Yahoo! JAPAN アメリカデータセンタとネットワーク変遷

ヤフー株式会社 サイトオペレーション本部 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・構築・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
  - ヤフーアメリカ子会社である Actapio, Inc に出向
  - アメリカデータセンタ現地でのネットワーク設計や
    運用をメインに、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の設計・構築・運用を担当

JANOG51.5 の振り返り

JANOG 51.5 振り返り

- なぜアメリカのデータセンタで運用を開始したのか
  - BCP / 電気代 / 空調
- アメリカデータセンタでのネットワーク構成の変遷
  - 各データセンタのネットワーク構成
  - 二つのデータセンタ間のサービス移行
  - ネットワーク拡張
- 現場での苦労話
  - 物理的な移設
  - コロナ禍でのデータセンタ建設
  - コロナ禍での旧データセンタの撤退

自分のいた5年間の経験はこんなものじゃない!

もっと話したいアメリカデータセンタのことや
アメリカならではの話がまだまだある!

今回の内容

- **Yahoo! JAPAN のアメリカデータセンタ**
  - なぜアメリカでデータセンタの運用を開始したのか
  - 電気代、空調、日本との違いについて
  - なぜ新アメリカセータセンタを建設したのか
  - アメリカの災害
- **アメリカデータセンタのネットワークの変遷**
  - ネットワーク構成を深掘り
- **苦労やトラブル**
  - 物理的な移設やコロナ禍の苦労話
  - 撤退時のトラブルやその解決策
  - 1000年に1度の熱波襲来
  - 回線障害
- **これから**
- **まとめ**

Yahoo! JAPANの
アメリカデータセンタ

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
- 運用体制 (2023/6 現在)
  - サーバエンジニア: 4名
  - ネットワークエンジニア: 3名
- https://actap.io

Yahoo! JAPANで初めてのアメリカデータセンタ運用の開始

- **アメリカでデータセンタを契約**し、2014/12 から運用を開始
- アメリカデータセンタを利用開始した目的
  - **BCP 強化**
  - **電気代が安い**
  - 水力発電による**100%再生可能エネルギー**
  - **直接外気冷房**の採用

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施設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/

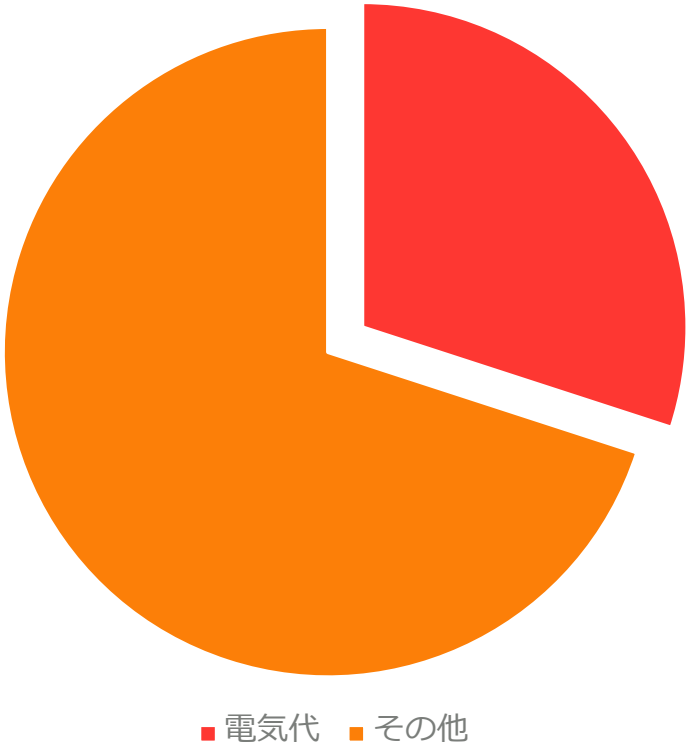
データセンタ主要コスト

- 建物減価償却費
- 回線費
- 業務委託費
- 電気料金
- 設備メンテナンス費

Yahoo! JAPANのアメリカデータセンタ

データセンタ主要コスト

- 建物減価償却費
- 回線費
- 業務委託費
- **電気料金**
- 設備メンテナンス費



電気料金は
全体の2-3割

Yahoo! JAPANのアメリカデータセンタ

データセンタ主要コスト

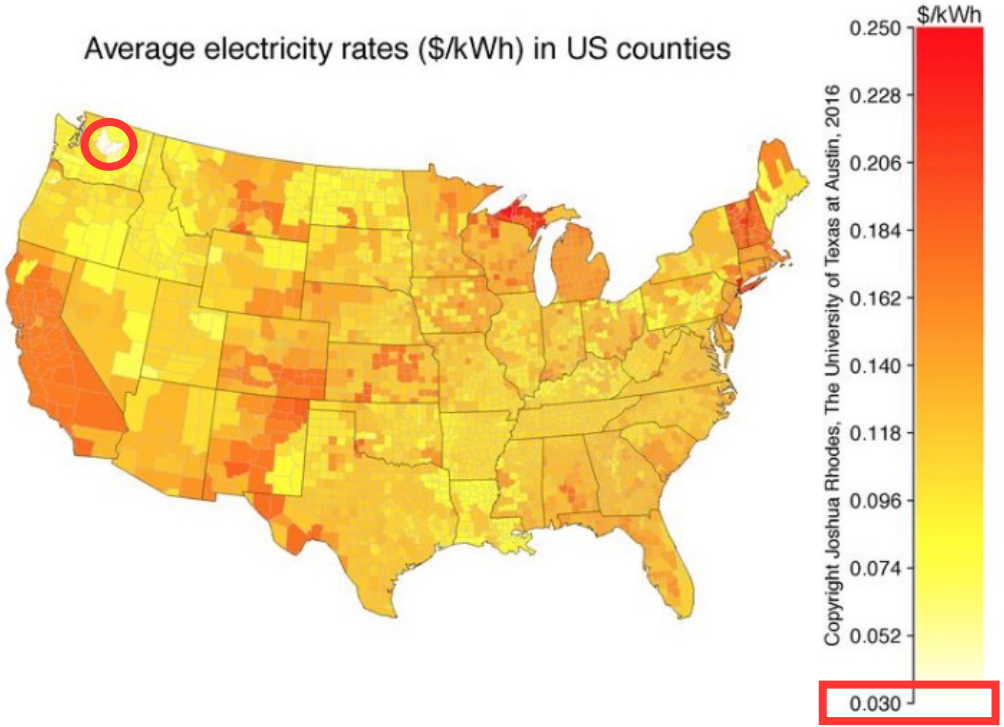
- 建物減価償却費
- 回線費
- 業務委託費
- **電気料金**
- 設備メンテナンス費

電気料金は
全体の2-3割

▼

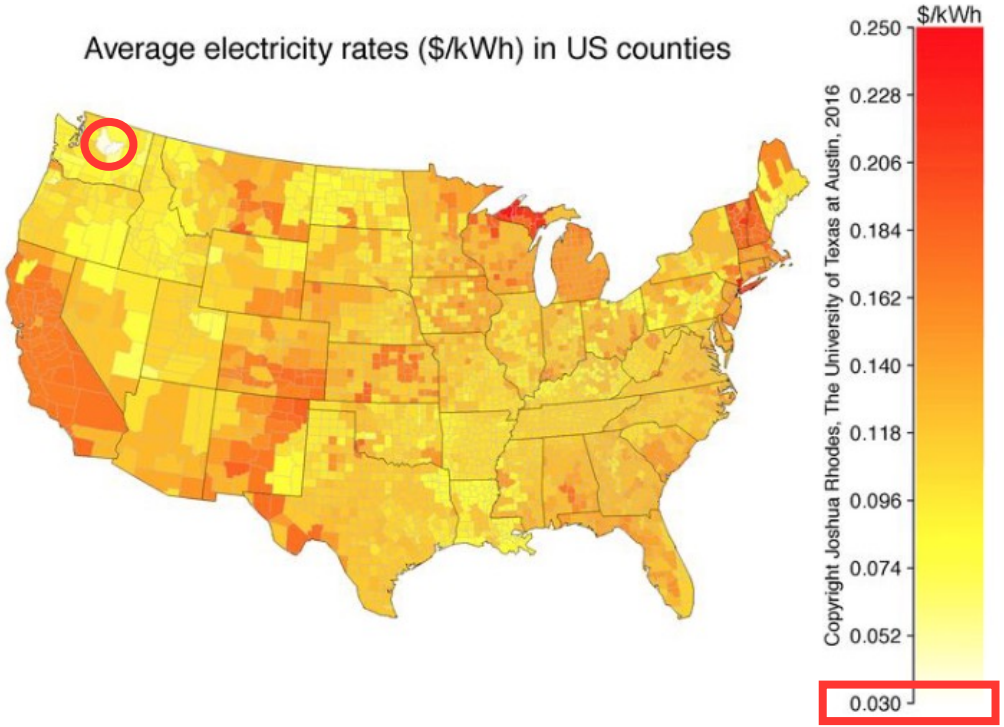
日本では
上昇傾向

アメリカの電気料金事情



https://www.renewableenergyworld.com/solar/when-will-rooftop-solar-be-cheaper-than-the-grid-here-s-a-map/#gref

アメリカの電気料金事情

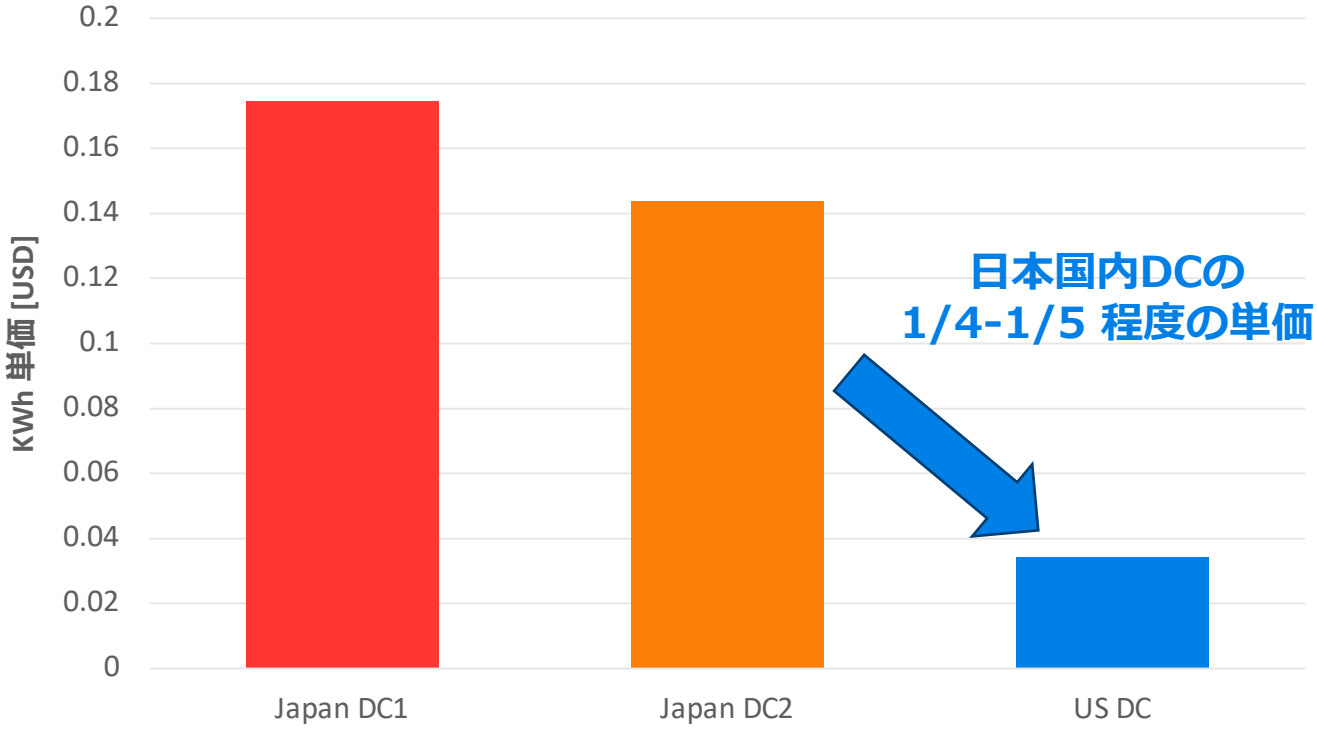


KWh 単価
3セント未満

全米 No.1
低電気代エリア

https://www.renewableenergyworld.com/solar/when-will-rooftop-solar-be-cheaper-than-the-grid-here-s-a-map/#gref

日米における電気料金比較 (2022年度)

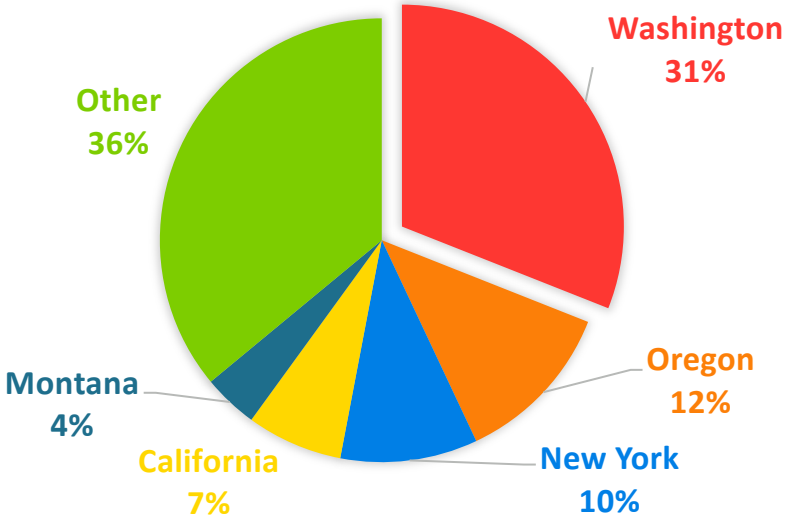


※ 各月の月間平均為替レートで計算

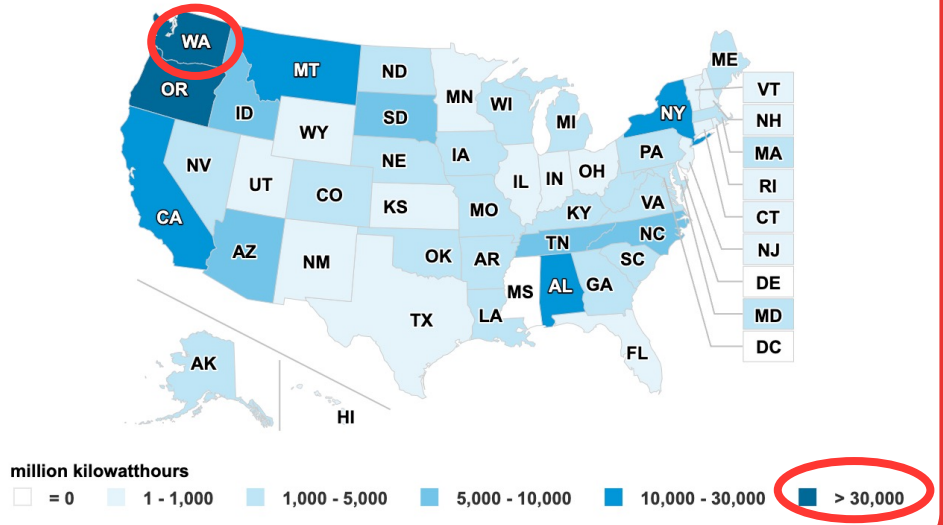
なぜ電気代が安いのか

- ワシントン州の**水力発電容量は全米一位** (全米の27%を占める)
- コロンビア川の豊富な水量を利用した大規模水力発電
- コロンビア川本流に**14**のダム（全米最大のグランドクーリーダム etc）

米国内における水力発電容量主要5州(2022年)



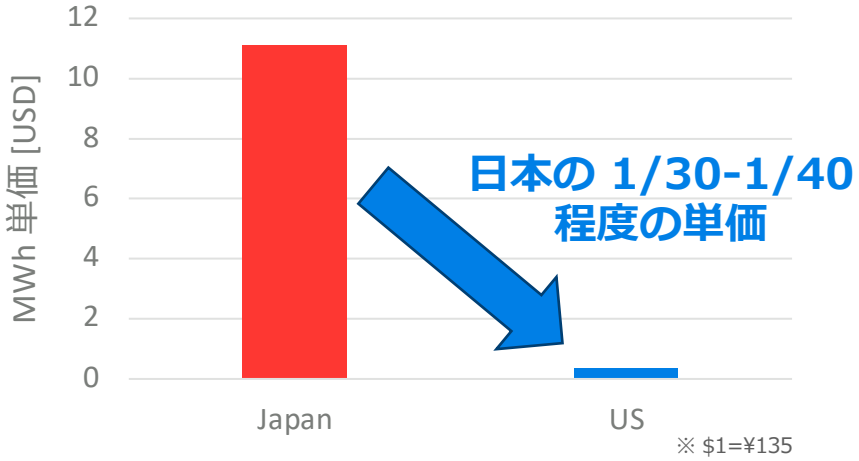
米国内における水力発電容量分布(2022年)



https://www.eia.gov/energyexplained/hydropower/where-hydropower-is-generated.php

電気料金の上昇と非化石証書(REC)

- アメリカでも電気料金の値上げはある
  - 水力発電の**設備メンテナンスのための値上げ**
  - **数年におよび計画された数%ずつ**で、原材料の高騰といった理由ではない
- 非化石証書(REC)も安い
  - 購入する場合でも**コストインパクトが低い**



新アメリカデータセンタの建設開始

- 4年近くアメリカでデータセンタを運用し、ノウハウが貯まってきた
- 最初のアメリカデータセンタはレンタル契約
  - 設備が古くなり、経年劣化から故障頻度が上がっていた

新アメリカデータセンタの建設開始

- 4年近くアメリカでデータセンタを運用し、ノウハウが貯まってきた
- 最初のアメリカデータセンタはレンタル契約
  - 設備が古くなり、経年劣化から故障頻度が上がっていた

新たな設備で、蓄積したノウハウを更に活かすべく
データセンタを自分達で設計・建築をスタート

新アメリカデータセンタ概要

- 最初のアメリカデータセンタと**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/

直接蒸発式外気冷房(100% 外気空調)

- 熱源を不使用なため、**非常に省エネルギーな空調システム**
  - 加湿冷却モード
    -> 空気を加湿し、気化熱を利用して気温を下げる
  - フリークーリングモード
    -> 室内より外気の気温が低い場合に、外気を混ぜる形で室内の気温を下げる
- アメリカデータセンタの周辺地域の気候特性
  - 夏：**日中最高気温40度近くなるが、湿度が10-20%**
  - 冬：**積雪があり、気温がマイナス**

直接蒸発式外気冷房(100% 外気空調)

- 熱源を不使用なため、**非常に省エネルギーな空調システム**
  - 加湿冷却モード
    -> 空気を加湿し、気化熱を利用して気温を下げる
  - フリークーリングモード
    -> 室内より外気の気温が低い場合に、外気を混ぜる形で室内の気温を下げる
- アメリカデータセンタの周辺地域の気候特性
  - 夏：**日中最高気温40度近くなるが、湿度が10-20%**
  - 冬：**積雪があり、気温がマイナス**

現地の気候特性にマッチした空調システム

PUEへの効果

https://www.raritan.com/blog/detail/what-is-power-usage-effectiveness-pue-and-how-is-it-calculated

- Power Usage Effectivenessの略
  - データセンターの電力使用効率を示す指標
  - PUE = (Total Facility Energy) / (IT Equipment Energy) で表される
  - 1 に近い数字であればあるほど良い
- Total Facility Energyには空調などIT機器以外の電力も含まれる

PUEへの効果

https://www.raritan.com/blog/detail/what-is-power-usage-effectiveness-pue-and-how-is-it-calculated

- Power Usage Effectivenessの略
  - データセンターの電力使用効率を示す指標
  - PUE = (Total Facility Energy) / (IT Equipment Energy) で表される
  - 1 に近い数字であればあるほど良い
- Total Facility Energyには空調などIT機器以外の電力も含まれる

空調の電力を削減することでPUEを低くできる

PUEへの効果

https://www.raritan.com/blog/detail/what-is-power-usage-effectiveness-pue-and-how-is-it-calculated

- Power Usage Effectivenessの略
  - データセンターの電力使用効率を示す指標
  - PUE = (Total Facility Energy) / (IT Equipment Energy) で表される
  - 1 に近い数字であればあるほど良い
- Total Facility Energyには空調などIT機器以外の電力も含まれる

空調の電力を削減することでPUEを低くできる

平均PUE 1.19 (2021年度)

※JDCC 第32回「省エネルギー小委員会」資料 では平均PUE 1.7

https://www.meti.go.jp/shingikai/enecho/shoene_shinene/sho_energy/pdf/032_08_00.pdf

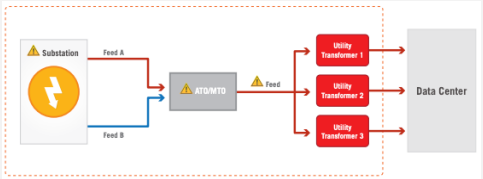
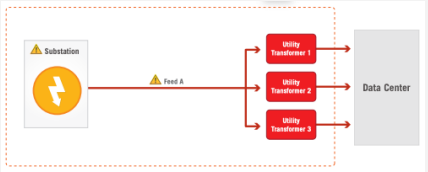
日本のデータセンタとのちがい(建築)

	Japan	USA
基礎	数10mの杭 + 鉄筋コンクリート基礎 ・杭打ちだけで数ヶ月	鉄筋コンクリート基礎(2m弱)のみ ・数日で基礎工事完了
耐震 免震構造	免震構造有り ・複雑な構造で工期・コストがかかる	免震構造無し ・シンプルな構造で工期短縮・コスト削減
資材	外壁は既製品のALC、押出成形セメント板を利用 ・プレキャストコンクリートは高価 小型資材を大量に現地へ輸送して設置 ・工期がかかる	外壁はプレキャストコンクリートを利用 ・プレキャストコンクリートが一番安価 ・特注のため空調用の開口などが柔軟に対応可能 大型資材を少数で現地へ輸送して設置 ・工期短縮

Yahoo! JAPANのアメリカデータセンタ

建築時の写真

日本のデータセンタとのちがい(設備)

	Japan	USA
受電構成	1系統 2経路受電が主流 •コスト高 •商用電源の信頼性が高い 	1系統 1経路受電が主流 •コスト低 •商用電源より自家発電機の方が信頼性が高い 
コミッショニング	導入設備の**一部**において、性能試験を実施	導入設備**全て**において、最大負荷での性能試験を実施
非常用燃料	優先供給契約	優先契約なし
納品	部品で輸送し、現地で組み上げ なるべく小型な完成品を輸送して設置	大型設備でも完成品をそのまま輸送して設置

https://www.datafoundry.com/media/documents/dual_power_feeds.pdf

竣工前の最大負荷での性能試験

- データセンタの設備で **Level One から Five** の試運転テストを実施
- **Level Five は Integrated Systems Testing (IST)** と呼ばれ、フル負荷を含む様々な負荷をかけて、設備の切り替え試験などを実施し、施設全体が問題無く機能するかの試験を行う
  - 例：フル負荷の状態で商用電源から非常用電源に切り替え・切り戻し

Level One – Factory Witness Testing
One of the key success factors in factory witness testing is to have a clearly defined test protocol in the purchase specification. This ensures that each supplier is providing a common test program allowing a better comparison of each manufacturer's value proposition. In addition it prevents substandard testing approaches from being submitted after award of the order. To achieve these goals, the Cx agent and INAP team will work with the project team to fully specify the testing that is desired at the manufacturer's facility.

