Proceedings of the ASME 2025 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems
InterPACK2025
October 28-30, 2025, Anaheim, CA, USA

INTERPACK2025-164278

DIRECT ON DIE TWO PHASE COOLING APPROACH FOR HIGH POWER GPUS

Akshith Narayanan
Accelsius
Austin, Texas

Qingyang Wang
Accelsius
Austin, Texas

Serdar Ozguc
Accelsius
Austin, Texas

Trevor Whitaker
Accelsius
Austin, Texas

Jacob Moore
Accelsius
Austin, Texas

Richard W. Bonner III
Accelsius
Austin, Texas

ABSTRACT

As the artificial intelligence boom accelerates, individual servers and racks have become exponentially more power-hungry. Over the past three years, there has been a 400% increase in the thermal design power (TDP) of both GPUs and CPUs. This surge in power demand has necessitated a transition to liquid cooling, as air cooling has reached its practical limits. Among liquid cooling solutions, two-phase direct-to-chip cooling has emerged as a promising approach to managing escalating TDPs and heat fluxes due to its high heat transfer coefficients and the advantageous properties of the boiling phenomenon. Additionally, the use of a dielectric working fluid provides a safeguard against potential damage to IT equipment in the event of a leak—an inherent risk in single-phase cooling solutions such as direct-to-chip water cooling.

In the industry, this cooling method is conventionally referred to as direct-to-chip cooling, even though the coolant does not come into direct contact with the chip surface; instead, a thermal interface material (TIM) and a cold plate are used. This study explores the effectiveness of a two-phase cooling approach in which the refrigerant directly impinges on the die, eliminating the need for both a TIM and a cold plate, thereby reducing the case-to-fluid thermal resistance. The working fluid used in this study was the medium-pressure refrigerant R515b. Testing was conducted on a thermal test vehicle (TTV) designed to mimic the die area of the NVIDIA Blackwell (B200) GPU, with a TDP of up to 3 kW, subjecting the system to heat fluxes exceeding 150 W/cm².

Experiments were performed on an enhanced surface featuring a skived fin base to simulate a heat sink that could be bonded directly to the chip surface. Various fluid manifold designs were evaluated to assess how different flow paths influence heat transfer performance. The results demonstrated that the direct-to-die approach significantly reduced thermal resistance (case to fluid). Additionally, this improvement in thermal resistance allows the CDU to operate effectively at higher facility water temperatures of 60-70°C. The enhanced performance and potential for heat reuse at elevated temperatures highlight the feasibility of integrating heat sinks directly onto the silicon chip surface, enabling a direct-to-die cooling strategy.

This study underscores the potential of two-phase cooling, particularly when using a dielectric fluid, as it enables direct impingement on the die without the risk of short-circuiting IT equipment. These findings suggest a promising pathway for GPU manufacturers to adopt this advanced cooling approach, facilitating more efficient thermal management in high-performance computing environments.

NOMENCLATURE

1P	single phase
2P	two phase
CDU	coolant distribution unit
CPU	central processing unit
F_h	fin height
F_p	fin pitch
F_w	fin width
GPU	graphical processing unit
GWP	global warming potential
HB1	heater block 1 (with skived fins)
HB2	heater block 2 (smooth, no surface enhancement)
HTC	heat transfer coefficient
PEEK	polyether ether ketone
QD	quick-disconnect
R_{FB}	flow boiling thermal resistance
R_{JI}	jet impingement thermal resistance
TDP	total dissipated power
TIM	thermal interface material
TTV	thermal test vehicle

1. INTRODUCTION

With the onset of Artificial Intelligence GPU and CPU TDP are skyrocketing with every new release. Just in the last three years the maximum TDP of these chips has increased nearly 400%. While the power requirements of these chips get higher and higher, often the junction and case temperature limits are just getting higher. This is resulting in data centers needing to provide cooler facility water to the CDUs in the racks. Already datacenters account for more than 2% of the world's energy usage, and water it and chips getting more powerful warrants the

need for colder facility water directly increasing that energy requirement. [1-3]

In the industry space colloquially, bring fluid to a heatsink mounted on a chip with plumbing is called direct-to-chip, but the fluid does not actually contact the silicon die, While this technology has proven to be the most efficient both thermally, and in terms of deployment in the field, serviceability and retrofitting into existing data centers, the thermal resistance stack up is fixed. It involves a cold plate and TIM, which can only get so efficient. With the ever-increasing power requirements, it's important we look at methodologies that explore direct-on-die cooling, i.e., eliminating the need for a TIM and cold plate and directly cooling a silicon die. This can only be done with a dielectric coolant to prevent short-circuiting the chips, so that rules out the use of single phase, water glycol-based solutions. While immersion cooling is technically bringing fluid to the die, its performance capabilities and difficulty to integrate into data centers has made it an edge-case solution not ready to hyper scale. [4]

To this end, we developed a testing bench to characterize the performance of our pumped two-phase direct-on-die idea. The facility was used to characterize the performance of two heater blocks mimicking silicon dies. Both blocks were designed to mimic a NVIDIA Blackwell B200 GPU, consisting of a skived fin heater block (HB1) and a smooth polished heater block (HB2) meant to mimic a bare die. Two popular flow methodologies in two-phase were tested for both blocks, namely a flow boiling and jet impingement approach. The TTV loads were varied from 0-3kW at various flow rates to characterize the HTC and thermal resistance of the solution. These tests give us a look at the potential for two phase performance, and the option to push manufacturers to install cold plates on their dies/etch microchannels into silicon and further reduce the need for colder facility water. Additionally, when the minimum FW temp required is well above ambient, it opens up avenues for heat reuse and net-zero operation 24/7.

