社　区　规　范

224G AI RACK 高速互连研究报告

(版本 V0.5)

224G HIGH SPEED INTERCONNECTION FOR AI RACK RESEARCH REPORT (VERSION 0.5)

编委会作者:

北京百度网讯科技有限公司
北京抖音科技有限公司
京东云计算有限公司
北京快手科技有限公司
上海哔哩哔哩科技有限公司
行吟信息科技（上海）有限公司
天翼云科技有限公司
莫仕（中国）投资有限公司深圳分公司
华为技术有限公司
锐捷网络信息股份有限公司
富加宜连接器（上海）有限公司
东莞立讯技术有限公司
软通动力信息技术股份有限公司
庆虹电子（苏州）有限公司
中兴通讯股份有限公司
上海曦智科技有限公司
中航光电科技股份有限公司
杭州新华三通信有限公司
中研益企（北京）信息技术研究院有限公司

全球计算联盟 Open AI Infra 社区

开放规范许可证（类型 B)

本许可证是本开放规范（本规范）贡献者与其使用者之间的法律协议。一旦使用本规范，即视为同意遵守本许可证条款。若不同意遵守本许可证，请勿使用本规范，否则可能构成侵权。若文档以压缩包（如"zip"文件）形式分发且包含本许可证，则该压缩包中的所有文件均受本许可证约束，除非另有明确排除条款。

1．定义

1.1 本许可证：是指本开放规范许可证(类型A)。除全球计算联盟外，其他主体无权对本许可证进行修改。

1.2 本规范：是指在本许可证下许可的设计材料或数字代码等信息，不论其置于任何载体上或以任何方式表达，例如示意图、机械图纸、流程图及说明性文字、电路或电路板布线图、集成电路布图设计（包括拓扑图/掩模作品）、元器件排列图及其他说明材料等。

1.3 符合产品：是指完全符合本规范中就对应产品形态所规定的所有规范性要求的产品。

1.4 贡献者：是指以明示或默示方式签署贡献者许可协议（类型 A），并参与开发本规范的主体。

1.5 被许可方：是指遵守本许可证，并实施(a)使用、复制、修改或分发本规范中的任一行为，或实施(b)制造或委托他人制造或分发硬件中任一行为的主体。

1.6 分发：是指通过任何媒介向第三方传输、发布、提供或以其他方式传递本规范的行为。

1.7 修改：是指对本规范进行改造、改编、翻译、组合等。

2．知识产权许可和不诉承诺

2.1 著作权许可：在被许可方遵守本许可证所有条款和条件，特别是条款2.2的前提下，每个贡献者根据《贡献者许可协议（类型A）》授予被许可方永久的、全球性的、非独占的、免费的、免版税的著作权许可，以复制、创作衍生作品、公开展示、公开演示、再许可及分发本规范及其修改， 但仅限用于全球计算联盟的开放规范项目。

2.2 著作权许可条件

在遵从以下条件的前提下，被许可方才可以在任何媒介中复制、分发本规范和/或其修改：

2.2.1 被许可方必须向本规范的接收者提供本许可证的副本；

2.2.2 被许可方必须保留本规范的所有著作权、集成电路布图设计专有权、专利、商标及归属声明，但与本规范任何部分均无关的除外；

2.2.3 如被许可方对本规范进行修改并分发，被许可方必须向全球计算联盟提供修改说明及贡献修改后的文件。

3．其他

3.1 除上述第2条许可条款之外，本许可证不提供任何其他知识产权的授权。

3.2 免责声明：许可方按"原样"提供本规范，不提供任何明示或默示的担保，包括但不限于任何所有权、不侵权、质量合格、适销性、特定用途及适用性的担保。被许可方应自行对自身使用、复制、修改、分发本规范及其修改或制造、使用、销售、许诺销售、进口基于上述文件开发的产品的适当性进行判断，并全权承担与之相关的所有风险，以及行使本许可证授予被许可方的权利相关的其他所有风险。

3.3 责任限制：在任何情况下，无论基于任何法律依据，全球计算联盟及任何贡献者均不对被许可方因使用、复制、修改、分发规范及其修改或其中的贡献或制造、使用、销售、许诺销售、基于上述文件开发的产品而引发的任何损害，包括但不限于利润损失或任何形式的间接、附带、特殊或后果性损害赔偿责任，无论是基于违约、侵权（包括疏忽）或其他原因，也无论被损害方是否已被告知该等损害的可能性。

4．通用条款

4.1 如果被许可方违反本许可证，则被许可方基于本许可证所获得的所有许可和承诺将自动终止，被许可方应立即删除并停止使用本规范及其修改。本许可证终止后，第3.2、3.3款继续有效。如被许可方基于本许可证所获得的许可和承诺按本款规定终止的，不影响从被许可方处取得副本或权利的下游接收者在本许可证下的授权。

4.2 法律适用：本许可证适用中国法律。

版本变更日志

版本号	版本描述	贡献者
V0.5	初始版本	编委会各成员

目 次

224G AI RACK 高速互连研究报告 6
1. 第一章 224G AI Rack 概述 6
1.1 AI Rack 系统架构演进趋势分析 6
1.2 单 Rack、双 Rack 和多 Rack 系统架构比较； 8
1.3 AI Rack 高速接口标准介绍； 8
1.3.1 PCIe Based 高速接口标准 8
1.3.2 Ethernet Based 高速接口标准 8
1.3.3 NVLink 私有高速接口标准 9
1.3.4 UALink 开放高速接口标准 9
1.4 AI Rack Scale Up 主流网络拓扑介绍； 11
1.4.1 Full Mesh（Single-Tier Switched Fabric）拓扑 11
1.4.2 Clos / Fat-Tree（Multi-Tier Fabric）拓扑 11
1.4.3 nD Torus 拓扑 12
1.4.4 Dragonfly 拓扑 12
2. 第二章 224G SerDes 和 Retimer 技术 13
2.1 SerDes 工作原理与 PAM4 基本原理 13
2.1.1 Serializer 串行器 14
2.1.2 Deserializer 解串器 14
2.1.3 时钟恢复机制 14
2.1.4 PAM4 原理 15
2.1.5 PAM4 与 NRZ 比较 15
2.1.6 BER 与 FEC 要求 15
2.2 224G SerDes 架构与关键技术模块 16
2.2.1 发射机架构（TX） 16
2.2.2 接收机架构（RX） 16
2.2.3 DSP 架构 16
2.2.4 FFE 前馈均衡器 17
2.2.5 CTLE 连续时间线性均衡器 17
2.2.6 DFE 判决反馈均衡器 17
2.2.7 ADC DSP 18
2.2.8 CDR 时钟恢复器 18
2.3 224G 信道信号完整性要求与分析方法 19
2.3.1 Channel Budget 分析 20
2.3.2 COM 分析方法 21
2.3.4 112G 与 224G 信号完整性要求对比 22
2.4 Retimer 基础原理和发展背景 22
2.4.1 Retimer 工作原理 23
2.4.2 Retimer 架构 23
2.4.3 Redriver 与 Retimer 比较 23

2.4.4 Retimer 系统影响 24
3. 第三章 单/双 Rack 高速互连方案 24
3.1 AI Rack 互连形态 24
3.1.1 单宽单柜/单宽双柜/双宽单柜/双宽双柜 25
3.1.2 Cable Tray 架构/正交中板/正交无中板 26
3.2 GPU Tray 互连方案 27
3.2.1 GPU Tray 内部典型互连拓扑 28
3.2.2 GPU Tray 内部连接方式； 29
3.3 Switch Tray 互连方案 29
3.3.1 Switch Tray 内部典型互连拓扑 29
3.3.2 Switch Tray 内部连接方式 30
3.4 背板与背板线缆连接方案 31
3.4.1 非正交架构 32
3.4.1.1 背板方案（板对板，线对板，线对线） 32
3.4.1.2 中置背板方案（板对板，线对板，线对线） 33
3.4.2 正交架构 34
3.4.2.1 中置背板方案（板对板，线对板，线对线） 34
3.4.2.2 无中板方案（板对板，线对板，线对线） 35
3.4.3 混合正交和背板架构 36
3.4.3.1 中置 Cable 背板方案 36
3.4.3.2 有源 Cable 背板方案 37
3.5 224G 背板和 Cable 背板设计要求 38
3.5.1 信号完整性要求 38
3.5.2 PCB 材料及工艺要求 38
3.5.3 机械和结构设计要求 38
3.6 224G 系统其他要求 39
3.6.1 热设计； 39
3.6.2 维护性要求 40
3.7 高速连接器技术要求 41
3.7.1 接口密度发展趋势； 41
3.7.2 2048、4096、8192 Differential Pair 设计需求 42
3.7.3 高速连接器性能指标； 43
3.7.4 高速连接器组装工艺要求； 44
3.8 双 Rack 互连方案 45
3.8.1 DAC 应用 45
3.8.1.1 技术特点； 45
3.8.1.2 应用场景 46
3.8.1.3 设计要求 46
3.8.2 AEC 应用 46
3.8.2.1 技术特点 47
3.8.2.2 应用边界 47
3.8.2.3 设计要求 47
3.8.3 ACC 应用 48

3.8.3.1 技术特点 48
3.8.3.2 应用场景 48
3.8.3.3 设计要求 49
4. 第四章 多 Rack 高速互连方案 50
4.1 多 Rack 网络拓扑分析 50
4.2 NPO/CPO 技术路线 51
4.2.1 系统架构 51
4.2.1.1 外置光源：硅光架构的必要妥协 51
4.2.1.2 封装内、板级与面板级三层布局 52
4.2.2 应用优势和挑战 52
4.2.3 关键技术指标 53
4.3 高密度光互连 XPO/NG IO 技术路线 54
4.3.1 系统架构 55
4.3.2 应用优势和挑战 56
4.3.3 关键技术指标 57
4.4 OCS/OXC 技术路线 57
4.4.1 系统架构 58
4.4.1.1 OCS 在 TPU 集群中的实际用法 58
4.4.1.2 OCS 器件路线横向对比 59
4.4.2 应用优势和挑战 59
4.4.3 关键技术指标； 60
4.5 光 Shuffle 技术 61
4.5.1 系统架构 63
4.5.2 应用优势和挑战 63
4.5.3 关键技术指标 64
5. 第五章 产业生态与未来发展趋势 66
5.1 全球 224G 产业生态分析 66
5.1.1 主流 GPU 厂商路线 66
5.1.2 主流交换芯片路线 67
5.1.3 高速连接器发展趋势 68
5.1.4 光模块及光引擎发展趋势； 69
5.2 448G PAM6 和 PAM4 技术展望 69
5.3 AI Factory 互连架构演进展望 70

224G AI RACK 高速互连研究报告

1. 第一章 224G AI Rack 概述

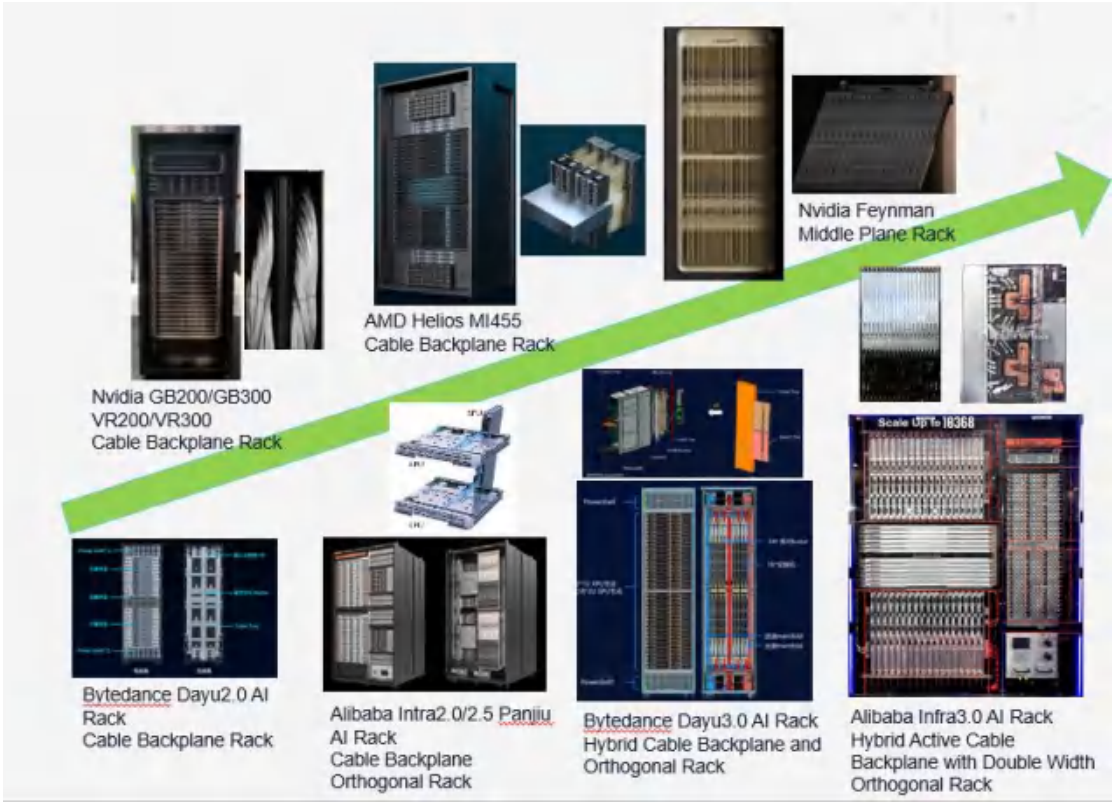
人工智能训练与推理工作负载持续推动数据中心由单机箱服务器级计算向机架级、集群级计算演进。传统服务器通常以 CPU 为计算和 I/O 的系统核心，GPU 作为辅助加速设备挂接在 CPU 周边；在高密度 AI 系统中，数十颗甚至上百颗 GPU 及通过高带宽互连组成计算域是系统核心，控制，网络、存储、供电和液冷等子系统围绕 GPU 计算域服务，机架级高速互连成为重点设计的对象，计算与互连资源的系统边界从单机箱延伸至机架及跨机架。

1.1 AI Rack 系统架构演进趋势分析

AI Rack 系统架构演进主要受三个因素驱动。其一，计算节点数量和模型并行程度提高，使加速器之间的数据交换成为系统效率的重要约束；其二，芯片 I/O Serdes 带宽与端口数量增长，使 PCB 走线距离、布线层数、信号损耗和热密度之间的冲突更为突出；其三，设备模块化、可维护性和规模化部署要求系统互连不仅追求电气性能，还必须兼顾装配公差、连接可靠性与服务操作空间。

国内外 112G 高速互连的 AI Rack 系统已经非常成熟并广泛应用，其主流的系统架构是采用铜缆互连的 Cable 背板架构或正交架构。

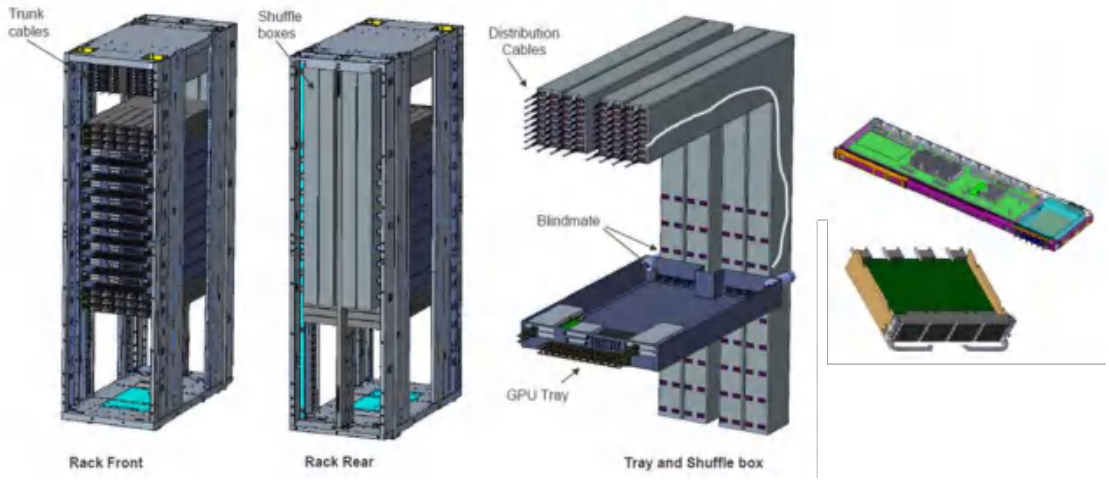
而 224G 高速互连的 AI Rack 系统伴随着 AI 模型规模的发展，GPU 集群规模的增加和 NPO 技术快速发展，基于铜缆互连的 Cable 背板架构或正交架构虽然仍会被广泛使用，但是基于 NPO 技术的集中式或分布式光互连架构已经成为更有发展潜力的新型架构。



基于铜缆互连的 Cable 背板架构和正交架构的 AI Rack 演进实例：

- 线缆背板超节点（Dayu2.0/NVL72 Oberon/Helios）
- 正交线缆背板超节点（Panjiu/Infra2.0/2.5）
- 中置背板超节点（Kyber NVL576）
- 正交中置线缆背板超节点（Dayu3.0）
- 正交后置双宽有源线缆背板超节点（Infra3.0）

基于光互连的 NPO，光 Shuffle 和 OCS 架构的 AI Rack 实例



1.2 单 Rack、双 Rack 和多 Rack 系统架构比较；

单 Rack 架构指同一计算域只有单个 Rack 内 GPU 通过实现互连，Rack 与 Rack 之间不组成同一计算域（Rack 与 Rack 之间没有 Scale Up 网络互连）；双 Rack 架构是同一计算域扩展到相邻的两个 Rack 的 GPU，相邻的两个 Rack 之间有高速 Scale Up 网络互连；多 Rack 架构指同一计算域扩展到多个 Rack 的 GPU，可以采用分层交换或将计算与交换功能分布于不同机架跨 Rack 布线等

维度	单 Rack	双 Rack	多 Rack
主要目标	机架内 GPU 集群	双倍 GPU 集群	多倍 GPU 集群
互连位置	Tray 内和 Tray 间	相邻的柜间	柜间、列间、网络层级间
技术选择	PCB 走线和线缆背板	DAC、AEC、ACC	以光互连为重要选项，短距铜互连仍可存在（<5m）

从互连设计看，单 Rack 通常更关注连接密度与机架内部线缆管理，主要采用铜互连；双 Rack 需综合评估被动铜缆可达性、主动均衡方案的功耗及线缆体积，也是主要材料铜互连；多 Rack 通常关注光纤布线、可插拔光模块、光引擎与网络分层，主要采用光互连，在 5m 互连长度范围内可采用 DAC、AEC、ACC 等铜互连方案。

1.3 AI Rack 高速接口标准介绍；

AI Rack 中的“高速接口”至少包含三个层次：系统协议及其链路层定义、SerDes/PHY 电气速率，以及承载该信号的物理互连组件。协议层的带宽数字不能直接转换为连接器的 224G 能力声明；不同标准对编码、纠错、训练、均衡及通道要求也不相同。

1.3.1 PCIe Based 高速接口标准

应用于 AI Rack Scale-Up 的 PCIe-based 高速接口主要包括 PCIe Gen6/Gen7/未来 Gen8、PCIe Switch Fabric、CXL 以及 PCIe 铜缆/光互连扩展。PCIe Gen6 采用 PAM4，速率 64 GT/s/lane，x16 理论带宽约 128 GB/s 单向、256 GB/s 双向；Gen7 提升至 128 GT/s/lane，x16 达 256 GB/s 单向、512 GB/s 双向，PCI-SIG 已将 AI/ML、HPC 和超大规模数据中心列为其重点应用，并支持向跨机架光互连扩展。 PCIe Gen8 目标进一步提升至 256 GT/s/lane，x16 双向理论带宽达到约 1 TB/s，但目前仍处于规范开发阶段。 在 Scale-Up 架构中，通过高端口数 PCIe Switch、GPU/xPU P2P 和多级交换，可将传统 Host-to-Device I/O 扩展成 GPU/xPU 间的开放式交换 Fabric；CXL 则基于 PCIe PHY 进一步提供内存扩展、池化和一致性能力。其核心优势是标准开放、生态成熟、兼容 CPU/GPU/NIC/SSD/CXL、成本较低且向后兼容。随着 Gen7/Gen8 以及铜缆、Retimer 和 Optical PCIe 的演进，PCIe 正从传统通用 I/O 向开放式 AI Rack Scale-Up Fabric 扩展。

1.3.2 Ethernet Based 高速接口标准

适用于 AI Rack Scale-Up 的 Ethernet-based 高速互连主要包括 SUE (Scale Up Ethernet) 和 ESUN（Ethernet for Scale Up Networking），底层复用高速 Ethernet PHY、SerDes、Switch、铜缆及光互连生态。SUE 1.0 面向 XPU 间的低延迟单跳交换，采用 200 Gb/s SerDes，典型单实例带宽 800 Gb/s，并可扩展至 1.6 Tb/s，支持低延迟、无损传输、信用流控及 XPU 内存事务，适合构建

GPU/XPU→Switch→GPU/XPU 的机架级紧耦合计算域。 ESUN 是 OCP 推动的开放 Scale-Up Ethernet 方向，重点针对 AI 加速器优化 Ethernet 的 L1/L2 及传输机制，兼顾低延迟、高带宽和跨厂商生态。物理层总体正从 100G/lane 向 200G/lane（224G-class SerDes） 演进，对应端口从 400G/800G 向 1.6T 升级。其核心优势是充分复用成熟的 Ethernet Switch、PHY、连接器、DAC/AEC 和光模块供应链，同时针对 AI Scale-Up 增强低延迟、Lossless、确定性及高带宽能力。相比专有的 NVLink，Ethernet-based Scale-Up 最大优势是开放生态、供应商多元化、成本与可扩展性好，并易于向铜缆和光互连演进。

1.3.3 NVLink 私有高速接口标准

NVLink 是 NVIDIA NVLink 面向 AI/HPC 加速计算设计的专有高速互连技术，核心目标是解决 GPU 之间传统 PCIe 带宽不足和通信延迟较高的问题，通过 NVLink + NVSwitch 构建高带宽、低延迟、全互联的 Scale-Up Fabric，使多颗 GPU 在服务器或整个 AI Rack 内形成紧耦合计算域。 NVLink 已从 GPU 点对点互连逐步演进为交换式机架级网络：第四代 NVLink 面向 Hopper，每颗 GPU 双向带宽最高 900 GB/s；第五代面向 Blackwell，提高到 1.8 TB/s，GB200/GB300 NVL72 可通过 NVSwitch 将 72 颗 GPU 构成全互联 Scale-Up 域，总聚合带宽约 130 TB/s；第六代面向 Vera Rubin，每 GPU 带宽进一步提升到 3 TB/s，Rubin NVL72 的 72 GPU 聚合带宽达到 216 TB/s。 NVSwitch 提供全线速 All-to-All GPU 通信，并集成 SHARP 网内归约和组播加速，特别适合 AllReduce、AllGather、Tensor Parallel、Expert Parallel 等通信密集型 AI 训练与推理。 NVLink 的主要优势是高带宽、低延迟、高 Bisection Bandwidth、GPU P2P 和与 NVIDIA GPU/CUDA 软件生态深度协同；主要局限则是其属于 NVIDIA 专有生态，相比 UALink、Ethernet-based Scale-Up 等开放标准，跨厂商通用性较弱。随着 NVLink Switch 和 NVLink Fusion 的发展，NVLink 正从节点内 GPU 互连持续向机架级和更大规模 AI Scale-Up 架构扩展

1.3.4 UALink 开放高速接口标准

UALink（Ultra Accelerator Link）是面向下一代 AI/HPC 的开放式 Scale-Up 高速互连标准，用于 GPU、NPU 等加速器之间以及加速器与 Scale-Up Switch 之间的高带宽、低延迟通信，是开放生态面向 NVLink 类专有互连的重要技术路线。UALink Consortium Specifications 显示，UALink 200G 1.0 支持单个 AI Computing Pod 内最多 1,024 个加速器互连。 UALink 1.0 有效数据速率为 200 Gb/s/lane，PHY 信令速率为 212.5 GT/s，支持 x1、x2、x4 Link，其中 x4 Station 可提供 800 Gb/s 单向、1.6 Tb/s 双向聚合带宽。 UALink 针对 Accelerator-to-Accelerator 通信优化，支持直接 Load、Store 和 Atomic 内存事务，强调低延迟、高带宽、确定性、低功耗和高带宽利用率，其设计目标可达到 93% 有效峰值带宽利用率。 相比 NVLink，UALink 的主要优势是标准开放、多供应商支持、可扩展性强，并可复用 Ethernet 产业链中的高速 SerDes、连接器、铜缆、Retimer 等基础技术，从而降低系统复杂度、功耗和总体拥有成本。 因此，UALink 是未来开放式 AI Rack/Pod Scale-Up Fabric 的重要技术方向。

4 类高速接口协议对比

维度	PCIe Based	Ethernet Based (SUE/ESUN)	NVLink / NVSwitch	UALink
定位	通用 I/O 向 Scale-Up Fabric 延伸	Ethernet 向 Scale-Up 优化	NVIDIA 原生 Scale-Up	开放原生 Scale-Up
典型 Lane Rate	Gen6 64 GT/s；Gen7 128 GT/s；Gen8 目标 256 GT/s	100G/200G/lane	专有 PHY，随代际演进	**200 Gb/s/lane**
典型端口/设备带宽	Gen7 x16 约 **512 GB/s 双向**	800G/1.6T 级端口	NVLink 5 **1.8 TB/s/GPU**	x4 Station **800 Gb/s 单向，1.6 Tb/s 双向**
延迟	低，但多级 PCIe Switch 增加延迟	较低，针对 Ethernet 进行 Scale-Up 优化	**极低**	**极低，专门为 Scale-Up 优化**
互连距离	PCB/机箱内为主，正向铜缆、光 PCIe 跨 Rack 发展	**最好**，铜缆可覆盖 Rack 级，光纤易扩展更远	Rack 内/紧邻系统 Scale-Up 为主	Rack/Pod 短距离 Scale-Up 为主
规模能力	中等，受 Switch 层级/radix 制约	**大，可充分利用 Ethernet Switch 生态**	当前成熟代表为 72 GPU NVL72	规范目标最高 1024 accelerators/Pod
内存语义	基础 PCIe 较弱，可结合 CXL 增强	SUE 等增强内存事务能力	**强，与 NVIDIA GPU 深度结合**	**强，支持 Load/Store/Atomic**
生态	**最成熟、开放**	**开放、供应链最广**	NVIDIA 专有	**开放、多厂商**

主要优势	兼容性、成本、成熟度	距离、开放性、Switch/光生态	**最高成熟度+高带宽+低延迟**	**开放+高性能+大规模**

1.4 AI Rack Scale Up 主流网络拓扑介绍；

AI Rack Scale-Up Network 是将多个 GPU/xPU 加速器通过超高带宽、超低延迟、低抖动且通常支持内存语义的互连 Fabric，组成一个紧耦合计算域，使多个物理加速器在系统和软件层面尽可能像一个更大的计算节点工作。其主要的特征是：High Bandwidth：超高带宽（GPU Memory Fabric 的带宽能力），Ultra-Low Latency：超低延迟，Lossless & Deterministic：无损、确定性；Memory Semantic：内存语义，Any-to-Any High Bandwidth：高 Bisection Bandwidth，Tight Coupling：把多个 xPU 形成一个计算域。

