

AIインフラ時代のデータセンター内光配線の実践知

～高密度実装の課題と解決策～



1. 自己紹介
2. これまでの議論の振り返り
3. さくらインターネットの取り組み
4. 構築の様子
5. いま見えてきている課題
6. 会場の皆さんとの議論



名前： 井上 喬視（いのうえ たかし）

所属： さくらインターネット株式会社
サービス統括本部 サービス基盤部

経歴：

2015年5月～ 入社 以降専用サーバーサービスに携わる

2023年5月～ 高火力 PHYメンバー

JANOG歴：

JANOG28 初参加 / 以後だいたい参加

JANOG54

生成AI向けパブリッククラウドサービスをつくってみた話

JANOG56

生成AIインフラを構築してわかったケーブリングの重要性



名前： 菊地 秀夫（きくち ひでお）

経歴： 1992年4月～2026年3月

株式会社フジクラ

光ケーブル及び光コンポーネットのエンジニア

2026年4月～

TWCCパートナーズ合同会社 代表

データセンターや生成AIの光配線コンサルティング

エクストリーク株式会社と業務提携

JANOG歴：

JANOG50より参加

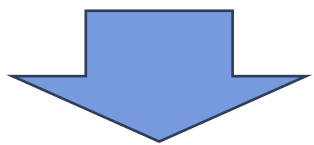
JANOG56

生成AIインフラを構築してわかったケーブリングの重要性

これまでの議論の振り返り

生成AIインフラを構築してわかったケーブルリングの重要性

- さくらインターネットの生成AIインフラ構築事例を紹介
- 制約のあるデータセンターにおいて、どのように設計、構築、運用をしていくか



- 生成AIのためのインフラ構築にあたっては、これまで以上に光配線設計が重要となっている

課題① 大量の光ケーブルをどのように敷設するか？

光ケーブルの細径化 & 多心化によるボリューム削減

MPOコネクタ付き細径多心ケーブルの採用

- 従来のトランクケーブルと比較し、細径且つ多心のケーブルの採用
- 同心数と比較して細径且つ軽量のためケーブルの取り扱いやすさを考慮

細径 & 多心ケーブル構造

SM 288心ケーブルの仕様

項目	仕様
ファイバ	シングルモードファイバ 200μm
リボン	12心間欠固定(250μmピッチ)
ケーブル外径	12.5mm
ケーブル質量	155kg/km
難燃性	Cca-sla,d0,a1, UL 1666, UL 1685

ケーブルの心数密度比較

288心ケーブル(細径多心タイプ) 144心トランクケーブル

2倍の心数を70%のケーブル断面積で実現
→ ケーブルボリュームを1/3まで削減

その他さまざまな工夫

ラック間の配線本数を削減する工夫

- GPUサーバーラックからの配線をMPOコネクタ付きSM288心ケーブルにすべて乗せることでケーブル配線数を削減

既設システムの構成(一例)

コンテナ内のケーブルリング構成

インターコネクト用とNorth-South NW
管理用のケーブルを別々に敷設

インターコネクト用とNorth-South NW
管理用のケーブルを集約

いざというときの保守用予備心線を設ける工夫

- 万が一、ケーブル敷設作業中に断線や破損が生じた場合でも予備心線へ切り替えられるようにリスク低減を図った

MPOコネクタの接続先

No.1 ~ No.16 - GPUサーバ - インターコネクトスイッチ
No.17 ~ No.20 - MPO/LC変換パッチパネル - MPO/LC変換パッチパネル
No.21 ~ No.24 - GPUサーバ - インターコネクトスイッチ

<https://www.janog.gr.jp/meeting/janog56/cabledesign/>

JANOG 56 in MATSUE

2025 07.30 - 08.01

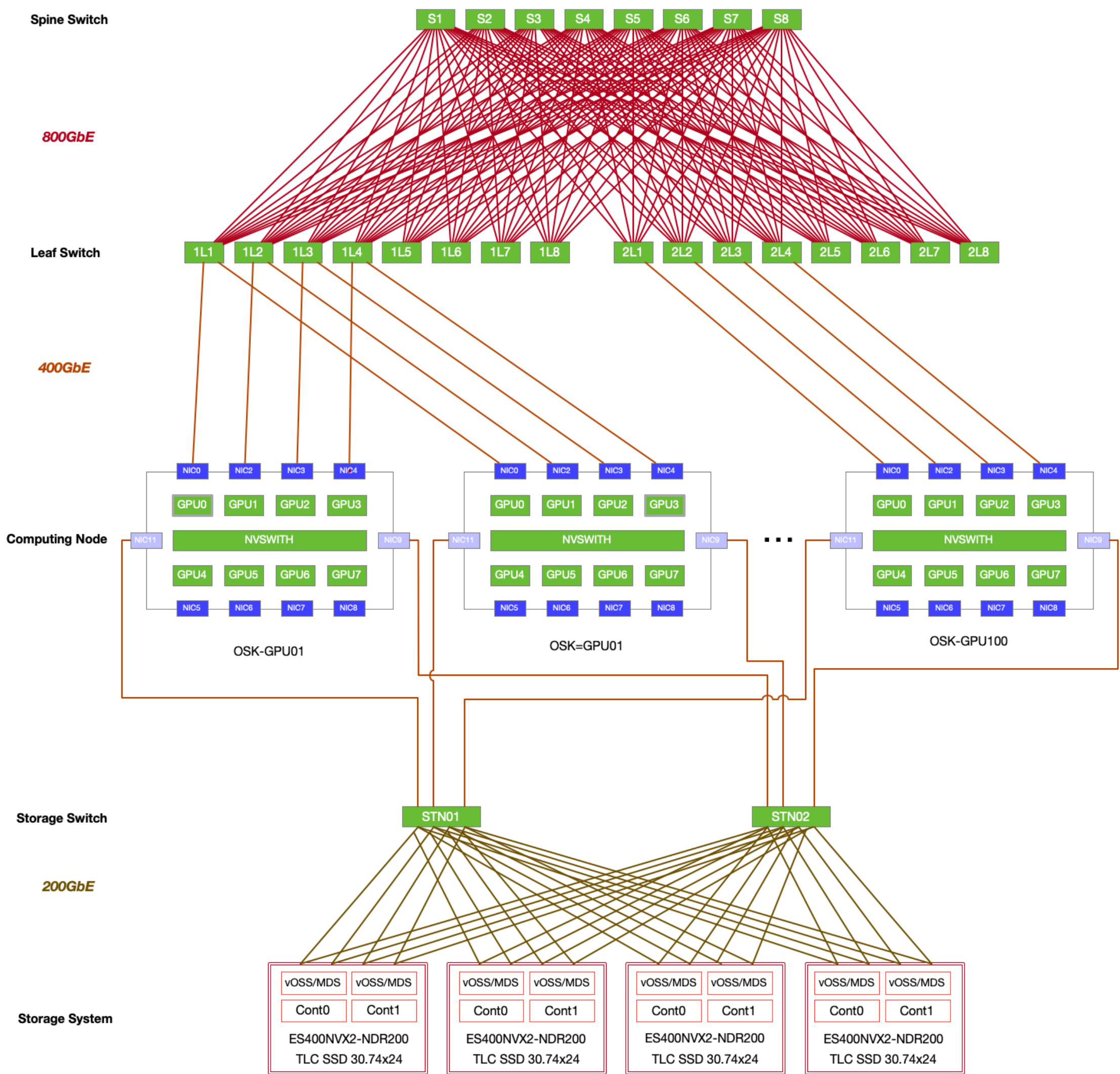
ケーブルリング BoF & Open Tranceiver / Optics BoF

- ケーブルやパッチパネルの管理は各組織でどのように行われているか
- MPOケーブルの取り回しについての難しさについて
- ケーブルタグとか整線部材はなにを使ってる？
- 光ケーブル/トランシーバ共に仕様が細分化しているので選定や設計に苦労
- 液冷に適したOSFPフォームファクタについて（IHS Open Top/Closed Top vs RHS）
- トランシーバーに関するトラブルシューティングネタ（苦労話）
- トランシーバーの部品供給についての話
- テックジャイアントにおける展望 … などなど



さくらインターネットの取り組み

「NVIDIA H100 SXM 80GB」で構成されたLLMの学習に特化したクラウド型のスパコン



さくらONEの特徴

- さくらインターネット研究所とプラナスソリューションズ株式会社との共同構築プロジェクト
- GPU間を接続するネットワークにはBroadcom Tomahawk5 800G 64ポートスイッチ 計24台によるLeaf/Spine構成、ネットワークOSとしてはSONiCを採用
- ISC2025で発表されたTOP500において世界49位を獲得
※ISC2026においては 世界69位



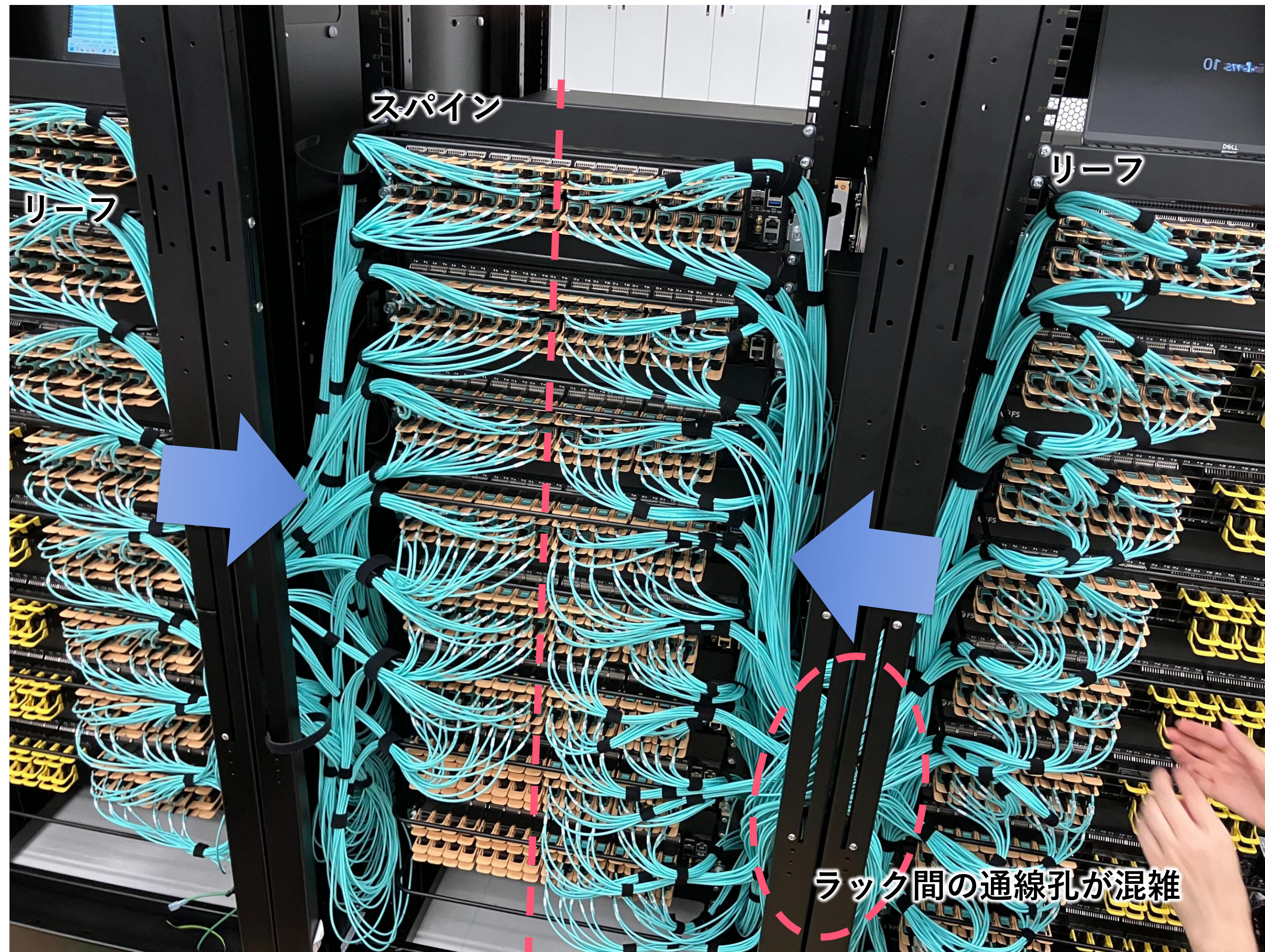
SONiC Workshop Japan 2025
SONiCで構築・運用する生成AI向けパブリッククラウドネットワーク
<https://sonic.connpass.com/event/345182/>

SONiC Workshop Japan 2026
SONiCで構築・運用する生成AI向けパブリッククラウドネットワーク～実装編～
<https://sonic.connpass.com/event/385386/>

SAKURAONE: Empowering Transparent and Open AI Platforms through Private-Sector HPC Investment in Japan
<https://arxiv.org/abs/2507.02124>

さくらインターネット研究所、クラウド型のスーパーコンピュータシステム「さくらONE」が処理性能ランキングTOP500で、世界49位を獲得
<https://www.sakura.ad.jp/corporate/information/newsreleases/2025/06/11/1968219771/>

スパインスイッチの両隣ラックにリーフスイッチを配置する設計



800G x 64ポートスイッチの採用

- 800G-SR8の採用による配線本数の削減
- 距離が近いためマルチモード規格を採用
 - MPO16 MMF APC研磨ケーブル 2m~3m

配線のしやすさを考慮したポート収容

- スパインスイッチのポートを真ん中で分割
 - 右半分は右側の配線用、左半分は左側の配線用
- 一方でケーブル余長の吸収には苦労した
 - 通線孔が1か所しかなかった
 - ケーブル不良時の交換が大変
 - 不良時は諦めて敷設しなおすことも視野に

2025年6月よりサービス提供を開始した水冷式コンテナデータセンター



コンテナ型データセンターによる構築

- さくらインターネット石狩データセンターの敷地内に水冷式のコンテナ型データセンターを構築
- 1コンテナあたり20ラック、合計40ラックをサービス提供中
- 計画から約1年6か月で竣工

直接液体冷却方式の導入

- 直接液体冷却方式（Direct Liquid Cooling）を導入し、冷却に必要な消費電力の削減
- NVIDIA H200 GPUを8基搭載した水冷モデルのサーバーを導入、合計約1000基を整備
- 2025年8月よりNVIDIA Blackwell GPU（B200）も同型コンテナにて提供開始

さくらインターネット、コンテナ型データセンターの稼働を開始
～ベアメタル型GPUクラウドサービス「高火力 PHY」にて、H200プランを提供開始～
<https://www.sakura.ad.jp/corporate/information/newsreleases/2025/06/11/1968219778/>

2026年2月より NVIDIA Blackwell GPU (B200) 約1,100基 の提供を開始



より大型のコンテナ型データセンターを構築

- 大型のコンテナ採用により、規模の大きいクラスタの構築が可能となった
- 直接液体冷却方式（Direct Liquid Cooling）を導入
- ISC2026で発表されたTOP500において**世界63位**を獲得
 - HPCGにおいては**世界22位**
- 昨年構築したコンテナデータセンターからの改善を盛り込んでいる



さくらインターネット、「NVIDIA Blackwell GPU」約1,100基を搭載したAIインフラの稼働を開始

<https://www.sakura.ad.jp/corporate/information/newsreleases/2026/02/25/1968223641/>

ISC2026で発表されたTOP500においては63位、HPCGでは22位にランクイン

SAKURAONE CYD - SUPERMICRO SUPERSERVER B200,
INTEL XEON 6960P 72C 2.7GHZ, NVIDIA B200 SXM 180GB,
GIGABIT ETHERNET, ROCKY LINUX 9.6

Site:	SAKURA Internet Inc. / Prunus Solutions Inc.
System URL:	http://one.sakura.ad.jp
Manufacturer:	Supermicro
Cores:	185,920
Processor:	Intel Xeon 6960P 72C 2.7GHz
Interconnect:	Gigabit Ethernet
Installation Year:	2026
Performance	
Linpack Performance (Rmax)	37.79 PFlop/s
Theoretical Peak (Rpeak)	43.18 PFlop/s
Nmax	4,669,440
Nhalf	2,334,720
HPCG	1,076.97 TFlop/s

ベンチマーク結果

- 37.79 PFLOP/s (HPL Rmax)
- 1076.97 TFLOP/s (HPCG)
- 9.1146 EFLOP/s (HPL-MxP with FP16) ※今年度のランキング未発表

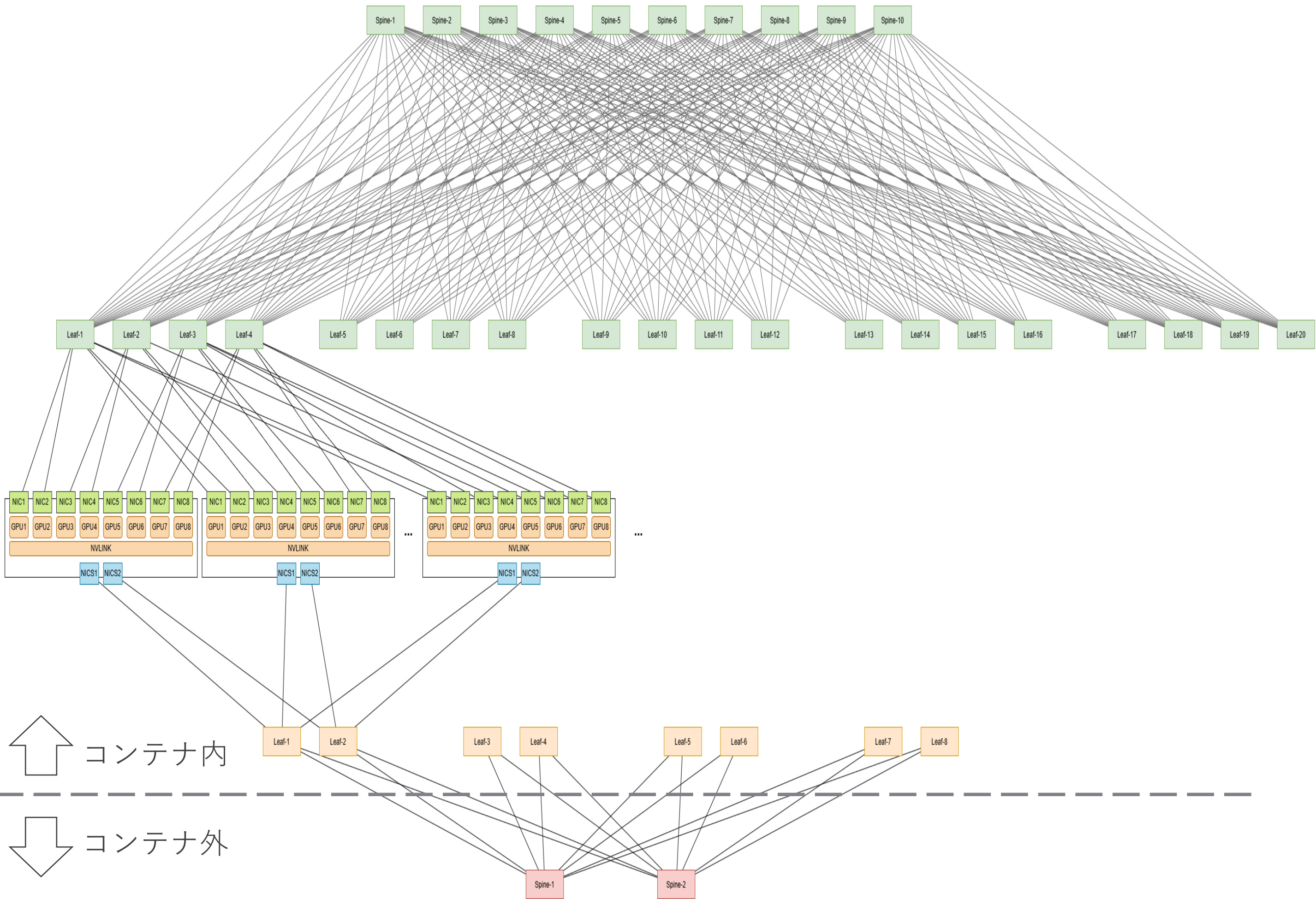
前年度のHPL-MxPランキング

November 2025							
Rank	Site	Computer	Cores	HPL-MxP (Eflop/s)	TOP500 Rank	HPL Rmax (Eflop/s)	Speedup
1	DOE/SC/LLNL	El Capitan	11,340,000	16.680	1	1.8090	9.2
2	DOE/SC/ANL	Aurora	9,264,128	11.643	3	1.0120	11.5
3	DOE/SC/ORNL	Frontier	9,066,176	11.390	2	1.3530	8.4
4	EuroHPC/FZJ	JUPITER Booster	4,801,344	6.255	4	1.0000	6.3
5	SoftBank	CHIE-4	662,256	3.304	17	0.1354	24.4

