Contents lists available at ScienceDirect

INTEGRATION, the VLSI journal

journal homepage: www.elsevier.com/locate/vlsi

Invited paper

Thermal modeling and analysis of 3D multi-processor chips

José L. Ayala [a,b,*], Arvind Sridhar [b], David Cuesta [a]

[a] DACYA - Complutense University of Madrid, Spain
[b] ESL - EPFL, Switzerland

ARTICLE INFO

Keywords:
Temperature
Thermal model
Floorplanning
3D
Nano-grid

ABSTRACT

As 3D chip multi-processors (3D-CMPs) become the main trend in processor development, various thermal management strategies have been recently proposed to optimize system performance while controlling the temperature of the system to stay below a threshold. These thermal-aware policies require the envision of high-level models that capture the complex thermal behavior of (nano)structures that build the 3D stack. Moreover, the floorplanning of the chip strongly determines the thermal profile of the system and a quick exploration of the design space is required to minimize the damage of the thermal effects.

This paper proposes a complete thermal model for 3D-CMPs with building nano-structures. The proposed thermal model is then used to characterize the thermal behavior of the Niagara system and expose the strong influence of the chip floorplanning in the thermal profile.

1. Introduction

The traditional chip fabrication technology in 2D is facing lots of challenges in utilizing the exponentially growing number of transistors on a chip. The wire delay and power consumption are increasing dramatically and achieving interconnect design closure is becoming a challenge. Vertical stacking of multiple silicon layers, referred to as 3D stacking, is emerging as an attractive solution to continue the pace of growth of Systems on Chips (SoCs). The 3D technology results in smaller footprint in each layer and shorter vertical wires that are implemented using Through Silicon Vias (TSVs) across the layers.

Despite the advantages of 3D ICs over 2D ICs, thermal effects are expected to be significantly exacerbated in 3D ICs due to higher power density and greater thermal resistance of the insulating dielectric, and this can cause greater degradation in device performance and chip reliability which have already plagued 2D ICs. Thus, it is essential to develop 3D-specific design tools that take a thermal co-design approach, so as to address the thermal effects and generate reliable and high performance designs. This work considers the design of a nano-structure to help on the thermal dissipation in 3D chips.

In 3D ICs, devices are fabricated on a number of active layers, which are separated by silicon dioxide and joined by an adhesive material. Within each device layer, interconnections among devices can be achieved with traditional interconnect wires and vias. Connections between active layers are facilitated by vertical interconnect vias that span through multiple layers, providing a means for electrically connecting wires in those layers. This type of via is different from a regular 2D via: in particular, it is significantly taller than conventional vias, and has a larger landing pad to maintain a viable aspect ratio. We refer to such vias as TSVs.

Due to the ultrashort lengths of the TSVs ($50-100\,\mu$m), they easily overcome RC delays of long, horizontal circuit traces in conventional 2D circuits, and they also provide a higher density of connections. The TSVs are good conductors of heat, and hence they can be effective in dissipating some of the temperature of the devices.

The design and placement of TSVs can be proposed as an effective mechanism for thermal dissipation in 3D chips. The thermal resistivity of the via diffuses the heat and can create a homogeneous thermal distribution in the stack if placed carefully. The aim of this work is to propose a nano-structure, built as a grid of TSVs, for the thermal dissipation and optimization in 3D stacks. The capability of selecting a higher density grid of TSVs in specific areas of the system will enable to cool down selectively those zones with a bigger power density.

The experimental work of this paper is carried out through a novel thermal analysis of a real 5-tier 3D stack (see Fig. 1). Then, the material layers and TSVs are modeled mathematically, and the effect of a non-homogeneous distribution of the vias for thermal control is analyzed and effective inclusion of localized TSVs conforming a grid of nano-structures for thermal control is proposed. Also, the effect of specific interface materials used as inter-layer glue is considered. These interfaces will expose unique characteristics due to the presence of aluminum dopants.

* Corresponding author at: DACYA - Complutense University of Madrid, Spain.
E-mail addresses: jayala@fdi.ucm.es (J.L. Ayala), arvind.sridhar@epfl.ch (A. Sridhar), dcuesta@fdi.ucm.es (D. Cuesta).

0167-9260/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.vlsi.2010.06.002

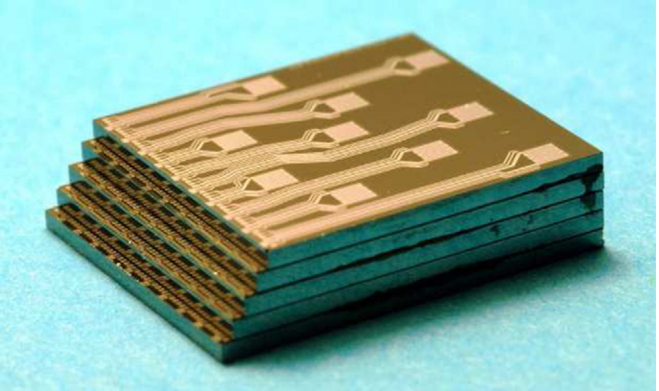

Fig. 1. View of the 3D chip.

The thermal model is also validated against the on-silicon implementation shown in Fig. 1, where the extensive measurements of several existing heaters and sensors per layer can be used to study the horizontal and vertical heat diffusion. The obtained results show interesting conclusions in the area of thermal modeling and optimization for 3D chips, as well as bringing new opportunities in the design of nano-structures based on TSVs for thermal balancing and control. Finally, the proposed thermal model is used to analyze the thermal profile of several 3D stacks based on the Niagara multi-processor. This analysis exposes several interesting conclusions in the area of chip floorplanning that can help on optimizing the thermal response of the system.

The structure of this chapter is as follows: next section presents the theory of numerical solvers, which will be used along the rest of the chapter. Section 2 reviews the previous works on the area of thermal modeling, Section 3 presents the configuration of the 3D stack developed for the experimental work, and the developed thermal model is explained in Section 4. Then, the validation of the model and the rest of the experimental work are covered in Sections 5 and 6, respectively. Section 7 presents the results of the thermal analysis of several multi-processor configurations, and the defined working scenarios are shown in Section 8. Finally, the conclusions of the work are drawn.

2. Related work

Three-dimensional (3D) integration consists of the vertical placement and interconnections of several layers of active circuits. The main interests of this technology are to reduce global interconnect lengths, to increase circuit functionality and to enable new 3D circuit architectures. They also facilitate the integration of heterogeneous materials, devices, and signals. However, the processes required to build circuits with multiple layers of active devices must be compatible with current state-of-the-art silicon processing technology [1–3].

Recently 3D die stacking is drawing a great deal of attention, primarily in embedded processor systems. Some previous works analyze the application of 3D stacks in system-on-chip designs: [4] compare wire length distributions, obtained for 2D and 2.5D implementations of benchmark circuits, [5] reviews the available fabrication technologies and testing solutions for this new integration technology and proposes a layout synthesis framework. Kgil et al. [6] show that a straightforward use of 3D stacking technology enables the design of compact energy-efficient servers, [7] proposes the addition of specialized analysis hardware built on separate active layers stacked vertically on the processor die using 3D IC technology. This provides a modular “snap-on” functionality that could be included with developer systems, and omitted from consumer systems to keep the cost impact to a minimum, that will help on the profiling and introspection of the system. Rahman and Reif [8] evaluate system-level performance metrics in 3D integrated circuits. Other works explore cache implementations: [9,10] (where the authors present a delay and energy model, 3DCacti, to explore different 3D design options of partitioning a cache) and [11]; design of 3D arithmetic circuits: [12] (it is focused on a 3D microprocessor test vehicle and demonstrates the speed advantages derived from the 3D integration) and [13] (they show how a barrel shifter implemented in 3D exhibits a 9% reduction in latency with a simultaneous 8% reduction in energy). Finally, other works evaluate wire benefits in full microprocessors: [14] (where a 3D structure of is examined and applied to a real x86 deeply pipelined high performance microprocessor. This paper shows that a 3D implementation can potentially improve the performance by 15% while improving power by 15%), [4,5,15] (where it is also shown that a 3D floorplan of a high performance microprocessor can simultaneously reduce power 15% and increase performance 15% with a 14 °C increase in peak temperature), and [16] (where the implementation of various microprocessor components using 3D technology is presented). This emerging 3D technology is considered as a very attractive method for integrating complex systems by the embedded industry. Furthermore, existing 3D products from Samsung [17] and Tezzaron [18] corporations demonstrate that the silicon processing and assembly of structures in 3D stacks are feasible in large scale industrial productions.

In the literature, the “1D” approximation is often assumed to evaluate the thermal behavior of 3D integration [2,19–21]. This means that the power is uniformly produced on “active levels” (or on part of it), one per stratum. This assumption may lead to strongly underestimated maximum temperature. Some authors [22] use this simplification but perform detailed simulation of 3D thermal effects due to the presence and localization of supervias. Other works [23] analyze the local (3D) and global (1D) modeling contribution to the maximum temperature, showing that thermal resistance can be higher than 1D thermal resistance due to local 3D effects.

Numerical thermal simulations have been carried out to convert power dissipation distribution into a temperature distribution in a 3D IC [24]. Based on the past work, the development of a fundamental analytical model for heat transport in 3D integrated circuits is highly desirable. Such an analytical model will provide a framework in which to analyze the general problem of heat dissipation in 3D ICs, and will offer simple thermal design guidelines.

A key component of 3D technology is a through-silicon via (TSV) that enables communication between the two dies as well as with the package. Some work has been reported on optimizing the problem of placement of vias for heat dissipation in 3D ICs [22,25]. Other works [26] propose analytical and finite-element models of heat transfer in 3D electronic circuits and use this model to analyze the impact of various geometric parameters and thermophysical properties (through silicon vias, inter-die bonding layers, etc.) on thermal performance of a 3D IC.

This is the first time that a nano-structure of TSVs is proposed on purpose as an effective way to optimize the thermal profile in 3D stacks. The closest work to our proposal is [27], where the authors analyze the impact of thermal through silicon vias (TTVs) in vertically integrated die-stacked devices. However, while the work presented in [27] performs a theoretical analysis, our approach proposes an accurate thermal modeling of the through-silicon vias and it is validated against measurements collected in a real chip. Finally, the thermal effect of the nano-structure of the TSVs will be examined.

3. Configuration of the 3D stack

The 3D chip manufactured for our experimental set-up is created as a multi-level chip, built by stacking silicon layers and fixed with an interface glue. In this configuration, we can find five silicon layers (Die 1–Die 5), the epoxy-based interface glue, and a bottom PCB layer (see Fig. 2). Each stack has an area of 1 cm^2.

