HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 4 (2023)

Reviewing Cooling Strategies for Post-Dennard Era Computing and Modern Electrothermal Phenomena

Tyler Sapasap[1] and Xicotencatl Rojas[1#]

[1]Redwood High School
[#]Advisor

ABSTRACT

Excessive thermal runaway typically manifests itself in computer component damage and other various negative side effects. As a mitigation strategy, end-users and consumers typically implement differing methods of component cooling, including fan-heatsink cooling, heatsink-only cooling, and liquid-based cooling. Different cooling methods remain impactful to modern computing, as it remains a core component in the thinking of thermal design and reliability engineering in the realm of semiconductor devices. The industry-wide acceptance of the end of Dennard scaling and the imminent end of Moore's law are major factors that are currently impacting CPU power consumption trends and modern cooling philosophies. Transistor packing and process refinement is beginning to push against atomic boundaries in combination with phenomena such as leakage current and high current density, causing a general trend of increasing temperatures generation-to-generation in microprocessors. As a result, thermal mitigation strategies and protections must be in place to reduce damage and catastrophic failure while increasing performance of the die package. In a high-heat scenario, liquid-cooling can provide up to 38% to 48% improvement over fan-heatsink variants depending on the type of workload executed by the processor. Fan-heatsink cooling faces thermal resistance limitations in the form of spreading and air convection resistance as a result of heatsink material composition, the resistance along the path of heat flow impeding conduction rate, and the lower thermal conductivity of air compared to liquid. Currently, the best performing variant appears to be liquid-based cooling while fan-heatsink combinations provide adequate levels of thermal dissipation based on these observations.

Introduction and Historical Context

The issue of transistor miniaturization posed many challenges to energy efficiency and improved performance in computing technology in the 1970s. Thermal and energy design was an important issue as manufacturers tried to manage ways to increase transistor density and decrease transistor size without the risks associated with excessive heat and reliability concerns. Robert H. Dennard was an electrical engineer who addressed these issues with a published work titled, "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions" which facilitated a roadmap of modern semiconductor manufacturing. The authors outlined a process in which metal-oxide-semiconductor field-effect transistors (MOSFETs) can be physically shrunk while still preserving the electric field strength, allowing no compromises to performance or design complexity (Bohr, 2007). In practicality, this allowed microprocessor devices to be designed with smaller lithographic process nodes while also enabling manufacturers to introduce more transistors on the device, therefore enabling improvements in transistor density, power efficiency, and performance per watt. These findings had impacted transistor and chip design philosophies for the following decades, leading to great advancements in semiconductor technology. However, these developments would meet difficulties in the mid-2000s; as transistor packages became more and more dense, power density to performance no longer scaled. The effects of quantum and direct band-to-band tunneling also initiated higher levels of leakage current (in which electrons at nanoscale could penetrate the gate oxide insulating layer causing superfluous energy consumption), leading to

power efficiency losses, which particularly became an issue as manufacturing processes began to implement smaller-scale processing nodes (Bohr, 2007; Su et al., 2003).

Scaling down other components, such as the gate oxide (SiO_2 dielectric) thickness, became a challenge to overall transistor downscaling as leakage current became a more prominent issue (Bohr, 2007). Scaling down the transistor node in accordance with Dennard scaling would also require that the SiO_2 dielectric atomic layers also become downscaled, which is an issue because smaller nodes need thicker SiO_2 dielectric layers to counter the leakage current side-effects (Leong et al., 2006; Su et al., 2003). However, thickening this layer would be counterproductive as thickening the SiO_2 dielectric would also thicken the dimensions of the transistor, leading to complications with transistor downsizing (Leong et al., 2006). Another issue is that SiO_2 dielectric downscaling is a finite process as transistors begin to push atomic limits, in which atoms required for the SiO_2 dielectric exceed that of what is required of the process node; this is crucial once process nodes begin to reach nanometer scale (Bohr, 2007).

The subsequent end of Dennard scaling marked the end of an era in which microprocessors no longer offered the benefits of increasing performance for the same amount of wattage drawn. As a response to this development, mainstream microprocessor design began to shift towards having multiple cores within the CPU package (Esmaeilzadeh, 2011). In turn, these new adoptions had decreased the power efficiency of the package for less generation-to-generation improvement as a result of increasingly expensive lithographic fabrication processes and rising watts per square millimeter power density. The culmination of these factors helped to contribute to the growing trend of power consumption for consumer and data-center processors (Esmaeilzadeh, 2011). In turn, increased power consumption leads to higher temperatures, ultimately emphasizing the need for active cooling for longevity in data-center, enterprise, and personal computing (Pedram & Nazarian, 2006).

