FRORE SYSTEMS

Do More.

Chip to Chiller

The Thermal Stack's Impact on AI Factory Efficiency

This page is intentionally blank.

Chip to Chiller

The Thermal Stack's Impact on AI Factory Efficiency

ABSTRACT

The world is power constrained and energy availability will limit AI Factory growth.

The AI Factory output is Tokens, the currency of intelligence. It is therefore essential that AI Factories are optimized to the system level efficiency metric of **Tokens/Watt**.

GPU Die temperature has a massive impact on Tokens/Watt.

The Thermal Stack can increase the Tokens/Watt by over 30% by maximizing thermal efficiency and lowering die temperature.

One of the key insights of semiconductor technology is that the Tokens/Watt metric is significantly impacted by the maximum junction temperature of the GPU die. **When operated at the same power, GPUs produce more Tokens when they run at cooler temperatures.** The infrastructure that determines the GPU die temperature in the AI Factory is the Thermal Stack.

The Thermal Stack comprises of the entire thermal chain for heat extraction from the GPU (GPU package architecture, thermal interface material (TIM), coldplate, coolant inlet temperature, and flow rate) and heat rejection into the atmosphere (free evaporative cooling towers and mechanical chillers).

In this paper, we analyze the profound impact the Thermal Stack has on GPU die max junction temperature and thereby the AI Factory efficiency metric of Tokens/Watt. A well-engineered Thermal Stack can boost AI Factory Tokens/Watt metric by over 30%.

AI Factories are optimized to Tokens/Watt metric

AI FACTORIES ARE POWER HUNGRY AND POWER CONSTRAINED

The binding operational constraint in virtually every AI Factory is not compute, memory bandwidth, nor capital investment — **it is power**.

Power available to AI Factories is finite and expensive to expand. Utility interconnection, transformer capacity, and physical power delivery infrastructure are all limiting factors on total facility power draw. In the current environment, grid capacity is increasingly scarce in major data center markets, and regulators in several jurisdictions are imposing consumption caps. Behind-The-Meter (BTM) power generation helps, but the huge gap between power demand and availability remains. Facing this power constraint, AI Factory operators understand that power is a strategic asset and every Watt allocated must be used to generate Tokens.

Every percentage point improvement in Tokens/Watt translates directly into a percentage point increase in profitability. This is the economic logic that makes **Tokens/Watt the central operational metric for AI Factory efficiency**.

A well-engineered Thermal Stack can boost AI Factory Tokens/Watt metric by over 30%.

ENERGY COST IS THE LARGEST COMPONENT OF AI FACTORY OPEX

Understanding the full composition of data center operating expenditure is essential for evaluating the true cost of Tokens delivered.

The AI Factory OPEX components include energy costs, facility maintenance, staffing and operations, networking and connectivity costs, hardware maintenance, software licensing, and facility lease.

Chip to Chiller

The Thermal Stack's Impact on AI Factory Efficiency

Energy cost is the largest and most variable factor, and encompasses GPU, CPU, memory, networking, storage, power delivery, and thermal stack power consumption. As the GPU is the primary consumer of the AI Factory power budget, the industry historically focused its investment and efforts on GPU design improvements. Moving from one process node to the next and transitioning toward increasingly inference-optimized designs has significantly improved GPU performance at equivalent power, which flows directly into improved Tokens/Watt at the system level.

However, the GPU efficiency for a given generation is set once its design architecture, process node, and floating-point precision support are fixed. Performance at the GPU level cannot improve until the next generation. This is where the Thermal Stack becomes the dominant remaining lever to pull for system level Tokens/Watt improvement.

GPU DIE TEMPERATURE HAS MASSIVE IMPACT ON TOKENS/WATT

Max junction temperature ($T_{j\,max}$) is the maximum temperature of the active transistor layer of the GPU die.

This is distinct from the temperature measured at the GPU packaging case surface, the coldplate, or the coolant.

It is the hottest point in the Thermal Stack since heat is generated by power flowing through the GPU die and all the power consumed by the GPUs converts into heat.

This heat must be extracted outward from the GPU through several layers — GPU package, TIM, coldplate, coolant — and ultimately rejected by the facility cooling system. Every GPU has a maximum rated junction temperature, typically in the 95°C range for modern GPUs.

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.

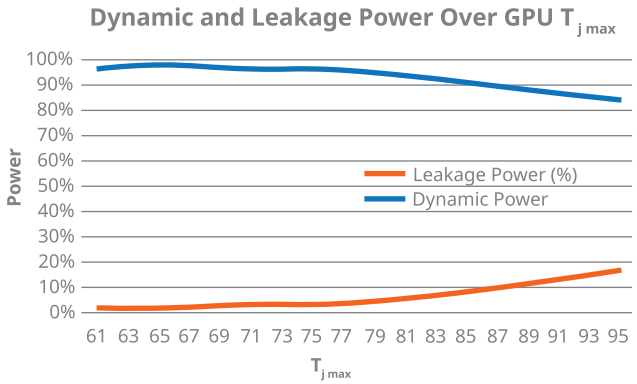


Figure 1: As max junction temperature increases, leakage power increases exponentially

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 maximum allowed.

Operating consistently near maximum junction temperature also accelerates electromigration and oxide degradation, shortening chip lifespan and increasing the risk of early failure in a fleet expected to operate continuously for years. The overall relationship between GPU max junction temperature ($T_{j\,max}$) and power efficiency is shown on the next page.