This 3D stack resembles the thermal effects that can be found in a 3D multi-processor systems on chip by the use of heaters that create the power dissipation. As the power dissipated in a chip is not uniform on its surface (microprocessors can dissipate between 200 and 300 W/cm^2 while memories only dissipate about 10 W/cm^2) each layer contains several microheaters located at different points to simulate the heat dissipated by the integrated components.

These microheaters are built as a serpentine wire created with thin-film technologies. The material used for the heaters is Platinum, due to its capability to operate at very high temperature and its long stability.

Some thermal sensors are also placed in specific places as detector devices to monitor the temperature inside of the stack and check the heat dissipated and the heat interactions between neighboring microheaters. Platinum has also been selected as the material to build the sensors; therefore, sensors and microheaters can be manufactured at the same time in a single step of the technology process. These sensors are resistance temperature detectors (RTDs). In this way, the temperature of the heater creates a variation in the resistance of the sensor. Then, the temperature can be obtained by the observation of the voltage drop at both extremities of the sensor (with a fixed current) and applying the resistivity temperature dependence of platinum:

$$R_T = R_0(1+\alpha T+\beta T^2) \qquad (1)$$

with R_T the resistance at temperature T, R_0 the nominal resistance at 0 °C, $\alpha = 3.9083\text{e}^{-3}\,^{\circ}\text{C}^{-1}$ and $\beta = -5.77\text{e}^{-7}\,^{\circ}\text{C}^{-2}$.

Heaters and sensors are connected to the PCB by wire bonding, allowing the direct access to perform the measurements. All pads are located in one side of the chip.

Each layer comprises 10 heaters of 1 mm^2 each, very similar to the area of common processing elements. These microheaters have been designed to resemble a hot-spot on the surface of the chip of 300 W/cm^2; therefore, each heater dissipates 3 W. The heaters are aligned in three vertical lines. The five layers of the stack have the same configuration, so the alignment of the heaters appears also out of the plane.

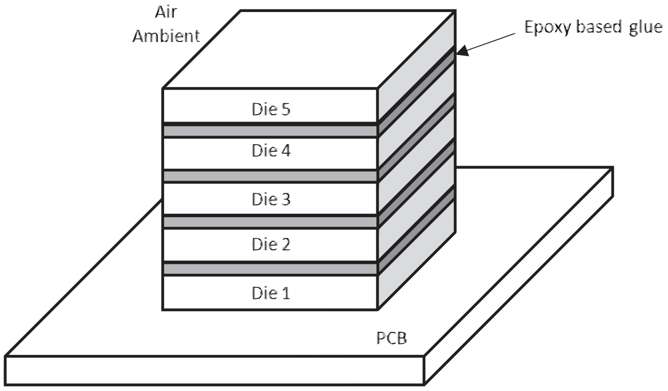

Fig. 2. The test 3D stacked structure.

In our configuration, RTDs are placed around the heaters. These sensors are designed for a value of 100 Ω and are driven with a current of 1 mA. A view of the placement of heat and sensors in each layer will of the 3D stack will be analyzed later in the text (see Fig. 7 in Section 5).

The fabrication process is schematically shown in Fig. 3 and outlined here. Starting with a 1000.5 mm in diameter and 52 525 μm double side polished wafer, a 200 nm wet oxide is placed as an insulating layer for heaters and resistance temperature detectors (RTDs). If these components were deposited directly on the silicon surface some current may flow in the semiconductor choosing non-appropriate way through the material (not following the metallic path) with an energy loss at the desired place. Other influences like schottky diode effects may appear.

For microheaters and sensors, all the design and dimensioning have been done for a 500 nm platinum evaporation on the surface. These structures are created using a lift-off process which involves a low-pressure vapor deposition (evaporation must be done instead of sputtering). To do this, a large distance between

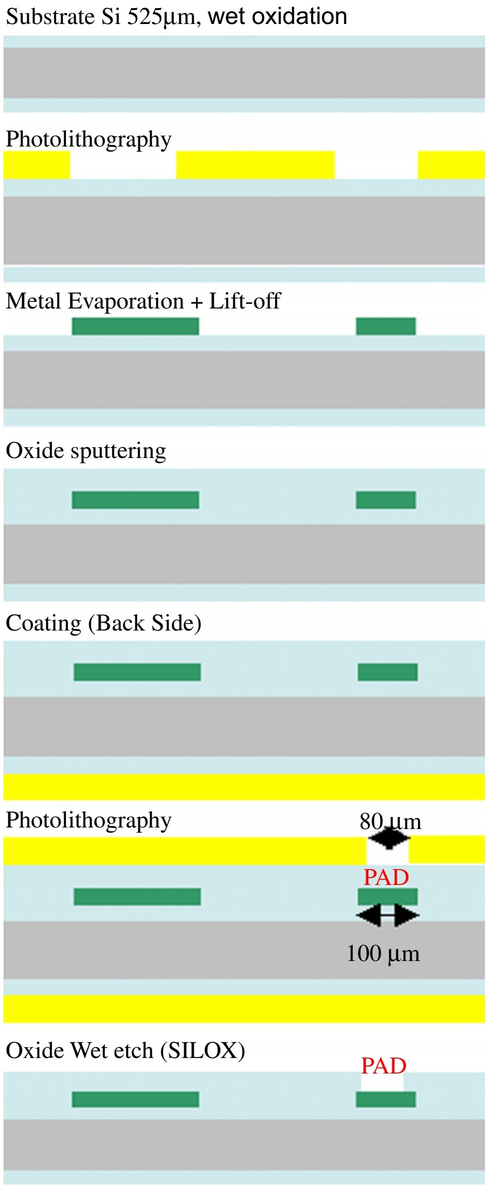

Fig. 3. Schematic view of the fabrication process.

Please cite this article as: J.L. Ayala, et al., Thermal modeling and analysis of 3D multi-processor chips, Integration VLSI J. (2010), doi:10.1016/j.vlsi.2010.06.002

source and substrates is needed for a lift-off process and then a larger amount of material is used (what is not welcomed here due to high cost of platinum). As 500 nm is quite large for platinum deposition, then for first trial aluminum has been evaporated instead of platinum to study the feasibility of the process.

Aluminum is suitable for wire bonding, if platinum would have been evaporated as it was planed at the beginning, a soft metal must be evaporated on it in the pad location. Aluminum metallization was done using titanium adhesion layer, 10 nm Ti +500 nm Al with lift-off process. Wet etching in SILOX bath with activation is done for pad etches to open contacts for heaters and sensors for wire bonding.

To build the stack, single layers have been glued on top of each other. To have a thin uniform thickness, glue has been deposited by screen printing. To do this, a semiautomatic machine (EKRA) dedicated to thick layers fabrication and suitable for gluing has been used. An epoxy resist material doped with alumina particles (Al_2O_3) has been used to avoid the creation of thermal insulating layer at the interface. Thickness deposited is around 30 μm. Stacking itself has been done manually. For alignment, layers have been leaned against a corner angle.

4. Thermal model

The test five layered 3D stack structure considered in this work is shown in Fig. 2. As seen in this figure, five silicon dies, stacked one on the top of another fixed with an interface epoxy glue, are placed on the printed circuit board (PCB). The bottom surface of the 3D stack attached to the PCB is assumed to be adiabatic; therefore, the heat will be exchanged through the vertical active and interface layers in the system.

Within each die, the aluminum resistor-based heaters are fabricated in the silicon dioxide layer on the top of the substrate, as shown in the cross-sectional view in Fig. 4. These heaters model the thermal effects of the hot-spot cores in an actual 3D multi-processor system-on-chip (3D MPSoC), where multiple cores act as thermal sources and the shared memories can behave as thermal sinks. The heat generated by these heaters flows through the body of the 3D stack, and ends at the environment interface (ambient) where it is spread through natural convection.

The heat flow inside this structure is diffusive in nature and hence, is modeled by its equivalence to an electronic RC circuit [28–30]. This is done by first dividing the entire structure into small cubical thermal cells as shown in Fig. 5. Each cell is then modeled as a node containing six resistances that represent the conduction of heat in all the six directions (top, bottom, north, south, east and west), and a capacitance that represents the heat storage inside the cell, as shown in Fig. 6. The conductance of each resistor and the capacitance of the thermal cell are calculated as follows:

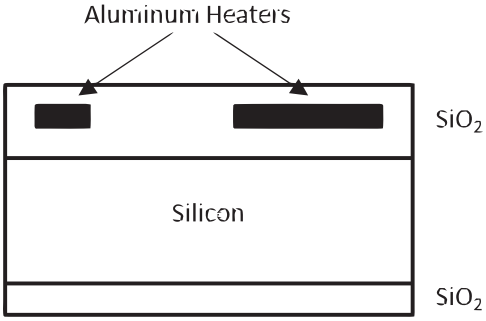

Fig. 4. Cross-sectional view of the layers in a single die.

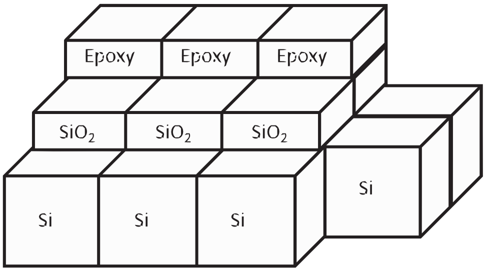

Fig. 5. The unitary thermal cells of the 3D stack.

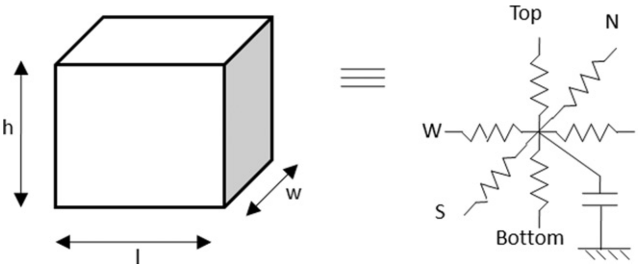

Fig. 6. Equivalent RC circuit of a single cell.

$$g_{top/bottom} = k_{th} \cdot \frac{l \cdot w}{(h/2)}, \tag{2}$$

$$g_{north/south} = k_{th} \cdot \frac{l \cdot h}{(w/2)}, \tag{3}$$

$$g_{east/west} = k_{th} \cdot \frac{w \cdot h}{(l/2)}, \tag{4}$$

$$c_{top} = sc_{th} \cdot (l.w.h). \tag{5}$$

Here, the subscripts *top*, *east*, *south*, etc., indicate the direction of conduction, k_{th} and sc_{th} are the thermal conductivity and the specific heat capacity per volume unit of the material, respectively.