AI Rack Scale-Up Network 有多种拓扑结构，其主要有以下几类：

拓扑	结构	Switch	Scale-Up 适用性	典型方向
Switched Full Mesh / Single-Tier Fabric	GPU-Switch-GPU	有	★★★★★	Rack-scale Supernode
Clos / Fat Tree	多级 Switch	有	★★★★★	Scale-Up 跨柜扩展
Torus / Mesh	GPU 邻接互联	可无	★★★★	TPU/HPC 类架构
Dragonfly / Grouped Fabric	组内高带宽+组间连接	有	★★★	超大规模 Scale-Up/Scale-Out 融合

1.4.1 Full Mesh（Single-Tier Switched Fabric）拓扑

Full Mesh（Single-Tier Fabric） 是一种单级交换式 Scale-Up 网络拓扑，所有 GPU/xPU 通过一层 Scale-Up Switch Fabric 互联，使任意两个加速器之间都能够通过交换网络进行通信，从而形成逻辑上的全互联计算域。通过单级 Switch Fabric 为所有 xPU 提供 Any-to-Any、接近全带宽、低延迟的通信能力，同时避免 Direct Full Mesh 随 xPU 数量增加产生的端口和链路数量爆炸。每个 GPU 不需要直接连接其他所有 GPU，而是连接一组 Scale-Up Switch。

1.4.2 Clos / Fat-Tree（Multi-Tier Fabric）拓扑

Clos / Fat-Tree Scale-Up Network 是一种采用两级或多级交换结构构建的 Scale-Up Fabric，通过 Leaf 与 Spine 等多级 Scale-Up Switch 将大量 GPU/xPU 连接为统一的紧耦合计算域，并利用多路径和高割面带宽实现大规模 Any-to-Any 加速器通信。Single-Tier 受到 Switch Radix 限制，Switch 端口数

量最终成为扩展瓶颈。

Clos / Fat-Tree Scale-Up：通过多级 Scale-Up Switch 和多路径互连突破单级 Switch Radix 限制，在保持高 Bisection Bandwidth 的同时，将紧耦合 xPU 计算域从 Rack Scale 扩展到 Multi-Rack Scale。

1.4.3 nD Torus 拓扑

nD Torus Scale-Up Network 是一种将 xPU 按 n 维坐标排列，每个 xPU 与各维度正、负方向上的邻居直接连接，并在每个维度首尾相连形成闭环的分布式 Scale-Up 拓扑。不依赖中心 Scale-Up Switch 实现 Any-to-Any，而是让 xPU 本身通过固定数量的邻居连接组成分布式 Fabric。其主要特性是：单颗 xPU 的拓扑邻接度不需要随总节点数线性增加（Fixed Node Degree），不需要大型中心化 Scale-Up Switch，Local 通信效率高。Torus 牺牲部分 Any-to-Any 通信的最小 Hop 能力，换取更低的节点 Radix、更规则的连接以及更好的规模扩展性。

nD Torus Scale-Up Network：将 xPU 组织为 n 维周期性邻接网络，每个 xPU 仅与各维度相邻节点直接互连，并通过多跳路由实现整个 Scale-Up Domain 内通信，从而以固定节点连接度获得良好的规模扩展能力。

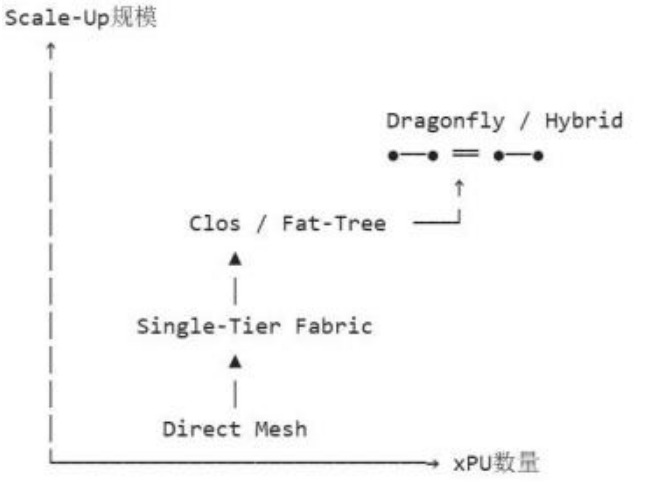
1.4.4 Dragonfly 拓扑

Dragonfly Scale-Up Network 是一种面向大规模紧耦合加速计算的分组式低直径网络拓扑，将 GPU/xPU 及 Scale-Up Switch 划分为多个 Group，Group 内采用高带宽、高连接度互连，Group 之间通过相对较少的 Global Link 直接互联，从而以较少的交换层级和全局链路实现大规模 Any-to-Any Scale-Up 通信。经典 Dragonfly 结构中，每个路由器连接终端节点，同时通过 Local Channel 连接 Group 内其他路由器，并通过 Global Channel 连接其他 Group。Dragonfly 通过“Group 化 + Global Direct Links”获得 Scalability。Group 内部使用高密度铜互连/短距光互连，Group 之间采用光互连。

4 种 Scale Up 网络拓扑对比

维度	Switched Full Mesh / Single-Tier	Clos / Fat-Tree	nD Torus	Dragonfly
基本结构	xPU→ 一层 Switch Fabric→xPU	多级 Switch Fabric	xPU/节点按多维网格邻接	Group 内高密互联+Group 间 Global Link
Switch 层级	1 级	2 级或更多	可不依赖集中式 Switch	分层/分组
网络跳数	最低	中等	随节点位置变化	较低
Bisection BW	极高，易做无阻塞	高，可构建无阻塞	中等	高，但依赖 Global Link 配置
延迟	最低、最确定	低～中	多跳，变化较大	较低

扩展性	中等，受 Switch Radix 限制	高	高	非常高
布线	集中、数量大	大量分层布线	规则但较复杂	Global Link 复杂
成本/功耗	中等	较高	较低	相对较低
路由复杂度	低	中	中~高	高
AI Scale-Up 适合度	★★★★★	★★★★★	★★★☆☆	★★★★☆
更适合	单 Rack/较小 Supernode	大型/跨 Rack Supernode	专用 ASIC/HPC	超大规模 Scale-Up/融合网络



2. 第二章 224G SerDes 和 Retimer 技术

2.1 SerDes 工作原理与 PAM4 基本原理

SerDes（Serializer/Deserializer，串行器/解串器）是高速芯片间互连的核心物理层技术，其作用是将芯片内部多路低速并行数据转换为少量高速串行信号，经 PCB、连接器、铜缆或光模块传输后，再在接收端恢复成并行数据。完整 SerDes 链路通常包括并串转换、编码与扰码、发射端驱动与 FFE 预加重、物理通道、接收端 CTLE/FFE/DFE 均衡、时钟数据恢复、采样判决、解码以及 FEC。高速信号经过封装、过孔、PCB、连接器和线缆时，会受到插入损耗、反射、串扰、抖动及码间干扰影响，因此 SerDes 并非简单的串并转换器，而是一套具备信号补偿、时钟恢复和链路训练能力的高速收发系统。当前 112G 和 224G 级接口普遍采用 PAM4 调制，即使用四个幅度电平表示两比特数据，使相同符号率下的数据率达到 NRZ 的两倍，从而降低对通道带宽的要求。代价是 PAM4 眼图由一只眼变成三只眼，相邻电平间距缩小，噪声裕量和线性度要求明显提高，需要更复杂的均衡、CDR 及 FEC 协同工作。OIF CEI 演示资料显示，CEI-112G 及 CEI-224G 均以 PAM4 作为主要调制方式，224G 接口实际测试信令速

率通常约为 212.5 Gb/s。

2.1.1 Serializer 串行器

Serializer 是 SerDes 发送端的并串转换模块，负责将芯片内部较宽、较低速的并行数据流转换成高速串行比特流。例如，交换 ASIC 内部可能以数百位宽的总线在较低时钟频率下处理数据，而外部高速接口只使用一对差分线传输。Serializer 通常采用多级复用结构，通过 2:1、4:1 等 Multiplexer 逐级提高数据速率，并利用 PLL 产生高速低抖动时钟。进入 PAM4 系统后，发送端还需要将两个比特映射成一个四电平符号，通常采用 Gray Coding，使相邻电平之间仅有一个比特不同，从而降低单次错误造成多个比特错误的概率。随后，DAC 或分段式电流驱动器将数字符号转换为四个模拟电平。为补偿通道高频损耗，发送端通常配置多抽头 FFE，通过 Pre-cursor 和 Post-cursor 预加重，提前抵消 PCB、连接器及线缆产生的码间干扰。Serializer 性能主要由输出抖动、上升/下降时间、差分摆幅、线性度、RLM、SNDR 和 FFE 能力衡量。在 224G 链路中，DAC 线性度、时钟抖动和封装寄生会直接压缩 PAM4 三只眼的裕量，因此必须将 ASIC 封装、PCB breakout、连接器和线缆作为完整通道协同优化。下一代高速 SerDes 发送端需要采用先进 DAC 架构、模拟前端及校准算法

2.1.2 Deserializer 解串器

Deserializer 位于 SerDes 接收端，负责从衰减且失真的高速串行波形中恢复数据，并将其转换回芯片内部并行格式。信号进入接收端后，首先经过终端匹配和模拟前端，再由 CTLE 补偿通道的高频衰减。对于损耗较大的链路，还会使用接收端 FFE、DFE 或 MLSD 消除前游和后游码间干扰。PAM4 接收机需要设置三个判决门限，将接收到的模拟幅度划分为四个电平，因此其判决电路、ADC 分辨率、线性度和噪声性能要求明显高于 NRZ。接收端同时通过 CDR 从数据跳变中恢复采样时钟，并在最佳相位对波形采样。判决后的 PAM4 符号再经过 Gray 解码、解扰、FEC 解码和齿轮箱处理，恢复成内部并行数据。现代高速 SerDes 接收端一般采用模拟判决式或 ADC/DSP 式架构。模拟架构功耗较低，但均衡灵活性有限；ADC/DSP 架构可以执行更复杂的 FFE、DFE、眼图监测和自适应算法，更适合 224G 高损耗通道，但功耗和延迟较高。接收端的目标并不是恢复理想模拟波形，而是在可接受功耗和延迟下，把判决前误码率控制在 FEC 能够纠正的范围内。内部 224Gbps 白皮书初稿_3753 也强调，端到端仿真应覆盖均衡、CDR、眼图、BER 及完整通道模型

2.1.3 时钟恢复机制

高速串行链路通常不会单独传输与数据同频的时钟，而是由接收端的 CDR（Clock and Data Recovery）根据数据跳变恢复采样时钟。CDR 一般由相位检测器、环路滤波器、压控振荡器或数字控制振荡器及反馈控制逻辑构成。相位检测器比较接收数据边沿与本地时钟之间的相位差，环路持续调整振荡器频率和相位，使采样点稳定在眼图中心附近。锁定后，CDR 既要跟踪低频频率漂移和 SSC，又要抑制超出环路带宽的高频输入抖动。PAM4 包含四个电平和多种跳变幅度，不同跳变穿越判决门限的方式不同，因此其 CDR 边沿检测和相位恢复比 NRZ 更复杂，通常需要结合数据判决、ADC 采样和 DSP 算法进行自适应优化。工程上重点关注锁定范围、锁定时间、抖动容限、抖动传递和抖动生成。Redriver 只有模拟均衡，不重建时钟，输入抖动会继续向后传递；Retimer 则包含 CDR，可对信号重新采样、整形和发送，从而切断前级抖动累积，但会增加功耗、成本和固定延迟，这一差别为 ACC 与 AEC 的核心区别， CDR 失锁会造成链路中断，因此必须提供锁定状态监测和恢复机制。

2.1.4 PAM4 原理

PAM4（Four-Level Pulse Amplitude Modulation）采用四个幅度电平传输数据，每个符号可以表示两个比特。例如四个电平可通过 Gray Coding 依次映射为 00、01、11、10，使相邻电平误判时通常只产生一个比特错误。由于每个符号携带 2 bit，PAM4 在相同符号率下可实现 NRZ 两倍的数据率；反过来看，在达到相同数据率时，PAM4 只需要 NRZ 一半的符号率和较低的 Nyquist 频率，因此能够降低 PCB、连接器和线缆在高频区的损耗压力。例如，106.25 Gb/s PAM4 对应约 53.125 GBaud，主要 Nyquist 频率约 26.56 GHz；212.5 Gb/s PAM4 则对应约 106.25 GBaud，Nyquist 频率约 53.125 GHz。PAM4 的代价是总电压摆幅被分成三个相邻电平间隔，形成上、中、下三只眼，每只眼的理想高度明显小于 NRZ；在相同总摆幅下，其理想幅度噪声裕量约缩小至 NRZ 的三分之一，对应约 9.5 dB 的幅度信噪比代价。 因此，PAM4 对发射端 RLM、接收门限、噪声、抖动、串扰和通道线性度高度敏感，必须结合 FFE、CTLE、DFE、CDR 与 FEC 实现可靠传输。

2.1.5 PAM4 与 NRZ 比较

NRZ 使用两个电平表示 0 和 1，每个符号承载 1 bit，结构简单、眼图裕量大、功耗和接收判决复杂度较低，适合较低速率或极短距离低功耗互连。PAM4 使用四个电平，每符号承载 2 bit，在相同数据率下仅需 NRZ 一半的符号率，因此可以降低 Nyquist 频率，减轻通道高频插入损耗、反射和材料带宽压力，这是 56G、112G 和 224G 高速接口采用 PAM4 的主要原因。 但 PAM4 的相邻电平距离只有 NRZ 整体电平间距的约三分之一，眼图从单眼变为三眼，信噪比、线性度、门限精度和抖动容限均更加严苛。NRZ 通常可以在较低原始误码率下工作，部分短链路可不依赖强 FEC；PAM4 则普遍需要均衡和 FEC 共同保障链路可靠性。此外，PAM4 发送端可能需要 DAC 和多抽头 FFE，接收端可能采用 ADC、CTLE、FFE、DFE 或 MLSD，导致功耗、面积、校准及测试复杂度提高。因此，两者的取舍本质上是“带宽与复杂度”的权衡：NRZ 用更高符号率换取较大噪声裕量，PAM4 用更小电平间距换取更高频谱效率。对于 AI Rack 的 112G/224G 电气互连，PAM4 已成为主流，但芯片到芯片的超短距低功耗接口仍可能根据系统目标采用 NRZ 或其他调制方式。

2.1.6 BER 与 FEC 要求

BER（Bit Error Rate，误码率）表示错误比特数与总传输比特数之比，是评估 SerDes 链路可靠性的核心指标。高速 PAM4 链路通常区分 Pre-FEC BER 与 Post-FEC BER：前者衡量均衡和判决后、FEC 纠错前的原始链路质量；后者衡量 FEC 解码后的最终数据可靠性。随着速率提高，通道损耗、串扰、随机抖动及 PAM4 较小眼高使原始误码率上升，因此系统允许一定数量的纠错前错误，再由 FEC 恢复。FEC 发送端将校验符号加入数据块，接收端利用冗余信息定位并纠正错误，常见方案包括 Reed-Solomon 类 FEC。FEC 能够显著放宽物理链路的原始 BER 要求，但会带来额外带宽开销、编码/解码延迟、逻辑功耗及突发错误敏感性，因此不能以 FEC 替代合理的通道设计。CEI-224G 示例中，不同 Reach 的 Pre-FEC BER 上限可从 5×10^{-6} 到约 10^{-4} 或 2.4×10^{-4}，而目标 Post-FEC BER 为 10^{-15} 量级；同一资料还以 RS(544,514,10)为外部 FEC 示例。 实际设计必须根据具体协议确认 FEC 类型、纠错门限和延迟预算，并同时验证随机错误、突发错误、误块率、FEC 余量以及误码平台效应。

2.2 224G SerDes 架构与关键技术模块

224G SerDes 是面向 1.6T Ethernet、AI Scale-Up、交换芯片、光模块及高速铜缆互连的单通道高速收发器。其典型架构由 TX、物理通道、RX、时钟系统和 DSP 控制系统组成。发送端依次完成数据编码、PAM4 映射、数字 FFE、并串转换和 DAC/Driver 输出；接收端则通过终端匹配、CTLE/VGA、时间交织 ADC、数字 FFE、DFE 或 MLSD、CDR、PAM4 判决及 FEC 接口恢复数据。224G PAM4 的符号率约为 112 GBaud，UI 约 8.93 ps，Nyquist 频率约 56 GHz，因此封装、PCB、过孔、连接器和线缆的微小反射、串扰与偏斜都会显著影响链路裕量。 相比 112G，224G 不仅要求模拟前端带宽翻倍，还需要更低抖动时钟、更高线性度 DAC/ADC、更复杂的自适应均衡和更严格的电源完整性设计。现代长距 224G SerDes 通常采用 ADC-DSP 架构，以数字方式完成 FFE、DFE、门限校准、CDR 和链路监控；短距应用则可能采用更简化的模拟架构，以降低功耗和延迟。CEI-224G Reach 覆盖约 50 mm 至 1 m、约 15 至 40 dB 通道损耗，不同 Reach 不能采用完全相同的功耗与均衡配置。

2.2.1 发射机架构（TX）

224G TX 负责把芯片内部低速并行数据转换为约 112 GBaud 的 PAM4 串行信号。典型数据路径为：编码/扰码→Gray Mapping→数字 FFE→并串转换器→高速 DAC→差分输出 Driver。Serializer 常采用多级 2:1 复用树，将并行数据逐级提升至目标速率；PLL、相位插值器和时钟分配网络则为各级复用器及 DAC 提供低抖动多相时钟。PAM4 发送机通常利用 6 ~ 7 bit 电流舵 DAC 或分段式 Driver 合成四个输出电平，同时通过数字 FFE 调整主抽头、前游和后游抽头，以预补偿封装、PCB 和连接器的高频损耗。公开的 224G 实现示例包含 9 抽头 FFE、Serializer、7 bit DAC、LCPLL 及输出匹配网络。 TX 的关键指标包括差分摆幅、RLM、SNDR、上升/下降时间、输出回损、随机抖动、确定性抖动和 FFE 调节范围。其中 RLM 反映 PAM4 四个电平的等间距程度，电平压缩会首先缩小最差子眼。FFE 虽可强化高频分量，但过强预加重会降低主游标幅度、放大串扰并增加 Driver 功耗。因此 TX 参数不能单独调优，必须依据封装、连接器和线缆组成的端到端 S 参数，以及 RX 均衡能力进行联合分配。

2.2.2 接收机架构（RX）

224G RX 的任务是在信号经历高损耗、反射、串扰和抖动后，恢复 PAM4 符号及时钟。典型接收路径包括差分终端、ESD 与输入匹配网络、CTLE、VGA、Sample-and-Hold、时间交织 ADC、数字 FFE、DFE/MLSD、CDR 和 PAM4 解码。CTLE 首先补偿通道高频衰减，VGA 把信号调整至 ADC 合适的满量程范围；ADC 将模拟波形转换为数字采样值，随后 DSP 消除残余码间干扰，并根据三个判决门限识别四个 PAM4 电平。公开的 224G RX 实现采用前端 CTLE/VGA、时间交织 SAR ADC、DCO-based CDR、最高约 30 抽头数字 FFE 以及可选 DFE。 RX 设计的关键不是把输入波形恢复成“完美眼图”，而是在有限功耗下获得足够的 SNR 和 Pre-FEC BER。高损耗通道需要更强 CTLE 及更多 DSP 抽头，但同时会放大噪声和串扰；高分辨率 ADC 能够改善量化精度，却会增加功耗、时钟负载和输入电容。在 40 dB 级通道中，性能往往受模拟前端噪声、线性度、时钟抖动和残余 ISI 共同限制，因此必须进行完整 PVT Corner、串扰、回损及 BER 验证，而不能只依据 Nyquist 频率处的单点插损值。

2.2.3 DSP 架构

224G SerDes 中的 DSP 负责将均衡、时钟恢复、门限校准和链路自适应从固定模拟电路转移到可编程数字域，是 224G 实现高损耗链路的关键。典型 DSP 数据路径为：ADC 采样→偏置与增益校准→数字 FFE→DFE 或 MLSD→PAM4 符号判决→Gray 解码→监测与 FEC 接口。控制路径则包括 AGC、CTLE 调节、FFE/DFE 抽头训练、PAM4 电平跟踪、CDR 环路、误码监控和后台自适应。数字 FFE 主

要消除线性前游及后游 ISI；DFE 使用已经判决的符号抵消后游 ISI，避免进一步放大噪声；MLSD 则通过序列概率寻找最可能的数据序列，适合反射明显或残余 ISI 较强的通道。公开 224G 设计显示，加入 MLSD 在部分高损耗或反射通道中可获得约 2 dB SNR 改善，但其标准化、功耗和延迟影响仍需系统评估。 DSP 通常依靠 LMS 等自适应算法更新抽头，并由片上微控制器管理训练、温漂补偿和失锁恢复。其优势是灵活、可校准并能适配多类通道；不足是功耗、面积和流水线延迟较高。对 AI Scale-Up 而言，不能只追求最大均衡能力，还要综合评估 DSP 引入的 pJ/bit、固定延迟及 FEC 前 BER。

2.2.4 FFE 前馈均衡器

FFE（Feed-Forward Equalizer，前馈均衡器）是一种线性 FIR 滤波器，可部署在 TX 或 RX，其基本功能是对当前符号及其前后符号进行加权，抵消通道产生的码间干扰。TX FFE 在信号进入通道前实施预加重，通过降低低频或增强跳变分量，预先补偿通道的高频衰减；典型配置包括若干 Pre-cursor、一个 Main-cursor 和若干 Post-cursor。RX 数字 FFE 则对 ADC 采样值进行多抽头卷积，可以同时处理前游和后游 ISI，并通过 LMS 等算法自适应更新系数。224G 通道中，短通道可能只需要少量 TX 抽头和 CTLE，而反射或长距通道通常需要更多 RX FFE 抽头。公开 224G 实现中，TX 采用 9 抽头 FFE，RX 数字均衡器可配置至约 30 抽头； CEI-224G LR 分析则采用多前游、主抽头、多后游及浮动抽头模型评估长尾反射。 FFE 的优势是线性、稳定且不发生错误传播，但它在增强高频信号的同时也会放大噪声与串扰，并可能降低有效主游标。因此抽头数量不是越多越好，应根据脉冲响应、反射位置、功耗和收敛速度配置，并与 CTLE、DFE 共同分配均衡任务。

2.2.5 CTLE 连续时间线性均衡器

CTLE（Continuous-Time Linear Equalizer，连续时间线性均衡器）位于 RX 模拟前端，通过构造随频率升高而增加的增益响应，补偿通道低通特性。其本质是在低频处提供较低增益或衰减，在 Nyquist 频率附近提供峰值增益，使被通道压低的高频分量重新接近低频分量。CTLE 通常由源退化差分对、可编程电阻电容网络或 LC 谐振结构实现，并与 VGA、终端匹配及 ADC 驱动电路协同设计。224G PAM4 的 Nyquist 频率约为 56 GHz，传统 RC 型 CTLE 难以同时满足带宽、增益、噪声和线性度要求，因此部分设计采用低 Q 值 LC 谐振或高、低频双路径架构。相关高速接收机研究表明，谐振型 CTLE 可通过电容阵列调节峰值频率，但过高 Q 值会产生时域振铃和额外后游 ISI。 CTLE 的优势是位于 ADC 之前，可减少 ADC 动态范围压力且功耗相对可控；不足是作为线性均衡器会同时放大高频噪声、串扰及反射，而且不能区分有效信号和干扰。工程上应避免只追求最大 Peaking Gain，而要综合评估频响平滑度、群时延、振铃、CMRR 和 PAM4 线性度，并通过 CTLE 与数字 FFE/DFE 分工降低总噪声和功耗。

2.2.6 DFE 判决反馈均衡器

DFE（Decision Feedback Equalizer，判决反馈均衡器）利用已经判决的历史符号预测并抵消当前采样点上的后游 ISI。其工作过程是把之前若干个符号判决结果分别乘以反馈抽头系数，再从当前输入中减去相应干扰。由于 DFE 只抵消由历史符号造成的确定性 ISI，而不会像线性 FFE 那样直接增强高频噪声，因此特别适合处理通道长尾和较强 Post-cursor。224G PAM4 DFE 需要处理四电平符号、三个判决门限及极短 UI，对反馈路径延迟提出严格要求。纯模拟 DFE 可以在采样判决附近快速工作，但抽头数量和可编程性受限；ADC-DSP 接收机可在数字域实现更多抽头和复杂自适应，不过需解决流水线延迟，常使用 Unrolling、Look-ahead 或 Speculative DFE 降低关键反馈环路延时。公开 224G 设计通常把 DFE 作为可选模块，与数字 FFE 或 MLSD 组合使用。 DFE 的主要优势是不显著放大随机噪声，且

后游 ISI 消除效率较高；主要风险是错误传播，即某次错误判决会通过反馈影响后续多个符号。因而 DFE 优化必须关注抽头收敛、门限漂移、Burst Error 和 Pre-FEC BER，而不能仅观察平均眼高。对于反射位置较远的通道，还需使用 Floating Tap 或 MLSD 避免抽头资源浪费。

2.2.7 ADC DSP

ADC-DSP 架构是 224G 长距 SerDes 的主流技术路线之一。RX 模拟前端首先通过 CTLE 和 VGA 完成有限度的增益与频响补偿，然后由时间交织 ADC 以多相方式采样高速 PAM4 信号，再在数字域执行增益校准、FFE、DFE、CDR、门限跟踪和序列检测。由于单个 ADC 难以直接达到 112 GS/s，通常使用数十路甚至更多子 ADC 交织工作，每个子 ADC 在较低速率下采样。时间交织可扩展总采样率，但会引入增益失配、偏置失配、采样时刻偏差和带宽不一致，这些误差必须通过前景或后台数字校准消除。公开 224G 接收机采用时间交织 SAR ADC，并结合数字 FFE、可选 DFE 及 DCO-based CDR。 ADC 分辨率通常需要在量化噪声、线性度、输入带宽和功耗之间取舍，增加 1 bit 并不一定能改善高损耗链路，因为系统可能已经受 AFE 噪声、抖动或串扰限制。ADC-DSP 的优势是均衡灵活、可提供眼图监测、链路诊断及跨通道自适应；缺点是功耗较高、数字流水线复杂且引入固定延迟。工程设计中应联合优化 ADC ENOB、采样率、交织数量、时钟偏差和 DSP 抽头数，而非单独追求最高 ADC 分辨率。

