

nuagenetworks

大規模オーバーレイSDNの 展開事例に見るユースケースと課題

日本アルカテル・ルーセント株式会社

毛利 元三

(オーバーレイ型の)
SDNはスケールするか？



**実際にスケール展開させた
事例と課題をご紹介します**

1. 展開事例・使われ方

- オーバーレイ型のネットワーク仮想化モデル
- 2つのトレンド（弊社実績より）
- 代表的な事例と導入モチベーションについて

2. 大規模展開時の課題

- マルチテナントによる面の重畳
- 見えないコントロールプレーン側の負荷
- 規模拡大に伴って苦しくなってくる事

3. 見落とされがちな問題

- オーバーレイとアンダーレイの相関性
- OPEX/CAPEX

オーバーレイネットワーク仮想化の1モデル

BGP EVPN for VXLANの場合：

データプレーン

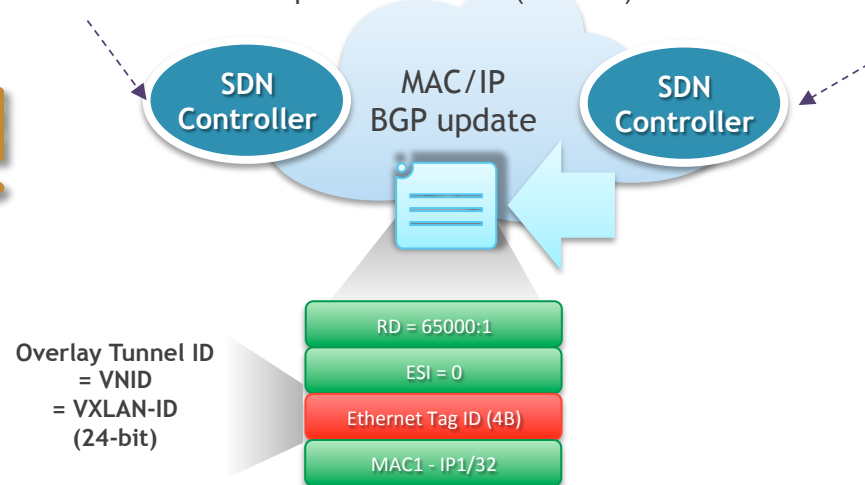
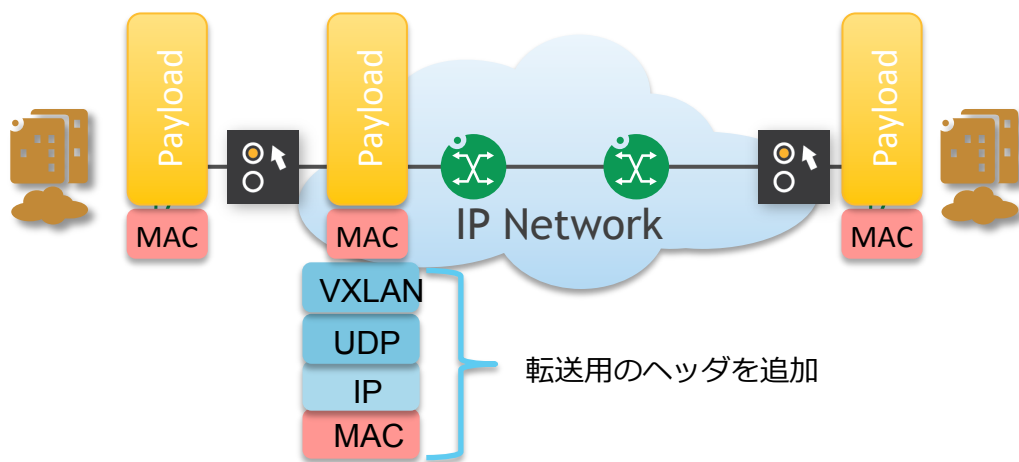
- 主にVXLAN(RFC 7348)を使用(但しUnicast)
- エッジ側でパケットカプセル化して対向装置に転送する
- 中継ネットワークでは転送用ヘッダを参照して通常のIPパケットとしてルーティングされる
- エッジ側はコントローラからプログラムされたflowのみを裁くため実装は極力シンプルな事も特徴

コントロールプレーン

- **Multi Protocol BGP EVPN**を使用
AFI = 25 (L2VPN) / SAFI = 70 (EVPN)
- VMなど仮想リソースのMAC/IP情報とVXLANのVNIを紐付けた情報を管理
- コントローラ間でBGP Peeringを行う事によりMAC/IP/VNI情報を交換
- BGP経由で学んだMAC/IP情報をエッジ装置にプログラム(Openflow)

例:

MAC: 00:00~, IP: 192.168.~ → Next-hop : VTEP 100
VTEP 100 → Next-hop : 10.1.1.100/32(HVなど)



1. クリティカルサービス化
2. 大規模化

1. クリティカルサービス化

2. 大規模化

ミッションクリティカルなサービスへの適用

事例 1/2 : 医療情報ネットワーク (アメリカ)

- 名称: UPMC (<http://www.upmc.com>)
アメリカ ペンシルバニア州の総合医療サービス機関
22棟の病院、62,000人のスタッフ・220万人の患者データを保持しており、先進的に医療のIT化に取り組む



- サービスイン: 2014年1月
- 対地: 約450拠点
- 取り扱いデータ量: 約6ペタバイト(電子カルテ, 医療画像データ(X-ray/MRI)など)
- 導入のモチベーション:

頻繁な拠点の増減や、多種の医療サービスの展開にあたり
既存VLANベースの運用に限界

=> ネットワーク設定を自動化したい

物理ストレージ装置と、サーバ仮想環境にあるVMとが
シームレスに接続できる必要がある

=> 物理・仮想環境を融合させたい (VXLAN-GW)

高可用性が求められる用途への適用

ミッションクリティカルなサービスへの適用 事例 2/2 : 金融サービス (イギリス)

- **名称:** 英国に拠点を置く大手金融グループ L SEP
対外向けオンラインサービスやグループ企業・社内システム用途に
自社DCを運用しており、リアルタイム株価情報を基にした
システムがDC内の仮想マシンで多数稼働 L SEP
- **サービスイン:** 2014年7月
- **対地:** 欧州全域
- **用途:** 株価情報の配信・オペレータ端末向けVDI L SEP
- **導入のモチベーション:** L SEP
 1. ネットワーク設定の**自動化**(サービス毎のVLANからの脱却)
 2. VDI端末に対する**セキュリティポリシーの厳格化**
 3. 中央配信システムから、**仮想ネットワーク**に所属する
多数のシステムに株価情報を**マルチキャスト配信**したい

