Experimental Study of Two-Phase Cooling to Enable Large-Scale System Computing Performance

Devdatta Kulkarni, Xudong Tang, Sandeep Ahuja, Richard Dischler, Ravi Mahajan
Intel Corporation
2111 NE 25th Ave.
Hillsboro, OR, U.S.A. 97124
Email: devdatta.p.kulkarni@intel.com

ABSTRACT

High Performance Computing (HPC) performance demands, which usually waterfall into mainstream servers later in time, show increasing need for higher power and hence pose cooling challenges for succeeding generations of compute elements. New generation processors have higher core count, multiple dies, and higher power to achieve generational performance CAGR. Typical high performance compute platforms have dual-socket boards that consume approximately 150-300 watts per socket. Each socket is directly fan-cooled or single phase liquid cooled. The thermal challenge is to cost-effectively increase the thermal envelope to enable greater performance and a target of doubling the power dissipation per socket. In order to push toward large scale system performance via thousands of compute nodes, there is a need to bring several high power processor nodes together with a fabric and a switch chip. Placing them on a single cost-effective motherboard enables high bandwidth fabric signaling and dense packaging, but with a high thermal load. A solution to this is transporting the heat away with two-phase liquid because it enables cooling of the silicon effectively and to a much lower junction temperature. In this paper, conclusive results from a prototype are shown using high efficiency Pumped Liquid Multiphase Cooling technology. 2kW of heat is moved from ten 200W devices (in series) while maintaining low, matched junction temperatures. The thermal management solution is designed for 1U height, keeping low weight as design constraint. The results from pumped two phase system are compared with single phase glycol water system as well and demonstrated advantages of the two phase system.

KEYWORDS: high power thermal management, Pumped Liquid Multiphase Cooling, large scale system performance, energy efficiency

NOMENCLATURE

T	temperature, °C
Psi	thermal resistance, °C/W

Subscripts

case	case temperature

INTRODUCTION

In High Performance Computing (HPC) community, there is a strong push to achieve exascale (1EFLOPs/S) performance from a computing system. As per the report by US DoE [1], target is to achieve exascale performance with 20MW power. Figure 1 below shows one example of exascale architecture presented by Barton [2]. As seen in this figure, there is a cluster of compute nodes which can range more than 50,000 or so which is connected via fabric and switches. The DOE report also outlines that to get the improvement in power efficiency, several technical areas in hardware design for exascale system need to be explored. These include, energy efficient hardware building blocks such as CPU, memory and interconnect, novel cooling and packaging as Si-photonic communication.

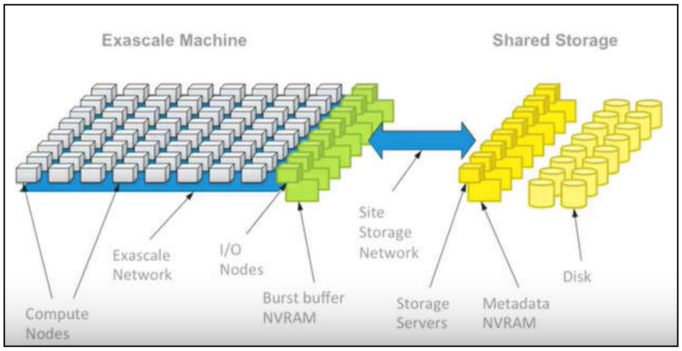

Fig. 1. Exascale I/O Architecture (taken from [2]).

To achieve higher performance from a processors, external memory dies are packaged next to CPU or FPGAs are being integrated with external memory dies which bring challenges in cooling multi-chip packages. Also, with the increasing core count of server CPUs, demand for power per socket has been on the rise to deliver the needed performance increases from one generation to another. The multi-chip nature of the packages makes it harder to maintain good thermal interface material on all dies under the IHS. Additionally some of the dies including the CPU itself could require lower junction temperatures (Tj) to lower leakage power or reduce refresh rates, which adds greater and greater demand on platform level cooling. In order to get to petascale and exascale levels of performance, high performance cooling solutions in densely packed boards and servers are essential.