2.2.8 CDR 时钟恢复器

CDR（Clock and Data Recovery）负责从接收到的 PAM4 数据中恢复采样频率和最佳采样相位，是 224G RX 能否稳定工作的核心模块。典型 CDR 包含相位检测器、数字环路滤波器、DCO 或 PLL、相位插值器和多相时钟分配网络。相位检测器根据数据过渡沿或均衡后的采样值产生早/晚误差信号，环路滤波后调整 DCO 频率或 PI 相位，使采样点保持在最差 PAM4 子眼的最佳位置。ADC-DSP 架构中，CDR 输入可来自原始 ADC 数据，也可来自数字 FFE 输出。前者延迟较小但 ISI 较重，后者眼图更清晰但会增加环路延迟和降低相位裕度。公开 224G 设计采用 DCO-based CDR，并允许由 ADC 或 FFE 输出驱动定时恢复。 CDR 设计需同时满足频偏跟踪、低频抖动跟随、高频抖动抑制、SSC 兼容、快速锁定与失锁恢复。环路带宽过宽会把输入高频抖动传入采样时钟，过窄则难以跟踪频率漂移；DSP 流水线延迟还可能造成 Jitter Peaking。系统验证应覆盖 JTOL、JTRAN、JGEN、锁定时间和 PVT 变化。对 Retimer 和 AEC 而言，CDR 还能切断前级抖动并实现信号再生，但会增加纳秒级延迟、功耗及互操作风险。

112G Serdes 与 224G Serdes 技术特征与对比

对比项目	112G SerDes	224G SerDes	224G 主要变化与挑战
典型调制	PAM4	PAM4	调制方式不变，主要通过提高 Symbol Rate 实现速率翻倍
Symbol Rate	约 53～56 GBaud	约 106～112 GBaud	约 2×
Nyquist Frequency	约 26.5～28 GHz	约 53～56 GHz	约 2×
TX 架构	Serializer + PAM4 Driver/DAC + TX FFE	高速 Serializer + 高线性 DAC/Driver + 强化 TX FFE	对 DAC 线性度、RLM、SNDR、Driver BW 和 Clock Jitter 要求更严
RX 架构	CTLE + DFE，或 CTLE + ADC/DSP	Broadband AFE + CTLE/VGA + ADC + DSP	ADC–DSP 在 MR/LR 等高损耗应用中的重要性明显提高

DSP 架构	DSP 用于 FFE/DFE、自适应和监控，复杂度适中	FFE + DFE/MLSD + CDR + ADC Calibration 等	DSP 承担更多 Channel Compensation，但功耗、面积和延迟增加
FFE	TX FFE 为主，部分 RX 采用数字 FFE	TX FFE + 多抽头 RX Digital FFE	需要处理更严重的 ISI 及反射，更多依赖 TX/RX 协同均衡
CTLE	传统 RC/Active CTLE 较成熟	高带宽 CTLE，更多考虑 LC/Resonant 结构	重点从“Peaking Gain”转向 BW + Noise + Linearity + Ringing 综合优化
DFE	广泛应用于 Post-cursor ISI 消除	DFE 仍重要，但反馈时序压力明显提升	~9 ps UI 使 Feedback Loop 更困难，更强调 Digital/Speculative DFE 或与 MLSD 结合
ADC-DSP	高性能LR中常用，但并非所有 Reach 必需	高损耗 224G RX 的重要架构路线	Time-Interleaved ADC、ENOB、Time Skew 及 Calibration 成为关键
CDR	Analog/Digital CDR 均成熟	DSP-assisted/DCO-based CDR 更加重要	UI 减半，对 PLL Phase Noise、Sampling Jitter 和 Loop Latency 更敏感
Channel Equalization	CTLE + FFE + DFE 即可覆盖较多应用	CTLE + TX/RX FFE + DFE/MLSD 联合	由单模块补偿转向系统级联合均衡
典型 LR 通道能力	CEI-112G LR 约 40 dB 级通道	CEI-224G 面向更严峻高损耗环境	224G 对 Package、PCB、Connector 和 Cable 提出更严格 SI 要求
架构复杂度	中等	高	Analog、Digital、Clock 及 Channel 必须 Co-design
功耗压力	中等	明显增加	ADC、高速 Clock 和 DSP 成为主要功耗挑战
不同 Reach 架构	VSR/LR 可采用不同配置	XSR/VSR 与 MR/LR 进一步分化	单一 SerDes Core 可能难以高效覆盖 XSR～LR 全部应用。

2.3 224G 信道信号完整性要求与分析方法

224G PAM4 信道的核心挑战来自符号率相对 112G 近似翻倍，Nyquist 频率从约 28 GHz 提升至约 56 GHz，使 PCB 介质/导体损耗、封装损耗、连接器反射、Via Stub 和串扰显著恶化。因此 224G 信号完整性不能只检查单一 Insertion Loss，而应进行端到端 Channel 分析。完整信道通常包括 ASIC Package、BGA Breakout、PCB Trace、Via、Connector、Backplane/Cable 及对端 Package。主要频域指标包括 SDD21 插入损耗、SDD11/SDD22 回损、ILD、NEXT/FEXT、PSXT 及 SCD21 模式转换；时域

还需检查 TDR 阻抗、Differential Skew、Impulse Response、ISI、Jitter 及 Eye Margin。最终应结合 TX FFE、RX CTLE/FFE/DFE 及 FEC 能力，通过 COM 或端到端 SerDes 模型判断链路裕量。你现有 224Gbps 白皮书初稿_3753 也明确提出 S 参数分析必须覆盖完整通道，并将 COM 与 ERL 作为核心 Channel 指标。

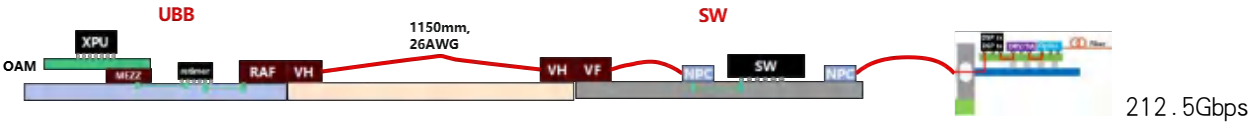
2.3.1 Channel Budget 分析

Channel Budget 用于把 SerDes 允许的总通道损耗分配给各物理组件，建立“每一段最多可以损耗多少”的设计边界。典型链路可表示为：TX Package → PCB → Via → Connector → Cable/Backplane → Connector → PCB → RX Package

分析首先根据目标 Reach 和 Nyquist 频率确定端到端 Insertion Loss Budget，再分配至 Package、PCB、Connector 和 Cable，同时为制造公差、温度、老化和系统 Margin 预留空间。需要特别强调，Loss Budget 不能只做简单 dB 相加，因为 Reflection、Crosstalk、Skew 及频响不平坦同样可能使链路失败。例如两个 IL 相同的 Channel，如果一个具有明显 Via/Connector 谐振，其 BER 可能明显恶化。因此工程流程应采用 IL Budget 作为第一层筛选，RL/Crosstalk/ERL 作为第二层约束，COM/BER 作为最终系统判据。现有 224Gbps 白皮书初稿_3753 指出 224G 长 PCB/背板损耗已成为主要限制，并需要通过低损耗材料、缩短 PCB 路径、Twinax 或 Retimer 改善链路。

案例：224G PAM4 Channel Budget 分析

场景：XPU 与 SW 通过 retimer--弯母连接器--直公 cable 背板--直母 cable 互联。



212.5Gbps

XPU—SW 场景（Cable Tray 互连）

表 212.5Gbps 链路概算 1（Cable Tray 互连 SW→Retimer）

212.5Gbps Channel bump-bump	SW Package	SW via	SW breakout	SW trace	NPC Conn	cable	VH-VF Conn	cable	VH-RAF Conn	UBB trace	Retimer Package	Total loss bump-bump (温度及波动影响)
Length inch/mm）; count	35mm	1	0.75inch	2inch	1	300mm	1	1150mm	1	4inch	15mm	38.72 @53.125GHz Margin:1.28dB
System loss dB @53.125GHz	7.5	0.6	0.90	2.12	1.50	3.59	2.50	8.78	3.50	4.23	3.5	

表 212.5Gbps 链路概算 1（Cable Tray 互连 Retimer→GPU）

212.5Gbps Channel bump-bump	Retimer Package	Retimer via	UBB trace	Mezz Conn	OAM trace	OAM via	GPU Package	Total loss bump-bump (温度及波动影响)
length（inch/mm）; count	15mm	1	8inch	1	3inch	1	25mm	23.14 @53.125GHz Margin:16.86dB
System loss dB @53.125GHz	3.50	0.60	8.47	1.80	3.18	0.60	5.00	

2.3.2 COM 分析方法

COM（Channel Operating Margin）是一种把物理 Channel 与参考 SerDes 均衡能力结合起来评估链路裕量的方法, 比单独观察 Insertion Loss 或 Eye Diagram 更接近实际系统性能。基本思路是输入 Victim Channel 及多个 Aggressor 的 S 参数，建立参考 TX/RX 模型，加入 TX FFE、RX CTLE/FFE/DFE，并计算 ISI、Reflection、Crosstalk、Jitter 及 Noise 共同作用后的有效信号裕量。分析时通常首先检查 SDD21、SDD11/22、ERL、ILD、NEXT/FEXT/PSXT 和 Mode Conversion，排除明显异常，再运行 COM 模型进行综合评价。如果 COM 不足，应回到物理通道定位问题，而不是简单增加 RX Equalization。优化顺序通常是减少 Via Stub 和阻抗不连续、改善 Connector Transition、降低 PCB/Cable Loss 与 Crosstalk，之后再优化 FFE/CTLE 参数。你现有白皮书已将 COM、ERL 和完整通道 S 参数列为关键评价指标，并要求 Cable、Termination、Cage 和 Connector 纳入同一 Channel Model。

案例：224G PAM4 通道 COM 分析

以 212.5 Gb/s PAM4 高速通道为例，符号率为 106.25 GBaud，Nyquist 频率为 53.125 GHz。IEEE P802.3dj 200G/lane KR 相关基线采用 die-to-die IL ≤40 dB @ 53.125 GHz，并定义参考 Die、Package 及 COM 参数。

分析项目	案例参数/结果	工程评价
Signaling Rate	212.5 Gb/s PAM4	224G 级 SerDes
Nyquist Frequency	53.125 GHz	SI 关键评价频率
Channel Boundary	Die-to-Die	Package + PCB + Via + Connector/Cable
TX Package IL*	4.0 dB	示例 Budget
PCB + Via IL*	14.0 dB	TX/RX 两侧合计
Connector IL*	2.0 dB	同时关注 RL/ERL 与串扰
Cable/Backplane IL*	14.0 dB	主要传输损耗
RX Package IL*	4.0 dB	示例 Budget
Total IL*	**38.0 dB**	低于案例 40 dB 参考边界
Channel 输入	SDD21 、 SDD11/22 、	建立 Victim/Aggressor 模型

	NEXT/FEXT 、 Mode Conversion	
Equalization	TX FFE + RX CTLE/FFE/DFE	优化 Residual ISI
综合干扰	ISI + Reflection + Crosstalk + Noise + Jitter	COM 综合评价
COM 结果*	3.5 dB	示例计算结果
案例判据*	3.0 dB	具体门限以采用的标准/COM 版本为准
Margin*	+0.5 dB	案例：PASS

案例结论： 224G 通道不能仅依据 Nyquist 频率处的 Insertion Loss 判断合规性。即使总 IL 满足 Channel Budget，Connector/Via 引起的 Reflection、Crosstalk、Mode Conversion 及 ISI 仍可能显著降低链路裕量。因此应采用 Channel Budget 初筛 → S 参数/ERL 分析 → COM 系统裕量评价 → BER 验证 的分层方法，并将 SerDes 均衡能力与 Package、PCB、Connector 和 Cable 作为完整端到端 Channel 进行 Co-design。224G 测试研究也显示，连接器/焊盘 Stub 可能在 57.3 GHz 等接近关键工作频带的位置形成明显谐振，因此单点 IL 不足以代表完整通道性能。

2.3.4 112G 与 224G 信号完整性要求对比

112G 和 224G 均主要采用 PAM4，但 224G 符号率近似翻倍，因此 SI 设计难度不是简单线性增加。112G PAM4 的 Nyquist 约 28 GHz，而 224G 约 56 GHz，这导致相同 PCB、Via 和 Connector 在 224G 下表现出更高的 Insertion Loss、Reflection 和 Crosstalk，同时原来位于带外的 Via Stub 或 Connector Resonance 可能进入关键频段。112G 系统中，通过低损耗 PCB、传统 Connector 和 CTLE/DFE 通常仍可覆盖较多板级应用；224G 则更加依赖超低损耗材料、短 PCB 走线、优化 Via/BGA Transition、高带宽 Connector 以及更强的 TX/RX Equalization。进入 224G 后 PCB 介质和导体损耗显著增加，需要使用更低损耗板材、缩短信号路径或引入 Retimer。

2.4 Retimer 基础原理和发展背景

Retimer（重定时器）是一类用于高速串行链路信号再生的有源器件，其核心作用是对经过高损耗通道后的信号进行均衡、时钟数据恢复（CDR）、重新采样和重新发送，从而将一条长距离、高损耗链路分割为两段或多段独立的 SerDes 通道。随着 SerDes 从 56G、112G 发展到 224G PAM4，Nyquist 频率提高至约 56 GHz，高频下 PCB、连接器和铜缆的插入损耗、反射、串扰及 ISI 显著增加，使纯无源铜互连的可达距离受到限制。OIF 针对 CEI-224G 定义了 XSR、VSR、MR 和 LR 等应用空间，其中 MR 目标覆盖最高 500 mm PCB 和一个连接器，LR 进一步面向最高 1000 mm Backplane 及两个连接器。在此背景下，Retimer 成为扩大铜互连 Reach 的重要技术之一，尤其适用于板载 Retimer、AEC、Backplane

以及 AI Scale-Up 互连。2026 年发布的 224G Retimer 产品已面向 1.6 Tb/s、224 Gb/s/lane PAM4 及 AEC/板载应用，反映出 Retimer 正在从 112G 向 224G AI 互连快速演进。

2.4.1 Retimer 工作原理

Retimer 的核心可以概括为 Equalize → Recover Clock → Re-sample → Re-transmit。高速 PAM4 信号经过 PCB、连接器和铜缆后，会产生衰减、ISI、Reflection 和 Jitter。Retimer 接收端首先通过 CTLE、FFE、DFE 或 ADC-DSP 对受损波形进行均衡，再利用 CDR 从输入数据中恢复时钟，并在优化后的 Sampling Point 重新判决 PAM4 Symbol。恢复后的数字数据随后进入另一侧发送 SerDes，经 TX FFE、DAC/Driver 重新生成新的 PAM4 信号。与 Redriver 单纯对模拟波形进行线性放大和均衡不同，Retimer 通过 CDR 和重新采样建立新的时序基准，因此可以显著降低前段链路累计的 Jitter 及 ISI 对后段链路的传递。Retimer 本质上相当于在物理通道中建立一个新的 SerDes 边界：SerDes → Channel A → Retimer RX/CDR → Retimer TX → Channel B → SerDes 这使Channel A和Channel B分别满足各自的Loss、COM 和 BER 要求，而不再要求原始 SerDes 直接覆盖整个端到端高损耗 Channel。Retimer 通常利用 CDR、Equalization 和 Signal Regeneration 扩展高速互连 Reach。

2.4.2 Retimer 架构

典型 224G Retimer 由双向 SerDes、模拟前端、ADC/DSP、CDR、数字数据通路、TX Driver 及管理控制单元组成。每个方向可抽象为：RX AFE → CTLE/VGA → ADC/Slicer → FFE/DFE → CDR → Digital Data Path → TX FFE → DAC/Driver RX 负责对前级通道进行均衡和数据恢复，CDR 建立新的采样时钟；内部数字路径将恢复后的数据转送至另一侧 TX，由发送端重新产生满足输出规范的 PAM4 信号。高性能 224G Retimer 越来越依赖 DSP 完成自适应均衡、通道训练及信号恢复，以应对铜通道中的衰减、回损和 ISI。 架构设计通常需要在 Equalization 能力、Latency、功耗和通道 Reach 之间权衡。短通道采用过强 DSP 会增加不必要的功耗，而 Backplane 或长铜缆则需要更强均衡能力。OIF 为 224G 分别设置 XSR、VSR、MR 和 LR 项目，本质上也反映了不同 Reach 在功耗、复杂度和 Channel Loss 方面具有明显不同的需求。

2.4.3 Redriver 与 Retimer 比较

Redriver 与 Retimer 都用于改善高速链路质量，但工作机制不同。Redriver 属于线性信号调理器，主要利用增益、CTLE 及 Pre/De-emphasis 补偿通道频率相关损耗。由于不执行完整 CDR 和数据重新判决，其优势是架构简单、功耗低、延迟小，但前级 Jitter、Noise 及部分信号失真仍会继续向后级传递。Retimer 则属于信号再生器，通过 Equalization、CDR、重新采样和 TX 重发建立新的信号与时钟，因此更适合高损耗或分段长距离通道。相关 224G Retimer 资料明确将 CDR、均衡和 Reach Extension 作为其关键作用。

对比项	Redriver	Retimer
信号处理	模拟/线性均衡	完整信号再生
CDR	通常无	有

Re-sampling	无	**有**
Jitter 隔离	弱	**强**
Channel 分段	否	**是**
功耗	较低	较高
延迟	很低	较高
适用场景	中短距、较低损耗	**高损耗、长距、AEC/Backplane**
224G 价值	Signal Conditioning	**Reach Extension + Signal Regeneration**

2.4.4 Retimer 系统影响

Retimer 能够显著提高高速铜互连的可达性, 但必须从系统角度评价, 而不能仅视为“增加 Channel Budget”的器件。其最大价值在于将一条高损耗端到端 Channel 分割成两个独立的 SerDes 链路，使每段分别满足 COM、BER 和均衡要求，同时通过 CDR 和重新采样降低前级 Jitter 与 ISI 继续累积的风险。在 224G AI Scale-Up 系统中，这可扩大 PCB、Backplane 和 AEC 的设计空间，减少 ASIC 必须紧邻 Connector 或 Backplane 布置的限制。224G Retimer 产品已经面向 AEC 和 On-board Retimer 应用，以解决高速铜通道中的衰减、Return Loss 及 ISI 问题。

另一方面，Retimer 会引入额外功耗、成本、Latency、PCB 面积和热设计压力，同时增加电源、时钟、Firmware、链路训练及互操作性的系统复杂度。因此 Retimer 应作为系统级 Channel Architecture 决策：当 Passive Channel 或 Redriver 方案仍具有足够 COM/BER Margin 时，应优先考虑低功耗方案；当通道损耗、距离或 Jitter 累积超过原始 SerDes 能力时，Retimer 才成为扩展 Reach 的重要手段。

3. 第三章 单/双 Rack 高速互连方案

3.1 AI Rack 互连形态

AI Rack 是 AI 计算集群的基本单元, 其内部及之间的互连形态直接决定集群的带宽、时延与可扩展性，是提升整体算力的关键。当前互连围绕机架内（Scale-Up）与机架间（Scale-Out）两个核心层级展开：机架内追求“一柜一台大电脑”的极致协同，机架间支撑大规模集群的资源池化与调度；整体正朝更高密度、更低延迟与光铜融合的方向演进。

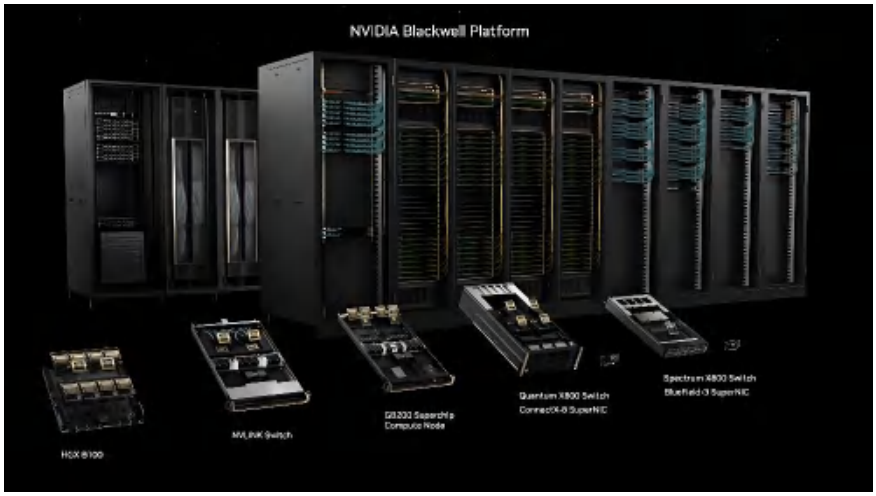


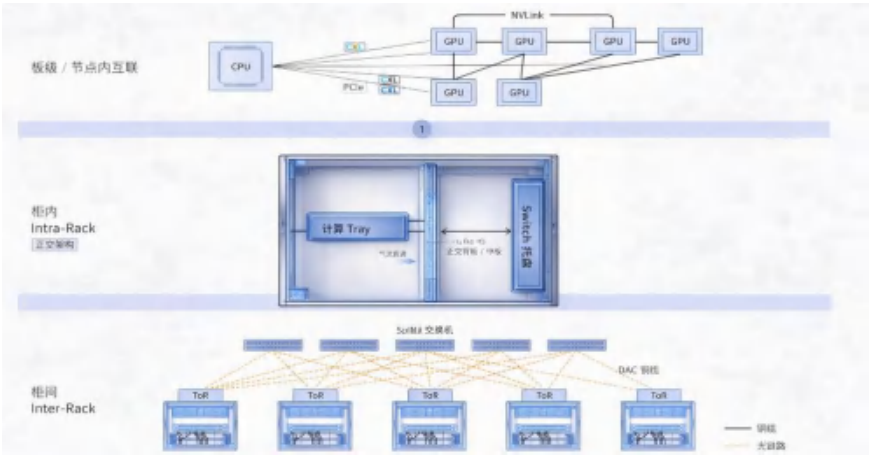
图 3.1NVIDA Blackwell Platform

互连形态可以从三个层级理解：

一是板级/节点内互连，包括 CPU-GPU、GPU-GPU（NVLink、PCIe、CXL 等）与内存一致性互连；

二是柜内互连（Intra-Rack），包括同一 Rack 内 Tray 之间、Tray 与 Switch 之间（铜缆、背板、中板、线缆托盘、交换托盘形态与 Cable Tray/正交互连架构等）；

三是柜间互连（Inter-Rack），包括 Rack 与 Rack 之间的 ToR/Spine 组网与光模块、AOC、DAC、光纤配线等。整体朝更高密度、更低延迟与光铜融合的方向演进。



AI Rack 高速互连的三个层级

3.1.1 单宽单柜/单宽双柜/双宽单柜/双宽双柜

柜内交换/计算托盘按占用槽位宽度与部署机柜范围形成四种组合：单宽单柜、单宽双柜、双宽单柜、双宽双柜。单宽指占用 1 个标准槽位宽度（如半宽 1RU 交换托盘），双宽指占用 2 个槽位宽度（常见于高密度 GPU 与 51.2T/102.4T 级交换托盘）；单柜指在单个机柜内部署，双柜指跨相邻两个机柜部署（常见于超长交换托盘与高密跨柜组网）。以 GB300 NVL72 为例，机柜内 18 个计算托盘（1U 全宽、每托盘 4 颗 B300）与 9 个 NVSwitch 托盘（每托盘 2 颗交换芯片）即属单柜内形态组合。

维度	单宽单柜	单宽双柜	双宽单柜	双宽双柜
槽位宽度	1 个标准槽位	1 个标准槽位	2 个槽位	2 个槽位
部署范围	单个机柜内	跨两个机柜	单个机柜内	跨两个机柜
典型特点	密度高、可并排多组	跨柜走线、协同部署	高密交换、单柜整机	超大托盘、跨柜组网
适用场景	通用交换/计算托盘	跨柜线缆集中管理	51.2T 级高密交换	102.4T/1.6T 时代交换

3.1.2 Cable Tray 架构/正交中板/正交无中板

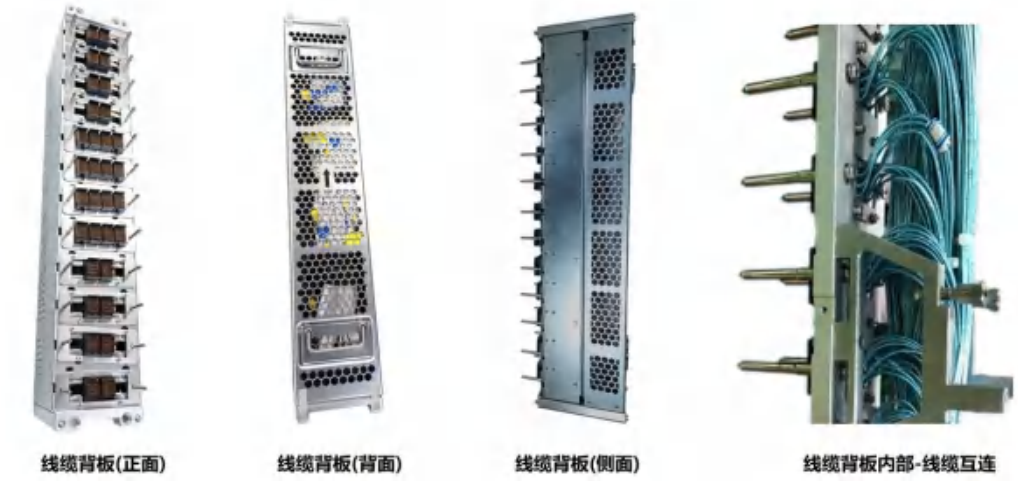
柜内互连架构分两类：一是 Cable Tray（线缆托盘）架构，通过机柜顶部或侧向线缆托盘集中承载铜缆与光纤，便于理线与扩容，AI 机柜中 GPU 与 Switch 之间的大量线缆依赖该方式管理；二是正交架构，接口板（计算板）与交换网板呈 90° 正交直接互连，其中正交中板经中板/背板连接器互连，结构成熟但中板增加信号跳变次数，正交无中板（正交直连）省去背板走线，信号路径最短、衰减最小、气流直通，已成为高带宽核心交换的主流方向（华为、锐捷、新华三、迪普等均采用）。NVIDIA GB300 NVL72 中 GPU 与 NVSwitch 经机柜背板连接，构成柜内无阻塞交换域。

柜内互连架构对比

维度	Cable Tray 架构	正交无中板	正交中板
原理与走线	机柜顶部/侧向线缆托盘集中走线	前后板卡 90° 正交直连，背板零走线	前后板卡经中板/背板连接器互连
信号与带宽	取决于线缆质量与布线	路径最短、衰减最小	经中板跳线，损耗略大
散热与气流	线缆占用空间，需规划风道	无中板阻挡，气流直通	中板阻挡气流，需开孔/导流
可维护性	易扩容、易理线	免背板、维护性好	结构成熟、抽拉更换
典型应用	AI 机柜 GPU↔Switch 布线	华为/锐捷/H3C 核心交换	传统框式核心交换

Cable Tray 架构是一种以高速铜缆模组为核心载体的高速互连方案。它通过采用高性能铜缆化连接器模组替代传统 AEC 线缆或 PCB 背板走线，成为计算节点与交换节点之间主要的高速互连物理通道，从而提供更高的互联带宽和更低的传输时延。在该架构下，计算节点单元与交换节点单元能够实现无源直接互连，形成灵活的网状拓扑结构，可承载多种互联协议，突破传统背板在尺寸、层数和走线长度方面的限制，支持高密度、大规模互连，并具备对多种 GPU 芯片的兼容能力。系统后续如需升级带宽，也可通过更换线缆或调整连接方式实现快速演进，降低整机系统的迭代成本。

