Approaching the Heat Limit with Liquid Immersion Technology

White Paper 1

Jimil M. Shah, Ph.D.
Robert Lipscomb, Ph.D.

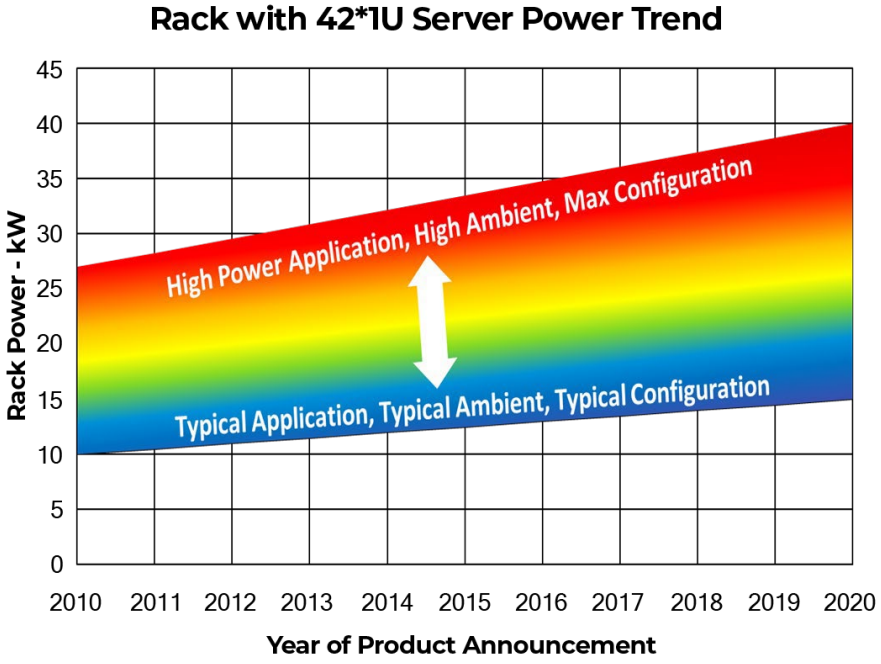
EXECUTIVE SUMMARY

The transference of heat away from processors persists as one of the most critical challenges to computing technology today. A variety of approaches have been employed to push against this “heat limit” as it remains a principal barrier to better and faster processing technologies that are otherwise prepared to realize greater and greater potential. This paper reviews the technology currently in use. It addresses the critical advances in the field of air cooling both at the server level and extending to the architectural and structural geometry found within data centers. While air cooling remains the most prevalent method of reducing processing temperatures, more recent advances have incorporated liquid cooling technologies. The balance of this paper considers the various opportunities and limitations of liquid immersion cooling in terms of heat transfer modes, convection modes, containment methods, and fluid chemistry. After a review of these technologies, it becomes apparent that the most significant approach to advancing the heat limit is found in liquid immersion cooling.

INTRODUCTION

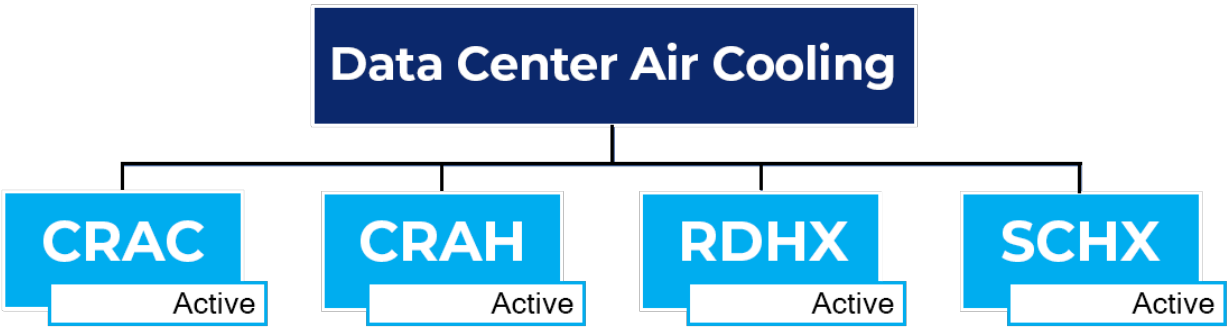
The Heat Limit and Today’s Datacenter Challenges

A key concept to help navigate the current field of knowledge concerning liquid cooling within data center environments can be distilled into one phrase: heat limit. The most significant advances in computing power, capacity, efficiency, and density are constantly confronted by the material issue of mitigating the heat produced from processors. From servers to data centers to on-the-ground applications, computing equipment is always constrained by cooling capacity.



