

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会 **IEIT**
SYSTEMS

2024开放计算中国峰会

互联技术探索

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

王海梦 赵伟康

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

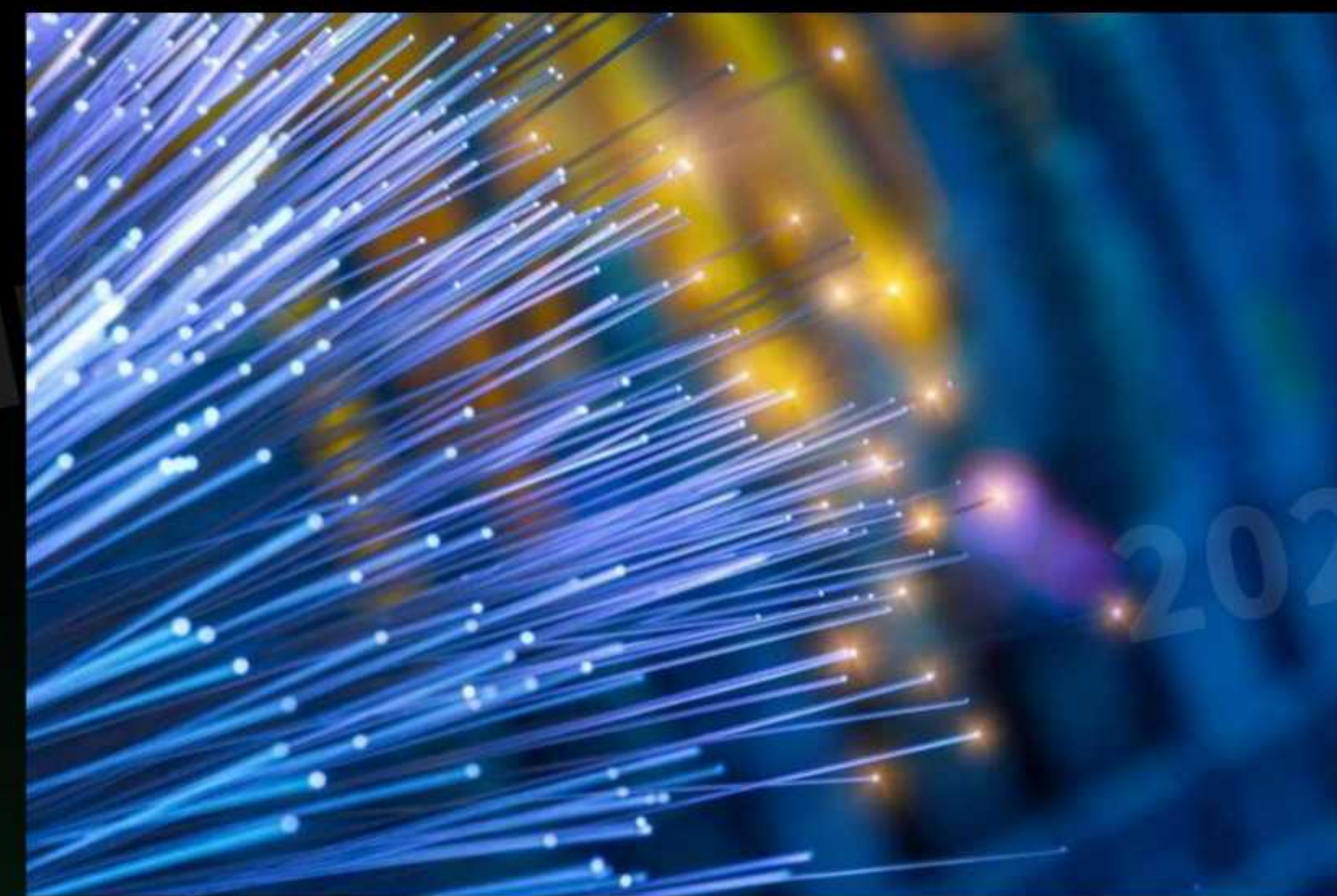
2024开放计算中国峰会

2024开放计算中国峰会 **IEIT**
SYSTEMS

CXL内存系统



PCIe 光互联



2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会 **IEIT**
SYSTEMS

2024开放计算中国峰会

互联技术探索：CXL内存系统

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

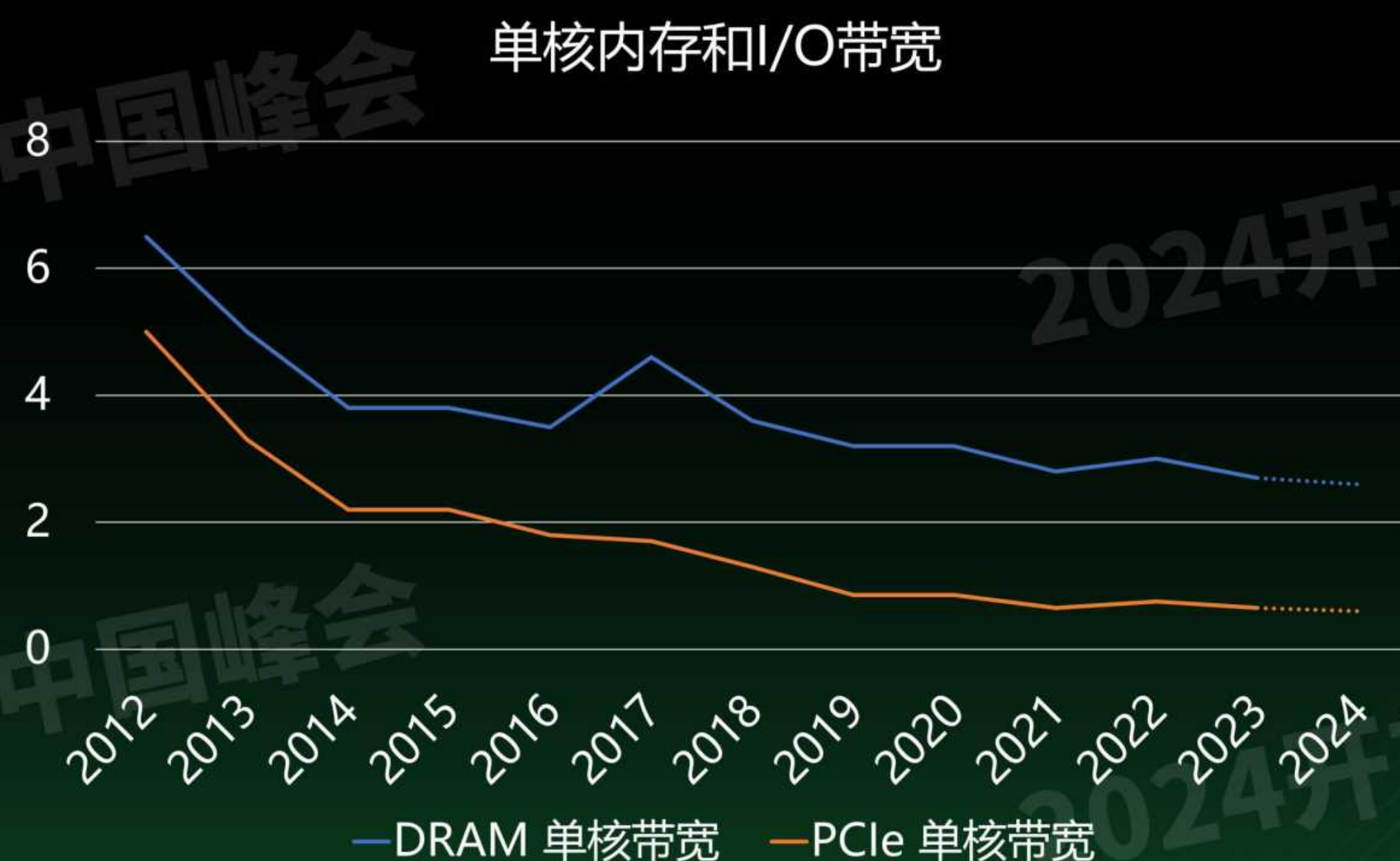
2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

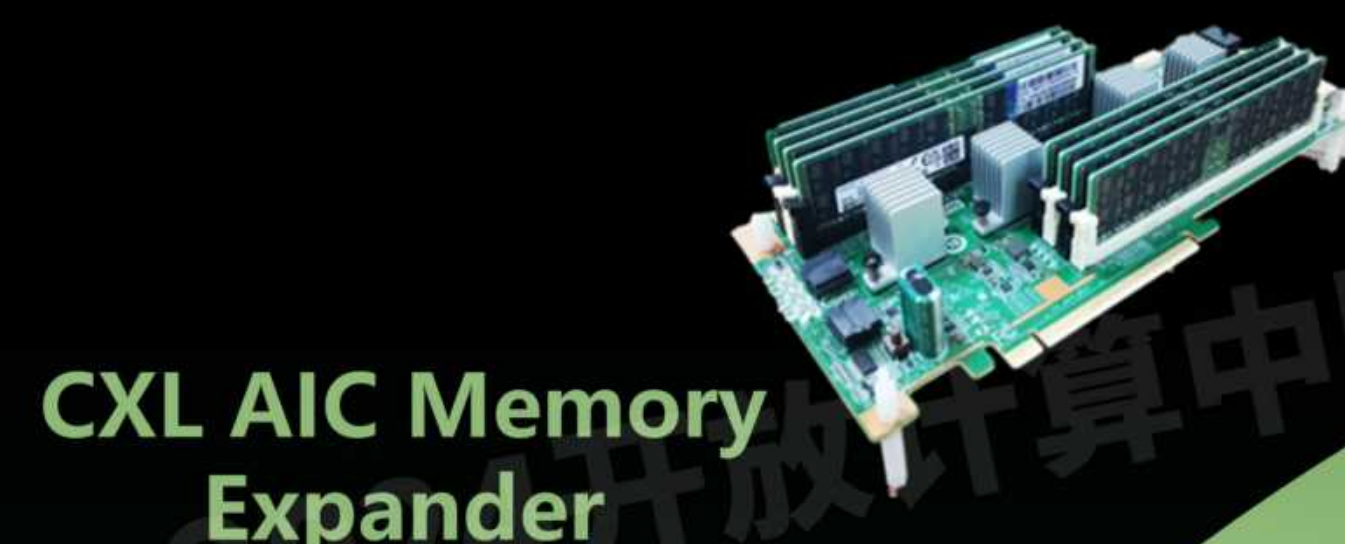
内存瓶颈、I/O瓶颈阻碍算力提升

- AIGC的到来，带来内存需求的大幅增加，也产生梯度数据聚合与分发等海量I/O通信需求；
- 内存和I/O的扩展能力远落后于计算密度的增长，平均到每个核心的内存和I/O带宽持续下降；
- 系统内存和I/O瓶颈需要新的架构来缓解，实现数据处理规模、并行处理能力和系统算力提升。



