

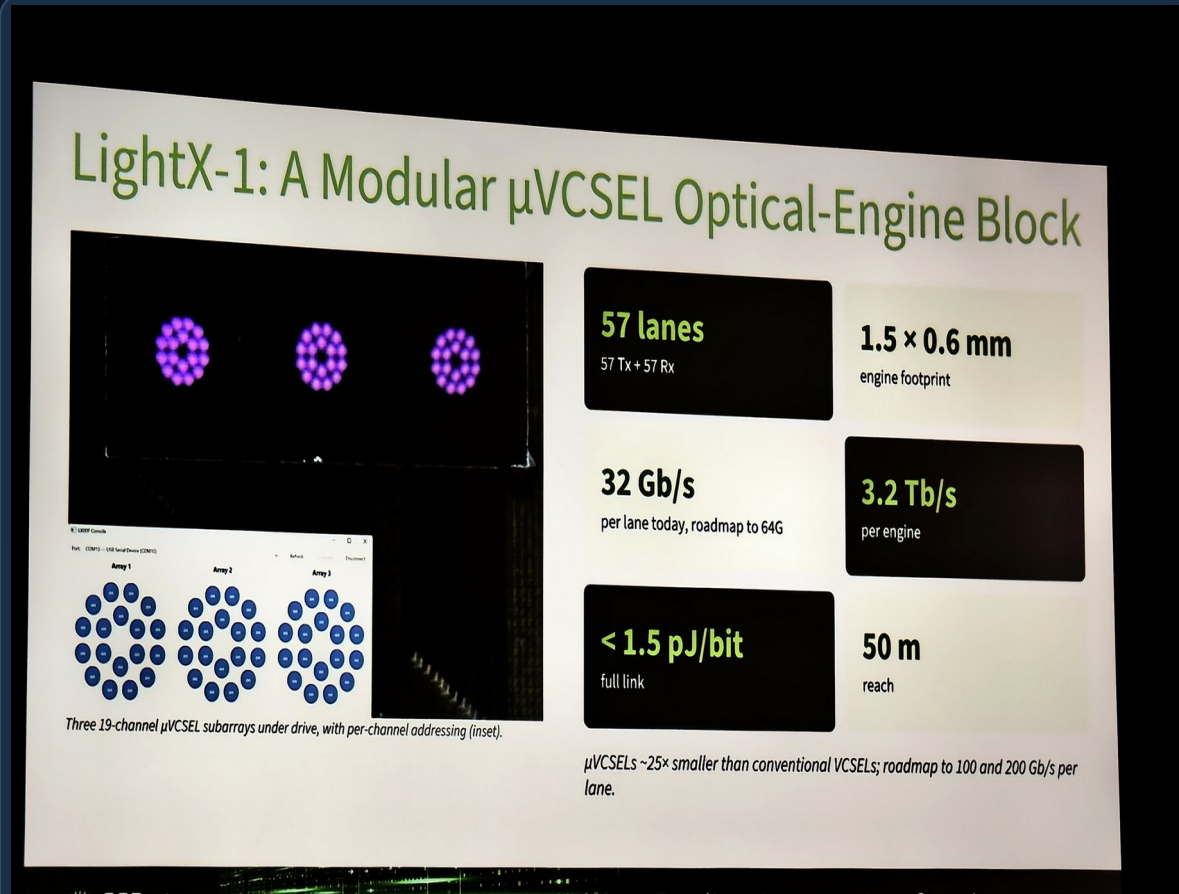
现场技术笔记

OCP 第二天： 光互连的工程真问题

MRC × XPO × NPO/CPO × Micro-VCSEL

AI Scale-up 互连、3×19 芯多芯光纤与 57 路光引擎

LightX-1: A Modular μ VCSEL Optical-Engine Block



57 lanes
57 Tx + 57 Rx

1.5 × 0.6 mm
engine footprint

32 Gb/s
per lane today, roadmap to 64G

3.2 Tb/s
per engine

< 1.5 pJ/bit
full link

50 m
reach

Three 19-channel μ VCSEL subarrays under drive, with per-channel addressing (inset).

μ VCSELs ~25× smaller than conventional VCSELs; roadmap to 100 and 200 Gb/s per lane.

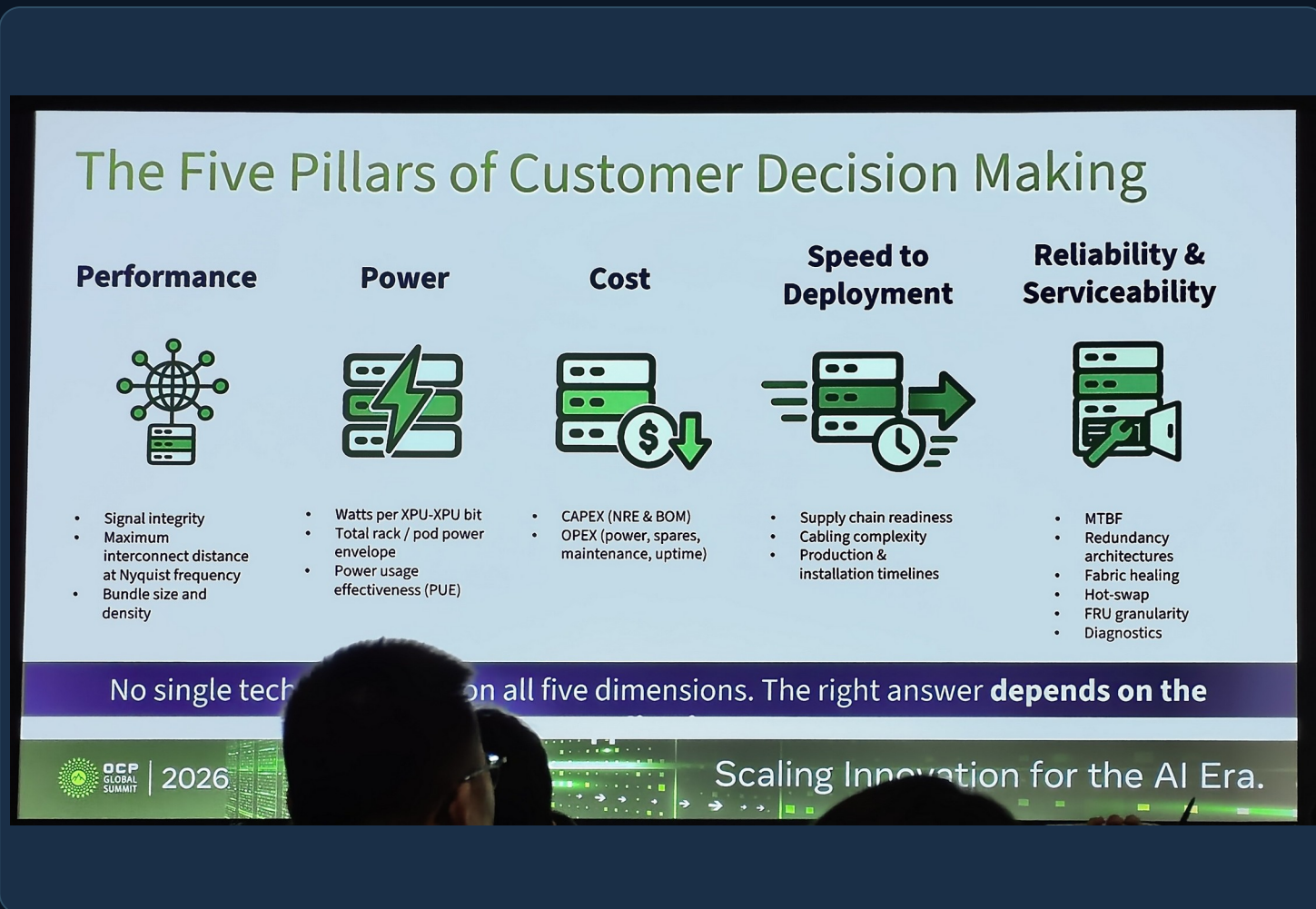
OCP | 2026 | Scaling Innovation for the AI Era

现场实拍 / LightX-1 Optical Engine

从 GPU 竞争到整个集群的效率竞争

同样算力，互连可能决定实际可用算力

- 性能 / 功耗 / 成本
- 部署速度 / 可靠性 / 可维护性
- 产业转向系统级协同优化



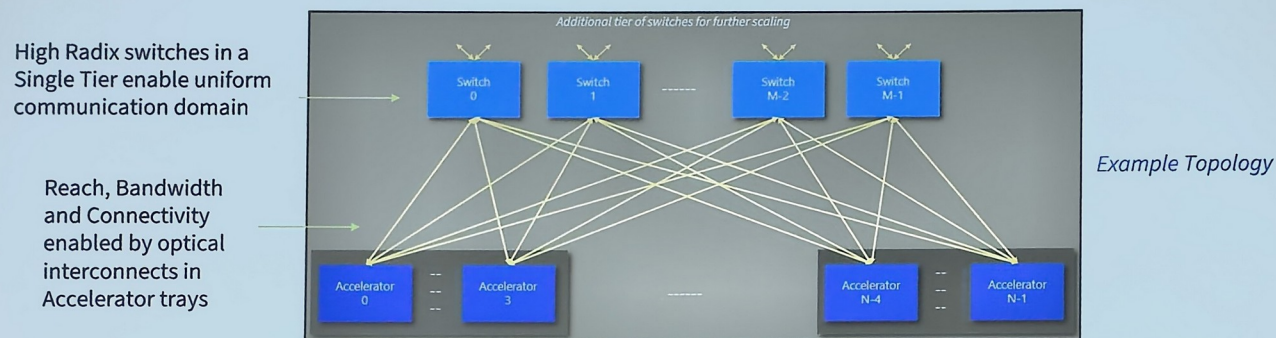
现场照片 G4-29 · OCP Global Summit 2026

Scale-up 光互连：光到底应该放在哪？

离 GPU 越近，电损耗越小，制造与维修越难

- 可插拔 XPO：可替换模块
- NPO：靠近芯片，保持可拆光引擎
- CPO：最短电路径，更复杂封装

AI Scale-up System Topologies



Richly connected, Multi-Rack and Multi Row Scale-up Domains

Several topology variations are possible depending on target AI workload performance and cost



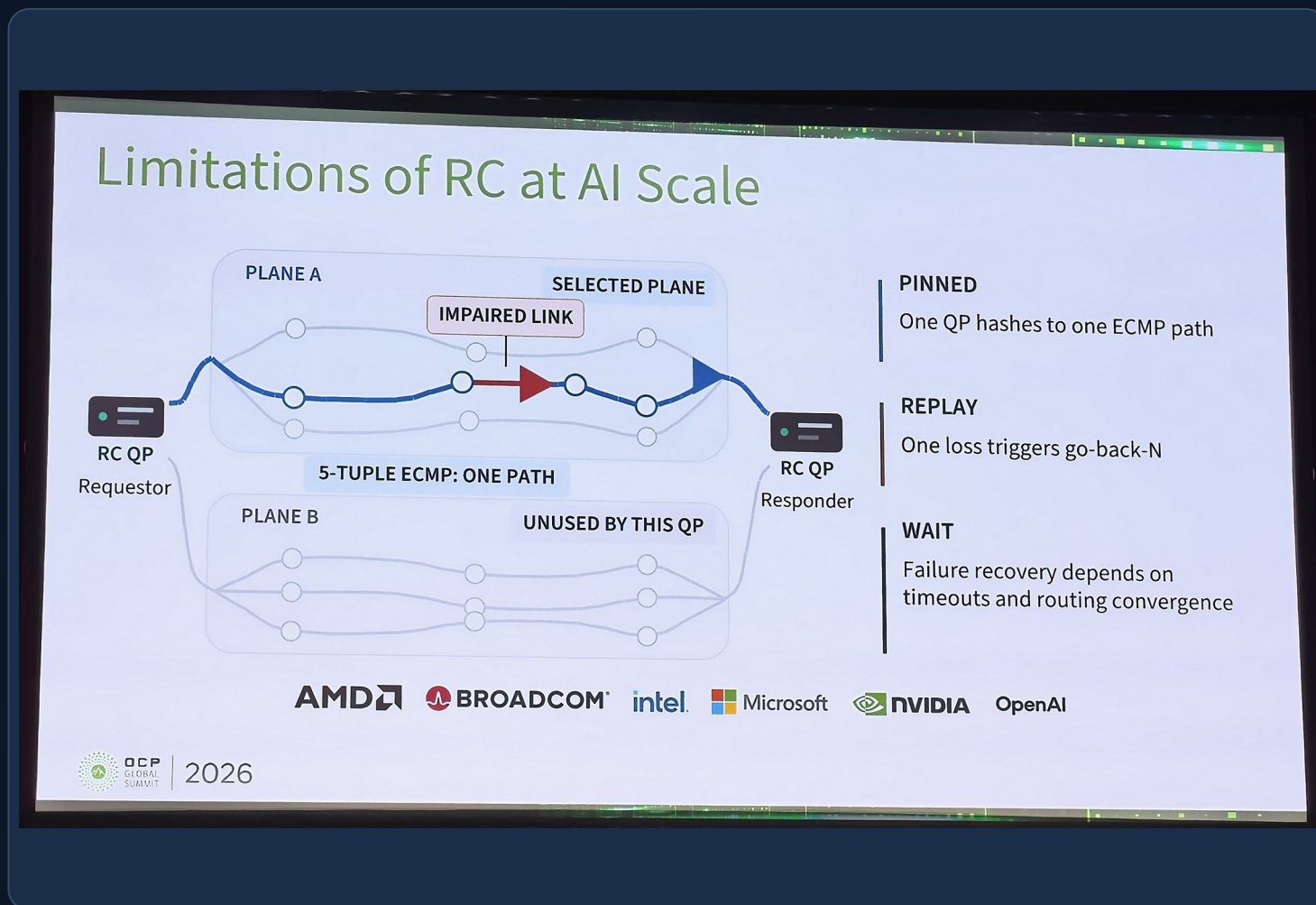
2026

Scaling Innovation for the AI Era.

