HUAWEI

LEADING NEW ICT

JANOG41 発表
中国OTT事業者のDCネットワーク事例検討 (その1)

発表者 自己紹介

南保 邦亮
- ソリューションマーケティング担当(何でも屋)
- 直近の経歴
  - シスコ 約11年、プリSE→営業
  - パンドウィット 約1年、DCファシリティの営業
  - 2013/8よりファーウェイ、営業→ソリューション→マーケ
- 猫ラブ、プロレス観戦

高木 圭一
- IPネットワーク製品担当SE
- 経歴
  - ネットワークのエンジニアとして32年
  - 国産メーカー、北米のメーカー、フリーランスを経て
  - 2015/2よりファーウェイ
- いろいろな国の人と働いてみたい

Disclaimer etc.

- 時間の関係上、かなりかいつまんだ内容となっております。
- （その2）の予定は今のところありません。
  アンケートの結果次第で考えます。
- 特定事業者の情報ではなく、大手OTT事業者の最大公約数的な
  特徴やトレンド情報で構成しております。
- 一部、弊社独自ソリューションによりインプリされている部分が
  ありますが、基本、ベンダーに依存しないジェネラルな内容で
  構成をしております。

アジェンダ

中国の大手OTTにおける

- DCネットワーク アーキテクチャ概説 - 南保
- ネットワークデザイン トピックス - 高木

- 質疑応答＆議論

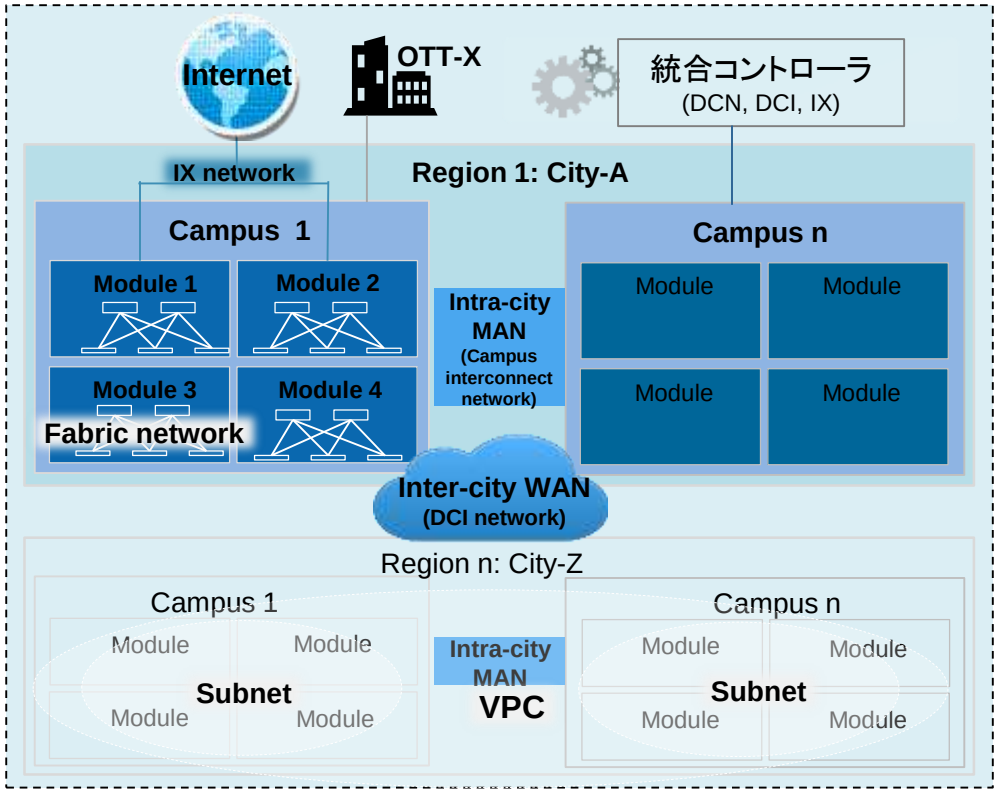
検討すべき観点

とにかく規模が大きいので、
- どういうアーキテクチャなら際限なくスケールアウトしていけるのか?
- いかにして運用の効率化・自動化を進めるのか?

成長のスピードがとにかく速いので、
- SDNなどを利用したタイムトゥーマーケット短縮の現実解は?
- 常にコスト性能比を最大にしてライバルに差をつける方法は?

オンライン決済など社会インフラとしての活用が進んでいるので、
- ベストエフォートとミッションクリティカルをどのように統合すればよいのか?
- コストとBCPのバランスをどう考えるのか?

