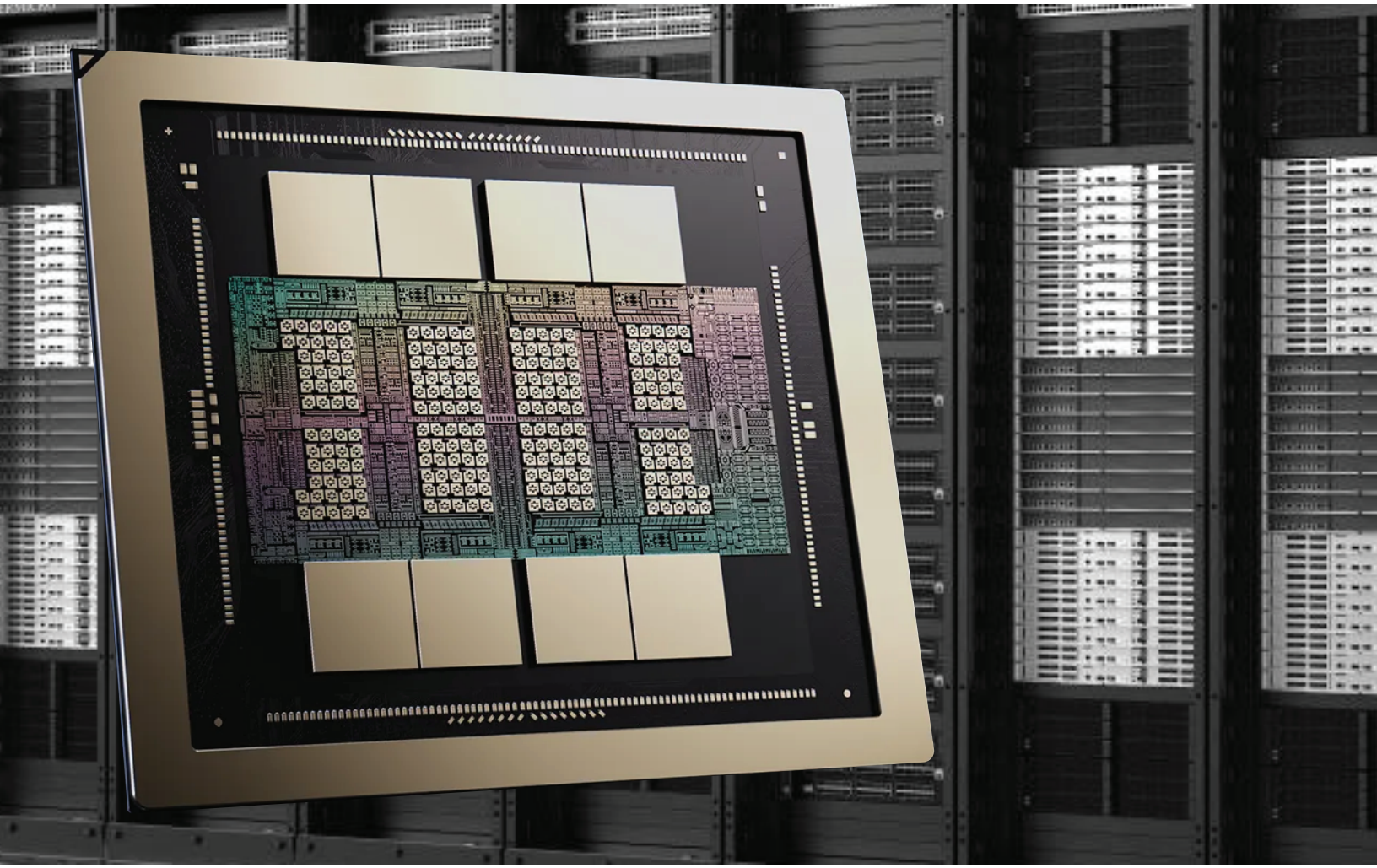
FRORE SYSTEMS

Do More.

