Identifying Redundant Inter-Cell Margins and Its Application to Reducing Routing Congestion

Woohyun Chung, Seongbo Shim, and Youngsoo Shin
Department of Electrical Engineering
KAIST, Daejeon 305-701, Korea
Email: wohychung@gmail.com

***Abstract*— A modern standard cell is embedded with extra space, called inter-cell margin, on its left and right ends. Margins are sometimes redundant, and so margins between some cell pairs can be removed for the benefit of area. Lithography simulations on whole layout to identify redundant margins take excessive amount of time, and thus are impractical. We propose to determine in advance the redundancy of margins between each cell pair; a few methods of approximation are introduced to accelerate the process, e.g. grouping cell pairs of similar boundary patterns, refining each group with geometry parameters, etc. Experiments indicate that the redundancy of margin is accurately determined in 93.7% of cell pairs; the remaining 6.3%, which are actually redundant, are declared irredundant by our method, so our method is inaccurate for those cell pairs yet is still safe. We take advantage of redundant margins and address the problem of routing congestion reduction. Placement is locally perturbed to identify more redundant margins; the cells in high congestion region are spread out after the margins in low congestion area are removed. The proposed method was evaluated on a few test circuits using 28-nm technology. The number of routing grids with congestion overflow was reduced by 43% with no impact on total wirelength.**

I. Introduction

As technology node shrinks down to 32-nm and below, a pattern failure from lithography process, e.g. contact bridge and metal short, is more likely to occur. A standard cell is carefully designed to avoid from any such failure through repeated layout modification, retargeting and OPC (optical proximity correction), and lithography simulation [1]–[3]. In addition, an extra space called inter-cell margin is attached to both sides of a cell as illustrated in Fig. 1(a) [4]; this is to prevent any inter-cell pattern failure that may arise along cell boundary. A pair of margins is typically the width of a single poly pitch and a dummy poly is inserted in the margin so that polys become regularly placed for better lithography [5], [6].

Margins are introduced to help lithography, particularly in contact and metal 1 layer, and lithography is pattern dependent. So, it is tempting to argue that margins between some cell pairs would be redundant and be taken away for the benefit of area. This is indeed the case, e.g. margins between NAND2 and INV are redundant. The challenge is that identifying redundant margins requires many lithography simulations thus takes a long time. Furthermore, there are numerous cell pairs if we want to determine in advance the redundancy of margins between all cell pairs.

We propose to approximate the process. For this purpose, we group cell pairs of similar patterns along cell boundary, in contact and metal 1 layer, respectively. Each group in turn is parameterized, e.g. vertical distance between two single contacts. A few lithography simulations are performed for each group and for each parameter value, which collectively yields the redundancy of margins as a function of parameter; it then serves as a basis to determine whether margins are redundant for each cell pair.

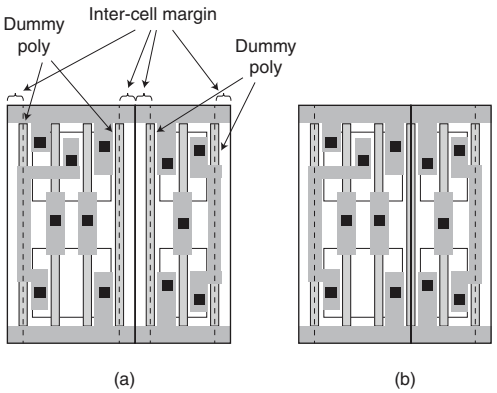

Fig. 1. (a) A typical layout of NAND2 and INV in 32-nm technology, and (b) after removing margins in-between.

