Proceedings of the ASME 2023 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems
InterPACK2023
October 24-26, 2023, San Diego, California

IPACK2023-112054

OPTIMIZATION OF A AIR-COOLED HEATSINK FOR IMMERSION COOLING APPLICATION

Gautam Gupta
The University of Texas at Arlington
Arlington, TX

Vivek Nair
The University of Texas at Arlington
Arlington, TX

Sai Abhideep Pundla
The University of Texas at Arlington
Arlington, TX

Pratik Bansode
The University of Texas at Arlington
Arlington, TX

Rohit Suthar
The University of Texas at Arlington
Arlington, TX

Joseph Herring
The University of Texas at Arlington
Arlington, TX

Jacob Lamotte-Dawaghreh
The University of Texas at Arlington
Arlington, TX

Krishna Bhavana Sivaraju
The University of Texas at Arlington
Arlington, TX

Dereje Agonafer
The University of Texas at Arlington
Arlington, TX

Poornima Mynampati
Silent-Aire USA Inc.
Gilbert, AZ

Mike Sweeney
Silent-Aire USA Inc.
Gilbert, AZ

ABSTRACT

Data centers have started to adopt immersion cooling for more than just mainframes and supercomputers. Due to the inability of air cooling to cool down recent high-configured servers with higher Thermal Design Power, current thermal requirements in machine learning, AI, blockchain, 5G, edge computing, and high-frequency trading have resulted in a larger deployment of immersion cooling. Dielectric fluids are far more efficient at transferring heat than air. Immersion cooling promises to help address many of the challenges that come with air cooling systems, especially as computing densities increase. Immersion-cooled data centers are more expandable, quicker installation, more energy-efficient, allows for the cooling of almost all server components, save more money for enterprises, and are more robust overall. By eliminating active cooling components such as fans, immersion cooling enables a significantly higher density of computing capabilities. When utilizing immersion cooling for server hardware that is intended to be air-cooled, immersion-specific optimized heat sinks should be used. A heat sink is an important component for server cooling efficacy. This research conducts an optimization of heatsink for immersion-cooled servers to achieve the minimum case temperature possible utilizing multi-objective and multi-design variable optimization with pumping power as the constraint.

A high-density server of 3.76 kW was modeled on Ansys Icepak that consists of 2 CPUs and 8 GPUs with heatsink assemblies at their Thermal Design Power along with 32 Dual In-line Memory Modules. The optimization is conducted for Aluminum heat sinks by minimizing the pressure drop and thermal resistance as the objective functions whereas fin count, fin thickness, and heat sink height are chosen as the design variables in all CPUs, and GPUs heatsink assemblies. Optimization for the CPU and the GPU heatsink was done separately and then the optimized heatsinks were tested in an actual test setup of the server in ANSYS Icepak. The dielectric fluid for this numerical study is EC-110 and the cooling is carried out using forced convection. A Design of Experiment (DOE) is created based on the input range of design variables using a full-factorial approach to generate multiple design points. The effect of the design variables is analyzed on the objective functions to establish the parameters that have a greater impact on the performance of the optimized heatsink. The optimization study is done using Ansys OptiSLang where AMOP (Adaptive Metamodel of Optimal Prognosis) as the sampling method for design exploration. The results show total effect values of heat sinks geometric parameters to choose the best design point with the help of a Response Surface 2D and 3D plot for the individual heat sink assembly.

Keywords: Single-phase, immersion cooling, optimization, thermal management, heat transfer coefficient, heatsink, dielectric fluid

NOMENCLATURE

SPIC	Single Phase Immersion Cooling
TDP	Thermal Design Power
CPU	Computer Processing Unit
GPU	Graphic Processing Unit
DIMM	Dual In-line Memory Module
TIM	Thermal Interface Material
LPM	Liter Per Minute
COP	Coefficient of Optimal Prognosis
AMOP	Adaptive Metamodel of Optimal Prognosis
DOE	Design of Experiment
PCB	Printed Circuit Board
OCP	Open Compute Project
h	Sensible enthalpy
k	Conductivity
k_t	Turbulence Transport Conductivity
v	Velocity vector
t	Time
T	Temperature
S_h	Volumetric heat source
ρ	Density
τ	Stress tensor

1. INTRODUCTION

Data centers are essential infrastructures for modern society, serving as the backbone of many online services and applications. As the volume of data being generated continues to grow, data centers have evolved to accommodate more electronics in smaller spaces, leading to higher heat generation and energy utilization. According to a report, global international bandwidth is at 997Tbps, which is a tripling of bandwidth since 2018 [1].

The increase in data processing and storage demands has led to higher densities of servers, storage, and networking equipment in data centers, creating a need for effective cooling systems to maintain optimal operating conditions [2]. A bottom-up study showed that data centers consume up to 1.1 to 1.5% of the world's total electricity supply [3]. This rapid increase in the energy consumption directly relates to the increase of heat generated in the data centers. As shown in Fig.1, for a 42U rack, the heat loads increase significantly [4].

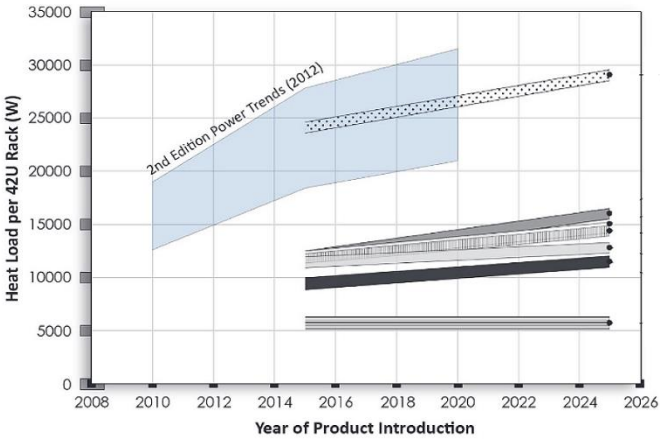

FIGURE 1: HEAT LOAD PER RACK ACROSS THE YEARS

Traditional air-cooling methods for data centers have limitations including inefficiency in dissipating heat from high-density servers, limited cooling capacity for densely packed racks, formation of hotspots impacting server performance, high energy consumption, space requirements, complex design and maintenance, and reduced scalability [5,6]. Currently, there is an increased focus on liquid cooling, immersion cooling, and hybrid systems to overcome these limitations and improve energy efficiency and cooling capabilities in modern data centers [7]. The effective thermal conductivity of various types of liquid cooling systems in Fig.2. illustrates the effective cooling capabilities of the systems as compared to air cooling [8].

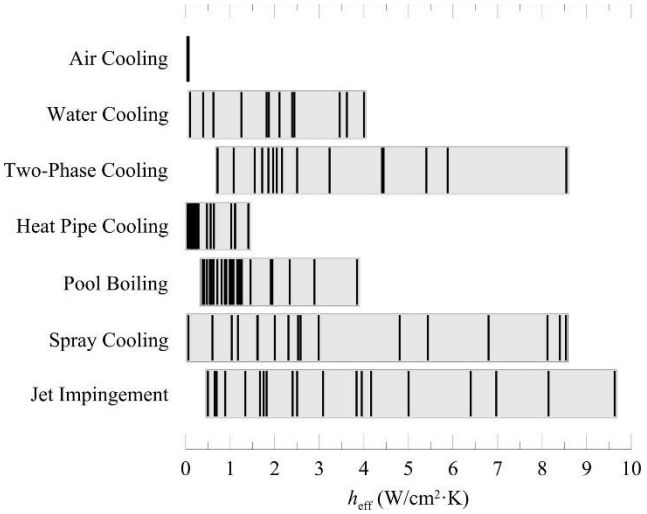

FIGURE 2: EFFECTIVE THERMAL CONDUCTIVITY OF VARIOUS TYPES OF LIQUID COOLING

Liquid cooling has proven to be an effective solution for high-power demand and offers several advantages over air cooling [9]. The power on multi-module chips has increased, and chip manufacturers are compensating for the fading effects of Dennard's scaling by increasing the number of transistors on the chip to achieve improved performance within the same chip area. Consequently, liquid cooling systems have seen significant adoption in data centers [10,11]. In addition to addressing power

density concerns, liquid cooling can greatly improve the energy efficiency of the system compared to air cooling, which requires a substantial amount of power [12]. The increase in heat dissipation poses challenges to chip performance and reliability in air cooling [13], potentially leading to thermal shutdown and system failure. In contrast, liquid cooling systems offer better heat dissipation, while also requiring fewer units and facilitating easier maintenance.

