Thermal analysis and model identification techniques for a logic + WIDEIO stacked DRAM test chip

Francesco Beneventi[†], Andrea Bartolini[†§], Pascal Vivet[‡], Denis Dutoit[‡] and Luca Benini[†§]
[†]Department of Electrical, Electronic and Information Engineering (DEI), University of Bologna, Italy
{francesco.beneventi, a.bartolini, luca.benini}@unibo.it
[§]Integrated Systems Laboratory, ETH Zurich, Switzerland
{barandre, lbenini}@iis.ee.ethz.ch
[‡]CEA-Leti, Grenoble, France
{pascal.vivet, denis.dutoit}@cea.fr

***Abstract*—High temperature is one of the limiting factors and major concerns in 3D-chip integration. In this paper we use a 3D test chip (WIDEIO DRAM on top of a logic die) equipped with temperature sensors and heaters to explore thermal effects. We correlated real temperature measurements with the power dissipated by the heaters using model learning techniques. The resulting compact thermal model is able to predict temperatures at chip locations far from the temperature sensors and to infer the power dissipation at any location of the chip. Results are verified by mean of an off-sample validation technique and show a high accuracy of the compact thermal model when compared with silicon measurements.**

I. Introduction and Related Work

3D-chip integration try to overcome scalability and performance bottleneck by stacking different silicon dies aiming to augment the silicon active area (e.g. number of processing element, memory banks, etc) with low latencies access time [3], [17]. Indeed inter-die wire connection in 3D stacking is significantly shorter and more efficient than in 2D planar integration. In 2D wire is longer and must traverse several I/O interfaces before reaching the end-point while in 3D-chip the wire can traverse the different die vertically by mean of Through Silicon Via (TSVs) [3].

Today several chips have been proposed with 3D technology, the majority of them stack memory dies on the top of an active one. These strategies increase the bandwidth to the memory while reducing the memory access energy with respect to the 2D planar integration. Recent JEDEC Wide I/O interface specs were defined for SDRAM interface to deliver twice the bandwidth of the LPDDR2 specification, at the same operating power and to allow for thinner, smaller system form factors specifically through the use of TSVs [12], [19].

Whereas 3D-chip integration significantly increases the performances of today devices, it has been shown to suffer from thermal related issues. Indeed removing heat from the bottom layers is difficult as they cannot be directly connected to a heat sink. It follows that heat flows in the stacked layers, inducing high temperature hot-spots and gradients. To overcome this issues different strategies have been proposed to augment the heat dissipation capability. Some of them exploit additional TSVs placement to serve as heat pipes for reducing the overall resistance toward the ambient [4], [9] while other solutions adopt liquid cooling through micro-channels on the silicon die [15], [16].

978-3-9815370-2-4/DATE14/©2014 EDAA

While different thermal simulators allow to estimate the thermal effects in 3D-chip [1], [11], practical evaluation is more difficult as the typical IR-cam approach has limitation in measuring the different die temperatures [7]. Recently, 3D test chips have been designed to evaluate 3D-integration thermal properties by embedding controllable heaters and thermal sensors [13]. This approach allows to estimate how the heat spreads in the silicon die, but limited only to the available thermal sensors positions.

Thermal interpolation strategies allow to estimate the temperature of the die locations different than those corresponding to the thermal sensors. Few works in state-of-the-art tackle this problem but are primarily evaluated on simulation and on planar devices [6], [20]. These techniques often rely on the knowledge of thermal models of the underlying HW. Under the same assumption, additional techniques solve the inverse problem and try to estimate the power consumption starting from the thermal sensors output. This problem is well-known to be ill-conditioned as consequence of that it is strongly affected by measurement noise and models errors [14].

In this paper we propose a technique that takes advantage of built-in heaters and thermal sensors of MAG3D, a 3D test chip designed by CEA-LETI, for extracting the static thermal model. This model can be coupled with floorplan data to extrapolate thermal information at chip locations not covered by a thermal sensor or heating source. Moreover, these strategies allow to estimate the power consumption of HW components by simply knowing their position.