DCネットワーク ビルディングブロック 概観



巨大なネットワークへの複雑なデプロイメント

- モジュールあたり**数万**の10Gサーバー
- キャンパスあたり**十万台規模**のサーバー

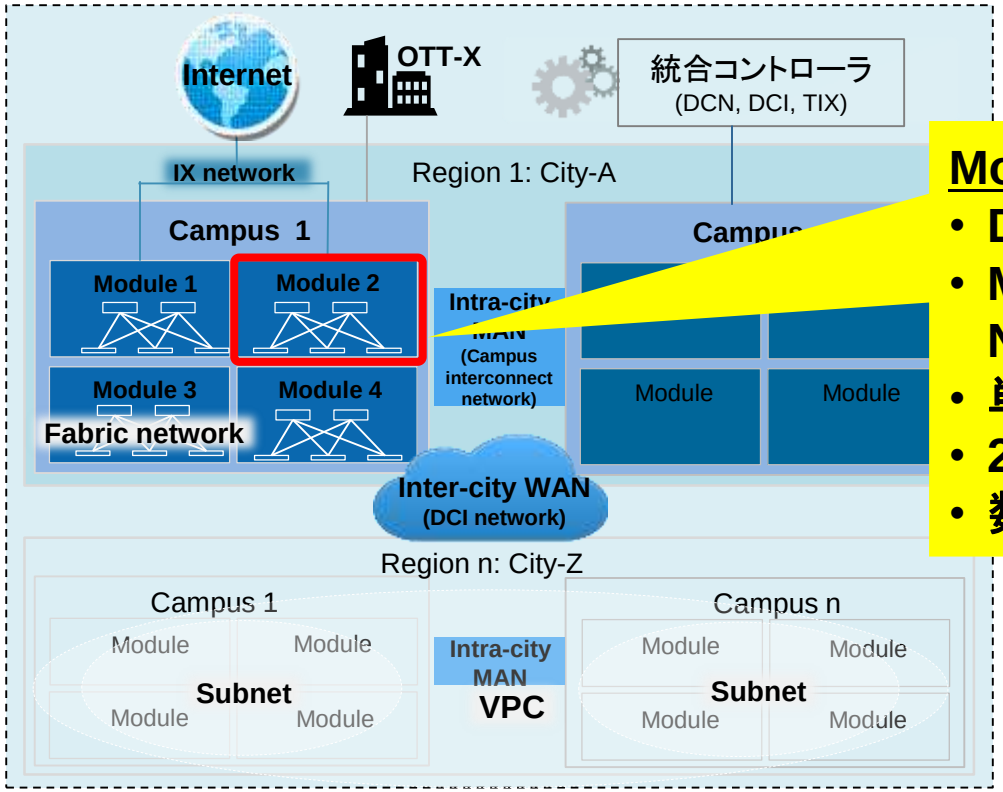
ハイパフォーマンス要求

- 高トラフィックのSNS、ゲーム、eコマース、ファイナンシャルサービスなどのアプリケーションが**高いネットワークパフォーマンス**を要求

DCIネットワークへの高い要求

- Intra-city MANによるキャンパス間接続
- Inter-city **DCI** ネットワークによるRegion間接続

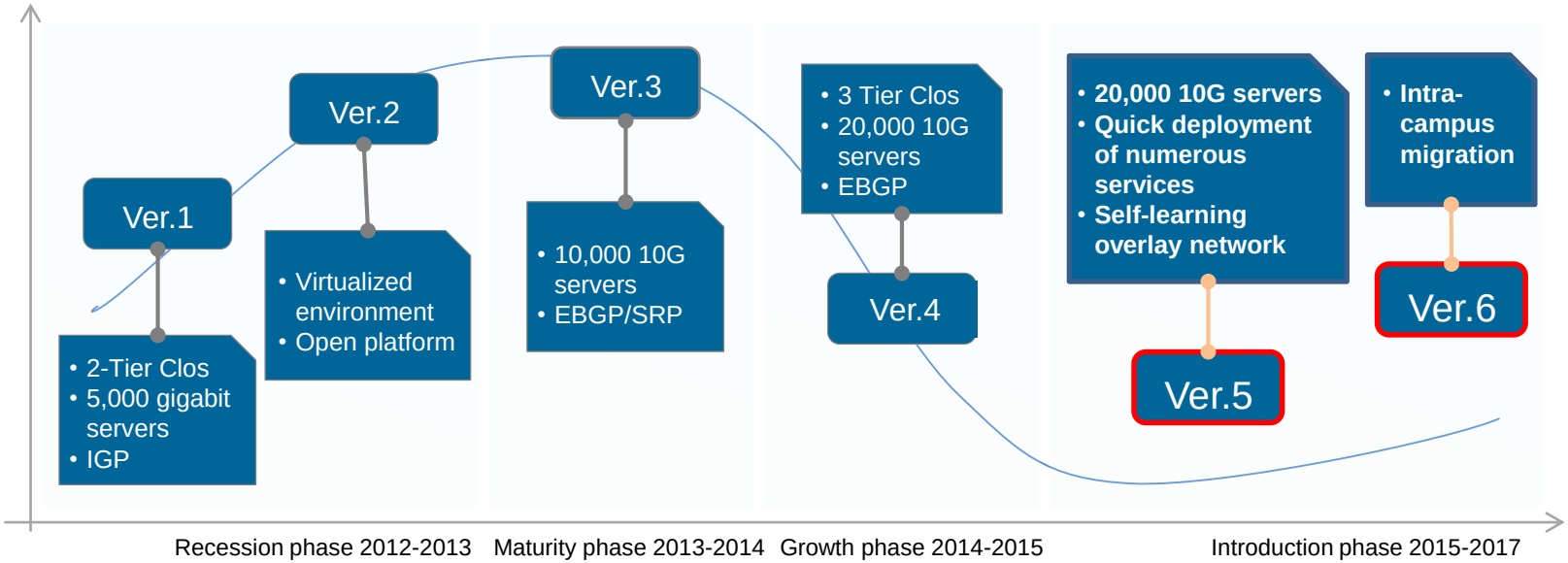
DCネットワーク ビルディングブロック - Module



Module

- DCファブリックNW
- Module単位で特定バージョンのNWアーキテクチャをインプリ
- 単一ベンダー、単一SDN環境
- 2モジュールによる冗長構成
- 数万の10G接続サーバーを収容

某OTT DCネットワーク アーキテクチャの推移



標準化以前 → “標準化” Ver.1 → “サーバー仮想化” Ver.2 → “プロトコル最適化” Ver.3 → “パフォーマンスアップグレード” Ver. 4 → “NW仮想化” Ver. 5 → “NWオートメーション” Ver.6

