CyberAgent

AI/ML基盤の400G DCネットワークを構築した話

JANOG52 Day3 10:15-10:45

前半: 内田 泰広 (Yasuhiro Uchida)

後半: 小障子 尚太朗 (Shotaro Koshoji)

2023/07/07

内田 泰広 (Yasuhiro Uchida)

- **2020 株式会社サイバーエージェント 中途入社**

  CIU(CyberAgent group Infrastructure Unit)

  Platform Div NWリーダー

- **業務内容**
  - AS運用・peering・拠点間接続
  - データセンターのネットワーク設計・構築・運用
  - 高負荷DCのGPU向け400G NW構築
  - ネットワーク監視・運用自動化
  - プロジェクト管理

小障子 尚太朗 (Shotaro Koshoji)

- **2019/4 株式会社日本レジストリサービス 新卒入社**

  DNSサーバーや周辺ネットワークの運用

- **2021/12 株式会社サイバーエージェント 中途入社**

  CIU Platform Div NWエンジニア

- **業務内容**
  - ネットワーク系業務全般
    - AS運用・peering・拠点間接続
    - データセンタのネットワーク設計・構築・運用
    - 高負荷DCのGPU向け400G NW構築
    - ネットワーク監視・運用自動化

Cycloud (AS24284)

- **プライベートクラウドとしてIaaS、 KaaS、 ML Platformを提供**
- **今回はML Platformで新しくデータセンターネットワークを構築したお話**

ここのお話

ML Platform

- **GPU環境の提供**
  - KubernetesのPodとしてGPUを提供
  - AI/MLジョブの学習環境 & モデルのデプロイ環境
  - マネージドJupyter Notebookなどを提供

- **オンプレを選定している理由**
  - クラウドより早いスピード感
  - 既存サービスとの接続
  - 長期利用時のコスト

ML Platformで直面していた課題

・大規模言語モデル(LLM)を様々な事業に活用している

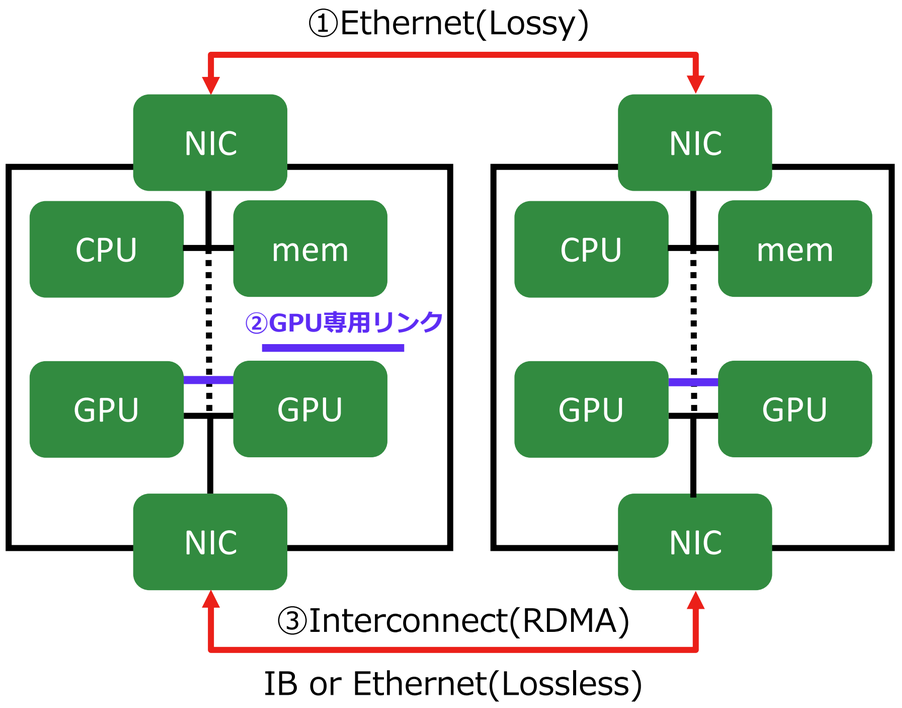
```
$ python3 OpenCALM-7B.py
大規模言語モデルとは、様々な自然言語で書かれたテキストをデータセットとして、機械的に学習したモデルです。
主に人工知能分野で利用される手法です。
```

※ “CA製の大規模言語モデル(OpenCALM)”に”大規模言語モデルとは”を聞いてみた結果

- 課題1: LLMの拡大により1台のGPUサーバーのメモリでの処理が困難(モデルパラレル)
- 課題2: 学習時間が長い (データパラレルの並列数を増やしたい)

複数台のGPUサーバーで並列分散学習できる環境

並列分散処理するためには



以下の3種類のNWを用意

① CPUセントリックなKubernetesなどの通信はLossyなEthernet

② サーバー内のGPU間通信はPCIeよりも広帯域なNVLinkで通信

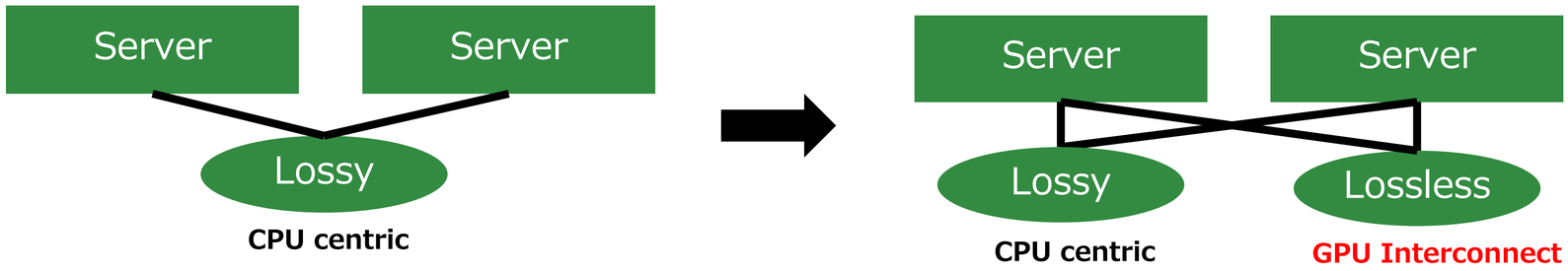
- **NVLinkの総帯域は900GB/sec**
- **GPU(AI/ML)の進化がネットワーキングにプレッシャーを与えている**

