香港中文大學
The Chinese University of Hong Kong

3D Mixed-Size Placement with GPU Acceleration

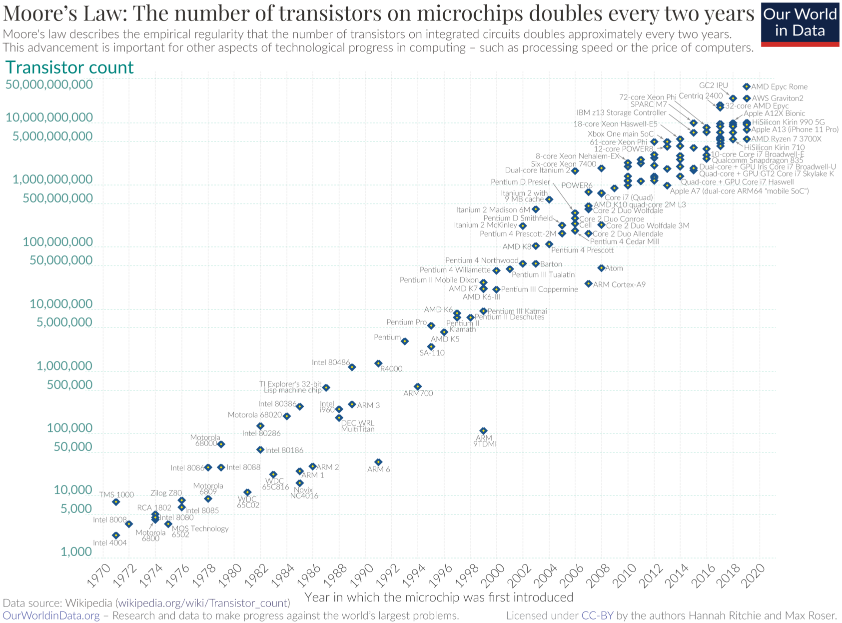
Yuxuan Zhao, Bei Yu
Department of Computer Science & Engineering
Chinese University of Hong Kong
yxzhao21,byu@cse.cuhk.edu.hk

October 23, 2025

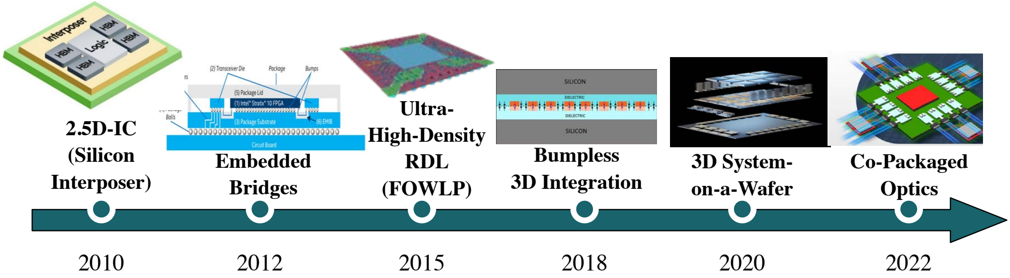
Outline

Introduction

Cramming More Components onto Integrated Circuits

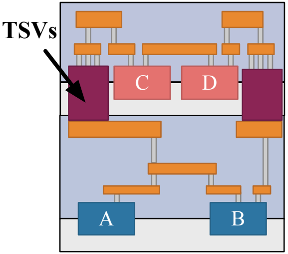


More-than-Moore Heterogeneous Integration



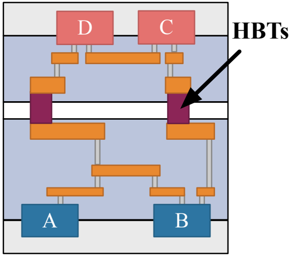
Source: Cadence

- Heterogeneous integration (HI) is a promising solution for better PPA, cost, and memory bandwidth.
- 3D integration is a key technology for HI to achieve Moore style gains.



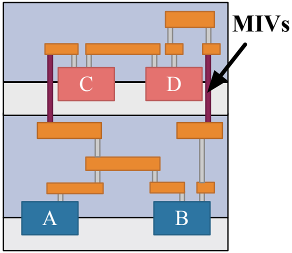
TSV based

- Commercially mature
- High area overhead
- High RC parasitics
- Pitch: > 10 um



Hybrid bonding

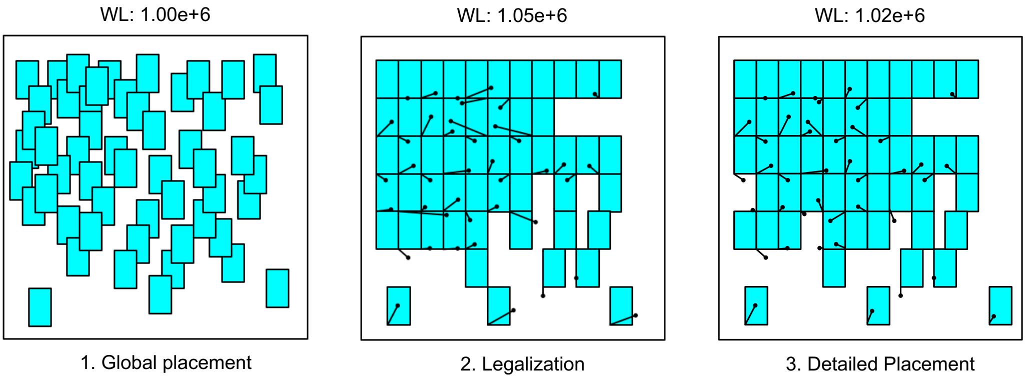
- Low fabrication cost
- No area overhead
- Coupling issues
- Pitch: 1 ~ 10 um



