



FRORE
SYSTEMS

实现更多



散热下一场产业革命

热管理技术栈对 AI 工厂盈利能力的影响

热管理技术栈对 AI 工厂 盈利能力的影响



摘要

世界正面临电力约束，能源供应将限制 AI 工厂的增长。

AI 工厂的产出是 Token，即智能产出的计量单位。因此，AI 工厂必须针对系统级效率指标 Tokens/Watt 进行优化。

半导体技术的关键洞见之一是，Tokens/Watt 指标深受 GPU 裸片结温的影响。在相同功率下运行，温度更低的 GPU 能产出更多 Token。

决定 AI 工厂中 GPU 裸片温度的基础设施是热管理技术栈。热管理技术栈涵盖从 GPU 提取热量的完整散热链（GPU 封装架构、热界面材料（TIM）、冷板、冷却液入口温度和流量）以及向大气排热的环节（自然冷却型蒸发冷却塔和机械冷水机组）。

本文分析热管理技术栈对 GPU 裸片结温的深远影响，以及由此对 AI 工厂效率指标 Tokens/Watt 的影响。设计精良的热管理技术栈可将 AI 工厂 Tokens/Watt 指标提升 40% 以上，直接带来 40% 的营收增长，且无需额外成本，因而全部转化为利润。

“AI 工厂针对 Tokens/Watt 指标 进行优化”

AI 工厂耗电巨大且电力受限

几乎每一座 AI 工厂的决定性运营约束都不是算力、内存带宽或资本投入——而是电力。

AI 工厂可用的电力是有限的，扩容成本高昂。公用电网接入、变压器容量以及物理输配电基础设施，都是限制

AI 工厂总功耗的因素。在当前的监管环境下，主要数据中心市场的电网容量日益稀缺，多个司法管辖区的监管机构正在实施用电上限。后备（BTM）发电虽有帮助，但电力需求与供给之间的巨大缺口依然存在。面对这一功率约束，AI 工厂运营商深知功率是战略资产，每一瓦分配的功率都必须尽可能高效地用于生成 Token。

Tokens/Watt 每提升一个百分点，盈利能力便直接提升一个百分点。正是这一经济逻辑，使 Tokens/Watt 成为衡量 AI 工厂效率的核心运营指标。

**“设计精良的
热管理技术栈可将
AI 工厂营收提升 40% 以上，
全部直接转化为利润。”**

能源成本是 AI 工厂 OPEX 中占比最大的部分

要评估所交付 Token 的真实成本，必须全面了解数据中心运营支出的构成。

AI 工厂 OPEX 的组成部分包括能源成本、设施维护、人员与运营、网络与连接成本、硬件维护、软件许可和设施租赁。

能源成本是其中占比最大、波动也最大的因素，涵盖 GPU、CPU、内存、网络、存储、供电以及热管理技术栈的功耗。由于 GPU 是 AI 工厂功率预算的主要消耗者，行业历来将投资和精力集中于 GPU 设计改进。从一个工艺节点迈向下一个工艺节点，并转向日益面向推理优化的设计，显著提升了同等功率下的 GPU 性能，并直接转化为系统级 Tokens/Watt 的提升。

然而，一旦某代 GPU 的设计架构、工艺节点和浮点精度支持确定，其 GPU 效率即已锁定。在下一代产品问世之

热管理技术栈对 AI 工厂盈利能力的影



前，GPU 层面的性能无法再提升。正是在这里，热管理技术栈成为提升系统级 Tokens/Watt 所剩的主要手段。

GPU 裸片温度对 Tokens/Watt 的巨大影响

GPU 裸片结温 T_j 是指 GPU 裸片有源晶体管层的最高温度。

它不同于在 GPU 封装壳体表面、冷板或冷却液处测得的温度。

它是热管理技术栈中的最热点，因为热量由流经 GPU 裸片的功率产生，且 GPU 消耗的所有功率最终都转化为热量。

这些热量必须从 GPU 向外逐层导出——经过 GPU 封装、TIM、冷板、冷却液——最终由设施冷却系统排走。每颗 GPU 都有最大额定结温（ T_j 限值），现代 GPU 通常在 95°C 左右。

GPU 功耗由两部分组成：

动态功耗是晶体管在计算过程中开关所消耗的有效功率，随时钟频率、电压以及任一时刻参与工作的有源晶体管数量而变化。

漏电功耗是被浪费的功率。无论是否在进行计算，它都持续流经晶体管，只产生热量而不贡献性能。

GPU 的功率效率高度依赖于裸片温度。根据阿伦尼乌斯方程， T_j 结温每升高约 10°C，漏电功耗呈指数增长，约翻一番。这形成连锁效应：温度越高，漏电越大，进而产生更多热量，推动温度进一步升高。随着漏电功耗增加，动态功耗相应减少。此外，晶体管开关所需的功率也随 T_j 结温每升高 10°C 增加约 2%，带来更低的时钟频率和更低的 GPU 性能。

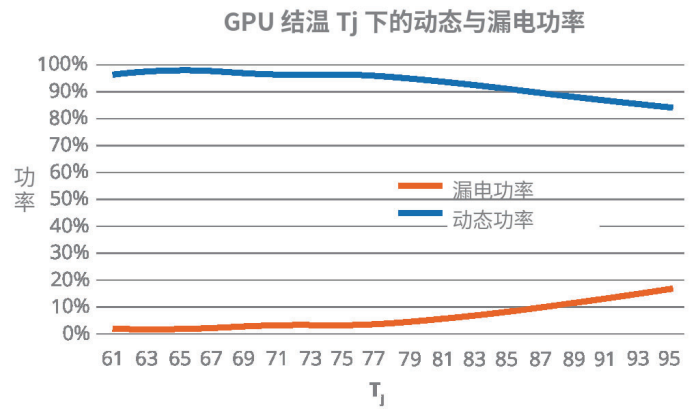


图1：结温升高时，漏电功率呈指数增长

在实际应用中，这些热损耗通过动态电压频率调节（DVFS）来管理。DVFS 是现代 GPU 的一种机制：当裸片温度升高时，提高电压并向 GPU 输送更高功率，以维持一定的时钟频率。

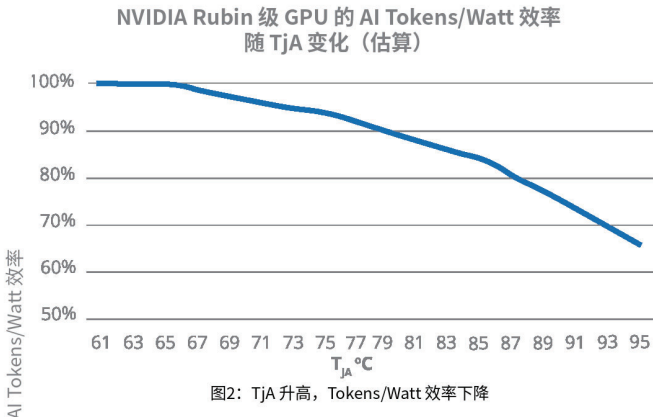
当结温接近最大额定限值（ T_j 限值）且电压达到最大允许值时，DVFS 会降低时钟频率，以减少动态功耗，使 T_j 保持在最大允许值以下。