On the basis of how the coolant interacts with electronic components, liquid cooling in data centers can be divided into two primary categories: direct and indirect cooling Direct cooling as the name suggests is a cooling method in which the liquid is directly in contact with the electronic package, some of the direct liquid cooling types are immersion cooling[14], pool boiling[15], submerged jet impingement, and spray cooling while indirect cooling involves the cooling when the liquid is not in contact with the electronic package, which include cold plates[16], heat pipes, and vapor chambers[17]. The fluid used in direct liquid cooling systems must meet specific requirements for efficient cooling. It should have high thermal conductivity, be non-corrosive, non-conductive, and have low viscosity [18,19]. The fluid should also have a high boiling point, low freezing point, and chemical stability. Environmental considerations may also be important. Common options include dielectric liquids and specialized coolant solutions.

Single-phase immersion cooling is a type of immersion cooling technology that involves immersing electronic components, such as servers, storage devices, and networking equipment in a non-conductive engineered dielectric fluid [20]. The fluid absorbs the heat generated by the electronic components, which is then transferred to a heat exchanger and dissipates into the environment [21]. Unlike two-phase immersion cooling, which uses a combination of liquid and vapor to cool the electronic components, single-phase immersion cooling uses only a liquid fluid, making it simpler to implement and maintain. Single-phase immersion cooling has been shown to provide high cooling efficiency, reduce energy consumption, and improve the reliability and lifespan of electronic components [22].

Using Ansys Icepak, a high-density server configuration with a power consumption of 3.76 kW is modeled, comprising 2 CPUs and 8 GPUs with 32 Dual In-line Memory Modules. Aluminum heat sinks are selected for optimization, with pressure drop and thermal resistance minimized as the objective functions. The design variables considered for optimization include fin count and fin thickness in the heat sink assemblies. Balancing thermal resistance and pressure drop is crucial for effective heatsink design. A well-designed heatsink should provide low thermal resistance to efficiently remove heat from the electronic component while also maintaining an acceptable pressure drop to ensure sufficient fluid or airflow. Achieving this balance ensures efficient cooling without compromising the system's energy consumption or performance.

Forced convection cooling is implemented using the EC-110 dielectric fluid Design of Experiment (DOE) is constructed employing a full-factorial approach to generate multiple design points within the specified range of design variables. The impact of these design variables on the objective functions is analyzed to identify parameters with significant influence on the optimized heat sink's performance. The optimization study is conducted using Ansys OptiSLang, utilizing the Adaptive Metamodel of Optimal Prognosis (AMOP) as the sampling method for design exploration. The Metamodel of Optimized Prognosis (MoP) is an automatic approach that aims to provide a user-friendly solution for parameter optimization. This methodology combines the use of metamodeling techniques and optimization algorithms to streamline the process of finding optimal parameter values for a given system or model.

2. MATERIALS AND METHODS

In this study, a high-density server conforming to the 3 OU Rackmount form factor was specifically selected. The server's chassis exhibits dimensions measuring 814 mm (L) x 531 mm (W) x 139 mm (H). It is structurally partitioned into two primary tiers: an upper tier and a lower tier. Within the upper tier, two central processing units (CPUs) coexist with 32 dual in-line memory modules (DIMMs), whereas the lower tier encompasses the installation of eight graphics processing units (GPUs). The focal components of interest, namely the heatsinks, CPUs with their underlying GPUs, DIMMs, and their associated smaller heatsinks, along with the printed circuit board (PCB) that accommodates all these constituents, were incorporated into the study. Conversely, the encompassing casing enclosing the components, the metallic framework serving as hosts for fans and hard disk drives (HDDs), as well as the screws, washers, nuts, and bolts utilized for assembly, were deliberately omitted from consideration due to their non-participation in heat transfer processes. This deliberate omission was undertaken as a means to simplify the model, diminish model complexity, and reduce the overall mesh count.

3.4 Server Base Model

The baseline configuration of the server was developed using the commercially available ANSYS Icepak software, as illustrated in Figure 3. The baseline model encompasses essential components, namely 2 CPUs and 8 GPUs, each equipped with their individual heatsink assembly, along with 32 DIMMs. To simplify the model and reduce the mesh count, only heat-dissipating components were included. Heat sources originating from the CPUs and GPUs were represented as two-dimensional solid obstructions within the model. This step significantly contributed to reducing the model's complexity. The thermal design power values of the CPUs, GPUs, and DIMMs are presented in Table 1. Considering all the components, the total power consumption of the server amounts to 3760 W. For our study, the selected dielectric fluid is EC-110, with its relevant properties discussed in Table 2. The fluid EC-110 used in this CFD study is temperature-dependent. This dielectric fluid possesses high thermal mass, rendering it thermally conductive yet electrically non-conductive. Its implementation aids in

efficiently dissipating heat from the server. The server's orientation is vertical, with the fluid entering from the bottom and flowing against the direction of gravity. To maintain an effective cooling, a flow rate of 5 gallons per minute (GPM) was maintained for the server setup.

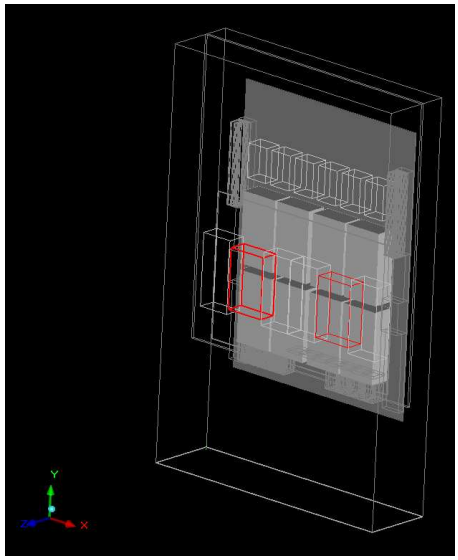

FIGURE 3: MODEL SETUP OF THE SERVER ON ICEPAK

TABLE 1: POWER CONSUMPTION OF EACH COMPONENT

Component	Quantity	Power (W)
TDP of each CPU	2	200
TDP of each GPU	8	400
DIMMs	32	5

TABLE 2: VARIATION OF THERMO-PHYSICAL PROPERTIES OF EC-110 AT DIFFERENT FLUID TEMPERATURES [22]

Temperature	Thermal Conductivity	Specific Heat	Dynamic Viscosity	Density
(°C)	(W/m-K)	(J/kg-K)	(Ps-S)	(kg/m^3)
10	0.138	2096	0.0148	852.5
20	0.137	2133	0.0100	845.9
30	0.136	2171	0.0072	839.3
40	0.136	2209	0.0053	832.7
50	0.135	2247	0.0041	826.1
60	0.135	2285	0.0033	819.5
70	0.134	2323	0.0026	812.9
80	0.134	2360	0.0022	806.3
90	0.133	2398	0.0019	799.7

The upper tier of the server is configured with two spread-core CPUs, accompanied by the installation of 32 DIMMs in three distinct sections, as visually depicted in Figure 4. Specifically, there are 8 DIMMs positioned to the left of CPU 1, 16 DIMMs situated between the two CPUs, and an additional 8 DIMMs positioned to the right of CPU 2. All these components are mounted on a single printed circuit board (PCB).

