Multi-Package Co-Design for Chiplet Integration

Zhen Zhuang[1], Bei Yu[1], Kai-Yuan Chao[2], and Tsung-Yi Ho[1]
[1]The Chinese University of Hong Kong
[2]Hong Kong Research Center, Huawei Technology Investment Co. Ltd.

ABSTRACT

Due to the cost and design complexity associated with advanced technology nodes, it is difficult for traditional monolithic System-on-Chip to follow the Moore's Law, which means the economic benefits have been weakened. Semiconductor industries are looking for advanced packages to improve the economic advantages. Since the multi-chiplet architecture supporting heterogeneous integration has the robust re-usability and effective cost reduction, chiplet integration has become the mainstream of advanced packages. Nowadays, the number of mounted chiplets in a package is continuously increasing with the requirement of high system performance. However, the large area caused by the increasing of chiplets leads to the serious reliability issues, including warpage and bump stress, which worsens the yield and cost. The multi-package architecture, which can distribute chiplets to multiple packages and use less area of each package, is a popular alternative to enhance the reliability and reduce the cost in advanced packages. However, the primary challenge of the multi-package architecture lies in the tradeoff between the inter-package costs, i.e., the interconnection among packages, and the intra-package costs, i.e., the reliability caused by warpage and bump stress. Therefore, a co-design methodology is indispensable to optimize multiple packages simultaneously to improve the quality of the whole system. To tackle this challenge, we adopt mathematical programming methods in the multi-package co-design problem regarding the nature of the synergistic optimization of multiple packages. To the best of our knowledge, this is the first work to solve the multi-package co-design problem.

1 INTRODUCTION

Semiconductor industries have driven the development of chips to follow the Moore's Law for over half a century. The economic advantages of integrating more and more transistors on a single die are decreasing for recent decades. Nowadays, limited foundries can afford the high costs. Therefore, many foundries, such as TSMC, Samsung, and Intel, are developing advanced package techniques, including heterogeneous integration with 2.5D packaging, to provide promising solutions to respond the increasing cost of more-Moore scaling. The heterogeneous integration roadmap, led by industries and academic societies, identifies the challenges and necessity to develop advanced package techniques [1].

ICCAD '22, October 30-November 3, 2022, San Diego, CA, USA

ACM ISBN 978-1-4503-9217-4/22/10…$15.00
https://doi.org/10.1145/3508352.3549404

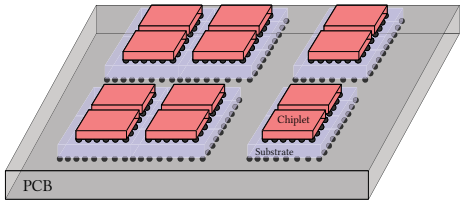


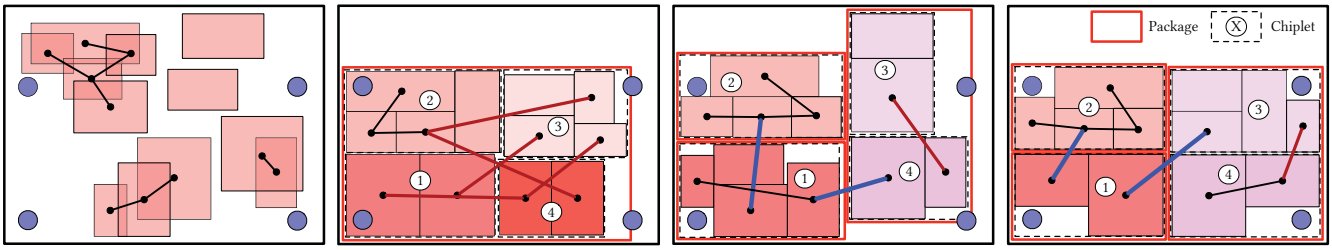
Figure 1: The illustration of the multi-package architecture.

Since chiplet enables several features, including low cost, heterogeneous integration, and drop-in design method, it is widely used for advanced package designs nowadays [2]. Feng *et al.* [3, 4] build effective cost models to explain the benefits of chiplet integration from the perspective of cost and yield. However, the large sizes of advanced packages become the bottleneck of the further saving of cost for chiplet integration [3]. The yield reduction caused by the reliability issues induced by the large size dramatically increases the cost of advanced packages. The higher risks of reliability issues are induced by the higher degree of coefficient thermal expansion (CTE) mismatch, such as bump joint reliability affected by stress, chiplet cracking, and substrate cracking affected by warpage [5–8]. Warpage is an unconventional bending out of the shape.

Maintaining reliability of packages is becoming more challenging as the sizes of the substrates continue to increase and greater stress is exerted on the substrates. To guarantee the ever-increasing computation, heterogeneous integration techniques have evolved with larger logic chiplets surrounded by different node chiplets mounted, or greater numbers of memories mounted around logic chiplets [5]. A system, which is constructed by multiple packages based on the chiplet integration, can efficiently improve the benefits of heterogeneous integration. The illustration of the multi-package architecture is shown in Fig. 1. The multi-package architecture includes four packages. Twelve chiplets are assigned to the four packages. To reduce the bump stress, which may cause the deformation of bumps, the macros should avoid be placed on the bumps. Due to the reduction of the package size by partitioning the whole system into several packages, the reliability and cost are effectively improved. Furthermore, the reuse of packages further reduces the cost. Although the interconnection cost between different packages is introduced in the multi-package systems, it will be effectively reduced in this work. The products of Siemens and Cadence, such as Xpedition and Cadence Allegro, have been designed for multi-package systems. The multi-package systems have been used for different applications [9].

1.1 Previous Work

The typical papers for different types of related work are selected and compared in this section. The work proposed for the physical design of on-chip problems, such as [10–14], cannot be directly scaled to solve package problems, and does not consider the reliability issues of



(a) The macros of input data. (b) One-package design. (c) Multi-package design considering bump stress. (d) Multi-package co-design.

Figure 2: The different designs of chiplet integration. (a) The macros, bumps, nets, and fixed-outline board generated from input data. (b) The one-package design with four chiplets. (c) The multi-package design including three packages and four chiplets. (d) The multi-package co-design, including three packages and four chiplets, with the optimization of interconnection cost, bump stress, and warpage.

advanced packages. The details of the differences between this work and the previous work are summarized in Section 3. The analysis shows the necessity to design new methods for the multi-package co-design problem. Therefore, the new mathematical programming methods are adopted regarding the nature of the problem.