③ サーバーを跨ぐGPU間通信は①と独立したInterconnectでRDMA

- **RDMA: CPUを介さずリモートホストのGPUメモリにダイレクトアクセス**

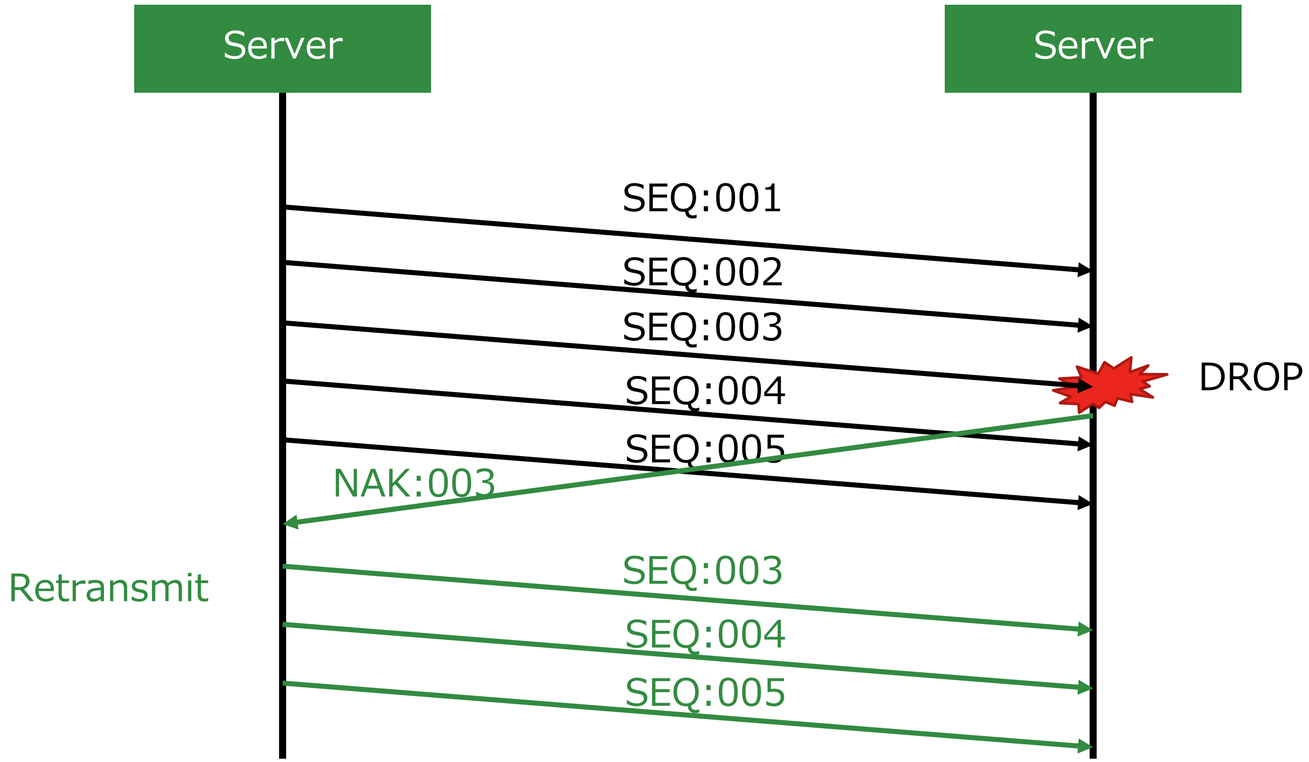
Interconnectの技術選定

- **サーバー間のRDMA専用NW(Interconnect)はInfiniband vs RoCEv2の選択肢**
  - RoCEv2(Routable RoCE)を採用: マルチテナント
  - **Infinibandの知見が少なくEthernetには安心感を感じた**
  - **Interconnectの帯域は400GbEを採用: ボトルネックの軽減**
- **RoCEv2には新しいデータセンターネットワークの要件が必要**
  - **Lossless Ethernet**
  - RDMAはLossが無いNWを前提としたプロトコルでLossがパフォーマンスに大きく影響



なぜLossがRDMAの性能に影響を与えるのか

- **RDMAではgo-back-N方式が多く採用されている**
  - **最近はSelective Repeatも実装されてきているが機能的制限あり**
- **Go-back-NはDROPしたSEQから遡って再送を実施し無駄が多い**
- **RDMAにはLossがないNWが求められる**

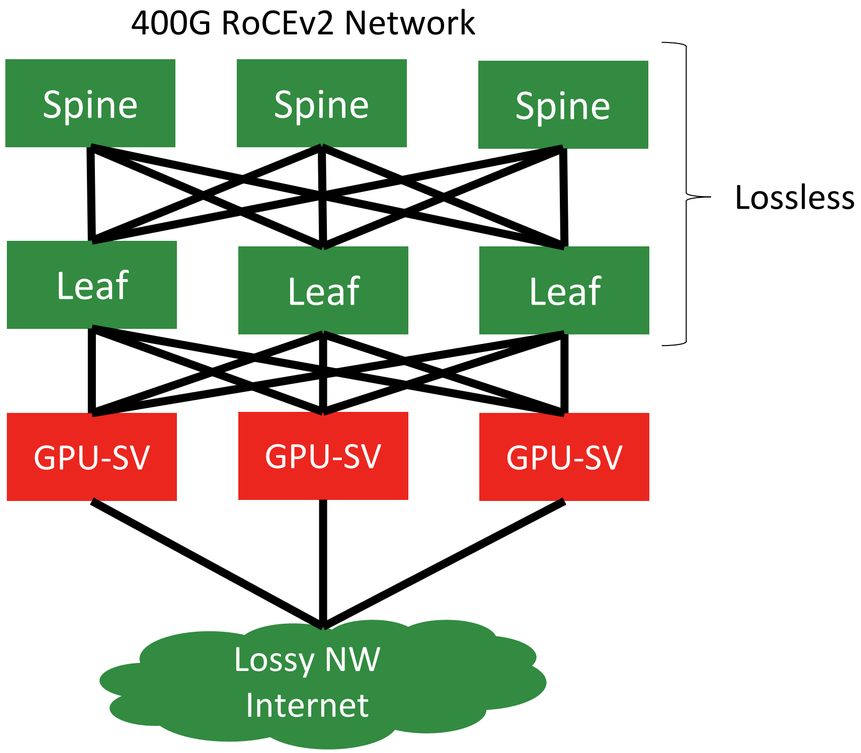


Lossless Ethernetとは

- **輻輳制御とQoSを使った信頼性の高いネットワーク：輻輳制御ではDCQCNを利用**
- **[IEEE DCB] PFC + ECN (CNP) + ETSを利用したロスが少ない信頼性の高いイーサネット**
- **PFC (Priority Flow Control)**
  - PAUSEフレームを使ったリンクレベルでの輻輳を抑制
  - トラフィックはCoSまたはDSCPで分類されキュー毎にPAUSEフレームで制御
- **ECN (Explicit Congestion Notification)**
  - エンドノード間の輻輳制御
  - IPヘッダのToSを利用し輻輳時に中継SWがパケットにマーキングを行い受信ノードに輻輳を通知
  - 受信ノードは送信元ノードにCNP(Congestion Notification Packet)パケットを送信し輻輳を通知
  - CNP受信後は送信ノードはトラフィック流量の制限
- **ETS (Enhanced Transmission Selection)**
  - プライオリティグループ毎にキューの優先順位を定義
  - RDMAとCNPのパケットだけを優先的に送信できるように設定
- **動作の詳細はAppendixに記載**