FIGURE 4: LOCATION OF CPUs AND DIMMs IN UPPPER TIER OF THE SERVER

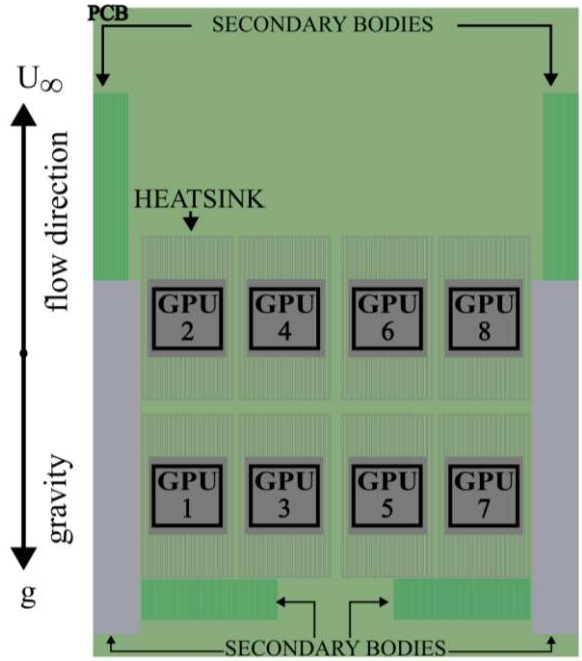

FIGURE 5: LOCATION OF GPUs IN LOWER TIER OF THE SERVER

Moving to the lower tier of the server, Figure 5 illustrates the presence of a PCB with 8 GPUs and their corresponding heatsink assemblies. These GPUs are divided into two sections: the front stack includes GPU 1, GPU 3, GPU 5, and GPU 7, while the rear stack encompasses GPU 2, GPU 4, GPU 6, and GPU 8. For the current study, an Indium foil measuring 0.5 mm in thickness was selected as the thermal interface material (TIM) of choice. The overall thermal resistance calculated takes into account the thermal resistance from case to TIM and TIM to heatsink. A combined thermal resistance is considered in this CFD study to assess the overall effectiveness of the heatsink in transferring the heat.

2.2 Mesh Sensitivity Analysis

A Grid Independence study was conducted on the server base model under specific operating conditions: a fluid inlet temperature of 40 °C and a flow rate of 5 (GPM). The study's objective was to ascertain the sensitivity of the CPU, front stack GPUs, and rear stack GPUs temperatures with varying element sizes, which correspond to different levels of mesh refinement ranging from a coarser mesh at 10 million elements to a finer mesh at 35 million elements. The results, depicted in Figure 6, revealed that the temperatures of these components remained constant across the range of element sizes. Consequently, for the server's baseline study, a grid count of 20 million elements was deemed sufficient. To further analyze mesh sensitivity, non-conformal meshing techniques were employed. The heatsink

stack-up was meshed separately using per object mesh parameters. Specifically, the mesh resolution was increased for the heatsink, encompassing the base, fins, and the interstitial cells between the fins. Slack settings were incorporated to ensure a smooth transition of fluid flow from the server region to the heatsink, promoting enhanced thermal performance.

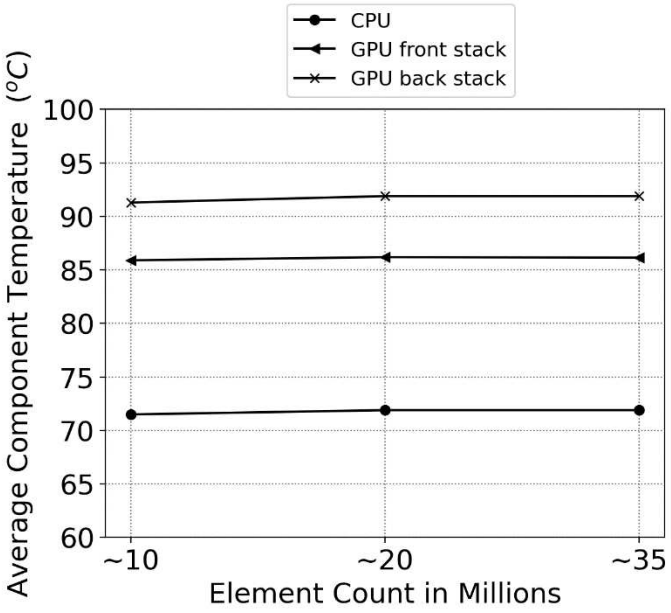

FIGURE 6: MESH SENSITIVITY ANALYSIS

2.5 Governing Equations

The CFD tool employed in this study utilizes the Navier-Stokes equations, encompassing the conservation equations for mass, momentum, and energy, to accurately model heat transfer under laminar flow conditions. If turbulence and radiation are involved in the flow and heat transfer phenomena, additional transport equations can be incorporated. However, for the present study, these additional equations were not considered. These equations are written as follows:

Mass conservation:

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho \vec{v}) = 0 \qquad (1)$$

The above equation reduces to $\nabla.(\vec{v}) = 0$ for incompressible fluids.

Momentum Equation:

$$\frac{\partial}{\partial t}(\rho \vec{v}) + \nabla.(\rho \vec{v}\vec{v}) = -\nabla p + \nabla.(\bar{\bar{\tau}}) + \rho \vec{g} + \vec{F} \qquad (2)$$

Energy Equation:

$$\frac{\partial}{\partial t}(\rho h) + \nabla.(\rho h \vec{v}) = \nabla.[(k + k_t)\nabla T] + S_h \qquad (3)$$

Here, the fluid energy equation is written in terms of sensible enthalpy, h. k is the molecular conductivity and kt is the turbulence transport conductivity. The source term S_h represents user-defined volumetric heat sources. For the solid regions, the energy equation due to conduction within the solid looks as follows:

$$\frac{\partial}{\partial t}(\rho h) = \nabla.(k \nabla T) + S_h \qquad (4)$$

Here, k is the thermal conductivity of the solid, ρ is the density, T is the temperature and Sh is the source term for volumetric heat sources.

2.3 Model Setup for CPU Heatsink Optimization

In order to conduct an optimization study on the CPU heatsink, a CFD model was developed using a test chamber within the ANSYS Icepak. The CPU dissipates a power of 200 W. The thermal stack representing the CPU consists of a two-dimensional (2D) heat source positioned atop a chip socket. On top of the CPU, there is a Thermal Interface Material (TIM), followed by the placement of a heatsink. The Indium TIM utilized in the model has a thickness of 0.5 mm and exhibits a thermal conductivity of 8 W/m-K. The baseline parameters for the optimization study can be found in Table 3. The flowrate for the optimization study is based of a certain temperature difference target and based on that 0.5 LPM is set as the constant flowrate for the CPU. The fluid flow within the system is directed opposite to the force of gravity in the positive z-direction. For this particular CPU heatsink optimization study, the chosen dielectric fluid is EC-110, which will serve as the heat transfer medium for the investigation.

