Liquid Cooling of Compute System

The continued demand for increasing compute performance results in an increasing system power and power density of many computers. The increased power requires more efficient cooling solutions than traditionally used air cooling. Therefore, liquid cooling, which has traditionally been used for large data center deployments, is becoming more mainstream. Liquid cooling can be used selectively to cool the high power components or the whole compute system. In this paper, the example of a fully liquid cooled server is used to describe different ingredients needed, together with the design challenges associated with them. The liquid cooling ingredients are cooling distribution unit (CDU), fluid, manifold, quick disconnects (QDs), and cold plates. Intel is driving an initiative to accelerate liquid cooling implementation and deployment by enabling the ingredients above. The functionality of these ingredients is discussed in this paper, while cold plates are discussed in detail. [DOI: 10.1115/1.4042802]

Introduction

A cooling solution is needed to keep power components in electronic applications under the required temperature limits. Traditionally, the cooling solution has been air cooling using fans and heat sinks. The heat sink is generally made out of copper and/or aluminum, and it transfers the heat from the high power components to the air. The fan is used to generate a significant airflow for increased heat transfer removal from the hot components. However, air cooling has its limitations, and as the power of components increase, a heat sink and fan can no longer meet the thermal requirements [1].

Liquid cooling is more efficient than air cooling in removing heat [2] and can be used to meet the increasing cooling demand, since liquid has higher heat capacitance and thermal conductivity properties than air. In the liquid cooling scenario, cold plates are used instead of heat sinks, and pumps are used to circulate the cooling medium/liquid instead of fans. Liquid cooling for electronic systems is not a new technology, it has been used for quite some time. Liquid cooling was initiated in the mid-1960s on large mainframe computers where high heat dissipation components using bipolar technology were water cooled to meet the temperature requirements [3–6]. The complementary metal-oxide-semiconductor (CMOS) transistor technology was introduced and adopted in the 1990s by the computer industry. This significantly reduced the heat dissipation and made it more cost effective to cool compute systems with air. However, the heat dissipation of CMOS components has been increasing over the years, and dissipation levels are again reaching the limits of air cooling [3–6]. Therefore, recently a refocus on liquid cooling technologies and more liquid cooled computer solutions are being reported [5–9].

Also, thermal design guidelines for liquid cooling have been published by ASHRAE [10]. The change from the early days is that liquid cooling is becoming more mainstream due to the increasing heat dissipation/powetrends and the inherent limitations with air cooling not allowing for higher density, efficiency, and performance systems.

Liquid cooling can be deployed to cool high power components selectively or the whole system. If the high power parts are liquid cooled, the rest of the system can generally be air cooled. This is a hybrid cooled system, and it requires both fans and pumps [11]. This can be an advantage if the environment around the compute system is already setup to handle heat removal through air. If the whole compute system is liquid cooled, all power components need to have a thermal pathway to a cold plate. This adds complexity to the plate designs, while the advantage is a more efficient cooling solution where all heat is removed through the liquid resulting in a relatively quiet thermal solution.

This paper uses a high-performance compute server (HPCS) as an example to discuss the liquid cooling ingredients needed to design a fully liquid cooled system. It is important to note that the whole system needs to be considered when designing a fully liquid cooled solution, and that there are different design constraints to consider when designing a fully liquid solution compared to traditional air cooling. This will become apparent in the discussions below.

The ingredients that make up a liquid cooling solution for the server are discussed along with their design challenges. The ingredients are: (1) a cooling distribution unit (CDU), which pumps the cooling liquid and exchange the heat to another liquid or air, (2) a fluid, which is the cooling medium, (3) a manifold, which distributes the cooling liquid to the server, (4) quick disconnects (QDs), which allow for disconnecting the server from the manifold for serviceability, and (5) cold plates, which transfer heat from components to the cooling liquid, with one or several fluid loops. Intel is driving a liquid cooling initiative to enable ingredients through suppliers to reduce the risk and accelerate implementation and deployment of liquid cooling at the server rack level. The enabling of common and compatible ingredients is the key. An
Overview of Liquid Cooling Ingredients

In the example of the HPCS application, which is used in this paper, the flow of the liquid cooling starts in the CDU and is transported to the server through the cold side manifold. The cooling fluid then passes through the QDs and enters the fluid loop on the server. The cold fluid is first directed to the central processing unit (CPU) cold plates, since the CPUs are the highest power components in this system. Thereafter, the fluid goes through the rest of the fluid loop to cool the lower power board components through board cold plates and other component solutions. The hot fluid then exits the server cooling loop through a QD on the warm side. Then, it is transported through the warm side manifold to the CDU, where the fluid is cooled. This loop continues to ensure cooling of the HPCS. An example of a liquid cooled installation concept is shown in Fig. 1.