Defining Excessive Thermal Load

Excessive thermal load is characterized by the state in which an implemented cooling solution is no longer capable of handling the heat load imposed by the underlying CPU package power consumption (Vassighi & Sachdev, 2006). Depending on the model utilized, a modern CPU package may include the integrated memory controller, onboard cache, integrated graphics, system agent, and other input/output devices that may contribute to heat generation. More modern chips have an underlying thermal junction maximum temperature (TjMAX), which states the highest operating temperature that would be considered safe and manageable under normal operating conditions for the processor cores. These normal operating parameters often are specified by the manufacturer of the chip and vary from model-to-model (Ramakrishnan et al., 2021). As of recent years, there is no de facto standard set by the modern chip manufacturers of today (Intel and AMD) that describe safe operating temperatures for all processors, meaning that proper documentation must first be consulted before use.

Once the TjMAX boundary has been crossed, there are three potentialities: 1) The chip may underclock — a mechanism that reduces the clock cycle frequency from the manufacturer's specified normal operating ranges, leading to reduced performance capability; 2) The chip may be undervolted — leading to the degradation in performance as a trade-off for system stability and safety; 3) The chip or hardware firmware may not have any protective measures in place to remediate the generated heat, causing catastrophic failure, reduced lifespan, or permanent performance loss (Mazouz et al., 2014; Vassighi & Sachdev, 2006). A normal safety response to high temperature in modern computers is typically a combination of 1 and 2, a widely known reliability design measure known as thermal throttling (Jalili et al., 2021). Computer shutdowns are typically a last resort response to prolonged excessive temperature gain. The lattermost possibility is more common in antiquated legacy systems that predated the existence of temperature-induced thermal throttling; they also predated C-state residencies in which the CPU cores during idle times would deactivate as a power saving mechanism (Ilsche et al., 2018). These safeguards ultimately prevent the possibility of damage to the computer components, but at the cost of potential end-user satisfaction as a result of the performance sacrifices made, emphasizing greater importance for power efficiency and the mitigation of overheating in its entirety (Schöne et al., 2019).

Defining Proper Cooling

First, to achieve proper cooling, the thermal conductivity of the transfer device must first be sufficient to handle the thermal output of the components it is cooling. This can be measured by the thermal design power (TDP) specifications of the cooling device and CPU; this describes the theoretical maximum amount of heat that a cooling device can dissipate and the theoretical maximum amount of heat that a component can generate, as measured in watts (Jalili et al., 2021). Second, the CPU components must stay under certain operating temperatures as specified by TjMAX in order to prevent damage. Third, in a typical cooling environment, the cooler and the CPU must reach steady state and thermal equilibrium while still maintaining temperatures under the safe operating limits of the particular CPU model (Pedram & Nazarian, 2006; Ramakrishnan et al., 2021).

In unmanaged, high heat computer workloads, steady state may not be observed in cases where the cooling solution is insufficient to the thermal demands of the processor; as mentioned earlier, in this scenario, temperatures would continue to rise until damage, or until the computer itself implements protective measures (thermal throttling, shutdown, etc.). The focus of this paper is centered on two main methods of cooling that are implemented to achieve a temperature state below TjMAX, including but not limited to: 1) fan-heatsink cooling; 2) liquid-based cooling.

Using Heat Exchanging Methods as a Distinction Between Heatsink Cooling and Cold Plate Cooling

Heatsink methods and liquid-based methods both have similar, yet distinct ways in which they transfer heat from the CPU package. They are similar in that they contact the integrated heat spreader (IHS) to transfer heat away from the package. However, there lies a crucial difference in where this heat may be transferred. The methods described below will outline the general cooling processes and may dismiss other critical aspects (such as thermal interface material and ambient temperature) for the sake of focusing on the general dichotomies of the two processes as well as for simplicity.

In heatsink cooling methods, thermal transfer is a passive process, strictly speaking, without the use of a fan. The heatsink must also make complete contact with the CPU IHS to transfer heat (Zhang et al. 2021). Typically, the heatsink has split heat pipes that lead to fins and other sources of additional surface area made of thermally conductive material like copper (Cu) or aluminum (Al) to draw away heat from the package (Ramakrishnan et al., 2021; Zhang et al., 2021). The passivity can be attributed to the dependence on natural convection in the air to dissipate heat away from the entire heatsink system, meaning that there are no mechanical or moving parts that are needed (Pedram & Nazarian, 2006; Zhang et al., 2021). However, this changes when a fan is added to the system, in which the system utilizes active convection and fluid flow through the heatsink fins to dissipate heat into the surrounding air (Fan et al., 2018). Essentially, the lack of moving parts in a heatsink-only system makes it a passive solution reliant on natural convection and pure surface area, while the addition of a fan makes it an active system, drawing air into and out of the heated surface area to reduce temperatures and increase efficiency.

