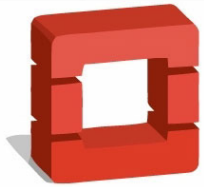


openstack™ DAYS TOKYO 2014

13 (Thu.) – 14 (Fri.) February 2014
Sola City Conference Center

The Expanding Open Cloud Ecosystem

OpenStackとVXLAN

アリスタネットワークスジャパン合同会社
兵頭 弘一

ARISTA

Redefining Data Center Switching

アリスタネットワークスの製品ラインナップ

Extensible Operating System

7048T

48ポート
データセンター
Gigabit Ethernet
スイッチ

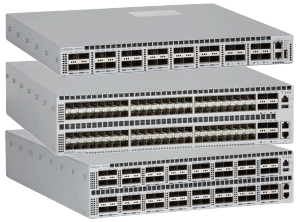
7050 S/T/Q

1/10G & 10/40G
データセンタ
スイッチ
10G SFP+ / 10G-T
高密度, 仮想化
10GbE / 40GbE DC

7150S

超低遅延
24,52,64ポート
SFP+ 1G-40GbE
スイッチ

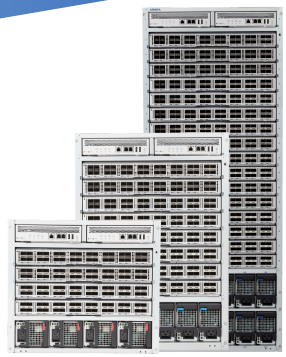
VXLAN G/W



7050X & 7250X

高密度, 低遅延
32 & 64ポート
QSFP+
96xSFP+/8xQSFP+

先進的機能
仮想化対応
高拡張性
可視化



7300X

超高密度,
モジュラシステム
512 40GbEポート

クラウド規模の
Leaf & Spine
10GbE-40GbE

7500E

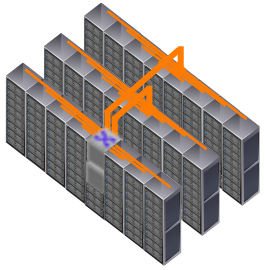
ロスレス,超高密度,
モジュラシステム 96
100Gbポート
288 40Gbポート
1152 10Gb ポート
(ワイヤスピード)

Spine
10-40-100GbE

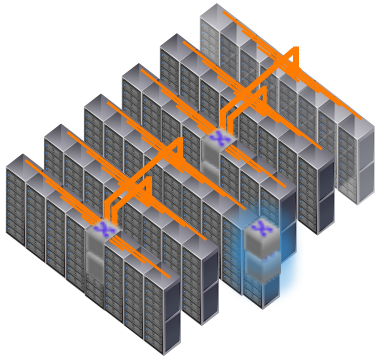
アジェンダ

- ✓VXLANのおさらい
- ✓なぜVXLANなのか？
- ✓ネットワーク設計上の要求事項
- ✓“OpenStack over VXLAN”によるネットワーク設計
- ✓今後の展望