Currently, most of the high performance HPC machines are being cooled by direct liquid cooling using water or a glycol/water mixture. This traditional single phase liquid (water based) cooling technology contains some inherent inefficiencies and areas of risk. The major ground rules in adoption of single phase liquid cooling are: Do not mix copper and aluminum in the liquid loop. Hence while designing and installing a liquid cooled solution, complete wetted material list needs to be inspected very carefully and should be compatible with copper and fluid inside. This forces one to use only copper based cold plates for processor cooling inside servers which adds weight and cost.

978-1-5386-1272-9/$31.00 ©2018 IEEE

Designer and operator need to make sure that water chemistry is perfect during operation especially for no bacterial growth and ionic dissolution, and have an extensive leak detection system in case of failure. In large scale performance systems, compute platforms usually have several processors on single motherboard and other components such as switches, voltage regulators, power supplies etc that need to be cooled via liquid. To gain energy efficiency and manageable tubing layout, liquid cooled components are connected in series and parallel loop. Single phase fluid picks up the heat from one processor and goes up in temperature before going to the second one. This translates to higher junction temperature (and hence lower performance) of the second CPU in line. This phenomenon is referred to as shadowing.

When water runs near CPUs, with high flow rates, it requires thicker tubing, and is prone to bio-growths. Additionally, the electrically conductive water poses leak risks. All these issues need to be carefully managed. Authors believe that the Pumped Liquid Multiphase Cooling (PLMC) technology described in this paper is energy efficient, uses small tubing and allows to use mixed metals, and does not pose leak risks since the working fluid is highly dielectric and evaporates immediately.

There are several literature reviews & studies available for different types of liquid cooling in data centers. As discussed by Capozzoli & Pimiceri [3], pumped two phase technology is emerging technology that has potential, to cool high density servers and data centers with higher energy efficiency. Saums et al. [4] also present a case study to cool high density power electronics converters using PLMC. They show, that using PLMC technology they can boost cooling capacity nearly (2-2.5X) when compared with air and liquid cooled solutions respectively. Also, the U.S. Government funded a collaborative research grant led by Clusters Systems Company, Inc. [5]. In this research, a team put together a highly dense rack system using a monolithic cold plate with micro-channels on top of the each server chassis. The heat from the major heat dissipating components are raised to the cold plate via solid aluminum blocks. The max heat dissipated per blade server is about 1 kW.

There are several advantages of two-phase liquid cooling technology. The fluid used for PLMC technology is a vaporizable dielectric fluid such as R134a or HFO 1234yf [6]. In the case of leaks or any failures, the fluid immediately vaporizes in open air without shorting-out any electrical components such as CPUs or VR FETs. Because the heat from the processors is absorbed using the latent heat of vaporization (converting fluid into vapor) the fluid cold-plate exit temperature is unchanged compared to the inlet. That means, with proper sizing of fluid flow rate for complete loop and lowest possible sub-cooling of the inlet fluid, one can have multiple processors in series and still achieve very high heat load dissipation keeping case or silicon junction temperatures nearly constant across all processors. Due to boiling phenomenon in the cold plate, one can achieve higher heat transfer coefficient with relatively lower flow rates as compared to single phase fluid. This provides the best cooling for all platform level components. This particular design feature allows PLMC to be a scalable thermal technology to any power rating with minimum design parameters such as total heat load on the loop, max inlet fluid temperature and maximum outlet fluid quality.

From rack and data center level perspectives available floor area is quite important. For such constraints, overall weight matters, and heavy copper is the choice of material for the water cold plates due to higher thermal conductivity and better corrosion resistance. PLMC technology enables the use of aluminum (lighter and more economical) cold plates and blocks with comparable thermal performance. Inert fluids such as R134a, HFO 1234yf, or 1234ze permit use of aluminum and other metals in the complete fluid loop including cold plates, manifolds, pump and Central Distribution Units (CDU).

