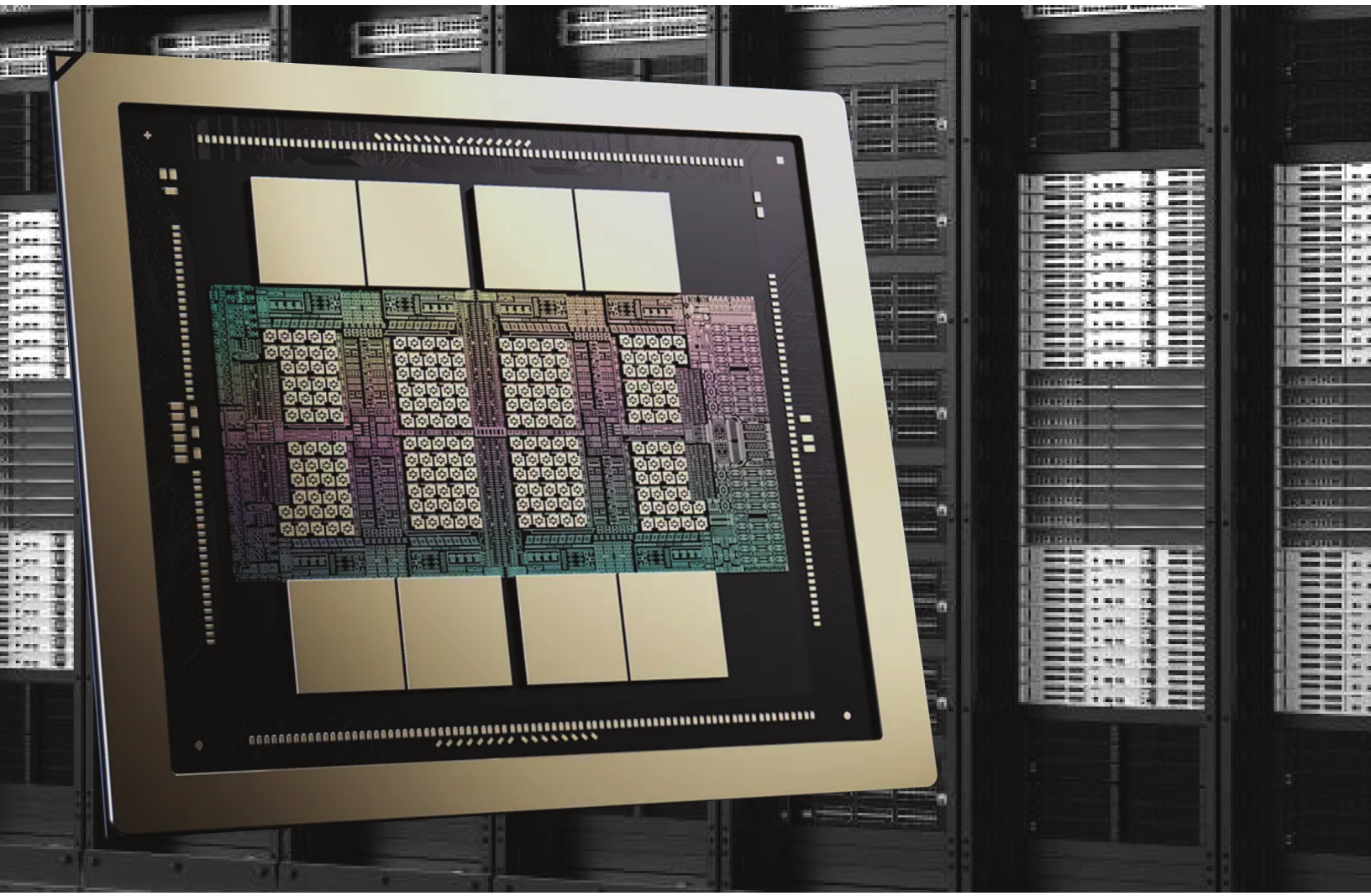
FRORE SYSTEMS

Do More.

