Yahoo! JAPAN アメリカデータセンタのネットワーク変遷

ヤフー株式会社 サイトオペレーション本部 インフラ技術1部 Clos ネットワーク 深澤 開

自己紹介

- 深澤 開 (ふかざわ かい)

- 2013/4 – 2018/6
  - ヤフー入社後、全社Hadoopの設計・運用をしつつ、データセンタネットワークの業務を兼務
- 2018/7 – 2023/3
  - ヤフー米国子会社である Actapio, Inc に出向
  - 米国データセンタのネットワーク設計や運用をメインに米国データセンタの建設プロジェクトや運用も担当
- 2023/4
  - ヤフーへ帰任
  - Clos ネットワークおよびバックボーンネットワークを担当

Actapio とは

- ヤフー株式会社 100%出資 US子会社
- 2014年 YJ America, Inc. として設立、 2017年 Actapio, Inc. へ社名変更
- ワシントン州内でデータセンタを運用
- クラウドプロバイダの業務形態でヤフーへコンピュートリソースを提供
- https://actap.io
- 2023/4 現在はヤフー出向および現地エンジニア 7名体制でアメリカデータセンタの運用を行っている

2014/12~

Yahoo! JAPANで初めてのアメリカデータセンタ運用の開始

2014/12~ Yahoo! JAPANで初めてのアメリカデータセンタ運用の開始

なぜアメリカデータセンタの運用を開始したのか

- BCP 強化
- 電気代が安い
  - 全米で最も電気代の安い地域で、日本の **1/4 – 1/6** 程度
- 水力発電による**100%再生可能エネルギー** (当時)
- 直接外気冷房を採用
  - 現地の乾燥した気候を利用することで、**より低いPUEを実現**できる

YAHOO! JAPAN　企業情報　ニュース　サービス　テクノロジー　CSR　採用情報

プレスリリース　2014.11.10

米国に保有するデータセンターについて

ヤフー株式会社
YJ America, Inc

～ サーバーの国外分散により、BCP強化と運用費削減を実現 ～

ヤフー株式会社（以下、Yahoo! JAPAN）の米国現地法人「YJ America, Inc」（以下、YJ America）は、米国ワシントン州において保有するデータセンターについて、来春（2015年4月予定）の正式稼働に向け、本年12月よりテスト稼動を開始することになりましたのでお知らせします。

Yahoo! JAPANでは、これまでに自社サービスの開発やコンテンツ格納において、連結子会社である株式会社IDCフロンティアが運営するデータセンターや自社で保有するサーバーなど、国内に設置している基盤を利用していました。来春からは従来の環境に加え、この米国データセンターの利用も開始する予定です。

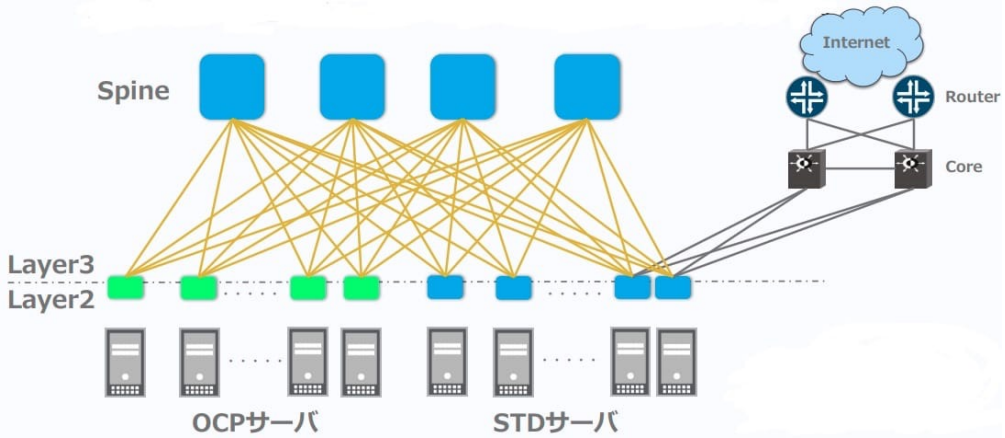
なお、本データセンターの利用により、Yahoo! JAPANグループで保有するデータセンターは国内での分散（東日本エリアと西日本エリア）のみならず国外への分散となるため、BCP（自然災害や大規模停電などに対応した事業継続計画）のさらなる強化が実現する見込みです。また、米国データセンターが設置されている地域の電気料金が非常に廉価であることから、Yahoo! JAPANグループ全体のサーバー運用費の削減もはかれる見通しです。