Technical Note



NVIDIA Rubin Tokens/Watt vs Max Die Junction Temperature

NVIDIA Rubin Tokens/Watt vs Max Die Junction Temperature

ABSTRACT

This paper develops an analytical chain from measured NVIDIA Hopper (H200) data to a projected NVIDIA Rubin (R200) inference Tokens/Watt efficiency curve.

Published measurements of LLM inference throughput at various power draw levels establishes the Hopper baseline; using a Hopper thermal stack model, we convert power draw to max junction temperature; and leakage current models calibrated to Hopper and Rubin TSMC 4N and N3P process nodes extends the result to Rubin.

The projection indicates that Rubin would reach its efficiency optimum at a slightly higher junction temperature than Hopper, but declines more steeply beyond it. For both Hopper and Rubin, the Tokens/Watt efficiency retained at 95°C max junction temperature is 68.7% of peak.

SUMMARY RESULTS

	H200 (measured)	R200 (projected)
Junction temperature at optimum efficiency	62.5°C	65.5°C
Decline in efficiency beyond optimum	-0.96%/°C	-1.06%/°C
Efficiency retained at 95°C max junction temperature	68.7% of peak	68.7% of peak

The sections in this paper rest on different classes of evidence, and are labelled accordingly throughout:

Section 1-2: Measured Hopper inference data

Section 3: Hopper Thermal Model

Section 4: Established Semiconductor physics

Section 5: Projected Rubin Tokens/Watt efficiency

Note:Relevant symbols and abreviations are included at the end of this paper

SECTION 1:

MEASURED HOPPER INFERENCE DATA

The measured Hopper inference data is cited from the published paper, Mayr, M., Wind, S., Schröder, L., Hager, G., Köstler, H. & Wellein, G. (2026), "AI Application Benchmarking: Power-Aware Performance Analysis for Vision and Language Models," arXiv:2603.16164. This work is from the performance-engineering group at NHR@FAU and Friedrich-Alexander-Universität Erlangen-Nürnberg.

Table 4 of that paper reports per-GPU inference throughput for LLaMA 3 8B served with SGLang, measured on a 4-GPU H200 node in power state P0, with the GPU power level sweep from 200 W to 700 W.

Dividing throughput by the active power — the power cap less a 70 W idle baseline — gives Tokens/Watt directly. (Subtracting the idle power baseline makes sense because idle power is drawn regardless of workload and produces no tokens; including it would understate efficiency most severely at the low-power end, where it is the larger share of the total.)

Power Level (W)	Tokens/GPU	Tokens/Watt	% of Peak
200	571.8	4.398	78.8%
300	1255.0	5.457	97.7%
350.4 (Interpolated)	—	5.583	100.0% (peak)
400	1804.2	5.467	97.9%
500	2063.7	4.799	86.0%
600	2289.7	4.320	77.4%
700	2417.5	3.837	68.7%

Frore Systems Copyright 2026. Specification subject to change. Patents Pending.

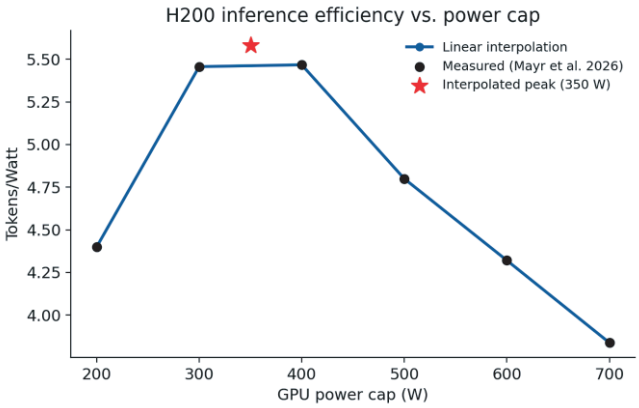


Figure 1: H200 LLM inference Tokens/Watt efficiency vs. power level, computed directly from measured inference throughput (Mayr et al. 2026, Table 4).

