Microsoft Azure

Microfluidic cooling:

Cooling at the micro level for Microsoft's datacenters

Up to 3x better cooling performance than cold plates[1]

Microsoft's lab-scale tests showed microfluidics performed up to three times better than cold plates at removing heat. This leads to a 65% reduction in maximum temperature rise of the silicon inside a GPU.

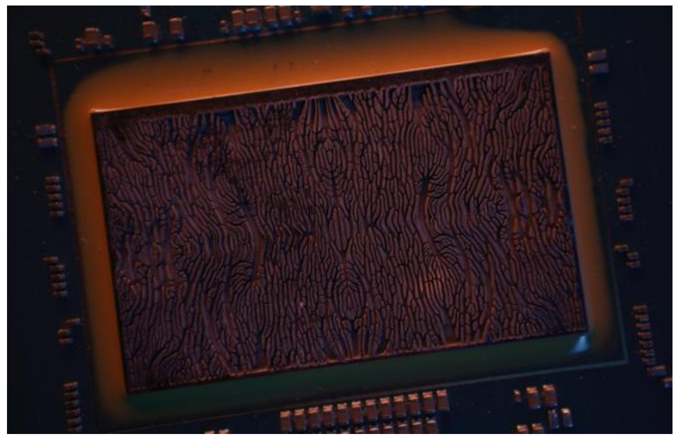

Figure 1: Chip with etched microfluidics cooling channels

Microsoft's unique approach to AI infrastructure considers every layer of the stack. As compute and AI workloads continue to grow, our team takes a holistic approach to optimizing silicon, servers, and the datacenters that bring them all together for the best performance and better power usage.

As we grow our cloud's physical footprint, we're doing so in a way that helps us progress toward our sustainability goals of being carbon negative and water positive, creating zero waste, and protecting more land than we use.

[1] 3x better cooling performance than cold plates depend on workloads and configurations involved

How microfluidics keeps chips cool

Microfluidic cooling is part of our systems approach to advancing AI infrastructure from chips to servers to data center all the way to the cloud. Through cooling innovations like this, we can open the door to new architectures that use less water and power to cool datacenters.

In addition to our advances in liquid cooling datacenters and open sourced, server-level cooling, Microsoft introduced our latest cooling innovation using microfluidics. Microfluidics is the science of moving liquid through microchannels, typically micrometers in size, similar to the size of a human hair.

Efficient and precise

Microsoft optimizes the channel design microfluidics uses for cooling with the help of AI. By mapping out the heat signatures of each chip, microfluidics brings liquid coolant directly to hot spots. Plus, embedding the liquid cooling inside the chip brings the coolant right next to the active silicon and cores.

Removes extreme heat flux

Dissipates more than 1 kW/cm^2, handling 2–3 times the heat flux of standard cold plates.[1, 2, 3]

Microsoft lab tests show coolant flowing inside the silicon chip reduces maximum temperature rise of the silicon inside a GPU by up to 65%.

Micropin-fin heat sink

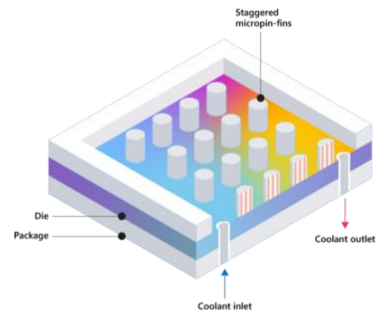

Figure 2: Render of microscopic pins where liquid coolant flow to absorb and dissipate heat

References

[1] A Review of Recent Developments in "On-Chip" Embedded Cooling Technologies for Heterogeneous Integrated Applications
[2] Moving Towards Microchannel-based Chip Cooling
[3] The ICECool Fundamentals Effort on Evaporative Cooling of Microelectronics

Optimizing performance with microfluidics

Proximity to active cores

Microfluidic cooling also addresses other challenges like keeping High Bandwidth Memory (HBM) cool and enables new designs that deliver power more efficiently from the back of the chip.