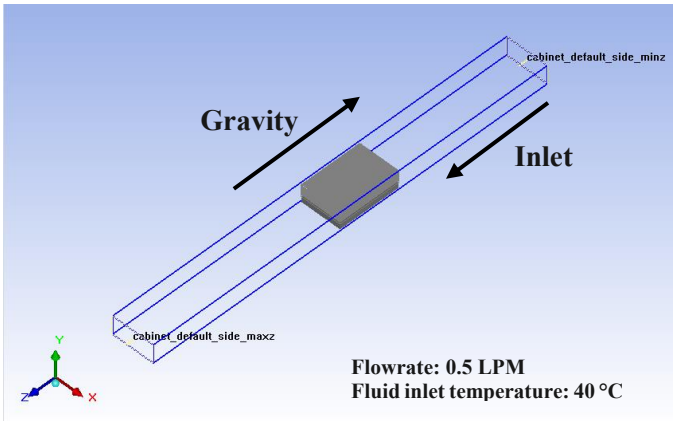

FIGURE 7: MODEL SETUP FOR CPU HEATSINK OPTIMIZATION

TABLE 3: INPUT PARAMETERS FOR CPU BASE MODEL

CPU Power	200 W
Fluid Inlet Temperature	40 °C
Flowrate	0.5 LPM
Heatsink Overall Height	24.4 mm
Heatsink Base Height	4 mm

Heatsink Fin count	58
Heatsink Fin Thickness	0.3 mm

2.4 Model Setup for GPU Heatsink Optimization

Similarly, a model setup for GPU and its heatsink assembly was created on ANSYS Icepak where a GPU was modeled inside a test chamber with a heatsink and a TIM in the middle. The dielectric fluid is EC-110. The input parameters for the baseline are discussed in Table 4.

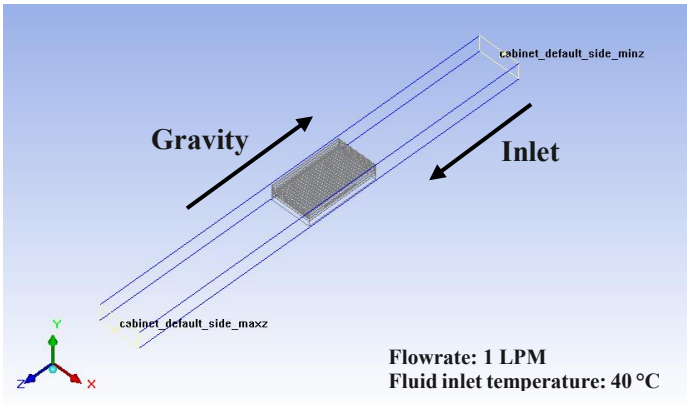

FIGURE 8: MODEL SETUP FOR GPU HEATSINK OPTIMIZATION

TABLE 4: INPUT PARAMETERS FOR GPU BASE MODEL

GPU Power	400 W
Fluid Inlet Temperature	40 °C
Flowrate	1 LPM
Heatsink Overall Height	20 mm
Heatsink Base Height	4 mm
Heatsink Fin count	20
Heatsink Fin Thickness	1.5 mm

2.6 OptiSLang Setup

In our study, OptiSLang serves as the designated design optimization tool. The study encompasses two distinct optimization simulations, each targeting the CPU heatsink and GPU heatsink, respectively, as previously discussed in sections 2.3 and 2.4. Following the completion of baseline simulations for both the CPU and GPU models, the parameter set comprising input and output parameters is exported and integrated into Workbench. The output parameters comprise primary and compound functions, including CPU/GPU temperature, pressure difference across the heatsink, and thermal resistance. The subsequent step involves identifying the design variables to be optimized. In our study, these variables pertain to the fin thickness, fin count, and heatsink height for both the CPU and GPU heatsinks, as detailed in Tables 5 and 6, respectively. These design variables are utilized within OptiSLang for design exploration purposes, aiming to enhance the performance of the heatsinks within an immersion-cooled server setup. The objective functions for this investigation revolve around minimizing thermal resistance and pressure drop. Figure 9 provides a visual representation of the integration process between Icepak and OptiSLang. To streamline the computational process, the design variables are chosen based on individual parameter studies, effectively narrowing down the number of variables and reducing computational time. Based on the parameters outlined in Table 5, a total of 168 design points are established (7 x 8 x 3), while for the GPU heatsink, based on the input parameters, a total of 180 design points are generated (6 x 6 x 5) as per table 6.

TABLE 5: INPUTS OF DESIGN VARIABLES FOR CPU

Parameters	Baseline value	Discrete Chosen Values	Total Variabes
Fin Count	58	18, 20, 22, 24, 26, 28, 58	7
Fin Thickness	0.3 mm	0.3, 1, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6	8
Heatsink Height	24.4 mm	20.4, 22.4, 24.4, 26.4	3

TABLE 6: INPUTS OF DESIGN VARIABLES FOR GPU

Parameters	Baseline value	Discrete Chosen Values	Total Variabes
Fin Count	20	18, 20, 22, 24, 26, 28	6
Fin Thickness	1.5 mm	1, 1.1, 1.2, 1.3, 1.4, 1.5	6
Heatsink Height	20 mm	20, 22, 24, 26, 28	5

OptiSLang, an integral component of ANSYS Workbench, offers a significant advantage by seamlessly integrating with ANSYS Icepak. This integration enables the independent solving of the simulation model targeted for optimization. Within the simulation module, the design parameters and their respective ranges or bounds for optimization are defined. Subsequently, these parameters are imported into OptiSLang for further analysis and optimization. OptiSLang employs a meta-modeling approach, specifically the adaptive meta-model of optimal prognosis (AMOP), to effectively sample the design space. This approach utilizes a Coefficient of Optimal Prognosis (COP) to approximate the quality of the model, aiding in the identification of optimal design solutions.

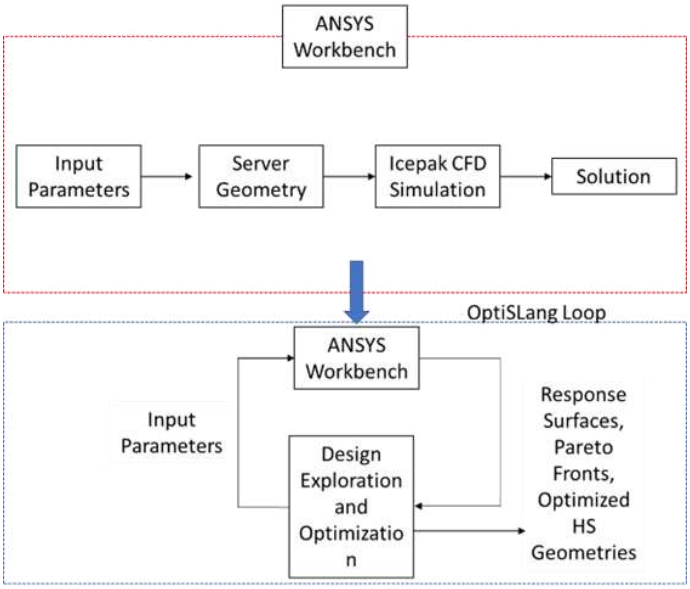

FIGURE 9: STEPWISE INTEGRATION OF ICEPAK WITH OPTISLANG SENSITIVITY ANALYSIS ON ANSYS WORKBENCH [23]

3. RESULTS AND DISCUSSION

The optimization results presented in this section are divided into three sections for the CPU, GPU and the server optimization results. The baseline results of the CPU setup, GPU setup and the server setup are compared with the optimized heatsinks based on the different input parameters.

3.4 Optimization Results of CPU

To identify the most optimal heatsink parameter with the lowest thermal resistance while maintaining a lower differential pressure, the baseline results for the CPU are compared against various design points. The baseline model exhibited a thermal resistance of 0.054 °C/W, with a maximum CPU temperature of 53.46 °C, as depicted in Figure 10. The pressure difference across the heatsink was measured at 15 Pa.

