GNN-MLS: Signal Routing in Mixed-Node 3D ICs through GNN-Assisted Metal Layer Sharing

Jiawei Hu[1], Pruek Vanna-iampikul[3], Zhen Zhuang[2], Tsung-Yi Ho[2], and Sung Kyu Lim[1]
[1]School of ECE, Georgia Institute of Technology, Atlanta, GA
[2]The Chinese University of Hong Kong; [3]Department of Electrical Engineering, Burapha University, Chonburi, Thailand;
{jiaweihu, limsk}@gatech.edu; zhuangzhen1995@gmail.com; tyho@cse.cuhk.edu.hk; pruek.va@eng.buu.ac.th;

***Abstract*—Native 3D Integrated Circuit (3D IC) design offers enhanced performance and density but faces challenges in signal routing due to limited true 3D EDA tool support. Pseudo-3D flows bridge this gap but lack cross-tier optimization, critical for both mixed-node and homogeneous designs. Metal Layer Sharing (MLS) addresses this by enabling cross-tier routing co-optimization but risks timing degradation if not applied strategically. Additionally, MLS creates open connections in hybrid-bonded 3D ICs, making chips untestable. We propose GNN-MLS, a Graph Neural Network-based framework for precise MLS net selection, combined with a tailored DFT solution for robust testability. Experiments show GNN-MLS reduces timing violations by 79% and improves WNS and TNS by 81% and 94%, and moves designs closer to true 3D ICs.**

I. Introduction

Three-dimensional Integrated Circuits (3D ICs) are a promising solution to extend Moore's Law by providing increased device density and performance through vertical integration. However, current Electronic Design Automation (EDA) tools, primarily designed for 2D ICs, face significant limitations in supporting true 3D integration. While dedicated 3D flows have been proposed for placement [1]–[3], true 3D routing remains underexplored, and these methods often fall short of achieving manufacturable quality. As a result, pseudo-3D flows, such as sequential-2D [4], Memory-on-Logic (MoL) [5], and pin-3D [6], have emerged as practical alternatives, leveraging 2D EDA tools to achieve commercially viable 3D designs. However, these flows fail to exploit fine-grained cross-tier optimization, leaving untapped opportunities for performance improvements [7]–[10].

Metal Layer Sharing (MLS) offers a promising pathway to bridge the gap between pseudo-3D and true 3D capabilities by enabling cross-tier routing co-optimization through 2D EDA tools. MLS facilitates the use of routing layers across tiers, as depicted in Figure 1, unlocking additional paths and enhancing timing performance. However, state-of-the-art (SOTA) approaches to MLS rely on heuristic strategies with limited net-level control, leading to indiscriminate MLS application that can degrade timing performance. For example, in our experiments with the MAERI 128PE design [11], SOTA MLS applied to specific nets worsened slack (e.g., net n146095 had slack degraded to -48ps), whereas disabling MLS improved slack to -45ps as shown in Table I. These findings underscore the need for more targeted MLS strategies. Figure 2 illustrates that while SOTA improves timing with MLS, our proposed method achieves far superior results.

Additionally, MLS introduces testability challenges, particularly in hybrid-bonded 3D ICs. Open connections created during fabrication render critical signals unobservable or uncontrollable, leading to non-testable designs. As illustrated in Figure 3, these open connections significantly affect testability in hybrid-bonded designs. Addressing these issues is crucial for realizing MLS's potential in cross-tier co-optimization and advancing pseudo-3D designs toward true 3D integration.

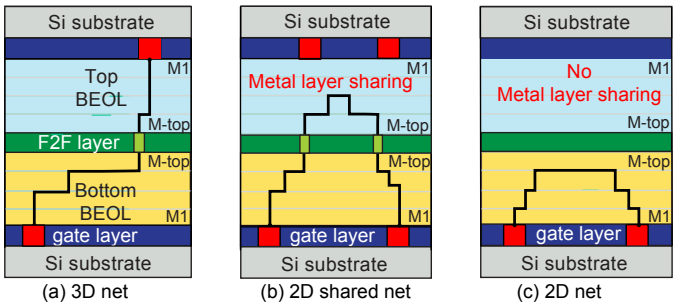


Fig. 1: Illustration of 3-d net, 2d-shared net, and 2d-net.

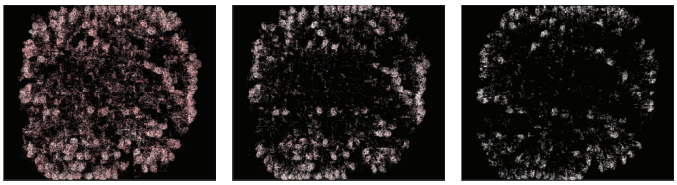
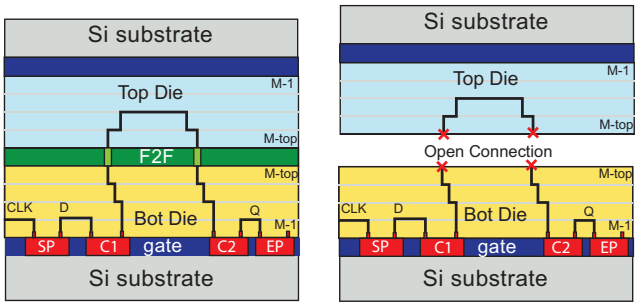


Fig. 2: Timing Violation Points: Registers with violations are highlighted. No MLS has the most violations, SOTA reduces violations by 68%, and GNN-MLS achieves an 80% reduction.

