Proceedings of the ASME 2022 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems
InterPACK2022
October 24-28, 2022, Anaheim, California

DRAFT-IPACK2022-97490

SINGLE-PHASE IMMERSION COOLING STUDY OF A HIGH-DENSITY STORAGE SYSTEM

Saket Karajgikar
Meta
Fremont, CA

Jasper Kidger
Iceotope
Sheffield, UK

Andrew Shaw
Iceotope
Sheffield, UK

Neil Edmunds
Iceotope
Sheffield, UK

Veerendra Mulay
Meta
Fremont, CA

ABSTRACT

In this paper, a standard air-cooled high-density storage system is reengineered to demonstrate the use of single-phase immersion cooling. The storage system consists primarily of seventy-two hard drives, two single socket nodes, two SAS expander cards, NIC and a power distribution board in a 4OU form factor. It is successfully demonstrated that the storage systems can be designed to support single phase immersion cooling while supporting hot swap and cooling redundancy requirement like an air-cooled system. In an air-cooled system, temperature gradient between the drives was as high as 19°C. The drives placed at the front of the system received cooler air while the drives placed in the rear received preheated air thus resulting in a temperature gradient. The drives used for the study were 20GB Helium filled sealed drives. For immersion cooling, seventy-two drives were cooled in parallel with temperature variance of less than 3°C. The other system components such as CPU, DIMMs, SAS chip and NIC had sufficient thermal margin. It was demonstrated that the system can operate reliably for facility coolant supply temperature as high as 40°C. The resulting power consumption of the pump was less than five percent of the total IT power. In addition, the proposed cooling solution may help mitigate acoustic vibrational issues for drives often encountered in air-cooling solution. The solution is virtually silent in operation.

Keywords: High density storage, Single-phase immersion cooling,

NOMENCLATURE

HDD	Hard Drive Disk
OCP	Open Compute Project
NIC	Network Interface Controller
BMC	Baseboard Management Controller
ASIC	Application Specific Integrated Circuits
TDP	Thermal Design Power
SAS	Serial Attached SCSI
SCSI	Small Computer System Interface

1. INTRODUCTION

Data center industry has witnessed continuous growth in high power components, primarily CPUs and GPUs. To support high performance, a multi-socket server in a limited U-space is often required. This has pushed the industry to evaluate or explore thermal management techniques beyond traditional air-cooling approach. Current air-cooled heatsink solution can approximately support 450W while requiring a large keep out volume over the high-powered ASIC [1]. Fig. 1 shows the approximate transition from air to liquid cooling based on the socket power along with ASHRAE operating temperature envelope [2].

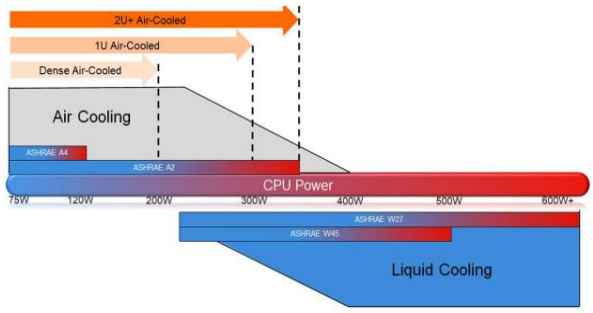
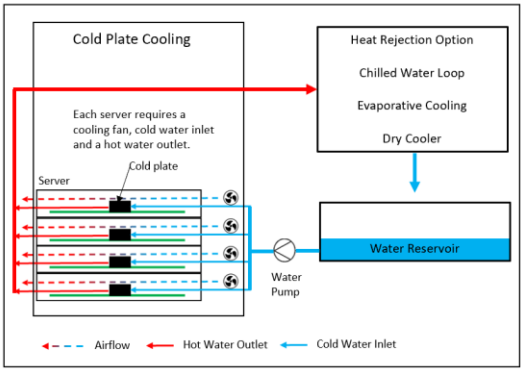


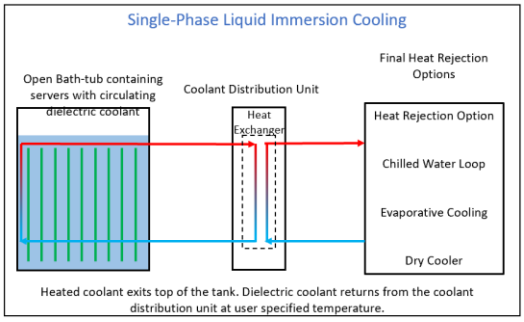
FIGURE 1: TRANSITION OF AIR COOLING TO LIQUID COOLING BASED ON SOCKET POWER [2].