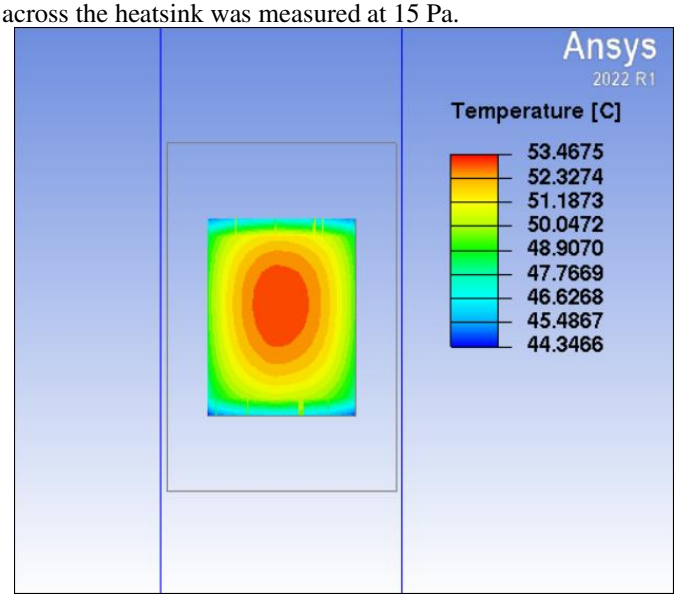

FIGURE 10: TEMPERATURE CONTOUR OF CPU BASELINE MODEL

ANSYS OptiSLang leverages advanced mathematical techniques to construct response surface models using extracted simulation data. These models provide an approximation of the relationship between the design variables and response variables, facilitating efficient optimization and sensitivity analysis. The initial phase of the optimization study focused on conducting sensitivity analysis for the design variables in relation to the objective functions. Figure 11 illustrates the total effects plots for the first optimization case, specifically at a CPU power of 200 W under forced convection flow. These plots serve to quantify the influence of each input variable on the corresponding outputs or post-processing functions. Fin thickness appears to be the most dominating factor (58.3%) in case of thermal resistance whereas fin count is 24.7 %. This also implicates that having a greater number of fins does not necessarily mean better heat transfer due to more surface area. Increasing the heatsink height can potentially enhance the convective heat transfer by providing additional surface area for heat exchange with the dielectric fluid. Heatsink height contributes to 13.4 % in the total effects plot. Notably, a linear regression CoP value exceeding 90% was achieved for the CPU model outputs. This indicates that the sample points were generated based on the design variable inputs, resulting in the creation of a highly robust model.

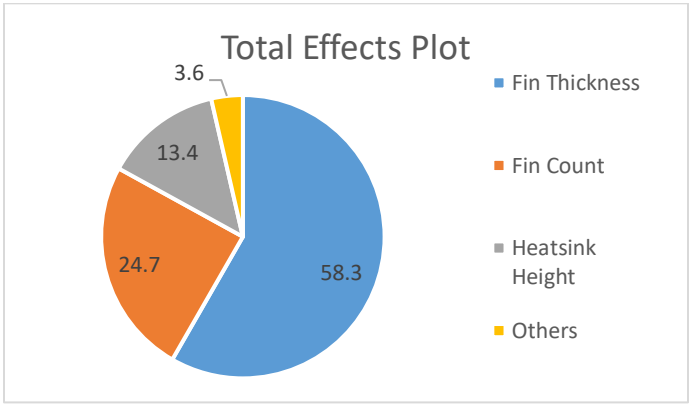

FIGURE 11: TOTAL EFFECTS PLOT OF CPU HEATSINK OPTIMIZATION

As detailed in section 2.4, a total of 168 design points were generated for the optimization of the CPU heatsink. To visualize the dependencies between the input design variables, namely fin thickness, fin count, and heatsink height, and thermal resistance, linear regression-based plots and response surfaces were employed. Figure 12, Figure 13, and Figure 14 showcase the 2D and 3D dependencies of these design variables on thermal resistance. In Figure 12, it is evident that fin thickness exhibits a significant impact on variations in thermal resistance. Specifically, a fin thickness of 1.5 mm yielded the lowest thermal resistance. Figure 13 illustrates the relationship between fin count and thermal resistance. It becomes apparent that a fin count of 26 resulted in the lowest thermal resistance. As the fin count decreases, the thermal resistance increases. This observation aligns with the notion that higher fin count does not necessarily

lead to improved heat dissipation, which explains why heatsinks designed for air-cooling may not perform optimally in an immersion-cooled setup. Figure 14 explores the association between heatsink height and thermal resistance. As expected, minimal changes in thermal resistance are observed when the heatsink height remains relatively constant. However, significantly increasing the heatsink height leads to a notable reduction in thermal resistance. It is important to note that maintaining the 1U form factor is a key consideration in the design process.

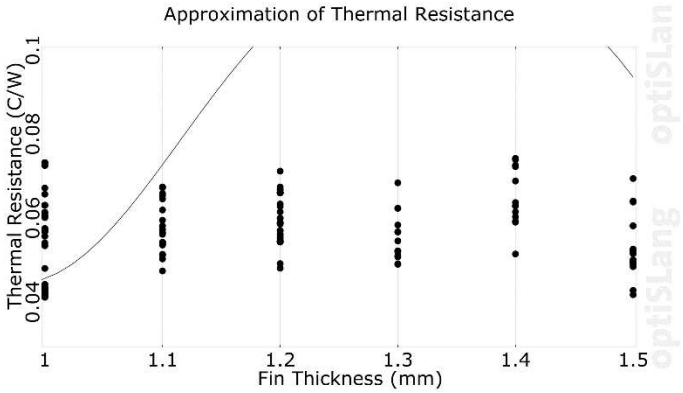

FIGURE 12: RELATION BETWEEN THERMAL RESISTANCE AND FIN THICKNESS FOR CPU

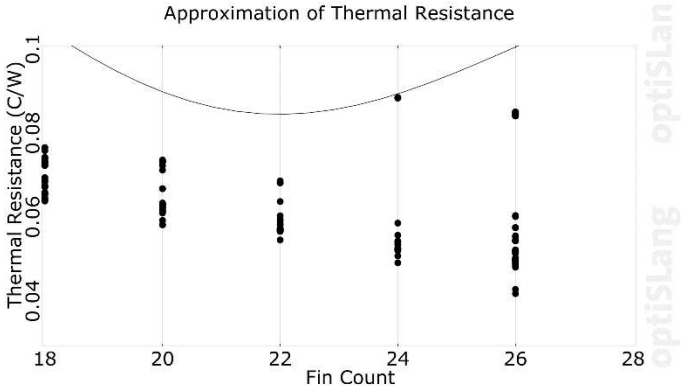

FIGURE 13: RELATION BETWEEN THERMAL RESISTANCE AND FIN COUNT FOR CPU

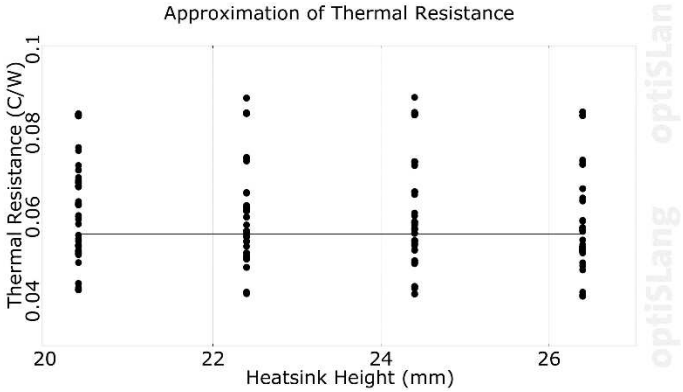

FIGURE 14: RELATION BETWEEN THERMAL RESISTANCE AND HEATSINK HEIGHT FOR CPU

After analyzing all the data of 168 design points, table 7 reflects the best design parameters of the heat sink with the lowest thermal resistance which includes the heat sink fin count, fin thickness and the heatsink height. The best design point 21 has shown the lowest resistance which is ~24 % lower than the baseline heatsink. The next best 4 design points also showed very similar thermal resistance.

