

構内配線を利用したNVIDIA 800G InfiniBandおよび イーサネット接続ソリューション AEN 185(改訂2版)

目次

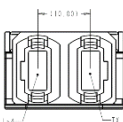
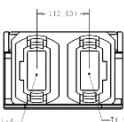
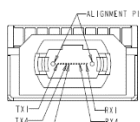
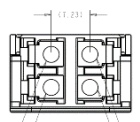
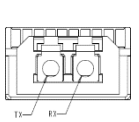
1.	トランシーバーの種類、ポートブレイクアウト、配線シナリオについて理解する.....	2
1.1.	シナリオ 1 - 800G および 400G - サーバーからスイッチへのアプリケーション	3
1.2.	シナリオ 2 - 800Gおよび400G - スイッチ間アプリケーション	4
1.3.	シナリオ 3 - 800G および 200G - サーバーからスイッチへのアプリケーション	5
1.4.	シナリオ 4 - 800Gおよび200G - スイッチ間アプリケーション	6
1.5.	シナリオ 5 - 800G および 400G - スイッチ間アプリケーション	7
1.6.	シナリオ 6 - 800Gおよび400G - スイッチ間アプリケーション	8
1.7.	各シナリオにおけるトランシーバーのオプションとポートブレイクアウト接続.....	9
2.	NVIDIA DGX SuperPOD配線アーキテクチャ参考ガイド	10
2.1.	DGX-H100ノード(サーバー)のコンピュータファブリック接続について理解する	11
2.2.	Quantum-2 9700 NDR IB スイッチについて理解する.....	12
2.3.	SuperPODのビルディングブロックであるスケーラブルユニット(SU)の概念について理解する.....	12
2.4.	NVIDIA DGX SuperPOD 内の配線シナリオの実装	13
2.4.1.	レベルA - サーバーからリーフへの配線	13
2.4.2.	レベルB - リーフからスパインへの配線	15
2.4.3.	レベルC - スパインからコアへの配線	17
2.5.	マルチモードvs シングルモード	19
2.6.	全体像	19
2.6.1.	4台のSUクラスタ/PODの配線	19
2.6.2.	8台のSUクラスタ/PODの配線	20
2.6.3.	16台のSUクラスタ/PODの配線	21
2.6.4.	32台のSUクラスタ/PODの配線	22
2.6.5.	64台のSUクラスタ/PODの配線	23
2.7.	結論	23
	付属書1 - 高密度ハウジング	24
	付属書2 - 極性図	25
	シナリオ 1 - 800G および 400G - サーバーからスイッチへのアプリケーション MPO-8/12 APC間のP2P配線.....	26
	シナリオ 2 - 800Gおよび400G - スイッチ間アプリケーション	27
	シナリオ 3 - 800G および 200G - サーバーからスイッチへのアプリケーション MPO-8/12 APC間のP2P配線	28
	シナリオ 4 - 800Gおよび200G - スイッチ間アプリケーション	30
	シナリオ 5 - 800G および 400G - スイッチ間アプリケーション LC-Duplex間のP2P配線	31
	シナリオ 6 - 800Gおよび400G - スイッチ間アプリケーション	31
	付属書3-参考文献と連絡先	32

1. トランシーバーの種類、ポートブレイクアウト、配線シナリオについて理解する

この「アプリケーション・エンジニアリング・ノート」では、200G、400G、800Gトランシーバーで動作するさまざまな光ファイバ接続および、400G NDR InfiniBand Quantum-2および400GbE Spectrum-4 イーサネット(400G IB/EN)スイッチ機能を用いた同一ラック/同列内/データセンター全体のブレイクアウトオプションについて説明します。

800Gテクノロジーの導入に伴い、NVIDIA 光トランシーバーモジュールは、8ファイバ2本を用いた新型Twin (デュアル)MPO-8/12 トランシーバーインターフェースを利用しています。コーニングのEDGE8®ソリューションは、トランシーバーで2、4、8、16本のファイバを使用することで、シングルモード、マルチモード両方の光インターフェースをサポートするように設計されています。

以下は、コネクタ種別のNVIDIAトランシーバーリスト(抜粋)です:

Twin MPO-8/12 APC インターフェース(OSFP)	MPO-8/12 APC インターフェース (QSFP112)	Twin LC Duplex UPC インターフェース(OSFP)	LCデュプレックス UPCインターフェース (QSFP-DD)	
シングルモード ^{a)} 800G-2xDR4 OSFP ^{c)} 2x8ファイバトランシーバー MMS4X00-NM MMS4X00-NS MMS4X00-NS-FLT ^{d)} MMS4X00-NS-T	シングルモード ^{a)} 400G-DR4 OSFP 8ファイバトランシーバー 2ポートのうち1ポートを使用 MMS4X00-NS400 MS4X00 NS400-T 200G-DR4 OSFP ^{e)} 4/8ファイバトランシーバー MMS1X00-NS400 MS1X00 NS400-T	シングルモード 800G-2xFR4 OSFP 2x2ファイバトランシーバー MMS4X50-NM MMS4X50-NM-T	シングルモード 400G-FR4 QSP-DD 2ファイバトランシーバー MMS1V50-WM MMS1V50-WM-T	
マルチモード ^{b)} 800G-2xSR4 OSFP ^{c)} 2x8ファイバトランシーバー MMA4Z00-NS MMA4Z00-NS-FLT ^{d)} MMA4Z00-NS-T	マルチモード ^{b)} 400G-SR4 OSFP 8ファイバトランシーバー 2ポートのうち1ポートを使用 MMA4Z00-NS400 MMA4Z00 NS400-T 200G-SR4 OSFP ^{e)} 4/8ファイバトランシーバー 2ポートのうち1ポートを使用 MMA4Z00-NS400 MMA4Z00 NS400-T	マルチモード ^{b)} 400G-SR4 QSFP112 8ファイバトランシーバー MMA1Z00-NS400 MMA1Z00 NS400-T 200G-SR4 QSFP112 ^{e)} 4/8ファイバトランシーバー MMA1Z00-NS400 MMA1Z00-NS400-T		
				

a) 「MPO-8/12 APCシングルモード」の光学部品は、黄色のブルタブおよび黄色の光ファイバが目印となります。樹脂カバーが緑色のMPO-8/12 APCコネクタは「斜め研磨コネクタ」であり、より低速ラインレートのトランシーバーで使用される「超フラット研磨コネクタ(UPC)」とは互換性がありません。

b) マルチモード光学部品は、タン(tan)色のブルタブと水色の光ファイバが目印となります。樹脂カバーが緑色のMPO-8/12 APCコネクタは「斜め研磨コネクタ」であり、水色カバーのHDR向け「超フラット研磨コネクタ(UPC)」とは互換性がありません。

c) 文書によっては800Gトランシーバーが「800G-2xDR4」ではなく「800G-DR8」、「800G-2xSR4」ではなく「800G-SR8」と表記されている場合があります。ただし、接続フットプリントは、「2x 8-ファイバMPO-8/12 APC」と表記しています。NVIDIA部品番号を主要リファレンスとしてご利用ください。

d) カードI/Oは、4つの800GTwin ポートOSFPに内部配線されています。

e) 400Gトランシーバーバージョンは、スプリッターケーブル(Y字ハーネス)を用いることで200Gにも対応。400Gトランシーバーの4レーンのうち2レーン(8ファイバのうち4ファイバ)をアクティブ化して200Gデバイスを作成できます。この構成は、上の表では「4/8-ファイバ」と表記しています。

f) T"で終わるトランシーバー品番はイーサネットのバージョンを示しています。

*NVIDIAのコンポーネントおよび設計の詳細については付属書3を確認するとともにNVIDIA Overview Whitepapersを参照ください。

表 1. トランシーバーのリスト。

1.1. シナリオ 1 - 800G および 400G - サーバーからスイッチへのアプリケーション

MPO-8/12 APC間のポイント・ツー・ポイント(P2P)配線

アプリケーション: Quantum-2 InfiniBandまたはSpectrum-4 イーサネットとa) Quantum-2 InfiniBandまたはSpectrum-4イーサネット、b) ConnectX-7およびBluefield-3、c) DGX H100/Cedar-7のリンク

シナリオ1は主に、ポイント・ツー・ポイント(P2P)配線アプリケーションに用いられ、これにはスケラブルユニット(SU)内部のサーバーや、同一ラック/同列内にあるリーフスイッチの接続などが挙げられます。また、この配線タイプは、リーフスイッチからスパインスイッチ、スパインスイッチからコアスイッチなど、異なるスイッチ間の接続に使用できる場合もあります。ただし、これらのスイッチがデータセンター内の異なるエリアに物理配置されている場合、P2P配線は推奨されません。

シナリオ1、2のユースケースA～Cに適用されるトランシーバーの一覧については、セクション1.7の表8を参照してください。SUの詳細およびDGX SuperPODの構築方法については、本文書セクション2を参照してください。

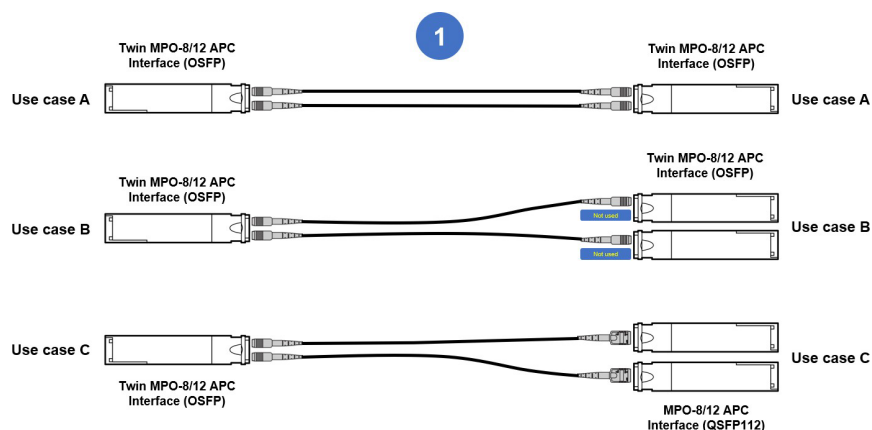


図 1.MPO-8/12 APC で動作する 800Gおよび400G トランシーバーを使用した P2P配線のユースケース

項目	参考品番	OS2品番 (米州)	OS2品番 (EMEA/アジア太平洋日本地域)	OM4品番(米州)	OM4品番 (EMEA/アジア太平洋日本地域)	説明
1	コーニング ^(a)	G-BND32-E8E8G-PN000-xxxF	G-BND32-E8E8G-LZ000-xxxM	G-BND32-E9E9Q-PN000-xxxF	G-BND32-E9E9Q-LZ000-xxxM	32x8ファイバメッシュバンドル、MTP® APC(ピン無し)～MTP APC(非ピン)、レッグ長78インチ(2m)、極性Bタイプ。「xxxF」はxxxフィート、「xxxM」はxxxメートルを示す。
		JE8E808GE8 NBxxxF	JE8E808GEZ NBxxxM	JE9E908QE8 NBxxxF	JE9E908QEZ NBxxxM	EDGE8®, 8ファイバジャンパー、MTP® APC(ピン無し)～MTP APC(ピン無し)、極性Bタイプ。「xxxF」はXXXフィート、「xxxM」はxxxメートルを示す。
	NVIDIA ^(b)	MFP7E30-Nxxx		MFP7E10-Nxxx		NVIDIA MPO-12/APC-to-MPO12/APC (8ファイバ)、ストレートクロスオーバーファイバ、ファイバケーブル製品、タイプB極性。「xxx」はメートル長を示す。

注

- a) コーニングのケーブルは、米州ではプレナムケーブル、EMEA/アジア太平洋日本地域では難燃性LSZH™/CPR定格ケーブルを使用。シングルジャンパー長は1～300m。バンドルジャンパーはメッシュスリーブを使用。バンドル長は10～300メートル(分岐点間長さ)まで、5m単位でご用意しています。バンドルは16、32レッグに加え特注レッグ数(直線/千鳥配列)にも対応。また、米州向けにメートル長でもご用意しています。
- b) NVIDIAケーブルは、OFNR/LSZHデュアル定格ジャケットを採用。SMFケーブル長は3、5、7、10、15、20、30、50、100、150m。MMFケーブル長は3、5、7、10、15、20、25、30、35、40、50m。
- c) コーニングとNVIDIAのケーブルはいずれもInfiniBand、イーサネット、NVLinkプロトコルに対応。
- d) コーニングの極性図(付属書 2)もご確認ください。

表 2. シナリオ1 - MPO-8/12 APCを用いたP2P配線の800Gおよび400G。品番スキーム

1.2. シナリオ2 - 800Gおよび400G - スイッチ間アプリケーション

トランクを用いたDC内MPO-8/12APC間の構内配線

アプリケーション: Quantum-2 InfiniBandまたはSpectrum-4 イーサネットとa) Quantum-2 InfiniBandまたはSpectrum-4イーサネット、b) ConnectX-7およびBluefield-3、c) DGX H100/Cedar-7のリンク

シナリオ2では、構内配線の実装にバックボーンとして光ファイバトランクを使用します。このアプリケーションは多くの場合、リーフスイッチとスパインスイッチのように異なるスイッチ間の接続に用いられ、スパインスイッチとコアスイッチの接続に実装することも可能です。2つの異なるアクティブデバイスがデータセンター内の異なるエリアに物理配置されている場合、このオプションが推奨されます。

シナリオ1、2のユースケースA～Cに適用されるトランシーバーの一覧については、セクション1.7の表8を参照してください。SUの詳細およびDGX SuperPODの構築方法については、本文書セクション2を参照してください。

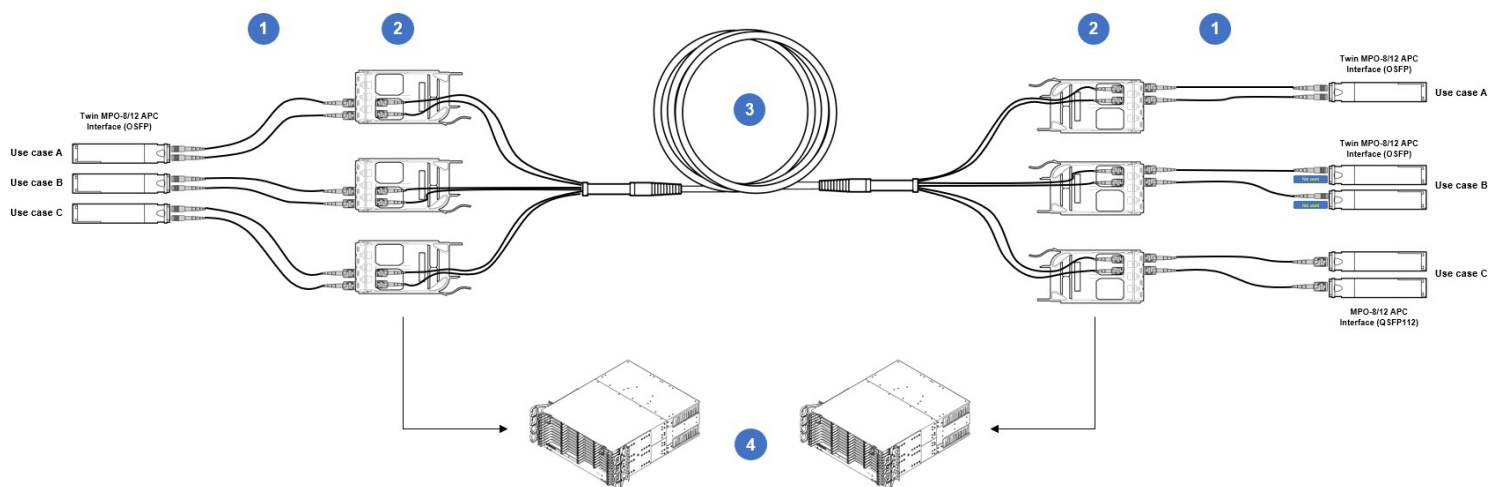


図 2.MPO-8/12 APCで動作する800Gおよび400Gトランシーバーを使用した構内配線のユースケース

項目	参考品番	OS2品番 (米州)	OS2品番 (EMEA/アジア太平洋日本地域)	OM4品番 (米州)	OM4品番 (EMEA/アジア太平洋日本地域)	説明
1	コーニング®	G-BND16-E8E8G- PN000-xxxF	G-BND16-E8E8G- LZ000-xxxM	G-BND16-E9E9Q- PN000-xxxF	G-BND16-E9E9Q- LZ000-xxxM	16x 8ファイバメッシュバンドル、MTP® APC (ピン無し) ~MTP APC (ピン無し)、レッグ長78インチ (2m)、極性Bタイプ。「xxxF」はxxxフィート、「xxxM」はxxxメートルを示す。
		JE8E808GE8 NBxxxF	JE8E808GEZ NBxxxM	JE9E908QE8 NBxxxF	JE9E908QEZ NBxxxM	EDGE8®, MTP® APC (ピン無し) ~MTP APC (ピン無し) 8ファイバジャンパー、TIA-568極性Bタイプ。「xxxF」はxxxフィート、「xxxM」はxxxメートルを示す。
		EDGE8-CP32-V1	EDGE8-CP32-V1	EDGE8-CP32-V3	EDGE8-CP32-V3	EDGE8 32-F MTPアダプターパネル (4ポート)
		GE7E7E4GPNDU xxxF ^(a)	GE7E7E4GLZDDU xxxM ^(b)	GE2E2E4QPNDU xxxF ^(b)	GE2E2E4QLZDDU xxxM ^(b)	EDGE8, MTPトランク、144-F, MTP APC (ピン有り) ~MTP APC (ピン有り)、レッグ長33インチ (840mm)、極性Bタイプ。プーリンググリップ付 (第一端のみ)。「xxxF」はxxxフィート、「xxxM」はxxxメートルを示す。
4		EDGE8-xxU	EDGE8-xxU	EDGE8-xxU	EDGE8-xxU	付属書1をご参照の上、ご用途に最適オプションを選択ください。

