# AI光互联发展技术发展趋势

立讯技术：彭小伟
2023-12-8

LUXSHARE TECH
立讯技术

# 目录

# AI人工智能计算特点

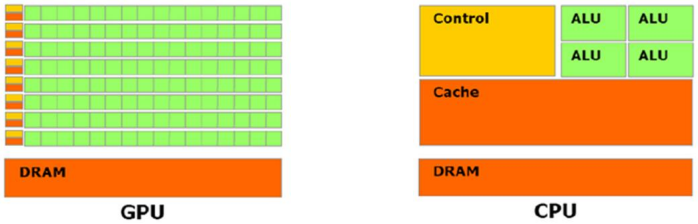
## ◘AI训练数据量巨大

✓GPT3.5: 1750亿训练参量，340T bit数据量

## ◘GPU服务器用于AI训练优势

| 主要特点 | GPU | CPU |
| --- | --- | --- |
| 架构核心 | 内核简单且众多，尽管只会简单的运算，却能并行处理大量数据。 | 内核强大而复杂，专为处理一项任务，这就像一个博学的教授 |
| 并行性 | 数量庞大的内核，它能并行处理更多的任务 | 更少内核，处理能力比CPU差100倍。 |
| 内存架构 | 配备专门的高带宽内存 | 更注重高效的缓存数据访问，对带宽的需求较低。 |
| 应用方向 | 处理大量可预测且相似的运算，如深度学习需要 | 快速响应的任务，如操作系统 |

GPU/CPU内核结构



# 算力时代AI集群对光互联的要求

## □传输速率快速增加

✓GPT3.5/4.0/5.0快速升级，带来训练参量的快速增加

✓GPU不断迭代，带宽增加，推动光模块速率提升

✓NVIDA下一代B100,即将采用1.6T

## □延时要求更为严苛

✓并行计算，以时延最大的计算结果为准

✓推动LPO，NPO, CPO等产品不断发展

## □可靠性要求更高

✓非实时保存，任何一数据出错，所有计算重来

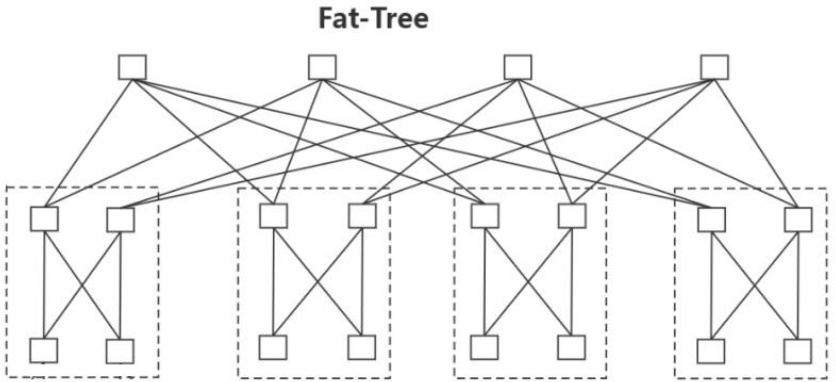
✓比电信网络和传统数据中心网络要求更高

✓推动AI互联产品向高可靠性设计

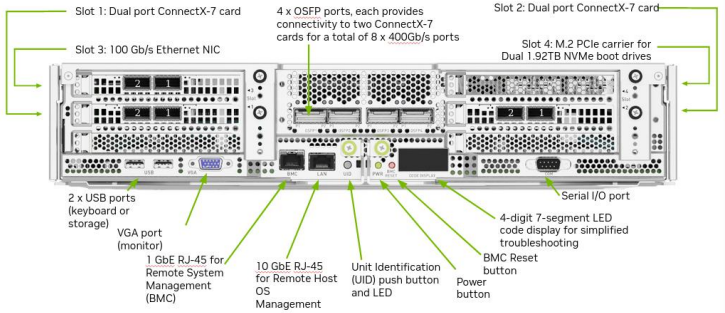
## □胖树(Fat-Tree)结构带来更多光模块需求

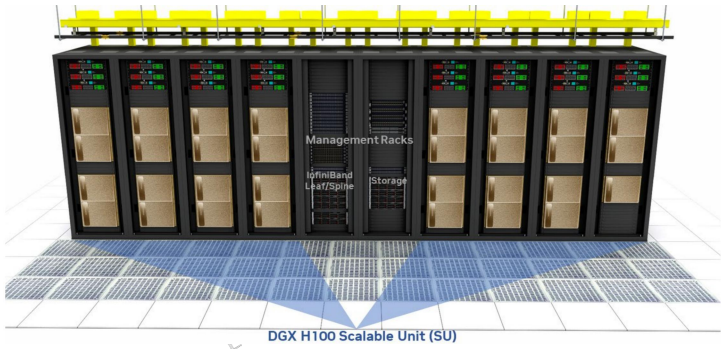
✓带宽无收敛,上行带宽和下行带宽相等

✓带来光模块数量快速增加



# 人工智能带来光互联新的成长动能：光模块的数量相对GPU数量呈倍数增长




## GPU与光模块的数量关系分析

- SU(扩展单元) = 20 node (GPU server) = 20x8 GPUs
- Super POD (超级传送点) = 7 SU
- AI Cluster System (AI集群系统) = 2 ~ 4 Super POD

