ヤフーのインフラ設計(概要編)

~ LINE Developer Meetup #74 - Private Cloud ~

2022/10/18

ヤフー株式会社 小林正幸

自己紹介

- 小林 正幸
- 所属：サイトオペレーション本部 インフラ技術1部
  - Clos Network Team
  - Private Cloud Network Team
- 業務内容：
  - データセンターネットワーク設計・運用・信頼性向上
  - プライベートクラウドの仮想ネットワーク運用
  - 次世代ネットワーク・インフラの技術検証・開発

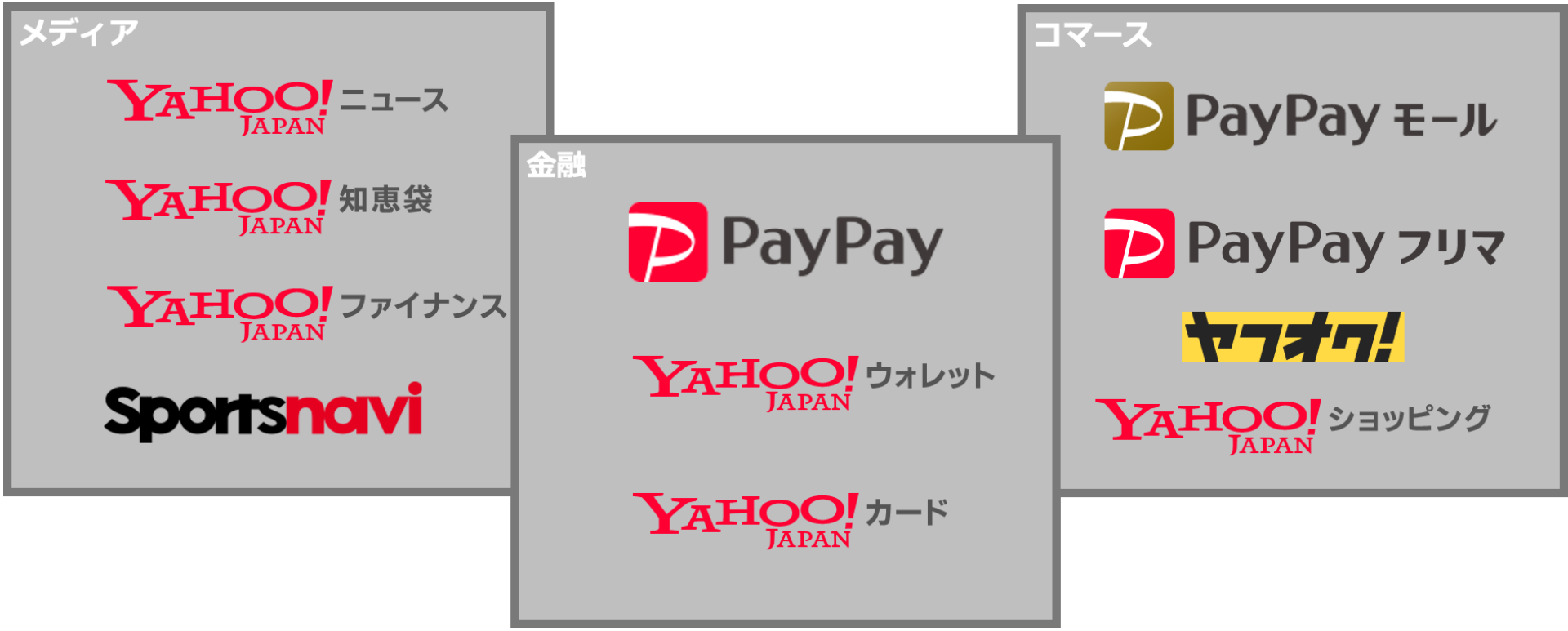
SiteOPS
Yahoo! JAPAN Site Operations

本日お話すること

- **Private Cloud であることの意味**
- **Private Cloud だからできること(ネットワーク編)**
- **Private Cloud の魅力・面白さ**

