Open AI Infra

下一代XPU模组设计的方案探讨

锐捷网络-CBG
程旭升-系统架构师

Global Computing Consortium
全球计算联盟

目录

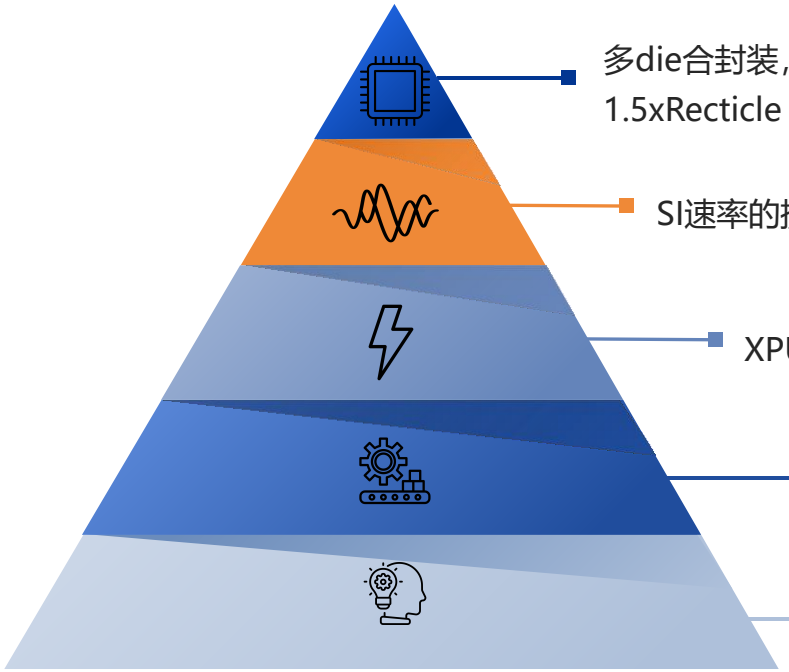
1、下一代XPU模组的挑战及设想

2、下一代XPU模组设计的方案探讨

3、基于XPU模组的超节点计算tray设计方案

4、单层光互联超节点设计探讨

下一代XPU模组面临的挑战



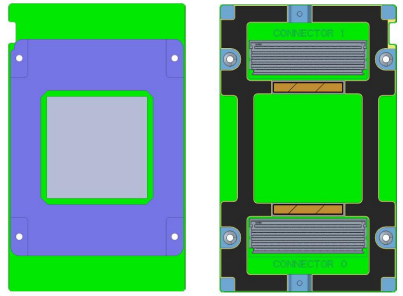
多die合封装，芯片尺寸的增大，
1.5xRecticle size->~3xRecticle size>更大尺寸

SI速率的提升，112G->224G->?

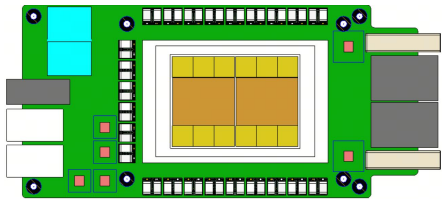
XPU功耗提升，700W->1400W->3000W+

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



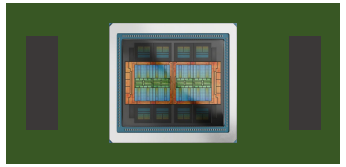
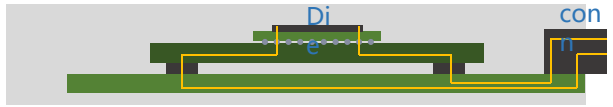
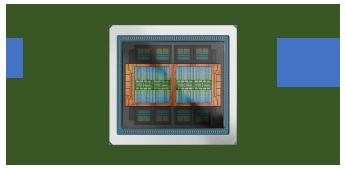
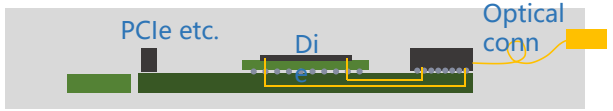
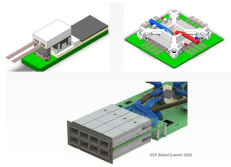
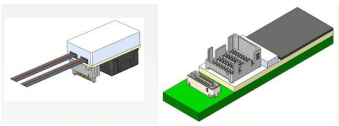
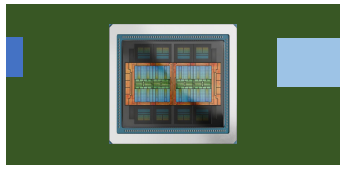
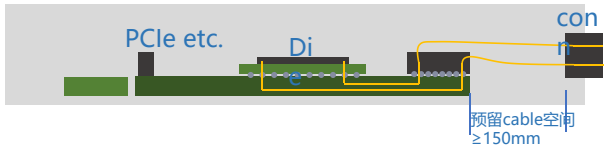
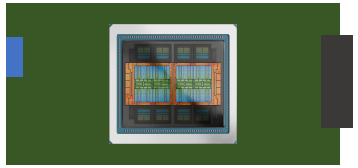
OAM 2.0



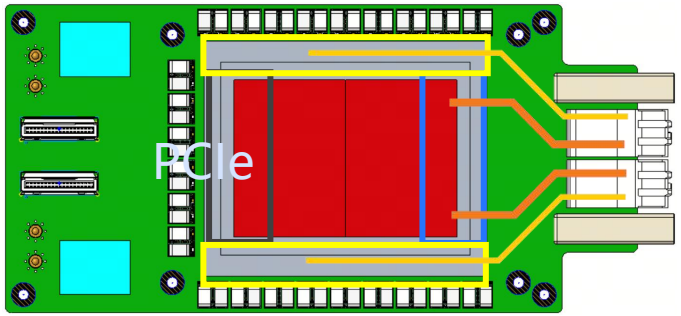
新XPU卡形态?

模组设计综合考虑：XPU尺寸、serdes扇出，供电、DC电源、散热、应用场景

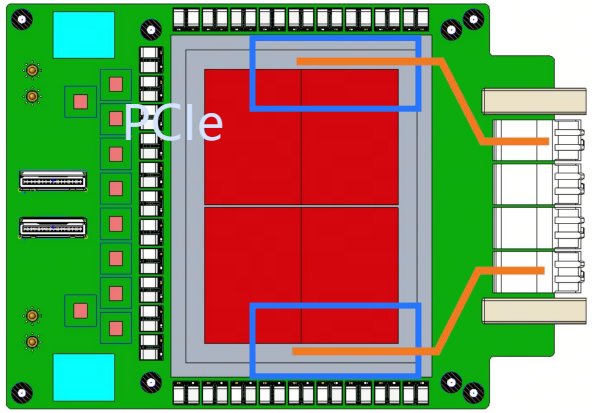
下一代XPU模组的变化设想

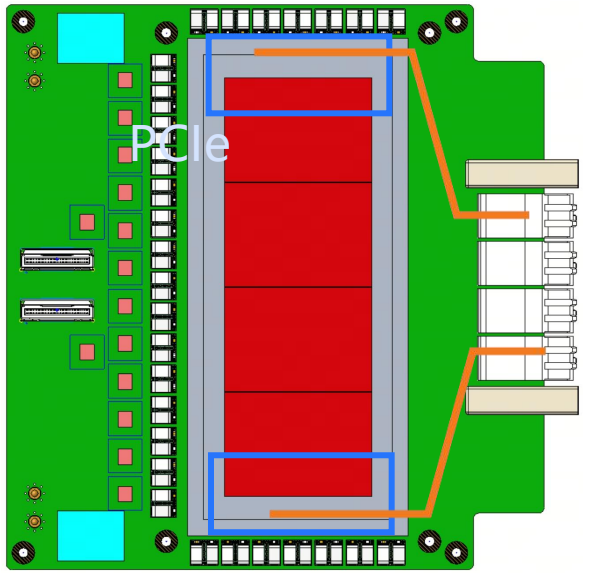
分类	示意图	超节点高速serdes扇出示意	趋势判断
OAM 2.0形态： Mezz扣板			1. 在UBB上，可良好支持传统Server的8卡方案，**在下一代XPU卡形态仍会继续**； 2. 在超节点方案中，该形态最大影响为SI，尤其是后Mezz连接器，增加了~4.5inch和2个via，对SI挑战较大；
自定义模组： 直出NPO光			1. 支持光/电封装兼容，灵活支持光或电连接； 2. 基于NPO方案，易通过光互连单层互连至512卡/1024卡； 3. 基于NPC方案，SI优于扣板形态；  
自定义模组： 直出NPC cable			
自定义模组： 直出BP Conn			1. 该形态铜互连的超节点，具备成本优势； 2. 在免retimer设计同时，可以降低UBB的成本； 3. 对于高速conn，需考虑具备P2P的B配套方案；