- 一个基本AI集群系统使用的光模块>10K (基于全光架构)
- H100/// GPU : 400G&800G Optical Transceiver > 1 : 3
- 2024O1即将发布的H20特供版，带宽增加到900GB/s
- 晟腾910B单卡算力0.6P，支持2000~3000张卡

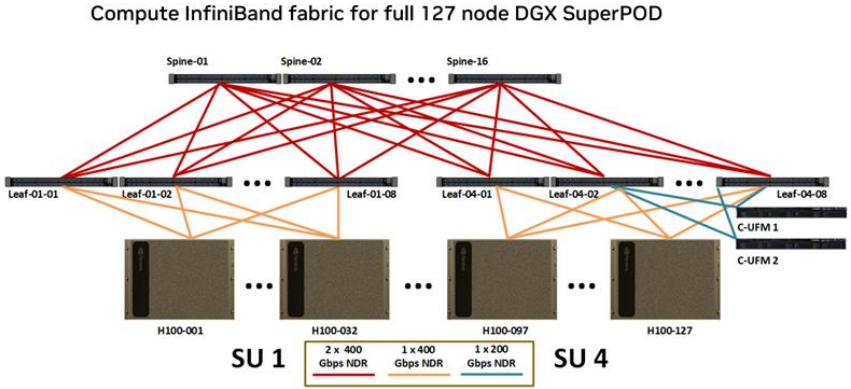
| GPU | H100 | H800 | H20 |
|---|---|---|---|
| 算力 | 2P | 2P | 0.148P |
| 带宽 | 600GB/s | 400GB/s | 900GB/s |
| 最大集群 | 5万 | 2万 | 2万 |
| 用途 | GPT 5 | GPT 4 | GPT 3.5 |

| 半光模塊架構 | 數量 | | 全光模塊架構 | 數量 |
|---|---|---|---|---|
| GPU (H100) | 4000 | | GPU (H100) | 4000 |
| IB NIC (400G NDR) | 4000 | | IB NIC (400G NDR) | 4000 |
| 銅纜 (Server to ToR) | 2000 | | 光模塊 (400G Server) | 4000 |
| | | | 光模塊 (800G ToR) | 2000 |
| 光模塊 (800G Spine/Core) | 8000 | | 光模塊 (800G Spine/Core) | 8000 |
| IB 交換機 (400G 64Ports) | 320 | | IB 交換機 (400G 64Ports) | 320 |

# AI架构: NVIDIA SuperPOD— DGX H100

## ◘ 计算网络架构—InfiniBand Fabric：

The compute fabric layout for the full 127-node DGX SuperPOD. Each group of 32 nodes is rail-aligned. Traffic per rail of the DGX H100 systems is always one hop away from the other 31 nodes in a SU. Traffic between nodes, or between rails, traverses the spine layer.




Compute fabric component count

| SU Count | Cluster Size # Nodes | Cluster Size # GPUs | Leaf Switch Count | Spine Switch Count | Compute + UFM Node Cable Count | Spine-Leaf Cable Count |
|---|---|---|---|---|---|---|
| 1 | 31[1] | 248 | 8 | 4 | 252 | 256 |
| 2 | 63 | 504 | 16 | 8 | 508 | 512 |
| 3 | 95 | 760 | 24 | 16 | 764 | 768 |
| 4 | 127 | 1016 | 32 | 16 | 1020 | 1024 |