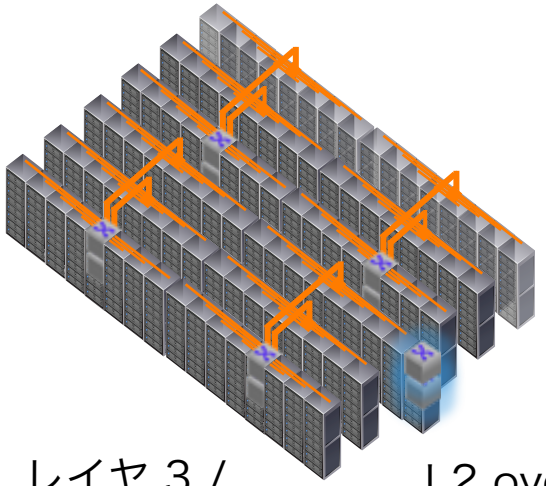
データセンターネットワークのトポロジ



Spline™

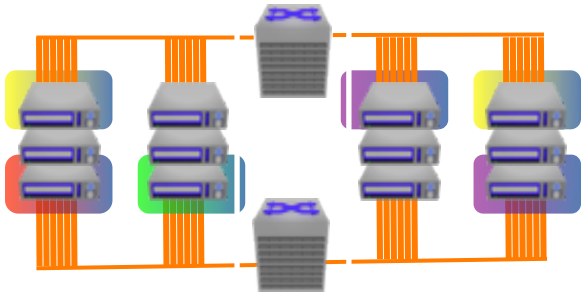


レイヤ 2 / MLAG

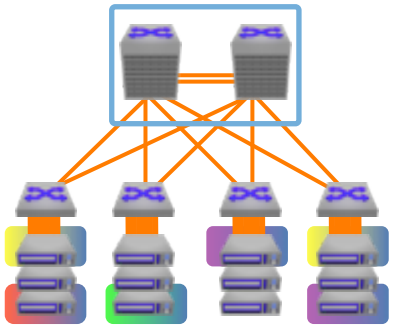


レイヤ 3 / ECMP

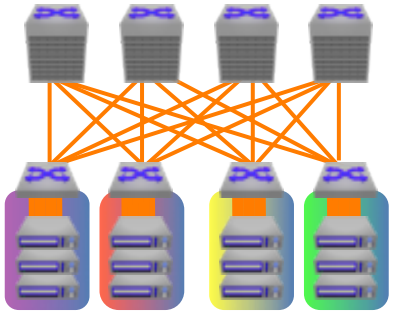
L2 over レイヤ 3 VXLAN



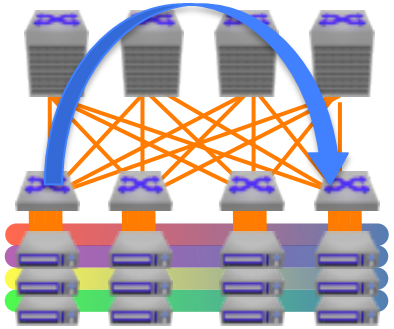
サーバ　Middle of Row　サーバ



サーバ



サーバ



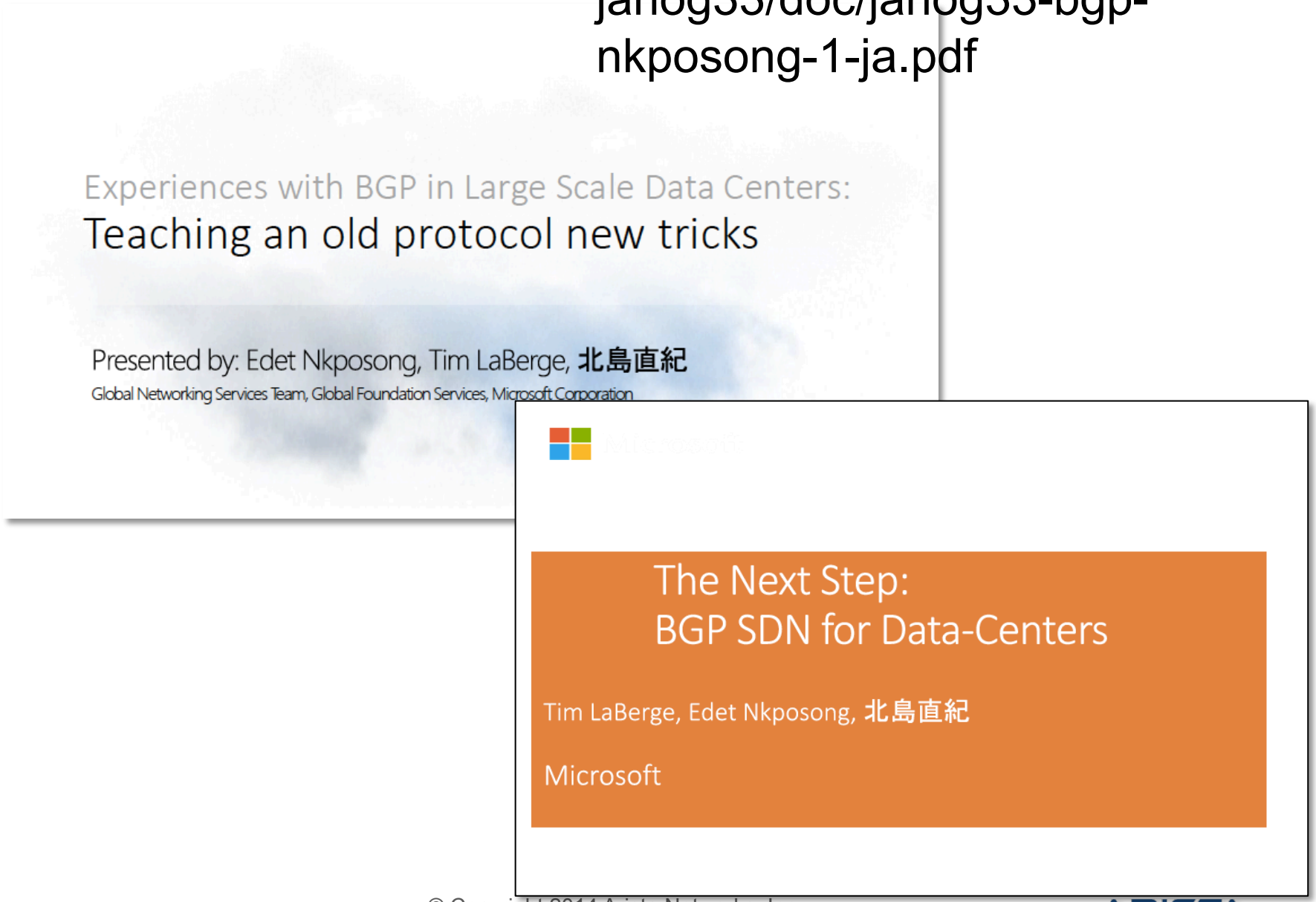
サーバ

サーバ数:	100 to 2,000	100 to 10,000	100 to 100,000+	100 to 100,000+

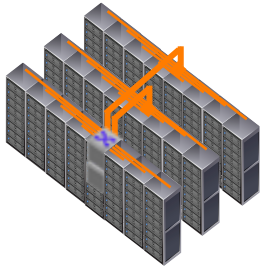
ARISTA

JANOG33にて

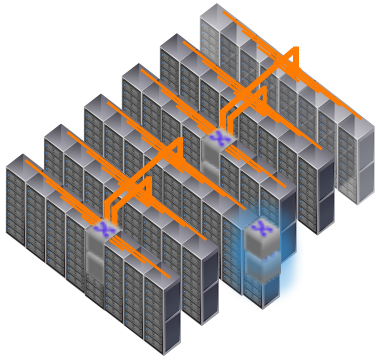
http://www.janog.gr.jp/meeting/janog33/doc/janog33-bgp-nkposong-1-ja.pdf



