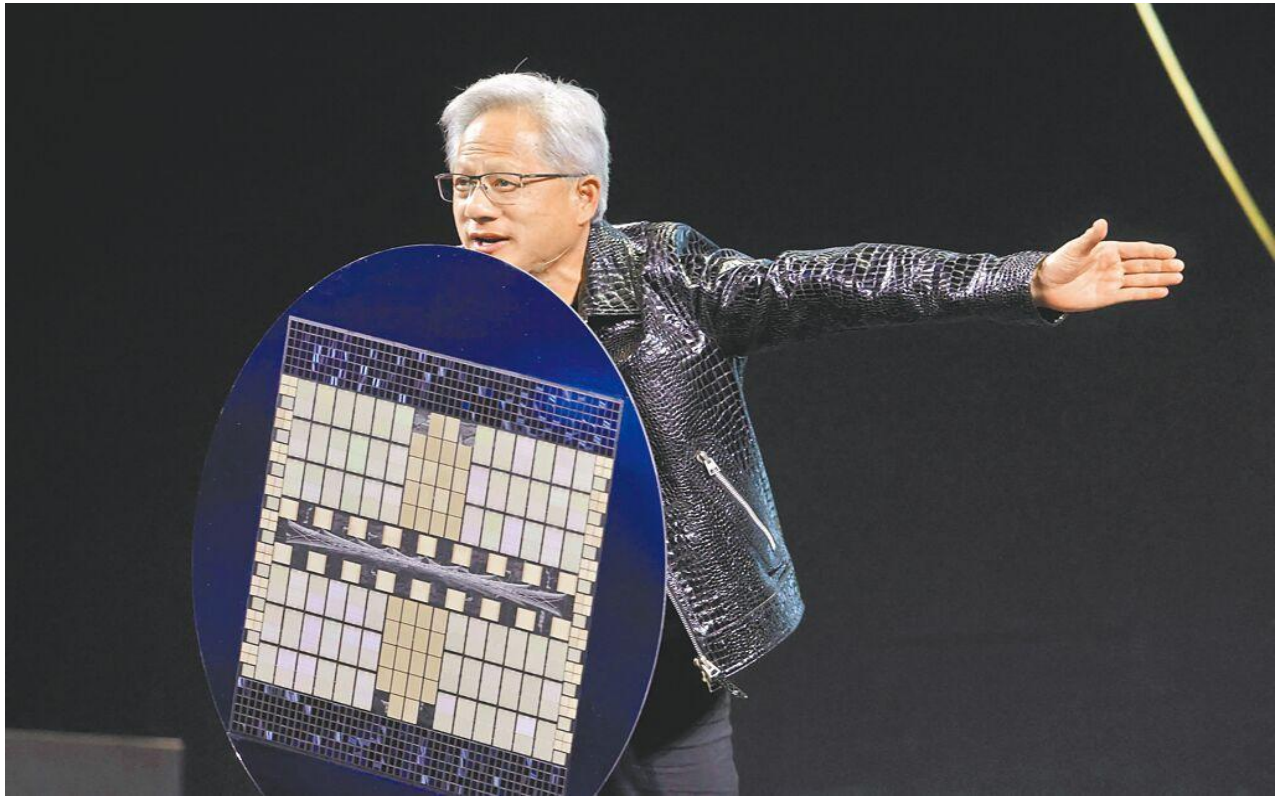
AI服务器液冷发展

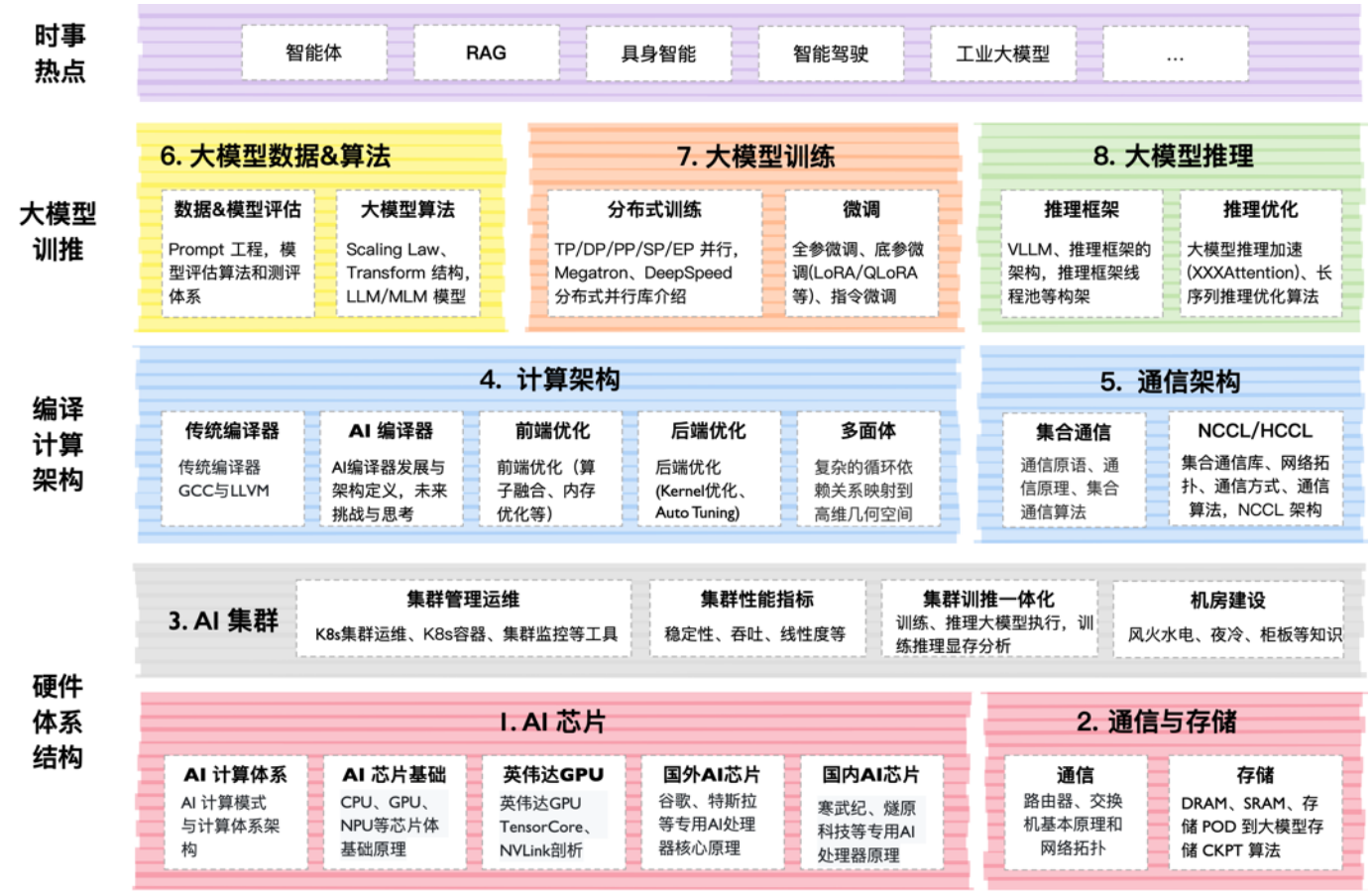
ZOMI

Content

- 未来的服务器能搞成这样的盾牌吗?

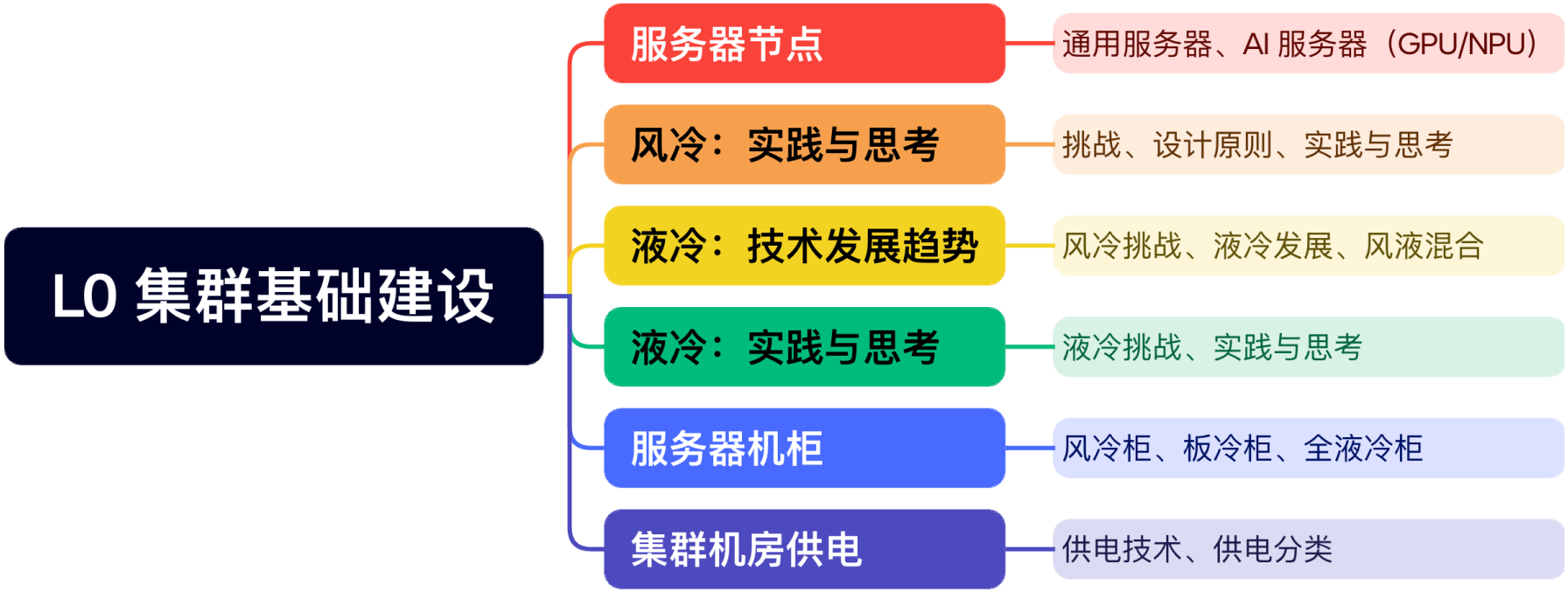


Content github.com/Infrasys-AI/AIInfra



GitHub github.com/Infrasys-AI/AIInfra book infrasys-ai.github.io

Content


L0 集群基础建设
服务器节点
通用服务器、AI 服务器（GPU/NPU）
风冷：实践与思考
挑战、设计原则、实践与思考
液冷：技术发展趋势
风冷挑战、液冷发展、风液混合
液冷：实践与思考
液冷挑战、实践与思考
服务器机柜
风冷柜、板冷柜、全液冷柜
集群机房供电
供电技术、供电分类