Chip to Chiller

The Thermal Stack's Impact on AI Factory Efficiency

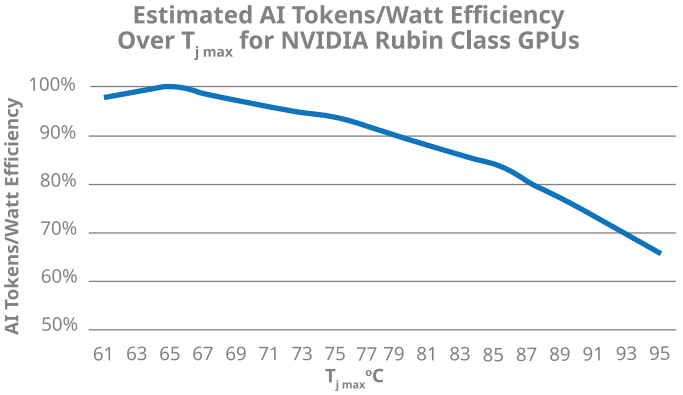
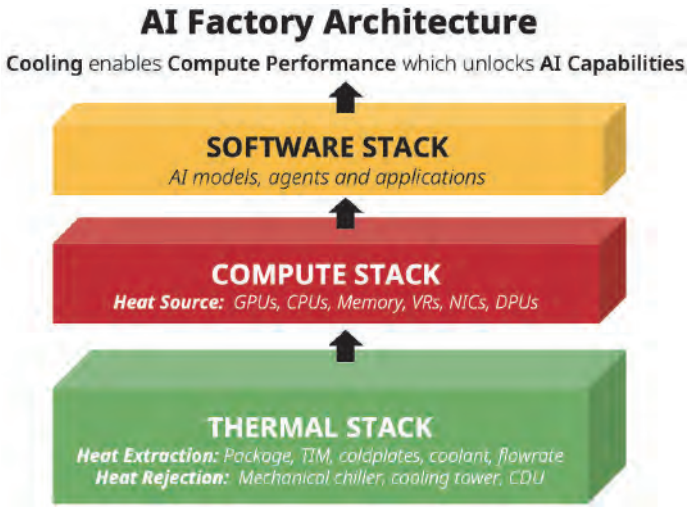


Figure 2: Tokens/Watt efficiency decreases as $T_{j\,max}$ increases

A 10°C decrease in GPU max junction temperature can result in a 15% increase in Tokens/Watt

AI FACTORY ARCHITECTURE

An AI Factory is a purpose-built system designed to produce one output: Tokens. Its architecture is made of three interdependent layers.



- **Software Stack** AI foundational models, Agents and Applications.
- **Compute Stack** consisting of every GPU, CPU, Memory and Network Switching chips plus Power delivery. Compute Stack generates heat since all the power it consumes converts into heat.
- **Thermal Stack** extracts the heat from the Compute stack and rejects it into the atmosphere. Heat extraction chain consists of chip packaging architecture, thermal interface materials, coldplates, coolant, and flow rate. Heat rejection consists of CDUs, free cooling towers and mechanical chillers.

The Thermal Stack is the foundational layer that makes everything possible. Without effective heat extraction and rejection, Compute Stack throttles and the Software Stack stalls. Thermal Stack efficiency determines how much useful computation the GPU delivers and the system level Tokens/Watt metric realized.

DEFINING THE THERMAL STACK

Heat generated by the Compute Stack must first be extracted before it is rejected into the atmosphere.

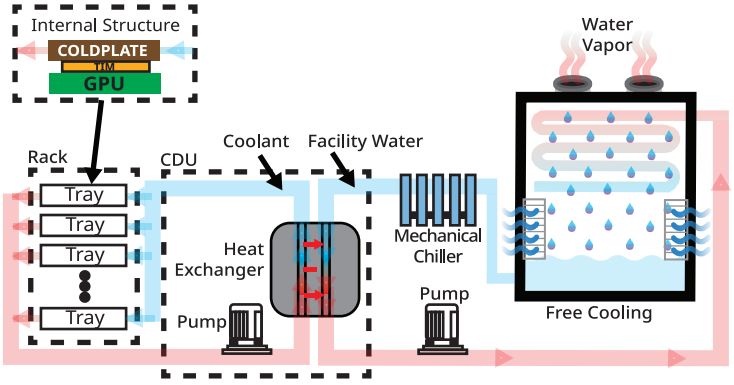


Figure 3: Thermal Stack Components

Direct-to-Chip liquid cooling (DLC) has become the industry standard for the latest generation of AI Factories.
This approach extracts heat from the GPUs using liquid coolant that circulates through a coldplate which is in direct thermal contact with the GPU.

Chip to Chiller

The Thermal Stack's Impact on AI Factory Efficiency

Cool liquid coolant (T_{inlet}) is pumped through microchannels in the coldplate to extract heat from the GPU. Warm coolant (T_{outlet}) exits the coldplate.

The coolant is recirculated through the Coolant Distribution Unit (CDU) which houses a heat exchanger where coolant at T_{outlet} expels heat into the facility water loop such that the coolant can return to T_{inlet} before it is recirculated back into the coldplates.

The heat that has been transferred to the facility water loop is either expelled into the outside atmosphere through evaporation in a cooling tower (free cooling), removed using a mechanical chiller, or both, depending on the local climate and required T_{inlet}. The coolant and facility water loops remain hydraulically isolated from each other within the CDU. This allows the high-purity coolant to be chemically controlled independently of the facility water supply.

GPU die max junction temperature ($T_{j\,max}$) which is the binding constraint on GPU performance efficiency as described in the previous section, is governed by the below equation:

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

Q is the GPU power consumed, T_{inlet} is the coolant temperature as it enters the coldplate, R_{total} is the cumulative thermal resistance between the GPU die and the coolant. R_{total} is determined by GPU package architecture, thermal interface material (TIM) selection, and coldplate design.

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

T_{inlet} is the sum of free cooling (T_{free}) which is determined by the wet bulb temperature — a function of atmospheric ambient temperature and relative humidity plus CDU heat exchanger efficiency — and mechanical chiller cooling ($T_{mechanical}$).

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

Mechanical chiller usage is power intensive. It adds to the facility-level power cost of holding T_{inlet} at a lower temperature. However, the better the mechanical chiller coefficient of performance (COP), the lower the additional facility-level power cost.

Together R_{total} and T_{inlet} connect the physical design of the Thermal Stack to GPU die max junction temperature ($T_{j\,max}$) and thus to Tokens/Watt.

GPU PACKAGE ARCHITECTURE

$R_{package}$ varies significantly based on whether the GPU package is lidded or un-lidded.

In a lidded package, an integrated heat spreader (Lid) sits over the die complex, adding mechanical protection and lateral heat spreading. A thin layer of graphene TIM is inserted in between the Lid and the Logic Die. Therefore, lidded packages have significantly higher thermal resistance than un-lidded packages.