可用性に加え、低遅延が要件化

1. クリティカルサービス化

2. 大規模化

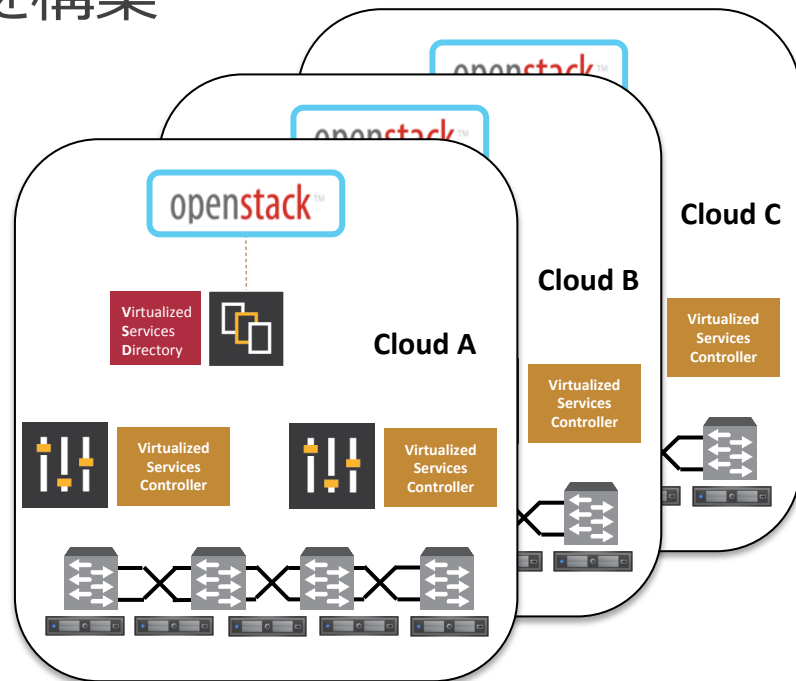
大規模展開の事例

事例 1/3 : プライベート・クラウド (ヨーロッパ地域クラウド事業者)

- 世界第3位のクラウド・ホスティング事業者
- 事業者IaaS上でユーザがセルフIaaSを構築

“Openstack as a Service”

- 商用サービスイン: 2014年10月
- サービス提供VM数: x万台
- フルオートメーション
- Openstack環境ホールセール(再販)
- Openstack + Nuage の
インテグレーション済パッケージの再販



フルオートメーション化によるIaaSの再販モデル

Find out more about **OVH.com**

The number 3 internet hosting company in the world

170,000 servers

15 datacentres

17 countries

18M web applications hosted

The key figures

- Datacentres
- Network
- Security
- Component
- Server prod
- Technical p



All You Need to Know About DevOps

Visit SDN Central's DevOps Topic Now!



- DIRECTORY
- RESOURCES
- PARTICIPATE
- SERVICES
- ARTICLES
- ABOUT US

Engage with us:

Channels: Brocade Cisco Data Center 2 Juniper Networks NEC Plexxi SDN Journey

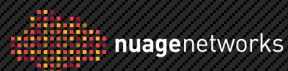
Topics: DevOps Network Virtualization NFV Open Source SDN Security White Box

700,000 customers

put their trust in us
around the world

Why choose OVH?

Innovative products and services, our own global
the OVH added value.



OVH To Deploy Nuage Networks VSP



Erin Moriarty, October 7, 2014

52 81 24



Nuage Networks — the Alcatel-Lucent subsidiary focused on software-defined networking (SDN) — is announcing another customer today: OVH, the largest European hosting provider

Related Articles

Google & Open Config Push an

Search SDN Central

Join us for DemoFriday™

November 14 at 10:00am PST
Register today.

WIND RIVER



Join SDN Central® to receive
exclusive access.

Sign Up Now!

大規模クラウドサービス展開事例

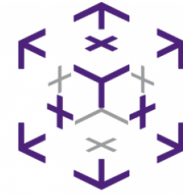
事例 2/3 : パブリック・クラウド (フランス)

- フランス最大のパブリック・クラウド事業者
“French national cloud”

- 商用サービスイン: 2014年3月末
- フランス国内DC: 10棟(建設中)
拠点間オーバレイ仮想ネットワーク

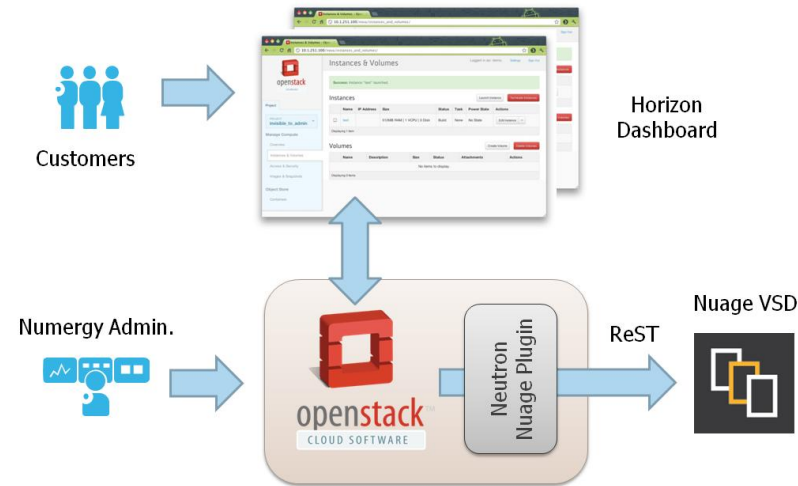
- サービス提供VM数: **100万VM 以上**
- クラウド管理プラットフォーム

Compute: Openstack
Network: Nuage VSP



NUMERGY
PRODUCEUR D'ÉNERGIE NUMÉRIQUE

<https://www.numergy.com/>



最大時100万台規模のVMスケール



PARIS, FRANCE // NOVEMBER 3-7, 2014

WWW.OPENSTACK.ORG/SUMMIT

Visit the [OpenStack Summit](#) page for the latest news, registration and hotels.

来週のOpenstack Summit Parisで講演があります！

Tuesday, November 4 • 11:15 - 11:55

Hybrid your Cloud with Numergy and Nuage

Sign up or log in to save this event to your list and see who's attending!

★ Add To My Sched

<http://sched.co/1vNm6WE>

Tweet 0

いいね!

Numergy is a public cloud provider, based on OpenStack and Nuage as SDN solution. We will share why we wanted to have SDN technology on top of OpenStack and how we chose Nuage. We will also share how we implement our architecture to answer the public cloud specificity and the key benefits of having the Nuage Network solution.

Speakers



Erik Beauvalot
COO, Numergy

Erik used to work at Colt Technology Service between 2001 and 2012 mainly on the managed service area. || Before that he spent his professional activities at AT&T and went to the CDC. || He is now the COO of Numergy.

