Thermal Modelling of 3D Multicore Systems in a Flip-Chip Package

Kameswar Rao Vaddina[*†], Tamoghna Mitra[+], Pasi Liljeberg[†] and Juha Plosila[†]
[*]Turku Center for Computer Science (TUCS), Joukahaisenkatu 3-5, 20520 Turku, Finland
[†]Department of Information Technology, University of Turku, Turku, Finland
[+]Chemical Engineering Department, Åbo Akademi University, Turku, Finland
Email: {vadrao, pakrli, juplos}@utu.fi and tmitra@abo.fi

***Abstract*—Three-dimensional (3D) technology offers greater device integration, reduced signal delay and reduced interconnect power. It also provides greater design flexibility by allowing heterogeneous integration. In this work, a 3D thermal model of a multicore system is developed to investigate the effects of hotspot, and placement of silicon die layers, on the thermal performance of a modern flip-chip package. In this regard, both the steady-state and transient heat transfer analysis has been performed on the 3D flip-chip package. Two different thermal models were evaluated under different operating conditions. Through experimental simulations, we have found a model which has better thermal performance. The optimal placement solution is also provided based on the maximum temperature attained by the individual silicon dies. We have also provided the improvement that is required in the heat sink thermal resistance of a 3D system when compared to the single-die system.**

I. Nomenclature

h_{eff}	= Effective heat transfer coefficient of the heat sink base (W/m^2K)
K	= Thermal Conductivity (W/mK)
T_A	= Ambient Temperature (°C)
T_J	= Junction Temperature (°C)
R_{JA}	= Junction-to-Ambient thermal resistance (°C/W)
Q	= Power dissipation that produced the change in the junction temperature (W)

II. Introduction

As technology scales down and power density increases, a lot of factors like power dissipation, leakage, data activity and electro-migration contribute to higher temperatures, larger temperature cycles and increased thermal gradients all of which impact multiple failure mechanisms [1]. This increase in temperature, increases interconnect delay due to the linear increase in electrical resistivity. These delay variations pose significant reliability problems with already dense interconnect structures. In order to overcome the problems associated with the interconnects and the limits posed by the traditional CMOS scaling, three-dimensional (3D) integrated circuits has been proposed. 3D integrated circuits take advantage of dimensional scaling approach and are seen as a natural progression towards future large and complex systems. They increase device density, bandwidth and speed. But on the other hand, due to increased integration, the amount of heat per unit footprint increases, resulting in higher on-chip temperatures and thereby degrading the performance and reliability of the system. In this case, heat sinks need to be very efficient in transferring the internally generated heat to the ambient. Although there is a dearth of design and layout tools for 3D technology, there is a significant amount of effort going on in that direction.

The ever expanding market for consumer electronics is driving innovation in packaging technology leading to newer packages which are smaller, more thermally efficient and cost effective at the same time. The technology related to wafer level packaging and 3D integration has recently outpaced ITRS roadmap forecasts [1]. One of the fastest growing packaging architectures is the wafer level packaging (WLP). It offers lower cost, improved electrical performance, lower power requirements and smaller size. Although several architectural variations are available, in this paper we will be discussing only the flip-chip packaging. The ITRS report projects that the power density for 14nm technology node will be greater than 100 W/cm^2 and the junction-to-ambient thermal resistance will be less than 0.2°C. It is very important to keep the thermal resistance at bay as this may increase the package cost and the overall cost of the product.

Guoping et al. [4] [5] have done thermal modelling of multicore systems and have investigated the effects of CPU power level, local hotspot power density, hotspot location and hotspot size on its thermal performance. But they stopped short of extending their work to 3D multicore systems. Ankur et al., [7] have proposed an analytical and numerical modelling of the thermal performance of three-Dimensional Circuits. In this paper we have chosen to model a 3D multicore system in a modern flip-chip package which is used mostly for high-performance processors. We have started our study with thermal modelling of a multicore processor and have investigated the effects of hotspots and their locations on the thermal performance of the package. We then proceeded to work on the 3D multicore systems. Due to the lack of space, only results pertaining to the 3D modelling are presented in this paper.

