JEITA講座『IT最前線』

ICTの力を、お客様のビジネスの推進力に。

Your ICT Force

SDN/OpenFlowの技術動向

ユニアデックス株式会社
山平　哲也
藤田　勝貫

UNIADEX
ユニアデックス株式会社

IT インフラにおける
アーキテクチャ面での変化

IT の普及速度は早くなっている

電話

75 年

ラジオ

38 年

テレビ

13 年

Facebook

3.5 年

LINE

1 年

Angry Birds Space

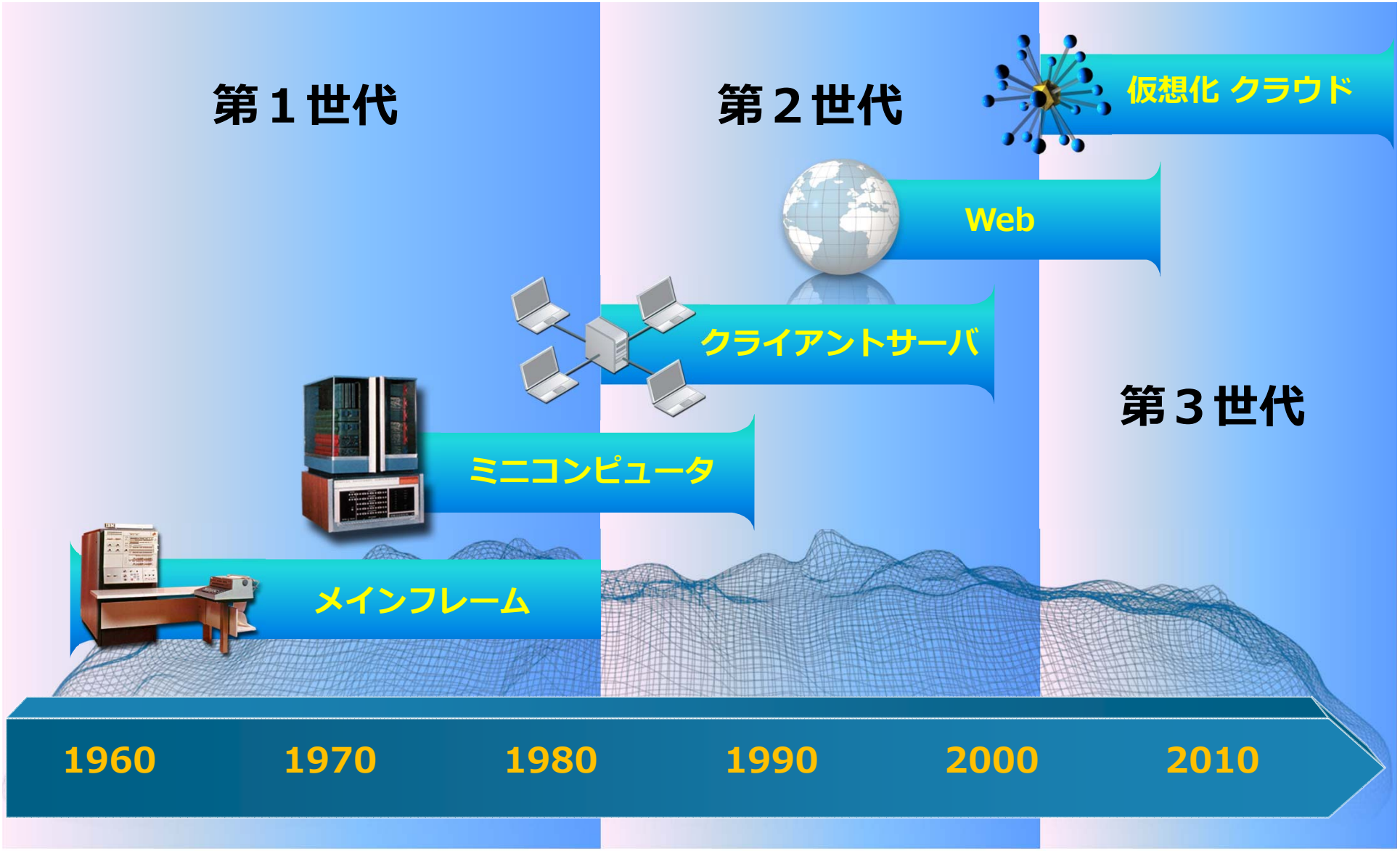
35 日

利用者 5000 万人に到達するまでの期間

http://innovation.gkofiannan.com/2012/05/01/radio-took-38-yrs-to-reach-50-million-users-o/ を参考に作成