The rest of the paper starts with an overview of the MAG3D architecture. Section III provides the details of our thermal modeling approach. In Section IV we show how the modeling framework can be applied to practical cases to augment the thermal and power introspection in 3D chip. In Section V we demonstrate the results of our analysis and Section VI concludes the paper.

II. MAG3D Architecture

MAG3D [5], [21] is a WideIO memory-on-logic 3D circuit (65nm technology node) where the dies are stacked in a face-to-back configuration and are connected through TSVs and μ-bumps. The SoC logic die implements a WIOMING circuit and is only 80um thick to accommodate the integrated TSVs. The WIOMING circuit (Fig.1a) includes:

- Four memory controllers, in the center, one per WideIO memory channel, plus the corresponding TSV and μ-bumps matrix to connect to a WideIO compatible DRAM memory. Each memory controller integrates one heater (H1,...,H4) and one thermal sensor (S1,...,S4).
- Four heaters (H5,...,H8) and three thermal sensors (S5,...,S7) in the bottom left corner, to emulate a quad-core processor (such as an ARM Cortex A9, for instance).

The thermal heaters are made of poly resistance and can dissipate up to 1W each. Each heater is independently controlled on the board via embedded software, while integrated thermal sensors are monitored in real time. Thermal sensor accuracy is $\pm 1^oC$ within the calibration temperature range (room temperature @27^oC), but with lower accuracy of up to $\pm 4^oC$ at 100^oC. Sensor resolution is 1^oC for the entire range.

III. Methods

In this paper we focus only on the steady-state thermal model. In the current section we formulate the thermal modeling problem.

A. Static gain matrix

The steady-state thermal model of the MAG3D chip can be obtained by applying a similar methodology as the one presented in [2] to the peculiar MAG3D Heater/Thermal sensor infrastructure. Given H heaters and S thermal sensors, the temperature field $T = (t_1,...,t_S)$ measured at the chip sensor locations due to the power $P = (p_1,...,p_H)$ dissipated by the heaters can be described by the following relation $T = T_{amb} + K \cdot P$ where K is the static-gain matrix. As in [2], in this work we obtain the K matrix by mean of measurements on the target chip. The necessary data are obtained with the following procedure. We apply a set of power vectors, composed by binary power stress (heater on/off) to the chip considering all the possible 2^H combinations. At the same time we store the corresponding temperature sensors readings after waiting the thermal response to be stabilized. The single power vector is maintained active for an amount of time enough to guarantee the steady-state condition. At the end of this procedure we obtain the power measurements matrix P_m and the temperature measurements matrix T_m. The elements of the T_m matrix are relative to T_{amb} such as $T_m = T - T_{amb}$. Finally we apply a least square optimization to compute the K matrix coefficients:

$$\hat{K} = \arg\min_K \|T_m - P_m \cdot K\|^2 \quad (1)$$

The static-gain matrix is a thermal resistance matrix since its coefficients have the units $(^oC/W)$. Thermal resistance is known to depend on the distance between the heat source and the sensing point [18]. Knowing the geometrical coordinates of the heaters and the temperature sensors placed on the chip it is possible to correlate the values of the K matrix with the reciprocal heater/sensor distance. The nature of this correlation is more clear by displaying the value of the coefficients of the K matrix versus the distance as showed in Fig.1b. From a practical point of view, knowing the analytical relation between the coefficients of the K matrix and the distance of the heat source from any location of the chip surface, enable to systematically calculate extra coefficients related to thermal sensing and heating sources positions not covered yet by the provided thermal sensors. More in detail, starting from the current K matrix and given the geometrical coordinates of the power source x_p and the thermal sensing point x_s, it is possible to learn a function that generates the new "virtual" coefficient of the K matrix as $K_v = f_K(x_p, x_s)$ To build the f_K function we use the approach described in [18]

$$f_K(x_p, x_s) = \frac{1}{((x_s - x_p)/a)^2 + 1} + \frac{b}{x_s - x_p} e^{\frac{\left(\ln \frac{x_s - x_p}{c}\right)^2}{d}} \quad (2)$$

We use a non linear least square to calculate the fitting parameters $\theta = \{a, b, c, d\}$ as

$$\hat{\theta} = \arg\min_\theta \|e_K\|_2^2 \quad (3)$$

where e_K is the error vector obtained by subtracting the coefficients of the original K matrix from those obtained using the f_K function.

