Title
Characterization, modeling and optimization of fills and stress in semiconductor integrated circuits

Author
Topaloglu, Rasit Onur

Publication Date
2008

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Characterization, Modeling and Optimization of Fills and Stress in Semiconductor Integrated Circuits

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in
Computer Science (Computer Engineering)

by

Rasit Onur Topaloglu

Committee in charge:

Professor Andrew B. Kahng, Chair
Professor Peter Asbeck
Professor Chung-Kuan Cheng
Professor Tajana Simunic Rosing
Professor Bang-Sup Song

2008

The dissertation of Rasit Onur Topaloglu is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

2008

To people who devote their lives to science.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my father Halil Topaloglu and mother Bilge Topaloglu for their support and having to bear the absence of their only child at a distance of halfway around the world.

I would like to thank all UCSD CSE and ECE professors that I have interacted with. In particular, I would like to thank Andrew B. Kahng for helping me perfect my CAD, mathematical and research skills; for providing me excellent counseling and state of the art problems to work on; for letting me attend, and present and discuss my research at top conferences and with top researchers; and for letting me leverage my Ph.D. experience while working at Advanced Micro Devices during the last stages of my Ph.D. degree. I would like to thank Alex Orailoglu for letting me build a very strong framework during the initial years of my Ph.D. studies, and having me see the problems that I should improve upon. I would like to thank Peter Asbeck for improving my TCAD skills and device physics knowledge, as well as being a temporary advisor. I would like to thank Bang-Sup Song for significantly improving my understanding of analog circuits and improving my design skills. I would like to thank Chung-Kuan Cheng and Tajana Simunic Rosing for being role models, showing interest in my research and providing me with increased motivation. I would like to thank the former and current VLSI CAD group members for various discussions and feedback (Puneet Gupta - now with UCLA, Kwangok Jeong, Swamy Muddu - now with Nextwave Semiconductor, Chul-Hong Park, Sherief Reda - now with Brown University, Kambiz Samadi, Puneet Sharma - now with Freescale Semiconductor).

Next, I would like to thank my AMD coworkers. I would like to thank my manager Ali Icel for letting me improve my skills, for being a role model for an excellent manager and providing me with flexibility to conduct research along my interests. I would like to thank Judy An (now with Sun Microsystems) for providing me research knowledge in device data analysis and experimentation. I would like to thank Jung-Suk Goo for letting me be more rigorous in my analyses and documentation. I would like to thank Frank Geelhaar for significantly improv-

ing my TCAD knowledge. I would like to thank John Faricelli for leveraging my RC extraction tool usage capabilities. I would like to thank our vice president, Nick Kepler, for supporting my Ph.D. studies. I would like to thank the rest of my group (Zhi-Yuan Wu, Ciby Thuruthiyil, Tran Ly, Martin Radwin, Brian Qiang Chen - now with Synopsys, Niraj Subba, Sushant Suryagandh, Mayank Gupta - no longer with AMD, Vineet Wason and Priyanka Chiney) for various discussions and feedback.

I would also like to thank my professors from Bogazici University. They have helped me build a very strong framework, led me to walk into the wonderful world of research, and even supported me during the down times of my Ph.D. I would like to thank in particular Yorgo Istefanopulos, Kadri Ozcaldiran, Oguzhan Cicekoglu and Gunhan Dundar.

The material in this thesis is based on the following publications.

- Chapter II is based on the following publications:
  - A. B. Kahng and R. O. Topaloglu, "DOE-based extraction of CMP, active and via fill impact on capacitances," *IEEE Trans. on Semiconductor Manufacturing*, 21(1), 2008, pp. 22-32.
  - A. B. Kahng and R. O. Topaloglu, "A DOE set for normalization-based extraction of fill impact on capacitances," **Best Paper Award**, *Proc. IEEE International Symposium on Quality Electronic Design*, 2007, pp. 467-474.
- Chapter III is based on the following publications:
  - A. B. Kahng and R. O. Topaloglu, "Performance-oriented interlayer-aware CMP fill pattern optimization," *under review in IEEE Trans. on Computer-Aided Design*, 2008.
  - A.B. Kahng and R.O. Topaloglu, "Performance-aware CMP fill pattern optimization," *Proc. International VLSI/ULSI Multilevel Interconnection Conference (VMIC)*, **Invited Paper**, 2007, pp. 135-144.

- R. O. Topaloglu, "Energy-minimization model for fill synthesis," *Proc. IEEE International Symposium on Quality Electronic Design*, 2007, pp. 444-451.

- Chapter IV is based on the following publications:

  - A. B. Kahng, P. Sharma and R. O. Topaloglu, "Chip optimization through STI stress-aware placement perturbations and fill insertion," *accepted for publication in IEEE Trans. on Computer-Aided Design*, 2008.
  - A. B. Kahng, P. Sharma and R. O. Topaloglu, "Exploiting STI stress for performance," *Proc. IEEE/ACM International Conference on Computer-Aided Design*, 2007, pp. 83-90.
  - R. O. Topaloglu, "Standard cell and custom circuit optimization using dummy diffusions through STI width stress effect utilization," *Proc. IEEE Custom Integrated Circuits Conference*, 2007, pp. 619-622.

- Chapter V is based on the following publication:

  - A. B. Kahng and R. O. Topaloglu, "A TCAD-based study of fill pattern and via fill impact on low-k dielectric stress," **Invited Paper**, *Proc. International Chemical-Mechanical Planarization for ULSI Multilevel Interconnection Conference (CMP-MIC)*, 2007, pp. 337-346.

My coauthors (Prof. Andrew B. Kahng and Dr. Puneet Sharma) have kindly approved the inclusion of the aforementioned publications in my thesis.

VITA

1979	Born, Istanbul, Turkey
2002	B.S., Electrical and Electronic Engineering, Bogazici University, Istanbul, Turkey
2005	M.S., Computer Science, University of California, San Diego
2006	C.Phil., Computer Science (Computer Engineering), University of California, San Diego
2008	Ph.D., Computer Science (Computer Engineering), University of California, San Diego

SELECTED PUBLICATIONS

1. A. B. Kahng and R. O. Topaloglu, "Performance-oriented interlayer-aware CMP fill pattern optimization," *under review in IEEE Trans. on Computer-Aided Design*, 2008.

2. A. B. Kahng, P. Sharma and R. O. Topaloglu, "Chip optimization through STI stress-aware placement perturbations and fill insertion," *accepted for publication in IEEE Trans. on Computer-Aided Design*, 2008.

3. R. O. Topaloglu, "Process variation characterization and modeling of nanoparticle interconnects for foldable electronics," *Proc. IEEE International Symposium on Quality Electronic Design*, 2008, pp. 498-501.

4. A. B. Kahng and R. O. Topaloglu, "DOE-based extraction of CMP, active and via fill impact on capacitances," *IEEE Trans. on Semiconductor Manufacturing*, 21(1), 2008, pp. 22-32.

5. R. O. Topaloglu, "Via chamfering modeling for improved MIM capacitance silicon correlation," *Proc. International VLSI/ULSI Multilevel Interconnection Conference (VMIC)*, 2007, pp. 362-364.

6. A. B. Kahng and R. O. Topaloglu, “Performance-aware CMP fill pattern optimization,” **Invited Paper**, *Proc. International VLSI/ULSI Multilevel Interconnection Conference (VMIC)*, 2007, pp. 135-144.

7. A. B. Kahng, P. Sharma and R. O. Topaloglu, “Exploiting STI stress for performance,” *Proc. IEEE/ACM International Conference on Computer-Aided Design*, 2007, pp. 83-90.

8. R. O. Topaloglu, “Standard cell and custom circuit optimization using dummy diffusions through STI width stress effect utilization,” *Proc. IEEE Custom Integrated Circuits Conference*, 2007, pp. 619-622.

9. A. B. Kahng and R. O. Topaloglu, “A DOE set for normalization-based extraction of fill impact on capacitances,” **Best Paper Award**, *Proc. IEEE International Symposium on Quality Electronic Design*, 2007, pp. 467-474.

10. R. O. Topaloglu, “Energy-minimization model for fill synthesis,” *Proc. IEEE International Symposium on Quality Electronic Design*, 2007, pp. 444-451.

11. A. B. Kahng and R. O. Topaloglu, “A TCAD-based study of fill pattern and via fill impact on low-k dielectric stress,” **Invited Paper**, *Proc. International Chemical-Mechanical Polishing for ULSI Multilevel Interconnection Conference (CMP-MIC)*, 2007, pp. 337-346.

12. R. O. Topaloglu, “Process variation-aware multiple-fault diagnosis of thermometer-coded current-steering DACs,” *IEEE Trans. on Circuits and Systems-II: Analog and Digital Signal Processing*, 54(2), 2007, pp. 191-195.

13. A. B. Kahng and R. O. Topaloglu, “Interconnect matching design rule inferring and optimization through correlation extraction,” *Proc. IEEE International Conference on Computer Design*, 2006, pp. 222-229.

14. R. O. Topaloglu, “Monte Carlo-alternative probabilistic simulations for analog systems,” *Proc. IEEE International Symposium on Quality Electronic Design*, 2006, pp. 249-253.

15. A. B. Kahng and R. O. Topaloglu, “Generation of design guarantees for interconnect matching”, *Proc. IEEE/ACM System Level Interconnect Prediction Workshop*, 2006, pp. 29-34.

16. R. O. Topaloglu, “Early, accurate and fast yield estimation through Monte Carlo-alternative probabilistic behavioral analog system simulations,” *Proc. IEEE VLSI Test Symposium*, 2006, pp. 136-142.

17. V. Wason, J. X. An, J.-S. Goo, Z.-Y. Wu, Q. Chen, C. Thuruthiyil, R. Topaloglu, P. Chiney and A. Icel, “Statistical compact modeling and Si verification methodology,” **Invited Paper**, *Proc. International Conference on Solid-State and Integrated-Circuit Technology (ICSICT)*, 2006, pp. 1198-1201.

18. E. S. Erdogan, R. O. Topaloglu, O. Cicekoglu, H. Kuntman and A. Morgul, “Novel multiple function analog filter structures and a dual-mode multifunction filter,” *International Journal of Electronics*, 93(9), 2006, pp. 637-650, DOI: 10.1080/00207210600711713. [**Listed as number 5 in 2006 most downloaded articles.**]

19. R. O. Topaloglu and A. Orailoglu, “Forward discrete probability propagation method for device performance characterization under process variations,” *Proc. Asia and South Pacific Design Automation Conference*, 2005, pp. 220-223.

20. R. O. Topaloglu and A. Orailoglu, “A DFT approach for diagnosis and process variation-aware structural test of thermometer coded current steering DACs,” *Proc. IEEE/ACM/EDAC Design Automation Conference*, 2005, pp. 851-856.

21. R. O. Topaloglu and A. Orailoglu, “On mismatch in the deep sub-micron era - from physics to circuits,” *Proc. Asia and South Pacific Design Automation Conference*, 2004, pp. 62-67.

22. E. S. Erdogan, R. O. Topaloglu, O. Cicekoglu and H. Kuntman, “New current-mode special function continuous time active filters employing only OTAs and OPAMPs,” *International Journal of Electronics*, 91(6), 2003, pp. 345-359.

23. R. O. Topaloglu, H. Kuntman and O. Cicekoglu, “Current-input current-output notch and bandpass analog filter structures as alternatives to active-R circuits,” *Frequenz*, 57(5-6), 2003, pp. 123-127.

24. R. O. Topaloglu, H. Kuntman and O. Cicekoglu, “Novel notch and bandpass filter structures using OTAs and OPAMPs,” *Proc. International Conference on Electrical and Electronics Engineering*, 2001, pp. 63-67.

ABSTRACT OF THE DISSERTATION

Characterization, Modeling and Optimization of Fills and Stress in Semiconductor Integrated Circuits

by

Rasit Onur Topaloglu
Doctor of Philosophy in Computer Science (Computer Engineering)
University of California, San Diego, 2008
Professor Andrew B. Kahng, Chair

To improve manufacturability and yield, a number of fill structures are used in semiconductor manufacturing. These structures are CMP, active region (diffusion) and via fills. CMP (dummy) fills are used to reduce metal thickness variations due to chemical-mechanical polishing (CMP). Via fills are used to improve neighboring via printability, to improve mechanical stability of low-k dielectrics, and to reduce via resistance variability. Active region fills are used for STI CMP uniformity and threshold voltage variation reduction. Contact fills may be used for contact printability enhancement and contact resistance variability reduction. In this thesis, we additionally utilize via fills for reliability improvement of low-k and ultra low-k dielectrics, active region fills for stress optimization, and contact fills along with via fills for device temperature reduction.

Although modern parasitic extraction tools accurately handle grounded fills and regular interconnects, such tools use only rough approximations to assess the capacitance impact of floating fills; these approximations include assuming that floating fills are grounded or that each fill is merged with neighboring ones. The accuracy of extractors must be improved without deviating from standard extraction flows. Furthermore, a thorough analysis of floating fill impact is needed by the industry to help select optimal fill shapes and pattern. In Chapter II, we show through 3D field solver simulations that the assumptions used in extractors result in significant inaccuracies. To reduce RC extractor inaccuracies, we provide a design of experiments (DOE) set for accurate extraction of coupling capacitances including interlayer coupling up to second neighboring layer. We provide a compact DOE structure from which multiple coupling capacitances can be obtained in one simulation to reduce the simulation time. We identify the relevant design and process parameters for fill analysis, and provide a thorough analysis of the floating fill impact on capacitances. Furthermore, we compare different fill algorithms and analyze how each parameter affects coupling. We provide this analysis methodology for designers to be able to analyze their own designs and BEOL process tuning. The final aspect of our contribution is a normalization-based integration methodology to improve RC extractor accuracy. We also extend our analyses and methodology to via fills and active region fills, which have more recently been introduced into semiconductor design-manufacturing methodologies and for which sufficient understanding is still lacking.

Classical methods to insert fills focus on metal density uniformity, but do not take into consideration or are unable to minimize the impact of fills on circuit performance. Furthermore, interlayer impact due to fills is not considered during fill synthesis. Fill insertion guidelines exist [1] [2], but automation is needed to apply them for large designs. In Chapter III, we develop a novel fill insertion methodology, starting from a physical analogy, that heuristically minimizes coupling capacitance increase due to fill. Our methodology uses a grid model with bonds for energy minimization during fill insertion. We provide models for the bond energy network to map given fill insertion guidelines into an energy minimization problem. Our optimization methodology can automate the application of known fill insertion guidelines. We extend our methodology to enable critical net-aware and interlayer-aware fill optimization. Furthermore, we provide a power-aware fill methodology for power-critical circuits. Experiments show that the proposed optimization methods can reduce fill impact on coupling capacitances by up to 96.10% for 30% pattern density and up to 15.64% for 60% pattern density cases. We have observed total capacitance reductions in an industrial dual core product testcase with no degradation of topography uniformity, relative to traditional fills.

Starting at the 65nm node, stress engineering to improve performance of transistors has been a major industry focus. An intrinsic stress source – shallow trench isolation (STI) – has not been fully utilized up to now for circuit performance improvement. Furthermore, the introduction of dual-stress liners (DSL) requires analyses and design guidelines for optimization. In Chapter IV, we present a new methodology that combines detailed placement and active-region fill insertion to exploit STI stress for performance improvement. We conduct process simulation of a production 65nm STI technology to develop mobility improvement models for STI width-mediated stress. We then utilize these models to perform STI stress-aware delay analysis of critical paths using a SPICE-based flow. We identify relevant layout parameters and automate the extraction process. We also achieve a timing-driven optimization of STI stress in standard cell designs, using detailed placement perturbation and active-region fill insertion to improve CMOS performance. We furthermore extend our optimization for intra-cell optimization. We assess the proposed analysis and optimization on small designs implemented with a 65nm production cell library and a standard synthesis, place and route flow. Our optimization improves clock frequency by 2.44% to 5.26% for inter-cell optimization and up to 11.32% for intra-cell STI width optimization. We also analyze DSL technology and provide design guidelines for layout optimization for DSL technology. The frequency improvement through exploitation of STI stress comes at practically zero cost in terms of design area and wirelength.

CMP fills and via fills can furthermore alter the local mechanical stresses in the BEOL (back-end of the line), which can improve or degrade reliability. Designers require guidelines for fill insertion for reliability. We provide test structures to analyze the impact of fills on local mechanical stress using a generic 65nm BEOL process flow. Our test structures replicate common layout configurations in the presence of fills. In Chapter V, we conduct TCAD simulations on our test structures and also conduct a process sensitivity analysis. We tie the resultant stress to BEOL reliability concerns through identifying stress measurement locations. We show that the design of fills can reduce the nominal stress or stress gradients by up to 5x, theoretically providing significant improvement to BEOL reliability. We provide design guidelines for BEOL reliability improvement with fills. In particular, we assess whether via fills can be used for BEOL reliability improvement. We

also show that connecting CMP fills to interconnects through via fills can improve BEOL reliability by more than 20%.

Device heating is also an important issue particularly for SOI (silicon on insulator) technology. Local modifications of a layout with fills can alter device temperatures. It is possible to reduce the temperature in a transistor channel through proper layout design style and structures such as via fills and contact fills. In Chapter VI, we evaluate the impact of CMP, via and contact fills on temperature at the transistor level. We provide a TCAD-based analysis through the test structures we have designed. We provide design guidelines to reduce device temperature and hence the channel leakage. We show that device temperature can be reduced by connecting the device diffusion region or interconnects connected to the device to CMP fills through contact and via fills. We have shown that the proposed guidelines can reduce the temperature by 5^{o}C with no penalty and device leakage up to and exceeding an order of magnitude if there is available area and performance flexibility to benefit the proposed guidelines.

I

Introduction

Design for Manufacturability (DFM) efforts have accelerated considerably starting with the $65nm$ technology node. DFM targets design analysis, modeling, optimization and automation to be able to manufacture working chips, and improve chip performance and reliability, in the face of mounting challenges (variability, power, etc.) in semiconductor manufacturing. With each technology generation, new problems arise or known problems start to affect performance and dominate.

Fills and FEOL (front-end of the line) and BEOL (back-end of the line) stress are very important topic areas in DFM. Efficient characterization, modeling and optimization of fills and stress are necessary to help reduce design pessimism and improve suboptimal design performance.

In this chapter, we first provide an introduction to integrated circuit manufacturing and fills. We introduce the DFM structures that we use throughout this thesis. We then describe how fills impact electrical performance and why they should be optimized. We then give an introduction to the thermal impact of fills. Finally, we introduce the phenomenon of stress in integrated circuits. We describe how FEOL stress is utilized, how BEOL stress is handled, and modern requirements for control and optimization of stress in semiconductor circuits.

I.A Integrated-Circuit Manufacturing and Fill

An integrated circuit consists of millions or billions of transistors and interconnects. Its manufacture starts with a layer of semiconductor devices, followed by multiple layers of interconnects, on a silicon wafer. Each semiconductor device is formed on an active silicon region and has a polysilicon or metal gate to control its operation. An active region consists of a doped silicon. The devices and interconnects are separated out by interlayer and intralayer dielectrics. The devices are planarly separated by an isolation layer called *shallow trench isolation* (STI), which usually is an oxide. The devices are connected to interconnects through vertical connections called *contacts.* The interconnects in different layers are connected through *vias.*

DFM structures used to improve manufacturability and yield are illustrated in Figure I.1. These structures are as follows.

- **Poly fills.** Polysilicon fills, or dummy gates, are inserted next to regular polysilicon gates to improve lithographic printability and enhance device current variability.

- **CMP fills.** CMP (chemical-mechanical polishing) fills, or (dummy) fills, are inserted to provide a flat post-CMP topography.

- **Diffusion fills.** Diffusion fills, or active region fills, are used to reduce the STI and diffusion region step height differences. The step height differences would result in threshold voltage variations. The device threshold voltage variations may result due to the STI shifting up (or down) the edge of the polysilicon, thereby altering the polysilicon doping and hence the channel control performance of the gate [3]. In this thesis, we also use diffusion fills for performance improvement through stress engineering.

- **Via fills.** Via fills are used to improve intralayer neighboring via printability. Via fills can exist between two CMP fills in neighboring layers or between a CMP fill in one layer and an interconnect in a neighboring layer. In this thesis, we also use via fills to improve BEOL (back end of the line) reliability.

- **Contact fills.** Contact fills, though rarely used, can be used to improve intralayer neighboring contact printability. In this thesis, we also use them for device temperature reduction.

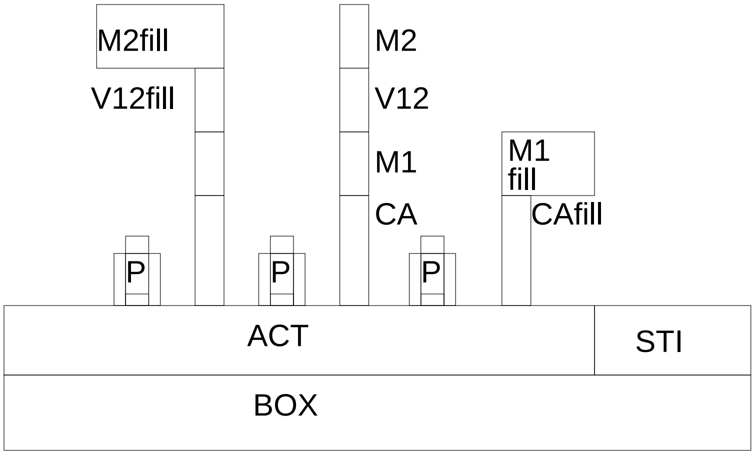


Figure I.1: DFM structures (poly fills, CMP fills, via fills and contact fills) are illustrated. *CA* is contact. *P* is polysilicon gate. *BOX* is built-in oxide for SOI (silicon on insulator) devices. *ACT* is active region.

After STI growth, it is possible that an STI CMP (chemical-mechanical planarization) step is applied to smooth out active region and STI height differences. The CMP process involves both chemical and mechanical elements, involving a pad and a slurry, to remove excess material from a given deposited layer. The CMP process is illustrated in Figure I.2.

A planar layer is needed so that upper layers can also be planar. The CMP process is a function of many process parameters in addition to some layout parameters [4]. These layout parameters include, at least, the density and widths of metal shapes. Design manuals provide ranges for metal density per layer, to which the design must conform. To make the metal density more uniform, CMP fills are inserted.

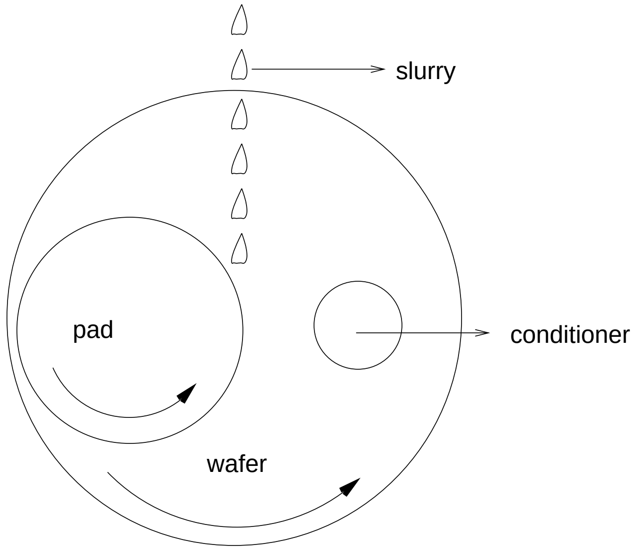


Figure I.2: CMP process.

CMP is applied after contact formation, as well as for each metal layer in a dual-damascane process. However, CMP does not guarantee a smooth layer: its performance is highly dependent on the pattern underneath. Regions with different density and width of metal will have different post-polish heights. The CMP (dummy) fills that are inserted to make metal density more uniform hence lead to a more uniform post-polish layer topography. If the fills were not inserted, the integrated-circuit would encounter CMP-related problems as shown in Figure I.3, which would result in catastrophic or parametric problems. Thus, CMP fills are an essential part of the alyout and/or post-tapeout phases of design. In this thesis, we use CMP fills to also improve timing and power dissipation performance. CMP fills can be connected to supply or ground, or can be left floating. The most common application is to leave them floating. Smaller fill structures are less restrictive to routing, but fills in general increase the capacitances between interconnects in a circuit.

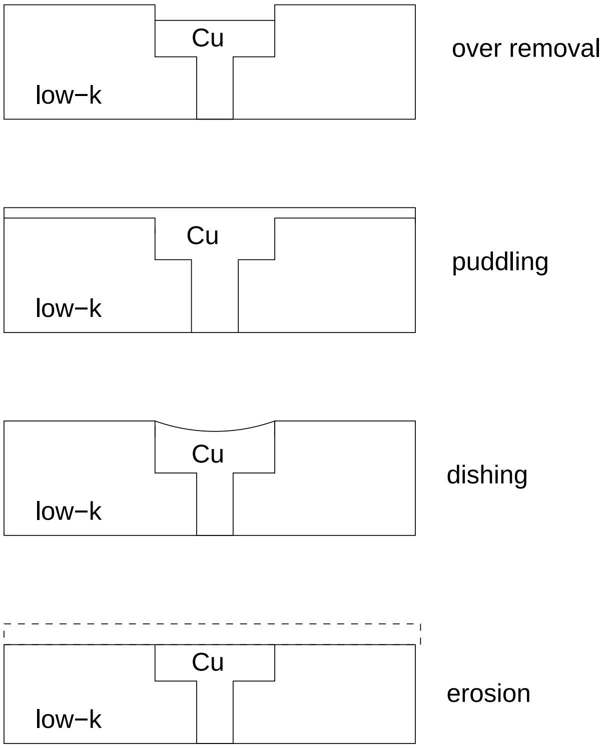


Figure I.3: CMP problems.

I.A.1 Fill Impact on Performance

Ideally, CMP fills should not alter the capacitances of or between interconnects. On the other hand, if the capacitances are going be altered by a non-negligible amount, they should be correctly accounted for in the RC extraction. Design rules help limit the increase in capacitances due to fill, but are by no means sufficient to eliminate the capacitance impact of CMP fills.[1] Furthermore, current extraction tools are inaccurate when it comes to extraction of floating fill impacts.

The semiconductor and design industries must accurately incorporate the impact of fills during extraction. In Chapter II, we present a parameterized DOE-based methodology to improve the accuracy of extraction in the presence of fills.

[1] For example, second neighboring layer coupling can be significant, yet there are no explicit design rules to restrict such a coupling component.

Addition of fills can be handled by either the design house or the foundry. In the latter case, while there is limited control of capacitance increase due to lack of design information (timing constraints and critical paths, functionality and stimuli, etc.), the capacitance impact analyses enabled by our DOE are still useful. When the design house inserts fills, some or all of the fills can be included in the RC extraction flow. In this case, our proposed DOE enables an accurate analysis of the impact. An integration methodology without deviation from industry standard flows is needed. Design and process parameters that alter the capacitances need to be identified. A compact DOE structure is needed to reduce simulation time. A thorough analysis is needed to understand how floating fills affect capacitances. And, different fill algorithms need to be compared. We target these requirements in Chapter II and provide designers with tools to conduct such analyses in arbitrary BEOL technologies, for any integrated-circuit design.

In the literature, Stine et al. have proposed CMP models in [5]. Lakshminarayanan et al. [6] have proposed electrical rules to reduce resistance variations due to CMP, and have proposed width-dependent spacing rules to be used. Zarkesh-Ha et al. [7] have linked the interconnect density to CMP models and circuit delay. Cueto et al. [8] have provided an algorithm to extract floating fills. Yu et al. [9] have introduced a field solver which can take into account floating fills by using floating fill conditions in the direct boundary element equations. Wang et al. [10] have used a minimum variance iterative method for density assignment. Nelson et al. [11] have determined fill size, shape and proximity to reduce parasitic capacitances using a 2D analysis. Batterywala et al. [12] have presented an extraction method, where fills are eliminated one by one using a graph-based random walk algorithm while updating the coupling capacitances. Kurokawa et al. [13] have shown that interlayer coupling can be more important than intralayer coupling. Lee et al. [14] have analyzed the impact of intralayer fills on capacitances. Kahng et al. [1] have provided design guidelines to reduce coupling. In [15], the authors have provided fill patterns to reduce interconnect coupling. Park et al. [16] have presented an exhaustive method to generate capacitance tables for fills. Chang et al. [17] have presented a charge-based capacitance measurement method to analyze the impact of fills. In [18], the impacts of fills have been analyzed using an effective permittivity model. Kim et al. [19] have proposed compact models for the impacts of fills.

A graph node reduction-based technique was proposed in [12] where, initially, capacitances between all metal shapes are computed. Figure I.4 illustrates such a graph node reduction-based technique. Metals are represented by the nodes of a graph, and the capacitances by the graph edges. For the capacitance computation, voltages of floating metals are first estimated by interpolation. Coupling capacitances are then computed assuming charge on a floating fill is zero. In the next step, capacitances between interconnects and fills and between two fills are eliminated using the following formulas:

$$Cik' = Cik + \frac{Cif \cdot Ckf}{Cff} \tag{I.1}$$

$$Cjk' = Cjk + \frac{Cjf \cdot Ckf}{Cff} \tag{I.2}$$

Following the reduction, fill nodes are eliminated one by one. This is a very slow process as the number of fills is large.

