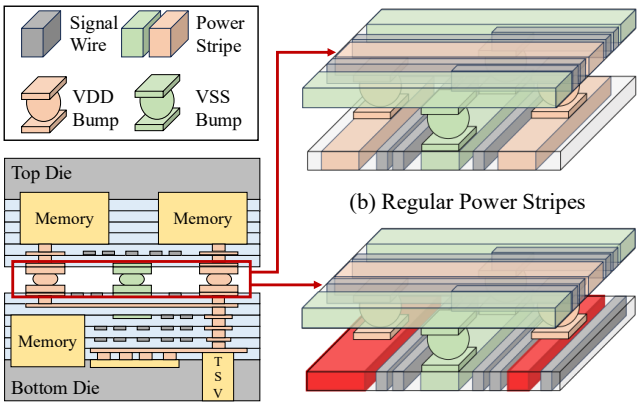
DPO-3D: Differentiable Power Delivery Network Optimization via Flexible Modeling for Routability and IR-Drop Tradeoff in Face-to-Face 3D ICs

Zhen Zhuang[1], Zheng Yang[2], Yuxuan Zhao[1], Jiawei Hu[2], Bei Yu[1], Sung Kyu Lim[3], Tsung-Yi Ho[1]
[1]The Chinese University of Hong Kong [2]Georgia Institute of Technology [3]University of Southern California

***Abstract*—With the advancement of technology nodes, increasing device density leads to serious routability challenges, especially for the promising face-to-face 3D ICs. Because most metal layer resources near the inter-die bonding layer are reserved by the stripes of power delivery networks (PDNs). Consequently, signal routing performance degrades as numerous nets must connect devices on both the bottom and top die through power-reserved metal layers. There are two primary challenges for 3D IC PDN optimization, including flexible power stripe optimization with global search capacity and addressing inter-die routability. Traditional 3D IC optimization methods using unified power stripes limit the search space, while single-die optimization methods cannot maintain the inter-die coherence. To address these challenges, we propose a differentiable PDN optimization framework, DPO-3D, for routability and IR-drop tradeoff in 3D ICs. At first, a novel modeling method that supports flexible 3D IC PDN optimization without compromising global search capability is proposed. Additionally, the modeling method effectively reduces design complexity. We further formulate the problem using an integer linear programming (ILP) model to address inter-die routability, achieving a good tradeoff between routability and IR-drop. Finally, we propose a GPU-accelerated differentiable method to solve the problem of overcoming the scalability issue of the ILP model. Compared with the state-of-the-art baseline, DPO-3D achieves a 7.10% DRC violation reduction, an 11.61% IR-drop reduction, and a 7.30× speedup on average.**



(a) Face-to-face 3D IC (b) Regular Power Stripes (c) Irregular Power Stripes

Fig. 1 The illustration of the PDN and signal nets in face-to-face 3D IC. (a) Face-to-face 3D IC. (b) The PDN constructed by regular power stripes. (c) The PDN constructed by irregular power stripes (red).

I. Introduction

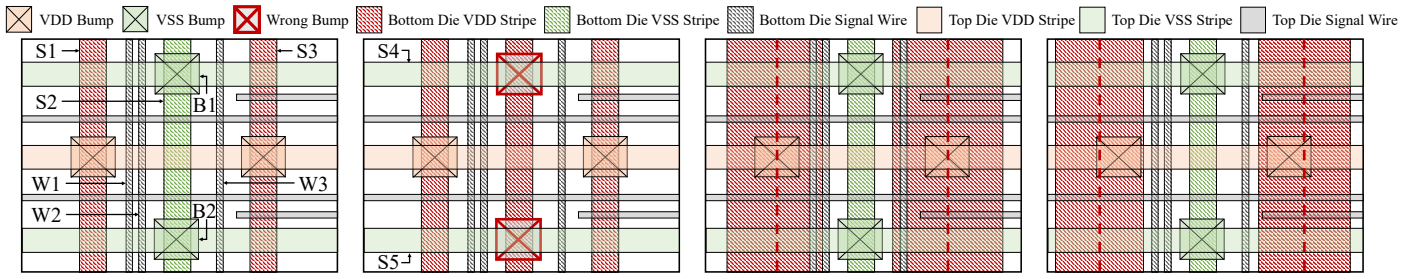
In the context of three-dimensional integrated circuits (3D ICs), the trade-off between routability and power integrity plays a crucial role in ensuring the overall performance and reliability of the power delivery network (PDN) [1]. The optimization of signal routability leads to thinner power delivery stripes, which can increase resistance and voltage drop, affecting power integrity. Power integrity, on the other hand, is the ability of the PDN to deliver stable and clean power to all components within the 3D IC. However, achieving optimal power integrity often requires wider power delivery stripes, which can impact signal routability. Balancing the trade-off between signal routability and power integrity involves making design decisions that prioritize both efficient routing of power delivery paths and maintaining the quality of power distribution. This may involve using advanced routing algorithms, optimizing power grid structures, and implementing power delivery techniques that minimize voltage drop while maximizing signal integrity.

Fig. 1 illustrates the PDN of face-to-face 3D IC. It is difficult to achieve a good trade-off between routability and power integrity since inter-die signal nets and PDNs are crowded into the metal layers close to the inter-die bump array. Nowadays, 3D ICs still use regular power stripes to construct PDN as shown in Fig. 1(b). In this paper, we use irregular power stripes to co-optimize routability and power integrity as shown in Fig. 1(c).