The two highest sampled points, 300 W and 400 W, are within 0.2% of each other, so the true optimum lies between them: cubic interpolation places it at 350.4 W.

SECTION 2:

HOPPER THERMAL MODEL

In this section, we model the Hopper direct-to-chip liquid cooling thermal stack and utilize its thermal resistance to calculate the GPU max junction temperature for each power level shown in Figure 1.

The Hopper thermal stack is modeled as five resistances in series:

1. The silicon die itself
2. TIM1 (die-to-lid)
3. The lid
4. PTM7950 TIM2, and
5. A coldplate — with PG25 coolant at 30°C inlet and 1 LPM flow rate. Die, TIM1, and lid together form the package-internal junction-to-case resistance (theta$_{jc}$).

The coldplate combines convective removal with the coolant's own sensible-heating (caloric) rise; the two are not separable as series elements, since both describe the same heat leaving the coldplate into the coolant stream.

The die resistance is computed as R = t/(k×A): 0.7 mm over an 814 mm^2 die at k ≈ 110 W/m·K — silicon's thermal conductivity at operating temperature, not the ~148 W/m·K figure quoted at 300 K, which would understate the die resistance by about a third.

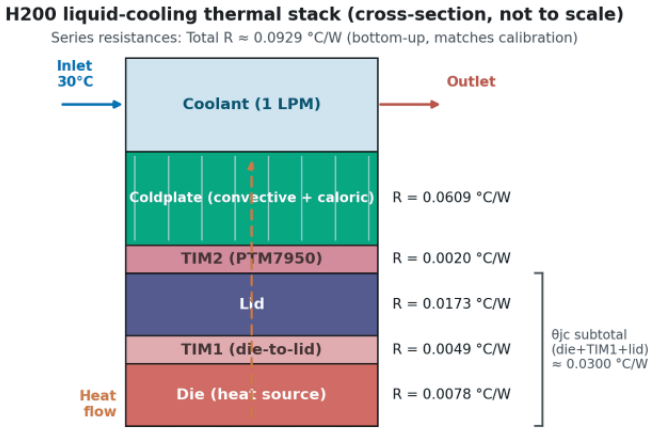


Figure 2: Hopper direct-to-chip liquid-cooling thermal stack (cross-section, not to scale). Series resistances sum to 0.0929 °C/W bottom-up, matching the calibrated total

The bottom-up thermal resistance estimate sums to 0.0929 °C/W (die 0.0078 + TIM1 0.0049 + lid 0.0173, a θjc subtotal of 0.0300, plus TIM2 0.0020 and coldplate 0.0609). Junction temperature follows as

$T_{j\,max}(P) = 30°C + P \times R_{total}$ °C/W.

Power Level (W)	$T_{j\,max}$ (°C)	Tokens/ Active Watt	% of Peak
200	48.6	4.398	78.8%
300	57.9	5.457	97.7%
350.4 (Interpolated)	62.5	5.583	100.0% (peak)
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 Copyright 2026. Specification subject to change. Patents Pending.

NVIDIA Rubin Tokens/Watt vs Max Die Junction Temperature

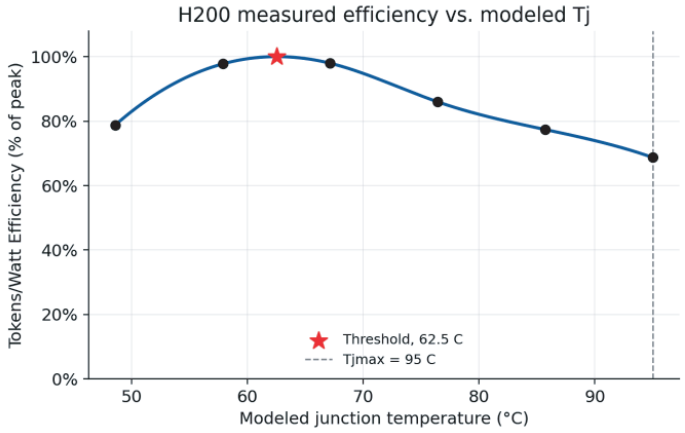


Figure 3: H200 measured inference efficiency as % of peak, against modeled junction temperature

