Microsoft

Liquid cooling heat exchanger units

Transforming air-cooled datacenters for AI workloads

Figure 1:
3D rendering of an HXU

Bringing liquid cooling to air-cooled datacenters

Microsoft is continuously architecting and optimizing every layer of the cloud and AI infrastructure stack to meet the demands of our AI advancements. Modern AI systems powering AI workloads demand higher power at higher densities, leading to a need to develop new methods of cooling to manage heat and power consumption.

To address this challenge, liquid cooling has emerged as the preferred solution, using circulating fluids to dissipate heat and prevent overheating to ensure efficient operation of AI hardware. At Microsoft, as we continue to expand our datacenter fleet to help enable the world's AI transformation, we're also developing methods for using air-cooled datacenters to provide liquid cooling capabilities for AI.

Microsoft is enabling advanced cooling innovations for modern AI systems, revolutionizing how datacenters housing GPUs and AI accelerators are cooled. We are integrating liquid cooling in legacy air-cooled datacenters, retrofitting existing facilities and quickly deploying to meet the cooling needs of AI systems.

The introduction of the dedicated liquid cooling "sidekick" for Microsoft's Maia 100 marked the first use of direct-to-chip (DTC) liquid cooling units in existing air-cooled datacenters. Building on this innovation, we continue to evolve our liquid cooling technology to enhance cooling capacity for modern AI processors. At the 2024 Open Compute Project (OPC) Global Summit, we showcased our advanced liquid cooling heat exchanger units (HXUs), designed to support next-generation GPUs and AI accelerators from Microsoft and other industry leaders.

Our approach focuses on innovating the entire cloud system—from silicon to software—to deliver improved performance, greater choice, enhanced resiliency, and broader availability for our customers. By co-designing and co-optimizing hardware, software, networks, and datacenters, we tackle the critical challenges of performance, energy efficiency, and capital investment.

Through the OPC, we are open-sourcing our HXU designs, sharing our development work to help enable closed-loop liquid cooling across AI datacenters industry-wide. This collaborative effort ensures that innovations in thermal management not only benefit Microsoft but also drive progress for the entire computing ecosystem.

Liquid cooling heat exchanger unit

To support higher densities of AI accelerators per rack, we are utilizing standalone liquid-to-air heat exchanger units. These units enable legacy datacenters—traditionally unequipped for using direct-to-chip liquid cooling—to efficiently handle the demands of new and advanced AI workloads.

