Rack-Scale GPUサーバーの NW設計と運用までの苦悩

ソフトバンク株式会社 (SoftBank Corp.)

内田 泰広 (Yasuhiro Uchida)

張 朝程 (Chaocheng Chang)

自己紹介

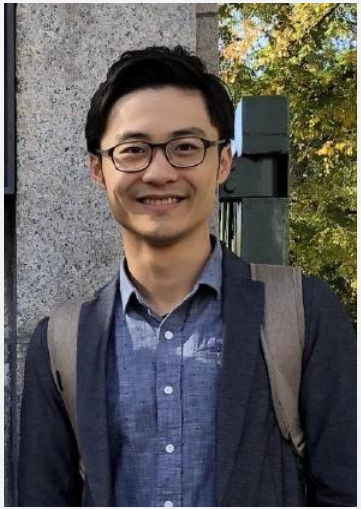
内田 泰広 (Yasuhiro Uchida)

ソフトバンク株式会社
AI&HPCインフラ統括部 AIネットワーク開発部
ネットワークアーキテクト

経歴
LBのプリセールスエンジニア
AS運用、プライベートクラウド、AI計算基盤
2024/10月より現職

担当業務
AI計算基盤のNW設計がメイン
Scale-Up/Scale-Out NW、Facility周り(冷却/電源)を勉強中

自己紹介



張 朝程 (Chaocheng Chang)

ソフトバンク株式会社
AI&HPCインフラ統括部 AIネットワーク開発部
ネットワークエンジニア

経歴
台湾出身
iOS ソフトウェアエンジニア
ホワイトボックススイッチメーカー SE担当
2025/1月より現職

担当業務
AI計算基盤のNW構築
NW機器の技術調査・検証 (Optics/Cable/Switch)

SoftBankのAI計算基盤の歴史

2023年からAI計算基盤の稼働を開始
2025年 スーパーコンピューターTOP500で世界17位(日本3位)を獲得

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
7	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
16	**ABCI 3.0** - HPE Cray XD670, Xeon Platinum 8558 48C 2.1GHz, NVIDIA H200 SXM5 141 GB, Infiniband NDR200, Rocky Linux 9, HPE National Institute of Advanced Industrial Science and Technology (AIST) Japan	479,232	145.10	181.49	3,596
17	**CHIE-4** - NVIDIA DGX B200, Xeon Platinum 8570 56C 2.1GHz, NVIDIA B200 SXM 180GB, Infiniband NDR400, Ubuntu 20.04.2 LTS, Nvidia SoftBank Corp. Japan	662,256	135.40	151.88	

TOP500 November 2025 (https://top500.org/lists/top500/2025/11/)

NEW

2023 NVIDIA A100

2024 NVIDIA H100

H1 2025 NVIDIA B200

H2 2025 NVIDIA GB200 NVL72

ソフトバンクとして初めての挑戦

「NVIDIA GB200 NVL72」を搭載した AI計算基盤が稼働開始

GB200 NVL72

01. サーバー
Rack-Scale GPUサーバー

02. 排熱方式
液冷対応 (Liquid to Liquid)

03. Scale-Out (Rack間通信)
Ethernet Switch / RoCEv2

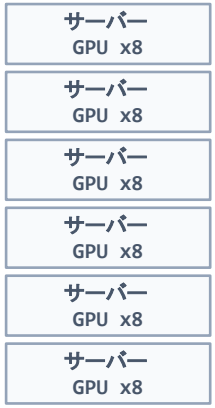
04. Scale-Out (Rack間通信)
EVPN/VXLAN

05. Scale-Up (Rack内通信)
NVLink Switch

06. Facility
OCP ORv3 Rack
RDHx/PowerShelf etc..

Rack-Scale GPUサーバー導入により変わったNW要件

従来のGPUサーバー



設計単位: 1GPUサーバー

空冷でRU数が増える

電源、冷却、運用設計が確立

1台 ~15kW

NW設計: Rack内、Rack間接続

Rack-Scale GPUサーバー

設計単位: 1Rack

Rack内で機器同士が密接に繋がる

電源、冷却、運用設計を再定義

1Rack 100kW~以上

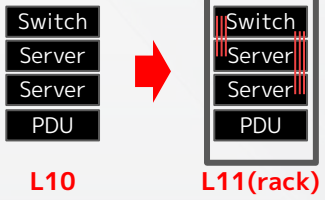
NW設計: 従来+Facility全体

Rack-Scale GPUサーバーの構成を理解しないと設計ができない

引用: NVIDIA DGX GB Rack Scale Systems User Guide
https://docs.nvidia.com/dgx/dgxgb200-user-guide/index.html

Rack-Scale GPUサーバーの物理アーキテクチャー (Rack)

ORv3 Rack(21inch)
【※画像はMGX Rack: 19inch】



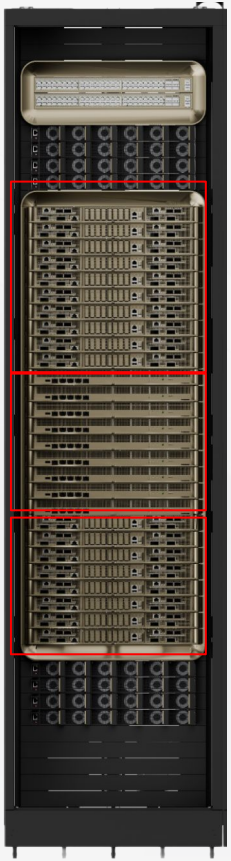
OCP ORv3 Rack 横幅 21inch

- **L11納品 (Rackに搭載・配線・検証された状態で納品)**
  - 組み上げ前に配線を確認
- 従来のRack: EIA 19inch
- NVIDIA規格 MGX v1.x (横幅 19inch Busbar)

Busbar

- **各機器に対して電源を供給する**
- 電圧 +48V
  一般的な19inch DC電源スイッチは-48V
  => DC-DCコンバータ + 電源Bar clipが必須

Rack-Scale GPUサーバーの物理アーキテクチャー (Tray)

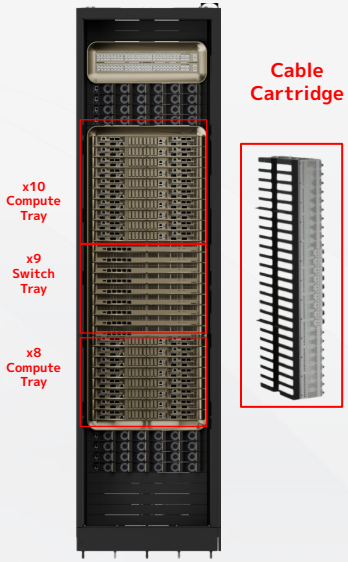


Compute Tray x18 (10 trays + 8trays)

- **Tray単位でOSが起動している**
- **各Tray B200 GPUが4枚搭載**
- OS側はBlueField3x2がInbandに接続される
- BMCはOOB接続

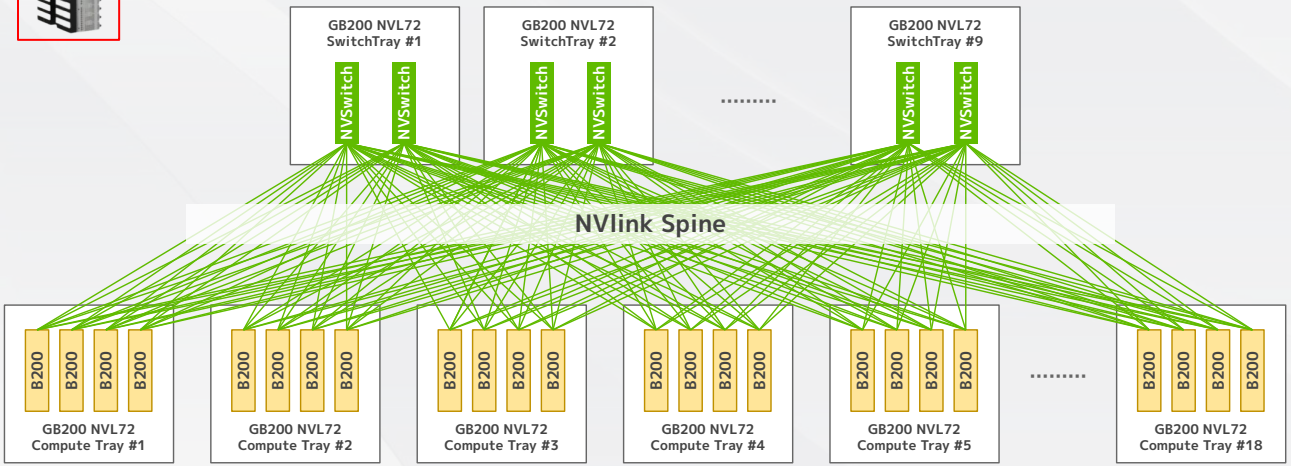
⚠️ **Rackの下が若番なのか、上が若番なのかは事前に確認が必要**
※認識の差異があると納品後に配線変更が発生

Rack-Scale GPUサーバーの物理アーキテクチャー (Tray)



Switch Tray x9

- Compute TrayをRack内で高速に繋ぐスイッチ
- 各Tray NVSwitchが2枚搭載
- Cable Cartridge(NVLink Spine)経由でCompute Tray間を接続



Rack-Scale GPUサーバーの物理アーキテクチャー (Facility)

PowerShelf

- **施設側電源を変圧しBusbarに電力を供給**
- **集中電源とも呼ばれる**
- **PowerShelf筐体の冗長**
  - **PowerShelf内のPSUも冗長(2N)**
- 電圧コンバータ (交流AC -> 直流DC48V変換)
- Busbarの電圧降下を考慮し搭載位置を決定
- 施設側電源の入力冗長も確認
- 機器毎の小型PSUより変換効率が良い

Rack-Scale GPUサーバーの物理アーキテクチャー (Facility)

CDU (Coolant Distribution Unit)