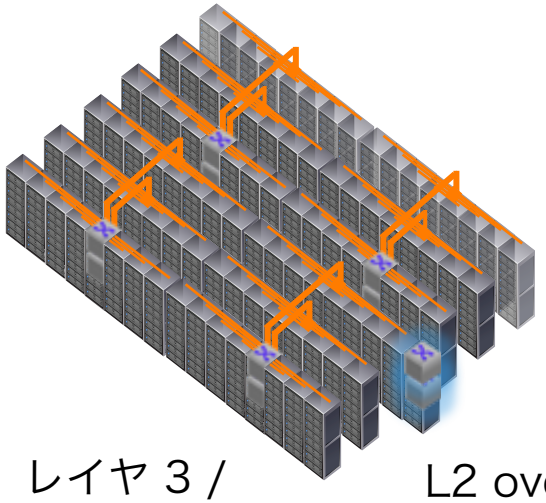
データセンターネットワークのトポロジ



Spline™

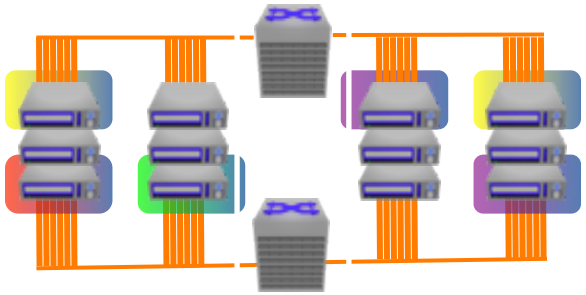


レイヤ 2 / MLAG

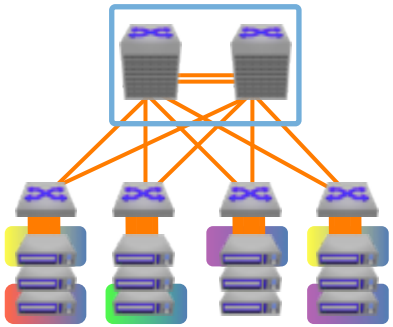


レイヤ 3 / ECMP

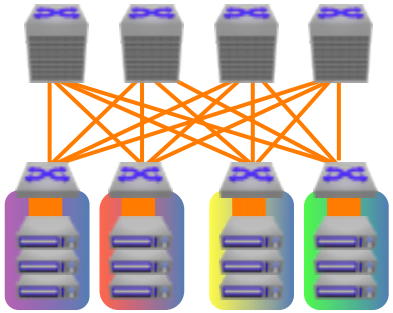
L2 over レイヤ 3 VXLAN



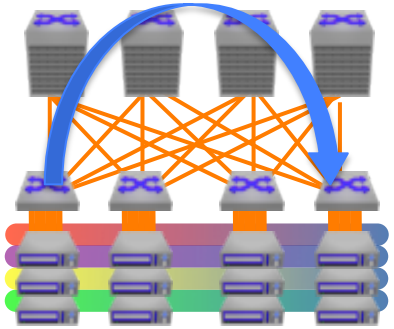
サーバ　Middle of Row　サーバ



サーバ



サーバ



サーバ

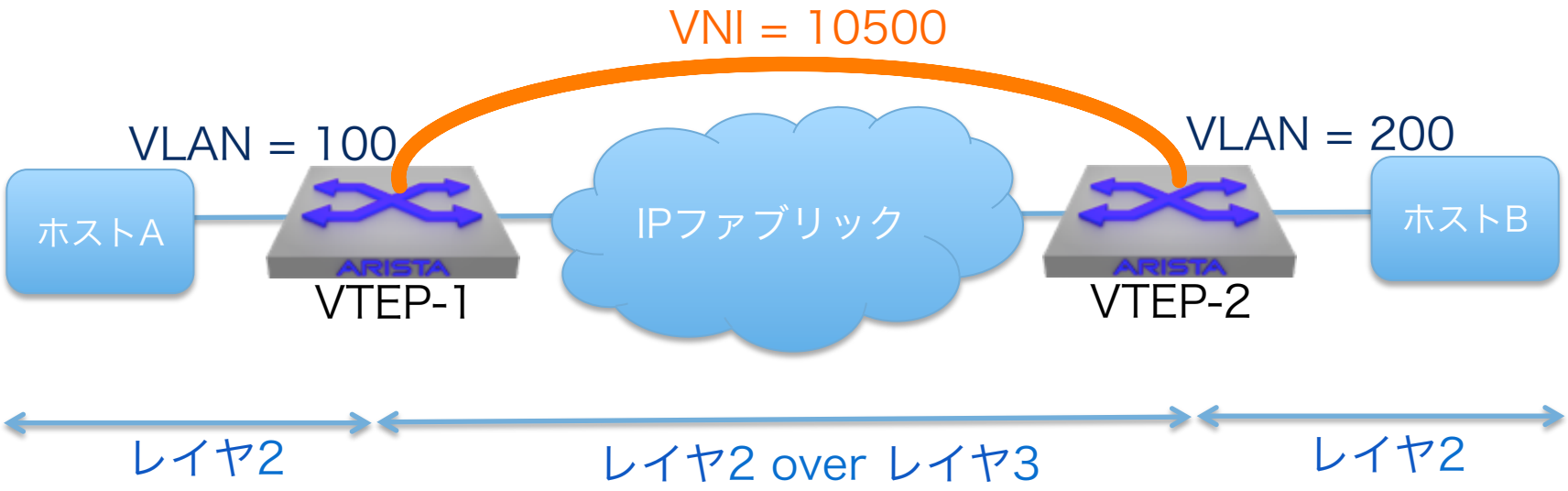
サーバ数:	100 to 2,000	100 to 10,000	100 to 100,000+	100 to 100,000+

ARISTA

なぜVXLANなのか？

- 4,000 VLANの制限に対する解決策。VXLANでは1,600万のテナントネットワークを作成可能
- コアネットワークにおけるMACアドレステーブルの肥大化の問題を解決
- L3 ECMPファブリックをコアとすることで、より高いスケーラビリティと耐障害性を実現
- VXLANはエンドポイントでのみサポートしていればよいため、ネットワーク内の機器選定の柔軟性が増す
- 汎用イーサネットコントローラのサポート

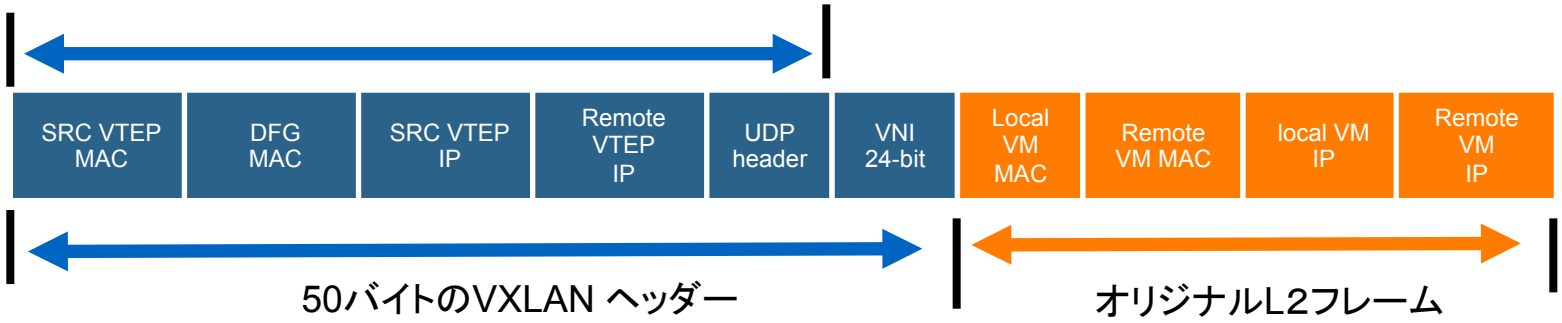
VXLANとは



- ✓ draft-mahalingam-dutt-dcops-vxlan-08.txt
- ✓ IPファブリック上でL2フレームを透過的に中継するためのオーバレイテクノロジ

VXLAN フレームフォーマット

- フレームフォーマット
  - オリジナルのL2フレームをUDP/IPでカプセル化
  - UDPヘッダー
    - ソースポートは、オリジナルフレームのイーサネットヘッダーに基づいたハッシュ値,
    - デスティネーションポートはIANA定義値
    - VXLANそのものを考慮していないIPファブリックにおいてもECMPによるロードバランスが可能
  - 8バイトのVXLANヘッダーは24ビットのVNIによって1600万超のL2ドメインを実現



