A semi-analytical approach for optimized design of microchannel liquid-cooled ICs

Arvind Sridhar, Mohamed M. Sabry, David Atienza
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{arvind.sridhar, mohamed.sabry, david.atienza}@epfl.ch

Abstract

The development of embedded and interlayer liquid cooling in integrated circuits (ICs) using silicon microchannels has gained interest in the recent years owing to the rise of on-chip heat uses that aggravate thermal reliability issues of the emerging 3D stacked ICs. Further development of such devices and their translation to commercial applications depend largely on the availability of tools and methodologies that can enable the "temperature-aware" design of liquid- cooled microprocessors and 2D/3D multiprocessor systems-on-chip (MPSoCs). Recently, two optimal design methods have been proposed for liquid-cooled microchannel ICs: one to minimize on-chip temperature gradients and the other, called GreenCool, to maximize energy efficiency in the coolant pumping effort. Both these methods rely upon the concept of channel width modulation to modify the thermal behaviour of a microchannel liquid-cooled heat sink. At the heart of both these methods is a new semi-analytical mathematical model for heat transfer in liquid-cooled ICs. Such a mathematical model enables the application of gradient descent approaches, such as non-linear programming, in the search for the most optimally performing channel design in a huge multi-dimensional design space. In this paper, we thoroughly quantify the impact and efficiency of the semi-analytical model, combined with non-linear programming, when compared against several numerical optimization mechanisms. Our experimental evaluation shows that non-linear programming, alongside the semi-analytical model, is up to 23x faster than conventional randomized/heuristic design approaches such as genetic algorithms and simulated annealing using fully-numerical thermal models.

Categories and Subject Descriptors

1.10 [**Three-Dimensional Electronics**]:

Keywords

Liquid-cooling of ICs, design optimization, channel width modulation

1. Introduction

Demands for high-performance and energy-efficiency in computing have encouraged research efforts in the development of compact liquid-cooled thermal packages for integrated circuits. Specifically, liquid cooling using intertier microchannels in stacked three-dimensional integrated circuits (3D ICs) has garnered interest in the recent years [1, 2]. Liquid-cooling has been chosen over air-cooling in high-performance thermal packages due to the superior thermal properties (thermal conductivities and heat capacities) of liquids compared to gases. Silicon microchannels here are etched directly on the back-side of individual IC dies before stacking them one over another and finally hermetically sealing the entire package, as shown in Fig. 1. Significant progress has been made in the development and validation of such advanced thermal packages [3, 4].

Large-scale adoption of liquid-cooling technology in electronics is only possible if design methodologies are in place that enable an early-stage *temperature-aware* exploration of the design-space and the optimization of various design parameters to provide a desired thermal response. Here, the term "temperature-aware" design or exploration indicates the inclusion of temperature or thermal response of the ICs as an important design metric during the design-space exploration in addition to the conventional metrics of electrical performance and energy consumption [5]. Such thermal-management methodologies can then be coupled with the existing vast pool of design-space exploration tools for performance enhancement and power management -such as floor planning, routing and DVFS (dynamic voltage and frequency scaling)- to obtain a holistic design cycle that realizes the full potential of this technology. Already, efforts have been made towards this end with the proposal for new compact thermal models for IC liquid cooling [6, 7] specifically meant for early-stage design. In addition, various dynamic thermal management schemes involving flow-rate control have been proposed to improve the energy-efficiency of microchannel liquid-cooling [8, 9].

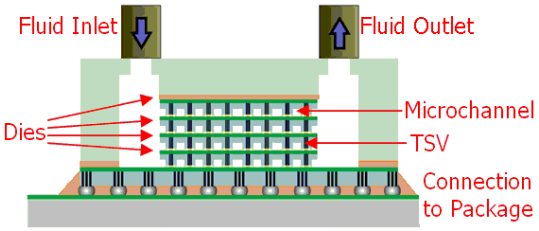


Figure 1: Stacked 3D IC with liquid cooling.

However, there is one avenue of optimized design specific to microchannel liquid-cooling that has not yet been fully explored but has the potential to open up various dimensions in early-stage temperature-aware design space exploration: *channel width modulation*. Channel width modulation entails the modification of the microchannel widths from inlet to outlet in specific forms and patterns to influence the heat-removal at different parts of the IC surface. The motivation for this comes from the well-known dependence of the local heat transfer coefficients on the aspect ratio of the microchannels [10]. By keeping the height of the microchannels- which is a function of the etching process during fabrication- constant, it is possible then to vary the microchannel widths from inlet to outlet to change the local aspect ratios (and hence the local heat removal capability) along the channel. It has minimal manufacturing overhead since it only involves using modulated channel masks instead of straight channel masks during the etching process. Using channel width modulation, thus, it is possible to obtain any desirable thermal property of the heat sink. [11] was the first attempt at using this concept to minimize temperature hotspots on the chip. But the methodology proposed here relies on heuristic metrics on increasing/decreasing channel widths with respect to the positions of hotspots along the channel and does not offer a robust solution to minimize a particular design cost (for eg. temperature gradients) by taking into account other design constraints (for eg. pressure-drops or pumping power).

