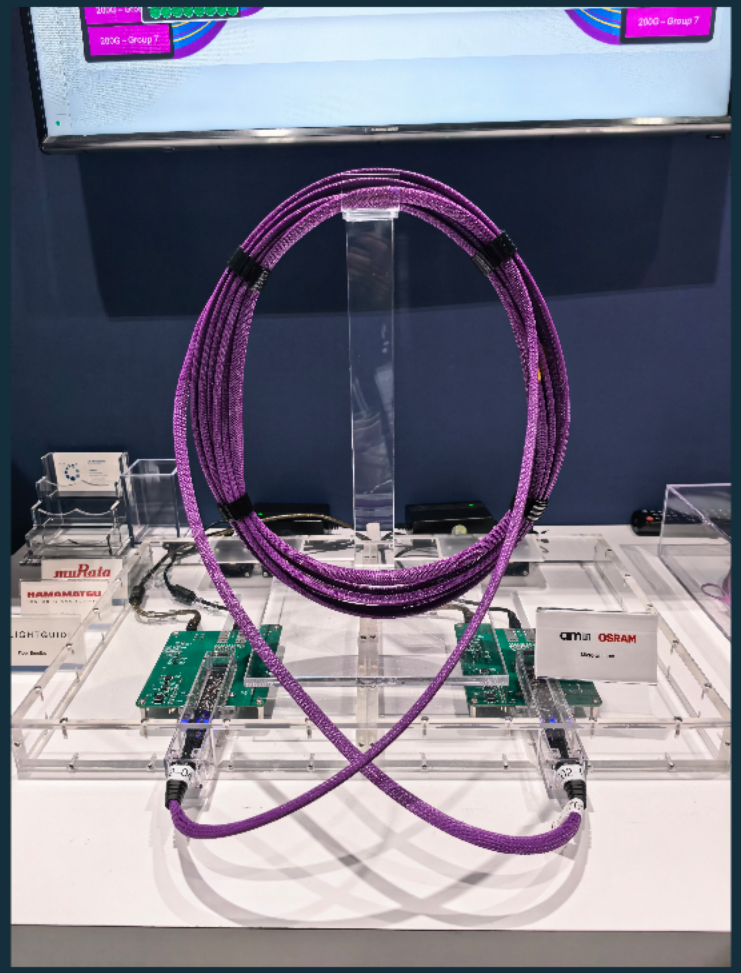
SYSTEM BOUNDARY 系统边界

OCP 首日现场记

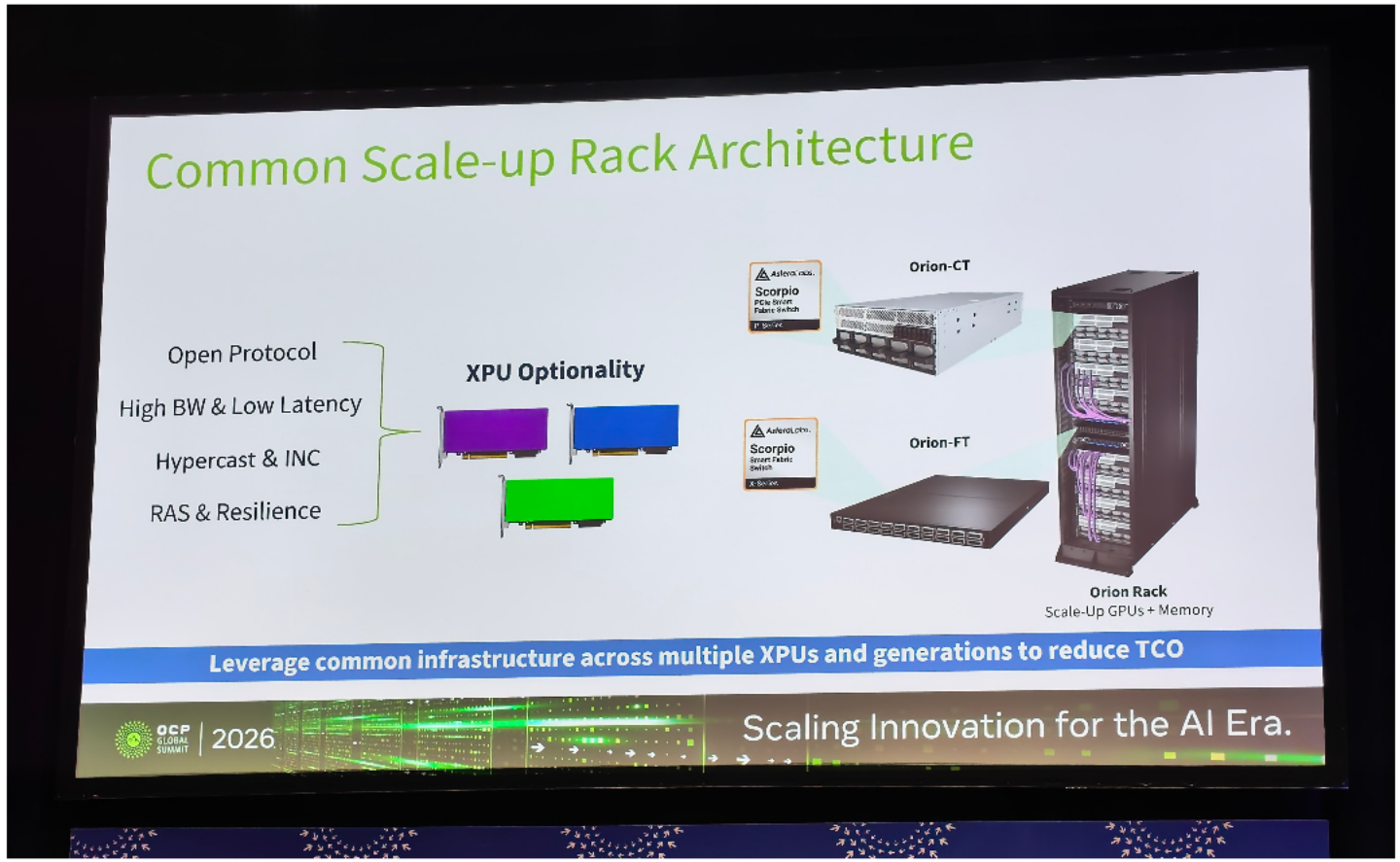
机架 互连与内存

Astera Labs · OIF 448G · Meta LPO
Credo microLED 与展会技术观察



Zach Yuan
Zach V2 本人授权合成声音
2026 年 10 月 13 日现场记录

Astera 的开放机架



公司位置

PCIe 与 CXL 连接芯片、交换 fabric 和管理软件

现场主张

计算、fabric、内存和链路可以分别演进

评估入口

模型、上下文和并发先决定扩展方式

PCIe fabric 的有效吞吐