Cooling Distribution Unit

The function of the CDU is to pump the fluid through the liquid cooling solution of the HPCS and to exchange the heat captured with another cooling medium. The CDU can be either local in relation to the HPCS or global servicing several HPCS units as seen in Fig. 1. Examples of external to the HPCS cooling media are facility air or facility liquid, which describes a liquid-to-air CDU or a liquid-to-liquid CDU, respectively. Some of the challenges with CDUs are the space constraints in compute systems and material compatibility with the cooling fluid. Space needs to be allocated for smaller scale local CDUs or larger scale global CDUs. Depending on the size of the liquid cooled installation, tradeoff analysis is needed to determine the type of CDU preferred. Material compatibility between the cooling fluid and the whole fluid loop (including the CDU) is also needed to avoid any potential issues over time such as corrosion.

Manifold

Manifolds provide the fluid supply and return between the liquid cooled unit and the CDU. There are a number of design parameters to consider when sizing the manifolds for the fluid loop. These include flow rate, fluid velocity, pressure drop, and maldistribution. The HPCS has a specific flow rate requirement necessary to cool the processors and other components. Maintaining a relatively low fluid velocity is important. ASHRAE [10] recommends not exceeding a maximum velocity of 1.8 m/s within the manifold to avoid potential erosion issues in the cooling loop.

Quick Disconnects

Quick disconnect fluid coupling valves are needed to quickly disconnect the system from the supply and return manifolds to remove or service it. In the example of the HPCS, it is critical that the systems be hot pluggable to minimize any downtime associated with servicing the equipment. Hot pluggable means that the server can be disconnected and reconnected while the system is still operating. The QDs must, therefore, be capable of shutting off both ends of the coupling without dripping when disconnecting and need to be double shutoff valves. In addition to the challenge of serviceability, there are several other engineering challenges to introduce this valve in a fluid loop including pressure drop, finger access, and termination types.

The pressure drop across the QD must be factored into the system design due to flow constrictions through the valve. QD suppliers typically publish a flow coefficient ($C_v$) rating. The $C_v$ rating is a relative measure of the QDs efficiency of fluid flow. The parameter $C_v$ is the flow rate ($Q$) multiplied by the square root of the specific gravity (SG) of the fluid medium divided by the pressure differential ($AP$) as shown in Eq. (1). The higher the $C_v$ rating, the
lower the pressure drop will be across the QD for a given flow rate

\[ C_v = \sqrt{\frac{SG}{\Delta P}} \] (1)

From a system design perspective, the higher \( C_v \), valves, i.e., lower pressure drop, will be advantageous because the resistance to flow will be lower and can affect overall pump sizing. Note that when the QD will be used in both flow directions, pressure drop data need to be provided for both flow directions to accurately model the system.

Serviceability also includes the ease of deployment, replacement, and upgrades. QDs simplify the process of extracting the servers from the rack. However, each QD supplier has a proprietary interface. This requires the engineers to source equipment with the specific suppliers QD mating parts. This excludes multi-sourcing of QDs and presents a challenge to the server manufacturers.

In the liquid cooling initiative, Intel is working on making QDs common and compatible for fluid loop and equipment suppliers. Intel is working with partners to develop a specification for a server universal quick disconnect (UQD) that will be fully interchangeable. The server UQD design is shown in Fig. 2.

**Fluid**

The fluid used in the cooling loop can be either glycol-based or nonglycol based water. The reason for selecting one or the other depends on the system requirements. If the requirement is to support system shipping temperatures as low as \(-40^\circ C\) without causing the cooling fluid to freeze, the glycol-based fluid should be selected. However, if the system will not be shipped charged (i.e., with the cooling fluid), then the nonglycol based water can be used.