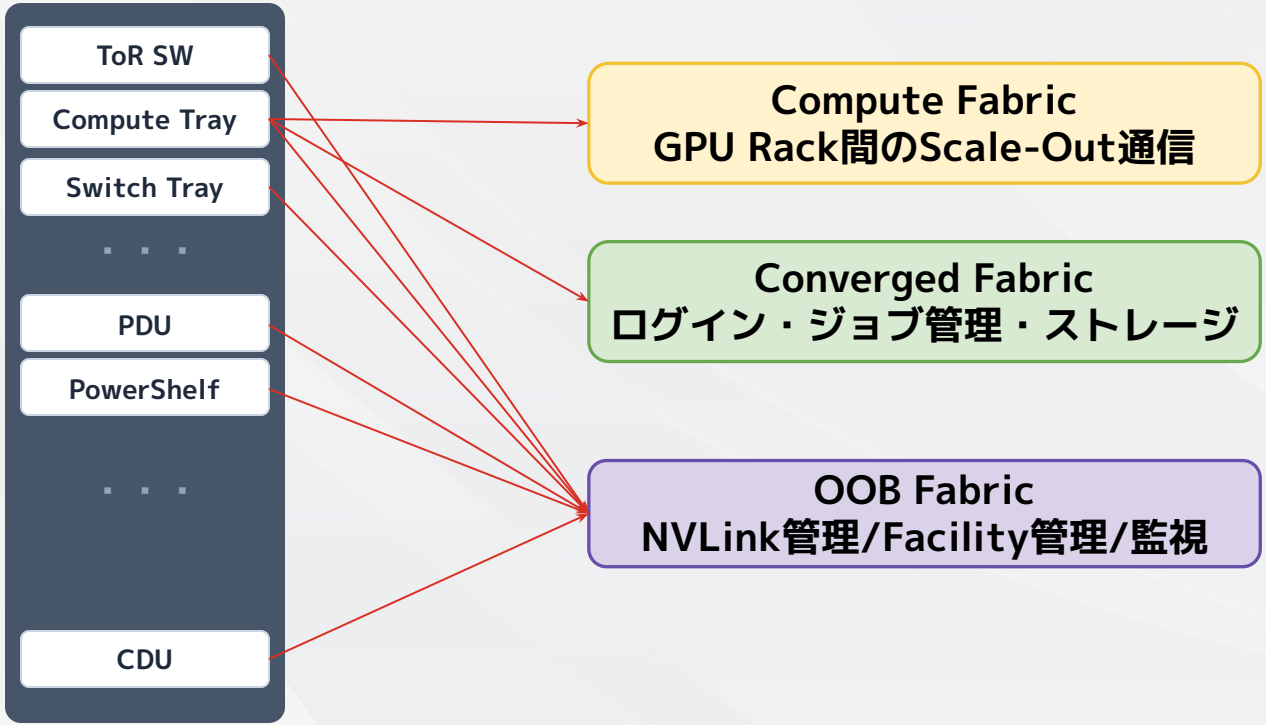
- **Rack内の冷媒を循環・制御する装置**
  - 二次冷却側の冷媒はPG25
- 冷却ループを制御・監視 (温度、流量、圧力、漏水)
- サーバー側(TCS)で温まった冷媒の熱を設備側(FWS)へ受け渡す
- CDUは監視・障害切り分けのためOOB接続の考慮が必要

Manifold

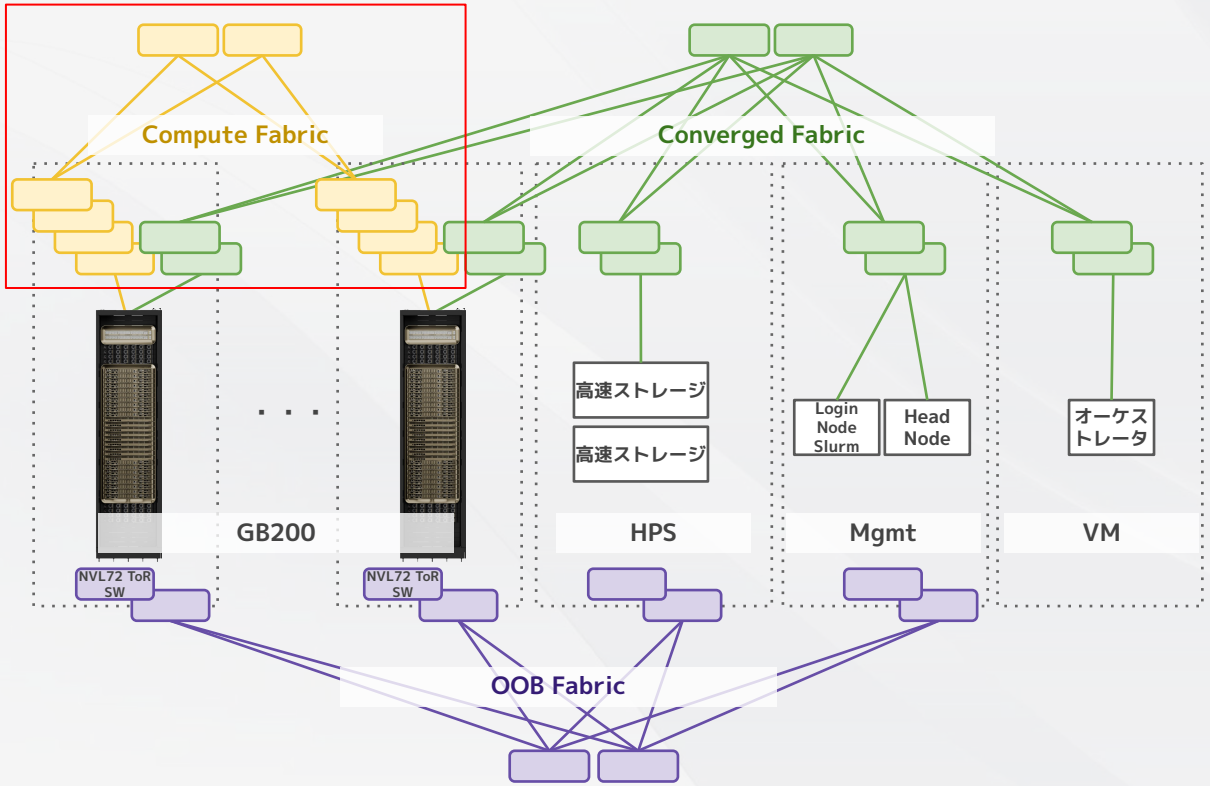
- **Rack背面にある配管**
- **CDUから供給された冷媒を各Trayのコールドプレートへ分配する**
- OCP規格 UQD,UQDB プラグ/ソケット
  - UQDB04はBlind mateのため、接続部を見なくても接続可
  - UQD04とUQDB04は別物: 互換性があるかは要確認
  - UQDB04とUQDB02,UQDB06は径が違うので接続不可
  - UQDv2でUQDとUQDBの統合をOCPで議論中
- NVIDIA規格 MQDプラグも存在

Rackに接続するNW Fabricの種類

Rack内の構成を理解出来たので、ネットワーク設計が出来る



AI計算基盤のNW構成：Compute Fabric

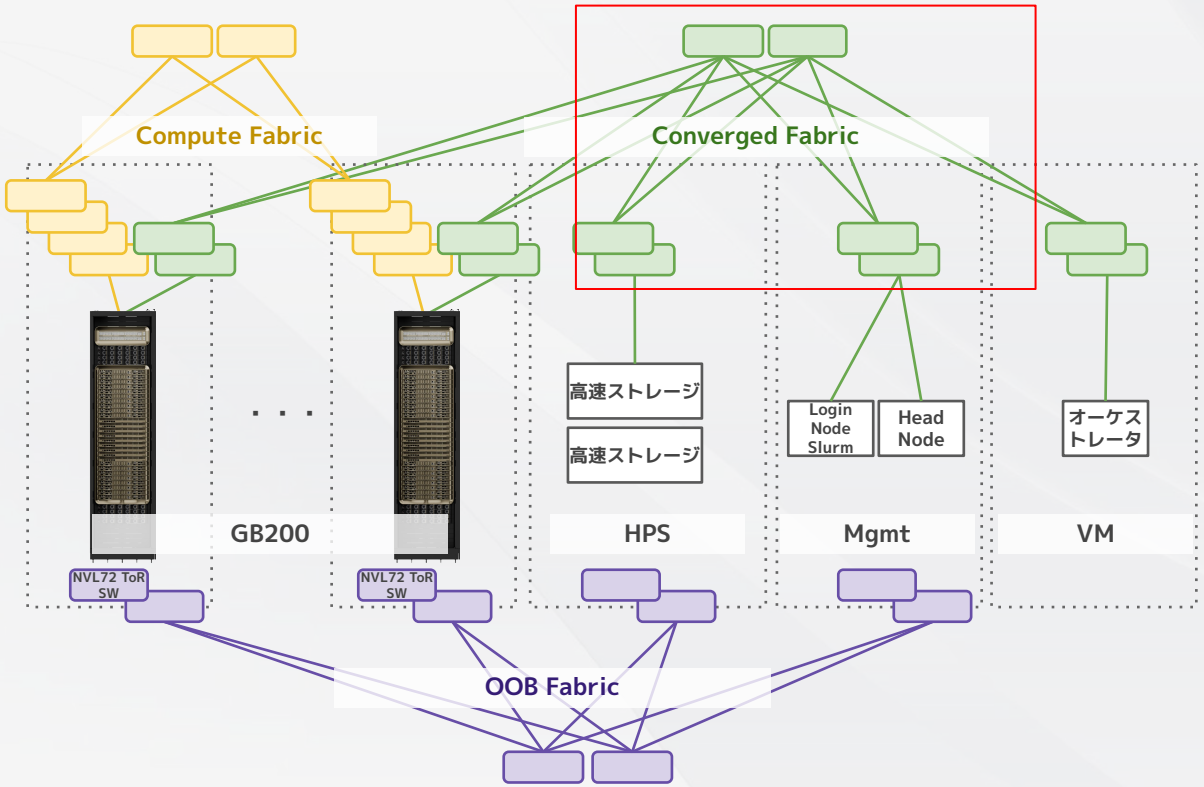


Compute Fabric

Back-end NWとも呼ばれる

- **GPU間を高速で通信する**
- **GPU-Leaf間の接続に工夫**
- マルチテナント
- Scale-Out NW
- 閉域な接続 (クローズド)
- GPU間を高速/高帯域で通信
  - RDMA(RoCEv2)
- BGP unnumbered
- BGP W-ECMP(Weighted ECMP)