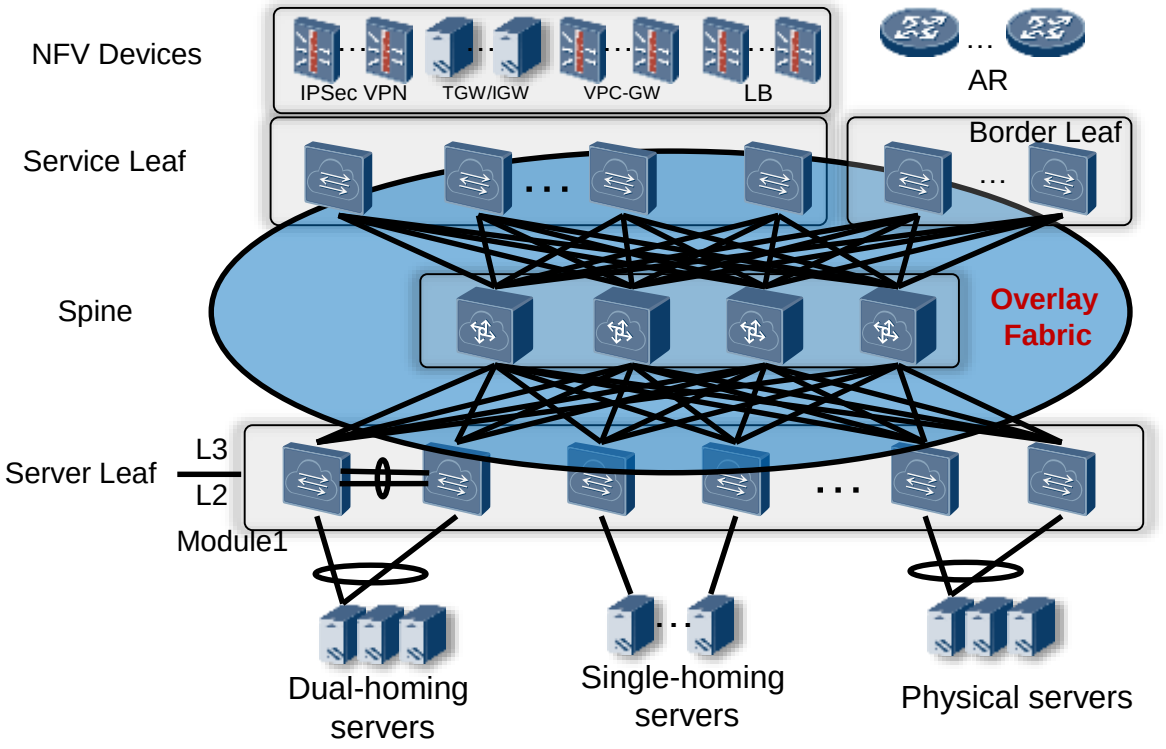
現在構築中のバージョン

パイロットフェーズ

標準化：
デバイス選定
NW設計アーキテクチャ
ソフトウェア機能

サービスデプロイのための迅速なネットワークデプロイ：大規模化と標準化へ

Ver. 5 アーキテクチャ

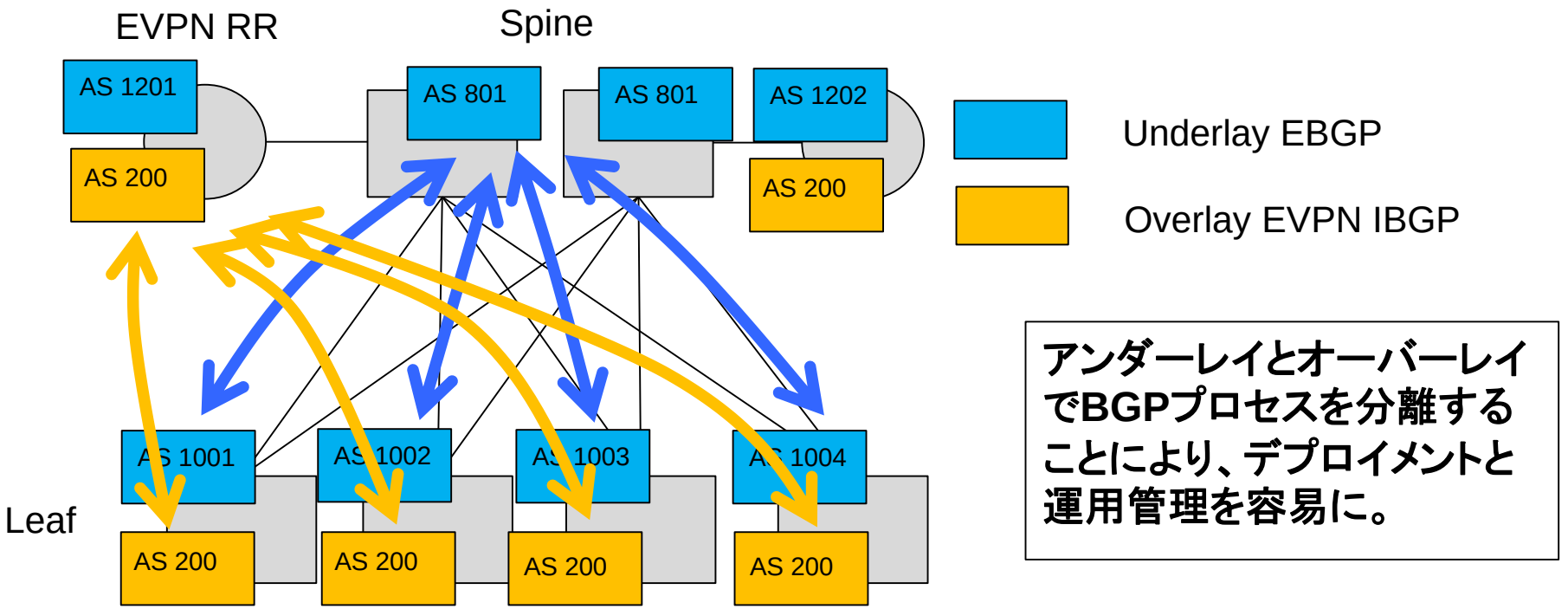


■3-tier CLOS アーキテクチャ
- 数万 10GEサーバーのスケーラビリティ.
- 標準化を進めて、モジュール増設のコストを低減.

■HWベース分散VXLAN GW：
- 大規模DCにおける集中VXLAN GWのARPテーブル拡大を緩和.
- NFVデバイスの拡張など、柔軟なスケールアウト性を確保.
- フォワーディングパスの最適化.
- Tident 2+チップセットのパフォーマンスイシューを解決.

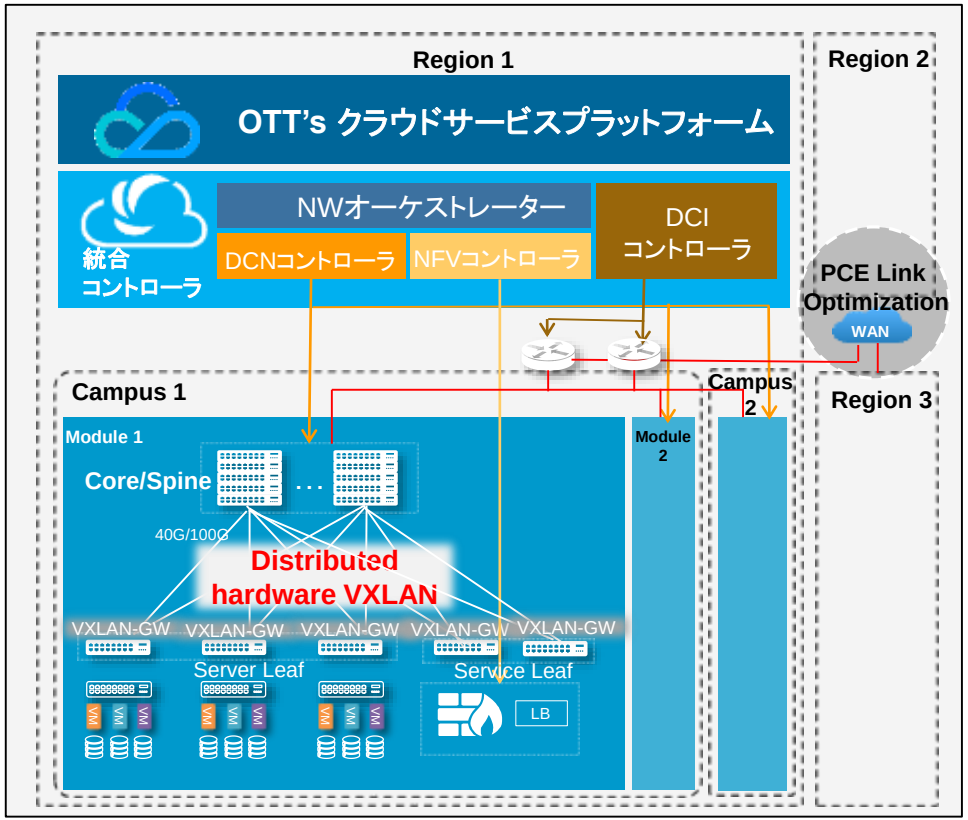
■分散VXLAN GWソリューションにおける各デバイスの役割：
- Server Leaf：サーバーアクセス
- Service Leaf：VASアクセス
- Border Leaf：DCIルーターへの接続
- Spine：リーフ間の相互接続