Numergy社 COO Erik Beauvalot氏

<Abstractの意識>
なぜ私たちがOpenstackとNuageを必要としたのか、サービスの実装方法と得られたメリットについてお話しします

- Keynotes
- Networking
- Operations
- OW2 Open Cloud Day

大規模クラウドサービス展開事例

事例 3/3 : ハイブリッド・クラウドサービス (中国)

- ナショナル・クラウド及び検索事業 (IaaS)
- 政府機関・巨大モールサイトや国内検索事業者向け

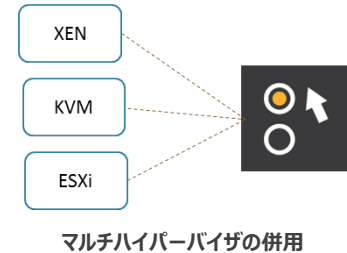
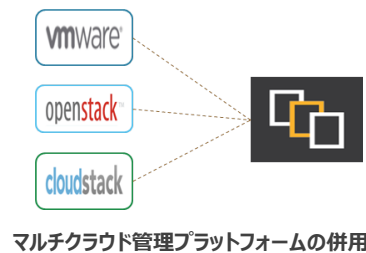
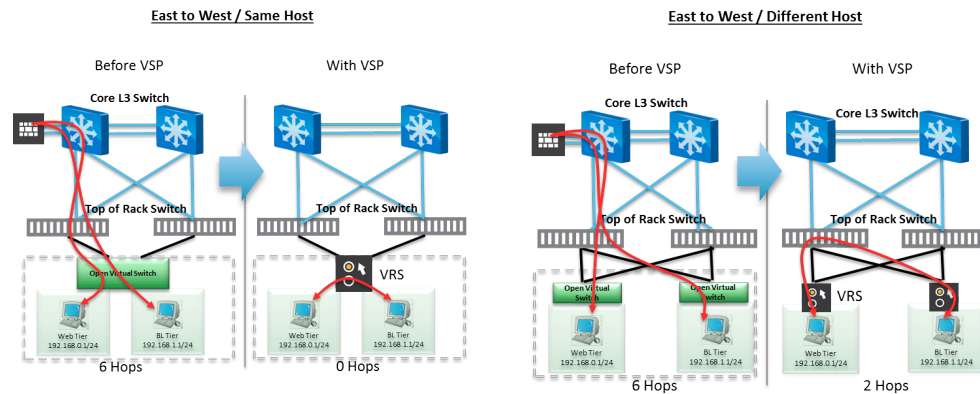
- 商用サービスイン: 2014年5月末
- 拠点数: 42棟の大規模DC
- 拠点間オーバーレイ仮想ネットワーク

- サービス提供VM数: **1,100万VM 以上**

- クラウド管理プラットフォーム

Compute: Cloudstack
Network: Nuage VSP

- 移行の主なモチベーション:
 - VRの性能(多機能使用時の性能劣化)
 - VPCネットワーク展開の柔軟性
 - マルチハイパーバイザ
 - ESXi, Xen, KVM, LXC
 - マルチクラウドプラットフォーム
 - Cloudstack, Openstack, vCloud



1. 大規模展開時の課題

- データプレーンとコントロールプレーン

2. 大規模展開を支える技術

- BGP方式の例

3. ネットワーク管理の限界

- 抽象化とプール化

1. 大規模展開時の課題

- データプレーンとコントロールプレーン

2. 大規模展開を支える技術

- BGP方式の例

3. ネットワーク管理の限界

- 抽象化とプール化

大規模展開時の課題: ネットワーク仮想化の構成要素 1/2

一般的にオーバーレイ方式の課題だと言われる3大要素

■ データプレーン

1. MTU問題

→ オーバーレイ通信用のヘッダ追加に伴う
最大フレームサイズの考慮が必要

2. トラフィックの負荷分散

→ フレームのカプセル化方式により、中継装置での
振り分け用のKeyが足りず、物理ネットワーク側の
リンク帯域に偏りが起きやすい傾向

3. フラッディング機構 (BUM対処)

→ 拠点間での不要なパケット送信を抑制する仕組み

これらは主にデータプレーン視点の話

■ コントロールプレーン

いわゆるSDNコントローラと言われる
箇所に実装される事が多い
規模拡大に伴って問題になるのはココ!