AI計算基盤のNW構成：Converged Fabric



Converged Fabric (Inband)

Front-end NWとも呼ばれる

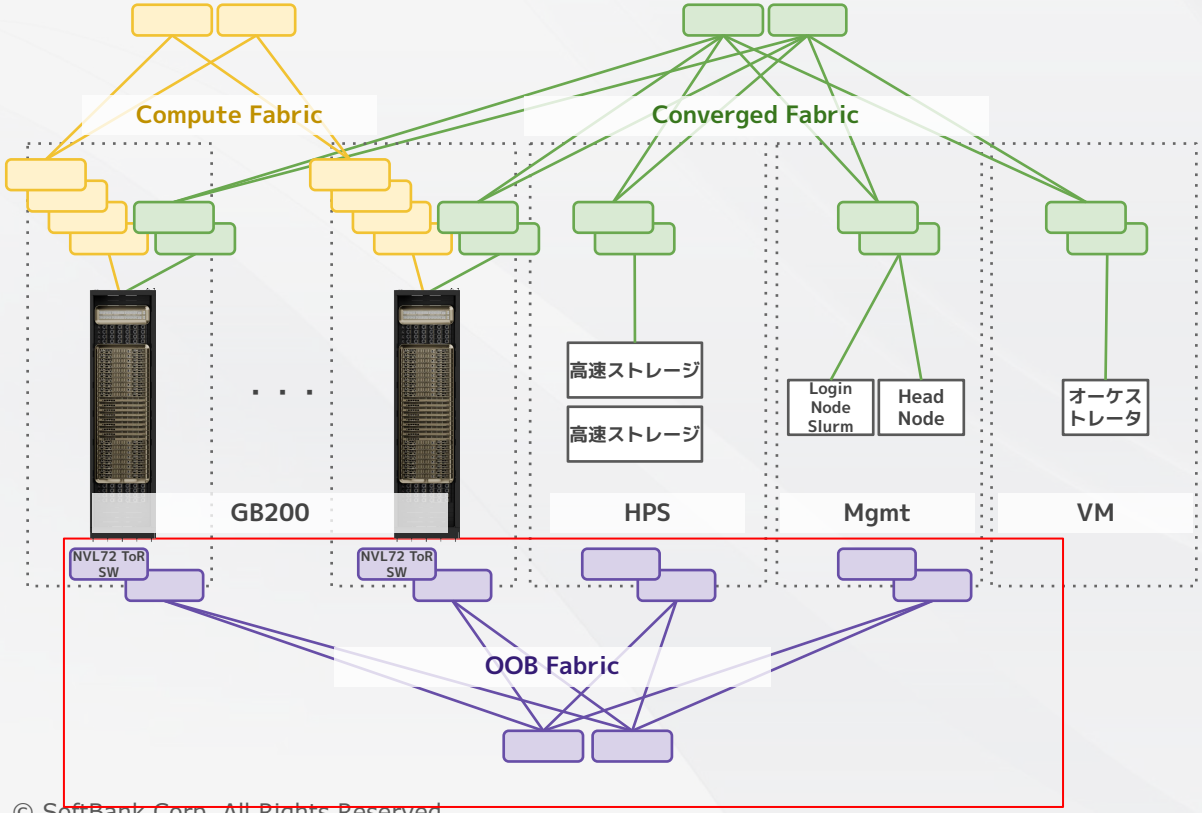
- **インターネットアクセス**
- **CPU主体の通信で使われる**
- マルチテナント
- NCCLのBootstrap通信

Converged Fabric (Storage)

学習に使うデータを保存

- **高速ストレージと通信するNW**
- **GPU1枚当たりのread/write性能を元に帯域設計**
- マルチテナント
- 学習中のスナップショットを保存
- RoCEv2

AI計算基盤のNW構成：OOB Fabric



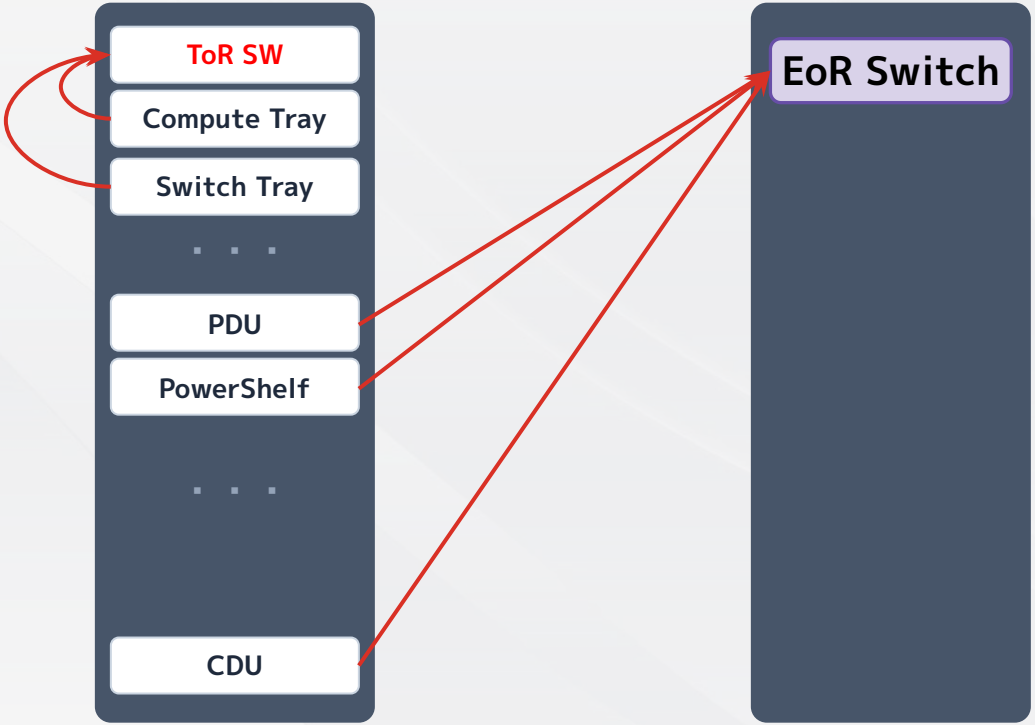
OOB Fabric

スイッチ・サーバーの管理系通信

- **Facility機器の監視**
  - CDU / RDHx / PowerShelfの監視
- **NVSwitchの管理系通信**
- **Rack-Scaleの場合、設計が重要**
- サーバー管理系
- マルチテナント
- NVL72 ToR SW

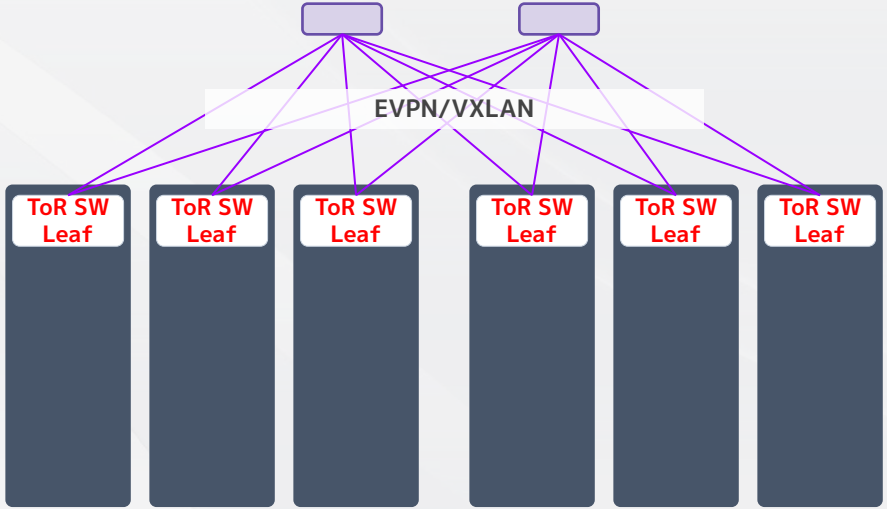
(参考) AI計算基盤のNW構成：OOB Fabric

Facility系機器(CDU/PowerShelf)はToR SWに接続するのではなくEoR SWへ収容
漏水・Rack障害時、ToR SWの二重障害時の切り分けを考慮



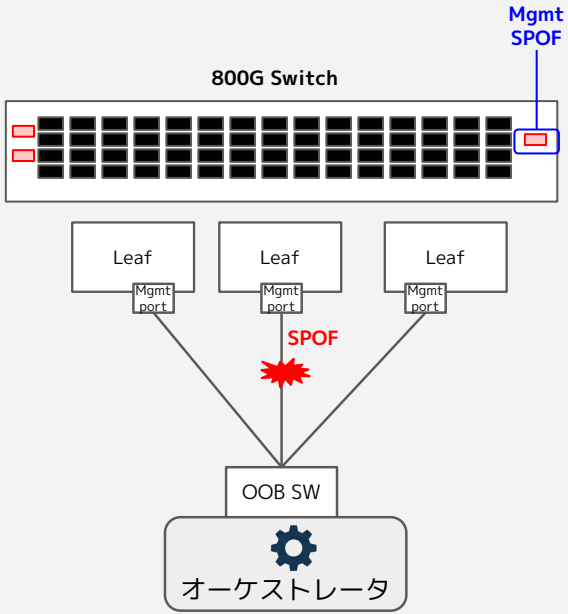
AI計算基盤のNW構成：OOB Fabric

ToR SWでマルチテナント(EVPN/VXLAN)したい時はスイッチ選定が重要。
マルチベンダでEVPN/VXLANをしたくない、L2 NWを作りたくない
※低速なスイッチのため高機能ではない可能性あり