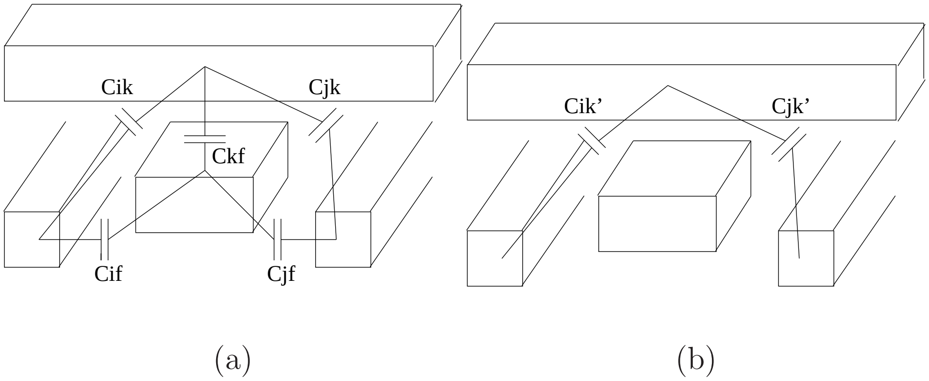


Figure I.4: Graph-based capacitance reduction. (a) Initially, capacitances between all metals are computed. (b) Capacitances to fills are eliminated one by one.

There also exist random walk-based methods which can account for floating fills. These methods are used by tools such as Magma Quickcap [20] and Syn-

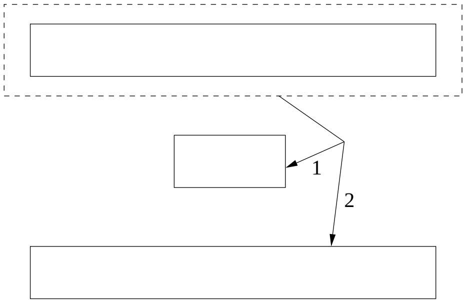


Figure I.5: Random walk-based methods.

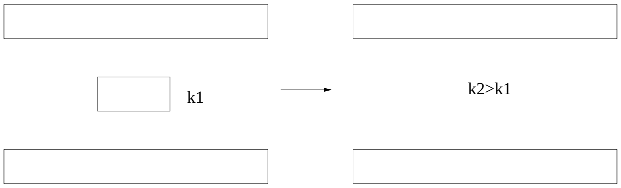


Figure I.6: Effective-k fill elimination method.

opsys Raphael NXT [21]. Electrostatic equations are solved using Gauss' Law. A random walk starts at a Gaussian surface and ends at the same or another interconnect, yielding a coupling capacitance estimate. Such a random walk is illustrated in Figure I.5. The potential voltage of fills is estimated using an integration over their surface. Application of this method within practical runtimes is limited to small layouts.

Another popular method for post-fill parasitic extraction is using an effective dielectric constant and eliminating the fill. Such a method is described in [22]. An effective dielectric constant method is illustrated in Figure I.6. Fills are dropped and the effective dielectric constant is increased to compensate. This approach can be considered as a compact modeling approach. Due to multi-layer interactions, compact models usually are insufficient except for very restrictive configurations.

I.A.2 Metal Fill Optimization

Maximal insertion of regularly-patterned floating fills is known to be inefficient from the standpoint of both pattern density variation control and post-fill electrical performance, even though this is still a method in use today. Previous works establish guidelines for fill insertion that can reduce capacitive coupling [1] and show the importance of constraining fill impact on capacitance of timing-critical signal nets [23]. Nevertheless, traditional fill synthesis methods still use simplistic (e.g., Boolean) pattern operations and fail to provide necessary automation of such complex fill optimization guidelines. In Chapter III, we propose a novel energy-minimization model for fill synthesis, through which complex fill guidelines can be implemented. Furthermore, interlayer-aware and power-aware fill synthesis methodologies are needed. We target these problems in Chapter III.

In the literature, Stine et al. [24] have targeted metal fill optimization to reduce dielectric thickness variations. Lee et al. [18] have analyzed the effects of certain parameters on parasitics. Recently, extraction of capacitance values for layouts which include floating fills have been important. Floating fills are preferable over grounded fills since they have less (albeit more variable) coupling and timing impact, do not interfere with power and ground network design, and yield more opportunity for performance optimization. Park et al. [16] and Cueto et al. [8] have developed field solvers which can take into account floating fills. Tugbawa et al. [25] have provided CMP models and a calibration methodology. Lee et al. [14] have proposed design guidelines for metal fill insertion and also proposed an RC extraction methodology. Kurokawa et al. [13] have presented a method for extraction in the presence of fills by mapping fill sizes to effective dielectric properties (thickness, permittivity). Kahng et al. [26] have proposed a window-based solution to the metal fill problem such that a density bound is preserved and variability across windows is reduced during metal fill insertion. Chen et al. [27] have introduced a performance-driven fill insertion method, either to minimize total delay impact or to maximize minimum slack. Kahng et al. [28] have proposed slotting algorithms in addition to filling. Xiang et al. have proposed a slot-based coupling-aware fill insertion method in [29] subsequent to [2]. However in [29], no timing impact is provided. Xiang et al. [30] have also presented a layout density analysis method.

A number of fill layout guidelines have been suggested for designers in [1]. These guidelines include centralization of fill locations with respect to neighboring interconnects, positioning of fills toward edges of available layout regions, maximization of the number of fill columns and minimization of the number of fill rows (between two vertically-oriented interconnects). We have provided additional fill impact analysis, including interlayer impacts, in [31]. We have identified an hour-glass fill pattern for optimal intralayer coupling in [2]. However, known fill synthesis algorithms cannot effectively apply these rules, and manual application is impractical. Hence, new fill synthesis algorithms are still required.

A force-directed scheme to optimize fill shapes and locations was first proposed in [2]. Chapter III presents a simpler version (e.g., without an “atomistic” physical analogy). We validate our methodology through parasitic extraction and timing analysis of standard-cell designs. Furthermore, critical net and interlayer awareness properties are added. Force-directed schemes have been proposed for many years in the realm of VLSI standard-cell and module placement; see, e.g., [32] for an overview. In the VLSI placement context, the circuit netlist provides connectivity information: energy between cells is a function of separation in the layout, and inter-cell distances can be used in the energy minimization. By contrast, in the CMP fill placement problem, no such connectivity is inherently present. Rather, given a layout of signal interconnects, fills must be inserted according to a guideline, with each floating fill shape considered independently.[2]

Traditional Fill Insertion. Traditional fill insertion does not take into consideration the coupling between interconnects, despite a number of “timing-driven fill” or “intelligent fill” approaches having been proposed over the years [4]. Typically, a window-based “fixed-dissection” scheme is used, where overlapping windows span

[2] In the VLSI high-level synthesis literature, Paulin et al. [33] proposed a heuristic model using force-directed energy minimization for scheduling. Our method also targets energy minimization, but as one would expect the models are significantly different.

an entire layout, and density constraints must be observed in each window. Linear programming frameworks can be used to determine target amounts of fill to be added into each portion of each window, so that overall layout density is as uniform as possible [23][28]. The actual fill pattern is then synthesized to realize these target amounts using, e.g., arrayed or staggered rectangular shapes. Fill shapes and pattern may also be selected to help satisfy a given target density in a window. More naive flows simply insert as many fills as possible, as long as fill dimension and spacing constraints are satisfied. Such "maximum-fill" solutions give reasonably uniform post-fill layout density, but fill impact on timing is usually higher than necessary.

I.A.3 Fill Thermal Effects

Temperature has an exponential dependency on leakage. A 2.5^{o}C increase in temperature can lead roughly to a 5% increase in leakage and the dependency of leakage on temperature is exponential. 10^{o}C and 30^{o}C temperature increases roughly correspond to 2x and 10x increase in leakage, respectively. There is significant opportunity to reduce temperature through efficient layout design. In particular, fills can be optimized to reduce temperature.

The thermal conductivity of materials is an important criterion for temperature gradients. Metals such as copper can conduct the temperature fairly efficiently. Polysilicon has a mid-range CTC (coefficient of thermal conductivity). On the other hand, dielectrics have very low CTC values. If CTC of a material is high, this means that the heat generated at a point can be transferred to another point on the wafer efficiently through this material, reducing the resultant temperature near the heat source. If the CTC of the material near the heat source is low, then the heat is trapped and the local temperature rises. Hence, dielectrics cause a significant restriction on the distribution of heat generated by a transistor.

It is possible, through efficient layout structures, to reduce the device temperature. In Chapter VI, we show certain guidelines to enable reducing the temperature through layout structures such as via and contact fills. We increase the volume of material connected to a device, thereby reducing the device temperature for the same amount of heat generated in the channel. Furthermore, test structures need to be identified so that common layout configurations in the presence of fills can be studied. We also target this problem in Chapter VI.

Most of the work in the literature for temperature optimization target higher level design such as placement, routing and architecture. However, devices and layout are important knobs for temperature and have not been fully utilized until now.

Zhang et al. [34] have presented a thermal-aware routing algorithm. Cong et al. [35] and Tsai et al. [36] have introduced thermal-aware placement algorithms. Cong et al. [37] have introduced thermal via placement. Goplen et al. [38] have proposed thermal-aware force-directed placement of standard cells. Tsai et al. [39] and Hung et al. [40] [41] have proposed temperature-aware floorplanning algorithms. Lee et al. [42] have provided thermal-aware partitioning. Cho et al. [43] have provided an algorithm for thermal-aware clock routing.

Joule heating in interconnects has been investigated by Li et al. in [44]. Casu et al. [45] have optimized power supply wires considering SOI self heating. Banerjee [46] has investigated the thermal effects in interconnects. Chiang et al. [47] have provided compact models for electro-thermal analysis of interconnects.

In the area of leakage modeling, Roy et al. [48] have analyzed the physical roots of transistor leakage. In the area of thermal simulation, Wang et al. [49] have presented a transient thermal simulator. Li et al. [50] have provided a grid-based full-chip analysis methodology. Zhan et al. [51] have provided a discrete cosine transform and table look-up based thermal analysis methodology.

In the area of self heating modeling, Zimin et al. [52] have analyzed heat generation and dissipation mechanisms in SOI MOSFET's. Oshima et al. [53] have analyzed the impact of buried layers to reduce self heating effects in SOI. Jin et al. [54] have proposed a characterization method for self heating. Olsson [55] has studied self heating in BJT's. Grasser et al. [56] have proposed a global model which is applicable to multiple transistor types. Tenbroek et al. [57] have analyzed the static and dynamic electro-thermal behavior in analog circuits. Workman et al. [58] have provided physical models for self heating. Semenov et al. [59] have analyzed the impact of self heating on reliability. Zheng et al. [60] have simulated the self heating. Hu et al. [61] have provided an analytical self heating model based on Poisson equation. Su et al. [62] have provided a measurement methodology.

Background on Heat Conduction. The law of heat conduction states that the time rate of heat transfer through a material is proportional to the negative temperature gradient and the surface area for heat flow. This relationship can be given through the following equation:

$$\frac{\partial Q}{\partial t} = -k \oint_s \nabla T \dot{d} T \tag{I.3}$$

In this equation, Q is the transferred heat amount, t is time, k is material heat conductivity, T is temperature and S is the area.

Chapter VI presents test structures to evaluate the impact of fills on temperature. As probing local temperature changes due to fills is difficult in silicon, we conduct TCAD simulations. We follow our TCAD (technology computer-aided design) simulations with guidelines for temperature reduction in circuits.

I.B Stress in Integrated Circuits

In this section, we begin with a brief introduction to stress, then detail how stress is exploited in FEOL and managed in BEOL.

I.B.1 Brief Introduction to Stress

Stress is average force per unit area. The stress unit is Pascals (Pa). Stress cannot be measured directly, but strain, which is the deformation caused by stress, can be measured. Based on the understanding in [63], the stress vector T_x acting normal to x is given as $T_x = \sigma_{xx}.\hat{x} + \sigma_{xy}.\hat{y} + \sigma_{xz}.\hat{z}$. The stress tensor is defined by the three stress vectors:

$$\sigma_{ij} = \begin{bmatrix} \sigma_{xx} & \tau_{xy} & \tau_{xz} \\ \tau_{yz} & \sigma_{yy} & \tau_{yz} \\ \tau_{zx} & \tau_{zy} & \sigma_{zz} \end{bmatrix}$$

In this equation, σ_{ii}'s are stress components normal to the unit cube faces, whereas τ_{ij} are shear components directed towards j on the orthogonal face to i. The σ_{ii}'s are used for analyzing the impact of stress. Individual stress components can be used for mobility computation for FEOL (front-end of the line) analysis. Computations in [64] can be used to convert the stress components into mobility improvement for device analysis.

Using the individual stress components, we can furthermore compute the average stress (volume averaged hydrostatic stress) and the von Mises stress as given in Equations (I.4) and (I.5) for BEOL (back-end of the line) analysis.

$$\sigma_{av} = (\sigma_{xx} + \sigma_{yy} + \sigma_{zz})/3 \tag{I.4}$$

$$\sigma_{vm} = \frac{1}{\sqrt{2}}\sqrt{\frac{(\sigma_{xx} - \sigma_{yy})^2 + (\sigma_{yy} - \sigma_{zz})^2 + (\sigma_{zz} - \sigma_{xx})^2}{2}} \tag{I.5}$$

We next provide introductions to FEOL and BEOL stress in integrated circuits.

I.B.2 FEOL Stress

At the 65nm process node and beyond, it is evident that stress-based techniques for mobility improvement will dominate traditional geometric scaling to maintain Moore's Law [65] trajectories for device performance. Enabling progress has been made in the manufacturing process and TCAD (modeling and simulation) to support stress. However, FEOL (front-end of the line) stress has not yet been exploited by layout optimizations to improve design performance. In Chapter IV, we present a new methodology that combines detailed placement and active-layer fill insertion to exploit shallow-trench isolation (STI) stress for IC performance improvement. We furthermore provide an analysis of the dual stress liner (DSL) technology by conducting TCAD simulations on the test structures we design. We present layout guidelines for DSL optimization. We conduct a process sensitivity analysis for DSL. Our methodology for STI stress optimization begins with process simulation of a production 65nm STI technology, from which we generate mobility models for STI stress. We develop STI stress-aware SPICE modeling and simulation of critical paths, and finally perform timing-driven optimization of STI stress in standard cell designs, using detailed placement perturbations to optimize PMOS performance and active-layer fill insertion to optimize NMOS performance.

Stress Modulation Techniques in the Process. In 65nm processes, a number of stress modulation techniques have been introduced:

- SiGe stress from underneath the channel
- Embedded SiGe from the source and drain
- (Dual) stress liner
- Stress memorization
- Hybrid orientation

Early stress modulation methods employed a silicon-germanium (SiGe) layer underneath the channel, which improves the channel mobility. More recently, embedded SiGe (e-SiGe) [66] [67] [68] has been used in the source and drain regions to exert stress along the channel, and improve PMOS speed. The stress liner technique involves deposition of stressed liners over the transistors on top of the polysilicon. Single stress liners may be used to cover the entire wafer with compressive or tensile liners, possibly followed by a masked doping step to destress one type of the liners. Alternatively, dual stress liners [69] [70] may be used to cover the NMOS devices with a tensile liner and PMOS devices with a compressive one. Stress memorization technique [71] [72], used typically to improve NMOS speed, relies on plastic deformation of certain materials due to a process step and the consequent *memorization* of the applied stress in the channel. Finally, in the hybrid orientation technique [73] [74], crystal orientations are used to enhance NMOS and PMOS speeds separately.

STI stress is the stress that is exerted by STI wells on device active regions and is generally compressive in nature. Irrespective of the use of stress modulation techniques in the process, STI stress is not negligible and its magnitude depends on the sizes of the STI wells and the active regions for a given process. In Chapter IV, we present a technique that modulates stress in timing critical devices to improve circuit delay by altering the STI widths (STIW) adjacent to the devices.

Stress Modeling Techniques. In the area of stress modeling and characterization, Rueda et al. [63] have provided general models for stress. Gallon et al. [75] have specifically analyzed the stress induced by STI. Bradley et al. [76] have characterized the piezo-resistance of CMOS transistors. Sheu et al. [77] have modeled well edge proximity effect on MOSFETs, and in [78] have modeled the effect of mechanical stress on dopant diffusion. Su et al. [79] have proposed a scalable model for layout dependence of stress. Miyamoto et al. [80] have provided a layout dependent stress analysis of STI. Recently, Tsuno et al. [81] have shown 65nm silicon data showing that STI width stress effect can impact drive current by up to 10%. However, no models or optimization methodology is presented. Without appropriate models, circuit level optimization would not be automated or accurate.

With respect to the STI process, several optimizations have been developed to reduce the STI stress, but they typically fail to completely eliminate layout-dependent stress impacts. Elbel et al. [82] have proposed an STI process flow based on selective oxide deposition. Lee et al. [83] have proposed an optimization for densification of the STI fill oxide to reduce the stress. Miyamoto et al. [80] also have proposed process innovations to reduce the active-area layout dependence of MOSFET characteristics. Looking forward, the introduction of e-SiGe (embedded SiGe) for certain 65nm nodes in the source and drain may reduce the stress variation due to STIW for PMOS, but NMOS performance can still be actively improved by exploiting the STI width effect. If these STI stress controlling techniques are used, we expect the improvements demonstrated in Chapter IV to reduce but remain non-negligible.

Stress TCAD simulations have been conducted by Moroz et al. in [84] and [85], and by Smith in [86]. The work of Moroz et al. is significant for indicating possible ways to enhance performance using STI stress; however, no circuit-level optimizations are explained. With respect to the current body of knowledge: (1) models are still needed to relate stress due to the STI width effect to transistor mobilities, and (2) there is still a lack of available stress optimization methods. A fundamental research goal is to develop novel and efficient simulation, modeling, analysis, and optimization methods to support next-generation stress-aware EDA technology. Our work strives to enable this.

I.B.3 BEOL Stress

BEOL (back-end of the line) stress is very important in terms of interconnect and dielectric reliability. Although substantial research has been conducted on CMP fill optimization for metal and dielectric height uniformity [24] [18] [16] [14], little is known about the impact of CMP fill patterns and via fill[3] on dielectric and interconnect stress and reliability. There has been recent interest in modeling pattern-dependent CMP effects [87]. We evaluate the impact of CMP fill pattern and location, as well as the presence of via fill, on dielectric and interconnect stress. We also suggest how to tie these results to certain BEOL reliability concerns.

Going toward the 32nm node, a key challenge is BEOL integration. In particular, dielectric reliability in the regime of low-k dielectrics is a major concern. As BEOL processing involves large amounts of stress, stress becomes a primary issue for BEOL reliability with low-k dielectrics, which are more fragile.

The dielectric and interconnect reliability depends not only on the process steps, but also on the layout pattern for metal layers. CMP fill shapes, which are

[3] Via fill is a via between a CMP fill and another CMP fill or an interconnect.

traditionally used to reduce metal and dielectric height variations due to CMP, can alter the BEOL reliability. Using via fills can similarly improve or degrade reliability issues.

Traditionally, CMP fills are used to optimize the density requirement so as to achieve a uniform metal and dielectric height. We provide optimizations to reduce coupling and total capacitances in Chapter III. With the dielectric breakdown issues facing the progress of BEOL integration, optimization to reduce BEOL stress is also required. In Chapter V, we evaluate the fill patterns, locations and the via fill approach in terms of their stress. As stress measurements are difficult to conduct on silicon and it is impractical to observe local changes due to stress, we conduct TCAD simulations. We have used a new TCAD tool from Synopsys, FAMMOS [88], to conduct this analysis. Impact of fills needs to be tied to BEOL reliability concerns. Test structures are required to investigate impact of fills in common layout configurations. Stress measurement points on the test structures need to be identified to relate stress to BEOL reliability concerns. Guidelines need to be then provided to the designers for reliability-aware fill insertion. We target these problems in Chapter V.

There have been quite a few modeling and simulation work for copper interconnect and low-k stress. Xu et al. [89] have analyzed interconnect stress through simulations. Paik et al. [90] have simulated the effect of low-k dielectric on stress. Srinivasan et al. [92] have provided insights for the impact of interconnect scaling on the reliability. Thurn et al. [93] have analyzed the impact of chemical vapor deposited dielectric films. Degryse et al. [94] have shown the impact of bonding on low-k. Besser et al. [95] have provided measurements and models of stress and strain evolution. For the measurement of stress, the relevant literature includes [96].

The main BEOL reliability concerns with copper interconnects and low-k dielectrics are stress migration, time-dependent dielectric breakdown (TDDB) or bias-temperature stability (BTS), delamination and crack formations, copper diffusion into low-k and electromigration. Particular references are given for each reliability concern below.

- On the topic of stress migration, Wang et al. [97] have provided simulations for stress-induced voiding. Alers et al. [98] have related the mechanical properties of copper to stress migration. Bentz et al. [99] have provided grain-based modeling of stress induced copper migration. Zhai et al. [100] have simulated and provided experiments for the stress migration. Shi et al. [101] have provided finite element models for stress migration. Stress migration results in metal voids as a result of stress gradients in the metal lines. Reduction of stress gradients results in improved stress migration reliability. The main causes of the stress gradients are the development of tensile stresses over multiple thermal cycles and the thermal mismatches due to the cooling from passivation temperatures [102].

- In the area of TDDB, Wu et al. [103] have provided models for TDDB breakdown in copper metallization. Hwang et al. [104] have investigated the intrinsic dielectric breakdown mechanisms in low-k. Yiang et al. [105] have analyzed the cap layer impact on TDDB. A physical model has been provided in [106]. There have been works in the literature showing the impact of stress on TDDB for gate dielectrics [107]. Following a similar reasoning, it is also expected that stress would influence the TDDB characteristics for low-k dielectrics.

- Delamination in patterned films has been analyzed by Liu et al. [108]. Stress components normal to material boundaries can result in crack starts and delamination. Reducing the stress would help to reduce the delamination. Impact of stress on delamination has been analyzed in the literature both for front and back end of the line. Stress components normal to the boundary between two materials will increase the delamination process.

- Copper ion diffusion in low-k has been studied by Aw et al. in [109] through simulations. Takeda et al. [110] have studied the copper-induced dielectric breakdown.

- Electromigration is mainly a result of electrical stress, but it can also get worse due to mechanical stress gradients. It has been indicated by Park et al. in [102] that electromigration can be impacted by mechanical stress, particularly in the interconnect extensions close to vias.

Beyond mechanical stress, concerns also include thermal and electrical stress. Cherault et al. [111] have analyzed the thermo-mechanical stress evolution using finite element method. Lim et al. [112] have analyzed impact of temperature stress on copper interconnects. Rhee et al. [113] have shown the effects of thermal stress on copper interconnects. Alam et al. [114] have provided design methodologies for electromigration.

Stress has an impact on all of these reliability concerns. In the literature, there has been a need to relate CMP fill and via fill impacts to stress and BEOL reliability concerns. In Chapter V, we target to fill this gap. It is necessary to understand how design layouts, in particular fills, affect BEOL stress and relate these to BEOL reliability. Stress cannot be measured directly, but strain can. However, its measurement is not trivial. Due to manufacturing and mask costs and measurement limitations resulting from local layout changes, it is rather difficult to observe impact of small features such as fills. Hence, a TCAD-based analysis is required. Furthermore, designers need guidelines for improving the layout for reliability. In Chapter V, we target these needs.

I.C Outline

We target the needs identified in this chapter by:

- designing test structures to cover known design and process combinations,
- conducting TCAD (technology computer-aided design) simulations,
- developing necessary models,
- providing design guidelines,
- optimizing circuits, and

- automating the optimization process.

In Chapter II, we provide a methodology and analyses based on TCAD simulations for fill impact on capacitances. In Chapter III, we provide an automated way to synthesize fills based on guidelines. We apply our technique to large testcases and achieve performance improvements in terms of both timing and power. In Chapter IV, we provide models for STI width stress effect for CMOS devices. We provide a standard-cell optimization method and use it to improve performance of circuits. In Chapter V, we conduct TCAD analyses on BEOL stress and relate the impact of fills on BEOL reliability. We provide design guidelines based on our observations. In Chapter VI, we analyze the impact of fills on thermal performance of devices and provide design guidelines.

II

Fill Impact Analysis

There is a need for public compact DOEs for analyzing fills and improving accuracy of current extraction flows. Furthermore, floating fill impact on capacitances needs to be thoroughly analyzed and different fill algorithms and options need to be compared. In this chapter, we target to achieve these and provide practical methods and parameterized DOEs for any design house or foundry to use on their technology to understand, analyze and characterize the impact of fills in their flow.

This chapter is organized as follows. Following motivations (Section II.A), where we identify inaccuracies in current extraction tools, we present in Section II.B our proposed DOE methodology. In Section II.C, we provide details of the simulation structures for our DOEs and show how the DOE algorithms are implemented. In Section II.D, we provide an insight on the keep-off distance, which is a key design rule related to CMP fills. We then provide a means to include the height variations due to CMP. In the experimental results section, we provide exhaustive simulation results for our experimental design for three types of fill algorithms: standard (traditional), staggered, and 2-pass. We show how much inaccuracy we would have observed, had we used approximations such as merged fills or grounded fills.

This chapter extends [31] with more detailed explanations of our methodology, and with new DOEs for via fills and active fills. Via fills are typically used between CMP fills of two neighboring layers to improve close-proximity via printability and dielectric reliability. With increasing use of via fills in 65nm and below process nodes, their impact on capacitances needs to be characterized more accurately. Active region fills are used in stress optimization and to control STI height variations due to STI CMP.[1] We provide a corresponding DOE which targets the impact of active region (AR) fills on capacitances. Since via and active region fills are relatively recent introductions to standard methodologies, our resultant observations open new grounds for designers and the design-manufacturing interface.

II.A Motivation

Current extraction tools have known inaccuracies for inclusion of floating fill impact on final coupling and total capacitances.[2] Most tools use simplifications to account for effects of fills. In this section, we review simplifications commonly used by extraction tools. Along with each simplification, we also indicate how much error can be introduced for a typical structure.

Assuming Floating Fills as Grounded. Before any other approach, extractors assumed that the floating fills are grounded. The same capacitance tables that are also used to extract the regular interconnect capacitances were initially used. Based on our data summarized in Table II.5, this assumption could result in up to 2x and 10x underestimation for first and second neighboring layer coupling capacitances, respectively, as well as almost eliminating the intralayer coupling capacitances.

Merging the Fills. A popular method is to merge all the neighboring fills within a layer into one large fill. The larger fill is constructed such that it is the smallest rectangular prism which can include all neighboring fills and does not include any part of an interconnect. This approach is illustrated in Figure II.1, where fills are

[1] If uncontrolled, STI step height variations can lead to, e.g., increased threshold voltage variability.

[2] The type of fill of interest to this chapter is floating fill, since grounded fills are not versatile due to routing and increased total capacitances, and since accuracy of their extraction is not a concern for current extraction tools.

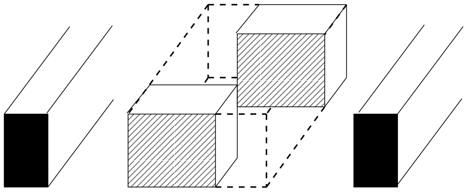


Figure II.1: Fill merging methodology. Fills are patterned and interconnects are represented by dark lines. Dashed box shows the merged fill, resulting from the convex hull of the intralayer neighboring fills.

replaced by the larger fill indicated by the dashed block.

Merging the fills may result in up to 10x average overestimation of the intralayer for small keep-off distances and underestimation of second neighboring-layer coupling capacitances up to 4x depending on the fill algorithm based on our data summarized in Table II.5. First neighboring-layer coupling capacitance can be underestimated up to 2x. Another extension of this assumption is accounting for fill density only. Some extraction tools take density of fills as input to their models. In this case, different fill patterns yielding the same density are assumed to yield the same results. However, different patterns yielding the same fill density are known to yield different coupling capacitances.

Other Inaccuracies in Consideration of Floating Fill Extraction Tools. Another important inaccuracy is related to first and second neighboring interlayer coupling, i.e., coupling between layers M and $M+1$, and between $M-1$ and $M+1$, respectively. Patterns on M and $M+1$ impact the coupling between the interconnects in these layers. As fills on layer M are introduced, couplings between interconnects on layers $M-1$ and $M+1$ are impacted according to the pattern in M. Hence, approximations such as the merging or grounding of fills will result in inaccuracies.

II.B Methodology

Current extraction tools do not contain accurate design of experiments for floating fills, although the DOEs for regular interconnects are sufficient. We provide an extensive DOE set for the floating fills. Our proposed method consists of a parameterized field solver DOE and normalization of results to enable a normalization-based extraction methodology for fills. In the traditional flow, after interconnects are designed and fills are automatically or manually inserted into the design, the extraction tool is run over the layout. As the extraction tools will each use one of the methods analyzed in the previous section, their results will not be sufficiently accurate.

According to our proposed flow illustrated in Figure II.2, the results are normalized to include the impact of fills. This flow makes it possible to compare impacts of different fill algorithms using results of the same extraction for interconnects with no fills in between. Essentially, we propose to first run an extraction tool over the interconnects with no fills; this step is accurately handled by current extraction tools. Then, using the fill DOE, we propose to update the impact of fills on coupling and total capacitances using a normalization step. The normalization is done with respect to the same structure and interconnect parameters without any fills in between interconnects. The capacitances with the fills are normalized with respect to capacitances without the fills. This results in normalized values close to and higher than 1, whenever the capacitance increases due to fills. The normalized couplings are mostly expected to be larger than 1, as fills usually increase coupling. The normalized data in the capacitance tables are then used to convert the result of extraction with no fills to accurate results accounting for the presence of floating fills. We use 3D field solutions for our DOE and hence the results will be much more accurate than known approximations.

