International Symposium on Physical Design

IDDA-3D: Inter-Die Delay Aware Timing-Driven Placement on Face-to-Face bonded 3D ICs

Zixian Yang, Shanyi Li,Leilei Jin, Tsung-Yi Ho, Chien-Nan Jimmy Liu

zxy.zixianyang@gmail.com

國立陽明交通大學
NATIONAL YANG MING CHIAO TUNG UNIVERSITY

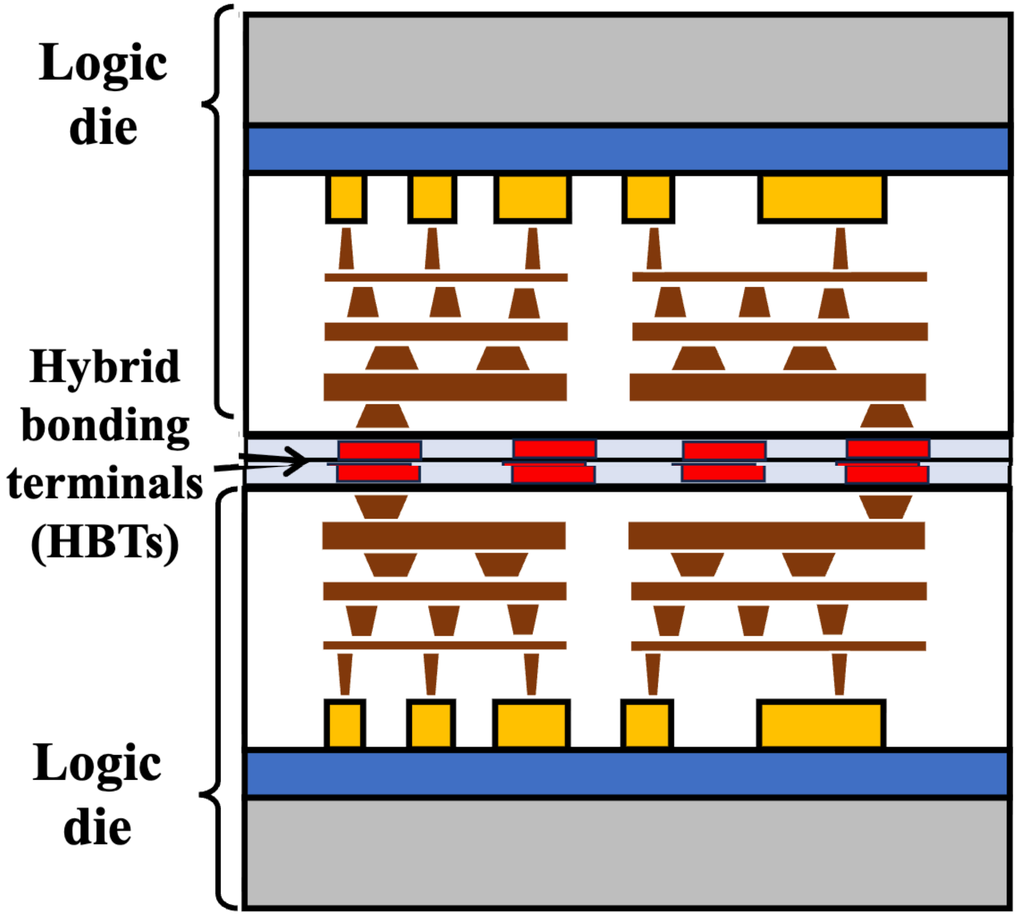
香港中文大學
The Chinese University of Hong Kong

Outline

- Introduction
- IDDA-3D
  - 3D Global Placement Flow
  - Integration of Timing Optimization
- Experimental Result