下一代XPU模组的形态探讨



单die、双die XPU模组





多die XPU模组

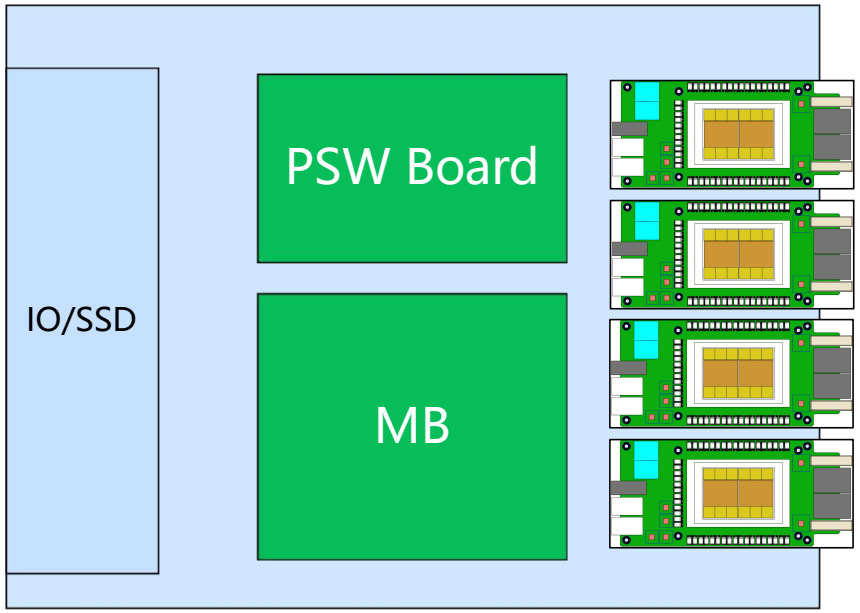
**XPU芯片封装可能大于102mm*

倡议：基于SI最优考虑，XPU芯片的serdes扇出：

- 南向扇出：scale up端口，（或从XPU芯片两侧扇出）
- 北向扇出：PCIe、scale out端口

基于下一代XPU模组的SVR设计设想

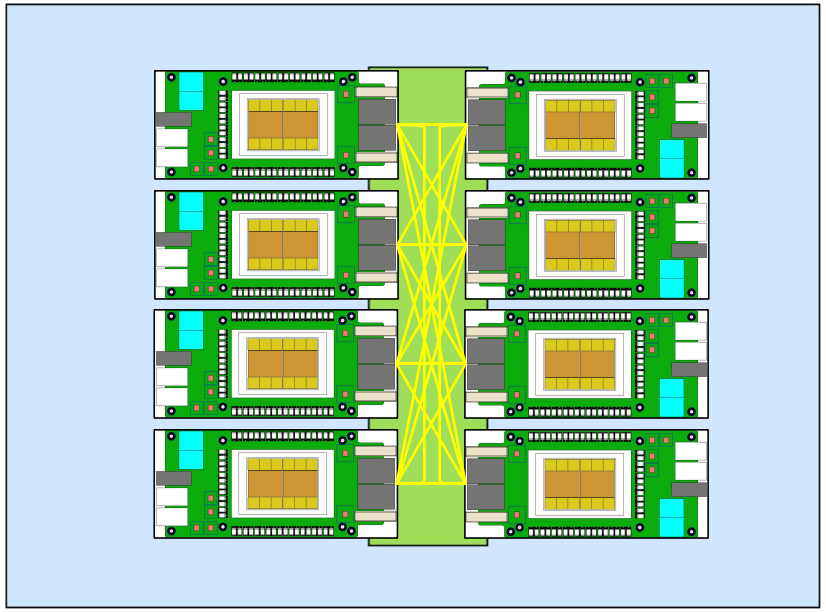
超节点计算Tray



俯视图

Scale up端口直出到背板

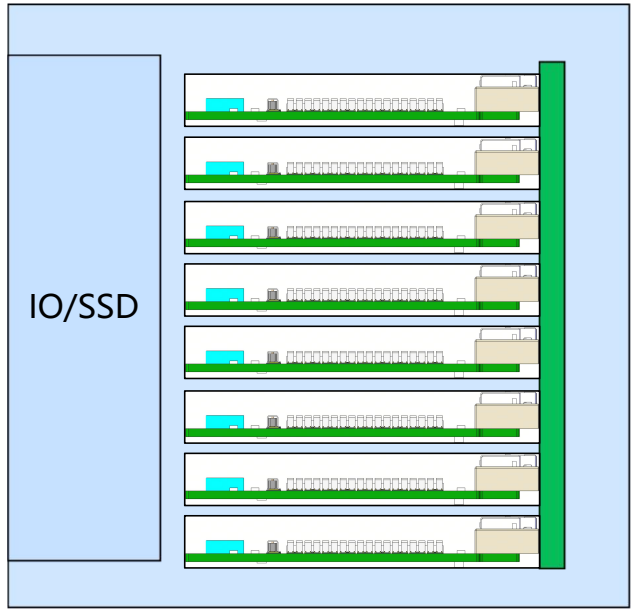
传统8卡服务器



俯视图

Fullmesh互联场景：采用一张小的转接板连接8个XPU模组

8卡/16卡服务器



俯视图

采用中置BP，将8-16张XPU模组连接至SW板

目录

1、下一代XPU模组的挑战及设想

2、下一代XPU模组设计的方案探讨

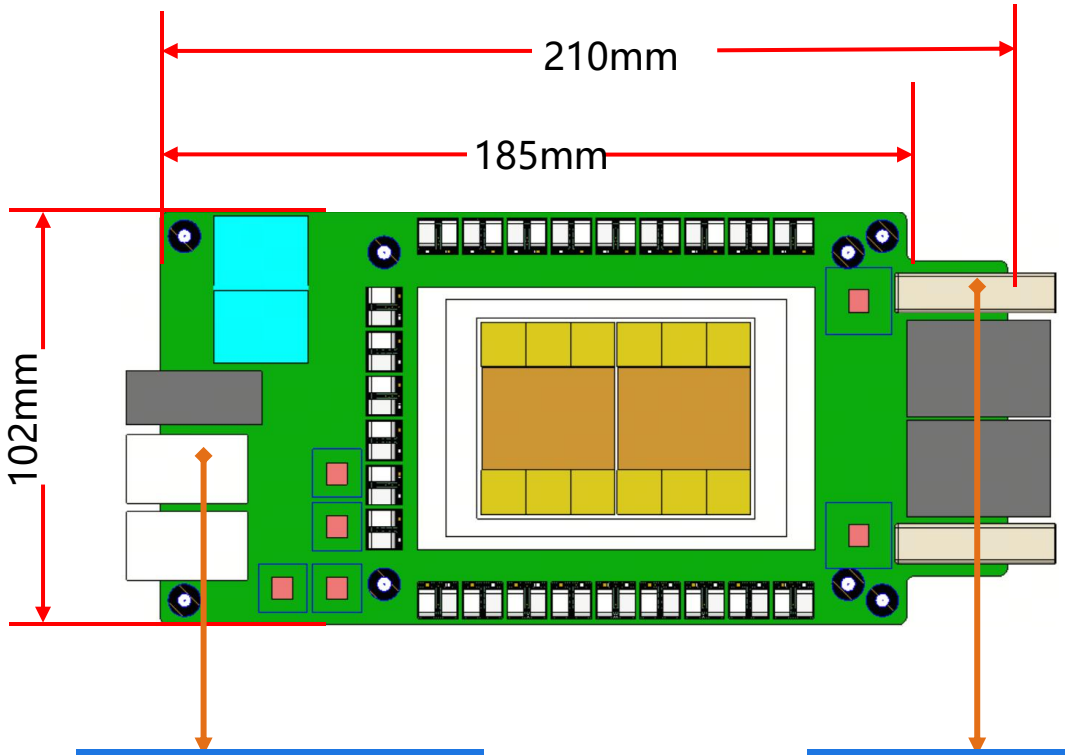
3、基于XPU模组的超节点计算tray设计方案

4、单层光互联超节点设计探讨

下一代XPU模组设计的方案探讨

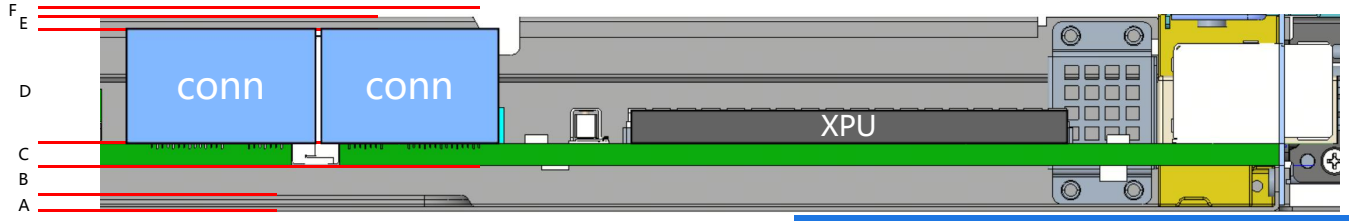
规格设定：

XPU芯片	设定 78mm x 100mm（双die）
模组尺寸	210mm x 102mm
模组功率	不低于1400W，44V-59.5V DC input
南向接口	Up to 64 Lanes per module(112G/224G serdes)
南向conn	支持RAR连接器，规格8pair x n或4pair x n
北向接口	1 x 16 host link (PCIe Gen5/Gen6)



下一代XPU模组设计—北向接口

水平BTB连接 (供电+信号):

