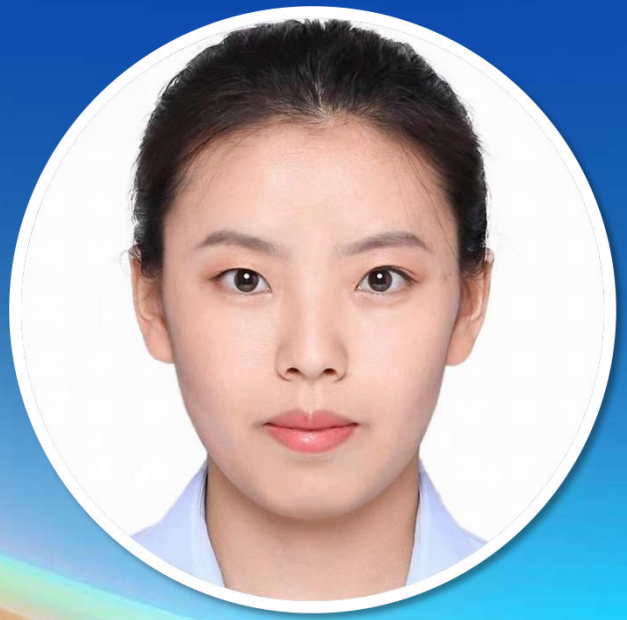


UXM模组在通用超节点中应用的探索

张环

锐捷网络　硬件架构师

AI范式跃迁推动超节点算力基础设施持续演进

Global Computing Consortium 全球计算联盟

行业落地加速

国产主要大模型日均Token调用量突破**500万亿**（截至 2026 年 6 月），半年增长**5倍**，场景渗透千行百业

算力重心转移

从训练为主转向**推理优先**，推理成AI价值兑现主战场，未来看持续产出效率，推理算力将远超训练。

超节点规模持续跃升

超节点从64/128卡向**千卡**演进，单机柜叠加更大的算力密度成为趋势，核心是TCO的收益

互联技术从铜走向光

铜互联带宽受限，2026光互联元年，预计30年Scale-Up光互联渗透率超**30%**，沿NPO→UPO→CPO



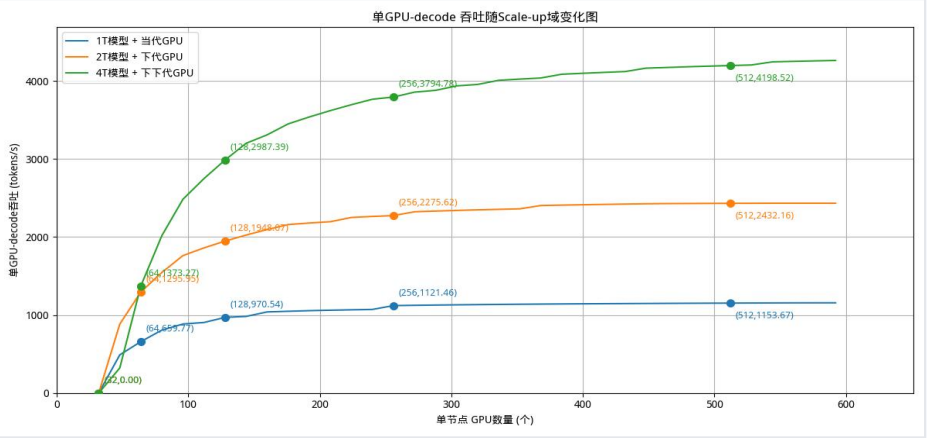
超节点推理规模收益

Global Computing Consortium 全球计算联盟

GPU	代际	时间	BF16 (TFLOPS)	FP8 (TFLOPS)	访存带宽	通信带宽(单向)	HBM
NVIDIA Rubin	N+1	2026	4375	8750	13TB/s	1.8TB/s	288GB
NVIDIA B200	N	2024	2250	4500	8TB/s	900GB/s	192GB
NVIDIA H100	N-1	2022	989	1979	3TB/s	450GB/s	80GB
N+2 国产GPU	N+2	2028	2000	4000	6.4TB/s	1.6TB/s	256GB
N+1国产GPU	N+1	2027	1000	2000	4.8TB/s	800GB/s	196GB
N国产GPU	N	2026	500	1000	3.2TB/s	400GB/s	144GB

1.9~2.3x　1.6~2.6x　2.0x　1.5~2.4x

2.0x　1.5x　2.0x　1.3x



不同参数量模型在不同HBD下收益仿真

模型代际	国产GPU代际	训推场景	SLO	输入上下文	输出上下文	scale-up域
模型参数量：1T	N GPU（算力：1000T@fp8）	推理	TPOT < 50ms	4k	1k	32~512
模型参数量：2T	N+1 GPU（算力：2000T@fp8）					
模型参数量：4T	N+2 GPU（算力：4000T@fp8）					