We present two state-of-the-art applications of performing robust design optimization using channel width modulation: Application 1 [12] that minimizes on-chip temperature gradients and Application 2 (GreenCool) [13] that minimizes pumping en-

*This work was partly funded by the EC FP7 STREP GreenDataNet project (no. 609000), and the YINS RTD project (grant no. 20NA21_150939) evaluated by the Swiss NSF and funded by NanoTera.ch with Swiss Confederation financing.

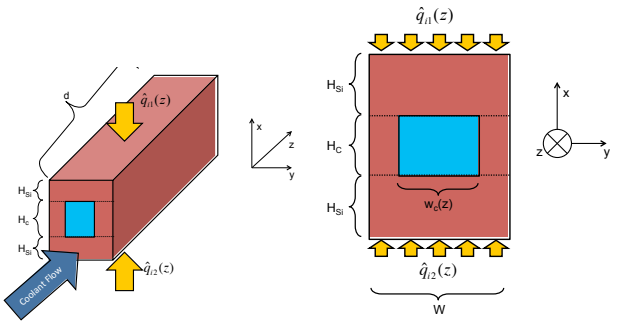


Figure 2: Heat transfer geometries: (a) Three-dimensional view of the test microchannel structure (b) Cross-sectional view of the test structure at a distance z from the inlet.

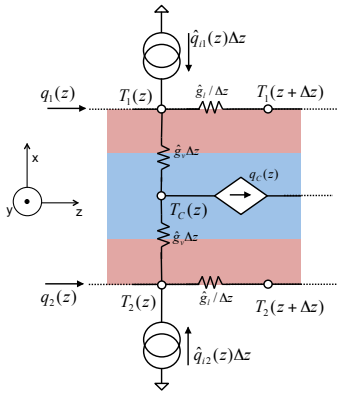


Figure 3: An infinitesimally small section of the test structure at a distance z from the inlet.

ergy for energy-efficient cooling with various design constraints. At the heart of these two applications lies a semi-analytical thermal model for heat transfer in liquid-cooled microchannels that has a state-space representation is suitable for the implementation of non-linear programming and other gradient descent methods for finding robust minimas in a design space. In this paper, we will present a detailed study of these applications and demonstrate how the semi-analytical model and gradient descent methods can provide better solutions for design optimizations using channel width modulation, compared to purely numerical models and heuristic methods such as genetic algorithms and simulated annealing. The contributions of this paper are summarized below:

1. The semi-analytical model for heat transfer in liquid-cooled microchannels for ICs will be briefly described.
2. Application 1 of the semi-analytical model, where on-chip temperature gradients in a liquid-cooled IC is minimized, will be described [12].
3. Studies on comparisons between the proposed method with traditional heuristic design methods such as genetic algorithms and simulated annealing will be presented.
4. Application 2 (GreenCool) of the semi-analytical model, where pumping power for a liquid-cooled IC is minimized for energy-efficient cooling, will be described [13].

2. A semi-analytical model for heat transfer

The basis of the semi-analytical model used in this study is well-known analogy between heat transfer and electric circuits, that is already being used in thermal models such as 3D-ICE [6]. However, in the semi-analytical model, there is no discretization along one dimension- i.e., along the dimensions where the channels are laid out. Hence the model and all the system variables along this dimension are continuous. To illustrate this, consider a single channel from a 3D IC with heat being generated on both the top and the bottom surfaces (corresponding to the active dies of an IC) as shown in Fig. 2. The semi-analytical model will be continuous along the z-direction and discretized along the x- and y- directions. Hence all system variables, such as temperatures on silicon surface T, heat flow q and heat inputs q_i, will be a series of functions along the z-dimension (multiple such functions can be written for each discretization point along the x- and y- directions).

By writing the equivalent electrical representation of heat conduction, convection and advection along the channel, for a small elemental section of this structure of size Δz at a distance z from the inlet (as shown in Fig. 3), the relationship between the temperatures and heat flows can be derived. By writing all system properties such as conductance $\hat{g}$ and heat inputs $\hat{q}_i$ as per unit length parameters and taking the limit $\Delta z \to 0$, we can obtain the following differential equations in the state-space form:

$$\frac{d}{dz}\mathbf{X}(z) = \mathbf{F}(\mathbf{X}(z), w_C(z), \hat{q}_i(z), T_{Cin}), \tag{1}$$

where

$$\mathbf{X}(z) = \begin{bmatrix} T_1(z) \\ T_2(z) \\ q_1(z) \\ q_2(z) \end{bmatrix}, \quad \hat{\mathbf{q}}_\mathbf{i}(z) = \begin{bmatrix} \hat{q}_{i1}(z) \\ \hat{q}_{i2}(z) \end{bmatrix}. \tag{2}$$

Here, $\mathbf{F}$ is the state-space function that defines the relationship between the state variables $\mathbf{X}$. $w_C(z)$ is the channel width expressed as a function of distance z, and $T_{C,in}$ is the constant channel inlet temperature. The full derivation of this function is beyond the scope of this paper. However, the following observations are made about this model:

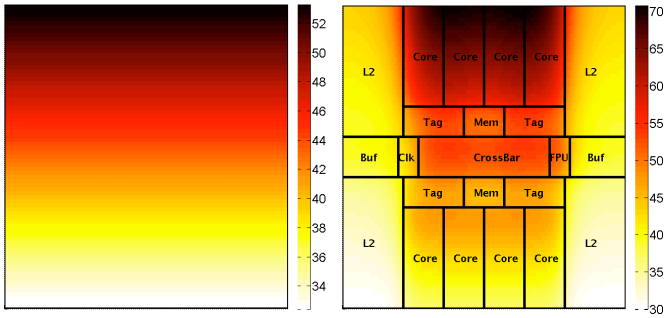
1. The model is extremely compact with only 4 system variables (functions) for the temperatures and heat flow in a two-die 3D IC per channel.
2. The accuracy of this model has been validated against numerical models such as 3D-ICE, which in turn have been validated against temperatures measurements from real liquid-cooled ICs.
3. Closed form solutions to this model can be derived for specific cases, such as when the heat inputs on either die are piece-wise constant functions along z, which is indeed the case for any realistic IC architecture.
4. The model can be extended for multiple channels stacked next to each other, to create a realistic IC die where an entire 2-dimensional area is cooled using multiple microchannels as in Fig. 1.
5. Multiple channels can be combined under one "channel block" by scaling the system properties accordingly in the semi-analytical model to create extremely compact representation of heat transfer in a 3D IC with liquid cooling.

The reader is referred to [12,13] for more details on the model. In the ensuing sections, two major design applications of this model will be presented in detail.

3. Application 1: Thermal gradient minimization

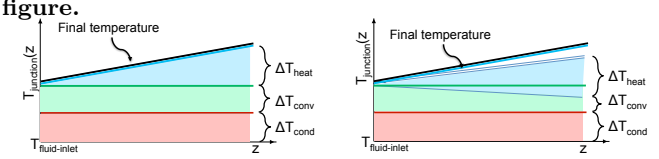
In this section, the first application of the semi-analytical model will be presented: channel width modulation for the minimization of on-chip temperature gradients in liquid-cooled ICs [12].

One serious challenge that comes with liquid cooling of ICs is the increased thermal gradient. Thermal gradients occur when different parts of the IC are at different temperatures. These



(a) Uniform heat flux (b) UltraSPARC T1

Figure 4: Temperature distribution from a 14mm x 15mm two-die 3D IC with (a) uniform heat flux density of 50W/cm^2 and (b) the UltraSPARC T1 (Niagara-1) chip architecture [14]- with nonuniform heat flux density ranging from 18W/cm^2 to 100W/cm^2. Direction of the coolant flow is from the bottom to the top of the figure.



(a) (b)

Figure 5: Junction temperature from inlet to outlet (for the case of uniform input heat flux distribution) for (a) an unmodulated channel (b) a modulated channel.

gradients cause uneven thermally induced stresses on different parts of the IC, significantly undermining overall system reliability and lifetime [15]. In the traditional air-cooled ICs, thermal gradients mainly because of nonuniform heat flux distribution of a heterogenous IC with multiple cores and memory blocks. Hotspot heat fluxes in these chips sometimes exceed the surrounding average heat flux by an order of magnitude. In liquid cooling, there is an additional component to increased thermal gradient that can come to dominate the temperature response of the IC: sensible heat absorption that occurs as the coolant flows along the microchannels increases the chip temperature from inlet to outlet. This second source of thermal gradients affects chips with uniform and nonuniform heat flux distributions alike as illustrated in Fig. 4. Fig. 5(a) illustrated the various contributions of the junction temperature of the chip from inlet to outlet. This illustration is for uniform input heat flux, but can easily be extended to non-uniform cases also. ΔT_{cond} is the temperature difference required for the conduction of heat in the silicon substrate from the junction to the channel. ΔT_{conv} is the contribution from convective heat transfer at the solid-liquid interface. Finally ΔT_{heat} is the contribution from the increasing sensible heat of the liquid from inlet to outlet, which ultimately contributes to the temperature gradient. One way to reduce its slope is to increase the flow rate so that heat is carried away from the chip more quickly, but this required additional pumping power. Another cost effective way to accomplish it is using *channel width modulation*.

3.1 Channel width modulation

Channel width modulation is the process of *modulating* or changing the width of a microchannel from inlet to outlet. This is mainly done to modify the local cooling properties of the microchannel. The motivation for channel width modulation comes from the study of heat transfer coefficients in microchannels. It is well known that for laminar flows in microchannels, very high aspect ratios (large heights and small widths) have higher heat transfer coefficients. Specifically, for fully developed

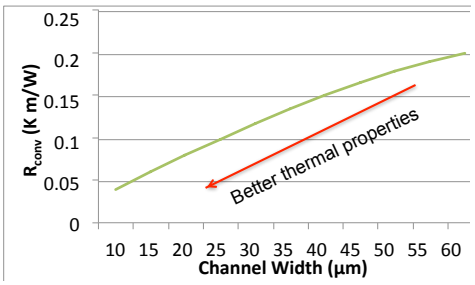


Figure 6: R_{conv} as a function of the channel width for a microchannel.

flows with isothermal channel perimeters, the nusselt number is given by the following expression [16]:

$$\begin{aligned}\text{Nu} = 8.235 \cdot (1 - 2.0421r + 3.0853r^2 \\ - 2.4765r^3 + 1.0578r^4 - 0.1861r^5)\end{aligned} \quad (3)$$

, where $r = w_C/H_C$, ratio width to height ratio of the microchannel cross-section. For a 100μm tall channel, the convective resistance as a function of the channel width according to the above relationship is illustrated in Fig. 6.