To demonstrate novel cooling technique to achieve large scale performance and with other logistics constraints a proof of concept project is designed. In this proof of concept project, a board was designed mimicking 20 processors/switches with several auxiliary components such as voltage regulators (VRs). The board was capable of dissipating > 4 kW of heat i.e. 200W each processor/switch. The layout of components were designed in such a way that a signal integrity analysis showed viability as a possible exascale configuration. The objective of this study was to demonstrate feasibility of

- cooling ~4kW of heat load in 1U blade configuration with multiple non coplanar processors tied in serial fluid loop
- in house analytical predictive capability for thermal performance of processors using PLMC
- cooling multiple processors in series and maintaining approximately the same junction (or case) temperature at maximum TDP of the processors.
- design Scalability for platform level products
- design of a thermal solution that meet the weight target of 12 lbs per blade; compare to non-PLMC and point out the weight advantage

The exit fluid-vapor mixture cooling happens in a separate heat exchanger that uses data center facility water. The pumped liquids near the CPUs (in all cases discussed here) are the method to move the heat away from the rack. The cold-plates are for heat transport; the actual heat sinking is usually done outdoors in cooling towers or heat exchangers.

EXPERIMENTAL TEST SETUP

Figure 2 illustrates the thermal test board which represents (or mimic) a next generation server blade with 16 high TDP processors (250 or 300W each) which are placed very close to each with a package to package gap of 2 mm. Additionally 2 high power fabric switches (200-300W each) were mimicked using thermal test vehicles (TTV) and 15% of miscellaneous heat loss (due to through VRs and other MB components) was accounted for in the design. Two types of TTVs were used for this demonstration project. Scott Depot TTV were used to represent processor and Switch chips while Higgins Peak TTVs were used to represent miscellaneous components such as VR FETS. Each TTV is powered using edge fingers on the sides of the board. The dimension of the board and cooling solution is designed such a way that it fits in regular 19”, 1U server blade chassis. Scott Depot TTVs are grooved on top of IHS, where thermocouples are soldered to measure Tcase temperatures during experimentations. This is the standard practice in the industry to measure Tcase temperature on any processors.

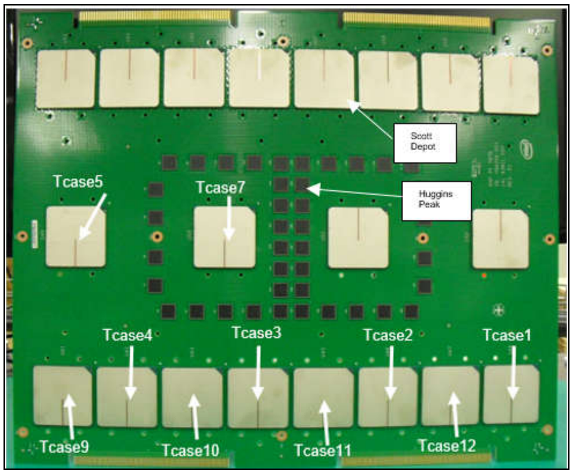

Fig. 2. Test board with multiple Thermal Test Vehicles (TTV) that was used for investigating performance of PLMC technology

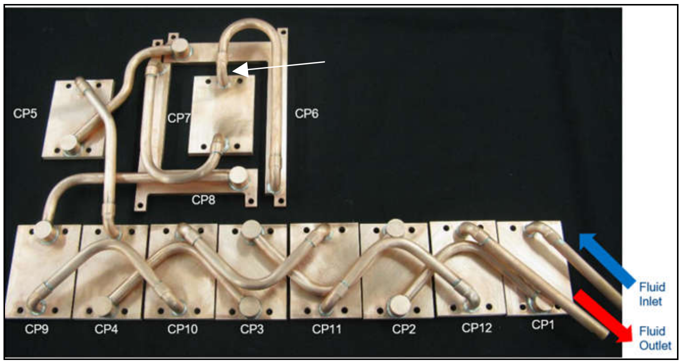

Fig. 3. Cold plate assembly that was used to cool half the components on the board and is symmetrical to cool rest half of the board.