To address these challenges, we propose **GNN-MLS**, a Graph Neural Network-assisted framework for fine-grained, net-level MLS decision-making focused on optimizing timing in 3D ICs. The core challenge lies in evaluating the timing impact of MLS on each net, which traditionally requires exhaustive Static Timing Analysis (STA) across all configurations—a computationally prohibitive process. GNN-MLS eliminates this need by leveraging a novel approach that converts the hyper-edge nature of nets into a node-centric problem, enabling efficient graph-based modeling of timing paths.

Using a transformer-based encoder, GNN-MLS captures the intricate interdependencies among nets along critical timing paths, ensuring precise MLS decisions that prioritize timing improvements. Additionally, we introduce a tailored Design-for-Test (DFT) solution to address testability challenges in hybrid-bonded 3D ICs, ensuring signal observability and controllability across tiers with minimal overhead. By strategically focusing on timing and testability, GNN-MLS advances pseudo-3D flows toward the performance and reliability of true 3D ICs, offering a scalable and practical solution for next-generation integration.

In this work, we propose a comprehensive framework for cross-tier routing co-optimization in 3D ICs, combining GNN-MLS for MLS decision-making, DFT for testability, and PDN design for heterogeneous stacking. Our main contributions are as follows:



(a) Timing Path using MLS (b) Pre-bond Testing

Fig. 3: Testability Issue in Hybrid-Bonded Design with MLS: Each MLS induces an open connection during individual die fabrication, which makes the chip untestable. The timing path consists of start-point (SP), combinational logic (C1-2), and endpoint (EP).

1) We propose GNN-MLS, integrating hyper-graph conversion, a Transformer for net interdependencies, contrastive pretraining, and supervised fine-tuning for MLS decision-making.
2) We design two DFT strategies for MLS in hybrid-bonded 3D ICs, ensuring high test coverage with minimal overhead.
3) We develop a multi-power domain framework and a PDN for mixed-node 3D ICs to support heterogeneous integration experiments.
4) Our framework achieves up to 75% improvement in WNS, 94% in TNS, and a 75% reduction in timing path violations compared to sequential 2D designs.

II. Motivation

A. Why MLS and What are its Limitations?

Metal Layer Sharing (MLS) is a critical innovation for routing efficiency in 3D ICs, addressing key challenges such as congestion and timing optimization. By enabling nets to utilize BEOL layers across tiers, MLS provides additional routing resources and unlocks new pathways for critical connections, which are otherwise constrained in 2D ICs. For example, in the MAERI architecture with 16PE, MLS improves critical path slack from -76 ps without MLS to -18 ps with selective MLS, demonstrating its potential to enhance timing and overall performance.

However, current EDA tools do not model or systematically implement MLS. The only systematic approach to applying MLS is through iterative STA, which evaluates the timing impact of MLS on a net-by-net basis. This process involves repeated disconnection, rerouting, and slack recalculation, making it computationally infeasible for large designs. Furthermore, MLS introduces severe testability challenges in hybrid-bonded 3D ICs, as inter-tier connections disrupt signal propagation and render dies non-testable. These limitations restrict the practical adoption of MLS despite its clear performance benefits.

To address these issues, we propose a comprehensive framework that combines machine learning-based MLS decision-making to eliminate exhaustive STA evaluations with tailored DFT solutions to ensure robust testability. This approach significantly improves the efficiency and practicality of MLS in modern 3D ICs.

B. Graph Neural Networks for Metal Layer Sharing

Metal Layer Sharing (MLS) decisions are essential for optimizing timing in 3D ICs, but systematically applying MLS presents significant challenges. Slack, a key timing metric, depends on the cumulative delays of all nets in a timing path, where changes to one net propagate through others, creating intricate interdependencies. The only systematic approach for evaluating MLS decisions is through Static Timing Analysis (STA), which requires exhaustively exploring the entire MLS configuration space. This process involves iterative disconnection, rerouting, and recalculating slack for each net, making it computationally prohibitive for large designs and delaying cross-tier routing optimization.

TABLE I: Metal layer sharing (MLS) impact on slack with single net of MAERI 128PE using the heterogeneous 3D integration ($16nm$ logic, $28nm$ memory).

Net name	before MLS		after MLS	
	slack (ps)	used metals	slack (ps)	used metals
n480132	-62	M1-6(bot)	-45	M1-6(bot)+M5-6(top)
n146095	-45	M1-4(bot)	-48	M1-6(bot)+M6(top)

To address the challenges of modeling timing paths, we leverage Graph Neural Networks (GNNs) to represent timing paths as directed acyclic graphs (DAGs), where nodes correspond to devices and edges to nets. GNNs provide a natural framework for propagating information across the graph, capturing local dependencies and structural relationships within the path. However, timing paths exhibit complex interdependencies that extend beyond immediate neighbors, requiring the ability to model both local interactions and long-range dependencies effectively. To tackle this, we integrate the Transformer [12] architecture into the GNN framework. Transformers, with their self-attention mechanism, dynamically weigh interactions across nodes, enabling them to model long-range relationships along the path. This ability to capture both sequential and global interactions complements the GNN's structural modeling, making the combined framework particularly powerful for understanding the cascading impact of MLS decisions and optimizing timing paths.