Ethernetインターコネクトのみ (TOP100以内)

Rank	Name	System	Interconnect
51	EYAS	NVIDIA H200 SXM5 142GB	400G Ethernet
59	Israel-1	NVIDIA H100	NVIDIA Spectrum-X Ethernet
63	SAKURAONE CYD	NVIDIA B200 SXM 180GB	400G Ethernet
69	SAKURAONE	NVIDIA H100 SXM5	400G Ethernet
98	KDDI GPU Cloud	NVIDIA DGX GB200 NVL72	NVIDIA Spectrum-X Ethernet

Arista 7060X6-64PEを採用したLeaf-Spine構成



ComputeNetwork構成

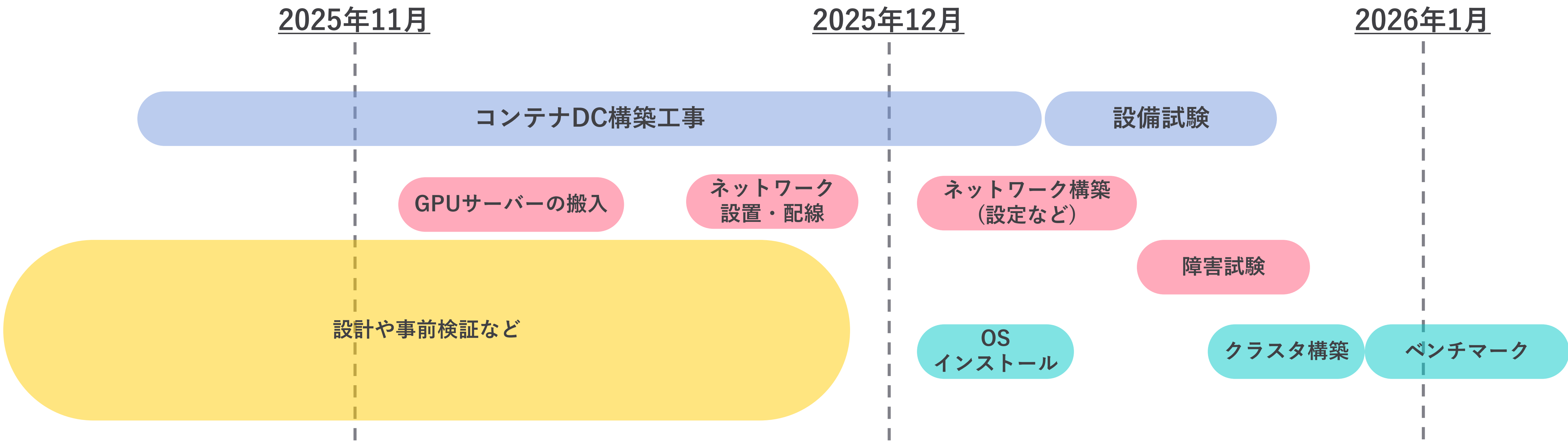
- Spine10台、Leaf20台構成
- 800G x 64ポート Tomahawk5 Arista 7060X6-64PE
- マルチテナンシ EVPN/VXLAN + RoCEv2
- 一部でLPOトランシーバーを採用



StorageNetwork構成

- Leafは2台ずつ、4セット構成で 合計8台の
- 400G x 54ポート Jericho2C+ Arista 7280DR3A-54
- シンプルにVLANによるマルチテナンシ + RoCEv2
- アップリンクはコンテナ外に設置のスイッチに接続
- やや距離があるためDeepBufferモデルを選択

迅速なサービスリリースに向けて納品後短期間で構築を実施



構築期間を短くするためにも事前の設計が重要となる

コンテナ内特有の環境に起因する「狭さ」と「密度」

狭さ

ラックの設置台数に限りがある

- あくまでもコンテナ内はGPUサーバーがメイン
- 狭いながらも以下ネットワークが必要
 - GPUインターコネクト（GPU間をつなぐための高速NW）
 - North-South / ストレージネットワーク
 - 管理用ネットワーク

配線スペースが限られる

- 天井高は一般的なデータセンターより低い
- コンテナ内にはネットワーク、電源配線以外にも水冷設備の配管が張り巡らされており、ラック上のスペースが限られる

密度

水冷設備導入により、1ラック当たりのGPUサーバー台数が増加



1ラック当たりの配線本数も増加

空冷環境の場合、1ラック当たり2台程度だったが
水冷設備導入により4台～5台に増加

- ☑ 光心数にして約30000心程度が必要
- ☑ 水冷配管などでラック内の密度も上昇

これまでのコンテナDC構築経験で見えてきた改善点をできるかぎり盛り込んで構築を実施

配線設計の重要性

- 配線の方法ひとつでネットワークシステムの運用性が大きく変化する
- コンテナのような狭小環境でも配線を工夫することで十分に構築することができる
- 生成AIインフラのような大量の配線を扱うシステムにおいてはより顕著に重要性が増す

これまで実施していた取り組み

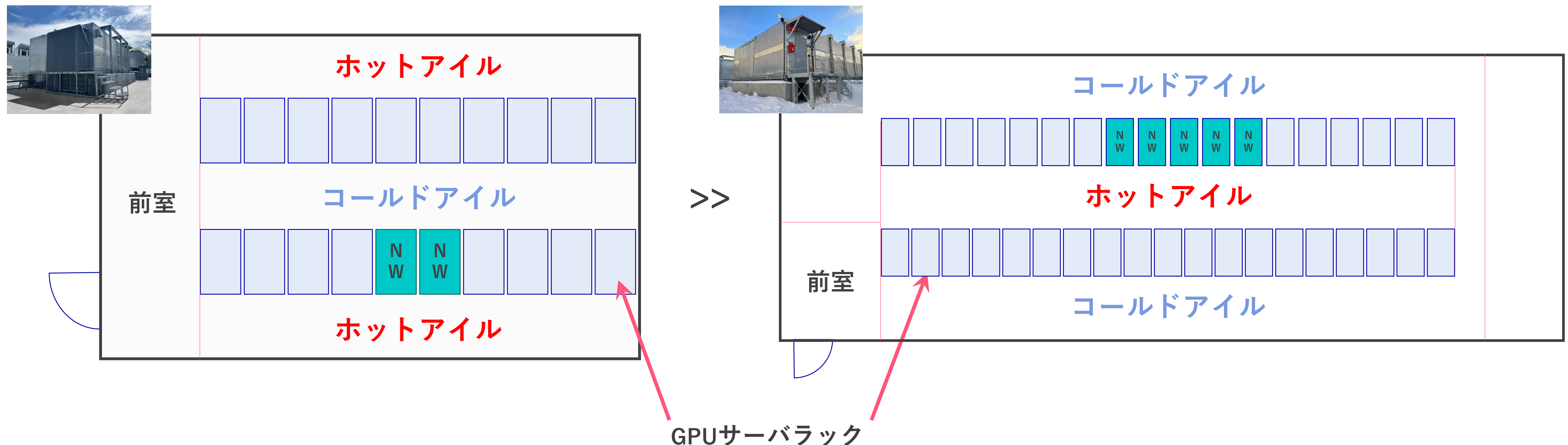
- MPOコネクタ付き細径多心ケーブルの採用
- ラック内スペースの有効活用
 - GPUラック内でケーブル余長を吸収
 - 集合型ケーブルによる直接配線
- 効率的な配線ルート of 工夫

新たに取り組んだこと

- GPU数の増加にともない、さらなるケーブルの最適化
 - 光ファイバ心数利用の効率化によるケーブル本数および配線部材数の抑制
- ラック内レイアウトと配線経路の工夫
 - 効率的なケーブルルーティング、光パッチコードの配線の最適化
- ネットワーク配線ラックの最適化
 - 大量の配線をさばけるラック選定、エアフローを考慮したカスタマイズ

コンテナのサイズが大きくなりラック数は2倍以上、GPUの数は3倍近くに

- シングルシャーシ構成 から リーフスパイン構成に 変化したことでスイッチの台数が大幅に増加
- リーフスパイン間の配線が増え、配線本数も大幅に増加
- コンテナDC自体のサイズは大きくなったものの、配線で利用できるスペースに大きなゆとりが生まれたわけではない
- 電気設備配線や水冷設備の配管についても規模に応じて増大している



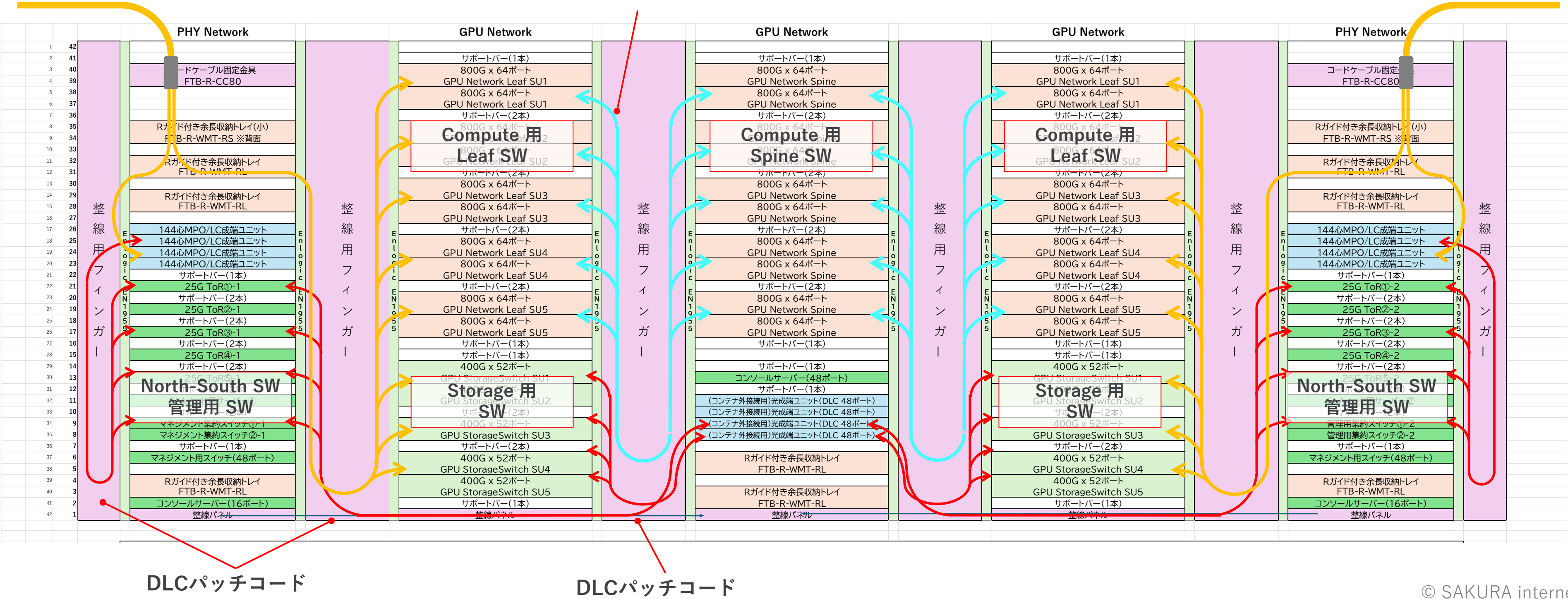
スパインスイッチを中心に左右対称となるように機材を設置、配線の流れを2分割化

- スイッチを搭載するネットワークラックは並べて配置
- ネットワークラック内の光配線の流れを全体的にデザインし、なるべく配線をクロスさせない、両端から中央に流れる配線設計とした
- 高集約パッチパネルを採用することでラック内の機材設置スペースを確保
- ケーブル余長吸収トレイを適切に利用することで、ラック間スペースで無駄に余長を取らない工夫を実施

MPOコネクタ付き288ケーブル(400G-DR4)

MMF MPO16パッチコード(800G-VR8)

MPOコネクタ付き288ケーブル(400G-DR4)



可能な限り配線距離を計算し、接続ポートに応じて配線距離を適正化

- 特にラック間のスペースに限りがあるため、Leaf-Spine間のケーブル長を個別に最適化
- 短すぎると配線ができないが、かといって長すぎるとラック間スペースに収まりきらないことも

↑
プラス0 = 同じ高さのポートに繋ぐ場合の本数

↑
プラス1 = 高さが1U違うポートに繋ぐ場合の本数

	プラス0	プラス1	プラス2	プラス3	プラス4	プラス5	プラス6	プラス7	プラス8	プラス9	片脇分	両脇分	予備込み
	30	54	48	42	36	30	24	18	12	6	300	600	
1.5m	30	54									84	168	185
2.0m			48	42	36	30	24				180	360	396
2.5m								18	12	6	36	72	81

パッチケーブルにはケーブル長がわかるように
長さ情報がわかるタグを取り付け

	A	B	C
1	長さ ☐	ランダム値 ☐	ラベル ☐
2	1.5m	WX97TPP4	(1.5m) WX97TPP4
3	1.5m	RTF4C34L	(1.5m) RTF4C34L
4	1.5m	RVF4JKJP	(1.5m) RVF4JKJP
5	1.5m	RRMWW3L	(1.5m) RRMWW3L
6	1.5m	TPWXF47T	(1.5m) TPWXF47T
7	1.5m	JJN3FJW4	(1.5m) JJN3FJW4
8	1.5m	HKT9V4VP	(1.5m) HKT9V4VP
9	1.5m	H7HFHFK3	(1.5m) H7HFHFK3
10	1.5m	9R73NWJK	(1.5m) 9R73NWJK
11	1.5m	HJHX7HTN	(1.5m) HJHX7HTN



リーフ

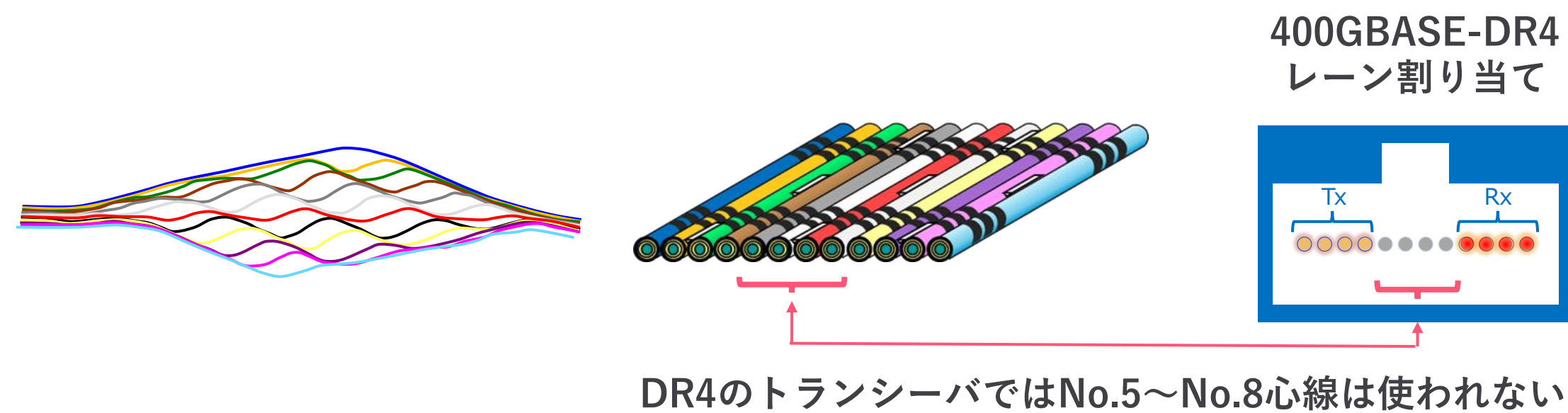
スパイン

リーフ

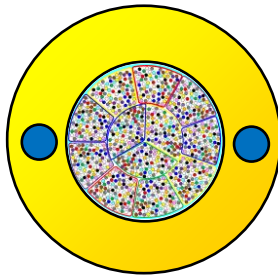


12心間欠固定リボン → 16心間欠固定リボンへ変更 同じ心数でも1.5倍の収容効率アップ

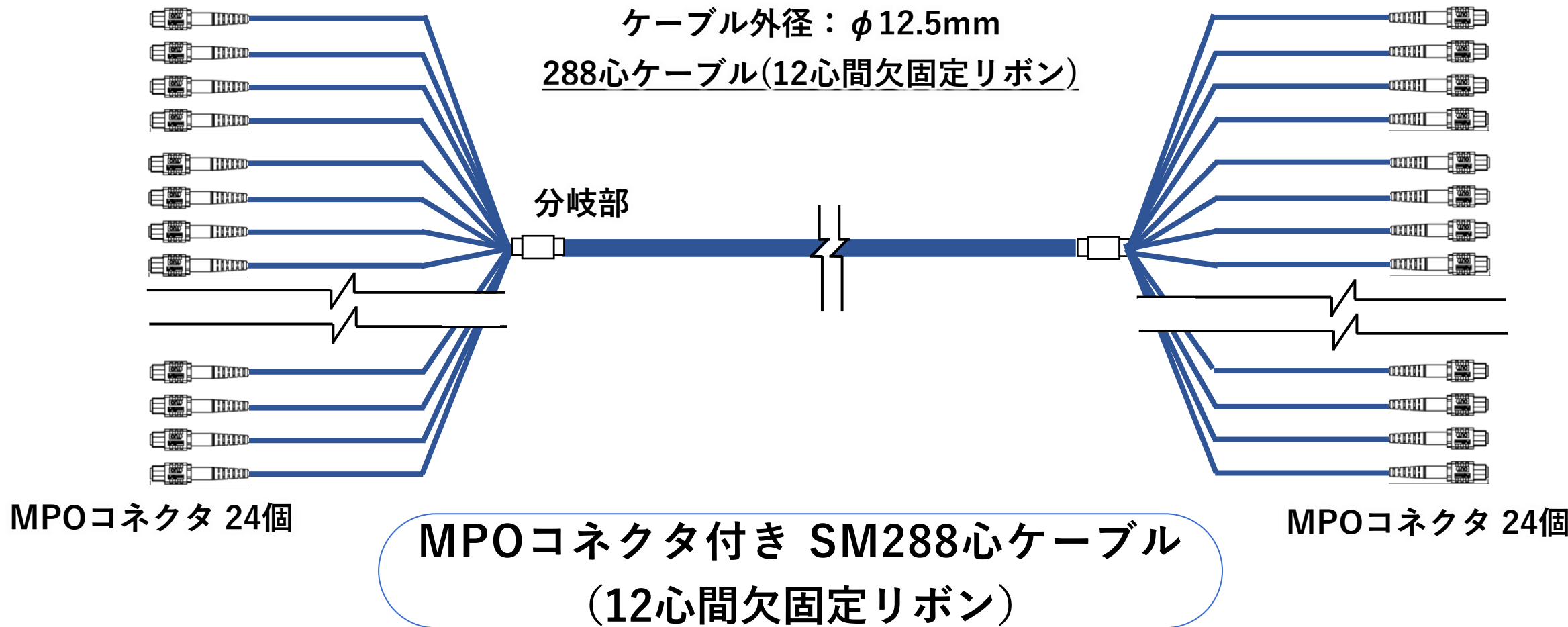
これまでのコンテナDCで使用していたケーブル



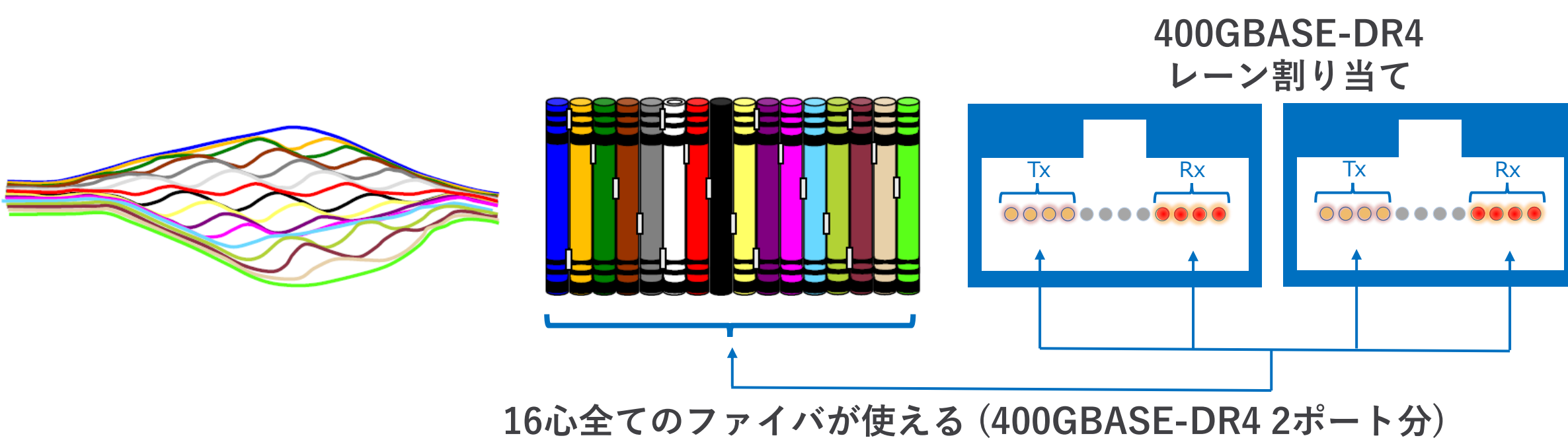
12心間欠固定リボン



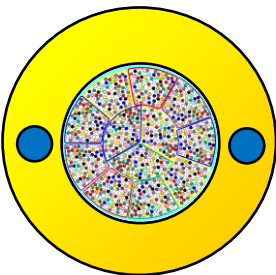
ケーブル外径：φ12.5mm
288心ケーブル(12心間欠固定リボン)