In traditional liquid-loop cooling methods, thermal transfer is an active process in which liquid coolant flows through thermally conductive microchannels within a cold plate (Fan et al., 2018; Ramakrishnan et al., 2021). The cold plate contacts the CPU IHS leading to heat transfer; this heat energy then transfers to the conductive microchannels to where the heat can be transferred to the coolant liquid. The coolant may be transported from a pump that maintains a certain flow rate to ensure that coolant can circulate throughout the looped system. This liquid may then be cooled through a heat exchange system, such as a radiator, where the heat energy can be dissipated through active convection, such as a fan. The coolant then loops around the system again to help transfer more heat away from the cold plate after passing through the heat exchanger, ultimately circulating coolant around in a looped fashion (Fan et al., 2018; Ramakrishnan et al., 2021).

A major distinction here is that heatsinks are used primarily for air cooling and air convection with fan-heatsink or standalone heatsink setups, while cold plates are used for liquid-based systems. This will help to clarify future terminological use of heatsinks and cold plates going forward in this paper, as they are very similar in function and application, but the fundamental nature of how they function remain the differentiative factors between them.

Die-Heatsink and Die-Cold Plate Transfer Process

Thermal transfer begins first within the CPU package itself. The heat generating components of the CPU are all placed in a single package, called the die, on the processor substrate and other printed circuit board (PCB) components. Once these components are in use and generating thermal load, the heat may be transferred from the CPU die to the first layer of thermal interface material (TIM1). This TIM1 is applied in between the CPU die and IHS and may come in the form of thermally conductive paste or thermally conductive phase change material (e.g., a thermal pad) (Nylander et al., 2018; Shia & Yang, 2020). This layer can help to fill the air gaps at a microscopic level in order to maximize the thermal transfer area from the die to the IHS. By maximizing this area, thermal resistance can also be reduced as the contact surface area is increased (Shia & Yang, 2020).

The heat energy, once saturated by the IHS, must then transfer to the heatsink/cold plate. To facilitate this transfer and to maximize transfer area, a second layer of TIM (TIM2) is used in between the IHS and heatsink/cold plate and is also made of thermally conductive pastes or phase change material (Shia & Yang, 2020). Once the thermal energy is absorbed by a heatsink, the energy is then dissipated through active convection that is initiated by a fan propelling air into or out of the heatsink fins. In the case of a cold plate, the energy would be dissipated through the coolant in the system, to then being expelled through a heat exchanger device, such as a radiator (Fan et al., 2018).

The IHS acts as a protective cover to the CPU die and facilitates heat transfer by increasing the contact size of the CPU with the cooler heatsink, rather than attaching the cooler to the package itself. However, implementing an IHS may also impede overall thermal transfer as the IHS can also act as a thermal resistance barrier in between the CPU die package and heatsink (Ramakrishnan et al., 2021; Vassighi & Sachdev, 2006). Overall, however, total thermal resistance can be measured by combining the total resistances of TIM1, TIM2, IHS, and heatsink thermal resistance. This is a crucial component of thermal transfer as it determines how much of the heat is resisted for the difference in junction and ambient (coolant) temperature (Ramakrishnan et al., 2021; Vassighi & Sachdev, 2006).

Fan-Heatsink Cooling

As stated above, this method relies solely on a properly mounted and attached heatsink to the IHS of the CPU. Heat is generated in the die package then transfers to the TIM1 and subsequently the IHS. Thermal energy then transfers from the IHS to the TIM2 to facilitate transfer to the heatsink (Nylander et al., 2018). Keep in mind that all of the layers of thermal transfer also lead to extra layers of thermal resistance (Ramakrishnan et al., 2021). Heat pipes from the heatsink carry thermal energy to the heatsink fins, which provide adequate surface area for heat to be dissipated. A fan is attached to the heatsink fins, which is used to further expel the thermal load through active convection and active fluid flow (Zhang et al., 2021).

