Journal of Engineering and Applied Sciences 13 (16): 6855-6859, 2018
ISSN: 1816-949X

A New Physical Design Approach for Setup Timing Optimization in 7 nm Global Routed Designs

[1]Mohamed Chentouf, [1]Lekbir Cherif and [2]Zine El Abidine Alaoui Ismaili
[1]Calypto Division, Mentor Graphics, Sala El Jadida, Morocco
[2]Information, Communication and Embedded Systems (ICES) Team, ENSIAS, University Mohammed V, Rabat, Morocco

Abstract: The scaling of semiconductor technologies to the atomic level has several important consequences on design performance. Today's System on Chips (SoCs) may operate in several Gigahertz (GHz) frequency. Particularly for 7 nm technology that offers more speeds compared to 10, 14 and 16 nm nodes with higher density, a greater number of I/OS and more metal layers (routing capability improvement). With all these technology advancements, new challenges have raised, one of these challenges is the inte rconnect domination of the overall circuit's performance, this new physical constraint should be taken into consideration as early as possible in the design development cycle to achieve design timing closure. In this study, we aim to close timing by reducing net delays using layer optimization during global routing. The idea is to use the difference in 7 nm metal layers electrical properties to meet its timing requirements. The method effectively optimizes timing in nets by faithfully decreasing the effective resistance and the coupling capacitance between wires. By the end of this research, we will propose a complete algorithm for timing closure in a 7 nm scale design. This approach has been shown to be effective with a gain of 35.5% in the Worst Negative Slack (WNS), 26.66% in the Total Negative Slack (TNS), 4.29% in the Total Hold Slack (THS) and 17.91% in the Worst Hold Slack (WHS) respectively, compared to the baseline flow.

Key words: Technology nodes, timing optimization, global route, detail route, wire delay, Worst Negative Slack (WNS), Total Negative Slack (TNS), Back-End-of-Line (BEOL), Non-Default-Rules (NDRs), Self-Aligned Double Patterning (SADP)

INTRODUCTION

As the IC fabs continue to reduce technology nodes to an atomic scale, interconnect delay has become the main contributor in the chip timing losses (Zhang *et al*., 2012). With this scale reduction, interconnect resistance per unit length increases and the capacitance per unit length remains constant while logic delay is continuously decreasing (Murgai, 2015). These features have made interconnect delay the most dominant factor to determine the system performances (Chen *et al*., 2017).

To optimize interconnect delay, there exist many techniques such as "Buffer insertion", "Buffer sizing" and "Wire sizing" which have been shown to be effective and have been well studied in the literature (Liu *et al*., 2014; Chen and Liu, 2007; Alpert and Devgan, 1997).

The first works on "Wire sizing" are approached in (Bhattacharya *et al*., 2016; Sarfati *et al*., 2017) and consists of chopping wire segments into several small segments. For and "Buffer sizing" techniques, we respectively aim to insert buffers between two segments and to modify their length in order to minimize interconnects timing consumption (Alaghi *et al*., 2017). However, it is no more suitable to apply these classical algorithms to reach the best timing optimization. As VLSI circuits are fast growing, searching for evolutionary solutions has been increased (Karimi and Ahmadi, 2014).

"Layer optimization" is an efficient solution that can solve a variety of interconnect timing optimization issues without impacting the IC area and consists of performing a net rerouting governed by some routing rules. In fact, the Self-Aligned Double Patterning (SADP) technology, which is one of the important features of 7 nm technology nodes, manifests a wide difference in layer resistance between SADP and No-SADP layers, this property will guide the global router to use No-SADP layers in performing nets rerouting. This new layers property is advantageous for two reasons. First, No-SADP layers are thicker, so, the cross sectional area of interconnect is more important and thus the resistance is much lower. Second, increasing the spacing between wires reduces the performance impact caused by the sidewall capacitance between parallel adjacent lines.

This study unfolds as follows. In Section 2, we will give a brief description of Place and Route (P&R) flow and highlight where timing optimization is performed using which methodology and tool. In Section 3, the "Layer

Corresponding Author: Mohamed Chentouf, Mentor Graphics Corporation, Calypto Division, Rabat, Morocco

optimization" algorithm will be explained and the experimental results of a comparison study between baseline and the new method is also included. Finally, conclusions and future works will be provided in Section 4.

MATERIALS AND METHODS

Background: The layout of an integrated circuit must not only satisfy the geometrical and routability requirements, but must also meet the timing constraints. For several years more attention had been paid to the delays of logical elements. Today, we give more attention to the delays introduced by the interconnections between logic elements of a design. The timing optimization tools must then make a good estimation of the delays of the circuit in order to satisfy setup and hold timing constraints.