The optimum efficiency sits at $T_{j\,max}$ = 62.5°C, and the reduction in efficiency at 95°C junction temperature is 31.3%.

SECTION 3:

ESTABLISHED SEMICONDUCTOR PHYSICS

This section establishes why there is a decline in Tokens/Watt efficiency at higher GPU die max junction temperatures.

GPU power consumption is the sum of two components:

1. **Dynamic power** is the useful power consumed by the transistors switching during computation, and it scales with clock frequency, voltage, and the number of active transistors performing work at any given moment.

2. **Leakage power** is wasted power. It flows continuously through transistors regardless of any computation occurring, contributing heat without contributing to performance.

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

This exponential temperature dependence traces to the leakage current itself: it flows through nominally-off transistors, gated by a threshold voltage with a negative temperature coefficient, so small increases in junction temperature translate into disproportionately larger leakage currents.

Thus, GPU power efficiency is heavily dependent on die temperature. As max junction temperature rises, as per the Arrhenius equation, leakage power increases exponentially and doubles for approximately every 10°C increase in $T_{j\,max}$. This creates a compounding effect: higher temperatures drive higher leakage, which in turn generates more heat, pushing temperatures higher still. As leakage power increases, dynamic power decreases proportionally.

Moreover, power required for transistor switching also increases by about 2% for every 10°C increase in $T_{j\,max}$, delivering lower clock frequencies and lower GPU performance for the same dynamic power.

These thermal penalties are managed in practice through Dynamic Voltage and Frequency Scaling (DVFS), the mechanism by which modern GPUs increase voltage and deliver higher power to the GPU as required to maintain a certain clock frequency when die temperature increases. As junction temperature approaches its maximum rated limit and voltage reaches its maximum allowed value, DVFS throttles clock frequency to reduce dynamic power and keep $T_{j\,max}$ below its maximum allowed.

Figure 4 makes this concrete for H200, expressed as each component's share of the total power drawn at a given junction temperature, so the two always sum to 100%. Below the threshold temperature of 62.5 °C, leakage is not yet the dominant term and holds at about 1% of total power. Above it the exponential takes over, and by 95 °C junction limit roughly 28% of every watt drawn is being spent on leakage rather than on work.

Idle power counts as dynamic here, not as leakage: the quiescent draw is overwhelmingly switching power — clocks, memory refresh, controllers, PLLs — so charging it to leakage would overstate leakage at every operating point. Dynamic power is therefore useful-work-equivalent power plus idle overhead, and at the efficiency optimum

Frore Systems Copyright 2026. Specification subject to change. Patents Pending.

those two account for essentially the entire draw (280.4 W + 70 W = 350.4 W). Leakage never vanishes, so a nominal floor of 1% is held below the threshold rather than exactly zero. At 700 W TDP and 95 °C the GPU is fully utilized, so dynamic power follows from the measured throughput and leaves 28.1% as leakage. The floor is assumed; the leakage and dynamic power values at the 95 °C junction temperature are derived from Figure 3.