Cable Tray 架构在散热适配方面也具有较强灵活性，可同时支持风冷和液冷部署。通过优化机柜风道，或灵活调整冷板、Manifold 以及互连总线的空间布局，能够在保证高速信号完整性的同时提升散热效率，满足整机柜、超节点等高密度算力场景对功耗管理和部署灵活性的要求。



线缆背板（Cable Tray）产品示意图

随着连接器与铜缆工艺的持续优化，线缆背板方案的链路损耗裕量得到显著提升，互连长度可扩展至约 2 米。其通常搭载盲插、自导向和浮动机构，能够有效消除传统硬连接因插拔偏差、公差累积或不完全互配导致的高速信号劣化问题，同时支持线对板、线对线等多种互连形态转换，具备较高的设计灵活性与连接可靠性。该方案尤其适合计算节点与网络节点采用水平布局或垂直堆叠的整机柜架构，因此成为当前整机柜超节点柜内应用较为广泛的高速互连方案。

为保障线缆背板在长期使用中的可靠连接，浮动机构与对接力设计通常需要重点关注以下方面：一是面板侧 X/Y 轴浮动能力，要求在指定承压力区间内实现 正负数毫米的自校正导引，机构运动全程不应出现卡滞、锁死或回弹失效等问题，行业常规设计中 XY 浮动能力约为 1.5~3 mm；二是面板侧 Z 轴行程浮动，即沿弹簧压缩方向，要求在节点出现过压状态时能够承受最大 数毫米 的弹簧压缩，且机构仍保持稳定，行业常规设计中 压缩行程约为 2 ~ 4 mm；三是弹簧弹力性能，弹簧是实现 X/Y/Z 三轴向浮动的核心部件，需要在不同结构设计和承压力区间内稳定提供符合要求的弹性响应，确保插拔力、保持力与浮动补偿能力之间取得平衡。

3.2 GPU Tray 互连方案

GPU Tray 是承载多颗 GPU 加速卡及其配套内存、网卡的计算托盘，是 AI Rack 中算力的核心载体。GPU Tray 内部连接方式主要解决两个问题：一是 GPU 与 GPU 之间的高速协同通信（Scale-Up），二是 GPU 与 CPU、内存、网卡及存储之间的控制面与数据面连接。

（1）NVLink / NVSwitch 直连（NVIDIA 生态）

NVLink 是 NVIDIA 推出的 GPU 间高速点对点互连总线，第六代 NVLink 单卡双向带宽可达 3.6 TB/s。在 GPU Tray 内部，多颗 GPU 通过 NVLink 直接互联，或经由板载 NVSwitch 芯片构建无阻塞全互联域。其优势在于带宽极高、延迟极低，专为 GPU 集合通信（AllReduce、AllGather 等）优化；劣势在于生态封闭，仅支持 NVIDIA 自家 GPU，且功耗与 PCB 布线难度较高。

（2）HCCS（华为昇腾生态）

HCCS（Huawei Cache Coherence System）是华为昇腾系列 AI 芯片的卡间高速互连总线，承担昇腾芯片之间的一致性通信与数据交换。其优势在于与昇腾软硬件栈深度整合，支持大规模集群级联；劣势同样是生态封闭，跨厂商兼容性受限。

（3）UALink（开放标准）

UALink（Ultra Accelerator Link）是由 AMD、ARM、Google、Intel、Meta、Microsoft、NVIDIA 等

厂商联合推动的开放加速器互连标准，1.0 版本支持单计算域最多 1024 个加速器，硬件预计 2026 年底出货、2027 年规模部署。其优势在于开放生态、多厂商兼容、可扩展性强；劣势在于标准尚在成熟过程中，商用产品与生态工具链有待完善。

（4）PCIe 交换

PCIe（Peripheral Component Interconnect Express）是通用 IO 总线，GPU Tray 内部 CPU 与 GPU 之间、GPU 与网卡/存储之间通常通过 PCIe 连接。PCIe 6.0 单通道速率为 64 GT/s（PAM4），PCIe 7.0 将提升至 128 GT/s。其优势在于标准化程度高、生态成熟、成本低；劣势在于带宽与延迟不如专用互连总线，且 PCIe 交换芯片会引入额外转发延迟。

GPU Tray 内部连接方式对比

连接方式	典型带宽	延迟	生态开放性	功耗	适用场景	224Gbps 支持
NVLink/NVSwitch	极高（单卡 TB/s 级）	极低	封闭（NVIDIA）	较高	NVIDIA GPU 训练协同	roadmap 支持
HCCS	高	低	封闭（华为）	较高	昇腾芯片集群	roadmap 支持
UALink	高	低	开放	中等	多厂商加速器域	标准规划中
PCIe 交换	中（PCIe 7.0 约 128GT/s）	中	完全开放	低	CPU-GPU、网卡、存储	PCIe 7.0 接近但未达

3.2.1 GPU Tray 内部典型互连拓扑

GPU Tray 内部多颗 GPU 的互连拓扑直接影响集合通信效率、带宽均衡性与可扩展性。典型拓扑包括以下四种：

（1）全互联拓扑（Full Mesh）

每颗 GPU 与其他所有 GPU 均有直接链路连接。其优势是任意两点间均为一跳直达，带宽无阻塞、延迟最低；劣势是链路数量随 GPU 数量呈平方级增长（N 颗 GPU 需要 N(N-1)/2 条链路），PCB 布线与连接器资源消耗巨大，通常仅适用于 4 卡及以下的小规模 Tray。在 224Gbps 速率下，全互联拓扑的布线密度与信号完整性挑战极为严峻，仅在极短距离（板内走线）下可行。

（2）环形拓扑（Ring）

各 GPU 依次连接形成闭合环路。其优势是链路数量少（N 条）、布线简单、成本低；劣势是远距离通信需多跳转发，带宽随跳转数衰减，单点故障会影响整个环。环形拓扑在 224Gbps 下信号路径可控，但多跳累积延迟不适合大规模训练的集合通信。

（3）胖树/分层拓扑（Fat Tree / Hierarchical）

GPU 分为多个组，组内全互联，组间通过上层交换节点（如 NVSwitch）连接。其优势是可扩展性好、带宽可通过胖树比例灵活配置，是 GB200/GB300 NVL72 等大规模系统的主流选择；劣势是需要额外交换芯片，成本与功耗较高，且组间通信存在竞争。在 224Gbps 下，胖树拓扑将高速链路集中在交换芯片附近，短距离走线有利于信号完整性，是最具可行性的大规模拓扑。

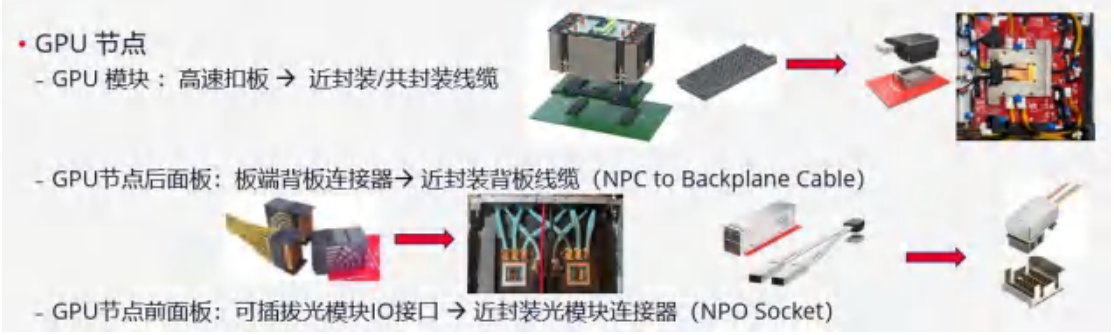
（4）混合拓扑

结合全互联与胖树的特点，如组内全互联 + 组间 NVSwitch。其优势是兼顾局部低延迟与全局可扩展性；劣势是设计复杂度高，需要仔细规划带宽比例。

GPU Tray 内部互连拓扑对比

拓扑类型	链路数量	最大跳数	带宽阻塞比	布线复杂度	适用规模	224Gbps 可行性
全互联 Full Mesh	N(N-1)/2	1 跳	无阻塞	极高	≤4 卡	仅板内极短距
环形 Ring	N	N/2 跳	有阻塞	低	小规模	可行但多跳延迟大
胖树 Fat Tree	可配置	2-3 跳	可配置无阻塞	中高	8-72 卡	推荐，短距集中走线
混合拓扑	可配置	1-3 跳	可配置	高	中大规模	可行，设计复杂

3.2.2 GPU Tray 内部连接方式；



3.3 Switch Tray 互连方案

Switch Tray 是 AI Rack 中承载交换芯片的托盘，负责柜内 GPU 间流量汇聚以及柜间流量转发。随着单芯片交换容量从 12.8T、25.6T、51.2T 向 102.4T 甚至 204.8T 演进，单颗芯片已难以满足端口密度需求，Switch Tray 内部常采用多芯片组合拓扑.

3.3.1 Switch Tray 内部典型互连拓扑

（1）单芯片交换拓扑

Tray 内仅放置一颗大容量交换芯片，所有端口直接连接到该芯片。其优势是结构最简单、无芯片间转发延迟、管理统一；劣势是受限于单芯片最大端口数与交换容量，当端口需求超出单芯片能力时无法扩展。在 224Gbps 下，单芯片方案的 SerDes 数量与功耗集中，散热压力大，但信号路径最短。

（2）多芯片级联拓扑（Cascade）

多颗交换芯片通过高速链路串联或树形级联，扩展端口数量。其优势是端口密度可线性扩展；劣势是级联链路会成为带宽瓶颈，跨芯片流量增加转发跳数与延迟，且级联链路本身也需要高速 SerDes，在 224Gbps 下板内级联走线的信号完整性是主要挑战。

（3）背靠背拓扑（Back-to-Back）

两颗交换芯片面对面放置，中间通过极短的高速链路直接互联，形成一个逻辑交换实体。其优势是芯片间距离极短、信号损耗小，可有效扩展容量；劣势是仅适用于双芯片组合，扩展能力有限，且机械结构设计特殊。在 224Gbps 下，背靠背的极短链路是最有利于信号完整性的多芯片方案。

（4）正交连接拓扑

交换芯片板与接口板呈 90 度正交布置，通过正交连接器直接互联。其优势是信号路径短、无背板走线、气流直通，是高带宽核心交换的主流方向；劣势是机械结构与连接器设计复杂，成本较高。正交拓扑在 224Gbps 下优势显著，是下一代高密交换的首选架构。

Switch Tray 内部互连拓扑对比

拓扑类型	芯片数量	端口扩展性	转发跳数	信号路径	散热难度	224Gbps 可行性
单芯片交换	1	受限于单芯片	1 跳	最短	高 (功耗集中)	可行，散热是瓶颈
多芯片级联	≥2	线性扩展	2-3 跳	较长	中	级联链路挑战大
背靠背	2	双倍容量	1-2 跳	极短	中	推荐，短链路优势
正交连接	≥2	可扩展	1-2 跳	短（无背板）	低（气流好）	强烈推荐

3.3.2 Switch Tray 内部连接方式

Switch Tray 内部连接方式主要包括交换芯片之间的互联、交换芯片与外部端口（光口/电口）的连接，以及为补偿信号损耗而引入的中间器件。

（1）交换芯片间互联

多芯片 Switch Tray 中，交换芯片之间通过板级高速走线或连接器互联。在 112Gbps 及以下速率，普通 FR-4 PCB 走线即可满足；进入 224Gbps 后，普通板材的介电损耗与导体损耗急剧增加，通常需要采用低损耗板材（如 MEGTRON 6/7、Nelco 4000-13）、缩短走线长度、或引入 Retimer 芯片进行信号重塑。

（2）芯片与光口/电口连接

交换芯片的 SerDes 端口需要连接到前面板的可插拔光模块（OSFP、QSFP-DD 等）或铜缆接口。连接路径包括 PCB 走线、连接器、光模块内部电气接口。在 224Gbps 下，光模块内部的电气接口损耗与连接器损耗成为瓶颈，通常需要在芯片与光口之间放置 Retimer 或 Redriver 进行信号补偿。Retimer 可完全重塑信号（时钟数据恢复），但增加功耗与延迟；Redriver 仅做信号放大均衡，功耗较低但补偿能力有限。

（3）线性直驱（Linear Drive / 无源通道）

在路径极短、损耗可控的场景下，可采用线性直驱方案，交换芯片 SerDes 直接连接光模块或对端芯片，中间无有源器件。其优势是功耗最低、延迟最小、BOM 成本最低；劣势是对通道损耗要求极为严格，在 224Gbps 下仅适用于极短路径（通常 ＜5 英寸 PCB＋ 连接器）。

Switch Tray 内部连接方式对比

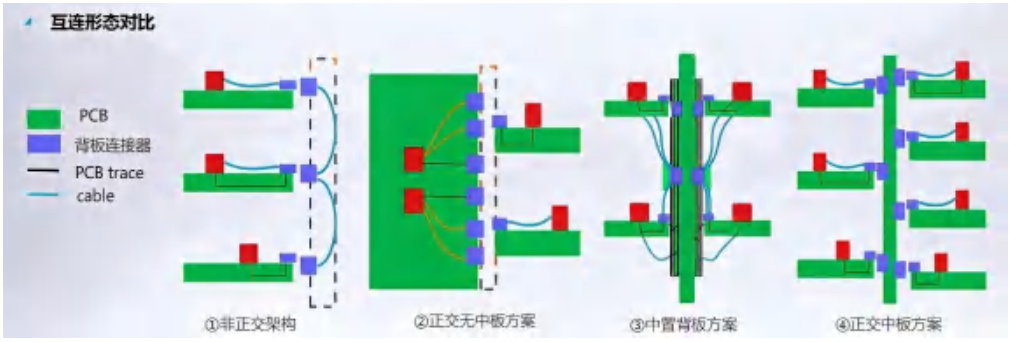
连接类型	实现要点	适用速率	优势	劣势	应用场景
交换芯片间互	PCB 走线 / 连接器；224G 采用低损耗板材或 Retimer	224G：低损耗板材	实现多芯片扩展	224G 损耗高，Retimer 增加功耗时延	托盘内交换芯片互连

连接类型	实现要点	适用速率	优势	劣势	应用场景
联					
芯片与光口/电口	PCB + 连接器，按需配置 Retimer/Redriver	112G/224G	接口标准化，模块可插拔	224G 接口损耗大，有源器件增加功耗	交换机对外端口，接 DAC/AOC/ 光纤
线性直驱无源通道	SerDes 直连，无有源器件	短通道可达 224G	功耗、时延、成本最低	通道损耗限制严格，224G 仅极短链路（<5 英寸）	托盘内近芯片超短走线



3.4 背板与背板线缆连接方案

背板（Backplane）是 AI Rack 中连接各功能托盘（GPU Tray、Switch Tray、IO Tray 等）的核心互连载体。背板方案的选择直接决定了柜内互连的带宽密度、信号质量、散热效率与可维护性。根据背板结构与连接方式，可分为以下五大类方案。



背板互连形态对比

（1）板对板背板方案:

指不同节点之间采用板级连接器与板级连接器直接对接的高速互连方式。该方案在铜互连通道中全部以 PCB 走线为主要信号载体，不引入线缆，因此能够充分发挥 PCB 走线在扇出布局上的灵活性。由于信号路径仍主要位于板内，单板可以较多复用既有 PCB 方案设计，节点结构设计、器件布局、单板组装工序以及检验机制也能较完整地参考传统案例经验，有利于缩短开发周期及实现整机设备全自动化装配检测的可行性,综合提升项目落地效率。但其主要挑战在于系统全链路损耗较为严峻，随着速率提

升和传输距离增加，PCB 走线损耗、插入损耗和回波损耗都会显著放大。为保障链路完整性，通常需要采用更高性能的芯片、更低损耗的 PCB 板材，或增加 Retimer 等信号补偿器件，从而在成本、功耗和散热方面带来更高要求。

（2）线对板背板方案：

指不同节点之间通过线缆化连接器与板级连接器对接实现高速互连。该方案在铜互连场景中能够压缩部分 PCB 走线长度，将原本较长的板内走线转化为局部线缆化传输，从而释放系统链路损耗余量，改善高速信号衰减问题。线缆化连接器通常可配合盲插、自导向和浮动机构使用，有利于提升装配容差和连接可靠性，降低现场维护难度。与此同时，部分单板仍可沿用现有 PCB 设计思路，节点结构设计、组装流程与检验机制也能部分参考传统方案，因此在性能优化与工程落地之间具有较好平衡。正因如此，线对板方案已成为全球行业内较为主流的高速铜互连应用类型。

（3）线对线背板方案：

指不同节点之间采用线缆化连接器与线缆化连接器直接对接的互连方式。该方案能够将 PCB 走线压缩到最短，最大化利用线缆完成高速信号传输，从而进一步降低板内走线带来的损耗，为长距离、高带宽铜互连提供更大的链路余量。其同样可结合盲插、自导向和浮动结构，提升连接可靠性和装配便利性。但线对线方案也面临较大的工程挑战：大量线缆会显著增加节点内部的结构设计复杂度，对走线路径、弯曲半径、固定方式、散热空间和电磁兼容布局提出更高要求；同时，线缆与单板的组装工序、检验标准和一致性管控也与传统 PCB 设计差异较大，需要建立更完善的工艺验证和质量控制体系。

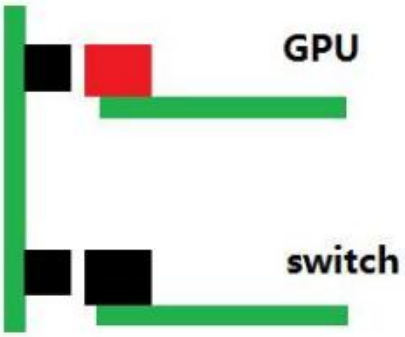
总体而言，板对板方案在布局灵活性和设计复用方面优势明显，但链路损耗压力较大；线对板方案兼顾损耗优化与工程可落地性，适合当前主流高速互连场景；线对线方案能够最大限度降低 PCB 走线损耗，但对线缆管理、结构设计和制造检验能力要求更高。实际选型需结合传输速率、链路长度、损耗预算、成本约束、装配维护及量产可行性综合评估。

3.4.1 非正交架构

非正交架构背板是传统框式设备最经典的结构：业务板卡与交换板卡平行插入，均通过同一平面的背板连接器连接，信号在背板 PCB 上走平面走线。

3.4.1.1 背板方案（板对板，线对板，线对线）

（1）板对板互配



业务板与交换板均通过背板连接器插接在同一块背板 PCB 上，信号经由背板上的铜箔走线完成连接。其优势是结构成熟、密度高、无需额外线缆；劣势是背板走线长度较长（通常 20-40 英寸），信号损耗大，且背板阻挡气流导致散热困难。在 224Gbps 速率下，普通 FR-4 背板的插损通常超过 40dB，

远超 SerDes 的接收能力，即使采用低损耗板材也难以满足长距离要求，仅在极短背板（< 10 英寸）下有条件可行。

连接器规格涉及弯母+直公规格

在 switch 与 GPU 互连通过背板内部走线交叉，叠层厚度高，成本相应增加

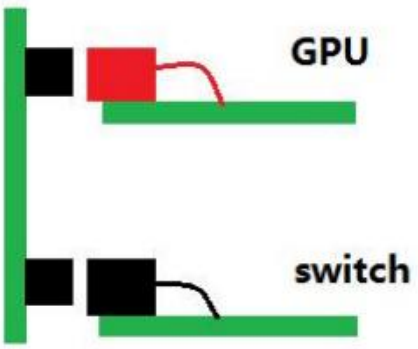
背板对整机风冷散热影响较大

结构设计精度要求高，连接器需具备高可靠性抵消受力对性能的影响， 单板尺寸过大时位置精度会受影响，进一步影响整柜 GPU 密度

节点内走线均通过 PCB，速率高时对芯片位置、板材等级要求高，PCB 成本受一定影响

典型机型：中兴天池超节点

（2）板对线互配



部分高速信号不经过背板 PCB 走线，而是通过高速线缆（如 Twinax 铜缆）直接连接业务板与交换板，背板仅承载电源与低速信号。其优势是高速信号走线缆，损耗远低于 PCB 背板，可支持更长距离；劣势是线缆管理复杂、密度受限、装配工时高。在 224Gbps 下，线对板方案通过优质 Twinax 线缆可支持 1-2 米距离，是传统背板向高速演进的重要过渡方案。

单侧使用 cable 浮动设计降低结构设计难度

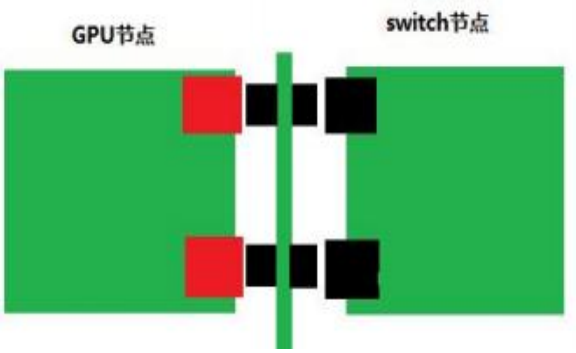
缩短芯片-Retimer-连接器在 PCB 板内的走线长度，可评估降低板材等级，减小整体链路损耗

部分交叉可以在 cable 内实现，降低单板内部走线交叉

3.4.1.2 中置背板方案（板对板，线对板，线对线）

中置背板方案将背板放置在机柜中部，前后两侧分别插业务板与交换板，信号从一侧板卡经中置背板连接到另一侧板卡。相比传统非正交背板，中置背板缩短了单侧走线长度，但信号仍需经过背板 PCB。

（1）板对板互配



前后板卡通过中置背板上的连接器与 PCB 走线互联。其优势是前后分区清晰、布线空间充裕；劣势是信号需穿越整个背板，路径仍较长，且中置背板阻挡前后贯通的气流。在 224Gbps 下，中置背板的板对板连接面临与非正交背板类似的损耗问题，通常需要 Retimer 补偿，且散热压力较大。

连接器规格涉及弯母+直公规格

在 switch 与 GPU 互连通过背板内部走线交叉，叠层厚度高，成本相应增加

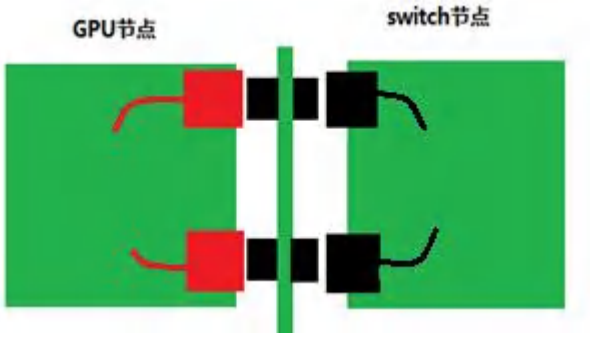
背板对整机风冷散热影响较大

结构设计精度要求高，连接器需具备高可靠性抵消受力对性能的影响，

节点内走线均通过 PCB，速率高时对芯片位置、板材等级要求高，PCB 成本受一定影响

典型机型：Kyber NVL144

（2）板对线互配



高速信号通过线缆跨接前后板卡，中置背板仅起结构支撑与电源分配作用。其优势是高速信号走线缆，损耗可控；劣势是中置背板的结构空间使线缆布放难度增加。在 224Gbps 下，线对板方案可通过短距离 Twinax 线缆满足要求，但机柜中部的线缆管理较为复杂。

单侧使用 cable 浮动设计降低结构设计难度

缩短芯片-Retimer-连接器在 PCB 板内的走线长度，可评估降低板材等级，减小整体链路损耗

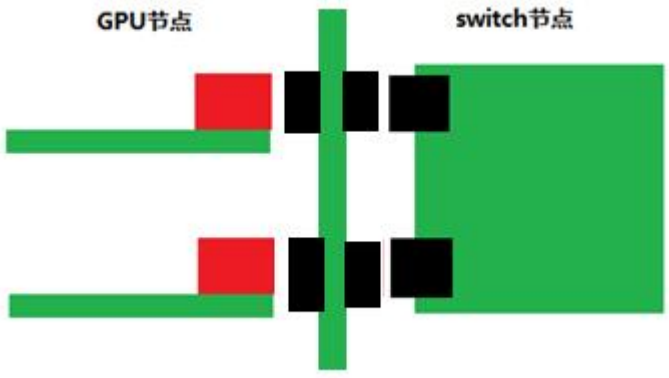
部分交叉可以在 cable 内实现，降低单板内部走线交叉

3.4.2 正交架构

3.4.2.1 中置背板方案（板对板，线对板，线对线）

正交架构中置背板方案结合了正交连接与中置背板的特点：业务板（计算板）与交换网板呈 90 度正交布置，两者通过中置背板上的正交连接器互联。

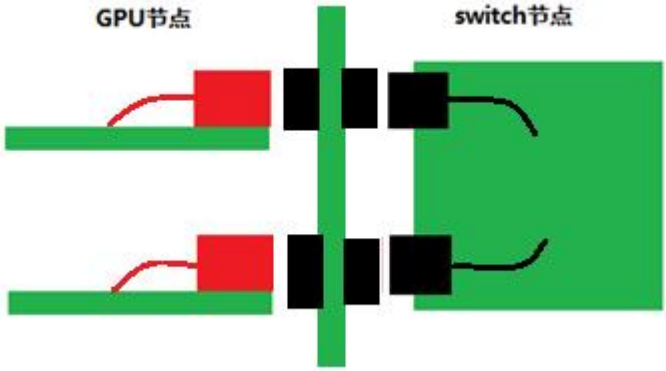
（1）板对板互配



正交连接器直接压接在中置背板上，业务板与交换板从前后两个方向正交插入，信号经正交连接器

+ 背板走线完成连接。其优势是正交连接器信号路径短、串扰小，相比平行连接器信号完整性更好；劣势是中置背板仍有一段走线，且背板阻挡气流。在 224Gbps 下，正交中置背板的板对板连接，因正交连接器的优良性能而优于传统背板，但背板走线仍是损耗瓶颈，适用于中等距离（< 15 英寸）场景。

（2）板对线互配

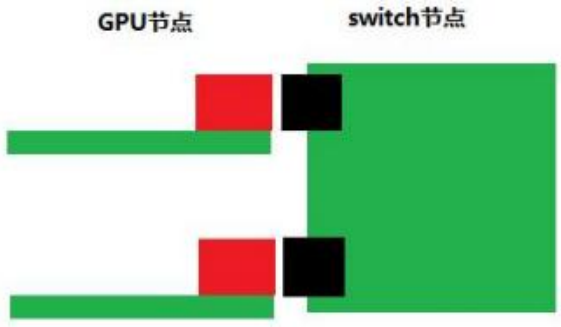


高速信号通过线缆从正交连接器引出，直接连接对端板卡。其优势是结合了正交连接的高密度与线缆的低损耗；劣势是结构设计复杂，正交连接器与线缆的集成难度高。在 224Gbps 下，该方案可支持较长距离，是高密机柜中兼顾密度与速率的可行方案。

3.4.2.2 无中板方案（板对板，线对板，线对线）

无中板方案（又称正交直连、零背板方案）完全取消了中置背板 PCB，业务板与交换板，通过正交连接器直接对插，信号路径仅包含板上走线 + 正交连接器，无任何背板走线。华为、锐捷、新华三、迪普等厂商的高端核心交换机均已采用此架构。