技术笔记



NVIDIA Rubin 每瓦 Token 数 vs 最高裸片结温

本页特意留空。

NVIDIA Rubin 每瓦 Token 数 vs 最高裸片结温

摘要

本文建立了一条从 NVIDIA Hopper（H200）实测数据出发、推导 NVIDIA Rubin（R200）推理每瓦 Token 数效率曲线的分析链条。

已发表的 LLM 推理吞吐量实测数据（在不同功耗水平下测得）确立了 Hopper 基线；利用 Hopper 热管理技术栈模型，我们将功耗换算为最高结温；再利用按 Hopper 与 Rubin 所用 TSMC 4N 和 N3P 工艺节点标定的漏电流模型，将结果外推至 Rubin。

预计结果表明，Rubin 达到效率最优点的结温将略高于 Hopper，但超过最优点后效率下降得更快。对于 Hopper 和 Rubin，在 95°C 最高结温下保留的每瓦 Token 数效率均为峰值的 68.7%。

摘要结果

	H200（实测）	R200（预计）
最优效率对应的结温	62.5°C	65.5°C
超过最优点后的效率下降	−0.96%/°C	−1.06%/°C
95°C 最高结温下保留的效率	峰值的 68.7%	峰值的 68.7%

本文各节所依据的证据类别不同，全文均按以下方式标注：

第 1–2 节：Hopper 实测推理数据

第 3 节：Hopper 热模型

第 4 节：已确立的半导体物理

第 5 节：预计 Rubin 每瓦 Token 数效率

注：相关符号与缩略语见本文末尾。

第 1 节：Hopper 实测推理数据

Hopper 实测推理数据引自已发表论文：Mayr, M.、Wind, S.、Schröder, L.、Hager, G.、Köstler, H. 与 Wellein, G.（2026），《AI Application Benchmarking: Power-Aware Performance Analysis for Vision and Language Models》，arXiv:2603.16164。该研究出自 NHR@FAU 与 Friedrich-Alexander-Universität Erlangen-Nürnberg 的性能工程团队。

该论文表 4 报告了 LLaMA 3 8B 在 SGLang 上服务的单 GPU 推理吞吐量，测得于一个 4 GPU H200 节点、功耗状态 P0，GPU 功耗水平从 200 W 扫描至 700 W。

用吞吐量除以有效功耗（即功耗上限减去 70 W 空闲基线）可直接得到每瓦 Token 数。（减去空闲功耗基线是合理的，因为空闲功耗无论负载如何都会产生，且不产生 Token；若将其计入，会在低功耗端最严重地低估效率，因为在低功耗端空闲功耗占总功耗的份额更大。）

功耗水平（W）	每 GPU Token 数	每瓦 Token 数	占峰值百分比
200	571.8	4.398	78.8%
300	1255.0	5.457	97.7%
350.4（内插）	—	5.583	100.0%（峰值）
400	1804.2	5.467	97.9%
500	2063.7	4.799	86.0%
600	2289.7	4.320	77.4%