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

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

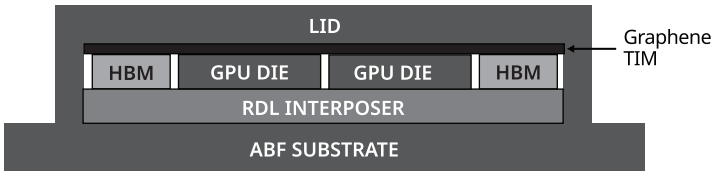


Figure 4: Lidded package design

In an un-lidded package, there is neither a lid nor a graphene TIM. Therefore un-lidded designs have lower R_{total} and achieve materially lower $T_{j\,max}$ at equivalent Q and T_{inlet}.

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

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

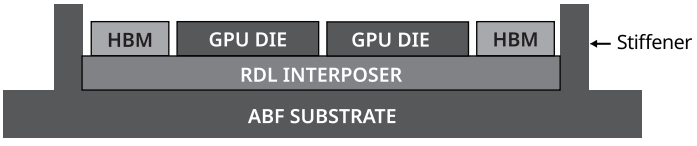


Figure 5: Un-lidded package design

For NVIDIA Rubin class GPUs, the difference in $T_{j\,max}$ between lidded and un-lidded packages can be as much as 20°C, resulting in up to a 35% improvement in Tokens/Watt efficiency.

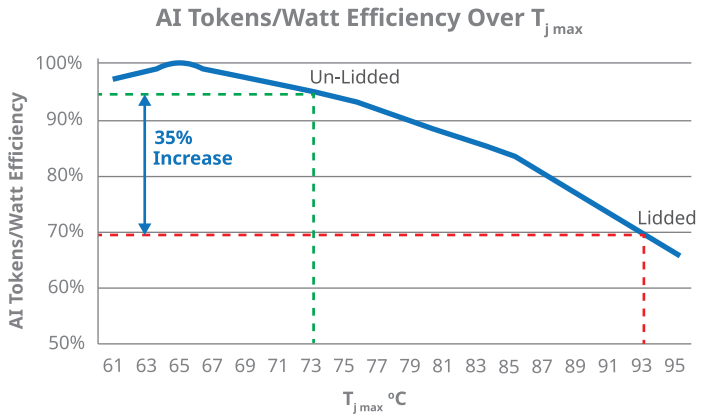


Figure 6: Due to lower thermal resistance, un-lidded packages can achieve higher Tokens/Watt efficiency at the same GPU power by enabling a lower max junction temperature

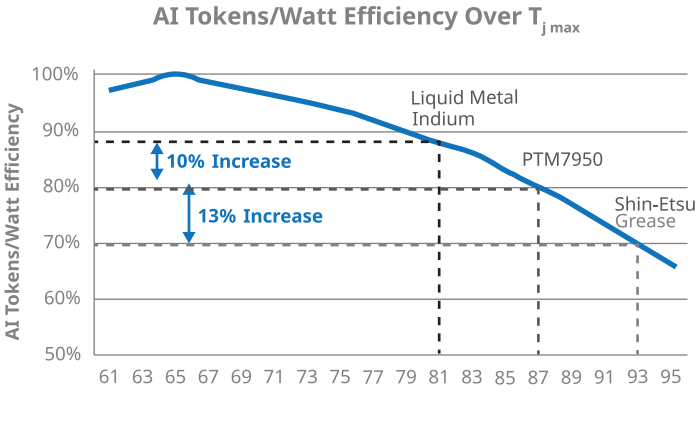


Figure 7: Due to lower thermal resistance, different TIMs can achieve higher efficiency at the same GPU power by enabling a lower max junction temperature

However, there could be a reliability trade-off. In un-lidded packages, the bare dies are directly exposed, so the risk of chipping or cracking during handling and coldplate installation are higher than with lidded packages. Additionally, the lid sets a precise, uniform surface. Without it, the coldplate may encounter non-uniform z-height across the die surface.

THERMAL INTERFACE MATERIALS

A thermal interface material (TIM) is used between the GPU package and the coldplate to ensure excellent thermal contact, but the TIM adds to the thermal resistance. Three TIM material classes are commonly used:

TIM Type	Typical Thermal Resistance	Notes
Shin-Etsu Grease	0.09°C cm²/W	Lowest cost, high resistance, easiest to rework
Honeywell PTM7950	0.06°C cm²/W	Acceptable cost, low resistance, high stability
Liquid Metal Indium	0.03°C cm²/W	High cost, lowest resistance, but requires containment design to prevent migration/corrosion over device lifetime

Liquid Metal Indium can be effectively used with lidded packages by electroplating the top of the Lid and the bottom of the coldplate with gold. Thus, the lidded package thermal resistance disadvantage can be somewhat neutralized by using Liquid Metal Indium TIM. Indeed, this is the strategy adopted by NVIDIA Rubin GPU.

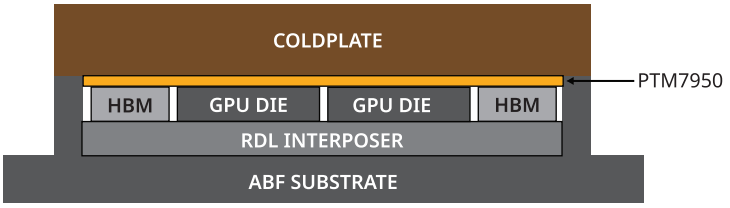


Figure 8: Un-Lidded package design with PTM7950 TIM

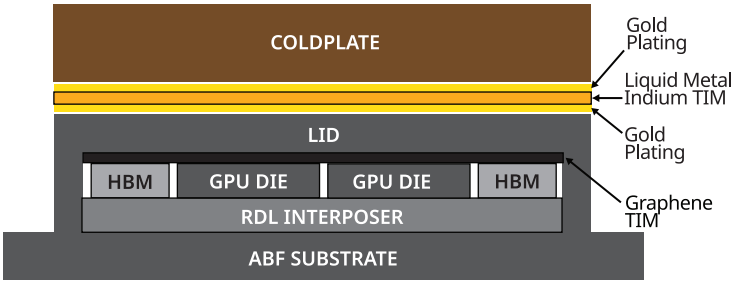


Figure 9: Lidded NVIDIA Rubin package with Liquid Metal Indium TIM

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

The difference in $T_{j\ max}$ for the two configurations is as much as 14°C resulting in up to a 28% delta in Tokens/Watt efficiency.