The most widespread technologies employed to cool processors involve forcing and extracting cooled air across thermal exchangers. Most servers in use today are designed to accommodate this strategy. As will be discussed in the next section, many data centers leverage an economies-of-scale-type approach by incorporating directional airflow into facility design, construction, and equipment. However, and regardless of its virtues, the very properties of this approach can only provide so much cooling while remaining cost-effective as the issue of heat and the ancillary role of dispersing heat from increasingly efficient processors consistently proves to be a limiting issue. Until new techniques are developed, refined, and implemented to cool processing equipment, real advances in computing technology will remain limited due to the persistent constraints of the heat limit.

AIR and AIR-HYBRID COOLING



CRAC	Computer Room Air Conditioner
CRAH	Computer Room Air Handler
RDHX	Rear Door Heat Exchanger
SCHX	Side Car Heat Exchanger

Fig 2: Approaches to Air Cooling

Air Cooling

Without question, air cooling provided a reliable, calculable, and relatively stable method for addressing the heat limit. Servers were (and are) commonly designed so that cooling air is driven through each unit to remove the heat from the processors. Design considerations demanded the addition of fans and blowers, which also required open pathways within servers so that air traveled effectively across the heatsinks and other such surfaces.

Almost immediately, data centers began leveraging a collective approach to server deployment. For example, many data centers forced cold air from beneath raised floors and into aisles where directionally stacked servers equipped with onboard fans pulled the cold air through the servers into spaces where the heat collectively rose and was captured by air handlers above. Among many, one key issue with the stacked-server approach was the fact that the servers toward the bottom of the stack had more access to cooler air than those toward the top. Some data centers eventually employed a “chicken coup” or silo design where the fronts of the servers were enclosed to better

retain the cold air, thus reducing the amount of cool air that was being lost by the entropic mixing with nearby warmer air. Free air cooling allowed for compressors to eventually be removed from the equation, though bulky air-handling equipment remained. Though many data centers modified configurations and improved efficiencies, the need to cool equipment by pushing (or pulling) chilled air persisted.

Regardless of configuration alterations and irrespective of modifications related to scaling, addressing the heat limit with forced air cooling remains unable to effectively address three key issues. For one, power density is constrained by, among other things, the space required for sufficient passage of air both within the servers themselves as well as the larger data center environments. Second, as power density increases, the degradation of energy efficiency emerges as an automatic consequence. Finally, in addition to increased energy expenditure, the cost of the required infrastructure is also quite high.