The work proposed for the physical design of package problems, such as [15, 16], is not suitable for the multi-package co-design problem. They are designed for the chiplet-level problems, which makes the assignment of macros impossible. Furthermore, the reliability issues are not considered. Fig. 2 shows the different designs of chiplet integration. The macros, which should be processed, are illustrated in Fig. 2a. The circles around the macros are the bumps, which should have no overlap with macros to avoid the deformation. Based on the conventional work [15, 16], the macros have been assigned to four chiplets with similar area, and the chiplets should be placed in one package as shown in Fig. 2b. Since all the chiplets near each other, the interconnection of one-package design is better. To optimize the area of package as much as possible, the area of the whole design is reduced to 1222 mm^2. However, the macros have overlaps with the bumps. The overlaps between bumps and macros increase the stress on the bumps, which may lead to the cracking of bumps. Furthermore, the warpage is not reduced to optimize the reliability. For the one-package design, the warpage is 696 μm.

Jung *et al.* [17] analyze the bump stress issue. Fig. 2c shows the multi-package architecture designed to optimize the bump stress issue. The four chiplets are distributed to three packages. Compared with the one-package design, the total area increases and the inter-package connections are induced. However, the overlaps between macros and bumps are avoided. Therefore, the bump stress can be effectively reduced. Furthermore, since the area of each package is reduced, the maximum area of packages and the warpage are reduced to 646 mm^2 and 379 μm, respectively. Compared with the one-package design, the area is reduced by 47%, and the warpage is reduced by 46%. Since the warpage is not the objective of this design, the warpage can be further reduced.

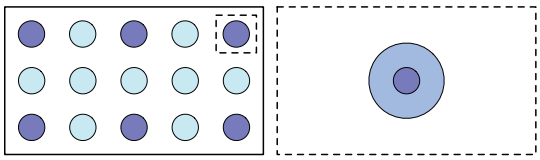
Fig. 2d shows the solution of multi-package co-design, including three packages. The four chiplets are distributed to the three packages. Compared with the design of Fig. 2c, it has the same interconnection cost and the similar maximum area of packages. However, the total area is reduced. The warpage is effectively reduced to 251 μm. Compared with the design only considering bump stress, the warpage is reduced by 34%. Therefore, the multi-package co-design can achieve the best quality for the optimization of both intra-package costs and inter-package costs.

1.2 Our Contributions

In this paper, the multi-package co-design problem is defined, and the methods for solving the problem are explored. For the solutions of the defined problem, both the intra-package costs and inter-package costs should be optimized. The major contributions are shown below:

- With the increasing of the size of advanced packages, it is difficult to maintain reliability and optimize cost. Therefore, the multi-package architecture is a popular alternative. To the best of our knowledge, this is the first work proposed to overcome the primary challenge of the multi-package co-design problem, which is the tradeoff between the inter-package cost and reliability.
- The critical reliability issues, including warpage and bump stress, are effectively optimized. An effective warpage model is formulated for the effective calculation. Then, the warpage can be effectively reduced based on the threshold specified by users. Furthermore, the design region is constrained to avoid the deformation of bumps.
- The mathematical programming (MP) method is adopted regarding the nature of the multi-package co-design problem. Firstly, a one-pass MP (**OPMP**) method and a partition-based MP (**parMP**) method are proposed as the baselines. Then, a hierarchical MP (**hieMP**) method is proposed to improve the efficiency compared with the two baselines.
- Experimental results show that the performance and reliability issues of large benchmarks can be efficiently optimized by hieMP. Furthermore, hieMP can also help for users to choose proper package architectures considering the bottleneck of systems, interconnection cost or reliability.

This paper is organized as follows. Section 2 presents the reliability issues and the problem formulation. Section 3 introduces the technical details of the proposed methods. Section 4 and Section 5 present the experimental results and conclusion, respectively.



(a) Bump array. (b) Bump margin constraints.

Figure 3: The illustration for the bump margin constraints.

2 PRELIMINARIES

2.1 Reliability Issues

2.1.1 Warpage. Warpage is an unconventional bending out of the shape. It is one of the most serious reliability issues for advanced packaging. Due to the CTE mismatch of different materials, the package will bend. The serious shape change will cause the cracking of chiplets and substrates. The warpage is measured by the vertical height variation of the positions of a package. In this work, we control the warpage of packages from the centers to all the edges. An effective warpage computing model introduced in [18, 19] is used for the multi-package designs. The calculation of warpage is shown below:

$$w(x) = \frac{t \cdot \Delta\alpha \cdot \Delta T}{2 \cdot \lambda \cdot D}\left[\frac{1}{2}x^2 - \frac{\cosh(kx) - 1}{k^2 \cosh(kl)}\right], \tag{1}$$

where $\Delta\alpha$ and ΔT are the difference of CTE between the chiplets and the substrate and the thermal load, respectively. t, λ, D, and k are the material-related coefficients. This model is built by taking the center of a package as the origin. x and l are half the length of a chiplet and half the length of a substrate, respectively.

The critical parameters affecting warpage include mold thickness, molding materials, and the ratio of chiplet to package area [7, 8]. Since the ratio of chiplet to package area, ar, is positive correlation with warpage, it is modeled as below:

$$ar = \frac{\sum_{c_i \in C} a_{c_i}}{\sum_{p_l \in P} a_{p_l}}, \tag{2}$$

where c_i and p_l represent the i-th chiplet in the chiplet set C and the l-th package in the package set P, respectively. a_{c_i} and a_{p_l} represent the area of c_i and p_l, respectively. The warpage can be reduced by decreasing ar.

2.1.2 Bump Stress. The stress induced by bumps is also one of the most serious reliability issues for advanced packaging. With the development of advanced package technology, the stress levels induced by bumps become higher, which should be reduced to avoid the deformation of components [20, 21]. According to the analysis of [17], the edges of bumps have the high stress caused by the placement of components near the bumps, which increases the risks of bump crack. In this work, the stress around the bump edges is controlled to reduce the risks of bump crack.

2.2 Problem Formulation

The inputs of this work include a netlist N, a set of macros M, a set of chiplets C, and a set of packages P. The output is the floorplanning solution which should assign macros to chiplets, assign chiplets to packages, place macros within chiplets, and place chiplets within packages.