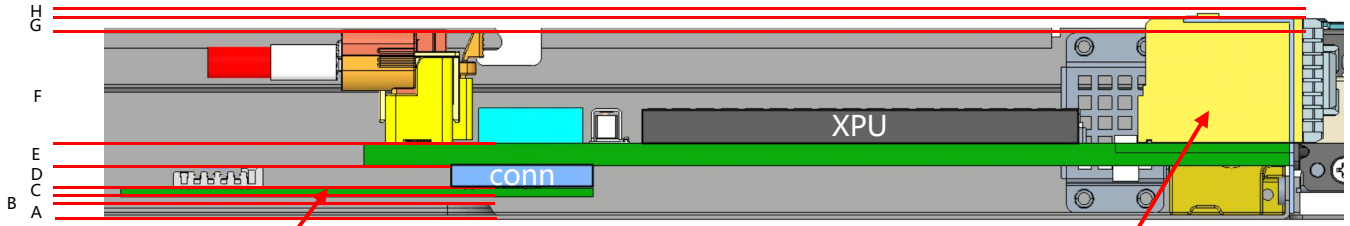


高度挑战小，为冷板设计和连接器高度保留余量

Linear Dimensions		
A	机箱底部到凸包顶面高度	2.50
B	凸包顶面到XPU板底面的间隙	9.50
C	XPU板厚度	5.00
D	**水平BTB连接器**高度	24.55
E	水平对插连接器到机箱顶部间隙	3.00
F	机箱顶盖厚度	1.00

③目标值结果	极限法	Nominal	Min.	Max.
	Worst Case	45.55	44.30	46.80
	统计法	**Mean**	**Min.(Z LSL=4)**	**Max.(Z USL=4)**
	RSS	45.55	44.94	46.16
	Monte Carlo	45.55	44.93	46.16

扣板连接 (供电+信号):



转接板厚度建议<2mm

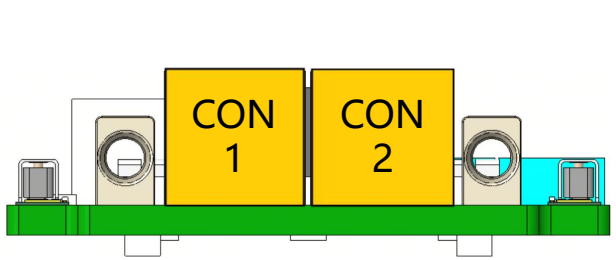
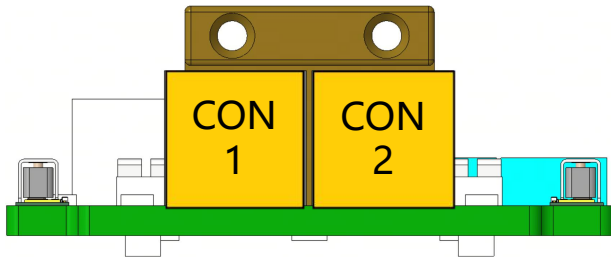
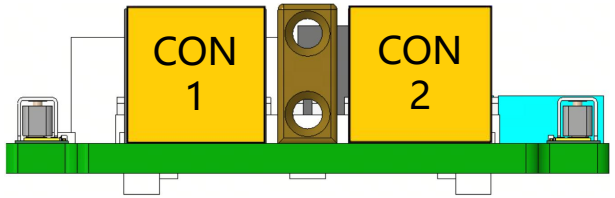
南向连接器高度超过20mm，会与后窗会干涉（高度方向）

Linear Dimensions		
A	机箱底部到凸包顶面高度	2.50
B	凸包顶面到MM转接板底面	3.50
C	扣板连接器转接板厚度	2.00
D	**扣板连接器连接器**互配高度	5.00
E	XPU板厚度	5.00
F	北向连接器高度	20.00
G	北向连接器顶部到后窗顶部高度	6.00
H	后窗顶部到机箱顶盖高度	1.50

③目标值结果	极限法	Nominal	Min.	Max.
	Worst Case	45.50	44.30	46.70
	统计法	**Mean**	**Min.(Z LSL=4)**	**Max.(Z USL=4)**
	RSS	45.50	45.15	45.85
	Monte Carlo	45.50	44.75	46.25

下一代XPU模组设计—导向设计

导套推荐：

左右侧导套	顶部导套	中部导套
		
优势： 导向能力最优 **劣势：** 导套占用宽度较多，挤占了连接器布局空间	**优势：** 导套不占用布局宽度； **劣势：** 占用TOP面高度，影响连接器高度，优先考虑4pair/col的conn	**优势：** 占用宽度较少50%； **劣势：** 导向能力稍弱，系统设计加强；

下一代XPU模组设计—供电设计

■ OAM模组电源解决方案：

① 54V/48V转12V（only Typical） + 12V转Core Power的电源架构。

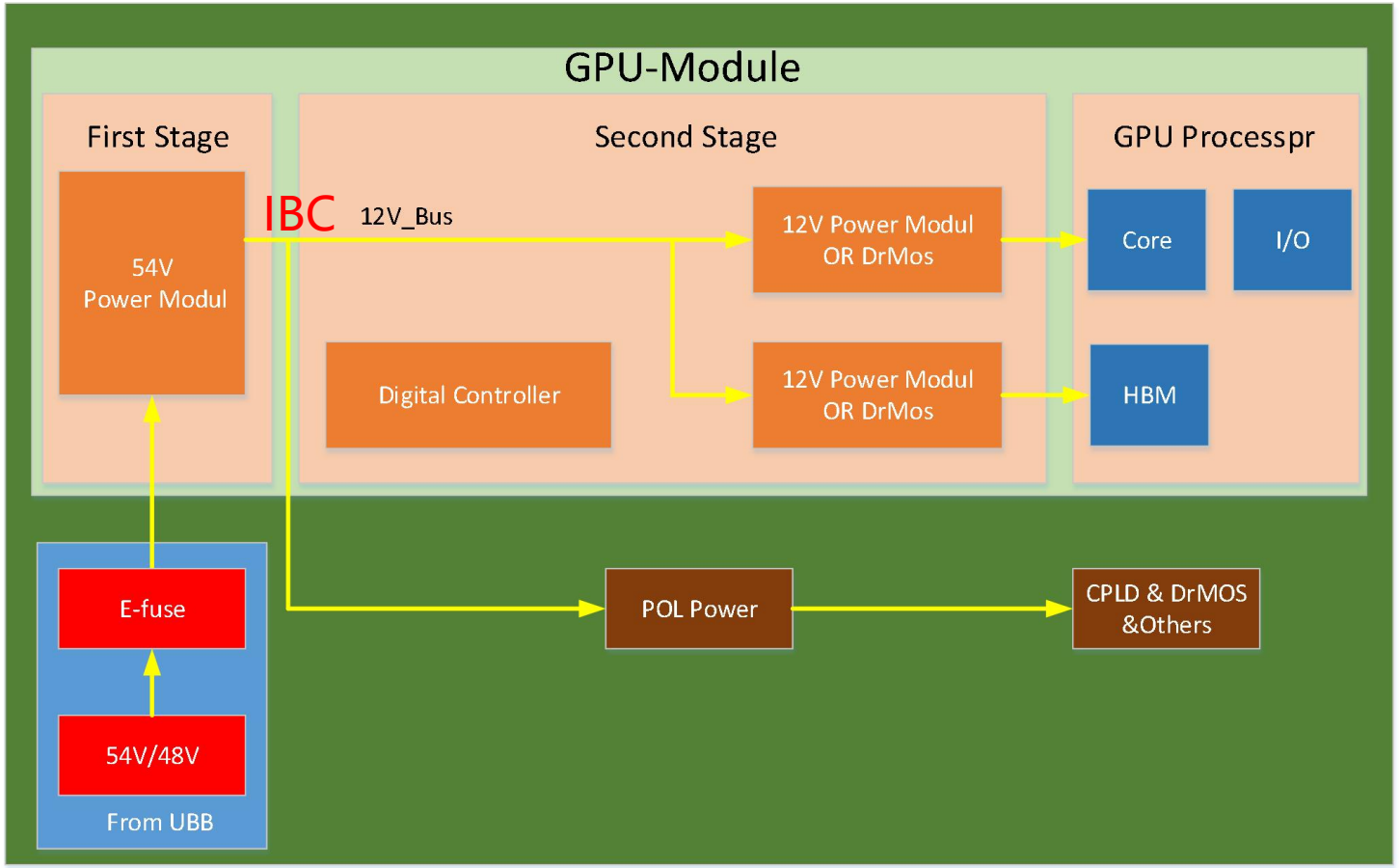
② 54V（40-60V）供电可采用4:1的转换比，输出电压10-15V（典型值12V）转换效率高，可大幅减小IBC中间转换电流。

③ 54V（40-60V）供电也可采用8:1的转换比，输出电压典型值6.75V，会翻倍增加IBC中间转换电流，PCB上的热会以I^2的比例上升，一般需要定制低压VRM，兼容性差，但也可以提高VRM级转换效率。

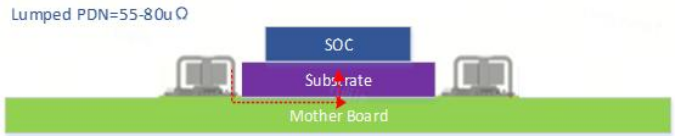
④ 采用水平供电（LDP），目前电源各级转换方案成熟可兼容Second supplier。