本日は時間の都合上、ヤフーの Private Cloud の概要と面白さについてお伝えします。
Private Cloud を構成するすべての要素技術や運用について触れることはできないため、
その他の詳細な話などは、次回以降のMeetupで各担当からお話できればと思います。

Yahoo! Japan



国内有数のコンテンツサービスプロバイダ

Public Cloud vs Private Cloud

Public Cloud の利点

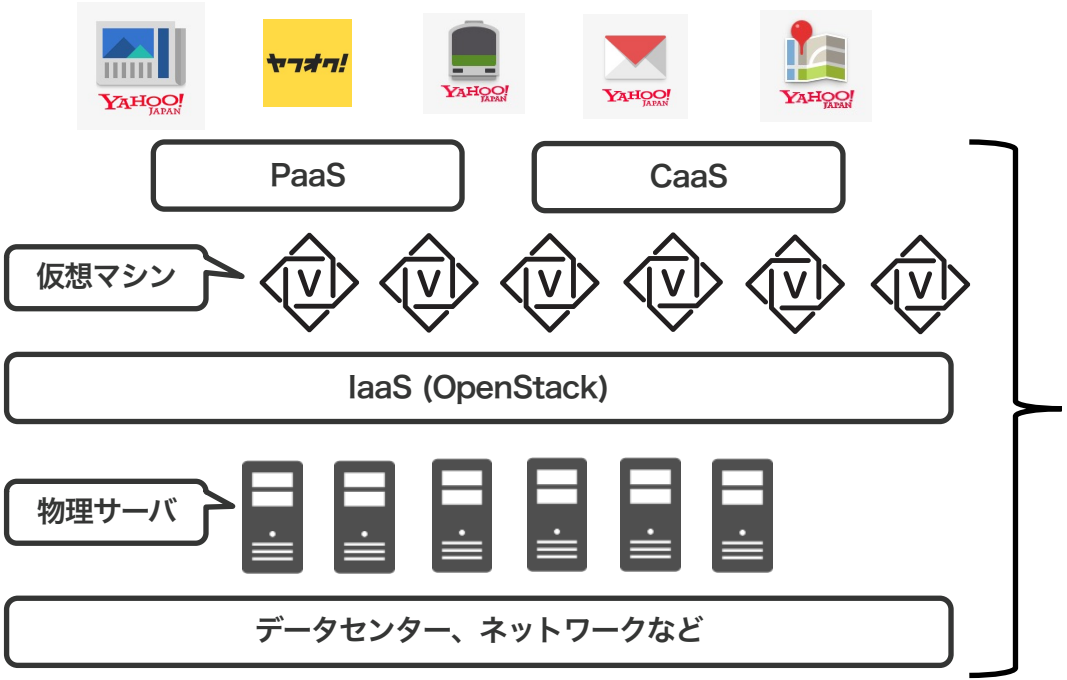
- 豊富なマネージドサービス
- 誰もが利用できる汎用的な設計

Private Cloud の利点

- 大規模なコンピューティング環境では圧倒的なコストメリット
- 自社のビジネスロジックに合わせた詳細な設計・開発・運用が可能
- 自分たちですべてのコードが読める・ブラックボックスが少ない
- 外部サービスの障害の影響を受けにくい

