



CORNING

H100 布线指南

基于结构化布线的NVIDIA 800G
InfiniBand和以太网
连接解决方案

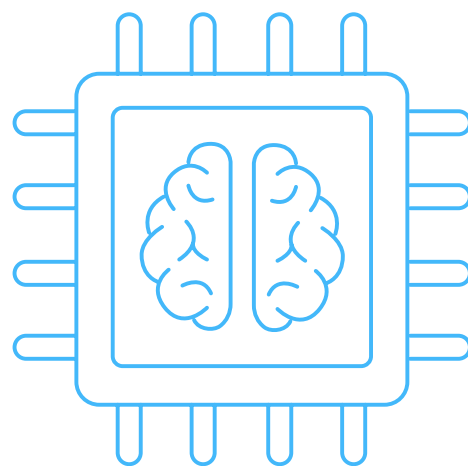
AEN 185



介紹

随着我们迅速迈入生成式人工智能 (Generative AI) 的时代, 为 NVIDIA 的 H100 GPU 系统提供量身定制的连接解决方案变得至关重要。我们致力于为您提供定制化解决方案, 帮助您始终保持领先地位, 同时提供灵活的定价选项和免费的设计支持。

在这个快速发展的领域, 时机就是一切。人工智能的世界将需要定制化的连接解决方案, 这些解决方案不仅具备前瞻性, 还能以数据中心行业前所未有的速度进行交付。



是否准备好购买?

需要免费的设计支持?

正在为您的项目寻找定制化解决方案?

[点击这里](#)

我们将协助您对接 HSDC 创新团队。

时间是您最宝贵的资源, 抓紧行动!



目录

1. 了解收发器类型、端口拆分和布线场景	4
1.1. 场景1 - 800G和400G - 服务器到交换机应用 MPO-8/12 APC	5
1.2. 场景2 - 800G和400G - 交换机到交换机应用 MPO-8/12 APC	6
1.3. 场景3 - 800G和200G - 服务器到交换机应用 MPO-8/12 APC	7
1.4. 场景4 - 800G和200G - 交换机到交换机应用 MPO-8/12 APC	8
1.5. 场景5 - 800G和400G - 交换机到交换机应用, 点对点 LC-双工UPC	9
1.6. 场景6 - 800G和400G - 交换机到交换机应用, 连接数据中心的 LC-双工UPC	10
1.7. 每个场景的收发器选项和端口拆分连接	11
2. NVIDIA DGX SuperPOD布线架构参考指南	12
2.1. 了解DGX-H100节点(服务器)的计算网络连接	12
2.2. 了解Quantum-2 9700 NDR IB交换机	13
2.3. 理解SuperPOD的构建块——可扩展单元的概念	13
2.4. 在NVIDIA DGX SuperPOD中实现布线场景	14
2.4.1. A级-服务器到叶节点的布线	14
2.4.2. B级-叶节点到脊节点的布线	16
2.4.3. C级-脊节点到核心的布线	18
2.5. 多模 vs 单模	20
2.6. 整体概览	20
2.6.1. 4可扩展单元(SU)集群/POD的布线	20
2.6.2. 8可扩展单元(SU)集群/POD布线	21
2.6.3. 16可扩展单元(SU)集群/POD的布线	22
2.6.4. 32可扩展单元(SU)集群/POD的布线	23
2.6.5. 64可扩展单元(U)集群/POD的布线	24
2.7. 结论	24
附件1 - 高密度机箱	25
附件2 - 极性图纸	26
场景1 - 800G和400G - 服务器到交换机应用MPO-8/12 APC	26
场景2 - 800G和400G - 交换机到交换机应用MPO-8/12 APC	27
场景3 - 800G和200G - 服务器到交换机应用MPO-8/12 APC	28
场景4 - 800G和200G - 交换机到交换机应用MPO-8/12 APC	30
场景5 - 800G和400G - 交换机到交换机应用, 点对点 LC-双工UPC	31
场景6 - 800G和400G - 交换机到交换机应用, 数据中心内的 LC-双工UPC	31
附件3 - 参考文献和联系方式	32

1. 1.了解收发器类型、端口拆分和布线场景

本应用工程文档将讨论与200G、400G和800G收发器一起使用的不同类型的光纤连接方案，以及在机柜内或同一排机柜之间，以及跨数据中心使用400G NDR InfiniBand Quantum-2和400GbE Spectrum-4 Ethernet (400G IB/EN) 交换机功能时的端口分解光纤连接方案。

随着800G的引入，NVIDIA光收发模块采用新的双MPO-8/12光收发接口，采用2根8芯MPO光缆。康宁的EDGE8®解决方案旨在支持基于2芯、4芯、8芯和16芯光纤收发器的单模和多模光接口。

以下是按连接器类型列出的部分NVIDIA收发器的清单

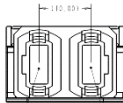
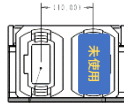
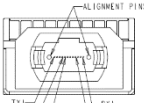
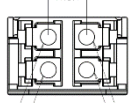
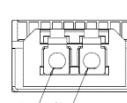
双 MPO-8/12 APC 接口 (OSFP)	MPO-8/12 APC 接口 (OSFP112)	双LC双工UPC接口 (OSFP)	LC双工UPC接口 (QSFP-DD)	
Single Mode ^{a)} 800G-2xDR4 OSFP ^{c)} 2x 8芯收发器 MMS4X00-NM MMS4X00-NS MMS4X00-NS-FLT ^{d)} MMS4X00-NS-T	Single Mode ^{a)} 400G-DR4 OSFP 8芯收发器 使用2个端口中的一个 MMS4X00-NS400 MS4X00 NS400-T 200G-DR4 OSFP ^{e)} 4/8芯收发器 使用2个端口中的1个 MMS4X00-NS400 MS4X00 NS400-T	Single Mode 800G-2xFR4 OSFP 2x 2芯收发器 MMS4X50-NM MMS4X50-NM-T	Single Mode 400G-FR4 QSFP-DD 2芯收发器 MMS1V50-WM MMS1V50-WM-T	
Multimode ^{b)} 800G-2xSR4 OSFP ^{c)} 2x 8芯收发器 MMA4Z00-NS MMA4Z00-NS-FLT ^{d)} MMA4Z00-NS-T	Multimode ^{b)} 400G-SR4 OSFP 8芯收发器 使用2个端口中的一个 MMA4Z00-NS400 MMA4Z00 NS400-T 200G-SR4 OSFP ^{e)} 4/8芯收发器 使用2个端口中的一个 MMA4Z00-NS400 MMA4Z00 NS400-T	Multimode ^{b)} 400G-SR4 QSFP112 8芯收发器 MMA1Z00-NS400 MMA1Z00 NS400-T 200G-SR4 QSFP112 ^{e)} 4/8芯收发器 MMA1Z00-NS400 MMA1Z00-NS400-T		
				

表1. 收发器列表(点击放大)

a) MPO-8/12 APC单模光学器件用黄色拉环和黄色光纤表示。MPO-8/12 APC连接器上的绿色塑料外壳表示角度抛光连接器，与用于较慢速率收发器的超平面抛光连接器(UPC)不兼容。

b) 多模光学器件用棕褐色的拉环和水蓝色的光纤表示。MPO-8/12 APC连接器上的绿色塑料外壳表示角度抛光连接器，与带有水蓝色外壳、用于HDR的超平面抛光连接器(UPC)不兼容。

c) 请注意，在一些文档中，800G收发器可能记为800G - DR8而不是800G - 2xDR4，或记为800G - SR8而不是800G - 2xSR4。然而，连接方式由2x 8-芯MPO-8/12 APC表示。请遵循 NVIDIA的部件编号作为主要参考。

d) 模块的I/O接口内部连接到四个800G双端口OSFP模块。

e) 400G收发器版本将能够通过分路电缆(Y型线束)支持200G，激活400G收发器中的四条通道中的两条(8芯光纤中的4芯)，从而创建一个200G设备。此配置在上表中表示为“4/8-芯”。

f) 以“-T”结尾的收发器部件号指的是以太网版本。

* 关于NVIDIA组件和设计的更多信息，请查阅附件3及NVIDIA概述白皮书的相关资料。

1.1.场景1 - 800G和 400G - 服务器到交换机应用

采用点对点布线连接MPO-8/12 APC到MPO-8/12 APC

应用: Quantum-2 InfiniBand或 Spectrum-4以太网到 a) Quantum-2 InfiniBand或 Spectrum-4以太网; b) ConnectX-7和Bluefield-3; c) DGX H100/Cedar-7链路。

场景1主要用于点对点布线应用, 它涉及将同一可扩展单元内的服务器连接到位于同一机架或机柜内的叶交换机。在某些情况下, 这种类型的布线也可以用于连接不同的交换机, 例如叶交换机到脊交换机或脊交换机到核心交换机。然而, 如果这些交换机物理上位于数据中心内的不同区域, 则不建议使用点对点布线。

请参考第1.7节表8, 以查看场景一和场景二中用例A至C的收发器列表。要了解有关可扩展单元以及如何构建DGX SuperPOD的更多信息, 请参阅本文档的第2节。

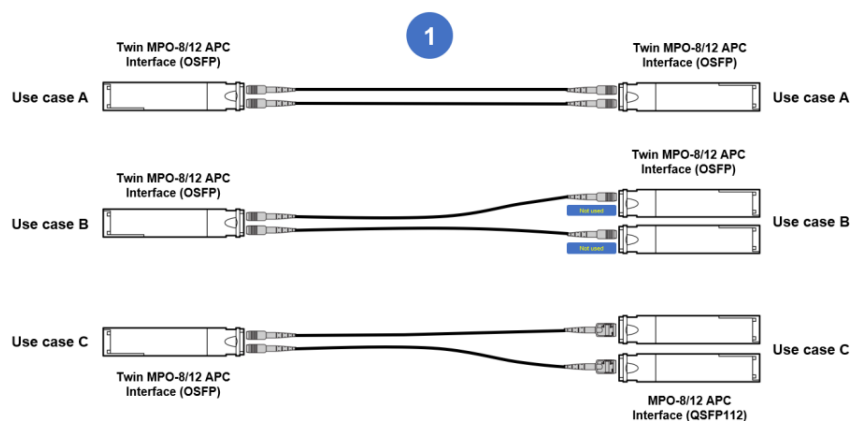


图1. 使用 800G和400G收发器, 结合MPO-8/12 APC的点对点布线用例

项目	参考部件号	OS2部件号 (美洲)	OS2部件号 (EMEA和亚太)	OM4部件号 (美洲)	OM4部件号 (EMEA和亚太)	描述
1	Corning ^{a)}	G-BND32-E8E8G-PN000-xxxF	G-BND32-E8E8G-LZ000-xxxM	G-BND32-E9E9Q-PN000-xxxF	G-BND32-E9E9Q-LZ000-xxxM	32x8芯束装线缆, MTP® APC (不带针) 到MTP APC (不带针), 78-in (2m), B型极性, xxxF (英尺) 或xxxM (米)
		或				
	NVIDIA ^{b)}	JE8E808GE8-NBxxxF	JE8E808GEZ-NBxxxM	JE9E908QE8-NBxxxF	JE9E908QEZ-NBxxxM	EDGE8®, 8芯跳线, MTP®APC (不带针) 到MTP APC (不带针), B型极性, xxxF (英尺) 或xxxM (米)
		或				
		MFP7E30-Nxxx		MFP7E10-Nxxx		NVIDIA MPO-12/APC到MPO-12/APC (8芯), 直通交叉连接光纤。光纤光缆产品, B型极性, xxx表示长度 (米)

a) 美洲地区的康宁光缆使用Plenum光缆, 而EMEA/APJ地区使用LSZH™/CPR额定光缆。单跳线长度从1米到300米不等。捆扎跳线采用网套套管。捆扎长度从10米

到300米不等(分叉点到分叉点), 以5米为增量, 光缆束可提供16根、32根以及自定义数量的光纤分支, 分支长度可以是相等的或不等的。美洲地区也提供以米为单位的长度选项。

b) NVIDIA光缆采用双额定OFNR/LSZH护套。SMF光缆长度有3、5、7、10、15、20、30、50、100和150米选项, MMF光缆长度有3、5、7、10、15、20、25、30、35、40和50米选项。

c) 康宁和NVIDIA光缆都支持InfiniBand、以太网和NVLink协议。

d) 请查看康宁的极性图纸, 详见附件2。

表2. 场景1 - 使用点对点布线与MPO-8/12 APC的 800G和 400G部件编号方案

1.2.场景2 - 800G和 400G - 交换机到交换机应用

使用结构化布线,通过主干光缆在数据中心内连接MPO-8/12 APC到MPO-8/12 APC

应用:Quantum-2 InfiniBand或Spectrum-4以太网到a) Quantum-2 InfiniBand或Spectrum-4以太网;b) ConnectX-7和 Bluefield-3;c) DGX H100/ Cedar-7链路。

在场景2中,利用光纤主干缆作为主干实现结构化布线。此应用主要用于连接不同的交换机,例如从叶节点到脊节点,也可以实现将脊节点连接到核心层。当两个不同的活动设备物理上位于数据中心内的不同区域时,建议使用这个选项。

请参考第1.7节表8, 查看场景1和场景2中适用于用例A至C的收发器列表。要了解有关可扩展单元以及如何构建DGX SuperPOD的更多信息,请参阅本文档中的第2节