好好想想

Content

1. AI 服务器冷却挑战
2. 液冷通用架构
3. 液冷技术发展
4. 风液混合方案

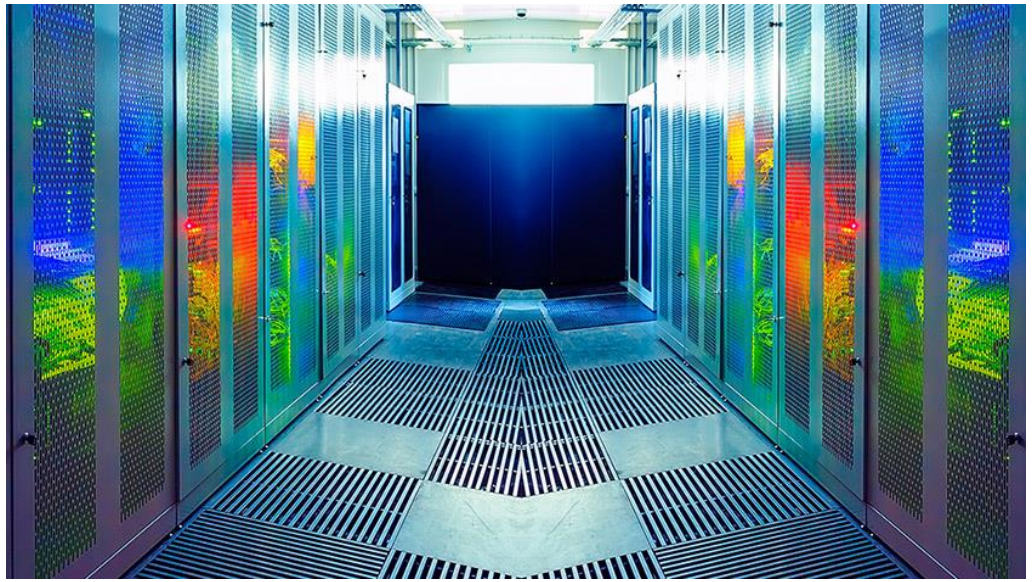
01

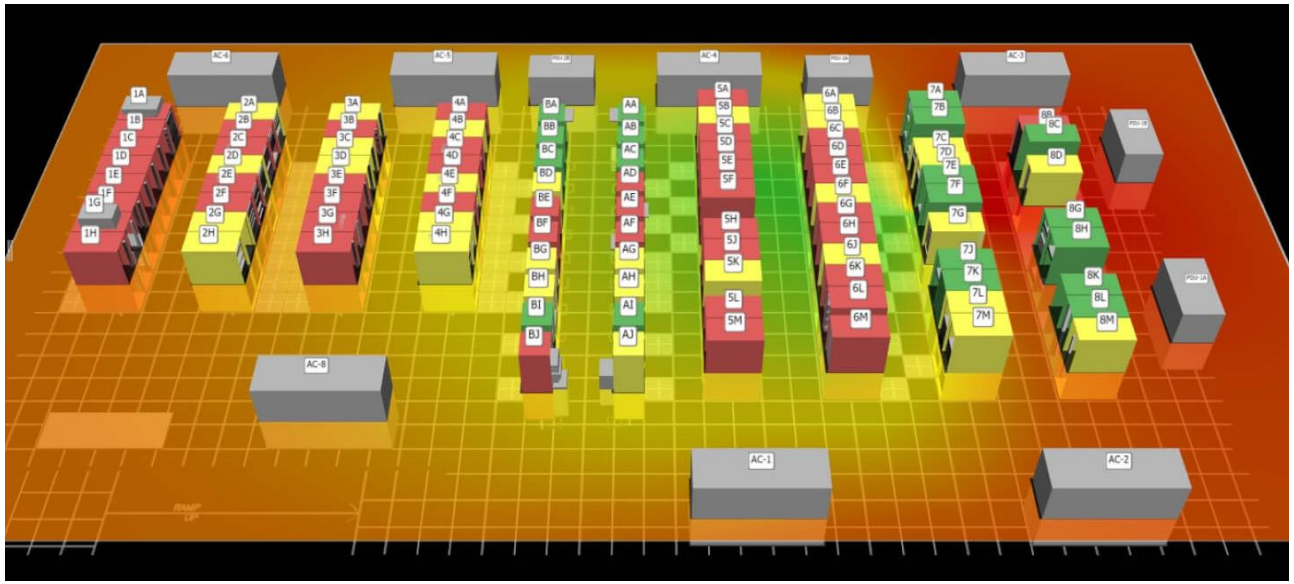
AI 服务器冷却挑战

芯片和单机柜功率密度增大

- **当前 AI 集群热功率达到风冷的极限：**

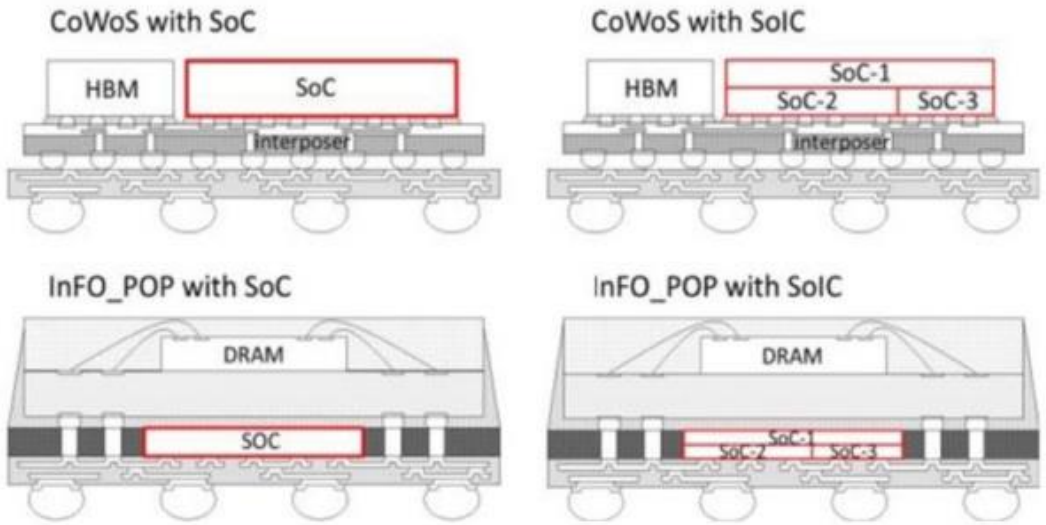
1. 单芯片 TDP（Thermal Design Power，热设计功耗）超 300w（90w/cm2）
2. 单机架整体功率密度超 >20KW

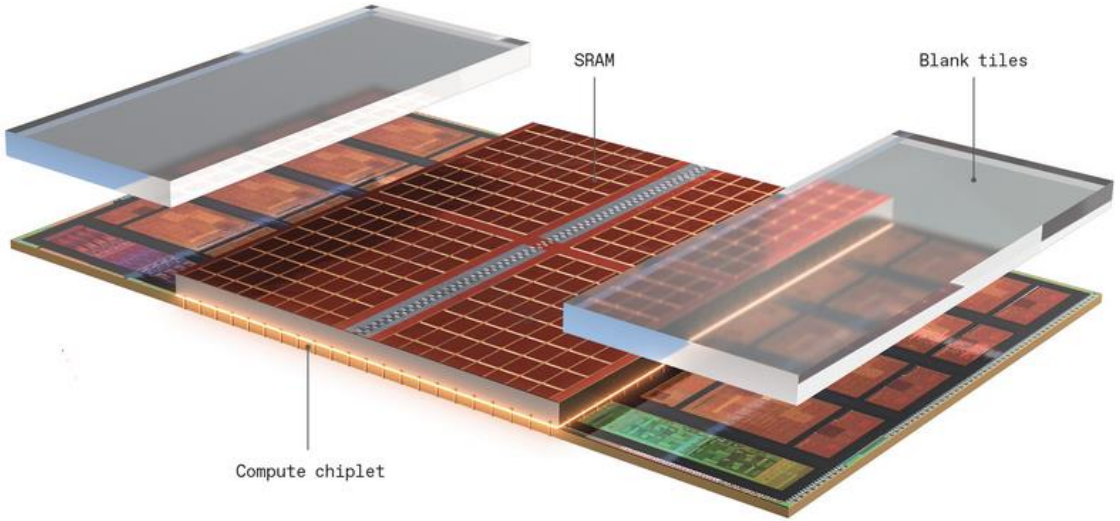




1 单芯 TDP > 300w

- 通过制程工艺缩小器件尺寸、研发新材料和电路结构来提升单位面积晶体管数，半导体制程接近物理极限，先进封装是延续摩尔定律重要路径。
- 多芯粒 2.5D/3D 封装在提升系统性能同时，架构堆叠 + 系统功率与热源密度提升，高效散热方案是刚需。