Single-die PDN optimization has been well studied. Lin *et al.* [2]–[5] propose several methods to optimize IR-drop and routability by inserting redundant power stripes. Chang *et al.* [6] propose a machine learning-based method to adjust power stripe width to optimize both routability and IR-drop. Lin *et al.* [5] propose a region-based method to refine and insert power stripes to optimize both routability and IR-drop. However, the single-die optimization method cannot maintain the inter-die coherence, i.e., the power stripes connected by inter-die bumps should belong to the same power net. Fig. 2(a) shows an original layout without optimization. For this region, the area of the VDD stripes in the bottom die should be increased to reduce IR-drop. Single-die optimization method may replace a VSS stripe with a VDD stripe in the bottom die as shown in Fig. 2(b). In this way, the power stripes $S2$, $S4$, and $S5$ connected by the bumps $B1$ and $B2$ belong to different power nets. Therefore, $B1$ and $B2$ cause the short of the VDD and VSS net. In conclusion, 3D IC needs specific methodologies to solve these special constraints.

The PDN optimization methods for 3D ICs use the regular power stripes to avoid the inter-die constraint violations, leading to limited search spaces. Zhu *et al.* [7] propose an iterative method to modify the width and pitch of unified power stripes to optimize both routability and IR-drop. However, the search space is limited, leading to a bad tradeoff between routability and IR-drop. In Fig. 2(c), the red dotted lines show the center lines of the VDD stripes $S1$ and $S3$. To increase the area of VDD nets, $S1$ and $S3$ are enlarged at the original positions in Fig. 2(a). Unfortunately, the VDD stripe $S1$ overlaps with the signal wires $W1$ and $W2$, while the VDD stripe $S3$ overlaps with the signal wire $W3$. Therefore, the routability is significantly degraded. In conclusion, more flexible optimization methods without search space limitations are necessary.

In this paper, we propose a differentiable PDN optimization framework, **DPO-3D**, for routability and IR-drop tradeoff in face-to-face 3D ICs. At first, we propose a novel and flexible modeling method



(a) Original layout. (b) Single-die optimization. (c) Previous 3D IC PDN optimization. (d) Flexible optimization.

Fig. 2 The comparison between different methods. (a) The original layout without optimization. The area of VDD nets should be increased in the bottom die to optimize IR-drop. (b) The single-die optimization methods replace a VSS stripe with a VDD stripe leading to the VDD and VSS short violations. (c) The previous work for 3D IC can only conduct unified optimization leading to signal wire short violations. (d) The proposed flexible optimization method for 3D IC can place any size power stripe in any position without design rule violations.

that can place any size power stripe in any position. Therefore, the search space is not limited. Then, an integer linear programming (ILP) model is formulated to achieve a good tradeoff between routability and IR-drop. However, the ILP-based method has scalability issues. Differentiable optimization methods have been widely used to efficiently solve electronic design automation (EDA) problems [8], [9]. Based on the proposed ILP model, we design a GPU-accelerated differentiable method to achieve scalable and efficient optimization. To the best of our knowledge, this is the first work to use the differentiable method to solve the routability and IR-drop co-optimization problem in 3D ICs. DPO-3D can achieve the best tradeoff between routability and IR-drop compared with other methods as shown in Fig. 2(d). The major contributions are shown below:

- We propose a novel modeling method to generate a new representation for 3D IC PDN optimization, which can transform the PDN routing problem into an assignment problem that can support flexible optimization, provide global search capacity, and reduce design complexity.
- We formulate the problem as an integer linear programming (ILP) model to simultaneously optimize routability and IR drop. The complicated 3D IC constraints, including the inter-die constraint, can be effectively solved.
- Based on the ILP model, we propose a GPU-accelerated differentiable optimization method to achieve scalable and efficient 3D IC PDN optimization for routability and IR-drop tradeoff.
- We propose a rank-based parameter calibration method to guide the co-optimization of routability and IR drop.
- Experimental results show that DPO-3D achieves a 22.75% DRC violation reduction and a 24.34% IR-drop reduction on average compared with the original flow. DPO-3D achieves a maximum speedup of 123.13× with an average speedup of 33.52× compared with the ILP-based method. Compared with the state-of-the-art baseline, DPO-3D achieves a 7.10% DRC violation reduction, an 11.61% IR-drop reduction, and a 7.30× speedup on average.

II. Preliminaries

A. Design Rule Constraints

The design rules are summarized below:

- Inter-die constraint: For each inter-die bump, if it is used by a power net, the power stripes on the top or bottom die connected to the bump should exist and belong to the same power net. This is a strict constraint that the PDN solution must satisfy to guarantee the connectivity of a power net.

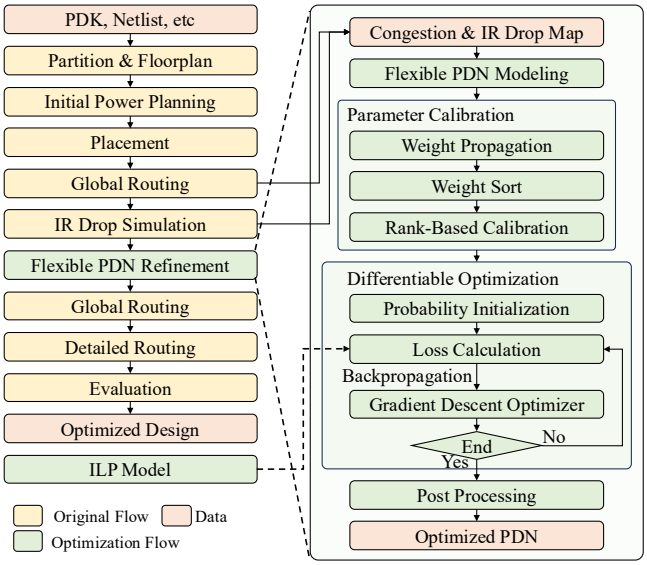


Fig. 3 The overview of DPO-3D.