VXLANにおけるMACラーニングとフラッディング

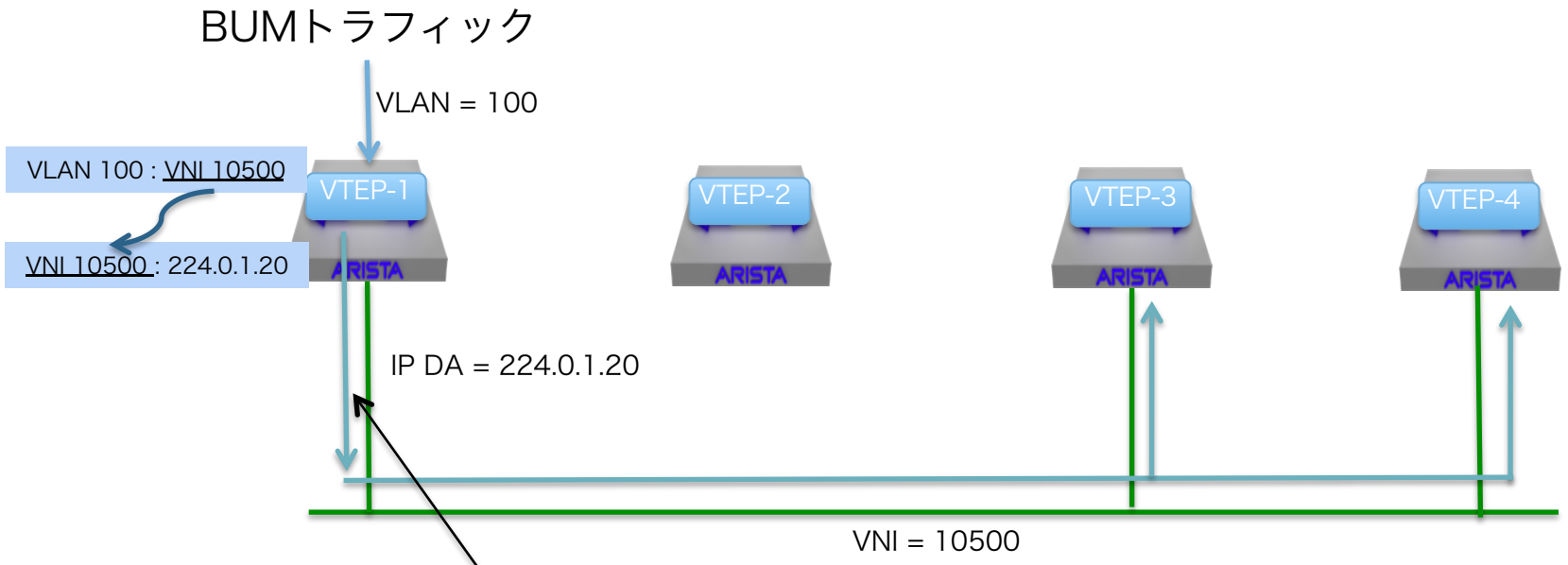
- MACラーニング
  - トンネルから受信されるフレームのSAを学習
  - MACアドレステーブルを配布するためのプロトコルを使用（併用）することも可能

- BUMトラフィックの取り扱い
  - BUM = ブロードキャスト(Broadcast) / 未学習アドレス宛ユニキャスト (Unknown Unicast) / マルチキャスト (Multicast)
  - BUMトラフィックのフラッディングに関する選択肢
    - IPマルチキャスト
    - ヘッド・エンド・レプリケーション / レプリケーション・ノード

VXLAN上でOpenStackを実現するための要望事項

- IPマルチキャストに対する抵抗
  - IPマルチキャストはBUMトラフィックをVTEP間にフラッドするには効率のよいメカニズムである
  - しかしながら… ネットワーク内でIPマルチキャストを動かすことに抵抗感を示す方が多数…
- ハードウェアベースのVXLANゲートウェイ
  - ノース・サウス方向のトラフィックの転送
  - 物理インフラ（ストレージ、非仮想化サーバなど）と仮想化されたネットワークのブリッジング
  - ソフトウェアベースのVXLANゲートウェイは、性能やポート密度において不十分

BUMトラフィックのフラッディング



VNIに対応するIPマルチキャストグループに対して
BUMトラフィックをフラッド

設計上での重要事項

- ソフトウェアベースのVTEPかハードウェアベースのVTEP
- レプリケーション・ノードかヘッド・エンド・レプリケーションによる完全分散か
- 外部のSDNコントローラかスタンドアローンのNeutronか