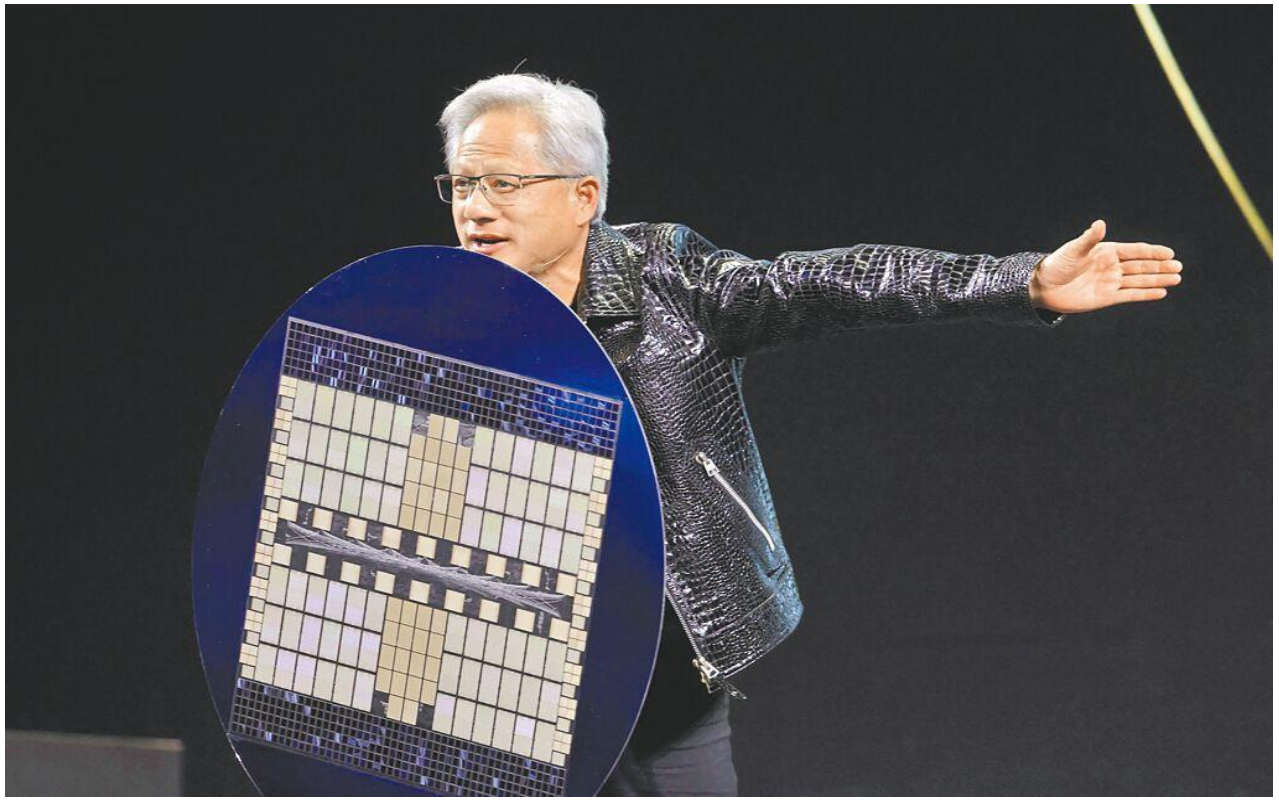


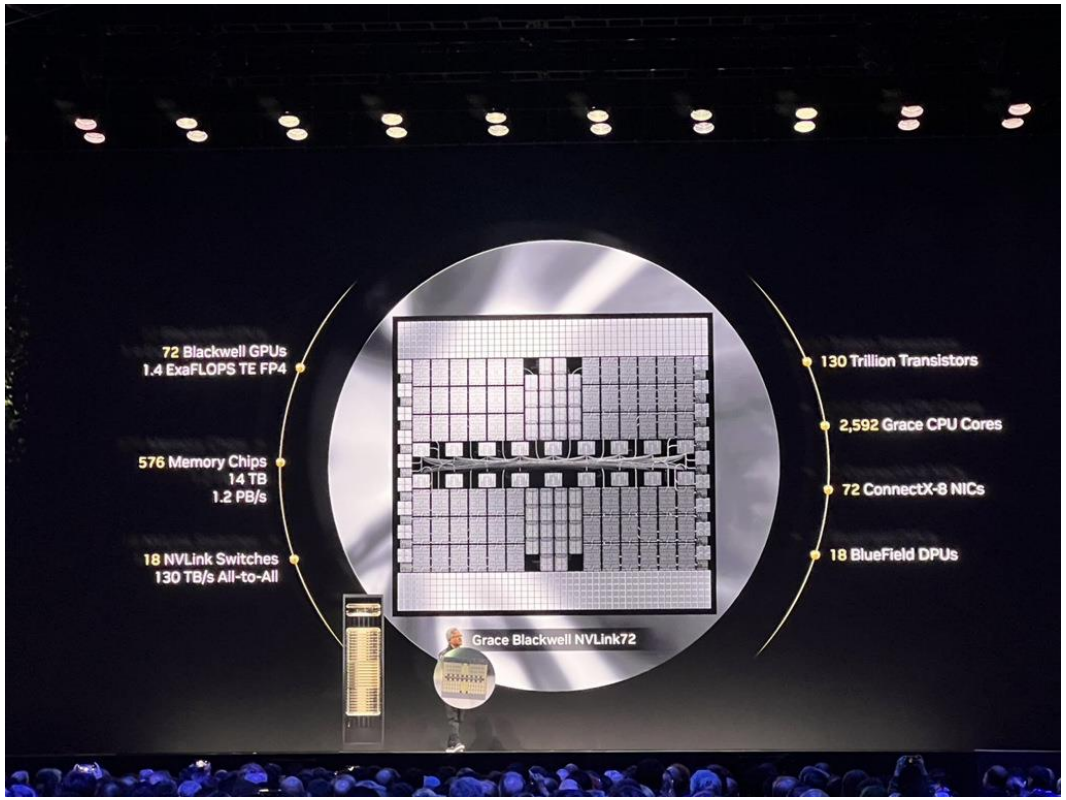


台积电 3D Fabric 路线，芯片热源密度提升

1 单芯 TDP > 300w

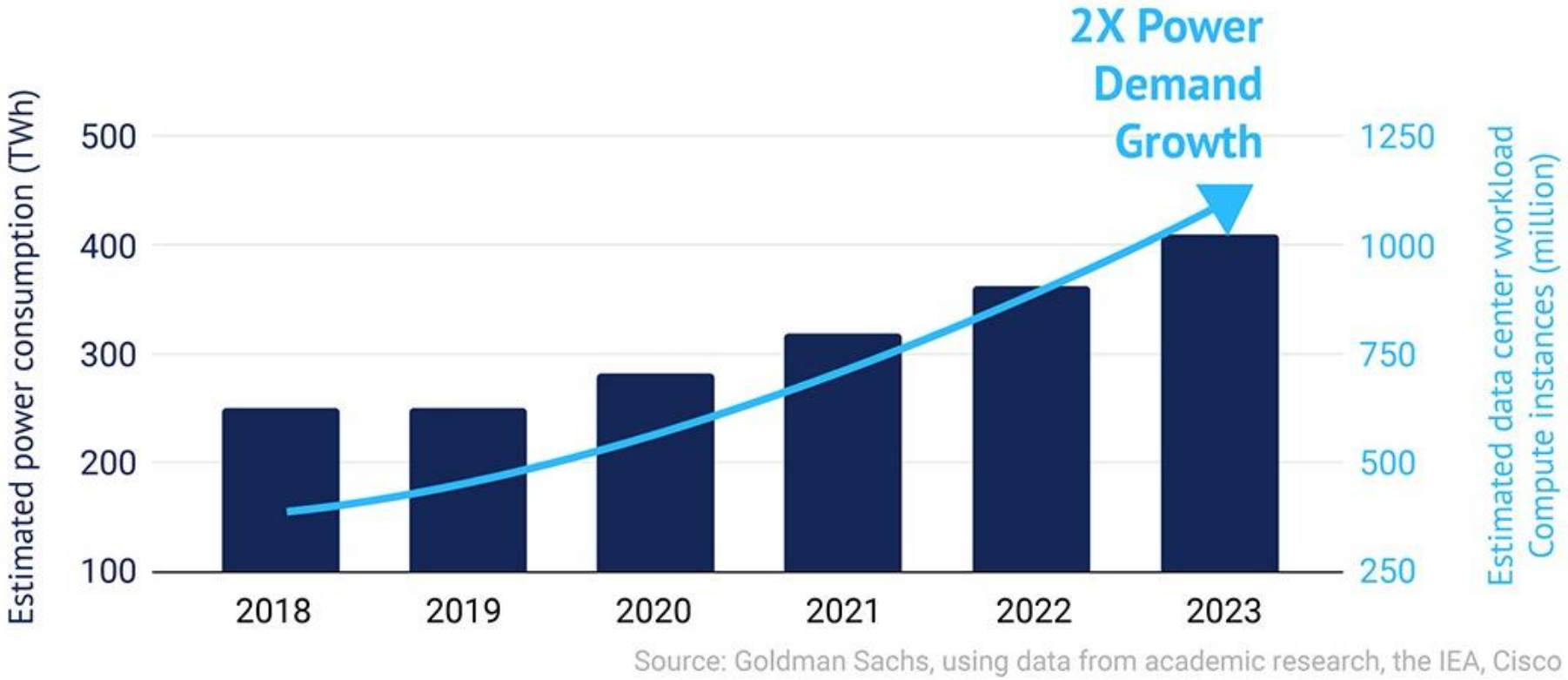
- H100 单卡功率 ~700W，8 卡 GPU 功率 ~5.6kW





1 单芯 TDP > 300w

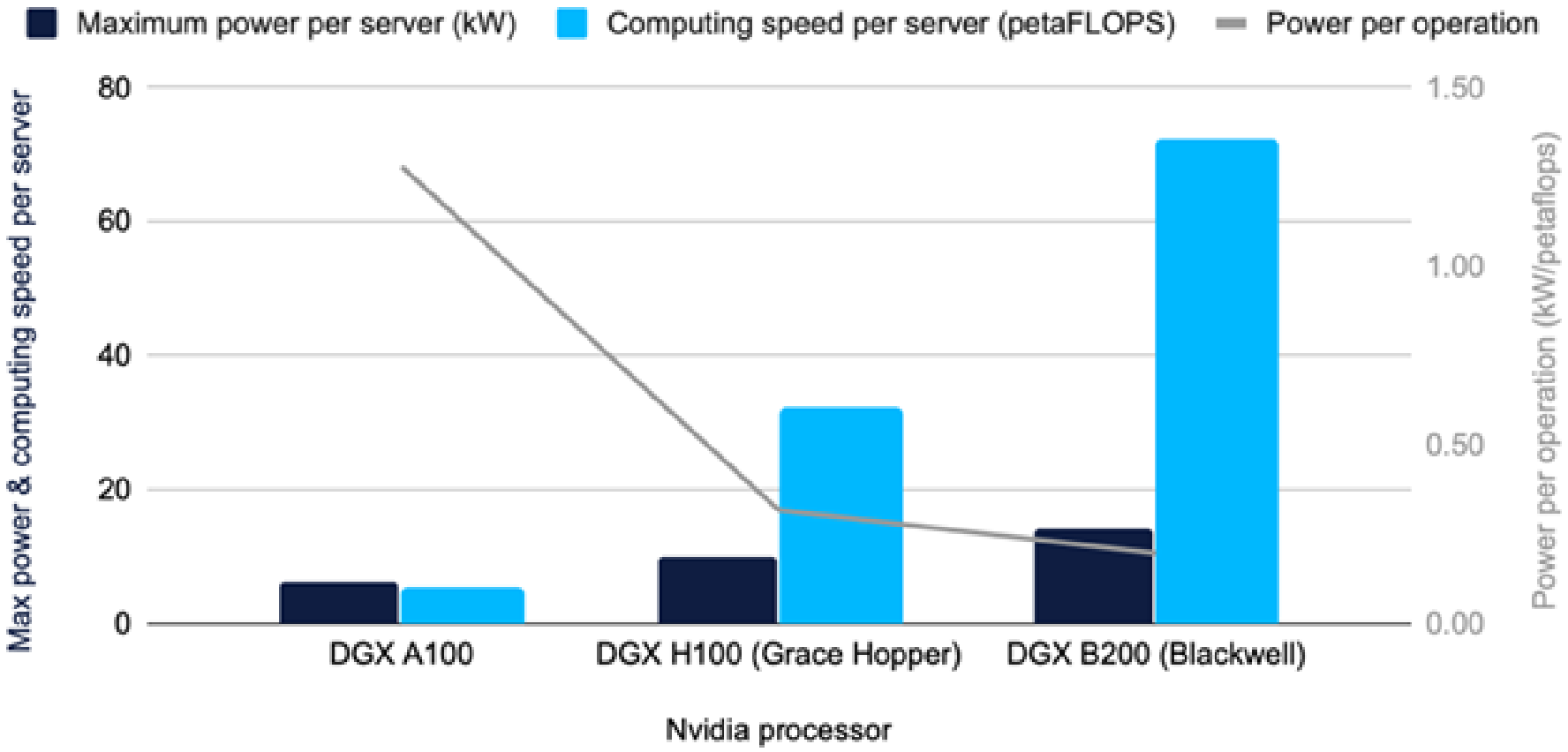
- 为满足算力需求，单芯片算力逐步提升，单芯片功耗，爆发式逐年递增
  - 20年到24年，单AI芯片功耗提升3倍



Nvidia’s Grace Hopper Runs at 700 W, Blackwell Will Be 1 KW.

1 单芯 TDP > 300w

- AI servers are becoming significantly more efficient, but the rise of AI usage requires greater power d ensities to be delivered