Monolithic

- High fabrication cost
- High integration density
- Small RC parasitics
- Pitch: ~ 0.1 um

Placement



- **Global placement:** Minimize the HPWL wirelength while satisfying the target density constraints.
- **Legalization:** Align cells to row and site without overlap.
- **Detailed placement:** Improve the objective based on the legal solution.

True-3D vs Pseudo-3D Placement

- **Pseudo-3D:** Pseudo-3D placers [TCAD'17][1][ISPD'18][2][ICCAD23-Zhao][3] separate the partitioning and placement phases, and adopt 2D placement tools to determine instance locations.
- **True-3D:** True-3D placers relax the discrete tier partitioning and adopt analytical approaches [TCAD'13][4][TCAD'13][5][ISPD'16][6].

---

[1] **Shreepad Panth et al. (2017). "Shrunk-2-D: A physical design methodology to build commercial-quality monolithic 3-D ICs".** In: *IEEE TCAD* 36.10, pp. 1716–1724.

[2] **Bon Woong Ku, Kyungwook Chang, and Sung Kyu Lim (2018). "Compact-2D: A physical design methodology to build commercial-quality face-to-face-bonded 3D ICs".** In: *Proc. ISPD*, pp. 90–97.

[3] **Xueyan Zhao et al. (2023). "iPL-3D: A Novel Bilevel Programming Model for Die-to-Die Placement".** In: *Proc. ICCAD*.

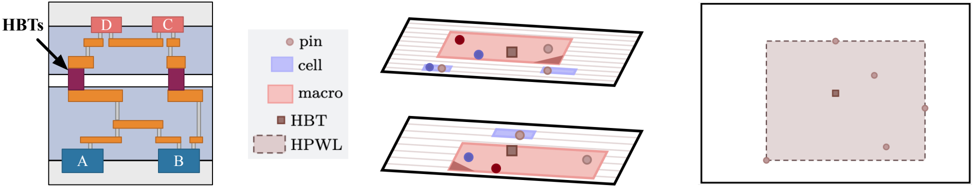
[4] **Guojie Luo, Yiyu Shi, and Jason Cong (2013). "An analytical placement framework for 3-D ICs and its extension on thermal awareness".** In: *IEEE TCAD* 32.4, pp. 510–523.

[5] **Meng-Kai Hsu, Valeriy Balabanov, and Yao-Wen Chang (2013). "TSV-aware analytical placement for 3-D IC designs based on a novel weighted-average wirelength model".** In: *IEEE TCAD* 32.4, pp. 497–509.

[6] **Jingwei Lu, Hao Zhuang, et al. (2016). "ePlace-3D: Electrostatics based placement for 3D-ICs".** In: *Proc. ISPD*, pp. 11–18.

3D Mixed-Size Placement with GPU Acceleration

Illustration of 3D Mixed-Size Placement Problem [ICCAD'23][7]



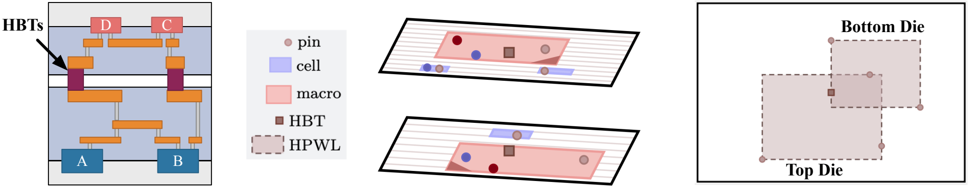
Hybrid bonding   3D View   Projected HPWL

- All cells and macros on both dies should be placed at legal positions. Macros can be rotated with 0°, 90°, 180°, and 270°.
- For any crossing-die net, one hybrid bonding terminal (HBT) is inserted.
- Conventionally, the wirelength objective of a net e is the projected HPWL:

$$W_e(\boldsymbol{x}, \boldsymbol{y}) = p_e(\boldsymbol{x}) + p_e(\boldsymbol{y}), \qquad p_e(\boldsymbol{x}) = \max_{i \in e} x_i - \min_{i \in e} x_i, \tag{1}$$

[7] Kai-Shun Hu et al. (2023). "2023 ICCAD CAD Contest Problem B: 3D placement with macros". In: *Proc. ICCAD*. IEEE, pp. 1–6.