Face-to-Face bonded 3D IC

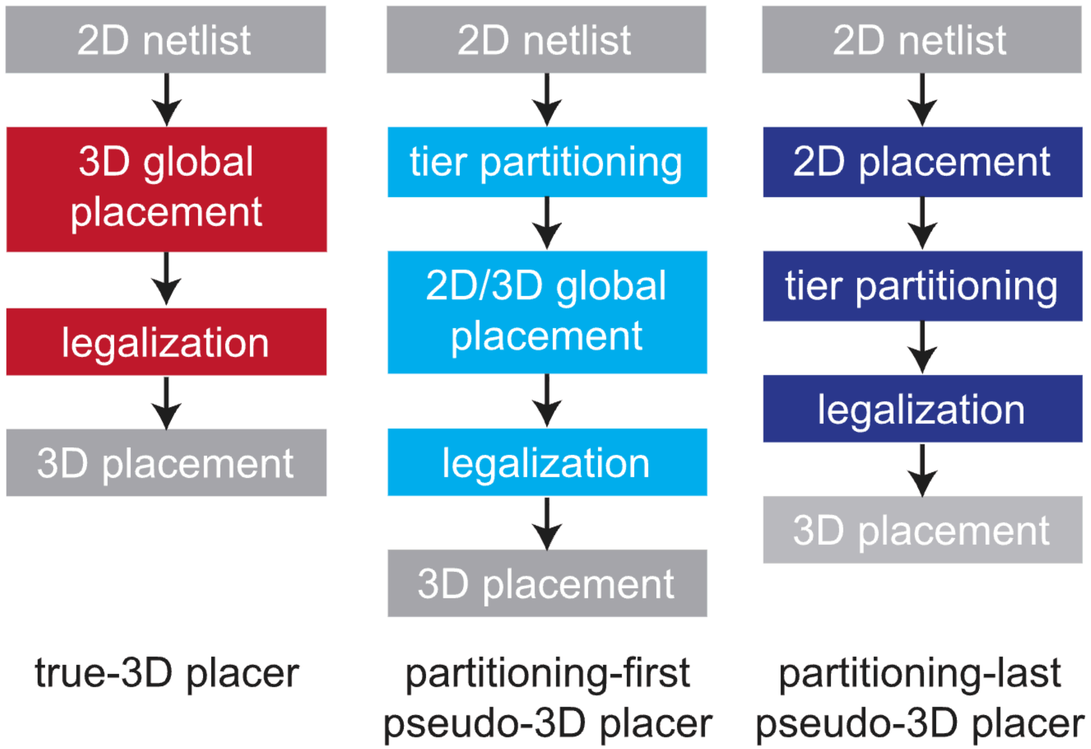
- High integration density
- Manufacturing Feasibility



Shi, Yunqi, et al. "Open3dbench: Open-source benchmark for 3d-ic backend implementation and ppa evaluation." arXiv preprint arXiv:2503.12946 (2025).

Motivation

- Pseudo-3D flows fail to incorporate timing during Placement
- True-3D placers remain wirelength driven