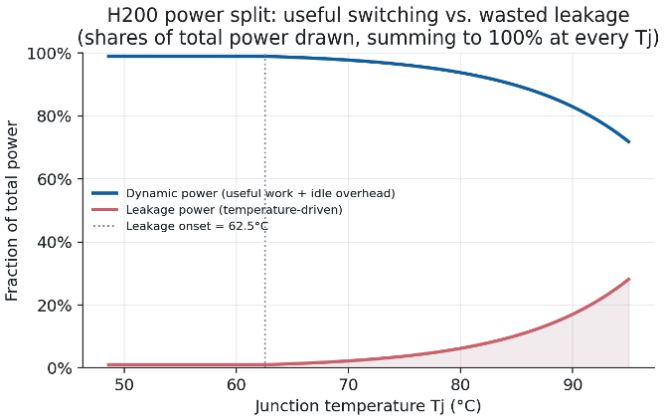


Figure 4: H200 power split into useful switching and temperature-driven leakage, as shares of the total power drawn at each junction temperature. The split is stable below the 62.5°C threshold and diverges above it, as leakage grows exponentially while dynamic power does not.

SECTION 4:

PROJECTED RUBIN TOKENS/WATT EFFICIENCY

Methodology

The R200 curve is derived from the measured H200 curve using a single free parameter that captures the increased temperature sensitivity of leakage in TSMC's N3P process (R200) relative to 4N (H200). The shift of the efficiency optimum to a higher junction temperature and the steeper decline beyond it are not two independent effects to be estimated separately — both follow from that one quantity. A separate process-gain factor sets the absolute efficiency level, but it does not affect the shape or position of the curve.

Leakage sensitivity ratio (s = 1.1). Per the Arrhenius relationship, leakage roughly doubles for every 10°C rise in junction temperature. A smaller node (N3P vs. 4N), with higher transistor density and shorter channel lengths, is modeled to carry greater aggregate leakage sensitivity: doubling over 9.1°C instead of 10°C, i.e. s = 1.1 times more sensitive. The same steeper exponential that punishes the smaller node above its threshold also means its leakage reduces faster below T_{jmax} — which is what buys it head-room to reach its efficiency optimum at a higher junction temperature in the first place.

Writing leakage below the limit as

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

and requiring both GPUs to reach their threshold at the same critical leakage fraction, gives:

$$T_{j\,max} - T_{j\,thresh}(H200) = s \times (T_{j\,max} - T_{j\,thresh}(R200))$$

The threshold therefore follows directly from s, and the thermal scaling factor λT used below to steepen the post-threshold decline is s — the same parameter, not a second one. This is a stronger internal consistency than the earlier formulation, which estimated the shift and the steepening independently and then checked that they agreed.

TSMC's N3P-vs-4N figures (for speed at equal power, or power at equal speed) are deliberately not used to set s. Those are measured at fixed power or fixed speed, not across a temperature sweep, so they characterize a different axis from the one being scaled here and cannot quantify how far the threshold moves. Absent a direct leakage-sensitivity measurement for N3P against 4N, s = 1.1 is an explicit modeling assumption, chosen conservatively: it credits the smaller node with a real but modest increase in leakage temperature-sensitivity across a single node step. It is the parameter to vary in any sensitivity study of this projection.

H200's threshold sits 32.5°C below the 95°C limit; dividing by s = 1.1 shortens that span to 29.5°C, placing R200's threshold at $T_{j\,max}$ = 65.5°C. In R200's own 45°C-coolant frame this is f = (65.5 − 45) / 50.0 = 0.41 of 2300 W TDP, or roughly 942 W.

Frore Systems Copyright 2026. Specification subject to change. Patents Pending.

NVIDIA Rubin Tokens/Watt vs Max Die Junction Temperature

Note that the threshold is treated as an intrinsic silicon characteristic in junction-temperature terms, and only then converted into each GPU's own power frame. Because R200 runs a warmer 45°C coolant inlet, it reaches that same junction temperature at a lower fraction of its rated power than H200 does — the difference is headroom consumed by the coolant loop, not by the silicon. Thermal-scaling factor (λT = 1.1). This is not an additional assumption: it is the same leakage sensitivity ratio s, appearing now as a scaling factor on the temperature axis. R200 covers only 29.5°C from threshold to limit against H200's 32.5°C, so reaching the same endpoint over the shorter span requires a proportionally steeper decline — and the ratio of those spans is s by construction:

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

A consequence of this construction is that both GPUs retain the same fraction of their own peak efficiency at the 95°C limit. That equal-efficiency endpoint was an assumed boundary condition in earlier formulations of this model; here it falls out of the single-parameter leakage construction rather than being imposed. It remains a reasonable expectation on independent grounds — 95°C is set by reliability qualification, and comparable end-of-life margins are targeted across generations.