UNIADEX

コンピューティングアーキテクチャの変遷



UNIADEX

コミュニケーション技術の変遷

UNIADEX

「アプリケーション」の変化

業務アプリケーション

データを処理

データを入力

ビックデータ データ分析

twitter　hadoop

facebook　SAP HANA

OPEN DATA
オープンデータ流通推進コンソーシアム

収益を増加する情報としてデータ分析

膨大なデータが発生

UNIADEX

毎日膨大な量のデータが生成されている

3.4 億ツイート　一日あたりのツイート数

60 億台　世界の携帯電話契約台数

309 件／秒　一秒あたりの最高注文数

6.65 億人　日間アクティブユーザ数

1.5 億人　総ユーザ数

UNIADEX

ICTからクラウドサービス連携に向けて

2000 – 2005 年
IT・通信の一体提供・運用が進展

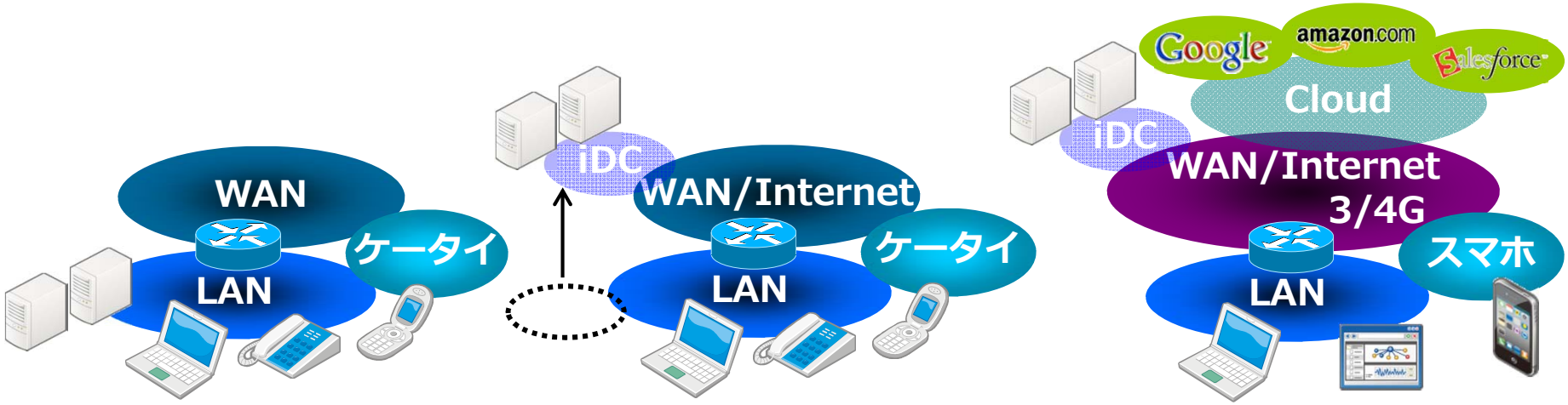
- 通信とITとの融合
- ITと通信との融合

2005 – 2010 年
データセンター集約アウトソーシング活用

- サーバ、ストレージをデータセンターに集約
- 運用のアウトソーシング

2010 年 -
クラウド本格活用サービス連携

- クラウド連携によるサービスインテグレーション
- ICT 利用基盤の拡大



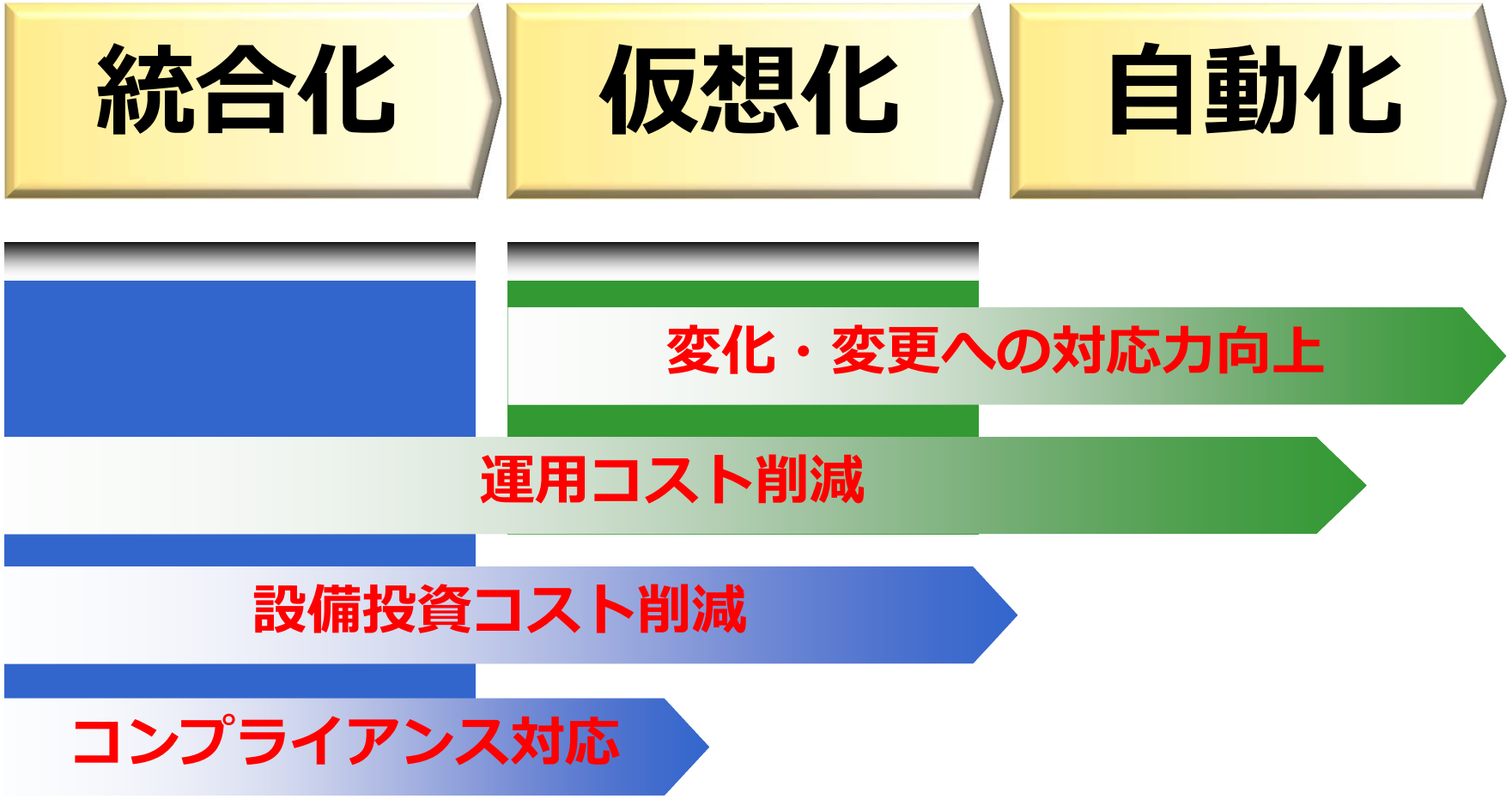
UNIADEX

クラウドの進化のためには

場所の制約
からの解放

ハードウェア
からの解放

「属人性」
からの解放



UNIADEX

現行のネットワークアーキテクチャ

ネットワークの基礎の基礎

現在使われているほとんどのネットワーク技術は次のどれかに分類される

通信相手を識別する
例：IP アドレス、ドメイン名

通信相手を探し出す
例：DNS、 ARP

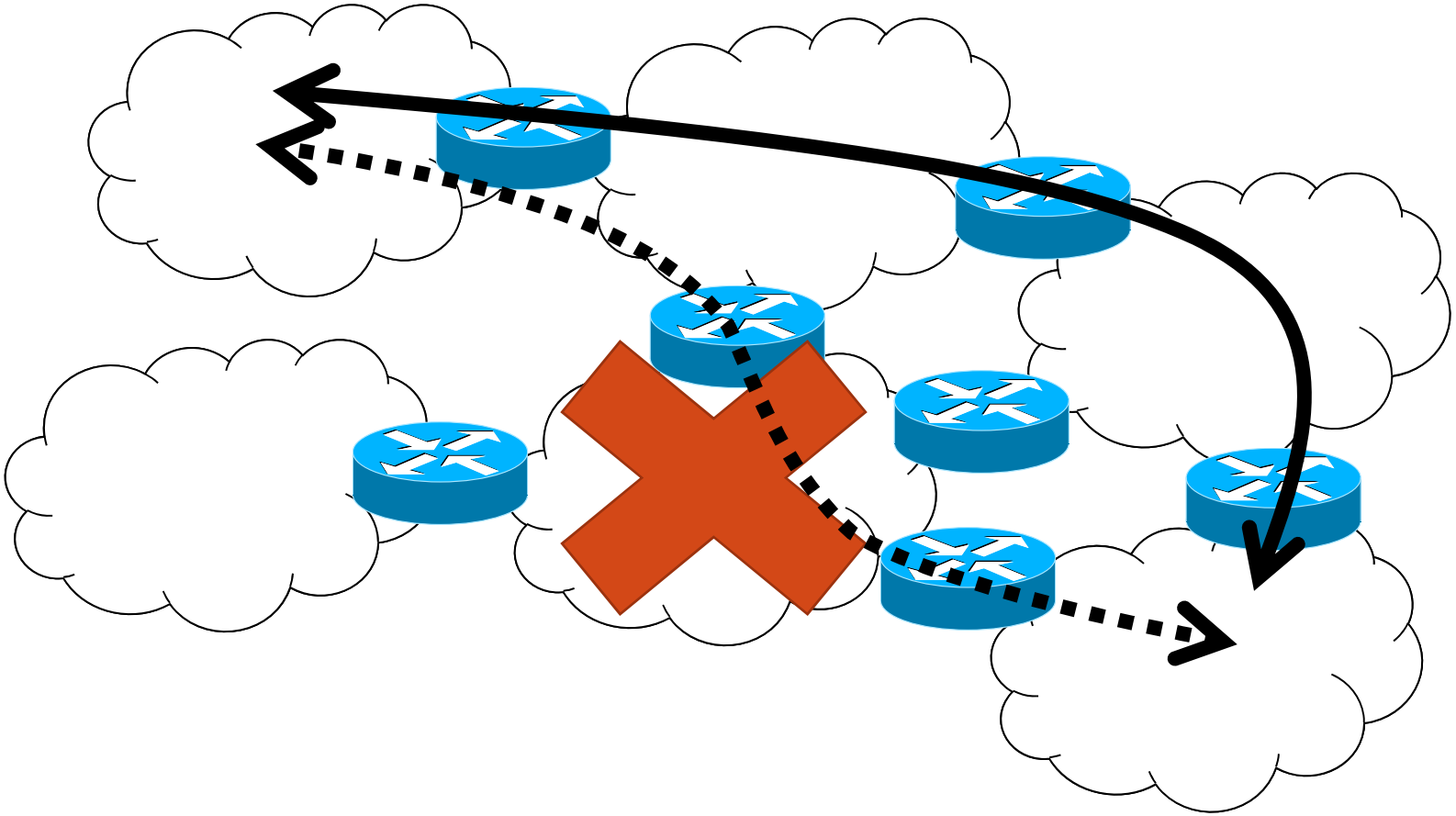
通信経路に関する情報を交換・共有する
例：OSPF、 BGP

データを送受信する
例：IP、 TCP、 UDP、 HTTP

通信の信頼性を向上する
例：STP、 VRRP

ARPANET からはじまったインターネット

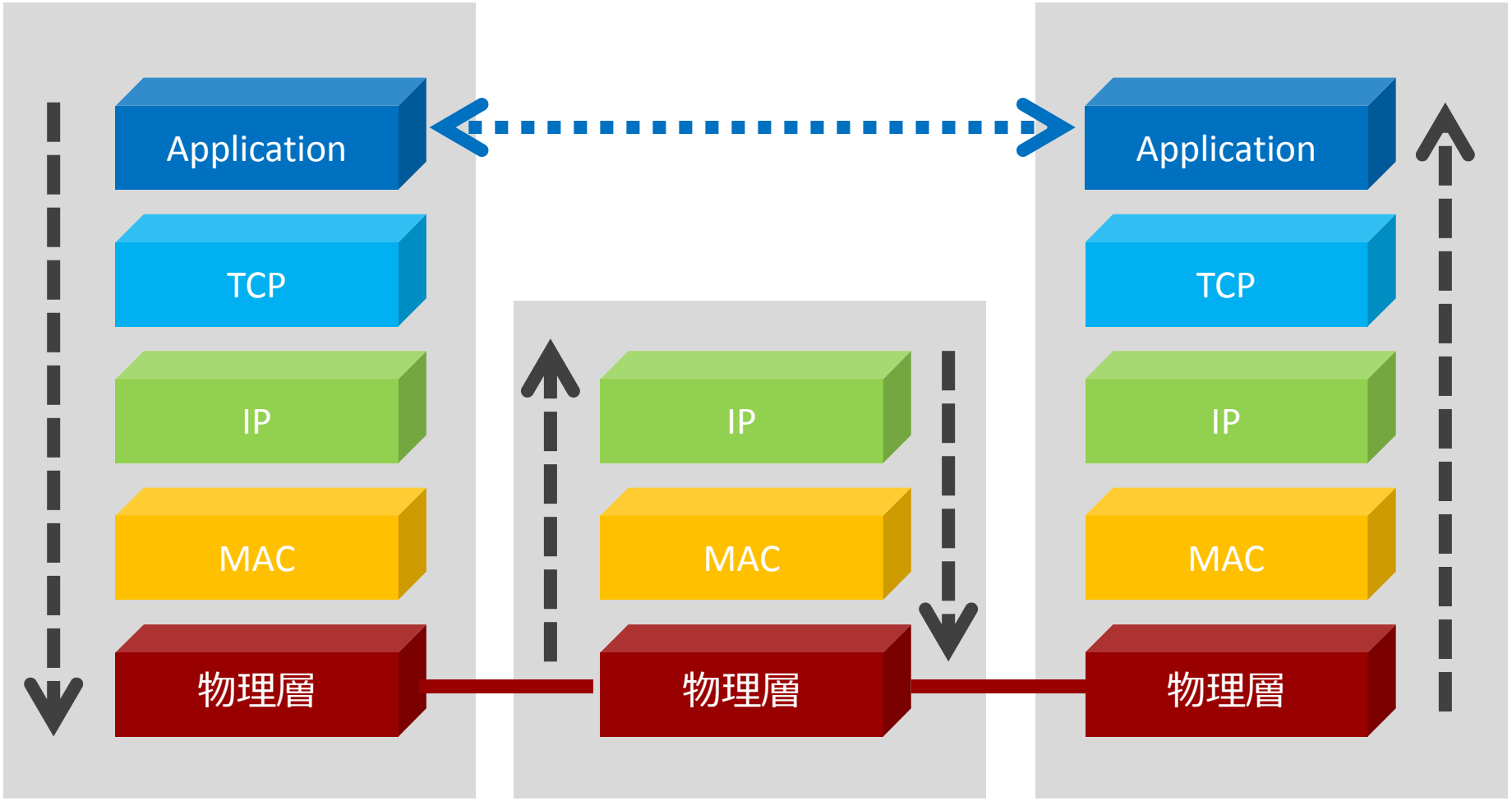
- 部分的な障害を全体に波及させない、という考え方



UNIADEX

階層構造：プロトコルスタック

各レイヤーに異なる役割を持たせ、全体が調和しながら動作



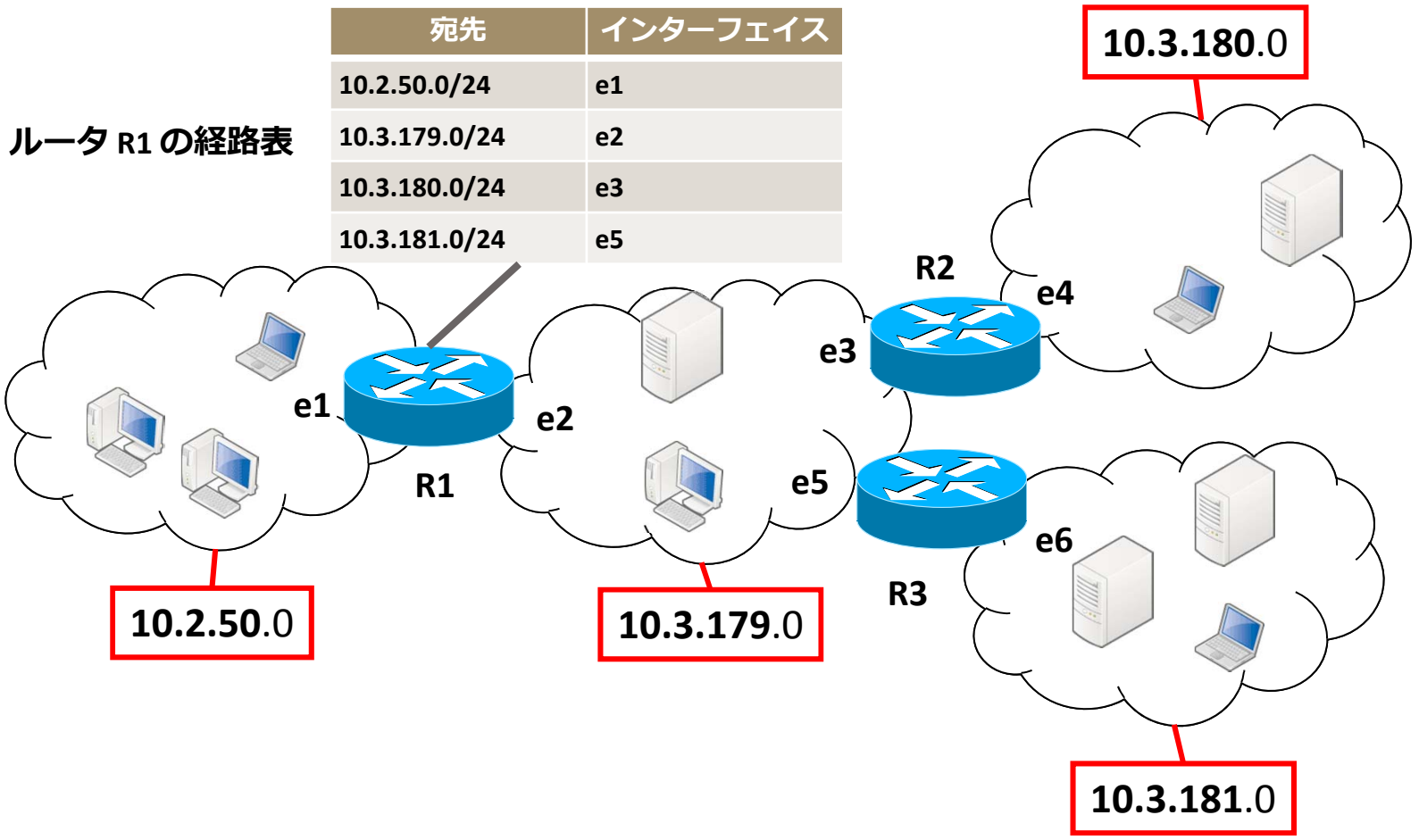
UNIADEX

ルータというもの

ルータの主な役割は２つ
１．ルーティングテーブル（経路表）を作ることでネットワークの形を部分的に知る
２．ルーティングテーブルを元に、データの送信／中継先を判断、データ転送する

ルータ R1 の経路表

宛先	インターフェイス
10.2.50.0/24	e1
10.3.179.0/24	e2
10.3.180.0/24	e3
10.3.181.0/24	e5



UNIADEX

ルーティングテーブルの作り方

- ネットワークの形（トポロジー）の変化に対応するべく、動的に作る

- 動的（自動的）に作るために使用されるのがルーティングプロトコル
  - 例　OSPF、BGP、EIGRP、etc （詳細は割愛します）
  - ほとんどのルーティングプロトコルがルーティングテーブルに含まれるトポロジー（ネットワークの形）についての情報を共有・交換するために使われる
  - **ルータ同士がルーティングプロトコルをつかって会話**する
  - ルーティングプロトコルを使用するメリットは**運用・管理が簡単になること（特に規模が大きくなるにつれて）**。

- トポロジー変化に関係なく、決めうちで静的につくる方法もある

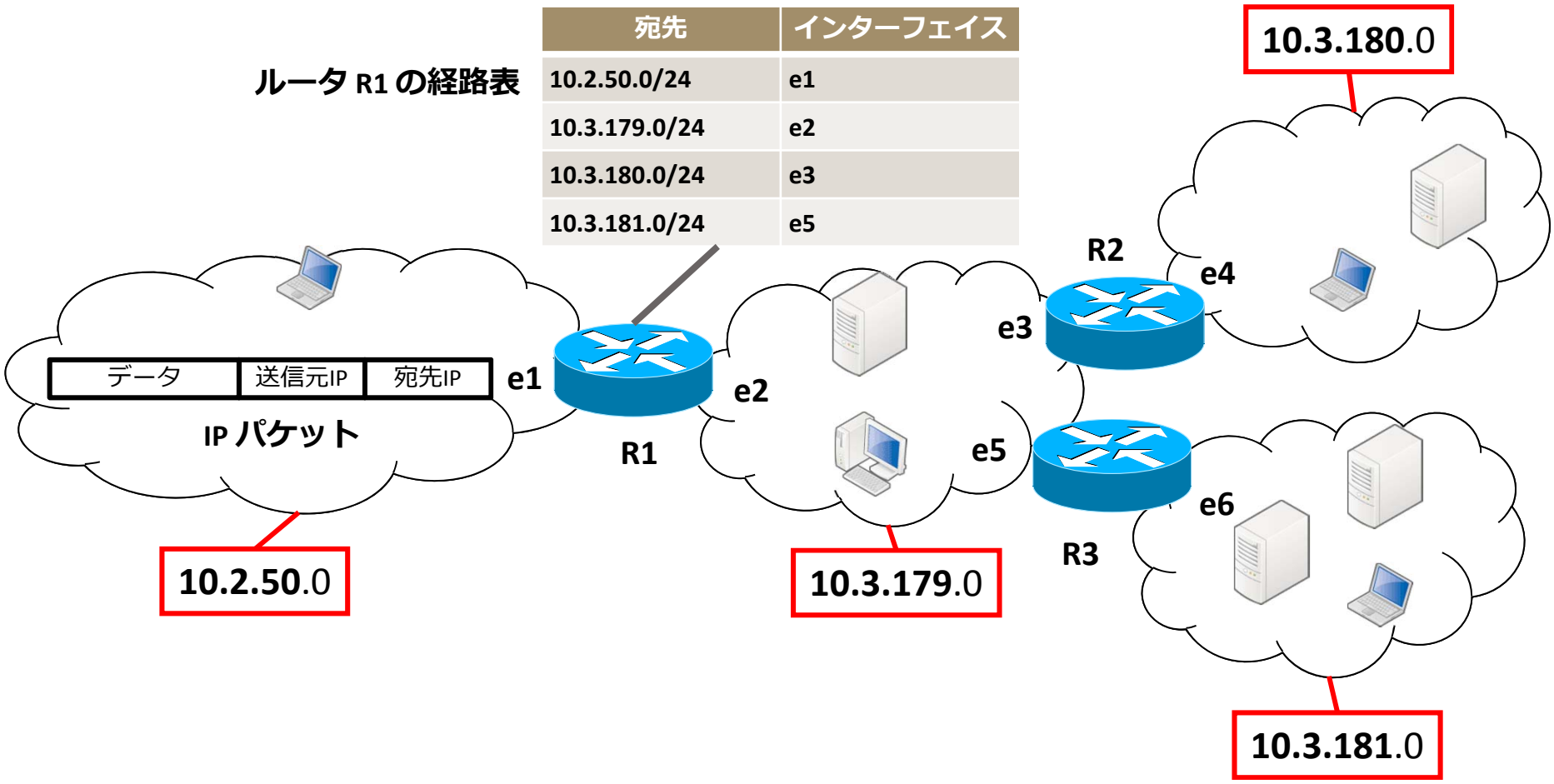
- 静的に設定するので、スタティックルーティングと呼ばれる
  - **運用・管理が煩雑になる**ので、よっぽどのことがない限り使われない。
  - つまり、スタティックルーティングがたくさんある構成はやっかいな案件になるケースがほとんど

データ(IPパケット)の転送

✓ルータに送られてくるのはIPパケット（宛先情報等とデータが含まれている）
✓宛先情報を見て、遠い（他のネットワークの）場合、次のルータにデータ転送、近い（直接接続のネットワークの）場合、宛先にデータ転送