Private Cloud であることの意味

ヤフーのサービスと Private Cloud



サービスの提供に必要な基盤を
Private Cloud で提供

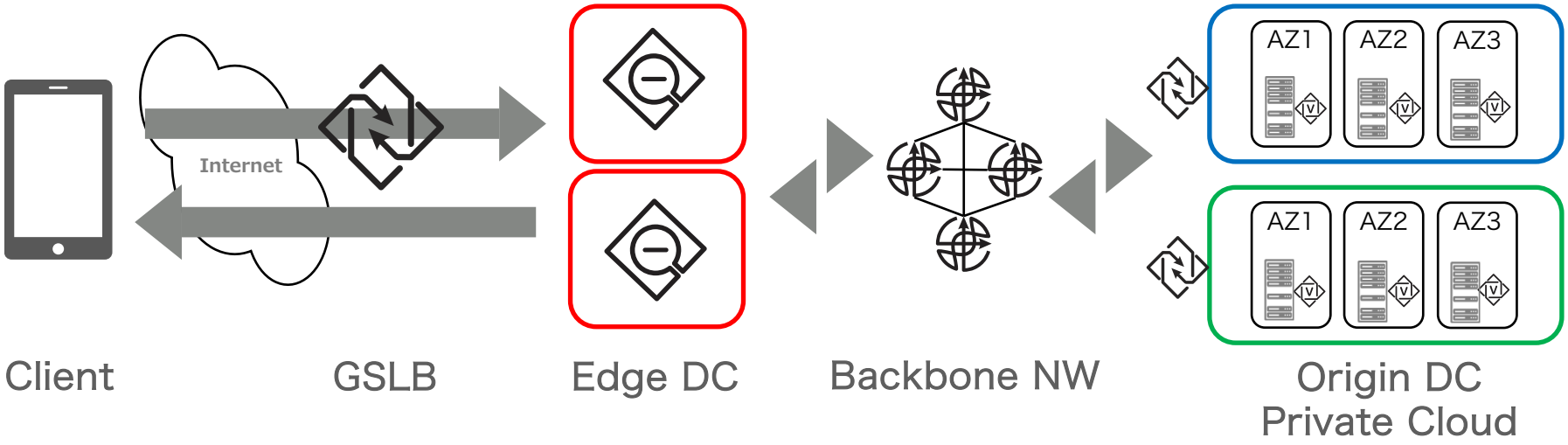
自社DCにてオンプレミスで構築

ヤフーのアーキテクチャと Private Cloud を選択する理由

ヤフーではクライアントのリクエストを直接受けるDCと、IaaSが配置されたバックエンドのDCを物理的に分離して運用

膨大な秒間リクエストを処理するシステムを Public Cloud で完結させることはコストやトラフィック制御の面で難しい

クラウドの利便性のメリットを享受しつつ、開発者の要望を反映し自分たちで完全な制御が効く形が理想



Private Cloud 環境を構成する設備規模

データセンタ数
17

物理サーバ台数
80,000+

ネットワーク機器台数
10,000+

ラック数
10,000+

仮想サーバ台数
170,000+

トラフィック量
800Gbps
※ CDNを除く

Private Cloud だからできること(ネットワーク編)

