Thermal Balancing of Liquid-Cooled 3D-MPSoCs Using Channel Modulation

Mohamed M. Sabry, Arvind Sridhar, and David Atienza
Embedded Systems Lab (ESL), Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland.
emails: {mohamed.sabry, arvind.sridhar, david.atienza}@epfl.ch.

***Abstract*—While possessing the potential to replace conventional air-cooled heat sinks, inter-tier microchannel liquid cooling of 3D ICs also creates the problem of increased thermal gradients from the fluid inlet to outlet ports [1, 2]. These cooling-induced thermal gradients can be high enough to create undesirable stress in the ICs, undermining the structural reliability and lifetimes. In this paper, we present a novel design-time solution for the thermal gradient problem in liquid-cooled 3D Multi-Processor System-on-Chip (MPSoC) architectures. The proposed method is based on channel width modulation and provides the designers with an additional dimension in the design-space exploration. We formulate the channel width modulation as an optimal control design problem to minimize the temperature gradients in the 3D IC while meeting the design constraints. The proposed thermal balancing technique uses an analytical model for forced convective heat transfer in microchannels, and has been applied to a two tier 3D-MPSoC. The results show that the proposed approach can reduce thermal gradients by up to 31% when applied to realistic 3D-MPSoC architectures, while maintaining pressure drops in the microchannels well below their safe limits of operation.**

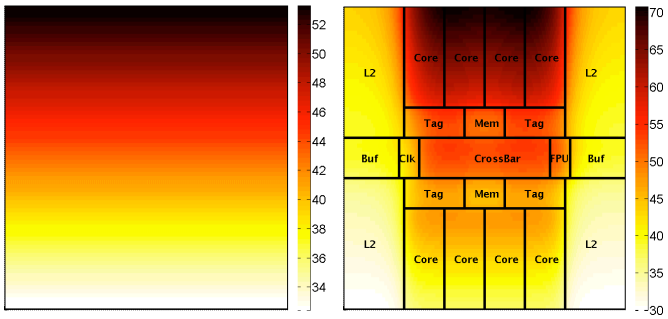
I. Introduction

Inter-tier liquid cooling is a recently proposed and a promising thermal packaging solution to counter the aggravated thermal issues arising from vertical stacking in 3D-multiprocessor ICs [3]. This new technology applies forced convective cooling in heat transfer geometries such as microchannels and pin fins, which are etched directly on the back of the 3D stacked silicon dies [3]. Thus, thermal resistances are reduced considerably, enabling the 3D ICs to operate at much lower temperatures than those with conventional heat sinks [4–6].

Inter-tier liquid cooling has given rise to a lot of benefits from a system-level perspective. Significant temperature reductions have been reported as an immediate consequence of using liquid cooling in 2D and 3D Multiprocessor Systems-on-Chip (MPSoCs) [3–6]. Moreover, using novel techniques such as run-time coolant flow rate variation enables high-precision control of the cooling effort based on the variable electronic-switching activity and time-dependent heat dissipation of the IC to achieve optimal energy consumption [4, 5].

However, inter-tier liquid cooling also brings with it new design-time and run-time challenges for the designers. For instance, the construction of the liquid-cooled heat transfer geometries on the silicon substrates limits the area for the allocation of through-silicon vias (TSVs)- the backbone of 3D integration of ICs- to the walls of the microchannels [5].

This research is partially funded by the Nano-Tera RTD project CMOSAIC (ref.123618), financed by the Swiss Confederation and scientifically evaluated by SNSF, and the PRO3D EU FP7-ICT-248776 project.



(a) Uniform heat flux (b) UltraSPARC T1

Fig. 1. Steady-state temperature distribution of a 14mm x 15mm two-die 3D IC with (a) uniform (combined) heat flux density of $50\mathrm{W/cm^2}$ and (b) the UltraSPARC T1 (Niagara-1) chip architecture [7]- the (combined) heat flux densities range from $8\mathrm{W/cm^2}$ to about $64\mathrm{W/cm^2}$. Direction of the coolant flow is from the bottom to the top of the figure.

Another serious challenge that single-phase liquid cooling brings is the increased thermal gradient. The sensible heat absorption that occurs as the coolant flows along the microchannels raises its temperature [8]. This results in an increase of coolant temperature from inlet to the outlet, which in turn, results in a thermal gradient on the IC surface even when the heat dissipation is uniform, as shown in Fig. 1 (a). More commonly, in ICs with non uniform heat dissipations, such as MPSoCs, the existing thermal gradients and hotspots are aggravated by this characteristic of liquid-cooled heat sinks, as shown in Fig. 1(b). As a result, thermal gradients proliferate in 3D-MPSoCs with liquid cooling [5]. These gradients cause uneven thermally-induced stresses on different parts of the IC, significantly undermining overall system reliability [9].