III. Futuristic virtualization platform

With the advent of cloud computing the systems of the future will become very complex with possibly thousands of cores running in parallel on a single silicon die. All of those cores could be tightly packed to form a data center on a chip which works on Cloud on a Chip (CoC) [2] paradigm.

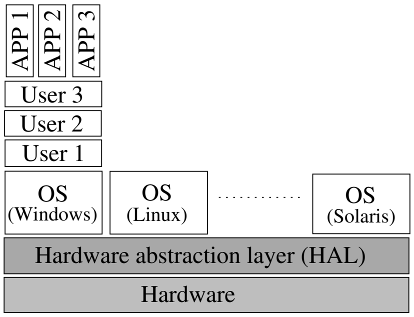

Fig. 1. Futuristic virtualization platform.

Virtualization platforms like the one shown in Fig. 1 can be an ideal solution for cloud computing. The hardware abstraction layer (HAL) is a small piece of software which interacts with the naked hardware and runs on top of it. Intel calls this hardware abstraction layer as Hypervisor, Microsoft calls it as Hyper-V and other vendors call it as Virtual Machine Monitor (VMM). There will be multiple operating systems running on the hardware simultaneously. Multiple users will be logged into those operating systems running multiple applications. The hardware abstraction layer provides access to the hardware resources and make them visible to the guest operating systems. The guest operating systems may not need to know the existence of other operating systems running in parallel. This increases the system robustness and stability. The time to deploy and debug new operating systems and applications without jeopardizing existing ones is a feature inherent to this technology. Such futuristic virtualization platforms would suffer from immense thermal challenges and needs dynamic thermal management techniques to be deployed.

IV. Flip-Chip package

Although IBM's Ball Grid Array packages have been in use since the 1970's, recent advances in packaging technology have lead to Flip-Chip Ball Grid Array (FCBGA) packages being extensively used. FCBGA allows for much higher pin count than the other package types by distributing the input-output signals through the entire die rather than being confined to the chip periphery. In an FCBGA the die is mounted upside-down (flipped) and connects to the package balls (lead-free solder bumps) via a package substrate.

The cross-sectional view of a modern 3D flip-chip package is shown in the Fig. 2 whose primary consideration will be its ability to transfer heat from the silicon die to the ambient. Unlike the traditional wire-bonding technology, the electrical connection of a face-down (or flipped) integrated circuit onto the substrate is done with the help of conductive bumps on the chip bond pads. The conductive bumps are initially deposited on the top-side of the die during the fabrication process. It is then flipped over so that its top side faces down, and aligned with the matching pads on the substrate. The solder is then flown to complete the interconnection. The advantages of flip-chip interconnect include reduced signal inductance, power/ground inductance, and package footprint, along with higher signal density [9].

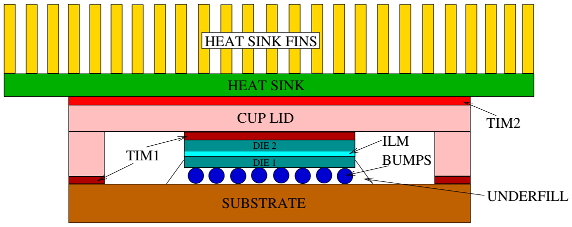

Fig. 2. Cross-Sectional view of a modern 3D Flip-Chip package.

V. Thermal modelling and Analysis

The high operating temperature of a semiconductor device, caused by the combination of device power density and ambient conditions is an important reliability concern. Instantaneous high temperature rises in the devices can possibly cause catastrophic failure, as well as long-term degradation in the chip and package materials, both of which may eventually lead to system failure [9]. Most modern flip-chip devices are designed to operate reliably with a junction temperature falling under a certain range. To ensure that the package can perform well thermally under this range a thermal model is simulated and tested. This thermal model can then be used to gauge the reliability of the package. This shortens the package development time and also provides an important analytical tool to evaluate its performance under different operating conditions.