2. EXPERIMENTAL ANALYSIS

2.1 Fluid Circulation Loop

A thermo-hydraulic fluid loop was designed and built to rapidly iterate and test different configurations of cold plates and TTVs in our facility. For this paper, the system was used to validate and characterize the performance of our innovative direct on die approach for high power TTVs, mimicking the latest and greatest GPU architecture. The system involves two sections, separated by liquid and vapor lines with QDs at each end, to simulate a data center environment. The first section is the CDU section, with a pump, brazed plate heat exchanger with facility water flow and reservoir. The server sled section consists of the TTV with plumbing of quality that of a real-life deployment. This is to ensure that any testing is accurate and representative of a real-life server package. These systems being isolated allow for quick iteration and testing of various systems at different flow rates, power levels and facility water temperatures. This loop is depicted in Figure 1.

The fluid chosen for these tests is the Honeywell Blended Solstice N15 (R515b). This is a medium-pressure refrigerant strictly developed for data center cooling. The system starts at the TTV, which in this case is varied between a finned and unfinned heater block. The fluid is impinged on this block with varying manifold designs, while thermal and fluid measurements are recorded at various stages of the system. The liquid-vapor mixture is then transported to the heat exchanger, through the vapor QD and line, where the facility water cools it down. The chosen condenser is sized for up to 10kW of heat rejection, therefore oversized for future tests.

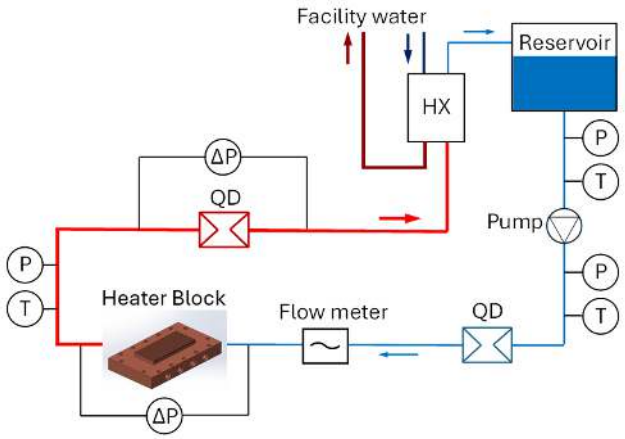


FIGURE 1: FLOW DIAGRAM OF EXPERIMENTAL SETUP

The condensed saturated liquid is then pushed to the reservoir, which is also carefully designed to have enough volume for the liquid and vapor volume throughout different operating conditions, to prevent both cavitation in the pump and a flooded condenser. Finally, the gear pump (Micropump L28640) returns the fresh, slightly sub-cooled liquid back to the TTV through the liquid QD. The liquid is just slightly sub-cooled due to the pressure drops through the system, but for 515b, it is very minimal. This system is a modified version of the loop used in studies [5-6,8]. The facility water temperature is controlled by a dry-cooler made in-house that supports loads of up to 15kW.

2.2 Thermal Test Vehicle and Methodology

The TTV design was innovative as it could not follow the standard procedure of having a two-phase cold plate with TIM between it and the heater block. To simulate a direct on die system, the heater block, mimicking the actual GPU die was directly impinged on by our two-phase refrigerant with different cooling manifolds that follow different flow paths. Two heater blocks, one with skived fins and one polished and flat were tested to understand the heat transfer performance of our system.

The heater block itself is made from copper 110 (highly conductive and 99.9% oxygen free) to distribute heat to the die area. For this study, the NVIDIA B200 (Blackwell) GPU was emulated, with a die area of (60 X 26mm). [7] As seen in previous studies, only the logic portion was modelled, to test with a conservative estimate, and essentially address a higher heat flux than in reality. HB1 is the heater block with skived fins designed to a 0.2mm F_p (fin pitch) 0.2mm F_w (fin width) and 1mm F_h (fin height). HB2 is the smooth heater block. The heat load was provided using 4 cartridge heaters to achieve TDPs of upto 3kW across this die, resulting in an effective heat flux of ~ 200W/cm^2, nearly triple that of the actual Blackwell rated GPU. A T-type thermocouple probe was inserted from the base of the block, just below the die area, with a gap of 1mm.

The various manifolds tested for both configurations were mounted onto heater block and hermetically sealed with the help of an O-ring. An inlet and outlet port flow the refrigerant over the heater block, with a 3D printed nylon insert used to direct the flow internally. This is shown in Figure 2. To maintain as little heat loss as possible, gasket material was used in between the aluminum lid and copper heater blocks to prevent loss through conduction, and PEEK insulation was also used around the heater block. Finally, the entire TTV assembly was insulated with foam to prevent heat loss to ambient.