Holding the junction limit common to both GPUs implies a substantially stronger cooling solution for Rubin: reaching 95°C from a 45°C inlet at 2300 W requires a stack resistance near 0.0217°C/W, roughly 4.3 times lower than the 0.0929°C/W calibrated for H200. This is consistent with the higher coolant flow rates and improved cold-plate designs expected for that power class, but it is an assumption of the projection rather than a derived result.

Consistent with the physics of Section 3, λT is applied only above the threshold. Below it, where leakage is not yet the dominant term, the measured H200 behavior is translated without scaling:

$$\%eff(T_{j\,max}, R200) = \%eff(T_{j\,equiv}, H200)$$

$$T_{j,equiv} = 62.5 + (T_{j\,max} - 65.5)$$
$$\text{for } T_{j\,max} \le 65.5°C \quad \text{(below threshold)}$$

$$T_{j\,equiv} = 62.5 + (T_{j\,max} - 65.5) \times \lambda T$$
$$\text{for } T_{j\,max} > 65.5°C \quad \text{(above threshold)}$$

The threshold derivation, step by step:

Step	Quantity	Value
1	H200 measured efficiency peak	350.4 W of 700 W TDP
2	H200 junction temperature at that peak	30 + 350.4 x 0.0929 = 62.54 C
3	H200 span from threshold to $T_{j\,max}$	95 - 62.54 = 32.46 C
4	R200 span, shortened by leakage sensitivity (/ s = 1.1)	32.46 / 1.1 = 29.51 C
5	R200 threshold junction temperature	95 - 29.51 = 65.49 C
6	Where R200 reaches it, in its own 45 C frame	(65.49 - 45) / 50.0 = 0.410 of TDP, ~942 W

The parameters, their sources, and their effects:

Parameter	Value	Source	Effect on the Curve
s	1.1	Leakage temperature-sensitivity ratio, N3P vs 4N (assumed)	Sets both the threshold shift and the post-threshold steepening
λT	1.1	Identically equal to s — not a separate parameter	Steepens the decline beyond the optimum
γderived	1.091	Derived diagnostic: implied ratio of temperature rises	None — reported for continuity only

With the R200 threshold and slope now fixed, the same construction can be read back out as a power split. Figure 5 shows the projected dynamic and leakage power curves for R200.

Frore Systems Copyright 2026. Specification subject to change. Patents Pending.

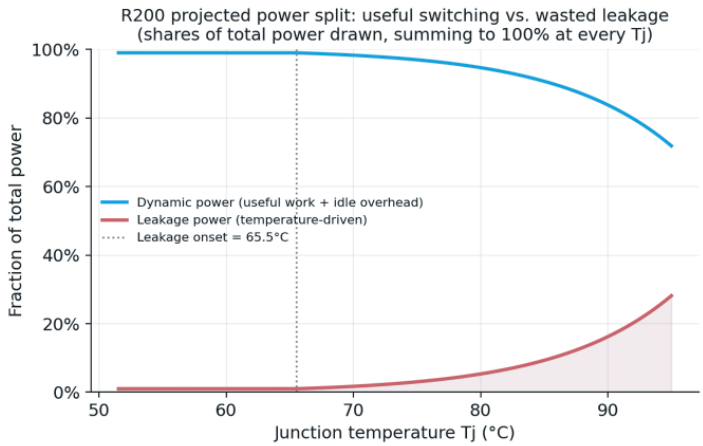


Figure 5: R200 projected dynamic and leakage power curves, as shares of the total power drawn at each junction temperature. The split holds at the 1% floor below the 65.5°C threshold and diverges above it, reaching the same endpoint as H200 but over a shorter span — 29.5°C against 32.5°C — which is the steeper rise s implies.

