FIT: Fill Insertion considering Timing

Bentian Jiang, Xiaopeng Zhang, Ran Chen, Gengjie Chen,
Peishan Tu, Wei Li, Evangeline F. Y. Young and Bei Yu

CSE Department, The Chinese University of Hong Kong

June. 6, 2019

香港中文大學
The Chinese University of Hong Kong

Outline

Dummy Fill Insertion

Dummy fill insertion (DFI) is a mandatory step in modern manufacturing process:

- Insert metal fills into layout;
- Reduce dielectric thickness variation;
- Provide nearly uniform pattern density;
- Highly-Related to the quality of chemical-mechanical polishing (CMP) process.

Mainstream works on dummy fill insertion (DFI) mainly focus on:

- Minimize the **density variation** [2] [Chen+, ISPD'02];
- Minimize the **fill amount** [3] [Feng+, TCAD'11];
- Hybrid objectives: **layer overlay, density variation, line hotspots**, **outlier hotspots, runtime** [5, 4] [Liu+, TODAES'16] [Lin+, TCAD'17]

Timing-aware Dummy Fill Insertion

Inserted metal $\Rightarrow$ Pros: impoves density, increases planarity

$\Rightarrow$ Cons: couples with signal tracks

- With the shrinkage of technology node, the **coupling effects** can severely affect the original layout timing closure:
- Need significant reduction of coupling capacitance impact during the insertion.

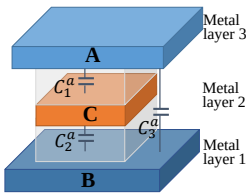
The coupling effect between metal fill and signal track (figure from [1] $*$)

$*$http://iccad-contest.org/2018/Problem_C/2018ICCADContest_ProblemC.pdf

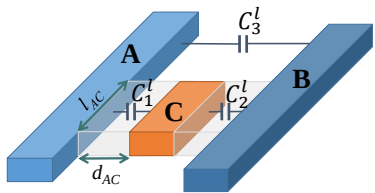
Capacitance Evaluation

There are three main types of capacitances to be considered for evaluation:

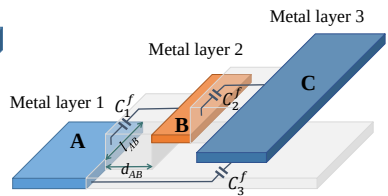
- **Area Capacitance**: Two conductor are on different metal layers, and their projections overlap $\Rightarrow C^a = P_{l_1,l_2}(s) \times s$.
- **Lateral Capacitance**: Two conductor are on same layer and have horizontal overlap $\Rightarrow C^l = P_l(d) \times l$.
- **Fringe Capacitance**: Two conductor pieces are on different layers, and have parallel edge overlap $\Rightarrow C^f = P_{l_1,l_2}(d) \times l + P_{l_2,l_1}(d) \times l$.



(a) Area cap



(b) Lateral cap



(c) Fringe cap

Problem Formulation

Given a design layout, insert metal fills to minimize:

- **Equivalent capacitance** †: The equivalent capacitance of the given critical nets.
- **Overall runtime**.

The insertion result must satisfy the hard constraints on:

- **Density criteria**: A running window of size $w \times w$ and a step size of $\frac{w}{2}$ is considered on each layer, the density inside the window **can not violate** the give density **lower** and **upper** bound.
- **Design rules**: Minimum spacing, minimum fill width, and maximum fill width.

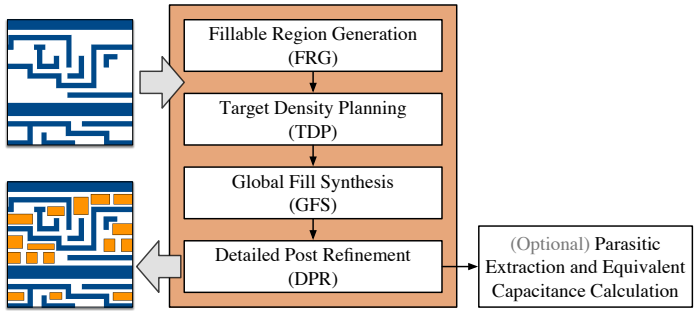
Additionally, the **total parasitic capacitance of all the signal nets** is also considered, since it will affect performance like power consumption, timing.

†Equivalent capacitance to the ground, obtained by network analysis [1].