The cold plate assembly, as shown in Figure 3, is designed such a way that the cooling solution, when mounted on the motherboard, still fits in 1U standard chassis. For demonstration purpose and quick turnaround in manufacturing of brazed cold plates, material of choice was copper for cold plates and tubing. Particular care is taken in connecting the cold plates such a way that the tube length between adjacent cold plates is at least 140 mm. This length is needed to provide some relaxation in location and height tolerance mismatch due to multiple soldered TTVs/processors on the board. Also the tubes used for fluid routing were made of annealed copper so as to not stress the tube material and joints and have enough flexibility while installation. As shown in Fig. 3, the cold plates are marked based on the fluid flow path. For designing cold plate performance analytical correlations presented by Lee & Mudawar [7] were used.

Figure 4 illustrates the test bed design that was used for investigation of the PLMC technology. The test bed includes a pump, a bypass valve to control fluid flow rate, flow meter, visual windows to visualize quality of inlet and outlet fluid, pressure and temperature sensors at inlet and outlet of the cold plate assembly, a reservoir with visual windows for fluid level visualization and a refrigerant to liquid compact heat exchanger to dissipate heat to the facility water. R134a refrigerant is used for this demonstration project as primary fluid for PLMC. Lab water was used to carry heat outside via this condenser. All the sensors are pre calibrated by vendor before supplying the test bed.

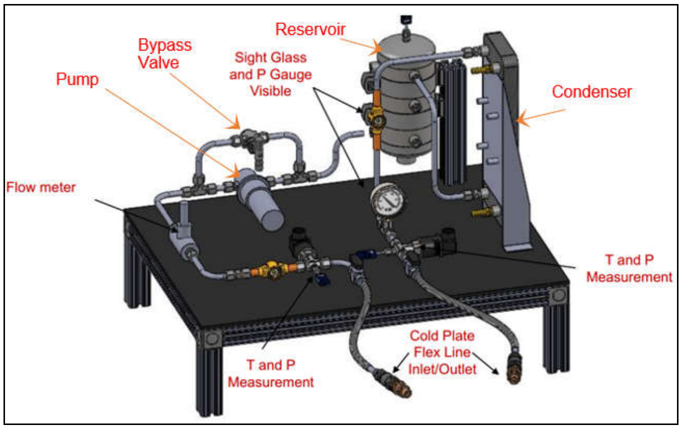

Fig. 4. Test bed and complete test assembly that was used for investigating performance of PLMC technology

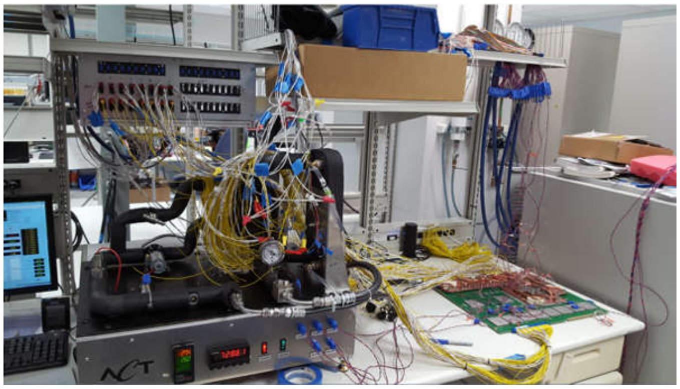

Fig.5. Full test assembly to investigate PLMC technology (no cold plates mounted for clarity)