1. 仮想リソース間の通信制御

→ リソース間の行き・帰りフローの考慮

2. セキュリティ担保

→ テナント間の通信閉域性。誰がどこにどう繋がるべきか。

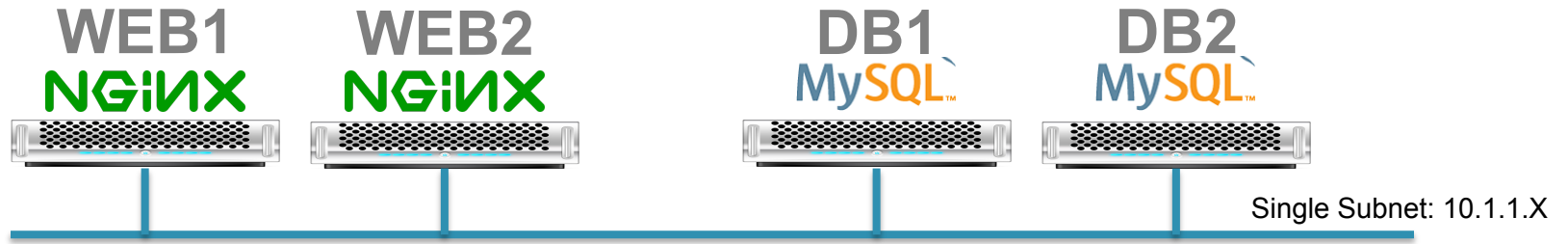
3. 別拠点追加時のスケールアウト

→ オーバーレイ方式の packets 転送は対地が増えても
IPレベルの接続性があれば成り立つ。

ただしコントローラはどう配置したら良いか。

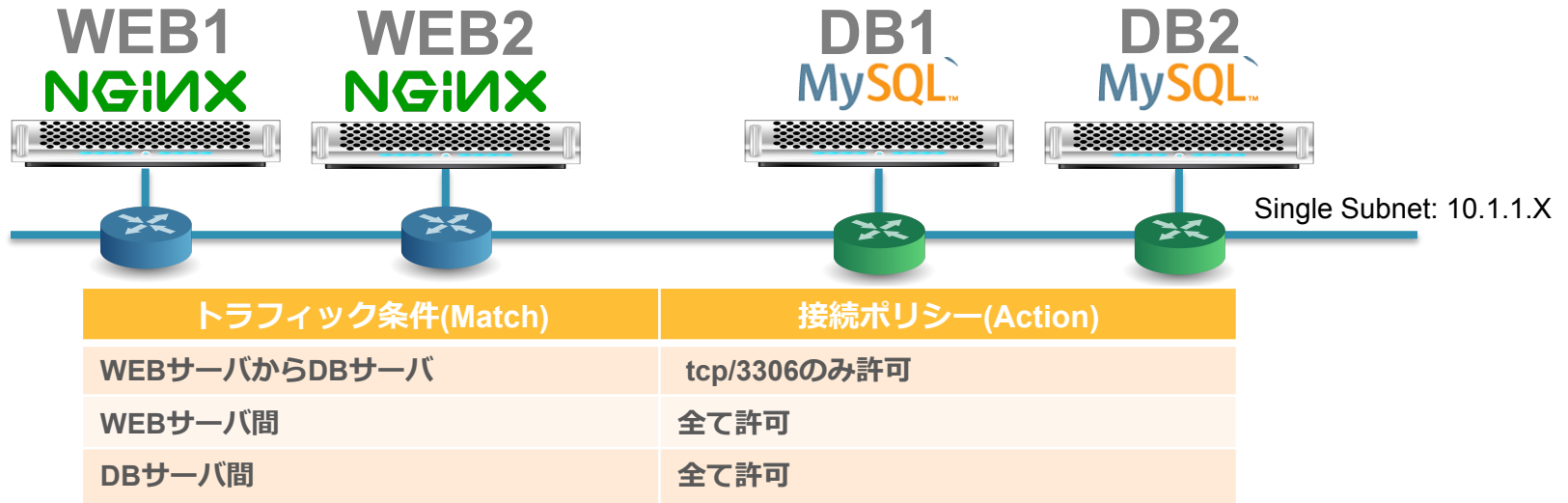
中央集中モデル vs 分散管理モデル・スケーリング・激甚災害対応

仮想リソース間 通信制御の例 (Security Groups)



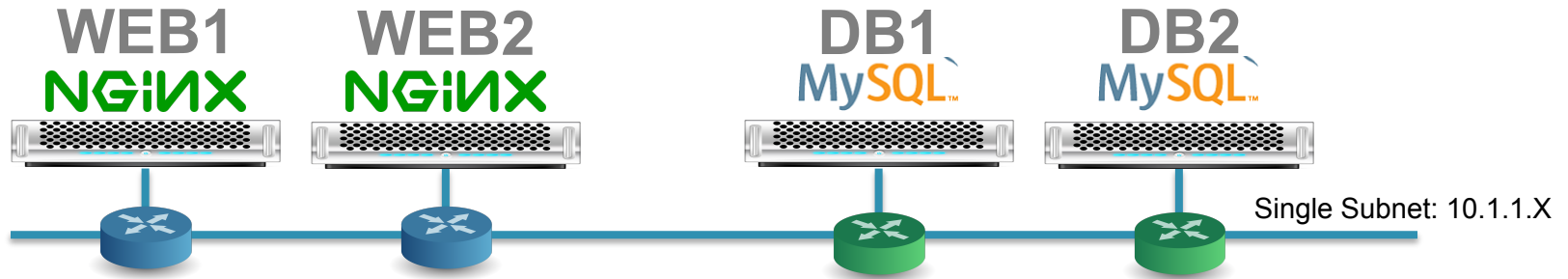
| トラフィック条件(Match) | 接続ポリシー(Action) |
|-----------------|----------------|
| WEBサーバからDBサーバ | tcp/3306のみ許可 |
| WEBサーバ間 | 全て許可 |
| DBサーバ間 | 全て許可 |

仮想リソース間 通信制御の例 (Security Groups)



- 分散ファイアウォール
VM間の接続性を制御
- 例: Openstack Neutron “Security Group”

仮想リソース間 通信制御の例 (Security Groups)



| トラフィック条件(Match) | 接続ポリシー(Action) |
|-----------------|----------------|
| WEBサーバからDBサーバ | tcp/3306のみ許可 |
| WEBサーバ間 | 全て許可 |
| DBサーバ間 | 全て許可 |

1. WEB1が起動

1. アクション無し

2. WEB2が起動

1. From WEB1 to WEB2 Permit
2. From WEB2 to WEB1 Permit

3. DB1が起動

1. From WEB1 to DB1 Permit TCP 3306
2. From WEB2 to DB1 Permit TCP 3306

4. DB2 が起動

1. From WEB1 to DB2 Permit TCP 3306
2. From WEB2 to DB2 Permit TCP 3306
3. From DB1 to DB2 Permit any
4. From DB2 to DB1 permit any

お馴染みの $O(N^2)$ 問題が発生

Order: O

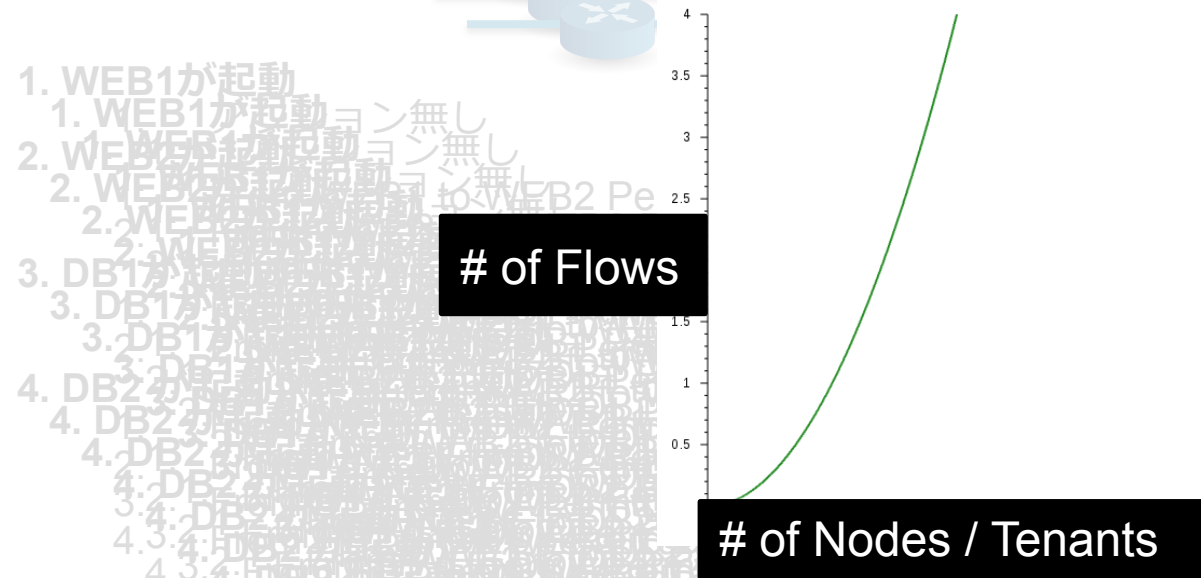
$O(N * (N-1)/2 * A)$

各FWでのState保持数:

