Benchmarking Study on the Thermal Management Landscape for Three-Dimensional Integrated Circuits: From Back-Side to Volumetric Heat Removal

Thomas Brunschwiler
IBM Research—Zurich,
Säumerstrasse 4,
Rüschlikon 8803, Switzerland
e-mail: tbr@zurich.ibm.com

Arvind Sridhar
IBM Research—Zurich,
Säumerstrasse 4,
Rüschlikon 8803, Switzerland
e-mail: rvi@zurich.ibm.com

Chin Lee Ong
IBM Research—Zurich,
Säumerstrasse 4,
Rüschlikon 8803, Switzerland
e-mail: ong@zurich.ibm.com

Gerd Schlottig
IBM Research—Zurich,
Säumerstrasse 4,
Rüschlikon 8803, Switzerland
e-mail: erd@zurich.ibm.com

An overview of the thermal management landscape with focus on heat dissipation from three-dimensional (3D) chip stacks is provided in this study. Evolutionary and revolutionary topologies, such as single-side, dual-side, and finally, volumetric heat removal, are benchmarked with respect to a high-performance three-tier chip stack with an aggregate power dissipation of 672 W. The thermal budget of 50 K can be maintained by three topologies, namely: (1) dual-side cooling, implemented by a thermally active interposer, (2) interlayer cooling with four-port fluid delivery and drainage at 100 kPa pressure drop, and (3) a hybrid approach combining interlayer with embedded back-side cooling. Of all the heat-removal concepts, interlayer cooling is the only approach that scales with the number of dies in the chip stack and hence enables extreme 3D integration. However, the required size of the microchannels competes with the requirement of low through-silicon-via (TSV) heights and pitches. A scaling study was performed to derive the TSV pitch that is compatible with cooling channels to dissipate 150 W/cm^2 per tier. An active integrated circuit (IC) area of 4 cm^2 was considered, which had to be implemented on the varying tier count in the stack. A cuboid form factor of 2 mm × 4 mm × 2.55 mm results from a die count of 50. The resulting microchannels of 2 mm length allow small hydraulic diameters and thus a very high TSV density of 1837 1/mm^2. The accumulated heat flux and the volumetric power dissipation are as high as 7.5 kW/cm^2 and 29 kW/cm^3, respectively.
[DOI: 10.1115/1.4032492]

Introduction

Vertical integration of ICs will be the key driver in the era of orthogonal system scaling [1,2], providing proximity and interconnectivity between components at enhanced latency and bandwidth.

The first 2.5D and 3D devices are currently available as low-volume product or engineering samples [3,4]. So far, packaging topologies with single-side electrical interconnects through the front-side and heat removal through the back-side of the chip stack have been successfully deployed, thanks to the single active layer on 2.5D [4] or the moderate power dissipation of less than 50 W of 3D memory applications [5]. The single-side topology can still be improved by reducing pitches and increasing the electromigration resistance [6] of the electrical interconnects as well as by liquid cooling and reduction of the thermal interfaces in the package [7,8]. However, the chip-stack footprint is invariant to the number of dies implemented, and hence single-side areal electrical interconnects and back-side cooling are not scalable solutions. Novel packaging platforms are required, which enable dual-side [9,10] and ultimately volumetric access for power delivery [11], signaling, and heat dissipation [12], so that 3D integration can be expanded to high-power devices with multiple tiers, such as microprocessors, caches, and accelerators.

Contributed by the Electronic and Photonic Packaging Division of ASME for publication in the Journal of Electronic Packaging. Manuscript received September 25, 2015; final manuscript received December 28, 2015; published online March 10, 2016. Assoc. Editor: Mehdi Asheghi.

In this paper, we aim to provide first an overview of the evolutionary and revolutionary heat-removal topologies reported so far in literature. The selection also includes a concept providing dual-side electrical interconnects to chip stacks enabled through thermal conductive substrates. In the second chapter, a high-performance 3D chip stack is defined, which is the basis in the subsequent chapter, to benchmark the thermal performance of the various heat-removal topologies. Finally, a scaling study considering true 3D integration by volumetric heat removal will illustrate the ultimate TSV density with integrated fluid channels in place.

Thermal Management Landscape

Back-Side Cooling Topologies. As long as possible, the industry will exploit existing cooling solutions developed for single-die packages and apply them to 3D chip-stack modules. In flip-chip packages, heat is removed through the die back-side and is absorbed in the heat-removing element, such as an air heat sink or a cold plate (CP). A copper lid typically provides mechanical protection for the chip and a defined interface to the CP (lid-attached CP, SS-1, Fig. 1(*a*)) [13]. The thermal coupling between solid elements is established through thermal interface materials (TIMs), which typically account for a significant portion of the total thermal resistance [14]. The number of TIMs can be reduced from two to one by considering a lid-less module with integrated direct-attached CP, which involves a tradeoff between mechanical robustness and thermal performance (direct-attached CP, SS-2, Fig. 1(*b*)) [7,15]. The embedding of microchannels into the back-side of the top die in the chip stack eliminates all TIMs in the

 Copyright © 2016 by ASME

Downloaded From: http://electronicpackaging.asmedigitalcollection.asme.org on 11/12/2018 Terms of Use: http://www.asme.org/about-asme/terms-of-use

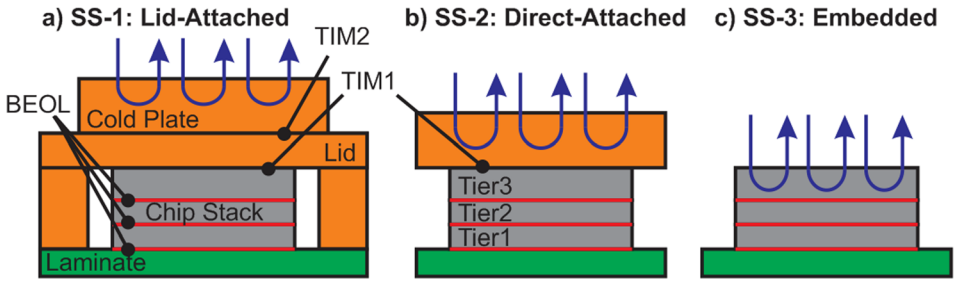

Fig. 1 Evolution of back-side cooling (single-side, SS): From (*a*) lid-attached to (*b*) direct-attached and finally (*c*) embedded convective cooling. With each generation, one TIM can be eliminated.

thermal path (embedded CP, SS-3, Fig. 1(*c*)), but requires leak-tight fluid interconnects from the system fluid loop to the silicon chip stack.

From Dual-Side to Volumetric Heat Removal. A topology change from single-side to dual-side and volumetric heat removal is a more disruptive option to reduce the thermal constraints on 3D chip stacks. Several studies [9,16,17] demonstrated approaches to enhance the thermal conductivity of organic substrates while maintaining electrical functionality. This enables the application of a second CP at the bottom side of the lidded module [18] (dual-side (DS) cooling with thermal laminate, DS-1, Fig. 2(*a*)). Such a solution still provides the mechanical robustness of the lid, but does not require the introduction of coolant into the module. Dual-side heat removal can also be established through the integration of fluid channels into a silicon interposer, while maintaining the TSV capability with sealing structures [19] (dual-side cooling with interposer, DS-2, Fig. 2(*b*)). The ultimate solution resulting in volumetric heat removal is the integration of microchannels in between the active dies in the chip stack [12,19]. This approach has to deal with the high density of TSVs at pitches below 100 μm, drastically reducing mass transport of the coolant compared with back-side CPs (volumetric (V) cooling, V-1, Fig. 2(*c*)).

For the interlayer cooling topologies with the constrained hydraulic diameters, the four-port approach was introduced to enhance the aggregate mass transport and the local coolant velocity in microchannels between dies [12] as compared to the two-port configuration (Fig. 3(*a*)). All four sides of the chip-stack area are used for feeding (e.g., north and south) and draining (e.g., east and west) coolant (Fig. 3(*b*)). The microchannels are arranged in a chevronlike pattern, with the shortest and the longest lengths in the chip-stack corners and the center, respectively. The maximal microchannel length in the four-port case is equivalent to the length of all microchannels in the two-port case for a chip stack with a square footprint. Hence, a four time larger mass flow rate is achieved for the same pressure boundary conditions and hydraulic diameter, owing to the reduction of the average channel length and increase of the channel count by a factor of two.