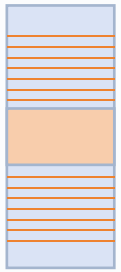
未来2~3年内，推理场景超节点域收敛到512卡左右

- 算力和模型如果超预期发展，超节点域需求会进一步增加
- 模型架构和训推技术也同样在影响，例如EPLB、MageMoE等新技术均对超节点有诉求

观点：对超节点Scale Up互连趋势的总结

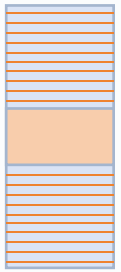
Global Computing Consortium 全球计算联盟

- **从模型演进看：**综合考虑训练/推理场景，256/512卡规模可以满足未来3年的需求
- **从组网技术角度：**超节点最适合采用scale-up单级组网，流控/时延等更优
- **从可靠性角度：**尽管“光进铜退”是业界长期演进的方向，但在未来1～3年的窗口期内，**“可用铜则用铜”**已成为超节点互联架构的核心设计原则。这一策略选择源于对成本、功耗与可靠性的综合权衡；
- **面向未来的角度：**随着未来需求演进，HBD增大，铜互联有上限，光互联是必然选择。如采用光互联，那么单层光互联，相比于两层组网，具备拥塞控制、成本等优势；在更大的HBD域，铜光融合组网更具备优势



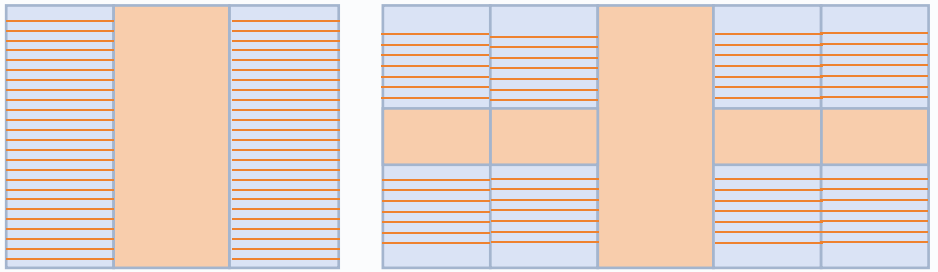
超节点1.0

- 整机柜-112G平台
- 单级铜互连
- HBD域-64~128卡



超节点2.0

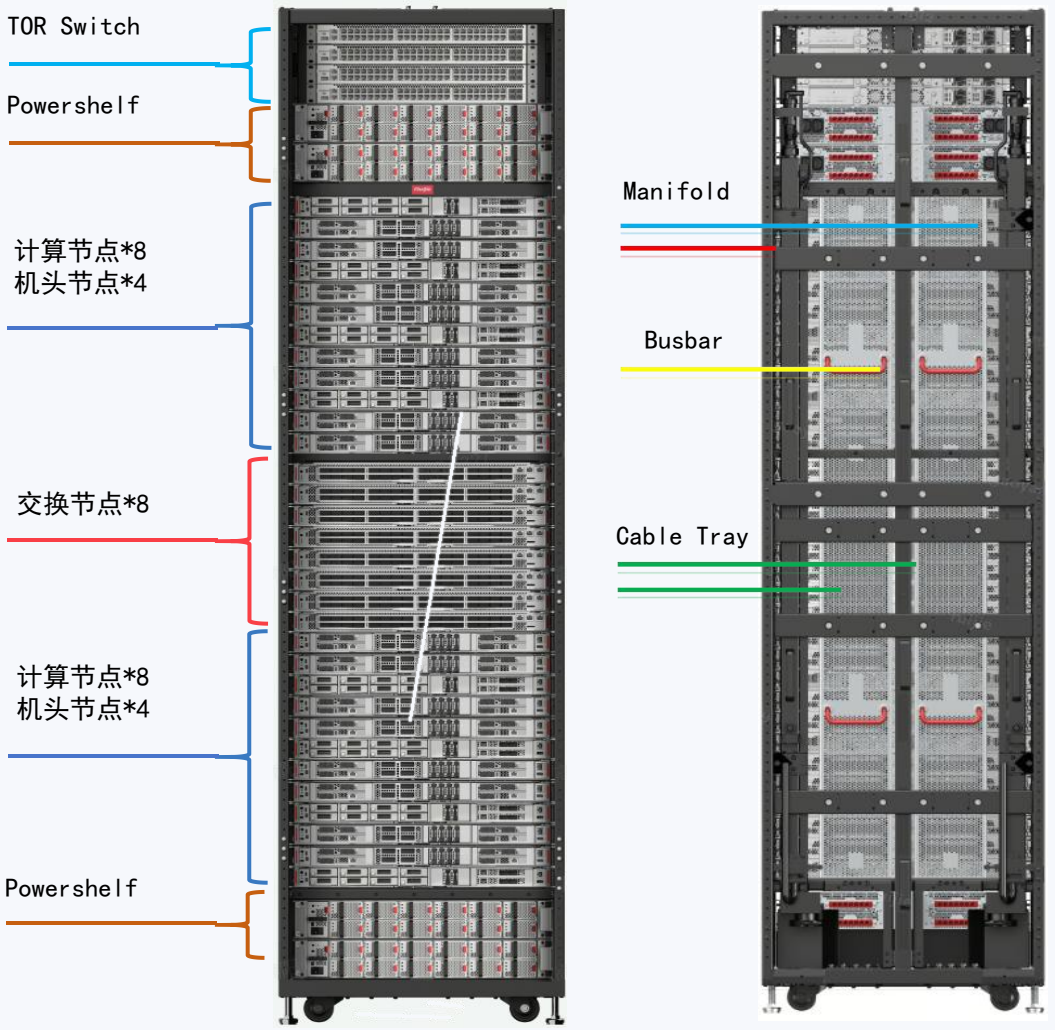
- 整机柜-224G平台
- 高密单级铜互连，极限密度
- 更大HBD域，128~256卡