Chip to Chiller

The Thermal Stack's Impact on AI Factory Efficiency

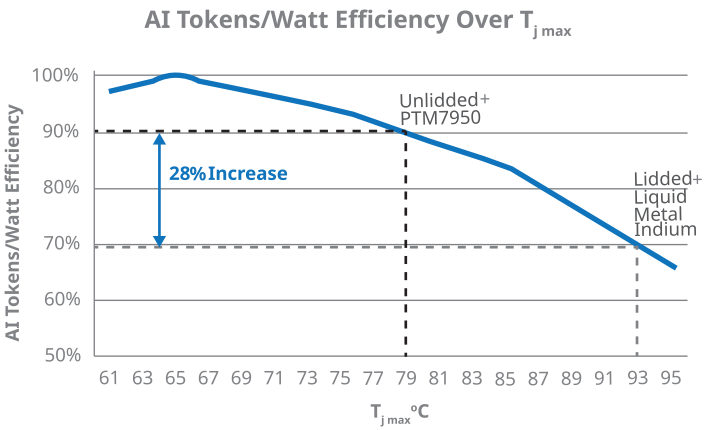


Figure 10: Un-lidded GPU package with PTM7950 TIM delivers higher efficiency at the same GPU power than lidded GPU package with Liquid Metal Indium TIM

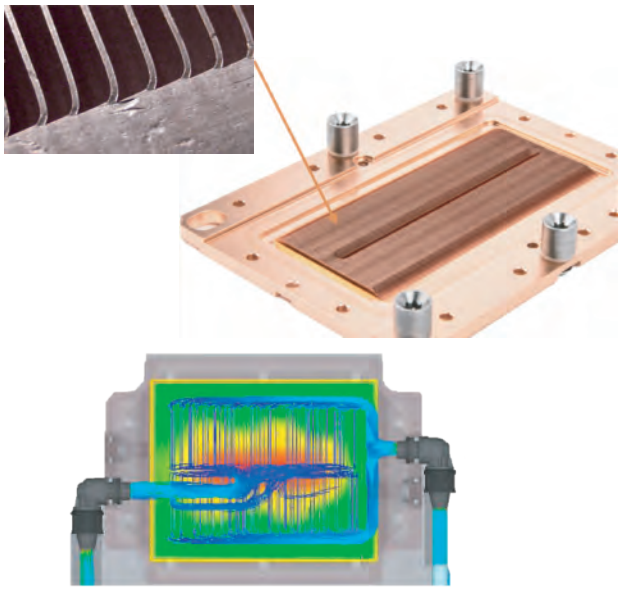


Figure 11: Skived Microchannels in Split-flow coldplate

COLDPLATE DESIGN

Coldplate design has a major role to play in reducing the thermal resistance R_{Total}. The coldplate thermal resistance, $R_{coldplate}$, is determined by three interacting design variables: microchannel geometry, flow routing, and flow rate.

Microchannel geometry

Microchannel length, fin pitch, fin height, and aspect ratio determine the convective heat transfer coefficient between the coldplate wall and the coolant flowing through the coldplate.

Narrower channels and higher aspect ratios increase surface area per unit volume and improve heat transfer, whereas long microchannel length reduces heat transfer efficiency because of *developed* coolant flow.

Therefore, designing ultra short length channels < 1mm (short-loop), significantly improves heat transfer efficiency by keeping the coolant flow in *developing* flow phase.

To construct short-loop microchannels, 3D microcells are required. But 3D microcells are impossible to manufacture using traditional machining manufacturing methods like skiving.

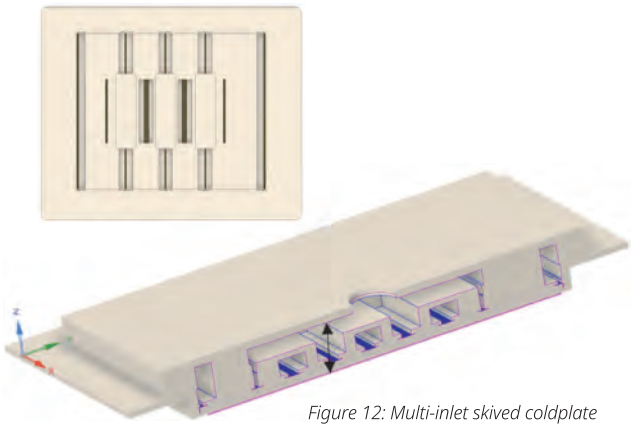


Figure 12: Multi-inlet skived coldplate

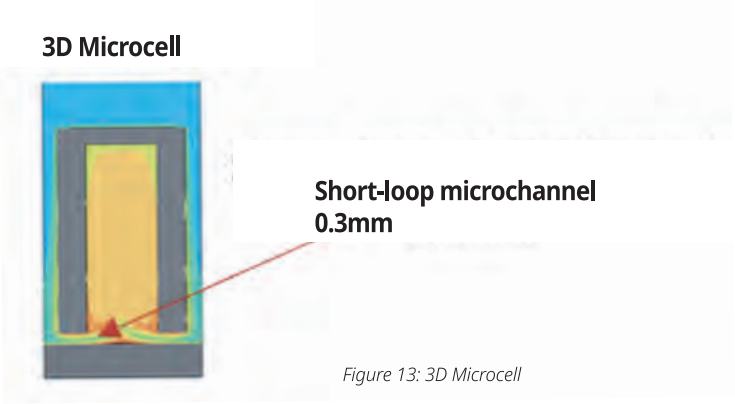


Figure 13: 3D Microcell

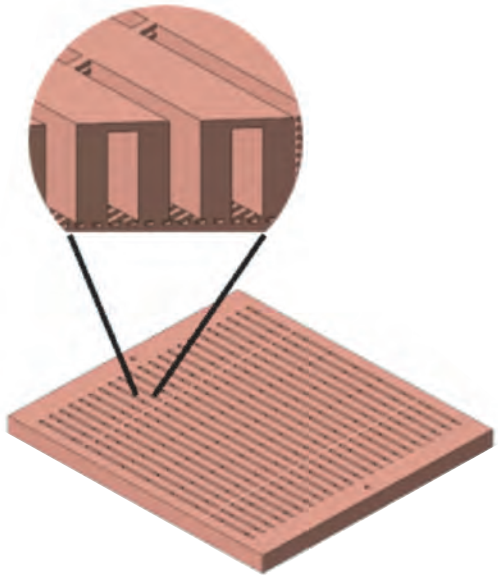


Figure 14: Coldplate with 3D Microcells

3D microcells, short-loops and multi-staging, together result in a far more efficient coldplate design