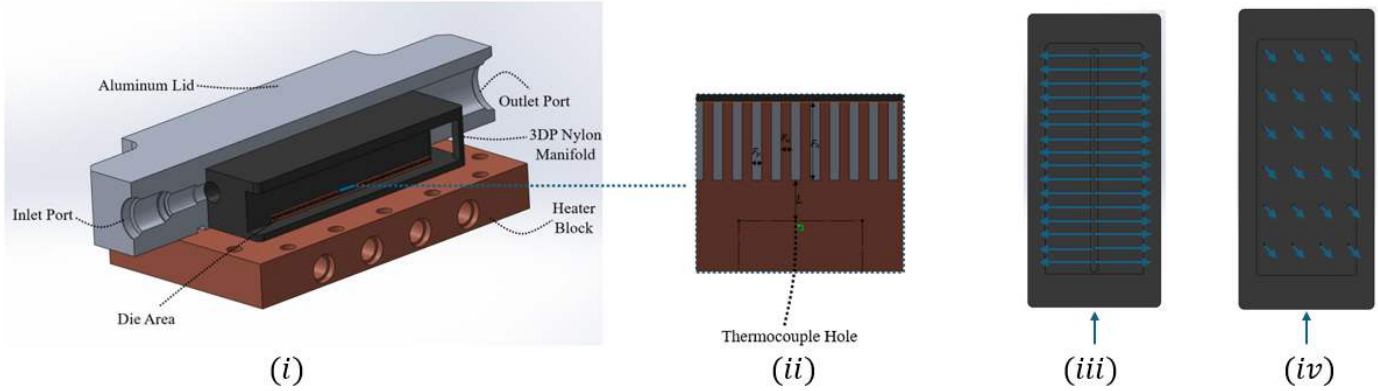


FIGURE 2: Thermal Test Vehicle, with detail on fin stack for HB1, and manifold flow paths

The nylon inserts are shown in Figure 2(iii) and 2(iv). 2(iii) depicts the flow boiling design, which impinges at the center of the die across the fins and forces the flow to boil and flow from both sides, thereby reducing the flow length in half, and reducing pressure drop. A gasket is used between the insert and the fins to ensure that there is no bypass across the fins stack. The insert also covers the rest of the heater block surface in contact with the refrigerant, with a gasket to ensure that the only surface contacting the two-phase coolant is the die area. 2(iv) uses a jet impingement design that would impinge the fluid directly on the die. The insert also insulates the heater block around the die from any refrigerant to prevent boiling at any other surface.

Testing was done with both flow methodologies, i.e., flow boiling and jet impingement, for the skive fin heater block. Just the jet impingement methodology was tested for HB2 due to its high HTC properties at the jets, which would be effective for cooling a surface without any heat transfer enhancement. Further methodology and results are discussed in section 3 of this paper.

2.3 Refrigerant properties

The two-phase working fluid was chosen to be refrigerant R515b also known, sourced from Honeywell, and known as Solstice N15. This is an azeotropic blend of R-1234ze and R-227ea, as a replacement for R134a which is a very environmentally hazardous refrigerant. R515b was specifically engineered for data center cooling due to its unique properties, and low global warming potential (GWP < 300) and non-ozone depleting properties. [9]. It is non-flammable, non-toxic, non-corrosive and has an ASHRAE A1 flammability rating. The properties are specified and depicted in Table 1. This working fluid has zero damage potential in the case of a leak due to its dielectric properties as compared to water. Additionally, its low vapor density allows for minimal pressure drop across complex components like cold plates and QDs improving performance further over other traditional refrigerants. [9

Properties		**Unit**	**R515B**
Global warming potential		-	293
Normal boiling point (@101.3 kPa)		°C	-18.9
Critical temperature		°C	108.7
Density @25 °C	Liquid	kg/m^3	1179.8
	Vapor		27.1
Viscosity @25 °C	Liquid	μPa·s	201.1
	Vapor		12.4
	Liquid	kJ/kg·K	1.66

Specific heat @25 °C	Vapor		1.12
Thermal conductivity @25 °C	Liquid	mW/m·K	73.1
	Vapor		13.9
Latent heat @25 °C		kJ/kg	141.3
Surface tension @25 °C		mN/m	8.8
Saturation pressure @25 °C		psi	72.1
Saturation pressure @50 °C		psi	144.7

TABLE 1: Thermo-Hydraulic Properties of R515b [9]

2.4 Data Instrumentation

All data is recorded using a data acquisition system, the Keysight DAQM901A, and Labview code. System heat was controlled using a thryristor based power controller (Gefran GFX4-IR) wired to the cartridge heaters. Pressure measurements are acquired using absolute and differential pressure transducers (Omega). These are used to measure the pressure drop across the TTV, liquid QD, vapor QD and the entire sled itself. Temperature measurements are made using T-type thermocouples (Omega) at the die, inlet and outlet of the TTV, reservoir and pump. The boiling temperature is controlled by regulating facility water flow rate, and the inlet fluid is minimally subcooled due to the pressure drops across the vapor lines, and heat exchanger, but is maintained well below 3 degrees C throughout the testing. Finally, ultrasonic flowmeters (Keyence FD-XS8) are used to measure flow across the system.