为什么一个故障会拖垮一整片训练网络？

传统 RC 对丢包与路径故障的恢复方式偏粗放

- 多条 plane 上数据并发前进
- 传统 go-back-N 可能重复发送已到包
- MRC：选择性恢复 + 路径级切换

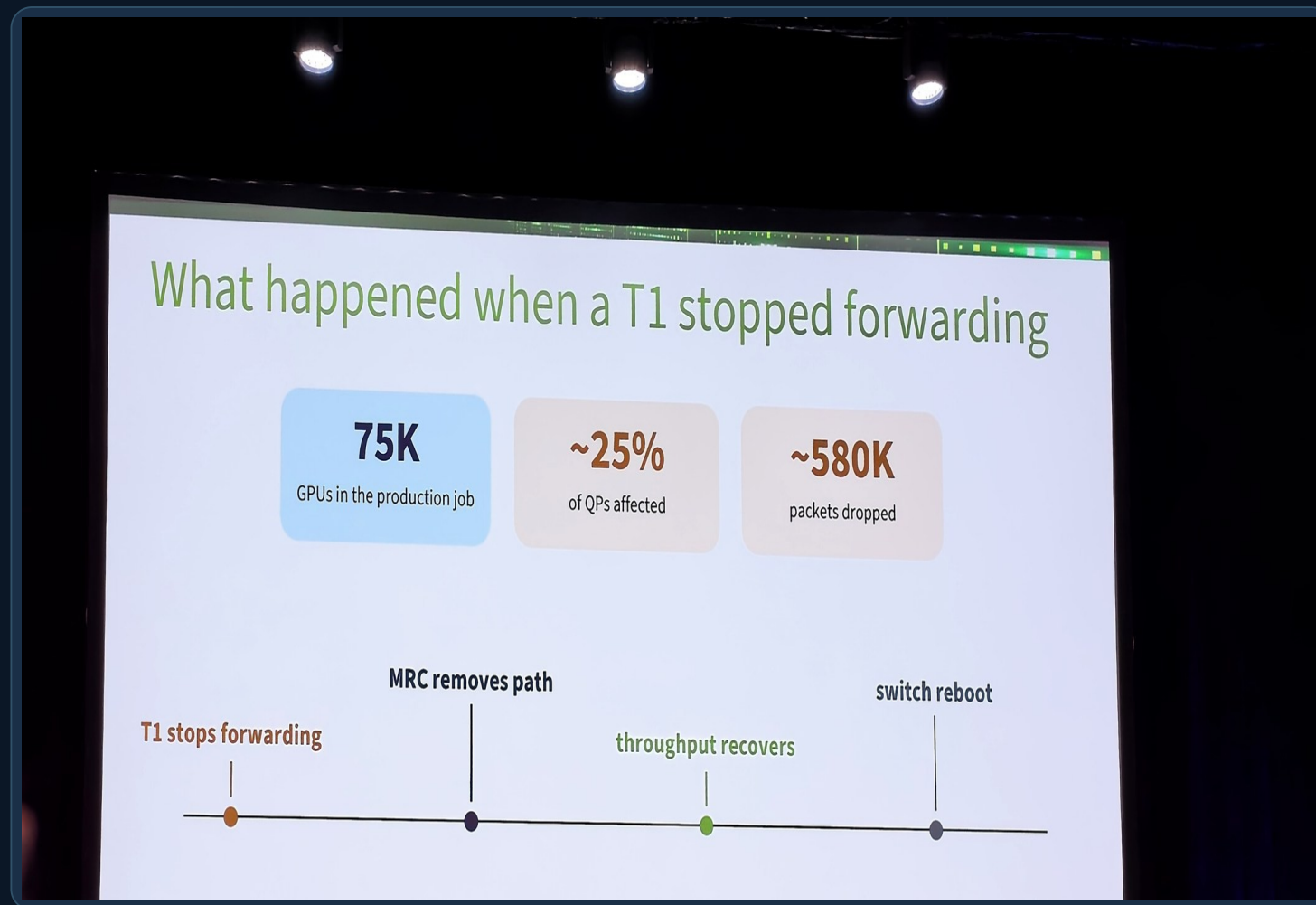


现场照片 G2-01 · OCP Global Summit 2026

现场案例：75K GPUs，约580K 丢包

读清楚三个数字，不能调换分母

- 75K：作业中的 GPU 数量
- ~25%：受影响的 Queue Pairs
- ~580K：被丢弃的数据包



现场照片 G2-07 · OCP Global Summit 2026

选择性 ACK：只补真正丢失的部分

比起整段重传，更重视对乱序的边界管理

- SACK/NACK 找到缺失序号
- 有界包序号窗口；直接写入目标内存
- 优先恢复最老的缺口

The slide is titled "Selective Loss Detection and Recovery". It compares two approaches to handling packet loss:

- CONVENTIONAL RC**
 - Go-back-N replays the tail**
 - One missing packet invalidates later progress
 - Sequence: 10, 11, ~~12~~, 13, 14, 15. A bracket under 12-15 indicates "RETRANSMIT 12-15".
- MRC**
 - Selective recovery preserves progress**
 - SACK identifies received packets and the actual gap
 - Sequence: 10, 11, ~~12~~, 13, 14, 15. A box labeled "SACK: 10 11 [GAP] 13 14 15" indicates "RETRANSMIT 12".

DETECT EARLY · RECOVER PRECISELY · GUARANTEE PROGRESS

- SACK + NACK** (with a checkmark icon): Explicit delivery state, Mandatory reliability feedback.
- Packet trim** (with a packet icon): Header survives the drop, Optional network acceleration.
- Probe** (with a probe icon): Elicit current SACK state, Optional proactive query.
- Timers + priority** (with a clock icon): Fallback and fast repair, Oldest gap retransmitted first.

Reordering is not loss. One loss should not erase later progress.

Logos at the bottom: AMD, BROADCOM, intel, Microsoft, NVIDIA, OpenAI.

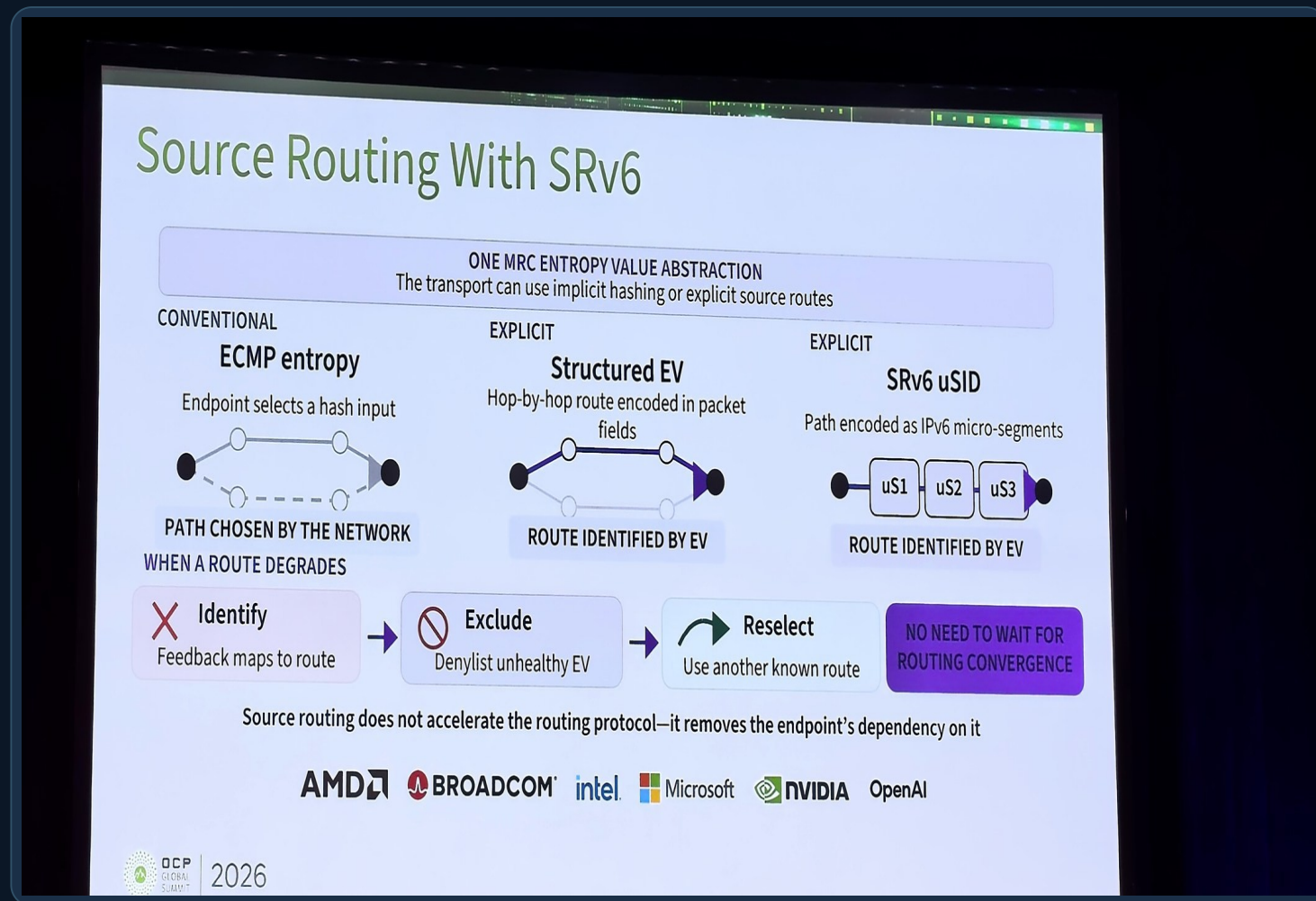
Footer: OCP GLOBAL SUMMIT 2026

现场照片 G2-04 · OCP Global Summit 2026

失败的 plane 需要被快速隔离

不是等网络慢慢重新收敛，而是路径表主动避障

- ECMP 与显式 SRv6 都能支持
- EV / Path liveness 识别异常
- 端点改变选路，不改变应用语义



现场照片 G2-02 · OCP Global Summit 2026

XPO：把可插拔模块推进高密度 AI 机架

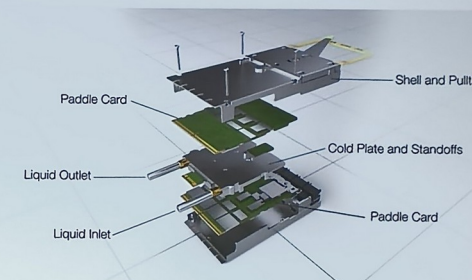
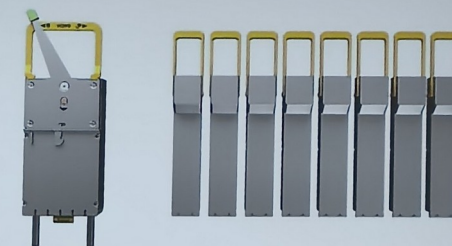
不是把 OSFP 简单做大，而是更彻底的系统接口

- XPO = eXtra Dense Pluggable Optics
- 12.8Tb/s 模块级设计目标
- 现场：液冷最高 500W / 模块 *

*500W：现场 25.6T coherent-lite 配置；3 月公告为 400W。

Intro to XPO

1. Highest Density Pluggable Optics form factor in the Industry
2. 204.8T per 10U with 200G/lane, 409.6T per 10U with 400G/lane
3. Supports all existing and emerging optics and interconnect standards
4. Liquid cooling up to 500W per module (for 25.6T coherent-lite)
5. Much lower failure rates with fewer components and lower temperatures
6. Vibrant XPO MSA ecosystem supported by all optics module vendors



XPO
OPTICS INNOVATION

XPO MSA Core Optics Module Members



2026

Scaling Innovation in the AI Era.