注

- a) コーニングのケーブルは、米州ではプレナムケーブル、EMEA/アジア太平洋日本地域では難燃性LSZH™/CPR定格ケーブルを使用。シングルジャンパー長は1~300m。バンドルジャンパーはメッシュスリーブを使用。バンドル長は10~300メートル (分岐点間長さ) まで、5m単位でご用意しています。バンドルは16、32レッグに加え特注レッグ数 (直線/千鳥配列) にも対応。また、米州向けにメートル長でもご用意しています。
- b) トランクのファイバ数は8~288本。
- c) コーニングのケーブルは、InfiniBand、イーサネット、NVLinkプロトコルに対応。
- d) コーニングの極性図 (付属書 2) もご確認ください。

表 3.シナリオ2 - MPO-8/12 APCを用いた構内配線の800Gおよび400G。品番スキーム。

1.3. シナリオ 3 - 800G および 200G - サーバーからスイッチへのアプリケーション

MPO-8/12 APC間のP2P配線

アプリケーション: Quantum-2 InfiniBandまたはSpectrum-4イーサネットとConnectX-7 / Bluefield-3のリンク

シナリオ3は、P2P配線アプリケーションに用いられます。このシナリオでは、400Gトランシーバーの4レーンのうち2レーン(8ファイバのうち4ファイバ)をアクティブにするスプリッターケーブル(Y字ハーネス)を使用することで、400Gトランシーバーバージョンでも200Gに対応することができ、実質的に200Gデバイスを作成可能です。P2P配線が推奨されるのはアクティブデバイスが同一ラックまたは同列内にある場合のみとなり、アクティブデバイスがデータセンター内の異なるエリアに物理配置されている場合は、構内配線が推奨されます。

シナリオ3、4 のユースケース A および B に適用されるトランシーバーの一覧については、セクション 1.7 の表 9 を参照してください。SUの詳細とDGX SuperPODの構築方法については、本文書セクション2を参照してください。

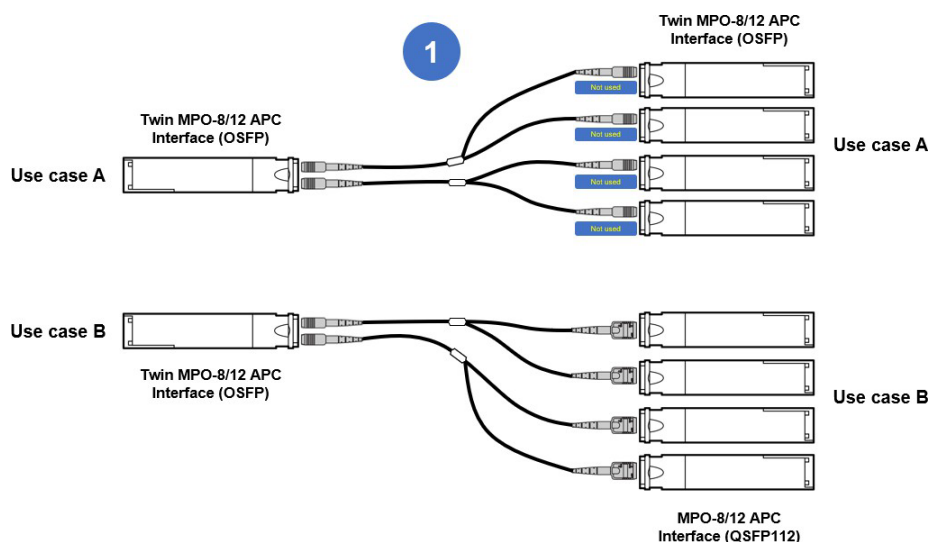


図 3.MPO-8/12 APC で動作する 800Gおよび200G トランシーバーを使用した P2P 配線のユースケース

項目	参考品番	OS2 品番(米州)	OS2品番(EMEA ／アジア太平洋日 本地域)	OM4 品番(米州)	OM4品番 (EMEA／アジア太平洋 日本地域)	説明
1	コーニング ^(a)	HE8E808GPH LBxxxF	HE8E808GLZ LBxxxM	HE9E908QPH LBxxxF	HE9E908QLZ LBxxxM	EDGE8®, 8ファイバ Y字ハーネス、MTP® APC(ピン無し) ~2x4-F(ピン無し) MTP APC、ブレイクアウトレッグ長36イン チ(910mm)、極性Bタイプ。「xxxF」はxxxフィート、「xxxM」は xxxメートルを示す。
	NVIDIA ^(b)	MFP7E40-Nxxx		MFP7E20-Nxxx		NVIDIA MPO-12/APC-to-MPO12/APC(8ファイバ)、4→2チャ ネルヘスプリッタファイバケーブル、極性Bタイプ。「xxx」はxxxメ ートルを示す。

注

- a) コーニングのケーブルは、米州ではプレナムケーブル、EMEA／アジア太平洋日本地域では難燃性LSZH™/CPR定格ケーブルを使用。米州向けにメートル長でもご用意しています。
- b) NVIDIAスプリッターケーブルは、OFNR/LSZHのデュアル定格ジャケットを使用。SMFおよびMMFスプリッターケーブル長は3、5、7、10、15、20、30および50m。
- c) コーニングとNVIDIAの両ケーブルは、InfiniBand、イーサネット、NVLinkプロトコルをサポートしている。
- d) 400Gトランシーバーバージョンは、スプリッターケーブル(Y字ハーネス)を用いることで200Gにも対応。400Gトランシーバーの4レーンのうち2レーン(8ファイバのうち4ファイバ)をアクティブ化して200Gデバイスを作成できます。
- e) コーニングの極性図(付属書 2)もご参照ください。

表 4. シナリオ3 - MPO-8/12 APCを用いたP2P配線の800Gおよび200G。品番スキーム。

1.4. シナリオ4 - 800Gおよび200G - スイッチ間アプリケーション

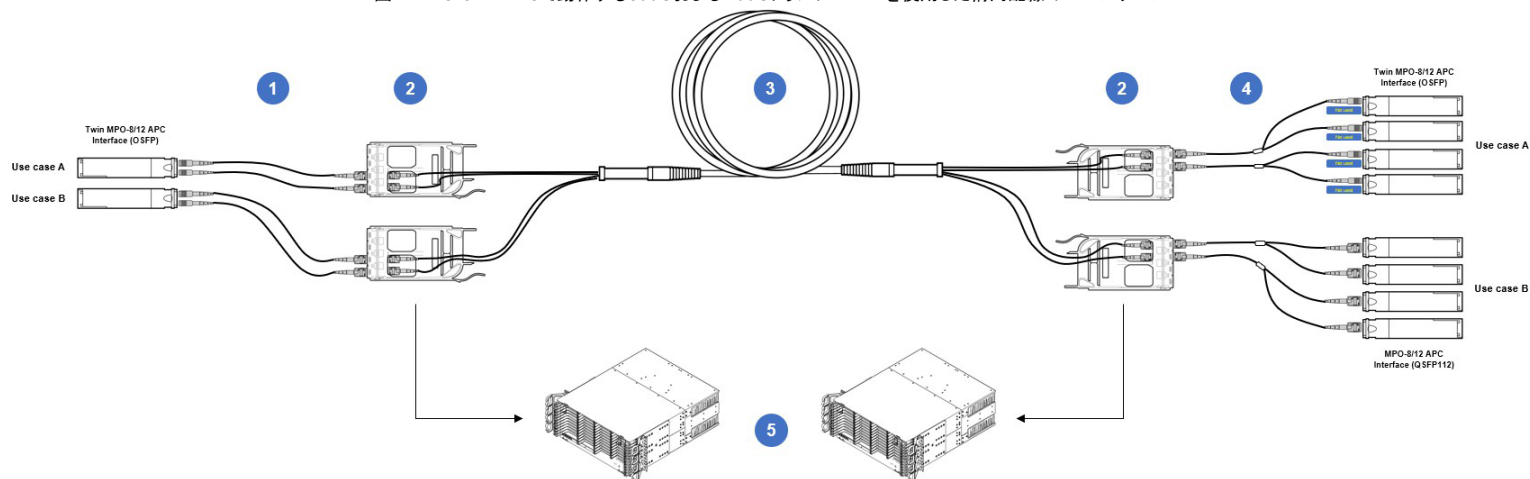
トランクを用いたDC内MPO-8/12APC間の構内配線

アプリケーション: Quantum-2 InfiniBandまたはSpectrum-4 イーサネットとConnectX-7 / Bluefield-3のリンク

シナリオ4は、構内配線コンポーネントで用いられます。このシナリオでは、400Gトランシーバーの4レーンのうち2レーン(8ファイバのうち4ファイバ)をアクティブにするスプリッターケーブル(Y字ハーネス)を使用することで、400Gトランシーバーバージョンでも200Gに対応することができ、実質的に200Gデバイスを作成可能です。このシナリオでは、アクティブデバイスはデータセンター内の異なるエリアに物理配置されます。

シナリオ 3、4 のユースケース A および B に適用されるトランシーバーの一覧については、セクション 1.7 の表 9 を参照してください。SUの詳細とDGX SuperPODの構築方法については、本文書セクション2を参照してください。

図 4.MPO-8/12 APCで動作する800Gおよび200Gトランシーバーを使用した構内配線のユースケース



項目	参考品番	OS2 品番(米州)	OS2品番(EMEA ／アジア太平洋 日本地域)	OM4 品番(米州)	OM4品番 (EMEA／アジア太平洋 日本地域)	説明
1	コーニング ^(a)	G-BND16-E8E8G- PN000-xxxF	G-BND16-E8E8G- LZ000-xxxM	G-BND16-E9E9Q- PN000-xxxF	G-BND16-E9E9Q- LZ000-xxxM	16x 8ファイバメッシュバンドル、MTP® APC (ピン無し) ~ MTP APC (ピン無し)、レッグ長78インチ(2m)、極性Bタイプ。「xxxF」はxxxフィート、「xxxM」はxxxメートルを示す。
		JE8E808GE8 NBxxxF	JE8E808GEZ NBxxxM	JE9E908QE8 NBxxxF	JE9E908QEZ NBxxxM	EDGE8®, MTP® APC (ピン無し) ~ MTP APC (ピン無し) 8ファイバジャンパー、TIA-568極性Bタイプ。「xxxF」はxxxフィート、「xxxM」はxxxメートルを示す。
2		EDGE8-CP32-V1	EDGE8-CP32-V1	EDGE8-CP32-V3	EDGE8-CP32-V3	EDGE8 32-F MTPアダプターパネル(4ポート)
3		GE7E7E4GPNDU xxxF ^(b)	GE7E7E4GLZDU xxxM ^(b)	GE2E2E4QPNDU xxxF ^(b)	GE2E2E4QLZDU xxxM ^(b)	EDGE8、MTPトランク、144-F、MTP APC (ピン有り) ~ MTP APC (ピン有り)、レッグ長33インチ(840mm)脚、極性Bタイプ。プーリンググリップ付き(第一端のみ)。「xxxF」はxxxフィート、「xxxM」はxxxメートルを示す。
4		HE8E808GPH LBxxxF	HE8E808GLZ LBxxxM	HE9E908QPH LBxxxF	HE9E908QLZ LBxxxM	EDGE8、8ファイバY字ハーネス、MTP® APC (ピン無し) ~ 2x 4-F (ピン無し) MTP APC、ブレイクアウトレッグ長36インチ(910mm)、極性Bタイプ。「xxxF」はxxxフィート、「xxxM」はxxxメートルを示す。
5		EDGE8-xxU	EDGE8-xxU	EDGE8-xxU	EDGE8-xxU	付属書1をご参照の上、用途に最適なオプションを選択ください。

注

a) コーニングのケーブルは、米州ではプレナムケーブル、EMEA／アジア太平洋日本地域では難燃性LSZH™/CPR定格ケーブルを使用。Y字ハーネス長は1~60m、シングルジャンパー長は1~300m。バンドルジャンパーは、メッシュスリーブを使用。バンドル長は10~300メートル(分岐点間長さ)まで、5m単位でご用意しています。バンドルは16、32レッグに加え特注レッグ数(直線／千鳥配列)にも対応。また、米州向けにメートル長でもご用意しています。

b) トランクのファイバ数は8~288本。

c) コーニングのケーブルは、InfiniBand、イーサネット、NVLinkプロトコルに対応。

d) コーニングの極性図(付属書 2)をご参照ください。

表 5. シナリオ4 - MPO-8/12 APCを用いた構内配線の800Gおよび200G。品番スキーム。

1.5. シナリオ 5 - 800G および 400G - スイッチ間アプリケーション

LC-Duplex間のP2P配線

アプリケーション: Quantum-2 InfiniBandまたはSpectrum-4 イーサネットと Quantum-2 InfiniBandまたはSpectrum-4 イーサネットのリンク

シナリオ5は主に、同一ラックまたは同列内にあるアクティブデバイスの接続などのP2P配線に用いられます。ただし、これらのデバイスがデータセンター内の異なるエリアに物理配置されている場合、P2P配線の使用は推奨されません。

Twin LC-Duplexトランシーバーは主に、Quantum-2またはSpectrum-4空冷スイッチの接続に用いられます。トランシーバー本体は黄緑色 (bright green) で、これは最長距離が2kmであることを示しています。

シナリオ 5、6 のユースケース A、B に適用されるトランシーバーの一覧については、セクション 1.7 の表 10 を参照してください。SUの詳細とDGX SuperPODの構築方法については、本文書セクション2を参照してください。

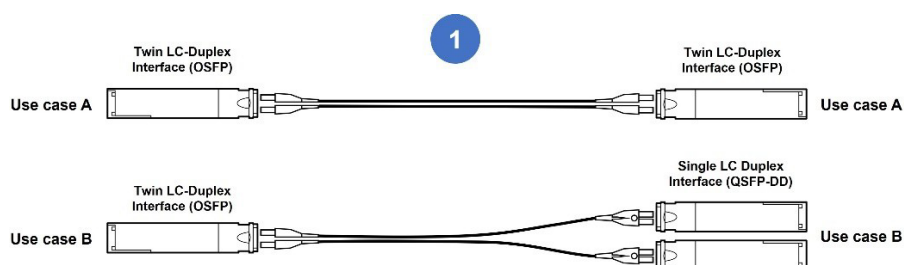


図 5. LC Duplex で動作する 800G および 400G トランシーバーを使用した P2P 配線のユースケース

項目	参考品番	OS2 品番 (米州)	OS2品番 (EMEA/アジア太平洋日本地域)	説明
1	コーニング ^(a)	787802GD120xxxF	E787802GNZ20xxxM	EDGE™ LC UPC Uniboot~LC UPC Uniboot Duplex ジャンパー。「xxxF」はxxxフィート、「xxxM」はxxxメートルを示す。

