Transistor Temperature Deviation Analysis In Monolithic 3D Standard Cells

M. Brocard, B. Mathieu, J-P. Colonna, C. Santos, C. Fenouillet-Beranger, V. LU Cao-Minh, G. Cibrario, L. Brunet, P. Batude, F. Andrieu, S. Thuries, O. Billoint
Univ. Grenoble Alpes, F-38000 Grenoble, France
CEA, LETI, MINATEC Campus, F-38054 Grenoble, France
melanie.brocard@cea.fr

Abstract—This study focuses on investigating the temperature deviation during the operation of transistors within a monolithic 3D standard cell built on two tiers. An early assessment of this topic is crucial to manage circuit design and requires both steady-state and transient thermal analysis at transistor level. A representative 3D standard cell in 14nm FDSOI technology is considered, using intermediate Back-End-Of-Line (iBEOL) and top tier BEOL. Steady-state and transient power dissipations in NMOS and PMOS are extracted from SPICE simulations with variable parameters such as output load capacitance and operating temperature. 3D thermal simulations are then performed to assess the impact of design parameters such as routing densities, thicknesses of iBEOL and BEOL, their number of metal layers and packaging techniques. Steady-state and transient thermal simulations enable accurate analysis to correlate the temperature deviation of transistors with these parameters. Design guidelines are provided to limit the temperature deviation between top and bottom tier which can reach up to 7°C during circuit operation in the worst case.

Keywords—3D monolithic integration; thermal simulation; transistor level; transient state; 3D VLSI; standard cells

I. INTRODUCTION

3D monolithic integration is a breakthrough technology raising great interest in the semiconductor industry for its promising performances in numerous applications [1]. Also called 3D Sequential technology, this process enables the fabrication of circuits using multiple stacked layers of devices connected at the transistor scale with vertical inter-tier via. It enables higher transistors density [2], reduced footprint and shorter routing wire length for a given technology node compared to 3D TSV-based technology. Fig.1 illustrates the 3D circuit general architecture with two active layers. The 3D inter-tier vias allow a very short vertical connection between active layers. In 14nm technology, 3D inter-tier via can be as small as 50nm wide and 150nm tall with a 100nm pitch [3].

Yet, 3D sequential is a disruptive technology which leads to a lot of technological and design challenges to lay ahead. In designer’s field, numerous studies on two-tiers 3D monolithic designs predict noticeable performance improvements [4][5]. The core of these studies (transistors and their electrical characteristics) relies on a strong hypothesis: it is supposed that at each tier, devices are operating at the same temperature. When considering the peculiarity of the 3D monolithic stacking, this assumption is not obvious and has to be proven especially in the case of 3D standard cells which are built on several tiers. Within a standard cell, all devices must be at the same temperature for right operations. A difference of a few degrees within a bitcell such as the one presented in [6] or within a standard cell would entail electrical characteristics (threshold voltage, rise and fall times, delays, etc.) deviations, resulting in heterogeneous characteristics in design tools.

Hence the need to predict the thermal behavior of this new 3D architecture is a crucial element to design a reliable system with high performances. In state of the art [7][8], 3D monolithic thermal studies at circuit level with simplified process layers and block-level power sources are reported. However, it is necessary to study the thermal behavior at the standard cell level to observe the temperature of every transistor in both steady and transient states. If a significant temperature difference between the two transistor layers is found, design guidelines will be required. In this work, we consider different parameters at the standard cell scale leading to temperature deviation variations. Besides, different design options (iBEOL and BEOL thermal parameters, packaging, etc.) are considered to evaluate the temperature deviations (ΔT) between transistors of a representative standard cell using a CFD thermal analysis tool.

The paper is organized as follows: in Section II the technology process, materials and geometries are detailed. In Section III we describe the test case and explain the different steps of the methodology we applied for an accurate simulation in a 3D thermal analysis tool. Section IV presents the simulation results in steady and transient states considering several design parameters. The last section draws conclusions and perspectives.

II. TECHNOLOGY PRESENTATION

The 3D monolithic integration process flow is described in Fig.2.
While several steps of the fabrication process are still under development because it requires new specific process considerations, concept and functionality were demonstrated in [9]. In this integration, bottom transistors are fabricated with a standard process while the top transistors need to be fabricated using a cold process [10] to preserve the bottom layers (bottom FEOL and iBEOL) electrical and mechanical properties and reliability.