Liquid cooling can counter the limitations of air-cooling and ensure reliable operation of the high-powered ASICs by operating below the thermal threshold. Two types of liquid cooling solution are generally considered – direct to chip aka cold plate and immersion cooling. In a hybrid cooling set-up, a cold plate is attached to the high-powered ASIC while rest of the components are air cooled. This requires system to be designed and optimized for both air and liquid which increases the server design complexity. In this case, the coolant is not in direct contact with the IT components. In immersion cooling, the entire server is dipped into a dielectric medium – thus all the IT components are in direct contact with the dielectric fluid which

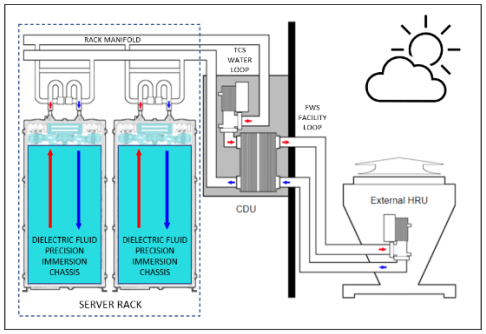
is cooled by facility water loop via a liquid-to-liquid heat exchanger. Immersion cooling can further be classified as single-phase immersion cooling and two-phase immersion cooling. In direct-to-chip cooling, about 60-80% of the server heat is removed via liquid while in immersion cooling, more than 95% of the heat is removed via liquid [3]. Figures 2a and 2b shows a schematic representation of both technologies.



(a) Cold plate cooling [4]



(b) Single-Phase Immersion cooling [4]



(c) Iceotope' s Single-Phase Precision Cooling Approach [5]

FIGURE 2: SCHEMATIC REPRESENTATION OF COLD PLATE AND IMMERSION COOLING TECHNOLOGIES.

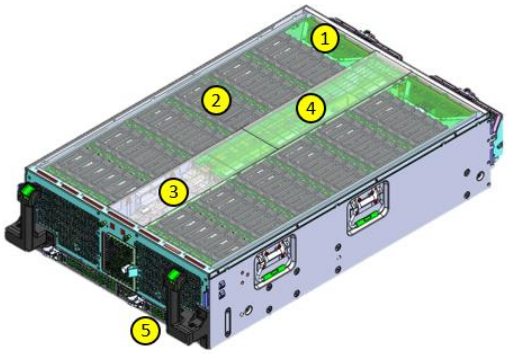
Another approach is Iceotope' s "Precision Cooling" approach. Each individual chassis has a dedicate dielectric loop connected to a liquid-to-liquid heat exchanger and a pump (fig. 2c). With manifolds to channel the dielectric, coolant is precisely pumped to individual IT components. The warm dielectric is then collected at the bottom of the chassis, referred as sump and circulated back via a liquid-to-liquid heat exchanger. Over here, the dielectric is cooled by the facility water loop. Benefit of this patented technology is that it minimizes the amount of dielectric required which reduces the overall volume of the dielectric required and hence the cost. In addition, this also ensures a good ΔT across the dielectric loop.

Several studies have demonstrated successful use of liquid cooling. There has been an increased interest in immersion cooling as it is believed to offer several benefits such as broad temperature support, high heat capture, high density, and flexible deployment options [1]. While discussing benefits and challenges of each technology is outside the scope of this study, in general, compared to air, single-phase immersion cooling can provide 1120-1400 times greater heat capacity by volume [5].

As mentioned earlier, driving factor for utilizing advance cooling solutions is to support high density servers utilizing high-TDP (Thermal Design Power) components. To the best of the authors knowledge, there is no published study of using single-phase immersion cooling for a high-density storage system. This could be attributed to the fact that traditionally hard drives were not hermetically sealed. Exposure of disk platters to any liquid would make the hard drive non-functional. It was only after the introduction of Helium filled, hermetically sealed drives, immersion cooling could be adopted. It was not a common practice until recently where the need for higher storage capacity drives has made the Helium filled drives mainstream.