Since multiple design factors are optimized in this work, and they have different levels of priority, the objective is set to a two-level formulation. The first-level objective is formulated as below:

$$\begin{aligned} min \quad & \beta_1 \times wl + \beta_2 \times oh_C + \beta_3 \times oh_P + \beta_4 \times \sum_{p_l \in P} a_{p_l} \\ \text{s.t.} \quad & wpg_{p_l} \leqslant wpgt_{p_l}, \ \forall p_l \in P, \end{aligned} \tag{3}$$

where wl, oh_C, oh_P, wpg_{p_l}, and $wpgt_{p_l}$ represent HPWL, inter-chiplet cost, inter-package cost, the warpage of p_l, and the warpage threshold of p_l, respectively. $wpgt_{p_l}$ is a user-specified threshold, which can be set based on the total area of macros, the estimated width of each package, and the estimated height of each package. β_i represents the user-defined coefficient which can be modified for the tradeoff of multiple objectives. oh_C is the number of inter-chiplet connections between each source-sink pair. oh_P is the number of inter-package connections between each source-sink pair. Since this work is dedicated to solving the problem in the early design stage, it is difficult and time-consuming for overhead analysis to use more practical interconnection. Therefore, oh_C and oh_P are used to calculate overhead cost.

The second-level objective is to minimize the ratio of chiplet to package area, ar. Since the warpage of each package is reduced in Equation (3), and the priority of ar is not higher than that of Equation (3), ar is set to the second-level objective.

In this work, there are three types of constraints. The first type is overlap constraints. Each macro cannot overlap with other macros. Each chiplet cannot overlap with other chiplets. Each package cannot overlap with other packages. The second type is warpage constraints. As shown in Equation (3), the warpage of each package p_l cannot be larger than the threshold. The third type is bump margin constraints. Margin spacing is defined to avoid the stress around bumps which may cause the deformation of bumps.

The illustration for the bump margin constraints is shown in Fig. 3. The circles represent bumps. Since it is difficult to completely avoid the location overlap between macros and bumps, bumps are classified to two types. The first type is the hotspot bumps represented by the purple circles. The second type is the regular bumps represented by the light blue circles. The bumps, which have higher risks of the deformation, are regarded as the hotspot bumps. The bump margin constraints are only suitable for the hotspot bumps. The dotted boxes shows the zoom-in region of a hotspot bump. For the bump margin constraints, the macros cannot overlap with the hotspot bumps and the spacing regions. The spacing region of a hotspot bump is represented by the blue circle in Fig. 3b.

3 TECHNICAL DETAILS

In this section, the proposed methods will be introduced in detail. The differences between this work and the previous partitioning or floorplanning work and the challenges are shown below:

- For the floorplanning solutions generated by the methods of this work, assigning macros to chiplets and assigning chiplets to packages should be optimized.
- For each chiplet or package, the outline is not fixed. The fixed outline will limit the optimization ability of reliability for the multi-package architecture.

- The locations of bumps are fixed for the board with fixed-outline. However, the bumps are non-fixed within the soft-outlines of chiplets and packages since the locations, widths, and heights of chiplets and packages are variables.
- Due to the existing of bumps, the bounding box of a board should have enough dead space, which makes the convergence of solutions difficult.

Considering the above analysis, the programming-based methods are proposed to solve the multi-package co-design problem. For each package $p_l \in P$, the soft-outline and location are controlled by the variables w_{p_l}, h_{p_l}, x_{p_l}, and y_{p_l}, which represent the width, height, x-coordinate of lower-left corner, and y-coordinate of lower-left corner of p_l, respectively. For each chiplet $c_k \in C$, the soft-outline and location are controlled by the variables w_{c_k}, h_{c_k}, x_{c_k}, and y_{c_k}, which represent the width, height, x-coordinate of lower-left corner, and y-coordinate of lower-left corner of c_k, respectively. The whole multi-package design is implemented on the board with bounding box. The size of the bounding box is $W \times H$. Since the problem defined in this work is complicated, it is difficult to formulate a linear, convex, or differentiable model. Therefore, the general MP models, including linear formulas, quadratic formulas, and non-differentiable formulas, are designed to solve the problem.

3.1 One-Pass MP Method

Firstly, the multi-package co-design problem is formulated as an MP problem, and solved by MP solver. To realize the desired design purposes, the following constraints should be satisfied.

For each macro $m_i \in M$, it cannot overlap with other macros and thus can be constrained as:

$$x_{m_i} + w_{m_i} \leqslant x_{m_j} + W \cdot (p_{i,j} + q_{i,j}), \tag{4}$$
$$y_{m_i} + h_{m_i} \leqslant y_{m_j} + H \cdot (1 + p_{i,j} - q_{i,j}), \tag{5}$$
$$x_{m_i} - w_{m_j} \geqslant x_{m_j} - W \cdot (1 - p_{i,j} + q_{i,j}), \tag{6}$$
$$y_{m_i} - h_{m_j} \geqslant y_{m_j} - H \cdot (2 - p_{i,j} - q_{i,j}), \tag{7}$$
$$p_{i,j},\ q_{i,j} \in \{0, 1\}, \tag{8}$$
$$1 \leqslant i < j \leqslant |M|, \tag{9}$$

where x_{m_i}, y_{m_i}, w_{m_i}, and h_{m_i} represent the x-coordinate of lower-left corner, y-coordinate of lower-left corner, width, and height of m_i, respectively. $p_{i,j}$ and $q_{i,j}$ are used to identify the relative locations of m_i and m_j. For example, the value $(0,\ 0)$ of $(p_{i,j},\ q_{i,j})$ is used to identify the situation, where m_i is on the left of m_j. The overlap of this situation is constrained as Equation (4). The overlap between chiplets or packages is also constrained by the formulations similar to Equations (4)–(9). The chiplets and packages are not necessary to be rotated since they are controlled by variables w_{c_k}, h_{c_k}, w_{p_l}, and h_{p_l}, however, all macros in this work can be rotated. Therefore, the w_{m_i} and h_{m_i} of each macro m_i can be formulated as:

$$w_{m_i} = r_{m_i} \cdot h^o_{m_i} + (1 - r_{m_i}) \cdot w^o_{m_i}, \tag{10}$$
$$h_{m_i} = r_{m_i} \cdot w^o_{m_i} + (1 - r_{m_i}) \cdot h^o_{m_i}, \tag{11}$$

where $w^o_{m_i}$ and $h^o_{m_i}$ represent the original width and height of input data, respectively. r_{m_i} is a 0-1 variable designed to identify whether m_i is rotated compared with the original input data.

For each macro $m_i \in M$, it should be assigned to a chiplet and thus can be constrained as:

$$\sum_{c_k \in C} t_{i,k} = 1,\ \forall m_i \in M, \tag{12}$$

where $t_{i,k}$ is a 0-1 variable representing whether m_i is assigned to a chiplet c_k. For each chiplet $c_k \in C$, it should be assigned to a package and thus can be constrained as:

$$\sum_{p_l \in P} u_{k,l} = 1,\ \forall c_k \in C, \tag{13}$$

where $u_{k,l}$ is a 0-1 variable representing whether c_k is assigned to a package p_l.

