Multi-Die Packaging and Thermal Superposition Modeling

Mike Kelly[(1)], Phillip Fosnot[(1)], Jonathan Wei[(2)], Max Min[(3)] and Jesse Galloway[(1)]
[(1)]Amkor Technology, 2045 E. Innovation Circle, Tempe AZ 85284
[(2)]eSilicon Corporation, 2130 Gold Street, San Jose, CA 95002
[(3)]Samsung Foundry, 3655 N. First Street, San Jose CA 95134

Abstract

Packaging density, electrical performance and cost are the primary factors driving electronic package architectures for high-performance server markets. Considerations such as thermal performance and mechanical reliability are equally important but tend to be addressed later in the design cycle. Presented in this paper is a historical view of the packaging trends leading to the current multi-die package options. Particular attention must be placed on the thermal limitations and benefits offered by each design. Since multi-die packages have many junctions of interest, a method for characterizing the package with arbitrary power conditions is required. A superposition method, using a matrix approach, is presented that will enable the end-user to investigate the effects of power levels on junction temperatures. Experimental data measured on a 2.5D package were taken to demonstrate the matrix approach for predicting junction temperature based on an independent power map. The agreement between the matrix model and data generated by an independent power map is within 8%.

Keywords

2.5D packaging, influence coefficient, thermal, modeling

Nomenclature

P	Power (W)
T	Temperature (°C)
ψ	Influence Coefficient (°C/W)

1. Introduction

Next generation multi-die integrated circuit (IC) package integration requires a flexible package technology portfolio. A key ingredient in this development path is qualifying module technologies that can meet a growing variety of customer needs. High performance system-level or multi-die packaging have been around for decades. IBM's server class products, for example the ES9000 [1], were among the first products to implement multiple die into a very high-performance module, see the thermal conduction module shown in Figure 1.

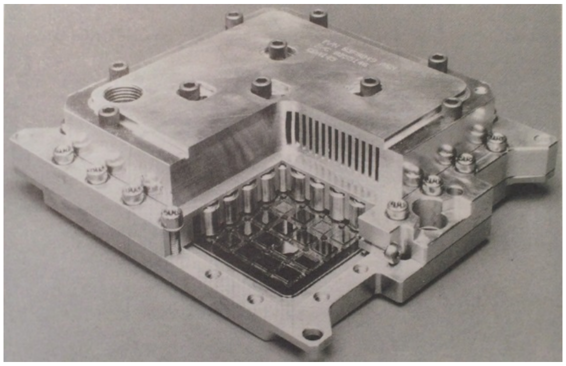


Figure 1: IBM Thermal conduction module.

More recently, the new compute performance levels enabled by cutting-edge FinFet central processing units (CPUs) and graphic processing units (GPUs), and a new class of high-density, high-bandwidth dynamic random-access memory (DRAM), such as the high bandwidth memory (HBM), has not only raised expectations but has also enabled a multitude of product possibilities. Market drivers for high-speed graphics, server-class compute in data centers and for artificial intelligence (AI) keeps raising the bar higher for what must be achieved in IC packages.

Several key technologies have come together to create this bold shift to higher performance. An important innovation has been thru-silicon via (TSV) technology, whether in the interposer, memory or logic devices. Shown in Figure 2 is the integration of the application-specific integrated circuit (ASIC) on the silicon interposer, where CuP refers to copper pillar interconnect and C4 refers to controlled collapse chip connection.

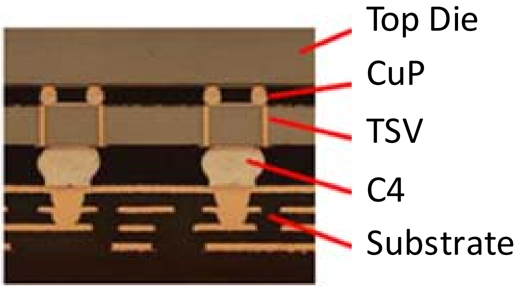


Figure 2: 2.5D package cross-section.