1. This is a 32 node per SU design, however a DGX Node must be removed to accommodate for UFM connectivity.

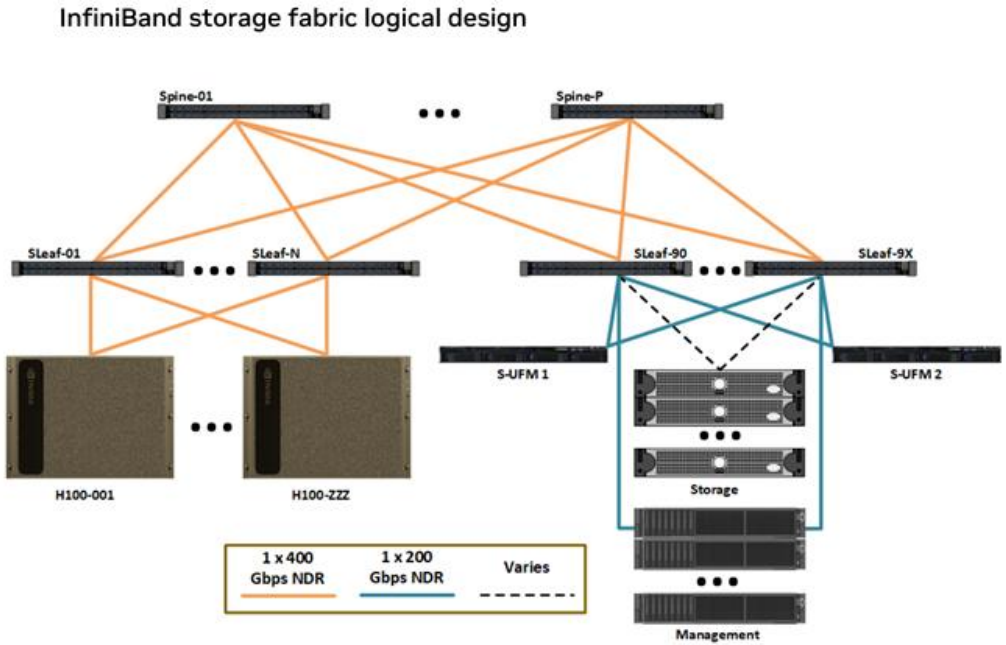
UFM: Unified Fabric Manager

# AI架构: NVIDIA SuperPOD— DGX H100

◘ 存储网络架构—InfiniBand Fabric：

The storage fabric employs an InfiniBand network fabric that is essential to maximum bandwidth. This is because the I/O per-node for the DGX SuperPOD must exceed 40 GBps. High-bandwidth requirements with advanced fabric management features, such as congestion control and AR, provide significant benefits for the storage fabric.





# AI大数据模型互连架构拆解： 基于线缆背板，连接器模组，Chip to BP 新一代的互连架构

光连接专注外部连接(柜间): 光跳线, AOC和光模块


1.Cable Backplane



112G Intrepid™ Backplane Cable Tray

2.Backplane Conn



112G Intrepid™ OD Backplane Conn

3.CHIP2IO/CPC/NPC/CHIP2CHIP & 4.CHIP2BP

112G CHIP2IO

OmniEdge™ CRE

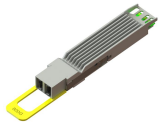
112G CHIP2BP

5.High Speed I/O

800G OSFP Conn & Cage

6.DAC & 7.AEC & 8.AOC/Transceiver/Patchcord

800G OSFP DAC &AEC



400G/800G Transceiver

# 英伟达 AI光互联场景

## □三种典型IB应用场景

✓交换机与交换机之间

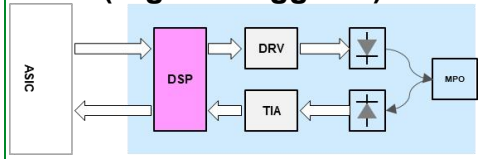
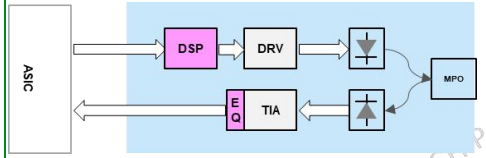
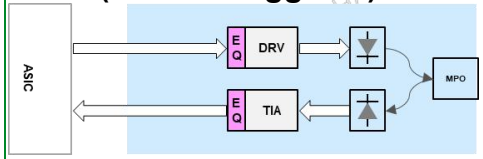
✓AI计算网络网卡与交换机之间