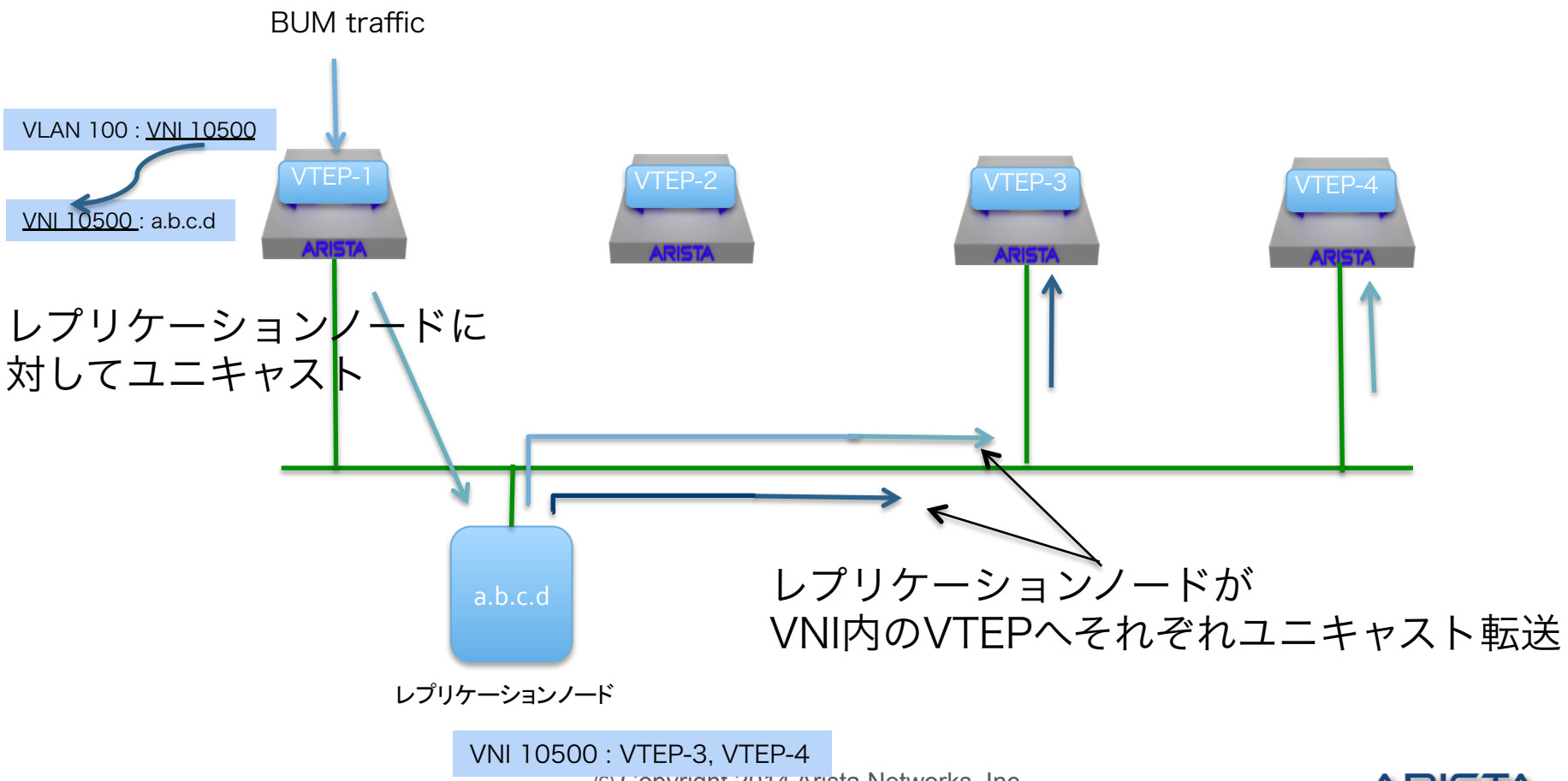
ソフトウェアベースのVTEPかハードウェアベースのVTEPか

- **ソフトウェアの柔軟性かハードウェアのパフォーマンスか**
  - ソフトウェアベースのVTEPはCPUパワーとメモリ容量により制限を受け、コンピュートノードごとに10～30%のオーバヘッドとなる
  - ハードウェアベースのVTEPは高いポート密度と性能を実現できる反面、ハードウェアのテーブルサイズによる制限を受ける
- **VXLAN環境におけるネットワークマネージメント**
  - 監視やトラブルシューティングのためのツール
    - sFlow
    - ミラーリング
  - エンキャプセレーション前のトラフィックに対する操作

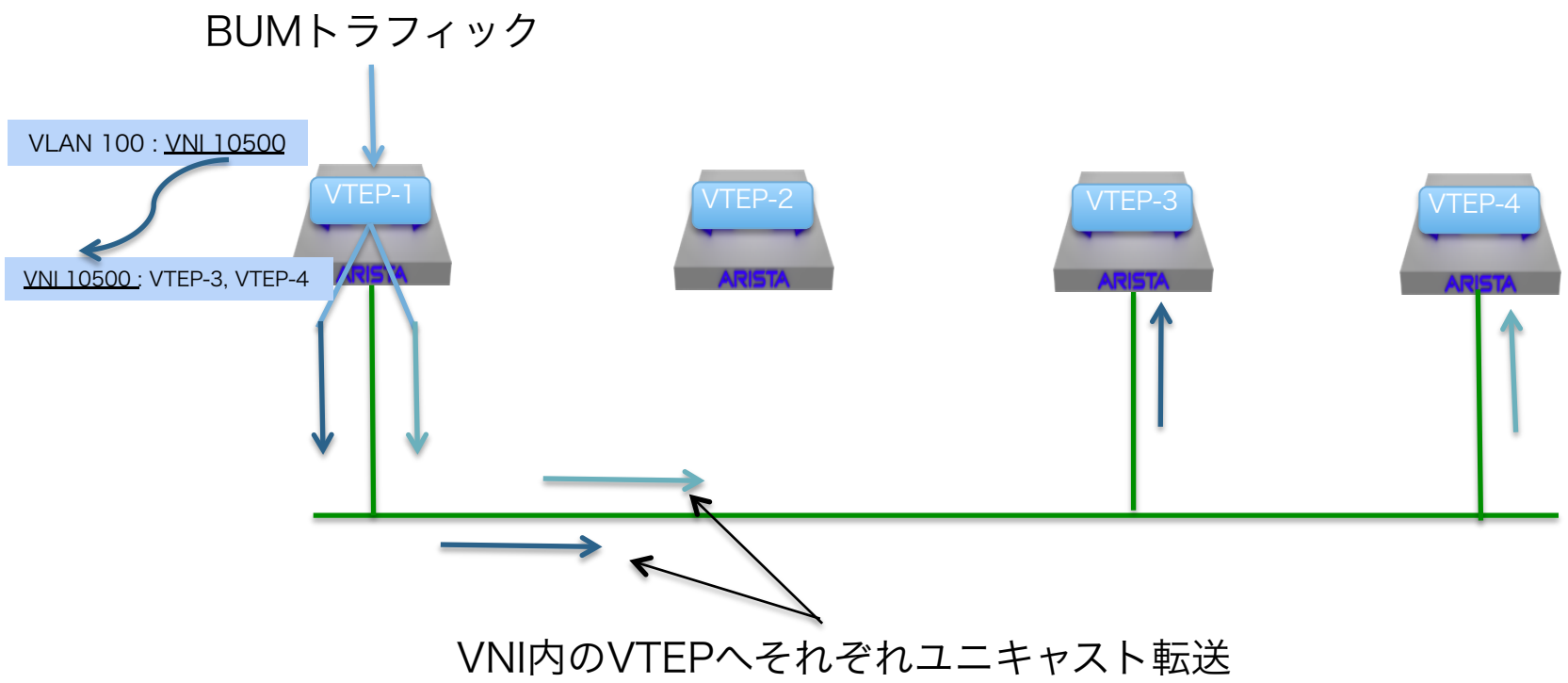
レプリケーション・ノードか
ヘッド・エンド・レプリケーションか

- **レプリケーション・ノードは、目的特化型で開発可能**
  - 複数のレプリケーション・ノードへのフローの分散が可能
  - レプリケーション・ノードの管理とHA機能が必須
- **各VTEPによるヘッド・エンド・レプリケーションであればHAは不要**
  - VTEPにレプリケーション動作の負荷

