新一代智算超节点技术趋势与挑战

高晓军

Open AI Infra社区管理委员会 联席主席

字节跳动 服务器架构师

新一代智算超节点技术趋势与挑战

高晓军

Open AI Infra社区管理委员会 联席主席

字节跳动 服务器架构师

AI新阶段：模型创新放缓、应用爆发，基础设施成核心壁垒

Open AI Infra
A Chapter of GCC

新一代智算超节点关键需求

Open AI Infra
A Chapter of GCC

计算

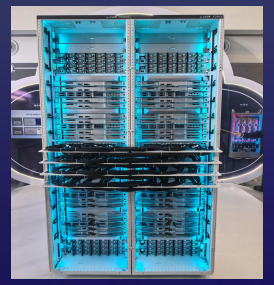
存储

网络互连

系统

能源

开放架构、多元算力



弹性扩展、极速部署

突破功耗墙	高带宽、低延迟	速率/带宽倍增	提升利用率	全液冷
异构计算	内存共享、弹性扩展	大Radix互连	弹性伸缩	能 - 算 - 碳协同
MOE原生适配	冷热数据分层	低延迟、高可用	故障自愈	全生命周期能效

异构计算与KV Cache：突破AI算力、内存瓶颈

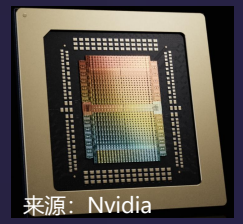
Open AI Infra
A Chapter of GCC

近存计算/ 存内计算减少数据搬运提升计算效率

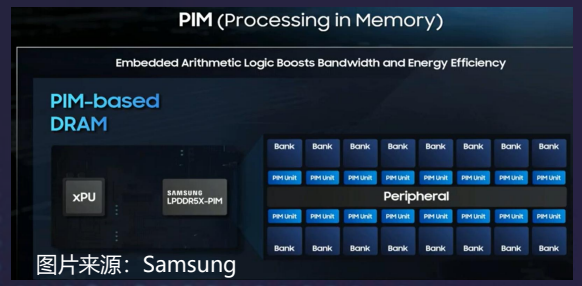
近存计算 (Near-Memory Computing)：计算单元贴近 HBM / 高带宽存储，适配低延迟大模型推理

Nvidia Groq 3 LPU：

- 500MB SRAM
- 150TB/s SRAM bandwidth
- 2.5TB/s scale-up bandwidth


来源：Nvidia

PIM(Processing-In-Memory) 、IMC(In-Memory Computing)技术逐步成熟，减少、消除数据搬运，协同提升系统算力。



图片来源：Samsung

图片来源：昕原半导体

KV Cache系统：AI推理的“新内存墙”与破局

应用：大模型推理 / 多轮对话 / Agent

KV Cache 全局调度与管理

L1：HBM　L2：DRAM（Host）　L3：Local SSD　L4：Network Storage

L1.5：HBF?　L2.5：Memory Pool　L3.5：ICMS

高速互连、协议支持（PCIe/CXL、KV原生语义、GDR、GDS）

- KV Cache 从推理优化手段升级为AI 基础设施核心组件
- 硬件与架构协同：
  - 异构算力、多级存储、高速网络
  - 不同分层存储介质的支持和兼容，特别是新的中间层
  - 新协议支持
  - 分层调度策略的系统性协同和优化
- 挑战：容量--成本--延迟的三角平衡，一致性、热设计、生态标准化

互连技术：从 "连接组件" 到 "计算核心"

Open AI Infra
A Chapter of GCC

01

超高密AI Rack

- 超高密机柜，柜内铜互连，高密、高速、供电、散热、可靠性等诸多技术突破和新建产业链。
- 机房基础设施的技术要求、部署。

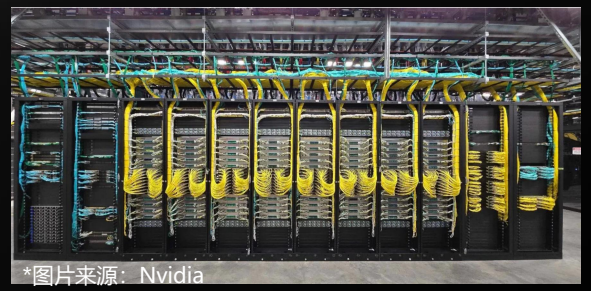
*图片来源：AMD


*图片来源：Nvidia

02

探索HBD的扩展边界

- 两层组网+光互连，多机柜实现更大的HBD

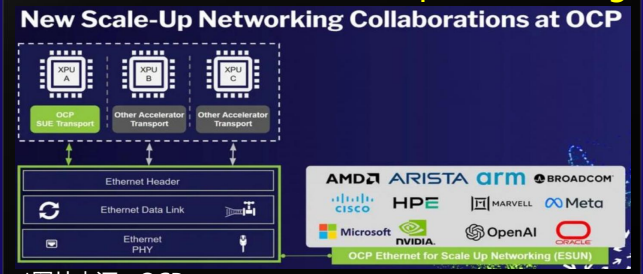

*图片来源：Nvidia

03

XPU语义和网络协议

- 生态：广泛应用、满足AI需求的带宽和容量演进节奏
- 规模扩展性
- Scale up、Scale out融合
- 铜互连、光互连兼容

ESUN(Ethernet for Scale-Up Networking)