Illustration of 3D Mixed-Size Placement Problem [ICCAD'23][7]



Hybrid bonding　　3D View　　Die-to-Die HPWL

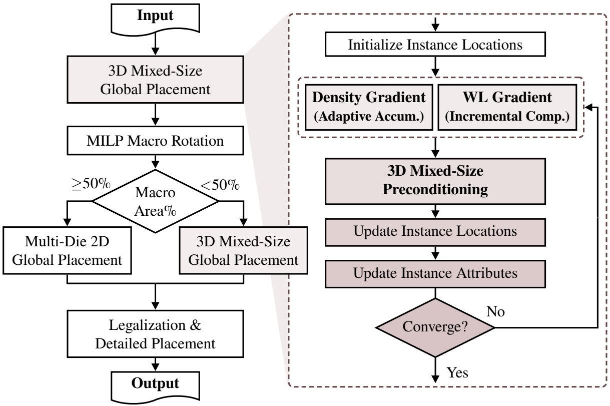
- **Die-to-Die** HPWL: the wirelength of a net e is the sum of the top net e^+ and bottom net e^- including the HBT,

$$W_e = W_{e^+} + W_{e^-}, \quad W_{e^+} = p_{e^+}(\boldsymbol{x}) + p_{e^+}(\boldsymbol{y}), \quad W_{e^-} = p_{e^-}(\boldsymbol{x}) + p_{e^-}(\boldsymbol{y}). \tag{1}$$

- **Heterogeneous** technology integration: instance layouts are different on different dies.

[7] **Kai-Shun Hu et al. (2023). "2023 ICCAD CAD Contest Problem B: 3D placement with macros".** In: *Proc. ICCAD*. IEEE, pp. 1–6.

Our 3D Mixed-Size Placement Framework[TCAD'25][8]



- 3D analytical formulation:

Constrained optimization

$$\min_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}} \ \sum_{e\in E} W_e(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) + \beta \sum_{e\in E} C_e, \quad \text{s.t.} \ D(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \le t_d$$

Nonlinear analytical placement

$$\min \underbrace{\sum_{e\in E} \hat{W}_e(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}_{\text{Wirelength model}} + \lambda \underbrace{\hat{D}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}_{\text{Density model}}$$

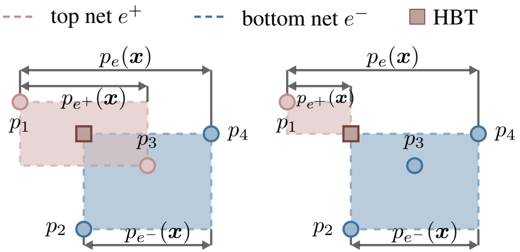
- GPU-accelerated wirelength and density algorithms.

[8] Yuxuan Zhao et al. (2025). "Analytical Heterogeneous Die-to-Die 3D Placement with Macros". In: *IEEE TCAD*.

- Differentiable approximation of HPWL.
- The de facto choice: weighted-average model [TCAD'13],

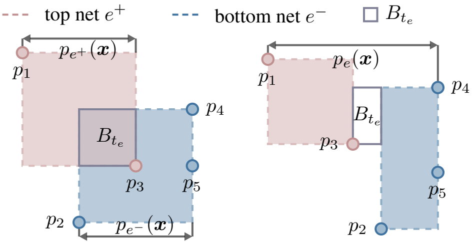
$$W_e(\boldsymbol{x}) \approx \hat{W}_{e,\mathrm{WA}}(\boldsymbol{x}) = \sum_{i=1}^{n} x_i \frac{e^{\frac{x_i}{\gamma}}}{\sum_{i=1}^{n} e^{\frac{x_i}{\gamma}}} - \sum_{i=1}^{n} x_i \frac{e^{-\frac{x_i}{\gamma}}}{\sum_{i=1}^{n} e^{-\frac{x_i}{\gamma}}}. \tag{2}$$

- The cut C_e can also be optimized by $\hat{W}_{e,\mathrm{WA}}(\boldsymbol{z})$.



Applying WA to $p_e(\boldsymbol{x})$ is not always equivalent to $p_{e^+}(\boldsymbol{x}) + p_{e^-}(\boldsymbol{x})$.

Bistratal Wirelength Model [TCAD'23][9]

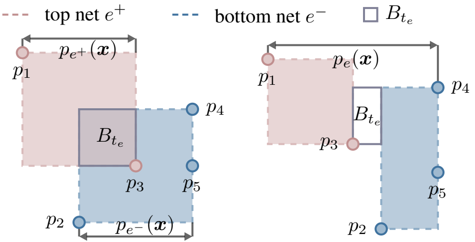


- Where to place the HBT t_e for the crossing-die net e?
- The optimal region $B_{t_e} = [x'_{\min}, x'_{\max}] \times [y'_{\min}, y'_{\max}]$ for the HBT t_e is defined as,

$$x'_{\min} = \min\left\{\max\left\{x^+_{\min}, x^-_{\min}\right\}, \min\left\{x^+_{\max}, x^-_{\max}\right\}\right\},$$
$$x'_{\max} = \max\left\{\max\left\{x^+_{\min}, x^-_{\min}\right\}, \min\left\{x^+_{\max}, x^-_{\max}\right\}\right\}, \quad (3)$$

[9] **Peiyu Liao et al. (2023). "Analytical Die-to-Die 3D Placement with Bistratal Wirelength Model and GPU Acceleration".** In: *IEEE TCAD*.

