FRORE SYSTEMS

实现更多

从芯片到冷水机组

热管理技术栈对 AI 工厂能效的核心影响

从芯片到冷水机组

热管理技术栈对 AI 工厂能效的核心影响

FRORE SYSTEMS

摘要

全球正面临电力约束，能源供应能力将成为限制 AI 工厂发展的关键因素。

AI 工厂的核心产出为 Token 算力，是人工智能领域的价值计量基准。因此，AI工厂必须围绕系统级核心能效指标 **Tokens/Watt（每瓦 Token 产出）** 完成全链路优化。

半导体行业有一项核心底层结论：GPU 裸片最高结温会显著 • • 每瓦 Token 产出。**在功耗不变的前提下，GPU 运行温度越低，单位功耗生成的 Token 数量越多。**而决定 AI 工厂中 GPU 芯片温度的核心基础设施，就是 Thermal Stack（热管理技术栈）。

Thermal Stack包含 GPU 散热的完整热传递链路，包括：GPU 封装架构、热界面材料 TIM、冷板、冷却液入口温度、冷却液流量，以及将热量最终释放至环境中的散热系统，包括：自然蒸发冷却塔、机械制冷系统。

本白皮书将深入分析 Thermal Stack 对GPU 最高结温的重大影响，以及其如何进一步影响 AI 工厂的核心效率指标 Tokens/Watt。经过优化设计的 Thermal Stack，可使 AI 工厂 Tokens/Watt指标提升超过 30%。

AI工厂的优化目标，是提升每瓦 Token 产出（Tokens/Watt）指标。

AI 工厂高功耗与电力约束

几乎所有 AI 工厂当前面临的核心运营约束，并不是计算能力、内存带宽或资本投入，而是**电力供应**。

AI 工厂可获得的电力资源是有限的，并且扩展成本高昂。电网接入能力、变压器容量以及实体电力输送基础设施，均成为限制数据中心整体功耗规模的关键因素。在当前环境下，主要数据中心市场的电网容量日益紧张，多个地区的监管机构也开始对能源消耗实施限制。厂内分布式发电（BTM）能够提供一定补充，但电力需求与可用供给之间仍存在巨大的缺口。面对这一电力约束，AI 工厂运营商已经认识到电力是一项战略资产，每一瓦电力都必须被用于产生更多Tokens。

Tokens/Watt每提升一个百分点，都会直接转化为盈利能力的相应提升。这正是为什么 Tokens/Watt成为衡量 AI 工厂效率的核心运营指标。

一套工程设计完善的热管理技术栈，可将 AI 工厂每瓦 Token 产出指标提升 30% 以上。

能源成本是 AI 工厂运营开支（OPEX）的最大组成部分

理解数据中心运营支出的完整构成，对于评估 Tokens输出的真实成本至关重要。

AI 工厂运营开支包含：能源成本、机房设施运维、人力运营、网络与专线通信、硬件维保、软件授权及场地租赁。

其中能源成本占比最高、波动幅度最大，覆盖 GPU、CPU、内存、网络、存储、配电及整套热管理系统的全部耗电。GPU 作为智算工厂主要功耗负载，行业长期将研发、资本投入集中于 GPU 芯片架构迭代。

由于 GPU 是 AI 工厂电力预算中的主要消耗来源，过去整个行业长期将投资和研发重点集中于 GPU 设计优化。通过不断推进制程节点演进，以及向更加针对推理任务优化的 GPU 架构转型，在相同功耗条件下，GPU 性能得到了显著提升，并直接推动系统级 Tokens/Watt指标改善。

但单代 GPU 的能效上限，在芯片架构、制程节点、浮点精度定型后便无法改变。芯片层面的性能提升只能等待下一代硬件迭代，此时热管理技术栈成为优化系统整体每瓦 Token 产出的最关键的• 段。

从芯片到冷水机组

热管理技术栈对 AI 工厂能效的核心影响

GPU 裸片温度对每瓦 Token 产出存在决定性影响

最高结温 $T_{j\,max}$ 指 GPU 裸片内部有源晶体管层的极限工作温度。

该温度与 GPU 封装外壳温度、冷板温度、冷却液温度属于完全不同的物理概念。

裸片是整套热管理链路中的最高温节点：电流流经 GPU 裸片产生功耗，所有电能最终全部转化为热量。

热量必须逐层向外传导：GPU 封装、热界面材料 TIM、冷板、冷却液，最终由机房制冷系统向外排出。当前主流 AI GPU 的额定最高结温普遍为 95℃。

GPU 总功耗分为两类：

1、动态功耗：晶体管运算切换产生的有效功耗，与时钟频率、电压、激活晶体管数量成正比，直接产出算力。

2、漏电流功耗：无效损耗，无论是否执行计算，电流持续穿过晶体管，仅产生热量、不贡献任何算力。

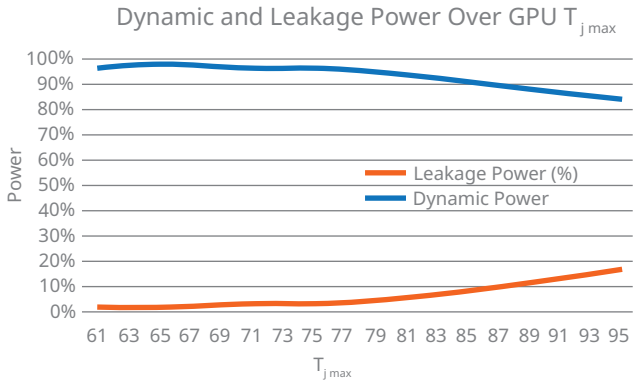


图 1：随着最高结温升高，漏电流功耗指数式上涨

温度与漏电流的指数级恶化循环： GPU 能效高度依赖裸片温度。根据阿伦尼乌斯方程（Arrhenius Equation），裸片最高结温升高时，漏电流功耗呈指数级增长；最高结温 $T_{j\,max}$ 每上升 10℃，漏电流功耗约翻倍，形成恶性循环：高温→漏电流激增→产热加剧→温度进一步攀升。伴随漏电流功耗占比提升，有效动态功耗占比同步下降；同时结温每升高 10℃，晶体管开关所需功耗提升约 2%，同等动态功耗下时钟频率被迫降低，GPU 性能下滑。

工程中通过动态电压频率调节（DVFS）缓解热损耗：温度升高时自动升压维持主频；一旦结温逼近额定上限，则下调电压、降低频率，强制约束 $T_{j\,max}$ 不超限值。

长期在接近最高结温区间运行，会加速芯片内部电迁移与氧化老化，缩短硬件使用寿命，大幅提升万卡集群多年连续运行下的故障概率。下图展示了 GPU 最高结温与能效的完整对应关系。