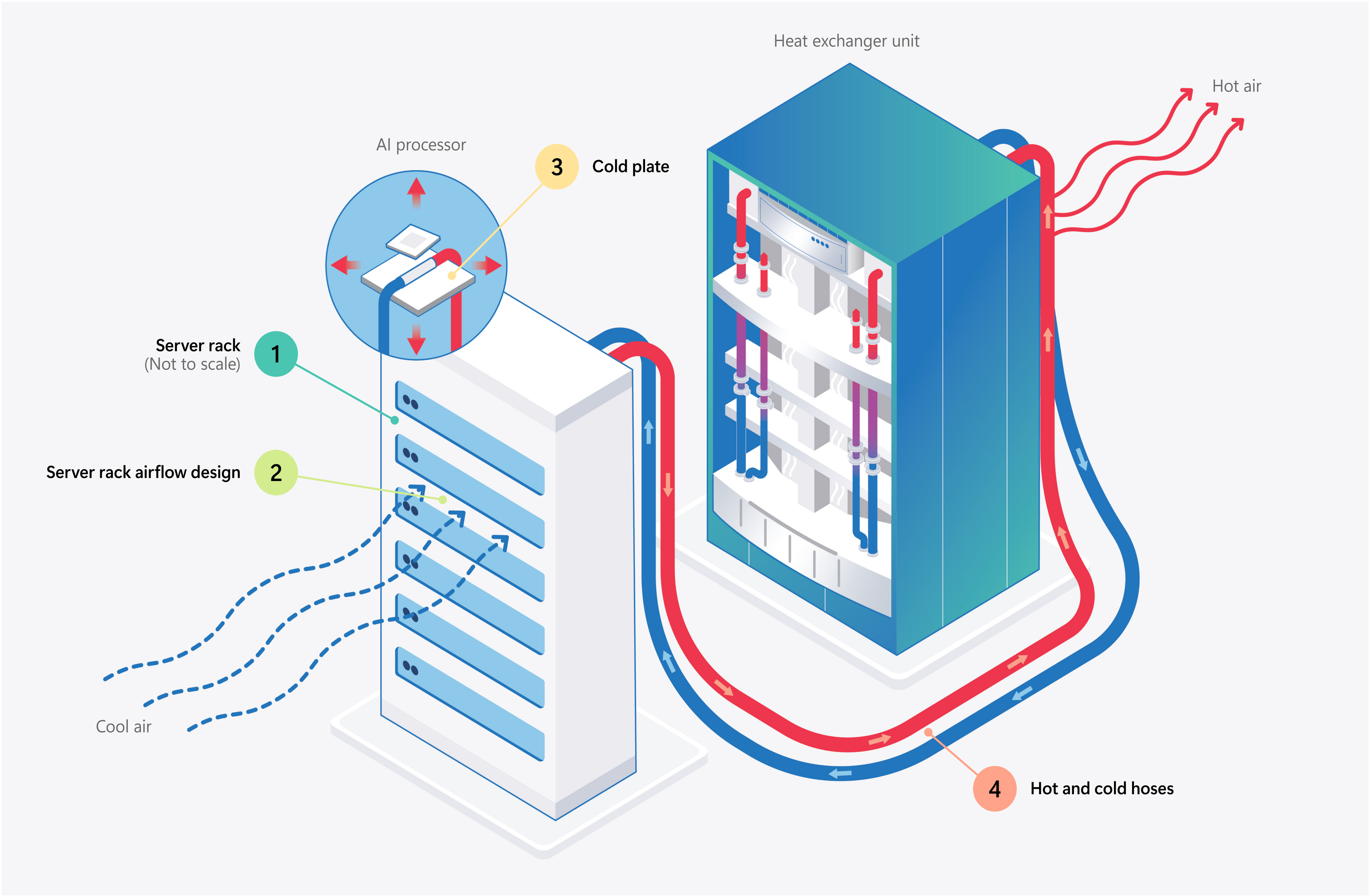

Figure 2: 3D illustrated depiction of air-to-liquid cooling steps

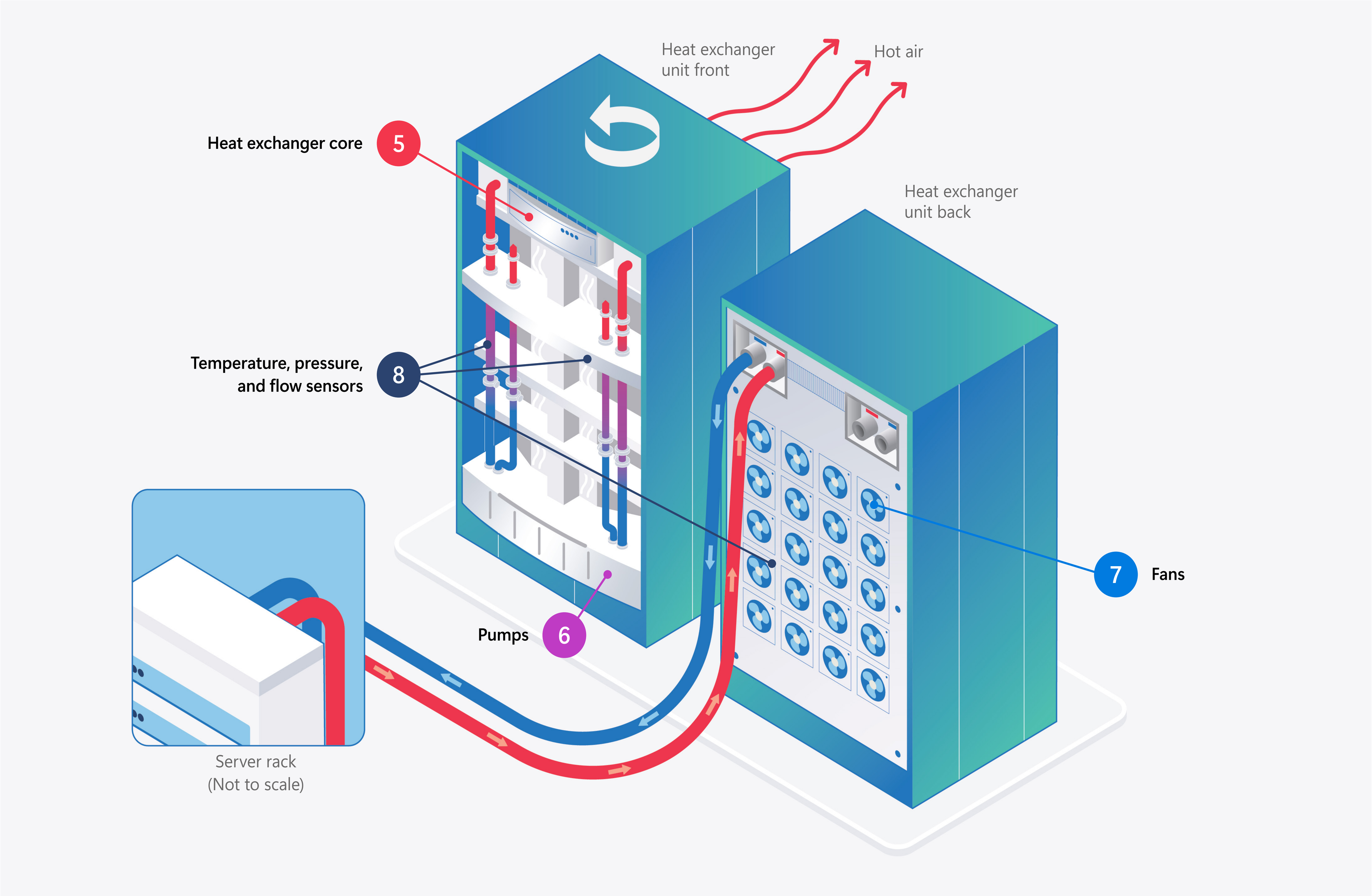

Figure 3: 3D illustrated depiction of an HXU

Components

1 **Server rack**
The process begins with single-wide racks designed to enhance heat dissipation capacity. These racks are securely packaged in leak-proof materials to ensure safe transport and installation, preserving the system's integrity before it's brought online.

2 **Server rack airflow design**
Cool air, usually supplied from underfloor vents or overhead ducts, is drawn into the front of the servers in the cold aisle. The servers' internal fans guide this cool air over their components, absorbing the heat generated by processors, memory, and other high-performance hardware.

3 **Direct-to-chip cooling and heat transfer**
Heat generated by high-performance chips is transferred directly to a liquid coolant via cold plates attached to the processors. The coolant effectively transfers the heat from the chip, bypassing reliance on traditional airflow for component-level cooling.

4 **Hot and cold hoses**
These hoses connect the cold plates in the server rack to the HXU and help maintain the flow of hot and cold liquid between the HXU and the server racks.

5 **Heat exchanger core**
The heat exchanger core is positioned to transfer heat effectively from the liquid coolant to the surrounding air. The liquid flows through the heat exchanger, where the heat is released and carried away by chilled air from the cold aisle. The liquid coolant is then recirculated, maintaining a continuous cooling loop.

6 **Pumps**
High-efficiency pumps are activated to control fluid flow precisely, adapting the cooling system to the specific thermal demands of each rack. The modular fluid connections between HXUs and server racks allow for flexible configurations, optimized to match the setup requirements.