现场照片 G4-03 · OCP Global Summit 2026

12.8 Tb/s 是怎样堆出来的？

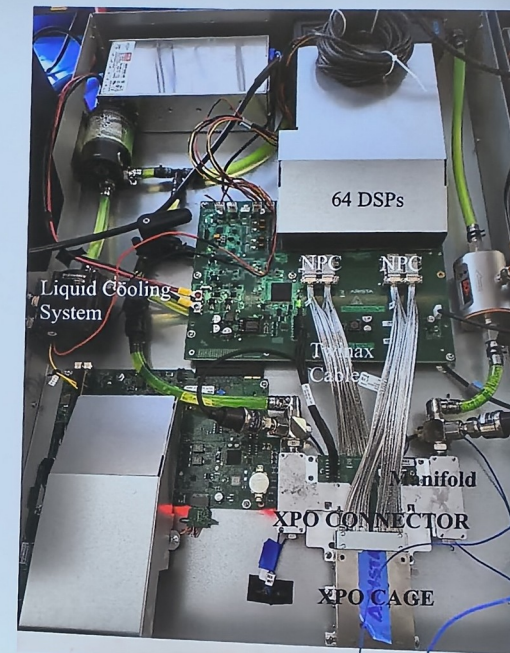
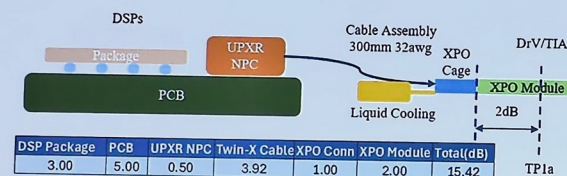
64 条高速电接口 lane × 200G 级

- $64 \times 200\text{G 级} = 12.8\text{Tb/s}$ (电接口)
- 现场: 204.8T/10U @200G/lane
- 500W 液冷能力* ; 不等于实测功耗

XPO Module Test Setup

An emulation box with 212 Gbps SerDes was designed by Arista to exercise the 64 channels on the XPO modules. This emulation platform featured the below characteristics.

- **x4 Agera3 Broadcom Retimers**
 - x16 lanes 212 Gbps each
- **NPC Copper Flyover Cables**
 - 300/600/900 mm long
 - 32 AWG Twinax: Amphenol XPO to Ultrapass-XR
- **Integrated Liquid Cooling**



2026

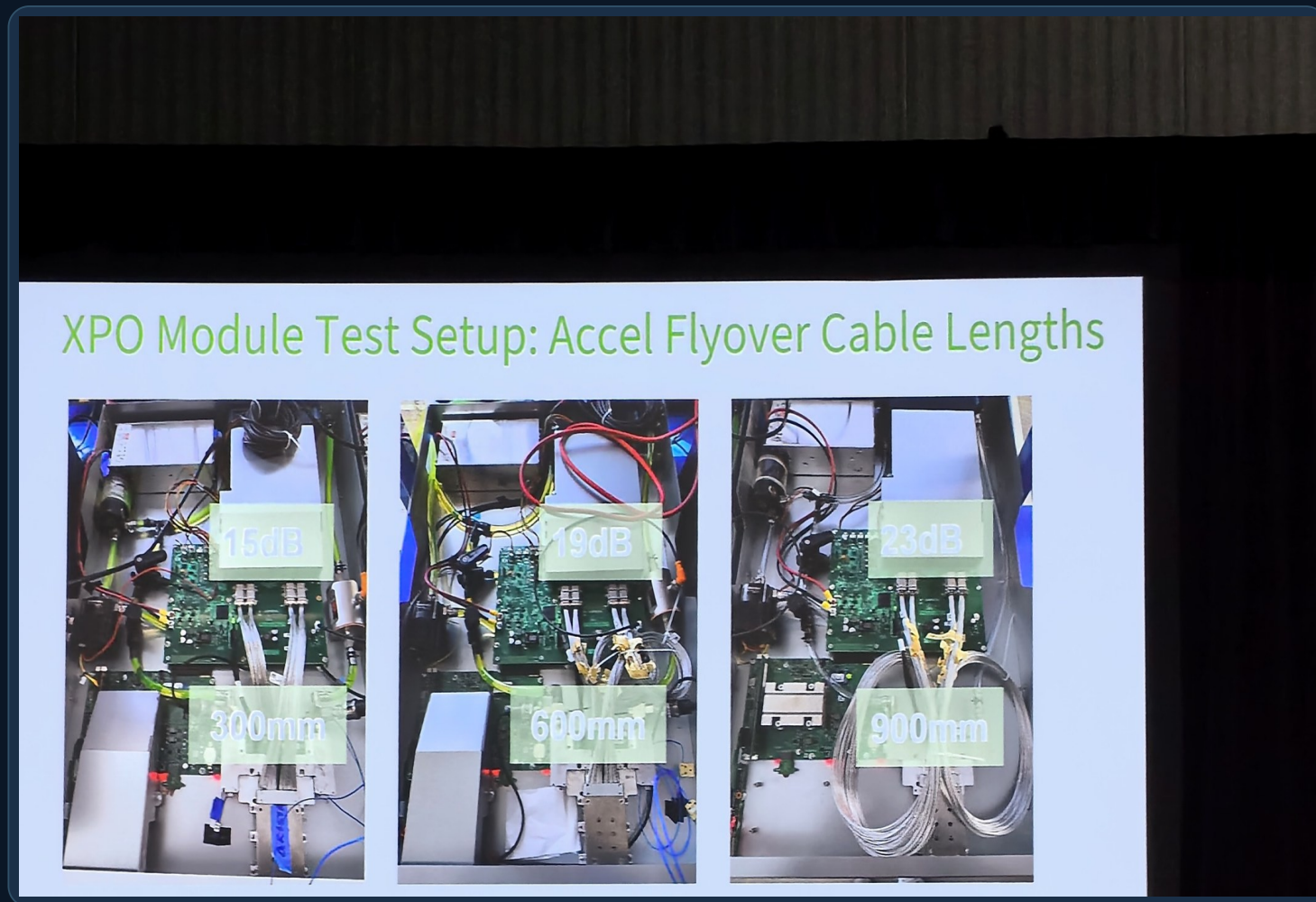
Scaling Innovation in the AI Era.

*500W : 现场 25.6T coherent-lite 配置 ; 3 月公告为 400W 。

一件容易忽略的事：先测电距离

现场用不同长度的电缆模拟真实系统

- 加速器侧：300 / 600 / 900 mm
- 对应约 15 / 19 / 23dB （特定配置）
- 交换侧固定 300mm / 15dB



现场照片 G4-07 · OCP Global Summit 2026

线性光模块能省下多少 DSP ?

图中配置：交换侧 LPO / 加速器侧 FRO

- 接收器均衡与 FEC 放在何处?
- BER 门槛和测试参考面要相同
- 更短电路径，才有机会减少有源均衡

LPO-FRO XPO Modules: BER/SNR

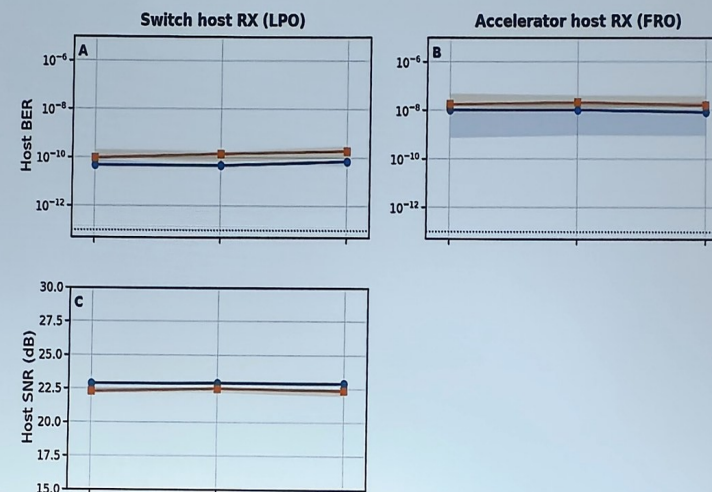
LPO-FRO signal integrity vs. bump-to-bump loss

— Optical attenuation: 0 dB — Optical attenuation: 3 dB

- Switch-side BER floor stays within 10^{-10}
- Accelerator-side BER remains within acceptable 10^{-7} range, primarily constrained by LPO Tx performance

Accelerator SNR omitted as it is not meaningful

Link performance expected to improve with standardized power drivers and CTLE settings