We have developed a thermal model of the modern flip-chip package using a commercial tool called COMSOL. It is a finite element based multiphysics modelling and simulation software. Our simulations are based on the heat transfer module of COMSOL multiphysics package. The size of the silicon die 1 and 2 is 20 mm x 20 mm x 0.6 mm which is being mounted on to the substrate of size 50 mm x 50 mm x 1.44 mm. The layers of silicon die are separated by an interlayer material whose thickness is around 0.02 mm. The cup lid which acts as the heat spreader and whose thermal conductivity is very high is placed on top of the silicon die. The thermal interface material (TIM1) which is some sort of a thermal grease and has very good adhesive properties is being used as the filler material in between the heat spreader and the silicon die. The heat sink base of size 100 mm x 100 mm x 5 mm is being used. A vapour chamber is used as the heat sink base and the detailed assumptions can be found in [4]. Instead of including the heat sink fins in our computational model, we have used an effective heat transfer coefficient (h_{eff}) as a boundary condition on the heat sink [5]. Other assumptions related to the geometry of the package and its components, material properties (like thermal conductivity, density and specific heat capacity) and the boundary conditions are taken from the literature [1] [3] [4] [5]. Some important model configuration parameters are represented in the tabular format as shown in Table 1. The parameter Q, which is the heat generated per unit volume is applied to the silicon die. The boundary condition for the substrate layer is assumed to be convective and the sides of the package are assumed to be adiabatic.

TABLE I
MODELLING PARAMETERS [1] [3] [4] [5].

MODEL CONFIGURATION	PARAMETERS	INPUT DATA
Boundary condition	T_{Amb} (°C)	25
	$h_{eff}(W/m^2K)$	840
Heat Sink Base [5]	Size (mm)	100x100
	t_{base} (mm)	5
TIM2	t_{TIM2} (mm)	0.1
	k_{TIM2} (W/mK)	3
Cup Lid (heat spreader)	Size (mm)	50x50
	t_{Lid} (mm)	2
	k_{Lid} (W/mK)	600
TIM1	t_{TIM1} (mm)	0.1
	k_{TIM1} (W/mK)	8
Silicon Die 1 and 2	Size (mm)	20x20
	t_{Die} (mm)	0.6
	k_{Die} (W/mK)	90
Interlayer Material	t_{ILM} (mm)	0.02
	k_{ILM} (W/mK)	4
Lead bumps and Underfill	k_{UF} (W/mK)	1
	t_{UF} (mm)	0.65
Substrate	Size (mm)	50x50
	t_{Sub} (mm)	1.44
	k_{Sub} (W/mK)	17
Boundary condition	$h_{Sub}(W/m^2K)$	10

A. Modelling interlayer material

Three effective thermal conductivities are used for the lead solder bumps/underfill layer, substrate layer and the interlayer material (ILM) respectively. The interlayer material in between the silicon dies is modelled as a homogeneous layer in our thermal model. We assumed a uniform through-silicon-via (TSV) distribution on the die and obtained the effective interlayer material resistivity based on the TSV density (d_{TSV}) values [3], where d_{TSV} is the ratio of total TSV's area overhead to the total layer area. Coskun et al. [3] have observed that even when the TSV density reaches 1-2%, the temperature profile of the silicon die is only limited by a few degrees, thus justifying the use of homogeneous TSV density in our thermal model. According to the current TSV technology [8], the diameter of each via is 10μm, and the spacing required around the TSV's is assumed to be around 10μm [3]. For our experiments we have assumed around 8 via's/mm^2, that is around 3200 vias spread across the 400 mm^2 area of the silicon die. Hence the TSV density is around 0.062$\%$ and the resistivity of the interlayer material is around 0.249 mK/W (i.e. thermal conductivity = 4.016 W/mK) [3].

B. Junction temperature and thermal resistance for a 3D system

The two most important thermal parameters for any semiconductor device are the junction temperature (T_J) and thermal resistance (R_{JX}). The junction temperature is usually the highest temperature on a silicon die, whereas the thermal resistance is quantified as the rate of heat transfer between two layers in a package. The junction-to-ambient thermal resistance (R_{JA}) which is a measure to evaluate the thermal performance of a flip-chip package is determined from equation (1).

$$R_{JA} = \frac{T_J - T_A}{Q} \qquad (1)$$

The single-valued junction-to-ambient thermal resistance which has been used traditionally to describe the thermal characteristics of a silicon die is not sufficient enough to describe the thermal performance of a 3D system, due to the presence of multiple heat sources and multiple thermal resistances. Hence, Ankur et al. [7] have suggested a matrix representation for the junction-to-ambient thermal resistance. In this regard R_{ij} represents the temperature rise in the *i*th layer per unit heat dissipation in the *j*th layer. This is represented in the equation (2).

