Proceedings of the ASME 2024 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems
InterPACK2024
October 8-10, 2024, San Jose, California

IPACK2024-141261

INVESTIGATION OF SERVER LEVEL DIRECT-TO-CHIP TWO PHASE COOLING SOLUTION FOR HIGH POWER GPUS

Akshith Narayanan
Accelsius
Austin, Texas

Qingyang Wang
Accelsius
Austin, Texas

Serdar Ozguc
Accelsius
Austin, Texas

Richard W. Bonner III
Accelsius
Austin, Texas

ABSTRACT

As artificial intelligence and technology advance, top end server CPU and GPUs are getting more and more powerful. Conventional air-cooling struggles to keep up with the surging processor power densities, leaving room for an efficient and scalable solution. This investigation focuses on the Nvidia H100 GPU architecture in a mimic 1U server with four-way GPUs. While direct to chip two phase cooling offers high heat transfer coefficients and heat capacity, there are technical challenges associated with pressure drop and flow maldistribution to understand the requirements and cost of its use. A single sled test bench is built to conduct systematic experiments with in-house designed Thermal Test Vehicles (TTVs) that replicate the form factor of Nvidia H100 GPUs. The facility is used to characterize pressure drops, die temperature, pumping power and thermal resistance at varying heat loads and flow rates, using R1233zd(E) as the working fluid. Flow restrictors were carefully chosen to balance the flow between the chips, making sure each of them receives sufficient liquid in all usage cases. The results show that the thermal resistance of the cold plate assembly varies between 0.02-0.03W/K. Higher power tests, up to 1.35kW, were conducted to understand how future-proofed the cold plate design is, in terms of rapidly scaling heat loads in the market. Results show that with sufficient flow, the system operates extremely well. The flow restrictors chosen scale quadratically with pressure drop and mass flow rate, and they efficiently suppress the maldistribution between chips on the server level and can be expanded to balancing flow on the rack level. The investigation concludes that pumped direct-to-chip two-phase cooling offers an elegant and scalable solution to the rapidly increasing power densities of CPUs and GPUs, while maintaining efficiency and minimizing the leak risks compared with using a water-based solution.

Keywords: direct-to-chip cooling, two-phase heat transfer, data center cooling, thermal test vehicle

NOMENCLATURE

1U	standard unit of measurement for height of server rack cabinets, which is equal to 1.75”
CDU	coolant distribution unit
CRAC	computer room air conditioner
CRAH	computer room air handler
GWP	global warming potential
PFAS	per- and polyfluoroalkyl substances
QD	quick-disconnect
TTV	thermal test vehicle

1. INTRODUCTION

Data centers are at the heart of technological advancement in today’s world. With the rapid onset and advances in artificial intelligence, CPU and GPU chips are getting more powerful with each passing day. The escalating demands of AI have resulted in chips getting extremely powerful, and data centers contributing significantly more towards the carbon footprint of the world. Currently, they account for about 1% of the global electricity, and that is slated to double over the next few years [1]. Of this 1%, the energy distribution states that most of it comes from the power hungry high-density servers. [2]

Traditionally, air cooling (CRAC or CRAH) units circulate cool air within the racks of a data center. Since this method is the least intrusive, it was widely adopted, when server loads were significantly lower, and overall racks were much lower power. With the increase in power density on both the server and rack level, along with servers becoming more compact, air-cooling methods struggle to keep up. This is where liquid cooling methods start to shine, significantly reducing power consumption of the data centers, due to their superior heat transfer. [3,4]

With the need for alternatives to air cooling becoming ever apparent, different cooling technologies emerged as potential solutions. Rear door heat exchangers were the first response to this problem since they integrate extremely well with existing air-cooled solutions. However, they have some of the same drawbacks that air cooling does and are limited by the heat loads that they can draw, often limited to not higher than 30kW per rack [4]. Different liquid cooling methods emerged in the form of Immersion cooling. It is another unique solution that involves submerging the servers themselves in pools of non-conductive liquid (often oil). The working fluid directly contacts the servers themselves enhancing heat removal capabilities, but they have incredible challenges in terms of cost, serviceability and sheer volume of working fluid [5].

Single-phase water emerged as one of the first direct-to-chip liquid cooling applications as a response to the limitations of air cooling and other liquid cooling technologies, due to its excellent thermal properties and ease of application [6]. However, it also faces its own issues with requiring very high flow rates as compared to two-phase and any leak in the system is disastrous for the electrical components in the server. As the power densities and complexity of modern chips has increased, these servers have become very expensive and even one leak can be very costly.

