

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

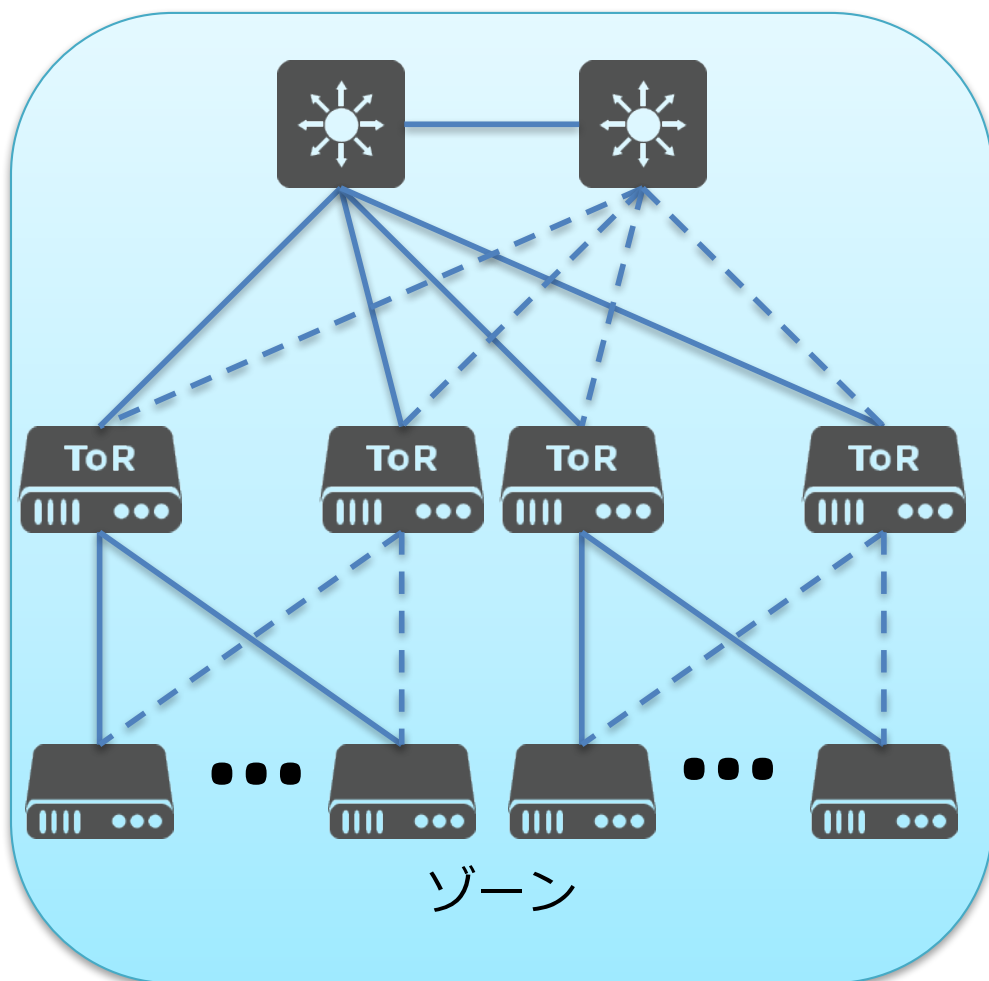
株式会社IDCフロンティア
見崎 徳仁

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

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での接続とし、ルーティング
プロトコル(OSPF,BGP等)で制御
 - eBGP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): VXLAN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）

続きは本会議にて