運用系通信の改善

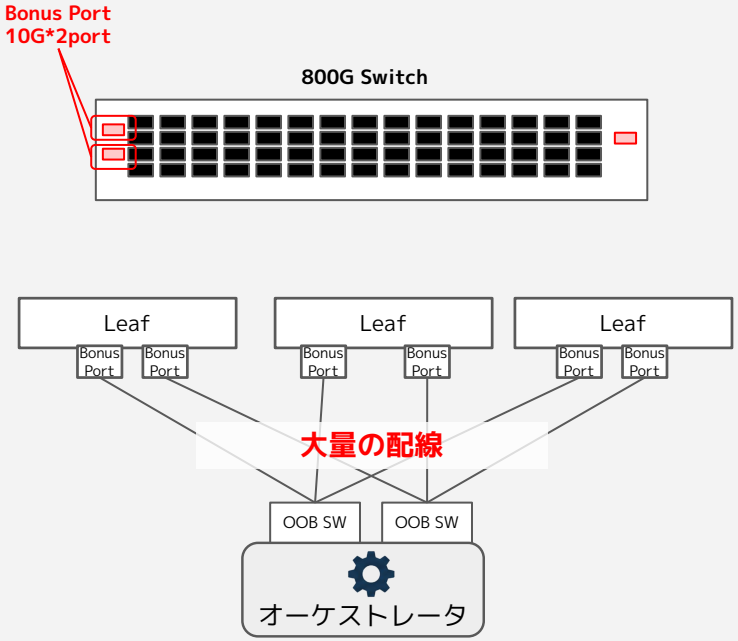
【運用系通信】Compute Fabricを使った到達性の冗長化



オーケストレータからLeafスイッチに設定を入れたい

- Mgmt port経由の通信は**単一障害点(SPOF)**
- Compute FabricはクローズドなNW
  - In-bandでの管理制御が難しい
- スイッチの管理系通信は高可用にしたい

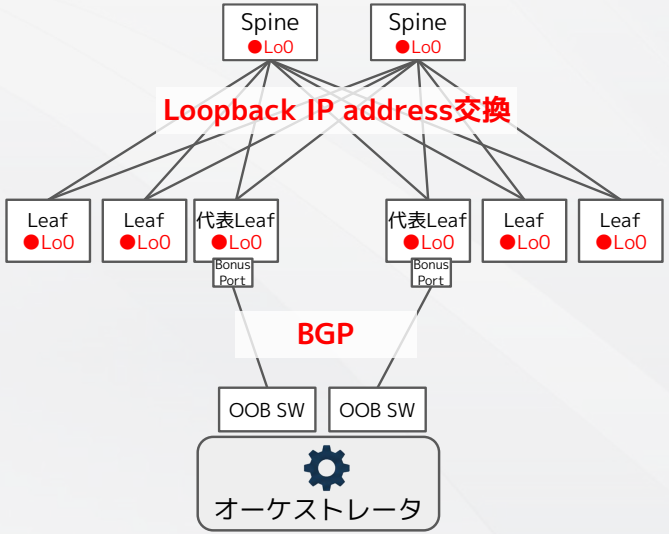
【運用系通信】Compute Fabricを使った到達性の冗長化



対策1: Bonus Portの活用

- 800Gスイッチは低速ポート(Bonus Port)が存在
- Bonus Portに2本ずつケーブルを結線し冗長性を担保できる
  しかし多くのOpticsとLANケーブルが必要
  Bonus Portが1 portしかないスイッチが存在し再現性がない
- 例: Leafが100台の場合
  - LANケーブル200本、Optics400本

【運用系通信】Compute Fabricを使った到達性の冗長化



対策2: Underlay Networkの活用

- Compute FabricはBGP Multipathで構成

1. UnderlayでLoopback IPアドレスを交換させる
2. 代表LeafのBonus Portに接続
   低コストでLo0到達性を冗長化

- **例: Leafが100台の場合**
  - **LANケーブル2本、Optics 4本**

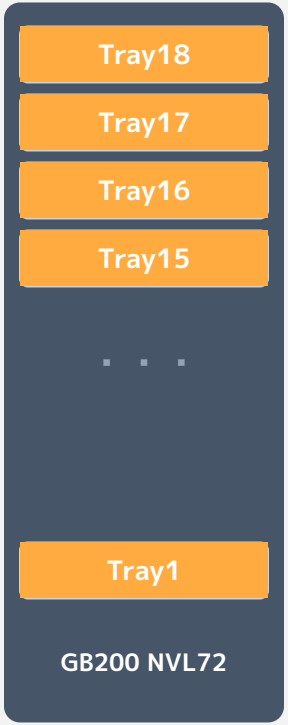
※運用時の到達性改善として利用。OOB接続の代替ではない

実際のサービス提供はどうするのか

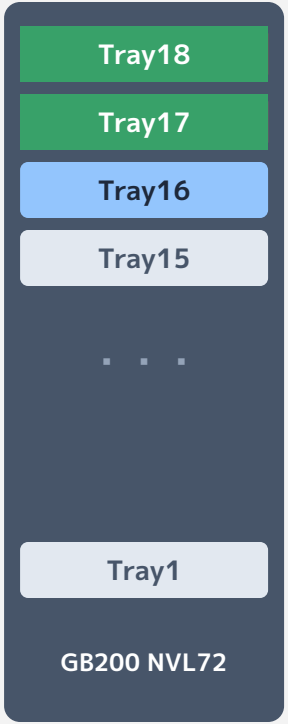
GB200 NVL72のサービス提供形態

どのサービス提供が適切であるか、サービスごとの特徴はなにか

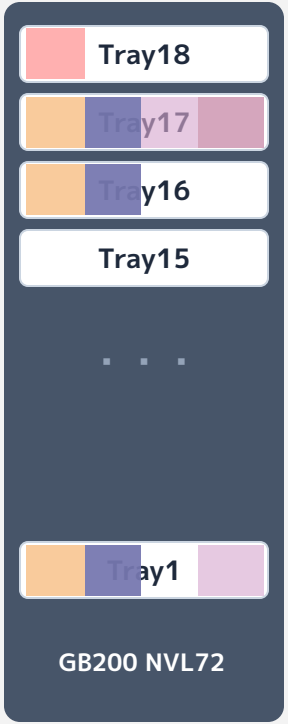
Rack単位の提供



Tray単位の提供

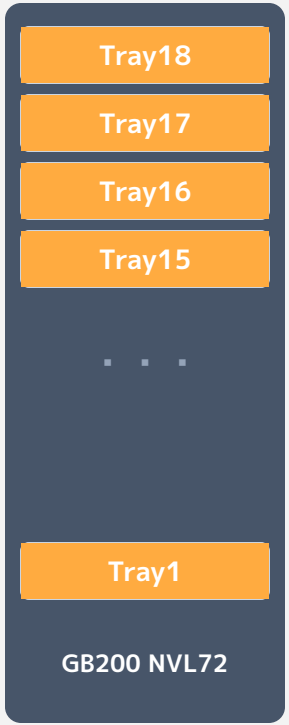


GPU単位の提供



サービス提供に共通する課題

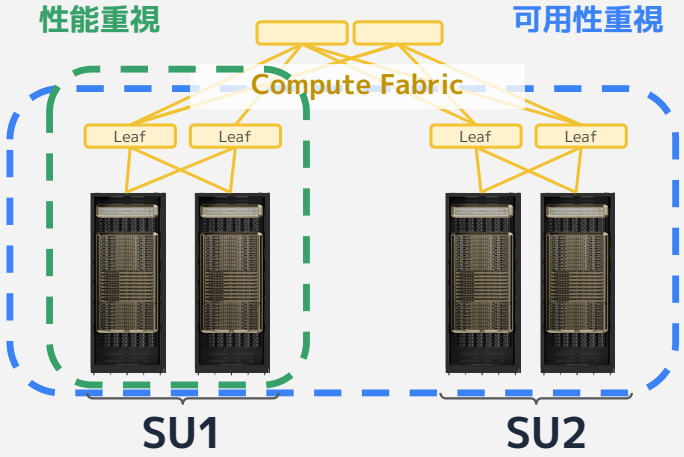
Rack/Tray/GPU単位の提供



共通課題

- 交換用Rackを用意する場合は大きなコストが発生
  - 検証環境やラボの用意も高価
- Rack内の単一障害点での障害影響が大きい
  - CDU障害は配管まで作業範囲が広がると長時間かかる

サービス提供に共通する課題



Scalable Unit (SU):
複数のNVL72Rackをまとめた単位・障害ドメイン

GPUアサインにおけるCompute Fabricのトレードオフ

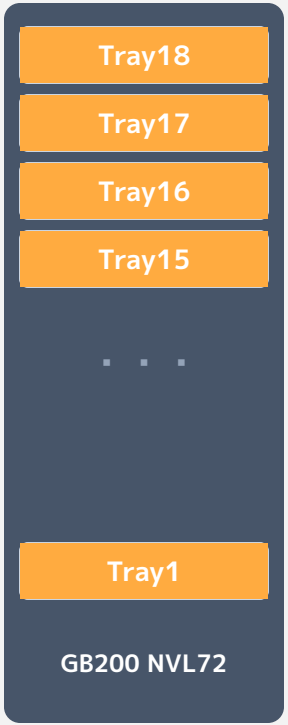
- 性能重視: 同じSU(障害ドメイン)にユーザーをアサイン
- 可用性重視: 別のSUにユーザーをアサイン
- ユーザーのGPUアサインはRackの位置情報の考慮が必要

OOB Fabric

- NVSwitchを提供する場合はマルチテナントが必須
  - NVL72のToRスイッチもマルチテナント
- BMCのユーザー提供の検討 (提供した場合は監視ができない)