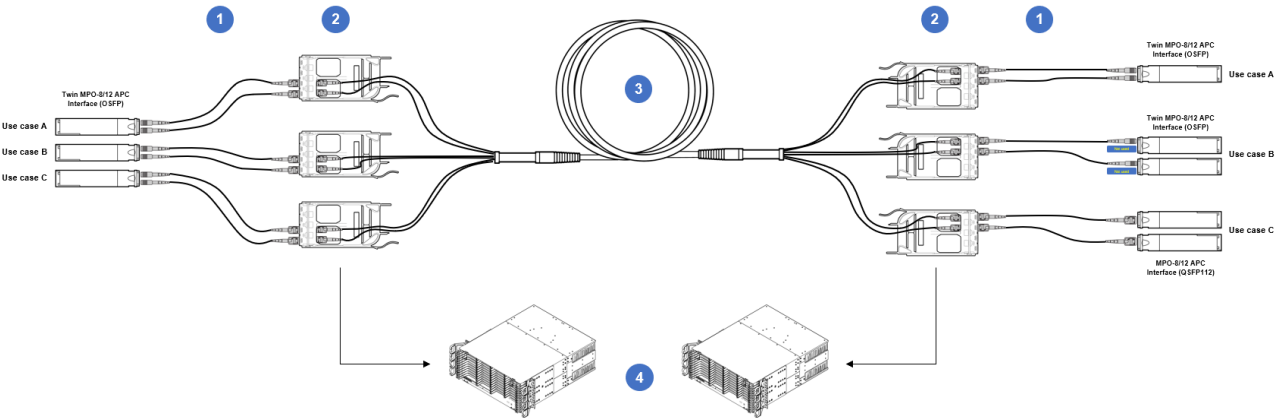


图2. 使用800G和400G收发器, 结合MPO-8/12 APC的结构化布线用例

项目	参考部件号	OS2部件号 (美洲)	OS2部件号 (EMEA和亚太)	OM4部件号 (美洲)	OM4部件号 (EMEA和亚太)	描述
1	Corning ^{a)}	G-BND16-E8E8G-PN000-xxxF	G-BND16-E8E8G-LZ000-xxxM	G-BND16-E9E9Q-PN000-xxxF	G-BND16-E9E9Q-LZ000-xxxM	16x8芯束装线缆, MTP® APC (不带针) 到MTP APC (不带针), 78-in(2m), 分支, B型极性, xxxF (英尺) 或xxxM (米)
		或				
		JE8E808GE8 NBxxxF	JE8E808GEZ NBxxxM	JE9E908QE8 NBxxxF	JE9E908QEZ NBxxxM	EDGE8®, MTP® APC (不带针) 到MTP APC (不带针) 8芯跳线, TIA-568 B型极性, xxxF (英尺) 或xxxM (米)
2		EDGE8-CP32-V1	EDGE8-CP32-V1	EDGE8-CP32-V3	EDGE8-CP32-V3	EDGE8 32-F MTP适配器面板 (4个端口)
3		GE7E7E4GPNDU xxxF ^{b)}	GE7E7E4GLZDDU xxxM ^{b)}	GE2E2E4QPNDU xxxF ^{b)}	GE2E2E4QLZDDU xxxM ^{b)}	EDGE8, MTP主干光缆, 144-F, MTP APC(带针)到MTP APC(带针), 33-in (840mm) 分支, B型极性, 仅在首端配有拉手, 手, xxxF (英尺) 或 xxxM (米)
4		EDGE8-xxU	EDGE8-xxU	EDGE8-xxU	EDGE8-xxU	请参阅附件1, 为您的申请选择最佳方案

注

a) 美洲地区的康宁光缆使用Plenum光缆, 而EMEA/APJ地区使用LSZH™/CPR额定光缆。单跳线长度从1米到300米不等。捆扎跳线采用网状套管。捆扎长度从10米到300米不等(分叉点到分叉点), 以5米为增量, 光缆束可提供16根、32根以及自定义数量的光纤分支, 分支长度可以是相等的或不等的。美洲地区也提供以米为单位的长度选项。

b) 主干缆有8到288芯光纤的规格。

c) 康宁和NVIDIA光缆都支持InfiniBand、以太网和NVLink协议。

d) 请查看康宁的极性图纸, 见附件 2

表3. 场景2 - 使用结构化布线与MPO-8/12 APC的 800G和 400G部件编号方案

1.3.场景3 - 800G和 200G - 服务器到交换机应用

使用点对点布线连接MPO-8/12 APC到MPO-8/12 APC

应用:Quantum-2 InfiniBand或Spectrum-4以太网到ConnectX-7 / Bluefield-3

场景3用于点对点布线应用, 其中400G收发器版本将能够通过使用分路光缆(Y型线束)来支持200G, 该光缆可以激活400G收发器中四条通道中的两条(8芯光纤中的4芯), 从而有效地创建一个200G设备。点对点布线 仅在主动设备位于同一机架或同一机柜内时推荐使用。然而, 如果主动设备在数据中心的不同区域, 则建议使用结构化布线。

请参阅第1.7节表9, 查看场景3和4中适用于用例A和B的收发器列表。要了解有关可扩展单元以及如何构建DGX SuperPOD的更多信息, 请参阅本文档中的第2节。

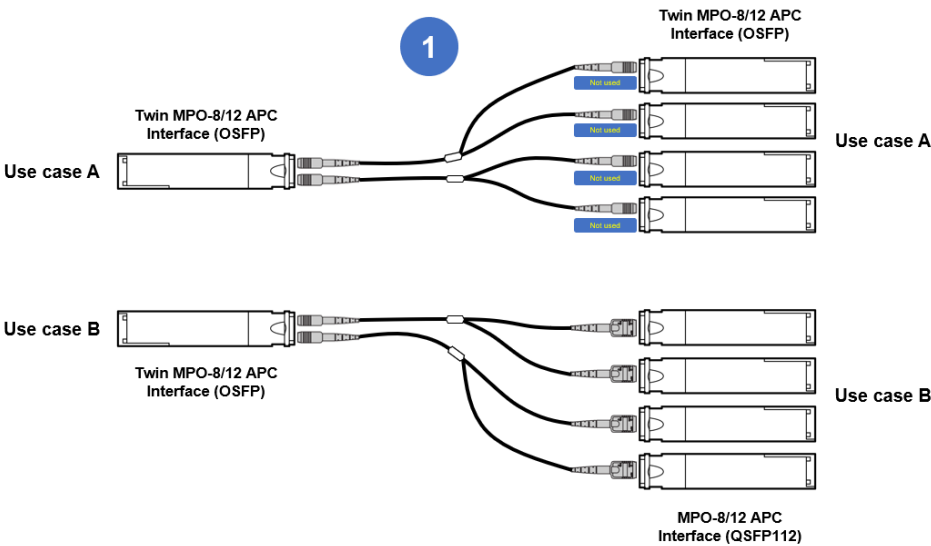


图3. 使用800G和200G收发器, 结合MPO-8/12 APC的点对点布线用例

项目	参考部件号	OS2部件号 (美洲)	OS2部件号 (EMEA和亚太)	OM4部件号 (美洲)	OM4部件号 (EMEA和亚太)	描述
1	Corning ^{a)}	HE8E808GPH LBxxxF	HE8E808GLZ LBxxxM	HE9E908QPH LBxxxF	HE9E908QLZ LBxxxM	8芯Y型跳线, MTP®APC(不带针)到2个4芯MTP APC(不带针), 36-in (910mm)分支长度, B型极性, xxxF (英尺) 或xxxM (米)
		或				
	NVIDIA ^{b)}	MFP7E40-Nxxx		MFP7E20-Nxxx		NVIDIA MPO-12/APC到MPO-12/APC (8芯), 4通道转两条2通道分支光缆, B型极性, xxx表示长度(米)

注

a) 美洲地区的康宁光缆使用Plenum光缆, 而EMEA/APJ地区使用LSZH™/CPR额定电缆。Y型线束长度可从1到60米。美洲地区也提供以米为单位的长度选项。

b) NVIDIA光缆采用双额定OFNR/LSZH护套。SMF和MMF光缆长度有3、5、7、10、15、20、30、50米。

c) 康宁和NVIDIA光缆都支持InfiniBand、以太网和NVLink协议。

d) 400G收发器版本将能够使用分路光缆(Y型线束)支持200G, 激活400G收发器中的四条通道中的两条(8芯光纤中的4芯), 从而创建200G设备。

e) 请查看康宁的极性图纸, 详见附件 2。

表4. 场景3 - 使用点对点布线与MPO-8/12 APC的 800G和 200G部件编号方案

1.4.场景4 - 800G和200G - 交换机到交换机应用

使用结构化布线,通过主干光缆在数据中心内连接MPO-8/12 APC到MPO-8/12 APC

应用:Quantum-2 InfiniBand或Spectrum-4以太网到ConnectX-7 / Bluefield-3

场景4与结构化布线组件一起使用,其中400G收发器版本将能够通过使用分路光缆(Y型跳线)来支持200G,该光缆可以激活400G收发器中四条通道中的两条(8芯光纤中的4芯),从而有效地连接200G设备。在此场景中,有源设备物理上位于数据中心内的不同区域。

请参考第1.7节表9,查看场景3和场景4中适用于用例A和B的收发器列表。要了解更多关于可扩展单元以及如何构建DGX SuperPOD的信息,请参考本文档中的第2节。

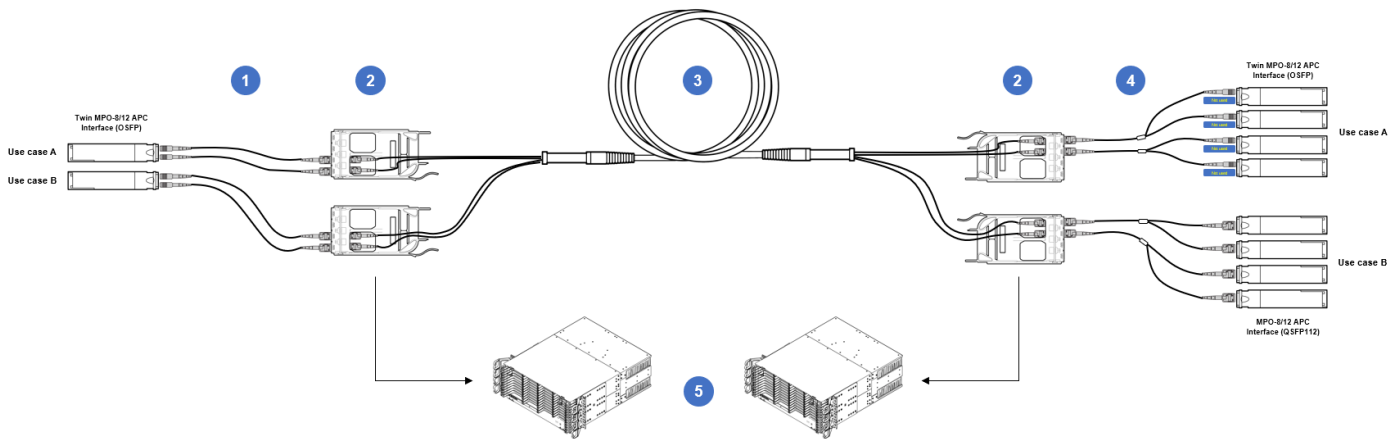


图4. 使用800G和200G收发器与MPO-8/12 APC适配工作的结构化布线用例

项目	参考部件号	OS2部件号 (美洲)	OS2部件号 (EMEA和亚太)	OM4部件号 (美洲)	OM4部件号 (EMEA和亚太)	描述
1	Corning ^{®)}	G-BND16-E8E8G-PN000-xxxF	G-BND16-E8E8G-LZ000-xxxM	G-BND16-E9E9Q-PN000-xxxF	G-BND16-E9E9Q-LZ000-xxxM	16x8芯 束装线缆, MTP®APC(不带针)到MTP APC(不带针), 78-in (2m), B型极性, xxxF (英尺) 或xxxM (米)
		或				
JE8E808GE8 NBxxxF		JE8E808GEZ NBxxxM	JE9E908QE8 NBxxxF	JE9E908QEZ NBxxxM	EDGE8®, MTP APC(不带针)到MTP APC(不带针) 8芯跳线, TIA-568 B型极性, xxxF (英尺) 或xxxM (米)	
2		EDGE8-CP32-V1	EDGE8-CP32-V1	EDGE8-CP32-V3	EDGE8-CP32-V3	EDGE8 32-F MTP适配器面板 (4个端口)
3		GE7E7E4GPNDU xxxF ^{b)}	GE7E7E4GLZDDU xxxM ^{b)}	GE2E2E4QPNDU xxxF ^{b)}	GE2E2E4QLZDDU xxxM ^{b)}	EDGE8, MTP主干, 144-F, MTP APC(带针)到MTP APC(带针), 33-in (840mm)分支, B型极性, 仅在首端配有拉拔握把, xxxF (英尺) 或xxxM (米)
4		HE8E808GPH LBxxxF	HE8E808GLZ LBxxxM	HE9E908QPH LBxxxF	HE9E908QLZ LBxxxM	EDGE8 8芯Y型跳线, MTP®APC(不带针)8芯到2个4芯MTP APC(不带针), 36-in (910mm) 分支长度, B型极性, xxxF (英尺) 或xxxM (米)
5		EDGE8-xxU	EDGE8-xxU	EDGE8-xxU	EDGE8-xxU	请参阅附件1, 为您的申请选择最佳方案

注

a) 美洲地区的康宁光缆使用Plenum光缆, 而EMEA/APJ地区使用LSZH™/CPR额定光缆。单跳线长度从1米到300米不等。束状跳线采用网状套管。长度从10米到300米不等(分叉点到分叉点), 以5米为增量, 光缆束可提供16根、32根以及自定义数量的光纤分支, 分支长度可以是相等的或不等的。美洲地区也提供以米为单位的长度选项。

b) 主干光缆有8到288芯光纤的规格。

c) 康宁光缆支持InfiniBand、以太网和NVLink协议。

d) 请查看康宁的极性图纸, 详见附件2。

表5. 场景4-使用结构化布线与MPO-8/12 APC的 800G和 200G部件编号方案

1.5.场景5 - 800G和400G - 交换机到交换机应用

使用点对点连接的LC双工到LC双工

应用:Quantum-2 InfiniBand或Spectrum-4以太网到a) Quantum-2 InfiniBand或Spectrum-4以太网。

场景5主要用于点对点布线应用,它涉及连接位于同一机架或同一机柜内的有源设备。但是,如果这些设备物理上位于数据中心内的不同区域,则不建议使用点对点布线。

LC双工收发器的主要应用是将Quantum-2或Spectrum-4风冷交换机连接在一起,收发器上的亮绿色标记表示最大连接距离可达2km。

请参阅第1.7节表10,查看场景5和6中适用于用例A和B的收发器列表。要了解有关可扩展单元以及如何构建DGX SuperPOD的更多信息,请参阅本文档中的第2节。

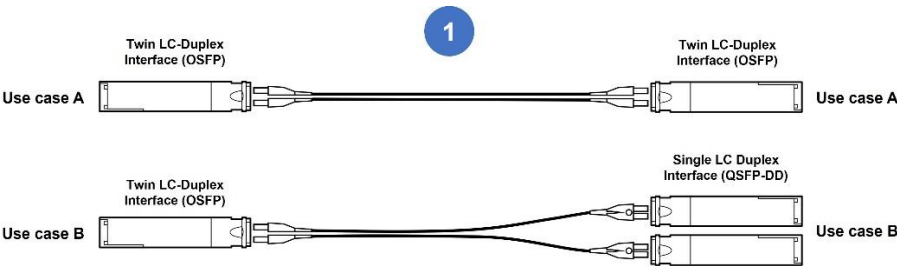


图5. 使用800G和400G收发器, 结合LC双工的点对点布线用例

项目	参考部件号	OS2部件号 (美洲)	OS2部件号 (EMEA和亚太)	描述
1	Corning ^{a)}	787802GD120xxxF	E787802GNZ20xxM	EDGE™ LC UPC Uniboot到LC UPC Uniboot双工跳线。xxxF(英尺)或xxM(米)