VXLANレプリケーションノード



VXLAN Head-end Replication



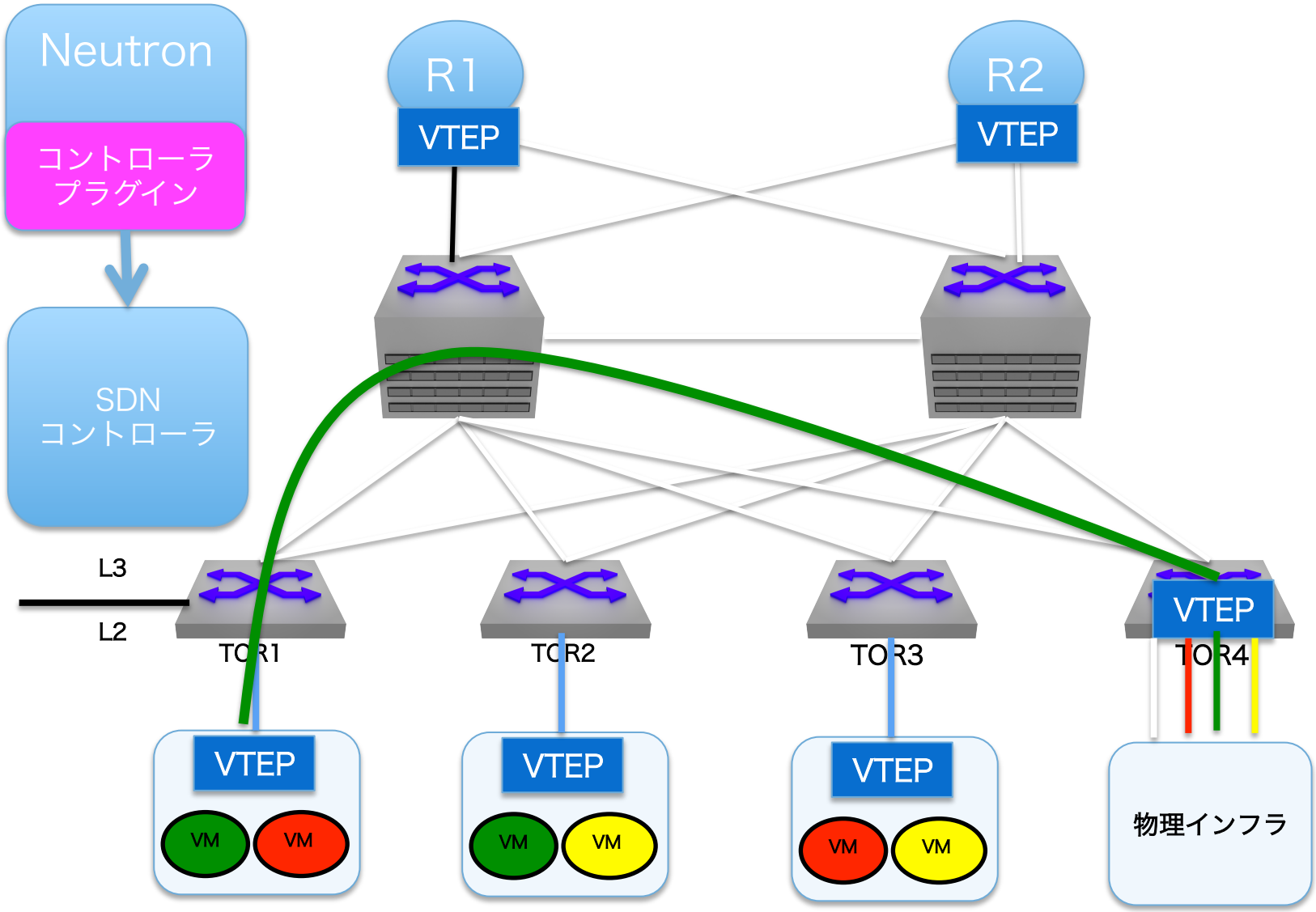
外部SDNコントローラかスタンドアロンNeutronか

- なかなか難しい選択
- 一般的には、機能とコストに応じて選択されることになるか

OpenStack over VXLAN

- **商用環境に向けた3つのデザイン**
  - ソフトウェアVTEPとハードウェアVTEPの混在環境を外部SDNコントローラで管理
  - スタンドアロンのNeutronとハードウエアVTEP
  - スタンドアロンのNeutronと、ソフトウェアVTEPとソフトウェアVTEPの混在

