

IDCフロンティア "データ集積地構想/Data Centric Cloud"を支える ネットワークのアーキテクチャ

株式会社IDCフロンティア
見崎 徳仁

2017年1月19日



- 見崎 徳仁（みさき のりひと）
- 株式会社IDCフロンティア(AS4694)
 カスタマーサービス本部
 プラットフォームエンジニアリング部
- BGP(バックボーン&CLOS)を中心に色々としています

IDCフロンティア、東日本地域に大規模クラウド拠点を新設

オールフラッシュクラウドに、さらなる高性能GPUサーバーを提供開始

[プレスリリースPDF](#) 

2016年11月25日

株式会社IDCフロンティア（本社：東京都千代田区、代表取締役社長：石田 誠司、以下IDCF）は、同社のクラウドサービスである「IDCFクラウド」で、東日本地域では2カ所目となるクラウド拠点を新設し、本日より稼働を開始します。

今回、新設したクラウド拠点では、物理的に機器が分離された2つのゾーンを同時に提供開始します。また、国内でも最大級となるIPファブリックネットワークを全面的に採用することで、従来比16倍以上となる広帯域かつシームレスなゾーン間、サービス間の相互接続が実現しています。本技術を採用した新ネットワーク構成では今後、さらなる帯域の拡張が容易に可能となります。



お知らせ

トピックス

プレスリリース

2017年

2016年

2015年

2014年

2013年

2012年

2011年

2010年

2009年

2008年

2007年

2006年

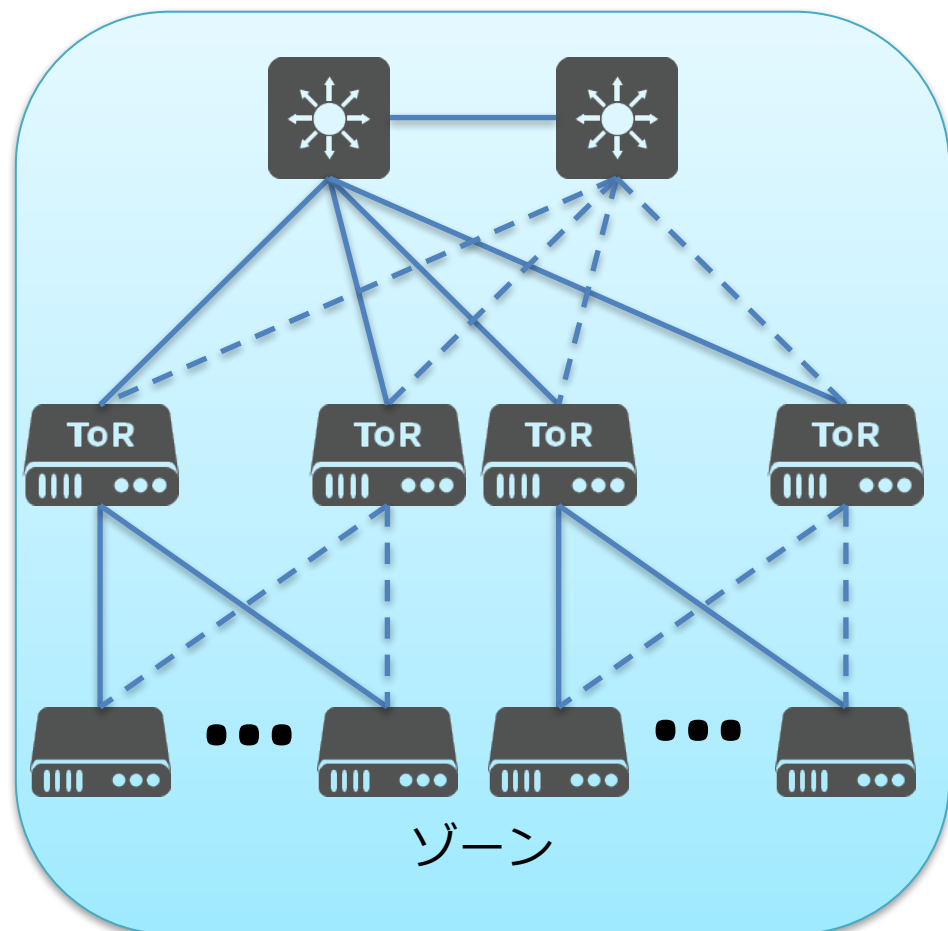
2005年

メディア・ブログ記事

イベント・セミナー

キャンペーン情報

拡張性・スケーラビリティに大きな課題



- レガシーなL2NWで構築
 - Active-Standby構成
- ゾーンごとにネットワークが分離
 - VLAN 4,000の壁
 - ゾーンをまたいでL2NWを拡張できない
- 作業でミスするとループが発生し大障害