技術的課題の解消や実装

自社データセンターの建設

データセンターを土地から選定して自社で建設、建物の設計からヤフー仕様にカスタマイズ

電力コストなどを考慮し、各リージョンに配置するサーバを調達・設計可能

各リージョンで自社の規模に見合うAvailability Zoneを定義することで可用性を向上

東日本リージョン

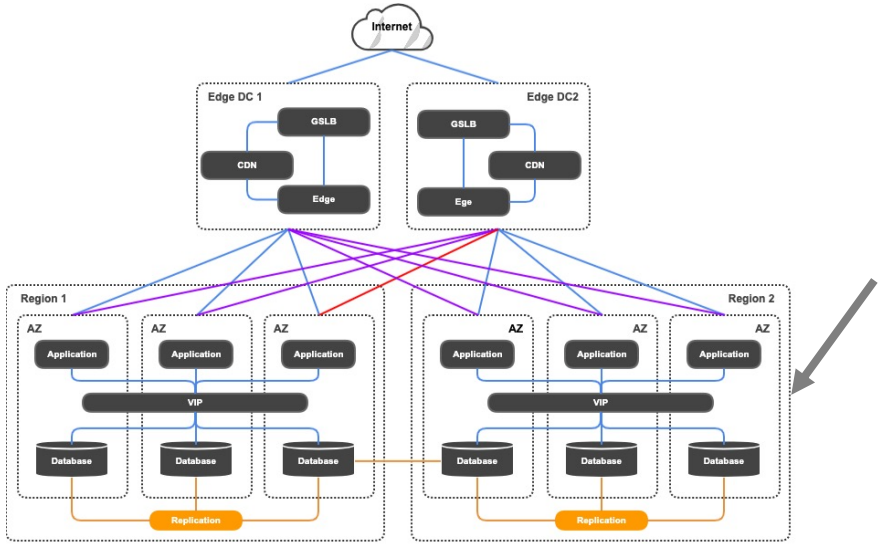
北米リージョン

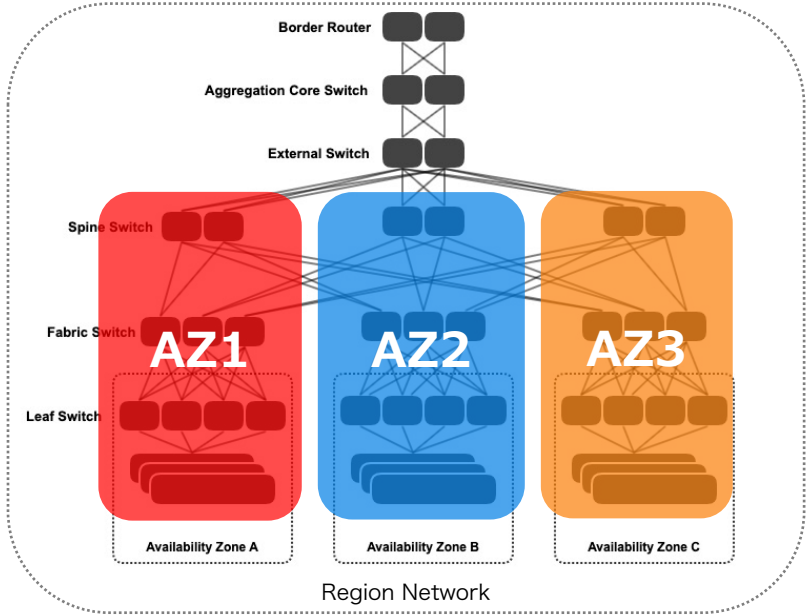
Private Cloud 環境に求められる冗長構成

1つのリージョンに3つ以上のAZを提供する Multi Region {Multi|Many} AZ (MRMAZ) 構成で冗長性を確保

AZは電源系統・空調設備・ネットワーク機器が物理的に独立した区画で、物理面の障害をAZに閉じ込める目的

サービスやプラットフォームのTier Levelに応じてリソースを配置することで、止まらないインフラを目指す





増え続けるトラフィック要求に対応するために

一般的なサーバのLinuxカーネルは汎用性重視のため、パケット処理に特化した用途で性能を求めることには不向き

割り込みによるコンテキストスイッチや、パケットデータのコピーがボトルネックになる

カーネルをバイパスする技術などボトルネックを回避する手法が様々なレイヤに存在し、それらを利用して高速化する

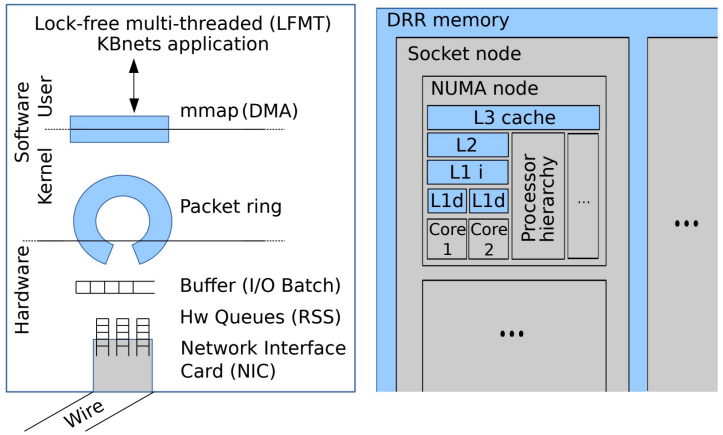
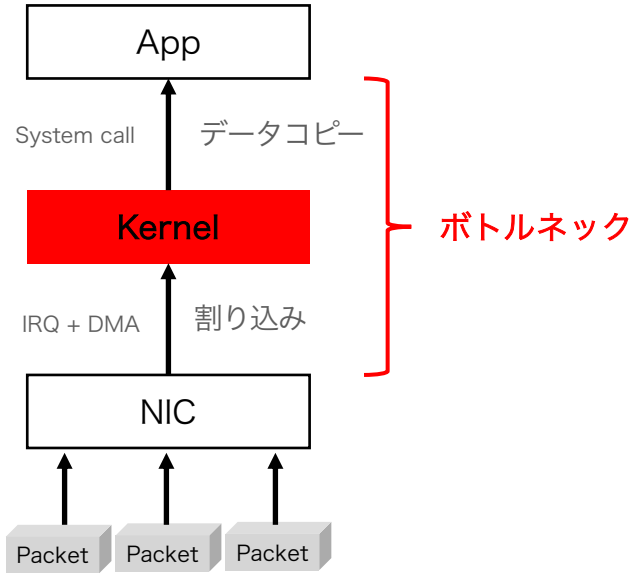