Rack単位のサービス提供の特徴

Rack単位の提供

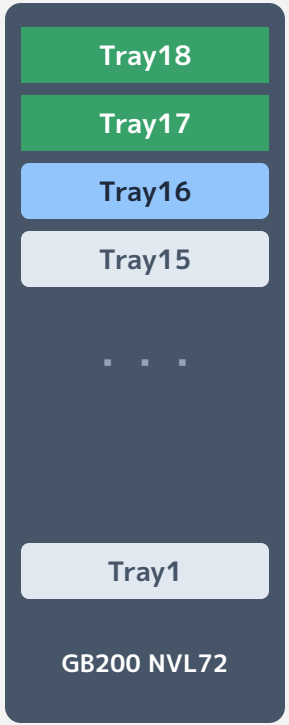


Rack単位提供の特徴

- NVL72 1Rack全体を利用
- 大規模ユーザーの利用を前提
- 構成がシンプル (Compute Trayの論理分割が不要)
- 他の提供形態と比べて一番容易にサービス提供が可能

Tray単位のサービス提供の特徴

Tray単位の提供



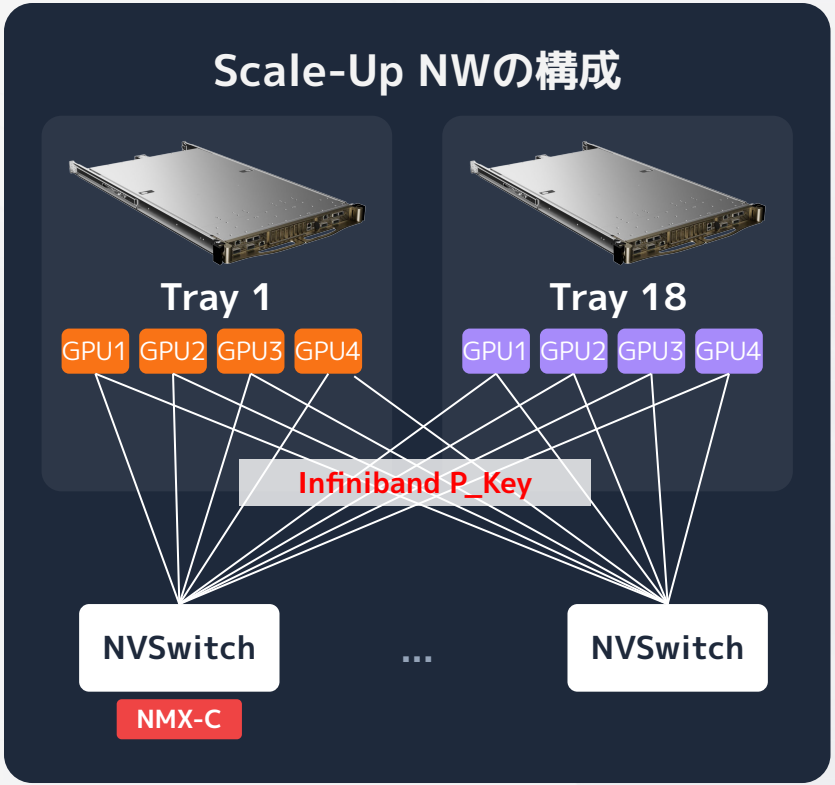
■Tray単位提供の特徴

- NVL72をCompute Tray単位で提供する
- 中規模ユーザーの利用を前提
- Compute Tray同士がNVSwitchで繋がっているため
  Tray間の論理分割が必要
  - テナントごとにNVLink Partitionを分離する必要あり

課題・考慮事項

- NVSwitchの管理プロセスはRack内で共通のため単一障害点
  - SLAを高めるためにはRack/SU単位の冗長が必要

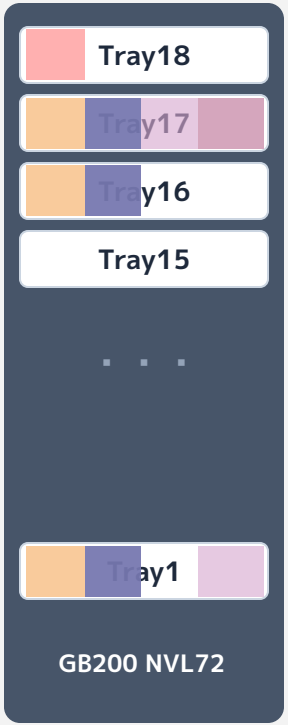
【Scale-Up NW】NVLink Partition Management



NVLink Partition Management

- **NVSwitchのマルチテナント機能**
  InfiniBand P_Keyを使ったマルチテナント
- **CLIまたはAPIを使ってSwitch操作**
  ⇒ VLAN操作のようにPartitionを分離
- **クラスタ管理プロセス**
  代表Compute Trayでプロセス(NMX-C)が稼働

GPU単位の提供



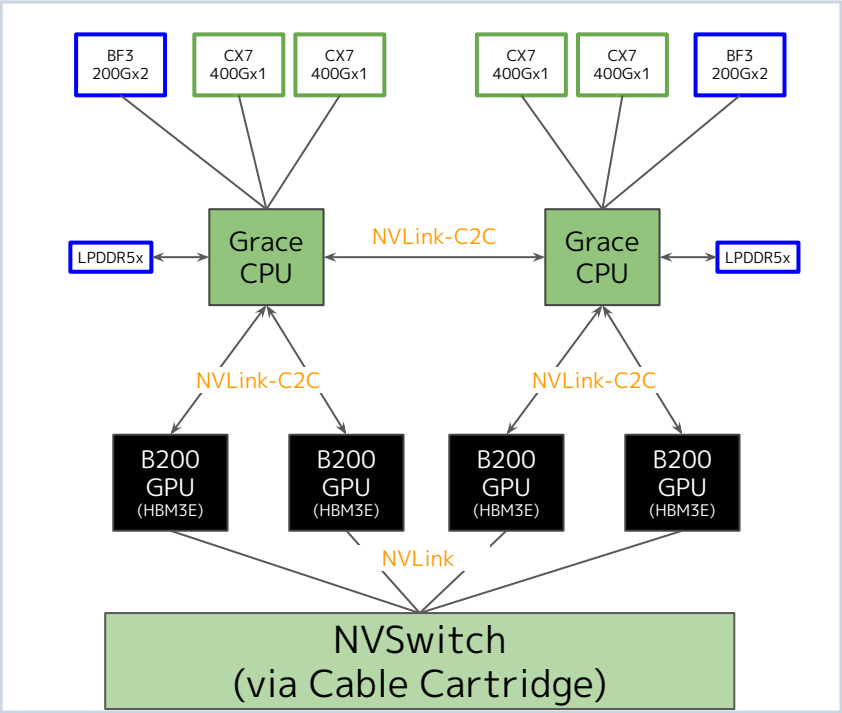
■GPU 単位提供の特徴

- NVL72をGPU単位で提供する
  - 理論上、1Rack最大72ユーザー
- 小規模ユーザーの利用を前提
- NVL72のRack-Scale性能を活かしきれない

■論理分割と仮想化要件

- **Tray間/GPU間(コンテナ)の論理分割を考慮する必要あり**
- VMの場合、PCIe Passthrough 等の仮想化方式を検討

GPU単位の提供の考慮事項



■ CPU・GPU間接続とメモリ制限

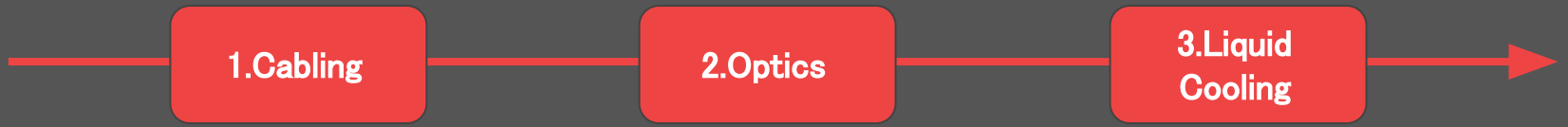
- **B200 GPUを1枚単位で切れても、**
- **Grace CPU・メモリ空間は同じ粒度で分割できない**
- Grace CPU : B200 GPU比 = 1:2
- コヒーレントメモリの考慮
  ※Grace CPUのメモリとB200 GPUのHBMを
  CPU/GPU双方から一貫性を保って扱える
- **対応コストが高い**

提供形態のサマリ

ソフトバンクではトレードオフを踏まえ、現時点ではRack単位とTray単位の提供

	Rack単位	Tray単位	GPU単位
論理分割	Compute Fabric Converged Fabric OOB	Rack単位 + Scale-Up(NVlink Partition)	Tray単位 + B200 GPU/ Grace CPU/ Memory etc..
対象顧客	大規模	中規模	小規模
対応コスト	小	中	大 テナント分離でリソース消費大
予備品の準備	予備ラックの準備	予備ラックの準備 予備Trayの準備	Tray単位とほぼ同様
障害の特徴	単一障害点が多い Compute Fabric/ CDU/NVSwitch/ Compute Tray/Busbar	Rack単位とほぼ同様 別Compute Trayの障害影響はなし (NVlinkはPartitionが別の場合)	Tray単位とほぼ同様