在裸片面积为 A 、总功耗为 Q 的真实 GPU 中，GPU 裸片面积上的功率分布 Q' 与结温分布 T_j 并不均匀。因此， T_{jA} ——按功率密度加权的 GPU 裸片面积平均温度——决定了漏电功耗。

$$T_{jA} = (\int_A (Q' \times T_{jx}) da) / Q$$

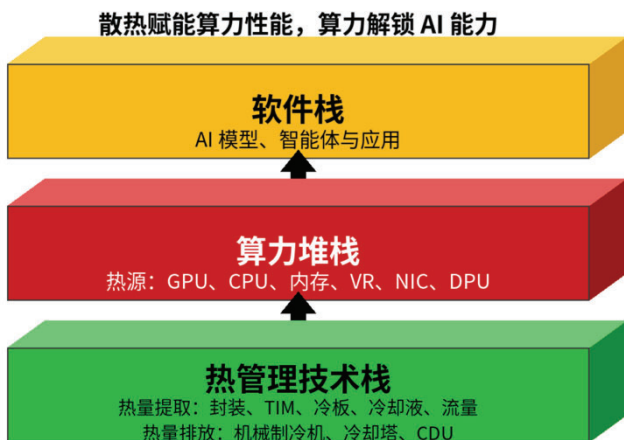
随着 T_{jA} 升高，无效漏电功耗增加，GPU 功率效率下降。GPU 加权结温 T_{jA} 与 AI 工厂效率指标 Tokens/Watt 之间的关系如下图所示。

热管理技术栈对 AI 工厂盈利能力的影



**“GPU 结温
每降低 10°C ,
Tokens/Watt 可提升 15%”**

AI 工厂架构



AI 工厂是为单一产出而专门构建的系统: **Token**。其架构由三个相互依存的层级组成。

软件栈: AI 基础模型、智能体与应用。

算力栈: 由所有 GPU、CPU、内存、网络交换芯片及供电组成。算力堆栈产生热量, 因为其消耗的所有功率都转化为热量。

热管理技术栈: 从算力堆栈提取热量并排入大气。热量提取链包括芯片封装架构、热界面材料、冷板、冷却液和流量。热量排放包括冷却液分配单元 (CDU)、自然冷却型冷却塔和机械冷水机组。

热管理技术栈是使一切成为可能的基础层。没有有效的热量提取与排放, 算力堆栈就会降频, 软件堆栈就会停滞。热管理技术栈的效率决定 GPU 能交付多少有效算力, 以及系统级 Tokens/Watt 指标的实际水平。

热管理技术栈的定义

算力栈产生的热量必须先被提取出来, 才能排入大气。

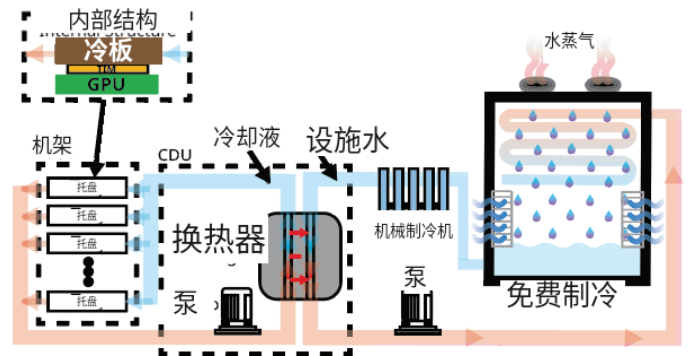


图3: 热管理技术栈组件

芯片级直接液冷 (DLC) 已成为最新一代 AI 工厂的行业标准。

该方案利用在冷板中循环的液态冷却液从 GPU 提取热量, 冷板与 GPU 直接热接触。

热管理技术栈对 AI 工厂 盈利能力的影响



低温冷却液 (T_{inlet}) 被泵入冷板内的微通道以提取 GPU 热量，升温后的冷却液 (T_{outlet}) 流出冷板。

冷却液经冷却液分配单元 (CDU) 循环，CDU 内置换热器，出口温度为 (T_{outlet}) 的冷却液在其中将热量排入设施水环路，使冷却液在回流至冷板之前恢复到 (T_{inlet})。

传递到设施水环路的热量，或通过冷却塔蒸发排入外部大气 (自然冷却)，或由机械冷水机组带走，或两者兼用，具体取决于当地气候、水资源条件和所需的 (T_{inlet})。冷却液环路与设施水环路在 CDU 内彼此保持水力隔离。

这使得高纯度冷却液可以独立于设施供水进行化学控制。

GPU 最大裸片结温 T_{jmax} 由以下公式决定：

$$T_{jmax} = T_{inlet} + Q \times R_{total}$$

其中 Q 为 GPU 消耗的功率， T_{inlet} 为冷却液进入冷板时的温度， R_{total} 为 GPU 裸片与冷却液之间的累计热阻 ($^{\circ}\text{C}/\text{W}$)。

GPU 裸片位置 x 处的结温由以下公式决定：

$$T'_{jx} = T'_{inlet\ x} + Q'_{\ x} R'_{\ x\ total}$$

其中 $Q'_{\ x}$ 为位置 x 处的功率密度， $T'_{inlet\ x}$ 为冷却液流经位置 x 时的温度， $R'_{\ x\ total}$ 为位置 x 处的累计热阻 ($^{\circ}\text{C}\cdot\text{cm}^2/\text{W}$)。

因此，如前一节所述，制约 GPU 性能的 GPU 加权结温 T_{jA} 由以下公式决定：

$$T_{jA} = (\int_A (Q'_{\ x} (T'_{inlet\ x} + Q'_{\ x} R'_{\ x\ total})) da) / Q$$

而 GPU 最大裸片结温为所有位置中最高的 T_{jx} 值：

$$T_{jmax} = \max (T_{jx})$$

T_{jmax} 通常出现在 GPU 裸片功率密度最高的位置处的 T_{jx} 。

R_{total} 由 GPU 封装架构、热界面材料 (TIM) 选型和冷板设计决定。

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

T_{inlet} 为自然冷却可实现温度 (T_{free}) 与机械制冷降温 ($T_{mechanical}$) 之和： (T_{free}) 取决于湿球温度——受大气环境温度、相对湿度决定，以及 CDU 效率共同决定； $T_{mechanical}$ 为机械冷水机组提供的降温。