Hence, it must be theoretically possible to find a modulated channel *channel width function* $w_C(z)$ such that the resulting local convective resistance $R_{conv}(z)$, and in turn the convective temperature drop $T_{conv}(z)$, compensate for the temperature gradients on a chip with any type of heat flux distribution as illustrated in Fig. 5(b). That is, the goal of channel width modulation here is to achieve the flattest temperature distribution possible for a chip by finding an appropriate function $w_C(z)$. One advantage of channel width modulation is that it comes a negligible manufacturing cost compared to other solutions, as the same processes used for etching uniform channels can be used to create modulated channels as well. On the other hand, modulated channels could also increase resistance to flow and hence pumping effort. While [11] provided a heuristic method to perform channel modulation targeted at chip hotspots, a robust solution to such a complex challenge with trade-offs can only be obtained by defining it as an optimization problem:

$$\min_{\mathbf{w_C}(z)} J = \int_0^d \left(\frac{dT}{dz}\right)^2 dz, \quad (4)$$

Subject to :
1. System equations (eg. Eq (1))
2. $w_{C,min} < w_C(z) < w_{C,max}, \forall z$
3. $\Delta P < \Delta P_{max}$

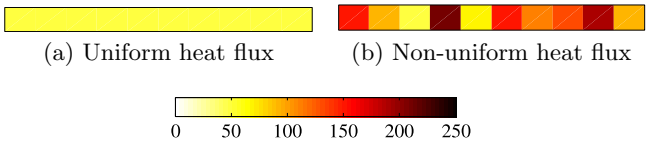
where J is a cost function to be minimized that integrates the temperature gradient in any selected or all junction surfaces of the chip dT/dz. The gradient terms are squared in the integration to accumulate both the rise and the fall of temperatures in the final cost function. The constraints 2. and 3. define the maximum and minimum bounds for the channel width $w_C(z)$ and the maximum allowable pressure drop across the channels. There are different methods of solving the above optimization problem which are discussed in the next subsection. The experimental results also demonstrate the efficacy of the semi-analytical model (Eq(1)) for this purpose.

3.2 Experimental results

One straightforward way to solve the optimization problem in Eq(4) is to do randomized or heuristic search of the entire allowable design space using methods such as Genetic Algorithms or Simulated Annealing [17]. However, such brute-force methods normally take a long time especially for large problem sizes with a large design space to scan. In addition, they do not provide a guarantee of minima even when they converge. In such scenarios where there are few local minima in the design space, gradient descent algorithms such as non-linear pro-

Table 1: Values of the system parameters

Parameter	Definition	value
k_{Si}	Silicon thermal conductivity	130 W/m · K
W	Channel pitch	100μm
H_{Si}	Silicon slab height	50μm
H_C	Channel height	100μm
c_v	Coolant volumetric heat capacity	$4.17 \cdot 10^6$ J/m^3 · K
$\dot{V}$	Coolant volumetric flow rate	4.8 ml/min/channel
$T_{C,in}$	Coolant inlet temperature	300 K
ΔP_{max}	Maximum pressure difference	$6.5 \cdot 10^5$ Pa
w_{Cmin}	Minimum channel width	10μm
w_{Cmax}	Maximum channel width	50μm



(a) Uniform heat flux (b) Non-uniform heat flux

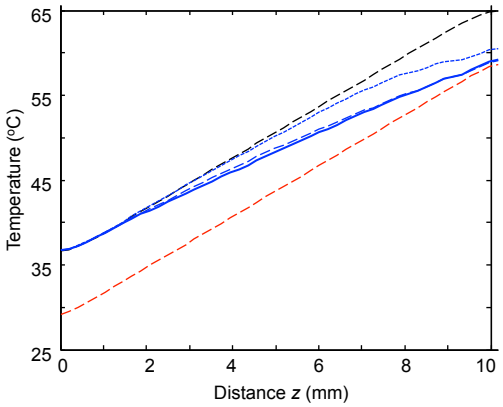
Figure 7: Heat flux distributions for (a) Experiment 1 and (b) Experiment 2.

gramming [18] offer a much more attractive solution. This is illustrated using the following two experiments. The structural and material properties used in all the ensuing experiments are tabulated in Table 1. All the experiments were performed on an Intel Corei7 3.40GHz processor with 32 GB RAM running Windows 7.