现场平台

Scorpio X 320 lane PCIe 6
与开放机架

软件条件

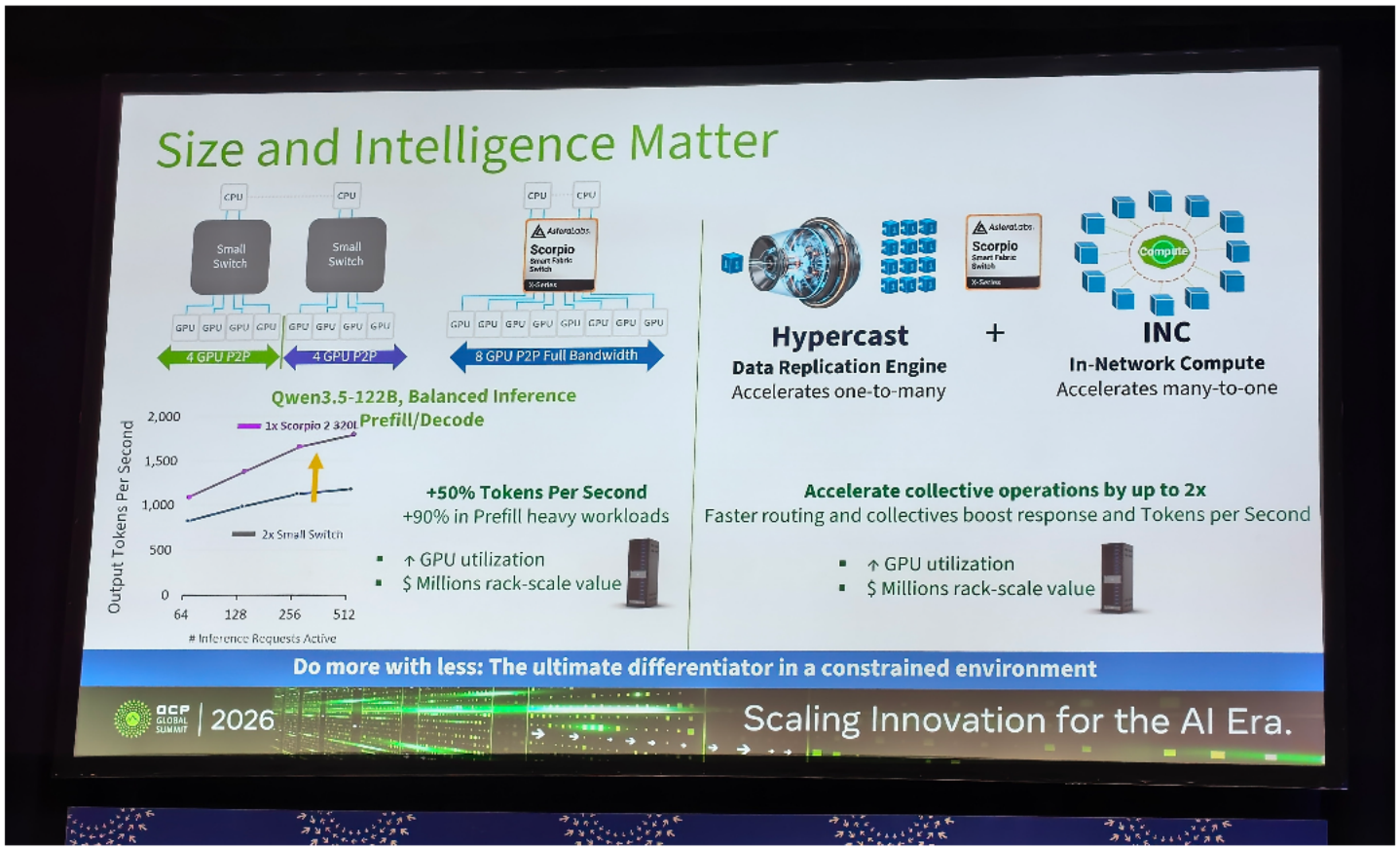
P2P、事务排序、流控和 collective

测量对象

有效带宽、拥塞与尾延迟

现场 P33–42；S08

MoE 通信中的多播与归约



Hypercast

有选择地向多个目标分发数据

In-network computing

在通信路径中处理归约类操作

任务依赖

expert 分布、batch 与拥塞决定收益

外部 KV cache 的访问路径



现场组合

Leo CXL、Scorpio 与 Wistron 平台

分层原则

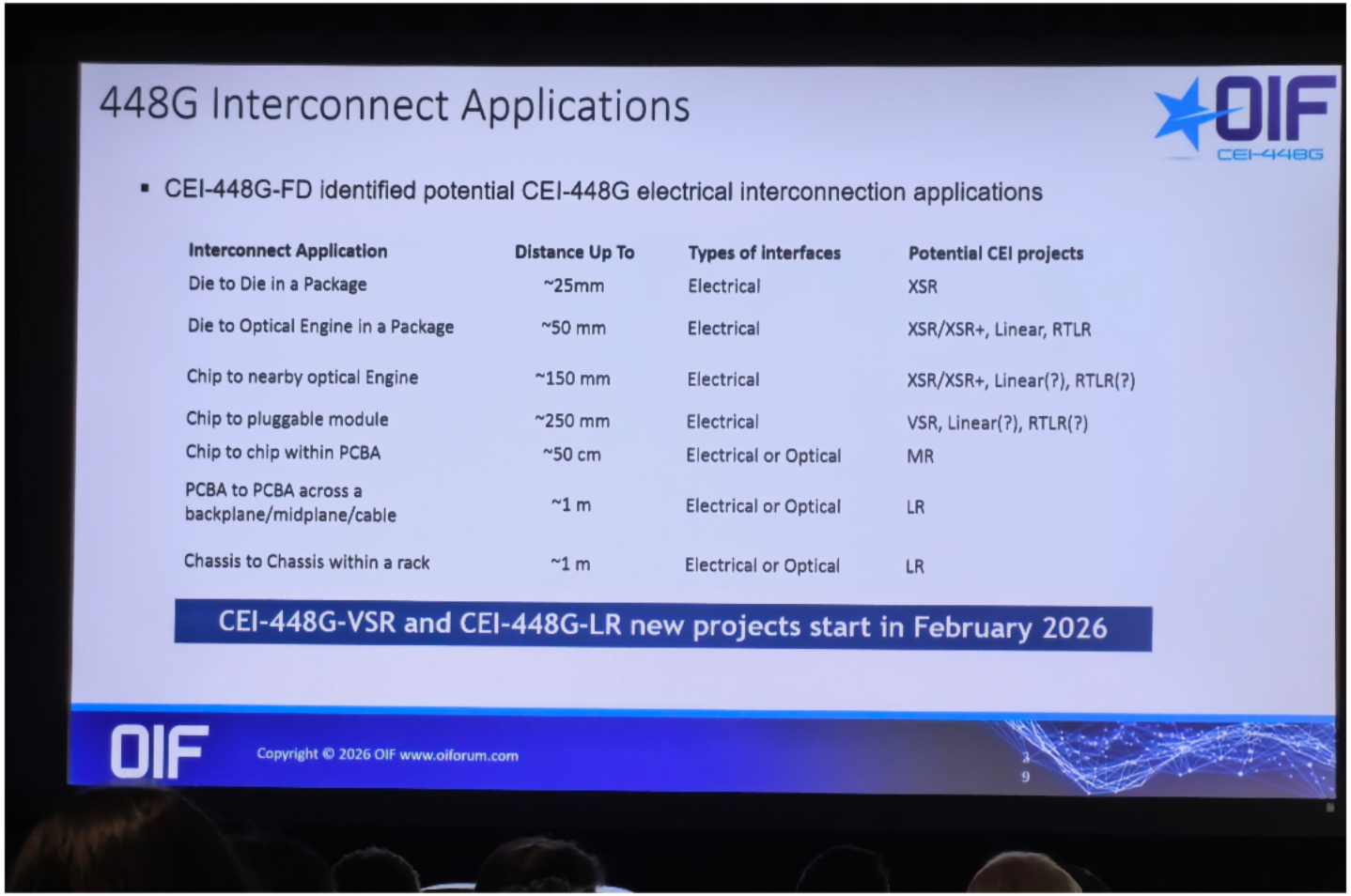