In this study, a commercially available storage system, Bryce Canyon [6] designed for air cooling was modified to demonstrate compatibility and benefits of single-phase, precision based, immersion cooling for storage systems. Helium filled and hermetically sealed hard drives were used for the study. Temperature variation among all the hard drives was observed and was compared to the air-cooled solution. In addition, cooling (pump) power was also measured.

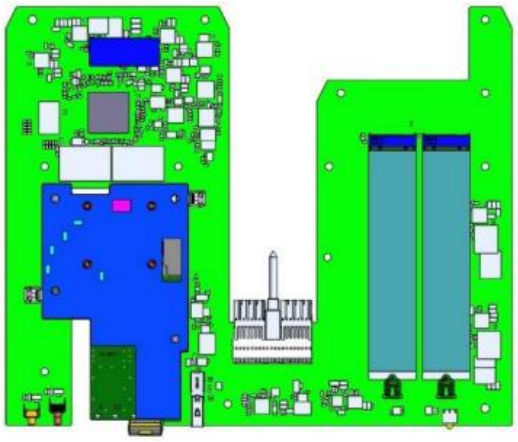
2. AIR-COOLED SYSTEM CONFIGURATION



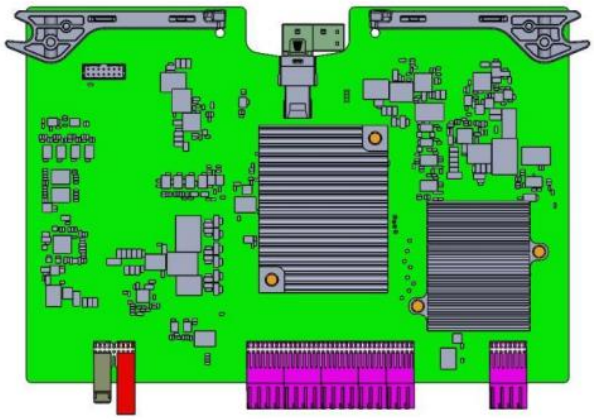
1 4x 92mm counter-rotating fans (two on each side)
2 72x hard drives, 36 on each side
3 2x Mono Lake 1S card
4 2x Storage controller card (hidden under the green air-shroud)
5 2x OCP NIC card, one on each side.

FIGURE 3: 3D RENDERING OF BRYCE CANYON SYSTEM.

Bryce Canyon is a 4-OU tall, single drawer storage chassis that is populated with seventy-two 3.5" hard drives and two compute modules. Compute module is a single socket (Broadwell-DE 16 core) Mono Lake Design [7]. Each compute module connects to a storage controller card that each control thirty-six drives. Majority of the components can be installed and serviced from the top -thus serviceability is a major aspect. In a traditional Bryce canyon system, cooling is provided by four 92mm, counter-rotating fans located at the back of the system. Air is pulled through the system from front to back. For complete system specification, please refer to [6]. Figure 3 shows a schematic representation of the air-cooled version of the Bryce Canyon. Figure 4 shows the key components of the system.



(a) I/O Module (OCP NIC Card)



(b) Storage controller card

FIGURE 4: BRYCE CANYON INTERNAL COMPONENTS.

Table 1 mentions the thermal design power for each key component. It is to be noted that power for the fans is the maximum and not operational. The fans are controlled to keep all the components under thermal threshold.

TABLE 1: THERMAL DESIGN POWER FOR BRYCE CANYON COMPONENTS

Sr. No.	Component	No. of Units	Component TDP (W)	Total TDP (W)
1	Mono Lake CPU (Broadwell DE)	2	65	130
2	Mono Lake DIMMs	4	5	20
3	HDD	72	9.6	691.2
4.	Storage Controller Card ASIC	2	33	66
5.	NIC Card ASIC	2	16	32
6.	Fans	4	56 (Max)	224