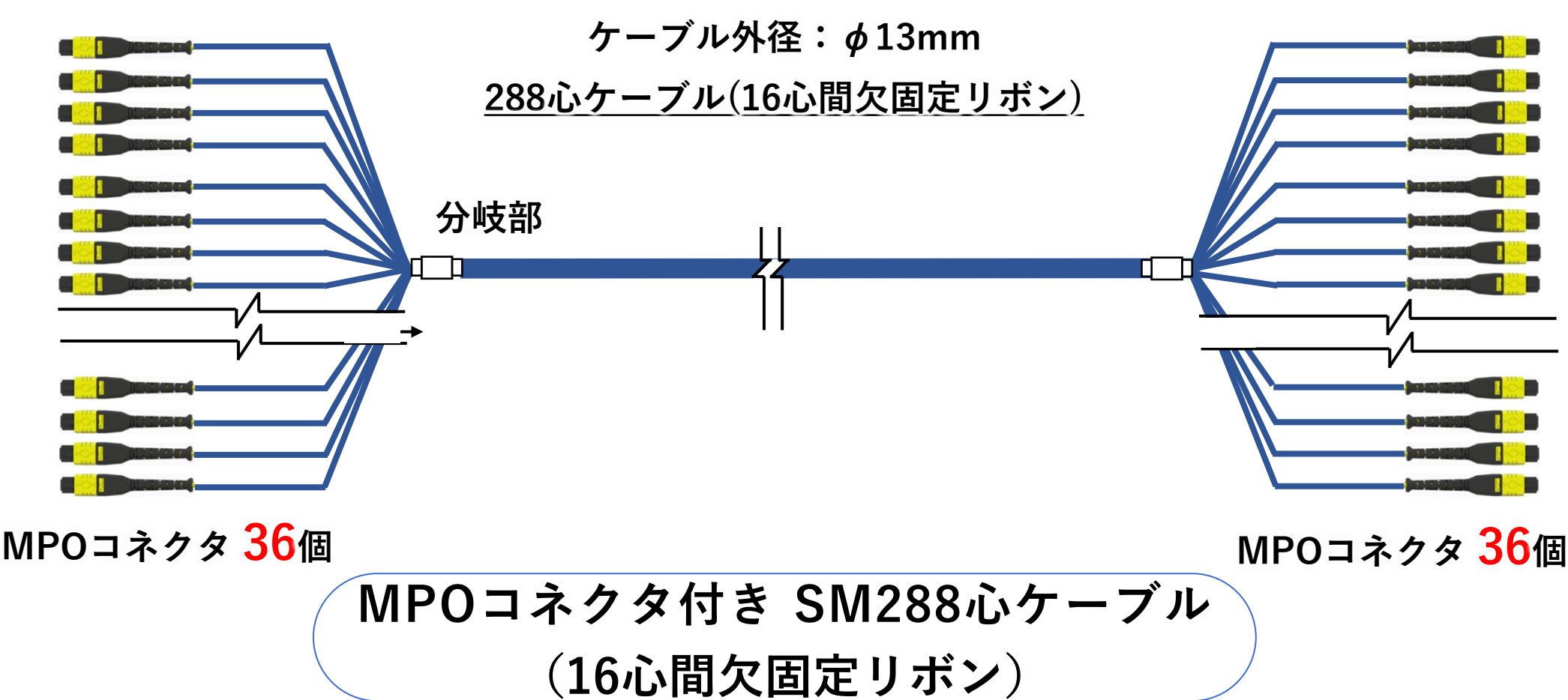
今回のコンテナDCで使用した光ケーブル



16心間欠固定リボン



ケーブル外径：φ13mm
288心ケーブル(16心間欠固定リボン)



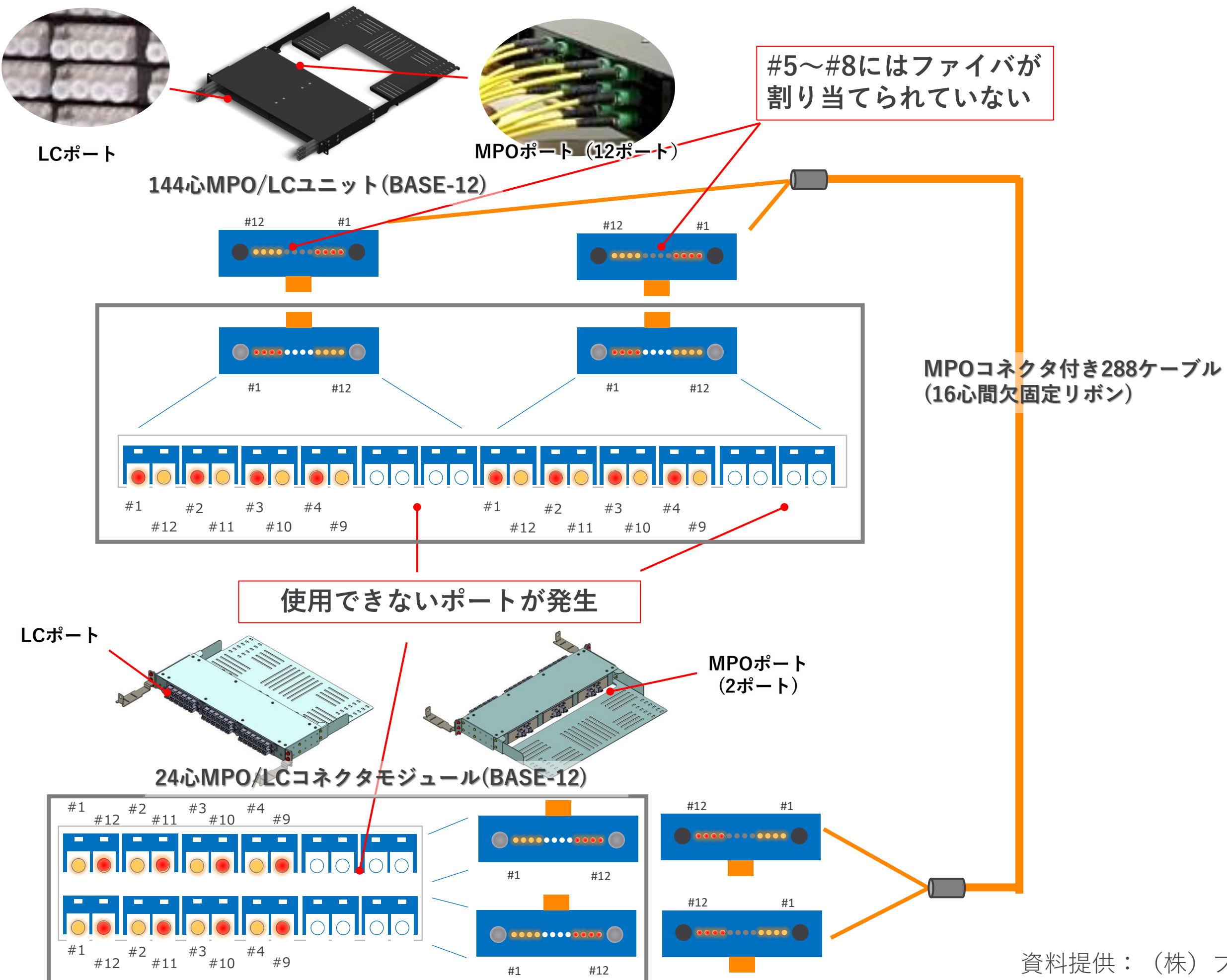
同じ288心ケーブルでもコネクタ数が1.5倍に



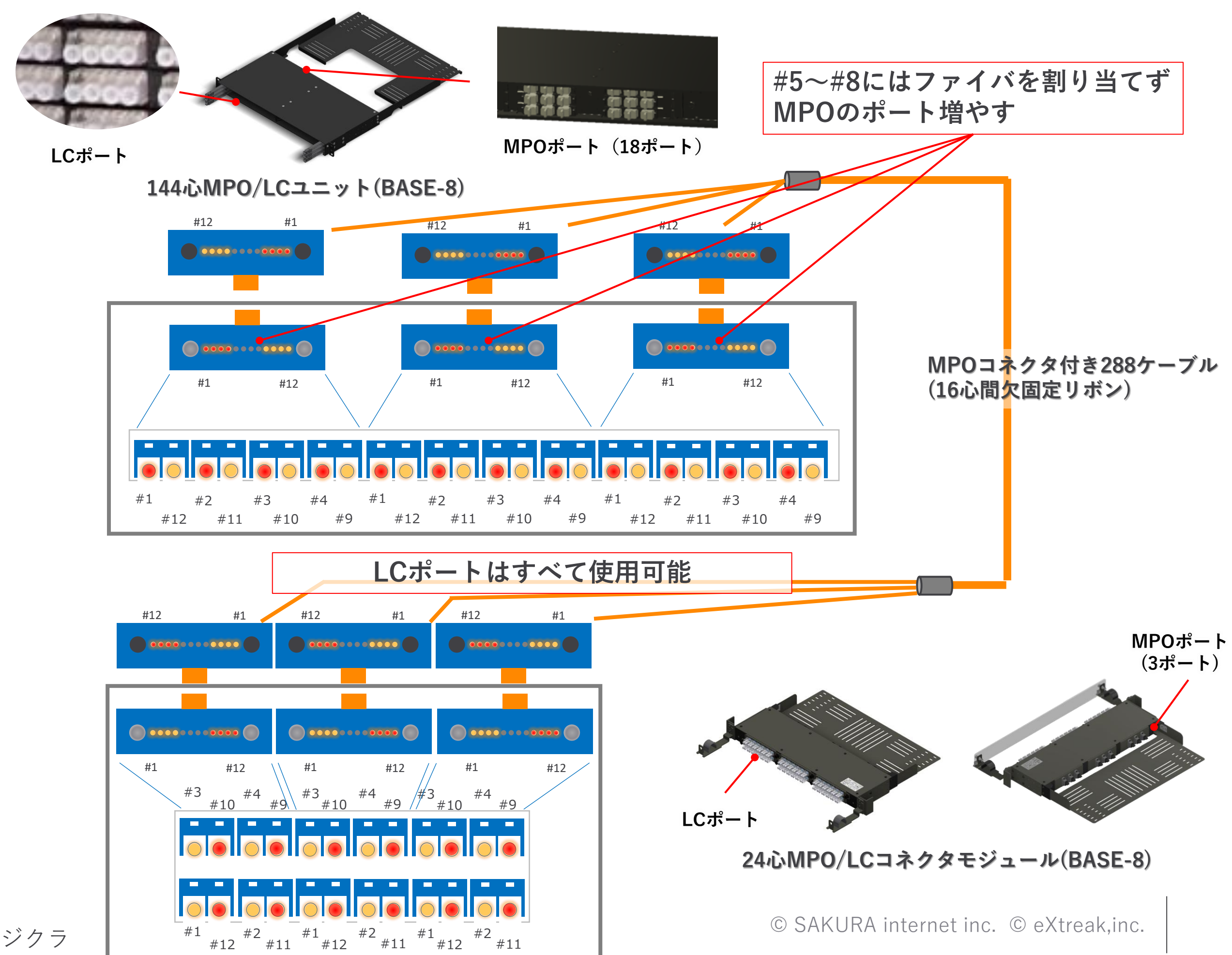
光ケーブルの16間欠固定リボンに合わせて MPO/LCカセット類をBASE-12→BASE-8へ変更

LCコネクタインターフェースのランシーバをつなぐネットワークにはBASE-8用 MPO/LC製品を使用することでコネクタ実装密度の低下を回避

BASE-12の伝送経路

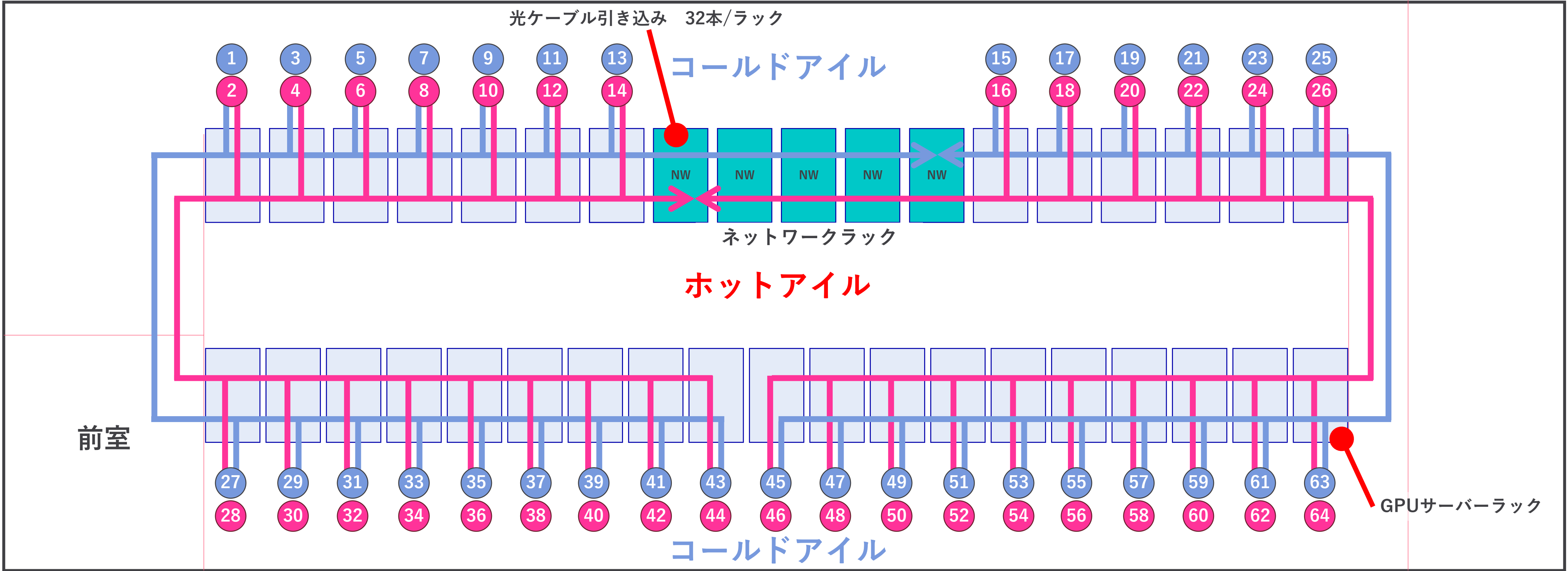


BASE-8の伝送経路



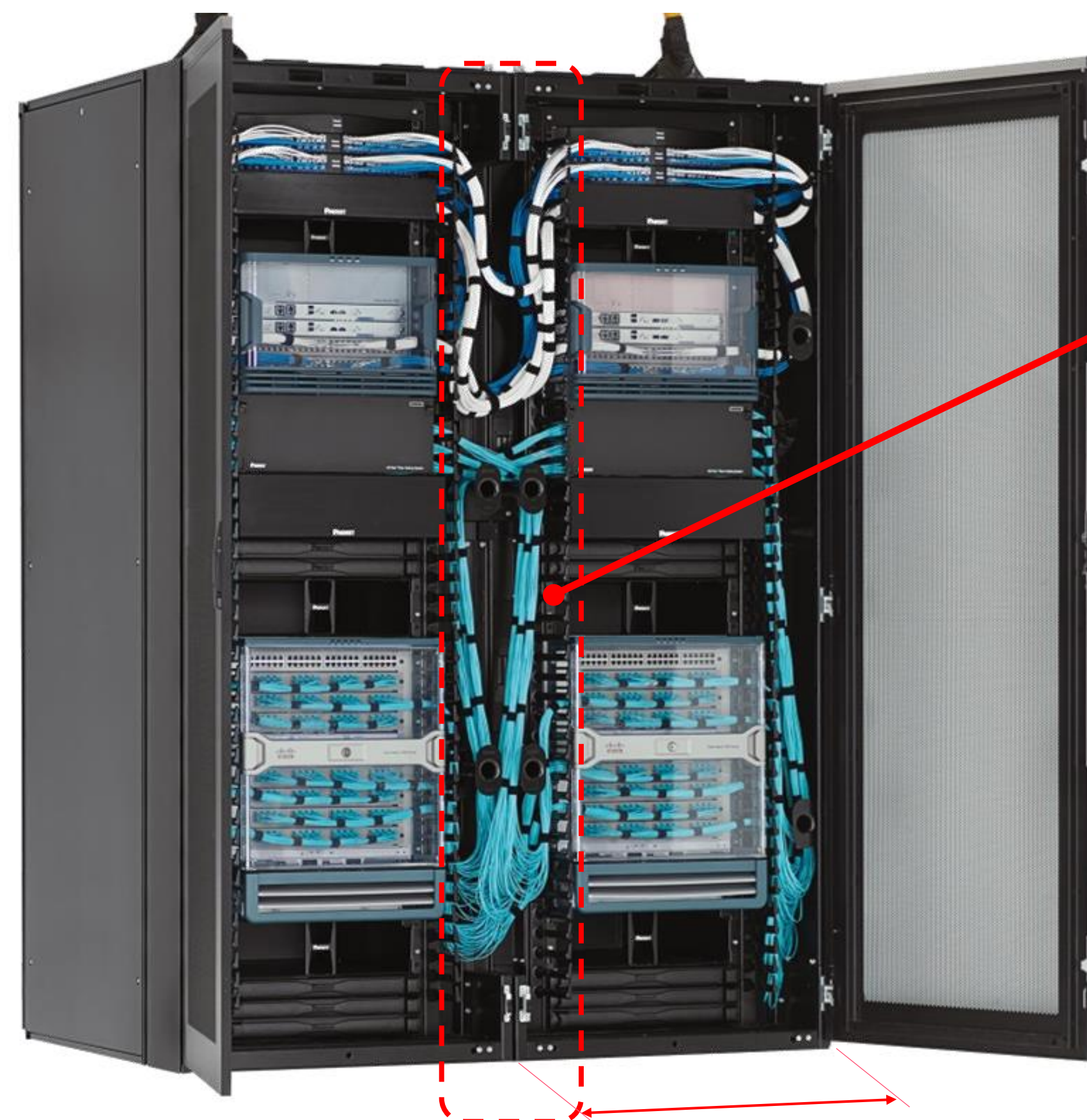
右回りルートと左回りルートに分けラダー上のケーブルを分散

- 1ラック当たり光ケーブル2本、計64本の光ケーブルをラダー上に敷設（これまでのコンテナDC構築規模と比べ約2倍の本数）
- ラダー上下の空間スペースに余裕がないため、光ケーブルができるだけ交差しないように配線位置とラックへのケーブル引き込み位置を調整