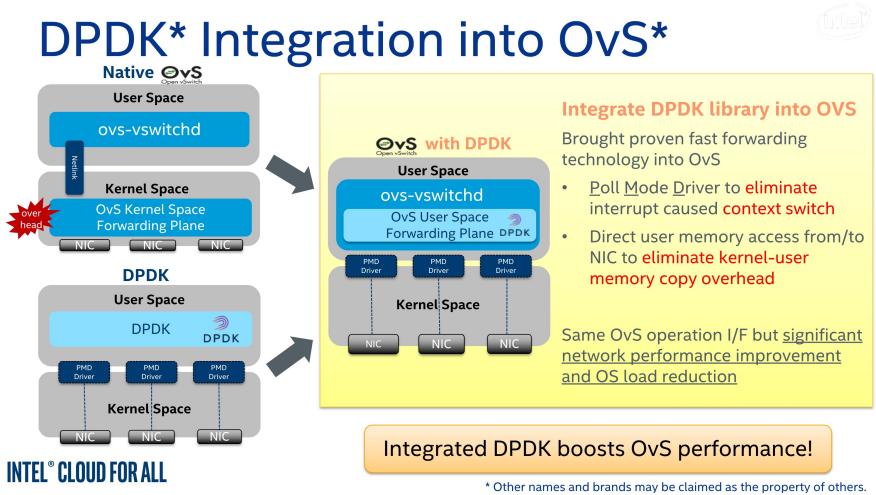
Figure 1: The architecture of a typical COTS software router with special attention the NIC (left) and the memory hierarchy (right).

画像引用元: High-speed Software Data Plane via Vectorized Packet Processing
https://perso.telecom-paristech.fr/drossi/paper/vpp-bench-techrep.pdf



高速で効率的なパケット処理を実現する取り組み

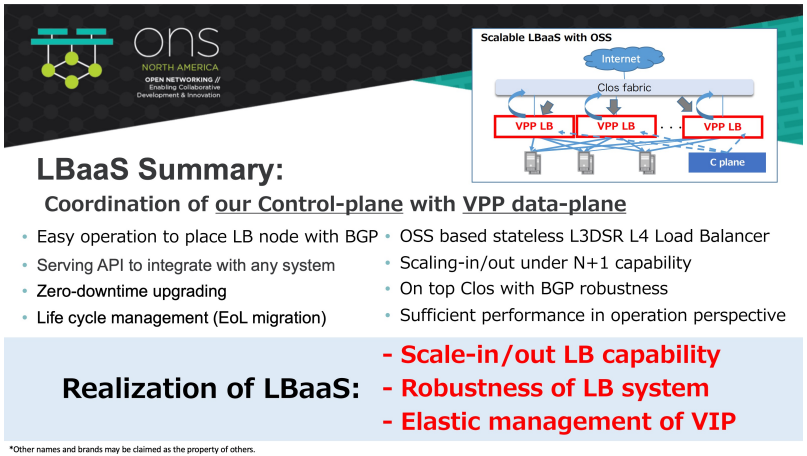
OpenStack Hypervisor Open vSwitch Integration



画像引用元: 2016 OpenStack Barcelona - Yusuke Tatsumi, Naoyuki Mori - Better Application SLA Using Open vSwitch
https://www.youtube.com/watch?v=NSeWzdulnTg

OvSがDPDKを足回りのライブラリとして利用

Software Load Balancer VPP (Vector Packet Processing)

