2023 22nd IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm) | 979-8-3503-2166-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ITHERM55368.2023.10177551

Liquid jet impingement cooling of high-performance interposer packages: a hybrid CFD – FEM modeling study

H. Oprins[1], Tiwei Wei[1,2], Vladimir Cherman[1], E. Beyne[1]
[1] imec, Leuven, Belgium
[2] Currently with School of Mechanical Engineering, Purdue University, West Lafayette, IN 47907, US
Herman.Oprins@imec.be

***Abstract*— Bare-die direct liquid jet impingement cooling is an efficient cooling technique that has been successfully applied with various materials including Si, ceramic, metal and plastics. In this paper, we assess the application of liquid jet impingement cooling for a 3D system-on-chip (SoC) interposer package for high performance computing applications (HPC) with a 26x26 mm^2 and 8 high-bandwidth memory modules (HBM-2e), using a modeling study based on experimentally validated thermal models. The simulation study shows that the local customization of the impingement nozzle jet plate with a differentiation of the nozzle density depending on the local power density enables the cooling of large interposer packages with more than 2kW power dissipation.**

Keywords—3D stacking 2.5D interposer, jet impingement cooling, thermal modeling.

I. Introduction

Traditional CMOS technology scaling will continue into the next decade through innovations in technology, materials and device architectures. CMOS scaling is increasingly being complemented by design-technology-co-optimization to improve the system's power, performance, area and cost (PPAC). But with increasing cost and technology complexity, these approaches no longer deliver sufficient gains at the system level. This is especially true for data-intensive high-performance applications that are challenged by the difficulty of accessing data quickly enough. To further optimize the system's PPAC metrics, the semiconductor industry is increasingly looking at multi-chip heterogeneous integration solutions [1]. Following this approach, the different functions on the chip can be optimized separately (using different (node) technologies), and shorter and faster connections can be realized between a system's sub-components. These 'heterogeneous' implementations mainly rely on 2.5 or 3D chiplet approaches involving separately designed and processed chiplet dies [2].

There are examples in high-performance compute space like AMD's V-cache technology where 3D integration is used to bring extra SRAM memory closer to the CPU [3-4]. Another example is the use of a silicon interposer bridge to connect two CPU die [5]. Fig. 1 shows an example of such a multi-chip heterogeneous integration application, including logic die with high power density, different types of memory and HBM modules, using different types of 3D stacking technologies depending on the different 3D partitions.

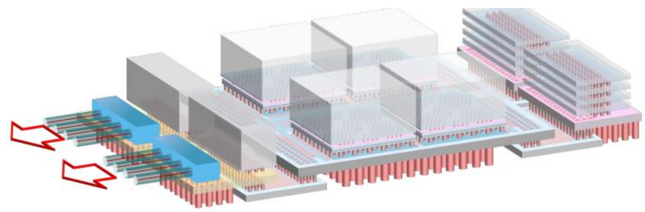

Fig. 1. Impression of a future high performance system including many heterogeneous components, using different flavors of 3D and 2.5D stacking.

Market demands for high-performance computing cause an increase in the power and power densities required for the high-performance applications, which requires the development of new, highly efficient cooling techniques [6]. Most applications for interposers combine high power components (such as logic chips) and temperature sensitive components (such as memory), that typically have a lower maximum allowable temperature. Since the components are thermally coupled in the package, the logic power will be limited by either the temperature limit of the logic or memory, whichever is reached first. This means there is a trade-off between the logic self-heating at the one hand and the thermal coupling with the memory chip at the other hand, which are impacted differently by the packaging materials and the cooling solutions [7]. Therefore, for an analysis of the thermal behavior of the package, both the self-heating of the chips as well as the thermal coupling to the other components are important parameters.

In this paper, a modeling study on the thermal behavior of a Si interposer with a large logic die and 8 HBM modules and the potential for direct liquid jet impingement of such system, assuming DI water as a coolant, is presented. In Section II, first the modeling methodology is explained and validated for a simplified interposer test vehicle. In section III, the test case of the HPC interposer package and the modeling details are introduced. Section IV presents the thermal modeling results for conventional cooling solutions. In Section V, the thermal performance of a customized liquid jet impingement cooler is predicted for the HPC interposer package.

II. Modeling Methodology & Validation

The objective of the thermal modeling study is to assess the thermal performance of the different cooling solutions for large interposer-based packages with multiple chips. The main results of the thermal analysis are the temperature distribution in the logic die and HBMs for different boundary conditions and power levels, allowing the comparison of the thermal performance. The two most common options for the package level thermal to simulate the temperature distribution in all materials are:

• Conduction-convection modeling using conjugated heat transfer with computational fluid dynamics (CFD);

• Thermal Conduction modeling using the finite element method (FEM).

In both approaches, the heat conduction in the solid structures is modeled using finite elements. In the conduction-convection models, the coolant surrounding the 3D package (liquid or air) is explicitly modeled, typically by means of CFD. With this approach, the velocity and temperature distribution of the fluid is included in the thermal model. However, this type of simulations requires a very large number of elements to correctly resolve the boundary layers and can very time consuming. In the conduction-only models at the other hand, the effect of the convection on the solid structure is included in the model by means of convective boundary conditions. The forced convection or natural convection heat transfer acting on a surface is represented by an equivalent heat transfer coefficient value or distribution. This is a simplification of the fluid flow and temperature distribution, since a single value or distribution is used for the ambient temperature and might lead to a significant error. However, this approach is useful to make a relative thermal behavior comparison for different configurations that have the same fluid conditions, as the main focus of this study is the internal heat conduction inside the package.

In this section, the modeling methodology is experimentally validated in 2 steps to assess the accuracy of the predicted temperature distribution within the interposer-based packages using a thermal test chip:

1. A conduction-only thermal model (FEM) of the interposer test vehicle to validate the internal heat spreading for different boundary conditions;
2. A hybrid CFD-FEM thermal model where the heat transfer coefficient distribution from the CFD model is used as a boundary condition for the FE model.

A. 2.5D interposer thermal test vehicle

The interposer packages used for the thermal evaluation of the modeling methodology are shown in Fig. 2. Two thermal test chips (8x8 mm^2) with integrated heaters and sensors, referred to as PTCQ (Packaging Test Chip version Q [8,9], are stacked face down on a 20×10 mm^2 Si interposer with a thickness of 100 µm using CuSn µbumps with 40 µm pitch. The interposer stacks are packaged in two versions of a 35×35 mm^2 ball grid array package (BGA): lidless packages (Fig.2a, allowing the cooling solutions to be directly applied to the backside of the chips, and lidded packages (Fig.2b) that require a TIM between chip and lid. A first measurement environment that is used is a plastic measurement socket that has been attached to the PCB (Fig 2.d). The BGA package and a plastic spacer are placed inside the socket. This socket configuration allows easy mounting and since there is no permanent connection, many test packages can be measured in the same socket. This socket is used to perform the thermal analysis of packages mimicking low power applications. Therefore, this socket is referred to as '*low power socket'*. To emulate high power applications, a dedicated socket with active cooling is used. In this socket, there is a direct contact between the backside of the chip package and the heat sink. On top of the heat sink, an 80x80 mm^2 fan is mounted to provide the forced convection cooling, resulting in a much lower overall thermal resistance for this socket configuration. This socket is referred to as '*high power socket*'. Using this socket, the main part of the heat generated in the package will be removed through the top side of the package, while in the case of the plastic socket without heat sink, the main part of the heat will be removed through the bottom side of the package, towards the PCB.