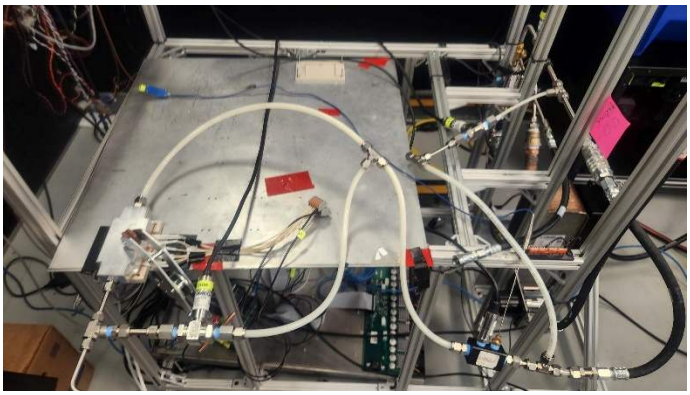


FIGURE 3: EXPERIMENTAL TEST SETUP

2.5 Mathematical Model

Traditionally one would compare the cooling solution performance using case to fluid thermal resistance, defined as the heat load divided by the difference in local fluid temperature and case temperature. To normalize for different die sizes, we can derive this in terms of die area, therefore defined as a resistance in mm^2K/kW.

$$R_{th} = A_{die} * \frac{(T_{case} - T^*_{sat})}{Q_{eff}} \quad (1)$$

Where A_{die} is the die area in mm^2, T_{case} is the measured case temperature and T^*_{sat} is the corrected saturation temperature. Finally, Q_{eff} is the effective heat load, and can be defined as,

$$Q_{eff} = Q_{tot} - Q_{sens} \quad (2)$$

Here, Q_{tot} is the measured heat load using the voltage and current readings supplied to the insertion heaters, and Q_{sens} is the sensible heat portion due to subcooling, which while minimal is not zero. This can be calculated as,

$$Q_{sens} = \rho_l V_l c_p (T_{sat} - T_{in}) \quad (3)$$

Where ρ_l is the density of the liquid, V_l is the measured volumetric flow rate and c_p is the specific heat at the inlet fluid conditions.

The case temperature measured is adjusted for the distance between the probe and the die surface, and we can assume that the heat transfer is 1D and use the conduction equation.

$$T_{case} = T_{\mathrm{TC}} - \frac{P_{GPU} * L}{A_{die} * k_{cu}} \quad (4)$$

Where, k_{cu} is the thermal conductivity of copper in W/mK, and A_{die} is the area of the die in m^2, and L is the vertical distance to the surface in m.

The adjusted saturation temperature includes components of both the two-phase latent heat of vaporization and the sensible heat from the sub-cooled single-phase fluid.

$$T^*_{sat} = \frac{Q_{sens}}{Q_{tot}} \frac{T_{in} + T_{sat}}{2} + \frac{Q_{tot} - Q_{sens}}{Q_{tot}} T_{sat} \quad (5)$$

From equations 2-5, we can derive the effective heat transfer coefficient, to understand what our R_{fb} (flow boiling thermal resistance) is. This is defined in equation 2, as the heat load divided by the area of the die in m^2 and the difference in saturation temperature and case temperature. This metric is used to quantify the effectiveness of the thermal solution, allowing us to also compare the R_{fb} (flow boiling thermal resistance) and R_{ji} (jet impingement thermal resistance) with other two-phase performance data.

$$h_{eff} = \frac{Q_{eff}}{(T_{case} - T^*_{sat}) * A_{die}} \quad (6)$$

Finally, the last important metric in two-phase heat transfer is understanding vapor quality of the mixture that exits after boiling at the heat source. The vapor quality x can be defined as,

$$x = \frac{Q_{eff}}{V_l * \rho_l * h_{fg}} \quad (7)$$

For equation 3, V_l is the volumetric flow rate, ρ_l is the working fluid density and h_{fg} is the latent heat of vaporization.