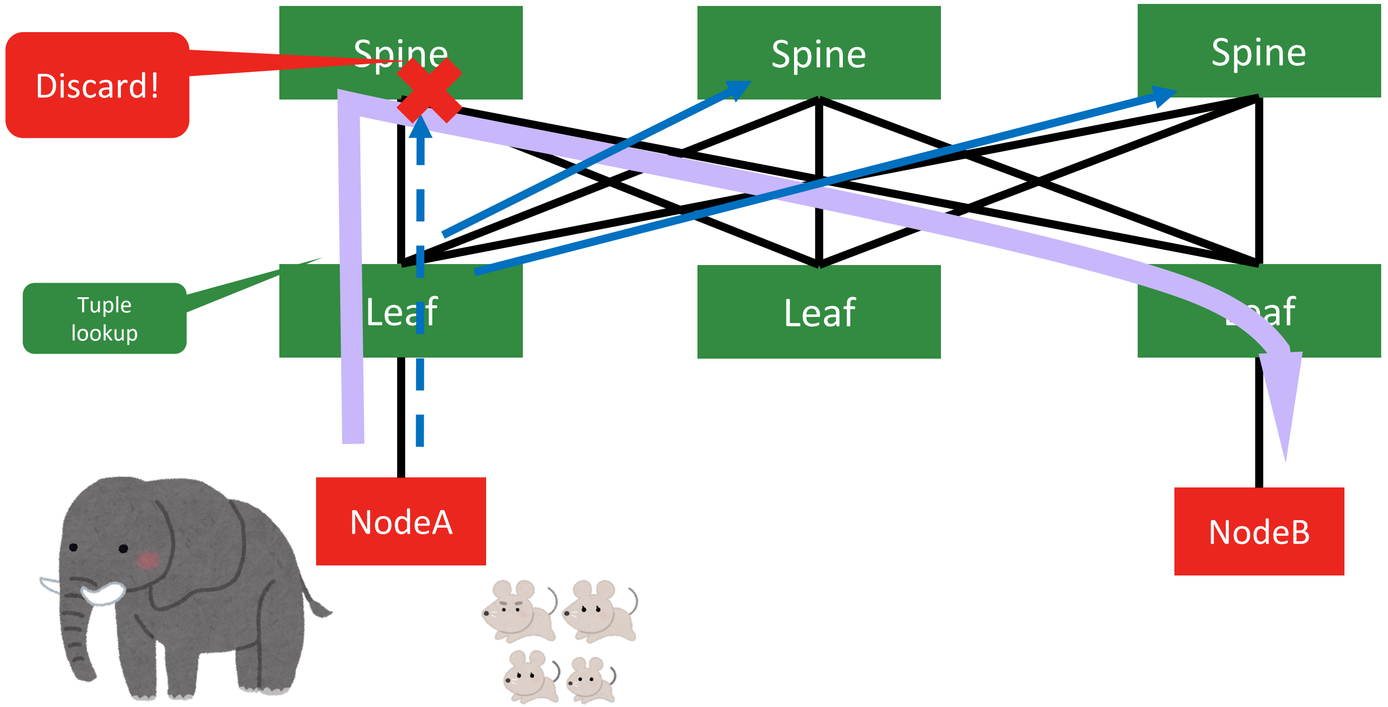
採用したInterconnect構成

- **Lossless NWは400GbEフルバイセクション構成**
- **Rail Optimized構成(GPU-SV:Leaf間)**
  - **■ 各Leafスイッチにケーブルを1本ずつ接続**
  - **■ アプリ側で最適なNICを選択する時の推奨構成**
- **400GbEで統一し異速度カットスルーの低減**
- **LAGの撤廃: 障害時などのフローの不均衡防止**
- **Adaptive Routingの採用**



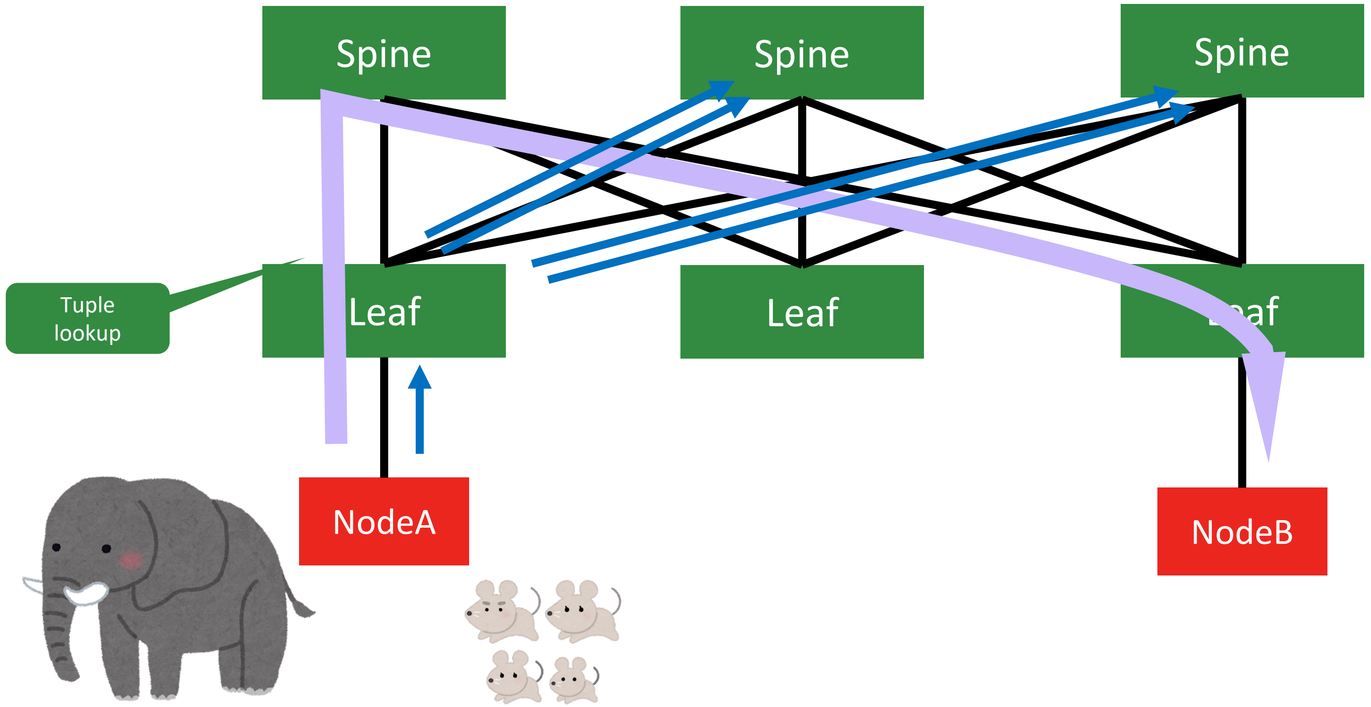
Elephant flow / Mice flow問題

- **5-tupleが同じパケットはスイッチのhashにより同じ経路を通過(static hashing)**
- **Elephant flowが発生するとMice flowなどにも影響が発生**
- **バッファを考慮しないhash方式のトラフィックの偏りが問題**



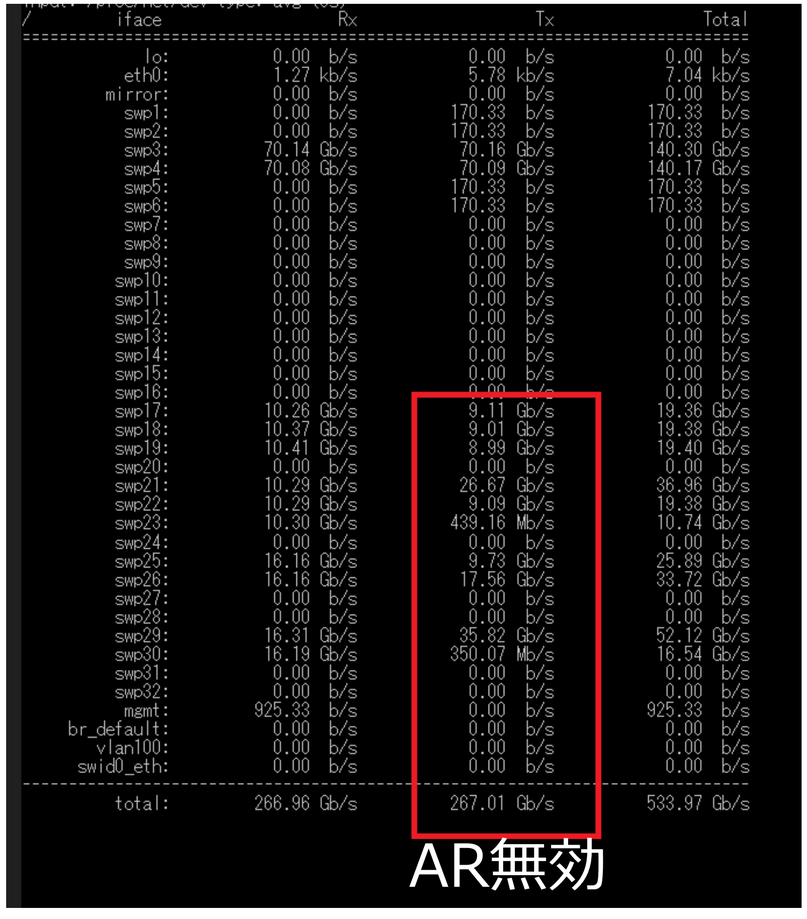
Elephant flow / Mice flow問題
- Adaptive Routing -