To address these growing differentiations of product needs, design options were developed for several key multi-die module developments. 2.5D TSV package construction remains a high-end bulwark with very predictable package reliability all the way to full reticle sized TSV-bearing interposers. With the next wave of product differentiation based on product test strategies, known good die (KGD) specifications and expected functional die size, TSV, the chip on substrate (COS) solution has also evolved to include both molded chip-on-wafer (COW) constructions as well. 2.5D packages have been qualified and are in production for sizes ranging up to 55mm x 55mm. To assist product development using these advanced features, a reliable means for predicting junction temperatures is required.

To date, 2.5D has been used nearly exclusively for combinations of the latest IC logic ASIC and HBM. The combination of ASIC and very high-bandwidth, high capacity DRAM, has been a winning combination for several product classes, including network switches and GPUs for gaming as well as high-end GPUs used in deep learning or AI algorithm optimizations in the data center.

The DRAM memory bus is a special, very wide parallel interface with 1,024 data bits on 1,024 physical signal lines. This parallel bus requires much lower power than a serial interface, but does requires very fine line signal routing, with line and space of 1 or 2 um and 3 or 4 layers being common. This type of interface is currently possible using only copper damascene back-end technology from the IC fabrication industry. This is why the silicon-based interposer using a 65nm or 90 nm copper back-end fabrication process is now common.

Even cutting-edge fine-line organic package substrates cannot provide the required level of signal routing density. Figure 3 shows a typical configuration, in this case one ASIC and 4 HBM stacked DRAM devices. These IC packages, which utilize HBM or HBM version 2, are possible today using 2.5D silicon interposer technology.

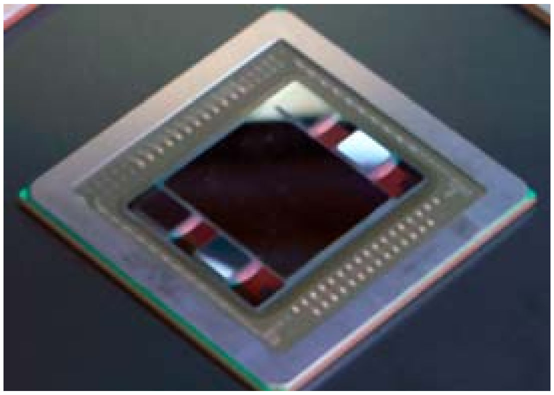


Figure 3. 2.5D package with ASIC and HBM.

The 2.5D TSV package construction is the first practical approach to integrating different ICs that require this extremely fine-signal routing. The reliability of this package type has been demonstrated in JEDEC level accelerated testing. One exceptional benefit for use of a silicon-based interposer is that the CTE mismatch to the ultra-low dielectric constant (ULK) layers in the functional die is very low. This, in turn lowers the stress in the functional die ULK layers as well as making the 2.5D construction a popular choice for upcoming 7nm silicon node products. In 2017, the most prominent product introductions using 2.5D packaging technology have come from the GPU arena, for both high end gaming and deep learning applications from leading suppliers. GPU architecture use hundreds of small parallel processing cores optimized for matrix math, an essential constituent in 3D graphics rendering. As it turns out, this very same processor architecture is also very well suited for AI database optimization now being used in large data centers for facial and voice recognition as well as other applications. 2.5D packaging is expected to remain a mainstay of the high-performance computing sector, mainly in high end gaming and AI and networking data center applications for years to come.

Integrating multiple die on the same silicon interposer creates thermal challenges because each device has its own temperature limitation and power requirements. Moreover, a specific multi-die package can be configured to operate at many different power levels depending on the end-use application and the nature of the programs being supported. For these reasons, it becomes a significant challenge to characterize the thermal performance of the package for many different applications anticipated during field operation. System architects need a method for predicting junction temperatures as a function of power map to support their given application.