B. Spatial adjustment

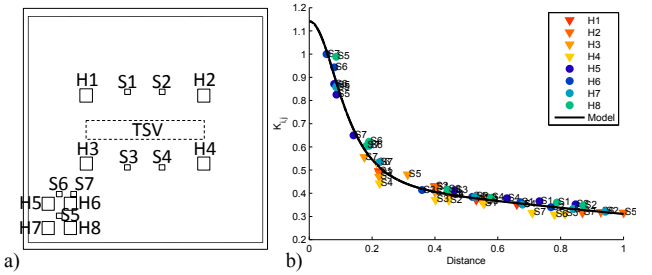

Fig. 1. (a) MAG3D floorplan; (b) Normalized function used to approximate the K matrix coefficients

By applying the previously described step to the MAG3D chip we obtain the resulting function showed in Fig.1b (Model) and it represents a good analytical approximation of the real coefficients. However, from the same figure, it is clear that this kind of "uniform" approximation neglects some interesting spatial properties of the coefficients. As instance, considering the value of the coefficients describing the iteration between $H1,...,H4$ and the sensors $S1,...,S4$; in Fig.1 can be noticed that even if they are at the same reciprocal distance, the corresponding coefficient value is quite different. This effect is mainly due to the different chip location where the heaters and the sensors are placed and so are subject to variability effects. Indeed in 3D-chips the floorplan is composed by spatial regions that involve different materials (i.e. TSVs). The coefficients of the K matrix promptly capture this behavior useful to describe accurately the temperature profile of the real silicon. Therefore, to improve the accuracy of the approximating function we need an "adjusting" factor able to consider also the effective position on the chip surface. The approach we propose is to appropriately weight the error k_{err} between the uniform approximation and the real coefficient. This weighting action is done using a nearest-neighbor approach, assuming that the properties of a new power source or sensor location are similar to the nearest existing one. Considering d_p as the distance of the power source and d_s as the distance of the sensing location from respectively the nearest known power source and temperature sensor, the adjusting factor can be defined as:

$$f_{ad} = K_{err} e^{-\frac{d_p + d_s}{\alpha}} \quad (4)$$

where α is a coefficient that regulates the spatial decaying of the adjusting term. Putting all together, the final function can be rewritten as $f_{K,ad} = f_K + f_{ad}$. In the results section (Section V) we will validate our modeling methodology with off-sample tests.

IV. Applications

In this section we discuss how these modeling strategies when combined with heaters and thermal sensors infrastructure, can be applied to augment the introspection of real 3D devices in the modeling of the thermal properties, estimating the real temperature and power consumption of not physically monitored HW components by knowing only their position.

A. Temperature to power

The following procedure describes how is possible to identify the power dissipated at unknown locations of the chip starting from the knowledge of the K matrix and the corresponding $f_{K,ad}$ function. We aim to calculate the power P_x of a further unknown unit in the chip that generates the measurable (by the available thermal sensors) steady state thermal response T_m. Given the position x_{p_x} of the new power source we start generating a new column entry in the original K matrix

$$K_{s,H+1} = f_{K,ad}(x_{p_x}, x_s) \quad s = 1, ..., S \tag{5}$$

obtaining the new matrix $K_N = \{K | K_{s,H+1}\}$. To calculate P_x we need to solve the problem $P = K^{-1} \cdot T$. We can use two approaches to obtain P_x.