⑤ 可根据OAM板卡实际布局情况选择分立器件或Power module 灵活兼容成本优势与PCB布局。

⑥ 水平供电痛点：XPU功耗越来越大导致PCB损耗越来越大，压降、PDN很难满足电源设计要求等。



• OAM电源架构框图



• 水平供电

• 垂直供电

下一代XPU模组设计—供电设计

■ VPD供电（垂直供电）解决方案：

➢ VPD垂直供电可以有效优化供电路径和PDN。

➢ **垂直供电痛点：**

① VRM的垂直布局会占用 ASIC底部原去耦电容空间，若电容设计不满足需求，需设计Interposer埋MLCC解决去耦电容问题（图4）。

② VRM自生重力原因带来的SMT生产工艺问题。

③ 自定义出线方式（图5，也可增加interposer board）底部拉开了垂直方向到机壳的距离，有利于垂直供电设计。

④ 大幅提高了PCB Pin-Pin的设计难度。

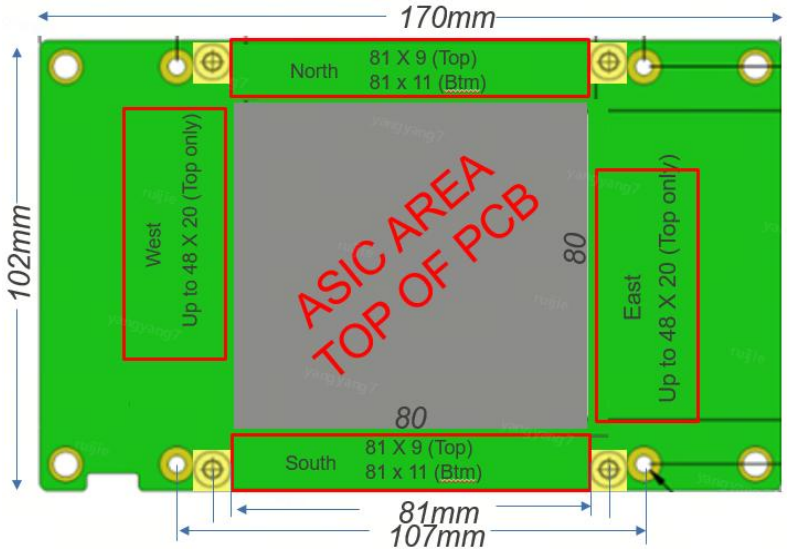


图1

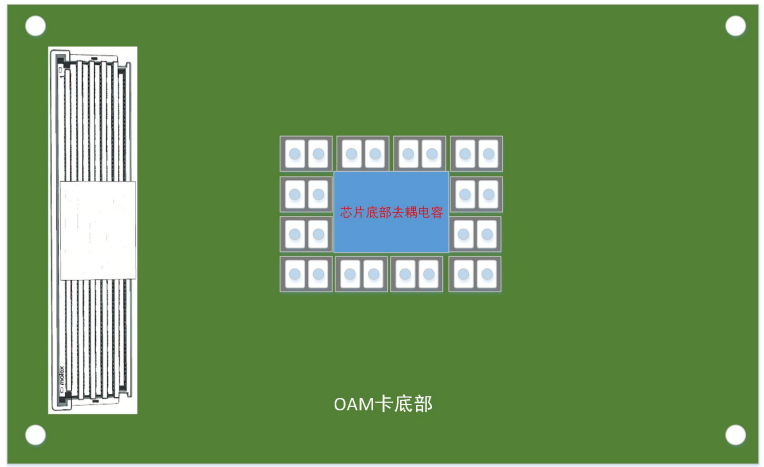


图2：VPD供电形态之一示意图

图3：MEZZ连接器VPD空间有限

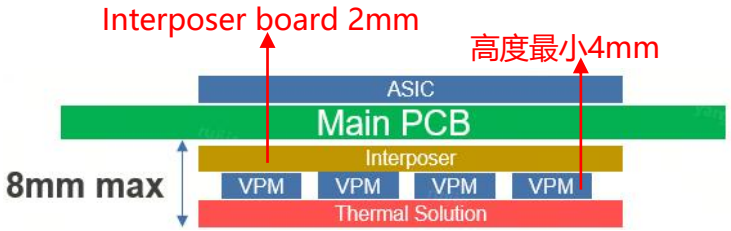


图4：标准OAM卡留给散热设计空间＜2mm

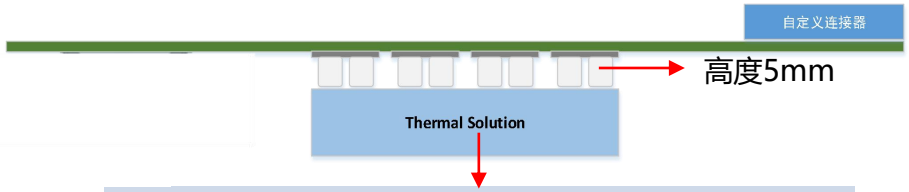


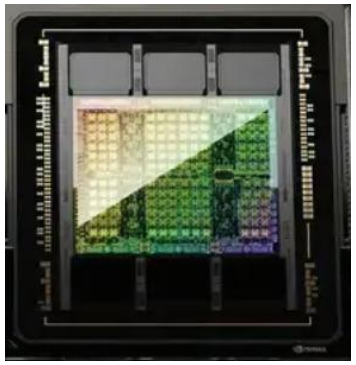
图5 底部空间＞10mm足够的高度用于解决电源散热

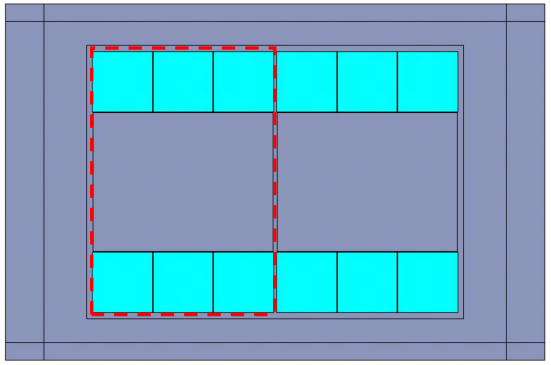
- 标准OAM卡留给电源热设计空间有限

- 自定义出线设计有足够的电源热设计空间

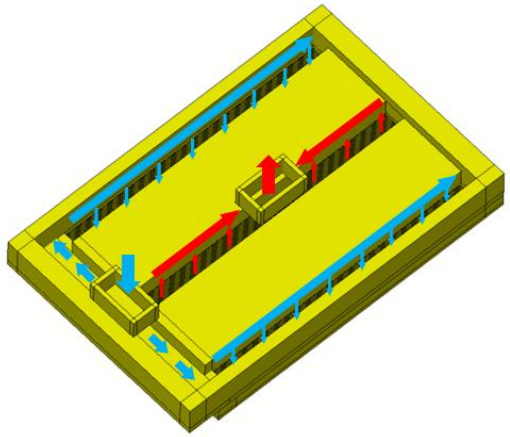
下一代XPU模组设计—散热设计

■ 针对高功耗 XPU 模组的散热需求，经仿真评估散热系统性能满足设计规格要求，为高算力场景下的设备稳定运行提供核心保障。

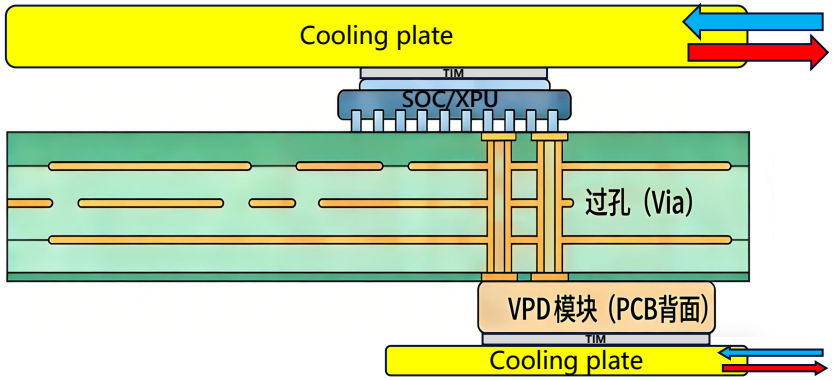




芯片模型预估



冷板流道示意图

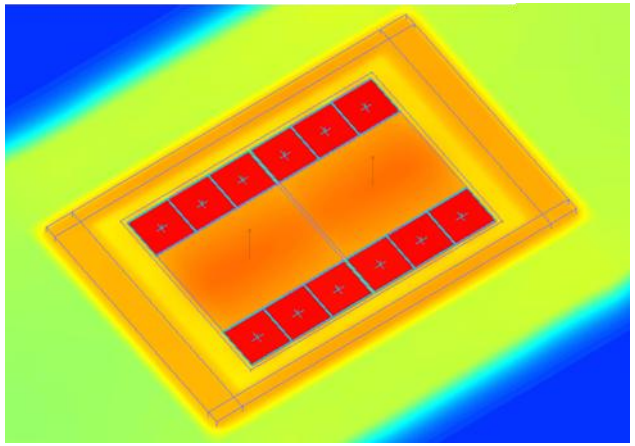


VPD液冷散热示意

XPU模组散热仿真

工况	1600W 40℃ 3.0LPM			2000W 40℃ 4.5LPM		
器件名称	功耗/W	规格/℃	温度/℃	功耗/W	规格/℃	温度/℃
ASIC1	563.6	105	*76.6*	704.4	105	*80.8*
HBM	39.4	**95**	*90.8*	49.25	**105**	*101*
ASIC2	563.6	105	*76.7*	704.4	105	*81.8*
HBM	39.4	**95**	*90.7*	49.25	**105**	*101*