容量收益要连同访问频率和搬运代价计算

验证内容

冷热迁移、首 token
延迟与多请求竞争

现场 P41–42；S08

448G 的通道分类



讲者视角

TE 连接器、Cisco
系统通道、Broadcom 芯片

应用范围

XSR、VSR、MR、LR 与 Linear

共同起点

封装、板材、连接器和长度组成通道预算

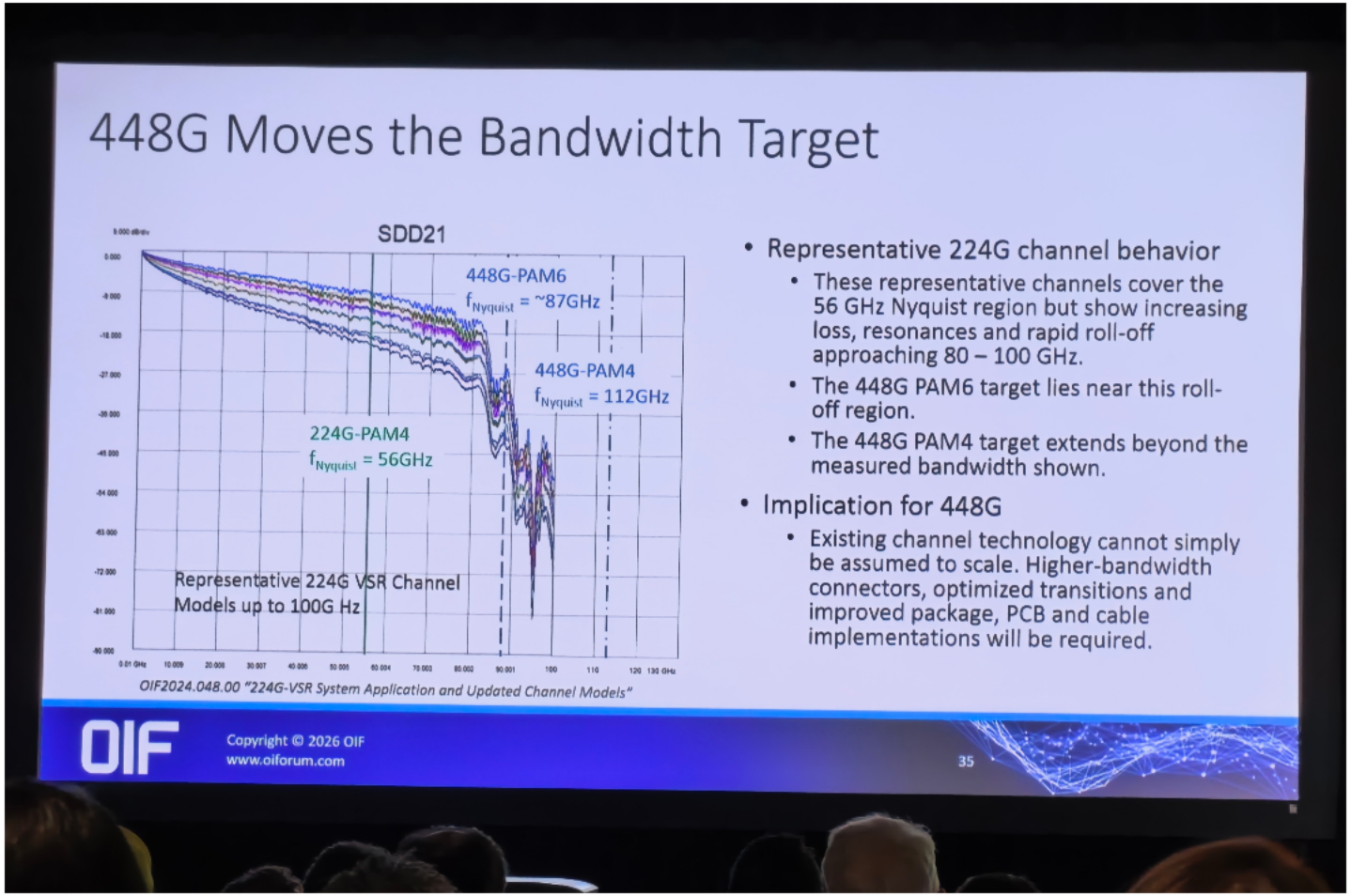
PAM4 与 PAM6 的带宽取舍

数据率与调制	理想 Nyquist	主要约束
224G PAM4	56 GHz	作为比较基准
448G PAM4	112 GHz	通道高频损耗
448G PAM6	约 86.7 GHz	更密电平所需的 SNR