Level Two – Site Acceptance Inspection
The acceptance of equipment as it is delivered to the construction site is the responsibility of the installing contractor, however, it is a great opportunity to verify that all components have been shipped and any loose items have been inventoried and stored in a secure location.

Level Three – Pre Functional Testing & Start Up
The pre functional check and startup of commissioned system is the responsibly of the installing contractor and manufacturer authorized technician. Project specific PFT checklists will be prepared by INAP and the Cx agent and distributed to the installing contractors.

Level Four – Functional Performance Testing
The verification of equipment and system operations during functional performance testing is the last point in the building commissioning process where major issues are expected to be unearthed. The majority of the issues should have been identified earlier in the commissioning program

Level Five – Integrated Systems Testing
The Integrated System Test (IST) is pinnacle of the commissioning program, and the performance of these activities demonstrates the performance of the facility as a whole against the owner's project requirements. The commissioned systems are operated at various loads and in various modes to demonstrate fully automated operation and proper response to equipment failures and utility problems.

https://www.inap.com/blog/integrated-systems-testing-for-data-centers/

日本のデータセンタとのちがい(データホール)

	Japan	USA
消火設備	ガス消火が主流 ・他の設置機器への影響を最小限にくい止める	スプリンクラー消火が必須 ・人命の安全確保が最優先、局所放水により消火 ・ガス消火の併用は可能だがオプション ・部屋の容積が大きく、ガスの充填は困難・高価
納品	現地でラックへの機器搭載が一般的 ・現地作業員等が搭載・配線を実施 ・現地作業員が多数必要	機器インテグレーション済みのラック納品が主流 ・インテグレーション専門企業も多数存在 ・納品からサービスインまでの時間を短縮 ・現地作業員は最小限
ラックへの給電	単相100/200V 入力 ※ 最近は三相の電源も増えてきている	三相208V 入力 •高効率

ラック納品

その他のアメリカデータセンタ特有の点

- **消防署にデータセンタ内のどこでも入れるマスターキーを渡さないといけない**
  - 火災時に消防が建物内に緊急で入れるように
  - ロックがかかっている扉には全て鍵穴がある
- **スプリンクラー消火の為、ネットワーク機器の配置により一層気を遣う**
  - 隣同士で機器を配置すると、局所スプリンクラーでも被害が出る可能性
  - 列などをばらして配置するように
- **データホールの床が若干斜めになっている**
  - スプリンクラー消火の水がはけるように
- **扉は全て中から外へのロックをかけず、押すだけで退出可能**
  - 人命第一
  - 緊急時に閉じ込められないように
- **平屋建て**
  - 納品物が運びやすい
  - データホールが増設されるたびに遠くなる

山火事と煙

山火事 / 煙

- あまりに乾燥しているので周辺地域で**山火事**が発生する
- その山火事による**煙**も発生する

山火事 / 煙

- あまりに乾燥しているので周辺地域で**山火事**が発生する
- その山火事による**煙**も発生する

山火事の影響による道路閉鎖、それに伴う配送遅延
煙の空気汚染による外での工事の停止
煙がDC内に入ることでの煙の誤検知
北はカナダ、南はカルフォルニア州の山火事の煙も流れてくる

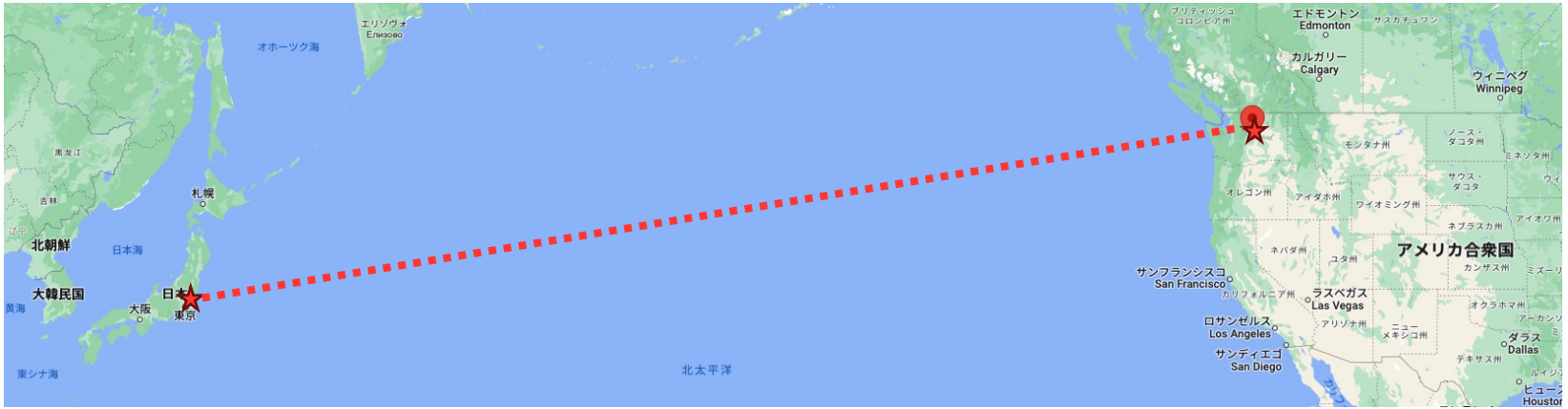
実際の山火事の写真

実際の煙の写真

アメリカデータセンタの
ネットワークの変遷

日米での通信遅延

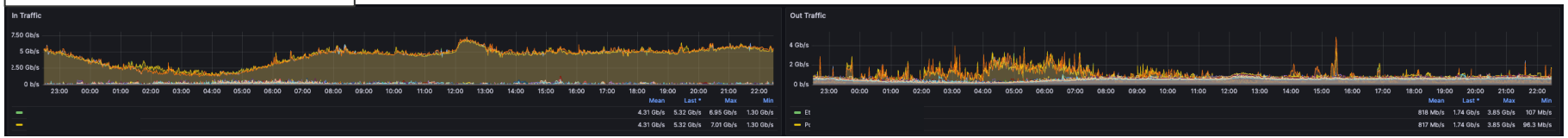
- 直線距離で**約8000kmの物理制約**
- 東京 – アメリカデータセンタ間で**RTT 約100msec**
- パケットの往復が多い**TCP通信の影響は大きい**



ネットワーク遅延に対する対応

- **電力需要が大きくインターネット向けの通信が少ないシステムを中心に収容**
  - データ分析基盤(ビックデータ、機械学習、ディープラーニング)
  - 開発環境
  - バックアップストレージ
- **East-West の通信がメイン**

インターネット向け通信

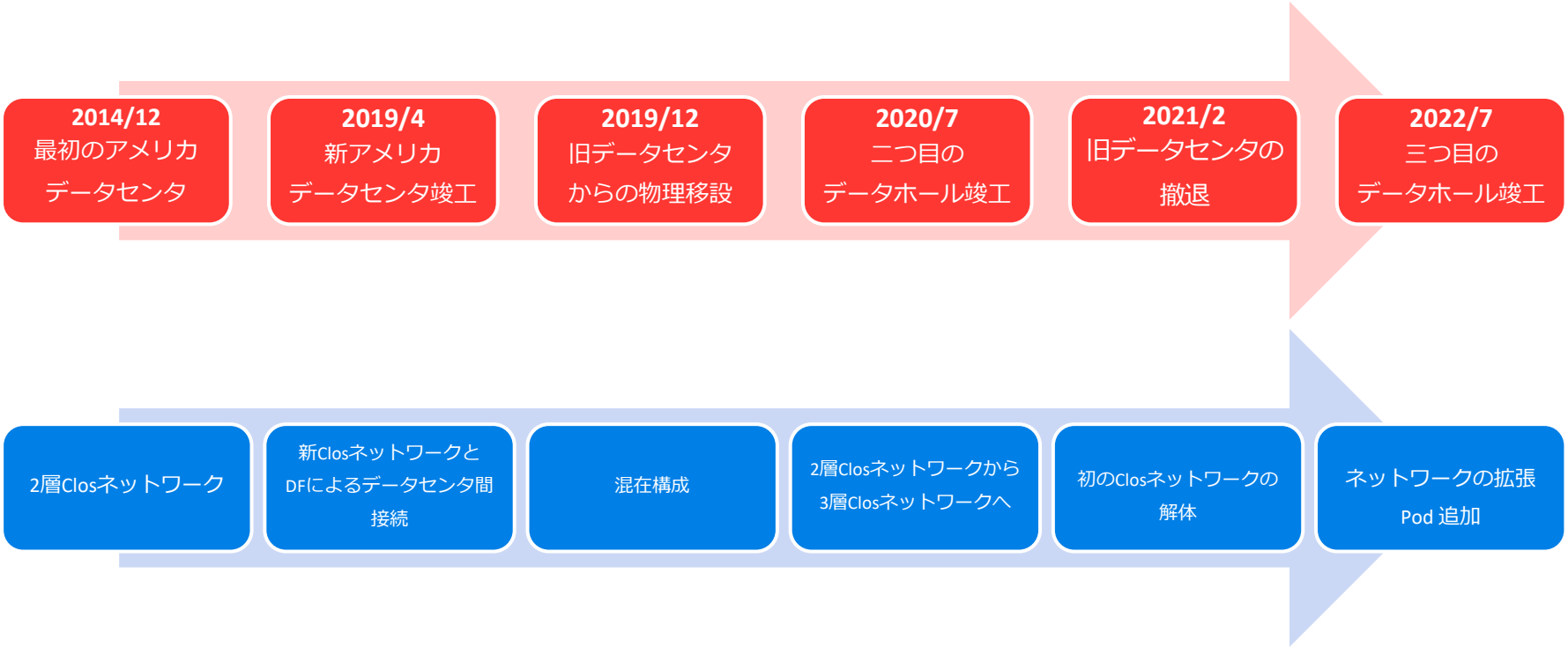


East-West の通信

ネットワークの変遷

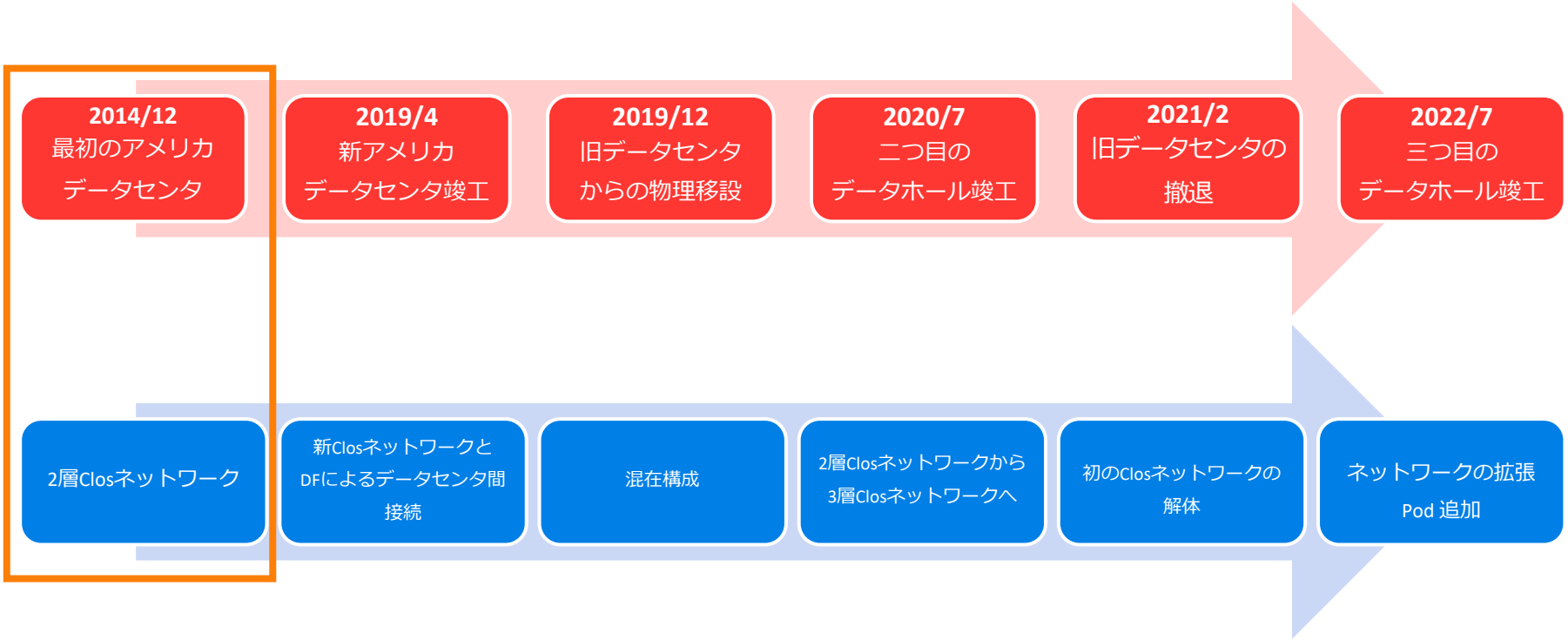
アメリカデータセンタのネットワークの変遷

ネットワークの変遷



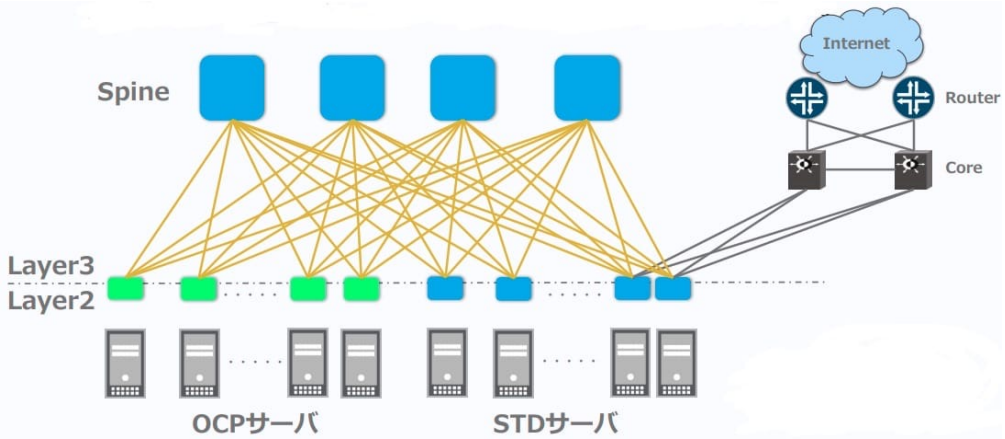
アメリカデータセンタのネットワークの変遷

ネットワークの変遷



最初のアメリカデータのネットワーク構成

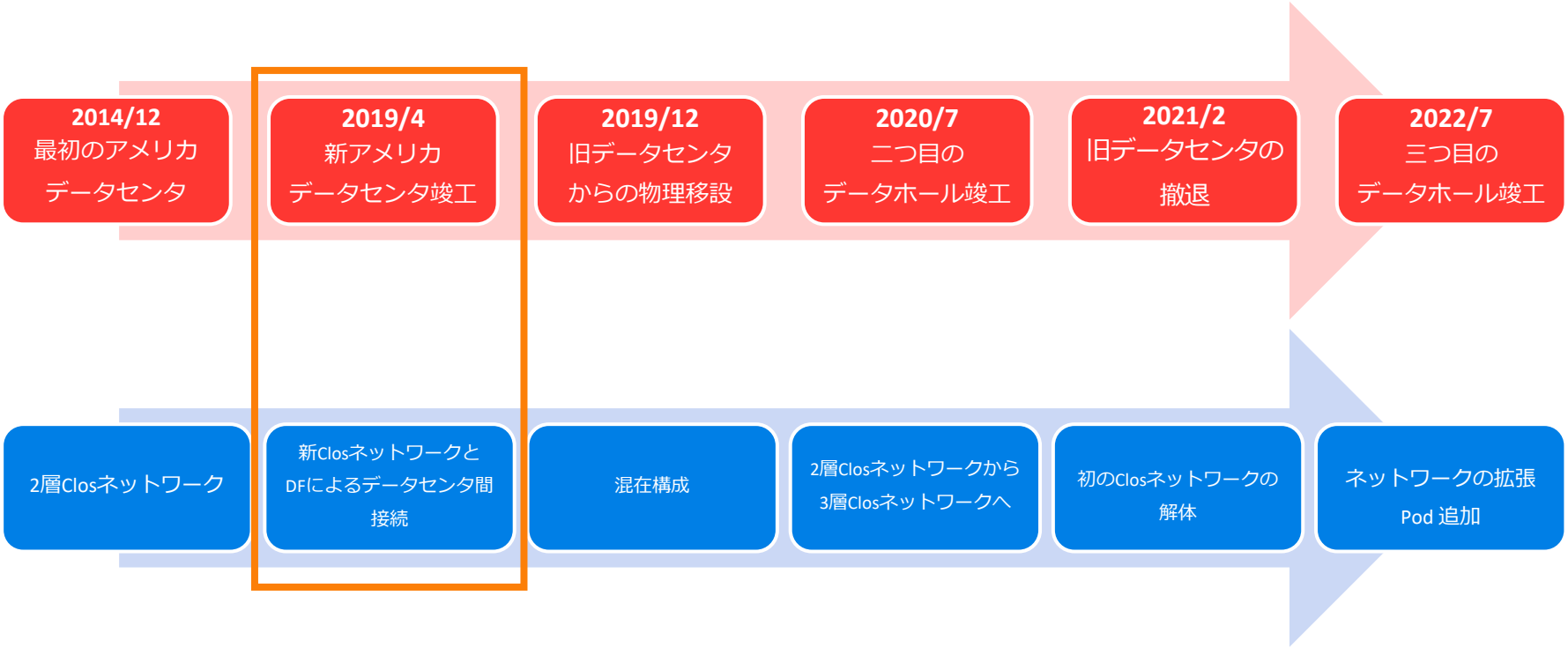
- Spineスイッチにシャーシタイプ、Leafスイッチにネットワークメーカースイッチとホワイトボックススイッチを採用
- Spine/Leaf間は40G-LRを4本
- 社内IPAMからコンフィグ生成+ZTPの本格利用を開始したデプロイ
- サーバは従来のラックマントサーバとOCPサーバを採用
- サーバのNICは10Gbps



https://techblog.yahoo.co.jp/entry/20200323819517/
https://www.janog.gr.jp/meeting/janog46/wp-content/uploads/2020/06/janog46-yahoo-clos-takahashi.pdf