Removing margins yields compact layout as shown in Fig. 1(b). In this paper, we take advantage of this to reduce routing congestion. Initial placement and routing, as well as clock tree synthesis, are performed with standard cells with margins. To minimize a disturbance of the layout, we aim to reduce congestion in row-by-row basis. Redundant margins are identified and removed from the cells of low congestion; in this process, a few cells are grouped and their positions are swapped to expose more number of redundant margins. This lets the cells with high congestion in the same row be spread out thereby reducing their congestion. In this process, the extent of spread out in all rows are made as equal as possible. An experiment on a few test circuits demonstrates on average of 43% reduction in the number of grids with congestion overflow with no impact on total wirelength; this comes at the cost of 21% increase in physical design runtime.

The remainder of this paper is organized as follows. In the next section, we address the method of characterizing the redundancy of margins and its validation in 28-nm technology. Reducing routing congestion and its application to test circuits

978-3-9815370-4-8/DATE15/©2015 EDAA

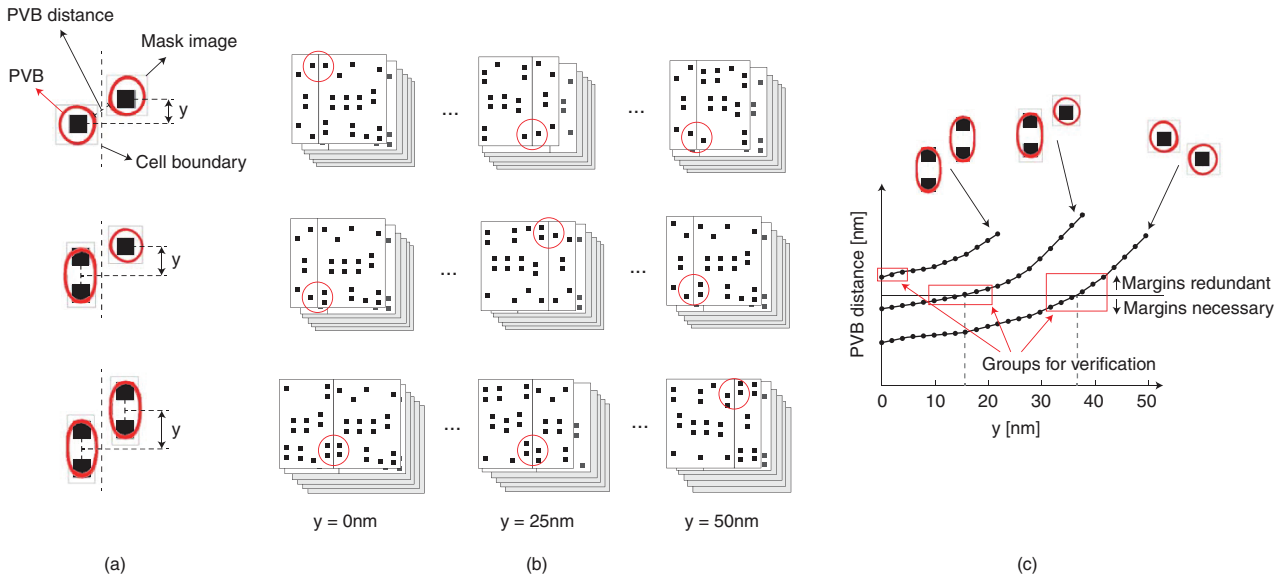

Fig. 3. (a) Adjacent contacts along cell boundary, (b) grouping cell pairs, and (c) PVB distance as a function of vertical contact distance y.

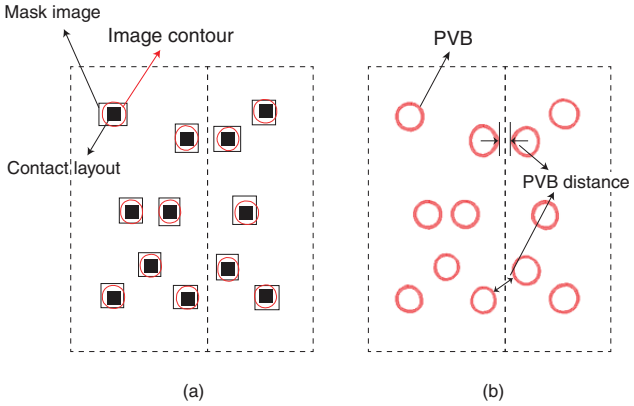

Fig. 2. (a) Contact layout, mask image after retargeting and OPC, and image contour from lithography simulation; (b) PVB from a set of lithography simulations with various lithography settings.

are presented in Section III. We draw conclusion in Section IV.