超节点3.0

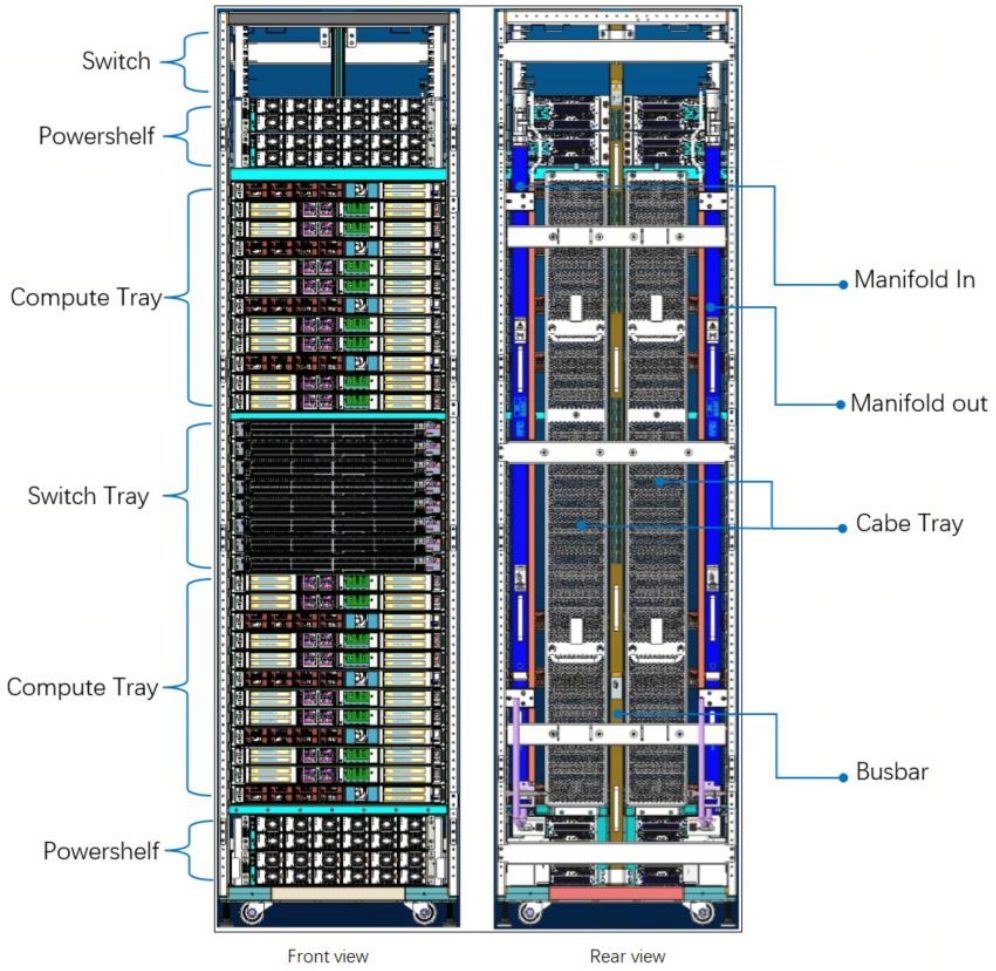
- 复用整机柜平台，集中式供电与供水
- 单级光互连、光电融合互连、2级光互连
- 超大HBD域，512~1024卡以上