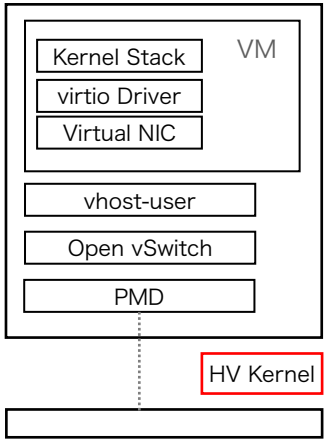


画像引用元: VPP Accelerated High Performance & Scalable L3DSR L4 Load Balancer on Top Clos
http://events19.linuxfoundation.org/wp-content/uploads/2018/07/ONS.NA_.2019.VPP_LB_public.pdf

「パケットの取得」をDPDKが担当
「パケットの処理」をVPPが担当

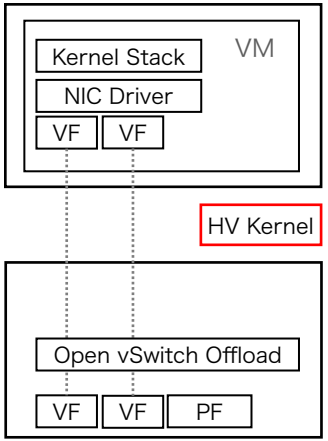
最適なパケット処理を行うNICの検証と導入

Open vSwitch with DPDK
10~25Gbps



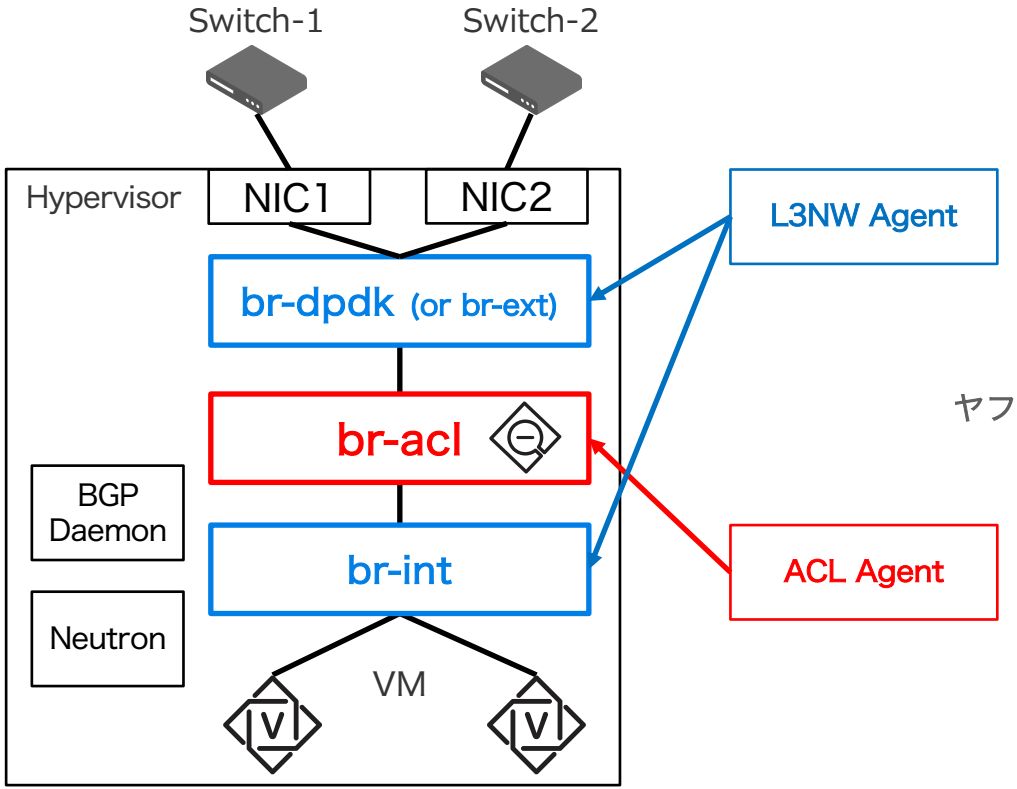
NICがパケットを受信したことをPMDが常にポーリングで監視
PMD用にCPUコアを占有させ、コンテキストスイッチを抑制
VMの収容効率とトレードオフにNWパフォーマンスを向上
現在ヤフー社内で最も利用されている構成