To balance the area of each chiplet $c_i \in C$, the constraints are formulated as:

$$\sum_{m_i \in M} (t_{i,k} \cdot a_{m_i}) \geqslant a^{lb}_{c_k},\ \forall c_k \in C, \tag{14}$$

where a_{m_i} represents the area of m_i. $a^{lb}_{c_k}$ represents the lower bound of chiplet area a_{c_k}, and it is calculated as:

$$a^{lb}_{c_k} = \frac{a_M \times rr}{|C|},\ \forall c_k \in C, \tag{15}$$

where a_M and rr represent the total area of all macros and the relaxation ratio of a_M, respectively. rr is a user-defined coefficient. Since the shapes and areas of macros are different, $a^{lb}_{c_k}$ without the relaxation ratio rr may lead to the failure of chiplet partition. The similar constraints are formulated to balance the chiplets in each package.

For each macro $m_i \in M$ belonging to chiplet $c_k \in C$, m_i should be completely placed within c_k and thus can be constrained as:

$$x_{c_k} \leqslant x_{m_i} + W \cdot (1 - t_{i,k}), \tag{16}$$
$$y_{c_k} \leqslant y_{m_i} + H \cdot (1 - t_{i,k}), \tag{17}$$
$$x_{m_i} + w_{m_i} \leqslant x_{c_k} + w_{c_k} + W \cdot (1 - t_{i,k}), \tag{18}$$
$$y_{m_i} + h_{m_i} \leqslant y_{c_k} + h_{c_k} + H \cdot (1 - t_{i,k}). \tag{19}$$

Equations (16)–(19) limit that m_i cannot exceed the four boundaries of c_k. For example, m_i will not exceed the left boundary of c_k if $t_{i,k}$ is 1 which means m_i should completely be placed within c_k. The location of c_k completely within the belonging package p_l is also constrained by the formulations similar to Equations (16)–(19).

To realize the desired reliability, warpage is constrained for each package $p_l \in P$. However, the warpage model, shown by Equation (1), is not the linear or quadratic formulation, which makes the solving of the programming model difficult. Therefore, an approximate upper bounding model is proposed as below:

$$w^{'}(x) = \frac{t \cdot \Delta\alpha \cdot \Delta T}{2 \cdot \lambda \cdot D} \left[\frac{1}{2} x^2 - \frac{((kx)^2 + 1)/2 - 1}{k^2 \cosh(kl)} \right]. \tag{20}$$

The curves of the original model and the approximate model are shown in Fig. 4. The two models have the similar curves. Based on the experimental results, the approximate model is effective to optimize the reliability within the practical value range.

For each package $p_l \in P$, the warpages are constrained as:

$$wpg^x_{p_l} \leqslant wpgt_{p_l},\ \forall p_l \in P, \tag{21}$$
$$wpg^y_{p_l} \leqslant wpgt_{p_l},\ \forall p_l \in P, \tag{22}$$

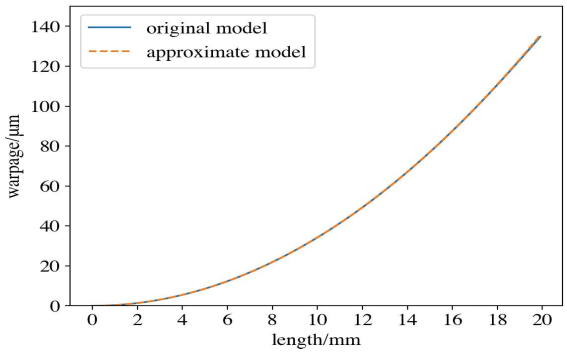


Figure 4: The curves of the warpage models.

where $wpg^x_{p_l}$ and $wpg^y_{p_l}$ represent the warpage on the direction of x-axis and the warpage on the direction of y-axis. They can be calculated based on Equation (20) as:

$$wpg^x_{p_l} = \frac{t\Delta\alpha\Delta T}{2\lambda D}\left[\frac{1}{2}(w_{p_l}/2)^2 - \frac{(k^2(w_{p_l}/2)^2+1)/2-1}{k^2\cosh(k(W/2))}\right], \quad (23)$$

$$wpg^y_{p_l} = \frac{t\Delta\alpha\Delta T}{2\lambda D}\left[\frac{1}{2}(h_{p_l}/2)^2 - \frac{(k^2(h_{p_l}/2)^2+1)/2-1}{k^2\cosh(k(H/2))}\right]. \quad (24)$$

Besides the warpage constraints introduced above, the spacing between macros and hotspot bumps cannot be less than the desired value. The bump margin constraints can be formulated as:

$$(x_{m_i} - x_{b_s})^2 + (y_{m_i} - y_{b_s})^2 \geqslant (d_{cc} + d_{mr})^2, \quad (25)$$

where x_{b_s} and y_{b_s} are the x-coordinate and y-coordinate of the center of hotspot bump b_s. d_{cc} is the radius of the circumscribed circle of the macro m_i. d_{mr} is the radius of the margin region. To maintain the proper spacing between macros and hotspot bumps, the circumscribed circles of macros are induced to satisfy the bump margin constraints as shown in Fig. 5.

Finally, the two-level objective is combined and formulated as:

$$\inf \ \beta_1 \times wl + \beta_2 \times oh_C + \beta_3 \times oh_P + \beta_4 \times \sum_{p_l \in P} a_{p_l} + \gamma \times \sum_{c_k \in C} a_{c_k}. \quad (26)$$

According to Equation (2), $\sum_{c_k \in C} a_{c_k}$ is added into Equation (3) to simultaneously optimize the two-level objective. Furthermore, since the priority of Equation (3) is higher than that of Equation (2), γ, which is a user-defined coefficient, is set to a tiny value.

3.2 Partition-Based MP Method

Due to the complexity of the one-pass MP (OPMP) method, it is difficult for the solutions to converge. Therefore, a partition-based MP method, parMP, is proposed corresponding to the conventional physical design flow. The first stage is the partitioning of macros. The macros should be partitioned into different chiplets. The chiplets should be partitioned into different packages. In this stage, an MP model is formulated and solved by MP solver. The second stage is the floorplanning of macros. Compared with OPMP, a simplified MP model is formulated and solved by MP solver in this stage.