ルータ R1 の経路表

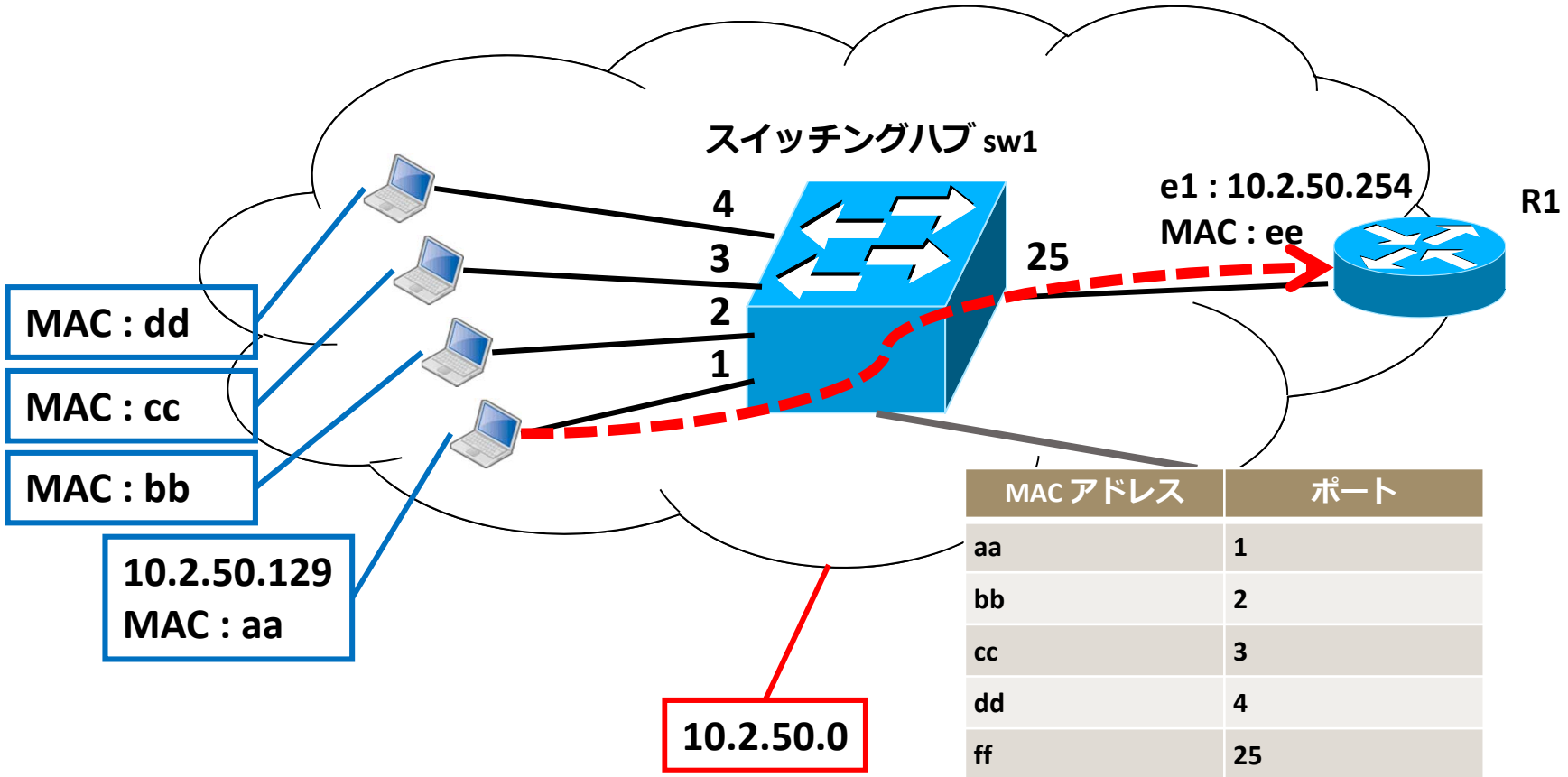
宛先	インターフェイス
10.2.50.0/24	e1
10.3.179.0/24	e2
10.3.180.0/24	e3
10.3.181.0/24	e5



UNIADEX

スイッチングハブというもの

スイッチングハブの主な役割は２つ
１．MAC アドレステーブル（経路表）を作ることでサブネットの形を部分的に知る
２．MAC アドレステーブルを元に、データの送信／中継先を判断、データ転送する



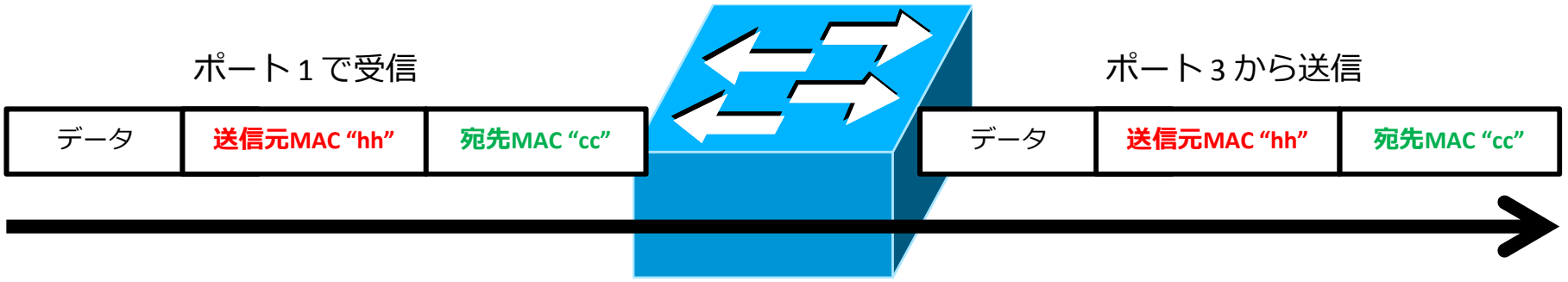
MAC アドレス	ポート
aa	1
bb	2
cc	3
dd	4
ff	25

スイッチングハブ sw1 の MAC アドレステーブル

UNIADEX

MAC アドレステーブルの作り方とデータ転送の仕方

- L2 での基本的な作り方は非常に簡単
- 流れてくるデータの送信元 MAC アドレスをテーブルに登録する
- L3 (IP) のようにプロトコルを使わないのは、接続されている端末を知るための仕組みだから



MAC アドレス
テーブルに追加

“cc” はポート 3
につながっている

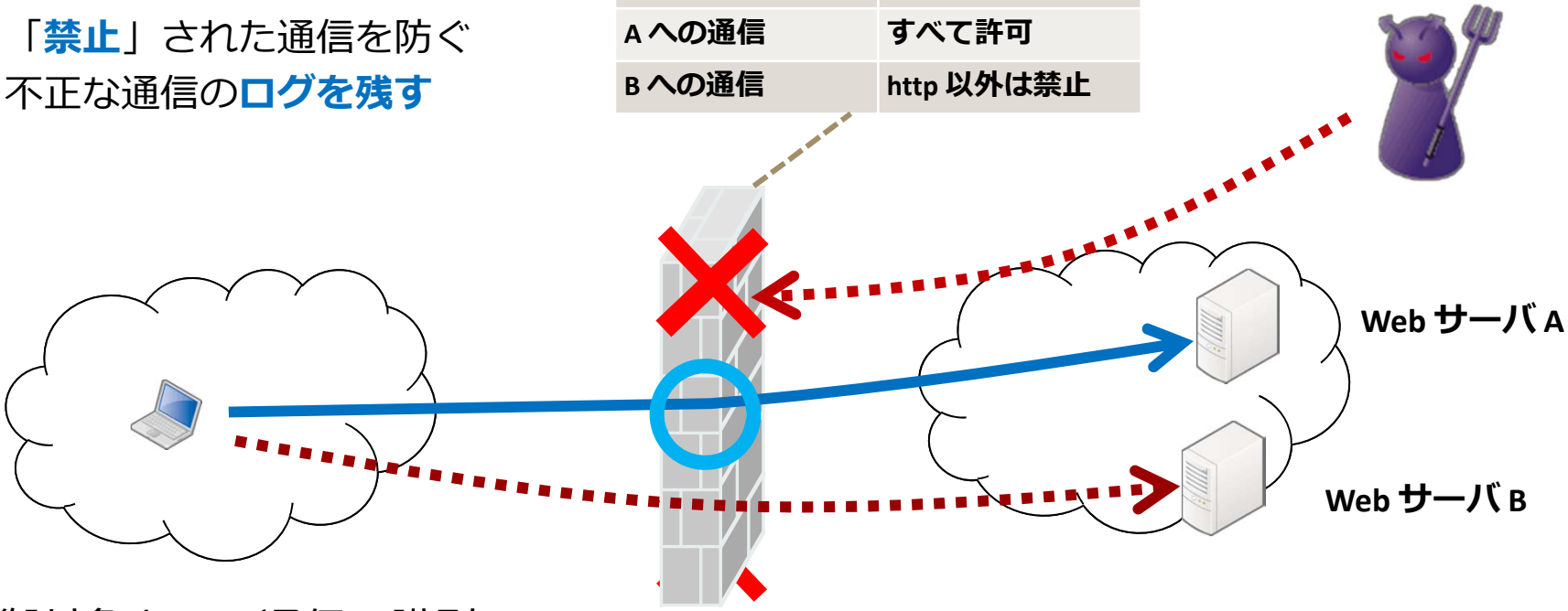
MAC アドレス	ポート
aa	1
hh	1
bb	2
cc	3
dd	4
ff	25

UNIADEX

ファイアウォール

- ファイアウォールの機能
  - 「**許可**」された通信を通す
  - 「**禁止**」された通信を防ぐ
  - 不正な通信の**ログを残す**

通信の種類	ポリシー
内部へのアクセス	すべて禁止
A への通信	すべて許可
B への通信	http 以外は禁止

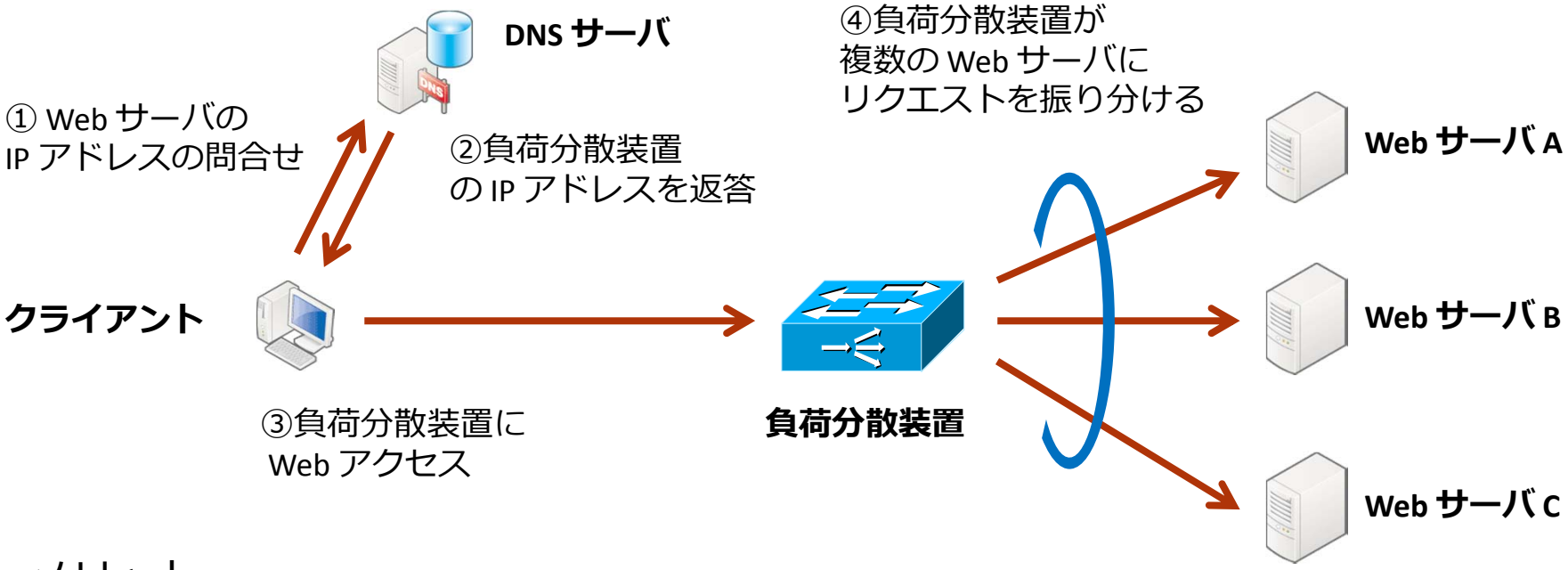


- 制御対象となる通信の識別
  - MAC アドレス、IP アドレス（Geo IP 含む）
  - TCP/UDP ポート
  - Web アプリケーション
  - User / Group 単位

UNIADEX

ロードバランサ:Web 負荷分散装置

- ロードバランサの役割 （「L4-7 スイッチ」と呼ぶ場合もある）
  - Web リクエストの振り分け → 同じクライアントからのリクエストは同じサーバに
  - Web サーバの死活監視 → 死んでいるときにはリクエストを振り分けない
  - Web サーバの性能監視 → トロいときにはリクエストを振り分けない