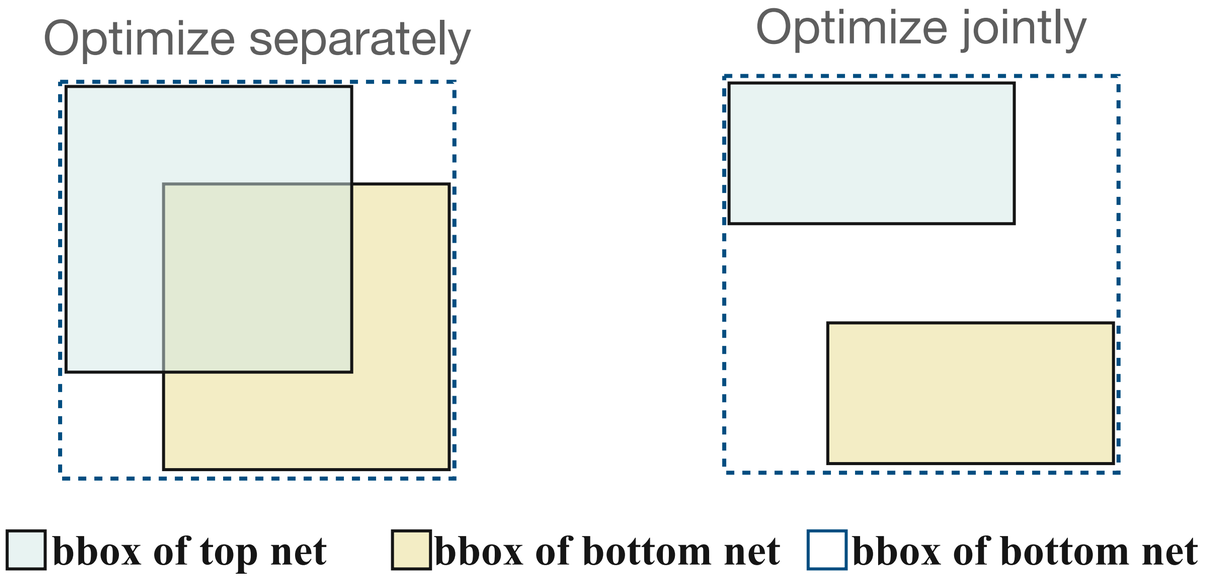


Heechun Park, Bon Woong Ku, Kyungwook Chang, Da Eun Shim, and Sung Kyu Lim. 2021. Pseudo-3D Physical Design Flow for Monolithic 3D ICs: Comparisons and Enhancements. ACM Trans. Des. Autom. Electron. Syst. 26, 5, Article 37 (September 2021), 25 pages

Related works - True 3D placer

- Limitation of Traditional **3D-HPWL** : Fails to capture the actual routing cost when Hybrid Bonding Terminals (HBTs) are inserted.

- The **Bistratal HPWL (BiHPWL)** Approach: Dynamically switches between minimizing the overall net span and optimizing sub-nets on each die independently.



Peiyu Liao, Yuxuan Zhao, Dawei Guo, Yibo Lin, and Bei Yu. 2024. Analytical Die-to-Die 3-D Placement With Bistratal Wirelength Model and GPU Acceleration. Trans. Comp.-Aided Des. Integ. Cir. Sys. 43, 6 (June 2024), 1624–1637.
Jingwei Lu, Hao Zhuang, Ilgweon Kang, Pengwen Chen, and Chung-Kuan Cheng. 2016. EPlace-3D: Electrostatics based Placement for 3D-ICs. In Proceedings of the 2016 on International Symposium on Physical Design (ISPD '16). Association for Computing Machinery, New York, NY, USA, 11–18.

IDDA-3D : Global Placement

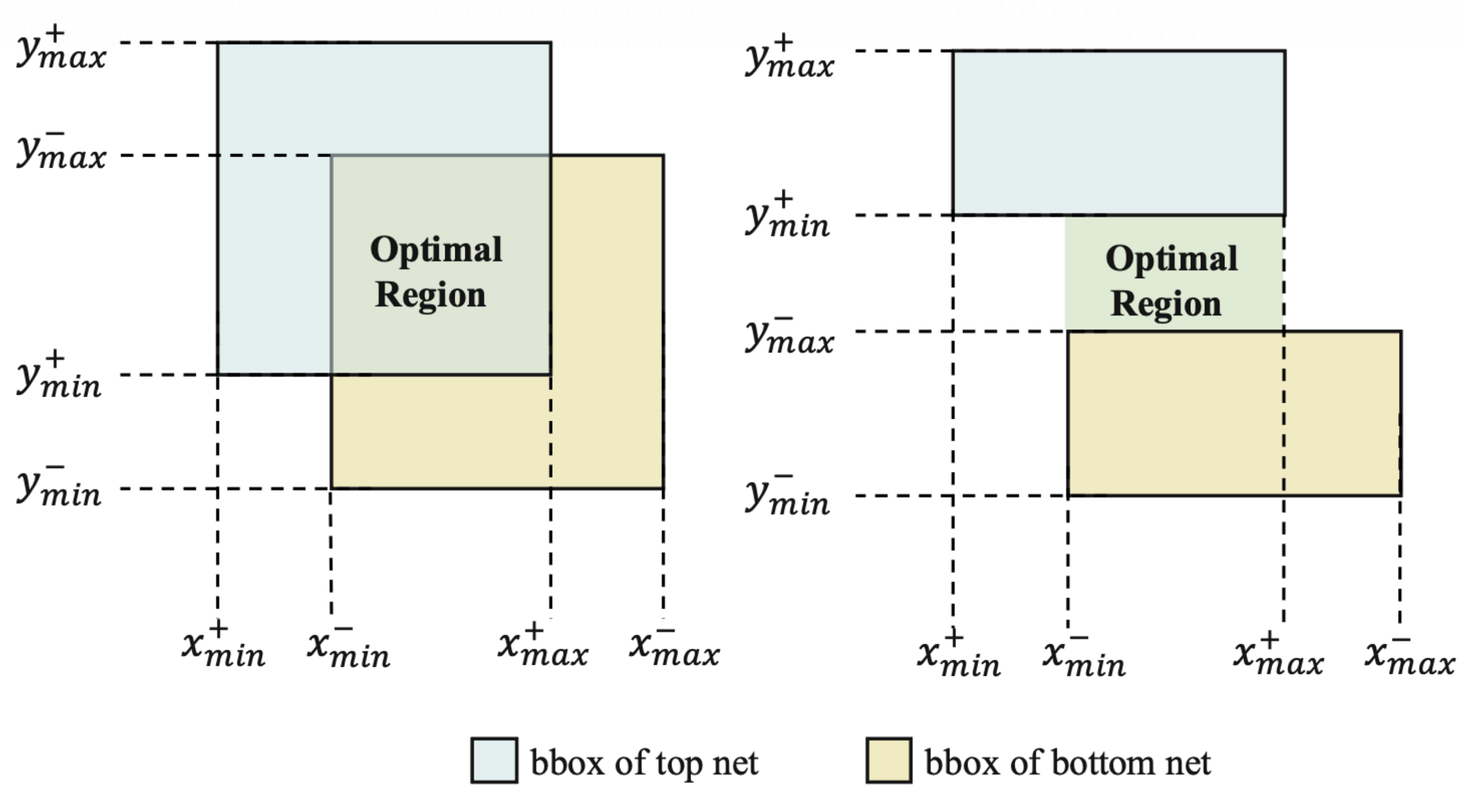
- **D2D WL** : HBT assigned, sum of WL on each die (HBT included)
- Primary design goal is to jointly minimize the D2D WL and the absolute value of TNS