大規模なデータセンターネットワークは CLOS(IP Fabric)がベストプラクティスらしい？

Googleさんとか…

Jupiter Rising: A Decade of Clos Topologies and
Centralized Control in Google's Datacenter Network

<http://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p183.pdf>

Facebookさんとか…

Introducing data center fabric, the next-generation Facebook data center network

<https://code.facebook.com/posts/360346274145943>

Microsoftさんとか…

Experiences with BGP in large Scale Data Centers/JANOG33 Meeting

<https://www.janog.gr.jp/meeting/janog33/program/bgp.html>

ヤフーさんとか…

ヤフーのIP CLOS ネットワーク / JANOG38 Meeting

<https://www.janog.gr.jp/meeting/janog38/program/clos.html>

- スイッチ間はL3での接続
 - L2で複数のサーバをつなげたい時は同じスイッチの下に接続しないといけない？
 - Leafのポート数以上のホストを同一L2配下に置きたい場合はどうする？
 - お客様毎にL2NWでセグメントを分離させたい場合はどうする？

**マルチテナントのネットワークを提供するためには
CLOSのネットワークを構築するだけでは不十分**

VXLAN over CLOS

[コンセプト]

CLOSでスケーラブルなアンダーレイを構築し、
VXLANオーバーレイで柔軟なL2ネットワークを実現する。

おさらい ～ CLOS編 ～

- 基本(最小)構成はSpine – Leaf構成
- Spine – Leaf間はL3で接続し、ルーティング
プロトコル(OSPF,BGP等)で制御
 - BGPを使うのが一般的と思われる

- 使い慣れた**BGP**で制御ができる
- スケールアウトが容易
- 冗長化の設計がしやすい
- ECMPが簡単に実現可能(BGP Multipath)

おさらい ～ VXLAN編 ～

- L3NW上に論理的なL2NWを構築するオーバーレイ技術
(トンネルプロトコル)
- 24bitの識別子により約1600万のNWを分離

[用語]

VNI(VXLAN Network Identifier): VXLANにおけるネットワークの識別子

VTEP(VXLAN Tunnel Endpoint): トンネルの始点・終点となる場所(IPアドレス)

宛先のMACを持つホストはどのVTEP配下にいるのだろうか？

- マルチキャスト
 - カプセル化したARP Requestをマルチキャストグループに対して送信
- Head End Replication
 - 送信先となるVTEPのリスト(Flood List)を作成し、カプセル化したARP Requestを複製しFlood ListのVTEPにフラッディング
- EVPN
 - MP-BGPを用いてMACやIPの情報を伝搬させる
- コントローラ
 - ベンダー独自のコントローラを利用し、VTEPにMACの情報を学習させる（ベンダー依存）

IDCフロンティアが採用した ネットワークアーキテクチャ

ネットワークをアンダーレイとオーバレイに分離

オーバレイ(VXLAN)