In this paper, we propose a novel design-time thermal balancing technique by modulating the microchannel width from inlet to outlet, without adding to the existing fabrication costs. This technique, referred to as *channel modulation* in the rest of this paper, relies on the well-known observation that the thermal resistance of microchannel heat sinks reduces with increasing aspect ratio of the channel cross-section [10]. In other words, keeping the height of the channels constant, while reducing the channel width increases the cooling efficiency. However, reduced channel widths imply larger resistance to the flow and hence, greater pumping effort increasing the cooling energy budget. In this work, we use optimal control based

techniques to judiciously modulate the channel width from inlet to outlet, cooling specific regions of the 3D IC more efficiently than the others, in order to minimize the overall thermal gradients in the 3D IC while meeting the design constraints. We apply this technique on test structures representing both random high-heat flux distributions and realistic 3D-MPSoC architectures. The results show that compared to constant channel width design, our proposed optimal channel width modulation minimizes the thermal gradient by 31% in 3D-MPSoCs at peak heat dissipation and 21% at average heat dissipation levels during their life-times.

The rest of this paper starts with an overview of the previous work on 3D-MPSoCs with inter-tier liquid cooling, and the various techniques proposed in the literature to overcome the thermal gradient problem in Section II. We briefly summarize the thermal modeling approach we follow in our proposal in Section III. In Section IV, we elaborate on the proposed optimal control design methodology for channel modulation, and the corresponding experimental results are reported in Section V. Finally, the conclusions are presented in Section VI.

II. Related Work

A considerable amount of research effort has gone into the study of effectiveness of inter-tier liquid cooling for high power density 3D ICs in the last decade and the related modeling methodologies. Most recent studies report that with a chip size of $1\mathrm{cm}^2$ and maximal difference of $60^{\circ}\mathrm{C}$ between the silicon junction and coolant temperatures, the heat-removal performance of single-phase cooling is more than $200\mathrm{W}/\mathrm{cm}^2$ for TSV pitches larger than $50\mu\mathrm{m}$. For further details, we refer the reader to [3, 8, 10–12].

Despite the obvious benefits of liquid cooling of 3D-MPSoCs, the problems of thermal gradients and the need for hotspot minimization and thermal balancing are also gaining attention in the recent years. Qian et al. [6] proposed a channel clustering methodology for this problem. In their method, microchannels are grouped into clusters of channels. Within a single cluster, the channels are injected with the same flow rate using micro-pumps, in order to customize the cooling effort to the demands of computing elements. This approach has not yet been shown to be feasible for large-scale production, in terms of the design complexity of the packaging and fluid networks.

Shi et al. [13] proposed a customized channel allocation technique, where the density of the etched microchannels reflects the cooling demands of various regions of the IC. This work primarily targets the improvement of energy efficiency and not thermal balance. Moreover, while the above two methods could work when the hotspots are lined up perpendicular to the coolant flow, the authors have not considered the case where multiple hotspots lie along a channel's pathway, compounding the thermal load on a single channel.

The authors in [12] used their thermal model to explore floorplanning solutions to homogenize temperature distributions in a 3D IC. But they consider a large number of identically sized functional blocks for this purpose, which is not the case in realistic designs. In addition, for this method to work, absolute flexibility in the placement of functional blocks in a complex IC architecture is required, which is seldom available given the constraints on the electrical performance of the system.

Recently, Sabry et al. [14] and Brunschwiler et al. [2, 15] investigated channel width modulation, four-port fluid access and the use of fluid guiding structure for hotspot optimization. Using some preliminary results, they have shown that it is possible to find a design-time solution for thermal balancing of 3D ICs using customized cooling. While this makes these approaches quite promising, none of them have been mathematically formulated to accurately predict their effect on the temperature distributions and to find the most optimal design solution.

This work focusses on the idea of channel modulation. To the best of our knowledge, our proposed work is the first that handles this approach in a systematic way and provides an optimal solution for thermal balancing and hotspot minimization. It is important to mention that our target is not to minimize the peak temperature, but to minimize the thermal gradients in the 3D-MPSoC. By minimizing thermal gradients, we are implicitly minimizing the peak temperature. This work contributes to providing an additional dimension of design-space exploration, in the form of channel modulation, to IC designers for the purpose of thermal balancing. Specifically, the main contributions of this work are:

1) We develop an analytical representation of heat transfer in a liquid-cooled 3D IC, to accurately predict the temperatures and heat flow in the 3D IC as a function of distance from inlet to outlet of the microchannels.
2) We use this model to formulate the problem of minimizing the thermal gradients as an optimal control design problem by selecting the channel widths (written as a function of the distance from inlet) as the control variable.

