Low-Chill™: Hierarchical Manifolds for Data Center Chip Cooling

David A. Smith, Christopher Roper, Jeremy Orosco, Adam Gross

HRL Laboratories

Inquiries please contact lowchill@hrl.com

Extending single-phase liquid cooling for AI hardware

High performance microprocessors enable the artificial intelligence (AI) revolution. Each successive generation of AI chips dissipates an increasing amount of thermal power. Traditional air cooling is no longer viable and, without a breakthrough, single-phase liquid cooling could soon be insufficient. The data center industry would be required to adopt unproven technologies and invest significant capital to complete the transition to two-phase cooling, creating a high barrier to entry and hindering profitability. HRL has developed the breakthrough technology necessary to extend the viability of single-phase liquid cooling for current and future generations of AI hardware. This technology is available for licensing and ready to enter the market, providing an immediate solution for the thermal management demands of the lynchpins behind the proliferation of AI.

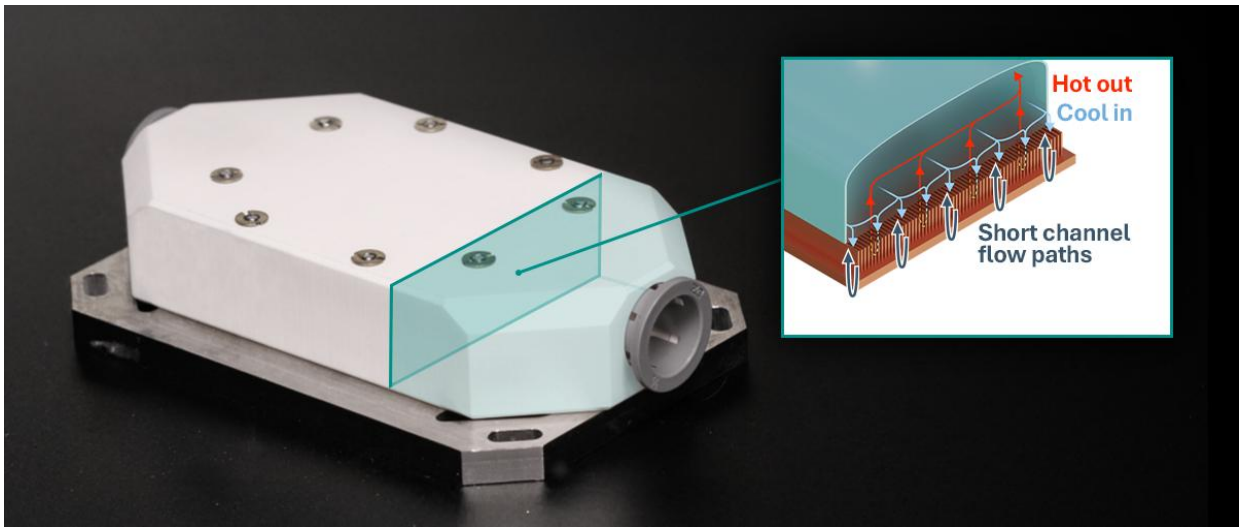

Figure 1: HRL's Low-Chill™ hierarchical cooling block, designed to be drop-in compatible with existing direct liquid cooling infrastructure. The cooling block version shown is compatible with a 1U server and sized to mount to a 10.6 cm² active chip area.

HRL's Low-Chill™ single-phase liquid cooling block produces ultra-low thermal resistance, thus extending the viability of single-phase liquid cooling for high power density

AI chips, servers, and data centers. By leveraging our expertise in design for additive manufacturing and novel materials, we achieve record-setting thermal resistances (8.2 ˚C/kW) at a very low pumping power (< 5 W pump power per 1kW thermal dissipation) to extend the viability of direct liquid cooling through the 2030's. Interweaving 3D hierarchical levels of inflow and outflow channels form a novel manifold structure that, when mated with a cold plate, creates a drop-in solution capable of handling the high heat flux density produced by the AI microprocessor chips of today and tomorrow.