Current sources, representing the sources of heat, are connected to the cells in the regions where thealuminum heaters are present. The entire circuit is grounded to the ambient temperature at the top and the side boundaries of the 3D stack through resistances, which represent the thermal resistance from the chip to the air ambient.

The behavior of the resulting RC circuit can be described using a set of first-order differential equations via nodal analysis [31] as follows:

$$\mathbf{G}\mathbf{X}(t)+\mathbf{C}\dot{\mathbf{X}}(t)=\mathbf{B}\mathbf{U}(t), \tag{6}$$

where $\mathbf{X}(t)$ is the vector of cell temperatures of the circuit at time t, $\mathbf{G}$ and $\mathbf{C}$ are the conductance and capacitance matrices of the circuit, $\mathbf{U}(t)$ is the vector of input heat (current) sources and $\mathbf{B}$ is a selection matrix. $\mathbf{G}$ and $\mathbf{C}$ present a sparse block-tridiagonal and diagonal structure, respectively, due to the characteristics and definition of the thermal problem.

In addition, $\mathbf{G}$ and $\mathbf{U}(t)$ are functions of the cell temperatures $\mathbf{X}(t)$, making the behavior of the circuit non-linear. This is because of the temperature-dependent thermal conductivity of silicon and the temperature-dependent electrical resistance of the aluminum heaters, respectively. In this work, a first-order dependence of these parameters on temperatures around 300 K is assumed. Some of these parameters are presented in Table 1 [32].

Table 1
Thermal properties of materials.

Silicon thermal conductivity	295–0.491 T W/mK
Silicon specific heat	$1.659\times10^6\,\mathrm{J/m^3\,K}$
SiO_2 thermal conductivity	1.38 W/mK
SiO_2 specific heat	$4.180\times10^6\,\mathrm{J/m^3\,K}$
Aluminum electrical resistivity	$2.82\times10^{-8}(1+0.0039\Delta T)\,\Omega\mathrm{m}$ $\Delta T=T-293.15\,\mathrm{K}$

For the validation of the thermal library, profuse temperature measurements on the 3D stack were performed with DC current inputs for the heaters. Hence, an efficient way to calculate the steady state response of the circuit described by Eq. (6) is required. The DC solution for the circuit can be found by solving the corresponding steady state equations,

$$\mathbf{GX}=\mathbf{BU}. \tag{7}$$

The above set of equations are solved by the inversion of the matrix **G** using the sparse LU decomposition method [33]. Because of the non-linearity of the circuit and the input sources, these equations were solved repeatedly, by updating the matrices after each iteration of solving, until convergence is reached. The description of this iterative algorithm is described in Pseudocode 4. In most of the test cases, 5–6 iterations were found to be sufficient to reach convergence within an error of 10^{-6}.

Pseudocode 1. Pseudocode for calculating steady state temperatures of the 3D stack.

```
1. Define:
2. $\mathbf{X}^r=$ vector of cell temperatures during the $r$th iteration,
3. $\mathbf{G}^r=$ conductance matrix during the $r$th iteration,
4. $\mathbf{U}^r=$ input vector during the $r$th iteration.
5. Set r=0. Generate an initial-guess for $\mathbf{X}^0$
6. Calculate $\mathbf{G}^0$ and $\mathbf{U}^0$ using the initial-guess $\mathbf{X}^0$
7. $\mathbf{X}^{r+1}=(\mathbf{G}^{\,r})^{-1}\mathbf{BU}^r$
8. Calculate $\mathbf{G}^{r+1}$ and $\mathbf{U}^{r+1}$ using the updated temperatures $\mathbf{X}^{r+1}$
9. If $\|\mathbf{X}^{r+1}-\mathbf{X}^r\|!$ (a predetermined error criterion), exit. Else set r=r+1 and go to step 4.
```

Further details about the mathematical background required for the development of the thermal model can be found in Appendix A.

5. Electrical measurements and validation of the model

Two different test cases were chosen for experimental measurements and validation of the thermal model: (A) lateral heat transfer measurement and (B) vertical heat transfer measurement. The two cases are described in detail in the ensuing of this section.

5.1. Lateral heat transfer measurement (layer 2).

In this test case, the lateral heat diffusion in a given layer of the 3D stack is characterized. For this, the heater in device D02 in Die 2 (as described in Section 3) is excited with different current levels. For each current level temperature several measurements are made at the sensors in devices D02, D04, D07, D09 and D10 within the same die. This test case is illustrated in Fig. 7. These measurements provide information on the behavior of the lateral temperature distribution as function of the distance of the sensor from the heat source within a layer.

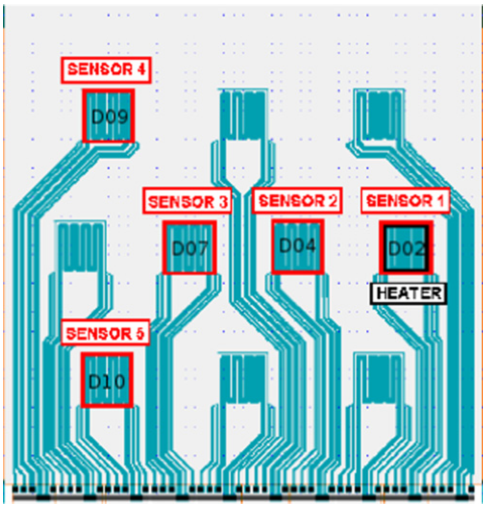

Fig. 7. Lateral heat transfer measurement in layer 2 (test case A).

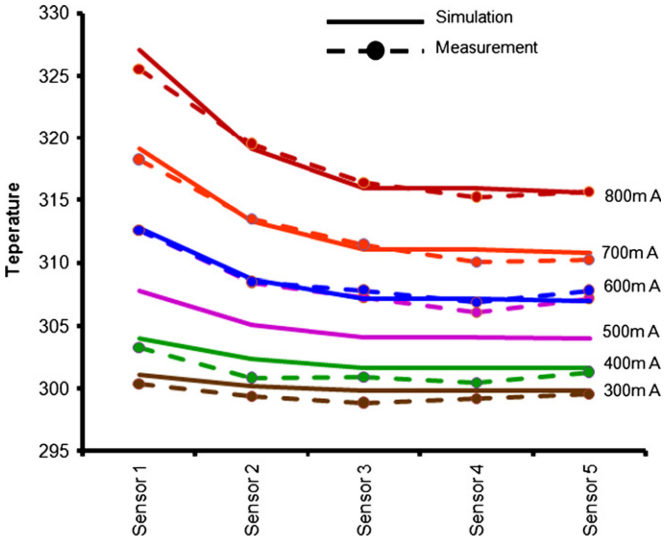

Fig. 8. Simulation and measurement results for lateral heat transfer (test case A).

Table 2
Average error for lateral heat transfer measurement (test case A).

Sensor	Average error between simulation and measurement (%)
1	0.669
2	0.823
3	0.240
4	0.426
5	0.314

Fig. 8 shows the comparative results between the measurements and the simulation for various heater current levels in the lateral heat transfer test case. The solid lines here indicate the simulation results and dashed lines indicate measurement results from the different sensors, respectively. Heater current levels from 300 mA (1.25 W/mm^2) to 800 mA (9 W/mm^2) were used in the experiments. As can be seen from this figure, the simulation results from the thermal model accurately predict the experimental results for all sensors. The percentage of average errors between measurement and simulation results for each sensor are tabulated in Table 2, being these errors $<1\%$ for all the sensor-heater configurations.

Please cite this article as: J.L. Ayala, et al., Thermal modeling and analysis of 3D multi-processor chips, Integration VLSI J. (2010), doi:10.1016/j.vlsi.2010.06.002

5.2. Vertical heat transfer measurement (device D02).

In this test case, the vertical heat flow from one layer to another in the 3D stack is characterized. For this, the heater in device D02 in layer 2 is again excited with different current levels. For each current level, temperature measurements are made at the sensors in devices D02 of Die 2, Die 3, Die 4 and Die 5. This test case is illustrated in Fig. 9. These measurements provide information about the behavior of the temperature distribution as function of the vertical distance of the sensor from the heat source in different layers of the 3D stack.

Fig. 10 shows the comparative results for the vertical heat transfer between measurements and the simulation, obtained for the same heater current levels as in test case A. Again, as in test case A, the solid lines indicate the simulation results and dashed lines indicate the measurement results for the different sensors, respectively. As can be seen in this figure, the simulation results from the thermal model accurately predicts the experimental results for all sensors. The percentage of average errors between measurement and simulation results for each sensor are tabulated in Table 3, being in this case always <2%.

6. Thermal TSVs and thermal grids

During the last years, many fabrication-based solutions for the thermal management in 3D integrated circuits have been proposed [34]. Thermal through silicon vias (TTSVs) have a prominent place among these solutions. Many times, it is more desirable to reduce the difference in the temperatures between various parts of the IC, rather than the reduction of the absolute temperature of the chip. This is because variations in operating temperatures affects performance of different parts of the IC (e.g., processor and memory) differently, leading to timing errors and chip failures. Moreover, thermal gradients have been observed as a determinant negative factor on system reliability.

To overcome the above mentioned challenges and to simulate the effects of on-chip metallizations on the thermal behavior of the 3D stack, thermal through silicon vias and thermal grids were introduced in the thermal model developed in the previous section. For the ensuing experiments, a 3-layered 3D stack was used instead of the 5-layered stack. Fig. 11 shows two test cases—(a) with two hot-spot cores in the same die of the 3D stack and (b) with three cores, one on the top of another, communicating each other through different layers (from the performance-enhancement perspective, it is desirable to place the most frequently communicating cores of a 3D IC one on the top of the other to reduce communication delay).

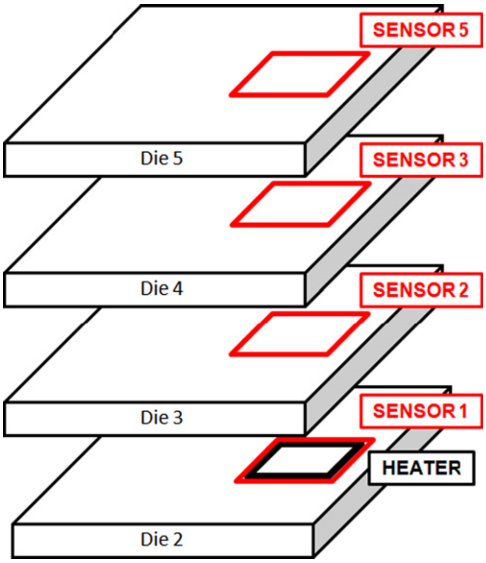

Fig. 9. Vertical heat transfer measurement in Device D02 (test case B).