Frore Systems 版权所有 2026。规格如有变更，恕不另行通知。专利申请中。

700	2417.5	3.837	68.7%

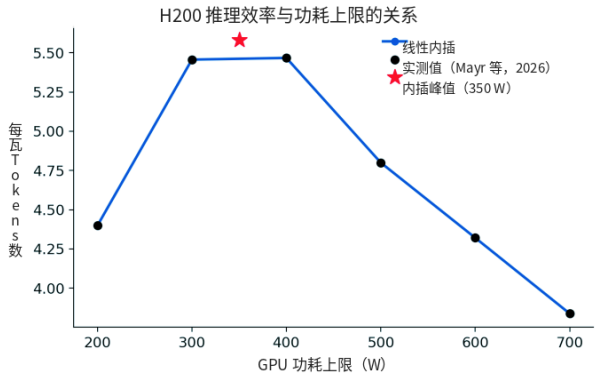


图 1：H200 LLM 推理每瓦 Token 数效率与功耗水平的关系，直接由实测推理吞吐量计算得出（Mayr 等，2026，表 4）。

采样最高的两个点 300 W 与 400 W 相差在 0.2% 以内，因此真实最优点位于两者之间：三次内插将其定位在 350.4 W。

第 2 节：Hopper 热模型

本节对 Hopper 直接芯片液冷热管理技术栈建模，并利用其热阻计算图 1 中每个功耗水平对应的 GPU 最高结温。

Hopper 热管理技术栈建模为五个串联热阻：

1. 硅裸片本身
2. TIM1（裸片至盖板）
3. 盖板
4. PTM7950 TIM2，以及
5. **冷板**——冷却液为 PG25，入液温度 30°C，流量 1 LPM。裸片、TIM1 与盖板共同构成封装内部结到壳热阻（θ_{jc}）。

冷板结合了对流散热与冷却液自身的显热温升；两者不可作为串联元件拆分，因为它们描述的是同一热量从冷板进入冷却液流的过程。裸片热阻按 R = t/(k×A) 计算：0.7 mm 厚度、814 mm² 裸片面积，k ≈ 110 W/m·K——这是硅在工作温度下的导热系数，而非 300 K 下引用的约 148 W/m·K，后者会使裸片热阻低估约三分之一。

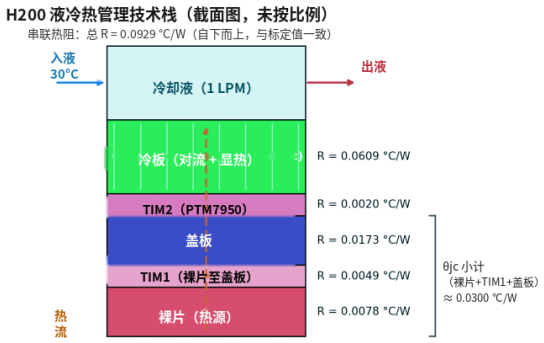


图 2：Hopper 直接芯片液冷热管理技术栈（截面图，未按比例）。自下而上串联热阻合计 0.0929°C/W，与标定总值一致。

自下而上的热阻估算合计为 0.0929°C/W（裸片 0.0078 + TIM1 0.0049 + 盖板 0.0173，θ_{jc} 小计 0.0300，再加 TIM2 0.0020 与冷板 0.0609）。

结温计算如下：

$$T_{jmax}(P) = 30°C + P \times R_{total}\ °C/W$$

功耗水平（W）	T_{jmax}（°C）	每有效瓦特 Token 数	占峰值百分比
200	48.6	4.398	78.8%
300	57.9	5.457	97.7%
350.4（内插）	62.5	5.583	100.0%（峰值）
400	67.1	5.467	97.9%
500	76.4	4.799	86.0%
600	85.7	4.320	77.4%
700	95.0	3.837	68.7%