Integration with Extraction Tools. Extractors use a DOE for regular interconnects. They may use an additional DOE for floating fills, for which the users have restricted inputs. With the proposed method, the DOE for the regular interconnects should still be used and the parasitics with no fills generated through

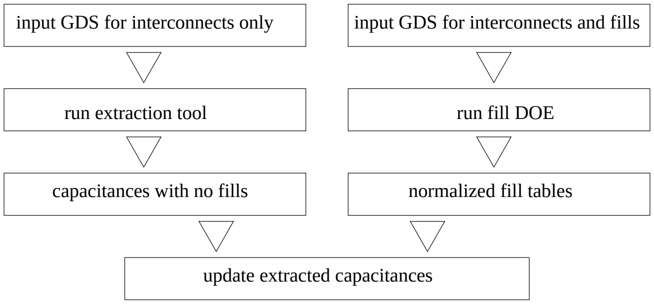

input GDS for interconnects only
input GDS for interconnects and fills
run extraction tool
run fill DOE
capacitances with no fills
normalized fill tables
update extracted capacitances

Figure II.2: Proposed flow to incorporate floating fill impact.

the extractor. The extractor DOE for floating fills can be replaced by the proposed DOE through wrapper scripts which utilize the normalized look-up tables generated through the field solver using the proposed fill DOE, the input netlist, input layout, and the parasitics file generated by the extractor. Alternatively, the proposed DOE may be integrated into the extractor through the tool provider.

II.C Fill DOEs

Basic CMP Fill DOE structure. In this section, we propose our parameterized DOEs. These DOEs can both be used for analysis and characterization of a process, as well as generating capacitance tables.

To reduce the runtime to a manageable amount, we have designed one structure for all DOEs, as shown in Figure II.3, except the parallel neighboring layers DOE, which uses a version where interconnects are parallel in each layer instead of orthogonal. We propose a 5-layer structure, with top and bottom plates grounded. Each layer consists of two parallel interconnects facing each other. Parallel interconnects rotated 90 degrees to each other are used in layers $M-1$, M and $M+1$. Here, layer M refers to the layer in the middle. In layers $M-1$, M and $M+1$, two parallel interconnects are present, with fills in between placed according to parameters and a selected fill algorithm, the end results of which may look like the ones in Figure 1 of [15], i.e., standard, staggered, 2-pass, etc. Layers $M+1$ and $M-1$ include orthogonally oriented interconnects with respect to layer M. Interconnects on layers $M-1$ and $M+1$ overlap with each other, though an additional parameter can be used to introduce shifting of the overlapped interconnects. The simulated structures are parameterized according to the particular fill pattern (algorithm) of interest.

In the figure, interconnects on layer M are drawn vertically, whereas interconnects on layer $M+1$ or $M-1$ are drawn horizontally as dark rectangles. We have included in the simulation window, indicated by dashed lines, half width of each interconnect to account for the Neumann boundary conditions. These boundary conditions enable the mirroring of each structure along the dashed lines. Hence, essentially part of a large regular pattern is simulated.[3,4]

The DOE structure is able to provide all the coupling capacitances of interest. For intralayer coupling, capacitances between lines on layer M are used in the proposed structure. For neighboring-layer coupling, capacitances between one line on layer M and $M+1$ each are used in the proposed structure. For second neighboring-layer coupling, capacitances between lines on layers $M+1$ and $M-1$ are used in the proposed structure. For neighboring layer parallel line capacitances, the structure has been modified such that there are two parallel lines on neighboring three layers.

[3] While implementing the DOE structures, interconnect lengths are selected long enough to enable a repetitive pattern according to Neumann boundaries. The given parameters otherwise define the simulation structure unambiguously.

[4] While constructing the fill tables, the capacitances, are normalized with respect to the interconnect length if the coupling is between parallel interconnects. If orthogonal, we have recorded the capacitance without normalization.

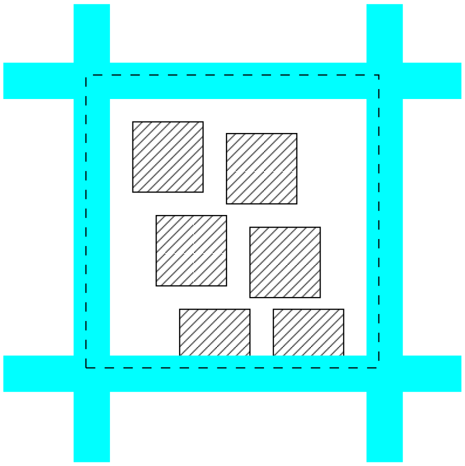


Figure II.3: Basic DOE structure. The structure consists of 5 layers. Top and bottom are ground planes. Three layers consist of parallel interconnects orthogonal to others across each layer. The structure enables observation of intralayer, first and second neighboring layer couplings in one simulation. Fills are patterned and interconnects are represented by solid dark lines.

II.C.1 Basic CMP Fill DOE for Intralayer Coupling

Our fill DOE is given below. Assuming there are four parameters of interest, the algorithm looks like the following:

```
1. foreach $w_f = w_f^{min} : w_f^{inc} : w_f^{max}$ {
2.   foreach $w_s = w_s^{min} : w_s^{inc} : w_s^{max}$ {
3.     foreach $c_f = c_f^{min} : c_f^{inc} : c_f^{max}$ {
4.       foreach $w_m = w_m^{min} : w_m^{inc} : w_m^{max}$ {
5.       $Run\ field\ solver\ over\ parameterized\ structure\ and\ add\ result\ to\ a\ table\}\}\}\}$
```

In this DOE, w_f and w_s refer to fill width and spacing between fills, respectively. c_f is the number of fill columns between two parallel interconnects for each of the layers $M-1$, M and $M+1$. w_m refers to metal width. w_f^{inc} corresponds to the increment and is equal to $(w_f^{max} - w_f^{min})/(num.\ of\ data\ points)$. Usually, four data points is sufficient to come up with reasonable data tables or compact models. min

and *max* for the fill parameters refer to the minimum and maximum values for a parameter, which usually can be decided using the design manual.

To enable updating of interconnect coupling and total capacitances with fills added, the fill capacitance models need to be normalized with respect to the same configuration including no fills. Hence, the same DOE structures are run with no fills present between the interconnects and the results with fills are normalized with respect to the results without fills. During extraction, when interconnects are seen in design, coupling capacitances between interconnects are multiplied by the normalized DOE results.

The runtime complexity of the algorithm is a function of the number of parameters and number of data points for each parameter. So, it is highly recommended to look for ways to reduce these. Herein, we provide a couple of guidelines. If a relationship between a parameter and the impact is known to be linear, then only two data points for that parameter should be selected. Certain parameters change at the same time as other parameters. For example, dielectric height changes with the dielectric constant. These kinds of parameters need to be tied to each other so that only one loop is executed for both. If sensitivity of coupling to a parameter is known to be low, then this parameter can be thrown out by setting it to a constant. Similar to field solver setups with current extraction tools, a careful selection at this step will be highly rewarding in terms of runtime.

II.C.2 CMP Fill DOE for Neighboring Layer

There are two types of interlayer couplings. One of them is first neighboring layer coupling. For an interconnect on layer M, neighboring layer refers to interconnects on layers $M-1$ and $M+1$. On the other hand, second neighboring layer refers to coupling between an interconnect on layer $M+1$ and an interconnects on layer $M-1$. Neighboring coupling is mainly of fringing type, whereas second neighboring coupling is of area overlap type, as the interconnect surfaces face each other.

Neighboring layer interconnects are most of the time orthogonal to each other to reduce coupling. A cross-over structure in 3D simulation yields exact coupling between the interconnects. However, the addition of fills around the interconnects increases this coupling.

There are two extreme cases for the location of these fills. For worst-case coupling, the fill can be overlapping the next layer interconnect from top view. This situation is shown in Figure II.4(b). In the figure, the shaded rectangles are the fills on layer M. The least coupling occurs when the fills on layer M are shifted. This is shown in Figure II.4(a).[5] Similarly, fills on layers $M + 1$ or $M - 1$ also have worst- and best-case coupling positions. The corresponding DOE consists of evaluating all incremental configurations between these worst and best cases for neighboring layers. Hence, one parameter is added to evaluate the fill shifts.

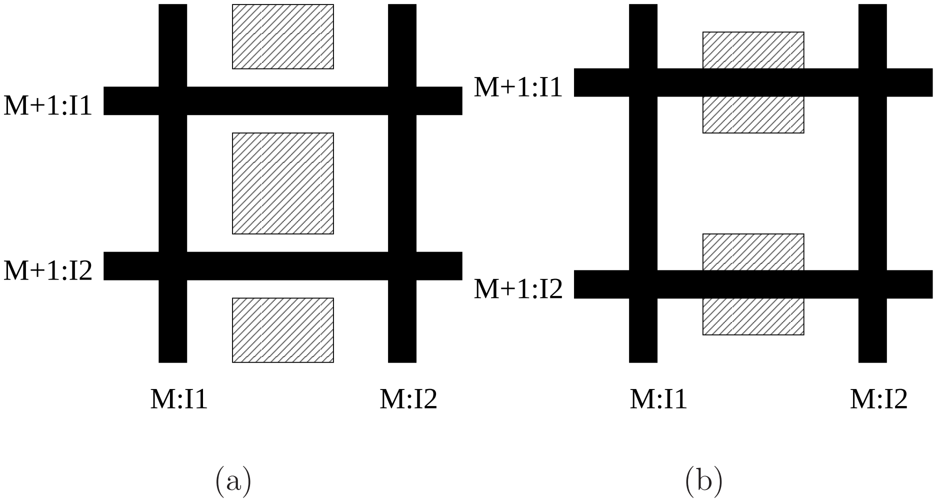


Figure II.4: Interlayer coupling for neighboring layers. (a) Fills on vertical layer M intersect minimally with horizontal interconnects on overlapping layers $M - 1$ and $M + 1$. (b) Fills on layer M are shifted and intersect maximally with interconnects on layers $M - 1$ and $M + 1$.

With respect to the originally defined DOE, we can augment the DOE by adding the following loop between lines 4 and 5.

[5] The shifting will impact coupling even with staggered fill patterns, especially if fill widths are large.

4.1. **foreach** $shiftM = shiftM^{min} : shiftM^{inc} : shiftM^{max}$ {}
Here, $shiftM$ denotes the amount of shift for layer M fills.

II.C.3 CMP Fill DOE for Parallel Neighboring Layer Coupling

It is possible that two consecutive layers have parallel lines. This condition is especially possible in lower layers as well as layers close to clock networks. To handle such a configuration, we have used a modified simulation structure as described above and illustrated in Figure II.5 from a side view. Worst- and best-case shifts again need to be implemented. The same DOE presented in the previous section is used with the pattern in Figure II.5.

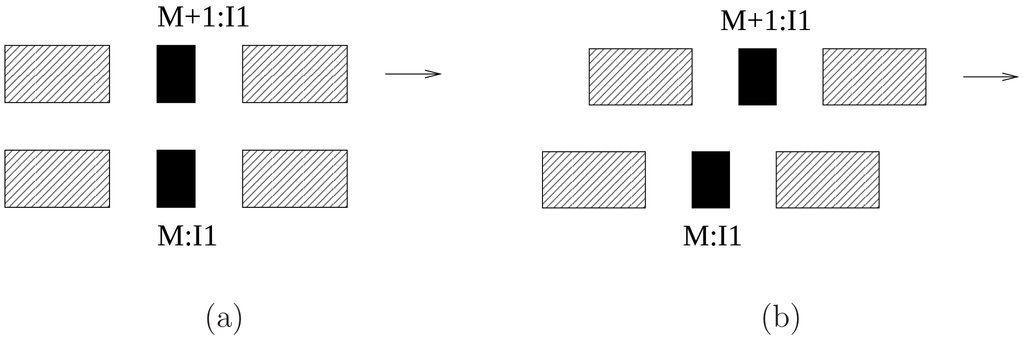


Figure II.5: Interlayer coupling for parallel neighboring lines. (a) Interconnects on layers M and $M+1$ intersect. (b) Layer $M+1$ shifted.

II.C.4 CMP Fill DOE for Second Neighboring Layer Coupling

To analyze the layer M fill impact on $M-1$ and $M+1$ coupling capacitances, the structure shown in Figure II.6 is used. Practically, we have used the same structure from Figure II.4 to reduce the number of simulations and hence handle both DOEs in one simulation. Similar to the previous DOE, positions for fills for best- and worst-case couplings should be identified.[6] Also, lines in $M-1$ and $M+1$ may not be overlapping. To account for these shifts, $M+1$ lines should be shifted by up to half the minimum spacing allowed between two interconnects. In our DOEs, we have only shifted the fills.

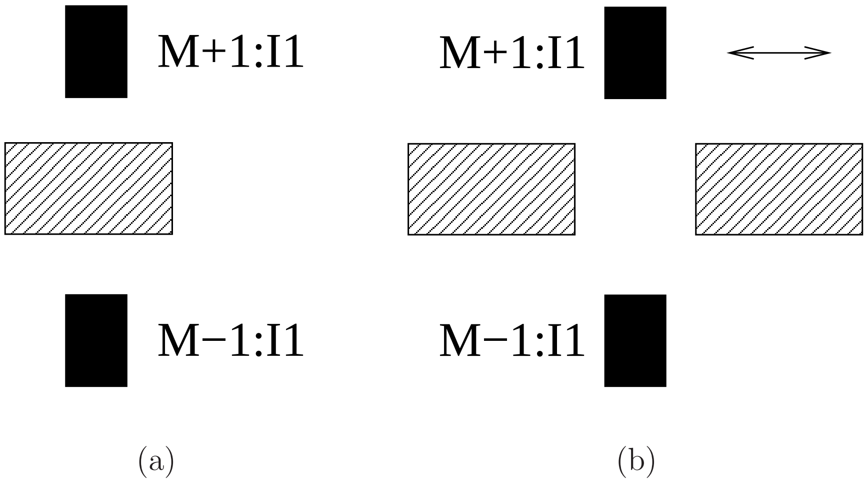


Figure II.6: Interlayer coupling for second neighboring layers. (a) Fills on layer M intersect maximally with interconnects on layers $M-1$ and $M+1$. (b) Fills on layer M are shifted and intersect minimally with interconnects on layers $M-1$ and $M+1$.

II.C.5 Implementation of Other Fill Patterns

The proposed DOE can be extended to other common fill patterns, such as staggered, two-pass or alternating rectangles. In this subsection, we briefly describe how we have implemented the DOE for staggered and two-pass methods.

Staggered Fill Algorithm. Staggered fill algorithm produces a shape similar to the standard fill algorithm, except that each row and column is *staggered* by a fixed distance.

Two-Pass Algorithm. Two or three-pass algorithms insert rectangles of two or

[6] For staggered patterns, these shifts are only important for line lengths on the order of the fill width.

three different sizes. Largest rectangles are inserted first, and are placed in the middle of two interconnects to minimize first neighboring layer coupling. Smaller fills are then inserted in the following steps.

II.C.6 Via Fill DOE

For the via fill DOE, we have used the simulation structure shown in Figure II.7. Overlapping fills laid out in traditional pattern are inserted between parallel interconnects in neighboring layers. On top and bottom layers, the structure is covered by ground planes representative of dense lines in an actual design. Via fills are inserted between neighboring layer fills to connect them.

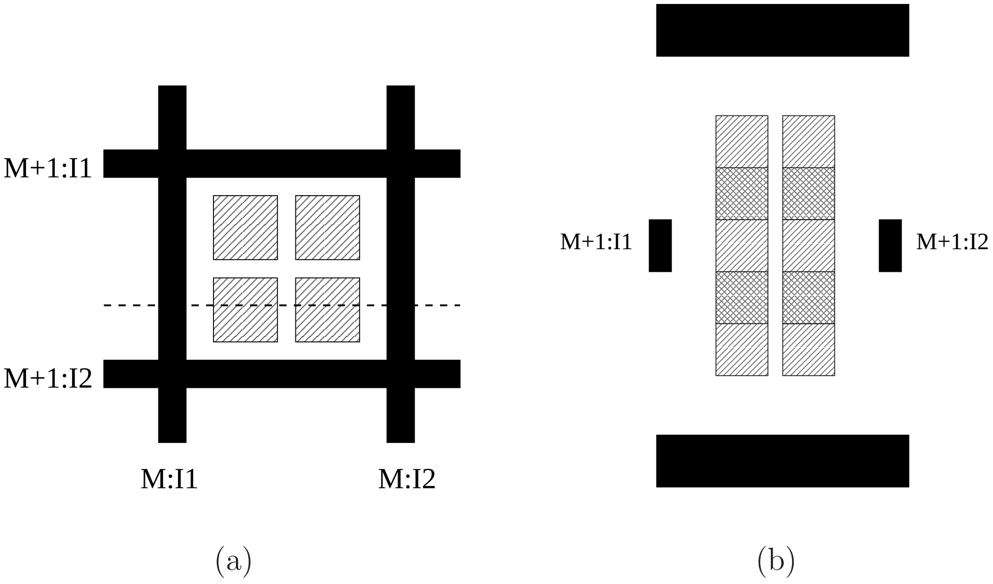


Figure II.7: Via fill DOE simulation structure. (a) Top view. CMP fills laid out in traditional pattern shown. $M + 1$ and $M - 1$ layer interconnects are overlapping. Top and bottom ground planes not shown. (b) Side view along dashed line. Via fills connecting the CMP fills on neighboring layers indicated by darker patterns.

II.C.7 Active Fill DOE

The simulation structure for the active fill DOE is shown in Figure II.8. We have in particular monitored the coupling between the control interconnect and an active region in the same layer as the AR fills, as well as the coupling between two active regions. The reflective simulation boundary is selected such that an AR fill and two half active region widths can fit into the window from both sides. Hence the width of the simulation window is given as $wact + 2 * act2act + wf$, where $wact$ is the active region width, $act2act$ is active to active spacing, and wf is the fill width.[7] For the upper layer, as many interconnects as can be fit are included into the window after placing the control interconnect in the middle of the simulation window.

We have used a shift parameter which shifts the interconnects in the layer above altogether as a fraction of the pitch. We have monitored the coupling to the control interconnect during these shifts.

Three AR fills have been used in the orthogonal direction as shown in Figure II.8 (b). Similarly sized and spaced three active regions are used. This orthogonal direction corresponds to the widths of transistors formed by the active regions. The reason of using same sized regions spaced out from each other is the fact that in standard cell designs, transistor rows of equal widths are used.

II.D Extensions for Keep-Off Rule and CMP Impacts

II.D.1 On the Keep-Off Design Rule

One of the design rules most relevant to floating fills is the keep-off, or exclusion, distance. This distance is defined as the minimum distance that a fill must be away from an interconnect. In this section, we provide some intuition

[7] Spacing between an active region and active region fill may be chosen larger than the spacing between two active regions to reduce coupling effects.

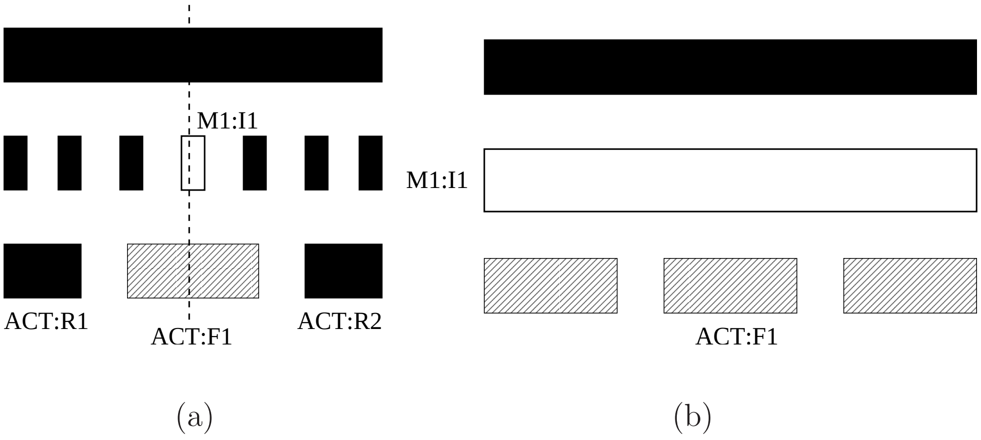


Figure II.8: Active fill DOE simulation structure. (a) Side view. (b) Side view along dashed line. Active region fills are shown as cross-hatched. Control interconnect is light-colored.

about this design rule.

This design rule is usually selected such that the coupling capacitance to an intralayer neighbor is negligible as compared to the total capacitance of a line. We have conducted an experiment on a layer with the values in Table II.1. We have changed the keep-off distance from $0.1\mu m$ to $0.9\mu m$ and observed the change in coupling capacitance over the total capacitance. This plot is shown in Figure II.9.

The coupling over the total capacitance is practically negligible (3%) around $0.5\mu m$, hence $0.5\mu m$ is likely to be selected as the keep-off distance for the layer for which this experiment has been conducted. As the fills are allowed to be closer to interconnects, corresponding to a lower keep-off distance, the coupling increases.

Having a large keep-off distance, although advantageous in terms of reducing intralayer coupling, has other issues. It becomes difficult to insert fills into certain regions to satisfy a density constraint, as the distance between two parallel interconnects has to be larger than two times the keep-off distance for fill insertion. Consequently, CMP results in more variations. A second issue is increased

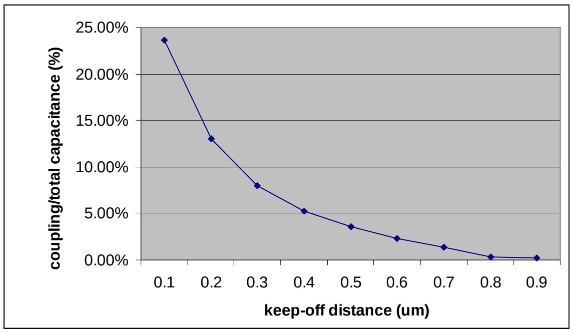


Figure II.9: Intralayer impact of keep-off distance.

Table II.1: Parameters Used in Keep-Off Distance Experiments.

metal height	dielectric height	dielectric constant	
$0.3\mu m$	$0.3\mu m$	3.1	
keep-off distance	metal width	fill spacing	fill width
0.1-$0.9\mu m$	$0.1\mu m$	$0.1\mu m$	$0.5\mu m$

coupling of interconnects to neighboring layers. As keep-off distance is increased, less electric flux is present between interconnects of the same layer. However, this flux is directed to interconnects on neighboring layers.

It is possible to have an edge over the design rule if accurate extraction is available. Historically, design rules appear before any analysis and optimization technique. With aggressive technologies, there is an unavoidable need to be able to analyze the effects of each interaction. In the context of the keep-off design rule, as accurate extraction has not been possible, the solution has been to restrict the proximity of fills to interconnects.

With the basic elements of an accurate extraction flow presented in this chapter, it is possible to analyze the impact of reduced keep-off distances on coupling and total capacitances as well. This permits greater flexibility of fill algorithms in regions where coupling between lines is not critical. Reducing the keep-off distance enables tighter metal density uniformity, as well as reduced interlayer coupling capacitances.

II.D.2 Incorporation of CMP Impacts

CMP is known to result in variations of copper height and hence dielectric height. CMP models exist which give metal heights in a tile within a layer. It is then necessary to tie these heights to the final capacitance values. We have run a set of experiments to evaluate the effect of height variations on the coupling and total capacitances. One such analysis is shown in Figure II.10. The x axis gives the multiplication factor we have used for the height. Values on the y axis are coupling capacitances. We have observed a linear relationship between height and both coupling and total capacitances. The implication is that by just running simulations for two different heights followed by linear interpolation or extrapolation, one can find the CMP-impacted capacitance.

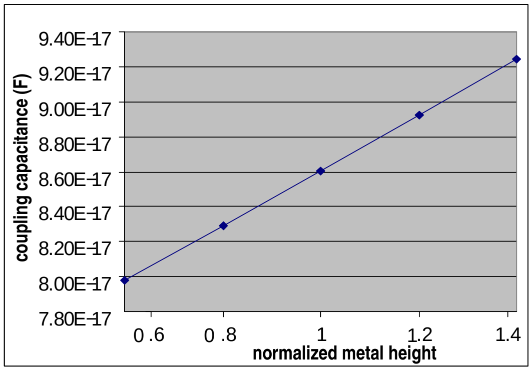


Figure II.10: CMP-induced height impact on coupling capacitance shows linear change. Nominal coupling capacitance in F as a function of normalized metal height is shown.

II.E Experimental Results: DOE Analyses

Using the proposed DOE, we provide an analysis of relationships we have observed. We have used three different fill algorithms. For each algorithm, we have repeated the simulations for merged fills and grounded assumptions for comparison. We have also simulated the structures with no fills for normalization. Each simulation takes between 10 to 120 seconds, depending on the selected parameters. All the DOEs take roughly 24 hours to 48 hours on a $2.4GHz$ 2 processor dual-core server with $2GB$ of memory, using the 3D field-solver Raphael [115]. We have used a minimum grid size of 100,000 nodes per each structure. We have used up to 10 licenses and 5 machines to further reduce the simulation time. The standard fill algorithm is parameterized using the values shown in Table II.2. Here, dielectric constant, metal and dielectric heights, changed at the same time, enable simulation of local, intermediate and global interconnects in the interconnect stack. Parameter names appended by a star sign are changed simultaneously to reduce the number of simulations as described above.

II.E.1 Analysis of Intralayer Coupling DOE for Standard Fills

Intralayer coupling exists, e.g., between same-layer interconnects such as $M : I1$ and $M : I2$ in Figure II.4. As fill width increases or fill spacing decreases, intralayer coupling increases as shown in Figure II.11. The increase is more pronounced if there are more columns, i.e. impact due to fills increases from 4.1x to 7.2x as the fill width is increased from 0.4 to $0.6\mu m$ for three fill columns [1]. With one column of fills only, the increase stays around 1.8x. As the fill to fill spacing is increased from 0.1 to $0.7\mu m$ as shown in Figure II.12, the impact decreases from 4.1x to 1.7x for three columns, and from 1.7x to 1.4x for one column.

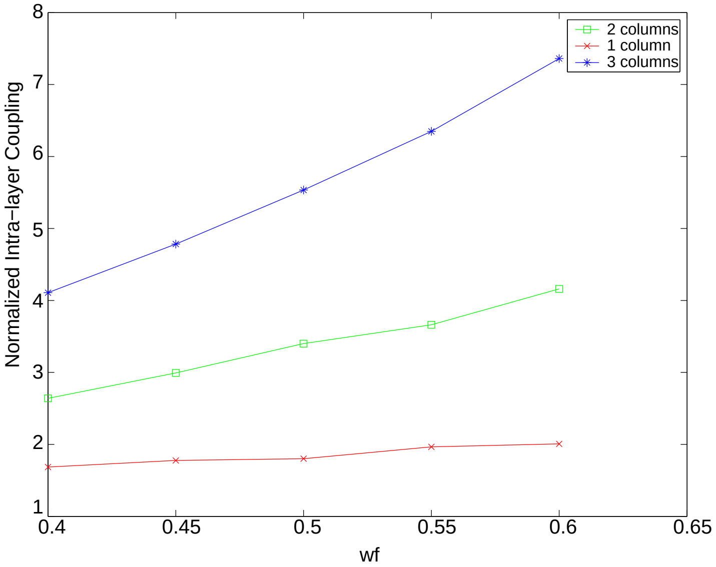


Figure II.11: Fill width dependency of intralayer coupling for different numbers of fill columns.

II.E.2 Analysis of First Neighboring Interlayer Coupling

To illustrate how much the shift can impact the coupling, we have used the representation as shown in Figure II.13. In the figure, each sample corresponds to a set of six simulations, where the shift parameter is changed from 0 to 1 in increments of 0.2. These numbers are multiplicative constants, which are multiplied by half the pitch. 125 samples are shown, corresponding to 750 field solver simulations. The corresponding sample is computed as follows:

$$s_I = max\ (v_i)/min\ (v_j) - 1\ :\ \forall\ v_{i,j}\ \epsilon\ v_I \qquad \text{(II.1)}$$

Here, I is a set of six experiments where the shift parameter is changed while keeping other parameters fixed; s_I is the corresponding sample value; and v_i and v_j

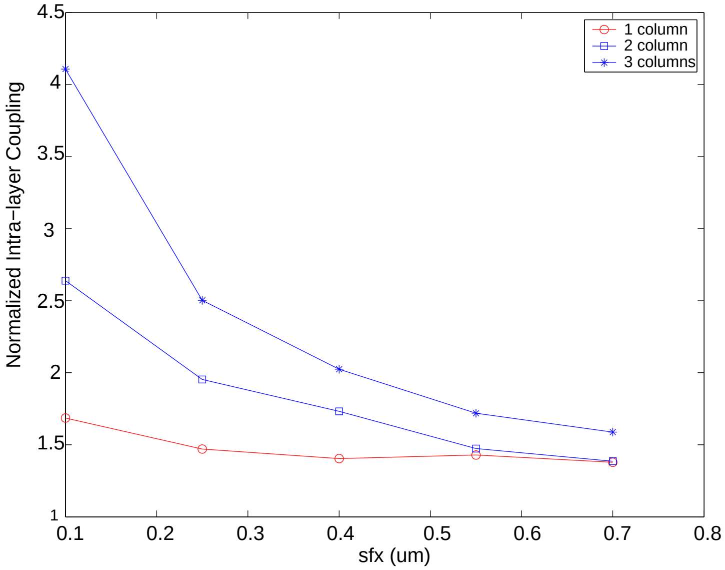


Figure II.12: Fill spacing dependency of intralayer coupling for different numbers of fill columns.

are values of the experiments in set I. Essentially, the maximum over the minimum of the values of a set gives the maximal change due to the shift operation. A 1 is subtracted to indicate the change.