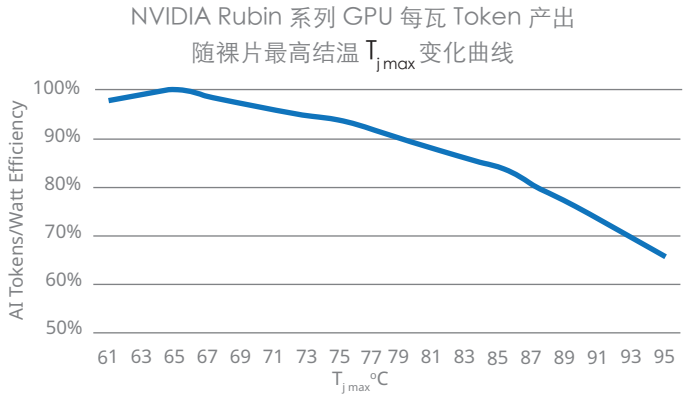


图 2：每瓦 Token 产出随 $T_{j\,max}$ 升高持续衰减

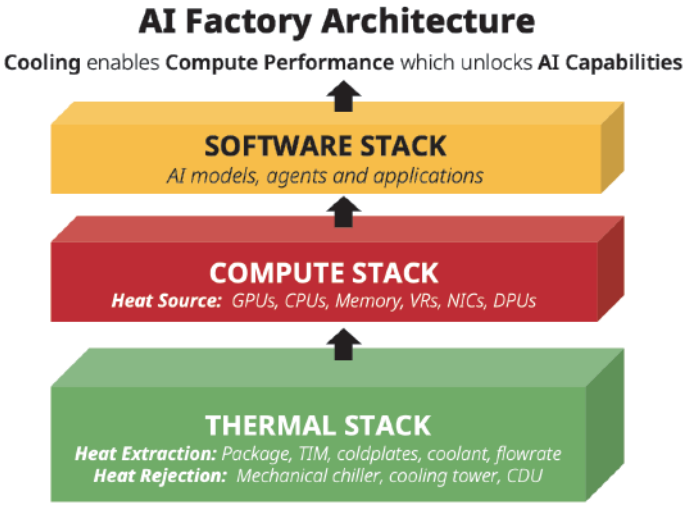
GPU 裸片最高结温每降低 10℃，可实现 15% 的每瓦 Token 产出提升

AI 工厂整体架构

AI 工厂是专为生成 Token 定制的专用系统，整套架构由三层相互耦合的层级构成：

从芯片到冷水机组

热管理技术栈对 AI 工厂能效的核心影响



- **软件栈（SOFTWARE STACK）**
  AI 基础大模型、智能体、各类上层应用程序
- **算力栈（COMPUTE STACK）**
  热源核心：GPU、CPU、内存、网络交换芯片、
  整套配电系统；算力栈消耗的全部电能都会转化为热量
- **热管理技术栈（THERMAL STACK）**
  取热环节：芯片封装、热界面材料、冷板、冷却液流速控制
  排热环节：机械式制冷机组、自然冷却塔、冷分配单元 CDU

热管理技术栈是整套系统的底层基础，是上层所有功能正常运行的前提。若热量无法高效传导并排出，算力栈会触发降频、软件栈算力停滞。热管理系统的效率，直接决定 GPU 可输出的有效算力，以及系统最终达成的 Tokens/Watt数值。

热管理技术栈定义

算力栈产生的热量，必须先由热管理链路带出并传导，再向外部环境排放。

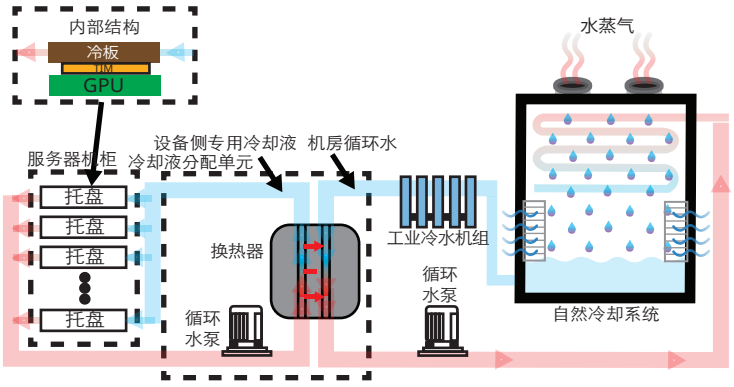


图 3：热管理技术栈完整组件流程图

芯片直冷液冷（DLC）已成为新一代 AI 工厂行业标准：该方案通过液体冷却剂从 GPU 中带走热量。冷却液在与 GPU 直接进行热接触的冷板内部循环流动，实现高效散热。

低温冷却液 (T_{inlet}) 通入冷板，并通过微通道结构带走 GPU 热量，升温后的冷却液 (T_{outlet}) 流出冷板。

升温介质循环进入冷分配单元 CDU，内置换热器将热量交换至机房水路，降温后的介质重新回流冷板。机房水路吸收的热量有两种排放路径：自然冷却：依靠冷却塔蒸发向大气排热，无需机械制冷；机械制冷：启动机械式制冷机组辅助降温。两者可根据当地气候、目标进水温度灵活选用。CDU 内部将高纯度冷却介质与机房水路水力隔离，冷却介质可独立进行化学水质调控，不受机房供水影响。

约束 GPU 性能的核心指标 —— 裸片最高结温 $T_{j\,max}$ ，遵循以下热学方程：

$$T_{j\,max} = T_{inlet} + Q \times R_{total}$$

Q是指GPU 整机功耗 T_{inlet} 是指冷板入口冷却液温度，R_{total} 则是指GPU 裸片到冷却液之间的总热阻，总热阻 R_{total} 由三大设计要素串联叠加，取决于 GPU 封装结构、热界面材料 TIM 选型及冷板设计。

$$R_{total} = R_{package} + R_{TIM} + R_{coldplate}$$

冷却液入口温度 T_{inlet} 由自然冷却部分 T_{free} 和机械制冷部分 $T_{mechanical}$ 共同决定。自然冷却温度取决于湿球温度，而湿球温度由环境温度与相对湿度共同影响，并进一步受 CDU 冷

从芯片到冷水机组

热管理技术栈对 AI 工厂能效的核心影响

却液分配单元）换热器效率影响；机械制冷温度 $T_{mechanical}$ 则取决于机械制冷系统的制冷能力。

$$T_{inlet} = T_{free} + T_{mechanical}$$

为机械制冷机组提供的额外降温幅度。启用机械制冷机组会增加机房整体功耗，用于压低 T_{inlet}；机组性能系数 COP 越高，降温带来的额外能耗损耗越低。总热阻（R_{total}）与冷却液入口温度（T_{inlet}）共同建立了热管理技术栈（Thermal Stack）的物理设计与 GPU 芯片最大结温（$T_{j\,max}$）之间的联系，并最终影响 AI Tokens/Watt（每瓦 Token 产出）表现。