Bistratal Wirelength Model [TCAD'23][9]



- Where to place the HBT t_e for the crossing-die net e?
- The optimal region $B_{t_e} = [x'_{\min}, x'_{\max}] \times [y'_{\min}, y'_{\max}]$ for the HBT t_e is defined as,

$$x'_{\min} = \min\left\{\max\left\{x^+_{\min}, x^-_{\min}\right\}, \min\left\{x^+_{\max}, x^-_{\max}\right\}\right\},$$
$$x'_{\max} = \max\left\{\max\left\{x^+_{\min}, x^-_{\min}\right\}, \min\left\{x^+_{\max}, x^-_{\max}\right\}\right\}, \tag{3}$$

Bistratal Wirelength

The partial nets e^+ and e^- are determined by the tentative partition $\boldsymbol{\delta} = P(\boldsymbol{z})$.

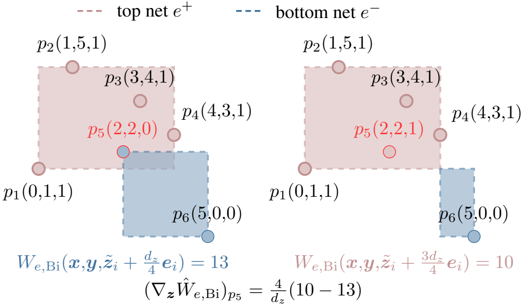
$$W_{e,\mathrm{Bi}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \max\{p_e(\boldsymbol{x}), p_{e^+}(\boldsymbol{x}) + p_{e^-}(\boldsymbol{x})\} + \max\{p_e(\boldsymbol{y}), p_{e^+}(\boldsymbol{y}) + p_{e^-}(\boldsymbol{y})\}. \tag{4}$$

- Apply WA to $W_{e,\mathrm{Bi}} \to \hat{W}_{e,\mathrm{Bi}}$.

[9] **Peiyu Liao et al. (2023). "Analytical Die-to-Die 3D Placement with Bistratal Wirelength Model and GPU Acceleration".** In: *IEEE TCAD*.

Bistratal Wirelength Model

- How to optimize z by $\nabla_{\boldsymbol{z}} \hat{W}_{e,\mathrm{Bi}}$? (non-smooth)



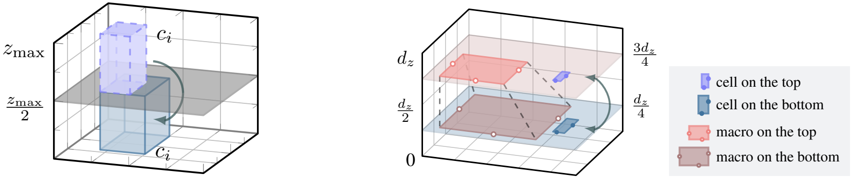
Finite Difference Approximation

The finite difference approximation (FDA) $(\nabla_{\boldsymbol{z}} \hat{W}_{e,\mathrm{Bi}})_i$ is defined by

$$(\nabla_{\boldsymbol{z}} \hat{W}_{e,\mathrm{Bi}})_i = \frac{4}{d_z}\left(W_{e,\mathrm{Bi}}(\boldsymbol{x}, \boldsymbol{y}, \tilde{\boldsymbol{z}}_i + \frac{3d_z}{4}\boldsymbol{e}_i) - W_{e,\mathrm{Bi}}(\boldsymbol{x}, \boldsymbol{y}, \tilde{\boldsymbol{z}}_i + \frac{d_z}{4}\boldsymbol{e}_i)\right), \tag{5}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are fixed, and $\tilde{\boldsymbol{z}}_i = \boldsymbol{z} \odot (\boldsymbol{1} - \boldsymbol{e}_i)$ and $\boldsymbol{e}_i \in \mathbb{R}^{|V|}$ is the unit vector with the i-th entry being 1 and others being 0.

- **Electrostatics-Based Density Model**. We adopt the eDensity3D [ISPD'16] with dynamically updating instance size for different technologies.



(a) (b)

- Let w_i^+ denote the instance width on the top die, and w_i^- on the bottom die. The dynamic width w_i can be derived as

$$w_i = \delta_i w_i^+ + (1 - \delta_i) w_i^- . \tag{6}$$

- For any macro $v_i \in V_M$, we propose to linearly transform macro width as

$$w_i = \left(\frac{2z_i}{d_z} - \frac{1}{2}\right) w_i^+ + \left(\frac{3}{2} - \frac{2z_i}{d_z}\right) w_i^- . \tag{7}$$

Density Model

- eDensity3D computes the density gradient by sovling the 3D Poisson's equation using spectral methods.