超节点1.0架构通用机电平台设计



机柜属性	规格参数
机柜尺寸	1400mm*600mm*2286mm （320U+12RU）
节点槽位	16个计算槽位 8个机头槽位（非解耦方案不使用） 8个交换槽位
满配净重	~1300kg
工作温度	进风温度35℃ @1800米 进液温度40℃ @1800米
机柜供电	-4个PowerShelf，最大支持整机供电240kW -PSU支持双路输入，支持N+1冗余 （64卡每柜时，GPU单卡1.5KW，整机最大功耗~150kW）

超节点2.0架构通用机电平台设计



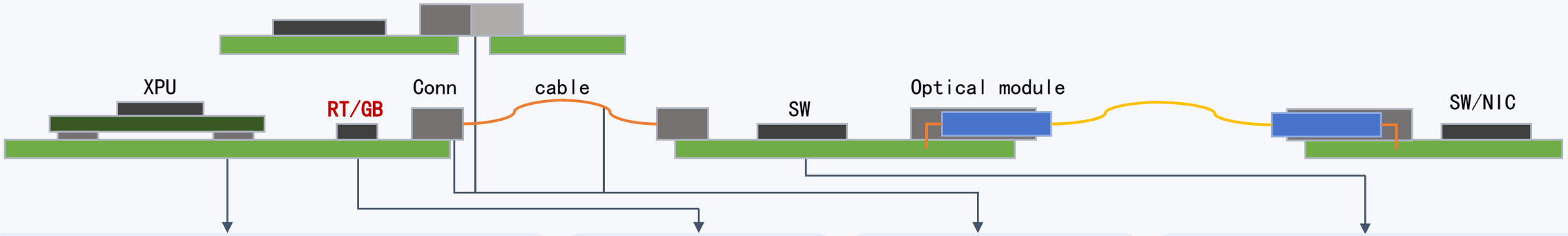
机柜正面如图所示

- 前窗布局16个计算节点+8个CPU节点
- 布局8~12个交换节点
- 支持4个powershelf，整机功率支持240kw以上
- 单节点支持的供电功率≥15kW，液冷散热支持全冷板式液冷散热

机柜背面

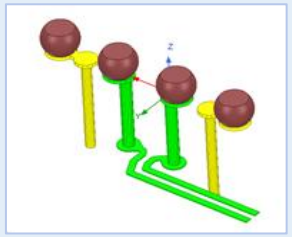
- 后窗布局2个224G cabletray
- 1个busbar，支持240kw的供电
- 1组manifold，支持240kW的液冷通流量

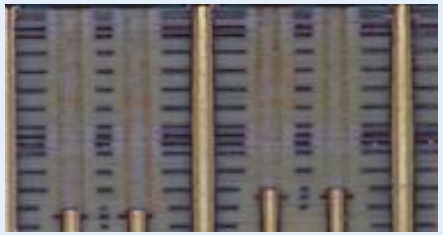
超节点2.0架构通用机电平台设计



PCB相关已完成产品化：

- 板材验证
- PCB相关工艺验证、测试、板厂能力摸底等
- 互连设计





224G信号中继：

- RT、GB方案如Brcm、Credo等已进入MP

224G连接：

- cable+conn方案，当前海外是主流，如AMP

SW已开始产品化：

- SI、散热、模拟等工程域挑战
- 制造工艺的挑战



224G Serdes产业链，在行业内已基本成熟，具备批量供货能力

超节点2.0架构通用机电平台设计

Global Computing Consortium 全球计算联盟

交换单板前窗和后窗的IL预估：

- 后窗：12dB~15.3dB（不含连接器）
- 前窗：~24dB（含overpass线缆）

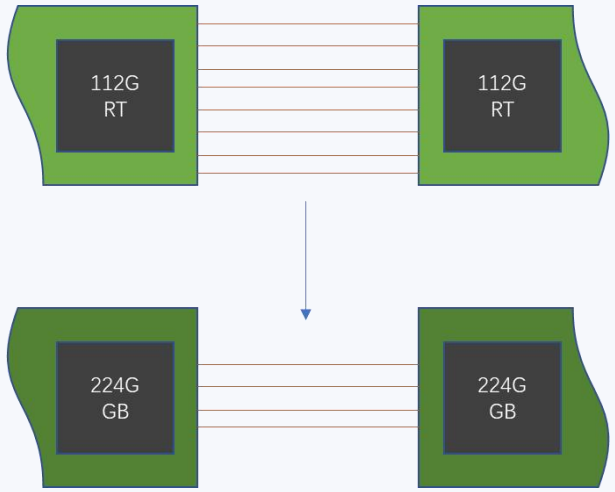
UBB RT-> MAC：

统计	Loss (dB)	Loss new	数量	备注
RT pkg	*2.3*		-	封装损耗-R
UBB Trace	*5*	8.3	5inch	2.5inch~5.0inch
UBB via	*1*		2	2个via
BP conn	*2.1*	2.5	1	含RAR与线端直公/4pair
cable	*11.04*	9.23	1.3	1.3m max（10+8）
BP conn	*2.88*	3	1	含RAR与线端直公/8pair
/	*0*	0	0	/
SW via	*1*		2	2个via
SW Trace	*8*	15.3	8	4.5~8inch
SW pkg	*7.5*		5	5~9.5dB-T
Total	*40.82*	38.33		