ネットワーク収容ラックには配線に適したラックを個別に選定

- 大量の光配線を収容可能なネットワークラックを選定
- 光配線部が見やすく、配線作業においてケーブルにアクセスしやすい（手が届きやすい）



連結前面のフロントサイドパネルを取り外すことで連結間の高い接続性を実現

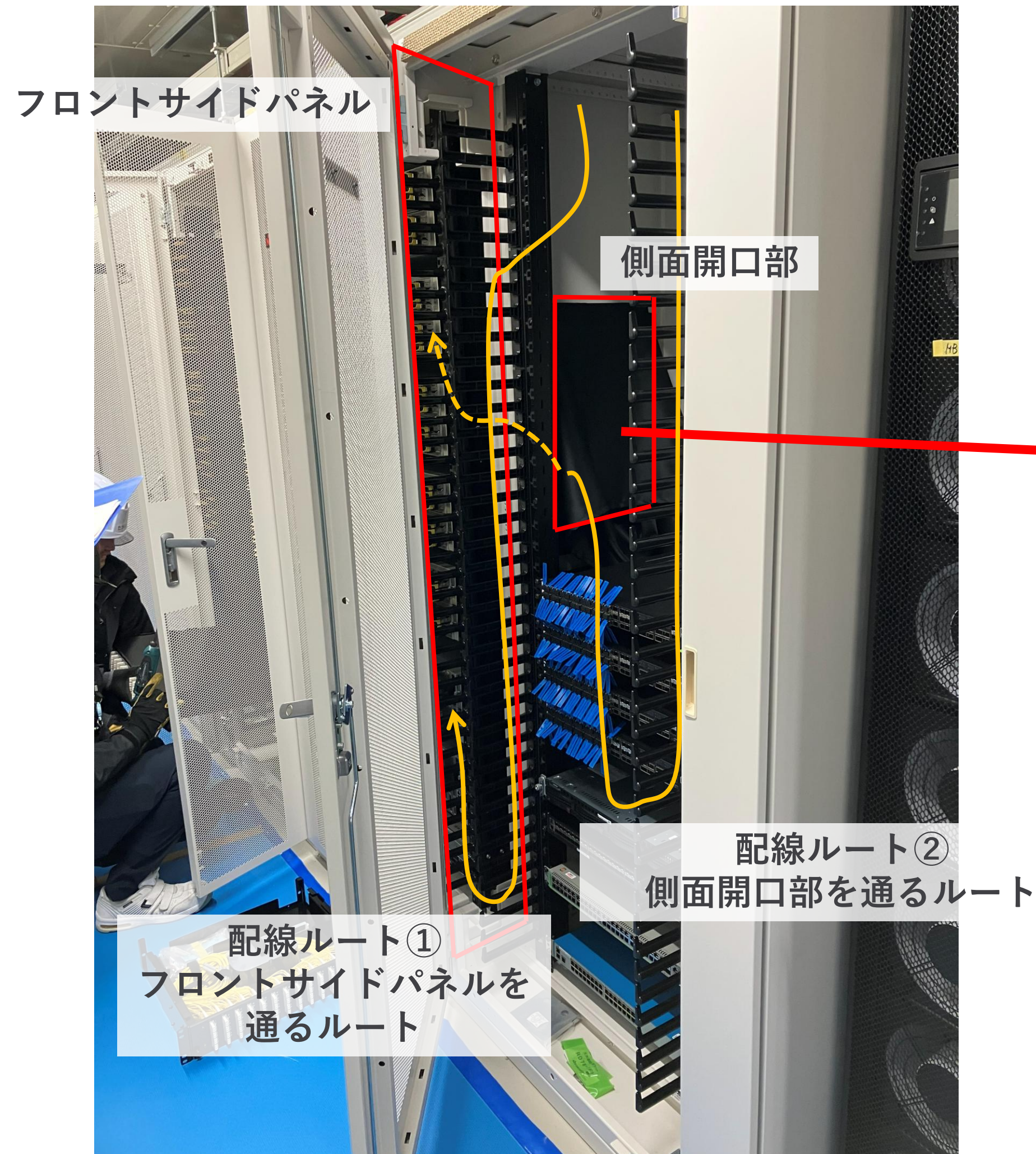
W800mmのラックを採用することで大量の光配線を行うスペースを確保

独立式フロントサイドパネル



配線を分散させるため、側面に開口部を設け配線ルートを確保

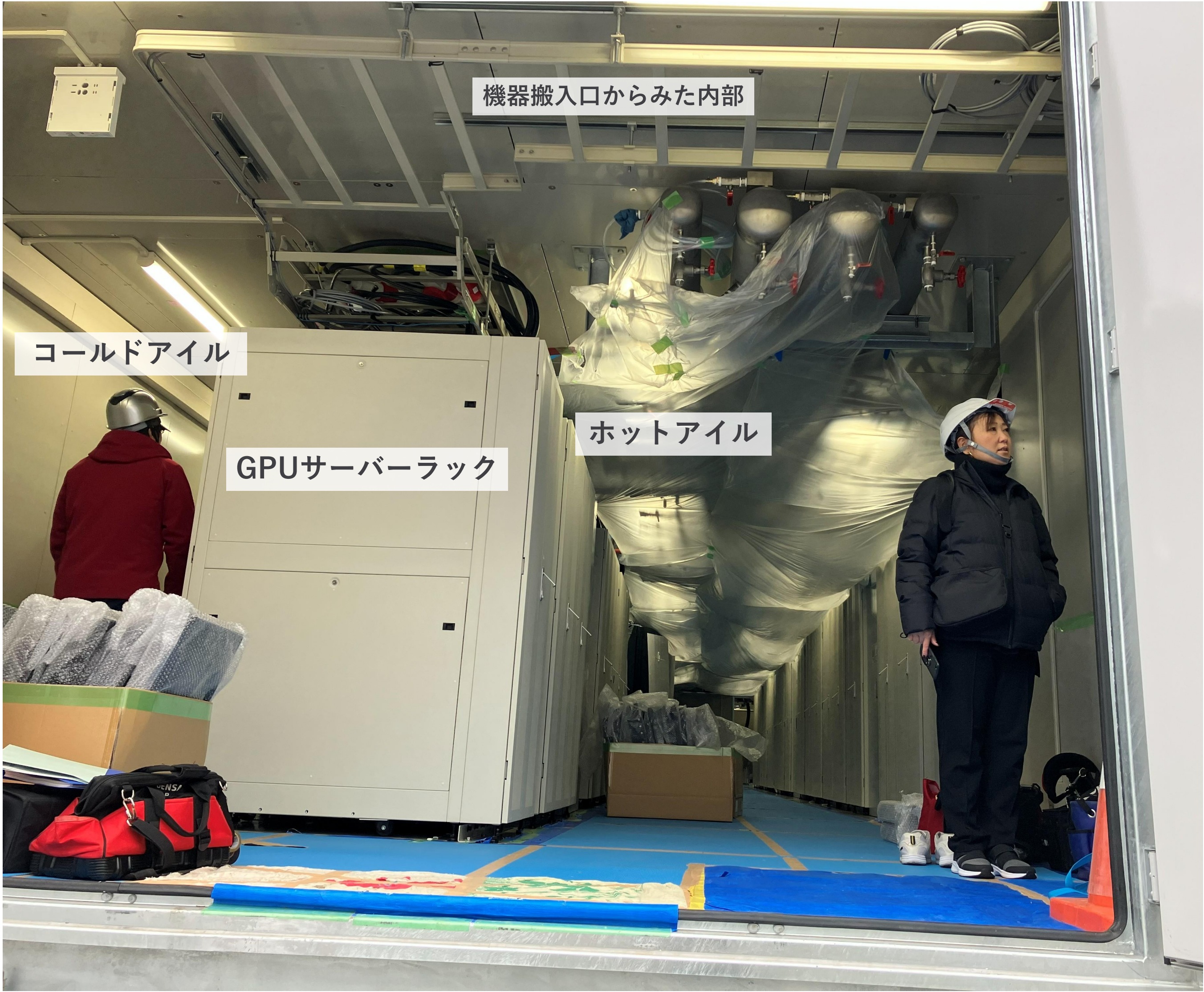
ネットワークラック（ケーブル引き込み用）



拡大



構築の様子



ホットアイル側

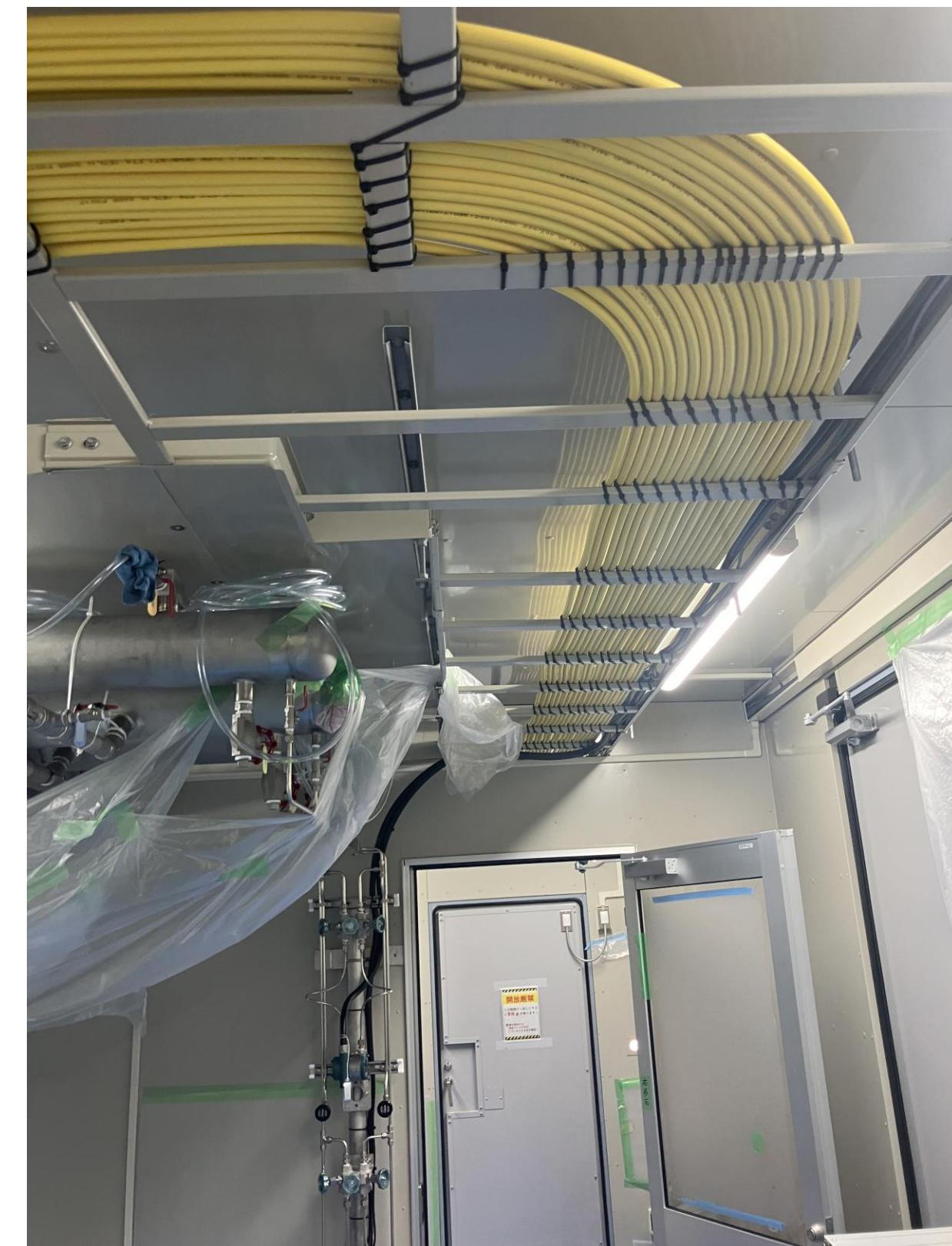


コールドアイル側

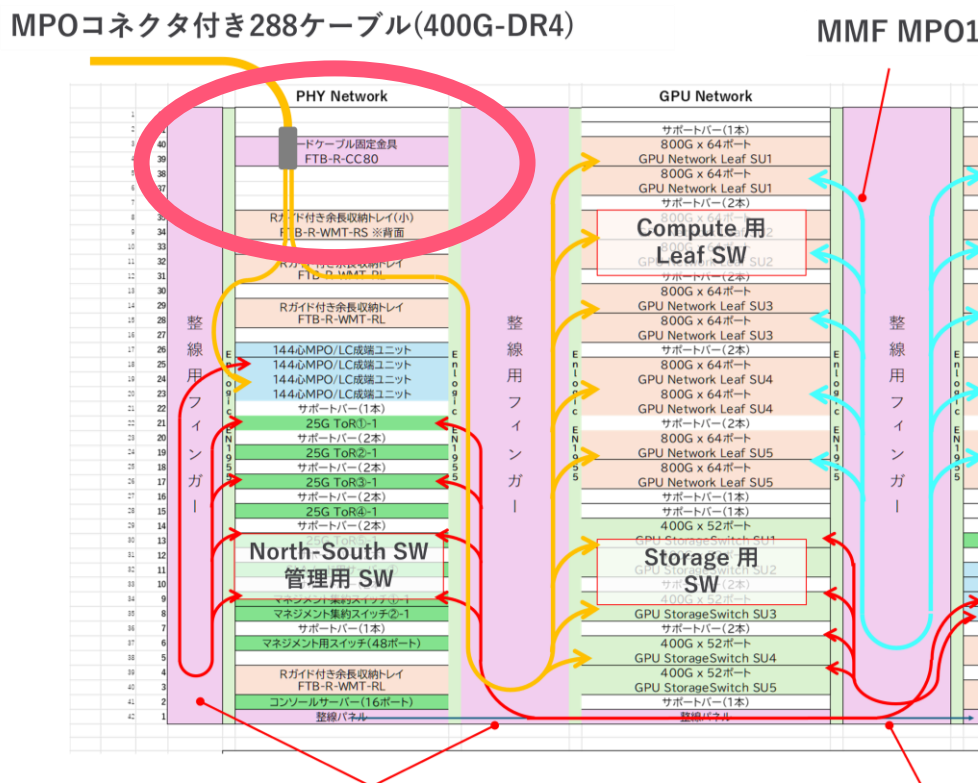


工事中の様子 – ケーブルラダー上の配線

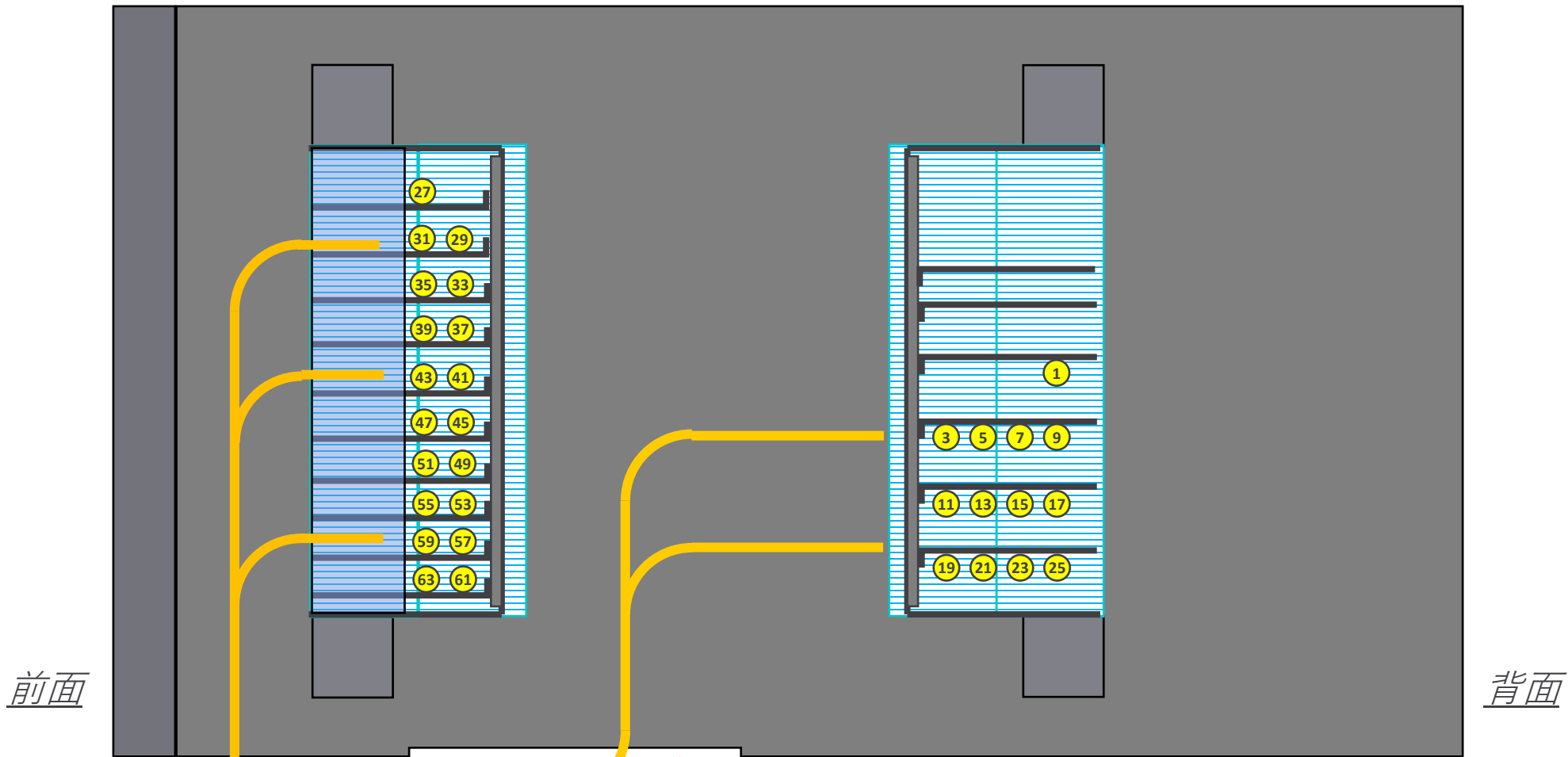
- MPOコネクタ付き光ケーブル(288心) が整列された状態で配線されている
- ケーブルラダー上の光ケーブルの敷設本数は最大32本



- ラック天井板の前面側と背面側の2箇所の開口部から32本の光ケーブルを引き込み
- Leafスイッチおよびストレージ用スイッチへの配線長を考慮してケーブルの引き込み&固定位置決定