Frore Systems 版权所有 2026。规格如有变更，恕不另行通知。专利申请中。

NVIDIA Rubin 每瓦 Token 数 vs 最高裸片结温

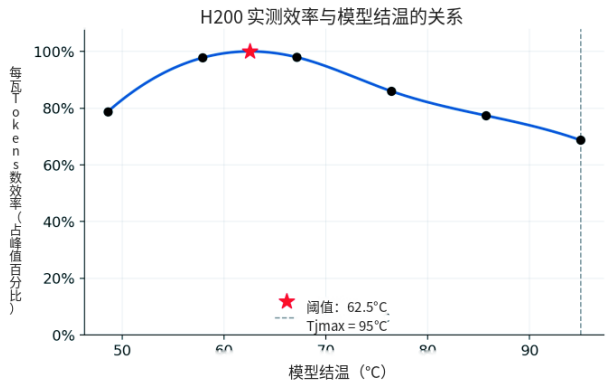


图 3：H200 实测推理效率（占峰值百分比）与模型结温的关系。

最优效率位于 T_{jmax} = 62.5°C，95°C 结温下的效率下降为 31.3%。

第 3 节：已确立的半导体物理

本节阐明 GPU 裸片最高结温升高时，每瓦 Token 数效率为何下降。

GPU 功耗由两部分组成：

1. 动态功耗：是晶体管在计算过程中开关所消耗的有用功耗，随时钟频率、电压以及任一时刻实际工作的晶体管数量而变化。

2. 漏电功耗：是浪费的功耗。它不论是否进行计算，都会持续流经晶体管，只产生热量而不贡献性能。

$$P_{total} = P_{dynamic} + P_{leakage}$$

这种指数型温度依赖性源于漏电流本身：漏电流流经名义上处于关断状态的晶体管，受具有负温度系数的阈值电压控制，因此结温的小幅升高会转化为不成比例的大幅漏电流增长。

因此，GPU 能效很大程度上取决于裸片温度。随着最高结温升高，根据阿伦尼乌斯方程，漏电功耗呈指数增长，T_{jmax} 每升高约 10°C 翻一倍。这会产生复合效应：温度越高，漏电越大，而漏电又产生更多热量，进一步推高温度。随着漏电功耗增加，动态功耗按比例下降。

此外，晶体管开关所需的功耗，T_{jmax} 每升高 10°C 也增加约 2%，在相同的动态功耗下带来更低的时钟频率和更低的 GPU 性能。

这些热损失在实际中通过动态电压频率调整（DVFS）来管理，即现代 GPU 在裸片温度升高时提高电压、向 GPU 提供更高功耗，以维持特定时钟频率的机制。当结温接近其最高额定限值、电压达到允许的最大值时，DVFS 会降低时钟频率，以减少动态功耗并使 T_{jmax} 保持在允许的最大值以下。

空闲功耗在此计为动态功耗而非漏电：静态功耗绝大部分是开关功耗——时钟、内存刷新、控制器、PLL——因此将其计入漏电会在每个工作点上高估漏电。因此，动态功耗等于等效有用功功耗加上空闲开销，在效率最优点处，两者基本构成全部功耗（280.4 W + 70 W = 350.4 W）。漏电永不消失，因此在阈值以下保持 1% 的名义下限，而非恰好为零。在 700 W TDP、95°C 下，GPU 处于满载，因此动态功耗由实测吞吐量推导，剩余 28.1% 为漏电。下限为假设值；95°C 结温下的漏电与动态功耗值图 3 推导得出。

Frore Systems 版权所有 2026。规格如有变更，恕不另行通知。专利申请中。

NVIDIA Rubin 每瓦 Token 数 vs 最高裸片结温

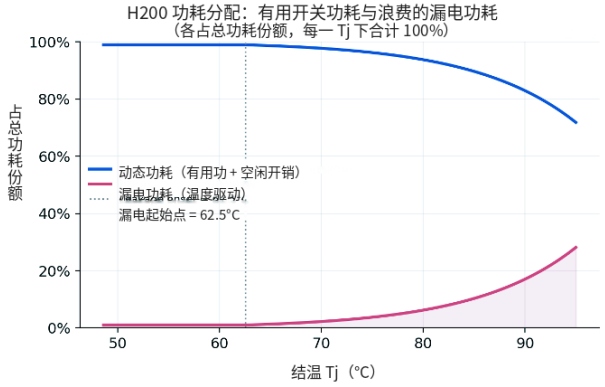


图 4 针对 H200 将这一点具体化，以各组成部分在给定结温下占总功耗的份额表示，因此两者之和始终为 100%。在 62.5°C 的阈值温度以下，漏电尚未成为主导项，约占总功耗的 1%。超过该阈值后，指数关系占据主导，到 95°C 的结温限值时，每消耗一瓦电约有 28% 花在漏电上而非做功。

第 4 节：预计 Rubin 每瓦 Token 数效率

方法

R200 曲线由实测 H200 曲线推导，使用单一自由参数，该参数表征 TSMC N3P 工艺（R200）相对于 4N（H200）漏电的温度敏感性提高。效率最优点向更高结温的移动，以及超过最优点后下降变陡，并不是两个需要分别估计的独立效应——两者都源于这一个量。