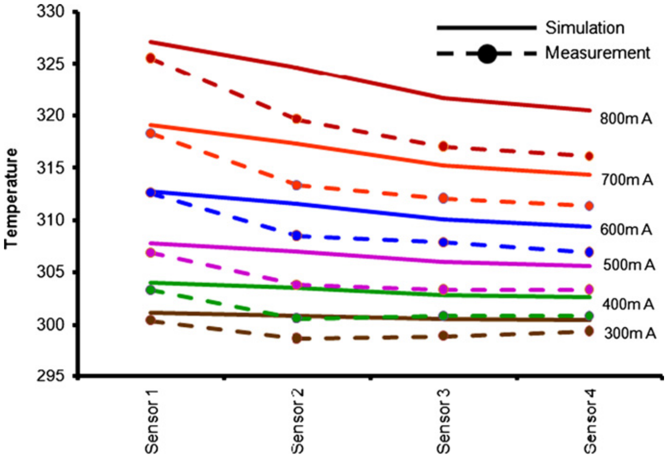

Fig. 10. Simulation and measurement results for vertical heat transfer (test case B).

Table 3
Average error for vertical heat transfer measurement (test case B).

Sensor	Average error between simulation and measurement (%)
1	0.669
2	1.344
3	1.556
4	1.783

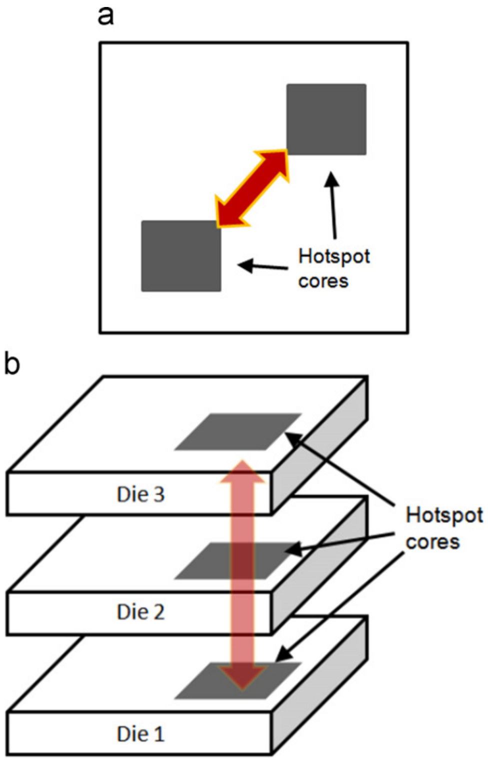

Fig. 11. Communication between active cores in a 3DIC: (a) within one layer and (b) between different layers.

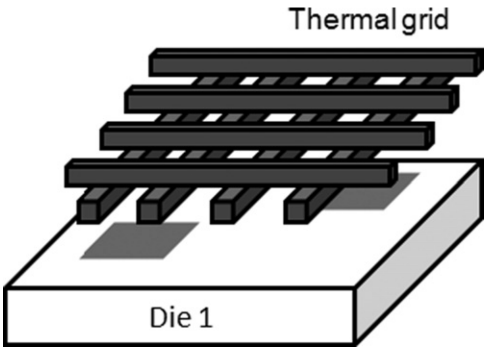

Fig. 12. Thermal grid for reducing temperature variation within a single layer.

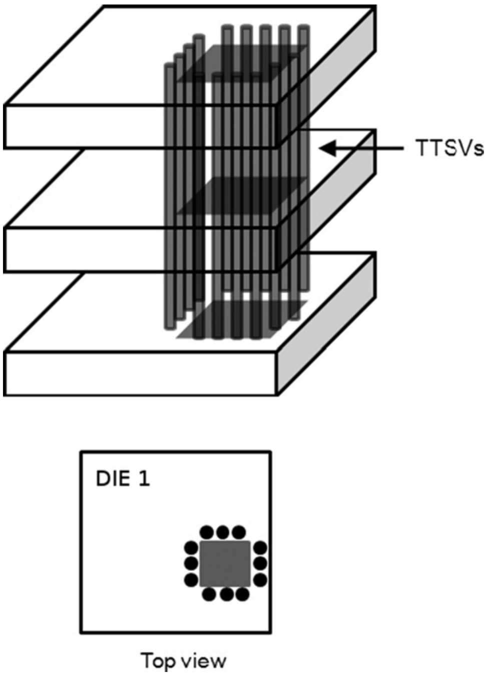

Fig. 13. TTSVs for reducing temperature variations along the different layers in a 3D IC.

For case (a), to reduce the temperature variations within the same layer, thermal grid networks-dedicated metallizations as well as existing metallizations for the electronic design, are proposed. These thermal grid networks lower the effective thermal conductivity of the dielectric material within the layer and hence, reduce the temperature variations in the layer. This is illustrated in Fig. 12, where the schematic configuration of the horizontal grid is shown.

For case (b), to address the temperature variations between different layers in regions where the communicating cores exist, TTSVs are placed around the active cores as shown in Fig. 13. This placement of TTSVs, in addition with the metallizations that naturally exist between the cores meant for electronic routing, reduces the effective thermal conductivity of this region. This, in turn, brings the temperature of different parts of this region closer to each other because of the favored thermal flow.

To incorporate both the thermal grid and the TTSVs in the thermal model, effective thermal conductivity was calculated for the cells in the region containing these metallizations, using the following relation:

$$k_{eff} = k_{cu}\omega + k_{th}(1-\omega), \qquad (8)$$

where, k_{cu} is the thermal conductivity of copper (the metal used for all metallizations in the IC), kth is the thermal conductivity of the surrounding material and ω is the wiring/via density in the region. In the case of TTSVs, a slight modification was made for the effective thermal conductance in the lateral direction. This parameter was calculated by computing the equivalent thermal resistance of the cells depending upon the path of heat flow while traversing it along north–south and east–west direction (a series/parallel combination of vias and surrounding material). Hence, anisotropic cells were created in order to more accurately capture the effects of TTSVs.

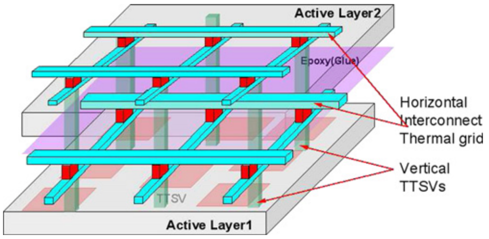

Fig. 14. Vertical and horizontal thermal grid.

Fig. 14 shows the devised nano-grid of horizontal interconnects and vertical TTSVs. The TTSVs that integrate the nanostructure improve the overall thermal conductivity of the active layer, provided the thermal coupling is good between the TTSVs. To improve the thermal coupling, these TTSVs must be placed as close as possible to each other but electrically isolated. On the other hand, the horizontal grid helps to spread the temperature along the die and also improve the thermal conductivity of the inter-layer material.

Two experiments were performed to measure the performance of these two strategies. In the first experiment, four heaters (in devices D02, D04, D07 and D10) in Die 1 of the 3-layered 3D stack were excited, each with a current of 300 mA (1.25 W/mm^2). First, this experimental set-up was simulated without any thermal grid. Next, thermal grid was added to Die 1 (with 50% wiring density) in the same experimental set-up and the resulting model was simulated again. The temperature distribution profile was drawn for each case. These histograms are shown in Fig. 15. As can be seen from this figure, the temperature spread within this layer has been reduced by the effect of the thermal grid, that easies the diffusion of the extra heat.

In the next experiment, the same set-up was used. TTSVs were laid around each of the active heaters in Die 1 as shown in Fig. 13. The resulting thermal circuit was then simulated, once without the TTSVs and then once with the TTSVs. Temperatures in the region covered by the TTSVs of one of the heaters (the region enclosed by the TTSVs encompassing all the three dies as shown in Fig. 13) were recorded in each case. The corresponding temperature distribution profiles for one such active heater regions are shown in Fig. 16. We find that the temperature spread was considerably reduced in this region along the vertical direction. Therefore, the grid of TTSVs can be considered as an effective mechanism to optimize the thermal profile in 3D stacks, both in the vertical and lateral directions.

7. Practical thermal analysis: experimental work

This section uses the proposed thermal model to analyze the thermal behavior of several system configurations. First, some layouts based in the Niagara architecture will be presented. Then, this configuration will be modified and replicated several times to increase the number of integrated cores. This experimental work

Please cite this article as: J.L. Ayala, et al., Thermal modeling and analysis of 3D multi-processor chips, Integration VLSI J. (2010), doi:10.1016/j.vlsi.2010.06.002

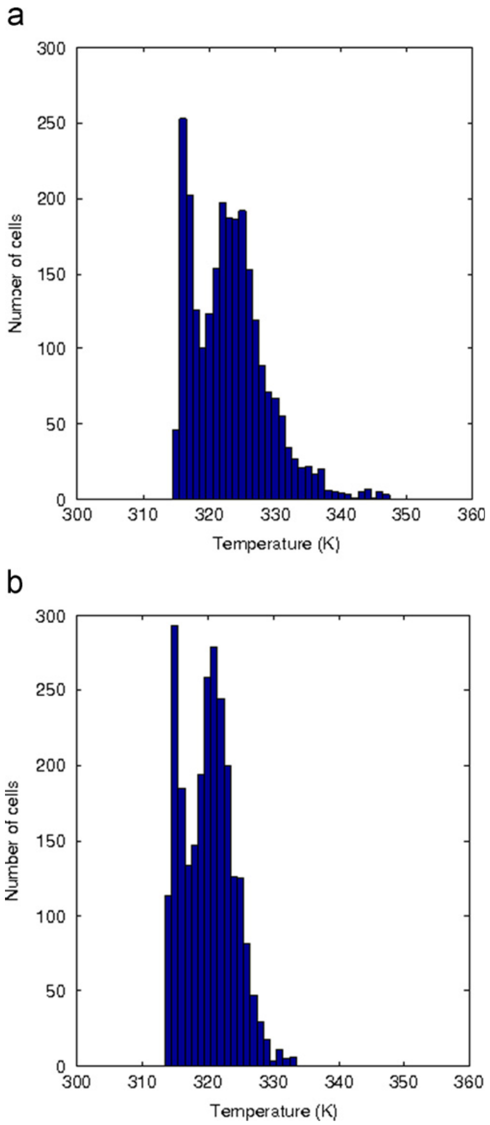

Fig. 15. Lateral temperature distribution profile for Die 1: (a) without thermal grid and (b) with thermal grid.