IEIT CXL 内存解决方案路标

2022



CXL Memory Expander CXL1.1



2023

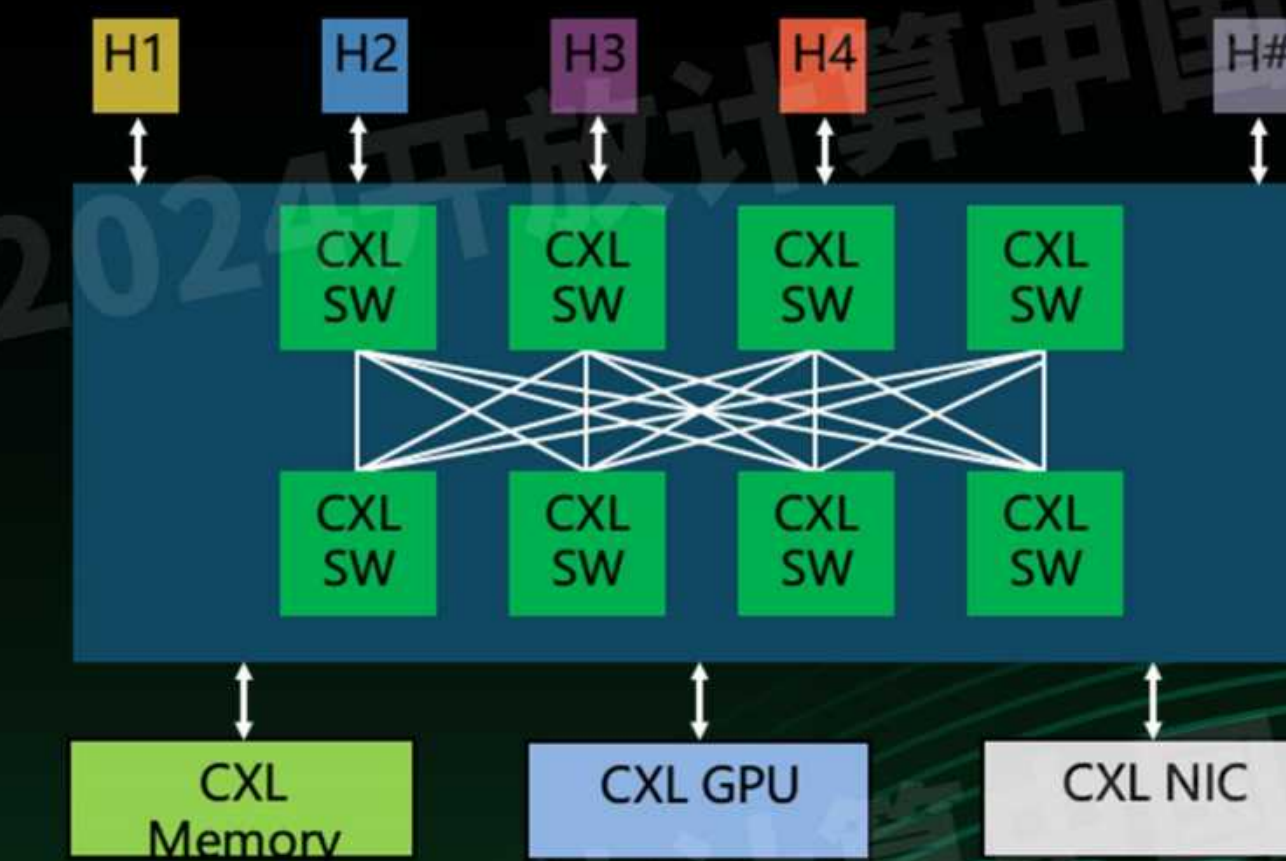


CXL Memory Pooling CXL2.0



2024

High-Performance CXL Switching Network CXL3.x



High-performance CXL switching network topology

内存远端扩展技术

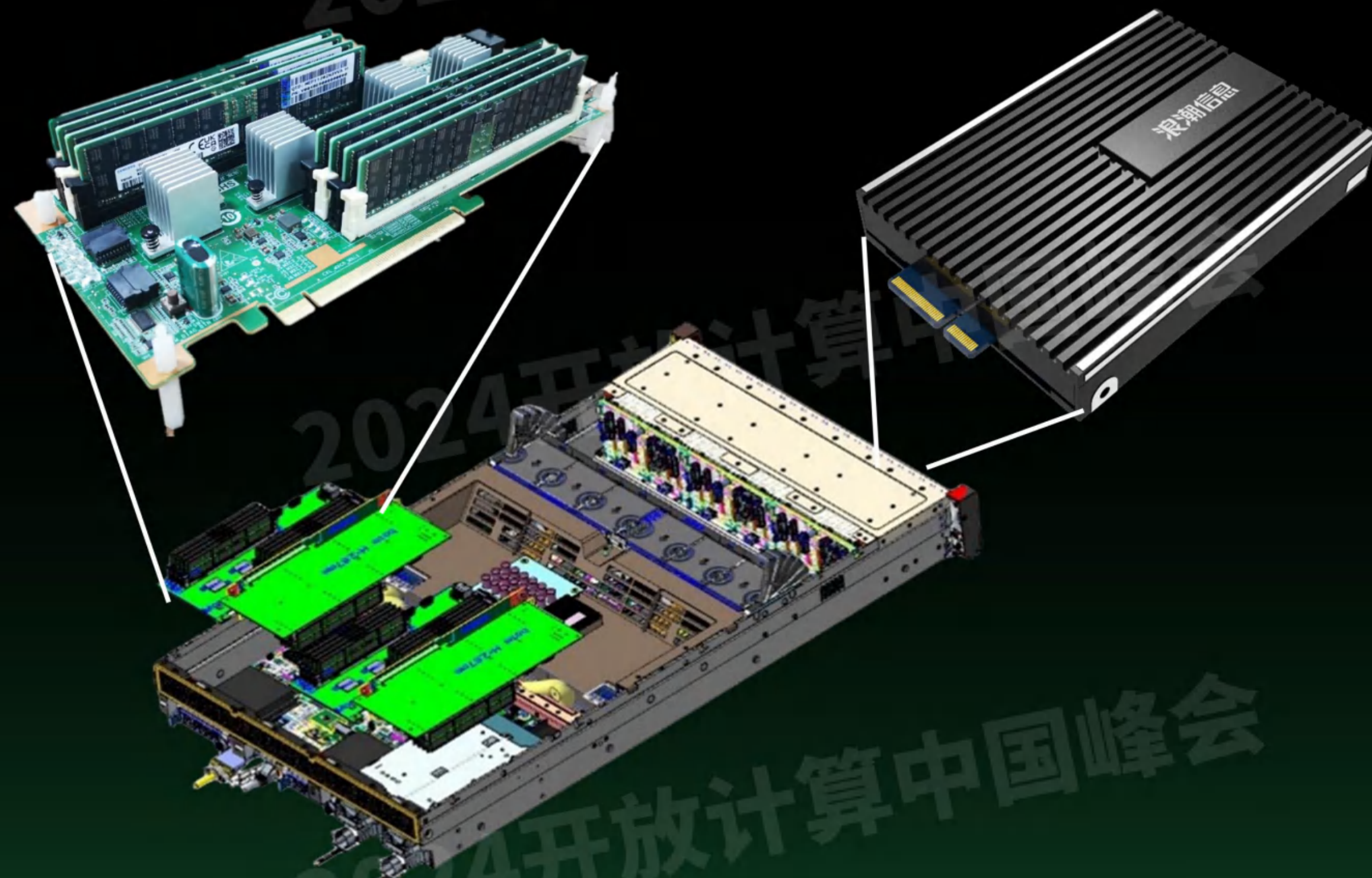
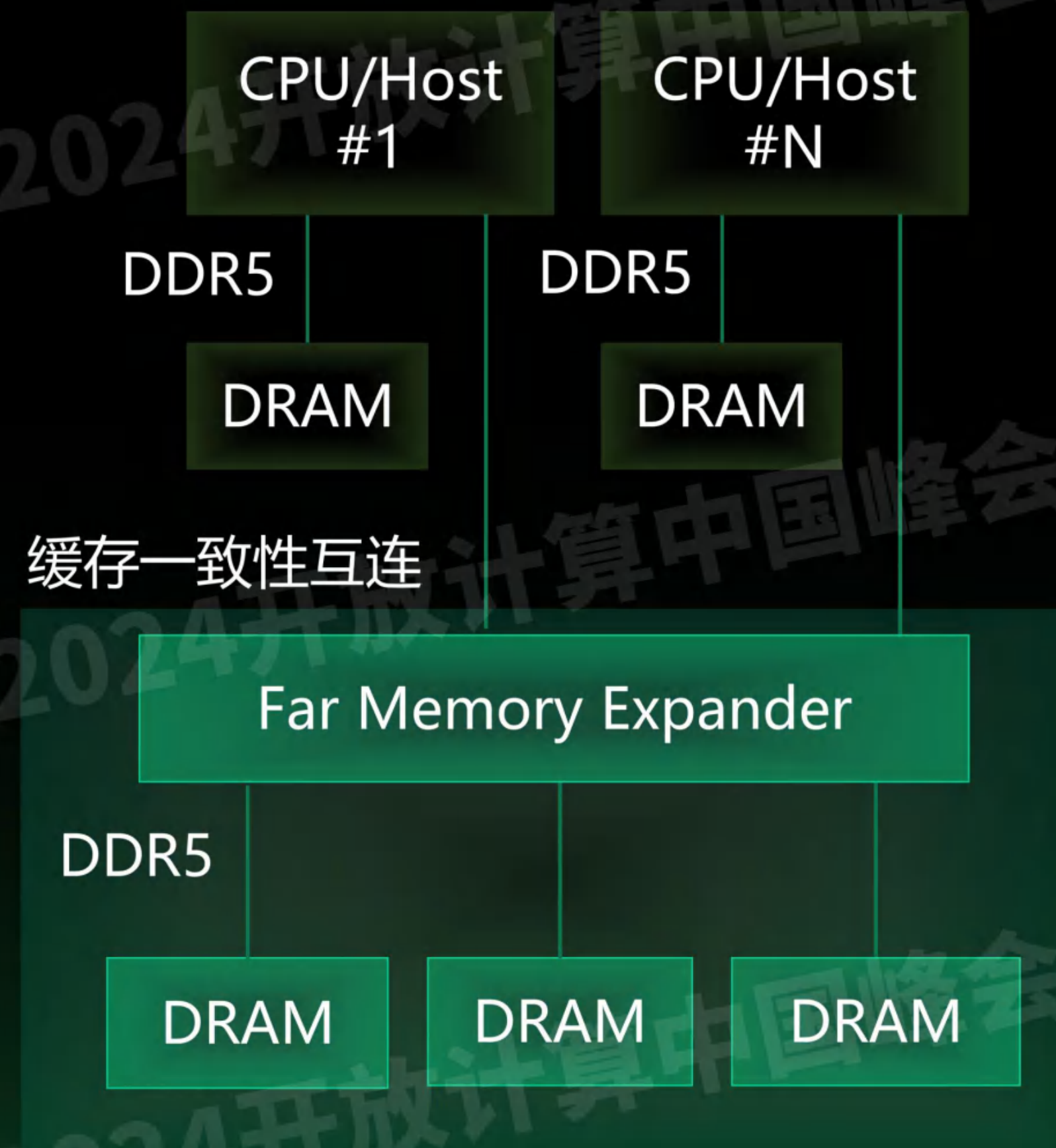
内存远端扩展技术，实现在百纳秒级访问延时的条件下，系统内存容量和带宽扩大一倍

CXL内存扩展卡

- 全高3/4长双宽 PCIE CEM卡；
- 支持8 DDR5-5600 RDIMM，单卡最高扩展容量2TB；

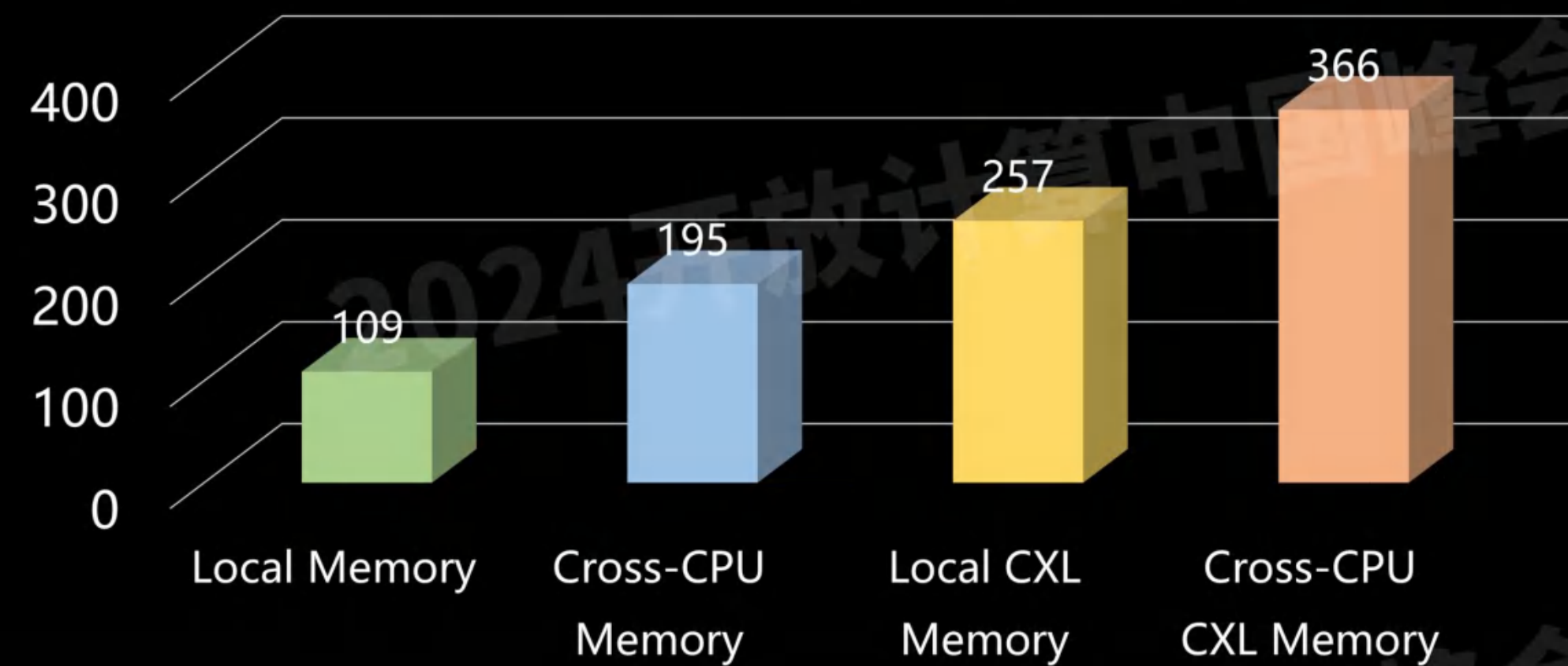
E3.S CXL内存模组

- E3.S 2T形态；
- 单模组容量支持64G/256G；

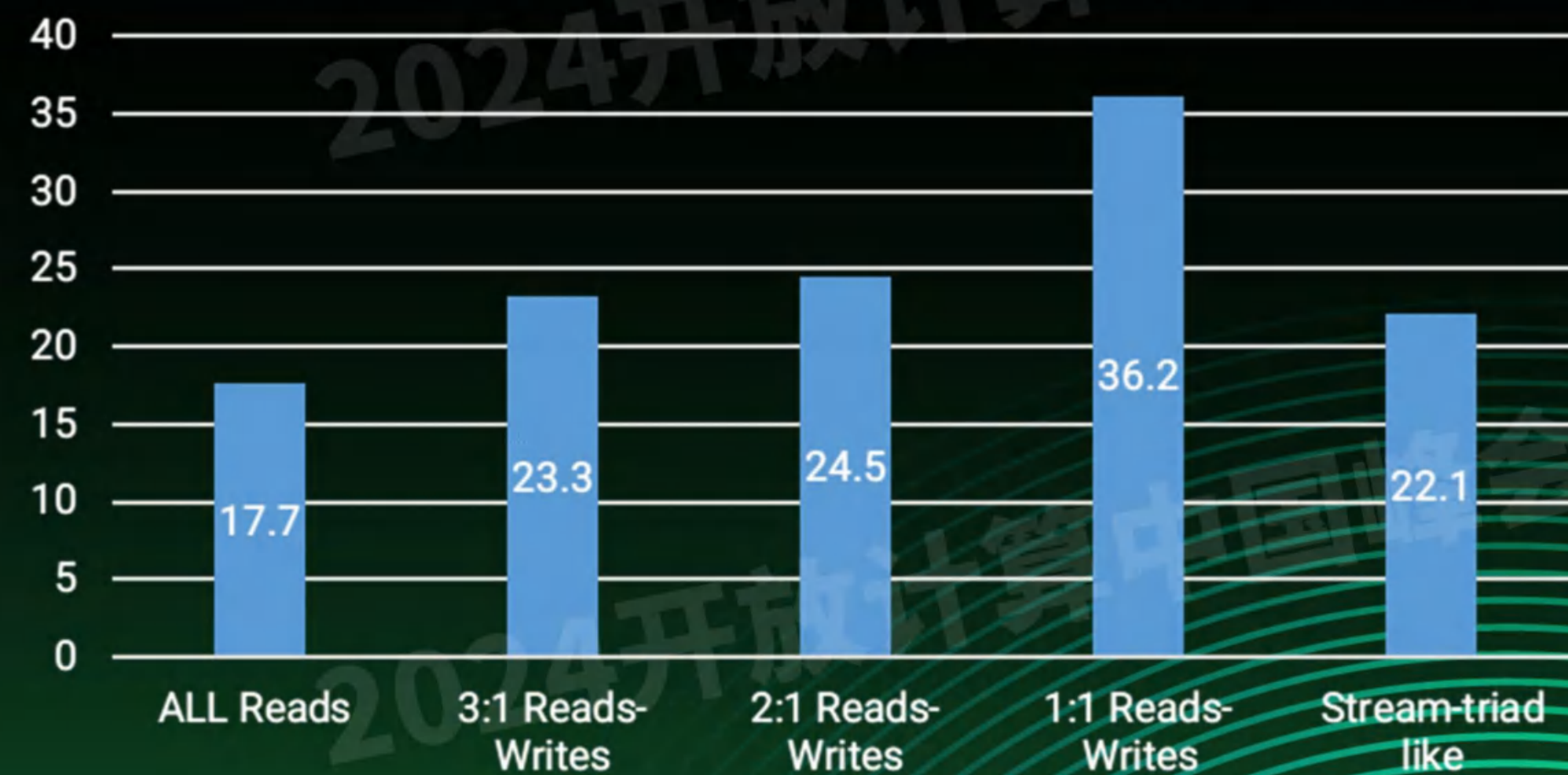


内存远端扩展原型系统

Memory latency performance comparison (ns)

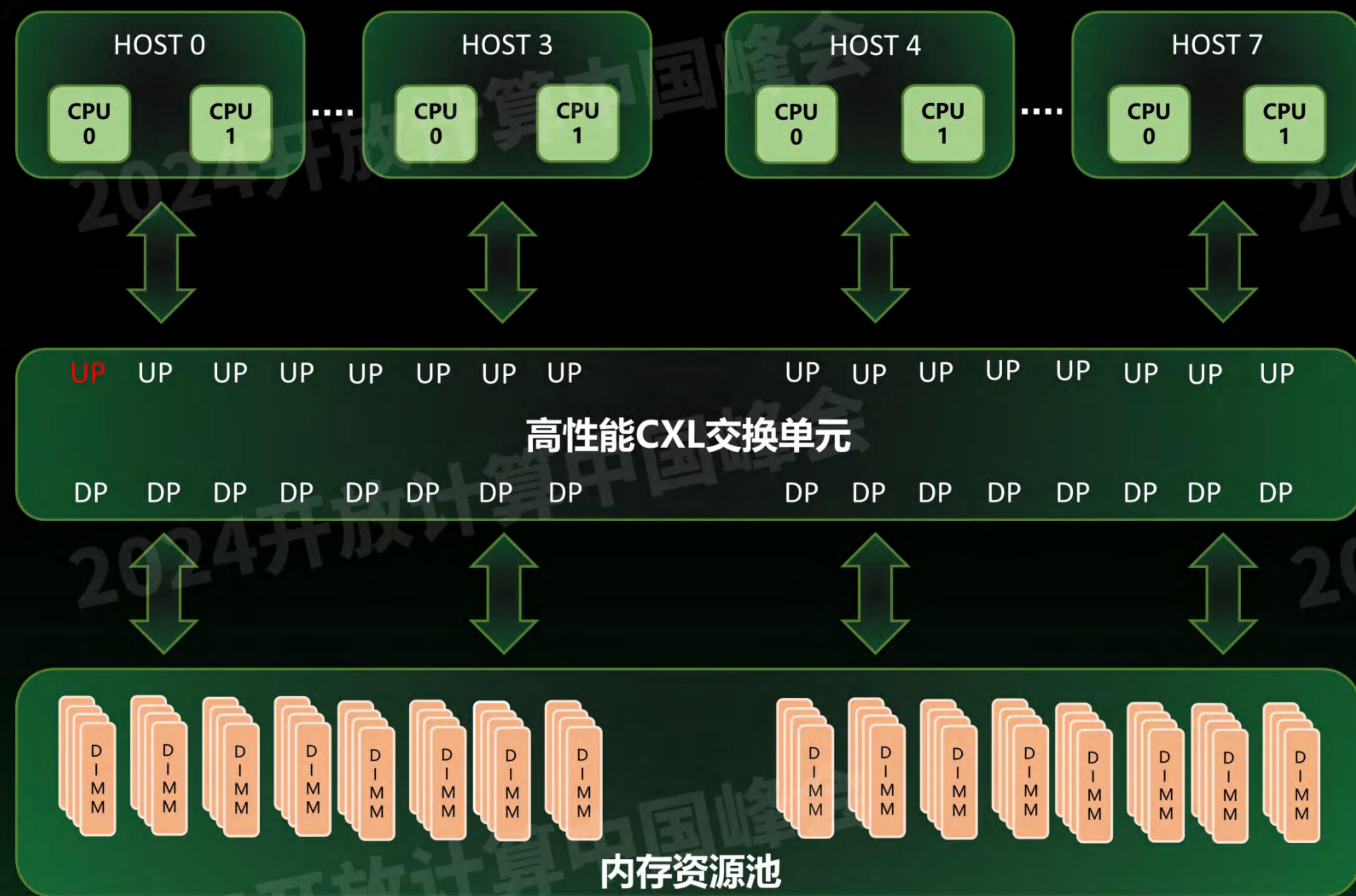


CXL Memory Bandwidth Test Results (GB/s)



