# Positioning of Thermal Via Regions for Reducing Hotspot Temperature in 3D ICs

C. Maj*, M. Galicia*, P. Zając* and A. Napieralski*

* Department of Microelectronics and Computer Science, Lodz University of Technology
Wolczanska 221/223, building B18, 90-924 Lodz, Poland
e-mail: cezary.maj.1@p.lodz.pl; phone: +48 42 631 27 24; fax: +48 42 636 03 27

## ABSTRACT

3D stacking of integrated circuits seems to be a main trend for increasing the processor performance. However, It results in power density increase inducing additional thermal issues that have to be overcome. In this paper we analyze the thermal behaviour of eight-core processor based on Intel Haswell architecture implemented as a 2D chip and as a 3D architecture with two layers. We implement special via regions in 3D structure in various configurations and we compare results to that obtained for 2D structure. We show that the peak temperature can be significantly reduced and locations of via regions notably affects temperature distribution inside the chip.

***Keywords***: processor design, 3D ICs, thermal simulation, hotspot, floorplanning

## 1 INTRODUCTION

The increase of the processor performance is the most desired result of technology evolution. As today's technology is pushed to its limits, some other methods has to be developed to uphold performance increase. One of the promising method is 3D stacking that should be commonly used in commercial products in near future. This idea was initially proposed for integrating memory and processor on the same die to decrease memory access latency [1]. However, it may be also used in reorganization of processor layout. In example an eight-core processor on two layer chip may be designed in following ways. The simplest approach is to implement two the same quad-core processors on each layer. The main advantage is that the base floorplan has to be slightly changed to implement interconnections between layers. Another simple approach assumes that first layer consists of all cores with cache memory and second layer consist of memory controller and uncore block. The last approach is the most ambitious. Each block of one core unit may be separated and placed on both layers. Its main advantage is the possibility of interconnection shortening but the entire floorplan of whole chip has to be reorganized.

The main problem in 3D stacked chips is increased power dissipation inside this chip [2]. In case of traditional heatsink cooling, the heat is transferred in vertical direction. If layers are multiplied, one need to remove more heat from the same area. Note that the layer thickness is much smaller than length/width of the package, so we can compare it to the 2D package dissipating the heat equal to the sum of heats dissipated in all 3D layers. A few solutions have been proposed to overcome this problem [3][4]. Nonetheless, the use of thermal vias [5] seems to be the most interesting. In general through-silicon vias (TSV) were designed to provide electrical connection between chip layers. However, the material used for its fabrication (copper) has much higher thermal conductivity than silicon. Thus, TSVs significantly improve the heat flow between chip layers resulting in lower maximal temperature in the die. It seems natural to implement TSVs also to provide lower thermal resistance. The optimal placement of TSVs is a subject of current research [5]. The best idea would be to place vias uniformly on the entire surface or even adjust it to the power density. However, from the designer point of view it would be necessary to reorganize the chip layout. Likely, it would elongate critical interconnection leading to latency increase what is unacceptable from the performance point of view. Therefore, the vias placement should not disrupt the optimized layout and should take into consideration critical interconnections. In our investigations we take into consideration a 22 nm technology Intel Haswell processor [6], concretely the high performance model i7-5960X (Fig. 1). Next, we analyze the hypothetical 3D implementation of this processor in two layers. Finally, following above mentioned considerations, we analyze 3D structure including regions with thermal vias. All these implementations are compared in terms of temperature distribution for the same power data. Moreover, a few localization design of TSVs regions are analyzed in order to provide the highest temperature reduction and the most uniform temperature distribution.

## 2 ANALYZED PROCESSOR

### 2.1 Real 2D Processor

As a baseline processor we take the 22 nm technology Intel Haswell i7-5960X. Its properties are presented in the Table 1 and its floorplan is shown in Fig. 1 (left).