注

- a) コーニングのケーブルは、米州ではプレナムケーブル、EMEA/アジア太平洋日本地域では難燃性LSZH™/CPR定格ケーブルを使用。ジャンパー長は1~300m。米州向けにメートル長でもご用意しています。
- b) コーニングのケーブルは、InfiniBand、イーサネット、NVLinkプロトコルに対応。
- c) コーニングの極性図 (付属書 2) をご参照ください。

表 6. シナリオ5 - LC Duplex を用いたP2P配線の800Gおよび400G。品番スキーム。

1.6. シナリオ6 - 800Gおよび400G - スイッチ間アプリケーション

トランクを用いたDC内LC-Duplex間の構内接続

アプリケーション: Quantum-2 InfiniBandまたはSpectrum-4 イーサネットとQuantum-2 InfiniBandまたはSpectrum-4 イーサネットのリンク

シナリオ6では、構内配線の実装にバックボーンとしてMPO-8/12接続の光ファイバトランクを使用します。このアプリケーションは、リーフスイッチとスパインスイッチのように異なるスイッチ間の接続に用いられます。2つの異なるアクティブデバイスがデータセンター内の異なるエリアに物理配置されている場合、このオプションが推奨されます。

Twin LC-Duplexトランシーバーは主に、Quantum-2またはSpectrum-4空冷スイッチの接続に用いられます。トランシーバー本体は黄緑色 (bright green) で、これは最長距離が2kmであることを示しています。

シナリオ 5、6 のユースケース A、B に適用されるトランシーバーの一覧については、セクション 1.7 の表 10 を参照してください。SUの詳細とDGX SuperPODの構築方法については、本文書セクション2を参照してください。

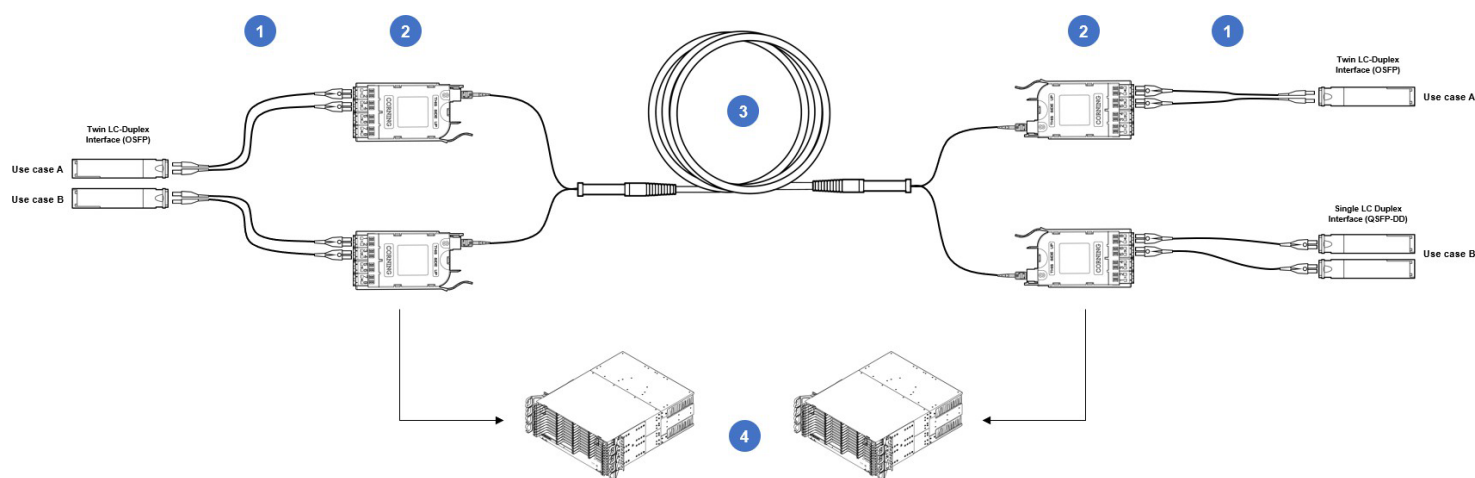


図 6.LC-DUPLEXで動作する800Gおよび400Gトランシーバーを使用した構内配線のユースケース

項目	参考品番	OS2 品番(米州)	OS2品番 (EMEA/アジア太平洋日本地域)	説明
1	コーニング ^(a)	787802GD120xxxF	E787802GNZ20xxxM	EDGE™ LC UPC Uniboot~LC UPC Uniboot Duplex ジャンパー。「xxxF」はxxxフィート、「xxxM」はxxxメートルを示す。
2		ECM8-UM08-04-E8G-ULL	ECM8-UM08-04-E8G-ULL	EDGE8® 超低損失モジュール、LC Duplex~MTP® (ピン無し)、8ファイバ、ユニバーサル極性。
3		GE7E7E4GPNDU xxxF ^(b)	GE7E7E4GLZDDU xxxM ^(b)	EDGE8、MTPトランク、144-F、MTP APC (ピン有り) ~MTP APC (ピン有り)、レッグ長33インチ (840mm) 脚、極性Bタイプ。ブーリンググリップ付き (第一端のみ)。「xxxF」はxxxフィート、「xxxM」はxxxメートルを示す。
4		EDGE8-xxU	EDGE8-xxU	付属書1をご参照の上、用途に最適なオプションを選択ください。

注

- a) コーニングのケーブルは、米州ではプレミアムケーブル、EMEA/アジア太平洋日本地域では難燃性LSZH™/CPR定格ケーブルを使用。ジャンパー長は1~300m。米州向けにメートル長でもご用意しています。
- b) トランクのファイバ数は8~288本。
- c) コーニングのケーブルは、InfiniBand、イーサネット、NVLinkプロトコルに対応。
- d) コーニングの極性図(付属書 2)をご参照ください。

表 7. シナリオ6 - アクティブデバイスでLC-DUPLEXを用いた構内配線の800Gおよび400G。品番スキーム。

1.7. 各シナリオにおけるトランシーバーのオプションとポートブレイクアウト接続

例を参考にして下記表をご利用ください。

例 1:「シナリオ 2、ユースケースB」の場合:表 8 に従い、2 x MPO-8/12 APC接続を備えた800G-2xDR4 MMS4X00-NS(図面左側)を、2個のOSFP トランシーバー(MMS4X00-NS400)に接続することができ、これによって個別に2つの400G 接続が可能となります(図面右側)。

例 2:「シナリオ 4、ユースケースB」の場合:表 9 に従い、2 x MPO-8/12 APC接続を備えた800G-2xDR4の MMS4X00-NS(図面左側)を、Y字ハーネスを使用して 4 個の 200G-DR4 QSFP112, MMS1X00-NS400 に接続することができ、これによって個別に4つの200G 接続が可能となります(図面右側)。

ユースケース	光ファイバ近端部(左)			光ファイバ遠端部(右)			距離	ファイバ種別
	速度	NVIDIAポート	フットプリント - ファイバトランシーバー	速度	NVIDIAポート	フットプリント - ファイバトランシーバー		
A	800G-2xDR4	MMS4X00-NM	OSFP - 2x 8F	800G-2xDR4	MMS4X00-NM	OSFP - 2x 8F	500m	シングルモード
	800G-2xDR4	MMS4X00-NM-T	OSFP - 2x 8F	800G-2xDR4	MMS4X00-NM-T	OSFP - 2x 8F	500m	シングルモード
	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	100m	シングルモード
	800G-2xDR4	MMS4X00-NS-FLT	OSFP - 2x 8F	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	100m	シングルモード
	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	100m	シングルモード
B	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	400G-DR4	MMS4X00-NS400	OSFP - 8F	100m	シングルモード
	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	400G-DR4	MS4X00 NS400-T	OSFP - 8F	100m	シングルモード
C	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	400G-DR4	MMS1X00-NS400	QSFP112 - 8F	100m	シングルモード
	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	400G-DR4	MS1X00 NS400-T	QSFP112 - 8F	100m	シングルモード
A	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	50m	マルチモード
	800G-2xSR4	MMA4Z00-NS-FLT	OSFP - 2x 8F	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	50m	マルチモード
	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	50m	マルチモード
B	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	400G-SR4	MMA4Z00-NS400	OSFP - 8F	50m	マルチモード
	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	400G-SR4	MMA4Z00 NS400-T	OSFP - 8F	50m	マルチモード
C	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	400G-SR4	MMA1Z00-NS400	QSFP112 - 8F	50m	マルチモード
	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	400G-SR4	MMA1Z00 NS400-T	QSFP112 - 8F	50m	マルチモード

表 8. シナリオ1、2のユースケースA～Cに適用可能なトランシーバー

ユースケース	光ファイバ近端部(左)			光ファイバ遠端部(右)			距離	ファイバ種別
	速度	NVIDIAポート	フットプリント - ファイバトランシーバー	スピード	NVIDIAポート	フットプリント - ファイバトランシーバー		
A	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	200G-DR4	MMS4X00-NS400	OSFP - 4F	100m	シングルモード
	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	200G-DR4	MS4X00 NS400-T	OSFP - 4F	100m	シングルモード
B	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	200G-DR4	MMS1X00-NS400	QSFP112 - 4F	100m	シングルモード
	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	200G-DR4	MS1X00 NS400-T	QSFP112 - 4F	100m	シングルモード
A	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	200G-SR4	MMA4Z00-NS400	OSFP - 4F	50m	マルチモード
	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	200G-SR4	MMA4Z00 NS400-T	OSFP - 4F	50m	マルチモード
B	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	200G-SR4	MMA1Z00-NS400	QSFP112 - 4F	50m	マルチモード
	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	200G-SR4	MMA1Z00 NS400-T	QSFP112 - 4F	50m	マルチモード

表9. シナリオ3、4のユースケースAおよびBに適用可能なトランシーバー

ユースケース	光ファイバ近端部(左)			光ファイバ遠端部(右)			距離	ファイバ種別
	速度	NVIDIAポート	フットプリント - ファイバトランシーバー	速度	NVIDIAポート	フットプリント - ファイバトランシーバー		
A	800G-2xFR4	MMS4X50-NM	OSFP - 2x 2F	800G-2xFR4	MMS4X50-NM	OSFP - 2x 2F	2km	シングルモード
	800G-2xFR4	MMS4X50-NM-T	OSFP - 2x 2F	800G-2xFR4	MMS4X50-NM-T	OSFP - 2x 2F	2km	シングルモード
B	800G-2xFR4	MMS4X50-NM	OSFP - 2x 2F	400G-FR4	MMS1V50-WM	QSFP-DD - 2F	2km	シングルモード
	800G-2xFR4	MMS4X50-NM-T	OSFP - 2x 2F	400G-FR4	MMS1V50-WM-T	QSFP-DD - 2F	2km	シングルモード

表 10. シナリオ5、6のユースケースA、Bに適用可能なトランシーバー

2. NVIDIA DGX SuperPOD配線アーキテクチャ参考ガイド

本セクションでは、NVIDIA DGX SuperPOD の配線方法について説明します。本ガイドではNVIDIA DGX H100 をリファレンスとして使用していますが、ここで重要なのは、DGX B200 および DGX GB200 を含むアプリケーションにも同様の配線コンポーネントおよびインフラを使用できるという点です。各 POD の検討事項は、お客様の設計によって異なります。

DGX SuperPODは、一定数のスケーラブルユニット(SU)からなり、各SUには一定数のノード(サーバー)が含まれます。この構成に応じてPODあたりのGPU総数が決まります。SUがSuperPODのビルディングブロックとなっています。

なお、お客様によって、PODは「クラスタ」とも呼ばれています。このクラスタ/PODのサイズに応じてInfiniBand リーフスイッチ、InfiniBandスパインスイッチ、InfiniBandコアスイッチの必要数が決まる、という点を理解しておくことが重要です。こうした構成に基づいて、必要なケーブル数、接続数を計算することができます。

例えば、SU数32の中型クラスタ/PODを例にとると(図7)、ノード(サーバー)総数は1,024となり、GPU数8,192に相当します。この構成の場合、PODにIBリーフスイッチ256台、IBスパインスイッチ256台、IBコアスイッチ128台が必要になります。本文書では、このPODサイズをリファレンスとして使用します。なお、「小型」、「中型」、「大型」という分類は NVIDIA によって定義されたものではありませんが、本ガイドではリファレンスとして使用します。

同じ図の「ケーブル数」欄は、システム接続に必要な各MPO-8/12 APC接続数を示しています。例えば、各ノード(サーバー)をIBリーフスイッチに接続するのに必要なMPO-8/12 APC接続数は8,192です。IBリーフスイッチとIBスパインスイッチを接続するには、さらに8,192のMPO-8/12 APC接続が必要です。最後に、IBスパインスイッチをIBコアスイッチに接続するには、さらに8,192のMPO-8/12 APC接続が必要となります。これにより、必要なMPO-8/12 APC接続数はPOD全体で24,576となります。

	SU Count	Node Count	GPU Count	Switch Counts			Cable Counts		
				InfiniBand Leaf	InfiniBand Spine	InfiniBand Core	Node-Leaf	Leaf-Spine	Spine-Core
Small Cluster or POD	4	128	1024	32	16	--	1024	1024	1024
	8	256	2048	64	32	--	2048	2048	2048
Medium Cluster or POD	16	512	4096	128	128	64	4096	4096	4096
	32	1024	8192	256	256	128	8192	8192	8192
Large Cluster or POD	64	2048	16384	512	512	256	16384	16384	16384

図7. SuperPODのコンポーネント数

2.1. DGX-H100ノード(サーバー)のコンピュータブリック接続について理解する

NVIDIA DGX-H100ノードは、InfiniBand コンピュータブリック、InfiniBand ストレージブリック、インバンド管理ネットワークおよびアウトオブ管理ネットワーク用の接続ポートを備えています¹。

このうち、本アーキテクチャガイドではInfiniBandコンピュータブリックを中心に説明します。同ブリックでは、各ノードに デュアルポートOSFP x 4つ、合計で8 x 400Gの接続が可能です。これらの接続には、図 8 に示すように、シングルモードまたはマルチモード MPO-8/12 APC コネクタを使用します。

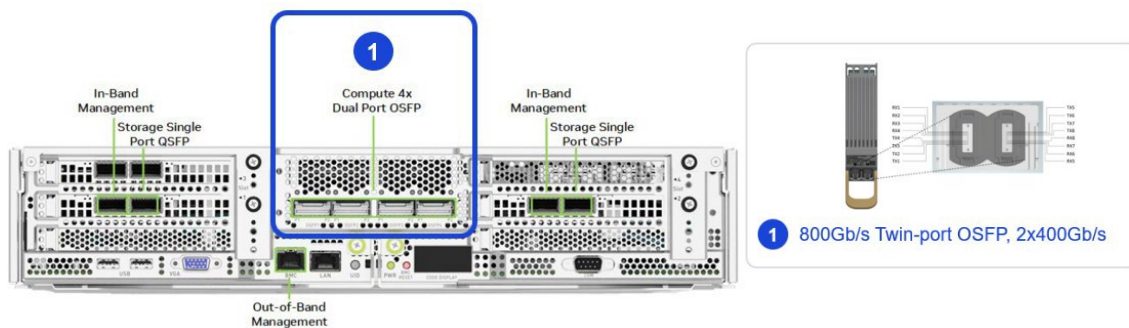


図8.DGX H100 CPUトレイの背面部と接続

¹ NVIDIA Reference Architecture Scalable Infrastructure: <https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastructure-h100/latest/network-fabrics.html>