注

a) 美洲地区的康宁光缆使用Plenum光缆, 而EMEA/APJ地区使用LSZH™/CPR额定光缆。单跳线长度从1米到300米不等。美洲地区也提供以米为单位的长度选项。

b) 康宁光缆支持InfiniBand、以太网和NVLink协议。

c) 请查看康宁的极性图纸, 详见附件 2。

表6. 场景5 - 使用点对点布线与LC双工的 800G和 400G部件编号方案

1.6. 场景6 - 800G和400G - 交换机到交换机应用

使用结构化布线,通过主干光缆在数据中心内连接LC双工到LC双工

应用:Quantum-2 InfiniBand或Spectrum-4以太网到a) Quantum-2 InfiniBand或Spectrum-4以太网。

在场景6中,结构化布线的实现利用MPO-8/12光缆作为主干光缆。该应用用于连接不同的交换机,例如叶交换机到脊交换机。当两个不同的有源设备物理上位于数据中心内的不同区域时,推荐使用该选项。

LC双工收发器的主要应用是将Quantum-2或Spectrum-4风冷开关连接在一起。收发器上的亮绿色标记表示最大连接距离可达2km。

请参阅第1.7节表10,查看场景5和6中适用于用例A和B的收发器列表。要了解有关可扩展单元以及如何构建DGX SuperPOD的更多信息,请参阅本文档中的第2节。

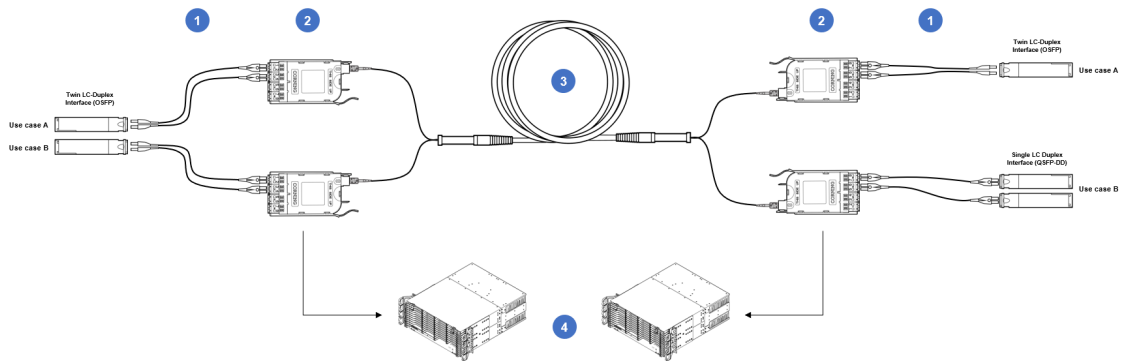


图6. 使用800G和400G收发器,结合LC双工的结构化布线用例

项目	参考部件号	OS2部件号 (美洲)	OS2部件号 (EMEA和亚太)	描述
1	Corning ^{a)}	787802GD120xxxF	E787802GNZ20xxXM	EDGE™ LC UPC Uniboot到LC UPC Uniboot双工跳线。xxxF (英尺)或xxXM (米)
2		ECM8-UM08-04-E8G-ULL	ECM8-UM08-04-E8G-ULL	EDGE8® 超低损耗模块, LC双工到MTP (不带针), 8芯, 通用极性
3		GE7E7E4GPNDDU xxxF ^{b)}	GE7E7E4GLZDDU xxxM ^{b)}	EDGE8, MTP干线, 144-F, MTP APC(不带针)到MTP APC (不带针), 33-in (840mm) 分支, B型极性, 仅在首端配有拉拔握把, xxxF (英尺) 或xxXM (米)
4		EDGE8-xxU	EDGE8-xxU	请参阅附件1, 为您的申请选择最佳方案

注

a) 美洲地区的康宁光缆使用Plenum光缆, 而EMEA/APJ地区使用LSZH™/CPR额定光缆。单跳线长度从1米到300米不等。美洲地区也提供以米为单位的长度选项。

b) 主干光缆有8到288芯光纤的规格。

c) 康宁光缆支持InfiniBand、以太网和NVLink协议。

d) 请查看康宁的极性图纸, 详见附件 2。

表7. 场景6 - 使用结构化布线, 结合LC双工连接的 800G和 400G在主动设备上的部件编号方案

1.7.每个场景的收发器选项和端口拆分连接

要阅读以下表格,请遵循示例:

示例1:在场景2,用例B中,使用表8,图中左侧的800G-2xDR4、MMS4X00-NS通过2个MPO-8/12 APC连接,可以连接到两个OSFP光模块MMS4X00-NS400,从而在右侧形成两个独立的400G连接。

示例2:在场景4,用例B中,使用表9,图中左侧的800G-2xDR4、MMS4X00-NS通过2个MPO-8/12 APC连接,可以利用Y型线束连接到四个200G-DR4、QSFP112、MMS1X00-NS400,从而在右侧形成四个独立的200G连接。

用例	近端光学组件(左边)			远端光学组件(右边)			距离	光纤类型
	速度	NVIDIA端口	收发器	速度	NVIDIA端口	收发器		
A	800G-2xDR4	MMS4X00-NM	OSFP - 2x 8F	800G-2xDR4	MMS4X00-NM	OSFP - 2x 8F	500m	单模
	800G-2xDR4	MMS4X00-NM-T	OSFP - 2x 8F	800G-2xDR4	MMS4X00-NM-T	OSFP - 2x 8F	500m	单模
	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	100m	单模
	800G-2xDR4	MMS4X00-NS-FLT	OSFP - 2x 8F	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	100m	单模
	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	100m	单模
B	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	400G-DR4	MMS4X00-NS400	OSFP - 8F	100m	单模
	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	400G-DR4	MS4X00 NS400-T	OSFP - 8F	100m	单模
C	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	400G-DR4	MMS1X00-NS400	QSFP112 - 8F	100m	单模
	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	400G-DR4	MS1X00 NS400-T	QSFP112 - 8F	100m	单模
A	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	50m	多模
	800G-2xSR4	MMA4Z00-NS-FLT	OSFP - 2x 8F	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	50m	多模
	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	50m	多模
B	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	400G-SR4	MMA4Z00-NS400	OSFP - 8F	50m	多模
	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	400G-SR4	MMA4Z00 NS400-T	OSFP - 8F	50m	多模
C	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	400G-SR4	MMA1Z00-NS400	QSFP112 - 8F	50m	多模
	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	400G-SR4	MMA1Z00 NS400-T	QSFP112 - 8F	50m	多模

表8. 场景1和2中用例A到C适用的收发器

用例	近端光学组件(左边)			远端光学组件(右边)			距离	光纤类型
	速度	NVIDIA端口	收发器	速度	NVIDIA端口	收发器		
A	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	200G-DR4	MMS4X00-NS400	OSFP - 4F	100m	单模
	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	200G-DR4	MS4X00 NS400-T	OSFP - 4F	100m	单模
B	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	200G-DR4	MMS1X00-NS400	QSFP112 - 4F	100m	单模
	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	200G-DR4	MS1X00 NS400-T	QSFP112 - 4F	100m	单模
A	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	200G-SR4	MMA4Z00-NS400	OSFP - 4F	50m	多模
	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	200G-SR4	MMA4Z00 NS400-T	OSFP - 4F	50m	多模
B	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	200G-SR4	MMA1Z00-NS400	QSFP112 - 4F	50m	多模
	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	200G-SR4	MMA1Z00 NS400-T	QSFP112 - 4F	50m	多模

表9. 场景3和4中用例A到C适用的收发器

用例	近端光学组件(左边)			远端光学组件(右边)			距离	光纤类型
	速度	NVIDIA端口	收发器	速度	NVIDIA端口	收发器		
A	800G-2xFR4	MMS4X50-NM	OSFP - 2x 2F	800G-2xFR4	MMS4X50-NM	OSFP - 2x 2F	2km	单模
	800G-2xFR4	MMS4X50-NM-T	OSFP - 2x 2F	800G-2xFR4	MMS4X50-NM-T	OSFP - 2x 2F	2km	单模
B	800G-2xFR4	MMS4X50-NM	OSFP - 2x 2F	400G-FR4	MMS1V50-WM	QSFP-DD - 2F	2km	单模
	800G-2xFR4	MMS4X50-NM-T	OSFP - 2x 2F	400G-FR4	MMS1V50-WM-T	QSFP-DD - 2F	2km	单模

表10. 场景5和6中用例A到C适用的收发器

2.NVIDIA DGX SuperPOD布线架构参考指南

在本节中,我们将回顾如何连接NVIDIA DGX SuperPOD。虽然本指南使用NVIDIA DGX H100作为参考,但重要的是要注意,类似的布线组件和基础设施也可用于涉及DGX B200和DGX GB200的应用。每个POD的具体考虑将取决于个别客户的设计需求。

DGX SuperPOD由一定数量的可扩展单元(SU)组成,每个SU包含一定数量的节点(服务器)。此配置决定每个POD的GPU总数。SU是SuperPOD的构建模块。

POD也可以被称为集群,这取决于每个客户的偏好。了解POD或集群的大小将决定InfiniBand叶交换机、InfiniBand骨干交换机和InfiniBand核心交换机所需的数量,这一点至关重要。基于该配置,我们可以计算所需的光缆或连接数量。

在图7中,如果我们以一个包含32个可扩展单元的中型集群或POD为例,这相当于1,024个节点或服务器,等同于8,192个GPU。这个设置将需要一个POD中有256个IB叶交换机、256个IB主干交换机和128个IB核心交换机。我们将在整个文档中以这个POD大小作为参考。请注意,小型、中型或大型的分类并不是由NVIDIA定义的,但我们将在本指南中将其作为参考。

“光缆数量”这一列显示了连接系统所需的单个MPO-8/12 APC连接的数量。例如,要将每个节点(服务器)连接到IB叶交换机,我们将需要8,192个MPO-8/12 APC连接。要将IB叶交换机连接到IB主干交换机,我们将需要额外的8,192个MPO-8/12 APC连接。最后,要将IB主干交换机连接到IB核心交换机,我们还需要另外8,192个MPO-8/12 APC连接,总共在整个POD中需要24,576个MPO-8/12 APC连接。

	SU 数量	节点数量	GPU 数量	交换机数量			光缆数量		
				InfiniBand 叶交换机	InfiniBand 脊交换机	InfiniBand 核心交换机	节点 叶节点	叶节点- 脊节点	脊节点- 核心
小型集群/POD	4	128	1024	32	16	--	1024	1024	1024
	8	256	2048	64	32	--	2048	2048	2048
中型集群/POD	16	512	4096	128	128	64	4096	4096	4096
	32	1024	8192	256	256	128	8192	8192	8192
大型集群/POD	64	2048	16384	512	512	256	16384	16384	16384

图 7. SuperPOD组件计数

2.1. 了解DGX-H100节点 (服务器) 的计算网络连接

NVIDIA DGX-H100节点具有InfiniBand计算网络、InfiniBand存储网络、带内管理网络和带外管理网络的连接接口。

本架构指南的重点将集中在InfiniBand计算结构上。这种结构允许在每个节点上连接 4个双端口OSFP, 总计8个400G连接。这些连接使用单模或多模 MPO-8/12 APC连接器 进行连接,如图 8所示。

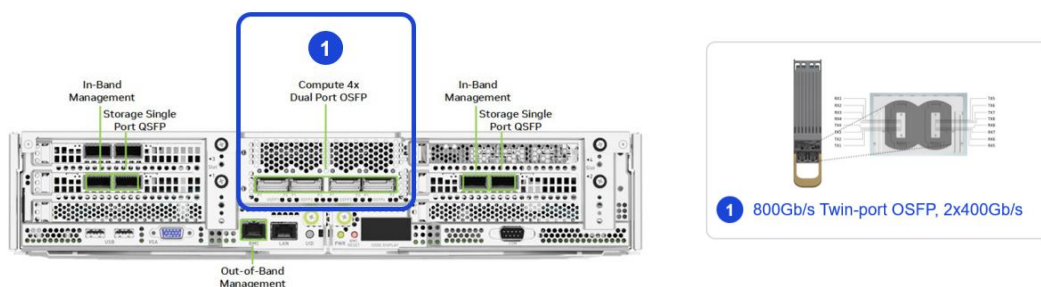


图8. DGX H100 CPU托盘的背面及其连接方式

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

2.2. 了解Quantum-2 9700 NDR IB交换机

Quantum-2交换机支持32个OSFP双MPO-8/12 APC端口。这意味着每台交换机最多可以提供64个 400Gbps的单端口。

图9展示了交换机级别端口映射的示例,用户可以在整个文档中将其作为参考。请注意,这只是客户进行端口映射的一种方法。

由于Quantum-2交换机将在网络中作为叶节点、脊节点或核心节点等不同角色进行部署,端口分配可以按照图9中建议的方法进行交换机间的连接。

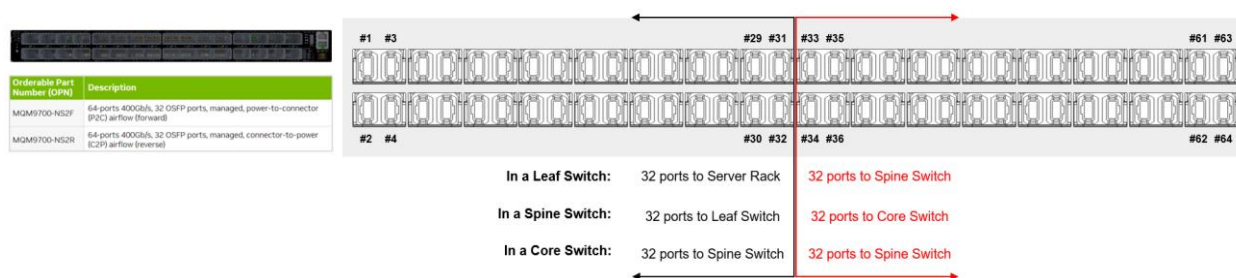


图9. Quantum-2交换机端口映射示例

需要注意的是,位于Quantum-2交换机底部端口的OSFP光模块将会倒置,从而使端口连接方向相反,如图10所示。在这个示例中,我们可以看到,双端口1提供了连接X和Y,而倒置的双端口2则提供了连接Y和X。区分这一点的物理方法是观察两个光模块的散热片,它们将呈相对的方向。