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

机械冷水机组耗电量大。为维持更低的 T_{inlet} ，会增加设施级电力成本。然而，机械冷水机组的性能系数 (COP) 越高，额外增加的设施级电力成本就越低。

R_{total} 与 T_{inlet} 共同将热管理技术栈的物理设计与 Tokens/Watt 联系起来。

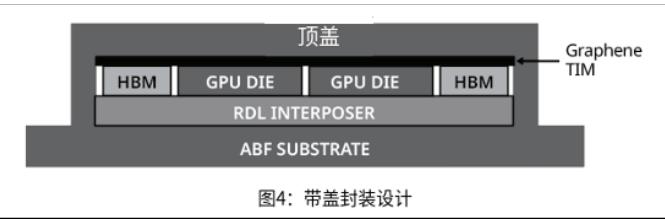
GPU 封装架构

$R_{package}$ 因 GPU 封装采用带盖还是无盖设计而有显著差异。在带盖封装中，集成均热盖板 (顶盖) 覆盖在裸片复合体之上，提供机械保护和横向均热。顶盖与逻辑裸片之间夹有一层薄薄的石墨烯 TIM。因此，带盖封装的热阻显著高于无盖封装。

热管理技术栈对 AI 工厂盈利能力的影

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

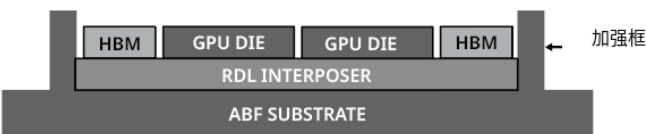
$R_{\text{total (lidded)}} = R_{\text{die}} + R_{\text{lid}} + R_{\text{graphene TIM}} + R_{\text{TIM}} + R_{\text{coldplate}}$



在无盖封装中，既没有顶盖，也没有石墨烯 TIM。因此，在相同的 Q 和 T_{inlet} 下，无盖设计的 R_{total} 更低， T_{jmax} 明显更低。

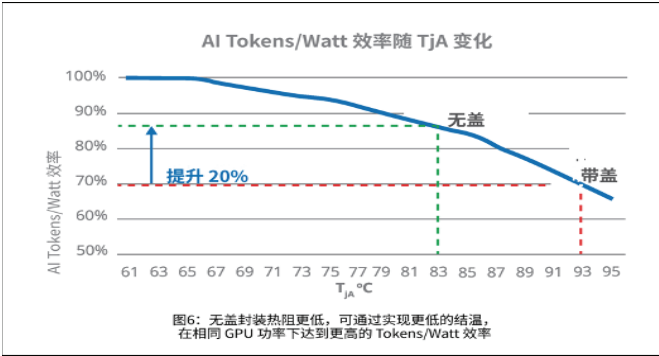
$R_{\text{package (un-lidded)}} = R_{\text{die}}$

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



对于 NVIDIA Rubin 级别的 GPU，带盖与无盖封装之间的 T_{jmax} 和 T_{ja} 差异可达 10°C，带来高达 20% 的 Tokens/Watt 效率提升。

	带盖封装	无盖封装
典型封装热阻 (°C·cm²/W)	0.097	0.047



但无盖封装可能存在可靠性方面的权衡。无盖封装中裸片直接暴露，在搬运和冷板安装过程中崩边或开裂的风险高于带盖封装。

此外，顶盖提供精确、均匀的表面。没有顶盖，冷板可能会遇到裸片表面 z 向高度不均匀的问题。

热界面材料

GPU 封装与冷板之间使用热界面材料 (TIM) 以确保优异的热接触，但 TIM 会增加热阻。常用的 TIM 材料分为三类：

TIM 类型	典型热阻	备注
信越 导热硅脂	0.09°C · cm²/W	成本最低、热阻最高、最易返工
Honeywell PTM7950	0.06°C · cm²/W	成本可接受、热阻低、稳定性高
液态金属钢	0.03°C · cm²/W	成本高、热阻最低，但需密封设计以防长期使用中迁移/腐蚀

热管理技术栈对 AI 工厂盈利能力的影

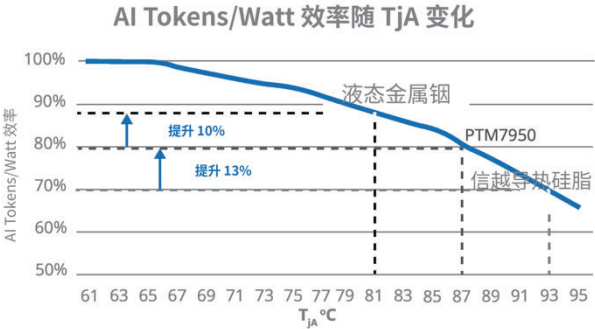


图7：不同 TIM 热阻更低，可通过实现更低的结温，在相同 GPU 功率下达到更高效率

液态金属钎可通过在顶盖顶部和冷板底部电镀金，有效用于带盖封装。由此，带盖封装热阻偏高的劣势可通过使用液态金属钎 TIM 得到一定程度的抵消。事实上，这正是 NVIDIA Rubin GPU 所采用的方案。

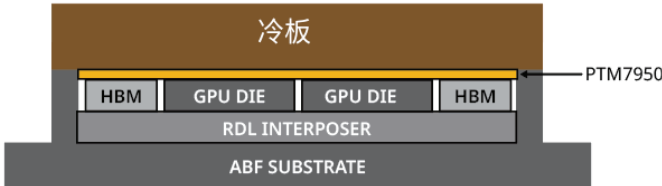


图8：采用 PTM7950 TIM 的无盖封装设计

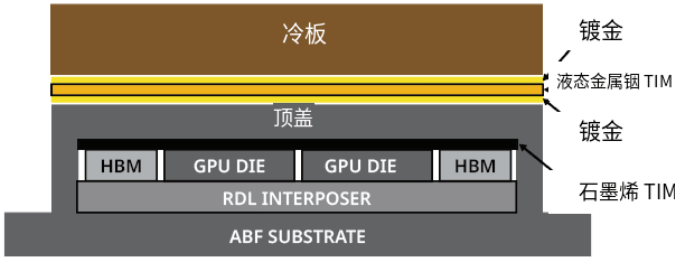


图9：采用液态金属钎 TIM 的带盖 NVIDIA Rubin 封装

然而，

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

两种配置的 $T_{j\text{max}}$ 差异为 4°C，带来高达 10% 的 Tokens/Watt 效率差异。

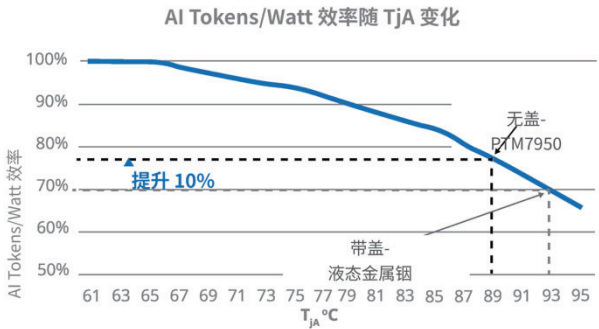


图10：采用 PTM7950 TIM 的无盖 GPU 封装，比采用液态金属钎 TIM 的带盖封装在相同功率下效率更高

冷板设计

冷板设计对降低热阻 R_{Total} 起着关键作用。冷板热阻 $R_{\text{coldplate}}$ 由三个相互影响的设计变量决定：微通道几何结构、流道布置和流量。

