Proceedings of the 11th World Congress on Momentum, Heat and Mass Transfer (MHMT 2026)
Paris, France - April, 2026
Paper No. ICMFHT 113 (The number assigned by the OpenConf System)
DOI: TBA

Two-Phase Micro Cooling Integrated In An Internal Heat Spreader

Ruben Betsema[1], Patrique Boerboom[1], Cor Rops[1]
[1]TNO High Tech Industry
High Tech Campus 25, Eindhoven, The Netherlands
ruben.betsema@tno.nl; patrique.boerboom@tno.nl; cor.rops@tno.nl

Abstract - The number of transistors per square centimeter approximately has doubled every two years over the last six decades. New applications such as AI and HPC push this semicon trend further for years to come. Along with this transistor density increase, since the start of the millennium change the energy density of the chips increases as well. Therefore putting the cooling abilities as a limiting factor for further developments. In this investigation a novel in internal heatspreader two-phase cooling method (IN-IHS two-phase microcooling) is presented and experimentally compared to standard air-cooling and water liquid cooling. The measured junction-to-fluid boiling temperature thermal resistance is 0.045K/W at a flow rate of 85ml/min and a socket power of 390W. Stable and robust flow boiling is observed for a wide range of flow rates.

***Keywords*:** two-phase microfluidic flow boiling, in-internal heatspreader cooling.

1. Introduction

Current chip developments, although pushed by the Artificial Intelligence (AI) employment and High-Performance Computing (HPC), are in line with a long-term trend. Over the last six decades the number of transistors per square centimeter has approximately doubled every two years. This observed trend is known as Moore's law as introduced in his IEEE speech in 1975 [1]. Currently, one of the main challenges for the future miniaturization of electronics is its thermal management. Although the number of transistors per square centimetre increased, the energy densities did not increase significantly before the millennium change. This is due to the fact that both the voltage and current scale down with decreasing length, also known as Dennard law [2]. However, since the start of this millennium Dennard scaling is not holding any longer, due to leakage current and threshold voltage. The generated heat fluxes of the electronic devices continue to increase over the years. Over the last 20 years a linear increase is observed from approximately 50W/cm^2 to almost 200W/cm^2 [3], see left graph Figure 1.

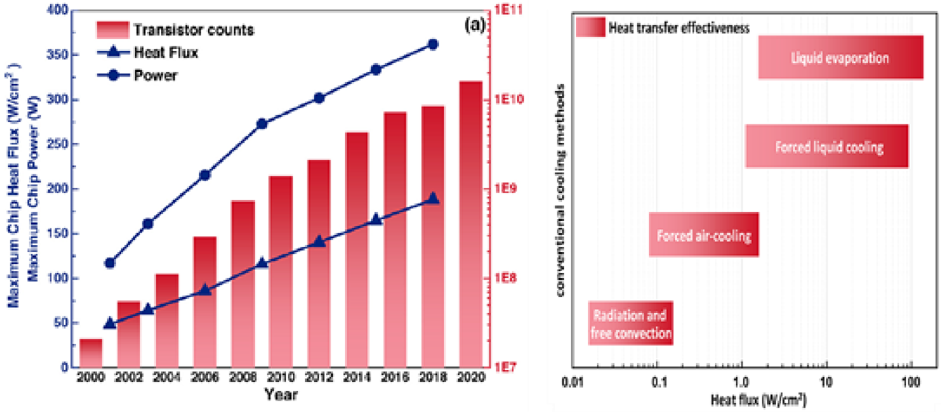

Figure 1: Left: Maximum chip heat flux (blue triangles) and maximum chip power (blue dots), and number of transistors per device (red bars). Right: Typical range of heat flux density of conventional cooling methods. After Zhang [3].

Various markets show the trend of increasing power densities. For example, the public acceptance of the electrification of vehicles depends on the ability to fast charge the battery. Similarly, during the operation of the electro-engine large portions of energy (300kW to 600kW) are converted from alternating to direct current or vice versa (AC/DC converters). Wide Band Gap (WBG) technologies such as SiC and GaN and increased operation voltages, increase the device's switching performance. Although the efficiency of these inverters is above 99.5% [4], the heat losses remain significant: 1 to 2kW.