✓AI存储网络网卡与交换机之间

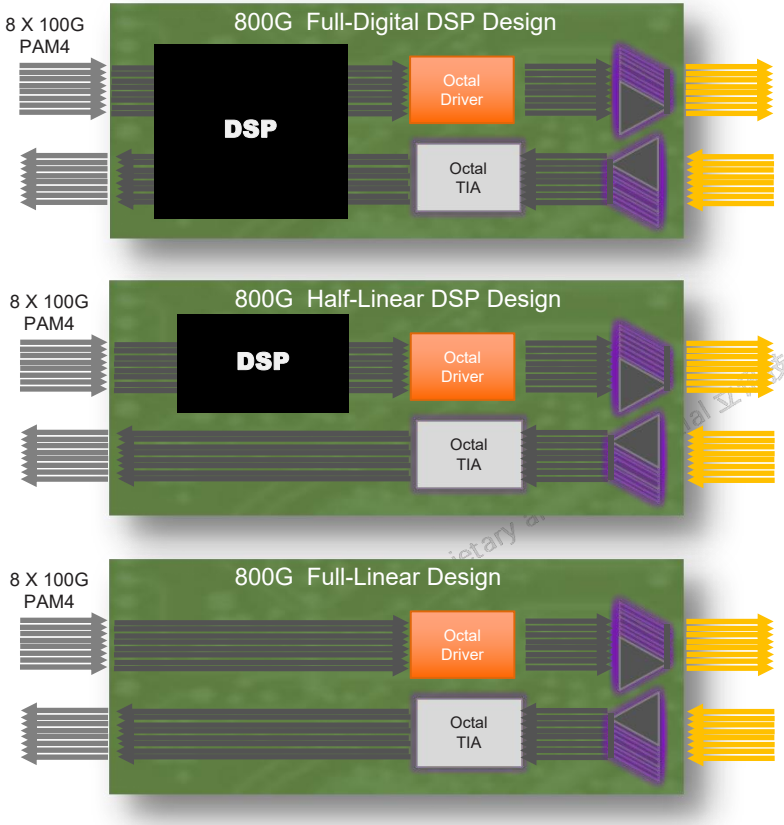
| 应用场景 | A端 | 产品形态 | SERDES速率 | 连接比例 | 交换机 | B端 | SERDES速率 | 产品形态 |
|---|---|---|---|---|---|---|---|---|
| NVIDIA Switch to Switch | 9700 | 800G OSFP Finned-top 2xSR4/DR4/AOC | 100G | 1:1 | 800G IB交换机 | 9700 | 100G | 800G OSFP Finned-top 2xSR4/DR4/AOC |
| AI Compute to NVIDIA Switch | DGX H100 | 800G OSFP RHS 2xSR4/DR4/AOC | 100G | 1:1 | 800G IB交换机 | 9700 | 100G | 800G OSFP Finned-top 2xSR4/DR4/AOC |
| AI Storage CX7 to NVIDIA Switch | CX7: OSFP | 400G OSFP RHS SR4/DR4/AOC | 100G | 2:1 | 800G IB交换机 | 9700 | 100G | 800G OSFP Finned-top 2xSR4/DR4/AOC |
| | CX7: Q112 | 400G Q112 SR4/AOC | 100G | 2:1 | 800G IB交换机 | 9700 | 100G | 800G OSFP Finned-top 2xSR4/DR4/AOC |

# AI可插拔光互联产品趋势-LPO/LRO

- DPO架構目前是AI光互联出貨的主力產品同時也持續在降功耗由 7nm DSP (2023)轉至 5nm DSP (2024)，
- LRO架構是一個新的方案選擇，相對於DPO功耗及時延都相對顯著降低，對於交換機通道間channel loss補償有較大的餘量
- LPO架構是一個終極的優化方案，功耗, 時延跟成本都是最低的，主要考量點是在不同交換機系統特性的兼容性並能否滿足IEEE規範的要求

| 800G DR8 | Power Consumption | Latency | PreFEC BER | Cost |
|---|---|---|---|---|
| **DPO (Digital Pluggable)**   | 14.0W*-16.0W | 113ns (*Data of QSFP112 DR4-DPO loopback with 10m Fiber) | BER ~1E-11 | 1x |
| **LRO (Linear RX Pluggable)**   | 10.0W*-12.0W | ~60ns | BER ~1E-11 | ~0.85x |
| **LPO (Linear Pluggable)**   | 8.0W | 2.4ns (*Data of QSFP112 DR4-LPO loopback with 10m Fiber ) | BER ~1E-08 | ~0.7x |