$$\min_{\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{x}_t,\mathbf{y}_t} \left(|\mathrm{TNS}|, \sum_{e \in E} W_{e,\mathrm{D2D}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, x_{t_e}, y_{t_e}) \right).$$

Peiyu Liao, Yuxuan Zhao, Dawei Guo, Yibo Lin, and Bei Yu. 2024. Analytical Die-to-Die 3-D Placement With Bistratal Wirelength Model and GPU Acceleration. Trans. Comp.-Aided Des. Integ. Cir. Sys. 43, 6 (June 2024), 1624–1637

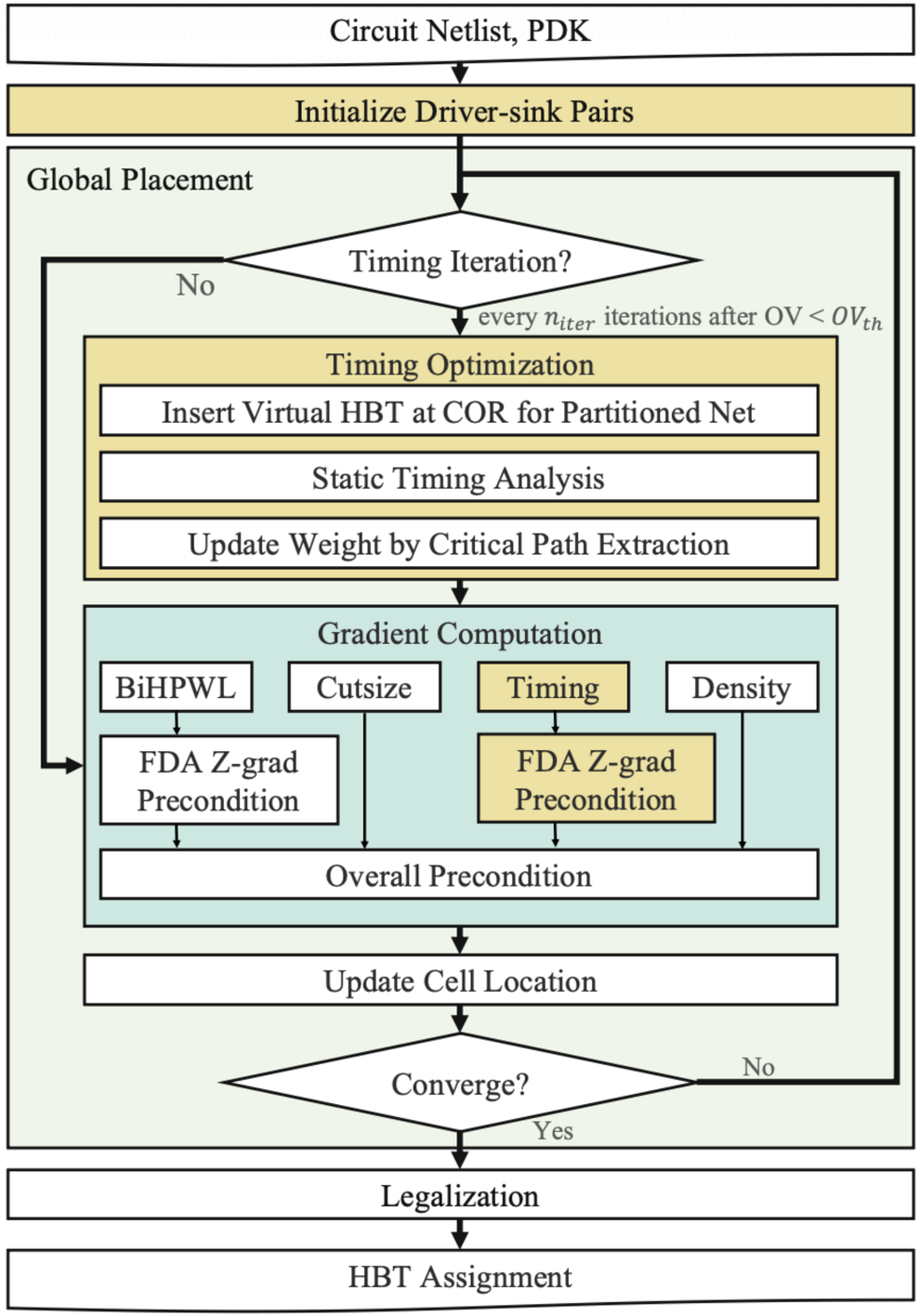
IDDA-3D : Global Placement

- **BiHPWL** : no HBT assigned yet, **surrogate model in GP** to estimate D2D WL without knowing HBT locations



IDDA-3D : Flow

- **Analytical Global Placement:**