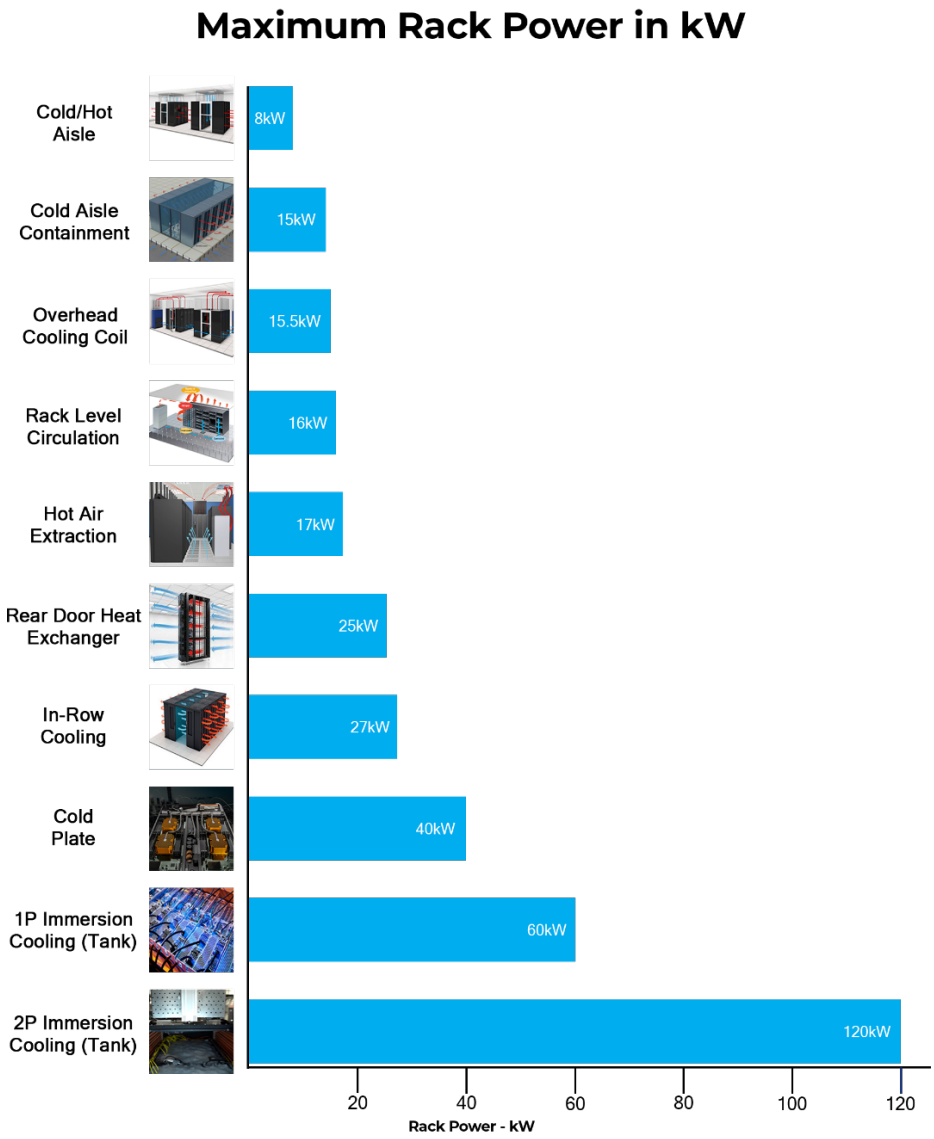


Fig. 3: Maximum Rack/Tank Power (KW) currently deployed by the real-world data centers using differentfacility-level cooling approaches

The Air-Liquid Hybrid

Another cooling intervention involved installing heat exchangers to the rear doors of the racks. These water-cooled exchangers, which combined the effectiveness of individual chassis fans to rack-mounted supplemental blowers, virtually eliminated the need for non-rack air handlers. This approach was further enhanced by the addition of supplemental heat exchangers situated amongst the tiers of servers. The use of these rear-door heat exchangers, or intercoolers, to capture heat from air-cooled hardware was limited in terms of efficiency and density as they required fans, blowers, heatsinks, and airflow paths.

A similar approach incorporated two-phase cooling by removing heatsinks and replacing them with spray modules that atomized the dielectric fluid, thereby cooling the CPU through evaporation. The vapor then flowed back through an apparatus to a cooling unit, where the heat transferred to facility water with the condensed dielectric fluid available for use again. This hybrid approach did not allow for the elimination of cooling components such as chassis fans, space that otherwise could have been utilized to support additional memory. In addressing the heat limit, these approaches provided only limited solutions.