OvS with HW Offload
25~100Gbps



OvSのflow ruleをNICにオフロード
CPUの占有がない
一部の特定用途向けに利用

独自のネットワーク機能の追加



- VMのBGP通信用flow ruleの設定を行うエージェント
- ネットワークの変更を監視、自動で更新
- Neutron自体への変更を少なくするために開発

ヤフー独自の機能は内製のエージェントに実装

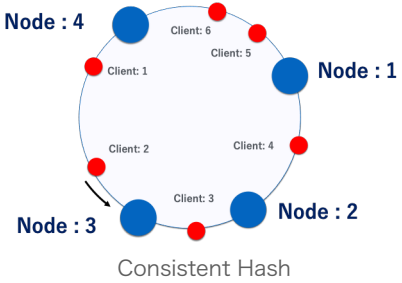
- セキュリティポリシーの設定を行うエージェント
- 利用サービス(テナント)の分離を実現

Stateless Software Load Balancer – L4LBの開発

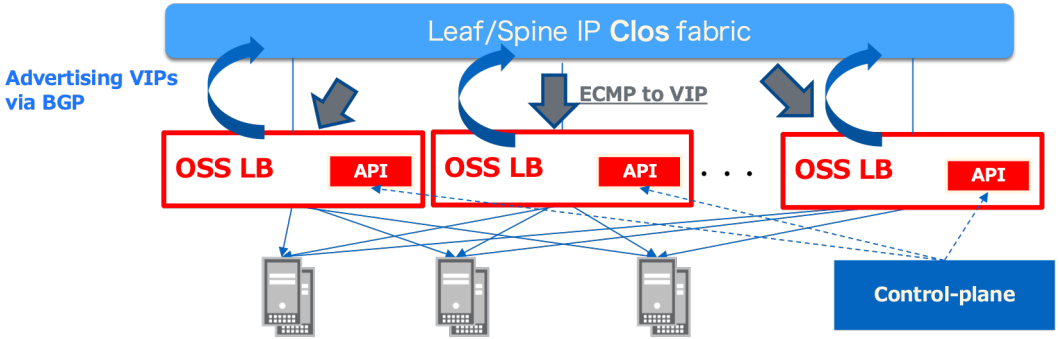
L3 NWのスケーラビリティを活用し、ネットワーク上のどこにでも配置できるmulti-active(N+1) LBを開発・展開