Another example of the increase in power densities can be found in the datacenter community [5]. In this 2021 publication the Thermal Design Power (TDP) in datacenters is identified to exceed 400W in 2030. However, 400W TPD values are already reported in 2025 for the NVIDIA Blackwell AI processors along with reports on overheating challenges of this chip. The cooling method of the chips affects the efficiency of the rack cooling loop [6]. Therefore, multiple investigations have been conducted to improve the cooling performance of cold plates. These solutions comprise conduction based technologies such as high conductive thermal films, thermal vias or thermoelectric cooling. Other cooling solutions are based on convection using IN-chip microfabricated structures, metal fins or two-phase flow boiling [7].

Conventional cooling technologies are not able to handle the currently obtained high heat fluxes. The right graph of Figure 1 shows that conventional free convection air cooling is suitable for heat fluxes up to 0.1 W/cm^2. By using forced convection, surface enhancing techniques and optimizing heat removal from the system, this number can be amplified one or two orders of magnitude [8]. However, taking into account the trend on heat flux increase, air cooling has reached its limits. Depending on the application more efficient cooling methods are explored, such as liquid cooling or fluid boiling cooling.

In 2021 Ramakrishnan et al. executed a performance assessment on three different CPU cooling methods: air-cooled heat sink, liquid cold plate and pin-fin boiling plate in an immersion tank [9]. Overclocking is used to tune the performance of the CPU, Intel i9 9900K. From the various stress test tools, Prime95 offered the maximum possible workload. Their main findings are represented in Figure 2

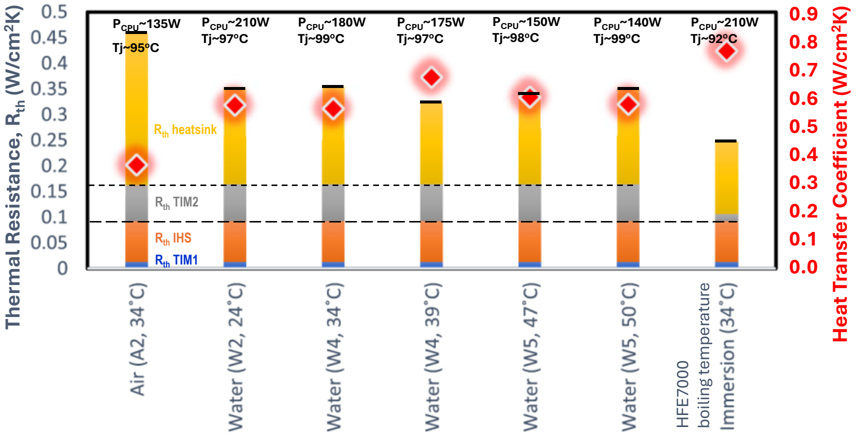

Figure 2: Thermal characterization various cooling methods (Air cooling, Water cooling, Immersive boiling HFE7000. Bars represent build-up thermal resistance, red dots represent the total heat transfer coefficient. After Ramakrishnan [9].

The above figure shows that the air cooling has the highest thermal resistance, while the immersion bath has the lowest total thermal resistance. However, the main improvement of the immersion bath as compared to the liquid cold plate is obtained by the reduction of the TIM2 (Thermal Interface Material) thermal resistance. This reduction is achieved by soldering the boiler plate to the internal heatspreader (IHS) instead of using a thermal grease to connect the air cooling or the liquid cold plate cooling to the IHS.

The liquid cold plate thermal resistance is not dependent on the entrance temperature of the liquid. Increasing the entrance temperature of the liquid cold plate leads to throttling of the CPU. The thermal throttling maintains the junction temperature its maximum allowed temperature (about 98 ^{o}C), but the employed socket power decreases. The immersion bath is able to remove all heat at the maximum CPU power of about 210W. Concluding, Ramakrishnan and co-workers show that reducing the TIM2 thermal resistance is possible and a complete removal would lead to a significant improvement. Additionally, it may be concluded that the casing-to-fluid thermal resistance needs minimization as well, to significantly improve the thermal performance of the cooling method.