3. IMMERSION COOLED SYSTEM CONFIGURATION

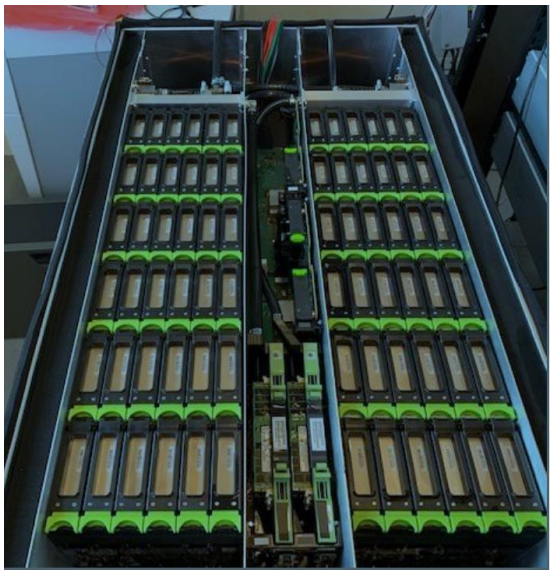


FIGURE 5: PRECISION IMMERSION LIQUID COOLED DESIGN OF BRYCE CANYON BY ICEOTOPE.

To study suitability and benefits of single-phase immersion cooling for a storage drive, an air-cooled Bryce Canyon system mentioned in earlier section was modified leveraging Iceotope' s precision based patented technology. Figure 5 shows the liquid cooled version of the system. In the air-cooled servers, fans were mounted external to the system to facilitate hot swap; for immersion cooling the depth of the chassis was extended by distance equivalent to the thickness of the fan module (56 mm). External total dimensions of the chassis remained the same. The base of the chassis serves as a reservoir for the dielectric coolant [8] used to remove heat from the components. It is to be noted

that the prototype tested is also compatible with other dielectric fluids that have lower global warming potential.

FIGURE 6: COMPACT PLATE HEAT EXCHANGER [9]

Fan assembly is replaced by two metal liquid-to-liquid heat exchanger located on the either side of the chassis behind the drives. Air shroud over the storage controller card (refer fig. 3, note 4) is removed as it would not facilitate any purpose in the absence of air as a cooling medium. The selected plate heat exchangers enable to have efficient heat transfer without allowing the dielectric and the technical water loop to mix with each other. The compact form factor of the heat sink, low weight (2.5 lbs) and ability to exchange heat up to 5kW under specific conditions [9] makes it suitable for the intended application. Figure 6 shows the picture of the plate heat exchanger. Hydraulic circuit for the dielectric loop is later explained in the paper.

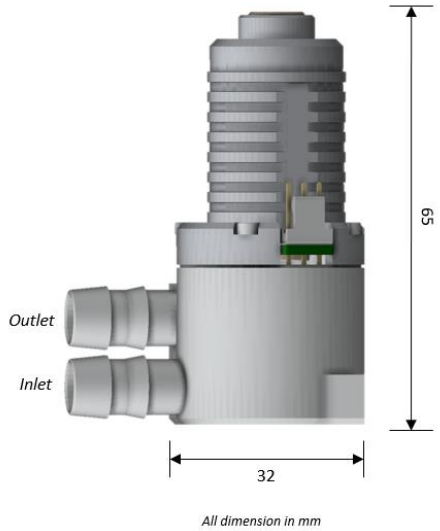


FIGURE 7: 3-D RENDERING OF MICROPUMP [10].

To circulate the dielectric through the system including heat exchanger, a compact micropump (170 grams) as shown in fig. 7 is used. The inlet port of the pump is left open and remains submerged under all conditions. Both ports are 8mm in outer diameter. The level of dielectric in the chassis is maintained to be slightly higher than the inlet port to eliminate air entrainment. It was also the reason to orient the pump in vertical direction rather than horizontal position. The outlet port is connected to inlet port of the secondary loop of heat exchanger. To account for pump failure and support N+1 redundancy, two such pumps are used – one for each heat exchanger. Each pump can provide maximum free flow of 8.7 l/min and consume 22.8W at rated conditions. An additional controller board (not shown) was also installed which was required to control the pumps. Refer to the pump specification sheet for further details [10].