- Minimum spacing: The minimum spacing between two wires is defined in each metal layer.
- Minimum width: The minimum width of each wire is defined in each metal layer.
- Detailed routing design rules: a set of design rules for signal detailed routing defined in the specification of practical designs.

B. Problem Formulation

Problem 1 (Flexible PDN Optimization for 3D IC)**.** *Given a set of metal layers, a set of inter-die bumps, a set of initial power stripes, a set of congestion maps, and a set of IR-drop maps, generate a new set of irregular VDD and VSS power stripes to minimize wirelength, DRC violations, and worst IR-drop.*

III. Technical Details

A. Overview

Fig. 3 shows the overview of DPO-3D embedded in the 3D IC design flow. Note that the proposed plug-in framework can be easily embedded into any 3D IC design flow. The flexible PDN optimizer is conducted after global routing. Therefore, it can obtain the congestion map generated from the global routing solution to optimize routability. The proposed optimizer also needs the IR-drop map generated from

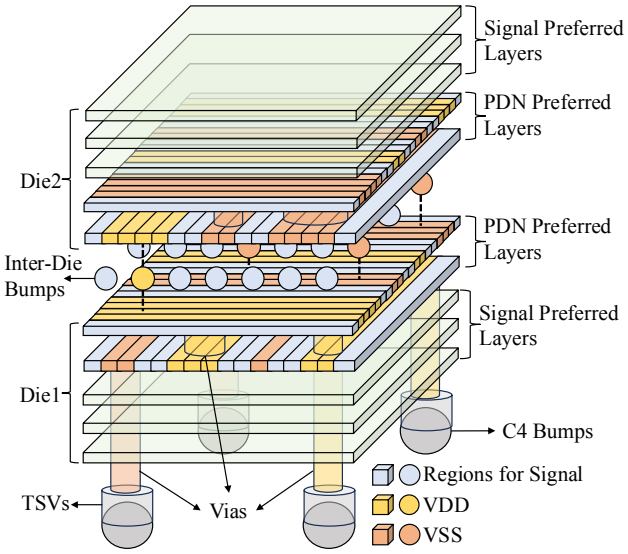


Fig. 4 The illustration for flexible 3D IC PDN modeling.

the IR-drop simulation to optimize power integrity. Based on the generated congestion and IR-drop maps, the routability and IR-drop can be co-optimized by the following techniques of the optimizer: flexible modeling, rank-based parameter calibration, the ILP model, differentiable optimization, and post processing.

B. Flexible 3D IC PDN Modeling

Flexible PDN routing should control several variables for each power net, including power stripe width, spacing, position, and number in each layer. Since the search space is too large, the metal layers used for PDN optimization are partitioned into a large number of small stripes among the preferred routing direction as shown in Fig. 4. The width of each small stripe is multiple times the sum of the minimum width and spacing of a signal wire. Therefore, the small stripe is called the *minimum stripe unit (MSU)*.

When the usage, i.e., VDD, VSS, or signal, of each MSU is determined, the PDN routing results are generated. Since the width of MSU is determined by the minimum width and spacing of signal wires, the search space is not limited. In other words, layout partition can still provide global search capacity. Each power stripe in the optimized results can be constructed by any number of adjacent MSUs. Hence, this modeling method can support flexible optimization with irregular power stripes. Furthermore, the new modeling changes the routing problem to the assignment problem, which effectively reduces the design complexity. In conclusion, the new modeling is more suitable for PDN optimization.

Layout partition is only used for metal layers. For the inter-die bump layer, each bump can be regarded as an MSU. Based on the new modeling method, an ILP-based method is proposed to solve the complicated constraints, including the inter-die constraint, and co-optimize routability and IR-drop.

C. ILP-Based Method

The flexible PDN refinement for 3D IC is formulated as an ILP problem, and solved by an ILP solver. To realize the desired design purposes, the following constraints should be satisfied.

For each MSU $u_{i,j} \in U_i$ in layer $l_i \in L$, it has three 0-1 variables $t_{i,j}^{vdd}$, $t_{i,j}^{vss}$, and $t_{i,j}^{signal}$. They are used to identify the usage of MSUs. $t_{i,j}^{vdd}$, $t_{i,j}^{vss}$, and $t_{i,j}^{signal}$ represent whether $u_{i,j}$ is used by VDD net, VSS net, and signal nets, respectively. Each MSU can only be used by one type of net and thus can be constrained as:

$$t_{i,j}^{vdd} + t_{i,j}^{vss} + t_{i,j}^{signal} = 1. \tag{1}$$

The inter-die constraint defined for each inter-die bump has two parts. One part constrains the relationship between the bump and the MSUs connected to the bump on the top die. The other part constrains the relationship between the bump and the MSUs connected to the bump on the bottom die. The two parts are similar. $I_{i,j}^{bot}$ is the set of k of the MSUs $t_{i-1,k}$ connected to the bump $u_{i,j}$ on the bottom die. $I_{i,j}^{top}$ is the set of k of the MSUs $t_{i+1,k}$ connected to the bump $u_{i,j}$ on the top die. Then, Equations (2) to (6) are used to formulate the inter-die constraint on the bottom die side.