Interlayer coupling exists between neighboring layer interconnects such as $M : I1$ and $M + 1 : I1$ in Figure II.4. We observe that the maximum of all the samples corresponds to a 5% change due to the shift of fills on layer M only. We consider this amount negligible, considering that we have used an almost best-case choice of $300nm$ keep-off distance for this analysis. The data set with largest impact corresponds to fill width, fill spacing and metal widths of $0.6\mu m$, $0.4\mu m$ and $0.4\mu m$, respectively, in our technology.

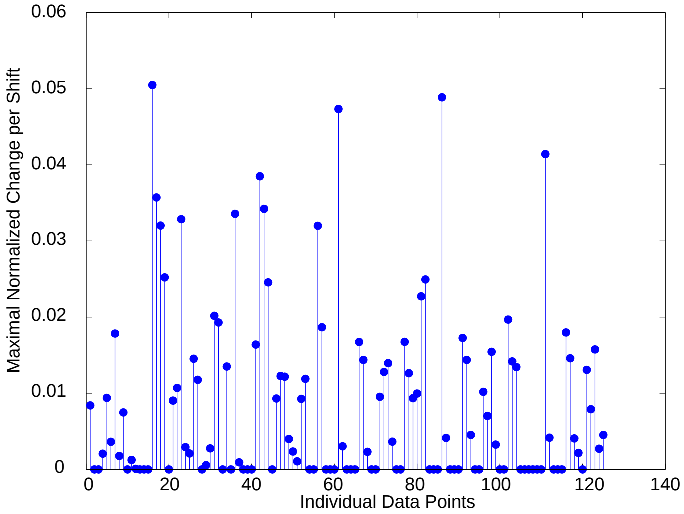


Figure II.13: Normalized data showing maximal change in coupling of neighboring lines due to shift.

II.E.3 Analysis of First Neighboring Layer Parallel Line Coupling

Parallel coupling exists between neighboring layer parallel interconnects, such as $M : I1$ and $M : I + 1$ in Figure II.5. A similar analysis has shown that the maximum of all the samples corresponds to 2.8% change due to parallel shifts of both interconnect and fills on layer $M + 1$. The data set with largest impact corresponds to fill width, fill spacing and metal widths of $0.4\mu m$, $0.55\mu m$ and $0.2\mu m$, respectively.

We observe that the neighboring layer coupling increases as a function of the shift parameter for this DOE up to 8% from no shift to half pitch shift as shown in Figure II.14. For small shifts, there is negligible impact. As shift is increased, field lines between interconnects on neighboring layers are blocked by a larger fill, which increases the coupling. Larger metal widths usually increase parallel coupling.

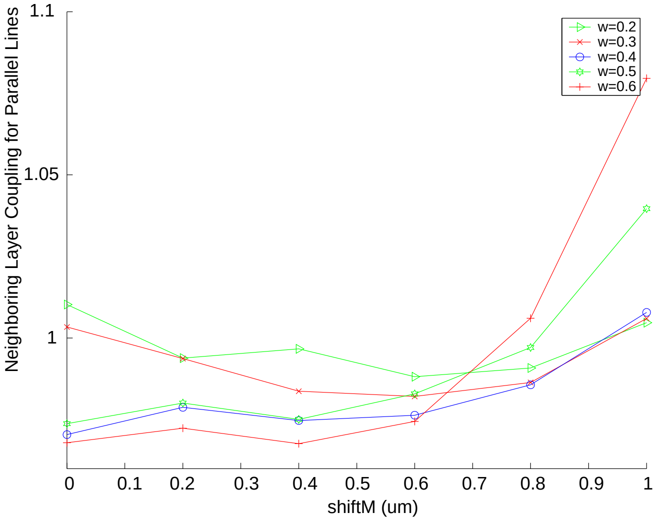


Figure II.14: Neighboring-layer parallel line coupling dependency on amount of M layer shift for different metal widths.

II.E.4 Analysis of Second Neighboring Interlayer Coupling

Second-layer coupling observes the coupling between interconnects such as $M-1 : I1$ and $M+1 : I1$ in Figure II.6. Keeping fill width and spacing at $0.4\mu m$, which corresponds to around 25% density, we have observed the fill shift dependency in Figure II.15.[8] At $shiftM = 0$, there is maximum overlap between

[8] Exact density depends on the window in which the density is calculated.

interconnects of layers $M-1$ and $M+1$, and up to 1.55 times the coupling is seen with respect to no fills. Shifting the layer M fills by changing the $shiftM$ parameter reduces the coupling around 20%. When fill width is small, there is negligible change due to the shift of fills, as field lines between larger interconnects on layers $M+1$ and $M-1$ can find a direct path without going through the fills. Changing the fill spacing and keeping fill and metal widths at $0.4\mu m$ and $0.2\mu m$ as in Figure II.16, we have observed that when the spacing between fills is small, the change in coupling due to the shift in layer M fills is negligible. On the other hand, increasing the spacing between fills (decreasing the metal density from 65% down to 25%) on layer M results in a 35% change, which is significant and requires extraction.

II.E.5 Analysis of Other Fill Patterns

We have parameterized the staggered fill algorithm by adding a parameter to define the stagger amount as 0.2, 0.25 or $0.275\mu m$ in addition to the parameters for the standard fill algorithm. For the 2-pass algorithm, we have used a two-pass ratio parameter value of 2 or 3 to define the larger fill width with respect to the narrower width which is inserted in the second step.

Table II.2: Parameters for Standard Fill Algorithm.

metal width	$0.1, 0.2, 0.3, 0.4$ (μm)
fill width	$0.4, 0.45, 0.5, 0.55$ (μm)
fill spacing	$0.1, 0.25, 0.4, 0.55$ (μm)
fill shift	$0.25, 0.5, 0.75, 1$ (x)
metal height*	$0.3, 0.4$ (μm)
dielectric height*	$0.3, 0.4$ (μm)
dielectric constant*	$3.1, 2.8$
number of fill columns	$1, 2, 3$
keep-off distance	$0.3, 0.5, 0.7$ (μm)

We have parameterized the via fill DOE using the variables in Table II.3. We have parameterized the AR fill DOE using the variables in Table II.4.

Table II.3: Parameters for Via Fill DOE.

metal width	$0.2, 0.35, 0.5$ (μm)
fill width	$0.4, 0.5$ (μm)
fill spacing	$0.1, 0.25, 0.4, 0.55$ (μm)
number of rows	$2, 3$
number of columns	$2, 3$
metal height	0.3 (μm)
dielectric height	0.3 (μm)
dielectric constant	3.1
keep-off distance	$0.3, 0.5$ (μm)

Table II.4: Parameters for Active Fill DOE.

active region width	$0.2, 0.3$ (μm)
active to active spacing	$0.15, 0.3, 0.45, 0.6$ (μm)
active fill width	$0.4, 0.6$ (μm)
metal spacing	$0.2, 0.35, 0.5$ (μm)
metal height	0.3 (μm)
shift factor	$0, 0.25, 0.5$
dielectric height	0.3 (μm)
dielectric constant	3.1

II.E.6 Comparison of CMP Fill DOE Results

Table II.5 contains a summary of all the simulations for standard, staggered and 2-pass algorithms. In order to compare the results, we have repeated

Table II.5: Comparison of DOE, Merged and Grounded Extraction for Standard, Staggered and 2-Pass Algorithms.

STANDARD	DOE	Merged	Grounded	Max. Coup. /Total	Min. Coup. /Total
intralayer	2.377	10.336	0.002	15.91%	0%
first-layer	1.083	1.123	0.492	22.25%	17.11%
second-layer	1.126	0.726	0.094	6.84%	2.38%

STAGGERED	DOE	Merged	Grounded	Max. Coup. /Total	Min. Coup. /Total
intralayer	2.579	25.9308	0.0021	23.33%	0%
first-layer	1.131	1.155	0.578	20%	16.32%
second-layer	1.153	0.559	0.107	6.870%	0%

2-PASS	DOE	Merged	Grounded	Max. Coup. /Total	Min. Coup. /Total
intralayer	5.308	34.607	5.998e-6	3.607%	0.909%
first-layer	1.110	0.531	0.546	19.562%	15.913%
second-layer	1.0160	0.284	0.147	7.776%	3.566%

the field solver simulations for the merged and grounded fills in addition to the proposed DOE. For merged fills, all same-layer neighboring fills are lumped into one, using the convex hull of the fills. For the grounded fills, the same fill pattern as the DOE is used, except each floating fill is connected to ground. In the table, the columns from left to right are the means of normalized DOE, merged and grounded fill results. These columns indicate the normalized increase in coupling capacitances due to fills. The normalization is with respect to the original interconnect structure with no fills. The last two columns indicate the magnitude of each coupling term as a percentage of the total capacitance.[9] We have included

[9] The default settings for most extractors cause the tool to neglect coupling capacitances below 1%.

both the maximum and minimum for this ratio. This ratio shows the importance of the given coupling capacitance.

The rows of Table II.5 show intralayer, first layer neighboring and second layer neighboring coupling, respectively, for each fill algorithm. Using the data from the 2-pass algorithm as an example, looking at the last two columns, we can say that the intralayer coupling shows less impact as compared to second-layer and first layer couplings. In terms of accuracy, we can see that, the increase in intralayer coupling due to fills can be 13.73 times greater using merged fill as compared to the DOE results, whereas this ratio can be almost negligible for the grounded fills.[10] The DOE results are closer to "actual results" than are to the outputs of approximate methods such as merged fill or grounded fill. Merged fills result in an overestimation of coupling capacitances, whereas grounded fills result in a significant underestimation. Although overestimation could be thought of as advantageous, there are two reasons why we believe it is not an advantage. The first reason is that the overestimation is significantly high. The second reason is that, as we observe the next two rows, we see that the overestimation for the intralayer coupling has resulted in an underestimation for both first and second neighboring layer couplings due to the fact that merged fills attract most of the flux which would otherwise go to the interconnects on the neighboring layers. Observing the first-layer coupling row, although we would expect an increase in coupling capacitances due to the insertion of fills, we see a reduction for the merged and grounded fills as indicated by normalized values lower than 1. This happens due to the flux reasoning above. Considering the fact that these coupling capacitances are large portions of the total capacitance, inaccuracies will be highly important. Standard and staggered algorithms also have shown similar inaccuracies, especially for the intralayer and second-layer couplings. As the proposed DOE uses accurate field solutions which take into consideration the pattern shapes and parameters, the results will be highly accurate with respect to known approximations.

Asymmetric Configurations. Although the simulations have been provided for

[10] Some of the high increase is due to allowing small keep-off distances, which can be helpful in achieving high-density fills.

symmetric parameters for neighboring layers, it is possible to find asymmetric configurations in real designs. For example, the metal width, fill width and spacing can be different between layers. Our simulations with asymmetric parameters indicate that it is common to see differences up to 40% with respect to the symmetric case. If two layers have similar parameters but the third layer has different, the difference between symmetric case may reduce down to 10%. As a general guideline, asymmetric design practices should be avoided as much as possible, as it would make the CMP process optimization harder. All local, intermediate and global interconnects should have a common set of fill width and spacing sizes within themselves for design uniformity. Even though it is possible to reduce the asymmetry in design using such guidelines, there will still be an asymmetry between transition metal layers, i.e. local to intermediate, or intermediate to global transitions. Such cases, and cases when design has known asymmetries, can easily be integrated into the proposed DOEs. The possible increase in the number of parameters should be controlled by restricting the number of parameters, may be through considering best- and worst-case parameter combinations only.

For the via fill DOE, Table II.6 shows the average normalized change in intralayer, first and second neighboring layer capacitances. These correspond to $M : I1$ to $M : I2$, $M+1 : I1$ to $M : I1$ and $M+1 : I1$ to $M-1 : I1$ in Figure II.7, respectively. When keep-off distance is $0.5\mu m$ for the parameters values given in Table II.3, the latter two change negligibly, however a 17% increase on the average is expected for intralayer capacitances using the parameters given in Table II.3. The same experiment for the case when keep-off distance is reduced to $0.3\mu m$ has not shown a significant change. Hence, we can conclude that keep-off distance has less control over the intralayer capacitance increases unlike the case for CMP fills.

For the AR fill DOE, Table II.7 shows the average normalized change in first neighboring layer and intralayer capacitances for the parameters values given in Table II.4. An average of 28% increase is expected for the first neighboring layer coupling between the control line ($M1 : I1$) and one of the active regions in the middle ($ACT : R1$) in Figure II.8 (b). This is not a negligible amount and needs to be incorporated using our DOE. The intralayer coupling between two middle active

Table II.6: Analysis of Via Fill DOE for Keep-Off of $0.5\mu m$.

$0.5\mu m$ keep-off	mean	max
intralayer	1.17	1.37
first-layer	1.00	1.02
second-layer	0.99	1.00

regions on either side of an AR fill ($ACT : R1$ to $ACT : R2$) though can reduce by 20% and increase by 17% depending on the parameters we have used. For example, when the active to active spacing and fill width are increased to $0.6\mu m$ and the upper layer interconnects are dense with minimum widths, a 20% reduction is seen as the field lines between the active region and the control interconnect are pulled away towards the farther away active region with the help of the AR fill. Hence, fills may be utilized to reduce certain couplings at the expense of others. Some coupling capacitances can be reduced by up to 20%, but other couplings from the same line will increase, which may overcome the benefit as the increase can be larger than 20%. Furthermore, for many configurations, couplings cannot be reduced by more than 2%. Meanwhile, total capacitance usually increases. Although it is not possible to reduce the total capacitance, it is possible to control the capacitance increase.

Table II.7: Analysis of Active Fill DOE.

	min	mean	max
first-layer	0.99	1.28	1.92
intralayer	0.8	1.00	1.17

II.F Conclusions

We have proposed a DOE set along with a compact DOE structure. We have identified the relevant parameters for DOE. We have thoroughly analyzed the impact of floating fills and compared different fill algorithms and configurations. We have proposed a normalization-based flow to integrate the proposed DOE with current extraction tools for accurate extraction of capacitances in the presence of floating fills. We have provided a parameterized design of experiments so that designers can implement to analyze their designs and technology. Our proposed field solver DOE set completes the DOE set that comes with the extractors, which is not optimal for floating fills. We have shown that the proposed field solver-based DOEs provide significant accuracy improvements over methods and assumptions used by current extraction tools. We have shown that keep-off distance may not be used as a controlling factor for via fills. Finally, we have found that active region fills results in non-negligible capacitance change. We believe that our contributions will enable a better overall analysis and extraction of the impact of fills on capacitances in interconnect technologies by virtue of its extensive and parameterized nature.

II.G Acknowledgments

We would like to thank David Overhauser and Sam Nakagawa for several valuable discussions. This chapter is based on or in part a reprint of the following publications.

- A. B. Kahng and R. O. Topaloglu, “DOE-based extraction of CMP, active and via fill impact on capacitances,” *IEEE Trans. on Semiconductor Manufacturing*, 21(1), 2008, pp. 22-32.
- A. B. Kahng and R. O. Topaloglu, “A DOE set for normalization-based extraction of fill impact on capacitances,” **Best Paper Award**, *Proc. IEEE International Symposium on Quality Electronic Design*, 2007, pp. 467-474.

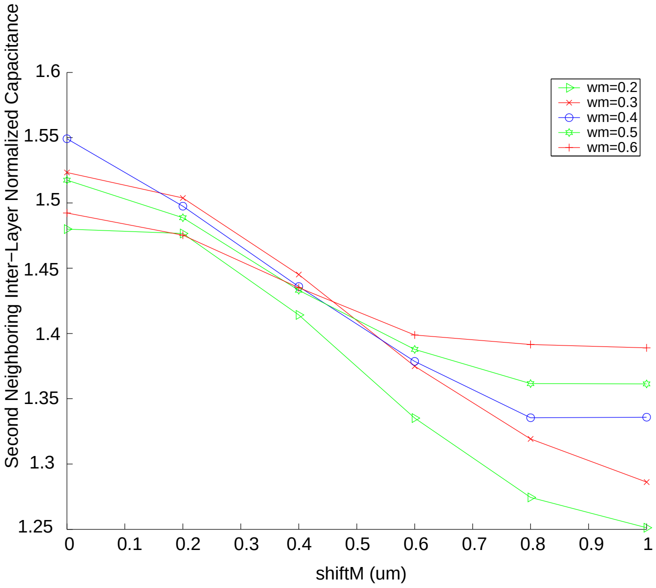


Figure II.15: Fill shift dependency of second neighboring layer coupling for different metal widths.

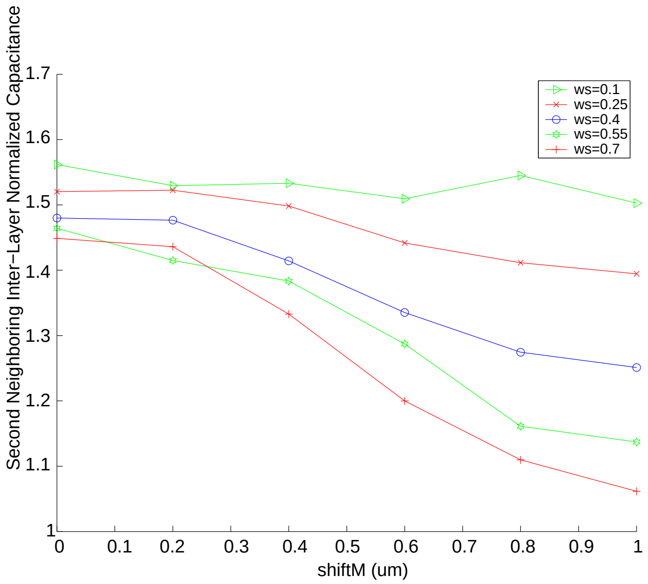


Figure II.16: Fill shift dependency of second neighboring layer coupling for different fill to fill spacings.

III

Metal Fill Optimization for Performance

Having a methodology for accurate fill impact analysis, one can estimate the impact of fills on capacitances. There is also a need to optimally place the fills so that not only the CMP uniformity is targeted, but the circuit performance can be improved. In particular, interlayer coupling needs to be considered. Insertion of fills needs to be automated utilizing available fill insertion guidelines. For power-critical circuits, power-aware fill synthesis methodology is needed. In this chapter, we target these problems and provide an automated fill synthesis framework by which circuit performance can also be improved.

In our fill synthesis framework, the possible locations for fill insertion have allocated energies. When a fill is to be inserted, the location with lowest energy is sought. Fills are inserted into a given region one by one, until a prescribed density constraint is satisfied. The energy network is designed so as to enable insertion according to design guidelines. In this framework, the design of the energy network becomes an important task, essentially migrating the fill synthesis problem from circuit design to the CAD algorithm domain. We propose energy models and give insights to ease the design of such energy networks.

The remainder of this chapter is organized as follows. We describe our fill optimization methodology in Section III.A, where we also present the heuristics used in our method. We provide optimizations targeting timing in Section III.B. We provide power-aware optimization in Section III.C. We provide experimental results in Section III.D. We conclude the chapter in Section III.E.

III.A Fill Optimization Methodology

We now describe our fill optimization methodology, covering the steps of region definition, grid definition, energy modeling within a grid, and fill insertion within a grid.

III.A.1 Adaptive Region Definition

Given a metal layer, we use a version of the scan-line algorithm to determine regions between facing interconnects with no other interconnect in between. A *region* consists of a maximal rectangular area such that two interconnects directly face each other in the horizontal (vertical) direction, when the main routing orientation in the given interconnect layer is vertical (horizontal).[1] Each region knows the interconnects or boundaries that define its edges from West and East directions. By convention, half of the width of each interconnect from the West and East directions are included in a region. We illustrate region definition with four interconnects in Figure III.1.

To obtain a uniform post-fill metal density while also considering performance impact of fill, we use a uniform target density for each region. Our region definition is adaptive in the sense that it follows the specific configuration of the interconnect design, in contrast to methods that simply dissect the layout into uniform square regions. A region can be formed from:

1. facing interconnects, e.g., $R2$-$R7$ in Figure III.1.

2. an interconnect and chip boundary, e.g., $R1$ in Figure III.1.

[1] We assume the main interconnect orientation is vertical in the rest of the chapter. We rotate the layers with horizontal orientation as necessary so that their orientation becomes vertical.

3. two facing chip boundaries.

In all cases, no interconnects other than half the width of the ones that define any of the West or East edges when the main routing is vertical (i.e., along North and South directions) can be present between facing edges.

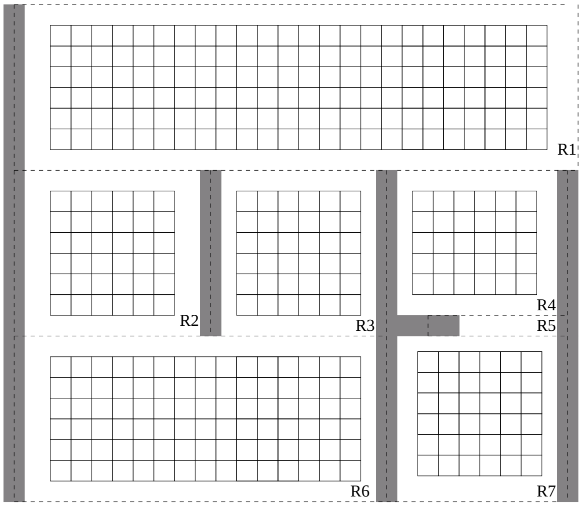


Figure III.1: Regions ($R1$-$R7$) are indicated by dashed lines and the grids of rectangles are shown in each region.

III.A.2 Converting a Region to a Grid

Converting a region into a grid, we can define possible grid rectangles for fill insertion. Each such region in Figure III.1 is converted into a *grid* after keep-off distances are stripped.[2] For each region, we form a grid of rectangles into which fills will be placed.[3]

A single region is shown in Figure III.2, where square boxes are the grid rectangles into which metal fills can be placed. Keep-off distances in the x and y

[2] A keep-off distance is a design rule which states the minimum distance of a fill to the nearest interconnect.

[3] Grids are preferable means for fill insertion due to the simplicity of algorithms and extraction flows, compressibility of data volume with GDS AREF constructs, controlled influence on wire line edge roughness and OPC side effects.

directions are indicated by Kx and Ky, respectively. The edges of the region are also indicated.

To obtain a grid, we define an *auxiliary frame* at the keep-off distance plus half the interconnect width away from the West and East edges and keep-off distance away from the North and South edges. The auxiliary frame forms the outer edge of the grid. The grid holds multiple rectangles in which fills may be inserted. The grid rectangles and the frame are connected together with *bonds*, which have adjustable energy values between them. Bonds that touch a grid rectangle are its *incident bonds*. These terms are illustrated in Figure III.3.

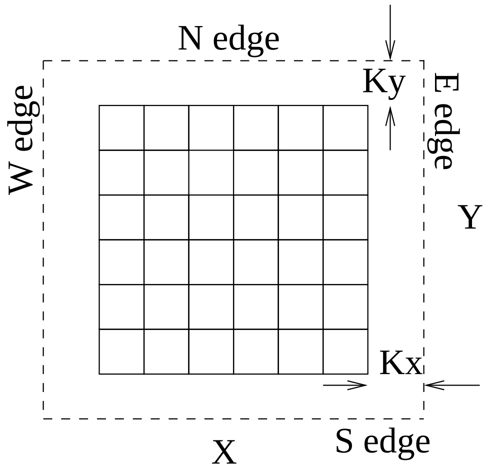


Figure III.2: A region is shown. A grid is formulated after keep-off distances are stripped out. There are 36 rectangles in this grid.

Converting Density To Fill Number. To convert target metal density to the number of fills to be inserted, we first compute the density in the grid:

$$D = (W \cdot A_r - A_i)/A_r \tag{III.1}$$

Here, $0 \leq W \leq 1$ is the target *feature* density assigned to a region. A_r and A_i are the area of the region and total area of interconnects in the given region, respectively. $0 \leq D \leq 1$ is the target *fill* density within the *grid*.

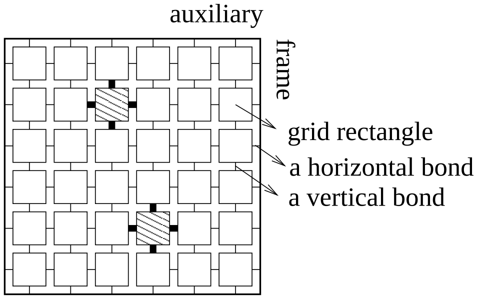


Figure III.3: An auxiliary frame (outer rectangle) holding grid rectangles (squares) with bonds (light lines) in between. Two filled rectangle locations (patterned) are shown along with the incident bonds used for energy calculation (bold lines).

Grid and Fill Size Selection. Assuming square fills, we compute the fill size, spacing and grid rectangle size as follows.

$$FW = (1+\epsilon)\sqrt{D} \tag{III.2}$$

$$FS = 1 - \sqrt{D} \tag{III.3}$$

In Equation (III.2), FW is a parameter proportional to the fill width. $0 < D < 1$ is the target fill density in a region. $0 < \epsilon < 1$ is an adjustment parameter, which helps select a fill size that will leave empty space in a region to enable movement of fills for improved optimization.

From Equation (III.3), FS is proportional to the fill spacing. For rectangular fills, both the proportionality constant and the fill aspect ratio are considered. Figure III.4 shows a diagram for the choice of fill width and fill spacing to accommodate the design rules. Design rule for minimum fill width ($FWmin$) and spacing ($FSmin$) are indicated by the vertical and horizontal lines, respectively. Minimum and maximum fill area (FA) rules are indicated by hyperbolas ($FAmin$ and $FAmax$, respectively). The valid region for the fill size and spacing choice is the shaded region. The line with the slope $m = FW/FS$ has constant density along the line.[4] To the left and right of the line are the lower and higher density regions, respectively. We choose FW and FS to lie on this line. We provide the final sizes that we have used in Table III.5. The number of fills to be inserted per region is then computed as $D * A_r$ divided by the area of a fill. Area of the grid rectangle is given as the square of the sum of fill width and spacing for a square fill.

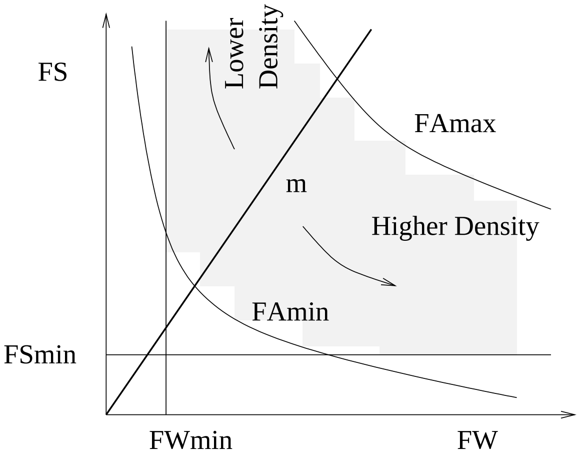


Figure III.4: Diagram showing fill choice to target design rules. Line with slope m is a constant density line.

III.A.3 Energy Modeling in a Grid

We have developed a parameterized model to choose bond values according to given fill insertion guidelines. In the following, vertical and horizontal bonds refer to bonds in the vertical and horizontal directions, respectively, in Figure III.3. Depending on a chosen guideline, fills may need to be inserted:

1. away from the interconnects, and

[4] CMP models are based on an average fill width and spacing per each window. Hence, choosing the same fill width and spacing across all regions yields a uniform height as long as the density requirements are also met.

2. centralized with respect to interconnects in a manner reminiscent of a hourglass shape.

Using 3D field solver experiments beyond those of [1], we have observed that these guidelines improve intralayer capacitance compared with other pattern guidelines. The former guideline seems to be more straightforward, as simply moving fills away from the interconnects reduces coupling. The intuitive reasoning for the second guideline seems to be that such a pattern pulls the flux in parallel to interconnects, thereby decreasing the coupling capacitance in certain interconnect configurations. We observe that the former guideline alone may be more beneficial for most general cases. Yet, pulling fills towards a given end of the interconnect, through a modification of the former guideline, may improve performance for longer interconnects. In a distributed capacitance model, capacitance at the end of the interconnect will be larger. Hence this may improve performance as suggested in [27].

To accommodate such guidelines as an example, we design our models to be as generic as possible. To implement these two guidelines in particular, horizontal bond energies start from 0 around the middle of the interconnects and increase as we approach the interconnects; energy minimization favors insertion closer to the center of the grid in the horizontal direction. For the latter case, additionally the vertical bond energies start from 0 around the middle of the interconnects and decrease as we approach the ends; energy minimization also favors fill insertion closer to the ends of the grid in the vertical direction.

The following generic models can implement these two particular guidelines with proper choice of their parameters. For the vertical bonds, the bond energy along the y-direction is given as

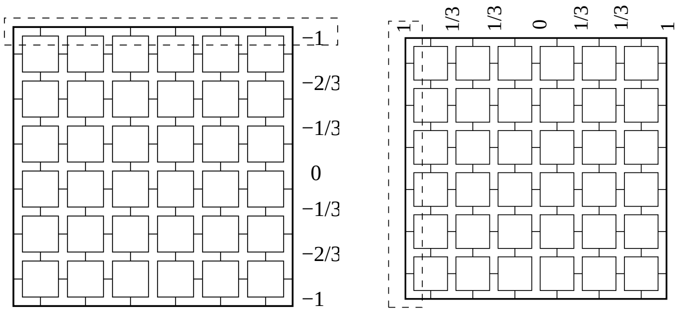
$$\gamma + (|(i_{mid}) - i| * \zeta/(i_{mid})) \tag{III.4}$$

where i is the row number starting from 0 and real number i_{mid} is the maximum row number divided by 2. For the horizontal bonds, energy along the x-direction is

$$\alpha + (|(j_{mid}) - j| * \beta/(j_{mid})) \tag{III.5}$$

where j is the column number starting from 0 and j_{mid} is the maximum column number divided by 2. α, β, γ and ζ are fitting parameters. The use of separate vertical and horizontal energy models allows flexibility in capturing different fill guidelines.