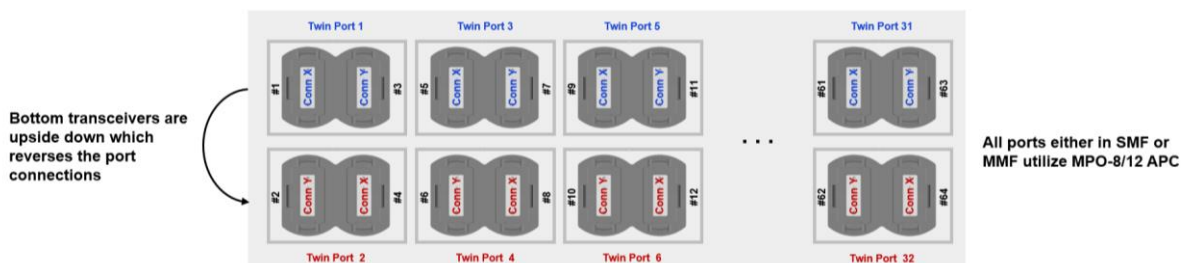


图 10. 双光模块端口注意事项

2.3. 了解SuperPOD的构建块——可扩展单元的概念

该系统建立在可扩展单元(SU)的构建块上,每个单元包含32个DGX H100系统,允许快速部署各种规模的系统。

每个SU包含256个GPU,分布在8个机架中的32个DGX H100系统上。每个可扩展单元(SU)有8个叶交换机。

该网络架构经过轨道优化,即来自每个节点的具有相同端口号的主机通道适配器(HCA)连接到同一个叶节点交换机。这意味着所有服务器的端口号1连接到叶节点交换机1,端口号2连接到叶节点交换机2,以此类推。

网络架构使用了Quantum-2 9700 InfiniBand交换机和800Gbps双OSFP光模块构建。

图11展示了基于以上信息的中间机架布局及其相关的计算网络组件。需要注意的是,还可以实现其他类型的布局。因此,在设计阶段与Corning工程团队保持密切联系至关重要,以验证您的设计选择的最佳布线方案。

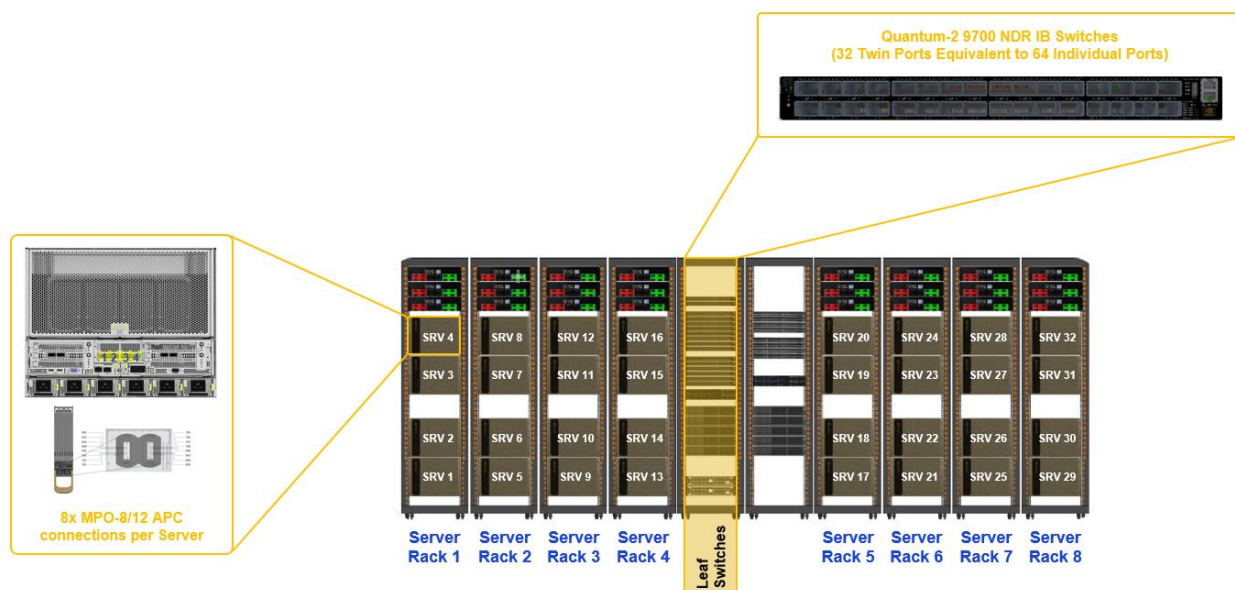


图11. 可扩展单元内的计算结构组件

2.4. 在NVIDIA DGX SuperPOD中实现布线场景

为了便于识别构建AI/ML集群时使用的不同布线组件, 康宁将在本指南中使用三个级别的连接。这些级别和交换机数量是基于32可扩展单元集群或POD的示例:

- A级-服务器到叶节点的布线;
- B级-叶节点到脊节点的布线;
- C级-脊节点到核心层的布线。

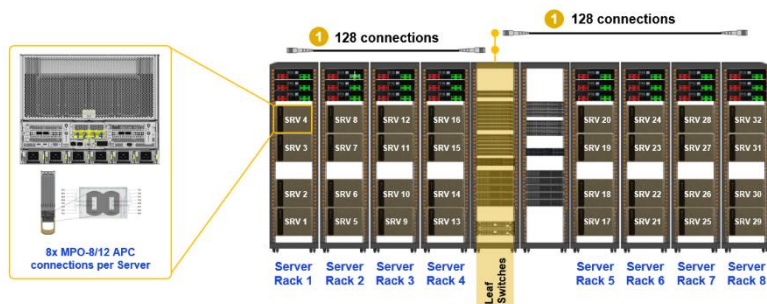
2.4.1 A级 -服务器到叶节点的布线

一个可扩展单元可以通过节点(服务器)和叶节点交换机之间的点对点连接来布线(见图1), 其中至少有两种布线产品选项可供选择(见图12)。在一些特定的定制设计中, 也可以在可扩展单元级别实施结构化布线(见图2)。

第一种方法是使用256根独立的8芯跳线来创建每个光模块的连接。第二种方法是使用8束跳线。束状跳线由一组8芯MPO-8/12 APC跳线组成, 作为一个256芯的单元, 并包裹有阻燃网套, 选择哪种方法将取决于客户设计的具体要求。无论选择哪种方法, 我们将总共连接256个MPO-8/12 APC端口到叶节点交换机。

在本例中, 交换机采用中间机架布局(见图12)。先前, 我们描述了计算网络架构为轨道优化设计, 共有 8条轨道。这些轨道对应于每台服务器的8个MPO-8/12 APC连接。在接下来的示例中, 我们将使用图12和图13中所示的颜色标记来指示每条轨道连接。

每个可扩展单元(SU)需要总共256条MPO-8/12 APC 8芯连接(每台服务器8条)至叶交换机



一个可扩展单元,多种方法:与客户合作共创成功未来

MPO-8/12 APC Jumpers
in SMF or MMF

1 SU = 256x 8F Individual
Jumpers



OR

MPO-8/12 APC Bundled Jumpers
in SMF or MMF

1 SU = 8x 256F Bundles

Bundled jumpers use a meshed sleeve



多轨架构

该网络结构经过多轨接入优化,意味着每个节点中具有相同端口号的主机通道适配器(HCA)将连接到同一个叶交换机。

也就是说,所有服务器的端口1连接将连接到叶交换机1,端口2连接到叶交换机2,以此类推。

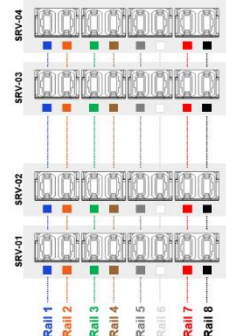


图12. A级-服务器到叶节点布线,描述连接数量和两种产品方案

对于A级布线,我们将使用束状跳线展示从服务器机架到位于可扩展单元中间机架(MOR)的InfiniBand叶节点交换机的连接。如果选择使用单根跳线,则可以按照轨道映射进行布线。

在图13的左侧,我们可以看到一个包含4台服务器的服务器机架,每台服务器对应其各自的轨道。在右侧,是叶节点交换机机架,内含8台叶节点交换机,每个交换机对应一个轨道。众所周知,这些Quantum-2交换机支持32个双MPO-8/12 APC端口,这意味着每台交换机可以支持多达64个独立连接。

每个可扩展单元(SU)需要总共256芯MPO-8/12 APC 8芯连接(每台服务器8条)至叶交换机

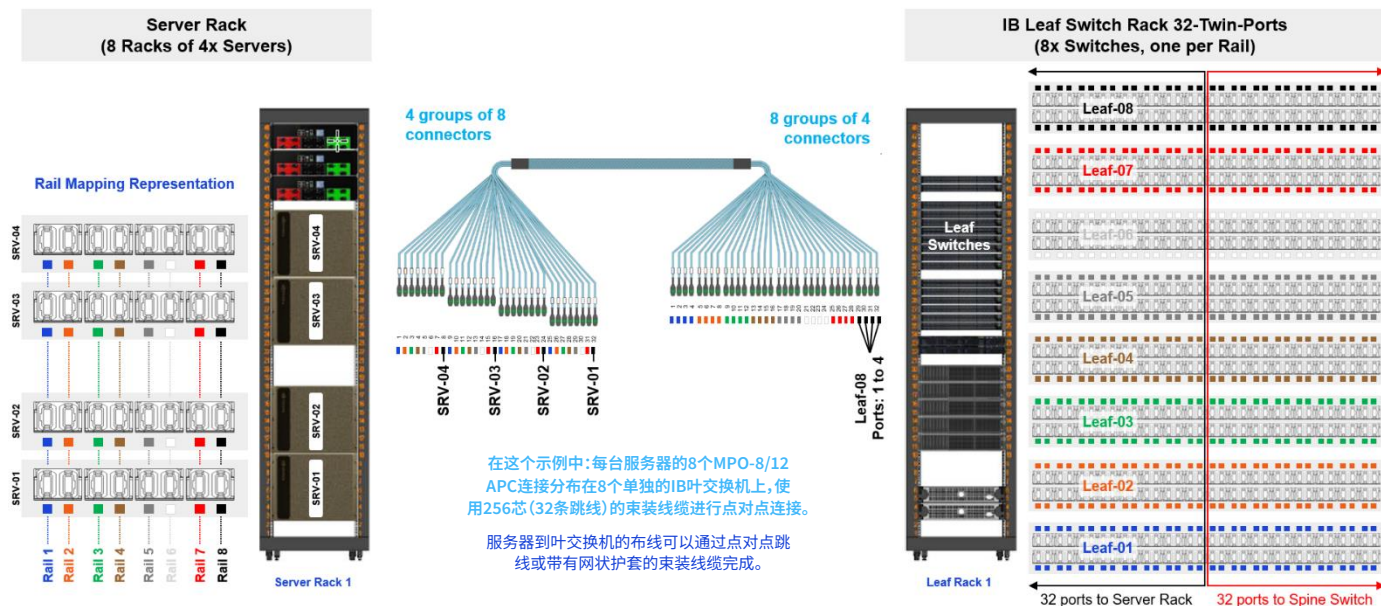


图13. A级-服务器到叶节点布线,使用束状跳线的点对点连接,基于32个可扩展单元群/POD的示例

布线通过每个服务器机架的一条束状跳线来执行,将每台服务器中相同颜色的所有轨道连接到相应颜色的叶节点交换机。在此示例中,每台服务器的轨道8(黑色轨道)被分配到叶节点08交换机的端口1至4。在一个可扩展单元中共有8个服务器机架,轨道8将按照图9所示的映射完成叶节点08的前32个端口连接,而剩余的32个端口连接将被路由到脊节点交换机。这个过程会对每个服务器机架和轨道重复,直到可扩展单元内的所有连接完成。

2.4.2 B级-叶节点到脊节点的布线

叶节点交换机和脊交换机之间的连接建议采用结构化布线方式。然而,在某些设计中,也可以使用点对点布线进行连接。需要注意的是,这些连接可以延长高达500米(基于800G-2xDR4, MMS4X00-NM收发器),这可能会给点对点配置的布线管理带来重大挑战。

正如架构部分开头所提到的,我们使用32可扩展单元集群或POD作为示例。参考图14所示的逻辑图和图7所示的组件数,我们可以观察到该设计共包含256个脊节点交换机,它们分布在16个脊节点机架上,说明每个机架包含16台脊节点交换机。

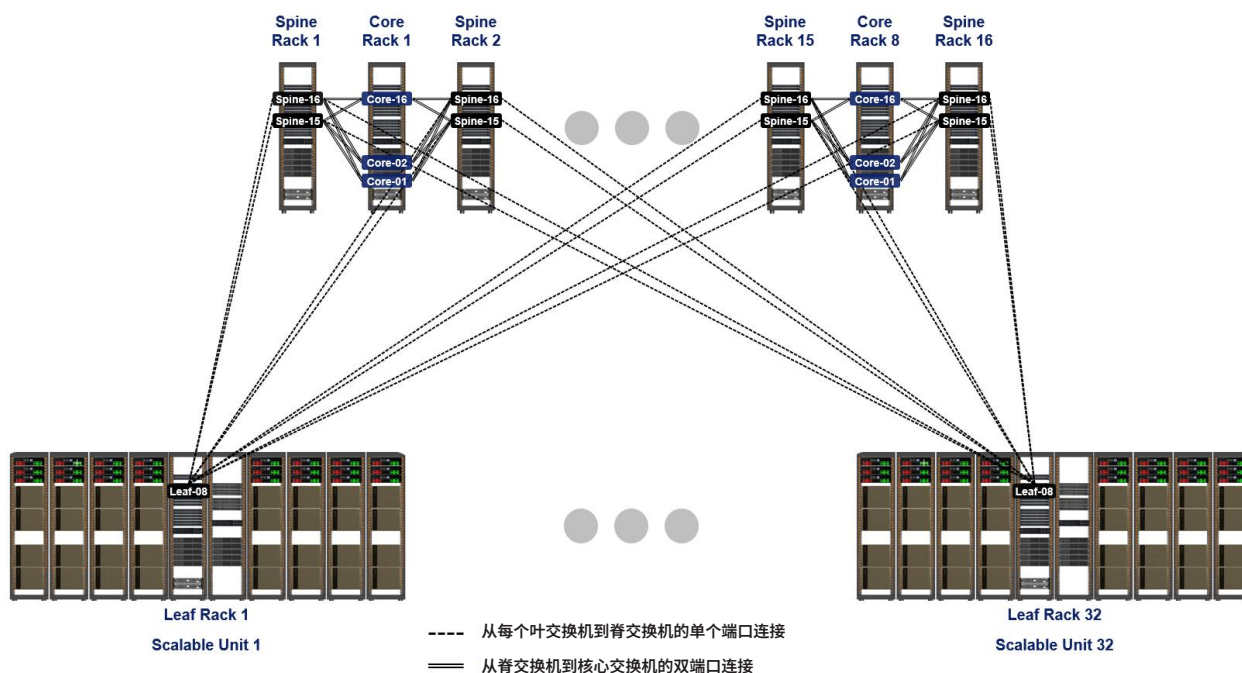


图14. 在32个可扩展单元的DGX H100 SuperPOD中, 轨道8 (黑色轨道) 的逻辑图

尽管这不是推荐的选择,我们还是先通过使用束状跳线进行点对点连接的布线示例来简化端口映射的说明。图15左侧显示了我们其中一个可扩展单元的叶节点交换机机架,其中每个交换机都标注了其对应的轨道颜色。右侧显示的是一个脊节点交换机机架(16个机架之一),包含16个脊节点交换机,每种轨道颜色对应两个脊节点交换机。