Material compatibility between the fluid and the cooling loop materials (i.e., wetted materials) is required to reduce any long-term issues with corrosion. Many different fluids with different chemistry can be used as long as material compatibility is ensured.

Maintenance are often associated with the use of a cooling fluid in terms of monitoring the health of the fluid to ensure no issues with bacterial growth, corrosion monitoring for material degradation, and/or the potential presence of large scale particles in the fluid loop. During the use of the fluid, additives might need to be added and the fluid quality monitored according to a predetermined schedule.

Corrosion can cause issues in the fluid loop and clog up the fluid pipes and cold plates. This can cause a reduced flow in sections of the fluid loop and potentially leading to thermal issues. The reduced flow issue can also occur if large scale particles are dislodged and accumulated. Therefore, the fluid loop might need to include a filter to capture such potential particles. This filter needs to be checked and replaced at regular intervals.

**Cold Plates**

Cold plates are used to provide a thermal path between hot components and the liquid. Cold plate designs can consist of a metal design with an incorporated fluid loop, where the cold plate is attached to the hot components through a thermal interface material (TIM). For a fully liquid cooled system, cold plate designs are needed for both high and low power components, since a sufficient conduction heat path and convective area enhancement are needed for every heat source within the system. The cold plate design is, therefore, dependent on the specific server/board design and will also be specific for high versus low power components to meet the thermal cooling requirement.

Examples of cold plate designs for the HPCS are one design for the high power component CPU and others for all lower power board components. The CPU cold plate can be a metal enclosure with internal microchannel fins to increase the surface area to the liquid and thereby increasing the heat transfer rate. The cold plate for lower power board component can be a metal plate with an incorporated fluid loop. The fluid loop needs to be designed to be located over any critical components with stringent temperature requirement, and the metal plate to touch down on components through a TIM to provide a heat transfer path. A separate cold plate design might be needed for components that are required to be easily serviceable and/or replaced in the field. It is important to note that the thermal performance analysis needs to include variability to ensure that the thermal requirements can be met.

This design objective presents four primary types of engineering challenges:

1. Design for high power components, where component performance is thermally limited and the best achievable cooling capability is needed (High Power Components section).
2. Design for components that must be field replaceable, where operator access and a separable thermal interface must be considered (Field Replaceable Components section).
3. Design for low power components, where traditional air cooling would provide sufficient cooling without any...
additional design consideration. In the absence of airflow, these components can overheat and contribute to the challenge of cooling adjacent components (Low Power Components section).

(4) Design for variance of flow distribution, wherein manufacturing variation contributes to the balance of fluid flow within parallel paths of the cooling network (Variability Analysis section).

High Power Components. For the case of thermally limited components, a great deal of design optimization work has been done around microchannel fin designs [12,13] and the use of two-phase coolant [14]. These design features can be shown to generate very high cooling capability. However, in many cases, the facility infrastructure limits the cooling design to single-phase coolant with larger fin channels that account for the risk of particle build-up that will clog the fluid path.

In the HPCS example, the fluid system filter is set to ~40 µm, with an industry guideline of 10x, the minimum fluid path dimension requirement is given as ≥400 µm. With these design constraints, two parallel plate fin design cold plates were investigated: one cross flow (Fig. 3) and one center flow cold plate (Fig. 4).

In this example, a CPU power of 350 W is used and a cold plate size of 83 × 82 × 9.3 mm with a flow rate of 0.38 gpm. The fluid temperature is 34 °C when entering the cold plate. The center flow cold plate fins are 0.4 mm thick, while the fins in the center flow cold plate are 1 mm. The increased thickness of the fins in the center flow cold plate is driven by structural requirement to withstand the load required. The cross flow cold plate is designed with a center load support feature, which is not feasible in the center flow design due to the flow. The gap between fins is 0.4 mm for both cold plate designs to adhere to the industry guideline mentioned earlier. The dimensions of the two cold plate designs are summarized in Table 1.