Flow routing

The coolant entering the coldplate is distributed across the microchannels. At its simplest, the flow is uniformly distributed across all the microchannels through a single inlet at one end and is collected at the far end. But this can produce localized hot spots since high-power die regions receive the same flow as the rest of the die. Even worse, the high power die regions could be located downstream of microchannel *developed* coolant flows that have already absorbed heat.

A somewhat better flow routing technique is the 'split-flow' configuration. In this design, flow is evenly split across two inlets on either end of the coldplate and warm coolant is collected at the outlet in the middle of the coldplate.

This cuts effective microchannel length by half, thus improving heat transfer efficiency. A simple manifold covers the microchannels to facilitate this flow routing. A logical extension of this approach is the multi-inlet coldplate. Flow is distributed in a weighted fashion across the multi-inlets using pressure balancing to favor higher power density regions of the GPU die.

Multi-stage

But in all these three coolant flow routing options, flow is only used once and not recycled. When incoming coolant is distributed across the microchannels, it flows linearly only once through the microchannels and is collected at the outlet.

However, in a multi-stage coldplate the coolant is recycled multiple times. Multi-stage coldplates partition the coldplate into multiple sections, and each section receives the entire coolant flow one after the other. For example, in a 3-stage coldplate, the entire coolant flow first enters stage 1, after exiting stage 1, the entire coolant flow sequentially enters stage 2, then stage 3. The stages are designed based on the GPU die power map with the early stages mapping to the high power die regions.

One should note multi-staging adds to the coldplate pressure drop, whereas short-loops reduce pressure drop. Therefore, using both these features together results in a far more efficient coldplate design with significantly lower $R_{coldplate}$ value without increasing coldplate pressure drop.

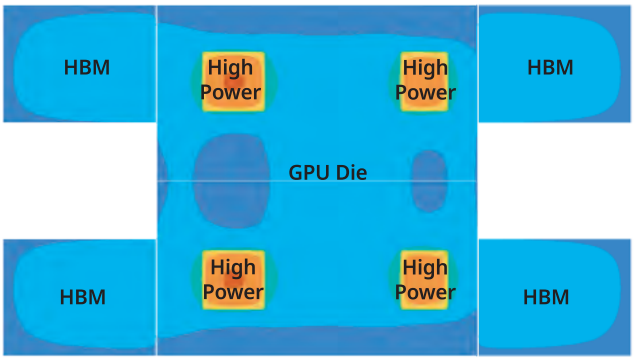


Figure 15: GPU power density map with four high power regions

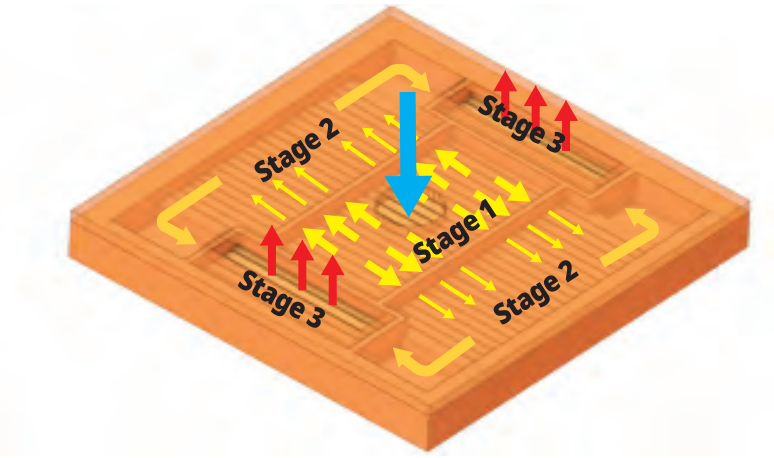


Figure 16:Coldplate with multi-stage design

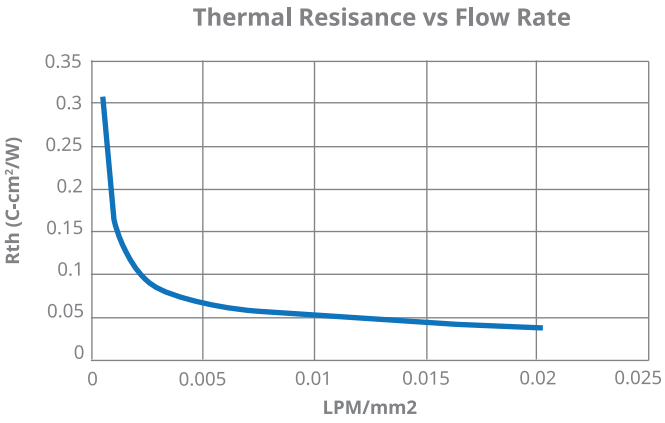


Figure 17: Coldplate thermal resistance vs flow rate for 3D Microcells

Coolant

In direct-to-chip liquid cooling systems, two types of coolants are used. De-ionized water, which has a heat capacity of 4.18 Joules/ml/°C and PG25 (25% propylene glycol, 75% water) which as a heat capacity of 3.85 Joules/ml/°C. To raise the temperature of 1 LPM of coolant by 1°C, 69.67 W and 64.17 W are required respectively.

Flow rate

Flow rate is the remaining operational degree of freedom once coldplate design is fixed. Increasing flow rate reduces $R_{coldplate}$. But the relationship is not linear: $R_{coldplate}$ falls steeply with flow rate at low flow and flattens as it approaches its geometry-limited ceiling, with each additional unit of flow buying progressively less thermal benefit.