For the first stage, Equations (12)–(13) are necessary to constrain the assignment of macros and chiplets. Equations (14)–(15) are also

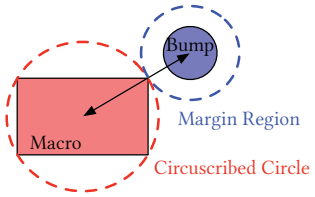


Figure 5: The illustration for the bump margin constraints.

necessary to keep the balance between different chiplets. The objective of this stage is formulated as:

$$\inf \quad \beta_2 \times oh_C + \beta_3 \times oh_P, \quad (27)$$

where oh_C and oh_P are calculated by the same method as those in Equation (26). The calculation based on the set of edges is shown below. The set of edges E is extracted from N. Each $e_{i,j} \in E$ connects the start macro m_i and the terminal macro m_j of a net. For each edge $e_{i,j} \in E$, the inter-chiplet cost $oh^c_{e_{i,j}}$, which is a 0-1 variable, is formulated as:

$$oh^c_{e_{i,j}} \geqslant t_{i,k} - t_{j,k}, \ \forall c_k \in C, \quad (28)$$

$$oh^c_{e_{i,j}} \geqslant t_{j,k} - t_{i,k}, \ \forall c_k \in C. \quad (29)$$

If the two macros belong to the same chiplet, $oh^c_{e_{i,j}}$ is 0. If the two macros belong to different chiplets, $oh^c_{e_{i,j}}$ is 1. Therefore, oh_C can be formulated as:

$$oh_C = \sum_{e_{i,j} \in E} oh^c_{e_{i,j}}. \quad (30)$$

For each edge $e_{i,j} \in E$, $v_{i,l}$ is used to identify whether macro m_i belongs to package p_l and can be formulated as:

$$v_{i,l} = max\left(0, \ \{t_{i,k} + u_{k,l} - 1 | \forall c_k \in C\}\right). \quad (31)$$

If $v_{i,l}$ is 1, m_i belongs to p_l; otherwise, m_i does not belong to p_l. Like Equations (28)–(29), the inter-package cost $oh^p_{e_{i,j}}$, which is a 0-1 variable, can be formulated as:

$$oh^p_{e_{i,j}} \geqslant v_{i,l} - v_{j,l}, \ \forall p_l \in P, \quad (32)$$

$$oh^p_{e_{i,j}} \geqslant v_{j,l} - v_{i,l}, \ \forall p_l \in P. \quad (33)$$

If the two macros belong to the same package, $oh^p_{e_{i,j}}$ is 0. If the two macros belong to different packages, $oh^p_{e_{i,j}}$ is 1. Therefore, oh_P can be formulated as:

$$oh_P = \sum_{e_{i,j} \in E} oh^p_{e_{i,j}}. \quad (34)$$

For the second stage, the constraints which are not considered in the first stage are formulated. The objective is formulated as:

$$\inf \ \beta_1 \times wl + \beta_4 \times \sum_{p_l \in P} a_{p_l} + \gamma \times \sum_{c_k \in C} a_{c_k}. \quad (35)$$

3.3 Hierarchical MP Method

According to the experimental results, the convergence speeds of the one-pass MP (OPMP) method and the partition-based MP (parMP) method are not fast enough. Therefore, a hierarchical MP method, hieMP, is proposed in this section. The design flow of hieMP is shown in Fig. 6. hieMP can be divided into two stages. The first stage is MP-based partition for chiplets and packages. In this stage, the same model as the first stage of parMP is built and solved by MP solver.

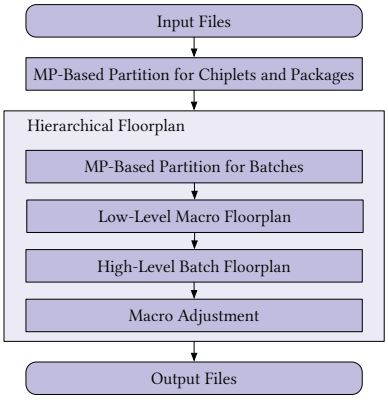


Figure 6: The design flow of hieMP.

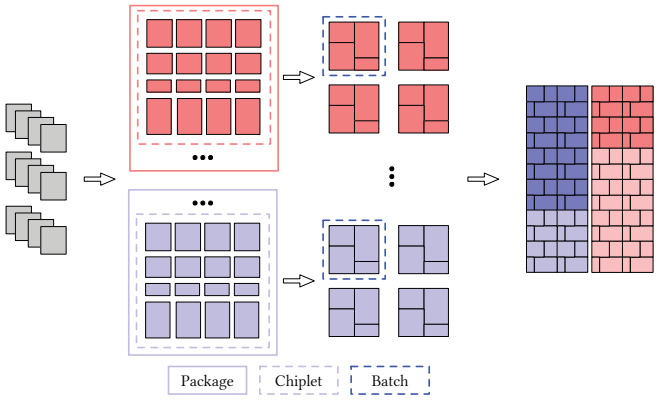


Figure 7: The illustration of hieMP.

The macros are assigned to chiplets, and the chiplets are assigned to packages. Both the inter-chiplet and the inter-package cost are optimized.

The second stage is hierarchical floorplan. This stage has four steps, including: (1) MP-based partition for batches; (2) low-level macro floorplan; (3) high-level batch floorplan; and (4) macro adjustment. The illustration of the design flow is shown in Fig. 7.

In the first step, a user-defined batch size bhs, which is the basic number of macros in a batch, is specified. bhs should be assigned a proper value for the tradeoff between runtime and solution quality. Users can set bhs according to the geometric property of each macro and the expected batch number of each chiplet. For each chiplet $c_k \in C$, the number of batches is calculated as:

$$|BH^{c_k}| = |M^{c_k}|/bhs, \tag{36}$$

where BH^{c_k} and M^{c_k} represent the set of batches and the set of macros belonging to c_k, respectively. For each macro $m_i^{c_k} \in M^{c_k}$, it should belong to one batch and thus can be constrained as:

$$\sum_{bh_j^{c_k} \in BH^{c_k}} w_{i,j} = 1, \ \forall m_i^{c_k} \in M^{c_k}, \tag{37}$$

where $w_{i,j}$ is a 0-1 variable representing whether $m_i^{c_k}$ is assigned to $bh_j^{c_k}$. To balance the number of macros in each batch, $w_{i,j}$ is constrained as:

$$bhs^{lb} \leqslant \sum_{m_i^{c_k} \in M^{c_k}} w_{i,j} \leqslant bhs^{ub}, \ \forall bh_j^{c_k} \in BH^{c_k}, \tag{38}$$

where bhs^{lb} and bhs^{ub} represent the lower bound and the upper bound of batch size. They are user-defined values. The objective of this step is to minimize the interconnection between different batches. The calculation of interconnection is similar to that of oh_C. The area of macros in each batch can also be balanced based on the similar inequality, however, it is not easy to control the area conditions since the differences of the macro areas can be very large.