III. Thermal Model

For the application of the optimal control techniques in order to modulate the channel widths for the minimization of thermal gradients in 3D ICs, it is first essential to find analytical formulation for the heat transfer problem in 3D ICs with microchannel liquid cooling. This analytical formulation must be in the form of an ordinary differential equation (ODE) providing the mathematical platform on which an optimal control algorithm can work. The goal of our optimization is to find a sequence of channel widths, as a function of the distance from the inlet, which minimizes the intended cost function: the temperature gradient. Hence, the **steady-state temperatures** of the 3D IC must be written as a function of this distance in the analytical formulation, with the channel widths as an input parameter. In other words, if the distance from the inlet is measured along the coordinate axis z, then we need to find an equation of the form:

$$\frac{d}{dz}\mathbf{T}(z) = \Phi(z, \mathbf{w_C}(z), \mathbf{T}(z)), \tag{1}$$

where $\mathbf{T}(z)$ is the steady-state temperatures vector on the IC and $\mathbf{w_C}(z)$ is a vector of width functions of different

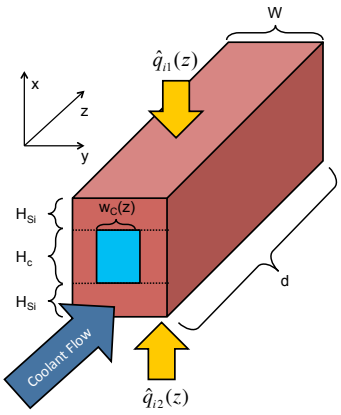


Fig. 2. Test structure- a single microchannel cooling two active silicon layers.

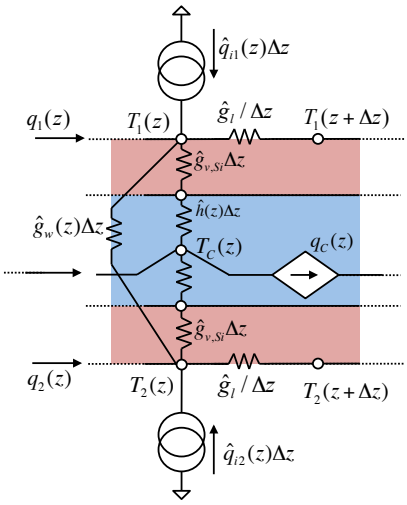


Fig. 3. A small element of the test structure, of length Δz and at a distance z from the inlet, with the equivalent electrical circuit.

microchannels written as a function of z. Our goal, then, is to find $\mathbf{w_C}(z)$ that minimizes the gradients in $\mathbf{T}(z)$.

In order to find the steady-state analytical model for heat transfer in an IC cooled by a microchannel, we consider a single microchannel structure of length d and width W shown in Fig. 2, between two silicon layers, and two silicon side walls. The width of the channel is a function of distance $w_C(z)$. The channel height (H_C) and the silicon die height above and below the channel (H_{Si}) are constants. Heat flux distributions $\hat{q}_{i1}(z)$ and $\hat{q}_{i2}(z)$ (measured here as the heat per unit length along z, $\mathrm{W/m}$) are applied to the top and the bottom layers of the silicon- referred to as the *active* layers. Coolant enters the microchannel at the inlet ($z = 0$) with a constant temperature T_{Cin}, absorbs heat from silicon along the way, and exits at the outlet ($z = d$). All exposed silicon surfaces are assumed to be adiabatic, hence the microchannel heat sink is the only way for the heat to exit the system. Using the electrical analogy for heat transfer in ICs with microchannel heat sinks, electric-circuit parameters can be written for a small element of this structure (Fig. 3), representing the following five heat transfers occurring in the structure [8, 11]:

(a) Longitudinal heat conduction inside the two active silicon layers, parallel to the microchannel ($\hat{g}_l$).
(b) Vertical heat conduction from the active silicon layers to surface of the top and bottom microchannel walls ($\hat{g}_w(z)$).
(c) Vertical heat conduction between the active silicon layers through the microchannel silicon side walls ($\hat{g}_{v,Si}$).
(d) Convective heat transfer from the surface of the microchannel walls into the bulk of the coolant ($\hat{h}(z)$).
(e) Convective heat transport downstream along the channel due to the mass transfer (flow) of the coolant ($q_C(z)$).