Pumped two-phase direct-to-chip is a superior solution to the power demand of servers and data centers, while addressing the limitations that the other liquid cooling solutions pose [7]. By bringing the working fluid directly to the chip and boiling the fluid in an attached cold plate, the heat transfer is greatly improved, allowing the removal of ultra-high heat fluxes. This solution cuts down the energy needed to cool as compared to air cooling and can be retrofit to existing servers with ease as compared to immersion cooling. [8]

Much research has been done to understand pumped two-phase heat transfer and its intricacies, but testing and implementation in data centers is very limited. Due to the number of parallel liquid branches in a single rack, and the complexities of thermal and fluid transport, it is hard to understand the performance and create a stable solution. It is of great importance to highlight solutions that understand and balance these complex flow paths and showcase the advantages of a working pumped two-phase direct-to-chip system.

To this end, we developed a testing bench at the server-level to characterize the performance of our pumped two-phase direct-to-chip solution. The facility has the capabilities to test various server arrangements, and this study focuses on mimicking a 1U four-way GPU server, modelled after the Nvidia H100 architecture. Tests were conducted up to the TDP of 700W per GPU, and further up to 1.35kW per GPU to understand the scaling of the solution and how it holds up against the higher thermal requirements of chips yet to be designed. Thermo-hydraulic measurements were taken to determine a case-to-fluid thermal resistance and understand flow distribution through the parallel paths across the server. The GPU loads were varied to extreme scenarios to quantify the flow balancing using flow restrictors across the server and showcase steady performance. The tests give us valuable insight into the capabilities of a well-designed pumped two-phase direct-to-chip system and the option to integrate them into data centers today.

2. EXPERIMENTAL ANALYSIS

2.1 Fluid Circulation Loop

A thermo-hydraulic fluid loop is developed and built to characterize the performance of the direct- to-chip solution. The facility involved the CDU section and the server sled section, and they were isolated enough to be able to rapidly iterate changes and test different servers at different power levels and flow rates. This flow loop is shown in Figure 1. For this experimental investigation, a 1U, four-way GPU was chosen.

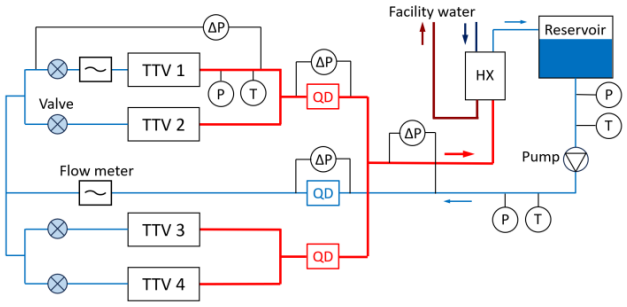

FIGURE 1: SCHEMATIC/FLOW LOOP OF EXPERIMENTAL TEST SYSTEM

The facility consists of a flat plate heat exchanger, with facility water on one side, a reservoir and a pump on the CDU section. The server sled then connects to this CDU via quick disconnects (QDs), to mimic a rack level solution. The brazed plate heat exchanger is designed to carry heat loads of up to 10kW, thereby future proofing this system to work with the ever-increasing heat loads of modern CPUs and GPUs. Refrigerant R1233zd(E) is the working fluid in this loop. The reservoir holds the fluid volume in the system and was carefully designed to have enough space for both the liquid and the vapor volume during operation to prevent flooding the condenser and cavitation in the pump. The pump pushes slightly sub-cooled liquid through the liquid QD, which then vaporizes over the four heated GPU TTVs. The two-phase fluid then exits through two vapor QDs and on to the condenser. The temperature of the condensate is maintained by a water-cooled chiller that is attached to the condenser. . Thermo-hydraulic measurements are carefully acquired at the TTVs, QDs, reservoir and pumps.

2.2 Refrigerant Properties

The two-phase working fluid was chosen to be refrigerant R1233zd(E) due to its unique properties, and ultra-low global warming potential (GWP = 1). Additionally, this fourth-generation refrigerant breaks down into naturally occurring compounds in the environment within a few days, unlike other forms of PFAS [9]. It is non-flammable, non-toxic, non-corrosive and has an ASHRAE A1 flammability rating. This working fluid has zero damage potential in the case of a leak when compared to water, due to its dielectric properties. Additionally, its low vapor pressure at typical operating conditions presents a low chance of leakage and alleviates safety concerns.