  - Jointly optimizes 4 objectives: WL, Density, Cutsize, and Timing.
- **Timing Optimization:**
  - Triggered periodically after density overflow drops below a threshold.
  - Path-based weighting via Static Timing Analysis (STA).



IDDA-3D : Objectives

- **Bistratal HPWL (BiHPWL):**

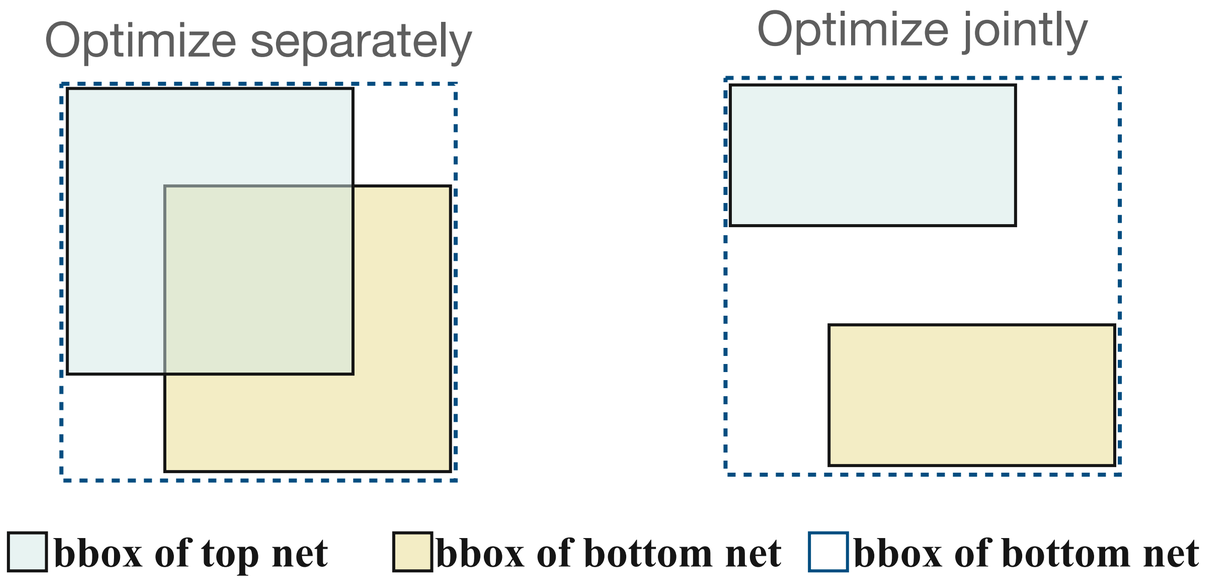
$$p_e(u) = \max_{i \in e} u_i - \min_{i \in e} u_i$$

$$W_{e_x}(\mathbf{x}, \hat{\mathbf{z}}) = \max\left\{p_e(\mathbf{x}),\ p_{e^+}(\mathbf{x}) + p_{e^-}(\mathbf{x})\right\},$$

$$W_{e,\mathrm{Bi}}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = W_{e_x}(\mathbf{x}, \hat{\mathbf{z}}) + W_{e_y}(\mathbf{y}, \hat{\mathbf{z}}).$$