A performance study was published by researchers of Microsoft's Azure, CO+I, and Research departments, analyzing the efficiency of fan-heatsink air cooling involving the use of overclocked systems in a stress-tested environment using utilities such as Prime95, Intel XTU, and Cinebench to test the maximum possible thermal load on a microprocessor (Ramakrishnan et al., 2021). To determine overclocking stability, Prime95 and XTU workloads were run for 5 minutes each while Cinebench looped for three rounds. The processor used was the Intel i9-9900k, with a TDP of 95W, which is normally exceeded in overclocking scenarios. The default TjMAX specification for the processor model was determined to be 95℃. The TIM2 they used for their air-cooled system was a thermal paste with a thermal conductivity of 8.5 W/mK and was standard throughout their testing. The TIM1 was undetermined but was

described as indium-based in their findings for the i9-9900k they tested, and the heatsink they used was the model Intel XTS100H. Fan speeds remained constant with an airflow volume of 0.158 CFM per watt; the air inlet temperature also remained constant at 34℃. The fan-heatsink system was noted to have a higher thermal resistance (0.56℃/W) (Ramakrishnan et al., 2021).

Power testing indicated that liquid-based cooling (also at 34℃ inlet temperature) allowed for a power draw increase of 38% in Prime95, 42% in XTU, and 48% in Cinebench while also maintaining similar junction temperatures as fan-heatsink cooling (Ramakrishnan et al., 2021). The Prime95 workload tested (with 34℃ inlet temperature) with fan-heatsink cooling appeared to be the most rigorous and taxing to the system as the highest achieved core voltage and frequency during the overclocks were the lowest of all the other workloads in the testing suite. The workload imposed by Prime95 limited the core clock frequency to below 3.5 GHz and the core voltage to 1.1 volts. The results demonstrate that fan-heatsink cooling was the worst performing overall due to its higher thermal resistance when compared to other forms of cooling, such as liquid-based cooling or 2-phase immersion cooling (Ramakrishnan et al., 2021; Zhang et al., 2021). To be precise, air has a higher thermal resistance and lower thermal conductivity when compared to liquid and other fluids as a result of its less efficient heat transfer properties. Overall, however, the fan-heatsink solution was sufficiently able to stay under TjMAX with a heavy load, albeit with lower power draw, meaning that the solution is not as capable in dissipating heat with higher power draws as observed in the liquid-cooled results (Ramakrishnan et al., 2021).

It was also concluded that cooling performance in a fan-heatsink configuration could be optimized by increasing the airflow capabilities of the fan and by also increasing the heatsink fin density to increase the dissipation surface area. Another improvement strategy included the integration of enhanced heat pipes and vapor chambers within the heatsink system; the disadvantages would be the potential cost of engineering a solution that integrates these devices (Ramakrishnan et al., 2021; Vassighi & Sachdev, 2006).

It is important to note that these observations for potential improvement did not consider the effect of static pressure (mmH_2O) and the non-standardized measurement of the TIM2 thermal conductivity. Static pressure is a more common phenomenon in aircraft flight, in which the shape of aircraft wings helps to manipulate pressure differentials between the upper and lower sides of the wings to generate lift and streamline airflow across the wings (Rodi & Leon, 2012). Strictly speaking, static air pressure, in the context of computer cooling hardware, is an important measurement for a fan's ability to push air against and through an impeding object while still maintaining steady airflow. The same aircraft principle also applies to a heatsink fan, in which the operating fan blades overcome pressure differentials within the heatsink fins to push steady airflow across. An example of static pressure would be a fan's ability to push air through heatsink fin gaps, radiator fin gaps, and other obstructions (Siddarth et al., 2018). Intrinsically, high airflow optimization does not necessarily mean high static pressure; a high airflow volume fan may not have the ability to exert enough pressure through restrictive spaces, leading to difficulties in cooling and maintaining an airflow path through the space due to potential pressure differences (Siddarth et al., 2018). This becomes especially critical for physically tight datacenters and server rooms that may receive improper airflow due to greater potential obstructions in the inlets or outlets (Kuzay et al., 2022). The method by which they measured the TIM2 thermal conductivity was also not specified for the fan-heatsink and liquid-cooled processors but was determined to be 8.5 W/mK. However, the thermal resistance of the materials used in the TIM2 may vary depending on operating parameters such as junction temperature and material purity (Shia & Yang, 2020; Vassighi & Sachdev, 2006). As a result, many of the thermal conductivity figures and operating parameters used to take the measurements are not industry regulated or standardized, making it harder to pinpoint the exact thermal compound they may have used. These considerations are merely deductions and do not take away from the conclusion of the study but may complicate the process of reproduction and determining the TIM2 in the findings.