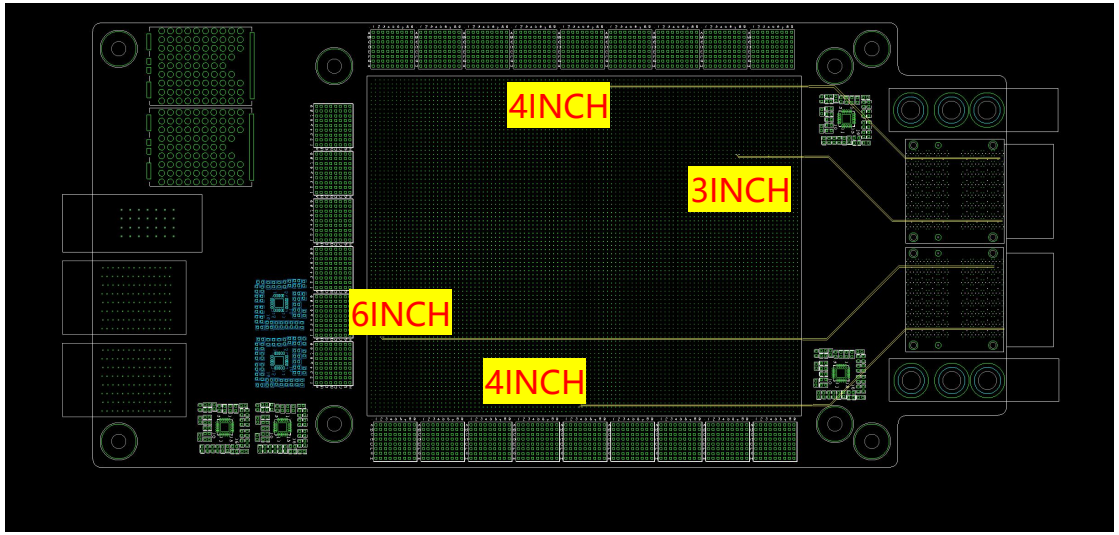
温度云图



- 40℃进液、3.0LPM工况下，HBM 95℃规格，1600W满足温度要求。4.5LPM工况下，HBM 105℃规格，2000W临界满足。
- VPD模组散热方案采用冷板背贴一体化结构设计，考虑单独走水，隔离热源、提升散热效率，保障模组长期稳定工作。

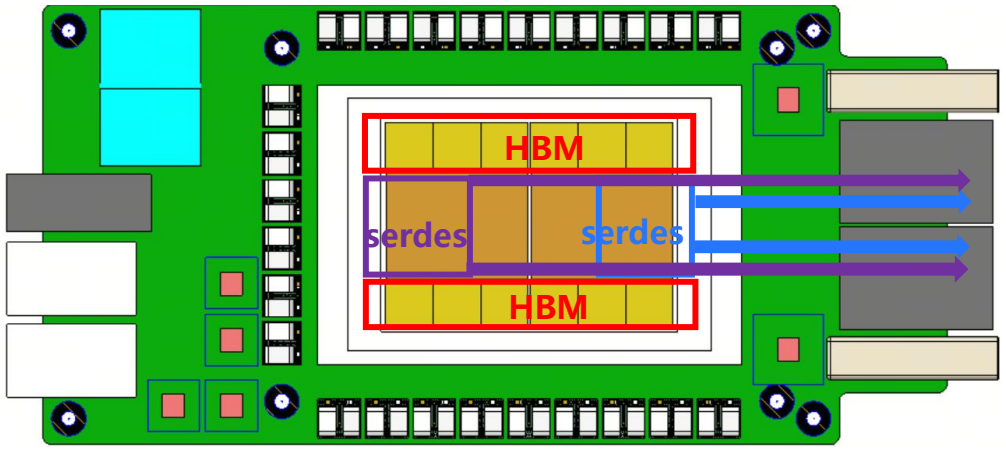
**芯片尺寸、功耗分布、温度规格等均基于已有信息进行假设分析；推测功耗≥2000W时，HBM规格调整为105℃*

下一代XPU模组设计—互连设计

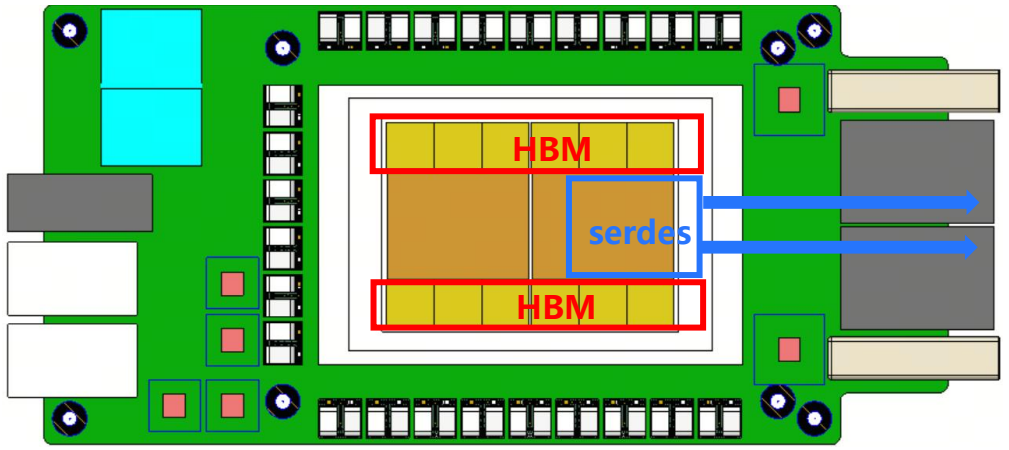


- 两侧出线到连接器预估走线长度6inch
- 单侧出线到连接器预估走线长度3inch
- 上下两侧出线到连接器预估走线长度4inch
- **建议**：XPU模组上serdes走线长度不超过3inch

XPU模组serdes两侧出线



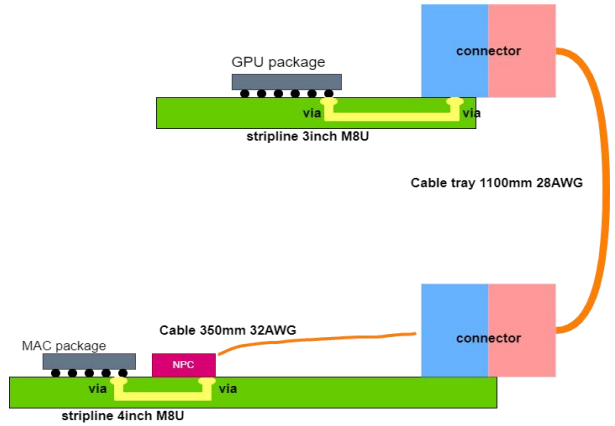
XPU模组serdes单侧出线



下一代XPU模组设计—超节点场景互连考虑

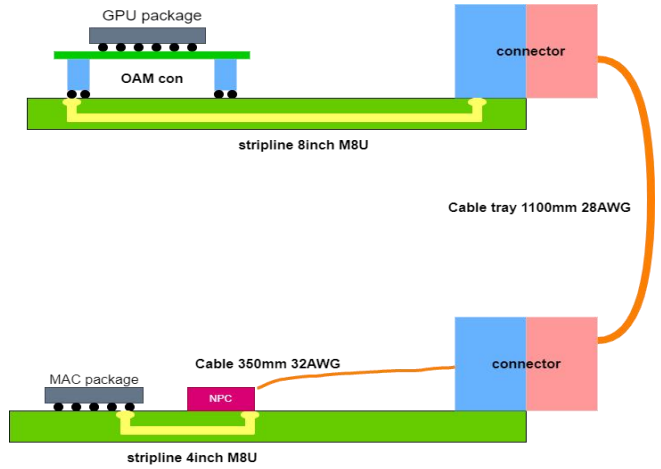
下一代XPU模组112G单侧出线

112G 不加retimer 单侧出线	MAC board (M8U 4inch+NPC350mm 32AWG+con)	Cabletray (28AWG 1100mm)	User defined XPU Board (M8U 3inch+con)
Package Loss(db)			
Routing Trace(db)	2.1		2.55
Routing Trace Neck mode(db)	1		
Via(db)	1		1
Capacitor(db)			
Cable(db)	3.08	7.15	
Connector(db)	1.2		1.23
Channel Total Insertion Loss @26.56GHz (db)	20.31		
Channel Total Insertion Loss (High Temp)@26.56GHz (db)	22.09		
Channael Total Insertion Loss spec @26.56GHz(ball to ball) (db) 112G PAM4 LR	28		