https://about.yahoo.co.jp/pr/release/2014/11/10a/

2014/12~ Yahoo! JAPANで初めてのアメリカデータセンタ運用の開始

データ分析基盤向けClosネットワーク

- Spineスイッチにシャーシタイプ、Leafスイッチにネットワークメーカースイッチとホワイトボックススイッチを採用
- Spine/Leaf間は40G-LRを4本
- 社内IPAMからコンフィグ生成+ZTPの本格利用を開始したデプロイ
- サーバは標準機とOCPサーバを採用
- サーバのNICは10Gbps



https://techblog.yahoo.co.jp/entry/20200323819517/

2019/4~

新アメリカデータセンタの運用の開始

旧アメリカデータセンタからのデータ移行

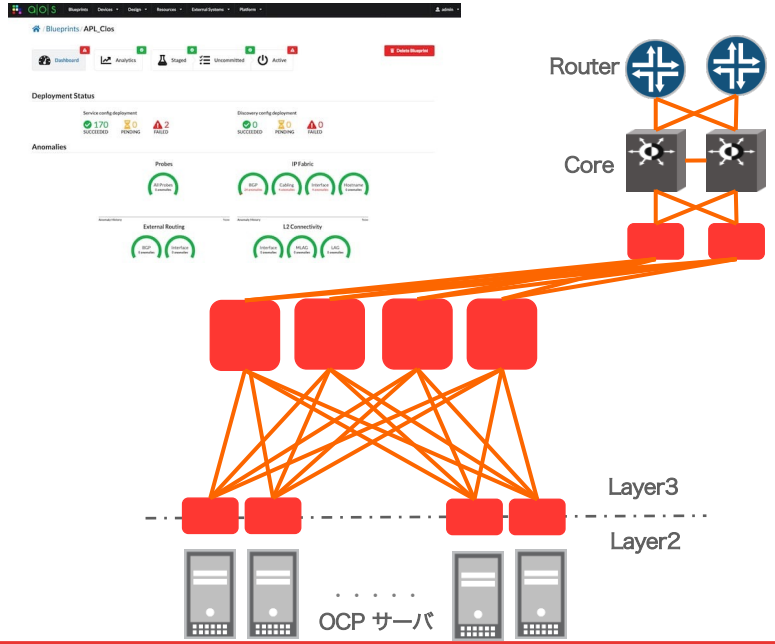
新アメリカデータセンタの建設

- 2014/12 に運用を開始したデータセンタと同じワシントン州内に新たにデータセンタを建築
- 2018/3 に着工し、2019/2 に竣工
- データセンタ概要
  - 建築面積 : 約9,300㎡
  - 敷地面積 : 約180,400㎡
  - 電力容量 : 16MW (竣工時は2MW)
  - ラック数 : 約1,600ラック (竣工時は約200ラック)
  - 建物構造/規模 : 鉄骨造/地上1階
  - 受電種別 : 100%再生可能エネルギー（水力発電）
  - 空調方式 : 直接蒸発式外気冷房（100%外気空調）
  - PUE : 1.2以下