THE LIQUID REVOLUTION

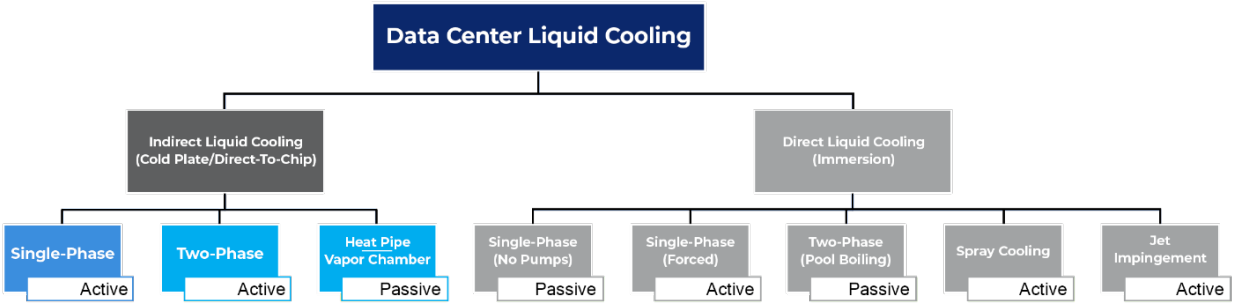


Fig. 4: Approaches to Liquid Cooling

Cooling with Liquid

To be clear, there is nothing new about a fluidic approach to transitioning heat away from processing equipment, but advances in both understanding and deployment have informed new directions for pushing the heat limit. As with all rapidly advancing technologies, approaches to liquid cooling come with both virtues and limitations.

Unlike the demands of air cooling, which require space and myriad apparatus to control flows and currents, liquid cooling, though less costly, nevertheless requires some consideration regarding the issue of containment. Indirect liquid cooling includes a cold plate or direct-to-chip approach in which water or refrigerant is pumped onto either each node or each server and through individual cold plates to remove the heat. In some configurations, cold plates are in contact with but not connected to servers and transfer heat to interfaces via heat pipes, vapor chambers, etc. Such technologies include complex cold plate assemblies, clamshells, hermetic connectors, manifolds, pumps, and braze joints. There is always the risk of water damage due to leakage via quick disconnects and joints.

The direct-to-chip method relies on cold plates to convey heat to water pumped directly into the servers. Considerations of apparatus and connection are further complicated by the fact that the water must then be conveyed to a separate location to be cooled. This method requires that water be connected to each server individually and is a consistent approach for any cooling method where the entire server is not immersion ready. The thermal interface required for the cold plate to function properly impacts thermal performance. Moreover, in single-phase designs, the fluid glide has a higher impact on thermal performance. Cold plates also require that the fluid be transferred to a cooling apparatus, or coolant distribution unit, with connections and pipes (or tubes) that function in an otherwise dry area. Direct-to-chip or clamshell-style (in which the entire server is hermetically sealed to accommodate the fluid) comes with not-insignificant maintenance needs due to numerous potential points of failure.

Perhaps the most proactive method of addressing technologies deployed in the current field of liquid immersion technology is by recognizing a matrix of overlapping and often-interconnected practices. Indeed, a variety of considerations, configurations, and classifications comprise the current deployment of liquid cooling technologies. Several factors including the increased options for the physical configuration of equipment and the type of liquid in use emerge as key considerations. Further, when it comes to both the physical configuration of equipment and the type of liquid that will be employed, how heat is being convected away from the equipment remains the most important consideration – and limitation. Table 1 provides an evaluation of traditional liquid cooling approaches from the perspective of efficiency, density, and mechanical design simplicity.

Technology	Efficiency	Density	Practicality
Use intercoolers or rear-door heat exchangers to capture all heat from air-cooled hardware.	a, d2, g	a, g	
Pump water or refrigerant onto each node and through individual cold plates on major heat generating devices.	(d2), (e)	b	b, c, (f)
Pump water or refrigerant onto each server and through its dedicated cold plates that mates to all devices.	d2, (e)		b, c, f
Cold plates in contact with but not connected to server. Transfer heat to interface via heat pipes, etc.	d3, (e)	g, b	
Pump dielectric liquid onto each server/node enclosed in direct immersion cooling clamshell.	d2, e, g	b, g	b, c, (h), j
Passive 2-phase immersion in pressure vessel.			b, j
Pump dielectric liquid through bath in which servers are immersed.	d2, e	g	(h), (j)