另有一个独立的工艺增益因子设定绝对效率水平，但它不影响曲线的形状或位置。

漏电敏感比（s = 1.1）

根据阿伦尼乌斯关系，结温每升高 10°C，漏电大约翻一倍。更小的工艺节点（N3P 对 4N）具有更高的晶体管密度和更短的沟道长度，模型设定其总体漏电敏感性更高：每 9.1°C 翻一倍而非 10°C，即敏感性为 s = 1.1 倍。同一条更陡的指数曲线在阈值以上对更小节点惩罚更重，同时也意味着其漏电在 T_{jmax} 以下下降得更快——而这正是它一开始就能在更高结温下达到效率最优点的原因。

将低于限值的漏电写作：

$$L(T_{jmax}) = L_{max} \times 2^{-(T_{jmax} - T_{jthresh})/D}$$

并要求两款 GPU 在相同的临界漏电份额处达到各自阈值，可得：

$$T_{jmax} - T_{jthresh}(H200) = s \times (T_{jmax} - T_{jthresh}(R200))$$

因此，阈值直接由 s 决定，下文用于加陡阈值后下降的热缩放因子 λ_T 就是 s——同一个参数，而非第二个参数。这比早期表述具有更强的内在一致性，早期表述先独立估计位移与加陡程度，再核对两者是否一致。

TSMC 的 N3P 与 4N 对比数据（相同功耗下的速度，或相同速度下的功耗）特意不用于设定 s。这些数据是在固定功耗或固定速度下测得的，而非在温度扫描下测得，因此表征的维度与此处缩放的维度不同，无法量化阈值移动了多远。在没有 N3P 相对 4N 的直接漏电敏感性实测数据的情况下，s = 1.1 是一项明确的建模假设，选取时持保守态度：它承认更小节点在漏电温度敏感性上确有实际但适度的提升，跨越一代工艺节点迭代。这一参数是在对该预计做任何敏感性研究时应当变化的参数。

H200 的阈值位于 95°C 限值以下 32.5°C；除以 s = 1.1 后，该跨度缩短为 29.5°C，从而 R200 的阈值位于 T_{jmax} = 65.5°C。在 R200 自身 45°C 冷却液温度的框架下，这相当于 f = (65.5 − 45) / 50.0 = 0.41（占 2300 W TDP），即约 942 W。

Frore Systems 版权所有 2026。规格如有变更，恕不另行通知。专利申请中。

注意，阈值在结温意义上被视为硅的内在特性，之后才换算到各 GPU 自身的功耗框架中。由于 R200 使用更热的 45°C 冷却液入液温度，它在额定功耗的更低比例处就达到同一结温——这一差异是冷却环路消耗的裕量，而非硅本身的差异。热缩放因子（λ_T = 1.1）这不是一项额外假设：它就是同一个漏电敏感比 s，现在作为温度轴上的缩放因子出现。R200 从阈值到限值只有 29.5°C，而 H200 为 32.5°C，因此在更短的跨度内达到同一终点需要按比例更陡的下降——而这两个跨度之比按构造就是：

$$\lambda_T = (95 - 62.5) / (95 - 65.5) = 32.500 / 29.500 = 1.1 = s$$

这一构造的一个结果是，两款 GPU 在 95°C 限值处保留的各自峰值效率份额相同。这一等效率终点在该模型的早期表述中是一项假设的边界条件；在这里，它由单参数漏电构造自然得出，而非强加。这一结果在独立的理由上仍然合理——95°C 由可靠性认证设定，各代产品都以可比的寿命末期裕量为目标。

两款 GPU 共用同一结温限值意味着 Rubin 需要强得多的散热方案：从 45°C 入液温度、在 2300 W 下达到 95°C，需要热阻接近 0.0217°C/W，约为 H200 标定的 0.0929°C/W 的 4.3 分之一。这一要求与该功耗级别预期的更高冷却液流量和改进的冷板设计一致，但它是该预计的一项假设，而非推导结果。