The state variables in this formulation are the temperatures in the two active silicon layers, $T_1(z)$ and $T_2(z)$, and the heat flowing in these layers parallel to the channel $q_1(z)$ and $q_2(z)$. $T_C(z)$ represents the temperature of the coolant as a function of the distance from the inlet. Assuming the silicon thermal conductivity to be k_{Si}, the volumetric heat capacity of the coolant to be c_v, and $\dot{V}$ its volumetric flow rate, we can write the following parameters for this circuit [8, 11]:

$$\begin{aligned}
\hat{g}_l &= k_{Si} \cdot W \cdot H_{Si} \quad (\mathrm{W \cdot m}) \\
\hat{g}_w(z) &= k_{Si} \cdot \frac{(W - w_C(z))}{(2H_{Si} + H_C)} \quad (\mathrm{W/m}) \\
\hat{g}_{v,Si} &= k_{Si} \cdot \frac{W}{H_{Si}} \quad (\mathrm{W/m}) \\
\hat{h}(z) &= h(z, w_C(z)) \quad (\mathrm{W/m}) \\
\hat{g}_v(z) &= (\hat{g}_{v,Si}^{-1} + \hat{h}(z)^{-1})^{-1} \quad (\mathrm{W/m}) \\
q_C(z) &= c_v \cdot \dot{V} \cdot T_C(z) \quad (\mathrm{W}).
\end{aligned} \tag{2}$$

The heat transfer coefficient $\hat{h}(z)$ is a function of various parameters, namely the Reynold's number of the flow, the coolant thermal conductivity, its viscosity, the distance from the inlet and, finally, the width of the channel $w_C(z)$. Our model is independent of the method used to estimate heat transfer coefficients: whether correlation studies based on experiments, or numerical techniques. For this work, we adopted the heat transfer coefficient calculated using the Nusselt number correlations (as a function of channel aspect ratio) presented by Shah & London [16].

Using the above parameters, the state-space analytical model for the heat transfer in the test structure can be derived in the following form:

$$\frac{d}{dz}\mathbf{X}(z) = \mathbf{F}(z, w_C(z), \mathbf{X}(z)) + \mathbf{G}(\hat{\mathbf{q}}_\mathbf{i}(z), T_{Cin}), \tag{3}$$

where

$$\mathbf{X}(z) = \begin{bmatrix} T_1(z) \\ T_2(z) \\ q_1(z) \\ q_2(z) \end{bmatrix}, \quad \hat{\mathbf{q}}_\mathbf{i}(z) = \begin{bmatrix} \hat{q}_{i1}(z) \\ \hat{q}_{i2}(z) \end{bmatrix}, \tag{4}$$

$\mathbf{F}(z, w_C(z), \mathbf{X}(z))$ is a non-linear function of distance, channel width and the states of the system, and $\mathbf{G}(\hat{\mathbf{q}}_\mathbf{i}(z), T_{Cin})$ is another vector function that is independent of the states and dependent solely on the input heat flux distributions and the inlet temperature of the coolant. Since all the exposed surfaces

of the IC are assumed to be adiabatic, the boundary conditions for the above analytical model can be defined as:

$$q_i(0) = q_i(d) = 0, \quad i = 1, 2. \tag{5}$$

The derivation of these functions are beyond the scope of this paper. This model has been validated against the numerical simulator 3D-ICE [8]. This model can be extended for the case of multiple channels adjacent to each other, by taking into account the additional heat spreading in the lateral (y) direction. Each added channel brings with it two additional nodes, namely, one for the top layer and one for the bottom layer of silicon (each node constitutes a temperature and a heat flow variable). It is also possible to combine two or more channels under a single set of top and bottom nodes to reduce the model complexity, by scaling the per-unit-length parameters in Eq (2) suitably.

IV. Optimal Channel Modulation Technique

In this section, we present the optimal control problem formulation for the thermal balancing of liquid-cooled 3D ICs. At the outset, we make the following assumptions in the problem formulation without a loss of generality:

1) Thermal gradients are dominant primarily *along* the path of fluid flow, which is a reasonable assumption based on observations in a multitude of ICs [1, 8].
2) Irrespective of the channel width function, the fluid is assumed to be under fully developed conditions with constant temperature-independent fluid parameters such as viscosity and density, for the computation of convective resistances [16]. Although the proposed technique is independent of the specific methods used for the computation of convective resistances, the study of these methods is beyond the scope of this work.
3) The volumetric flow rate in all channels is constant. Hence, we leave the pressure drop to be computed (constrained by an upper limit) during the search of the optimal solution, determining the final pumping effort.