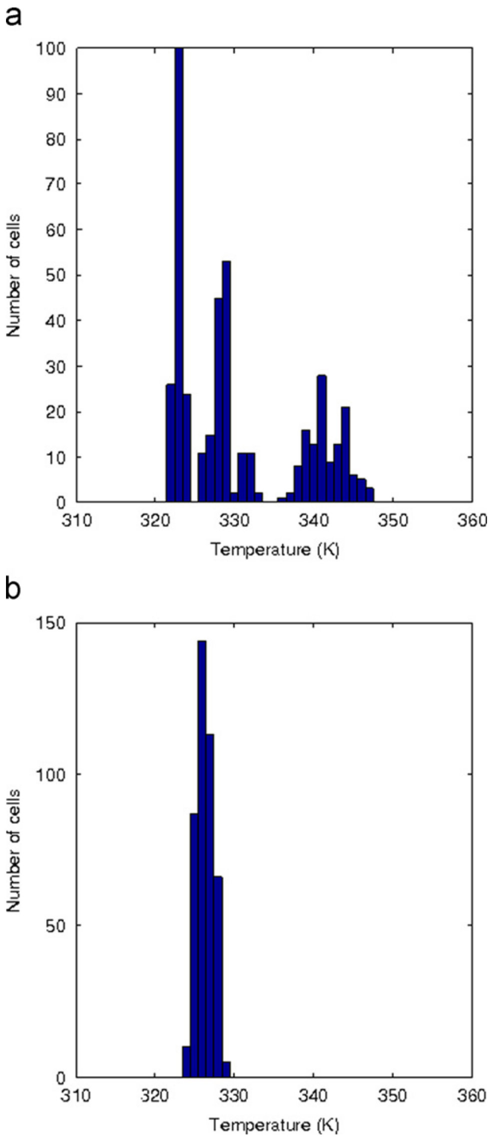

Fig. 16. Vertical temperature distribution profile for region around D06: (a) without TTSVs and (b) with TTSVs.

will allow to extract some thermal rules that must be satisfied when an optimal thermal behavior of the chip is required.

7.1. Layout configuration

The experiments made in this work are based on the Niagara architecture, fabricated in 90 nm technology. The system is composed of eight multithreaded UltraSparc T1 cores [35], and their two level cache memories. In the center of the layout there is a crossbar that allows the cores to communicate, as well as the level two shared memories placed at the corner of the active area.

To increase the number of cores up to 64, we replicated and modified the original distribution of the functional units, creating seven different 2D layers (as shown in Fig. 17). These 2D layers will be used as patterns to create our 3D system, stacking them vertically. These patterns consider just a limited number of the functional units available in the original Niagara layout. Therefore, only memories, cores, will be considered for the design of the layouts.

The patterns have been designed with the purpose to create different power distributions in the chip, and the thermal behavior analysis can be addressed.

Each 3D configuration will be designated by a number from the conversion Table 4. The pattern sequence field specifies which patterns compose the stack, starting with the bottom layer, and ending with the top one.

7.2. Thermal analysis

The experiments have been run for three different configurations, attending to the number of cores in the system, starting with eight cores and following the current tendency in the design of embedded systems of increasing this parameter. These experiments also keep the tier area constant.

Therefore, the first set of experiments takes a standard Niagara system with eight cores to extract the thermal characteristics of 2D and simple 3D systems.

Then, the second set of our experiments designs a 16-core architecture by replicating the Niagara configuration. This set-up allows to observe some thermal phenomena that only happens when the middle layers come into play in 3D stacks.

Please cite this article as: J.L. Ayala, et al., Thermal modeling and analysis of 3D multi-processor chips, Integration VLSI J. (2010), doi:10.1016/j.vlsi.2010.06.002

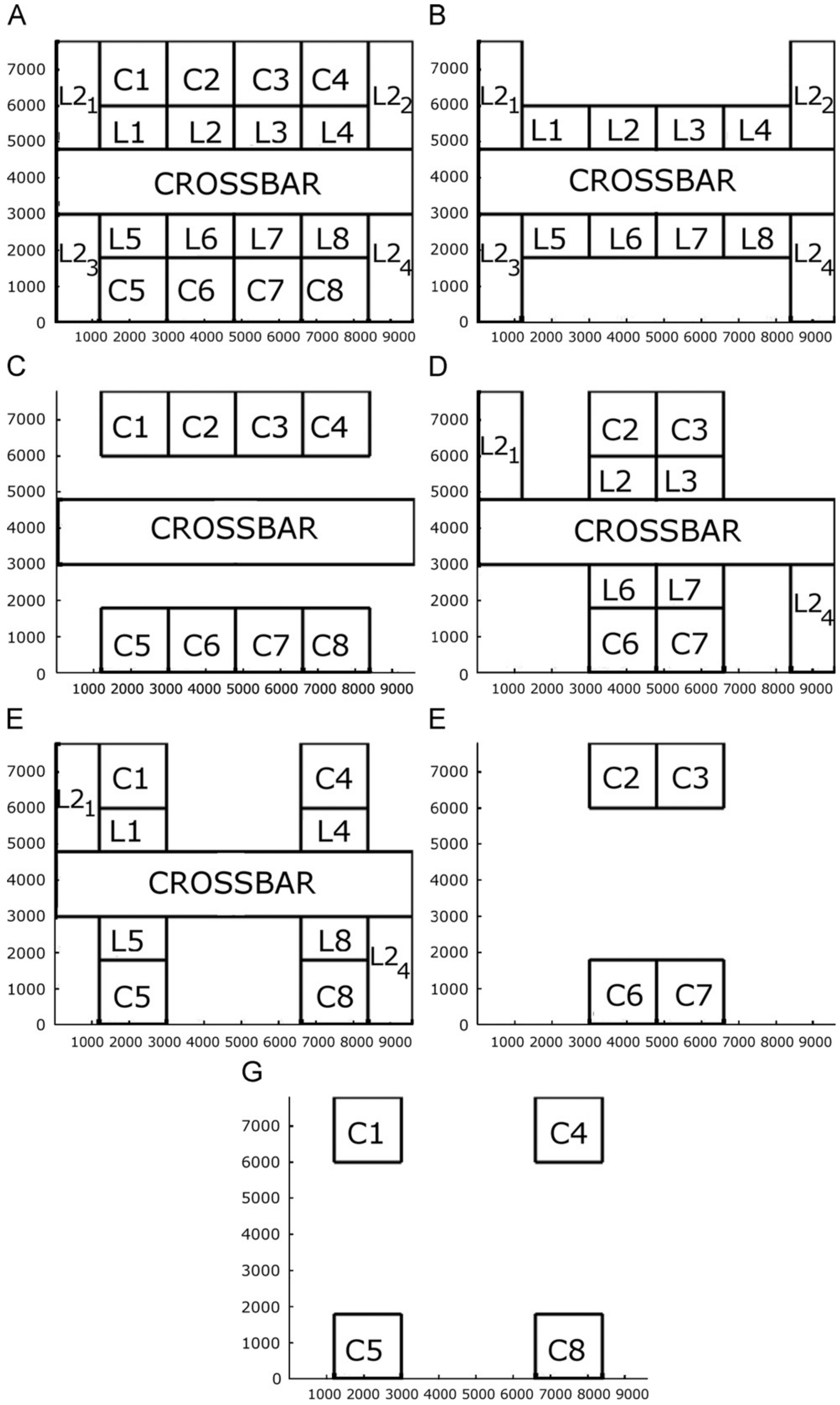

Fig. 17. Floorplan patterns.

Finally a 64-core system was tested. This configuration shows the thermal and technological problem of integrating a big number of cores in a 3D stack.

The metrics considered for the analysis of the experimental results are the mean temperature, the maximum temperature and the thermal gradients that can appear on the surface of the chip.

Table 4
3D layout description.

Layout	Pattern sequence	Core number	Layer number
1	A	8	1
2	DD	8	2
3	DE	8	2
4	ED	8	2
5	BFF	8	3
6	BFG	8	3
7	BGF	8	3
8	AA	16	2
9	DDDD	16	4
10	EDED	16	4
11	BFFBFF	16	6
12	BFGBFG	16	6
13	AAAAAAAA	64	8
14	BBBBBBBBAAAAAAAA	64	16
15	BCBCBCBCBCBCBCBC	64	16

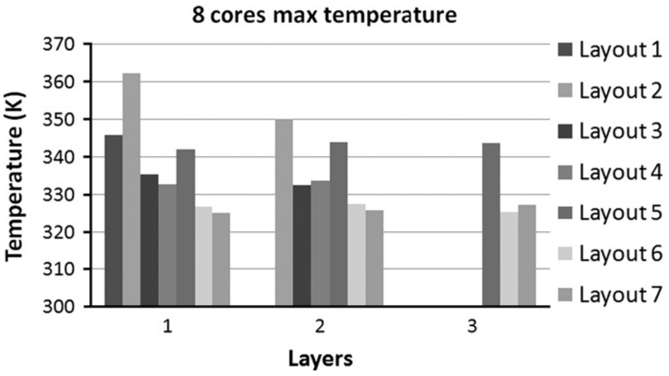

Fig. 18. Eight cores maximum temperature.

These metrics are usually found in every thermal-related analysis. We also included the study of the wire length in every layout to compare the delay introduced by our thermal aware rules in the considered scenarios. The results have been collected for an ambient temperature of 295 K.

7.3. 8-Cores configuration

The results for the 8-core configuration are shown in Figs. 18–21.

As can be seen in Fig. 18, the effect of increasing the number of layers is not always a good thermal-aware strategy, as the comparison of Layout 1 with Layouts 2 and 5 shows. In these configurations, the cores are placed close to each other and exhibit an increase in the maximum temperature of the chip. This can be explained by the difficulty of the cores to spread heat to sinks due to the placement of heat sources vertically aligned. When the heat sources are placed in this way, the vertical heat flow feedbacks positively, producing a high increase in the maximum temperature of the chip. On the other hand, if the cores are placed as we do in Layouts 6 and 7, we can reduce the maximum temperature in 20 K because now the cores can diffuse heat in every lateral direction. A similar behavior can also be observed for the thermal gradient (Fig. 19). Placing the cores close to each other, creates a big hot-spot and consequently a high thermal gradient. Therefore, a lateral and vertical spread of the cores is always a good policy.

Increasing the number of layers makes the mean temperature of the chip lower, as shown in Fig. 20. Heat can be diffused to upper or lower layers, but the adiabatic PCB base makes difficult to cool down the lower layers of the chip. Adding new layers also creates new heat sinks that help on reducing the temperature. Besides, this effect relieves the problem of the thermal gradients.