超节点2.0架构通用机电平台设计



基于224G速率升级，整机柜的scale-up带宽提升一倍，**交换槽位数不变，柜内的计算节点槽位数不被占用**



以整柜6.4T@112G*64卡 Scaleup带宽的XPU连接成本对比：

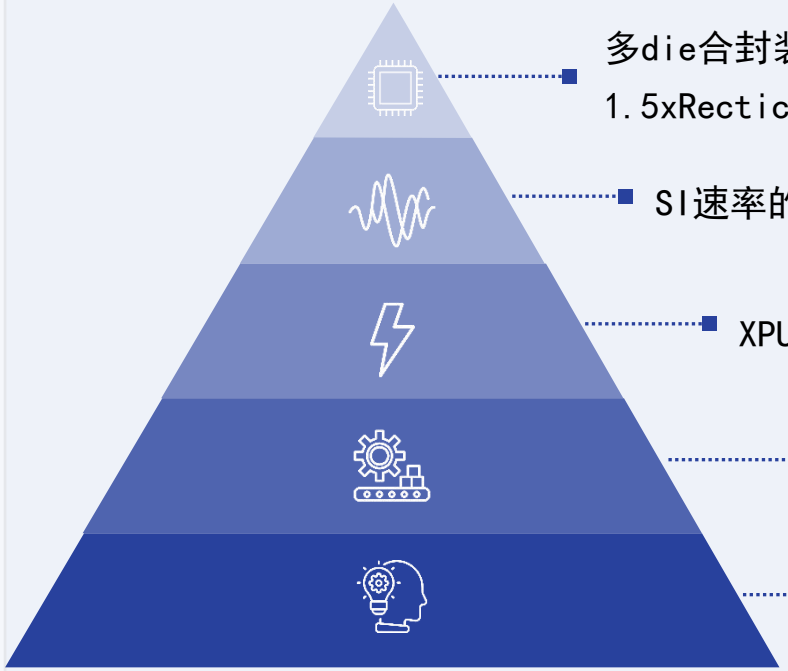
互连器件	112G	成本	224G	成本	备注
XPU侧板材	M8	2E	M9	2.8E ↑	
RT芯片	1.6T *4 16lane	2A	3.2T*4 16lane	2.6A	224G链路成本会更低，主要收益来自cable的数量减少；
Conn	112G/128pair	2C	224G/64pair	1.8C ↓	
Cable数量	128pair	2B	64pair	1.4B	
Conn pair	112G/128pair	2C	224G/64pair	1.8C ↓	
交换槽位数	8/16	2D	4/8	1.8D	

基于224G速率+Cable tray方案，单XPU卡的互连成本相比于112G方案，降低~8%

国产GPU快速演进对模组、系统生态带来的挑战

Global Computing Consortium 全球计算联盟

尺寸　功耗　散热



多die合封装，芯片尺寸的增大，
1.5xRecticle size->~3xRecticle size>更大尺寸

SI速率的提升，112G->224G->?

XPU功耗提升，700W->1400W->3000W+

模组集成的挑战，大功率、垂直供电、背向散热、全液冷等

系统设计的HBD域提升对整机工程域的技术难点

OAM 2.0标准的限制

芯片模组生产的挑战

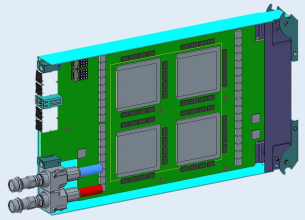
从PCIe/OAM转向UXM，适配从铜到光，从Box到整机柜多生态、多场景

Global Computing Consortium 全球计算联盟

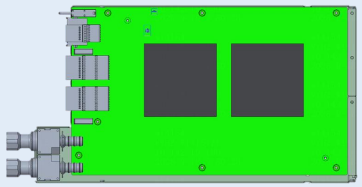
PCIe卡

OAM模组

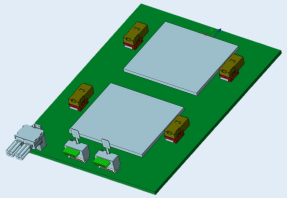
OPEN UXM Type1：
PCIe 原生液冷、高密、易维护



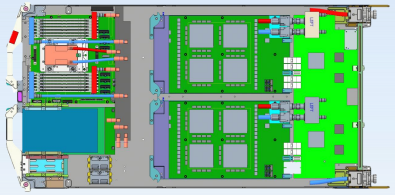
OPEN UXM Type2：
高功耗，大尺寸芯片，免下架维护



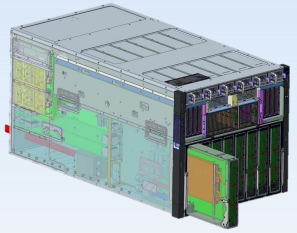
OPEN UXM Type3：
高功耗，大芯片，高带宽大Radix互连（ NPO/NPC ）



