3D-ICE: Fast Compact Transient Thermal Modeling for 3D ICs with Inter-tier Liquid Cooling*

Arvind Sridhar[1], Alessandro Vincenzi[1], Martino Ruggiero[1], Thomas Brunschwiler[2], David Atienza[1]

[1] Embedded Systems Laboratory (ESL), École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[2] Advanced Thermal Packaging Group, IBM Research Laboratory, Rüschlikon, Switzerland
{arvind.sridhar, alessandro.vincenzi, martino.ruggiero, david.atienza}@epfl.ch, tbr@zurich.ibm.com

ABSTRACT

Three dimensional stacked integrated circuits (3D ICs) are extremely attractive for overcoming the barriers in interconnect scaling, offering an opportunity to continue the CMOS performance trends for the next decade. However, from a thermal perspective, vertical integration of high-performance ICs in the form of 3D stacks is highly demanding since the effective areal heat dissipation increases with number of dies (with hotspot heat fluxes up to 250W/cm^2) generating high chip temperatures. In this context, inter-tier integrated microchannel cooling is a promising and scalable solution for high heat flux removal. A robust design of a 3D IC and its subsequent thermal management depend heavily upon accurate modeling of the effects of liquid cooling on the thermal behavior of the IC during the early stages of design. In this paper we present 3D-ICE, a compact transient thermal model (CTTM) for the thermal simulation of 3D ICs with multiple inter-tier microchannel liquid cooling. The proposed model is compatible with existing thermal CAD tools for ICs, and offers significant speed-up (up to 975x) over a typical commercial computational fluid dynamics simulation tool while preserving accuracy (i.e., maximum temperature error of 3.4%). In addition, a thermal simulator has been built based on 3D-ICE, which is capable of running in parallel on multicore architectures, offering further savings in simulation time and demonstrating efficient parallelization of the proposed approach.

Categories and Subject Descriptors

2.5 [**CAD for systems**]: Reliable and Alternative Systems

Keywords

Compact modeling, microchannel, 3D IC, inter-tier cooling

*This research has been partially funded by the Nano-Tera RTD project CMOSAIC (ref. 123618)- which is financed by the Swiss Confederation and scientifically evaluated by SNSF, and the PRO3D project- financed by the European Community 7th Framework Programme (ref. FP7-ICT-248776).

ICCAD '10 San Jose, California USA

1. INTRODUCTION

With the ever increasing demand for higher data rates and performance as well as multi-functional capabilities in circuits, vertical integration of IC dies using through-silicon vias (TSV) is envisioned to be one of the most viable solutions for the development of new generation of electronic products [1]. 3D integration of multi-core processors, for instance, offers massive bandwidth improvements, but reduces the effective chip footprint. However, this leads to a tremendous increase in heat dissipation per unit area of the chip. This in turn results in higher chip temperatures and thermal stress, hence, limiting the performance and reliability of the chip-stack [2].

Conventional back-side heat removal strategies like heat sinks, air cooling and microchannel cold-plates prove to be insufficient for 3D ICs and can only scale partially with the die size [3]. On the contrary, inter-tier microchannel-based liquid cooling can scale with the number of dies and is compatible with area-array TSVs, thus, is capable of removing heat from multi-processor 3D ICs [4]. However, the introduction of microchannel cooling in 3D stacks necessitates a cooling-aware design of 3D ICs, where the effects of forced convective liquid cooling are taken into account during the early-stages of design in order to achieve optimal performance under safe operating temperatures. Therefore, accurate thermal models are required for calculating the costs of operating the liquid cooling (pumping power), determining the overall energy budget and performing run-time thermal management. In this paper we propose 3D-ICE, a compact transient thermal model (CTTM) for liquid cooling for fast thermal simulation of 3D ICs with inter-tier microchannel cooling. The major contributions of the proposed model are:

1. The 3D-ICE model is transient and can accurately predict the temporal evolution of chip temperatures when operational system parameters (heat dissipation, flow rate of coolant etc.) change during dynamic thermal management. We have validated the accuracy of the model with a commercial computational fluid dynamics (CFD) simulation tool as well as measurement results from a 3D test IC (Fig. 1) with microchannel heat transfer geometry (a maximum error of 3.4% in temperature).

2. The 3D-ICE model is compatible with the conventional transient compact thermal models of IC modeling tools, like HotSpot [5]. Thus, no major computational overhead is introduced due to the introduction of microchannels in the proposed model. This is critical

for speeding up the performance-thermal optimization cycles during the design of 3D ICs (the proposed model shows up to 975x speed-up with respect to commercial CFD tools).

3. The proposed 3D-ICE model offers the designer the freedom to incorporate any suitable heat transfer geometry, such as microchannels or pin fin arrays. Only the empirical correlation set, representing the convective heat transfer has to be adjusted accordingly. These values can be derived from literature or can specifically be computed by conjugate heat and mass transfer CFD modeling of a unit-cell of the heat transfer structure.

4. A thermal simulator based on the proposed 3D-ICE model was implemented for multithreaded CPU. It was found that the parallelization of the simulation resulted in considerable time-savings, especially for large problem sizes (i.e., detailed thermal models for large 3D stacks with liquid cooling).

The rest of the paper is organized as follows. The previous work on the thermal modeling of ICs with microchannel cooling is discussed in Section 2. Then, Section 3 presents the architecture of a typical 3D stacked IC with inter-tier microchannel cooling and describes the problem to be solved. Next, Section 4 presents the development of the proposed compact thermal model for liquid cooling. Section 5 presents the details about the implemented thermal simulator. The experimental results are detailed in Section 6 and finally, the conclusions are presented in Section 7.

2. PREVIOUS WORK

A considerable amount of research has gone into developing compact thermal models for 3D ICs with conventional heat-sink based cooling [5, 6, 7, 8, 9]. Most of these methods (except [5] and [6]) present simplified thermal models for steady state simulations and provide no information about the transient thermal behavior of the ICs. The methods in [5, 6] use a finite-difference based method to generate a compact thermal model for the IC. In addition, while a high-level interconnect model is developed in [5] to simulate the effects of interconnect self-heating, [6] applies the alternative direction implicit (ADI) technique for obtaining fast and stable transient results. However, none of the above methods have provision for handling forced convective liquid cooling.