Figures III.5(a) and (b) show the energies assigned to bonds in the vertical and horizontal directions, respectively for an example assignment to satisfy the two guidelines. All vertical (horizontal) bonds aligned in the same horizontal (vertical) direction are assigned the same values in Figures III.5(a) and (b). The dashed boxes exemplify this assignment for a single horizontal or vertical alignment.



(a) (b)

Figure III.5: Example energies assigned to bonds. (a) Energies assigned to vertical bonds. (b) Energies assigned to horizontal bonds.

III.A.4 Insertion of Fills into a Grid

We map the fill insertion problem to an energy-minimization problem. Fills are placed in available locations to minimize an energy criterion. In Figure III.3, the rectangles are fixed in the grid, and hence the locations of bonds are fixed. We find the grid rectangle for fill insertion with minimum sum of energies for its incident bonds. We implement metal fill insertion using greedy optimization. We provide our fill insertion algorithm in Figure III.6.

The input to the algorithm consists of a grid on interconnect layer l with grid number i and the assigned number of fills NF to be inserted. The output is a grid with the placed fills. In the algorithm, t is the current grid rectangle location, t_{next} is the next grid rectangle, e is energy, b_t is the set of bonds incident to grid rectangle t, e_{min} is minimum stored energy, t_{min} is the grid rectangle location with minimum energy, and G_{li} is the grid itself. Lines 1 to 1.1.2.2 search for a grid rectangle with minimum energy, and the fill is inserted in Line 1.3. This algorithm is run for all regions identified by the scanline algorithm in a given metal layer. Although we describe the algorithm in this form for clarity, in the actual implementation, we first rank the grid rectangles in terms of their energies. We then insert fills using this ranked list. This eliminates otherwise repeated energy computations. Hence, runtime is approximately linear in the number of grid rectangles per region.

III.A.5 Summary

We summarize the overall algorithm in Figure III.7. Step 1.1 is handled using the scanline algorithm as described in Section III.A.1 for all interconnect layers l. This algorithm has complexity of O($nlogn$). Step 2.1.1 is handled in Section III.A.2. Steps 2.1.2 and 2.1.3 are handled in Section III.A.3. Step 2.1.4 is handled in Section III.A.4.

III.B Timing-Oriented Optimization

In this section, we describe timing criticality-aware and interlayer-aware fill optimization methodologies.

III.B.1 Critical Net-Aware Fill

Interconnect delay has been catching up with gate delay and even exceeding it in certain cases. While STA (Static Timing Analysis) with gate delays alone would be sufficient previously, sub-90nm technologies require accurate incorporation of critical nets into the STA. Hence, similar to critical gates being important in a critical path, critical nets are important during fill optimization and require

Table III.1: Critical Net-Aware Fill Flow.

1. Place, synthesize clock network and route design.
2. Extract SPEF parasitics from DEF.
3. Run static timing analysis using SPEF file from Step 2.
4. Use Perl scripts to obtain top critical net names.
5. Input net names to MFO for energy calculation.
6. Insert critical net-aware fills.

special attention. In fact, chip performance can be improved by optimizing the fills around the critical nets.

The first step in critical net-aware fill is to identify the critical nets. We handle this step by using the STA timing reports and extracting critical net names. Fills should be placed away from critical nets. In order to account for this, we update bond energies so that the energies favor locations away from the critical nets. Our example is updated in Figure III.8, assuming that the East edge of the region overlaps a critical interconnect. Essentially, j_{mid} is updated to

$$j_{mid} = j_{max} * [0.5 + (P_E - P_W)/2.0] \tag{III.6}$$

In Equation (III.6), j_{max} is the maximum column number for bonds, while P_E and P_W are *priorities* for East and West interconnect that may overlap on the East and West edges of a region. If the East or West edge does not overlap a critical interconnect, or overlaps a non-critical interconnect, the corresponding priority is set to 0, otherwise set to 1. In the case when both priorities are 1, Equation (III.6) gives the original definition of j_{mid}. Our critical net-aware flow is given in Table III.1.

III.B.2 Interlayer-Aware Fill

A valuable advance would be to achieve awareness of fills in neighboring layers. Traditional fill insertion rules are based on keep-off distances, which are

Table III.2: Interlayer-Aware Fill Flow.

1. Place, synthesize clock network and route design.
2. Extract SPEF parasitics from DEF.
3. Run static timing analysis using SPEF file from Step 2.
4. Use Perl scripts to obtain top critical net names.
5. Check critical nets on neighboring layers for each net.
6. Update energy values for bonds.
7. Insert interlayer-aware fills.

intrinsically based on intralayer coupling minimization. Interlayer awareness is a valuable performance upgrade for fill insertion.

The updates necessary to implement interlayer-aware fill are illustrated in Figure III.9. For each grid rectangle, upper and lower bonds are added so that each grid rectangle is *aware* of neighboring layer interconnects. These newly added bonds augment the list of incident bonds for each grid rectangle, to a total of six bonds.

The newly added bonds are assigned values as shown in Figure III.10. Essentially, if a grid rectangle overlaps an interconnect in the upper layer, its upper bond is set to 1. Furthermore, upper bonds for grid rectangles next to a predefined buffer zone away from the overlapped grid rectangles are also set to 1 to account for minimization of fringing effects. A selection of 1 reduces the change of such a grid rectangle to be selected for fill insertion.

Due to the number of neighboring layer nets to check when inserting fills, it is beneficial to restrict the number of nets to check in neighboring layers. Hence, to bound the fill insertion and optimization time, we use interlayer awareness only in the context of the critical net-aware flow. That is, we only check a neighboring net for overlaps only if it is a critical net. Our flow is given in Table III.2.

III.C Power-Oriented Optimization

Power consumption is a very important consideration in integrated circuits. Dynamic power increases as larger capacitances are charged and discharged every cycle. Power can be reduced by reducing the capacitances of power-critical nets.

Similar to the critical-aware timing optimization, we determine the power-critical nets and optimize the fills around them. A net becomes power-critical if its dynamic power is high. To determine such nets, we rank the nets according to their power consumption. We select a number of most critical nets and optimize the fills around these nets so that the capacitances seen by such nets are minimized. We use interlayer-aware fill for this purpose.

III.D Experimental Study

In the following subsections, we describe our experimental setup and results.

III.D.1 Experimental Setup and Protocol

Our experiments validate the efficiency and relevance of the proposed fill synthesis method by comparing layouts filled with the proposed method against those filled with a traditional fill method. We have implemented the proposed method using approximately 4000 lines of custom C++ code. Input is a GDSII (Graphic Data Stream) file converted into OpenAccess format [123], and output is the fill-optimized GDSII and DEF (Design Exchange Format).

In our flow, we use *Cadence SOC Encounter v5.2* [118] for placement, clock tree synthesis and routing, and *Synopsys Star RCXT 2007.06* [119] to extract post-fill parasitics. We compare layouts filled with the proposed scheme versus a *Blaze IF*-implemented [120] or *Calibre*-implemented [121] fill scheme representing traditional fills.[5] We then run the *Synopsys Primetime 2005.12* [122] static timing

[5] It is important to note that henceforth, while we use "Blaze IF" or "Calibre" to denote a

Table III.3: Analysis Flow.

1. Obtain filled GDS, e.g., from flow in Table I or II.
2. Use Perl scripts to obtain DEF containing fills.
3. Extract SPEF parasitics from DEF containing fills.
4. Run static timing analysis using SPEF file from Step 3.
5. Use Calibre scripts to obtain density histograms.
6. Use Perl scripts to obtain slack histograms.

Table III.4: Design Rules.

	Width	Spacing	MinDensity	Max Density	Min Area
$M2$-$M6$	$0.1\mu m$	$0.1\mu m$	0.15	1.0	$0.04\mu m^2$
$M7$-$M8$	$0.4\mu m$	$0.4\mu m$	0.20	1.0	$0.056\mu m^2$

analysis tool and compare the slack distributions reported by timing analysis. Our C++ code uses a subset of the *OpenAccess 2.2.4* API. For CMP prediction, we use Cadence CCP 1.4 with models optimized for TSMC and AMD 65nm processes.

Our analysis flow is given in Table III.3. For the optimized case, fills are inserted using our code. We use Blaze IF scripts for traditional fill insertion. We use unfilled GDS for the no-fill case. The $SPEF$ file on Line 4 is in Standard Parasitic Exchange Format, and the DEF file is in Design Exchange Format. For the power-oriented fill, instead of running static timing analysis, we run scripts to compute the interconnect switching dynamic power.

III.D.2 Results and Discussion

We have used TSMC 65nm GPlus 8-layer metal and AMD 65nm BEOL technologies in our experiments. Table III.4 provides the relevant design rules for the former. Table III.5 provides the relevant fill sizes. We have applied our tool

traditional fill pattern, this reflects only on the merits of the traditional strategy. The *Blaze IF* and *Calibre* tools themselves are industrial tools with powerful features.

Table III.5: Fill Sizes for Traditional Fill.

	Width (30%)	Spacing (30%)	Width (60%)	Spacing (60%)
M2-5	$0.22\mu m$	$0.12\mu m$	$0.52\mu m$	$0.1\mu m$
M6-8	$0.88\mu m$	$0.48\mu m$	$2.06\mu m$	$0.4\mu m$

Table III.6: Testcases Used in Experimental Validation.

Circuit	Source	#interconnect segments
S38417	ISCAS'89	43,076
ALU	opencores.org	72,423
AES	opencores.org	135,304
industrial	AMD	>50,000,000

to the ISCAS89 S38417 and OpenCores ALU and AES benchmark circuits using a $2.2GHz$ computer. These benchmarks contain 8 metal layers. S38417 contains 43,076 interconnect segments, i.e., straight pieces of interconnects between gates, ports and flip-flops which may be on different metal layers. Our program uses up to 160 MB of RAM for the S38417 benchmark. ALU and AES contain 72,423 and 135,304 interconnect segments, respectively. The testcases are summarized in Table III.6. We have applied fill optimization to these benchmarks (layers $M2$ through $M8$) and have compared the solutions against fills inserted traditionally, as well as with no-fill conditions, for different target densities, keep-off distances and optimization parameters.

Impact of Critical and Interlayer Awareness. Figure III.11 shows the worst 200 critical path setup slacks for the S38417 benchmark. *Slack* is a parameter which gives the difference between the actual signal arrival time at a flip-flop input or primary output, and the corresponding required arrival time. A large negative slack indicates that setup timing is not met by a particular critical path, and that hence either the clock period must be increased (i.e., frequency decreased), the clock skew between sequentially adjacent flip-flops must be adjusted, or the combinational cells on the critical timing path must be sped up. We present the results using histograms so as to show that not only a single path is modified by optimization, but the whole distribution shifts and its shape changes in most cases.

Figure III.11 shows the comparison between our tool MFO (Metal Fill Optimizer) using the density only (DO), intralayer critical-aware (CA), interlayer-aware (ILA) fill options and no fill case for 30% target density. To facilitate an equal comparison, we use a fixed target density, fill size and number. Total number of fills is 699,084 for all cases. Parameters b and z correspond to the model parameters β and ζ, respectively. Any unspecified parameters out of α, β, γ, ζ is taken to be 0. We take $\epsilon = 0.5$ by default unless otherwise stated, as we have seen that this value gives a reasonable tradeoff between final density uniformity and timing improvement. Density-only fill targets density optimization only and does not target performance. It uses the same number and size of fills per a region. We observe that intralayer critical-aware fill, CA, performs better than DO fill.[6] Also, interlayer critical-aware fill, ILA, performs much better than CA fill for performance. In fact, results with ILA are very close to the no fill case, where no impact on timing due to fill is present. All fill types result in the same final density uniformity, since the same fill sizes and numbers are used for all cases. We have observed that hold timing slacks were similar in all three cases. We use square fills in our analysis, although our tool can place rectangular and even more complex fill shapes.[7]

Setting $\beta = 1$ only, we can further improve the optimization as shown in Figure III.12 (note that the x-axis is scaled down as compared to Figure III.11). Table III.7 summarizes the power-oriented fill results. We use 100 power-critical nets for this analysis. Power is given for $1GHz$ operating frequency in the table. For large circuits, a non-negligible 2.4% total interconnect switching power

[6] The S38417 testcase contains many flip-flops with short critical paths. Hence the optimization results are highly optimistic. With longer critical paths, there will be more interconnects involved, not all of which can be optimized simultaneously. Fill optimization will result in smaller timing impact on average.

[7] Parasitic extraction tools have known inaccuracies with respect to floating fills. These inaccuracies may skew the results in terms of nominal values. However, a relative comparison should still be reasonably dependable.

Table III.7: Power-Oriented Fill Results.

Circuit	Power with fills	Power with power-oriented fills	% Improvement
S38417	1,746	1,121	35.8
ALU	6,337	6,185	2.4
AES	11,670	11,350	2.7

reduction is obtained with zero area loss.

Impact of Optimization Parameters. We next provide results to show how the fill shape and circuit performance depend on the optimization parameters. In Figure III.13, we show how the β parameter impacts performance. Choosing $\beta = 1$ pulls fills away from the interconnects. We also see in Figure III.13 how turning on the critical-awareness, thereby also pulling fills away from critical nets, improves performance.

Figure III.14 shows the impact of the ϵ parameter, which is indicated by e in the legend. A smaller value of ϵ (closer to 0) restricts the available space for the movement of fills during optimization. The fills end up being inserted closer to the interconnects, thereby increasing the coupling. A smaller value of ϵ decreases the fill width to spacing ratio. The fill size and spacing for the $\epsilon = 0$ case are 0.215 and $0.175\mu m$, respectively. The number of fills is 644,220. Notice that even though the number of fills has decreased, the performance worsens due to the proximity of fills to the interconnects.

Impact of Target Density. Setting the target density to 60%, the number of fills decreases and the fill size to spacing ratio increases. Overall, the performance degrades. Figure III.15 shows a comparison for the AES benchmark. The AES benchmark optimization runs in 30.25s.

Impact of Keep-Off Distance. We note that keep-off distance is the same as the fill spacing in our experiments. It is possible to see in actual designs that keep-off distances are up to 5 times the minimum fill spacing. Although originally introduced to reduce intralayer coupling, such a design rule is known to result in inferior density uniformity, as certain regions will not be able to be filled. As an example, increasing the keep-off distance to 2x for all layers degrades performance as shown in Figure III.16 for the ALU benchmark. The ALU benchmark optimization runs in 13.86s.

We provide snapshots from the S38417 GDS in Figure III.17(a) and (b). Fills are pulled away from interconnects and towards the interconnect ends in Figure III.17(a) and away from interconnects only in Figure III.17(b).

Comparison with Standard Methods. To compare our tool against industry standard tools, we have used the Blaze IF tool. The minimum fill (min-fill) feature of the tool results in a minimum number of fills being inserted for a given target density. The minimum variance (min-var) feature targets minimization of the variance in density across windows. We use the maximum fill feature and match the number of fills to MFO by random deletion per each window. The results are shown in Figure III.18. Min-fill and min-var conditions have the same fill size and spacing but slightly higher fill count (741,190 and 741,206, respectively). The results indicate that reducing the number of fills and thereby matching the number to those of MFO in the matched case improves performance for Blaze IF matched case; however, MFO performs better as the fills are also pulled away from each interconnect.

We have not used CMP-aware RC extraction in our analysis due to lack of such a setup. However, if such a setup is available, it can directly be used with MFO with no change. CMP-aware RC extraction would result in different critical lines. MFO can account for such changes by default during bond-value assignments.

We furthermore have not used a feedback loop based on CMP topography. Such a feedback loop would increase the runtime of the overall optimization. Simple solutions exist whereby regions with abrupt density gradients are identified and refilled with different fill size and spacings in a multiple-pass scheme. Our focus in this chapter has rather been improving performance given the same CMP topography with respect to traditional filling.

For the S38417 testcase, the runtime for the MFO-optimized fill is $172.43s$ including printing out the GDS file. The runtime decreases as fill size is increased, and as the number of grid rectangles decreases. Similarly, the runtime for interlayer-aware fill is $180.49s$ on a $2.2GHz$ computer. In comparison, Blaze IF takes $152.75s$ for the minimum variance fill on a $2.4GHz$ computer.

Final Height Uniformity. To evaluate CMP uniformity, we predict the final copper and interlayer dielectric (ILD) thickness topographies using the Cadence CCP tool. Pre-fill and post-fill copper height topographies (for layer $M4$ as an example) are shown in Figures III.19(a) and (b), respectively. Dimensions on the x and y-axes show the region size in μm. The copper height is shown in the legend in angstroms. The mean-normalized 3σ variation for copper height is 3.15% before fill insertion, with the distribution being roughly uniform. The differences for the topmost row and rightmost column with respect to the mean are, we believe, due to the computational inaccuracies in the tool when the window size is smaller than $20\mu m$.

The pre-fill and post-fill copper height histograms for the layer $M4$ are shown in Figures III.20(a) and (b), respectively. The inaccuracy mentioned above can also be seen towards the right in the histogram. Otherwise, the copper height distribution is very tight, with a 3σ standard deviation of $20nm$, which is 1.17% of the mean height. The distribution is closer to Gaussian as compared to the pre-fill case. Distributions in other layers are similarly tight.

An industrial testcase. We have demonstrated the applicability of our algorithms to an industrial testcase. The testcase is a dual-core microprocessor with a size of $8mm$ by $14mm$. Below, we report the summary for core1 of the testcase for layer $M6$. The core2 and core1 are symmetric to each other, hence we report only one of the cores in Table III.8. We perform the fill optimization using a server with two dual-core processors. We fill the testcase by dividing it into 16 horizontal stripes and filling the stripes in parallel. This reduces the runtime significantly in a multi-core system. If a single core system is used, the runtimes for each stripe would need to be summed to represent the total runtime. We observe that choos-

Table III.8: Runtimes for the Industrial Testcase Core1 $M6$ Layer.

stripe no.	1	2	3	4
no. of interconnects	609,910	535145	597,705	735,479
no. of regions	1,253,826	1,222,835	1,308,515	1,527,573
region ident. time (s)	2,177	1,811	2,166	2,776
filling and W/O time (s)	4.490	8.070	2.910	2.460
stripe no.	5	6	7	8
no. of interconnects	622,660	571,512	395,358	117,608
no. of region	1,411,223	1,304,571	776,952	324,552
region ident. time (s)	1,443	872.210	594.160	11.680
filling and W/O time (s)	8.930	11.380	11.240	25.720

ing the number of stripes as a small multiple of the number of processing cores is beneficial in terms of reducing runtime.

Figure III.21(a) shows the pre-fill copper height topography as output by Cadence CCP with a model optimized for the $65nm$ AMD BEOL process that is used for the testcase. Figure III.21(b) shows the corresponding histogram for the final copper height. Notice that the histogram shows a large variation.

Figures III.22(a)-(c) show the post-fill copper height topography for Calibre-filled traditional, matched and MFO fill, respectively. The matched fill case, i.e., MFO DO, uses the same number of fills per region and same fill sizes as MFO fill. The only difference is that MFO fills are optimized for placement in each region as compared to the matched fill case. Traditional fill also targets density. After setting fill size and spacing to yield a final density, fills are inserted as long as there is space to insert a fill. We use a 30% target metal density. As we do not have access to the critical net information, we do not enable optimizations to account for critical timing paths. We see that the predicted post-CMP topographies are similar between the traditional fill and the performance-oriented MFO fill.

For MFO, we have chosen $\beta = 1$ and $\zeta = 1$ as our optimization parameters to pull fills away from interconnects and towards the interconnect ends. We use fills of the same width as interconnects for tight regions, and fill widths and fill to fill spacings of 4x the interconnect size for sparse regions using MFO. We identify whether a region is sparse based on the number of initial rectangles in a grid. We set the limit to 10,000, and if the number of rectangles is more than this limit, we enlarge the size by 4x. This helps speed up the filling process significantly. For traditional Calibre fills, we set the fill spacing for 30% target density and keep the fill sizes the same. We utilize a 2-step fill methodology similar to the matched and MFO cases.

In our testcase there are circuit blocks to the right, where the interconnect orientation is orthogonal to the rest of the layer. Such regions should be filled at the block level before the full-chip integration to improve the filling performance. Otherwise, we observe that the target density is obtained reasonably closely as a result of our filling algorithm.

Figures III.23(a)-(c) show the corresponding histograms for the final copper height. The histograms this time show a primary peak for MFO filled and matched traditional fill cases as compared to that of Figure III.21(b), indicating that the copper heights are centered close to the target. We can see that the histograms are similar between the traditional and performance-oriented MFO fills. The traditional Calibre fills result in a smaller density range. However, the results show two peaks, which may not be desirable, e.g., in a statistical timing optimization context.

Input: An empty grid.

Output: Filled grid.

```
insertFills($l$,$i$,$NF$) {
[0] Initialize $t_{min} = 0$ $e_{min} = inf$
[1] do
[1.1]  do
[1.1.1]    $e = \sum b_t$
[1.1.2]    if $(e \leq e_{min})$
[1.1.2.1]     $e_{min} \leftarrow e$
[1.1.2.2]     $t_{min} \leftarrow t$
[1.1.3]    $t \leftarrow t_{next}$
[1.2]  until all $t \in G_{li}$ are evaluated
[1.3]  Place fill at $t_{min}$
[2] until all $NF$ fills are inserted }
```

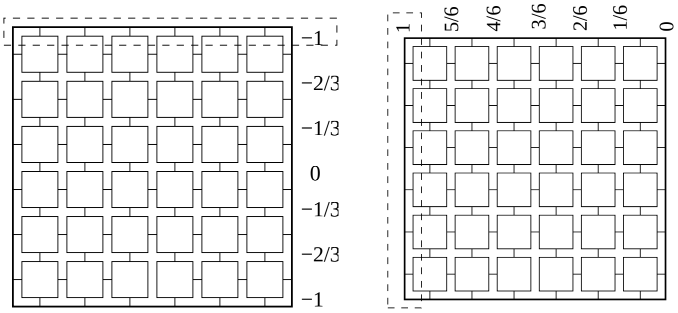
Figure III.6: Pseudo-code for fill insertion in a grid.

Input: GDS to be filled.

Output: Filled GDS (or DEF).

```
[1] forall $l$
[1.1]   identifyRegions($l$)
[2] forall $l$
[2.1]   forall $i$
[2.1.1]     generateGrid($l$,$i$)
[2.1.2]     $NF \leftarrow$ computeFills(l,i)
[2.1.3]     assignEnergies($l$,$i$)
[2.1.4]     insertFills($l$,$i$,$NF$)
```

Figure III.7: Pseudo-code for fill insertion in a GDS.



(a) (b)

Figure III.8: Example energies assigned to bonds when East edge overlaps a critical interconnect. (a) Energies assigned to vertical bonds. (b) Energies assigned to horizontal bonds.

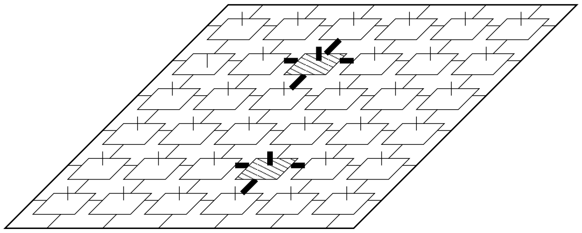


Figure III.9: Interlayer-aware feature. Additional bonds for upper interconnect layer awareness shown. There also exist bonds for lower interconnect layer awareness (not shown). Two fills are shown as patterned with dark incident bonds used for energy calculation.

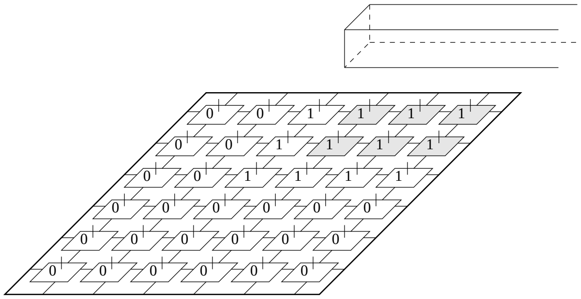


Figure III.10: Upper layer interconnect overlaps lightly shaded grid rectangles. Upper bonds are set to 1 for overlapped rectangles and rectangles within buffer distance from overlapped rectangles. Remaining upper bonds and all lower bonds are set to 0.

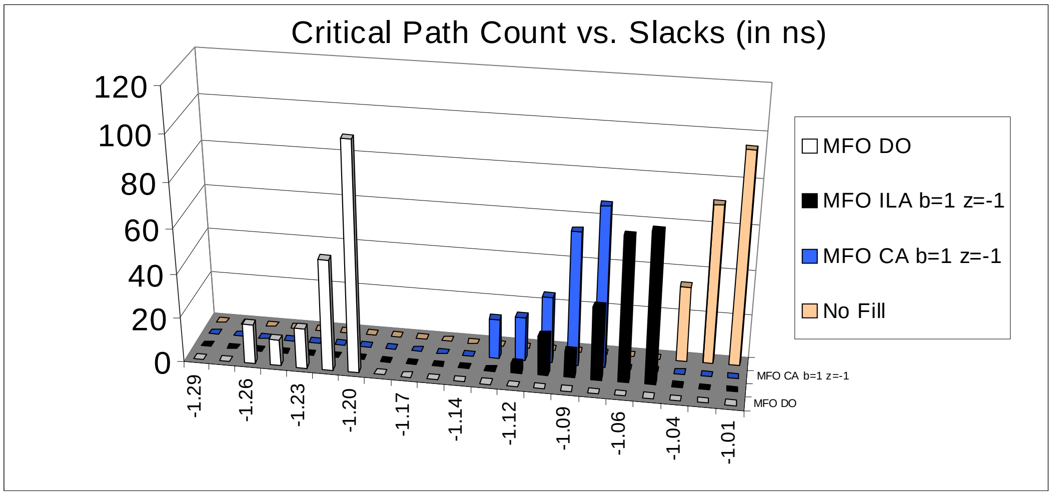


Figure III.11: Timing slacks for S38417 benchmark, for density-only fill, critical-aware, interlayer-aware fill and no-fill cases. 30% fill density.

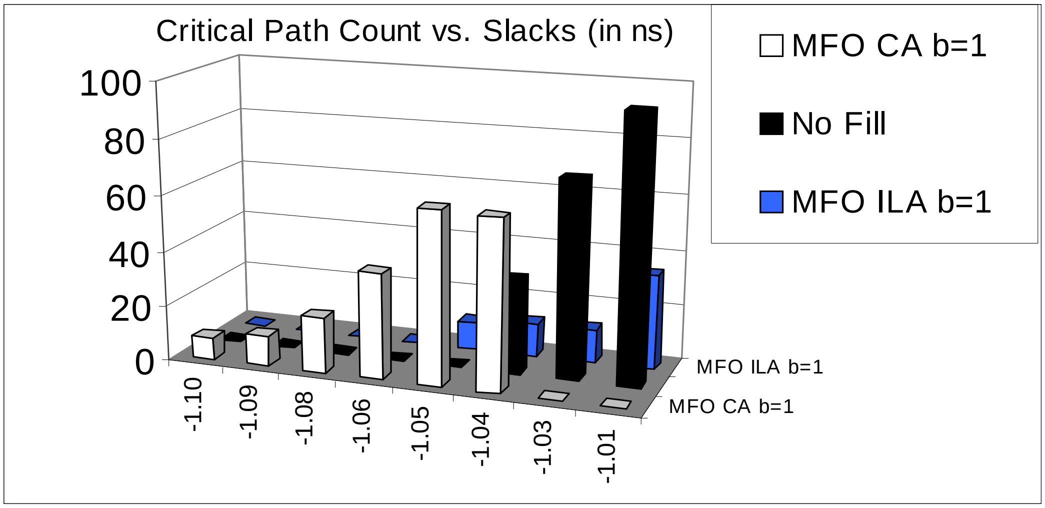


Figure III.12: Timing slacks for S38417 benchmark, for critical-aware, no-fill and interlayer-aware fill cases for $\beta = 1$ with 30% target density.

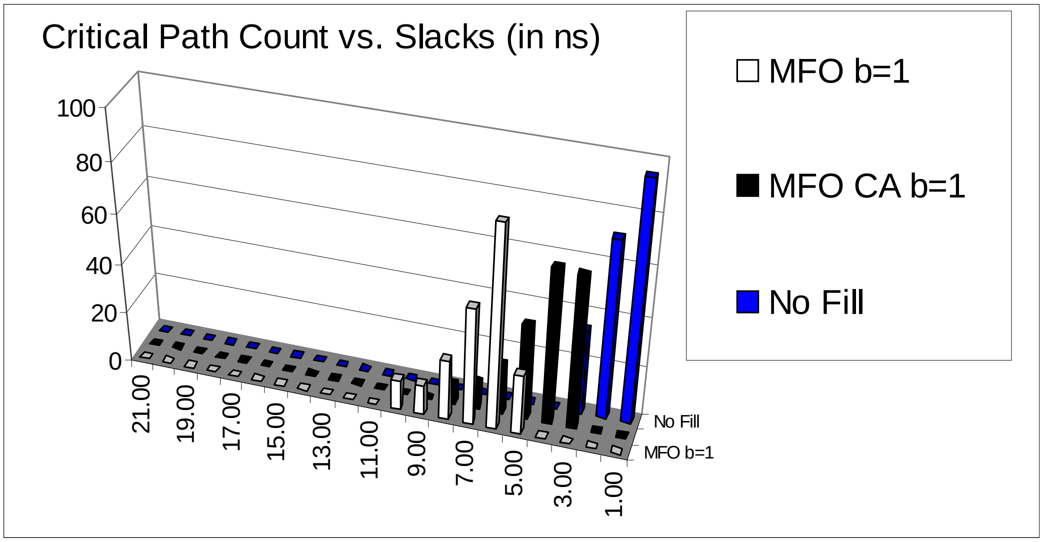


Figure III.13: Timing slacks for S38417 benchmark for various optimization parameters. 30% density case.