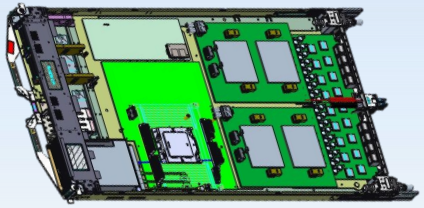
整机柜



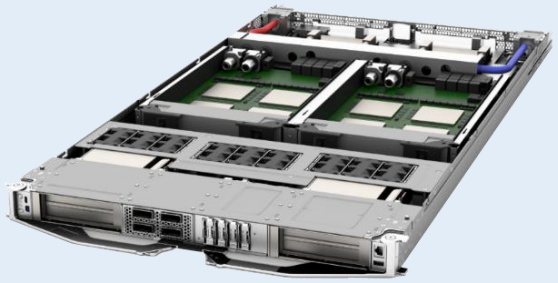
服务器Box



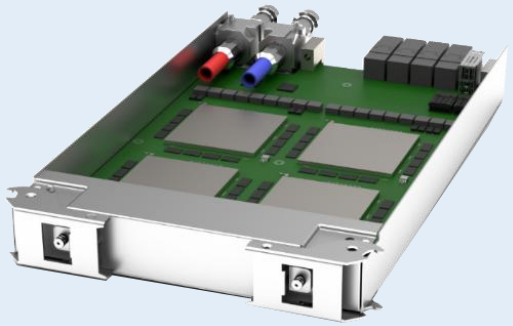
整机柜



基于UXM Type1验证系统的设计

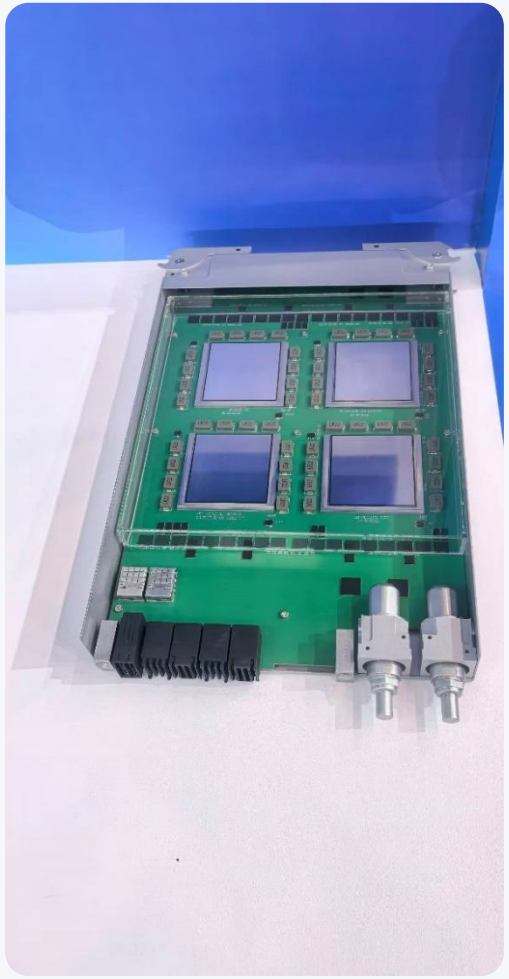


GPU Tray

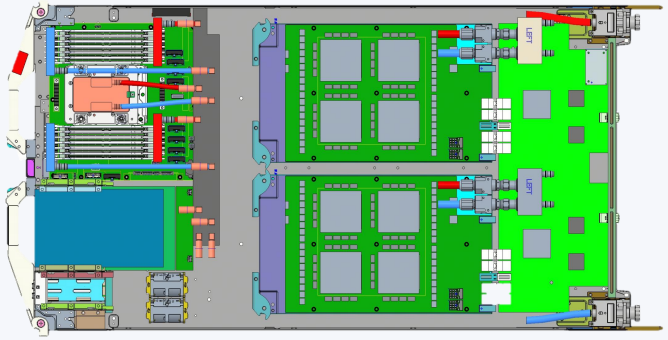


UXM Type1







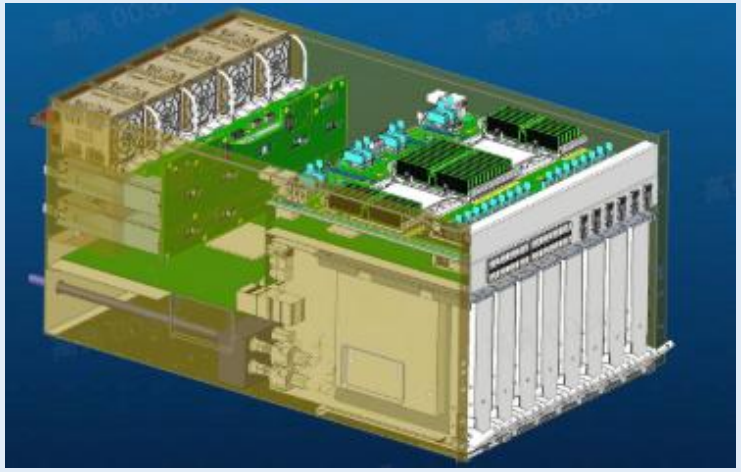


基于Type 1进行完整设计，推出Demo产品，UXM 的接口定义、散热接口、供电接口、维护接口都落实到了可插拔的模组和机箱里

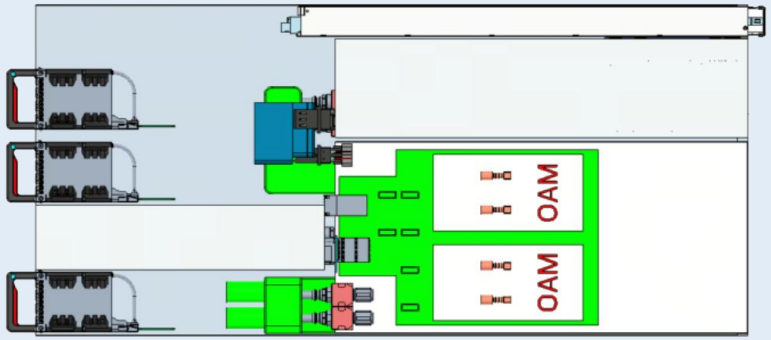