微通道几何结构

微通道长度、翅片间距、翅片高度和深宽比决定冷板壁面与流经冷板的冷却液之间的对流换热系数。

更窄的通道和更高的深宽比可增加单位体积的表面积并改善换热，而过长的微通道长度会因冷却液流动已充分发展而降低换热效率。

因此，设计长度 $< 1\text{mm}$ 的超短通道（超短流道），可使冷却液流动保持在发展流阶段，从而显著提升换热效率。

要构建超短回路微通道，需要 3D 微单元。但采用铲齿等传统机械加工方法无法制造 3D 微单元。

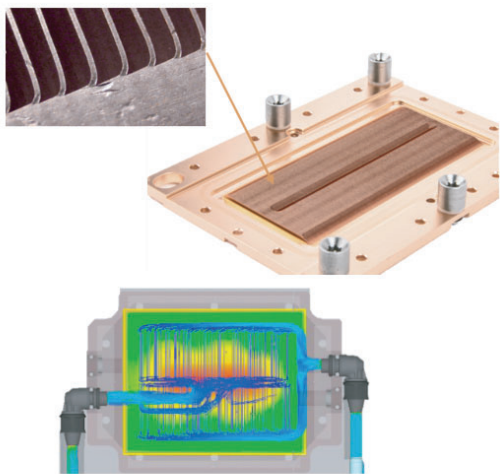


图11：分流式冷板中的锯齿微通道

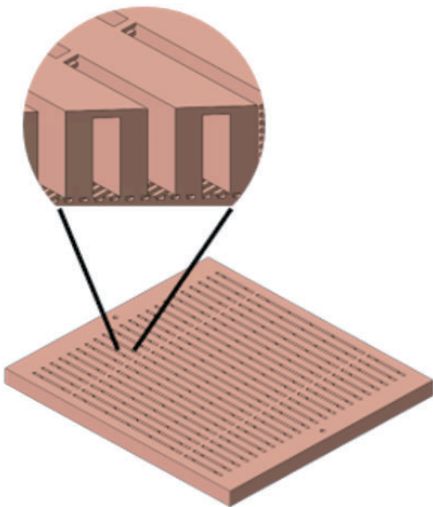


图14：带 3D 微单元的冷板

3D 微单元

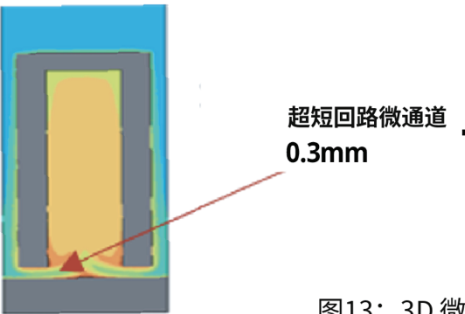


图13：3D 微单元

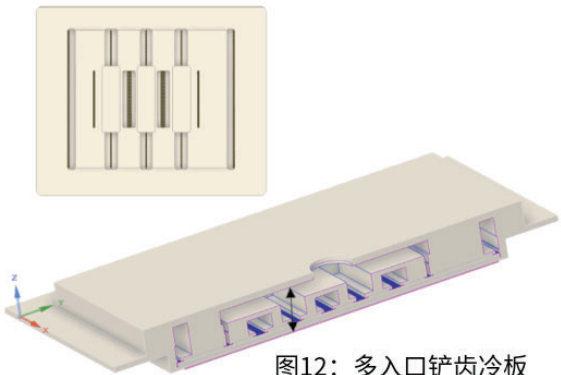


图12：多入口锯齿冷板

**“3D 微单元、
超短流道与多级设计相结合，
带来远为高效的
冷板设计”**

流道布置

进入冷板的冷却液被分配到各微通道。最简单的形式是：通过一端单个入口将流量均匀分配到所有微通道，再在远端汇集。但这会产生局部热点，因为高功率裸片区域分到的流量与其余区域相同。更糟的是，高功率裸片区域可能位于冷却液已充分发展、且已吸收热量的微通道下游。

稍好一些的流道布置方式是“分流式”结构。该设计中，流量从冷板两端的两个入口平均分入，升温后的冷却液在冷板中部的出口汇集。

这使微通道的有效长度减半，从而提升换热效率。一个简单的歧管覆盖在微通道上即可实现这种流道布置。

热管理技术栈对 AI 工厂盈利能力的影

这一思路的进一步延伸是多入口冷板。流量按权重分配到多个入口，这些入口正好位于 GPU 裸片功率密度较高的区域上方。

多级结构

但在上述三种冷却液流道布置方案中，流量都只被使用一次，不循环。当进入的冷却液分配到各微通道后，只沿微通道直线流过一次，便在出口被汇集。

而在多级冷板中，冷却液被多次循环利用。多级冷板将冷板分成多个区段，每个区段依次承接全部冷却液流量。例如，在 3 级冷板中，全部冷却液先进入第 1 级，从第 1 级流出后，全部冷却液依次进入第 2 级，再进入第 3 级。各级的划分依据 GPU 裸片功率分布图，前面的级对应高功率密度区域。多级设计显著降低 GPU 的 T_{jmax} 和 T_{jA} 。

需要注意的是，多级结构会增加冷板压降，而超短流道则降低压降。因此，将这两种特性结合使用，可在不增加冷板压降的前提下，获得 $R_{coldplate}$ 值显著更低、高效的冷板设计。