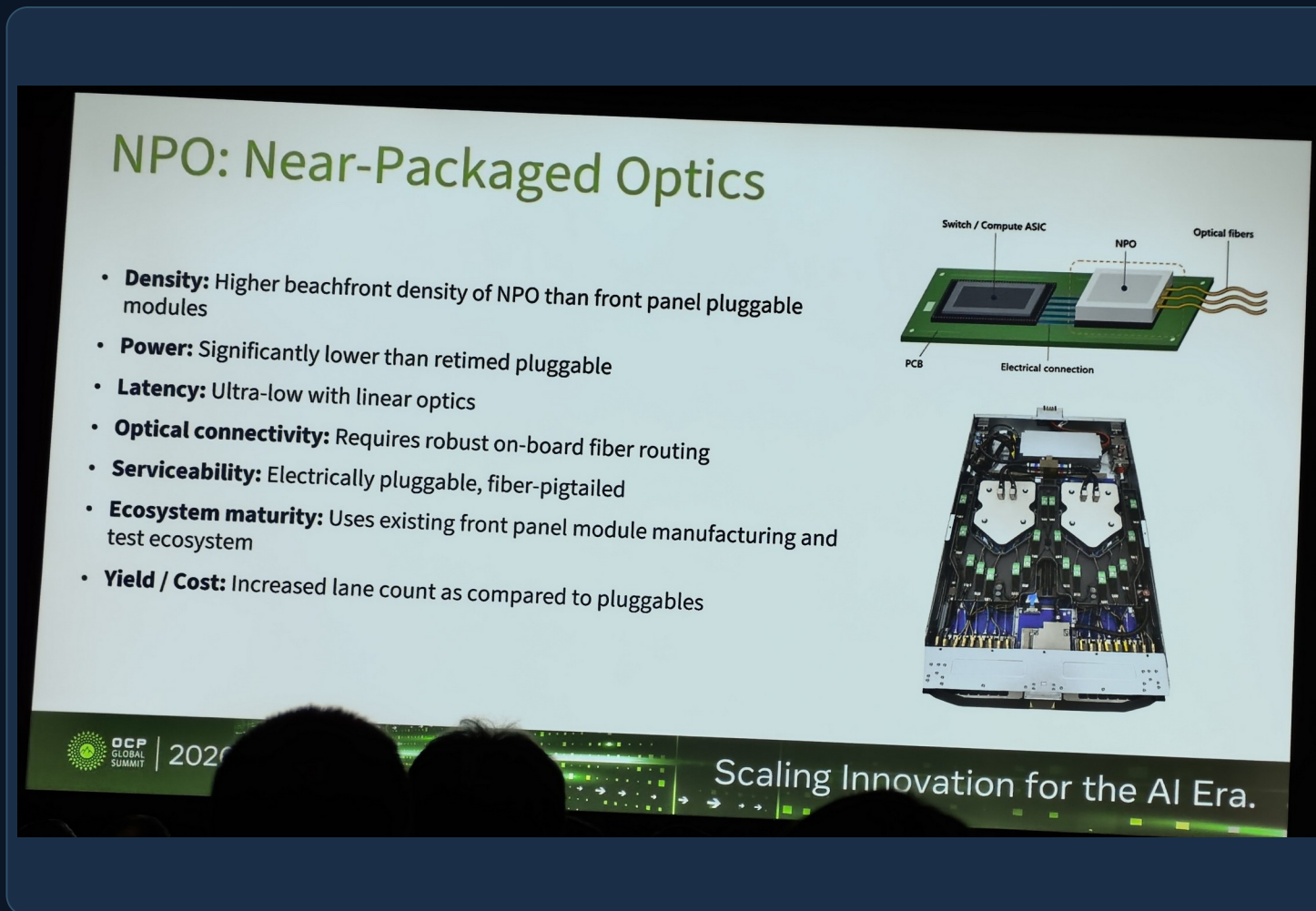


现场照片 G4-10 · OCP Global Summit 2026

NPO 的价值，不只是让光离 ASIC 更近

光引擎靠近芯片，但最好仍可维修、可替换

- Near-Packaged Optics 将高损耗电链路缩短
- Socket 必须兼顾信号完整性与热设计
- 要有第二供应商，接口得真正统一

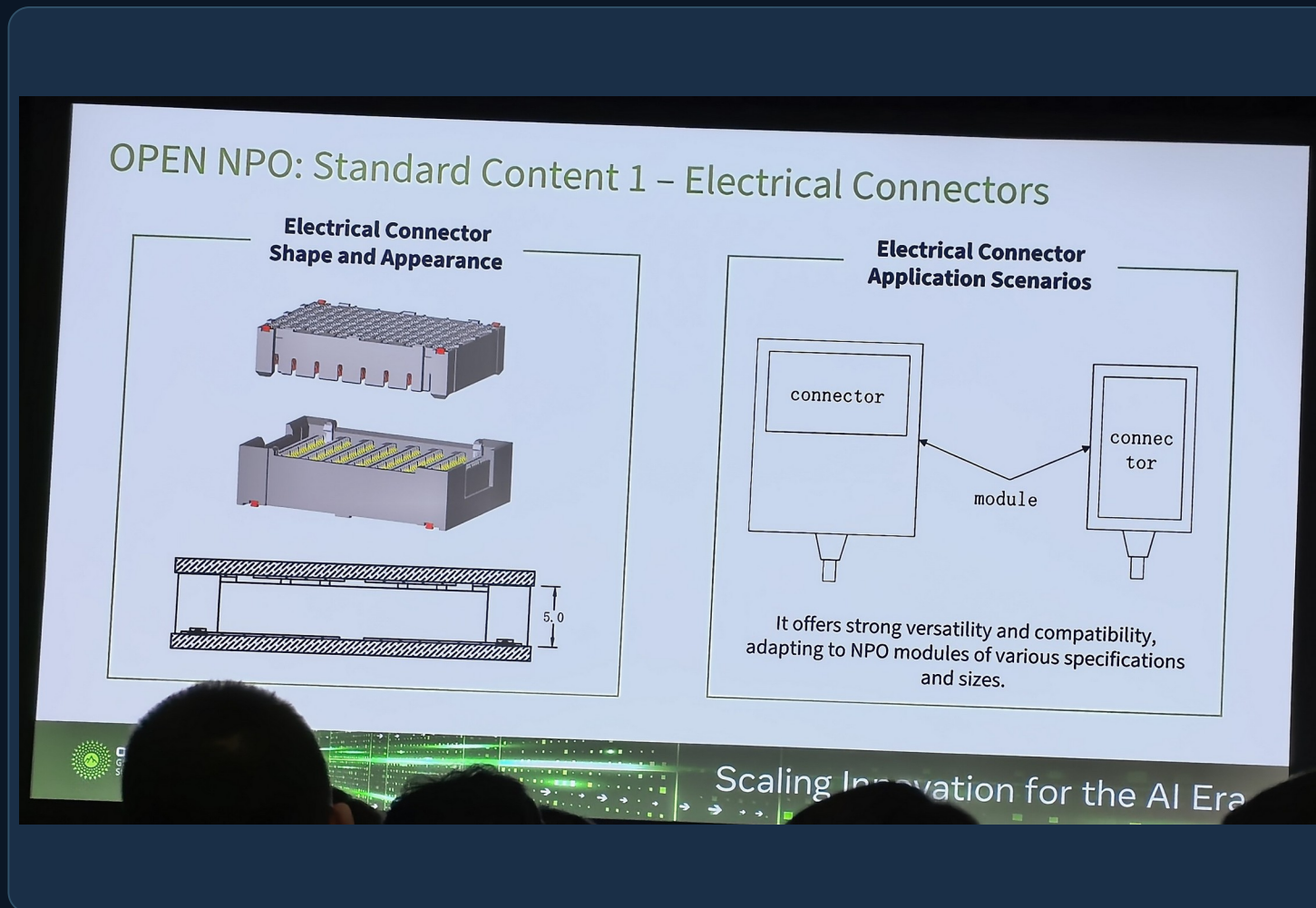


现场照片 G4-34 · OCP Global Summit 2026

开放接口必须从 connector 细节做起

机械、SI、供电、温度与测试边界共同定义

- 72 差分对（36 通道）+96 电源 / 低速 pin
- 现场 IL $\leq 1.5\text{dB}$: 10MHz-56GHz
- 供电地回流、ESD 与管理信号



现场照片 G4-21 · OCP Global Summit 2026

One size does not fit all

训练与推理需求不同；铜和光共存

- 训练示例：长链路、高对分带宽
- 推理示例：短突发、延迟与 TCO 敏感
- 按场景定义接口与可互换范围

One Size Does NOT Fit All

Training Cluster

Longer Links

- Long sustained all-reduce collectives
- High bisection bandwidth is critical
- Large physical distances between nodes

Inference Cluster

Shorter Links

- Short burst, latency sensitive traffic
- TCO sensitivity, prioritizes bandwidth per dollar
- Shorter physical differences between nodes (but growing!)

— Copper
— Optics

Technology choice is not binary – **copper and optics will coexist**

OCP GLOBAL SUMMIT | 2026

Scaling Innovation for the AI Era.

现场照片 G4-30 · Astera Labs 演讲页

XPO / NPO / CPO : 三条路线并存

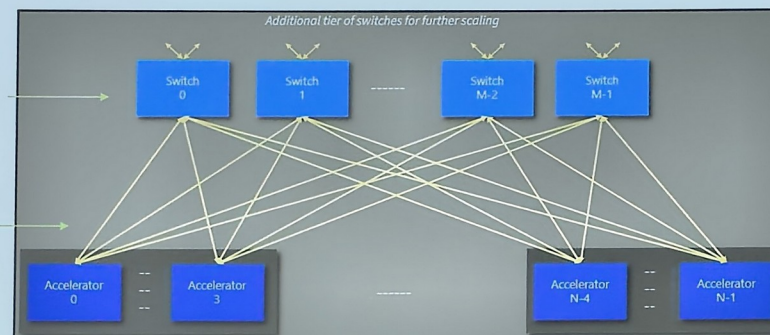
判断的对象是系统约束，不是缩写

- XPO : 密度 + 维修便捷
- NPO : 短电段 + 可替换光引擎
- CPO : 极短电段 + 封装耦合复杂

AI Scale-up System Topologies

High Radix switches in a Single Tier enable uniform communication domain

Reach, Bandwidth and Connectivity enabled by optical interconnects in Accelerator trays



Richly connected, Multi-Rack and Multi Row Scale-up Domains

Several topology variations are possible depending on target AI workload performance and cost



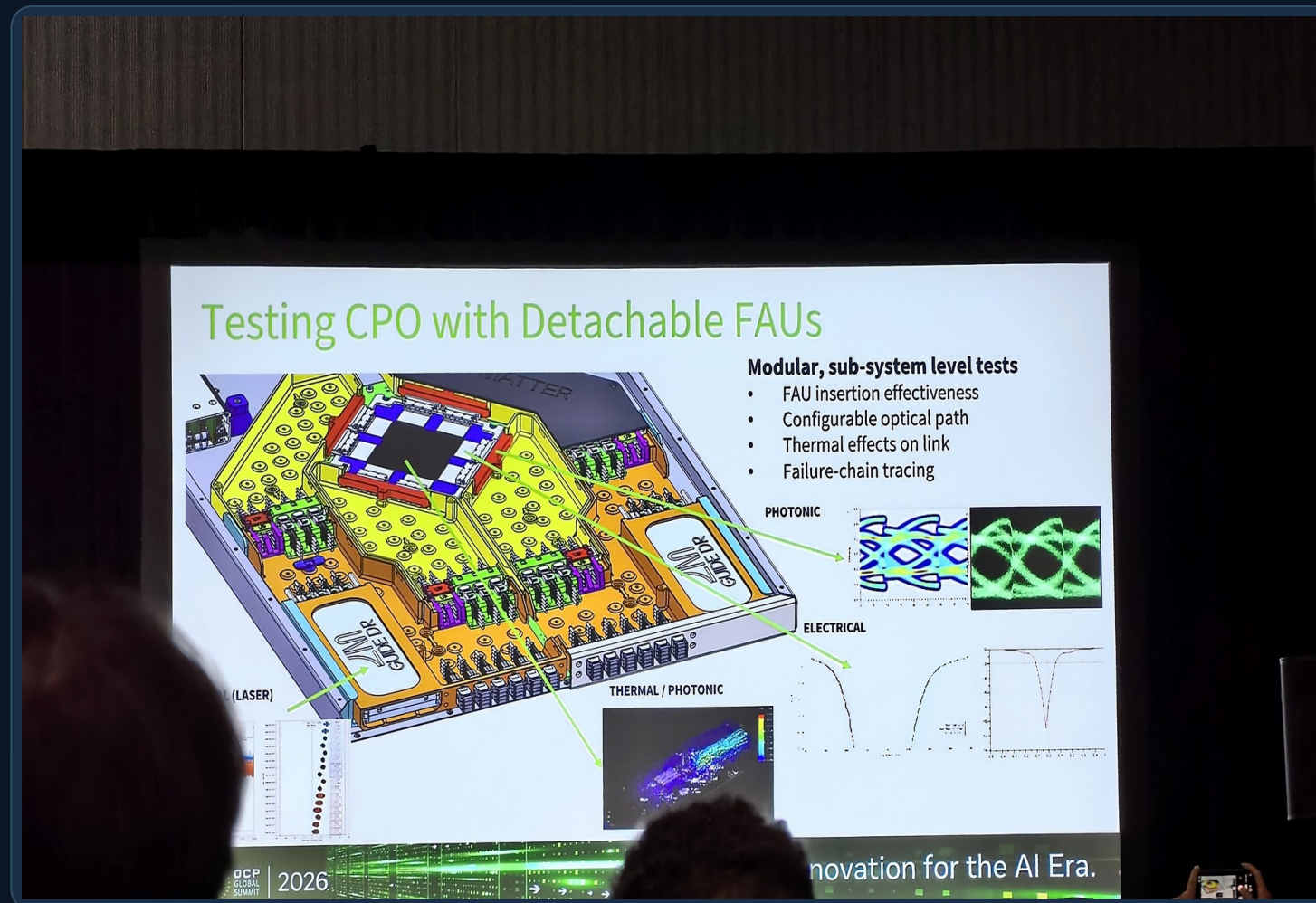
2026

Scaling Innovation for the AI Era.

从一张光学示意图到一条产线

FAU——Fiber Attach Unit——是关键工艺单元

- 被动对准要取代逐根主动找位置
- 可拆卸设计改变测试与维修流程
- 必须测分布、而不只是最佳样品

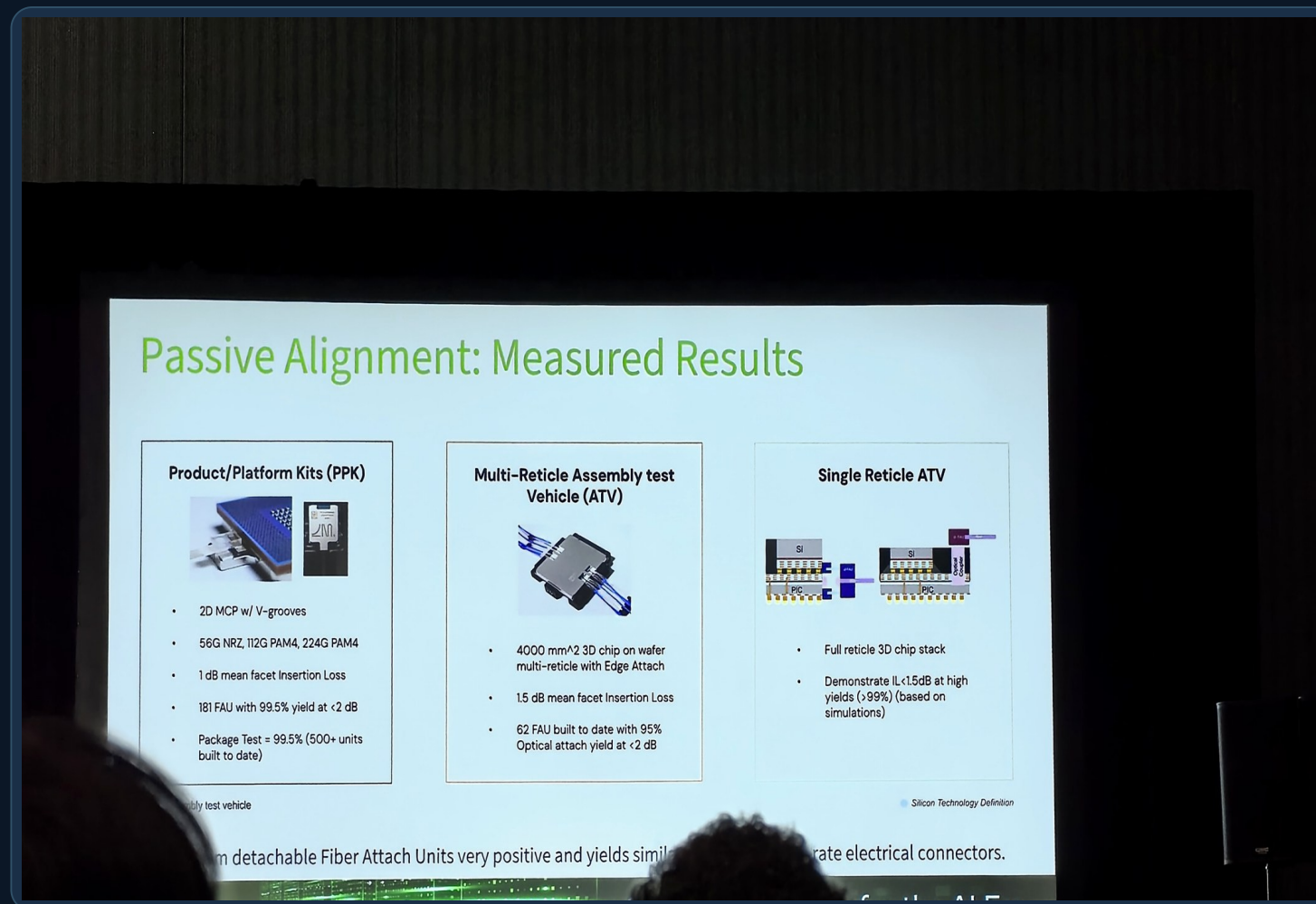


现场照片 G3-02 · OCP Global Summit 2026

181 个 FAU , 99.5% 达到 低于 2dB

对比实验必须区分不同 test vehicle 和分母

- 181 FAU / 99.5% / <2dB (PPK)
- 62 FAU / 95% / <2dB (另一载具)
- 单 reticle >99% 是模拟预测