Liquid-Based Cooling

Liquid-cooling is a method that relies on an active cold plate to contact the CPU IHS, instead of a passive heatsink (Fan et al., 2018). The transfer processes are similar to fan-heatsink cooling until the thermal energy meets the cold plate. As the heat saturates the cold plate, heat is then transferred to the flowing coolant inside of the conductive cold plate microchannels (Ramakrishnan et al., 2021; Zhang et al., 2021). The coolant is propelled through the system with the use of a pump with either static or variable flow rate, depending on end-user preference. Heated coolant then meets a heat exchanging device, such as a radiator made of thermally conductive Cu, which acts as surface area in which the heat can be transferred away from the coolant. The saturated heat energy in the heat exchanger can then be dissipated away into the surrounding ambient air, typically facilitated with the use of fans (Fan et al., 2018).

The same study, conducted by researchers from Microsoft's CO+I, Azure, and Research teams analyzed the performance of liquid-cooling in overclocked data-center applications (Ramakrishnan et al., 2021). The stress-testing software suite remained unchanged to maintain consistency with their comparison with the fan-heatsink results, using Prime95 and Intel XTU for 5 minutes each, and Cinebench for three looped rounds. The same i9-9900k was used to also maintain consistency with a TjMAX of 95℃. The coolant used was polypropylene glycol likely to prevent galvanic corrosion as a result of its corrosion inhibitive properties and its application near oxide-prone metals such as Cu (Shia et al., 2021). The specific cold plate used was a CoolIT R4 passive cold plate with a split flow arrangement for the inner microchannels to improve coolant distribution across the IHS contact area of the plate. The TIM2 used was the same as the fan-heatsink system, having a measured thermal conductivity of 8.5 W/mK. To propel fluid around the loop, the team used the CoolIT AHx2 coolant distribution unit (CDU); within the CDU was an air-liquid heat exchanger for thermal transfer. The ambient air inlet temperature was sustained at 34℃ to maintain safe operating parameters for other integrated circuitry components (Ramakrishnan et al., 2021).

The same Prime95 workload with the ambient air inlet temperature of 34℃ resulted in core clock frequencies of above 3.5 GHz and core voltage of around 1.25v (Ramakrishnan et al., 2021). Higher core frequencies and higher core voltage can increase the phenomenon of leakage current, which in turn should increase the amount of power and heat that is generated (Thomas & Shanmugasundaram, 2018). The ability of the liquid-based system to sustain a higher clock frequency and core voltage (>3.5 GHz and ~1.25v, respectively) compared to the fan-heatsink based system (<3.5GHz and ~1.1v, respectively) supports the finding that liquid-based systems are objectively more efficient at transporting heat away from the processor, likely due to the coolant having higher thermal conductivity compared to air (Vassighi & Sachdev, 2006). The amount of dissipative surface area in the heat exchanging unit of the CDU may have also impacted the findings of these results but were not specified conditions mentioned in the study. The calculated thermal resistance of the cold plate (0.35-0.45℃/W) was also consistently lower than that of the fan-heatsink variant (0.56℃/W). These factors could explain how the cold plate solution was able to maintain safe operating temperatures despite dealing with a heavier thermal load (Ramakrishnan et al., 2021).

Reproduction of these results may depend on the silicon quality of the chip used, as there may be some process variation and silicon quality variation factors at play. Due to the use of an overclocked chip in the study, voltage and frequency results may vary significantly depending on the silicon quality of the die as core clock frequencies are pushed above the manufacturer designated limits and are by no means guaranteed. Thermal results may also vary depending on manufacturing tolerances (Ramakrishnan et al., 2021).

Effective Solutions in the Post-Dennard Age

Fan-heatsink cooling and liquid-based cooling appear to be effective methods for dissipating heat away from the CPU die package. Although fan-heatsink setups are less capable than liquid-based solutions, they are often more cost-effective than liquid-based solutions as a result of less design complexity, less mechanical parts, and ease of maintenance (Jalili et al., 2021). They are effective in that they are a sufficient thermal method to combat thermal throttling

and subsequently the slowdown protection mechanisms (core frequency and voltage decrease) that are incurred once the TjMAX barrier is reached under non-overclocked operating parameters (Zhang et al., 2021). This makes it particularly suitable for all types of computing in the current age, such as personal, enterprise, and data-center computing. However, this type of solution is not advised for systems that are overclocked or will experience heavy load over a long period (Jalili et al., 2021). As noted in the study, the benefits of overclocking (increased clock frequency and performance) could hardly be observed because of hardware instability due to the cooling system being unable to handle the thermal load (TDP and heat flux) of the die package (Ramakrishnan et al., 2021). Increasing core clock frequency (overclocking) often requires an increase in the operating voltage, which in turn leads to greater power consumption. Oftentimes, this increased TDP leads to thermal throttling in the case of traditional fan-heatsink systems, making them ill-suited for these types of scenarios (Jalili et al., 2021; Thomas & Shanmugasundaram 2018).