For the simulation of forced convective liquid cooling there are a number of empirical correlation-based methods available in the literature [10, 11, 12, 13, 14, 4]. All these methods are meant only for steady state simulations and are not specifically suitable for microchannel cooling.

More recently, a steady state thermal model for integrated microchannel cooling in 3D ICs was presented in [15]. In this method, the microchannels are discretized into small blocks along the direction of the flow and four new temperature nodes are added for each such block just inside each of the 4 channel walls. The temperature change in each of these nodes in the downstream direction is then calculated as a linear function of the heat flux entering a given upstream node using numerical presimulation. These linear functions, which model the "thermal wake" (the rise in temperature at a location downstream due to heating at a location upstream), are then incorporated in the heat conduction resistive model for the 3D stack. The three main drawbacks of this approach, which are addressed in the proposed model(see Section 6.2 for more details), are: First, steady state conditions are assumed and no transient information is obtained. Second, the problem size is very huge with four new nodes per microchannel block. And finally, extensive numerical presimulation calculations are needed for every thermal simulation executed.

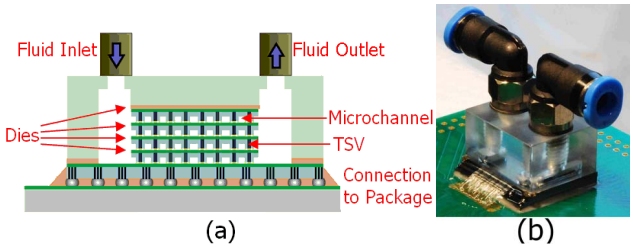


Figure 1: (a)3D stacked IC with interlayer microchannel cooling, (b) A 3D IC test vehicle, with lateral wire-bonds for electrical IO, fabricated with fluid manifold mounted on a printed circuit board.

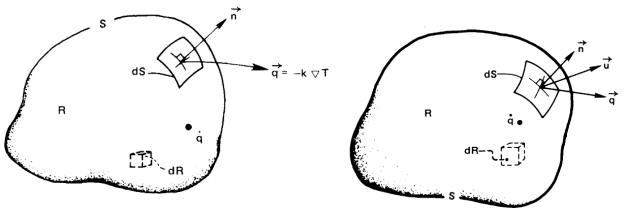
3. 3D STACKED IC ARCHITECTURE

Fig. 1(a) shows the schematic architecture that we consider for a typical 3D stacked IC with four active tiers and inter-tier cooling. The microchannel cavities are etched into the back of the die with TSVs running through the channel walls between the individual dies. A fluid manifold with inlet and outlet cavities is attached to the silicon carrier, resulting in the fluid containment and delivering the coolant to the individual fluid cavities. A 3D IC thermal test vehicle was manufactured including microchannels. To simplify the setup, lateral wire-bonding was utilized as electrical IO, instead of TSV as shown in Fig. 1(b). Heat dissipated from the IC tiers is predominantly removed through the coolant in the microchannels.

The goal of this paper is to build 3D-ICE, a compact thermal model based on finite-difference approximation, which takes into account the effects of the the inter-tier cooling through microchannel heat sinks. Considering the application of inter-tier cooling, with typically ten times less coolant mass flow rate compared to back-side heat removal, a fluid temperature increase of up to 30K was measured. Therefore the model has to track the fluid temperature increase from inlet to outlet. The footprint of the IC stack is assumed to be 10mm $\times$ 10mm with microchannel etched on the rear of a die in the silicon substrate having channel and wall cross sectional dimensions of about $50 - 100\mu m$ and $50 - 100\mu m$, respectively.

4. DEVELOPMENT OF THE PROPOSED CTTM

In this section, the proposed compact transient thermal model for microchannel cooling is presented. In the ensuing subsections, we first describe the conventional compact modeling of heat conduction in solids. Then, a compact model for fluids is derived from the first principles, and the analogies between the two compact models are drawn. Finally, we describe the incorporation of the compact fluid model in a complete 3D IC test case.



(a) (b)

Figure 2: Control volume of (a)solid (b)fluid [16].

4.1 Conventional compact modeling of solids

The conventional compact modeling for heat conduction in solids is done by applying finite-difference approximation to the governing equations of heat transfer in solids [5, 16, 17]. Consider a control volume of a solid R as shown in Fig. 2(a). The energy conservation equation for this control volume can be written as [16]:

$$\frac{d}{dt}\int_R \rho\ \hat{u}\ dR + \int_S (-k\nabla\ T)\ .\vec{n}\ dS = \int_R \dot{q}\ dR, \qquad (1)$$

where ρ is the density of the material, $\hat{u}$ is the enthalpy, S is the surface area of the control volume, k is the thermal conductivity of the material, $\vec{n}$ is the unit normal vector on the surface of the volume and $\dot{q}$ is the volumetric rate of generation of heat inside the volume. In the above equation, the first term on the left hand side is a volume integral representing the amount of heat energy stored in the volume. The second term is a surface integral representing the loss of heat from the volume due to heat conduction. The term on the right hand side is a volume integral representing the rate of generation of heat inside the volume due to conversion from another form of energy (chemical, electrical etc.). Thus, taking the limit $R \rightarrow\ 0$, applying Stoke's theorem [16], and assuming the material has isotropic thermal conductivity, the above equation reduces to:

$$C_v\frac{dT}{dt} + \left(-k\nabla^2\ T\right) = \dot{q}, \qquad (2)$$

where C_v is the volumetric specific heat of the material and T is the temperature of the control volume. The above partial differential equation can be converted into an ordinary differential equation by applying the finite difference approximation to the spatial derivative (the second term on the left hand side) in the above equation [16, 17]. To this end, the given volume of solid is discretized along the 3 Cartesian coordinates with discretization lengths $\Delta\ x$, $\Delta\ y$ and $\Delta\ z$, respectively, to generate a thermal grid. If the temperature of each node in the grid is represented by its location as $T_{i,j,k}$, then the finite difference approximation for the above equation at the location (i, j, k) can be written as,

$$\begin{aligned} C_v\Delta\ x\Delta\ y\Delta\ z\frac{dT}{dt} &- k\frac{T_{i+1,j,k} - 2T_{i,j,k} + T_{i-1,j,k}}{\Delta\ x^2} \\ &- k\frac{T_{i,j+1,k} - 2T_{i,j,k} + T_{i,j-1,k}}{\Delta\ y^2} \\ &- k\frac{T_{i,j,k+1} - 2T_{i,j,k} + T_{i,j,k-1}}{\Delta\ z^2} \\ &= \dot{q}\Delta\ x\Delta\ y\Delta\ z. \end{aligned} \qquad (3)$$

