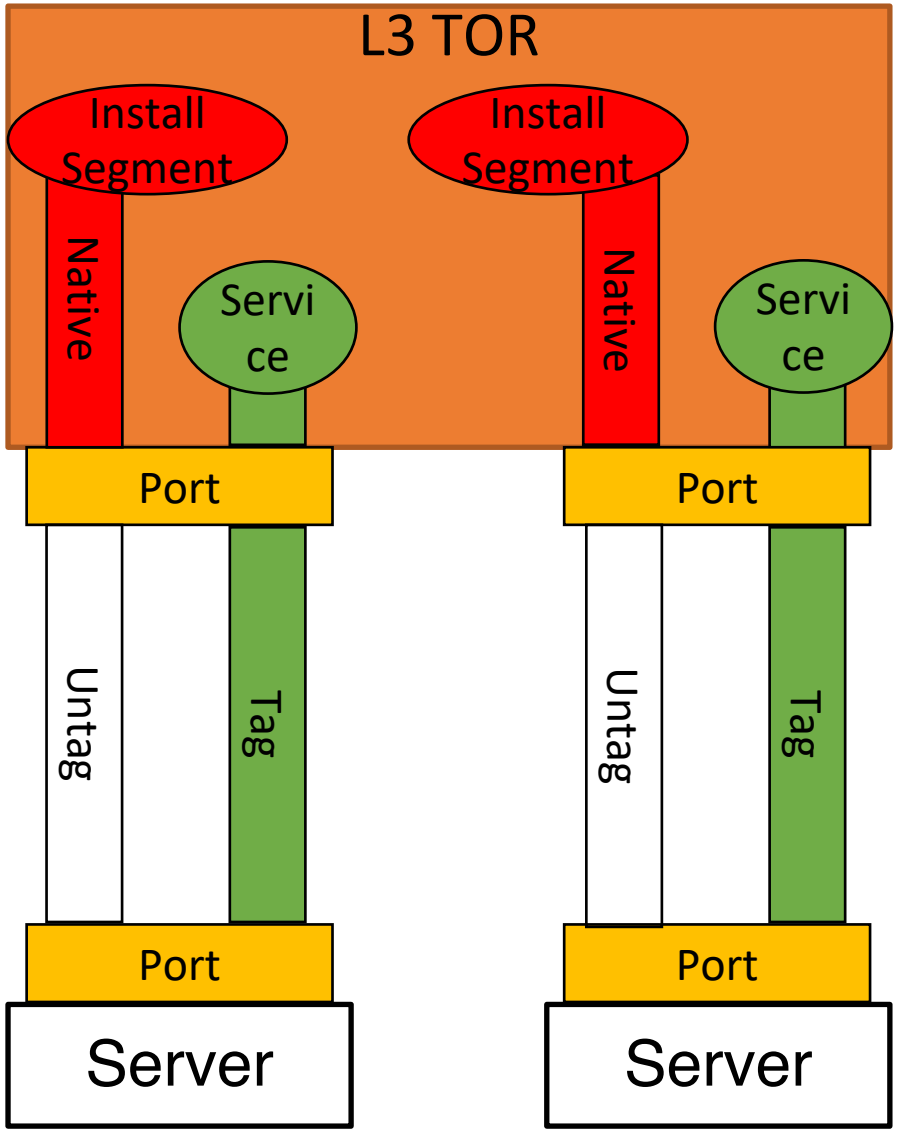
		サーバのVLAN設定
Trunk Port + Native VLAN		サーバごとにVLANがバラバラ プロビジョニングで意識する必要がある
VLANあり Routed Port		VLANは全て一致 ただしサーバ側にVLAN設定が必要
VLANなし Routed Port		VLAN設定が一切不要