在这个示例中,我们以黑色轨道(轨道8)为参考。如果我们考虑从叶节点机架1中的叶节点交换机08(这是可扩展单元1中的叶节点机架)连接,并使用叶节点交换机的端口33和34(如图9中的端口映射所示),我们可以将它们路由到脊节点机架1,连接到脊节点15的端口1和脊节点16的端口1。由于同一轨道上的每个叶节点交换机都需要连接到同一轨道上的每个脊节点交换机(每个脊节点机架上有两个),所以来自不同可扩展单元的不同叶节点交换机的端口33和34也会连接到脊节点机架1,分别连接到脊节点交换机16的端口2和脊节点交换机15的端口2,依此类推。目标是建立所有可扩展单元中的叶节点交换机与所有脊节点机架中的两个脊节点交换机之间的连接,如图15所示。以轨道8为参考,这将需要从可扩展单元1中的叶节点交换机08使用16根束状跳线连接到16个不同的脊节点机架,从而连接到总共32个轨道8(黑色轨道)的脊节点交换机,如图14所示。



在图16左侧,我们可以看到我们其中一个可扩展单元的叶节点交换机机架,其中每个交换机都标注了其对应的轨道颜色。右侧显示的是脊节点机架1(16个机架之一),包含16个脊节点交换机,每种轨道颜色对应两个脊节点交换机。



仍然以黑色轨道(轨道8)为参考,如果我们从叶节点机架1中的叶节点-08交换机获取连接,并使用叶节点交换机的端口33和34(如之前所做),我们可以使用遵循相同映射的束状跳线进行路由。然而,将会添加结构化布线组件,如适配器面板和机箱,如图2所示。

目标是使用高芯数干线作为主干光缆,在不同区域之间进行高芯数的连接。我们将使用适配器面板将我们的连接从叶节点到脊节点进行映射。

在图16的示例中,我们可以看到来自叶节点-08的33和34端口的束状连接跳线进入位置A的适配器面板1号托盘。其中前两个适配器端口分配给黑色轨道。多个适配器面板随后使用主干布线将所有这些连接路由到数据中心另一部分的不同机箱,遵循相同的映射。最后,我们使用束状跳线完成到脊节点15的端口1和脊节点16的端口1的连接,结果与点对点布线相同。

如果我们遵循这种映射类型,可能会有一些位置(例如我们的配线面板中的位置E和F)保持未使用。然而,在行业中,即使这些备用适配器和连接在我们的映射示例中未必使用,利用备用位置仍然是一种常见做法。这些备用位置可以用来管理客户可能需要的备用连接,从而提供灵活性并满足其特定要求。

实施这种结构化布线的主要优点是最大化预留空间并减少安装时间。它还允许轻松进行移动、增加和更改(MAC),尤其是在叶节点交换机和脊节点交换机位于不同区域时。此外,它提供了未来扩展的灵活性,这是点对点布线无法支持的。

2.4.3 C级-脊节点到核心的布线

脊节点到核心交换机的布线可以通过点对点连接来实现。在一些特定的定制设计中,也可以在脊节点到核心连接级别实施结构化布线,这将取决于核心交换机相对于脊节点的物理位置。另外一个选项是使用直连铜缆(DAC)。

以32可扩展单元集群或POD为例,该设计包含总共128个核心交换机,分布在8个核心机架中,每个机架包含16个核心交换机。以图14中的逻辑图为例,为了举例说明,核心机架被放置在两个脊节点机架之间。

为了简化端口映射说明,我们使用束状跳线进行点对点连接。在图17中,左侧显示的是脊节点机架1,右侧显示的是脊节点机架2。正如我们所知,这些机架各自有16个脊节点交换机,每种轨道颜色对应两个交换机。在中间,我们有一个核心机架,包含16个核心交换机。每个核心交换机将聚合来自各种脊节点交换机的不同轨道。

以本示例为目的,并基于图9中描述的端口映射,我们假设脊节点机架2在物理上“镜像”脊节点机架1。这意味着连接到脊节点机架2中的叶节点交换机的端口是33到64,而不是1到32,连接到核心交换机的端口是1到32,而不是33到64。通过将核心交换机机架置于中间,我们可以通过束状跳线进行直接路由。

首先,关注来自脊节点机架1的连接。脊节点机架1中的所有脊节点交换机的端口33和34将被路由到核心交换机01的端口1到32;所有脊节点交换机的端口35和36将被路由到核心交换机02的端口1到32,依此类推,直到脊节点交换机的端口63和64被路由到核心交换机16的端口1到32。

接下来,我们需要处理来自脊节点机架2的连接。所有脊节点交换机的端口1和2将被路由到核心交换机01的端口33到64;所有脊节点交换机的端口3和4将被路由到核心交换机02的端口33到64,依此类推,直到脊节点交换机的端口31和32被路由到核心交换机16的端口33到64。

IB 脊交换机机架, 32个双端口
(16个交换机, 每个轨道配备两个)

IB核心交换机机架, 32个双端口 (16个交换机: 所有脊交换机的33/34端口连接到核心交换机#1; 所有脊交换机的35/36端口连接到核心交换机#2; ...所有脊交换机的63/64端口连接到核心交换机#16)。

IB 脊交换机机架, 32个双端口
(16个交换机, 每个轨道配备两个)

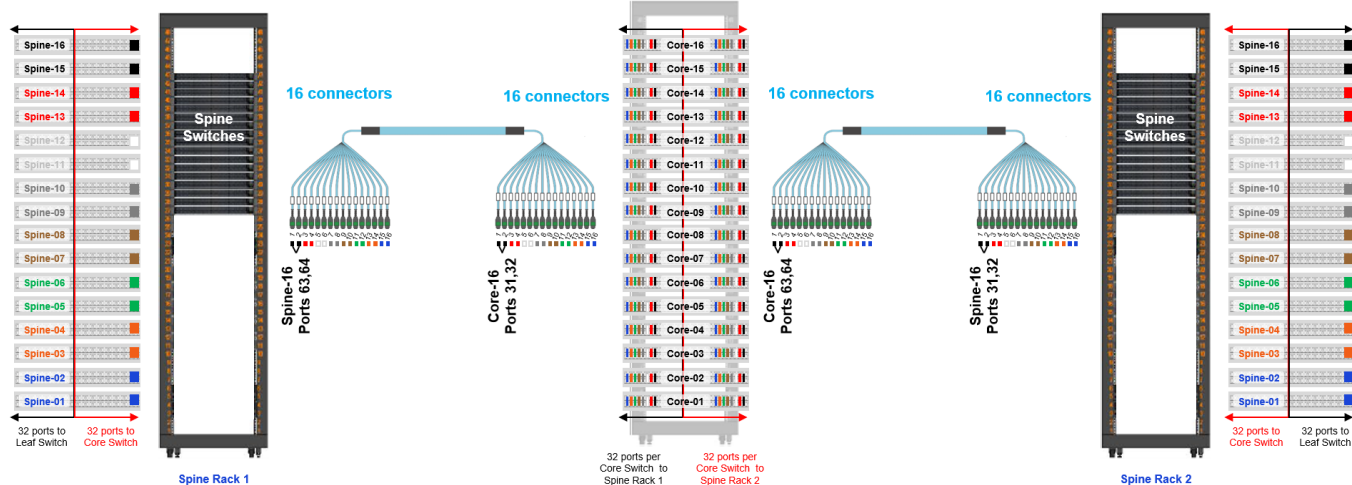


图17. C级 - 脊节点到核心布线, 使用束状跳线的点对点连接, 基于32可扩展单元集群/POD的示例

在图18中, 我们可以更详细地查看如何使用束状跳线进行布线。这显示了脊节点机架1, 其中所有脊节点交换机的端口33和34被路由到核心交换机Core-01的端口1到32。可以使用垂直线缆管理器 (在图18中仅作为示例呈现为机架) 来方便地进行从脊节点到核心的束状跳线路由。

Bundled Jumper Routing Example

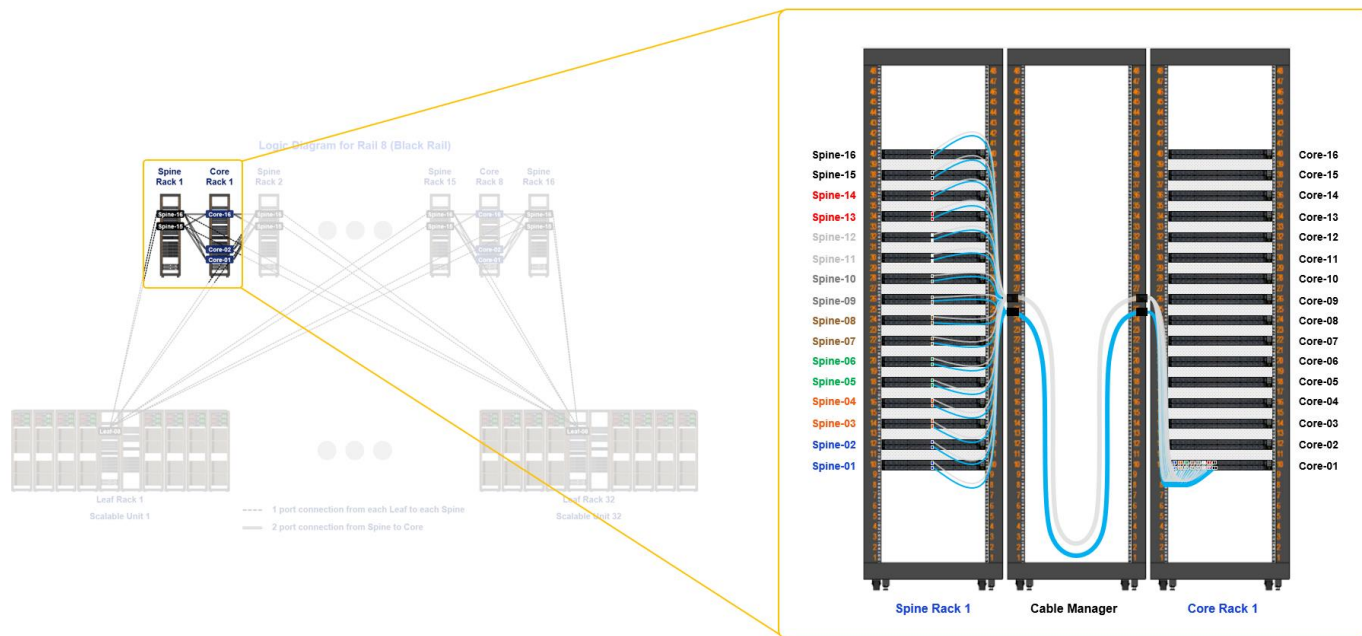


图18. C级 - 所有脊节点交换机的端口33和34路由到核心交换机Core-01的端口1至32, 基于32可扩展单元集群/POD的示例

2.5 多模 vs 单模

在网络中使用多模光纤还是单模光纤的决策将取决于具体的设计考虑。多模光纤的传输距离最多为50米,通常限制其应用于可扩展单元或脊节点到核心的连接。然而,这种情况并不总是实用,因为可扩展单元的设计可能不会遵循中间行布局,或者物理位置不相邻。同样,脊节点交换机和核心交换机在物理距离上可能也不接近,这使得整个设计中更倾向于使用单模光纤。

2.6. 整体概览

现在我们已经了解了不同的集群或POD尺寸,以及如何在计算布线系统的活动设备之间进行布线,让我们通过图示总结一下可以使用的组件。这些组件将取决于具体的设计,但主要基于我们在本文中回顾的不同产品和部件编号。

2.6.1. 连接4个可扩展

单元(SU)集群/POD的布线

在图19中,将4个SU POD与128台服务器、32台叶节点交换机和16台脊节点交换机进行布线需要:

- 1,024个从节点到叶节点的连接
- 1,024个从叶节点到脊节点的连接
- 如果有核心交换机的话,还需要1,024个连接(可选)

SU 数量	节点数量	GPU 数量	交换机数量			光路数量		
			InfiniBand 叶交换机	InfiniBand 脊交换机	InfiniBand 核心交换机	节点-叶节点	叶节点-脊节点	脊节点-核心
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** 服务器到叶交换机的布线可以使用点对点跳线或束装跳线组件
- B** 叶交换机到脊交换机的布线使用结构化布线
- C** 脊交换机到核心交换机的布线可以使用点对点跳线、束装跳线组件或直连铜缆(DAC)

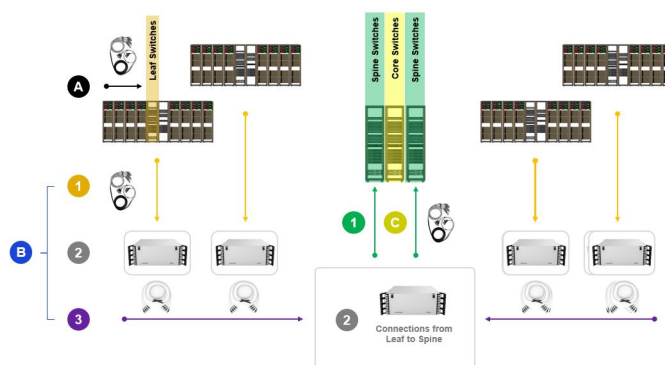
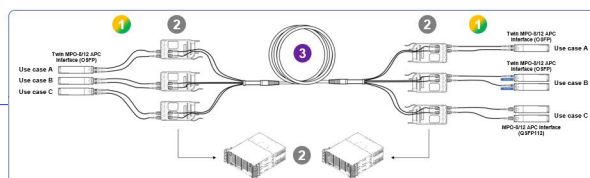


图19. 包含A级、B级、C级布线的4个可扩展单元(SU)集群/POD

通过应用我们已探讨的不同布线层级,可以得出以下结论:

- 数据中心运营商可以在可扩展单元中使用A级布线,进行服务器到叶节点的布线,使用点对点跳线或束状跳线组件。这将需要每个可扩展单元256个独立的8芯跳线(长度可变),或者仅需8个每个含256芯的束状跳线(长度可变)。
- 数据中心运营商可以使用B级布线,进行叶节点到脊节点的布线,采用结构化布线。这将需要安装单根跳线或束状跳线、适配器面板、配线架和高芯数主干光缆作为主干布线。通过这种主干布线,我们可以轻松使用64条主干光缆(每个叶节点16条主干光缆,每条144芯),这将替代点对点连接中的单根跳线和束状跳线。这种方法允许使用备用端口,更好地管理复杂性,并提高数据中心中的通道空间。
- 如果部署核心交换机,数据中心运营商可以使用C级布线,进行脊节点到核心的布线,使用点对点跳线、束状跳线组件或DAC(直连铜缆)。