1) Model1: The first approach "Model1" assumes that an external power gauges is available to monitor the full chip power consumption [10]. This translates in a constrained optimization problem where the complete chip power vector $P_T = \{P_m, P_x\}$ is the unknown. The optimization problem is:

$$\hat{P}_T = \arg\min_{P_T} \|K_N \cdot P_T - T\|^2 \quad s.t. \ \{P_T \geq 0, \ \sum_i p_i = P_{MAX}\}$$

2) Model2: A second approach "Model2" assumes that a limited number of power gauges are integrated on-chip to monitor the power consumption of a small subset of HW components, while others are not monitored [8] . This translates in the following least square optimization where the vector P_x is the only unknown:

$$\hat{P}_x = \arg\min_{P_x} \|K \cdot P_m + K_{S,H+1} \cdot P_x - T_m\|^2 \quad s.t. \ \{P_x \geq 0\}$$

B. Thermal sensors virtualization

The approximating function $f_{K,ad}$ can be used to infer the temperature at positions of the silicon surface not covered by any temperature sensors. Formally the problem in this case translates into the evaluation of the temperature T_x at the location x_{s_x} given K, the function $f_{K,ad}$ and the input power map P_m. To solve this problem first we evaluate the new row of the K matrix which correlates the existing power sources to the new virtual temperature sensor:

$$K_{S+1,h} = f_{K,ad}(x_{s_x}, x_h), \quad h = 1, ..., H$$

Finally, after reconstructing the new matrix $K_s = \{K | K_{S+1,h}\}$ the temperature at the unknown location x_{s_x} of the chip is:

$$T_x = \sum_{h=1}^{H} K_{S+1,h} \cdot P_{m,h} \tag{6}$$

In next section we evaluate the performance of the proposed application directly on the MAG3D chip.

V. Results

1) Test 1: In the first test we want to measure the robustness of the proposed methods toward the lack of heaters available on the chip. In our test chip there are $H = 8$ heaters and we emulate the scenario of obtaining the K_{red} matrix using only a reduced subset of heaters but maintaining the same number of temperatures sensors $(S = 7)$. The missing coefficients are thus recovered using the $f_{K,ad}$ function. Finally we validate the accuracy of this model comparing the results with the full K obtained using all the heaters. Initially, as described in section III-A, we calculate the reduced K_{red} matrix using 2^{H-n} power vectors and temperature vectors; n indicates the number of removed heaters. Successively, using the method showed in the section IV-A we calculate the missing parameters of the K_{red} matrix. In this case the approximating function used in (5) is obtained using the K_{red} matrix and so it has less accuracy. Finally we compare the obtained coefficients with the corresponding one in the original K matrix. In Fig.2 is showed the norm of the error calculated for the cases with $n = 1, .., 4$ missing heaters. Since the performance of the approximating function are strictly correlated to which heater is removed, for each of the four tests we calculated the average error considering all the *n-choose-H* combinations. In Fig.3 is showed the case *3-choose-8*. The results confirm that the adjusted approximating function behaves better in coefficient recovery than the uniform one since it can consider spatial variability. There is also a linear dependency between the error in the matrix coefficients and the number of removed heaters.