### 2.2 Hypothetical 3D Implementation

There are many possibilities in designing 3D version of the processor shown in Fig. 1 (left). However, we want to avoid significant changes in original design. Therefore, the simplest design that requires least changes in layout is presented in Fig. 1 (right). Cores and L3 cache are placed in one layer, so the optimal interconnections between these blocks remain undisrupted. Queue/Uncore/I/O and Memory Controller are located on the second layer. The core layer is placed closer to the heatsink as it generates more heat. This shortens the distance the heat must pass and reduces heating of the second layer. Moreover, this solution is also better in terms on interconnections. Memory Controller together with I/O block requires connections to the package pins and they will be shorter in such configuration.

| Model | i7-5960X |
|---|---|
| Cores | 8 |
| Technology | 22 nm |
| CPU clock rate | 3.0 GHz |
| Width/height | 20.2 mm/17.6 mm |
| Size | 355.5 mm$^2$ |
| Thermal design | 140 W |

Table 1: Processor parameters

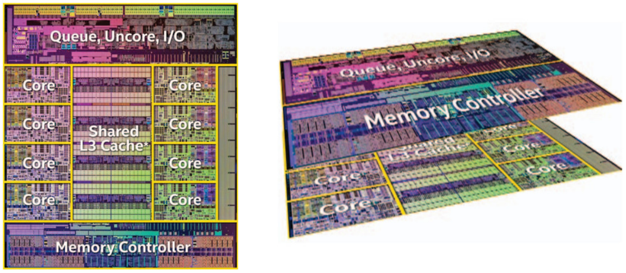

Figure 1: Intel Haswell i7-5960X die photo (left) and its hypothetical 3D design (right)

## 2.3 3D Implementation with Via Regions

The 3D structure must exhibit worse thermal behaviour than the 2D structure due to increased power density. Therefore, to minimize this effect in our previous works [7] we proposed an alternative design with special regions with vertical thermal vias. These via regions (VR) are located next to the cores (Fig. 2) to minimize the impact on original design. Moreover, vias go through both active layers (Fig. 3) forcing some minor changes to the design of Queue/Uncore/I/O and Memory Controller. But these changes are more easily applied than modification of strictly optimized core layer. The density of vias inside regions is set to 25% and regions width is equal to 0.5 mm. Five floorplans which include thermal via regions placed in various locations are compared. The first one is the basic one from our previous work (V0) [7]. Because the temperature distribution was not optimal and high gradient was observed, we put additional via regions in order to obtain higher peak temperature reduction and its more uniform distribution. Fig. 2 shows the layouts of all analysed cases. Note that all cases will be compared also to the case without vias to illustrate benefits from via regions implementation.

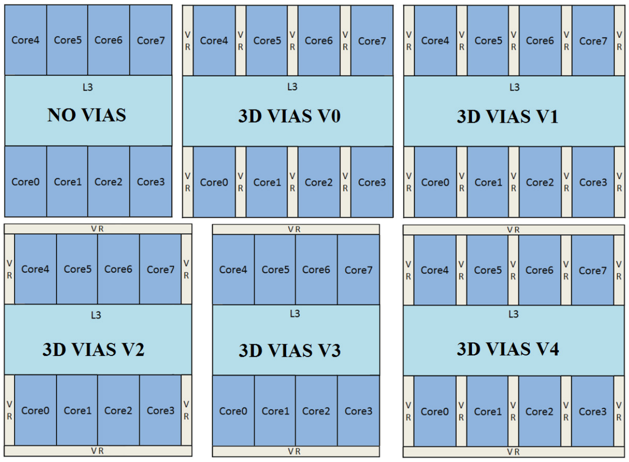

Figure 2: Thermal via regions (VR) placement (vias are present in all chip layers) for all analysed cases.