内存池化技术

开发内存池化原型系统，实现多主机内存资源池化与共享，最大支持8HOST主机共享访问内存资源池，支持百纳秒级远端池化内存访问



CXL交换单元



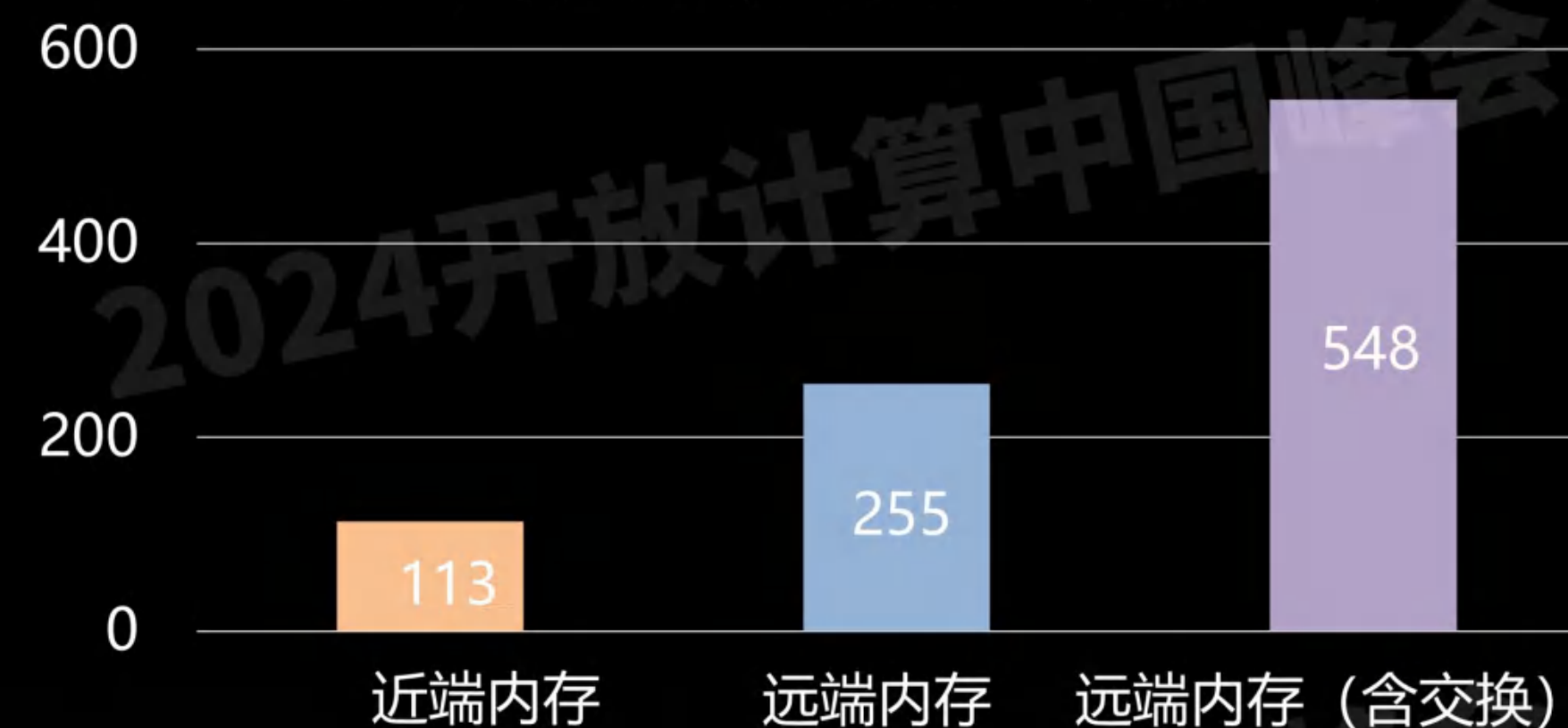
- 19英寸 2U机箱
- 支持32*CDFP x16 port, 支持CXL 2.0
- mCPU用于带内管理, RunBMC用于带外管理

内存资源池

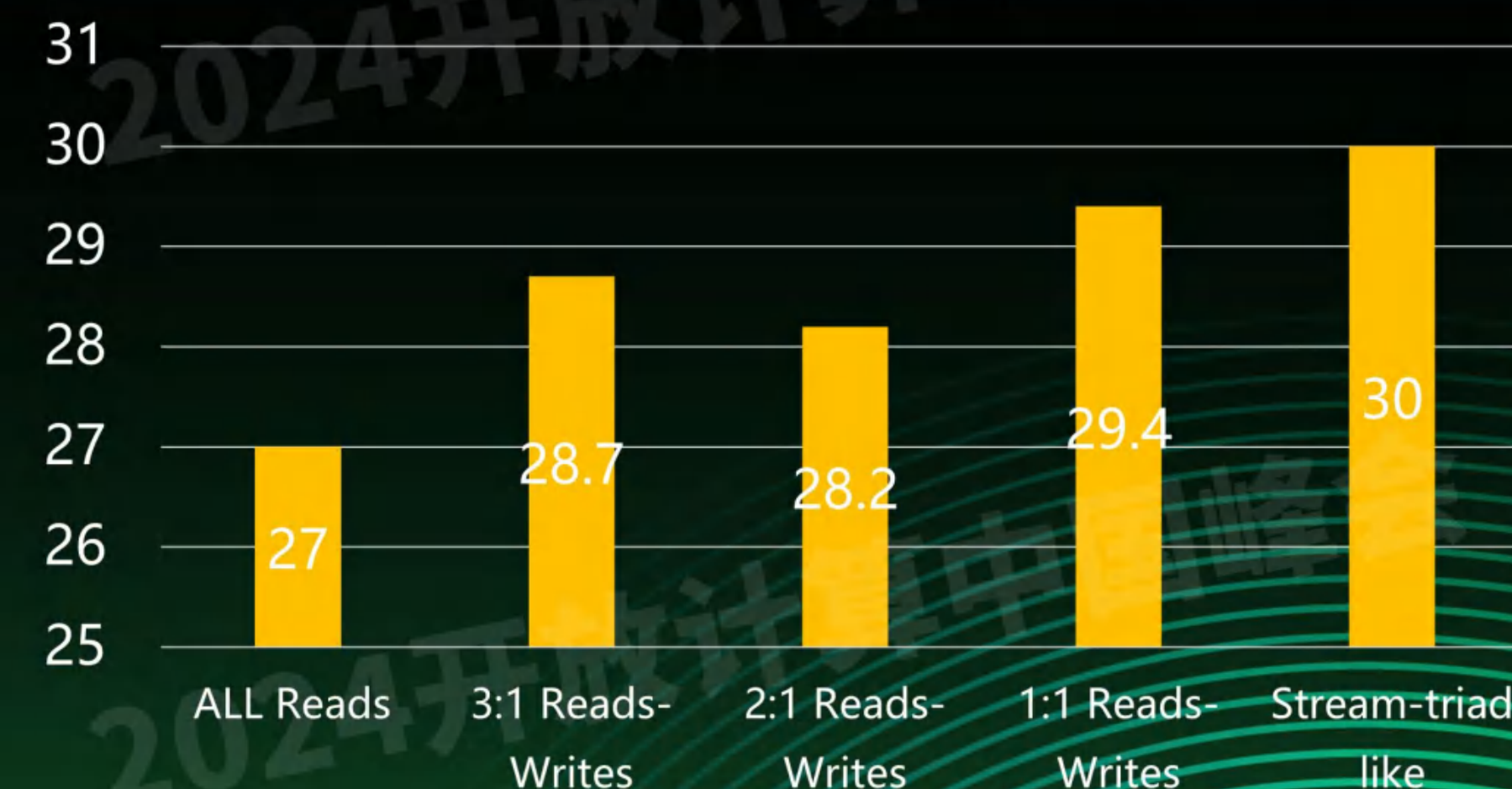


- 19英寸 2U机箱
- 支持 64*DDR5 RDIMM (5600MT/s), 单机最高支持16TB内存扩展;
- 支持16*CDFP x16 port, 支持 CXL 2.0

内存访问延时性能对比 (单位: ns)

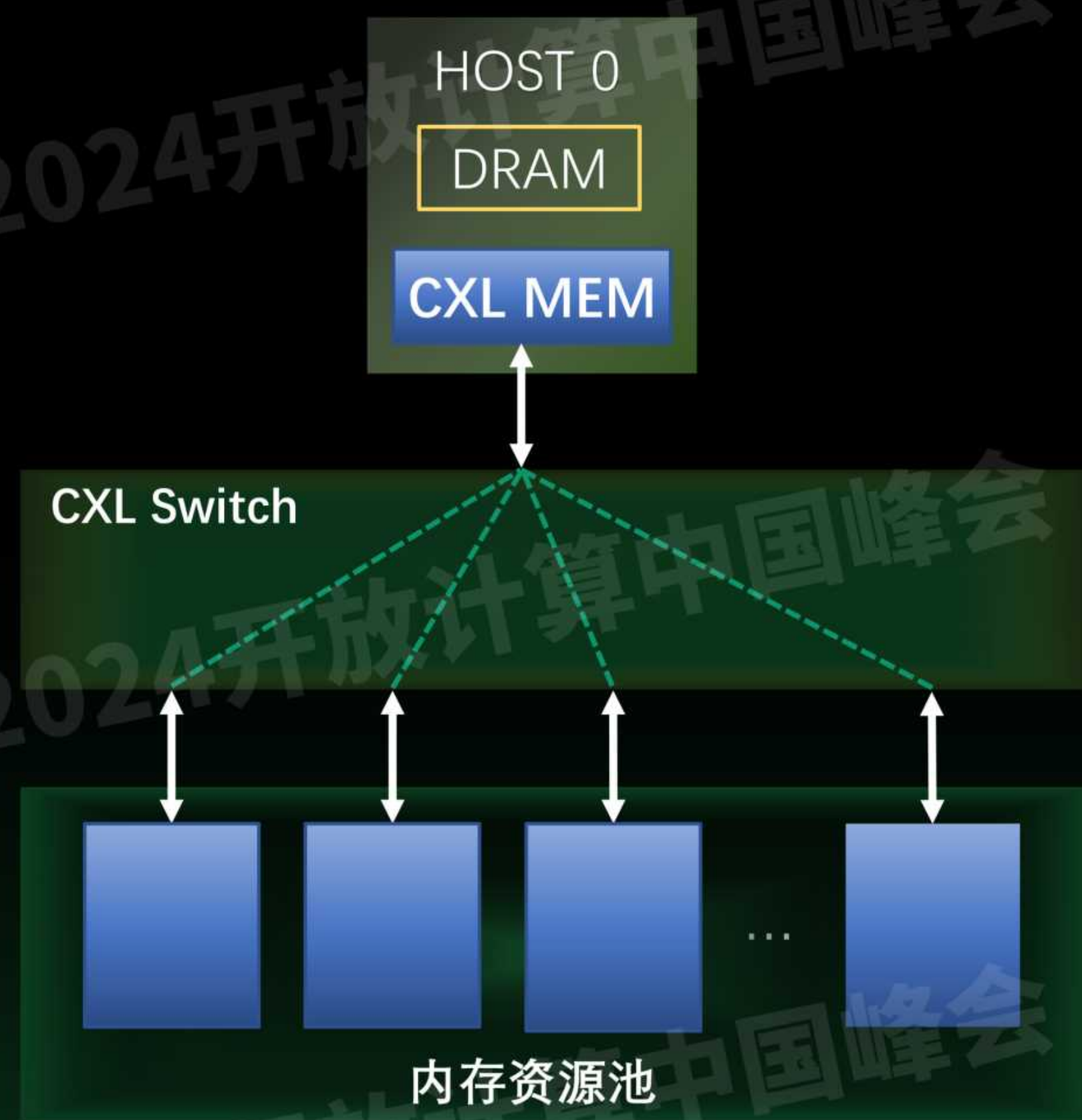


池化内存远端访问带宽性能 (单位GB/s)

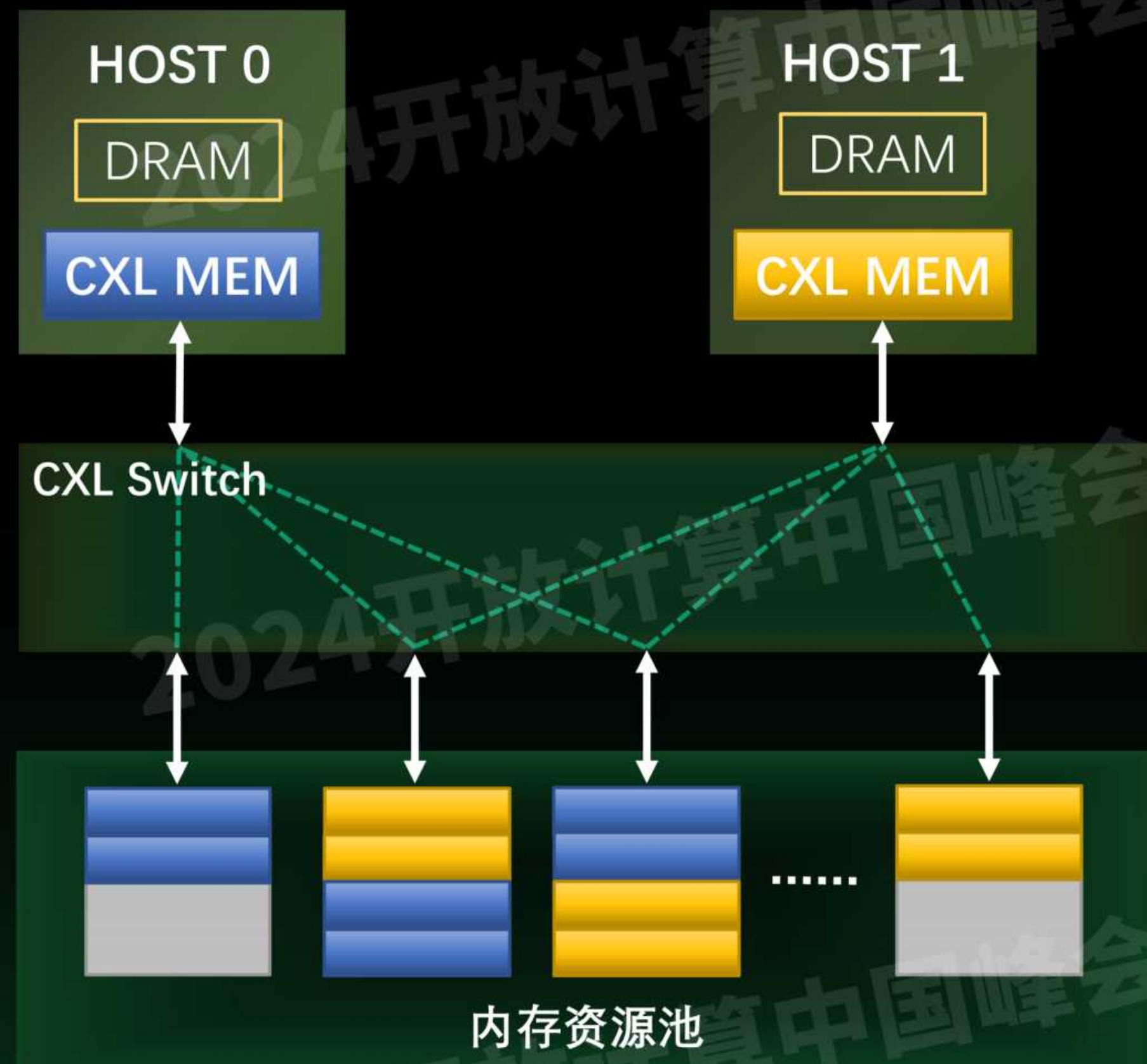


内存池化系统功能

开发业界首款CXL交换机，支持CXL2.0&CXL1.1单元，单节点最大支持32路32GT/s x16对外互连端口，聚合带宽达到4TB/s，支持任意上下行配置，支持direct mapping/pooling/sharing功能