带宽 PAM6 降低所需符号率

信号质量 更多电平缩小垂直裕量

封装与连接器消耗眼图裕量



物理细节

BGA、过孔、反射、skew 与模式转换

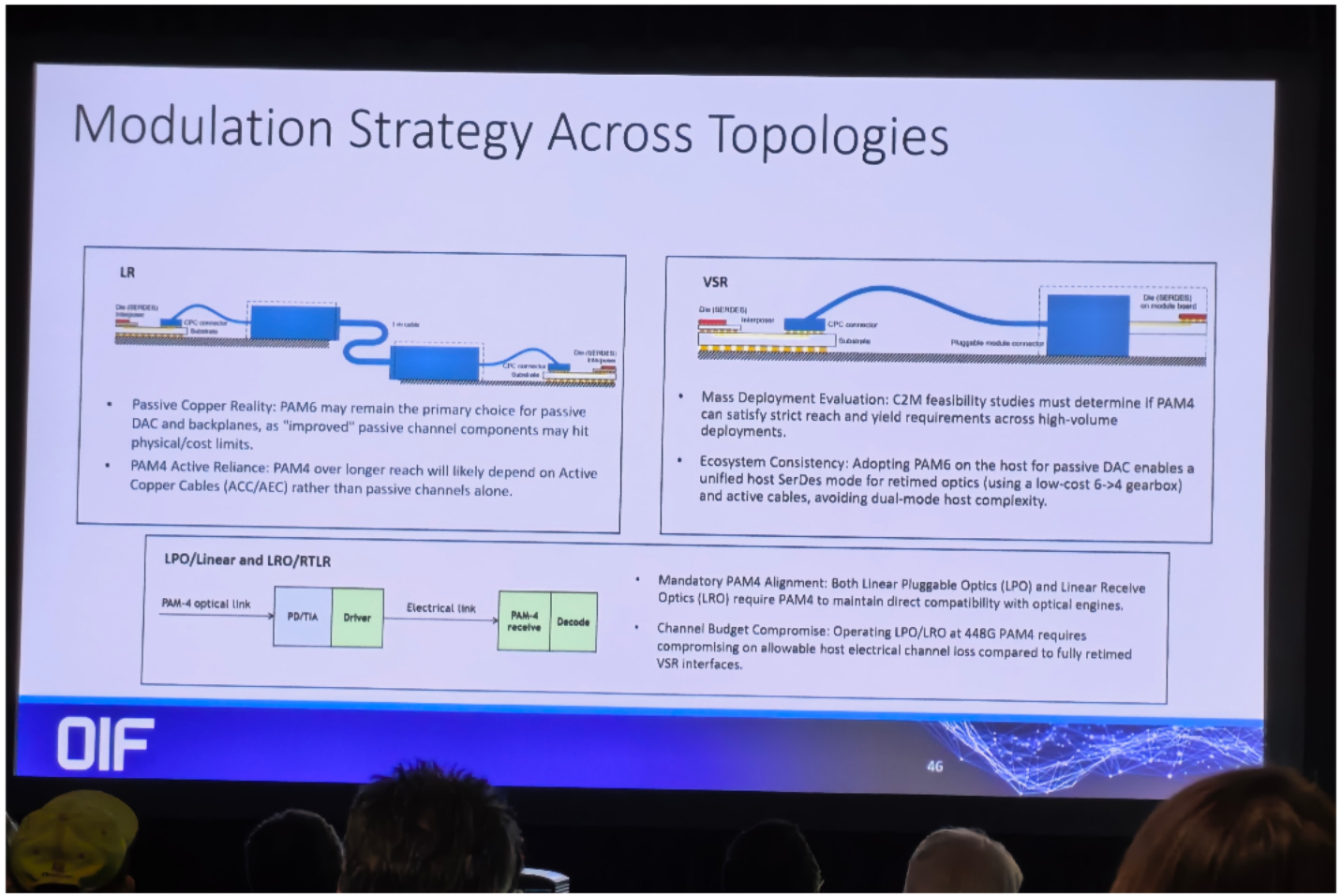
CPC

电缆靠近芯片，减少 PCB 路径

系统代价

机械装配、维修与测试方式随之改变

接收机复杂度与系统延迟



处理链

FFE、DFE、MLSE 与 inner FEC

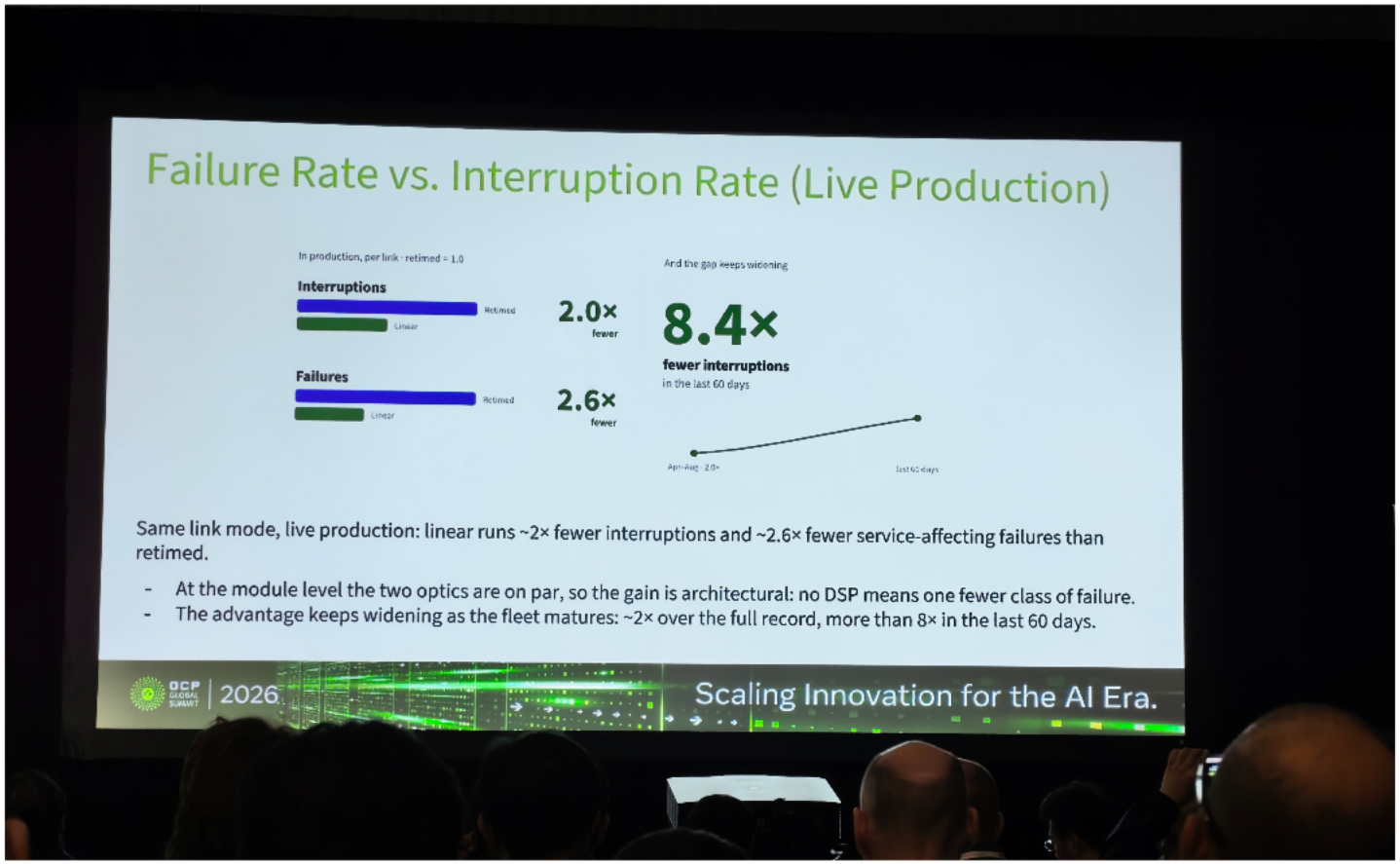
比较口径

有效吞吐、整跳功耗、突发误码和恢复