2.2. Quantum-2 9700 NDR IB スイッチについて理解する

Quantum-2スイッチにはOSFP Twin MPO-8/12 APCポートが32個あります。つまり、各スイッチで400Gbpsのシングルポートを最大64ポート利用可能です。

図9に、スイッチレベルでのポートマッピング例を示します。本文書を通じて、この例をリファレンスとして利用いただけますが、あくまでお客様によるポートマッピングの一例であることにご留意ください。

Quantum-2スイッチは、リーフ、スパイン、コアなど様々な役割でネットワーク全体に配置されるため、ポート分配は図9に示したスイッチ間接続アプローチに沿って行うことが可能です。

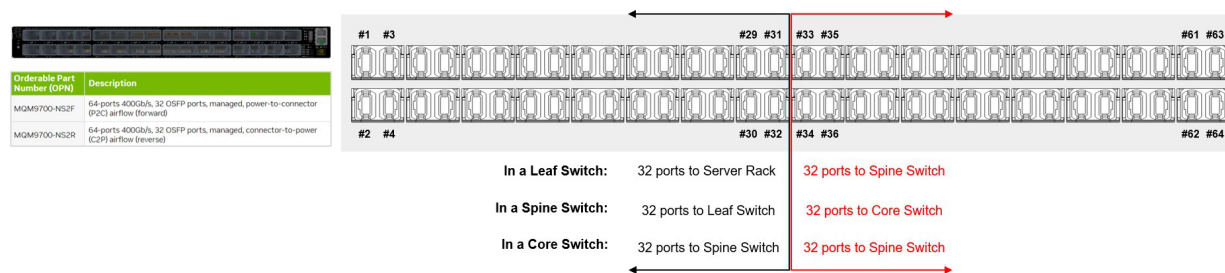


図 9. Quantum-2 スイッチのポートマッピング例

Quantum-2 スイッチの下段ポートにある OSFP Twin トランシーバーは、図 10 に示すように逆さまになり、ポートの接続が逆になることに注意が必要です。この例では、Twin ポート1が「X」、「Y」接続となるに対して、下段のTwin ポート2では逆になっているため、「Y」、「X」の接続となることがわかります。これを物理的に見分けるには、両方のトランシーバーのフィンを確認してください(互いに反対になっています)。

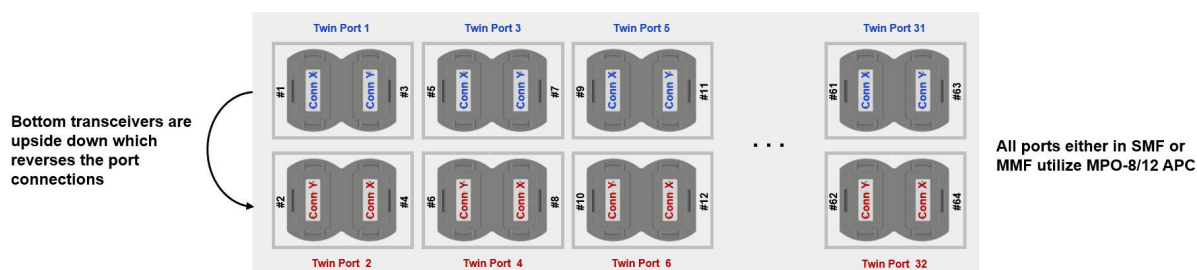


図10. Twin トランシーバーポートに関する考慮事項

2.3. SuperPODのビルディングブロックであるスケーラブルユニット(SU)の概念について理解する

システムは、それぞれ32個のDGX H100システムを搭載するスケーラブルユニット(SU)というビルディングブロックで構成されており、これによって大小さまざまな規模のシステムを迅速に導入することが可能になっています。

各SUは256個のGPUを備えており、8つのラックにある32個のDGX H100システムに分散配置されています。また、各SUには8つのリーフスイッチがあります。

ファブリックはレーン最適化されており、各ノードから、同じポート番号のホストチャネルアダプタ(HCA)が同じリーフスイッチに接続されます。すなわち、すべてのサーバーからのポート番号1の接続はすべてリーフスイッチ番号1、ポート番号2の接続はリーフスイッチ番号2...と接続されます。

ファブリックは、800Gbps/Twin ポートOSFPトランシーバーを備えたQuantum-2 9700 InfiniBandスイッチを用いて構築されています。

図11は、上述の情報に基づき、SU内に関連コンピュータファブリックコンポーネントを配置したMoR(ミドルオブロー)レイアウトを示したものです。もちろん、これ以外のレイアウトも実装可能です。したがって、設計の選択に最適な配線オプションを検証する上で、設計段階でコーニングのエンジニアリングチームと緊密にコミュニケーションを図ることが非常に重要になります。

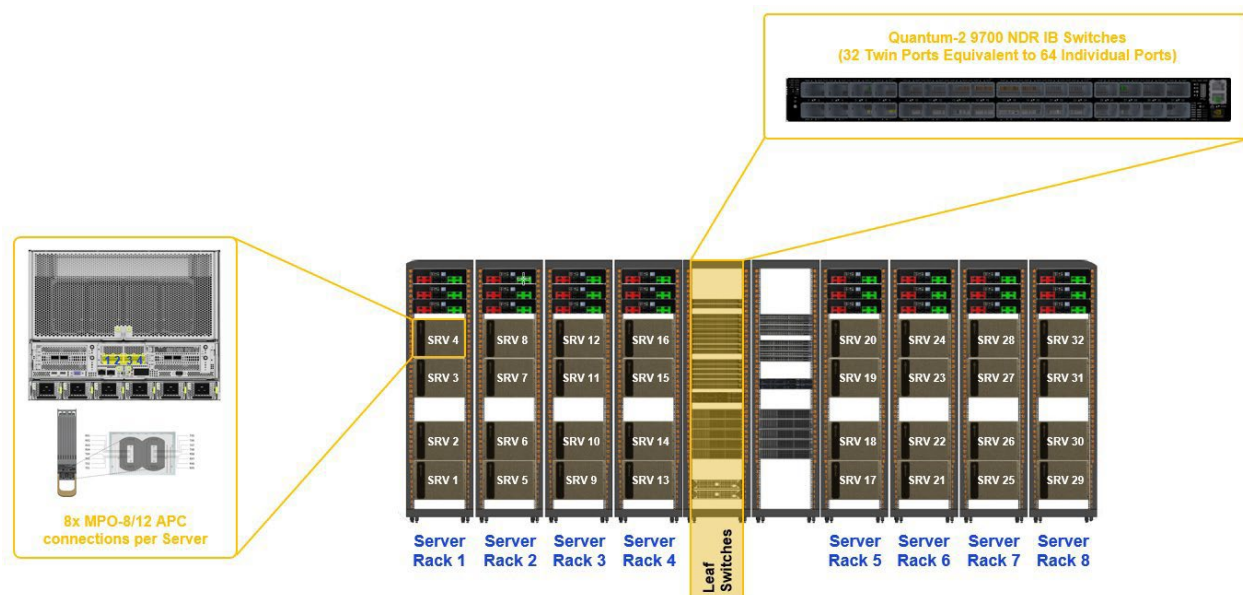


図11. SUのコンピュータファブリック関連コンポーネント

2.4. NVIDIA DGX SuperPOD 内の配線シナリオの実装

AI/MLクラスタの構築に使用される配線コンポーネントを識別しやすくするために、コーニングは本ガイドで3つの接続レベルを使用します。各レベルとスイッチ数は、32個のSUクラスタ/POD例をベースとしています。

レベルA - サーバーからリーフへの配線

レベルB - リーフからスパインへの配線

レベルC - スパインからコアへの配線

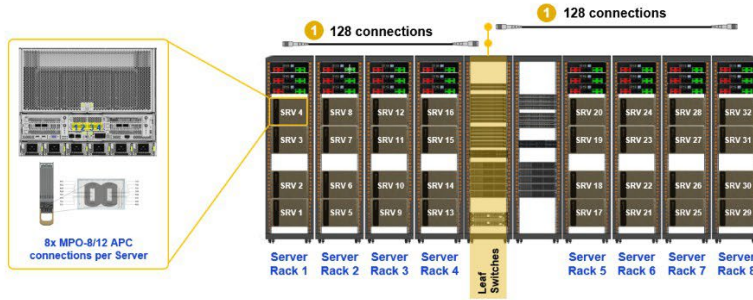
2.4.1. レベルA - サーバーからリーフへの配線

SUは、ノード(サーバー)とリーフスイッチをP2P接続で配線できます(図1参照)。また、2つ以上の配線製品オプションも利用可能です(図12参照)。一部のカスタム設計では、SUレベルで構内配線を実装することもできます(図2参照)。

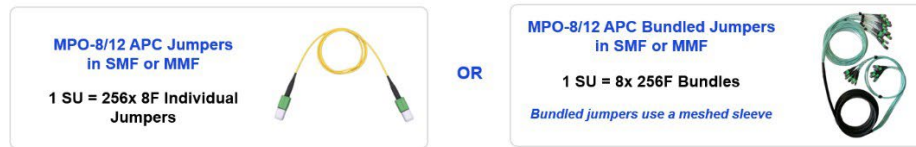
この配線における1つ目のアプローチは、各ツイントランシーバーからの接続を構築するために256本の個別8ファイバジャンパーを用いる方法です。2つ目のアプローチとして、8本のバンドルジャンパーを使用する方法もあります。バンドルジャンパーは、8ファイバMPO-8/12 APCパッチコードを難燃性ラップアラウンドメッシュで256ファイバユニットとしてグループ化したものです。どちらを選択するかはお客様の具体的設計要件によって異なりますが、いずれの場合も、MPO-8/12 APCとリーフスイッチの接続数は合計で256になり、この例ではMoR レイアウトになっています(図12参照)。

コンピュータファブリックはレーン最適化されており合計8つのレーンがあることはすでに説明しましたが、これらのレーンは、各サーバーからの8つのMPO-8/12 APC接続に対応しています。以下の例では、図12と図13に示すように、カラーリファレンスを使用して各レーン接続を示します。

Each SU requires a total of 256 MPO-8/12 APC 8-fiber connections (8 per Server) to the Leaf Switches



One Scalable Unit, Multiple Approaches: Partnering with Customers for Future Success



Rail Architecture

The fabric is Rail-optimized, meaning that Host Channel Adapters (HCAs) with the same port number from each node are connected to the same Leaf Switch.

This means all port number 1 connections from all servers will be connected to Leaf Switch number 1, port number 2 to Leaf Switch number 2, and so on.

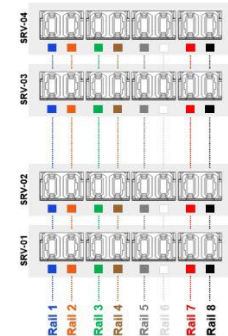


図12. レベルA - サーバーからリーフスイッチへの配線 接続数と2つの製品アプローチについて説明

まずレベルAについて、サーバーラックからSUの中央列 (MoR) のInfiniBandリーフスイッチへのバンドルジャンパーによる接続を説明します。個別のジャンパーを選択する場合は、レールマッピングに従ってください。

図13の左側には、4台のサーバーを収めたサーバーラックがあり、各ラックのレールが示されています。右側には、各レールに対応する8台のリーフスイッチを収めたリーフスイッチサーバーがあります。ご存知のように、Quantum-2スイッチは32個のTwin MPO-8/12 APCポートを備えており、各スイッチには最大64の個別接続が可能です。

Each SU requires a total of 256 MPO-8/12 APC 8-fiber connections (8 per Server) to the Leaf Switches

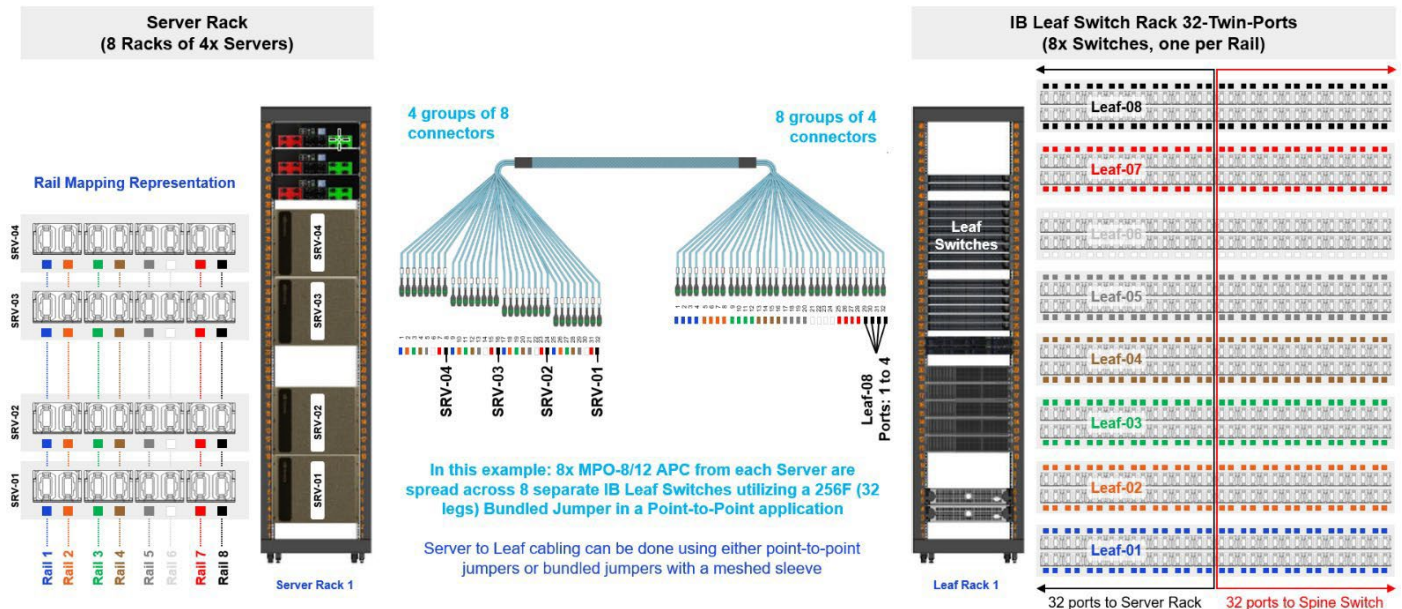


図 13. レベルA - バンドルジャンパーを用いたP2P接続によるサーバーからリーフスイッチへの配線 - 32台のSUクラスPODの場合

各サーバーからの同色レールをすべて対応色のリーフスイッチにルーティングするために、サーバーラックごとに1本のバンドルジャンパーを用いて配線を行います。この例では、各サーバーからのレール8 (黒レール) がリーフ08スイッチのポート1~4に分配されます。SUには合計8台のサーバーラックを備えており、レール8は、図9で示したマッピングに従い、最初の32ポートをリーフ08に接続した後、残りの32ポート接続はスパインスイッチにルーティングします。

SU内のすべての接続が完了するまで、各サーバーラックとレールにこのプロセスを繰り返します。

2.4.2. レベルB - リーフからスパインへの配線

リーフスイッチとスパインスイッチの接続には構内配線を推奨していますが、設計によっては、P2P配線で接続することも可能です。なお、これらの接続距離は最大500m(800G-2xDR4、MMS4X00-NMTランシーバーの場合)まで可能であることから、P2P構成を配線管理する上で大きな課題になり得る点に注意が必要です。

アーキテクチャのセクション冒頭で記したとおり、例として32台のSUクラスタ/PODを使用します。図14のロジック図と図7のコンポーネント数を参照すると、この設計は合計256台のスパインスイッチで構成され、16台のスパインラックに分散していることがわかります。これは、各ラックに16台のスパインスイッチが搭載されていることを示しています。