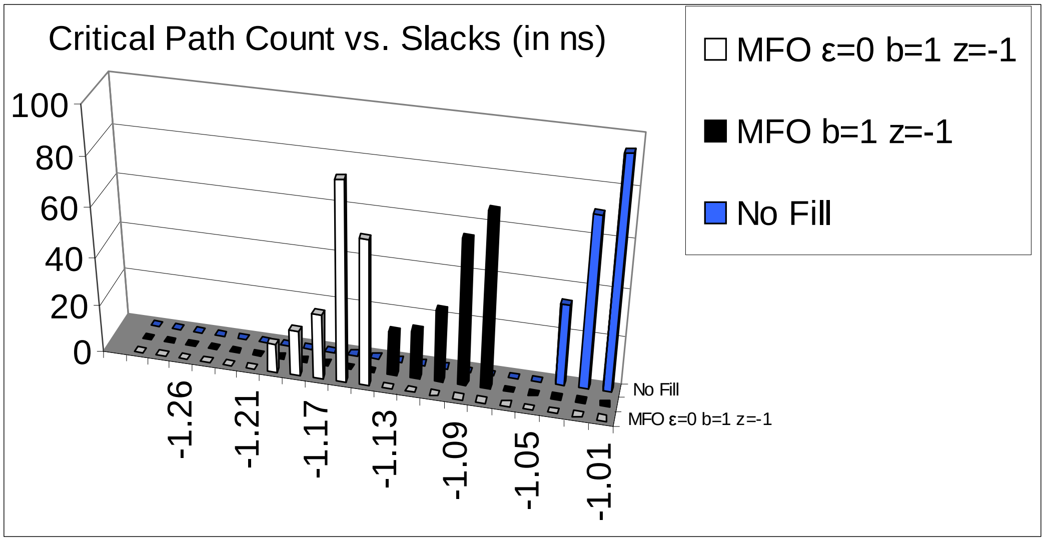


Figure III.14: Timing slacks for S38417 benchmark to show the impact of ϵ parameter. 30% density case.

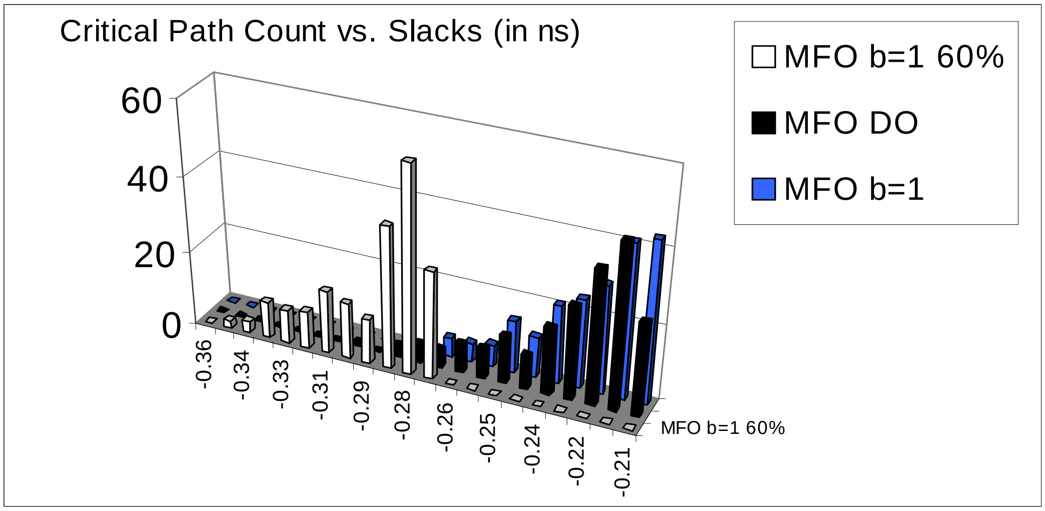


Figure III.15: Timing slacks for AES benchmark for 60% MFO with $\beta = 1$, 30% MFO density only, and 30% MFO with $\beta = 1$ cases.

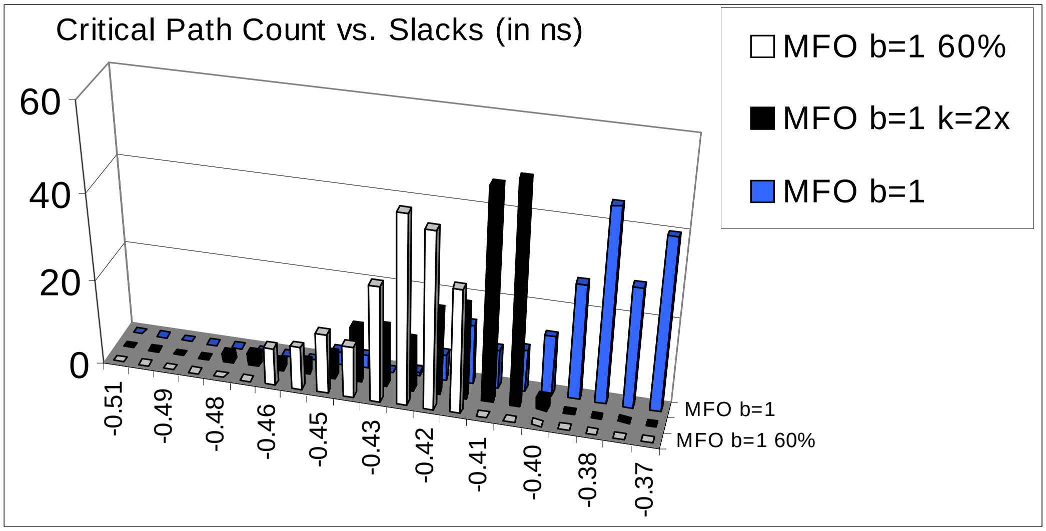
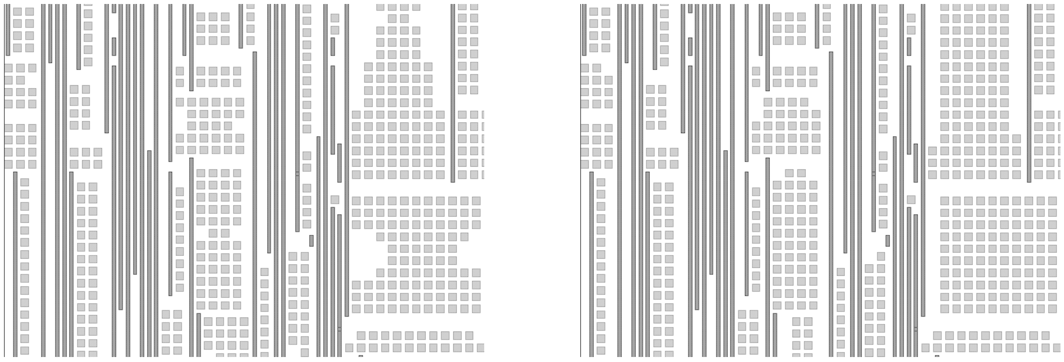


Figure III.16: Timing slacks for ALU benchmark for 60% MFO with $\beta = 1$, 30% MFO with $\beta = 1$ and 2x keep-off distance, and 30% MFO with $\beta = 1$ cases.



(a) (b)

Figure III.17: Snapshot of the 30% optimized fills in the S38417 benchmark layer $M4$. Window width is 10 μm at location (95,100μm). (a) MFO fill with $\beta = 1$ and $\zeta = -1$. (b) MFO fill with $\beta = 1$.

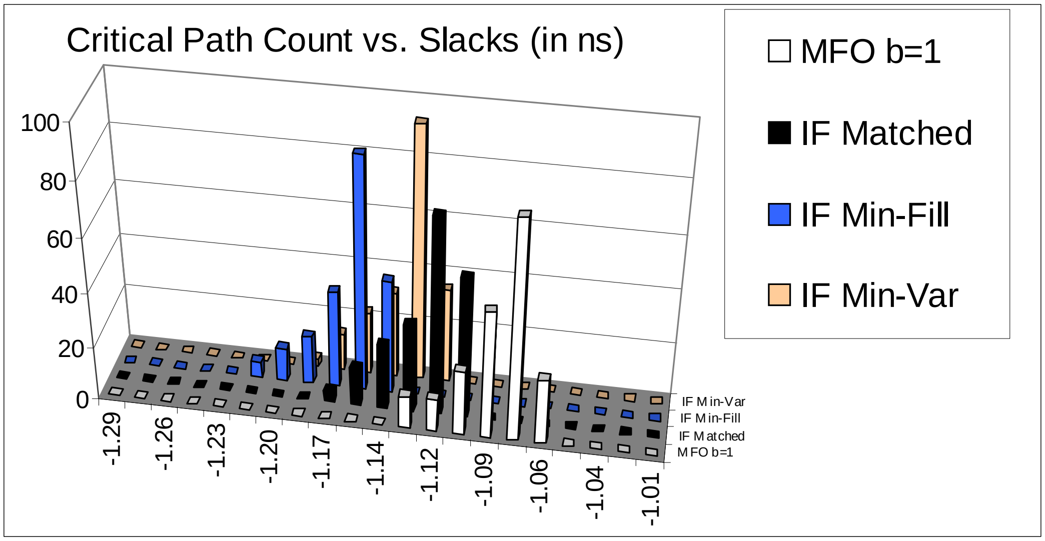
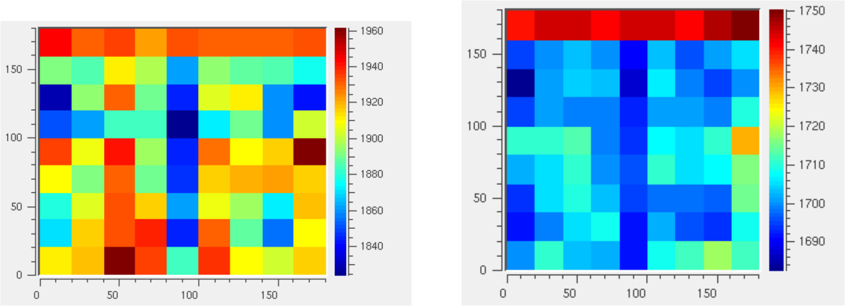
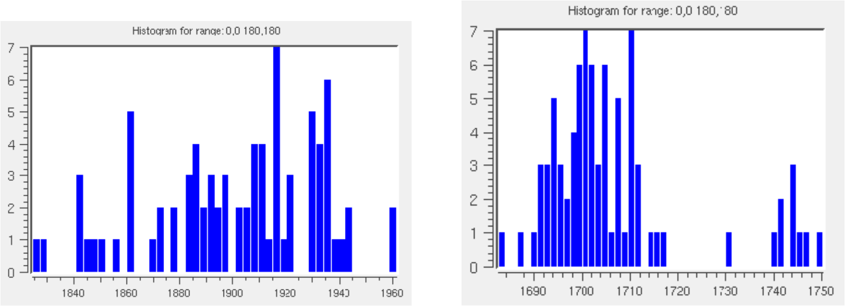


Figure III.18: Timing slacks for S38417 benchmark for MFO, Blaze IF matched to the fill number of MFO, Blaze IF min-fill, and Blaze IF min-var cases. 30% density case.



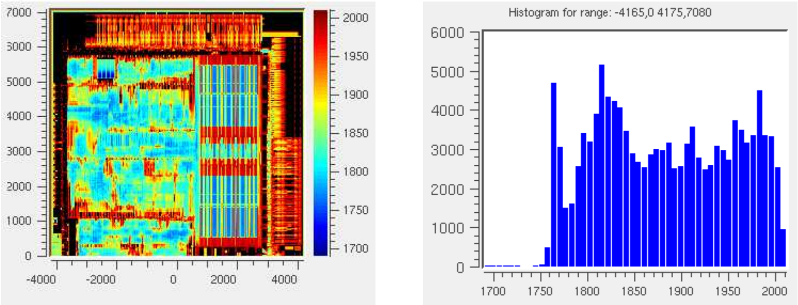
(a) (b)

Figure III.19: (a) Pre-fill copper height topography on layer $M4$. (b) Post-fill copper height topography on layer $M4$.



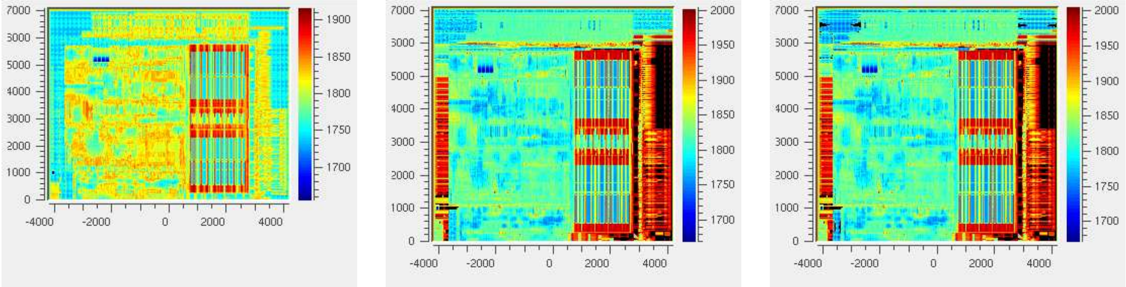
(a) (b)

Figure III.20: (a) Pre-fill copper height histogram. (b) Post-fill copper height histogram.



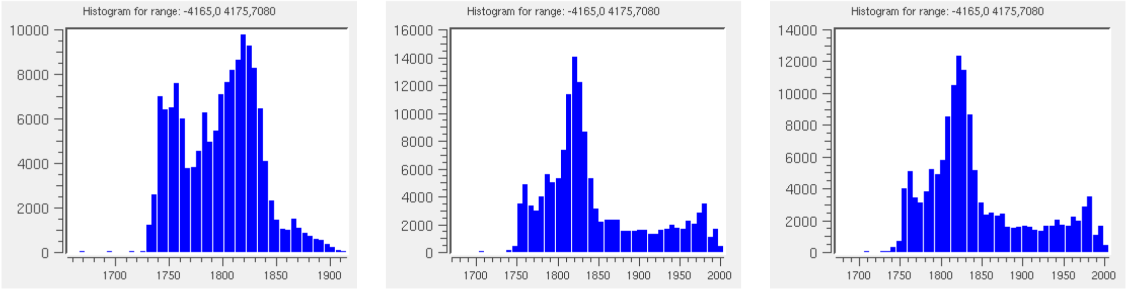
(a) (b)

Figure III.21: (a) Pre-fill copper height topography on layer $M6$ for core1 of the industrial testcase. (b) Pre-fill copper height histogram.



(a) (b) (c)

Figure III.22: Post-fill copper height topography on layer $M6$ for core1 of the industrial testcase for (a) traditional Calibre fill, (b) matched fill, and (c) MFO fill.



(a) (b) (c)

Figure III.23: Post-fill copper height histogram for (a) traditional Calibre fill, (b) matched fill, and (c) MFO fill.

Table III.9: Post-Fill Copper Height Topography Differences.

$\mu(M-U)$	$\mu(T-U)$	$\mu(M-T)$
171.1850A	174.1775A	2.9925A

$\mu(C-U)$	$\mu(C-T)$	$\mu(C-M)$
178.5232A	4.3457A	7.3382A

Table III.9 provides error differences for the copper topography random variables for unfilled (U), MFO filled (M), and traditional filled with Calibre (C) and matched traditional filled (T) layouts. To compute these differences, copper height difference at each grid is first computed, followed by the mean computation. Core1 contains 384 by 354 grids for the CMP analysis. As can be seen from the mean values, the mean difference between MFO filled and traditional filled layouts is only about 3 and 7 Angstroms for matched and Calibre versions, respectively.

Figure III.24 shows the histogram for the percentage reduction in total capacitance as we use MFO fill as opposed to matched traditional fill, the topography and histograms of which are shown in Figures III.22 and III.23, respectively. In Figure III.24, the y-axis is the number of nets and the x-axis shows the bins as a range of percentage capacitance reduction. For example, 10-20 means that the corresponding bin has nets with percentage capacitance reduction between 10 to 20%. Hence, even though the histogram and the topography stays similar, there is a total capacitance reduction observed with the MFO fill.

Figure III.25 shows the histogram for the percentage reduction in total capacitance as we use MFO fill as opposed to traditional fill, the topography and histograms of which are shown in Figures III.22 and III.23, respectively. In Figure III.25, the y-axis is the number of nets and the x-axis shows the bins as a range of percentage capacitance reduction. Even though the histogram and the topography stays similar, there is a total capacitance reduction observed with the MFO fill with respect to traditional fill.

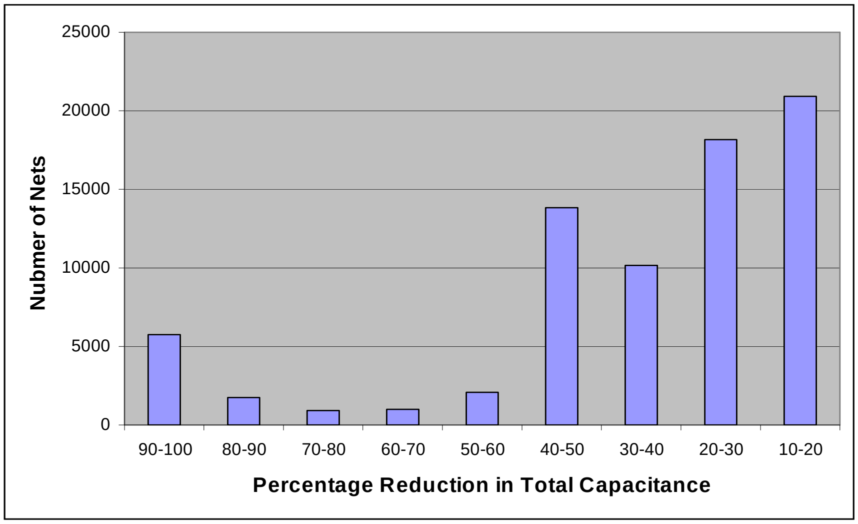


Figure III.24: Histogram for the percentage reduction in total capacitance as MFO fill is used instead of traditional matched fill. Total number of nets is 4,185,361.

III.E Conclusions

In this chapter, we have introduced a fill synthesis framework that heuristically accepts and realizes complex, performance-driven CMP fill guidelines. Our heuristic approach uses an energy minimization framework to achieve metal fill insertion in a given region. We have presented a parameterized bond-energy model, along with insights to guide energy modeling. With the proposed framework, complex guidelines can be implemented efficiently. We have extended our tool to optimize for critical-aware and interlayer-aware cases. We have achieved 2.5% reduction in total interconnect switching power using power-aware fill insertion. We have also obtained results for circuit timing optimization. Experimental results with the proposed optimization methods show that with 65nm testcases, it is possible to reduce the fill impact on coupling capacitances by up to 96.10% for 30% pattern density, and by up to 15.64% for 60% pattern density. We have shown the applicability of MFO on an industrial dual-core microprocessor testcase

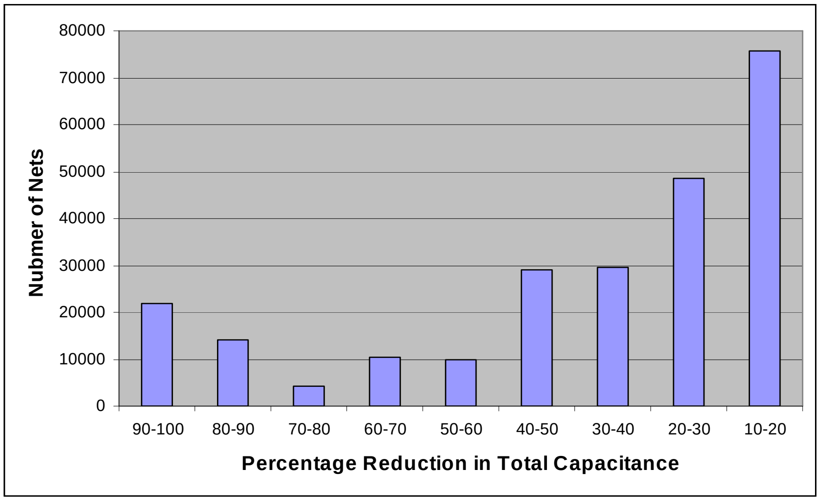


Figure III.25: Histogram for the percentage reduction in total capacitance as MFO fill is used instead of traditional Calibre fill. Total number of nets is 4,185,361.

and shown that while maintaining CMP uniformity similar to that achieved by traditional fills, the capacitances decrease.

III.F Acknowledgments

This chapter is based on or in part a reprint of the following publications.

- A. B. Kahng and R. O. Topaloglu, "Performance-oriented interlayer-aware CMP fill pattern optimization," *under review in IEEE Trans. on Computer-Aided Design*, 2008.
- A. B. Kahng and R. O. Topaloglu, "Performance-aware CMP fill pattern optimization," *Proc. International VLSI/ULSI Multilevel Interconnection Conference (VMIC)*, **Invited Paper**, 2007, pp. 135-144.
- R. O. Topaloglu, "Energy-minimization model for fill synthesis," *Proc. IEEE International Symposium on Quality Electronic Design*, 2007, pp. 444-451.

IV

FEOL Stress Modeling and Optimization

Stress engineering is being used starting with $65nm$ and onward to improve channel mobility and hence device performance. There is an increased need for modeling of stress at a level appropriate to drive design optimizations. Furthermore, there is a need to improve the circuit performance through layout optimizations by utilizing all stress sources. In this chapter, we target these objectives.

Shallow trench isolation (STI) is an important and well-studied stress source that has not been fully exploited until now for design quality improvement. STI usually exerts a compressive stress along the channel (i.e., the current flow direction), which improves PMOS device mobility. The opposite type of stress, tensile stress, degrades the PMOS performance in this direction. NMOS is in general complementary to PMOS in terms of how it is affected by stress, and its mobility degrades because of STI stress.

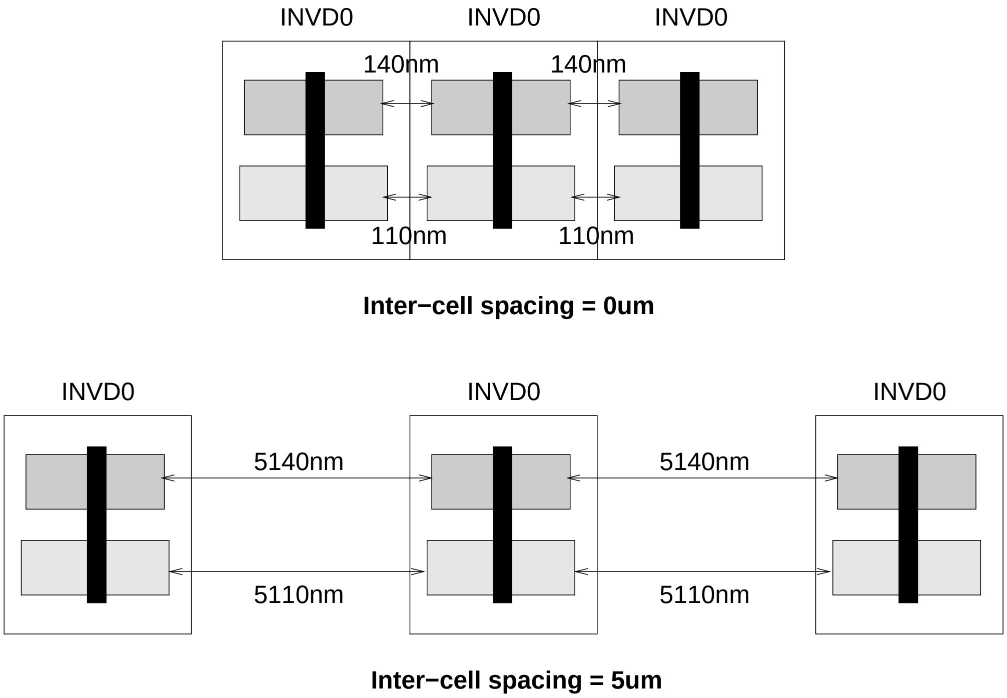
Device mobility increase corresponds to switching speed increase. Hence, it is possible to utilize STI, which is used to separate NMOS and PMOS regions, to improve performance. Table IV.1 shows the impact of STIW on rise and fall delays (averaged over all timing arcs) of several 65nm standard cells using the models developed in this chapter. Figure IV.1 illustrates the change in STI width for the INVD0 cell when the inter-cell spacing is increased from $0\mu m$ (i.e., abutting neighbors) to $5\mu m$. Impact of placement on STI width and consequently on rise (*R-Delay*) and fall (*F-Delay*) delays for a few cells are provided. For each cell in the table, three instances of the cell are placed with different spacings between them, and the delay of the center instance is reported. In Table IV.1, *Spacing* is the spacing between cells and *PMOS* $STIW_L$ (*NMOS* $STIW_L$) and *PMOS* $STIW_R$ (*PMOS* $STIW_R$) are the STI widths next to the left and right sides of p-type active (n-type active) regions of the center cell. It is possible to both speed up and slow down cells by controlling the STIW and, thereby, the stress that is applied to a cell. In particular, larger STI width will generate more stress in neighboring transistors. In this chapter, we propose placement perturbation and the insertion of active-layer fills to control the STI width in a performance-driven manner.

The proposed active-layer fill insertion and placement perturbation do not require additional process steps or add complications to resolution enhancement techniques. Active-layer fill insertion is a standard process step that is performed in all designs to control active-layer density. Placement perturbation yields a new valid placement. We ensure that the design is design rule correct after we perform these two steps.

Organization of the chapter. The remainder of this chapter is organized as follows. In the next section, we describe our STI width-induced stress models that we have developed. We then review the process steps we have simulated using TCAD tools, and our proposed stress models. In Section IV.B, we present our STI stress-aware timing analysis approach. Section IV.C describes our timing optimization methodology. In Section IV.D, we present our circuit-level optimization results. We conclude the chapter with Section IV.E.

IV.A STI Width Stress Modeling

This section describes a generic STI process flow, along with the STIW models we propose. The STIW parameter is as shown in Figure IV.2.



Inter−cell spacing = 0um

Inter−cell spacing = 5um

Figure IV.1: Change in STI width with increase of inter-cell spacing from $0\mu m$ to $5\mu m$.

IV.A.1 Process Steps

The process recipe that we use for simulation of STI stress is summarized in Table IV.2. We have simulated the structure up to the gate deposition step using the Synopsys Sentaurus 2005.12 process simulator.[1] We make the following observations.

- We use a high mesh density, especially between the STI and underneath the channel, to obtain accurate finite-element calculations close to the channel.

- Temperature cycling (Steps 7, 17, and 19) and densification steps (Steps 10-12) are responsible for the stress build-up. Due to visco-elastic material behavior, materials cannot recover to their original state after stress is

[1] When foundry models are used, the exact process steps are not known except for hints provided in the literature or various collateral documentation. Foundries should provide STI width impact models in such a scenario.

Table IV.1: Impact of STI Width on Performance of Several Standard Cells.

Cell	Spacing	PMOS	PMOS	NMOS	NMOS	R-Delay	F-Delay
		$STIW_L$	$STIW_R$	$STIW_L$	$STIW_R$		
	(nm)	(nm)	(nm)	(nm)	(nm)	(ps)	(ps)
INVD0	0um	140	140	110	110	27.27	21.96
	5um	5140	5140	5110	5110	23.65	23.70
BUFFD0	0um	140	140	125	125	45.56e	46.11
	5um	5140	5140	5125	5125	43.84	43.53
NR2D0	0um	140	140	110	110	51.12	23.06
	5um	5140	5140	5110	5110	42.77	24.69
ND2D0	0um	140	140	110	110	29.63	35.36
	5um	5140	5140	5110	5110	25.77	38.81

withheld. Thermal cycles result in stress due to thermal mismatch between different materials, which have different thermal expansion coefficients. As a result, stress builds up in the STI oxide and this stress remains there even at room temperature at the end of the process. Final stress shows its effects all the way into the channel of neighboring transistors, in a space-dependent trend, during the lifetime of a chip.

- In Step 14, STI CMP is applied. At the end of this step, the top of the STI is left above the active region on purpose. The basic reason is to avoid defectivity such as delamination of the STI oxide. At the edges of the channel, this step height difference would introduce threshold voltage variations and so-called “width effects” [3].

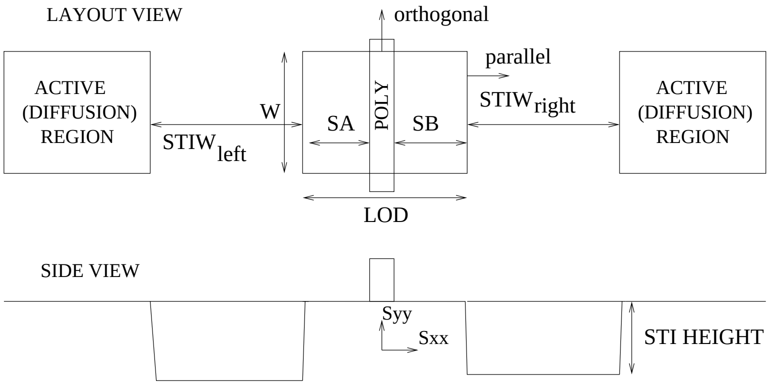


Figure IV.2: STIW parameter. LOD is accounted for in BSIM models. STIW impact is not modeled. Parallel and orthogonal distances with respect to a transistor are also indicated in the figure.