- パケットをflowletとして認識し、バッファの状況を見てdynamicにhashを実施
- Elephant flowが発生してもSpine-Leaf間を均等にhash
- flowlet awareな処理なためout-of-orderの発生を抑制
- IP NWでは厳密なノンブロッキングは不可能なのでAdaptive Routingによるアプローチを選択



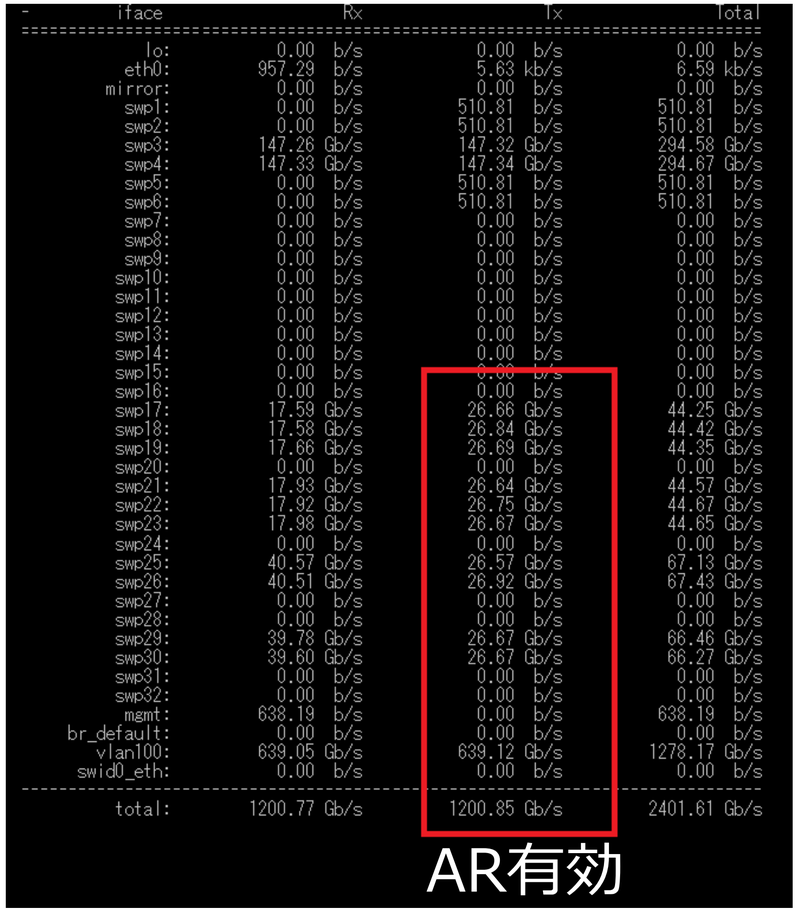
Adaptive Routing(AR)の検証

- **ARを有効/無効にしたLeaf SWが混在した環境でMPIを実行**
- **ARを有効にしたスイッチに置いてSpine向けのTX側トラフィックが均等に分散**



iface	Rx	Tx	Total
lo:	0.00 b/s	0.00 b/s	0.00 b/s
eth0:	1.27 kb/s	5.78 kb/s	7.04 kb/s
mirror:	0.00 b/s	0.00 b/s	0.00 b/s
swp1:	0.00 b/s	170.33 b/s	170.33 b/s
swp2:	0.00 b/s	170.33 b/s	170.33 b/s
swp3:	70.14 Gb/s	70.16 Gb/s	140.30 Gb/s
swp4:	70.08 Gb/s	70.09 Gb/s	140.17 Gb/s
swp5:	0.00 b/s	170.33 b/s	170.33 b/s
swp6:	0.00 b/s	170.33 b/s	170.33 b/s
swp7:	0.00 b/s	0.00 b/s	0.00 b/s
swp8:	0.00 b/s	0.00 b/s	0.00 b/s
swp9:	0.00 b/s	0.00 b/s	0.00 b/s
swp10:	0.00 b/s	0.00 b/s	0.00 b/s
swp11:	0.00 b/s	0.00 b/s	0.00 b/s
swp12:	0.00 b/s	0.00 b/s	0.00 b/s
swp13:	0.00 b/s	0.00 b/s	0.00 b/s
swp14:	0.00 b/s	0.00 b/s	0.00 b/s
swp15:	0.00 b/s	0.00 b/s	0.00 b/s
swp16:	0.00 b/s	0.00 b/s	0.00 b/s
swp17:	10.26 Gb/s	9.11 Gb/s	19.36 Gb/s
swp18:	10.37 Gb/s	9.01 Gb/s	19.38 Gb/s
swp19:	10.41 Gb/s	8.99 Gb/s	19.40 Gb/s
swp20:	0.00 b/s	0.00 b/s	0.00 b/s
swp21:	10.29 Gb/s	26.67 Gb/s	36.96 Gb/s
swp22:	10.29 Gb/s	9.09 Gb/s	19.38 Gb/s
swp23:	10.30 Gb/s	439.16 Mb/s	10.74 Gb/s
swp24:	0.00 b/s	0.00 b/s	0.00 b/s
swp25:	16.16 Gb/s	9.73 Gb/s	25.89 Gb/s
swp26:	16.16 Gb/s	17.56 Gb/s	33.72 Gb/s
swp27:	0.00 b/s	0.00 b/s	0.00 b/s
swp28:	0.00 b/s	0.00 b/s	0.00 b/s
swp29:	16.31 Gb/s	35.82 Gb/s	52.12 Gb/s
swp30:	16.19 Gb/s	350.07 Mb/s	16.54 Gb/s
swp31:	0.00 b/s	0.00 b/s	0.00 b/s
swp32:	0.00 b/s	0.00 b/s	0.00 b/s
mgmt:	925.33 b/s	0.00 b/s	925.33 b/s
br_default:	0.00 b/s	0.00 b/s	0.00 b/s
vlan100:	0.00 b/s	0.00 b/s	0.00 b/s
swid0_eth:	0.00 b/s	0.00 b/s	0.00 b/s
total:	266.96 Gb/s	267.01 Gb/s	533.97 Gb/s