In order to check the timing convergence through the design development cycle many techniques are used, each has its advantages and disadvantages. One of the widely used techniques is the Static Timing Analysis (STA). Indeed, the STA is a technique that permits to estimate and approximate the timing of a design and quickly propagate the required times and arrival times to the pins of the different cells and logic gates, identifies timing violations and validates if the circuit may or not operate at the nominal frequency defined during the "Specifications" step of the design flow without any timing violation and under all possible conditions (modes and corners). This technique is fast and exhaustive, it allows to analyze all the critical paths of a design and has also a high processing capacity of the chip, since, it does not require any stimulus vector. The STA is called at each stage of the design flow, especially during the physical implementation phase. The Physical Design (PD) or Place and Route (PnR) is a step that comes after the synthesis and before the fabrication steps, it takes as input a gate-level netlist and deliver a GDSII file to fabs for mask fabrication. Its sub-steps consist of floor planning, Pre-CTS, Clock Tree Synthesis (CTS), post-CTS, routing and post-routing.

The timing in the PnR flow can be optimized in pre-CTS, post-CTS and post-routing. During the pre-CTS, the timing is fixed by making changes in the design logic, this optimization mainly aims to reduce setup violations. During the PostCTS the timing is improved by using the various methodologies of "Buffer insertion", "Buffer sizing" and "Wire sizing" to reduce hold violations. Finally, during the post-routing step more optimization passes are performed to eliminate residual timing violations and electrical Design Rule Constraints (eDRC) errors.

But as the technology advances, the total Resistance (R) and Capacitance (C) of interconnect have increased significantly due to the decreasing width and spacing between the wires and have become a dominant factor affecting the design performance (delay and power consumption). And hence, the need for additional innovative timing optimization techniques that takes into consideration the opportunities offered by the new enabled nodes to tackle their challenges.

Before presenting the "Layer optimization" solution, which is a new solution to face the timing problems of 7 nm designs, it is necessary to recall the wire delay model. Net delay (or wire delay) is the amount of time it takes for a signal to propagate from the output of a gate to the input of the next gate and it is proportional to the conductor (wire) parasitic resistance and capacitance and to the input capacitance of the receiver gate (Eq. 1):

$$T = Rwire*(\frac{Cwire}{2}+CL) \quad (1)$$

To estimate the interconnect delay, there are many models which are based on interconnect equivalent resistance and capacitance, the resistance is deduced from the interconnect traces in various metal layers and vias in the design's physical implementation, the capacitance is also extracted from the metal traces and is comprised of the coupling capacitance with the ground as well as the coupling capacitance with neighboring routes (Bhasker and Chadha, 2009). The RC interconnect can be represented by various simplified models which are described in the subsections as.

Assuming that N Nodes of a network is initially discharged to GND and that a step input is applied at node s which is the diver on the N Nodes at time t = 0. The elmore delay at eac node I of the network is then given by the Eq. 2:

$$\tau_{Di} = \sum_{k=1}^{N} C_k R_{ik} \quad (2)$$

This expression is equivalent at the first moment of the impulse response of the network and represents a simple approximation of the actual delay between the source node and node i. It is true that it is more accurate to calculate the delay using complex timing models such as Non-Linear Delay Model (NLDM) and Composite Current Source (CCS) but this approximation can still be reasonable and acceptable. It provides a quick estimate of the delay of a complex network.

The resistance of a wire is proportional to its length L and inversely proportional to its crosssection A. The resistance of a rectangular conductor can be expressed by Eq. 3 (Rabaey *et al*., 2003):

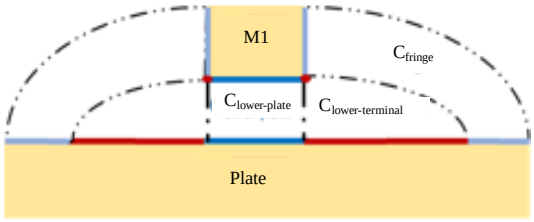

Fig. 1: Different parasitic capacitances seen in circuit interconnect

$$R = \frac{\rho L}{A} = \frac{\rho L}{HW} \quad (3)$$

Where:

• = The resistivity of the material (• -m)
H = A techno constant
L and W = Theheight, the length and the width of the wire

The net delay second parameter is the wire capacitance (Fig. 1) which has three components:

- Plate capacitance: between two parallel metal surfaces (Rabaey *et al*., 2003)
- Fringe capacitance: from the sidewall of the wire to another perpendicular surface, e.g., the ground plate (Rabaey *et al*., 2003)
- Terminal capacitance: from the corner of the wire to other metal surfaces (Rabaey *et al*., 2003)

In the next section, we will present an innovative approach to reduce wire delay, it is a new routing technique that aims to optimize timing, by modifying only the routing topology. It uses the layer resistance differences between SADP and No-SADP layers to distinguish the routing of critical and non-critical.