- メリット
  - Web サーバの負荷分散が可能＝スケーラブルな Web 基盤
  - 物理 Web サーバの停止、交換などをノンストップで

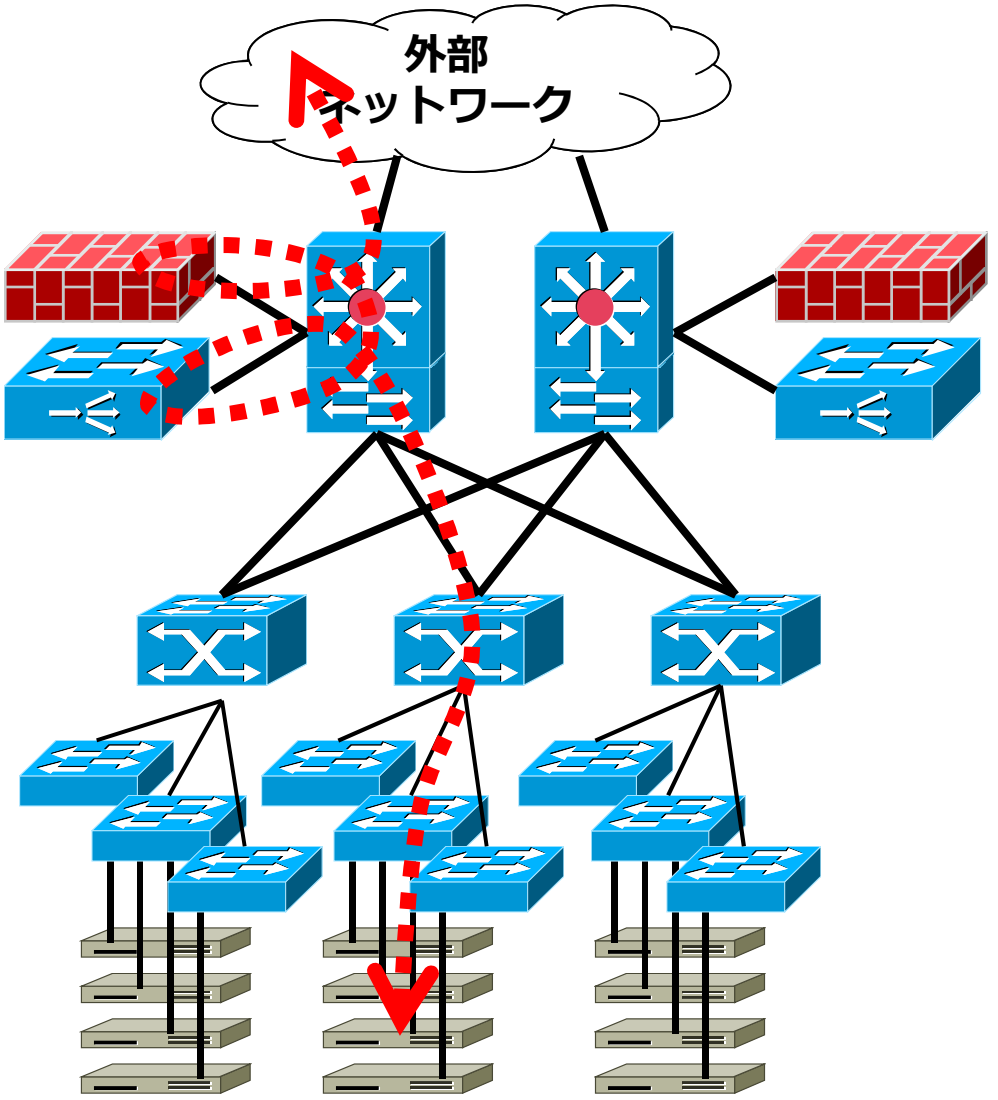
UNIADEX

現在標準的なサーバサイドネットワークの構成

コアスイッチ（ルータ）
&
サービスアプライアンス
（ファイアウォール、
ロードバランサ等）

アグリゲーション
&
アクセス
スイッチ（L2スイッチ）

各種サーバ



UNIADEX

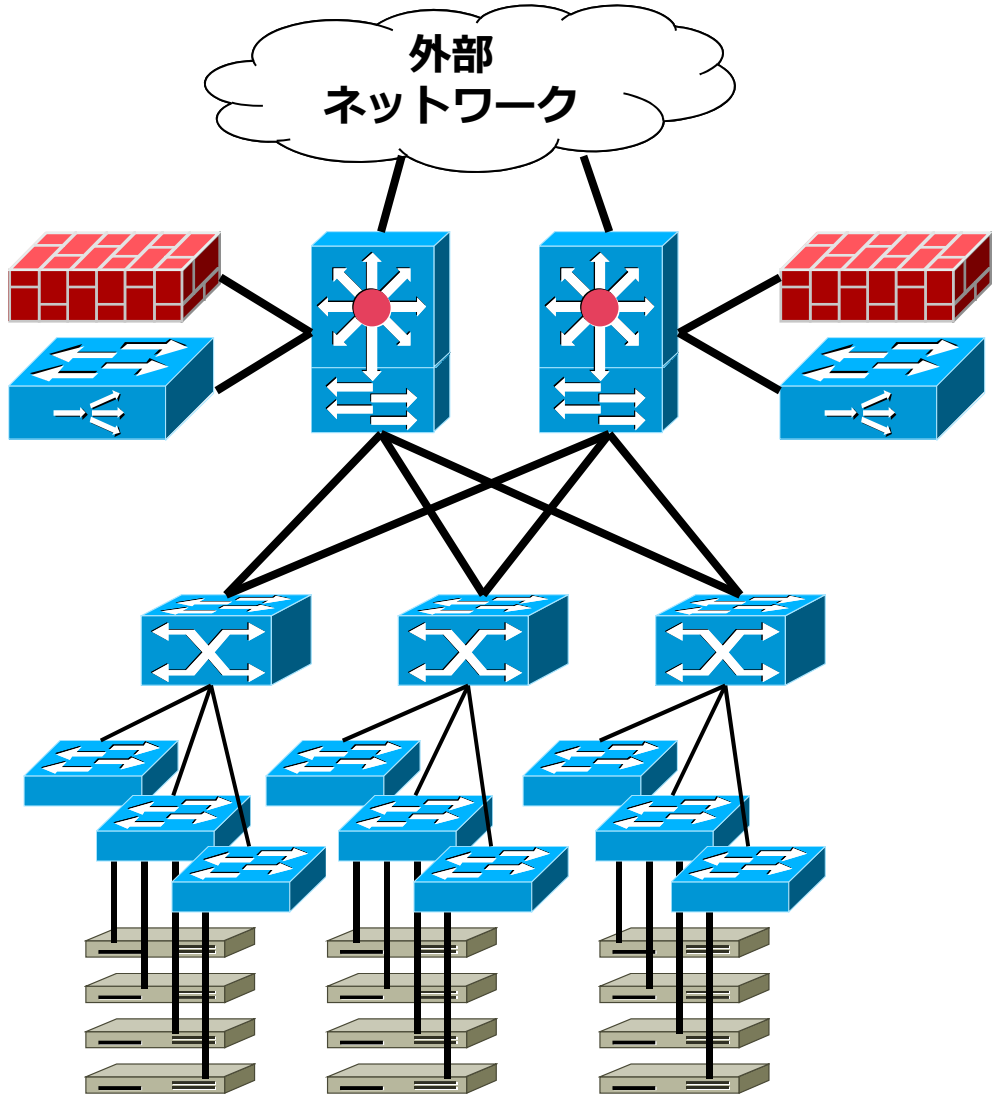
現行のネットワークアーキテクチャの課題・問題点

前提

- サーバの処理能力向上
- サーバ仮想化が普及

課題・問題点

- 迅速な構成追加・変更が難しい（特にサーバ仮想化環境で）
- 垂直トラフィックの増減に強いが、水平トラフィックの増減に弱い
- そもそもマルチテナンシーな考え方がない
- 管理対象の機器数が多い
- 複数の種類の os に対するスキルが必要



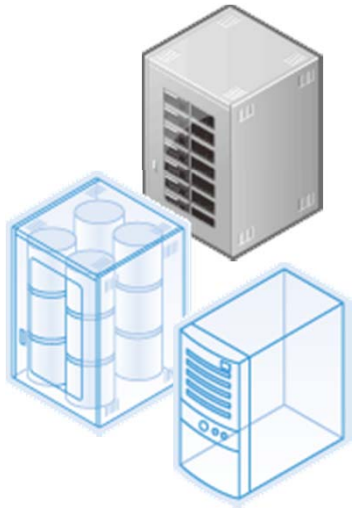
UNIADEX

クラウド時代のネットワークに何が求められるのか

サーバーコンピューティングの変化

- 仮想化によるサーバー統合（リソースの論理統合と論理分割）
- クラウドコンピューティングによる統合管理・セルフサービス化

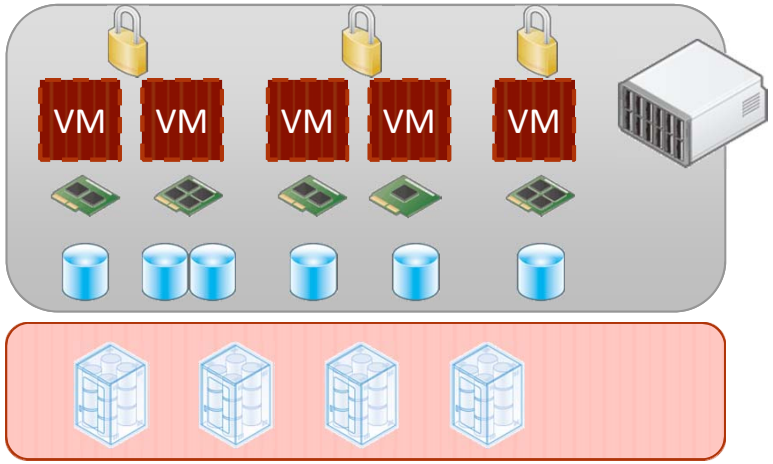
物理 → **サーバー仮想化** → **クラウド** → **マルチクラウド**



UNIADEX

クラウド管理ツール：CloudStack

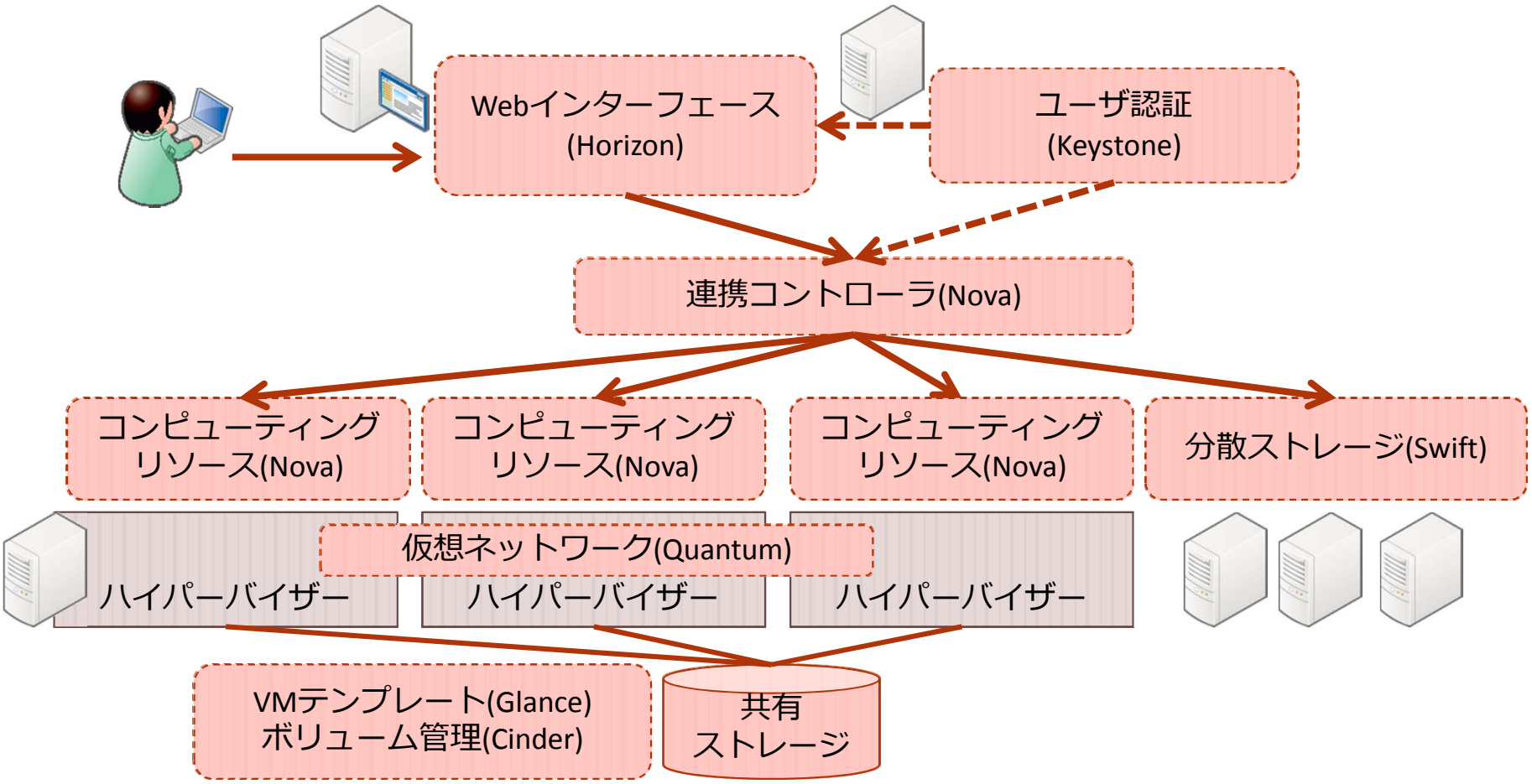
- CloudStackはOSSベースのクラウド構築ソフトウェア
  - セルフプロビジョニング
  - マルチテナントリソース管理
  - マルチハイパーバイザー対応



UNIADEX

クラウド管理ツール：OpenStack

- OpenStackは、2010年にRackspace CloudとNASAによって始められたクラウドコンピューティングプロジェクト。OSS。



クラウドとは

- クラウド・コンピューティング　（NIST定義 2011）

  コンピューティング資源(ネットワーク・サーバー・ストレージ・アプリケーション・サービス)の共有プールへの、オンデマンドなネットワークアクセスを可能にするモデル。・・・

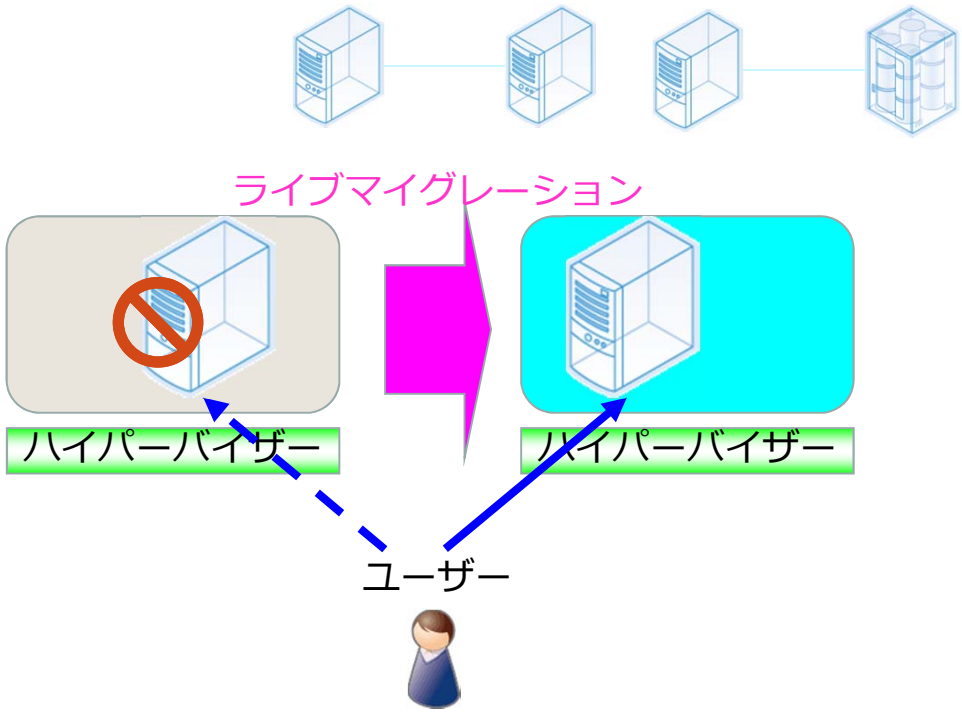
  【基本的な特徴】

  - **オンデマンド・セルフサービス（On-demanded self-service ）→必要に応じて、自動的に**
  - **幅広いネットワークアクセス　（Broad network access）**
  - **リソースの共有　（Resource pooling ）→マルチテナント**
  - **スピーディな拡張性（Rapid elasticity ）→素早くスケール変更（スケールイン＆アウト）が可能**
  - **サービスが計測可能であること　（Measured Service ）**

新たにネットワークに期待されることは？

仮想化、クラウド化でネットワークに何が求められるか？

- 仮想マシン同士の通信
- ライブマイグレーション対応
  - 同一ＬＡＮ
  - 拠点間



- テナント分離
  - VLAN数の制限 4,094 (IEEE802.1Q)
- セルフサービス化による自動化対応
  - API アプリケーションプログラムインタフェース