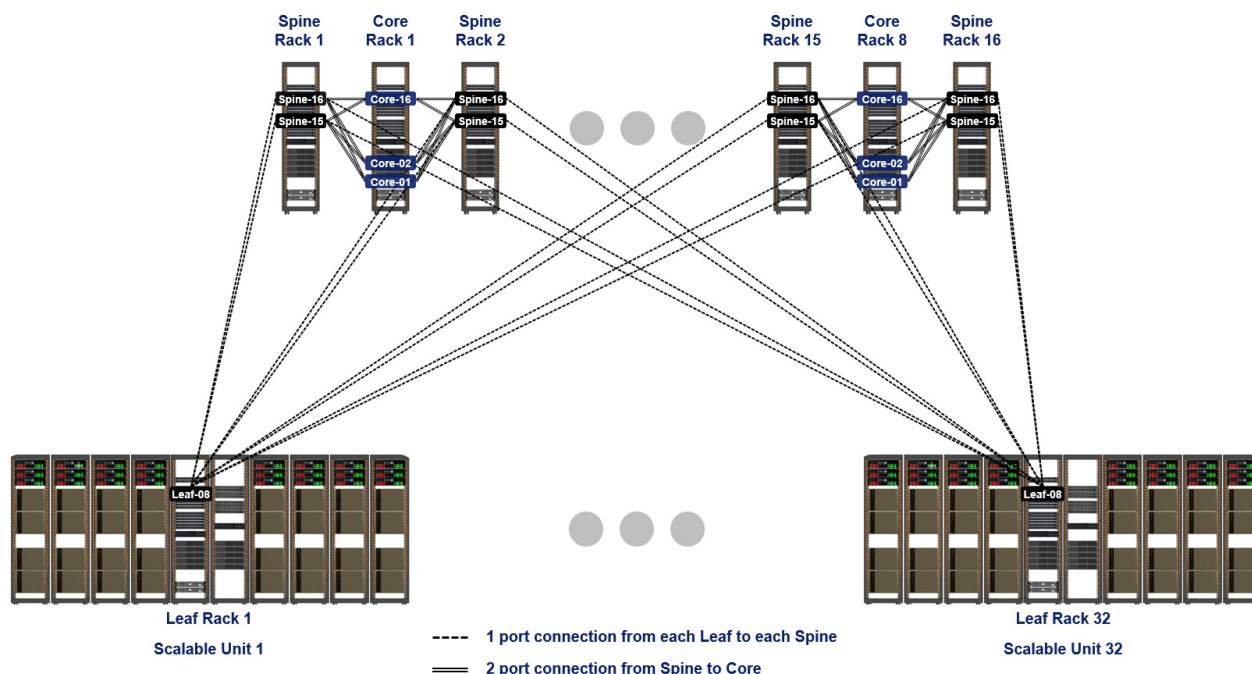
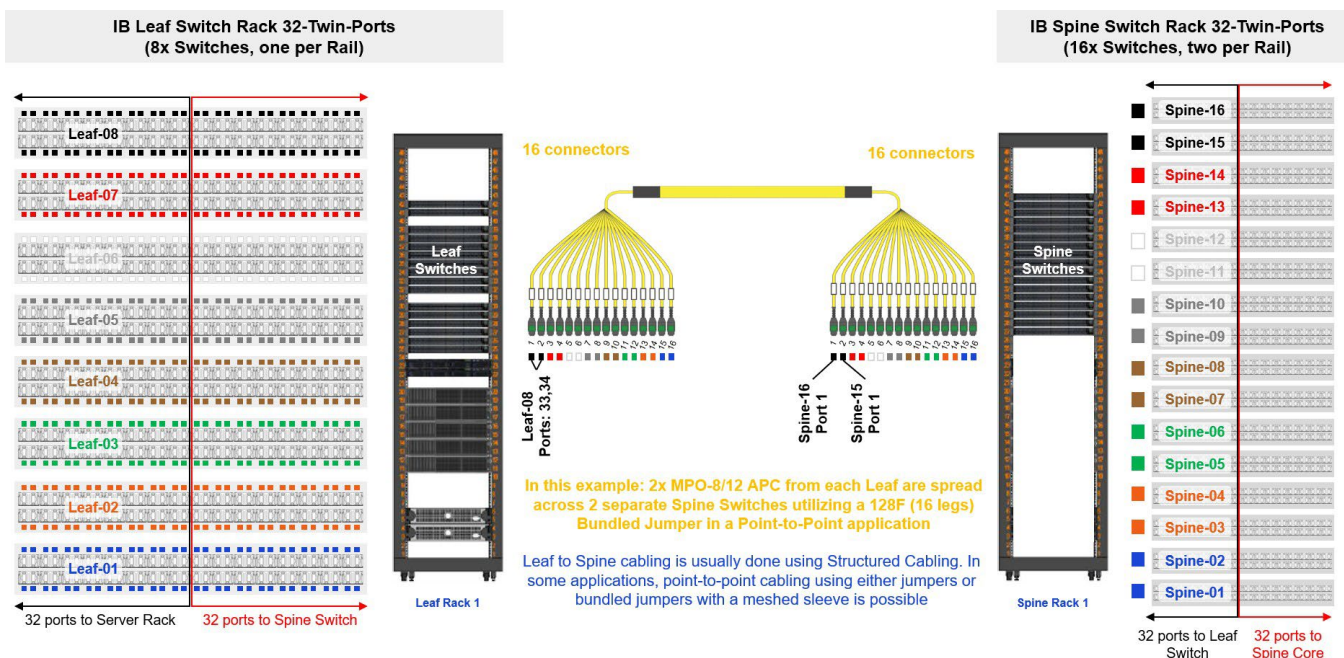


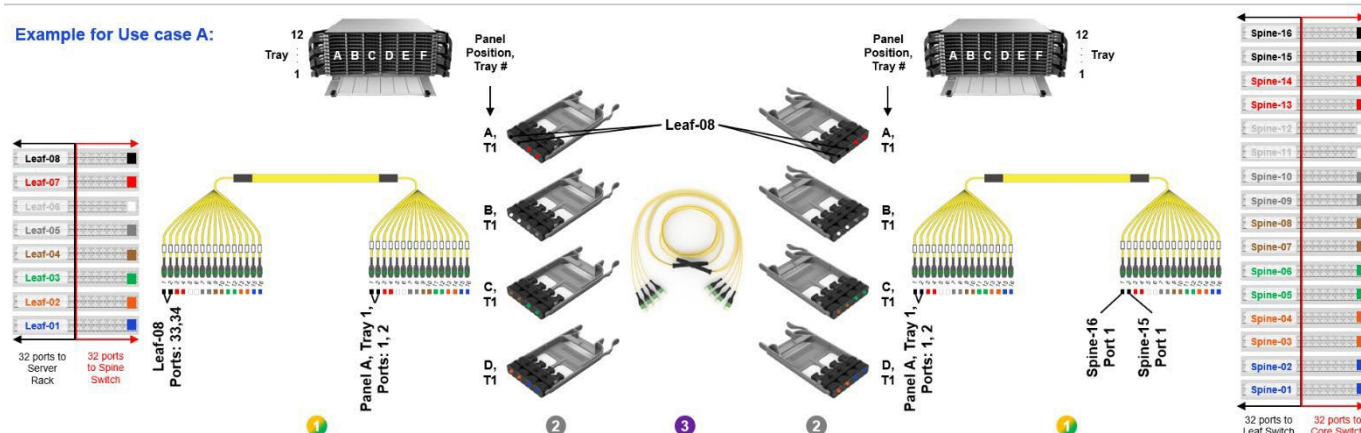
図 14. 32 台のSU GX DGX H100 SuperPOD レール 8(黒レール)のロジック図

推奨オプションではありませんが、ポートマッピングをわかりやすく説明するために、バンドルジャンパーを用いたP2P接続の配線例から確認してみましょう。図 15 の左側には、SUのリーフレックがあり、各スイッチは各レールの色で示されています。右側に示したのは、16 台のスパインスイッチを収納したスパインラック(16 台中の1台)であり、各レールの色に対応したスパインスイッチが2 台ずつ収納されています。

この例では、黒レール(レール 8)をリファレンスとします。リーフレック 1(SU1のリーフレックと考える)のリーフ 08からの接続について、(図 9 のポートマッピングで示したとおり)リーフスイッチポート 33 と 34 を使用すると考えてみましょう。この場合、スパインラック 1 にルーティングし、スパインスイッチ15 のポート 1 とスパインスイッチ16 のポート 1 に接続できます。同一レールにあるすべてのリーフスイッチ、同一レールにあるすべてのスパインスイッチに接続する必要があるため(スパインラックあたり2つ)、別のSUのリーフスイッチのポート33、34もスパインラック1に接続し、スパインスイッチ16のポート2、スパインスイッチ15のポート2に接続します。この目的は、図15に示すように、すべてのSUのリーフスイッチと、すべてのスパインラックの2台のスパインスイッチとの接続を確立することです。レール8をリファレンスとすると、SU1の1つのリーフ08スイッチから16本のバンドルジャンパーを使用して16台のスパインラックに接続する必要があります。結果的に、図14に示すように、合計32台のレール8(黒レール)スパインスイッチに接続することになります。



バンドルジャンパーを用いたP2P接続によるリーフスイッチからスパインスイッチへの接続マッピングがどのように機能するかを理解した上で、同じロジックを構内配線オプションに適用してみましょう。



ここでも黒ルール(ルール8)をリファレンスとして、(上と同様に)リーフラック 1 のリーフ 08 からの接続のためにリーフスイッチポート 33 と 34 を使用する場合、P2P配線で説明したのと同じマッピングに従って、バンドルジャンパーを用いてルーティングできます。ただし、図2で説明したように、アダプターパネルやハウジング(ファイバエンクロージャー)などの構内配線用コンポーネントを追加することになります。

この目的は、マルチファイバトラックをバックボーンとして用いて、異なるエリア間で多芯ファイバ接続を実行することです。リーフからスパインへの接続は、アダプターパネルを使ってマッピングします。

図16の例では、リーフ08のポート33と34からの接続に用いるバンドルジャンパーが、ポジションA、トレイ1(A.T1)にあるアダプターパネルに入り、最初の2つのアダプターポートが黒ルールに割り当てられていることがわかります。その後、複数のアダプターパネルがバックボーン配線を用いて、同じマッピングに従いデータセンターの別セクションにあるハウジングにすべての接続をルーティングします。最後に、バンドルジャンパーを用いてスパインスイッチ15のポート1とスパインスイッチ16のポート1に接続します。

この種類のマッピングに従うと、パッチパネルのポジションEとFが未使用の状態になる場合があります。しかし、予備のアダプターや接続を活用するのは業界の常識となっています(このマッピングの例では未使用のままになる可能性もあります)。こうした予備のポジションは、顧客にとって必要となるかもしれない予備の接続を管理する際に活用できます—より柔軟に、顧客の特定の要求に対応することが可能になるのです。

この接続に構内配線を実装する主なメリットは、スペースを最大限に活用できること、また設置時間を短縮できることです。リーフスイッチとスパインスイッチが異なるエリアに物理配置されている場合は特に、MAC作業(移動、追加、変更)を実施しやすくなります。また、P2P配線では対応できない将来的な拡張にも柔軟に対応できます。

2.4.3. レベルC - スパインからコアへの配線

スパインスイッチからコアスイッチにはP2P接続による配線が可能です。カスタム設計によっては、この接続に構内配線を実装できる場合もあります。その他のオプションとして、DAC(Direct Attach Copper)ケーブルを用いることも可能です。

本文書では32台のSUクラスタ/PODを例として使用しているため、この設計には8つのコアラックに分散配置された合計128台のコアスイッチが含まれ、各ラックには16台のコアスイッチが収納されています。図14のロジック図をリファレンスとして説明すると、コアラックは2台のスパインラックの間に配置されています。

ポートマッピングを分かりやすく説明するために、バンドルジャンパーによるP2P接続を用いましょう。図17の左側にはスパインラック1、右側にはスパインラック2があります。各ラックには、各ルールの色に対応したスパインスイッチが2 台ずつ(合計16台)収納されています。図の中央にあるのがコアラックで、16台のコアスイッチがあります。各コアスイッチが、さまざまなスパインスイッチからのルールすべてを集約しています。

ここでは、図9のポートマッピングに基づき、スパインラック2がスパインラック1を物理的に「ミラーリング」していると仮定してみましょう。すなわち、スパインラック2のリーフスイッチに接続されるポートは1~32ではなく33~64となり、逆にコアスイッチに接続されるポートは33~64ではなく1~32となります。コアスイッチラックを中央に配置することで、バンドルジャンパーを用いて両側のスパインラックからシンプルなルーティングを行うことができます。

まず、スパインラック1からの接続に注目してみましょう。すべてのスパインスイッチのポート33と34はコア01のポート1~32に、すべてのスパインスイッチのポート35と36はコア02のポート1~32に、すべてのスパインスイッチのポート63と64は、コア16のポート1~32にルーティングされます。

次に、スパインラック2からの接続を処理する必要があります。すべてのスパインスイッチのポート1と2はコア01のポート33から64に、すべてのスパインスイッチのポート3と4はコア02のポート33~64に、そしてすべてのスパインスイッチのポート31と32は、コア16のポート33~64にルーティングされます。

IB Spine Switch Rack 32-Twin-Ports
(16x Switches, two per Rail)

IB Core Switch Rack 32-Twin-Ports
(16x Switches: Port 33/34 in all Spines go to Core Switch #1;
Port 35/36 in all Spines go to Core Switch #2; ... Port 63/64 in
all Spines go to Core Switch #16)

IB Spine Switch Rack 32-Twin-Ports
(16x Switches, two per Rail)

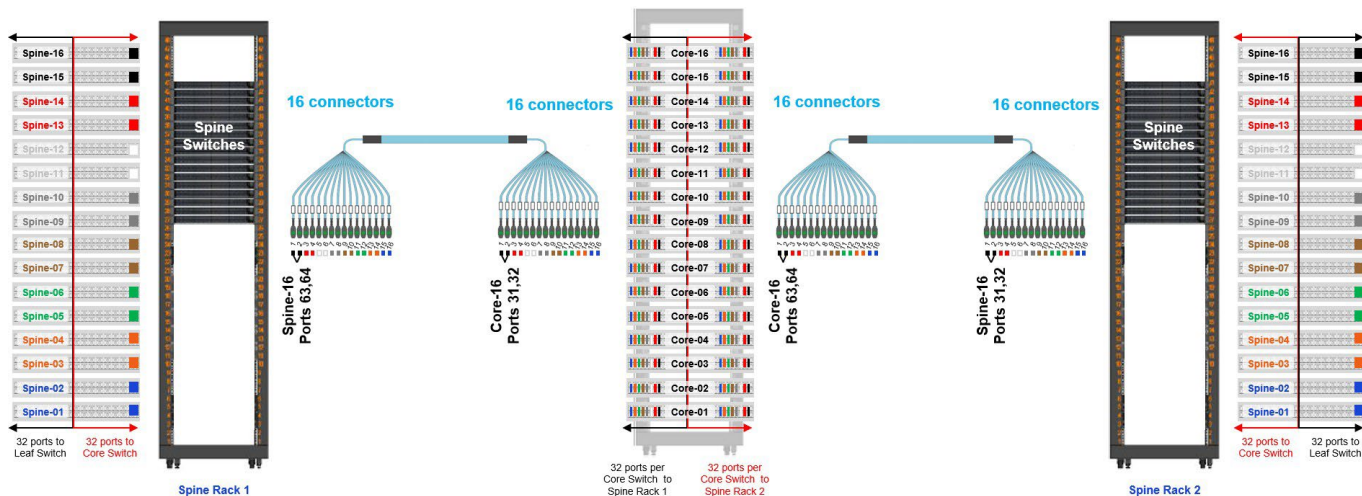


図 17. レベルC - バンドルジャンパーを用いたP2P接続によるスパインからコアへの配線 - 32台のSUクラスタPODの例

図 18は、バンドルジャンパーを用いた配線をより詳細に示しています。この図はスパインラック1を示しており、すべてのスパインスイッチのポート33と34が、コア01のポート1~32にルーティングされています。スパインからコアへのバンドルジャンパーによるルーティングを容易にするため、垂直ケーブルマネージャー（図18では、分かりやすく例示するためにラックとして示しています）を使用することができます。