- サービス用のネットワークをVXLANで定義
- 柔軟性に優れたネットワーク

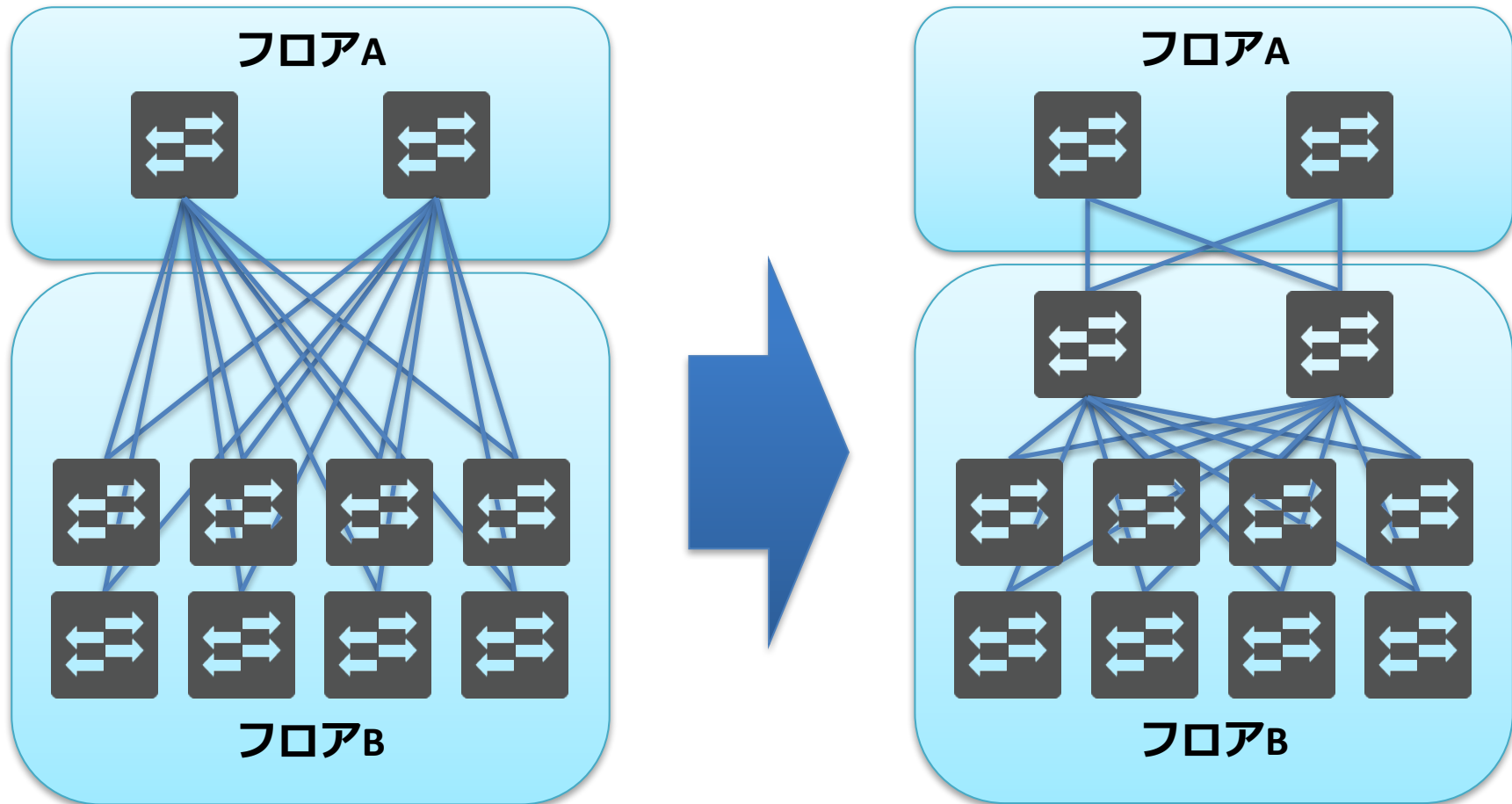
アンダーレイ(CLOS)