Shortcomings () = depends on fluid or thermal interface technology

a. Fans, blowers, and airflow paths

b. Complex cold plate assemblies, clamshells, hermetic connectors, manifolds, pumps, braze joints, etc.

c. QD, clamshell, or electrical via leakage risks water damage in case of water and refrigerant loss

d. Secondary (2) or ternary (3) thermal interface(s) impact thermal performance

e. Fluid glide impacts thermal performance (single-phase)

f. Cost and global warming emissions from fluid loss at intractable sites (refrigerant)

g. Need extended surfaces on heat generating devices

h. Fire risk due to combustible coolant

i. Performance limited by inability to use best boiling enhancement technologies

j. Difficult to open and service clamshell/pressure vessel

Table 1: Evaluation of Traditional Liquid Cooling Approaches

Immersion Cooling

Submerging servers in a dielectric liquid allows for significant energy savings today while also accommodating future load densities. The effectiveness and energy savings for new construction and even retrofits at the facility level have been demonstrated by adapting to submersion cooling in numerous case studies. Among the many conclusions is the fact that liquid cooling reduces the root causes of many problems and improves operating conditions and dependability all while advancing cooling technology. In other words, it is already proven to be a successful approach to pushing the heat limit.

Immersion cooling technology within data centers extends the prospects for improved reliability in operations as it minimizes common issues and eliminates the root causes of failure such as solder joint failures. It also allows for lower operating temperatures for boards and components, eliminates oxidation and corrosion of electrical contacts, removes certain moving parts such as fans within device enclosures, mitigates exposure to electrostatic discharge, and vitiates sensitivity to ambient particulate, humidity, or temperature conditions. The advances in reliability include reducing corrosion and electrochemical migration, lessening environmental contamination like dust, debris, and particulates, and mitigating tin and zinc whiskers.
Unlike indirect cooling that can only be used on those components where cooling distribution units are attached, immersion cooling also allows removing heat directly from the chip(s) and all other components with no intervening thermal conduction resistance (other than what is needed between the device heat sources and the chip surfaces in contact with the liquid).

It is best to consider four overarching categories when discussing liquid immersion cooling: the mode of heat transfer, the convection mode, the containment method, and the chemistry of the various fluids. That said, a selection in one category does not necessarily lock the possibilities in another. For example, when configuring equipment to either use single- or two-phase cooling for heat transfer, either a passive or forced convection mode can be employed. This paper provides a review of the key considerations and deployments of immersion cooling to establish the groundwork for future analyses regarding fluidic innovations being made to press against the heat limit.