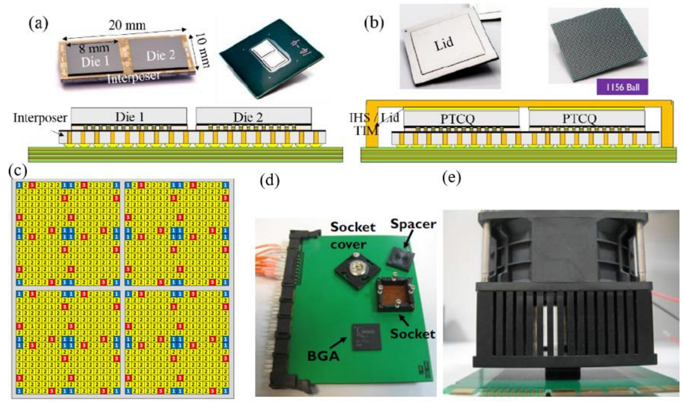

Fig. 2. Configurations of 2.5D Si interposer packages with the 8x8 mm^2 thermal test chips: (a) Bare die (lidless) package; (b) Lidded package with TIM and lid. (c) Layout of 8x8 mm^2 thermal test chip where the yellow cells represent the heater areas. (d) Low power socket. (e) High power socket.

B. 2.5D interposer conduction model (FEM)

For the validation of the finite element model, detailed parameterized 3D finite element models for both the molded and the exposed 3D packages have been constructed for the socket configuration with and without cooling and for the soldered configuration by using a general purposed finite element software package. By means of steady state thermal simulations, the temperature distribution inside the package configuration is modeled for the different power dissipation patterns. The detailed FE models include the package laminate, both chips in the die stack and the interconnections between the chips and the package and the locations of all heating elements on the test chip. Figure 3 shows the constructed models for the molded and bare die 3D packages. To reduce the complexity of the thermal models, several structures have been replaced by elements with equivalent thermal properties. The µbump and Cu pillar arrays embedded in underfill material are replaced by a material with equivalent in plane and out of plane thermal properties. Using a unit cell modeling approach for a µbump with a certain geometry and pitch the equivalent in plane and out of plane thermal conductivity can be found taking into account the local thermal spreading and constriction resistances. Experimental values for the equivalent resistance of such mixed layers of interconnections and underfill material can be extracted from the temperature measurements with uniform power dissipation [10]. The simplification of the die-die and die-package interconnections is shown in Figure 3.

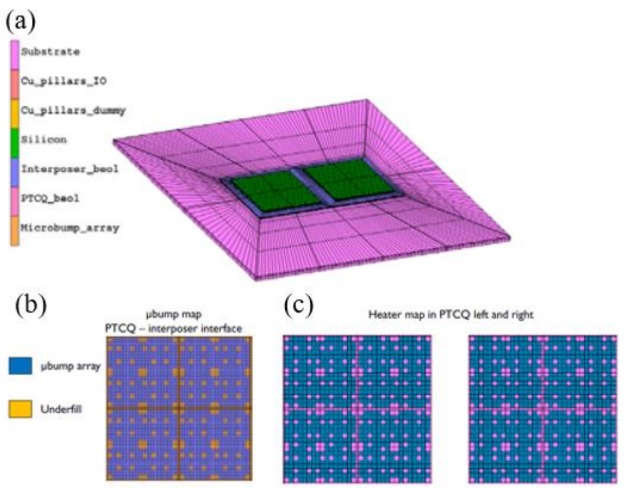

Fig. 3. Finite element model of the interposer package. (a) Mesh of the exposed die package. (b) Distribution of equivalent properties for the die to interposer 3D interface. (c) Power map in the two test chips.

TABLE I. COMBINATIONS OF INTERPOSER PACKAGES AND MEASUREMENT CONDITIONS.

R_{ij} (K/W)	High power socket		Low power socket	
	Lidded package	Bare die package	Lidded package	Bare die package

For the validation measurements of the combinations listed in Table I, quasi-uniform power is dissipated in one test chip, while the temperature distribution is measured in both chips. The measured and modeled temperature distribution are shown in detail in Fig. 4 for the bare die package in the high-power socket in terms of a temperature contour plot and temperature profile in both dies. For this configuration the lowest temperature increase is observed of the four configurations: 1.1K/W for self-heating, and a very low coupling of 0.5K/W. Figure 5 shows a summary of all four measurement results, the temperature profile is plotted along a line crossing the two dies on the interposer (indicated in Figure 4). Figure 5 shows a very good agreement between the modeling and measurement results of all four test configurations of the interposer package, indicating that the FE conduction model is capable to accurately predict the chip self-heating in the package as well as the thermal coupling to the other chip in the package.

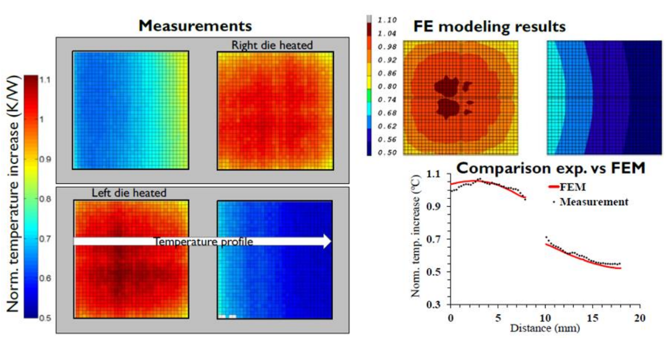

Fig. 4. Validation of the condution FE model for the bare die interposer packagea in the high power socket: temperature contour plots and temperature profile.