As shown in Fig. 7, the macros of each chiplet are partitioned into different batches in the first step. Then, the floorplanning solution of each batch is generated in the second step. An MP method is formulated and solved by MP solver for the second step. To avoid the overlap between macros, the constraints like Equations (4)–(9) should be formulated. The objective is to minimize the area of each batch. For each batch $bh_j^{c_k} \in BH^{c_k}$, the objective is formulated as:

$$\inf \quad w_j^{c_k} \times h_j^{c_k}, \tag{39}$$

where $w_j^{c_k}$ and $h_j^{c_k}$ represent the width and height of $bh_j^{c_k}$, respectively. This step is an iterative process with the increase of dead space ratio. The fixed-outline of each batch $bh_j^{c_k}$ is calculated based on the dead space ratio dsr as:

$$w_{bh} = h_{bh} = \sqrt{a_M^{bh} \times (1 + dsr)}, \tag{40}$$

where w_{bh}, h_{bh}, and a_M^{bh} represent the width of the fixed-outline, the height of the fixed-outline, and the total area of the macros belonging to $bh_j^{c_k}$, respectively. The initial dsr is 0.1. For each iteration step, dsr increases by 0.5. If the feasible floorplanning solution is generated by the MP method, the iteration will end.

In the third step, all the batches are processed to construct the multi-package architecture. An MP model is formulated like that of the second stage of parMP and solved by MP solver. However, the complexity of this step is efficiently reduced compared with parMP because of the hierarchical structure. The transform from macros to batches reduces the number of the processed components.

Finally, the absolute coordinates of macros are calculated in the last step. Because the macros can only get the relative coordinates corresponding to batches in the second step.

4 EXPERIMENTAL RESULTS

The proposed methods in this work are implemented in C++ language on a Linux server with 64GB memory. Since the related work, such as [15] and [16], only processed the entire chiplets, the macros were not considered. Therefore, these benchmarks are not suitable for the problem in this work. The conventional benchmarks GSRC, which are also adopted in [10], are chosen in this work. Due to the limitation of pages, the details of GSRC can be obtained from [22]. The Gurobi optimizer [23] is adopted in this work to solve the MP models. In this paper, the proposed methods are tested for two package architectures, including multi-chiplet single-package (MCSP) architectures and multi-chiplet multi-package (MCMP) architectures. Each MCSP architecture includes two chiplets and one package, and each MCMP architecture includes four chiplets and two packages.

Table 1: The comparison of the proposed methods.

Method	Benchmark	HPWL (μm)	oh_C	PA (μm^2)	ar (%)	WPG ($\times 10^{-4}$ μm)	Cost	Time (min)	Ratio	Speedup
OPMP	n10	52189	28	233070	99.66	7.68	285541.32	483.26	1.00	1.00
	n30	164024	79	250120	97.47	6.52	414936.44	482.46	1.00	1.00
	n50	229227	171	274032	96.64	5.87	504971.65	480.65	1.00	1.00
	n100	N/A	N/A	N/A	N/A	N/A	N/A	480.07	N/A	N/A
	n200	N/A	N/A	N/A	N/A	N/A	N/A	480.14	N/A	N/A
	n300	N/A	N/A	N/A	N/A	N/A	N/A	480.22	N/A	N/A
parMP	n10	52929	18	239030	97.67	8.22	292141.33	34.85	1.02	13.87
	n30	160951	37	242638	99.59	6.90	403961.42	422.18	0.97	1.14
	n50	202828	105	279760	96.85	6.05	483640.71	420.27	0.96	1.14
	n100	N/A	N/A	N/A	N/A	N/A	N/A	426.34	N/A	N/A
	n200	N/A	N/A	N/A	N/A	N/A	N/A	480.14	N/A	N/A
	n300	N/A	N/A	N/A	N/A	N/A	N/A	480.16	N/A	N/A
hieMP	n10	60709	18	259598	94.36	9.75	320489.45	1.21	1.12	398.24
	n30	179152	39	266505	97.95	7.03	446049.61	60.89	1.07	7.92
	n50	233783	107	262980	98.59	6.09	497835.59	43.63	0.99	11.02
	n100	275231	140	227052	86.58	4.40	503684.97	73.03	N/A	N/A
	n200	525278	286	201600	98.85	6.54	729739.99	139.32	N/A	N/A
	n300	737555	331	315840	99.47	14.03	1056708.14	141.96	N/A	N/A

4.1 The Convergence of the Proposed Methods

In this section, the convergence of the proposed methods is analyzed. The data of this section is generated from n10 which is a benchmark with 10 macros. The package architecture is MCSP. To illustrate the difference of the convergence of the proposed methods, the gap between the feasible solution of each optimization step and the final optimized solution is normalized as shown in Fig. 8. The ratio represents the normalized gap. The ratio is 1 means that the current feasible solution is an initial solution. The ratio is 0 means that the final optimized solution has been generated. As shown in Fig. 8, the solutions gradually converge to the final optimized solution with the increase of time. hieMP has the best convergence speed compared with the other methods. hieMP can get the final optimized solution within 200 seconds. However, parMP and OPMP cannot converge within 1000 seconds. For parMP and OPMP, parMP has better convergence speed. In conclusion, the efficiency of hieMP is better than that of parMP and OPMP.

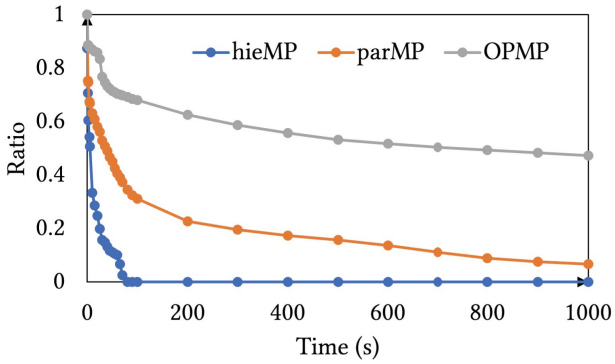


Figure 8: The convergence curves of n10.