（1）板对板互配



业务板与交换板的正交连接器直接对接，信号不经过任何中间 PCB。其优势是信号路径最短、损耗最小、无背板阻挡气流（散热最佳）、可维护性好；劣势是正交连接器成本高，机械精度要求严格，前后板卡的深度空间需要精确配合。在 224Gbps 速率下，无中板方案的通道损耗通常可控制在 20dB 以内，是所有背板方案中最能满足 224Gbps 要求的架构，也是下一代高密 AI Rack 的首选。

连接器规格涉及弯母+弯公 OD

在 switch 与 GPU 互连 serdes pinmap 灵活度低，可实现性较差

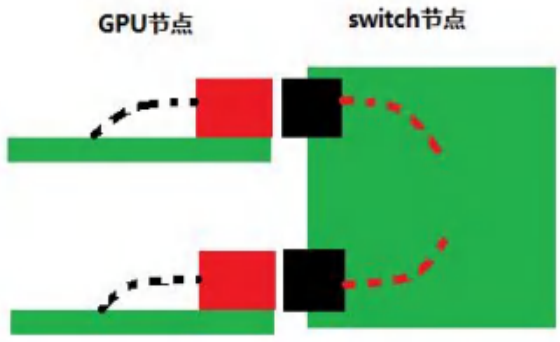
无背板对整机风冷散热影响相对较小

连接器需具备高可靠性抵消受力对性能的影响，节点增多导致 switch PCB 过长时连接器位置精度变差，且整个节点插拔力较大

节点内走线均通过 PCB，速率高时对芯片位置、板材等级要求高，PCB 成本受一定影响

典型机型:曙光 scale X40

（2）板对线互配



在无中板架构中，部分跨柜或长距离信号可通过线缆直接从正交连接器面板引出。其优势是结合了无中板的短距直连与线缆的长距能力；劣势是线缆从正交面板引出的空间有限。在 224Gbps 下，无中板 + 线对板的组合可灵活覆盖短距与中距场景。

节点内走线内侧可通过 cable，降低 PCB 损耗带来的影响，芯片布局、板材等级选用，相对灵活

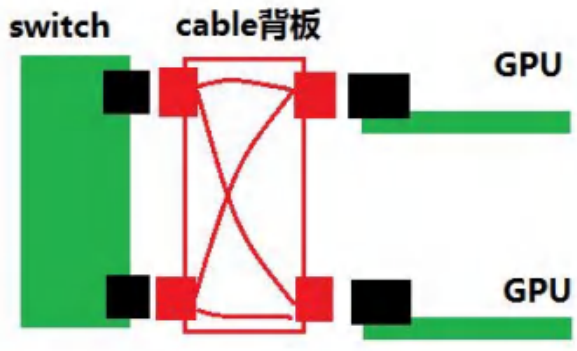
采用单侧 cable 或双侧 cable 设计，双侧 cable 设计时一侧连接器为固定状态，一侧浮动即可

典型机型：阿里磐久 64/128、HW 384 900 A3 超节点

3.4.3 混合正交和背板架构

混合正交和背板架构在同一机柜中结合正交连接与背板/线缆连接，根据信号速率与距离灵活选择不同的连接路径，兼顾高密度、高速率与可扩展性。

3.4.3.1 中置 Cable 背板方案



中置 Cable 背板方案以线缆矩阵替代传统 PCB 背板：在机柜中部设置线缆配线架，各板卡的高速信号通过预设长度的 Twinax 线缆在配线架内交叉连接。其优势是高速信号全部走线缆，损耗远低于 PCB 背板，可支持 224Gbps 下 1-2 米距离；线缆可按需配置，灵活性高；无 PCB 背板阻挡，散热较好。劣势是线缆密度大、装配与维护复杂、成本较高，且线缆的弯曲半径与气流影响需要精细设计。该方案是 AI 机柜中从传统背板向全光互连演进的重要过渡形态。

连接器规格涉及弯母+cable 公端

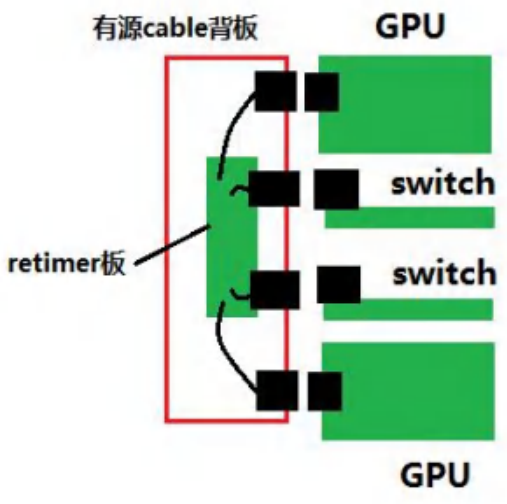
机柜 GPU 密度高，深度会加长，对于 cable 背板的维护相对较难

整机风冷散热影响相对较小

连接器单侧浮动设计，减小硬偏位带来的应力。结构设计难度相对较低。

部分交叉可在 cable 背板内部完成，减小 PCB 叠层压力

3.4.3.2 有源 Cable 背板方案



有源 Cable 背板方案在中置 Cable 背板基础上，于线缆两端或中间集成有源信号补偿器件（Retimer 或 Redriver），进一步延长 224Gbps 信号的传输距离。其优势是通过有源补偿可将 224Gbps 传输距离延长至 3-5 米，覆盖更大机柜范围；劣势是有源器件增加功耗（每条链路约 1-3W）、成本上升、故障点增加，且有源器件的散热需要额外考虑。在 224Gbps 下，有源 Cable 背板是长距离柜内互连的可靠方案，但功耗与成本 trade-off 需要仔细评估。

连接器规格涉及弯母+cable 公端

Cable 背板内放置 retimer 有源芯片，维护简单

整机风冷散热影响相对较小

连接器单侧浮动设计，减小硬偏位带来的应力。结构设计难度相对较低。

部分交叉可在 cable 背板内部完成，减小 PCB 叠层压力

背板与线缆连接方案综合对比

方案类型	信号路径	典型插损 (224Gbps)	散热	可维护性	成本	224Gbps 可行性
非正交背板-板对板	长 PCB 走线 (20-40")	30-40dB	差（背板阻挡）	中	低	困难，需 Retimer
非正交背板-线对板	线缆 + 短 PCB	15-25dB	中	中低	中	可行（短距线缆）
中置背板-板对板	穿越背板 PCB	28-38dB	差	中	中低	困难，需 Retimer
中置背板-线对板	线缆跨接	15-25dB	中	中低	中	可行
正交中置背板-	正交连接器+	20-30dB	中	中	中高	有条件可行

板对板	短背板					
正交中置背板-线对板	正交连接器+线缆	15-25dB	中	中	中高	可行
无中板-板对板	正交连接器直连	< 20dB	优（气流直通）	优	高	强烈推荐
无中板-线对板	正交面板+线缆	15-25dB	优	优	高	推荐
中置 Cable 背板	全线缆矩阵	15-25dB	良	中低	高	推荐（1-2m）
有源 Cable 背板	线缆+有源补偿	< 15dB（等效）	中	中低	很高	推荐（3-5m）

3.5 224G 背板和 Cable 背板设计要求

3.5.1 信号完整性要求

3.5.2 PCB 材料及工艺要求

3.5.3 机械和结构设计要求

支撑连接器的支架强度要满足变形要求，仿真变形推荐≤0.3mm。如连接器采用弹簧浮动设计，面板变形要求可降低。

Cable Tray 面板变形量需进行仿真模拟，根据面板变形量的仿真结构确定结构件面板材质及厚度。

如面板变形量过大，为减少面板变形量可通过在 cable tray 内部增加加强筋的方式支撑面板。

背板连接器设计成浮动连接器，单边浮动量建议≤0.5mm，连接器本身导向能力 – 单边浮动量≥1.0mm。

线缆连接器和支架前后方向设计较小的活动间隙保证连接器浮动，如下图，活动间隙设计在对端连接器插入时活动连接器后退到底和支架接触的另外一侧，活动间隙建议≤0.25mm。

连接器和线缆数量较多时，连接器出线处增加线缆支撑杆，用以克服线缆重力和拉力对连接器模组的拉扯，利于连接器浮动顺畅，防止极限情况下连接器卡死。

高度超过 600mmCable tray，建议在 Cable tray 中间设计固定位置与机柜横梁固定减少插拔过程中 cable Tray 面板的变形。

尽量减少 2 个 CABLE TRAY 同时和一个单板对插的设计。如果必须使用这种设计，2 个 TRAY 之间的位置公差要尽量控制小，浮动连接器浮动量需要增加，连接器浮动增加量要满足 2 个 TRAY 位置公差的增加量。

连接器 Z 向不受力的自由状态下，X 向和 Y 向浮动阻力≤15kgf/pcs（对 Cable 绑线进行约束）。

每个连接器浮动 X/Y 方向推荐满足单边 0.15mm min 浮动，Z 向浮动范围根据设计中过插要求制定。

相邻节点浮动结构件尺寸需要保证在最大浮动量时不能出现干涉。

浮动模块浮动量设计需进行机柜级公差分析，计算出浮动模块浮动量要求。

浮动结构件推荐采用不锈钢材质机加工零件，保证连接器与浮动结构件的位置精度以及强度。如有减重要求可使用高强度铝合金

为保证连接器 full mating，机柜节点连接器浮动结构件上可增加弹簧设计，插入一般按照过插设计，保证在过插时连接器不会损坏且处于完全配合状态或小于 SI 性能要求的最大 demating。

弹簧数量需根据浮动结构件尺寸均匀分布，保证过插时不同位置弹簧受力均匀，避免出现倾斜/扭

转状态。

连接器对插到位后继续施加压力，弹簧被压缩变形，实现过插；为保证弹簧在连接器对插到位后才发生变形，这里弹簧被压缩的力要大于连接器最大插入力。

弹簧压缩量一般需要满足 3mm 以上且弹簧的选择需根据连接器插拔力以及扳手设计的力相匹配，并提供弹簧在不同压缩量下的力便于节点扳手设计。

弹簧力设计时厂家需提供如下关键位置的弹簧力，便于确认弹簧设计合理性，且最大压缩量下需保证连接器不被插坏，对端连接器不被插坏且无掉件：

CABLE TRAY 比较重，需要考虑人工或者机械安装入机箱的方法。机箱上设计导向销辅助装配，导向销要可以承受 TRAY 重量不变形。

CABLE TRAY 和机箱之间精定位设计，可直接借用导向销或单独设计精定位。

CABLE TRAY 上需设计把手，辅助人工。把手的位置和尺寸要符合人工工程力学要求，把手可以承受 TRAY 重量不变形。

CABLE TRAY 建议使用不脱螺钉锁紧在滑道以及机柜的横梁上，紧固螺钉无遮挡，位置清晰，同时由于使用松不脱，还可以避免在紧固过程螺钉掉落遗失问题，便于安装维护 。

安装 CABLE TRAY 时需检查是否有其他零部件或线缆进入 CABLE TRAY 安装面内，影响节点与 CABLE TRAY 互配损坏连接器。

安装 CABLE TRAY 时一般需要将节点均拔出一定距离，建议拔出距离大于导向销长度，避免由于安装过程中的偏斜导致导向销撞击节点连接器 mating 面。

在保证结构件强度的前提下 Cable Tray 面板增加散热通风孔。

3.6 224G 系统其他要求

224Gbps 速率不仅对信号完整性提出挑战，也对系统级的热设计、可维护性、供电与结构设计提出了更高要求。本章从热设计与维护性两个维度，分析 224G 系统的特殊约束与设计要点。

3.6.1 热设计；

224Gbps 系统的热设计面临三重压力：一是 GPU 与交换芯片本身功耗持续攀升（单颗 GPU 功耗已达 1000W 以上，单颗 51.2T/102.4T 交换芯片功耗达 500-800W）；二是 224Gbps SerDes 每通道功耗较 112Gbps 显著增加（单通道功耗从约 10-15mW/Gbps 上升至 15-25mW/Gbps），高密 Tray 内数百条高速链路的聚合功耗不可忽视；三是高密连接器与背板区域气流通道狭窄，形成局部热点。

（1）风冷散热的局限与增强

传统风冷方案依赖前后贯通的气流通道，但在 224G 高密系统中，中置背板、线缆矩阵、正交连接器等结构均会不同程度地阻挡气流。非正交背板与中置背板方案因背板 PCB 横亘在风道中，风阻最大，通常需要提高风扇转速（增加噪音与功耗）或采用分区散热。无中板正交方案因气流直通，风阻最小，是风冷条件下的最优架构。增强措施包括：采用高风压风扇、优化机柜内部导流结构、提高进风温度标准（如 ASHRAE A3/A4 级）、在高功耗芯片处采用局部增强散热（如散热器内嵌均热板）。

（2）液冷散热的引入

当单机柜功耗超过 30-50kW 时，风冷已难以满足散热需求，液冷成为必然选择。液冷方案包括冷板式液冷（间接液冷）与浸没式液冷（直接液冷）。冷板式液冷将冷却液通过冷板直接贴合高功耗芯片（GPU、交换芯片、CPU），散热效率高且对现有架构改动较小，是当前 AI 机柜的主流方案；浸没式液冷将整个 Tray 浸没在绝缘冷却液中，散热能力最强但对连接器、线缆与器件的耐液性要求极高，且维护方式完全改变。在 224G 系统中，高速连接器与有源器件（Retimer、光模块）的散热是液冷设计

的难点：这些器件功耗密度高但体积小，冷板覆盖困难，部分方案采用导热垫和金属散热块将热量导至冷板。

（3）高密连接器区域的局部散热

224Gbps 高速连接器（尤其是正交连接器与高密背板连接器）通常布置在板卡边缘或机柜中部，周围空间紧凑、气流不畅。连接器自身的接触电阻在大电流下会产生热量，而相邻连接器的密集排列会形成热叠加。设计要求包括：连接器区域预留散热通道、采用金属外壳连接器增强热传导、在连接器附近布置温度传感器进行闭环监控、避免将高功耗有源器件紧邻高密连接器布置。

224G 系统散热方案对比

散热方案	散热能力	适用机柜功耗	对互连结构影响	维护复杂度	成本	224G 适用性
传统风冷（增强）	中	< 30kW/柜	依赖气流直通，背板方案受限	低	低	无中板方案可行
冷板式液冷	高	30-100kW/柜	需预留冷板与管路空间	中	中高	推荐，主流方案
浸没式液冷	极高	> 100kW/柜	连接器/线缆需耐液认证	高	高	高密场景可选
风冷+液冷混合	高	30-80kW/柜	芯片液冷+其余风冷	中高	中高	灵活，过渡方案

3.6.2 维护性要求

224G 系统的高密化与高速化使维护性设计更加关键。高速连接器的插拔次数、对中精度、更换时间直接影响系统可用性。AI Rack 通常要求 7×24 小时不间断运行，关键部件应支持在线更换（热插拔），且更换操作不应影响相邻业务。

（1）在线更换与热插拔

GPU Tray、Switch Tray、电源模块、风扇模块均应支持热插拔。对于高速互连链路，热插拔时需要确保：插拔过程中信号引脚的有序连接/断开（先接地、后信号、最后电源，或反之），避免瞬态高压损坏 SerDes；更换后链路自动重新训练（link training），无需人工配置；单 Tray 拔出不应导致整个柜内互连域瘫痪（应有冗余路径或流量快速切换机制）。224Gbps 链路的训练时间通常在数百毫秒至秒级，需要在设计中纳入高可用切换流程。

（2）前后维护与单侧维护

传统框式设备通常支持前后维护：业务板从前侧插拔，交换板与电源从后侧插拔。但在 AI 高密机柜中，液冷管路、线缆矩阵、正交连接器等结构可能限制单侧维护空间。无中板正交方案通常支持前后双向维护，业务板与交换板分别从前后两侧插入，互不干扰；中置 Cable 背板方案因线缆集中在机柜中部，前后维护时需注意线缆拉扯与弯曲半径。部分高密场景采用单侧维护设计（所有部件从前方插拔），可节省机柜后部空间，但对连接器与线缆的正面可达性要求更高。

（3）线缆管理与可访问性

224G 系统中高速铜缆与光纤的数量巨大（单柜可达数百条），线缆管理直接影响维护效率与散热。设计要求包括：采用线缆托盘（Cable Tray）分层布放，强弱电分离、高速与低速分离；线缆预留足够的维护冗余长度，但不影响气流；关键链路采用标签与色标管理，便于快速定位；采用模块化线缆组件（如预端接光缆、铜缆束），更换时整束插拔而非单根操作。对于中置 Cable 背板方案，线缆矩阵的可访问性是维护难点，通常需要设计可旋转或可抽出的配线架。

（4）故障定位与更换时间

224G 链路故障可能源于芯片、连接器、线缆或光模块等多个环节，需要完善的故障定位机制：SerDes 内置眼图监测与误码率统计、链路训练状态上报、连接器插拔次数记录、光模块数字诊断（DDM/DOM）。维护性指标通常要求：单 Tray 更换时间 ＜5 分钟，单条高速线缆更换时间 ＜2 分钟，故障定位到具体链路的时间 ＜1 分钟。

不同背板架构的维护性对比

维护维度	非正交背板	中置背板	正交中置背板	无中板正交	中置 Cable 背板
热插拔支持	支持	支持	支持	支持	支持（线缆需断电）
前后维护	前侧为主	前后双向	前后双向	前后双向	前后受限
单 Tray 更换时间	3-5 分钟	3-5 分钟	3-5 分钟	2-3 分钟	5-10 分钟
线缆可访问性	好	中	中	好	差(中部密集)
故障定位难度	中	中	低	低	高
对相邻业务影响	中	中	小	最小	中

3.7 高速连接器技术要求

高速连接器是 224G 系统中信号路径的关键节点，其性能直接决定通道的信号完整性与可达距离。随着交换芯片容量从 51.2T 向 102.4T、204.8T 演进，单芯片的高速差分对（Differential Pair, DP）数量从 1024 对向 2048、4096 乃至 8192 对扩展，对连接器的密度、性能与工艺提出了前所未有的要求。

3.7.1 接口密度发展趋势；

高速连接器的接口密度通常以每英寸差分对数（DP/inch）或每平方英寸差分对数（DP/inch²）衡量。过去十年间，背板连接器的密度从约 10-20 DP/inch 提升至 40-60 DP/inch，正交连接器密度可达 60-80 DP/inch。进入 224Gbps 时代，密度提升面临两大约束：一是高速信号对串扰的敏感性要求相邻差分对之间有足够的接地屏蔽，密度不能无限提升；二是连接器机械结构的物理极限（引脚直径、间距、塑料壁厚）制约了进一步微缩。

当前主流高速连接器产品的密度水平如下：传统背板连接器（如 Amphenol XCede、TE Strada Whisper）约 30-50 DP/inch；正交连接器（如 Molex Impel）约 50-70 DP/inch；新一代超高速正交连接器（如 Samtec AcceleRate、Amphenol ExaMAX2）目标密度达 80-100 DP/inch。在 224Gbps 下，为控制串扰与插入损耗，密度提升速度将放缓，更多通过优化接地结构与材料来实现。

厂商	产品系列	传输速率	节点密度	产品形态	产品状态
中航	GN7A	112G	2.2 × 4.0mm	1 件式	已量产
中航	GN9A	224G	3.2 × 3.0mm	1 件式	样品 / 推广中

厂商	产品系列	传输速率	节点密度	产品形态	产品状态
中航	GN9B	224G	3.0 × 2.5mm	1 件式	样品 / 推广中
安费诺	ultrapass	224G	3.2 × 3.0mm	1 件式	待确认
安费诺	XtremePass	300G	1.8 × 2.0mm	1 件式	样品 / 量产初期（224/448G CPC）
莫仕	impress	224G+	2.0 × 1.6mm	2 件式	样品 / 量产初期（压接式近 ASIC 方案，公开推广中）
立讯	KOOLIO	224G+	1.9 × 1.7mm	1 件式	样品 / 量产（DesignCon 2026 已展示 448G 演进）
华为	omnistack	224G+	1.8 × 2.0mm	2 件式	内部自研 / 未公开
华为	LUNA	112G－224G	1.8 × 2.0mm	2 件式	内部自研 / 未公开

3.7.2 2048、4096、8192 Differential Pair 设计需求

交换芯片的端口容量与差分对数量直接相关。以 224Gbps 每通道计算：102.4T 交换芯片需要约 1024 对高速差分对（1024 × 224Gbps ≈ 229T 双向，折合单向约 114T，考虑编码开销后匹配 102.4T）；204.8T 芯片需要约 2048 对；409.6T 芯片需要约 4096 对；未来 819.2T 芯片将需要 8192 对。如此大规模的差分对需要在有限的板卡边缘空间内通过连接器引出，对连接器密度与板级布线提出了极端要求。

（1）2048 DP 规模（对应 ~204.8T 交换）

2048 对差分对是 224Gbps 时代中高端交换芯片的典型规模。以 60 DP/inch 的正交连接器计算，需要约 34 英寸的连接器边缘长度，这在单块交换板上已接近极限。通常采用多连接器分段布置（如 4 个 512 DP 连接器），或采用双面连接器（板卡正反两面均布置连接器）。板级布线方面，2048 对差分对需要至少 16-24 层高密 PCB，走线区域集中在芯片与连接器之间，需严格控制等长与阻抗。

（2）4096 DP 规模（对应 ~409.6T 交换）

4096 对差分对已超出单块板卡的物理承载能力，通常需要采用多芯片级联或交换网板分离架构。一种方案是将 4096 DP 分布在 2-4 块交换板上，通过板间高速互联组成逻辑交换实体；另一种方案是采用正交中板架构，计算板与交换板分别承载部分差分对，通过正交连接器交叉连接。4096 DP 规模对 PCB 层数要求达 28-32 层，板材需采用超低损耗材料（如 MEGTRON 7、Rogers RO4000 系列），且过孔设计需采用背钻、盘中孔（VIPPO）等技术减少 stub 效应。

（3）8192 DP 规模（对应 ~819.2T 交换，前瞻）

8192 对差分对是下一代超大规模交换的前瞻需求，目前尚无单芯片可达到此规模。可能的技术路径包括：多芯片封装（如 3D 封装、chiplet 架构）将多颗交换芯片集成在同一封装内，减少板级互连需求；光互连前移（CPO/NPO/OIO）将高速电信号在芯片附近转换为光信号，减少电连接器需求；以及超高密正交连接器（目标 > 100 DP/inch）配合多板协同架构。8192 DP 规模下，纯电连接器方案将面临密度、散热与信号完整性的三重极限，光电融合是必然方向。

不同 DP 规模的设计需求对比

DP 规模	对应交换容量	所需连接器长度 DP/inch	PCB 层数	典型架构	224G 可行性
1024 DP	~102.4T	~17 英寸	12-16 层	单芯片+单连接器组	成熟商用
2048 DP	~204.8T	~34 英寸	16-24 层	单芯片+多连接器/双面	可行，设计紧凑
4096 DP	~409.6T	~68 英寸	28-32 层	多芯片级联/正交中板	有挑战，需超低损板材
8192 DP	~819.2T	~136 英寸	>32 层	3D 封装/光电融合/多板协同	前瞻，需技术突破

3.7.3 高速连接器性能指标；

224Gbps 高速连接器的核心性能指标包括插入损耗（Insertion Loss）、回波损耗（Return Loss）、串扰（Crosstalk）、特性阻抗（Characteristic Impedance）、时延（skew）等。这些指标共同决定了连接器在 224Gbps 下的信号传输质量。

1）插入损耗（Insertion Loss, IL）

插入损耗是信号通过连接器时的功率衰减，以 dB 为单位。224Gbps 下要求连接器在 Nyquist 频率（112GHz）处的插入损耗控制在 1.5-2.5dB 以内（具体取决于连接器长度与类型）。插入损耗主要来源于导体损耗、介质损耗与辐射损耗，优化手段包括：采用低介电常数、低损耗因子的塑料材料（如 LCP、PEEK）、优化接触件表面处理（如镀金厚度与粗糙度控制）、缩短信号路径。

（2）回波损耗（Return Loss, RL）

回波损耗反映连接器阻抗不连续导致的信号反射。224Gbps 下要求在 DC 至 112GHz 范围内回波损耗优于 15-20dB。阻抗不连续主要发生在接触点、过孔、引脚弯曲处，优化手段包括：精确控制差分对的几何尺寸与间距、采用阻抗补偿设计（在接触区域调整线宽）、减少过孔 stub（采用背钻或盘中孔）。

（3）串扰（Crosstalk）

串扰是相邻差分对之间的信号耦合，包括近端串扰（NEXT）与远端串扰（FEXT）。224Gbps 下要求串扰控制在 -35dB 至 -45dB 以下。串扰优化手段包括：差分对之间设置接地屏蔽针（ground shield）、采用正交排列（相邻差分对呈 90 度，耦合最小）、优化接地针的分布密度。正交连接器因信号方向正交，天然具有低串扰优势。

（4）特性阻抗与眼图

差分对特性阻抗目标值通常为 85Ω 或 100Ω（差分阻抗），偏差控制在 ±10% 以内。眼图是信号质量的综合体现，224Gbps PAM4 信号要求眼高（Eye Height）与眼宽（Eye Width）满足接收端阈值，通常要求眼图张开度 > 30%。连接器的性能最终需通过系统级眼图仿真与实测验证。

高速连接器关键性能指标要求

性能指标	符号	224Gbps 要求	优化方向

插入损耗	IL	IL≤3.6dB@53.125GHz(3.5mm 孔长，板材：M8 级别)，带宽~70GHz	低损材料、短路径
回波损耗	RL	≤-15dB(0-28GHz) ≤-12dB(28-56GHz) ≤-10dB(56-70GHz)	阻抗补偿 背钻
近端串扰	NEXT	≤-65dB(0-28GHz) ≤-50dB(28GHz-56GHz) ≤-40dB(56GHz-70GHz)	优化屏蔽 正交排列
远端串扰	FEXT	≤-65dB(0-28GHz) ≤-50dB(28GHz-56GHz) ≤-40dB(56GHz-70GHz)	优化屏蔽 介质均匀
差分阻抗	Zdiff	nominal：90ohm demating 0mm：-6/+6 Ohm demating 1mm：-6/+10 Ohm offset X/Y=0.5mm，阻抗变化-7+/7 Ohm	调整间距 公差控制
对内时延	Skew	≤1ps	等长设计

3.7.4 高速连接器组装工艺要求；

高速连接器的性能不仅取决于设计，更高度依赖组装工艺的一致性。224Gbps 下，微小的组装偏差（如共面度超标、对中偏移、焊接空洞）都可能导致阻抗不连续或接触电阻增大，从而劣化信号完整性。

（1）压接 vs 焊接

高速连接器与 PCB 的连接方式主要有压接（Press-fit）与焊接（SMT/Solder）两种。压接连接器通过弹性引脚压入 PCB 镀通孔实现电气连接，无需焊接，适合背板连接器与正交连接器，优点是可返修、无焊接热应力，但压接孔的加工精度与压入力控制要求高；SMT 焊接连接器通过表面贴装焊接在 PCB 焊盘上，适合密度更高的板对板连接器，优点是信号路径短（无过孔 stub），但焊接质量（空洞率、共面度）控制难度大，且返修困难。224Gbps 下，SMT 连接器因无过孔 stub 而具有信号完整性优势，但对 PCB 焊盘平整度与回流焊工艺要求极为严格。