7 **Fans**
Fans are placed at the rear of each HXU to maintain uniform airflow. They prevent preheating within the HXU and help maintain consistent cooling performance.

8 **Temperature, pressure, and flow sensors**
Finally, strategically placed sensors—including quick disconnects, leak detection ropes, and drip pans at the HXU base—are engaged to detect, contain, and manage any potential leaks, ensuring that temperatures remain precisely regulated for consistent performance.

? **Did you know?**
Our HXUs feature modular designs that enable field servicing of critical components such as pumps, fans, filters, printed circuit board assemblies, and sensors. This approach ensures reliable cooling while minimizing downtime, keeping operations running smoothly and efficiently.

Microsoft's sustainability goals

Microsoft designs and operates our datacenters to support our climate goals: to become carbon negative, water positive, and zero waste all by 2030, while enabling others to do the same.

Carbon negative

By 2030, we'll be carbon negative, and by 2050, we'll remove our historical emissions since our founding in 1975.

Water positive

By 2030, we'll replenish more water than we consume across our direct operations.

Zero waste

By 2030, we'll be zero waste across our direct waste footprint.

Protect and preserve ecosystems

We'll protect more land than we use by 2025 and build a Planetary Computer.

Sustainable by design

Our quest to innovate across every part of our cloud infrastructure to deliver more sustainable cloud services has led to many changes across how we design, build, and operate our datacenters.

Learn more through the Sustainable by design blog series