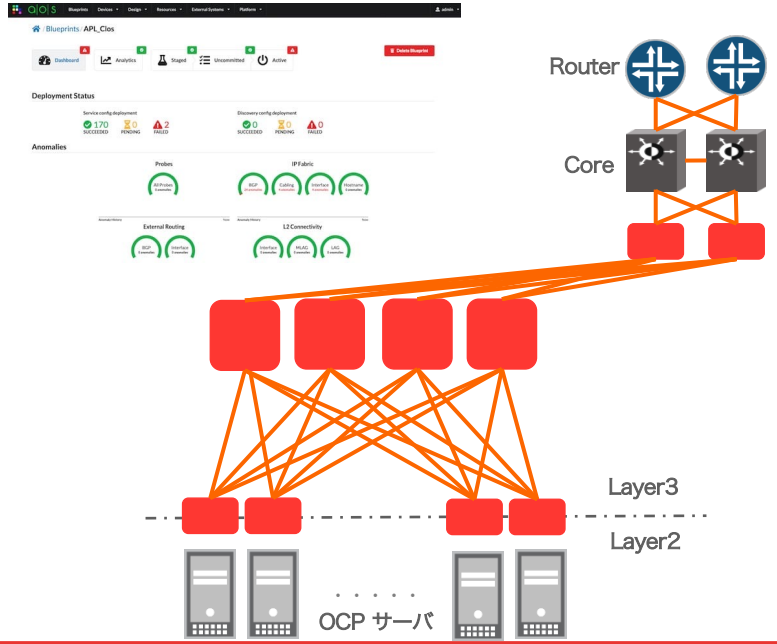
アメリカデータセンタのネットワークの変遷

ネットワークの変遷



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



2019/4 新アメリカデータセンタ竣工

Shared IPMI の導入

- IPMI 用のポート設定には **Dedicated(専用ポート)**と**Shared(NICと共用)** の2種類がある
- Shared(NICと共用) を導入決めた理由
  - **初期投資費用削減**
  - **スイッチ運用コスト低減**
  - **減らしたスイッチ分の電力をサーバにまわせる**
  - **ケーブル削減による現地のオペ効率化**

※ IPMI での大規模サーバー管理
https://techblog.yahoo.co.jp/operation/2014-infra-ipmi/

2019/4 新アメリカデータセンタ竣工

Shared IPMI の導入

Dedicated (専用ポート)

Shared (NICと共用)

2019/4 新アメリカデータセンタ竣工

Shared IPMI の導入

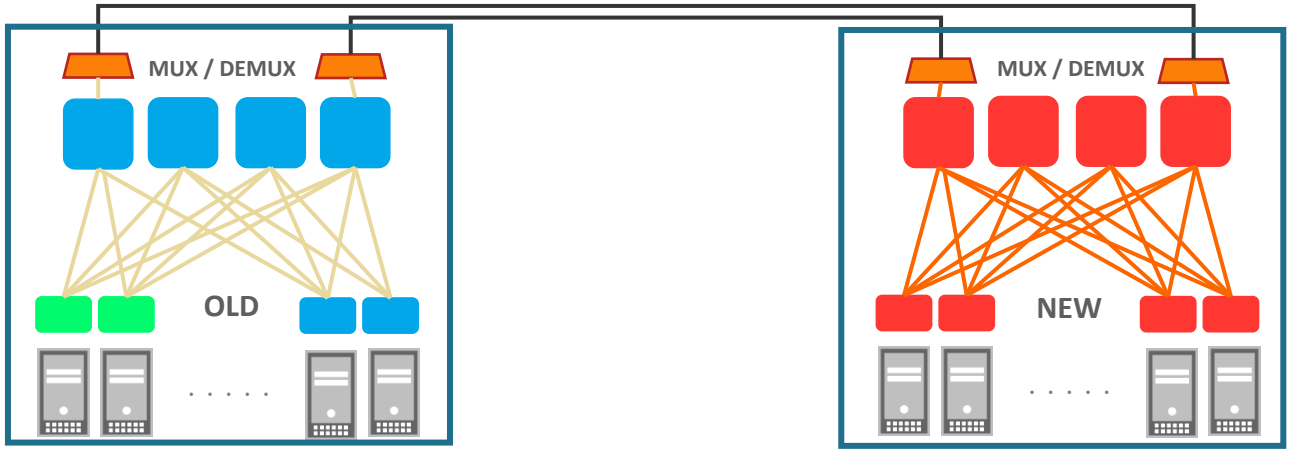
Dedicated (専用ポート)

Shared (NICと共用)