A. Cost Function Definition

The primary goal of this work is to minimize the thermal gradient, along the fluid flow path, function expressed as $\frac{dT}{dz}$. From the analytical model discussed in the preceding section, we know that two temperature nodes are added for each channel in the system. Hence, temperature gradient variables to be minimized in a 3D IC with M cavity layers interleaved between $M+1$ active layers and each cavity containing N channels can be written as the vector:

$$\mathbf{T'} = \left[\frac{dT_1}{dz}, \frac{dT_2}{dz}, ..., \frac{dT_{(M+1)\cdot N}}{dz}\right]^T \tag{6}$$

The cost function for our optimal problem should represent the thermal gradients present in the target 3D IC. In addition, we must remember that temperatures rise and fall along the path of the channels, and we must somehow "accumulate" the effect of all the gradients (positive or negative) along the length of the IC. Thus, we define our cost function as the square of the euclidean norm of $\mathbf{T'}$. Our optimal control design problem can be formulated as:

$$\begin{aligned} \min_{\mathbf{w_C}(z)} J &= \int_0^d \|\mathbf{T'}\|^2 dz \\ \text{Subject to :} &\quad \text{1. Eq (3) and Eq (5)} \\ &\quad \text{2. Design constraints} \end{aligned} \tag{7}$$

Note that the temperature gradients in Eq (6) are directly related to the corresponding heat flow variables in Eq(4), from the laws of heat conduction. Hence, $\|\mathbf{T'}\|^2$ in the cost function above can be replaced by $\|\mathbf{q}\|^2$, where $\mathbf{q} = \left[q_1, q_2, ..., q_{(M+1)\cdot N}\right]^T$.

B. Design Constraints

All optimized design problems have constraints that limit the space in which the optimal solution must be found. In the case of channel modulation, we consider the following constraints:

1) Boundedness of Channel Widths: Area array through-silicon vias (TSVs) are the driving factors of 3D stacked integration and the microchannel heat sinks implemented in 3D ICs must be compatible with them. Hence, it must be ensured that maximum channel width is bounded to give clearance for the etching processes involved in the fabrication of TSVs. This bound depends upon the TSV pitch and diameter. On the other hand, channels cannot be arbitrarily thin. In addition to being difficult to fabricate, they offer excessive resistance to coolant flow requiring huge pumping effort [2]. These considerations require a minimum channel width to be defined. Thus, our optimal design problem is constrained with the following inequality for N channels:

$$\begin{aligned} w_{Cmin} \leq w_{Ci}(z) \leq w_{Cmax} \\ \forall z \in [0, d], \quad i = 1, 2, ..., N. \end{aligned} \tag{8}$$

2) Maximum Pressure Drop: Excessive pressure drops undermine the structural reliability of the 3D IC. In addition, given constant volumetric flow rates, large pressure drops directly imply large pumping effort. Hence, an upper bound for the pressure drop must be specified (note that the optimal design problem can alternatively be stated as minimizing the pumping effort, with an upper bound for the temperature gradient). We assume a 3D IC structure where all the channels are connected to a single coolant reservoir [3]. Thus, for N modulated channels, the pressure drop across each channel ($\Delta P_i, i = 1, 2, ..., N$) can be written as a function of the channel width function (based on the Darcy-Weisbach equation) and should satisfy the constraints:

$$\Delta P_i = \int_0^d 8\mu\dot{V}\frac{(H_C + w_{Ci}(z))^2}{(H_C \cdot w_{Ci}(z))^3} dz \leq \Delta P_{max}, \tag{9}$$

$$\Delta P_i = \Delta P_j, \forall i, j = 1, 2, ..., N, \tag{10}$$

where μ is the fluid dynamic viscosity and ΔP_{max} is the maximum allowable pressure drop in the system.