Pyramidlike Chip Stacks. A further challenge is the removal of power dissipated from pyramidlike chip stacks with dies of different sizes. In the case of back-side cooling, the lid or CP needs to be adapted to maintain a small TIM bondline also for the extended area of the larger dies (Fig. 4(*a*)). Because of manufacturing tolerances, a lateral gap between the heat-removing element and the subsequent smaller silicon die will remain [20]. Heat dissipated in this gap region relies on heat spreading in the thinned silicon die toward the TIM zones. Accordingly, the power density in these areas needs to be throttled. Bottom-side and dual-side heat removal, however, are compatible with pyramid chip stacks as they extract the heat through the bottom side of the module even in gap areas (Fig. 4(*a*)).

Intrastack Hot Spot Mitigation. Improvements in heat conduction in the chip-stack can further reduce thermal gradients and mitigate hot spots resulting from power maps with high heat-flux concentrations within the tiers. This can be achieved by a thermally aware placement of TSVs, thermal underfills, or the integration of near-junction spreaders, such as graphene or diamond [21] (SS-2:sidd, Fig. 4(*b*)). The main challenge of integrated spreaders, however, is compatibility with current chip-manufacturing processes and in particular TSV integration.

Dual-Side Electrical Interconnects. With advanced cooling technologies, power delivery and signaling from and to the chip stack might become the limiting factors rather than the heat

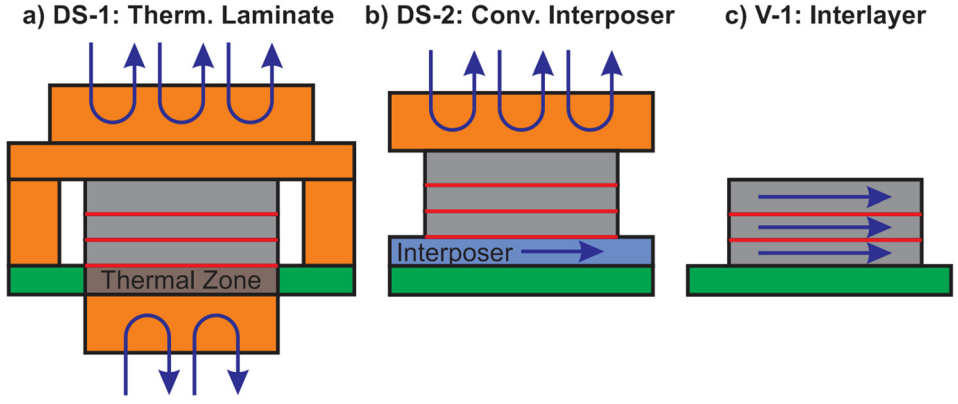

Fig. 2 Roadmap of disruptive cooling approaches, ranging from DS cooling, considering (*a*) thermal zones in laminates or (*b*) thermally active interposers, to volumetric (*V*) heat removal with fluid channels integrated between active dies (*c*)

Downloaded From: http://electronicpackaging.asmedigitalcollection.asme.org on 11/12/2018 Terms of Use: http://www.asme.org/about-asme/terms-of-use

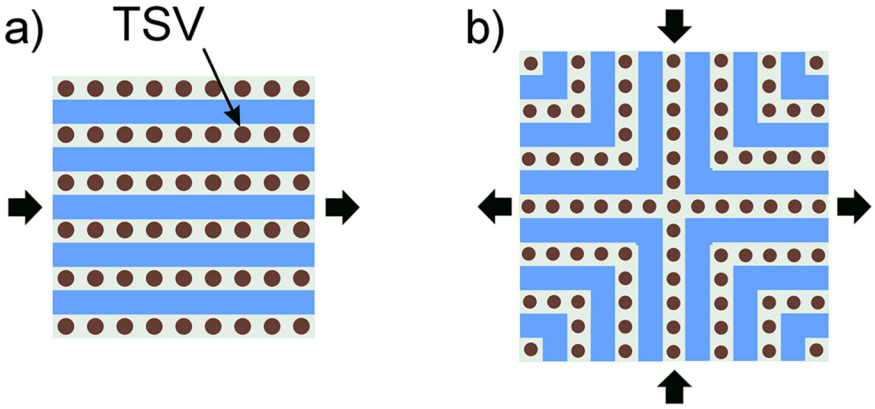

Fig. 3 Top view of (*a*) a two-port and (*b*) a four-port microchannel fluid-delivery architecture compatible with interlayer cooling

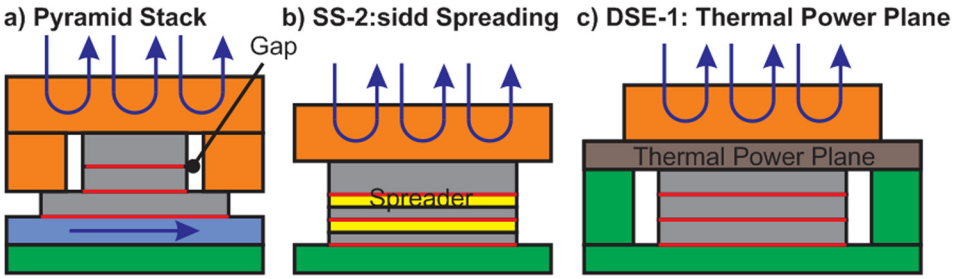

Fig. 4 Module sketches illustrating (*a*) the DS cooling of a pyramid chip stack, (*b*) the integration of intrastack spreading layers (SS-2:sidd), and (*c*) the implementation of an electrically active lid (DSE-1)

removal itself. Here, functional thermal management can provide solutions: the integration of thermally conductive laminates enables dual-side electrical interconnects as demonstrated with the thermal power plane [9] (dual-side electrical interconnects, DSE-1, Fig. 4(*c*)). Such an electrically active lid allows power to be delivered also from the chip stack back-side, so that additional electrical interconnects are available for signaling on the chip stack front-side.

High-Performance 3D Chip Stack

Chip-Stack Thermal Characteristics. The cooling topologies described above and their performance were presented in various publications considering different chip-stack configurations and heat inputs. In this paper, we intend to benchmark these topologies (Table 1) with respect to a common high-power, three-tier chip stack serving as a strawman application (Fig. 5). The bottom die, tier 1, features a 5×4 array of accelerators and the input/output ports providing the communication to and from the entire chip stack. Tier 2 contains the shared cache layer used by the accelerators of tier 1 and the 12 cores on the microprocessor die, tier 3. At a chip-stack footprint of 4 cm^2 and heat-flux levels of 300, 150, 80, and 20 W/cm^2 for the core hot spots, the core itself, the accelerators, and the other areas (including the cache), respectively, we get an aggregated power dissipation of 672 W. In general, the dies are stacked in downward-facing orientation.

Table 1 Nomenclature of the benchmark topologies considered in this study

Abbreviation	Benchmark topologies
Single-side cooling	
SS-1	Lid-attached CP
SS-2	Direct-attached CP
SS-3	Embedded convective cooling
SS-2:sidd	Direct-attached CP with diamond replacing silicon of the two thinned dies
SS-2:ddd	Direct-attached CP with diamond replacing silicon of all the three dies
Dual-side cooling	
DS-1	Dual-side cooling with thermal laminate
DS-2: 0.3	Dual-side cooling with thermal active interposer
DSE-1	Dual-side electrical interconnects
DSE-2	Dual-side cooling and electrical interconnects
Volumetric cooling	
V1: base, 0.3	Interlayer cooling with two-port fluid delivery and 0.3 bar pressure drop
V1: base, 1.0	Interlayer cooling with two-port fluid delivery and 1.0 bar pressure drop
V1: 4-port, 1.0	Interlayer cooling with four-port fluid delivery and 1.0 bar pressure drop
V1: half, 1.0	Interlayer cooling with two-port fluid delivery and 1.0 bar pressure drop considering only half of the chip-stack foot print
V1: +SS-3, 0.3	Interlayer cooling with two-port fluid delivery and embedded convective cooling, with 0.3 bar pressure drop

Downloaded From: http://electronicpackaging.asmedigitalcollection.asme.org on 11/12/2018 Terms of Use: http://www.asme.org/about-asme/terms-of-use