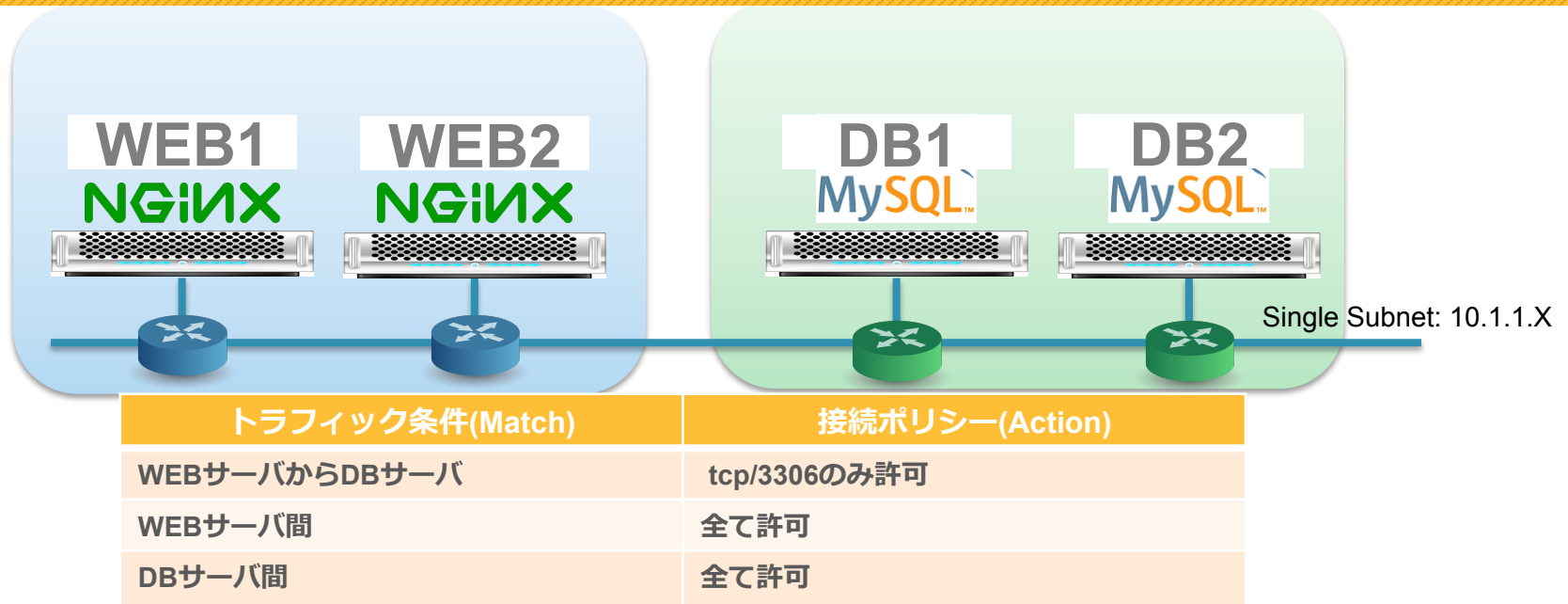
$O(N*A)$

管理フロー数の指数関数的な増加

$O(N^2)$ 問題により
テナント数が増えれば増えるほど
指数関数的にFlow数が増加



参考: Nuageの実装例



1. WEBサーバをグループ化
2. DBサーバをグループ化
 1. WEB – DB間のTCP 3306をpermit
 2. 同一グループ内はPermit any

制御対象の仮想ポートに論理的なタグ付けを行ってグループ化を実施

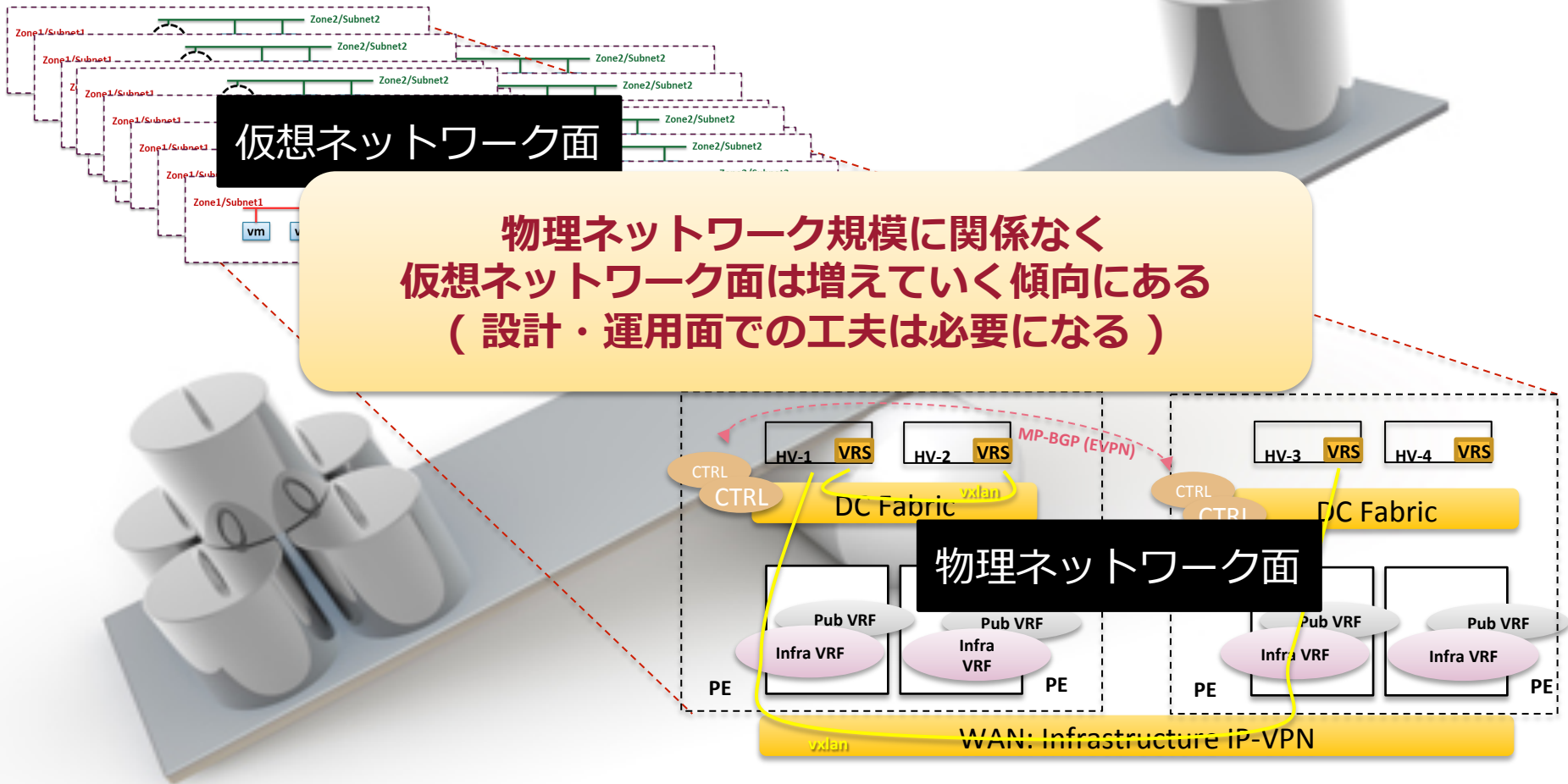
従来のACLの様に特定インターフェースやIPアドレスを直接の制御対象とせず、抽象化されたグループ単位にアクションを定義する