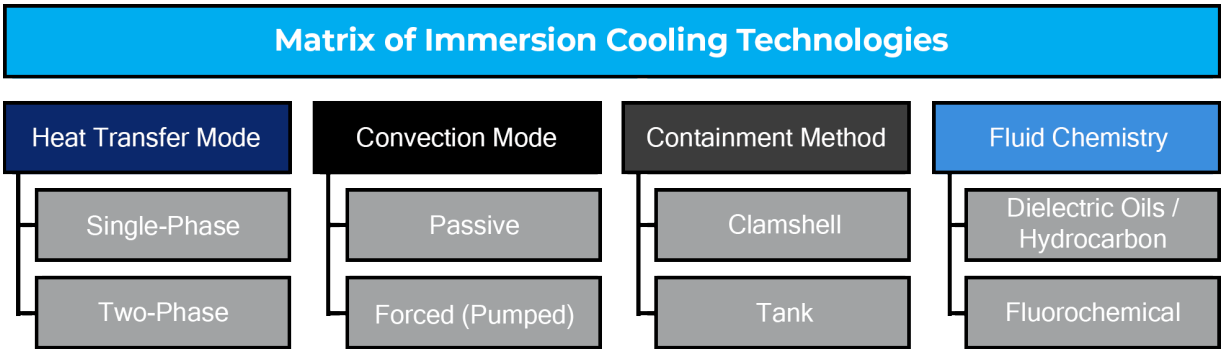


Fig. 5: Matrix of Immersion Cooling Technologies

LIQUID COOLING CAPACITY and POTENTIAL

Heat Transfer Modes: Single-Phase vs. Two-Phase

Two general categories determine the heat transfer mode of liquid immersion cooling: single-phase and two-phase. In technology that employs single-phase cooling, computing equipment is placed in a fluidic environment that remains in a liquid state throughout the cooling process. In single-phase immersion cooling, the fluid does not change state from liquid to vapor. Single-phase immersion cooling can use either natural convection or forced convection. The forced convection approach requires that a pump move the fluid inside the tank. As a consequence, the pump is a function of tank power and fluid temperature rise. For single-phase immersion, as with air cooling, engineering and power requirements increase with processing density as managing flow becomes more challenging. Eventually, a limit is reached.

Single-phase immersion cooling can use high boiling point hydrocarbons. Many providers make blends of mineral oil and other synthetic hydrocarbons such as Polyalphaolefin and Gas-To-Liquid (GTL) as well as natural and synthetic esters. They are dielectric and have been used in transformers for decades. As these fluids or blends tend to be oleaginous, it can be difficult to service the server hardware inside the tank. Moreover, they can cause slippery surfaces, which may introduce a hazard to the work environment.

Single-phase cooling offers advantages for today's processing capacity needs.

When it comes to pushing the heat limit, two-phase cooling addresses the data center needs of the future.

In two-phase immersion cooling, fluid boils and condenses; thereby, changing its state of matter from liquid to vapor and back again. It is a passive process and does not require pumps. The dielectric fluids used are Perfluorocarbons (PFC), Perfluoropolyether (PFPE), Hydrofluoroeter (HFE), and Fluoroketone (FK). These fluidshave high dielectric strength so that they can be in contact with a larger amount of electronics. Two-phase fluids are non-combustible, non-flammable, and non-toxic. In fact, Perfluorocarbons have been used in electronic cooling for more than 40 years.

Without question, single-phase liquid cooling offers significant advantages for today's processing capacity needs. For one, it offers greater cooling capacity over air. Yet when it comes to pushing the heat limit, two-phase cooling addresses the data center needs of the future as it accommodates configurations that can operate with increased power density and efficiency. Considering the potential afforded by two-phase cooling, the capability for as of yet unrealized extreme power densities is remarkable.

Much higher heat transfer coefficients can be achieved through two-phase (evaporation and condensation) than single-phase. Two-phase immersion cooling with fluorochemicals allows for energy efficiencies over forced convection, water, oil, and air cooling by degrees of magnitude. Experiments with two-phase cooling show that 100 times the density of a typical server cooling can be achieved. Ultimately, extremely

high densities with extremely high efficiencies can be accomplished using two-phase liquid immersion technologies.