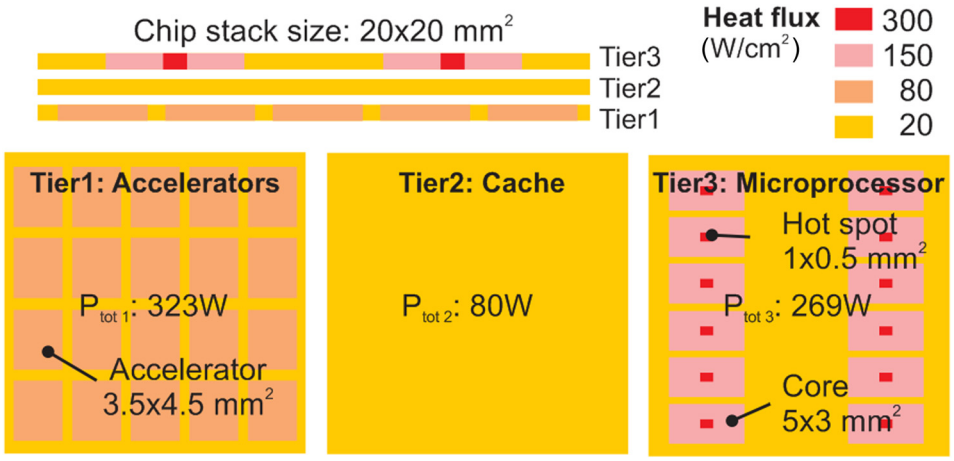

Fig. 5 Power map of the three-tier chip-stack strawman showing the functional blocks, dimensions, and heat fluxes

Table 2 Thickness and thermal properties of the layers in the chip stack and the respective cooling elements

Element	Layer	t (μm)	k (W/m K)	R_{th} (K mm^2/W)
Chip stack	Si top die	780	140	5.6
	Si other dies	50	140	0.4
	BEOL	10	2	5.0
	Micro solder balls	20	2	10.0
	Diamond spreader	780	1800	0.4
CP	Convection			10.0
	Cu base plate	500	390	1.3
Lid	Cu lid	2000	390	5.1
TIMs	1	50	4	12.5
	2	80	4	20.0
	3	80	4	20.0

Values from Refs. [14,15] and [22–25].

The state-of-the-art dimensions and thermal properties were defined for the elements in the chip stack and the cooling components of the module (Table 2). The element thermal resistance R_{th} is the result from the division of the element thickness t with its thermal conductivity k. The silicon die thickness is assumed to be 50 μm and 780 μm for the dies with TSVs and the top die without TSVs, respectively. A thermal resistance of 5 and 10 K mm^2/W for the back-end-of-line (BEOL) wiring layers and the micro-solder-ball layers are defined, as discussed by Wakil and coworkers [22,23]. For the back-side CP, a convective thermal resistance of 10 K mm^2/W is assumed for a 30 kPa pressure drop [14,15]. Adiabatic boundary conditions were applied to the chip stack bottom-side, as heat dissipation through the organic substrate can be expected to be negligible. Typical values for the TIM [24,25] are listed as well. The improved thermal conductivity of the thinned silicon slab considering integrated TSVs was not included in the model, as the thermal resistance of the silicon slab was already negligible compared to the TIM and BEOL values.

Table 3 Thermal performance of layers in the dual-side cooling approaches

Element	Layer	t (μm)	k (W/m K)	R_{th} (K mm^2/W)
Power insert	Solder rails	60	10	6.0
	Laminate with power insert	1200	143	8.4
Thermal	Solder rails	60	10	6.0

Values from Refs. [9] and [18].

Thermal Conductive Substrates. The substrates with low thermal impedance were demonstrated, such as a power insert [18] and a thermal power plane [9] connected through solder rails, enabling dual-side heat removal or dual-side electrical interconnects (Table 3). The power insert consists of copper lamellas laminated into an organic substrate in the chip-stack shadow area. It is vertically oriented to provide power and dissipate heat through the substrate. The thermal power plane is a coreless substrate with an array of stacked built-up vias, spanning the entire thickness of the substrate, again to provide power and extract heat. To improve thermal contact, both elements are mainly bonded through solder rails (elongated solder shapes) to the chip stack.

Fluid Channel Geometries. The channel geometries in the thermally active interposer or in the interlayer cooling topologies need to accommodate the required TSV dimensions (Table 4). A TSV density equivalent to the solder-ball density of the package can be expected for the interposer. In contrast, a smaller TSV pitch must be used within the chip stack [26] to facilitate the required power distribution and communication bandwidth in the chip stack.

Thermal Management Benchmarking

Compact Thermal Modeling. The temperature field of the chip stack for the different cooling options was derived by 3DICE, an open-source compact thermal-modeling framework capable of accounting for heat conduction and convection [27]. Heat conduction in solid parts and across the solid–liquid interface is modeled by an equivalent resistor network. The convective part is implemented by temperature-dependent current sources. The convective resistance of the back-side CP is applied as a boundary condition, as defined in Table 2. The single-phase heat and mass transport in the silicon interposer and the chip stack are computed with pressure boundary conditions for microchannels considering developed fluid boundary layers according to the correlations derived by Shah and London [28] with a Nusselt number, Nu, of

Table 4 Fluid channel dimensions for interposers and interlayer cooling

Channel dimensions (μm)	Interposer	Interlayer
Pitch	250	100
Height	300	100
Width	150	50

Downloaded From: http://electronicpackaging.asmedigitalcollection.asme.org on 11/12/2018 Terms of Use: http://www.asme.org/about-asme/terms-of-use

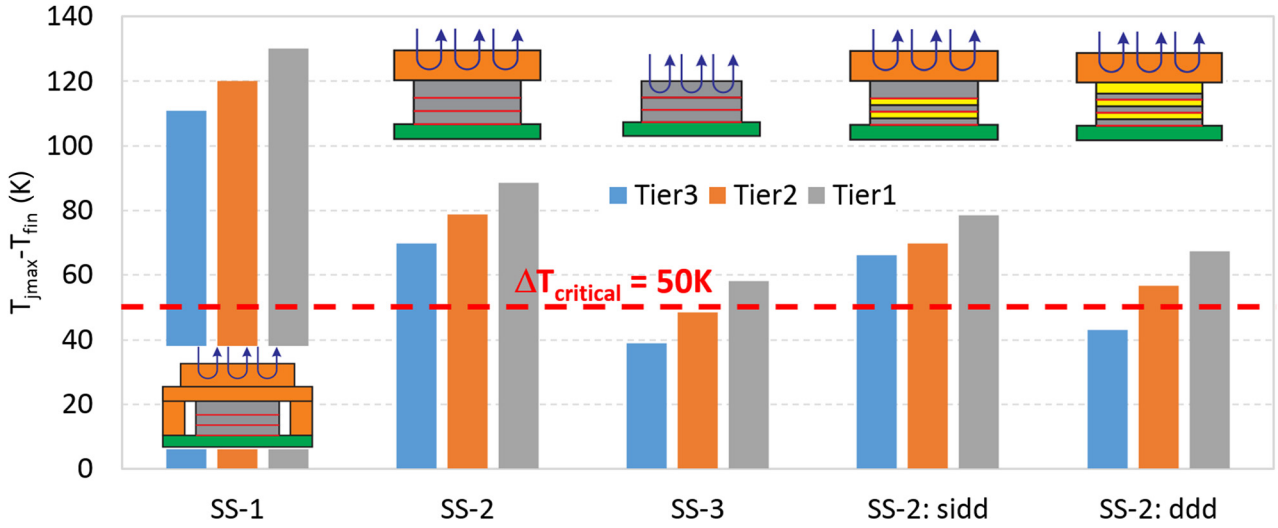

Fig. 6 Computed thermal gradients from fluid inlet to the maximal junction temperature for every tier for the back-side cooling approaches (lid-attached SS-1, direct-attached SS-2, and the embedded CP SS-3), including the replacement of silicon (gray) by diamond (yellow) in the chip stack for the thinned dies only (SS-2:sidd) and all the dies (SS-2:ddd) for the direct-attached CP case

$$\mathrm{Nu} = 8.235 \cdot (1 - 2.0421\mathrm{AR} + 3.0853\mathrm{AR}^2 - 2.4765\mathrm{AR}^3 + 1.0578\mathrm{AR}^4 - 0.1861\mathrm{AR}^5) \quad (1)$$

and the product of the friction coefficient, fr, with the Reynolds number Re

$$\mathrm{fr} \cdot \mathrm{Re} = 24 \cdot (1 - 1.3553\mathrm{AR} + 1.9467\mathrm{AR}^2 - 1.7012\mathrm{AR}^3 + 0.9564\mathrm{AR}^4 - 0.2537\mathrm{AR}^5) \quad (2)$$

considering a channel aspect ratio AR (channel height divided by channel width). This modeling approach was previously validated against experimental results with an error of less than 10% [27], proofing the validity of the approach. Water was considered as a coolant, due to its superior heat transfer properties in single-phase compared to dielectric fluids. Especially, its high sensible heat and low viscosity are of importance for the low hydraulic diameter application. However, special sealing structures as proposed by Madhour et al. [19] are required to prevent electrical shorting between TSVs. The thermal gradient from the fluid inlet temperature to the maximal junction temperature (i.e., silicon slab temperature at the interface to the BEOL layer, where transistors are located) per tier was computed. A 50 K thermal budget is considered to comply with the free-cooling standards in datacenters.