T_j (°C)	R200% of Peak
55	88.3%
60	96.9%
65.5 (threshold)	100.0% (peak)
75	90.7%
81	82.5%
95 (TDP)	68.7%

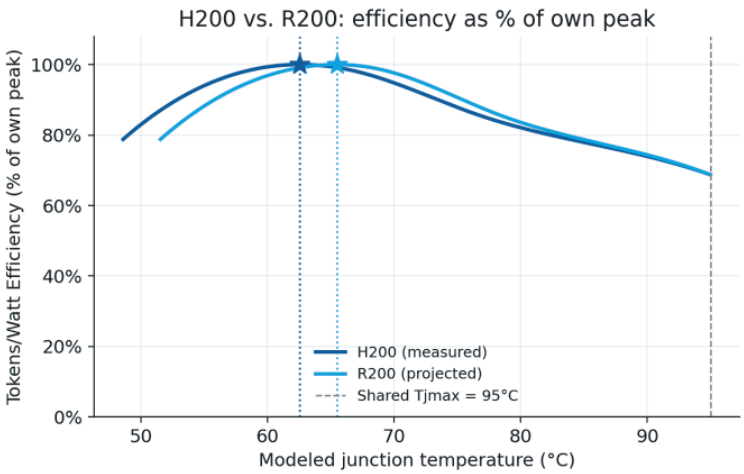


Figure 6: H200 vs. R200, inference efficiency as % of own peak. Both the later peak and the steeper post-threshold decline trace to a single quantity — the leakage-temperature-sensitivity difference between the two nodes. Because each curve is normalized to its own peak, the process gain cancels and has no effect on either shape.

Analysis

Because the smaller node's leakage falls off more steeply below the junction limit, it reaches its efficiency optimum at a higher junction temperature before leakage takes over — moving the optimum from 62.5°C to 65.5°C, a shift of 3.0°C. And because that same exponential is steeper on the way up, the decline past the optimum steepens to -1.06%/°C against H200's -0.96%/°C. A wider efficient window traded for a sharper cliff is the signature of a single leakage-sensitivity difference, not two independent effects.

The magnitude of both is modest, by construction. A 10% increase in leakage temperature-sensitivity across one node step is a deliberately conservative assumption, and the resulting shift is correspondingly small; a more aggressive assumption would widen the efficient window and deepen the cliff proportionally, but there is no measured basis to prefer one. Process- and architecture-level efficiency improvements are handled separately: they raise the absolute efficiency level without moving the peak.

This transformation isolates the expected first-order impact of one physical effect — the leakage temperature-sensitivity difference between the two nodes — while intentionally neglecting architectural changes, memory-system differences, clock-frequency policy, and future power-management algorithms. The projected Rubin curve should be interpreted as a single-parameter sensitivity study rather than a prediction of the final product.

Frore Systems Copyright 2026. Specification subject to change. Patents Pending.

NVIDIA Rubin Tokens/Watt vs Max Die Junction Temperature

SECTION 5:

PROJECTED R200 CURVE

The final projected curve: R200's inference efficiency, isolated from H200, expressed as % of its own peak against junction temperature.