ケーブルもスイッチも減ったことでスッキリ

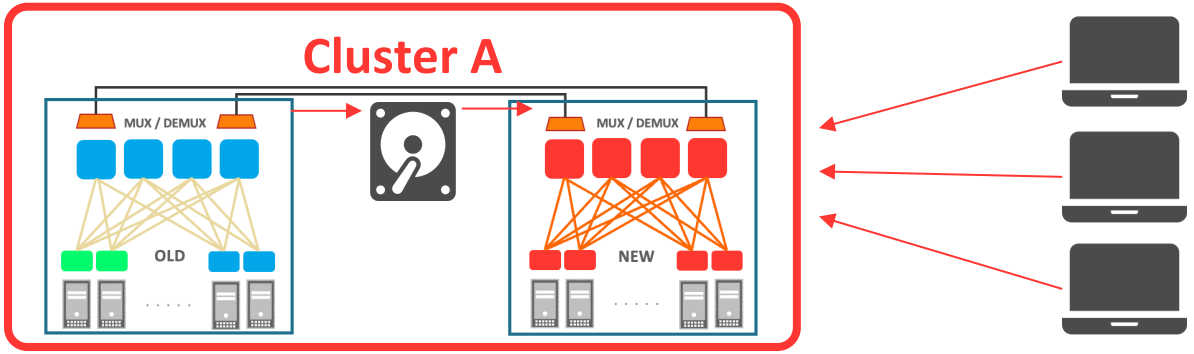
新旧アメリカデータセンタ間の接続

- 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
  - **1.2Tbps** 冗長



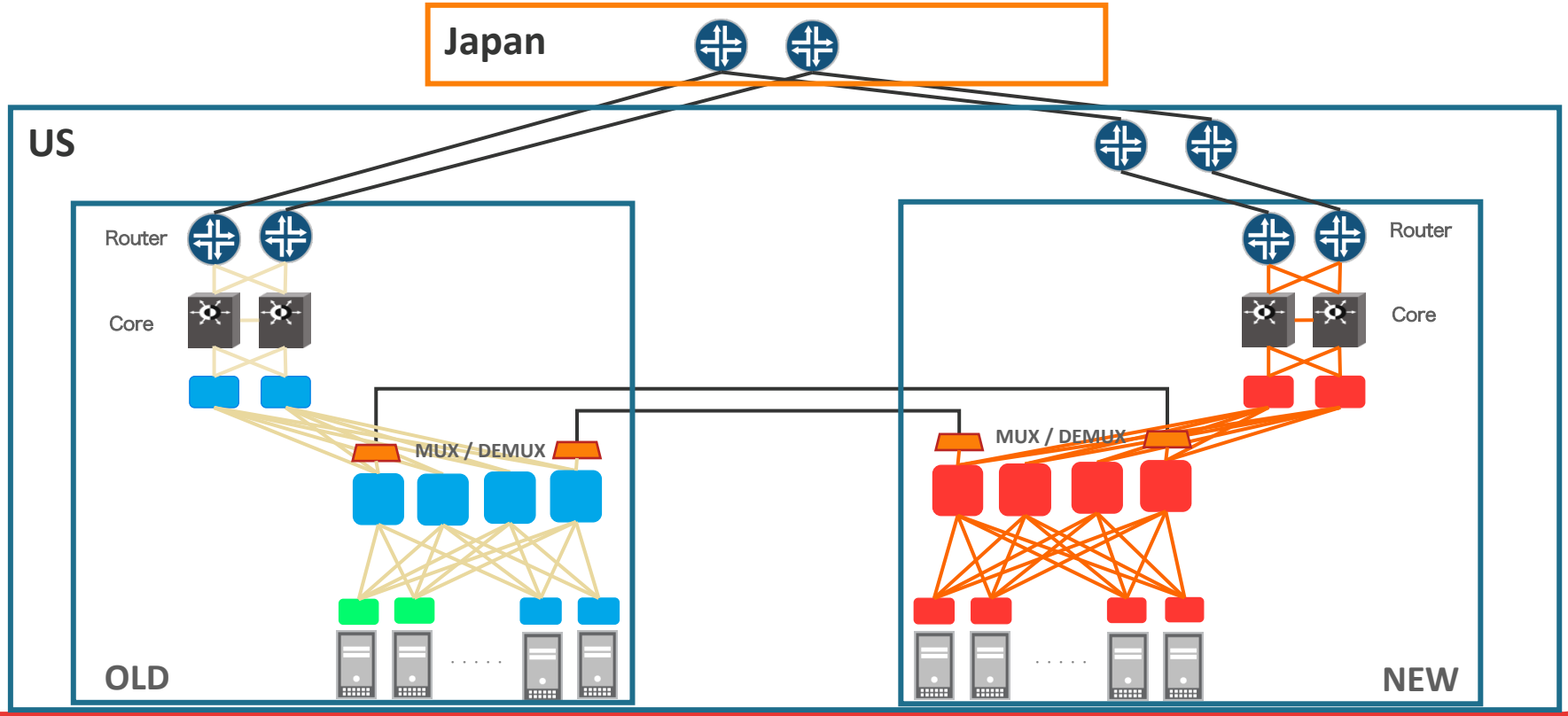
ダークファイバでのデータセンタ間接続の理由

- **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



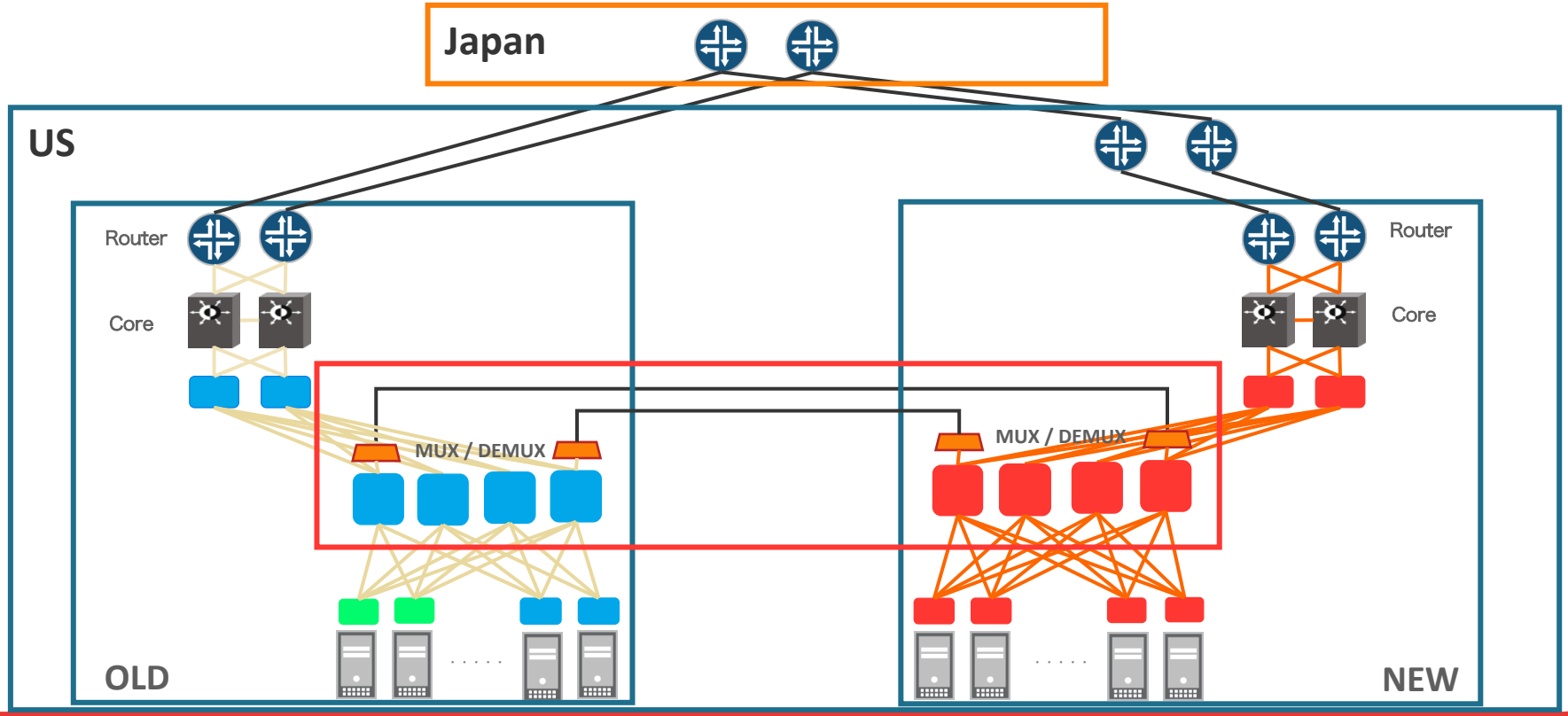
2019/4 新アメリカデータセンタ竣工

日米回線含めたデータセンタ間の構成

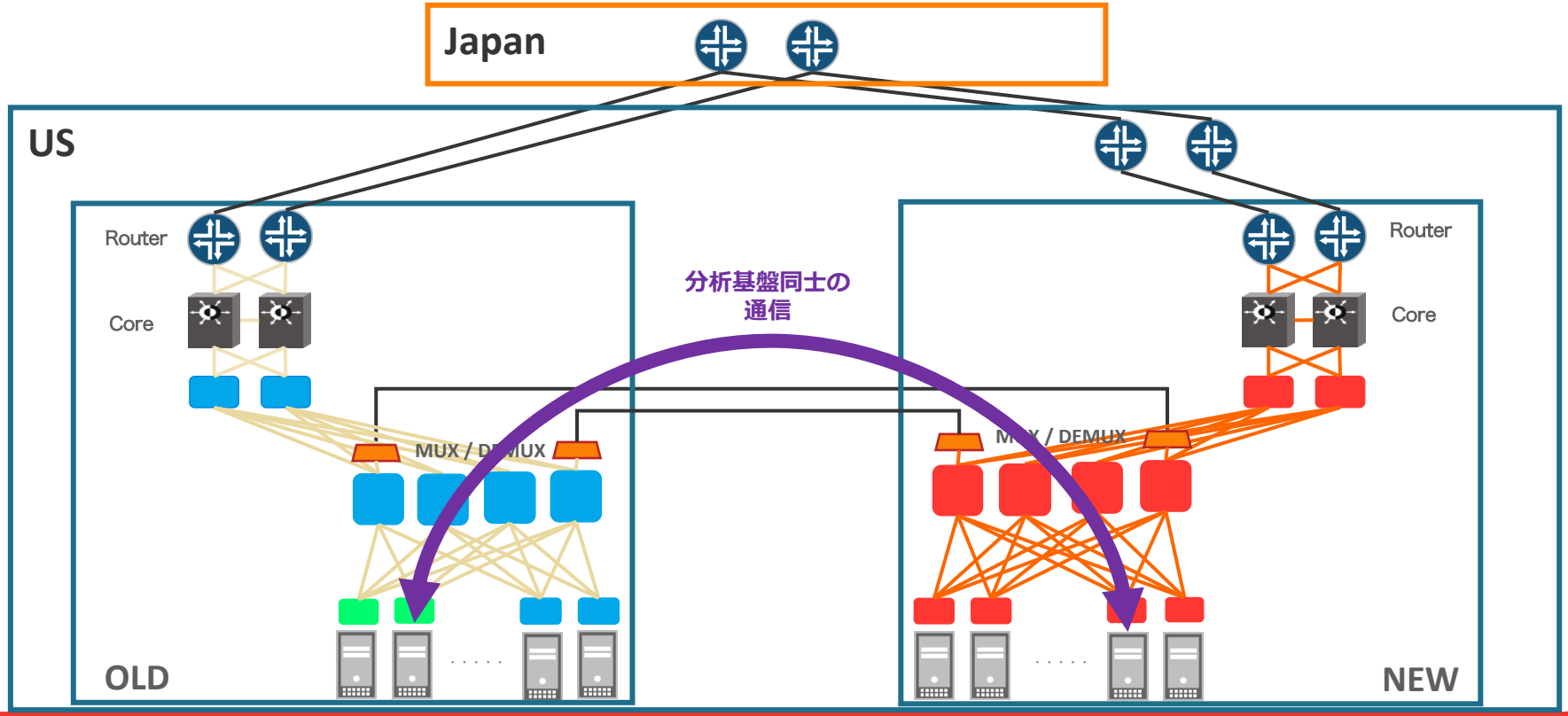


2019/4 新アメリカデータセンタ竣工

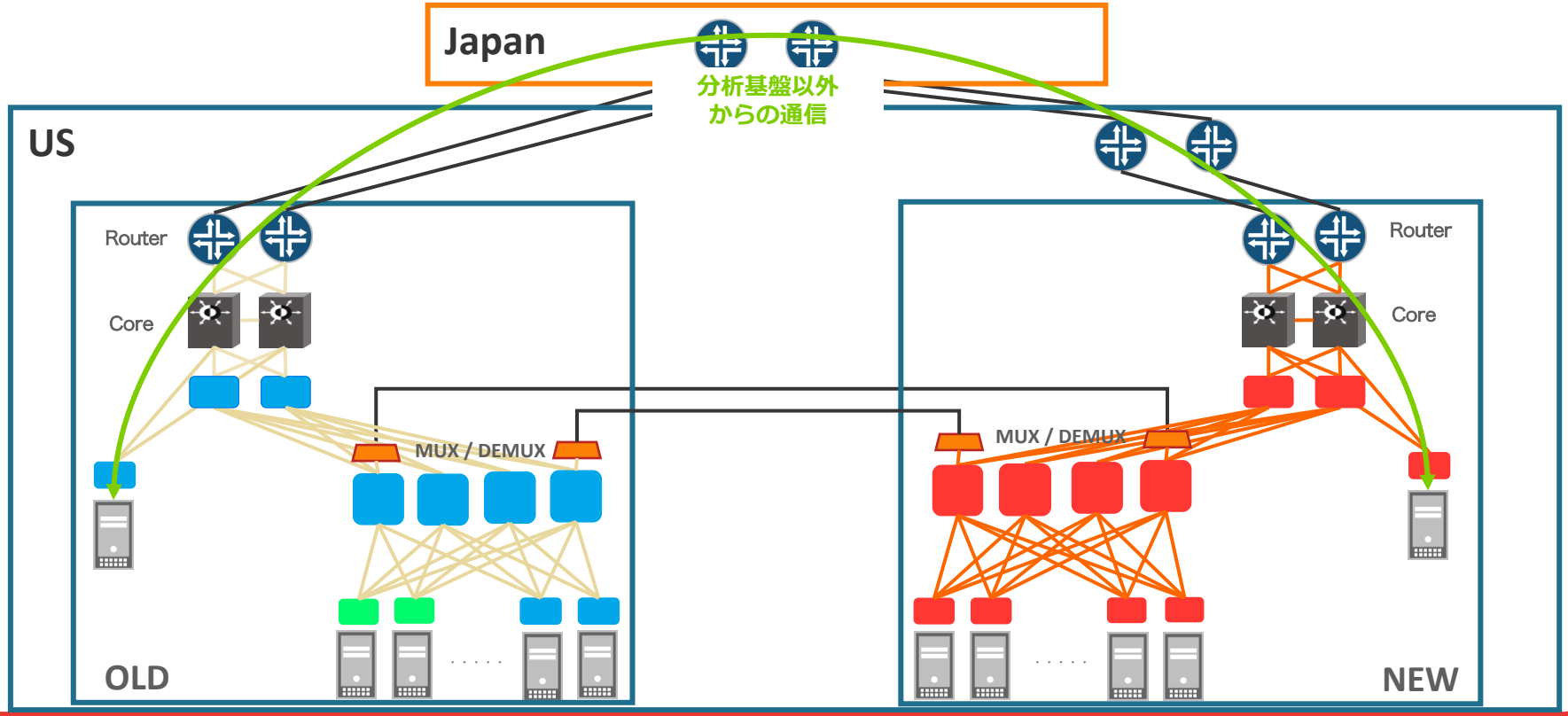
日米回線含めたデータセンタ間の構成



日米回線含めたデータセンタ間の構成

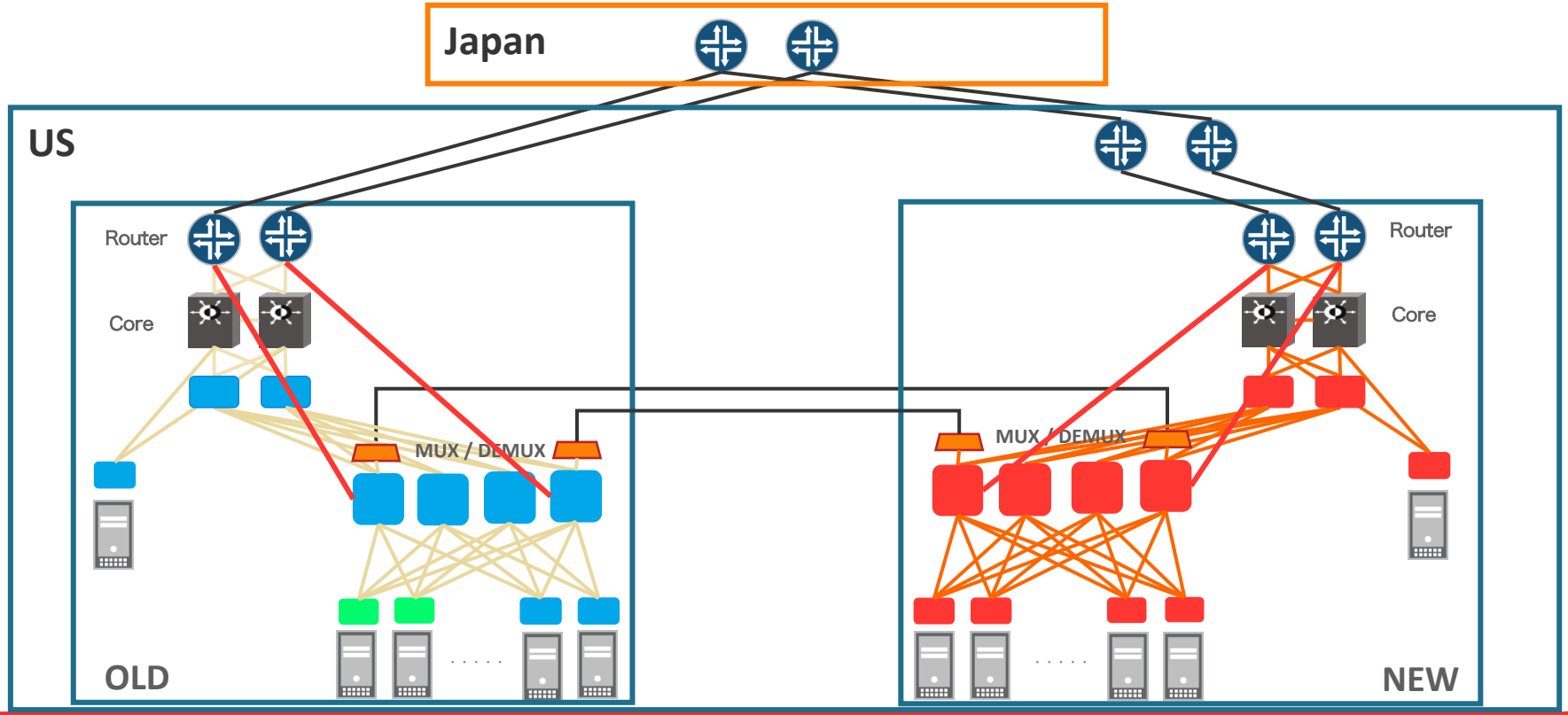


日米回線含めたデータセンタ間の構成



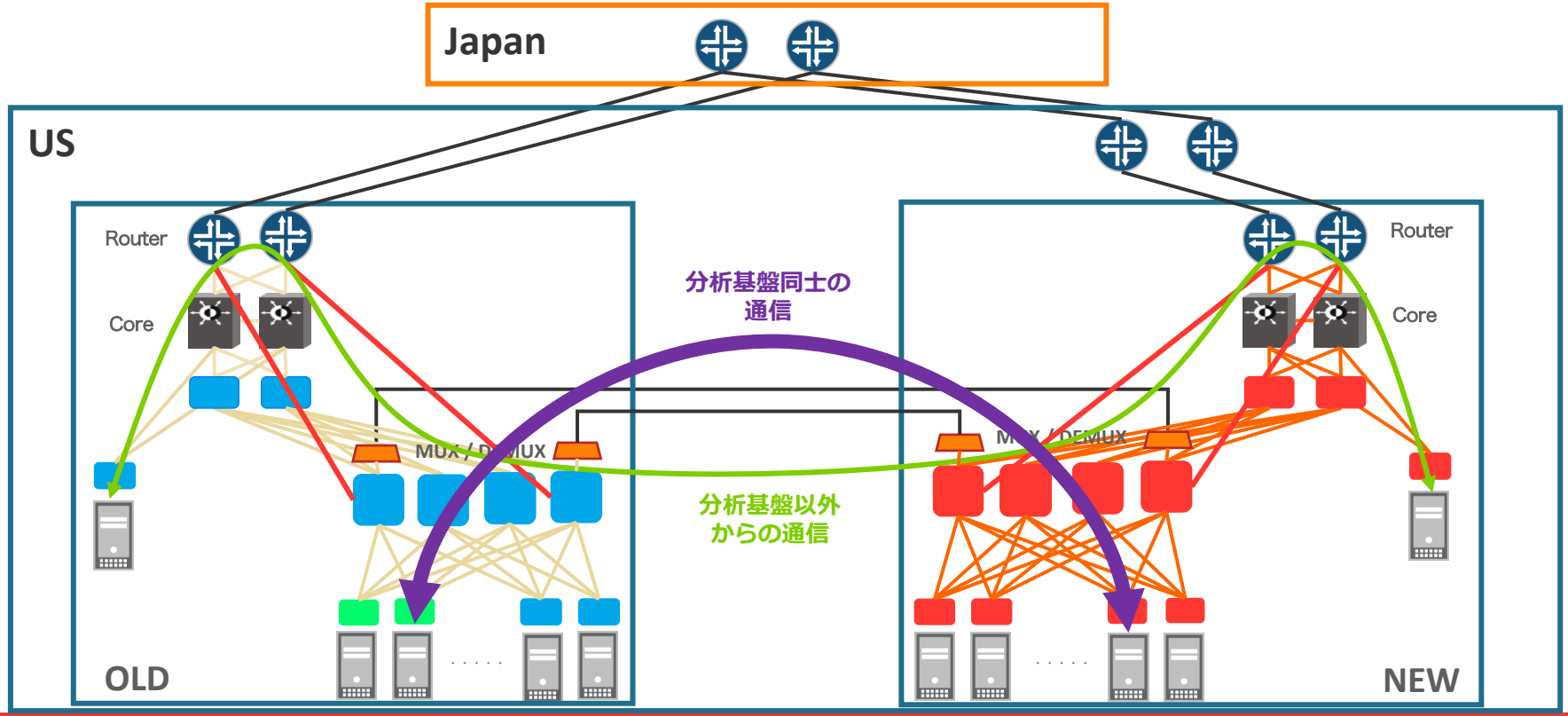
2019/4 新アメリカデータセンタ竣工

日米回線含めた旧データセンタ含めた構成変更



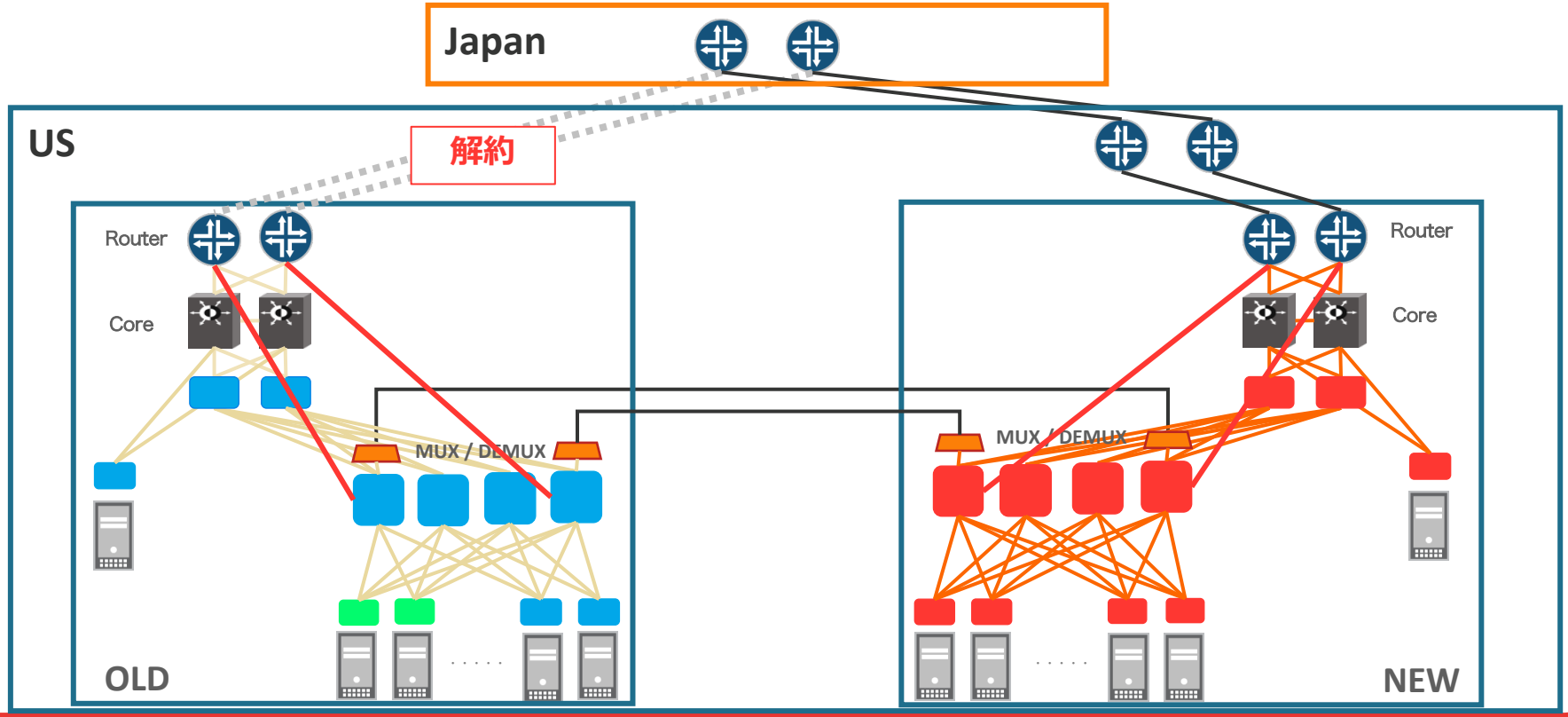
2019/4 新アメリカデータセンタ竣工

日米回線含めた旧データセンタ含めた構成変更



2019/4 新アメリカデータセンタ竣工

日米回線含めた旧データセンタ含めた構成変更



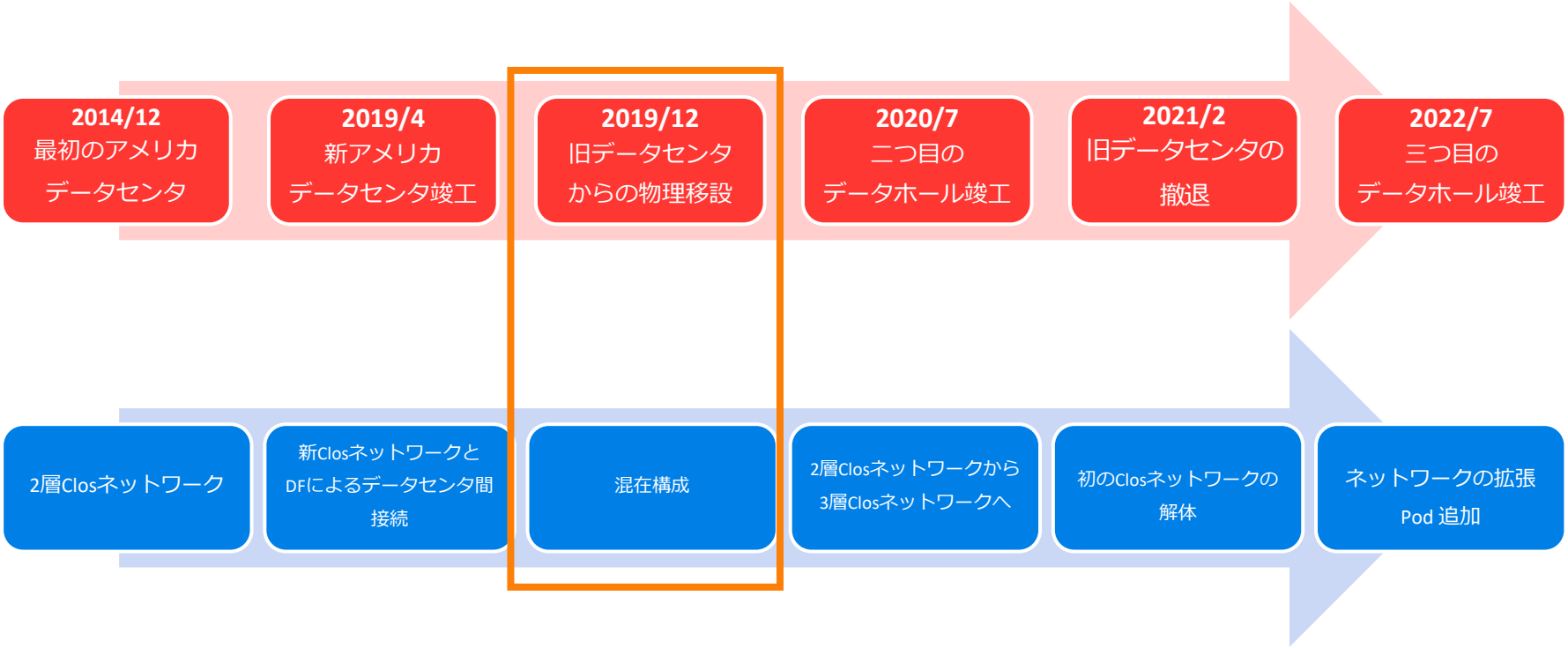
新旧アメリカデータセンタでのネットワークの違い(物理)

	Old	New
構内配線	Multi / Single / UTP	Single / UTP
配線経路	基本床下配線 ネットワークラックのみラック上配線	ラック上配線
ネットワークラック	パッチパネル含め全て2ポスト	基本4ポスト パッチパネルのみ2ポスト
コンソール	ロールオーバ	ストレートケーブル

2019/4 新アメリカデータセンタ竣工

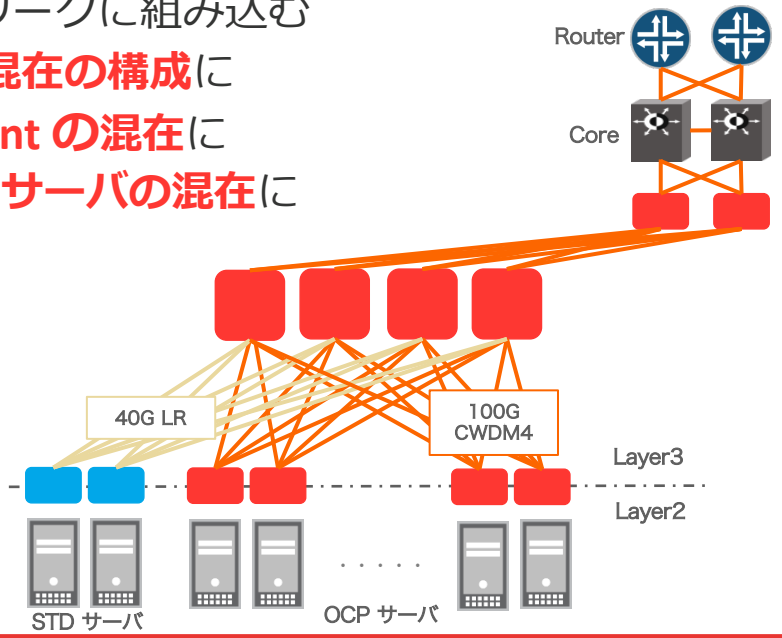
吊り下げラック

ネットワークの変遷



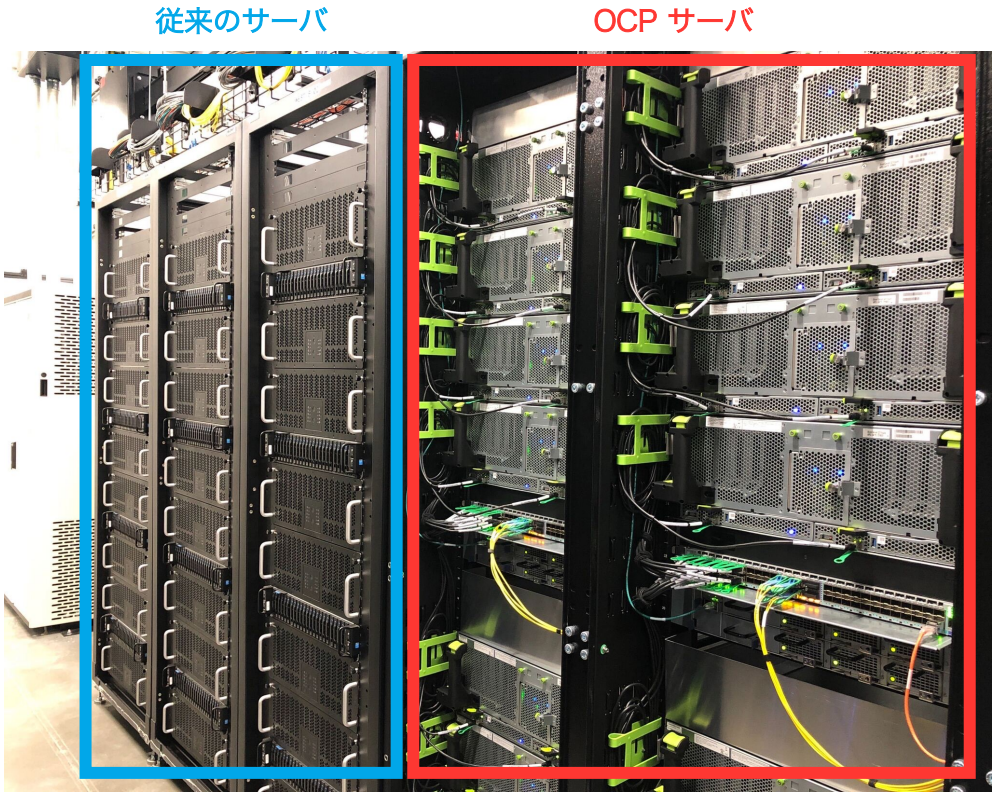
混在構成

- 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
- 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
- サーバも**従来のラックマウントサーバとOCPサーバの混在**に