RESULTS AND DISCUSSION

Implementation: Traditional interconnect delay optimization techniques are no longer as effective as they use to be in old technologies, due to the shrinking in node size, the growing area and power dissipation, the increasing complexity and speed and the advancement in fabrication material properties.

In 7 nm technology, the spectrum of metal layer resistivity is large from the top (Non-SADP) less resistive layers to the bottom (SADP) layers (M0-M3). As shown in Table 1, we can notice that if we perform routing using metal 4 instead of metal 2, we will reduce 54% of the sheet resistance and so on for upper layers.

Our proposed optimization technique uses this propriety as an advantage to optimize wire delay by performing a nets re-routing on No-SADP layers using Non-Default-Rules (NDRs).

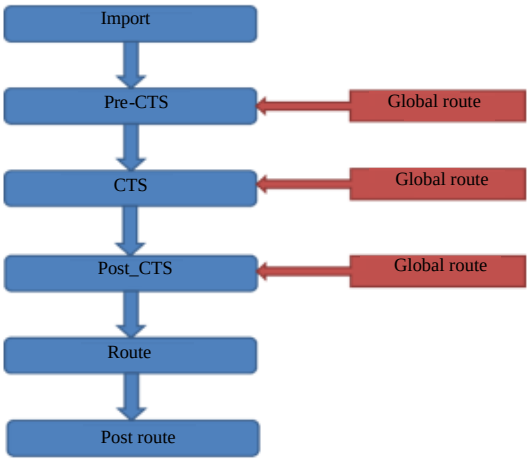

Fig. 2: PnR traditional flow and route global calls

Table 1: Layers sheet resistance in 7 nm technology

Layers	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
R (W/sq)	2.7	2.3	1.2	0.9	0.9	0.9	0.9	0.3	0.3	0.3

To implement our solution, we have developed a TCL script and executed it using Nitro-SoC (Mentor Graphics PnR tool). The TCL script solution can be applied at each global routing stage, since, the global routing is performed many times in the PnR flow as we can see in Fig. 2.

New solution script description: Our improved timing optimization solution is a wrapper that drives the Nitro-SoC (Nitro-SoC™ and Olympus-SoC ™user's Manual, 2017; Nitro-SoC™ and Olympus-SoC™ advanced design flows guide, Software Version 2017; Nitro-SoC™ and Olympus-SoC™ Software Version 2017.1.R2, 2017.) global router to perform an SADP timing driven routing and to generate a routing topology that prioritizes critical timing nets by routing them on upper non-SADP layers and by applying on them additional spacing and width to reduce their overall resistance and capacitance and consequently, reducing their timing component. The proposed wrap proceeds as follows:

- Create NDRs to drive the global route
- Identify the best net targets for optimization
  - Collecting at a first all nets that violated setup timing nets
- Filter nets with a delay greater than a delay threshold
- Apply the non-default-rule that governs the routing remove old global routing and perform a new one on the target nets list
- Measure gain
  - If the gain is negative, revert changes
  - If the gain is positive, keep changes

Table 2: NDR impact on design timing

NDR	WNS (Ps)	TNS (Ps)	WHS (Ps)	THS (Ps)
Upper 4 double space	-61.8	-10942.3	-104.7	-14724.3
Upper 5 double space	-38.2	-6579.8	-104.8	-14187.3
Upper 6 double space	-50.1	-12125.0	-105.1	-14171.0
Upper 7 double space	-79.7	-20780.5	-104.7	-14669.4

Table 3: Net rerouting results

Net	U_{27}-U_{39}	U_{39}-U_{44}	U_{44}-U_{45}	U_{45}-c_{319}	c_{319}-c_{23}	c_{23}-U_{161}	U_{161}-U_{228}	U_{228}-U_{181}
Reference (Ps)	0.5	0	4.8	1.1	4.4	2.2	0.5	7.1
Solution (Ps)	0.5	0	1.8	1.1	1.5	3.1	0.5	1.6
Difference (%)	0	0	63	0	66	-41	0	77

Code implementation: It is very important to choose the right NDR to apply, so, we can see the real impact of rerouting on wire delays. To do this, we have performed an automatic optimization with a transformation that is already used by Nitro-SoC (Bhattacharya *et al*., 2016; Chen and Liu, 2007; Chen *et al*., 2017) then, we have tested different types of NDRs. Table 2 summarizes the results of the NDRs that gives logical improvement and which are neither too restrictive nor too relaxed from routing resources point of view.