- カプセル化されたトラフィックの転送のみを行う
- スケーラビリティに優れたネットワーク

アンダーレイの構築

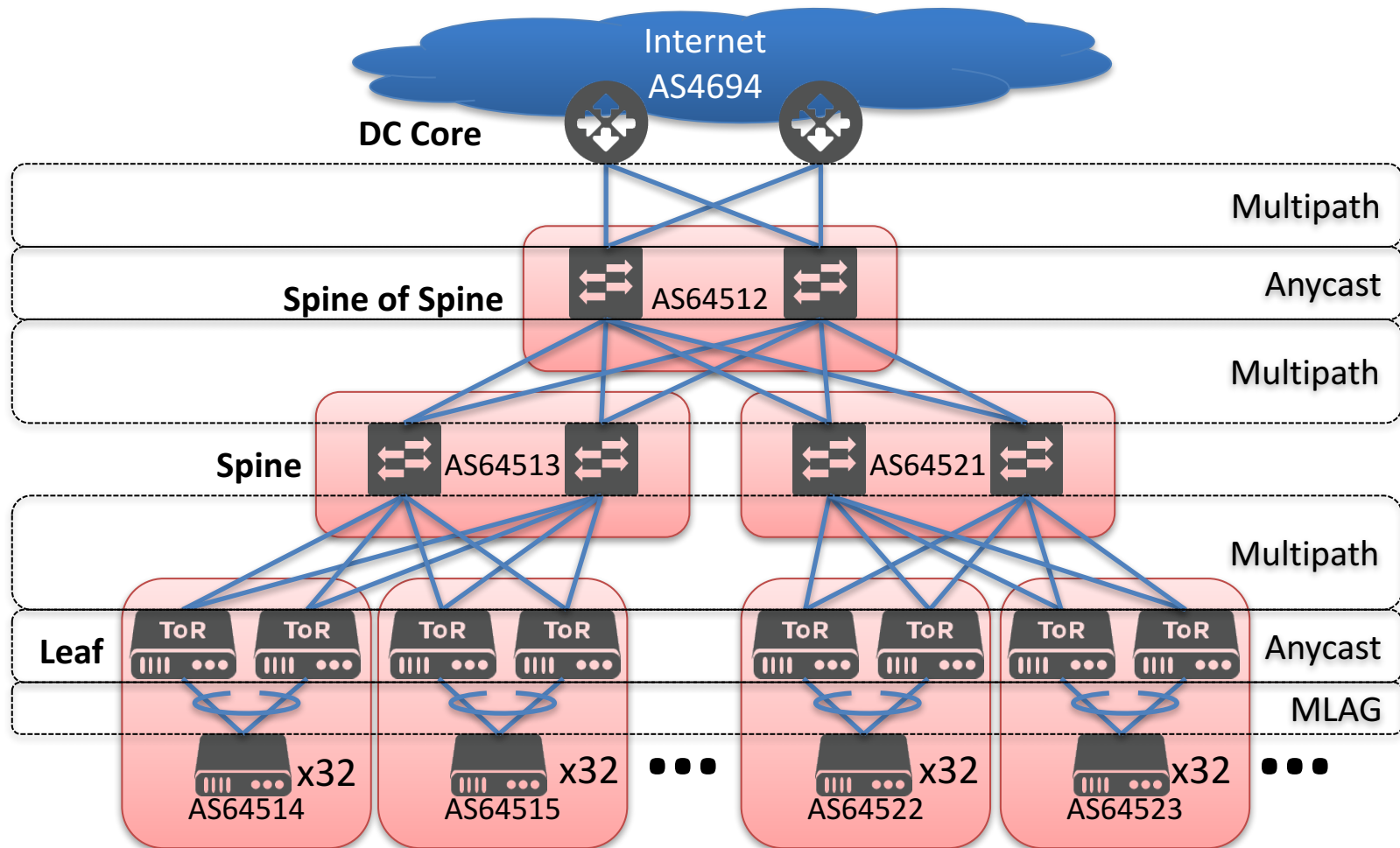
- 3層構成
 - Spine of Spine – Spine – Leafの構成
 - フロア・サービス毎にSpineを作ること、棟間・フロア間ケーブルを削減
- ルーティングプロトコルはeBGPを採用
 - 1ラック1ASで設計
 - CLOS内に再配する経路はLoopbackアドレスのみ
 - Loopbackアドレス = VTEP
 - Connectedアドレスは再配しない