TABLE 7: COMPARSION OF BEST DESIGN POINTS WITH THE BASELINE PARAMETERS

Design Point Number	Fin Count	Fin Thickness (mm)	Heatsink Height (mm)	R_{th} (°C/W)	CPU Temp (°C)
Baseline	**58**	**0.3**	**24.4**	**0.054**	**53.46**
21	26	1.5	26.4	0.041	51.27
115	26	1.5	24.4	0.041	51.34
87	26	1.5	22.4	0.041	51.38
5	26	1.1	22.4	0.042	51.46
58	24	1.3	22.4	0.043	51.65

3.2 Optimization Results of GPU

In the second phase of the optimization study, a comparison is made between the baseline results for the GPU and 180 different design points to identify the most efficient heatsink configuration. The objective is to minimize thermal resistance while ensuring that the pressure difference across the heatsink does not increase. The baseline model exhibited a thermal resistance of 0.051 °C/W, with a maximum GPU temperature of 65.42 °C at a power of 400 W, as depicted in Figure 15. The pressure difference across the heatsink was measured at 12 Pa.

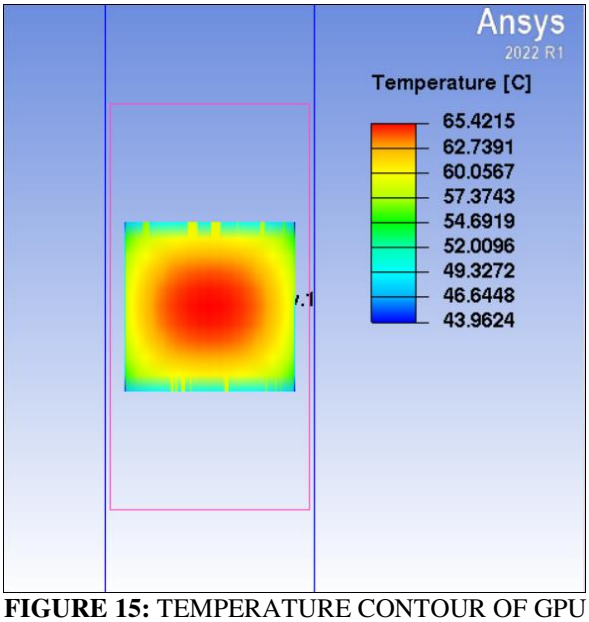

FIGURE 15: TEMPERATURE CONTOUR OF GPU BASELINE MODEL

Figure 16 presents the total effects plots for the GPU heatsink optimization case under forced convection flow at a power of 400 W and a flow rate of 1 LPM. Similar to the CPU model, a linear regression CoP value exceeding 90% was achieved for the GPU model outputs. This indicates that the sample points were generated based on the design variable inputs.

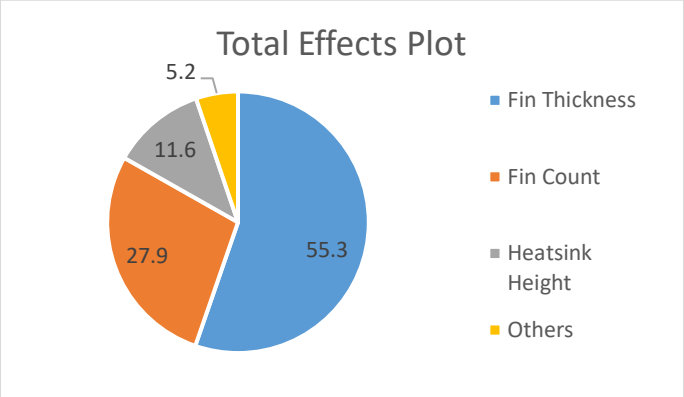

FIGURE 16: TOTAL EFFECTS PLOT OF GPU HEATSINK OPTIMIZATION

Response surface plots, as illustrated in Figure 17, Figure 18, and Figure 19, depict the influence of heatsink fin thickness, fin count, and heatsink height on thermal resistance. Increasing the fin thickness from 1 to 1.5 mm generally leads to a reduction in thermal resistance, with the lowest thermal resistance observed at a fin thickness of 1.4 mm. Likewise, a fin count of 26 yielded the best thermal resistance, which aligns with the findings from the CPU heatsink optimization. Finally, a heatsink height of 28 mm exhibited the lowest thermal resistance, as previously discussed, as the heatsink height is inversely proportional to thermal resistance.

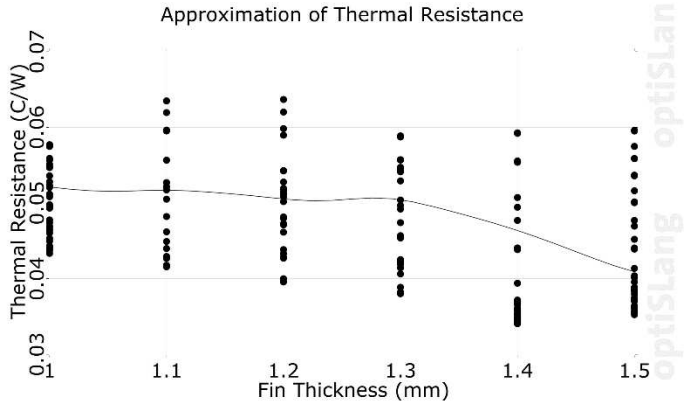

FIGURE 17: RELATION BETWEEN THERMAL RESISTANCE AND FIN THICKNESS

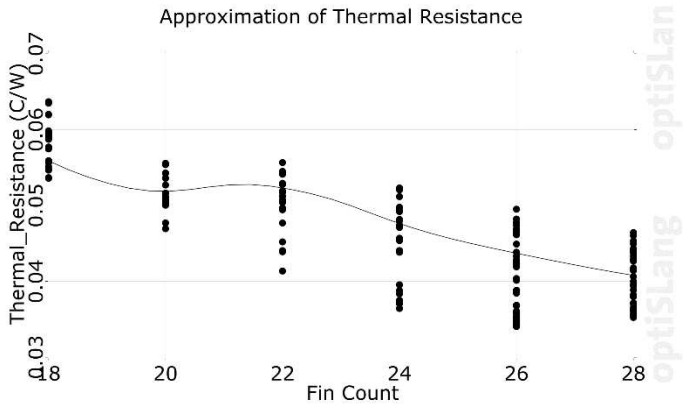

FIGURE 18: RELATION BETWEEN THERMAL RESISTANCE AND FIN COUNT FOR GPU

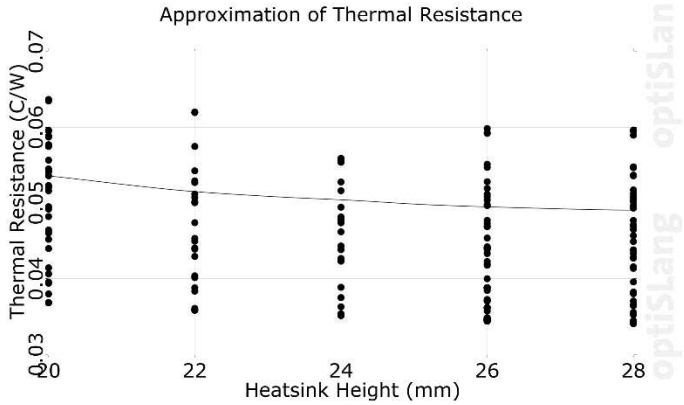

FIGURE 19: RELATION BETWEEN THERMAL RESISTANCE AND HEATSINK HEIGHT FOR CPU

The 5 best design points among the 180 design points are discussed in table 8. The best design point 96, showed the lowest thermal resistance of 0.034 °C/W which is ~33% lower than the baseline heatsink thermal resistance.