2 单机架功率密度 >20KW

- 为满足集群算力规格需求，减少互联成本和延迟，单机柜功耗也在逐步提升
  - 从 20kw 逐步突破到 130kw，提升 6 倍多

NV GH200 NVL32

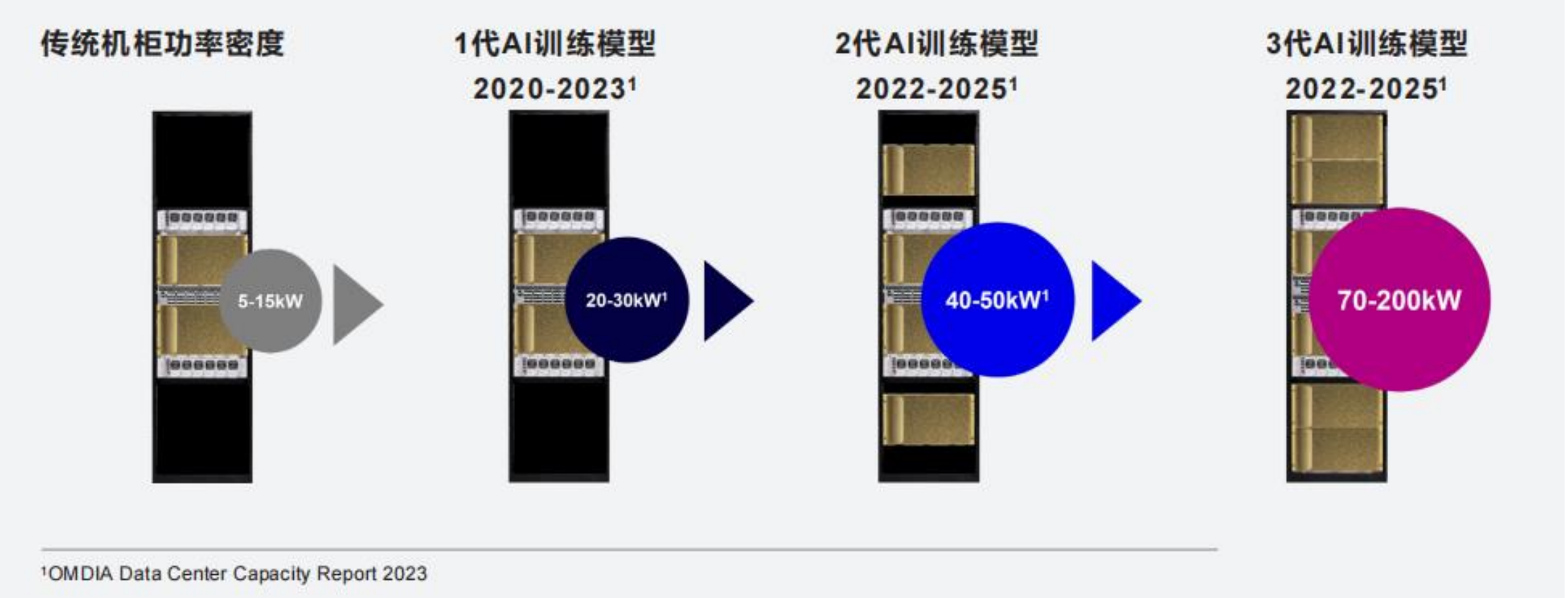
~40 KW

NV GB200 NVL72

~130 KW

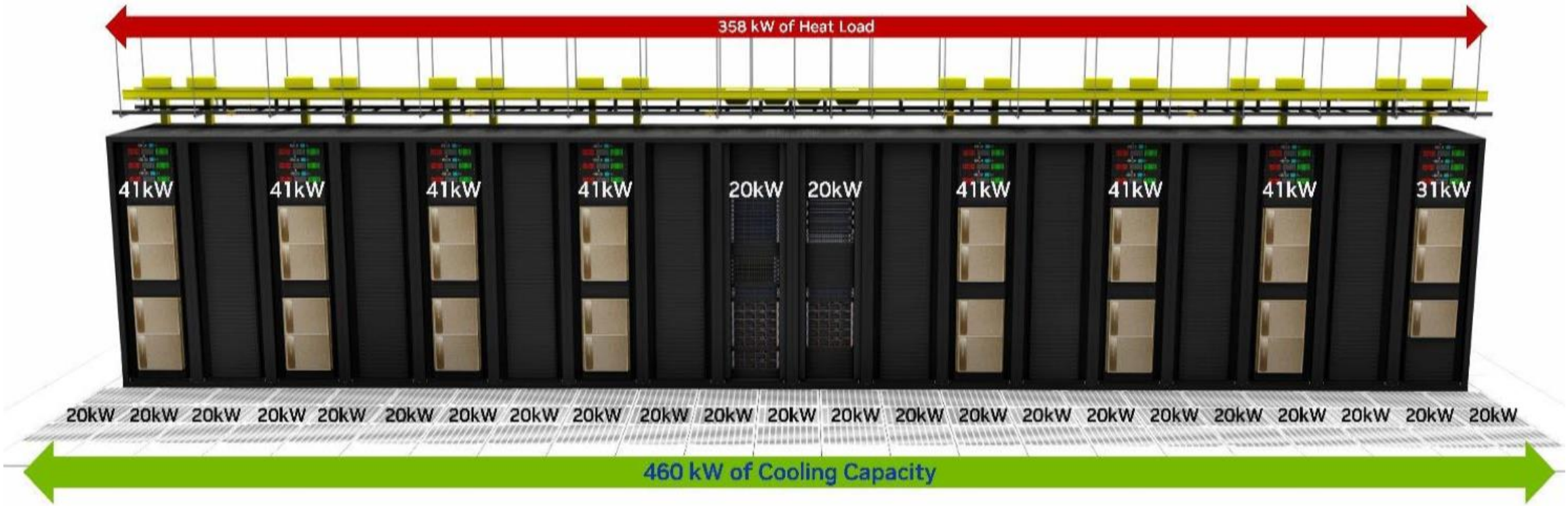
❷ 单机架功率密度 >20KW

- 单机柜功耗从传统 DC 4～6KW 逐渐增加至 AI 智算中心 20～40kW，未来逐步发展至 40～120kW，智算中心机柜呈现高密度化趋势。



❷ 单机架功率密度 >20KW

- AI 大模型训推重构算网架构，大模型参数量的增速 >GPU 内存增速，高集成度 + 大内存 + 多 NPU 互联系统更适配大模型的形态，单机架进行组合形成超节点。



以 NVIDA 为例

1. NVIDIA HGX 服务器，一台机器包含 8/4 个 GPU.以 8 个 H100 为例，单台服务器功耗 ~10.2kW，选用 B200 单台服务器设计功耗 14.3kW

2. NVIDIA 从 NVL32 到 NVL72 机柜，单机柜部署 4 台服务器至 9 台服务器，GPU 数量从 32 到 72 颗，总功耗也从 44kW 增加到 120kW

架构	HGX A100	HGX H100	HGX H200	HGX B100	HGX B200
	8 x A100 SXM	8 x H100 SXM	8 x H200 SXM	8 x B100 SXM	8 x B200 SXM
	Ampere	Hopper		Blackwell	
显存大小	640GB	1.1TB	1.1TB	1.44/1.5TB	1.44/1.5TB
显存宽带	8 x 2TB/s	8 x3.35TB/s	8 x 4.8TB/s	8 x 8TB/s	8 x 8TB/s
FP16稠密算力（FLOPS）	2.4P	8P	8P	14P	18P
INT8稠密算力（OPS）	4.8P	16P	16P	28P	36P
FP8稠密算力（FLOPS）	X	16P	16P	28P	36P
FP6稠密算力（FLOPS）	X	X	X	28P	36P
FP4稠密算力（FLOPS）	X	X	X	56P	72P
GPU-to-GPU宽带	600GB/s	900GB/s	900GB/s	1.8TB/s	1.8TB/s
NVLink宽带	4.8TB/s	7.2TB/s	7.2TB/s	14.4TB/s	14.4TB/s
以太网网络	200Gb/s	400Gb/s+200Gb/s	400Gb/s+200Gb/s	2 x 400Gb/s	2 x 400Gb/s
IB网络	8 x 200Gb/s	8 x 400Gb/s	8 x 400Gb/s	8 x 400Gb/s	8 x 400Gb/s
GPU功耗	3.2kW	5.6kW	5.6kW	5.6kW	8kW
总功耗	6.5kW	10.2kW	10.2kW	10.2kW	14.3kW
备注	ConnectX-6 NIC	ConnectX-7 NIC	ConnectX-7 NIC	BlueField-3 DPU ConnectX-7 NIC	BlueField-3 DPU ConnectX-7 NIC

02

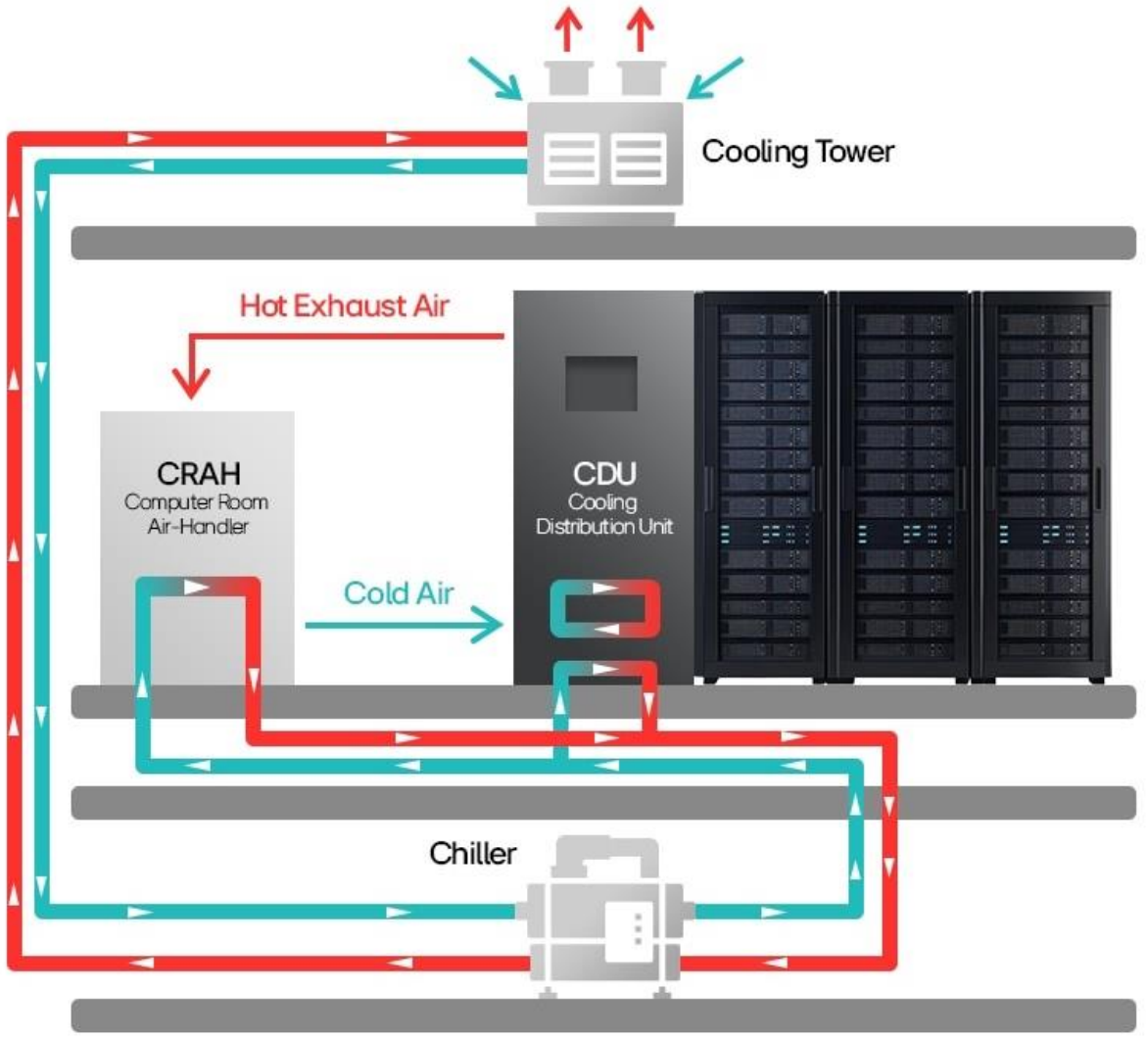
液冷通用架构

Question

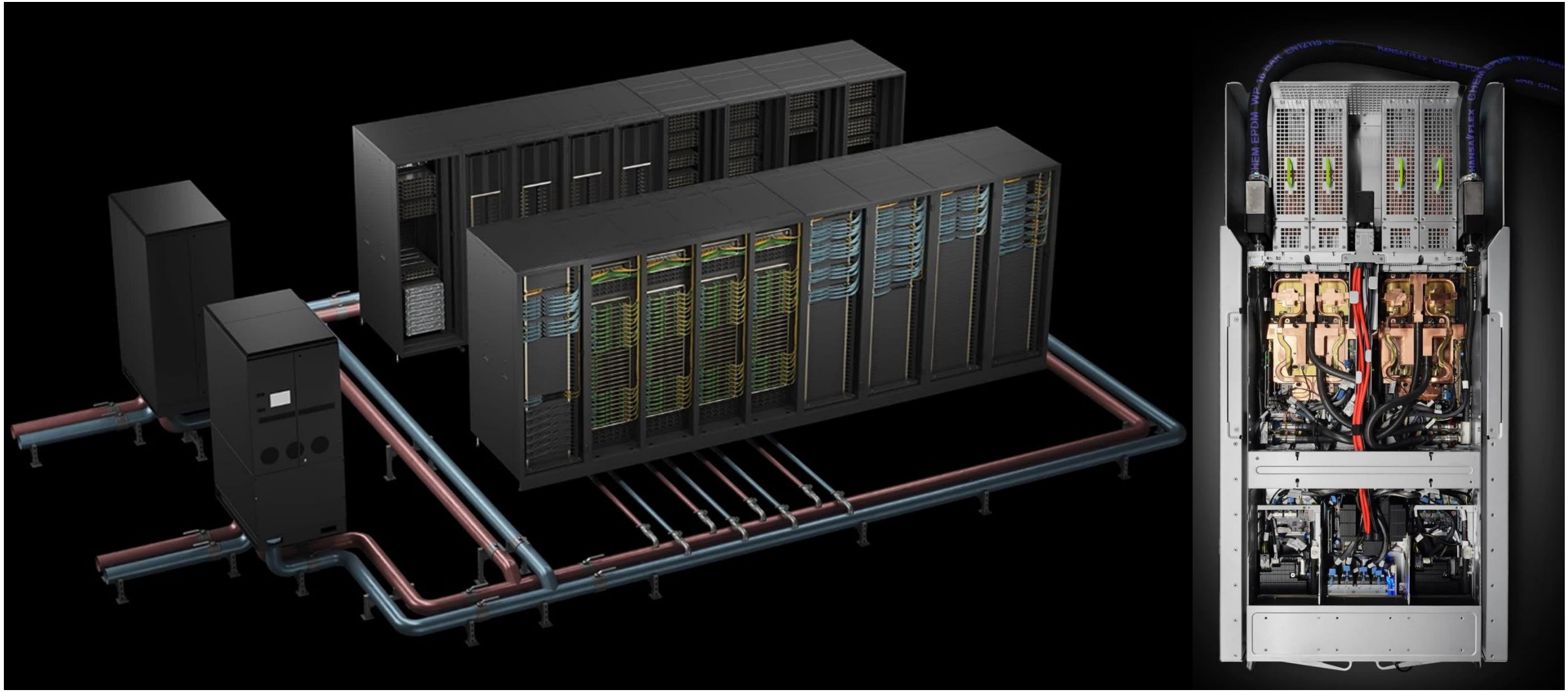
- 液冷听说有冷板式、浸没式、相变式、单相式、喷淋式各种各样的液冷设备和方案，这都有通用的架构?