16ビット	3ビット	1ビット	**12ビット**
TPID	TCI		
	PCP	CFI	**VID**

ネットワークの動的な変更ができることが重要

UNIADEX

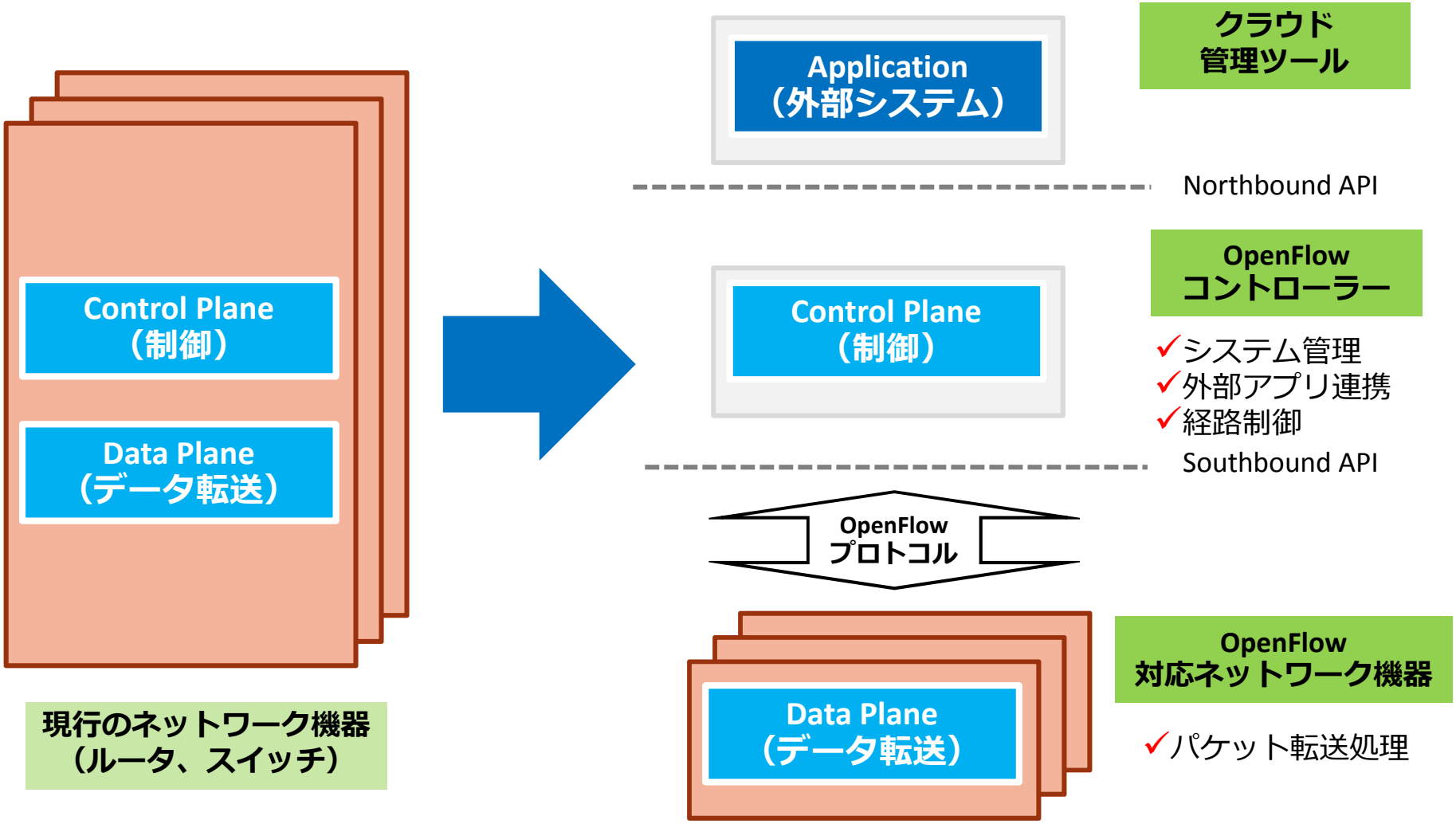
SDN/OpenFlowの技術

SDNとは？

- SDN（Software Defined Network）とは，ネットワークの構成や機能をソフトウェアで動的に制御できるアーキテクャー
  - 従来のネットワーク機器に密に結合されていたコントロールプレーンとデータプレーンを分離し，ネットワークの状態管理や高度な制御を一元的に実現する。
  - コントロールプレーンとデータプレーン
    - コントロールプレーン　・・・　ネットワーク制御機能
    - データプレーン　・・・　データ転送機能

SDNでプログラマブルなネットワークを

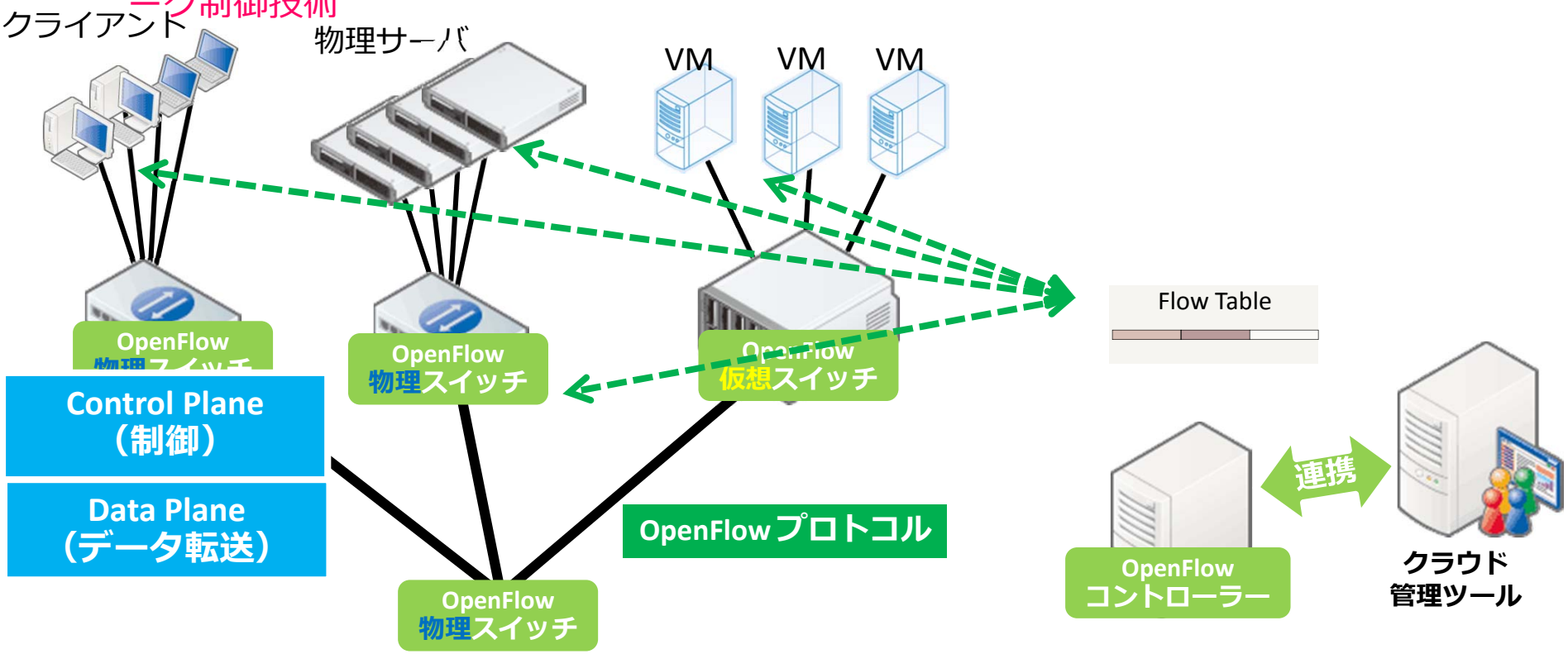
- ソフトウェアで動的に制御できる＝プログラマブル



UNIADEX

OpenFlow とは？

- スタンフォード大学を中心とした「OpenFlowスイッチングコンソーシアム」によって提唱され、Open Networking Foundation(ONF)によって標準化されている次世代ネットワーク制御技術



従来のネットワーク制御方式

- ✓**自律**的に**分散**して制御するアーキテクチャ
- ✓MAC アドレス、IP アドレス等による経路制御

OpenFlow

- ✓トラフィックを**集中制御**するアーキテクチャ
- ✓「**フロー**」単位の経路制御

UNIADEX

OpenFlow 仕様

- OpenFlow 関連仕様
  - Specification OpenFlow プロトコルの仕様本体
  - OF-Config　OpenFlow スイッチセットアップ時のコンフィグレーション
  - OF-Test　OpenFlow をサポートする機器間の相互接続性試験の仕様

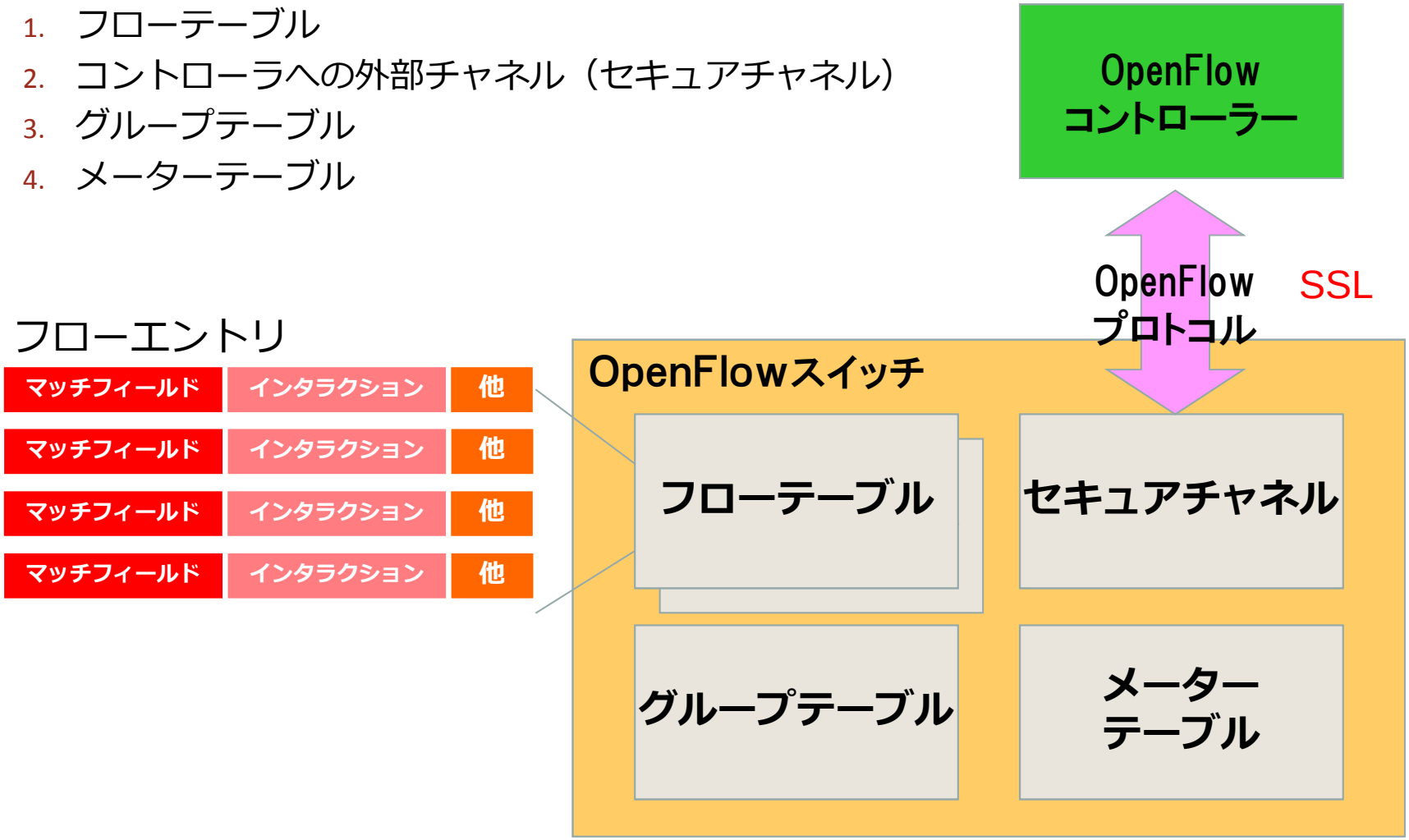
2012年6月ドラフト版提供予定

- Specification リリースロードマップ

バージョン	リリース時期	サポート機能
1.0	2010年1月	MACアドレス、802.1Dスパニングツリー、IPv4アドレス、VLANタグ、ARP、フローテーブル(シングルインスタンス)
1.1	2011年2月	MPLS、フローテーブル (マルチインスタンス)
1.2	2011年12月	IPv6(フローテーブルの条件として),拡張マッチング,実験用拡張 複数コントローラー対応
1.3	2012年3月	IPv6（ルーティング等のサポート), QoS, VPN, PBB

OpenFlowスイッチの構造

- OpenFlowスイッチは以下で構成される
  1. フローテーブル
  2. コントローラへの外部チャネル（セキュアチャネル）
  3. グループテーブル
  4. メーターテーブル



フローエントリの構造1

- フローエントリは以下の項目から構成される
  - マッチフィールド
  - プライオリティ値
  - カウンタ値
  - インストラクション
  - タイムアウト値
  - クッキー

- マッチフィールド（一部）
  - 送信元MACアドレス
  - 宛先MACアドレス
  - IPプロトコル
  - IPv4送信元アドレス
  - IPv4宛先アドレス
  - TCP送信元ポート番号
  - TCP宛先ポート番号
  - UDP送信元ポート番号
  - UDP宛先ポート番号

フローエントリの構造2

- インストラクションの種類
  - Meter パケットをメーターエントリに送る
  - Apply-Actions パケットに対し、アクションを実行する
  - Clear-Actions アクションセットの中のアクションを削除する
  - **Write-Actions アクションセットにアクションを追加する**
  - Write-Metadata メタデータを利用するフィールドに対してメタデータを書き込む
  - **Goto-Table パケットを送る次のフローテーブルを指定する**

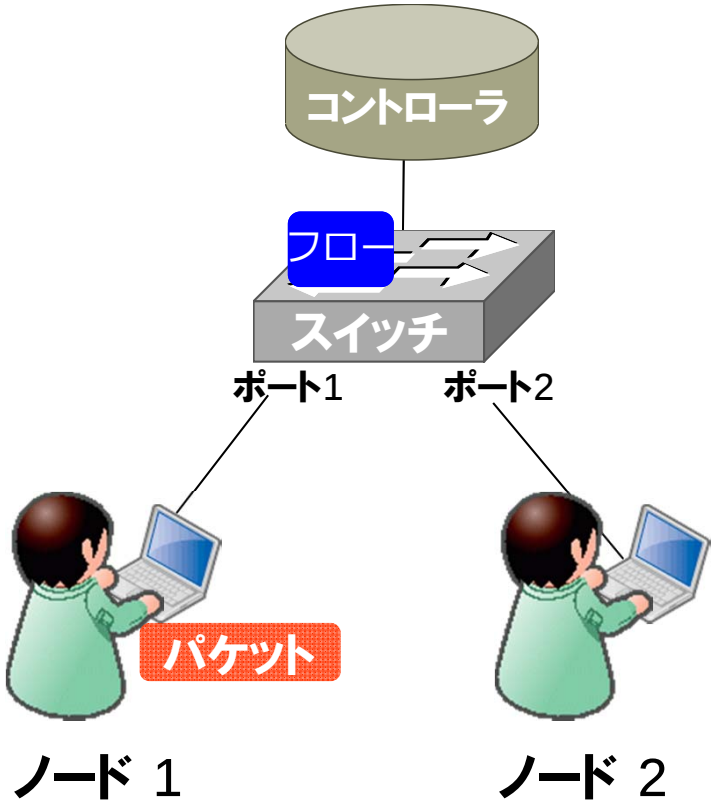
- **アクションの種類（一部）**
  - **Output パケットを出力する**
  - **Drop パケットを破棄する**
  - **Group パケットを特定のグループエントリを使って処理する**