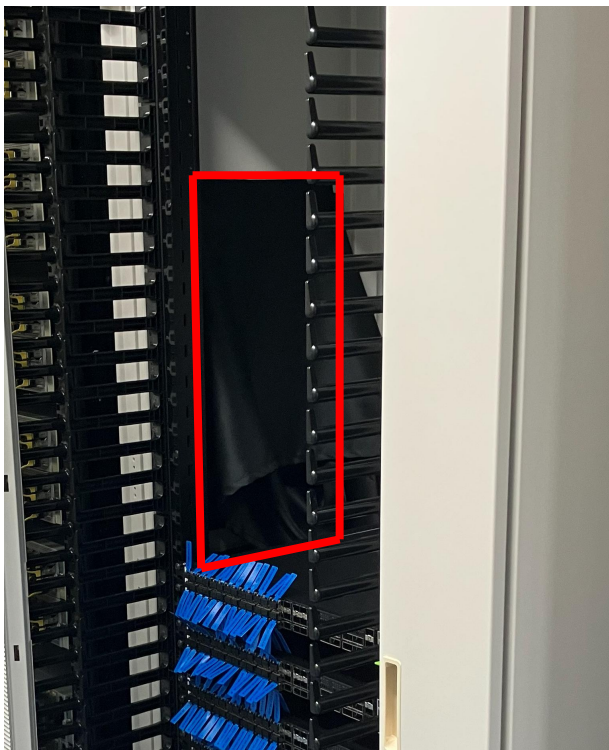


コード長が一律3mのコネクタがスイッチに届くように固定位置を決定



カスタマイズした側面開口部

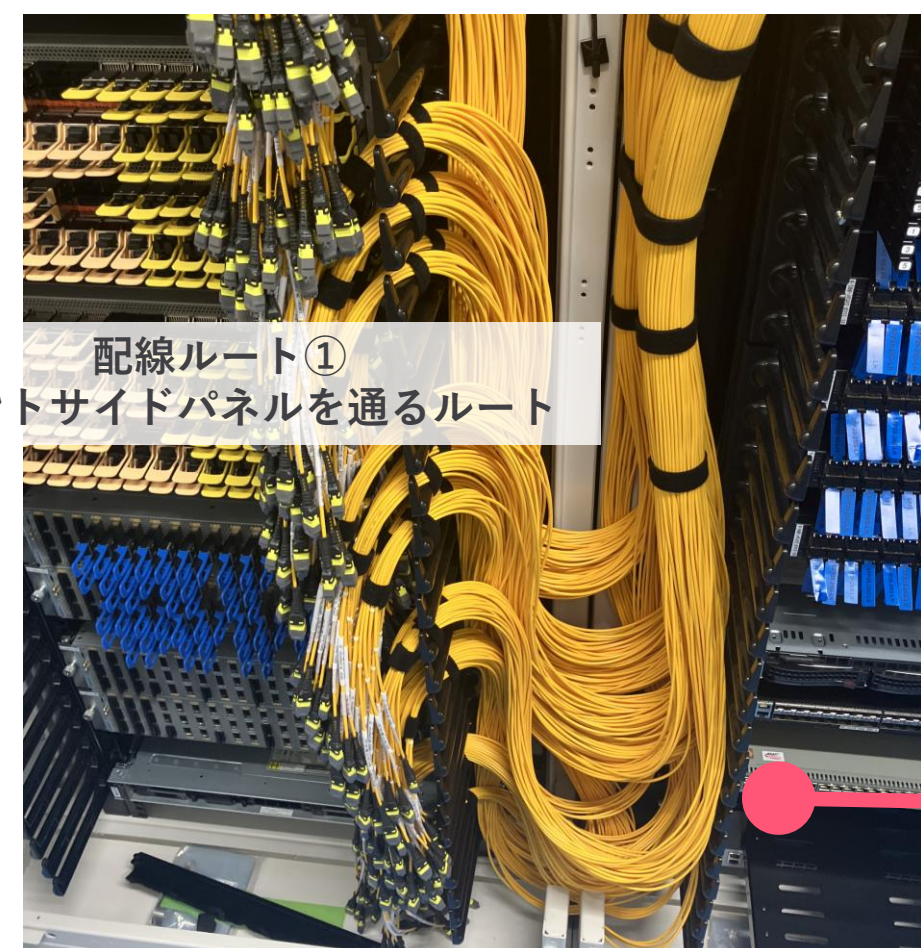
Leaf スwitch/ストレージスイッチ へ



工事中の様子 – ネットワークラックの配線

- コードの余長がラック下に溜まらないようにケーブル固定部の直下の余長収容トレイで余長を吸収
- ラック前面側の側板が外せるため、大量のコードをラック間に配線が可能

リーフスイッチラック



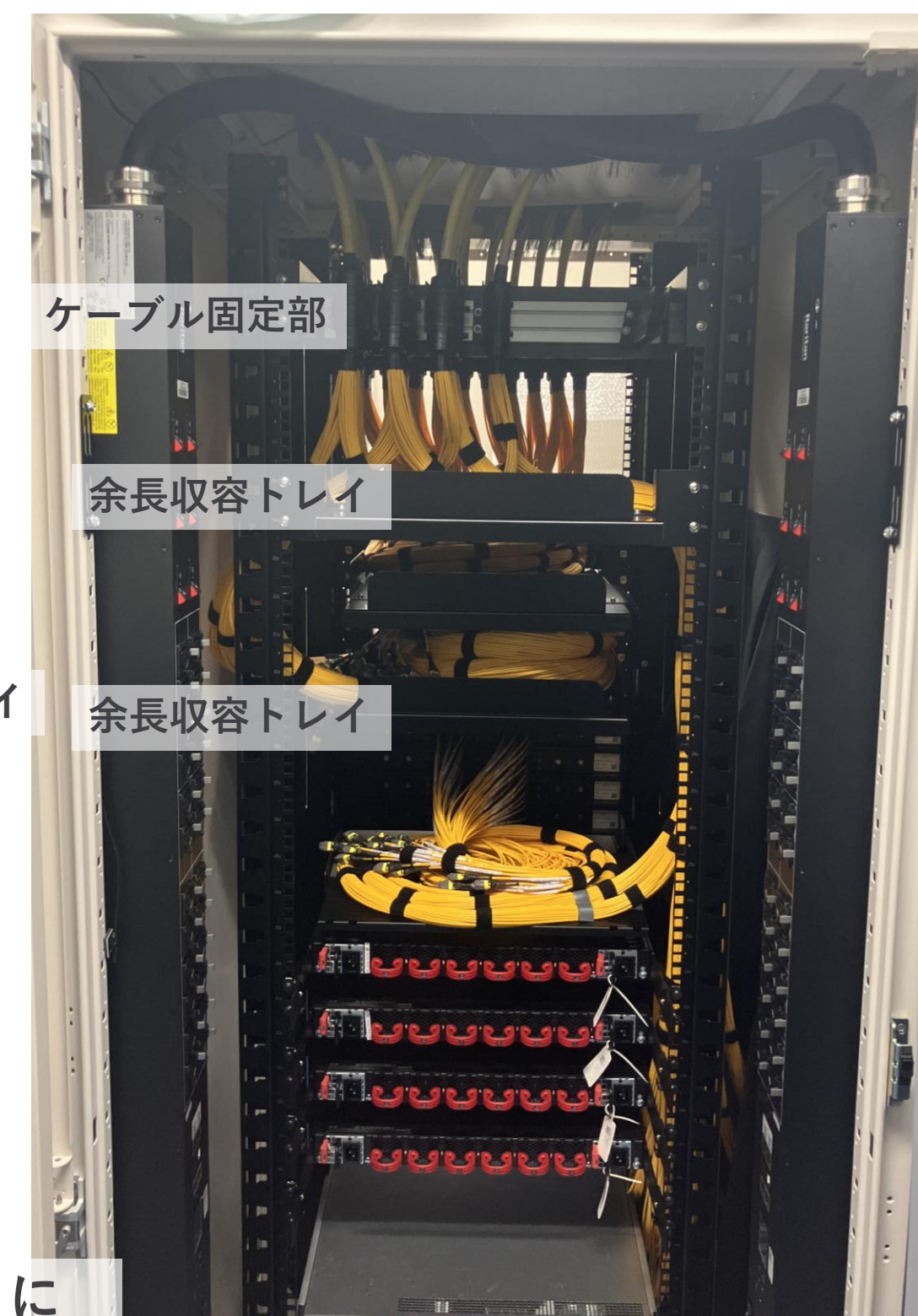
ケーブル引き込みラック



前面側



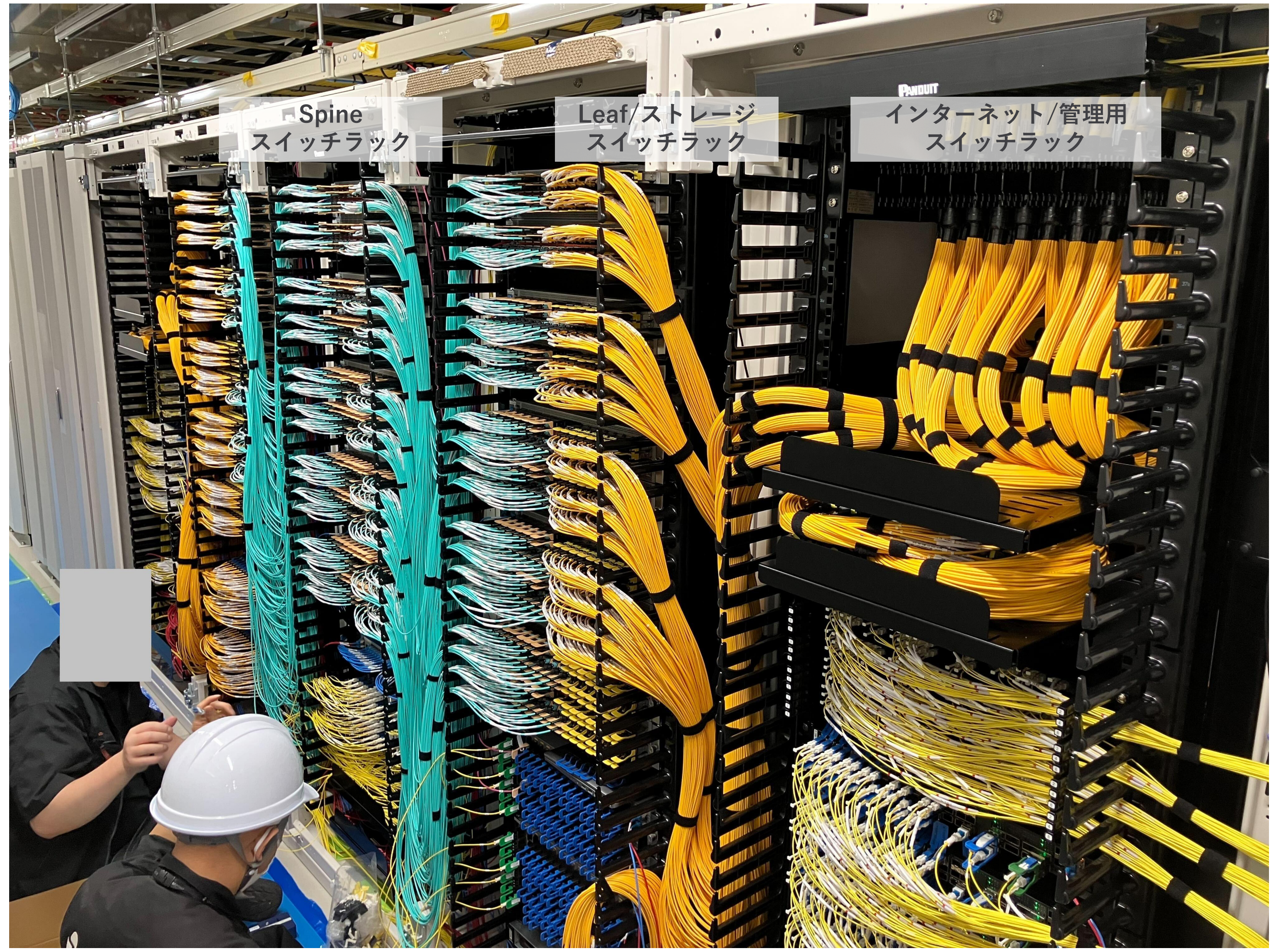
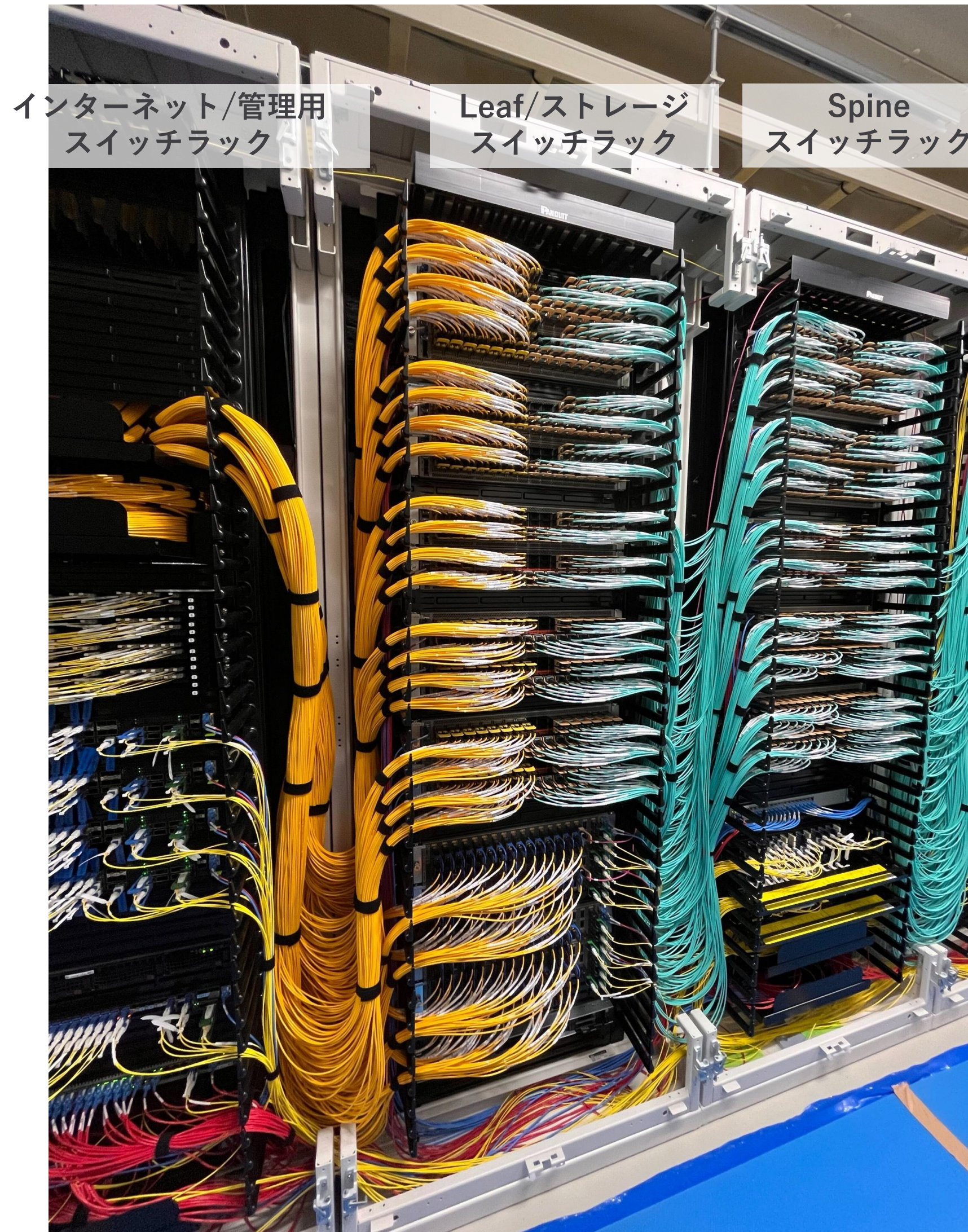
背面側



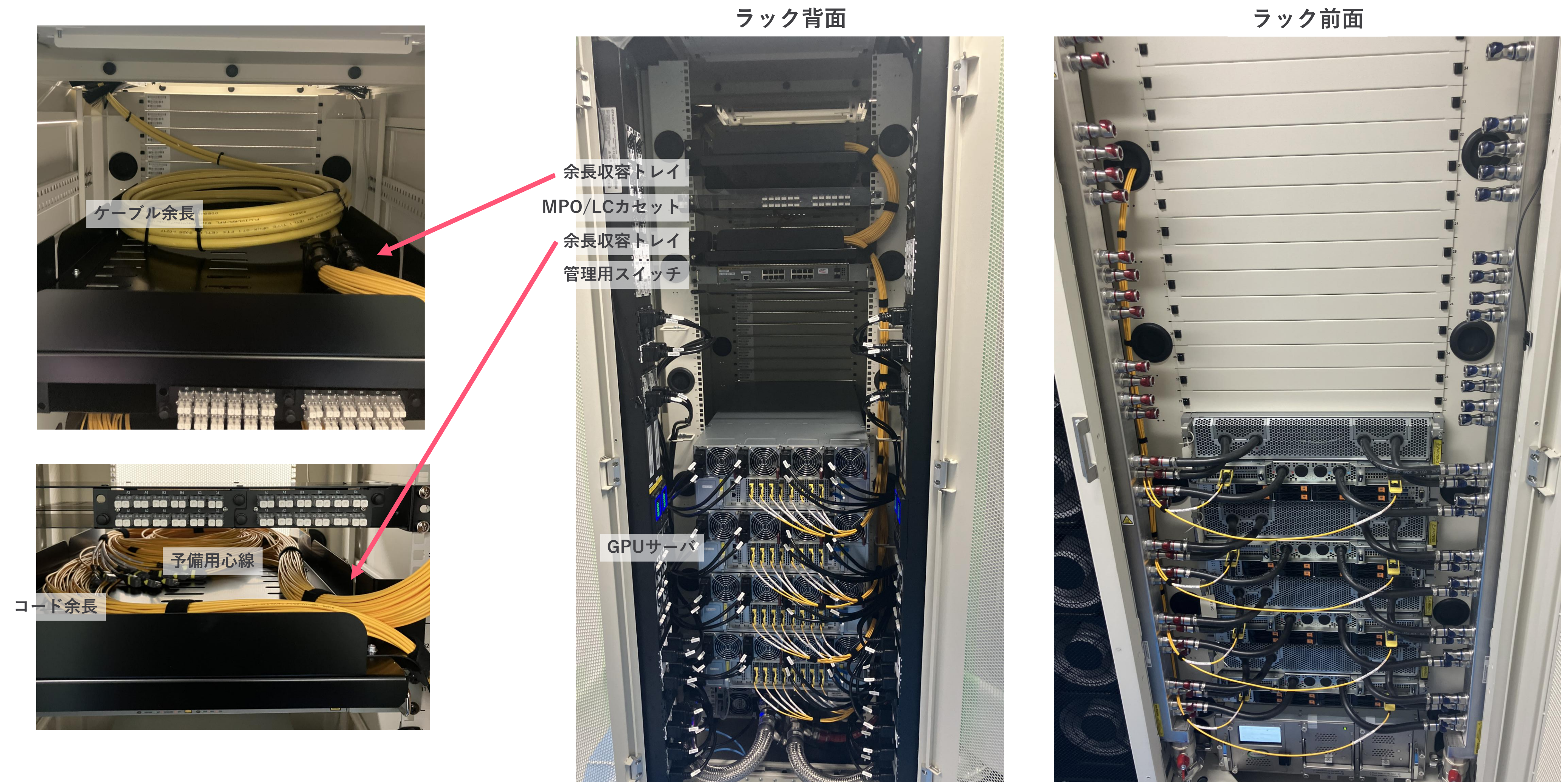
コード余長がラック下に溜まらないように
余長収容トレイ上で余長を調整

工事中の様子 – ネットワークラック全景

- 5つのネットワークラックに約3万心分のコードを配線し収容することができた



- ケーブル余長はGPUサーバラック内に收容、コード余長を余長收容トレイに收容することでラック内の配線は整然としている



ネットワークラック、配線経路を緻密に設計することで無理無駄のない配線に

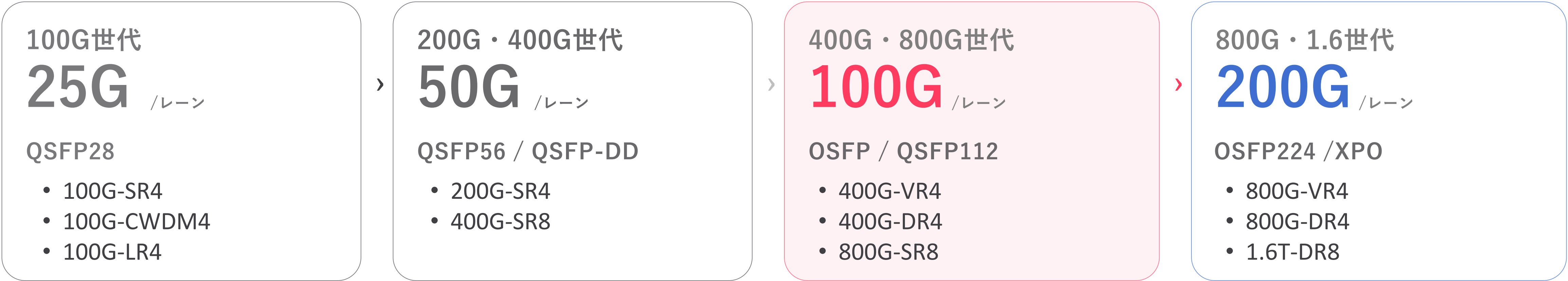


いま見えてきている課題

- 200GSerDes
- LRO・LPO・CPO
- マルチプレーン
- シャフルアーキテクチャ
- 光配線技術の多様化

AIインフラ向けとしては今がちょうど変化の過渡期

- Broadcom Tomahawk5 から Tomahawk6 へ、 NVIDIA Spectrum4からSpectrum6、 Connect-X7/8 から ConnectX-9 へ
- スイッチやNICのSerDes、あるいはフォームファクタは何を採用しているのかを気に掛ける必要がある
- Gearboxによって変換されるトランシーバーもあるのでトランシーバーの選択肢が増える