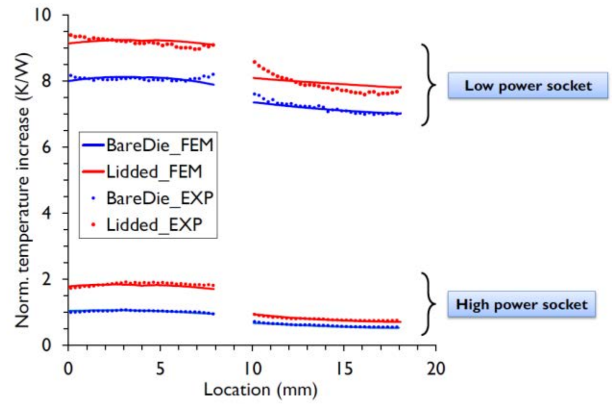

Fig. 5. Validation of the condution FE model for all four configurations

C. Hybrid CFD-FEM thermal model

To allow a parameterized thermal modeling study of an interposer package cooled by direct liquid jet impingement cooling, a hybrid modeling methodology has been introduced and experimentally validated in [11]. In this approach, the flow rate dependent heat transfer distribution, calculated from the detailed, package level CFD simulation is applied as a convective boundary condition to the FE conduction model. The schematic of the impingement coolers on the lidded package and lidless packages, used for the experimental validation study are shown in Fig.6, in which the four main parts can be distinguished: inlet plenum, outlet plenum, nozzle plate and impingement cavity, covering two PTCQ die. The cooler has been designed to match the 35×35 mm^2 footprint of the BGA packages and to fit the Si interposer and bonded chips in the cooler cavity. For the cooler design, unit cells of 2×2 mm^2 are used with 600 µm diameter nozzles, resulting in a 4×4 array of inlet nozzles and staggered 5×5 array of outlet nozzles centered on top of each chip. The cooler was 3D printed using the water resistant Watershed material [12], and characterized experimentally for flow rates between 300 ml/min and 1 L/min.

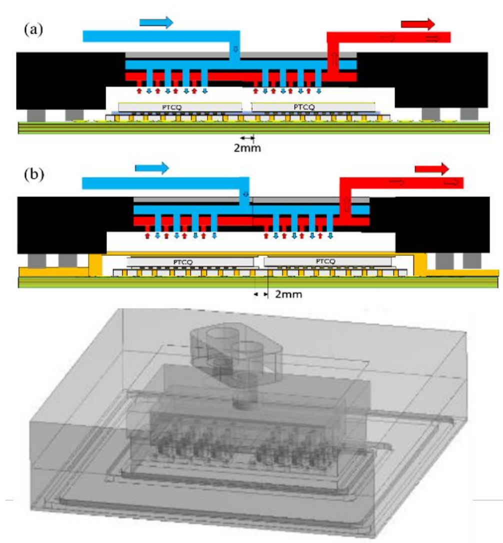

Fig. 6. Demonstrator of the package level jet impingement cooler for the interposer package with two PTCQ thermal test vehicles [11].

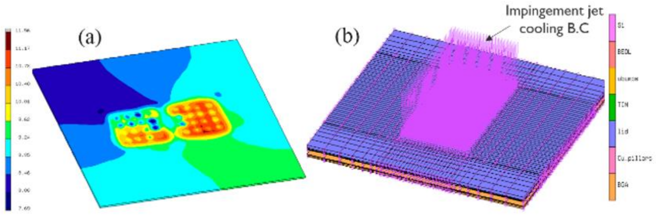

Fig. 7. Hybrid CFD-FE modeling approach: extracted heat transfer coefficient from CHT CFD model (a) is applied as convective boundary condition to the conduction FE model (b).

For the modeling study, a package level conjugate heat transfer CFD simulation (Ansys Fluent) with a transition SST turbulence model was performed. The grid sensitivity study using the Richardson extrapolation resulted in a low truncation error of 0.3 %, indicating a grid independent solution. From the CFD simulation the heat transfer distribution maps as function of flow rate were extracted, on top of lid in case of lidded package, or on top of bare die in case of bare die package. This heat transfer distribution was used as a convective boundary condition for the parameterized finite element model (Hexagone Marc) containing 400,000 including the PCB, the solder balls, the package laminate, the interposer, logic and memory chip, the interconnections between the chips and the package, such as BEOL, micro-bump layer, Cu pillars and underfill. Fig. 7a shows the heat transfer coefficient extraction results from the full cooler level CFD model. The extracted heat transfer coefficient map is applied on the corresponding lid surface in the FEM model as a convective boundary condition shown in Fig.7b.

Fig.8 shows the FE modeling results for the lidded package using the extracted heat transfer coefficient from the CFD models at different flow rates as a boundary condition. The comparison with the experimental PTCQ measurements shows a good agreement for both the active heat chip as well as the passive chip. Therefore, the CFD model and FE models for the lidded package cooler are successfully validated (8x8 mm^2) and can be used for the extrapolation of the cooling conditions.

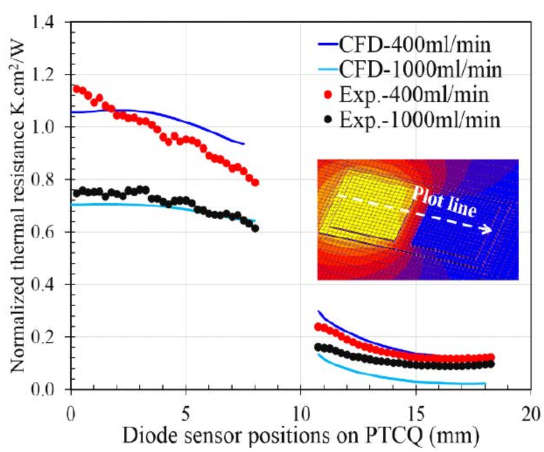

Fig. 8. Experimental validation of the hybrid CFD-FE modeling approach for impingement cooling on the interposer package with two thermal test chips, of which one is heated.

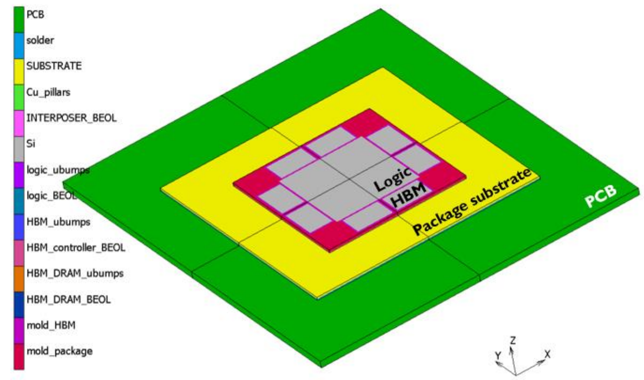

Fig. 9. HPC interposer test case with large logic die and 8 HBM modules.