A significant challenge in implementing MLS is the scarcity of labeled data, as generating MLS labels through STA is computationally expensive and impractical at scale. To overcome this, we employ self-supervised learning with Deep Graph InfoMax (DGI) [13] to pretrain the Transformer. DGI maximizes mutual information between global path representations and local node embeddings, enabling the extraction of meaningful features from unlabeled timing path data. These pre-trained embeddings are then fine-tuned using a limited labeled dataset, which includes 500 timing paths each from the heterogeneous A7 single-core and MAERI 128PE designs, as well as 500 timing paths from the corresponding homogeneous configurations. This approach significantly reduces reliance on STA while ensuring efficient and accurate MLS predictions.

By combining the sequential modeling capabilities of Transformers, the graph-structured representation of GNNs, and the data efficiency of DGI, our framework addresses the challenges of interdependencies, computational overhead, and data scarcity. This enables early, efficient, and accurate MLS decision-making, advancing cross-tier routing optimization in 3D IC designs.

III. GNN-MLS Algorithms

Cross-tier co-optimization in 3D ICs with Metal Layer Sharing (MLS) poses significant computational challenges. Evaluating the timing impact of MLS relies on post-routing Static Timing Analysis (STA), which requires iterating through multiple MLS configurations. This process is infeasible for large designs due to its high computational cost. To address this, we propose GNN-MLS, a machine learning-based framework that predicts MLS decisions early in the design flow. By moving MLS decisions to the routing stage, GNN-

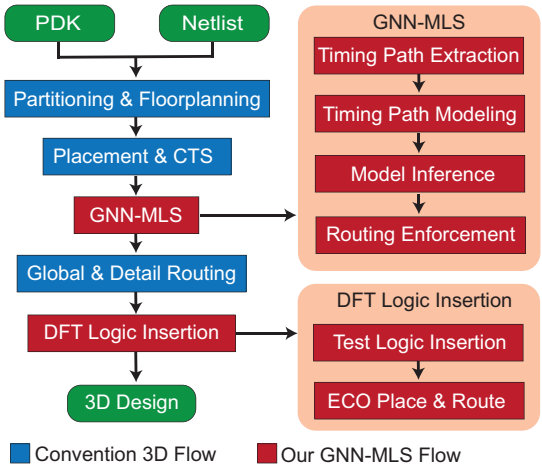


Fig. 4: GNN-MLS Design Flow

MLS reduces dependency on STA and enables efficient cross-tier routing optimization.

A. Design Flow

Our design flow integrates GNN-MLS to tackle timing optimization for monolithic and hybrid-bonded 3D ICs, while also ensuring testability in hybrid-bonded designs. Without MLS-specific tools, current methods rely entirely on post-routing STA to determine MLS decisions, restricting EDA tools from leveraging MLS during earlier stages. This late-stage dependency results in limited optimization opportunities and increased computational overhead. By enabling MLS predictions during routing, GNN-MLS integrates seamlessly into the design flow, facilitating timing optimization and ensuring testability at an earlier stage.

Figure 4 illustrates the key stages of the flow:

1) MLS Decision-Making with GNN-MLS: The GNN-MLS module generates binary MLS decisions for individual nets, indicating whether MLS should be applied. These decisions identify whether MLS should be applied based on a learned node representations. By providing these decisions early, GNN-MLS enables efficient routing optimization, allowing EDA tools to consider MLS during cross-tier routing.
2) Targeted Routing and DFT Integration: Using the MLS decisions from GNN-MLS, targeted routing enforces selective metal layer sharing, ensuring that only specific nets share metal layers across tiers to maximize timing performance. Following routing, custom Design-for-Test (DFT) logic is integrated to ensure testability. This addresses challenges unique to MLS-enabled designs, ensuring that cross-tier connections remain verifiable.

This design flow reduces the computational cost of iterative STA while achieving practical cross-tier co-optimization and addressing testability challenges.

B. Timing Path Modeling

Evaluating the timing impact of MLS on each net before routing is essential to optimize timing performance and eliminate iterative steps in the design flow. This fine-grained decision problem can be formulated as:

***Problem** 1 (Net-Level MLS Decision):* Given a netlist $\mathcal{N}$ and a timing path set $\mathcal{P}$, each net in a timing path should decide whether it uses MLS to maximize the slack of its timing path.

TABLE II: Hand-crafted Features used in our GNN-MLS.