If the bump $u_{i,j}$ belongs to the VDD net, at least one of the MSUs connected to $u_{i,j}$ must belong to the VDD net. Thus, it can be constrained as:

$$t_{i,j}^{vdd} \leqslant \sum_{k \in I_{i,j}^{bot}} t_{i-1,k}^{vdd}. \tag{2}$$

If the bump $u_{i,j}$ belongs to the VSS net, at least one of the MSUs connected to $u_{i,j}$ must belong to the VSS net. Thus, it can be constrained as:

$$t_{i,j}^{vss} \leqslant \sum_{k \in I_{i,j}^{bot}} t_{i-1,k}^{vss}. \tag{3}$$

The following constraints can guarantee that if a power net does not use the bump $u_{i,j}$, the MSUs connected to $u_{i,j}$ cannot be selected by the corresponding power net.

$$t_{i,j}^{vdd} \geqslant t_{i-1,k}^{vdd},\ \forall k \in I_{i,j}^{bot}, \tag{4}$$

$$t_{i,j}^{vss} \geqslant t_{i-1,k}^{vss},\ \forall k \in I_{i,j}^{bot}. \tag{5}$$

The combination of Equations (2) to (5) can maintain the power consistency between the bump and the corresponding MSUs.

For signal nets, the constraints are not as strict as those of power nets. Because if the bump $u_{i,j}$ belongs to a power net, signal MSUs connected to $u_{i,j}$ can exist since it only means the stripe is reserved for signal nets. Signal nets cannot connect $u_{i,j}$ but can use the other regions of the stripe. Therefore, we only need to guarantee if the bump $u_{i,j}$ belongs to signal nets, all of the MSUs connected to $u_{i,j}$ should belong to signal nets as below:

$$t_{i,j}^{signal} \leqslant t_{i-1,k}^{signal},\ \forall k \in I_{i,j}^{bot}. \tag{6}$$

Unbalanced power stripe distribution always leads to the degradation of IR drop. To achieve balanced power stripe distribution, the ratio of VDD nets, VSS nets, and signal nets are limited in each metal layer as below:

$$LB_T \leqslant \sum_{j=0}^{|U_i|} t_{i,j}^{T} \leqslant UB_T, \tag{7}$$

where LB_T represents the lower bound of MSU number for net $T \in \{vdd, vss, signal\}$. UB_T represents the upper bound of MSU number for net $T \in \{vdd, vss, signal\}$.

To achieve more robust control of power stripe distribution balance, The ratio of each net type is also limited in multiple sub-regions of each metal layer as below:

$$LB_T \leqslant \sum_{u_{i,k} \in U_i^m} t_{i,k}^{T} \leqslant UB_T, \tag{8}$$

where $U_i^m \subset U_i$ represents the MSUs belonging to a sub-region of metal layer $l_i \in L$. In our implementation, each metal layer is equally partitioned into three sub-regions.

If two adjacent MSUs belong to different power nets, the two power nets will violate the minimum spacing design rule and thus can be constrained as:

$$t_{i,j}^{vdd} + t_{i,j+1}^{vss} \leqslant 1, \tag{9}$$

$$t_{i,j}^{vss} + t_{i,j+1}^{vdd} \leqslant 1. \tag{10}$$

Finally, the objective is formulated as:

$$\min \sum_{i=0}^{|L|} \sum_{j=0}^{|U_i|} (w_{i,j}^{vdd} \times t_{i,j}^{vdd} + w_{i,j}^{vss} \times t_{i,j}^{vss} + w_{i,j}^{signal} \times t_{i,j}^{signal}), \tag{11}$$

where $w_{i,j}^{vdd}$, $w_{i,j}^{vss}$, and $w_{i,j}^{signal}$ represent the weights of $u_{i,j}^{vdd}$, $u_{i,j}^{vss}$, and $u_{i,j}^{signal}$ to guide the optimization, respectively. Based on the given weights, the ILP model can be used to find the optimal solution.

D. GPU-Accelerated Differentiable Optimization

Since the proposed ILP-based method should tackle a lot of constraints, the efficiency will significantly decrease as the design scale grows. Therefore, the scalability of the ILP-based method is limited. In this section, we propose a differentiable method to efficiently solve the optimization problem for 3D IC PDN.

The original objective in the proposed ILP model shown in Equation (11) cannot be directly optimized by differentiable methods due to the binary variables and constraints. At first, we define new unconstrained trainable variables $p_{i,j}^T$ instead of the original constrained binary variables $t_{i,j}^T$, where $T \in \{vdd, vss, signal\}$. Then, we introduce a `softmax` layer to map unconstrained trainable variables to probabilities. To avoid the deterministic nature of the `softmax` function leading to local optima, we employ the `gumbel_softmax` function [10] as below:

$$q_{i,j}^{type} = \frac{exp((p_{i,j}^{type} + g_{i,j}^{type})/\tau)}{\sum_{T \in \{vdd,vss,signal\}} exp((p_{i,j}^{T} + g_{i,j}^{T})/\tau)}, \tag{12}$$

where $g_{i,j}^{type}$ is the noise sampling from the Gumbel distribution. $type \in \{vdd, vss, signal\}$. This function is a stochastic variant of the `softmax` function. τ can be set to a small value to ensure that the final sampling of the MSU assignment is discrete. In this way, the final probabilities associated with MSUs closely approximate either 0 or 1.

Based on the `gumbel_softmax` function, the binary selection problem in the original ILP model is transformed into a continuous probability optimization between the three choices for each MSU. In this way, the unconstrained trainable variables can be used for differentiable optimization. The constraints in Equation (1) can also be reduced.