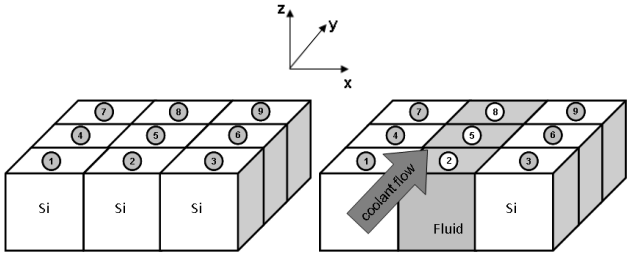
The well-known analogy between heat and electrical conduction is invoked here with the temperature represented as voltage, heat flow represented as electric current [5], the first term on the left hand side in the above equation represented as a capacitor and the rest of the terms on the left hand side represented as conductances, giving rise to an RC circuit [16]. Then, for an IC the compact thermal model is generated considering a single silicon layer of a die divided into 9 different "thermal cells", as shown in Fig. 3(a). Each thermal cell has a length l, width w and height h, as shown in Fig. 4, modeled as a node containing six resistances representing the conduction of heat in all the six directions (top, bottom, north, south, east and west), and a capacitance representing the heat storage inside the cell. The conductance of each resistor and the capacitance of the thermal cell are calculated as follows:

$$\begin{aligned} g_{top/bottom} &= k_{Si}\cdot\ \frac{l\cdot\ w}{(h/2)},\ g_{north/south} = k_{Si}\cdot\ \frac{l\cdot\ h}{(w/2)}, \\ g_{east/west} &= k_{Si}\cdot\ \frac{w\cdot\ h}{(l/2)},\ c_{cell} = C_{vSi}\cdot\ (l\cdot\ w\cdot\ h). \end{aligned} \qquad (4)$$

Here, the subscript *top*, *east*, *south* etc. indicate the direction of conduction (i.e., "north" here represents conduction in the $+y$ direction, "west" represents conduction in $-x$ direction and so on). Current sources of value $(\dot{q}\cdot\ l\cdot\ w\cdot\ h)$, representing the sources of heat, are connected to the cells wherever there is heat dissipation. Next, the nodes of these thermal cells are connected to the nodes of their neighboring cells through the interfaces by computing the equivalent conductances between them. Hence, the following system of ordinary differential equations are created:

$$\mathbf{G}\mathbf{T}(t) + \mathbf{C}\dot{\mathbf{T}}(t) = \mathbf{U}(t), \qquad (5)$$

where $\mathbf{T}(t)$ is the vector of all node temperatures (as a function of time) ordered according to their numbering in Fig. 3(a), $\mathbf{C}$ is a diagonal matrix of all cell capacitances calculated using Eq (4), $\mathbf{U}(t)$ is a vector of inputs (heat sources



(a) (b)

Figure 3: Discretization of a single layer of silicon-(a)without microchannel and (b)with microchannel-into "thermal cells".

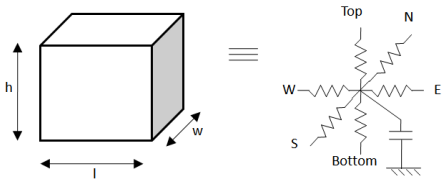


Figure 4: Equivalent circuit of a solid thermal cell.

as a function of time) wherever they exist.**G** is a symmetric block tri-diagonal conductance matrix, where non-zero non-diagonal elements respresent the connections between neighboring nodes and the diagonal term corresponding to a given node is equal to the sum of all conductances between that node and its neighbors. The formulation of heat flow equations, as described above, can be extended to structures containing multiple layers of thermal cells. Indeed the structure of the **G** will still remain block tri-diagonal and sparse. Then, the boundary conditions are given as source terms in the **U** vector. For example, if the top layer is connected to the ambient via some silicon-air-thermal resistance, then the nodes on the top layer are grounded to the ambient temperature via that conductance term. This method can be used to generate a compact thermal model for any general heterogeneous structure like an IC die, and the three-dimensional temporal evolution of heat inside the 3D IC can be accurately modeled.

4.2 Compact modeling of fluids

The energy conservation equation for heat transfer in a control volume of liquid, similar to the one described in the previous subsection, and can be written as (Fig. 2(b)) [16]:

$$\frac{d}{dt}\int_R \rho\ \hat{u}\ dR + \int_S (-k\nabla\ T).\vec{n}\ dS + \int_S \left(\rho\ \hat{h}\right)\ \vec{u}\cdot\ \vec{n}\ dS = \int_R \dot{q}\ dR. \tag{6}$$

In this case, when compared to Eq (1), the above equation contains an added term on the left hand side. Here, $\vec{u}$ is the velocity of outflow of the fluid at the surface of the control volume. This term indeed represents the net outflow of heat from the control volume due to convection. As in the conventional compact model, taking the limit $R \rightarrow\ 0$ and applying Stoke's theorem we get,

$$C_v\frac{dT}{dt} + \nabla\ \cdot\ (-k\nabla\ T) + C_v\vec{u}.\nabla\ T = \dot{q}. \tag{7}$$