Features	Description	Unit
cell location (x,y)	location of the cells	μm
cell delay	delay of the cells	ps
pin capacitance	pin capacitance of the output pin	pF
wirelength	early-global routing wirelength of the net	μm
wire capacitance	estimated wire capacitance of the net	pF
wire resistance	estimated wire resistance of the net	Ω

The optimal slack for each path p_i is given by:

$$\text{slack}_{\text{opt}}(p_i) = \text{slack}_{\text{2D}}(p_i) + f(\delta(n_1), \delta(n_2), \ldots, \delta(n_{|p_i|})), \quad (1)$$

where $slack_{2D}(p_i)$ represents the slack of p_i in the sequential-2D design (no-MLS). $\delta(n_i)$ is a binary variable to identify if the net $n_i \in \mathcal{N}$ uses MLS. $f(\cdot)$ represents the slack variation caused by the MLS decision compared with the sequential-2D design. The optimization objective then becomes:

$$\max_{\delta(n_i)\in\{0,1\}} \sum_{p_i \in \mathcal{P}} \text{slack}_{\text{opt}}(p_i) \quad (2)$$

However, solving this optimization directly is infeasible due to the exponential number of MLS configurations and the need for STA evaluations to compute $f(\cdot)$. Instead, GNN-MLS approximates this solution by learning to predict the optimal MLS decisions $\delta(n_i)$ based on timing path dependencies. By framing the problem in this manner, GNN-MLS allows efficient decision-making while indirectly optimizing the total slack objective.

We propose a timing path modeling methodology tailored to the unique characteristics of timing paths, as shown in Figure 5. Circuits are first modeled as directed acyclic graphs (DAGs) based on signal transmission. However, multi-pin nets, which connect three or more pins and interconnect multiple devices, introduce significant challenges for MLS decision-making due to their complexity.

To simplify this, each multi-pin net is treated as a hyperedge with a single source node, corresponding to the output pin in the netlist. This transformation allows net-related MLS decisions to be reformulated as node-specific decisions, effectively combining edge features with node features. Graph Transformer architectures then process these features, transforming the initial node attributes into informative embeddings for accurate MLS predictions. The fused features are shown in Table II.

C. Graph Transformer-Based MLS Decisions

The Transformer architecture plays a central role in GNN-MLS, as it effectively captures the interdependencies among MLS decisions across nets within a timing path. The slack improvement for a timing path depends on how MLS is applied to all nets along the path. This interdependence, combined with the sequential nature of timing paths, necessitates a model that can capture both local and global relationships. The entire graph Transformer-based framework is shown in Figure 5.

Traditional GNNs, which aggregate information from immediate neighbors, are insufficient for this task because they rely on fixed neighborhood structures. The Transformer's self-attention mechanism dynamically weighs interactions between all nodes in a path, regardless of distance, making it ideal for capturing long-range dependencies. Additionally, the sequential nature of timing paths is preserved through positional encodings, which ensure that the order of nodes is explicitly encoded in the model. To balance expressiveness and efficiency, the proposed Transformer architecture has three layers. Each layer consists of a three-head self-attention

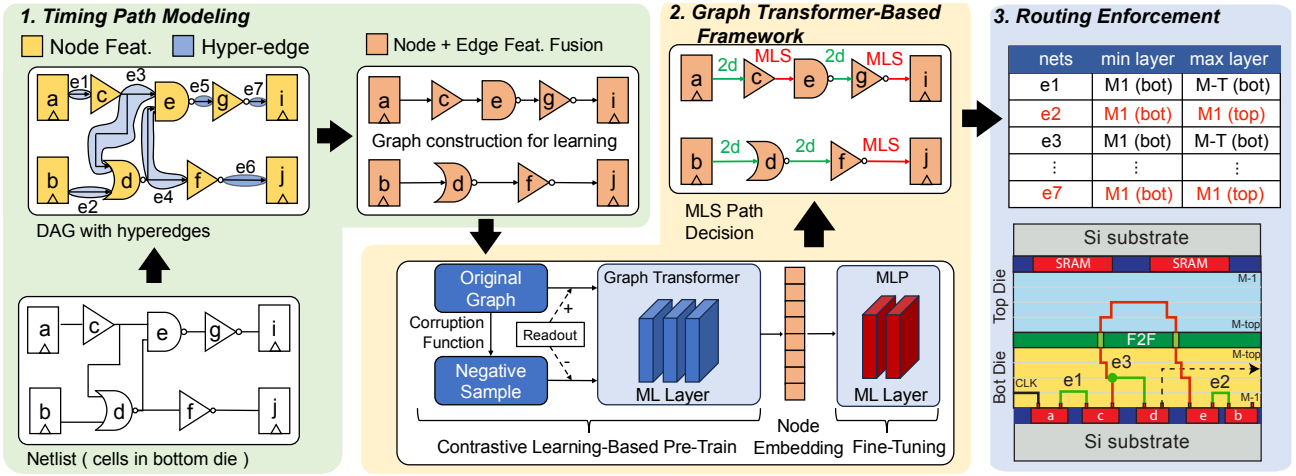


Fig. 5: The GNN-MLS framework converts netlist into hypergraph. It embeds hyperedge features into output-pin-associated nodes and extracts timing paths. The Graph Transformer, pretrained with DGI, processes these paths. MLP fine-tunes the embeddings to make MLS decisions.

mechanism. This architecture is tailored to the complexity of MLS decision-making, addressing the combinatorial dependencies among nets while maintaining computational efficiency.