与第 3 节的物理机制一致，λ_T 仅在阈值以上应用。在阈值以下、漏电尚未成为主导项的区间，实测 H200 行为不经缩放直接平移：

$$\%_{eff}(T_{jmax}, R200) = \%_{eff}(T_{jequiv}, H200)$$

$$T_{jequiv} = 62.5 + (T_{jmax} - 65.5)$$

$$Tj_{equiv} = 62.5 + (Tj_{max} - 65.5)，当 Tj_{max} \le 65.5°C（阈值以下）$$

$$Tj_{equiv} = 62.5 + (Tj_{max} - 65.5) \times \lambda T，当 Tj_{max} > 65.5°C（阈值以上）$$

阈值推导（分步）

步骤	量	值
1	H200 实测效率峰值	350.4 W（占 700 W TDP）
2	H200 在该峰值处的结温	30 + 350.4 × 0.0929 = 62.54°C
3	H200 从阈值到 T_{jmax} 的跨度	95 − 62.54 = 32.46°C
4	R200 跨度，按漏电敏感性缩短（/ s = 1.1）	32.46 / 1.1 = 29.51°C
5	R200 阈值结温	95 − 29.51 = 65.49°C
6	R200 在自身 45°C 框架中达到该阈值的位置	(65.49 − 45) / 50.0 = 占 TDP 的 0.410，约 942 W

参数、来源与对曲线的影响

参数	值	来源	对曲线的影响
s	1.1	漏电温度敏感比，N3P vs 4N（假设）	同时设定阈值位移与阈值后加陡程度
λ_T	1.1	与 s 完全相等——不是独立参数	加陡最优点后的下降
$\gamma_{derived}$	1.091	派生诊断量：温升比的隐含比值	无——仅为保持连续性而报告

Frore Systems 版权所有 2026。规格如有变更，恕不另行通知。专利申请中。

在 R200 阈值与斜率现已确定的情况下，同样的构造可以反向读出为功耗分配。图 5 展示了 R200 预计的动态与漏电功耗曲线。

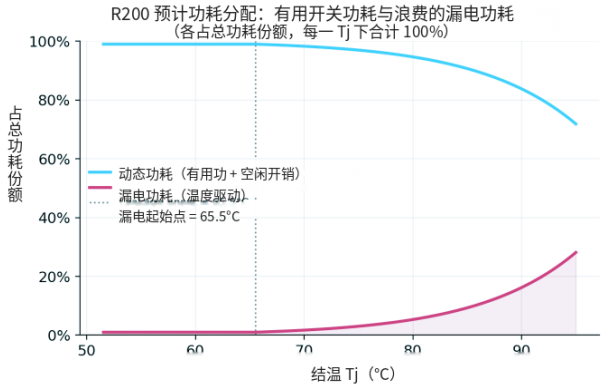


图 5：R200 预计的动态与漏电功耗曲线，按各结温下占总功耗的份额表示。该分配在 65.5°C 阈值以下保持 1% 的下限，超过阈值后发散，达到与 H200 相同的终点，但跨度更短——29.5°C 对 32.5°C，这正是 s 所隐含的更陡上升。

分析

由于更小节点的漏电在结温限值以下下降得更陡，它在漏电占据主导之前，于更高的结温处达到效率最优点——最优点从 62.5°C 移至 65.5°C，移动了 3.0°C。而由于同一条指数曲线在上升段同样更陡，最优点过后的下降加陡至 −1.06%/°C，而 H200 为 −0.96%/°C。以更尖锐的悬崖换取更宽的高效窗口，正是单一漏电敏感性差异的特征，而非两个独立效应。

两者的幅度按构造都不大。漏电温度敏感性在一代工艺节点迭代中提高 10%，是一项刻意保守的假设，相应的位移也较小；更激进的假设会按比例拓宽高效窗口并加深悬崖，但没有实测依据支持偏向某一种假设。工艺与架构层面的效率提升另行处理：它们提高绝对效率水平，但不移动峰值。