（2）共面度与对齐精度

多引脚高速连接器的共面度要求通常控制在 0.05-0.10mm 以内，确保所有引脚同时与 PCB 良好接触。对于 SMT 连接器，共面度超标会导致部分引脚虚焊；对于压接连接器，共面度超标会导致部分引脚压入深度不足。对中精度要求正交连接器的公母座对中偏差 ＜0.1mm，否则会导致接触电阻增大或引脚变形。组装时需采用高精度定位治具与视觉对位系统。

（3）自动化装配与检测

224Gbps 高密连接器的引脚数可达数千，人工装配已无法满足质量与效率要求，必须采用自动化装配：自动压接机（压接连接器）控制压入力与深度，高速贴片机（SMT 连接器）实现精密贴装，回流焊炉实现精确温度曲线控制。装配后的检测包括：X 射线检测（检查焊接空洞率与压接质量）、AOI 自动光学检测（检查贴装偏移与桥连）、TDR/阻抗测试（抽检信号完整性）、高压绝缘测试。关键批次应进行 100% 高速信号测试（如眼图或误码率测试）。

（4）材料与表面处理

连接器接触件通常采用铜合金 (如铍铜、磷青铜) 作为基材, 表面镀镍金 (Ni/Au) 或镀钯金 (Ni/Pd/Au) 以提高耐磨性与导电性。224Gbps 下，镀层的粗糙度与厚度会影响高频信号的集肤效应损耗，要求镀

金层厚度均匀（通常 0.05-0.15μm）、表面粗糙度 Ra < 0.2μm。连接器塑料本体需采用低介电损耗材料（如 LCP 液晶聚合物，Df < 0.002；或 PEEK，Df < 0.003），且在高温回流焊后保持尺寸稳定。

高速连接器组装工艺对比

工艺维度	压接（Press-fit）	SMT 焊接	混合工艺
信号完整性	有过孔 stub，损耗略大	无 stub，路径短，优	按需分配
共面度要求	≤ 0.10mm	≤ 0.05mm	分别控制
返修性	好（可退针重压）	差（需拆焊）	部分可返修
装配效率	中（需专用压接机）	高（SMT 产线）	中
检测方式	X 射线+压入力监测	X 射线+AOI+SPI	组合检测
适用连接器	背板、正交、大引脚数	板对板、高密小引脚	复杂背板系统

3.8 双 Rack 互连方案

3.8.1 DAC 应用

DAC 为纯无源方案，电缆组件内不含任何有源器件（无 CDR、无 Retimer、无均衡芯片），通过线缆线对的物理特性与连接器端接实现信号传输。业内 DAC 产品已覆盖 SFP、DSFP、SFP-DD、QSFP、QSFP-DD、OSFP、OSFP-XD 等全系列封装，单通道速率支持 112G PAM4 并正向 224G PAM4 演进，采用超低损耗氟塑料绝缘与优化绞合节距设计，有效控制 IL（插入损耗）与 RL（回波损耗）。

3.8.1.1 技术特点；

维度	特性描述
功耗	近零功耗（<0.1W/端口），是三种方案中功耗最低的互连形式，无需考虑散热设计
成本	单位带宽成本最低，且不含易失效的有源器件
延迟	信号直接传播，无信号处理时延（0ps/m）
可靠性	MTBF 高，无源结构抗振动、抗电磁干扰能力强
传输距离	受插入损耗预算严格限制，典型链路预算下距离最短
灵活性	支持 26AWG、28AWG、30AWG、32AWG 等多种线规组合，可实现线缆外径与弯折半径的灵活平衡

3.8.1.2 应用场景

DAC 是机柜内互连（intra-rack）的绝对主力，典型场景包括：

（1）ToR 交换机与服务器/NIC 互联：同机柜或相邻机柜的叶子交换机-服务器直连，距离通常<=3m；

（2）邻近交换机间互联：同个柜内的两个交换机间，或跨一两个柜的交换机间互联；

（3）存储阵列、GPU 集群内部短距互联；

（4）5G 基站 AAU-BBU 短距传输（部分前传场景）；

（5）实验室验证、产线测试与设备管理口等对成本敏感、距离可控的场景。

3.8.1.3 设计要求

（1）链路预算与长度设计

DAC 链路预算须满足 IEEE 802.3 标准规定的通道插入损耗与回波损耗模板（如 802.3ck 对应 112G/lane 的规范）；

长度选择：在 112G 速率下最长可支持到 2 米，在 224G 速率时最长可支持到 1 米。设计时须预留 3~5dB 的系统裕量（含连接器损耗，如 112G 时每端连接损耗约 1~1.5dB@28GHz）；

通过 S 参数（S4P 模型）进行全通道仿真，确保在奈奎斯特频率点（112G PAM4 约 28GHz）的 COM（通道工作裕量）> 3dB。

（2）PCB 与连接器协同设计

主机侧 PCB 走线应采用低损耗板材（如 Megtron 6/7 级别），走线长度尽量压缩至 <3 英寸；

连接器 footprint 采用 GSSG/GSG 优化焊盘结构，降低阻抗不连续；

线缆端接处采用 360°屏蔽端接，保证 EMC 性能满足 CISPR 32 Class B。

（3）机械与散热设计

弯折半径：静态弯折半径≥线缆外径的 5 倍，动态弯曲半径≥线缆外径的 10 倍；

机柜内线缆布放须避免与 220V 电源线平行走线（间距>30mm），必要时采用屏蔽隔板；

无需散热设计，但须保证空气流通，避免长期工作温度超过 70℃（导体温度上升会导致 IL 劣化）。

（4）兼容性要求

须与主流交换芯片（Broadcom Tomahawk 系列、NVIDIA Spectrum 系列、Marvell Teralynx 等）的 SerDes 配置矩阵进行兼容性验证，包括 CTLE 档位与 DFE 抽头容限。

（5）可靠性验证

执行 EIA、Telcordia、SFF 等行业可靠性标准：插拔寿命（Telcordia GR-1217-CORE & EIA 364-09）、循环湿热（EIA-364-31, method III, condition B, omitting 7b vibration test）、热冲击（EIA 364-32 test condition I）等测试

3.8.2 AEC 应用

AEC 在线缆两端集成有源信号处理芯片（通常为全功能 DSP/CDR），对信号进行重定时与整形，从而将铜互连的有效传输距离大幅延伸。业内 AEC 产品已覆盖 QSFP、QSFP-DD、OSFP、OSFP-XD 等主流封装，单通道速率支持 112G PAM4 并正向 224G PAM4 演进

3.8.2.1 技术特点

维度	特性描述
功耗	>8W/端口（随速率与芯片架构变化），显著高于 DAC 和 ACC，但仅为同规格 AOC 的 约 50%～70%
传输距离	借助有源 DSP 的时钟数据恢复处理，可达 7m 以上长度（112G/224G 速率），覆盖机柜间互连需求
信号完整性	内置 DFE/CDR 主动补偿通道损耗，对主机侧 PCB 损耗容忍度更高，降低主板设计难度
成本结构	较 DAC 增加芯片 BOM 成本，但省去 AOC 的光引擎与激光器，在 <10m 区间具有最优性价比
延迟	增加 DSP 处理延迟（通常每端 <50ns），对大多数以太网业务无感知，对 RDMA 场景需纳入时延预算
供电	通过接口金手指从主机取电（3.3V），需关注主机的端口供电能力（单端口供电预算）
诊断能力	高端 AEC 支持 DDM 数字诊断功能，可实时监测温度、电压、BER 预 FEC，支持预测性维护

3.8.2.2 应用边界

AEC 定位于"铜互连的极限延伸"，其应用边界清晰：

（1）适用边界：

a. 链路长度 1～7m 的 ToR 交换机 ↔ 服务器互连（跨机柜/同排机柜）；

b. 112G/224G 高密度部署中，DAC 无法满足 IL 预算或小 OD 要求的场景；

c. 主机板 PCB 损耗较大、无法为 DAC 提供足够通道裕量的存量系统升级；

d. Spine-Leaf 架构中相邻机柜的 Leaf-Spine 短距上联。

（2）不适用边界（建议转向 LPO/AOC）：

a. 超过 10m 的场景：铜缆衰减随频率与距离呈指数恶化，AEC 芯片的均衡能力存在物理极限，>10m 后 BER 裕量与功耗急剧恶化，应选用 LPO 或 AOC；

b. 严苛延迟敏感场景：如部分 HPC 的 MPI 紧耦合通信、金融低延迟网络，DSP 引入的固定延迟需逐案例评估；

c. 端口供电受限场景：当交换机 ASIC 或线卡为低功耗设计（LPO 交换机）且端口供电预算紧张时，双端取电的 AEC 可能与系统供电策略冲突；

d. 高温环境：有源器件结温受限（通常 ≤85℃），高温机柜（>45℃进风温度）会降低芯片寿命与均衡余量

3.8.2.3 设计要求

（1）功耗与热设计

系统级热仿真须纳入 AEC 芯片功耗（按最恶劣工况 18W/端口估算），确认交换机散热器与风道设计满足芯片结温要求；

采用 V 型或平行分支结构（breakout）时，功耗随分支数线性增加，须校核线卡单槽位总功耗。

（2）供电与电源完整性

校核主机端口 3.3V 供电能力，多根 AEC 同时插拔时的浪涌电流需满足热插拔规范；

AEC 内部 DC-DC 布局应靠近金手指端，电源平面完整性通过 PDN 阻抗分析验证（目标阻抗 <10mΩ@10MHz ~ 100MHz）。

（3）互操作性与固件管理

AEC 须符合 CMIS（Common Management Interface Specification）规范，支持 EEPROM 信息上报与固件在线升级（部分方案）；

与交换机 NOS 的端口配置协同：主机侧 SerDes 通常需设置为"长通道"模式，关闭部分 CTLE 加重以避免过均衡。

（4）EMC 与结构设计

有源器件工作频率高（DSP 内核数 GHz 级），PCBA 须采用 10 层板以上设计并整体金属屏蔽罩包裹，确保辐射发射满足 CISPR 32；

外壳散热孔设计兼顾散热与屏蔽效能（SE > 60dB@1GHz）。

（5）可靠性验证

除行业标准测试外，还要执行 GR-468 有源器件可靠性标准：高低温循环 (-40℃ ~ +85℃, 96cycles)、高温寿命（70℃，2000H）、ESD（HBM 2kV）等测试

3.8.3 ACC 应用

ACC 在行业内通常指代与 AEC 同属有源铜互连范畴、但采用线性放大（Linear Redriver / Equalizer）或特定高集成度均衡架构的有源铜缆方案。与 AEC 侧重"重定时"不同，ACC 侧重以更低功耗、更低延迟的方式延长铜链路，是 LPO（线性可插拔光学）浪潮下铜互连的有源化演进方向。业内 ACC 产品已覆盖 SFP、QSFP、QSFP-DD、OSFP 等少数封装，单通道速率支持 112G PAM4 并正向 224G PAM4 演进。ACC 采用线性均衡芯片（Linear EQ / Redriver）架构，对信号进行模拟域的连续时间线性均衡（CTLE）与增益补偿，不含全功能 DSP 与 CDR 重定时环节。

3.8.3.1 技术特点

维度	特性描述
功耗	<= 2W/端口，介于 DAC 与 AEC 之间，较 DSP 架构 AEC 降低 60%以上
延迟	信号处理时延极低（每端 <100ps），无 CDR 重定时，几乎不引入额外抖动积累
传输距离	有效延伸铜互连长度至 3~4m (112G 速率下最长 4m，224G 速率下最长 3m)
成本	芯片 BOM 成本低于 AEC，可作为 DAC 与 AEC 之间的过渡选择
抖动特性	保留主机 SerDes 的原始抖动特性，依赖主机均衡能力与线性芯片的协同，对系统设计匹配性要求更高
标准趋势	与 IEEE 802.3 LPO 标准及 OIF Linear Equalization 规范协同发展，是 800G/1.6T 短距互连的重要候选方案

3.8.3.2 应用场景

800G/1.6T AI 集群机柜间互连：当 800G DAC 因 IL 预算无法覆盖 3 ~ 4m 链路时，ACC 以可接受的功耗增量实现无缝覆盖；

（2）LPO 交换机配套场景：与 LPO 光模块技术路线对应的"线性化"铜互连需求，保证端到端通

道均为线性架构，规避 DSP-DSP 串联带来的稳定性风险；

（3）延迟敏感型算力网络：RDMA/RoCEv2 网络、GPU 集群参数同步通信等对亚微秒级时延敏感的场景；

（4）高密度机柜（功率密度>30kW/rack）：散热压力大的场景下，ACC 的低功耗特性优于 AEC；

（5）成本敏感的 112G、224G 速率的大规模部署：替代 AOC 实现 3～4m 覆盖，降低光互连采购成本与备件复杂度。

3.8.3.3 设计要求

（1）线性链路协同设计（最关键要求）

ACC 的线性均衡架构要求主机 SerDes 的 TX FFE（前馈均衡）与芯片 CTLE 进行端到端联合优化，用户主机侧 SerDes 配置参数表（FFE 抽头系数、摆幅设置），并在系统验证阶段进行眼图与 BER 扫描；

须进行全通道 COM/眼高眼宽仿真，纳入线性芯片的 S 参数模型（含封装与 PCB 寄生效应）；

链路两端芯片参数须匹配（对称链路设计），非对称布放会显著压缩裕量。

（2）功耗预算管理

按 2W/端口进行系统级功耗核算，大规模部署时（如 32 端口线卡）单卡新增 64W 功耗须纳入整机散热设计；

支持按温度自适应调节均衡增益的功能（Adaptive EQ），高温下自动优化功耗与性能平衡。

（3）机械与线缆工艺要求

ACC 线缆段仍须满足无源铜缆的工艺标准：发泡度 >40% 的 FEP 绝缘以降低介电常数（ε_r <1.7）；

屏蔽结构采用双层编织+铝箔复合屏蔽，转移阻抗 <50mΩ/m@30MHz，抑制芯片高频辐射通过线缆向外辐射；

连接器端壳体须为压铸金属并整体导电搭接，保证 360°端接屏蔽连续性。

（4）验证与认证要求

眼图测试：802.3 规范眼罩（Eye Mask）合规性测试须在最差工艺角（Worst-case PVT）下验证；

误码测试：224G 链路 BER < 1e-6（pre-FEC 误码率）、FEC 纠后 BER < 1e-12，测试时长 ≥24h 覆盖温度循环工况；

互操作性测试（IOP）：与主流交换芯片平台完成多批次、多温度点的互通矩阵验证；

热插拔与 DDM 管理：符合 CMIS 5.x 规范，热插拔冲击电流 < 规定限值，EEPROM 信息包含厂商 ID、线缆长度、温度告警门限。

（5）可靠性验证

除行业标准测试外，还要执行 GR-468 有源器件可靠性标准：高低温循环 (-40℃～+85℃，96cycles)、高温寿命（70℃，2000H）、ESD（HBM 2kV）等测试。

4. 第四章 多 Rack 高速互连方案

本章节主要集中在多 Rack 高速互连方案进行，尤其是当 AI 计算系统突破单机柜功率与散热的物理约束后，如何在多个机柜（Rack）之间构建高速互连体系涉及到的拓扑、技术路线及前沿技术积累等。

多 Rack 架构将多个机柜通过 Scale-out 网络组织为大规模 AI 集群，机柜间互连的规模、带宽与拓扑结构决定集群的整体可扩展性与训练效率。多 Rack 网络通常采用 Clos（Leaf-Spine/Fat-Tree）、Dragonfly 等规整拓扑，将各 Rack 作为叶子节点接入汇聚/核心交换层，实现无阻塞或近无阻塞的大规模互连。

4.1 多 Rack 网络拓扑分析

首先，机柜间互连的物理距离通常在 3 至 15 米量级，远超机柜内铜缆的可用范围。在 112G 世代，6 类 DAC/AEC 尚可勉强覆盖这一距离；但在 224G 世代，单通道 PAM4 的介质与导体损耗显著上升，被动 DAC 必须通过加粗线规（低 AWG）来压降阻抗，这会同时恶化线缆的体积、重量与最小弯曲半径，最终使被动铜缆的可行距离落入 2 至 3 米的量级。换言之，跨机柜互连在 224G 下从"介质选择问题"变成了"必然的光互连问题"——这一转换不是工程偏好的选择，而是信道预算的数学结果。

其次，Lambda 在评估 NVIDIA GB300 NVL72 规模 CPO 交换机时给出了一组值得反复引用的数据：在三层 800G 架构中，后端网络已占数据中心网络总功耗的 86%。当网络功耗达到该比例时，网络不再是"配套系统"，而是决定集群 token 吞吐量的主要负载之一。这意味着多 Rack 拓扑的设计目标中，降低每比特传输能耗与降低跳数同等重要。

另外，跨机柜互连同时承载两类性质完全不同的流量：一是集合通信（all-reduce、all-to-all），要求高带宽、低时延、对称性；二是 Scale-out 侧的存储与参数加载流量，要求高并发、低成本。二者对最优拓扑的偏好并不一致。

由此，多 Rack 拓扑设计可以形式化为一个受约束的优化问题：在给定的 scale-up 域边界、SerDes 速率、交换芯片 radix 与端口速率约束下，选择拓扑族与收敛比，使得集群在给定作业负载分布下的模型算力利用率（MFU）最大化，而约束是端口总数、每节点功耗与光模块数量最小。

多大的计算规模应当被纳入一个强耦合的 scale-up 域？这个边界一旦确定，域内拓扑与域外（scale-out）拓扑就基本解耦。该边界由物理层直接决定。铜缆在机柜内可以提供 TB/s 级的低功耗互连，但一旦出柜就必须转光，转光环节引入约 200ns 量级的转换开销与一次光电转换功耗。因此，在 224G 世代，scale-up 域边界的决定因素正在从"物理距离约束"转向"协议地址空间约束"。物理上，跨柜必须转光；但只要协议能提供足够的地址空间与带宽，光域并不必然成为规模扩张的障碍。

多 Rack 网络设计需重点权衡三个维度：一是带宽与无阻塞能力，保障任意两个 Rack 之间的通信带宽；二是跳数与时延，降低跨 Rack 通信的端到端时延；三是可扩展性与成本，使网络规模增长时端口与布线开销可控。Clos 架构在理想无收敛配置（收敛比 1:1）下，任意两点通信均可获得全带宽，但对交换芯片、链路与光模块的资源需求最高；工程落地中通常接受轻度带宽收敛（收敛比 1.2:1~2:1），以适度牺牲部分带宽保障换取硬件成本优化，设计中需明确收敛比，并对网络热点流量实施监控。

多 Rack 网络与机柜内 Scale-up 域协同设计：Scale-up 域以 NVLink/UALink 等高带宽低时延互连满足紧耦合计算，Scale-out 域以以太网/RoCE（或 NPO/CPO 光互连）满足跨 Rack 数据交换。两者共同构成“Scale-up + Scale-out”的混合组网架构，是 AI 超节点与多 Rack 集群的主流形态。

当前主流拓扑族包括：

·Clos /胖树（Fat-tree）拓扑。以 leaf-spine 两层或三层结构实现无阻塞，流量特征是对称、可预测，工程实现与运维复杂度最低。代价是跳数随规模线性增加，且 Spine 层端口数需求随节点数线性增长。在 224G/1.6T 端口速率下，Clos 的端口需求量是成本的主要驱动力。

·nD Torus 拓扑。Google TPU 谱系的核心选择。TPU v4 采用 4×4×4 = 64 芯片为一个 cube 的 3D

torus 结构，每芯片 6 条 ICI 链路以缩短网络直径；整个 pod 含 64 个 cube、4096 颗芯片。Torus 的优势是链路数固定（与节点数成线性而非平方关系）、光纤用量低，且非常规网络直径小；代价是存在对角线（bisection）带宽受限，且必须引入 wrap-around 链路，而 wrap-around 在物理上要求跨机柜互连。

4.2 NPO/CPO 技术路线

随着 AI 超节点规模扩大与互连带宽持续提升，传统可插拔光模块与高速铜互连面临功耗、端口密度与传输距离的多重瓶颈。NPO（Near-Packaged Optics，近封装光学）与 CPO（Co-Packaged Optics，共封装光学）通过将光引擎靠近甚至与交换芯片共封装，缩短高速电链路、降低功耗与时延，是下一代多 Rack/AI 超节点全光互连的核心技术路线。

224G 世代最关键的架构变化：光电转换级的物理位置。其核心：224G SerDes 是把光互连从"交换机前面板"推向"芯片封装内"的物理触发点。

在 112G 世代，光模块作为可插拔模块安装在交换机前面板，交换 ASIC 通过主板走线（15 至 30 cm）连接到模块内的 DSP。DSP 对这条长电通道做均衡与重定时。当单通道速率翻倍至 224G，这条通道的插损预算迅速恶化——可插拔方案的信道损耗可达 20 dB 以上，必须依赖 DSP/Retimer 进行补偿，而 DSP 本身带来 15 至 25 pJ/bit 量级的功耗与固定时延。

NPO（Near-Packaged Optics，近封装光学）与 CPO（Co-Packaged Optics，共封装光学）的共同做法，都是把光电转换级向交换 ASIC 靠拢，从而缩短这条通道。

4.2.1 系统架构

NPO 将硅光光引擎板载部署于主板、紧邻交换芯片，光引擎与交换芯片物理独立、可热插拔更换，通过板内短距高速电链路与交换芯片相连；CPO 则将硅光光引擎与交换 ASIC 共同封装在同一基板（采用 2.5D/3D 先进光电共封装技术），将电链路压缩至毫米级，实现光电深度集成。

在系统架构上，NPO/CPO 部署于超节点/AI Rack 内部，填补高速铜互连（112G/224G/448G）与传统可插拔光模块之间的技术空白。超节点内部 Scale-up 域优先评估高速铜互连，当电链路插损无法满足信号完整性要求时，切换至 NPO/CPO 光互连；跨机柜/多 Rack 的 Scale-out 域则以光互连承载更长距离传输。以国产 3.2T NPO 交换芯片为例，系统采用硅光光引擎配合先进封装，端到端路径由交换 ASIC→基板/板内走线→光引擎→光纤/连接器→对端光引擎构成。

ASIC 与光电转换级的距离是影响系统架构的关键，功耗、密度、良率、维护成本等所有其他差异，都受这个关键因素影响，而不是并列的独立选择。理解这一点，NPO/CPO 的技术选择就从多选一变成了单一参数寻优。

关于 NPO 的定义边界，行业内存在两种并行口径：口径 A (广义，含 OBO) 指光引擎在同一 PCB 上但未与 ASIC 同封装，走线 5 至 7 cm，代表为 OIF 的 3.2T NPO 标准封装（LGA 连接器，22.5 × 35.1 mm）；口径 B（狭义）指光引擎与 ASIC 同基板封装、走线小于 1 至 2 cm。本章采用"光引擎紧邻 ASIC、板级可插拔"作为工作定义，并在涉及具体产品时注明其所属口径——这一处理是必要的，因为两个口径下的功耗与维护性数据不可直接比较。

4.2.1.1 外置光源：硅光架构的必要妥协

硅是间接带隙材料，不能高效发光，因此所有硅光链路都必须外挂 III-V 族（InP）激光器。CPO 把光引擎放到发热 ASIC 旁边之后，激光器成为热管理死结——激光器的效率与寿命对温度高度敏感，

而它恰好被放在了温度最高的区域。主流解法是把激光整体移出封装，做成前面板可插拔的外置光源模块（ELSFP）。这是 CPO 架构中唯一把可维护性重新找回来的设计。

存在三种并存的光源布局，分别如下：

· 外置可插拔 ELSFP。代表为 Broadcom TH6-Davisson 与 Lumentum ELSFP。可现场更换，激光远离热源；代价是前面板仍占位，且需额外管理光纤阵列。

· 外置高功率连续波光源 + 波分复用。代表为 Lumentum UHP 平台。一台光源可扇出更多引擎，通过波长复用提升密度。其难点在于光纤阵列复杂度显著上升。

· 内置光源（光引擎自带激光）。代表为华为 Hi-ONE。免除了外部光纤耦合与保偏光纤要求；代价是光源热量压在引擎内部，散热压力大。

4.2.1.2 封装内、板级与面板级三层布局

封装内布局是 CPO 的主流形态，包含两种集成方式。2.5D 方式将 EIC 与 PIC 并排贴装在无源硅中介层上，经微凸块或铜柱互连，其寄生参数优于 PCB，但信号仍需跨越两个凸点界面，限制了电路径的进一步缩短。3D 混合键合方式将 EIC 直接堆叠在 PIC 上，封装寄生最小，TSMC COUPE 与 NVIDIA 的合作路线属此类。

板级布局是 NPO 的典型形态。典型做法是通过 LGA 插座连接，使光引擎与主芯片在物理与电气上解耦。参考实现是 ODCC 的 102.4T NPO 参考交换机（TCS9610-128O0）：采用 Broadcom TH6 ASIC, 4RU 高度, 128 个 800G MPO16 接口, die-to-die 通道损耗低于 16 dB VPD, 支持 3.2T 级 NPO 模块，多个 ELSFP 远程光源可现场更换，配 2+2 冗余电源与 7+1 冗余风扇，液冷为后续版本。

4.2.2 应用优势和挑战

NPO/CPO 的核心优势在于：缩短高速电链路，降低信号损耗与时延；提升端口密度，突破前面板端口带宽瓶颈；降低光互连功耗与整机功耗。其挑战在于：先进封装良率与散热管理、光引擎与交换芯片的协同设计、标准化尚未完全收敛（业界多套规范并存）、以及产业链成熟度与成本控制。CPO 把电/光域耦合进同一封装，良率与维护是主要风险；NPO 保留可插拔模组，是更稳健的过渡路线。

CPO 的能效收益已经被验证时可交付的。

· 整机级对比。NVIDIA Quantum-X InfiniBand Photonics Q3450-LD 为 4U 液冷交换机，配备 144 个 800G InfiniBand 端口，交换容量 115.2 Tb/s 无阻塞转发，采用四颗 Quantum-X800 ASIC 构成多平面交换架构（每颗提供 28.8 Tbit/s）。传统可插拔方案同类交换机功耗约 7.0 kW，Q3450-LD 为 3.95 kW，单台节省 3.05 kW，能效提升约 1.77 倍。

· 集群级换算。Lambda 给出的集群级换算：41,472 GPU 集群规模下，采用 1,440 台 CPO 交换机可释放 4,392 kW 电力，按 Blackwell Ultra GPU 1,400 W TDP 折算，相当于额外运行 3,137 块 GPU；在 12.8 万 GPU 数据中心规模下，传统方案需要约 65.5 万个独立光收发模块，CPO 显著减少了光学组件数量，从而减少潜在故障点。

· 信道损耗的量级跃迁。CPO 将电路径从厘米级缩短至微米级，信号损耗从约 20 dB 降至约 4 dB，同时 DSP 被移除，每一跳 fabric 的功耗与时延贡献同时下降。