As mentioned above in multiple literature studies integration of the cooling into the internal heatspreader is a promising technology to significantly reduce the thermal resistance. In 2009 TNO has developed a microfluidic flow boiling method which ensures stable flow boiling in a microchannel structure [10]. Over the last decade this method has developed to a microcooling device which has experimentally shown to cool 1kW half bridges for fast charging test setups at a heat flux up to 250W/cm^2 [11].

In the current investigation the microfluidic flow boiling method is integrated into the internal heatspreader (IN-IHS microcooling) of an AMD EPYC chip. The results of the IN-IHS two-phase micro cooler will be compared to the standard air-cooling solution as present on the server board and an off-the-shelve liquid cooler. In the next chapter the experimental setup and analysis method will be presented. In the third chapter the experimental results will be shown and discussed, and in the final chapter conclusions will be drawn and a further outlook will be given.

2. Experimental setup and analysis method

In order to test at current standard conditions which are present in datacenters, the test chip should be able to reach a socket power of 400W. The EPYC High Performance Computing (Zen 4 cores) is able to reach a socket power of 390W. The thermal management system of the DELL server board manages the junction temperature to remain below 95 ^{o}C. Above this temperature thermal throttling starts reducing significantly the electrical power supply to the chip. As a power stress test the Prime95 program is used. This program requires all system resources and is commonly used as overclocking tool. The AMD micro-prof analysis tool is used to identify the runtime performance and monitor thermal and power characteristics of the system.

Three cooling methods have been implemented on the server board. First the performance of the standard installed air cooling is tested. Next, after removing the air cooling, a commercially available liquid cold plate is mounted on the IHS. Finally, the EPYC chip (with IHS) is replaced for a similar EPYC chip with a reworked IHS containing the microfluidic flow boiling technology, see Figure 3.

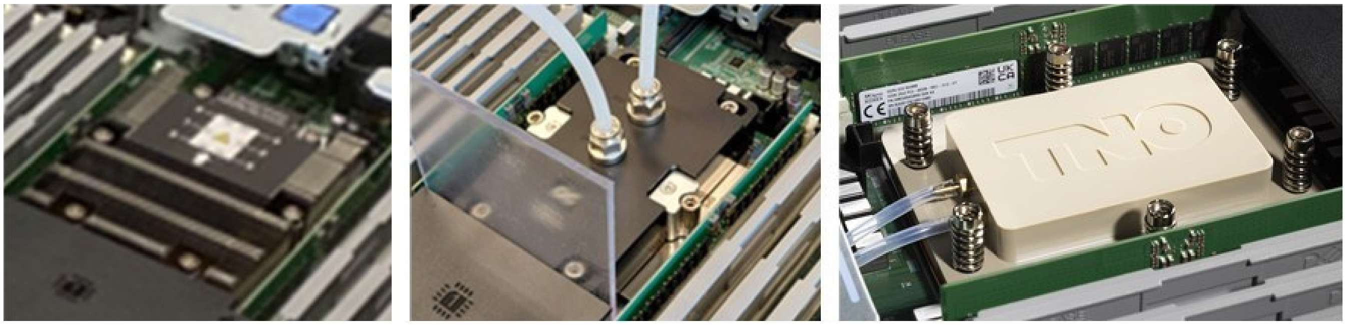

Figure 3: Three cooling methods. Left: standard air cooling; Middle: water cooling cold plate; Right: IN-IHS two-phase microcooling.