另一条路线

低速并行把压力转向通道数与组装

Meta 的 LPO 生产配置



规模

约 4,000 个模块运行 4 个月

拓扑

DTSW–EDSW 一跳两端 LPO

条件

2×400G FR4 以 1×800G 模式运行，FBOSS

Meta 上传照片与录音；生产配置页

LPO 的生产中断统计

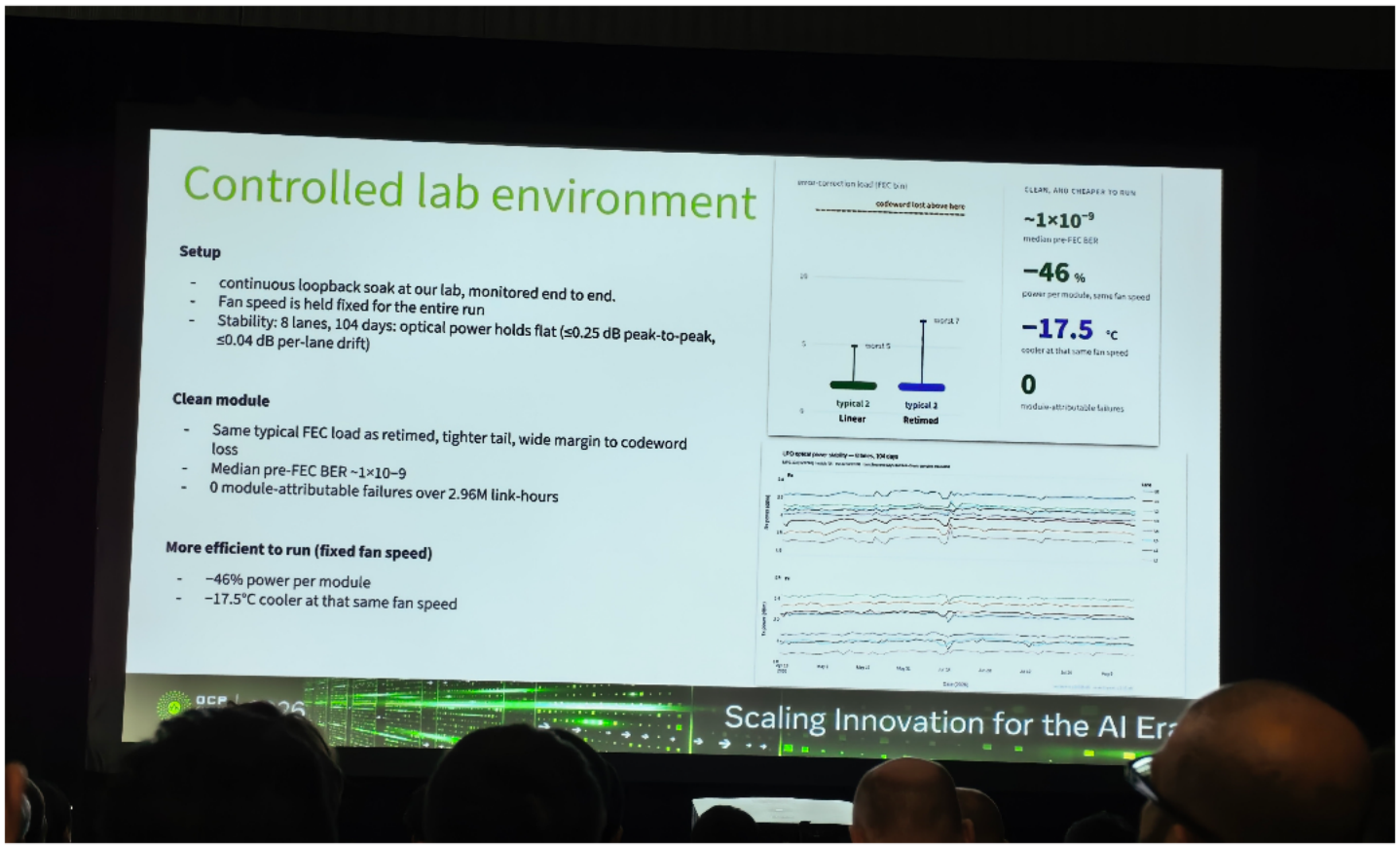
每链路指标	Retimed 基准	LPO 相对值
整个观察期中断	1	约 1/2
影响业务的失败	1	约 1/2.6
最后 60 天中断	1	约 1/8.4