Thermal predictions of the two cold plate designs were performed using the computational fluid dynamics (CFD) software ANSYS ICEPAK. The cross flow cold plate performed better than the center flow one with a lower case temperature and lower pressure drop as shown in a relative comparison in Table 2. This is an example of performance analysis needed together with cost and high volume manufacturability for design tradeoff analysis. Here, the necessary cooling performance could be achieved using the cross flow parallel plate fin design in a copper cold plate shown in Fig. 3.

This cold plate design architecture requires fins to be cut or skived out of the copper base and a copper cap over the top to encapsulate the fluid channels. It is important to minimize the bypass flow over the fins to ensure max efficiency of the cold plate. All materials used must be considered for compatibility with the working fluid and corrosion inhibition. This optimized design gives an effective convection coefficient of 25,000 W/m²·K for the area of the fin bank.

Field Replaceable Components. For components that must be field replaceable, the cooling design must enable access and a separable thermal interface between the replaceable component and the system without adding risk to fluid leaks where cooling loop break points are used. In the case of dual inline memory module (DIMM) cooling on the HPCS system, this also presents the challenge of creating a cooling solution that fits within typical server DIMM pitch of 10 mm.

In Fig. 5, the field replaceable design shown utilizes a flattened heatpipe on each side of the DIMM to transport heat to cold blocks at each end of the DIMM. Here, the heatpipes interface the cold block through a reusable or replaceable gap pad TIM. Minimal contact thermal resistance is achieved with a clamp that maintains pressure along the interface, since thermal resistance decreases with pressure. The use of the flattened heatpipe enables separation between the replaceable component and the cooling loop without the necessity of a break in the fluid path and preserves the minimum DIMM spacing. However, this design strategy adds an additional thermal interface that reduces overall thermal performance. It also requires that the DIMM heat spreader and heatpipe assembly must be replaced with the DIMM. In this HPCS example, the target DIMM power to be cooled is 18 W (9 W per cold block). The cold blocks are downstream of the high power CPU components and the corresponding caloric temperature rise brings reference fluid temperature for these cold blocks to 48 °C. At the interface of the replaceable gap pad TIM, an average surface temperature of ≤ 62 °C is maintained to enable the 18 W thermal design power per DIMM. The cooling capability can be enhanced further by extending the length of the heatpipes and the cold blocks to increase the replaceable gap pad TIM contact area and reduce thermal resistance between the cold block and the reference fluid itself. A thermal solution is needed for all DIMMs in a fully liquid cooled system, but the cooling design can potentially be modified for low versus high power DIMMs. Some

<table>
<thead>
<tr>
<th>Table 1</th>
<th>CPU cold plate dimensions and conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cold plate designs</td>
<td>Cross flow</td>
</tr>
<tr>
<td>Manifold dimensions (mm)</td>
<td>5 × 10</td>
</tr>
<tr>
<td>Fin dimensions (mm)</td>
<td>0.4 × 5 × 57</td>
</tr>
<tr>
<td>Fin gap (mm)</td>
<td>0.4</td>
</tr>
<tr>
<td>Base thickness (mm)</td>
<td>1.5</td>
</tr>
<tr>
<td>Reynolds number</td>
<td>145</td>
</tr>
<tr>
<td>Total fin area (m²)</td>
<td>0.04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>CPU case temperature and cold plate pressure drop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cold plate designs</td>
<td>Cross flow</td>
</tr>
<tr>
<td>Temperature difference between CPU case and fluid inlet (°C)</td>
<td>17.4</td>
</tr>
<tr>
<td>Cold plate pressure drop ratio</td>
<td>1</td>
</tr>
</tbody>
</table>
of the design tradeoffs to consider are heat pipe on one or both sides of the DIMM, and heat pipe attach to the cold block design, i.e., cold block temperature required, to meet the DIMM temperature requirements. These are examples of design tradeoffs that need to be analyzed for each system solution.

Low Power Components. By eliminating the need for open airflow pathways through the system, liquid only cooled systems provide a significant density advantage. However, without any appreciable airflow, low power electrical components that would otherwise be inconsequential to the thermal design are now critical for design consideration.