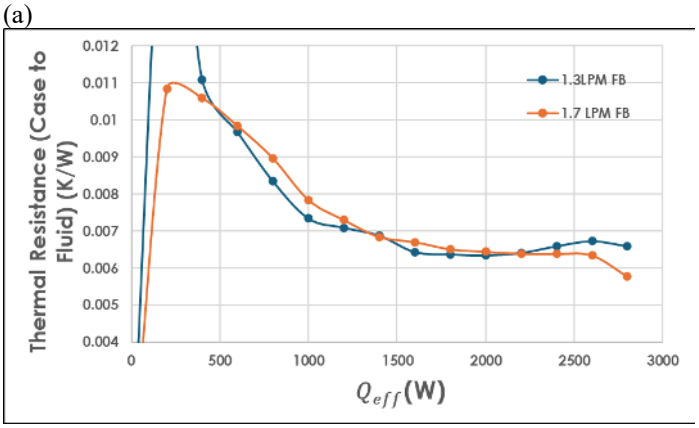
3. RESULTS AND DISCUSSION

3.1 Flow Boiling

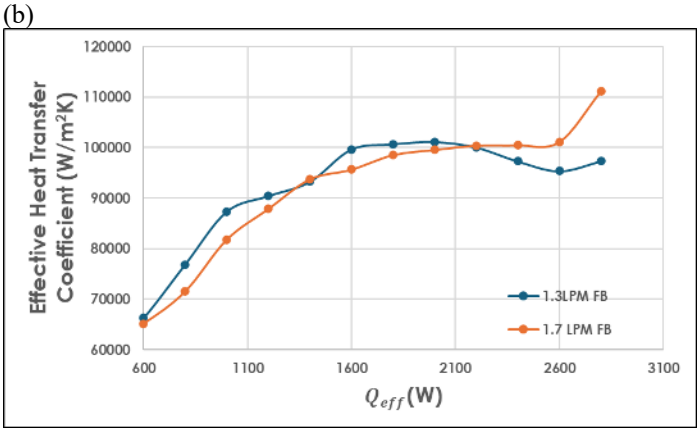
HB1 is tested at constant flow rates, varying the power configuration up to heat flux of over 200 W/cm^2. Various flow rates are tested for both flow manifolds, to understand the effectiveness of both a flow boiling and jet impingement approach. To ensure that the results are recorded at steady state, each power level is maintained for >1 minute, which is more than enough time to achieve a steady state result. Additionally, the readings over 10 seconds of steady state are averaged to account for any other variation. The power is ramped from 100W all the way up to 2.7kW per TTV for two different flow rates per manifold design.

Firstly, let's look at the flow boiling manifold thermal performance for this approach. This manifold is designed with a 2mm slot at the center of the fins, forcing the flow across the length towards the outlet.

(a)



(b)



(c)

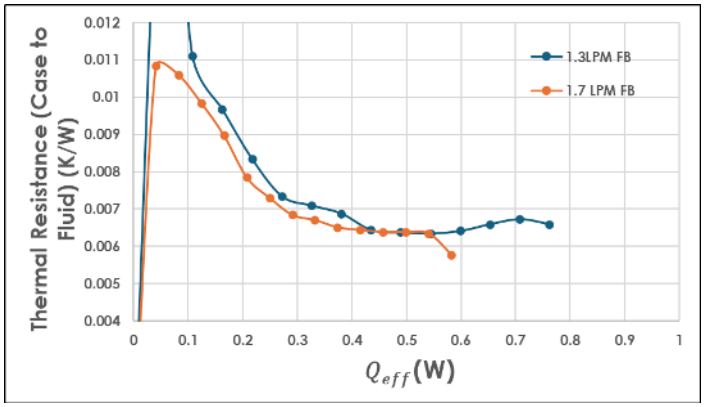


FIGURE 4: (a) R_{FB} vs Q_{eff}, (b) h_{eff} vs Q_{eff}, (c) h_{eff} vs x

Figure 4(a) and 4(b) show the thermal resistance and effective heat transfer coefficient plot against the effective power supplied, and Figure 4(c) shows the same vs exit quality, x.

Here we observe at low power, from 100-300W, the thermal resistance, and effective heat transfer coefficient are both dominated by the 1P heat transfer. This is where we are not able to identify a clear correlation between the heat transfer performance and flow rate. This is probably due to the combined performance at these extremely low vapor qualities of both the 1P convection and 2P boiling incipience. We cannot correlate a higher flow rate to a lower thermal resistance due to the counteracting nature of this multi-phase regime. A higher HTC is observed at higher flow rates, resulting in a larger portion of the overall HTC, being 1P convection. As you approach higher power levels, and higher qualities, the heat transfer regime is dominated by 2P latent heat and the correlation with flow rate tends to disappear for a flow boiling approach. The slight variation in performance can be attributed to the experimental uncertainty within this study.

From exit quality of 0.1 – 0.3 the thermal resistance rapidly trends downwards, as the 2P latent heat transfer starts to dominate the regime. From exit vapor qualities of 0.3 – 0.7 the system performance is essentially the same, maintaining a value of ~0.0062 K/W. In general, the performance is better due to an increased nucleate boiling HTC with increase in supplied power, i.e., higher heat flux, and we can observe this phenomenon in Figure 4(c).

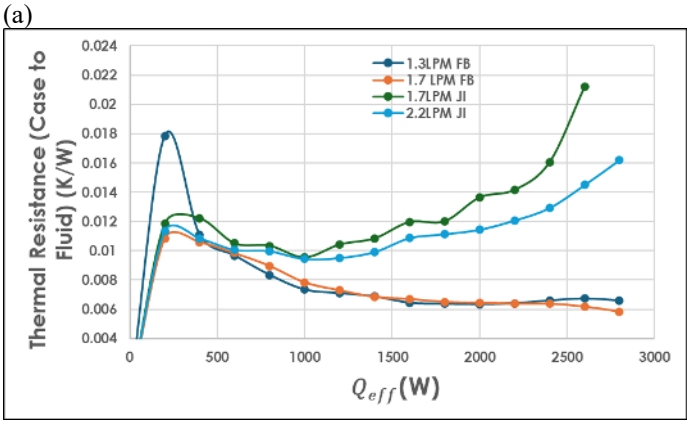
At the rated power for the NVIDIA Blackwell GPU of 1200W, the effective thermal resistance of 11.38 mm^{2}K/W, with HTC of over 80000W/m^{2}K. As you up the power to more than double the Blackwell spec, thereby future proofing this study, the system performs even better. We observe extremely high HTC at the fins, nearing over ~110000 W/m^{2}K at the highest heat flux tested (185 W/cm^2). This was the limit of the insertion heaters we tested with, and future work should explore higher power to see what the theoretical limit of this cooling solution would be. It is also observed that no dry out or critical heat flux limits were achieved at these power levels for these flow rates

further encouraging an exploration of high-power tests [10-15]. It is observed that when the R_{FB} is isolated from a traditional cold plate with the overall thermal resistance being a combination of that, and TIM+copper baseplate conduction, fellow reseachers observed similar values of both HTC and flow boiling thermal resistance in mm^{2}K/W. [10-14]

