Two-Phase Micropillar Evaporators to Enable Cooling of Next-Generation GPU Servers

Qingyang Wang
Accelsius
Austin, TX, USA
qwang@accelsius.com

Richard W. Bonner
Accelsius
Austin, TX, USA
rbonner@accelsius.com

Dereje Agonafer
The University of Texas at Arlington
Arlington, TX, USA
agonafer@uta.edu

Sai Abhideep Pundla
The University of Texas at Arlington
Arlington, TX, USA
saiabhideep.pundla@mavs.uta.edu

Braxton J. Smith
The University of Texas at Arlington
Arlington, TX, USA
bjs4224@mavs.uta.edu

Damena Agonafer
University of Maryland
College Park, MD, USA
agonafer@umd.edu

Vivek Vardhan Manepalli
University of Maryland
College Park, MD, USA
vmanepal@umd.edu

Kidus Jeglalo Guye
University of Maryland
College Park, MD, USA
kguye@umd.edu

Hermann von Drateln
Celestica
Alviso, CA, USA
hvondrateln@celestica.com

Robert Fernandez
Celestica
Alviso, CA, USA
rfernandez@celestica.com

Bill Boudsamad
Celestica
Toronto, ON, Canada
bboudsa@celestica.com

***Abstract*—The rapid development of AI technologies has been driving surging power densities in data centers. The next generation of AI servers will have densely packed GPUs or AI accelerators with high heat flux dissipation, posing significant challenges for thermal management. Two-phase direct-to-chip cooling offers great thermal performance at the processor level to address the pressing cooling needs, with the ability to dissipate high heat flux with minimized temperature differential due to the intrinsic high heat transfer coefficient of liquid-vapor phase-change heat transfer. Ultralow thermal resistance has been demonstrated on a Direct-to-Chip Evaporative Cooler (DCEC) developed through research efforts funded by the ARPA-E COOLERCHIPS program. Accelsius has also developed MR250, a 250kW in-row two-phase coolant distribution unit (CDU), demonstrating high power capacity and thermal performance. The combination of cutting-edge evaporator technology and industry-leading two-phase CDU offers untethered potential to address the thermal needs of next-generation AI servers and processors using direct-to-chip two-phase cooling. In this paper, we analyze the system-level performance metrics when a state-of-the-art 8-way GPU server is cooled by the DCEC, along with Accelsius' MR250 CDU. The estimated GPU case temperature can be maintained to be below 68 °C even with 40 °C facility coolant supply and 10 °C facility coolant temperature rise, when the CDU is cooling a full IT load of 250 kW. The results suggest great potential for energy saving by allowing high facility coolant supply temperature, and sufficient headroom of direct-to-chip two-phase cooling for future generations of processors/servers/IT racks with even higher power densities and heat fluxes.**

Keywords—liquid cooling, direct-to-chip, two-phase cooling, data center, thermal management

I. Introduction

The rapid development of AI technologies has created an insatiable demand for computational power and efficiency, which leads to increasingly densely packed racks and servers in data centers. The thermal design power of AI optimized processors has also been increasing drastically, driving the need for more efficient thermal solutions to dissipate the resulting high chip-level heat fluxes.

Single-phase direct-to-chip cooling using water/glycol mixture is among the earliest adopted liquid cooling solutions to replace traditional inefficient air cooling. However, the use of water in data centers could cause catastrophic damage to IT equipment in the event of leakage. More importantly, as AI development keeps driving increased heat fluxes of processors, single-phase direct-to-chip solutions will struggle to meet the cooling demand due to insufficient thermal performance.

Two-phase direct-to-chip cooling offers superior thermal performance [1-4], especially at the processor level, due to the intrinsic high heat transfer coefficient of liquid-vapor phase-change heat transfer. Additional benefits offered by two-phase cooling include the isothermality across the processor surface which alleviates silicon warpage, the small flow rate requirement which reduces pumping cost and mitigates erosion, and the use of dielectric refrigerant which prevents IT damage.

In order to tackle the thermal challenges of data centers, the US Department of Energy's ARPA-E has established a COOLERCHIPS research program. As part of the program, UT Arlington and University of Maryland developed a Direct-to-Chip Evaporative Cooler (DCEC) with ultralow thermal resistance [5]. Accelsius has recently joined the research project and complemented the team with an MR250 product, which is a multi-rack, in-row, 250kW two-phase CDU. The combination of cutting-edge evaporator technology and the industry-leading high-capacity two-phase CDU presents exciting opportunities for addressing the high power and high heat flux requirements posed by next-generation advanced AI processors and servers.

In this work, we analyze the performance metrics of the two-phase cooling solution combining the DCEC evaporator with the MR250 CDU, used to cool a state-of-the-art server consisting of 8-way AMD MI355X GPUs. The estimated

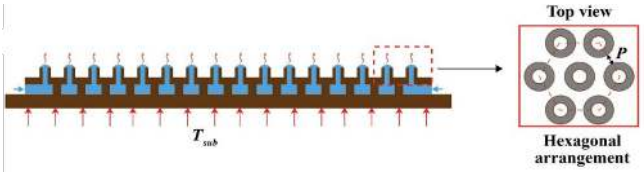

Fig. 1 Schematic showing the micropillar-based evaporator design [5].

GPU case temperature can be maintained to be below 68 °C even with 40 °C facility coolant supply and 10 °C facility coolant temperature rise, when the CDU is cooling 250 kW of thermal load. The results suggest great potential for energy saving by allowing high facility coolant supply temperature. It also suggests sufficient headroom of direct-to-chip two-phase cooling for future generations of processors/servers/IT racks with even higher power densities and heat fluxes.