Thus, the conduction term is no longer written using a Laplacian operator because heat conduction in a flowing fluid is not necessarily isotropic. The new divergence (third) term on the left hand side represents a "temperature controlled heat source". From Eq (6), it can be deduced that this convection term can be calculated for each surface of a small cuboidal thermal cell as a product of the velocity of the fluid flowing out, the surface temperature, the area of the surface and the volumetric heat capacity of the fluid. If at a particular surface the fluid is flowing in, then a negative sign must be assigned to the term. Finally, these convection terms must be summed up to calculate the cumulative effect of flow of fluid on the convective heat transfer inside the volume. Thus, by applying finite difference approximation, similar to the case of solids, for a given liquid cell with unidirectional fluid flow (towards the "north" or $+y$ direction,

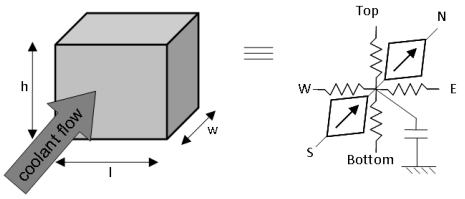


Figure 5: Equivalent circuit of a fluid thermal cell.

as shown in Fig. 5) we obtain:

$$\begin{aligned} C_v\Delta\ x\Delta\ y\Delta\ z\frac{dT}{dt} &- k_{xx}\frac{T_{i+1,j,k} - 2T_{i,j,k} + T_{i-1,j,k}}{\Delta\ x^2} \\ &- k_{yy}\frac{T_{i,j+1,k} - 2T_{i,j,k} + T_{i,j-1,k}}{\Delta\ y^2} \\ &- k_{zz}\frac{T_{i,j,k+1} - 2T_{i,j,k} + T_{i,j,k-1}}{\Delta\ z^2} \\ &+ C_v u_{avg,y}\Delta\ A_y\ (T_{S2} - T_{S1}) \\ &= \dot{q}\Delta\ x\Delta\ y\Delta\ z. \end{aligned} \tag{8}$$

In the above equation, the terms k_{xx}, k_{yy} and k_{zz} are the conductivity of the fluid in the x, y and z directions, respectively. $u_{avg,y}$ is the average velocity of the fluid in through the cell in the y direction, namely, the only non zero component of the velocity, with fluid entering from the front end and exiting from the rear end of the fluid cell as indicated in Fig. 5. The terms T_{S2} and T_{S1} represent the surface temperatures of the rear and the front ends, respectively.

4.3 Compact model for microchannels

At this point the theory developed in the previous subsection for compact modeling of liquids can be applied to the case of microchannel cooling of 3D ICs. For this, consider a single microchannel layer of an IC divided into 9 thermal cells, as before, as shown in Fig. 3(b). The microchannel is laid out in the middle with silicon walls on either side and hence cells 2, 5, and 8 are the "microchannel cells". Then the coolant flow is in the $+y$ direction. Since convection dominates conduction in this direction by many orders of magnitude, the second conduction term in Eq (8) can be neglected for these cells. Hence, g_{north} and g_{south} components of the microchannel cells do not exist in the proposed model. The convection term, which appeared as a "temperature controlled heat source" term above, can be consequently translated into a *"voltage-controlled current source"* in the equivalent RC circuit.

To this end the first step is to compute this convection term. It can be seen from Eq (8) that, given the nature of discretization of the channel layer of the IC, all the terms except those in the brackets in the convection term, i.e. $C_v u_{avg,y}\Delta\ A_y$, are constants. Here, $A_y = l\cdot\ h$ and coolant velocity $u_{avg,y}$ can be calculated as,

$$u_{avg,y} = \dot{V}\cdot\left(\frac{1}{A_y}\right), \tag{9}$$

where $\dot{V}$ is the volumetric flow rate of the coolant per microchannel in the 3D IC. By applying central differencing scheme [18] the interface temperatures, T_{S2} and T_{S1}, for each cell can be calculated using a first order approximation by interpolating the node temperatures of two microchannel cells which share that interface. For example, for cell 5 in Fig. 3(b), the interface temperature T_{S2} of the north face can

be calculated as $\frac{T_5+T_8}{2}$ and the interface temperature T_{S1} of the south face can be computed as $\frac{T_5+T_2}{2}$, assuming uniform discretization in the y direction. Hence, the convection term for cell 5 in Eq (8) becomes:

$$C_v u_{avg,y} \Delta\ A_y \left(T_{S2,5} - T_{S1,5}\right) = -c_{conv} \cdot \left(T_8 - T_2\right), \quad (10)$$

where $c_{conv} = C_v u_{avg,y} \frac{\Delta\ A_y}{2}$. These *"voltage controlled current sources"* model the transport of heat from the inlet to the outlet of the microchannel and, hence, account for the rise in temperature of the coolant as it flows through the microchannel. For the boundary surfaces, i.e, front surface of cell 2 and rear surface of cell 8, the surface temperatures are calculated as follows: the front surface of cell 2 is where the coolant enters the microchannel, which means that the coolant here is at a constant inlet temperature obtained from the refrigeration design. Hence, the temperature of the front surface of cell 2 is set to a constant T_{inlet}. As seen from Eq (8), this is multiplied by $2c_{conv}$ and taken to the right hand side of Eq (8) serving as an input source for the equivalent RC circuit (Dirichlet boundary condition [19]). Since there is no heat flux into the coolant beyond the outlet of the microchannel, i.e., the rear end of cell 8, it can be assumed that there is no rise in the coolant temperature at the outlet. Thus, the spatial derivate of temperature along the $+y$ direction at this surface can be approximated to a first order by setting the difference $T_8 - T_{S2,8}$ to zero (Neumann boundary condition [19]). In this case, although the temperature gradient is set to zero at the microchannel outlet, the heat flux out of the microchannel is not zero. This is because, unlike heat conduction, the convection term given by Eq (10) depends upon the absolute temperature of the surfaces and not on the temperature gradient.