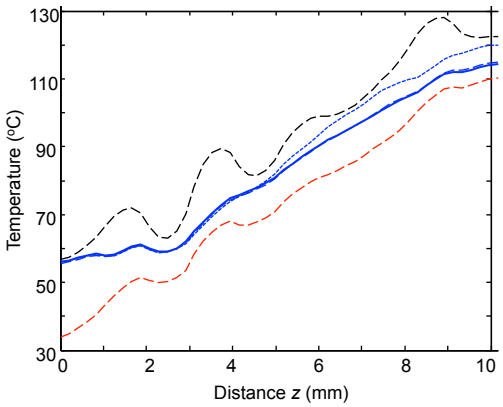
Experiment 1: In the first experiment we consider a simple case of a single microchannel as shown in Fig. 2 of length $d =$ 1cm. A uniform heat flux density of 50W/cm^2 is applied to both the top and the bottom junctions as shown in Fig. 7(a). First the junction temperatures were simulated using uniform minimum and maximum microchannel widths ($w_{C,min} = 10\mu$m and $w_{C,max} = 50\mu$m) that would the lowest possible and the highest possible convective resistance respectively. This would give us the lower bound and upper bound of the temperature distributions in the design space as shown in Fig. 8(a) using red and black lines. The temperature profile of any modulated channel with these constraints must lie within these bounds. Next, the optimization problem in Eq(4) was solved using the following three search algorithms:

1. Genetic Algorithm (GA), where starting from an initial guess design variables are randomly changed in every "generation". After thermal simulations, the best possible alternative (i.e. the one with the lowest cost function) is chosen as the search moves forward.
2. Simulated Annealing (SA), where in every iteration the design variables are changed based on both the change in cost function and also a predefined probability function, to find a global minimum in the design space that may have multiple local minima.
3. Gradient Descent Algorithm (GDA) using non-linear programming applied directly to the cost function in Eq(4).

For 1. and 2., the genetic algorithm and bounded simulated annealing packages in Matlab were used. For 3. the **fmincon** non-linear programming-based optimization package in Matlab was used [19]. In all the three cases a fully numerical thermal model called 3D-ICE [6] was used to solve the heat transfer system equations for each iteration during the design-space search. The discretization along the direction of the coolant flow was set to 10μm (which is more than sufficient for accuracy purposes [6]) and the gradients dT/dz in Eq(4) were calculated from the discrete temperature results of the simulations using forward difference approximation. The resulting temperature profiles and the channel width profiles from each of the three methods are shown in Fig. 8(a) and 9(a) using various blue lines. As can be seen, except the Genetic Algorithm, the *theoretical minimum* temperature gradient (the flattest possible



(a) Experiment 1



(b) Experiment 2

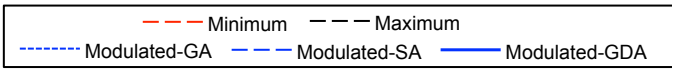


Figure 8: Temperature change from inlet to outlet for Experiments 1 and 2.

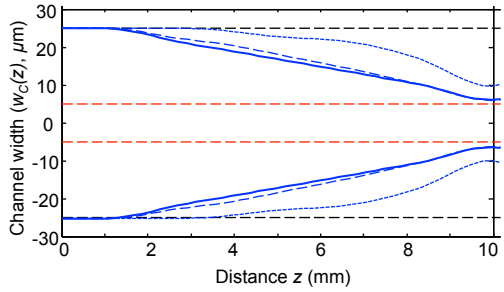
Table 2: CPU Execution times (min:sec) for Experiments 1 and 2

Optimization Method	Experiment 1	Experiment 2
Genetic Algorithm	16:24	17:41
Simulated Annealing	14:48	14:36
Gradient Descent	00:53	00:46

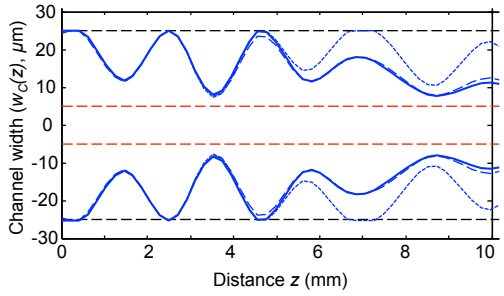
temperature profile) was obtained in all cases. The execution times for each algorithm are tabulated in Table 2. As can be seen, the gradient descent method outperforms the other two by up to 18X.

Experiment 2: The above experiment was repeated for the same single microchannel strip, but with non-uniform heat fluxes. For this, the strip was divided into 10 segments and in each segment (in both the top and the bottom layers) a random heat flux density in the interval [50 250]W/cm^2 was applied as shown in Fig. 7(b). Again, the temperature distributions using uniformly minimum and uniformly maximum channel widths are plotted using red and black lines in Fig. 8(b). These plots give the lower and upper bounds to the temperature distributions during the search. The channel width modulation was again performed using all the three algorithms (GA, SA and GDA). The temperature distributions and channel width profiles are plotted in Fig. 8(b) and 9(b). As before, we see that while the results show agreement, the execution time for the gradient descent approach (Table 2) is up to 23X lower than the other two methods.

The above experiments demonstrate that gradient descent methods such as **fmincon** are significantly superior to randomised or heuristic search algorithms for the purpose of chan-



(a) Experiment 1



(b) Experiment 2

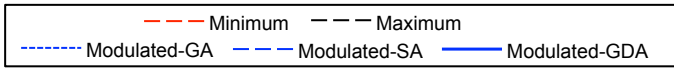


Figure 9: Channel width profile as a function of distance from the inlet for Experiments 1 and 2.