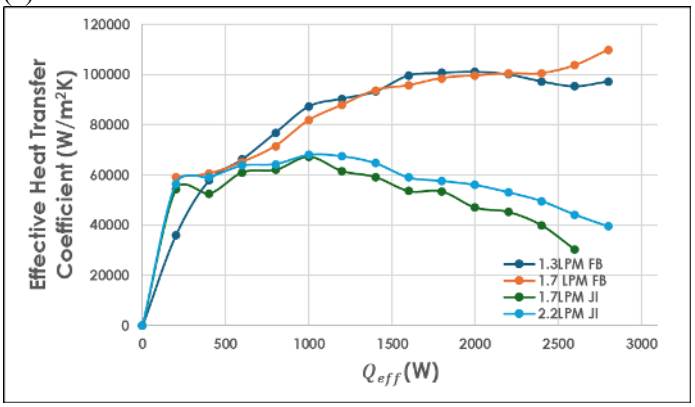
At these higher heat fluxes, there is potential that we depart from traditional nucleate boiling, and there is future study that can explore optimizing the fin and channel width to extract the maximum out of the cold plate. If you take into account this performance alone, a chip manufacturer that can etch channels into their silicon die, like illustrated here [11], we can bring up to 50% warmer facility water by removing the large resistance contributions from TIM and Copper baseplate thickness. At these higher heat fluxes, and smaller dies we can estimate that the effective contribution to overall thermal resistance by flow boiling is only around ~30%. This direct on die methodology would allow data centers to bring much warmer facility water temperatures to their IT space. For an H100, the known throttle temperature is 87C [15]. With a direct on die approach, we can use a saturation temperature at the cold plate of up to 80C.

3.2 Jet Impingement

(a)



(b)



(c)

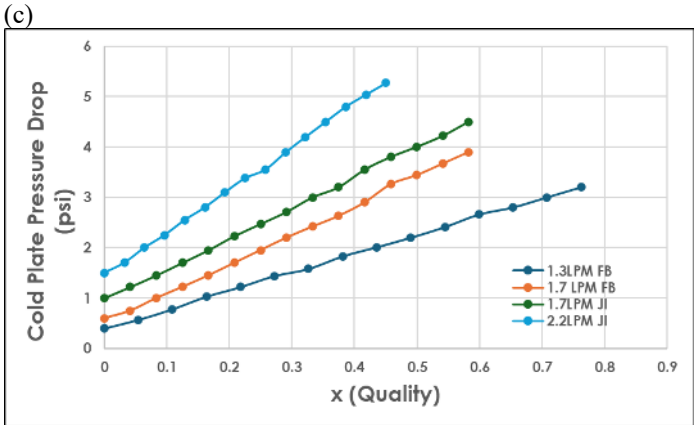


FIGURE 5: (a) R_{FB} vs Q_{eff}, (b) h_{eff} vs Q_{eff}, (c) ΔP_{CP} vs x

We now compare the jet impingement manifold design to the flow boiling approach, and where the differences in performance lie. This manifold is designed with a grid of 5X4 jets, each 0.5mm wide, impinging across the die area. The size of the jets were chosen to ensure high velocity at the jets, with a reasonable pressure drop across the TTV.

At low qualities, of 0.1-0.3, the thermal resistance for the jet impingement manifold is comparable if not better than the values from the flow boiling approach. This is again because the low quality zones are dominated by a 1P cooling regime, and jet impingement results in extremely high velocities at the jets so the 1P HTC is improved dramatically. Despite this there is only a small improvement since the jet velocities are still relatively low as seen by the pressure drop comparison is Figure 4(c). At the rated power of the Blackwell GPU, the system has an effective thermal resistance of 14.82 mm^{2}K/W, with an HTC of close to 68000 W/m^{2}K. As we approach higher power, the jet impingement method suffers in steady state thermal performance. This is a reflection of reaching critical heat flux, especially since the flow is not being forced through the fin stack and has the potential to bypass over the fins and go to the outlet. This theory is further justified by seeing the increase in thermal resistance as you cross power loads of 1500W, which corresponds to a quality of 0.3 and 0.25 for the two flow rates respectively. This showcases that the two phase regime is not being sufficiently used due to bypass, and a lack of forced flow through the channels. Despite this, the HTC >55000W/m^{2}K until loads of 2000W, at a heat flux of close to 150 W/cm^2. This again validates the concept of directly impinging the fluid onto the die surface.

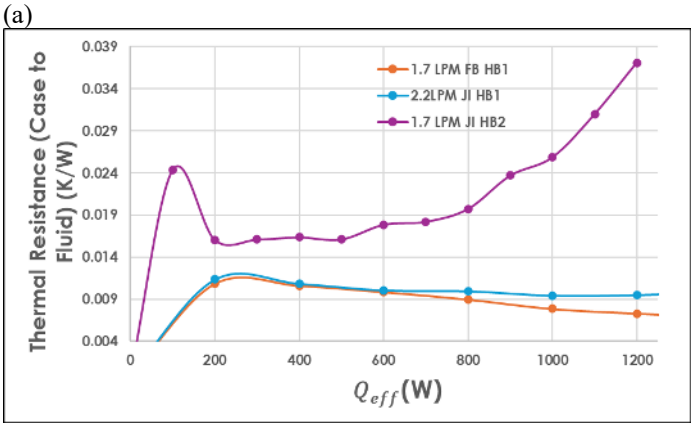
Finally, an analysis of the cold plate pressure drop further pushes the requirement for an optimization of the jet size to ensure high velocities at the local exits, which would greatly improve the heat transfer performance. Especially since the refrigerant used, R515b, does not suffer that greatly from a pressure drop at the cold plate, with approximately a 3-psi drop resulting in just a 1 degree increase in your local saturation temperature. For the same overall flow rate, the jet impingement