比较单位 以 retimed 的每链路统计归一

工程意义 中断次数要连同恢复时间和业务损失记录

Meta 生产结果原始照片；相对值按现场标注

Meta 的长期实验室验证



长期测试

296 万 link-hours，观察到 0 次模块归因失败

功耗与温度

相同固定风扇下 −46% 与 −17.5 °C

测试代际

100G/lane，MiniPack3 光回环

Meta 实验室结果照片；上传演讲资料

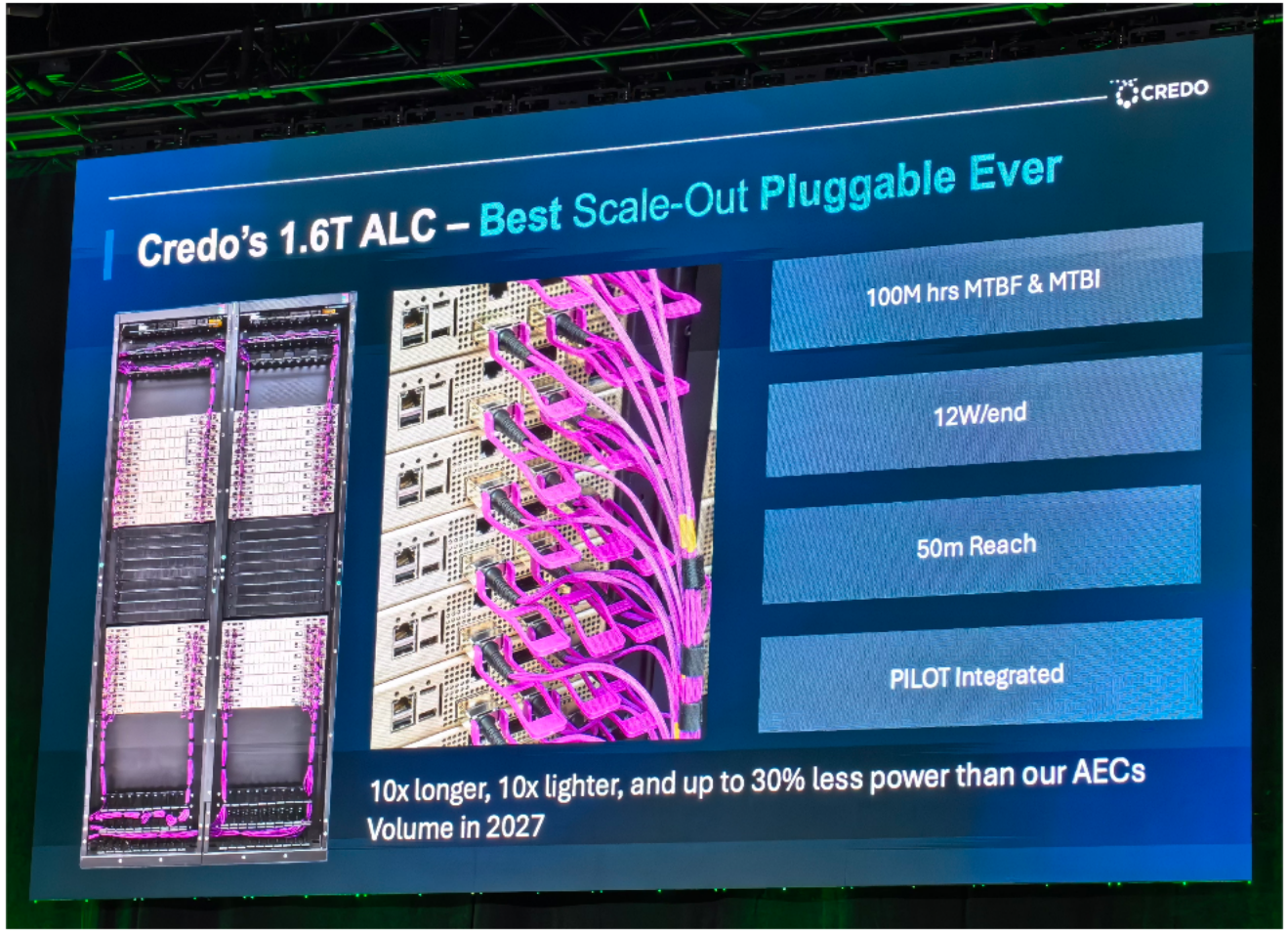
误码与光功率的长期漂移

观察项	现场数值
Pre-FEC BER 中位数	约 10^{-9}
FEC bin 典型值	两组均为 2
最差 FEC bin	Linear 5，Retimed 7
8 lane 104 天光功率	峰峰 ≤0.25 dB，漂移 ≤0.04 dB