III. HPC INTERPOSER TEST CASE

A. Test case description

This section describes the test case that is considered for the thermal analysis of the interposer-based package with large logic die and multiple HBM modules. Fig. 9 shows a representation of the test cases with a 26x26 mm^2 logic die and 8 HBM modules (consisting of a logic controller and 12 memory dies) on a Si interposer. The gap between the logic-memory stack and the HBM is assumed to be filled with mold material. The dimensions and thickness values of the components and layers in the package are listed in Table II. The same height is assumed for the logic die and the HBM module, assuming planarization after assembly. The TIM 1 layer can then be applied on a flat surface of memory on logic stack, HBM and mold. For the power dissipation, a uniform power dissipation is assumed for the logic die and each memory die of the HBM module.

TABLE II. HPC INTERPOSER TEST CASE PARAMETERS.

Component	X (mm)	Y(mm)	Z (µm)
Logic Si	26	26	710 (matched to HBM height)
Logic BEOL	26	26	10
HBM	12	8	720
HBM ctrl Si	12	8	50
HBM ctrl BEOL	12	8	3
HBM DRAM 1-11 Si	11.5	7.5	35
HBM DRAM 12 Si	11.5	7.5	66
HBM DRAM 1-12 BEOL	11.5	705	3
µbump interface			15
Logic / HBM to interposer µbump interface thickness			15
Si interposer	42	32	100
Si interposer BEOL	42	32	10
Cu pillars			50
Package substrate	65	65	500
PCB (Jedec test board size)	114.3	101.6	1570

B. FE thermal model details

Based on the symmetry present in the package layout, a quarter symmetry has been used for the model generation in order to reduce the required computational time. A procedure has been developed to parametrize the model, allowing easy changes of the geometrical parameters, material properties and boundary conditions in the model. For this model, roughly 700,000 elements are used to create an accurate thermal model that can include all the required detail for the logic power map, the HBM details and the many vertical layers in the model. The typical calculation time for this model is 200s for a single iteration. In case temperature dependent material properties are considered, several iterations are required (typically 3 or 4). Fig. 10 shows an exploded view of the relevant Si parts in the model in the logic and HBM layers. In Fig. 11, a cross section of the structure is shown to reveal the layers in the structures and the equal height of the logic die and HBM module.

The heat is generated in the thin active region of the chip. These heat fluxes have been used to apply the heat generation for the logic die, the memory dies and the different layers of the HBMs. For the PCB side, and equivalent thermal resistance of 20 K/W is considered. For the cooling solution, the boundary condition is applied to the top side of the structure.

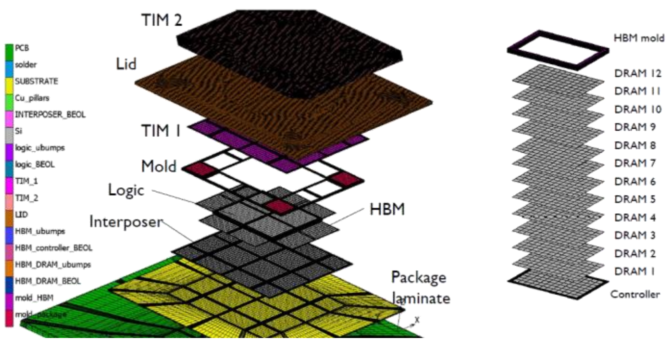

Fig. 10. Exploded view of the FE model visualizing the different layers in the test case.

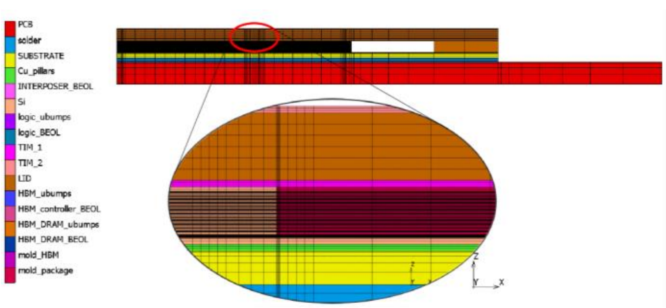

Fig. 11. Cross-section of the FE model.

IV. Modeling Analysis

In the first analysis step, the thermal finite element model will be used to assess the thermal behavior of the interposer package for conventional cooling solutions including air cooling and liquid cooling. For this study, a typical thermal resistance of 0.1 K/W is assumed for the air-cooling solution, while a thermal resistance value of 10 mm^2-K/W is assumed for both TIM1 and TIM 2. The lid is assumed to be 1 mm Cu and the ambient temperature is 25ºC. Fig.12 shows the self-heating and thermal coupling temperature increase for the logic die and the HBM module, showing a large thermal coupling between the logic and HBM, mainly through the Si interposer.

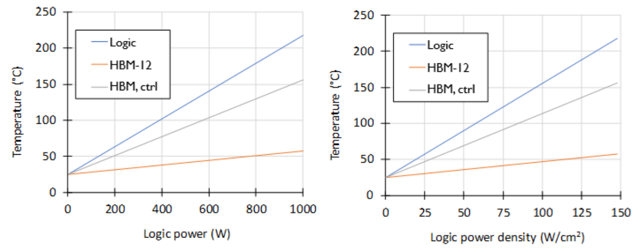

Fig. 12. Temperature increase in the logic and HBM module (logic controller and top memory) as function of logic and HBM power.

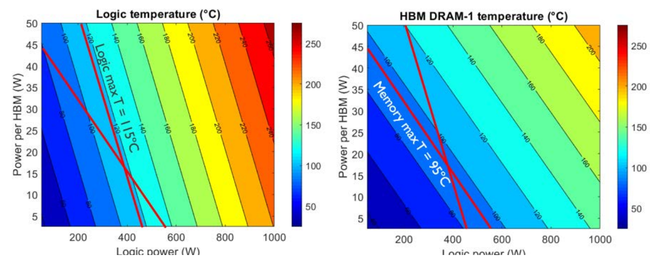

Fig. 13. Contour plots of the logic and HBM (DRAM die 1) maximum temperature as function of logic and HBM power for air cooling conditions with a thermal resistance of 0.1 K/W.

The results from Fig. 12 can be used to extract equivalent self-heating and thermal coupling resistance values. Using superposition, the maximum allowed logic and HBM power can be estimated for a given maximum allowable logic and HBM temperature, as illustrated in the temperature contour plots of Fig. 13. Assuming a maximum logic temperature of 115ºC and a maximum memory temperature of 95ºC, an estimated maximum power of 400W for the logic and 15 W per HBM module is found at the intersection of the iso-temperature lines. A detailed simulation for a logic power of 400W (60 W/cm^2) and 10W per HBM module, results is a maximum logic temperature of 112.2°C and a HBM temperature of 89.1°C, 97.2°C and 97.2°C for the HBM controller, the bottom memory die and the top memory die respectively. Fig. 14 shows the vertical temperature profiles for this case, at the locations of the logic die and the HBM module. The vertical temperature profile on the left-hand side shows that the main gradients are caused by the TIM 1 and TIM 2 layers. Using more advanced thermal interface materials, these vertical gradients can be reduced. The graph on the right shows the gradient with the HBM module with the highest temperature at the bottom of the module due to the strong thermal coupling with the logic die.