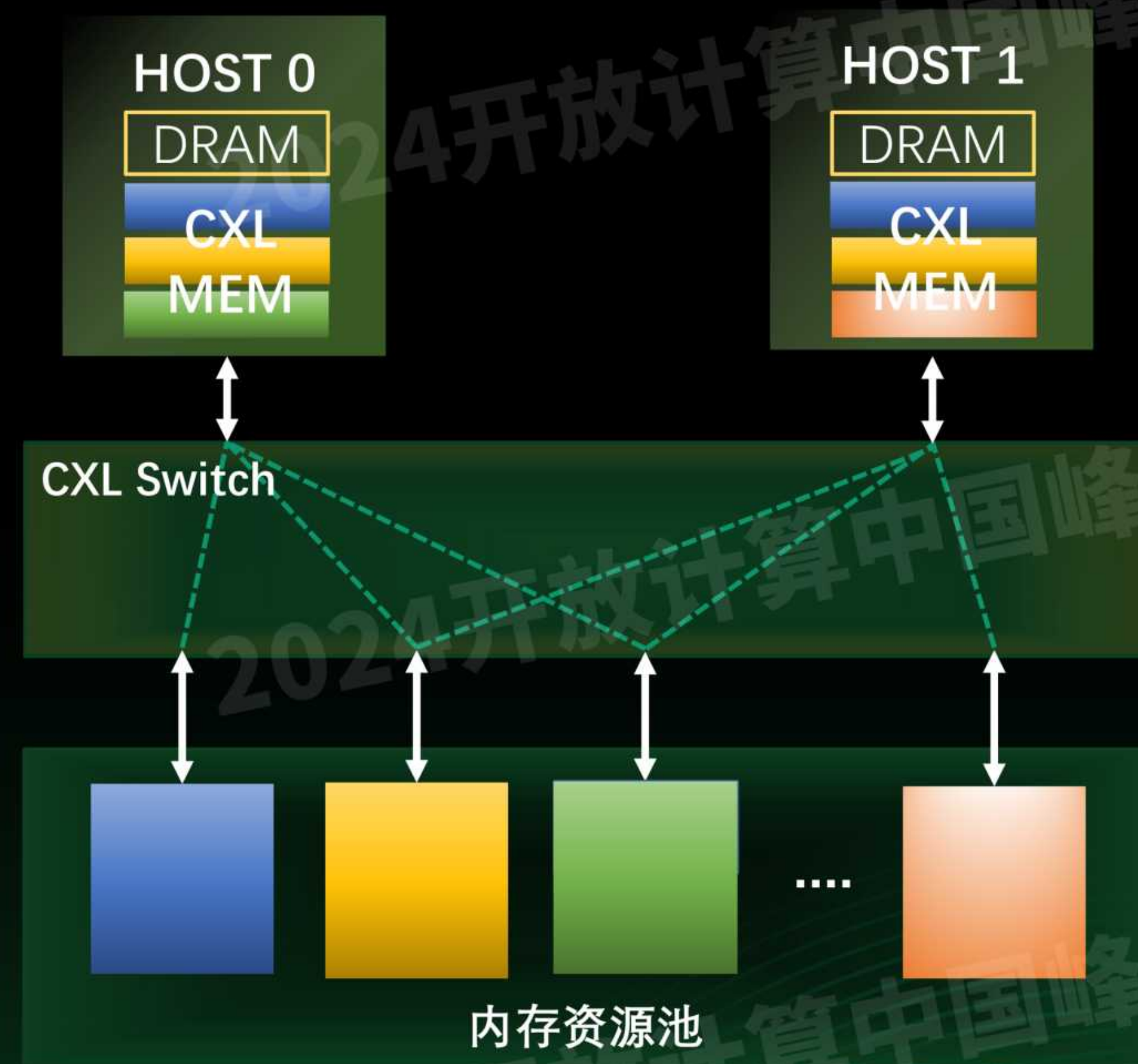


Direct Mapping Mode



Pooling Mode

支持16 分片内存 (64GB/片)



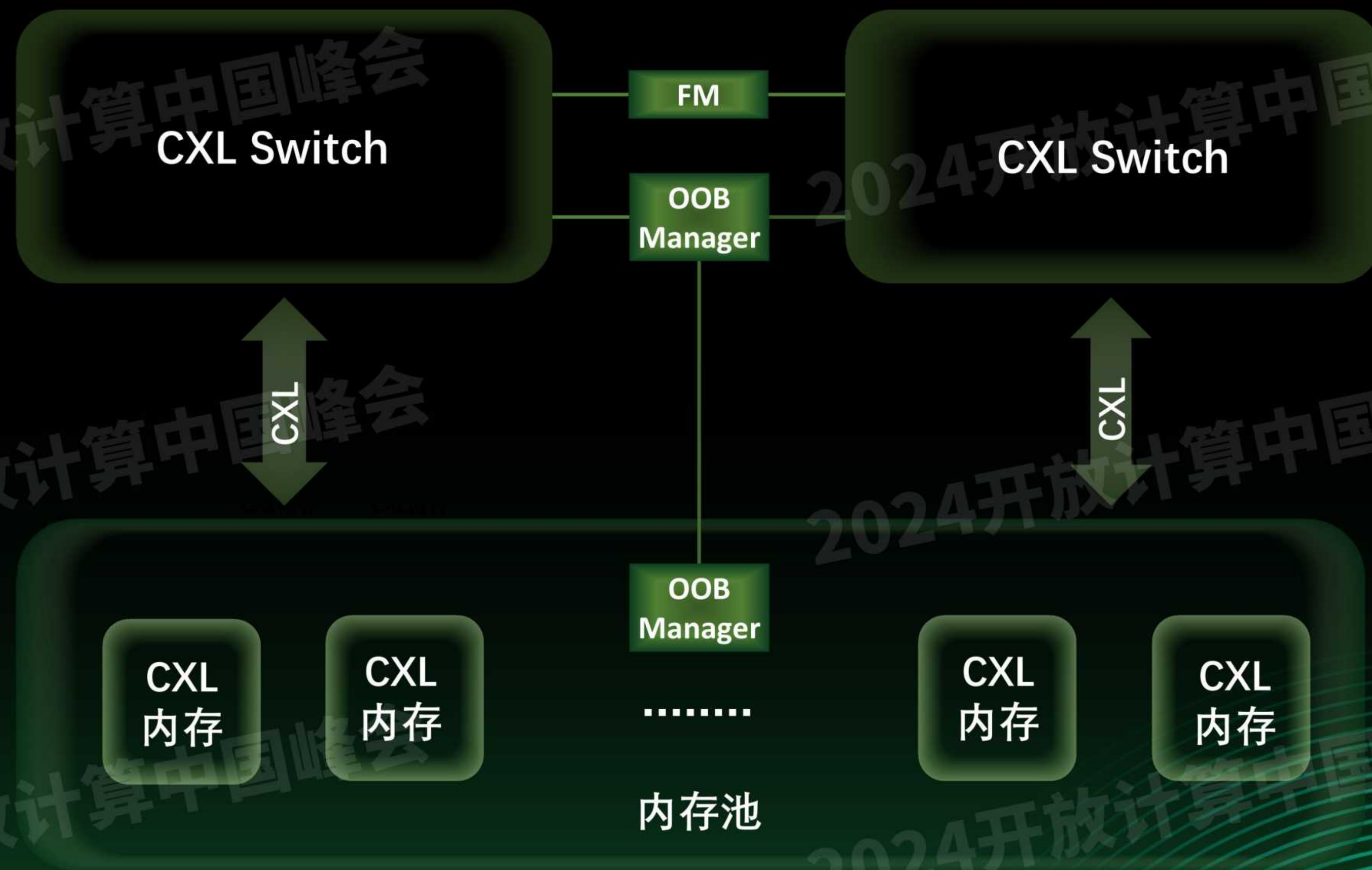
Sharing Mode

8主机共享内存读写一致

系统管理架构

CXL 内存池管理系统包括如下两个部分：

- Fabric Management(FM)：负责内存资源管理
- OOB Manager：负责带外监控以及健康监控



资源管理软件架构

FM作为管理系统的主体包含基础配置，端口配置，内存资源配置以及系统监控四大模块；并对用户提供CLI以及API接口

应用层

CLI

API

驱动层

基础配置

寄存器读写

系统管理

电源管理

错误监控

端口管理

端口类型配置

端口速率配置

端口带宽配置

端口reset

内存资源管理

Directmap配置

Poolmap配置

资源配置表管理

拓扑管理

Linux Kernel & Drivers

物理层

PCIe

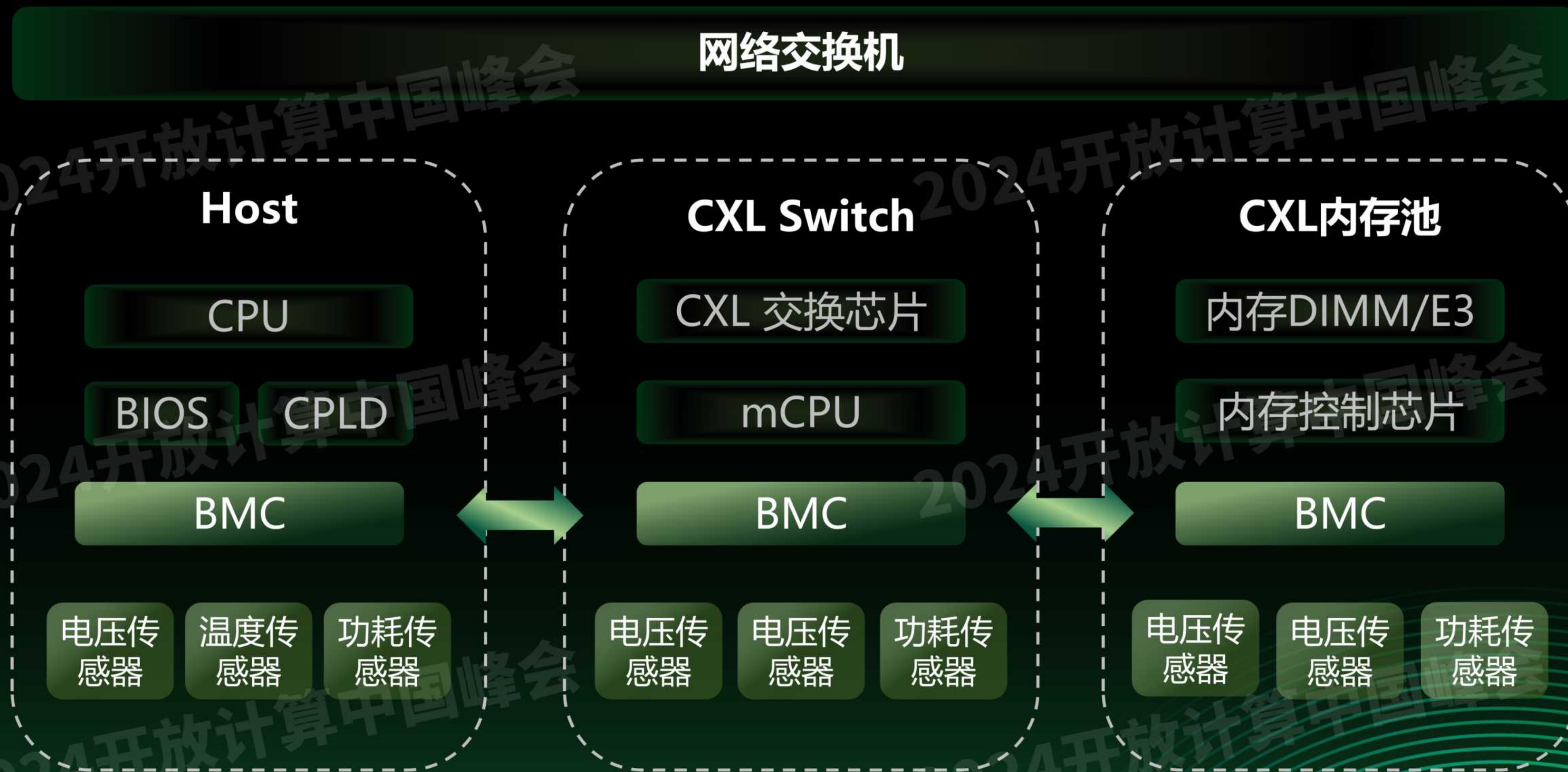
I2C/UART

带外管理

带外管理系统提供四大功能：

- **监控**：监控整个系统的状态
- **错误管理**：错误检测、错误报告、故障诊断和错误报告模块
- **电源管理**：系统协调电源控制和监控
- **其他基础管理**：如固件管理、用户管理和服务管理