グループテーブルの構造

- グループテーブルは、グループエントリを保持する。
- グループエントリは、複数のフローに対して同一の一連の処理を施すために利用される
- グループエントリの構造は以下の通り
  - グループＩＤ
  - グループタイプ
  - カウンタ値
  - アクションバケット

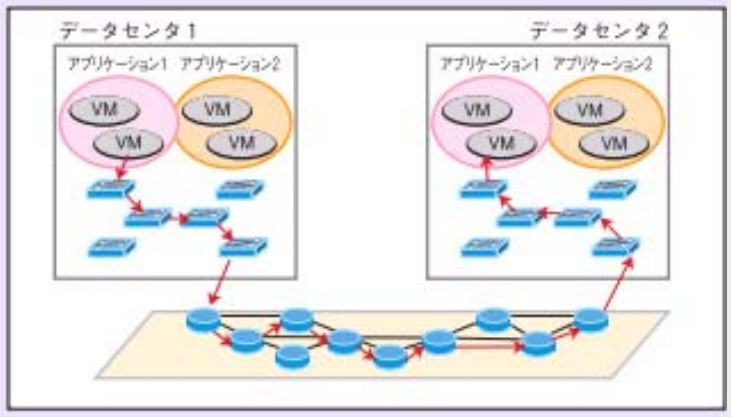
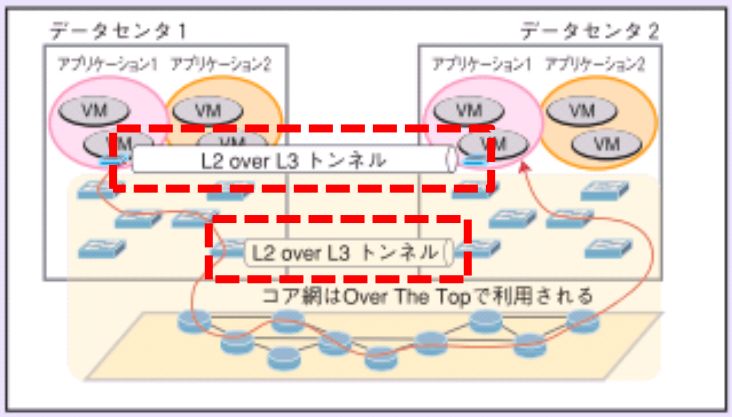
OpenFlow 動作概要

- 動作例



① ノード1は、ノード2に対しパケットを送信
② スイッチは、受信パケットに対応するフローの存在を確認
(対応するフローが、フローテーブル内に存在しない場合)
スイッチは、パケットをバッファに保存し、コントローラへ問合せを実施 (Packet-In メッセージ)
④ コントローラは、パケットの内容に基づき、制御方法を決定
⑤ コントローラは、パケットを ポート2から転送し、(Packet-out)
フローテーブルにエントリを追加する指示 (Flow-Mod)
⑥ スイッチは、ポート2からパケットを送信
指示されたフローをフローテーブルに追加
⑦ ノード2は、スイッチ2よりパケットを受信

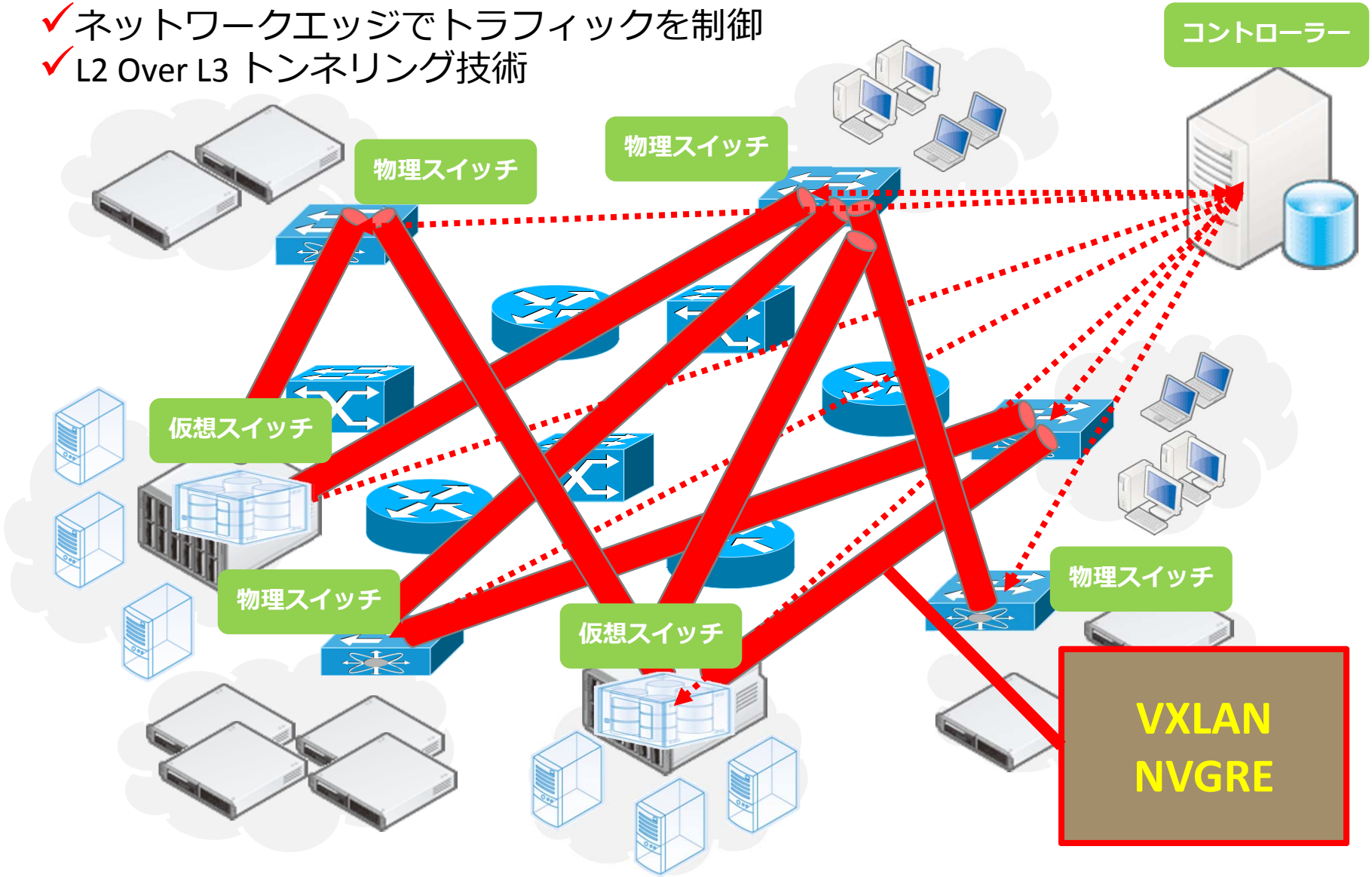
OpenFlow ： 2**種類のアーキテクチャ**

	ホップ・バイ・ホップ	オーバーレイ
概要	●**すべてのスイッチ（物理・論理）**に対してフローを設定して、トラフィックを制御 ●コントローラは**フロー**を管理	●物理ネットワーク上に**トンネリング技術**による論理ネットワークを構成して、トラフィックを制御 ●コントローラは**トンネル**を管理
長所	●OpenFlow の特徴を存分に活用可能	●既存網を利用し、レイヤ2ネットワークを柔軟に構成可能
短所	●すべてのスイッチで OpenFlow 対応が必要	●トンネリングのスケーラビリティが不明
監視方法	●全機器の一元管理が可能	●既存網とオーバレイ部分が分離
採用シナリオ	●新規にDCネットワークを構築する場合	●既存ネットワークと混在する場合
構成イメージ		

UNIADEX

オーバレイ・アーキテクチャ

✓ネットワークエッジでトラフィックを制御
✓L2 Over L3 トンネリング技術

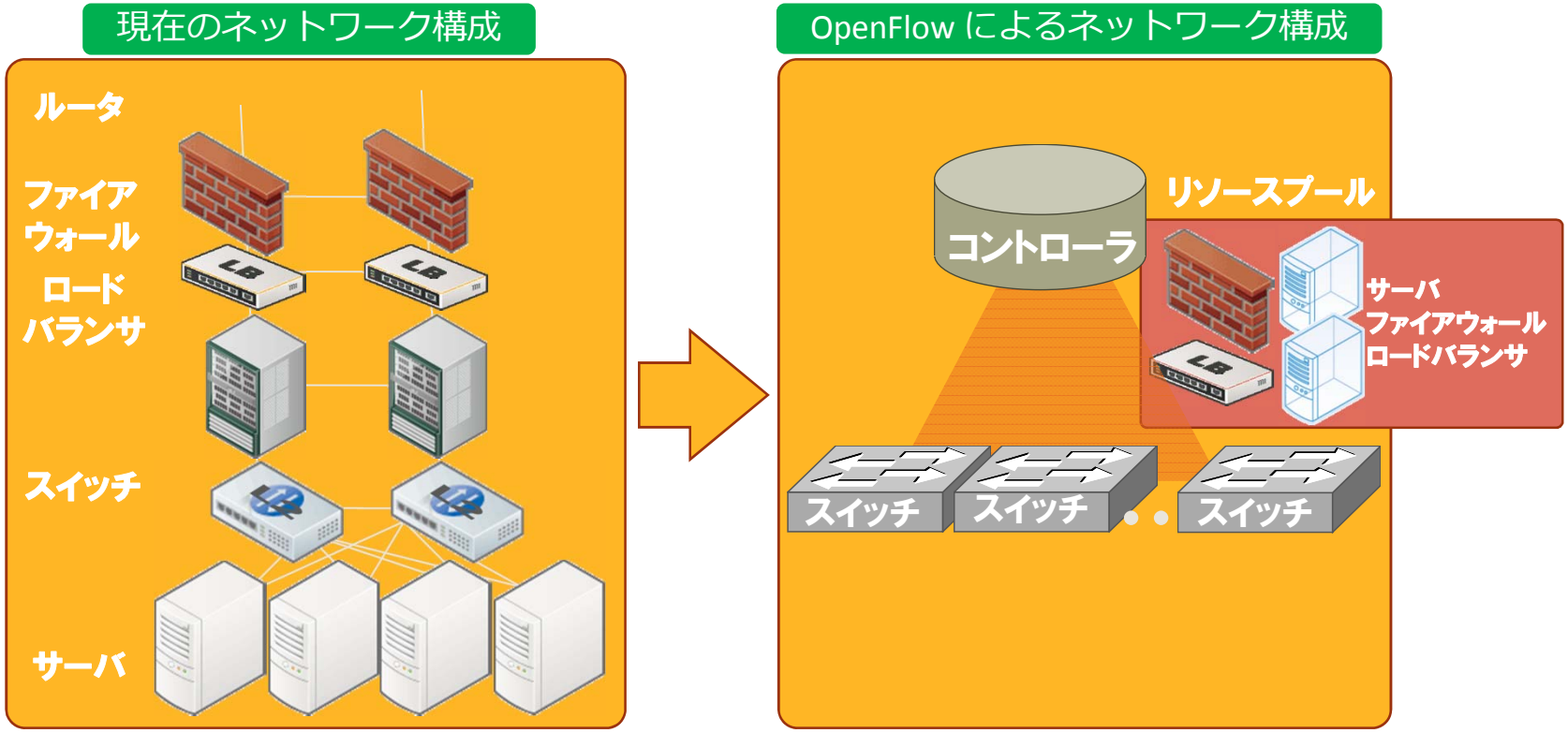


UNIADEX

OpenFlowへの期待と最新動向

OpenFlowへの期待

- シナリオ１　ネットワーク構成の簡素化

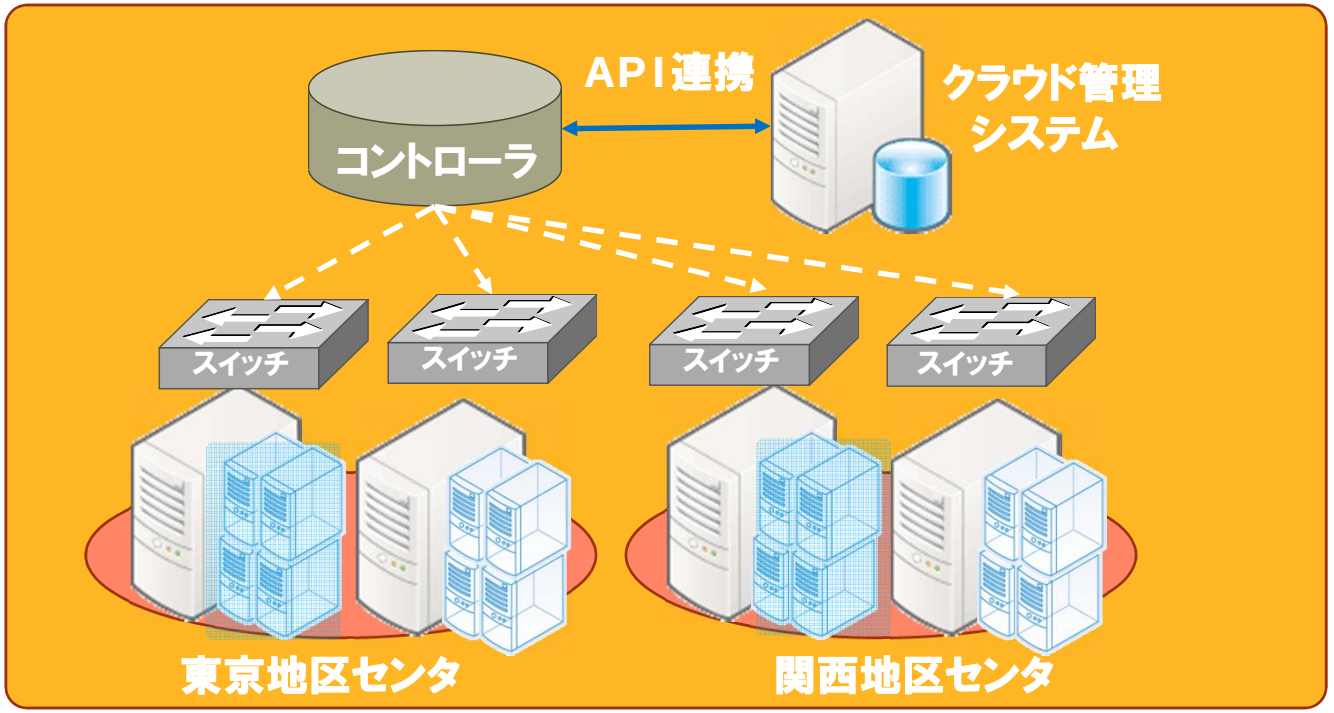


・ネットワーク階層のフラット化
　→高価なコアスイッチを削減可能
・ネットワークリソースのプール化
・管理インタフェイスの集中化

UNIADEX

OpenFlowへの期待

- シナリオ２　クラウドインフラの高度化
  - 課題:CMS/セルフポータルで準備できるのは仮想サーバインスタンスのみ
  - 解決策:仮想サーバ～ネットワークが連動した、一元的なプロビジョニング、新規仮想環境作成・設定変更の実施