T_j（°C）	R200 占峰值百分比
55	88.3%
60	96.9%
65.5（阈值）	100.0%（峰值）
75	90.7%
81	82.5%
95（TDP）	68.7%

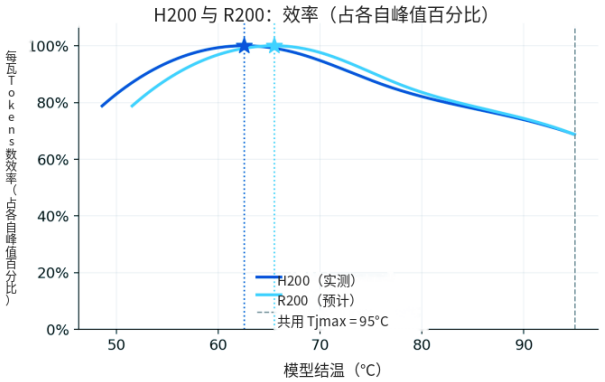


图 6：H200 与 R200 推理效率（占各自峰值百分比）。更晚的峰值与阈值后更陡的下降都源于同一个量——两个工艺节点之间的漏电温度敏感性差异。由于每条曲线都按各自峰值归一化，工艺增益被抵消，对曲线形状没有影响。

这一变换分离出了一个物理效应的预期一阶影响——两个工艺节点之间的漏电温度敏感性差异——同时有意忽略了架构变化、内存系统差异、时钟频率策略以及未来的电源管理算法。预计的 Rubin 曲线应理解为一项单参数敏感性研究，而非对最终产品的预测。

Frore Systems 版权所有 2026。规格如有变更，恕不另行通知。专利申请中。

第 5 节：预计 R200 曲线

最终预计曲线：R200 的推理效率，从 H200 中分离出来，以占自身峰值的百分比表示，对应结温。

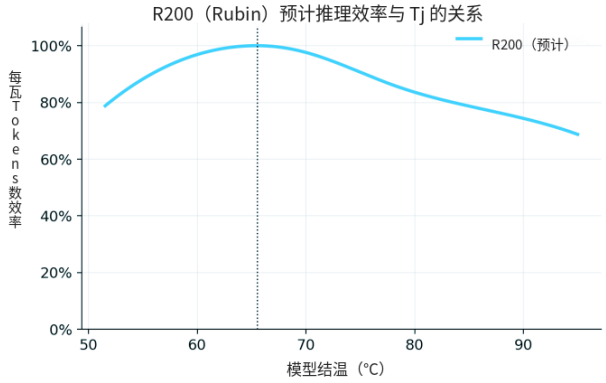


图 7：R200（Rubin）预计推理效率与 T_j 的关系（占自身峰值百分比）。更晚的峰值与更陡的下降都源于更小节点更高的漏电温度敏感性；曲线经过归一化，因此工艺增益不进入曲线。

第 6 节：结论

主要结论如下：

- 以每有效瓦特表示的 H200 实测推理效率，在约 350 W 处达到峰值——在标定的热管理技术栈下对应 T_{jmax} = 62.5°C。
- 由于基线针对目标负载实测并以每有效瓦特表示，效率曲线既不需要从计算基准测试换算的代理转换，也不需要厂商吞吐量比值。
- 在所述假设下，预计 Rubin 将在高于 Hopper 的结温处达到效率最优点，但超过该点后效率下降更快——两者都是单一参数（两个工艺节点之间的漏电温度敏感性差异）的结果。工艺层面的效率增益提高绝对效率水平，但不移动峰值或改变曲线形状。
- 该预计对热管理技术栈的误差具有稳健性。漏电敏感比 s 是输入量，阈值按其与 T_{jmax} 之间的跨度比例定位，因此 R_{total} 的误差会使所有绝对结温一起移动，而不改变任一条曲线的形状、两个峰值的相对位置或 λ_T。