Obtaining labeled data for MLS decisions, such as STA-based ground truth, is computationally expensive. To address this, we use Deep Graph Infomax (DGI) for self-supervised pretraining, allowing the model to learn effective node embeddings from unlabeled timing path graphs. DGI maximizes the mutual information between global path representations $g(Y)$ and local node embeddings v, ensuring that embeddings capture both structural and timing-critical dependencies. Negative sampling, achieved by perturbing node features, enhances robustness by forcing the model to distinguish true node-path relationships from noise. We adopt sigmoid function σ to map inner product to probability and aid training stability. The DGI loss for a graph Y is defined as:

$$I(Y) = \sum_{v \in Y} \log(\sigma(\langle v, g(Y) \rangle)) + \sum_{v^* \in C(Y)} \log(\sigma(\langle v^*, g(Y) \rangle)) \tag{3}$$

Following DGI pretraining, the Transformer-generated embeddings are fine-tuned using a two-layer Multi-Layer Perceptron with collected labeled data. The MLP maps each node embedding to a binary decision $\delta(n_i)$, indicating whether MLS should be applied. The entire training scheme is shown in Algorithm 1.

Algorithm 1: Training Process for MLS Decision-Making

Input: Set of timing path graphs $\{Y_i\}_{i=1}^{M}$; node features X; binary MLS labels for fine-tuning.
Output: Trained Transformer and MLP.

DGI Pretraining:
for *each timing path graph* Y_i **do**
  Compute node embeddings v using graph Transformer.
  Generate global summary vector $g(Y_i)$.
  Create perturbed graph $C(Y_i)$ for negative samples v^*.
  Compute DGI loss to maximize mutual information.
MLP Fine-Tuning:
for *each labeled timing path graph* Y_i **do**
  Pass DGI-pretrained node embeddings by a 2-layer MLP.
  Train MLP on labels to predict MLS decisions $\delta(n_j)$.

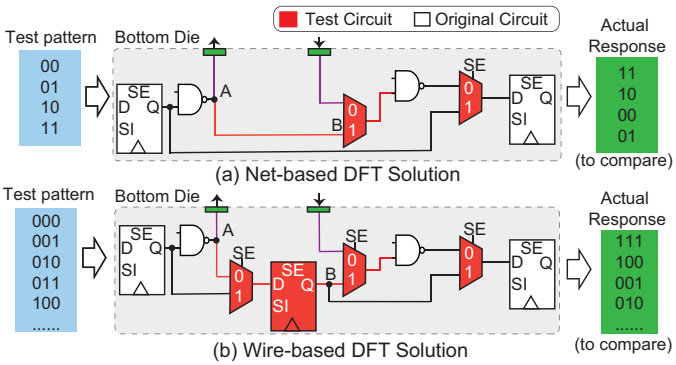


Fig. 6: Two DFT solution for MLS: (a) Net-based DFT where fault can be detected if either outgoing and incoming signal from F2F pads are incorrect, (b) Wire-based DFT where the mechanism can detect if outgoing or incoming F2F pads are fault.

D. DFT Strategy for MLS-Enabled Designs

While Metal Layer Sharing (MLS) significantly improves 3D routing by enabling cross-die connections, it introduces critical testability challenges. Each MLS connection creates an open circuit, disrupting signal propagation and preventing observability and controllability across dies, as shown in Figure 3. This renders upstream signals non-observable and downstream signals non-controllable, making MLS-enabled designs non-testable in hybrid-bonded 3D ICs. Conventional Design-for-Test (DFT) strategies fail to address these challenges, necessitating tailored solutions for MLS testability.

We propose two targeted DFT strategies designed for MLS connections, inserted post-routing to align with their precise locations. The first method, illustrated in Figure 6(a), adds a MUX at the MLS connection, toggling between functional and test modes. In test mode, the scan chain redirects signal flow across the MLS path, enabling observability and controllability. The second method, depicted in Figure 6(b), introduces a scan flip-flop (FF) at the MLS connection, improving test coverage by registering the upstream

DFT method	Total fault	Detected fault	WNS
Net-based DFT	444,296	438,152	-21ps
Wire-based DFT	444,346	438,276	-23ps

TABLE III: Two DFT approaches comparison for MLS nets: (a) net-based MLS DFT, (b) wire-based MLS DFT.

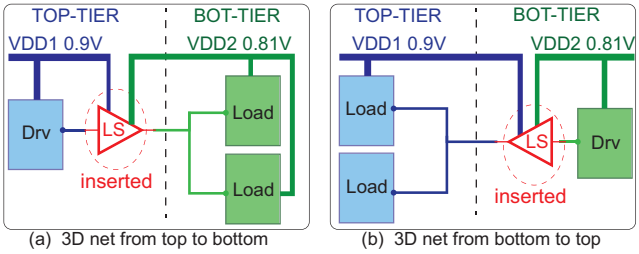


Fig. 7: Power Design for Hetero-Integration.

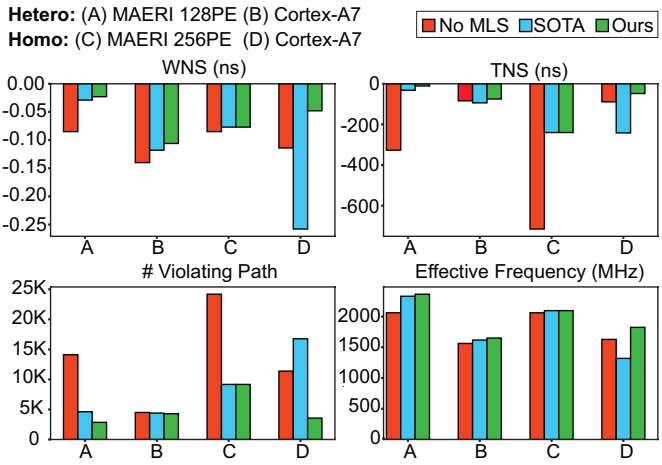


Fig. 8: Timing Metric Comparison on different design benchmark.

signal and driving the downstream signal during testing.