Liquid-based cooling systems are generally more effective and efficient in transferring heat away from the CPU package than fan-heatsinks as a result of the higher thermal conductivity of coolant compared to air (Ramakrishnan et al., 2021). Generally, they are capable of handling high-heat workloads as they have higher heat capacity and can transfer heat more efficiently, meaning that it may take longer for the coolant to reach its highest temperature or heat saturation point (known as steady-state). This in turn would lower overall temperature in short term high-heat workloads. These benefits would be less apparent for more persistent workloads, but generally the overall minimum achievable temperature should still be lower in liquid-cooling compared to air cooling resulting from improved heat flow due to an increase in the heat transfer coefficient of the ambient fluid, measured to be ~0.3 (°C/W) in the overclocking study; this is higher than that of the heat transfer coefficient of fan-heatsink solution measured at ~0.2 (°C/W). Note that the higher resistance of the fan-heatsink solution is not purely based on air convection resistance, but a culmination of other factors that vary depending on the material and purity of the heatsink used, such as conduction and spreading resistance as heat flows throughout the heatsink prior to the active heat-air convection (Ramakrishnan et al., 2021; Vassighi & Sachdev, 2006). Coolant and cold plate based liquid systems offer a heavy advantage compared to fan-heatsink systems as a result of their increased effectiveness in heat transfer capabilities and ability to handle more persistent and extreme workloads and conditions such as those commonly seen in overclocked computers and datacenter/server environments (Zhang et al., 2021). However, they may meet disadvantages in terms of system cost, design complexity, and risk with the handling of electrically conductive fluid such as coolant around sensitive components (Jalili et al., 2021; Vassighi & Sachdev, 2006).

Both fan-heatsink and liquid-based solutions appear to be sufficient to handle the modern needs of computing. Fan-heatsinks have their limitations in the modern day in the realm of high-performance enthusiast and data-center computing, where thermal demands have gotten increasingly higher proceeding the fall of Dennard scaling (Zhang et al., 2021). Higher temperatures increase the risk of common failure modes in semiconductor devices, such as time dependent dielectric breakdown (TDDB) and electromigration (Vassighi & Sachdev, 2006). The increasing temperature trend in semiconductor devices poses a great risk from a reliability and fabrication design standpoint. For example, one challenge chip designers face is electromigration in which higher junction temperatures reduce transistor reliability, which may necessitate the need for limiting the maximum current density on a chip. This issue is typically resolved by also limiting transistor density on the semiconductor device (Pedram & Nazarian, 2006). However, this also limits the number of transistors that can be packed on an integrated circuit (IC), leading to difficulties in maintaining significant and consistent generation-to-generation performance improvements over time.

The slowdown of transistor packing can be attributed to fundamental obstacles such as increased leakage current and power density, leading to a growing trend in chip manufacturing called “dark silicon”; this is a method used by chip designers that aims to limit or power-off portions of the die package to stay within reasonable TDP limits to avoid excessive heat (Taylor, 2012). The end of Dennard scaling and these electrothermal limits of transistors have complicated chip design to the extent to which manufactured ICs may no longer comply with Moore’s empirical law; Moore’s law specifies that the number of transistors on a chip double every year, but in recent years has varied from periods ranging from 18 months to 3 years (Cavin et al., 2012). This may lead to a slowdown in improvements in microprocessor performance over the coming years as well.

The growing concern over rising temperatures from the end of Dennard scaling has extended into the realm of semiconductor progression, performance, and reliability. Thermal mitigation is now an important issue that must also be addressed to ensure that technology can remain reliable and effective. As of currently, the best way for consumers and enterprise users to maximize these critical factors is to keep thermals within safe operating limits using a proper cooling method to decrease the chance of thermal throttling and premature failure.

Conclusion

Current physical and atomic limitations are beginning to affect transistor process refinement as demonstrated by the end of Dennard scaling and the near end of Moore's law. As the boundaries of transistor downscaling are pushed, numerous side effects, such as increased current and thermal density are to occur, enabling the growing trend of higher TDP and operating junction temperatures in microprocessors. This likely will have some impact on both performance and reliability philosophy in process-transistor design as thermal limitations also come into play (e.g., the more recent widespread growth of dark silicon to combat excess power draw). Modern thermal protections now must be in place in modern processors as they begin to run hotter, such as thermal throttling, the process in which the processor may underclock and undervolt to stay within the TjMAX boundaries. The takeaway is that the increased thermal demands of modern computing necessitate the use of proper cooling solutions in order to mitigate the drawbacks associated with these growing industry trends.