外部コントローラと
ソフトウェアVTEP&ハードウェアVTEP



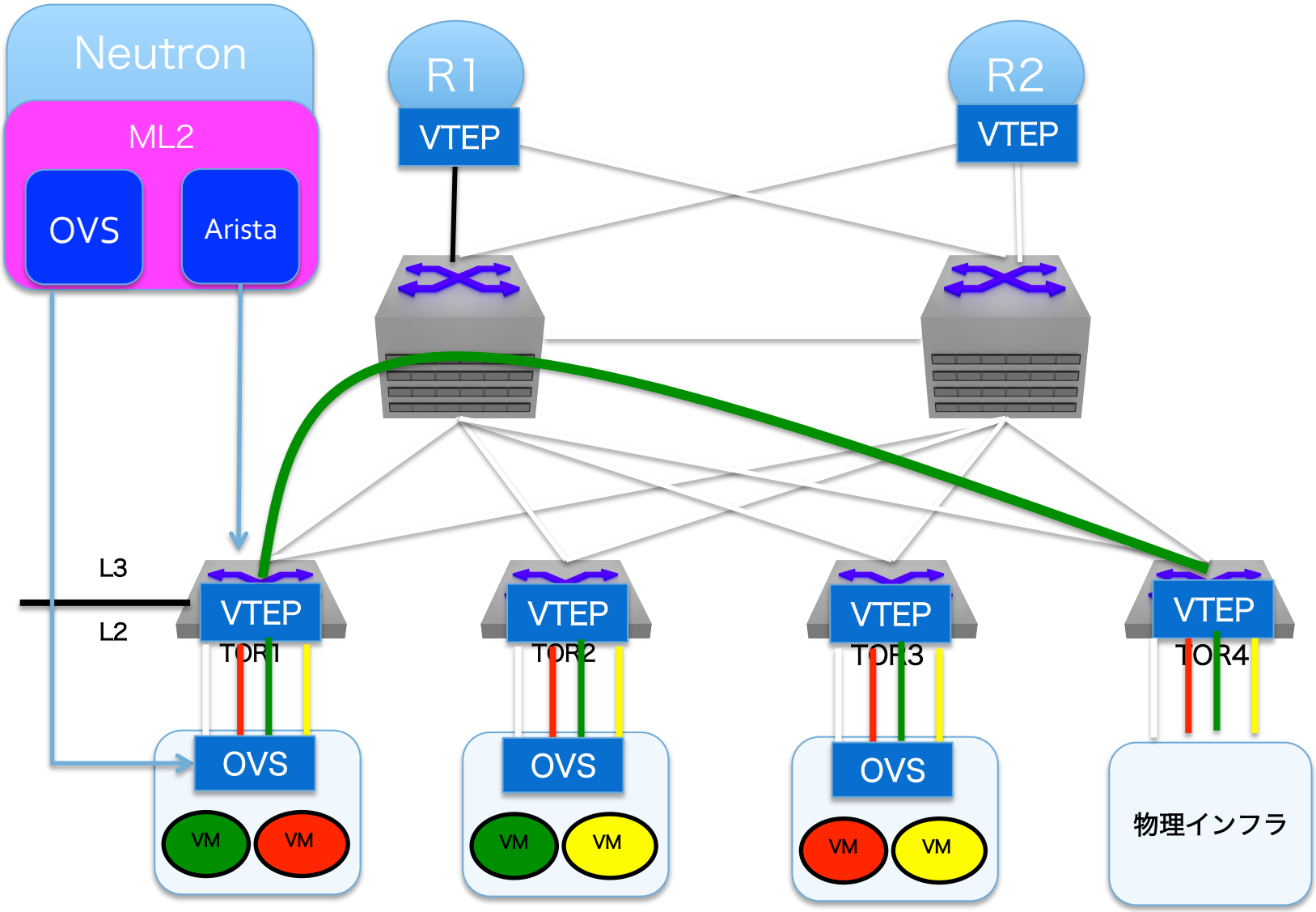
ソフトVTEP&ハードVTEPと外部SDNコントローラ

- SDNコントローラ
  - 仮想VTEPとその配下にあるVMを管理
  - ハードウェアVTEPと統合して、Neutronによるエンド・トゥ・エンドのプロビジョニングにおけるゲートウェイ機能のプロビジョニング
  - 物理VTEPと仮想VTEPの間でのVXLAN MACアドレステーブルの交換を行い、マルチキャストレスのVXLANを実現

Neutron ML2

- Modular Layer 2 plugin
- Havanaリリースで対応した新しいNeutronのプラグイン
  - テナントネットワークのステータスと、そのステータスがネットワーク上でどのように実現されるか… を分ける
  - 物理ネットワークと論理ネットワークがいかに管理されるかに関して柔軟
  - 複数のメカニズム・ドライバにより複数のベンダーのネットワーク機器を同時にサポート
- Type DriverとMechanism Driver
  - Type Driverはネットワークのタイプごとに必要なステート情報を維持
  - Mechanism DriverはType Driverで設定された情報を、各メカニズムに適用する

NeutronとハードウェアVTEP



スタンドアロンNeutronとハードVTEP

- ハードウェアの機能、性能を活用して、コンピュート・ノードのCPUの使用を削減
- VLANによる4Kテナントネットワークの制限
  - ラックごとのVLAN指定により4K以上のテナントネットワークの作成も可能だが、そのためには若干の開発とML2のマルチセグメントサポートが必要

今後の展望

- スタンドアロンNeutronとソフトVTEP、ハードVTEPの混在は、現時点では困難
  - VXLANの接続性に関する情報を、物理インフラと論理インフラの間で共有する仕組みが必要
  - ML2へのL2ポピュレーション・メカニズムの取り込みが正しい方向か
- NeutronによるVXLANゲートウェイノードの一般化モデルが必要
  - テナントネットワークへの、物理インフラの動的な接続、分離