We evaluated these methods on the MAERI 16PE 4BW design with 16 MLS nets. As shown in Table III, while the MUX-based method minimizes additional logic, the FF-based method achieves superior fault coverage, detecting 94 more faults at the cost of 50 additional faults due to added logic. Post-routing ECO adjustments ensure that the timing impact of these solutions remains minimal. These results validate both approaches, with the FF-based method offering enhanced testability for MLS-enabled designs.

E. Power Delivery for Mixed-Node 3D ICs

In mixed-node 3D IC heterogeneous integration setup, we introduce a unique 3D power domain configuration, shown in Figure 7. The top-level operates at 0.9V, with the 28nm memory sub-domain also at 0.9V and the 16nm logic sub-domain at 0.81V. Level shifters are inserted for each 3D signal connection to manage voltage differences across tiers, facilitating reliable communication between the 16nm and 28nm technologies.

The PDN is implemented with specific width and pitch to ensure that the IR-drop of all design are within 10% of lowest VDD (0.81V) as defined in Table IV. The remaining routing resource are utilized for the 2D or MLS nets.

IV. Experimental Results

A. Experimental Setup

We evaluate the GNN-MLS framework on homogeneous and heterogeneous 3D integrations with Face-to-Face (F2F) bonding using three benchmarks: 128PE 32BW MAERI-ARCH, dual-core Cortex A7, and 256PE 64BW MAERI-ARCH. Heterogeneous designs use TSMC $28nm$ for the logic die and $16nm$ for the memory die, while homogeneous designs use $28nm$ for both dies. 3D Place-and-Route (PnR) is performed with Cadence DDI 23.12.000 (Innovus) and the Macro-3D flow [5]. F2F via parameters are configured as size $0.5\mu m$, pitch $1.0\mu m$, resistance 0.5Ω, and capacitance 0.2fF.

We compare GNN-MLS with the state-of-the-art (SOTA) MLS technique [9] and a sequential 2D (no MLS) stacking method, analyzing timing, routing efficiency, and testability. Runtime, based on timing path depth, includes transformer-based training and inference. Results highlight GNN-MLS's advantages in both integration contexts.

B. Results for Heterogeneous 3D ICs

In heterogeneous 3D ICs integrating 16nm logic with 28nm memory, GNN-MLS achieves substantial timing improvements (Figure 8, Table IV). For the MAERI 128PE design, which has a balanced logic-memory ratio, GNN-MLS reduces TNS to -11 ns from -32 ns (SOTA) and -327 ns (No MLS), and decreases violating paths to 2.8K from 4.6K (SOTA) and 14K (No MLS). WNS improves to -23 ps, outperforming both SOTA (-29 ps) and No MLS (-85 ps).

Similarly, in the A7 Dual-Core configuration, GNN-MLS reduces TNS to -75 ns versus -94 ns (SOTA) and -84 ns (No MLS). Violating paths decline to 4.2K from 4.4K (SOTA) and 4.5K (No MLS), with WNS improving to -106 ps compared to -118 ps (SOTA) and -140 ps (No MLS).

Notably, GNN-MLS applies MLS more selectively in heterogeneous designs. For MAERI 128PE, MLS usage drops to 2.4K nets from 9.5K (SOTA); for A7 Dual-Core, usage decreases to 2.6K nets from 3.5K (SOTA). This selective approach demonstrates GNN-MLS's adaptability, applying MLS strategically to maximize timing benefits and minimize overhead.

These results highlight GNN-MLS's ability to optimize timing in heterogeneous designs through targeted and efficient MLS application, meeting the specific demands of hybrid-bonded 3D ICs.

C. Results for Homogeneous 3D ICs

In homogeneous 3D ICs stacking 28nm memory on 28nm logic, GNN-MLS significantly improves timing (Figure 8, Table V). For MAERI 256PE, selected for balanced logic-memory ratio, GNN-MLS reduces TNS to -240 ns from -715 ns (SOTA) and -513 ns (No MLS). Violating paths drop to 9,173 from 24,195 (SOTA) and 16,037 (No MLS), and WNS improves to -0.077 ns, surpassing SOTA (-0.085 ns) and No MLS (-0.083 ns).

In the A7 Dual-Core design, GNN-MLS lowers TNS to -48 ns compared to -242 ns (SOTA) and -89 ns (No MLS). Violating paths decrease to 3,569 from 16,770 (SOTA) and 11,391 (No MLS), with WNS significantly improving to -0.048 ns versus -0.258 ns (SOTA) and -0.114 ns (No MLS).

GNN-MLS also expands MLS usage strategically in homogeneous designs. For MAERI 256PE, MLS coverage increases to 1,600 nets from 870 (SOTA); in A7 Dual-Core, it rises sharply to 73K nets from 8.4K (SOTA). This adaptive approach highlights GNN-MLS's ability to balance extensive MLS coverage with optimal timing.