↓ Smooth by Weighted-Average (WA) model

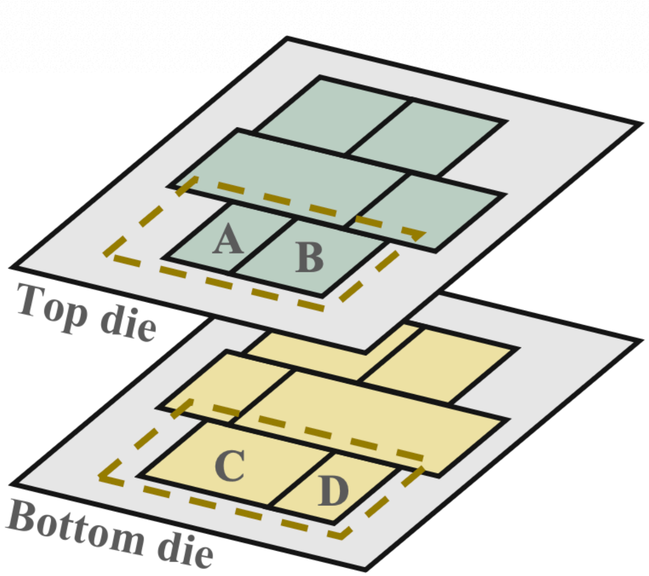
$$p_{e,\mathrm{WA}}(\mathbf{u}) = \frac{\sum_{i \in e} u_i \exp(u_i/\gamma)}{\sum_{i \in e} \exp(u_i/\gamma)} - \frac{\sum_{i \in e} u_i \exp(-u_i/\gamma)}{\sum_{i \in e} \exp(-u_i/\gamma)}$$

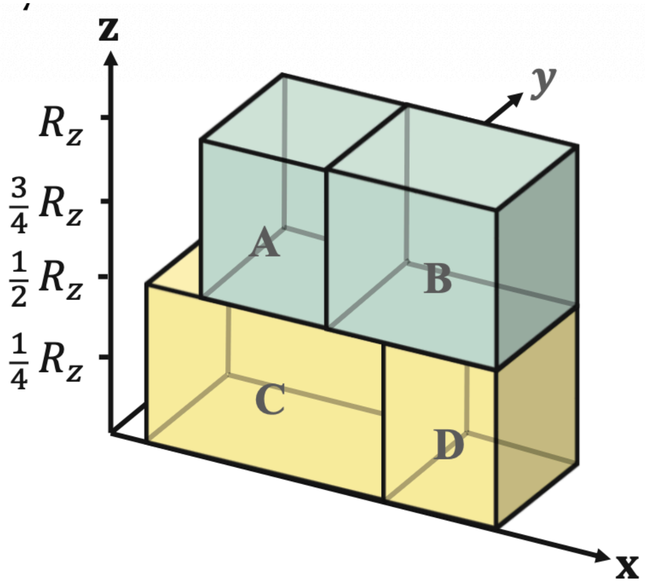


IDDA-3D : Objectives

- **3D density**
  - Placement density overflow as electrostatic potential energy.
  - The volume of an instance is the value of the charge.

$$U(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \sum_{i \in V} q_i \, \Phi_i(x_i, y_i, z_i)$$





Jingwei Lu, Hao Zhuang, Ilgweon Kang, Pengwen Chen, and Chung-Kuan Cheng. 2016. EPlace-3D: Electrostatics based Placement for 3D-ICs. In Proceedings of the 2016 on International Symposium on Physical Design (ISPD '16). Association for Computing Machinery, New York, NY, USA, 11–18.