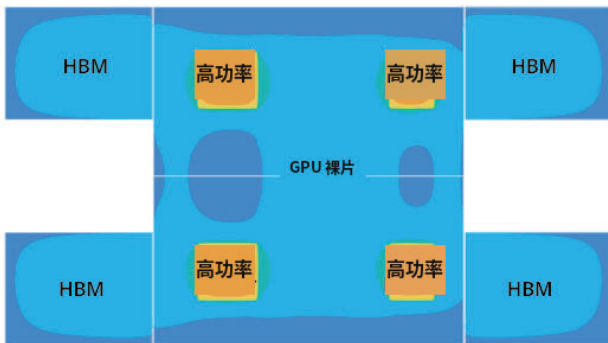


图15：含四个高功率区的 GPU 功率密度图

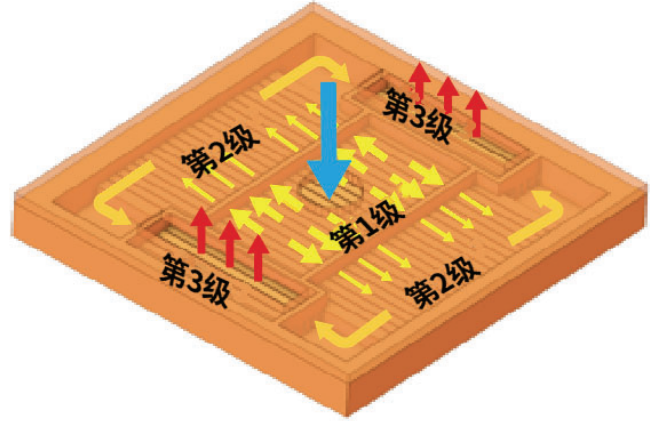


图16：多级设计的冷板

3D 超短回路多级 LiquidJet 冷板，可在无盖封装 + PTM7950 TIM、70°C 温差 ($T_{jmax} - T_{inlet}$) 条件下，支持 5mm × 5mm 大热点 670 W/cm² 的功率密度。针对带盖 NVIDIA Rubin GPU 功率分布定制的 LiquidJet 冷板，与最先进的铲齿多入口冷板相比，可将最大裸片结温 T_{jmax} 和 T_{jA} 降低 6°C 以上。

冷却液

在芯片级直接液冷系统中，使用两种冷却液：去离子水（热容 4.18 焦耳/毫升/°C）和 PG25（25% 丙二醇 + 75% 水，热容 3.85 焦耳/毫升/°C）。要使 1 LPM 的冷却液升温 1°C，两者分别需要 69.67 W 和 64.17 W。

流量

冷板设计确定后，流量是剩下的可调运行自由度。提高流量可降低 $R_{coldplate}$ ，但两者并非线性关系： $R_{coldplate}$ 在低流量时随流量急剧下降，接近其几何极限后趋于平坦，每增加一单位流量所换来的热收益递减。

更高的流量也会增大压降和泵功耗，泵功耗与流量呈立方关系（泵相似定律）。

热管理技术栈对 AI 工厂 盈利能力的影



实际中，相对于 GPU 和机械冷水机组的功耗，这一功率代价很小，因此将流量提高到 $R_{coldplate}$ 趋于平坦的点几乎总是值得的——但运行点通常由设施 CDU 容量以及歧管、软管和快换接头的防漏规格决定。

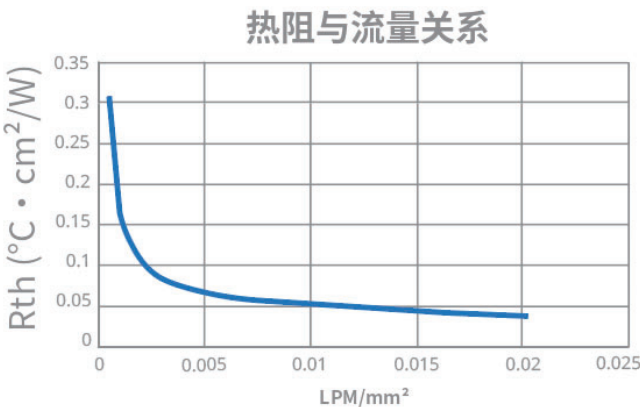


图17：3D 微单元冷板的热阻与流量关系

	分流式冷板	多入口冷板	LiquidJet冷板
热点处典型冷板热阻 (°C·cm²/W)	0.10	0.07	0.04
冷板整体典型热阻 (°C/W)	0.021	0.0192	0.0165

制造工艺

制造工艺决定了哪些冷板设计能够实现。

传统的机械铲齿加工技术是当今大批量冷板生产普遍采用的方法。在铲齿加工中，翅片直接从整块铜上铲削成形，形成连续、长条、高深宽比的翅片结构。铲齿结合歧管，可生产简单的微通道冷板、分流式冷板和多入口冷板。但超短回路微通道、多级等先进结构超出了铲齿的能力范围，需要新的制造技术。

Frore Systems 开发了独特的制造工艺：将半导体制造工艺应用于铜晶圆，以实现复杂的 LiquidJet® 冷板设计。LiquidJet® 冷板采用窄轮廓微通道、由 3D 微单元实现的超短回路、多级以及一系列针对高功率裸片区域、降低 $R_{coldplate}$ 的新颖技术。

Frore Systems 首先针对 GPU 功率分布图生成定制 LiquidJet® 冷板的 CAD 设计，然后利用掩模版生成、光刻、湿法刻蚀、单片键合等半导体制造工艺，以经过验证的可靠性实现 LiquidJet® 冷板的大规模低成本制造。

Frore Systems 的 LiquidJet® 冷板可将 T_{JA} 降低 6–12°C，从而将 Tokens/Watt 显著提升 10–25%。

“将半导体制造工艺
应用于铜晶圆，
得以实现复杂的
LiquidJet® 冷板设计。”

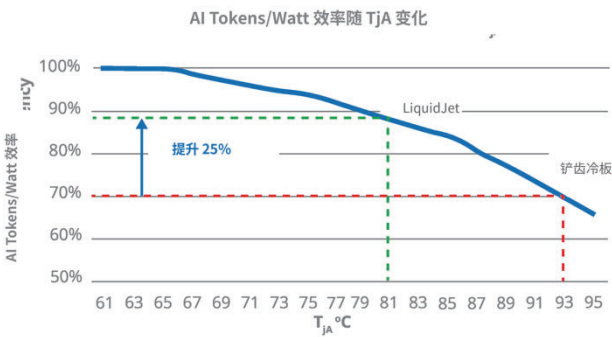


图18：LiquidJet 热阻更低，可通过实现更低的结温，在相同 GPU 功率下达到更高效率

热管理技术栈对 AI 工厂 盈利能力的影响

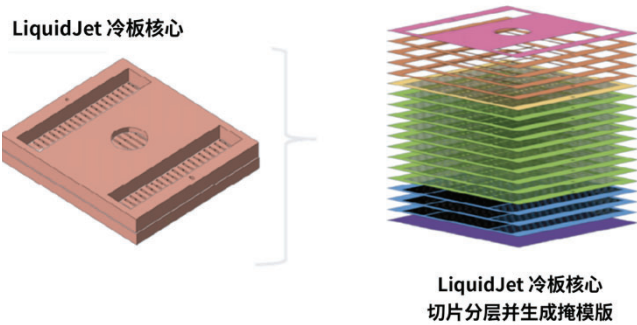


图19: LiquidJet® CAD 设计 → 掩模版生成

也有公司尝试用金属 3D 打印等增材制造技术构建类似的短回路冷板，但这些方案的灵活性、可靠性和可扩展性尚未得到验证。

LiquidJet Diamond

针对 GPU 功率分布定制的 LiquidJet Diamond 冷板，与纯铜 LiquidJet 冷板相比，可将最大裸片结温 T_{jmax} 再降低 10°C。