$$R_{ij} = \frac{\theta_i}{Q_j} \qquad (2)$$

Where, θ_i is the temperature rise above ambient of the *i*th node and Q_j is the heat generated at the *j*th node. The equation (2) can be rewritten as follows.

$$R_{ij} = \frac{T_i - T_A}{Q_j} \qquad (3)$$

Where, T_i is the junction temperature of the *i*th layer. So, for a simple two-die stack, where one layer is the processing layer (denoted by subscript '*p*') and the other a memory layer (denoted by subscript '*m*'), we have 4 different thermal resistance values namely $R_{pp},\ R_{pm},\ R_{mp}\ and\ R_{mm}$ and the junction-to-ambient thermal resistance can be represented as shown below.

$$R_{JA} = \begin{bmatrix} R_{pp} & R_{pm} \\ R_{mp} & R_{mm} \end{bmatrix}$$

VI. EXPERIMENTAL RESULTS

We have built a generic two-die stack in a flip-chip package using COMSOL. The layer where the hotspot is generated is considered as a processing die and the other layer is considered as the memory die in our simulations. In the first instance (model-I) the processing die is placed near the substrate, and the memory die is placed next to the heat spreader and the heat sink. In the second instance (model-II) the memory die is placed near the substrate and the processing die is placed near the heat spreader and sink. We have assumed that the total power consumed by both the processing layer and the memory layer is 100 W. Guoping Xu [5] has varied the size of the hotspot from 0.5 mm to 2 mm in his work related to the thermal modelling of multicore systems. In our work the power density of the hotspot which is being generated at the center of the multicore processing layer is fixed at 100 W/cm^2 and the dimensions are fixed at 1mm x 1mm x 0.6mm. We

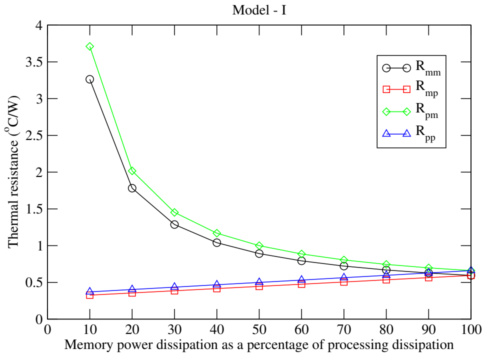

Fig. 3. Thermal resistance measurements for both the dies in model-I at steady-state.

have performed both the steady-state and transient heat transfer analysis on the flip-chip package.

A. *Steady-state heat transfer analysis*

In the steady-state the heat generated by the memory and the processing layer is equal to the heat leaving the flip-chip package. During the measurements we have assumed that the power is gradually applied to the chip until the chip has reached the maximum working temperature (i.e. steady state). We have then measured thermal resistance which is the reluctance of the die to transfer heat when it reaches steady state. Fig. 3 and 4 show different thermal resistance plots for the dies in both the models. They are plotted against the memory power dissipated (as a percentage of processing power dissipation). It can be clearly seen that the overall thermal performance of model-II is much better than that of model-I. It can also be noted that, when both the layers consume equal amount of power then there is not much difference in the thermal resistance values in both the models. That is, the stacking order of the silicon dies does not influence the thermal resistance values.

When both the layers are consuming equal amount of power, then in model-II, it can be noted that there is no difference in the thermal resistance values of the processing and memory layers even though a hotspot is present in the processing layer. This shows that the heat sink is efficient in removing the heat generated by the hotspot, thereby maintaining constant thermal resistance values.

Fig. 5 shows the maximum temperature attained on the processing and the memory die for both models at steady state. The maximum temperature is plotted against the memory power dissipation (as a percentage of processing power dissipation). In the case where the memory die consumes around 10% of the processing die power, it can be observed that the difference in the maximum temperature of memory and processing die layers is around 4°C for model-I and 0.3°C for model-II. This goes on to say that the model-II is the optimized one which places the most heat generating layer, i.e. the processing layer near the heat sink for efficient heat transfer to the ambient.