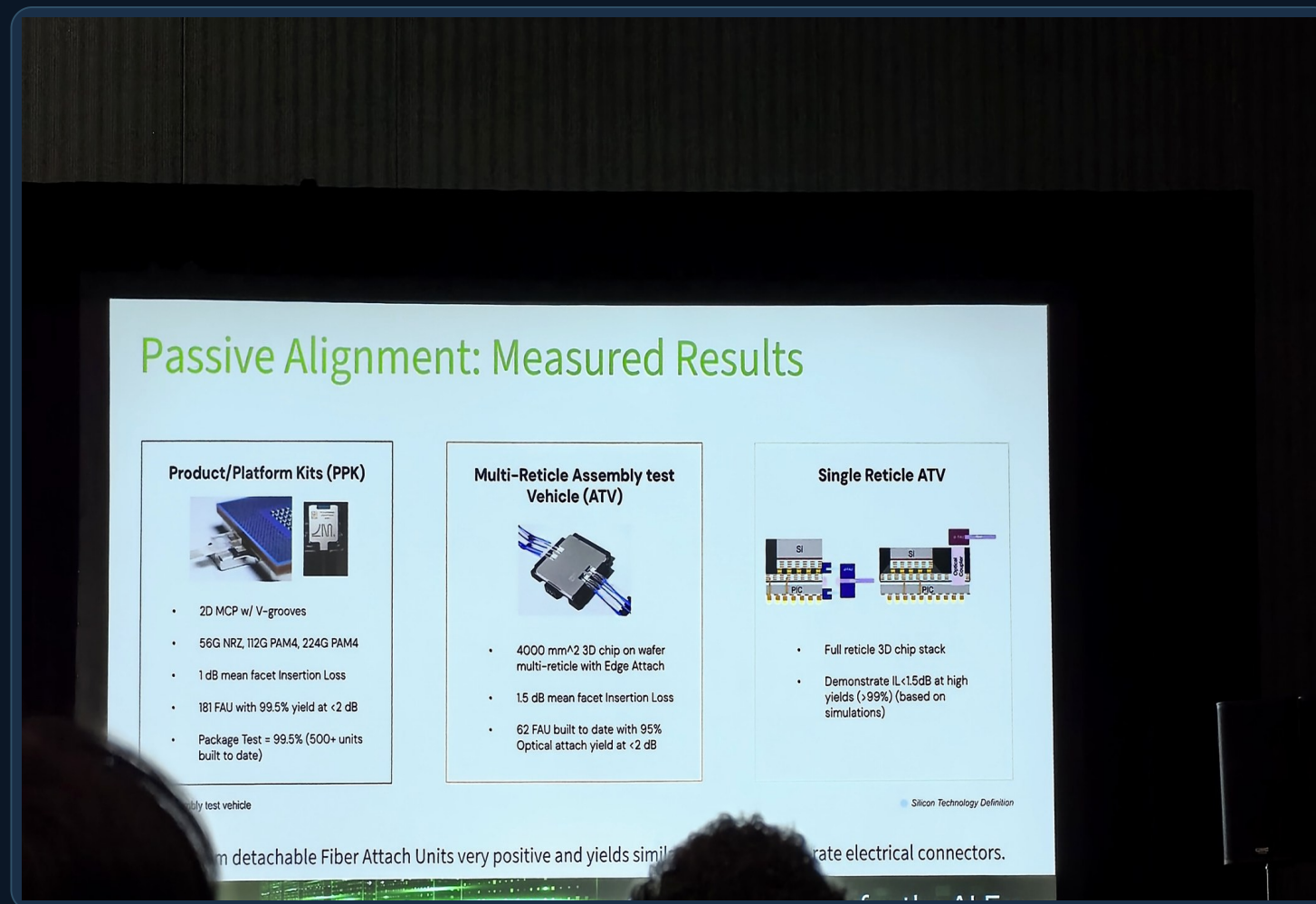


现场照片 G3-01 · OCP Global Summit 2026

从平均插损转向真实的 sign-off

公差链与最差分布，比平均值更接近量产难题

- IL P95/P99 / worst-channel
- 热循环、弯折、重复插拔
- 封装测试 99.5% / 500+ 件：另一分母

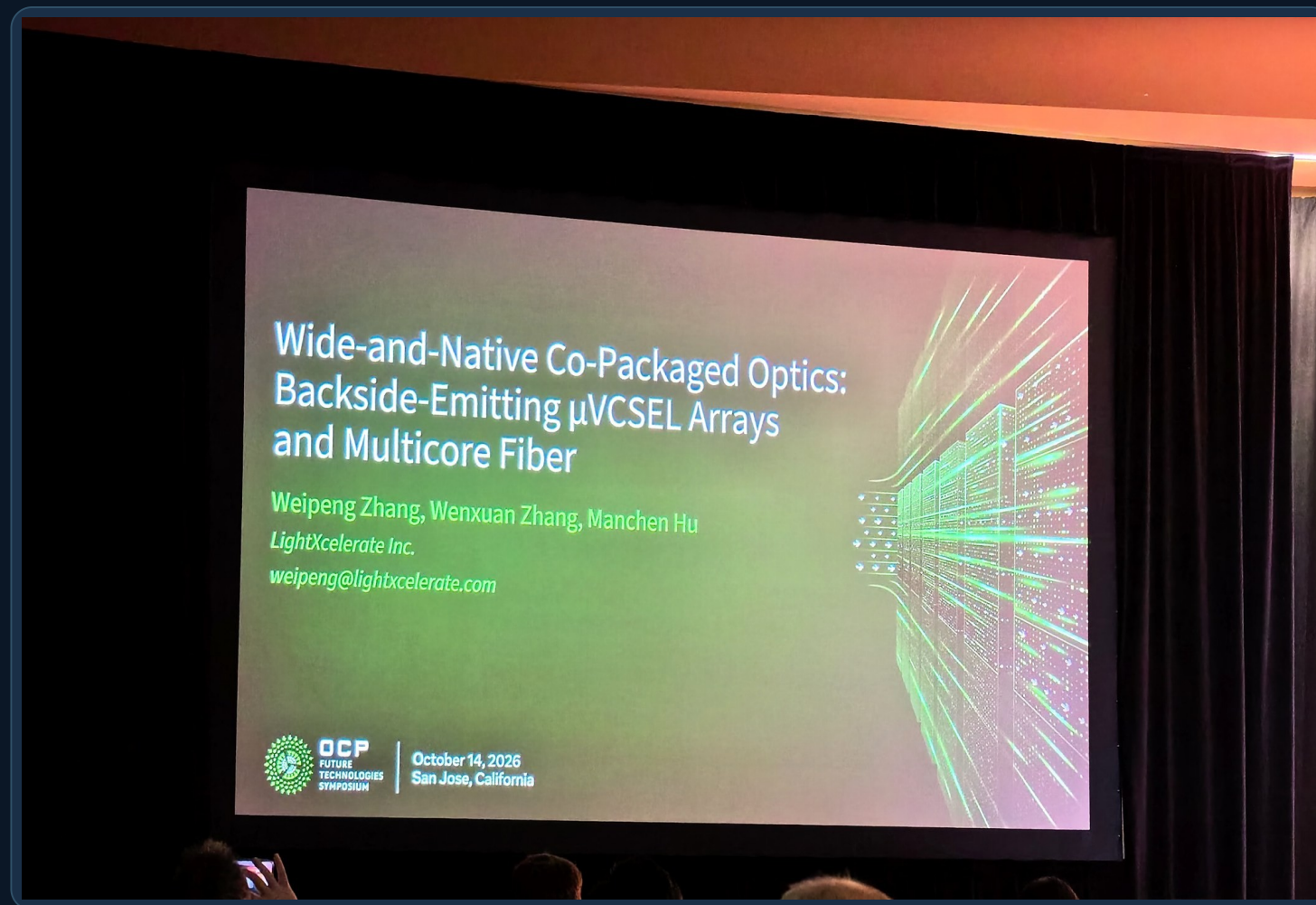


现场照片 G3-01 · OCP Global Summit 2026

LightXcelerate：新公司，明确的技术下注

把微型 VCSEL / PD / EIC 与多芯光纤做成可复制光引擎

- 核心团队：Manchen Hu / Weipeng Zhang
- OFC 2026 展示 19-core MCF hybrid chiplet
- OCP 现场 LightX-1 演示



现场照片 G1-06 · OCP Global Summit 2026

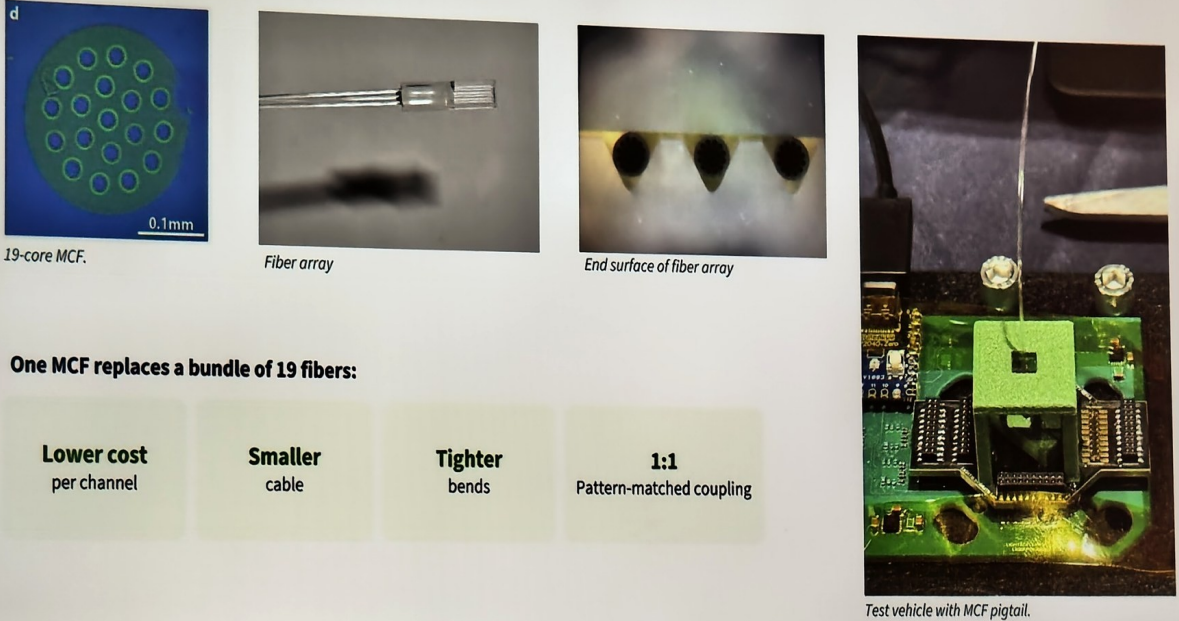
为什么是 57 路，而不是 64 路？

答案来自物理光纤结构：3 根 × 19 芯

- 三根 19-core MCF = 57 个芯
- 19 芯按几何阵列和光学器件一一配对
- 不是 57 根独立单芯光纤

μVCSEL Array × Multicore Fiber

Partnership with **YOFC**
Smart Link Better Life.



19-core MCF. 0.1mm

Fiber array

End surface of fiber array

Test vehicle with MCF pigtail.

One MCF replaces a bundle of 19 fibers:

Lower cost per channel	Smaller cable	Tighter bends	1:1 Pattern-matched coupling
----------------------------------	-------------------------	-------------------------	--

OCP FUTURE TECHNOLOGIES SYMPOSIUM | 2026

Scaling Innovation for the AI Era.

现场照片 G1-11 · OCP Global Summit 2026

LightX-1 的公开规格究竟说了什么？

演示页给出 57 Tx + 57 Rx，但不能等同全链验收

- 规格页标注：57 Tx + 57 Rx
- 32Gb/s/lane 当前演示，64G roadmap
- 1.5×0.6mm 尺寸定义有冲突；50m / <1.5pJ/bit

The slide titled "LightX-1: A Modular μVCSEL Optical-Engine Block" features a large photograph of the device with three glowing circular patterns. Below the photo is a schematic diagram of three 19-channel μVCSEL subarrays, each labeled "Array 1", "Array 2", and "Array 3". To the right of the photo, a table lists key specifications:

57 lanes 57 Tx + 57 Rx	1.5 × 0.6 mm engine footprint
32 Gb/s per lane today, roadmap to 64G	3.2 Tb/s per engine
< 1.5 pJ/bit full link	50 m reach

Three 19-channel μVCSEL subarrays under drive, with per-channel addressing (inset).

μVCSELs ~25× smaller than conventional VCSELs; roadmap to 100 and 200 Gb/s per lane.

OCP logo and "Scaling Innovation for the AI Era" text are visible at the bottom of the slide.