From the design point of view, it is useful to precise that 3D monolithic technology is versatile and allows CMOS on CMOS, unipolar (one MOS type) on CMOS, unipolar on unipolar. In this paper, we describe bottom FDSOI PMOS FET comprising thin silicon active areas, buried oxide related to the SOI technology and STI (Silicon Trench Isolation, generally made of SiO$_2$) that fills the gaps between active zones. The top NMOS FET are described with a thin silicon layer and a polysilicon back gate used to bias and control the device channel. The thick silicon layers exhibit significantly lower thermal conductivity compared to bulk. TABLE I presents the thermal characteristics and thicknesses of material layers we set in the thermal analysis tool. The iBEOL and BEOL are made with different dielectrics for thermal budget stability reasons [11]. Indeed, the iBEOL should be fabricated with a dense oxide such as SiO$_2$ in order to avoid shrinking during multiple thermal anneals of the top MOS process while the BEOL is made with porous high-k dielectric to reduce parasitic capacitances. This implies different thermal properties with thermal conductivity and specific heat being divided by two compared to iBEOL values in our simulations.

Among all the possible technological, electrical, thermal and geometrical parameters, we selected the thicknesses and thermal conductivities of the iBEOL and BEOL layers as they are the most relevant for our study since they are design dependent. In the next part, we will present the steps we followed to describe transistor devices of a 3D monolithic test circuit in our simulation tool with realistic thermal and electrical conditions.

### III. MODELING IN THERMAL SIMULATION TOOL

In this section we aim at modeling the thermal behavior of a standard cell at transistor scale, using multiple relevant design parameters. Thermal simulations have been carried out using Mentor Graphics FloTHERM to accurately modeling the heat dissipation within the 3D structure. Indeed, the structures we want to study are not all located in the same plane and we cannot therefore consider a unidirectional heat flow in the 3D monolithic geometry.

#### A. Test Case Geometry Description

The first step of our thermal assessment strategy was to define a piece of 3D monolithic design that would be representative of any circuit regardless the physical implementation flavor (NMOS on PMOS, CMOS on CMOS). A 3D NMOS on PMOS buffer function was selected to embody the thermal behavior of 3D monolithic designs. We have built a two active layers structure with the typical dimensions of a 14nm standard cell, i.e. a fixed y-axis cell height of 880nm (delimited by the power supply rails) and an x-axis width measuring two CPPs (Contacted Poly Pitch). Two devices are implemented on the bottom layer and two on the top one with different gate widths (150nm and 280nm as shown Fig. 3) in order to vary the power density per device. The silicon bulk is reduced from its 750 m-real-thickness to 3 m to reduce run time while keeping the accuracy of the simulation results.

![Fig. 3. Geometry structure of the test case and schematic of the buffer](image-url)
with P1 and N1 w=150nm and P2 and N2 with w=280nm.

Fig. 4. 3D view of the test case, with two bottom PMOS, two top NMOS and their power mapping located in the drain-channel areas.

The solution domain, i.e. the region of space within which the different thermal equations have to be solved, is strictly limited to the solid shape where heat transfer occurs only by conduction.

B. Side Boundary Conditions

This standard cell-like geometry needs an accurate description of its thermal environment which is set through the boundary condition parameters.

Fig. 5 shows a top view of a typical digital circuit containing standard cells. The optimized cell placement leaves very few empty spaces. We assume the same in a 3D monolithic digital circuit where the physical implementation will also be optimized to result in a high area utilization. Our 3D test case is designed to represent a little piece of a digital circuit. As we want to work as much as possible with realistic worst case conditions, we assume that our test case is surrounded by similar neighbors. Therefore, the four sides in y-z and x-z planes have been set with symmetry boundary conditions in the simulation tool. With top and bottom boundary conditions, we are able to emulate different packaging techniques as described in the next section.

D. iBEOL and BEOL Thermal Modelization

iBEOL and BEOL metallization layers are surrounded by dielectrics with low thermal conductivities as described in section II and include tungsten or copper vias and metal lines with much higher thermal conductivities (see TABLE I). In standard cells, transistors are interconnected using the two first metal levels. Inter-cell routing and power distribution use 3 to 10 metal levels. For each BEOL and iBEOL metallization level, metal density can vary a lot across the circuit. Thus, the iBEOL and BEOL are a mix of material entailing heterogeneous thermal properties.

