Challenges on DTCO Methodology Towards Deep Submicron Interconnect Technology

Heechun Park[1], Kyungjoon Chang[2], Jooyeon Jeong[2], Jaehoon Ahn[2], Ki-Seok Chung[3] and Taewhan Kim[2]
[1]Education and Research program for Future ICT Pioneers, Seoul National University
[2]Department of Electrical and Computer Engineering, Seoul National University
[3]Department of Electronic Engineering, Hanyang University

2021 18th International SoC Design Conference (ISOCC) | 978-1-6654-0174-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISOCC53507.2021.9614026

***Abstract*—Design-technology-co-optimization (DTCO) is essential in deep submicron technologies (e.g., 14nm and below) to co-optimize process technology and design rules and obtain more benefit from advanced node. As the process technology shrinks to deep submicron, the importance of back-end-of-line (BEOL) interconnect in a full chip design drastically grows since its less-than-micron width brings unexpected critical design rules that requires novel design techniques. In this paper, we provide a comprehensive survey on recent challenging issues and cutting-edge design methodologies for DTCO in deep submicron interconnect technology, which includes: offset assignment for pin accessibility; monolithic 3D integration; middle-of-line (MOL) utilization for routing; BEOL-aware representative critical path circuit synthesis; and buried power rail (BPR).**

I. Introduction

Design-technology-co-optimization (DTCO) is the concept to co-optimize the development of process technology and design rules. It is widely adopted by advanced semiconductor companies to overcome the limitation of sequential development process. As the technology node downscales towards deep submicron region, however, DTCO methodology is facing new challenges on different design issues. Especially when the width of back-end-of-line (BEOL) interconnects shrink down to the scale of nano-meters, extra unexpected but critical design rules are arisen and it becomes impossible to apply existing design solutions used for previous nodes and evaluate process technology. In this paper, we provide a comprehensive survey on several new challenging issues and also introduce future directions that enables DTCO for deep submicron interconnect technology.

Precisely, we cover the following issues and directions:

(1) Differences on contacted poly pitch (CPP) in standard cells and metal pitch (MP) in BEOL metal layers incur pin accessibility problems of standard cells;

(2) Shifting the 3D IC interconnect type from through-silicon-via (TSV) to nanoscale monolithic inter-tier via (MIV) for monolithic 3D (M3D) design enables significantly denser vertical integration;

(3) To cope with complex design rules in chip routing stages, novel routing techniques to make good use of middle-of-line (MOL) layers (e.g., AIL, GIL) for chip routing, i.e., CB over RX, MOL routing, are introduced;

(4) As BEOL interconnect variation occupies more proportion on the post-Si performance of critical paths in advanced nodes, BEOL-aware representative critical path (RCP) circuit synthesis is proposed to predict chip performance more accurately including BEOL interconnects;

(5) For sub-5nm technology node, the concept of buried power rail (BPR) is introduced to bury power rails within Si substrate and free up routing tracks, which in turn achieves standard cell height reduction and area scaling.

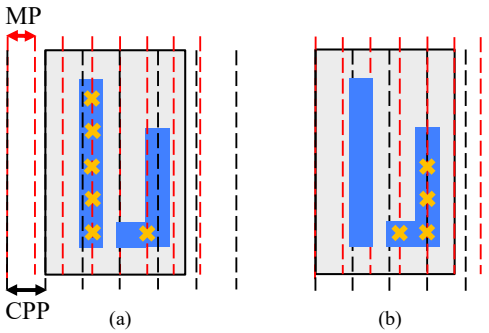

Fig. 1. (a) An example of pin access points in a standard cell. (b) Pin access points changed when the cell in (a) is moved left for 1 CPP.