Outline

Overview of FIT Flow

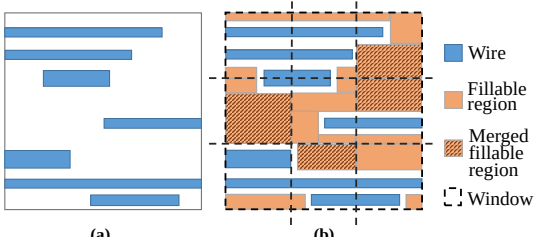
- **Efficient**: Strong runtime performance on ICCAD 2018 benchmarks.
- **Effective**: Outperforms the contest winner by all metrics.
- **Extendable**: Separate modules, easy to further integrate other optimization flow.



Overall dummy fill insertion flow.

Fillable Region Generation

- Extract fillable **polygons** of the entire layer.
- Polygon decomposition: polyons with thousands of vertices and maybe holes inside are difficult to handle $\Rightarrow$ decompse them to **rectangles**, assign rectangles into different windows ($\frac{w}{2} \times \frac{w}{2}$).
- Rectangle aspect ratio fits the layer preferred direction. Merge rectangles locally by sweep line.
- Significantly expand the solution spaces for later procedures.



Fillable region generation

Window density upper bound comparison

	Case 1	Case 2	Case 3	Case 4	Case 5
I-PTR[5]	0.7196	0.7339	0.7196	0.6990	0.6940
Ours	0.7866	0.8009	0.7725	0.7632	0.7642
Improvement	9.30%	9.13%	7.34%	9.18%	10.13%

Target Density Planning

Objective: distribute the target density for each window (under density constraint):

- Divide original window ($w \times w$) into 4 sub-windows with size of $\frac{w}{2} \times \frac{w}{2}$.
- Reduce the critical nets capacitance and total wire capacitance.

$$\min_{\mathbf{D}} \quad \sum_{i,j} \Omega_{i,j} D_{i,j} - \min_{i,j}\{\tau(D_{i,j}^{max} - D_{i,j})\} \tag{1a}$$

$$\text{s.t.} \quad \text{Sub-windows density constraints} \tag{1b}$$

$$\text{Max fillable area constraints} \tag{1c}$$

Weight $\Omega_{i,j}$ measures $\Rightarrow$ the criticality of window W_{ij} (the ratio of critical wires enclosed in),

$$\Omega_{i,j} = \begin{cases} \epsilon, & \text{if} \quad a_{ij}^{c} = 0,\ a_{ij}^{nc} = 0, \\ \omega^{c} \cdot a_{ij}^{c} + \omega^{nc} \cdot a_{ij}^{nc}, & \text{else.} \end{cases} \tag{2}$$

Need to introduce auxiliary variable and constraint to linearize formula (1).

Global Fill Synthesis and Legalization

Global fill synthesis flow (GFS):

- An efficient heuristic window-based flow for high quality initial solution.
- Guided by the target density scheduling result.
- Only performing the GFS flow can already **beat** the contest winner results.

Not dive into details, but list 3 most important criteria:

- Increase the **spacing** and reduce the **parallel overlap lengths** between any two metal conductors.
- Forbid any **area overlap** between fill and the given critical wire.
- Order-Sensitive process, obtain a better insertion order:
  - Sort the windows order by the density gap $D_{max} - D_t$.
  - Sort the fillable rectangles by weighted score of their shape, area, distance and parallel overlap to/with surrounding wires.

$$\alpha \cdot h + \beta \cdot A + \gamma \cdot \sqrt{d_e} + \eta \cdot \frac{1}{l}. \tag{3}$$

Global Fill Synthesis and Legalization

A design rules checker (RTree) is maintained to perform legalization and record density

- Naive implementation: Insert all wires and fills into checker $\Rightarrow$ Time consuming
- Pruning: Global checker + local checkers.
- Local checker responsible for insertion/legalization inside a specific window, discard when finished.
- Global checker keeps wire locations for the entire layer (or one partitioned region of the layer), success insertion of a window only commits those fills that close to window border to global checker.
- Significant runtime improvement for large scale benchmarks.