TABLE I
VALUES OF THE SYSTEM PARAMETERS

Parameter	Definition	value
k_{Si}	Silicon thermal conductivity	$130\ \mathrm{W/m \cdot K}$
W	Channel pitch	$100\mu\mathrm{m}$
H_{Si}	Silicon slab height	$50\mu\mathrm{m}$
H_C	Channel height	$100\mu\mathrm{m}$
c_v	Coolant volumetric heat capacity	$4.17 \cdot 10^6\ \mathrm{J/m^3 \cdot K}$
$\dot{V}$	Coolant volumetric flow rate	$4.8\ \mathrm{ml/min/channel}$
$T_{C,in}$	Coolant inlet temperature	$300\ \mathrm{K}$
ΔP_{max}	Maximum pressure difference	$10 \cdot 10^5\ \mathrm{Pa}$
w_{Cmin}	Minimum channel width	$10\mu\mathrm{m}$
w_{Cmax}	Maximum channel width	$50\mu\mathrm{m}$

C. Solving The Optimal Control Problem

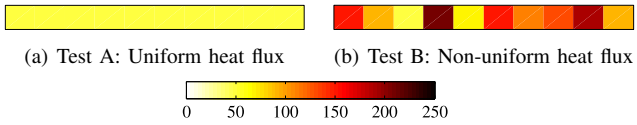
Once the problem statement in our optimal control design has been formulated, as discussed in the preceding subsections, we deploy the *direct sequential* solving method [17, 18] in determining the optimal solution. We discretize the channel width computation by enforcing piece-wise constant functions on w_C. However, the analytical state-space representation in Eq (3) is still preserved and no discretization is applied on it during the solving. There are various other solving methodologies available in the optimal control literature that can be used. However, studying the trade-offs between different solving techniques is beyond the scope of this work.

V. Experimental Results

To quantify the effects of our proposed optimal channel modulation technique, we apply it to various 3D IC structures to minimize the thermal gradients. First, we present the results from the single channel test structures shown in Fig. 2. Then, we show the impact of channel modulation on the thermal gradient of a 2-die 3D-MPSoC. In our evaluation, we compare the optimally modulated channels for the cases of using uniformly minimum and maximum possible channel widths. The values of the parameters used in our experiments are shown in Table I [2, 4]. It is important to note that our used maximum channel width ($w_{Cmax} = 50\mu m$) is the most common channel width used in previous works [3–5].

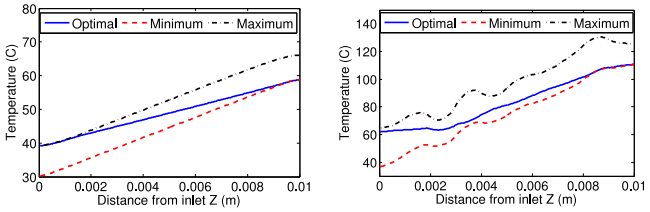
A. Thermal Gradient Analysis Using 3D IC Test Structures

The optimal channel modulation design is applied to the test structure in Fig. 2 with two different heat flux distributions as shown in Fig. 4. In Test A, Fig. 4(a), an input heat flux of $50\ \mathrm{W/cm^2}$ is applied to both the top and bottom surfaces of the silicon. While in Test B, Fig. 4(b), the top and the bottom silicon dies are divided into a group of segments, such that in each segment, a random heat flux value in the interval $[50, 250]\ \mathrm{W/cm^2}$ is applied, which is the range of power densities typically used to model the non-uniform heat dissipation of ICs in the literature [3]. Although Test B



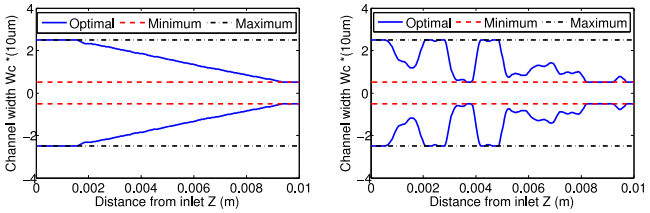
(a) Test A: Uniform heat flux (b) Test B: Non-uniform heat flux

Fig. 4. Planar view of the heat flux distributions in the our case studies (used range $[0 - 250]\mathrm{W/cm^2}$). The length of the strip is $d = 1\mathrm{cm}$.



(a) Test A (b) Test B

Fig. 5. Temperature change from inlet to outlet for the two tests . Each plot shows the change in temperature when using optimally-modulated, minimum and maximum channel widths respectively.



(a) Test A (b) Test B

Fig. 6. Channel width profile as a function of distance from the inlet for the two tests. The plots show the channel boundaries for optimally modulated, maximum and minimum width cases.

(Fig. 4(b)) is not a realistic power map, we use this case to illustrate the effectiveness and versatility of our proposal.