# 3 SIMULATION METHODOLOGY

The power modelling methodology combines a full-system simulator gem5 [8] and McPAT [9] power model. Gem5 executes SHA cryptographic benchmark and return the detailed information about accesses to particular processor units. These data are an input for McPAT which computes power dissipation in each processor unit for 22 nm technology. Then, using the technological parameters McPAT calculates the static power dissipation. The same data are used for each of simulated cases mentioned in previous section. The detailed methodology was described in [10]. Among particular processor units was the same for all cases mentioned in the previous section. This, it is possible to meaningfully compare these six designs in terms of thermal behavior. Hotspot [11] thermal modelling tool was used for thermal simulation analyzed chips. In case of 3D chip a special extension was applied [12]. Package parameters for both cases are presented in Table 2.

As in real cores the power dissipation is not uniform, our core model is divided into two blocks: the first one with very high power dissipation (up to 1W/mm$^2$) and the second one with less active core areas like L2 cache. For more details please refer to [7]. The simulated chip is built from the following layers: chip, thermal interface material (TIM), heat spreader and heat sink. 3D structure contains another chip and TIM layers. Thermal vias goes through active silicon layers and TIM layers between them as shown in Fig. 4.

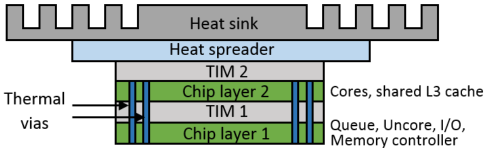

Figure 3: Simulated 3D processor package (not to scale).

| Common package parameters | |
|---|---|
| Chip layer thickness | 0.15 mm |
| TIM thickness | 0.08 mm |
| Silicon thermal conductivity | 100 W/mK |
| TIM thermal conductivity | 4 W/mK |
| Ambient temperature | 40ºC |
| Convection resistance (assuming liquid cooling) | 0.015 K/W |
| Heat spreader thermal conductivity | 400 W/mK |
| Heat spreader thickness | 1 mm |
| Heat spreader size | 30 x 30 mm |
| 3D package parameters | |
| Inter-layer TIM thickness | 0.08 mm |
| Inter-layer TIM thermal conductivity | 4 W/mK |
| Thermal via density | 25% |
| Via thermal conductivity | 400 W/mK |

Table 2: Package parameters

## 4 RESULTS

Fig. 4 shows the temperature map of bottom layers (this one farther from the heatsink) as they have higher temperatures. Note, the floorplan shown in figures coresponds to the top layer. For all cases the power distribution in each processor unit is the same. Fig. 5 shows the chip temperature across cross-section where the peak temperature occurs (depends on analyzed case).

The results for basic 3D layout (V0) showed that via regions reduces the peak temperature from 351.3 K (no vias case - NV) to 345.6 K (reduction of 5.7 K). It is almost the same level as for classical 2D implementation [7]. However, the temperature distribution showed that via regions induced unwanted temperature gradient. Moreover, high temperature on the right side of the chip proves the lack of via region in that location. Therefore, case V1 eliminates this disadvantage. It can be seen that now peak temperature occurs in one of the middle cores and is 344.5 K (6.8 K less than NV). The core on the edge is now much colder by about 2.5 K. Nevertheless, the gradient remains high and the temperature difference in the area of two neighbouring cores is about 7.5 K. In terms of thermal stresses it is undesirable because this worsens chip reliability and reduces its lifetime. Thus, we tried to reduce the gradient and put via regions only outside the cores (case V2) because most of gradient occurs between the cores. The results shows that the temperature distribution in the middle of the core zone is almost uniform. As expected, the peak temperature slightly increased and now is 346.8 K (2.3 K more than case V1 and 4.5 K less than NV). Note that we analyze gradient only in horizontal direction. In vertical direction the gradient is inevitable due to L3 cache existance between core zones. It is illustrated in Fig. 6. Moreover, one can observe high gradient on the edges (almost 13 K of difference) due to existence of via regions there. Thus, in the next case (V3) these via regions on the edges were removed. The results show that the temperature distribution is almost uniform and 1 K of difference is cause by the assymetry of cores placement in horizontal direction. It has to be emphasized that the peak temperature is almost the same as for case V2 and is 347 K. The last case (V4) combines all previous cases, so via regions are located in all possible locations. As expected, the peak temperature is the lowest from all cases and is 342 K (9.3 K less than NV). However, the gradient is very high. The temperature difference for cores on the edge is about 8.5 K and for these in the middle is about 6.5 K.