Bundled Jumper Routing Example

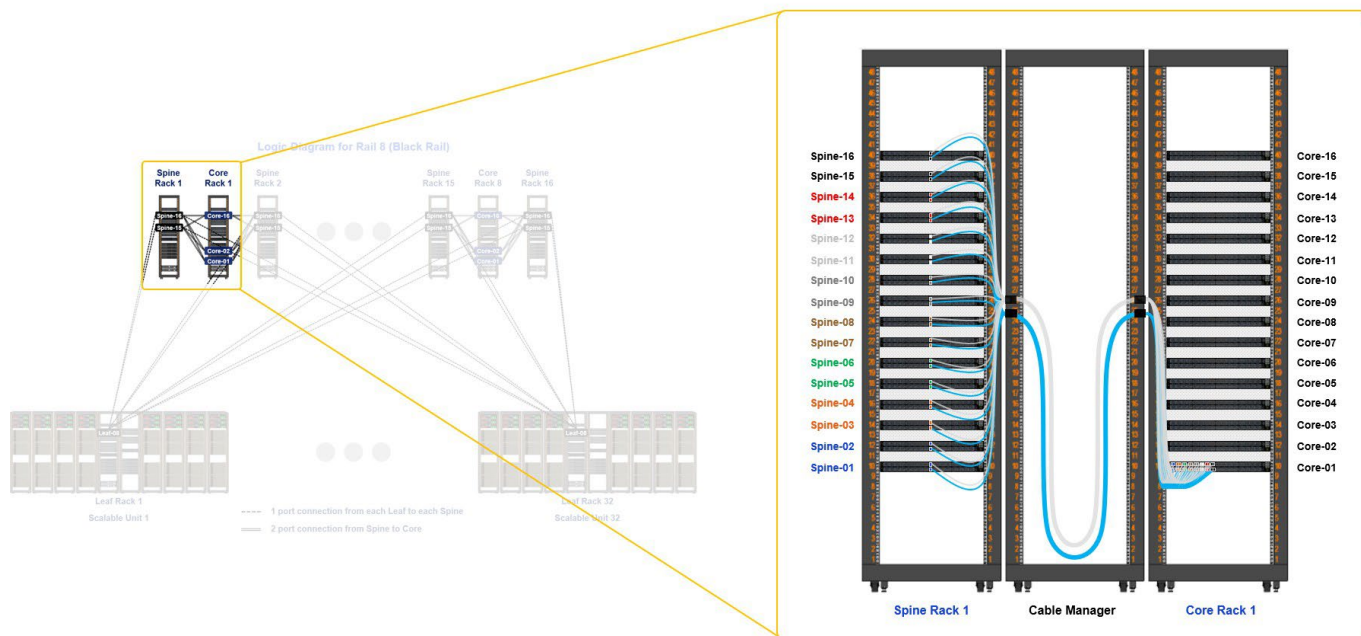


図18. レベルC - すべてのスパインスイッチのポート33と34をCore-01のポート1から32にルーティング

2.5. マルチモードvs シングルモード

ネットワーク内でマルチモードとシングルモードどちらのファイバを使用するかは、設計上の考慮事項によって決まります。マルチモードファイバの最長距離は50mで、通常、SUまたはスパインスイッチからコアスイッチへの接続に用いられます。しかし、SU設計がMoRレイアウトでない場合や、物理的に近接していない場合など、マルチモードファイバの使用が現実的ではない場合もあります。同様に、スパインスイッチとコアスイッチも物理的に近接していない可能性があり、こうした場合も、設計全体にわたりシングルモードファイバの使用を検討することになります。

2.6. 全体像

ここまでクラスタ/PODのサイズ、そしてコンピュートファブリック内のアクティブデバイス間の配線について学んできましたが、以下、使用可能なコンポーネントを図示してまとめてみましょう。コンポーネントは設計によって異なりますが、概ね本文書で検討したさまざまな製品と品番がベースとなります。

2.6.1. 4台のSUクラスタ/PODの配線

図19のように、128台のサーバー、32台のリーフスイッチ、16台のスパインスイッチを備えた、4台のSU PODの配線には以下が必要となります。

- ノードからリーフへの1,024の接続
- リーフからスパインへの1,024の接続
- オプションでコアが利用可能な場合はさらに1,024の接続

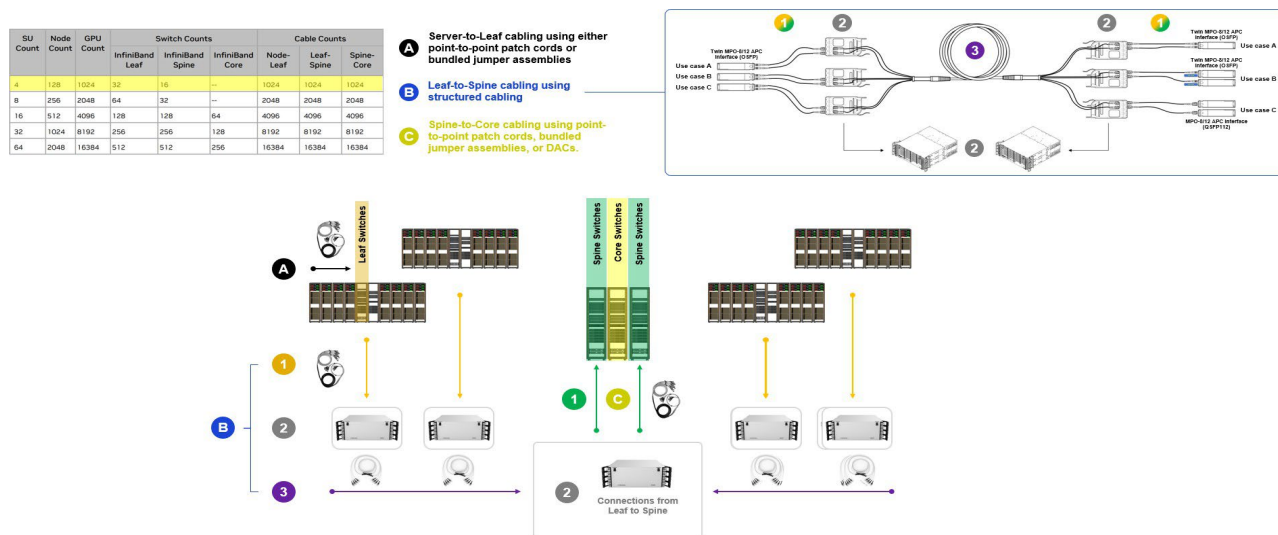


図19. 配線レベルA、B、Cの4台SUクラスタ/POD

上で確認した各レベルの配線を活用することで、以下を理解できます：

- データセンター運営者は、SUのレベルAを用いて、P2Pパッチコードまたはバンドルジャンパーアセンブリのいずれかを使用して、サーバーからリーフへの配線を行うことができます。この場合、SUごとに、長さの異なる256本の8ファイバジャンパー、または長さの異なる256本のファイバを束ねた8本のバンドルジャンパーのいずれかが必要となります。
- データセンター運営者は、レベルBを用いて、構内配線によるリーフからスパインへの配線を行うことができます。この場合、バックボーン配線として、個別のジャンパーまたはバンドルジャンパー、アダプターパネル、ハウジング、マルチファイバートランクを実装する必要があります。このバックボーン配線では64本のトランク(リーフあたり、144本ファイバのトランクx16本)を簡単に使用できるため、P2P接続において個別のパッチコードやバンドルジャンパーを使用する代替方法となります。また、この方法により、予備ポートを確保し、複雑な配線の管理が容易になるのに加えて、データセンター全体の経路スペースを改善することができます。
- コアが実装されている場合、データセンター運営者はレベルCを使用して、P2Pパッチコード、バンドルジャンパーアセンブリ、またはDACを用いてスパインからコアへの配線を行うことができます。

2.6.2. 8台のSUクラスター/PODの配線

図20のように、256台のサーバー、64台のリーフスイッチ、32台のスパインスイッチを備えた8台のSU PODの配線には以下が必要となります。

- ・ ノードからリーフへの2,048の接続
- ・ リーフからスパインへの2,048の接続
- ・ オプションでコアが利用可能な場合さらに2,048の接続

SU Count	Node Count	GPU Count	Switch Counts			Cable Counts		
			InfiniBand Leaf	InfiniBand Spine	InfiniBand Core	Node-Leaf	Leaf-Spine	Spine-Core
4	128	1024	32	16	—	1024	1024	1024
8	256	2048	64	32	—	2048	2048	2048
16	512	4096	128	64	—	4096	4096	4096
32	1024	8192	256	128	—	8192	8192	8192
64	2048	16384	512	256	—	16384	16384	16384

A Server-to-Leaf cabling using either point-to-point patch cords or bundled jumper assemblies

B Leaf-to-Spine cabling using structured cabling

C Spine-to-Core cabling using point-to-point patch cords, bundled jumper assemblies, or DACs.

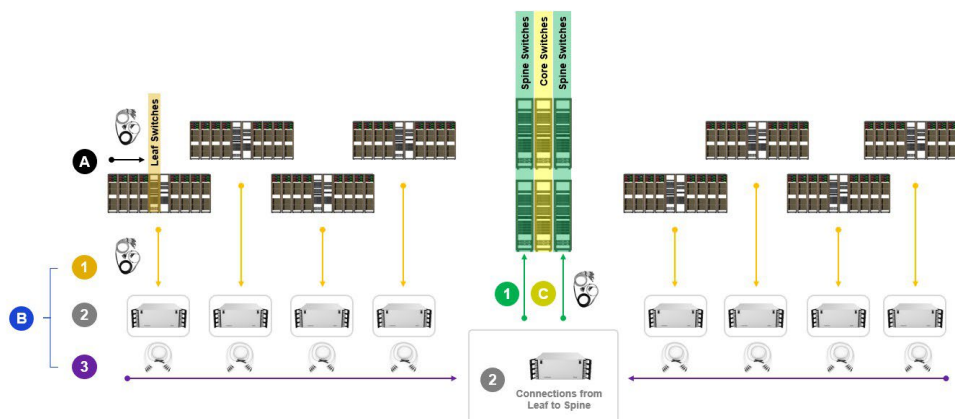
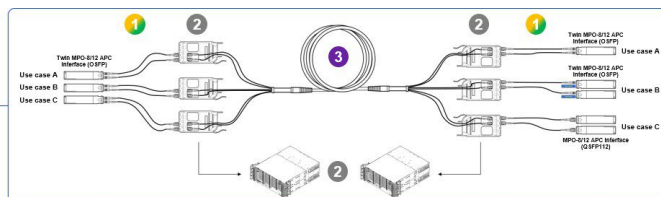


図20.8 配線レベルA、B、Cの8台SUクラスター/POD

上で確認した各レベルの配線を活用することで、以下を理解できます：

- ・ データセンター運営者は、SUのレベルAを使用して、P2Pパッチコードまたはバンドルジャンパーアセンブリのいずれかを使用して、サーバーからリーフへの配線を行うことができます。この場合、SUごとに、長さの異なる256本の8ファイバジャンパー、または長さの異なる256本のファイバを束ねた8本のバンドルジャンパーのいずれかが必要となります。
- ・ データセンター運営者は、レベルBを使用して、構内配線によるリーフからスパインへの配線を行うことができます。この場合、バックボーン配線として、個別のジャンパーまたはバンドルジャンパー、アダプターパネル、ハウジング、マルチファイバートランクを実装する必要があります。このバックボーン配線では、128本のトランク（リーフあたり、144本ファイバのトランクx16）を簡単に使用できるため、P2P接続において個別のパッチコードやバンドルジャンパーを使用する代替方法となります。また、この方法により、予備ポートを確保し、複雑な配線の管理が容易になるのに加えて、データセンター全体の経路スペースを改善することができます。
- ・ コアが実装されている場合、データセンター運営者はレベルCを使用して、P2Pパッチコード、バンドルジャンパーアセンブリ、またはDACを用いてスパインからコアへの配線を行うことができます。

2.6.3. 16台のSUクラスタ／PODの配線

図21のように、512台のサーバー、128台のリーフスイッチ、128台のスパインスイッチ、および64台のコアスイッチを備えた16台のSU PODの配線には以下が必要となります。

- ノードからリーフへの接続: 4,096
- リーフからスパインへの接続: 4,096
- スパインからコアへの接続: 4,096

SU Count	Node Count	GPU Count	Switch Counts			Cable Counts		
			InfiniBand Leaf	InfiniBand Spine	InfiniBand Core	Node-Leaf	Leaf-Spine	Spine-Core
4	128	1024	32	16	—	1024	1024	1024
8	256	2048	64	32	—	2048	2048	2048
16	512	4096	128	64	—	4096	4096	4096
32	1024	8192	256	128	—	8192	8192	8192
64	2048	16384	512	256	—	16384	16384	16384

A Server-to-Leaf cabling using either point-to-point patch cords or bundled jumper assemblies

B Leaf-to-Spine cabling using structured cabling

C Spine-to-Core cabling using point-to-point patch cords, bundled jumper assemblies, or DACs.

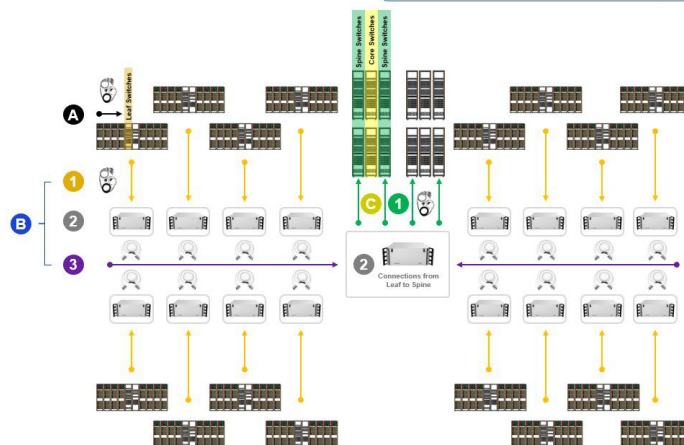
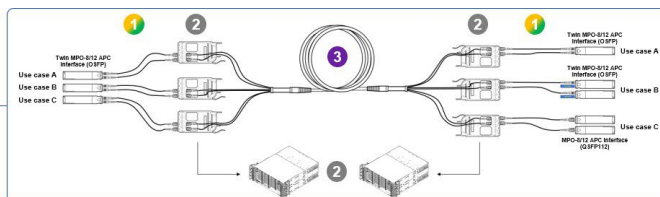


図21. 配線レベルA、B、Cの16台SUクラスタPOD

上で確認した各レベルの配線を活用することで、以下を理解できます：

- データセンター運営者は、SUのレベルAを使用して、P2Pのパッチコードまたはバンドルジャンパーアセンブリのいずれかを使用して、サーバーからリーフへの配線を行うことができます。この場合、SUごとに、長さの異なる256本の個別の8ファイバジャンパー、または長さの異なる256本のファイバを束ねた8つのバンドルジャンパーのいずれかが必要となります。
- データセンター運営者は、レベルBを使用して、構内配線によるリーフからスパインへの配線を行うことができます。この場合、バックボーン配線として、個別のジャンパーまたはバンドルジャンパー、アダプターパネル、ハウジング、マルチファイバートランクを実装する必要があります。このバックボーン配線では、256本のトランク（リーフあたり、144本ファイバのトランクx16）を簡単に使用できるため、P2P接続において個別のパッチコードやバンドルジャンパーを使用する代替方法となります。また、この方法により、予備ポートを確保し、複雑な配線の管理が容易になるのに加えて、データセンター全体の経路スペースを改善することができます。
- データセンター運営者は、レベルCを使用して、P2Pパッチコード、バンドルジャンパーアセンブリ、またはDACを用いてスパインからコアへの配線を行うことができます。

2.6.4. 32台のSUクラスタ/PODの配線

図22のように、1,024台のサーバー、256台のリーフスイッチ、256台のスパインスイッチ、128台のコアスイッチを備えた32台のSU PODの配線には以下が必要となります。