TABLE IV: PPA metrics comparison between GNN-MLS, State-of-the-Art [9], and SOTA without MLS in Heterogeneous Integration. The frequency is in MHz, power is in mW, run time is in minutes. M-T represents the top-most layer of the memory die where W, P, and U denote the width, pitch, and utilization of the PDN on this metal layer. The highlighted entry denotes the best value among all design options.

	16nm Logic + 28nm Memory					
	MAERI 128PE			A7 Dual-Core		
	No MLS	SOTA	Ours	No MLS	SOTA	Ours
BEOL	6+6			8+8		
Target Freq.	2,500			2,000		
FP (mm^2)	0.38			1.11		
WL (m)	5.23	5.18	5.16	7.60	8.30	8.10
WNS (ps)	-85	-29	-23	-140	-118	-106
TNS (ns)	-327	-32	-11	-84	-94	-75
#Vio. Paths	14,000	4,600	2,800	4,500	4,400	4,200
#MLS Nets	0	9,500	2,370	0	3,542	2,621
Run-Time	-	-	20	-	-	15
Pwr	1,472	1,404	1,389	1,008	1,061	1,052
IR-drop (%)	10.00	9.50	9.40	1.90	2.00	1.98
M-T:W/P/U	2.00μm / 7μm / 14%			2.70μm / 9μm / 30%		
L.S Pwr	40	45	46	31	32	33
Eff. Freq.	2,061	2,330	2,363	1,562	1,618	1,650

TABLE V: PPA metrics comparison between GNN-MLS, State-of-the-Art [9], and SOTA without MLS in Homogeneous Integration. The frequency is in MHz, power is in mW, run time is in minutes. The highlighted entry denotes the best value among all design options.

	28nm Logic + 28nm Memory					
	MAERI 256PE			A7 Dual-Core		
	No MLS	SOTA	Ours	No MLS	SOTA	Ours
Target Freq.	2500			2000		
FP(mm^2)	1.42			1.11		
WL.(m)	14.5	14.6	15.5	14.5	12.1	11.2
WNS (ps)	-83	-85	-77	-114	-258	-48
TNS (ns)	-513	-715	-240	-89	-242	-48
#Vio. Paths	16K	24K	9K	11K	16K	3.5K
#MLS Nets	0	870	1.6K	0	8.4K	73K
Run-Time	-	-	35	-	-	15
Pwr.(mW)	4680	4747	4804	1425	1412	1442
Eff. Freq.	2,070	2,061	2,096	1,628	1,319	1,824

These results confirm GNN-MLS's effectiveness in enhancing timing through targeted MLS application in homogeneous designs.

D. Design-for-Test Results

Hybrid-bonded 3D ICs face testability issues from open MLS connections, affecting signal observability. We introduce a Design-for-Test (DFT) solution using scan flip-flops at critical MLS points with power gating for minimal overhead. DFT is applied to No MLS and GNN-MLS designs (excluding SOTA due to impractical probe pad needs).

As shown in Table VI, the combined GNN-MLS and DFT framework delivers robust testability and timing improvements. MAERI 128PE achieves 98.38% coverage, 75% fewer violating paths and WNS, 94% lower TNS, and a 15% frequency increase. A7 Dual-Core maintains 97.49% coverage, reduces violating paths by 17%, TNS by 10%, WNS by 5%, and frequency increases by 4.3%.

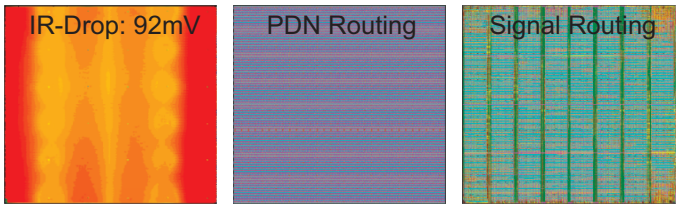
These results confirm the framework effectively addresses MLS testability and enhances timing in hybrid-bonded 3D ICs.

E. Power Delivery Network Design Results

We integrate the PDN into the heterogeneous 3D design, summarized in Table IV, carefully selecting PDN width and spacing to maintain IR-drop within 10% of V_{DD}. Specifically, MAERI 128PE reaches a 10% IR-drop due to its higher power consumption, whereas the A7 Dual-Core achieves a lower IR-drop of 2%, as shown in Figure 9(a). Figures 9(b)–(c) illustrate the sharing of top metal layer resources between PDN and Metal Layer Sharing (MLS) nets, demonstrating effective balancing of power delivery and signal routing. Level shifter (LS) power overhead remains comparable across designs, slightly increasing in SOTA and GNN-MLS due to additional MLS nets, resulting in a total power overhead of about 2–3%.

TABLE VI: Comparison of MLS in Testable Designs between GNN-MLS and State-of-the-Art (28nm / 16nm Technology Node). The frequency is in MHz, power is in mW, run time is in minutes. The highlighted entry denotes the best value among all design options.