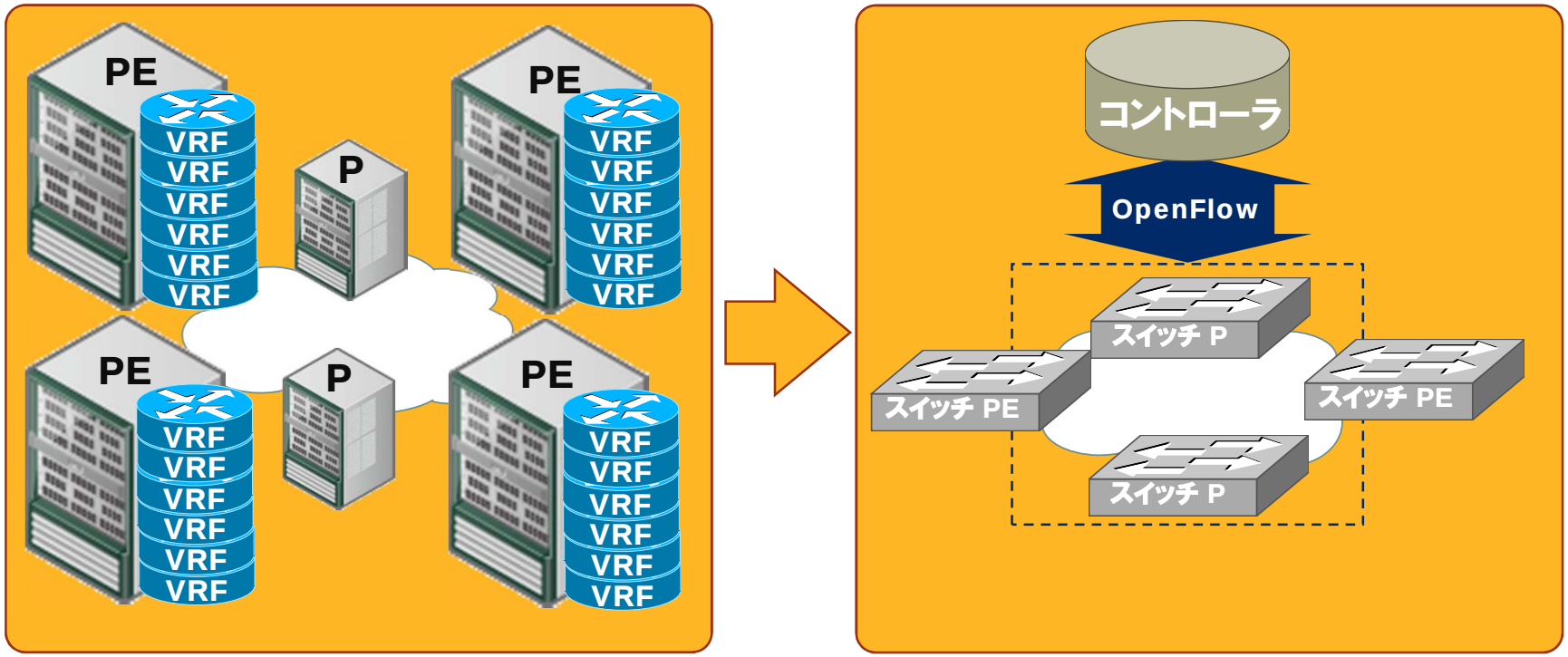


・CMSと運用の一元化
・機械化、自動化の実現

UNIADEX

OpenFlowへの期待

- シナリオ３　高コスト機器のリプレース
  - 今後、高価なMPLS-VPN設備を、安価な OpenFlow 環境で置換可能

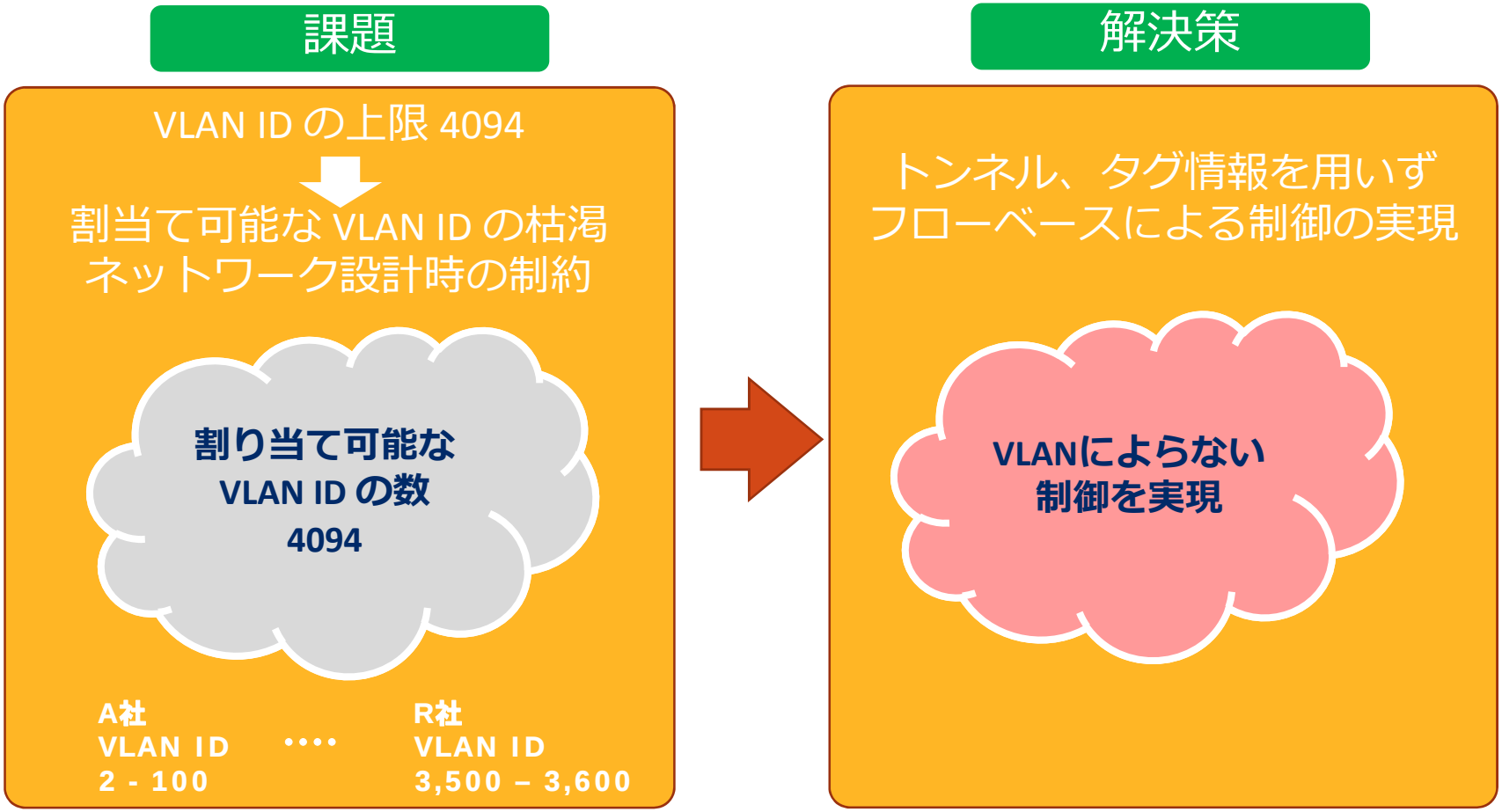


コスト低減の実現

UNIADEX

OpenFlowへの期待

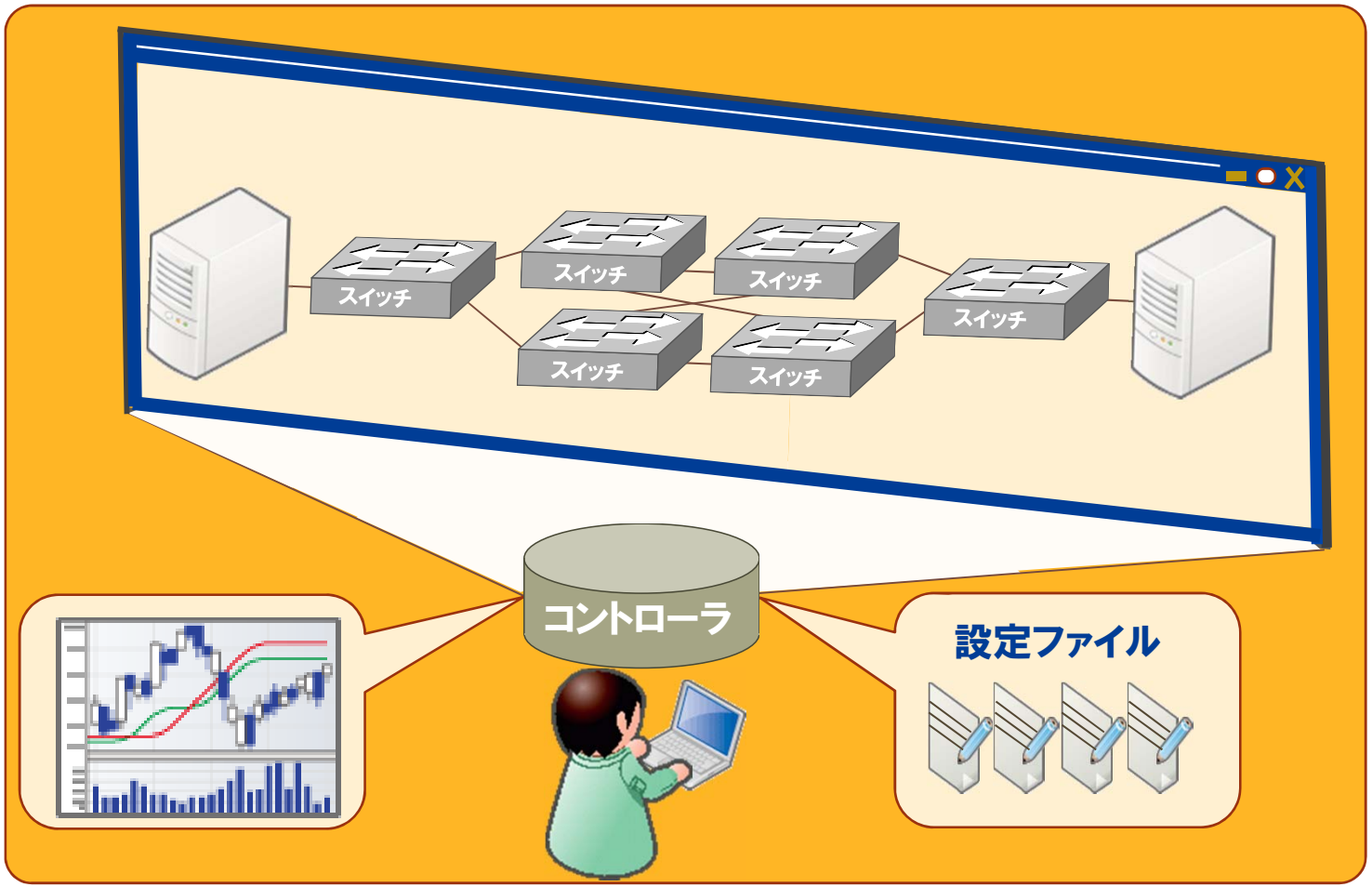
- シナリオ４ VLAN IDの上限数に制約されない、マルチテナンシの実現



UNIADEX

OpenFlowへの期待

- シナリオ５　ネットワークの可視化と一元管理
  - 一元的なネットワーク設定環境の提供
  - フロー単位の可視化による、きめ細かな運用管理の実現



UNIADEX

トンネリング技術紹介

VXLAN

- 概略
  - Virtual eXtensible Local Area Network
  - 2011/8/26 に IETF に Internet Draft (Experimental) として提出、現在 Draft 0.2 (2012/8/22)
  - 提案者：Cisco, Vmware, Citrix, Red Hat, Broadcom, Arista
  - サポートベンダー：**増加中**
    - Cisco Nexus1000V、VMware vCloud Suite、F5 BIG-IP、Arista Arista 7150、日立電線 Apresia
    - Broadcom StrataXGS Trident II (スイッチ用 LSI、NVGREも対応)

- プロトコル概要

⇒ Layer 3 ネットワーク上でのLayer 2 ネットワークのオーバーレイ

⇒ MAC in UDP によるアプローチ

→ UDP：トンネル管理がステートレス

⇒ **VXLAN Network Identifier (VNI)** 24 ビット ID で VXLAN セグメントを識別

→ **1600 万セグメントまでスケール可能**

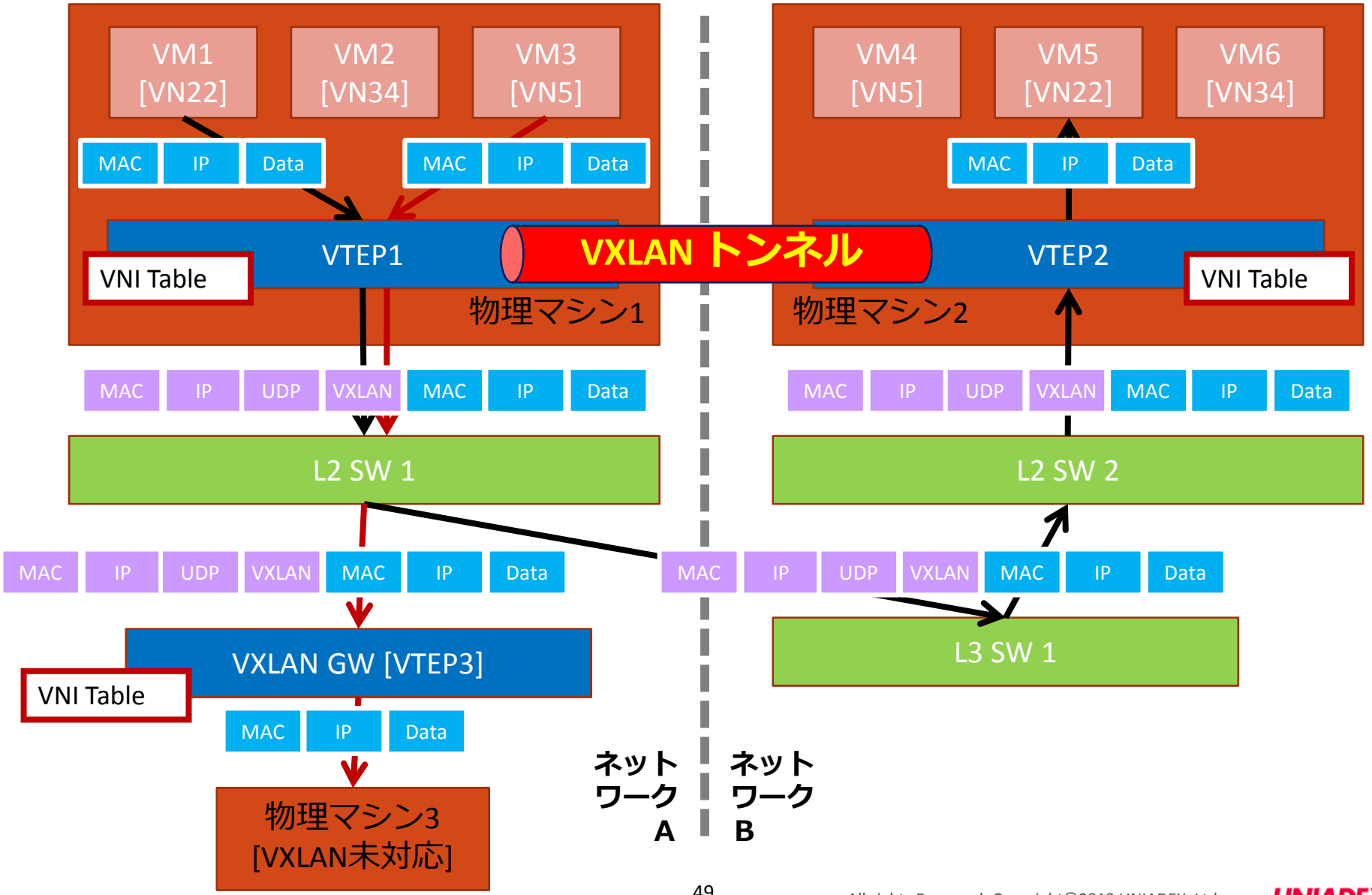
⇒ ハイパーバイザーあるいは物理サーバ単位に設置された VTEP がトンネルの終端ポイントとなる