Ver. 5 BGPデザイン



アンダーレイとオーバーレイでBGPプロセスを分離することにより、デプロイメントと運用管理を容易に。

Ver. 5 SDNアーキテクチャ



統合SDNネットワーク

- SDNコントローラにより、DCN、DCI、NFVの**一元プランニング**を実現。ネットワークデプロイを数分で完了可能。.
- **複数DCの統合管理・運用**により、リソース利用効率を改善。
- **一元サービスチェインオーケストレーション**をNSHベースで実現。

ハイパフォーマンスハードウェアオーバーレイ

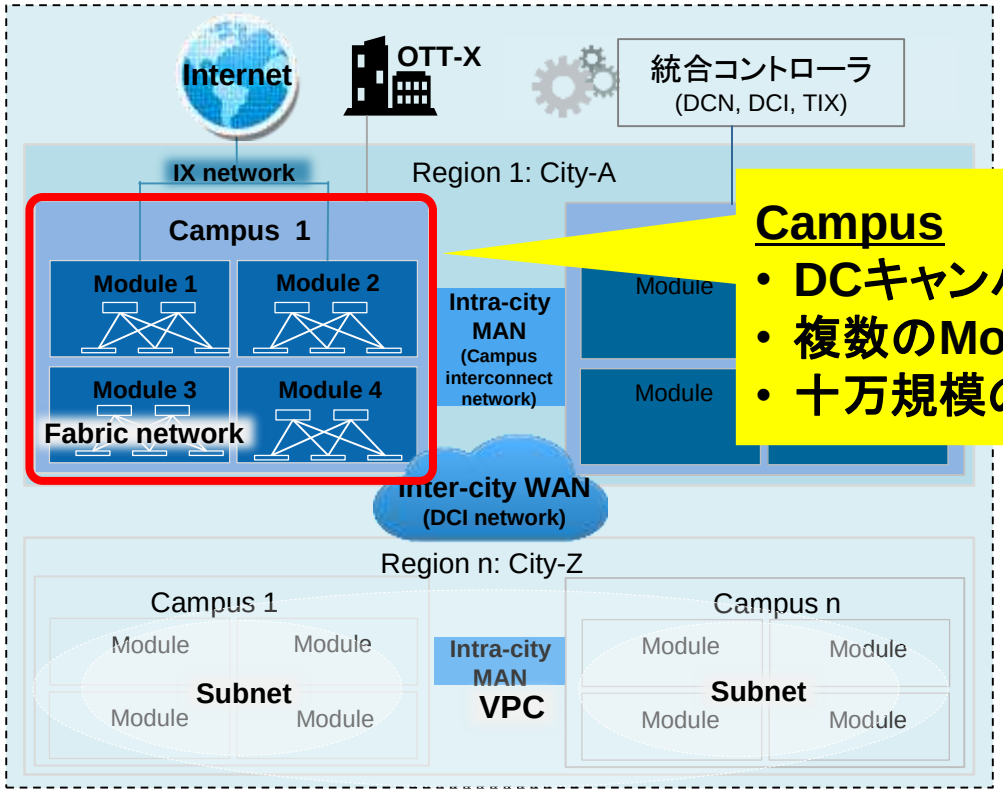
- 分散型ハードウェアEVPN/VXLANネットワークにより、高効率フォワーディングと**オンデマンドスケールアウト**を実現。
- **40G/100G** のSpine-Leaf接続による、高速フォワーディング。
- アンダーレイEBGPによる、ミリ秒レベルのコンバージェンス。

DCIリンク最適化

- DCI SDNコントローラによるリンクの可視化とリソースプール化。
- PCE(Path Computation Element)ベースのパス最適化ソリューションにより、リンク利用率を**80%、以前の3倍**に向上。