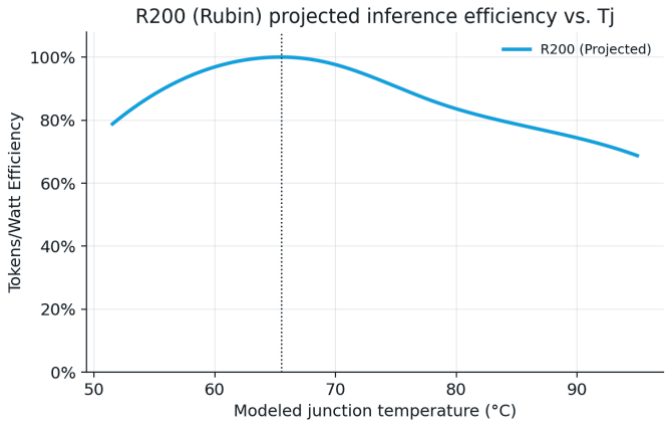


Figure 7: R200 (Rubin) projected inference efficiency vs. Tj, as % of its own peak. Both the later peak and the steeper decline follow from the smaller node's greater leakage-temperature sensitivity; the curve is normalized, so the process gain does not enter it.

SECTION 6:

CONCLUSION

The key conclusions are:

- Measured H200 inference efficiency, expressed per active watt, peaks at approximately 350 W — corresponding to $T_{j\,max}$ = 62.5°C under the calibrated thermal stack.
- Because the baseline is measured on the workload of interest and expressed per active watt, the efficiency curve requires neither a proxy conversion from a compute benchmark nor a vendor throughput ratio.
- Under the stated assumptions, Rubin is projected to reach its efficiency optimum at a higher junction temperature than Hopper, but to lose efficiency faster beyond that point — both are consequences of a single paramater, the leakage-temperature-sensitivity difference between the two nodes. Process-level efficiency gains raise the absolute efficiency level but do not move the peak or change the shape.
- The projection is robust to errors in the thermal stack. The leakage sensitivity ratio s is an input, and the threshold is located as a fraction of the span between it and $T_{j\,max}$, so an error in R_{total} shifts every absolute junction temperature together without changing the shape of either curve, the relative position of the two peaks, or λT.

Frore Systems Copyright 2026. Specification subject to change. Patents Pending.

SYMBOLS AND ABBREVIATIONS

P - GPU power cap (W)

TDP - Thermal design power, the rated maximum (700 W - H200; 2,300 W - R200)

f - Power as a fraction of TDP (P / TDP)

$\mathbf{f}_{\mathbf{thresh}}$ - Power fraction at which efficiency peaks (0.501 - H200)

$\mathbf{T}_{\mathbf{j}}$ - Junction temperature, the on-die silicon temperature (°C)

$\mathbf{T}_{\mathbf{coolant}}$ - Coolant inlet temperature (30 °C - H200; 45 °C - R200)

$\mathbf{T}_{\mathbf{j\ thresh}}$ - Junction temperature at which efficiency peaks, and beyond which leakage drives it down (62.5 °C - H200; 65.5 °C - R200)

$\mathbf{T}_{\mathbf{j\ max}}$ - Junction-temperature limit at TDP (95 °C, taken as common to both)

$\mathbf{T}_{\mathbf{j\ equiv}}$ - Point on the H200 curve used to read off a given R200 temperature (°C)

$\mathbf{R}_{\mathbf{total}}$ - Total junction-to-coolant thermal resistance (0.0929 °C/W - H200)

s - Leakage temperature-sensitivity ratio, N3P vs 4N — the model's single free parameter (1.1)

λT - Thermal scaling factor on the temperature axis; identically equal to s (1.1)

D - Leakage doubling interval (Arrhenius) (10 °C - H200; 9.1 °C - R200)

%eff - Efficiency as a percentage of that GPU's own peak (%)

TIM1, TIM2 - Thermal interface materials: die-to-lid and lid-to-coldplate (°C/W)

LPM - Coolant flow rate, litres per minute (1 LPM assumed)

REFERENCES

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/

Sections 1–2 rest on published measurements and the semiconductor physics of Section 3 is established; the thermal stack and all Rubin projections are derived from explicit assumptions documented in this note.

Frore Systems Copyright 2026. Specification subject to change. Patents Pending.