In Fig. 6, a comparison of an air cooled board (top) and a liquid only (bottom) cooled board illustrates the heat path for various components. Figure 6 shows a primary component that is directly cooled via a heatsink or the liquid cold plate and adjacent low power components that are not directly cooled. In the case of an air cooled system, the airflow over the board is sufficient to remove a significant amount of heat from low power adjacent components. However, in the liquid only cooled design, the power from indirectly cooled components must primarily transfer to the cooling fluid by conducting through the board and up through the nearest directly cooled component. This creates a high resistance heat transfer path for the indirectly cooled components and effectively adds to the thermal load of the directly cooled component.

In the example illustrated in Fig. 6, the additional power transferred through the directly cooled components adds \( \frac{10 W}{60 W/m \cdot K} \) thermal design power, which contributes to underestimation of component temperature for the directly cooled component if not accounted for in the thermal design.

In a typical example of a liquid cold plate, board component cooling is accomplished by contacting selected components across the board through a gap pad TIM to a simple aluminum plate with a fluid heat transfer path routed across the plate. Components selected for direct cooling contact are prioritized based on power dissipation and upper operating temperature limit. It may not be necessary or cost effective to provide cold plate contact for all components on the board. Therefore, a threshold for power density must be established to ensure adequate cooling.

Figure 7 showcases a generalized thermal model of an area of the board with a directly cooled component and a set of adjacent low power components without a direct thermal path to the cold plate. The thermal models described below were generated using the CFD software ANSYS ICEPAK v16. The model is symmetric along the right edge in Fig. 7. This right edge represents the centerline of the directly cooled component. This model assumes a high board conductivity of 60 W/m \cdot K in plane and 6 W/m \cdot K orthogonal to the board surface and no radiation heat transfer. The directly cooled component has a fixed power of 10 W and a fixed effective heat transfer coefficient on the top surface to represent the cold plate thermal resistance. The power of the indirectly cooled components is all equal and swept through a sensitivity range along with the effective convection heat transfer from the surface of these components and the board.

Figure 8 shows the impact of reduced airflow heat transfer to the indirectly cooled components. As an example, a 5 \( \times \) 5 mm field-effect transistor (FET) is used as a basis with a functional temperature limit of 105 °C. In the analysis, some amount of natural convection heat exchange between the board and the cold plate...
is assumed to be 1–2 W/m²·K. This results in a threshold range of 0.3–0.4 W for a liquid only cooled design. By comparison, a typical air-cooled server will have an effective convection heat transfer rate on the surface of the board of 10–20 W/m²·K or greater, and this corresponds to a 1.0–1.3 W limit for FET power dissipation with indirect cooling design. This is a significant difference and the lower power threshold for the liquid only cooled system will encompass a much larger number of components that require thermal design consideration. These results are specific to the model assumptions stated earlier and will change significantly between different system designs, but demonstrates an approach to determine the power threshold. In the generalized example used here, with a low effective heat transfer coefficient of 2 W/m²·K on the board surface, the sensitivity to in-plane board conductivity and the distance between the indirectly cooled component and the directly cooled component are relatively low as shown in Fig. 9. However, this will also be dependent on the target system design.

In the example of the HPCS, the impact of not directly cooling a pair of 0.15 W 3 × 3 mm FETs is shown in the board temperature plot in Fig. 10. Here, without direct cooling, the component temperature exceeds the maximum operating temperature limit of 105°C despite relatively low component power and nearby directly cooled adjacent components. This is due to several factors: (1) the nearest directly cooled components are already operating at a relatively high temperature, (2) there is a long conduction path through the board to the nearest low temperature directly cooled component, and (3) the local power delivery losses in the board contribute to the high component temperature. This illustrates the importance of establishing a threshold power for
Fig. 9  Indirectly cooled FET junction temperature as a function of board plane conductivity and distance to the directly cooled component

Fig. 10  Thermal contour of predicted temperatures for an area of an HPCS board
with a complete background covered by Xiu [17], the state of the art covered by Eldred [18] and Eldred and Burkardt [19], and examples of cooling solution optimizations under uncertainty by Bodla et al. [20]. At a fundamental level, polynomial chaos expansion works by fitting a polynomial response surface to the numerical model for the input parameters that are varied. This polynomial response surface can then be used as a basis for a random sampling method, such as Monte Carlo, without the burden of long solve times. Using the response surface as a stand-in for the numerical model does add some amount of error in the predicted distribution of the system performance. However, the method of generalized polynomial chaos expansion minimizes this error by choosing the orthogonal polynomial series with the weighting function that matches the probability density function of the input variables as the basis for the response surface. Xiu [17] shows (Sec. 3.3.2) that this method results in an exponential convergence in the model approximated by a polynomial response surface and one based on the original numerical model as the polynomial order is increased for a function that is sufficiently smooth.