Openstack Group based Policy Abstractions for Neutron

<https://github.com/openstack/neutron-specs/blob/master/specs/juno/group-based-policy-abstraction.rst>

アンダーレイ vs オーバーレイ

マルチテナント環境
仮想ネットワークのリソース消費量
(FDB flow, security-policy)



1. 大規模展開時の課題

- データプレーンとコントロールプレーン

2. 大規模展開を支える技術

- BGP EVPN for VXLAN

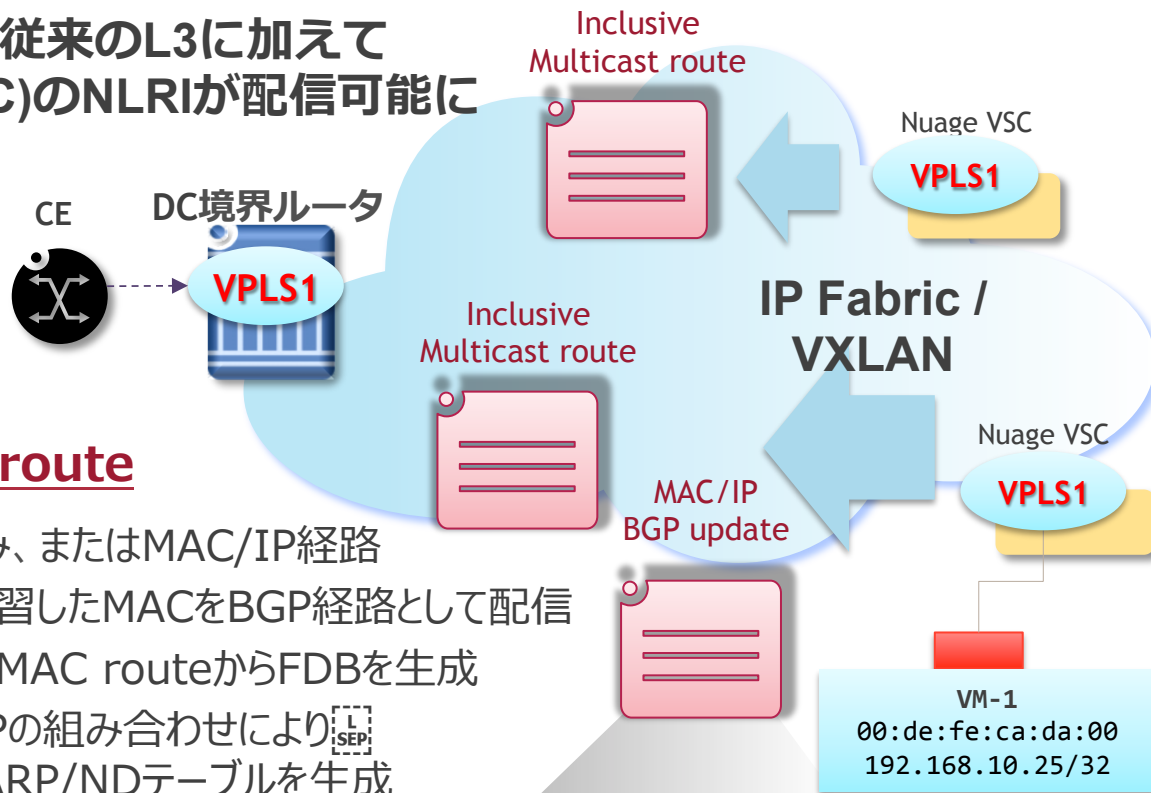
3. ネットワーク管理の限界

- 抽象化とプール化

大規模展開を支える技術

EVPN for VXLAN (draft-sd-l2vpn-evpn-overlay)

- BGPで従来のL3に加えてL2(MAC)のNLRIが配信可能に



MAC/IP route

- MACのみ、またはMAC/IP経路
- Local学習したMACをBGP経路として配信
- 交換したMAC routeからFDBを生成
- MACとIPの組み合わせによりProxy ARP/NDテーブルを生成