IV.A.2 STI Stress Modeling

The popularly used BSIM SPICE model (revision 4.3 and higher) contains an explicit STI model. However, only the impact of the distance from transistor channel to the STI boundary is modeled. Hence, the dependency on the STI width is not present in the BSIM4 model. Our simulations, as well as simulations and data in the literature [81], show that STIW impact cannot be neglected. Thus, as noted above, our present work not only models STIW impact, but also builds upon this modeling to improve circuit performance at no area cost.

The STI impact as a function of length of diffusion (LOD) (refer to Figure IV.2) is already incorporated into the BSIM4 model. Our objective is to isolate and correct for the impact of STIW, in a manner that can be applied on top of existing BSIM4 stress modeling. Using 2D simulations, we have developed the model given in Equations (IV.1)-(IV.4) to capture the STIW effect in the parallel direction (shown in Figure IV.2). The LOD parameter still appears in the equation, as the STIW impact differs according to the choice of LOD. Also, for purposes of this chapter, we do not require or discuss STIW impact modeling in the orthogonal direction (shown in Figure IV.2), as the STI width effects are blocked in the orthogonal directions by active regions for the type of standard cells we have used. At the end of TCAD simulations, we obtain stress values in Pascals. We then convert the stress values to mobilities using the methodology in [64] and normalize the mobilities. The NMOS equation is given as:

$$MOB_{L,R} = \zeta + (1 - (STIW_{L,R}/2)^{\alpha})/S\{A,B\}^{\beta} \tag{IV.1}$$

$$MOB = [MOB_L * MOB_R]^{0.26} \tag{IV.2}$$

In Equation (IV.2), MOB is the mobility multiplier. Parameters L and R indicate left and right directions with respect to the channel. The equation states that the final mobility multiplier (i.e., MOB) is the product of the mobility multipliers from the left and right directions (i.e., MOB_L and MOB_R). The PMOS equation is given as:

$$MOB_{L,R} = \zeta + ((STIW_{L,R}/2)^{\alpha})/S\{A,B\}^{\beta} \tag{IV.3}$$

$$MOB = [MOB_L * MOB_R]^{0.14} \tag{IV.4}$$

The model and data comparison is shown in Figure IV.3. In the figure, the x axis is a given data point in the DOE, i.e. a given SA, SB, $STIW_L$ and $STIW_R$ combination, and the y axis is the mobility multiplier. The models provide an average of 7.5% accuracy with respect to the data, along with preserving physical intuition.

The physical intuition is simply based on a relational understanding of how distance to a stress source impacts the stress depending on the type of transistor. As $S\{A,B\}$ increases in Equation (IV.3), the stress source becomes farther away from the channel. Hence mobility should decrease for PMOS as STI has compressive stress in this technology. This is captured in the equation. Increasing STIW in the Equation (IV.3), should improve the mobility. This also can be seen from the equation. To the contrary, NMOS equation, Equation (IV.1), has these physical terms to yield the opposite behavior. The mathematical constants are used to enable a fitting for the particular process technology.

When BSIM models are being generated from test structures, assuming that device modeling test structures with worst-case corners for STI width are present, the slow N - slow P corner would correspond to data from test structures with large STI widths near NMOS modeling devices, and test structures with minimum STI widths near PMOS devices. Our normalization step should consider the slow corner conditions as described herein, as well as the presence of STI models in BSIM models. The latter is necessary to eliminate any possibility of the double counting for the channel to STI edge stress, which is already covered by the BSIM models.

In order to enable an accurate normalization, we have set the slow N or P devices to a mobility of 1. Other STIW values result in an increase in the mobility. As multiple SA and SB values which can share the same STIW are present in the DOE, there are multiple devices which have mobilities of 1 after normalization. The models are then fit to the normalized results of the process simulation DOE. The corresponding parameters of the model are given in Table IV.3.

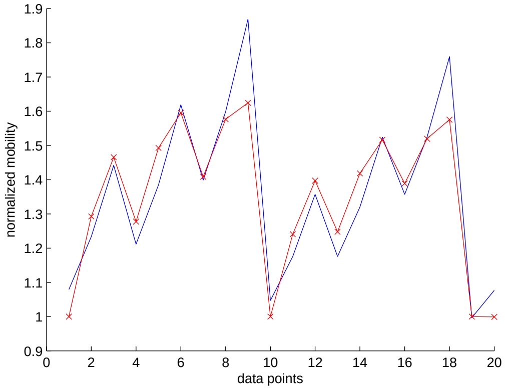


Figure IV.3: Model vs. data. Plot with crosses corresponding to the data. Each data point corresponds to a DOE instance.

Consideration of Non-rectangular Active Regions. For non-rectangular active regions, an average distance to the channel from STI boundary and for STIW can be computed as follows:

$$S\{A,B\} = \frac{\Sigma_i w_i * S\{A,B\}_i}{W} \tag{IV.5}$$

$$STIW_{L,R} = \frac{\Sigma_i t_i * STIW_{L,Ri}}{W} \tag{IV.6}$$

In Equations (IV.5) and (IV.6), i is an enumeration over active region edges, w_i and t_i are the width of each such edge parallel to the channel for distance from channel to the active edge and STIW, respectively. The parameter W is the width of the channel. Similar width-weighted averaging is used by BSIM 4.3.0 for calculation of stress effects without explicitly using parameters to capture irregular active region shapes.

A related discussion pertains to the choice of active diffusion fill widths. Fringing effects due to STI next to PMOS may degrade NMOS speed. However, we consider the fringing effects to be small because NMOS and PMOS are separated by hundreds of nanometers.

STI CMP. Traditionally, active region fills are usually inserted at the tape-out stage to minimize active region density variations. When active region fills need to be utilized for performance, a number of observations needs to be made. A buffer distance can be determined, such that above this distance, the stress width has no further effect on channel. This distance can be selected as $10\mu m$ for the process we have studied for example. It is advisable that such post-fill insertion algorithms use such buffer distances and do not insert fills inside buffer regions near critical gates. Furthermore, these fills are inserted using a window-based scheme to minimize density variations across windows. With optimization, fills will be inserted next to NMOS but not next to PMOS. Hence, there will be approximately 50% active region fill in an optimized window. If the density is lower and needs to be increased, then users can start to insert fills near non-critical gates as well. This approach will have negligible impact on timing closure.

Impact of Stress Liner. To evaluate the impact of stress liners on STI stress, we have used a similar simulation setup with the parameters given in Table IV.4. A nitride liner is shown in Figure IV.4. The nitride liner height and intrinsic stress are varied to observe the influence on STI stress. The parameters are shown in Figure IV.2. Combinations of all parameters are simulated individually. We have observed that, in the given parameter range, the addition of a $1GPa$ $0.1\mu m$ thick stressed nitride layer increases the impact of STI by 9.9% in terms of an average stress under the channel. A $0.2\mu m$ stressed nitride layer, on the other hand, increases the STI width impact by 12.9%. A $2GPa$ liner can increase impact by 10.3%. These changes seem negligible and also indicate that STI stress width effect will still be important across such process variants. The 10% improvement based on $65nm$ silicon data reported in [81] supports our findings, i.e., STI stress width effect is still important even in the presence of stressed nitride liners. The nominal stress values usually increase for the Syy, which is shown in Figure IV.2, as this component is in the direction of the nitride liner. The Sxx stress component typically reduces with these changes.

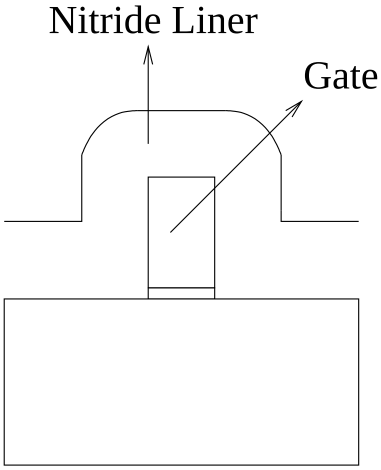


Figure IV.4: Stressed nitride liner as mobility enhancer.

Impact of STI Height. Parameters used to analyze the impact of STI height on stress are shown in Table IV.5. The height of the STI trench is changed to see the impact of STI height on stress. We have observed that increasing the STI height can result in a reversal of the STI stress, i.e., changing it from compressive to tensile for the Syy component. This observation is in line with what has been presented in [116]. Comparison of average stress values shows a reduction of up to 6% for the STI width impact due to variation of STI height. As with stress liner impact, we believe that our conclusions below remain valid across STI heights used by different processes.

Analyses and optimizations proposed in the rest of the chapter are not tethered to the models developed in this section, and can be used with other models after appropriate modifications. For example, there are known STI processes which may induce tensile instead of compressive stress. This may be due to STI trench height, or material and thermal processing differences, such as HDP (High Density Plasma) CVD (Chemical Vapor Deposition) as used in [117]. The optimization procedure presented below can be adopted to such a scenario with minor changes. Furthermore, the proposed models show monotonic response with respect to the STI proximity and widths. This results in an optimization scheme where maximum or minimum allowed dimensions will improve performance. With models that are non-monotonic, the optimization algorithm would need to be altered to provide an optimal solution.

Even though new mobility enhancement techniques show substantial variation in mobility, STI has been a mainstream methodology for over a decade. STI processes are much better controlled than the new stress engineering methods. Hence, there is much smaller observed variability in the STI mobility impact. Lithography effects will show negligible impact on the mobility change due to STI, as long as the layout is designed considering DFM rules such as using regular active regions without unnecessary corners. Finally, since active-layer fill insertion is a knob that we propose to exploit, we comment on additional process considerations for active-layer filling.

- There exist design rules which restrict the maximum active layer density, with this constraint arising for reasons of STI CMP uniformity. Such rules must be observed.
- Insertion of active-layer fills can potentially increase the total capacitance of inter-cell $M1$ routing, and may induce additional RCX modeling and charac-

terization for a given process technology. However, our methodology should not affect extraction of intra-cell $M1$ routing. As our active-layer fills are floating, their impact is also smaller.

- Reduced STI widths may slightly increase leakage between NMOS and PMOS transistors as well as between devices of the same type. However, the active to active design rules are typically set such that this leakage is minimized to a negligible level.

IV.A.3 DSL Stress Analysis

Stressed liner technologies improve transistor mobility by depositing a stressed nitride liner instead of a neutral liner on top of the gate and spacers of a transistor. Depending on the technology, either a compressive liner on top of PMOS devices or a tensile liner on top of NMOS devices can be used, with the remaining device type having a neutral liner on the top. The liner can be neutralized by doping. Dual stress liner technology, on the other hand, uses both types of stressed liners and targets to improve both NMOS and PMOS mobilities at the same time.

We illustrate the DSL technology in layout and side views in Figure IV.5. Usually, at least two masks are required, one defining the compressive regions and the other defining the tensile regions. Looking at the layout view, a PMOS transistor is present on the left and an NMOS transistor is present on the right. In the side view, we can see that PMOS is covered with a compressive liner and NMOS with a tensile liner. The boundaries of these liners are defined by the dashed lines in the layout view. The dashed lines correspond to the edges of the masks that define the compressive and tensile regions. Whenever a compressive region ends, a tensile region starts. A small overlap is possible at the boundary of compressive and tensile regions to avoid collapse of the dielectric to be deposited on top of the liners.

A region may contain multiple number of similar type of transistors. To demonstrate this option, we have left the boundaries open. As an example in standard-cell design, usually alternating two rows of PMOS and NMOS devices are present. There could be other PMOS and NMOS devices that exist towards these open boundaries. The boundaries usually meet on top of STI unless there is a specific design need for additional optimization.

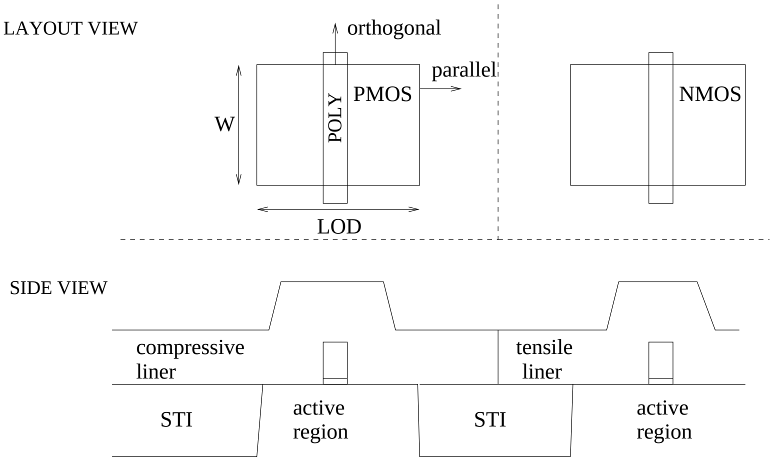


Figure IV.5: Illustration of dual stress liner. PMOS and NMOS transistors have compressive and tensile nitride liners, respectively, above their polysilicon gates.

DSL Process Flow. In order to simulate the effects of DSL, we have used the 65nm process flow in Table IV.6 on an SOI (silicon on insulator) wafer. In Table IV.6, we have implicitly indicated the temperature ramp up or downs by changing the temperatures in a consecutive process step. In Step5, we deposit the tensile liner to cover both devices. Tensile liner is etched away from PMOS devices in Step6 using a mask. We then deposit the compressive liner on top of PMOS devices using a mask. The actual process continues with dielectric deposition, contact etch and deposition and metal deposition steps. We bring the device to the room temperature and measure the channel mobility 10nm below the gate oxide in the middle of the channel. Notice that we have used multiple deposition and etch steps in Steps5-7. We have done this to ensure more accurate stress computations. We have selected the multiple step number as 5 after observing that further increasing the step number did not change the results.

We provide the process parameters for the DSL flow in Table IV.7. We assume that there is a built-in compressive polysilicon stress of -500MPa. We assume that the required compressive liner stress is twice that of the tensile liner to attain mobility benefits in both NMOS and PMOS.

Table IV.2: STI Process Steps.

1. Deposit pad oxide
2. Deposit pad nitride
3. Deposit photoresist for STI lithography
4. Anisotropically etch nitride and oxide
5. Strip photoresist
6. Directional etch at a rate of $0.01 \mu m/s$ for 40 seconds at 86^o angle
7. Ramp up temperature to 600^oC
8. Deposit TEOS oxide at a rate of $0.01 \mu m/s$ for 20 seconds
9. Deposit trench fill oxide
10. Temperature ramp up from 600^oC to 1000^oC at a rate of $50^o \ C/min$
11. Hold temperature for 1 minute at 1000^oC
12. Temperature ramp down from 1000^oC to 600^oC at a rate of $50^o \ C/min$
13. Diffuse oxide
14. STI CMP
15. Etch nitride isotropically at a rate of $0.015 \mu m/s$ for 15 minutes
16. Etch oxide isotropically at a rate of $0.02 \mu m/s$ for 1 minute
17. Temperature ramp up from 600^oC to 800^oC at a rate of $40^o \ C/min$
18. Diffuse for 5 minutes using dry O_2
19. Temperature ramp down from 800^oC to 600^oC at a rate of $40^o \ C/min$
20. Deposit polysilicon gate
21. Ramp down to room temperature

Table IV.3: Model Parameter Table.

	ζ	α	β
NMOS	1.03	0.076	0.48
PMOS	0.49	0.48	0.57

Table IV.4: Liner Stress Analysis DOE Parameters.

SA	SB	$STIW_L$	$STIW_R$	$LinerHeight$	$LinerStress$
0.2,4μm	0.2,4μm	0.1,2μm	0.1,2μm	0.1,0.2μm	1,2GPa

Table IV.5: STI Height Analysis DOE Parameters.

SA	SB	$STIW_L$	$STIW_R$	$STIHeight$
0.2,4μm	0.2,4μm	0.1,2μm	0.1,2μm	0.2,0.4μm

Table IV.6: DSL Process Steps.

1. Deposit active region silicon at 500^oC
2. STI fill at room temperature
3. Deposit gate oxide at 900^oC
4. Deposit polysilicon gate at 600^oC
5. Deposit tensile liner (in five steps) at 450^oC
6. Etch tensile liner using compressive liner mask (in five steps) at room temperature
7. Deposit compressive liner (in five steps) at 450^oC

Table IV.7: DSL Process Parameter Table.

Built-in Oxide Thickness	Active Region Thickness	Poly Thickness
0.15 μm	0.08μm	0.1μm
Built-in Poly Stress	Tensile Liner Stress	Compressive Liner Stress
-500MPa	1GPa	2GPa

DSL Teststructures. In order to analyze the impact of DSL, we have designed teststructures and ran TCAD simulations on these teststructures. We use FAMMOS 2007.09 [88] for the TCAD simulations. This simulator can handle 3D configurations and mainly conducts a finite element analysis for thermal mismatches.

We provide the teststructure template in Figure IV.6. We use the same teststructure for both PMOS and NMOS by replacing all compressive regions with tensile and vice versa. The dimensions x and y denote the distances from the edge of the active region to the opposite stress region in the parallel and orthogonal directions, respectively.

We conduct two types of experiments. The first type is proximity effect and the second type is process-impact. Proximity effect pertains to the effect of the neighboring liner on a given transistor. We provide the parameters for a subset of the proximity effect DOE we have used in Table IV.8 for PMOS and in Table IV.9 for NMOS. We provide the percentage mobility improvement as calculated by the FAMMOS tool at the last column of the tables. Starting with DOE structure 3 and going towards DOE structure 1, i.e. bringing the opposite stress liner closer to the channel in the parallel (x) direction, decreases the mobility. On the other hand, starting with DOE structure 3 and going towards DOE structure 5, i.e. bringing the opposite stress liner closer to the channel in the orthogonal (y) direction, increases mobility. We observe that there is not much impact from transistor width and length of oxide definition (LOD) based on DOE structures 6-9. However, we know that these parameters result in larger impact in silicon. Similar observations reveal

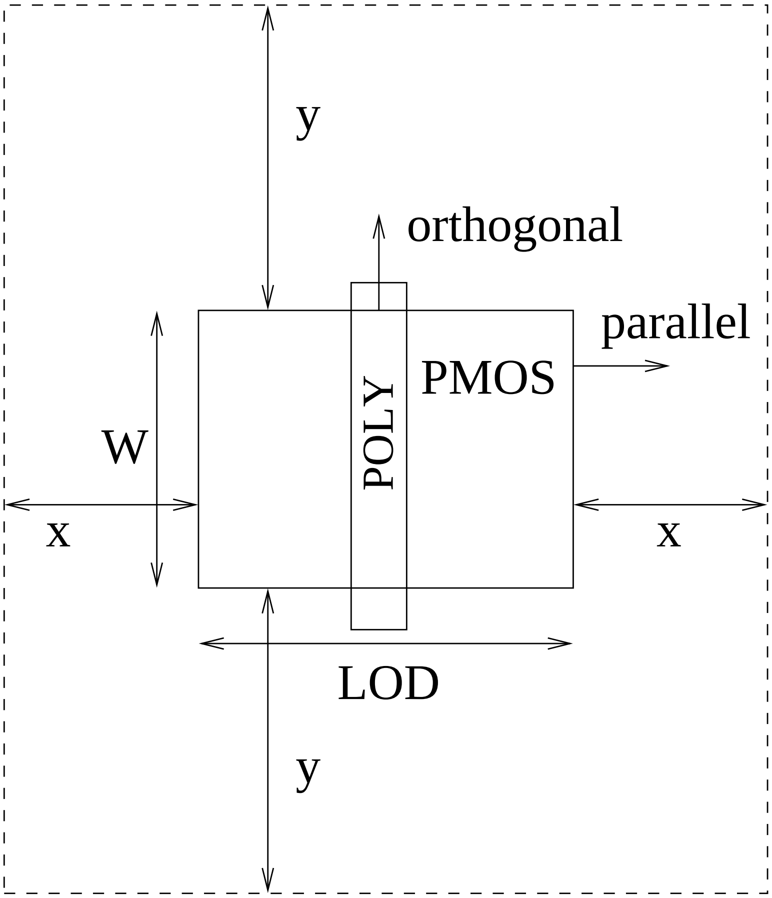


Figure IV.6: Dual stress liner test structure. x is parallel proximity to opposite stress liner; y is orthogonal proximity to opposite stress liner.

that NMOS speed is degraded when the opposite stress liner is brought closer to the channel in either parallel or orthogonal directions, i.e. going from DOE structure 11 towards 10 and 12, respectively.

Sensitivity to Process Parameters. To analyze the impact of process parameters of DSL on final mobility, we have conducted additional experiments. We change the DSL liner thickness and intrinsic stress and observe the change in channel mobility.

Analysis of the Proximity Effect. We have observed that placing a tensile region close to PMOS in the parallel direction has reduced the PMOS mobility.

Table IV.8: DSL Proximity Effect DOE for PMOS.

DOE No.	x	y	W	LOD	Mob. Improve.
1	$0.05\mu m$	inf	$0.5\mu m$	$0.5\mu m$	32.23%
2	$0.5\mu m$	inf	$0.5\mu m$	$0.5\mu m$	58.92%
3	inf μm	inf	$0.5\mu m$	$0.5\mu m$	65%
4	inf μm	$0.05\mu m$	$0.5\mu m$	$0.5\mu m$	77%
5	inf μm	$0.5\mu m$	$0.5\mu m$	$0.5\mu m$	73%
6	0.5 μm	inf	$0.5\mu m$	$0.75\mu m$	57.14%
7	0.5 μm	inf	$0.5\mu m$	$1.0\mu m$	56.27%
8	inf μm	$0.5\mu m$	$0.75\mu m$	$0.5\mu m$	73.22%
9	inf μm	$0.5\mu m$	$1.0\mu m$	$0.5\mu m$	76.87%

Table IV.9: DSL Proximity Effect DOE for NMOS.

DOE No.	x	y	W	LOD	Mob. Improve.
10	$0.05\mu m$	inf	$0.5\mu m$	$0.5\mu m$	-33.76%
11	inf μm	inf	$0.5\mu m$	$0.5\mu m$	-5.4%
12	inf μm	$0.05\mu m$	$0.5\mu m$	$0.5\mu m$	-29.07%

On the other hand, placing a tensile region close to PMOS in the orthogonal direction improves the PMOS mobility.[2]

NMOS mobility is degraded by the proximity of compressive liners in either direction. Notice that for both NMOS and PMOS, the degradation due to the proximity of the opposite stress liner is still less than the benefit obtained from the liner right above the transistor. A rule of thumb based on silicon data is 15-30% improvement due to the liner above the transistor and a possible additional

[2] Notice that percentage change in drive current is usually less than that of the mobility. A rule of thumb drive current improvement would be one half the mobility improvement. Furthermore, there is a saturation effect; when mobility is further increased or reduced, the drain current may saturate.

5-10% due to the proximity effect.

Analysis of the Process. In order to analyze the impact of process parameters, we conduct the DOE given in Table IV.10.

Table IV.10: DSL Process Sensitivity DOE Parameters.

Str. No	Tensile Stress	Compressive Stress	Liner Thickness	Mob. Improve.
3	$1GPa$	-$2GPa$	$0.125\mu m$	61%
1	$1.5GPa$	-$2GPa$	$0.1\mu m$	15.62%
1	$0GPa$	$0GPa$	$0.1\mu m$	2.14%

In Table IV.10, the first column is the structure number used from Table IV.8. Taking the given structure as reference, we alter the liner stresses or thicknesses. The last column shows the mobility improvement with respect to the structure given in the first column. The last row indicates that although a nitride liner is deposited over the transistors, this nitride liner is not stressed. This corresponds to a technology with no stress liners.

Proposed Design Guidelines. In order to improve the performance in a DSL technology, the following design guidelines needs to be adhered.

1. Place tensile liner away from PMOS in the parallel direction.
2. Place tensile liner close to PMOS in the orthogonal direction.
3. Place compressive liner away from NMOS in the parallel direction.
4. Place compressive liner away from NMOS in the orthogonal direction.

We illustrate these guidelines in Figure IV.7. By default, masks may not be designed to follow these guidelines. Reasons would include eliminating additional optimization efforts, finding such impact negligible, or reducing mask verification costs and design rules by deriving stress liners directly from well boundaries. However, the proposed guidelines can improve performance by 5 to 10% and such efforts are worth to pursuit.

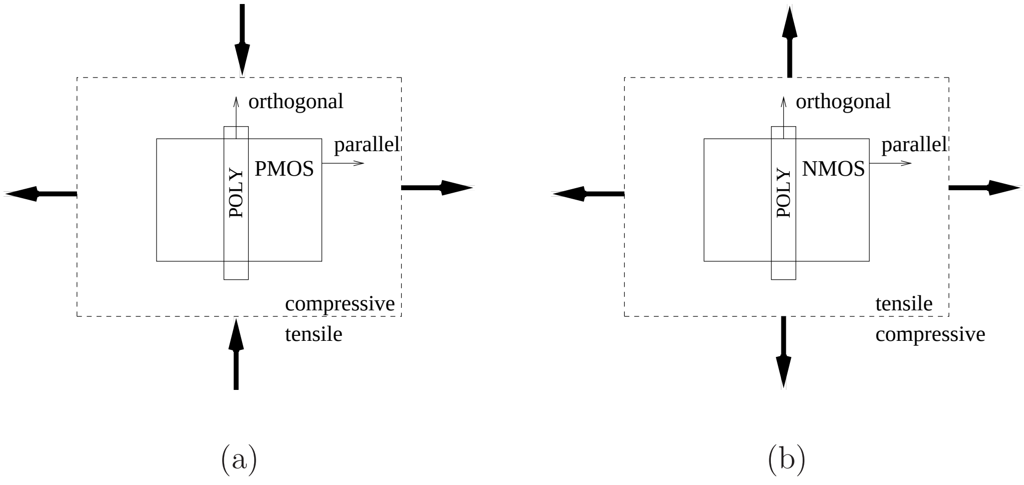


Figure IV.7: Proposed guidelines for optimizing the DSL boundaries. (a) PMOS: bring opposite liner close in the orthogonal direction; push away the opposite liner in the parallel direction. (b) NMOS: push away opposite liners in both directions.

IV.B Stress-Aware Timing Analysis

In this section we describe our STI stress-aware timing analysis methodology. We adapt the traditional SPICE-based timing analysis flow to consider stress induced by STI widths.

IV.B.1 Traditional SPICE-Based Timing Analysis

Cell-level static timing analysis tools such as PrimeTime offer a good tradeoff between accuracy and analysis speed. Full designs or their blocks are typically analyzed and signed off with circuit-level static timing analysis (STA). However, if greater accuracy is desired, SPICE-based analysis, which has better accuracy but substantially slower analysis speed, is employed. Since running full-chip SPICE analysis is not feasible, critical paths are first identified with static timing analysis and then simulated with SPICE. A typical netlist input to SPICE is layered into the following three tiers:

- *Device-level models* which contain transistor parameters in the form of coefficients of functions defined in BSIM or equivalent formats. Device-level models allow output waveforms for PMOS and NMOS devices to be simulated.

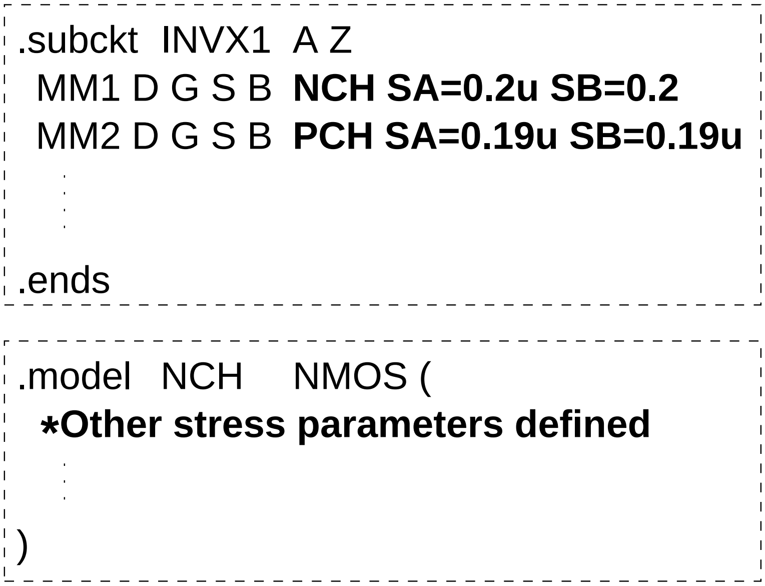
- *Cell-level netlists* which describe the connectivity of the devices that comprise individual cells. Cell-level netlists instantiate device-level models and allow SPICE to simulate waveforms at the outputs of cells in the library when subjected to given stimuli.

- *Critical path netlists* which describe the connectivity between the cells for each critical path. Critical path netlists instantiate cell-level netlists and can be simulated to calculate the delay of the critical paths.

As noted above, stress-induced device mobility change is determined by (1) the separation between the gate and the active edges, and (2) by the size of the STI region that surrounds the active region of the device. Fortunately, the separation between gate and active edges is fixed when the cells are designed, and the contribution of this separation to stress and mobility can be modeled at the cell level. Specifically, in the BSIM 4.3.0 device-level models, stress parameters *SA* and *SB* have been introduced to model the stress effect as a function of gate and active edge separation. In cell-level netlists these parameters are passed with the instantiation of the device-level models. Cell-level netlists are used in library characterization to generate gate-level timing models for use in STA. An example of device-level instantiation with stress parameters is shown in Figure IV.8.