II. Pin Accessibility by Pitch Difference

Conventionally, contacted poly pitch (CPP) is an integer multiple of the vertical minimum metal pitch (MP), so that metals could be aligned right on the transistors using on-track vias. However, in the era of deep submicron technology, CPP and minimum MP are shrunk in different ratios and this rule of integer of multiple is broken, which results in the contacted poly not align to the BEOL metal layer. When there is an offset between CPP and MP and not considered during standard cell placement, cell pins are not well aligned to the metal tracks. Inserting an off-track via [1] for each misalignment consumes more routing resources, and also loses design quality in terms of electrical features [2],

Fig. 1 shows an example of pin accessibility changes in a cell with two pins. Each yellow point represents a pin access point, which is an intersection spot of a vertical metal track and the cell pin polygon. Before the cell movement (Fig. 1(a), the left pin has five access points and the right pin has only one access points. On the other hand, when this cell is moved to the left for one CPP (Fig. 1(b)), the left pin has no access point at all as it does not intersect with vertical metal stack,

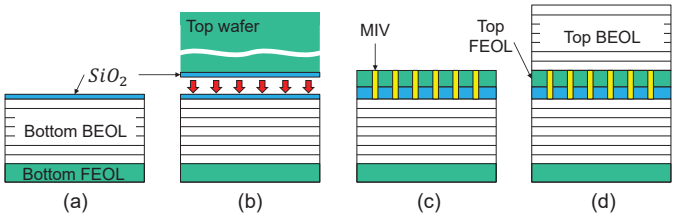

Fig. 2. Illustration of M3D technology.

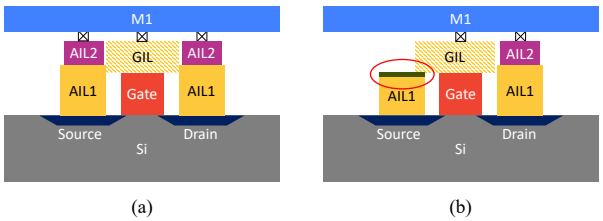

Fig. 3. (a) Cross section view of cell routing without SAC. (b) Cross section view of cell routing with SAC.

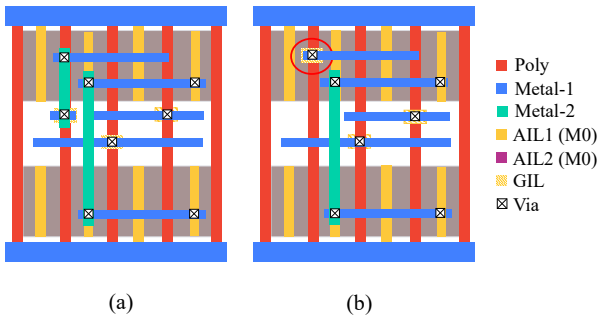

Fig. 4. (a) Cell without CB over RX. (b) Cell with CB over RX.

while the right pin now has three more access points than before, which means it became much easier to be accessed.

As shown in the example, pin accessibility is profoundly affected by the difference between CPP and minimum MP. Especially in the incremental placement stage, CPP, which is the unit of cell movement, is not an integer multiple of metal track pitch, and thus cell relocation causes a huge changes on pin accessibility. In the future research, the issue of CPP-MP difference should be considered thoroughly to accurately measure pin accessibility and improve design routability.

III. Monolithic 3D Integration

The concept of monolithic 3D (M3D) integrated circuit (IC) [3] is emerged to solve the limitation of through-silicon-via (TSV) based 3D IC by using nanoscale monolithic inter-tier via (MIV) to enable high density vertical integration. A MIV is two order of magnitudes smaller than a TSV, and is more like a normal BEOL signal via. This enables fine-grained gate-level 3D implementation that makes the best use of 3D structure with sufficient amount of MIVs. Experiments in [4] showed that the M3D IC with 50nm MIVs outperforms TSV-based 3D IC designs with 2μm and 5μm TSVs in PPA using commercial RTLs and process design kits (PDKs).