Detailed Post Refinement

(1) Timing-aware Fill Relocation:

- Relocate those fills (obtained from GFS) with high-impact on timing:

$$\begin{aligned} \min_{\mathbf{A}} \quad & \sum_i \gamma_i A_i \\ \text{s.t.} \quad & \text{Density constraints} \\ & \text{Max fillable area constraints} \end{aligned} \tag{4}$$

- Weight $\gamma_i = \sum_k \frac{l_{ik}}{d_{ik}^2}$ estimates the timing-impact of the fill insertion in ith fillable rectangle $\Rightarrow$ minmize high-impact fills.
- d_{ik} and l_{ik} measure the distance and the parallel overlap between the fill and the closest critical wire.
- Can be solved efficiently by greedy method.

Detailed Post Refinement

(2) Timing-aware Fill Shifting:

- To capture the lateral and fringe capacitance with respect to critical wires.
- $d(f,c)$ and $l_{fc} \rightarrow$ distance and overlap between fill and surrounding critical wire.

$$
\begin{aligned}
\min_{d} \quad & D = \sum_{f \in F} \sum_{c \in C} \frac{l_{fc}}{d^2(f,c)}, \\
\text{s.t.} \quad & d(f,c) = |x_f - x_c| - \frac{1}{2} w_f - \frac{1}{2} w_c, \\
& L + \frac{1}{2} w_f \leq x_f \leq R - \frac{1}{2} w_f, \qquad \rightarrow \text{ Boundary constraint} \\
& |x_f - x_{f'}| \geq \frac{1}{2} w_f + \frac{1}{2} w_{f'} + S_{min}, \ \rightarrow \text{ Fix order constraint} \\
& |x_f - x_b| \geq \frac{1}{2} w_f + \frac{1}{2} w_b + S_{min}, \quad \rightarrow \text{ Fix order constraint} \\
& \forall f, f' \in F, \text{ and } f \neq f', \ c \in C, \ b \in B.
\end{aligned}
\tag{5}
$$

Timing-aware Fill Shifting

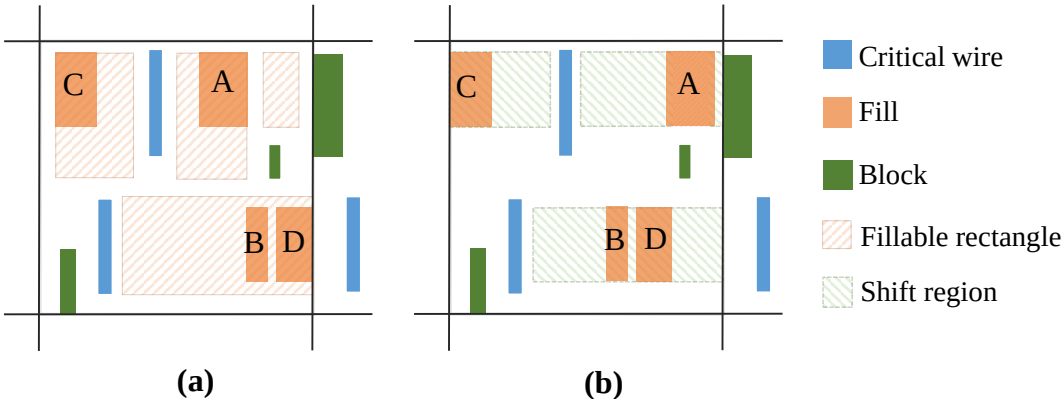
- Relative order between any conductors is fixed to smooth the objective function.
- Use traditional gradient descent.
- Alternatively optimize X-dimension and Y-dimension.

$$\frac{\partial D}{\partial x_f} = \begin{cases} \sum_{c \in C} \dfrac{-2 \cdot l_{fc}}{(x_f - x_c - \frac{1}{2}w_f - \frac{1}{2}w_c)^3}, & \text{if } x_f \geq x_c, \\ \sum_{c \in C} \dfrac{-2 \cdot l_{fc}}{(x_f - x_c + \frac{1}{2}w_f + \frac{1}{2}w_c)^3}, & \text{if } x_f < x_c, \end{cases} \tag{6}$$

$$x_f^{(t+1)} \leftarrow x_f^{(t)} - \alpha \frac{\partial D}{\partial x_f^{(t)}}, \quad f \in F,$$

Timing-aware Fill Shifting

The shifting refinement is regardless of original fillable region limitation.



(a) The positions of fills A and C are limited by the fillable rectangles; (b) Fills A and C jump out of the fillable rectangles.