後半: AI計算基盤の構築時の苦労話



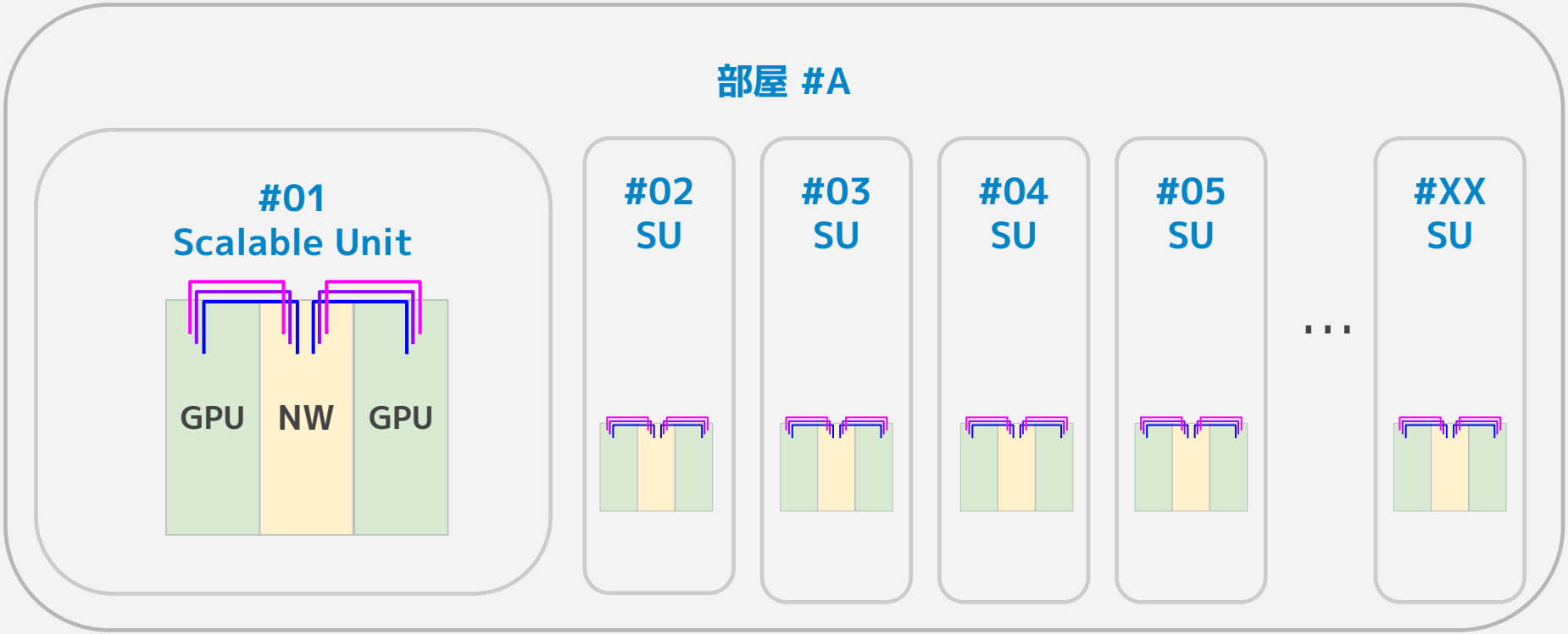
Cabling

大規模AI計算基盤の構築は物理工事に大幅な時間がかかる

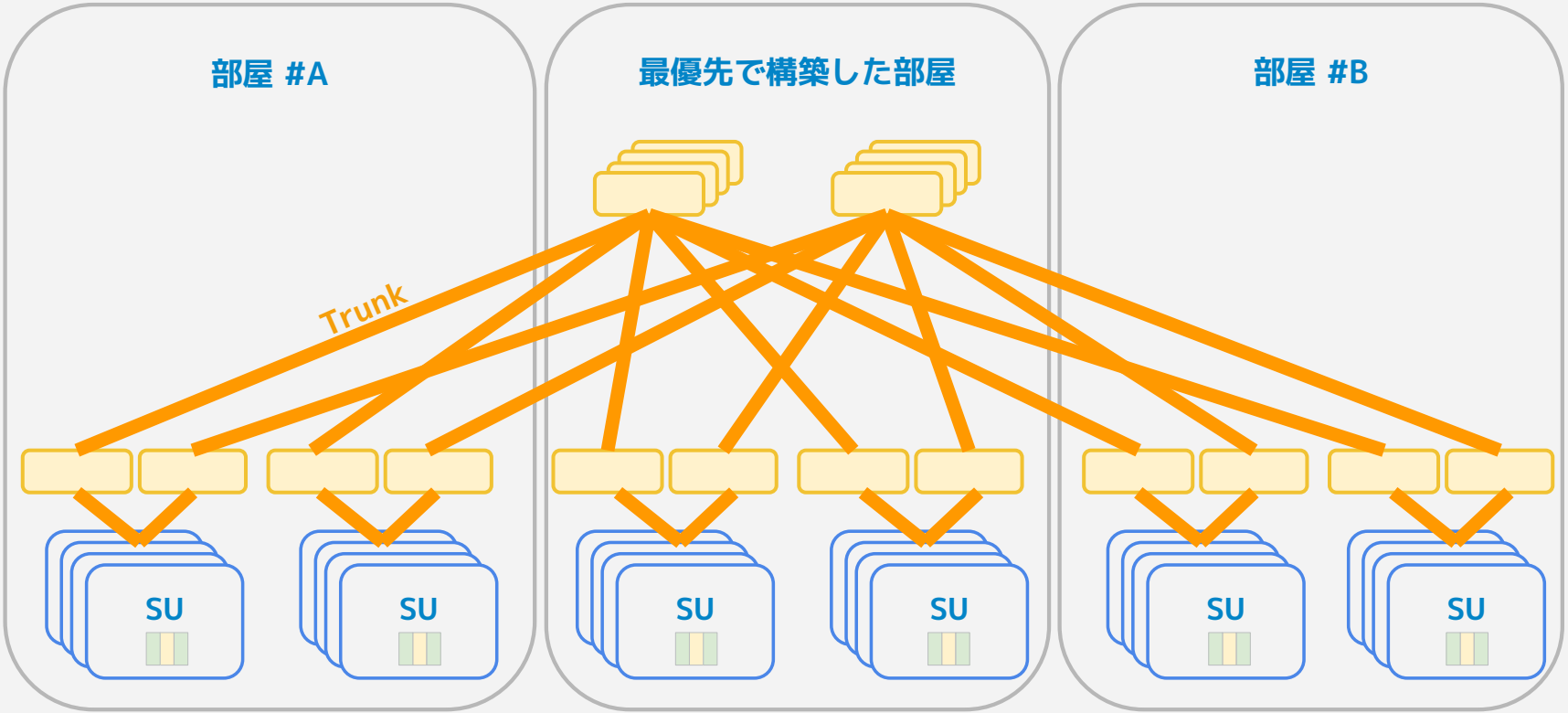
短納期で 完遂すること	膨大な結線表の 作成・修正	自動化のしやすさ

シンプルな物理構成をコンセプトに設計

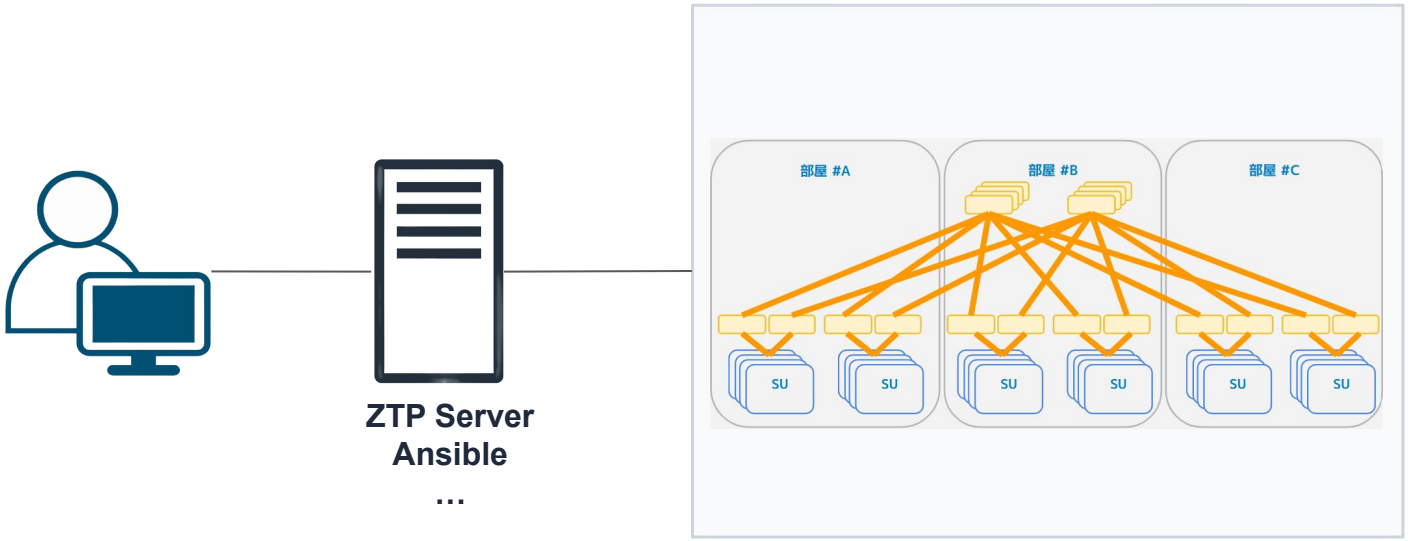
案①: 部屋内の機器の対称



案②: 部屋間の機器の対称

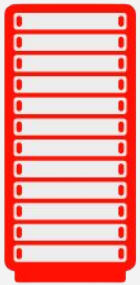


POD型の対称構成により
大規模AI計算基盤を短納期で効率的にデプロイ



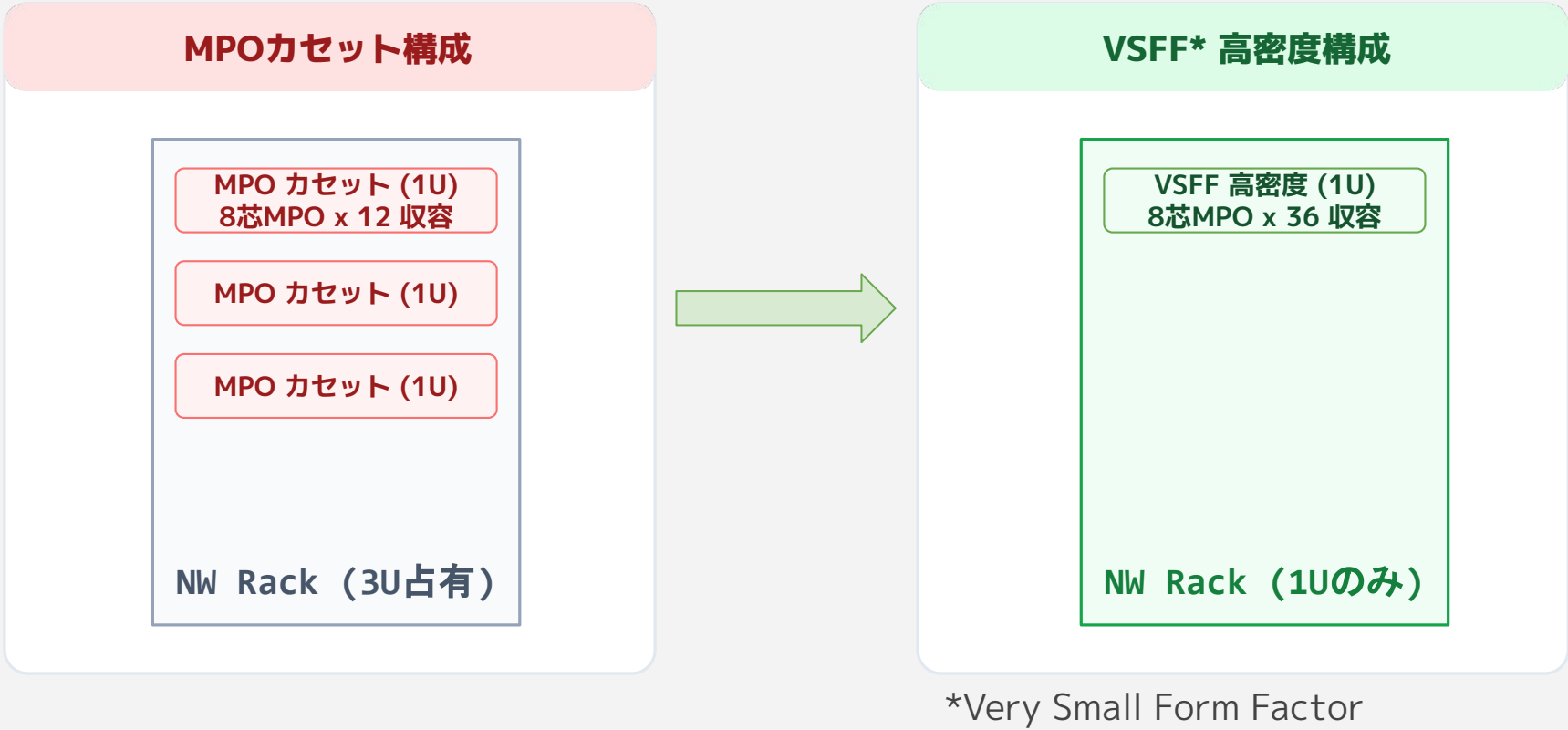
省スペース化とコスト削減の両立が求められる

ラック内の搭載スペースが
限られている



ケーブリング関連部材の
調達コストも高い

【ケーブリング】パッチパネルの例



*Very Small Form Factor

VSFF の選定で、省スペース化とコスト削減を実現

Optics

数万本の Interface との戦い

ケーブリングが無事に完了した後、、

- 結線ミス
- 初期不良
- コネクター汚れ
- **特定の Link でスループットが出ない??**

発生プロセスと現場対応の限界

01. 全リンクの性能確認

NCCLベンチマークにより、全リンクの通信性能を測定

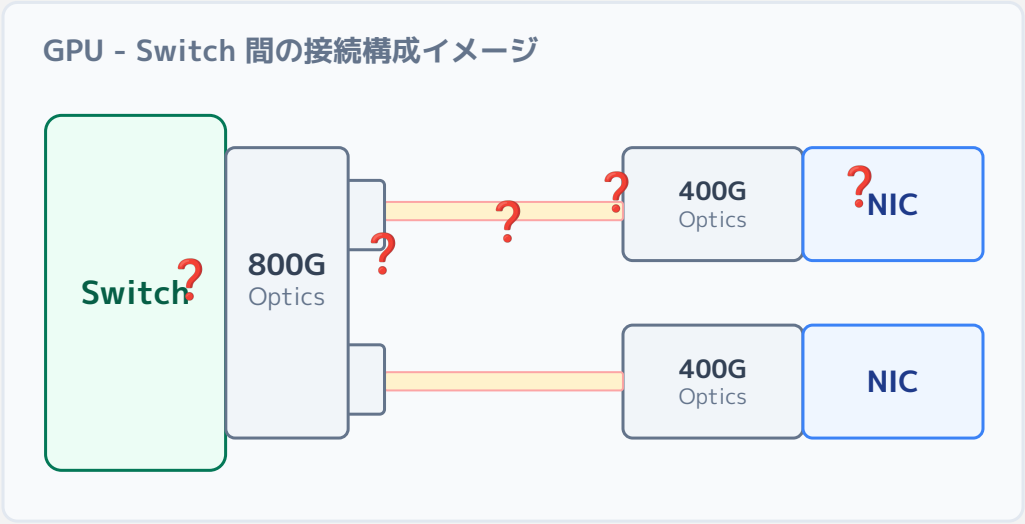
02. 異常発見

特定のリンクで、通信スループットが本来の数分の一程度まで低下する事象が発生