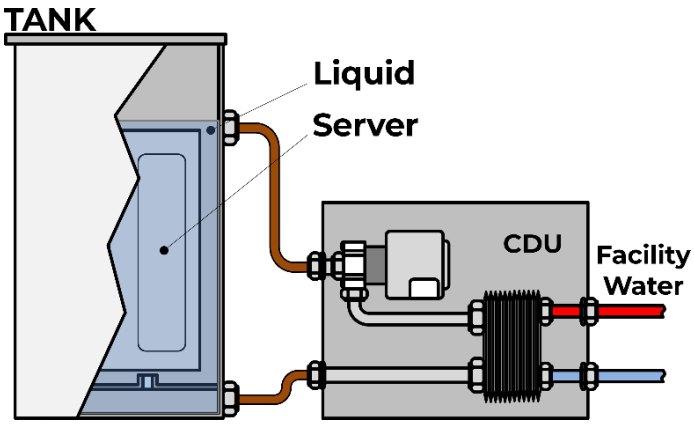


Fig. 6: Single-Phase Immersion Cooling

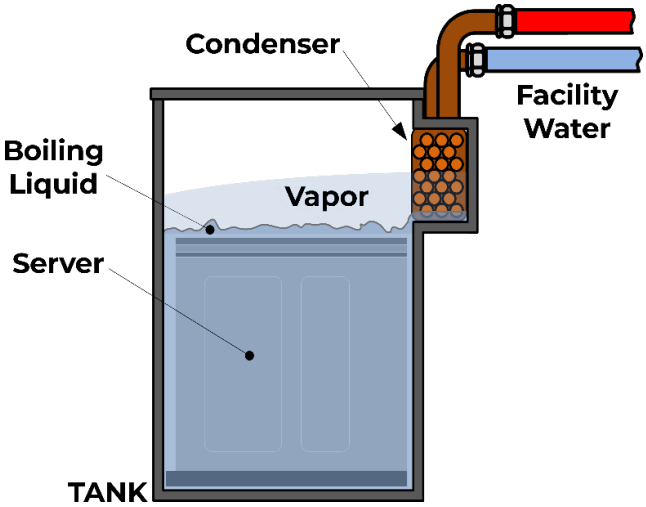


Fig. 7: Two-Phase Immersion Cooling

Convection Modes: Passive vs. Forced

Though intrinsically more efficient than air cooling, the liquid environment also works by convecting heat away from processing components. In the liquid environment, the convection mode is either passive or forced. In a passive system, the transference and dispersal of heat rely upon naturally occurring convection currents and gravity. In single-phase cooling, a passive system requires space and an increased volume of fluid to accommodate the dispersal of heat. A forced convection mode relies on pumps to convey heated fluid or vapor into a heat exchanger. Both the passive and forced convection modes can be used in either single- or two-phase fluidic environments.

Regardless of the convection mode, the use of liquid as a cooling agent comes with new considerations. In a single-phase system, forced cooling is common and requires that the fluid being pumped from the tank and into heat exchangers. In a two-phase system, condensers aid the natural cooling and conversion of vapor back into liquid.

Containment Methods

As referenced above, in the clamshell configuration, the entire server chassis is enclosed, and the fluid is forced into one location on the server and then out another. This method does eliminate some of the complications associated with applying intricate piping and tubing into and out of each server as the server itself becomes a sealed vessel. That said, the modifications to the server casing are complex – potential points of failure, such as connection locations, must be actively monitored. From direct-to-chip to clamshells (and designs in between), all containment methods that do not involve immersing entire servers require complex designs that are inherently difficult to maintain by virtue of the very fact that there are so many potential points of failure and leakage.

One containment method proven to provide a stable containment environment is also the most common in use today: the tank. As the name implies, the tank is a larger apparatus into which servers are submerged. Tanks, irrespective of their design to accommodate either single- or two-phase cooling, typically require connection to the facility. However, a stable connection location and a reduced need to attach and detach the water connection lessens the risk of leaks being created operationally. That said, servers require minimal modification as compared to the apparatus required for cold plate cooling.

Additionally, the implementation of two-phase tanks creates opportunities to reconfigure sites. Though a sealed vessel is required to vitiate potential fluid loss from the two-phase process, the elimination of either the volume of fluid or the pumps needed for the single-phase process means that not only is the fundamental design simpler, but that simplicity also allows for greater freedom when it comes to design possibilities. As stated, due to the greater efficiency at which heat disperses into a fluidic environment, the amount of space required for equipment can be significantly reduced. Further, the latency of connections from the processors to the points of transference can also be reduced. These two factors allow the processors, which function in a diminished capacity due to the confined thermodynamic availabilities in air-cooled units, to be turbo boosted as the properties of the liquid environment allow for overclocking. The lowering of cost over time, a factor already present in single-phase cooling, is further decreased in the two-phase process. Clearly and despite an increase in certain considerations – especially in terms of fluid hygiene and the need to control vapor loss – two-phase immersion cooling counterintuitively allows for a more elegant process arising from less intricate designs.

Such compositional and configurational considerations certainly extend to overall data center architecture as well. For one, the imperative to account for massive air handling equipment combined with the need for servers to be situated so that they can collectively take advantage of the larger processes means that the rack-stacked placement of servers can be reduced, if not eliminated. Not only will less space be required in data centers for the same amount of processing, but new and heretofore unrealized possibilities in data center design will be identified, utilized, and maximized. Further, it is possible that currently unused spaces in older structures, once considered too small for data centers, will be reclaimed.