$$\begin{aligned} \Delta\phi(\boldsymbol{v}) &= -\rho(\boldsymbol{v}), \quad \boldsymbol{v} \in \Omega \\ \hat{\boldsymbol{n}} \cdot \nabla\phi(\boldsymbol{v}) &= 0, \quad \boldsymbol{v} \in \partial\Omega. \end{aligned} \tag{8}$$

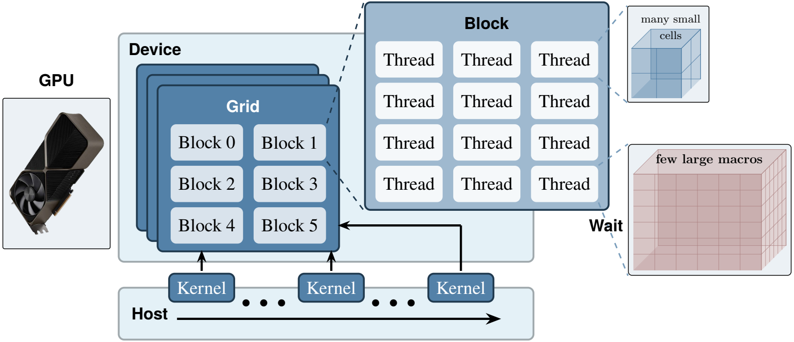
- Let $(\omega_j, \omega_k, \omega_l) = (\frac{j\pi}{d_x}, \frac{k\pi}{d_y}, \frac{l\pi}{d_z})$ denote the frequency indices. The density frequency coefficients a_{jkl} are computed as

$$a_{jkl} = \frac{1}{N} \sum_{x,y,z} \rho(x,y,z) \cos(\omega_j x) \cos(\omega_k y) \cos(\omega_l z), \tag{9}$$

- The electric field $\boldsymbol{E}(x,y,z) = (E_x, E_y, E_z)$ is computed as below

$$\begin{aligned} E_x &= \sum_{j,k,l} \frac{a_{jkl}\omega_j}{\omega_j^2 + \omega_k^2 + \omega_l^2} \sin(\omega_j x) \cos(\omega_k y) \cos(\omega_l z), \\ E_y &= \sum_{j,k,l} \frac{a_{jkl}\omega_k}{\omega_j^2 + \omega_k^2 + \omega_l^2} \cos(\omega_j x) \sin(\omega_k y) \cos(\omega_l z), \\ E_z &= \sum_{j,k,l} \frac{a_{jkl}\omega_l}{\omega_j^2 + \omega_k^2 + \omega_l^2} \cos(\omega_j x) \cos(\omega_k y) \sin(\omega_l z). \end{aligned} \tag{10}$$

Adaptive 3D Density Accumulation

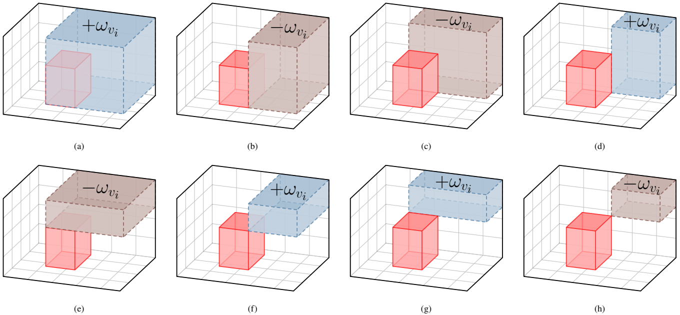


- For each bin $b \in B$ as a cuboid with size $w_b \times h_b \times d_b$, the density is calculated as,

$$\rho_b = \sum_{v_i \in V} \omega_{v_i} \frac{\mathrm{vol}(D_{v_i} \cap b)}{\mathrm{vol}(b)}, \tag{11}$$

  where D_{v_i} indicates the region of the instance v_i in 3D space.
- The computation workload can be very imbalanced for standard cells and macros in 3D scenarios.

3D Prefix Sum for Density Accumulation

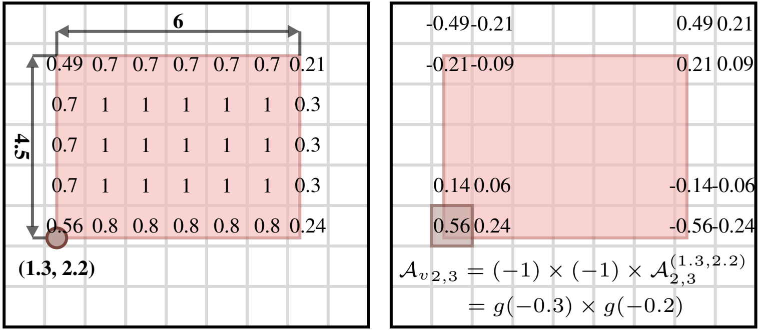


- The 3D density map calculation for macro v_i colored in red is decomposed to the weighted sum of 8 corner maps, performed by 3D prefix sum.
- The *3D prefix sum* is a function $\varphi : \mathbb{R}^{N_x \times N_y \times N_z} \to \mathbb{R}^{N_x \times N_y \times N_z}$ such that

$$\varphi(\mathcal{A})_{ijk} = \sum_{i'=1}^{i} \sum_{j'=1}^{j} \sum_{k'=1}^{k} \mathcal{A}_{i'j'k'} \tag{12}$$

holds for any 3D map $\mathcal{A} \in \mathbb{R}^{N_x \times N_y \times N_z}$ and valid index tuple (i, j, k).