03. 状態の矛盾

・LED：Green
・Link Status: UP
・光パワー：正常範囲内

【Optics】スループット低下



切り分けポイントが多い...

- Switch?
- Optics?
- NIC?
- ファイバー?
- コネクター汚れ?

Optics 起因の問題を疑い、Optics 開発者と議論

BER / FECの値がスループット低下に密接に関係している

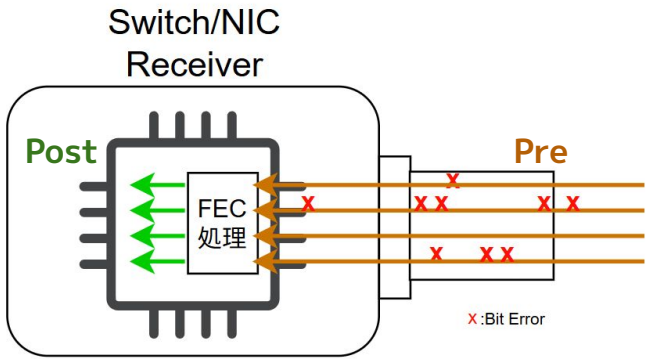
BER（Bit Error Rate）

■ **Layer1の伝送品質を示す基本指標**
BERとは、一定時間内に「受信した全 Bit 数」と「Bit Error 数」の比率

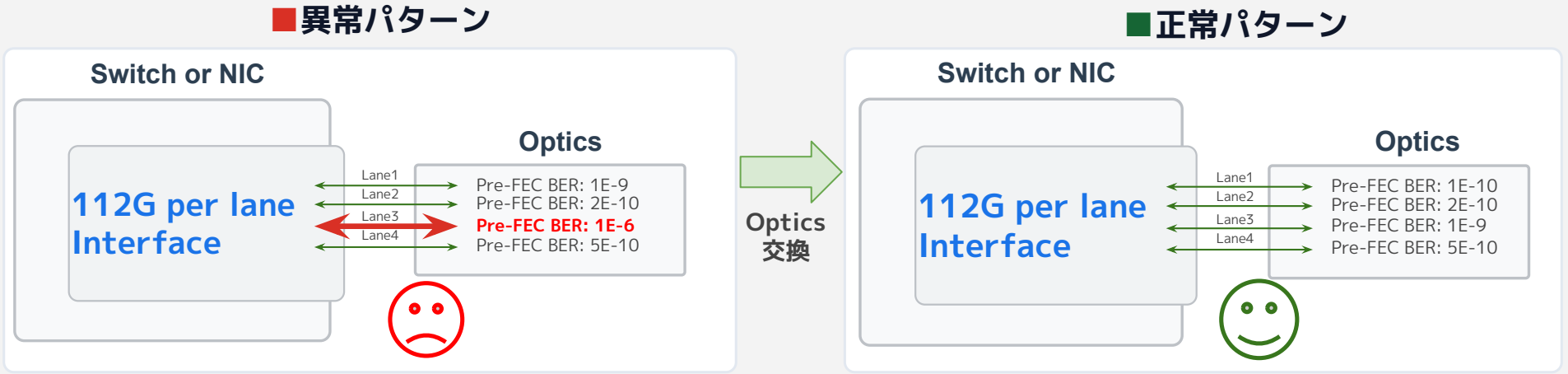
■ **表記**
指数表記（例: 1E-9 = 10億分の1）で表現され、**値が小さいほど通信品質が良好**と判断

■ **Pre-FEC BER & Post-FEC BER**

Pre-FEC BER: FEC訂正前のエラー率、**Optics から受け取った**生データ
Post-FEC BER: FEC訂正後のエラー率、0になるべき



【Optics】スループット低下の切り分け方法①



⚠️ **【切り分けポイント】 Pre-FEC BER が悪化したか確認**

例: Pre-FEC BER の値が**1E-8より小さい時**が正常とすると、**Lane3 の 1E-6** は異常と判断

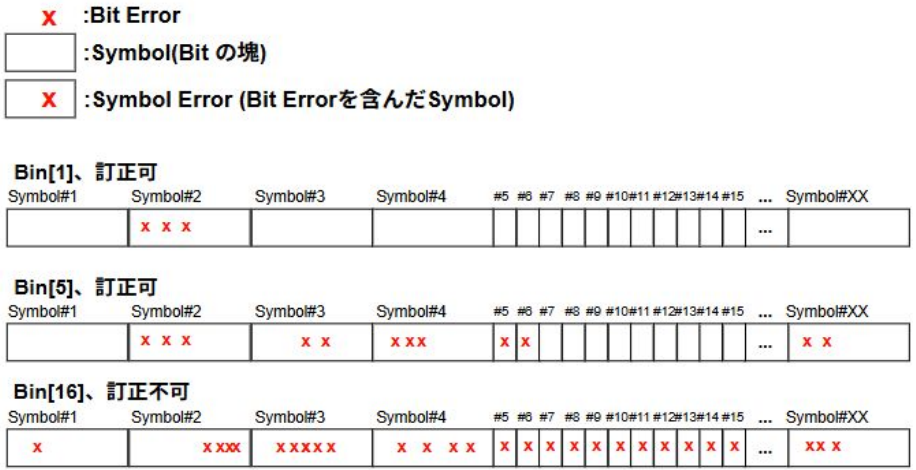
FEC (Forward Error Correction)