IDDA-3D : Objectives

- **Cutsize**
  - HBTs are limited resource
  - Z-axis peak-to-peak function **balances WL and HBT usage**

$$p_e(u) = \max_{i \in e} u_i - \min_{i \in e} u_i$$

IDDA-3D : Objectives

- **Timing Cost**
  - Must be differentiable
  - Consider both die information
  - Limitations of previous timing models :
    - **net-weighting** : indirect optimization of driver-sink connections
    - **RSMT** : 3D steiner topology change when z change

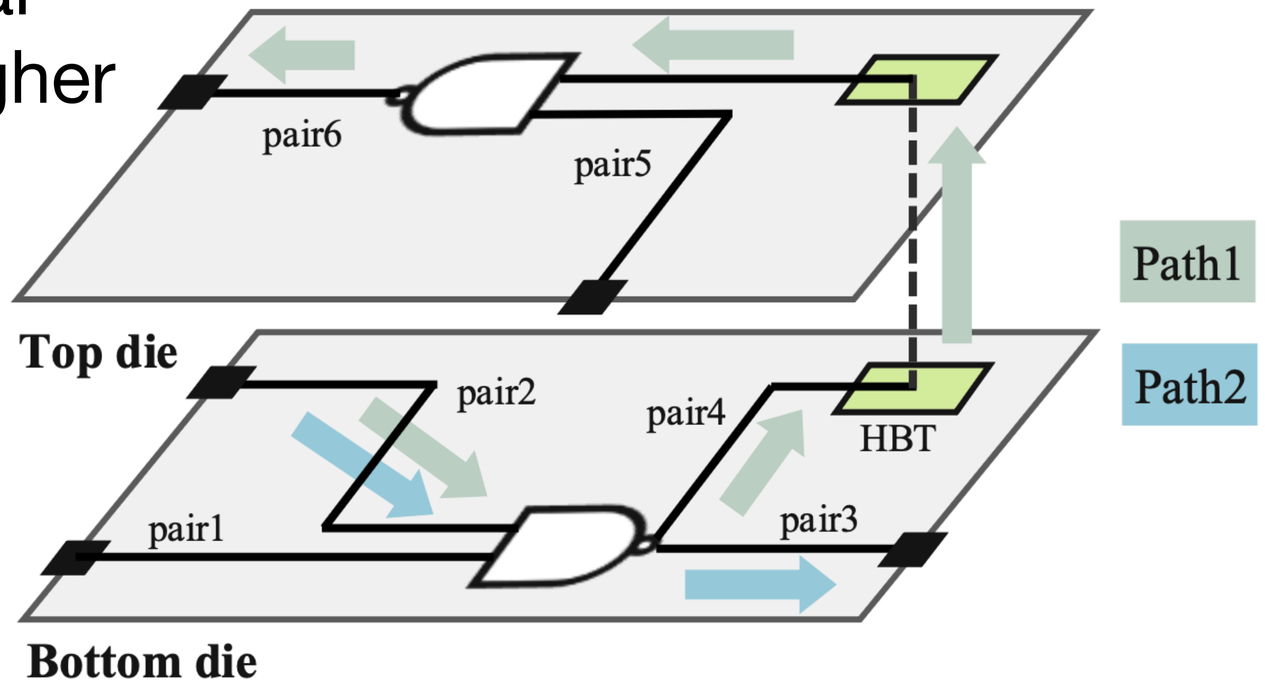
IDDA-3D : Timing

- **Timing Cost = weighted Intra-/Inter-Die Delay**
- **Intra- / Inter-Die Delay Model :**
  - Elmore approximation : RC delay is **quadratic** in the wirelength
  - For driver-sink pair in same die : placer will reduce its intra-die delay
  - $\frac{1}{2}rc\,\|(x_i, y_i) - (x_j, y_j)\|^2$
  - For cross-die driver-sink pair : placer will reduce its inter-die delay
  - $\frac{1}{2}rc\,\|(x_i, y_i) - (x_{t_e}, y_{t_e})\|^2 + \|(x_j, y_j) - (x_{t_e}, y_{t_e})\|^2$

IDDA-3D : Timing

- **Timing Cost = weighted Intra-/Inter-Die Delay**
- **Path-based Weighting :**
  - Pairs shared by multiple critical paths naturally accumulate higher weights from all paths.

$$+ \; w_{\text{cum}} \cdot \frac{\text{slack}(p)}{\text{WNS}}$$



IDDA-3D : Timing

- Timing Cost is not continuous on Z-axis, so we adopt Finite difference approximation (FDA) to obtain the gradient.

$$\left(\widetilde{\nabla}_z C_{\text{tim}}\right)_i = \Delta_{\frac{R_z}{4}} C_{\text{tim}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$$

$$= \frac{C_{\text{tim}}(\mathbf{x}, \mathbf{y}, \mathbf{z}^{(\text{top})}) - C_{\text{tim}}(\mathbf{x}, \mathbf{y}, \mathbf{z}^{(\text{bot})})}{R_z/4}.$$

IDDA-3D : Experimental Results

- Use different node to prove our RC-based delay cost.
- OpenSTA as Evaluator.