However, other constraints must be reduced while satisfying the design rules to achieve efficient differentiable optimization. We categorize the remaining constraints into three types: bonding type, region ratio type, and spacing type. The bonding type corresponds to the inter-die constraint as shown in Equations (2) to (6). The region ratio type corresponds to Equations (7) and (8). The spacing type corresponds to Equations (9) and (10). For each type, we design a loss function to transform the constraints into an unconstrained differentiable minimization objective. Therefore, the constrained ILP model is transformed into an unconstrained optimization problem.

To enable the loss functions to serve as the original constraints, the following two principles should be satisfied as much as possible. ① The optimal boundaries of the loss function correspond to the legal solutions. In other words, the minimum value of the loss function corresponds to the solution that satisfies the constraints. ② The process of minimizing the loss function matches the trend of design rule violation reduction. Considering the two principles, we design the following loss functions.

The loss function for bonding-type constraints is shown below:

$$\mathcal{L}_{Bond} = \sum_{i=0}^{|\mathcal{G}|} [1 - (\frac{1}{|\mathcal{G}_i|} \sum_{j=0}^{|\mathcal{G}_i|} q_j^{vdd} - \frac{1}{|\mathcal{G}_i|} \sum_{j=0}^{|\mathcal{G}_i|} q_j^{vss})^2 - (\frac{1}{|\mathcal{G}_i|} \sum_{j=0}^{|\mathcal{G}_i|} q_j^{signal})^3], \tag{13}$$

where $\mathcal{G}$ represents the groups of MSUs. The MSUs in a group $\mathcal{G}_i \in \mathcal{G}$ are connected via inter-die bumps. For the inter-die constraint, the three legal situations in one group $\mathcal{G}_i$ are shown below. ① All the MSUs belong to signal nets, since the inter-die bumps are signal bumps. ② Most MSUs belong to VDD nets and few MSUs are reserved for signal nets, since the inter-die bumps are VDD bumps. No MSU belongs to VSS nets. ③ Most MSUs belong to VSS nets and few MSUs are reserved for signal nets, since the inter-die bumps are VSS bumps. No MSU belongs to VDD nets. When the solution is legal, the loss is the minimum value 0. The more illegal the solution, i.e., the more VDD and VSS nets are mixed in the group $\mathcal{G}_i$, the greater the loss value.

The loss function for region ratio-type constraints is shown below:

$$\mathcal{L}_{RR} = \sum_{i=0}^{|\mathcal{R}|} \sum_{T \in \mathcal{T}} [(\frac{1}{|\mathcal{R}_i|} \sum_{j=0}^{|\mathcal{R}_i|} q_j^T - LB_T)^2 + (\frac{1}{|\mathcal{R}_i|} \sum_{j=0}^{|\mathcal{R}_i|} q_j^T - UB_T)^2], \tag{14}$$

where $\mathcal{R}$ represents the groups of MSUs to limit the net ratio in different regions. $U_i \in \mathcal{R}$ and $U_i^m \in \mathcal{R}$. $\mathcal{T} = \{vdd, vss, signal\}$ represents net types. In each region, when the average probability of one type of net is $[LB_T, UB_T]$, i.e., the net ratio is within the required interval, the loss is the minimum value. Furthermore, the loss value is directly proportional to the difference between the net ratio and the required interval.

The loss function for spacing-type constraints is shown below:

$$\mathcal{L}_S = \sum_{i=0}^{|L|} \sum_{j=1}^{|U_i|} (q_{i,j}^{vdd} \times q_{i,j-1}^{vss} + q_{i,j}^{vss} \times q_{i,j-1}^{vdd}). \tag{15}$$

The loss is calculated for each pair of adjacent MSUs. For each pair, when the solution is legal, i.e., a VDD net is not adjacent to a VSS net, the loss is the minimum value 0. Furthermore, the loss value is directly proportional to the illegal probability.

The original objective in the ILP model is used to build the basic loss function as below:

$$\mathcal{L}_{Basic} = \sum_{i=0}^{|L|} \sum_{j=0}^{|U_i|} \sum_{T \in \mathcal{T}} (w_{i,j}^T \times q_{i,j}^T). \tag{16}$$

Finally, the sum of all the loss functions yields the final loss function as below:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{Basic} + \beta \mathcal{L}_{Bond} + \gamma \mathcal{L}_{RR} + \theta \mathcal{L}_S. \tag{17}$$

We use the deep learning toolkit *PyTorch* to implement our differentiable solver. All data calculated in Equations (12) to (17) are organized as vectors $\mathbf{p}$, $\mathbf{q}$, and $\mathbf{w}$. $\mathbf{p}$ is the vector of $p_{i,j}^T$. $\mathbf{q}$ is the vector of $q_{i,j}^T$. $\mathbf{w}$ is the vector of $w_{i,j}^T$. $T = \{vdd, vss, signal\}$. Leveraging *PyTorch*, we benefit from its robust and efficient support

for vector operations, automatic gradient derivation, and optimization. This toolkit seamlessly integrates GPU acceleration, further enhancing computational efficiency and enabling rapid optimization.

E. Post Processing

When the maximum epoch count is reached, a PDN optimization solution is generated by determining the net type of each MSU based on the associated probabilities. However, three post-processing steps should be performed to ensure that the generated solution is feasible.