In the air-cooled version of the system, HDDs in each row are spaced at 5 mm apart (fig. 8). Fans pull air through this gap to cool the drives. For immersion cooling, the gap is filled with a weir – five of them in between the HDDs and two at each side between the drive and chassis. In front of each HDD row, an exit weir is made towards the top. This arrangement creates a mini dielectric reservoir for each HDD. Each side has three manifolds which supply the dielectric coolant to each HDD reservoir. As the level of each reservoir rises, heat is removed from the HDD. Towards the top, when the level reaches the exit weir, it discharges the heated coolant into the sump on the bottom PCB. From here, the flow is directed towards the back of the chassis to pump. Fig. 9 shows the flow path for dielectric coolant along with plate heat exchangers and micropumps. Note that HDDs on both sides have supply and return path. In fig. 9 for the purpose of explanation they are not shown on each side.

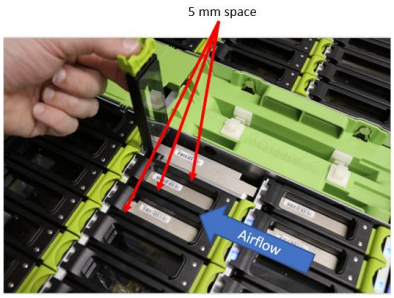
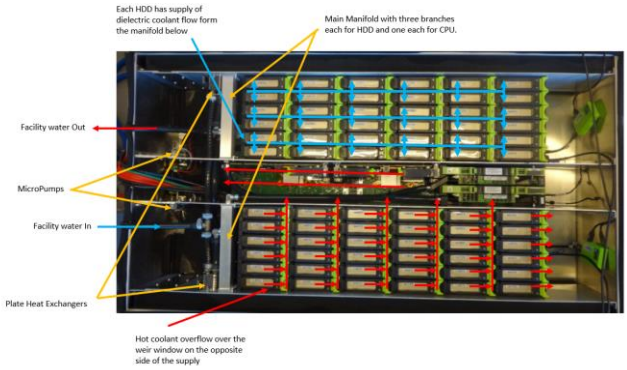
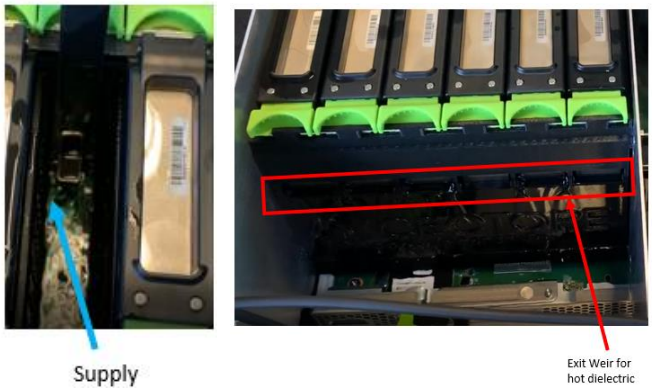


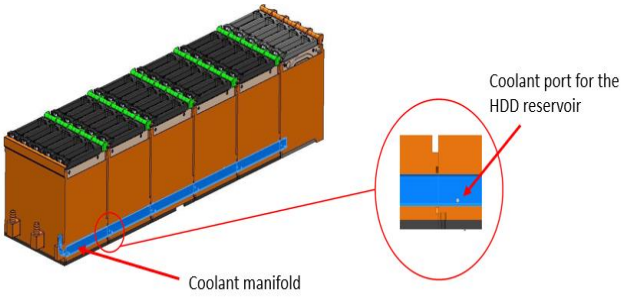
FIGURE 8: AIRFLOW THROUGH 5 MM GAP BETWEEN HDDS.



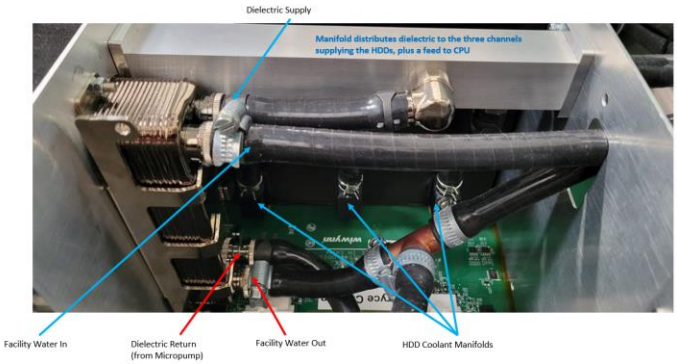
(a) Overlay of coolant distribution (blue) and return path for hot dielectric (red).