ベンダーSDNコントローラのNBI経由でOTT独自サービス管理システムと連携

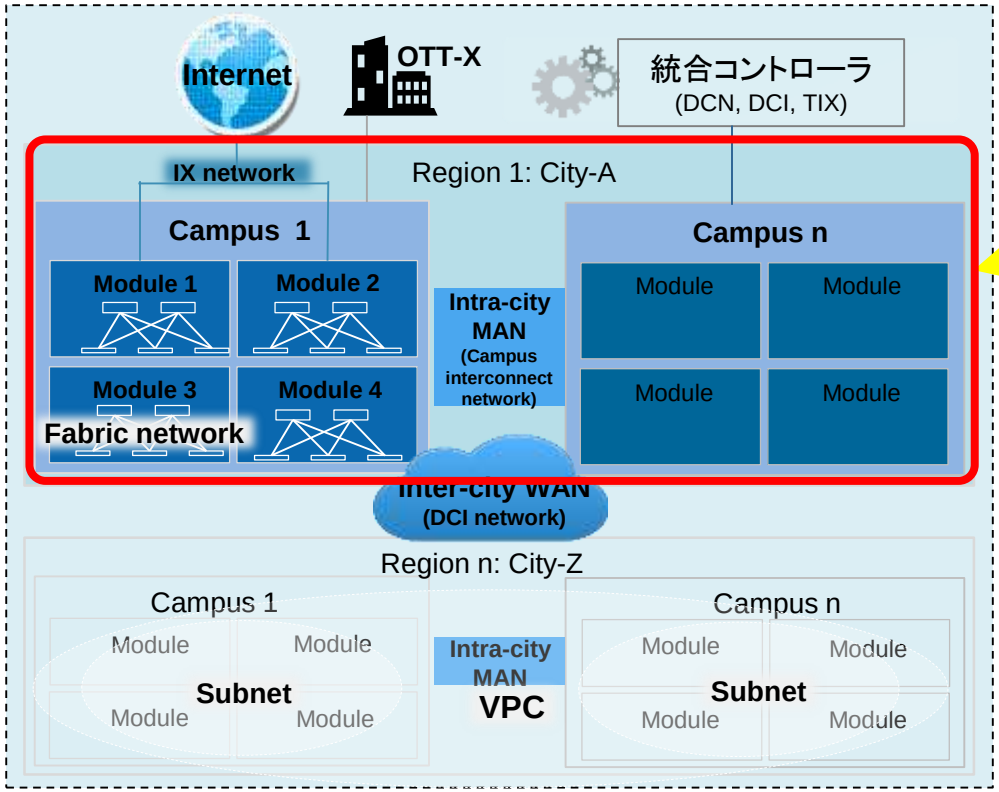
DCネットワーク ビルディングブロック - Campus



Campus

- **DCキャンパス**
- **複数のModuleで構成**
- **十万規模のサーバー**

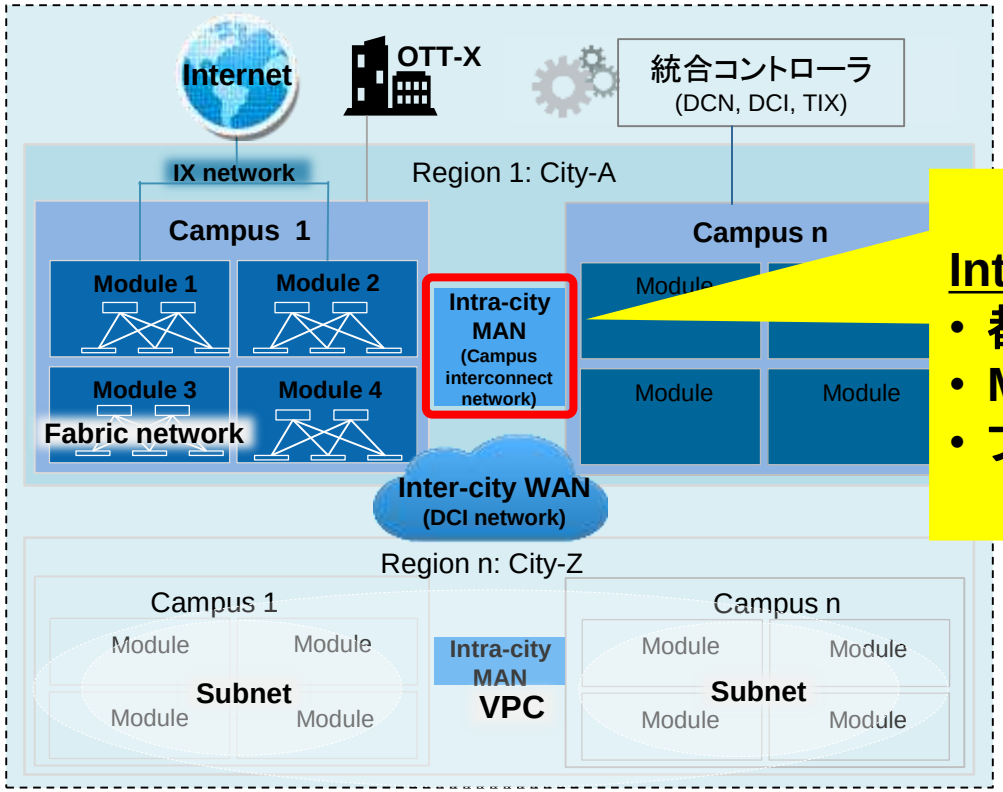
DCネットワーク ビルディングブロック - Region



Region

- **都市/地域**
- **都市(リージョン)内の複数DCをMANで接続**

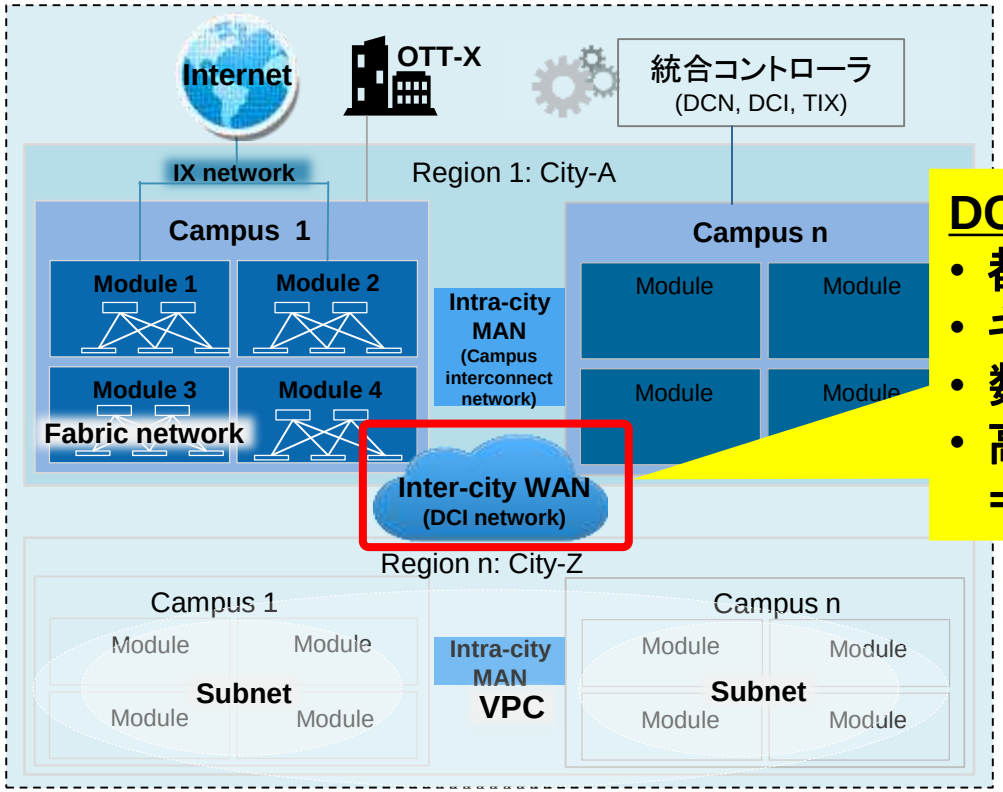
DCネットワーク ビルディングブロック - MAN



Intra-City MAN

- 都市(リージョン)内のDCI
- Metro WDM
- フルメッシュ、高速

DCネットワーク ビルディングブロック - DCI



DCIネットワーク

- 都市(リージョン)間接続
- キャリア提供のWAN（専用線）
- 数百～数千km
- 高価、遅延要因、冗長パス
  ⇒パス最適化・利用率の向上が重要

ネットワークデザインのトピックス（かなり抜粋版）

1. スケーラビリティの話題：ARPが限界
2. ミッションクリティカル保護の話題：DCNとWANそれぞれ
3. 自動化を目指す話：コントローラとテレメトリー

1. スケーラビリティ①

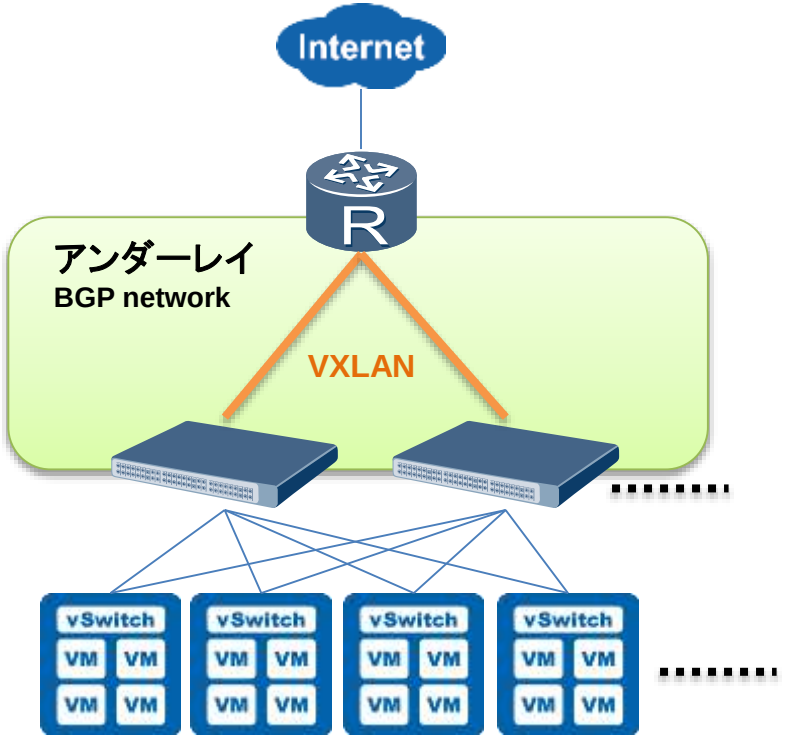