In order to compare the various cooling methods to each other the thermal resistance between the electrical junctions and the ambient fluid is determined. Therefore the junction temperature, $T_{junction}$, and the socket power, Q_{socket}, are obtained using the AMD micro-prof analysis tool. As ambient fluid temperature, $T_{ambient\,fluid}$, the ambient air temperature is used for the air cooling. In case of the liquid cooling, the entrance temperature of the water is used. Finally, for the IN-IHS two-phase microcooling the fluid boiling temperature is used.

$$\mathrm{R_{th}} = \frac{Q_{socket}}{T_{junction} - T_{ambient\ fluid}} \tag{1}$$

The measured ambient air temperature is 20 °C, and since no additional thermal conditioning of the water for the liquid cooling is done the entrance temperature for the liquid cooling is 20 °C as well. The applied non-toxic, non-PFAS, high-latent heat fluid for the IN-IHS two-phase microcooling has a boiling temperature 67.5 °C.

Both the liquid cooling and the IN-IHS two-phase microcooling are closed fluid loops. From the liquid pump the cooling liquid flows via a particle filter (15micron) into the chip cooler. Downstream of the chip cooler the cooling liquid, or two-phase flow, flows towards a heat exchanger bringing the cooling fluid back to the ambient temperature. In between the heat exchanger and the pump an expansion vessel is placed to maintain the operating pressure. Therefore the pump always receives the cooling liquid always at ambient temperature and ambient pressure.

3. Performance comparison

The three cooling methods are compared to each other using an equal torture test, Prime95, to obtain a maximum socket power. It is noted that if the temperature of the junctions exceeds 95 °C , the thermal management system of the server throttles the power towards chips. The observed socket power decreases significantly in case of this thermal throttling. The temperatures reported in this investigation are the stable temperatures reached after several minutes of running the torture test. The reported socket powers are obtained during this steady state situation, see Table 1.

Table 1: Measured junction temperature and socket power for the three investigated cooling methods.

	Flow rate [ml/min]	**Socket power, *Prime95* [W]**	**Junction temperature [°C]**	**Fluid temperature [°C]**	**Thermal resistance [K/W]**
Air cooling	*- fans full power -*	392	87.7	20	0.17
Water cooling	900	379	49.3	20	0.08
2-phase cooling	85	390	84.9	67.5	0.04

For the air cooling it is not relevant to denote a flow rate, since the system is an open system with fans. Which portion of the generated air flow passes through the cooling fin structure cannot be determined. It is noted that at the maximum socket power, the thermal management system set the fans at full power in order to keep the junction temperature below 95 °C. Although the temperature is highest and the fans are run at full power, the chips are able to operate at full power with the standard air cooling without thermal throttling.

For the liquid water cooling, it is observed that the obtained socket power is slightly less than for the other two cooling methods. The junction temperature is well below the 95 °C, thus no thermal throttling is expected. During the liquid cooling and 2-phase cooling testing, the standard air fans are still operational, in order to guarantee proper functioning of the remainder of the server board. However, for an objective cooling performance comparison, in front of the EPYC chip a small plate is placed to block the air flow. This ensures only the liquid water cooling, or the IN-IHS 2 phase microcooling, is actively cooling the chip. It is noted that the results of both the liquid water cooling and the IN-IHS two-phase microcooling do not change if the plate blocking the air is removed. This indicates that both cooling methods are dominant over the (remaining) air cooling.

The last column of Table 1 shows the thermal resistance of the three cooling methods. The thermal resistance of the air cooling, R_{th_air}=0.17K/W, is the largest. The liquid water cooling, R_{th_liq}=0.08K/W, is roughly twice as good as the air cooling. The IN-IHS two-phase microcooling is again almost twice as good as the liquid cooling, R_{th_2ph}=0.04K/W. The flow rate of the two-phase cooling, 85ml/min, is about ten times less than the applied liquid flow rate, 900ml/min, of the liquid cooling. This corresponds to the typical ratio between the latent heat of evaporation and the sensible heat needed to bring a liquid to its boiling point. This ratio also allows to run the two-phase cooling loop at a higher temperature. Since the condenser only needs to cooldown the two-phase exit fluid to just below the boiling temperature, the secondary cooling loop can generate higher temperatures. This results in valuable high quality waste heat which can be directly applied in other processes such as district heating or industrial processes.