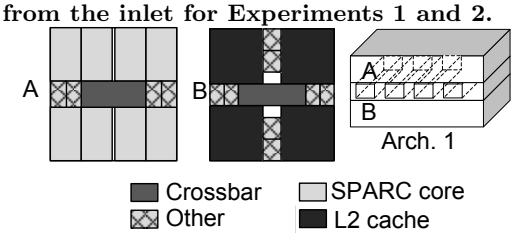
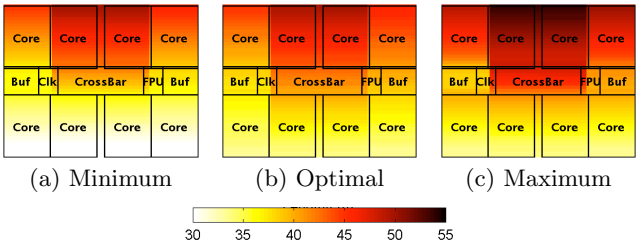


Figure 10: Layout of the UltraSPARC T1 3D-MPSoC used in Experiment 3.

nel width modulation.

Experiment 3: Owing to the extremely compact representation of the semi-analytical model compared to fully numerical models like 3D-ICE, it is ideally suited for solving the above optimization problem using a gradient descent approach. To demonstrate this, in this experiment, we consider a realistic two-die 3D IC with a single microchannel cavity between them. For this we use the 90nm UltraSPARC T1 (Niagara-1) multiprocessor system-on-chip (MPSoC) [14] architecture. Fig. 10 shows the layout of the 3D MPSoC. The dies are of size 1 cm × 1.1 cm and the heat flux densities range from 8 W/cm^2 to 64 W/cm^2 in the two dies [8, 14, 20]. There are 100 channels (with structural and material properties in Table. 1) in this configuration. They are grouped into 16 blocks of channels in the simulations. The gradient descent algorithm (GDA) using **fmincon** was run using both the semi-analytical model and the fully numerical 3D-ICE model as the underlying system model.

In both cases the channel width modulation using GDA achieves a thermal gradient reduction of 31% (23°C to 16°C) compared to uniform channels at peak heat dissipation of the MPSoC. Thermal maps of the top-tier (Fig. 10 A) using uniformly minimum, maximum and optimally modulated channel widths are plotted in Fig. 11 to illustrate the ameliorating effect the proposed method has on the thermal gradients. The execution times for GDA using the semi-analytical model and GDA using 3D-ICE are tabulated in Table 3. As can be seen using the



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

Figure 11: Thermal maps for Experiment 3 (Fig. 10) top tier when minimum, maximum and optimally modulated channel widths are used.

Table 3: CPU Execution times for Experiment 3

Optimization Method	CPU time (hr:min:sec)
Gradient Design with semi-analytical	00:48:31
Gradient Design with 3D-ICE	21:50:45

compact semi-analytical model gives speed ups as high as 27X compared to 3D-ICE in the search for the optimally modulated channel profiles for a complete 3D IC. Thus, the efficacy of both a gradient descent algorithm for solving the thermal gradient minimization problem and the efficacy of the semi-analytical model for this purpose have been demonstrated using these experiments.

4. Application 2: Energy-efficient cooling

In this section, the second application of the semi-analytical model will be presented: *GreenCool*-channel width modulation and optimization for minimizing cooling power in liquid cooled ICs [13].

A second challenge in the heat-removal and thermal packaging in ICs is the maximization of energy-efficiency by minimising the energy spent on cooling devices. Liquid cooling of electronics using microchannels has been advanced as possessing the potential to reduce convective resistances and improve energy efficiency, especially in high-performance computing installations and data centres [21]. However, for high-performance multiprocessor system-on-chips with high heat flux densities, coolant pumping power can be expensive especially with stringent thermal design constraints. To address this issue, we present another state-of-the-art application of channel width modulation and the semi-analytical model called the GreenCool [13] aimed at the minimization of coolant pumping power in liquid-cooled ICs.

This optimization problem begins, as before, with a cost function to minimize via channel width modulation, as follows:

$$\min_{\mathbf{w_C}(z),\dot{\mathbf{V}}} J = \mathbf{\Delta P} \cdot \dot{\mathbf{V}}^T. \quad (5)$$

Subject to :

1. System equations Eq. (1)
2. $w_{C,min} < w_C(z) < w_{C,max}, \forall z$
3. $\max(\mathbf{T}) < T_{max}$
4. $\max(\Delta\mathbf{T}) < \Delta T_{max}$

Here, $\mathbf{\Delta P} \cdot \dot{\mathbf{V}}^T$ is a measure of the pumping power where $\mathbf{\Delta P}$ is the vector of pressure drops in all channels of the IC and $\dot{\mathbf{V}}$ is the vector of volumetric flow rates. As before the channel widths are bounded by the design constraints $w_{C,min}$ and $w_{C,max}$. In addition there are the maximum absolute temperature and maximum temperature gradient constraints specified by the design T_{max} and ΔT_{max} (here, temperature gradient is defined the as the difference between the maximum and minimum temperatures in the IC). The pressure drop for each channel can be calculated using the Darcy-Weisbach equation:

$$\Delta P = \int_0^d 8\mu\dot{V}\frac{(H_C + w_C(z))^2}{(H_C \cdot w_C(z))^3} dz, \quad (6)$$

where μ is the dynamic viscosity and H_C is the channel height.