Linear superposition is a common technique used for predicting the temperature field as a function of the power map for the various devices operating on the multi-die package. Lall et.al. [2] provided an experimental method for predicting the junction temperature for a multi-die package mounted in a quad flat pack (QFP) package with a drop-in heat spreader. The overall solution was based on the average package temperature. Individual chip temperatures were predicted as rise above the package average temperature determined by power factors for each chip. Following this approach individual die temperatures for many different power configurations could be predicted using data from a limited number of tests. Emigh [3], Fan [4] and Zhang [5] presented an application of the classical superposition approach using a matrix of influence coefficients. The temperature for a given device was calculated by adding the contribution made by each device powered. The contribution was controlled by influence coefficients. The value of the influence coefficient was larger if the contributing chip was in close proximity to the device in question or if there was a connection with a low thermal resistance path.

The application of linear superposition requires that the heat transfer process become independent of power applied to the package. For limited sets of conditions, this assumption is valid. For example, if the temperature rise is not too large, then the thermal conductivity will not change significantly and the nonlinear effects will not be present. In forced convection regimes, the heat transfer coefficient does not change significantly when the power levels are increased. Superposition methods may work well for high power applications that require an external heat sink with a cooling fan such as a CPU cooling fan.

However, for natural convection, power changes may have a significant influence on air circulation around the package and mother board thus changing the heat transfer coefficient. Superposition methods will not work well under natural or still air conditions due to the nonlinearity response of the heat transfer coefficient with the applied power.

A systematic calculation method must be developed for predicting die temperature as a function power maps applied to multiple power blocks in a 2.5D application. Experimental data and a superposition calculation method is provided in this study to demonstrate the limitations and applications for predicting temperatures in a multi-zone heated package.

2. Experimental Methods

A 2.5D thermal test vehicle (TTV) was constructed to determine the thermal response of the ASIC and HBM to different power maps. Both the ASIC and HBM share a common silicon interposer. The silicon interposer is connected to the substrate and the substrate is connected to the mother board. Power connections, sense lines and temperature sensor connections are made through pin headers on the mother board. A standard CPU cooler is used as a heat sink to dissipate the rejected heat. An overview of the 2.5D package and mother board is shown in Figure 4.

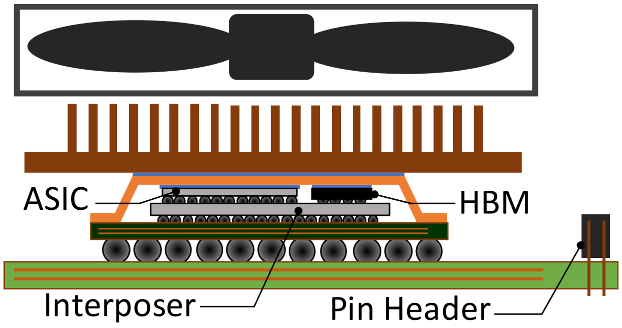


Figure 4. TTV system.

The power zones for the TTV are shown in Figure 5 with the corresponding heat fluxes (relative values shown) applied to each block. Each zone on the ASIC included independent heaters and temperature sensors. The HBM also included heaters and thermal sensors.

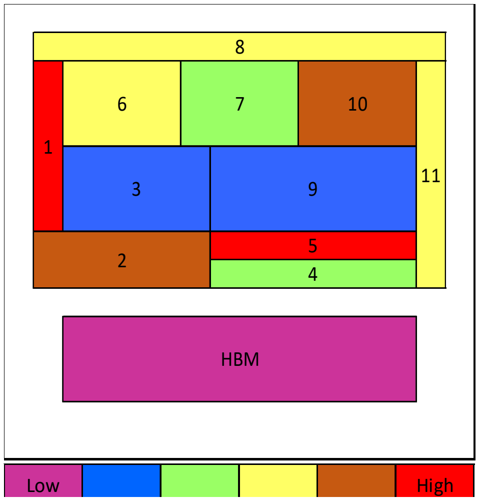


Figure 5. Power block relative heat flux.

The CPU fan was run at the same speed for all tests. The modeling results presented here are for one fan speed. Additional testing would be required to model the effect of fan speed on die temperatures. One set of power conditions is shown in Figure 6. The data were normalized by dividing each individual power block by the total power supplied to the TTV.