Trunk Port + Native VLAN



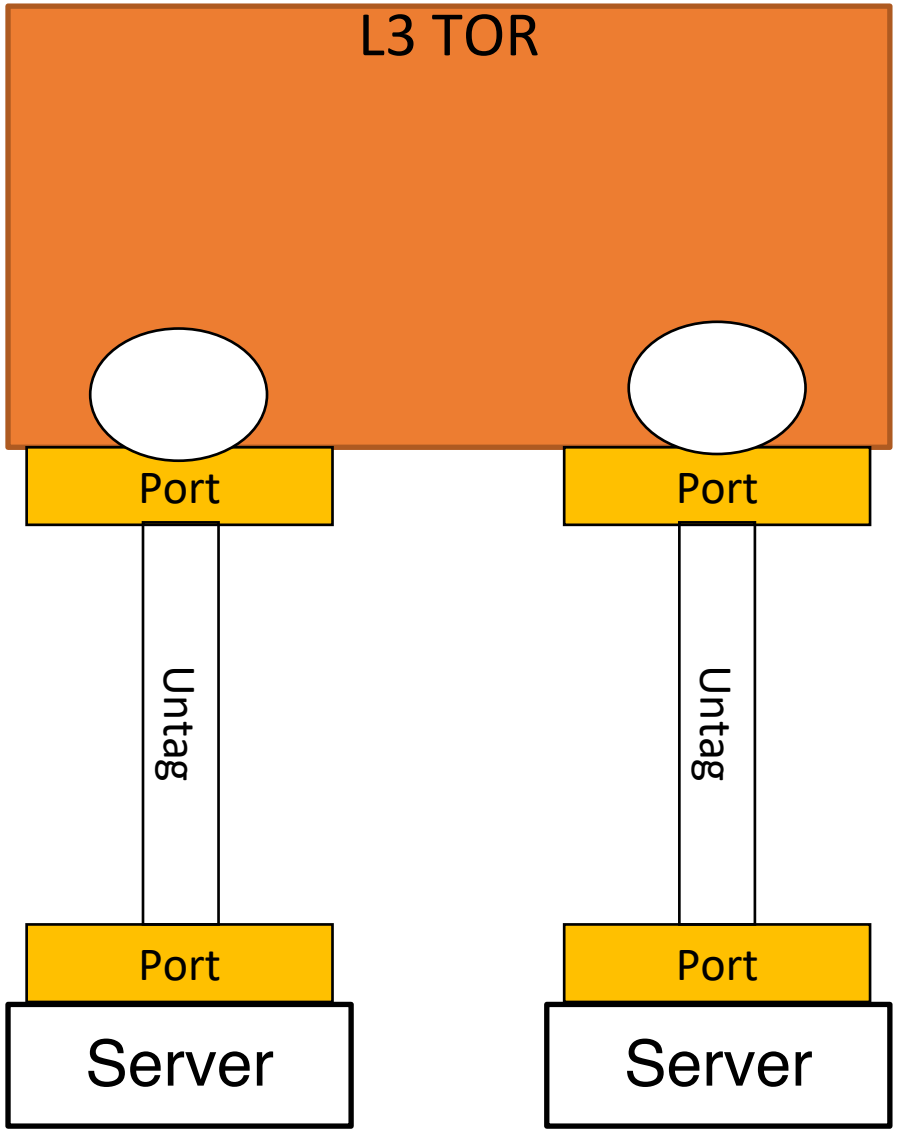
- Native VLANを利用してインストールはUntagで
- サービス用はTagged
  - サーバごとにVLANを分ける必要がある
  - スイッチのポートとサーバの関係が1対1になってしまう
  - プロビジョニングで意識する必要がある ==> 管理対象増

VLANありRouted Port



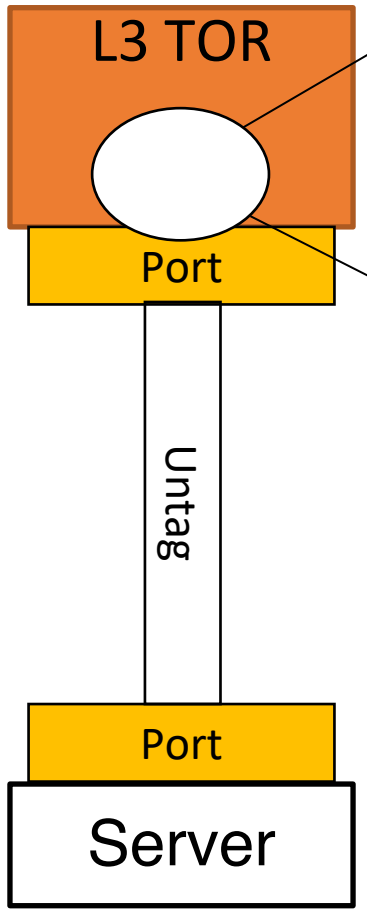
- Subinterface を利用する
- VLAN IDは共通のものを使うことができる
- ただし、サーバ側でVLAN を設定する必要がある ==> 管理対象増

VLANなしRouted Port



- インストールもサービスも共通のRouted Port を使う
- サーバでは一切のVLAN設定を無くせる
- そのために様々な仕掛け

どうやるか



- IPv4 Address
- DHCPv4 relay
- IPv6 link local address
- RA + Stateless DHCPv6

- tftp via IPv4
- インストール via IPv6 with NAT64
- サービス利用 with RFC5549

弊社環境の都合

- 弊社ではStatic NATでグローバルIPをアサインしている
  - マルチクラウド前提なので所謂Elastic IPと同じ抽象度
  - 昔の発表参照
- インストール用のIPv4 も全部mappingします？
  - 無駄使い
  - /32 でmapping 設定するのもサーバ台数的に厳しい
- インストールにおけるIPv4利用を以下に制約する
  - インターネットに接続しない
  - tftpでインストーラを取得するためだけに利用

tftp via IPv4

- 全ての対サーバRouted Port に/30を設定
- DHCP relay を利用してサーバでDHCPdを起動
- DHCPdの設定が恐ろしい量
  - サーバー台数分の全セグメント(/30)設定を記載
- IPv4 のnameserver は渡さない！！