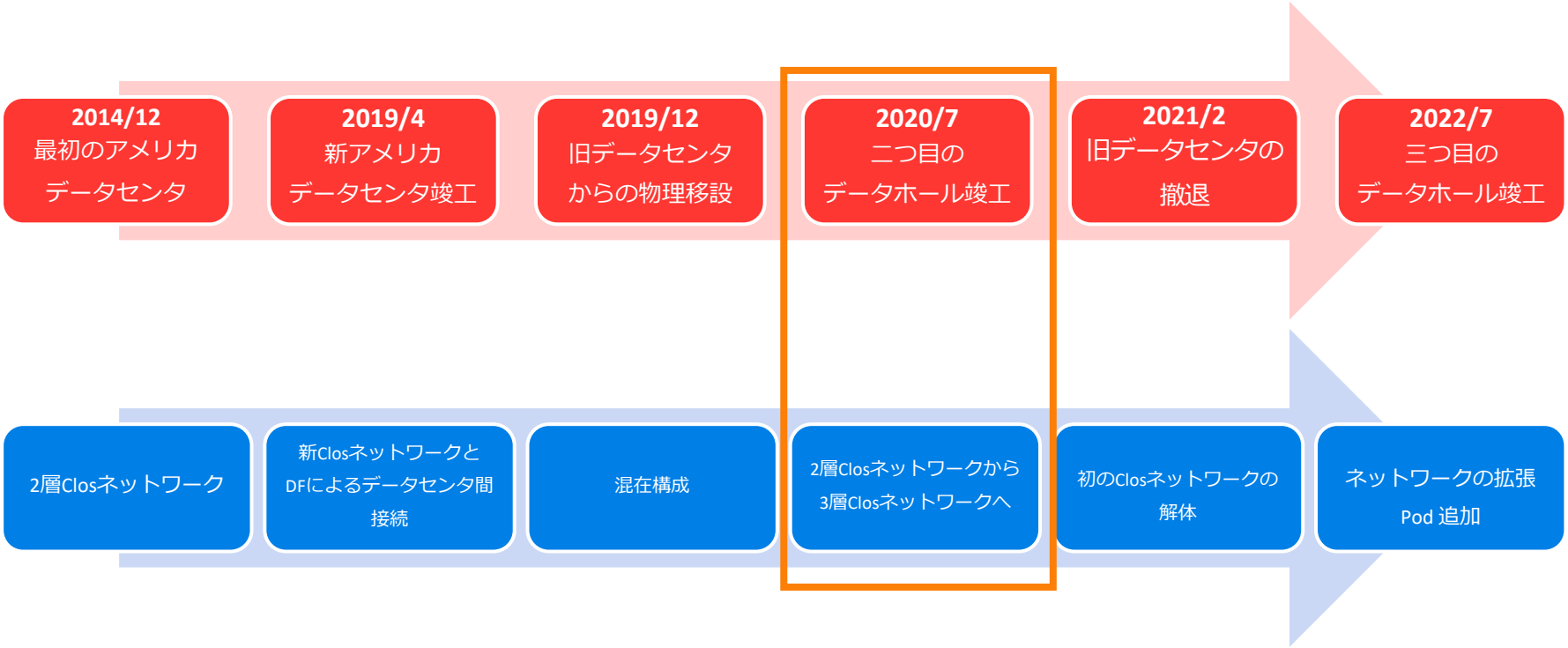
2019/12 旧データセンタからの物理移設

ラックを物理的に移動



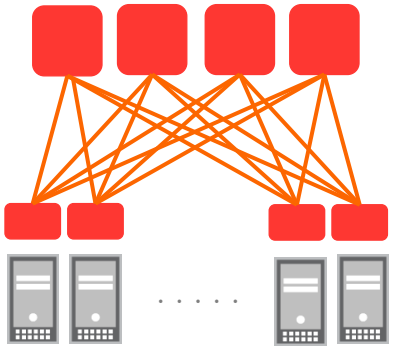
アメリカデータセンタのネットワークの変遷

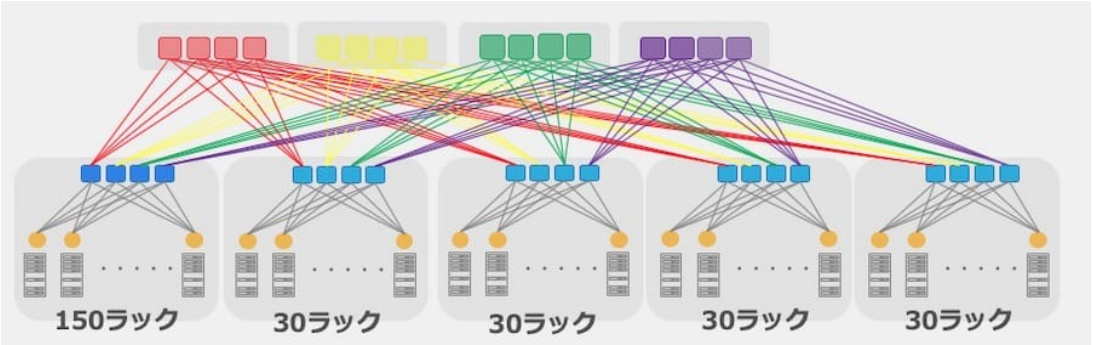
ネットワークの変遷



Clos ネットワークの構成の変更と配線作業

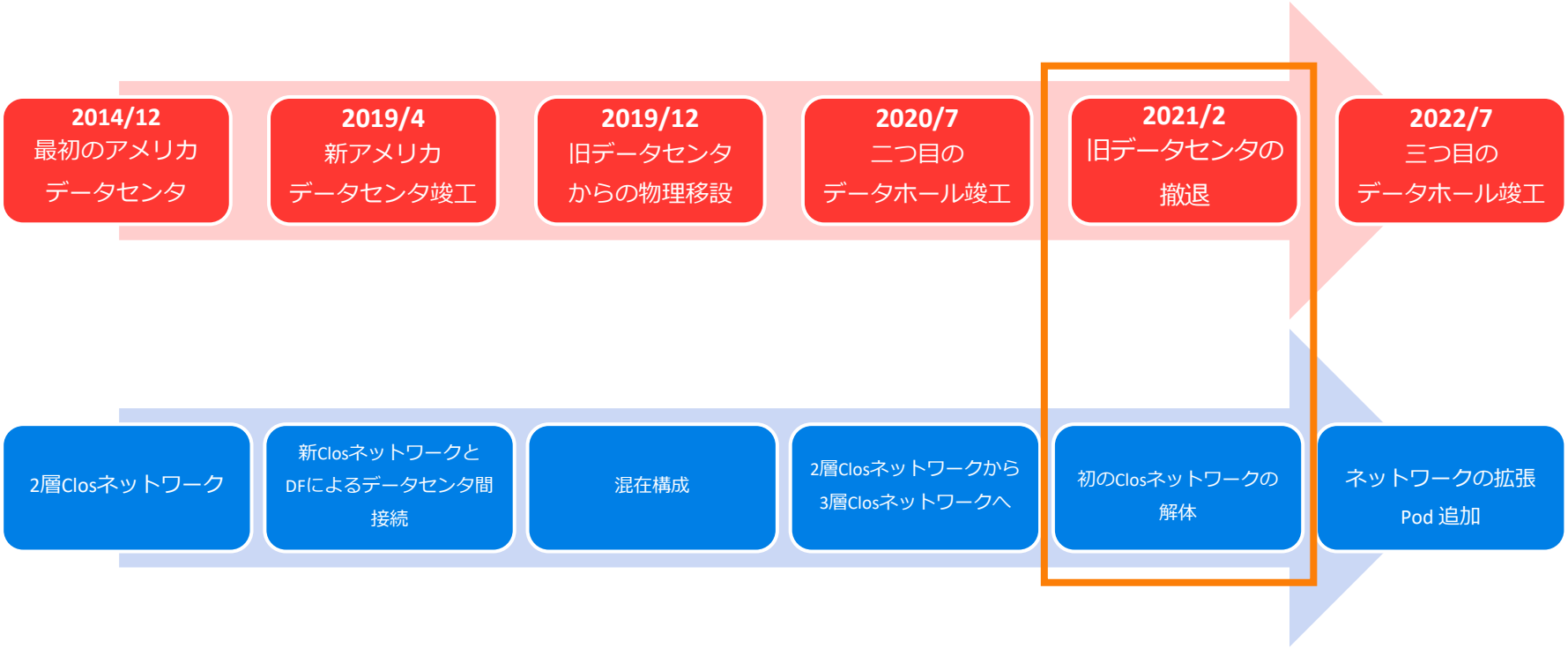
- **二つ目のデータホール竣工**に伴い、Closネットワークも拡張のため、構成変更が必要となった
  - **2層構成**では収容しきれなくなるため、**3層構成**へ変更





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

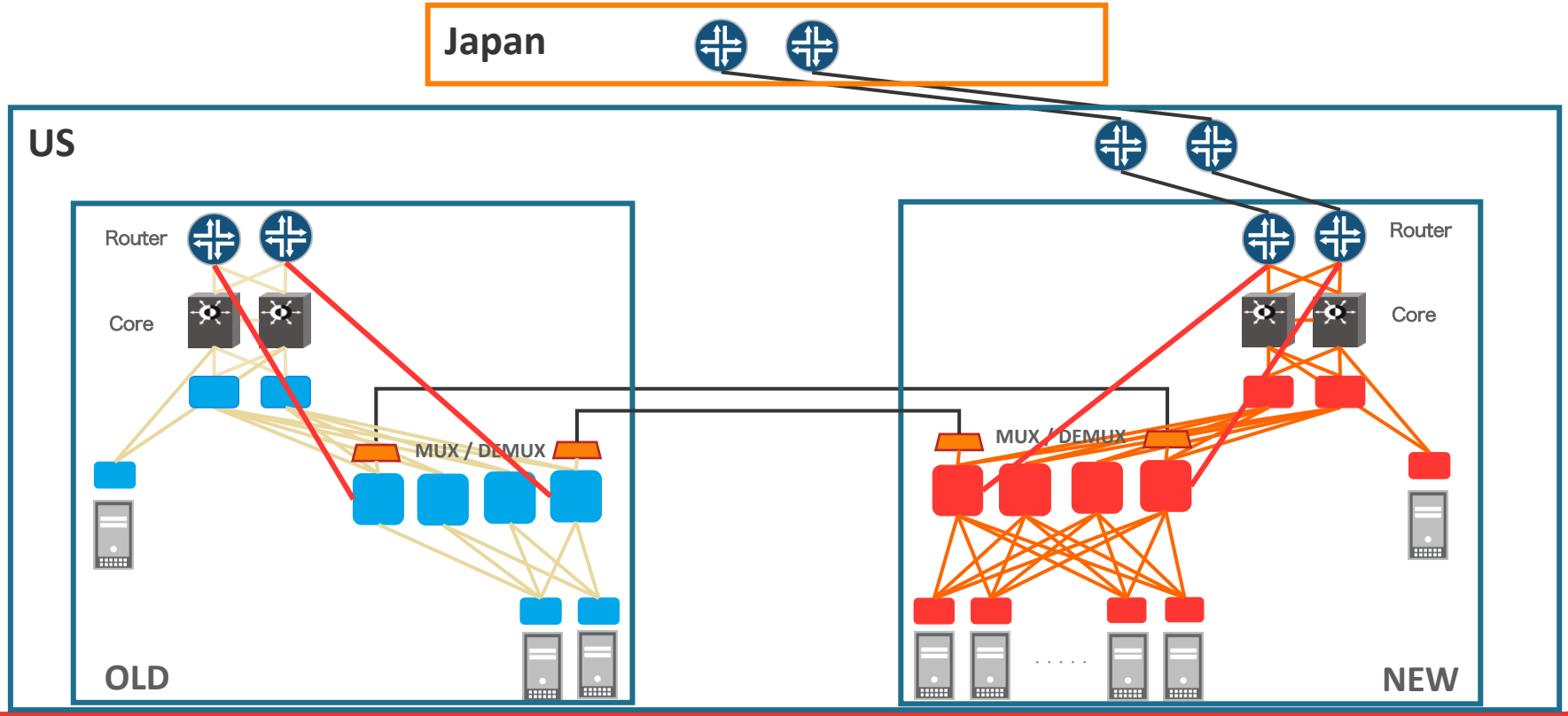
ネットワークの変遷



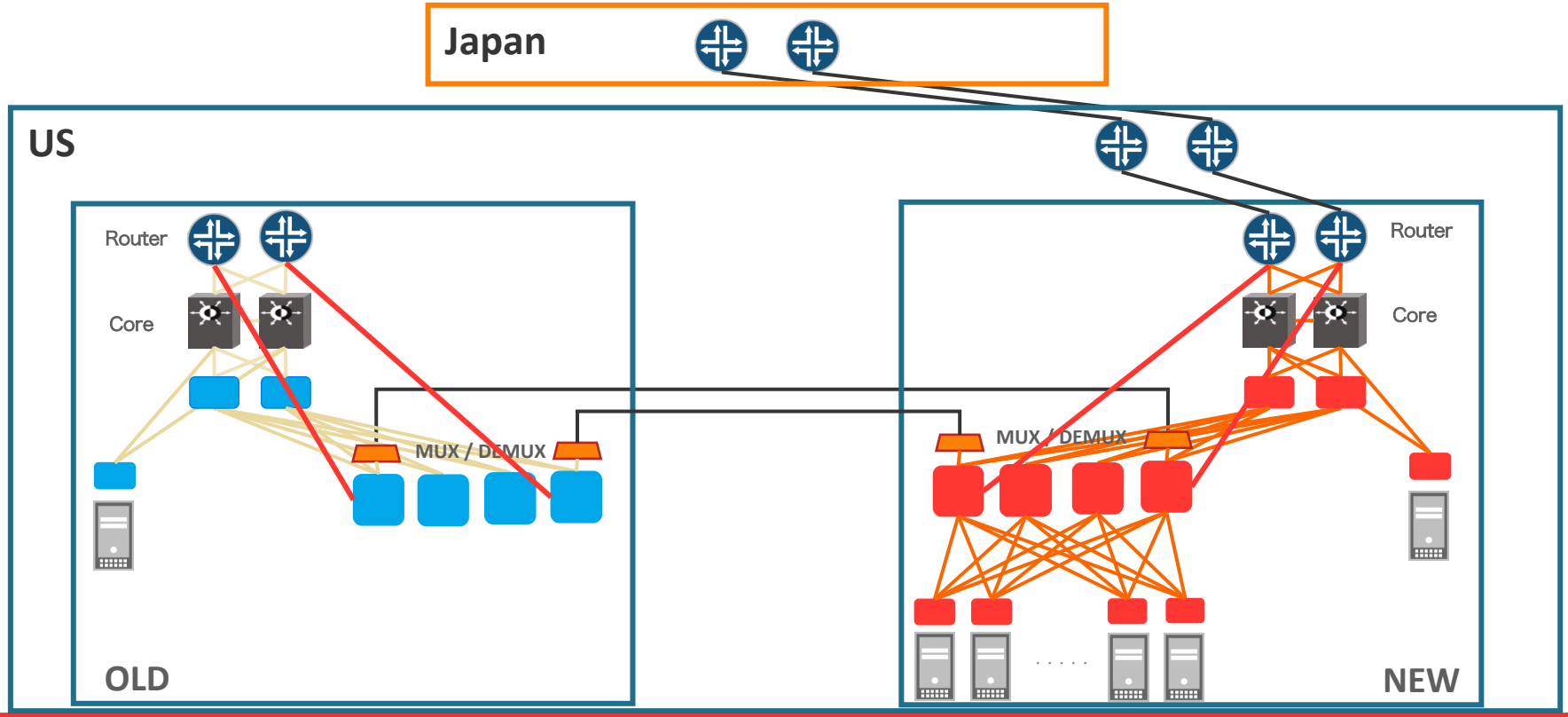
撤退時の考慮

- 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
- **作業は割とシンプル**
  - 使わなくなったラックの電源を落とす
  - ネットワーク毎に広報を止めると言ったことはせず、
    下から順番に落としていき、最後は上流スイッチを落とす
- **落とす順番を間違えないように注意**
  - DWDMを搭載したSpineスイッチは最後に落とす
  - DNS や NTP といった運用系のサーバも最後