The first step is the net type selection for the MSU groups $\mathcal{G}$ associated with the inter-die constraint, since the inter-die constraint is the most complicated hard constraint. For each group, the VDD probability, VSS probability, and signal probability are calculated for all MSUs. The net type with the largest probability is selected as the type of the inter-die bumps in the group. Then, the types of illegal MSUs are changed to signal or feasible power net types according to probabilities. The second step is the net type selection for other MSUs based on the corresponding probabilities. The third step is to check the adjacent MSUs. If the adjacent MSUs belong to different power net types, one of the MSUs will be reassigned to a signal type. This modification has no conflict with other constraints.

F. Rank-Based Parameter Calibration

At first, the VDD IR-drop and VSS IR-drop of each power stripe are extracted from the simulation results of the initial PDN configuration. The congestion information, i.e., routing overflow in each tile, is extracted from the global routing result. Then, the MSUs that overlapped with objects, i.e., congestion tiles or power stripes, should extract initial weights from the input data. Since the size of an object is not the same as that of an MSU, an area ratio-based calculation method is proposed to update the MSU weights according to the overlapped objects. According to the coordinates and shapes of an MSU $u_{i,j}$ and an overlapped object o_k, the area ratio of $u_{i,j}$ on o_k can be calculated as $ar_{i,j,k}$. The initial weight $w_{i,j}^T$ of $u_{i,j}$ can be updated as:

$$w_{i,j}^T = w_{i,j}^T + p_k \times ar_{i,j,k}, \tag{18}$$

where $T \in \{vdd, vss, signal\}$ and p_k represents the input data of o_k. If o_k is a congestion tile, it will be used to update the signal weights where $T = signal$. If o_k is a VDD power stripe, it will be used to update the VDD weights where $T = vdd$. If o_k is a VSS power stripe, it will be used to update the VSS weights where $T = vss$.

The $w_{i,j}^{vdd}$ and $w_{i,j}^{vss}$ of many MSUs are zero as plenty of regions do not have any power stripes. Furthermore, the $w_{i,j}^{signal}$ of many MSUs are zero as plenty of tiles do not have any routing overflow. Therefore, the initial weights only include sparse information, degrading the quality of the co-optimization of routability and power integrity. We propose a weight propagation method to enrich the sparse data.

The closer an MSU is to an object, i.e., a congestion tile or a power stripe, the more degradation risks, i.e., congestion or IR-drop, will be caused in the MSU. Hence, if we want to enrich the sparse data, the non-zero data can be spread to the regions with zero data. The weights are propagated to the adjacent MSUs one by one in the left and right directions until the next MSU has a non-zero weight. Since the degradation risk of each MSU decreases with increasing distance, the weight is updated as below:

$$w_{i,j-k} = w_{i,j} \times wr^k, \tag{19}$$

$$w_{i,j+k} = w_{i,j} \times wr^k, \tag{20}$$

where wr represents the weight relaxation coefficient that should be smaller than 1. k represents the step number away from the MSU with a non-zero weight $u_{i,j}$. An MSU with zero weight may get relaxed weights from the two directions. In this case, it will choose the maximum one.

When the informative data has been obtained, a method should be designed to measure the importance of different metrics for each MSU since the number of congestion and IR-drop values have different orders. Therefore, the weights of one metric in one layer are sorted in descending order. Then, the rank $r_{i,j}$ of each MSU $u_{i,j}$ can be obtained. Ranks can effectively make different metrics in the same measurement system. To better measure each MSU, the ranks, i.e., $r_{i,j}^{vdd}$, $r_{i,j}^{vss}$, and $r_{i,j}^{signal}$, are normalized as below:

$$w_{i,j}^T = \frac{r_{i,j}^T}{r_{i,j}^{vdd} + r_{i,j}^{vss} + r_{i,j}^{signal}},\ T \in \{vdd, vss, signal\}. \tag{21}$$

Based on the proposed rank-based parameter calibration method, the updated weights can be effectively used to guide the co-optimization of congestion and IR-drop.

IV. Experimental Results

A. Experimental Setup

In this work, the proposed ILP model is implemented in C++ language. The proposed differentiable optimization method is implemented using *PyTorch*, a popular deep learning toolkit. The program of the ILP-based method is run on a PC with an Apple M1 chip and 8GB memory. The program of the differentiable optimization method is run on a NVIDIA GeForce RTX 3090 GPU with 24GB memory. The input data is generated based on a practical 3D IC design flow [11]. The output data is embedded in the design flow to evaluate performances. Within the design flow, the entire 3D Place-and-Route (PnR) flow is executed using *Synopsys IC Compiler II* [12]. *ANSYS RedHawk* [13] is used for IR drop analysis. Both PnR flow and IR drop simulations are performed on a Linux server equipped with 128 Intel Xeon Gold 6454S 2.20GHz CPU processors. The ILP model is solved by Gurobi [14].

We conduct experiments on four testcases. The four testcases are generated from three designs: one MAERI accelerator with 64 processing elements and 16 SRAM buffers [15], one MAERI accelerator with 128 processing elements and 28 SRAM buffers [15], and OpenPiton [16] utilizing Face-to-Face (F2F) 3D integration. "testcase1" pertains to the OpenPiton design. "testcase2" and "testcase3" correspond to two distinct MAERI-64PE designs. "testcase4" corresponds to the MAERI-128PE design. The designs are implemented using an open-source predictive 3nm Process Design Kit (PDK).

Each die in the 3D stack utilizes six metal layers for routing. Specifically, metal layers M5 and M6 in the bottom die, along with metal layer M6 in the top die, are allocated for PDN routing. In this work, the PDN optimization is conducted for M6 of both the bottom die and the top die since M5 should maintain the original layout to provide stable power delivery for instances. The PDN configurations of the original design flow for all four testcases use uniform power stripes with a width of 1.5 μm and a pitch of 4 μm.