II. Redundant Inter-Cell Margin

Suppose that two cells are placed without margins in-between as shown in Fig. 1(b). For each layer of contact and metal 1, we apply retargeting and OPC to obtain mask image, and then perform lithography simulation, which returns image contour (see Fig. 2(a)). The process is repeated while we change lithography settings (scanner focus, exposure energy, and mask error [7], [8]). The result is aggregated image contours, called PVB (process variation band), as shown in Fig. 2(b). The minimum distance between adjacent PVB pairs along cell boundary, which we call PVB distance, is a parameter of our concern. If PVB distance is larger than foundry-provided threshold [7], [8], all images can be safely patterned out without any defects, implying that margins are redundant in that particular layer. Margins are declared redundant if they are redundant both in contact and metal 1 layer. Otherwise margins have to be kept as in original cell layout. The whole process takes roughly three seconds for one cell pair. If a library contains 1000 cells, total number of cell pairs are 4 million, and runtime reaches 100 days, which is impractical.

A. Contact Layer

There are two forms of contact: single or double. A double contact is two contacts laid out very close, which eventually become one big contact after manufacturing. Double contacts are used to improve lithography yield of contact layer.

Two adjacent contacts along cell boundary are thus one of the three shown in Fig. 3(a). The distance in horizontal direction is constant (at minimum value) because polys are placed in regular pitch and contacts are located in between two polys. Thus each case of Fig. 3(a) is further classified with the vertical distance between contacts as a parameter (see Fig. 3(b)). One cell pair may have more than one adjacent contact pair at minimum horizontal distance, so it may belong to more than one group.

In each group of Fig. 3(b), we randomly pick 50 cell pairs and obtain PVB distances through lithography simulations. The minimum value of all PVB distances is then compared to foundry-provided threshold, which guides us to declare whether the margins of all cell pairs in that group are redundant or not. We have observed the followings in our 28-nm technology library, which are also illustrated in Fig. 3(c):

- The margins between two double contacts are always redundant.
- The margins between a double and a single contact are redundant if they are separated by more than 17nm in y direction.
- The margins between two single contacts are redundant if y distance is larger than 38nm.

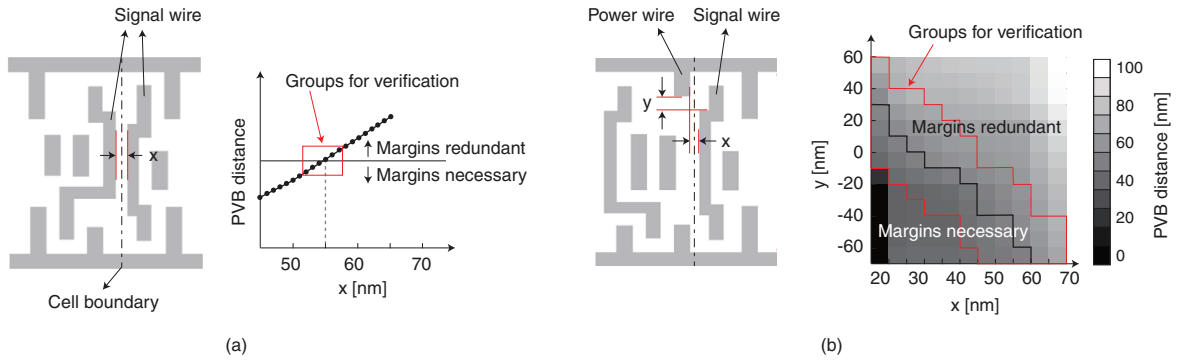

Fig. 4. Assessment of margins (a) when two signal wires face each other and (b) power wire and signal wire are adjacent at cell boundary.