Fröre Systems 正在推出 LiquidJet 冷板的新变体——“LiquidJet Diamond”。LiquidJet Diamond 可将 T_{jmax} 再降低 10°C，将 Tokens/Watt 提升 30–35%。LiquidJet Diamond 通过在冷板底部均热板嵌入金刚石晶圆实现更优的散热性能。参考图 15 中带热点的 GPU 功率分布图，LiquidJet 通过多级设计瞄准这些热点；若在热点正下方的冷板底部精确集成金刚石晶圆，还可进一步改善。

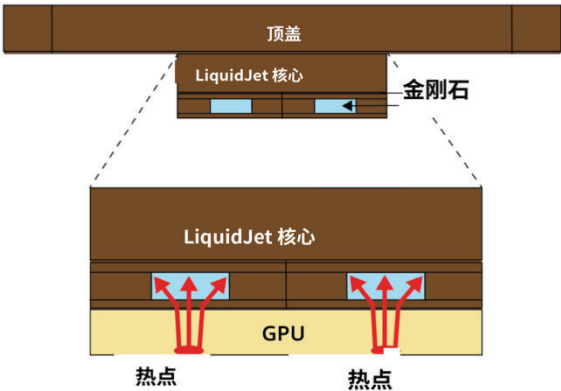


图20: LiquidJet Diamond

金刚石晶圆的导热系数为铜的 5 倍，可在 x-y 方向快速横向扩散热量，弥散热点、降低其等效功率密度，而冷板热阻几乎不增加。因此，LiquidJet Diamond 可在无盖封装 + PTM7950 TIM、70°C 温差 ($T_{jmax} - T_{inlet}$) 条件下，支持 5mm × 5mm 大热点 770 W/cm² 的功率密度。

	LiquidJet	LiquidJet Diamond
热点处典型冷板热阻 (°C·cm ² /W)	0.04	0.03
冷板整体典型热阻 (°C·cm ² /W)	0.0165	0.0122

图 21 : LiquidJet 冷板热阻

热管理技术栈对 AI 工厂盈利能力的影

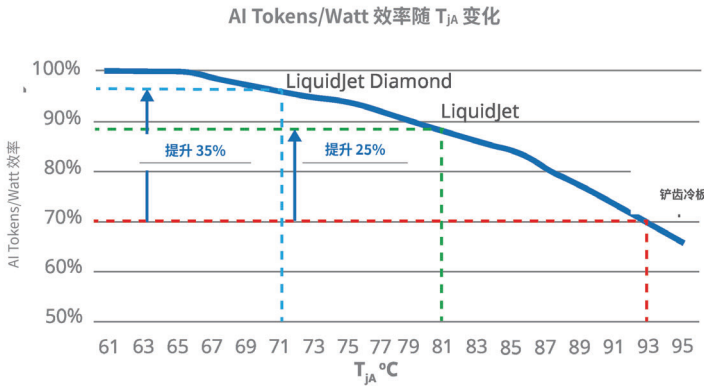


图22: LiquidJet Diamond 通过实现低 10°C 的 T_{JA}, 在相同 GPU 功率下达到更高效率

**“LiquidJet Diamond 将
Tokens/Watt 指标
提升 35%。”**

冷却液入口温度

冷却液入口温度 (T_{inlet}) 与 GPU 加权结温 (T_{JA}) 呈直接的 1:1 关系。在 GPU 功率 (Q) 与 R_{Total} 保持不变时, T_{inlet} 每降低 1°C, T_{JA} 同步降低 1°C。而 T_{JA} 每降低 1°C, GPU 的 Tokens/Watt 效率都会提升。

NVIDIA 已规定基于 Rubin GPU 的 AI 工厂按 45°C 入口温度运行。该规定的意义在于设施级功耗: 在全球大多数数据中心气候条件下, 仅靠自然冷却即可实现 45°C 入口温度, 完全省去机械冷水机组。

话虽如此, NVIDIA Rubin GPU 平台能在 45°C 入口温度下运行, 并不意味着每处部署都应该如此。对于部署在 AI 工厂中的 GPU, 鉴于 Tokens/Watt 指标对 GPU 裸片

结温的敏感度, 强烈建议采用更低的 T_{inlet} 。但将 T_{inlet} 降到 T_{free} 以下需要使用耗电量大的机械冷水机组, 可能拉低 AI 工厂整体 Tokens/Watt 系统功率效率指标, 因为分配给机械冷水机组的功率必须从 GPU 功率中扣除。因此, 除非制冷机 COP 足够高, 否则这种取舍并不划算。

机械冷水机组对 Tokens/Watt 的影响

当 T_{inlet} 被维持在 T_{free} 以下时, 需用机械冷水机组补足差值。机械冷水机组的效率用性能系数 (COP) 表征:

$$COP = Q / W$$

其中 Q 为带走的热量, W 为机械冷水机组消耗的电功率。COP 为 5 意味着制冷机每消耗 1 千瓦电功率, 可带走 5 千瓦热量。假设用机械冷水机组将 T_{inlet} 降低 1°C, 则入口温度降低 1°C 直接转化为 T_{JA} 降低 1°C, 进而可带来 1.5% 的 Tokens/Watt 提升 (T_{JA} 从 89°C 降至 88°C)。

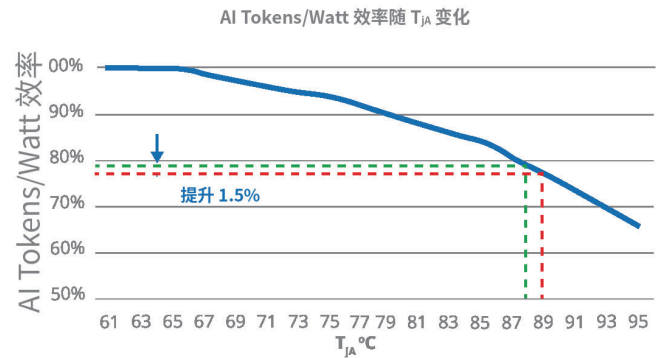


图23: 结温每降低 1°C, AI Tokens/Watt 提升 1.5%

热管理技术栈对 AI 工厂盈利能力的影



然而，降低 T_{inlet} 并非免费，机械冷水机组要耗电，当冷却液为去离子水时：

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

因此，若总功率预算 x 保持不变，当一部分功率从 GPU 转给机械冷水机组，留给 GPU 的功率预算就相应减少。好的一面是， T_{JA} 降低 1°C 可带来 $y\%$ (T_{JA} 从 89°C 降至 88°C 时为 1.5%) 的 Tokens/Watt 提升。于是，只要机械冷水机组的 COP 足够高，Tokens/Watt 就可能提升。使 GPU 的 Tokens/Watt 在功率转给机械冷水机组后保持不变的 COP，称为 $\text{COP}_{\text{breakeven}}$ (盈亏平衡 COP)。