Fan-heatsink cooling designs have been a dependable solution for computers for a while even during the time of Dennard-scaling. Modern designs still largely rely on the main principle of transferring heat away from the CPU IHS by using TIM2 and a thermally conductive heatsink. The process of dissipation through active air convection then occurs when a fan is attached to this heatsink, blowing air into the conductive fins to transfer heat into the surrounding air. This method of cooling remains effective as modern computers can remain under TjMAX even in an overclocked, increased voltage scenario. However, overclocking is still not advised with this type of cooling as performance is limited due to electrothermal limitations and the potential for overheating as observed in a study conducted in overclocked data-centers. This can be owed to the observation that this solution has a higher thermal resistance and lower conductivity as a result of heatsink spreading and air convection resistance, leading to poorer thermal transfer. Reproduction of these results may cause variance due to the type of TIM2 used and the measured static pressure of the fan used. Despite these drawbacks, this solution can be economical and sufficient for computers that do not run persistent, high heat workloads and configurations due to setup simplicity, upfront costs, and costs on maintenance.

Liquid-cooling designs are a newer concept in thermal mitigation that is implemented by transferring heat away from the CPU IHS through direct contact via a cold plate with conductive microchannels for coolant to flow through. The heated coolant is propelled through a pump, which then leads the coolant to a heat-exchanging device such as a Cu radiator. Once thermal transfer is completed between the coolant and radiator, the residual heat is then dissipated into the air with active air convection from an attached fan. The process restarts as coolant loops around the system again to transfer heat away from the cold plate. This cooling is capable of handling high-heat scenarios where intensive applications are run on a long-term basis. It also was able to extract more performance out of an overclocked i9-9900k system as it was able to run higher clock frequencies and higher power draw with lower temperatures, making liquid-cooling a decent choice for high-heat computer configurations (such as overclocking) as well. These lower temperatures could be due to the observation that liquid-cooling has a lower overall thermal resistance and higher thermal transfer coefficient; the process of heat transfer from the die package to the coolant is also more efficient than that of fan-heatsink solutions, as spreading resistance and air convection resistance are less significant factors to consider in liquid-cooling. Liquid-cooling may also allow for greater dissipative surface area at the site of coolant-to-air heat exchange, helping to lower the coolant temperature more effectively. However, a thermal solution such as this may incur higher upfront cost and cost of maintenance due to equipment and the required expertise needed to install and maintain these systems. This may also introduce increased risk due to more points of failure within the loop and the handling of electrically conductive fluid near sensitive electronic components.

References

Bohr, M. (2007, Winter). A 30 Year Retrospective on Dennard's MOSFET Scaling Paper. *IEEE Solid-State Circuits Society,* 12(1), 11-13. doi:10.1109/N-SSC.2007.4785534

Cavin, K. R., Lugli, P., Zhirnov, V. V. (2012, April). Science and Engineering Beyond Moore's Law. *Proceedings of the IEEE,* 100(Special Centennial Issue), 1720-1749. doi:10.1109/JPROC.2012.2190155

Esmaeilzadeh, H., Blem, E., Amant, S. R., Sankaralingam, K., Burger, D. (2011, June). Dark silicon and the end of multicore scaling. In *ISCA '11: Proceedings of the 38th annual international symposium on Computer architecture* (pp. 365-376). Association for Computing Machinery. doi:10.1145/2000064.2000108

Fan, Y., Winkel, C., Kulkarni, D., Tian, W., (2018, Summer). Analytical Design Methodology for Liquid Based Cooling Solution for High TDP CPUs. In *2018 17th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)* (pp. 582-586). Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/ITHERM.2018.8419562

Ilsche, T., Schöne, R., Joram, P., Bielert, M., Gocht, A. (2018, May). System Monitoring with Io2s: Power and Runtime Impact of C-State Transitions. In *2018 IEEE International Parallel and Distributed Processing Symposium Workships (IPDPSW)* (pp. 712-715). Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/IPDPSW.2018.00114

Jalili, M., Manousakis, I., Goiri, I., Misra, A. P., Raniwala, A., Alissa, H., Ramakrishnan, B., Tuma, P., Belady, C., Fontoura, M., Bianchini, R. (2021, June). Cost-Efficient Overclocking in Immersion-Cooled Datacenters. In *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)* (pp. 623-636). Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/ISCA52012.2021.00055