B. Metal 1 Layer

When M1 wire is used as a signal wire within a cell, it connects two contacts on different diffusion layers so it runs vertically along cell boundary as shown in Fig. 4. If it is used as a power wire, it connects a contact to nearby power rail so it runs in short distance.

Cell pairs can be grouped in two cases in metal 1 layer. If two signal wires face each other at cell boundary (Fig. 4(a)), the horizontal distance x is a parameter that determines the redundancy of margins. If signal wire faces power wire as shown in Fig. 4(b), both horizontal and vertical distance (between the two rectangular wire segments closest to cell boundary) become parameters. Two power wires may face each other, but we do not consider this case because they always connect to the same power rail.

Similar to Section II-A, we randomly pick 50 pairs from each group, obtain PVB distances, choose minimum value, and compare it to foundry-provided threshold. We have observed the followings:

- In Fig. 4(a), margins are redundant if x is larger than 55nm.
- In Fig. 4(b), margins are redundant if (x, y) is located in the upper right region.

C. Experiments

The proposed procedure was applied to our library in 28-nm technology, which consists of 1043 cells. Note that there are more than 4 million combinations of cell pairs and cell orientations (cell pairs for short). Our method, which is approximation, has to be conservative in a sense that redundant margins may be declared as irredundant but irredundant margins have to be declared correctly. For this purpose, we heuristically increase the foundry-provided threshold by 5% in contact and metal 1 layer, respectively.

The results are illustrated in Fig. 5. Each bar indicates the case we have classified in each layer; the white region of bar corresponds to redundant margins. The collective percentage of redundant margins is 77% in contact and 41% in metal 1 layer. The margins are eventually redundant if they are redundant in both layers; 35% of cell pairs belong to this category.

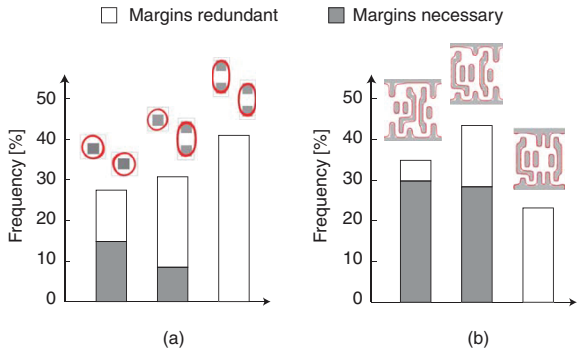

Fig. 5. Experimental results in 28-nm technology (a) in contact layer and (b) metal 1 layer.

To assess the accuracy of our method and to make it sure that our method is indeed conservative, we verified whole cell pairs in 6 groups near the threshold, which are indicated as red boxes in Fig. 3(c), Fig. 4(a) and (b). About 110000 cell pairs were picked, and the redundancy of margins of each cell pair was determined through lithography simulations, and compared to the redundancy estimated by our method. Out of the whole pairs, 31.1% have irredundant and 68.9% have redundant margins. All irredundant margins were correctly estimated. Out of 68.9%, 62.6% were correctly estimated to be redundant, and the remainders (6.3%) were estimated incorrectly but in conservative way as we expect.

III. Post-Placement Optimization to Reduce Routing Congestion

A. Overview

The overall flow to reduce routing congestion (congestion for brevity) is shown in Fig. 6. Standard steps of placement, clock tree synthesis, and routing are first performed using conventional cells having inter-cell margins. Layout is divided into a set of square grids (see Fig. 7); congestion is the number of occupied routing tracks divided by the number of available tracks, and is calculated for each grid; the grids whose congestion exceeds 100% are called overflow grids.