基于UXM Type2 打造16卡通用BoX平台

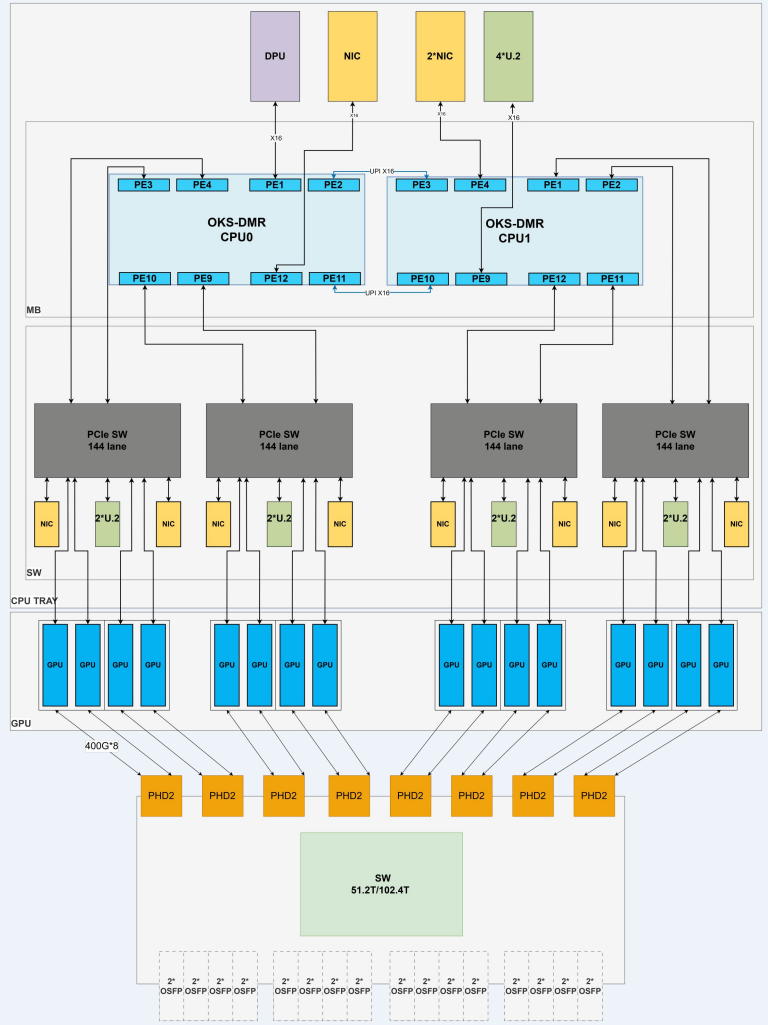
Global Computing Consortium 全球计算联盟

Type2:Switch in Server
实现8模组16芯，
模块化设计，支持多交换协议、多CPU平台



基于Type2 Form Factor
适配传统OAM *2 ，
提供灵活的产品过渡；





散热一致性：一致性不是「公差更严」，是「分布更收敛」

大芯片面积↑ → 界面局部不均 → 局部接触热阻↑ → **热点** → **整片散热能力↓**

TIM 导热垫 · 材料层

管均一性

核心失效：与上下界面接触不佳、间隙填充不充分

- 混料不均 → 导热系数本身就不均一
- 相分离（长期高温 / 反复相变）→ 性能衰减
- 回弹率不足 → 间隙变化填不满，接触差
- 破损 / 尺寸偏小 / 承压碎裂 → 有效面积↓

