

OPEN UXM项目组规划分享

武正辉

Open AI Infra社区OPEN UXM项目组组长
百度智能云技术专家

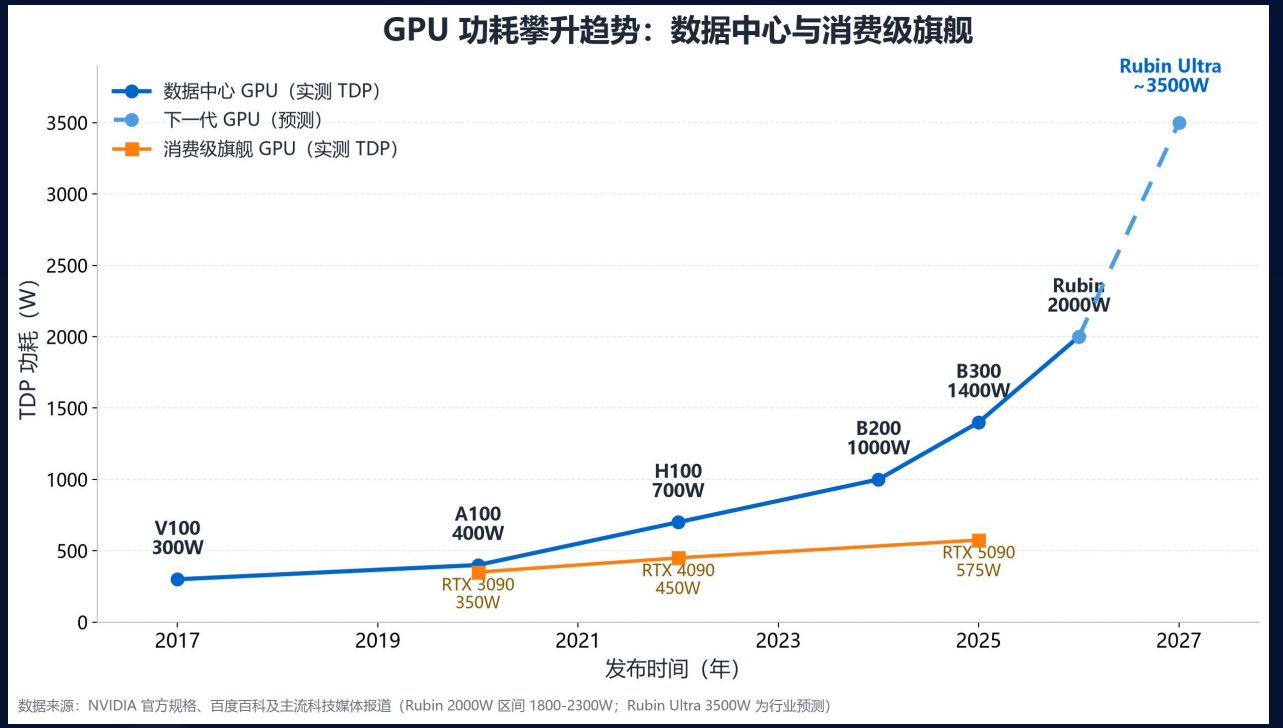
01

国内算力模组标准化的迫切需求

GPU 功耗未来趋势

Open AI Infra
A Chapter of GCC

GPU TDP 演变（2017–2027，单位 W）



H100 • 700W 2022
上代 AI 训练主力

B200 1000W+ 2024
NVL72 整机柜 120kW

Rubin 1800–2300W 2026
下一代架构，功耗近 3 倍于 H100

RTX 5090 • 575W 2025
消费级旗舰也在攀升

6 年接近 9 倍：V100 300W（2017）→ Rubin Ultra ~3500W（2027 预测）；模组→计算节点→整机柜供电架构面临重构

XPU芯片现状分析

Open AI Infra
A Chapter of GCC

- 国外GPU生态已基本收敛，国内处于蓬勃发展阶段，目前国内GPU厂家至少超过20+家；
- 亟需可以快速落地的整机柜全生态的标准化方案，加速国产GPU厂家的快速产业化落地。