Our technology

HRL's Low-Chill™ cooling block leverages 3D printing to enable cooling architectures and flow paths that are not possible today with conventional manufacturing methods. The manifold contains complex flow paths that separate the inflow and outflow channels while routing fluid from a singular inlet/outlet connection. This results in a record-breaking thermal resistance and pressure drop across the cooling block. Figure 1 depicts a schematic of the flow distribution pathways within HRL's cooling block, printed using high-temperature plastic for a 1U server form factor, mated to a conventional fin plate. **The cooling block is designed to be drop-in compatible with existing single-phase fluidic connections and data center infrastructure.** Additionally, Low-Chill™ has been designed with elevated fluid temperatures in mind, allowing fluid temperatures as high as 70 ˚C. Our low thermal resistance extends the capability of warm water cooling to higher power densities and heat fluxes for years to come. The freedom of design and manufacturing from 3D printing allows designing cooling blocks that can fit within any server form factor and sized to any chip architecture.

Low-Chill™ is ready today to extend the capacity for single-phase cooling to handle MW racks. Figure 2 shows HRL's cooling block performance compared to a commercial off-the-shelf (COTS) cooling block with a similar microchannel fin array. Both blocks were tested on an AM5-sized die simulator, dissipating 1 kW of heat. **HRL's cooling block shows >10X lower pressure drop or 40% higher power dissipation compared to the COTS cooling block.** The low pressure drops result in low pumping powers (< 5 W) required to achieve the above performance, typically <1% of the IT load. Ultra-low thermal resistance, 8.2 ˚C per 1 kW on a 10.6 cm^2 area (equivalent to an AM5-socket size and comparable to an NVIDIA GB202 die), allow for a higher fluid temperature while keeping the chips below a critical temperature (typically <80 ˚C to avoid thermal throttling). HRL's experience with design for additive manufacturing allows the technology to be readily scalable to any chip size and arrangement. **Ultimately, HRL's Low-Chill™ technology can handle heat flux densities up to 400 W/cm^2**, or 3 kW for a single 750 mm^2 die, while using <1% of the IT load for pumping power.