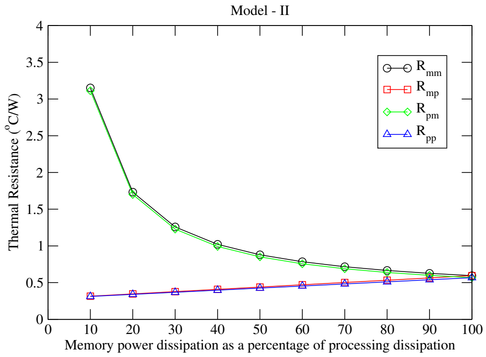

Fig. 4. Thermal resistance measurements for both the dies in model-II at steady-state.

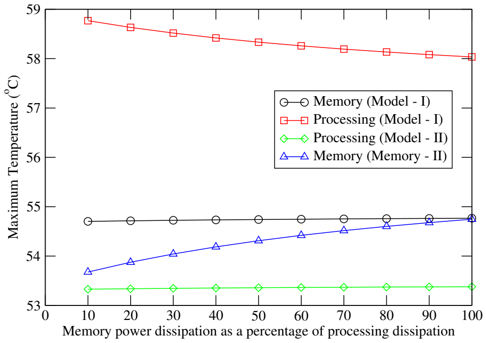

Fig. 5. Maximum temperature on the processing and memory die for both models.

B. *Transient heat transfer analysis*

The dalliance in reaching the steady state is measured in transient analysis, wherein the temperature responses are continually recorded within a short time interval for the given power consumption of the silicon dies. Transient analysis is necessary to observe the steady-state behaviour and also the thermal profile of different configurations that might change over time as the maximum temperature is reached.

Fig. 6 and Fig. 7 shows the maximum temperature and the thermal resistance curves plotted against time for both the models when the memory layer is consuming around 10% of the processing power consumption. It can be seen from those curves that the heat sinks of the two models are efficient enough to take the heat out of the system irrespective of the placement of the processing die. By the time steady-state is reached the processing cores of model-I is 6.5°C hotter than model-II. It can also be noted that the thermal resistance of the memory die (R_{mm}) in Model-I is lower by 0.45°C/W when compared to model-II, whereas the thermal resistance of the processing die (R_{pp}) in both the models is almost the same.

In order to find out the improvement that is required in the heat sink thermal resistance for a 3D system when compared to the single die system, a transient percentage reduction plot of the heat sink thermal resistance (R_{hs}) has been plotted as shown in Fig. 8. The single die package system whose power consumption is 100 W, and has a hotspot of 100 W/cm^2 power density at the center of the silicon die has been used for comparison purposes. The curves have been plotted for both

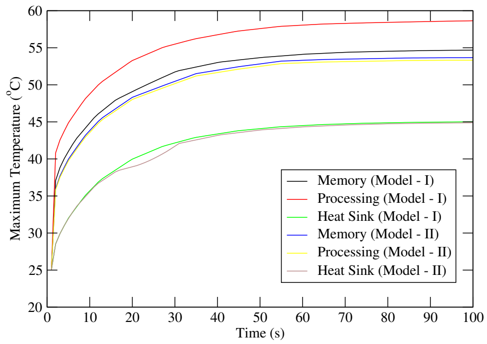

Fig. 6. 10% Maximum temperature on the processing and memory die for both models.

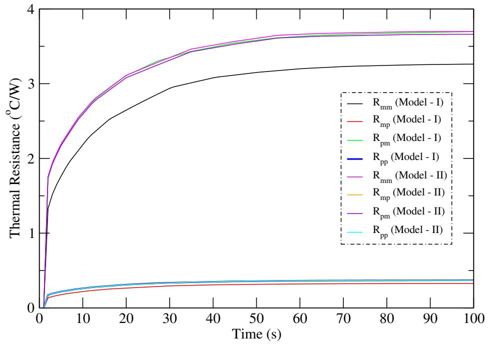

Fig. 7. 10% Thermal Resistance on the processing and memory die for both models.

the models and for different power consumption's of processing and memory layers. All those plots have showed some similarities in nature and hence could be easily segmented into three distinct durations or stages. In this paper we have presented only one plot (Fig. 8) where in the memory die is consuming around 50% of the processing die power.

In the first stage the percentage reductions in R_{hs} is approximately the same for both the models, suggesting that the heat sink behaves identically for both the models for short durations of time.