In order to characterise the IN-IHS two-phase microcooling more elaborately, the thermal performance of the cooling method are measured for different flow rates. It is assured that the exit flow of the micro cooler remained in the a proper two-phase flow regime. A too low supply flow could lead to potential dryspots in the micro cooler, which may lead to undesired hotspots. In Figure 4 the main thermal performance indicators are shown for a wide range of flows. All presented flows exhibit stable flow boiling, and no thermal throttling of the system is observed.

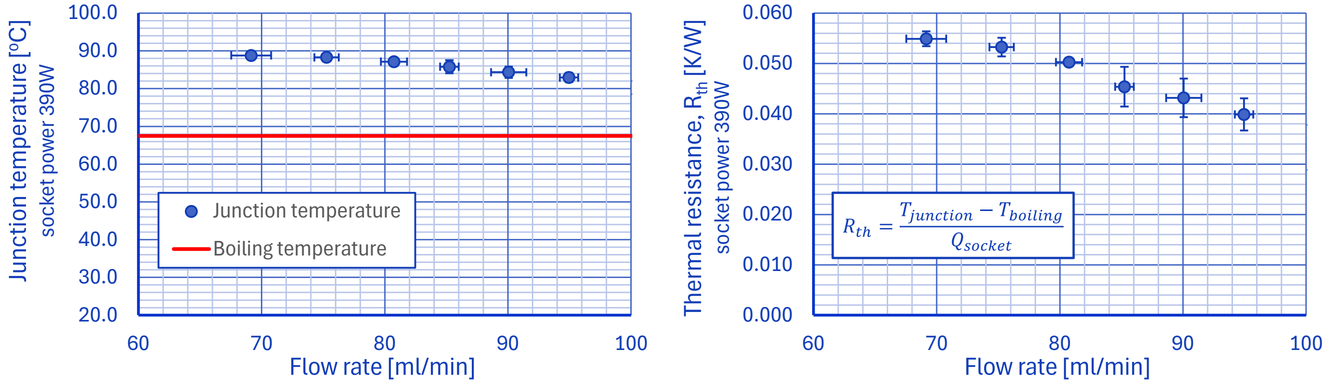

Figure 4: Main thermal performance indicators IN-IHS two-phase microcooling. Left: Junction temperature as a function of the flow rate; Right: Junction-to-fluid boiling temperature thermal resistance as a function of the flow rate.

The left graph of Figure 4 shows that the junction temperature is hardly dependent on the flow rate of the IN-IHS two-phase microcooling loop. Increasing the flow rate by roughly 50%, only changes the junction temperature by about 5 degrees. An increase in wall temperature leads to more intense flow boiling, which increases the fluid heat transfer coefficient. This directly stabilizes the junction temperature, even at highly variable operation loads. Therefore, by the laws of nature the junction temperature is controlled in a robust and stable manner. This implies that the IN-IHS two-phase microcooling is able to run at a constant liquid supply flow rate, which simplifies the cooling fluid loop by excluding any (feedback/feedforward) flow control.

In the right graph of Figure 4 the junction-to-fluid boiling thermal resistance is shown. The obtained values vary between 0.040K/W to 0.055K/W depending on the supply flow rate. The high localized heat flux at the junction cores are expected to contribute significantly to the junction-to-fluid boiling temperature difference and thus to the total thermal resistance. Since this internal chip resistance is not known, similar as the thermal properties of the TIM1 layer between the IHS and the silicon die, it is not possible to allocate a justified contribution of these bodies total thermal resistance. Therefore the thermal resistance of the IN-IHS two-phase microcooling method (case-to-fluid thermal resistance) cannot be deduced. This could be done performing a thermal torture test on a Thermal Test Vehicle (TTV). However, testing with the actual chip-silicon with HPC like workloads obtain results which are difficult to acquire using a TTV. However, these actual measured thermal resistances are relevant for actual thermal system design.