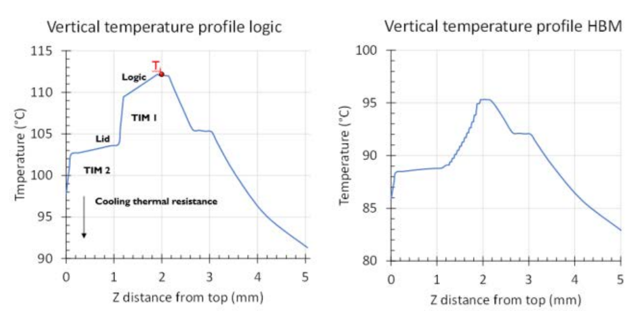

Fig. 14. Vertical temperature profiles for air-cooling modeling case of the HPC interposer package.

.

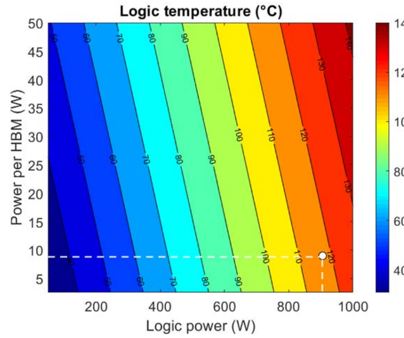

Fig. 15. Contour plot of the logic die maximum temperature as function of the logic and HBM power for liquid cold plate conditions with a thermal resistance of 0.03 K/W.

A similar analysis can be performed for a liquid cold plate cooling configuration with an assumed thermal resistance of 0.03 K/W. The contour plot in Fig. 15 of the superposition of logic and HBM modules heating, shows that for those cooling conditions, a maximum logic power of 950W can be dissipated for a total HBM power of 80W (10W per module) in the interposer package.

V. Impingement Cooling Analysis

Direct liquid jet impingement cooling on the backside of the Si chip is a highly efficient cooling method as the thermal interface material between the cooling and chip is eliminated and the parallel, vertical supply of liquid coolant results in a uniform chip temperature distribution. In [13], we experimentally demonstrated this cooling concept with DI water as a coolant, using a 3D printed impingement cooler assembled to a 23 x 23 mm^2 thermal test die with a power dissipation of 1 kW. In this section, the cooling liquid jet impingement cooling performance will be extrapolated for the HPC interposer test case introduced in Section III.A.

A. Modeling methodology

The starting point for the analysis is the fabricated impingement cooler from [13] shown in figure 16. The jet impingement array for this 23x23 mm^2 test chip consists of an 11x11 array of jets and an array of 12x2 outlets distributed in between the inlet nozzles, with an inlet nozzle pitch of 2 mm and a nozzle diameter of 600 µm. A conjugate heat transfer CFD model (Ansys Fluent) using a transition SST turbulence model has been built for the full cooler geometry and the results for the thermal and hydraulic behavior have been experimentally validated by measurements, as shown in Fig. 17. The validated CFD model is then used to predict the thermal performance (in terms of heat transfer coefficient distribution on the chip surface, or on the lid in case of a lidded package) and pressure drop over the complete cooler and over the nozzle plate. The extrapolated results for 1 mm pitch are shown in Fig. 17. This indicates that the thermal performance for a fixed flow rate can be improved by reducing the nozzle pitch, at the expense of a higher pressure drop.

Next, a modeling parameter sensitivity study has been performed for range of parameter values for the nozzle diameter, nozzle pitch and flow rate. From these simulation coefficient and pressure drop as function of the local flow rate, the nozzle pitch and the nozzle diameter:

$$h = 8440 \cdot d^{-0.4157} \cdot \dot{V}_{nz}^{\,(0.7843-0.6624\cdot d)} \quad (1)$$

$$\Delta p = 0.655 \cdot d^{-4} \cdot \dot{V}_{nz}^{\,1.76} \quad (2)$$

where h is the average heat transfer coefficient [W/m^2-K], d is the nozzle diameter [mm], V_{nz} is the local nozzle flow rate [ml/min] and Δp is the pressure drop over the nozzle plate [bar]. Fig. 18 shows the measured normalized thermal resistance as function of flow rate per jet nozzle for two different coolers for different test chip sizes, with the same nozzle unit cell geometry (2 mm nozzle pitch and 600 µm diameter) This shows that the thermal performance is scalable for a fixed flow rate per nozzle and can easily be recalculated for different chip sizes. This concept is now exploited to simplify the modeling methodology for the interposer package with the large logic die and HBMs.

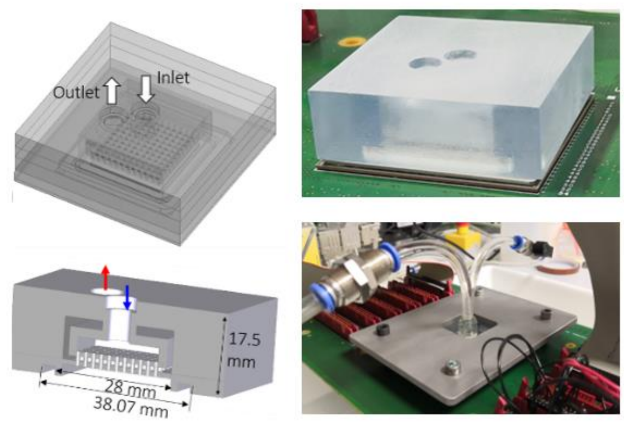

Fig. 16. Experimental demonstrator of the 3D printed impingement cooler on the large die 23x23mm^2 thermal test vehicle.

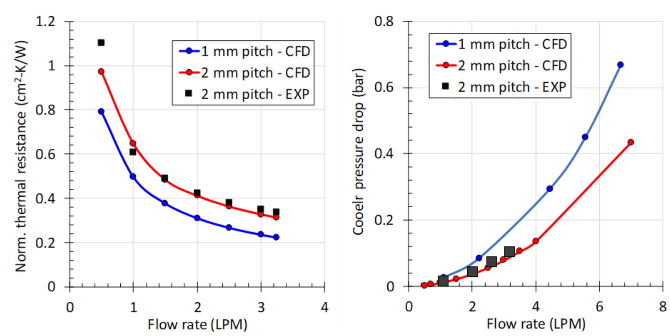

Fig. 17. Experimental characterization of the thermal and hydraulic performance of the bare die jet impingement cooler on the large die test vehicle.