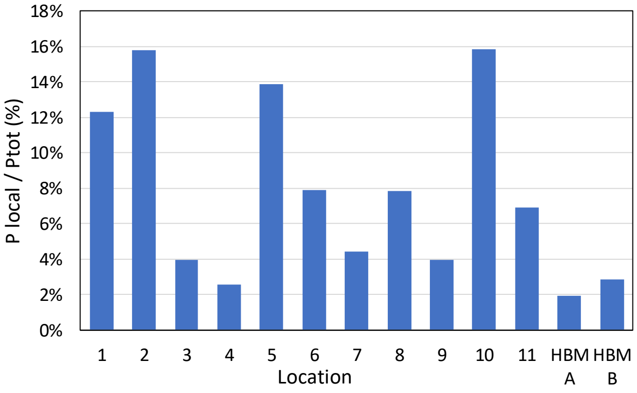


Figure 6. Nomalized power distribution.

The corresponding temperature rise above ambient is shown in Figure 7. The temperature response is not only a function of the power applied to the power block but also depends on its location and its neighboring power blocks. Even though power at block 1 was similar in magnitude and size compared to block 5, the corresponding temperature at location 1 is significantly higher than at location 5.

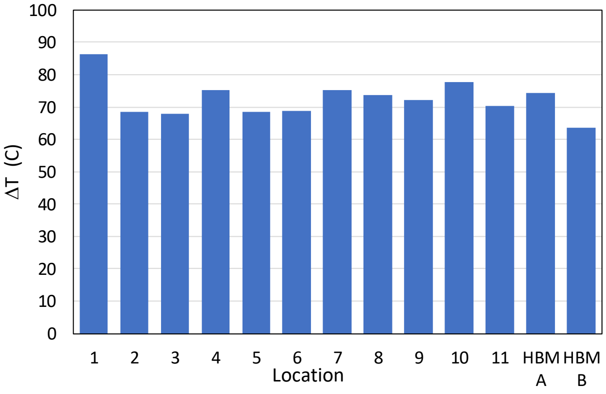


Figure 7. Temperature rise above ambient for power blocks shown in Figure 6.

3. Linear Superposition Model

A training set of power conditions was used to power individual heaters one at a time while reading all temperature sensors. To predict the temperature-rise above ambient for a new power set, influence coefficients, ψ_{ij}, were extracted from a linear model as represented in equation (1).

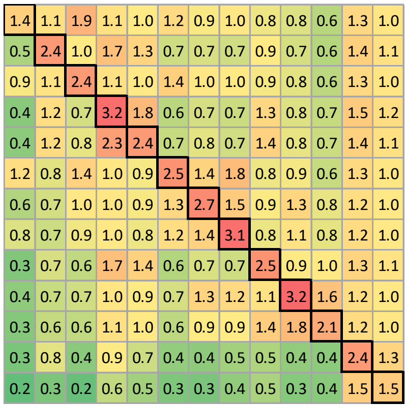
$$\begin{bmatrix} \psi_{1,1} & \cdots & \psi_{1,N} \\ \vdots & \ddots & \vdots \\ \psi_{N,1} & \cdots & \psi_{N,N} \end{bmatrix} \begin{bmatrix} P_1 \\ \vdots \\ P_N \end{bmatrix} = \begin{bmatrix} \Delta T_1 \\ \vdots \\ \Delta T_N \end{bmatrix} \quad (1)$$

Power devices were turned on for all 13 heaters while recording temperatures 1 – 11 and temperatures for HBM-A and HBM-B. A total of thirteen independent power combinations were tested experimentally to generate a training set necessary for the evaluation of the ψ matrix. A simplification can be made to solve the ψ matrix by powering the heaters one at a time. For this limiting case, it follows that the influence coefficients can be calculated as

$$\boldsymbol{\psi_{i,j}} = \frac{\boldsymbol{\Delta T_i}}{\boldsymbol{P_j}} \tag{2}$$

where “j” refers to the power block location and “i” refers to the location of the temperature sensor.