2.3 Thermal Test Vehicle

A TTV (thermal test vehicle) is designed to mimic the Nvidia H100 GPU architecture. The TTV design matched the actual chip in terms of both die size and total power dissipation and can be seen in Figure 2.

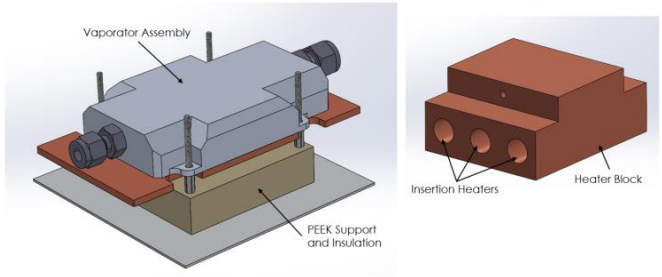

FIGURE 2: CAD DIAGRAM OF THERMAL TEST VEHICLE

A copper heater block is used to distribute the heat from cartridge heaters to the die area (approximately 26mm×35mm) similar to a Nvidia H100 GPU. A T-type thermocouple probe is inserted into the heater block 2mm below the surface to estimate the case temperature of the die. The heaters can output heat loads up to a combined 2.5kW per TTV, nearly three times as high as the 700W TDP of the H100 chip, allowing us to test extreme heat loads and understand is the design margins available for future processors. The cold plate assembly is mounted onto the heater block with a recommended mounting force and thermal interface material.

2.4 Data Instrumentation

All data is recorded using a data acquisition system, the Keysight DAQM901A, and the vendor provided data logging software. Pressure measurements are acquired with the help of both 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, to predict pumping power requirements for a rack level cooling system. T-type thermocouples (Omega) are used to measure temperatures of the saturated liquid at the reservoir, TTV inlet and outlet temperatures, and case temperature of the chip. Finally, ultrasonic flowmeters (Keyence) are used to measure flow across one of the parallel paths, and the total flow in the system. These are selected due to their unique properties that allow them to clamp on to the external tubing and estimate the flowrate without inhibiting flow and causing additional pressure drops that in-line or Coriolis flowmeters create. This setup is seen in Figure 3.

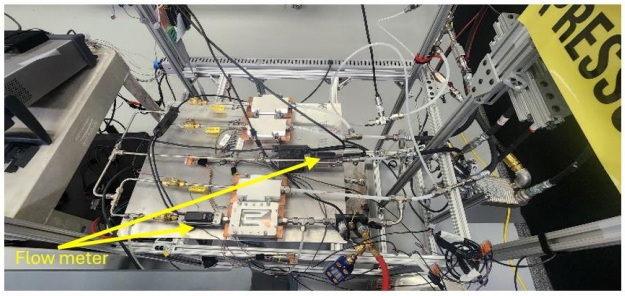

FIGURE 3: EXPERIMENTAL TEST SETUP

2.4 Mathematical Model

The case-to-fluid thermal resistance is used to quantify performance of the cooling solution and is defined as the heat load of the die divided by the difference in case and local fluid saturation temperature. Traditional single-phase heat transfer uses fluid inlet temperature, but that is because all of the heat transfer is sensible. In the case of two-phase heat transfer, it is mostly latent heat, and the sensible heat is near zero, hence we use saturation temperature. [7]

$$R_{th} = \frac{T_{case} - T_{sat}}{P_{GPU}} \quad (1)$$

Here, P_{GPU} is the power of the GPU and is measured using the voltage and current readings supplied to the insertion heaters. The case temperature is measured using a 1D conduction equation to the surface from the thermocouple inserted near the end.

$$T_{case} = T_{\mathrm{TC}} - \frac{P_{GPU} * L}{A_{die} * k_{cu}} \quad (2)$$

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.

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,

$$x = \frac{P_{GPU}}{Q * \rho_l * h_{fg}} \quad (3)$$

For equation 3, Q is the volumetric flow rate, ρ_l is the working fluid density and h_{fg} is the latent heat of vaporization. Again, the sensible heat contribution is neglected here.