https://about.yahoo.co.jp/pr/release/2018/03/09a/

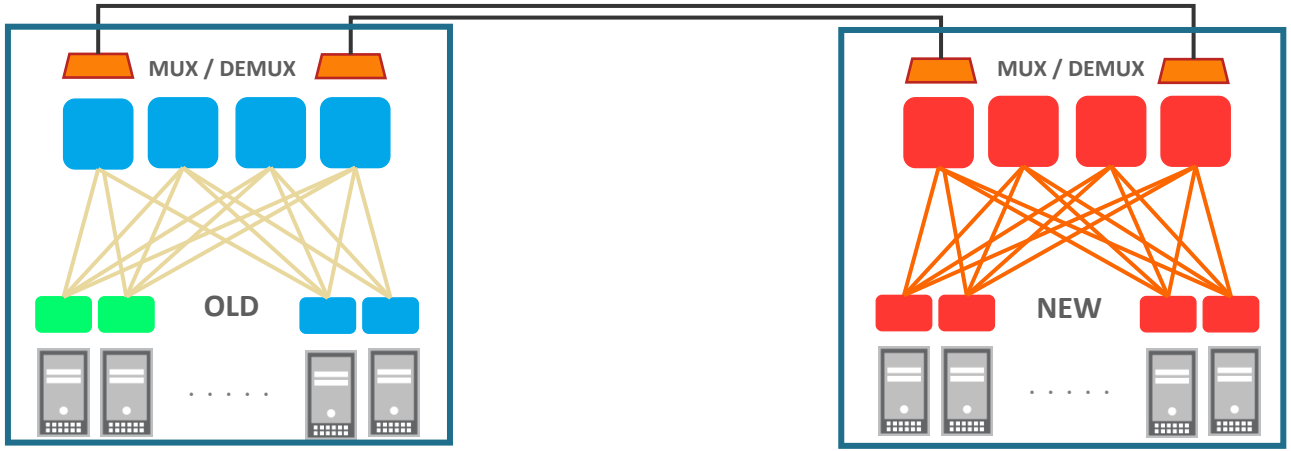
新アメリカデータセンタのネットワーク構成

- 基本的な構成は旧アメリカデータセンタの構成と同じ
- Spineスイッチにシャーシタイプ、Leafスイッチにネットワークメーカースイッチを採用
  - **Deep Buffer** のモデルを採用したため
- Spine/Leaf間は**100G-CWDM4**を4本
- デプロイには **Apstra AOS**を採用
- サーバは**全てOCPサーバ**を採用
- サーバのNICは**25G**bps
  - 一部GPUサーバ向けに**100G**bpsもあり
- **サーバ間も含めたL3構成**や **Shared IPMI** の導入を開始



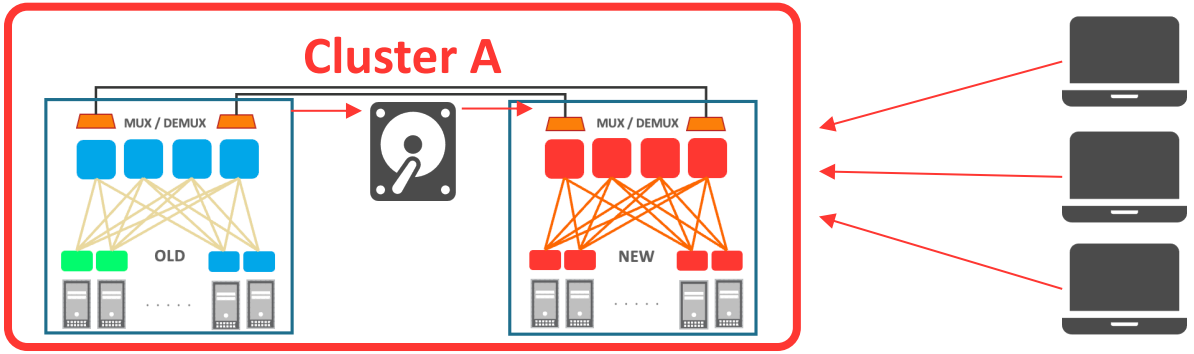
新アメリカデータセンタのネットワーク構成

- 新データセンタと旧データセンタ間に**ダークファイバ**を引き、各データセンタのClosネットワークの4Spineのうち2Spineのシャーシに **DWDMモジュール**を挿入し新旧データセンタ間を接続
  - 新旧データセンタ間の距離は約**20**km
  - **1.2T**Bps 冗長



ダークファイバでのデータセンタ間接続の理由

- **ユーザが意識しなくても、データ分析基盤の移行をするため**
- 旧データセンタから新データセンタのデータ分析基盤へユーザに移ってもらう必要があった
- 新旧データセンタ間をダークファイバでつなぐことで、拠点を跨いでも1つのデータ分析基盤と扱って問題ないだけの帯域を確保できるようにした
- これにより、**ユーザが移行作業すること無く**、旧データセンタからデータ転送やサービスアウトをすることで新データセンタのデータ分析基盤へ処理やデータを移すことを可能にした