AR無効



iface	Rx	Tx	Total
lo:	0.00 b/s	0.00 b/s	0.00 b/s
eth0:	957.29 b/s	5.63 kb/s	6.59 kb/s
mirror:	0.00 b/s	0.00 b/s	0.00 b/s
swp1:	0.00 b/s	510.81 b/s	510.81 b/s
swp2:	0.00 b/s	510.81 b/s	510.81 b/s
swp3:	147.26 Gb/s	147.32 Gb/s	294.58 Gb/s
swp4:	147.33 Gb/s	147.34 Gb/s	294.67 Gb/s
swp5:	0.00 b/s	510.81 b/s	510.81 b/s
swp6:	0.00 b/s	510.81 b/s	510.81 b/s
swp7:	0.00 b/s	0.00 b/s	0.00 b/s
swp8:	0.00 b/s	0.00 b/s	0.00 b/s
swp9:	0.00 b/s	0.00 b/s	0.00 b/s
swp10:	0.00 b/s	0.00 b/s	0.00 b/s
swp11:	0.00 b/s	0.00 b/s	0.00 b/s
swp12:	0.00 b/s	0.00 b/s	0.00 b/s
swp13:	0.00 b/s	0.00 b/s	0.00 b/s
swp14:	0.00 b/s	0.00 b/s	0.00 b/s
swp15:	0.00 b/s	0.00 b/s	0.00 b/s
swp16:	0.00 b/s	0.00 b/s	0.00 b/s
swp17:	17.59 Gb/s	26.66 Gb/s	44.25 Gb/s
swp18:	17.58 Gb/s	26.84 Gb/s	44.42 Gb/s
swp19:	17.66 Gb/s	26.69 Gb/s	44.35 Gb/s
swp20:	0.00 b/s	0.00 b/s	0.00 b/s
swp21:	17.93 Gb/s	26.64 Gb/s	44.57 Gb/s
swp22:	17.92 Gb/s	26.75 Gb/s	44.67 Gb/s
swp23:	17.98 Gb/s	26.67 Gb/s	44.65 Gb/s
swp24:	0.00 b/s	0.00 b/s	0.00 b/s
swp25:	40.57 Gb/s	26.57 Gb/s	67.13 Gb/s
swp26:	40.51 Gb/s	26.92 Gb/s	67.43 Gb/s
swp27:	0.00 b/s	0.00 b/s	0.00 b/s
swp28:	0.00 b/s	0.00 b/s	0.00 b/s
swp29:	39.78 Gb/s	26.67 Gb/s	66.46 Gb/s
swp30:	39.60 Gb/s	26.67 Gb/s	66.27 Gb/s
swp31:	0.00 b/s	0.00 b/s	0.00 b/s
swp32:	0.00 b/s	0.00 b/s	0.00 b/s
mgmt:	638.19 b/s	0.00 b/s	638.19 b/s
br_default:	0.00 b/s	0.00 b/s	0.00 b/s
vlan100:	639.05 Gb/s	639.12 Gb/s	1278.17 Gb/s
swid0_eth:	0.00 b/s	0.00 b/s	0.00 b/s
total:	1200.77 Gb/s	1200.85 Gb/s	2401.61 Gb/s

AR有効

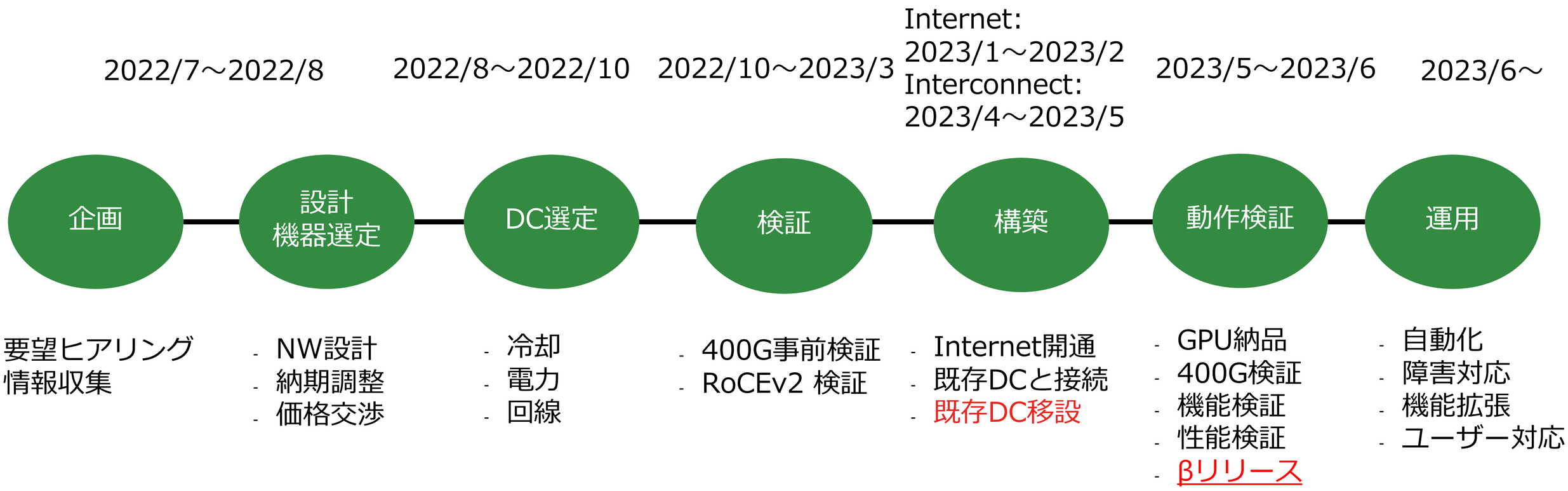
ML Platform: ラック写真

Lossy

Interconnect
Lossless

後半は現場目線の話

データセンター構築の流れ

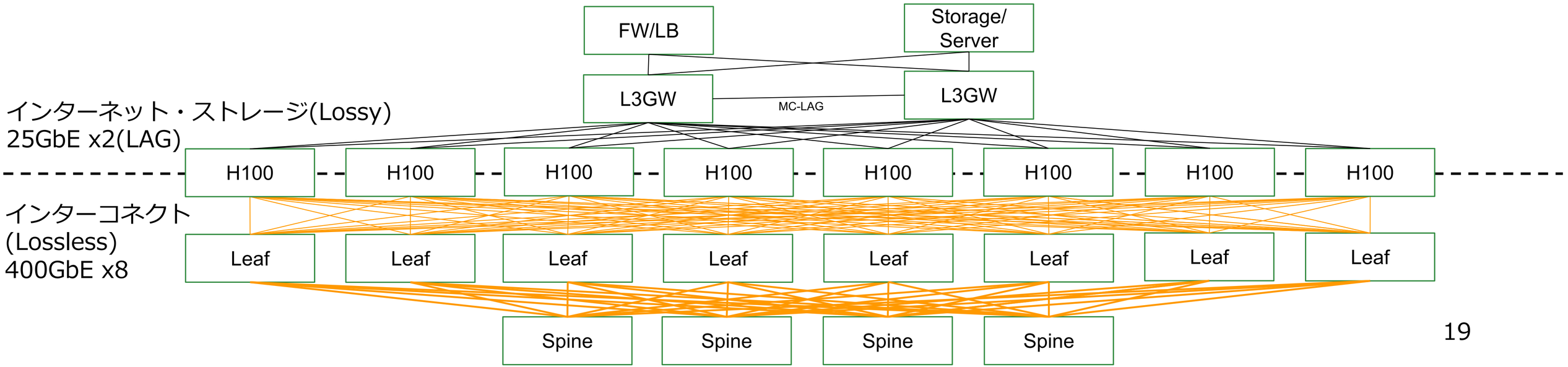


GPUサーバー用データセンター選定

- **NVIDIA DGX H100/HGX H100を国内初導入**
  - これまではA100を使っていたがLLMで分散学習環境が必要なためH100を導入
  - 既存DCでは様々な課題があった
- **新たにGPUサーバー用データセンターを選定**
  - 電力 -> 1ラック 35kVA以上
    - 消費電力: 定格12.5kVA(A100と比べて約1.5倍)
  - 冷却性能 -> リアドア空調DCを採用
  - ラックの拡張性 -> 最大200ラック
    - 既存DCの資産を移設