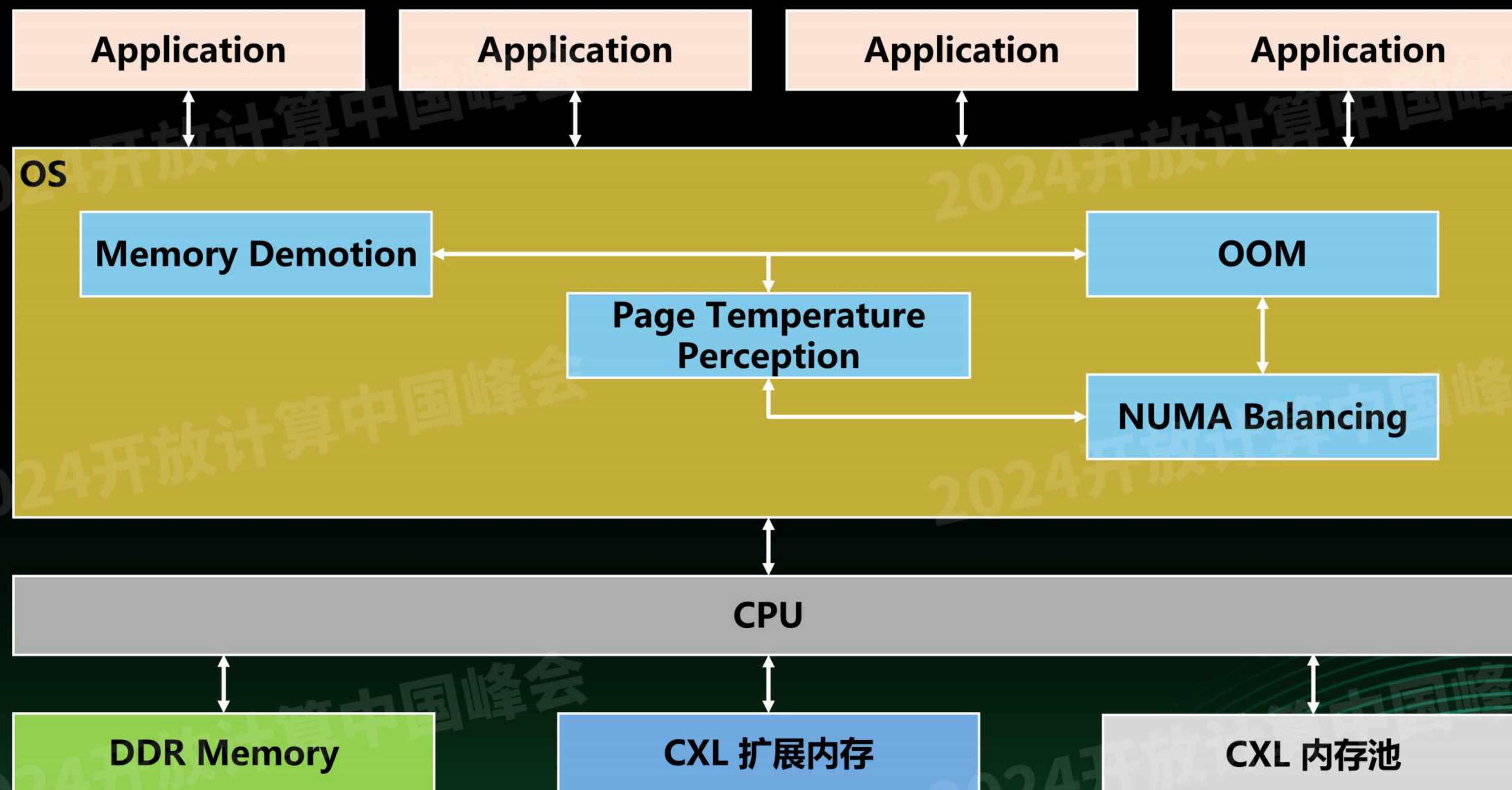
系统通过Redfish, GUI和IPMI提供接口供用户调用



内存分层管理

Linux Kernel开源社区提供的内存分层管理包含以下四个模块

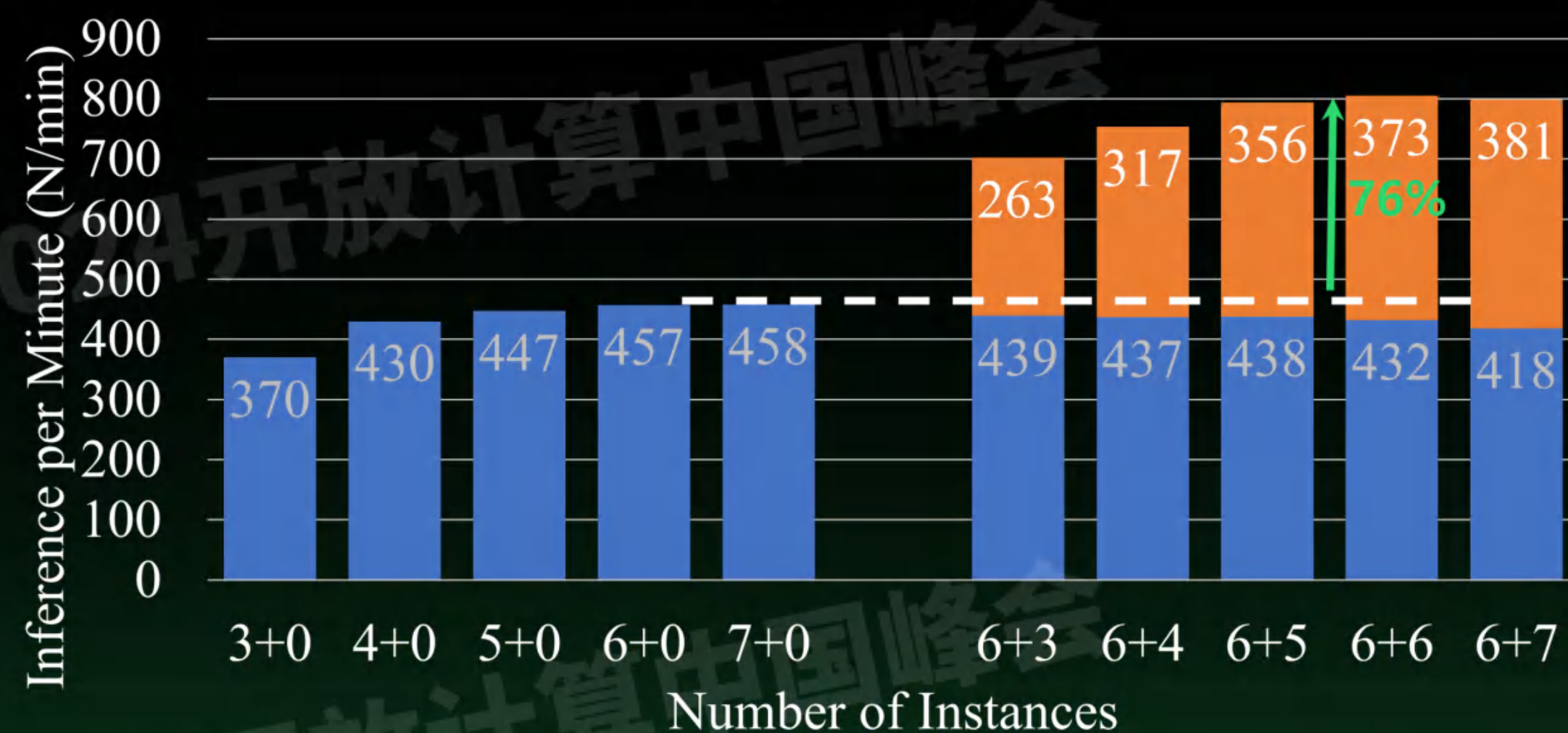
- **Page Temperature Perception** : 基于 NUMA balance 页表扫描和提示页故障之间延迟的算法来测量系统热/冷内存页面
- **Memory Demotion**: 将DDR内存中的冷页面降级至CXL内存
- **NUMA Balancing**: 将CXL内存中的热页提升至DDR内存
- **OOM**: 确保节点有足够的可用页面用于升级/降级



内存池化应用场景

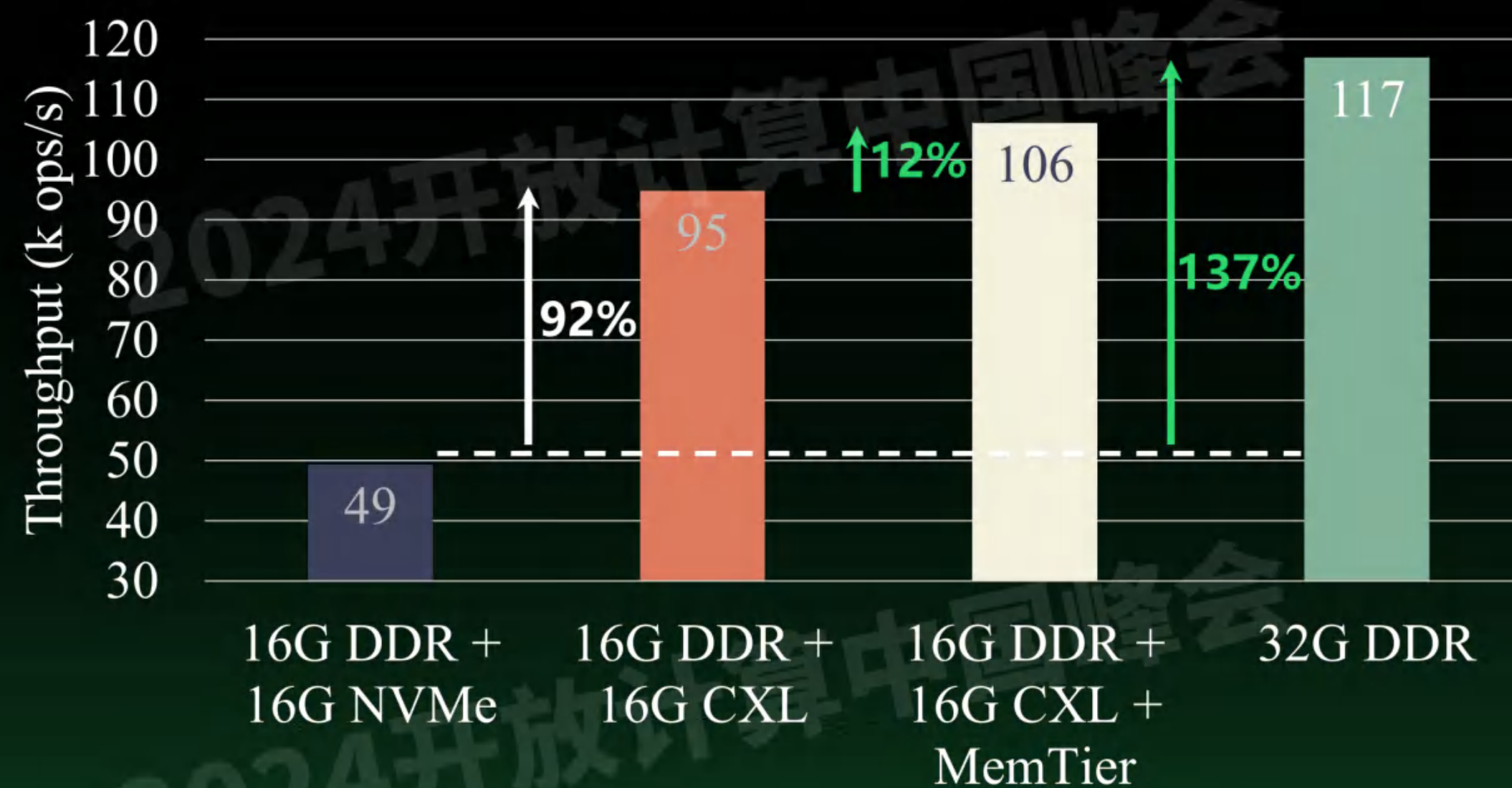
- AI应用场景下，CXL内存可额外扩展内存通道，增加AI应用可用带宽，提升应用吞吐量，BERT推理速度提升76%
- 内存数据库应用场景下，CXL内存可扩展数据库容量，相比无CXL架构，数据库吞吐量提升92%；Keyarch OS内存分层功能（无需修改应用程序）实现本地内存、CXL内存不同层次间内存页面冷热迁移，相比原生OS可提升吞吐量12%
- 向量数据库应用场景下，结合内存分层软件（无需修改内核和应用程序），充分利用DRAM和CXL存储器设备提供的带宽，相比全DDR配置，调整应用使用的DDR、CXL内存占比可提升应用吞吐量24%

AI应用 (BERT)



6+7：在DDR上运行6个BERT实例，在CXL上运行7个BERT实例，每个实例运行1000推理数

内存数据库 (Redis)



core: 20, read: update=7:3, distribution=latest

向量数据库 (Weaviate)



3200-4800MHz: DDR内存速率, core=40

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会 **IEIT**
SYSTEMS

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

THANKS

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会 **O**
E
I
T
SYSTEMS

2024开放计算中国峰会

互连技术探索：PCIe光互连

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

赵伟康

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

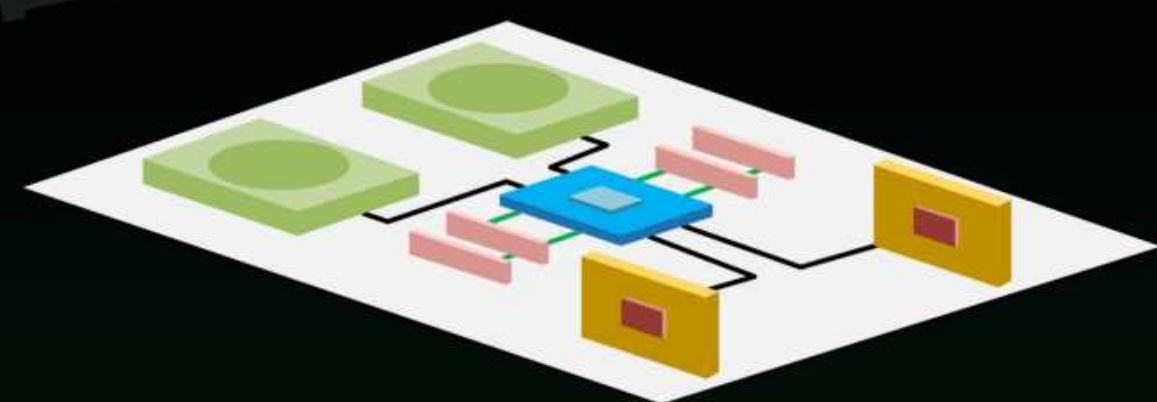
2024开放计算中国峰会

2024开放计算中国峰会

PCIe互连网络需支持远距离信号传输

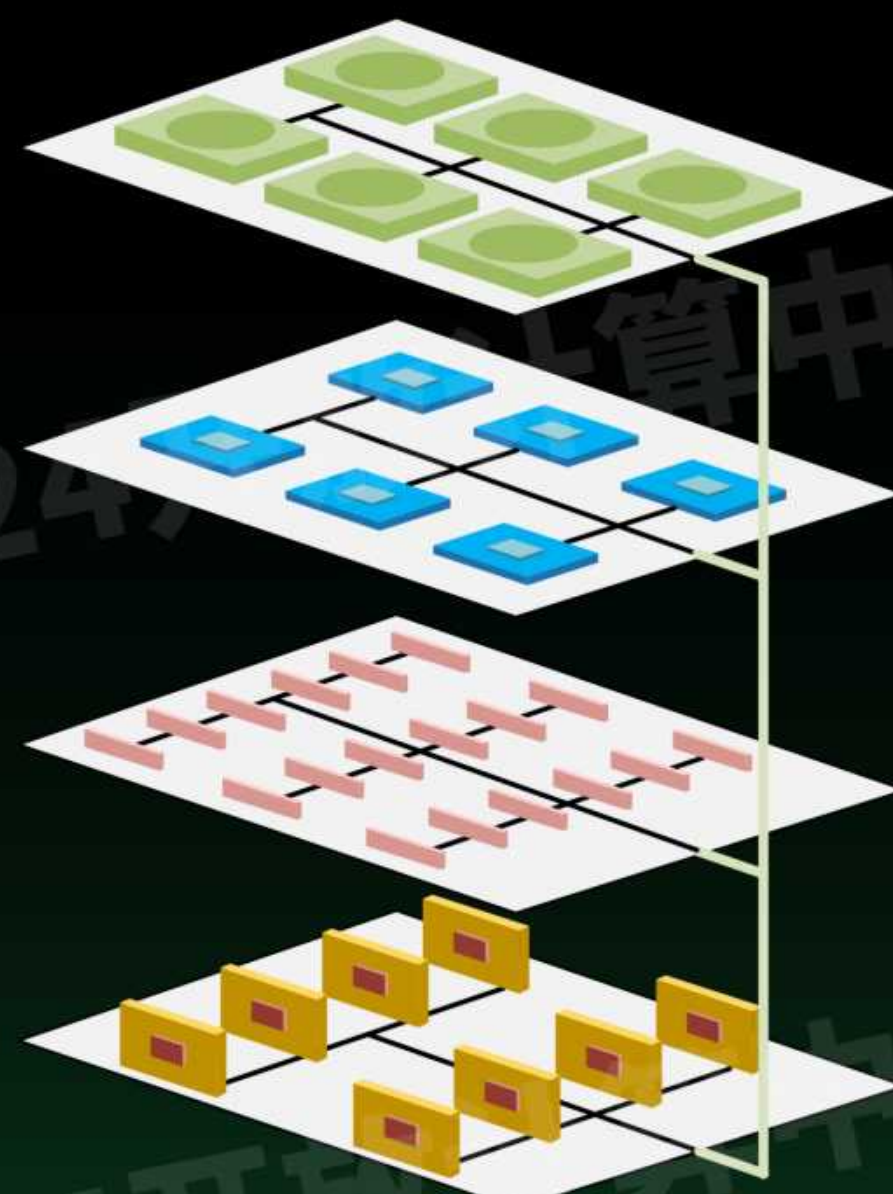
随着传统资源集中式服务器逐渐向资源解耦池化的融合架构服务器转变，驱动PCIe技术逐渐从单机内部互连，向机柜内互连，进一步向数据中心内跨机柜互连发展