B. The Comparison between Different Methods

TABLE I shows the experimental results of the original design flow [11], the ILP-based method, and the proposed differentiable method, DPO-3D, about wirelength, design rule constraint (DRC) violation, and worst instance IR-drop. Compared with the original design flow, DPO-3D achieves a 0.11% wirelength reduction, a 22.75% DRC violation reduction, and a 24.34% IR-drop reduction on average. Although the ILP-based method has a better ability to reduce wirelength

TABLE I The comparison between the ILP-based method and DPO-3D. "WL" and "# DRC Vio." represent wirelength and the DRC violation number, respectively. The units of "WL", "Worst IR drop", and "Runtime" are m, mV, and min, respectively.

Testcase	Original Flow [11]			ILP-Based Method				DPO-3D			
	WL	# DRC Vio	Worst IR-Drop	WL	# DRC Vio	Worst IR-Drop	Runtime	WL	# DRC Vio	Worst IR-Drop	Runtime
testcase1	1.46	21247	119.30	**1.44**	**13255**	133.80	0.90	1.45	14060	**103.66**	**0.85**
testcase2	0.71	13704	338.20	0.71	**10593**	142.20	4.10	**0.71**	12189	**137.79**	**1.10**
testcase3	0.71	17441	99.80	0.71	**12843**	97.50	7.52	**0.71**	12857	**93.91**	**1.22**
testcase4	1.73	63746	223.00	1.72	**47665**	182.81	346.76	**1.72**	51107	**180.43**	**2.82**
Ratio	1.00	1.31	1.48	1.00	**0.94**	1.09	33.52	**1.00**	1.00	**1.00**	**1.00**

TABLE II The comparison between the baseline [7] and DPO-3D. "WL" and "# DRC Vio." represent wirelength and the DRC violation number, respectively. The units of "WL", "Worst IR drop", and "Runtime" are m, mV, and min, respectively.

Testcase	Baseline [7]				DPO-3D			
	WL	# DRC Vio	Worst IR-Drop	Runtime	WL	# DRC Vio	Worst IR-Drop	Runtime
testcase1	1.45	15248	159.30	7.48	**1.45**	**14060**	**103.66**	**0.85**
testcase2	0.72	12982	140.10	11.00	**0.71**	**12189**	**137.79**	**1.10**
testcase3	0.72	14093	103.20	8.38	**0.71**	**12857**	**93.91**	**1.22**
testcase4	1.72	54211	182.00	9.93	**1.72**	**51107**	**180.43**	**2.82**
Ratio	1.01	1.08	1.17	7.30	**1.00**	**1.00**	**1.00**	**1.00**

TABLE III The detailed analysis on an AI accelerator design.

Metrics	MAERI 128PE	
	Original Flow	Ours
Effective Frequency (GHz)	2.14	**2.15 (+0.39%)**
WNS (ps)	-67.45	**-65.63 (+2.70%)**
TNS (ps)	-60743.06	**-33705.42 (+44.51%)**
# Violation Path	5392	**3332 (-38.20%)**
Wirelength (m)	1.73	**1.72 (-0.56%)**
# DRC Violation	63746	**51107 (-19.83%)**
Worst IR drop (mV)	223.00	**180.43 (-19.09%)**

and DRC violations, DPO-3D achieves up to 22.53% reduction on IR-drop compared with the ILP-based method. Furthermore, the IR-drop of testcase1 is not optimized by the ILP-based method, while DPO-3D efficiently reduces the IR-drop of testcase1. Therefore, DPO-3D achieves a better tradeoff between routability and IR-drop compared with the original flow and the ILP-based method. In addition, DPO-3D achieves a 33.52× speedup on average. DPO-3D even achieves a 123.13× speedup on the largest testcase, testcase4.

In this section, DPO-3D is compared with the state-of-the-art (SOTA) baseline for routability and IR-drop tradeoff in 3D IC [7]. The experimental results are shown in TABLE II. DPO-3D achieves a 0.89% wirelength reduction, a 7.10% DRC violation reduction, and an 11.61% IR-drop reduction on average. DPO-3D also achieves a 7.30× speedup on average.

C. Detailed Case Study

TABLE III shows a more detailed case study on testcase4, a MAERI-128PE design. In addition to its impact on routability and IR-drop, DPO-3D also has the capacity to enhance timing performance. Compared with the original flow, DPO-3D achieves a 44.51% total negative slack improvement and a 38.20% timing violation path reduction. Fig. 5 shows the IR-drop and DRC violation distributions of the original flow and DPO-3D of testcase4. In addition to optimizing the worst IR-drop, DPO-3D has effectively eliminated the crimson hotspot areas observed in the original IR drop distribution. The grid with more than 80 violations is defined as the DRC violation hotspot. In Fig. 5, the original flow has 95 DRC violation hotspots, while DPO-3D has 63 DRC violation hotspots. DPO-3D achieves a 33.68% DRC violation hotspot reduction. The detailed case study shows the comprehensive optimization ability of DPO-3D.