Prefix Sum Illustration



- Let a 3D map $\mathcal{A}^{(x,y,z)}$ be induced according to the following mechanism,

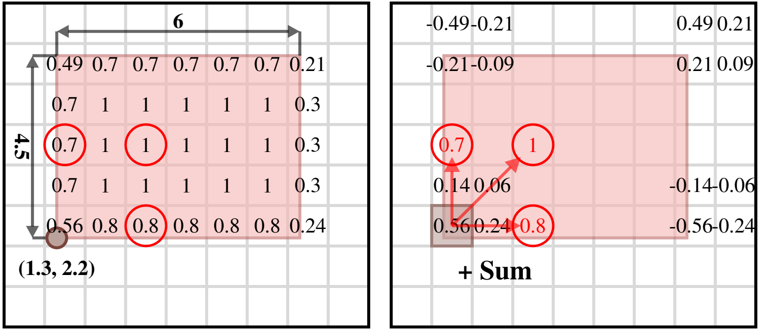
$$g(a) = \max\{1 - |a|, 0\}, \quad \mathcal{A}_{ijk}^{(x,y,z)} = g\left(i - 1 - \hat{x}\right) g\left(j - 1 - \hat{y}\right) g\left(k - 1 - \hat{z}\right), \tag{13}$$

where the normalized coordinates $(\hat{x}, \hat{y}, \hat{z}) := \left(\frac{x}{w_b}, \frac{y}{h_b}, \frac{z}{d_b}\right)$.

- Then the desired 3D map $\mathcal{A}_{v_i}$ for macro v_i can be obtained by

$$\mathcal{A}_{v_i} = \sum_{\sigma_x, \sigma_y, \sigma_z \in \{-1,1\}} -\sigma_x \sigma_y \sigma_z \mathcal{A}^{\left(x_i + \sigma_x \frac{w_i}{2}, y_i + \sigma_y \frac{h_i}{2}, z_i + \sigma_z \frac{d_z}{4}\right)}. \tag{14}$$

Prefix Sum Illustration



- Let a 3D map $\mathcal{A}^{(x,y,z)}$ be induced according to the following mechanism,

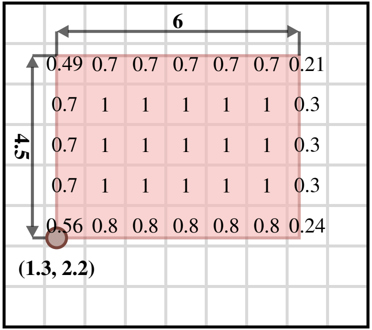
$$g(a) = \max\{1 - |a|, 0\}, \quad \mathcal{A}^{(x,y,z)}_{ijk} = g\left(i - 1 - \hat{x}\right) g\left(j - 1 - \hat{y}\right) g\left(k - 1 - \hat{z}\right), \tag{13}$$

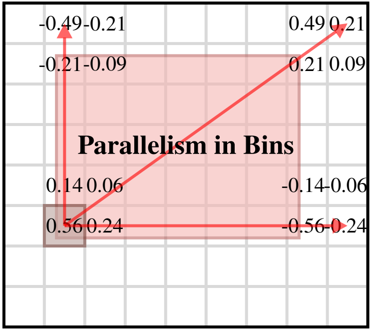
  where the normalized coordinates $(\hat{x}, \hat{y}, \hat{z}) := \left(\frac{x}{w_b}, \frac{y}{h_b}, \frac{z}{d_b}\right)$.

- Then the desired 3D map $\mathcal{A}_{v_i}$ for macro v_i can be obtained by

$$\mathcal{A}_{v_i} = \sum_{\sigma_x,\sigma_y,\sigma_z \in \{-1,1\}} -\sigma_x\sigma_y\sigma_z \mathcal{A}^{\left(x_i+\sigma_x \frac{w_i}{2}, y_i+\sigma_y \frac{h_i}{2}, z_i+\sigma_z \frac{d_z}{4}\right)}. \tag{14}$$

Prefix Sum Illustration





- Let a 3D map $\mathcal{A}^{(x,y,z)}$ be induced according to the following mechanism,

$$g(a) = \max\{1 - |a|, 0\}, \quad \mathcal{A}^{(x,y,z)}_{ijk} = g\left(i - 1 - \hat{x}\right) g\left(j - 1 - \hat{y}\right) g\left(k - 1 - \hat{z}\right), \tag{13}$$

  where the normalized coordinates $(\hat{x}, \hat{y}, \hat{z}) := \left(\frac{x}{w_b}, \frac{y}{h_b}, \frac{z}{d_b}\right)$.