manifold results in just a 30% overall larger pressure drop at the cold plate. We know that the pressure drop in the flow boiling arrangement is from the fin stack, and flow manifold due to the gasket forcing flow through the channels. For the jet impingement design, the largest pressure drop is expected at the jets, and if a larger flow rate was utilized, or compensated using smaller jets, there would be higher velocities and a performance improvement here. Literature has seen exceptional performance utilizing fine tunes jet impingement methods. [16,17]

3.2 Jet Impingement on smooth die

This section goes over the tests for direct impingement on smooth die with no fins/heat transfer surfaces on it. This is to draw a direct comparison to a flat silicon die and understand if just impinging two phase fluid on a die would be able to cool it. While a machined and polished copper block does not have the same flatness as a silicon die, the authors tried their best to maintain a very smooth, low roughness copper block for this test. As seen before at super low power of 100W, the HTC is dominated by 1P convection, and HB1 with fins performs a lot bettwe than HB2. As the power increases, the finned HB1 performance starts to shoot up, with much higher thermal performance due to the existence of that fin stack. However, for a die without any enhancement, the jets seem to move to a nucleate boiling regime, and the performance beings to improve. [18]

(a)



(b)

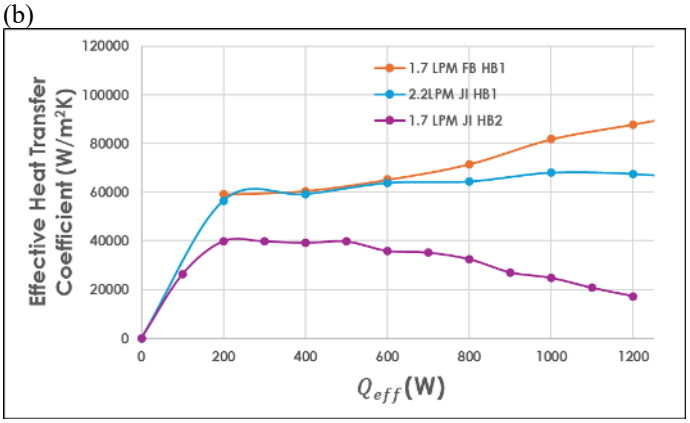


FIGURE 6: (a) *Thermal Resistance* vs Q_{eff} for both flow boiling and jet impingement on HB1 and HB2, (b) h_{eff} vs Q_{eff} for the same.

Until a power level of 700W, current H100/H200 power level the jet cooled smooth block has a thermal resistance of under 0.02K/W.

Here onwards a drop off in performance is observed, going up to the power level of the Black well die at 1200W, the system has a net thermal resistance under 0.04K/W, which is still competitive when compared to the overall thermal resistance of a complete direct-to-chip package including a traditional cold plate and TIM. The steady state measurement here was stable, with signs of nearing critical heat flux as the performance deterioration was significant. Additionally, there is potentially local dry out, that can be validated with more thermocouple measurements in future testing. Similar to the previous jet impingement conclusions, a further optimization of the jet size to increase the velocity through them is essential to extract the maximum performance out of this cooling solution. By optimizing said jets, we can further improve the operating window of this manifold potentially being able maintain HTC of ~40000 W/m^{2}K until that higher heat flux of 80W/cm^2.

4. SUMMARY AND CONCLUSION

In this study, a two-phase direct on die solution for high power GPUs is proposed and tested. A quick iterating tester was commissioned and utilized to rapidly iterate different flow manifolds and heater configurations. A novel heater block and lid is designed with skived fins on the “die-area” mimicking a NVIDIA Blackwell B200 GPU. Two different two-phase flow principles are tested, i.e., flow boiling and jet impingement. The thermal performance and considerations are listed below

- The skived fin copper block is designed with a 0.2mm F_p (fin pitch) 0.2mm F_w (fin width) and 1mm F_h (fin height). This system is inserted in a custom lid, with the ability to use different manifold inserts to direct the flow. A flow boiling insert was optimized with a 2mm slot at the center to distribute the flow evenly through the channels, reducing flow length and pressure drop in half. The jet impingement insert was designed with a 5X4 grid of 0.5mm jets spread out over the die area to impinge onto the heater block directly.
- The case-to-fluid thermal resistance was calculated for varying heat loads, and at the B200 TDP of 1200W, we find that R_{FB} = 0.007 K/W, and R_{JI} = 0.009 K/W.
- For higher heat loads up to 2.8kW, the flow boiling thermal resistance continues to improve as the nucleate boiling HTC increases with heat flux, up to a HTC of over 110000 W/m^{2}K. The jet impingement design starts to degrade deeming an investigation to optimize the jet sizes, and flow path for future TDPs.
- For a data center use case, this study justifies to chip manufacturers to try and design dies with microchannels etched into the silicon, allowing for much warmer facility water temperatures into the CDUs. With a R_{FB} of just 0.007K/W, we will only

observe a 7-8 degree rise between saturation temperature and die temperature at that Blackwell TDP.