All results are summarized in Table 3. Note that the maximal temperature difference is calculated for the closest local minimal and maximal temperatures. Therefore, the highest value occurs for case V2 but it does not mean that the gradient is the highest because the distance between analyzed two points is much larger than in other cases. It is clearly visible in Fig. 6 where temperature maps in 3D charts for cases V2 and V4 are presented. The slope is much steeper for case V4. Chip area penalty is the parameter that says how much more area of silicon wafer is needed to implement via regions. Note that case V2 requires least additional area for via regions, even it provides most advantageous result in terms of peak temperature and thermal gradient.

| Case | Peak temperature [K] | Temperature difference [K] | Chip area penalty |
|---|---|---|---|
| NV | 351.3 | 3 | - |
| V0 | 345.6 | 9.26 | +9.8% |
| V1 | 344.5 | 11.3 | +12.2% |
| V2 | 346.8 | 13.6 | +5.7% |
| V3 | 347 | 1.5 | +11.1% |
| V4 | 342 | 9.3 | +19.1% |

Table 3: Summarized results

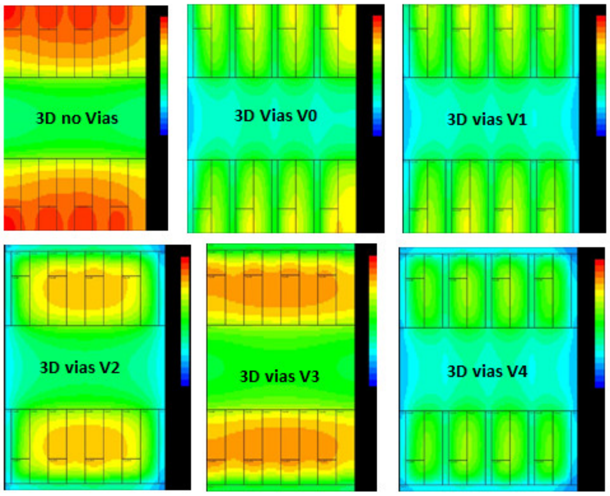

Figure 4: Temperature map for all analysed cases of 3D implementations.

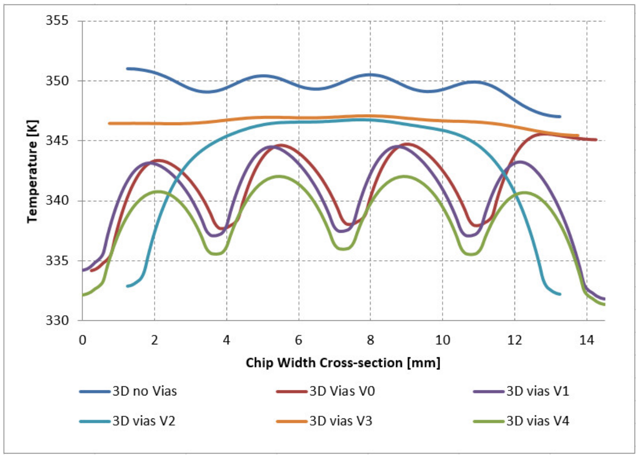

Figure 5: The chip temperature profile across chip cross-section for all analysed cases.

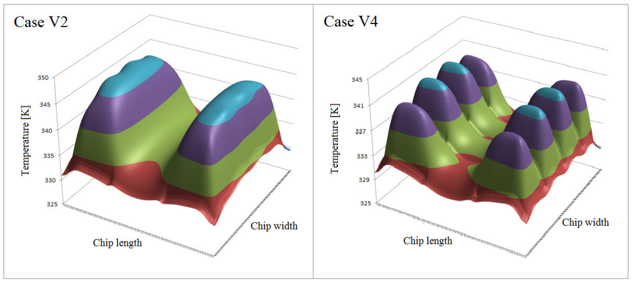

Figure 6: Temperature distribution in whole chip for cases V2 (left) and V4 (right).