The overall package temperature rise is expected to be smaller when heaters are powered individually due to the total power being smaller compared to the power map where all heaters are activated at the same time. For individually powered heaters, the temperatures ranged between 4 to 30°C. This temperature range is lower than the condition when all power zones were run at the same time, see Figure 7. After running 13 power maps corresponding to heaters 1-13 powered one at a time, numerical values for ψ_{ij} were determined using equation (2). Shown below in Figure 8 is a summary of the influence coefficient matrix, ψ. Notice that the main-diagonal elements (shown in heavy outlined boxes in Figure 8) have higher values compared to off-diagonal elements. The reason for this observation is that the influence of a power block on the temperature sensor at the same location is expected to be greater than at neighboring temperature sensors.



1.4	1.1	1.9	1.1	1.0	1.2	0.9	1.0	0.8	0.8	0.6	1.3	1.0
0.5	2.4	1.0	1.7	1.3	0.7	0.7	0.7	0.9	0.7	0.6	1.4	1.1
0.9	1.1	2.4	1.1	1.0	1.4	1.0	1.0	0.9	0.8	0.6	1.3	1.0
0.4	1.2	0.7	3.2	1.8	0.6	0.7	0.7	1.3	0.8	0.7	1.5	1.2
0.4	1.2	0.8	2.3	2.4	0.7	0.8	0.7	1.4	0.8	0.7	1.4	1.1
1.2	0.8	1.4	1.0	0.9	2.5	1.4	1.8	0.8	0.9	0.6	1.3	1.0
0.6	0.7	1.0	1.0	0.9	1.3	2.7	1.5	0.9	1.3	0.8	1.2	1.0
0.8	0.7	0.9	1.0	0.8	1.2	1.4	3.1	0.8	1.1	0.8	1.2	1.0
0.3	0.7	0.6	1.7	1.4	0.6	0.7	0.7	2.5	0.9	1.0	1.3	1.1
0.4	0.7	0.7	1.0	0.9	0.7	1.3	1.2	1.1	3.2	1.6	1.2	1.0
0.3	0.6	0.6	1.1	1.0	0.6	0.9	0.9	1.4	1.8	2.1	1.2	1.0
0.3	0.8	0.4	0.9	0.7	0.4	0.4	0.5	0.5	0.4	0.4	2.4	1.3
0.2	0.3	0.2	0.6	0.5	0.3	0.3	0.4	0.5	0.3	0.4	1.5	1.5

Figure 8. Influence coefficients generated from the first training set, heaters powered one at a time.

One should not expect a symmetric matrix. For example, heater 1 may have a different effect on temperature 2 compared to heater 2 on temperature 1. To test the accuracy of the super-position model, an independent power set, [P_i], was experimentally tested and modeled using equation (1) and the influence coefficients reported in Figure 8. A comparison of the experimental data and model is shown in Figure 9.

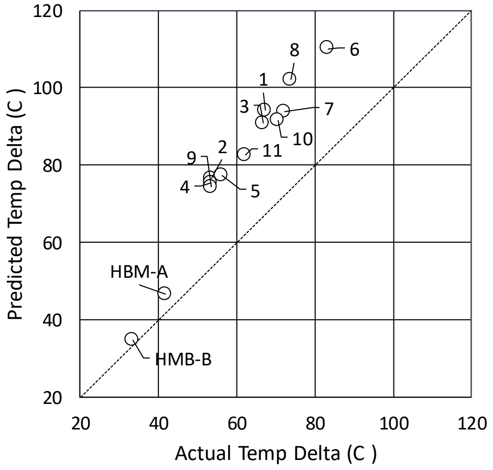


Figure 9. Comparison of super-position model to experimental data.

The model overpredicted the actual temperatures by approximately 26%. It is well known that material properties used to construct the package (i.e. thermal conductivity) are temperature dependent. The training power map set was taken at a condition that produced lower die temperatures compared to the test set when all heaters were activated.