Fig. 2 briefly illustrates the M3D technology [5], which is a sequential fabrication from bottom to top tier. After the bottom tier is fabricated as a common wafer (Fig. 2(a)), a sliced empty wafer is laid on and bonded to the BEOL of the bottom wafer (Fig. 2(b)). The top surface is then planarized (e.g., chemical mechanical polishing (CMP)), and MIVs are created using lithography (Fig. 2(c)). Then the top wafer is fabricated sequentially to finalize the M3D fabrication (Fig. 2(d)).

The main challenge of M3D fabrication process is the thermal limitation of top wafer fabrication. Top wafer processes should be performed under low temperature (i.e., room temperature for bonding, $\sim$ 400°C for the following steps) to maintain the bottom wafer integrity, while the fabrication of sub-14nm FinFET and deep submicron interconnects requires much higher temperature (e.g., $>$ 1100°C). There are various efforts [6], [7] and prototypes [8] to realize the M3D IC fabrication, which are guiding us towards a bright future. Although the commercial electronic design automation (EDA) tool for M3D design is also not yet available, an alternative *pseudo-3D* approach, which utilizes commercial 2D EDA tool for M3D design, provides commercial quality RTL-to-GDS design flow for M3D ICs. Several pseudo-3D design flows have been proposed for different design requirements, which are summarized in [9], [10].

IV. Utilization of Middle-of-Line Layers

Middle-of-line (MOL) layers are located between the bottom-most BEOL metal layers and the Si substrate (i.e., source, drain). Two main MOL layers in recent PDKs are active interconnect layer (AIL) and gate interconnect layer (GIL). In this section, we introduce two routing techniques to utilize MOL layers, which are: CB over RX; and intensive use of MOL.

Firstly, CB over RX means 'GIL over active region'. Commonly, GIL is placed only on the non-active region, as GIL above active area causes current short between source/drain and gate through GIL and AIL as shown in Fig. 3(a). Recently, a fabrication technique to avoid short between gate and AIL called self-aligned diffusion contact (SAC) is introduced [11]. It disconnects gate and AIL by recessing the gate metal and then depositing a silicon nitride etch-stop on the top of the gate metal. Similar technique which is red circle is applied to avoid short between GIL and AIL as shown in Fig. 3(b).

As GIL over active region, i.e., CB over RX, is available by SAC, routing for standard cell requires less routing resource from higher BEOL metal, which incurs more routing resource for chip routing. In Fig. 4(a), cell routing without CB over RX requires M2 layer to connect M1 metal and gate poly through GIL on non-active region. In contrary, CB over RX removes this connection by directly connect M1 and gate poly using GIL on the active region, as in Fig. 4(b).

Another technique is the intensive use of MOL layer, which increase the usage of AIL/GIL layers during cell routing to reserve more spaces on BEOL layers (e.g., M1, M2 layers) for later chip routing. Fig. 5 shows an example of differences between with and without intensive MOL usage. Compared to Fig. 5(a) that used six M2 tracks for cell routing, we were able to use only four M2 tracks to finish cell routing thanks

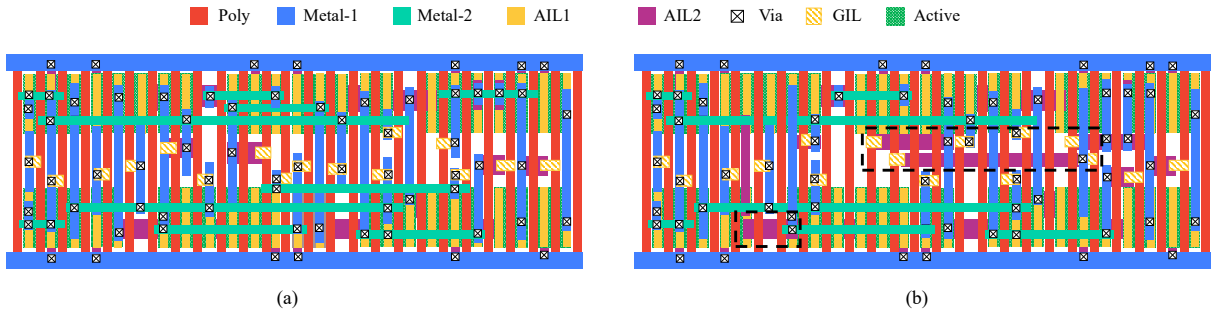

Fig. 5. Comparison of cell routing with (a) common routing using M1/M2 and (b) intensively using MOL layers.