Higher flow rates also increase pressure drop and pump power, following a cubic relationship between pump power and flow rate (the pump affinity laws).

In practice, this power penalty is small relative to GPU and mechanical chiller power draw, so running flow rate up to the point where $R_{coldplate}$ flattens is almost always worth the cost — but the operating point is typically set by facility CDU capacity and manifold, hoses and quick disconnect specifications for leak avoidance.

MANUFACTURING APPROACH

The manufacturing approach shapes which coldplate designs are achievable.

The traditional machining skiving manufacturing technique is universally used today for high volume coldplate production. In skiving manufacturing, fins are shaved directly from a solid copper block, producing continuous, long, high-aspect-ratio fin structures. Skiving combined with manifolding is used to produce simple microchannel coldplates, split-flow coldplates, and multi-inlet coldplates. But advanced techniques like short-loop microchannels and multi-staging are out of reach for skiving. New manufacturing techniques are required.

Frore Systems has developed a unique manufacturing approach by applying semiconductor manufacturing processes to copper wafers to realize their complex LiquidJet® Coldplate designs. LiquidJet® Coldplate is designed with narrow contoured jetchannels, short-loops enabled by 3D microcells, multi-staging and a host of other novel techniques that target high power die regions and reduce $R_{coldplate}$.

Frore Systems first generates CAD design for a custom LiquidJet® Coldplate optimized for the GPU power map. Frore Systems then uses semiconductor manufacturing processes like maskset generation, photolithography, wet etching, and monolithic bonding to manufacture the LiquidJet® Coldplate cost effectively at scale with proven reliability.

Chip to Chiller

The Thermal Stack's Impact on AI Factory Efficiency

LiquidJet® Coldplates reduce $T_{j\,max}$ by 6 – 12°C, thus significantly improving Tokens/Watt by 10-25%.

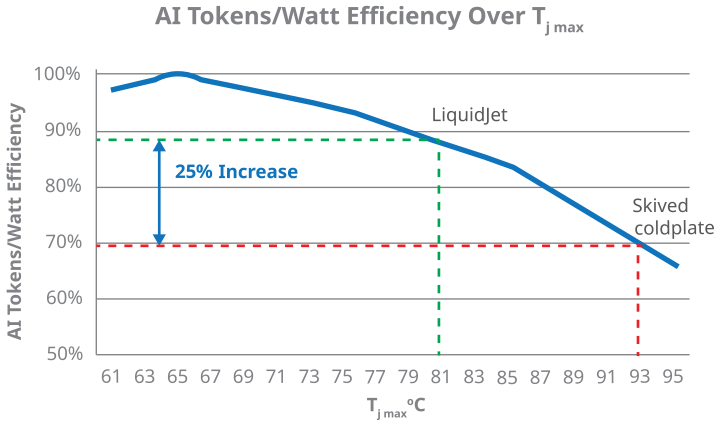


Figure 18: Due to lower thermal resistance, LiquidJet can achieve higher efficiency at the same GPU power by enabling a lower max junction temperature

Semiconductor manufacturing processes applied to copper wafers enable complex LiquidJet® Coldplate designs.

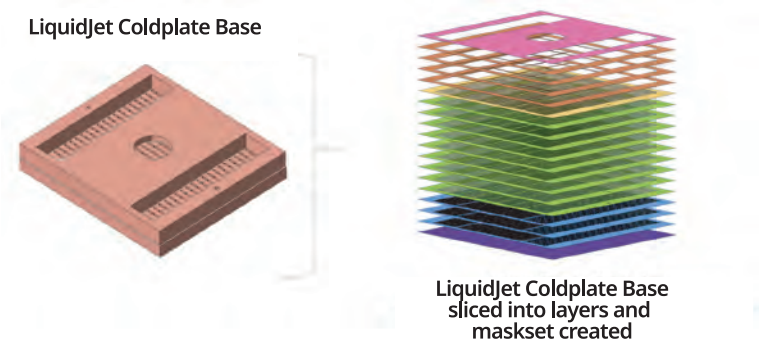


Figure 19: LiquidJet® CAD design → Maskset generation

Other companies are attempting to apply additive manufacturing techniques like metal 3D printing to build similar short-loop coldplates, but the versatility, reliability, and scalability of these approaches is unproven.

COOLANT INLET TEMPERATURE

Coolant inlet temperature (T_{inlet}) has a direct 1:1 relationship with max junction temperature ($T_{j\,max}$). Holding GPU power (Q) and R_{total} constant, every 1°C reduction in T_{inlet} produces exactly a 1°C reduction in $T_{j\,max}$. And every 1°C reduction in $T_{j\,max}$ improves the Tokens/Watt efficiency of the GPU.

NVIDIA has specified operation at a 45°C inlet temperature for Rubin GPU based AI Factories. The significance of that specification is facility-level power consumption: at 45°C, cool water can be produced by free cooling alone across most global data center climates, removing the need for mechanical chillers entirely.

That said, the NVIDIA Rubin GPU platform's ability to run at 45°C inlet does not mean every deployment should.

While a higher T_{inlet} may suffice for some workloads, for GPUs deployed in AI Factories, given the sensitivity of Tokens/Watt metric to GPU die max junction temperature, a lower T_{inlet} is highly recommended. But taking T_{inlet} below T_{free} requires using mechanical chillers which are power intensive and could negatively impact the overall Tokens/Watt power efficiency metric of the AI Factory. So the trade-off is not worthwhile unless the chiller COP is sufficiently high.

MECHANICAL CHILLER IMPACT ON TOKENS/WATT

When T_{inlet} is held below T_{free}, a mechanical chiller is used to close the gap. Mechanical chiller efficiency is characterized by its Coefficient of Performance (COP):

$$COP = Q/W$$

Where Q is the heat removed and W is the electrical power consumed by the mechanical chiller. A COP of 5 means the chiller removes 5 Kilowatts of heat for every 1 Kilowatt of electrical power it consumes. Let's assume a mechanical chiller is used to reduce T_{inlet} by 1°C. Then that 1°C lower inlet temperature directly translates into 1°C lower $T_{j\,max}$, which in turn can result in a 1.5% improvement in Tokens/Watt (89°C to 88°C $T_{j\,max}$ decrease).