- Then the desired 3D map $\mathcal{A}_{v_i}$ for macro v_i can be obtained by

$$\mathcal{A}_{v_i} = \sum_{\sigma_x,\sigma_y,\sigma_z \in \{-1,1\}} -\sigma_x\sigma_y\sigma_z \mathcal{A}^{(x_i+\sigma_x\frac{w_i}{2}, y_i+\sigma_y\frac{h_i}{2}, z_i+\sigma_z\frac{d_z}{4})}. \tag{14}$$

3D Mixed-Size Preconditioning

- The eDensity families [TODAES'15][10] [ISPD'16] adopt the Jacobi preconditioner,

$$(\boldsymbol{H}_f)_{ii} = \sum_e \frac{\partial^2 \hat{W}_e}{\partial x_i^2} + \lambda \frac{\partial^2 \hat{D}}{\partial x_i^2}. \tag{15}$$

- We propose the 3D mixed-size preconditioner as follows,

$$(\boldsymbol{H}_f)_{ii} \approx \begin{cases} \max\{1, |E_i| + \lambda q_i\}, & \text{if } v_i \in V_M, \\ \max\{1, \lambda q_i\}, & \text{otherwise.} \end{cases} \tag{16}$$

[10] **Jingwei Lu, Pengwen Chen, et al. (2015). "ePlace: Electrostatics-based placement using fast fourier transform and Nesterov's method".** In: *ACM TODAES* 20.2, pp. 1–34.

- **Multi-Die 2D Placement**: We model the top die, bottom die, and bonding terminal layer as independent 2D electrostatic fields,

$$\min_{\boldsymbol{x},\boldsymbol{y}} \sum_{e \in E} \hat{W}_e(\boldsymbol{x}, \boldsymbol{y}) + \langle \boldsymbol{\lambda}, \hat{\boldsymbol{D}} \rangle, \tag{17}$$

- **Legalization**: Die-by-die legalization by Tetris [US Patent'02][11] and Abacus [ISPD'08][12].
- **Detailed Placement**: Die-by-die detailed placement by ABCDPlace [TCAD'20][13].

[11] **Dwight Hill (2002).** ***Method and system for high speed detailed placement of cells within an integrated circuit design*****.** US Patent 6,370,673.

[12] **Peter Spindler, Ulf Schlichtmann, and Frank M. Johannes (2008). "Abacus: fast legalization of standard cell circuits with minimal movement".** In: *Proc. ISPD*, pp. 47–53.

[13] **Yibo Lin et al. (2020). "ABCDPlace: Accelerated batch-based concurrent detailed placement on multithreaded CPUs and GPUs".** In: *IEEE TCAD* 39.12, pp. 5083–5096.

Experimental Results

Experimental Setup

- **Benchmarks:** 3D mixed-size placement benchmarks from ICCAD 2023 contest.
- **Metrics:** Wirelength, HBT count, and runtime. The raw score = HPWL + β#HBTs ($\beta = 10$).
- **Platform:** Intel Xeon Silver 4210R CPU (2.40GHz), 1 GeForce RTX 3090Ti GPU.

Table: The statistics of ICCAD 2023 contest benchmarks [ICCAD'23].

Bench.	#Cells	#Macros	#Nets	RH^+	RH^-	w'	r_{MA}
case2	13901	6	19547	33	33	92	0.88
case2h1	13901	6	19547	33	48	92	0.88
case2h2	13901	6	19547	33	48	92	0.88
case3	124231	34	164429	33	48	56	0.71
case3h	124231	34	164429	36	48	58	0.67
case4	740211	32	758860	92	115	54	0.36
case4h	740211	32	758860	55	69	32	0.36

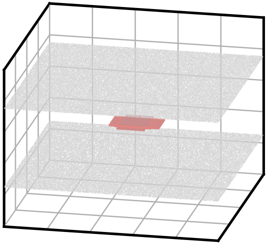
Experimental Results

Table: Placement results compared to the top-3 winners on the ICCAD 2023 contest benchmark.

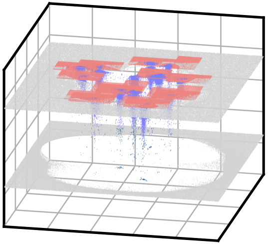
Bench.	1st Place			2nd Place			3rd Place			Ours-CPU			Ours-GPU		
	HPWL	#HBTs	RT	HPWL	#HBTs	RT	HPWL	#HBTs	RT	HPWL	#HBTs	RT	HPWL	#HBTs	RT
case2	16 490 836	1523	67	16 277 152	993	32	16 537 596	2153	144	**15528810**	1128	76	15 622 062	1329	38
case2h1	18 121 844	120	39	19 047 767	821	43	21 156 596	2435	160	16 708 363	1135	80	**16556213**	1349	35
case2h2	18 123 283	120	42	19 193 899	821	41	21 640 714	2426	162	**16748148**	1091	80	16 807 840	1312	36
case3	98 706 330	22 189	534	105 386 847	26 112	104	116 022 515	29 457	602	**97281904**	10 704	236	98 081 778	12 446	92
case3h	122 271 798	18 761	262	120 770 382	5038	104	117 633 295	25 641	612	109 386 719	13 224	239	**108028790**	13 798	86
case4	1 046 106 185	160 993	3605	1 108 969 124	188 137	615	1 130 211 865	138 762	5309	1 040 202 500	132 109	3070	**1036364973**	131 119	335
case4h	654 962 287	156 586	1567	680 554 407	167 686	592	702 244 786	141 916	4492	634 654 510	133 734	2640	**633920946**	133 853	361
Average	1.058	0.981	4.000	1.096	1.019	1.289	1.156	1.660	7.830	1.000	0.907	4.047	**1.000**	1.000	1.000

- **5.8%** HPWL improvement with similar #HBTs compared to the 1st place.
- **4.0**$\times$ runtime speedup compared to the 1st place by GPU acceleration.