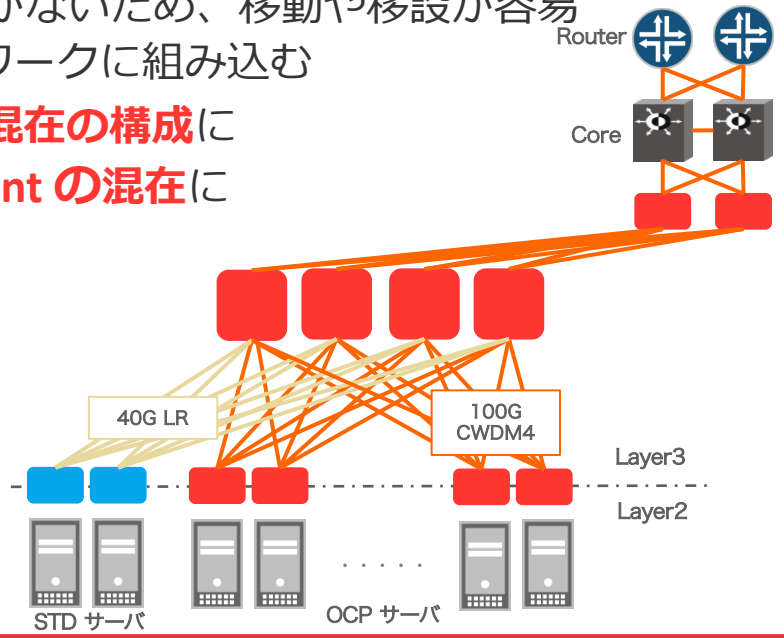


2019/12~

旧アメリカデータセンタから物理移設

ラックを物理的に移動

- 再利用したいサーバやストレージ(46ラック)を物理的に新データセンタへ移動
- 利用が一旦完了しているため、電源を落とし、ラックまるごと移動
  - 日本と違いラックの床への打ち付け固定がないため、移動や移設が容易
- 移動後、配線・機器の設定を行い、ネットワークに組み込む
- Fabric/Leaf 間の帯域が **40Gbps と 100Gbps 混在の構成**に
- Leaf スイッチも **Front-to-RearとRear-to-Front の混在**に
- 輸送の影響でディスクもそれなりに壊れた
  - A社 SAS 4TB : 0.075% (6 / 7980)
  - B社 SATA 4TB : 4.6% (28 / 600)
  - B社 SATA 8TB : 0.3% (24 / 7980)
  - 全体 : 0.3% (58 / 15960)



2019/12~ 旧アメリカデータセンタから物理移設

ラックを物理的に移動

STD サーバ　　OCP サーバ

2020/7~

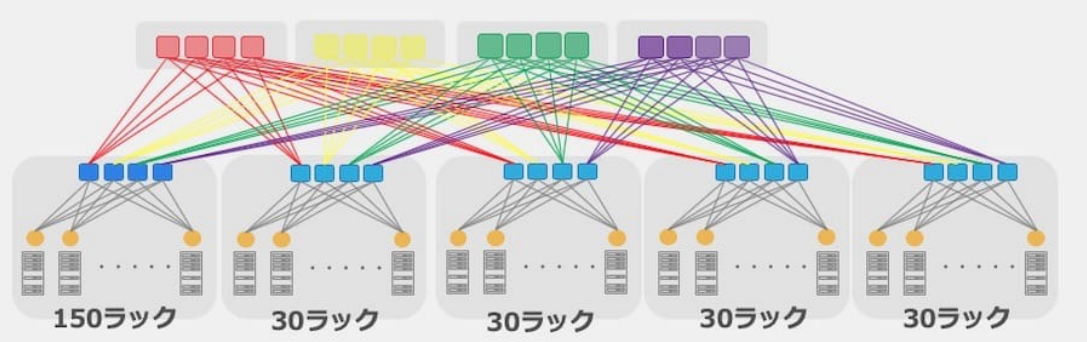
コロナ禍での新アメリカデータセンタの拡張とネットワークの構成変更

コロナ感染による、拡張工事停止のリスク

- 新データセンタの二つ目のデータホールの建設が始まった頃にコロナ禍に
- 工事業者やデータセンタ運用メンバのコロナ感染により**工事が停止するリスク**の中でネットワーク拡張の現地作業をする必要があった
- データセンタ運用メンバの中でのシフト勤務だけでなく、**工事業者の作業時間や移動経路を考慮しての勤務**をすることとなった
  - 設置作業や配線作業が**深夜時間帯メイン**に

Clos ネットワークの構成の変更と配線作業

- 二つ目のデータホール建設に伴い、Closネットワークも拡張のため、構成変更が必要となった
  - **2層構成**では収容しきれなくなるため、**3層構成**へ変更
- 構成変更に伴い、新規配線が大幅に増えた
  - SpineスイッチとFabricスイッチ間を新規で配線
  - 拡張分のFabricスイッチとLeafスイッチ間も当然必要
  - 夜な夜な配線作業をこなした



https://techblog.yahoo.co.jp/entry/20200323819517/

2020/10~

コロナ禍での旧アメリカデータセンタの撤退

旧データセンタの契約満了と撤退時の苦労やトラブル

- 旧データセンタの契約が **2021/2末** に満了
  - 機器などを全て撤去した状態でデータセンタを返却する必要あり
- ネットワーク的には使わなくなったラックのスイッチを落とすだけだったり、ネットワーク毎に広報を止めると言ったことはせず、上流スイッチを落とすだけといった割とシンプルな作業だったが、、、