机内互连
<1m



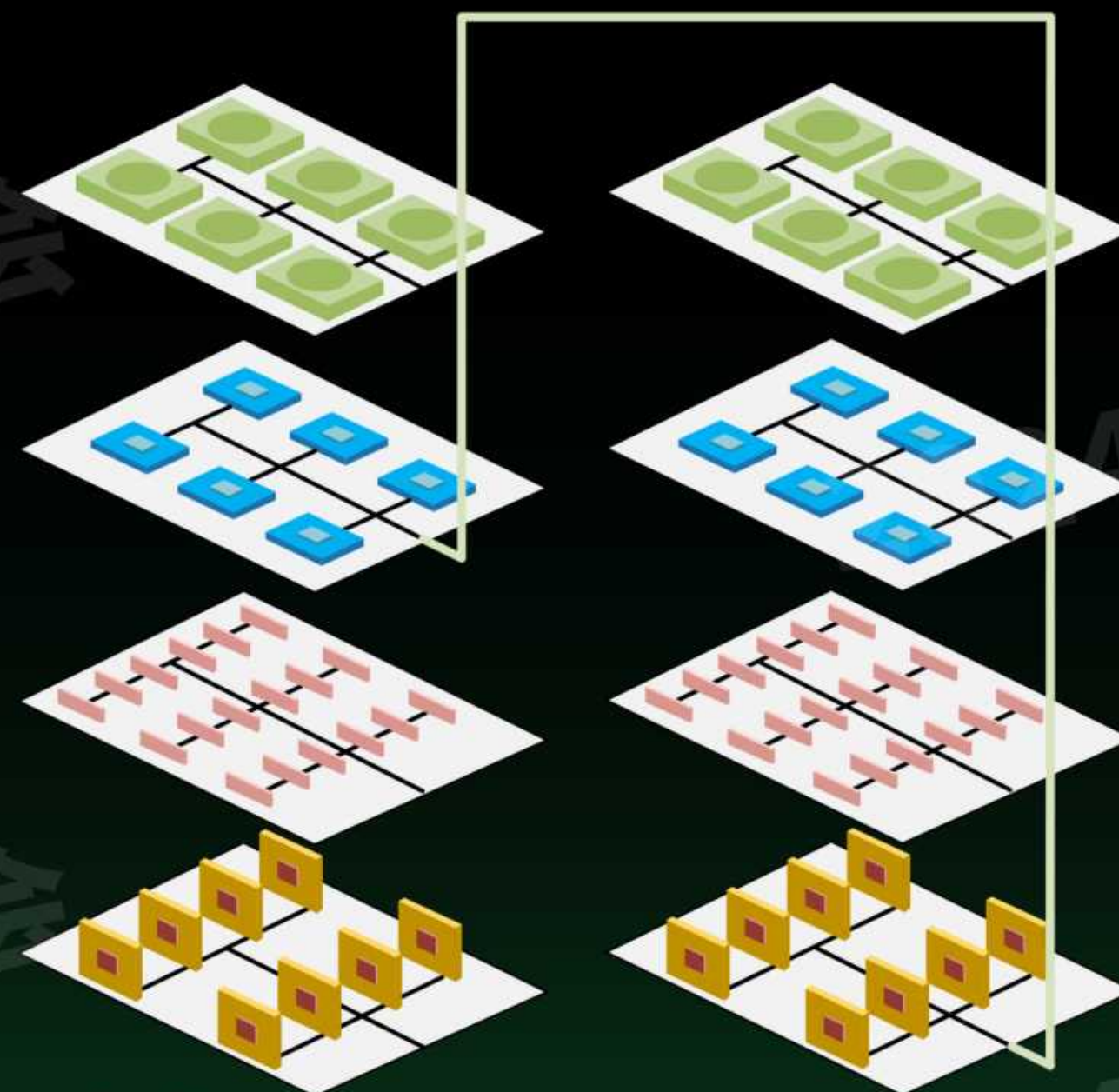
Converged Server

柜内互连
<3m

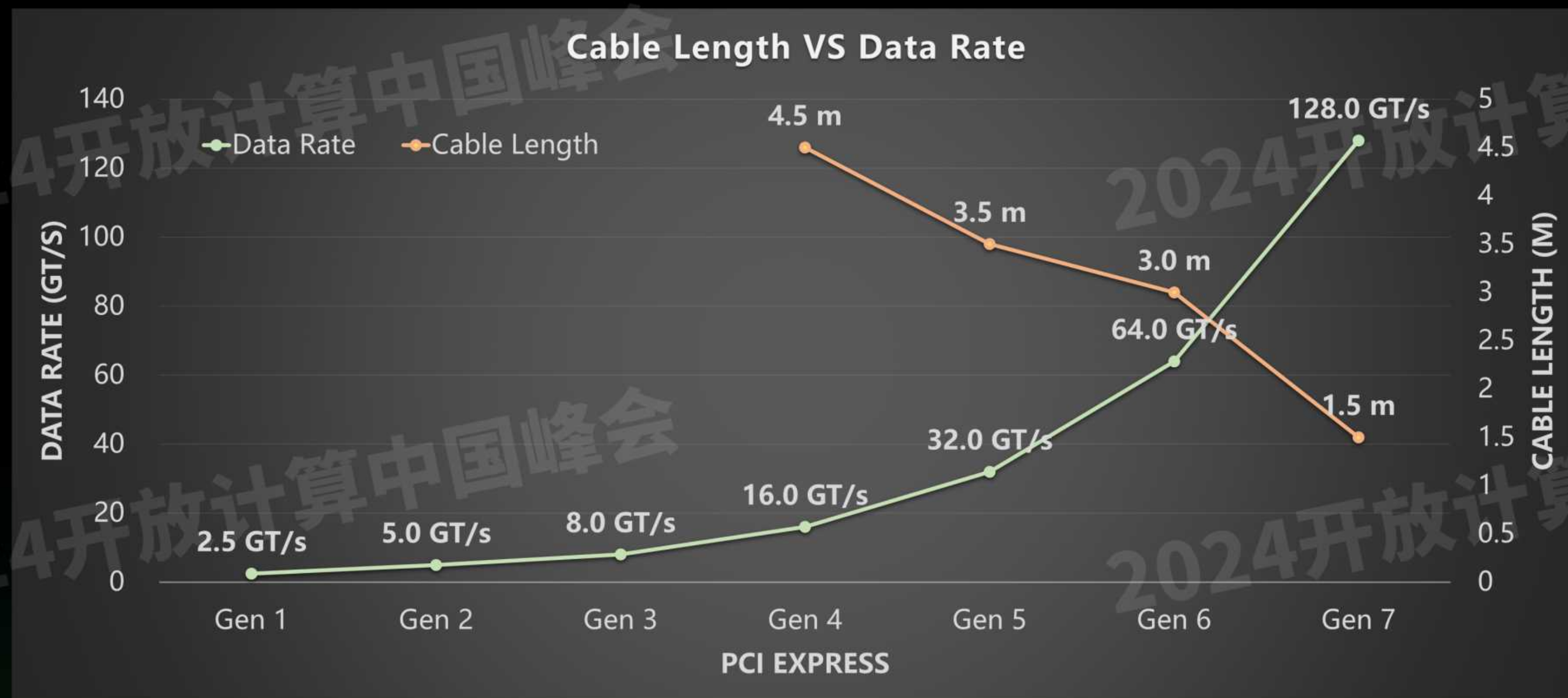


Disaggregated Server

柜间互连
<100m

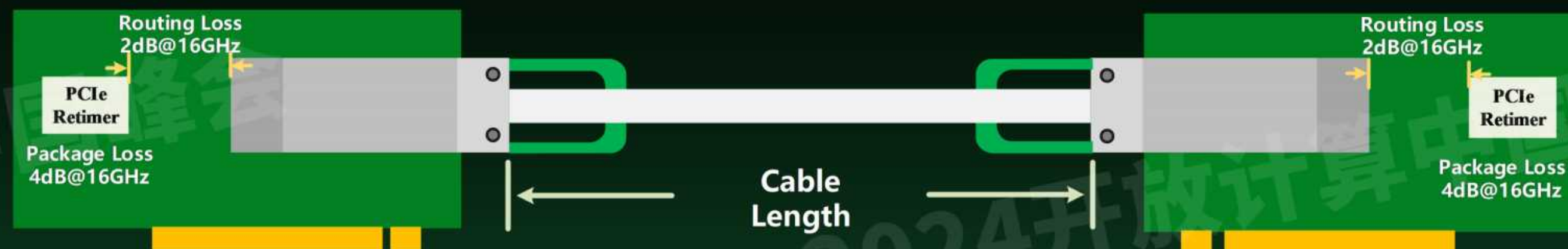


铜互连传输距离不断缩短，光互连优势凸显



- ✓ 传统铜互连因**损耗**问题限制了PCIe信号的传输距离，最远传输距离在**不断缩短**
- ✓ 光纤链路的低损耗特性可以支持高速信号远距离传输，成为构建跨机柜远距离PCIe互连网络的重要技术手段

铜缆长度评估：



Ref. PCIe 7.0 External Cable Spec.

$$IL(dB) = -2.0 - 0.28125 * f(GHz)$$

30AWG铜缆 $IL_{\text{Typical}} = 4.25\text{dB/m} @ 8\text{GHz}$

$6.5\text{dB/m} @ 16\text{GHz}$ $11\text{dB/m} @ 32\text{GHz}$

PCIe协议信号光传输难点

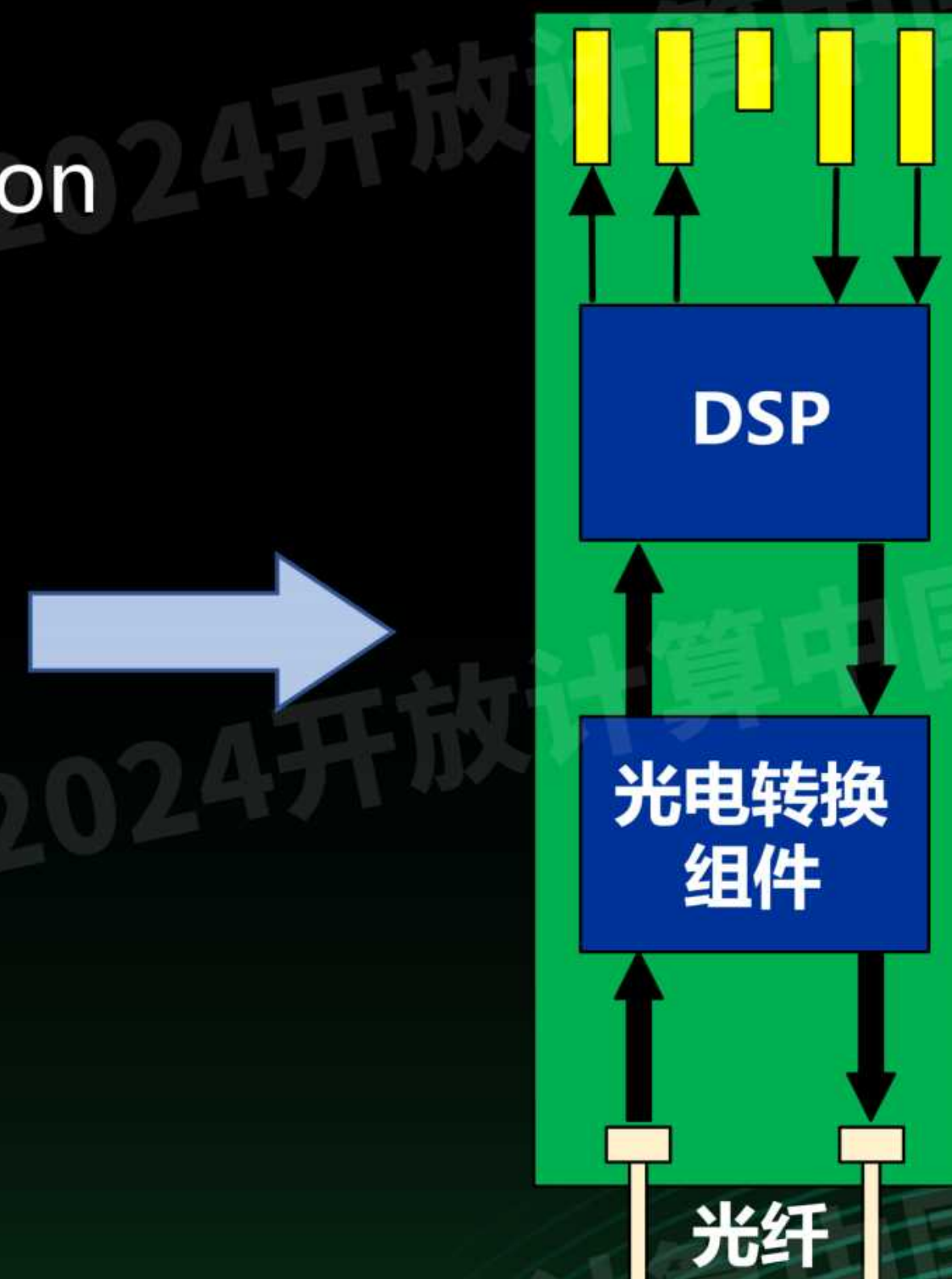
- ✓ PCIe协议最初主要针对计算机内部芯片和设备之间的互连，电互连技术成熟且能满足高速数据传输需求，光传输技术不成熟且成本较高，因此早期没有考虑光传输方式
- ✓ 光传输难点：
 - 1) PCIe协议向下兼容，链路速率范围较宽；
 - 2) 训练过程中的RD和EI易造成建立失败和误码；
 - 3) 辅助信号需要传输以实现了对对端设备的状态控制；