# 400G/800G 可插拔AI光互联产品




DPO: Digital-Drive Pluggable Optics

① 800G 2xDR4
② 800G 2xFR4
③ 800G 2xSR4
④ 400G DR4
⑤ 400G FR4
⑥ 400G SR4

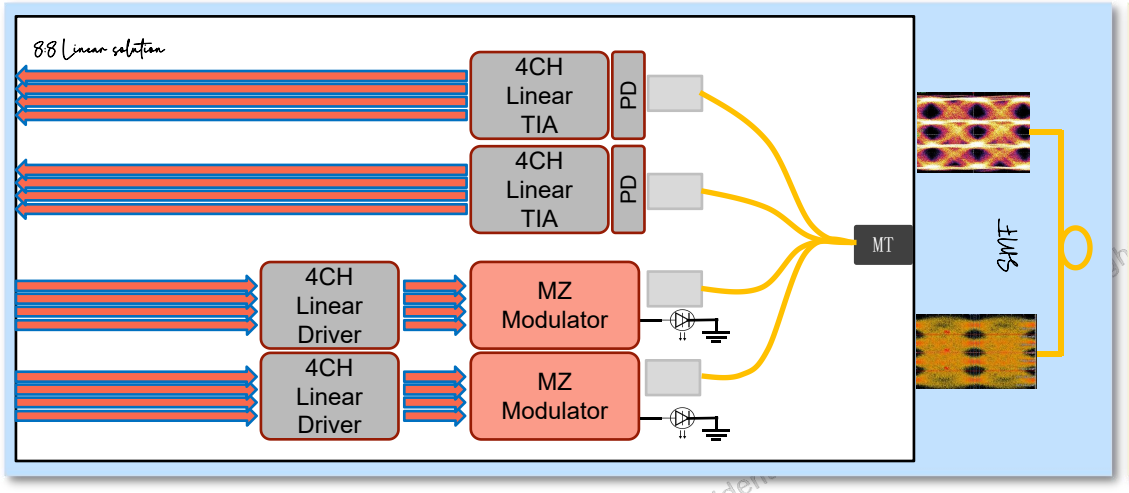
LRO: Half-Linear Pluggable Optics

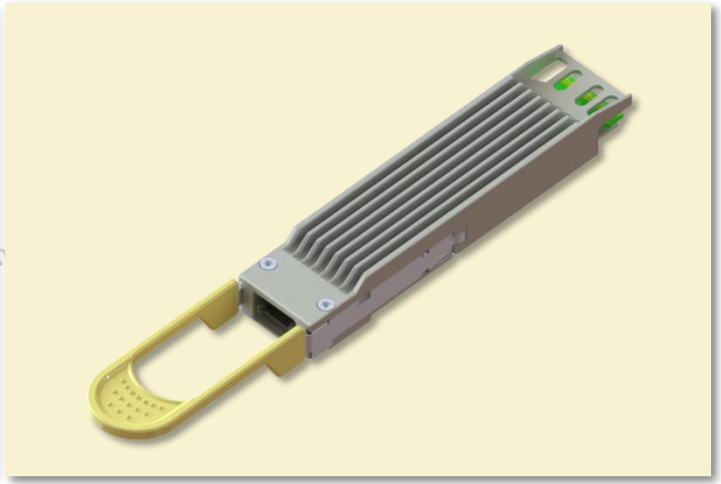
① 800G 2xDR4
② 800G 2xSR4
③ 400G DR4
④ 400G SR4

LPO: Linear-Drive Pluggable Optics

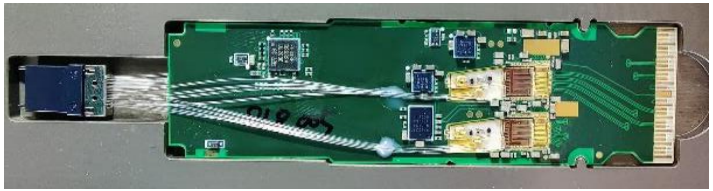
① 800G 2xDR4
② 800G 2xSR4
③ 400G DR4
④ 400G SR4

# 800G OSFP 2xDR4/DR8 LPO



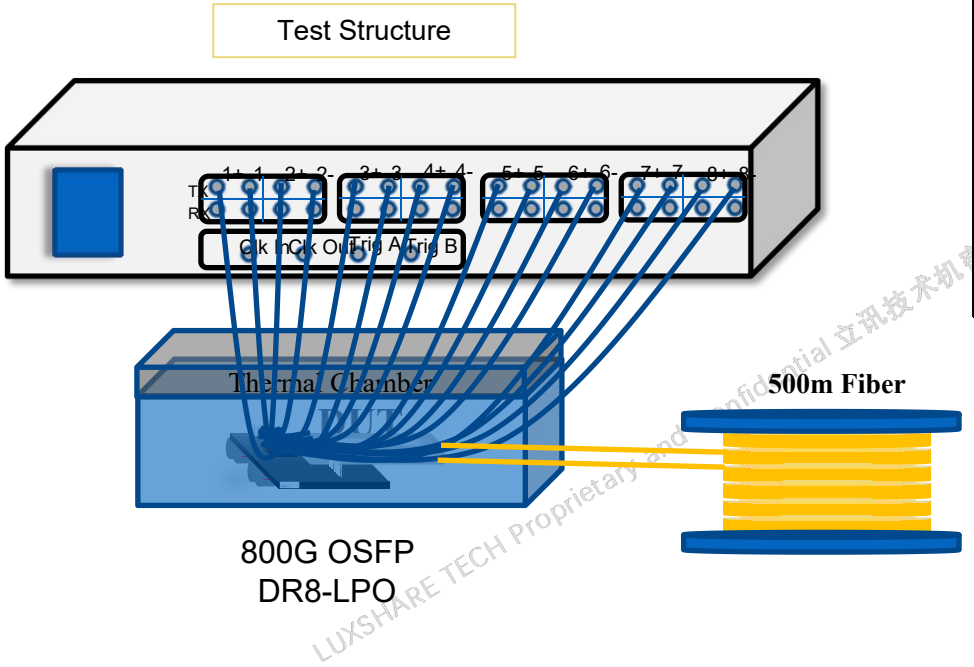
- Power consumption≤8W;
- 2x Silicon photonic MZ Modulators
- 2x 4CH Drivers, 2x4CH TIAs & 2x 4CH PD arrays
- High Power CW Laser