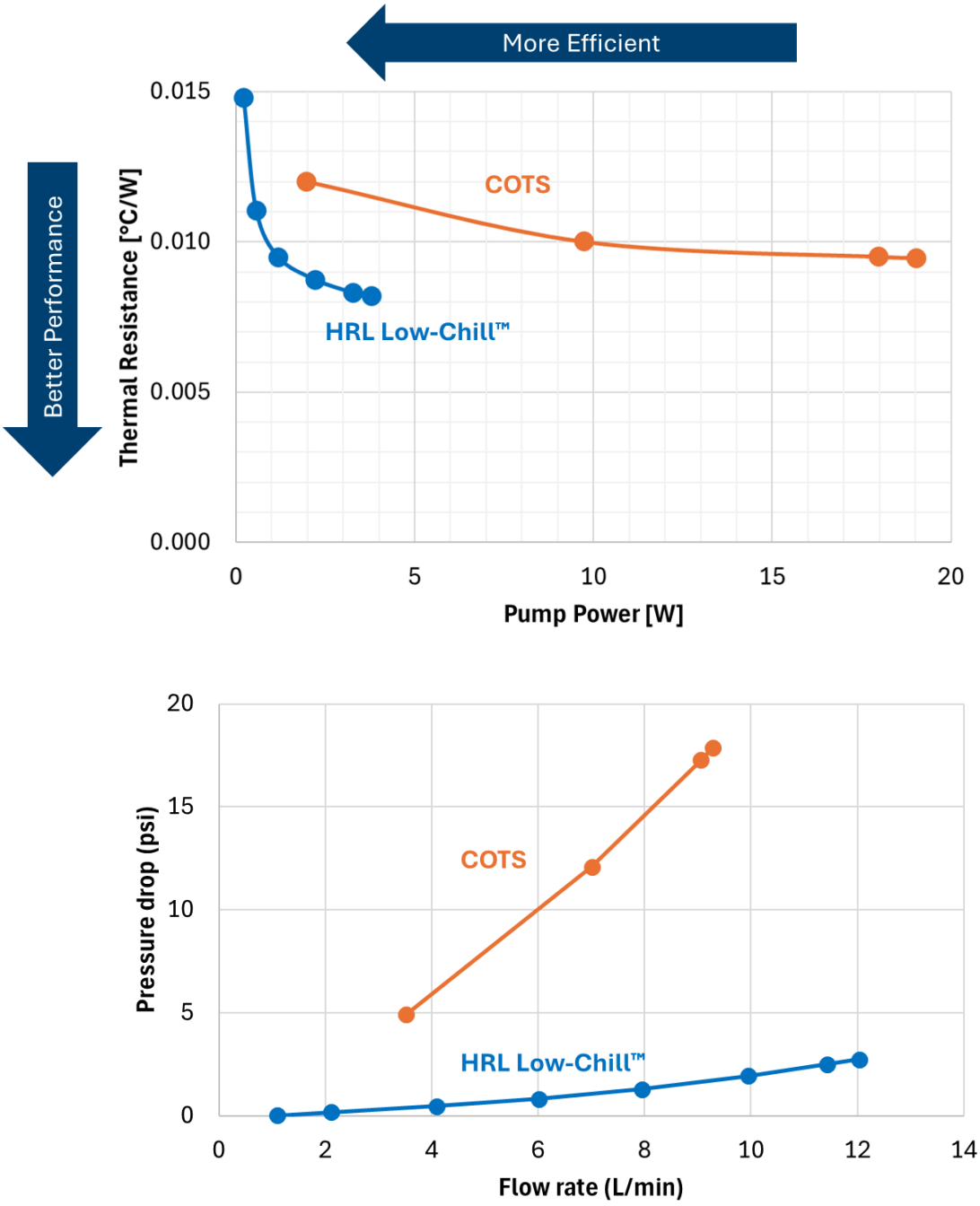

Figure 2: (Top) Thermal resistance vs. pump power (product of pressure drop and volumetric flow rate) and (bottom) cooling block pressure drop vs. volumetric flow rate for HRL's Low-Chill cooling block and a near-equivalent COTS cooling block. Tests were performed on a 1 kW, 10.6 cm^2 thermal test vehicle. The lines are guides to the eye.

Ready for transition

HRL has designed the manufacturing and assembly of our cooling block to be compatible with the existing single-phase liquid cooling hardware ecosystem. We have worked with multiple 3D printing suppliers to validate our designs across multiple

manufacturing systems and materials to meet the anticipated high-volume production necessary for data centers. **The manifold can be 3D printed from a wide range of materials, including copper alloys, stainless steels, and high temperature polymers.** All other components are standard and the necessary supply chains have been established. A conventional skived fin plate, made from copper or aluminum depending on the performance needed, is attached to the manifold. Manufacturing the cooling block is cost competitive to existing solutions, verified in partnership with multiple external manufacturing houses. A foundational and forward-looking comprehensive IP portfolio related to this technology has been established.

Why does it matter to you?

Cooling block manufacturers have the opportunity to license a superior cooling block design with an established and scalable design and manufacturing workflow for multiple chip architectures for a future-proof product line. They can immediately offer 40% higher power dissipation or 10x lower pumping power than the competition with a ready-to-transition technology. This technology can support future 400 W/cm^2 power densities to mitigate future generations of hot spots.

Server manufacturers can employ already proven single-phase liquid cooling for future generations of AI servers, fulfilling the needs of data center operators and hyperscalers for years to come. The cooling blocks are compatible with all existing direct liquid cooling components but require 1/10th the pumping power of existing technology, thus providing ease of install and more AI Tokens per Watt (TPW) of power. Furthermore, the cost of the Low-Chill™ system is similar to midrange direct liquid cooling options. Thus, combined with drop-in compatibility with existing hardware, Low-Chill presents no additional barriers to customers.