ARPテーブルが限界

アンダーレイ：BGPでスケールします！

VXLAN：4000VLANの壁を超えます！

サーバ仮想化：VMの集積度がどんどん向上します！

ARPが限界です…



（例）

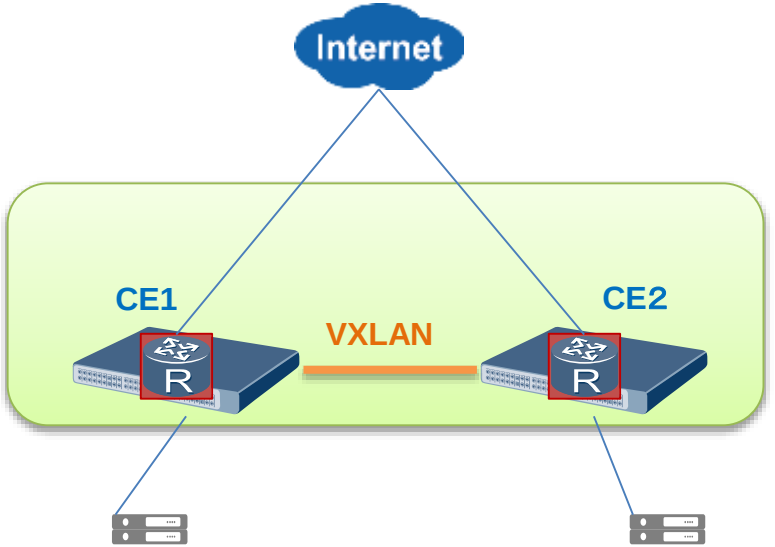
- **サーバが1万台**
- **1サーバあたり100VM**
- **1VMあたり2つの仮想NIC**
- **ARPテーブルは200万行**

1. スケーラビリティ②

対策：GWを分散させる

- ✓ ToRスイッチをオーバーレイのL3-GWにする

しかし、
単にGWを分散するだけだと、
サブネットが細分化されてしまう。



Host 1とHost 2は

- ➢ 異なるネットワークアドレス
- ➢ 異なるデフォルトゲートウェイ

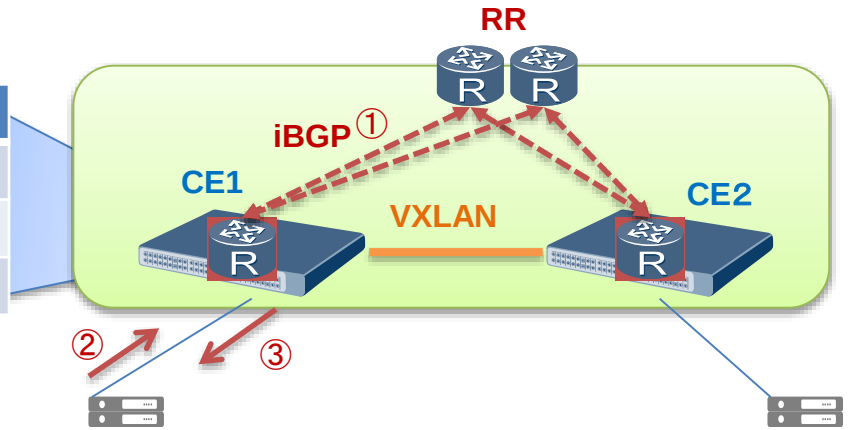
1. スケーラビリティ③

対策：GWを分散させる

- ✓ ToRスイッチをオーバーレイのL3-GWにする
- ✓ BGPを使ってホスト経路を交換する
- ✓ Virtual-subnet（RFC7814）を使う

① CE1とCE2はBGPでホスト経路を交換
② Host1からHost2のARP要求を送信
③ CE1がARPに代理応答して自身のMACアドレスを通知

Prefix	Next-hop	Protocol
192.0.2.0 /24	192.0.2.1	Direct
192.0.2.2 /32	192.0.2.2	Direct
192.0.2.3 /32	CE2	iBGP