型号：现场 LightX-1；公告 LXC-1，关系待确认。

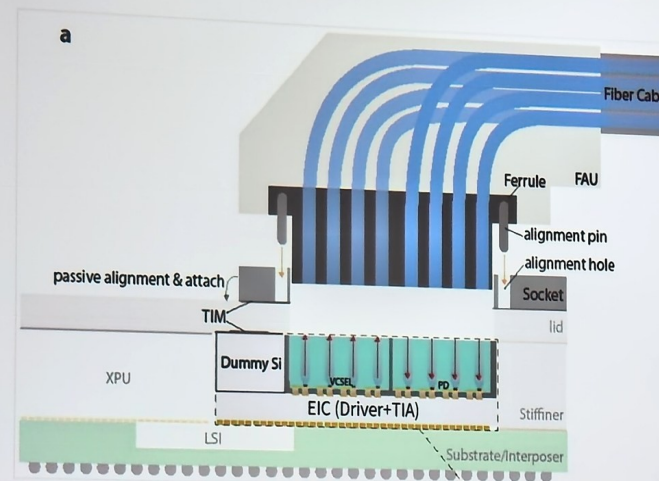
现场照片 G1-10 · OCP Global Summit 2026

为什么选择背面出光与背面入光？

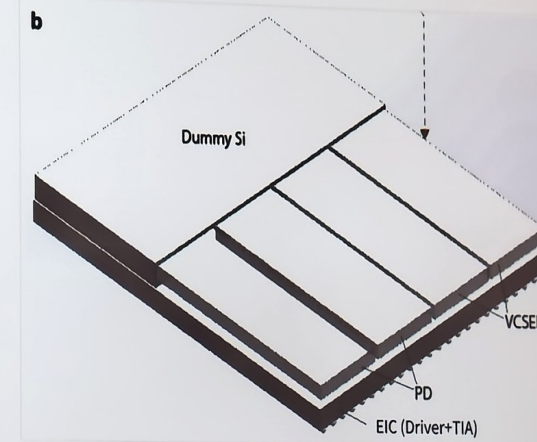
把高速电子链路尽量缩短

- BSE micro-VCSEL 与 BSI-PD
- Driver/TIA EIC 通过 Cu pillars 接近光芯片
- FAU socket + alignment 支持被动耦合

Optical Engine Architecture



(a) Cross-section: engine beside the XPU on a shared substrate/interposer.



(b) Chiplet: VCSEL and PD dies tiled on the EIC.

1 Passive fiber attach
FAU pins drop into a socket on the lid.

2 Backside optics
BSE VCSELs + BSI PDs flip-chipped face-down.

3 EIC underneath
Driver + TIA; μm -scale Cu-pillar paths.

4 Thermal path
Dummy Si + TIM carry heat into the lid.



2026

Scaling Innovation for the AI Era.

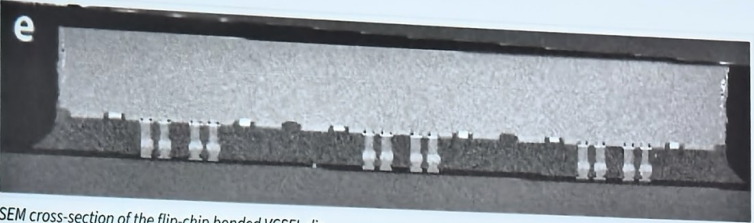
现场照片 G1-09 · OCP Global Summit 2026

Cu pillar、散热和 2.5D： 真正的工程活

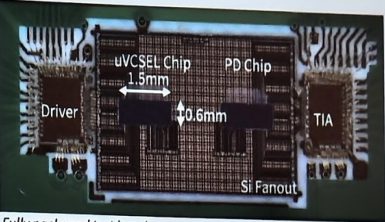
高频信号路径短，装配与热路径就更重要

- flip-chip + Cu pillars 缩短高频互连
- dummy silicon / TIM / lid 引导热流
- 1.5×0.6mm 在实物图中标于 μ VCSEL 芯片

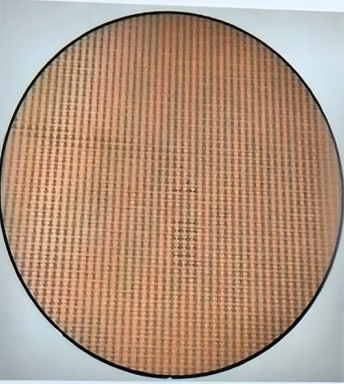
Flip-Chip Assembly with Cu Pillars



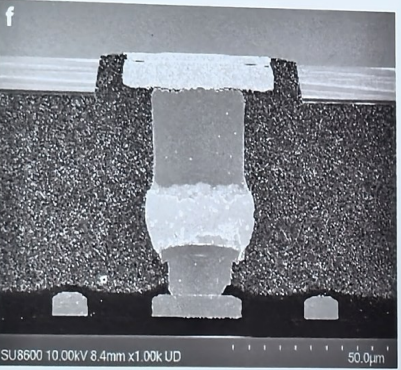
SEM cross-section of the flip-chip bonded VCSEL die.



Fully packaged test bench



12in Interposer wafer.



Magnified view of a bonded Cu pillar.

Built on OSAT-ready steps

- Cu-pillar flip-chip bonding of III-V dies
- Underfill for mechanical robustness
- TSVs route vertical I/O through the EIC
- Passive mechanical fiber alignment
- Conventional package integration

2026

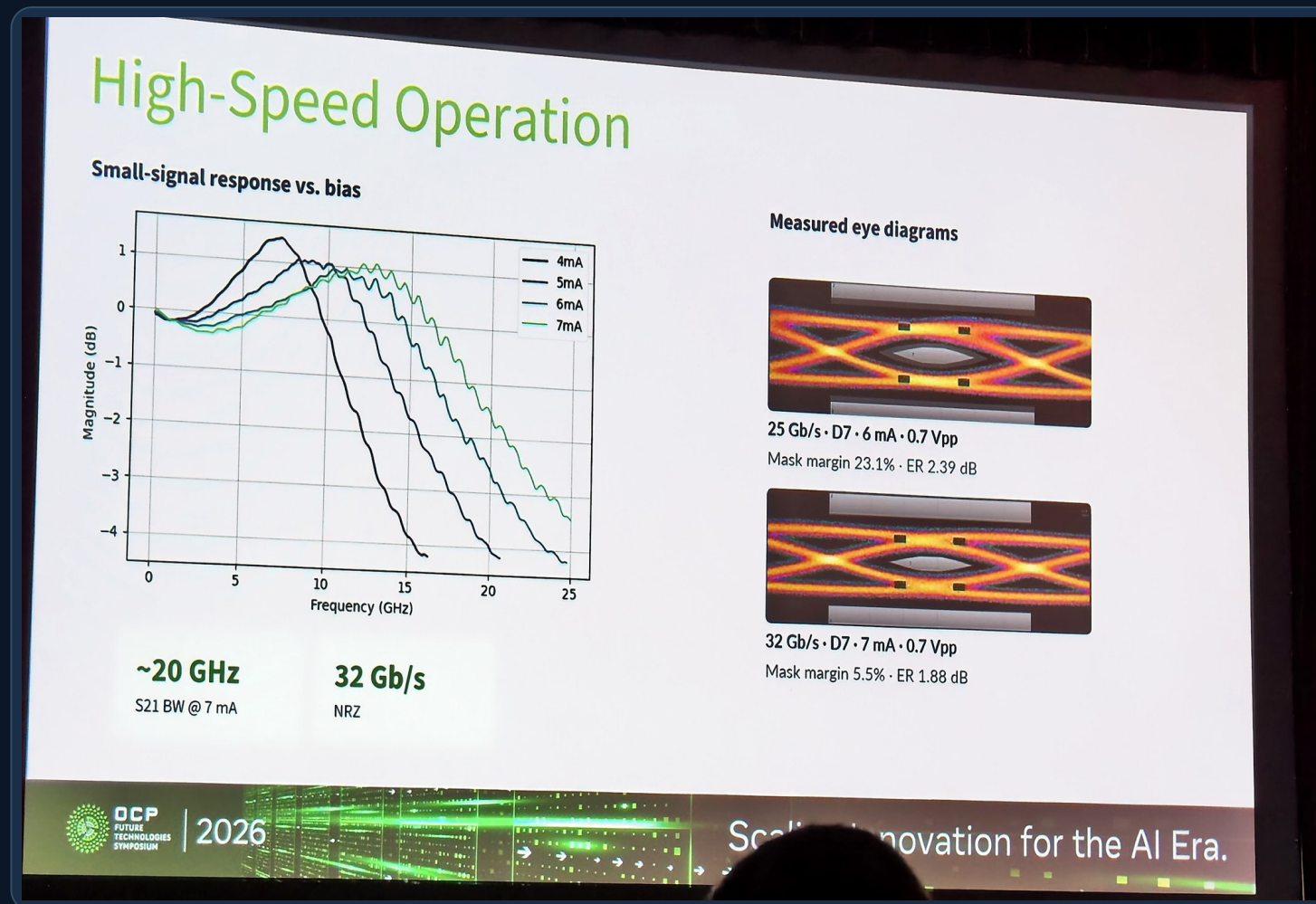
Scaling Innovation for the AI Era.

现场照片 G1-12 · OCP Global Summit 2026

25G 到 32G：眼图没有那么“免费”

D7 两测试点：偏置也由 6mA 变为 7mA

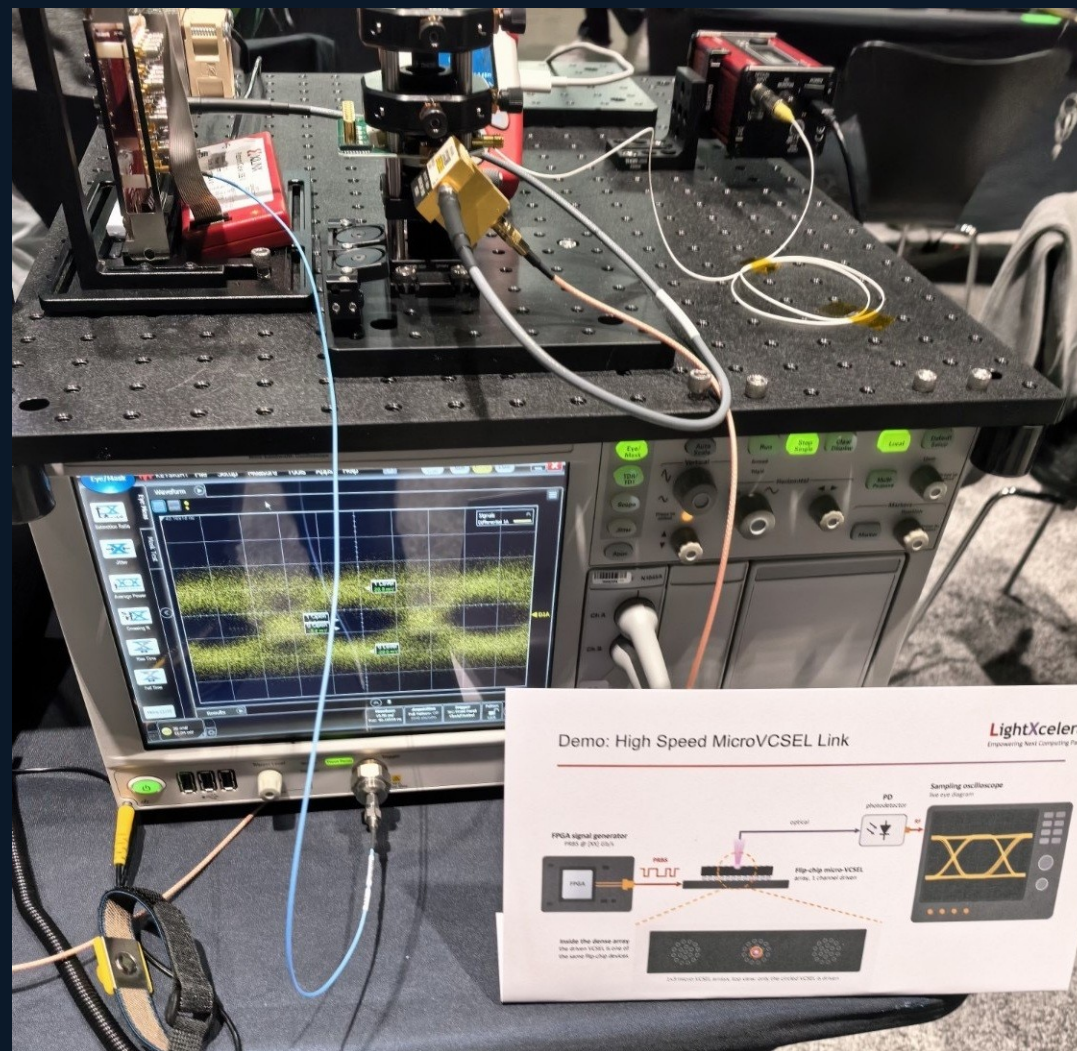
- 25G NRZ：Mask margin ~23.1%，ER ~2.39dB
- 32G NRZ：Mask margin ~5.5%，ER ~1.88dB
- 代表性单路，不等同全阵列 PVT