(b) Inlet manifold for coolant supply (left) and outflow of hot dielectric through exit weir (right).



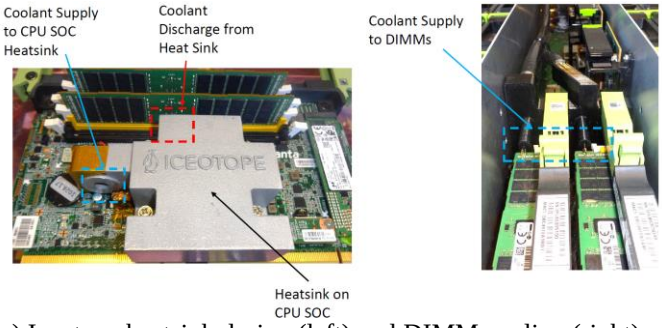
(c) Sectional CAD view to show coolant manifold and supply port to reservoir



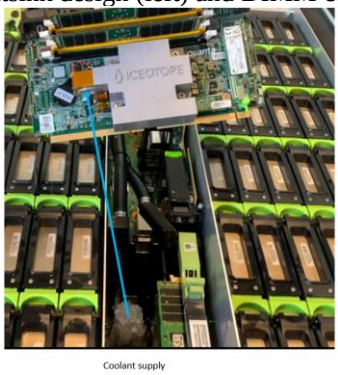
(d) Coolant distribution and facility water connections

FIGURE 9: PRECISION COOLING APPROACH FOR HDD COOLING

Figures 9a and 9d shows the dielectric feed for Monolake servers (CPU and DIMM) from the main manifold. Traditional Aluminum heat sink for air-cooled version was replaced by a custom heatsink. The heatsink essentially is a cold plate with supply and discharge ports for dielectric coolant. The feed from the manifold is further divided into two channels. One channel is connected to the supply port of cold plate. Second channel opens over the memory modules (DIMMs). Discharge from the cold plate and from the memory module flows into the sump (due to gravity) and back towards the pump. Fig. 10 shows the details of cooling Monolake card.



(a) Iceotope heatsink design (left) and DIMM cooling (right)



(b) Dielectric fountain when the Monolake card is removed

FIGURE 10: DETAILS OF MONOLAKE COOLING

The OCP NIC card is located towards lower front side of the system. As a result, it remains submerged in the dielectric all the time. However, compared to other components, temperature of dielectric coolant is relatively (about 4-5°C) higher than at the main manifold. This is because it is located on the opposite side of the pump and coolant received is overflow from HDD reservoir and Monolake cards. Similarly, chips on the expander card are in path of return flow of dielectric in sumps. For this purpose, heat sink with larger fins was used to increase the wetted surface area as shown in fig. 11.

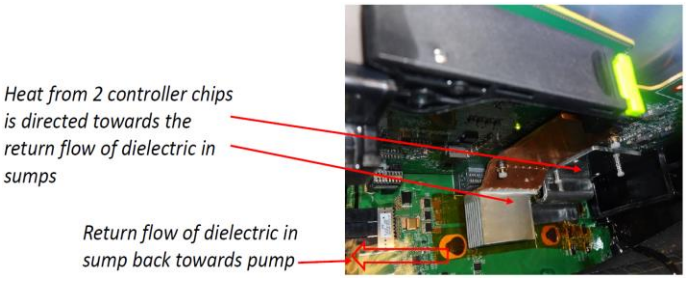


FIGURE 11: DETAILS FOR EXPANDER CARD COOLING

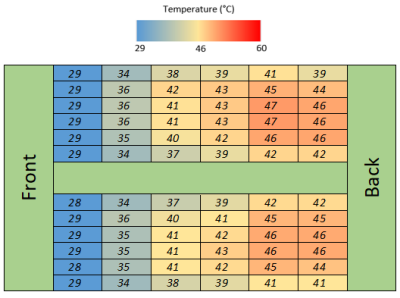
3. RESULTS AND DISCUSSION

Before the system was reworked for immersion cooling set-up, thermal and power data for air-cooled system was collected at 20°C and 45°C inlet air condition at sea level. Although air-cooled systems are typically not operated at 45°C inlet conditions, intention was to compare the test results for 40°C facility water temperature in case of immersion cooling. It was back calculated assuming a 4-5°C approach temperature for the Koolance plate heat exchanger (fig. 6). This means, the supply temperature of dielectric coolant was expected to be between 44-45°C.