112G OAM2.0模组

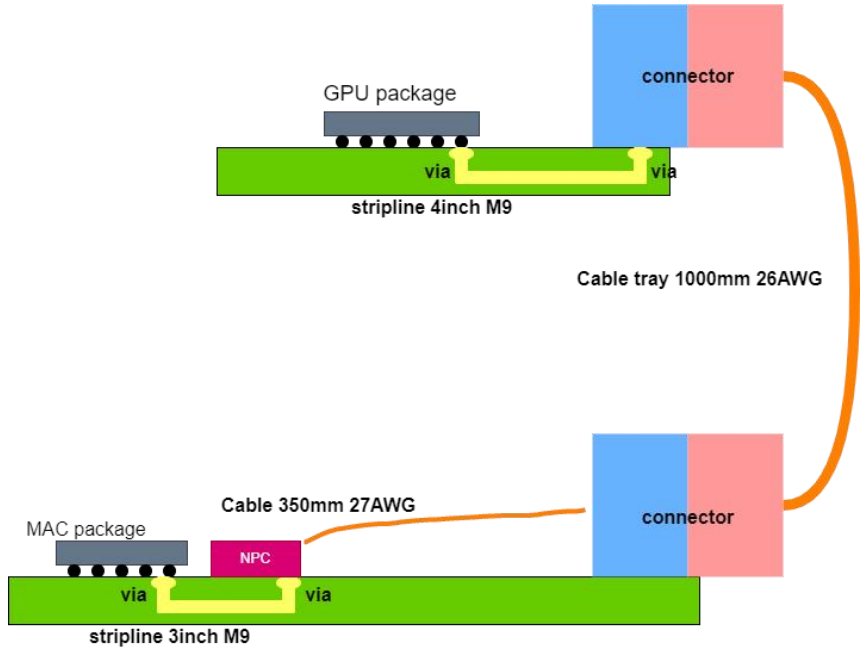
112G OAM2.0	MAC board (M8U 4inch+NPC350mm 32AWG+con)	Cabletray (28AWG 1100mm)	UBB (M8U 8inch+con)	OAM (M8U 3inch+MEZZ)
Package Loss(db)				
Routing Trace(db)	2.1		5.6	2.55
Routing Trace Neck mode(db)	1			
Via(db)	1		1	1
Capacitor(db)				
Cable(db)	3.08	7.15		
Connector(db)	1.2		1.23	1
Channel Total Insertion Loss @26.56GHz (db)	27.91			
Channel Total Insertion Loss (High Temp)@26.56GHz (db)	30.14			
Channael Total Insertion Loss spec @26.56GHz(ball to ball) (db) 112G PAM4 LR	28			



下一代XPU模组设计—超节点场景互连考虑

XPU模组224G不加retimer

224G不加retimer单侧出线	MAC Board (M9 3inch+NPC 350mm 27AWG+Con)	Cabletray (26AWG 1000mm)	User defined XPU Board (M9 4inch+Con)
Package Loss(db)	6.5		4
Routing Trace(db)	2		4.8
Routing Trace Neck mode(db)	1.45		
Via(db)	1.6		1.6
Capacitor(db)			
Cable(db)	3.22	8.1	
Connector(db)	1.9		1.9
Channel Total Insertion Loss @53.12GHz (db)	**37.07**		
Channel Total Insertion Loss (High Temp)@53.12GHz (db)	**39.027**		
Channael Total Insertion Loss @53.12GHz(die to die) (db) 224G PAM4 LR	40		



- 长通道需要采用更粗线径(优先26AWG)降低cabletray插损或缩短cabletray长度
- 如果XPU无法支持die to die -40db@53.12GHz,需要在交换节点内部或计算节点内部增加224G retimer,建议优先在计算节点内部增加224G retimer
- XPU芯片PINMAP需要将serdes引脚分配到靠近连接器侧，单侧出线优先（走线长度3inch以内）
- XPU芯片基板面积尽量缩小，降低package插损
- XPU侧连接器优先选用4pair，能降低连接器插损