旧データセンタの契約満了と撤退時の苦労やトラブル

- **スケジュールが相当タイト**
  - **2020/12末** まで使い続ける予定のサービスがいた
  - **2021/1中** の機器の引き取りを予定していたため、電源を落とせるラックがあったら、 1日たりとも遊ばせずすぐ電源落とす
- **日本からのヘルプ無し**
  - コロナ禍で移動制限がかかっており、日本からヘルプに来てもらえず
  - **17時間の時差を乗り越えた**日本とのコミュニケーションを実施
  - 約17,000本のディスクを現場エンジニア6人で抜き取り作業し、**ディスクを留めているネジに恨みを覚える**日々
- **DNS リゾルバの変更漏れ**
  - 新データセンタ側のDNS リゾルバの向き先が旧データセンタに向いていることが**2020/12** に発覚し、緊急で対応
- **引き取りトラックのキャンセル**
  - 引き取った機器を載せる予定のトラックとドライバーが**急にキャンセル**になる

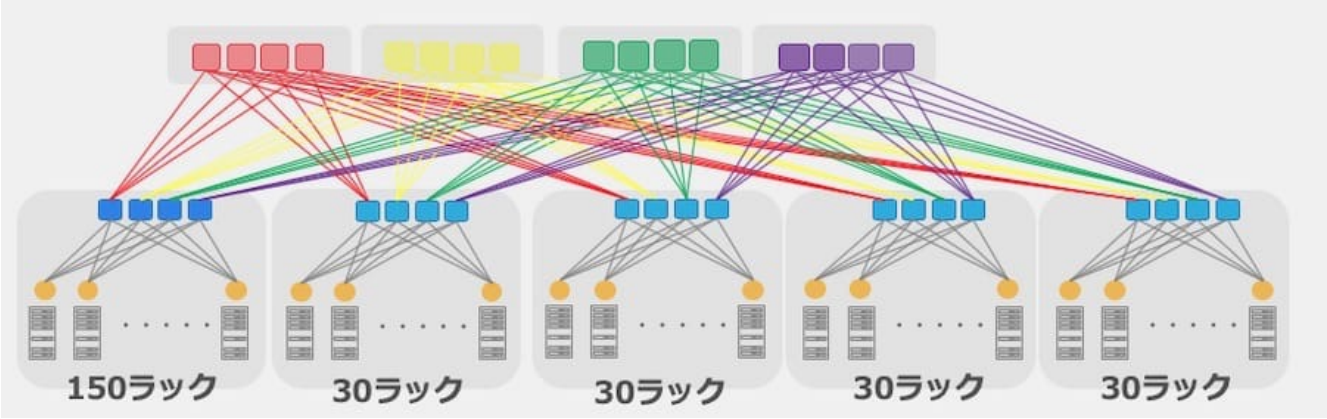
2021/3~

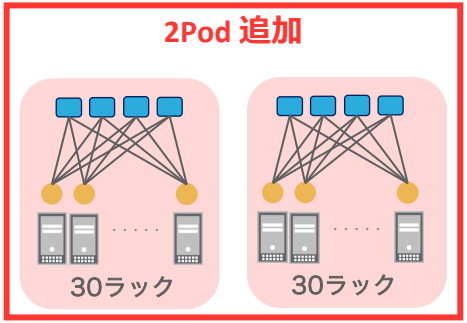
ネットワーク拡張

2021/3~ ネットワーク拡張

三つ目のデータホールの竣工と納期遅延

- 建設済みのデータホールで順次ネットワークを拡張しつつ、 **2022/7** に **三つ目**のデータホールが竣工
  - それに伴い、新たにPodを追加
- サーバ、ネットワーク機器に限らずデータセンタの設備関連のものに納期の遅れが発生し、各所の調整に本当に苦労した
- そのほか、旧データセンタから物理移設したラックの撤退も実施





2023/4~

これから

2023/4~

これから

- 現在もデータセンタを順次拡張している
  - **四つ目のデータホール** の建設計画が進行中
- Clos ネットワークも拡張
  - 今の構成のままでは収容しきれないため **Spineの追加** を計画中
- アメリカデータセンタの利点をより活かすために
  - 安価な電気代をより活かせるように **GPU サーバ** や **400G** のネットワークにも挑戦していく