液冷通用架构：能量逐级传递

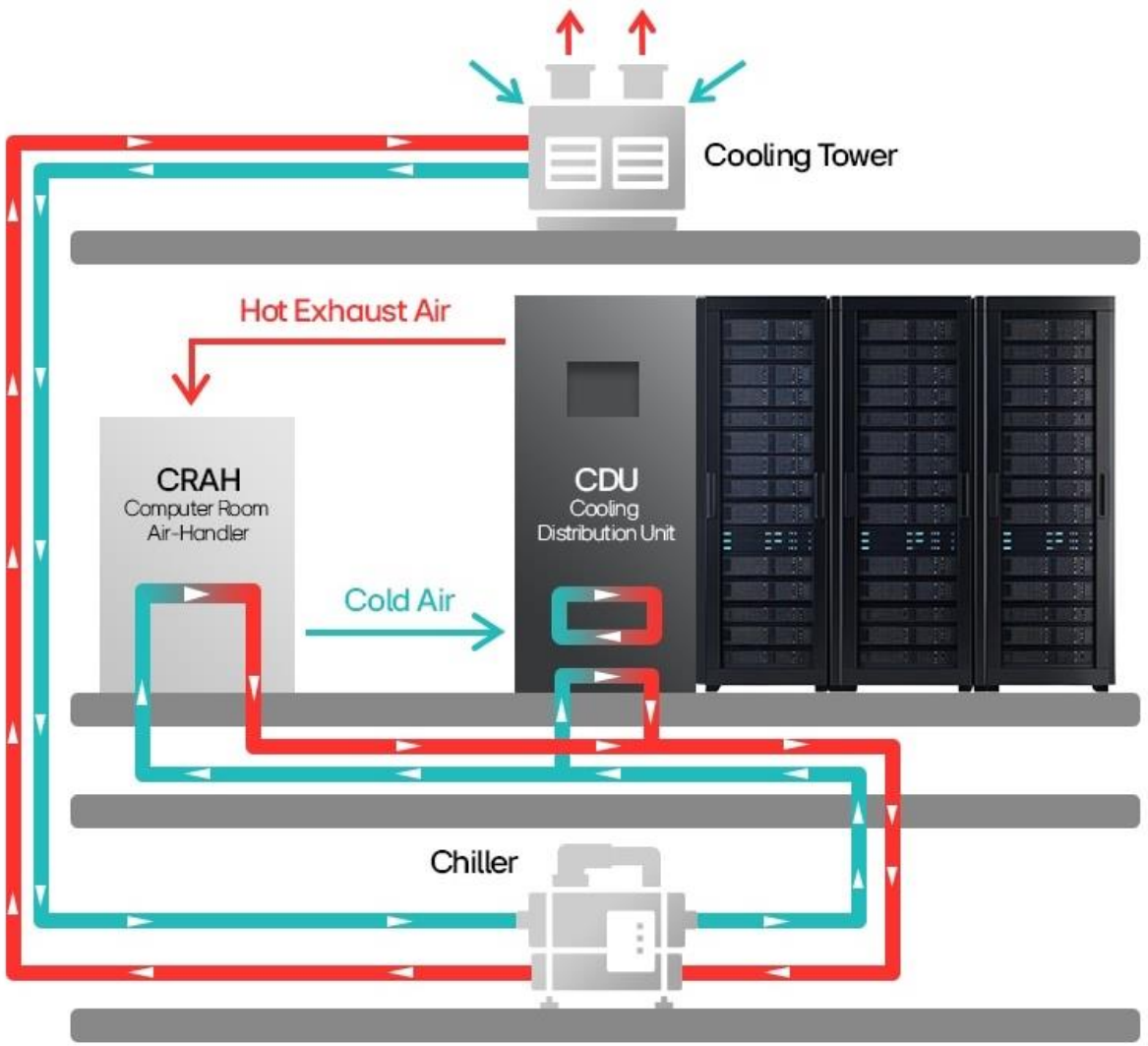
- **液冷架构主要包括三个部分：**
  1. **热捕获：**发生在液冷机柜内，指使用冷却液体将 GPU 产生的热量带走。
  2. **热交换：**连接液冷系统一次侧和二次侧的桥梁，通过 CDU（冷量分配单元）对资源进行分配与交换。
  3. **冷源：**布局在 DC 外部，热量在这一部分与自然环节交换，完成处理。



液冷流程

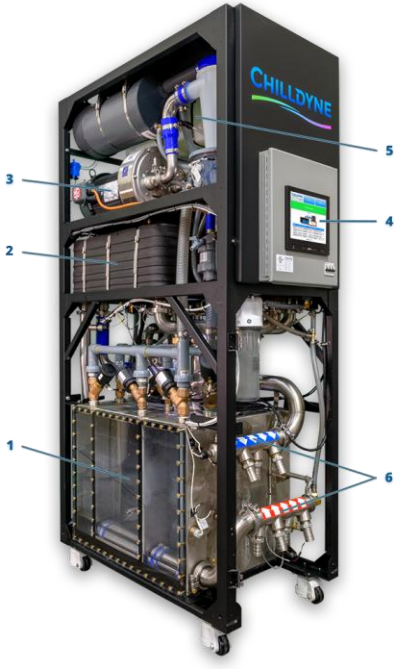


整体链路图

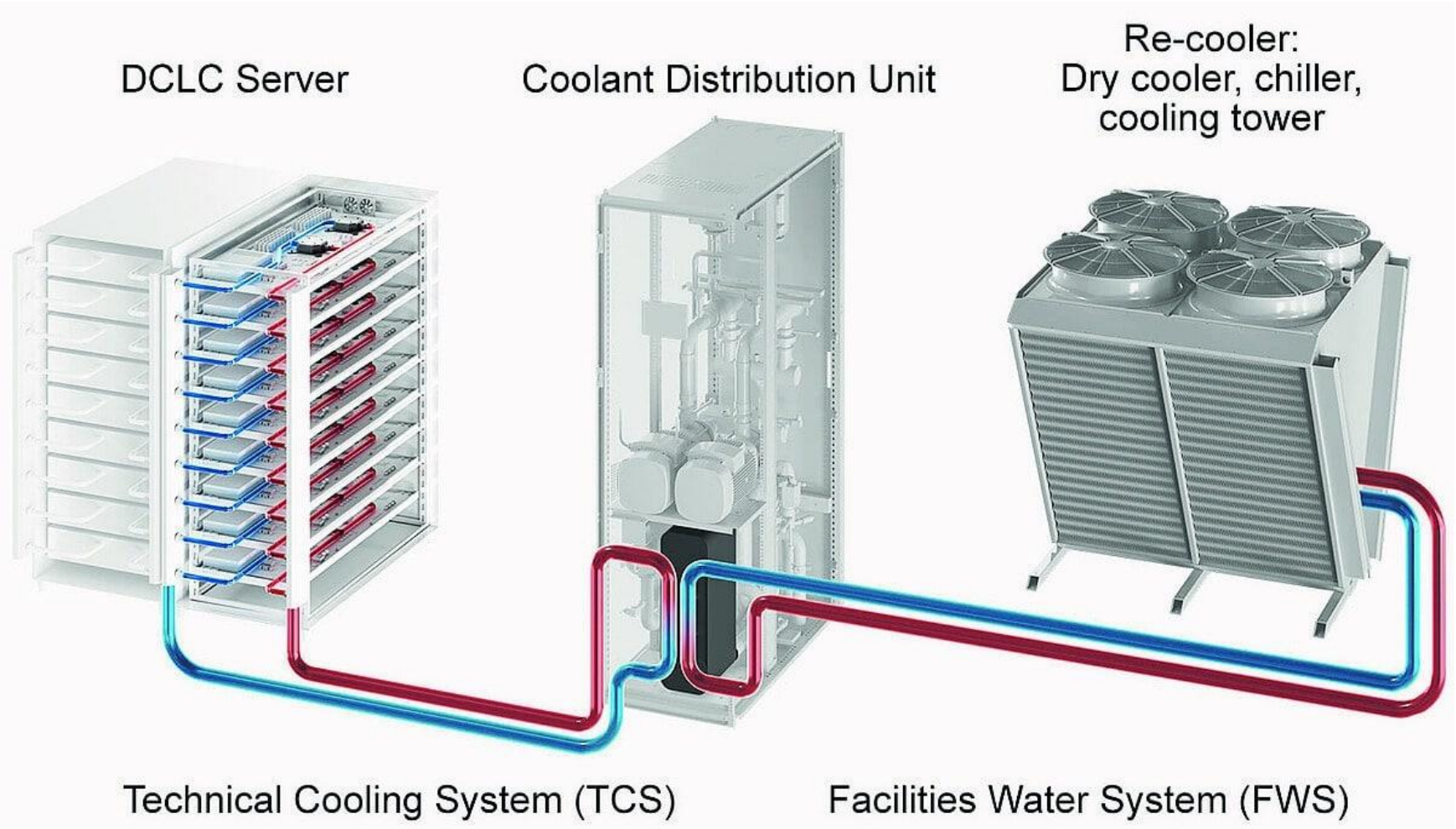

Cooling Tower
Hot Exhaust Air
CRAH
Computer Room
Air-Handler
CDU
Cooling
Distribution Unit
Cold Air
Chiller

什么是 CDU?

- CDU（Cooling Distribution Unit 冷却分配单元），液冷系统中核心组件。负责将冷却液从主冷却系统（如冷水机组或冷却塔）输送到服务器，吸收热量后再将升温的液体送回主冷却系统进行再冷却。



什么是 CDU?