Outline

Experimental Results

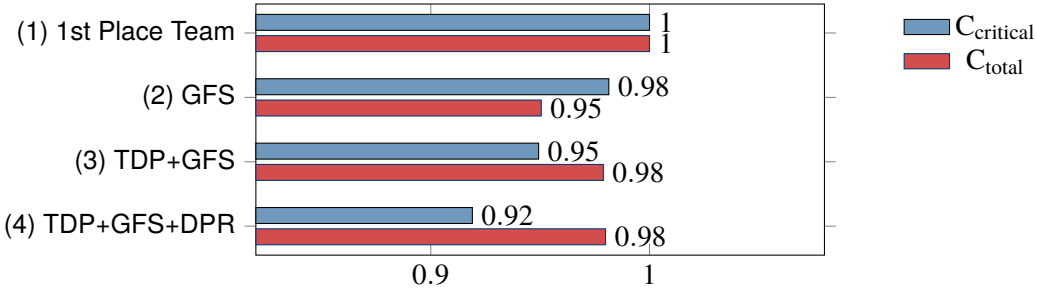
- On ICCAD 2018 Contest Benchmarks
- Capacitance evaluation tool is released by the contest organizers [1]

case	# wires	# critical wires	1st place team				FIT			
			RT-s (s)	RT-m (s)	$C_{critical}$ (pF)	C_{total} (pF)	RT-s (s)	RT-m (s)	$C_{critical}$ (pF)	C_{total} (pF)
case1	305667	12897	19.10	19.10	33.11	11313.69	8.80	5.16	30.94	10883.66
case2	750166	33325	61.83	61.83	79.41	39612.26	31.44	18.41	73.53	39523.68
case3	64903	5307	3.51	3.51	13.01	1669.72	1.59	1.09	11.80	1558.17
case4	149464	11896	7.52	7.52	25.46	3136.48	3.55	2.43	23.25	2969.20
case5	275425	22813	15.14	15.14	50.89	6150.53	6.97	4.62	45.97	5705.39
Total	-	-	107.10	107.10	201.88	61882.68	52.34	31.71	185.48	60640.10
Ratio	-	-	1.000	1.000	1.000	1.000	**0.489**	**0.296**	**0.919**	**0.980**

* RT-s denotes overall runtime in single thread mode, RT-m denotes overall runtime in 8-threads.

- FIT framework outperforms the contest winner in all metrics.
- **8%** reduction on critical nets capacitance, **2%** reduction on total capacitance of all nets. **2**× runtime speedup in single-thread execution, and **3.37**× in multi-thread execution.

Experimental Results



- Global fill synthesis stage is very effective, already beats the contest winner.
- Target density planning and detailed post refinement stages can significantly further reduce critical capacitance.
- Target density planning improves the **critical** capacitance at the expanse of **total** capacitance.

Outline

Future Works

(1) Better multi-thread scheduling:

- Original layer-based implementation $\Rightarrow$ Wait for the slowest layer, inefficient!
- Conflict-free window-based scheduling $\Rightarrow$ Higher resource utilization.

(2) Fast incremental optimization framework:

- ML-based capacitance evaluator to extract parasitic cap locally.
- More comprehensive objective function for optimization.

(3) Fill sizing:

- Originally determine fill size by heuristic.
- Formulate fill sizing problem related to layer overlay?

Q&A

Thanks and Questions?


Bo, Y., and Sriraaman, S.

ICCAD-2018 CAD contest in timing-aware fill insertion.

In *Proc. ICCAD* (2018).


Chen, Y., Kahng, A. B., Robins, G., and Zelikovsky, A.

Closing the smoothness and uniformity gap in area fill synthesis.

In *Proc. ISPD* (2002), pp. 137–142.


Feng, C., Zhou, H., Yan, C., Tao, J., and Zeng, X.

Efficient approximation algorithms for chemical mechanical polishing dummy fill.

IEEE TCAD 30, 3 (2011), 402–415.


Lin, Y., Yu, B., and Pan, D. Z.

High performance dummy fill insertion with coupling and uniformity constraints.

IEEE TCAD 36, 9 (2017), 1532–1544.


Liu, C., Tu, P., Wu, P., Tang, H., Jiang, Y., Kuang, J., and Young, E. F. Y.

An effective chemical mechanical polishing filling approach.

In *Proc. ISVLSI* (2015), pp. 44–49.