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

在功率受限的 AI 工厂中，只有当 COP 高于 $\text{COP}_{\text{breakeven}}$ 时，把功率从 GPU 转给机械冷水机组才有意义。

NVIDIA Rubin GPU 功率为 2300W，估算流量为 3.25 LPM，得出 $\text{COP}_{\text{breakeven}}$ 为 6.7。因此，只有当 $\text{COP} > 6.7$ 时，使用机械冷水机组才划算。

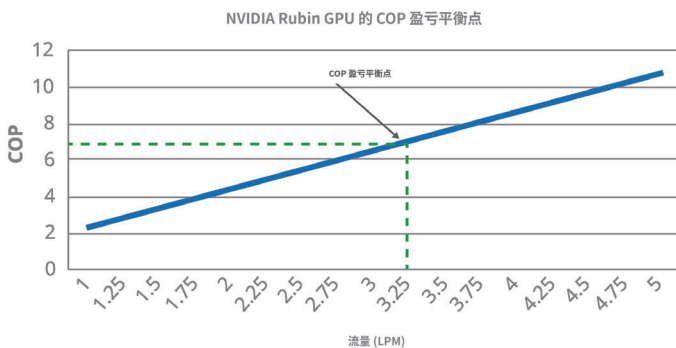


图24：NVIDIA Rubin 的 COP 盈亏平衡点

GPU 可靠性

除提升 Tokens/Watt 效率外，较低的 GPU 最大裸片温度对 GPU 可靠性也有积极作用，尤其在对 AI 工厂至关重要的几个可靠性维度上。

GPU MTBF：

在较高结温下运行会加速电迁移和氧化层退化，缩短芯片寿命，并增加一批需连续运行数年的 GPU 集群早期失效的风险。在 95°C T_{jmax} 下持续运行的 GPU，其寿命可能比在 65°C 下运行缩短一半以上。

GPU 静默比特翻转：

静默比特翻转会导致静默数据损坏 (SDC)，即 GPU 内部某个比特——寄存器、运算单元或内存中的比特——在计算过程中翻转为错误值，而系统中的任何环节都检测不到。芯片会静默输出错误计算结果，并像正确结果一样交给下游。

在拥有 10 万颗以上 GPU、训练任务持续数周的 AI 工厂中，这是真实存在的担忧。与漏电功耗一样，SDC 概率随 GPU 最大裸片温度升高呈指数上升。

GPU 集群频率分布：

在大型 AI 工厂的 GPU 集群中，由于半导体制造差异，每颗 GPU 的实际最高频率围绕均值分布，其标准差与 GPU 裸片温度相关。例如，最高频率为 2250MHz，但每颗 GPU 的实际值在 T_{jmax} 为 65°C 时的最小/最大偏差范围为 300MHz，而在 T_{jmax} 为 95°C 时偏差范围翻倍至 600MHz。

因此，当 T_{jmax} 从 65°C 升至 95°C ，集群中最慢的 GPU 从

热管理技术栈对 AI 工厂 盈利能力的影响



2100MHz 降至 1950MHz。这对 GPU 集群性能影响显著，尤其是训练负载采用分布式同步方式，在所有 GPU 完成各自任务并到达检查点之前无法继续推进——最慢的 GPU 决定机群性能。

结论

AI 工厂是功率受限的系统，其主导效率指标——Tokens/Watt——决定每一次部署的盈利能力。

GPU 一经制造和部署，其硅片级效率即已固定。能实质性改善 Tokens/Watt 的唯一手段，只剩下热管理技术栈的配置。

**“将热管理技术栈
视为核心设计方向的
超大规模运营商，
可在同等投入下，
产出更高的每瓦 Token 数，
进而提升利润。”**

本文已证明，GPU 最大结温 T_{jmax} 与加权结温 T_{JA} 并非芯片设计的结果，而是可主动设计优化的变量。 T_{JA} 与 Tokens/Watt 之间存在直接关系，它由热管理技术栈各相互依存层级的累计配置决策所控制：GPU 封装架构、热界面材料（TIM）选型、冷板设计、冷却液入口温度、冷却液流量和机械冷水机组的部署。

热管理技术栈的每一层都贡献热阻或增加设施功率预算，每一层都呈现独特的优化机会。

GPU 封装架构给热管理技术栈引入显著热阻。带盖设计

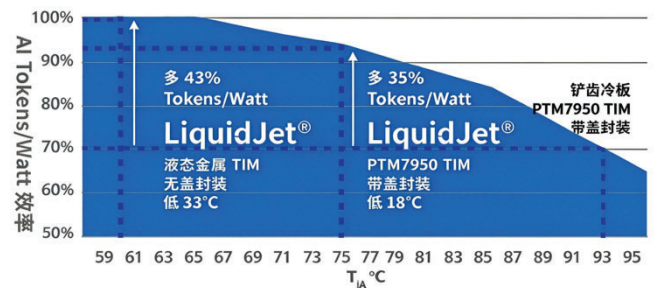
带有石墨烯 TIM 和顶盖，在热阻中增加 $R_{graphene}$ 和 R_{lid} ，而无盖设计消除了这两项，但代价是在搬运和安装过程中机械风险更高。

**“仅 LiquidJet Diamond 冷板
即可将 GPU 裸片结温
降低 16–22°C，
将 Tokens/Watt 指标
提升 30–35%。”**

热界面材料的选型决定热量从封装向冷板传递的效率，液态金属钎的热阻最低。无盖封装配上合适的 TIM，可将 Tokens/Watt 提升高达 25%。

Frore Systems 的 LiquidJet 冷板设计，利用传统铲齿制造工艺无法实现的、由半导体制造技术支撑的 3D 超短回路与多级设计，可大幅降低 $R_{coldplate}$ 。

仅 LiquidJet Diamond 冷板即可将 GPU 最大结温降低 16–22°C，将 Tokens/Watt 指标提升 30–35%。



将 T_{inlet} 设在何处、是否启用机械冷水机组，最终是一个系统级功率分配问题，由机械冷水机组 COP、流量以及 T_{JA} 每降低一度带来的边际 Tokens/Watt 增益共同决定。