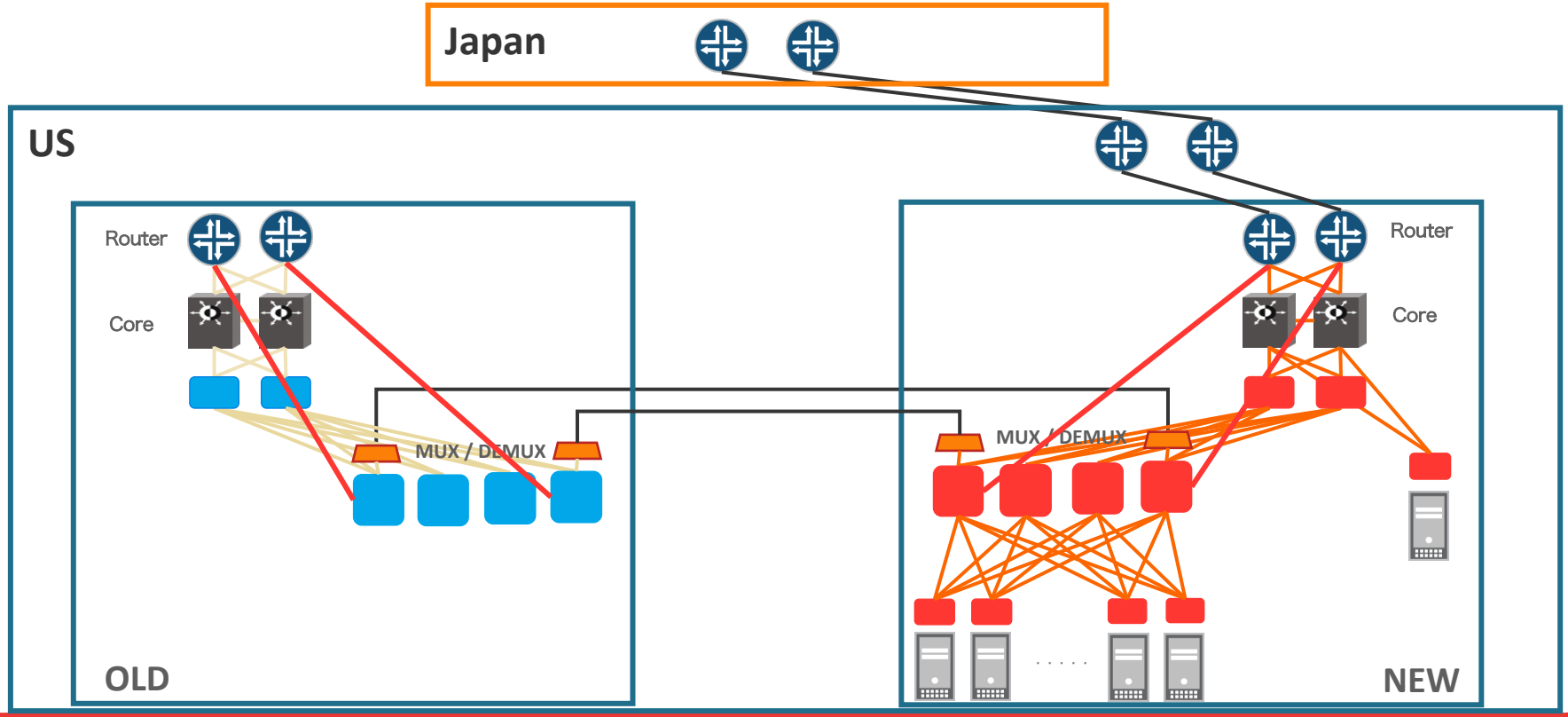
撤退前の構成



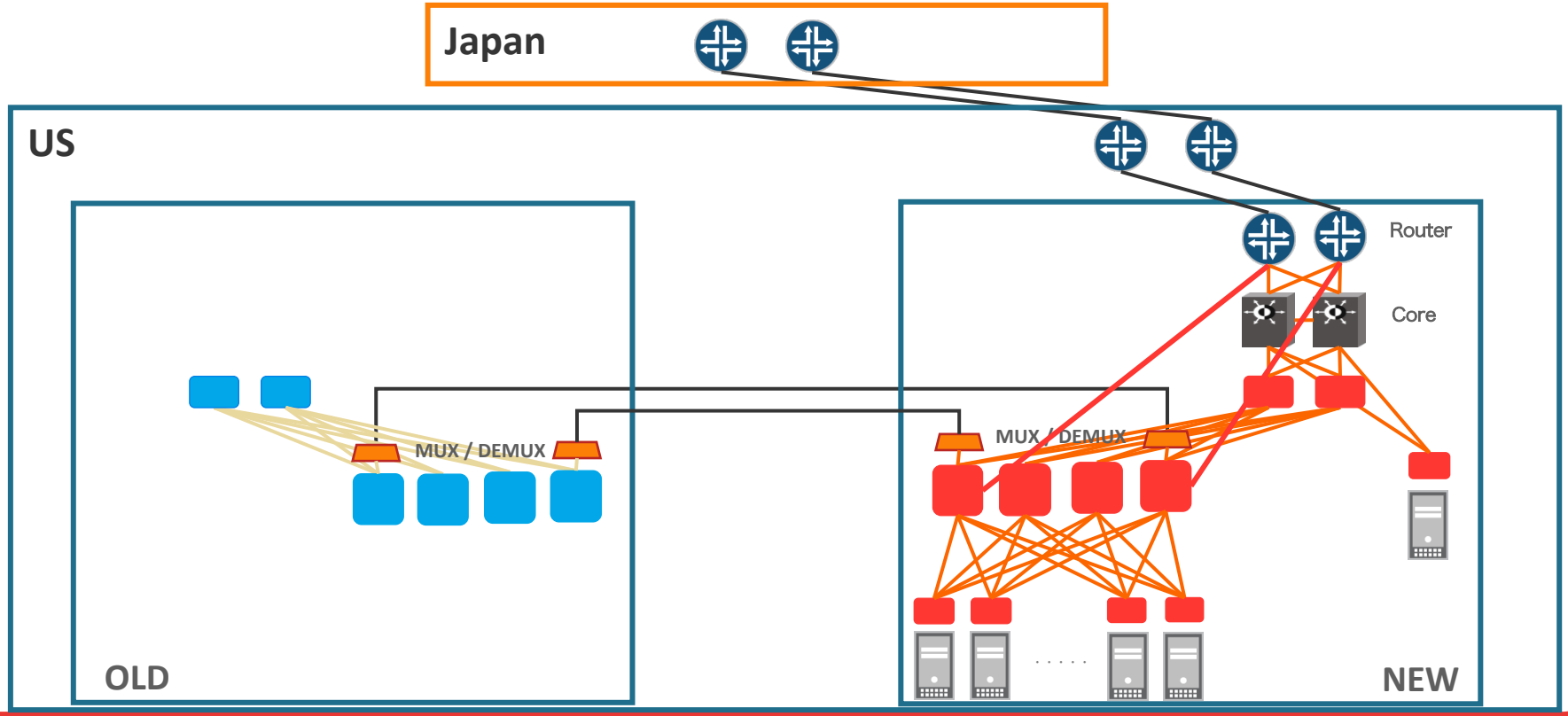
Clos ネットワーク Leafの撤退



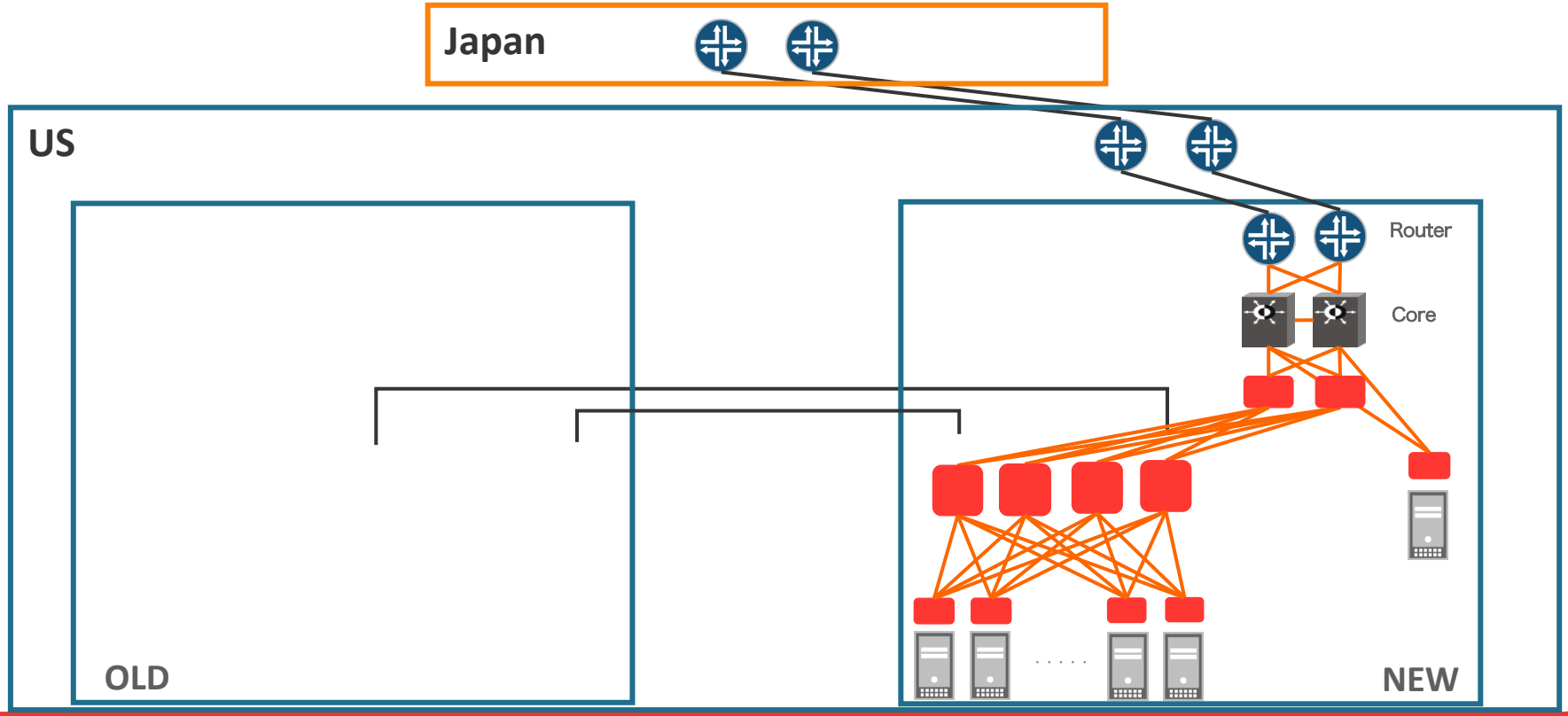
ToR の撤退



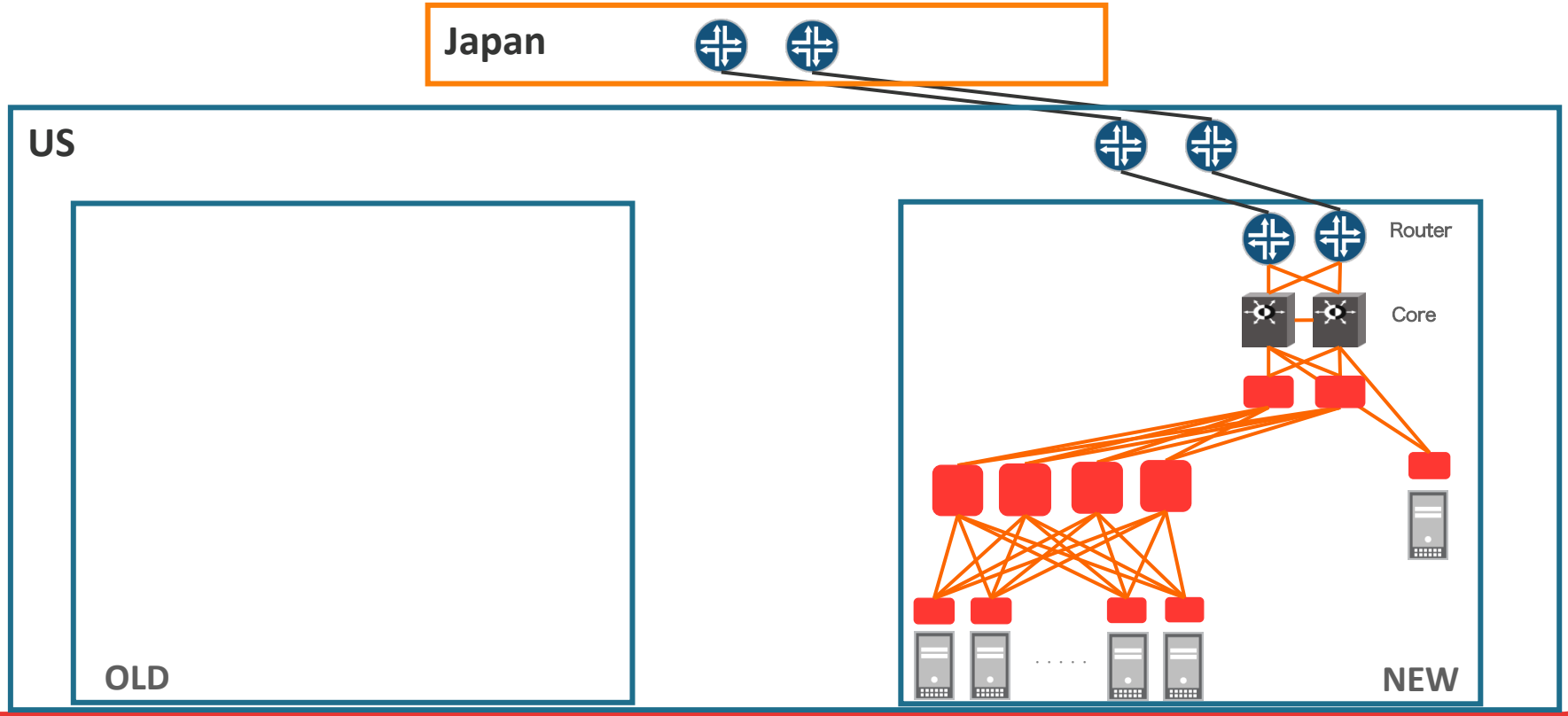
ルータ・コアの撤退



SpineとDWDMの撤退



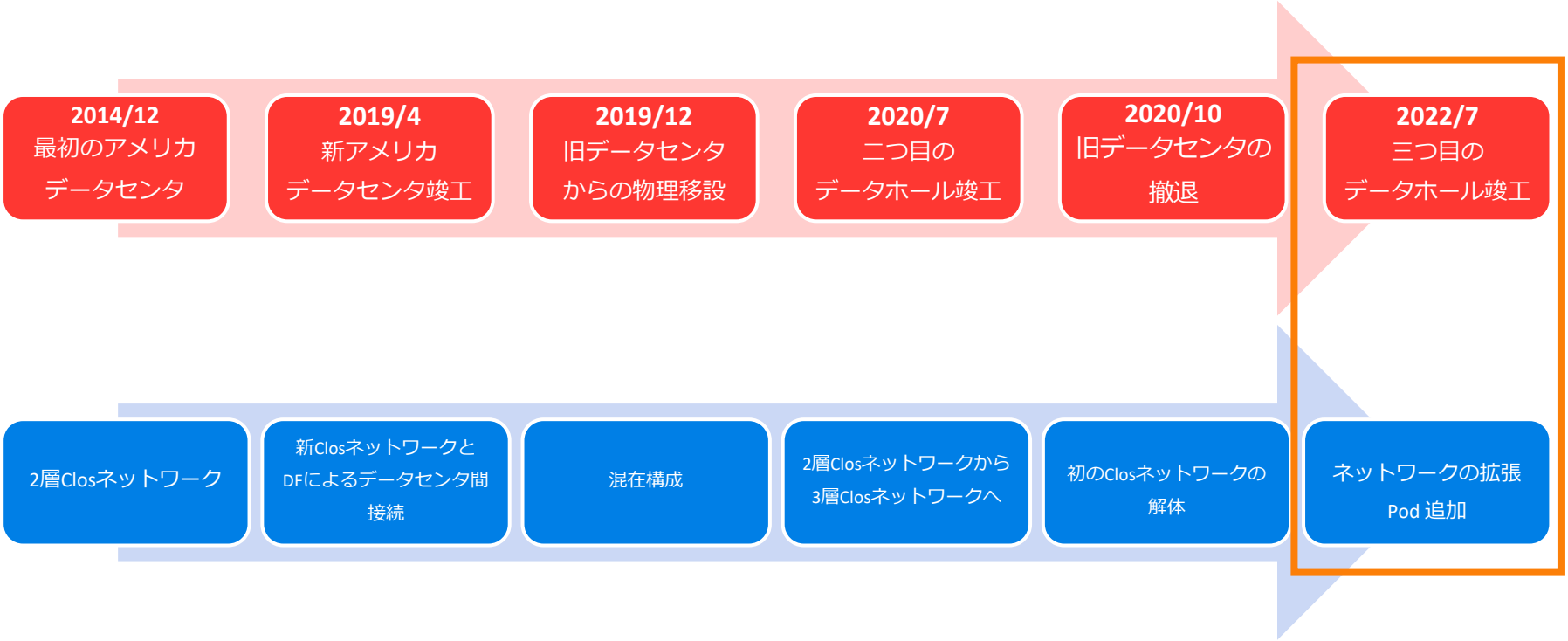
ダークファイバの解約



2021/2 旧データセンタの撤退

無事撤退

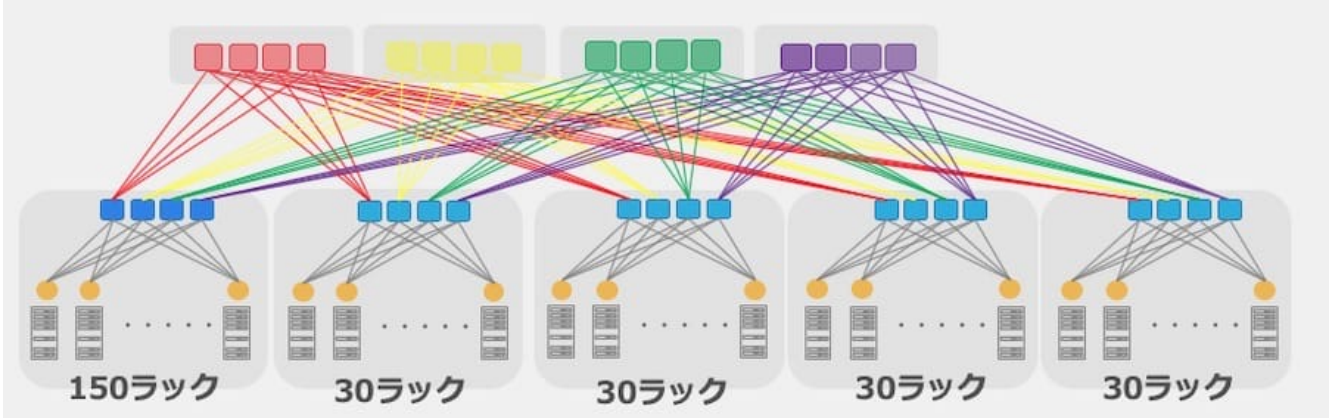
ネットワークの変遷

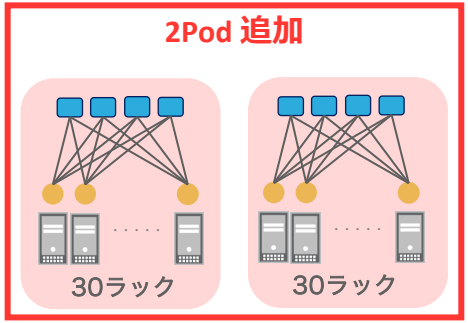


2022/7 三つ目のデータホール竣工

ネットワークの拡張 Pod 追加

- 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2Podを追加





ネットワーク障害

BMC firmware アップデート時の作業ミス

- オペレーションミスで誤った対象にBMC firmwareアップデート作業を実施
  - **誤った対象に誤ったBMCを焼いてしまう**
- BMC が中途半端に起動しては落ちるを繰り返すことに

BMC firmware アップデート時の作業ミス

- オペレーションミスで誤った対象にBMC firmwareアップデート作業を実施
  - **誤った対象に誤ったBMCを焼いてしまう**
- BMC が中途半端に起動しては落ちるを繰り返すことに

その再起動の繰り返しがネットワークのChipも巻き込んで起こってしまい、NIC の Down/Up も発生

BMC firmware アップデート時の作業ミス

- オペレーションミスで誤った対象にBMC firmwareアップデート作業を実施
  - **誤った対象に誤ったBMCを焼いてしまう**
- BMC が中途半端に起動しては落ちるを繰り返すことに

その再起動の繰り返しがネットワークのChipも巻き込んで起こってしまい、NIC の Down/Up も発生

サーバへ疎通ができない状態に

BMC firmware アップデート時の作業ミス

- 電源抜き差しをしても、BMCが再起動を繰り返し続けるだけで改善せず
- BMCのチップを取り出して直接焼き直す方法も試してみるも、改善したが別の問題が

BMC firmware アップデート時の作業ミス

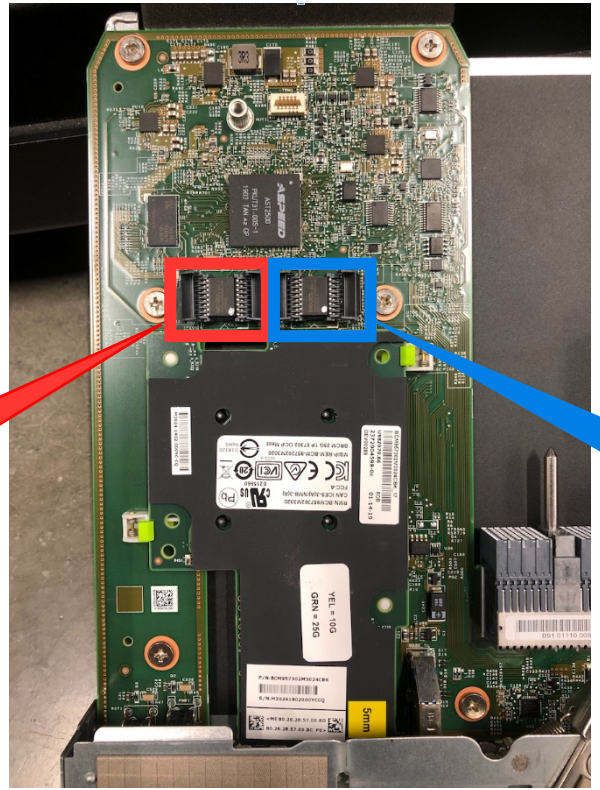
- 電源抜き差しをしても、BMCが再起動を繰り返し続けるだけで改善せず
- BMCのチップを取り出して直接焼き直す方法も試してみるも、改善したが別の問題が

1Chipあたり5分かかってしまう
しかも、1Chipずつしか焼き直せない
作業対象が数100台レベルであり、復旧に時間がかかる

BMC firmware アップデート時の作業ミス

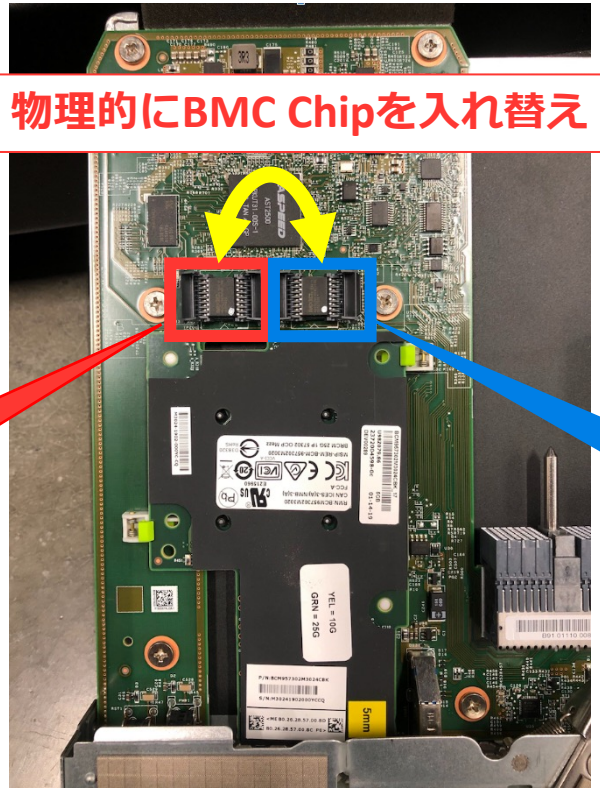
1つのサーバにBMC Chipが2枚搭載されてるけど、
1枚は前のバージョンのまま無事なのでは？

解決策



解決策

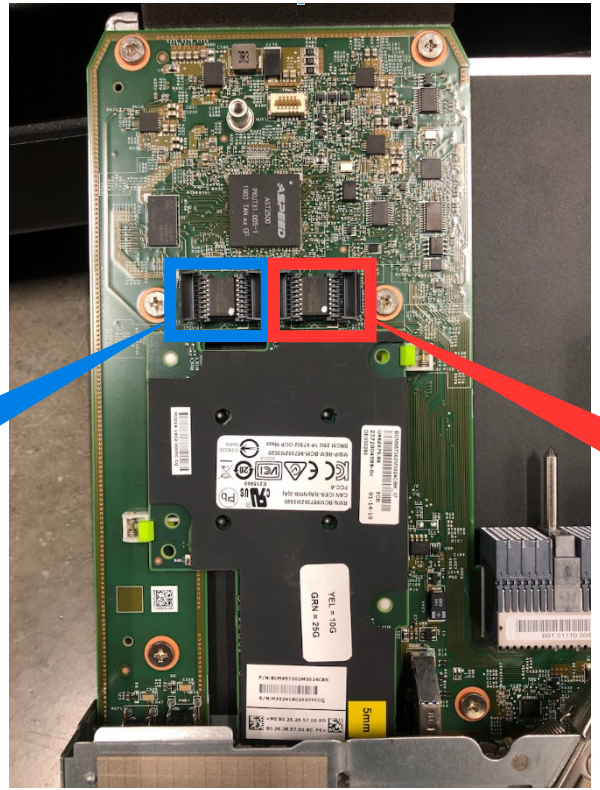
物理的にBMC Chipを入れ替え



誤ったfirmwareが
入っているBMC Chip

作業前のfirmwareが
入っているBMC Chip

解決策



解決策

誤ったBMC firmwareを焼いてしまった全サーバの
BMC Chipを入れ替えることで復旧

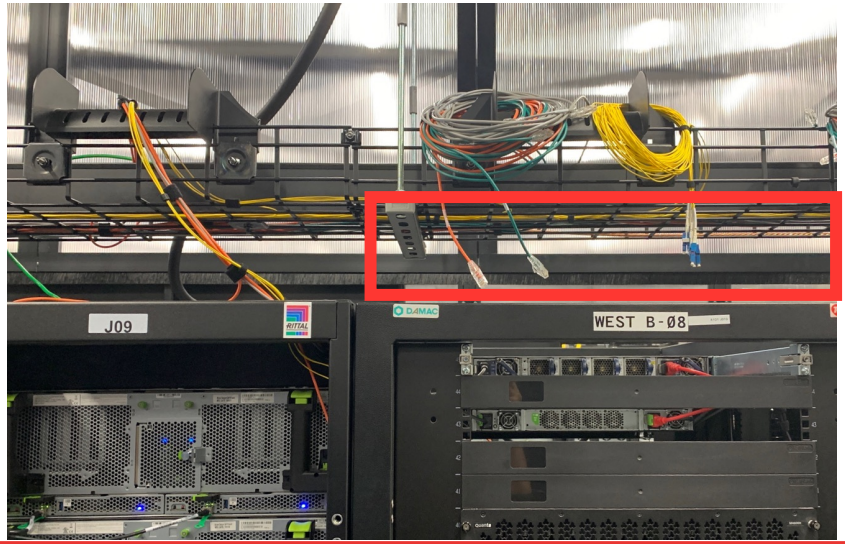
旧データセンタからの物理移設時の苦労

ラック丸ごとの移動

- 床打ち付けのラック固定をしていないため、**ラック丸ごとの移動**が可能
  - 輸送のためにサーバやスイッチ、ラックPDUを取り外す作業はなし
- 配送時の振動で、**筐体を固定するネジが緩んでしまう**といった事象あり
- 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)

ネットワークの機種の混在

- コールドアイル側(Front)側にしかケーブルを下ろす想定をしていなかったため、**普段とは違う経路**でラック外からのケーブルを通さなくてはいけなくなった
- ネットワークの作業で**ホットアイル(Rear)側**に行かないといけなくなった
  - コールドアイル側での作業になれてしまった人間には辛い