Back-Side Cooling Performance. First, the back-side cooling options, from lid-attached (two TIMs) to direct-attached (one TIM) and finally to the embedded cooling module, are assessed. The thermal gradient can be reduced drastically by eliminating individual TIMs in the package (Fig. 6, SS-1–SS-3). The thermal gradient difference between tiers in the chip stack is moderate compared with the total gradient. This results from the arrangement of the cores, with the highest power densities in the top tier (T3). However, all solutions still violate the free-cooling thermal budget of 50 K. The performance benefits of replacing the silicon body with diamond of the thinned dies (T1 and T2) are moderate because of the low heat flux contrast in those dies (Fig. 6, SS-2:sidd). Only the additional substitution of even the thick silicon with diamond in T3 will efficiently spread the heat dissipated from the cores and thus achieve a significant thermal improvement (Fig. 6, SS-2: ddd). In all the cases, the junction temperature of the bottom die is most critical, despite the moderate power dissipation levels of that tier. This is a result of the accumulating heat flux, which is dissipated toward the chip-stack back-side. This cross-talk results in a modulation of the temperatures of the bottom dies by the power map of the subsequent tiers (Fig. 7).

Dual-Side Cooling and Electrical Interconnects. The dual-side cooling approaches (Fig. 8, DS-1 and DS-2) provide a second heat-removal path and hence invert the maximal junction temperature distribution between the tiers. Still, the thermal coupling between elements and the low power dissipation of tier 2 result in a moderate thermal gradient within the chip stack. The module with the thermally active silicon interposer below tier 1 (DS-2) complies with the thermal budget of 50 K. A pressure difference of 30 kPa was considered, to be compatible with the fluid feed in current liquid-cooled servers. This number is based on the pressure available from compact and reliable pumps minus the pressure drops in other system components, such as heat exchangers, connectors, filters, and distribution piping. Practical back-side CPs are operated in the same pressure regime [23].

The implementation of the electrically active lid, called thermal power plane, results in a slightly lower thermal performance than with the passive lid (compare SS-1 with DSE-1 and DS-1 with

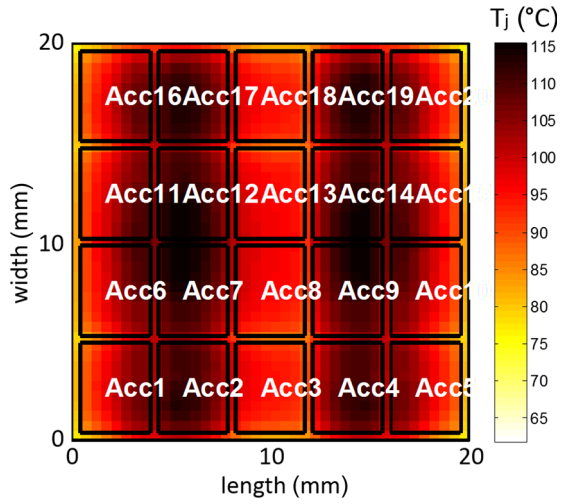

Fig. 7 Junction temperature of tier 1 for the direct-attach case (SS-2) considering a fluid inlet temperature of 27 °C

Downloaded From: http://electronicpackaging.asmedigitalcollection.asme.org on 11/12/2018 Terms of Use: http://www.asme.org/about-asme/terms-of-use

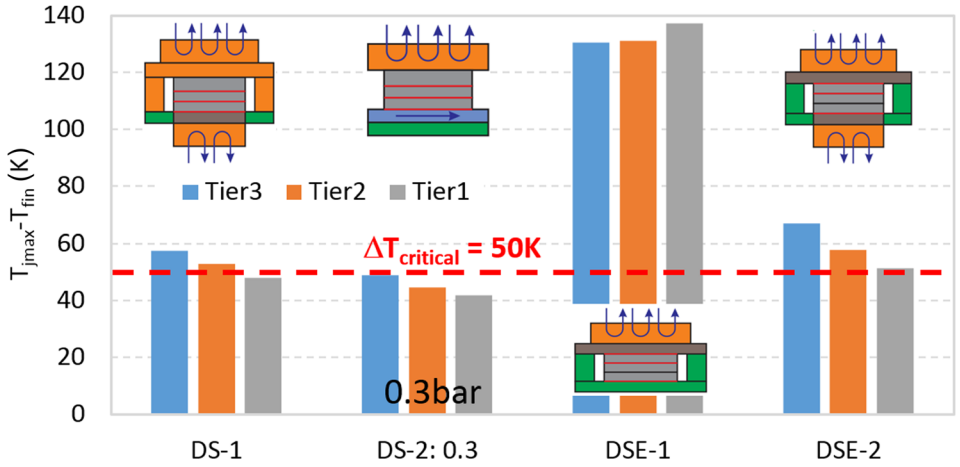

Fig. 8 Computed thermal gradients from fluid inlet to the maximal junction temperature for every tier for the dual-side cooling approaches, considering dual-side cooling with a thermal laminate (DS-1) and a thermally active silicon interposer (DS-2). The thermal performance of the dual-side electrical interconnect package with the thermal power plane in the single (DSE-1) and the dual-side (DSE-2) heat-removal configuration is also shown.

DSE-2), mainly because of the orientations of the dies in the stack. The two top dies face upward to receive power through the thermal power plane and thus minimize the number of power TSVs in the stack. Accordingly, the high heat fluxes of tier 3 need to be dissipated through the BEOL wiring layers of the same die.

Interlayer Cooling Performance. The interlayer cooling approach allows the integration of three microchannel layers within the chip stack, one per active die. The resulting thermal gradient is largely caused by the fluid temperature increase, due to the low coolant mass flow rates. These are result from the small hydraulic diameter of 66 μm for the microchannels with 100 μm height and 50 μm width. In contrast, hydraulic diameters of 150 μm and split-flows [14,15] are implemented in back-side CPs, resulting in an order-of-magnitude higher flow rates and respective thermal gradients. Hence, a pressure drop of 100 kPa instead of 30 kPa is required for interlayer cooling topologies, to increase the mass flow rate by 3.3-fold. Accordingly, the peak thermal gradient can be reduced from 259 °C to below 83 °C (Fig. 9, V1: base, 0.3 versus V1: base, 1.0). Even in the high-pressure case, the thermal response is dominated by the fluid temperature increase, see Fig. 10, where a large thermal gradient is visible in the flow direction. The increase in pressure drop compared with that of the state-of-the-art back-side CPs will affect the choice of fluid pumps. Currently, pumps with up to 220 kPa pressure differential were successfully implemented in server systems [29] and indicate the feasibility of increased pressure drops.

The benefits of the four-port fluid delivery options can be significant, especially for nonuniform power maps. In the strawman case, the coolant from each channel dissipates the heat from two cores in the two-port case, compared with a single core only in the four-port case (Fig. 11). Accordingly, the temperature response of the four-port interlayer-cooled chip stack with a pressure budget of 100 kPa is compatible with free-cooling (Fig. 9, V1: four-port, 1.0). In general, the interlayer cooling approach benefits from chip stacks with at least one short edge to minimize the microchannel length in two-port mode. The thermal performance improves by close to four times considering half the strawman chip stack in the flow direction, so that only one row of cores needs to be cooled there (Fig. 9, V1: base, 1.0 versus V1: half, 1.0).