GPU 封装架构

封装热阻 $R_{package}$ 会因 GPU 封装结构采用有顶盖（Lidded）或无顶盖（Un-lidded）设计而产生显著差异。

在有顶盖封装中，集成式散热盖简称 Lid 覆盖于芯片裸晶组件（Die Complex）之上，用于提供机械保护并促进横向热扩散。在散热盖（Lid）与逻辑裸晶（Logic Die）之间，需要插入一层薄型石墨烯导热界面材料（TIM）。由于有顶盖封装增加了额外的热传递层级和界面，其整体热传导路径更长，因此相比无顶盖封装，有顶盖封装具有显著更高的封装热阻。

$$R_{package\ (lidded)} = R_{die} + R_{lid} + R_{graphene\ TIM}$$

$$R_{total\ (lidded)} = R_{die} + R_{lid} + R_{graphene\ TIM} + R_{TIM} + R_{coldplate}$$

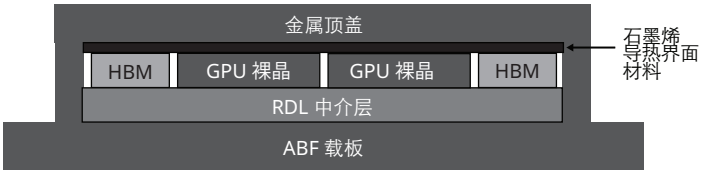


图 4：带金属均热盖封装结构剖面示意图

无顶盖封装取消了传统封装中的散热盖以及石墨烯导热界面材料，减少了热传递路径中的额外界面与热阻。因此，在相同功耗热负载和冷却液入口温度（T_{inlet}）条件下，无顶盖封装能够显著降低总热阻（R_{total}），从而实现更低的 GPU 芯片最大结温（$T_{j\,max}$）。

$$R_{package\ (un\text{-}lidded)} = R_{die}$$

$$R_{total\ (un\text{-}lidded)} = R_{die} + R_{TIM} + R_{coldplate}$$

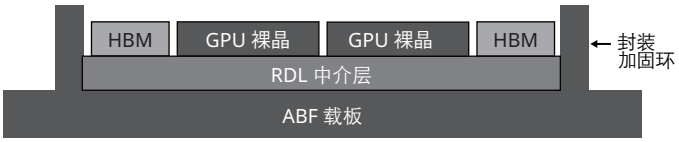


图 5：无金属均热盖封装结构剖面示意图

以 NVIDIA Rubin系列 GPU 为例：带盖、无盖封装的裸片最高结温差值最高可达 20℃，对应每瓦 Token 产出最高提升 35%。

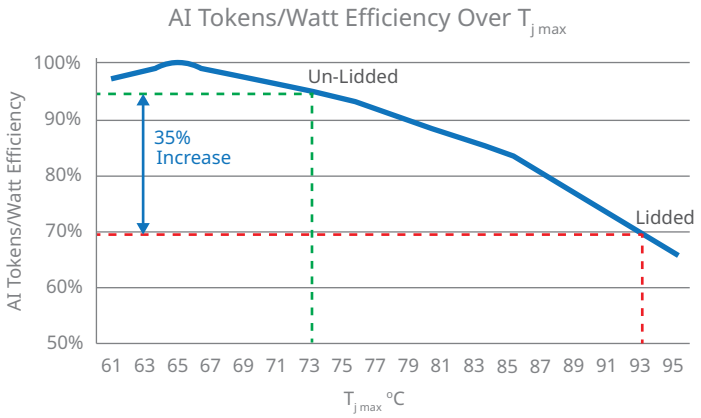


图 6：由于无顶盖封装具有更低的热阻，在相同 GPU 功耗条件下，其能够支持更低的最大结温，从而实现更高的每瓦 Token 产出效率。

然而，无顶盖封装也存在一定的可靠性权衡。在无顶盖封装中，裸露的芯片裸晶直接暴露于外部环境，因此在搬运过程以及冷板安装过程中，发生芯片边缘崩裂或裂纹的风险高于有顶盖封装。此外，散热盖（Lid）能够提供一个精确且高度一致的平整接触表面。取消散热盖后，冷板可能会面对芯片表面不同区域之间存在的非均匀 Z 高度问题，从而影响冷板与芯片之间的接触一致性以及整体散热性能。

导热界面材料

导热界面材料（TIM）用于 GPU 封装与冷板（Coldplate）之间，以确保两者之间形成良好的热接触。然而，TIM 本身也会增加热传递路径中的热阻。目前行业中常用的 TIM 材料主要分为三类：

从芯片到冷水机组

热管理技术栈对 AI 工厂能效的核心影响

TIM 类型	典型热阻值	备注
信越导热硅脂	0.09°C cm²/W	成本最低，热阻偏高，返修操作便捷
霍尼韦尔 PTM7950相变片	0.06°C cm²/W	热阻适中，稳定性优异
液态金属铟	0.03°C cm²/W	热阻最低，但封装防护要求高，需规避长期腐蚀渗漏

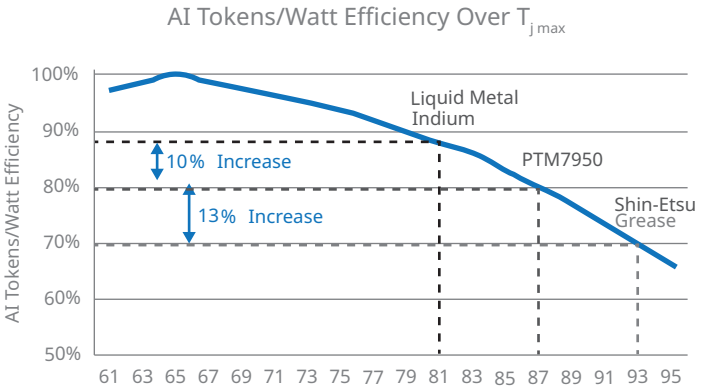


图 7：由于不同导热界面材料具有不同的热阻特性，在相同 GPU 功耗条件下，低热阻 TIM 可支持更低的最大结温，从而实现更高的 Tokens/Watt能效。

液态金属铟可应用于有顶盖封装，通过在散热盖顶部及冷板底部进行镀金处理，实现稳定、高效的热界面连接。因此，采用液态金属铟 TIM 可以部分缓解有顶盖封装因封装结构导致的热阻劣势。NVIDIA Rubin GPU正采用这一技术路线。

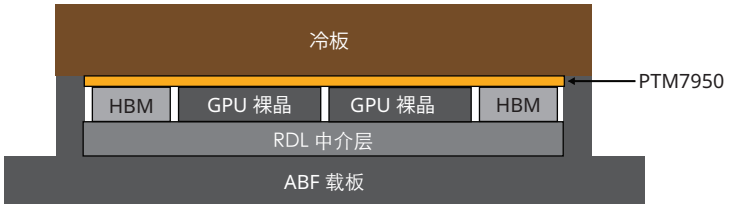


图 8：采用 PTM7950 TIM 的无顶盖封装设计结构示意图

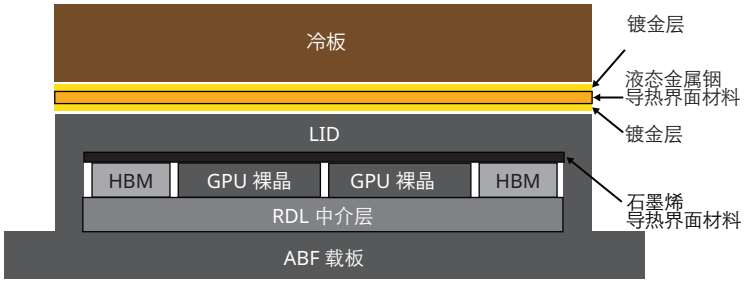


图 9：搭载液态金属铟导热界面材料的 NVIDIA Rubin GPU有顶盖封装设计

然而，$R_{package\ (un\text{-}lidded)} + R_{PTM7950} < R_{package\ (lidded)} + R_{Liquid\ Metal\ Indium}$