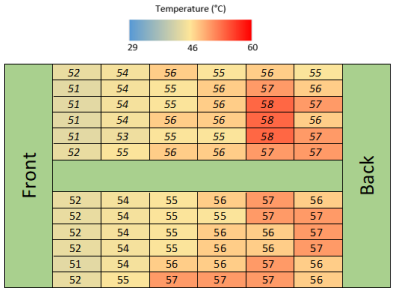
For immersion cooling, testing was performed on a lab bench. The chassis was reworked as mentioned in the previous section to make it suitable for immersion cooling. Foam insulation was used on both sides (left and right) and base of the chassis to minimize heat losses to ambient. Lauda VC3000 pony chiller was used to maintain the facility inlet water temperature of 40°C. As per the specification, chiller has a cooling output of 3kW at 20°C [11]. Nalco CCL100 was used a facility coolant. Temperature data for all the components was collected using the system BMC which would poll the data from the component sensors every three seconds. For dielectric and facility water temperature measurements, 30 AWG T-type thermocouples were used. Pump was powered through base board. Thus, the system power reported by the BMC included pump power as well. CPUs were stressed using MPRIME while HDDs were stressed using FIO script. Steady state temperatures are further reported in fig. 12 and table 2.

TABLE 2: TEST DATA OF CRITICAL COMPONENTS.

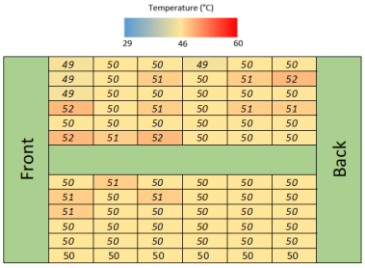
Description	Component	Temeprature Threshold (°C)	Test Data		
			Air@20°C	Air@45°C	Facility Coolant@40°C
Server 1	CPU (°C)	102	78	79	66
	Max. DIMM (°C)	85	38.3	52	49
	Expander Card -SAS Chip Temeprature (°C)	105	53.9	73	87
	Power (W)	--	615	636	639
Server 2	CPU (°C)	102	78	79	67
	Max. DIMM (°C)	85	40.9	53	58
	Expander Card -SAS Chip Temeprature (°C)	105	61	74	88
	Power (W)	--	617	640	639
Net Power (W)	$(Power)_{Server\ 1}+(Power)_{Server\ 2}$	--	1,232	1,335	1,278
Net Flow Rate		--	100 CFM	366 CFM	3 l/min



(a) Air cooling @20°C ambient



(c) Air cooling @45°C ambient



(c) Immersion cooling @40°C facility coolant

FIGURE 12: SUMMARY OF HDD TEMPERATURES.

For air cooling, fan speed was controlled by default FSC to maintain component temperature below their threshold. From fig. 12a, although HDDs are below their critical threshold of 60°C, there is about 18°C variation in HDD temperature. This is expected as fans are spinning at lower speed pushing about 100 cfm (table 2) under stress at 20°C ambient air. Different fan speeds also result in different ΔT between ambient air and maximum HDD temperature. The ΔT for 20°C and 45°C ambient condition is 26°C and 13°C respectively. Except for the HDDs in first row, all other drives receive preheated air and thus the increase in HDD temperature from front to back. Air over the two OCP NIC cards (fig. 1) eventually finds its way to mix with the relatively warmer air in the HDD section. Thus, not much variation in temperature is observed for HDDs in rows three and four (from front). Since the paper is more to discuss about immersion cooling, readers are encouraged to refer to Bryce Canyon specification [6] where additional details could be found. For 45°C ambient condition (fig. 12b), HDDs in the back row are operating with low margin of 2-3°C. There is about 7°C variation in HDD temperatures from front to back. From table 2, although all the components are below their threshold limits, system airflow is about 3.6x the previous test case (20°C ambient). The difference in system power is partially due to the fans spinning at maximum speed thus consuming more power. It is unusual for any data center to operate at 45°C ambient. The test was performed as a reference to compare the performance to that of immersion cooling.