Reducing congestion should be done for the grids that are neighbors of overflow grids as well as for overflow grids them-

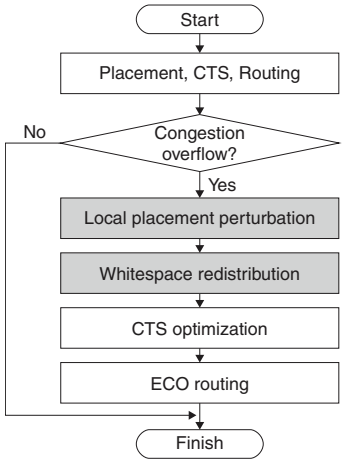

Fig. 6. Overall flow to reduce routing congestion.

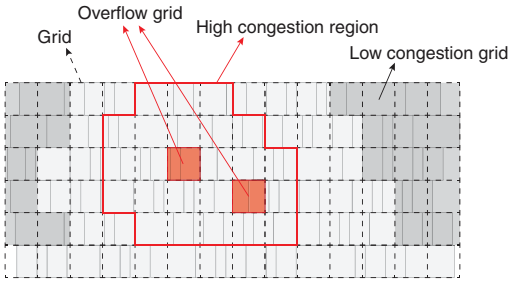

Fig. 7. Overflow grids, high congestion region and low congestion grids.

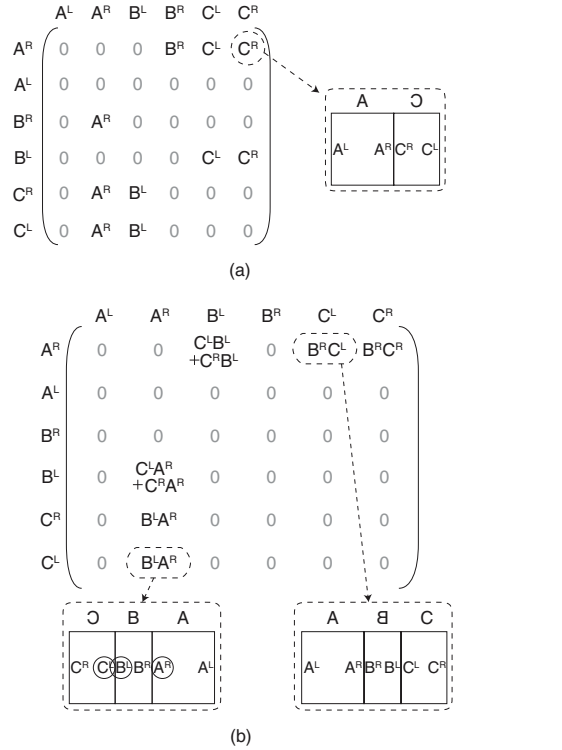

Fig. 8. (a) A matrix to denote redundancy of margins, and (b) taking a square of matrix yields the best placement.

selves. For this purpose, we calculate the average congestion of the grids that are within a circle with each overflow grid at a center; we gradually increase the size of the circle until the average congestion does not increase anymore; corresponding set of grids are treated as high congestion region as illustrated in Fig. 7.

The cells in the high congestion region should be spread out uniformly, because substantially different amount of spread out in different rows may lead to large increase of wirelength. In order to prevent this, we calculate the amount of spread out for each row so that when actual spread out is performed it is done uniformly in all the cells of high congestion region, which corresponds to how many redundant margins are required in each row.

In each cell row that contains the grids of high congestion region (the top five rows in Fig. 7), we define the grids as low congestion ones if their congestion is smaller than some threshold (70% in our experiment) and if they do not lie in high congestion region. We then scan the cells, which belong to the grids with low congestion, from left to right within a cell row; group a few of them and adjust their placement, i.e. swap their locations and flip the orientation, in a way that more redundant margins are identified (this is named local placement perturbation in Fig. 6 and its details are addressed in Section III-B). We repeat this process until required number of redundant margins are generated. The redundant margins are then taken out so that the cells in high congestion region can now be spread out (called whitespace redistribution in Fig. 6).

Some cells may have been displaced after local placement perturbation and whitespace redistribution, so clock tree is re-built through CTS optimization. ECO routing is finally performed to re-connect disconnected wires.