Frore Systems 版权所有 2026。规格如有变更，恕不另行通知。专利申请中。

NVIDIA Rubin 每瓦 Token 数 vs 最高裸片结温

符号与缩略语

P——GPU 功耗上限（W）

TDP——热设计功耗，即额定最大值（H200 为 700 W；R200 为 2300 W）

f——功耗占 TDP 的比例（P / TDP）

$\mathbf{f_{thresh}}$**——**效率达到峰值时的功耗比例（H200 为 0.501）

$\mathbf{T_j}$**——**结温，即芯片上的硅温度（°C）

$\mathbf{T_{coolant}}$**——**冷却液入液温度（H200 为 30°C；R200 为 45°C）

$\mathbf{T_{jthresh}}$**——**效率达到峰值、超过该点后漏电导致效率下降的结温（H200 为 62.5°C；R200 为 65.5°C）

$\mathbf{T_{jmax}}$**——**TDP 下的结温限值（95°C，两款 GPU 共用）

$\mathbf{T_{jequiv}}$**——**用于读取给定 R200 温度的 H200 曲线上的点（°C）

$\mathbf{R_{total}}$**——**结到冷却液总热阻（H200 为 0.0929°C/W）

s——漏电温度敏感比，N3P vs 4N——模型的单一自由参数（1.1）

$\boldsymbol{\lambda_T}$**——**温度轴上的热缩放因子；与 s 完全相等（1.1）

D——漏电翻倍间隔（阿伦尼乌斯）（H200 为 10°C；R200 为 9.1°C）

$\mathbf{\%_{eff}}$**——**效率占该 GPU 自身峰值的百分比（%）

TIM1、TIM2——热界面材料：裸片至盖板与盖板至冷板（°C/W）

LPM——冷却液流量，升/分钟（假设为 1 LPM）

参考文献

1. AI Application Benchmarking: Power-Aware Performance Analysis for Vision and Language Models. Mayr, M.、Wind, S.、Schröder, L.、Hager, G.、Köstler, H.、Wellein, G.（2026）。arXiv:2603.16164 [cs.PF]。https://arxiv.org/abs/2603.16164

2. Honeywell / Solstice Advanced Materials. PTM7950 Phase Change Material datasheet. https://ws.solstice.com/us/en/products/advanced-semiconductor-materials/thermal-management/phase-change materials/ptm7950

3. DataCenterKnowledge. Skived Coldplates Technical Brief. https://www.datacenterknowledge.com/cooling/skived-coldplates-technical-brief

4. Amkor Technology. Characterizing Junction-to-Case Thermal Resistance and Its Impact on End-Use Applications. https://www.circuitnet.com/news/uploads/2/Amkor_Characterizing_thermal_resisitance_White_Paper.pdf

5. TSMC 2022 Technology Symposium, as reported in: Tom's Hardware, TSMC Outlines 3nm Roadmap: Five FinFlex Nodes — N3E versus N5 at 18% higher speed at equal power, or 34% lower power at equal speed. https://www.tomshardware.com/news/tsmc-outlines-3nm-roadmap

6. SemiWiki. Three Key Takeaways from the 2022 TSMC Technical Symposium — corroborating the N3E-versus-N5 PPA figures. https://semiwiki.com/semiconductor-manufacturers/tsmc/314262-three-key-takeaways-from-the-2022-tsmc-technical-symposium/

7. Rambus Blog. Understanding Dennard Scaling. https://www.rambus.com/blogs/understanding-dennard-scaling-2/

第 1–2 节依据已发表的实测数据，第 3 节的半导体物理已确立；热管理技术栈与所有 Rubin 预计均由本文记录的明确假设推导得出。

Frore Systems 版权所有 2026。规格如有变更，恕不另行通知。专利申请中。

NVIDIA Rubin 每瓦 Token 数 vs 最高裸片结温

Frore Systems 版权所有 2026。规格如有变更，恕不另行通知。专利申请中。



Do More.



V1.1