ネットワーク設計

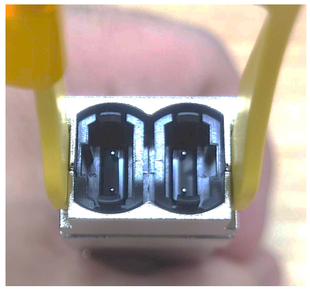
- **ラック構成はEnd of Row(EoR)を採用**
  - 1ラックあたり最大でGPUサーバー3台搭載となるためToR構成は収容効率が悪い
  - サーバーへの配線は全てパッチパネルを経由
- **要件の異なるネットワーク(LossyとLossless)を分離**



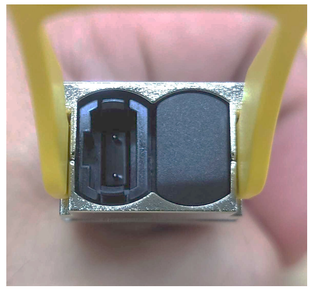
インターコネクト設計

- **構築・運用しやすい構成**
  - ネットワーク構成はFat-Treeを採用
  - BGP Unnumberedを採用
  - メンテナンス時はG-shut communityによる迂回
  - L2 延伸しない(L2は1台のLeafで閉じる)
- **フローの偏りの抑止**
  - LAGを使わない
  - Adaptive Routing によるフローの偏りを解消
- **RoCEv2への対応**
  - フルバイセクション
  - Lossless Ethernetへの対応
    - PFC、ECN(CNP)、ETS
  - 400GbEへの対応

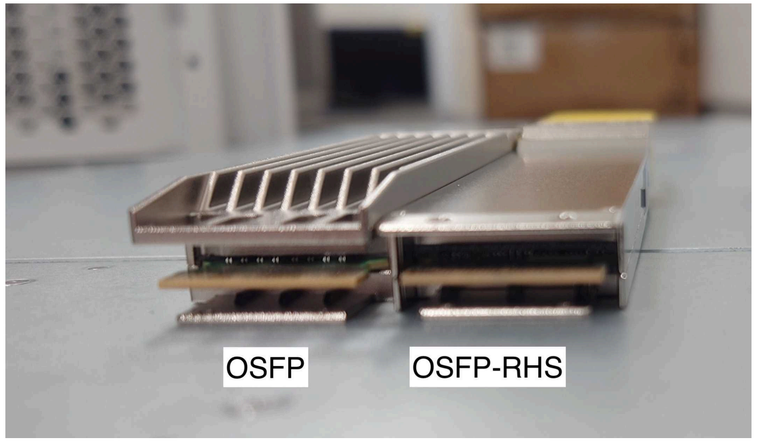
GPUサーバー用トランシーバーの選定



OSFP-RHS
400G x2



OSFP-RHS
400G x1



- **サーバーはConnectX-7で接続**
  - ConnectX-7の接続方式は2パターン
    - OSFP-RHS 800G (400G x2)
    - OSFP-RHS 400G x1
- **検証時に用意していた<u>トランシーバーが刺さらない問題発生</u>**
  - MSAのOSFP SpecificationでOSPF-RHSの存在を知る
    - OSPF-RHSはOSPF Riding Heat Sinkの略
  - OSPF-RHSはヒートシンクをNIC側に外付けしたもの
    - <u>Riding Heat Sinkだがヒートシンクが乗っているわけではない</u>
  - 他社では取り扱いが無いため純正トランシーバーを採用

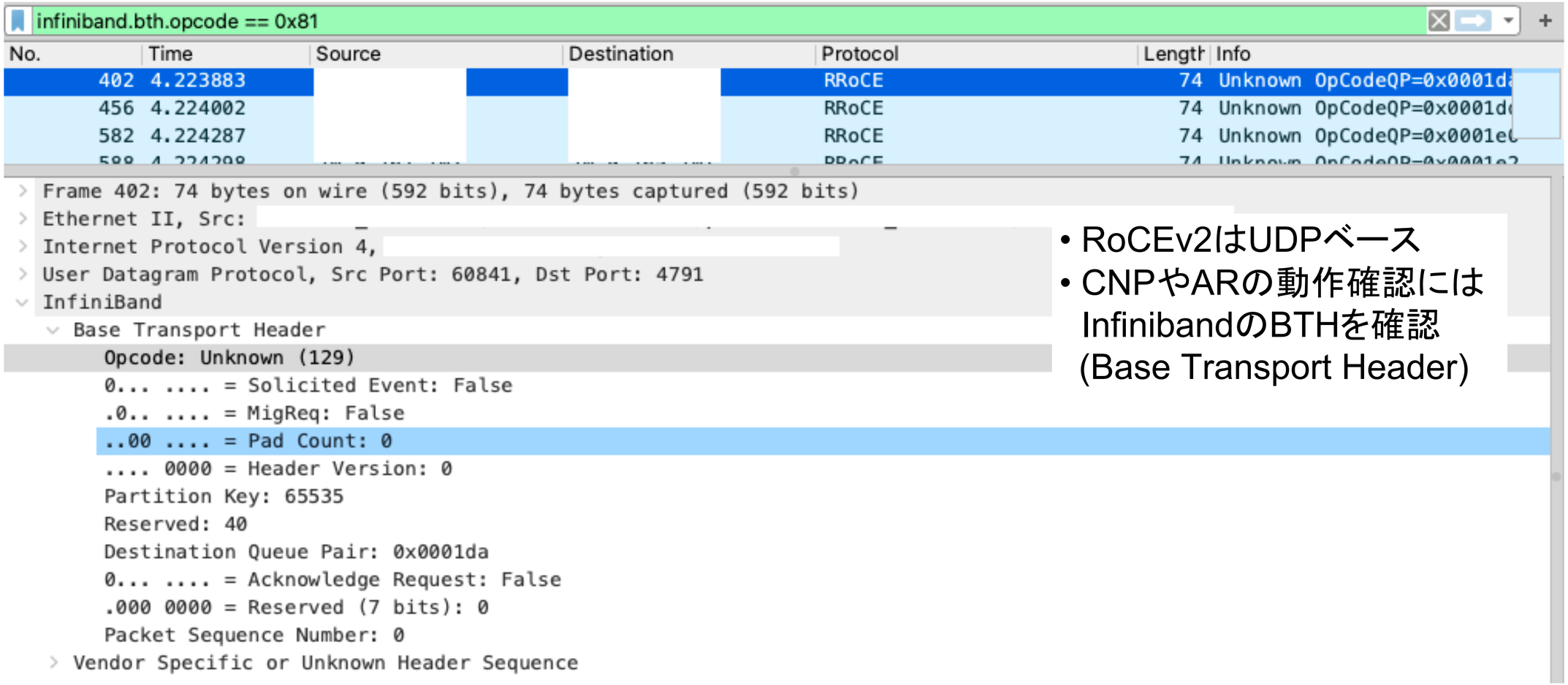
スイッチ用トランシーバーの選定

- **GPUサーバーとSW間は400G-DR4を選定**
  - GPUサーバー: OSFP-RHS / SW側: QSFP-DD
  - ファイバーはMPO-12/SMFを初採用
- **スイッチ間はAOCを選定**
  - EoR構成のためSpine-Leafの物理的距離が近い
  - トランシーバー x2 + MPOケーブルより安価に構築可能
- **400Gトランシーバーは純正を選定**
  - 3rd party トランシーバーでトラブルが発生したため
    - 純正と比べてLinkupまで時間がかかる(純正:20～50sec、3rd party:90sec)
    - shutdown -> no shutdown で Linkupしなくなる症状
  - スイッチ/NICで運用上必要な情報を取得出来るか確認
    - 光レベル、シリアルNo、温度