Another option to improve the thermal response is a hybrid approach with embedded microchannels in the topmost die and interlayer cooling in the two residual dies. The thermal budget can already be observed with a pressure budget of 30 kPa for the microchannels (Fig. 9, V1: + SS-3, 0.3).

To summarize, our benchmarking study has identified three cooling options that satisfy the free-cooling thermal budget and

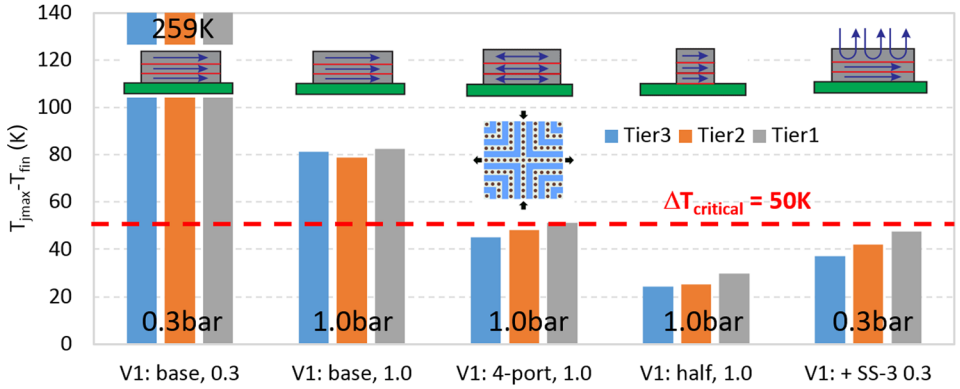

Fig. 9 Computed thermal gradients from fluid inlet to the maximal junction temperature for every tier for the interlayer cooling cases (V1). Pressure drops of 30 kPa (0.3) and 100 kPa (1.0) in two (base) and four-port (4-port) configuration (all other two-port), for large (base) and small (half) footprint stacks are considered. A hybrid solution with embedded and interlayer cooling is also presented (+SS-3).

Downloaded From: http://electronicpackaging.asmedigitalcollection.asme.org on 11/12/2018 Terms of Use: http://www.asme.org/about-asme/terms-of-use

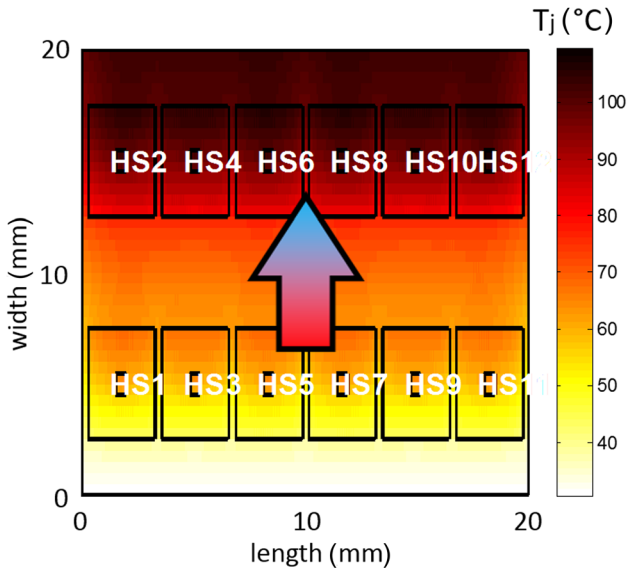

Fig. 10 Junction temperature of tier 3 for the two-port interlayer cooling case at 100 kPa (V1: base, 1.0) considering a fluid inlet temperature of 27 °C

the strawman power map: DS-2, V1: 4-port, 1.0, and V1: +SS-3, 0.3. The dual-side cooling approach with a thermally active interposer (DS-2) affects the chip-stack design itself the least, compared with the interlayer-cooled options, in four-port (V1: 4-port, 1.0) and hybrid mode (V1: +SS-3, 0.3), with the resulting constraints of a TSV pitch of 100 μm, respectively.

In general, the available cooling options need to be assessed for each individual chip-stack design, and rearrangements of tiers might be required. Back-side cooling approaches benefit from the integration of high-power dies in higher tiers, close to the CP. In dual-side cooling approaches, high-power dies will ideally be placed at the bottom and the top of the stack. Only the interlayer approach results in lower temperatures for high-power dies placed within the stack, as heat can be absorbed by the adjacent top and bottom channels. The complexity of the thermal design certainly increases and also requires an electrothermal codesign strategy to benchmark the system performance for all the thermally viable solutions.

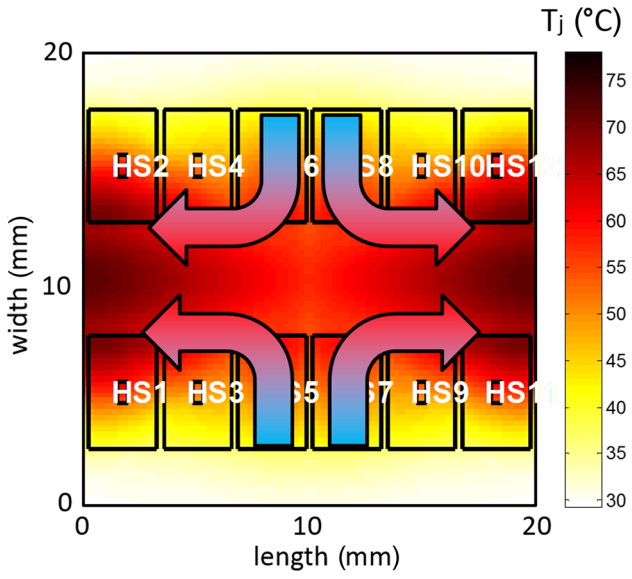

Fig. 11 Junction temperature of tier 3 for the four-port interlayer cooling case at 100 kPa (V1: four-port, 1.0) considering a fluid inlet temperature of 27 °C

Interlayer Cooling Enabling Extreme 3D

As discussed, interlayer cooling is the only cooling solution that scales with the number of dies integrated in the chip stack (Fig. 12) and hence enables extreme 3D integration. The integration of TSVs and microchannels and the resulting thermal performance were reported in several studies [12,19,30].

The biggest issue degrading the interlayer cooling performance remains the combination of high TSV densities at pitches below 100 μm at large chip-stack footprints of more than 4 cm^2, which results in long microchannels with small hydraulic diameters. Hence, the research community is trying to mitigate the pressure drop by means of various heat-transfer structures, such as pin-fins, nonuniform fluid cavities, and the implementation of high aspect ratio TSVs [12,31].

First 3D chip stacks will exploit technology and intellectual property (IP) separation, e.g., by implementing the cache on a different tier and node size than the cores. The ultimate goal is to minimize the communication distance between functional elements at a high interconnect density. Extreme 3D, i.e., the stacking of many tiers into a cubic chip stack, comes close to the spherical optimum, with minimal communication distance (Fig. 13).

Scaling Study. In this scaling study, our aim is to predict the maximal interconnect density for an interlayer-cooled stack with multiple tiers, considering a current microprocessor with a total active silicon area (A_{tot}) of 20×20 mm^2 and a uniform power dissipation of 150 W/cm^2, resulting in a total power dissipation of 600 W. A tier width of twice the tier length was chosen as a compromise to obtain a shorter channel length but still maintain proximity of elements. A maximum junction temperature rise of 60 K (compared to the former 50 K) relative to the fluid inlet temperature is allowed for a pressure drop of 100 kPa and water as coolant, considering a data center with chilled water supply. Correlations of Shah and London [28] are applied to determine the mass and heat transport analytically, considering fully developed boundary conditions.