アリスタネットワークスの最新状況

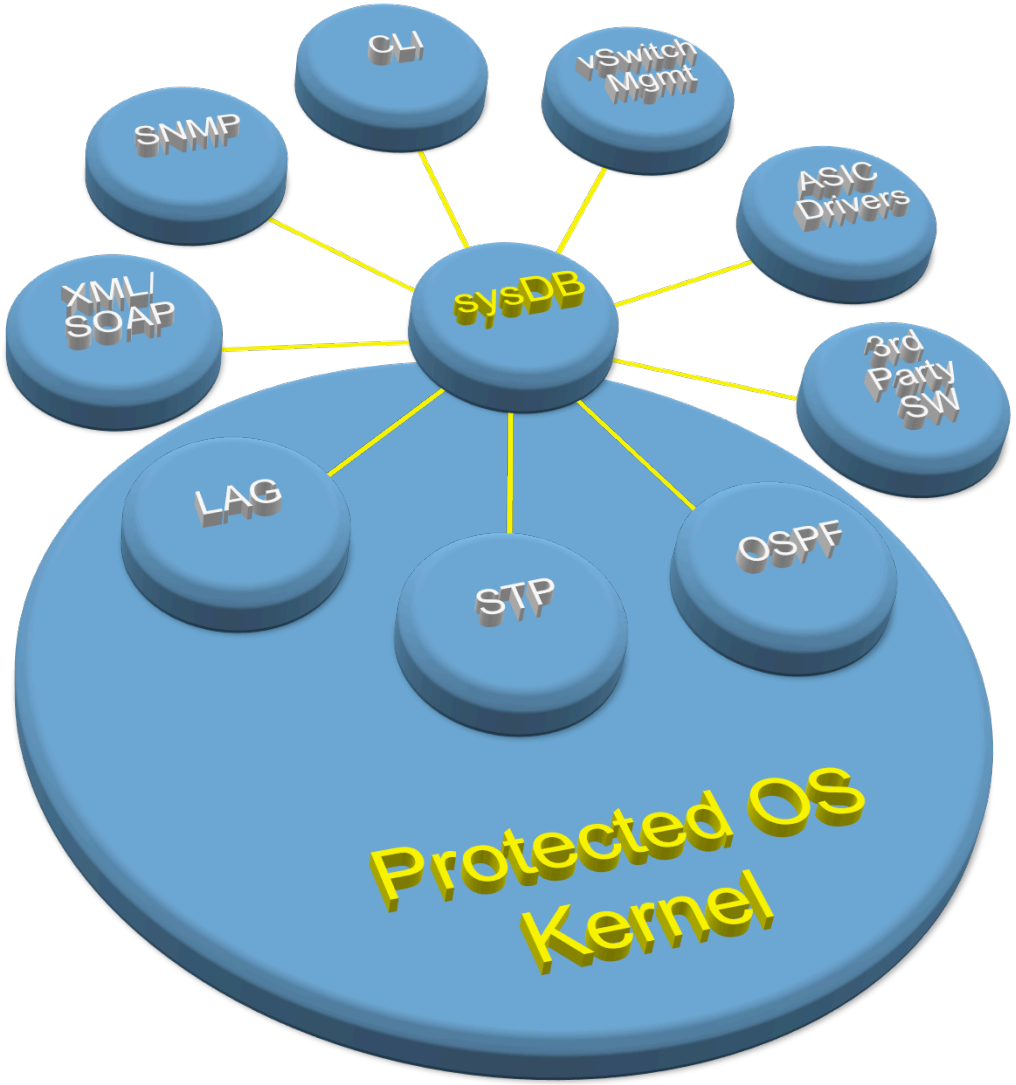
仮想化、クラウド、データセンター向け
10/40/100GbEネットワーク

- ✓ 2004年創業
- ✓ 2008より製品出荷開始
- ✓ 2200社を超えるユーザ
- ✓ 従業員750名

次世代のネットワークOS “EOS”

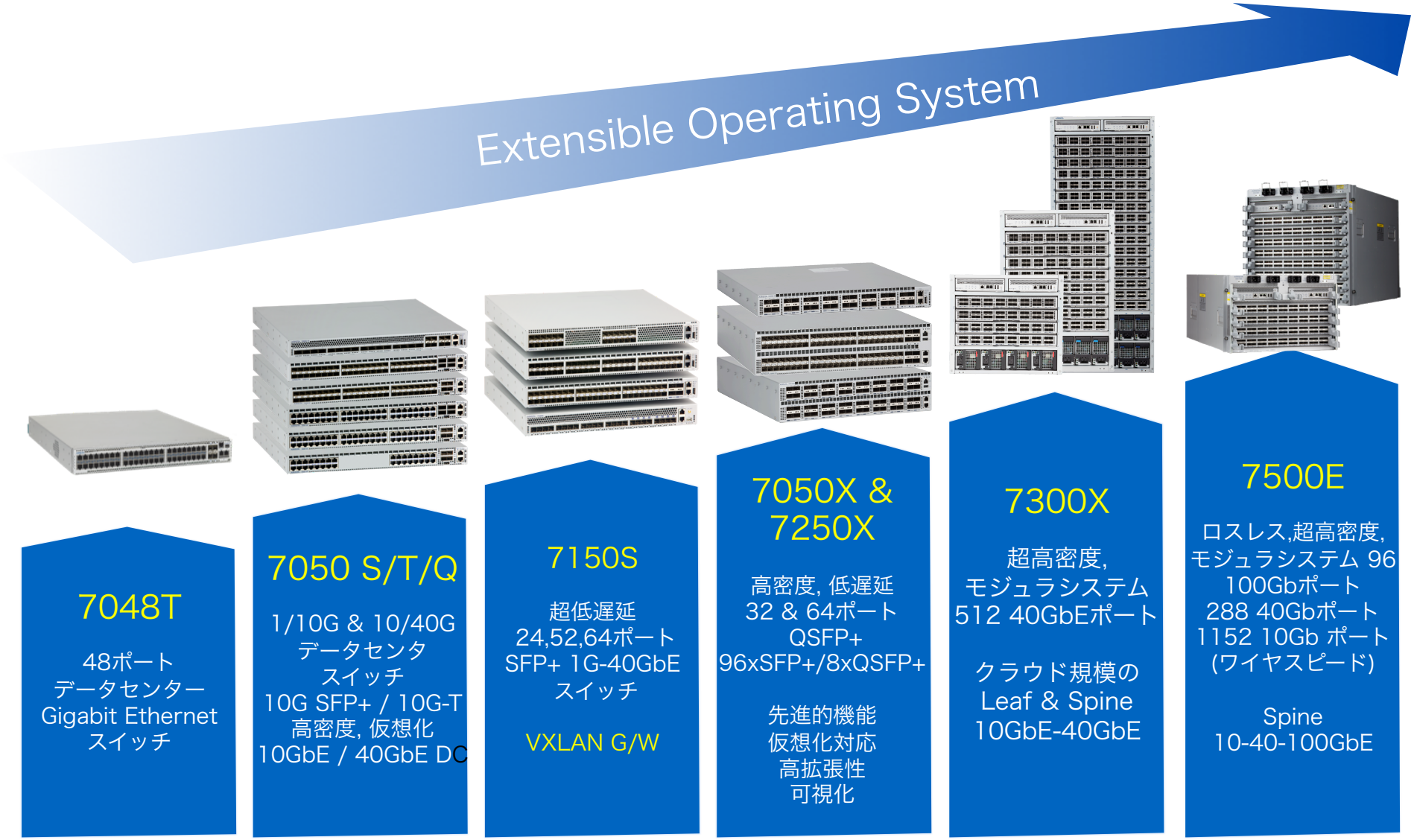
- ✓ Linuxベース
- ✓ プログラマビリティ
- ✓ オートメーション(Ansible / Chef / Puppetとの統合)
- ✓ JNOSベースのAPI

EOS – Extensible Operating System
アーキテクチャと特徴



- ✓ 完全にモジュール化されたネットワークOSで、ステートフルなリスタートを実現
- ✓ 全てのプロセスのステートを管理しプロセス間通信を仲介するsysDBを中心としたアーキテクチャ
- ✓ インサービスソフトウェアアップグレード（ISSU）
- ✓ サードバーティのアプリケーションを実行可能な拡張性
- ✓ ネットワークの運用をよりシンプルに
- ✓ 全てのアリスタ製品に単一のソフトウェアイメージ

アリスタネットワークスの製品ラインナップ



ご清聴ありがとうございました

www.aristanetworks.com/jp