コロナ禍でのネットワーク拡張の苦労

2020/7 二つ目のデータホール竣工

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

2020/7 二つ目のデータホール竣工

コロナ感染による、拡張工事停止のリスク

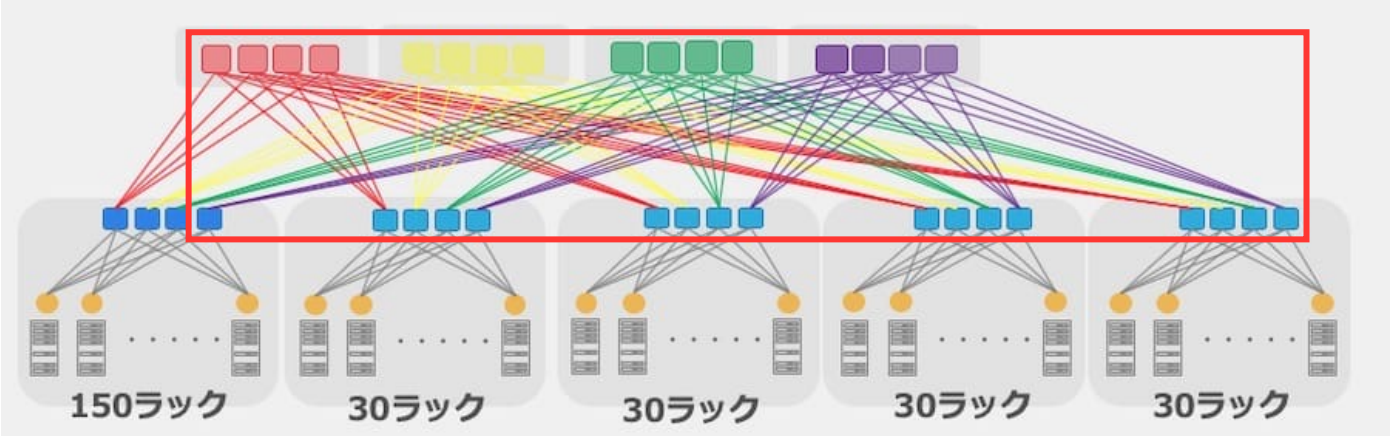
- 新データセンタの**二つ目のデータホール**の建設が始まった頃にコロナ禍に
- 工事業者やデータセンタ運用メンバのコロナ感染により**工事が停止するリスク**の中でネットワーク拡張の現地作業をする必要があった
- データセンタ運用メンバの中でのシフト勤務だけでなく、**工事業者の作業時間や移動経路を考慮しての勤務**をすることとなった
  - 設置作業や配線作業が**深夜時間帯メイン**に

シンプルに行動制限や普段と違う時間帯での勤務で大変だった

2020/7 二つ目のデータホール竣工

Clos ネットワークの構成の変更と配線作業

- **構成変更に伴う新規設置作業と配線作業**
  - Fabric層の追加による機器の増加
  - SpineスイッチとFabricスイッチ間の新規配線
- **夜な夜な設置と配線作業をこなした**
  - ローカルメンバと駐在員の当時4名で対応



三つ目のデータホールの竣工時の納期遅延

- **様々なものの納期遅れが発生**
  - スイッチ、サーバ(パーツ)、データセンタ設備
- **すべてインテグレーションしてから納品のため、何かがかけるとそれがクリティカルパスになって全て遅れる**
  - インテグレーション時に簡単な動作試験なども行うため、**何か足りない状態で納品されると現場で動作試験を行う必要**がある
  - 遅延した部材を**在庫で機器でまかなう**
- **利用ユーザとの調整**
  - ユーザと密に連携して、いつまでにどのくらいのリソースが必要かを把握し、影響与えないように引き渡しを実施

旧データセンタ撤退時の苦労

様々な苦労

- **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

日本からのヘルプ無し

- 日本からヘルプが来ることが出来なかったため、**現地メンバーのみで対応**
- **17時間の時差(日本の朝がアメリカ現地の夕方)**との戦い
  - 夕方にMTGして方針を決め、その後**夜から深夜にかけて行動**
- 現地でやるべきことは**全てやる**
  - 機器の電源シャットダウン
  - ディスク抜き
  - 配線撤去
  - ラックの引き取り手配、立ち会い
  - ゴミ処理の手配
- **最終的に CEO も動員**

ディスク抜きで感じたOCPサーバの恩恵

- 従来のサーバでは、**ディスクはマウンタにネジ止め**されている
- しかし OCP サーバでは、様々なパーツが**基本、ツールレスで交換可能**
  - **ディスクも一切ネジ止めされていない**
- ネジ止めされていたディスクは、ネジ穴が潰れて(なめて)回せないといった理由で取り外せないものもあり、**OCPに比べて作業負荷が高かった**

ネジとの格闘の後

撤退時に見つけたダミーサーバ

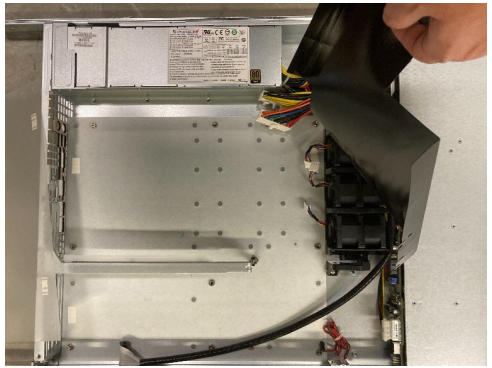
- ラックにサーバがマウントされているが、ネットワーク機器もなく、ラックPDUも搭載されていないラックが見つかった
- ラックがサーバ搭載されているのに**異様に軽い**
- そのサーバにはマザーボード含めて**何もパーツが搭載されていない**
  - ただなぜか、**電源だけは搭載**されていた
- **発見当時は何のためのサーバかわからず混乱**

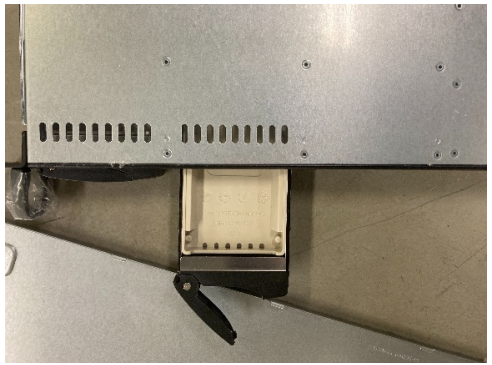
撤退時に見つけたダミーサーバ

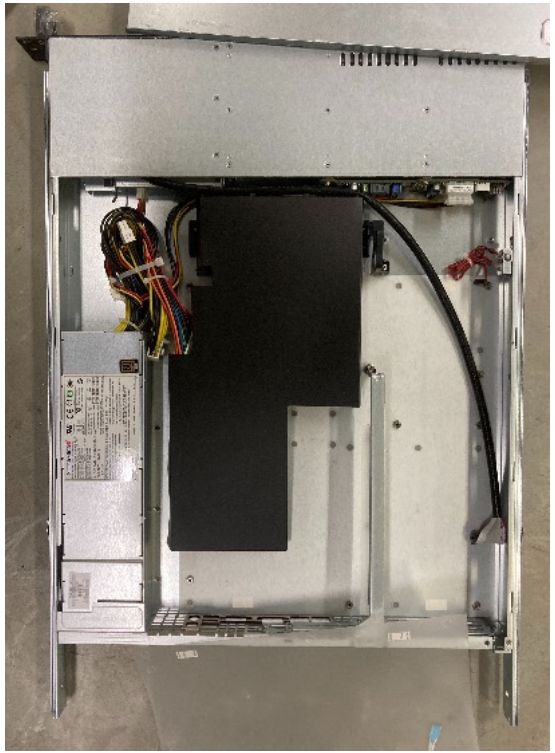
- ラックにサーバがマウントされているが、ネットワーク機器もなく、ラックPDUも搭載されていないラックが見つかった
- ラックがサーバ搭載されているのに**異様に軽い**
- そのサーバにはマザーボード含めて**何もパーツが搭載されていない**
  - ただなぜか、**電源だけは搭載**されていた
- **発見当時は何のためのサーバかわからず混乱**

- **ラック移動やサーバのマウントなど問題ないか確認するために利用したダミーのサーバと判明**

撤退時に見つけたダミーサーバ







DNS リゾルバの変更漏れ

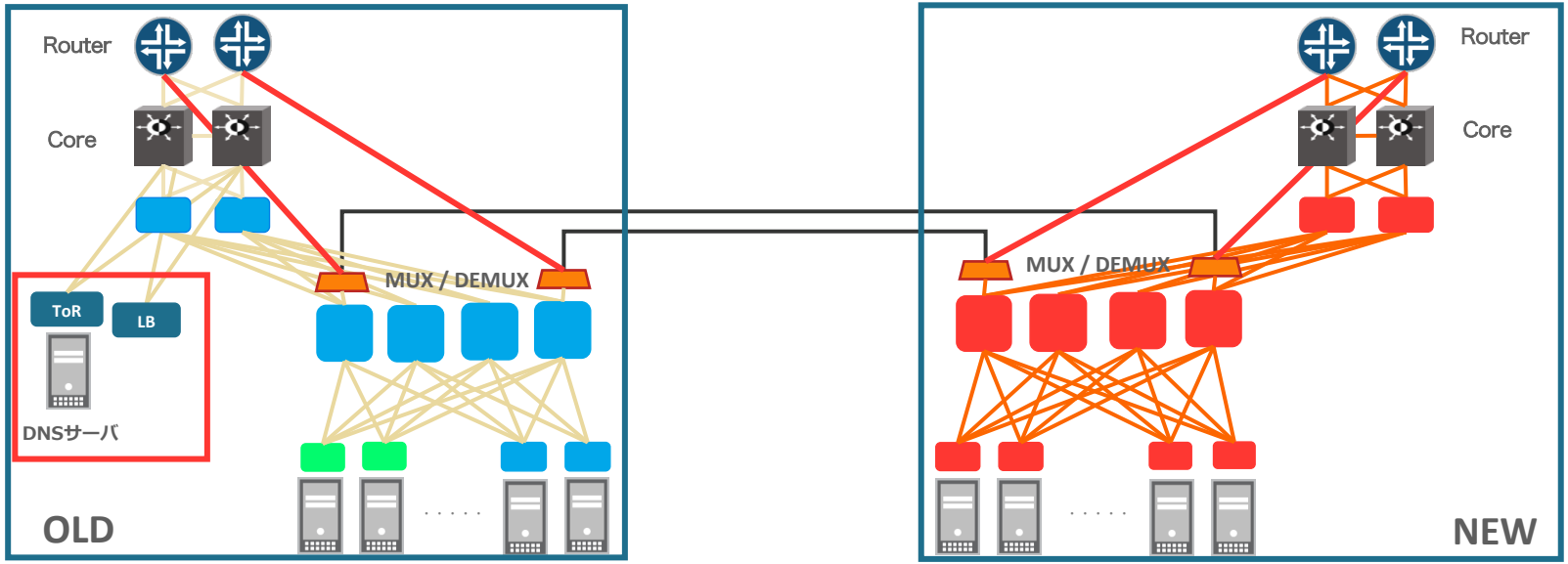
- 新データセンタ側のDNSリゾルバの向き先が旧データセンタに向いていることが**撤退完了の1ヶ月前に発覚**
- Resolve.conf に設定されているのは 2VIP
  - ともに L2DSR のため、**DNSのサーバとロードバランサは同じVLAN**
  - **2VIPは別VLAN**
- Hadoop では resolve.conf の設定を反映するのに**プロセスの再起動が必要**
  - Hadoopクラスタを停止して全て再起動できればいいが、それには**利用者とメンテナンスの調整が必要**
  - メンテナンスを要さず、サービスアウトとプロセス再起動をやろうとすると**数ヶ月かかってしまう**

DNS リゾルバの変更漏れ

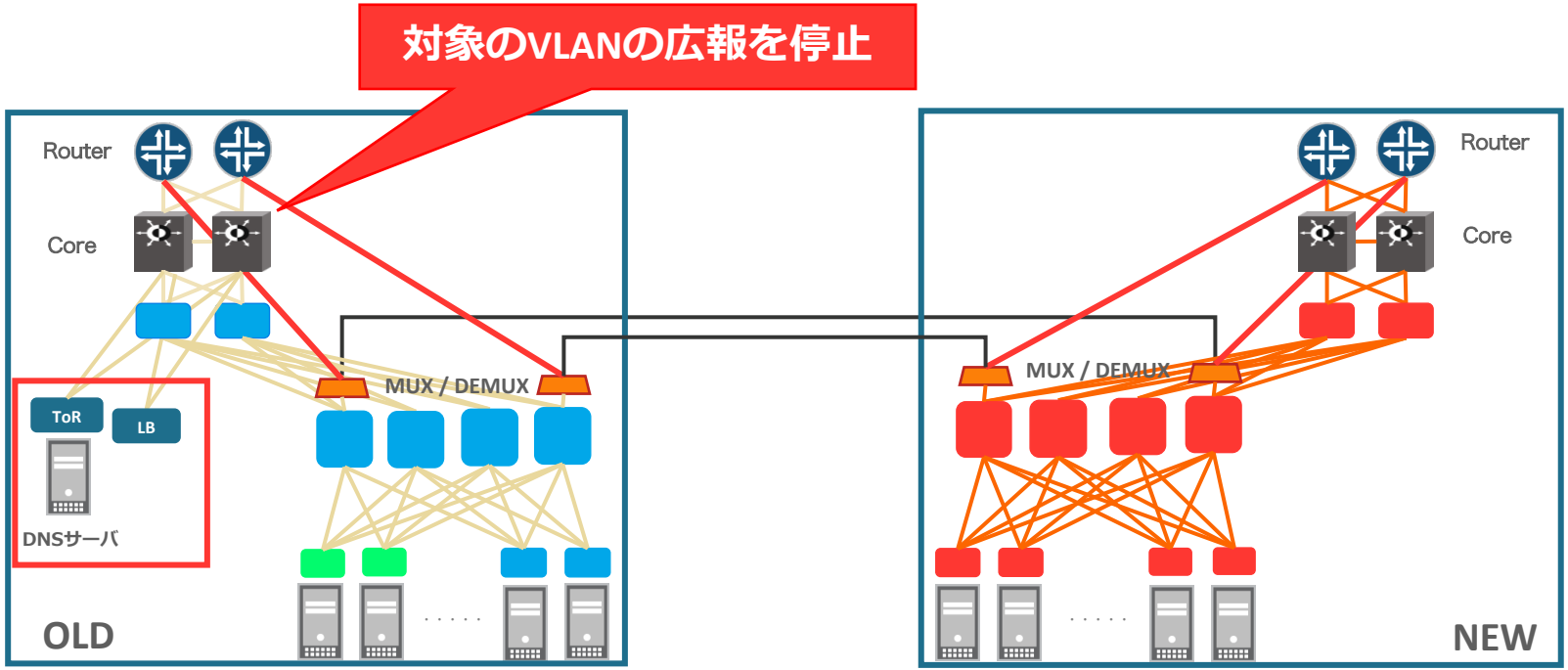
旧データセンタのDNSサーバ、ToRスイッチ、
ロードバランサ、VLAN など関連する全てを
新データセンタへVIP毎の2回に分けてお引っ越し