分布优于平均 逐 lane BER 与 FEC bin 分布

趋势优于快照 温度、光功率与漂移同步记录

Credo 的产品与技术背景



既有能力

SerDes、gearbox、有源线缆与 PILOT 管理

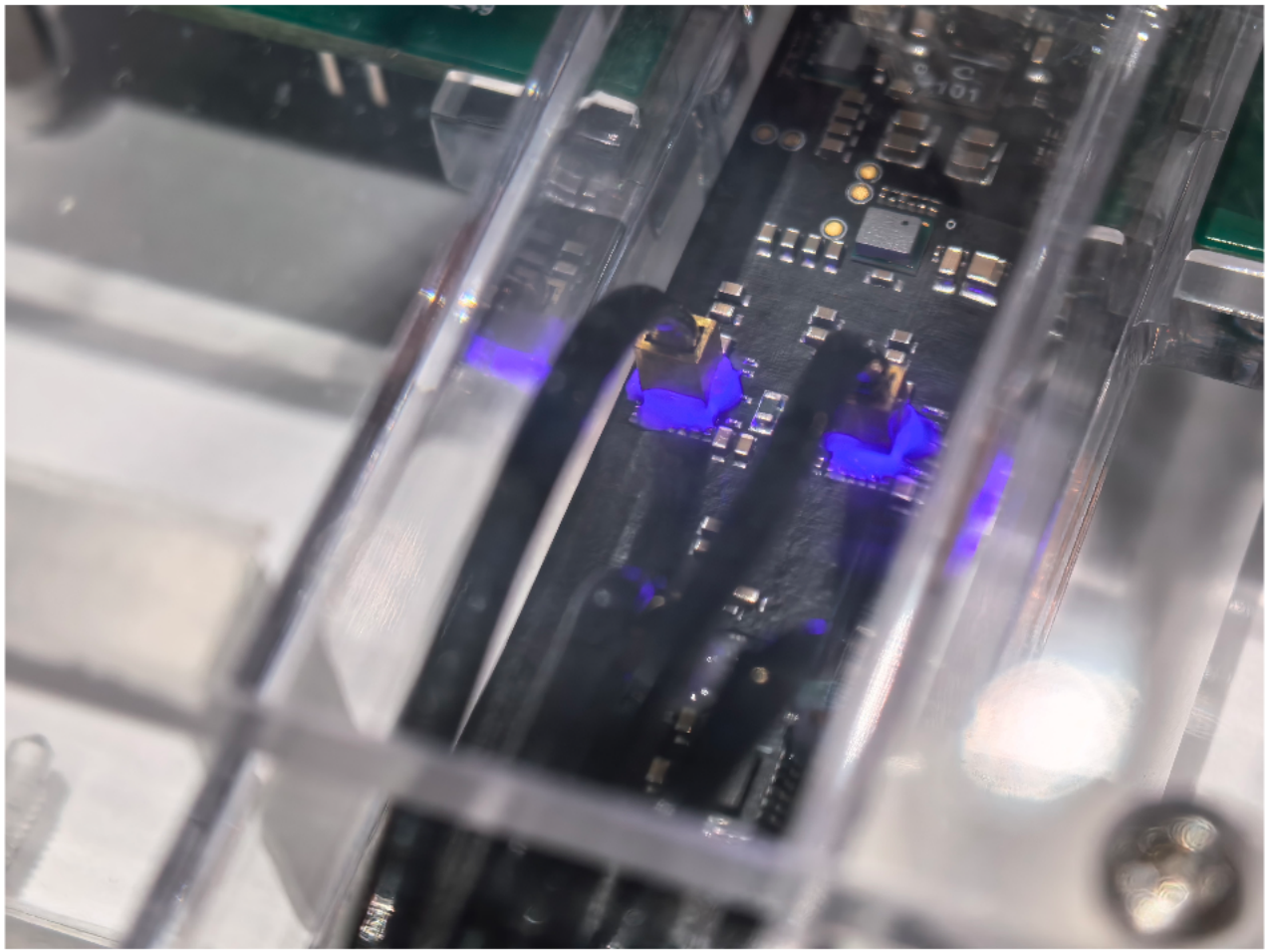
技术补充

2025 年收购 Hyperlume

ALC 目标

1.6T、50 m、12 W/end，2027 年交付计划

蓝色发光区域与光纤接口



现场近照

保留蓝紫色光、连接头、引出线和周边电路

器件线索

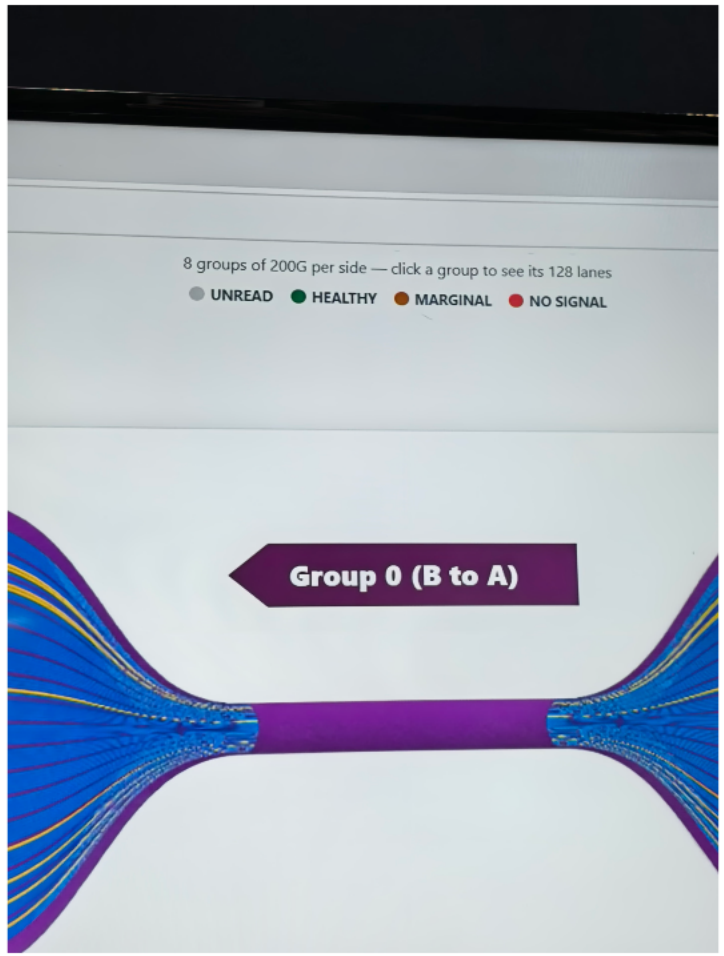
旁侧展牌标有 ams OSRAM Micro Emitters

工程观察

耦合、定位、应力释放与可装配性

P23–25；现场工程师交流

8 组 128 lane 的带宽组织



界面标注

每侧 8 组，每组 200G、128 lane

带宽复算

8×200G = 1.6 Tb/s 每方向

数量级

200/128 = 1.5625 Gb/s 每位置

P26、P31–32；现场口述与算术复核

Gearbox 与通道重映射



数据路径

Host、gearbox、crossbar 与 emitter array

管理路径

MCU、健康监测与备用通道

后续资料

工艺节点、模拟数字划分与时钟结构

P30；现场口述；S02

单通道能耗与整条线缆功耗

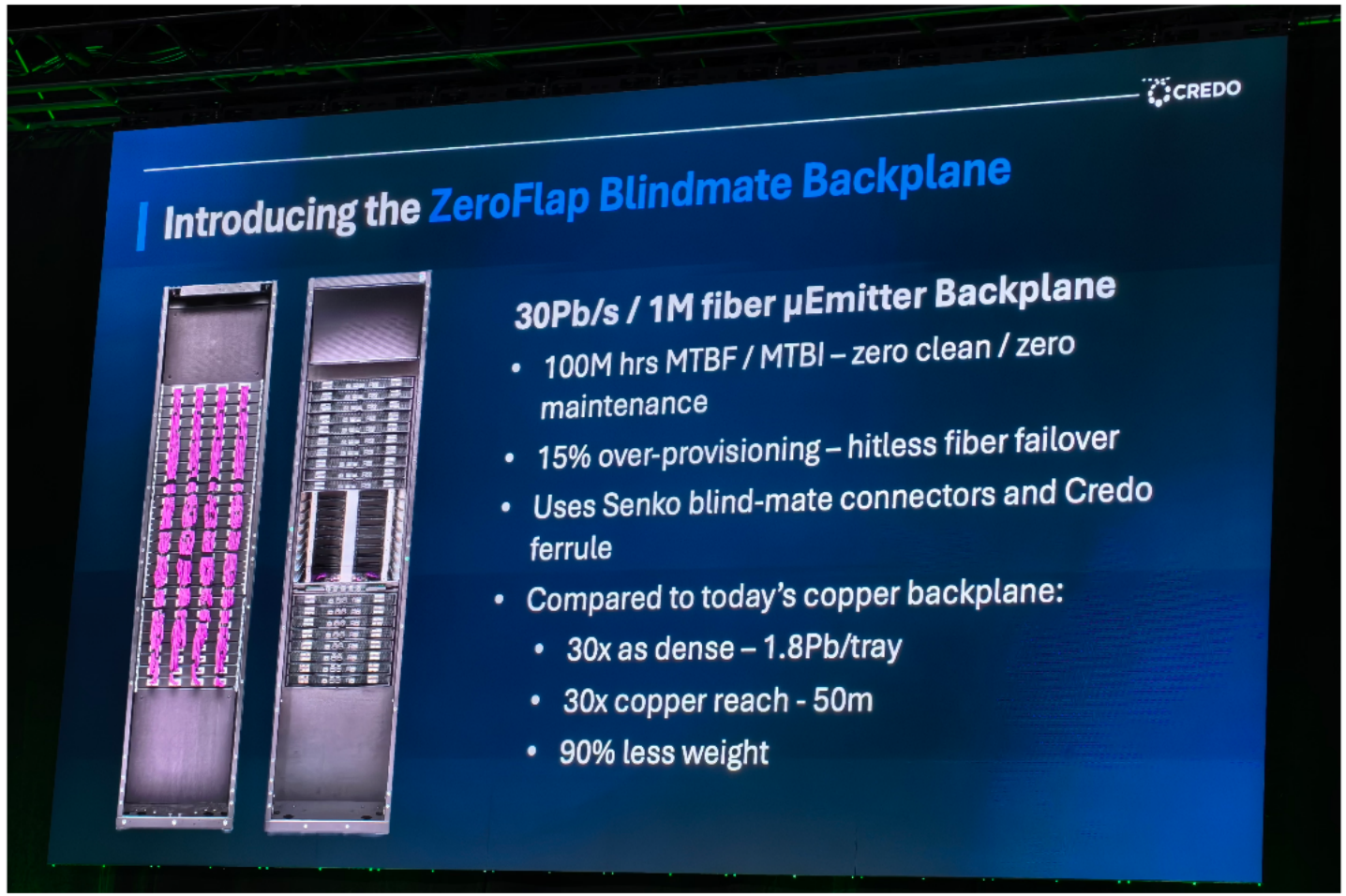
统计范围	算式	结果
单通道口述指标	1 pJ/bit × 1.6 Gb/s	1.6 mW
假定 1 / 2 / 3 V 供电	1.6 mW ÷ V	1.6 / 0.8 / 0.53 mA
两端 双向满载	24 W ÷ 3.2 Tb/s	7.5 pJ/bit
两端 仅计单向带宽	24 W ÷ 1.6 Tb/s	15 pJ/bit