现场照片 G1-16 · OCP Global Summit 2026

高速链路演示：整体测试台与展板

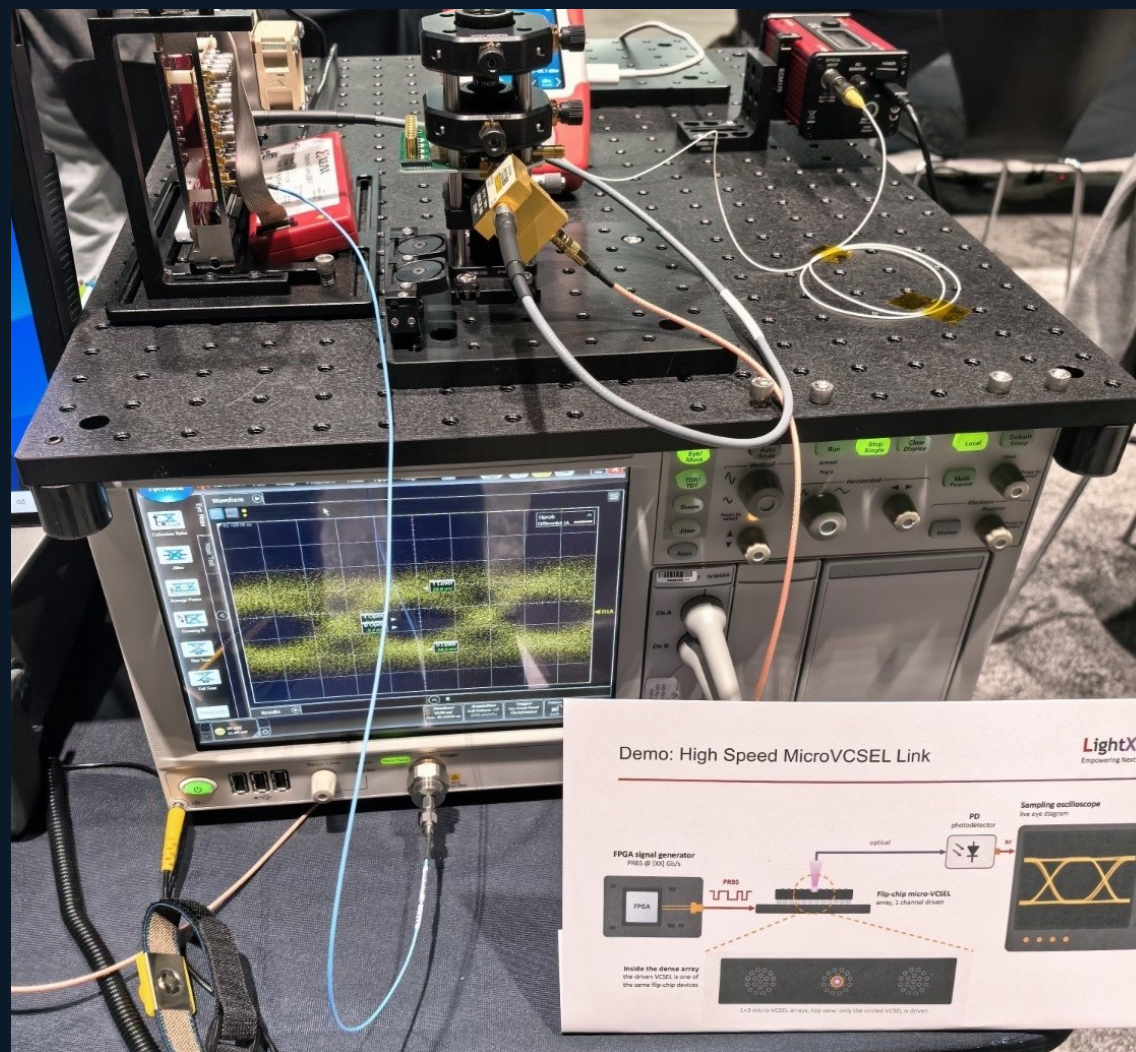
展板标注“1 channel driven”。速率栏为“[XX] Gb/s”占位符，未写明演示速率。单路眼图不能证明全部 57 路同时高速运行。



现场补图 78702 · 仅截取技术画面

高速链路演示：测试台 补充视角

第二视角保留器件、探测端、示波器和连接关系。照片没有给出可独立核验的完整测试配置，速率、BER 与全阵列工作状态应以正式测试记录为准。

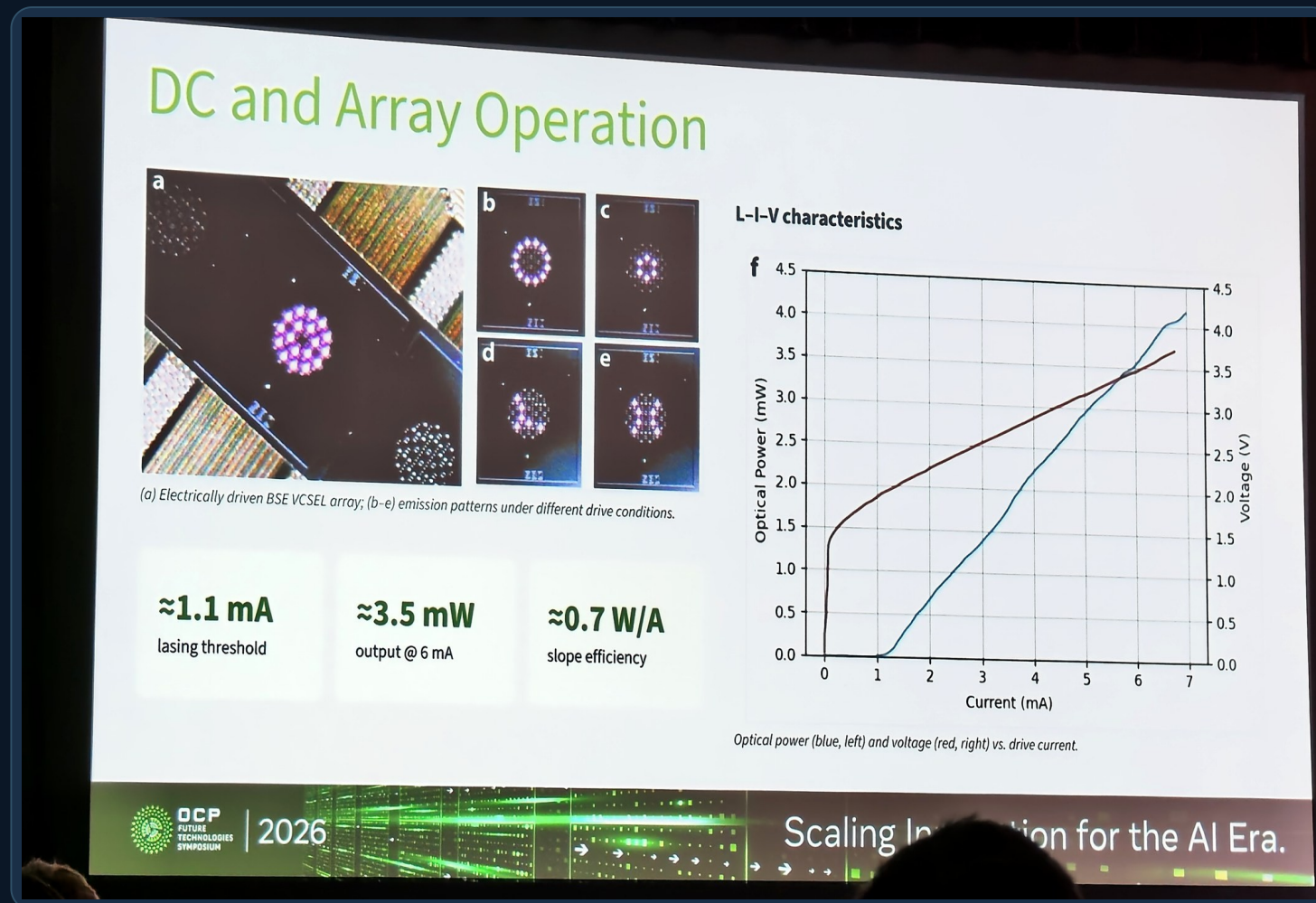


现场补图 78703 · 仅截取技术画面

DC、斜率效率与高频带宽的关系

看一颗器件能不能亮，不够；看阵列分布才有意义

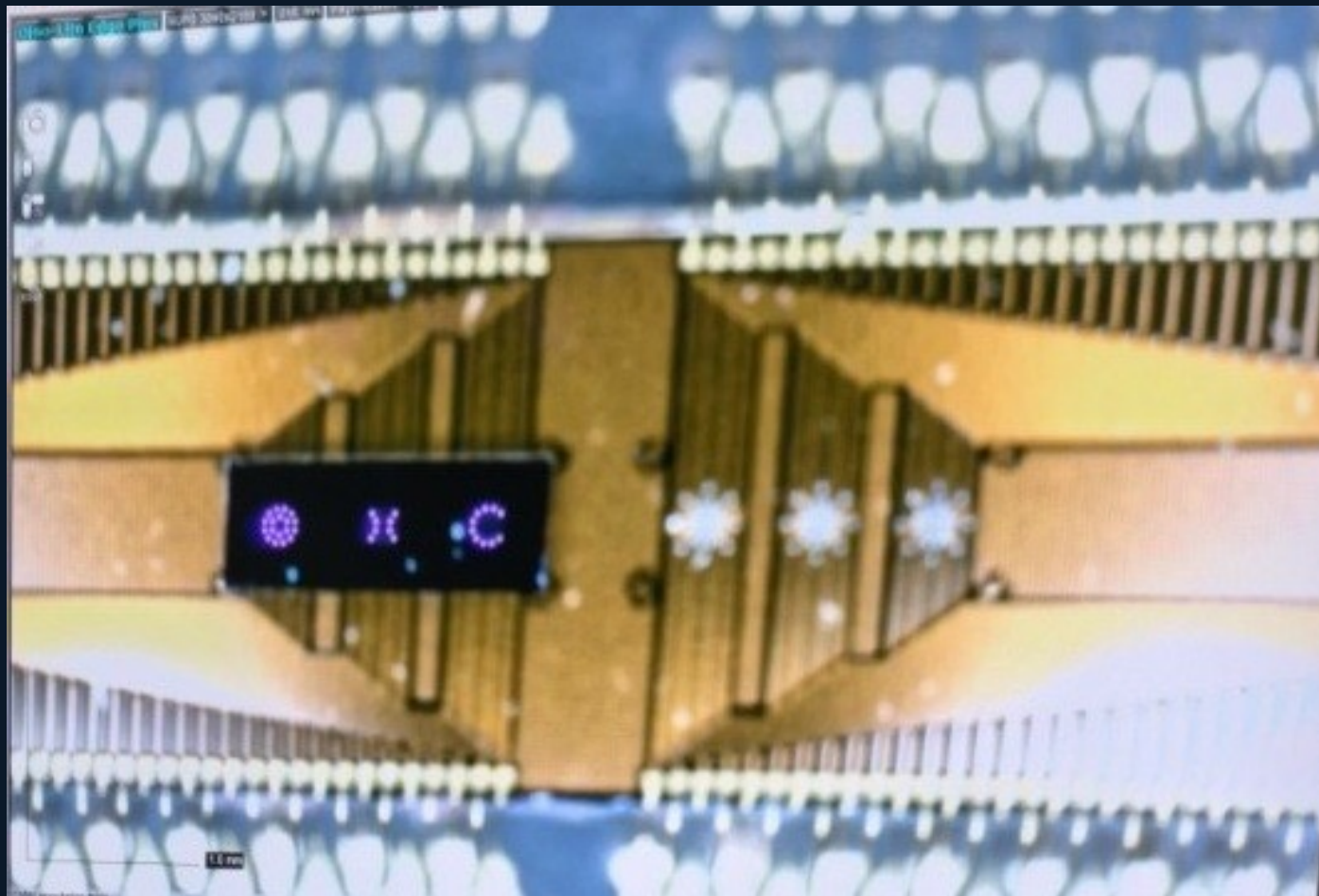
- VCSEL threshold $\sim 1.1\text{mA}$
- 约 3.5mW @ 6mA ；斜率效率 $\sim 0.7\text{W/A}$
- 20GHz 量级响应示例 @ 7mA