链路训练过程

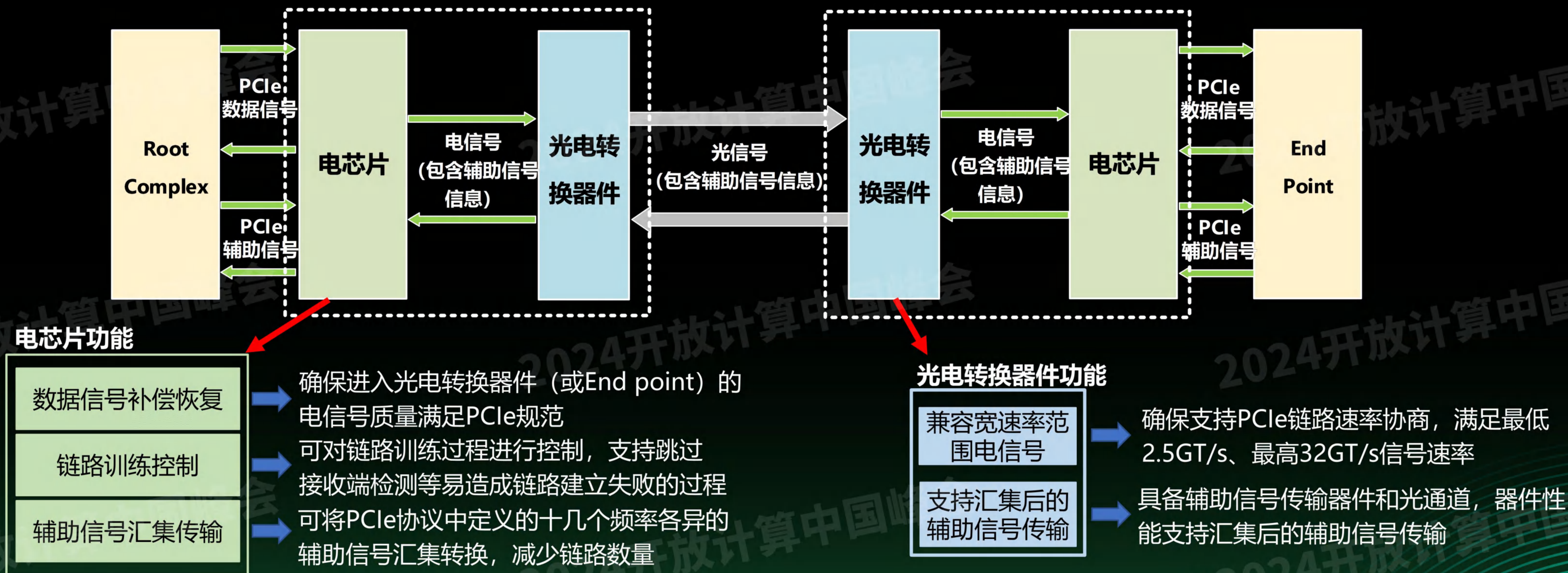
- Link Rate Negotiation (2.5GT/s~32GT/s)
- Electrical Idle
- Receiver Detect

辅助信号传输

- PERST#
- SMBDAT
- SMBCLK
- PRSNT# 等



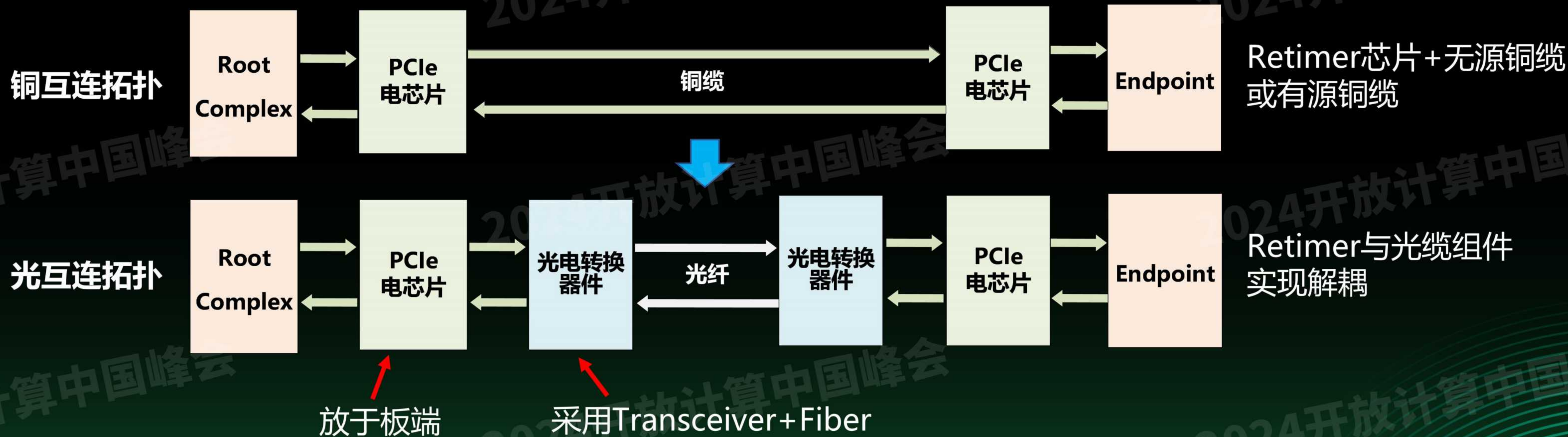
PCIe光互连链路方案功能分析



备注：Root Complex，根复合体、End Point，端点设备

PCIe光互连链路拓扑形式分析

验证链路拓扑形式参照外部铜缆互连拓扑，将PCIe电芯片放置于AIC卡，光电转换器件采用Transceiver光模块，链路各组件实现解耦，组件选择更灵活，便于调试测试



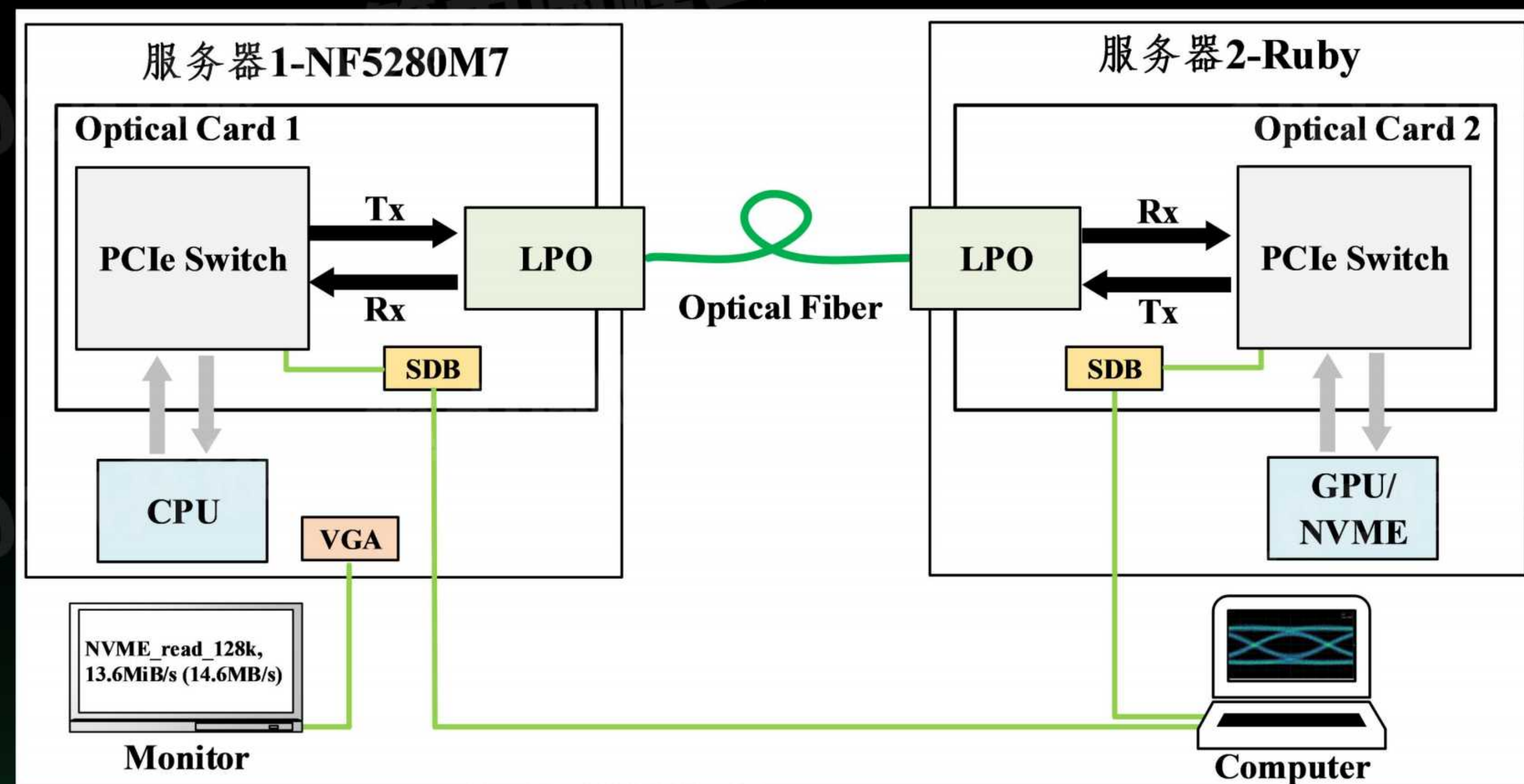
PCIe Gen5光互连原型样机



样机规格:

1. 具备PCIe x16/x32 金手指
2. 支持 PCIe 5.0
3. 支持 x4, x8, x16链路配置
4. 支持PERST#等辅助信号的双向传输
5. QSFP接口, 支持不少于30m传输距离

传输性能及应用测试

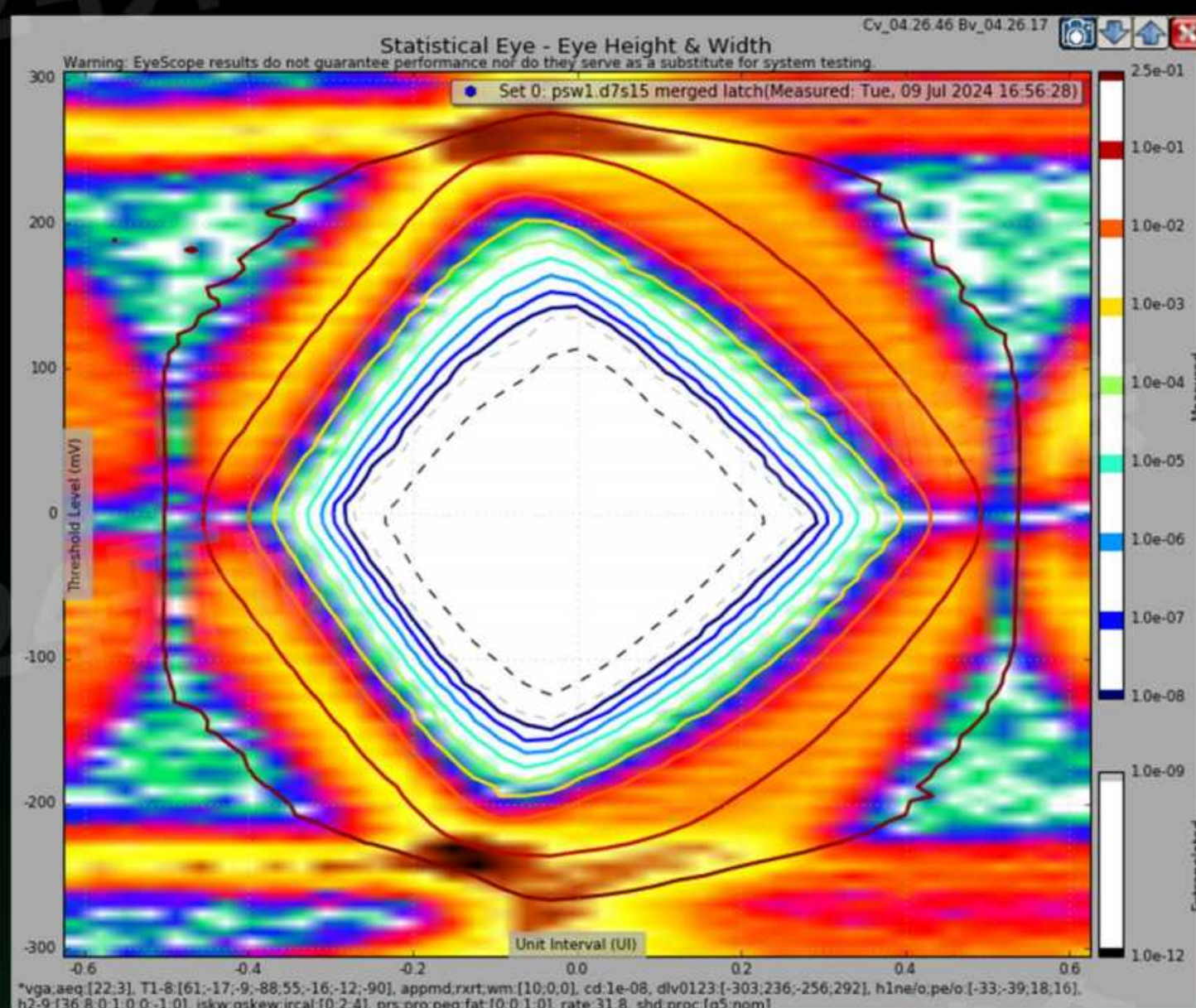


测试简介:

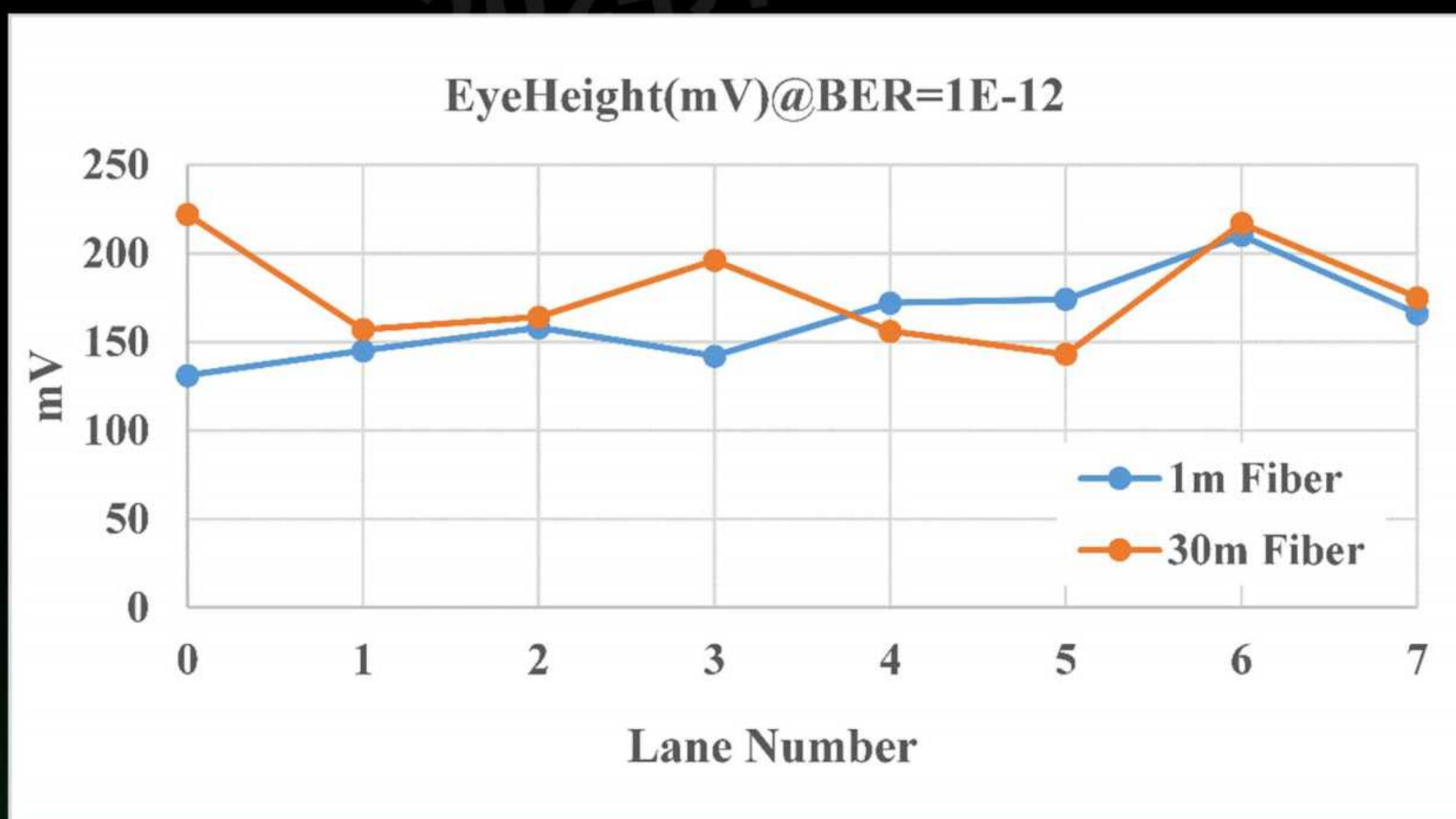
- 1.测试平台服务器: 1为NF5280M7, 2为自研的Ruby
- 2.两种不同长度的测试光纤: 1米和30米
- 3.测试项包含NVMe硬盘读写测试和GPU带宽测试
- 4.误码测试采用安立MP1900, 搭配QSFP112G MCB板