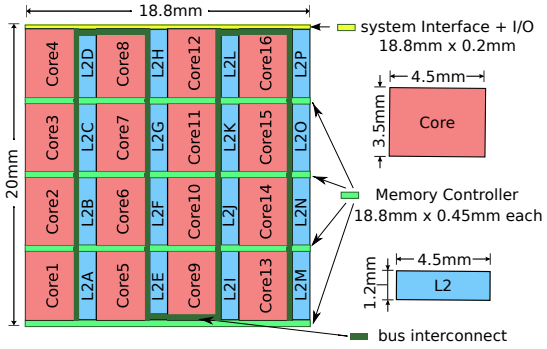


Figure 12: The layout of the logic layer of the target 3D system.

Table 4: Design constraints used in the GreenCool experiment

Symbol	Constraint	Value
T_{max}	Peak temperature constraint	60°C
ΔT_{max}	Peak temperature gradient constraint	12°C
$w_{C,min}$	Minimum channel width	30μm
$w_{C,max}$	Maximum channel width	80μm

The above optimization problem can be solved using the gradient descent algorithm **fmincon** with the semi-analytical model to obtain the optimal modulated channels for an IC that minimises cooling power while respecting the various thermal constraints of the design. This is demonstrated in the next subsection with experiments using the example of a realistic liquid-cooled IC.

4.1 Experimental results

In our experiments, we used a two-die 16-core 3D MPSoC with a DRAM-on-multicore architecture and a microchannel cavity in between. All the processing cores and caches in this 3D MPSoCs are on a logic layer, and the DRAM layer is stacked below it. TSVs vertically connect the logic and DRAM layers. The core layer architecture is based on the AMD Magny-Cours processors as in [22] as shown in Fig. 12. The processor is manufactured using 45nm technology and has a total die area of 376mm^2. The structural and the material properties of the microchannels used in these experiments were the same as those tabulated in Table 1

Extensive architectural exploration was performed to evaluate both electrical performance and power analysis. Various types of architectures were studies: with/without L2 caches; single-, 4-way and 8-way parallel TSV buses connecting the two layers. For each case, the peak performance heat flux densities were recorded. Next this data was fed into the GreenCool optimization problem (Eq(5)) to minimize cooling power via channel width modulation using the gradient descent algorithm **fmincon** and the semi-analytical model. This was repeated for each test case in the architectural design space. The design constraints used in these extensive experiments are tabulated in Table 4.

Studies from our experiments showed that performing channel width modulation results in pumping power savings up to 98% compared to the case with uniform minimum width or uniform maximum width channels. To further motivate the need for channel width modulation for this application, two different types of optimization problems were solved for each architectural test case:

1. Optimal modulated channel widths were computed for pumping power minimization according to Eq(5), as described above (designated as GreenCool with modulation).

2. Alternatively, optimal *uniform* channel widths were computed (i.e. channel widths no longer comprise a vector mathematical functions of distance from the inlet $\mathbf{w_C}(z)$, but are a vector of simple numbers $\mathbf{w_C}$) according to Eq(5) (designated as GreenCool without modulation).

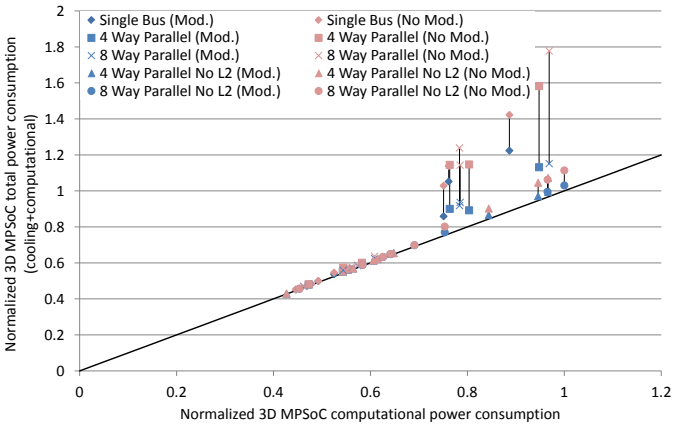


Figure 13: 3D MPSoC energy efficiency for GreenCool with and without modulation. Here, "Mod" indicates that channels have been optimally modulated.