Next, the four conductances g_{top}, g_{bottom}, g_{east} and g_{west} for the microchannel cell, which account for the heat transfer from the walls of the microchannel to the fluid, must be computed. They can be calculated using heat transfer coefficients obtained from either empirical correlations or numerical presimulation, as follows:

$$g_{top/bottom} = h_{f,vertical} \cdot \left(l \cdot w\right),\ g_{east/west} = h_{f,side} \left(h \cdot w\right), \quad (11)$$

where $h_{f,vertical}$ and $h_{f,side}$ are, respectively, the vertical and side heat transfer coefficients for microchannel forced convection, and are functions of the channel dimensions and coolant velocity. For the proposed model, the heat transfer coefficient on all sides are calculated using the formula:

$$h_{f,vertical} = h_{f,side} = \frac{k_{coolant} \cdot Nu}{d_h}, \quad (12)$$

where $k_{coolant}$ is the thermal conductivity of the coolant and d_h is the hydraulic diameter of channel, defined as $\frac{2h \cdot\ l}{(h+l)}$. Nusselt number (Nu) correlations considering imposed axial heat flux and radial isothermal conditions are assumed to represent the given heat transfer mode for the implemented microchannels with low aspect ratio fins most appropriate and were derived by London and Shah [20], as follows:

$$\begin{aligned} Nu &= 8.235 \cdot (1 - 2.0421AR + 3.0853AR^2 \\ &\quad - 2.4765AR^3 + 1.0578AR^4 - 0.1861^5) \\ fr \cdot\ Re &= 24(1 - 1.3553AR + 1.9467AR^2 \\ &\quad - 1.7012AR^3 + 0.9564AR^4 - 0.2537AR^5) \end{aligned} \quad (13)$$

In this context, AR is the aspect ratio of the channel h/l.

In this study the microchannel geometry was assumed to be constant, defined through the technological constraints from TSV fabrication and the assumption of uniformly distributed TSVs with a fixed pitch and diameter. With the approximation of developed hydrodynamic and thermal boundary layers the Nusselt number becomes invariant to coolant velocity in case of microchannels, as shown in Eq (13).

Finally, the capacitance for the microchannel cell is calculated in the same way as for solids in Eq (4). Once these model parameters are obtained, the silicon and microchannel cells are connected as described in the Subsection 4.1 and the ordinary differential equations similar to Eq (5) can be obtained. The new $\mathbf{G}$ retains the same block tri-diagonal structure of the $\mathbf{G}$ matrix for the case of a pure solid structure. However, the non-zero non-diagonal elements of the new conductance matrix corresponding to two neighboring liquid cells is given by (10) and this convection term gets added to the diagonal elements corresponding to cells at the microchannel inlet and outlet. The resulting matrix is no longer symmetric because of the unidirectional flow of heat due to convection (i.e., this asymmetry reflects the non-reciprocity of the system).

5. IMPLEMENTATION FEATURES

A software thermal library was built in C based on the compact thermal modeling discussed in the previous section. This 3D-ICE design exploration tool with inter-tier cooling is available at [21]. This thermal library is flexible and accepts a variety of 3D IC stack descriptions as input and produces time domain chip temperatures waveforms as outputs. The inputs to the Library are: a) the physical description of the IC- layers (comprising the stack and their material properties) b) floorplan information of each individual die (reflecting location of various circuit blocks and their power dissipation values) c) the discretization parameters (thermal cell size, time-step and time of simulation). The physical composition and the floorplan are given through netlist files. For the power dissipation values, the user has the option of either specifying them as a function of time in a text file or feed them from an external device (e.g., an emulation platform) via a network socket.

The netlists are parsed with functions generated by Bison [22] and Flex [23] tools. From the data structure so filled, a three dimensional "Thermal Grid" matrix is generated with the properties of individual thermal cells, as described in Section 4.1. Size of the thermal cell for the channel layers depend upon the channel dimensions since, as can be seen from Fig. 3(b), the fluid cells must comprise of the entire cross section of the channel. The user has the freedom to choose any cell dimension in the direction along the channel. In our experiments, we found that cell sizes of few hundred micrometers are sufficient for accuracy. The source terms are computed for each thermal cell using the power values from the user and the boundary conditions. Because a fluid manifold encloses the 3D IC as shown in Fig. 1, the exposed surfaces of the IC stack are assumed to be adiabatic.

The next step is the formulation of equations for the simulation of the Thermal Grid. For this, Eq (5) is integrated numerically using the backward Euler method as follows:

$$\begin{aligned} \left(\mathbf{G} + \frac{1}{h}\mathbf{C}\right)\mathbf{X}(t_{n+1}) &= \mathbf{U}(t_{n+1}) + \frac{1}{h}\mathbf{C}\mathbf{X}(t_n) \\ \Rightarrow \mathbf{A}\mathbf{X}(t_{n+1}) &= \mathbf{B}_{n+1}, \end{aligned} \quad (14)$$

where h is the time-step used for the numerical integration, $\mathbf{A} = \mathbf{G} + \frac{1}{h}\mathbf{C}$ and $\mathbf{B}_{n+1} = \mathbf{U}(t_{n+1}) + \frac{1}{h}\mathbf{C}\mathbf{X}(t_n)$. Here, t_n denotes the n^{th} time point during the transient simulation. The matrices $\mathbf{A}$ and $\mathbf{B}$ are generated from the Thermal Grid and stored in a compressed column format [24]. Next, the matrices so generated are fed to a sparse linear system solver. For the proposed 3D thermal library, two different matrix solver libraries were tested: KLU [25] and SuperLU [26, 30]. In our experiments we found the latter to be faster.

The user has the option of specifying the simulation with default initial temperatures for the Thermal Grid ($X(t_0)$, with the initial temperatures being the ambient temperature), or resume from the end of a previous simulation. Finally, temperatures for one or more circuit blocks, as indicated by the user as the observation points, are recorded as the output temperatures waveforms and are either printed out in the form of a text file or can be sent through a network socket for possible HW-SW cosimulation. The sparse system solver libraries are the most computationally dominant part of the thermal library and hence, possible future work in this direction could include exploring different solver techniques, which can exploit parallel computing architectures like GPUs.

6. EXPERIMENTAL RESULTS

In this section we evaluate the accuracy and CPU performance of the proposed 3D-ICE thermal model.