It can be determined from the wire length analysis that adding new layers sometimes implies adding more wire density for a constant area and number of cores. However, this impact is very low. For a reduction of 20 K of maximum temperature between Layouts 1 and 7, there is only an overhead of 3.5% of wires. Adding a new layer is not always detrimental for the wire length and the temperature as can be seen in Fig. 21 in Layouts 1, 3 and 4. While the temperature is reduced in 10 K approximately, the wire length is also decreased in a 0.1%.

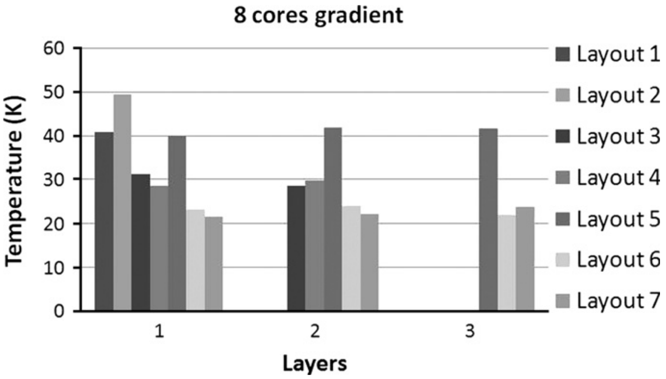

Fig. 19. Eight cores thermal gradient.

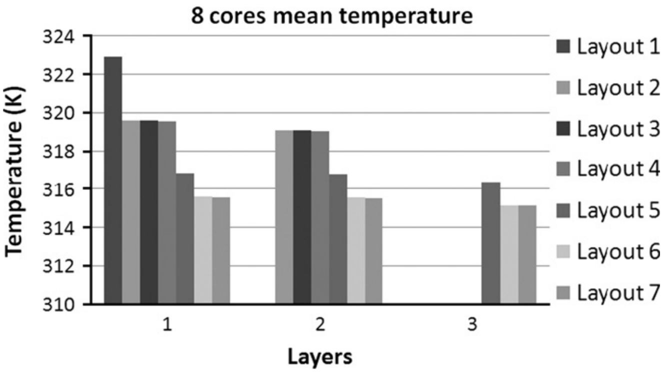

Fig. 20. Eight cores mean temperature.

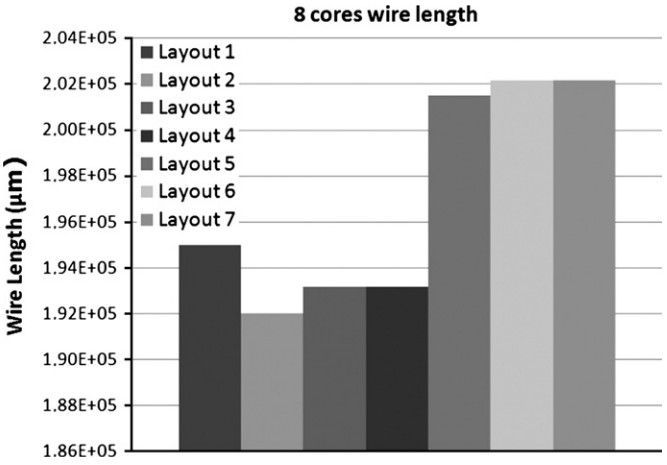

Fig. 21. Eight cores delay.

7.4. 16-Cores configuration

For the 16-core configuration we will refer to Figs. 22–25.

The analysis of this configuration will provide us some guidelines about the placement of heat sinks according to the placement of the sources. If we consider the maximum temperature, the same effects reported before for the increase in the number of layers can be observed; this time from Layouts 8 to 9 and from Layouts 10 to 11. The placement of heat sinks (memories) has to consider the placement of heat sources (cores) in order to reduce the extra wire density and the probability of hot-spots.

Including a bottom layer of memories creates a layer of heat sinks that is able to absorb heat from the upper layers, considering that the PCB does not allow heat flow through it. This effect can be observed in Fig. 22, Layout 12, where a bottom layer of memories is placed and then the stack is built alternating core and memory layers. With this approach, a reduction of 20 K in maximum temperature is achieved.

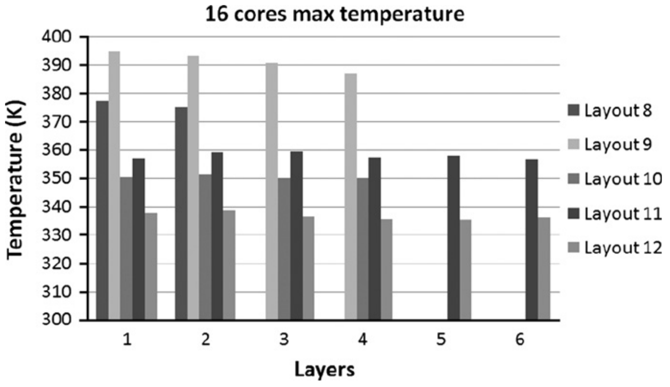

Fig. 22. The 16 cores maximum temperature.

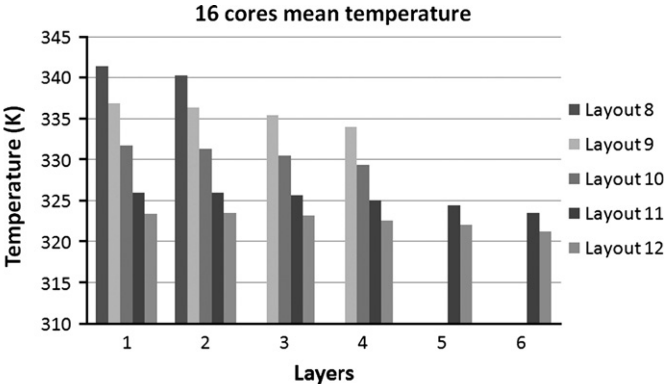

Fig. 23. The 16 cores mean temperature.

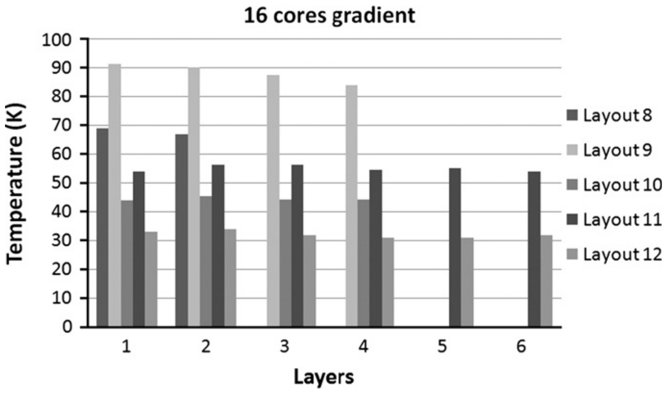

Fig. 24. The 16 cores thermal gradient.

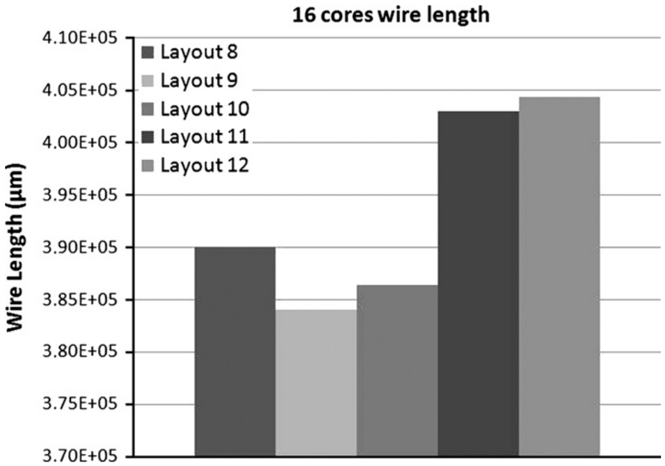

Fig. 25. The 16 cores delay.

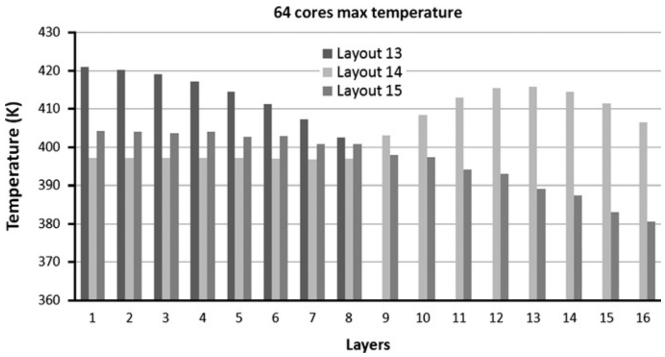

Fig. 26. The 64 cores maximum temperature.

The mean temperature follows a similar trend (Fig. 23). The more layers we place, the lower the mean temperature is, because both lateral and vertical heat diffusions are permitted. Besides, the effect of including memories between the core layers stabilizes the temperature, and the vertical thermal gradient is reduced (Fig. 24).

The effect observed for the wire length is very similar to the previous analysis, as shown in Fig. 25. Keeping the number of cores and adding two more layers in order to decrease the temperature has also a positive impact in the wire length (it is reduced). However, adding extra layers is not longer profitable after some point, as can be deduced from Layouts 5–7 of the same figure, where the overhead incurred is shown.

The complexity of the floorplanning problem increases with the number of cores to place. Moreover, as the number of cores increases, new thermal effects appear, as the vertical dissipation, presence of hot spots and placement highly dependent with the mean temperature. Therefore, the use of the thermal-aware floorplanning tool is required during the design process.

7.5. 64-Cores configuration

This set of experiments explores the effect of placing several sink layers together. As these layers warm up jointly, their temperature remains constant and the heat can flow from upper layers to the PCB.

Please cite this article as: J.L. Ayala, et al., Thermal modeling and analysis of 3D multi-processor chips, Integration VLSI J. (2010), doi:10.1016/j.vlsi.2010.06.002

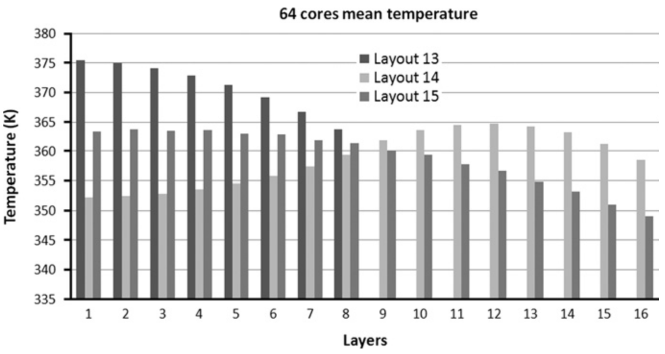

Fig. 27. The 64 cores mean temperature.