3. RESULTS AND DISCUSSION

3.1 Uniform Heating

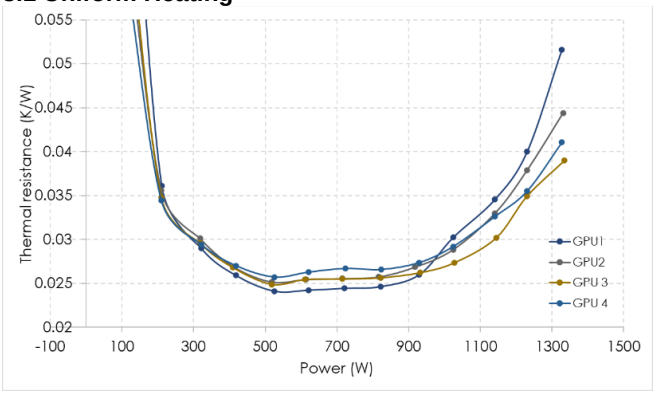

FIGURE 4: Case to fluid thermal resistance vs GPU power with all four TTVs uniformly heated

To achieve steady state results, the server sled is tested at a constant flow rate and varying power configurations. In this section we discuss these results, with the next section focusing on performance and flow distribution when a single TTV is pushed to high power, while keeping the others at zero power.

The first set of tests were conducted by steadily increasing the power of all the GPU TTVs at the same rate, while supplying constant flow rate of approximately 450-500 mL/min per TTV. The results of this test can be seen in Figures 4 and 5. Figure 4 showcases the relationship between case-to-fluid thermal resistance and GPU power for all four TTVs. Figure 4 also shows the flow distribution between the four parallel channels to showcase the effect of the flow restrictor.

From Figure 5, we can see that all four GPUs perform extremely similarly, ensuring that the flow distribution through the system has been carefully balanced. The four GPUs are heated from 100W to the H100 rated TDP of 700W, and then up to 1.35kW per GPU to understand its performance past the H100 threshold. At lower power (<100W), we can see that the single-phase heat transfer is quite poor, and therefore results in an extremely high thermal resistance. This is since the heat fluxes are not significant enough to cause nucleation, and the cold plate design only shines when these sites become active. As the heat loads increase, we see a rapid drop in the thermal resistance, down to the average low value of 0.02-0.03K/W. This is due to the superior heat transfer characteristics of flow boiling. This rapidly reduces the case temperature, and proportionally reduces thermal resistance.

Figure 5 shows the case temperature of the GPUs as a function of power supplied. We can see that for the rated TDP of 700W, the case temperature reaches 65°C. We can also estimate the junction temperature based on the calculated case-to-fluid thermal resistance of the cold plate. For all four GPUs, the max junction temperature is around 80°C, which is 7 degrees lower than the throttling limit of 87°C for the H100 GPU.

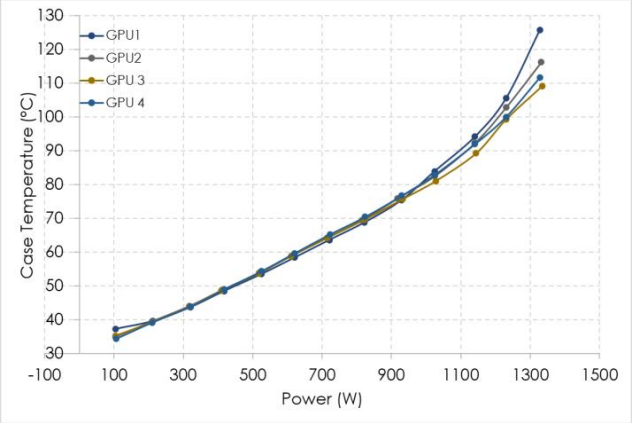

FIGURE 5: Case temperature of the four GPUs vs power

This showcases the ability of this generation of cold plate to very effectively and efficiently cool an H100 GPU, at a relatively low flow rate of just 500mL/min per GPU. Increasing said flow rate would allow the technology to extend its cooling abilities past even the 1,000W mark. All these results also maintain the quality of the mixture well below 0.55. With two-phase heat transfer, dry -out phenomenon can result in the cold plate space being flooded by 100% vapor, resulting in extremely poor heat transfer characteristics and thermal runaway. In this case, we can maintain that fraction well below that limit, giving the system significant margin to a failure state. In this system, we try to maintain that vapor exit quality to 0.7, to allow for said margin.

3.2 Non-uniform Heating

The next set of experiments conducted simulated an extreme real-application case, involving a single GPU pushed to maximum power while the other parallel stems are switched off. This is an extreme case that can occur when different GPUs on the same server are pulsed or engaged at varying power, and the flow distribution can severely affect throttling conditions. To simulate this, we pushed a single GPU up to 1.35kW, and kept the remaining GPUs at zero power, and then plotted the variation of thermal resistance of the single system as a function of power and vapor quality. This was done keeping the total sled flow rate fixed to 2000 mL/min; 500 mL/min per GPU during regular operation.