6.1 Accuracy evaluation of the CTTM model

In the first set of experiments, we have validated the accuracy of the proposed model. To this end, we compared the results from the proposed model with a commercial computational fluid dynamics (CFD) simulation tool, i.e., Ansys CFX [27], which is a finite volume method based solver. In all our experiments, the fluid flowing through the channel was assumed to be hydrodynamically fully developed. Therefore, periodic hydrodynamic boundary conditions with a fixed pressure gradient (10^5Pa)were imposed, to derive the velocity field of the coolant. In a subsequent run, the velocity field was used as initial conditions and the energy equation with an imposed power step was computed.

Two main structures were simulated in our experiments: (a) **Test Stack 1** with two active dies and one microchannel cavity between them and (b) **Test Stack 2** with three active dies and four microchannel cavities adjacent to them. Both **Test Stack 1** and **Test Stack 2** have a footprint of 10mm × 10mm. A small slice of both the ICs, showing the composition of stack and the cross section of the channels is shown in Fig. 6. The material and structural properties used in our experiments for both experimental stacks are tabulated in Table 1 [28]. To minimize the model complexity , only half of the microchannel and microchannel wall with symmetry boundary conditions were taken into account in the CFD model. This 50μm wide and 10mm long computational domain for the four cavity test stack resulted in an unstructured hexahedral mesh with 170k nodes. In invariant heat fluxes transversal to the flow direction (i.e. along x direction if the microchannels are laid out in the y direction) were imposed, resulting in a periodic temperature variation along the x direction. However, the entire stack was simulated in our thermal library since our discretization (a single Thermal Cell had a footprint of around 50μm × 50μm) resulted in much smaller problem sizes and CPU times.

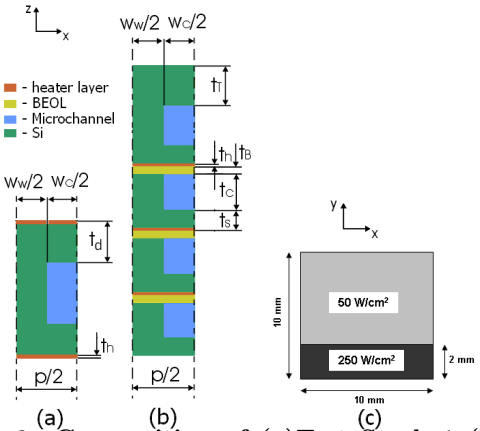
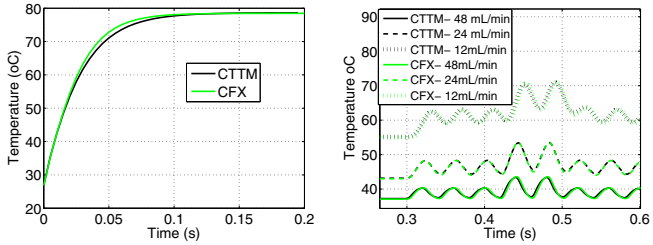


Figure 6: Composition of (a)Test Stack 1 (b)Test Stack 2; (c)Top view of the 2^{nd} die in Test Stack 2 showing the hotspot.

Table 1: Geometrical and material properties of the Test Stacks [28]

IC size	10mm × 10mm
Number of layers-**Test Stack 1**	5 (2 active dies)
Number of layers-**Test Stack 2**	15 (3 active dies)
Channel height-t_c	100μm
Channel width-w_c	50μm
Channel pitch-p	100μm
Die height- t_d	300μm
Heater height- t_h	2μm
Si slab height-t_s	50μm
Back-end of line (BEOL) height- t_B	12μm
Top Si layer height-t_T	100μm
Silicon thermal conductivity, heat capacity	130W/m · K, 702J/kg · K
BEOL thermal conductivity, heat capacity	2.25W/m · K, 517J/kg · K
Fluid thermal conductivity, heat capacity	0.6069W/m · K, 4181J/kg · K
Fluid density, viscosity	998kg/m^3, 8.89 × 10^{-4}Pa · s
Fluid flow rate per cavity	48mL/min
Fluid velocity per channel	1.62m/s



(a) (b)

Figure 7: Comparison of temperature waveforms: (a) Uniform Dissipation Case (b) Non-uniform Dissipation Case

Uniform Dissipation Case: In the first experiment, **Test Stack 1** was simulated with a *uniform* heat flux of 100W/cm^2 in both the dies, which was switched on at time t = 0s, until reaching steady state at time t = 0.1s. A discretization of 50μm × 50μm per Thermal Cell was applied for the proposed model. The resulting junction temperature plot measured at the center of the top die from both the proposed compact model and the CFX simulations are shown in Fig. 7(a). The maximum error between the two temperature waveforms at any point during the entire simulation time interval was found to be 1.6K (3% w.r.t. peak temperature change). Similar results were found for other observation points in the IC.

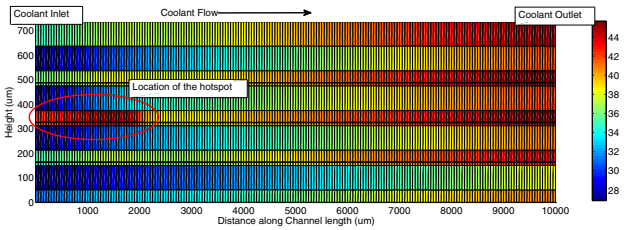
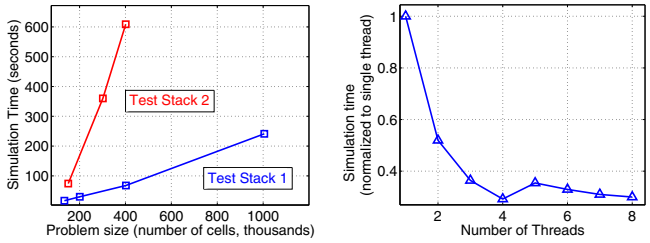


Figure 8: Temperature map of Test Stack 2 in the yz plane along the channel direction.