Kuzay, M., Dogan, A., Yilmaz, S., Herkiloglu, O., Atalay, S. A., Cemberci, A., Yilmaz, C., Demirel, E. (2022, August). Retrofitting of an air-cooled data center for energy efficiency. *Case Studies in Thermal Engineering,* 36. doi:10.1016/j.csite.2022.102228

Leong, M., Narayanan, V., Singh, D., Topol, Anna., Chan, V., Zhibin R. (2006, June). Transistor scaling with novel materials. *Materials Today,* 9(6), 26-31. doi:10.1016/S1369-7021(06)71540-1

Mazouz, A., Laurent, A., Pradelle, B., Jalby, W. (2013, Summer). Evaluation of CPU frequency transition latency. *SICS Software-Intensive Cyber-Physical Systems,* 29, 187-195. doi:10.1007/s00450-013-0240-x

Nylander, A., Darmawan, C. C., Boyon, B. A., Divay, L., Samani, K. M., Ras, A. M., Fortel, J., Fu, Y., Ye, L., Ziaei, A., Liu, J. (2018, September). Thermal Reliability Study of Polymer Bonded Carbon Nanotube Array Thermal Interface Materials. In *2018 24th International Workshop on Thermal Investigations of ICs and Systems (THERMINIC)* (pp. 1-5). Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/THERMINIC.2018.8593282

Pedram, M., & Nazarian, S. (2006, August). Thermal Modeling, Analysis, and Management in VLSI Circuits: Principles and Methods. *Proceedings of the IEEE,* 94(8), 1487-1501. doi:10.1109/JPROC.2006.879797

Ramakrishnan, B., Alissa, H., Manousakis, I., Lankston, R., Bianchini, R., Kim, W., Baca, R., Misra, A. P., Goiri, I., Jalili, M., Raniwala, A., Warrier, B., Monroe, M., Belady, C., Shaw, M., Fontoura, M. (2021, August). CPU Overclocking: A Performance Assessment of Air, Cold Plates, and Two-Phase Immersion Cooling. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 11(10), 1703-1715. doi:10.1109/TCPMT.2021.2106026

Rodi, R. A., Leon, C. D. (2012, Fall). Correction of static pressure on a research aircraft in accelerated flight using differential pressure measurements. *Atmospheric Measurement Techniques,* 5(11), 2569-2579. doi:10.5194/amt-5-2569-2012

Schöne, R., Ilsche, T., Bielert, M., Gocht, A., Hackenburg, D. (2019, July). Energy Efficiency Features of the Intel Skylake-SP Processor and Their Impact on Performance. In *2019 International Conference on High Performance Computing & Simulation (HPCS)* (pp. 399-406). Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/HPCS48598.2019.9188239

Shia, D., & Yang, J. (2020, July) Analytical, Numerical and Experimental Study of Phase Change Material in TIM2 Application for High-Power Server CPUs. In *2020 19th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)* (pp. 158-165). Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/ITherm45881.2020.9190178

Shia, D., Yang, J., Sivapalan, S., Soeung, R., Amoah-Kusi, C. (2021, June). On Cold Plate Corrosion with Propylene Glycol/Water Coolant. In *2021 20th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm)* (pp. 212-219). Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/ITherm51669.2021.9503231

Siddarth, A., Eiland, R., Fernandes, E. J., Agonafer, D. (2018, June). Impact of Static Pressure Differential Between Supply Air and Return Exhaust on Server Level Performance. In *2018 17th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)* (pp. 953-961). Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/ITHERM.2018.8419536

Su, H., Liu, F., Devgan, A., Acar, E., Nassif, S. (2003, August). Full chip leakage-estimation considering power supply and temperature variations. In *Proceedings of the 2003 International Symposium on Low Power Electronics and Design, 2003. ISLPED '03.* (pp. 78-83). Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/LPE.2003.1231839

Taylor, B. M. (2012, June). Is dark silicon useful?: harnessing the four horsemen of the coming dark silicon apocalypse. In *DAC '12: Proceedings of the 49th Annual Design Automation Conference* (pp. 1131-1136). Association for Computing Machinery. doi:10.1145/2228360.2228567

Thomas, D., & Shanmugasundaram M. (2018, March). A Survey on Different Overclocking Methods. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/ICECA.2018.8474921

Vassighi, A., & Sachdev, M. (2006). *Thermal and Power Management of Integrated Circuits.* Springer.

Zhang, Z., Wang, X., Yan, Y. (2021). A review of the state-of-the-art in electronic cooling. *e-Prime – Advances in Electrical Engineering, Electronics and Energy,* 1. doi:10.1016/j.prime.2021.100009