目录

1、下一代XPU模组的挑战及设想

2、下一代XPU模组设计的方案探讨

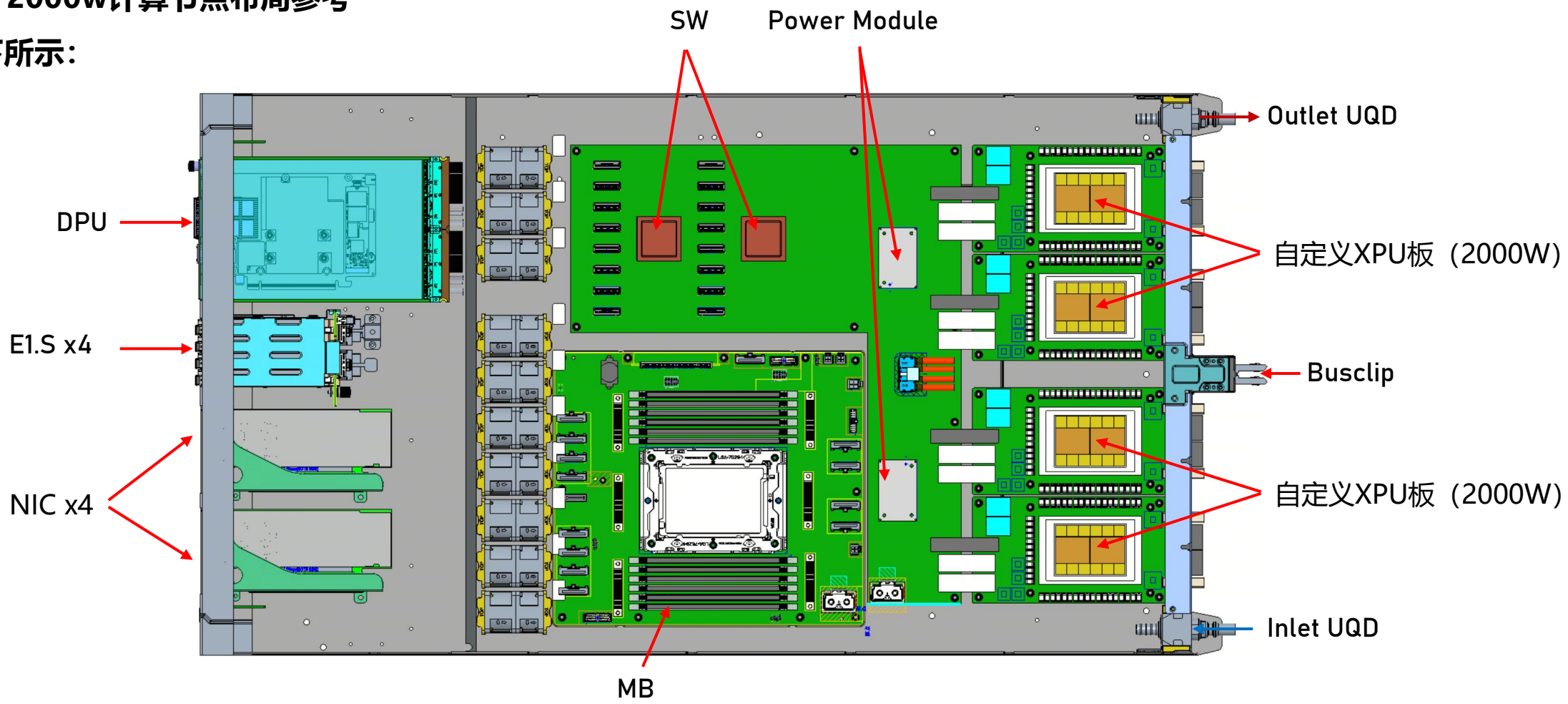
3、基于XPU模组的超节点计算tray设计方案

4、单层光互联超节点设计探讨

基于下一代XPU模组的计算tray设计方案

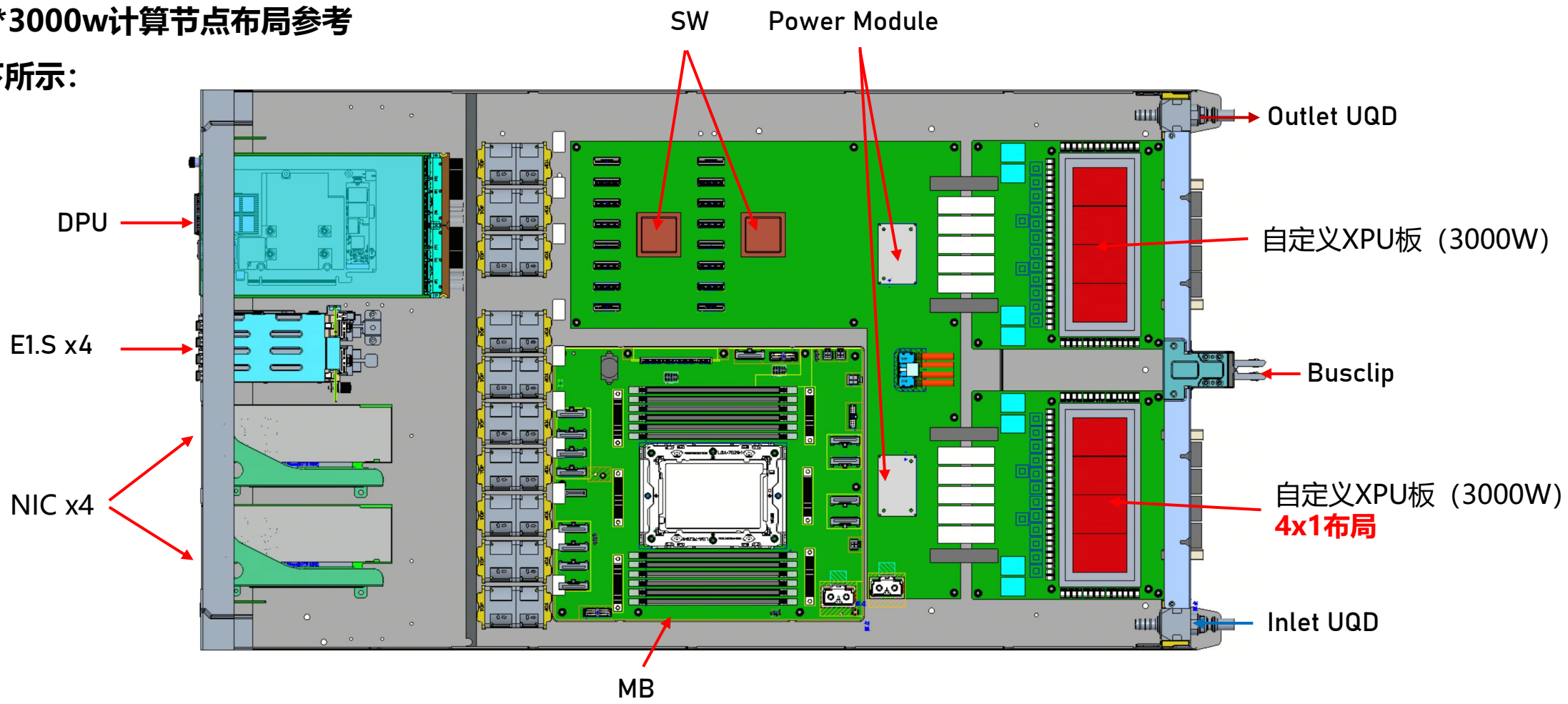
4卡*2000w计算节点布局参考

如下所示：



基于下一代XPU模组的计算tray设计方案

2卡*3000w计算节点布局参考如下所示：



目录

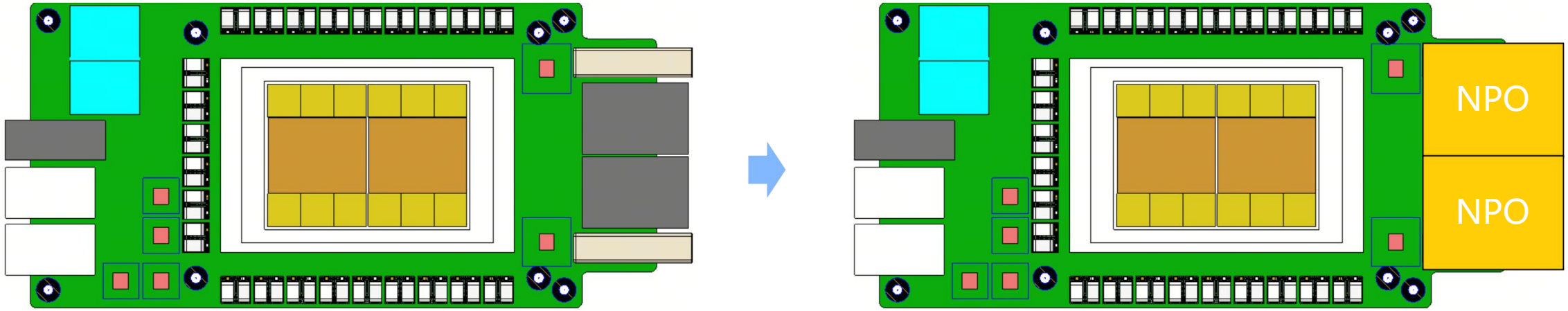
1、下一代XPU模组的挑战及设想

2、下一代XPU模组设计的方案探讨

3、基于XPU模组的超节点计算tray设计方案

4、单层光互联超节点设计探讨

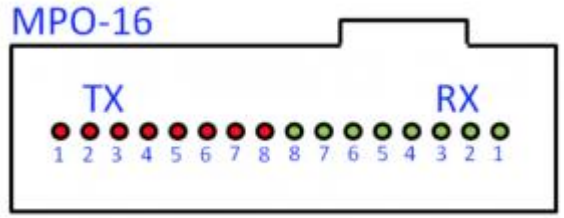
下一代XPU模组设计—NPO方案



基于NPO单级光互联方案探讨

光交叉(32x200G NPO规格，XPU假设为x1每port)

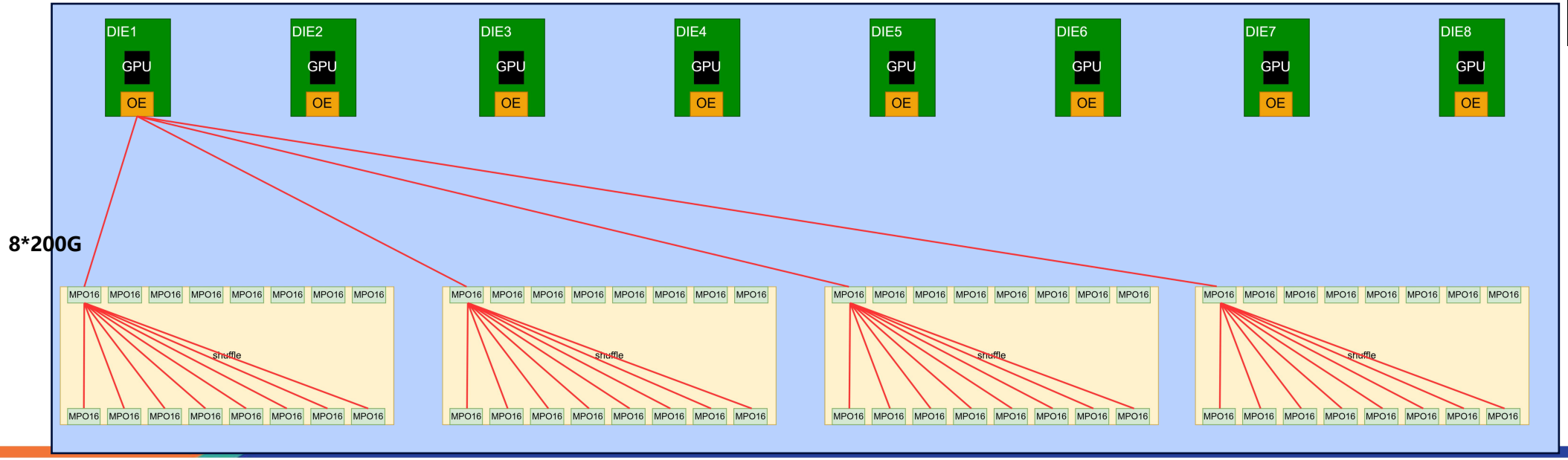
- 每XPU模组1个OE，每个OE(32通道X200G) **6.4Tbps**;
- 每个OE包含4xMPO16+1xMPO12(for ELSFP);
- 前窗每个MPO16的信号分别来自于**8**个XPU（**8*200G**）；



MPO-16 定义

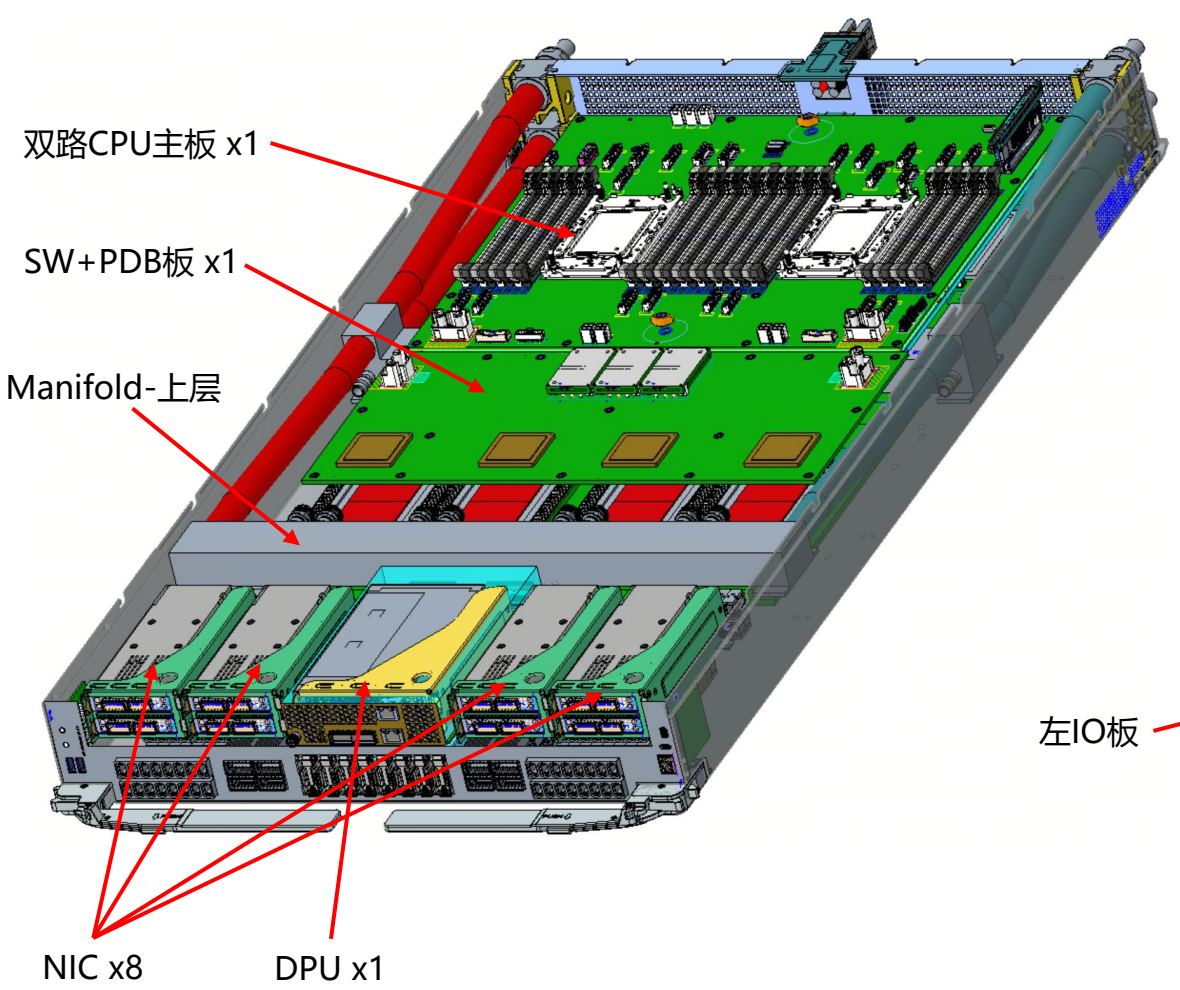
XPU1	1
XPU2	2
XPU3	3
XPU4	4
XPU5	5
XPU6	6
XPU7	7
XPU8	8