To model such a complex thermal behavior while covering the wide range of routing density in circuits, we decided to homogenize the iBEOL and BEOL layers. We rely on the reference [12] in which authors compute and provide equivalent orthotropic thermal conductivity for several densities of metal and vias. More precisely, [12] reports thermal conductivity $k_x$ and $k_y$ (in the x-y plane as defined Fig. 4) higher than vertical one $k_z$. Indeed, the x-y plane contains the metal lines made with copper, while in the z-axis, there are the iBEOL and BEOL vias acting like a heat flow barrier. This creates a thermal discontinuity and entails a higher temperature gradient along z-axis. Thus thicknesses of iBEOL and BEOL, i.e. the number of metal levels, are also part of our parameters to study this thermal phenomenon. Therefore, we have defined in TABLE II three iBEOL and BEOL profiles that can be typically observed in 14 nm circuits.
IV. Power Dissipation in Transistors

To achieve transistor scale thermal analysis, we need to define power sources. In a circuit, the dissipated power comes mainly from devices and is generated by Joule effect. More specifically it occurs in the channel-drain junction area so the power will be mapped there (Fig. 4).

To know exactly the dissipated power, which is equal to the power consumption in the transistors, a SPICE simulation of the buffer was performed. We assume that the operating temperature of the transistors and their loads are two main parameters impacting power consumption so they need to be considered. Three typical output loads of 1fF, 5fF and 10fF were selected in the lookup tables of the buffer in the foundry .lib file for digital place and route. Simulations were run at 25°C, 40°C and 80°C for each load value at an operating frequency of 500MHz. For each transistor, drain current (Fig. 7) and drain to source voltage were plotted as a function of time and multiplied to obtain the power dissipation. The average power was computed for the steady state thermal analysis (TABLE III) while instantaneous power was used for the transient analysis.

Power consumption stays nearly the same when device temperature ranges from 25°C to 80°C. In contrast, the power consumption of the output stage (P2, N2) of the considered buffer is dramatically impacted by the load capacitance value.

Thus we have kept the load capacitance as a parameter of the thermal study and chose to use the associated dissipated power at 40°C (values in blue boxes in TABLE III). Power dissipation of N2 and P2 could represent any power dissipation that occurs in N and P transistor outputs of standard cells. In parallel, power dissipation of N1 and P1 is generalizable to any transistor within a standard cell. In the next part, we will perform thermal simulations to investigate the impact of the aforementioned parameters on the transistor temperature during circuit operations.

V. Transistor Thermal Deviation Analysis

Thermal simulation results of the 3D buffer cell are now discussed according to various metallization thermal properties, output load of the standard cell and packaging options. In the simulated testcases, four monitor points are located in the P1, N1, P2, N2 channels in order to measure each transistor temperature. TABLE IV brings the simulation results where \( T_{N1} \) is the reference temperature for transistor N1 and temperature deviation for transistors P1, P2, N2 is calculated as \( \Delta T_i = T_{i,x} - T_{N1} \).
Hence if $\Delta T_x$ is negative, it means that $x$ is cooler than N1, if it is positive, $x$ is hotter than N1. Note that N2, P2 are hotter than N1, P1 when the load is 10fF because power density is higher, while with 1fF, power densities are lower for N2 and P2. Colors of the “case” boxes refer to routing density (TABLE II).

A. Power Dissipation

In this part, effects of average dissipated power in each transistor on the $\Delta T$ in a thermal steady-state are analyzed. The $\Delta T$ between top and bottom devices stays below 2°C when the load capacitance is 1fF for all considered sets of parameters. The $\Delta T$ is accentuated when the standard cell load capacitance increases and reaches 7°C in the worst simulated case such as the number 13 in TABLE IV. A first design guideline could be to set a maximum load capacitance value in digital place & route flows.

B. iBEOL and BEOL properties

In this part we investigate the effect of iBEOL and BEOL metal densities and thicknesses on the temperature deviation of each transistor in the test case.