データプレーン側でUnknown Floodingさせる必要がない

当該MAC/IPに到達するためのVNI情報を配信

RD = 65000:1

ESI = 0

Ether-Tag ID = VNI-1

MAC 00:de:fe:ca:da:00
IP 192.168.10.25/32

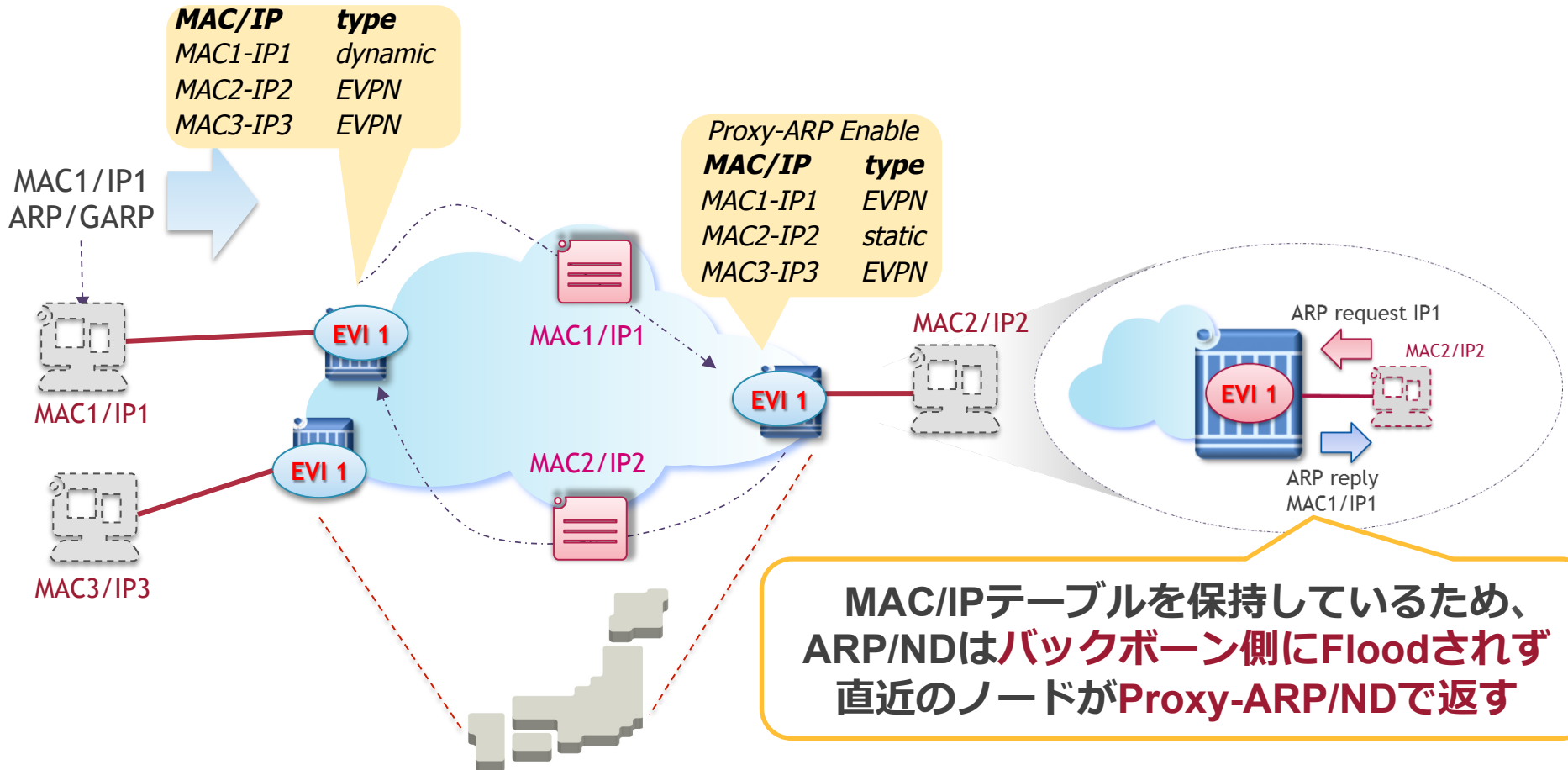
Inclusive multicast route

- VXLAN VTEPの自動探索
- BGPによるメンバーシップ情報の自動配信
- コア側のMulticast statesの保持不要

大規模展開を支える技術

EVPN for VXLAN (draft-sd-l2vpn-evpn-overlay)

- 物理的に離れた拠点間のLayer2仮想ネットワーク
- 仮想ホスト間のARP抑制の例



なぜBGP(MP-BGP-evpn)を使うのか

1. 実績

- ・ I/Oロジックや既存の最適化技術が転用可能

2. 自動探索

- ・ 通信先になるVTEPやReplicationしてよいGroup
- ・ 大規模展開時の運用ではVTEPの自動探索が必須

3. スケールアウト可能

- ・ BGPピアリングによる分散管理が可能

4. 既存運用テクニックの再活用

- ・ コミュニティ属性による経路フィルタリング・カラーリング
- ・ 更に大規模構成ではルートリフレクタを使ったクラスタ管理

5. 軽くて速い！

マルチベンダ環境下でのOverlay EVPNインターオペラビリティもIETFで議論中(Cisco/Alcatel/Juniper)

Interconnect Solution for EVPN Overlay networks / draft-rabadan-bess-dci-evpn-overlay
<http://tools.ietf.org/html/draft-rabadan-bess-dci-evpn-overlay-00>

1. 大規模展開時の課題

- データプレーンとコントロールプレーン

2. 大規模展開を支える技術

- BGP方式の例

3. 設計管理の限界

- 抽象化と雛形化

■ ネットワーク利用の短命化 / ephemeral network

大規模かつ複雑な仮想ネットワーク構成だが、
わずかな時間しか必要とされない利用形態の出現

- > 設定 = 数秒で開通(自動)
- > 設計 = 数時間(手動)

■ 設計の繰り返し要素への対処

クラウドサービスの通信パターンは複雑だが、
共通要件はある程度パターン化できそう

そこで抽象化と雛形(テンプレート)化

ネットワークの抽象化・雛形化

仮想ネットワーク テンプレート

Gold Customer Template

Subnet Policy

- WEB (DHCP enable)
- LBaaS (DHCP enable)
- FWaaS (DHCP disable)

Security Policy

Permit LB to WEB
Permit FW to LB
....

Service Chaining

Rule-01:
From Internet
To any
-> Redirect to FW-External
Rule-02:
....

QoS Policy

Rule-01:
DSCP 0x00 -> PIR 500M / CIR 0
.....

Floating IP Policy

Up to "16" IPs

多項目かつ汎用的なネットワーク設定を雛形として管理



実体の生成
&&
ポリシーの継承

テンプレートから生成された 各テナント単位のネットワークインスタンス

Virtual Tenant A / Domain-01

Virtual Tenant B / Domain-01

Virtual Tenant C / Domain-01

Subnet Policy

- WEB
- LBaaS
- FWaaS

Security Policy

Permit LB to WEB
Permit FW to LB
....

Service Chaining

Rule-01:
From Internet
To any
-> Redirect to FW-External
Rule-02:
....

QoS Policy

Rule-01:
DSCP 0x00 -> PIR 500M / CIR 0
.....

Floating IP Policy

Up to "16" IPs

Subnet Policy

- WEB (DHCP enable)
- LBaaS (DHCP enable)
- FWaaS (DHCP disable)

Security Policy

Permit LB to WEB
Permit FW to LB
....

Service Chaining

Rule-01:
From Internet
To any
-> Redirect to LB-External
Rule-02:
....

QoS Policy

Rule-01:
DSCP 0x00 -> PIR 800M / CIR 0
.....

Floating IP Policy

Up to "16" IPs

ネットワーク仕様の差分は
インスタンス側で吸収

参考: Nuageの実装例

The screenshot displays the Nuage Domain Designer interface for a domain named '00_Gold_Customers'. The interface is divided into several sections:

- Left Panel:** Lists domain templates and domains. The '00_Gold_Customers' template is highlighted with a red box and labeled '設計済みのテンプレート' (Design-ready template). Below it, 'MY DOMAINS' and 'DOMAINS SHARED...' are visible.
- Top Panel:** Navigation tabs include Dashboard, Domains, L2 Domains, Users, Groups, Gateways, and Events. The 'Domains' tab is active.
- Center Panel:** Shows the domain design for '00_Gold_Customers'. It includes components like DBaaS, FWaaS, Internet, LBaaS, and nginx, each with associated network and gateway configurations.
- Right Panel:** Shows the configuration for a 'Gold_Class-QoS' object, including settings for Rewrite table, Information Rate, Burst Size, BUM Information Rat, and BUM Burst Size.