- ノードからリーフへの8,192の接続
- リーフからスパインへの8,192の接続
- スパインからコアへの8,192の接続

SU Count	Node Count	GPU Count	Switch Counts			Cable Counts		
			InfiniBand Leaf	InfiniBand Spine	InfiniBand Core	Node-Leaf	Leaf-Spine	Spine-Core
4	128	1024	32	16	—	1024	1024	1024
8	256	2048	64	32	—	2048	2048	2048
16	512	4096	128	64	—	4096	4096	4096
32	1024	8192	256	128	128	8192	8192	8192
64	2048	16384	512	256	256	16384	16384	16384

A Server-to-Leaf cabling using either point-to-point patch cords or bundled jumper assemblies

B Leaf-to-Spine cabling using structured cabling

C Spine-to-Core cabling using point-to-point patch cords, bundled jumper assemblies, or DACs

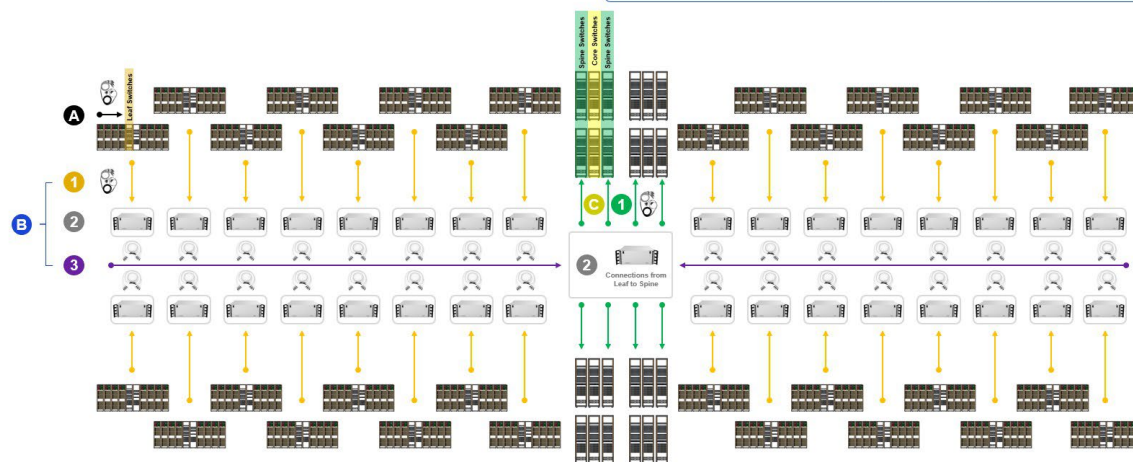
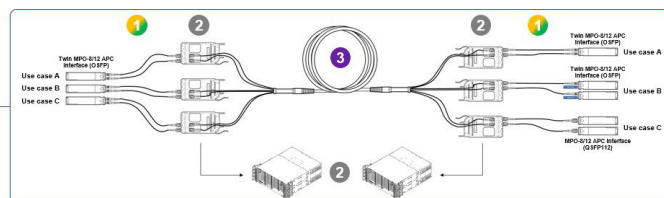


図22. 配線レベルA、B、Cの32台SUクラスタ/POD

上で確認した各レベルの配線を活用することで、以下を理解できます：

- データセンター運営者は、SUのレベルAを使用して、P2Pのパッチコードまたはバンドルジャンパーアセンブリのいずれかを使用して、サーバーからリーフへの配線を行うことができます。この場合、SUごとに、長さの異なる256本の個別の8ファイバジャンパー、または長さの異なる256本のファイバを束ねた8つのバンドルジャンパーのいずれかが必要となります。
- データセンター運営者は、レベルBを使用して、構内配線によるリーフからスパインへの配線を行うことができます。この場合、バックボーン配線として、個別のジャンパーまたはバンドルジャンパー、アダプターパネル、ハウジング、マルチファイバートランクを実装する必要があります。このバックボーン配線では、512本のトランク（リーフあたり、144本ファイバのトランクx16）を簡単に使用できるため、P2P接続において個別のパッチコードやバンドルジャンパーを使用する代替方法となります。また、この方法により、予備ポートを確保し、複雑な配線の管理が容易になるのに加えて、データセンター全体の経路スペースを改善することができます。
- データセンター運営者は、レベルCを使用して、P2Pパッチコード、バンドルジャンパーアセンブリ、またはDACを用いてスパインからコアへの配線を行うことができます。

2.6.5. 64台のSUクラスター/PODの配線

図23のように、2,048台のサーバー、512台のリーフスイッチ、512台のスパインスイッチ、および256台のコアスイッチを備えた64台のSU PODの配線には以下が必要となります。

- ・ ノードからリーフへの16,384の接続
- ・ リーフからスパインへの16,384の接続
- ・ スパインからコアへの16,384の接続

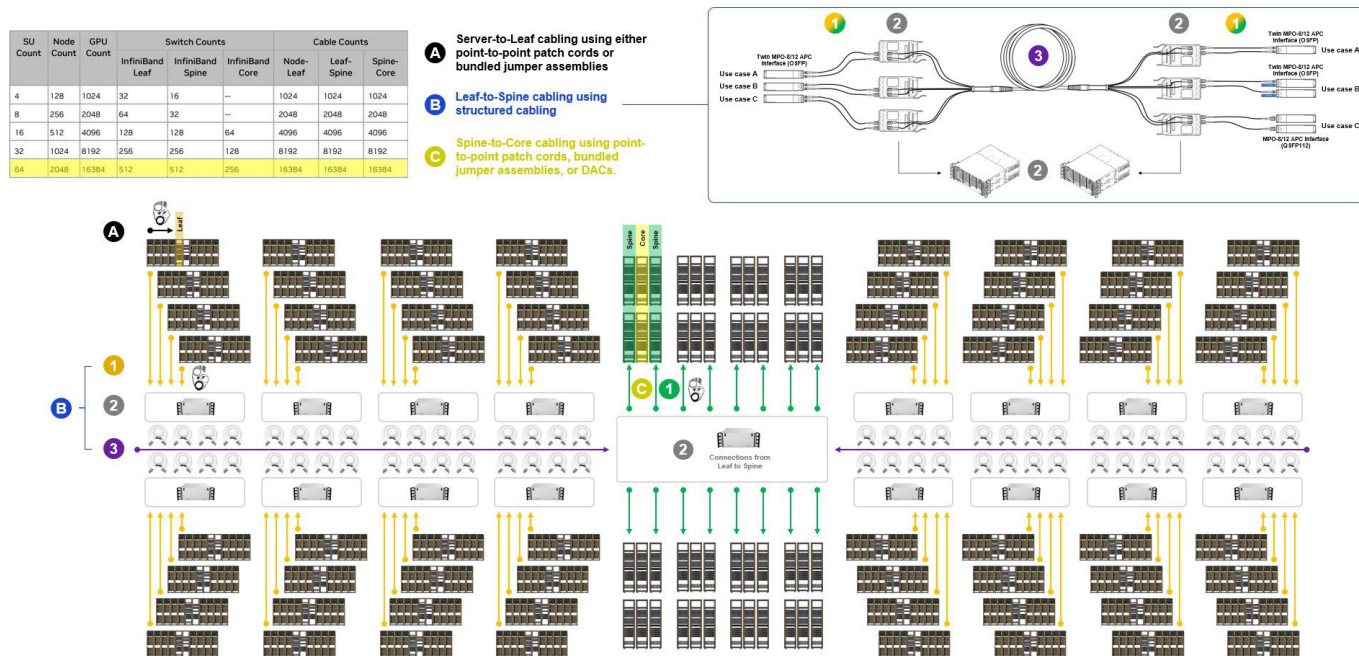


図23. 配線レベルA、B、Cの64台SUクラスター/POD

上で確認した各レベルの配線を活用することで、以下を理解できます：

- ・ データセンター運営者は、SUのレベルAを使用して、P2Pのパッチコードまたはバンドルジャンパーアセンブリのいずれかを使用して、サーバーからリーフへの配線を行うことができます。この場合、SUごとに、長さの異なる256本の個別の8ファイバジャンパー、または長さの異なる256本のファイバを束ねた8つのバンドルジャンパーのいずれかが必要となります。
- ・ データセンター運営者は、レベルBを使用して、構内配線によるリーフからスパインへの配線を行うことができます。この場合、バックボーン配線として、個別のジャンパーまたはバンドルジャンパー、アダプターパネル、ハウジング、マルチファイバートランクを実装する必要があります。このバックボーン配線では、1204のトランク(リーフあたり、144本ファイバのトランクx16)を簡単に使用できるため、P2P接続において個別のパッチコードやバンドルジャンパーを使用する代替方法となります。また、この方法により、予備ポートを確保し、複雑な配線の管理が容易になるのに加えて、データセンター全体の経路スペースを改善することができます。
- ・ データセンター運営者は、レベルCを使用して、P2Pパッチコード、バンドルジャンパーアセンブリ、またはDACを用いてスパインからコアへの配線を行うことができます。

2.7. 結論

まとめると、DGX SuperPODの敷設を最適化するためには、各レベル(A、B、C)の詳細な配線要件を理解することが非常に重要です。さらに、可能な限り構内配線を実装することで、特に大規模な設置において管理性と効率性を高めることができます。

設計段階でコーニングのエンジニアリングチームと連携することにより、各データセンターのレイアウトや顧客要件に適した配線戦略を練ることが可能になります。

付属書1 - 高密度ハウジング

19インチラックやキャビネットにマウント可能なEDGE8® HDハウジングは、EDGE8モジュール、パネル、ハーネス、トランク、ジャンパーと組み合わせることで、業界トップクラスの超高密度接続を実現します。お客様やプロジェクトごとにニーズは異なるため、最も適したハウジングをBOMIに追加ください。

	品番 (グローバル)	モジュール ／パネル 最大数	最大ファイバ密度		高さ
	EDGE8-01U-SP	18	LC	144本	1U
			MTP	576本	
	EDGE8-02U	36	LC	288本	2U
			MTP	1,152本	
	EDGE8-04U	72	LC	576本	4U
			MTP	2,304本	

表11 - 高密度ハウジング

付属書2 - 極性図

極性図は「光ファイバ極性図」と呼ばれることも多く、光ファイバ配線によるデータセンターのリンク設計・実装に不可欠です。極性図は、適切な接続性、信号の整合性、および異なるネットワークコンポーネント間の互換性を確保する上で重要な役割を果たします。