以一个XPU为例

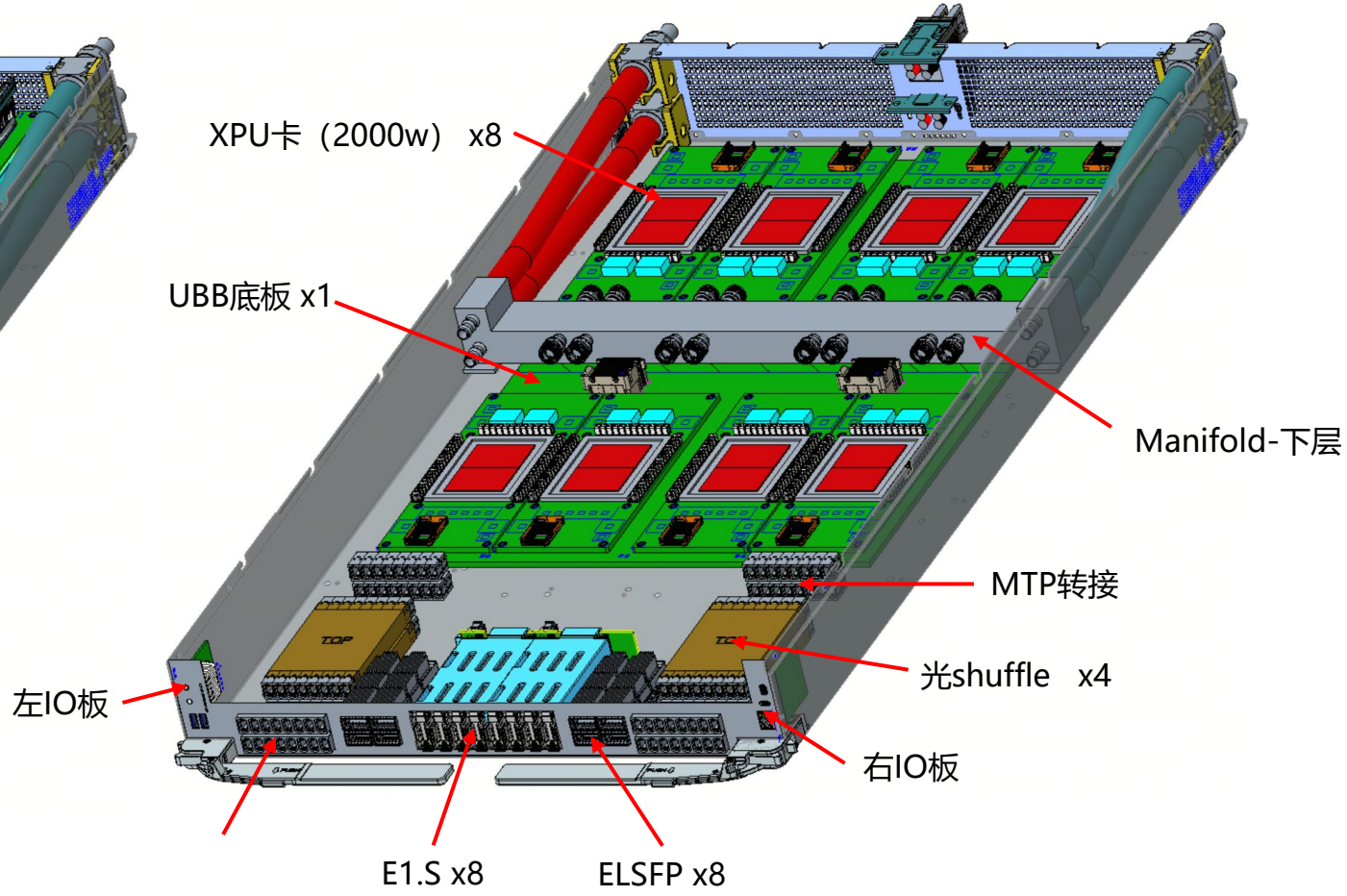


基于NPO单级光互联方案探讨—SVR布局参考

计算节点结构布局--（8卡-2000W）



上层布局

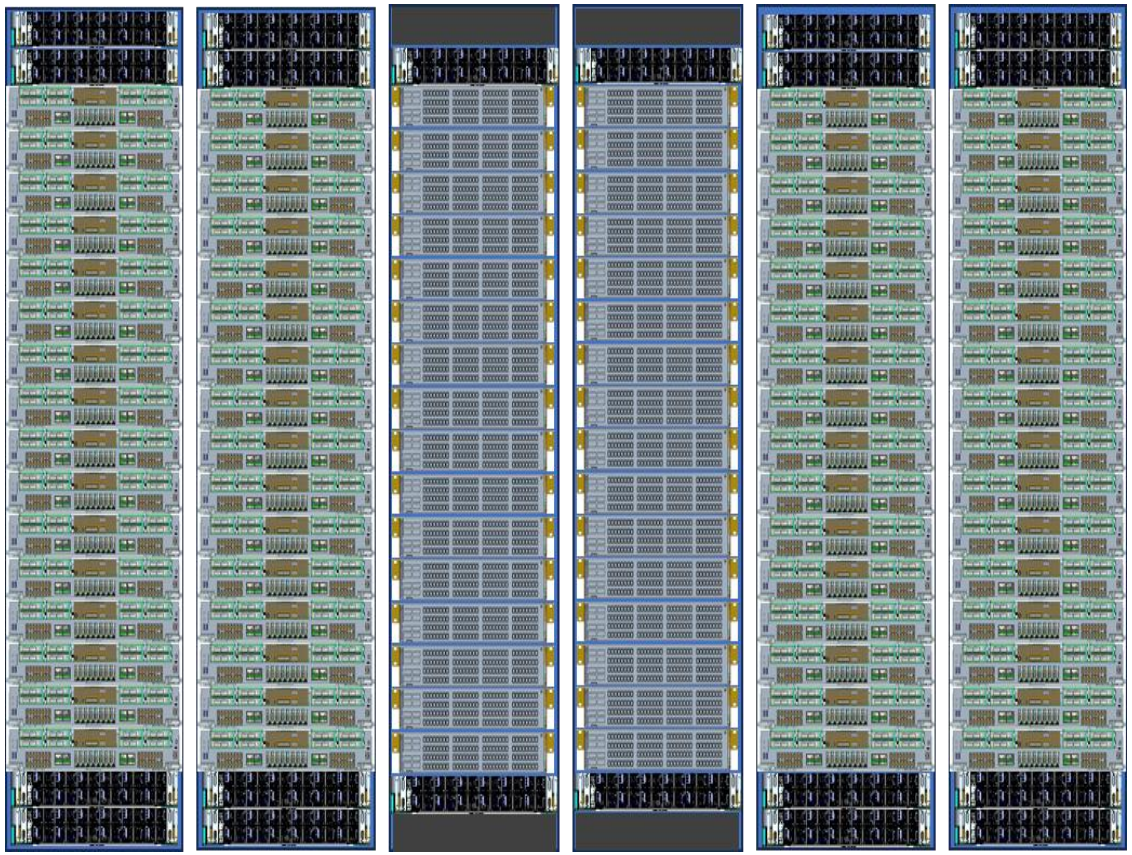


下层布局

基于NPO单级光互联方案探讨

规格	参数
机电平台	采用平台化设计机柜， 支持光电融合AI服务器， 以及NPO交换机
供电	最大支持**4**个Powershelf 支持整机供电>**240KW**
计算柜	最大支持**16**个2U 光电融合AI服务器
交换柜	最大支持**16**个2U NPO交换机
散热	液冷散热，外置CDU
机柜尺寸	1400mm*600mm*2300mm（**内高44U**）

光互联512卡方案示意图



计算柜　计算柜　交换柜　计算柜　计算柜

Open AI Infra

Thank You !

Global Computing Consortium
全球计算联盟

Open AI Infra

Thank You !

Global Computing Consortium
全球计算联盟

Open AI Infra

Thank You !

Global Computing Consortium
全球计算联盟