旧データセンタ撤退時の苦労

DNS リゾルバの変更漏れ

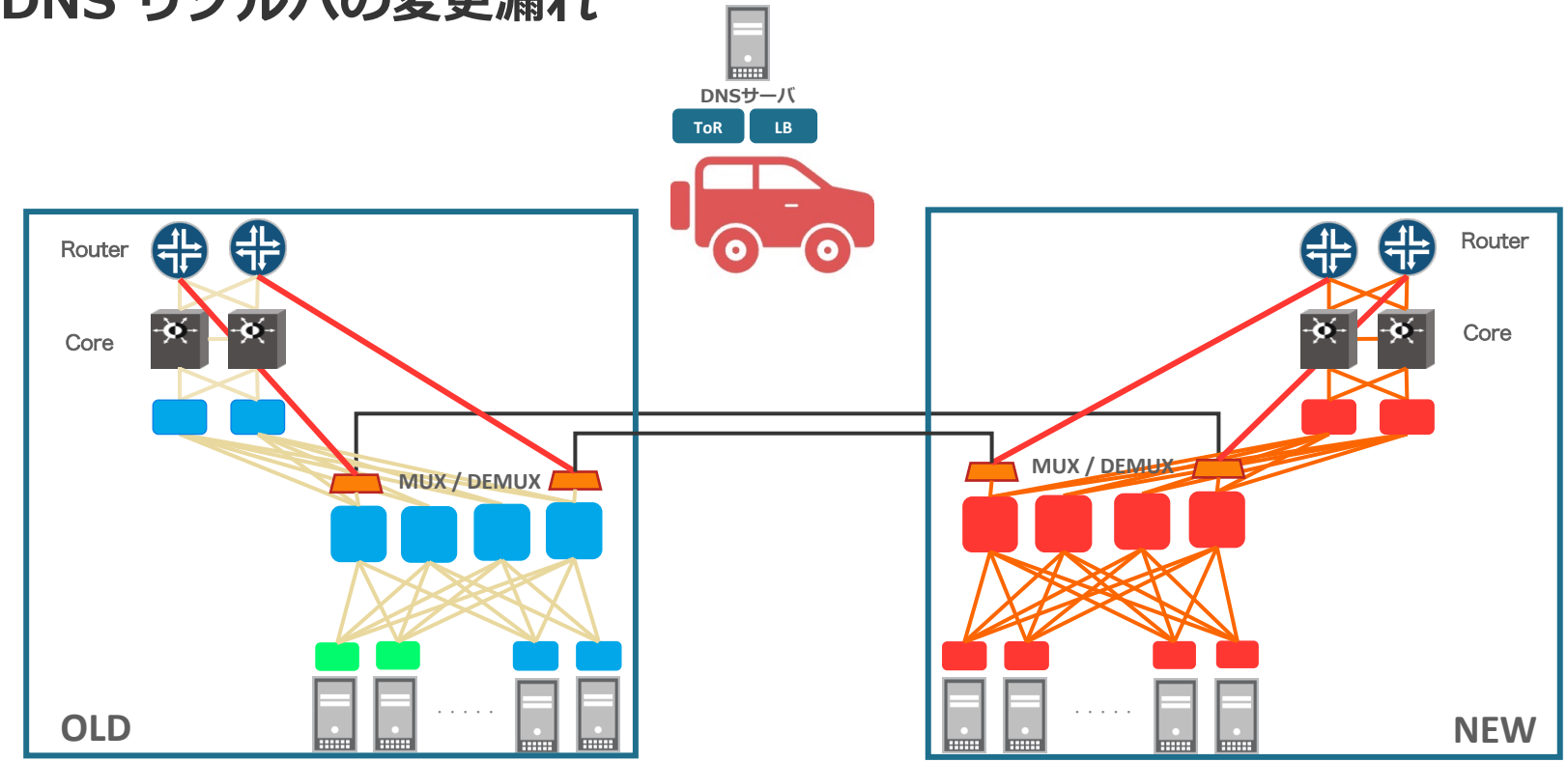


DNS リゾルバの変更漏れ



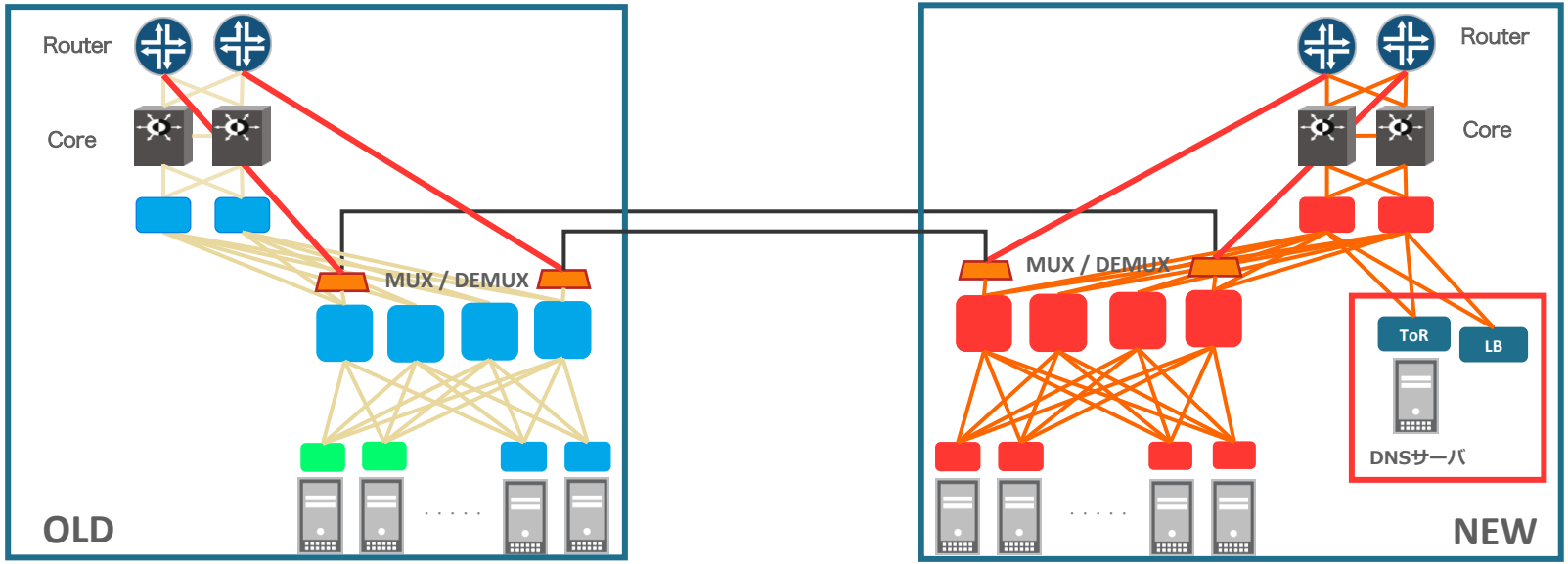
旧データセンタ撤退時の苦労

DNS リゾルバの変更漏れ



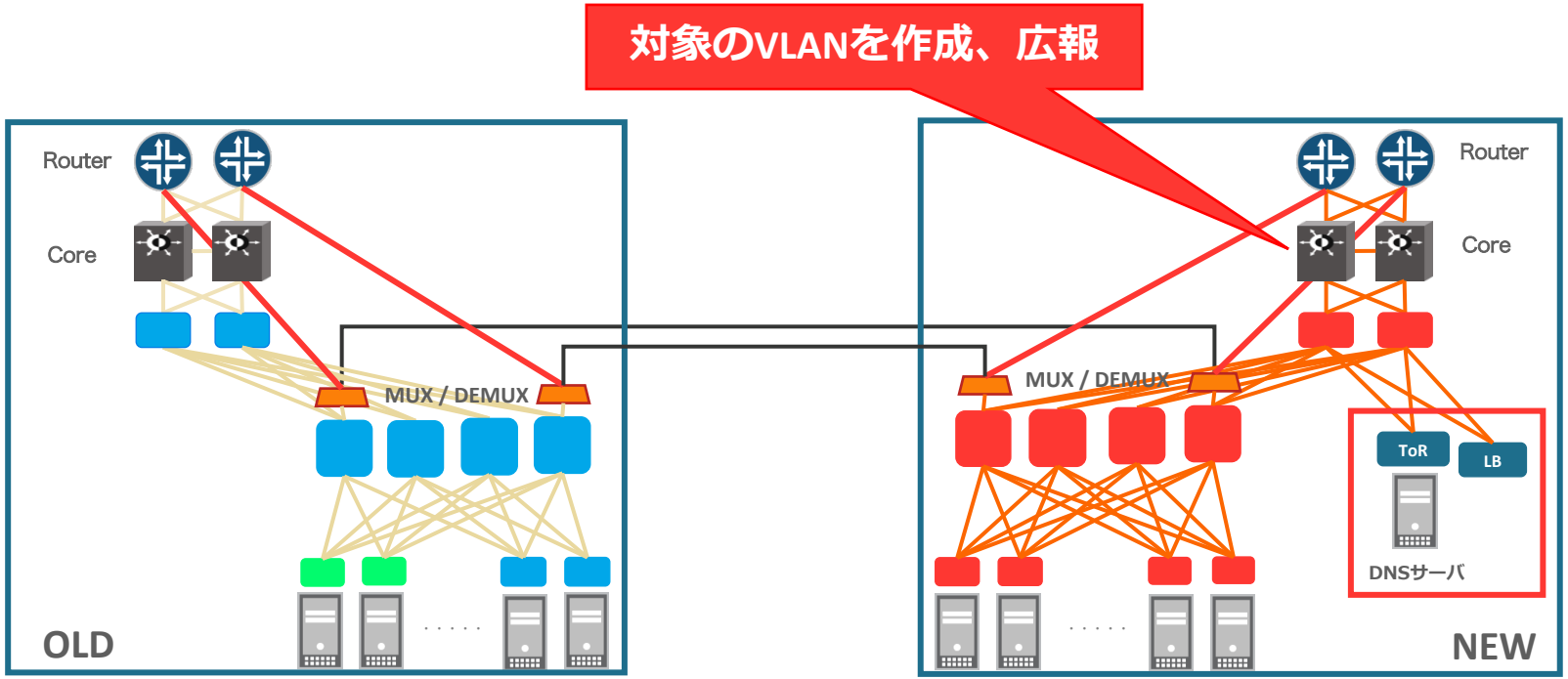
旧データセンタ撤退時の苦労

DNS リゾルバの変更漏れ



旧データセンタ撤退時の苦労

DNS リゾルバの変更漏れ



旧データセンタ撤退時の苦労

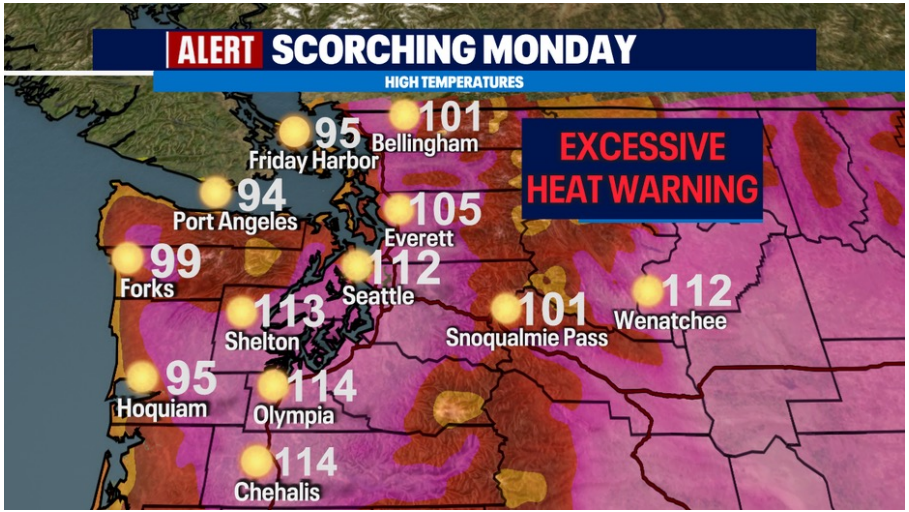
DNS リゾルバの変更漏れ

無事DNSが移行された

1000年に1度の熱波

1000年に1度の熱波って何？

- 2021年6月末頃に北米を襲った熱波(Heat Wave)
  - 北米に「1000年熱波」カナダ西部は47.9度で2日連続の国内記録
- 北米(カナダ含む)の各所で過去最高気温を更新



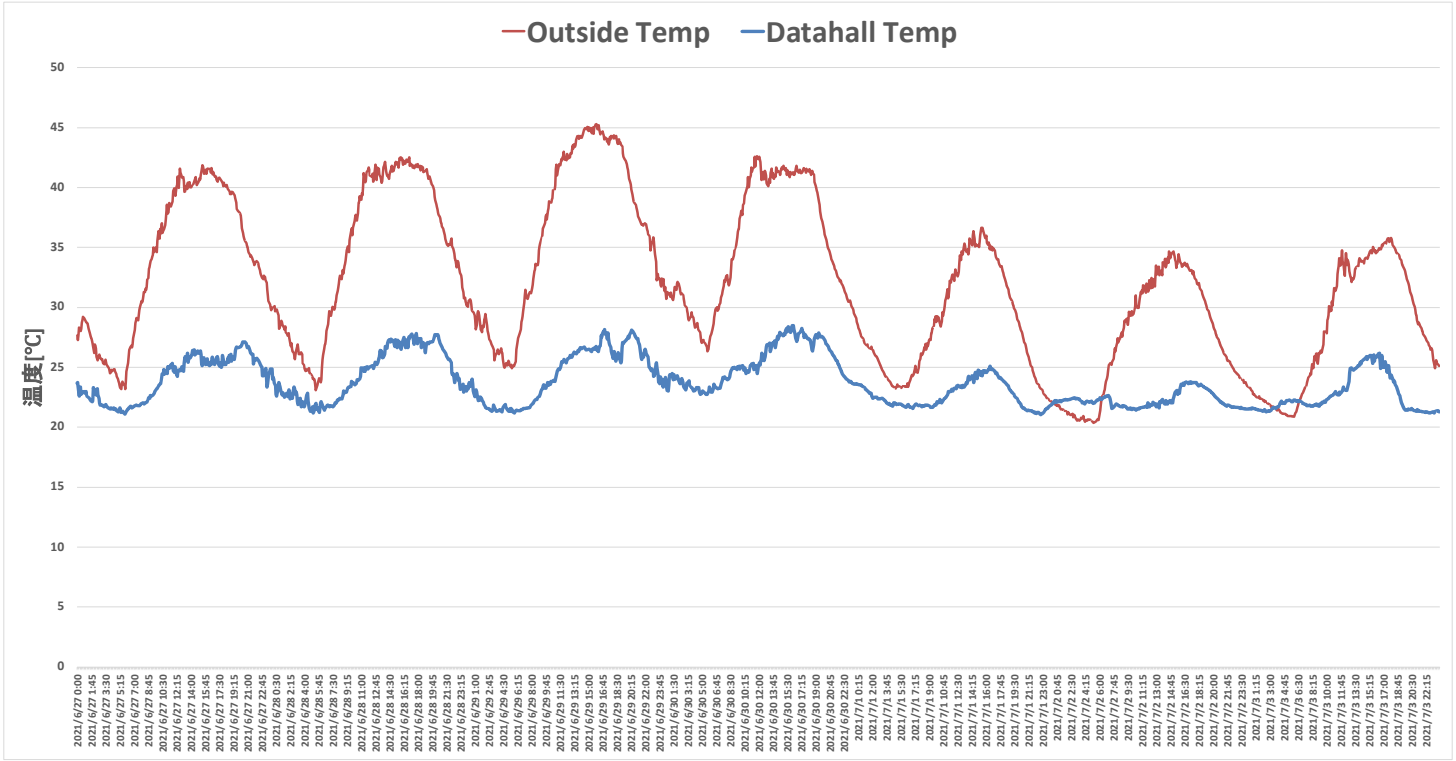
https://www.fox13seattle.com/weather/dangerous-heat-continues-across-washington-state

現場での熱波対策

- 事前準備
  - 予定されていた設備メンテナンスを**全てリスケ**
  - 緊急パターンを複数想定し、**現場対応訓練の実施**
- 当日
  - サーバルーム用の空調機の前で散水し、取り込む外気を気化熱で冷却
    -> **取り込む前に先に外気を冷やしておく作戦**
  - サーバルーム付近にFANを配備し、**非常時にはすぐに動かせる状態を準備**
  - 温度ピーク前後は**ITエンジニア、設備エンジニアが現場に常駐し、設備に異常がないか注視**

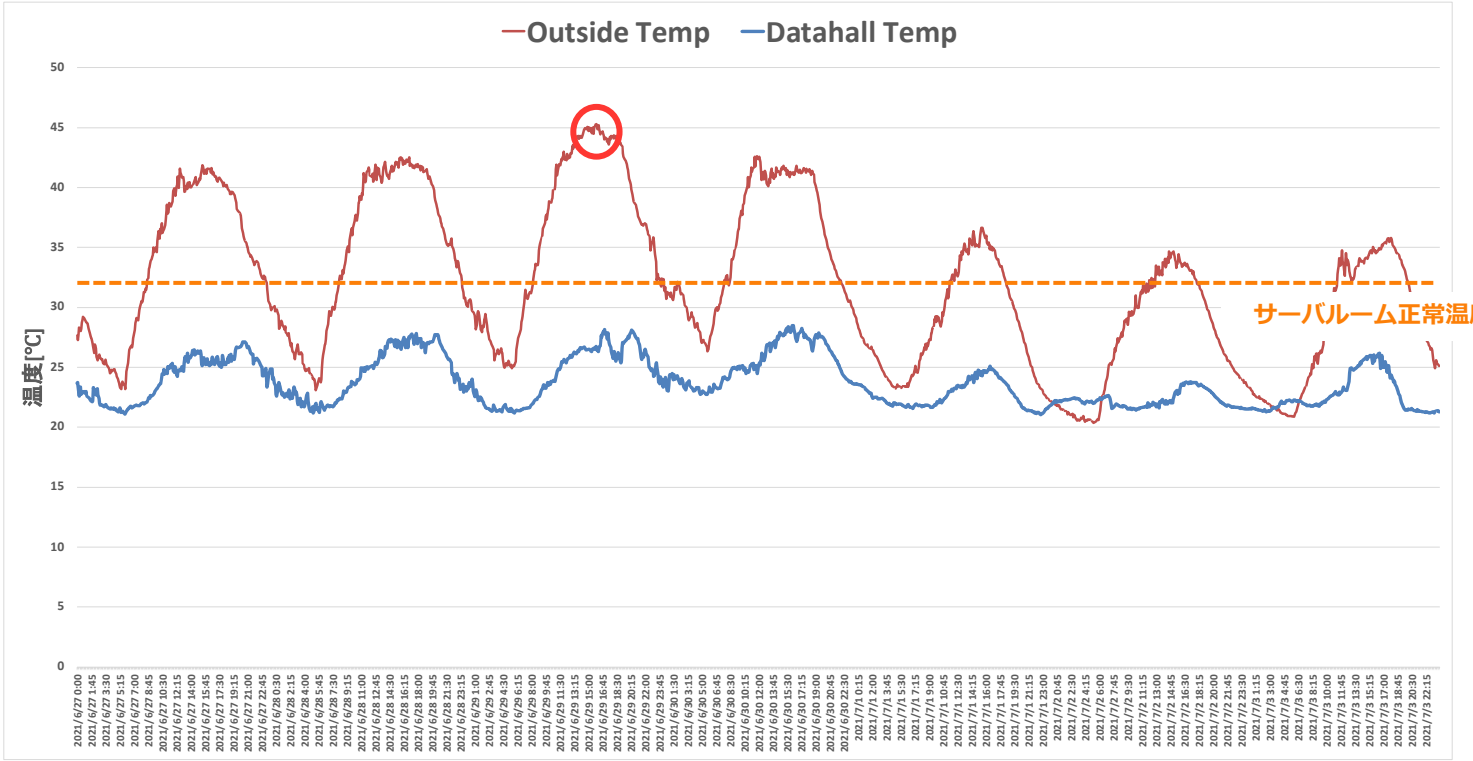
1000年に1度の熱波

外気温とデータホールの気温の推移



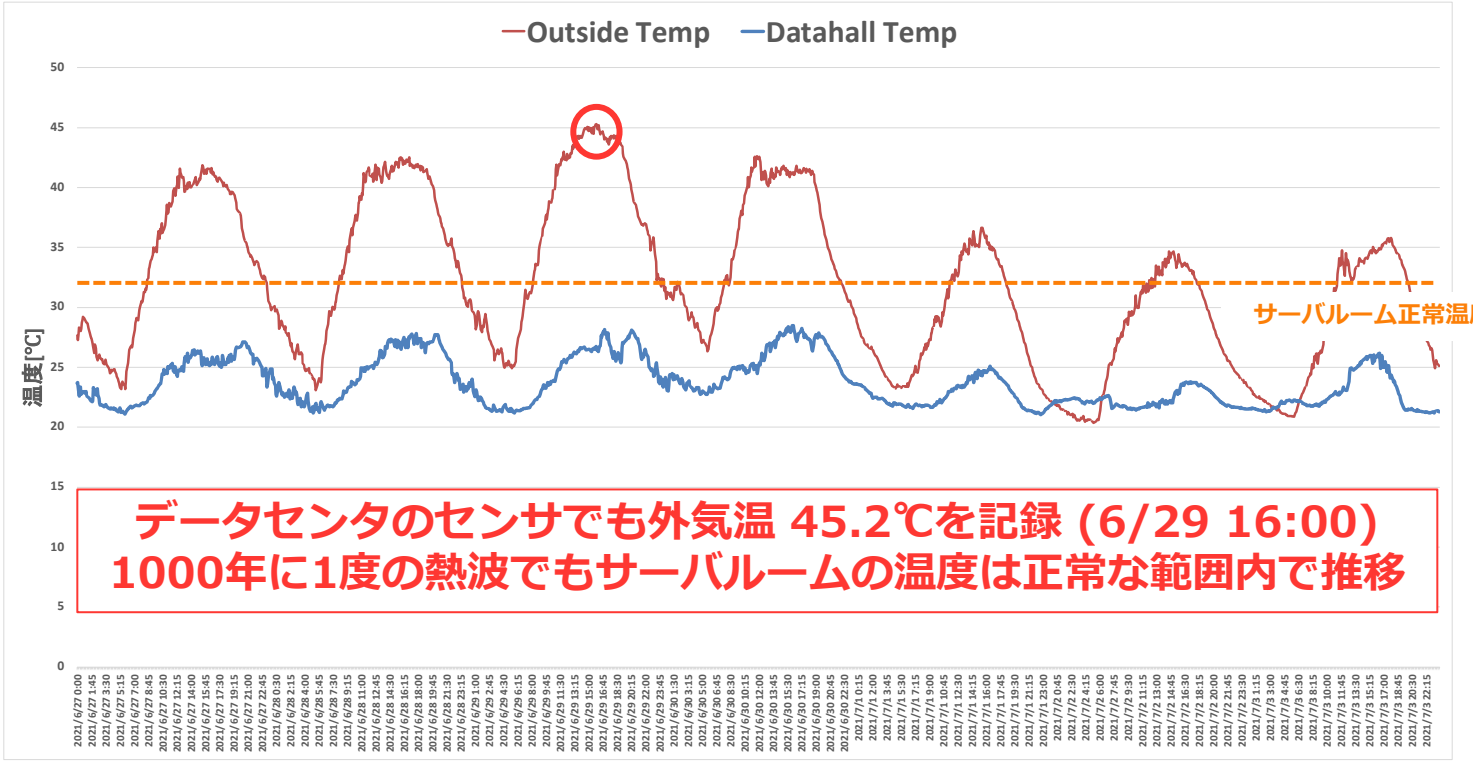
1000年に1度の熱波

外気温とデータホールの気温の推移



1000年に1度の熱波

外気温とデータホールの気温の推移



1000年に1度の熱波

熱波の実体験

- 外に出ると**「あ、外にいちゃ駄目だな」**とすぐに思う
- 高温が原因で**散発的な停電**が発生
- 小売店の**コンプレッサーが故障し、冷凍、冷蔵商品をすべて廃棄**
- 赴任者が契約している**アパートのインターネット回線がダウン**
- アパートの3階だと、広い部屋の**冷房が効かず**、冷房のある別の狭い部屋に避難
  - 1階だと大丈夫だった模様
- **携帯キャリアの電波がダウン**
- 通信障害でCostcoやファーストフード店の**システムがダウン**

熱波で回線障害

熱波で局舎がダウン

- アメリカ国内の回線の**局舎(ラストワンマイル)の空調が故障**
  - 設計値を超えた気温だったため
- 空調が壊れたことで、**局舎内の気温が上がり続けて**しまったため、回線の機器に影響が発生
  - 障害チケットによると**79度近く**まで温度が上がったとのこと
  - 局舎内のケーブルラダーやパッチパネルなどプラスチック部分が溶けて、曲がってしまったりもした
- Spot Cooler などで空調が復旧した後も、高気温による影響で**断続して機器の再起動や交換作業**が発生

銃で回線障害

銃で回線障害

- 回線に**銃の弾が当たったため**、回線が切れてしまった

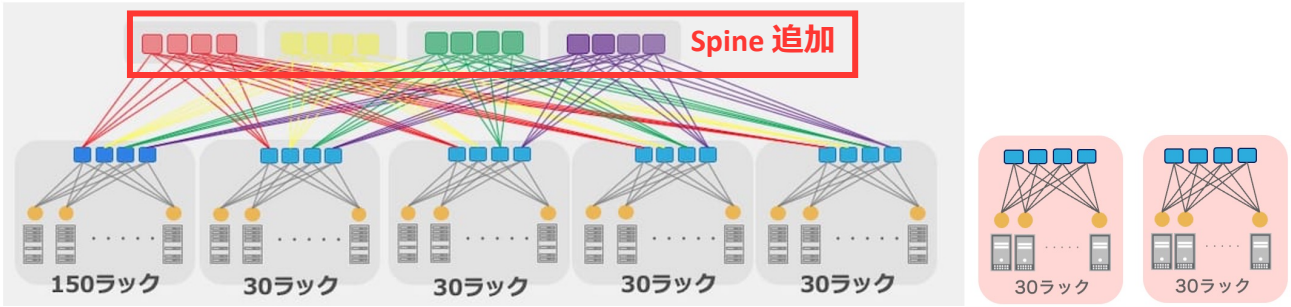
実際の障害チケットのコメント

> We are seeing some services restore at this time but our partner provider has not confirmed if they are hands-off at this time or not. We are reaching out to them for confirmation and will relay that information once known. This was a fiber event due to a **“gunshot”** on the affected span which required patching around by partner provider.

2023/6 これから

2023/6 これから

- **四つ目のデータホール** の建設計画が進行中
- **Clos ネットワークの拡張**
  - 今の構成のままでは収容しきれないため **Spineの追加を含めた構成変更を検討**
- **デプロイツールの見直し**
  - 5年経ったことによる、運用体制などの変化に対応するために
- **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



まとめ

まとめ

- **アメリカの利点を活かして、クリーンエネルギーかつ高効率なデータセンタを運用**
  - 電気料金、水力発電、OCP、建設 etc.
- **状況に合わせてネットワークが様々な構成に**
  - 2層Closネットワーク -> DFを利用した拠点間接続
    -> 旧データセンタからの移設 -> 3層Closネットワーク
    -> 旧データセンタネットワーク撤退 -> Pod追加 -> Next Design
- **新しいことにチャレンジすると色々なことが起きる**
  - 1000年に1度のことも起きるし、回線を銃で撃たれることもある

最後に

- **アメリカのデータセンタのイメージつかめましたか？**
  - 想像通りだった、想像とはまるで違ったなど
  - もっと知りたいことがあれば是非教えてください！
- **アメリカにデータセンタ建ててみたくなりましたか？**
  - 建ててみたいという方は是非！
- **ネットワーク、回線のトラブルや対応に関して**
  - あの時の対応、こうすればよかったんじゃないかな？など
    あれば教えてください！