DCLC Server
Coolant Distribution Unit
Re-cooler:
Dry cooler, chiller,
cooling tower
Technical Cooling System (TCS)
Facilities Water System (FWS)


好好想想

03

液冷分类

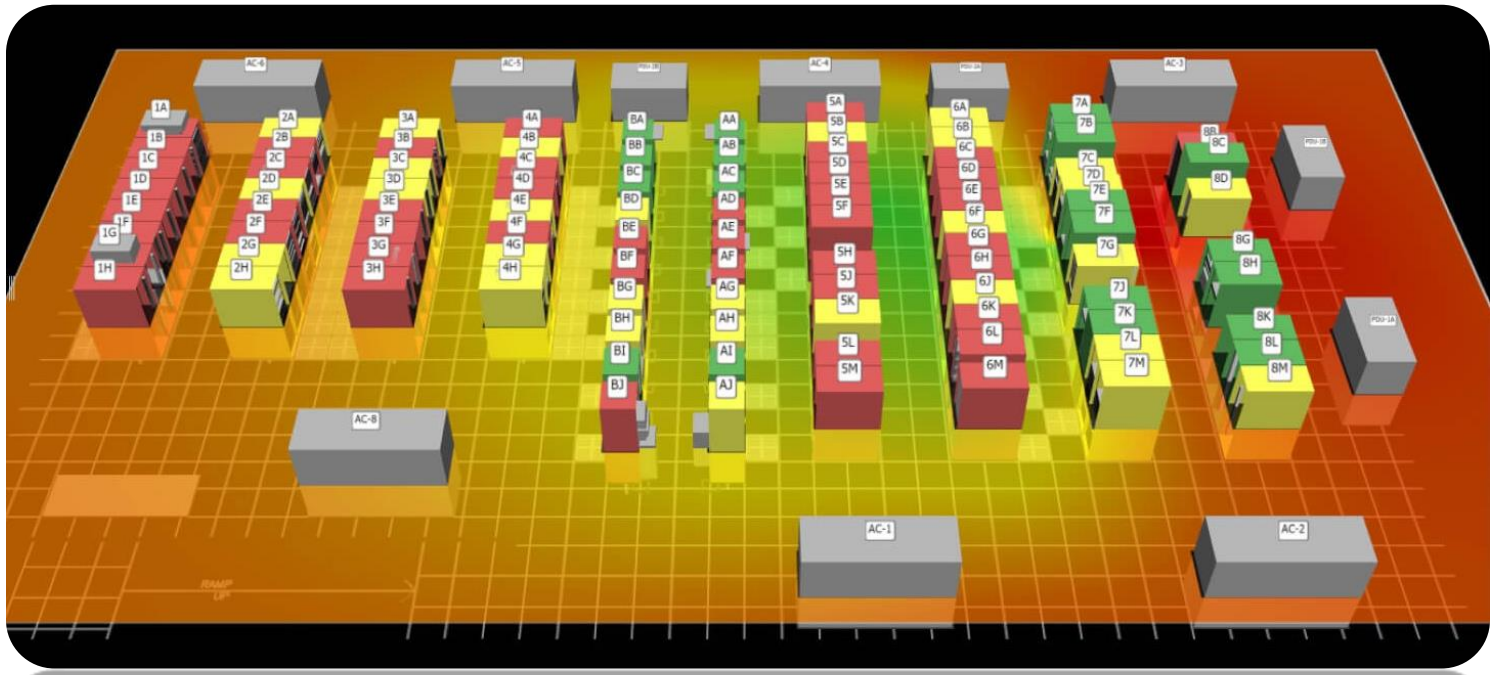
空气 vs 水

- **空气 vs 水，散热效率对比：**
  - 同等体积下，水的带热能力是空气 **~3000 倍**
  - 水的导热系数是空气 **~20 倍**
  - 水的比热容是空气 **~4 倍**

- 25°C 常压下，水的密度为 1000kg/m³ vs 空气密度 1.29 kg/m³
- 25°C 常压下，水的导热系数为 0.58W/mK vs 空气导热系数为 0.027W/mK（每开尔文温度差）
- 25°C 常压下，水的比热为 4.2 kJ/kg°C vs 空气比热为 1.0 kJ/kg°C

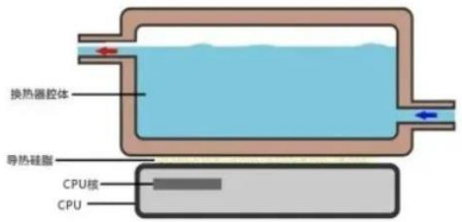
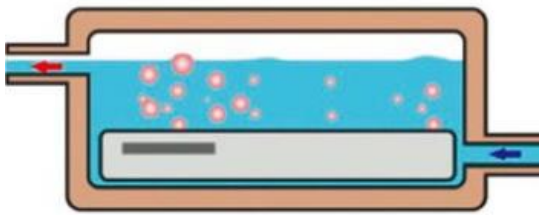
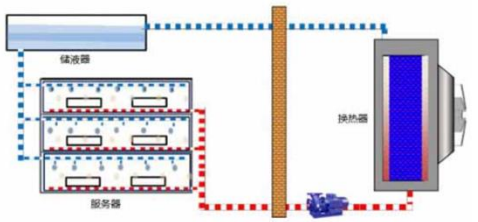
液冷的出现

1. **芯片层面：** 芯片典型功耗 >300W，需要使用液冷才能保证算力性能释放
2. **整机层面：** AI 服务器单柜功率激增 >50kW，算力密度激增迫切需要液冷渗透
3. **机房层面：** IDC PUE 从 1.5 以上降至 1.2 最有效办法选择液冷



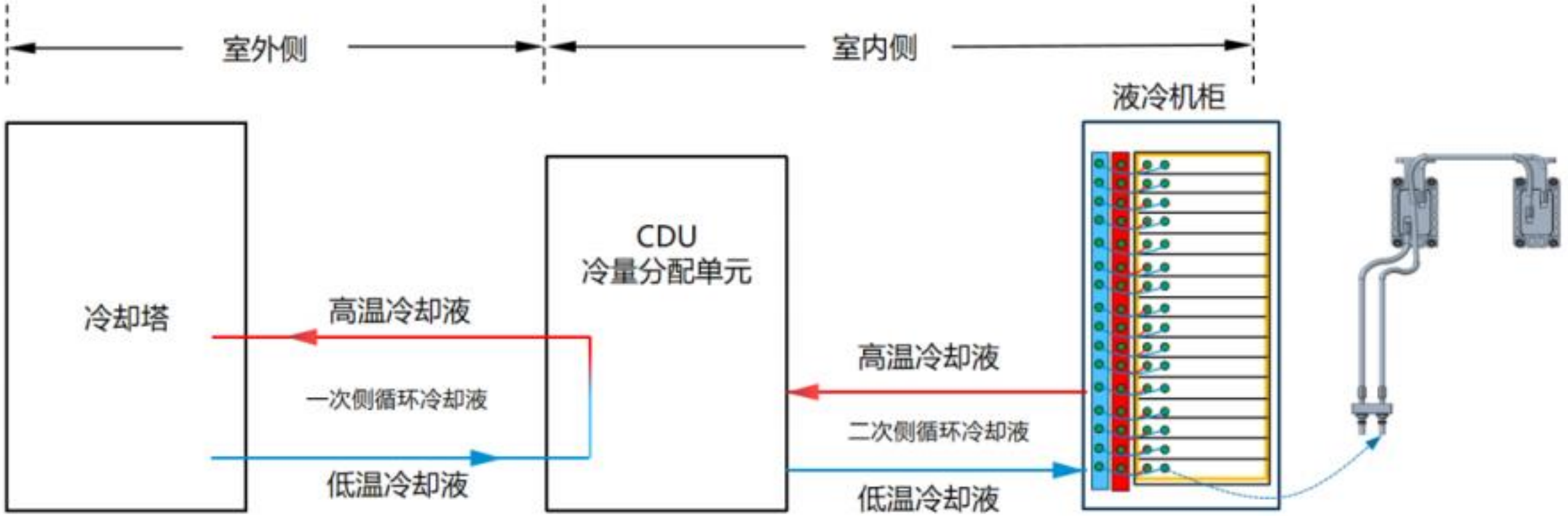
机柜功率密度 >40kW 风冷系统会失去有效性，此时可采用液冷方法

液冷散热技术

	冷板式	相变浸没	单相浸没	喷淋式
原理	• 冷板贴近热源 xPU，利用冷板中的冷却液带走热量	• 服务器完全浸没在冷却液中，冷却液蒸发冷凝相变带走热量	• 服务器完全浸没在冷却液 中，冷却液循环流动带走 热量	• 冷却液从服务器顶部喷淋，对流换热降温
特点	• 硬件系统改造小，维护简 单； • 单相+相变接头、密封件多，可靠性要求高	• 散热能力强、功率密度高，静音； • 服务器刀片式，专用机柜，管理控制复杂	• 散热能力强、功率密度高，静音； • 清理拆装难，较少运维经 验	• 静音，节省液体； 运维复杂，排液补液复杂， 密封结构
生态	• IT、冷媒、管路、供配电等不统一； • 服务器多与机柜深度耦合	• 定制化，光模块兼容待验证		• 较少
形态				