两种封装方案之间的最大结温 $T_{j\,max}$ 差异最高可达 14°C，并由此导致 Tokens/Watt能效差异最高达到 28%。

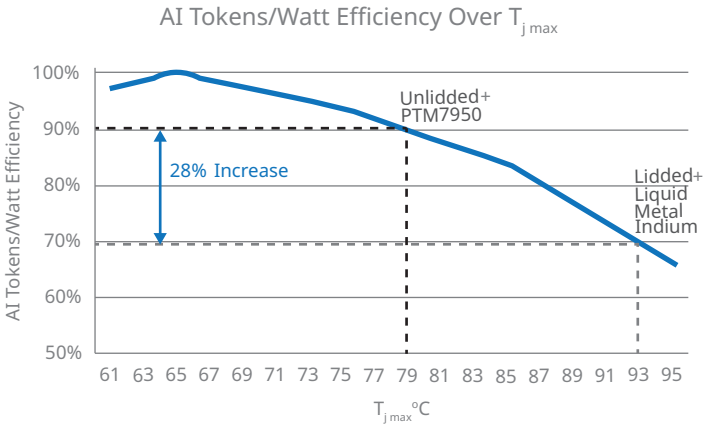


图 10：在相同 GPU 功耗条件下，采用 PTM7950 TIM的无顶盖 GPU 封装方案，相较于采用液态金属铟 TIM 的有顶盖 GPU 封装方案，可实现更高的能效

冷板设计

冷板设计在降低总热阻（R_{total}）方面发挥着关键作用。冷板热阻（$R_{coldplate}$）由三个相互影响的设计变量共同决定：微通道结构设计、流体路径设计、流量。

微通道结构设计

微通道长度、鳍片间距、鳍片高度以及深宽比共同决定冷板壁面，与流经冷板内部的冷却液之间的对流换热系数。

- 窄的通道以及更大的深宽比能够增加单位体积内的换热面积，从而提升热传递效率；然而，较长的微通道会由于冷却液流动逐渐进入充分发展流态，导致换热效率下降。

因此，通过设计超短长度微通道（<1 mm，Short-loop），可以显著提升换热效率，使冷却液始终保持在流动发展阶段，从而增强热交换能力。

为了实现短循环微通道，需要采用三维微单元结构。然而，采用传统机械加工工艺（例如 Skiving铲齿工艺）无法制造这种三维微单元结构。

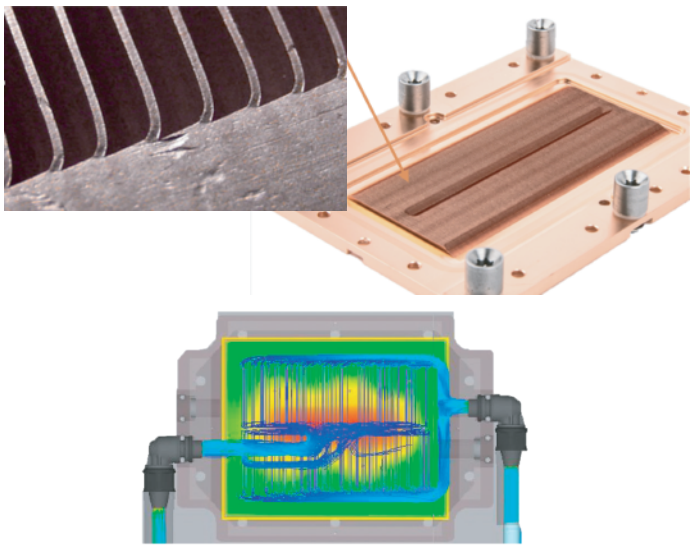


图 11：采用分流式冷板设计的铲齿微通道结构

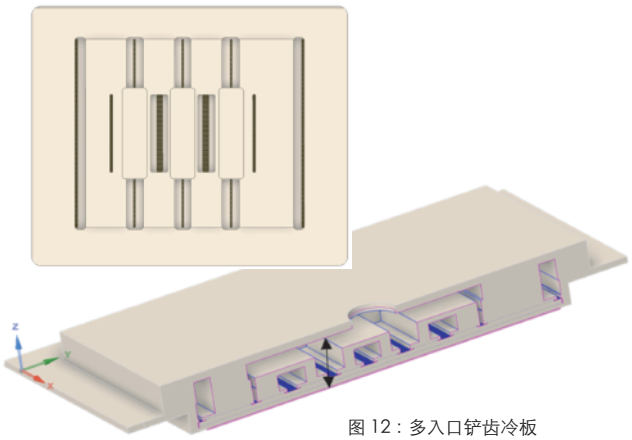


图 12：多入口铲齿冷板

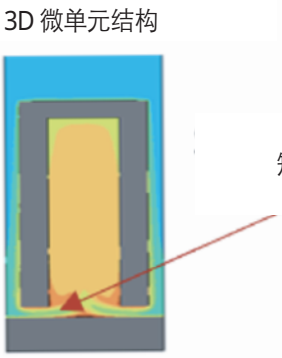


图 13：3D 微单元结构

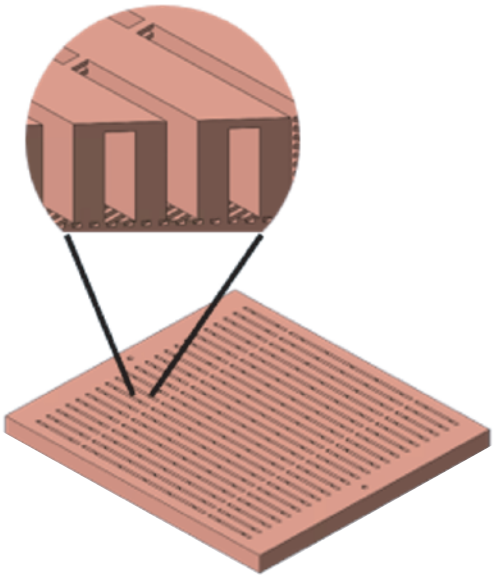


图 14：搭载 3D 微单元阵列的冷板整体结构图

3D 微单元、短回路流道与多级流道结构结合，可打造换热效率大幅提升的冷板方案

流体路径设计

进入冷板的冷却液需要被分配至各个微通道中。最简单的流体设计方式是：冷却液通过冷板一端的单一入口进入，并均匀分布至所有微通道，随后在另一端汇集并流出。然而，这种设计可能会产生局部热点，原因在于：高功耗芯片区域获得的冷却流量与芯片其他区域相同，无法针对更高热负载区域提供额外散热能力。更严重的是，高功耗芯片区域可能位于微通道下游位置，而此处的冷却液已经在上游吸收了一部分热量，并进入充分发展流态，导致其进一步吸热能力下降，从而加剧局部温度升高。