性能测试结果

眼高眼宽结果符合PCIe Gen5眼图规范，30m光纤链路眼高和眼宽结果与1m光纤链路基本一致，表明传输性能几乎不随链路长度增加而衰减，这是铜互连链路所不具备的优势



典型30m光纤眼图



1m和30m光纤眼高眼宽测试结果

性能测试结果

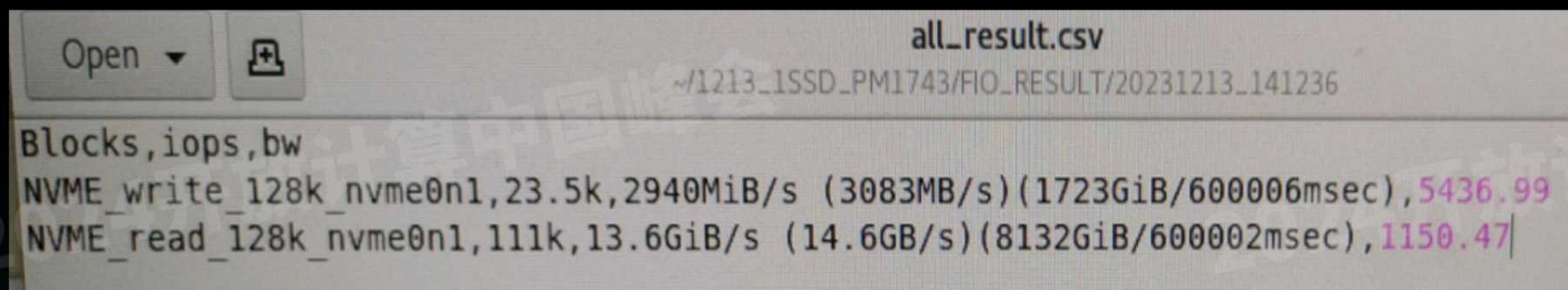
序号	速率	光纤长度	ISI板 IL (@16GHz)	测试结果
1	Gen5	1m	-	Pass
2	Gen5	1m	5.4dB	Pass
3	Gen5	1m	7.4dB	Pass
4	Gen5	1m	9.5dB	Pass
5	Gen5	1m	11.5dB	Pass
6	Gen5	1m	12dB	Pass
7	Gen5	1m	12.5dB	Pass
8	Gen5	1m	13dB	Fail
9	Gen5	30m	12dB	Pass
10	Gen5	30m	12.5dB	Pass
11	Gen5	30m	13dB	Fail

串接ISI误码测试可以得到当前PCIe光互连设计整个互连链路仍有12.5dB的损耗裕量。互连链路的剩余损耗裕量可以用于PCB板端设计，增强互连链路设计的灵活性

应用测试结果

1) NVMe-based FIO测试

2小时读取测试结果 (1.92TB三星PM1743)



all_result.csv	
~/1213_1SSD_PM1743/FIO_RESULT/20231213_141236	
Blocks,iops,bw	
NVME_write_128k_nvme0n1,23.5k,2940MiB/s (3083MB/s)(1723GiB/600006msec),	5436.99
NVME_read_128k_nvme0n1,111k,13.6GiB/s (14.6GB/s)(8132GiB/600002msec),	1150.47

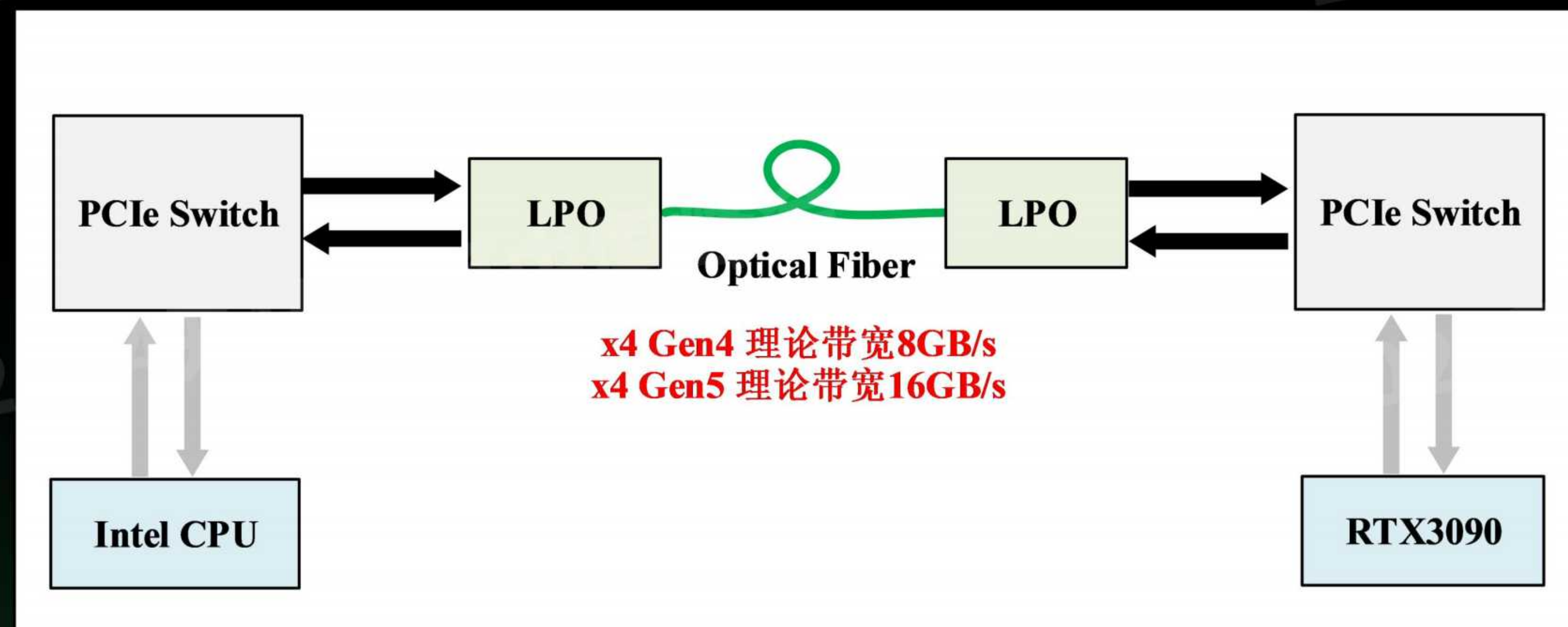
*30m光纤

1. 128k顺序读取速率为**14.6GB/s**，128k顺序写入速率为**3.08GB/s**，作为参照PM1743 Spec直连CPU 128k顺序读写速率分别为14GB/s和3GB/s，光互连链路测试结果与铜互连相当，测试结果符合预期，满足应用需求
2. 硬盘测试中无AER错误，无掉盘、掉速、降带宽、重启、宕机等异常现象，光互连长期传输链路稳定

应用测试结果

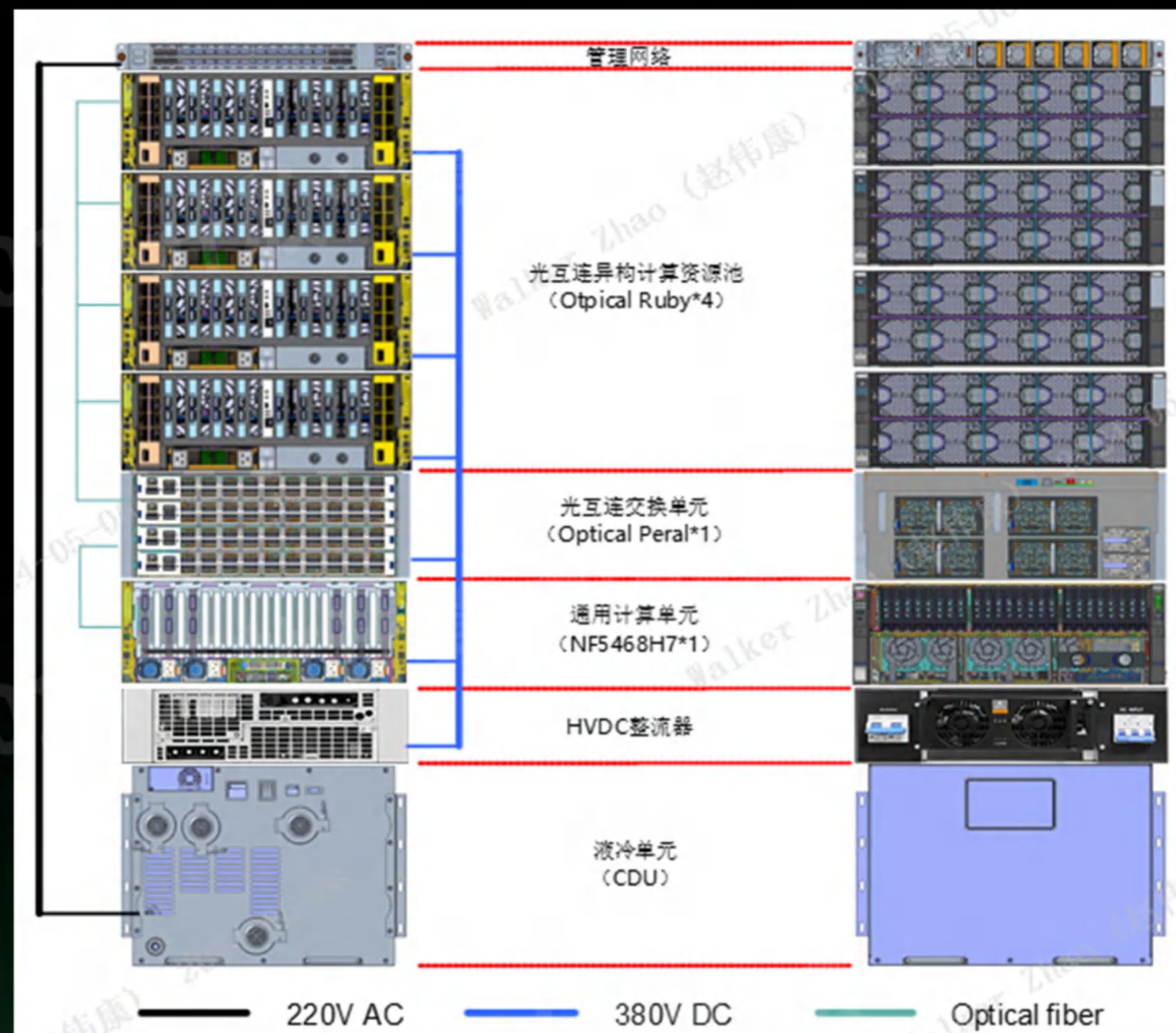
2) GPU带宽测试

x4光互连链路GPU带宽测试结果 (RTX3090, PCIe Gen4)

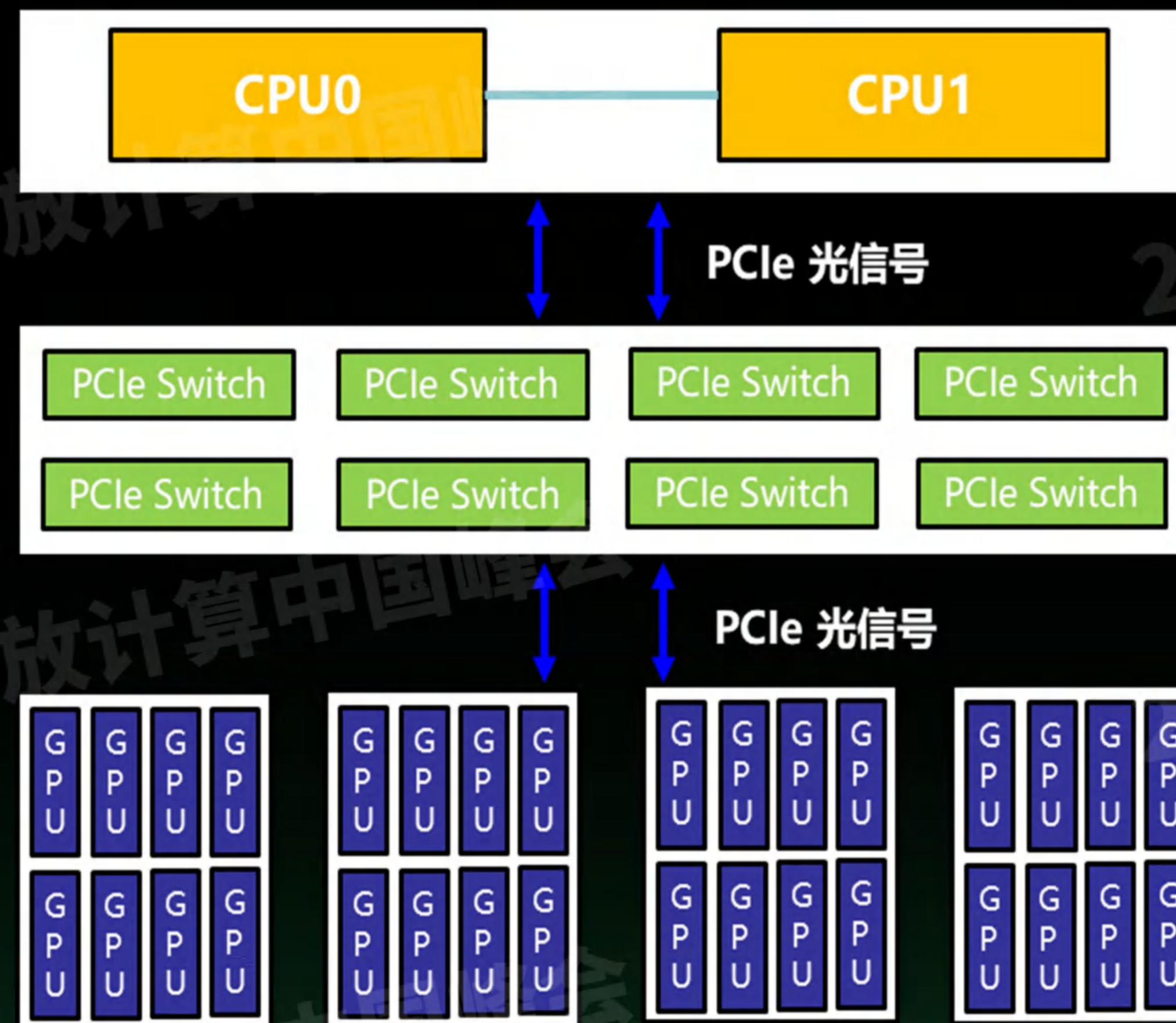


- 使用CUDA进行24小时带宽测试， Gen4 x4实测带宽为 6.7GB/s (84%)， Gen5 x4实测带宽稳定在 13.5GB/s (84%)， 作为参照RTX3090直连CPU Gen4 x16带宽为 26.2GB/s (82%)， **光互连链路测试结果与铜互连测试结果相当，测试结果符合预期**
- CUDA测试期间全链路均未出现AER错误，链路稳定

聚焦光互连系统开发测试和兼容性测试

通用计算
单元

交换单元

异构
资源池

1) 实现通用计算单元与异构加速资源池之间PCIe全光互连, 完成实际应用场景测试

2) 联合业内头部光模块/芯片等厂商推进PCIe光互连传输兼容性测试

PCIe融合架构光互连系统

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会 **IEIT**
SYSTEMS

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

THANKS

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会

2024开放计算中国峰会