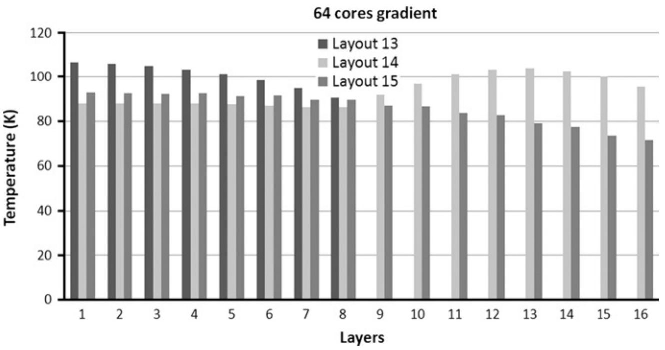

Fig. 28. The 64 cores thermal gradient.

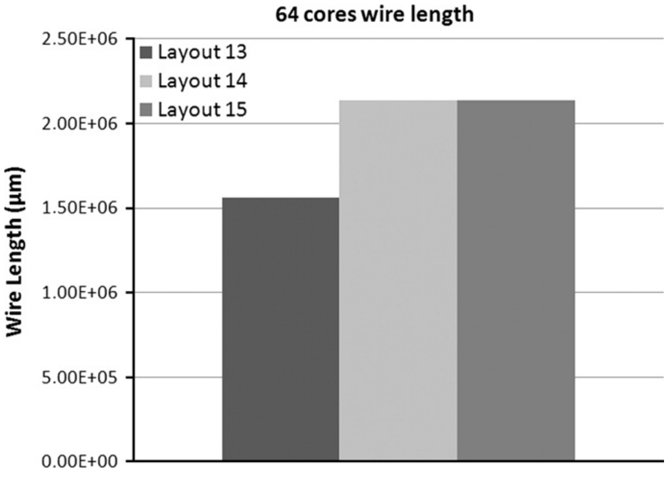

Fig. 29. The 64 cores delay.

On the other hand, the placement of core layer near to each other increases the temperature of the area surrounding the core (this effect can be observed in Figs. 26 and 27, Layout 14). The first eight bottom layers are composed of memories (sinks) and their temperatures are always comprised in a narrow range. However, the core layers show both an increase and decrease in the temperature, depending on the proximity to the ambient. The highest temperature is found in the middle core layer (Layer 13) because it can only exchange heat with its close neighbor.

Once more, the best thermal results are obtained when sink and core layers are combined. The temperature remains the same as observed in Layout 14, not exceeding 1 K of difference between two neighbor layers, and the decrease in temperature due to the effect of the top layer cooling can seen from layer 10 on (Fig. 28). Fig. 29 shows the impact of the layout configurations in the wire length for the 64-cores system.

As has been proved, increasing the core number implies more aggressive thermal effects, such as thermal gradients or the high increase of the temperature, reaching values that are out of technology limits. Hence again, a thermal model must be included in the floorplanning tool.

8. Working scenarios

This section classifies the floorplan configurations into several working scenarios, according their thermal behavior. As these different implementations must be studied thermally, the inclusion of a thermal model in the simulation flow of multi-processor systems is required.

Depending on the preferred working scenario specified by the designer, a thermal metric is defined to weight and select a floorplan. This metric incorporates the effect of the thermal factors that determine the optimality of the working scenario, weighted by the corresponding parameters.

$$G = \alpha T_{mean} + \beta T_{max} + \delta T_{grad} + \varepsilon W_{length}, \tag{9}$$

where T_{mean} is the mean temperature of the layer, T_{max} is the maximum temperature found in the layer and T_{grad} is the maximum thermal gradient. α, β, δ and ε are the factors of proportionality to be tuned experimentally and depending on the working scenario (i.e., they will be increased to prioritize the corresponding metric).

We will define three different scenarios, defined by the temperature and wire length metrics. These scenarios are:

- *Scenario* 1: *Low temperature*: In this scenario, a minimum of temperature is targeted as a working condition. These working conditions can be found in portable electronics such as mobile telephones, where cooling capabilities are a major thermal restriction.
- *Scenario* 2: *High performance*: This scenario is found in systems with high computation requirements. To achieve the delay constraints imposed, the wire length must be minimized, and cores must be placed as close as possible penalizing thermal improvement in favor of reducing communication delays.
- *Scenario* 3: *High reliability*: In this scenario, we will try to minimize thermal gradients and maximum temperature, metrics that have a great influence on the reliability of the circuits.

We will classify each of the 16 floorplans attending to their behavior in the different working scenarios. Table 5 presents three groups for the best, average and worst behavior for each working scenario described before. In this way, we can find which floorplan configuration is able to satisfy both performance and thermal constraints.

The analysis of this table presents that the layouts whose cores are close to each other exhibit much better performance (as happens with Layout 9) but, on the contrary, these configurations have a bad thermal behavior. On the other hand, if the thermal

Table 5
Scenario classification.

	Scenario 1	Scenario 2	Scenario 3
BEST	12	9	12
AVERAGE	11 10	10 8	10 11
WORST	9 8	11 12	8 9

behavior is optimized as in Layout 12, a worse performance is obtained. This fact has to be consider in the floorplan design to trade-off carefully temperature and performance according to the global constraints of our system. The average behavior is found for Layout 10, where a memory layer is firstly placed and then the core layers rotate their positions to avoid placing cores vertically aligned. This layout achieves frilly good marks for all the analyzed metrics.

These results support the thesis of this paper in which the integration of a thermal model for 3D architectures with a run-time transient thermal emulator can help on improving the results of the thermal-aware floorplanners.

9. Conclusion

This paper presents an accurate thermal model for the analysis of complex 3D multi-processor systems. This model has been profusely validated with real measurements on a silicon chip. The measurements performed in a real 5-layered 3D chip manufactured on purpose confirm the validity of the model with an error lower than a 2% in all the cases (lateral and vertical heat diffusion).

Later, the proposed model has been used to evaluate the capability of a nano-structure of thermal through-silicon vias to improve the thermal response of the 3D system. Also, the model has been able to characterize the strong influence that the floorplanning has in the thermal profile of a real 3D system as the Niagara.

Acknowledgements

This work was partly supported by the Spanish Government Research Grant TIN2008-00508 and by the Nano-Tera.ch NTF Project CMOSAIC (ref. 123618), which is financed by the Swiss Confederation and scientifically evaluated by SNSF.

Appendix A. Numerical solution of initial value problems

This section covers the mathematical background required to understand the solution of the thermal model. The numerical methods here presented determine the accuracy of the solution and the computation time until convergence. The mechanism implemented in our thermal model considers these constraints to obtain an efficient response in terms of performance and accuracy.

A.1. Definitions

An ordinary differential equation (ODE) is defined as a differential relationship between a dependent variable (y) and an independent variable (x). For this class of equations, we can define the initial value problem (IVP) attending to the specification of external data (initial conditions) that provide a unique solution to the differential equation. For the IVP, all the initial conditions are specified at one point.

Some of the key concepts associated with the numerical solution of IVPs are the local truncation error, the order and the stability of the numerical method. The numerical methods analyzed in this section will expose different behavior according to these metrics. Also, these numerical methods will be classified into implicit or explicit ones.

We are interested in the numerical solution of the IVP

$$\frac{dy}{dt} = f(y,t), \quad y(t=0) = y_0. \tag{10}$$

In particular, if $f(y,t) \equiv g(y)$, the IVP above is called autonomous and if $g(y\ ) = ky$ where k is a constant, the IVP is linear. We assume that a unique solution exists and denote that solution by $y^e(t)$. Therefore, $y(t)$ will refer to the numerically computed solution, which is only an approximation to $y^e(t)$, the analytical solution to the problem.

A.2. Forward and backwardEuler methods

Let us denote the time at the **n**th time-step by t_n and the computed solution at the **n**th time-step by y_n, i.e., $y_n \equiv y(t=t_n)$. The step size of the numerical solver h (assumed to be constant for the sake of simplicity) is then given by $h\ =\ t_n - tn - 1$. As we will present shortly, the quality of the solution is compromised by the selection of this parameter. Given (tn, yn), the forward Euler method (FE) computes y_{n+1} as

$$y_{n+1} = y_n + hf(y_n,t_n), \quad \text{(explicit) forward Euler method.} \tag{11}$$

The forward Euler method is based on a truncated Taylor series expansion, i.e., if we expand y in the neighborhood of $t=t_n$, we get

$$\begin{aligned} y(t_n+h) \equiv y_{n+1} &= y(t_n) + h\left.\frac{dy}{dt}\right|_{t_n} + \mathrm{O}(h^2) \\ &= y_n + hf(y_n,t_n) + \mathrm{O}(h^2). \end{aligned} \tag{12}$$

It is clear the error induced in Eq. (12) due to the truncation of the Taylor series. This error is known as the local truncation error (LTE) of the method. For the forward Euler method, the LTE is O(h^2), as happens in the first-order numerical techniques.[1] Higher order techniques provide lower LTE for the same step size, but at the cost of a higher computation effort.

The truncation error is different from the global error g_n, which is defined as the absolute value of the difference between the true solution and the numerically computed solution, i.e., $g_n = |y^e(t_n) - y_n|$. In most cases, the exact solution is not known and hence the global error is not possible to be evaluated. However, the global error can be approximated at the **n**th time step as $n \times LTE$, since **n** is proportional to 1/**h**, g_n should be proportional to **LTE/h**. According to this fact, the global error scales as h^k for a **k**th order method.

The forward Euler method is an explicit numerical method, i.e., y_{n+1} is given explicitly in terms of known quantities such as y_n and $f(y_n,t_n)$. Explicit methods are very easy to implement; however, their main drawback is the selection of a time step size that ensures numerical stability in the computation of the solution. It can be proved that for a linear IVP given by $dy/dt = -ay$, $y(0)=1$ with $a<0$, in order to prevent the amplification of the errors in the iteration process, it is required $|1-ah|<1$ or for stability of the forward Euler method, we should have $h<2/a$.

The conditional stability, i.e., the existence of a critical time step size beyond which numerical instabilities manifest, is typical of explicit methods such as the forward Euler technique. Implicit methods can be used to replace explicit ones in cases where the stability requirements of the latter impose stringent conditions on the time step size. However, implicit methods are more expensive to be implemented for non-linear problems since y_{n+1} is given

[1] In general, a method with O(h^{k+1}) LTE is said to be of **k**th order.