- 基于国外某GPU单GPU芯片尺寸120*150mm>OAM2.0
- 国产GPU国内制程受限，非先进制程，尺寸较大，功率较大
- 部分国产推理GPU和高端训练GPU共用封装，芯片尺寸~120mm。

国 外

NVIDIA AMD intel

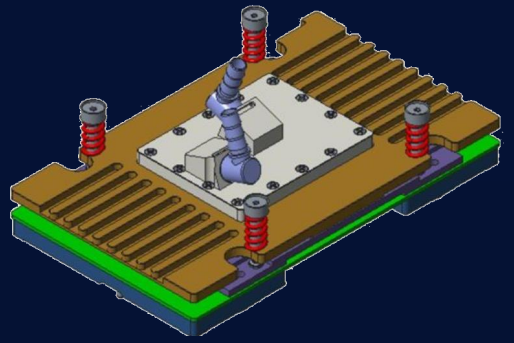
国 内

20+

XPU模组行业现状



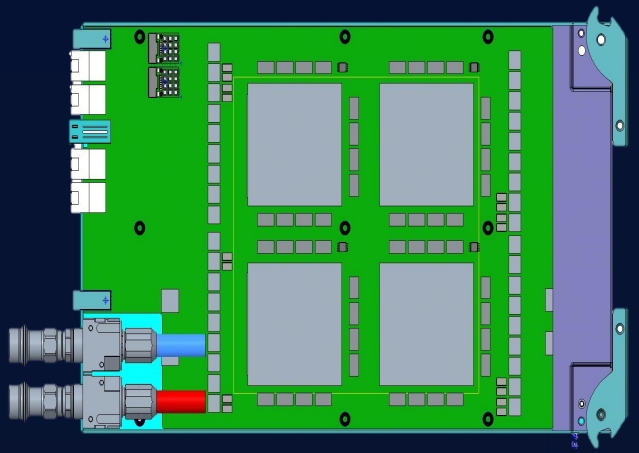
	CEM 限制	升级难度
功率(TBP)	~600W	高
散热	风冷优先 液冷改造复杂	高
PCB 厚度	1.57mm OR 阶梯板	高
PCIE接口	X16 lane	高
高速接口	无	无法升级



	OAM 2.0限制	升级难度
功率(TBP)	~1400W	高
散热	~1400W	高
尺寸	102mm*170mm	高
垂直供电	不支持	高
高速接口	112G/224G	高
最新版本	2023年	高