In the second stage, when the maximum temperature on the heat sink starts to increase before attaining steady-state, model-I demands less reduction in the heat sink thermal resistance. This is due to the fact that the heat could not be transferred from the processing layer below the ILM to the heat sink. If the configuration of model-I tends to work in this stage, then instead of improving the heat sink one should concentrate on improving the effective thermal conductivity of the ILM layer.

In the third stage when both the models are attaining steady-state, they exhibit expected behaviour, as the configuration with the processor layer near the heat sink (model-II) behaves more efficiently. This is because the required reduction in thermal resistance is less. This plot not only shows the dependence on the stacking sequence but also shows that the observations should not be made strictly on the basis of the steady-state [7] analysis, as in some cases the chips might not reach steady state due to various dynamic thermal management techniques that are employed.

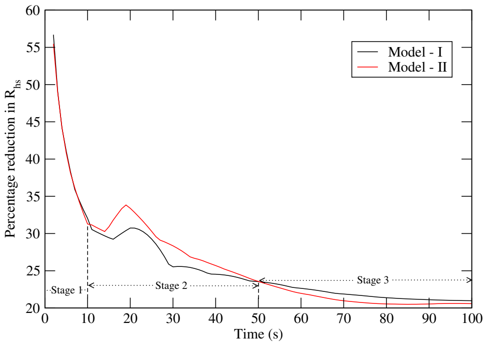

Fig. 8. Improvement required in heat sink thermal resistance for a 3D system (both models) whose memory layer is consuming 50% of the processing die power. It has been compared with a single die package system.

VII. Conclusions

A thermal model of a 3D multicore system in a modern flip-chip package is developed in order to investigate the effects of hotspot, and placement of silicon die layers, on the thermal performance of a multicore system. We have used a finite-element based method to run our simulations. Both the steady-state and transient heat transfer analysis has been performed on the 3D flip-chip package we built. Two different thermal models were evaluated under different operating conditions. We have found that in steady-state for the case where the memory layer dissipates around 10% of the power consumed by the processing core, an overall improvement of 0.6°C/W is obtained in the thermal resistance by placing the silicon layers optimally. For the same case, it has been observed that the difference in the maximum temperature of memory and processing die layers is around 4°C for model-I and 0.3°C for model-II. An improvement that is required in the heat sink thermal resistance for a 3D system when compared to a single-die system has been quantified.

VIII. Acknowledgement

The authors would like to thank the Academy of Finland and Nokia for their financial support.

References

[1] The International Technology Roadmap for Semiconductors, 2007.

[2] Prof. Hannu Tenhunen, Personal communication.

[3] Ayse K. Coskun, José L. Ayala, David Atienza, Tajana Simunic Rosing, and Yusuf Leblebici, "Dynamic Thermal Management in 3D Multicore Architectures," *In Proceedings of Design Automation and Test in Europe (DATE)*, pp. 1410-1415, 2009.

[4] Guoping Xu, Bruce Guenin, Marlin Vogel, "Extension of Air Cooling for High Power Processors," *In Proceedings of 9^{th} Inter Society Thermal Phenomena in Electronics Systems (ITherm) Conference*, 2004.

[5] Guoping Xu, "Thermal Modeling of Multi-core Systems," *In Proceedings of 10^{th} Inter Society Thermal Phenomena in Electronics Systems (ITherm) Conference*, pp. 96-100, 2006.

[6] Ravi Kandasamy, and A.S. Mujumdar, "Interface Thermal Characteristics of flip-chip packages - A numerical study," *Applied Thermal Engineering*, Volume 29, Issues 5-6, pp. 822-829, April 2009.

[7] Ankur Jain, Robert E. Jones, Ritwik Chatterjee, and Scott Posder "Analytical and Numerical Modeling of the Thermal Performance of Three-Dimensional Integrated Circuits," *IEEE Transactions on Components and Packaging Technologies*, Volume 33, No. 1, March 2010.

[8] C. Zhu and et al. "Three-Dimensional Chip-Multiprocessor Run-Time Thermal Management," *IEEE Transactions on CAD*, August 2008.

[9] Texas Instruments, "Flip Chip Ball Grid Array Package Reference Guide," Literature Number: SPRU811A, May 2005.