Chip to Chiller

The Thermal Stack's Impact on AI Factory Efficiency

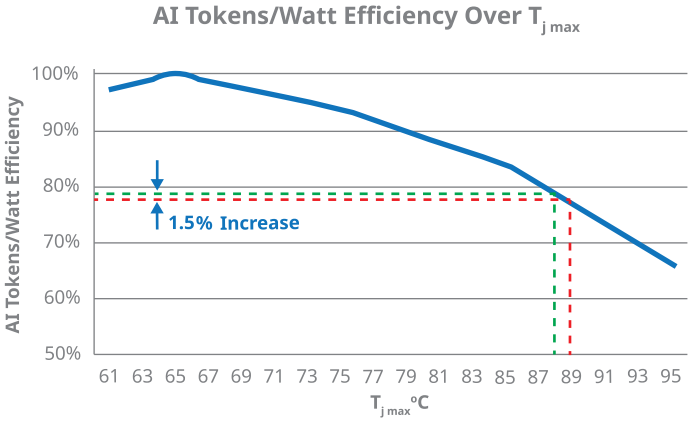


Figure 20: AI Tokens/Watt efficiency 1.5% increase for every 1°C drop in max junction temperature

However, this T_{inlet} temperature reduction is not free, as mechanical chillers consume power. When the coolant is de-ionized water:

Chiller Power (W) for 1°C reduction = (LPM * 69.67/COP)

So, if the total power budget *x* is held constant, while some power is diverted from the GPU to the mechanical chiller, that much less power budget remains for the GPU. On the plus side, 1°C decrease in $T_{j\,max}$ yields a *y*% (1.5% from 89°C to 88°C $T_{j\,max}$ decrease) increase in Tokens/Watt. Therefore, Tokens/Watt could increase if the mechanical chiller COP is high enough. COP at which GPU Tokens/Watt remains the same when power is diverted to the mechanical chiller is $COP_{breakeven}$.

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

In a power constrained AI Factory diverting power away from the GPUs to a mechanical chiller only makes sense when the COP is higher than $COP_{breakeven}$ and the lower the flow rate required by the coldplate for a given GPU power, the lower the $COP_{breakeven}$.

NVIDIA Rubin GPU Max-P power is 2300W and estimated flow rate when using a standard skived coldplate is 3.25 LPM, resulting in a $COP_{breakeven}$ of 6.7. On the other hand, with LiquidJet 3D microcell short-loop multi-stage coldplate, only 2 LPM is required to maintain the same GPU die max junction temperature, resulting in a $COP_{breakeven}$ of only 4.1.

Thus, a more thermally efficient coldplate design like LiquidJet, which lowers the flow rate required, renders mechanical chillers more useful in improving the AI Factory Tokens/Watt metric.

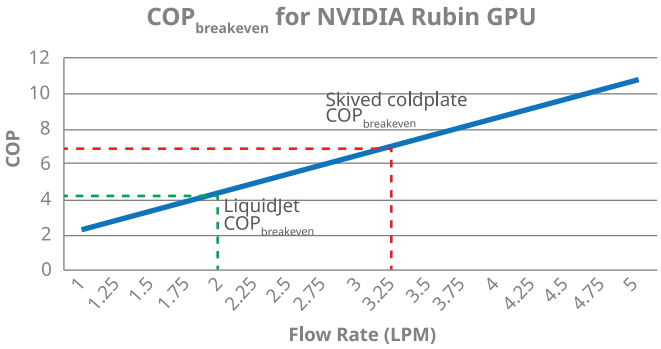


Figure 21: COP_breakeven is lower for LiquidJet as its superior design reduces required flow

CONCLUSION

AI Factories are power-constrained systems where the governing efficiency metric — Tokens/Watt — determines the profitability of every deployment.

Once a GPU is manufactured and deployed, its silicon-level efficiency is fixed. The only remaining lever to meaningfully improve Tokens/Watt is the Thermal Stack configuration.

Hyperscalers who treat the Thermal Stack as a first-class design discipline will extract materially more Tokens/Watt from the same investment.

Chip to Chiller

The Thermal Stack's Impact on AI Factory Efficiency

LiquidJet Coldplate alone can reduce GPU die max junction temperature by 6 – 12°C, improving Tokens/Watt metric by 10 – 25%.

This paper has demonstrated that GPU max junction temperature ($T_{j\,max}$) is not a passive outcome of chip design, but an actively engineered variable. There is a direct relationship between $T_{j\,max}$ and Tokens/Watt, and it is controlled by the cumulative configuration decisions made across several interdependent layers of the Thermal Stack: GPU package architecture, thermal interface material (TIM) selection, coldplate design, coolant inlet temperature, coolant flow rate and mechanical chiller deployment.

Each Thermal Stack layer contributes thermal resistance or adds to the facility power budget. Each presents a distinct optimization opportunity.

GPU package architecture introduces significant thermal resistance into the Thermal Stack. Lidded designs have a graphene TIM and a Lid and add $R_{graphene}$ and R_{Lid} to the thermal resistance, while un-lidded designs eliminate both, but at the cost of higher mechanical risk during handling and installation.

Thermal interface material selection determines how effectively heat transfers from the package to the coldplate with Liquid Metal Indium delivering the lowest resistance. Un-lidded package and the right TIM can improve Tokens/Watt by as much as 25%.

Frore Systems' LiquidJet coldplate design can dramatically lower $R_{coldplate}$ by using semiconductor manufacturing techniques enabling 3D microcell short-loops, and multi-stage designs, that traditional skiving manufacturing techniques cannot achieve.

LiquidJet Coldplate alone can reduce GPU max junction temperature by 6 – 12°C, improving Tokens/Watt metric by 10 – 25%.

The decision on where to set T_{inlet} — and whether to engage mechanical chillers — is ultimately a system-level power allocation problem governed by mechanical chiller COP, flow rate, and the marginal Tokens/Watt improvement of each degree of $T_{j\,max}$ reduction.

The AI infrastructure landscape is converging on this reality: hyperscalers who treat the Thermal Stack as a first-class design discipline will extract materially more Tokens/Watt from the same CAPEX and OPEX investment.