Please cite this article as: J.L. Ayala, et al., Thermal modeling and analysis of 3D multi-processor chips, Integration VLSI J. (2010), doi:10.1016/j.vlsi.2010.06.002

only in terms of an implicit equation. The implicit analogue of the explicit FE method is the backward Euler (BE) method. This is based on the following Taylor series expansion:

$$y_n \equiv y(t_{n+1}-h) = y(t_{n+1}) - h\left.\frac{dy}{dt}\right|_{t_{n+1}} + \mathrm{O}(h^2) \qquad (13)$$

which gives

$$y_{n+1} = y_n + hf(y_{n+1},t_{n+1}), \quad \text{(implicit) backward Euler method.} \qquad (14)$$

$f(y_{n+1},t_{n+1})$ is not known, hence it gives us an implicit equation for the computation of y_{n+1}. For instance, let $f(y,t) \equiv p(y) = y\cos(y)$. This means that to obtain y_{n+1}, we need to solve the non-linear equation $y_{n+1} - hy_{n+1}\cos(y_{n+1}) = y_n$ at any given time step **n**. A suitable root finding technique such as the Newton–Raphson method can be used for this purpose. This is evidently much more time consuming than the explicit FE method where, for the problem above, we have $y_{n+1} = y_n + hy_n\cos(y_n)$; but implicit techniques are stable for all $h > 0$. Also, the accuracy of the computed solution deteriorates as **h** is increased, and we expect the global error to scale linearly with **h**.

A.3. Higher order methods

FE and BE methods are first selection methods due to the simplicity of the implementation and low computational effort. However, they present poor convergence to the solution and this rate of convergence scales linearly with **h**. Several efficient higher order techniques exists in the literature and can be selected when the accuracy of the solution is the main constraint and the simulation time can be sacrificed.

Among other higher order methods, we can distinguish:

- **Runge–Kutta methods** (**RK**): a class of methods that use the information on the *slope* ($dy/dt|_{t_n}$) at more than one point to extrapolate the solution to the future time step. RK methods of order 2, 3, 4 or 5 can be derived, improving the accuracy of the calculated solution. RK methods are explicit techniques, hence they are only conditionally stable.
- **Adams methods** (**AM**): Adams methods are based on the idea of approximating, once the IVP is defined in an integral form, the integrand with a polynomial within the interval (t_n, t_{n+1}). Using a **k**th order polynomial results in a **k+1**th order method. There are two types of Adams methods, the explicit (known as Adams–Bashforth method) and the implicit (known as Adams–Moulton method).
- **Predictor–corrector methods**: which are defined as a suitable combination of an explicit and an implicit technique to obtain a method with better convergence characteristics. A configuration that is very often found consist of the FE and the second order AM (AM2) methods. In this case, the FE method is used as a predictor equation to get y^p_{n+1} and AM2 is used subsequently as a corrector equation to get the final computed solution y_{n+1}.

References

[1] S. Das, A. Chandrakasan, R. Reif, Design tools for 3-D integrated circuits, in: Proceedings of the 2003 Asia and South Pacific Design Automation Conference, 2003, pp. 53–56.

[2] K. Banerjee, S.J. Souri, P. Kapur, K.C. Saraswat, 3-D ics: a novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration, Proc. IEEE (2001) 602–633.

[3] A.W. Topol, et al., Three-dimensional integrated circuits, IBM J. Res. Dev. (2006) 494–506.

[4] Y. Deng, W.P. Maly, Interconnect characteristics of 2.5-D system integration scheme, in: Proceedings of the 2001 International Symposium on Physical Design, 2001, pp. 171–175.

[5] Y.S. Deng, W. Maly, 2.5D system integration: a design driven system implementation schema, in: Proceedings of the 2004 Asia and South Pacific Design Automation Conference, 2004, pp. 450–455.

[6] T. Kgil, S. D'Souza, A. Saidi, N. Binkert, R. Dreslinski, T. Mudge, S. Reinhardt, K. Flautner, Picoserver: using 3D stacking technology to enable a compact energy efficient chip multiprocessor, SIGOPS Oper. Syst. Rev. 40 (2006) 117–128.

[7] S. Mysore, B. Agrawal, N. Srivastava, S.-C. Lin, K. Banerjee, T. Sherwood, Introspective 3D chips, SIGARCH Comput. Archit. News 34 (2006) 264–273.

[8] A. Rahman, R. Reif, System-level performance evaluation of three-dimensional integrated circuits, IEEE Trans. Very Large Scale Integr. Syst. 8 (2000) 671–678.

[9] P. Morrow, et al., Wafer-level 3D interconnects via Cu bonding, in: Proceedings of the 2004 Advanced Metalization Conference, 2004, pp. 125–130.

[10] Y.-F. Tsai, Y. Xie, N. Vijaykrishnan, M.J. Irwin, Three-dimensional cache design exploration using 3Dcacti, in: Proceedings of the 2005 International Conference on Computer Design, 2005, pp. 519–524.

[11] A.Y. Zeng, J. Lu, R. Gutmann, K. Rose, Wafer-level 3D manufacturing issues for streaming video processors, in: Proceedings of the Advanced Semiconductor Manufacturing Conference, 2004, pp. 247–251.

[12] J. Mayega, O. Erdogan, P.M. Belemjian, K. Zhou, J.F. McDonald, R.P. Kraft, 3D direct vertical interconnect microprocessors test vehicle, in: Proceedings of the 13th ACM Great Lakes Symposium on VLSI, 2003, pp. 141–146.

[13] K. Puttaswamy, G. Loh, The impact of 3-dimensional integration of the design of arithmetic units, in: Proceedings of the IEEE International Symposium on Circuits and Systems, 2006, pp. 4951–4954.

[14] B. Black, D.W. Nelson, C. Webb, N. Samra, 3D processing technology and its impact on iA32 microprocessors, in: Proceedings of the IEEE International Conference on Computer Design, 2004, pp. 316–318.

[15] D.W. Nelson, et al., A 3D interconnect methodology applied to iA32-class architectures for performance improvement through RC mitigation, in: Proceedings of the 21st International VLSI Multilevel Interconnection Conference, 2004, pp. 453–464.

[16] Y. Xie, G.H. Loh, B. Black, K. Bernstein, Design space exploration for 3D architectures, J. Emerg. Technol. Comput. Syst. 2 (2006) 65–103.

[17] Samsung, Samsung 3D, 〈http://www.samsung.com〉, 2009.

[18] Tezzaron, Tezzaron 3D, 〈http://www.tezzaron.com〉, 2009.

[19] S. Im, K. Banerjee, Full chip thermal analysis of planar (2-D) and verticallyintegrated (3-D) high performance ICs, in: IEDM Technical Digest, International Electron Devices Meeting, 2000, pp. 727–730.

[20] A. Rahman, R. Reif, Thermal analysis of three-dimensional (3-D) integrated circuits (ICs), in: IITC Conference, 2001, pp. 157–159.

[21] T.-Y. Chiang, S.J. Souri, C.O. Chui, K.C. Saraswat, Thermal analysis of heterogeneous 3-D ICs with various integration scenarios, in: IEEE International Electron Devices Meeting, pp. 31.2.1–31.2.4.

[22] B. Goplen, S. Sapatnekar, Placement of thermal vias in 3-D ICs using various thermal objectives, IEEE Trans. Computer-Aided Des. Integrated Circuits Syst. 25 (2006) 692–709.

[23] P. Leduca, et al., Challenges for 3D IC integration: bonding quality and thermal management, in: IEEE International Interconnect Technology Conference, 2007, pp. 210–212.

[24] K. Puttaswamy, G.H. Loh, Thermal analysis of a 3D die-stacked high-performance microprocessor, in: Proceedings of the 16th ACM Great Lakes Symposium on VLSI, pp. 19–24.

[25] J. Cong, Y. Zhang, Thermal via planning for 3-D ICs, in: Proceedings of the 2005 IEEE/ACM International Conference on Computer-aided Design, 2005, pp. 745–752.

[26] A. Jain, R. Jones, R. Chatterjee, S. Pozder, Z. Huang, Thermal modeling and design of 3D integrated circuits, in: Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems, 2008, pp. 1139–1145.

[27] V. Natarajan, A. Deshpande, S. Solanki, A. Chandrasekhar, Thermal and power challenges in high performance computing systems, in: International Symposium on Thermal Design and Thermophysical Property for Electronics, 2008, pp. 78–83.

[28] S. Heo, K. Barr, K. Asanovic, Reducing power density through activity migration, Proc. ISPD (2003) 217–222.

[29] K. Skadron, M.R. Stan, K. Sankaranarayanan, W. Huang, S. Velusamy, D. Tarjan, Temperature-aware microarchitecture: modeling and implementation, Trans. Archit. Code Optim. 1 (2004) 94–125.

[30] H. Su, F. Liu, A. Devga, E. Acar, S. Nassif, Full chip leakage estimation considering power supply and temperature variations, Proc. ISPD (2003) 78–83.

[31] J. Vlach, K. Singhal, Computer Methods for Circuit Analysis and Design, Springer, 1983.

[32] F.P. Incropera, D.P. Dewitt, T.L. Bergman, A.S. Lavine, Fundamentals of Heat and Mass Transfer, John Wiley and Sons, 2007.

[33] T.A. Davis, I.S. Duff, An unsymmetric-pattern multifrontal method for sparse LU factorization, SIAM J. Matrix Anal. Appl. (1997) 140–158.

[34] Y.-J. Lee, R. Goel, S.K. Lim, Multi-functional interconnect co-optimization for fast and reliable 3D stacked ICs, in: International Conference on Computer-Aided Design, 2009, pp. 645–651.

[35] Sun, Sparc t1, 〈http://www.sun.com/processors/UltraSPARC-T1/〉, 2010.

Jose L. Ayala got his MS in Telecommunication Engineering and his Ph.D. in Electrical Engineering in 2001 and 2005, respectively, both from Politecnica University of Madrid. He also has a MS in Physics from the Open University of Spain. Prof. Ayala is currently an Associate Professor in the Complutense University of Madrid (Spain) and a Permanent Visiting Professor at EPFL (Switzerland). His research interests include thermal-aware electronic design and thermal-aware compilation, thermal-modeling and power-efficient computer architectures.

Arvind Sridhar completed a MS of Applied Sciences in Electronics at Carleton University in 2009. He is currently a Ph.D. student at EPFL (Switzerland), and his research interests include simulation and optimization for CAD tools, thermal modeling, 3D integrated circuits and computational fluid dynamics.

David Cuesta completed got a MS in Physics and a MS in Electrical Engineering from Complutense University of Madrid, both in 2008. He is currently a Ph.D. student at the same university and his research interests include thermal-aware run-time techniques for multi-processors, thermal simulation and thermal-aware floorplanning.

Please cite this article as: J.L. Ayala, et al., Thermal modeling and analysis of 3D multi-processor chips, Integration VLSI J. (2010), doi:10.1016/j.vlsi.2010.06.002