3層構造にすることで棟間・フロア間のケーブルを削減



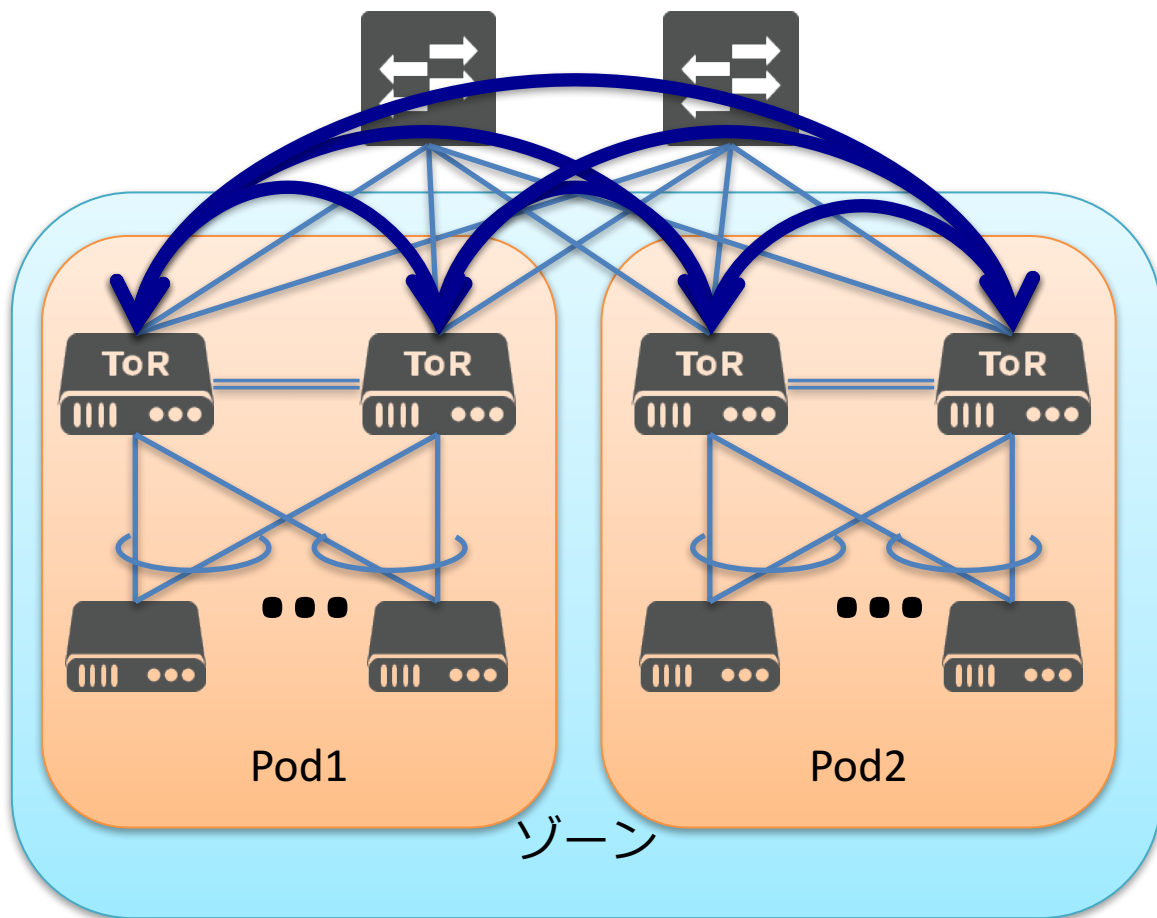
※ケーブリングのしやすさからLCコネクタのケーブルが使えるOpticsを利用

- 1ラック1ASで設計
- 全てのLink, 機器をActive-Activeで利用



オーバーレイの構築

ゾーン内の全Leafは同じマッピングルールでプリコンフィグ



VLAN 10 <-> VNI 15010

VLAN 11 <-> VNI 15011

VLAN 12 <-> VNI 15012

VLAN 13 <-> VNI 15013

(略)