XPU模组新标准规范适应大模型及芯片发展需求

Open AI Infra
A Chapter of GCC



1. XPU芯片功耗升高OAM ΔIR增大电源完整性凸显
2. XPU芯封装尺寸变化需要新的模组与其匹配
3. 推理大模型需要多卡多机、高带宽、低时延
4. 易安装，易运维，易替换、易运输
5. 标准化、弹性扩展，弹性互联

01

OPEN UXM项目典型需求场景

OPEN UXM 目前3类Type

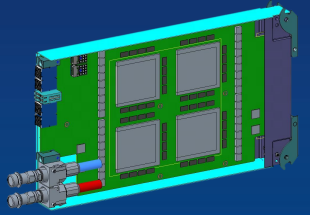
Open AI Infra
A Chapter of GCC

PCIe卡

OAM模组

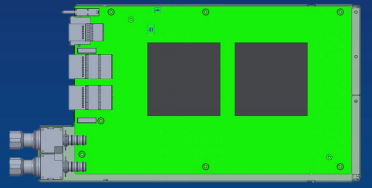
OPEN UXM Type1：

PCIe **原生液冷、高密、易维护**



OPEN UXM Type2：

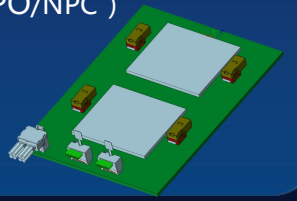
高功耗，大尺寸芯片，免下架维护



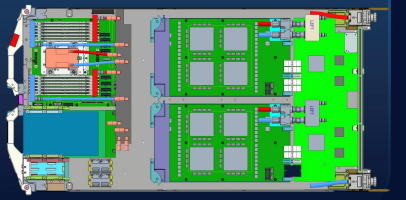
OPEN UXM Type3：

高功耗，大芯片，**高带宽大Radix互连**

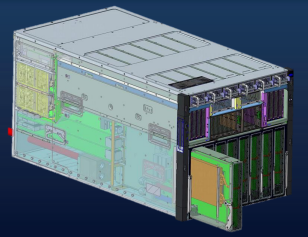