4. Conclusion and outlook

This investigation compares the two-phase microfluidic cooling method embedded in the internal heat spreader (IN-IHS two-phase microcooling) to standard air cooling and water liquid cooling. An EPYC High Performance Computing chip is successfully subjected to a thermal torture test by running Prime95. The thermal performance indicators are measured on a functional DELL datacenter server board using the dedicated AMD micro-prof analysis tool. The standard air cooling and the liquid cooling are mounted onto the internal heatspreader according to device specifications. The IN-IHS two-phase microcooling is successfully retrofitted into the internal heatspreader of another EPYC High Performance Computing chip.

All cooling methods are able to cool 390W with a junction temperature below the thermal throttling limit of 95 ^{o}C. The water liquid cooler has a thermal resistance which is half of the thermal resistance obtained using the standard air cooling. The IN-IHS two-phase microcooling has again a factor of two better thermal resistance than the water liquid cooler. Depending on the flow rate the junction-to-fluid boiling temperature thermal resistance of the IN-IHS two-phase microcooling varies between 0.040K/W and 0.055K/W. The evaporative latent heat employment allows that a 50% change in flow rate affects the junction temperature only by roughly 5 degrees.

The microfluidic flow boiling technology can be miniaturized further such that it is integrated inside the top layer of the silicon die. This will remove the thermal resistance of the TIM1 layer between the IHS and the silicon die. Additionally, it will reduce the internal thermal resistance of the die. It is expected that junction-to-fluid boiling temperature thermal resistances below 0.02K/W are possible to achieve.

References

[1] G. Moore, *IEEE International Electron Devices Meeting*, Washington, D.C., USA, 1975.

[2] B.G. Streetman, S.K. Banerjee, "Solid state electronic devices." Boston: Pearson, 2016.

[3] Z. Zhang, X. Wang, Y. Yan, "A review of the state-of-the-art in electronic cooling", *Advances in Electrical Engineering, Electronics and Energy*, vol.1, 2021.

[4] J. Zhu, H. Kim, H. Chen, R. Erickson, D. Maksimović, "High efficiency SiC traction inverter for electric vehicle applications", *IEEE Applied Power Electronics Conference and Exposition (APEC)*, San Antonio, USA, 2018.

[5] S.K. Rajan, B. Ramakrishnan, H. Allisa, W. Kim,C. Belady, M. Bakir, "Integrated silicon microfluidic cooling of a high-power overclocked CPU for efficient thermal management", *IEEE Access Electronics packaging*, vol.10, 2022.

[6] A. Heydari, A. R. Gharaibeh, M. Tradat, Q. Soud, Y. Manaserh, V. Radmard, B. Eslami, J. Rodriguez, B.Sammakia, "Experimental evaluation of direct-to-chip cold plate liquid cooling for high-heat-density data centers", *Applied Thermal Engineering*, vol. 239, 2024.

[7] S. Rangarajan, S. N. Schiffres, B. Sammakia, "A Review of Recent Developments in ''On-Chip" Embedded Cooling Technologies for Heterogeneous Integrated Applications", *Engineering*, vol.26, 2023.

[8] Mohammed, Razuqi, "Effect of air fan position on heat transfer performance of elliptical pin fin heat sink subjected to impinging air flow", *Journal of thermal engineering*, vol.7(6), 2021.

[9] B. Ramakrishnan , H. Alissa , I. Manousakis, R. Lankston, R. Bianchini, W. Kim, R. Baca, P. A. Misra, I. Goiri, M. Jalili, A. Raniwala, B. Warrier, M. Monroe, C. Belady, M. Shaw, M. Fontoura, "CPU Overclocking: A Performance Assessment of Air, Cold Plates, and Two-Phase Immersion Cooling", *IEEE transactions on components, packaging and munfacturing technology*, vol.11(10) 2021.

[10] C.M. Rops, J.F.M. Velthuis, F. Graaf, L.F.G. Geers, "Multiple connected channel micro evaporator", *WO2009/082230*, 2009.

[11] G.D. Mensink, R. Betsema, C.M. Rops, R.B.J. Koldeweij, P.B.T.H Boerboom, "High heat transfer chip cooling: capillary based flow boiling model and a novel experimental analysis method", *10th World Conference on Experimental Heat Transfer, Fluid Mechanics and Thermodynamics*, Greece, 2024.