*图片来源：OCP

Scale up光互连：突破铜互连物理极限

Open AI Infra
A Chapter of GCC

XPU Scale up：高带宽36lane/72lane，224G+
OE通道数现状：16ch/32ch

OE封装、socket选型与定制、封装结构、板级布线设计
OE管理、光链路诊断、光功率监控
OE散热、液冷设计

OE规格与Scale up带宽平衡

XPU模组直接出光

适合AI的光连接走向成熟

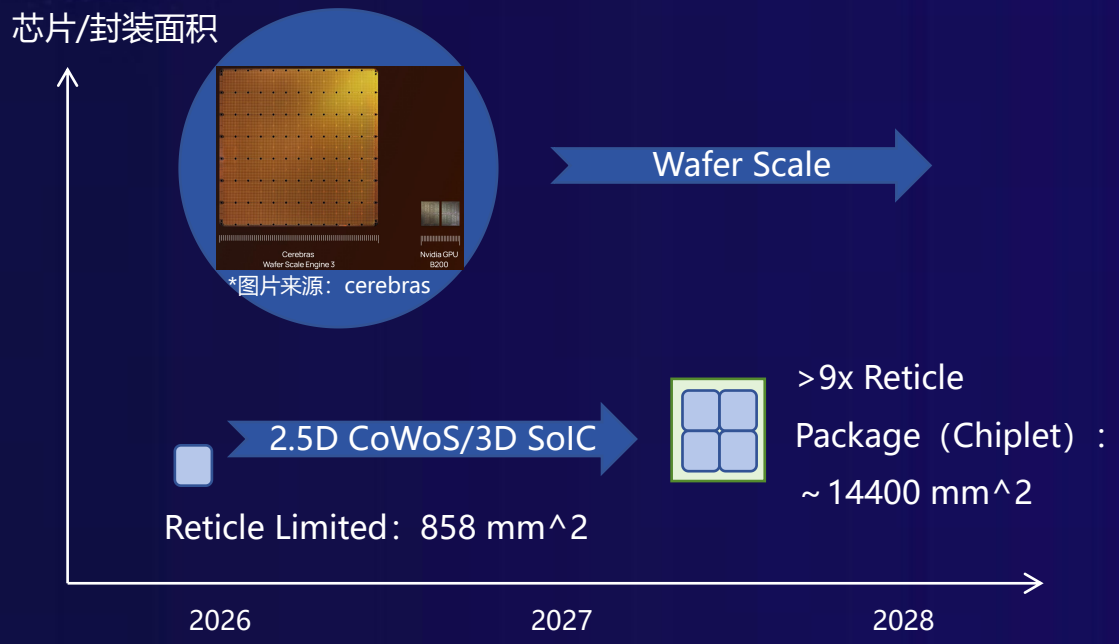
开放标准百花齐放

硅光 SiPh + 外置 CW 光源
EML（InP 基）电吸收调制激光器
硅光集成多波长光源 + 片上 CWDM 波分复用

OIF CEI：底层电气、管理、框架规范
Open CPX：物理封装 / 接口，socket、连接器、机械、热、光引擎接口
OCI MSA：光 PHY / 链路层，调制、波分、激光器、链路协议

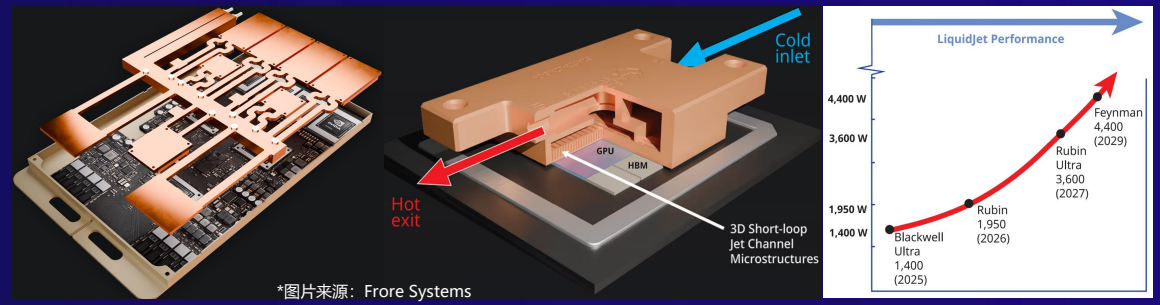
提升散热能力，支持高密度MW级系统散热

Open AI Infra
A Chapter of GCC



- 芯片功耗不断提升
- 冷板Fin Gap： 0.15mm → 0.10mm → ?mm
- 冷板铲齿加密，冷板流阻增加，工质及环网脏污造成堵塞风险提升

- 单芯片功耗推高到3000W+， MLCP冷板/射流冷板/相变冷板等进一步提升冷板散热能力
- 芯片级精准散热、半导体级的精密制造



- 节点/机柜：
  - 更大流量及可用压降
  - 更高效的快接头
  - 热管理
- 机房液冷系统长期运行挑战：
  - 工质
  - CDU过滤
  - 高洁净度管路
  - 在线监测