The cold plate assembly is mounted on the board using in-house designed retention and back plate. Figure 5 shows the full test assembly with fully populated edge connectors, thermocouples connected to data acquisition system and power supplies are connected to heaters of each TTV and power in each TTV is controlled using labview. Dow thermal grease, TC 5630 (TC 5622 Dow chemical name) was used as thermal interface material (TIM) for Scott Depot TTV (mimicking the CPU and switches) and PCM45F was used for Huggins peak TTV (mimicking the miscellaneous components on motherboard that need to be cooled as well) to accommodate tolerance stack up heights. The power to each TTV is supplied using a rack of 110V power supplies. Thermocouples are attached at the center of all Scott Depot TTVs to measure Tcase temperature. All measurements and power settings were recorded using a DAQ and the Labview lab measurement program. The measurement instruments such as DAQ, power supplies are calibrated using professional vendors every year.

TEST PLAN

Below is the test plan that was used to investigate pumped two phase cooling for this this project.

1. The baseline thermal performance comparison was done with single phase liquid cooling. This was achieved by allowing high fluid flow rate and very small heat load is applied on only 1st TTV. This way, boiling was kept to minimum & achieve single phase performance.
2. To investigate effect of varying flow rate on thermal performance (Psi case-fluid_inlet), all TTVs are powered to half power (total ~1kw) and recorded the case temperatures (Tcase)
3. Effect of flow rate (or quality) due to full load on TTVs
4. Effect of variation of TTV power on Tcase of each TTV

EXPERIMENTAL TEST RESULTS

Initially only the 1st TTV in the loop was powered up at a very low power (~20W) to set up the baseline for PLMC performance. The thermal resistance from case to fluid inlet temperature (Psi_case-fluid inlet) was ~0.1°C/W (at the lowest quality). As expected by intentionally disallowing the two-phase fluid to boil causes the set-up to expectedly under-perform; at the same level as water-glycol cold-plates. As seen in Fig. 6, at lower flow rate of 180 ml/min, thermal resistance was 0.08°C/W and exit quality (amount of vapor on the mixture of fluid and vapor) was 14%. As flow rate increases, the exit quality reduces and thermal performance starts increasing. At the highest flow rate (800 ml/min) thermal performance matched with off the shelf cold plate performance with glycol-water based fluids.

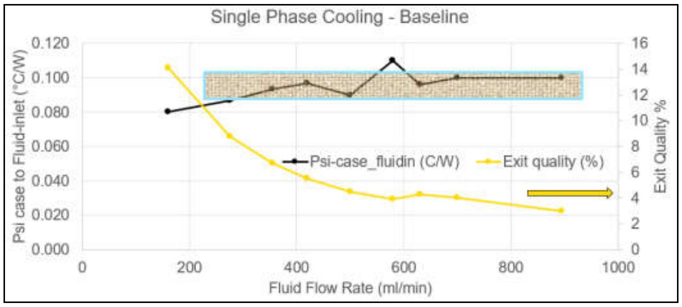

Fig. 6. Effect of varying flow rate on thermal performance of cold plate and exit quality of the mixture