As GPU power densities continue their upward trajectory beyond NVIDIA Rubin toward future architectures, the performance gap between thermally optimized and thermally unoptimized deployments will only widen.

The Thermal Stack is not support infrastructure. It is a performance system and, in a power constrained world, it is essential for optimizing AI Factory Tokens/Watt output.

A well-engineered Thermal Stack can boost AI Factory Tokens/Watt metric by over 30%.

The Thermal Stack is not support infrastructure. It is a performance system and, in a power constrained world, it is essential for optimizing AI factory Tokens/Watt output.

Chip to Chiller

The Thermal Stack's Impact on AI Factory Efficiency

GLOSSARY

AI Factory — A large-scale data center built specifically to run artificial intelligence workloads, where the primary output is AI-generated results (Tokens) rather than traditional computing tasks.

Arrhenius Equation — A mathematical formula describing how the rate of a thermally-activated process increases exponentially with temperature. In the context of GPU power, it explains why leakage current rises sharply as max junction temperature increases — the physical basis for leakage power roughly doubling with every 10°C rise in $T_{j\,max}$. Named after Swedish chemist Svante Arrhenius.

Behind-the-Meter (BTM) Power Generation — On-site power generation located at the data center facility itself, such as solar panels, backup generators, or fuel cells. "Behind the meter" means the power is generated and consumed on-site before it ever passes through the utility grid meter — reducing dependence on grid capacity and helping offset the gap between AI Factory power demand and utility supply.

COP (Coefficient of Performance) — A measure of mechanical chiller efficiency. A COP of 5 means the chiller removes 5 kilowatts of heat for every 1 kilowatt of electricity it consumes. Higher is better.

$COP_{breakeven}$ — The minimum mechanical chiller efficiency above which diverting power from GPUs to the chiller still results in a net improvement in Tokens/Watt. Below this threshold, running the chiller hurts overall efficiency.

DVFS (Dynamic Voltage and Frequency Scaling) — An automatic mechanism inside GPUs that adjusts voltage and clock speed with changes in workload and max junction temperature.

Developed vs. Developing Coolant Flow — When coolant first enters a microchannel it is in a "developing flow" state, meaning it is actively mixing and exchanging heat with the channel walls at maximum efficiency. As it travels further along the channel it reaches "developed flow" — a stable, uniform state where heat transfer efficiency drops significantly. Short-loop microchannel designs keep coolant in the developing flow phase by limiting channel length. Shorter channels transfer heat more efficiently than longer ones.

Electromigration — A gradual degradation process in chip circuitry caused by sustained high temperatures, which shortens GPU lifespan.

Free Cooling — Using evaporative cooling towers to cool facility water without mechanical refrigeration. Lower cost and lower power than mechanical chillers.

$T_{j\,max}$ (Max Junction Temperature) — The temperature at the active transistor layer inside the GPU die. The governing variable for both performance and lifespan. Everything in the Thermal Stack is ultimately trying to control this number.

Leakage Power — Wasted electrical power that flows through transistors continuously regardless of whether computation is occurring. Increases exponentially with temperature.

LiquidJet® — Frore Systems' proprietary coldplate technology, manufactured using semiconductor processes to achieve coldplate designs that traditional manufacturing, like skiving, cannot replicate.

Mechanical Chiller — A powered refrigeration system used to cool facility water below the temperature free cooling alone delivers.

Pump Affinity Laws — A set of engineering principles that describe how pump power scales. The critical insight for the AI Factory is that pump power increases with the cube of flow rate — meaning doubling the coolant flow rate requires eight times the pump power.

Q — GPU power consumption, measured in Watts. The total heat load the Thermal Stack must remove.

R_{die} — The thermal resistance of the GPU die silicon wafer.

$R_{package}$ — The thermal resistance contributed by the GPU package architecture — whether lidded or un-lidded, and what materials sit between the GPU die and TIM.

$R_{coldplate}$ — The thermal resistance contributed by the coldplate itself, determined by microchannel geometry, flow routing, and flow rate.

$R_{graphene\ TIM}$ — Thermal resistance added by the graphene TIM between the lid and the die in a lidded package.

R_{lid} — Thermal resistance added by the integrated heat spreader (lid) in a lidded GPU package.

R_{total} — Total thermal resistance of the Thermal Stack between the GPU transistor layers and the coolant. Lower R_{total} means lower $T_{j\,max}$ for the same Q and T_{inlet}.

R_{TIM} — Thermal resistance contributed by the thermal interface material between the GPU package and the coldplate.

TIM (Thermal Interface Material) — A substance applied between the GPU package and the coldplate to improve thermal contact.

R (Thermal Resistance) — A measure of how much a material or component impedes the flow of heat. Lower resistance means heat moves more efficiently and max junction temperature stays lower.

Thermal Stack — The complete chain of components and systems responsible for moving heat away from the GPU and out of the facility — from the chip transistor layer all the way to the cooling tower.

T_{free} — The lowest T_{inlet} coolant temperature achievable using free cooling alone, determined by ambient temperature, relative humidity, and CDU heat exchanger efficiency. No mechanical chiller power required to reach this point.

T_{inlet} — The temperature of liquid coolant entering the coldplate. Every 1°C reduction in T_{inlet} produces exactly 1°C reduction in $T_{j\,max}$.

$T_{mechanical}$ — The additional cooling contribution from mechanical chillers, used when T_{free} is not low enough to hit the target T_{inlet}.

Token — The fundamental unit of AI output. Every word, image, or result generated by an AI model consists of Tokens.

Tokens/Watt — The key efficiency metric for AI Factories. Measures how many units of AI output (Tokens) are produced for every Watt of power consumed. Higher is better.

T_{outlet} — The temperature of liquid coolant as it exits the coldplate, after absorbing heat from the GPU.

Wet Bulb Temperature — A measure of how much cooling is achievable through evaporation in a given climate. It accounts for both air ambient temperature and relative humidity. The lower the wet bulb temperature, the more effective free cooling can be — and the lower the T_{inlet} achievable without mechanical chillers. It is the key climate variable that determines how much free cooling is available at any given data center location.

Do More.