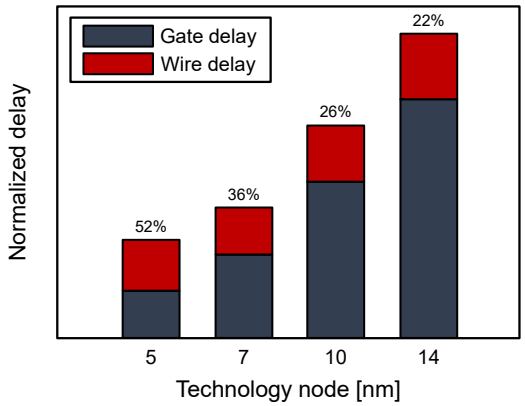

Fig. 6. Transistor and interconnect delay of an inverter driving 150 units of contacted gate pitch of wire length and fan-out of 3 [14].

to the intensive use of MOL layer, marked in the dotted box of Fig. 5(b).

V. BEOL-AWARE PERFORMANCE MONITORING

As transistors and interconnects scale down, process variation grows larger and the affected chips are operating on wide range of performance. Conventionally, industries apply voltage binning [12] to make slow chips usable. They classify the manufactured chips according to their performance, and assign higher supply voltage to the chips that are rated as 'slow'. To measure the performance of each chip and decide which bin to be placed, a representative critical path (RCP) circuit is installed on the target chip, and the delay of RCP circuit represents the overall post-Si chip performance. A widely acceptable candidate of RCP circuit is the critical path replica (CPR) [13], which is a duplicated circuit of the slowest timing path at nominal parameter values.

However, when the process scales down to the deep submicron region, CPR no longer represents the critical path delay as the unpredictable process variation becomes too large and CPR delay does not match to the actual critical path delay of the chip. More importantly, as in Fig. 6, the BEOL delay occupies more than 20% of the critical path delay starting from 14nm technology, and finally becomes as significant as that of FEOL at 5nm technology. The up-to-date CPR synthesis method, which is based on iterative gate sizing to fine tune

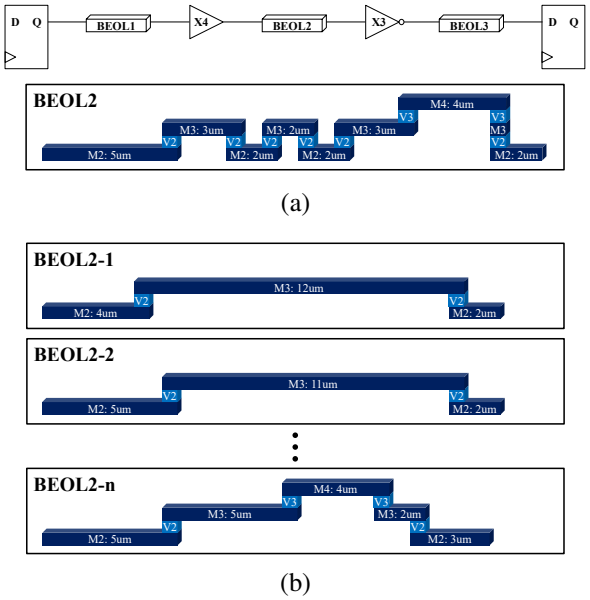

Fig. 7. (a) An example of critical path with 3 BEOLs and the routing details of BEOL2. (b) An example of BEOL reconfigurations for BEOL2 [15].

the nominal critical path circuit, only take the FEOL gates into account, and thus shows low accuracy when predicting the critical path delay in deep submicron environment.