一种相对更优的流体路径设计方式是“分流式结构”。在这种设计中，冷却液从冷板两端的两个入口进入，并被均匀分配至微通道中；吸收热量后的温热冷却液则在冷板中央的出口处汇集并排出。

这种设计将有效微通道长度缩短了一半，从而提升了换热效率。通过在微通道上方覆盖一个简单的歧管结构，可以实现这种流体路径分配方式。这一方法的自然延伸是多入口冷板。在多入口冷板设计中，通过压力平衡机制，对多个入口之间的冷却液流量进行加权分配，使更多冷却液优先流向GPU裸晶中功率密度更高的区域，从而更有效地冷却热点区域。

多级结构

然而，在上述三种冷却液流路设计方案中，冷却液都只被利用一次，并不会被循环重复利用。当进入冷板的冷却液被分配到微通道后，会沿线性方向仅流经一次微通道，并最终在出口处汇集排出。

相比之下，多级冷板能够实现冷却液的多次利用。多级冷板将整个冷板划分为多个区域，每个区域依次接收完整的冷却液流量。例如，在一个三级冷板中：全部冷却液会首先进入第一级冷板；离开第一级后，冷却液继续依次进入第二级，随后进入第三级。每一级的设计都基于 GPU 裸晶功率分布图进行优化，使前级区域对应 GPU 裸晶中功率密度最高的区域，从而优先对高热负载区域进行散热。

需要注意的是，多级结构会增加冷板压降，而短循环流道能够降低压降。通过将这两种设计特性结合，可以实现更加高效的冷板方案，在不增加冷板压降的情况下，大幅降低冷板热阻（$R_{coldplate}$），从而实现更高效的散热性能。

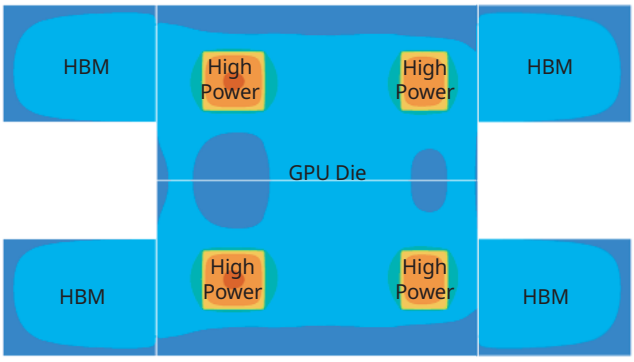


图 15：包含四个高功率密度区域的 GPU 裸晶功率分布图

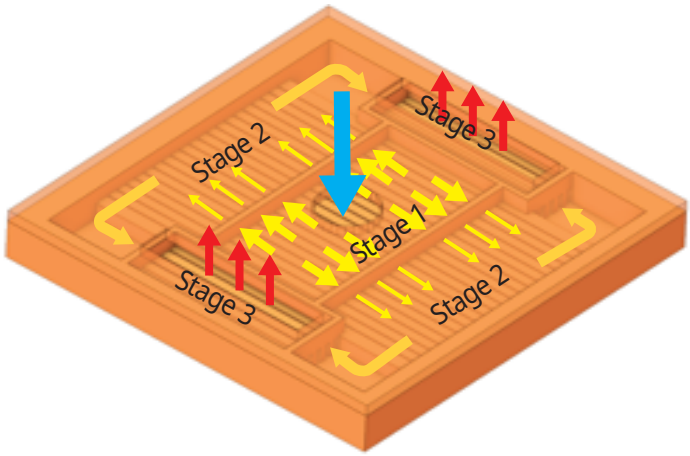


图 16：采用多级结构设计的冷板

冷却液

在芯片直冷液冷系统中，通常使用两类冷却液：去离子水热容为 4.18 J/ml/°C。PG25（25% 丙二醇 + 75% 水）热容为 3.85 J/ml/°C。当冷却液流量为 1 LPM 时，使 1 升冷却液升高 1°C 所需吸收的热量分别为：去离子水为 69.67 W，PG25为64.17 W

流量

当冷板设计确定后，流量成为影响散热性能的主要可调参数。提高流量可以降低冷板热阻（$R_{coldplate}$），但两者并非线性关系：在低流量阶段，提高流量能够显著降低热阻；随着流量逐渐增加并接近冷板结构所决定的性能极限后，热阻下降幅度逐渐减小，每增加单位流量带来的散热收益递减。