AI 基础设施格局正在向这一现实收敛：将热管理技术栈作为一流的设计方向对待的超大规模运营商，将从相同

热管理技术栈对 AI 工厂 盈利能力的影响



的 CAPEX 和 OPEX 投入中获得显著更高的
Tokens/Watt。

随着 GPU 功率密度在 NVIDIA Rubin 之后向未来架构持续攀升，散热优化与未优化部署之间的性能差距只会越拉越大。

热管理技术栈不是辅助基础设施。它是一套性能系统，在功率受限的世界里，是优化 AI 工厂 Tokens/Watt 产出的关键。

设计精良的热管理技术栈可将 AI 工厂 Tokens/Watt 指标提升 40% 以上，直接带来 40% 的营收增长，且无需额外成本，因而全部转化为利润。

**“热管理技术栈不是
辅助基础设施。
它是一套性能系统，
在功率受限的世界里，
是实现 AI 工厂盈利
最大化的关键。”**

热管理技术栈对 AI 工厂盈利能力的影



术语表

AI Factory (AI 工厂) ——专为运行人工智能负载而建的大规模数据中心，其主要产出是 AI 生成的结果 (Token)，而非传统计算任务。

Arrhenius Equation (阿伦尼乌斯方程) ——描述热激活过程速率随温度呈指数上升的数学公式。在 GPU 功耗语境中，它解释了漏电流为何随最大结温升高而急剧上升——即漏电功耗约随 T_{jmax} 每升高 10°C 翻一番的物理基础。以瑞典化学家 Svante Arrhenius 命名。

Behind-the-Meter (BTM) (表后发电) ——位于数据中心设施现场的发电，如太阳能板、备用发电机或燃料电池。“表后”指电力在经过公用电网电表之前即在现场发电并消纳——减少对电网容量的依赖，帮助弥合 AI 工厂电力需求与公用事业供给之间的缺口。

COP (Coefficient of Performance) (性能系数) ——衡量机械冷水机组效率的指标。COP 为 5 意味着制冷机每消耗 1 千瓦电功率，可带走 5 千瓦热量。越高越好。

COP_{breakeven} (盈亏平衡 COP) ——机械冷水机组的最低效率阈值：高于该值，把功率从 GPU 转给制冷机仍能带来 Tokens/Watt 的净提升；低于该值，运行制冷机会损害整体效率。

DVFS (Dynamic Voltage and Frequency Scaling) (动态电压频率调节) ——GPU 内部的自动机制，随负载和最大结温变化调整电压与时钟频率。

Developed vs. Developing Coolant Flow (充分发展流与发展流) ——冷却液刚进入微通道时处于“发展流”状态，即与通道壁主动混合、以最高效率换热。沿通道继续前行后，进入“充分发展流”——稳定均匀的状态，换热效率显著下降。3D 超短回路设计通过限制通道长度使冷

却液保持在发展流阶段。短通道的换热效率高于长通道。

Electromigration (电迁移) ——芯片电路中由持续高温引起的渐进退化过程，会缩短 GPU 寿命。

Free Cooling (自然冷却) ——利用蒸发冷却塔冷却设施用水，无需启用机械冷水机组。相比机械冷水机组，运行功耗与成本更低。

T_{μ} (加权平均结温) ——按功率密度加权的 GPU 裸片面积平均温度。

T_{jmax} (最大结温) ——GPU 裸片内有源晶体管层处的温度。决定性能与寿命的主导变量。热管理技术栈中的一切最终都是为了控制这个数值。

Leakage Power (漏电功耗) ——持续流经晶体管的无效电功率 (漏电功耗)，无论是否在计算。随温度呈指数上升。

LiquidJet® ——Frore Systems 的专有冷板技术，采用半导体工艺制造，实现传统铲齿等制造方式无法复制的冷板设计。

Mechanical Chiller (机械冷水机组) ——压缩式制冷设备，用于将环路水冷至单纯依靠自然冷却无法达到的温度。

Pump Affinity Laws (泵相似定律) ——描述泵功耗变化规律的一组工程原理。对 AI 工厂的关键启示是：泵功耗随流量的立方增长——即冷却液流量翻倍需要 8 倍的泵功耗。

Q——GPU 功耗，以瓦特计。热管理技术栈必须带走的总热负荷。

热管理技术栈对 AI 工厂 盈利能力的影响



R_{die} ——GPU 裸片硅晶圆的热阻。

$R_{package}$ ——GPU 封装架构贡献的热阻——取决于带盖还是无盖，以及 GPU 裸片与 TIM 之间有哪些材料。

$R_{coldplate}$ ——冷板自身贡献的热阻，由微通道几何结构、流道布置和流量决定。

$R_{graphene\ TIM}$ ——带盖封装中顶盖与裸片之间石墨烯 TIM 增加的热阻。

R_{lid} ——带盖 GPU 封装中集成均热盖板（顶盖）增加的热阻。

R_{total} ——GPU 晶极管层与冷却液之间热管理技术栈的总热阻。 R_{total} 越低，在相同的 Q 和 T_{inlet} 下 $T_{j\ max}$ 越低。

R_{TIM} ——GPU 封装与冷板之间热界面材料贡献的热阻。

TIM (Thermal Interface Material) (热界面材料) ——涂覆在 GPU 封装与冷板之间以改善热接触的物质。

R (Thermal Resistance) (热阻) ——衡量材料或部件阻碍热量流动程度的指标。热阻越低，热量传递越高效，最大结温越低。

Thermal Stack (热管理技术栈) ——负责将热量从 GPU 带走并排出设施的完整部件与系统链条——从芯片晶极管层一直到冷却塔。

T_{free} ——仅靠自然冷却可达到的最低 T_{inlet} 冷却液温度，由环境温度、相对湿度和 CDU 换热器效率决定。达到该温度点无需机械冷水机组耗电。

T_{inlet} ——进入冷板的液态冷却液温度。 T_{inlet} 每降低 1°C ， $T_{j\ max}$ 恰好降低 1°C 。

$T_{mechanical}$ ——机械冷水机组的额外制冷贡献，在 T_{free} 不足以达到目标 T_{inlet} 时使用。

Token (词元) ——AI 输出的基本单位。AI 模型生成的每一个词、每一张图、每一个结果都由 Token 组成。

Tokens/Watt ——AI 工厂的关键效率指标。衡量每消耗一瓦特功率产生多少单位 AI 输出 (Token)。越高越好。

T_{outlet} ——冷却液在吸收 GPU 热量后流出冷板时的温度。

Wet Bulb Temperature (湿球温度) ——衡量在给定气候下通过蒸发可实现多少制冷量的指标，综合考虑空气环境温度与相对湿度。湿球温度越低，自然冷却越有效，无需机械冷水机组即可实现的 T_{inlet} 越低。它是决定任一数据中心选址可用自然冷却量的关键气候变量。



实现更多.