Host 1とHost 2は
- ➢ 同じネットワークアドレス
- ➢ 同じデフォルトゲートウェイ

Host 1
IP: 192.0.2.2/24
GW: 192.0.2.1

Host 2
IP: 192.0.2.3/24
GW: 192.0.2.1

2. ミッションクリティカルの保護①

なんでミッションクリティカル？

TencentやAlibabaのネットワークには
日本の全銀行の取引量を上回る決済トラフィックが流れている

最近はテレビなどでも報道されていますが、
中国ではモバイル決済なくして生きていけないレベルになってます。

屋台でも

お賽銭も

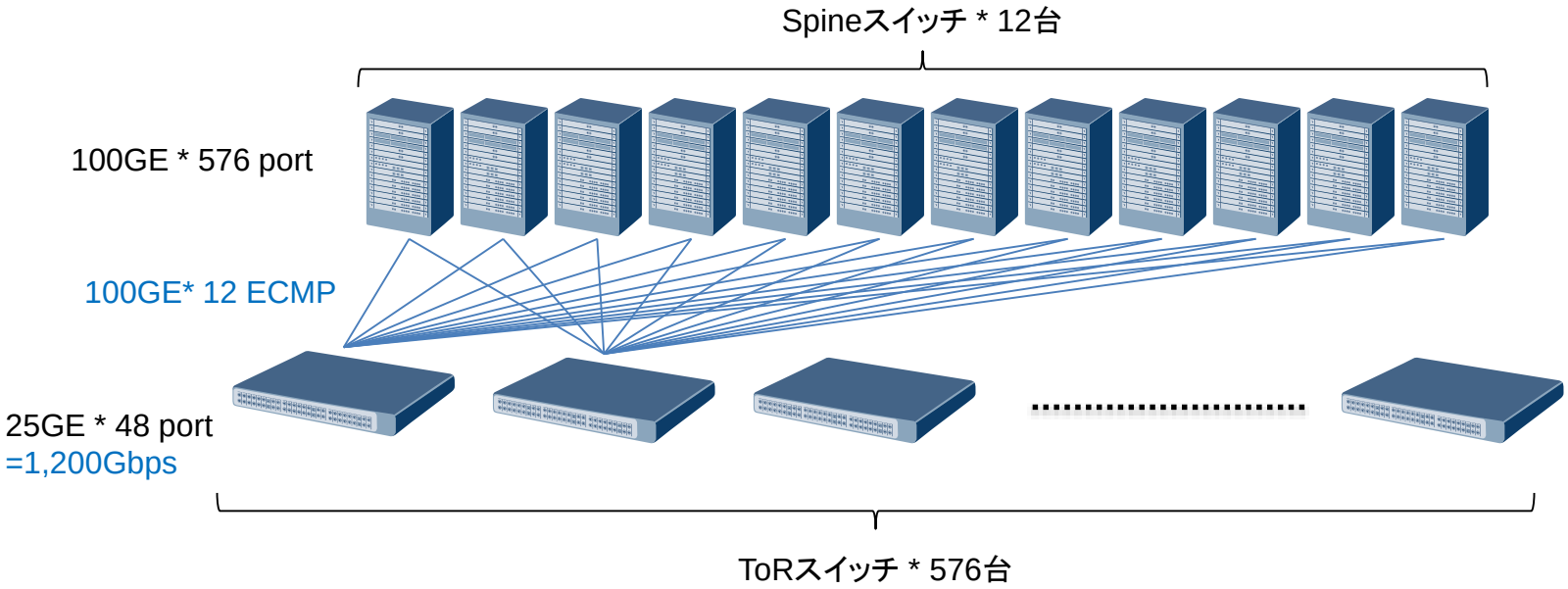
ご祝儀も！

2. ミッションクリティカルの保護②

DC内はノンブロッキングと多マルチパス

- ✓ ボトルネックがなければQoSやTEの必要性は低い
- ✓ 多マルチパスだから一つ死んでもインパクトが小さい

（構成例）
25GE * 27,648 port
ノンブロッキングなファブリック



2. ミッションクリティカルの保護③

で、こんなの作りました。

※この製品は国内では販売しておりません

2. ミッションクリティカルの保護④

WANには厳しい中国のネットワーク事情

- ✓ 国土面積は日本の約26倍
- ✓ さすがに自力でのファイバー敷設は無理
- ✓ 専用線の費用が運用コストを圧迫

リンク帯域の利用率を最大化したい

しかし、

重要なトラフィックは保護しなければいけない

そこで、

- SDN controller
- Segment routing

など先進的な技術に投資して

トラフィックを最適化したい

というモチベーションが高い

2. ミッションクリティカルの保護⑤

コントローラによるエンド-エンドの最適化を実施

- ✓ 重要トラフィックの保護
- ✓ リンク利用率の均等化

【実現方法】

①情報を収集する

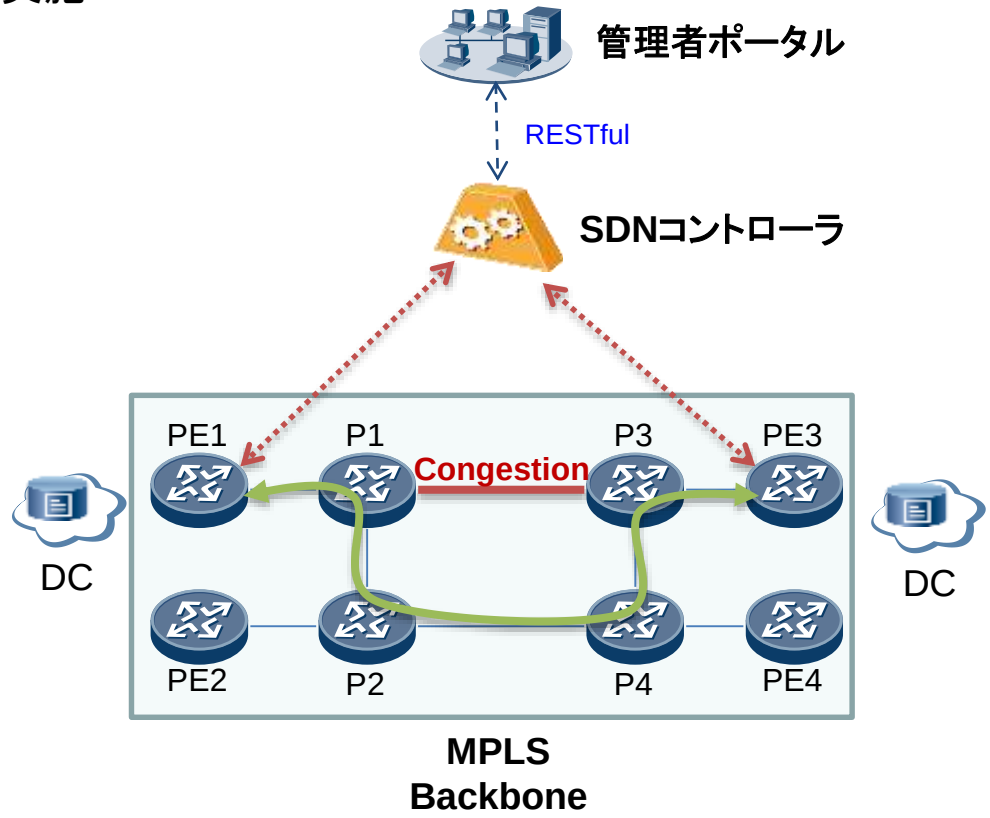
- トラフィック統計：Netflow
- トポロジ情報：IGP、BGP-LS
- トラフィック品質：TWAMP