Overclocking feasibility

Overclocking microchips involves increasing the clock speed of components, which can lead to increased performance. The cooling and thermal margin allowed by microfluidics enables additional advantages to increase performance.

Part of a greater system

Taking a holistic approach and systems thinking approach, Microsoft looks at cooling through the lens of interactions across the workload, silicon, coolant, server, and the datacenter to unlock new architectures for advanced workloads and

Microfluidics cooling system

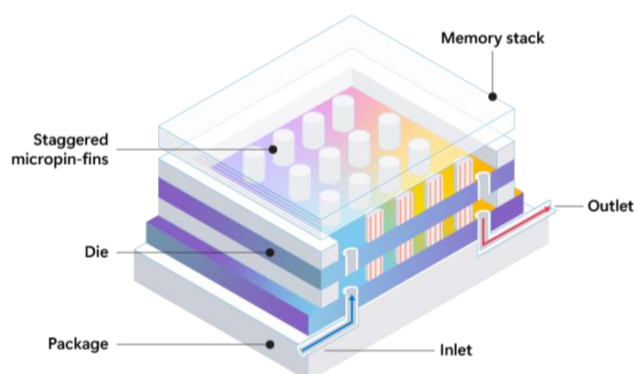

Figure 3: Render of stacked chips with a microfluidic cooling system

enhancing the efficiency of existing rack designs like in Figure 3.

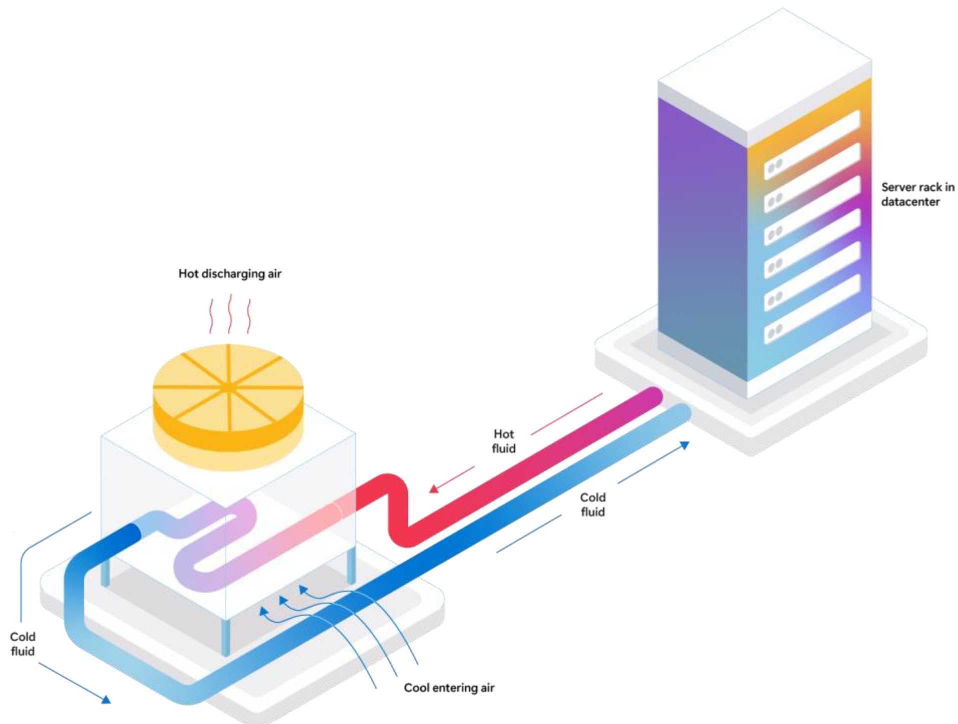

Figure 4:
Render of how microfluidics enables the use of dry coolers

Explore More

Watch Microsoft's cooling breakthrough with etched microchannels in our video

Discover how microfluidics is reshaping datacenter cooling—and what it means for the future of cloud and AI in our Microsoft Source Blog

Read our silicon microfluidic cooling research in the IEEE and Research Square.

Figure 5: A chip enclosed in casing with coolant pipes within a server blade