（NPO/NPC）



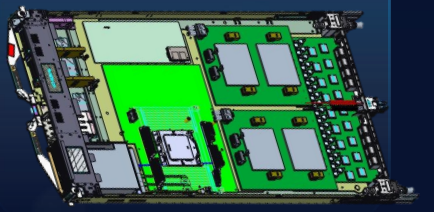
整机柜



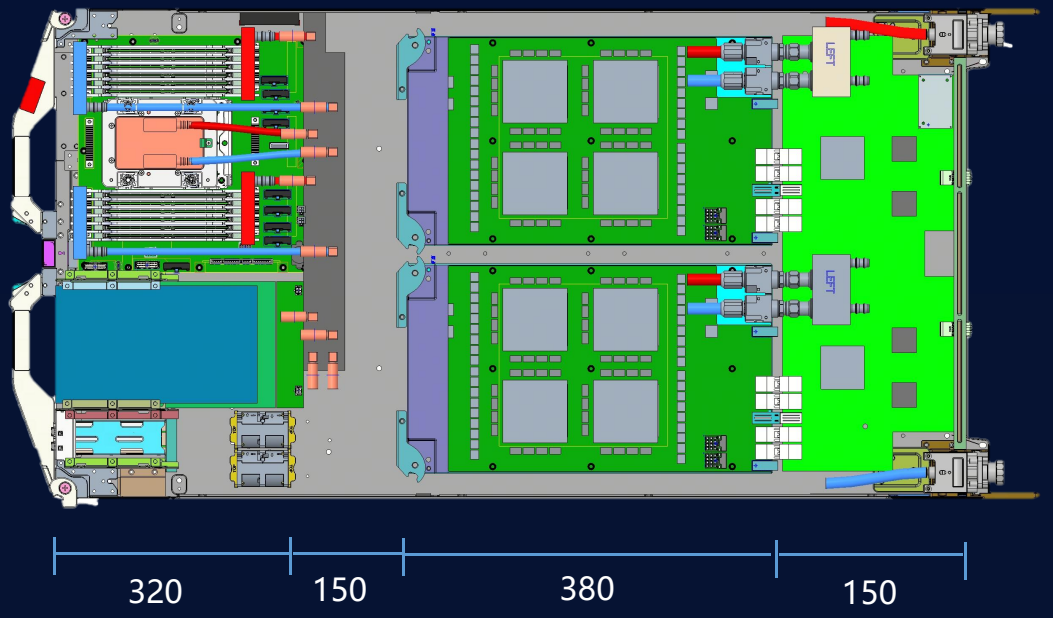
服务器Box



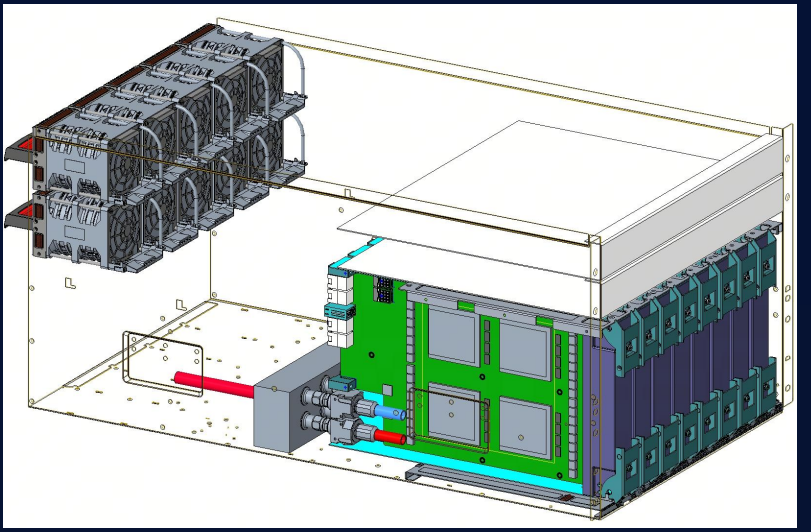
整机柜



OPEN UXM-Type1 场景



- **主力场景1：rack场景，纯pcie 无对外高速scaleup**
- 长度约束：380mm
- 高度约束： 42mm



- 主力场景2：标准机

设计约束：pcie连接器位置同type2保持一致，可以插到标准机

实现8模组16芯/32芯，模组内可选fullmesh架构

OPEN UXM-Type1 规格

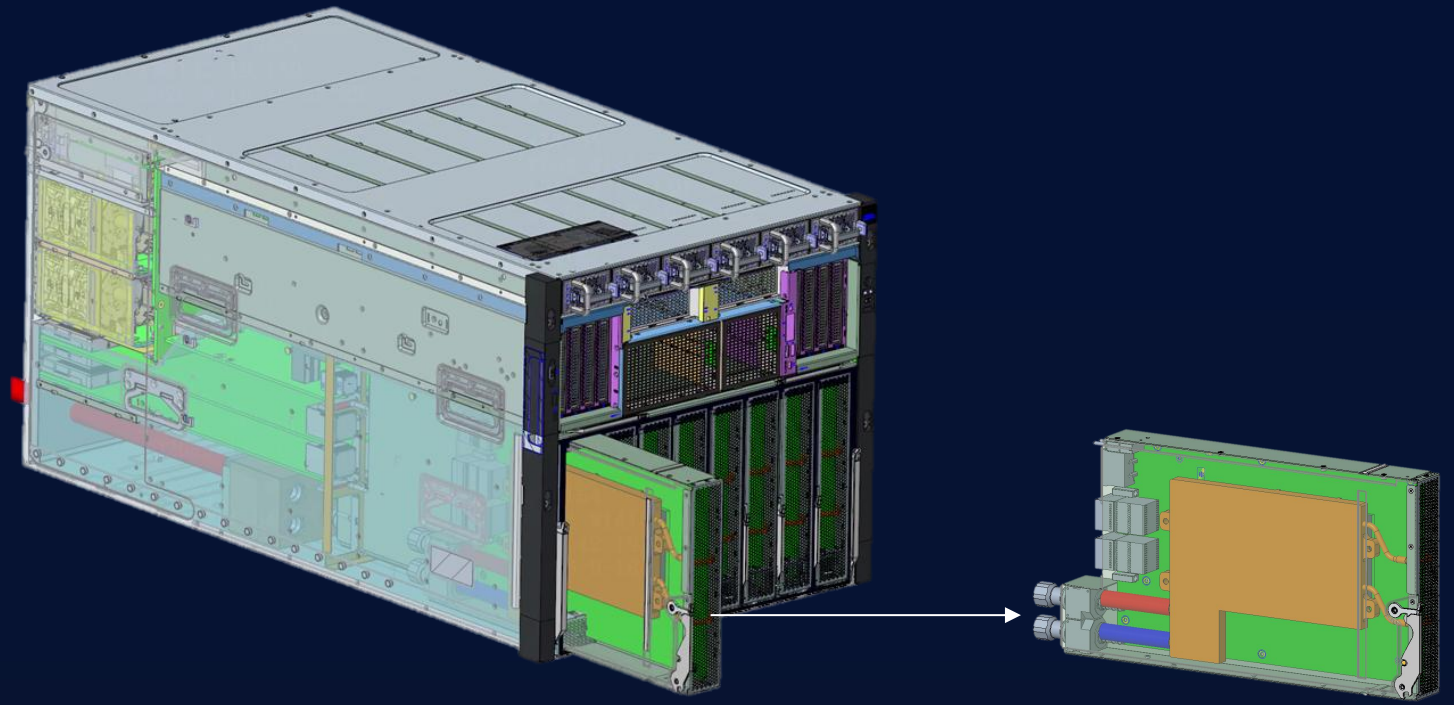
Open AI Infra
A Chapter of GCC

V0.3规格

模 组 尺 寸	230*380mm*42mm（pcb）
芯 片 尺 寸	4*XPU，单XPU 80mm*80mm
GPU最大功率	4*1000W
液冷最大功率	4000W UQDB04
P C I E 接 口	4*Pcie6.0/7.0 X16
供 电 接 口	54V 90A
高 速 接 口	无
维 护	支持盲插