VLAN 3999 <-> VNI 18999

VLAN 4000 <-> VNI 19000

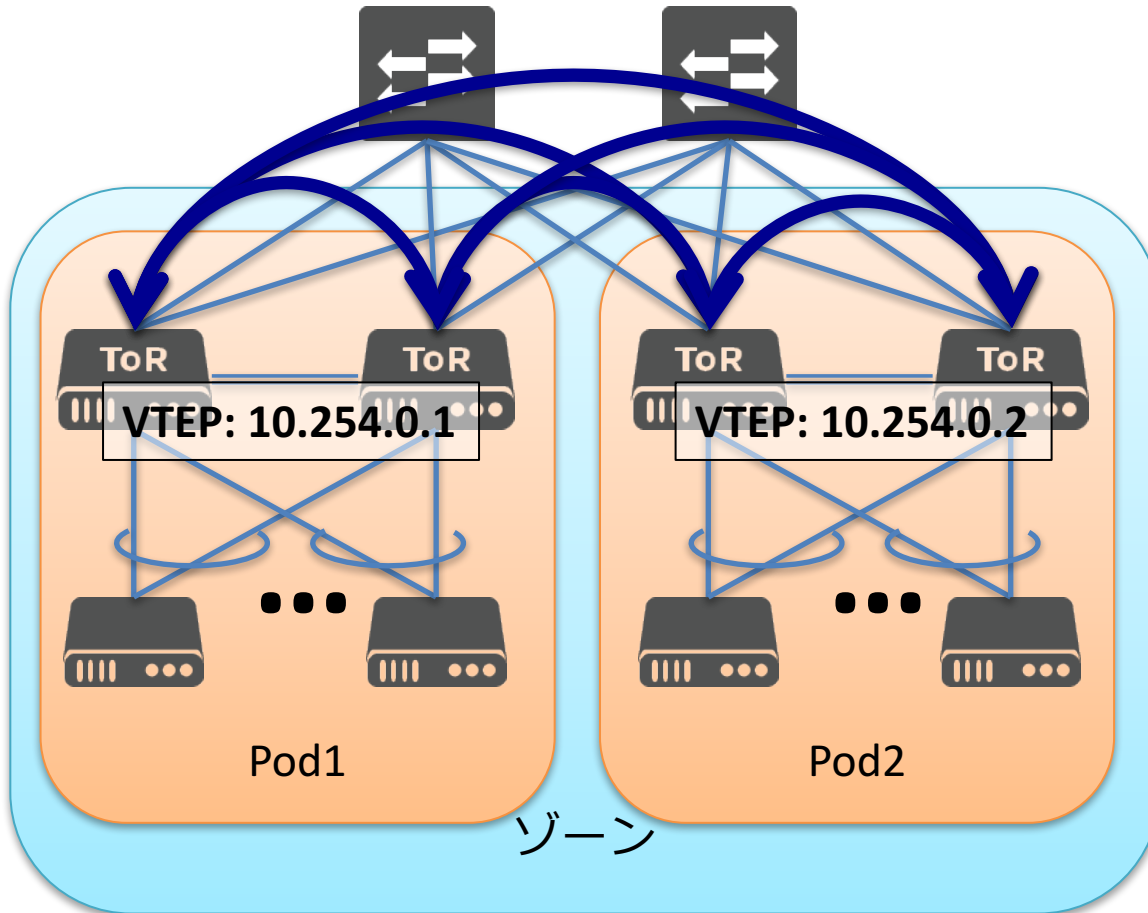
※あらかじめ約4,000個のマッピング設定

クラウド基盤ソフトウェアにネットワークを意識させない

Head End Replication方式を採用

	メリット	デメリット
マルチキャスト	VXLANの標準なので大体のベンダーが対応している	- CLOS上でマルチキャストを有効にしないといけない - LeafをActive-Activeで利用しようとする問題あり(ARPの重複)
Head End Replication	- アンダーレイ(VTEP)のActive-Active構成を生かせる - BUMトラフィックの送信範囲を容易に限定できる	ベンダーによってサポート状況が異なる
EVPN	アンダーレイのBGPと親和性が高い	- Interoperability - ベンダーによってサポート状況が異なる
コントローラ	機器との親和性が高い	Interoperability