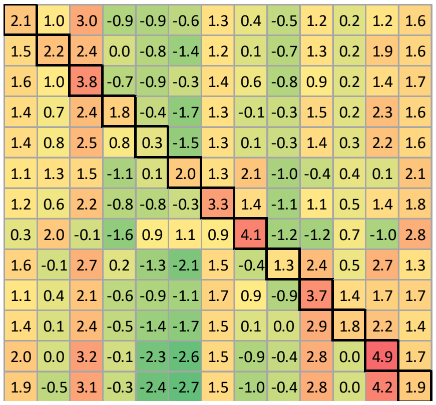
A different approach was needed to develop a training set that keeps the average temperature of the package similar to the temperatures produced by a desired power map. Thirteen power training sets were developed by randomly varying the power for each block about a nominal value with the constraint that these values did not produce a condition that led to an over-heat condition (i.e. exceed safe material temperature limits). Data taken on 13 different training sets were loaded into the matrix listed in equation (3). The superscript for the power and temperature vectors indicates a specific training set, 1-13. The subscript indicates the location of the power blocks and temperature sensors, see Figure 5.

$$\begin{bmatrix} \boldsymbol{\psi_{1,1}} & \cdots & \boldsymbol{\psi_{1,13}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\psi_{13,1}} & \cdots & \boldsymbol{\psi_{13,13}} \end{bmatrix} \begin{bmatrix} \boldsymbol{P_1^1} & \cdots & \boldsymbol{P_1^{13}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{P_{13}^1} & \cdots & \boldsymbol{P_{13}^{13}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Delta T_1^1} & \cdots & \boldsymbol{\Delta T_1^{13}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Delta T_{13}^1} & \cdots & \boldsymbol{\Delta T_{13}^{13}} \end{bmatrix} \tag{3}$$

Numerical values for the ψ_{ij} matrix elements were calculated by taking the inverse of the power matrix as show in equation (4).

$$\begin{bmatrix} \boldsymbol{\psi_{1,1}} & \cdots & \boldsymbol{\psi_{1,13}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\psi_{13,1}} & \cdots & \boldsymbol{\psi_{13,13}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Delta T_1^1} & \cdots & \boldsymbol{\Delta T_1^{13}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Delta T_{13}^1} & \cdots & \boldsymbol{\Delta T_{13}^{13}} \end{bmatrix} \begin{bmatrix} \boldsymbol{P_1^1} & \cdots & \boldsymbol{P_1^{13}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{P_{13}^1} & \cdots & \boldsymbol{P_{13}^{13}} \end{bmatrix}^{-1} \tag{4}$$

A new set of influence coefficients were calculated and reported in Figure 10. Notice that the off-diagonal elements in some cases are negative whereas the off-diagonal elements in the first training set, Figure 8, were all observed to be positive.



2.1	1.0	3.0	-0.9	-0.9	-0.6	1.3	0.4	-0.5	1.2	0.2	1.2	1.6
1.5	2.2	2.4	0.0	-0.8	-1.4	1.2	0.1	-0.7	1.3	0.2	1.9	1.6
1.6	1.0	3.8	-0.7	-0.9	-0.3	1.4	0.6	-0.8	0.9	0.2	1.4	1.7
1.4	0.7	2.4	1.8	-0.4	-1.7	1.3	-0.1	-0.3	1.5	0.2	2.3	1.6
1.4	0.8	2.5	0.8	0.3	-1.5	1.3	0.1	-0.3	1.4	0.3	2.2	1.6
1.1	1.3	1.5	-1.1	0.1	2.0	1.3	2.1	-1.0	-0.4	0.4	0.1	2.1
1.2	0.6	2.2	-0.8	-0.8	-0.3	3.3	1.4	-1.1	1.1	0.5	1.4	1.8
0.3	2.0	-0.1	-1.6	0.9	1.1	0.9	4.1	-1.2	-1.2	0.7	-1.0	2.8
1.6	-0.1	2.7	0.2	-1.3	-2.1	1.5	-0.4	1.3	2.4	0.5	2.7	1.3
1.1	0.4	2.1	-0.6	-0.9	-1.1	1.7	0.9	-0.9	3.7	1.4	1.7	1.7
1.4	0.1	2.4	-0.5	-1.4	-1.7	1.5	0.1	0.0	2.9	1.8	2.2	1.4
2.0	0.0	3.2	-0.1	-2.3	-2.6	1.5	-0.9	-0.4	2.8	0.0	4.9	1.7
1.9	-0.5	3.1	-0.3	-2.4	-2.7	1.5	-1.0	-0.4	2.8	0.0	4.2	1.9

Figure 10. Influence coefficient matrix generated from the second training set using random power settings.