OPEN UXM-Type2 场景

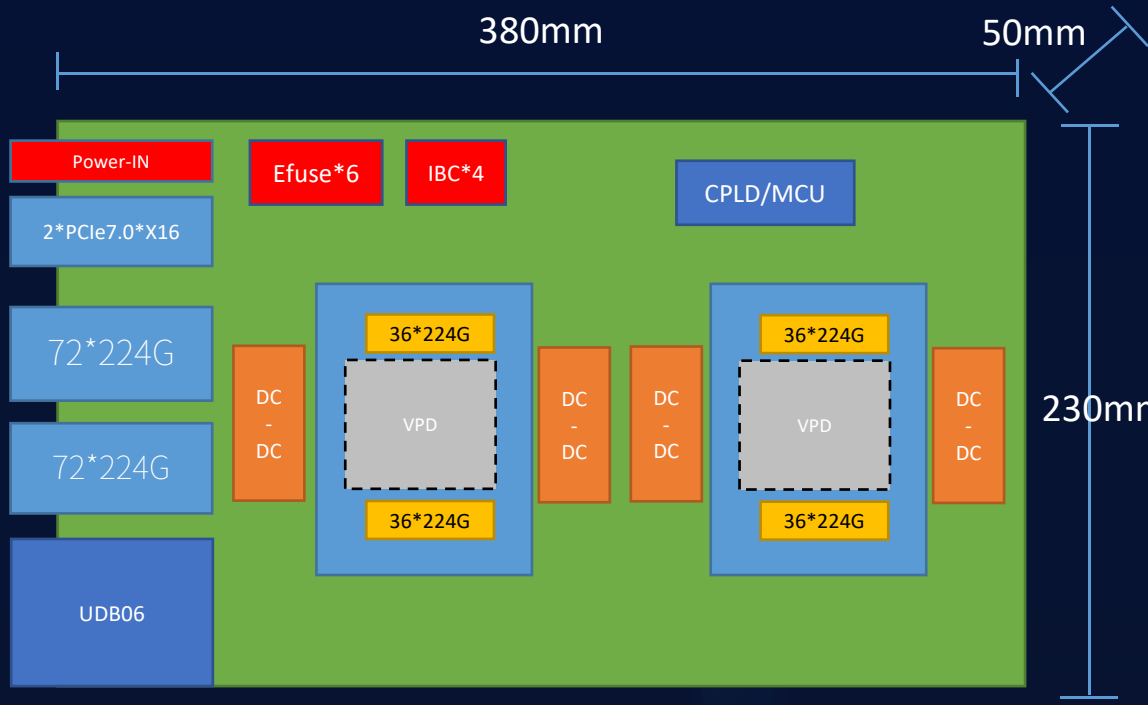


原生UXM服务器，可复用到传统6U XPU Tray

OPEN UXM-Type2 规格

Open AI Infra
A Chapter of GCC

模 组 尺 寸	230*380mm*50mm
芯 片 尺 寸	2*XPU，单XPU 150mm*120mm
GPU最大功率	2*2500W，单卡6000W
液冷最大功率	6000W UQDB06
P C I E 接 口	2*Pcie6.0/7.0 X16
供 电 接 口	54V 6000W
高 速 接 口	2*72*224G
维 护	支持盲插