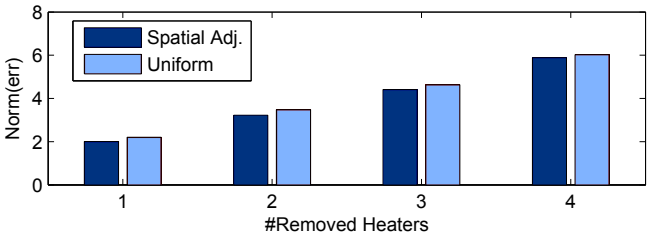

Fig. 2. Adjusted and uniform approximating function comparison

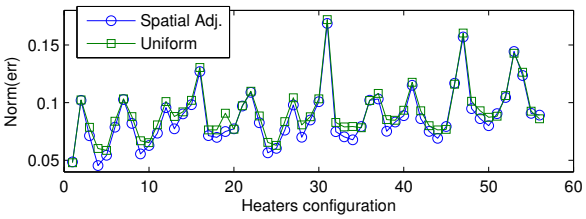

Fig. 3. Error comparison in the *3-choose-8* heater configuration test

2) Test 2: In a second test we want to measure the ability of the proposed methods in recognize the power dissipated in a unknown location of the chip. We use the approach described in Section IV-A. In particular we removed the heater H4 and we use its power trace as a reference for the method validation. Executing all the described steps using both the "Model1" and "Model2" approaches we obtain two identified power vectors compared in Fig.4. The real validation trace is indicated as "Actual". "Model2" clearly has better performance as also

confirmed in Fig.5 where are showed the residuals histograms.

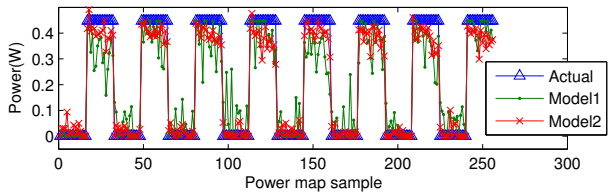

Fig. 4. Power traces recovered using two different methods to invert the K matrix

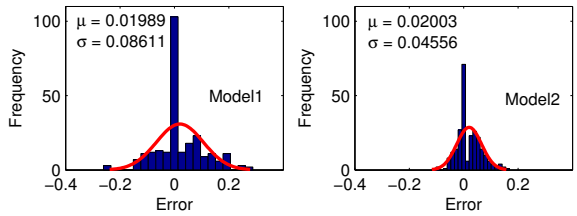

Fig. 5. Residuals of the Model1 and Model2 output obtained using the real heater power trace

3) Test 3: The final test evaluates the performance of the approximating function emulating virtual thermal sensors located at chip position not covered by real temperature sensors. The procedure is explained in section IV-B. In this particular test we removed the sensor S6 and used its real value as a validation trace. We calculate the error as the difference between the real sensor traces and the traces generated by two models: the "Full model" obtained using all the chip resources (and so even S6) and the recovered model (Rec. Model) obtained reconstructing the missing coefficients using the approximating function. In Fig.6 are compared the two models and it is possible to quantify the accuracy of the sensor virtualization approach. Using this procedure it is possible to

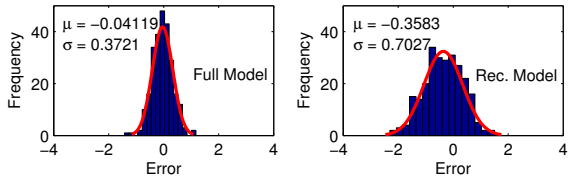

Fig. 6. Residuals of the Recovered model and the Full model output obtained using the real temperature trace

generate a detailed thermal map of the whole chip surface. Assuming to subdivide the chip surface in a MxL grid where each point is considered as a virtual thermal sensor point. Executing MxL times the procedure described in IV-B we can generate a larger K matrix K_l composed by (MxL)xH coefficients, able to calculate the temperature at each point of the chip surface. Setting M=L=64 we reconstructed the steady-state thermal field of the chip corresponding to a generic power map (i.e. H1,H4 and H5 = on). The resulting thermal map is showed in Fig.7.

VI. Conclusion

In this paper we estimated the static thermal model of a real 3D test chip. We exploit the on chip heaters and thermal sensors to apply accurate power map to the silicon and to gather the corresponding temperature filed. This allows to accurately calculate the steady state thermal model. Moreover we show several methods enabling more possibilities like recovering the power dissipated by unknown units in the chip or reconstructing the temperature field of the whole chip starting from limited temperature and power information. The accuracy of the proposed techniques is improved considering in the models the chip variability effects. Finally we prove the effectiveness of our methods with off-sample validation techniques.