2.6.2 连接8个可扩展单元(SU)集群/POD的布线

在图20中, 将8个SU POD与256台服务器、64台叶节点交换机和32台脊节点交换机进行布线需要:

- 2,048个从节点到叶节点的连接
- 2,048个从叶节点到脊节点的连接
- 如果有核心交换机的话, 还需要2,048个连接(可选)

SU 数量	节点数量	GPU 数量	交换机数量			光纜数量		
			InfiniBand 叶交换机	InfiniBand 脊交换机	InfiniBand 核心交换机	节点-叶节点	叶节点-脊节点	脊节点-核心
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 服务器到叶交换机的布线可以使用点对点跳线或束装线缆组件

B 叶交换机到脊交换机的布线使用结构化布线

C 脊交换机到核心交换机的布线可以使用点对点跳线、束装线缆组件或直连铜缆(DAC)

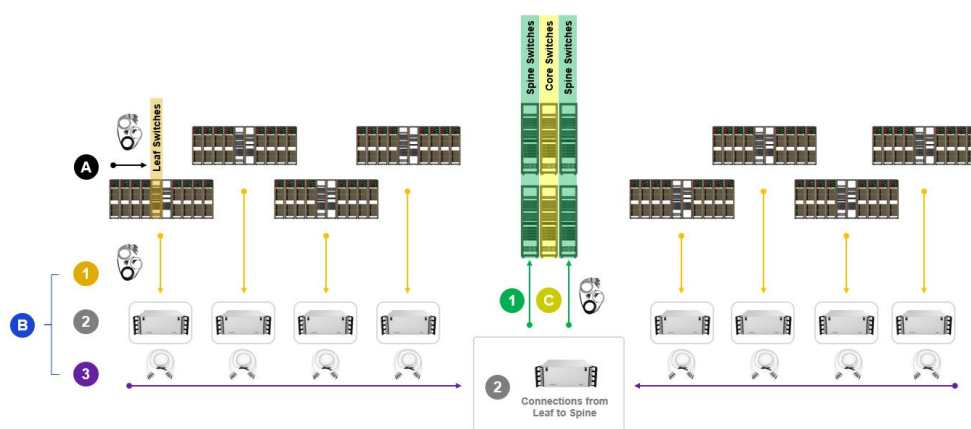
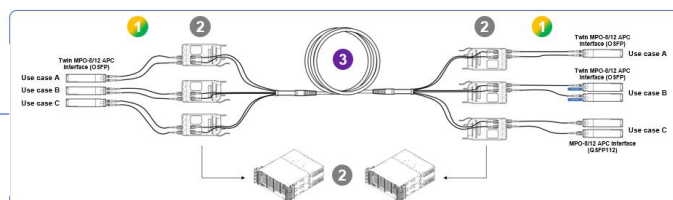


图20. 包含A级、B级、C级布线的8个可扩展单元(SU)集群/POD

通过应用我们已探讨的不同布线层级, 可以得出以下结论:

- 数据中心运营商可以在可扩展单元中使用A级布线, 进行服务器到叶节点的布线, 使用点对点跳线或束状跳线组件。这将需要每个可扩展单元256个独立的8芯跳线(长度可变), 或者仅需8个每个含256芯的束状跳线(长度可变)。
- 数据中心运营商可以使用B级布线, 进行叶节点到脊节点的布线, 采用结构化布线。这将需要安装单根跳线或束状跳线、适配器面板、配线架和高芯数主干光缆作为主干布线。通过这种主干布线, 我们可以轻松使用128条主干光缆(每个叶节点16条主干光缆, 每条144芯), 这将替代点对点连接中的单根跳线和束状跳线。这种方法允许使用备用端口, 更好地管理复杂性, 并提高数据中心中的通道空间。
- 如果部署核心交换机, 数据中心运营商可以使用C级布线, 进行脊节点到核心的布线, 使用点对点跳线、束状跳线组件或DAC(直连铜缆)。

2.6.3 连接16个可扩展单元(SU)集群/POD的布线

在图21中, 将16个SU POD与512台服务器、128台叶节点交换机和128台脊节点交换机进行布线需要:

- 4,096个从节点到叶节点的连接
- 4,096个从叶节点到脊节点的连接
- 如果有核心交换机的话, 还需要4,096个连接(可选)

SU 数量	节点数量	GPU 数量	交换机数量			光缆数量		
			InfiniBand 叶交换机	InfiniBand 脊交换机	InfiniBand 核心交换机	节点 叶节点	叶节点- 脊节点	脊节点- 核心
4	128	1024	32	16	--	1024	1024	1024
8	256	2048	64	32	--	2048	2048	2048
16	512	4096	128	128	64	4096	4096	4096
32	1024	8192	256	256	128	8192	8192	8192
64	2048	16384	512	512	256	16384	16384	16384

A 服务器到叶交换机的布线可以使用点对点跳线或束装线缆组件

B 叶交换机到脊交换机的布线使用结构化布线

C 脊交换机到核心交换机的布线可以使用点对点跳线、束装线缆组件或直连铜缆(DAC)

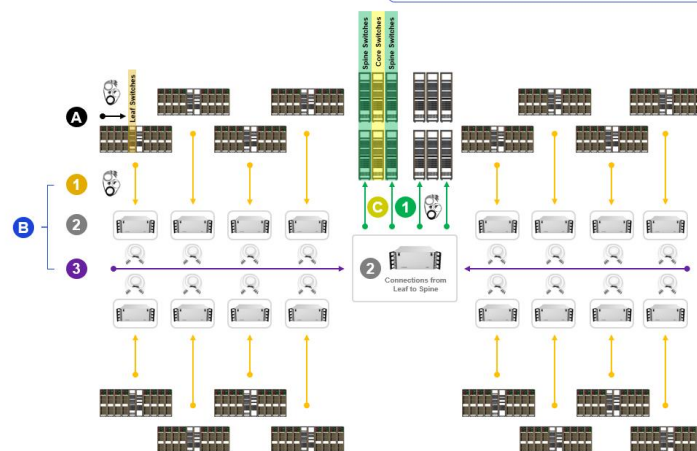
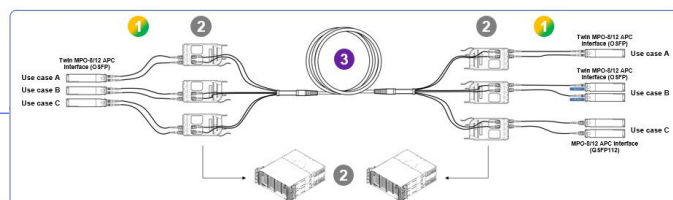


图 21. 包含A级、B级、C级布线的16个可扩展单元(SU)集群/POD

通过应用我们已探讨的不同布线层级, 可以得出以下结论:

- 数据中心运营商可以在可扩展单元中使用A级布线, 进行服务器到叶节点的布线, 使用点对点跳线或束状跳线组件。这将需要每个可扩展单元256个独立的8芯跳线(长度可变), 或者仅需8个每个含256芯的束状跳线(长度可变)。
- 数据中心运营商可以使用B级布线, 进行叶节点到脊节点的布线, 采用结构化布线。这将需要安装单根跳线或束状跳线、适配器面板、配线架和高芯数主干光缆作为主干布线。通过这种主干布线, 我们可以轻松使用256条主干光缆(每个叶节点16条主干光缆, 每条144芯), 这将替代点对点连接中的单根跳线和束状跳线。这种方法允许使用备用端口, 更好地管理复杂性, 并提高数据中心中的通道空间。
- 如果部署核心交换机, 数据中心运营商可以使用C级布线, 进行脊节点到核心的布线, 使用点对点跳线、束状跳线组件或DAC(直连铜缆)。

2.6.4 连接32个可扩展单元(SU)集群/POD的布线

在图22中, 将32个SU POD与1,024台服务器、256台叶节点交换机、256台脊节点交换机和128台核心交换机进行布线需要:

- 8,192个从节点到叶节点的连接
- 8,192个从叶节点到脊节点的连接
- 8,192个从脊节点到核心的连接

SU 数量	节点数量	GPU 数量	交换机数量			光缆数量		
			InfiniBand 叶交换机	InfiniBand 脊交换机	InfiniBand 核心交换机	节点-叶节点	叶节点-脊节点	脊节点-核心
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 服务器到叶交换机的布线可以使用点对点跳线或束装线缆组件

B 叶交换机到脊交换机的布线使用结构化布线

C 脊交换机到核心交换机的布线可以使用点对点跳线、束装线缆组件或直连铜缆 (DAC)

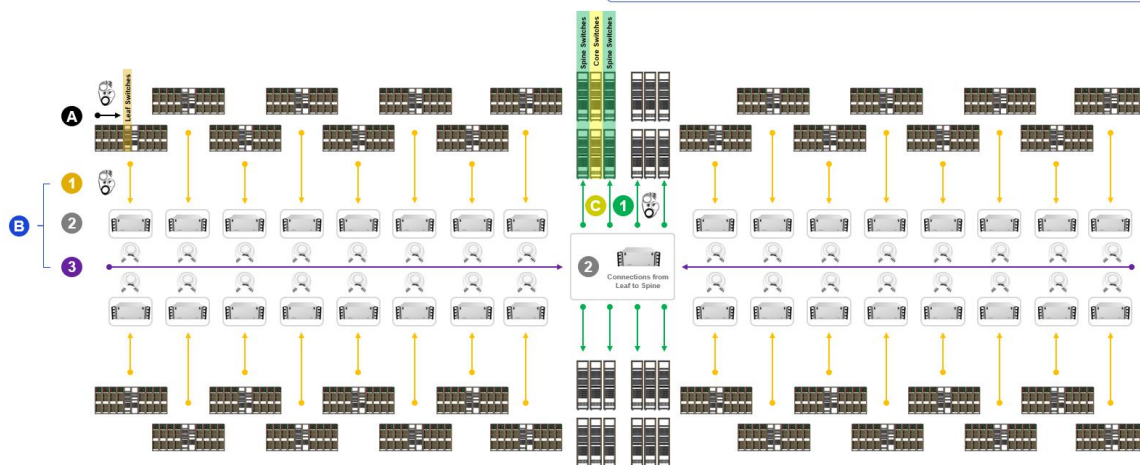
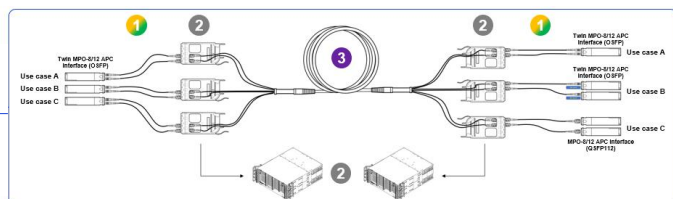


图 22. 包含A级、B级、C级布线的 32个可扩展单元 (SU) 集群/POD

通过应用我们已探讨的不同布线层级, 可以得出以下结论:

- 数据中心运营商可以在可扩展单元中使用A级布线, 进行服务器到叶节点的布线, 使用点对点跳线或束状跳线组件。这将需要每个可扩展单元256个独立的8芯跳线(长度可变), 或者仅需8个每个含256芯的束状跳线(长度可变)。
- 数据中心运营商可以使用B级布线, 进行叶节点到脊节点的布线, 采用结构化布线。这将需要安装单根跳线或束状跳线、适配器面板、配线架和高芯数主干光缆作为主干布线。通过这种主干布线, 我们可以轻松使用512条主干光缆(每个叶节点16条主干光缆, 每条144芯), 这将替代点对点连接中的单根跳线和束状跳线。这种方法允许使用备用端口, 更好地管理复杂性, 并提高数据中心中的通道空间。
- 如果部署核心交换机, 数据中心运营商可以使用C级布线, 进行脊节点到核心的布线, 使用点对点跳线、束状跳线组件或DAC (直连铜缆)。

2.6.5 连接64个可扩展单元(SU)集群/POD的布线

在图23中, 将64个SU POD与2,048台服务器、512台叶节点交换机、512台脊节点交换机和256台核心交换机进行布线需要:

- 16,384个从节点到叶节点的连接
- 16,384个从叶节点到脊节点的连接
- 16,384个从脊节点到核心的连接

SU 数量	节点数量	GPU 数量	交换机数量			光路数量		
			InfiniBand 叶交换机	InfiniBand 脊交换机	InfiniBand 核心交换机	节点-叶	叶-脊	脊-核心
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 服务器到叶交换机的布线可以使用点对点跳线或束状跳线组件

B 叶交换机到脊交换机的布线使用结构化布线

C 脊交换机到核心交换机的布线可以使用点对点跳线、束状跳线组件或直连铜缆 (DAC)

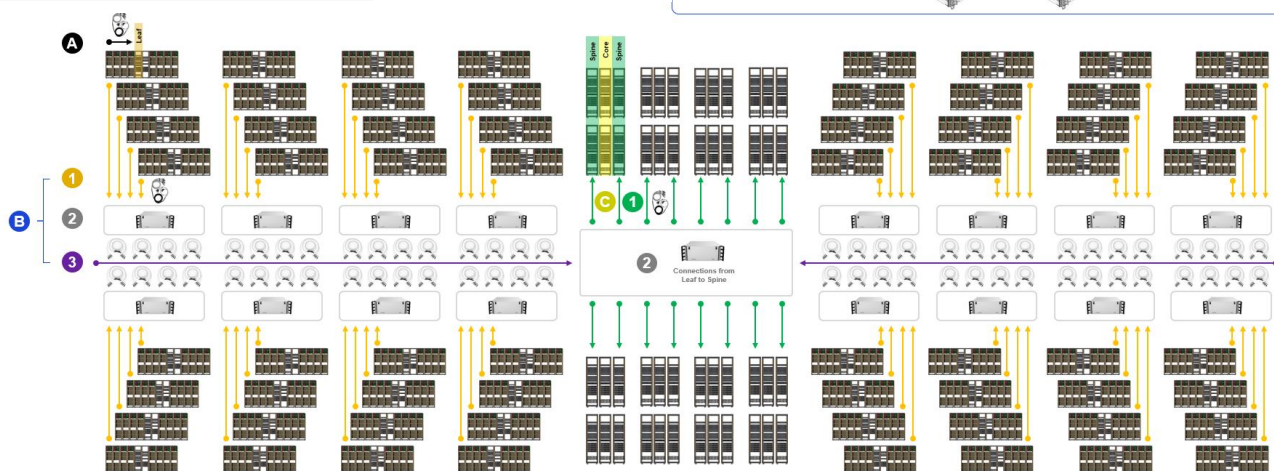
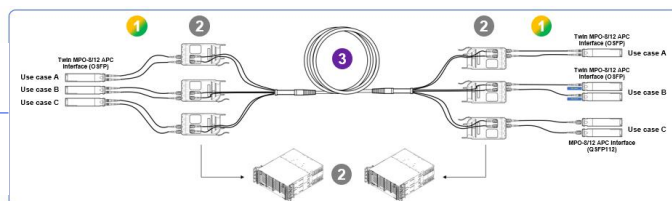