· NPO 侧的量产实证。华为在昇腾 960 超节点上采用自研 NPO 光引擎 Hi-ONE，这是目前公开信息中全球最完整的 NPO 量产实证：单个超节点使用 5,500 个 Hi-ONE 模块，替代原本需要的 4.8 万颗 800G 光模块，功耗降低超过 550 kW，系统无故障运行时间提升一倍，系统可用度达到 99.8%。Hi-ONE 单引擎传输容量达 7.2 T，公开口径为业界首个量产 NPO 产品与唯一实现内置光源的 NPO 产品【来源：华为全联接大会，2026-09-17】。华为 NPO 价值在"省电"的同时实现了"高带宽高可靠"与"低时延低功耗"的统一——其中"系统无故障运行时间提升一倍、可用度 99.8%"这一项，是把光互连

的可靠性问题从"单个模块 MTBF"提升到"系统级可用度"来衡量的范例。

与能效收益同等重要的是，CPO 的挑战同样具体。

· 挑战一：Retimer 存废由损耗阈值决定，而非由"CPO"这个标签决定。224G 下的损耗阈值约为 10 至 13 dB。NPO 的典型 die-to-die 损耗（约 13 至 16 dB）恰好压线甚至越线，CPO（小于 7 dB）则有余量。

· 挑战二：热管理与光纤管理的耦合复杂度。Google 在 Hot Chips 2026 报告早期试点中，CPO 与 OCS 组合引入了热管理与光纤管理的新问题，导致规模化进程放缓。

· 挑战三：标准与生态碎片化。同一"NPO"概念下至少存在三套并行标准或接口定义（基于 LGA 的 NPO 封装、OIF 3.2T CPO Module IA、ELSFP 外置光源接口），叠加华为 OPEN NPO 与欧美 OIF 路线，产业生态呈现明显分化。

· 挑战四：产能与供应链约束。高功率连续波激光器的产能仍是硬约束；CPO 将大量光学组件集成到单一封装内，使光学供应链的产能成为整机产能的直接约束，而非并行约束。

4.2.3 关键技术指标

关键技术指标（以可插拔方案作为基线）

指标类别	指标项	可插拔基线	NPO	CPO
带宽	单端口带宽	1.6T	1.6T 至 7.2 T（Hi-ONE 单引擎 7.2 T）	整机 115.2 至 409.6 Tb/s
带宽	交换容量	—	—	NVIDIA SN6810 102.4 Tb/s；SN6800 409.6 Tb/s
信道	ASIC 至光引擎损耗	约 20 至 22 dB	13 至 16 dB（ODCC 参考设计 <16 dB VPD）	约 4 dB
信道	Retimer 需求	必需	压线，视配置而定	可取消
功耗	单光引擎 / 模块功耗	800G OSFP 12 至 18 W	3.2T 模块约 20 至 21 W	单引擎约 7 W；ELS 约 2 W/端口
功耗	整机对比基线	约 7.0 kW（144×800G）	—	3.95 kW（Q3450-LD），节省 3.05 kW
功耗	整机级能效	约 15 至 20 pJ/bit（数量级）	约 6 至 7 pJ/bit（数量级）	约 5 至 6 pJ/bit（数量级）
可靠性	可用度	—	华为 Hi-ONE 案例：系统可用度 99.8%—	Meta 实测 Bailly：>100 万小时无不可纠正码字
散热	冷却方式	风冷为主	风冷 / 液冷混合	液冷基本强制
维护	激光器可更换性	—	视架构而定（内置光源则不可）	外置 ELSFP 可现场更换

维护	光引擎可更换性	面板级	板级 / 子组件级	子组件级或不可更换
标准化	成熟度	成熟	标准并存，尚在收敛	OIF 已有 3.2T IA，生态分化中

NPO 与 CPO 对比

对比项	NPO（近封装光学）	CPO（共封装光学）
集成形态	光引擎板载主板、紧邻芯片，可热插拔	光引擎与交换 ASIC 共封装同一基板
电链路长度	短（板内近距离）	极短（毫米级）
功耗与密度	较传统可插拔显著优化	最优
可维护性	故障模组可单独更换，维修成本低	光电绑定，维修复杂，良率要求高
成熟度	高，已进入试点部署	中，以原型样机/工程验证为主
适用阶段	当前主力落地（2026-2028）	远期演进（2028 年后）

NPO/CPO 系统的关键技术指标包括：光引擎聚合带宽（如单封装集成多颗 3.2T/6.4T 光引擎，聚合带宽可达 51.2Tbps 量级）、单 lane 速率（基于 224G SerDes 的 200G/lane）、端到端误码率（纠错前 BER<1e-6、纠错后 BER<1e-15）、光互连功耗节省幅度、端口密度与能效（pj/bit）。

NPO/CPO 关键技术指标

指标	量级/目标	说明
单 lane 速率	200G/lane（224G SerDes）	面向 Scale-up/Scale-out 的物理基础
光引擎聚合带宽	51.2Tbps 量级	单封装多颗 3.2T/6.4T 光引擎
纠错前 BER	<1e-6	依赖 DSP/均衡能力
纠错后 BER	<1e-15	满足协议可靠传输
功耗节省	较可插拔显著下降	缩短电链路、省去模块 DSP
能效	pj/bit 量级	每比特能耗持续下降

产业标准方面，业界多个组织推进 NPO/CPO 接口规范，包括 OIF、CPX、ODCC、DORA、OCTC、OAII 与 NPX 等，各方案在连接器形态（LGA/Mezzanine）、Pitch、低速管理接口（I2C/SPI）与通道配置上尚存差异，整体尚未收敛。产业实践上，国内头部云厂商已开展 NPO 超节点试点，CPO 则以国产整机原型样机验证为主。

4.3 高密度光互连 XPO/NG IO 技术路线

高密度光互连并非单一技术，而是若干条并行路线。按光电转换位置可归为三派。

· 外接派：LPO、LRO、XPO。保留前面板可插拔形态与可维护性，通过提升单模块密度解决空间问题。

· 内建派：NPO、CPO、OIO（光学 I/O，芯粒形式）。把转换点移入封装，换取极致带宽密度。

· 铜线派：CPC（共封装铜）、DAC、AEC、LACC。在光占优前维持低成本方案。

架构层面的共识是：未来是混合部署，而非单一技术的淘汰赛。一个典型形态是交换机前面板插高密度光模块承担 scale-out，机箱内部署 NPO 承担芯片侧极速光电转换，而机柜内最近的两颗加速器之间仍保留铜互连。

4.3.1 系统架构

XPO（扩展封装光学/新一代高密度光互连）在 NPO/CPO 框架内进一步提升光引擎集成密度、降低每比特功耗，推动光引擎、交换芯片、连接器与光纤阵列（FA）的一体化设计；NG IO（下一代 IO）聚焦下一代 SerDes 速率（224G→448G）与光互连标准的演进。XPO 通常与 USFP（Universal Small Form Factor Pluggable，通用小型可插拔光模块）路线互补：USFP 在可插拔形态内持续提升单模块容量，XPO 则向共封装与板载方向演进，二者面向不同部署代际。

在系统架构上，XPO/NG IO 以更高速率 SerDes 与更高密度光引擎为核心，构建支持万卡级以上超大规模 AI 训练集群的 Scale-up/Scale-out 互连。光引擎采用先进光电共封装与高密度光纤互连，交换芯片与光引擎深度协同，配合 448G SerDes 与下一代光互连接口，实现超高带宽密度与极低每比特功耗的全光互连。

特征项	XPO 规格
单模块带宽	12.8 Tbps
电通道数	64 lane（双 paddle card，每 card 32 lane）
单通道速率	224 Gbps PAM4（信令 200G/lane 净数据）
引擎配置	同一 paddle card 可装 4×1.6T / 2×3.2T / 1×6.4T 光引擎
光学制式覆盖	SR、DR、FR、LR、ZR/ZR+、coherent-lite、铜、RF
接口架构	兼容 retimed / half-retimed / linear（可对接 LPO 低功耗方案）
光连接器	8 × MPO-16
模块外形	约为 OSFP 宽度的 2.7 倍
功耗与散热	集成原生液冷冷板，功耗超过 400 W，48V 供电，仅支持液冷
单 U 密度	16 模块/1U = 204.8 Tb/s，约为 1.6T OSFP 方案的 4 倍

XPO 架构中最具工程价值的设计是"一个外形通吃所有介质"：同一 paddle card 可适配多种光引擎配置与多种制式，涵盖短距多模、常规单模、相干轻量以及铜互连。对 224G 世代而言，这一能力具有特殊价值：224G 之后还存在 400G/lane 的演进预期，形态兼容性直接降低了换代风险。

可插拔、NPO/CPO、XPO/USFP 路线对比

维度	可插拔光模块	NPO/CPO	XPO/USFP（下一代）
集成度	面板可插拔	光引擎近/共封装	更高密度、标准与共封装并存
单端口容量	800G/1.6T	多引擎聚合	持续提升，向 3.2T+演进
电链路	长（前面板）	短	更短/毫米级
功耗	高	低	最低

维度	可插拔光模块	NPO/CPO	XPO/USFP（下一代）
标准化	成熟	收敛中	标准定义中
部署代际	当前	当前主力/远期	远期（448G 时代）

-

4.3.2 应用优势和挑战

XPO/NG IO 的应用优势包括：突破 NPO/CPO 的规模与带宽密度限制，支撑更大规模 AI 超节点与 AI Factory；单 lane 速率向 448G 演进，显著提升互连带宽；光引擎与交换芯片一体化设计，降低互连功耗、时延与每比特成本；面向下一代算力基础设施，提供可持续演进的光互连路线。

· 密度优势（直接且可算）。XPO 单模块 12.8 Tbps 等于 8 颗 1.6T OSFP；单 1U OCP 开放机架单元可容纳 16 个模块，提供 204.8 Tb/s 交换容量，前面板密度约为 1.6T OSFP 方案的 4 倍，交换机机柜占地可降低约 75%。

· 系统级账。Arista 口径：12 万 GPU 数据中心采用 XPO 后，网络机柜可从 1,400 个降至 350 个；百万 XPU 数据中心约需 1.28 亿颗 scale-up 光学部件，全产业年需求超过 10 亿颗。在超大规模场景中，互连部件的数量级已接近芯片本身，形态变革带来的总量效应极为可观。

· 能效优势。中国移动白皮书给出了一组值得警惕的增速剪刀差数据：模型参数规模约每两年增长 400 倍，芯片算力约每两年提升 3 倍，而芯片间互连能力仅约每两年提升 1.4 倍。同一白皮书指出，在 576 GPU 以上的超大规模集群中，高速铜互连系统功耗约占集群总能耗的 30%。若该趋势延续，互连将成为算力扩张的首要瓶颈。

· 可维护性优势。XPO 方案直击 NPO/CPO 的核心痛点——光引擎损坏即导致整机报废或需要整机断电更换——通过标准化插槽实现不停机更换。这代表业界对 NPO/CPO 维护性问题的一种主流回应方向。

XPO 面临挑战的核心是：铜光边界尚未收敛，且生态正处分化期。

其挑战包括：光电共封装的良率与可靠性、散热与热管理难度大；448G SerDes 对 PCB 板材、连接器、DSP 与光互连提出极高要求，远超 224G；光引擎与交换芯片的协同设计复杂度高；高速光互连的标准化与多厂商互通尚未成熟；从原型验证走向规模量产的产业链成熟度与成本控制是主要制约。

关于铜光边界，当前可确认的工程经验判断是：近期 1 至 3 年内，距离 7 m 及以内优先采用铜缆（如 AEC 增强型铜缆），发挥低功耗、低成本、高可靠优势；超过 7 m 优先采用光互连；3 至 5 年长期向端到端 CPO 演进。另一侧判断是：在 200G/lane 以下，铜线仍具成本优势，铜光并进预计持续至 2027 至 2028 年。其机理在于：224G 级 PAM4 下介质与导体损耗显著上升，被动 DAC 需加粗线规以压降阻抗，进而恶化体积、重量与最小弯曲半径，超出可行距离后必须转向 LACC、AEC 或光模块。阿里云的判断是"光与铜本质互补而非竞争"，采用近端铜、远端光的组合方案。

关于生态分化，2026 年 OFC 是多路线同期落地节点，三大 MSA 分别占据不同生态位。

MSA	主导方	定义范围	关键技术特征
XPO	Arista	高密度可插拔模块形态	12.8 T/模块，64 lane，原生液冷冷板
OCI	AMD、Broadcom、Meta、Microsoft、NVIDIA、OpenAI	Scale-up 芯片间光信号格式与传输协议	NRZ 调制免除 DSP 延迟；协议无关性；短距 VCSEL/MicroLED，跨机架硅光加外置 CW 激光

Open CPX	Coherent、Marvell	可维修式共封装光学插槽接口	基于 Samtec CPC 连接器，铜光共接口；当前实现 6.4T@200G/lane，为 400G/lane 预留

生态分化的风险是可预见的：XPO 的"一换八"设计将压缩传统 OSFP 形态的出货量，不具备整合能力的中低端模组厂将面临可寻址市场的结构性收缩；同时，无硅光制程能力的代工厂可能被系统性排除。

4.3.3 关键技术指标

XPO/NG IO 的关键技术指标包括：单 lane 速率（224G→448G 演进）、光引擎聚合带宽与端口密度、每比特功耗（pj/bit）、端到端误码率（纠错前/纠错后 BER）、光电共封装的良率与可靠性、光互连标准化程度与多厂商互通性，以及从样片到量产的产业链成熟度。

• 指标项	数值	来源与口径
XPO 单模块带宽	12.8 Tbps	Arista / XPO MSA，2026-03
XPO 电通道数	64 lane（32 lane/paddle card，双 card）	XPO MSA 官方
XPO 单通道速率	224 Gbps PAM4（信令 200G/lane）	Converge Digest，2026-03
XPO 单模块功耗	超过 400 W，48V 供电，仅液冷	Arista / XPO MSA
XPO 单 U 密度	16 模块/1U = 204.8 Tb/s（约 4× 1.6T OSFP）	Arista，2026-03
XPO 光连接器	8 × MPO-16	公开技术分析，2026-06

- XPO 主要规格（数据来源：Arista 与 XPO MSA，2026-03）

4.4 OCS/OXC 技术路线

OCS（Optical Circuit Switching，光电路交换）以光为交换介质，通过光开关在光通路层面建立端到端连接，不在电域进行逐包转发；OXC（Optical Cross-Connect，光交叉连接）在光域实现多端口之间的任意交叉连接。OCS/OXC 系统由光开关阵列(基于 MEMS 微镜、LCoS 液晶或波长选择开关 WSS)、光交换矩阵、控制/编排平面与光电混合交换网络组成。

在系统架构上，OCS/OXC 部署于 AI 集群多 Rack 网络的汇聚/核心层，与电分组交换（EPS）构成光电混合交换架构：电交换承担动态、不可预测的突发流量，光交换承载规律性强、带宽需求高的大流量。OCS 通过光通路将交换/计算节点间的固定大流量直接贯通，实现光域直连，降低网络层级与时延。OCS 适合周期性、可预测且持续时间长的通信（如 All-to-All、All-Reduce 的固定阶段），其毫秒级切换无法适配微突发或每请求的动态路径。

维度	OCS（光电路交换）	OXC（光交叉连接）

英文全称	Optical Circuit Switch	Optical Cross-Connect
交换语义	在端口之间建立点到点物理光路，不解析数据包	在 DWDM 波长层面做上下路与交叉连接
交换粒度	端口对端口	波长对端口
出现位置	数据中心内 spine / super-spine、AI 集群 ICI	骨干网与城域的 ROADM 节点、CDC、WSS
速率透明性	完全透明（不感知速率、协议与波长）	仅在指定频带内工作，波长栅格固定
设计约束	必须对波长完全透明，否则无法跨代复用	必须精确控制波长，否则无法做波长分配
代表系统	Google Apollo / Palomar、华为 DC808	WSS + ROADM、存量电信级 WSS 产品线

4.4.1 系统架构

3D MEMS 是当前唯一具备大规模商业可交付性的 OCS 器件路线。其原理是：输入与输出各配置一组二维光纤准直器阵列，中间放置两组二维 MEMS 反射镜阵列，光束经两次反射在自由空间中完成三维交叉。

Google Palomar 的公开规格是理解该架构的范本：136×136 非阻塞配置 (每个 MEMS 阵列有 176 面镜，按良率启用 136 面，对应 $136^2 = 18{,}496$ 种输入输出组合）；插入损耗低于 2 dB；回波损耗典型值 −46 dB (规格小于 −38 dB)；整机功耗最大 108 W，作为对照的是同量级电分组交换设备约 3,000 W；采用 850 nm 监测光束经相机成像并通过图像处理闭环反馈微镜以优化插损；全双工通过光环形器在单根光纤上实现，从而把光纤数量减半。规格中的 136 端口实际由 128 个业务端口加 8 个用于链路测试与维修的备用端口构成。

Palomar 的 108 W 对比电分组交换的约 3,000 W，这一个数字说明：OCS 是"带服务等级承诺的可编程配线架"，而不是"更聪明的接入交换机"。其内部没有 DSP、没有缓存、没有 SerDes、不做逐包处理。

4.4.1.1 OCS 在 TPU 集群中的实际用法

OCS 的真实价值需要通过实际部署数据来判断。TPU v4 公开资料显示：TPU v4 的部署结构是：4,096 颗 TPU 组成 64 个 $4\times4\times4 = 64$ 芯片的 cube，每 cube 独占一个机架；每 cube 6 个面共 16 条光链路，即 96 条光链路；由于 3D torus 需要 wrap-around 互联，每 cube 恰好连接到 $6\times16\div2 = 48$ 台 OCS。论文明确给出：OCS 及其底层光器件占系统成本低于 5%、占系统功耗低于 3%，并称其相比 InfiniBand 更便宜、更低功耗、更快。更关键的是，用户可以选择 twisted 3D torus 拓扑，此时 OCS 的作用类似"插线板"，可跳过失效的 cube，从而实现故障绕行。

演进到 TPU v7x Ironwood，结构保持一致但规模扩大：单 superpod 含 9,216 颗 TPU，由 144 个 cube 组成，每 cube 仍接 48 台 OCS，官方口径为每芯片约 1.5 个光收发器（cube 内走铜、cube 间走光）。

值得重点关注的是 TPU v8 世代的分化。TPU v8i Zebrafish 改用 Boardfly 拓扑（受 Dragonfly 启发的高基数分层结构），单 pod 含 1,152 颗芯片，OCS 负责连接 36 个分组，网络直径从 3D torus 的 16 跳降至 7 跳（降低 56%），平均传输时延降低 50%。但 TPU v8t Virgo 刻意不使用 OCS 承担

scale-out，Google 的表态是"训练扩展尽量不用 OCS，除非必要，因为它比传统以太网交换芯片更贵也更稀缺"；Virgo 采用扁平非阻塞两层结构，单 fabric 支持 13.4 万颗 TPU 8t 与 47 Pb/s 非阻塞二分带宽，Apollo OCS 的定位转向"把 TPU 集群连接到数据中心内其他计算与存储资源"。

4.4.1.2 OCS 器件路线横向对比

器件路线	端口规模	插入损耗	切换速度	偏振敏感性	功耗	量产成熟度
机器人 / 机械臂配线架	千级	极低	秒级	无	极低	已商用
3D MEMS	数百至千级	低于 2 至 3 dB	毫秒级	中	数十至二百瓦	已商用，唯一具备大规模交付能力
DLC 数字液晶	数百级	中	毫秒级	高（需偏振控制）	低	已商用（电信 WSS 存量巨大，近年平移至数据中心）
压电 DLBS	百级	中	微秒至毫秒	高	低	研发阶段
硅光 MZI / 微环	受限	中（耦合损耗高）	微秒级	低	极低	未在量产系统实现
超表面	受限	低	微秒级	低	极低	新兴

硅光路线在理论指标上全面占优——端口成本、速度、与电子系统集成度、驱动电压均优于 MEMS。但 OCP 与 Google 均明确指出：这些优势尚未在量产系统中实现，核心障碍是光纤耦合损耗与端口数限制。

4.4.2 应用优势和挑战

OCS/OXC 的核心优势在于：其一，光域交换无需光电转换，功耗与时延极低；其二，协议透明，支持任意速率与格式的透明传输，带宽大；其三，光通路可灵活重构，能按训练拓扑动态调整连接；其四，在 AI 训练通信（周期性 All-to-All、All-Reduce）高度确定性场景下，可通过预配置光通路建立确定性低时延通道，减少电交换转发层级。

优势一：跳数下降。跳数是 OCS 在 AI 集群中最可量化的收益，且已被两代产品独立验证。Google Jupiter 数据中心网络层面，约 60% 流量走直连路径，全网平均路径长度降至 1.4 跳；产品级口径为 FCT 降低 10%、吞吐提升 30%、功耗降低 40%、成本降低 30%、停机时间降至 1/50。TPU ICI 层面，TPU v8i 配合 OCS 后直径从 16 跳降至 7 跳、平均传输时延降低 50%。此外 Google 将 OCS 拆分为 4 个各占 25% 的故障域，保证单一 OCS 机架故障时集群仍保留不低于 75% 的容量。

优势二：跨代复用带来的 CapEx 节奏优势。OCS 不感知下游端口速率与协议，光模块从 400G 演进到 800G 再到 1.6T 时，OCS 本体与光纤可以原样保留。Google Jupiter 因此可以错峰升级网络，而电分组交换设备每次交换 ASIC 换代都意味着一轮全网 CapEx 投入。

优势三：故障不中断训练。这一点可能是 OCS 最不可替代的价值。TPU v4 的 twisted torus 中，OCS 像插线板一样跳过失效单元。OCP 白皮书的测算给出量化对比：当服务器可用度为 99.9% 时，1,024 颗 TPU 的静态拓扑切片只能获得 25% 的有效吞吐，而具备可重构能力的方案可获得 75%。这条论断的意义在于：在大规模训练中，集群可用度对有效算力的影响是非线性的，OCS 的价值应按"有效吞吐"而非"交换容量"来计价。

其挑战包括：光开关切换时间通常达毫秒级，难以匹配微突发流量与快速重路由需求；光层调度需与控制平面/编排系统深度协同，实现拓扑与光通路的自动化管理；光通路故障定位与快速倒换复杂；MEMS/LCoS 等核心光器件的成熟度、寿命与成本仍是规模部署的制约；以及与电交换的协同联动和标准化尚需完善。

· 挑战一：毫秒级切换与带宽中断。切换时间存在四个不同口径，必须避免混用：器件微镜机械时间约 1 至 10 ms；学术口径的 3D MEMS 硬件时间为 10 至 100 ms；端到端重配置实测为 204.3 ms，其中约 204 ms 花在收发器锁定与同步上；厂商产品宣传口径为单端口小于 50 ms、全 fabric 小于 200 ms。

· 挑战二：需预调度，不适合数据包突发流。电路复用一旦接通即独占通路，不适合统计复用的突发包流；且无逐包 QoS 与包级遥测能力。这意味着 OCS 需要一个能够准确预测流量矩阵的控制平面，而这正是 Google 所说"控制平面工作量可能超过硬件本身"的原因。

· 挑战三：插损与跳数上限。OCS 每台插入损耗约 0.8 至 3.0 dB，且 OCP 明确指出在无光放大条件下实用 OCS 跳数通常限制在 1 至 2 级。这一约束与"OCS 可替代 spine 实现扁平化"并不矛盾（扁平化本就只需 1 跳 OCS），但明确反驳了"OCS 可级联成深层次 OXC"的想当然。对 224G 世代而言，这一约束的严峻性上升——因为光模块侧本身留给 OCS 的插损余量已进一步收窄。

· 挑战四：与无损流控的原理性冲突。InfiniBand 依靠链路级 credit-based flow control 实现零丢包，而光路是单向的。若反向 credit 路径因拓扑重配而不存在或发生拥塞，发送方将停发并引发队头阻塞，导致性能崩塌。相关研究 (ReFly) 将该问题建模为带 credit 约束的二分图最大权匹配，在 Nekbone trace 与 0.9 负载条件下将 tFCT 降至 Edmonds 基线的 58.1%，LULESH 场景吞吐提升 30.1%。

4.4.3 关键技术指标；

厂商 / 系统	端口数	器件技术	插入损耗	切换时间	功耗	状态
Google Palomar	136×136（128+8 备用）	3D MEMS	低于 2 dB	毫秒级	最大 108 W	生产部署，数千套
华为 OptiXtrans DC808 / OCS800	256×256 非阻塞	全光交叉	未公开	未公开	低于 200 W，6U	商用
Coherent	300×300（另有 64×64）	数字液晶 DLC	未公开	毫秒级	未公开	已量产，驱动电压低于 10 V
Lumentum R300	300×300	MEMS	约 3 dB	毫秒级	未公开	已商用，两家超大规模

						客户
Calient S320	320（每端口 TX/RX 对），7RU	3D MEMS	0.8 dB min / 1.5 dB typ / 3.0 dB max	单端口低于 50 ms，全 fabric 低于 200 ms	低于 45 W	商用，累计部署超百万端口
光迅科技	192×192	自研 MEMS	不高于 1.2 dB（隔离度不低于 50 dB）	低于 50 ms	未公开	商用
Telescent G5	1008×1008	机械臂配线架	极低	秒级	latching	商用

OCS/OXC 关键技术指标

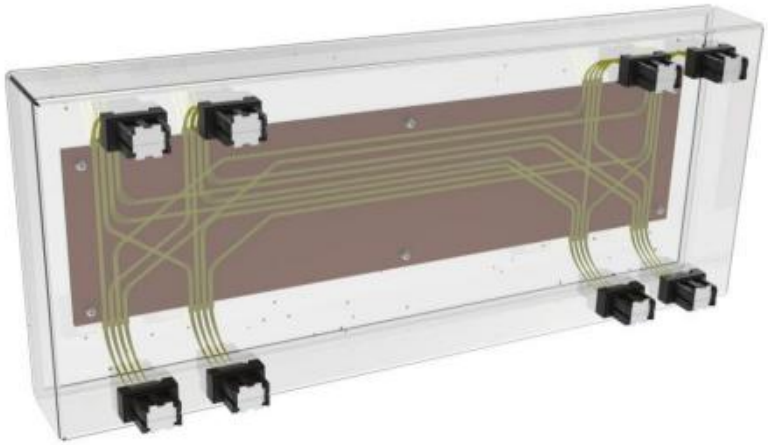
指标	量级/目标	说明
端口规模	数十至数百端口	光交叉端口数量
切换时间	毫秒级	MEMS/LCoS/WSS
插入损耗	数 dB 量级	决定功率预算与传输距离
通道隔离/串扰	满足链路预算	通道间隔离度
重配置周期	随规模增长	拓扑重构频率
寿命与可靠性	长期稳定	MEMS 开关寿命等

4.5 光 Shuffle 技术

光 Shuffle（光交叉/光跳线）通过无源光交叉组件（如光纤跳线矩阵、光背板、光交叉模组）将光信号在多个端口与链路之间重新分配，实现光域的拓扑重构。其系统架构由光交叉模组、光纤阵列 (FA)、连接器与布线体系组成，部署于 NPO/CPO 系统、多芯片交换架构与 AI 超节点内部，完成光信号的高密度交叉连接。

在系统架构上，光 Shuffle 通常与 NPO/CPO 光引擎及交换芯片协同：交换芯片/光引擎端口信号经光 Shuffle 按拓扑需求交叉互联，替代部分电 Shuffle 功能，将板内/柜内光信号路径重新分配，缩短电互连路径。典型实践如国产 3.2T NPO 多芯片交换机采用系统级光互连搭配 Shuffle 光交叉模组，实现芯片间光交叉与细粒度端口拆分。光 Shuffle 通过预置的交叉映射（Passive Shuffle）把多路光信号重排到不同目的地，无需光电转换，插损是决定链路预算的首要因素。