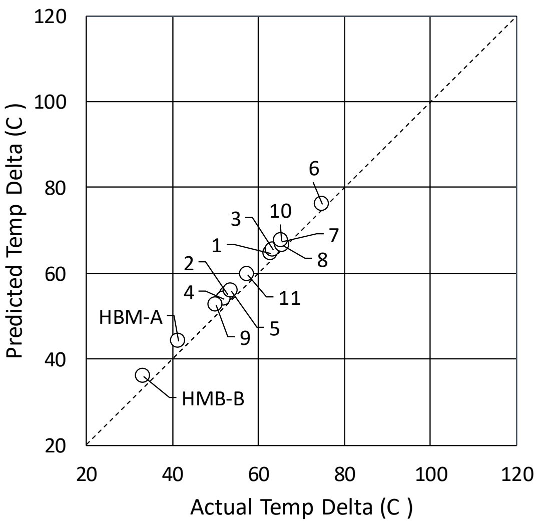


Figure 11. Equal power training set for super-position model.

The positive influence coefficients in the first training set resulted because the power was applied to one heater at a time and the ψ values were calculated independently from other power tests. A positive value is expected since the temperature should rise at all locations when power is applied. When all boundary conditions are implemented at the same time using the inverse matrix approach for the second training set written in equation (4), the effect of experimental variations cause some coefficients to be positive while others are negative. It is the sum of all coefficients multiplied by power that is used to calculate temperature. The off-diagonal elements balance in the sum to calculate a positive temperature. The main diagonal elements should be positive. This is a requirement; for example, since adding power to block 5 should cause an increase in temperature at location 5. The improved accuracy of the model is noted by the predicted results falling closer to the diagonal line as show in Figure 11. The average difference between the superposition model and the actual temperature difference are 3% with a maximum difference of 8%.

A method for representing the thermal performance of the 2.5D package subject to an arbitrary power map that is within the temperature range of the training set, produces accurate temperature predictions. This approach may be useful in helping package designers predict junction temperatures as a function of different power maps.

Conclusions

Linear superposition models can accurately predict junction temperatures for arbitrary power maps provided the package average temperature is similar. A matrix inversion method was introduced using influence coefficients to predict die temperatures. [N] number of training sets are required for [N] number of heat sources with a corresponding [N] number of temperature sensors. When training power maps produce temperatures that vary significantly greater than the case of interest, material property non-linearities may make it difficult to recover a linear model that will accurately predict the temperature response for an applied power map.

Acknowledgments

The thermal test vehicle was designed by eSilicon corporation and Samsung. Assembly of the test vehicle was supported by Amkor's K4 package assembly facility. The authors greatly appreciate their assistance in making the thermal data collection possible.

References

1. IBM Archives, https://www-03.ibm.com/ibm/history/exhibits/mainframe/mainframe_FS9000.html
2. Lall B., Guenin B., and Molnar R, "Methodology for thermal evaluation of multichip modules", IEEE Transactions on Components, Packaging and Manufacturing Technology – Part A, Vol. 18, No. 4, December 1995, pp. 758-764.
3. Emigh, R., "Thermal behavior and data processing for multi-chip packages: lateral, stacked, PoP and PiP, 2007, www.meptec.org/Resources/07%20STATSChipPAC%20pres.pdf
4. Fan X, "Development, validation and application of thermal modeling for a MCM power package, 11-13 March 2003, pp. 144-150.
5. Zhang H., Zhang X., Lau B., Lim S., Ding L. and Yu M., "Thermal characterization of both bare die and overmolded 2.5-D packages on through silicon interposers"; IEEE Transactions on Components, Packaging and Manufacturing Technology, Vol. 4, No. 5, May 2014, pp. 807-816.