このセクションでは、すでに説明した各シナリオに適用される具体的な極性図について説明します。

シナリオ 1 - 800G および 400G - サーバーからスイッチへのアプリケーション

MPO-8/12 APC間のP2P配線

ユースケース A

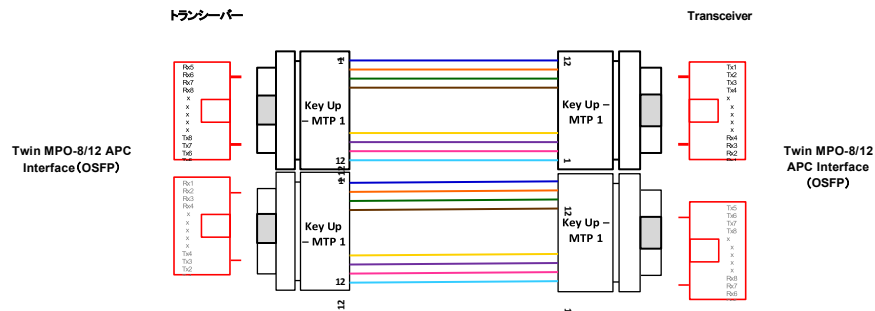


図24 - シナリオ1 - 800Gおよび400G - スイッチからサーバーローカルへの接続 - ユースケースA

ユースケースB

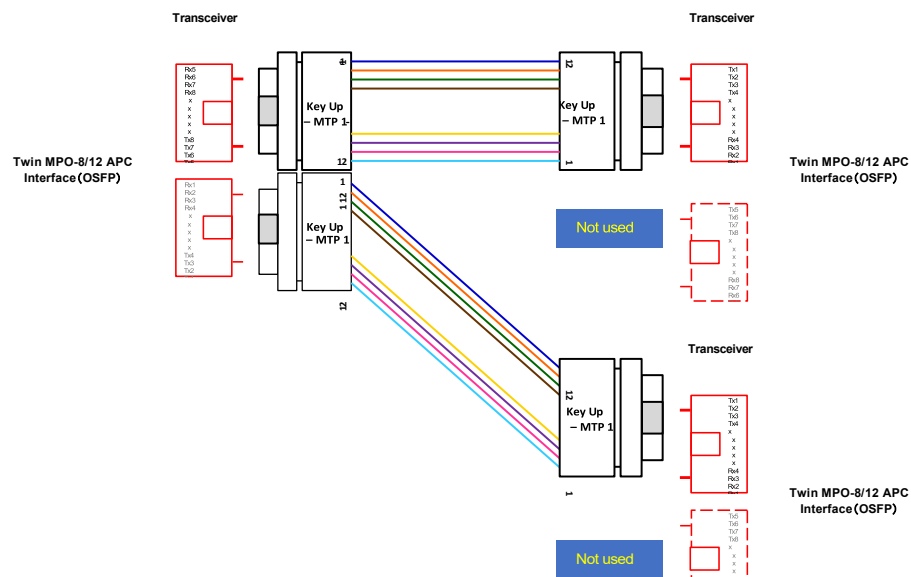


図25 - シナリオ1 - 800Gおよび400G - スイッチからサーバーローカルへの接続 - ユースケースB

ユースケースC

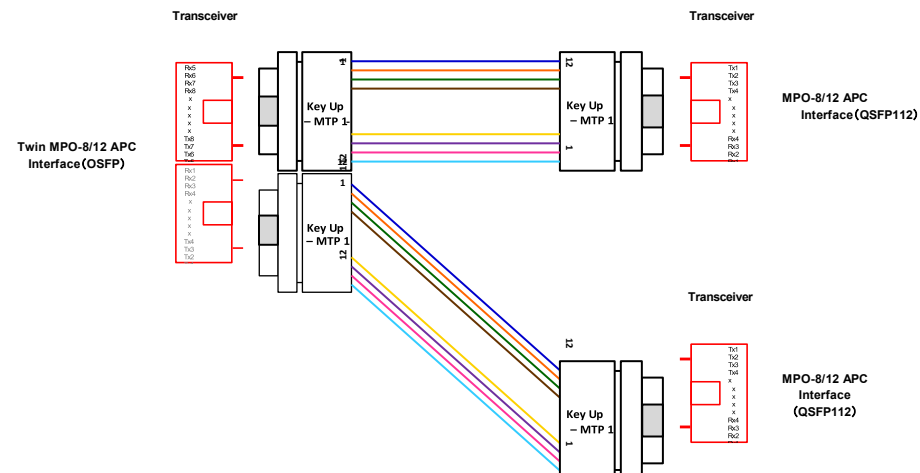


図26 - シナリオ1 - 800Gおよび400G - スイッチからサーバーローカルへの接続 - ユースケースC

シナリオ 2 - 800Gおよび400G - スイッチ間アプリケーション

トランクを用いたDC内MPO-8/12APC間の構内配線

ユースケースA

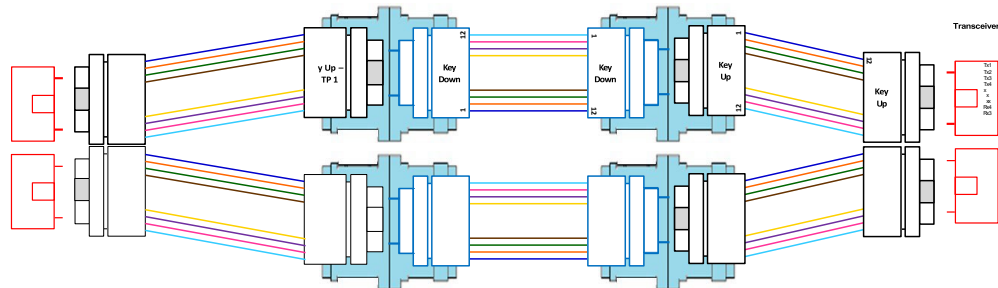


図27 - シナリオ2 - 800Gおよび400G - トランクを用いたDC内スイッチ間接続 - ユースケースA

ユースケースB

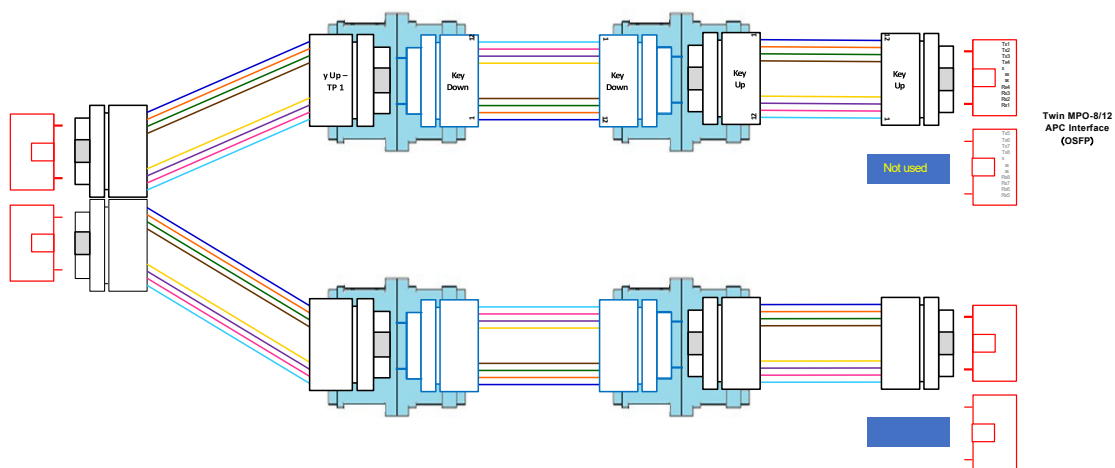


図28 - シナリオ2 - 800Gおよび400G - トランクを用いたDC内スイッチ間接続 - ユースケースB

ユースケースC

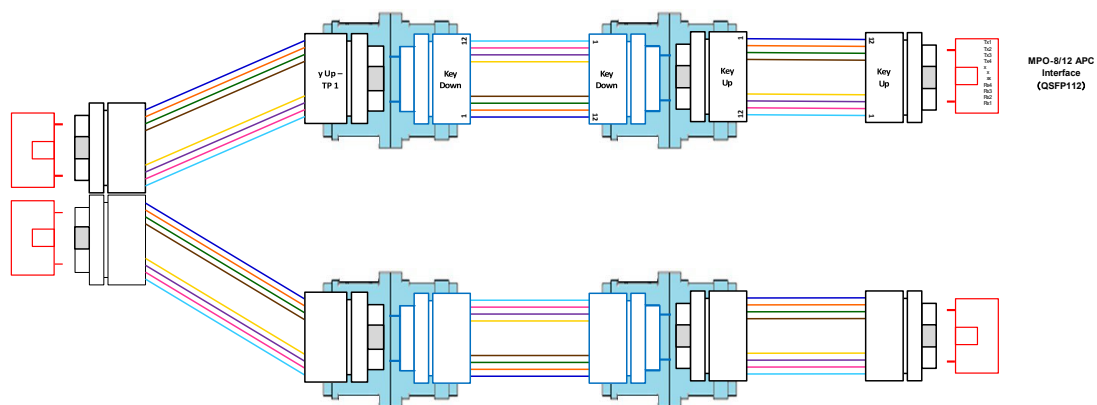


図29 - シナリオ2 - 800Gおよび400G - トランクを用いたDC内スイッチ間接続 - ユースケースC

シナリオ 3 - 800G および 200G - サーバーからスイッチへのアプリケーション

MPO-8/12 APC間のP2P配線

ユースケース A

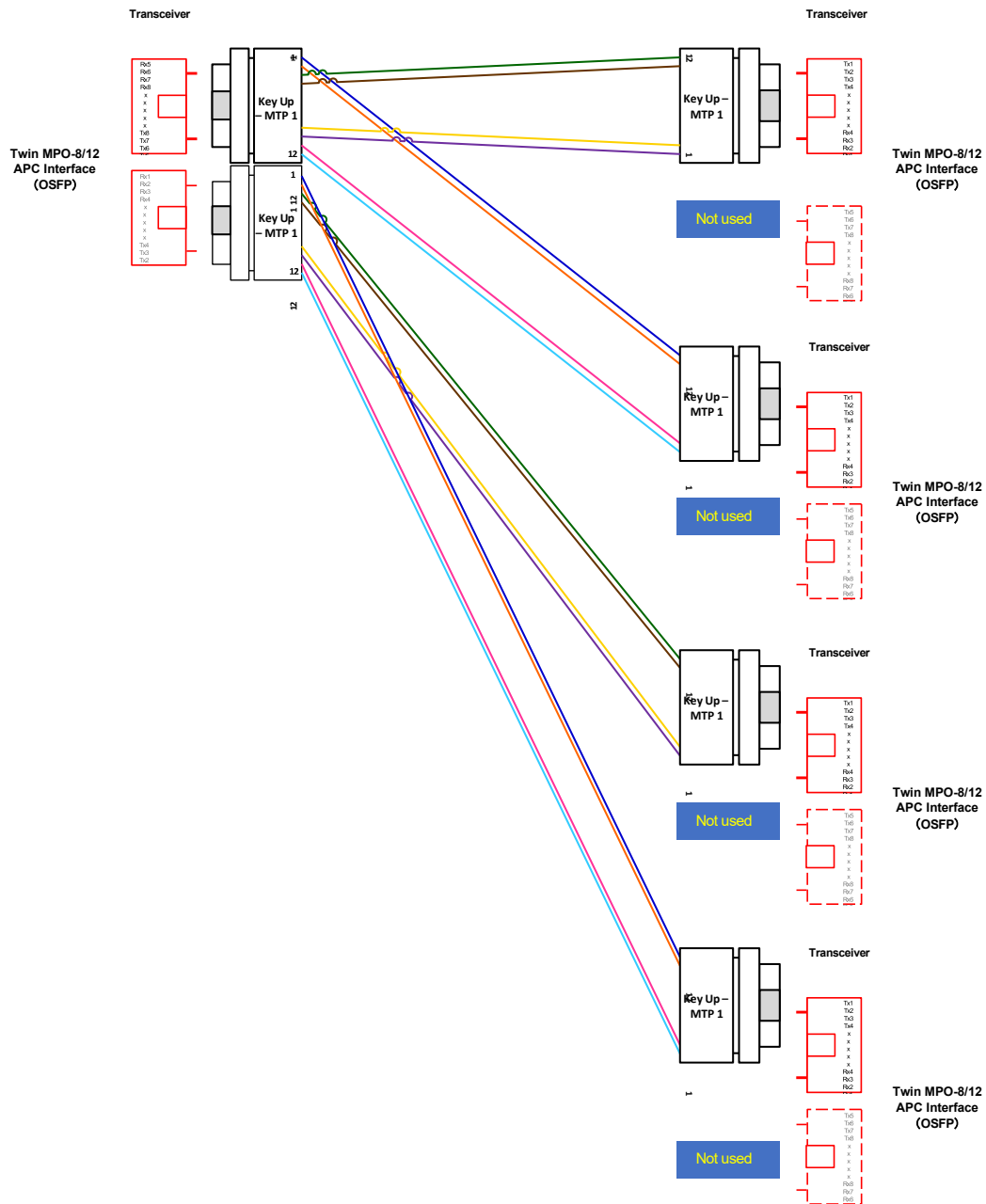


図30 - シナリオ3 - 800Gおよび200G - スイッチからサーバーローカルへの接続 - ユースケースA

ユースケースB

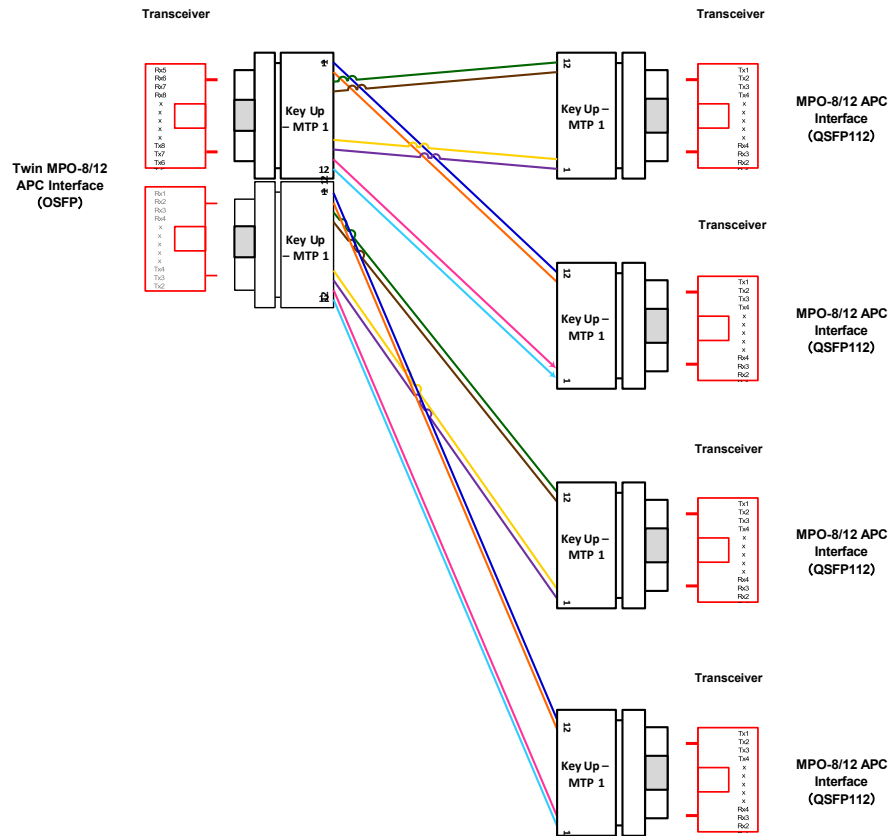


図31 - シナリオ3 - 800Gおよび200G - スイッチからサーバーローカルへの接続 - ユースケースB

ユースケースA

図 32 - シナリオ 4 - 800G および 200G - トランクを用いたDC内スイッチ間接続 - ユースケースA

図 33 - シナリオ 4 - 800G および 200G - トランクを用いたDC内スイッチ間接続 - ユースケース B

シナリオ 5 - 800G および 400G - スイッチ間アプリケーション

LC-Duplex間のP2P配線

ユースケースA

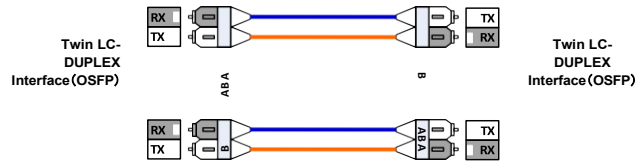


図34 - シナリオ5 - 800Gおよび400G - スイッチ - スイッチローカル接続 - ユースケースA

ユースケースB

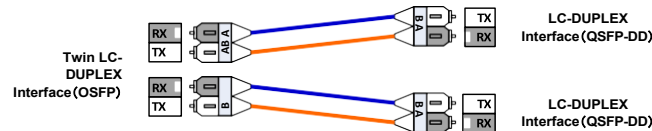


図35 - シナリオ5 - 800Gおよび400G - スイッチ - スイッチローカル接続 - ユースケースB

シナリオ 6 - 800Gおよび400G - スイッチ間アプリケーション

トランクを用いたDC内LC-Duplex間の構内接続

ユースケースA

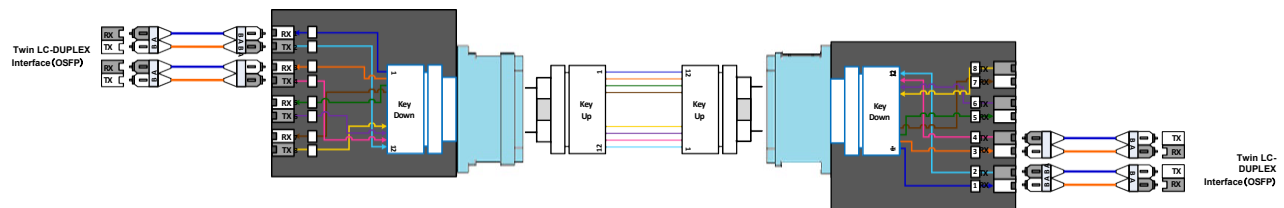


図36 - シナリオ6 - 800Gおよび400G - トランクを用いたDC内スイッチ間接続 - ユースケースA

ユースケースB

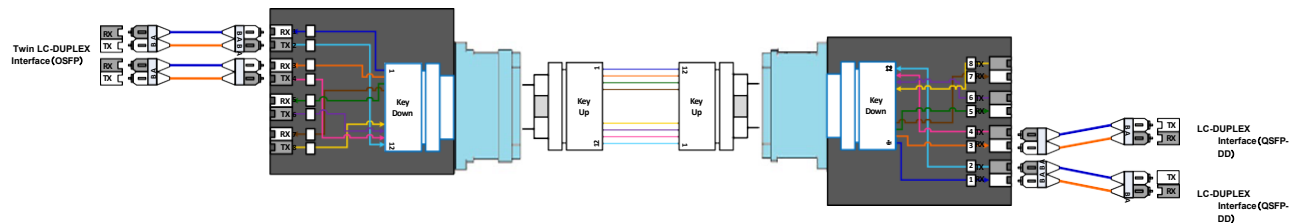


図37 - シナリオ6 - 800Gおよび400G - トランクを用いたDC内スイッチ間接続 - ユースケースB

付属書3-参考文献と連絡先

ここでは、NVIDIA Overview Whitepapersへの参考文献の一部リストを紹介しています。NVIDIA製品のより詳細な情報については、<https://docs.nvidia.com/>にアクセスしてください。

トランシーバー:

- **MMS4X00-NM** 800Gbps Twin ポートOSFP 2x400Gb/s シングルモード 2xDR4、500m
<https://docs.nvidia.com/networking/display/mms4x00nm800g500m/application+overview>
- **MMS4X00-NS** 800Gbps Twin ポートOSFP 2x400Gb/s シングルモード 2xDR4、100m
<https://docs.nvidia.com/networking/display/800gmms4x00ns/overview>
- **MMA4Z00-NS** 800Gb/s Twin ポートOSFP、2x400Gb/s マルチモード2xSR4、50m
<https://docs.nvidia.com/networking/display/800gmma4z00ns/overview>
- **MMS4X50-NM** 800Gb/s Twin ポートOSFP、2x400Gb/s シングルモード2xFR4、2km
<https://docs.nvidia.com/networking/display/mms4x50nm800g2kmpub>

NVIDIAケーブル:

- **MFP7E30-Nxxx**、シングルモード、ストレートクロスオーバー ファイバケーブル
<https://docs.nvidia.com/networking/display/mfp7e30nxxxpub/specifications>
- **MFP7E40-Nxxx**、シングルモード、スプリッタークロスオーバーファイバケーブル
<https://docs.nvidia.com/networking/display/mfp7e40nxxxpub/specifications>
- **MFP7E10-Nxxx**、マルチモード、ストレートクロスオーバー ファイバケーブル
<https://docs.nvidia.com/networking/display/mfp7e10nxxx/specifications>
- **MFP7E20-Nxxx**、マルチモード、スプリッタークロスオーバーファイバケーブル
<https://docs.nvidia.com/networking/display/mfp7e20nxxx/specifications>

DGX SuperPODアーキテクチャ:

- **DGX H100**
<https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastructure-h100/latest/dgx-superpod-architecture.html>
- **DGX B200**
<https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastructure-b200/latest/dgx-superpod-architecture.html>
- **DGX GB200**
<https://www.nvidia.com/en-us/data-center/dgx-superpod-gb200/>

お問い合わせ

さらにサポートが必要な場合は、当社専門チームまでお気軽にお問い合わせください。[こちら](#)をクリックするとメールを作成いただけます。本文にはお名前(フルネーム)、ご所属、お問い合わせ内容をご記入ください。プロフェッショナルサービスチームのメンバーが迅速に対応いたします。