The thickness of the silicon die (tSi + ch) and the channel width (cw), height (ch), and pitch (cp) depend primarily on the TSV manufacturing capabilities. Thus, in this scaling study, the dimensions of the microchannels were defined in ratios relative to the TSV diameter (dTSV) (Table 5, third column). We assumed a TSV aspect ratio (hTSV to dTSV) of 6:1. The TSV pitch in streamwise direction (pTSVx) is half of that in the transversal (pTSVy) direction to allow both high interconnect densities and hydraulic diameters (Fig. 14). Only the thickness of the BEOL wire layers was defined in absolute dimensions, i.e., 2 μm. The baseline case dimensions are listed in Table 5, last column, to give an idea of the geometries considered. The minimal cp

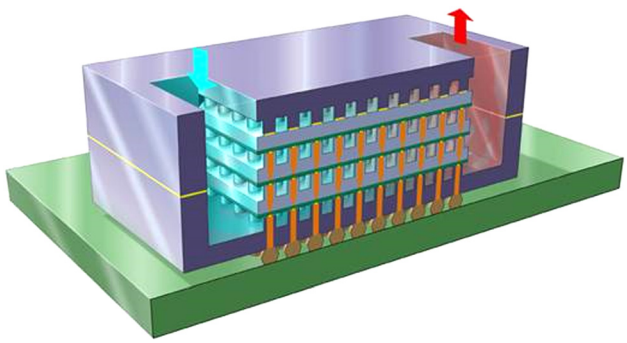

Fig. 12 Three-dimensional view of an interlayer-cooled chip stack with embedded TSVs in the microchannel walls

Downloaded From: http://electronicpackaging.asmedigitalcollection.asme.org on 11/12/2018 Terms of Use: http://www.asme.org/about-asme/terms-of-use

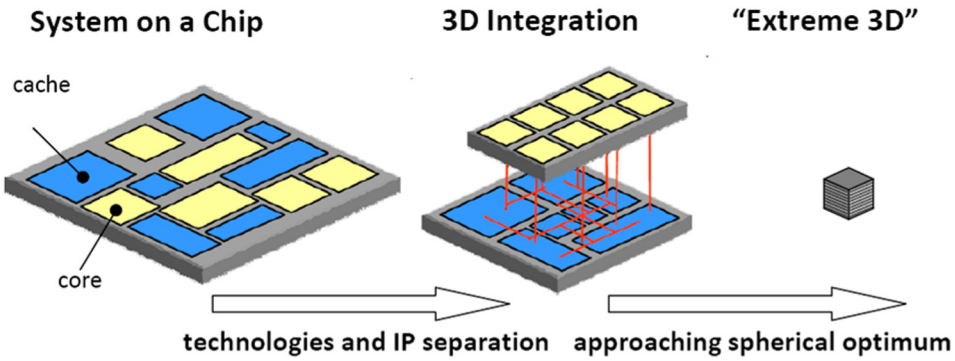

Fig. 13 Three-dimensional roadmap: from 3D stacking with technology and IP separation to extreme 3D with minimal distance between functional elements

resulting in the maximally allowed junction temperature rise at the channel outlet for a stack with n dies was computed; all other dimensions result from the ratios defined in Table 4.

Results. To achieve the active silicon area (A_{tot}) required, the lateral dimensions of the chip stack are reduced as the number of tiers increases (Fig. 15(*a*)—dark and light blue curves). The hydraulic diameter and thus the pitch of the microchannels can be shrunk at smaller chip dimensions and channel lengths, see Fig. 15(*b*)—dark blue curve. At the aspect ratio given (Table 5), the reduction in channel pitch is accompanied by an increase in the TSV density (Fig. 15(*b*)—dashed red line), but a reduction in channel height and, accordingly, chip-stack volume (Fig. 15(*a*)—dashed red line).

The chip stack is close to cubic, with a volume of 20.4 mm^3 at a die count of 50 (die thickness 51 μm), resulting in a channel length of 2 mm and a 3D chip-stack width and height of 4 mm and 2.55 mm, respectively. The TSV density is 1837 TSVs/mm^2, which is equivalent to a uniform 23-μm TSV pitch. The total mass flow through all the cavities of the chip stack is lower at shorter channel lengths because of the simultaneous reduction in hydraulic diameter (Fig. 15(*b*)—green curve).

The formation of the close to cubic stack with 50 tiers results in a maximal wire length of 8.55 mm, considering the Manhattan distance from the top right to the bottom left corner of the stack, compared with 40 mm in the case of a single die. The cuboid form factor of 2 mm enables the integration of a high TSV count, as required from an electrical perspective, and still allows a uniform power dissipation of 150 W/cm^2 per tier and a total of 600 W per stack. This corresponds to a volumetric power density of 29 kW/cm^3 and an aggregate power density of 7.5 kW/cm^2, representing the extreme 3D character. Despite these benefits of extreme 3D, the technological challenges, in particular the yield for building a chip stack with tens of tiers, are enormous. Accordingly, yield resilient IC-designs are required. However, only the future will tell whether such form factors will be economically viable and enable a new class of systems with ultimate densities and performance.

Table 5 Definition of dimensional ratios considered in the scaling study

Dimension	Variable	Dimension relative to TSV diameter	Baseline case (μm)
TSV diameter	dTSV	1	25
TSV pitch streamwise	pTSVx	2	50
TSV pitch transversal	pTSVy	4	100
TSV height	hTSV	6	150
Channel pitch	cp	4	100
Channel height	ch	4	100
Channel width	cw	2	50
Fin width	fw	2	50
Silicon slab thickness	tsi	2	50

A baseline case is also provided.

Conclusion

A cooling topology roadmap for 3D chip stacks was proposed. We envision a potential progression to start from the established lidded back-side approaches to direct-attached CPs and finally the embedding of microchannels into the chip-stack back-side. On this evolutionary track, the number of TIMs can be reduced and completely mitigated. Dual-side and volumetric cooling are revolutionary approaches using thermally conductive substrates, thermally active interposers, or microchannels within active tiers. The characteristics listed in Table 6 summarize the value proposition of the various topologies.

In the benchmark study using the three-tier strawman chip stack, the dual-side cooling approach with thermally active interposer, interlayer cooling in four-port, and as hybrid solution with embedded back-side microchannels resulted in thermal gradients below 50 K, i.e., sufficiently low for free cooling.

The scaling study of the interlayer-cooling approach shows the benefits of chip stacks with cuboid form factor, with short channel lengths, but multiple tiers, resulting in extreme 3D integration. The TSV density can be scaled up to 1837 1/mm^2 considering an active IC area of 4 cm^2, distributed over 50 tiers, with a final form factor of 2 mm × 4 mm × 2.55 mm. On each layer, 150 W/cm^2 can

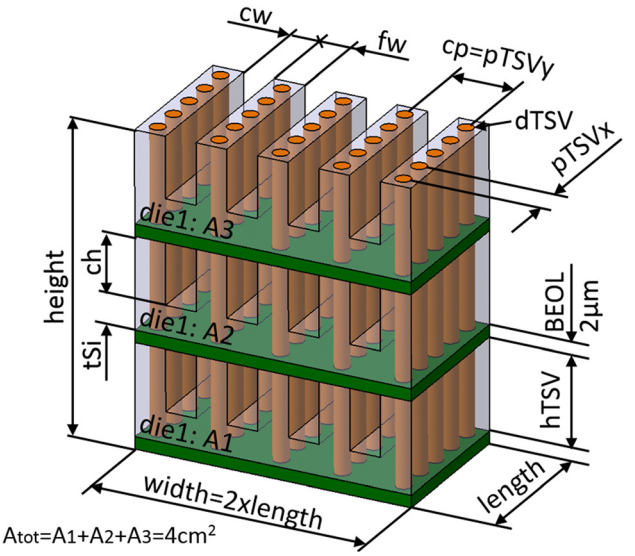

Fig. 14 Sketch of chip stack depicting the parameters considered in the scaling study (green: BEOL layer, gray: silicon, and orange: TSV)

Downloaded From: http://electronicpackaging.asmedigitalcollection.asme.org on 11/12/2018 Terms of Use: http://www.asme.org/about-asme/terms-of-use