Comparing the several $\Delta T$ for a light routing case as described in TABLE II to a medium routing case ( cases 6, 7 and 8), with four available metal layers for routing in iBEOL, the temperature deviation between transistors remains the same (up to 6.7°C). For these cases thicknesses between active layers and $k_x$ are the same. If routing is set as dense, with higher $k_x$, temperature deviation is considerably lowered (below 2.8°C for 2 or 4 metal levels, cases 3, 5, 9 and 14). Actually, whatever the lateral conductivities $k_x$ and $k_y$, the important factor that determines $\Delta T$ is $k_z$, which is linked to the via density. The $k_x$ and $k_y$ are linked to the metal line density and therefore have an effect on the homogenization of temperature for each layer, as shown in Fig. 8. Moreover, BEOL only contributes to heat dissipation capacity of the circuit through the top and is shown to have a negligible impact on $\Delta T$ (comparing cases 7 and 8 or 11 and 13) contrary to iBEOL. Thickness of iBEOL (number of metal levels) has a slighter impact. Therefore, to improve heat coupling between top and bottom layers, the number of via should be maximized in order to optimize the conductivity in the z-axis.

C. Packaging impact

Two options are now explored: wire bonding and flip chip. In both cases no significant heat flux was observed at the boundary “air” (set at 20W/m².K heat transfer) so the external temperature and the heat transfer coefficient have negligible impact on $\Delta T$.

Second, as expected, the top device temperature elevation is higher in a wire bond package because it suffers from heat confinement between iBEOL and BEOL with low thermal conductivities, whereas the bottom layer is in direct contact with the bulk which efficiently spreads the heat to the PCB. In a flip chip package, this tendency is reversed (Fig. 9), the top devices are the closest to the PCB heat spreader.

Nevertheless, results do not show a noticeable influence of the packaging on $\Delta T$. The packaging technique determines only where heat escapes from the chip, being through silicon bulk for wire bonding and through BEOL for flip chip technique.

---

**TABLE IV. STEADY-STATE THERMAL SIMULATION RESULTS**

<table>
<thead>
<tr>
<th>Case</th>
<th>Packaging</th>
<th>Load (fF)</th>
<th>iBEOL metal levels</th>
<th>BEOL metal levels</th>
<th>$T_{N1}$ (°C)</th>
<th>$\Delta T_{N2}$</th>
<th>$\Delta T_{P1}$</th>
<th>$\Delta T_{P2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Wire bonding</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>41.4</td>
<td>0</td>
<td>-0.9</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>Wire bonding</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>44</td>
<td>0.9</td>
<td>-2.8</td>
<td>-2.5</td>
</tr>
<tr>
<td>3</td>
<td>Wire bonding</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>42.4</td>
<td>0.5</td>
<td>-1.2</td>
<td>-0.9</td>
</tr>
<tr>
<td>4</td>
<td>Wire bonding</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>43.9</td>
<td>1.6</td>
<td>-2.8</td>
<td>-2.2</td>
</tr>
<tr>
<td>5</td>
<td>Wire bonding</td>
<td>10</td>
<td>4</td>
<td>2</td>
<td>44</td>
<td>0.5</td>
<td>-2.8</td>
<td>-2.5</td>
</tr>
<tr>
<td>6</td>
<td>Wire bonding</td>
<td>10</td>
<td>4</td>
<td>2</td>
<td>47.7</td>
<td>0.7</td>
<td>-6.6</td>
<td>-6.3</td>
</tr>
<tr>
<td>7</td>
<td>Wire bonding</td>
<td>10</td>
<td>4</td>
<td>2</td>
<td>47.9</td>
<td>0.9</td>
<td>-6.7</td>
<td>-6.3</td>
</tr>
<tr>
<td>8</td>
<td>Wire bonding</td>
<td>10</td>
<td>4</td>
<td>4</td>
<td>47.8</td>
<td>0.8</td>
<td>-6.7</td>
<td>-6.3</td>
</tr>
<tr>
<td>9</td>
<td>Wire bonding</td>
<td>10</td>
<td>4</td>
<td>4</td>
<td>43.8</td>
<td>0.4</td>
<td>-2.7</td>
<td>-2.4</td>
</tr>
<tr>
<td>10</td>
<td>Wire bonding</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>41.4</td>
<td>0</td>
<td>-0.9</td>
<td>-0.1</td>
</tr>
<tr>
<td>11</td>
<td>Flip chip</td>
<td>10</td>
<td>4</td>
<td>2</td>
<td>54.5</td>
<td>0.5</td>
<td>6.4</td>
<td>6.8</td>
</tr>
<tr>
<td>12</td>
<td>Flip chip</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>45.7</td>
<td>0</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>13</td>
<td>Flip chip</td>
<td>10</td>
<td>4</td>
<td>4</td>
<td>65.9</td>
<td>0.7</td>
<td>6.7</td>
<td>7.1</td>
</tr>
<tr>
<td>14</td>
<td>Flip chip</td>
<td>10</td>
<td>4</td>
<td>4</td>
<td>44.5</td>
<td>0.35</td>
<td>2.7</td>
<td>3</td>
</tr>
</tbody>
</table>