推进原生液冷部件生态发展，提升液冷占比

Open AI Infra
A Chapter of GCC

提升液冷占比

风液混合=>全液冷

单机柜功率提升到300kW~500kw，风液混合，90%液冷占比下，风侧负荷超30kW。

风液比：全液冷，免风扇

冷板：减重、易维护、成本优化

液冷部件设计

网卡：插拔设计、Form Factor定义
SSD、内存：颗粒布局、系统密度
PSU：降噪声
快接头：小型化、标准化、高性能

原生液冷部件

供电刚需：800V HVDC、AI计算负载功率波动治理

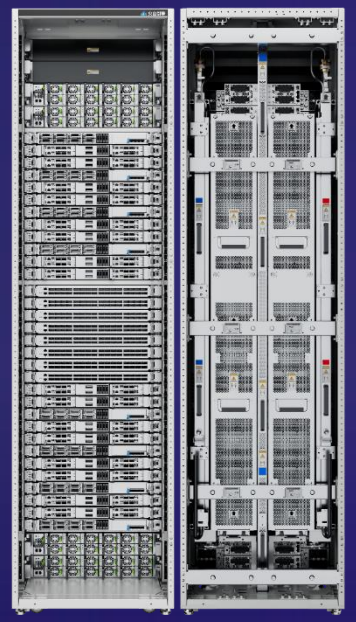
Open AI Infra
A Chapter of GCC

单机柜300kW+

单芯片功率倍增

单柜芯片数量增长

有限的机柜空间



功率 vs 机柜空间

应对工程挑战

直流系统的单点故障扩散

故障隔离有效性

机柜宽度压力

IT运维：800VDC操作

Reliability

Availability

Safety

TCO

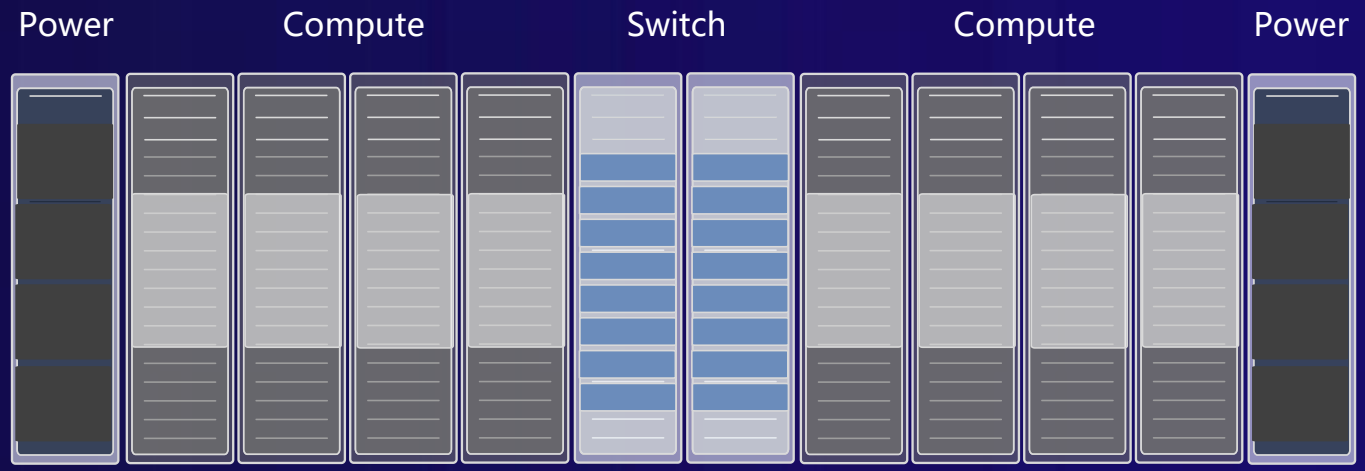
Regulatory compliance

Efficiency

EDPp

依托Open AI Infra社区 共同推进超节点生态发展

Open AI Infra
A Chapter of GCC



算力及部件

- 多样性算力：

-新一代FormFactor

- 部件

-原生液冷

系统互连

- 更大HBD：64/128 --> 256/512

硬件资源池：Memory Pool、SSD、DPU

- 速率：112G bps--> 224G bps
- 互连方式：高速铜互连、NPO
- 机柜标准

制冷

- 冷板进化：

-3000W级别的芯片散热

- 提升液冷占比：

-DIMM、SSD、NIC、光模块

- 基础设施接口

-流量：快接头、冷板Fin Gap

-水质在线监测

供电

- 系统级：

-800VDC：直流保护，运维操作

-标准与生态

-EDPp治理

- 板级

-垂直供电

-高密、高性能

开放共创AI Infra未来