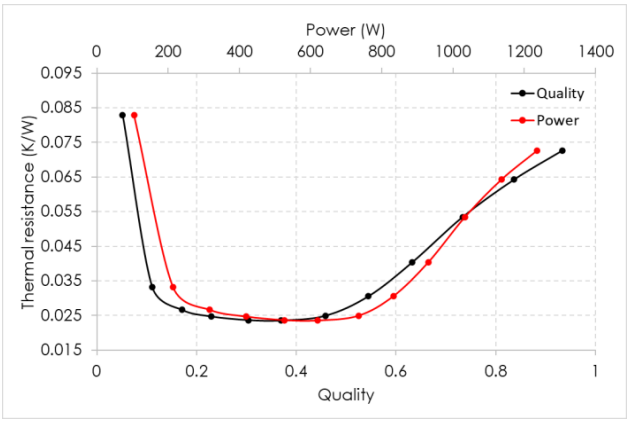

FIGURE 6: Case to fluid thermal resistance vs GPU power and vapor quality

As explained above, at low power (<100W) the single-phase heat transfer is poor, resulting in high thermal resistance which is seen in Figure 6. The system operates extremely well within the 100-800W range, at a vapor quality below 0.6. This shows that for that power range, 500mL/min is adequate flow rate, keeping that thermal resistance, and proportionally case temperature below throttling limits. For higher power, higher flow rate and optimized flow paths are necessary. Despite this, when the system is heated uniformly, it performs exceptionally well even at the higher heat loads. Even in this extreme case, with a single GPU running close to dry-out, and the remaining three parallel channels running 100% liquid volume the system

can maintain said case and junction temperatures well below the throttling limit.

3.3 Flow Distribution

A crucial component of two-phase direct-to-chip when it comes to these power dense GPU servers is the flow distribution between the chips on the same server, and across the rack [10]. In this case, there are four GPUs per sled, and each GPU is supplied flow via parallel paths. This study focuses on the server-level, so balancing flow between these chips is of utmost importance to ensure steady thermal performance regardless of the power load each chip has. Here, four GPUs are connected in parallel paths, and each individual cold plate has approximately 250 parallel channels. To balance the flow amongst each, in-house designed flow restrictors are used at the inlet of each GPU. These are carefully optimized such that this pressure drop dominates every other pressure drop, i.e., at the TTV, tubing, QDs etc., in the system.

Figure 7 depicts the flow ratio of a single GPU to the total flow for both uniform and nonuniform heating conditions. The flow ratio here is defined as the flow rate in a single GPU divided by the overall flow rate. Therefore, a flow ratio of 0.25 indicates a perfectly balanced flow distribution. When each GPU is uniformly heated, the flow ratio is almost constant at that 0.25 mark, ensuring that the system is well balanced. The data acquired also shows a <10% flow maldistribution between GPUs on the same server for any given power level.

When we heat up a single GPU, while keeping the other three at a zero-power level, we can see that the flow ratio starts dropping as the power increases. The flow resistance starts increasing significantly in the heated TTV as compared to the unheated ones due to vapor pressure drop and this can be seen in the graph. Despite this, there is still only about a 20% deviation in flow, and the system does not dry out. More interestingly, at the 700W rated TDP power level, the system only has about a 9% flow imbalance, and the thermal resistance is still very good, at ~0.028K/W.

Finally, looking at Figures 4 and 5 allows us to determine how effective these flow restrictors are in balancing the flow. The case temperatures between the four GPUs are never more than 5% in variation. This results in a very balanced and uniform system that ensures no single chip can overheat or dry out before others. The thermal resistances are also in-line with the others ensuring that this system performs uniformly regardless of the power output of each GPU.

Balancing each parallel flow channel in each cold plate and extending that to each GPU is a monumental task, given the instabilities of two-phase flow. It is a phenomenon still being investigated [11,12] and this system ensures that each system gets sufficient flow, ensuring uniform performance.