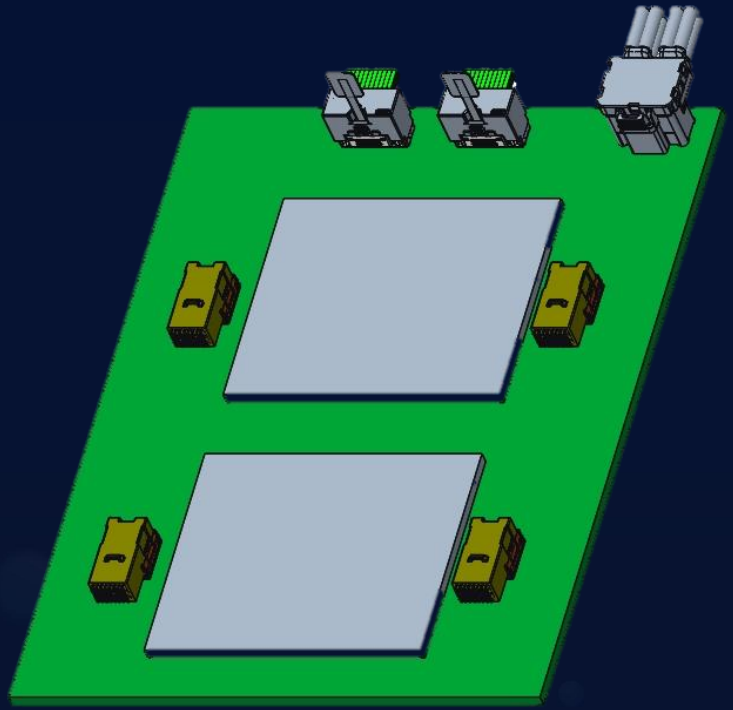


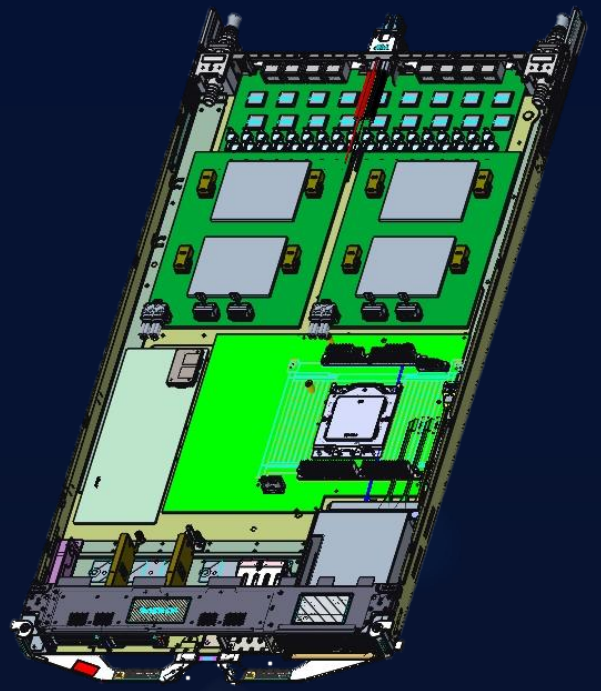
OPEN UXM Type3 场景

Open AI Infra
A Chapter of GCC

Type3 应用场景：整机柜为主，兼容标准机设计；

限制因素：模组厚度受到整机柜计算节点 1OU（46.5mm机箱高度限制）；



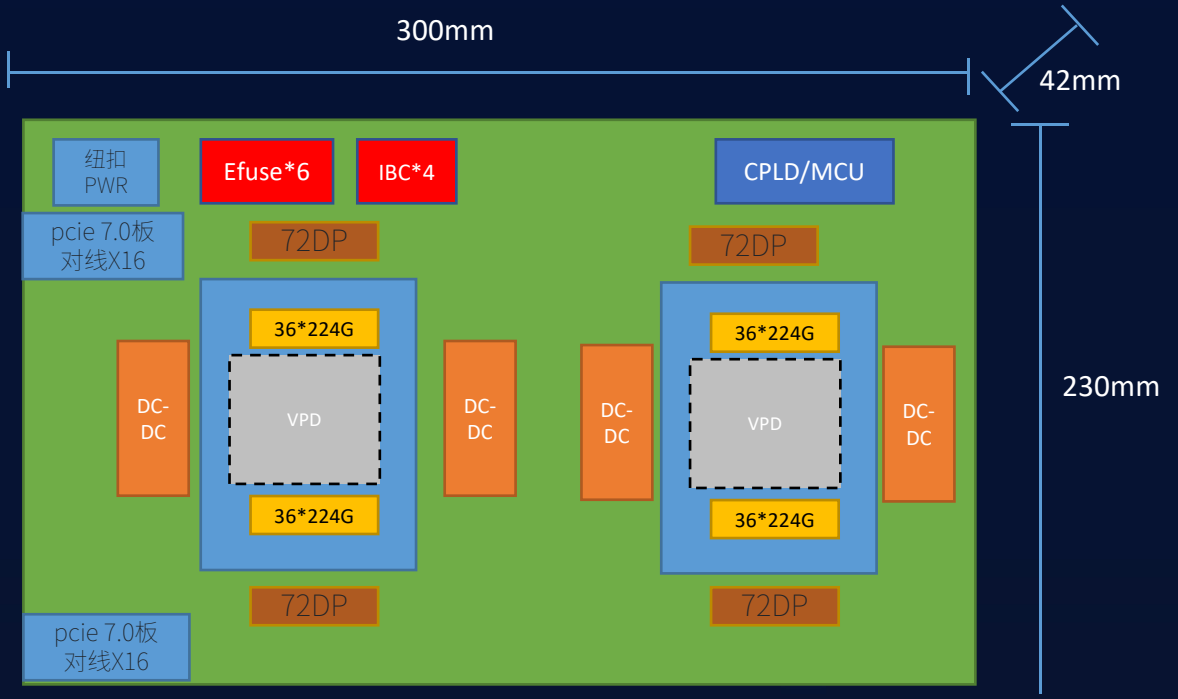


OPEN UXM-Type 3 规格