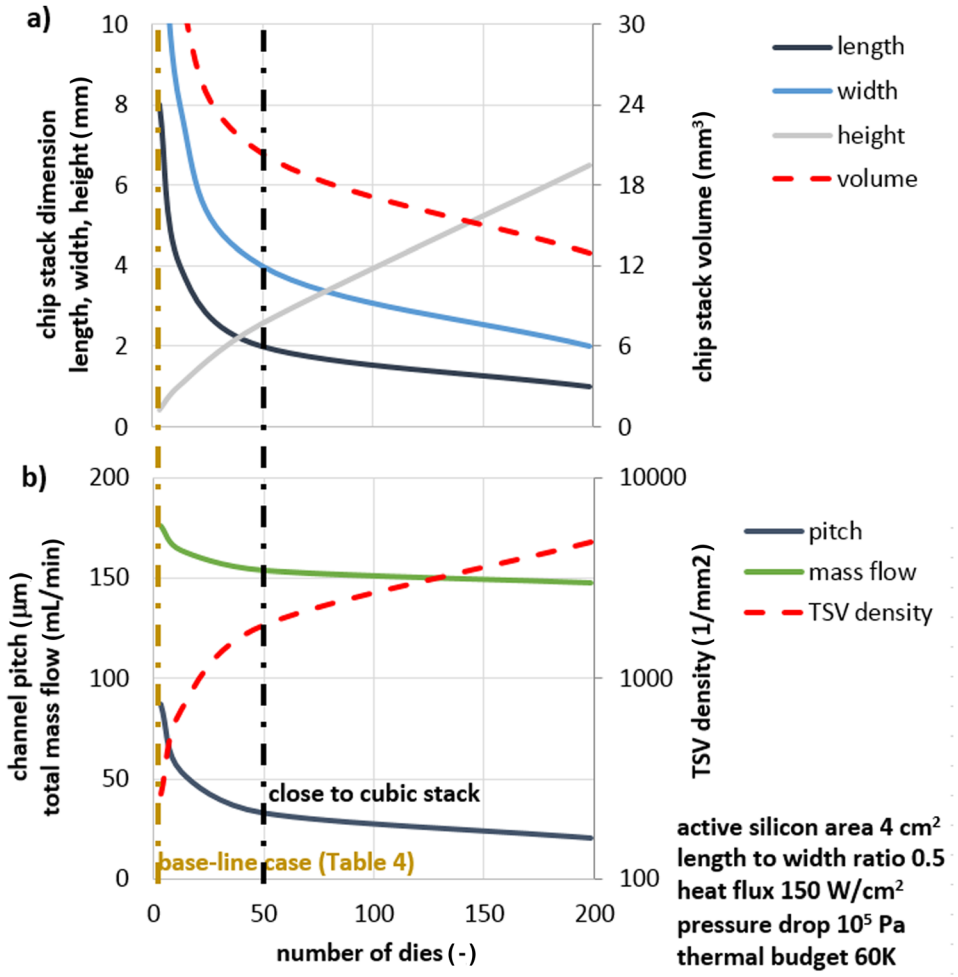

Fig. 15 Results from the scaling study: (*a*) chip-stack dimensions and (*b*) resulting channel pitch, total mass flow, and TSV density versus number of dies in the stack at constant active silicon area

Table 6 Value proposition of various cooling topologies

Topology	Pros	Cons
Single-side cooling	Compatible with current cooling solutions, ease of implementation	Poor performance, requires high-power die to be on top tier
Lid-attached	Defined interface, mechanically protected die	Two thermal interfaces
Direct-attached	Only one TIM	Controlling formation of TIM1 is delicate
Embedded	Best thermal performance	Fluidic sealing to chip stack required
Dual-side cooling	High-performance die on top and bottom side of chip stack; supports the cooling of pyramid chip stack	Higher complexity, additional cost
Thermal laminate	Dual-side cooling, without the need of fluidic sealing to silicon	Compromise on electrical performance of thermal laminate
Convective interposer	Dual-side cooling, no electrical penalty in laminate	Fluidic sealing to interposer required
Volumetric cooling Interlayer cooling	Scalable with the number of tiers	Large pumping pressure needed for large dies, limitations on TSV density

be dissipated, resulting in an accumulated heat flux of 7.5 kW/cm^2 (projected on the footprint) and a volumetric power dissipation of 29 kW/cm^3.

For such chip stacks, power delivery will become the main limitation. One could consider four faces of the cube to be populated with electrical interconnects for power delivery, with an area fill factor of 50%, still leaving two faces for fluid delivery and drainage and some area for signaling. A current density of 66 A/mm^2 would result in the power interconnects that exceed the electromigration limit of solder joints. Accordingly, scalable power delivery concepts are required, such as on-chip voltage conversion or electrochemical power delivery as described by Andersen et al. [32] and Sridhar et al. [33].

Acknowledgment

We acknowledge Bruno Michel and Walter Riess for their continuous support of this research activity. Part of this work was

Downloaded From: http://electronicpackaging.asmedigitalcollection.asme.org on 11/12/2018 Terms of Use: http://www.asme.org/about-asme/terms-of-use

carried out within the European CarrICool Project under the Seventh Framework Program for Research and Technological Development (FP7-ICT-619488).

Nomenclature

AR = microchannel aspect ratio (height divided by width)
ch = channel height (μm)
cp = channel pitch (μm)
cw = channel width (μm)
dTSV = TSV diameter (μm)
fr = friction factor
fw = fin width (μm)
hTSV = TSV height (μm)
k = thermal conductivity (W/m K)
Nu = Nusselt number
pTSVx = TSV pitch streamwise (μm)
pTSVy = TSV pitch transversal (μm)
R_{th} = thermal resistance (K mm^2/W)
Re = Reynolds number
t = thickness (μm)
T_{fin} = fluid inlet temperature (K)
T_{jmax} = maximal junction temperature (K)
tsi = silicon slab thickness (μm)
$\Delta T_{critical}$ = thermal budget (K)

References

[1] Ruch, P., Brunschwiler, T., Escher, W., Paredes, S., and Michel, B., 2011, "Toward Five-Dimensional Scaling: How Density Improves Efficiency in Future Computers," IBM J. Res. Dev., **55**(5), pp. 1–13.

[2] Iyer, S. S., 2012, "The Evolution of Dense Embedded Memory in High Performance Logic Technologies," IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, Dec. 10–13, pp. 33.1.1–33.1.4.

[3] Jeddeloh, J., and Keeth, B., 2012, "Hybrid Memory Cube New DRAM Architecture Increases Density and Performance," Symposium on VLSI Technology (VLSIT), Honolulu, HI, June 12–14, pp. 87–88.

[4] Erdmann, C., Lowney, D., Lynam, A., Keady, A., McGrath, J., Cullen, E., Breathnach, D., Keane, D., Lynch, P., De La Torre, M., De La Torre, R., Peng, L., Collins, A., Farley, B., and Madden, L., 2015, "A Heterogeneous 3D-IC Consisting of Two 28 nm FPGA Die and 32 Reconfigurable High-Performance Data Converters," Solid-State Circuits, **50**(1), pp. 258–269.

[5] Khurshid, M. J., and Lipasti, M., 2013, "Data Compression for Thermal Mitigation in the Hybrid Memory Cube," IEEE 31st International Conference on Computer Design (ICCD), Asheville, NC, Oct. 6–9, pp. 185–192.

[6] Minhua, L., Da-Yuan, S., Lauro, P., Sung, K., Goldsmith, C., and Sun-Kyoung, S., 2009, "The Effects of Ag, Cu Compositions and Zn Doping on the Electromigration Performance of Pb-Free Solders," 59th Electronic Components and Technology Conference (ECTC 2009), San Diego, CA, May 26–29, pp. 922–929.

[7] Schultz, M., Gaynes, M., Parida, P., and Chainer, T., 2014, "Experimental Investigation of Direct Attach Microprocessors in a Liquid-Cooled Chiller-Less Data Center," IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITHERM), Orlando, FL, May 27–30, pp. 729–735.

[8] Tuckerman, D. B., and Pease, R., 1981, "High-Performance Heat Sinking for VLSI," IEEE Electron Device Lett., **2**(5), pp. 126–129.

[9] Brunschwiler, T., Heller, R., Schlottig, G., Tick, T., Harrer, H., Barowski, H., Niggemeier, T., Supper, J., and Oggioni, S., 2014, "Thermal Power Plane Enabling Dual-Side Electrical Interconnects for High-Performance Chip Stacks: Concept," Electronics System-Integration Technology Conference (ESTC), Helsinki, Finland, Sept. 16–18, pp. 1–6.

[10] Marcinkowski, J., 2014, "Dual-Sided Cooling of Power Semiconductor Modules," PCIM Europe 2014: International Exhibition and Conference for Power Electronics, Intelligent Motion, Renewable Energy and Energy Management, Nuremberg, Germany, May 20–22, pp. 1–7.