In next experiment, the system was powered to dissipate a total of 840W (essentially 80W on each Scott Depot TTVs (qty 10) and 20W on each row of Huggins Peak TTVs (qty 2), and the corresponding case temperature measurements of each TTV are presented in Figure 7. As seen in the figure when the fluid flow rate is 900 ml/min, Tcase temperature on 1st TTV is 34°C and Tcase on the last TTV in series was 32°C with the difference between inlet and outlet fluid temperature is less than 1°C. The results again point that as fluid passes from 1st cold plate to last, thermal performance of cold plate downstream is equal or better compared to upstream performance. As fluid (or mixture of fluid and vapor) passes through each cold plate, the amount of vapor in the mixture (i.e. vapor quality) increases and it results in increased effective flow rate in every subsequent cold plate leading to same or improved heat transfer coefficient. If the incoming liquid flow rate is decreased to the first cold plate, the exit vapor quality increases from 27 to 65% for flow rate of 900 to 350 ml/min. But even though fluid flow rate changes, exit fluid temperature does not change. At highest exit quality, as the effective heat transfer coefficient is higher, and hence thermal performance of the cold plates near the exit of fluid is also higher (as long as quality is less than 100%). For single phase liquid (water/glycol mixture), there is a 1.3°C temperature rise from cold plate inlet to outlet when 70W of power is dissipated in each TTV and it will increase to 2.8°C at 150W at 800 ml/min. As single phase liquid goes through each cold plate in series, its temperature will continue to rise and thereby increasing the inlet temperature to the downstream cold plate. For 10 cold plates in series, this will amount to 15°C increase in inlet temperature for total of 840W. With two-phase cooling, the effect of increase in power is fairly limited as is primarily due to conductive resistance between the IHS and the boiling surface.

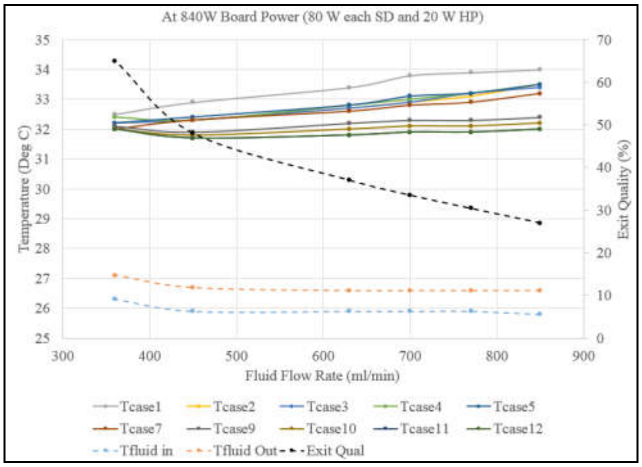

Fig. 7. Effect of varying fluid flow rate on thermal performance of cold plate with 840W of heat dissipation

To demonstrate effect of varying heat load and its effect on consequent TTV case temperatures in series, heat on the first TTV is increased to 150W from original 80W with the same fluid flow rate of 360 ml/min. As the heat load on the 1st cold plate increases, Tcase temperature jumps to 36°C from original 33°C as shown in Fig. 8. But the Tcase temperatures on the downstream cold plates does not alter. The only difference is the exit quality increases from 65% to 70% to accommodate 70W of extra heat. If this happened to be single phase liquid cooling, Tcase temperatures of downstream TTVs with increase with same temperature rise as 1st TTV. In real systems, due to this shadowing effect, due to increased junction temperature compute performance of shadowed part may decrease with same power on the processor due to leakage.

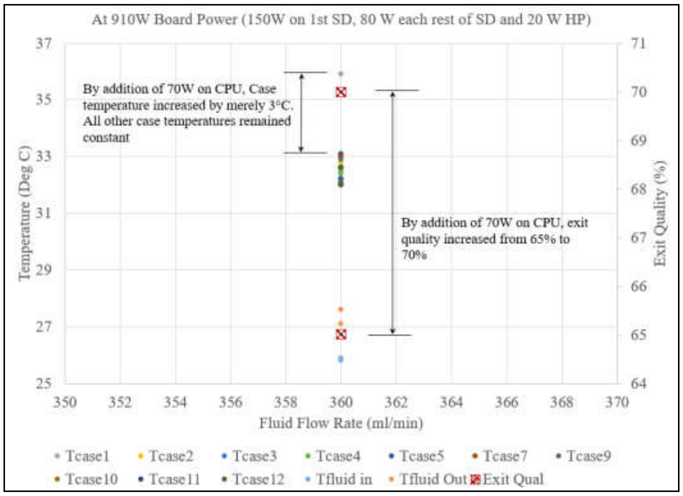

Fig. 8. Effect of increasing heat load on 1st node and its impact on the downstream Tcase TTV temperatures