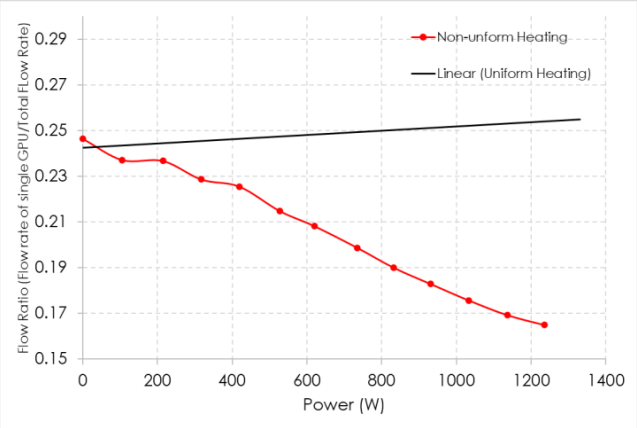

FIGURE 7: Flow ratio of single GPU during both uniform and non-uniform heating conditions

4. SUMMARY AND CONCLUSION

In this study, a two-phase direct-to-chip liquid cooling solution for a power dense 1U four-way GPU server is analyzed. A thermo-hydraulic facility with the ability to hot-swap different server arrangements is built to characterize the performance of this solution. Thermal test vehicles are designed to mimic the Nvidia H100 GPU architecture, and four of them are placed in a 1U server space, with the microchannel vaporator assembly attached to the heater block with a thermal interface material. The thermal performance and flow distribution across the server are illustrated.

- The case-to-fluid thermal resistance was calculated for varying heat loads, and at the H100 TDP of 700W, there is an average resistance of 0.025K/W across the four GPUs. For higher heat loads up to 1.35kW, the thermal resistance starts to degrade deeming an investigation to optimize liquid and vapor flow paths necessary for future chips.
- The case temperatures of the GPUs at the 700W TDP are well below 65°C. With the calculated thermal resistance, the estimated junction temperature if 7-10°C lower than the throttling limit of the GPUs. Future iterations with optimized cold plate geometry can even yield lower thermal resistances and increase the margin even further.
- The system performs extremely well even under imbalanced heat loads, where a single GPU is pushed to 1.35kW, while the other three are “off”, showcasing a real-life extreme condition for application in data centers.
- Custom designed flow restrictors are able to maintain flow imbalance under 10% for the defined TDP of 700W and maintain them below 20% up to 1.35kW. The system does not reach dry out and maintains stable case temperatures throughout.

Future work to optimize cold plate geometry to improve flow distribution between liquid and vapor pathways would ensure performance to ultra-high heat fluxes. Additionally, modifying the system facility to better characterize pressure drops across different components like QDs and tubing would help us understand the flow regime of two-phase solutions. This optimized vaporator design would address the high-power superchips incoming in the GPU and CPU space as they get more power dense. Our work showcases the novelty and advantages of a two-phase direct-to-chip solution over traditional air cooling and other liquid cooling methods. While there are challenges regarding flow distribution, balancing it effectively allows the solution to target ultra-high heat flux chips without the need to increase flow rate beyond safe operating conditions. The solution is robust, scalable and ready to retrofit onto existing server and rack platforms, while reducing overall power consumption by over 40%.

ACKNOWLEDGEMENTS

The authors thank Aurelio Munoz, Tyler Richardson, Levi Jordan, Jacob Moore and William Allai 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] Simon, S., Modi, V., Sivaraju, K.B., Bansode, P., Saini, S., Shahi, P., Karajgikar, S., Mulay, V., & Agonafer, D. (2022). *Feasibility Study of Rear Door Heat Exchanger for a High Capacity Data Center.* Proceedings of the ASME 2022 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems, V001T01A018

[6] Heydari, A., Gharaibeh, A. R., Tradat, M., Soud, Q., Manaserh, Y., Radmard, V., Eslami, B., Rodriguez, J., & Sammakia, B. (2024). *Experimental evaluation of direct-to-chip cold plate liquid cooling for high-heat-density data centers.* Applied Thermal Engineering, 239, 122122. ISSN 1359-4311.

[7] 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).

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

[9] Bonner, R., Grieco, B., Cruz, L. (2024). *Understanding PFAS Concerns for Two-Phase Cooling of Data Centers*. https://www.datacenterfrontier.com/sponsored/article/33035570/understanding-pfas-concerns-for-two-phase-cooling-of-data-centers

[10] Ozguc, S., Wang, Q., Narayanan, A., & Bonner, R. W. (2024). *Investigation of Flow Restrictors for Rack Level Two-Phase Cooling Under Nonuniform Heating.* In 2024 40th Semiconductor Thermal Measurement, Modeling & Management Symposium (SEMI-THERM) (pp. 1-6). IEEE.

[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] Abdollahi, A., Sharma, R. N., & Vatani, A. (2017). *Fluid flow and heat transfer of liquid-liquid two phase flow in microchannels: A review.* International Communications in Heat and Mass Transfer, 84, 66-74. ISSN 0735-1933.