参考: Nuageの実装例

The screenshot displays the Nuage management console for a domain named "00_Gold_Customers". The interface is divided into several sections:

- Left Panel:** Lists domain templates and domains. The "00_Gold_Customers" domain is selected.
- Design View:** Shows a hierarchical network design with components like DBaaS, FWaaS, Internet, LBaaS, and nginx, each with associated network and gateway configurations.
- Instantiation Dialog:** A "New Domain Instance" window is open, showing the "Name" field set to "SDN Japan" and the "Description" field set to "Instance-01". A "Create" button is visible at the bottom right of the dialog.
- Buttons:** A red arrow points from the "Instantiate" button in the top right to the "Create" button in the dialog. Another red arrow points from the "00_Gold_Customers" domain in the left panel to the "Instantiate" button.

事前設計済みのネットワークの実体を生成

参考: Nuageの実装例

The screenshot displays the Nuage SDN interface for a domain named "SDN Japan Instance-01". The interface is divided into several sections:

- Left Panel:** Lists domain templates (00_Gold_Custo..., 01_Silver_Custo..., 02_Bronze_Cust..., NuageStdDemo, P2V-BaseTemplate) and domains (NuageStdDemo-D..., SDN Japan Instance-01, P2V-Customer-A, P2V-Customer-B). The "SDN Japan Instance-01" domain is highlighted with a red box.
- Top Panel:** Navigation tabs for Dashboard, Domains, L2 Domains, Users, Groups, Gateways, and Events.
- Design View:** A central diagram showing the network topology. It includes a "24 Hour Activity" graph, "Virtual Machines" (Active: 0, Inactive: 0), and "VPorts" (VM VPorts: 0, Host VPorts: 0, Bridge VPorts: 0). The network components include:
 - Network auto (Hosts: auto)
 - FWaaS (Network: auto, Hosts: auto)
 - Internet (Addressing is managed)
 - LBaaS (Network: auto, Hosts: auto)
 - nginx (Network: auto, Hosts: auto)
 - FW-External Segment-01 (Network: 10.86.114.0/24, Gateway: 10.86.114.1)
 - FW-Internal Segment-01 (Network: 10.47.77.0/24, Gateway: 10.47.77.1)
 - LB-External Segment-01 (Network: 10.116.115.0/24, Gateway: 10.116.115.1)
 - NGINX WEB Servers-01 (Network: 10.27.60.0/24, Gateway: 10.27.60.1)
- Right Panel:** Configuration details for a selected component, showing settings for DHCP (DHCP Managed by VSP, Relay DHCP Disabled), Maintenance (Maintenance Mode), Application Deployment (Deploy applications as zones), and Advanced Networking Information (Route Target: 65534:273, Route Distinguisher: 65534:273, Tunnel Type: VXLAN).

事前設計済みの
ネットワークの実体を生成

参考: Nuageの実装例

The screenshot displays the Nuage SDN interface for a domain named "SDN Japan2". The interface includes a sidebar with domain templates and instances, a central design canvas, and a configuration panel on the right.

Domain Designer - SDN Japan2

Design | Policy Groups | Redirection Targets | Ingress Security Policies | Egress Security Policies | Forwarding Policies

10 objects

DOMAIN TEMPLATES

- 00_Gold_Custo... Gold class tenant
- 01_Silver_Custo... Silver class tenant
- 02_Bronze_Cust... Bronze class tenant
- NuageStdDemo No description given
- P2V-BaseTemplate Base template for ...

MY DOMAINS

- NuageStdDemo-D... No description given
- SDN Japan Instance-01**
- SDN Japan2 Instance-02**

DOMAINS SHARED...

- P2V-Customer-A V1005001001
- P2V-Customer-B V1005001002

24 Hour Activity

Bytes in | Bytes out

Virtual Machines

Active VMs: 0
Inactive VMs: 0

VPorts

VM VPorts: 0
Host VPorts: 0
Bridge VPorts: 0

Network auto
Hosts auto

FWaaS
No description given
Network auto
Hosts auto

Internet
No description given
Addressing is managed

LBaaS
No description given
Network auto
Hosts auto

nginx
No description given
Network auto
Hosts auto

FW-External Segment-01
No description given
Network 10.86.114.0/24
Gateway 10.86.114.1

FW-Internal Segment-01
No description given
Network 10.47.77.0/24
Gateway 10.47.77.1

LB-External Segment-01
No description given
Network 10.116.115.0/24
Gateway 10.116.115.1

LB-Internal Segment-01

DHCP

DHCP Behavior
DHCP Managed by VSP

DHCP Address
Relay DHCP Disabled

Maintenance

Maintenance Mode

Application Deployment

Deployment Policy
Deploy applications as zones

Advanced Networking Information

Route Target
65534:281

Route Distinguisher
65534:281

Tunnel Type
VXLAN

Update

同じ基本設計を持つ
仮想ネットワークの
量産が可能

参考: Nuageの実装例

雛形から継承した
Service Chainingの
トラフィックフロー

The screenshot displays the Nuage SDN management interface. On the left, there are sections for 'DOMAIN TEMPLATES' (including 00_Gold_Customers, 01_Silver_Customers, 02_Bronze_Customers, NuageStdDemo, and P2V-BaseTemplate) and 'MY DOMAINS' (including NuageStdDemo-Domain01, SDN Japan Instance-01, and SDN Japan2 Instance-02). The main area shows 'Forwarding Security Policies' with one object, 'POLICIES FROM TEMPLATE' (including Service Chaining), and 'USER POLICIES'. The 'Forwarding Policies' tab is active, showing a 'Read Only' view of a policy. The 'Origin Location' column lists various network elements like 'Non IP', 'Any', 'Endpoint's own Subnet', 'Endpoint's own Zone', 'Endpoint's own Domain', 'alu-intranet', 'Internet', 'DBaaS', 'DB Segment-01', 'FWaaS', 'FW-External Segment-01', 'FW-Internal Segment-01', 'LBaaS', 'LB-External Segment-01', 'LB-Internal Segment-01', 'nginx', 'NGINX WEB Servers-01', 'DB Servers', 'Firewalls', 'HAProxy', and 'WEB Servers'. The 'Destination Network' column lists similar elements. A blue line connects 'Internet' in the Origin Location to 'Internet' in the Destination Network. A green line connects 'DB Servers' in the Origin Location to 'DB Servers' in the Destination Network. A blue line connects 'WEB Servers' in the Origin Location to 'WEB Servers' in the Destination Network. The right sidebar shows 'Location: Internet', 'Network: Any', and 'Redirects-to-FWs' with details for 'Internet' (Src. Port: N/A, Ethertype: IPv4 - 0x0800, Protocol: Any, DSCP: Any, IP Match: No information, Dst. port: N/A) and 'Anywhere'.

個別要件など雛形との差分は
インスタンス側の実装する



THANK YOU

オーバーレイの可視化モデル (物理・論理マッピング)