光背板 Shuffle 是实现全光互连系统内光路交叉互连的核心部件，高密度布线且占用空间极小，解决柜内不同节点间的互连，克服传统光缆布线空间受限的困难，为 CPO/NPO 互连架构提供轻量化、高可靠性的光互连解决方案。



光背板 Shuffle

光 Shuffle 是实现 NPO 架构下光路交叉互连的核心部件，高密度布线且占用空间极小，实现光引擎到面板的光路交叉，直接解决传统光缆布线空间受限的矛盾，为 CPO/NPO 互连架构提供轻量化、高可靠性的光互连解决方案。



图 光Shuffle

光 Shuffle 是一个真实存在但极为年轻的术语。其权威定义来源是 2026 年 2 月 24 日 IEEE 802.3 以太网人工智能（E4AI）评估 adhoc 会议上，Panduit 的 Jose Castro 提交的同名报告《Optical Shuffle Architectures for Large AI Networks》。该报告的核心定义是：扁平化的网络架构需要更少的交换机与收发器，从而降低功耗与通信时延；而扁平网络依赖光通道洗牌（lane shuffling，也称 meshing），通过光 breakout 实现；breakout 让一个收发器可连接到多个端点，而非单一对端；典型例子是 1.6T 等于 8×200G lane，每条 lane 可被分发到不同的交换机端口。

换言之，光 Shuffle 的本质是：在物理层把一个高速光模块的多条并行光通道确定性地重新映射、分发到多台不同交换机端口乃至不同交换机的无源光纤重排器件。这里的关键词是无源与静态——它没有开关元件，不消耗功率，映射关系在部署时一次固化。

关于 breakout 与 shuffle 的区分：Castro 的原始定义中，breakout 只是"一根拆多根"（例如 8 芯 MPO 拆为 4×2 芯），而 shuffle 才是"重新决定哪一路信号送往何处"——映射关系本身就是产品价值所在。

关于与 OCS 的关系，二者不等价而是互补，且恰好处于系统设计生命周期的两个不同时间点。

维度	光 Shuffle（被动洗牌）	OCS（光电路交换）
决策时机	部署前（工厂预制）	运行时（可重构）

有源器件	无（无供电、无管理接口、无告警）	有（MEMS 驱动与控制器）
可重构拓扑数量	1（一次固化）	N×N 任意置换
切换时间	不适用	毫秒级
插入损耗	极低，仅连接器、弯曲与余长	约 1.5 至 2.7 dB（商业产品实测）
解决的问题	布线密度、网络层数、故障域	拓扑弹性、带宽按需分配
成熟度	已商用（2026 年有多家产品）	已商用部署

二者在 AI 集群中可串联使用（例如节点侧采用 shuffle、spine 层采用 OCS），不存在冲突。需要指出的是，概念上二者共享同一个祖先——都在做"端口置换"。置换等于电路交换这个等式在 OCS 语境下成立，但不适用于光 Shuffle，因为光 Shuffle 是固定的、非交换的。

4.5.1 系统架构

光 Shuffle 有三种实现形态。

· Shuffle harnesses（洗牌线缆）。把 breakout 与 shuffle 结点集成在线缆错开的支腿里，光纤直接落到不同交换机与端口。优势是省机架空间，代价是更难排查故障。

· Mesh cassettes（网格盒 / 小型光柔性电路）。小尺寸、模块化、易安装易排查，但大规模部署时空间利用率较差。

· Shuffle boxes / modules（洗牌盒/ 模块）。密度高于 cassette，简化安装，损坏时整模块更换。这是当前最大规模 AI 集群设计中最常出现的形态。

从网络架构层面看，光 Shuffle 的作用机制可概括为三点。

· 第一，消除一层网络。其核心价值在于把传统三层胖树压为两层。IEEE 报告给出的聚合统计是：交换机数量减少 40%、fabric 收发器减少 50%、总收发器减少 33%，并省去两跳。需要谨慎对待这组数字：它们是"两层对比三层"的聚合统计，在 8K 规模下 spine 数量并不发生变化（64 保持 64），改善体现为端口利用率提升 4 倍，而非设备数量减少。

· 第二，跨层扩展而硬件不变。以 NVL72 GB300 为例，可从 1,152 GPU（64 个 leaf 加 18 个 spine）扩展至 18,432 GPU（1024 个 leaf 加 288 个 spine），硬件不变而规模扩大 16 倍。规模阶梯为 8K（128 leaf / 64 spine）→ 32K（512 leaf / 64 spine，spine 由 256 减至 64）→ 131K → 524K，每级约 4 倍递增。

· 第三，限制故障损失。由于 1.6T 可拆分为 8 个独立通道，单点故障仅损失 12.5% 带宽，集群可维持 87.5% 容量继续运行 RDMA。这一特性与 4.4 节 OCS 的故障旁路能力形成呼应，且实现成本远低于 OCS。

4.5.2 应用优势和挑战

成本优势。插入损耗极低是光 Shuffle 最突出的优势。以厂商规格书数据为准，低损耗型产品插入损耗典型值不高于 0.25 dB，低损耗等级不高于 0.15 dB，对比 3D MEMS OCS 的约 1.5 至 2.7 dB，差距在两个数量级。这一差异直接决定链路预算是否需要光放大。

· 无源带来的部署简化。器件无供电、无管理接口、无告警，故障模式极少，部署与运维复杂度显著低于有源光路交换。

· 速率无关性。同一光纤既可跑 1.6T 亦可跑 200G，映射一旦确定即与速率解耦，兼具跨代兼容性。

· 故障域隔离。将单点故障损失限制在 1/8，而非整个端口或整机失效，对大规模训练的可用性贡献直接。

光 Shuffle 的应用优势包括：无源实现、功耗极低；拓扑灵活，可适配多种互连拓扑与端口拆分需求；插损可控，利于高密度光布线；可减少电 Shuffle 与高速电互连，降低功耗与信号损耗，提升系统能效。在 AI 超节点柜内互连中，光 Shuffle 是实现大规模光交叉、平滑扩展算力规模的重要使能技术。

· 挑战一：不可重构。映射关系固化于部署时，无法适应负载变化。因此对 MoE（混合专家）等每次迭代通信模式都不同的负载，光 Shuffle 原理上无法跟随。需要明确指出的是，业界针对 MoE all-to-all 的答案是 OCS 路线（如 MixNet，SIGCOMM 2025），而非光 Shuffle——这是两个容易混淆的技术选择，其分野在于动态与静态。

· 挑战二：部署灵活性与故障排查。Shuffle harness 省空间但难排查；mesh cassette 易排查但空间利用率差；shuffle box 折中但损坏时需整模块更换。三种形态各有代价，选型需按运维场景权衡。

· 挑战三：与下一代收发器的链路训练要求耦合。下一代收发器将要求 optical link training，这约束了 lane 在各 PMD 之间的分配，可能改变当前映射规则。这意味着今天预制的映射未必能在未来适配新收发器。

· 挑战四：市场规模与数据可靠性。厂商宣传数字与规格书数字存在系统性偏差，需注意甄别。典型案例：某厂商新闻稿称插入损耗"低至 0.05 dB"，但其自家规格书（IEC 61753-1）标注为典型值不高于 0.25 dB、低损耗等级不高于 0.15 dB，两者相差 3 至 5 倍。本章建议正文引用规格书数据而非新闻稿数据。

其挑战包括：光交叉的插入损耗预算受限，需严格控制链路总插损；端口管理与可维护性要求高，回插一致性与长期可靠性需保障；高密度光纤布线对光纤阵列（FA）与连接器的对准精度提出严苛要求；相关标准化尚未完全收敛，多厂商互通性与供应链成熟度有待提升。

4.5.3 关键技术指标

光 Shuffle 的关键技术指标包括：交叉端口数量与芯数（决定可支持的互连规模）、单通道插入损耗（影响链路功率预算）、回波损耗与串扰、光纤阵列（FA）对准精度与一致性、端口拆分的粒度和灵活度、可维护性与插拔寿命，以及整体系统的功耗与集成密度。

指标项	数值	来源与口径
定义权威来源	Jose Castro, Optical Shuffle Architectures for Large AI Networks, IEEE 802.3 E4AI	2026-02-24，16 页
基本映射单元	1.6T = 8×200G lane，每 lane 可分发至不同交换机端口	IEEE 802.3 E4AI 报告
规模阶梯	8K（128L/64S）→ 32K（512L/64S）→ 131K → 524K，每级约 4 倍	IEEE 802.3 E4AI 报告
扩展实例	NVL72 GB300：1,152 GPU（64L+18S）→ 18,432 GPU（1024L+288S），硬件不变，16 倍	IEEE 802.3 E4AI 报告

两层化收益	交换机减 40%，fabric 收发器减 50%，总收发器减 33%（聚合统计口径）	IEEE 802.3 E4AI 报告
端口利用率提升	同规模下由 1 提升至 4 倍（spine 数量不变时的利用率改善）	IEEE 802.3 E4AI 报告
单点故障损失	12.5%（1.6T 拆分为 8 通道），集群维持 87.5% 容量	IEEE 802.3 E4AI 报告
插入损耗	典型不高于 0.25 dB，低损耗等级不高于 0.15 dB	厂商规格书（IEC 61753-1）
单层最大纤芯数	1,200 芯以上	厂商规格
Shuffle Box 形态	4U 空间容纳 1,024 芯，插入损耗低于 0.35 dB，8 芯并行光架构，EIA 19 英寸兼容	厂商规格
合规标准	GR-63/ GR-1435、IEC 61300、IEC 61753-1、IEC 61754-7、TIA-604-5	厂商规格
支持的纤芯类型	MPO / MTP 8 至 24 芯、MMC、SN-MT（VSFF）、裸纤	厂商规格

5. 第五章 产业生态与未来发展趋势

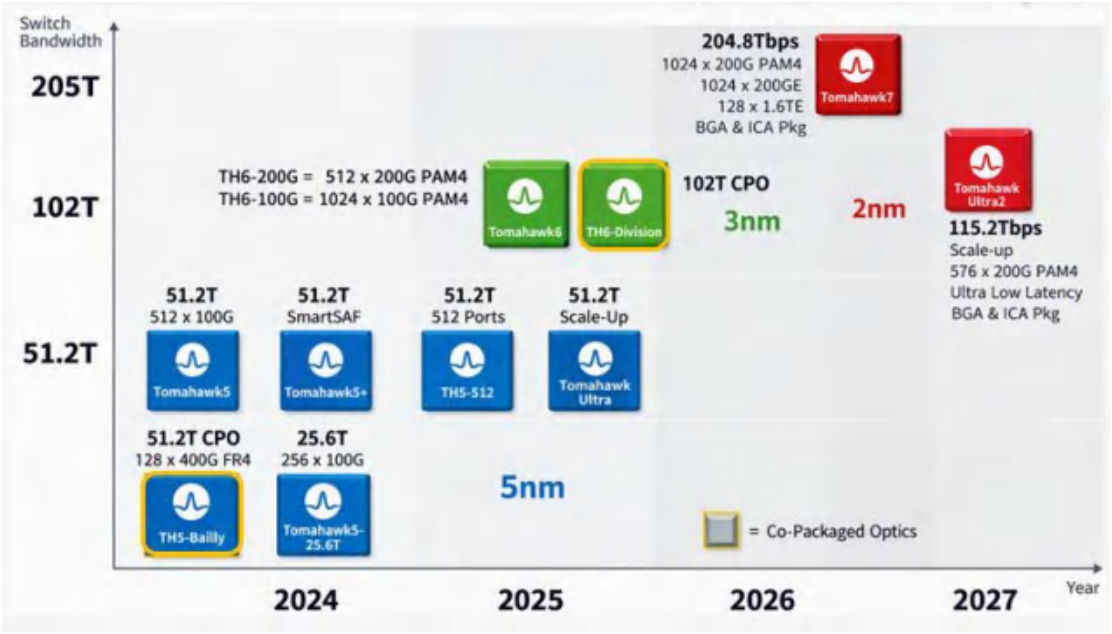
5.1 全球 224G 产业生态分析

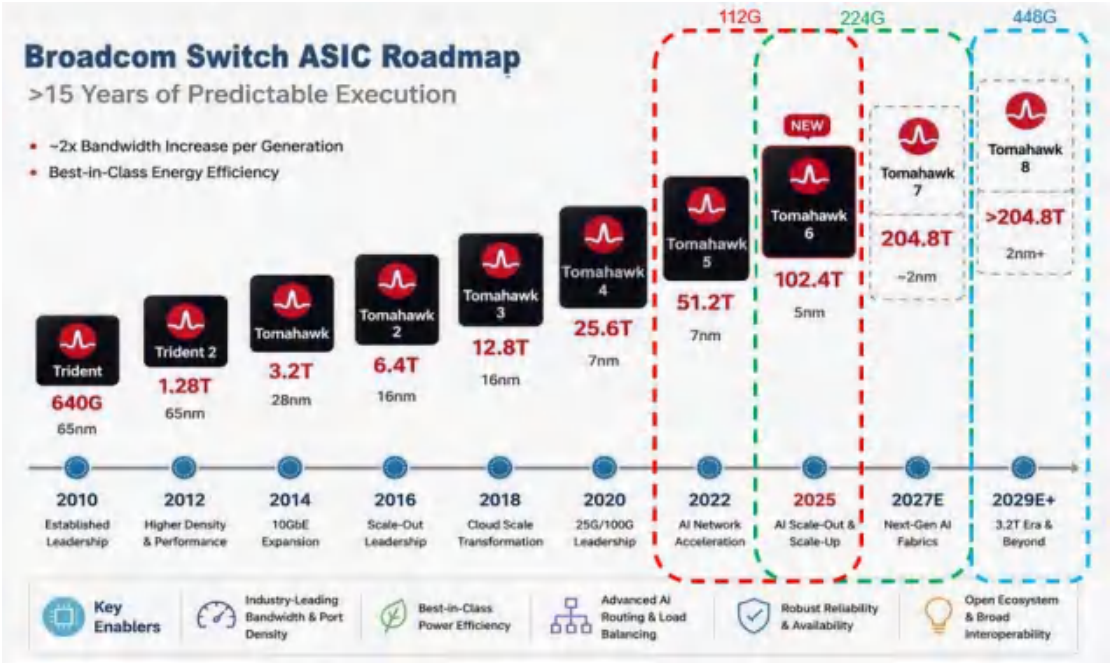
5.1.1 主流 GPU 厂商路线



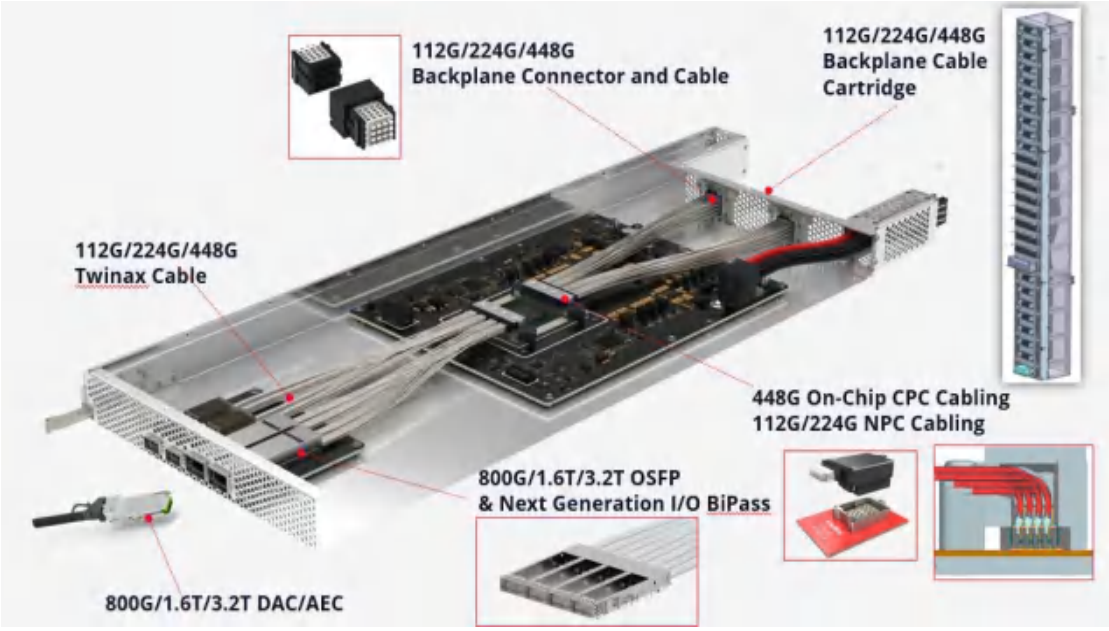


5.1.2 主流交换芯片路线





5.1.3 高速连接器发展趋势

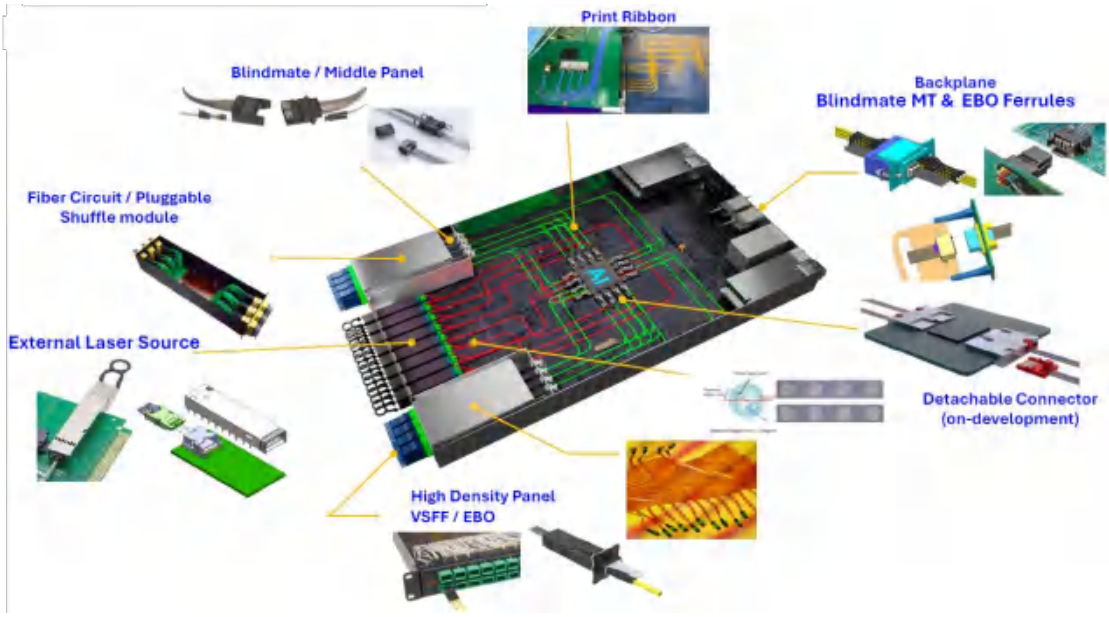


高速铜互连趋势预测

112G ：节点内互连采用板对板+PCB 走线，柜内节点间互连采用背板线缆（Cable Backplane），跨柜互连采用无源铜缆（DAC）

224G ：节点内互连采用板对板+PCB 走线或近封装线缆(NPC Cable)，柜内节点间互连采用背板线缆（Cable Backplane），跨柜互连采用有源铜缆（AEC）

448G ：节点内互连采用共封装线缆(CPC Cable)，柜内节点间互连采用背板线缆 (Cable Backplane)，跨柜互连采用超低损耗有源铜缆（AEC）



NPO/CPO 光互连方案应用趋势预测

112G ： NPO 在 GPU 节点内，光纤 Shuffle 位于 GPU 节点内或是独立 Shuffle Box，MPO 是主流光纤接口

224G ： NPO 在 GPU 节点内，光纤 Shuffle 位于 GPU 节点内或是独立 Shuffle Box ，MMC/SN-MT/EBO 开始挑战 MPO 的主流光纤接口地位，集成多插芯的 Truck Cable 开始用于系统间互连。

448G ：CPO 在芯片封装上，Shuffle 位于 GPU 节点内， MMC/SN-MT/EBO 成为主要的光纤接口，集成多个 EBO 插芯的 Truck Cable 会被广泛选择

5.1.4 光模块及光引擎发展趋势；

光铜协同互连的出现与灵活互补

随着 AI 算力集群、超节点和高带宽交换架构持续演进，单通道速率正从 112G 向 224G 乃至 448G 推进，系统对带宽密度、时延、功耗和链路可靠性的要求同步提升。在此背景下，铜互连与光互连不再是简单的替代关系，而是需要根据传输距离、损耗预算、散热条件和可维护性进行分层协同：机柜内短距场景继续发挥铜互连在成本、时延和工艺成熟度方面的优势，长距或高密度场景则更多依赖光互连突破带宽和损耗瓶颈。过去，铜互连与光互连在硬件形态、封装方式和测试流程上长期不兼容，芯片厂商和系统厂商往往需要为 CPC、CPO、NPO、前面板可插拔等不同方案分别投入设计与验证资源。随着产业上下游对共封装技术路线的讨论不断深入，CPX（Co-packaged Extended）这类统一互连平台逐渐受到关注。CPX 的核心价值在于通过统一的插座和夹层连接结构，使同一封装侧接口能够兼容共封装铜缆 CPC 与共封装光引擎 CPO，从而让系统设计在铜、光之间保留切换空间。这不仅可以降低前期方案锁定的风险，也有助于减少多套硬件平台的重复验证投入，提升后续带宽升级和架构迭代的灵活性。

以及随着高速化与高密化的互联需求，NPO 的应用也逐步在产品化，在 NPO 应用中，光源放置位置直接影响可靠性、散热设计和运维成本，行业分为内置光源与外置光源两大类型，内置光源的优势在于集成度更高，光引擎体积更紧凑，光电链路更短，有利于提升空间利用率和信号完整性；但其劣势也较为突出：激光器紧邻高温 ASIC 或 GPU 芯片，长期工作在 80—100℃ 附近时，温度升高可能影响寿命与性能；一旦激光器失效，往往需要连带更换光引擎甚至整板，维护成本较高。外置光源则将激光器独立部署，通过保偏光纤向光引擎提供连续光载波，能够实现较好的热隔离，使激光器工作温度更易于控制在 50—55℃左右，从而提升长期可靠性。外置光源支持独立更换，运维成本更低，也便于光电技术分别迭代。不过，外置方案会引入额外的光纤布线和高密度连接器，例如 FAU、ELSFP 等，对精密耦合、光路管理以及系统级散热布局提出更高要求。当前行业内两种路线仍在并行探索，选择通常取决于场景对空间密度、可维护性和长期可靠性的侧重。

在高速互连架构中，电到光的衔接是实现数据从计算域进入传输域的关键枢纽。GPU、ASIC 等计算芯片只能处理电信号，而光纤传输依赖光信号，因此光引擎承担着电光转换的核心职能。随着单通道速率向 224G、448G 演进，PCB 走线损耗、串扰和均衡功耗迅速上升，电光接口必须不断向芯片侧靠近。传统可插拔光模块通常依赖 DSP 完成重定时和均衡，虽然兼容性好，但功耗和时延较高；LPO 通过移除模块内 DSP，改由主机 SerDes 直接驱动，降低了板级电光衔接的功耗，但仍受限于前面板较长的 PCB 走线；NPO 将光引擎移至 ASIC 附近的 PCB 上，使电信号传输距离缩短至厘米级；CPO 则进一步将光引擎与 ASIC 集成到同一封装基板，电信号路径压缩到毫米级，可显著降低 SerDes 功耗和链路损耗。总体来看，电光衔接正从“板级”向“封装级”“芯片级”持续压缩，未来光铜协同的重点也将围绕更低功耗、更高密度、更强可维护性和更灵活的铜光切换能力展开。

5.2 448G PAM6 和 PAM4 技术展望

448G 单通道传输是 224G PAM4 之后的重要演进方向，当前主要候选方案为 224GBaud PAM4 和

174GBaud PAM6。PAM6 相较 PAM4 可将 Nyquist 频率从约 112GHz 降低至约 90GHz，因此对 PCB、连接器、线缆和封装带宽要求更低，能够延长铜缆传输距离并降低高速通道设计难度，更适合作为 448G 时代的“过渡方案”。

然而，PAM6 需要 6 电平判决，信噪比（SNR）裕量明显低于 PAM4，对 DSP 均衡、ADC 精度和 FEC 提出更高要求，且目前主流光模块和交换芯片生态普遍以 PAM4 为目标，光传输对 PAM6 支持较弱。

从应用来看，PAM6 更适合 AI 超节点内部、GPU 机柜间、背板及短距铜互连等 Scale-Up 场景，可加快 448G 铜互连落地；而 PAM4 凭借成熟的光互连生态、更好的兼容性和标准化基础，更适用于交换机、光模块和 Scale-Out 网络。当前 OIF 也在同步推进 CEI-448G PAM6 与 PAM4 规范，产业趋势普遍认为：短期 PAM6 有望率先实现 448G 铜互连，中长期 448G PAM4 将成为统一的电光互连主流方案。



5.3 AI Factory 互连架构演进展望

AI Factory 正在成为继云计算数据中心之后的新一代基础设施形态，微软、Google、Meta、AWS、xAI、OpenAI 等均在建设超大规模 AI Factory，以支撑万卡到百万卡级 AI 集群

AI Factory 正从传统 CPU 数据中心向超大规模 GPU/AI 加速器集群演进，互连架构也经历了由“铜互连+电交换”向“光互连+光交换”的发展过程。当前主流 AI Factory 采用 Scale-Up 与 Scale-Out 两级网络架构：柜内或超节点内通过 NVLink、UALink 等高带宽低时延互连构建 GPU 资源池，柜间则通过 800G/1.6T Ethernet InfiniBand 实现大规模扩展。

未来 3~5 年，随着单 GPU 带宽突破 20TB/s、机柜功耗超过 600kW，传统铜缆和 PCB 信道面临带宽、功耗及布线密度瓶颈。448G PAM6 将率先推动短距铜互连升级，随后 448G PAM4 成为统一标准，支持 1.6T/3.2T 端口演进。

在架构层面，AI Factory 将逐步从 GPU 与 Switch 共柜的 Cable Backplane 架构，向 GPU 机柜、网络机柜与光纤互连解耦架构发展。Shuffle Box、Optical Shuffle、FlexPlane 及 OCS (Optical Circuit Switch) 将大量替代传统线缆背板，实现高密度、低功耗、可重构的光互连网络。

更长期来看，NPO（Near-Packaged Optics）和 CPO（Co-Packaged Optics）将把光接口推进至交换芯片和 AI ASIC 附近，显著降低高速电信号传输距离与系统功耗。同时，光互连将从点对点连接进一步演进为光 Fabric 和光交换网络，实现计算、网络和存储资源的动态重构。

总体趋势： AI Factory 互连将沿着“224G→448G→896G SerDes、铜缆→光纤、固定拓扑→可重构光 Fabric、分立光模块→NPO/CPO”的路线演进，最终形成以光互连和光交换为核心的大规模 AI 基础设施