Gearboxによる変換イメージ



SerDes(Serializer/Deserializer) = 光トランシーバーにおける 1電気レーンあたりのスピード

トランシーバーの消費電力を下げる流れで普及しつつある

方式	特徴	消費電力
FRO Fully Retimed Optics (従来型)	送受信ともにDSPが搭載された従来型	高
LRO Linear Receive Optics	送信側のみDSPを残したもの 運用に注意は必要なものの、まだ相互運用可？	
LPO Linear Pluggable Optics	DSPを完全に排除したもの スイッチごとにカスタム、ベンダーロックイン強め	
CPO Co-Packaged Optics	完全にスイッチに固定	低

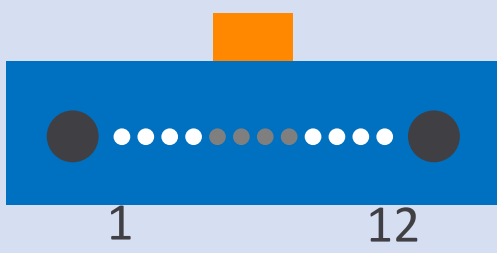
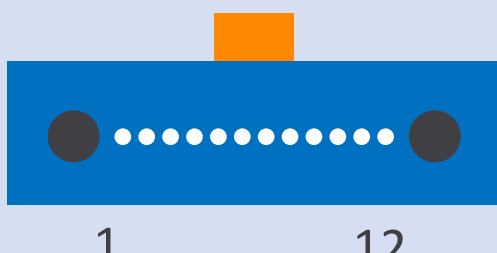
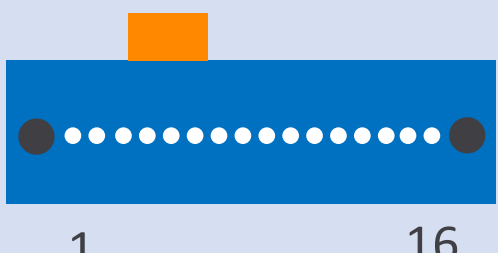
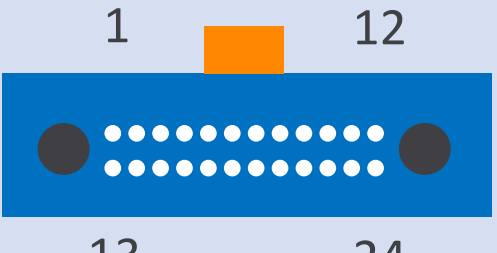
● LPOの対向はLPO？ FRO+LPO、LRO + LPO などの組み合わせは？

- LPOの対向はLPO？ FRO+LPO、 LRO + LPO などの組み合わせは？
- サードパーティLPOは？
- CPOの運用性は？

MPOの構造化配線の構築と運用には高い知識が必要

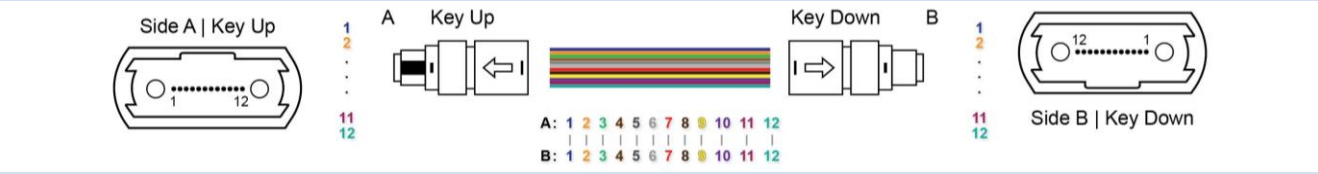
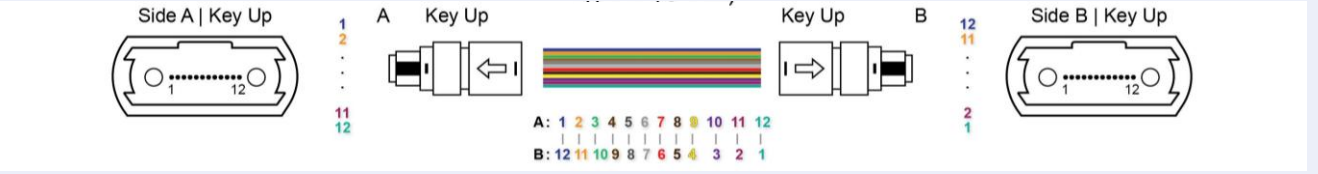
MPOの構造化配線は多種多様で選択肢が多いことから、生成AI等のスケールアウトのシステムに採用する場合は高い知識を必要とする

- 基本単位となるファイバ心数およびMPOコネクタの選定を行う

	BASE-8	BASE-12	BASE-16	BASE-24
MPO	12MPO(8心) 	12MPO 	16MPO 	24MPO 
主な用途	2心伝送 8心伝送(SR4, DR4)	2心伝送 8心伝送(SR4, DR4) 20心伝送(SR10等)	8心伝送(SR4, DR4) 16心伝送(SR8, DR8)	2心伝送 20心伝送(SR10等)

- 各配線方式により製品の組み合わせが異なる

BASE-12を用いたペアリンク(2心)における各配線方式の製品組み合わせ

方式	MPOトランク	両端カセット(MPO/LC)	両端パッチコード(Duplex LC)
Method A	Type A 	左右同一	左右で異なる(Type A, Type B)
Method B	Type B 	左右で異なるカセット	Type B
Universal	Type B ↑	左右同一(ユニバーサル配線)	Type B
Method C	Type C 	左右同一	Type B

MPOの構造化配線の構築と運用には高い知識が必要

- MPOのジェンダー、研磨方式、MPOアダプタとの関係
 - MPOコネクタにはピン無しとピン有りがある、これらをMPOアダプタを介して接続する
 - MPOコネクタの研磨方式にはフラットPC研磨とAPC研磨がある
 - 昨今、マルチモード用MPOコネクタAPC研磨が登場し、接続には注意が必要

MPOコネクタのジェンダー

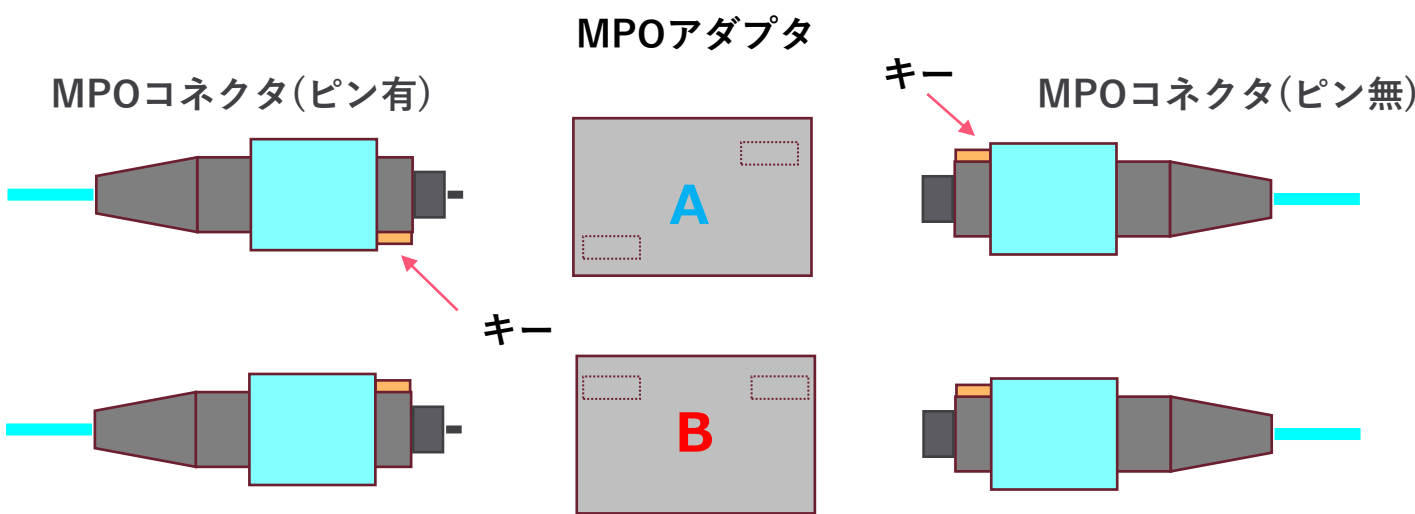


MPOコネクタと研磨方式

	フラットPC研磨	APC研磨
MPO(MM用)	○	○
MPO(SM用)	—	○

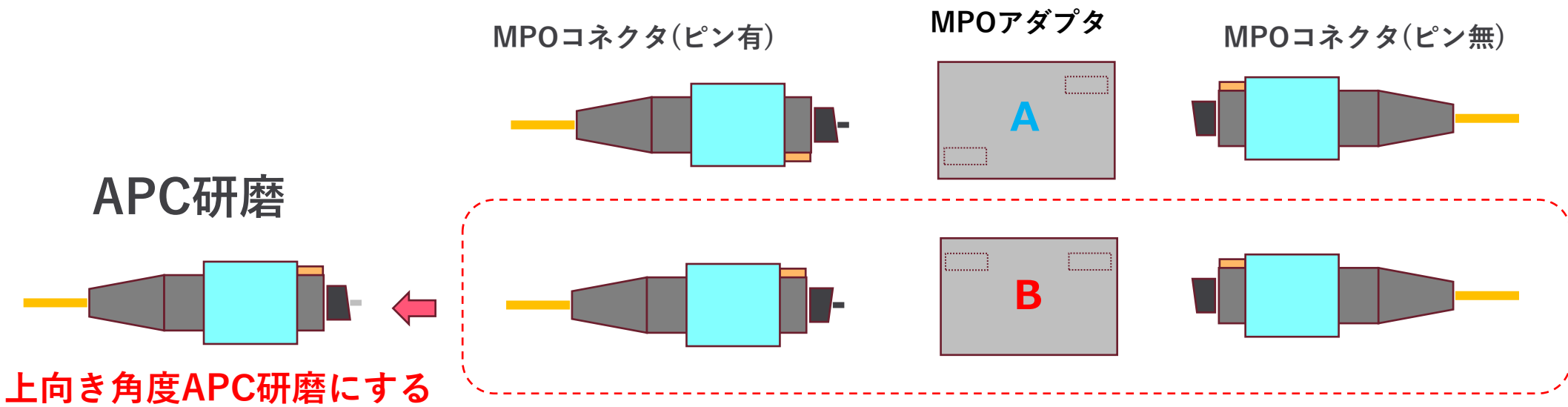
MPOコネクタと研磨方式、その適用アダプタ

フラットPC研磨



Type A、Type Bで接続可能

APC研磨



Type Bのアダプタにおいて、下向き角度APC研磨同士ではフェルールがコンタクト出来ない！！

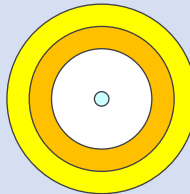
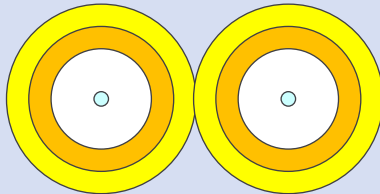
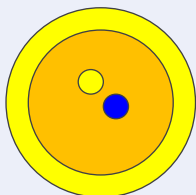
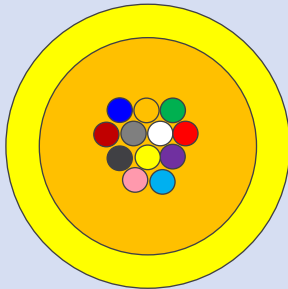
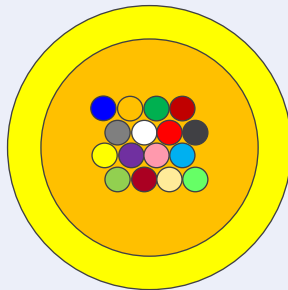
MPOの構造化配線の構築と運用には高い知識が必要

MPO配線システムの注意点

項目	説明
極性・ジェンダーの管理が複雑	・ MPOコネクタやLCコネクタの極性、ジェンダーを正しく合わせる必要がある ・ 極性のタイプ、アダプタのタイプ、ケーブルの接続順に注意が必要 ・ 間違えると、通信不能、コネクタの破損に繋がる
高精度なアライメントが必要	・ MPOコネクタは多心一括接続のため、ファイバの軸ズレにシビア ・ 製品には高い品質が求められる
可視性が低い	・ 多心を一括で接続しているため、問題が生じているファイバを特定しにくい ・ MPO専用テスターや専用測定コードが必要になる ・ トラブルシューティングには高い知識と多様な機材が必要になる（テスター、コード類等）
ラインナップ不足、納期の長期化	・ 品種によっては、ラインナップが限られる（例：16MPOソリューション） ・ 納期もかかる製品もある
清掃が非常に重要	・ MPOの多心コネクタは埃や汚れの影響を受けやすい（1心でも不具合があるとダメ） ・ クリーナを使いこなす知識が必要（乾式、湿式） ・ 端面観察器の必要性が高い
ケーブル余長対策	・ MPO付きトランクケーブルは現場で長さ変更ができないため、正確な条長設計が必要 ・ MPO付きトランクケーブルの余長吸収対策の検討が必要

FR、DR、SRによる配線ボリュームの比較

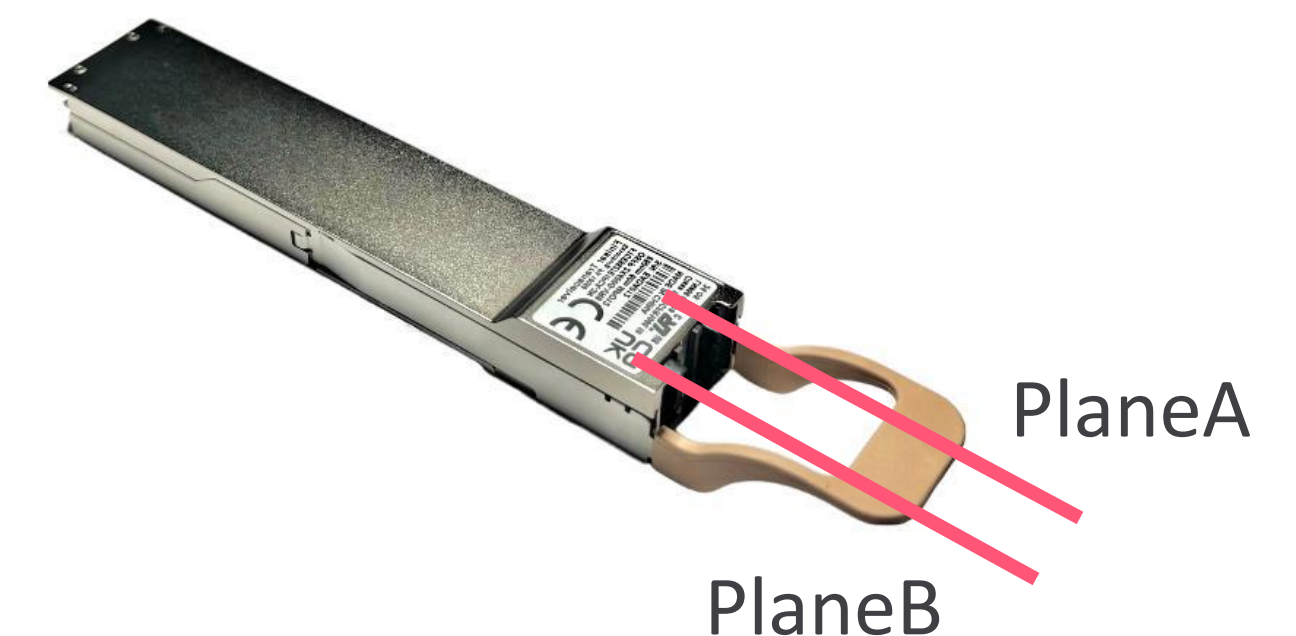
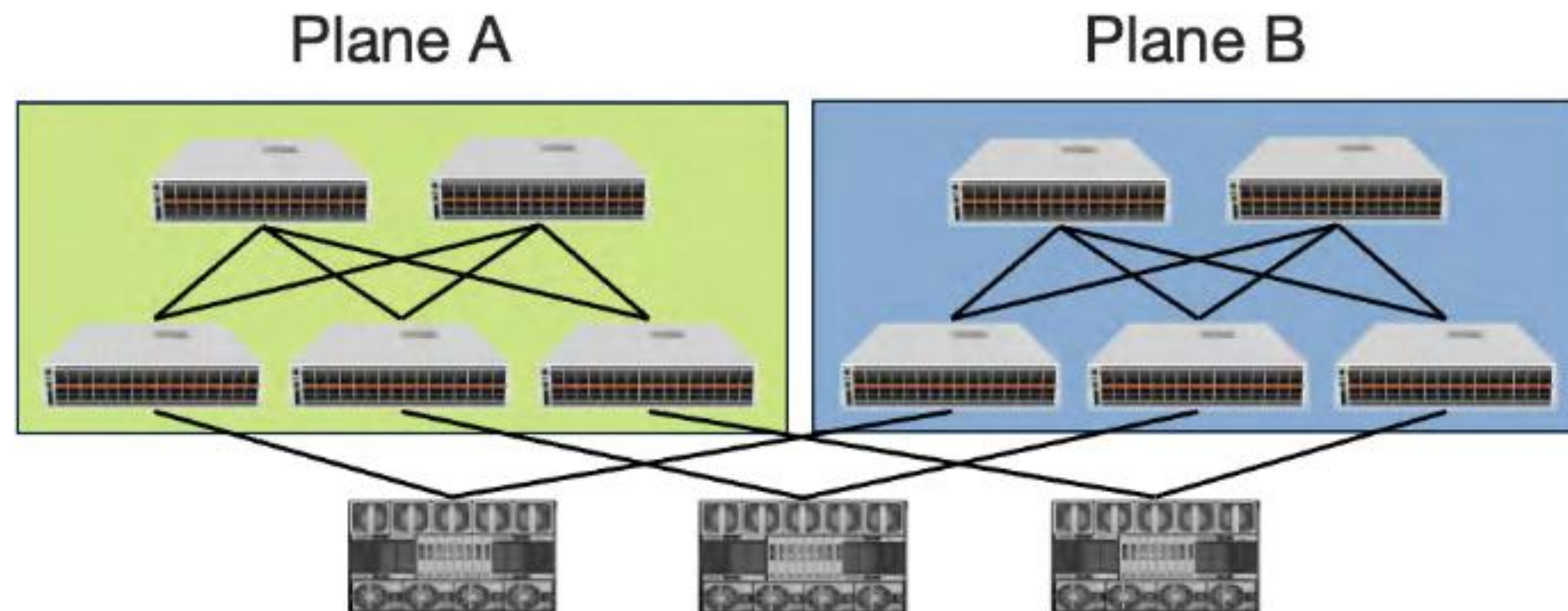
- MPOは扱いにくいのでWDMオプティクスによるLC2心規格を利用すればよいのでは？ → ファイバ数は1/4や1/8に削減可能
- 配線ボリュームで比較すると…？