Non-uniform Dissipation Case: In the second experiment, **Test Stack 2** was simulated for 0.4s seconds, with a uniform background heat flux of 50W/cm^2 for all the three dies switched on for the first 0.3s. At this point, a 2mm wide strip of the heated layer on the 2^{nd} die (as shown in Fig. 6(c)) was switched alternatively between the powers 250W/cm^2 and 50W/cm^2, and between 450W/cm^2 and 50W/cm^2 at different time intervals until 0.6s to create a realistic hotspot switching activity (i.e. *non-uniform* heating of the die). A discretization of 50μm$\times$ 75μm per Thermal Cell was applied for the proposed model. The side view (zy plane) temperature map of the IC is shown in Fig. 8 with the location of the hotspot indicated. The experiment was repeated for two other flow rates- 24mL/min and 12mL/min per cavity (resulting from a pressure drop of 5×10^4Pa and 2.5×10^4Pa respectively) to demonstrate the proposed model's ability to handle different flow rates. The temperature waveforms from the proposed model and the CFX simulations for all the three cases are shown in Fig. 7(b) (as can be seen, the average temperature of the chip increases with decreasing flow rate). For all the simulations, the maximum difference between the temperatures obtained from CTTM and CFX was found to be 1.5K (3.4% w.r.t. peak temperature change). Similar results were obtained for other observations points.

6.2 Performance evaluation of 3D-ICE

Simulation speed is indeed a critical aspect for such kind of tools, since it can support or discourage its adoption by chip designers. For example, one of the possible exploitations of the thermal library can be the HW-SW cosimulation. The thermal library could be interfaced with other tools for system emulation or simulation. The resulting overall framework can be used to both system design space explorations and early thermal-aware software development. Clearly, very long simulation time makes the considered scenarios impractical and impossible to handle.



(a) (b)

Figure 9: (a)Simulation times for Test Stacks, (b) Parallelization of the proposed thermal model.

Table 2: Comparison of simulation times for Test Stack 1 and Test Stack 2 using CTTM and CFX

Problem	Discretization Size	CFX (min:sec)	CTTM (milli sec)	Speed-up
Test Stack 1	50μm × 50μm	04:10	16	260.42
Test Stack 2	50μm × 75μm	146:07	150	974.11

First we evaluated the speed-ups obtained using our thermal library implementing the proposed CTTM with respect to the CFD simulation tool. For this purpose, we simulated the same reduced size problems for **Test Stack 1** and **Test Stack 2** on the thermal library, as was simulated using the CFD model (as described earlier). These simulations were run on an Intel Core2Duo 2 GHz machine with 3 GB RAM. The recorded simulation times and the speed-ups are tabulated in Table 2. As can be seen, the proposed model is up to about 1000 times faster than the CFD model.

Next, to obtain CPU times for realistic 3D ICs, the complete **Test Stack 1** and **Test Stack 2** were simulated using the thermal library for different discretization sizes (25μm to 200μm along the channel direction). These simulations were run on Intel(R) Core(TM) i7 920 2.67 GHz processor with 8 cores and 6 GB RAM. The recorded simulation times are plotted against the problem sizes in each case in Fig. 9(b). It can be seen from this plot that the simulation times for even large problems is contained and scales approximately linearly with the problem size. Also, it can be seen that the slope of this scaling increases as the number of layers in the stack is increased (**Test Stack 2**).

In order to observe the performance of the proposed CTTM with respect to realistic 3D ICs, the 4-die stack studied in [29] was simulated using our thermal library and the thermal map of each layer was plotted as shown in Fig. 10. This structure resulted in a thermal grid consisting of 167k nodes and the thermal library took 221 seconds (approximately 3.5 minutes) to simulate it for a time interval of 70ms with a step size of 1ms (a total of 70 time points).

At this point, it is important to highlight certain crucial differences between the nature and performance of proposed

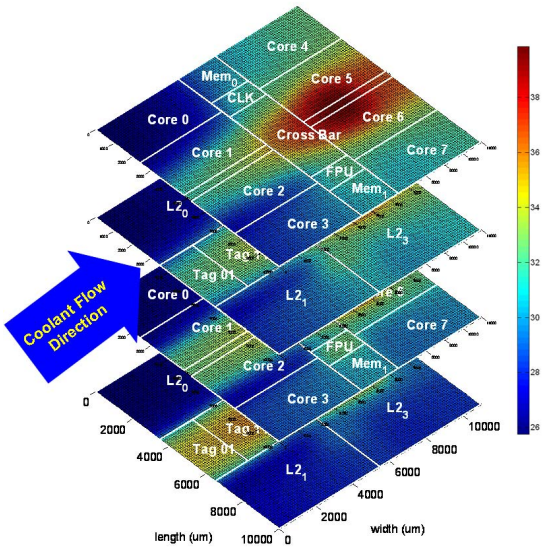


Figure 10: Temperature maps of the dies in the 3D stack described in [29].

model and that of the thermal model developed in [15]. The proposed model is capable of performing both transient and steady-state analysis (for steady-state, simply open circuit all the thermal capacitances and solve the resistive mesh directly), while the model in [15] pertains only to steady-state thermal analysis. Also, for every thermal cell in the fluid domain, the proposed model requires only one node for the simulation while the model in [15] generates 4 nodes. For example, in the **Test Stack 2** considered above, the 3D-ICE model contained 33k cells, and hence 33k nodes, in the fluid domain. For the same structure with the same discretization, the model developed by [15] would have resulted in 132k nodes in the fluid domain resulting in a 60% increase in the total problem size. These additional nodes contain thermal information of little value to an electronic designer and considerably increase the computational load. Finally, the model in [15] requires numerical presimulation of the structure to compute the thermal wake functions every time flow conditions or channel dimensions are changed. It can be seen from [15] that this numerical presimulation is orders of magnitude more computationally expensive than the actual thermal simulation. The 3D-ICE model removes this need for numerical presimulation and allows for the incorporation of any reliable correlation-based fluid heat transfer coefficient of the designer's choosing, drastically improving the CPU performance of the simulator. Albeit, the option of using a numerical presimulation also exists in 3D-ICE.