更高流量也会增加压降以及泵功耗。根据泵相似定律，泵功耗与流量呈三次方关系。

在实际应用中，相比 GPU 和机械制冷系统的整体功耗，增加流量所带来的泵功耗增加通常较小。因此，将流量提升至热阻下降趋缓的临界点，通常能够获得更高的系统效率。

不过，实际运行流量通常受数据中心 CDU 容量、歧管、管路和快速接头规格限制，以确保系统可靠运行并降低泄漏风险。

从芯片到冷水机组

热管理技术栈对 AI 工厂能效的核心影响

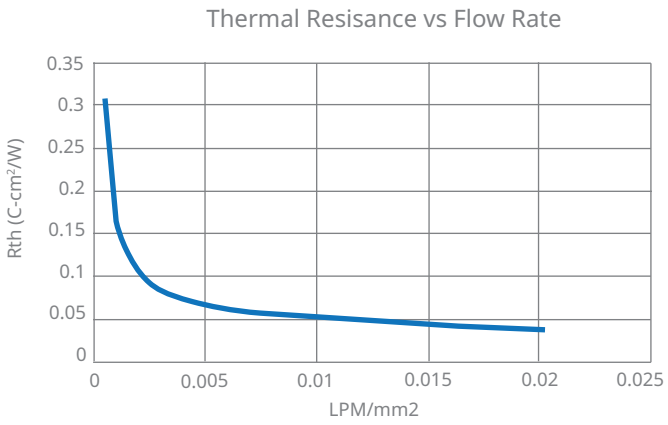


图 17：三维微单元冷板热阻随流量变化曲线

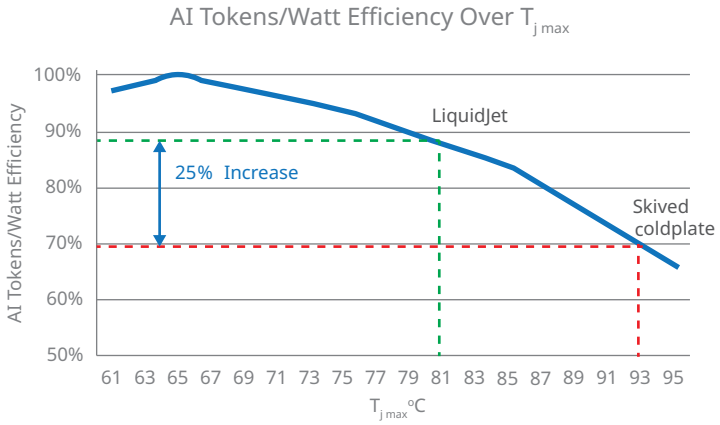


图 18：同等功耗下，LiquidJet冷板对比传统铲齿冷板，每瓦 Token 产出提升 25%

制造工艺

制造工艺决定了冷板设计能够实现的结构和性能边界。

目前，规模量产冷板普遍采用传统刨削加工工艺制造。在这种工艺中，鳍片直接从整块铜材上一体刨削成型，形成连续、细长且高宽比的鳍片结构。

结合歧管设计，刨削加工工艺能够制造传统微通道冷板、分流式冷板以及多入口冷板。然而，对于短循环微通道、三维微单元和多级流路等先进冷板结构，传统刨削加工工艺已无法实现，因此需要采用全新的制造技术。

Frore Systems创新性地将半导体制造工艺应用于铜晶圆加工，开发出独有的 LiquidJet® 冷板制造技术。LiquidJet® 冷板采用了窄型喷射流道、基于三维微单元的短循环微通道、多级流路等多项创新设计，并针对 GPU 高功率密度区域进行优化，从而显著降低冷板热阻。

Frore Systems首先根据 GPU 功率分布图，为不同 GPU 定制优化的 LiquidJet® 冷板 CAD 设计。

随后，通过光罩制作、光刻、湿法蚀刻以及键合等成熟的半导体制造工艺，实现 LiquidJet® 冷板的大规模、低成本制造，同时确保产品具备经过验证的高可靠性。

与传统冷板相比，LiquidJet® 冷板可将 GPU 最大结温降低 6–12°C，从而使 Tokens/Watt提升 10%–25%。

将半导体晶圆加工工艺应用于铜基材，打造结构复杂的 LiquidJet® 冷板

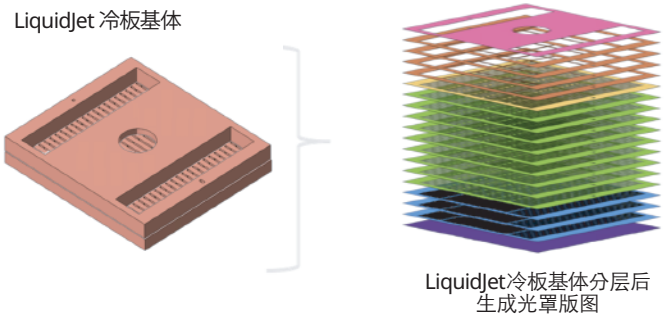


图 19：LiquidJet® 冷板 CAD →光罩版图生成流程

目前，其他厂商也在尝试采用金属 3D 打印等增材制造技术，制造类似的短循环冷板。然而，这些制造方法在设计灵活性、可靠性以及大规模量产能力方面，仍有待进一步验证。

从芯片到冷水机组

热管理技术栈对 AI 工厂能效的核心影响

冷却液入口温度

冷却液入口温度（T_{inlet}）与 GPU 最大结温（$T_{j\,max}$）呈直接的一一对应关系。在 GPU 功耗（Q）和总热阻（R_{total}）保持不变的情况下，T_{inlet} 每降低 1°C，$T_{j\,max}$ 便会降低 1°C。而 $T_{j\,max}$ 每降低 1°C，都会进一步提升 GPU 的 Tokens/Watt能效。

NVIDIA为 Rubin GPU AI工厂平台规定的冷却液入口温度为 45°C。这一温度设定的重要意义在于数据中心整体能耗：当入口温度达到 45°C 时，在全球大多数地区的数据中心，仅依靠自然冷却即可提供所需冷却能力，无需启动机械制冷系统。

然而，这并不意味着所有部署场景都应采用 45°C 的入口温度。

对于部分应用负载而言，较高的入口温度已经能够满足运行需求；但对于部署在 AI 工厂中的 GPU，由于 Tokens/Watt 对 GPU 最大结温高度敏感，因此仍建议尽可能采用较低的入口温度，以获得更高的算力能效。需要注意的是，当入口温度低于自然冷却所能提供的最低温度时，就必须依赖机械制冷系统进行降温。机械制冷系统本身耗电较高，会增加数据中心整体能耗，并可能降低 AI 工厂整体的 Tokens/Watt 能效。因此，只有当机械制冷机具有足够高的 COP（性能系数） 时，进一步降低入口温度所带来的收益才具有实际价值。

机械制冷系统对 Tokens/Watt的影响

当需要将冷却液入口温度控制在自然冷却温度以下时，必须通过机械制冷系统提供额外制冷能力。

机械制冷机的效率通常采用性能系数（COP）进行衡量，其定义为：

$$COP = Q / W$$

其中，Q 指制冷冷量，W 指机械制冷机消耗的电功率，例如，当 COP = 5 时，表示机械制冷机每消耗 1 千瓦电力，能够移除 5 千瓦热量。假设通过机械制冷系统将冷却液入口温度降低 1°C，那么 GPU 的最大结温（$T_{j\,max}$）也将同步降低 1°C。以 GPU 最大结温由 89°C 降至 88°C 为例，这一温度下降可带来约 1.5% 的 Tokens/Watt能效提升。
因此，在评估是否进一步降低冷却液入口温度时，需要综合考虑温度下降带来的算力能效提升与机械制冷增加的能耗成本之间的平衡。只有当机械制冷系统具备较高的 COP 时，降低入口温度才能带来更优的整体能效和投资回报。

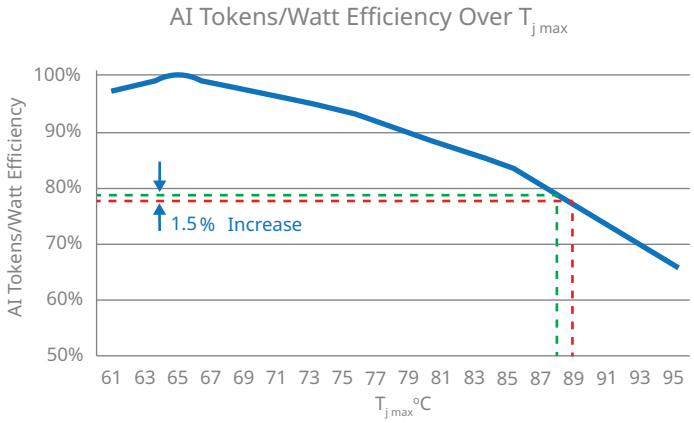


图 20：GPU 最大结温每降低 1°C，AI Tokens/Watt效率提升 1.5%

然而，降低冷却液入口温度（T_{inlet}）并非没有代价，因为机械制冷系统本身需要消耗电力。当冷却液采用去离子水时，每降低 1°C 所需的机械制冷功率为：

$$\text{Chiller Power (W) for 1°C reduction} = (LPM * 69.67/COP)$$

假设 AI 工厂的总功率预算保持不变，那么分配给机械制冷系统的电力越多，可分配给 GPU 的电力就越少。另一方面，GPU 最大结温（$T_{j\,max}$）每降低 1°C，即可提升约 y% 的 Tokens/Watt能效（例如，Tj max从 89°C 降至 88°C 时，Tokens/Watt可提升约 1.5%）。因此，只要机械制冷系统的 COP 足够高，即使部分电力从 GPU 转移到制冷系统，整体 Tokens/Watt仍有可能获得提升。维持 GPU Tokens/Watt不增不减所需的最低 COP 称为盈亏平衡性能系数（$COP_{breakeven}$），其计算公式如下：