管控方向

验收从「标称值 + 公差」改为「分布指标」：导热系数分布、厚度分布、回弹率下限

相分离与老化验证写进寿命曲线，而不只做来料合格判定

冷板装配 · 结构层

管均匀性

核心失效：冷板 / MAC 板压合力不均匀

- 冷板制造偏差 → 扣合力不均匀
- 弹簧力偏差 / 老化 → 弹力不足或过大
- 螺钉未锁紧 → 局部压力缺失
- 结构件平面度 / 变形 → 局部间隙不均

管控方向

总扣合力合格 ≠ 压力分布合格：用压力膜 / 压敏纸测「分布」

弹簧力分组选配，锁付顺序与扭矩固化到工艺文件

抓手不是加严公差，而是三件事

分布可定义　压力可测量　过程可追溯

大尺寸芯片焊接：窗口是常数，变形是变量

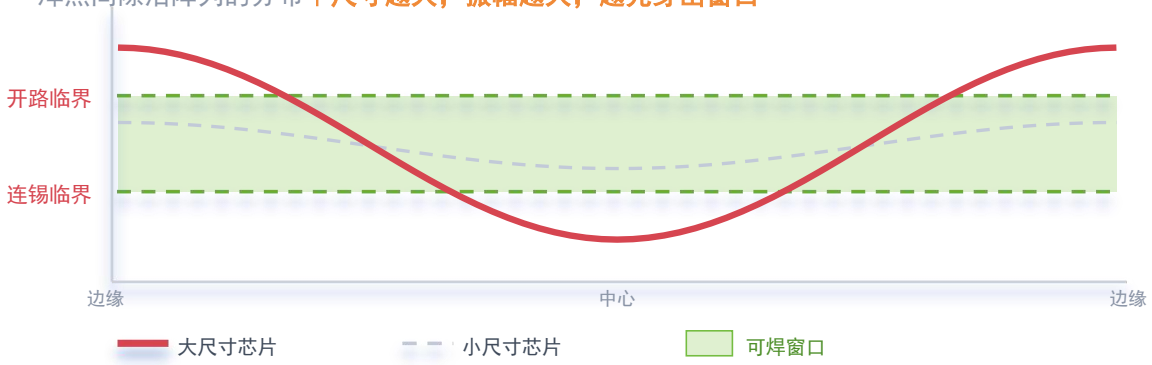
Global Computing Consortium 全球计算联盟

焊球塌落后的间隙由锡量决定，与芯片尺寸基本无关；芯片与 PCB 的翘曲却随尺寸放大 —— 间隙不涨、变形在涨。

容差被单向吃掉：边缘先开路、中心先连锡。尺寸越大，可焊窗口越先消失。

一 · 难点：变形吃掉了窗口

焊点间隙沿阵列的分布 | **尺寸越大，振幅越大，越先穿出窗口**



01 **双向失配**

芯片与 PCB 在回流温区反向翘曲，边缘被拉开、中心被压紧。

02 **窗口收窄**

设备精度换不回被变形吃掉的余量，窗口只能靠管住变形来保。

03 **判据缺失**

焊点在芯片正下方，二维投影看不出形貌；没有量化判据就拦不住。

二 · 做法：把变形管在焊之前

焊前 **先判可焊性**

先用“芯片翘曲 × PCB 翘曲”把可焊边界算出来。
焊不了的板子不进产线，而不是进了产线再调炉温去救。

焊中 **分区补偿**

变形是空间分布的，补偿也必须分区：锡量按变形量给，约束按变形方向加。
用一组统一参数对付变形，必然失效。

焊后 **量化判据**

无损检测只能看到投影，判据必须由切片标定。
把焊点形貌变成可判定的等级，缺陷才拦在制造端。

失效 **闭环返修**

大芯片热容大、热梯度难控。
能否高良率拆装，决定这套工艺是“能焊”还是“能用”。

大芯片焊接的成熟度，不取决于设备精度，而取决于变形是否被完整管住。

设计端算准 · 工艺端补准 · 检测端判准 · 失效端兜住

Global Computing Consortium
全球计算联盟

2026全球计算大会（CGC2026）

COMPUTING GLOBAL CONFERENCE 2026

谢谢！