TABLE 8: COMPARSION OF BEST DESIGN POINTS WITH THE BASELINE PARAMETERS

Design Point Number	Fin Count	Fin Thickness (mm)	Heatsink Height (mm)	R_{th} (°C/W)	CPU Temp (°C)
Baseline	**20**	**1.5**	**20**	**0.051**	**65.42**
96	26	1.4	28	0.034	58.29
129	26	1.4	26	0.034	58.44
162	28	1.5	28	0.035	58.75
59	28	1.5	26	0.035	58.94
108	28	1.5	24	0.036	59.16

3.4 Optimization Results at Server Level

The baseline results without optimized heatsinks are compared with optimized heatsinks in a server setup. In the top tier of the server, where the two CPUs are located, the baseline heatsink was replaced by the best optimized heatsink that showed the lowest thermal resistance. Similarly, in the lower tier

of the server, the baseline GPU heatsinks were replaced by the best optimized heatsink. It was found out that case temperature of the CPUs was reduced by 3.5 °C from 72.3 °C to 69.75 °C which is about 4 % reduction. In case of GPUs, the front GPUs temperature saw a decrement of 5.5 °C from 86.88 °C to 81.38 °C which is reduction of 6.33 %. The rear GPUs saw a declination of 6 °C. This proves that the optimization of both the CPU and GPU heatsink in a test chamber helps reduce the temperature in an actual server setup.

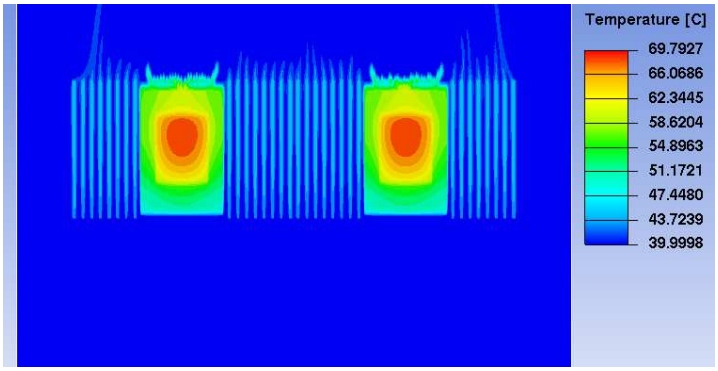

FIGURE 20: CPU CONTOUR WITH OPTIMIZED HEATSINKS

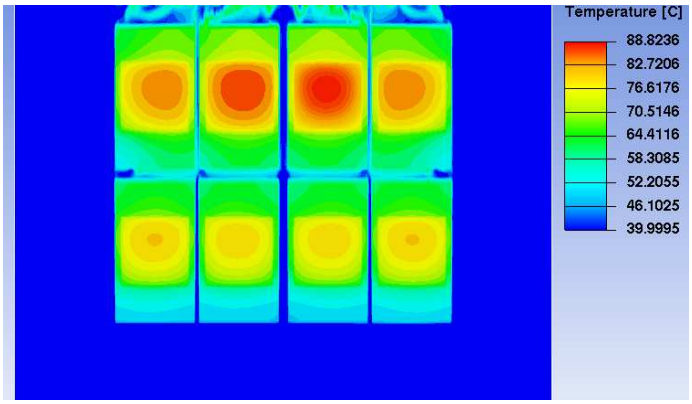

FIGURE 21: GPU CONTOUR WITH OPTIMIZED HEATSINKS

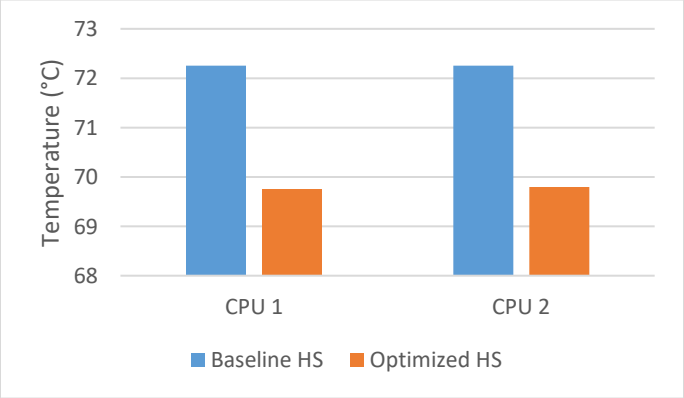

FIGURE 22: COMPARSION OF BASELINE HEATSINK WITH OPTIMIZED HEATSINK OF CPUs IN THE SERVER SETUP

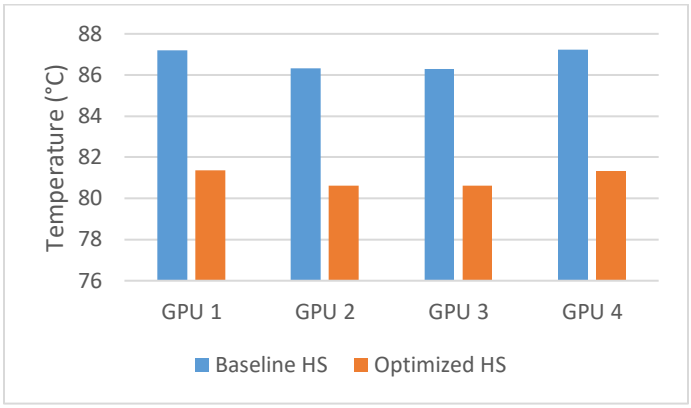

FIGURE 23: COMPARSION OF BASELINE HEATSINK WITH OPTIMIZED HEATSINK OF FRONT STACK GPUs IN THE SERVER SETUP

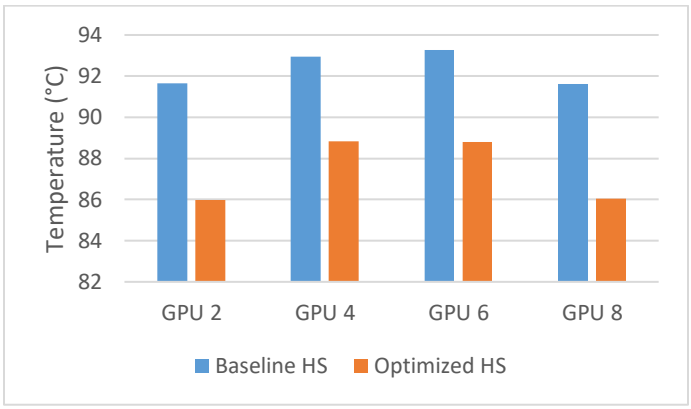

FIGURE 24: COMPARSION OF BASELINE HEATSINK WITH OPTIMIZED HEATSINK OF REAR STACK GPUs IN THE SERVER SETUP

4. CONCLUSION

With the increasing power densities of high-performance processors, the need for efficient cooling technologies has become more prominent. Single-phase immersion cooling has emerged as a promising solution, offering advantages over traditional air-cooling and liquid-cooling methods. It provides higher thermal mass, simplifies cooling infrastructure, mitigates airborne contamination concerns, and is well-suited for edge data center deployments. This study delves into the optimization of heat sinks in immersion-cooled servers, exploring various multi-objective and multi-design variable optimization schemes. The geometric parameters of the heat sink, such as heatsink height, fin thickness, and fin count, were systematically varied. The objective of the optimization study was to minimize thermal resistance and pressure drop while maintaining a constant pumping power. Compared to the baseline heat-sink design, the optimized heat sinks demonstrated significant improvements. In the CPU test chamber setup, the optimized heat sink achieved a 24% reduction in thermal resistance, while in the GPU test chamber setup, the reduction reached 33%, all while maintaining a lower pressure drop. These results highlight the enhanced

performance of the optimized heat sinks in both test chamber and real server setups.