As presented in Table 2, the best NDR is upper 5 double space, it gives better results in term of setup and hold timing compared to other NDRs. Upper 5 double space application means that the rerouting will be performed from metal 5 and up and with double spacing between wires.

In the first results observation, we have noticed that the delays of some nets became bigger after the script optimization. The problem is that the rerouting with NDR is performed on nets that were already routed by upper layers (M4-M9), so when we apply the script optimization, the routing on these nets is modified and increase the total wire length which introduces timing losses.

As we can see in Table 3, the net C23-U161 delay had increased from 2.2-3.1 ps after layer optimization. As a remedy of this problem, new nets filtering criteria should be added, to choose only the nets with high improvement probability.

As a solution, we implemented a filtering mechanism that selects only the nets that were originally routed by a certain percentage of SADP layers (Metals from M1-M3) in order to do not target nets that had already an optimized delay value (already routed with non-SADP). The next step is to define the threshold value to use, this threshold value represents the percentage of the net wires routed in Non-SADP divided by the total wirelength. During our tests, we have tried to apply several thresholds to deduct the best targets list.

As we can see in Table 4 results, the best non-SADP (M4-M9) threshold is 70%. It means that the rerouting will be performed only on nets that were already routed with more than 30% with SADP layers.

Table 4: Non-SADP threshold impact on design timing (Units)

	Timing impact			
Threshold (%)	WNS (Ps)	TNS (ps)	WHS (ps)	THS (ps)
30	-75.4	-33374.1	-104.7	-14227.3
50	-77.7	-12476.3	-104.7	-14150.1
70	-47.5	-11824.1	-104.8	-14436.6
80	-54.9	-13221.7	-104.9	-14366.4

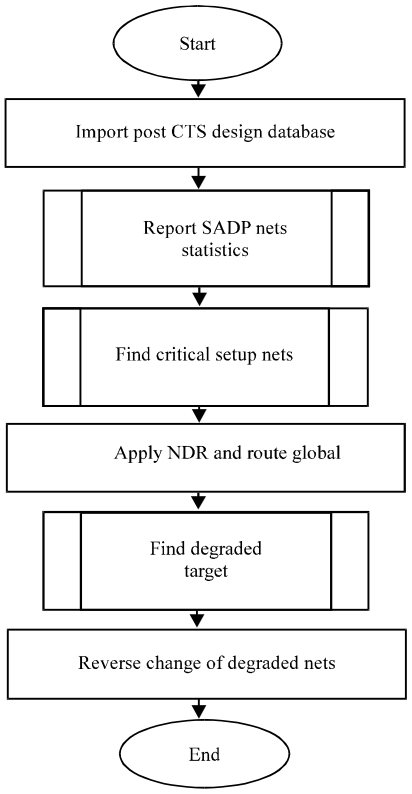

Fig. 3: Final net rerouting solution

Final code diagram flow: After defining all the parameters needed, we implemented our rerouting solution as shown in the diagram represented by Fig. 3. The algorithm starts by loading the design post CTS data base, then it calculates the SADP profiling for each data net in the design, the outcome of this step is a list of nets characterized by their non-SDAP wire percentage. After characterizing all nets a new pass of timing extraction is made to find the critical timing nets. Both data are combined to filter the good targets nets which we apply the NDR upon in the subsequent step.

After re-routing the filtered targets a new pass of timing extraction is needed to estimate the changes impact on timing and to evaluate the hold impact on the design, for the nets that did not benefit from the change or that shows hold timing degradation, they are reverted to their original routing topology.

Table 5: New approach gain on different design metrics

Parameters	WNS (ps)	TNS (ps)	WHS (ps)	THS (ps)	Util (%)	Buff/Inv
Baseline	-103	-35461.2	-33.5	-1926.2	62.12	85748
New algo	-66.8	-26006.7	-27.5	-1843.6	61.95	85063
Gain (%)	35	27	18	4	0.002	1

Experimental results: To view the final results and to compare timing parameters before and after adding the TCL solution to our 7 nm design after the post CTS stage, we first run the routing script then launch all timing reports to see the solution's gain. After applying the TCL script on the post CTS design database, we compared the results after the routing step as shown in Table 5. We note on the table a timing optimization gain of 35% for the WNS as well as an important optimization gain of 27% in the TNS without making any degradation in the hold time (WHS, THS). The utilization of the chip has also been improved with a gain of 0.027%, since, less additional buffers and inverters are required to optimize the timing. The area of the standard cells has been improved by 0.68% for the same reason.

