Bringing 2-Phase Immersion Cooling to Hyperscale Cloud

Ashish Raniwala
Azure, Microsoft, One Microsoft Way, Redmond, WA 98052, USA
Ashish.Raniwala@microsoft.com

Abstract: The advent of high-TDP chips is pushing cloud providers towards liquid cooling. Microsoft is advancing and deploying 2-phase immersion cooling in Azure. We discuss the trends, the challenges, and the opportunities in this space. © 2022 Ashish Raniwala

1. The Need for Liquid Cooling

Cloud providers typically use air-based solutions (e.g., chillers, water-side economized, direct evaporative cooling) for cooling servers, as the wide availability of expertise and equipment makes it easy to install, operate, and maintain such solutions. Unfortunately, air cooling has many downsides. Its low heat dissipation efficiency necessitates large heat sinks and fans and increases costs. Operating at higher component junction temperatures results in higher leakage power, which in turn negatively impacts energy efficiency. It may also degrade performance, due to hitting thermal limits resulting in reduced clock frequency. Most importantly, the trend of increasing transistor counts, coupled with the end of Dennard scaling, will result in chips with thermal design power (TDP) that is beyond the capabilities of air cooling in the near future [1]. For example, manufacturers expect to produce CPUs and GPUs capable of drawing more than 500W in just a few years.

2. Liquid Cooling in Datacenters

Liquid Cooling. To tackle rising chip temperatures, providers have started to explore liquid cooling solutions for their most power-hungry workloads [2, 3]. These technologies keep chip temperatures at a lower and narrower range than air cooling, reducing leakage power, eliminating the need for fans, and reducing datacenter Power Usage Effectiveness (PUE), i.e., the ratio of total power to IT (e.g., server, networking) power.

Cold Plates. The initial efforts typically placed cold plates on the most power-hungry components. Fluid flows through the plates and piping to remove the heat produced by those components. For example, Google cools its Tensor Processing Units (TPUs) with cold plates [2]. Although efficient, each cold plate needs to be specifically designed and manufactured for each new component, which increases engineering complexity and time to market. Moreover, cold plates remove localized heat from power-hungry components and typically still require air cooling for other components.

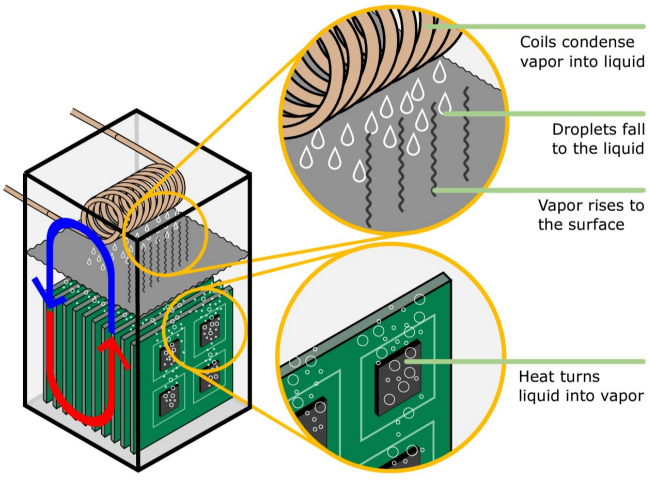

Fig. 1. Two-phase immersion cooling. IT is submerged into a dielectric liquid that changes phase (boils). Vapor rises to the top where it rejects heat and condenses back to liquid form. This process requires no additional energy. Figure rights belong to Allied Control Limited. Permission to modify has been granted to the author of this work.

Immersion cooling. An alternative to cold plates is immersion cooling, where entire servers are submerged in a tank and the heat is dissipated by direct contact with a dielectric liquid specifically engineered to transfer heat from electronics. There are no recurring engineering overheads. The heat removal can happen in a single-or two-phase manner. In single-phase immersion cooling (1PIC), the tank liquid absorbs the heat and circulates using pumps, whereas in two-phase immersion cooling (2PIC) a phase-change process from liquid to vapor (via boiling) carries the heat away. As Figure 1 shows, in 2PIC the vapor naturally rises to the top of the tank where a colder coil condenses it back to liquid. No liquid is lost and the heat transfers to the coil condenser secondary loop. The heat carried in the coil is finally rejected with a dry cooler (not shown). There are options for immersion fluid that provide different tradeoffs between properties such as boiling point, dielectric constant, reactivity to moisture and other contaminants.

Alibaba introduced 1PIC tanks in their datacenters and showed that it reduces the total power consumption by 36% and achieves a PUE of 1.07 [3]. The BitFury Group operates a 40+ MW facility that comprises 160 tanks and achieves a PUE of 1.02 [4] with 2PIC.

3. Two-phase Immersion Cooling in Microsoft Azure

Until recently the adoption of 2-phase immersion cooling had been limited to the crypto mining domain. At Microsoft, we are developing/deploying 2PIC for cooling general-purpose hardware in Azure. The first immersion tank we deployed in Azure hosts 36 Open Compute 2socket server-class blades [5]. Figure 2 shows three views of this prototype: (a) outside the tank; (b) a server being removed; and (c) all servers in place. These servers are standard air-cooled servers modified for immersion through modifications such as fan removal, replacement of CPU heatsinks with boilerplates, PSU removal, overall depth reduction, and some firmware modifications. The tank has power, cooling, and networking interfaces to the data center.