4.2 The Comparison of the Proposed Methods

In this section, the optimization of different factors is compared between the proposed methods. The methods are analyzed based on the MCSP architecture. The experimental results are shown in Table 1, where "HPWL", "oh_C", "PA", "ar", "WPG", "Cost", "Time", "Ratio", and "Speedup" represent the estimated wirelength based on the half perimeter wirelength model, the inter-chiplet cost, the total area of packages, the ratio of chiplet to package area, the maximum warpage from the centers of packages to all the edges of packages, the total cost, the runtime, the ratio of the total cost compared with OPMP, and the speedup ratio compared with OPMP, respectively. "Cost" is calculated based on Equation (26). The coefficients β_1, β_2, β_3, β_4, and γ are set to 1, 10, 100, 1, and 0.00001, respectively. These settings are also adopted in the proposed methods.

Since the problem solved in this work is complicated, it is difficult for solutions to converge to the final optimized solution. Therefore, the runtime of the MP solver is limited to generate the best-effort solutions. The runtime limitation of OPMP is 8 hours. The runtime limitations of the two stages in parMP are 1 hour and 7 hours, respectively. The runtime limitation of the first stage in hieMP is 1 hour. For the second stage of hieMP, the runtime limitations of the partition of batches for each chiplet, the floorplan of each batch, and the high-level batch floorplan are 100 seconds, 50 seconds, and 1 hour, respectively.

For the two baselines, OPMP can achieve better solutions for n10 compared with parMP. However, since the convergence of parMP is better than that of OPMP, parMP can achieve better solutions compared with OPMP within the limited runtime for larger benchmarks. Furthermore, both OPMP and parMP cannot generate feasible solutions for large benchmarks, including n100, n200, and n300, within the limited runtime.

hieMP can generate feasible solutions for all benchmarks within the limited runtime. Since the macros are partitioned into multiple batches in hieMP and the batches are placed within limited bounding boxes, hieMP cannot achieve better solutions compared with OPMP

Table 2: The comparison of different package architectures.

Benchmark	MCSP				MCMP			
	oh_C	oh_P	PA (μm^2)	WPG ($\times 10^{-4}$ μm)	oh_C	oh_P	PA (μm^2)	WPG ($\times 10^{-4}$ μm)
n100	140	0	227052	4.40	266	152	205503	3.88
n200	286	0	201600	6.54	613	329	201701	4.82
n300	331	0	315840	14.03	666	379	323199	9.30
Ratio	1.00	N/A	1.00	1.00	2.02	N/A	0.98	0.76

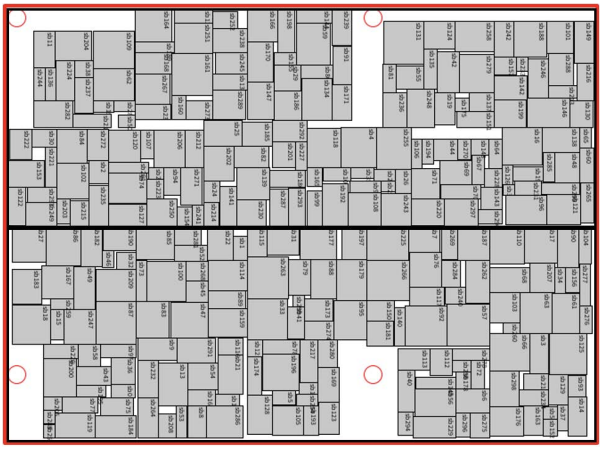


Figure 9: The MCSP architecture of n300.

and parMP for n10 and n30. For these small benchmarks, OPMP and parMP are applicable. However, hieMP can achieve better solutions with shorter runtime for n50, which shows the efficiency of hieMP. The solution of hieMP for n300 with the MCSP architecture is shown in Fig. 9. The red box and the black boxes represent the package and the two chiplets in the MCSP architecture, respectively.

4.3 The Comparison of Different Architectures

The proposed methods are also applicable for the MCMP architectures. The experimental results of hieMP are shown in Table 2. Since n10–50 are not larger enough to construct two packages, these benchmarks are not tested. According to the experimental results in Table 1, OPMP and parMP cannot converge within the limited runtime, therefore, these methods are not tested. The "oh_P" represents the inter-package cost. Since the small benchmarks are not suitable for the MCMP architectures, the three representative benchmarks, n100–300, are chosen to compare the differences of the two package architectures. As shown in Table 2, the factors related to the reliability issues, including "PA" and "WPG", are effectively optimized by the MCMP architectures. Compared with the MCSP architectures, the MCMP architectures can achieve 2% reduction of package area on average. The MCMP architectures can also achieve 24% reduction of warpage on average. However, it is predictable that the optimized "oh_C" and "oh_P" of the MCMP architectures are larger than those of the MCSP architectures. The users can choose proper package architectures based on the proposed methods according to the bottleneck of systems. For example, if the bottleneck of a system is reliability, such as warpage, the MCMP architectures can be chosen and optimized by hieMP. If the bottleneck of a system is interconnection cost, the MCSP architectures can be chosen and optimized by hieMP. Therefore, on the one hand, hieMP is effective to optimize the reliability issues for the multi-package co-design problem. On the other hand, hieMP can also help the users to choose proper package architectures with the tradeoff between performance and reliability.

Table 3: The effectiveness of warpage control.

Item	PA (mm^2)	Ratio	WPG (μm)	Ratio
Non-Control	215716	1.00	62.56	1.00
Control	205425	0.95	52.11	0.83

4.4 The Effectiveness of Warpage Control

For hieMP, the ability of warpage control is analyzed in this section. Since the unit of benchmarks, which is μm, is too small compared with the practical packages, the experimental results shown in Table 1 and Table 2 are not generated based on the challenging threshold of warpage. Furthermore, the approximate warpage model, shown in Equation (20), is not suitable for tiny package size, which is not practical. Therefore, the warpage control of hieMP is tested based on the modified unit, which is 50 μm, in this section. The threshold of warpage, which is the $wpgt_{p_l}$ of Equations (21)–(22), is set to 60 μm. The benchmark n200, where the $wpg^x_{p_l}$ and $wpg^y_{p_l}$ of Equations (21)–(22) have the largest gap, is chosen to test hieMP. The experimental results are listed in Table 3. Compared with the hieMP without the warpage control, the complete hieMP can achieve 5% reduction of package area, and 17% reduction of warpage. In conclusion, hieMP can effectively optimize reliability.

5 CONCLUSION

The multi-package architecture is a popular alternative to optimize the reliability and cost for advanced packages. To the best of our knowledge, this is the first work to overcome the primary challenge of the multi-package co-design problem, which is the tradeoff between interconnection cost and reliability. Three mathematical programming methods, including OPMP, parMP, and hieMP, are proposed to optimize the interconnection cost and reliability for the proposed problem. Compared with the two baselines, OPMP and parMP, the hierarchical mathematical programming method (hieMP) has the better efficiency to effectively optimize the interconnection cost and reliability issues, including warpage and bump stress, for large benchmarks. Furthermore, hieMP can also help for users to choose proper package architectures considering the tradeoff between performance and reliability.