冷板式液冷

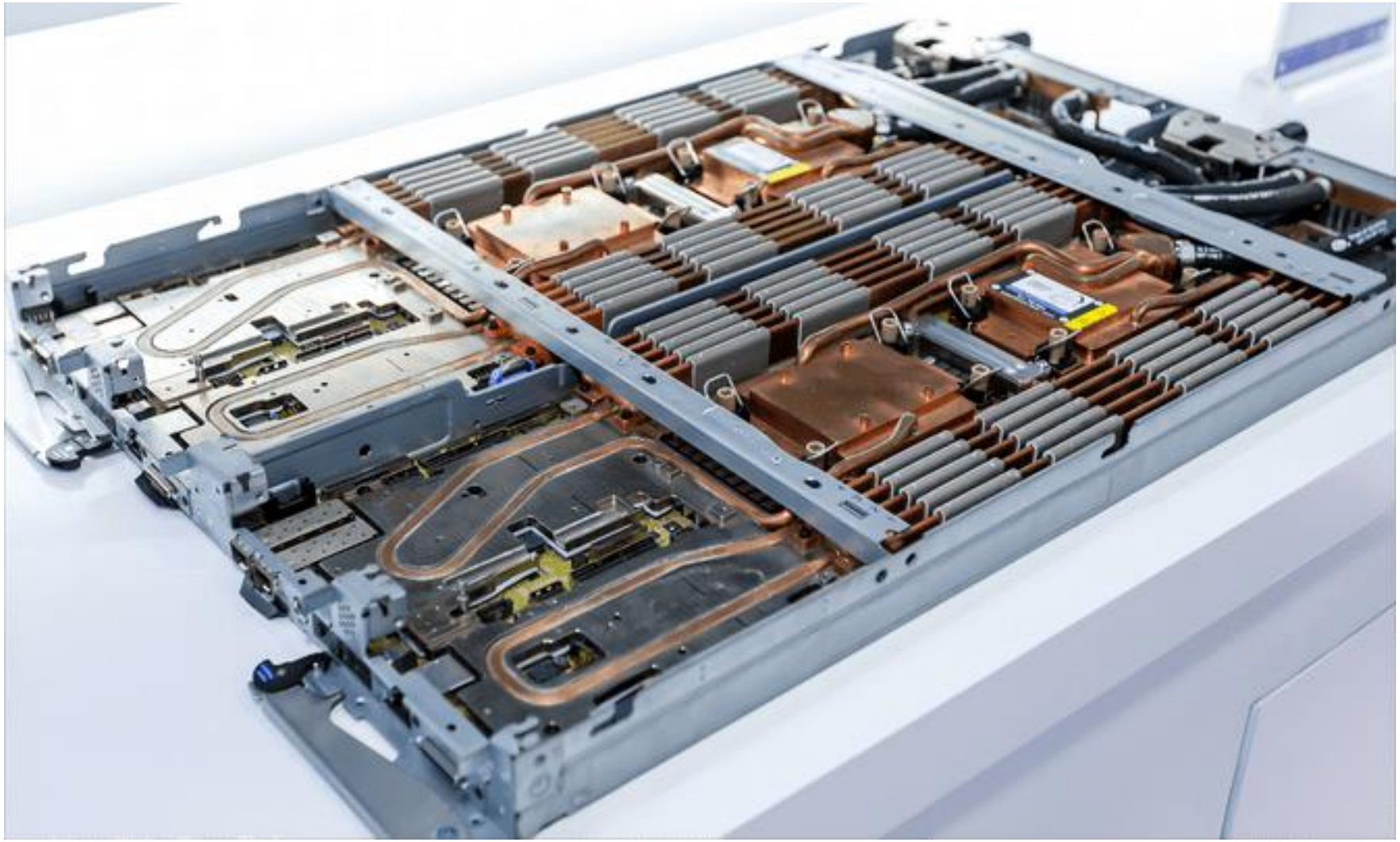
- 冷板贴近热源 xPU，利用冷板中的冷却液带走热量



冷板式液冷

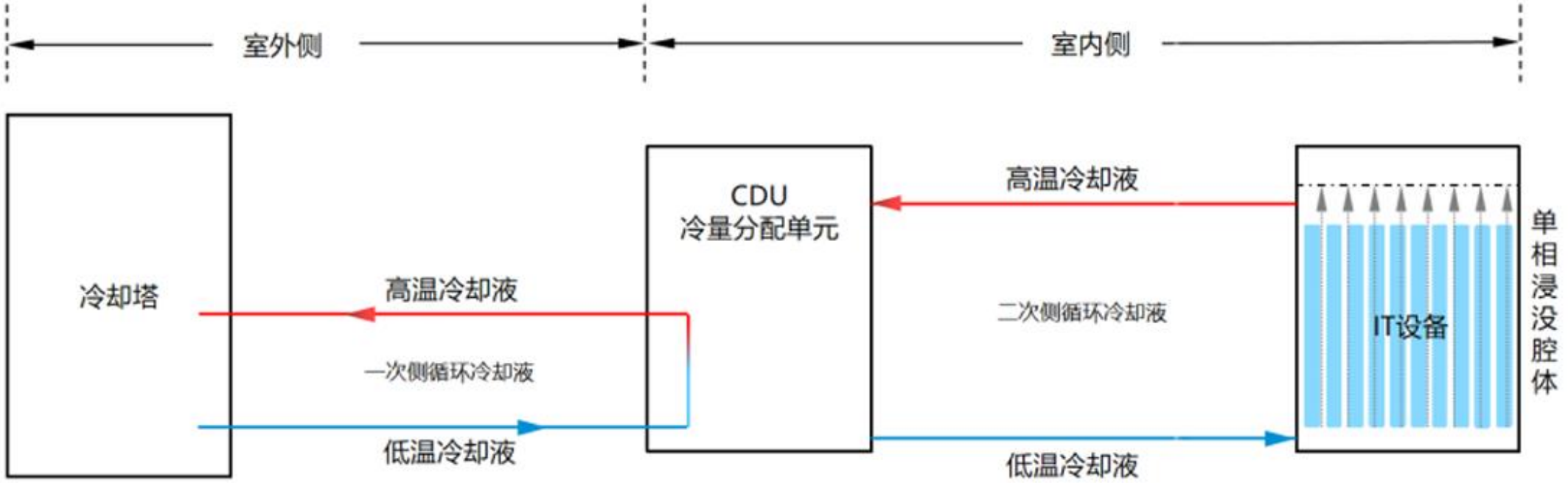
冷板式液冷

冷板式液冷



浸没式液冷

- 单相浸没式液冷的二次侧冷却液在热交换过程中不发生相态变化，仅依靠物质的显热变化进行热 量传递。

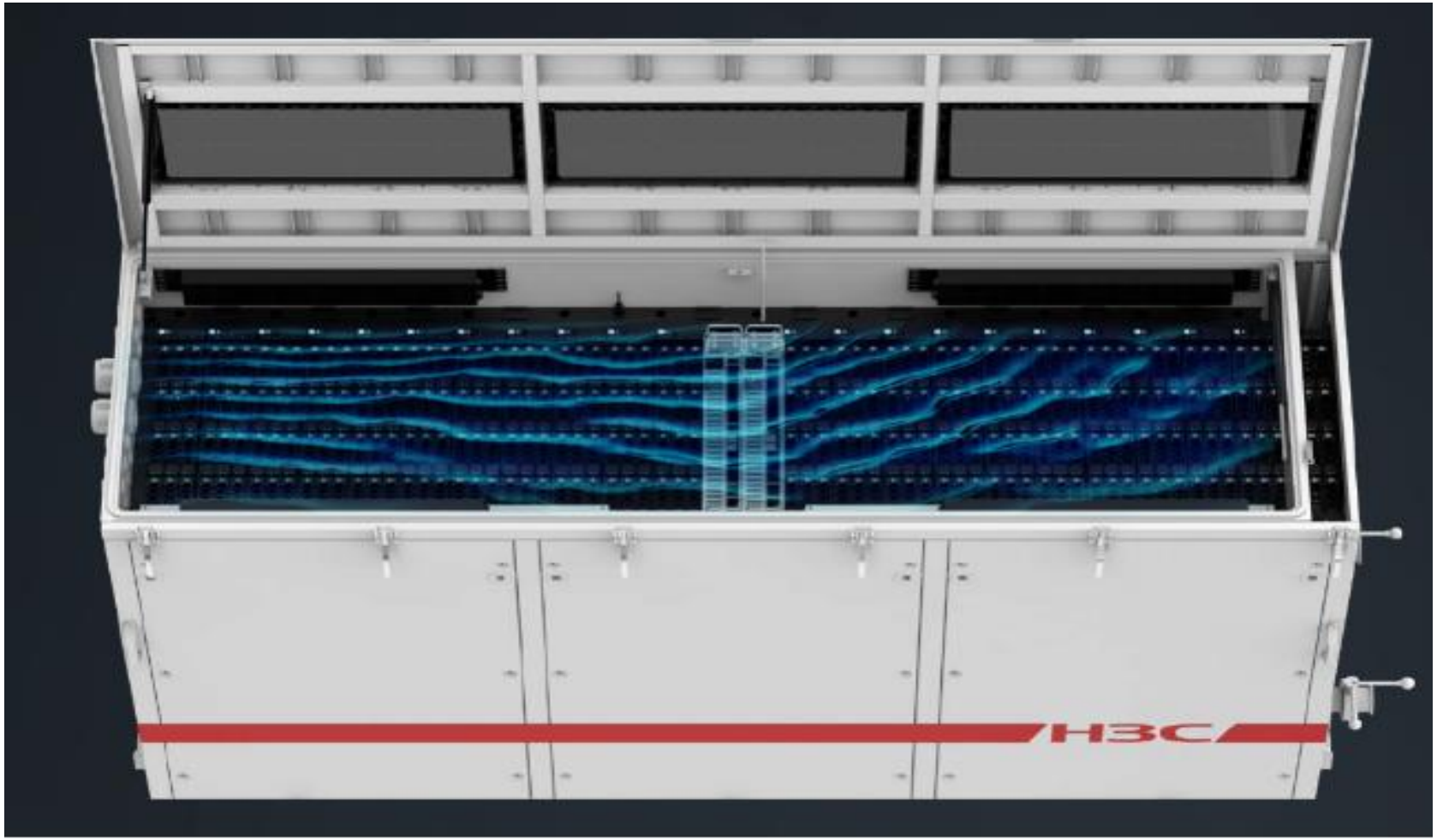


浸没式液冷

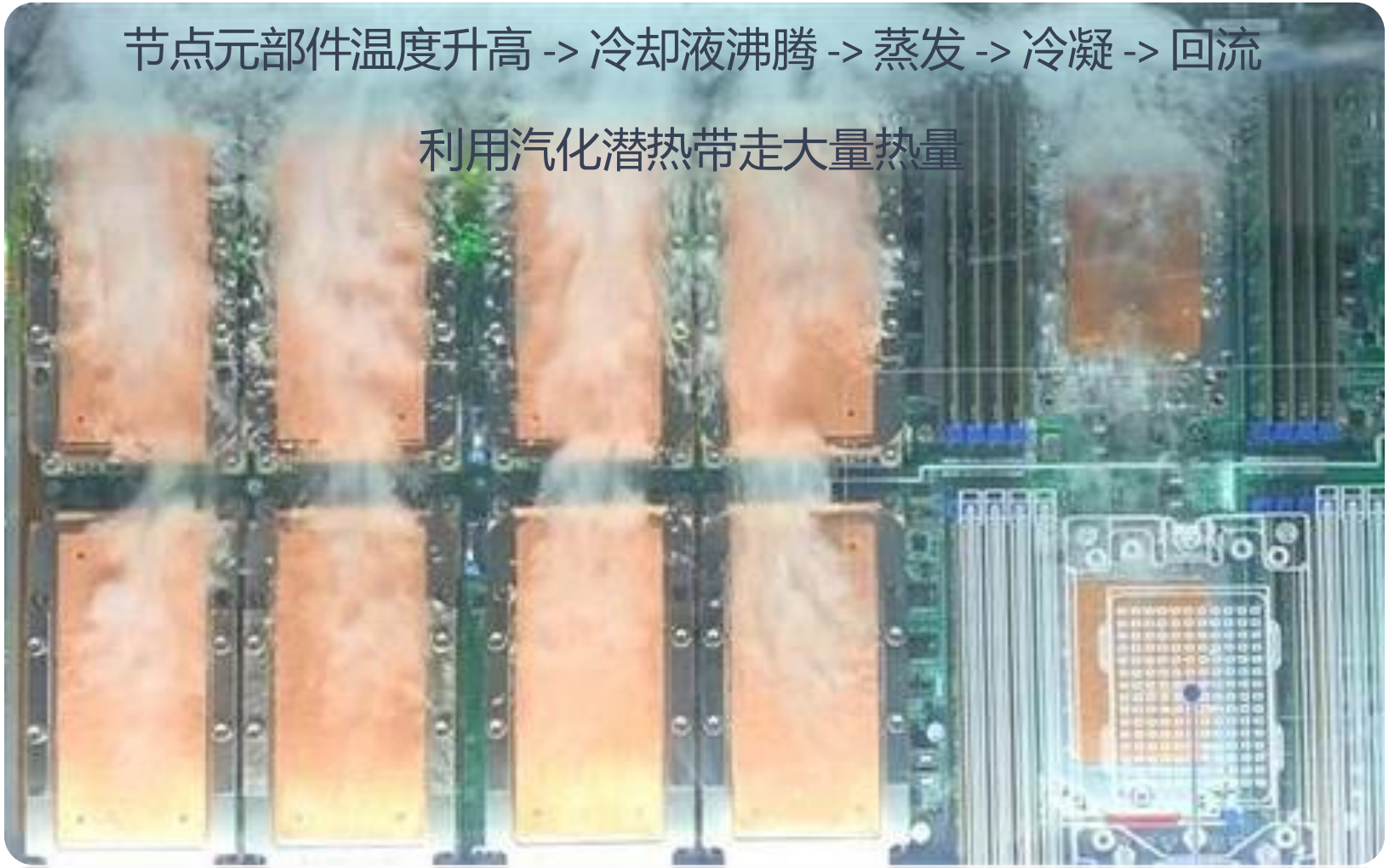
- 服务器完全浸没在冷却液中，冷却液蒸发冷凝相变带走热量

浸没式液冷散热

- H3C 52 台 1U 服务器（交换机 + 服务器）

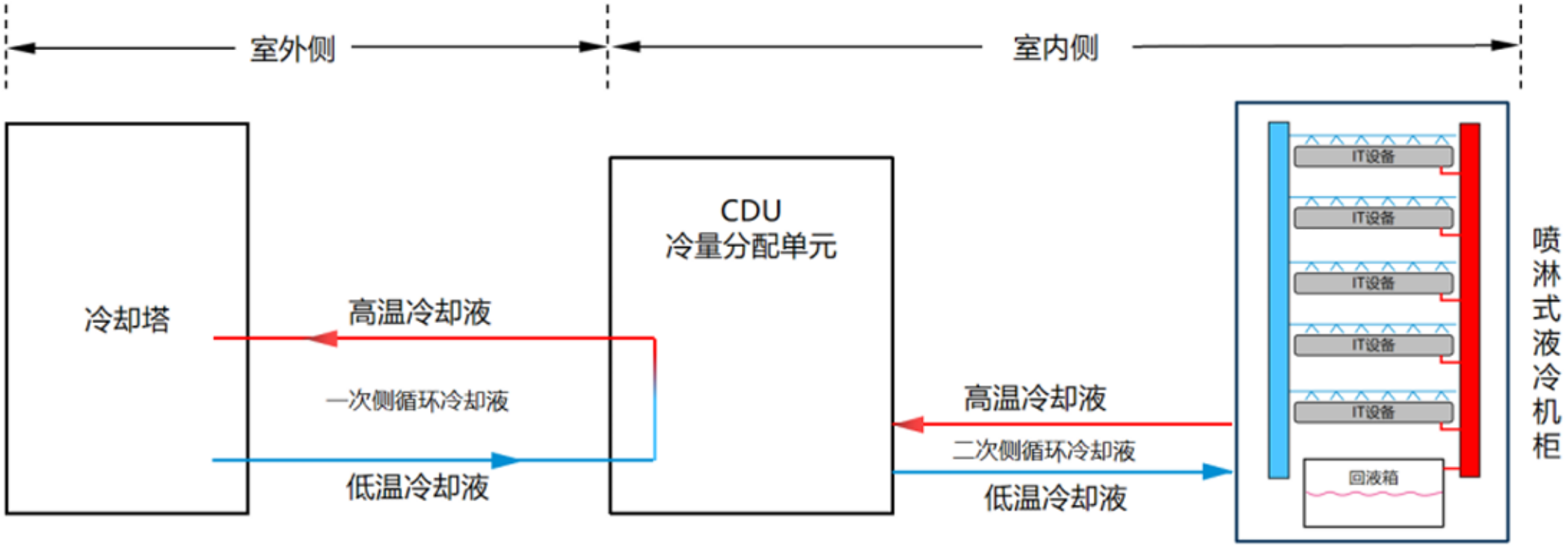


相变浸没液冷散热



喷淋式液冷

- 直接面向芯片级部件，冷却液喷洒至发热器件实现精准冷却。



冷却液不能使用水吧?