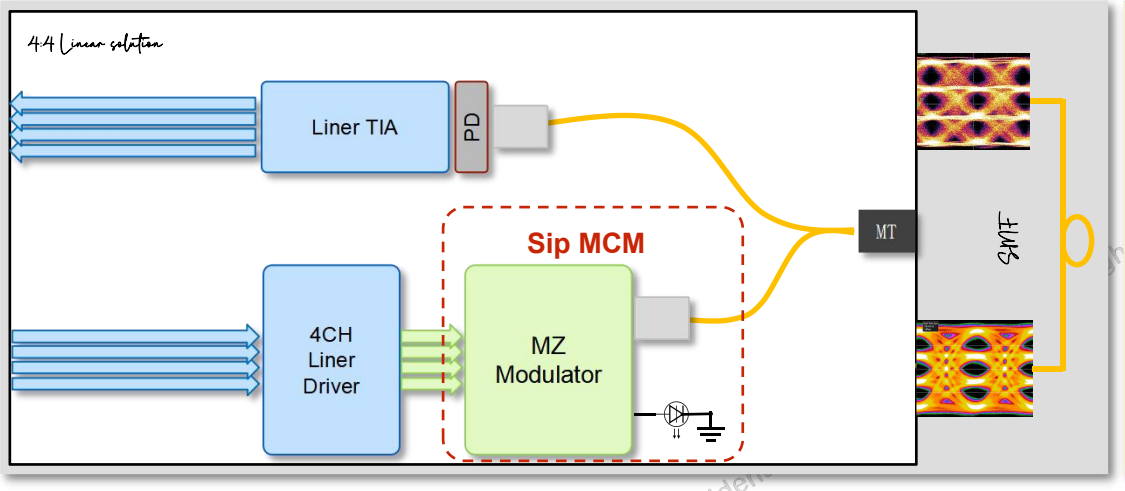
◆ **Sample available**

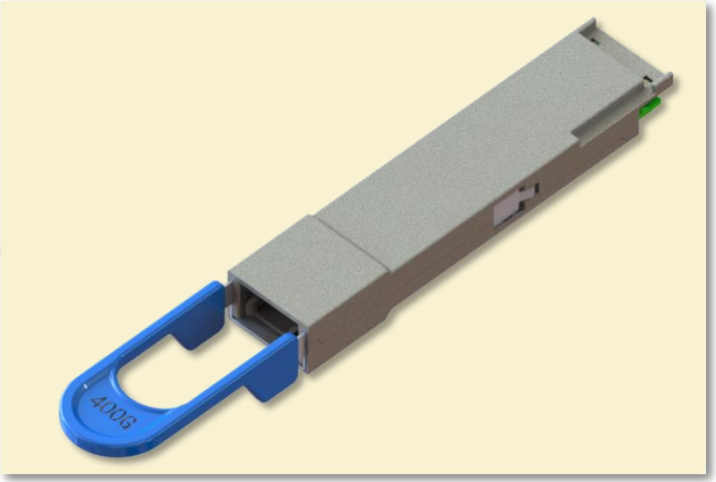
# 800G OSFP 2xDR4/DR8 LPO误码测试




| | TX参数 | | RX参数 | | 功耗 |
|---|---|---|---|---|---|
| **Channel** | TX power | TDECQ | SEN | BER | Power Consumption(W) |
| CH1 | 2.2 | 2.78 | -7.2 | 6.00E-09 | 7.21 |
| CH2 | 2 | 2.62 | -7.3 | 3.00E-09 | |
| CH3 | 1.3 | 2.38 | -7.6 | 2.80E-09 | |
| CH4 | 2.45 | 2.35 | -6.9 | 1.70E-09 | |
| CH5 | 2.5 | 3.5 | -7.3 | 1.70E-09 | |
| CH6 | 2.8 | 2.5 | -7.1 | 3.30E-09 | |
| CH7 | 3.4 | 2.25 | -7.2 | 1.60E-08 | |
| CH8 | 3.3 | 2.41 | -7 | 5.40E-10 | |