$$COP_{breakeven} = \frac{69.67 * (1 + y) * LPM}{x * y}$$

对于受电力限制的 AI 工厂而言，只有当机械制冷系统的 COP 高于 $COP_{breakeven}$ 时，将部分电力从 GPU 分配给机械制冷系统才具有实际意义。此外，在相同 GPU 功耗条件下，冷板所需的冷却液流量越低，$COP_{breakeven}$ 也越低。这意味着，高散热效率的冷板能够降低机械制冷系统达到经济运行条件的门槛。

以 NVIDIA Rubin GPU Max-P（2300W） 为例：

采用传统铲齿冷板时，为维持相同的 GPU 最大结温，预计需要 3.25 LPM 的冷却液流量，对应的 $COP_{breakeven}$ 为 6.7。

从芯片到冷水机组

热管理技术栈对 AI 工厂能效的核心影响

相比之下，采用 LiquidJet® 3D 微单元短循环多级冷板时，仅需 2 LPM 的冷却液流量即可维持相同的 GPU 最大结温，因此 $COP_{breakeven}$ 降低至 4.1。

因此，像 LiquidJet® 这样具有更高散热效率的冷板设计，通过降低冷却液流量需求，不仅提升了冷板本身的散热性能，也使机械制冷系统更容易实现经济高效运行，从而进一步提升 AI 工厂整体的 Tokens/Watt能效。

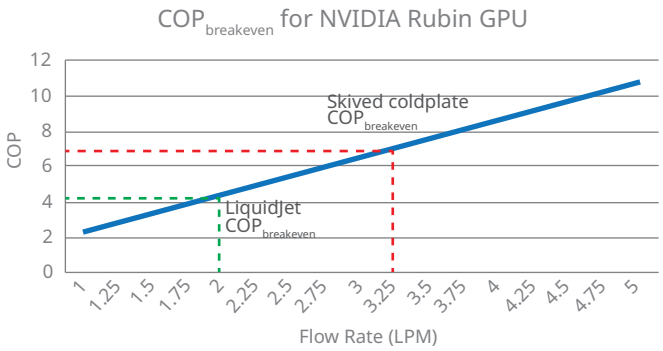


图 21：LiquidJet®通过降低所需流量实现更低的 $COP_{breakeven}$

本文论证：GPU 裸片最高结温（$T_{j\,max}$）并非芯片设计的被动结果，而是可主动工程调控的变量。$T_{j\,max}$与每瓦 Token 产出存在直接线性关联，由热管理技术栈多层级联动设计共同决定：GPU 封装架构、热界面材料选型、冷板设计、冷却液进水温度、介质流速、机械式制冷机组部署。

LiquidJet®冷板单独即可实现 GPU 最大结温降低 6–12°C，并将 Tokens/Watt能效提升 10–25%。

结论

AI 工厂本质上是一个受电力约束的计算系统，而衡量其运营效率的核心指标——Tokens/Watt，直接决定了每一项 AI 基础设施部署的盈利能力。

当 GPU 完成制造并部署之后，其芯片本身的能效已经固定，无法再改变。此时，能够持续提升 Tokens/Watt的关键变量，只剩下 Thermal Stack（热管理技术栈） 的设计与配置。

超大规模云服务商若将热管理技术栈作为一级核心设计维度，同等资本与运营投入下，可显著提升每瓦 Token 产出

热管理技术栈每一层结构都会引入额外热阻、占用机房电力预算，每一层都具备独立优化空间。GPU 封装架构会在 Thermal Stack中引入显著热阻。采用盖板封装的设计包含石墨烯 TIM 和盖板，会增加 $R_{graphene}$ 与 R_{Lid} 热阻；而无盖板封装设计则可以消除这两部分热阻，但会增加在运输、操作和安装过程中的机械风险。

热界面材料的选择决定了热量从封装传递至冷板的效率。其中，液态金属铟 TIM 可实现最低热阻。无盖板封装结合合适的 TIM 选择，最多可使 Tokens/Watt提升 25%。

Frore Systems的 LiquidJet冷板设计，通过采用半导体制造工艺，实现传统刨削加工工艺无法制造的三维微单元短循环结构和多级流路设计，从而显著降低 $R_{coldplate}$。

仅 LiquidJet冷板即可将 GPU 最大结温降低 6–12°C，使 Tokens/Watt指标提升 10–25%。

冷却液入口温度（T_{inlet}）的设定，以及是否启用机械制冷系

统，最终是一个系统级功率分配问题，其取决于机械制冷机 COP、冷却液流量，以及每降低 1°C GPU 最大结温所带来的 Tokens/Watt增益。

AI 基础设施行业正在逐渐认识到这一现实：将 Thermal Stack 作为核心设计体系的超大规模云服务提供商，将能够在相同资本投入（CAPEX）和运营投入（OPEX）下，获得显著更高的 Tokens/Watt产出。

随着 GPU 功率密度持续提升，从 NVIDIA Rubin延伸至未来架构，经过热优化的部署方案与未经优化的部署方案之间的性能差距将进一步扩大。

Thermal Stack并非辅助型基础设施，而是一套决定系统性能的核心体系。在电力受限的时代，它是优化 AI 工厂 Tokens/Watt输出的关键。

一套工程化完善的热管理技术栈，最高可将AI工厂每瓦 Token 产出提升 30% 以上。

热管理技术栈（Thermal Stack）并非辅助性基础设施，而是一套决定系统性能的核心体系。在电力受限的时代，它是优化 AI 工厂 Tokens/Watt产出的关键。

从芯片到冷水机组

热管理技术栈对 AI 工厂能效的核心影响

术语表

AI Factory AI 工厂

专为承载人工智能工作负载搭建的规模化数据中心，核心产出为AI生成结果（Token），区别于承载传统通用计算任务的数据中心。

Arrhenius Equation阿伦尼乌斯方程

用于描述热激活反应速率随温度呈指数增长的数学公式。在GPU功耗分析场景中，该方程解释了裸片最高结温升高时，漏电流功耗急剧攀升的物理底层原理：芯片结温每升高10℃，漏电流功耗约翻倍。该公式由瑞典化学家斯万特·阿伦尼乌斯命名。