インターコネクト動作検証

- **RoCEv2**
  - L3ルーティング可能か/E2Eの帯域測定(400Gbps)/BGPによる迂回可能か
  - ETS設定時のQueueの優先度確認
- **Adaptive Routing**
  - Spine-Leaf間でflowが分散しているか
  - ON/OFF時のパケットの中身の変化を確認
- **ECN/CNP**
  - 輻輳発生時にECN/CNPによって輻輳制御が実施されているか
    - 検証時にCongestion Point(200GbE)を作成
  - CNPパケットの確認

RoCEv2のパケットフォーマット



- RoCEv2はUDPベース
- CNPやARの動作確認にはInfinibandのBTHを確認(Base Transport Header)

CNPのパケットをフィルタリングした結果

インターコネクトを構築した感想

- **Ethernetを選定したメリット・デメリット**
  - メリット: これまでのEthernetとほぼ同じ運用が出来る。BGPによる迂回など
  - デメリット: **RoCEv2にはInfinibandの知識が必須。QoSチューニングが大変。**
- **Lossyと分離したメリット・デメリット**
  - メリット: 要件が全く異なるトラフィックのQoSチューニングを考慮しなくていい
  - デメリット: Lossy/Losslessの構築・運用コストがかかる
- **400GbEを選択したメリット・デメリット**
  - メリット: Ethernetのボトルネックを気にしなくて良くなった
    - **AI/MLはサービス側から常に最大の性能を要求される**
  - デメリット: 規格と部材の選定の調査にコストがかかる

次に挑戦したいこと

- **800GbEの導入検討**
- **GPUサーバーの拡張検討**
- **水冷/液浸の導入検討**
- **Dragonfly トポロジの検証**
- **3rd party トランシーバーの導入**
- **GPUDirect Storage**
- **マイクロバーストの検知(Telemetry)**

ML Platformの詳細

https://cadc.cyberagent.co.jp/2023/sessions/distributed-ml-with-kubernetes/

まとめ

- **AI/ML時代の400G Lossless ネットワークを構築**
- **RDMAによりデータセンターネットワークの要件が変化してきている**
- **400GbEは規格が複数あり選定・検証が大変**
  - DR4だけでも大量に規格が存在: DR4、DR4+、XDR、DR4++
  - OSFP、OSFP-RHS、QSFP-DDの相互接続性
- **今までのネットワークエンジニアの守備範囲を超えた知識が必要**
  - Lossless、サーバー側のSR-IOVやCNIなどに加えて、AI/MLのワークロード側の知識も必要

良いものを作るには全てを外注せずに社内のエンジニアで構築することが大事

議論ポイント

1. **これからのデータセンターネットワークの要件の変化にどのように対応するべきか**
2. **ネットワークとコンピューティングの情報のキャッチアップ方法**
3. **GPU/HPC向けデータセンターを運用・検討している方の意見・感想の共有**

その他気になることがあれば質問お願いします

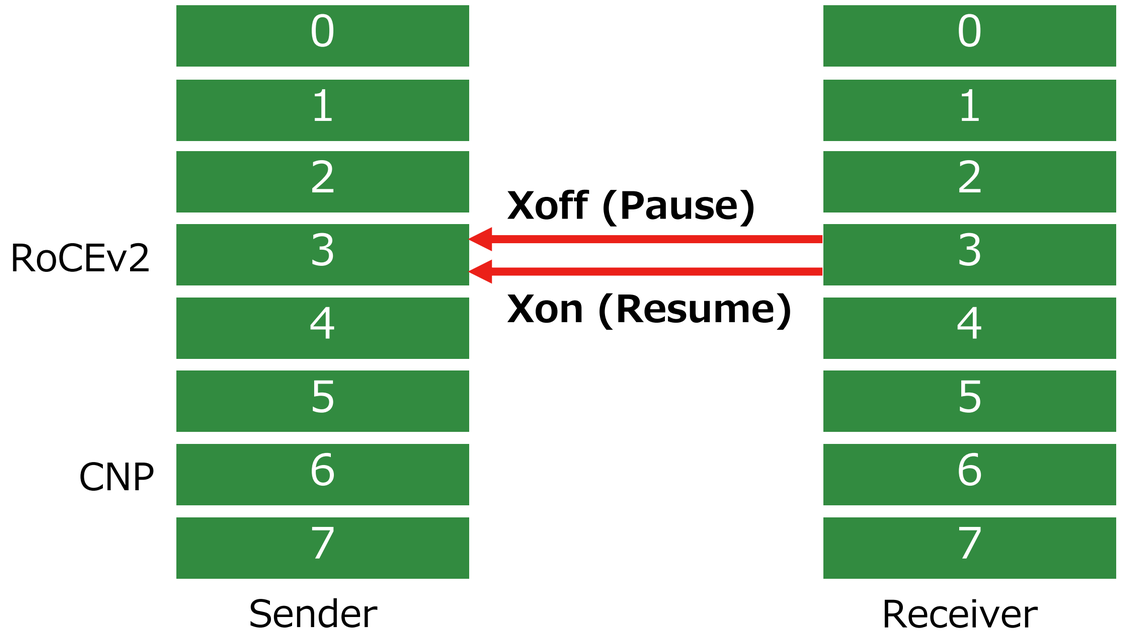
APPENDIX

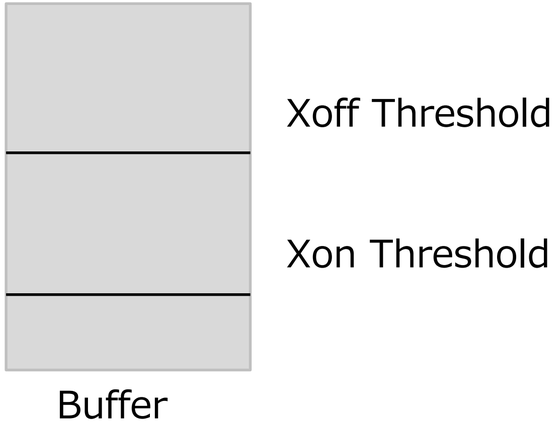
PFC (Priority Flow Control)

- **VLAN CoS(PCP)またはIP ToS(DSCP)を参照しMapping tableでパケットを分類**
- **輻輳時にキュー毎にXoff(一時停止)、Xon(送信再開)をPauseフレームで制御**
- **閾値を超えたらXoff、閾値を下回ったらXonを送信**

CoS	DSCP	Priority
0	0-7	0
1	8-15	1
2	16-23	2
3	24-31	3
4	32-39	4
5	40-47	5
6	48-55	6
7	56-63	7