Temperature change in the silicon from the inlet to the outlet for the cases of optimally modulated, uniformly maximum and uniformly minimum channel widths for Tests A and B are plotted in Fig. 5. Using both uniformly maximum and uniformly minimum channel widths result in very similar thermal gradients (defined as the difference between the maximum and minimum temperatures in the IC): $28^{\circ}\mathrm{C}$ in Test A and $72^{\circ}\mathrm{C}$ in Test B. Together, the temperature distributions for these two cases define the upper and lower bounds of all possible temperature distributions than can be physically obtained using any channel modulation scheme for a given heat flux distribution. Hence, these figures show that the proposed technique achieves the minimum possible temperature gradient in both tests. In this case our proposed optimal design manages to reduce the thermal gradient by 32% ($19^{\circ}\mathrm{C}$ in Test A and $48^{\circ}\mathrm{C}$ in Test B) even for these high-heat flux scenarios.

The channel width profiles from inlet to outlet for the two tests are shown in Fig. 6. This figure shows that the optimal channel width selection is a function of both the distance from the inlet and the heat flux from the surrounding silicon dies. In Test A, with uniform heat flux (Fig. 6(a)), the channel width gradually decreases from the inlet to the outlet, to compensate for the increasing coolant temperature using increased heat transfer coefficients. On the other hand, in Test B (Fig. 6(b)) the channel width has to be made smaller in regions with higher heat fluxes than the immediate surroundings (e.g., between $0.4\mathrm{cm}$ and $0.5\mathrm{cm}$ from the inlet, as shown in Fig. 6(b)), in addition to the global trend of decreasing channel widths.

B. Channel Modulation Applied to 3D-MPSoCs

In this experiment, we apply the proposed method to different liquid-cooled 3D-MPSoC architectures to demonstrate

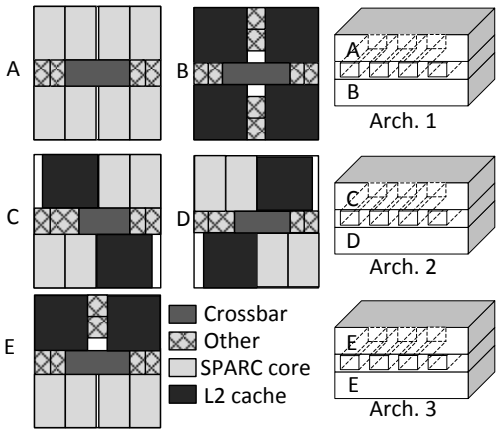


Fig. 7. Layout of the 3D-MPSoCs used in our experiments.

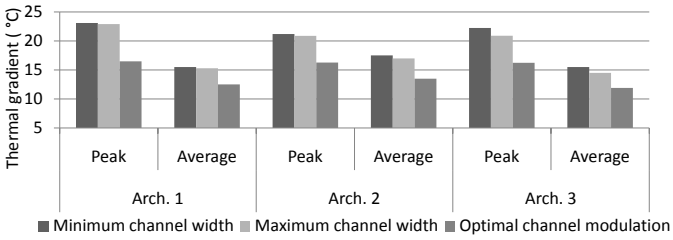
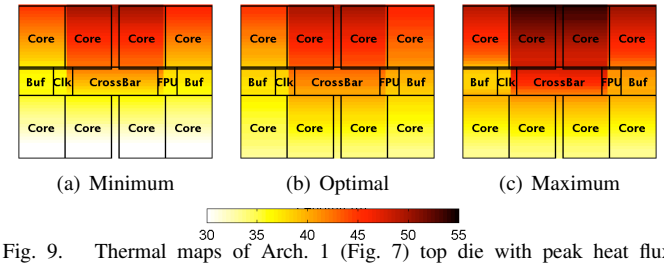


Fig. 8. Thermal gradients observed in the different 3D-MPSoC architectures dissipating peak and average level heat fluxes, using maximum, minimum and optimally modulated channel widths.

how the optimal channel modulation technique can be used in conjunction with the conventional floorplan exploration to obtain the desired thermal behavior during the IC design. For this we use different configurations of the 90nm UltraSPARC T1 (Niagara-1) processor [7] architecture. Fig. 7 shows the layout of the different two-dies 3D-MPSoCs used in this experiment. The dies are of size $1\ \mathrm{cm} \times 1.1\ \mathrm{cm}$ and the heat flux densities range from $8\ \mathrm{W/cm^2}$ to $64\ \mathrm{W/cm^2}$ in the two dies. For further details about the floorplan and power dissipations we refer the reader to [4, 5, 7].