In the design of the package level impingement cooler for the HPC interposer package (Section III.A), the nozzle array configuration (local nozzle pitch and diameter) can be optimized, allowing for a different nozzle pitch at the location of the logic (with higher power density) die and the HBM (with lower power density, but stricter temperature limit), respectively. For the modeling study of the HPC interposer package with impingement cooling, the finite element model for the bare die package introduced in Section III.B is used. As a cooling boundary conditions, an equivalent heat transfer coefficient is applied to the backside surface of the logic die and the HBM modules. In the first part of the analysis, it is assumed that the coolant flow supplied to the cooler is divided between the different chips on the interposer package in the manifold layer of the cooler. To estimate the respective equivalent heat transfer coefficient, following procedure is followed:

1. A total coolant flow rate for the package level cooler is chosen.
2. Jet array dimensions are chosen for the logic die and HBM. Figure 19 shows an example of jet distribution on the logic die (2 mm pitch) and HBM (4 mm pitch).
3. An analytical 1D flow network using equation 2 is used: assuming an equal pressure drop across the nozzle plate for each jet in parallel, the local flow can be estimated using equation 2 for each nozzle.
4. Equation 1 is used to estimate the average heat transfer coefficient for each nozzle array using the estimated local flow rate in step 3, to be applied as a local boundary condition in the package level finite element model.

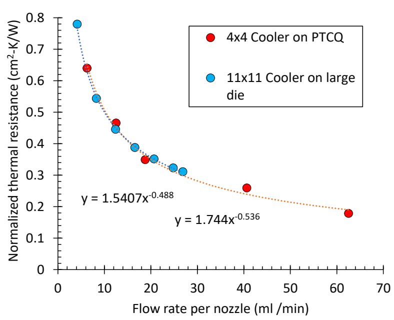

Fig. 18. Measured normalized thermal resistance of direct liquid mulit-jet impingement cooling on the Si backside for two different test vehicle sizes: 8x8 mm^2 PTCQ test chip and 23x23 mm^2 large die test vehicle.

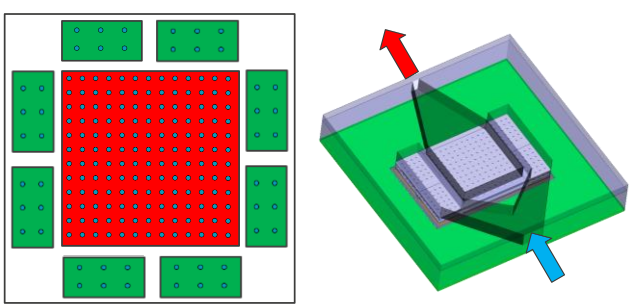

Fig. 19. Visualisation of the inlet nozzle arrays on the logic die (red) and HBM modules (green). The outlet nozzles are distributed between the inlet nozzles (not shown here).

B. Analysis for parallel cooling

In this section, the cooling performance of a direct jet impingement cooler with parallel cooling paths for the logic die and the HBM modules is assessed using the methodology introduced in the previous section. For the test case, a power of 1500W is assumed in the 26x26 mm^2 logic die (225 W/cm^2). For the HBM, a power of 8W (8.33 W/cm^2) is assumed in the logic controller of the HBM, while a power of 1W is assumed for each of the 12 memory dies in the HBM (1.16 W/cm^2), resulting in a total power of 20W for each of HBM modules. For the cooling, a flow rate of 3.5 LPM is considered. The considered jet pitch is 1 mm for the nozzles above the logic die and 2 mm for the nozzles above the HBM modules. Figure 20 shows the temperature distribution in the package for this power and cooling conditions. Fig. 20.a shows the complete ¼ symmetry thermal model, while Fig. 20.b shows the temperature distribution in the logic and HBM modules only. The temperature gradient in the HBM is shown in Fig. 20.c (vertical temperature profile) and Fig. 20.d (contour plot). The simulations show a significant vertical temperature gradient over the HBM module. The top memory die, DRAM 12, is cooled directly by the liquid coolant. The controller, at the bottom of the HBM module has the highest temperature. The peak temperature of the controller is almost 30ºC higher than the top memory die. A large part of the internal thermal resistance of the HBM module is caused by the 3D bonding interface, in this case µbumps and underfill with an equivalent thermal resistance of 3.8 mm^2-K/W [10]. When the thermal properties of hybrid bonding (0.9 mm^2-K/W) are used [14,15] for the 3D bonding interface, the temperature difference between DRAM-1 and DRAM-12 is reduced by 40%. The temperature of DRAM-1 itself reduces by 12%, as shown in Fig. 21.

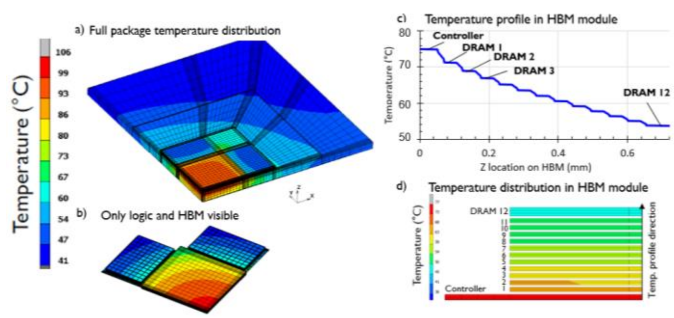

Fig. 20. Temperature distribution in the HPC package: cooling flow rate 3.5 LPM, 1500 W logic power, 20W per HBM. 1 mm nozzle pitch logic, 2mm nozzle pitch HBM.

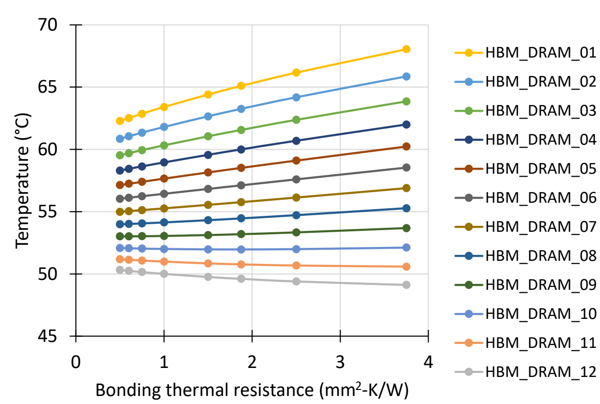

Fig. 21. Impact of the HBM 3D bonding thermal interface on the temperature gradient in the HBM module.

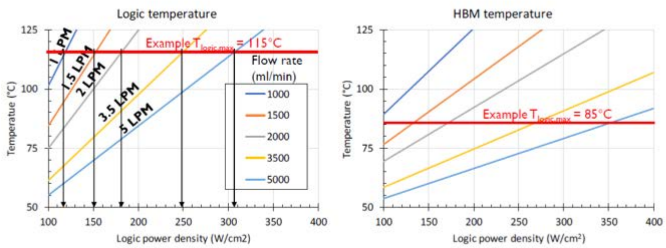

Fig. 22. Logic self-heating and thermal coupling to HBM module for different cooler flow rates, as function of the logic die power.