As illustrated in Fig. 9, as long as the exit quality is under 80%, a small amount of pumped refrigerant (~780 ml/min) was able to easily carry away heat from all of these 200W series connected loads. This was done while keeping them within a very tight temperature range and impervious to any variations in the loading mix. Three heat loads were applied i.e. 840, 1860 and 2060W at constant fluid flow rate. As seen in Figure 9, even though the load is increased from 840W to 2060W (>2X), temperature rise across inlet and outlet fluid temperature is merely 2°C out of which 1°C comes from sub cooling of refrigerant.

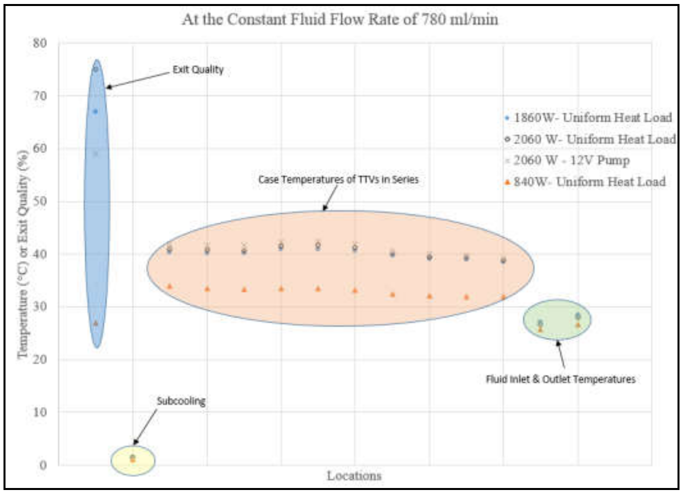

Fig. 9. Effect of varying heat load on the processor and its corresponding temperature rise at case or junction

As the heat load increases from 840W to 2060W (or 80W to 200W for each TTV), Tcase temperatures of Scott Depot TTV increases from 37°C to 41°C. The exit quality changes from 27% to 75% for the same flow rate of 780 ml/min. As explained earlier, downstream TTVs are getting better cooling due to increased vapor quality and hence shows the lower case/junction temperatures. But this improvement is within 1 to 2°C and could be within thermocouple error.

In real scenarios encountered at typical data center, all processors need not be running at the same power as they may be running different workloads. To simulate such a real scenario, alternate Scott Deport TTVs are powered to full 200W (2060 W total) and 80 W (840W total) respectively. Figure 10 illustrates the results of this test. As seen in the figure, TTVs which were running at 80W, showed Tcase temperatures of 36°C, as if they were all having same 80W each i.e. 840W heat load condition. The TTVs which were running at 200W, Tcase were 41°C same as 2060W case where all TTVs were running at 200W represent all processors are running at turbo mode. The small difference in temperature between uniform and non-uniform power at 200W Scott Depot TTV is due to higher saturation temperature of closed loop due to total higher power dissipation. This proves that each CPU will be running at its respective junction temperature depending upon workload without having any shadow effect due to upstream processors or will not impact junction temperature of downstream processors as well.