Behind-the-Meter (BTM) Power Generation厂内分布式发电

部署于数据中心场地内部的发电设施，包含光伏组件、备用发电机、燃料电池等设备。“厂内分布式”指电力就地生产、就地消纳，不经过公共电网计量设备，可降低机房对电网容量的依赖，弥补智算工厂用电需求与市政供电之间的供需缺口。

Coldplate冷板

经精密工艺加工的金属散热构件，直接贴合GPU芯片表面，通过内部微通道流动的液态冷却介质持续带走芯片热量，是液冷散热系统的核心换热部件。

COP (Coefficient of Performance性能系数

衡量机械式制冷机组能效的核心指标。COP数值为 5 时，代表机组每消耗1千瓦电能，可移除 5 千瓦的设备热量，数值越高代表制冷能效越优异。

$COP_{breakeven}$ COP 盈亏平衡性能系数

机械式制冷机组的最低能效阈值。当机组COP高于该数值时，将电力从GPU算力供给分配至制冷机组，仍可实现系统每瓦Token产出的净增长；低于该阈值时，运行制冷机组会拉低智算工厂整体能效。

DVFS (Dynamic Voltage and Frequency Scaling动态电压频率调节

GPU内置的自适应调控机制，可根据设备负载、芯片最高结温的实时变化，动态调整工作电压与时钟频率，实现性能与散热的动态平衡。

Developed vs. Developing Coolant Flow充分发展流／发展流

冷却介质进入微通道的两种流态。介质刚流入通道时为发展流，流体充分混合，与通道壁面换热效率达到峰值；介质沿通道持续流动后，转变为充分发展流，流态趋于稳定均匀，换热效率大幅衰减。短回路微通道设计通过限制通道长度，使冷却介质始终维持在高效发展流状态，换热性能显著优于长通道结构。

Electromigration电迁移

芯片电路在长期高温工况下产生的渐进式老化损耗现象，会逐步损伤电路结构，直接缩短GPU硬件使用寿命。

Free Cooling自然冷却

依托蒸发式冷却塔对机房水路进行降温的散热模式，全程无需机械式制冷设备参与，能耗与运维成本均显著低于机械制冷方案。

GPU (Graphics Processing Unit图形处理器

AI 工厂的核心计算芯片，最初应用于图形渲染领域，现已成为AI模型训练、推理任务的核心算力硬件。

$T_{j\,max}$ 最高结温

GPU裸片内部有源晶体管层的极限工作温度，是决定芯片运行性能与使用寿命的核心变量。热管理技术栈的所有工程设计，最终目标均为精准管控该温度指标。

Leakage Power 漏电流功耗

GPU晶体管产生的无效损耗功耗，无论芯片是否执行计算任务均会持续产生，仅生成热量、不贡献算力输出，且功耗随芯片结温升高呈指数级增长。

LiquidJet® 液态喷射微通道液冷技术

Frore Systems自研专利冷板技术，采用半导体精密制造工艺，可实现传统铲齿工艺无法复刻的高精度微通道冷板结构，是面向高密度AI算力场景的新一代液冷解决方案。

Mechanical Chiller机械式制冷机组

主动式制冷设备，用于将机房冷却水温降至纯自然冷却无法达成的低温区间，补足自然冷却的温控上限。

Pump Affinity Laws泵相似定律

流体工程核心准则，用于界定水泵功耗与介质流速的变化关系。针对智算工厂场景的核心结论：水泵功耗与介质流速的三次方成正比，流速翻倍，泵耗功率提升至原来的8倍。

Q GPU整机功耗

GPU运行总功耗，单位为瓦（W），是热管理系统需要持续移除的总热负载。

R_{die} 裸片热阻

GPU硅基裸片本体产生的固有热阻。

$R_{package}$ 封装热阻

由GPU封装结构带来的热阻，区分带盖封装与无盖封装两种架构，包含裸片至热界面材料之间所有结构的热阻叠加。

$R_{coldplate}$冷板热阻

从芯片到冷水机组

热管理技术栈对 AI 工厂能效的核心影响

冷板本体产生的热阻，由微通道几何结构、流道布局、冷却介质流速三大因素共同决定。

$R_{graphene\ TIM}$石墨烯热界面层热阻

带金属均热盖的GPU封装中，裸片与均热盖之间的石墨烯热界面材料所产生的界面热阻。

R_{lid}均热盖热阻

带盖GPU封装结构中，集成金属均热盖所引入的额外热阻。

R_{total}总热阻

GPU晶体管层至冷却介质之间，热管理技术栈全链路的累积总热阻。在GPU功耗、冷却液进水温度恒定的前提下，总热阻越低，芯片最高结温越低。

R_{TIM}热界面材料热阻

填充于GPU封装与冷板之间的热界面材料，所产生的界面接触热阻。

TIM (Thermal Interface Materia)热界面材料

填充在GPU封装表面与冷板之间的导热介质，用于消除接触面缝隙、强化热传导效率，降低界面热阻。

R (Thermal Resistance)热阻

用于衡量材料或构件阻碍热量传导能力的物理指标。热阻数值越低，热量传导效率越高，芯片最高结温控制效果越好。

Thermal Stack热管理技术栈

一套完整的层级化散热架构体系，覆盖从GPU芯片晶体管层取热、逐级导热，到最终通过冷却塔向外界排热的全链路硬件与系统，是决定智算工厂算力能效的核心底层架构。

T_{free}自然冷却极限水温

仅依靠自然冷却可实现的最低冷却液进水温度，由环境温湿度、冷分配单元换热器效率共同决定，达成该温度无需启动机械式制冷机组。

T_{inlet}冷却液进水温度

冷却介质进入冷板时的初始温度。进水温度每降低1℃，GPU最高结温同步降低1℃，二者呈线性对应关系。

$T_{mechanical}$机械制冷附加温降

当自然冷却无法达到目标进水温度时，通过机械式制冷机组实现的额外降温幅度。

Token算力单元

AI模型输出的基础计量单元，AI生成的文本、图像、运算结果等所有智能输出，均由Token构成。

Tokens/Watt

每瓦Token产出智算工厂核心能效与经济性指标，用于衡量每消耗1瓦电力可产出的AI有效算力单元数量，数值越高代表算力利用效率、资本回报率越优。

T_{outlet}冷却液出口温度

指冷却液在冷板中吸收 GPU 热量后，从冷板流出的温度。

Wet Bulb Temperature湿球温度

指在特定气候条件下，通过蒸发冷却能够实现的降温能力指标。它综合考虑空气环境温度和相对湿度的影响。湿球温度越低，自然冷却的效果越好，并且在无需使用机械制冷系统的情况下，可以实现更低的 T_{inlet} 温度。湿球温度是决定任意数据中心所在地可利用自然冷却能力的关键气候变量。

Do More.

V1.3