B. Local Placement Perturbation

We are given n adjacent cells, and we want to determine the best placement (the order of cells and the orientation of each cell along y-axis) maximizing the number of redundant margins. There are $n!2^n$ combinations we need to examine; we introduce a matrix to simplify the process.

Let $n = 3$ and A, B, and C be the name of cells. The redundancy of margins between all cell pairs is arranged as a matrix shown in Fig. 8(a). The margin on the right and left of each cell is denoted by superscript R and L, respectively. If margins between X and Y are redundant, the entry (X, Y) takes Y as a value; otherwise it takes 0. Note that the entry (A^R, C^R) implies C flipped along y-axis; note also that the row and column of matrix are indexed in different order, i.e. the column starts with A^L while the row starts with A^R.

To obtain the best placement of three cells, we multiply the matrix by itself as shown in Fig. 8(b). It is easy to see that the entry (X, Y) takes ZY as a value if both margins between X and Z and between the opposite side of Z and Y are redundant; otherwise it takes 0. Therefore, all the entries with non-zero value correspond to the best placement. We pick the one that is

TABLE I

COMPARISON OF STANDARD PLACEMENT AND ROUTING, AND OUR METHOD TO REDUCE ROUTING CONGESTION

Circuit			Standard P&R			Proposed method		
Name	# Cells	Util.	Wirelength (mm)	# Overflow grids	Runtime (sec)	Wirelength	# Overflow grids	Runtime
tv80	4612	0.70	46	287	327	0.96	0.47	1.13
mem_ctrl	5006	0.82	55	186	313	0.99	0.80	1.23
b15	5234	0.65	49	91	379	0.97	0.36	1.18
ac97	7005	0.54	78	26	310	0.99	0.27	1.19
usbf	7194	0.64	105	605	437	0.97	0.61	1.27
b14	7771	0.8	67	172	372	0.98	0.66	1.24
b20_1	10650	0.75	94	42	488	0.95	0.55	1.23
b21_1	10879	0.8	85	61	435	0.98	0.61	1.26
aes_cipher	13070	0.56	164	554	736	0.98	0.38	1.19
b17	15837	0.80	142	192	688	0.97	0.71	1.10
b20	16056	0.75	136	101	750	0.96	0.66	1.19
b22_1	16462	0.75	143	251	759	0.97	0.49	1.27
b21	16467	0.75	134	199	731	1.00	0.66	1.27
b22	24214	0.70	207	201	999	0.96	0.65	1.24
Average						0.98	0.57	1.21

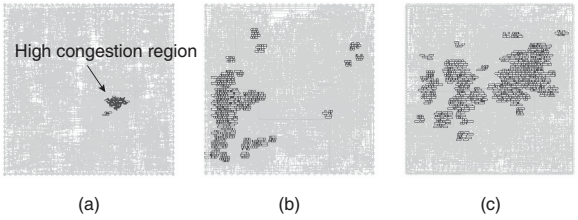

Fig. 9. High congestion regions: (a) ac97, (b) b15, and (c) mem_ctrl.

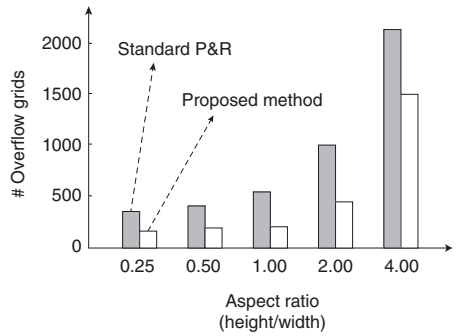

Fig. 10. Congestion reduction in aes_cipher with varying aspect ratio of placement region. Only M3 is used for vertical routing, so the number of overflow grids increases as aspect ratio increases.

closest to the initial placement of cells. Note that some entries have two values added up, e.g. the entry (A^R, B^L) and (B^L, A^R), which implies there are two solutions.