II. Direct-to-Chip Evaporator

The DCEC represents a state-of-the-art thermal management technology designed to address the escalating heat dissipation demands of high-performance computing and AI-driven platforms. The system utilizes arrays of silicon-based hollow micropillars integrated with liquid delivery layer to enable microscale evaporation of dielectric coolants [5] as shown in Figure 1, which can achieve heat fluxes up to 400 W/cm².

The DCEC is engineered for minimal pressure drop and high scalability, which enables effective heat dissipation across large chip modules while managing inlet pressure requirements under high thermal loads. The high thermodynamic utilization of the dielectric fluid (exit vapor quality of 1) ensures extremely small flow rates to dissipate high heat loads. Compared to existing technologies, this novel evaporator offers enhanced reliability with dielectric fluids, supporting future electronic systems where traditional cooling methods fall short.

III. Next-Generation AI Server

The next generation of AI servers will require closely packed GPUs and switches to enable efficient computation. DrMOS technology is already a staple in high-performance GPUs and motherboards due to its efficiency and power delivery capabilities. Liquid cooling is essential for advanced, high-density DrMOS GPUs and CPUs due to the intense heat they generate under heavy workloads. These components operate at high frequencies and voltages, making traditional cooling solutions insufficient for maintaining optimal temperatures. Figure 2 shows the perspective view of an advanced AI server, showing 8 AMD MI355X GPUs packed inside the server tray, with cold plate-based liquid cooling as a default cooling solution. A near-term transition from single-phase to two-phase direct-to-chip cooling requires minimal re-design as the cold plates can be shared without any two-phase specific modifications, and two-phase still offers better performance [6]. Cold plates specifically optimized for two-phase cooling, such as the DCEC described above, can further enhance the cooling performance.

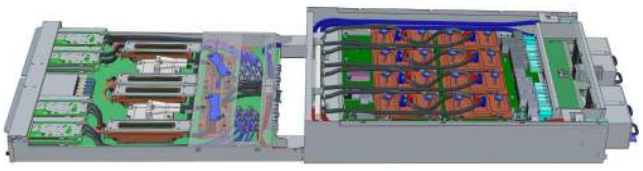
Fig. 2 Perspective view of the server tray with dual CPUs and 8-way GPUs.

IV. Cooling Performance

Lab-scale testing of the DCEC has demonstrated ultralow case-to-fluid thermal resistance, which includes contributions from thermal interface materials, conduction through the structures, and phase change heat transfer. The flux-based thermal resistance can be considered independent of the working conditions with the same evaporator structure. We then estimate the GPU case temperature when the GPUs in the 8-way AI server is cooled with evaporators designed specifically for the GPU package, and the servers are populated in a rack connected to an Accelsius' MR250 two-phase CDU. Given the heat flux of the GPU package and the MR250 thermohydraulic data, the estimated GPU case temperature is below 68 °C with a facility coolant (PG25) supply temperature of 40 °C and a full load on the MR250 CDU. This 250 kW of load can be from a single rack or multiple racks. The facility coolant flow rate is assumed to be 1.5 L/min per kW of cooling load, to maintain a 10 °C facility coolant temperature rise. A lower CDU power load can further reduce the case temperature.

Maintaining a 68 °C GPU case temperature using a 40 °C inlet facility water temperature is noteworthy. With this level of thermal headroom, it is expected that a future generation of GPUs with approximately double the heat flux and power could be maintained below the typical GPU maximum case temperatures. Furthermore, if facility water temperatures could be reduced to 30 °C, the power and heat flux could be nearly tripled over the MI355X TDP and heat flux levels.

V. Conclusion

As the next-generation AI servers with advanced processors become increasingly power intensive, two-phase direct-to-chip cooling offers the ability to dissipate the high heat flux at the chip level with minimized temperature differential, which allows high facility coolant temperature while maintaining low case/junction temperatures. Thin film evaporators in particular promise to show the most optimal two-phase cooling performance. By combining a thin film DCEC developed by UT Arlington and UMD along with the two-phase CDU MR250 developed by Accelsius, the two-phase cooling solution applied on an advanced AI server allows the GPU case temperature to be maintained below 68°C with 40°C facility coolant supply at 250 kW of cooling load. The results suggest that two-phase direct-to-chip cooling offers sufficient headroom to address the cooling demands of the next-generation processors/servers/racks, whose heat fluxes and power densities will continue to increase in the next several years.

References

[1] Q. Wang, et al. "A server-level test system for direct-to-chip two-phase cooling of data centers using a low global warming potential fluid", IEEE ITherm, 2024.

[2] Q. Wang, et al. "A practical metric for cold plate thermal performance in two-phase direct-to-chip cooling", Semi-Therm, 2025.

[3] Q. Wang, et al. "Performance comparison of R1233zd(E) and R515B for two-phase direct-to-chip cooling", IEEE ITherm, 2025.

[4] Q. Wang and R. W. Bonner. "High-performance two-phase cooling under different cold plate orientations." OCP EMEA Summit Future Technologies Symposium, 2025.

[5] K. Guye, et al. "Design and modeling of hollow micropillars evaporator for thermal management in high heat flux applications: Numerical analysis." Applied Thermal Engineering 262 (2025): 124977.

[6] Q. Wang, et al. "Universal direct-to-chip cold plates for single- and two-phase cooling." OCP Global Summit Future Technologies Symposium, 2024.