DPDKを利用したOSSである VPP を通常のサーバで利用、 L3DSRで10G/25Gでほぼwire-rateを達成

複数ノードから同一VIPを広報、Consistent-HashとLocal session tableによりTCPパーシステンスを安定維持する設計



Consistent Hash

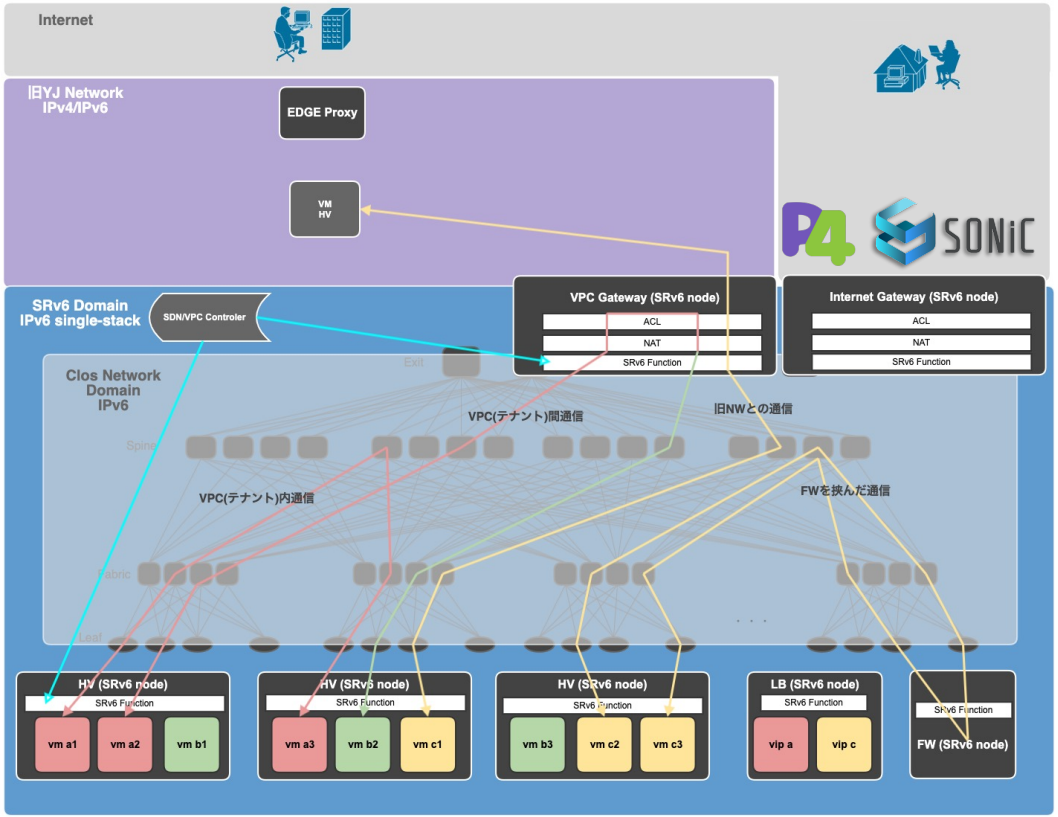


Control Planeは独自実装

- VPP設定
- VIP広報
- Health Check
- L3DSR Check

すべてossベースなので、自分たちで読める・運用者目線の改修が可能

VPC (Virtual Private Cloud) の実現に向けた取り組み ~検証中~



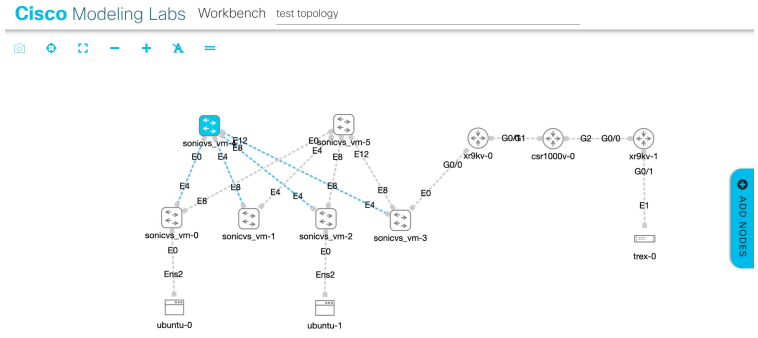
VPCの社内提供に向けてSDNの検証・開発を実施中

論理的に分離された環境で仮想的な機能を提供予定

SRv6などのデータプレーン技術を活用

プライベートクラウド基盤の高度化

利用者が真にインフラを意識しない世界へ



Private Cloud は エンジニアリングの総合格闘技

- データセンター建設、ファシリティ管理
- 機器選定、見積もり、調達交渉
- 回線敷設、サーバ・ネットワーク構築
- デプロイ
- 検証、テスト
- DevOps

物理インフラから、アプリケーション開発、交渉のコミュニケーションまで
あらゆるスキルが活きる!

Private Cloud に関する取り組みの詳細

Cloud Native Telecom Operator Meetup 2022
10/25 16:20-16:50

「CloudNative」の前に考えたい「DevOps」の話
https://cntom.jp/online

Tech-Verse 2022
11/17 16:00-17:00

ヤフーが実践するプロダクション環境でのカオスエンジニアリング
https://tech-verse.me/ja/sessions/175

We are Hiring!

ネットワークエンジニア

https://about.yahoo.co.jp/hr/job-info/role/1454/

インフラ向けツール開発エンジニア

https://about.yahoo.co.jp/hr/job-info/role/1348/

キャリア登録(まずは情報収集したい方)

https://career-user.blm.co.jp/registration/yahoo

Q&A

アンケートの回答よろしくお願いします。