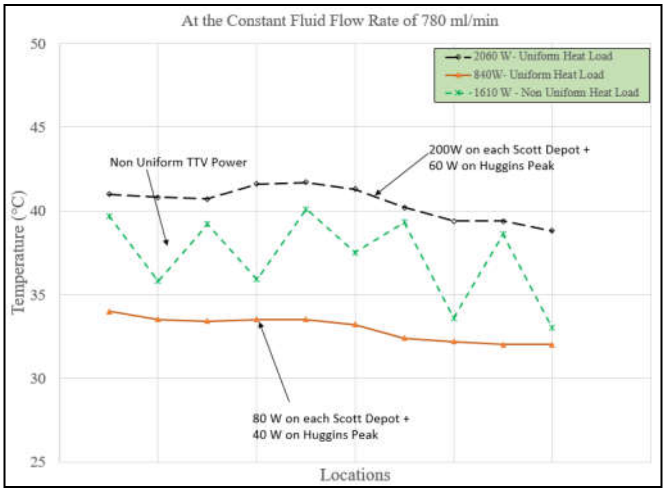

Figure 10. Effect of varying heat load on Tcase of the thermal test vehicles

SUMMARY

In this work, it was demonstrated that two phase cooling can be utilized to cool total heat load of 4kW of power on a single motherboard via cooling multiple CPUs. The effect of shadowing that is seen in single phase cooling with cold plates when connected in series can be eliminated with two-phase cooling. All cold plates in series can have effectively the same fluid inlet temperature and hence all CPUs can run at the same case or junction temperature.

In comparison, single phase cooling suffers from inlet temperature increase from one cold plate to the next one in series. In order to get the required cooling for the last CPU in series, one has to provide much lower glycol/water inlet temperature to the first cold plate to compensate for the temperature rise. This can drive lower facility water temperature requirement resulting in higher cost of cooling. The other way to overcome the shortcoming of single phase cooling would be to provide parallel connection to each cold plate. However, that would require significant amount of liquid flow coming to each server requiring larger diameter piping and the amount of piping in the system would prohibit achieving the density of CPUs on a board. In comparison, the tubes used with two phase cooling can be smaller in diameter. Copper tubing was used to connect alternate cold plate and not the adjacent cold plates to make the tube connection longer in length to provide enough flexibility to handle any co-planarity issues among the different TTV IHS surfaces. The cold plates were light weight and met the 12 lbs weight target. This can be further reduced by ~68% if aluminum is used instead of copper for cooling loop manufacturing.

In all, results proves the Pumped Liquid Multiphase Cooling (PLMC) technology is easy to design, scalable, modular, offer light weight option, no thermal shadow effect, able to dissipate more than 2 kW heat load per loop, can cool multiple processors on board even with no coplanar surfaces, reliable, energy efficient and can be adopted for thermal management of serves with exascale performance. To get to wider adoption of this technology, design maturity need to be improved as well as component ecosystem & availability via high volume manufacturing need to be satisfied.

ACKNOWLEDGEMENTS

Authors would like to thank Ian Williams, Mike Patterson, Eric Buddrius and Jeff Smalley for their support in designing the test prototype. Authors would also like to acknowledge Devin Pellicone from Advanced Cooling Technologies, Lancaster, PA for support in building cooling loop for this project.

REFERENCES

[1] "The Opportunities and Challenges of Exascale Computing," Summary Report of the Advanced Scientific Computing Advisory Committee, US Dept of Energy, Fall 2010.

[2] E. Barton, "Eric Barton Progress Updae on the Fast Forward I/O & Storage Program." jan-2013 [Online]. Available: http://www.youtube.com/watch?v=ynyR8Pjq8W4&feature=youtube_gdata_player. [Accessed: 11-Nov-2017]

[3] G Capozzoli, A., Primiceri, G., "Cooling Systems in data centers: state of art and emerging technologies," Energy Procedia, 2015, Vol. 83, pg. 484-493.

[4] Saums, D., Levett, D., Howes, J.C., Marsala, J., "vaporizable dielectric fluid cooling for IGBT power semiconductors," IMAPS conference, May 2009.

[5] Hughes, P., Lipp, R., "Recovery ACT: Development of a very dense liquid cooled compute platform," Energy Research & Development Division Final Project Report, February 2013.

[6] Marcinichen, J., "Reasons to Use Two Phase Refrigerant," Electronics cooling, March 2011.

[7] Lee, J., Mudawar, I., "Low Temperature two phase microchannel cooling for high heat flux thermal management of defense electronics," IEEE Transactions of Components & Packaging Technologies, Vol. 32, No. 2, June 2009.