ACKNOWLEDGEMENTS

This work is supported by NSF IUCRC Award No. 2209751.

REFERENCES

[1] K. Lee, "State of the network: Piecing together telecom trends in 2023," TeleGeography, https://blog.telegeography.com/state-of-the-network-piecing-together-telecom-trends-in-2023 (accessed May 18, 2023).

[2] Modi, Himanshu, Pardeep Shahi, Lochan Sai Reddy Chinthaparthy, Gautam Gupta, Pratik Bansode, Vibin Shalom Simon, and Dereje Agonafer. "Experimental Investigation of the Impact of Improved Ducting and Chassis Re-Design of a Hybrid-Cooled Server." In *International Electronic Packaging Technical Conference and Exhibition*, vol. 86557, p. V001T01A019. American Society of Mechanical Engineers, 2022.

[3] J. G. Koomey,"Growth in data center electricity use 2005 to 2010" (Analytics Press for the New York Times, 2011).

[4] ASHRAE TC 9.9, "2011 Thermal Guidelines for Data Processing Environments – Expanded Data Center Classes and Usage Guidance", ASHRAE, 2011

[5] Patterson, Michael K., Randall Martin, J. Barr Von Oehsen, Jim Pepin, Yogendra Joshi, Vaibhav K. Arghode, Robin Steinbrecher, and Jeff King. "A field investigation into the limits of high-density air-cooling." In *International Electronic Packaging Technical Conference and Exhibition*, vol. 55768, p. V002T09A013. American Society of Mechanical Engineers, 2013.

[6] Modi, Himanshu, Pardeep Shahi, Vibin Shalom Simon, Lochan Sai Reddy Chintaparthy, Gautam Gupta, Akiilessh Sivakumar, Satyam Saini, Pratik Bansode, and Dereje Agonafer. "Experimental Study of Improved Chassis and Duct Redesign for Air-Cooled Server." In *2023 22nd IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (Itherm)*, pp. 1-8. IEEE, 2023.

[7] Heydari, Ali, Pardeep Shahi, Vahideh Radmard, Bahareh Eslami, Uschas Chowdhury, Akiilessh Sivakumar, Akshay Lakshminarayana et al. "Experimental Study of Transient Hydraulic Characteristics for Liquid Cooled Data Center Deployment." In *International Electronic Packaging Technical Conference and Exhibition*, vol. 86557, p. V001T01A009. American Society of Mechanical Engineers, 2022.

[8] Kheirabadi, Ali C., and Dominic Groulx. "Cooling of server electronics: A design review of existing technology." *Applied Thermal Engineering* 105 (2016): 622-638.

[9] Ellsworth, M. J., L. A. Campbell, R. E. Simons, M. K. Iyengar, R. R. Schmidt, and R. C. Chu. "The evolution of water cooling for IBM large server systems: Back to the future." In *2008 11th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, pp. 266-274. IEEE, 2008.

[10] Iyengar, Madhusudan, Milnes David, Pritish Parida, Vinod Kamath, Bejoy Kochuparambil, David Graybill, Mark Schultz et al. "Server liquid cooling with chiller-less data center design to enable significant energy savings." In *2012 28th annual IEEE semiconductor thermal measurement and management symposium (SEMI-THERM)*, pp. 212-223. IEEE, 2012.

[11] Modi, Himanshu, Pardeep Shahi, Akiilessh Sivakumar, Satyam Saini, Pratik Bansode, Vibin Shalom, Amrutha Valli Rachakonda, Gautam Gupta, and Dereje Agonafer. "Transient CFD Analysis of Dynamic Liquid-Cooling Implementation at Rack Level." In *International Electronic Packaging Technical Conference and Exhibition*, vol. 86557, p. V001T01A012. American Society of Mechanical Engineers, 2022.

[12] Patterson, Michael K., Shankar Krishnan, and John M. Walters. "On energy efficiency of liquid cooled HPC datacenters." In *2016 15th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (Itherm)*, pp. 685-693. IEEE, 2016.

[13] Daniel, Abishai, and Nishi Ahuja. "Impact of data center cooling strategies on component reliability." In *2014 Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM)*, pp. 197-201. IEEE, 2014.

[14] A. Bar-Cohen, M. Arik and M. Ohadi, "Direct Liquid Cooling of High Flux Micro and Nano Electronic Components," in *Proceedings of the IEEE*, vol. 94, no. 8, pp. 1549-1570, Aug. 2006, doi: 10.1109/JPROC.2006.879791.

[15] El-Genk, M. "Nucleate boiling enhancements for immersion cooling of high power electronics." In *2010 3rd International Conference on Thermal Issues in Emerging Technologies Theory and Applications*, pp. 5-5. IEEE, 2010.

[16] Osman, O. S., R. M. El-Zoheiry, M. Elsharnoby, and S. A. Nada. "Performance enhancement and comprehensive experimental comparative study of cold plate cooling of electronic servers using different configurations of mini-channels flow." *Alexandria Engineering Journal* 60, no. 5 (2021): 4451-4459.

[17] Koito, Yasushi, Hideaki Imura, Masataka Mochizuki, Yuji Saito, and Shuichi Torii. "Numerical analysis and experimental verification on thermal fluid phenomena in a vapor chamber." *Applied Thermal Engineering* 26, no. 14-15 (2006): 1669-1676.

[18] Bansode, Pratik V., Jimil M. Shah, Gautam Gupta, Dereje Agonafer, Harsh Patel, David Roe, and Rick Tufty. "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* 142, no. 1 (2020): 011010.

[19] Gupta, Gautam. *Experimental Analysis of A Single-Phase Direct Liquid Cooled Server Performance at Extremely Low Temperatures for Extended Time Periods*. The University of Texas at Arlington, 2018.

[20] Sivaraju, Krishna Bhavana, Pratik Bansode, Gautam Gupta, Jacob Lamotte-Dawaghreh, Satyam Saini, Vibin Simon, Joseph Herring, Saket Karajgikar, Veerendra Mulay, and Dereje Agonafer. "Comparative Study of Single-Phase Immersion Cooled Two Socket Server in Tank and Sled Configurations." In *International Electronic Packaging Technical Conference and Exhibition*, vol. 86557, p. V001T01A010. American Society of Mechanical Engineers, 2022.

[21] Bansode, Pratik V., Jimil M. Shah, Gautam Gupta, Dereje Agonafer, Harsh Patel, David Roe, and Rick Tufty. "Measurement of the thermal performance of a single-phase immersion cooled server at elevated temperatures for prolonged time." In *International Electronic Packaging Technical Conference and Exhibition*, vol. 51920, p. V001T02A010. American Society of Mechanical Engineers, 2018.

[22] Murthy, Prajwal, Gautam Gupta, Joseph Herring, Jacob Lamotte-Dawaghreh, Krishna Bhavana Sivaraju, Pratik Bansode, Himanshu Modi, Dereje Agonafer, Poornima Mynampati, and Mike Sweeney. "CFD Simulation-Based Comparative Study of Forced Convection Single-Phase Liquid Immersion Cooling for a High-Powered Server." In *International Electronic Packaging Technical Conference and Exhibition*, vol. 86557, p. V001T01A006. American Society of Mechanical Engineers, 2022.

[23] Saini, Satyam, Tushar Wagh, Pratik Bansode, Pardeep Shahi, Joseph Herring, Jacob Lamotte-Dawaghreh, Jimil M. Shah, and Dereje Agonafer. "A Numerical Study on Multi-Objective Design Optimization of Heat Sinks for Forced and Natural Convection Cooling of Immersion-Cooled Servers." *Journal of Enhanced Heat Transfer* 29, no. 8 (2022).

-