Fluid Chemistry: Hydrocarbons vs. Fluorochemicals

When considering commercially significant heat transfer fluids, there is a variety that cannot be used for immersion cooling. Aqueous coolants, for example, tend to be very poor dielectrics and are usually corrosive. Many flammable materials cannot be used. Chlorinated and brominated organics tend to be Ozone-depleting and many of them have a very toxic profile. However, in hydrocarbon fluids, the hydrocarbon structure has important implications. One way of mitigating some of those effects is through the addition of fluorine, which produces fluorochemicals and which tend to be more inert and stable, though considerably more expensive.

The fluids that are put to use in liquid immersion cooling are not new. It is decades-old technology that has evolved both in terms of its molecular structure and in use. Dielectric Oils, also known as hydrocarbons, represent the early generations of this technology and are themselves better suited to the single-phase process because they have a higher boiling temperature. Therefore, the heat is contained within one phase of matter, the liquid phase. The hydrocarbon fluids that are used for immersion cooling are usually characterized as oils and the reason for this is straightforward. If one considers the alkanes as the representative of the class, one must go to carbon numbers of ten or more before the hydrocarbon is classified as combustible and therefore safe for immersion. Proceeding from ten, one eventually encounters oils that are simply too viscous to be pumpable and therefore not useful for immersion. The low vapor pressure of these oils means that they are much easier to contain but also means that the hardware is oily when it comes out of the tank, which tends to be a little messier.

Fluorochemicals with higher boiling points can also be used for single-phase immersion cooling. It is beneficial to use fluorochemicals for single-phase immersion cooling because of their low viscosity and long-term stability.

Generally speaking, with a lower boiling point predominantly used for two-phase immersion cooling, fluorochemical fluids present a significantly higher heat transfer efficiency as compared to single-phase fluids using convection heat transfer modes. In two-phase immersion cooling, heat is removed by boiling the fluid which yields much lower and uniform component temperatures and enables operation at higher heat densities. Flourochemical fluids are typically colorless, odorless, non-oil-based, non-flammable, non-combustible, and non-corrosive. Further typical attributes include wide operating temperature ranges, low toxicity, outstanding thermal/chemical stability, and exceptional dielectric properties.

CONCLUSION

As will be discussed at length in a future paper, the most important consideration for transitioning to liquid cooling is the reduced environmental impact, which becomes apparent almost immediately. Immersion cooling liquids are effective, efficient, and environmentally friendly, and non-flammable. No pumps and jets are required to keep hardware cool.

the natural process of evaporation and without spending any extra energy. It is this simplicity that eliminates conventional cooling hardware and results in better cooling efficiency. Compared to traditional air or direct-to-chip cooling, this passive process results in the use of much less energy.

Additionally, the improved consistency and reliability manifest myriad savings opportunities; the decrease in temperature variation, elimination of bulky chassis fans, and the elimination of vibration reduce both short-term performance degradation and long-term wear and tear of components. The elimination of fans and other considerations required for air cooling opens physical space to more densely pack computing-focused equipment. Finally, such considerations translate into the workspace, allowing liquid immersion cooling equipment to be maintained in tighter, remote, and even extreme locations.

REFERENCES

1. Tuma, P., & Shah, J. M., 2020, “Recent Developments in Immersion Cooling with Fluorochemical Fluids by Open Compute Community,” May 2020, Open Compute Project Virtual Summit.
2. Day, T., Lin, P., Buger, R., 2019, “Liquid Cooling Technologies for Data Centers and Edge Applications,” Schneider Electric White Paper No. 265. https://download.schneider-electric.com/files?p_Doc_Ref=SPD_VAVR-AQKM3N_EN
3. Wen, Fangzhi, 2018, “Best Practice of Alibaba Datacenter-Immersion Cooling Escorts
Cloud Computing,” Presented at OCP Global Summit 2018, March 20-21, San Jose, CA, https://www.opencompute.org/files/Immersion-Cooling-for-Green-Computing-V1.0.pdf.