インストール via IPv6 with NAT64

- RA でアドレス付与
- OフラグをON にする (ipv6 nd other-config-flag)
  - DHCPv6 でIPv6 nameserver を付与！！
- 後で気づいたがRFC6106でも良い(検証中)
  - IPv6 Router Advertisement Options for DNS Configuration
- DNS64 + NAT64 でインターネット接続
  - IPv4のパッケージ取得が可能に

Arista + FRR

- ToRにArista社を選択
  - RFC5549 (4.17.0F)
  - BGP IPv6 link-local peering support(4.20.5F)
  - BGP IPv6 Link Local Peers Discovery(4.21.3F)
  - AS Ranges for Dynamic BGP Peer Groups(4.17.0F)
- ホスト側のルーティングプロトコルデーモンにFRRを選択

- IPv6 link-local のみで完結可能
  - お互い対抗のASも知る必要ない
  - 管理対象削減！！

実際の設定(Arista側)

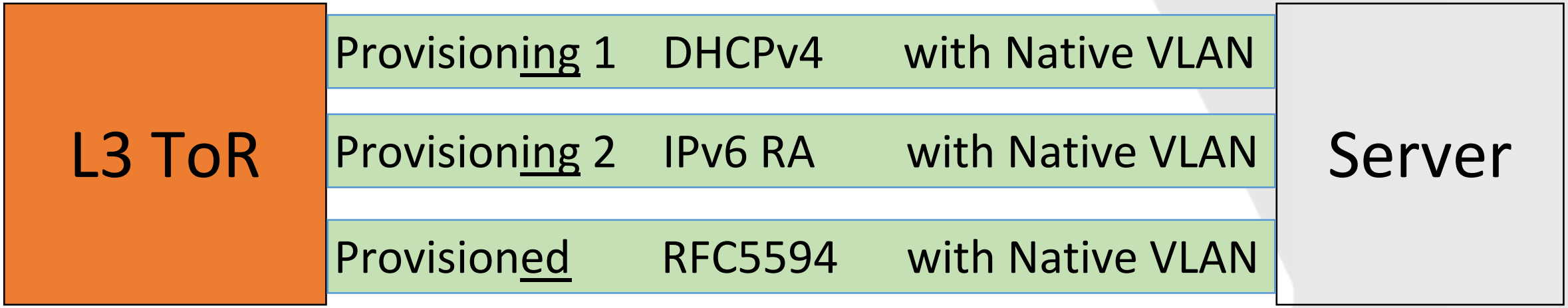
```
peer-filter leaf
   10 match as-range 4000000000-4294967295 result accept

router bgp 65000
   bgp default ipv4-unicast transport ipv6
   bgp listen range fe80::/10 peer-group HOST peer-filter leaf
   neighbor SERVER peer-group
   neighbor SERVER maximum-routes 12000
   address-family ipv4
      neighbor SERVER activate
      neighbor SERVER next-hop address-family ipv6 originate
   address-family ipv6
      neighbor SERVER activate
```

実際の設定(FRR側)

```
router bgp 4200000001
  bgp router-id 198.18.0.1
  bgp bestpath as-path multipath-relax
  network 198.18.0.1/32
  neighbor ToR peer-group
  neighbor ToR remote-as external
  neighbor ToR capability extended-nexthop
  neighbor eno1 interface peer-group ToR
  neighbor eno2 interface peer-group ToR
```

流れのまとめ



- インストーラ取得まではDHCPv4
- インストール中はIPv6 RA + NAT64
- 起動後はRFC5594

インストール後の工夫

- RA受信フィルタ
  - netplanならば、インターフェイス設定に”accept-ra: no”
- もちろんDHCP clientは動かさない

- いかに
  - BGP以外で経路を渡さないか
  - autoconf以外のアドレスを付けないか

本当のゴール

- IPv6 Router Advertisement Option for Network Boot
  - RAでインストールに必要な情報が渡せるようになる
  - Arista ではすでに実装済み(4.22.0F)
- サーバ側のIPv6 ブートが課題
- これが実現すればToRで完全にGoodbye IPv4 が可能に

まとめ

- 管理対象最小でサーバまでのL3化が実現
  - サーバに紐づくパラメーターはサービス用IPアドレスのみ
- サーバの物理的位置制約が一切ない
  - どのポートに挿してもOSインストール可能でBGPを貼ってIP を広報することが可能
  - 今後コンテナ化などを対応する上でも強みになる

Thank You!!!