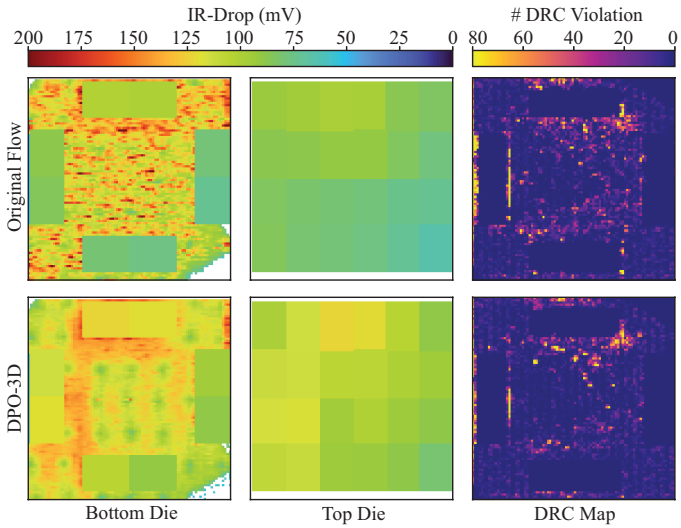


Fig. 5 The IR-Drop and DRC distributions of testcase4.

V. Conclusion

In this work, we propose a differentiable PDN optimization framework, DPO-3D, for routability and IR-drop tradeoff in face-to-face 3D ICs. The proposed framework uses a novel modeling method, which can support flexible optimization to generate irregular power stripes and provide global search capacity. Based on the assignment modeling, an ILP-based model is formulated to solve the complicated constraints, including the special inter-die constraint for 3D IC, and achieve the routability and IR-drop co-optimization. To improve the scalability and efficiency, a GPU-accelerated differentiable optimization method is proposed instead of the ILP model.

Acknowledgment

The research work described in this paper was conducted in the JC STEM Lab of Intelligent Design Automation funded by The Hong Kong Jockey Club Charities Trust. This work is jointly supported by the Research Grants Council of Hong Kong SAR No. CUHK14211324 and No. CUHK7010840.

REFERENCES

[1] A. Agnesina, M. Brunion, J. Kim, A. Garcia-Ortiz, D. Milojevic, F. Catthoor, M. Perumkunnil, and S. K. Lim, "Power, performance, area and cost analysis of memory-on-logic face-to-face bonded 3d processor designs," in *2021 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*. IEEE, 2021, pp. 1–6.

[2] J.-M. Lin, Y.-T. Kung, Z.-Y. Huang, and I.-R. Chen, "A fast power network optimization algorithm for improving dynamic ir-drop," in *Proceedings of the 2021 International Symposium on Physical Design*, 2021, pp. 91–98.

[3] J.-M. Lin, Y.-T. Chen, Y.-T. Kung, and H.-J. Lin, "Voltage-drop optimization through insertion of extra stripes to a power delivery network," in *Proceedings of the 2023 International Symposium on Physical Design*, 2023, pp. 35–43.

[4] C.-P. Lu, I. H.-R. Jiang, C.-C. Peng, M. M. Mohd Razha, and A. Uber, "Power sub-mesh construction in multiple power domain design with ir drop and routability optimization," in *Proceedings of the 2024 International Symposium on Physical Design*, 2024, pp. 205–212.

[5] J.-M. Lin, J.-S. Syu, and I.-R. Chen, "Macro-aware row-style power delivery network design for better routability," in *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 2018, pp. 1–8.

[6] W.-H. Chang, L.-D. Chen, C.-H. Lin, S.-P. Mu, M. C.-T. Chao, C.-H. Tsai, and Y.-C. Chiu, "Generating routing-driven power distribution networks with machine-learning technique," in *Proceedings of the 2016 on International Symposium on Physical Design*, 2016, pp. 145–152.

[7] L. Zhu, C. Jo, and S. K. Lim, "Power delivery solutions and ppa impacts in micro-bump and hybrid-bonding 3d ics," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 12, no. 12, pp. 1969–1982, 2022.

[8] W. Li, R. Liang, A. Agnesina, H. Yang, C.-T. Ho, A. Rajaram, and H. Ren, "Dgr: Differentiable global router," in *Proceedings of the 61st ACM/IEEE Design Automation Conference*, 2024, pp. 1–6.

[9] Y.-C. Lu, K. Kunal, G. Pradipta, R. Liang, R. Gandikota, and H. Ren, "Lego-size: Llm-enhanced gpu-optimized signoff-accurate differentiable vlsi gate sizing in advanced nodes," in *Proceedings of the 2025 International Symposium on Physical Design*, 2025, pp. 152–162.

[10] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[11] L. Bamberg, A. García-Ortiz, L. Zhu, S. Pentapati, S. K. Lim *et al.*, "Macro-3d: A physical design methodology for face-to-face-stacked heterogeneous 3d ics," in *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2020, pp. 37–42.

[12] "Synopsys ic compiler II," https://www.synopsys.com/implementation-and-signoff/physical-implementation/ic-compiler.html.

[13] "Ansys redhawk-sc," https://www.ansys.com/products/semiconductors/ansys-redhawk-sc.

[14] "Gurobi optimizer," https://www.gurobi.com/.

[15] H. Kwon, A. Samajdar, and T. Krishna, "Maeri: Enabling flexible dataflow mapping over dnn accelerators via reconfigurable interconnects," *SIGPLAN Not.*, vol. 53, no. 2, p. 461–475, Mar. 2018. [Online]. Available: https://doi.org/10.1145/3296957.3173176

[16] J. Balkind, M. McKeown, Y. Fu, T. Nguyen, Y. Zhou, A. Lavrov, M. Shahrad, A. Fuchs, S. Payne, X. Liang *et al.*, "Openpiton: An open source manycore research framework," *ACM SIGPLAN Notices*, vol. 51, no. 4, pp. 217–232, 2016.