	MAERI 128PE		A7 Dual-Core	
	No MLS	GNN-MLS	No MLS	GNN-MLS
Target Freq.	2500	2500	2000	2000
FP. (mm^2)	0.38	0.38	1.11	1.11
WL. (m)	5.95	5.93 (0.3%)	9.40	9.30 (1.1%)
Test Cover.(%)	98.25	98.38 (0.1%)	97.32	97.49 (0.2%)
WNS (ps)	-86	-21 (75%)	-159	-132 (17%)
TNS (ns)	-358	-20 (94%)	-112	-76 (32%)
#Vio. Paths	15,321	3,766 (75%)	6,055	5,267 (13%)
#MLS Nets	0	2,425	0	2,536
Run-Time	-	20	-	15
Pwr	1,539	1,523 (1.0%)	1,157	1,152 (0.4%)
Eff. Freq.	2,062	2,375 (15.2%)	1,517	1,582 (4.3%)



(a) (b) (c)

Fig. 9: (a)Hetero-MAERI 128PE IR-drop of 92mV. (b)PDN routing Layout. (c)Signal routing Layout.

V. Conclusion

This work presents GNN-MLS, a data-driven framework that enables efficient and effective cross-tier co-optimization through Metal Layer Sharing (MLS), significantly improving timing performance in 3D IC designs. Complementing this, our tailored Design-for-Test (DFT) solution ensures the testability of MLS connections, addressing a critical challenge in hybrid-bonded 3D ICs. Additionally, our power design supports heterogeneous integration with mixed-node configurations, facilitating robust experimentation in diverse 3D IC setups. Together, these contributions advance pseudo-3D flows toward true 3D IC performance by bridging the gap between 2D-based EDA tools and the demands of native 3D designs. This framework lays the groundwork for manufacturable 3D methodologies, marking a significant step toward fully optimized, high-density 3D ICs for commercial adoption.

VI. Acknowledgement

This work was supported by the Semiconductor Research Corporation under the JUMP 2.0 Center Program (CHIMES), Samsung Advanced Institute of Technology (SAIT) under the AI for Semiconductors Program, and the Ministry of Trade, Industry & Energy of South Korea (1415187652, RS-2023-00234159).

REFERENCES

[1] Peiyu Liao, Yuxuan Zhao, Dawei Guo, Yibo Lin, and Bei Yu. Analytical die-to-die 3-d placement with bistratal wirelength model and gpu acceleration. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 43(6):1624–1637, 2024.

[2] Xueyan Zhao, Shijian Chen, Yihang Qiu, Jiangkao Li, Zhipeng Huang, Biwei Xie, Xingquan Li, and Yungang Bao. ipl-3d: A novel bilevel programming model for die-to-die placement. In *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, pages 1–9, 2023.

[3] Jai-Ming Lin, Yu-Chien Lin, Hsuan Kung, and Wei-Yuan Lin. Hyplace-3d: A hybrid placement approach for 3d ics using space transformation technique. In *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, pages 1–8, 2023.

[4] Nesara Eranna Bethur, Anthony Agnesina, Moritz Brunion, Alberto Garcia-Ortiz, Francky Catthoor, Dragomir Milojevic, Manu Komalan, Matheus Cavalcante, Samuel Riedel, Luca Benini, and Sung Kyu Lim. Hier-3d: A methodology for physical hierarchy exploration of 3-d ics. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 43(7):1957–1970, 2024.

[5] Anthony Agnesina, Moritz Brunion, Jinwoo Kim, Alberto Garcia-Ortiz, Dragomir Milojevic, Francky Catthoor, Manu Perumkunnil, and Sung Kyu 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)*, pages 1–6, 2021.

[6] Sai Surya Kiran Pentapati, Kyungwook Chang, Vassilios Gerousis, Rwik Sengupta, and Sung Kyu Lim. Pin-3d: A physical synthesis and post-layout optimization flow for heterogeneous monolithic 3d ics. In *2020 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, pages 1–9, 2020.

[7] T. Kim et al. Multi-Stack Wafer Bonding Demonstration utilizing Cu to Cu Hybrid Bonding and TSV enabling Diverse 3D Integration. In *2021 IEEE 71st Electronic Components and Technology Conference (ECTC)*, pages 415–419, 2021.

[8] Christopher Netzband, Kevin Ryan, Yuji Mimura, Son Ilseok, Hirokazu Aizawa, Nathan Ip, Xuemei Chen, Hideyuki Fukushima, and Shinichi Tan. 0.5 μm pitch next generation hybrid bonding with high alignment accuracy for 3d integration. In *2023 IEEE 73rd Electronic Components and Technology Conference (ECTC)*, pages 1100–1104, 2023.

[9] Sai Pentapati and Sung Kyu Lim. Metal layer sharing: A routing optimization technique for monolithic 3d ics. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 30(9):1355–1367, 2022.

[10] H. Park et al. Pseudo-3D Physical Design Flow for Monolithic 3D ICs: Comparisons and Enhancements. *ACM Trans. Des. Autom. Electron. Syst.*, 26(5), jun 2021.

[11] Hyoukjun Kwon, Ananda Samajdar, and Tushar Krishna. MAERI: Enabling flexible dataflow mapping over dnn accelerators via reconfigurable interconnects. *ACM SIGPLAN Notices*, 53(2):461–475, 2018.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[13] Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *arXiv preprint arXiv:1809.10341*, 2018.