For immersion cooling, comparing results from fig 12c to 12b, the HDDs are operating at lower temperatures. Variation in HDD is also lowered to 3°C compared to air-cooled test data. Since each HDD is placed in its own reservoir with a coolant

manifold towards the bottom, it was expected that all the HDD would exhibit variance of 1°C or lower. It could be due to the manufacturing tolerance of the manifold opening, there is a variation in volume of coolant supplied to each reservoir. None the less the HDD temperatures are quite below their temperature threshold. From table 2, except for SAS chips for both servers and the DIMM for server 2, all other components operate at lower temperature compared to its air-cooled version. It was expected for SAS chips to be at higher temperatures as heat from the chips is directed towards the return flow of dielectric in sump. For DIMM modules in server 2, the dielectric supply was slightly misaligned with the card that resulted in a higher temperature. It was a design problem which was later fixed. The dielectric supply temperature was measured about 44.2°C while the dielectric return temperature was about 46.9°C. The pump power was approximately 46 W which is just shy of five percent of the total system power.

The data indicates that benefits of liquid cooling (single phase immersion) can also be realized for a high-density storage system by operating at a facility water temperature of 40°C. Another benefit of this test set-up was its virtually silent operation. In the absence of fans, there would be no acoustic vibrations which could potentially hinder the drive performance. The air-cooled versions of the chassis have several features such as honeycomb structures, fan guards etc. to lower the acoustic performance. In the prototype tested (first iteration), each half of the chassis (36 HDDs) was served by its own pump and heat exchanger. There was no cross-over nor spare capacity for the other pump to serve the full chassis. However, in the recent prototype minor modifications to the design loop were made to support N+1 redundancy. Since the design utilizes single phase immersion cooling in a chassis form factor, it could be mounted on rails like air-cooled chassis. The design allows for HDDs to be hot swapped in case of disk failures without the need for draining the system.

4. CONCLUSION

In this study, an air-cooled, high-density storage system was re-engineered to utilize single phase immersion cooling. The design enabled all seventy-two HDDs to be cooled in parallel at 40°C facility coolant supply. The design achieved minimal variation in drive temperatures and majority of the components (including HDDs) operated at lower temperatures compared to air-cooling under comparable conditions. System level cooling power was less than five percent of the total power consumption. Apart from efficiency, the design approach could potentially eliminate acoustic vibration issues often encountered in an air-cooled solution.

DISCLAIMER

The project was a research collaboration between the authors. Designs presented contains Iceotope' s patented technology and no license is granted to copy or replicate any element of the design without express permission of Iceotope.

REFERENCES

[1] Panigraphy, A., Subramanyam, P., Pang, Y.-F., Sahan, R. and Xia, A., "Optimizing Closed-Loop Liquid Cooling Solution for Extreme High Power Multi-Packages," Proceedings of the ASME 2021 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems, InterPACK2021.

[2] "Emergence and Expansion of Liquid Cooling in Mainstream Data Centers," White Paper Developed by ASHRAE Technical Committee 9.9 for Mission Critical Facilities, Data Centers, Technology Spaces, and Electronic Equipment, 2021.

[3] Vertal, L. and Hughes, P., "WP#70-Liquid Cooling Technology Update", The Green Grid, 2016.

[4] Varma D., "Data Center Cold Wars -Part 3: Single-Phase Immersion Cooling Versus Cold Plate," GRC Technical Blog, 2020.

[5] Bansode, P., Shah, J., Gupta, G., Agonafer, D., Patel, H., Roe, D., and Tufty, R., "Measurement of the Thermal Performance of a Custom-Build Single-Phase Immersion Cooled Server at Various High and Low Temperatures for Prolonged Time," Journal of Electronic Packaging Vol. 142, 2020.

[6] Bryce Canyon Storage System Specification V1.0, Open Compute Project, 2018.

[7] Mono Lake 1S Server Design Specification V0.4, Open Compute Project.

[8] https://www.appliedthermalfluids.com/product/galden-ht-70-perfluorinated-pfpe-heat-transfer-fluid-5kg-bottle/

[9] Specification Sheet for Koolance P/N: HXP-193

[10] DS32-M510 S Data Sheet, Rev 2 Small, TCS Micropumps.

[11] https://www.lauda.de/pimimport/assets/context/pdmarticle/85/8571/8571/attachments/Export.8571.2018-10-24-16-15-24.11e5359f.pdf