# 400G QSFP112 DR4 LPO



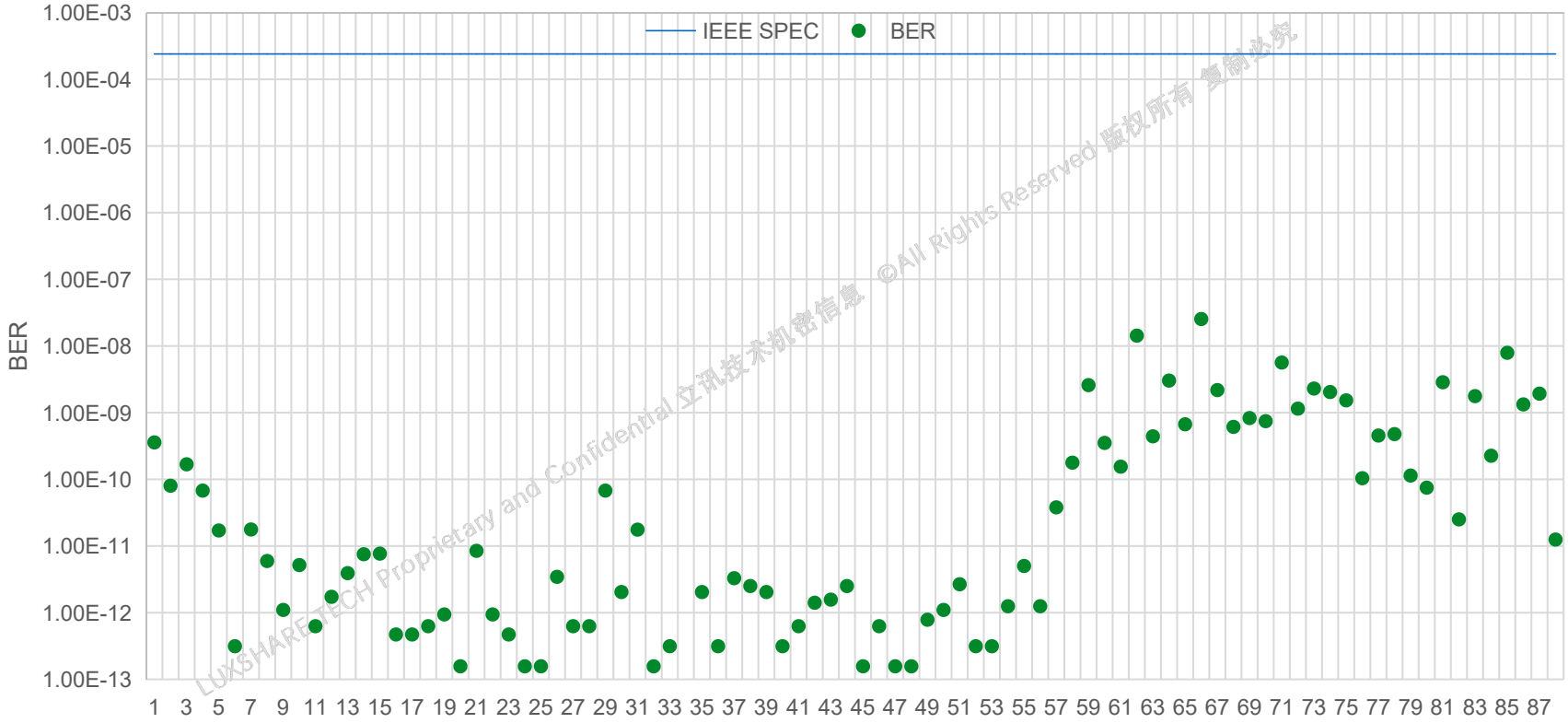




- Power consumption≤5W;
- Silicon photonic MZ Modulator
- 4CH driver, 4CH TIA & 4CH PD array
- High Power CW Laser



◆ **Samples available**

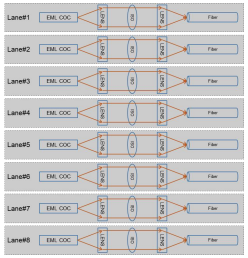
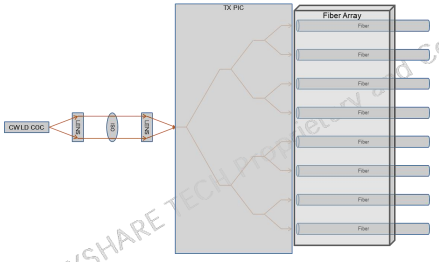
# 400G QSFP112 DR4 LPO性能测试@TH5交换机平台