Data center operators looking to retrofit and upgrade their hardware for existing single-phase liquid cooling see a <1 year ROI from using HRL's drop-in compatible solution that requires lower powered pumps and reduces system downtimes from the lower flow rates and pressures. Furthermore, as flow rates and pump pressures are determined by the most resistive component attached to a mixed server equipment fluid loop, any Low-Chill™ upgrade will require lower pressures and flows than existing equipment and thus never increase operational requirements. Despite the expected year-on-year generational growth, HRL's Low-Chill™ cooling block allows the same pump to be used without upgrading and redesigning coolant distribution units.

New build data center operators can dramatically lower their operational expenses and time to return on investment by enabling megawatt-scale racks using reliable single-

phase cooling and reducing or eliminating their water usage by increased fluid temperatures and dry air coolers, which offers a high ROI versus installing evaporative coolers. When paired with an evaporative cooler, HRL's Low-Chill™ solution enables MW-scale compute racks with required pumping powers <1% of the associated IT load (Figure 3).

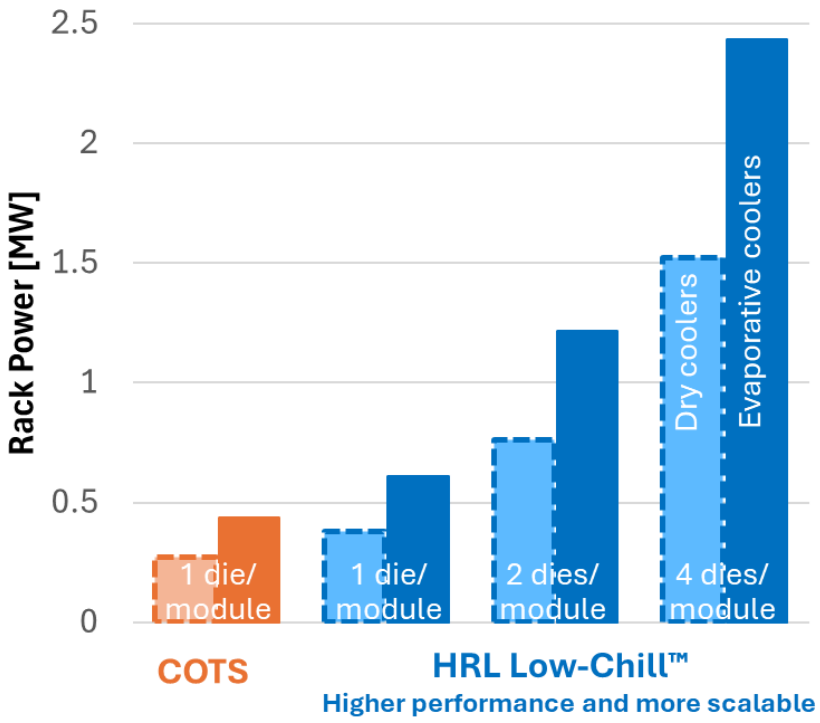

Figure 3: HRL's cooling block enables rack power densities higher than the competition, surpassing the MW-rack scale when combined with an evaporative cooler and multichip modules. We assume 4 multi-chip modules (MCMs) each containing a number of GPU dies per server, 42 servers per rack, 23% additional power for CPUs and memory, and a 2 mm^2K/W thermal interface material (TIM), with a coolant to ambient rejection temperature delta of 25 °C for dry coolers (dashed lines) and 40 °C for evaporative coolers (solid lines).

Who are we?

HRL Laboratories LLC is a research and development lab can trace its legacy back to Hughes Research Lab. Located in Malibu, CA, USA and jointly owned by The Boeing Company and General Motors, we have a long-standing history of delivering innovations in materials and microsystems for automotive, aerospace, defense, and energy applications to advance the critical missions of our customers. Using our previous experience of designing architected systems and materials for additive manufacturing in aerospace, we have developed this high-performance cooling block for data centers as part of the Department of Energy's ARPA-E COOLERCHIPS program. Click here to learn more.