→ VM 側では VTEP あるいは VXLAN を意識する必要がない

※ VTEP : VXLAN Tunnel End Point

UNIADEX

VXLAN 動作概要



UNIADEX

NVGRE

- 概略
  - Network Virtualization using Generic Routing Encapsulation
  - 2011/9 に IETF に Internet Draft (Informational) として提出
  - 提案者：HP, Dell, Microsoft, Intel, Broadcom, Emulex, Arista
  - サポートベンダー：**まだ少ない**
    - Microsoft Server 2012 Hyper-V
    - Broadcom StrataXGS Trident II (スイッチ用 LSI、VXLANも対応)

- NVGRE 概要

⇒ Layer 3 ネットワーク上でのLayer 2 ネットワークのオーバーレイ

⇒ MAC in GRE によるアプローチ

→ GRE：トンネル管理がステートレス → ただ IP Fragmentation 制御などでステートフルにする必要も……

⇒ **Tenant Network Identifier (TNI)** 24 ビット ID で NVGRE セグメントを識別

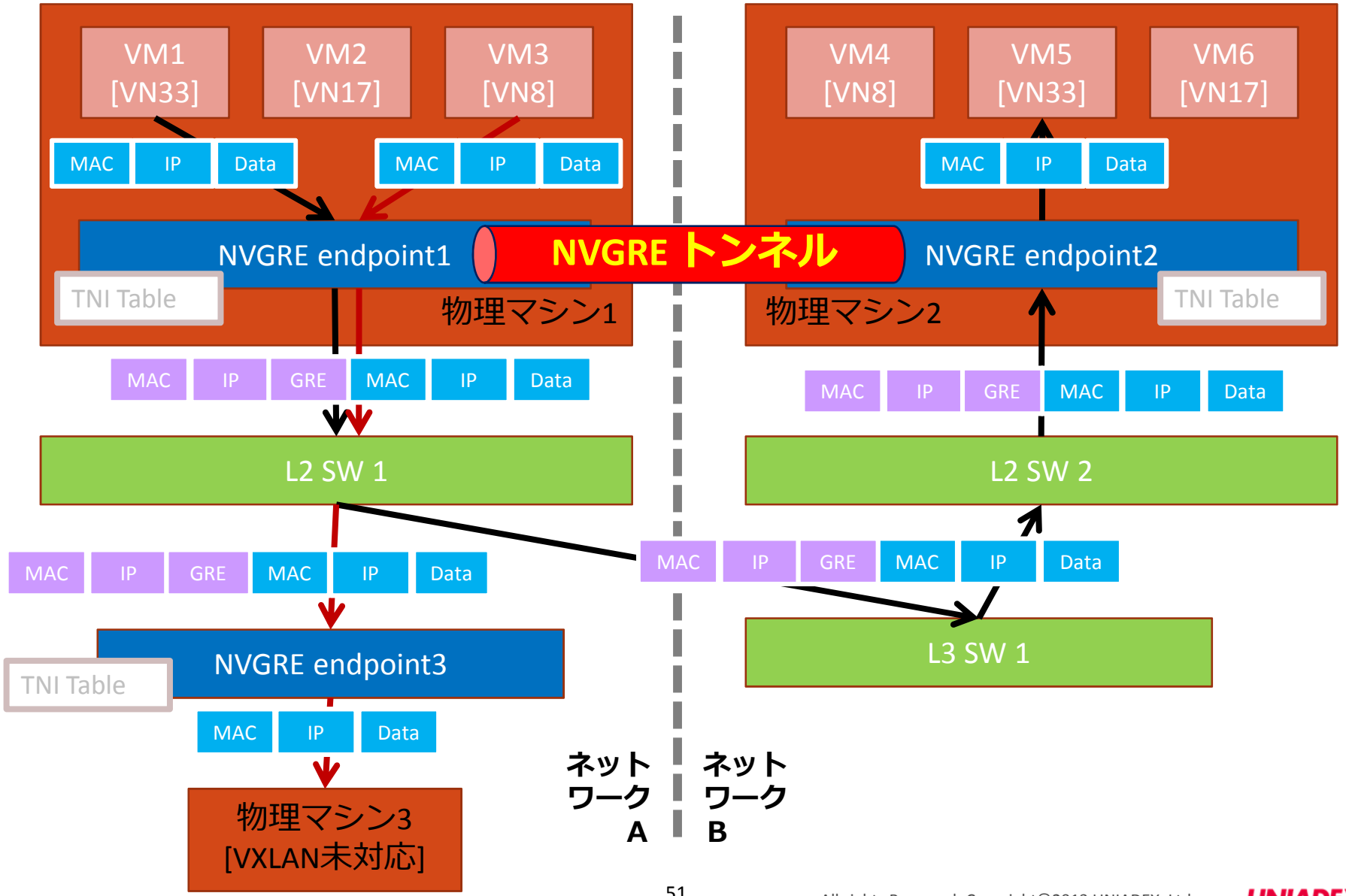
→ **1600 万セグメントまでスケール可能**

⇒ TNI ID を含むヘッダには RFC2890 で定義されている “Key and Sequence Number Extensions to GRE” (一応 Proposed Standard) を活用

⇒ ハイパーバイザーあるいは物理サーバ単位に設置された NVGRE endpoint がトンネルの終端ポイントとなる

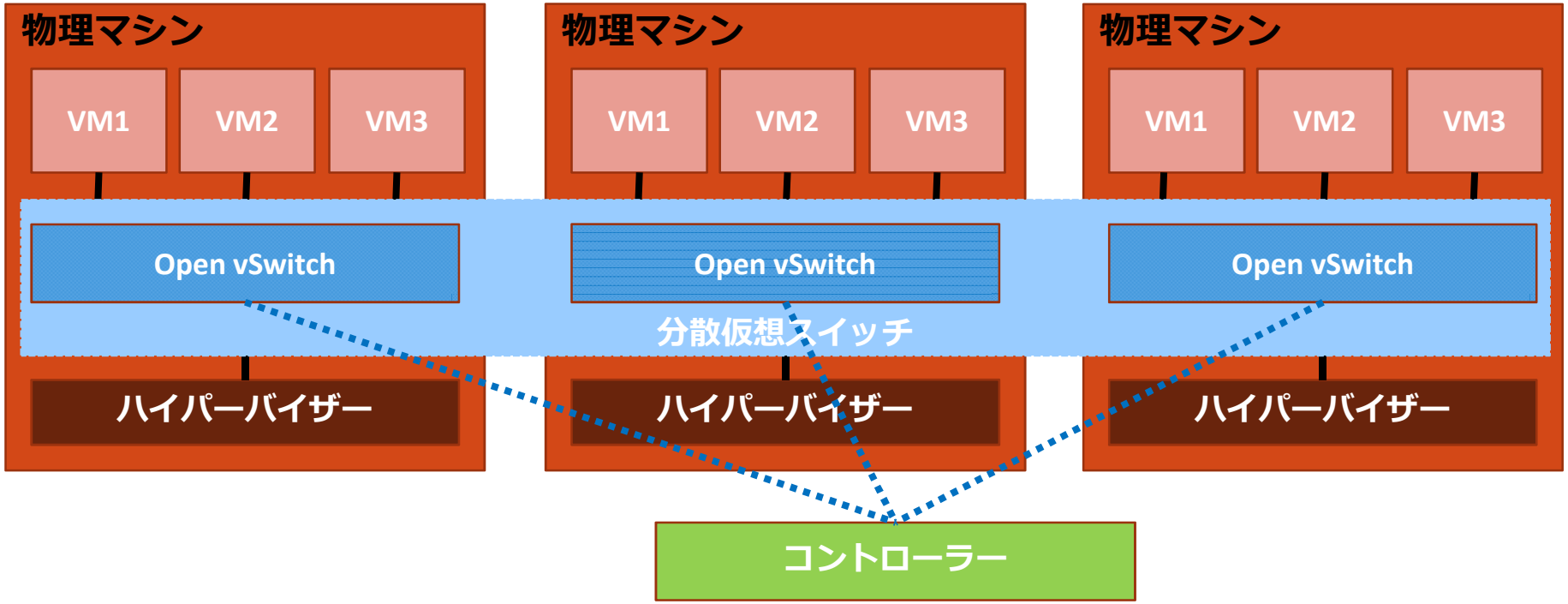
→ VM 側では NVGRE あるいは NVGRE endpoint を意識する必要がない

NVGRE 動作概要



UNIADEX

Open vSwitch

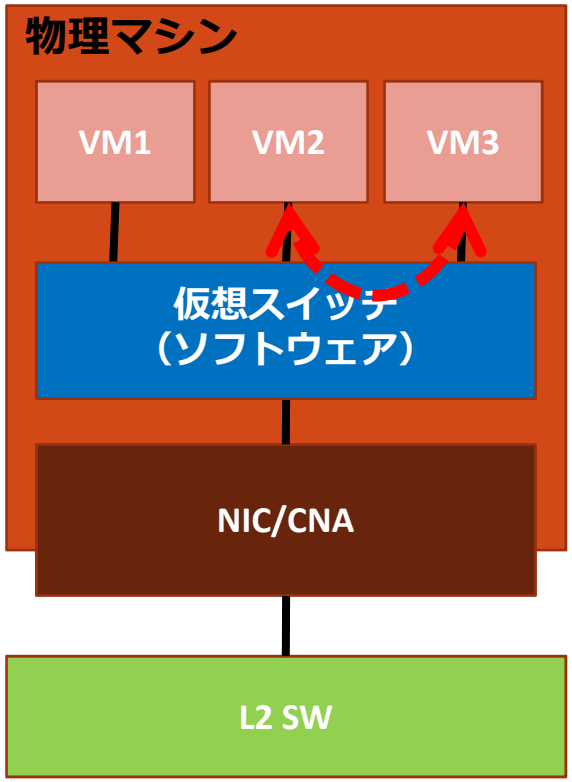


✓ソフトウェアで実装された L2/3 スイッチ
✓パケット転送だけでなく各種機能を実装（ACL, QoS, LACP, NetFlow/Sflow/Mirroring etc）
✓複数ノードに搭載されたスイッチを**１台の仮想スイッチ**として利用可能
✓コントローラーによる**運用管理の集中化**
✓オープンソースで Linux ベースのハイパーバイザーに対応

- Xen, KVM, Virtual Box
- Xen Server 6.0 からは標準のネットワークスタックに統合

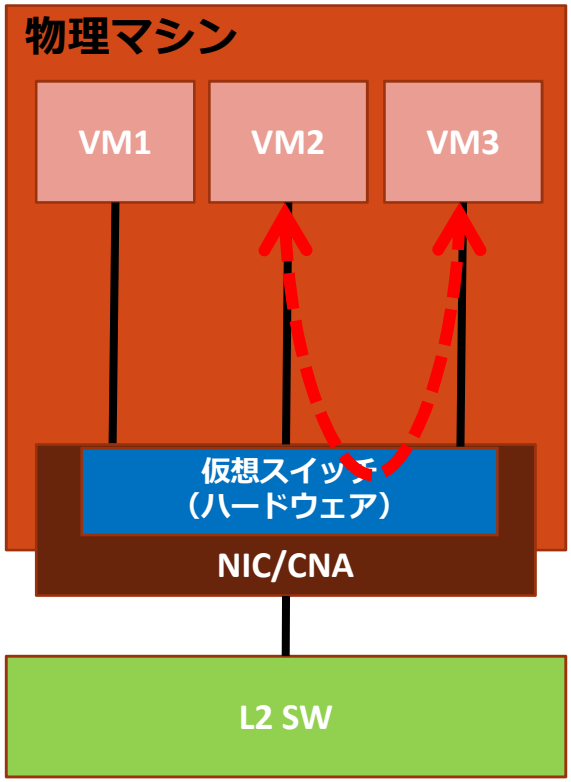
UNIADEX

仮想スイッチ：ソフトウェア or ハードウェア あるいはタグ付け



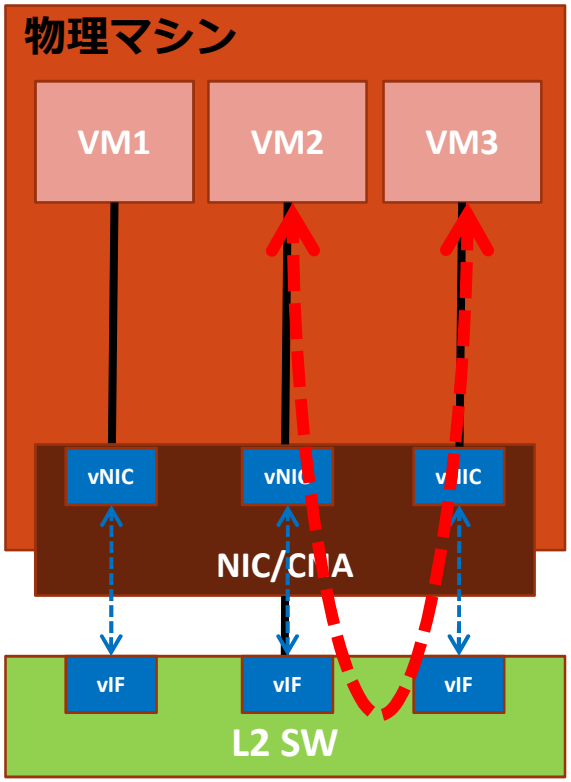
✓CPU へのインパクトあり
✓ACL等で性能インパクト？
✓低遅延
✓高い柔軟性

Open vSwitch, Nexus1000V



✓CPU は使用しない
✓ACL等は機能に制限（HW実装）
✓低遅延
✓柔軟性に乏しい
✓対応 NIC/CNA の追加必要

802.1Qbg, Edge Relay, EVB,
VEB



✓CPU は使用しない
✓対応 NIC/CNA、スイッチが必要
✓遅延が大きい
✓それなりに柔軟性あり

802.1Qbg, Edge Relay, EVB
VEPA / Cisco VNTag

※ CNA: Converged Network Adaptor, EVB: Edge Virtual Bridge, VEB: Virtual Ethernet Bridge, VEPA: Virtual Ethernet Port Aggregator

UNIADEX

ユニアデックス株式会社