Mapping table

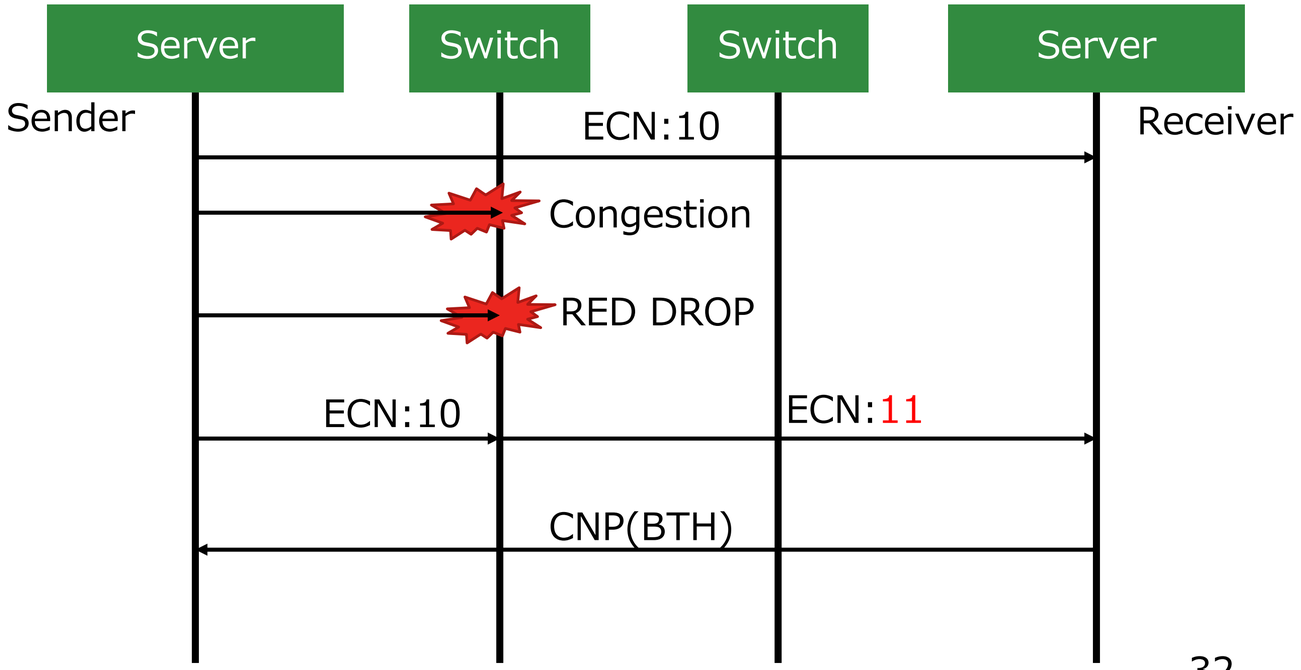




ECN (Explicit Congestion Notification)

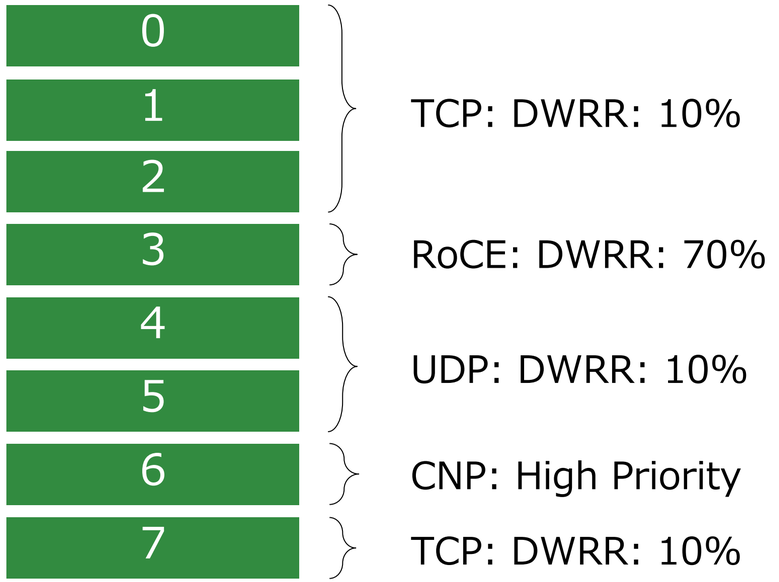
- **IP ToS(ECN)の2bitで輻輳制御を実施**
- **輻輳を検知したらRED (Random Early Detection): [REDはoptionであることが多い]**
- **輻輳時は全てのパケットにECN CE(11)をマーキング**
- **ReceiverはSenderにCNPを送信、Senderは送信レートの制限を実施**
- **エンドノード間の帯域制御によって輻輳を根本的に解決**

bit	Description
00	ECN非対応
01	ECN対応 ECT[1] (CNPの時に利用/CNPはBTHで定義)
10	ECN対応 ECT[0] (一般的に利用)
11	輻輳発生 / CE

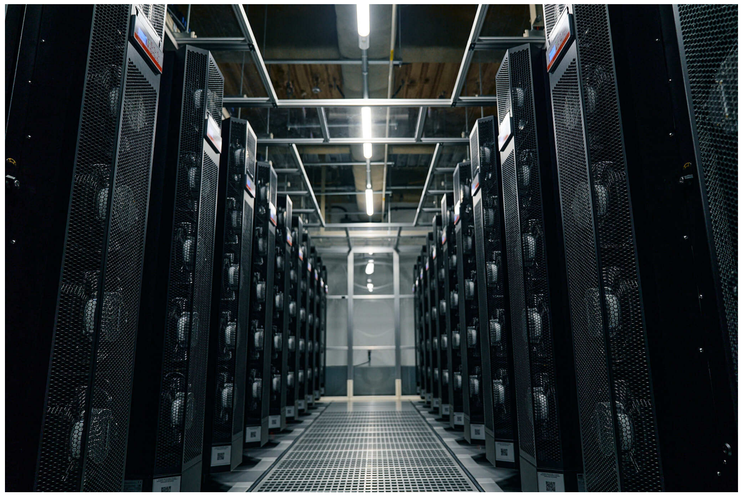


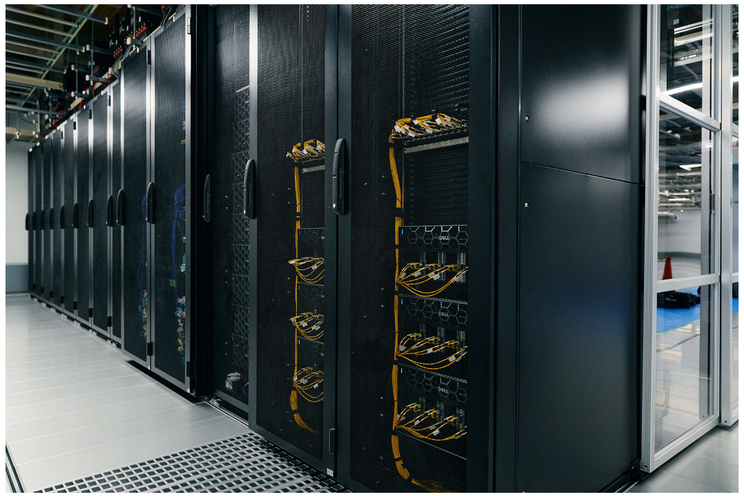
ETS (Enhanced Transmission Selection)

- **データ転送側のキューの制御**
- **各PriorityGroup毎にQoSを使いキューの優先処理を決定する**
- **RoCEv2やCNPのキューの優先度を上げてAI/MLワークロードに最適なチューニングを実施**



リアドア空調ラック





- **空冷で高負荷サーバーを高密度に設置可能**
  - リアドア(ラック背面)に冷却機器を搭載
  - ラック毎に冷却制御可能
- **フロントドアを拡張**
  - QSFP-DD/OSPFトランシーバーに合わせて拡張
- **作業時にイヤーマフ必須**
  - リアドアのファンが高速回転しているため騒音が大きい
    - 約100dB

FIN