Tech	Benchmark	#Cells	#Nets	T_{cycle} (ns)
ASAP7	NV_NVDLA_partition_m	24,324	24,504	0.40
	NV_NVDLA_partition_p	76,352	76,783	0.50
	aes_256	278,265	278,663	0.30
	hidden1	38,090	38,304	0.35
	hidden4	259,511	260,457	0.65
	hidden5	246,797	247,118	0.30
Nangate45	ibex	15,874	18,537	2.80
	jpeg_encoder	60,745	73,102	2.00

IDDA-3D : Experimental Results

- IDDA-3D improves **TNS by 75%** and **WNS by 23%** compared to the 2D.
- IDDA-3D improves **TNS by 44%** and **WNS by 22%** compared to the 3D.

	2D [21]				Analytical-3D [16]					Ours				
Benchmark	WL	TNS	WNS	RT	WL	TNS	WNS	#HBTs	RT	WL	TNS	WNS	#HBTs	RT
NV_NVDLA_partition_m	**60,290**	-17.72	-367	4	60,700	-17.48	**-347**	599	60	61,779	**-16.85**	-351	690	283
NV_NVDLA_partition_p	**285,748**	-293.70	-384	17	319,814	-133.87	**-354**	632	129	333,531	**-97.46**	-412	786	406
aes_256	1,166,061	-66.54	189	61	1,109,336	-36.67	-156	2,325	1128	**1,065,214**	**-35.84**	**-71**	1571	2623
hidden1	129,622	-45.44	-218	6	**121,609**	-35.68	-183	542	77	124,836	**-25.80**	**-144**	533	214
hidden4	**1,198,507**	-792.40	-373	72	1,435,713	-332.88	**-330**	1,331	846	1,402,333	**-175.95**	-383	1363	1326
hidden5	996,534	-41.83	-207	48	905,401	-23.43	-188	1,839	973	**880,923**	**-6.52**	**-64**	1297	1542
ibex	**263,318**	-46.84	-255	3	274,959	-3.02	-60	454	28	275,012	**-1.79**	**-50**	414	78
jpeg_encoder	816,389	-153.40	-340	14	643,738	-63.00	-300	1605	128	**641,722**	**-0.11**	**-30**	1435	215
Ratio	1.03	4.05	1.30	0.03	1.02	1.79	1.27	1.15	0.50	1.00	1.00	1.00	1.00	1.00

IDDA-3D : Ablation Study on FDA Gradients

- **Convergence Failures** on two designs without FDA grad.
- **Timing Degradation :** For converged designs, TNS worsened by 35% and WNS by 18% on average without FDA grad.

	Our w/o C_{tim} FDA z-gradient					Our w/ C_{tim} FDA z-gradient				
Benchmark	WL	TNS	WNS	#HBTs	RT	WL	TNS	WNS	#HBTs	RT
NV_NVDLA_partition_m	63,370	-23.78	-453	446	287	**61,779**	**-16.85**	**-351**	690	283
NV_NVDLA_partition_p	**330,843**	-129.54	**-405**	815	386	333,531	**-97.46**	-412	786	406
aes_256	NA	NA	NA	NA	NA	**1,065,214**	**-35.84**	**-71**	1571	2623
hidden1	**121,609**	**-23.88**	-150	542	189	124,836	-25.80	**-144**	533	214
hidden4	1,433,505	-259.70	-386	1568	1133	**1,402,333**	**-175.95**	**-383**	1363	1326
hidden5	NA	NA	NA	NA	NA	**880,923**	**-6.52**	**-64**	1297	1542
ibex	**273,338**	**-1.53**	**-40**	503	75	275,012	-1.79	-50	414	78
jpeg_encoder	659,478	-48.08	-220	1580	194	**641,722**	**-0.11**	**-30**	1435	215
Ratio	1.02	1.53	1.21	1.04	0.90	1.00	1.00	1.00	1.00	1.00