ゾーン内の全Leafは同じFlood Listでプリコンフィグ



Headend replication flood vtep list

VLAN10 10.254.0.1 10.254.0.2

VLAN11 10.254.0.1 10.254.0.2

VLAN12 10.254.0.1 10.254.0.2

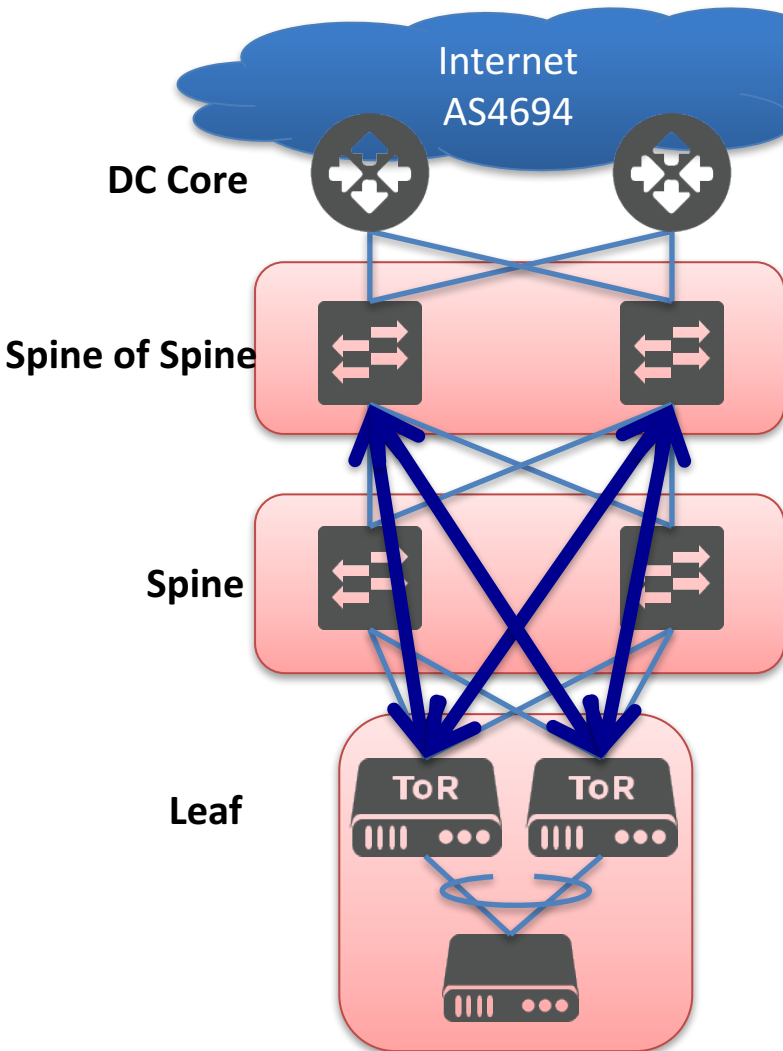
VLAN13 10.254.0.1 10.254.0.2

(略)

VLAN3999 10.254.0.1 10.254.0.2

VLAN4000 10.254.0.1 10.254.0.2

インターネット向けの通信もVXLANオーバーレイで実現



Anycast+VXLAN Routing+Virtual MAC

- Spine of SpineにてVXLANからIPアドレスを振ったVLAN Interfaceへdecap
- Virtual MACを利用し両系のVLAN IFのMACアドレスを揃える
 - どちらのスイッチにカプセル化されたパケットが届いても矛盾が生じない
- 完全にアンダーレイを隠蔽できる
 - セキュリティ面でのメリットも

- 全ての機器・全てのリンクをActive-Activeで利用
 - 障害が発生しても縮退のみ
 - キャパシティが十分なら疎通影響は出ない
- VXLANをふんだんに使い、柔軟なL2ネットワークを実現
 - VXLANの設定だけで任意の場所のVLAN同士を接続可能
 - ゾーンやサービスをまたぐL2接続も可能
 - VLAN4,000の壁をあまり意識しなくてよくなる
- No snowflakes
 - コンフィグはSpine of Spine / Spine / Leafの3種類のみ
 - どんなサービスでも同一のネットワーク構成で実現可能

- VLAN4,000の壁は未だ存在しているが、ほぼ意識しなくてよくなったので設計が楽になった
- どんなサービスでも載せられる構成なので、構築にあまり時間がかからなくなった
 - 一番時間がかかるのがハードウェアの納期…
- 奇をてらったことをやっていないが、この構成を実現できるスイッチ(チップセット)はかなり限定されてしまう
- MLAG&anycast&multipathでパケットがどこのスイッチやリンクを通過したのかが全くわからないため、トラブルシューティングで難航する可能性がある
 - このあたりを可視化する方法ってないですか？

どの方式が一番ハッピーになれるのだろうか？

- マルチキャストネットワークを構築したい人は少数では？
- Head End Replicationはゾーン間をまたぐVXLANが増えた際に管理が煩雑化するかも？
- EVPNのInteroperabilityは改善されている？
- コントローラを使った場合のInteroperabilityって？