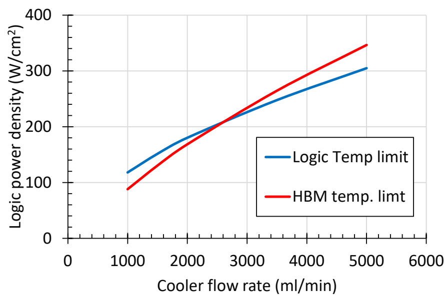

Fig. 23. Logic power limit (20W per HBM) as function of cooler flow rate for the impingement cooler with a nozzle pitch of 1 mm at the location of the logic and 2 mm at the location of the HBM modules.

The temperature in the chips on the interposer is partially caused by self-heating of the chip itself, and partially by thermal coupling to the neighboring chips. Fig. 22 shows the maximum temperature of the logic die and HBM module for different logic power values, assuming of fixed power of 20W per HBM, and for different cooler flow rates. It can be seen that the increase of flow rate results in a lower logic and HBM temperature. From these graphs, it is furthermore possible to determine the maximum logic power limit for specified maximum allowable temperature limits for the logic and HBM. In this study, 115ºC and 85ºC are used for the maximum logic and HBM temperature respectively. By increasing the flow rate from 3.5LPM to 5 LPM, the logic power density can be increased from 250 W/cm^2 (1590W logic power) to 300 W/cm^2 (2028W), enabling the cooling of 2kW+ interposer packages. Fig. 23 show this maximum power density for both the logic die as well as the HBM module as function of the cooler flow rate, where the logic power will be determined by the minimum of those curves. It can be seen that for this cooler configuration (2 mm nozzle pitch on the HBM, 1 mm nozzle pitch on the logic), the logic power is limited by the HBM maximum temperature at low flow rate values, while it is limited by the logic temperature for high flow rate values. This is caused by the higher thermal coupling between logic and HBM at low flow rate values, and a significant reduction for this coupling at higher flow rates.

Since the design of the nozzle sizes and locations is flexible, the cooler design can be further improved by changing the relative nozzle density for the logic and HBM modules. In that way, the relative flow rate and the local cooling performance can be optimized in order to maximize the logic die power before either temperature limit is met. For the test case, the logic power is now increased to 2000 W (300 W/cm^2) to investigate for which conditions the requirements can be met. A better thermal performance is achieved for a higher nozzle density. Therefore, a high nozzle density is desirable at the location of the logic die with high power density and a lower nozzle density at the HBM location with lower power density. As a results, a higher flow rate flow rate will be directed towards the logic die, resulting in a significant reduction of the logic temperature. Despite the lower flow rate at the HBM location, the temperature at the HBM controller and DRAM-1 still reduces due to the thermal coupling between HBM and logic. In this way, an optimized distribution of the coolant flow between logic and HBM modules, can reduce the temperature of all components. Fig. 24 shows the peak temperature of the logic die and the HBM (controller, bottom DRAM die, top DRAM die) for four different nozzle configurations for a total flow rate of 3.5LPM:

1. Logic: 2 mm pitch, HBM: 2 mm pitch
2. Logic: 1 mm pitch, HBM: 1 mm pitch
3. Logic: 1 mm pitch, HBM: 2 mm pitch
4. Logic: 1 mm pitch, HBM: 4 mm pitch

The modeling results show that for a parallel cooling configuration of the 2kW logic die and HBM modules, the maximum temperature requirements can be met for a fixed flow rate of 3.5 LPM, using a tailored design of the nozzle pitch depending on the local power density. Reducing the local coolant flow rate at the location of the HBMs, in favor of the logic location, can reduce the logic temperature increase by 40%, as well as the HBM peak temperature increase by 25% due to the thermal coupling in the package.

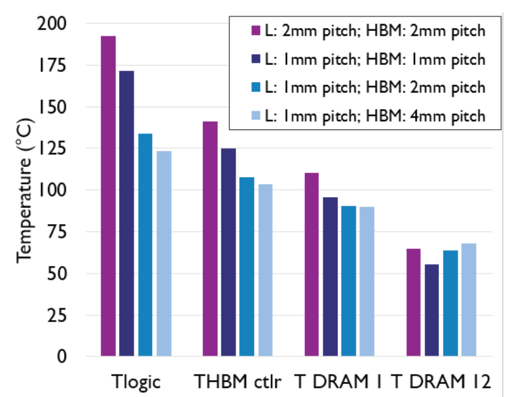

Fig. 24. Impact of nozzle configurations on the maximum temperature in the logic and HBM modules of the HPC interposer for a logic power of 2kW and HBM power of 20W per modules and a coolant flow rate of 3.5LPM.

C. *Further improvement – serial cooling configuration*

In the previous section, the coolant flow was divided over the logic die and the HBM modules, to cool those components in a parallel way, with the same reference inlet coolant temperature. Since the coolant temperature does not increase that much for relatively high flow rate values, it is possible to design the cooler manifold in such a way that the components are cooled consecutively in a series configuration instead of a parallel configuration. The local outlet of the cooling part of one component is connected to the inlet of the other component. This results in a higher local flow rate per component for the same system flow rate at the cost of a higher pressure drop. For the analysis of the series cooling configurations, the analytical model for the estimation of the convective boundary conditions is extended with a parameter for the inlet temperature for each component, based on the coolant temperature increase in the previous component based on the energy balance.

Fig. 25 shows the impact of this series connection for the test case with a nozzle pitch of 1 mm above the logic and 2 mm above the HBM. The figure compares the logic and HBM peak temperature for the parallel cooling configuration with the 2 series configurations: 1) the coolant first cools the logic die and then the HBM, 2) the coolant first cools the HBM modules and then the logic. For this analysis, the analytical model is extended with a parameter for the inlet temperature for each component, based on the coolant temperature increase in the previous component based on the energy balance. When the logic die is cooled first, a temperature reduction of 9ºC is observed for the logic die, due to the higher local flow rate compared to the parallel cooling configuration. The temperature in the HBM controller is however slightly higher due to the higher local inlet temperature. When the HBM modules are cooled first, there is only a very small change for the logic temperature (+0.5ºC) due to the relatively low HBM power. There is however a significant impact on the HBM temperature. Compared to the parallel connection, the series cooling connection can achieve an additional temperature reduction of 20% and 50% of the bottom and top DRAM layer of the HBM respectively, however at the expense at a higher pressure drop.