CONCLUSION

Timing closure is a major challenge in physical design. In this study, we have proposed a complete solution for the reduction of interconnect delay in post-CTS stage with a gain of 35.5% in WNS, 26.66 % in the TNS, 4.29% in the THS and 17.91% in the WHS. The proposed algorithm has helped also in area and power reduction by reducing the number of buffers and inverters needed to optimize the timing violation.

The compliance of the results suggests a promising future for VLSI design because even if the commercial IC fab industry continues reducing device features to the atomic-scale, high performances still could be reached. At the end, interconnect delays are still an open field of research that requires more optimization and adjustment.

ACKNOWLEDGEMENTS

This research was supported by Mentor Graphics a Siemens Business. We thank our colleagues from the IC Design Solutions (ICDS) Division who provided insight and expertise that greatly assisted this research. We thank Dr. Hazem El Tahawy (Mentor Graphics, Managing Director MENA Region) for initiating and supporting this work. From the place-and-route solutions group in ICDS division, we thank David Chinnery (Architect, Optimization), Sarvesh Bhardwaj (Group Architect, Optimization) and Nikitin Nikita (Member of consulting staff, ICDS R&D CTS) for assistance, help and guidance through this research.

REFERENCES

Alaghi, A., W.T.J. Chan, J.P. Hayes, A.B. Kahng and J. Li, 2017. Trading accuracy for energy in stochastic circuit design. ACM. J. Emerging Technol. Comput. Syst., 13: 1-30.

Alpert, C. and A. Devgan, 1997. Wire segmenting for improved buffer insertion. Proceedings of the 34th Annual Conference on Design Automation, June 09-13, 1997, ACM, Anaheim, California, USA., ISBN:0-89791-920-3, pp: 588-593.

Bhasker, J. and R. Chadha, 2009. Static Timing Analysis for Nanometer Designs: A Practical Approach. Springer, New York, USA., ISBN:978-0-387-93819-6, Pages: 572.

Bhattacharya, S., D. Das and H. Rahaman, 2016. Delay minimization of multilayer graphene nanoribbon based interconnect using wire sizing method. Proceedings of the 2016 International Conference on Microelectronics, Computing and Communications (MicroCom'16), January 23-25, 2016, IEEE, Durgapur, India, ISBN: 978-1-4673-6622-9, pp: 1-6.

Chen, S. and X. Liu, 2007. A low-latency and low-power hybrid insertion methodology for global interconnects in VDSM designs. Proceedings of the 1st International Symposium on Networks-on-Chip (NOCS'07), May 7-9, 2007, IEEE, Princeton, New Jersey, USA., pp: 75-82.

Chen, X., X. Huang, Y. Xiang, D. Zhang and R. Ranjan *et al*., 2017. A CPS framework based perturbation constrained buffer planning approach in VLSI design. J. Parallel Distrib. Comput., 103: 3-10.

Karimi, G. and A. Ahmadi, 2014. Buffer insertion for delay minimization using an improved PSO algorithm. Appl. Math. Inf. Sci., 8: 2277-2285.

Liu, L., Y. Zhou and S. Hu, 2014. Buffering single-walled carbon nanotubes bundle interconnects for timing optimization. Proceedings of the 2014 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), July 9-11, 2014, IEEE, Tampa, Florida, ISBN: 978-1-4799-3765-3, pp: 362-367.

Murgai, R., 2015. Technology-dependent logic optimization. Proc. IEEE., 103: 2004-2020.

Rabaey, J.M., A.P. Chandrakasan and B. Nikolic, 2003. Digital Integrated Circuits: A Design Perspective. 2nd Edn., Pearson, New York, USA., Pages: 761.

Sarfati, E., B. Frankel, Y. Birk and S. Wimer, 2017. Optimal VLSI delay tuning by space tapering with clock-tree application. IEEE. Trans. Circuits Syst. I Regul. Pap., 64: 2160-2170.

Zhang, H., M.D. Wong, K.Y. Chao and L. Deng, 2012. A practical low-power nonregular interconnect design with manufacturing for design approach. IEEE. J. Emerging Sel. Top. Circuits Syst., 2: 322-332.