If the best placement for the three cells is determined, we include the next cell to the cell group and continue this process for the four cells. When the fourth cell is not available, we restart this process for the next cell group. Before proceeding to next group, we select one among the best solutions, which is the most similiar to the original placement (the entry (A^R, C^L) in Fig. 8(b)).

In general, $2n$-by-$2n$ matrix is set up for n cells. Multiplying the matrix by itself $n-1$ times yields a few non-zero entries, which correspond to the best placement. In our experiment, n is restricted to 4, which is empirically determined for the sake of runtime and minimizing the extent of cell displacement.

C. Experiments

The proposed method was assessed by using 14 test circuits from OpenCores [9] and ITC99 [10], which are listed in Table I. All the experiments are done automatically with the aid of the commercial EDA tool [11]. For each circuit, a few iterations of placement were tried to determine the minimum amount of whitespace while circuit timing constraints are satisfied; corresponding area utilization is shown in the third column. Routing layers up to metal 4 were used; this was intended to cause non-negligible amount of routing congestion, which we then try to reduce using our method. Columns 4–6 correspond to total wirelength, the number of overflow grids, and runtime, respectively, when standard placement and routing are applied.

Corresponding figures (as ratios) after the proposed method is applied are shown in the last three columns. The number of overflow grids and total wirelength are reduced to 57% and 98% on average, respectively. Circuit ac97 benefits most. As shown in Fig 9(a), its high congestion region is localized in a small area with enough number of low congestion grids nearby. In b15, substantial area corresponds to high congestion region as shown in Fig. 9(b). Fortunately, the region is tall in vertical direction, and since congestion is reduced in row-by-row basis, enough number of redundant margins could be removed in most rows, which explains the great reduction of congestion. Fig 9(c) corresponds to high congestion region of mem_ctrl. The region is widespread in horizontal direction, which explains why the circuit benefits least.

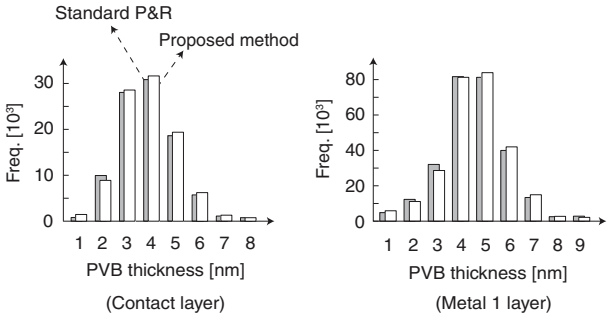

Fig. 11. Histogram of PVB thickness of contact and metal 1 layer for aes_cipher.

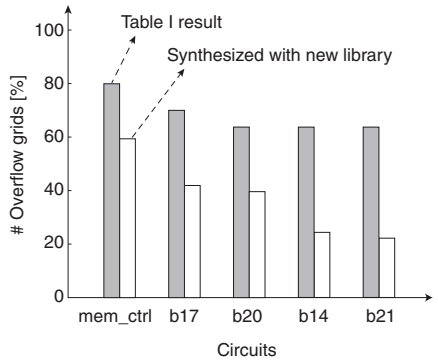

Fig. 12. The number of overflow grids after the proposed method when circuits are synthesized using original library and new library.

The proposed flow shown in Fig. 6 takes 21% more runtime than standard placement and routing. CTS optimization and ECO routing are responsible for 16%; local placement perturbation and whitespace redistribution take only 5%.

In our method, whitespace is redistributed in row-by-row basis. So if placement is done in taller region and each row becomes shorter, our method becomes less effective; the opposite is true as placement region becomes fatter. This is experimentally verified with aes_cipher as shown in Fig. 10 (Table I is based on aspect ratio of 1.0). This can be alleviated to some extent if the cells in high congestion region are allowed to be spread out in non-uniform fashion; this however comes at the cost of wirelength increase. An improved solution is left for future investigation.