Recently, Han and Kim [15] proposed a BEOL-aware methodology of synthesizing RCP circuit, which iteratively tunes routing patterns as well as FEOL gate sizing. While its gate sizing follows the same method in [14], it reconfigures the BEOL structure as well, i.e., metal length and via usage, to generate multiple candidates (Fig. 7(b)) for each BEOL in the critical path (Fig. 7(a)). Then, Monte Carlo simulation is applied on all RCP circuit candidates to find best correlated RCP circuit under statistical BEOL random variables. From the experiments in [15] with industrial benchmarks, RCP from BEOL-aware synthesis method is shown to reduce the prediction error by 54% and 19% on average over that of conventional CPR and gate sizing based CPR, respectively.

VI. BURIED POWER RAILS

In the advanced technology nodes, downscaling the standard cells with pitch scaling or fin depopulation is almost at its

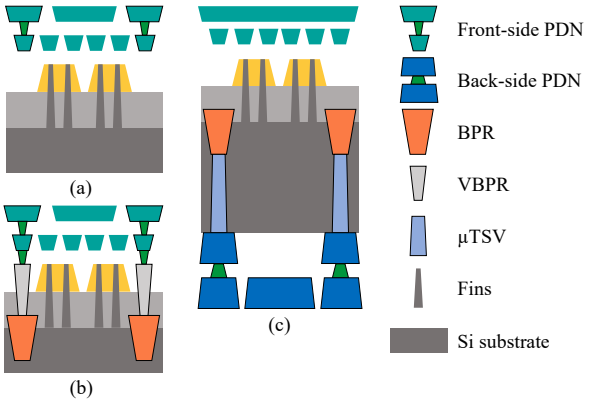

Fig. 8. Examples of types of PDN. (a) Conventional front-side PDN without BPR. (b) Front-side PDN with BPR. (c) Back-side PDN with BPR.

limit. The novel concept of buried power rail (BPR) is then introduced to achieve further area scaling by standard cell height reduction [16]. The main idea of BPR is to bury power rails (VDD, VSS) into the Si substrate, so as to reduce standard cell height by around two tracks without any decrease of drive strength.

Fig. 8 illustrates and compares three different types of PDN. Conventional PDN (Fig. 8(a)) uses BEOL metal layers to delivering power and also uses BEOL track to feed to logic gates. Front-side PDN with BPR (Fig. 8(b)) also uses BEOL metal layers to delivering power, while it uses BPR to supply power to logic gates. A new type of via called VBPR is introduced to connect BPR and the Front-side PDN. Lastly, Back-side PDN with BPR [17] (Fig. 8(c)) uses metals on the opposite side of the chip even to deliver power. It is then connected to BPR using micro through-silicon-via(μTSV), which supplies power to logic gates [18].

BPR is expected to provide design advantages such as area scaling by removing BEOL power rails [19], extra routing space in M0 layer, and improved IR drop [17]. Still, there are more considerations for realizing BPR-based chip design. Front-side PDN with BPR (Fig. 8(b)) has a main issue of power tap cell frequency and back-side PDN with BPR (Fig. 8(c)) should consider μTSV pitch and BPR aspect ratio, both of which make BPR adoption much more complex than the conventional PDN.

VII. Conclusion

DTCO is a prominent methodology in deep submicron technology to achieve design improvements by utilizing the nanoscale features. Nevertheless, downscaling the process node also incurs various new design issues especially on the interconnects, which are not taken into account in the past node. In this paper, we introduced new challenges for DTCO under deep submicron interconnect technology. We believe that our survey on these rising issues and the latest solutions will guide us to a correct direction towards the application of DTCO in deep submicron technology.

Acknowledgement

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MEST) under Grant 2020R1A4A4079177, in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1I1A1A01049399), and in part by the BK21 Four Program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2021.