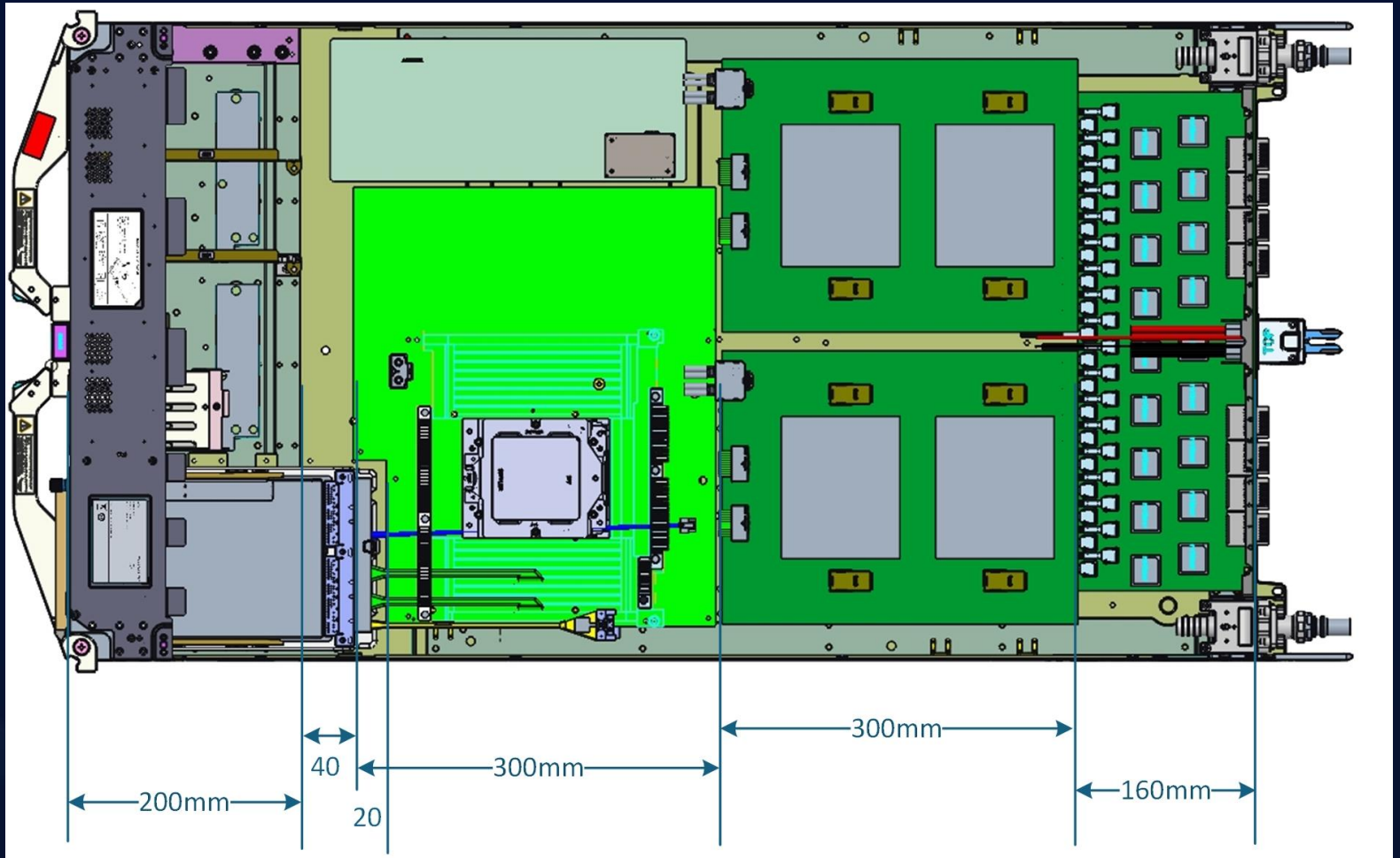
Open AI Infra
A Chapter of GCC

Type3 双芯

- 单卡尺寸230*300*42mm
- 芯片尺寸：2*XPU，单XPU 150mm*120mm
- 芯片尺寸：1*XPU，单XPU 200mm*160mm
- 单卡功耗：3000W*2
- PCIe接口： 2* PCIe6.0/7.0 X16
- 高速接口： 2*72lane，最高224G，NPO/NPC
- 支持VPD



OPEN UXM-Type3 计算节点



计算节点总深度=240+300+300+160=1000mm

OPEN UXM 规范项目计划

Open AI Infra
A Chapter of GCC



诚邀您继续参与撰写规范，提供宝贵反馈，携手助力生态系统的发展

zhangguangbin@gccorg.com
wanghaifeng@gccorg.com

Global Computing Consortium 全球计算联盟

Open AI Infra
A Chapter of GCC

Thank You !

高晓军 0279