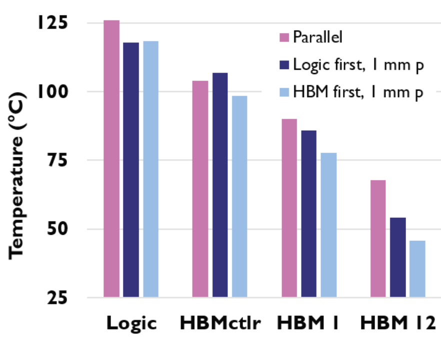

Fig. 25. Comparison of parallel and series cooling configurations in the impingement cooler for a 2 kW logic die and 8 HBM modules for a cooler flow rate of 3.5 LPM.

VI. Conclusions

In this paper, we assess the application of liquid jet impingement cooling for a 3D system-on-chip (SoC) interposer package for high performance computing applications (HPC) using a modeling study based on experimentally validated thermal models. The considered SoC consists of a 26 x 26 mm^2 logic die with a power dissipation of 2 kW and 8 high-bandwidth memory modules (HBM-2e) consisting of 12 DRAM layers. In the presented study, a hybrid modeling approach is used for the thermal analysis of the 3D-SOC interposer package. First, a full conduction-convection model is performed using conjugated heat transfer computational fluid dynamics (Ansys-Fluent) to simulate the heat transfer in the package and the convective heat transfer in the impinging coolant. The modeling results have been experimentally validated with the measurement data (up to 1 kW) from both thermal test vehicles for fabricated coolers with 2 mm and 1 mm nozzle pitch. In the second step, a FE model with boundary conditions extracted from the CFD model is used to perform a sensitivity study.

The modeling results show that for a parallel cooling configuration of the 2kW logic die and HBM modules, the maximum temperature requirements can be met for a fixed flow rate of 3.5 LPM DI water, using a tailored design of the nozzle pitch depending on the local power density. For a constant overall flow rate, the reduction of the local coolant flow rate at the location of the HBMs, in favor of the logic location, can reduce the logic temperature increase by 40%, as well as the HBM peak temperature increase by 25% due to the thermal coupling in the package. The cooler performance can be further improved by cooling the logic and HBMs in a series cooling configuration, since the coolant does not heat up significantly. This results in a higher local flow rate per component for the same system flow rate and an additional temperature reduction of 20% and 50% of the bottom and top DRAM layer of the HBM respectively, however at the expense at a higher pressure drop. Additional sensitivity studies in the package level FE model show that the HBM peak temperature and temperature gradient can be reduce by 12% and 40% respectively considering wafer-to-wafer hybrid (Cu / dielectric) bonding instead of conventional 3D bonding with µbumps.

Acknowledgment

This work was supported as part of the imec Industrial Affiliation Program on 3D System Integration and has been strongly supported by the imec partners and the imec Reliability, Electrical testing, Modeling and 3D technology teams.

References

[1] D. C. H. Yu, C. -T. Wang and H. Hsia, "Foundry Perspectives on 2.5D/3D Integration and Roadmap," *2021 IEEE International Electron Devices Meeting (IEDM)*, 2021, pp. 3.7.1-3.7.4, doi: 10.1109/IEDM19574.2021.9720568.

[2] E. Beyne, "The 3-D Interconnect Technology Landscape," in *IEEE Design & Test*, vol. 33, no. 3, pp. 8-20, June 2016, doi: 10.1109/MDAT.2016.2544837.

[3] R. Agarwal *et al.*, "3D Packaging for Heterogeneous Integration," *2022 IEEE 72nd Electronic Components and Technology Conference (ECTC)*, 2022, pp. 1103-1107, doi: 10.1109/ECTC51906.2022.00178.

[4] J. Wuu *et al.*, "3D V-Cache: the Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, 2022, pp. 428-429, doi: 10.1109/ISSCC42614.2022.9731565.

[5] R. Mahajan *et al*., "Embedded Multidie Interconnect Bridge—A Localized, High-Density Multichip Packaging Interconnect," in *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 9, no. 10, pp. 1952-1962, Oct. 2019, doi: 10.1109/TCPMT.2019.2942708.

[6] C. J. Wu *et al*., "Ultra High Power Cooling Solution for 3D-ICs," *2021 Symposium on VLSI Circuits*, 2021, pp. 1-2, doi: 10.23919/VLSICircuits52068.2021.9492489.

[7] H. Oprins and E. Beyne, "Generic thermal modeling study of the impact of 3D -interposer material and thickness options on the thermal performance and die-to-die thermal coupling,"

Fourteenth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2014, pp. 72-78, doi: 10.1109/ITHERM.2014.6892266.

[8] T.-W. Wei, H. Oprins, V. Cherman, G. Van der Plas, I. De Wolf, E. Beyne, M. Baelmans, experimental characterization and model validation of liquid jet impingement cooling using a high spatial resolution and programmable thermal test chip, Applied Thermal Engineering, Volume 152, 2019, Pages 308-318.

[9] Oprins, H., Cherman, V., Van der Plas, G., De Vos, J., and Beyne, E. (March 10, 2016). "Experimental Characterization of the Vertical and Lateral Heat Transfer in Three-Dimensional Stacked Die Packages." ASME. *J. Electron. Packag*. March 2016; 138(1): 010902.

[10] Oprins, H, Cherman, V, Van der Plas, G, De Vos, J, & Beyne, E. "Experimental Characterization of the Vertical and Lateral Heat Transfer in 3D-SIC Packages." *Proceedings of the ASME 2015 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems collocated with the ASME 2015 13th International Conference on Nanochannels, Microchannels, and Minichannels*. *Volume 1: Thermal Management*. San Francisco, California, USA. July 6–9, 2015. V001T09A068.

[11] T.-W. Wei *et al*., "Thermal Analysis of Polymer 3D Printed Jet Impingement Coolers for High Performance 2.5D Si Interposer Packages," *2019 18th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, 2019, pp. 1243-1252, doi: 10.1109/ITHERM.2019.8757402.

[12] DSM Somos Watershed data sheet: https://www.dsm.com/content/dam/dsm/additive-manufacturing/en_US/documents/somos-watershed-xc-11122-leaflet-v2.pdf

[13] T. Wei *et al*., "Experimental and Numerical Study of 3-D Printed Direct Jet Impingement Cooling for High-Power, Large Die Size Applications," in *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 11, no. 3, pp. 415-425, March 2021, doi: 10.1109/TCPMT.2020.3045113.

[14] H. Oprins *et al*., "Thermal characterization of the inter-die thermal resistance of hybrid Cu/dielectric wafer-to-wafer bonding," *2016 15th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, 2016, pp. 1333-1339, doi: 10.1109/ITHERM.2016.7517703.

[15] H. Oprins *et al*., "3D Wafer-to-Wafer Bonding Thermal Resistance Comparison: Hybrid Cu/dielectric Bonding versus Dielectric via-last Bonding," *2020 19th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, 2020, pp. 219-228, doi: 10.1109/ITherm45881.2020.9190392.