[11] Sridhar, A., Sabry, M., Ruch, P., Atienza, D., and Michel, B., 2014, "PowerCool: Simulation of Integrated Microfluidic Power Generation in Bright Silicon MPSoCs," IEEE/ACM International Conference on Computer-Aided Design (ICCAD), San Jose, CA, Nov. 2–6, pp. 527–534.

[12] Brunschwiler, T., Paredes, S., Drechsler, U., Michel, B., Cesar, W., Toral, G., Temiz, Y., and Leblebici, Y., 2009, "Validation of the Porous-Medium Approach to Model Interlayer-Cooled 3D-Chip Stacks," IEEE International Conference on 3D System Integration (3DIC 2009), San Francisco, CA, Sept. 28–30, pp. 1–10.

[13] Xiaojin, W., Marston, K., and Sikka, K., 2008, "Thermal Modeling for Warpage Effects in Organic Packages," 11th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITHERM 2008), Orlando, FL, May 28–31, pp. 310–314.

[14] Colgan, E., Furman, B., Gaynes, M., Graham, W. S., LaBianca, N. C., Magerlein, J. H., Polastre, R. J., Rothwell, Beth, M., Bezama, R. J., Choudhary, R., Marston, K. C., Toy, H., Wakil, J., Zitz, J. A., and Schmidt, R. R., 2007, "A Practical Implementation of Silicon Microchannel Coolers for High Power Chips," IEEE Trans. Compon. Packag. Technol., **30**(2), pp. 218–225.

[15] Schlottig, G., De Fazio, M., Escher, W., Granatieri, P., Khanna, V. D., and Brunschwiler, T., 2015, "Lid-Integral Cold Plate Topology Integration, Performance, and Reliability," ASME Paper No. IPACK2015-48427.

[16] Schacht, R., Wunderle, B., May, D., Abo Ras, M., Faust, W., Michel, B., and Reichl, H., 2008, "Effective Thermal Modelling Evaluation and Non-Destructive Tests for Thermal Via-Structures in Organic Multi Layer PCBs," 2nd Electronics System-Integration Technology Conference (ESTC 2008), Greenwich, UK, Sept. 1–4, pp. 999–1008.

[17] Matsumoto, K., Ibaraki, S., Sueoka, K., Sakuma, K., Kikuchi, H., Orii, Y., Yamada, F., Fujihara, K., Takamatsu, J., and Kondo, K., 2013, "Thermal Design Guidelines for a Three-Dimensional (3D) Chip Stack, Including Cooling Solutions," 29th Annual IEEE Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), San Jose, CA, Mar. 17–21, pp. 1–6.

[18] Gschwend, D., Tick, T., Oggioni, S., Paredes, S., Matsumoto, K., Tiwari, M., Poulikakos, D., and Brunschwiler, T., 2014, "Laminate With Thermal-Power Insert for Efficient Front-Side Heat Removal and Power Delivery," 8th International Conference on Integrated Power Systems (CIPS), Nuremberg, Germany, Feb. 25–27, pp. 1–6.

[19] Madhour, Y., Zervas, M., Schlottig, G., Brunschwiler, T., Leblebici, Y., Thome, R., and Michel, B., 2013, "Integration of Intra Chip Stack Fluidic Cooling Using Thin-Layer Solder Bonding," IEEE International 3D Systems Integration Conference (3DIC), San Francisco, CA, Oct. 2–4, pp. 1–8.

[20] Sikka, K., Wakil, J., Toy, H., and Hsichang, L., 2012, "An Efficient Lid Design for Cooling Stacked Flip-Chip 3D Packages," 13th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITHERM), San Diego, CA, May 30-June 1, pp. 606–611.

[21] McDonald, J. F., Erdogan, O., Jacobs, P., Belemjian, P., Gutin, A., Zia, A., Chu, M., Kim, J.-W., Clarke, R., DeSimone, N., Liu, S., and Kraft, R. P., 2009, "Thermal Analysis for a SiGe HBT 40 Watt 32 GHz Clock 3D Memory Processor Chip Stack Using Diamond Heat Spreader Layers," IEEE International Conference on 3D System Integration (3DIC 2009), San Francisco, CA, Sept. 28–30, pp. 1–7.

[22] Colgan, E., Andry, P., Dang, B., Magerlein, J., Maria, J., Polastre, R., and Wakil, J., 2012, "Measurement of Microbump Thermal Resistance in 3D Chip Stacks," 2012 28th Annual IEEE Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), San Jose, CA, Mar. 18–22, pp. 1–7.

[23] Colgan, E., Polastre, R., Knickerbocker, J., Wakil, J., Gambino, J., and Tallman, K., 2013, "Measurement of Back End of Line Thermal Resistance for 3D Chip Stacks," 29th Annual IEEE Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), San Jose, CA, Mar. 18–22, pp. 23–28.

[24] Sikka, K., Toy, H., Edwards, D., Iruvanti, S., Ingalls, E., and DeHaven, P., 2002, "Gap-Reduced Thermal Paste Package Design for Cooling Single Flip-Chip Electronic Modules," Eighth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems, (ITHERM 2002), San Diego, CA, May 29–June 1, pp. 651–657.

[25] Larson, L., Yin, T., Durfee, L., Hale, L., Plante, C., Iruvanti, D., Wagner, S., Davis, R., Longworth, T., Lavoie, H., Langois, A., and Tables, R., 2014, "Engineered Thermal Interface Material," IEEE 64th Electronic Components and Technology Conference (ECTC), Orlando, FL, May 27–30, pp. 236–241.

[26] Bernstein, K., Andry, P., Cann, J., Emma, P., Greenberg, D., Haensch, W., Ignatowski, M., Koester, S., Magerlein, J., Puri, R., and Young, A., 2007, "Interconnects in the Third Dimension: Design Challenges for 3D ICs," 44th ACM/IEEE Design Automation Conference (DAC'07), San Diego, CA, June 4–8, pp. 562–567.

[27] Sridhar, A., Vincenzi, A., Atienza, D., and Brunschwiler, T., 2014, "3D-ICE: A Compact Thermal Model for Early-Stage Design of Liquid-Cooled ICs," IEEE Trans. Comput., **63**(10), pp. 2576–2589.

[28] Shah, R., and London, A., 1978, *Laminar Flow Forced Convection in Ducts: A Source Book for Compact Heat Exchanger Analytical Data*, Academic Press, New York.

[29] Ellsworth, M., 2014, "Flow Network Analysis of the IBM Power 775 Supercomputer Water Cooling System," IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), Orlando, FL, May 27–30, pp. 715–722.

[30] Yue, Z., Dembla, A., Joshi, Y., and Bakir, M., 2012, "3D Stacked Microfluidic Cooling for High-Performance 3D ICs," IEEE 62nd Electronic Components and Technology Conference (ECTC), San Diego, CA, May 29–June 1, pp. 1644–1650.

[31] Zheng, L., Zhang, Y., Huang, G., and Bakir, M., 2014, "Novel Electrical and Fluidic Microbumps for Silicon Interposer and 3-D ICs," IEEE Trans. Compon., Packag. Manuf. Technol., **4**(5), pp. 777–785.

[32] Andersen, T., Krismer, F., Kolar, J., Toifl, T., Menolfi, C., Kull, L., Morf, T., Kossel, M., Brandli, M., Buchmann, P., and Francese, P., 2014, "4.7 A Sub-NS Response On-Chip Switched-Capacitor DC-DC Voltage Regulator Delivering 3.7W/mm^2 at 90% Efficiency Using Deep-Trench Capacitors in 32nm SOI CMOS," Solid-State Circuits Conference Digest of Technical Papers (ISSCC), San Francisco, CA, Feb. 9–13, pp. 90–91.

[33] Sridhar, A., Sabry, M., Ruch, P., Atienza, D., and Michel, B., 2014, "PowerCool: Simulation of Integrated Microfluidic Power Generation in Bright Silicon MPSoCs," Computer-Aided Design (ICCAD), San Jose, CA, Nov. 2–6, pp. 527–534.

Downloaded From: http://electronicpackaging.asmedigitalcollection.asme.org on 11/12/2018 Terms of Use: http://www.asme.org/about-asme/terms-of-use