■ **FEC Histogram**
Bit Error の訂正状況の統計情報

■ **Symbol と Symbol Error**
Symbol: FEC処理用の Bit の塊
Symbol Error: Bit Errorを含んだ Symbol

■ **Bin[XX] の表記**
何個の Symbol Error がカウントされたか

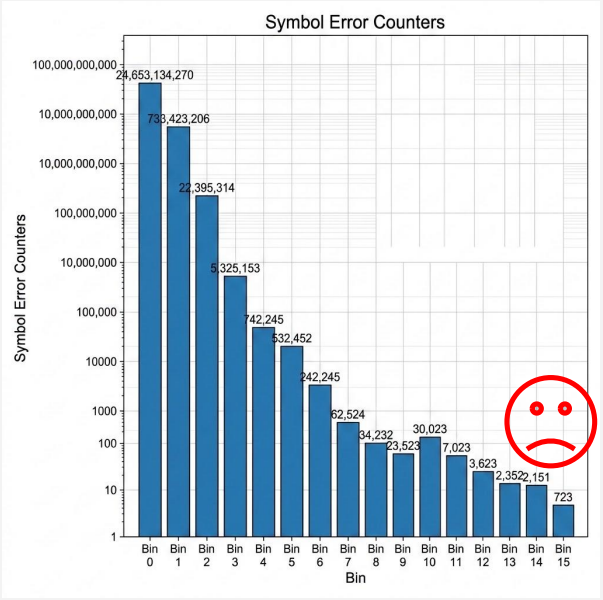
■ **Ethernet FEC RS の訂正範囲**
Bin[0] ~ Bin[15]: 訂正可
Bin[16] ~ : 訂正不可



■ **Bin[値] が増えている場合は、リンク品質が悪化している兆候**

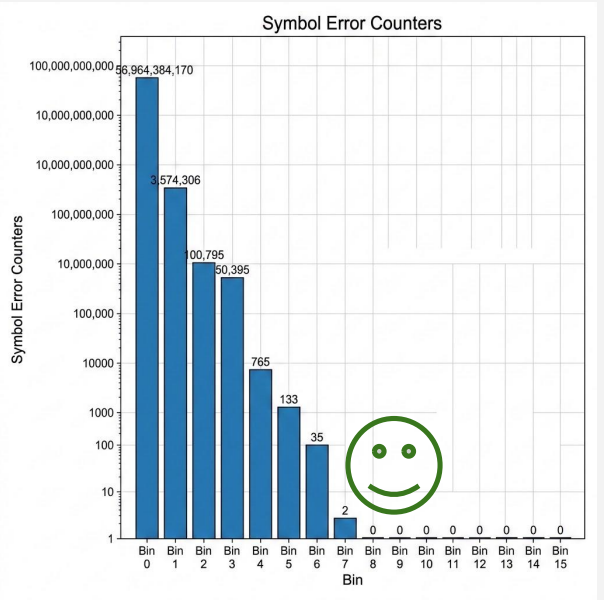
【Optics】スループット低下の切り分け方法②

■異常パターン



Optics 交換

■正常パターン



⚠️ **【切り分けポイント】Bin の [値]はどこまでカウントされたのか注意**

例: **Bin [15]までカウントされて**、訂正可能な閾値を超える可能性が高いためスループット低下の原因になる

【Optics】Opticsの性能評価の見直し

■物理層への理解

BERやFECなどの物理指標を重視し、Opticsの構造への理解を深める

■Optics評価の見直し

従来 → 見直し後

Link UP = OK
不十分!!!

物理層の品質をもとに評価

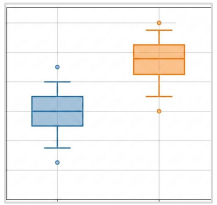
AI 計算基盤NWの安定稼働に貢献

■障害検知から劣化の予兆へ

Interface Down のみの監視体制は不十分

単なる Link down の検知だけではなく、劣化の予兆できる監視体制へシフトする

■ Interface の障害傾向を把握



週/月単位で Interface と Opticsログを収集・分析し、障害傾向を把握

【Optics】Optics 以外の選択肢

障害が発生しにくい部材を選定すればよい？

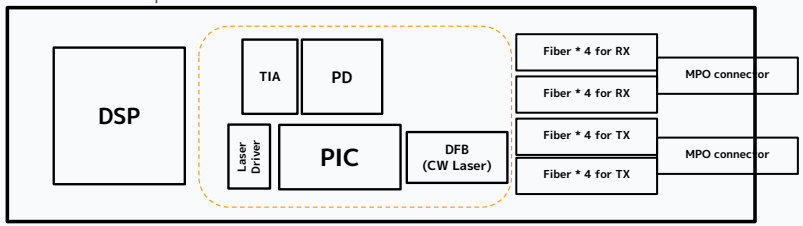
【Optics】Optics 以外の選択肢 - AEC

Optics
- 故障ポイントは多い
- レーザ特有の Flapping 不具合は多発

AEC（Active Electrical Cable）
- Copperと電気回路構成による安定品質
- 調達コストは安い

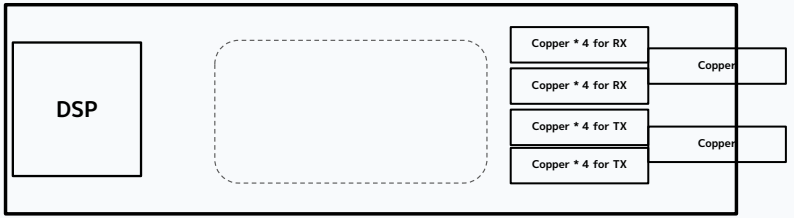
Optics Module

※ SiPh 800G Optics module



AEC Module

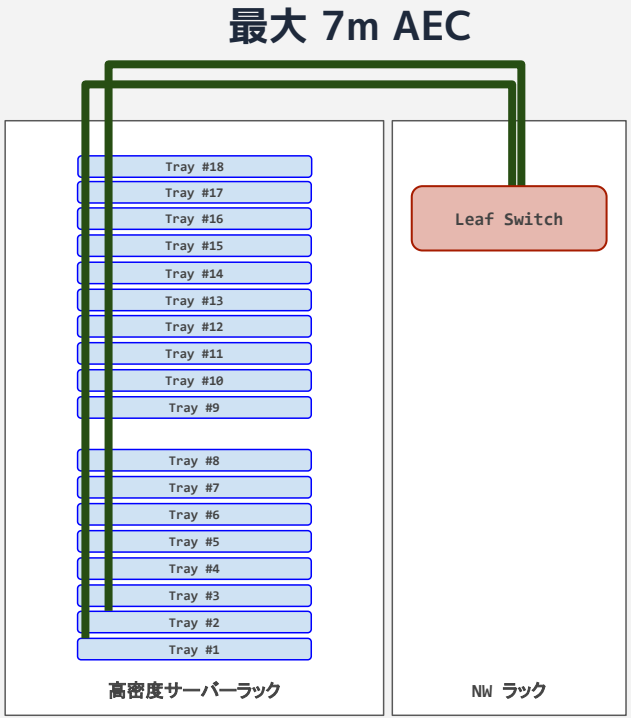
※ AEC 800G module



参考リンク / 技術発表実績

[SONiC Workshop Japan 2025] SONiC で 800G AEC ケーブルを検証してみた
[Janog57 LT] 安定した800G DCNWを求めて~トランシーバのBER計測~

【Optics】Optics 以外の選択肢 - AEC



■ AECと高密度サーバー

- 800G AECの最大 7m
- 配線の工夫は必要
- 高密度のGPUサーバーと親和性が高い
- Copper でも Scale-Out 実現できる

Liquid Cooling

【液冷機器の課題】Tray 交換時のトラブル

構築時に直面した課題

今後液冷スイッチの導入に向けて

- 液冷機器の満水確認
- 水温の監視
- 水圧など..

まとめ

01. Facilityの知識が必須

AI計算基盤の構築には、電源、冷却、サーバーのコンポーネントなどFacilityの知識が必須

02. Rack-Scale GPUサーバー内の理解

提供形態を決定する際は、サーバー独自の特徴を十分に考慮する必要がある

03. 最速デプロイのための工夫

最速でデプロイを完了するためには、従来のやり方にとらわれない工夫が必要

04. 開発者との対話の必要性

Opticsやケーブルメーカーの開発者と直接会話して深い技術的知見を得ることで、困難な切り分けも迅速に解決できる

Rack-Scale GPU サーバーの NW は ”苦悩” と ”挑戦” の積み重ね

01. 注目しているOCP以外の標準化団体

OCP以外で注目している標準化団体はあるか

液冷、電源、NW、Optics、ストレージ

02. 注目しているScale-Up / Scale-Out技術

Scale-Up:SUE, ESUN, UALink, CXL(memory pooling), NVLink Fusion

Scale-Out:UEC/UET, MRC, OCI(Optical Compute Interconnect), CPO

03. NWエンジニアの再定義

AI基盤時代のNWエンジニアの再定義

- スイッチ構造・プロトコル・アーキテクチャーの理解がメイン
- これから何を学ぶべきか、どのように学ぶべきか

04. データセンタの最速デプロイ手法

構築プロセス迅速化への取り組み

データセンタを最速でデプロイするために、現在どのような工夫や取り組みを行っているか

その他、議論したいことがあれば質問お願いします


SoftBank