Layouts of all test circuits have been verified by strict lithography simulations in contact and metal 1 layers. PVB distances of all cell pairs in the layout have been examined, and no lithography defect was detected. PVB thickness is also a measure of lithography defect [12], so we extracted its distribution as a histogram from both standard placement and routing and our method. As shown in Fig. 11, the two methods achieve almost identical histograms although many inter-cell margins are removed in the proposed method, which is understandable consequence of removing only the redundant margins.

There are 4172 cell pairs that contain 1X NAND2 gate; the margins of 74% cell pairs are redundant. A gate 11X BUF is also involved in 4172 cell pairs but only 11% pairs have redundant margins. We dropped such gate as 11X BUF from the library if percentage of cell pairs (containing the gate) whose margins are redundant are less than 40%; 578 cells (out of 1043) were dropped. A design was synthesized with new library, followed by reducing routing congestion through our method. The resulting number of overflow grids (white bars) are shown in Fig. 12 for 5 test circuits, and are compared to corresponding figures of Table I shown in grey bars. Further reduction in overflow grids has been expected since circuits are newly synthesized using only cells that are likely to have redundant margins and thus are more likely to benefit from our method of reducing routing congestion. Circuit area increases due to less choice of gates during synthesis, but only marginally (about 3%).

IV. Conclusion

A systematic method has been proposed to approximately determine, in advance, whether margins between each cell pair are redundant or not. In experiments with 28-nm technology library, the redundancy of margins has been correctly estimated in 93.7% cell pairs; the remaining 6.3% cell pairs have redundant margins but the method has estimated them to be irredundant, which on the other hand is safe in lithography perspective.

We have addressed the problem of routing congestion reduction by taking advantage of redundant margins. The overflow grids have been reduced by 43% on average of test circuits. The current method is applied in row-by-row basis in an effort to minimize a perturbation in initial placement and routing. More aggressive methods would benefit future investigation.

References

[1] K. Peter, R. Marz, S. Grondahl, and W. Maurer, "Litho-friendly design (LfD) methodologies applied to library cells," in *Proc. SPIE*, Mar. 2013, pp. 8 684 031–8 684 039.
[2] C. Andrus and M. R. Guthaus, "Lithography-aware layout compaction," in *Proc. Great Lakes Symp. on VLSI*, May 2012, pp. 147–152.
[3] S. Paek, J. Kang, N. Ha, and B. Kim, "Yield enhancement with DFM," in *Proc. SPIE*, Oct. 2012, pp. 128–138.
[4] L. Oscar, J. Manoj, T. Kong-Beng, and C. Harry, "Standard cell architecture and methods with variable design rules," U.S. Patent 0 155 783A1, June 24, 2010.
[5] J. Wang, A. K. Wong, and E. Y. Lam, "Standard cell design with regularly placed contacts and gates," in *Proc. SPIE*, Mar. 2004, pp. 55–66.
[6] V. Dal Bem, P. Butzen, F. Marranghello, A. Reis, and R. Ribas, "Impact and optimization of lithography-aware regular layout in digital circuit design," in *Proc. Int. Conf. on Computer Design*, Oct. 2011, pp. 279–284.
[7] L. Liebmann, S. Mansfield, G. Han, J. Culp, J. Hibbeler, and R. Tsai, "Reducing DfM to practice: the lithography manufacturability assessor," in *Proc. SPIE*, Feb. 2006, pp. 786–798.
[8] J. P. Cain, "Design for manufacturability: a fabless perspective," in *Proc. SPIE*, Mar. 2013, pp. 8 684 031–8 684 039.
[9] "Opencores," http://www.opencores.org/.
[10] "ITC99," http://www.cerc.utexas.edu/itc99-benchmarks/bench.html.
[11] Synopsys, "IC Compiler Implementation User Guide, Version H-2013.03-SP4," Sept. 2013.
[12] J. Andres and T. Robles, "Integrated circuit layout design methodology with process variation bands," U.S. Patent 0 251 771A1, Nov. 10, 2005.