- **冷板式冷却液：**冷却液流经安装在芯片上的金属冷板，间接带走热量。
- 使用：去离子水、水乙二醇等单相液体。

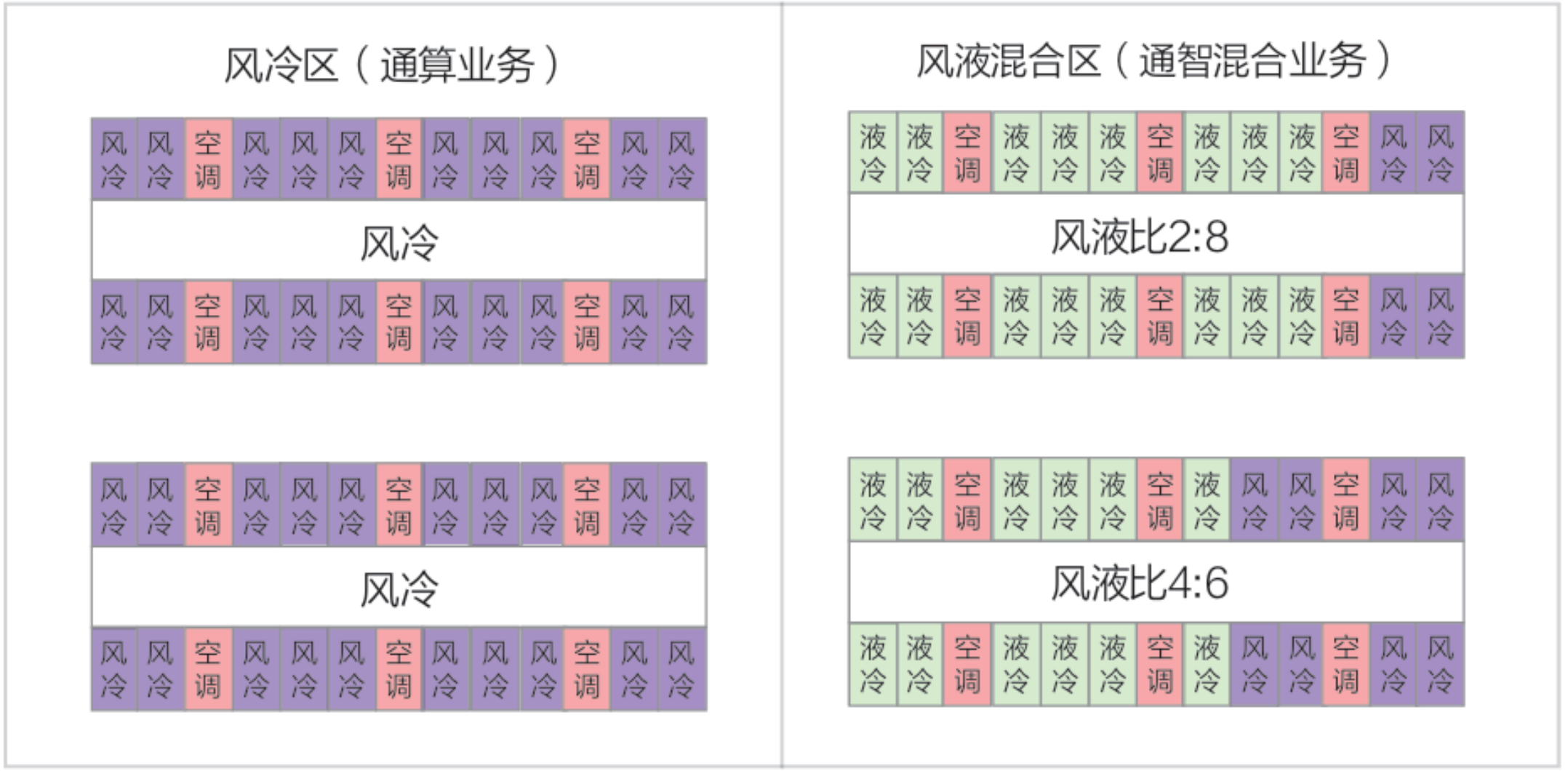
- **单相浸没式冷却液：**冷却液始终为液体，通过泵循环散热。
- 使用：矿物油、硅油等液体。

- **相变浸没冷却液：**冷却液在芯片表面沸腾，蒸汽在顶部冷凝回流。
- 使用： FC-72、FC-87 等低沸点氟化液。

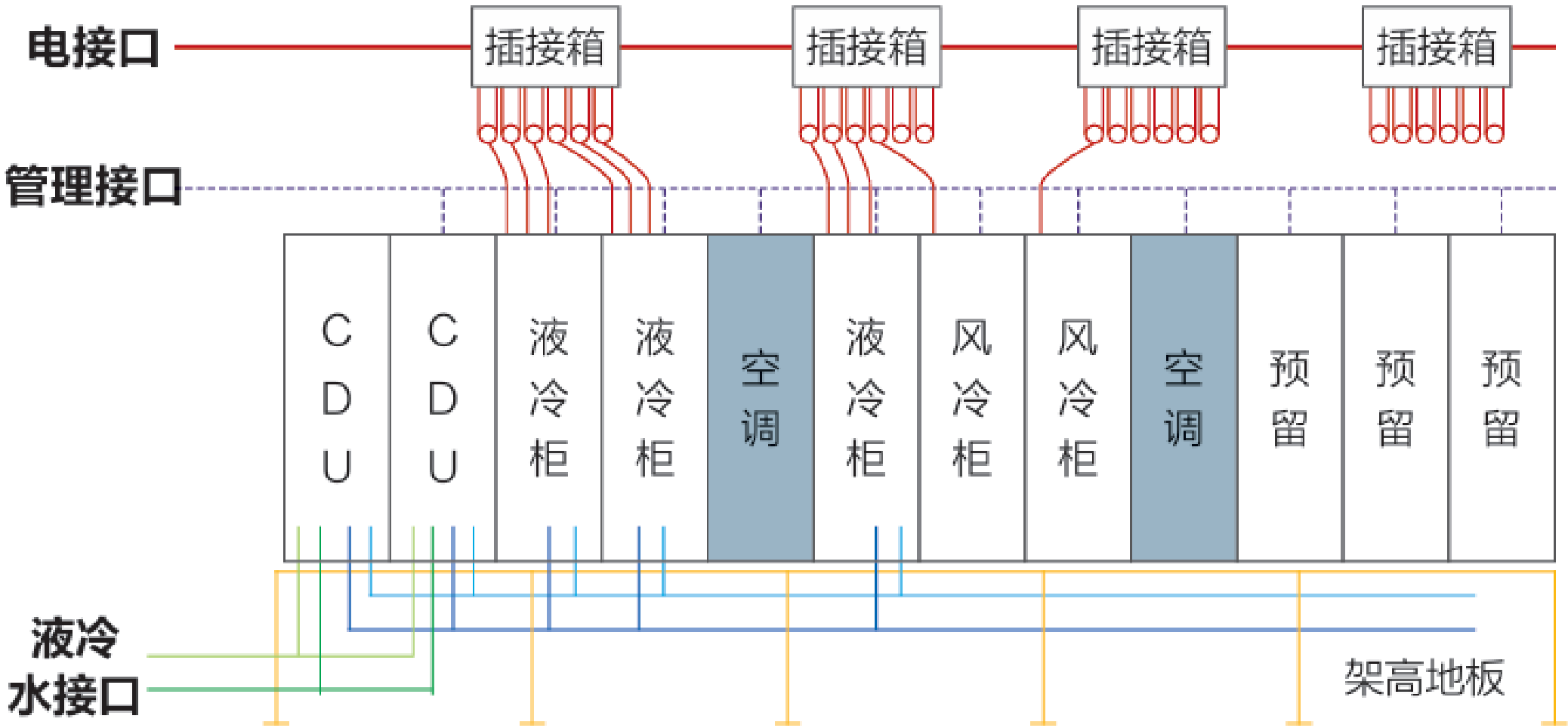
04

风液混合

大型AI DC 机房分区示意图

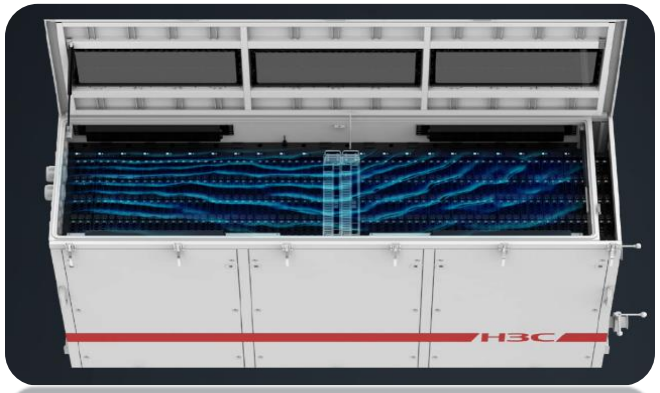


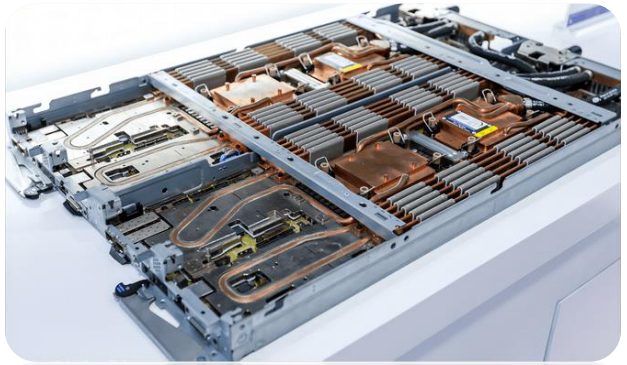
风液混合微模块示意图


电接口
管理接口
液冷
水接口
插接箱
插接箱
插接箱
插接箱
CDU
CDU
液冷柜
液冷柜
空调
液冷柜
风冷柜
风冷柜
空调
预留
预留
预留
架高地板

总结与思考

液冷发展的目的是什么?





通用服务器 vs AI 服务器

- 当前超节点时代，AI DC 功率密度已跃过液冷经济性的临界点，促使行业爆发的核心在于算力和电力的关系，而非仅液冷本身的制造成本。

Thank you

把 AIInfra 带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI Infra to every person, home and organization for a fully connected, intelligent world.

ZOMI

GitHub github.com/Infrasys-AI/AIInfra

Book infrasys-ai.github.io

引用与参考

1. https://navitassemi.com/nvidias-grace-hopper-runs-at-700-w-blackwell-will-be-1-kw-how-is-the-power-supply-industry-enabling-data-centers-to-run-these-advanced-ai-processors/
2. https://www.eet-china.com/mp/a357098.html

PPT 开源在：https://github.com/Infrasys-AI/AIInfra