For each test case the savings in pumping power and improvement in energy efficiency was compared between GreenCool with modulation and GreenCool without modulation. To quantify and better visualize the energy efficiency in these experiments, we use a metric called the power usage effectiveness (PUE), which is a metric often used for quantifying the efficiency of data centers. In this work, we define PUE as follows:

$$PUE = \frac{\text{3D MPSoC total power (computation+cooling)}}{\text{3D MPSoC computational power}}. \quad (7)$$

The closer PUE is to 1, the lesser the relative power spent on cooling, and hence, the more efficient the system. The energy efficiency in each case is visualised using the scatter plot in Fig. 13. Here the total power consumption (computational + cooling) is plotted against the computational power of the 3D MPSoC for the various experiments. Hence, the line x=y in this plot represents the case where PUE=1 (ideal energy efficiency). The power values are normalised to the power consumption of the most power intensive architecture encountered during the initial architectural exploration (corresponding to a heat flux density of 125W/cm^2). The red data points represent the results from using GreenCool without modulation for various architectures and the connected blue data points represent the corresponding results from GreenCool with channel width modulation. As can be seen from these results, in each case, GreenCool with channel width modulation gives better energy efficiency with smaller pumping power when compared to GreenCool without modulation. In fact, the energy savings were up to 35% without modulation and 6% on an average for all cases combined. These experiments demonstrate the need for channel width modulation in the maximisation of energy-efficiency in high-performance liquid-cooled ICs. Furthermore, they also underline the importance of the proposed modeling and optimization methods in such design-space explorations. Detailed analyses of these results can be found in [13].

5. Conclusions

In this paper, we presented a semi-analytical model for microchannel liquid-cooled ICs. The semi-analytical model provides an extremely compact state-space representation of the heat-removal process in a liquid-cooled ICs. Two different applications of this semi-analytical model, involving channel width modulation, in the "temperature-aware" design of liquid-cooled ICs has been presented. The first application minimizes on-chip thermal gradients using gradient descent algorithm and nonlinear programming. Experiments were performed to show that gradient descent methods outperform randomised or heuristic

design-space exploration methods such as genetic algorithms and simulated annealing by a factor of 23X. Furthermore, the semi-analytical model was shown to have a better performance compared to the conventional fully numerical thermal models with speed ups as high as 27X. The second application (GreenCool) minimizes coolant pumping power for high-performance liquid-cooled ICs and multiprocessor system-on-chips (MPSoCs) using gradient descent methods and the semi-analytical model. Power savings as high as 98% were obtained using the GreenCool applications. Experiments were also performed to demonstrate that GreenCool with channel modulation provides up to 35% more power savings compared to GreenCool without channel modulation, demonstrating the need for channel width modulation and the efficacy of the semi-analytical model in design optimization. The unique properties of the semi-analytical model may in the future give rise to other design appellations for liquid-cooled ICs.

6. References

[1] The CMOSAIC project, http://www.nano-tera.ch/projects/67.php, 2010.

[2] The ICECool project, http://www.darpa.mil/Our_Work/MTO/Programs/Intrachip/Interchip_Enhanced_Cooling_(ICECool).aspx, 2010.

[3] T. Brunschwiler *et al.*, "Heat-removal performance scaling of interlayer cooled chip stacks," in *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2010 12th IEEE Intersociety Conference on*, june 2010.

[4] T. Brunschwiler *et al.*, "Validation of the porous-medium approach to model interlayer-cooled 3d-chip stacks," in *International Conference on 3D System Integration (3DIC), San Francisco, California, USA*, 2009, pp. 1–10.

[5] S. Murali *et al.*, "Temperature-aware processor frequency assignment for mpsocs using convex optimization," in *CODESS+ISSS*, 2007.

[6] A. Sridhar *et al.*, "3D-ICE: Fast compact transient thermal modeling for 3D-ICs with inter-tier liquid cooling," in *ICCAD*, 2010, pp. 463–470.

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

[8] M. M. Sabry *et al.*, "Fuzzy control for enforcing energy efficiency in high-performance 3D systems," in *ICCAD*, 2010, pp. 642–648.

[9] M. M. Sabry *et al.*, "Energy-Efficient Multi-Objective Thermal Control for Liquid-Cooled 3D Stacked Architectures," *IEEE Transactions On CAD*, vol. 30, no. 12, pp. 1883–1896, 2011.

[10] F. Incropera, *Liquid cooling of electronic devices by single-phase convection.* John Wiley and Sons, 1999.

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

[12] M. M. Sabry *et al.*, "Thermal balancing of liquid-cooled 3d-mpsocs using channel modulation," in *DATE*, 2012.

[13] M. Sabry *et al.*, "Greencool: An energy-efficient liquid cooling design technique for 3-d mpsocs via channel width modulation," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 32, no. 4, pp. 524–537, 2013.

[14] A. Leon *et al.*, "A power-efficient high-throughput 32-thread SPARC processor," *ISSCC*, vol. 42, no. 1, pp. 7 – 16, 2007.

[15] A. K. Coskun *et al.*, "Utilizing predictors for efficient thermal management in multiprocessor socs," *IEEE Transactions on CAD*, vol. 28, no. 10, pp. 1503–1516, 2009.

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

[17] W. H. Press *et al.*, *Numerical Recipes: The Art of Scientific Computing (Chapter 10).* New York: Cambridge University Press (3rd Edition), 2007.

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

[19] Matlab Optimization Toolbox, http://www.mathworks.ch/ch/help/optim/index.html.

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

[21] T. Brunschwiler *et al.*, "Toward zero-emission data centers through direct reuse of thermal energy," *IBM Journal of Research and Development*, vol. 53, no. 3, pp. 11:1 –11:13, may 2009.

[22] J. Meng *et al.*, "Optimizing energy efficiency of 3d multicore systems with stacked dram under power and thermal constraints," in *DAC*, 2012.