现场口述 约 1 pJ/bit，统计边界尚需器件资料补充

产品目标 12 W/end，共 24 W

现场口述；P54；S01–S02；基于明确假设复算

连接器与背板决定扩展密度



现场架构展示

15% overprovisioning 与 blindmate

产业协作

Credo、ams OSRAM、SENKO 与纤路伙伴

SRMO

信号、功耗、模块和连接器接口，目标 2027 年规范

P22、P55；S03；C02

VCSEL 并行光连接



联合展示

Lumentum、Astera Labs 与 Corning

现场路线

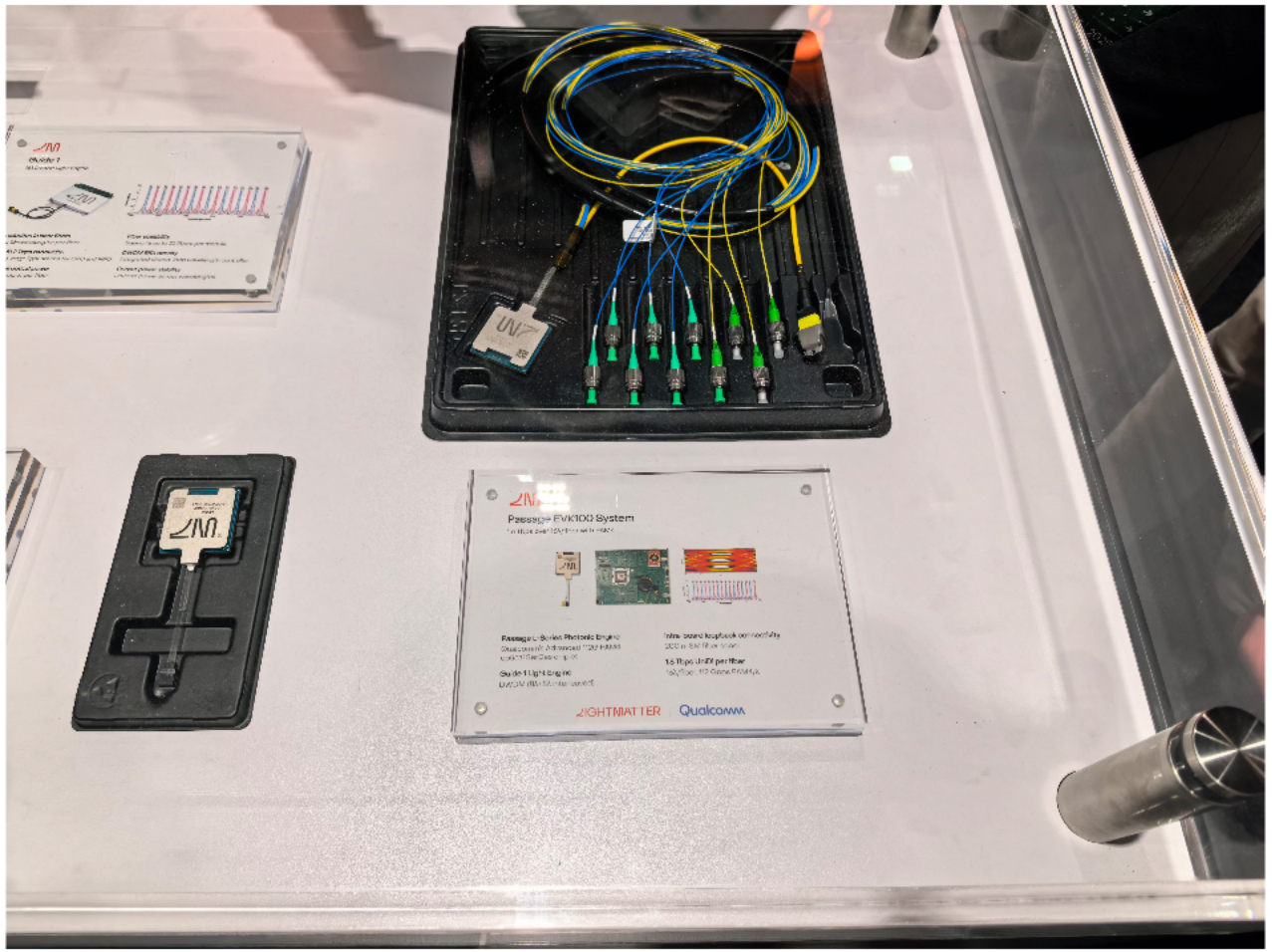
1060 nm VCSEL、32 Gb/s NRZ 与 PCIe 5

比较焦点

每通道带宽、阵列耦合与纤路数量

P35–40；S05

Lightmatter 用 WDM 提高每纤带宽



现场系统

Passage EVK100、Guide 光源与 Qualcomm

每根光纤

16 个波长，1.6 Tb/s 单方向

系统边界

公开目标能效包含光源

P11–13；S06–S07

模块制造与机架散热



InnoLight 与 TeraHop

光模块、光引擎、测试与交付

Ciena

XPO、CPO 与光连接的系统放置

nVent

机架供电、液冷与维护空间

P4–10；S16–S20

推理系统的内存层级

01 **热数据**

紧邻计算的高带宽内存

02 **容量扩展**

CXL 内存与外部 KV cache

03 **读密集数据**

HBF 与 SSD 后备层的不同位置

P1–3、P41–53；S11–S15

XCENA 的计算内存



MX1

CXL 3.2、PCIe 6 与 DDR5 容量扩展

计算下沉

RISC-V、向量和压缩等数据处理

软件条件

SDK、数据布局与冷热命中率

P1–3；S11–S12

SK hynix 与 Marvell 的 CXL 方案



CMM-Ax

512 GB、4 通道 DDR5 与近内存 Arm 计算

CMM-DDR5

扩容控制器与模组

访问路径

本地 DDR
带宽与主机接口带宽分别预算

P47–49、P51–53；S13–S14

FAISS 演示的能效复算

方案	QPS	功耗	QPS/W
64 核服务器 本地 DRAM	42,500	900 W	47.22
4×Structera A	66,400	440 W	150.91
比值	1.56×	0.489×	3.20×

相同质量 两组 recall 均为 95%

复算结果 QPS 约 1.56 倍，QPS/W 约 3.20 倍

P53；按现场表格复算

HBF 面向读密集容量层



现场形态

8Hi 与 16Hi NAND 堆叠

目标参数

最高 512 GB、0.4–3.0 TB/s 档位、UCIe

软件影响

访问粒度、写入与耐久性进入数据放置策略

P45；S15

把现场观察变成可测的实验

01 **系统实验**

固定任务与质量，测有效吞吐和尾延迟

02 **链路实验**

逐 lane 误码、温漂、故障注入和重映射

03 **合作输入**

接口资料、样品、SDK 与测试条件

本文工程分析；三组 session 与展台资料