References

[1] M. Ahrens *et al.*, "Detailed Routing Algorithms for Advanced Technology Nodes," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 4, 2015, pp. 563–576.
[2] K. Y. Lee and T. C. Wang, "Post-routing Redundant Via Insertion for Yield/Reliability Improvement," in *Asia and South Pacific Conference on Design Automation*, 2006, pp. 303–308.
[3] S. Wong *et al.*, "Monolithic 3D Integrated Circuits," in *International Symposium on VLSI Technology, Systems and Applications*, 2007, pp. 1–4.
[4] S. K. Samal *et al.*, "Monolithic 3D IC vs. TSV-based 3D IC in 14nm FinFET technology," in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference*, 2016, pp. 1–2.
[5] P. Batude *et al.*, "3-D Sequential Integration: A Key Enabling Technology for Heterogeneous Co-Integration of New Function With CMOS," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, no. 4, pp. 714–722, 2012.
[6] M. M. Shulaker *et al.*, "Monolithic 3D integration of logic and memory: Carbon nanotube FETs, resistive RAM, and silicon FETs," in *IEEE International Electron Devices Meeting*, 2014, pp. 27.4.1–27.4.4.
[7] M. M. Shulaker *et al.*, "Three-Dimensional Integration of Nanotechnologies for Computing and Data Storage on a Single Chip," *Nature*, vol. 547, no. 8, pp. 74–78, 2017.
[8] DARPA ERI Summit 2019. https://eri-summit.darpa.mil/2019-archive-keynote-slides.
[9] H. Park *et al.*, "Pseudo-3D Approaches for Commercial-Grade RTL-to-GDS Tool Flow Targeting Monolithic 3D ICs," in *International Symposium on Physical Design*, New York, NY, USA, 2020, p. 47–54.
[10] H. Park *et al.*, "Pseudo-3D Physical Design Flow for Monolithic 3D ICs: Comparisons and Enhancements," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 26, no. 5, Jun. 2021.
[11] C. Auth *et al.*, "A 10nm High Performance and Low-Power CMOS Technology Featuring 3rd Generation FinFET Transistors, Self-Aligned Quad Patterning, Contact over Active Gate and Cobalt Local Interconnects," in *IEEE International Electron Devices Meeting*, 2017, pp. 29–1.
[12] V. Zolotov *et al.*, "Voltage Binning under Process Variation," in *International Conference on Computer-Aided Design*, 2009, p. 425–432.
[13] Q. Liu and S. S. Sapatnekar, "Capturing Post-Silicon Variations Using a Representative Critical Path," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, no. 2, pp. 211–222, 2010.
[14] T. Huynh-Bao *et al.*, "Statistical Timing Analysis Considering Device and Interconnect Variability for BEOL Requirements in the 5-nm Node and Beyond," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 25, no. 5, pp. 1669–1680, 2017.
[15] C. Han and T. Kim, "Synthesis of Representative Critical Path Circuits Considering BEOL Variations for Deep Sub-micron Circuits," *Integration*, vol. 78, pp. 1–10, 2021.
[16] J. Ryckaert *et al.*, "Extending the Roadmap Beyond 3nm Through System Scaling Boosters: A Case Study on Buried Power Rail and Backside Power Delivery," in *Electron Devices Technology and Manufacturing Conference*, 2019, pp. 50–52.
[17] D. Prasad *et al.*, "Buried Power Rails and Back-side Power Grids: Arm® CPU Power Delivery Network Design Beyond 5nm," in *IEEE International Electron Devices Meeting*, 2019, pp. 19.1.1–19.1.4.
[18] M. O. Hossen *et al.*, "Power Delivery Network (PDN) Modeling for Backside-PDN Configurations With Buried Power Rails and μ TSVs," *IEEE Transactions on Electron Devices*, vol. 67, no. 1, pp. 11–17, 2020.
[19] B. Chava *et al.*, "DTCO exploration for efficient standard cell power rails," in *Design-Process-Technology Co-optimization for Manufacturability XII*, vol. 10588, 2018, p. 105880B.