---

Fig. 8. (a) Case 4 (2 metal levels), description with isotropic thermal properties in iBEOL ($k_x=k_y=k_z=1.2$) (b) Case 7 (four metal level in iBEOL), description with orthotropic properties, transistors on each layer are isothermal in case 7, not in case 4.

Fig. 9. Case 12 flip chip, NMOS are now operating at lower temperature than PMOS.
**D. Transient State**

During transient thermal analysis, the tool also takes into consideration specific heat and materials density properties (TABLE I). The specified dissipated power is no longer averaged but instead described as transient power peaks using a time step distribution. The power peaks in each device for the chosen clock frequency of 500 MHz are largely spaced as observed in SPICE simulation (Fig. 7) and for higher frequencies they would still do.

Fig. 10 shows the time dependency of the temperature for all 4 transistors at 500MHz. This figure shows that when a transistor switches, the other transistors do not see any temperature variation. Hence the transient thermal coupling does not need to be taken into account to model the cell behavior and a steady state thermal simulation is sufficient. This is due to the combination of a high frequency and a low thermal diffusivity of the homogenized iBEOL. Temperature variations in nearby transistors would be seen at much lower frequencies but the average dissipated power would also be strongly reduced as well as the induced temperature variations. Hence the transient thermal coupling would also have no measurable effect at lower frequencies.

Fig. 10. Case 8, time dependency of the temperature for all 4 transistors at 500Mhz.

![Fig. 10. Case 8, time dependency of the temperature for all 4 transistors at 500Mhz.](image)

At the operating frequency of 500MHz and with the selected homogenized properties of the iBEOL, temperature fluctuations due to a power spike in a transistor are negligible at a distance larger than 30nm of the transistor, as shown in Fig. 11.

However, the thermal diffusivity of the homogenized iBEOL is probably underestimated compared to the real structure. Indeed the metal lines provide a very fast path for the heat to reach the other transistors, and the average thermal diffusivity of the iBEOL is different from the thermal diffusivity computed from homogenized thermal properties. A more accurate description of the iBEOL taking into account the layout of the metal lines will probably lead to longer propagation distances of the temperature spikes generated by the transistors.

**VI. CONCLUSION**

In this paper, we have investigated the impact of the main design parameters to find out design solutions to overcome temperature deviation in 3D monolithic technology in operating condition.

In order to extrapolate the conclusions to the wide variety of standard cells we can find in digital circuits, we have designed a test case with symmetry conditions on the sides, various metallization and power dissipation profiles linked to the standard cell activity and load. Higher drive standard cells should have the same thermal behavior since we assume that we have covered the whole range of possible power densities in transistors. The impact of packaging has been also investigated by the means of changing the top and bottom boundary conditions. The thermal investigation has been performed by isolating the above considerations to individually evaluate each parameter effect.

Results show low and acceptable temperature deviations between top and bottom active tiers during operations. Two main parameters are mainly responsible to the vertical temperature gradient but design guidelines can be applied to mitigate such effects. We also demonstrated that this deviation occurs in steady state and that transient state exhibits local heating in the neighborhood of each transistor without affecting the others. The study provides design and technology guidelines to deal with possible thermal issues and enable more accurate circuit performance prediction.

Finally, a more detailed iBEOL model taking into account the real layout would be required to overcome the limitations of the homogenization approach.

**REFERENCES**