Fig. 2. Microsoft's first 2-phase immersion cooling production deployment.

Table 1. Value proposition of 2-phase immersion cooling

Lower TCO	Improve Sustainability	Enable Innovation
Simplified climate-agnostic rapidly-deployable infrastructure Reduce hosting costs via efficient cooling and smaller DCs Increase reliability by preventing oxidation/contaminants	Enable smaller land parcels DC Reduce GHG emissions and Energy per VM Enable water-less cooling Enable energy recovery	Enable higher-performance and higher-power CPUs (>500W), FPGAs, GPGPUs and ASICs Enable stable operating temperatures for components Simplify and improve interconnect Enable physical disaggregation

2PIC can provide significant value by lowering TCO and enabling innovation while providing more sustainable DC operations. Table 1 lists these potential benefits. To realize these benefits, we are refining the core technology as we scale the deployments – cost optimizing the tanks, finding the right immersion fluids, redesigning the server blades to leverge the high heat removal, and optimizing networking gear. In the rest of the abstract, we focus on networking-specific aspects of immersion.

4. Networking in Immersion: Challenges and Opportunities

There are a few challenges with networking in an immersion environment:

1. **Getting the network in/out of the tank**: During normal operations, the immersion tanks are kept sealed to minimize vapor escape and keep out any excess humidity. This brings up the basic problem of getting a high-capacity network in and out of the tank while minimizing tank perforations. To that end, we put the TOR (top of rack) switch inside the tank so the DAC cables connecting servers to TOR can be routed internally. The TOR itself needed to be modified slightly to not run fan and went through a material compatibility study. The connection from TOR to MOR (middle of row) is terminated on the tank wall using an array of IP68-rated connectors. A second cable is then used to patch the other side of the connector to the MOR.
2. **Material compatibility:** Certain materials such as PVC present in QSFP and RJ45 cables dissolve in the immersion fluid and precipitate by distillation thereby affecting thermal performance. Although the immersion tank employs filtration for such contaminants, it is crucial to eliminate these sources of contamination as much as possible, e.g., replacing PVC cables with LSZH (Low Smoke Zero Halogen) cables.
3. **Optical path in immersion:** Fluid has a different refractive index than air. This may necessitate design changes in some cases, e.g., a transceiver with an air gap that is now filled with fluid may not work in immersion unless the air gap is sealed with epoxy.
4. **Electrical traces in immersion:** Fluid also has different dielectric constant (D_k) than air (D_k=1). This typically does not pose a problem, but electrical traces for networking components usually operate at much higher data rates (56G PAM4) compared to rest of the servers (< 8Gbps). This necessitates use of a lower dielectric fluid (< 2.5) eliminating some of the otherwise promising immersion fluids.

Immersion also opens up several opportunities for simplifying design, improving performance, and reducing cost of network components such as chips, switches and transceivers:

1. **Thermally unconstrainted environment**: Due to extremely high volumetric heat capacity [6], 2-phase immersion environment can support high thermal design power (TDP, Watts) and heat flux (W/cm^2) components.
2. **Narrow operating temperature**: Unlike air, immersion provides a narrow operating temperature range around the boiling point of the immersion fluid.
3. **No oxidation and contamination:** Immersion also provides an inert operating environment free of oxidation and contamination.

6. Acknowledgements

I would like to thank my colleague Rich Baca for helping significantly with the networking specifics. I would also like to thank all my colleagues at Microsoft, 3M, and Wiwynn, who are collaborating on this work.

7. References

[1] Y. Sun, N. B. Agostini, S. Dong, and D. Kaeli, "Summarizing cpu and gpu design trends with product data," 2019, arXiv
[2] T. Norrie, N. Patil, D. H. Yoon, G. Kurian, S. Li, J. Laudon, C. Young, N. P. Jouppi, and D. Patterson, "Google's training chips revealed: TPUv2 and TPUv3," in Proceedings of the Hot Chips Symposium, 2020
[3] Y. Zhong, "A Large Scale Deployment Experience Using Immersion Cooling in Datacenters." Alibaba Group: OCP Summit, 2019.
[4] 3M, "Two-Phase Immersion Cooling: A revolution in data center efficiency." Technical Report, 2015.
[5] M. Jalili, I. Manousakis, I. Goiri, P. Misra, A. Raniwala, H. Alissa, B. Ramakrishnan, P. Tuma, C. Belady, M. Fontoura, R. Bianchini, "Cost-Efficient Overclocking in Immersion-Cooled Datacenters," in Proceedings of the Intl Symposium on Computer Architecture (ISCA) 2021.
[6] B. Ramakrishnan , H. Alissa , I. Manousakis, R. Lankston, R. Bianchini, W. Kim, R. Baca, P. Misra, I. Goiri, M. Jalili, A. Raniwala, B. Warrier, M. Monroe, C. Belady, M. Shaw, M. Fontoura, "CPU Overclocking: A Performance Assessment of Air, Cold Plates, and Two-Phase Immersion Cooling," in IEEE Transactions on Components, Packaging, and Manufacturing Technology, Vol. 11, No. 10, Oct 2021.