图23. 包含A级、B级、C级布线的64个可扩展单元 (SU) 集群/POD

通过应用我们已探讨的不同布线层级, 可以得出以下结论:

- 数据中心运营商可以在可扩展单元中使用A级布线, 进行服务器到叶节点的布线, 使用点对点跳线或束状跳线组件。这将需要每个可扩展单元256个独立的8芯跳线 (长度可变), 或者仅需8个每个含256芯的束状跳线 (长度可变)。
- 数据中心运营商可以使用B级布线, 进行叶节点到脊节点的布线, 采用结构化布线。这将需要安装单根跳线或束状跳线、适配器面板、配线架和高芯数主干光缆作为主干布线。通过这种主干布线, 我们可以轻松使用1,024条主干光缆 (每个叶节点16条主干光缆, 每条144芯), 这将替代点对点连接中的单根跳线和束状跳线。这种方法允许使用备用端口, 更好地管理复杂性, 并提高数据中心中的通道空间。
- 如果部署核心交换机, 数据中心运营商可以使用C级布线, 进行脊节点到核心的布线, 使用点对点跳线、束状跳线组件或DAC (直连铜缆)。

2.7 结论

总之, 了解每个级别(A、B和C)的详细布线要求对于优化DGX SuperPOD的部署至关重要。此外, 在可能的情况下实施结构化布线可以提高可管理性和效率, 特别是在大规模设置中。

在设计阶段与康宁工程团队合作, 确保布线策略与特定的数据中心布局和客户要求保持一致。

附件1 -高密度配线架

EDGE8®HD配线架安装在19英寸机架或机柜中,与EDGE8模块、面板、线束、中继器和跳线结合使用时,可提供业界领先的超高密度连接。

由于每个客户和项目都有特定的需求,请在BOM中添加最适合您需求的配线架:

	部件编号	最大模块/ 面板数	最大光纤密度		高度
	EDGE8-01U-SP	18	LC	144芯	1U
			MTP®	576芯	
	EDGE8-02U	36	LC	288芯	2U
			MTP®	1,152芯	
	EDGE8-04U	72	LC	576芯	3U
			MTP®	2,304芯	

表11. 高密度配线架

附件2 - 极性图纸

极性图, 通常被称为光纤极性图, 在使用光纤布线设计和实施数据中心链路时是必不可少的。它们在确保不同网络组件之间的适当连接、信号完整性和兼容性方面发挥着至关重要的作用。

本节将介绍适用于前面描述的每种场景的特定极性图。

场景1—800G和400G—服务器到交换机应用 MPO-8/12 APC到MPO-8/12 APC使用点对点布线

用例A:

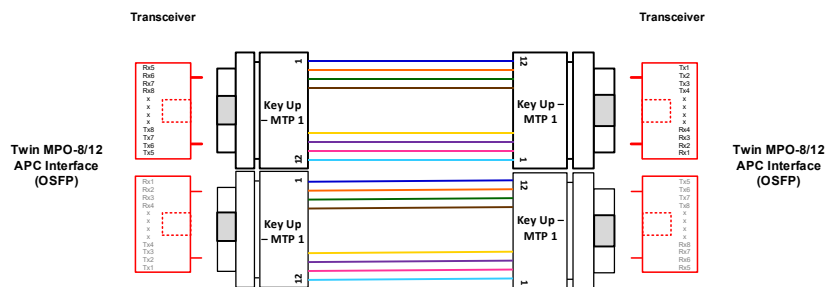


图24. 场景1-800G和400G-交换机到服务器本地 - 用例A

用例B:

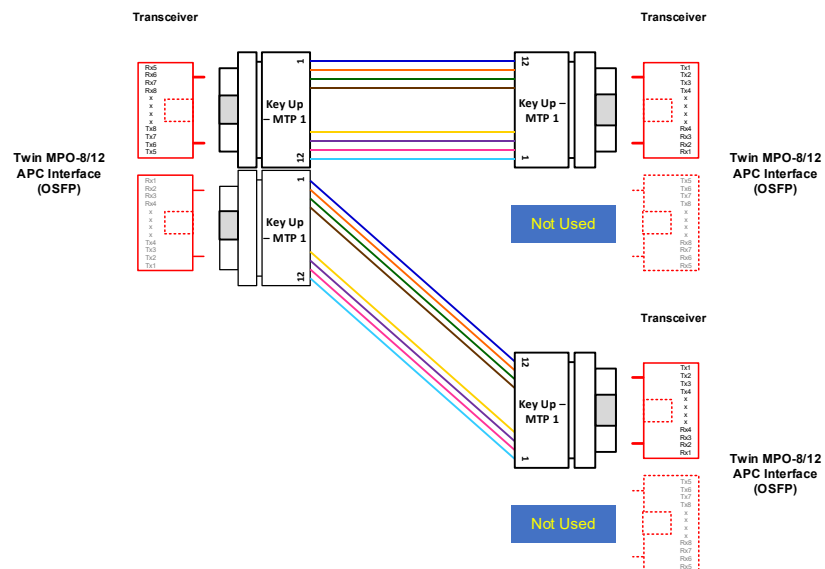


图25. 场景1-800G和400G-交换机到服务器本地 - 用例B

用例C:

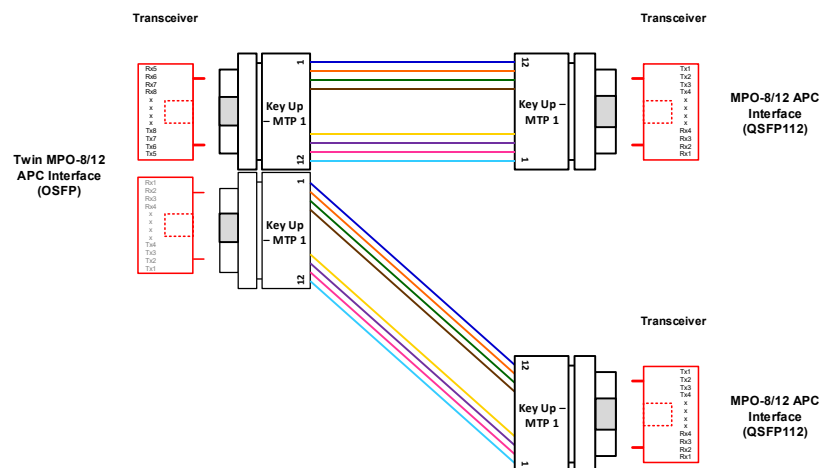


图26. 场景1 - 800G和400G - 交换机到服务器本地 - 用例C

场景2 - 800G和400G -交换机到交换机应用 MPO-8/12 APC到MPO-8/12 APC通过主干结构化布线跨数据中心连接

用例A:

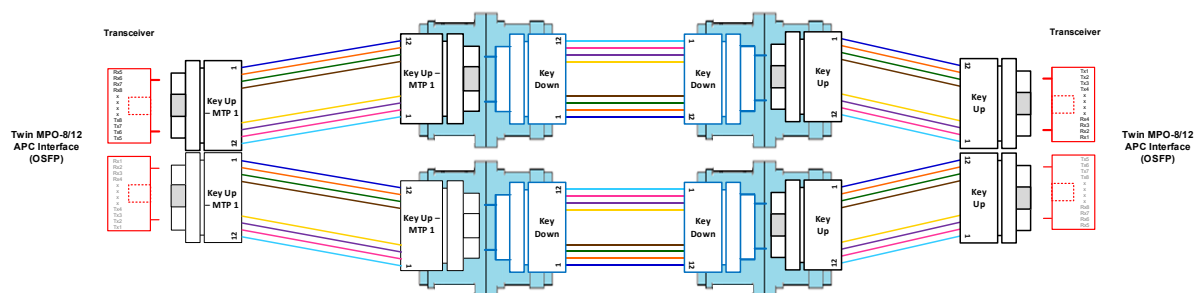


图27. 场景2 - 800G和400G -交换机到交换机通过主干跨数据中心连接 - 用例A

用例B:

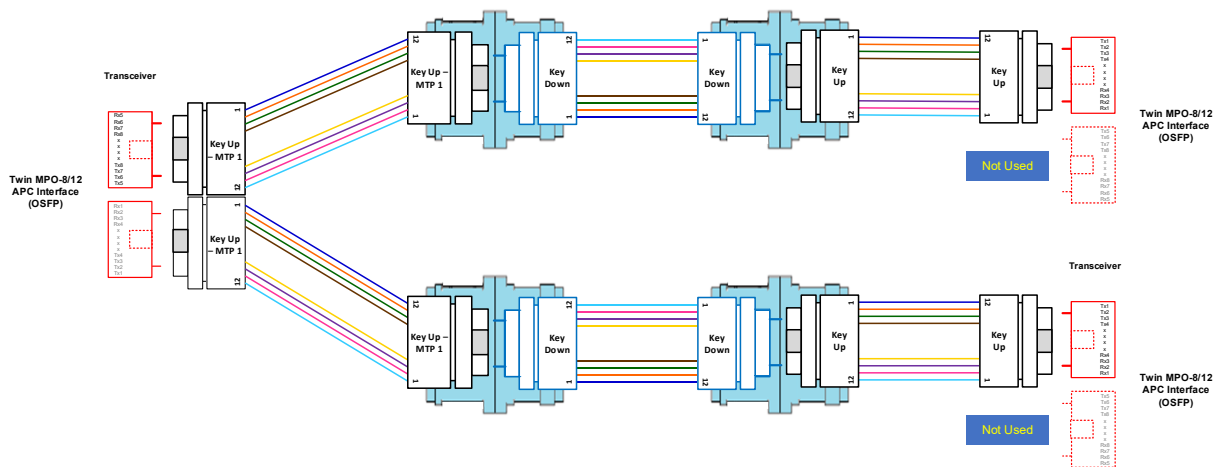


图28. 场景2 - 800G和400G -交换机到交换机通过主干跨数据中心连接 - 用例B

用例C:

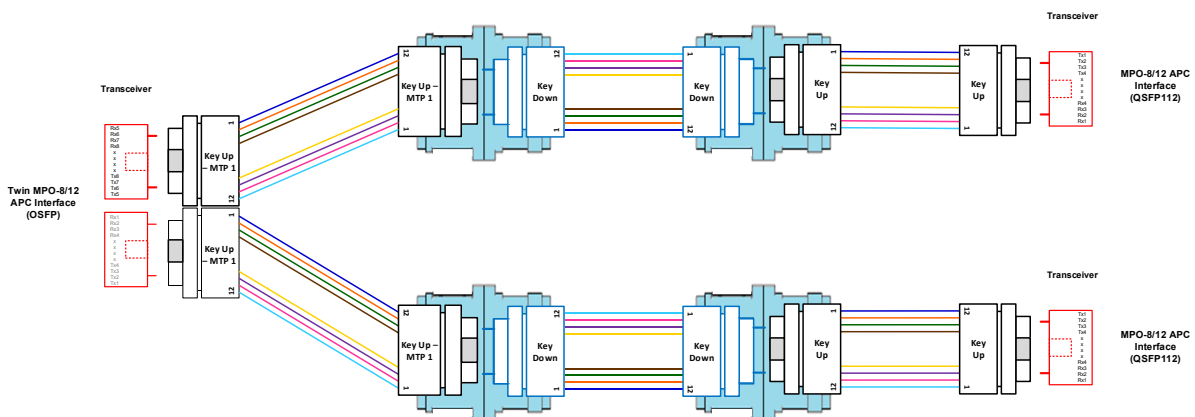


图 29. 场景2 - 800G和400G -交换机到交换机通过主干跨数据中心连接 - 用例C

场景3 - 800G和200G -服务器到交换机应用 MPO-8/12 APC到MPO-8/12 APC采用点对点布线

用例A:

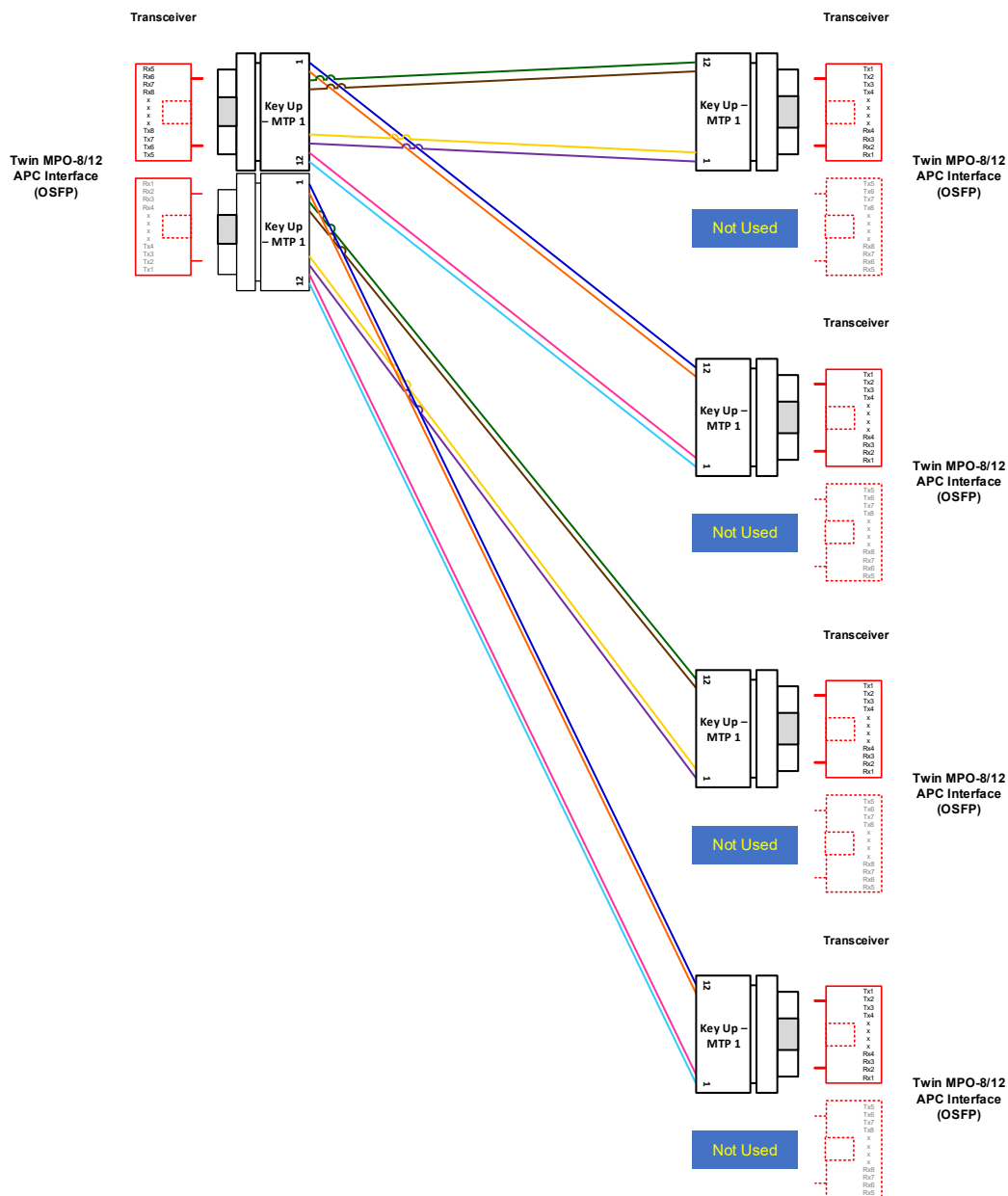


图30. 场景3 - 800G和200G -交换机到本地服务器-用例A

用例B:

