Special Analysis

Microfluidics: Potentially Cooling the Future of AI

Jaclyn Ludema and Mark Nossokoff
November 2025

HYPERION RESEARCH OPINION

The ongoing demand for more powerful Artificial Intelligence (AI) and High-Performance Computing (HPC) is creating significant thermal challenges in many data centers. As AI chips continue to advance in capability, they are generating substantial levels of heat, pushing conventional cooling technologies closer to their operational limits. Rack power densities have increased considerably, with high-density racks in 2014 typically around 10 kW per rack, compared to current high-density configurations often exceeding 100 kW per rack. Moreover, industry discussions suggest that megawatt-scale racks may emerge in the coming years. This thermal challenge could potentially constrain performance gains, increase operational costs, and complicate future hardware innovation.

A developing technology, microfluidic cooling, has emerged as a candidate solution to this challenge. Recent tests by Microsoft, in collaboration with Swiss startup Corintis, have demonstrated an innovative approach: moving cooling from the outside of the chip package to the interior of the silicon chip itself. By etching small channels directly onto the chip, this approach may deliver more targeted cooling with improved efficiency. Microfluidics could represent a meaningful shift that may help address thermal constraints on next-generation AI hardware.

FIGURE 1

Microfluidic Channels on Chip

Source: Corintis, 2025

SITUATION ANALYSIS

The leading chip-level liquid cooling approach, Direct-to-Chip (DTC) cold plates, involves mounting a metal plate containing liquid channels on top of the chip package. While DTC represented progress from air cooling, its limitations are becoming more apparent. A key limitation of DTC is that the layers of packaging material between the silicon and the cold plate can act as thermal insulators, potentially impeding heat transfer. These layers function somewhat like "blankets," constraining the cooling system's effectiveness. With future AI chips likely to generate even more heat, the industry may face scenarios where performance could be limited more by heat dissipation capabilities rather than processing power.

A developing alternative liquid cooling technology, microfluidics, involves manipulating fluids in microscopic channels etched directly on or within the silicon itself. Several companies and research institutions are exploring variations of this approach. In particular, the collaboration between Microsoft and Corintis provides an interesting proof-of-concept. Their approach demonstrates two notable innovations:

- **Direct Silicon Contact:** By etching microchannels directly onto the back of the silicon, the cooling liquid makes direct contact with the heat source, potentially bypassing the insulating packaging layers.
- **AI-Optimized Design:** The system uses AI and simulation to create complex, bio-inspired channel networks, somewhat like veins in a leaf. This approach allows coolant to be routed more precisely to the chip's specific "hot spots," potentially improving efficiency.

The reported performance gains from the Microsoft-Corintis approach appear substantial. The tests suggest that this microfluidic cooling system may be as high as three times more efficient at removing heat than current DTC cold plates and could lower chip temperatures by a significant margin compared to traditional air cooling (reported as over 80% in their tests).

Likewise, JetCool Technologies has developed an alternative microconvective cooling approach that uses tiny jets of fluid directed at chip hotspots. Their technology has been tested in collaboration with the U.S. Air Force Research Laboratory and other partners for HPC applications. Cooler Master Technology has also investigated embedded cooling solutions incorporating microchannels, though their primary focus has been on gaming and consumer electronics rather than data center-scale deployments. IBM Research has historically explored on-chip cooling as part of their 3D chip stacking initiatives.

Beyond commercial efforts, significant research is underway at academic and government laboratories. The COOLERCHIPS program, a DARPA-funded initiative involving Georgia Institute of Technology (Georgia Tech), NREL, and Sandia National Laboratories, is developing advanced microfluidic cooling solutions that integrate directly with silicon chips to achieve heat removal capabilities of at least 1,000 watts per square centimeter. Stanford's NanoHeat Lab has also been conducting fundamental research on nanoscale and microscale thermal transport, exploring how to optimize heat transfer at the chip level.

Market Perspective

These technical metrics suggest several potential market implications. Cooler operating temperatures generally allow chips to be run at higher frequencies to handle peak demand, which may improve both performance and cost-effectiveness.

- **Potential Performance Enhancement:** For HPC/AI datacenters, this could translate to improved performance from existing hardware, potentially enhancing both speed and cost-efficiency.
- **Possible Reliability Improvements:** Lower temperatures tend to correlate with reduced failure rates, which may decrease downtime and replacement costs.
- **Sustainability Considerations:** The reported efficiency of microfluidics suggests the coolant may not require aggressive chilling, potentially reducing the datacenter's overall energy consumption (PUE). Additionally, the targeted nature of the cooling could reduce the total volume of liquid needed, which may help address environmental concerns about water usage in large-scale AI facilities.

Implementation Challenges

Manufacturing Complexity and Scale

The precision required for microfluidic cooling presents considerable manufacturing challenges. Creating microscopic channels with dimensions measured in micrometers typically requires specialized fabrication techniques and equipment that differ significantly from conventional cooling system production. These manufacturing processes may require substantial investment in new facilities and capabilities. As with many emerging technologies, early production costs tend to be higher, and achieving the necessary economies of scale to make microfluidic cooling cost-competitive with established alternatives may take time.

Integration and Standardization

Integrating microfluidic cooling into existing data center infrastructure and workflows can present some practical challenges. Unlike established cooling solutions that benefit from decades of standardization, microfluidic systems may require custom interfaces, connections, and operational protocols. Ensuring reliable, leak-free connections between the microscale cooling channels and conventional data center plumbing systems can be complex. Additionally, the lack of industry-wide standards for microfluidic components may complicate procurement, maintenance, and component interchangeability. For organizations with significant existing investments in liquid cooling infrastructure, the transition to microfluidic technology would likely need to demonstrate clear return on investment to justify the changeover costs and potential operational disruption.

FUTURE OUTLOOK

Microfluidics appears to be transitioning from a laboratory concept into a technology with potential applications in the AI and HPC infrastructure market. The diversity of approaches being pursued, from Corintis's etched channels to JetCool's microconvective jets, suggests that multiple technical pathways may prove viable for different applications. For data center operators, chip manufacturers, system vendors, and cloud service providers, microfluidic cooling represents an approach worth evaluating. Assessing and potentially adopting this technology may become an important strategic consideration for organizations seeking to address thermal challenges in AI and HPC workloads as the market continues to evolve.

The adoption trajectory for microfluidic cooling technology appears to be developing across two phases. In the near term, Corintis is scaling manufacturing to produce one million copper cold plates with integrated microfluidic channels by the end of 2026, offering data centers an enhanced alternative to conventional DTC systems for current infrastructure. Looking further ahead, Microsoft has successfully tested a more advanced approach during internal demonstrations where microfluidic channels are etched directly into the silicon chip itself during manufacturing. While still in testing, this chip-integrated approach could potentially deliver substantially greater cooling performance and may enable new chip architectures, such as 3D-stacked designs, that are currently constrained by thermal management limitations.

About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology, and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). Hyperion Research provides thought leadership and practical guidance for users, vendors, and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com

Copyright Notice