②最適パスを計算する

- 予め定義したルールに基づく自動計算
- 管理者による手動設定

③トンネルを設定しトラフィックを制御する

- デバイスの基本設定：Netconf
- ラベル情報の配布：PCEP
- トラフィックの転送：SR-TE



3. 自動運転を目指して①

まずは、コントローラから一元制御できる仕組みを作っている。

が、課題もあります。

マルチベンダー化

- コントローラと制御対象デバイスを異なるベンダーとした場合の相互接続性が低い

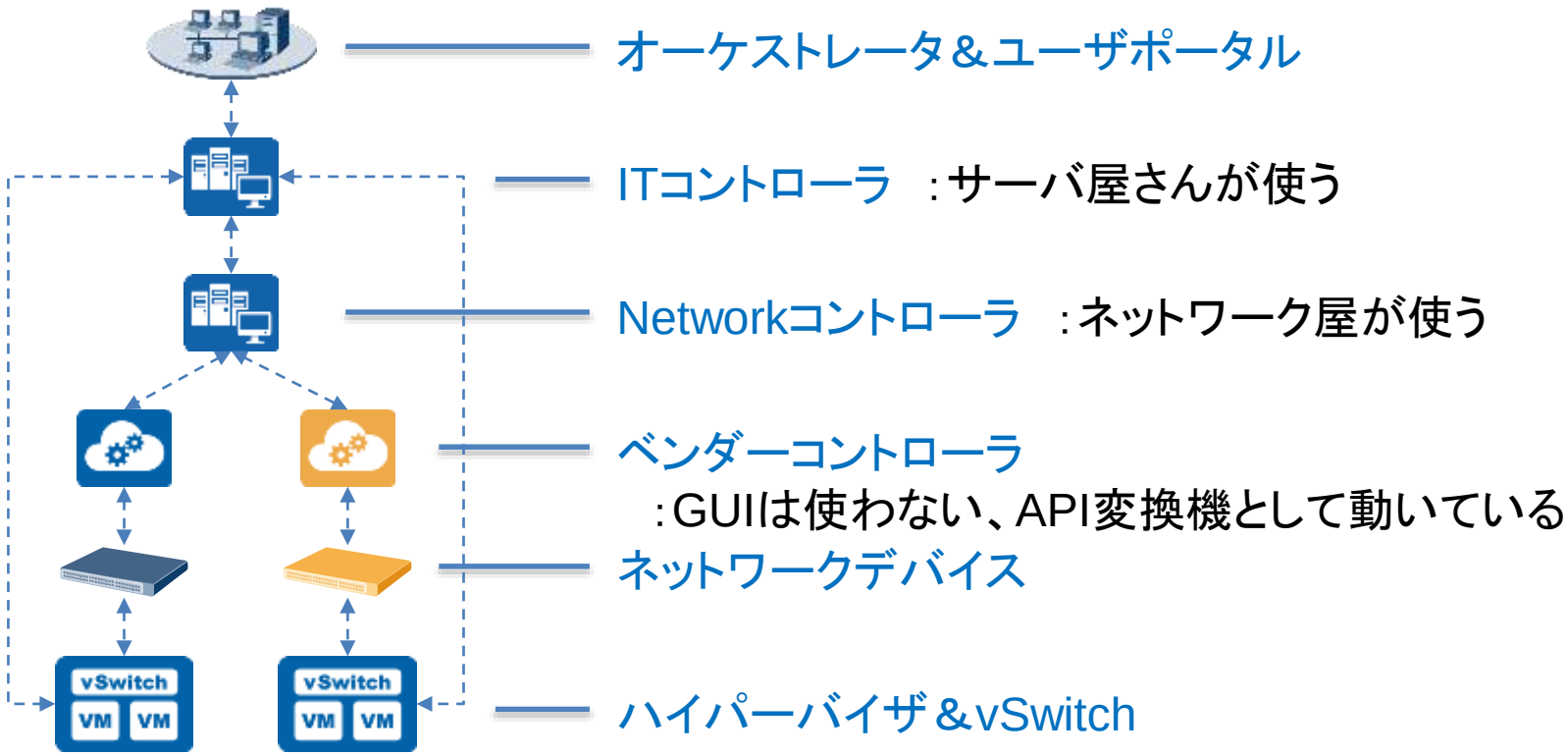
スケーラビリティ

- 1つのコントローラから制御できるデバイス数や論理リソースに限界がある

職務分掌（サーバ屋vsネットワーク屋）

- ハイパーバイザ（vSwitch）はどっちの担当？
- 属人化した作業プロセスがある場合、まずは標準化が必要（これが結構大変）

現状の解決策：コントローラの階層構造



オーケストレータ＆ユーザポータル

ITコントローラ　：サーバ屋さんが使う

Networkコントローラ　：ネットワーク屋が使う

ベンダーコントローラ
　：GUIは使わない、API変換機として動いている

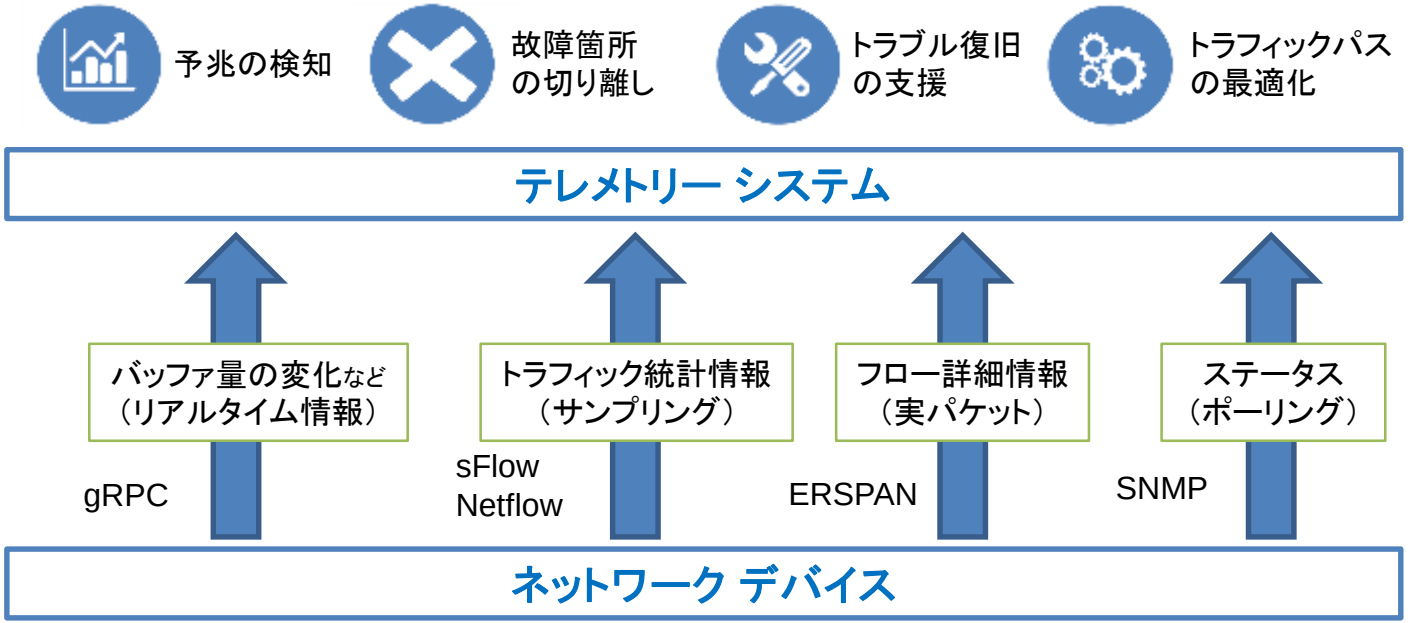
ネットワークデバイス

ハイパーバイザ＆vSwitch

3. 自動運転を目指して③

テレメトリー　:情報収集と分析のプラットフォーム

OSSベースで内製化　＞各社で開発中



※サポートするテレメトリー情報の種類と粒度がネットワーク機器選定の重要な判断基準になっている

まとめ

- その時の最新のハードウェアとNW（ファブリック）技術を標準アーキテクチャに取り込んで適用
- ネットワークをスケールアウトさせる工夫
  - ✓ GWをToRに分散
- ミッションクリティカルの保護

  ⇒コストとの兼ね合いで適材適所な方法を選択
  - ✓ DC内部：ノンブロッキング
  - ✓ WAN：高度なトラフィックエンジンニアリング
- 自動運転を目指した取り組み
  - ✓ コントローラからの一元管理
  - ✓ テレメトリーの開発

質疑応答 & 議論

- ネットワークをスケールアップさせる上での課題や工夫していること
- ネットワークが社会インフラ化している故の課題や工夫していること
- ネットワークの自動運転へ向けた課題や工夫していること

THANK YOU