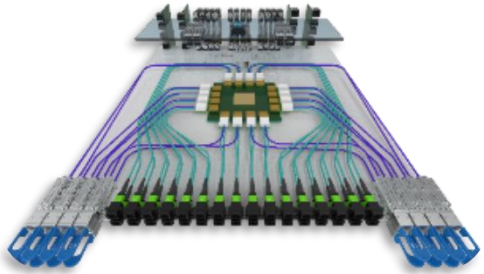



*Based on SL PIC solution

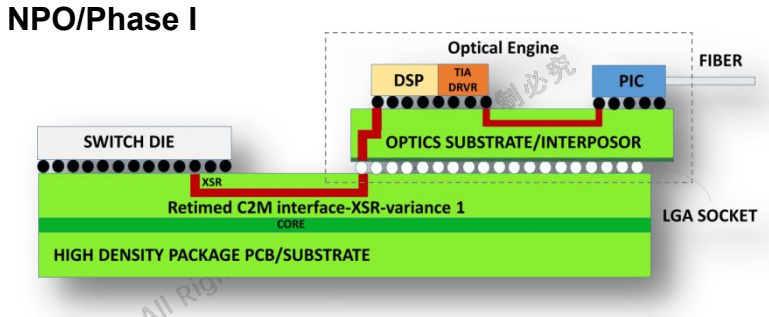
# 硅光演进方向-高集成度高可靠性

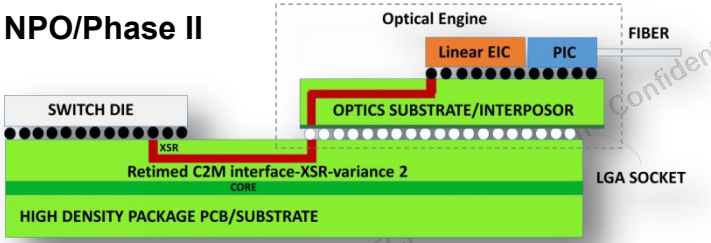
| 800G DR8 | Product Architecture | Production Scalability | Assembly Cost/UPH | Reliability |
|---|---|---|---|---|
| 传统<br>分离器件方案<br>(EML) |  | ① EML供应链受限制<br>② 封装制程复杂使用非常多的分离光学镜片组<br>③ EML拥有较高的带宽 | ① EML需要耦合8次<br>② 产业链更成熟 | ① 需要8个激光器，失效可能性8倍<br>② 分离器件，工艺复杂，可靠性影响因素多 |
| 硅光集成方案 |  | ① 硅光波导方案制程相对简化<br>② 大功率激光器工艺成熟<br>③ 封装制程简单，高密度硅光波导及硅调制器已经集成在硅光芯片上面 | ① 硅光方案(1分8)只需要耦合1个激光器和1组FA<br>② 产业链处于发展期，但已有批量长期应用 | ① 1个激光器，可靠性更好<br>② 硅光集成器件，集成度高，可靠性有优势 |

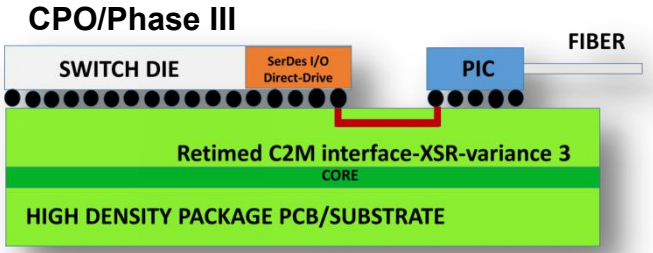
# 下一代AI光互联-NPO/CPO



NPO Switch




- 交换机芯片带宽持续提升，可插拔式光模块功耗持续上升，交换机系统设计面临挑战，光互连设计复杂度大幅提升，推动下一代光互联向**NPO/CPO**演进

# 总结

- AI计算数据量巨大，采用GPU并行计算可满足深度学习应用
- 人工智能并行计算特性，在光互联传输速率、时延、可靠性和模块数量带来新的要求
- 英伟达GPU服务器胖树拓扑，带来大于1:3的光模块需求
- 可插拔AI光互联演进方向-LPO/LRO，以满足严苛的时延要求
- 硅光集成方案的应用，带来更高可靠性
- 交换机芯片带宽，可插拔式光模块功耗，交换机系统设计和光互连设计复杂度的变化，推动AI光互联向CPO/NPO演进

# Thank You