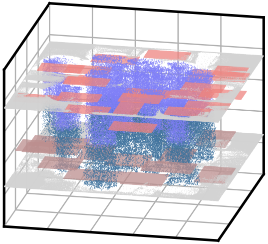
Placement Visualization



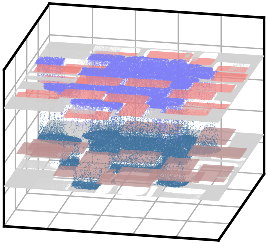
Iter.	WL (10^7)	#HBTs	OVFL
0	3.60	88507	1.00



Iter.	WL (10^7)	#HBTs	OVFL
1000	4.49	1254	0.89

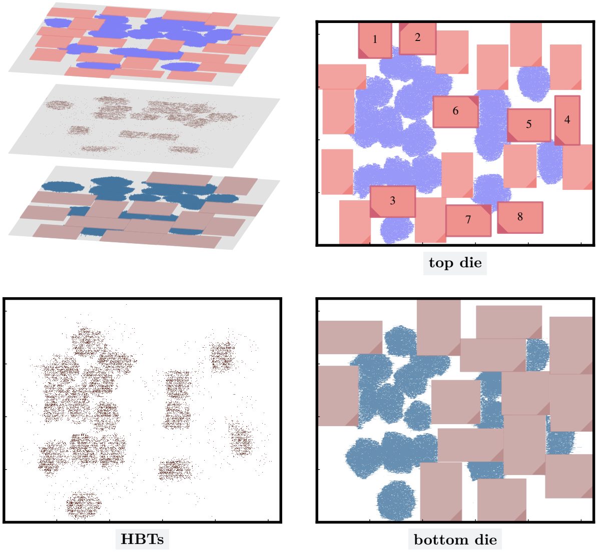


Iter.	WL (10^7)	#HBTs	OVFL
1800	10.30	30961	0.34

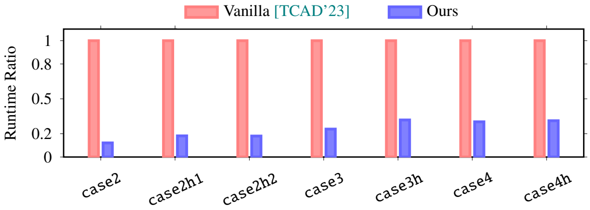


Iter.	WL (10^7)	#HBTs	OVFL
2142	10.51	13798	0.07

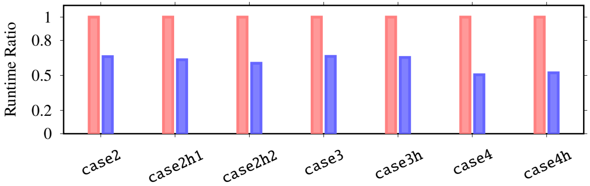
Placement Visualization



Acceleration of Density and Wirelength Algorithms



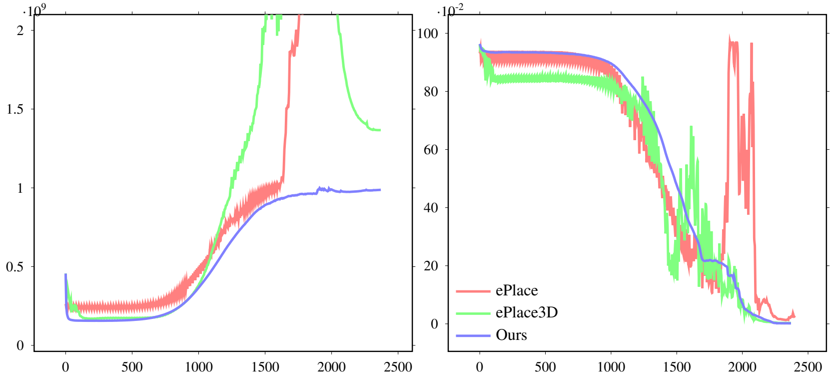
(a) Density algorithm runtime comparison.



(b) Wirelength algorithm runtime comparison.

- **4.7**$\times$ speedup by adaptive 3D density accumulation.
- **1.7**$\times$ speedup by incremental wirelength gradient algorithm.

Ablation Study on Preconditioning



(a) Wirelength curve. (b) Macro Density Overflow.

Ablation study on the mixed-size preconditioner.

- Our mixed-size preconditioner enables stable optimization.

Q&A