The application of polynomial chaos expansion can be applied with any one of a number of available software packages. At the time of writing, the most well-known packages are DAKOTA, OpenTurns, UQLab, and ChaosPy, among others. The results shown here are produced using ChaosPy [21] and the CFD software ANSYS ICEPAK v16.0 as the numerical model, but similar results can be obtained using any of the other packages listed earlier.

In the example of the HPCS, the board cooling cold plate design consists of a flattened copper tube that routes fluid in a series of parallel pathways across an aluminum plate that is thermally connected to various components across the board. This design is a cost effective method for cooling many low power components. However, the pressure drop through each leg of the fluid loop is sensitive to the variation in the flattening process for each of the tube components. The flattening of the tube is used to alter the pressure drop of the fluid loop to ensure that the desired flow distribution is achieved. An example of an HPCS fluid loop solution for cooling board components and pipe flattening used in that solution is shown in Fig. 11.

To demonstrate the process of implementing general polynomial chaos, an investigation of a section of the flattened tube in the board cold plate fluid loop, where the outer diameter of the flattened height and the tube wall thickness are chosen as the input parameters of interest to vary. A second-order orthogonal polynomial is used to approximate the response curve and a Normal distribution for each of the inputs described is assumed.

**Variability Analysis.** It is important to take into account the variability of thermal performance of the liquid cooled ingredients to ensure that the thermal requirements are still met. For example, one of the ingredients that show a large variance in part geometry and material properties due to manufacturing variation and use case application is the fluid loop. This variance results in a distribution of expected system and component performance. In liquid cooled systems, there are often many parallel fluid pathways that must maintain an expected range of flow in order to ensure expected thermal performance.

Assessing performance variation without building and testing, a large number of components can be accomplished using statistical methods or random sampling models given that a simple mathematical model exists to evaluate the impact of variation in these input parameters. However, for fluid flow and pressure drop evaluation of complex designs that rely on CFD modeling, the duration of model solve times prohibits random sampling methods due to the many thousands of model runs needed to achieve convergence. Fortunately, recent advances in the field of uncertainty quantification have developed methods based on general polynomial chaos expansion that enable random sampling methods to be used with only 10’s of numerical model solutions. The detailed mathematical basis for these methods is not discussed here; the reader can find a high level overview by Barth [15] and Perez [16].

---

**Table 3 Input parameters**

<table>
<thead>
<tr>
<th>Input dimension variance</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flattened tube height</td>
<td>4.6 mm</td>
<td>0.05 mm</td>
</tr>
<tr>
<td>Tube wall thickness</td>
<td>0.7 mm</td>
<td>0.033 mm</td>
</tr>
</tbody>
</table>

**Table 4 Model solutions needed**

<table>
<thead>
<tr>
<th>Model</th>
<th>Tube height (mm)</th>
<th>Wall thickness (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.51</td>
<td>0.64</td>
</tr>
<tr>
<td>2</td>
<td>4.60</td>
<td>0.64</td>
</tr>
<tr>
<td>3</td>
<td>4.69</td>
<td>0.64</td>
</tr>
<tr>
<td>4</td>
<td>4.51</td>
<td>0.70</td>
</tr>
<tr>
<td>5</td>
<td>4.6</td>
<td>0.70</td>
</tr>
<tr>
<td>6</td>
<td>4.69</td>
<td>0.70</td>
</tr>
<tr>
<td>7</td>
<td>4.51</td>
<td>0.76</td>
</tr>
<tr>
<td>8</td>
<td>4.60</td>
<td>0.76</td>
</tr>
<tr>
<td>9</td>
<td>4.69</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Fig. 11 Example of an HPCS fluid loop solution for cooling board components and pipe flattening used.
Within ChaosPy, the Golub-Welsh algorithm [19] is used to determine the Gauss quadrature points and weights based on the input parameters shown in Table 3. This yields nine model solutions needed to define the response curve polynomial coefficients, as shown in Table 4.