主な規格	ファイバ数	コード形状 (一般的なサイズ)		断面積比較 (リンクあたり)
400GBASE-FR4 (*WDMオプティクスを使用)	2心 対SR4/DR4：1/4 対SR8/DR8：1/8	 単心コード (φ2mm)	 Zipコード (φ2mm×2)	対Duallinkコード：2倍 対12心コード：0.89倍 対16心コード：0.89倍
		 Duallinkコード (φ2mm)		対単心/Zipコード：1/2 対12心コード：0.44倍 対16心コード：0.44倍
400GBASE-SR4 400GBASE-DR4 800GBASE-DR4	8心 対FR：4倍 対SR8/DR8：1/2		12心コード (φ3mm)	対単心/Zipコード：1.1倍 対Duallinkコード：2.3倍 対16心コード：1倍
400GBASE-SR8 800GBASE-SR8 800GBASE-DR8	16心 対FR：8倍 対SR4/DR4：2倍		16心コード (φ3mm)	対単心/Zipコード：1.1倍 対Duallinkコード：2.3倍 対12心コード：1倍

ラック内配線やラック間のコネクタ付きコード集合タイプのケーブル配線の場合は配線ボリュームはあまり変わらない

NVIDIA B300世代 ConnectX-8からマルチプレーンという選択肢が登場

- 複数の独立したClos Fabric(プレーン) によって構成されるネットワーク
 - プレーンは物理的にも論理的にも独立している
 - いずれかのプレーンで発生した障害は他のプレーンに影響を与えない = レジリエンシの向上が期待できる
 - 対応したNICが必要 (Packet Splay / Packet Reordering / パス監視… など)
 - NICベンダーのプロプライエタリな仕組みに依存していた



マルチプレーンを前提としたオープンなプロトコル

- OpenAIが主導し、AMD / Broadcom / Intel / Microsoft / NVIDIA と連携して開発したプロトコル
- UECによるUETの仕様も一部取り込まれている (packet spraying / congestion notification / packet trimming)
- SRv6のソースルーティングを活用
 - 各プレーン内のパス障害や輻輳シグナル検知による迅速な迂回やパス除外
- すでにOpenAIやMicrosoft、Oracleなどによる大規模な実装例もあり

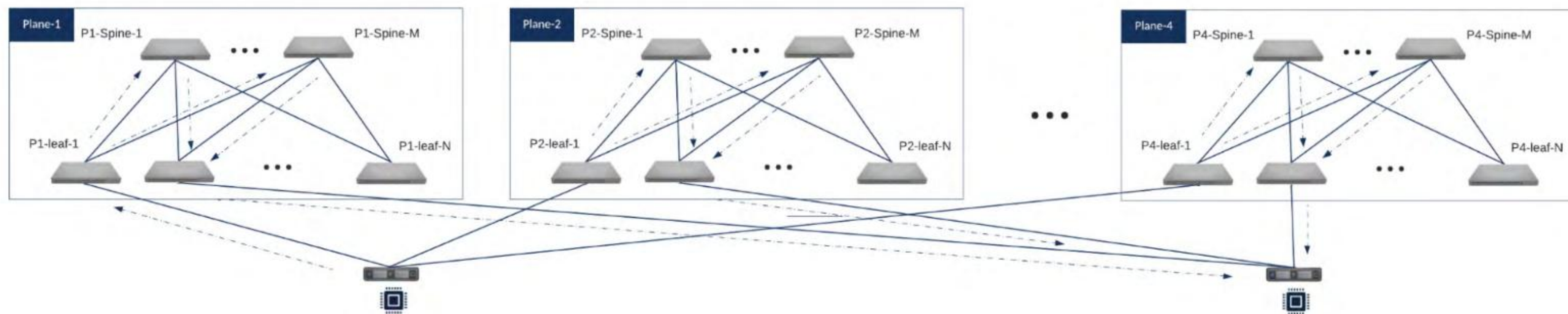


Figure 3: Diagram of the multi-planar transport of MRC

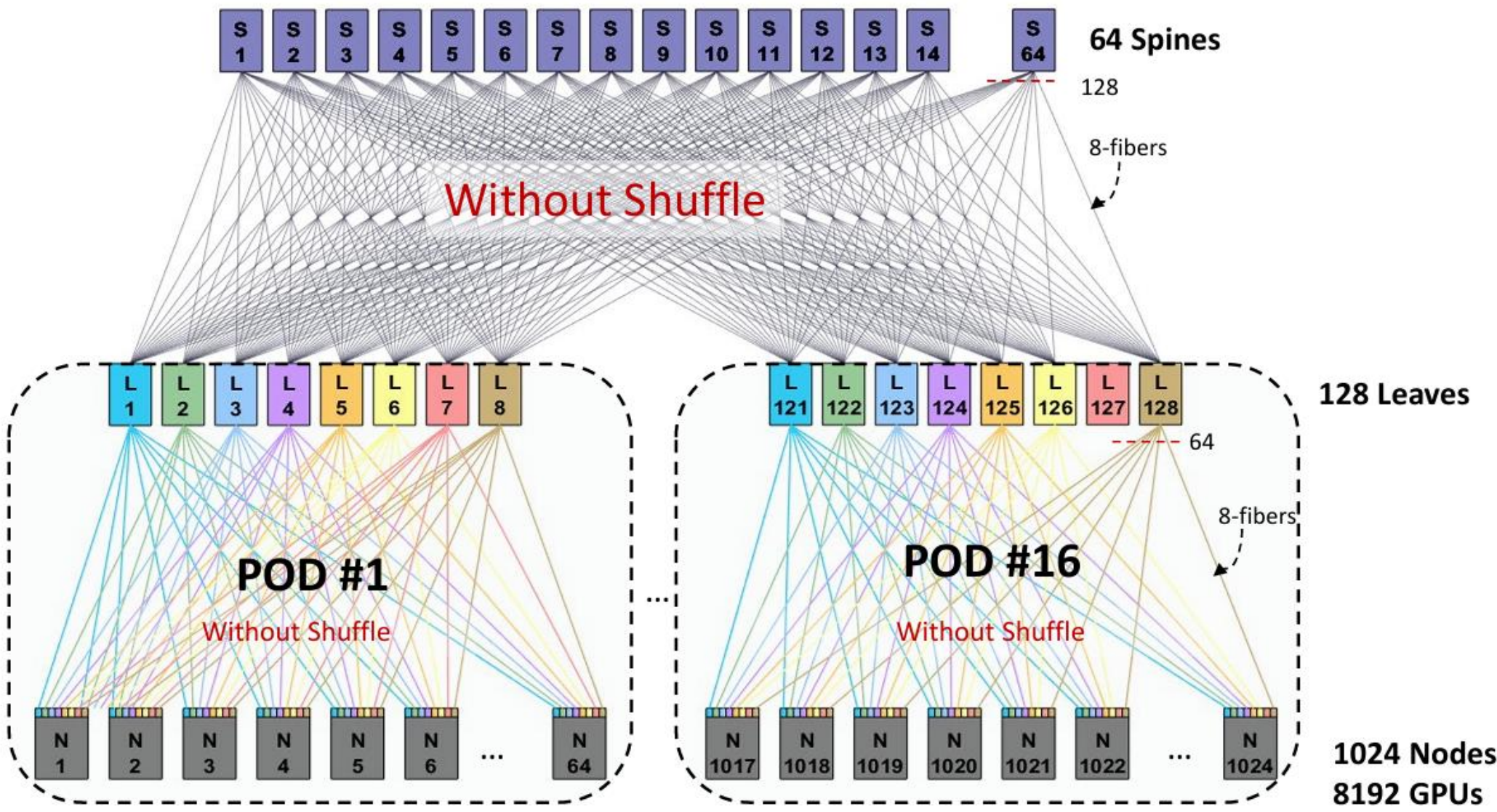
光ファイバーを「シャフル」することでGPUクラスタの大規模化を実現

- ・ シャフル配線によりバックエンドのスイッチ層を増やさずにGPUクラスタの大規模化が可能
- ・ GPUクラスタ大規模化と低遅延の両立を実現

Leaf – Spine Network (シャフル配線無し)

AI Cluster with 8192 GPUs

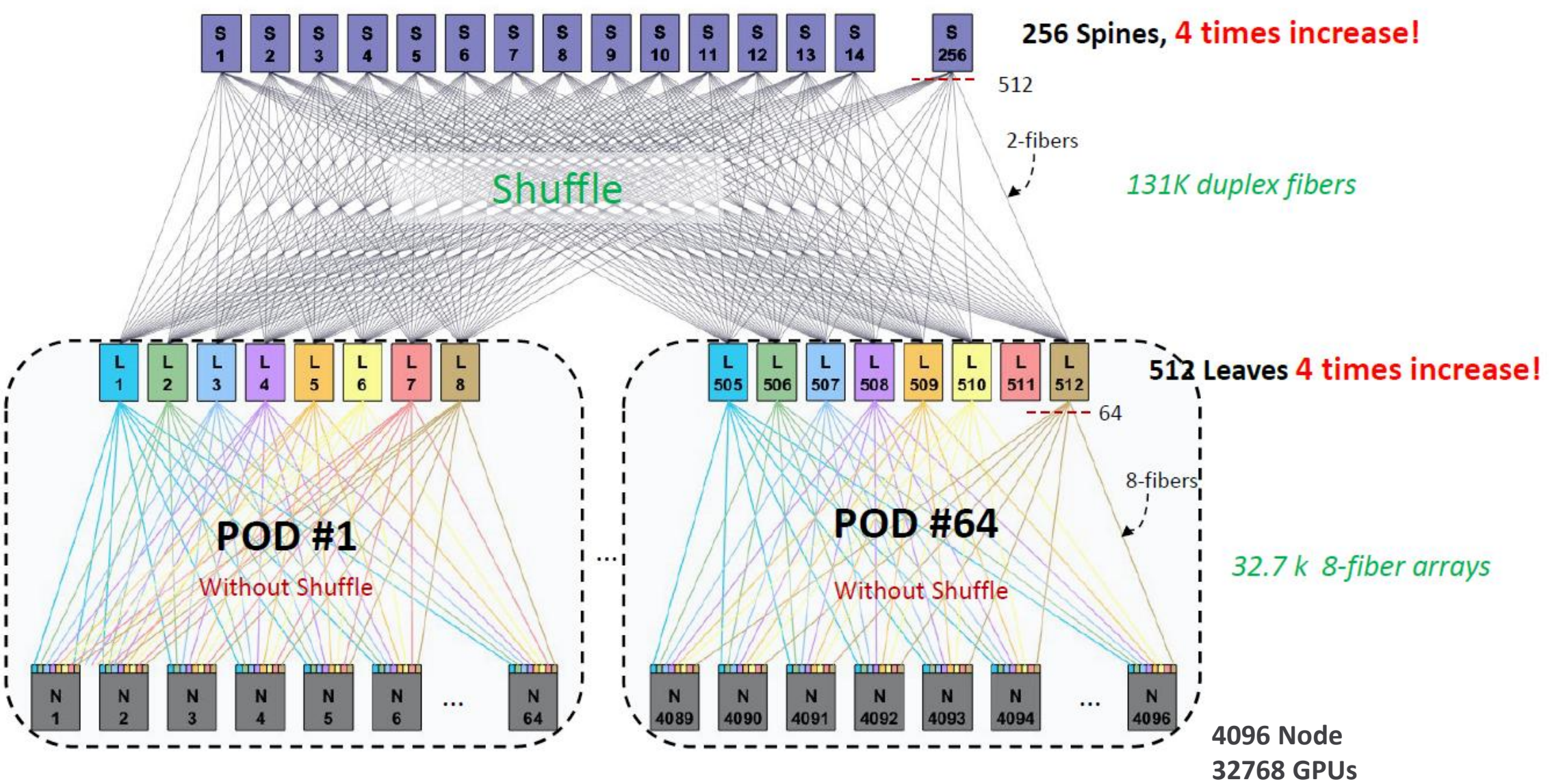
N: Node with 8 GPUs
L: Leaf , S: Spine
All switches have a radix of 512



Leaf – Spine Network (シャフル配線有り)

AI Cluster with 32768 GPUs

N: Node with 8 GPUs
L: Leaf , S: Spine
All switches have a radix of 512

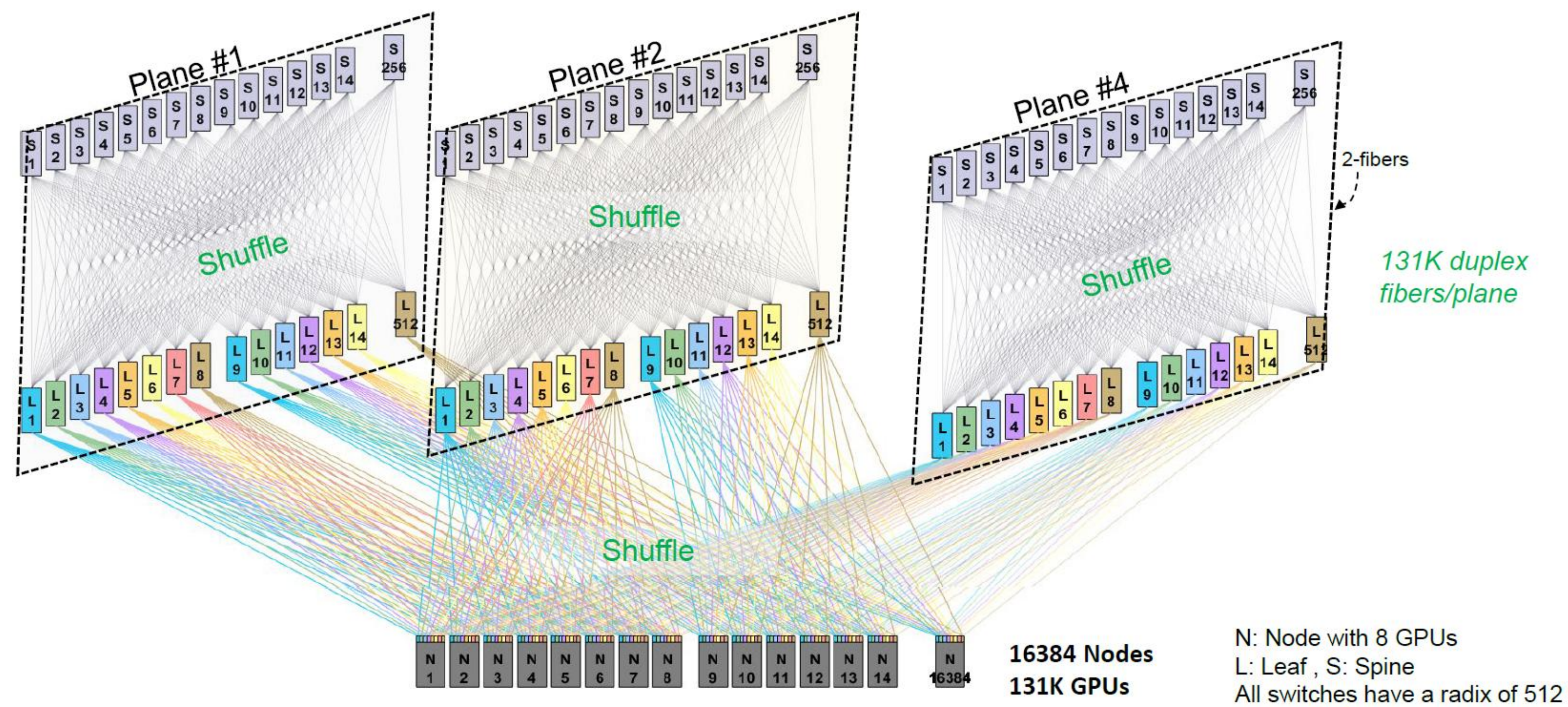


Source : Optical Shuffle Architectures for large AI Networks
https://www.ieee802.org/3/ad_hoc/E4AI/public/26_0224/castro_e4ai_02_260224.pdf

シャフルとマルチプレーンにより耐障害性向上とGPUクラスタの大規模化が可能に

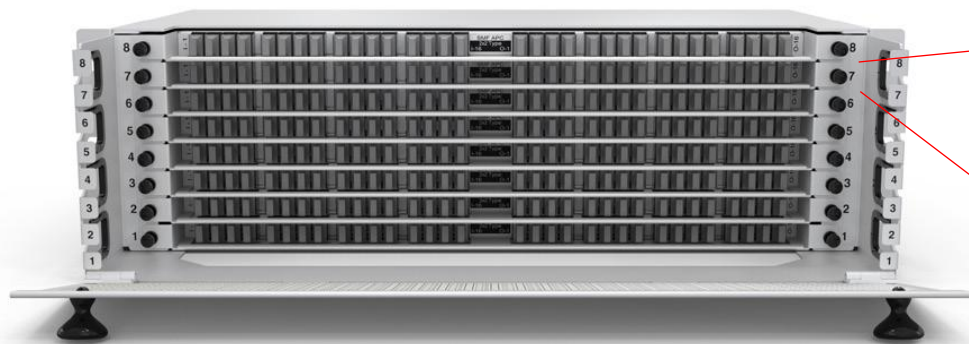
2-tier leaf-spine network (quad plane)

AI Cluster with 131K GPUs
Adding Node to Leaf Lane Breakouts increases the number of GPUs 4 times



Source : Optical Shuffle Architectures for large AI Networks
https://www.ieee802.org/3/ad_hoc/E4AI/public/26_0224/castro_e4ai_02_260224.pdf