- The cold plate pressure drop for both fluid delivery methods are under 5psi, with the jet impingement manifold having approximately 30% higher pressure drop for the same overall flow rate. This further justifies a need to investigate smaller jets to improve thermal performance further at the cost of cold plate pressure drop. This is feasible due to the unique properties of R515b, which only increases by a degree in saturation temperature for every 3 psi of pressure drop.
- A smooth die setup was also tested to validate the idea of impinging high-speed two-phase coolant directly on the die and cool it without needing any heat transfer enhancement surfaces. At H100 power level, the block was able to maintain thermal resistance below 0.018K/W, and at max B200 power levels of 1200W the system performance deteriorates to a thermal resistance of 0.039 K/W.

Future work would involve both optimizing the jet sizes to better compare the two heat transfer methods and understand the tradeoff them. The next steps to further normalize the idea of direct-on-die cooling, would be to use this optimized jet manifold on the heater block with no fins on it. This would help us understand how the system would perform without having to rely on silcon manufacturers to have heat sinks bonded to the chips.

ACKNOWLEDGEMENTS

The authors thank Aurelio Munoz and Clifford Pauley for their support in system setup and experimental testing.

REFERENCES

[1] Davis, A. (2024). *Thermal management in high-density data centers: Challenges and solutions.* Journal of Advanced Computing, 58(3), 204-219.

[2] Smith, B., & Lee, C. (2022). *Power consumption in next-generation data centers.* Journal of Power and Energy Engineering, 40(2), 134-145.

[3] Jones, A., et al. (2023). *Global trends in data center energy consumption.* Energy Policy, 141, 112-121.

[4] Kumar, R., & Zhao, Y. (2023). *Comparative analysis of cooling technologies in data centers.* HVAC&R Research, 29(4), 450-468

[5] A. Narayanan, Q. Wang, S. Ozguc, R. W. Bonner. "Investigation of Server Level Direct-to-Chip Two-Phase Cooling Solution for High Power GPUs." International Electronic Packaging Technical Conference and Exhibition. American Society of Mechanical Engineers, 2024.

[6] Q. Wang, D. P. Kulkarni, R. W. Bonner, J. C. Gulick. "Universal Direct-to-Chip Cold Plates for Single- and Two-Phase Cooling." 2024 OCP Global Summit, Future Technologies Symposium.

[7] "NVIDIA H100 Tensor Core GPU," Nvidia. Accessed: Sep. 26, 2024. [Online]. Available: https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet

[8] Wang, Q., Ozguc, S., Narayanan, A., & Bonner, R. W. (2024). *A Server-Level Test System for Direct-To-Chip Two-Phase Cooling of Data Centers Using a Low Global Warming Potential Fluid*. 2024 23rd IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm).

[9] Honeywell International Inc. "Solstice® N15 (R-515B) Refrigerant." *Honeywell Refrigerants Europe*. https://www.honeywell-refrigerants.com/europe/product/solstice-n15-r-515b/ (accessed May 5, 2025).

[10] White, S., et al. (2023). *Efficiency of direct-to-chip liquid cooling systems.* Applied Thermal Engineering, 179, 1156-1164.

[11] Mudawar, I. (2011). *Two-Phase Microchannel Heat Sinks: Theory, Applications, and Limitations.* ASME Journal of Electronic Packaging, 133(4), 041002. December 8, 2011.

[12] S. G. Kandlikar, et al. Heat transfer and fluid flow in minichannels and microchannels. Elsevier, 2005.

[13] K. P. Drummond, et al. "A hierarchical manifold microchannel heat sink array for high-heat-flux two-phase cooling of electronics." International Journal of Heat and Mass Transfer 117 (2018): 319-330.

[14] C. Woodcock, et al. "Ultra-high heat flux dissipation with Piranha Pin Fins." International Journal of Heat and Mass Transfer 128 (2019): 504-515.

[15] "NVIDIA H100 PCIe 80 GB." TechPowerUp. Accessed: Sep. 26, 2024. [Online]. Available: https://www.techpowerup.com/gpu-specs/h100-pcie-80-gb.c3899#:~:text=Since%20H100%20PCIe%2080%20GB,mm%C2%B2%20and%2080%2C000%20million%20transistors

[16] Myung Ki Sung, Issam Mudawar, Single-phase and two-phase heat transfer characteristics of low temperature hybrid micro-channel/micro-jet impingement cooling module, International Journal of Heat and Mass Transfer, Volume 51, Issues 15–16, 2008, Pages 3882-3895, ISSN 0017-9310,

[17] Shailesh N. Joshi, Ercan M. Dede, Two-phase jet impingement cooling for high heat flux wide band-gap devices using multi-scale porous surfaces, Applied Thermal Engineering, Volume 110, 2017, Pages 10-17, ISSN 1359-4311

[18] Matthew D. Clark, Justin A. Weibel, Suresh V. Garimella, Identification of nucleate boiling as the dominant heat transfer mechanism during confined two-phase jet impingement, International Journal of Heat and Mass Transfer, Volume 128, 2019, Pages 1095-1101, ISSN 0017-9310,