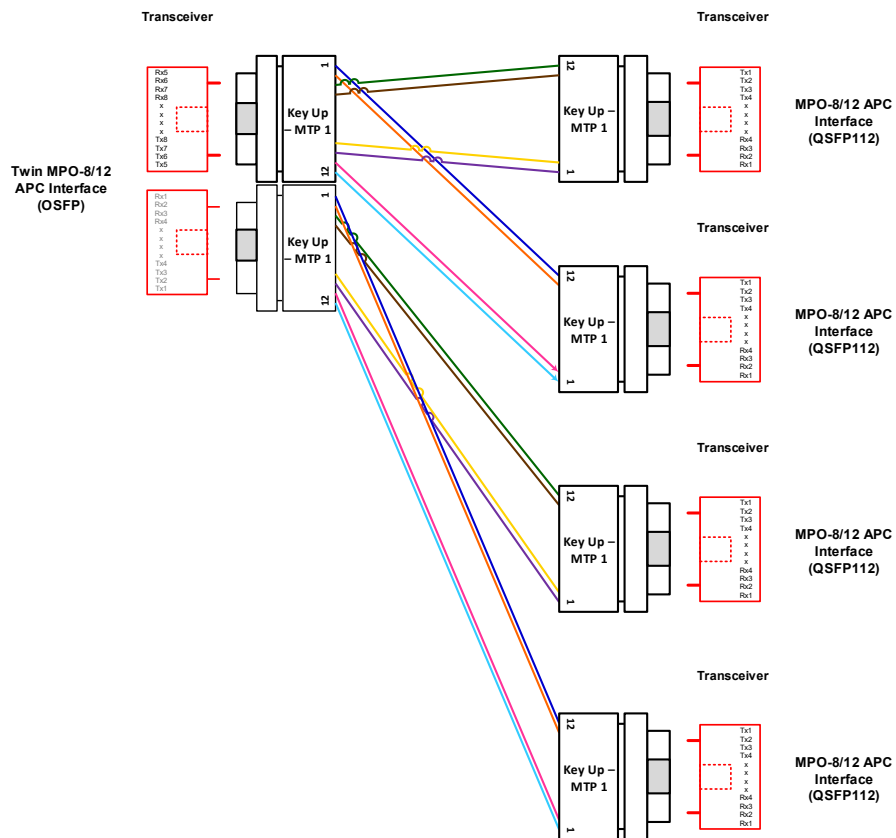


图31. 场景3 - 800G和200G - 交换机到本地服务器-用例B

场景4—800G和200G—交换机到交换机应用
MPO-8/12 APC到MPO-8/12 APC, 使用结构化布线跨数据中心连接

用例A:

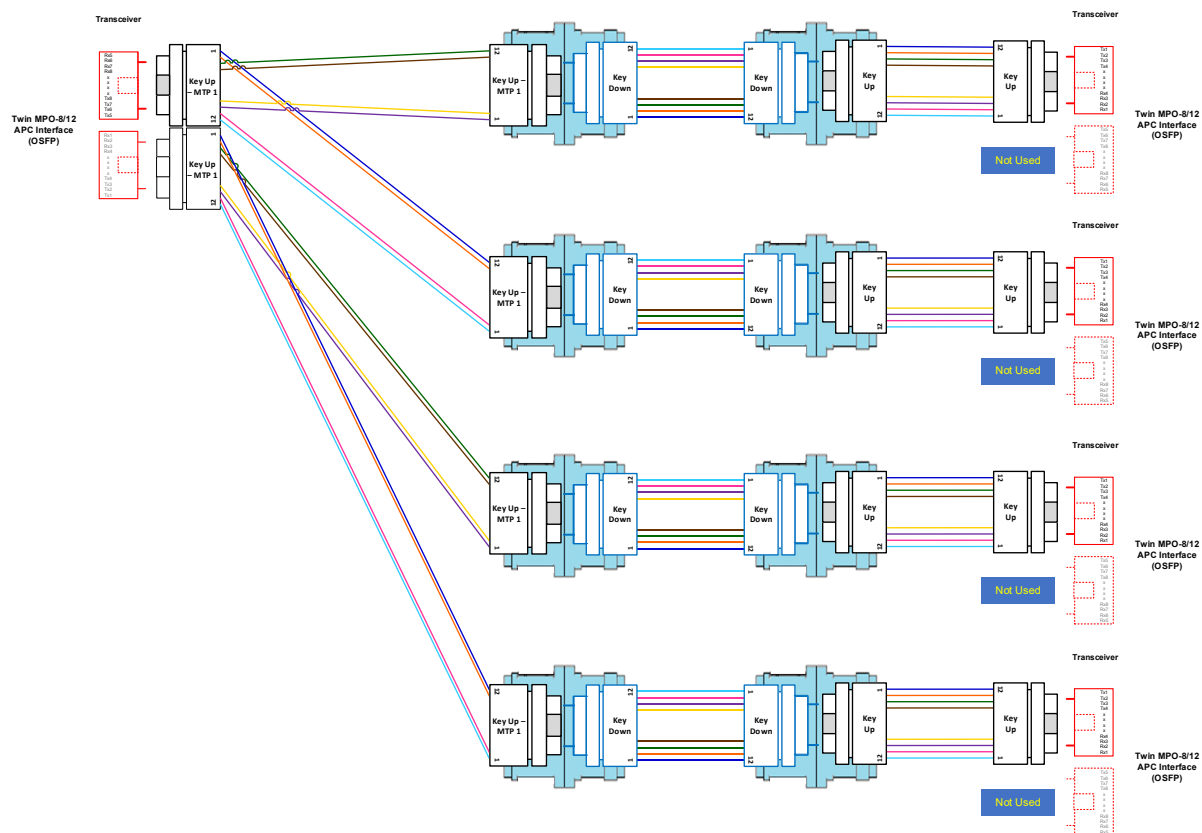


图32. 场景4 - 800G和200G - 交换机到交换机通过主干跨数据中心连接 - 用例A

用例B:

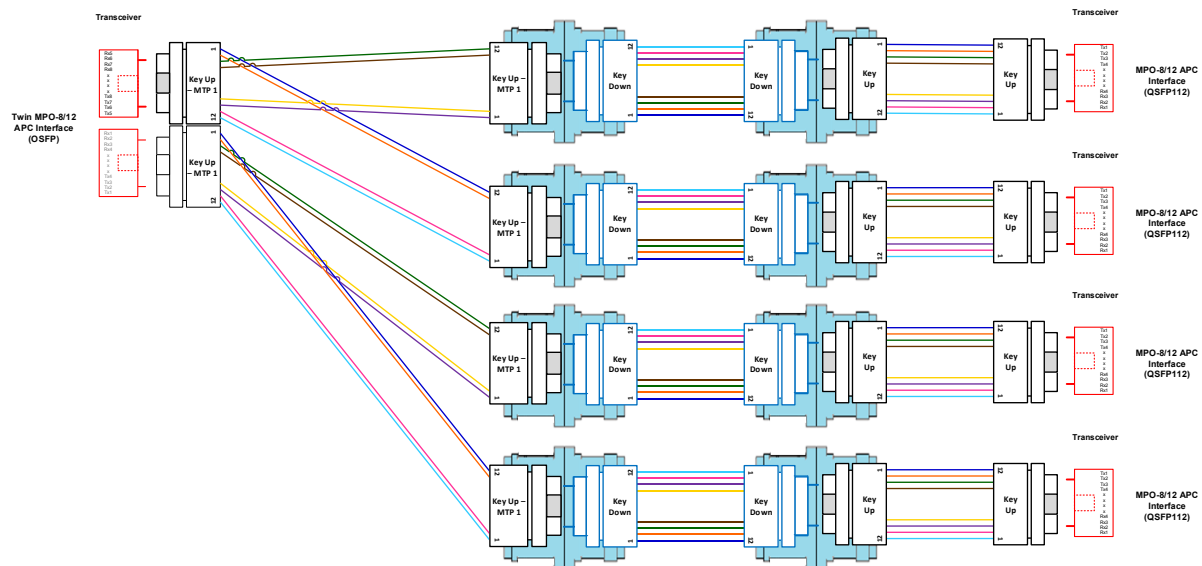


图33. 场景4 - 800G和200G - 交换机到交换机通过主干跨数据中心连接 - 用例B

情景 5 - 800G 和 400G - 交换机到交换机应用 LC-双工到 LC-双工使用点对点布线

用例A:



图34. 场景5 - 800G和400G - 交换机到本地交换机-用例A

用例B:

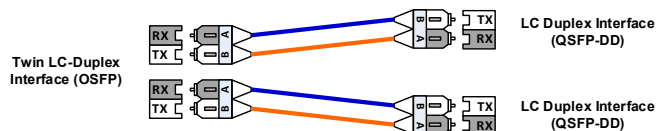


图35. 场景5 - 800G和400G - 交换机到本地交换机-用例B

场景6 - 800G和400G - 交换机到交换机应用

LC-双工UPC到 LC -双工UPC使用结构化布线通过数据中心的主干光缆连接

用例A:

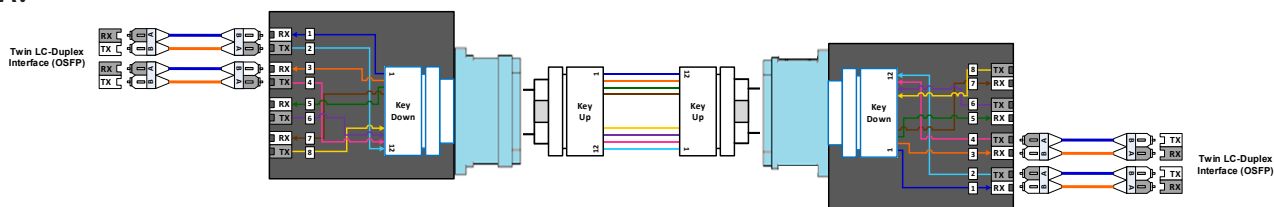


图36. 场景6 - 800G和400G - 交换机到交换机通过主干跨数据中心连接 - 用例A

用例B:

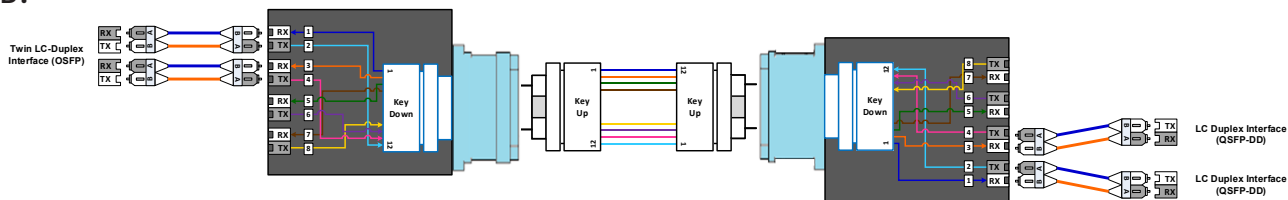


图37. 场景6 - 800G和400G - 交换机到交换机通过主干跨数据中心连接 - 用例B

附件3 -参考文献和联系人

本节包含NVIDIA概述白皮书的部分参考列表。有关NVIDIA产品的更多详细信息,请访问<https://docs.nvidia.com/>

收发器:

- MMS4X00-NM 800Gbps Twin-port OSFP 2x400Gb/s Single Mode 2xDR4, 500m
<https://docs.nvidia.com/networking/display/mms4x00nm800g500m/application+overview>
- MMS4X00-NS 800Gbps Twin-port OSFP 2x400Gb/s Single Mode 2xDR4, 100m
<https://docs.nvidia.com/networking/display/800gmms4x00ns/overview>
- MMA4Z00-NS 800Gb/s Twin-port OSFP, 2x400Gb/s Multimode 2xSR4, 50m
<https://docs.nvidia.com/networking/display/800gmma4z00ns/overview>
- MMS4X50-NM 800Gb/s Twin-port OSFP, 2x400Gb/s Single Mode 2xFR4, 2km
<https://docs.nvidia.com/networking/display/mms4x50nm800g2kmpub>

英伟达光缆:

- MFP7E30-Nxxx, 单模直交叉光纤线缆
<https://docs.nvidia.com/networking/display/mfp7e30nxxxpub/specifications>
- MFP7E40-Nxxx, 单模分路交叉光纤线缆
<https://docs.nvidia.com/networking/display/mfp7e40nxxxpub/specifications>
- MFP7E10-Nxxx, 多模直交叉光纤线缆
<https://docs.nvidia.com/networking/display/mfp7e10nxxx/specifications>
- MFP7E20-Nxxx, 多模分路交叉光纤线缆
<https://docs.nvidia.com/networking/display/mfp7e20nxxx/specifications>

DGX SuperPOD架构:

- 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/>

CORNING

康宁光通信中国·上海市漕河泾高科技开发区桂箐路111号立明大厦3楼 (200233)

电话 :+86 21 5450 4888 · 传真 :+86 21 5427 7898 · www.corning.com/opcomm/cn

康宁光通信保有改进、提高和修改康宁光通信产品的功能和规格的权利, 恕不另行通知。康宁光通信的完整商标列表可在www.corning.com/opcomm/trademarks上获得。
其他所有商标均为其各自所有者所有。康宁光通信通过了ISO 9001认证。© 2025康宁光通信版权所有 AEN185-A4-ZH / December 2025