Updating the CFD model of the flattened tube section with each of the geometries listed in the table above gives a solution for each model that can be fit to the second-order response curve. This can be done with any number or performance results of interest. Even the flow vector field for the entire model can be fit to the response curve, if the resulting flow vectors can be defined independently of mesh changes between model geometry. For this example, the pressure drop across the tube section is considered for a range of flow rates. Once a response curve is created, it can be applied to a statistical random sampling model, such as Monte Carlo, to generate a distribution of predicted pressure drop for a large number of parts based on the known manufacturing variation.

Figure 12 shows the resulting pressure drop distribution based on the variance in the input parameters shown in Table 3. This shows that, at the intended operating flow rate of 0.75 gpm, even maintaining ±3σ geometric manufacturing tolerances of ±150 μm on the tube flattening operation and ±100 μm on the wall thickness results in ±19% swing in pressure drop through this tube section. These results can then be used as inputs for evaluating the expected thermal performance at the extremes of the expected flow conditions. Furthermore, using the fitted response surface equation, the impact of each of the parameters can be compared using sensitivity analysis. Based on the distributions used, it can be calculated that the variance in the wall thickness accounts for 70% of the total variation in the pressure drop. The flattened height of the tube accounts for 26.3% and the remaining 2.7% of the variance is due to interaction between the tube height and wall thickness varying together.

Essentially, the generalized polynomial chaos expansion methodology does not enable a broad application of stochastic computation for all design parameters in a large system design. However, it does enable design studies for targeted parameter sensitivity and estimating variance in system performance for engineering designs that require numerical models to analyze a discrete solution point. This adds a significant capability when designing for large-scale liquid cooling systems such as an HPCS solution.

**Conclusion**

Liquid cooling offers increased thermal capacity allowing for higher density, efficiency, and performance systems compared to traditional air cooling. Liquid cooling is quickly becoming more mainstream due to the continued demand for increasing compute performance, and therefore, increasing system power. The liquid cooling ingredients needed for a fully liquid cooled thermal solution are CDU, manifold, QDs, fluid, and cold plates.

To enable liquid cooling ingredients, Intel is driving a liquid cooling initiative with the intent to reduce the risk and accelerate the implementation and deployment of manifold distributed liquid cooling at the server rack level. These enabled ingredients include in-rack CDUs, server UQD, multiple suppliers for glycol-based cooling fluid, CPU cold plates, and design guidelines for manifolds. These are all common ingredients that can be used for multiple liquid cooled solutions.

General enabling of ingredients that is solution specific such as cold plates may not always be feasible. Therefore, analysis methods for evaluating cold plate solution designs are demonstrated. The cold plate examples given are design analysis of high and low power components, and field replaceable units. Two high power component cold plate designs are numerically evaluated for performance benefits. For low power components, the analysis specifically shows the importance to determine power thresholds and distance to nearest directly cooled component for a fully liquid cooled solution to ensure that the temperature requirements will be met. Design considerations for a thermal solution that needs to be field replaceable are discussed. A variability analysis is performed to highlight the need of taking manufacturability variances and tolerances into account when evaluating cooling solutions to ensure that the expected thermal performance can be achieved. These methodologies are all helpful to analyze and characterize liquid cooled compute designs.

**Acknowledgment**

The authors want to thank Emery Frey, Juan Cevallos, Roger Flynn, and Joseph Broderick for thermal analysis and design of CPU cold plates. Suchismita Sarangi for fluid loop analysis, Phi Thanh for DIMM investigation, Brian Jarrett for board cold plate design, and Jason Chesser for review.

**Nomenclature**

- \( C_v \) = flow coefficient (US gpm/psi\(^{1/2}\))
- \( Q \) = flow rate (US gpm)
- \( SG \) = specific gravity
- \( T_j \) = junction temperature (°C)
- \( \Delta P \) = pressure difference (psi)
- \( \Delta T \) = temperature difference (°C)

**References**