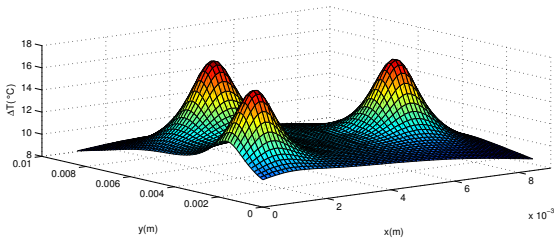

Fig. 7. Full chip thermal map

Acknowledgments

This work was supported by the FP7 ERC project MULTITHERMAN (g.a. 291125).

References

[1] I. Arnaldo et al., Fast and scalable temperature-driven floorplan design in 3d mpsocs. In 13th Latin American Test Workshop (LATW), 2012.
[2] F. Beneventi, A. Bartolini, and L. Benini. Static thermal model learning for high-performance multicore servers. *ICCCN 2011*, pp. 1–6.
[3] E. Beyne. The rise of the 3rd dimension for system intergration. In *IITC 2006*, pp. 1–5.
[4] James Burns. TSV-Based 3D integration. In *Three Dimensional System Integration*, pp. 13–32. Springer, 2011.
[5] F. Clermidy, F. Darve, D. Dutoit, W. Lafi, and P. Vivet. 3D embedded multi-core: Some perspectives. In *DATE 2011*, pp. 1–6.
[6] R. Cochran and S. Reda. Spectral techniques for high-resolution thermal characterization with limited sensor data. In *DAC 2009*, pp. 478–483.
[7] R. Cochran, A. N. Nowroz, and S. Reda. Post-silicon power characterization using thermal infrared emissions. In *ISLPED 2010*, pp. 331–336.
[8] H. David et al., RAPL: memory power estimation and capping. In *ISLPED 2010*, pp. 189–194.
[9] K. Ganeshpure and S. Kundu. Reducing temperature variation in 3d integrated circuits using heat pipes. In *ISVLSI 2012*, pp. 45–50, 2012.
[10] J. Howard and et al., A 48-core IA-32 message-passing processor with DVFS in 45nm CMOS. In *ISSCC 2010*, pp. 108–109.
[11] W. Huang et al., Hotspot: A compact thermal modeling methodology for early-stage VLSI design. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol.14, no.5, pp.501–513, 2006.
[12] SST JEDEC. Association. *Wide I/O SDR DRAM Standard, JESD229*, 2012.
[13] H. Oprins et al., Steady state and transient thermal analysis of hot spots in 3D stacked ICs using dedicated test chips. In *SEMI-THERM 2011*, pp. 131–137.
[14] Zhenyu Qi et al., Temperature-to-power mapping. In *ICCD 2010*, pp. 384–389.
[15] M. M. Sabry, A. Sridhar, J. Meng, A. K. Coskun, and D. Atienza. Greencool: An energy-efficient liquid cooling design technique for 3-d mpsocs via channel width modulation. *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol.32, no.5, pp.524–537, 2013.
[16] D. Sekar et al., A 3d-ic technology with integrated microchannel cooling. In *IITC 2008*, pp. 13–15.
[17] A. W. Topol et al., Three-dimensional integrated circuits. *IBM Journal of Research and Development*, vol.50, no.4.5 pp.491–506, 2006.
[18] C. Torregiani et al., Compact thermal modeling of hot spots in advanced 3D-stacked ICs. In *EPTC 2009*, pp. 131–136.
[19] C. Weis, I. Loi, L. Benini, and N. Wehn. An energy efficient DRAM subsystem for 3D integrated SoCs. In *DATE 2012*, pp. 1138–1141.
[20] Y.Zhang et al., Chip level thermal profile estimation using on-chip temperature sensors. In *ICCD 2008*, pp. 432–437.
[21] D. Dutoit et al., A 0.9 pJ/bit, 12.8 GByte/s WideIO memory interface in a 3D-IC NoC-based MPSoC, In *VLSIT 2013*, pp. C22-C23.