现场照片 G1-13 · OCP Global Summit 2026

阵列点亮：三个子阵列的选择性图案

屏幕展示三个阵列可选择性点亮的图案。这是阵列寻址 / 控制的现场证据，不等同于所有发射与接收通道同时完成高速传输。图片仅截取技术画面。



现场补图 78704 · 仅截取技术画面

MicroVCSEL 与多芯光纤耦合

实物显示 19 芯端面、光纤与测试载具。三根 19 芯光纤合计 57 个物理芯；这些芯与 57 Tx + 57 Rx 的映射仍需拓扑图。

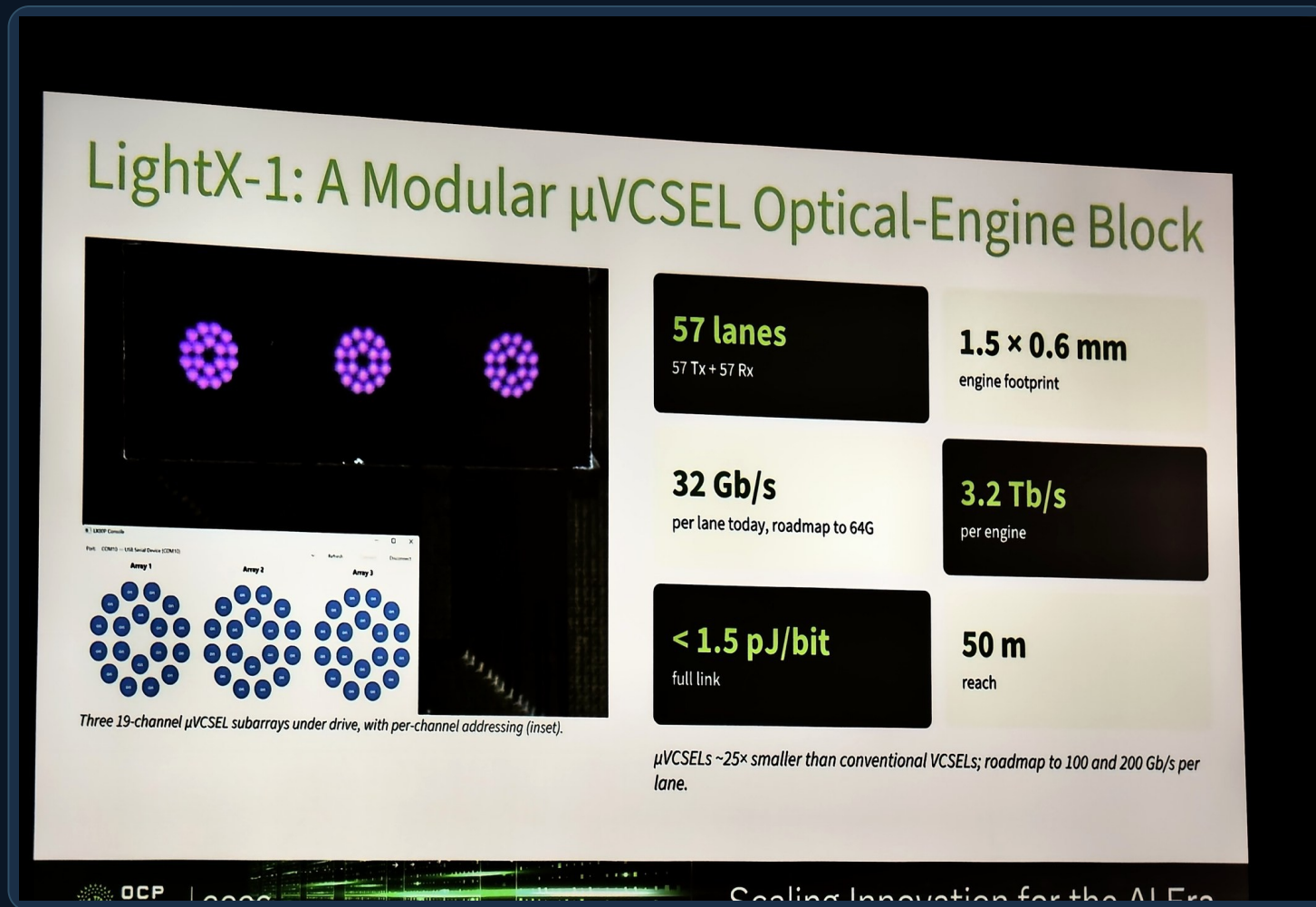


现场补图 78705 · 仅截取技术画面

3.2 Tb/s：最该追问的带宽口径

57 × 32G 与演示页宣称的 3.2T 不能直接画等号

- 条件计算： $57 \text{ Tx} \times 32\text{G} = 1.824\text{Tb/s}$
- 若两向各 57 路同速，算术和为 3.648Tb/s
- 3.2Tb/s per engine 的定义待确认



The slide titled "LightX-1: A Modular μVCSEL Optical-Engine Block" features a central photograph of three glowing purple circular light arrays. Below the photo is a schematic diagram of three 19-channel μVCSEL subarrays, each labeled "Array 1", "Array 2", and "Array 3". To the right of the photo, six key performance indicators are listed in colored boxes: "57 lanes" (57 Tx + 57 Rx), "1.5 × 0.6 mm" (engine footprint), "32 Gb/s" (per lane today, roadmap to 64G), "3.2 Tb/s" (per engine), "< 1.5 pJ/bit" (full link), and "50 m" (reach). At the bottom right, a note states: "μVCSELs ~25× smaller than conventional VCSELs; roadmap to 100 and 200 Gb/s per lane." The slide footer includes the OCP logo and the text "Scaling Innovation for the AI Era".

LightX-1: A Modular μ VCSEL Optical-Engine Block

57 lanes
57 Tx + 57 Rx

1.5 × 0.6 mm
engine footprint

32 Gb/s
per lane today, roadmap to 64G

3.2 Tb/s
per engine

< 1.5 pJ/bit
full link

50 m
reach

Three 19-channel μ VCSEL subarrays under drive, with per-channel addressing (inset).

μVCSELs ~25× smaller than conventional VCSELs; roadmap to 100 and 200 Gb/s per lane.

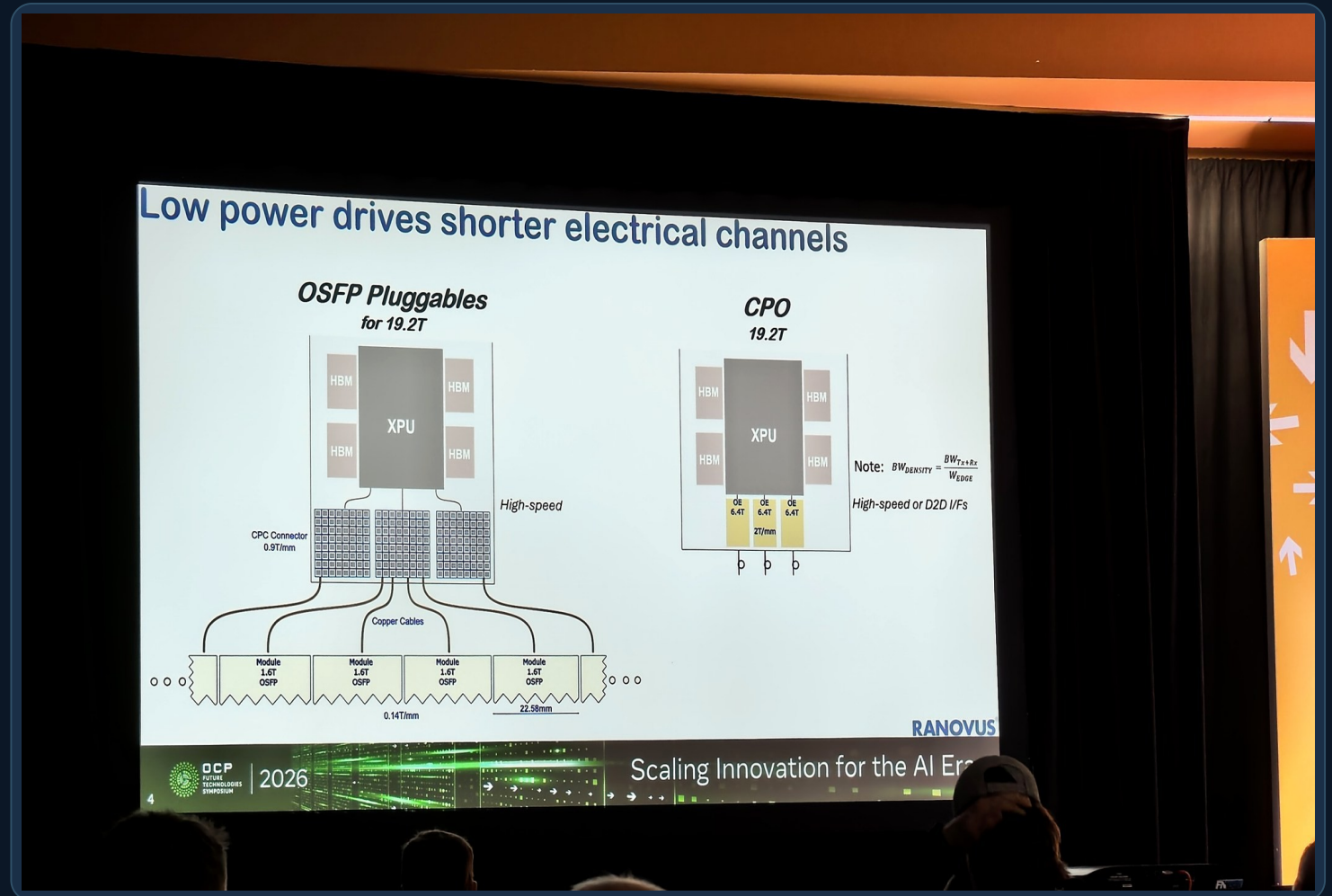
OCP Scaling Innovation for the AI Era

现场照片 G1-10 · OCP Global Summit 2026

Wide & Parallel 有什么不同于高速串行?

SerDes 省下的功耗，会不会被其它模块拿走?

- 低每线速率减少电损耗与均衡需求
- 但更多 lanes 增加 bias / 控制 / 测试复杂度
- 不能跳过 gearbox、FEC、封装的整体账



现场照片 G1-01 · Ranovus 演讲页

整机可靠性是一条贯穿线

MRC、FAU、VCSEL 看似无关，其实都在谈故障成本

- 单路眼图 → 全阵列 BER
- FAU 装配良率 → 维护与返修
- 链路 flap → 整个 AI 作业的吞吐恢复

Why reducing link errors for scale-up is important

Highly simplified scale-up example using CBFC (credit-based flow control)

Time

t = 0: Sender Transmits Block #1

t = 1: Receiver Receives Block (Correct)

t = 1: Sender Transmits Block #2

t = 2: Receiver Receives Block (ERROR)

t = 2: Sender Re-transmits Block #2

t = 3: Receiver Receives Block (Correct)

Matrix Multiplication: $C = A \cdot B$
 $C_{ij} = \sum_k A_{ik} \cdot B_{kj}$

Job partitioned into tiles assigned to accelerators (exchanging large blocks of data)

- ▶ The following block is only transmitted once the current block is received correctly
- ▶ Periodic retransmissions drive tail latency & reduces effective interconnect BW
- ▶ Link flaps occur when the number of retries exceeds the retry limit → focus on minimizing burst errors

Minimizing burst errors to prevent link flaps is a key objective to reduce tail latency for scale-up

5 OCP 2026 | 2026

Scaling Innovation for the AI Era.

现场照片 G1-02 · Ranovus 演讲页

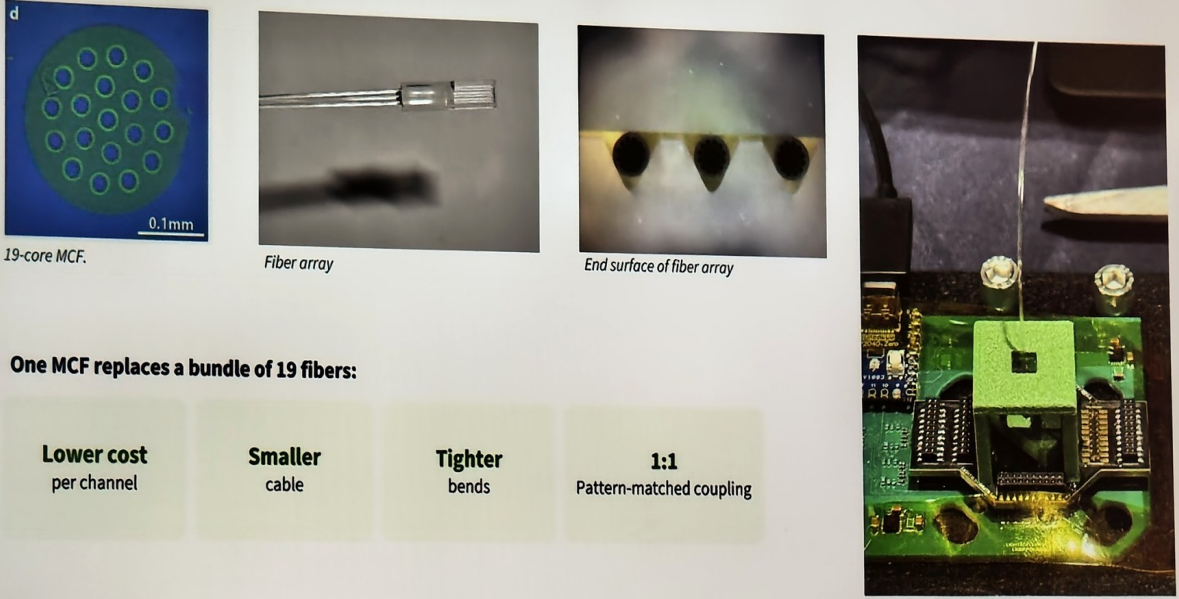
带走六个可验证的问题

从技术营销，回到统一测量定义与工程可行性

- 3.2Tb/s 的方向和编码口径？
- 全 57 路并行 BER、PVT、串扰？
- <1.5pJ/bit 包含哪些模块？
- 3×19 芯 MCF 收发拓扑、对准良率？
- 维修 / 热循环 / FAU 寿命怎么测？
- XPO/NPO/CPO 哪个场景经济性最好？

μ VCSEL Array $\times$ Multicore Fiber

Partnership with **YOFC**
Smart Link Better Life.



One MCF replaces a bundle of 19 fibers:

Lower cost per channel	Smaller cable	Tighter bends	1:1 Pattern-matched coupling
----------------------------------	-------------------------	-------------------------	--

OCP FUTURE TECHNOLOGIES SYMPOSIUM | 2026

Scaling Innovation for the AI Era.

现场照片 G1-11 · OCP Global Summit 2026