REFERENCES

[1] Heterogeneous Integration Roadmap (HIR) 2021 Edition, IEEE. Available: https://eps.ieee.org/technology/heterogeneous-integration-roadmap/2021-edition.html.

[2] I. H.-R. Jiang, Y.-W. Chang, J.-L. Huang, and C. C.-P. Chen, "Intelligent Design Automation for Heterogeneous Integration," in *Proceedings of International Symposium on Physical Design (ISPD)*, pp. 105–106, 2022.

[3] Y. Feng and K. Ma, "Chiplet Actuary: A Quantitative Cost Model and Multi-Chiplet Architecture Exploration," in *Proceedings of Design Automation Conference (DAC)*, 2022. arXiv preprint arXiv:2203.12268.

[4] M. Ahmad, J. DeLaCruz, and A. Ramamurthy, "Heterogeneous Integration of Chiplets: Cost and Yield Tradeoff Analysis," in *Proceedings of International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems (EuroSimE)*, pp. 1–9, 2022.

[5] T. Hayashi, P. Y. Lin, R. Watanabe, and S. Ichikawa, "Development of Highly Reliable Crack Resistive Build-up Dielectric Material with Low Df Characteristic for Next-Gen 2.5D Packages," in *Proceedings of IEEE Electronic Components and Technology Conference (ECTC)*, pp. 570–576, 2021.

[6] C.-C. Lee, C.-W. Wang, C.-C. Lee, C.-Y. Chen, Y.-H. Chen, H.-C. Lee, and T.-S. Chou, "Warpage Estimation of Heterogeneous Panel-Level Fan-Out Package with Fine Line RDL and Extreme Thin Laminated Substrate Considering Molding Characteristics," in *Proceedings of IEEE Electronic Components and Technology Conference (ECTC)*, pp. 1500–1504, 2021.

[7] S. C. Chong, S. S. B. Lim, W. W. Seit, T. C. Chai, and D. C. Sanchez, "Comprehensive Study of Thermal Impact on Warpage Behaviour of FOWLP with Different Die to Mold Ratio," in *Proceedings of IEEE Electronic Components and Technology Conference (ECTC)*, pp. 1082–1087, 2021.

[8] F. X. Che, D. Ho, M. Z. Ding, and X. Zhang, "Modeling and Design Solutions to Overcome Warpage Challenge for Fan-Out Wafer Level Packaging (FO-WLP) Technology," in *Proceedings of IEEE Electronics Packaging and Technology Conference (EPTC)*, pp. 1–8, 2015.

[9] M. O. Hossen, J. L. Gonzalez, and M. S. Bakir, "Thermomechanical Analysis and Package-Level Optimization of Mechanically Flexible Interconnects for Interposer-on-Motherboard Assembly," in *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 8, no. 12, pp. 2081–2089, 2018.

[10] J.-M. Lin, T.-T. Chen, Y.-F. Chang, W.-Y. Chang, Y.-T. Shyu, Y.-J. Chang, and J.-M. Lu, "A Fast Thermal-Aware Fixed-Outline Floorplanning Methodology Based on Analytical Models," in *Proceedings of IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 1–8, 2018.

[11] M. Healy, M. Vittes, M. Ekpanyapong, C. Ballapuram, S. K. Lim, H.-H.S. Lee, and G.H. Loh, "Microarchitectural Floorplanning Under Performance and Thermal Tradeoff," in *Proceedings of Design Automation & Test in Europe Conference (DATE)*, pp. 1–6, 2006.

[12] W.-P. Lee, H.-Y. Liu, and Y.-W. Chang, "Voltage-Island Partitioning and Floorplanning Under Timing Constraints," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 28, no. 5, pp. 690–702, 2009.

[13] K. Blutman, H. Fatemi, A. Kapoor, A. B. Kahng, J. Li, and J. P. d. Gyvez, "Logic Design Partitioning for Stacked Power Domains," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems (TVLSI)*, vol. 25, no. 11, pp. 3045–3056, 2017.

[14] A. Henzinger, A. Noe, and C. Schulz, "ILP-Based Local Search for Graph Partitioning," in *ACM Journal of Experimental Algorithmics*, vol. 25, no. 1, pp. 1–26, 2020.

[15] W.-H. Liu, M.-S. Chang, and T.-C. Wang, "Floorplanning and Signal Assignment for Silicon Interposer-based 3D ICs," in *Proceedings of ACM/EDAC/IEEE Design Automation Conference (DAC)*, pp. 1–6, 2014.

[16] S. Osmolovskyi, J. Knechtel, I. L. Markov, and J. Lienig, "Optimal Die Placement for Interposer-Based 3D ICs," in *Proceedings of Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 513–520, 2018.

[17] M. Jung, D. Pan, and S. K. Lim, "Chip/Package Co-Analysis of Thermo-Mechanical Stress and Reliability in TSV-based 3D ICs," in *Proceedings of Design Automation Conference (DAC)*, pp. 317–326, 2012.

[18] R. Irwin, K. Sahoo, S. Pal, and S. S. Iyer, "Flexible Connectors and PCB Segmentation for Signaling and Power Delivery in Wafer-Scale Systems," in *Proceedings of IEEE Electronic Components and Technology Conference (ECTC)*, pp. 507–513, 2021.

[19] M.-Y. Tsai and Y.-W. Wang, "A Theoretical Solution for Thermal Warpage of Flip-Chip Packages", in *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 10, no. 1, pp. 72–78, 2020.

[20] S. Wen, J. Goodelle, V. Moua, K. Huang, and C. Xiao, "Cu Pillar Bump Design Parameters for Flip Chip Integration," in *Proceedings of IEEE Electronic Components and Technology Conference (ECTC)*, pp. 211–216, 2021.

[21] K. Sakuma, M. Farooq, P. Andry, C. Cabral, S. Rajalingam, D. McHerron, S. Li, R. Kastberg, and T. Wassick, "3D Die-Stack on Substrate (3D-DSS) Packaging Technology and FEM Analysis for 55 μm - 75 μm Mixed Pitch Interconnections on High Density Laminate," in *Proceedings of IEEE Electronic Components and Technology Conference (ECTC)*, pp. 292–297, 2021.

[22] GSRC. Available: http://vlsicad.eecs.umich.edu/BK/GSRCbench/.

[23] Gurobi Optimizer. Available: https://www.gurobi.com/.