## 5 CONCLUSIONS

In this paper the impact of localized thermal vias in 3D processor was analyzed. The thermal vias were placed in regions that should not induce major changes in optimized processor design. It has been shown that this idea may provide the temperature reduction, such that the peak temperature is the same as in classical 2D design. Moreover, the proper placement of via regions significantly reduces the drawback of this sollution which is the higher temperature gradient. It was shown that via regions between the cores are not advantageous in temperature distribution point of view. It seems that the least harmful solution is the placement of one via region close to the each core zone. The number of locations with high thermal gradient is reduced and peak temperature reduction is still notisable. Moreover, the thermal gradient is then lower than in case without vias. Summarizing, via thermal regions may be a good solution for overheating problem and proper localization of these regions may significantly reduce peak temperature together with smoothing of thermal gradient.

## REFERENCES

[1] G.H. Loh, "3D-stacked memory architectures for multi-core processors". In Inter. Symposium on Computer architecture (ISCA'08), pp.453-464, (2008).

[2] K. Puttaswamy, G. H. Loh, "Thermal Herding: Microarchitecture Techniques for Controlling Hotspots in High-Performance 3DIntegrated Processors", Proceedings of the 2007 IEEE 13th International Symposium on High Performance Computer Architecture, p.193-204, February 10-14, 2007.

[3] A. K. Coskun , J. L. Ayala , D. Atienza , T. Simunic Rosing, Y.Leblebici, "Dynamic thermal management in 3D multicore architectures", Proceedings of the Conference on Design, Automation and Test in Europe, April 20-24, 2009, Nice, France.

[4] C. Zhu, Z.P. Gu, L. Shang, R.P. Dick and R. Joseph, "ThreeDimensional Chip-Multiprocessor Run-Time Thermal Management", IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, vol. 27, no. 8, pp. 1479-1492, Aug. 2008.

[5] B. Goplen, S. Sapatnekar, "Thermal via placement in 3D ICs", Proceedings of the 2005 international symposium on Physical design, April 03-06, 2005, San Francisco, California, USA

[6] P. Hammarlund et al., "Haswell: The Fourth-Generation Intel Core Processor", Micro, IEEE, vol.34, no.2, pp.6, 20, Mar.-Apr. 2014

[7] P. Zajac, M. Galicia, C. Maj, A. Napieralski, "Investigation of Localized Thermal Vias for Temperature Reduction in 3D Multicore Processors", Mixed Design of Integrated Circuits & Systems (MIXDES), 2015 22nd International Conference, 25-27 June 2015.

[8] N. Binkert et al., "The gem5 simulator," ACM SIGARCH Computer Architecture News, vol. 39, no. 2, p. 1, Aug. 2011

[9] S. Li et al., "Mcpat: An Integrated Power, Area, and Timing Modeling Framework for Multicore and Manycore Architectures," Proc. Symp. IEEE Micro, pp. 469-480, 2009

[10] P. Zajac, M. Galicia, C. Maj, A. Napieralski, "Optimizing temperature distribution in modern processors through efficient floorplanning", 20th International Workshop on Thermal Investigations of ICs and Systems (THERMINIC), pp.1-6, 24-26 Sept. 2014

[11] K. Skadron et al., "Temperature-Aware Microarchitecture." In Proceedings of the 30th International Symposium on Computer Architecture, June 2003.

[12] J.Meng, K. Kawakami, A. K. Coskun, "Optimizing energy efficiency of 3-D multicore systems with stacked DRAM under power and thermal constraints", Proceedings of the 49th Annual Design Automation Conference, June 03-07, 2012, San Francisco, California