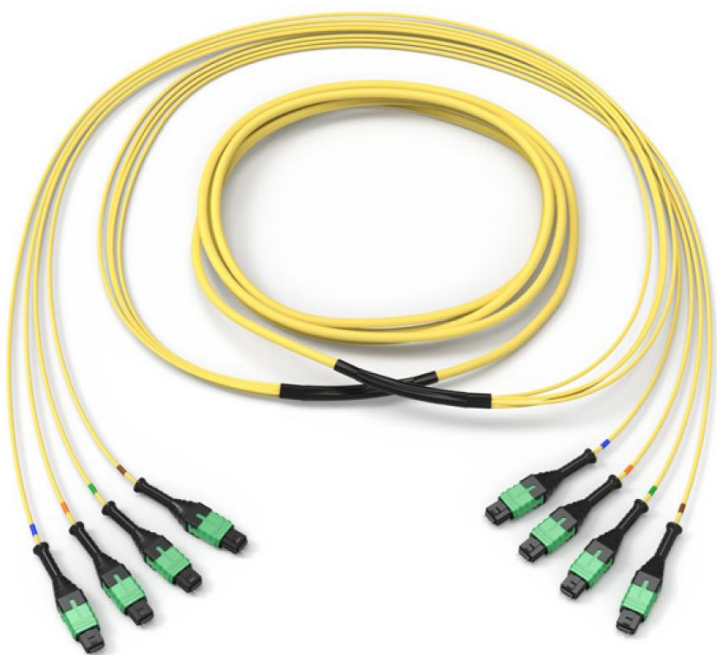
シャフル配線を実現する2種類の仕組み BOXタイプとコネクタアッセンブリタイプ



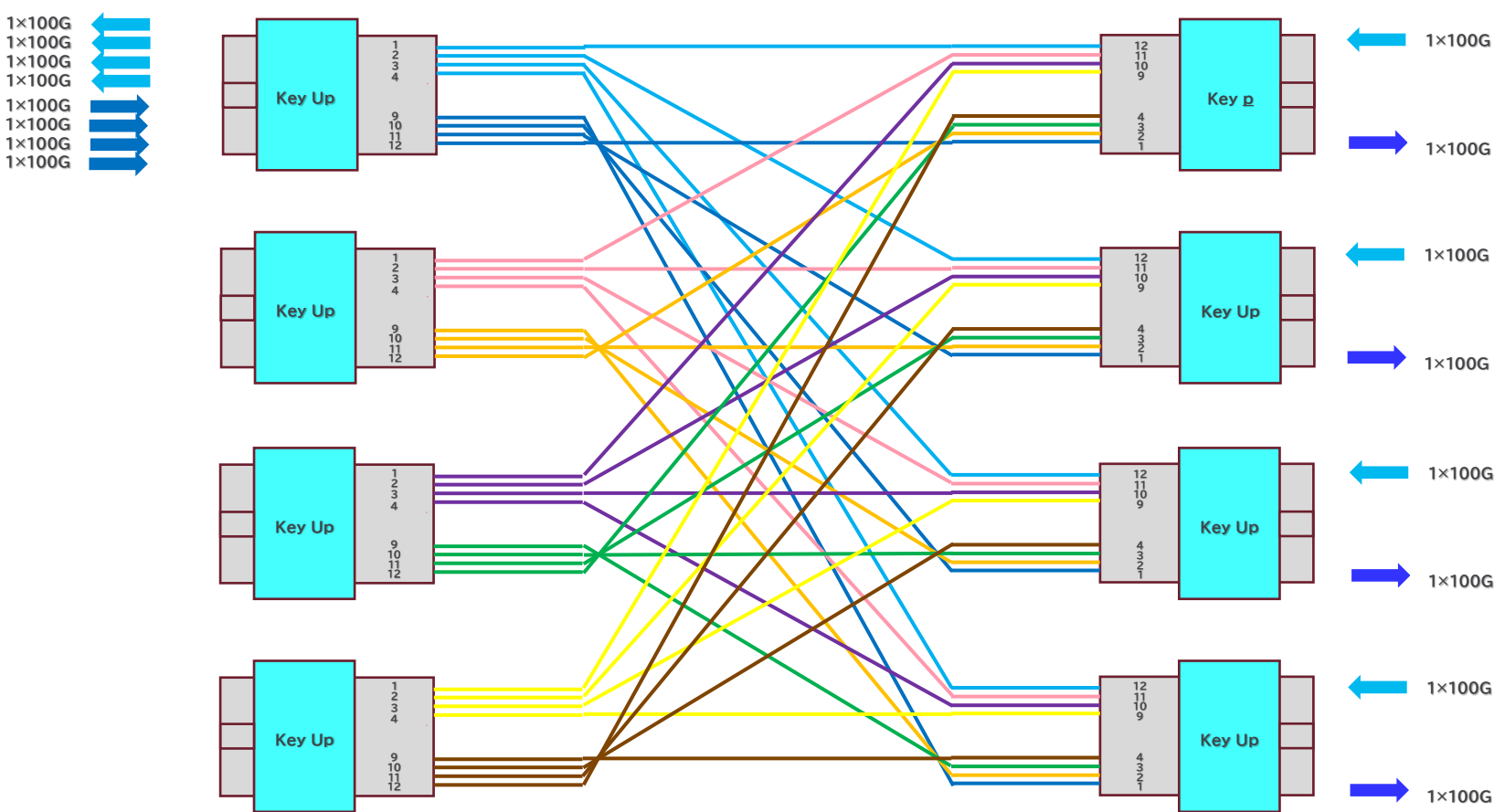
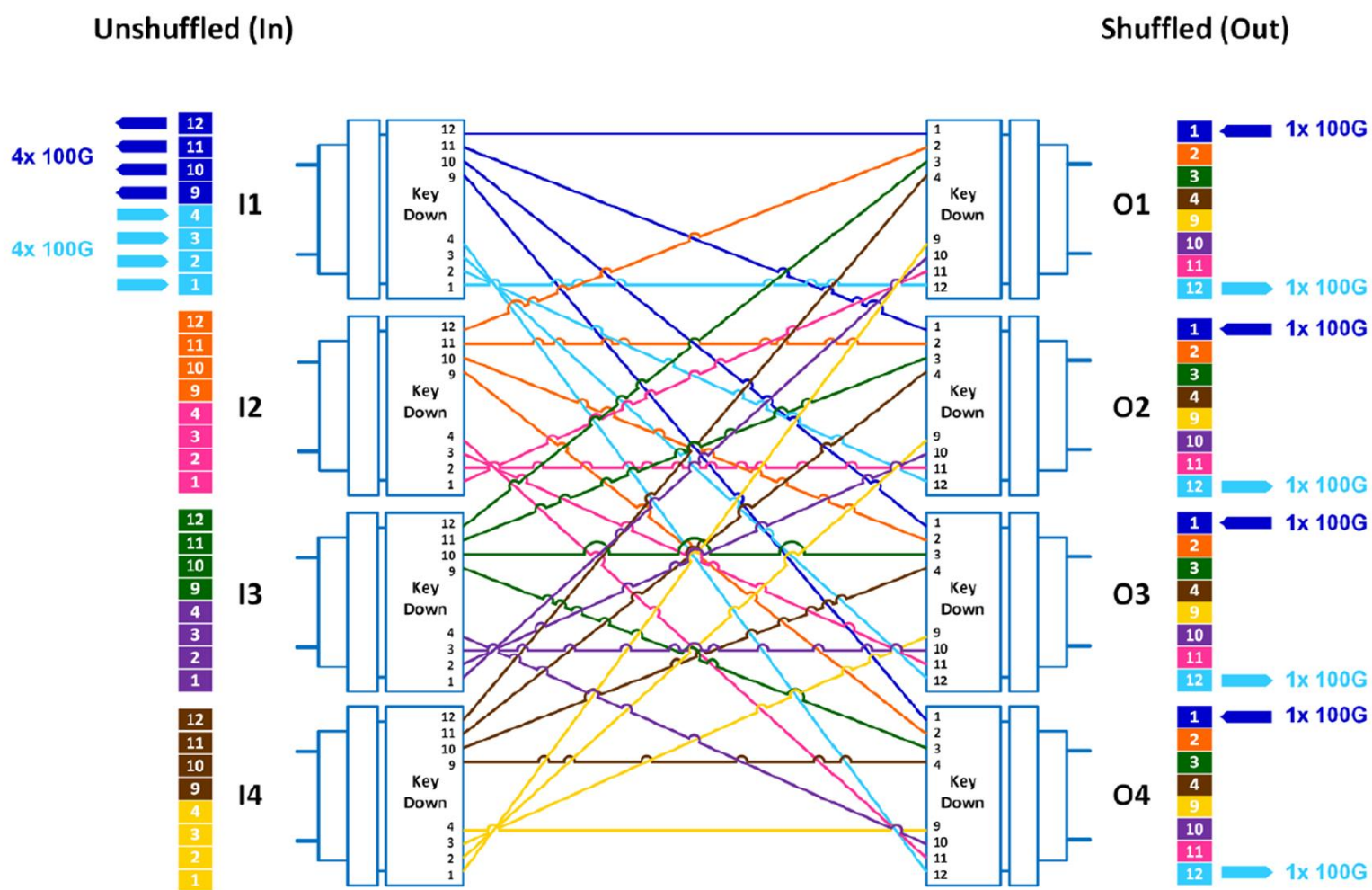
シャフルBox



シャフルモジュール

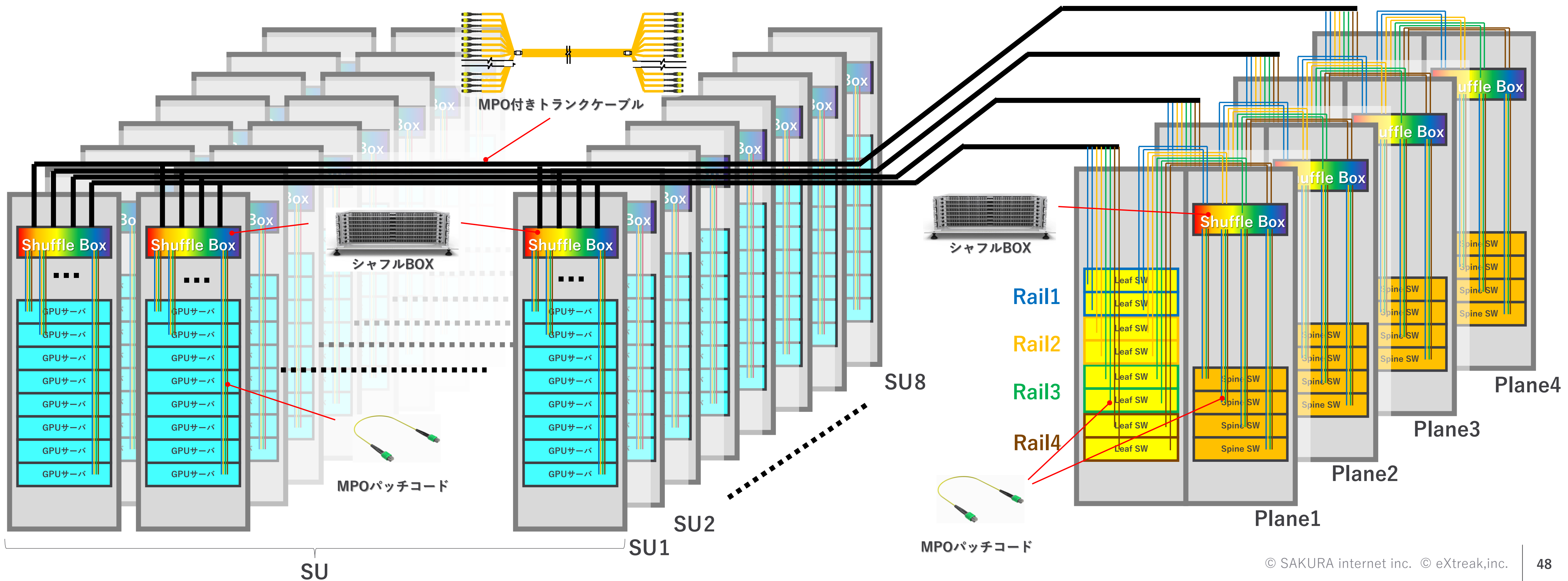


4×4 シャフルアッセンブリ



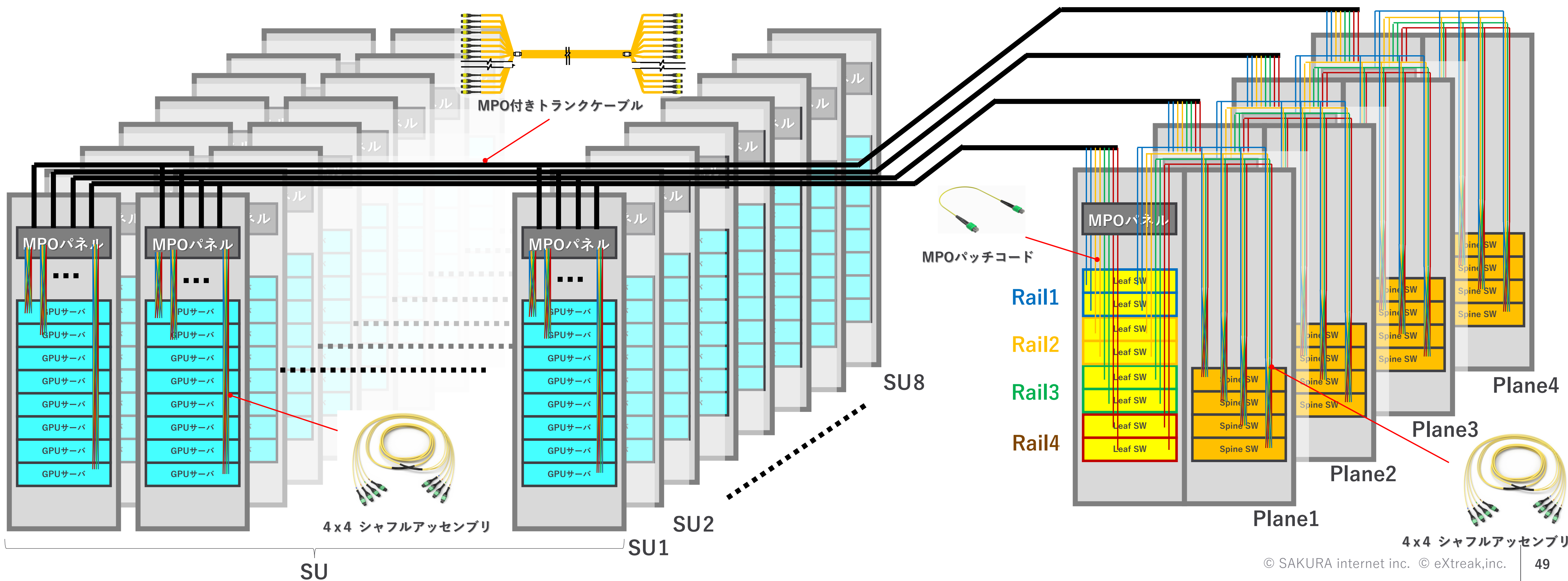
シャフル配線を用いた生成AIの光配線構成例（シャフルBOXを用いた場合）

- シャフルBOXはラック内に収容
- ラック内は通常のパッチコードを配線するため余長処理はシャフル配線無しの場合と変わらない



シャフル配線を用いた生成AIの光配線構成例（シャフルアッセンブリを用いた場合）

- シャフルアッセンブリはラック内配線および隣接ラック間配線に使用
- シャフルアッセンブリは端末がBreakout構造のため、Breakoutの余長処理方法を検討しておく必要がある



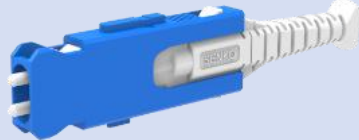
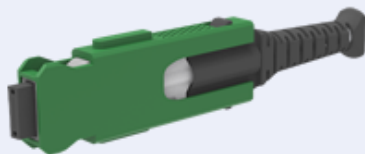
シャフル配線の導入にあたり考えるべきこと

- ファイバがシャフルされるため伝送経路が複雑になる（ファイバのマッピングを理解する必要がある）
- 光源パワーメーター法による設置後の損失測定においては、測定知識が必要、測定の手間も通常の配線より時間を要する
- 伝送経路に不具合が生じた場合、不具合個所を特定に時間を要する
- シャフル配線のコネクタ 1 か所が不具合だった場合でも、シャフル配線すべてが使えなくなる（不具合時の影響が大きい）
- シャフルBOXとシャフルアッセンブリの長所・短所を理解した上で選定することが必要

項目	シャフルBOX/Module	シャフルアッセンブリ
構成変更のしやすさ	高 パッチコードによるクロスコネクトにより変更は容易	低 パッチコードがシャフルされているため変更は難し
保守運用時の視認性	高 シャフルのポートが見えやすい	低～中 シャフル部がケーブル内に隠れる
コネクタ接続点	増える	増えにくい
伝送路損失への影響	接続点増加で高くなりやすい	伝送路損失への影響は少ない
施工性	通常パッチコードの配線と同様	パッチコードがBreakout構造のため余長処理が難しい
設置スペース	シャフルBOX/Moduleの実装スペースが必要	設置スペースへの影響は少ない

小型化 & 多心化の流れ → VSFFの浸透
生成AIのGPUクラスタには今後どの光コネクタが使われていくのか

- データコム市場においてはLCコネクタおよびMPOコネクタが主に使われてきた
- Duplex ConnectorタイプのVSFFインターフェースをもつ光トランシーバが徐々に浸透
- CPOスイッチにはMMCコネクタが採用
- GPUクラスタの大規模化による高密度配線の浸透と Multi Fiber ConnectorタイプのVSFFの需要の高まり

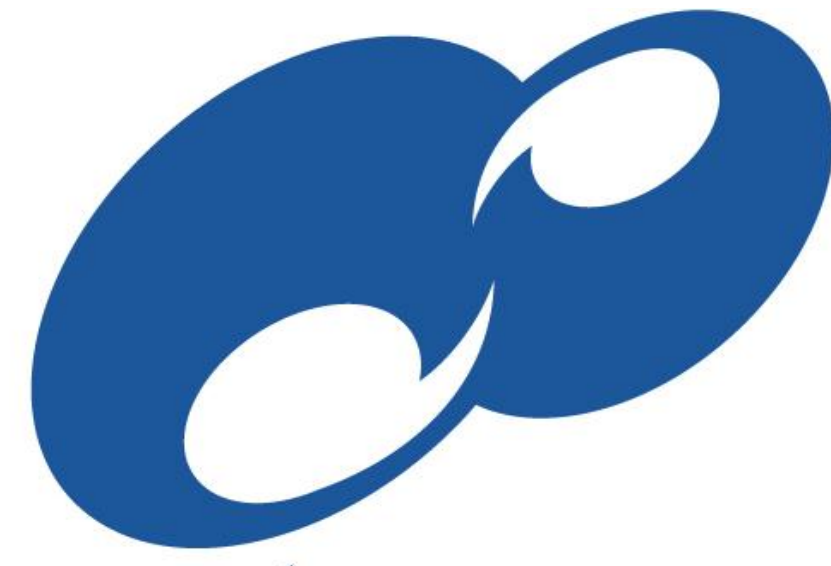
	従来型	VSFF		
Duplex Connector (ϕ 1.25mm ferrule) 高密度実装 対LC	 LCコネクタ	 CSコネクタ (LCの2倍)	 SNコネクタ (LCの3倍)	 MDCコネクタ (LCの3倍)
Multi Fiber Connector (MT type ferrule) 心数 高密度 対MPO	 12, 16, 24, 32心 (最大72心まで可能) MPOコネクタ	 12, 16, 24, 32心 SN-MTコネクタ (MPOの3倍)	 12, 16, 24, 32心 MMCコネクタ (MPOの3倍)	

今後AIインフラをどのように作っていくか

- 今後はより多心化 & 高密度化が進む
 - GPU/CPU/NICの高速化(400G/800G/1.6T) 、光ファイバー技術の進化
 - 水冷ラックへのGPUサーバ搭載数の増加による高集約化
- トポロジーが複雑化
 - コンピュート、ストレージ、インターネット、管理用と複数のネットワークで構成
 - コンピュートのネットワークにマルチプレーン/シャフルアーキテクチャが登場
- 配線設計の重要性が増す
 - 物理配線そのものが生成AIのシステム性能と拡張性を左右する重要な設計要素のひとつ
 - 光ファイバ、光コネクタ等の構成部材に加え複雑なトポロジーを理解の上、最適な配線設計を行う必要がある
- 施工現場との連携
 - 複雑は配線設計を確実に現場で施工するため、設計側と施工側の連携が必須
 - VSFFやシャフル配線等の最新技術導入に伴い、測定技術やトラブルシューティングへの対応も必要
- AIインフラに対応できる技術者について
 - 発展著しいAIインフラに対応できる次世代技術者育成が必要
 - 「ソブリン」は生成AIモデルだけでよいのか？インフラ構築技術のソブリン性は？



SAKURA internet



eXtreak