6.3 Parallelization of the thermal library

To explore the potential of parallelization of the proposed CTTM, we included a new multi-thread version of the SuperLU solver [30] in the thermal library that we developed and used it to simulate a 3mm × 10mm version of the **Test Stack 2** on Intel(R) Core(TM) i7 920 2.67 GHz processor with 8 cores and 6 GB RAM. The resulting normalized CPU times against the number of threads are shown in Fig. 9(b). This shows the potential for parallelization of the proposed modeling technique. It must be noted that the Super-LU method is a direct method of solving simultaneous equations and hence, parallelization is inherently limited. More significant gains in CPU times could be obtained using iterative techniques and employing other parallel computing platforms such as GPUs. Exploring these techniques could be a possible candidate for future research in this direction.

7. CONCLUSIONS

In this paper, we have proposed 3D-ICE, a compact transient thermal model (CTTM) for fast thermal simulation of 3D ICs with inter-tier microchannel cooling. The model is transient and can accurately predict the temporal evolution of chip temperatures with changing operational system paratmeters due to dynamic thermal management. A software thermal library was developed based on 3D-ICE. This thermal library can be easily interfaced with other tools for system emulation. The accuracy and speed of the 3D-ICE model has been validated against a commercial CFD simulation tool for various realistic 3D test chips with inter-tier cooling. The proposed model shows a maximum temperature error of only 3.4% and speed-ups up to 975x with respect to the CFD simulation tool, enabling a very fast early-stage thermal design exploration of 3D ICs with inter-tier cooling.

8. REFERENCES

[1] "International technology roadmap for semiconductors (ITRS)," *2009 Edition- ERD*, http://www.itrs.net/Links/2009ITRS/Home2009.htm.

[2] F. Li *et al.*, "Design and management of 3D chip multiprocessors using network-in-memory," in *Proc. ISCA*, pp. 130–141, 2006.

[3] T. Brunschwiler *et al.*, "Forced convective interlayer cooling in vertically integrated packages," in *Proc. ITHERM*, 2008.

[4] T. Brunschwiler *et al.*, "Interlayer cooling potential in vertical integrated packages," *Microsystem Technologies: MNSISPS*, vol. 15, no. 1, pp. 57–74, 2009.

[5] W. Huang *et al.*, "HotSpot: A compact thermal modeling methodology for early-stage VLSI design," *IEEE Trans. VLSI Sys.*, vol. 14, pp. 501–513, May 2006.

[6] T. Wang and C. Chen, "3-D thermal-ADI: a linear-time chip level transient thermal simulator," *IEEE Trans. CAD for ICs*, vol. 21, pp. 1434–1445, December 2002.

[7] S. Im and K. Banerjee, "Full-chip thermal analysis of planar (2D) and vertically integrated (3D) high performance ICs," in *IEDM Technical Digest*, pp. 727–730, 2000.

[8] P. Li *et al.*, "Efficient full-chip thermal modeling and analysis," in *Proc. ICCAD*, pp. 319–326, 2004.

[9] Y. Cheng *et al.*, "ILLIADS-T: An electrothermal timing simulator for temperature-sensitive reliability diagnosis of CMOS VLSI chips," *IEEE Trans. CAD for ICs*, vol. 17, pp. 668–681, August 1998.

[10] D. Tuckerman and R. Pease, "High-performance heat sinking for VLSI," *IEEE ED Letters*, vol. 2, no. 5, pp. 126–129, 1981.

[11] W. Qu and I. Mudawar, "Thermal design methodology for high-heat flux single-phase and two-phase microchannel heat sinks," *IEEE Trans. CPT*, vol. 26, pp. 598–609, 2003.

[12] X. Wei and Y. Joshi, "Optimization study of stacked micro-channel heat sinks for micro-electronics cooling," *IEEE Trans. CPT*, vol. 26, no. 1, pp. 55–61, 2003.

[13] N. Lei *et al.*, "Modeling and optimization of multilayer minichannel heat sinks in single-phase flow," in *Proc. IEEE InterPACK*, 2007.

[14] J. Koo *et al.*, "Integrated microchannel cooling for three-dimensional electronic circuit architectures," *ASME Journal HT*, vol. 127, pp. 49–58, 2005.

[15] H. Mizunuma *et al.*, "Thermal modeling for 3D-ICs with integrated microchannel cooling," in *Proc. ICCAD*, pp. 256–263, November 2009.

[16] J. Lienhard-IV and J. Lienhard-V, *A heat transfer textbook*. Cambridge, Massachusetts: Phlogiston Press, 2006.

[17] F. Incropera *et al.*, *Fundamentals of heat and mass transfer*. New York: John Wiley and Sons, 2007.

[18] M. Ozisik, *Finite Difference Methods in Heat Transfer*. CRC Press, 1994.

[19] A. Cheng and D. T. Cheng, "Heritage and early history of the boundary element method," *Engg. anal. with boundary elements*, vol. 29, pp. 268–302, 2005.

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

[21] A. Sridhar *et al.*, 3D-ICe , http://esl.epfl.ch/3D-ICE .

[22] http://www.gnu.org/software/bison/.

[23] http://flex.sourceforge.net/.

[24] I. S. Duff *et al.*, "Sparse matrix test problems," *ACM Trans. MS*, vol. 15, no. 1, pp. 1–14, 1989.

[25] T. A. Davis and E. P. Natarajan, "Algorithm 8xx: KLU, a direct sparse solver for circuit simulation problems," *ACM Trans. MS*, vol. 5, no. 1, pp. 1–14.

[26] J. W. Demmel *et al.*, "A supernodal approach to sparse partial pivoting," *SIAM Journal MAA*, vol. 20, pp. 720–755, 1999.

[27] http://www.ansys.com/products/fluid-dynamics/cfx/.

[28] T. Brunschwiler *et .al.*, "Heat-removal Performance scaling of Interlayer Cooled Chip Stacks," *Proc. ITHERM 2010*, pp. 1-12, June 2010.

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

[30] J. W. Demmel *et al.*, "An asynchronous parallel supernodal algorithm for sparse gaussian elimination," *SIAM Journal MAA*, vol. 20, no. 4, pp. 915–952, 1999.