In our optimization technique, we are using the worst-case (peak) power dissipation of the 3D-MPSoC functional elements [4, 5, 7] (obtained using measurements). Our proposed method achieves a thermal gradient reduction of 31% ($23^{\circ}\mathrm{C}$ to $16^{\circ}\mathrm{C}$). When the peak heat flux levels were replaced by average values, this same optimal channel modulation configuration manages to reduce the thermal gradient by 21% compared to the uniform channel width case. In addition, we observe that the peak temperature in the optimally modulated channel case equals to the peak temperature of the minimum channel width case, which is lower than the peak temperature of the maximum channel width case. Thus, our proposal implicitly minimizes the peak temperature to the lowest value achievable within a given channel width range. The thermal gradients obtained for the different cases and for various channel types are plotted in Fig. 8. Sample thermal maps of the *Arch. 1* top-die, for the case of peak heat flux are also plotted in Fig. 9 to illustrate the ameliorating effect our proposed method has on the thermal gradients. The direction of coolant flow is from bottom to top of the figures.



(a) Minimum (b) Optimal (c) Maximum

Fig. 9. Thermal maps of Arch. 1 (Fig. 7) top die with peak heat flux levels, when minimum, maximum and optimally modulated channel widths are applied. All the thermal maps are drawn with identical temperature scale ($[30 - 55]^{\circ}\mathrm{C}$).

VI. Conclusion

In this paper, we have proposed a thermal balancing methodology for liquid-cooled 3D ICs. Our method is based on channel width modulation to meet the cooling demands of different heat flux distributions in an IC. We model the system's thermal response along the coolant flow direction as an analytical state-space equation. Then, we formulate the channel width modulation as an optimal control design problem to deduce the optimal channel width trajectory, subject to the system dynamics and constraints. When applied to a realistic 3D-MPSoC, our method reduced the thermal gradient by 31%.

References

[1] B. Agostini et al. State of the art of high heat flux cooling technologies. *Heat Transfer Engineering*, 28(4).

[2] T. Brunschwiler et al. Hotspot-optimized interlayer cooling in vertically integrated packages. In *MRS Fall Meeting*, 2009.

[3] T. Brunschwiler et al. Interlayer cooling potential in vertically integrated packages. *Microsyst. Technol.*, 15(1):57 – 74, 2009.

[4] A. K. Coskun et al. Energy-efficient variable-flow liquid cooling in 3D stacked architectures. In *DATE*, pages 111–116, 2010.

[5] M. M. Sabry et al. Energy-Efficient Multi-Objective Thermal Control for Liquid-Cooled 3D Stacked Architectures. *IEEE Transactions On CAD*, 30(12):1883–1896, 2011.

[6] H. Qian et al. Cyber-physical thermal management of 3D multi-core cache-processor system with microfluidic cooling. *ASP Journal of Low Power Electronics*, 7(1):1–12, 2011.

[7] A. Leon et al. A power-efficient high-throughput 32-thread SPARC processor. *ISSCC*, 42(1):7 – 16, 2007.

[8] A. Sridhar et al. 3D-ICE: Fast compact transient thermal modeling for 3D-ICs with inter-tier liquid cooling. In *ICCAD*, pages 463–470, 2010. http://esl.epfl.ch/3d-ice.html.

[9] A. K. Coskun et al. Utilizing predictors for efficient thermal management in multiprocessor socs. *IEEE Transactions on CAD*, 28(10):1503–1516, 2009.

[10] D. B. Tuckerman and R. F. W. Pease. High-performance heat sinking for VLSI. *IEEE Electron Device Letters*, 5:126–129, 1981.

[11] A. Sridhar et al. Compact transient thermal model for 3D-ICs with liquid cooling via enhanced heat transfer cavity geometries. In *THERMINIC*, pages 1–6, 2010.

[12] H. Mizunuma et al. Thermal modeling and analysis for 3D-ICs with integrated microchannel cooling. *IEEE Transactions On CAD*, 30(9):1293–1306, 2011.

[13] B. Shi et al. Non-uniform micro-channel design for stacked 3D-ICs. In *DAC*, pages 658–663, 2011.

[14] M. M. Sabry et al. Towards thermally-aware design of 3D MPSoCs with inter-tier cooling. In *DATE*, pages 1–6, 2011.

[15] T. Brunschwiler et al. Angle-of-attack investigation of pin-fin arrays in nonuniform heat-removal cavities for interlayer cooled chip stacks. In *SEMI-THERM*, pages 116–124, 2011.

[16] R. Shah and A. London. *Laminar flow forced convection in ducts*. New York: Academic Press, 1978.

[17] J. T. Betts. *Practical methods for optimal control using nonlinear programming*. Siam, 2001.

[18] K. L. Teo et al. *A Unified Computational Approach to Optimal Control Problems*. Longman Scientific and Technical, 1991.