The stress effect due to STI width is not modeled primarily for the following two reasons:

- STI width is determined by the placement of the cells, so that stress effect due to STI cannot be captured in library characterization. A new methodology that analyzes a placed design and annotates STI width information for use in timing analysis is required.



```
.subckt  INVX1  A Z
  MM1 D G S B  NCH SA=0.2u SB=0.2
  MM2 D G S B  PCH SA=0.19u SB=0.19u
    .
    .
    .
.ends
```

```
.model   NCH     NMOS (
  *Other stress parameters defined
    .
    .
    .
)
```

Figure IV.8: Instantiation of device-level models in a standard-cell SPICE netlist. Parameters added in BSIM 4.3.0 to partially model stress are shown in bold.

- Stress effect due to STI is of smaller magnitude than gate and active edge separation.

IV.B.2 STI Stress-Aware Timing Analysis

Our approach analyzes the placement of a design and the standard-cell layouts to calculate the STI widths for all critical cells in the design. The STI widths are then passed as parameters which are used in the models developed in the previous section.

We modify the cell-level netlists such that parameters PL, PR, NL, NR which capture the STI width are passed to them. Parameter PL is the spacing between the boundary of a cell and the neighboring active region to the left of its p-type active region. Similarly parameter PR is the spacing between the boundary of a cell and the neighboring active region to the *right* of its p-type active region (PRX). Parameters NL, and NR are similarly defined for *n-type* active regions (NRX). The parameters are set in the critical path netlists when cells are instantiated as shown in the example in Figure IV.9.

The PL, PR, NL, NR parameters can be calculated from the placement and the cell's layouts, specifically, the cell boundary to active spacings. Computation of PL for a cell, which is the spacing between the cell's boundary and the p-type active region of the cell to its left, is as follows. The spacing between the cell and its left neighboring cell is found from the placement. The spacing between the p-type active region of the neighbor and its cell boundary is found from layout analysis of the neighbor. The two spacings are then added, with correct consideration of the orientations of the cell and its neighbor. Other parameters PR, NL, and NR are calculated similarly. Figure IV.10 illustrates the calculation.

We note that our flow needs modifications to work for cells with complex active shapes such as flip-flops and multiplexors. Active shape complexities include non-rectangular shapes and non-continuous shapes. To model STI stress impact for non-rectangular active shapes, modifications such as those employed by BSIM to handle non-rectangular active may be used. For cells with non-continuous active shapes, devices can be completely shielded from STI width outside the cell and our flow should not alter their mobility. In our analysis and optimization, we focus on the cells with simple active shapes and do not change the mobilities for cells with complex active shapes (i.e., use traditional analysis and no optimization for them). Fortunately, the most frequently used cells such as inverters, buffers, NAND's, NOR's, AND's, and OR's have simple active shapes so we consider and optimize most cells in our designs.

IV.B.3 Alternative Flow

STI stress aware timing analysis can also be performed by cell-level STA. Towards this standard cells in the library can be characterized for different STI width configurations around them. Since stress dependence on STI width is relatively gradual, STI widths can be binned into a small number of bins to reduce the total number of STI width configurations. For each standard cell, variants may be created corresponding to each STI width configuration. Given the STI width, models presented in the previous section are used in library characterization. The STI width of a cell in a design can be computed from the placement and standard-cell layouts, and can be used to find the variant that has the closest STI width configuration. The cell can then be bound to the variant in the library and cell-level STA run to perform STI stress aware timing analysis.

IV.C Timing Optimization

In this section we present our timing optimization methodology. The basic idea exploited in our optimization is that STI widths of devices can be altered to change their mobility and improve performance. Specifically, the alteration involves increasing the STI widths for PMOS devices and decreasing them for NMOS devices. We identify the timing critical cells and alter their STI widths to improve the circuit performance. In our approach we use the following two knobs to alter the STI widths:

- Placement perturbation. The placement of a layout can be changed to increase or decrease the spacing between neighboring cells which directly increases or decreases the STI width. Additionally, cells can be spaced apart to allow fills, for which initially insufficient space exists.

- Active-layer fill (*RX fill*) insertion. Active-layer fills are rectangular dummy geometries inserted on the active (RX) layer primarily to improve planarity after chemical mechanical polishing (CMP). However, such geometries also reduce the STI width of the devices next to which they are inserted. The STI width after insertion of an RX fill next to a device is the spacing between the active region of the device and the fill.

We now present the details of the above two knobs.

IV.C.1 Active Layer (RX) Fill Insertion

The effect of RX fills on stress is identical to that of active regions of devices. When inserted next to the active region of an NMOS device, fills substantially reduce the STI width and stress of the device, and consequently improve the performance of the NMOS device. On the other hand, fills inserted next to a PMOS device reduce STI width and stress but consequently degrade the performance. Hence, inserting fill next to the NMOS devices but not next to the PMOS devices of a cell improves performance.

Circuit delay improves when the delay of setup-critical cells is reduced. So, we insert rectangular RX fills next to the NMOS devices, to the left and right of the cell. No RX fills are inserted next to the PMOS devices; so the PMOS remains exposed to a large STI width and stress. The devices closer to the active boundary experience the maximum benefit of this optimization. Since the most frequently used cells in the designs are small, a large fraction of devices in the design benefit from fill insertion. Our technique can also be employed for hold-time critical cells in the reverse manner, i.e., insert fills next to the PMOS devices but not next to NMOS devices to slow down the cell.

Figure IV.11 shows an example standard cell with PRX (active regions for PMOS devices) and NRX (active regions for NMOS devices). As can be seen, active regions exist under the top and bottom cell boundary that completely shield the cell from STI stress effects in the direction orthogonal to the carrier (current) flow direction. Hence, we only apply our optimization in the parallel direction by inserting fill to the right and left of a cell. Figure IV.12 illustrates fill insertion for a setup-critical cell; NRX fills are inserted next to the NRX region to reduce stress and *fasten* the NMOS devices. Figure IV.13 illustrates the approach for a few setup-critical cells in a standard-cell row.

All fills are inserted subject to the design rule constraints (DRCs) and introduce no DRC violations. No additional mask step is required, and that $M1$ capacitance impact is likely negligible. Since the fill insertion knob can only decrease STI width, NMOS performance can be improved but PMOS performance can at best be kept constant. However, neighboring cells which have very small spacing and between which fills cannot be inserted can be spaced apart by placement perturbation to allow fills to be inserted.

IV.C.2 Intra-Row Placement Optimization

We now present the placement perturbation knob which can increase (decrease) the STI width and improve PMOS (NMOS) performance. Placement of a cell determines its location (consequently its neighbors and spacings with them) and its orientation. In our optimization we change the location of the cells such that spacings are altered but the ordering of cells in a standard-cell row is not affected. Increased spacing next to a cell, increases the STI width and improves the delay of the PMOS devices. However, the delay of the NMOS devices increases with increased spacing. Fortunately, we can utilize our first knob, RX fill insertion, to reduce the NMOS STI width and improve its delay as well. In fact, if the spacing between cells is too small for fill insertion, placement can facilitate fill to be inserted by creating additional space for it. The placement perturbation just reorganizes the whitespace in the standard-cell row of the cell without requiring any additional space.

Minimizing delay increase due to wirelength increase. The perturbation of detailed placement from the original placement results in small wirelength change, which can impact wire parasitics and consequently timing. Even though our localized placement perturbations do not significantly affect timing, small changes in the timing of critical paths can affect the minimum clock cycle time. To minimize the timing change of critical paths, we fix the cells and nets in the critical paths. Fixed cells are not moved during optimization and fixed nets are not changed during engineering change order (ECO) routing that is performed after optimization. Since the nets in the critical path are fixed, all cells connected to these nets should also be marked as fixed and not moved during optimization. We note that the delay of such nets can marginally change due to the coupling capacitance with neighboring nets, the routing for which may change. We also fix all flip-flops, clock buffers, and clock nets to avoid any impact on the clock tree. So our list of *fixed cells* comprises timing critical cells, their fanout cells, flip-flops, and clock buffers.

Our intra-row placement optimization attempts to create space on the right and left sides of each timing critical cell. In the process, the minimum number of cells are displaced to minimize the wirelength impact. Figure IV.14 presents the pseudo-code for our intra-row placement optimization. For each timing critical cell, right and left spacings are increased by functions *createRightSpace* and *createLeftSpace* respectively to attain a spacing of up to S. The spacing, S, may not always be attainable because of the presence of fixed cells and availability of limited space in the row. For the right side, the function *cellsToMoveRight* finds the minimum number of cells to move. Then the function *moveCellsRight* flushes the computed number of cells to the right as much as possible.

Our algorithm sequentially processes critical cells in decreasing order of their criticality. Cells displaced in an iteration to create space are added to the list of fixed cells to lock them for successive iterations. This can limit the optimization of critical cells processed later in the algorithm. Therefore, we run the algorithm multiple times with increasing value of S. This enhancement allows a fair distribution of whitespace among all critical cells. We increase the value of S from $0.6\mu m$ to $1.8\mu m$ in steps of $0.2\mu m$. Starting with a smaller value of S leads to a more equitable distribution of whitespace at the expense of runtime. For designs with high utilization ratios, starting S as less than $0.6\mu m$ may be desirable.

Our second enhancement is perturbing the critical cells to balance the space on the right and left sides of them. Since the stress effect decays rapidly with space, nearly-equal spacings on both sides are desirable. We limit the perturbation to $0.6\mu m$ to minimize wirelength and the associated delay increase. The space required to insert RX fill is typically very small and in the $0.2\mu m$ range. Therefore, if the optimization creates any space for PMOS optimization, fill can always be inserted to improve the deteriorated NMOS performance. Figure IV.15 illustrates placement perturbation and fill insertion for setup-time optimization on a standard-cell row.

While it is possible to perform fill insertion without placement perturbation, we have found the associated performance benefits to be very small. Both knobs complement each other: placement creates space for fill insertion and fill insertion improves the performance of the NMOS devices that are slowed down by placement perturbation. Our overall STI stress-aware placement and fill optimization flow is as follows:

1. Identify critical paths and critical cells
2. Perform intra-row placement optimization
3. Perform ECO routing followed by parasitic extraction
4. Perform RX fill insertion
5. Evaluate the optimized layout with STI stress-aware timing analysis

IV.C.3 Post-Layout Optimization

Intra-row optimization may be sufficient for standard-cell designs. However, for custom blocks, post layout optimization is beneficial. Furthermore with intra-row optimization, there may be unused intra-cell optimization opportunity that may have been lost. For post layout optimization, we leave the optimization to a later stage, i.e, after we obtain a GDS layout. This approach brings the cost of geometric layout processing instead of faster algorithms based on cell placement manipulation.

IV.D Experimental Study

We now present our experiments to evaluate the proposed optimization methodology. Our experiments assess the impact of our optimization on the minimum clock cycle time, delay of top critical paths, and final routed wirelength.

Table IV.11: Testcases Used in Experimental Validation.

Circuit	Source	#cells	Utilization	MCT (ns)
C5315	ISCAS'85	1,408	82%	0.912
ALU	opencores.org	11,106	78%	4.333
S38417	ISCAS'85	8,514	79%	3.086
AES	opencores.org	21,000	78%	4.738

IV.D.1 Experimental Setup

The details of the testcases used in our experiments are presented in Table IV.11. In Table IV.11, **MCT** is the minimum cycle time. We use *Synopsys Design Compiler vW-2004.12.SP3* [124] for synthesis, *Cadence SOC Encounter (v5.2)* [118] for placement, clock tree synthesis, routing, and parasitic extraction, *Synopsys PrimeTime vW-2004.12.SP2* [122] for cell level timing analysis, and *Synopsys HSPICE vY-2006.03* [125] for SPICE simulations. For our experiments, we use the 50 most frequently used cells from high-V_{th} and nominal-V_{th} 65nm high-speed libraries. SPICE device models and cell netlists were supplied by a foundry. We built our optimizer on top of *OpenAccess API v2.2.4* [123].

IV.D.2 Experimental Results

We first compare the proposed stress-aware timing analysis with traditional analysis. Since traditional analysis does not account for STI stress and must correctly analyze all STI configurations, it is conservative. Traditional analysis is corner-based and uses the worst-case cell delay which reflect worst-case STI stress effects in addition to worst-case process variations. Worst-case analysis, while correct, leaves valuable performance on the table. Stress-aware timing analysis reduces pessimism in analysis by explicitly accounting for STI stress. We therefore expect stress-aware timing analysis to report circuit delays that are smaller than those of traditional analysis.

Table IV.12: Traditional vs. Stress-Aware Timing Analysis.

Circuit	Traditional		Stress-Aware			
	MCT (ns)	TPD (ns)	MCT (ns)	MCT (%)	TPD (ns)	TPD (%)
C5315	0.977	87.43	0.915	6.31	81.93	6.29
ALU	1.885	185.50	1.778	5.68	175.24	5.53
S38417	1.068	104.95	1.018	4.68	99.58	5.11
AES	1.739	165.82	1.655	4.83	158.88	4.19

Table IV.12 presents the comparison between traditional timing analysis and stress-aware timing analysis on four testcases. We study two delay metrics: (1) minimum cycle time (MCT), (2) and *top paths delay* (TPD), which is the sum of the delays of top 100 critical paths. While MCT determines the maximum speed at which the circuits can be run, TPD determines the robustness to variations. We observe that stress-aware analysis reduces MCT by 5.75%, and TPD by 5.28% on average. We use stress-aware analysis to evaluate our optimization in the remainder of this section.

In Section IV.C we presented two optimization knobs: fill insertion and placement perturbation. Although, the two techniques complement each other, we evaluate the fill insertion knob separately. Placement perturbation, without fill insertion, is not interesting because it slows down the NMOS devices while speeding the PMOS. Table IV.13 presents the improvements in MCT and TPD due to fill insertion. In Table IV.13, **MCT** is the minimum cycle time and **WL** is the wirelength. **TPD** stands for top paths delay and is the sum of the delays of the top 100 critical paths. Since we optimize several critical paths, TPD reduces. However, we observe that reductions in MCT and TPD are typically under 1%.

We now evaluate the simultaneous use of the proposed placement perturbation and fill insertion knobs. In addition to comparing MCT and TPD results, we also compare the wirelength which changes because of placement perturbation.

Table IV.13: Timing Optimization Results With Fill Insertion.

Circuit	Original		Fill Opt			
	MCT (ns)	TPD (ns)	MCT (ns)	MCT (%)	TPD (ns)	TPD (%)
C5315	0.915	81.83	0.903	1.32	81.35	0.71
ALU	1.778	175.24	1.771	0.39	174.53	0.40
S38417	1.018	99.58	1.010	0.79	99.92	0.39
AES	1.655	158.88	1.651	0.24	158.55	0.21

After placement perturbation several nets are left dangling; we perform ECO routing to route them, and follow by RC extraction and stress-aware timing analysis to accurately report the MCT and TPD results for the optimized case. The runtime of our placement and fill optimization is generally small; it depends on the circuit size and the number of critical paths to be optimized. In our experiments, the runtime was under one minute for all testcases on a 2.2GHz AMD Opteron/8GB RAM machine running Linux 2.6.

Table IV.14 presents our results for our four testcases. In Table IV.14, **MCT** is the minimum cycle time and **WL** is the wirelength. **TPD** stands for top paths delay and is the sum of the delays of the top 100 critical paths. For negligible increase in wirelength, we observe 4.37% and 5.15% reductions in (stress-aware) MCT and TPD averaged over the testcases C5315, ALU, and AES. The testcase S38417, however, demonstrates smaller improvements. We attribute this to the fact that S38417 is a testcase with over 50% of its cells being flip-flops. We do not allow our optimization to change the locations of flip-flops, so as to avoid changes to the clock tree; hence, in the S38417 testcase, we can perturb the placement of fewer cells. Figure IV.16 shows the histograms for the delays of top 200 critical paths of our testcase AES before and after optimization. As can be seen, the delay distribution has shifted to the left (lower delay) substantially.

We also tried our technique to optimize hold-critical paths but found

Table IV.14: Timing Optimization Results With Placement and Fill Insertion.

Circuit	Original			Placement & Fill Opt					
	MCT (ns)	TPD (ns)	WL (mm)	MCT (ns)	MCT (%)	TPD (ns)	TPD (%)	WL (mm)	ΔWL (%)
C5315	0.915	81.93	17.8	0.879	3.97	75.50	7.85	17.9	+0.67
ALU	1.778	175.24	196.1	1.709	3.88	168.14	4.05	196.8	+0.36
S38417	1.018	99.58	96.4	0.993	2.44	97.94	1.65	96.64	+0.23
AES	1.655	158.88	374.7	1.568	5.26	153.21	3.56	3.75	+0.08

negligible improvement in hold slack for our testcases. This is because stress optimization can only change cell delays by 10%-20% and for hold-critical paths the cell delays are very small. So the change in the delay of hold-critical paths is insignificant with our approach and traditional delay increase methods such as insertion of delay elements or wire snaking must be used.

Post-Layout Optimization. To evaluate the benefit of full layout optimization, we have designed a miscellaneous ring oscillator with 33 stages and fanout of 2. We have used NANDs, NORs, OAIs and MUXes. We have manually optimized the layout. We have written PERL scripts to extract the parameters required for our STI models. We have observed 11.32% improvement with active fill insertion. The ring oscillator delay improves from 412.6ps to 365.9ps. The improvement for this testcase may slightly be overestimated as we do not have STIW models for orthogonal impact due to 2D TCAD limitations.

RC Impact of Active Fills. A practical consideration is that the insertion of fills may slightly increase the total capacitances. However, as the fills are floating and not grounded, this increase will be negligible. Furthermore, the insertion of floating active layer fills can reduce the line to line coupling for the $M1$ routing lines, as the fill will draw some of the electrical flux between the overlying lines.

In order to understand the impact due to active fill insertion, we have con-

Table IV.15: Comparison of Capacitances Due to Active Fill.

$F/\mu m$	case1	case2	case3
$M1$ total	1.71E-16	1.60E-16	1.64E-16
$M1$ coupling	6.120E-17	6.85E-17	6.86E-17

ducted 2D field solver simulations with Raphael. Dense $M1$ layer with $M2$ routing on top and underlying active layer is simulated. For $M1$ routing, we have compared the following three cases: an underlying active layer which is grounded (case1), no underlying active layer (case2) and an underlying active layer fill (case3). The results, as given in Table IV.15 indicate that the total capacitance for $M1$ is between grounded fill and no fill for the floating active fills, closer to the latter one. The impact on the coupling capacitance is same as the case when there are no active layers underneath.

IV.E Conclusions

We have conducted TCAD process simulations to generate models that relate the dependence of transistor mobilities to stress induced by STI width. We have proposed an STI width-aware design methodology for standard-cell place-and-route designs. The proposed stress-aware timing analysis technique reduces pessimism in delay analysis. Over traditional corner-based analysis, delays reported by stress-aware analysis were on average 5.75% lower. We have also devised an optimization methodology, based on cell placement perturbation, to create extra space around critical cells; this is followed by dummy diffusion insertion. The proposed optimization flow, while demonstrated with our models, can be generalized to other STI stress models. We have applied the proposed optimization flow on a number of testcases implemented with industrial 65nm libraries. Our data shows that STI width optimization can increase performance by 2.44% to 5.26% with no area penalty. We have analyzed DSL technology and proposed guidelines for optimization. The proposed optimization scheme can form the basis of circuit optimization that exploits upcoming stress-engineered transistor technologies in 65nm and below processes.

IV.F Acknowledgments

We would like to thank Prof. Hoong-Joo Lee of Sangmyung University, Korea, and Frank Geelhaar of AMD for very useful discussions, and Swamy Muddu for help with the implementation and the testcases. This chapter is based on or in part a reprint of the following publications.

- A. B. Kahng, P. Sharma and R. O. Topaloglu, "Chip optimization through STI stress-aware placement perturbations and fill insertion," *accepted for publication in IEEE Trans. on Computer-Aided Design*, 2008.

- A. B. Kahng, P. Sharma and R. O. Topaloglu, "Exploiting STI stress for performance," *Proc. IEEE/ACM International Conference on Computer-Aided Design*, 2007, pp. 83-90.

- R. O. Topaloglu, "Standard cell and custom circuit optimization using dummy diffusions through STI width stress effect utilization," *Proc. IEEE Custom Integrated Circuits Conference*, 2007, pp. 619-622.

```
*Critical path 00001
 X01 N1 N2 INVX1  PL=0.08u PR=4.08u NL=0.06u NR=4.06u
  X02 N2 1 N2 NAND2X1  PL=5.0u PR=5.0u NL=5.0u NR=5.0u
  X03 N3 N4 BUFFX1  PL=2.1u PR=5.0u NL=2.04u NR=5.0u
     ⋮
```

```
.subckt  INVX1  A Z
  .param PMOB = Our_PMOS_Model (PL, PR, NL, NR)
  .param NMOB = Our_NMOS_Model (PL, PR, NL, NR)
  MM1 D G S B  NCH SA=0.2u SB=0.2  MOB=NMOB
  MM2 D G S B  PCH SA=0.19u SB=0.19u  MOB=PMOB
     ⋮
.ends
```

```
.model NCH  NMOS (
  *Other stress parameters defined
     ⋮
)
```

Figure IV.9: Critical paths instantiate cell-level netlists which instantiate device-level models. Our modifications to the traditional flow to model STI width-dependent stress are shown in bold.

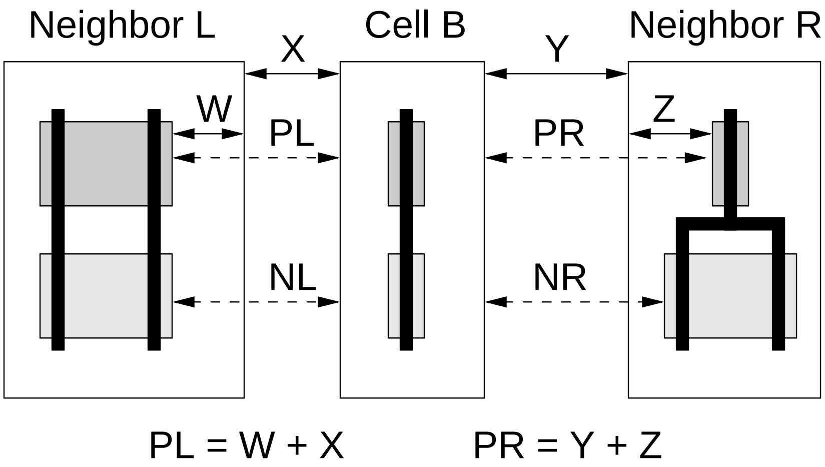


Figure IV.10: Calculation of parameters PL, PR, NL, and NR from inter-cell spacings and active to cell boundary spacings.

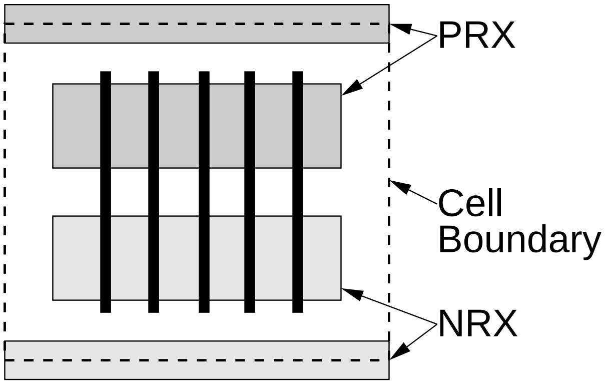


Figure IV.11: Generic standard cell showing polysilicon, p-type active regions (PRX), n-type active regions (NRX), and cell boundary.

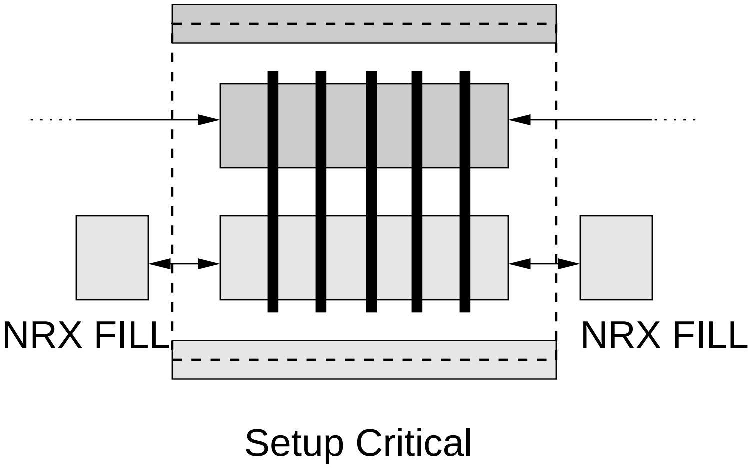


Figure IV.12: The generic cell of Figure IV.11 optimized with fill insertion for setup-time criticality.

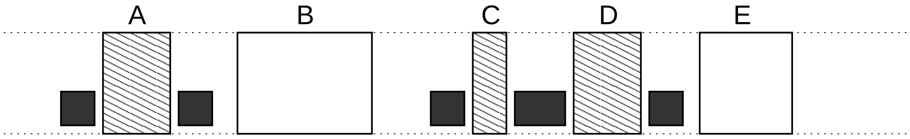


Figure IV.13: A row of standard cells after active-layer fill insertion for setup-time improvement. Cells patterned with diagonal lines are the setup-critical cells and filled rectangles are the inserted active layer fills.

Input: Placed design; set of timing-critical cells, T; set of fixed cells, F; maximum spacing to create, S

Output: New placement with altered inter-cell spacings

```
[1] forall cells $t \in T$
[1.1]    createRightSpace(t)
[1.2]    createLeftSpace(t)

createRightSpace(t)
[1] n = cellsToMoveRight(t);
[2] moveCellsRight(t, n);

cellsToMoveRight(t)
[1] i ← t;
[2] j ← cellToRightOf(t);
[3] accumulatedSpacing = 0;
[4] cellsToMove = 0;
[5] while accumulatedSpacing $\leq S$ and $j \notin F$ and cellToRightOf(j) $\notin T$
[5.1]    accumulatedSpacing += interCellSpacing(i, j);
[5.2]    cellsToMove++;
[5.3]    i ← j;
[5.4]    j ← cellToRightOf(i);
[6] return cellsToMove;

moveCellsRight(t, n)
[1] // flush n cells to the right of Cell t towards the right to create space S
[2] F ← F ∪ {n cells to the right of Cell t}

createLeftSpace(t)
[1] // similar to createRightSpace(t)
```

Figure IV.14: Pseudo-code for intra-row placement optimization.

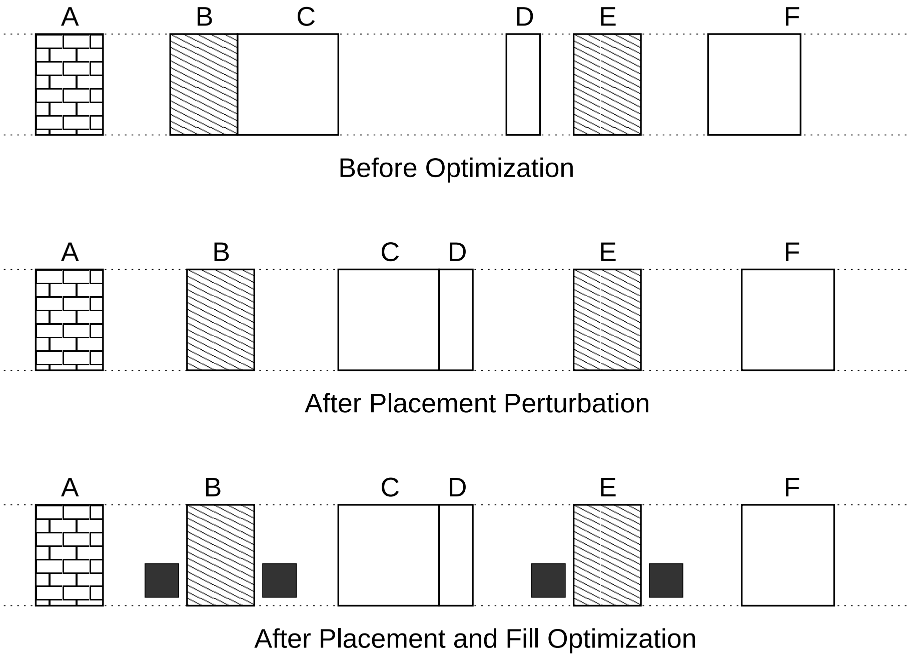

A
B
C
D
E
F
Before Optimization
A
B
C
D
E
F
After Placement Perturbation
A
B
C
D
E
F
After Placement and Fill Optimization

Figure IV.15: Placement change and fill insertion for setup-time optimization. A standard-cell row is shown before optimization, after placement perturbation, and after fill insertion. Cells patterned by diagonal lines are the setup-critical cells for which timing is optimized. Fixed cells are patterned with the brick pattern and their placement cannot be changed.

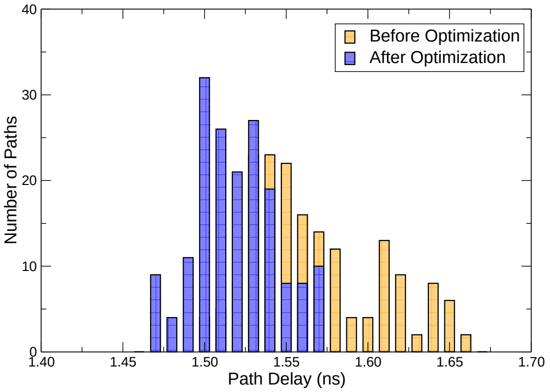


Figure IV.16: Path delay histograms for the top 200 critical paths of testcase AES, before and after optimization.

V

BEOL Stress Analysis

There is a need to understand the reliability impact of fills in the BEOL. Fill insertion should be reliability aware and we should have means to improve reliability of BEOL utilizing fill structures. In this chapter, we conduct a theoretical analysis of BEOL reliability due to fills using TCAD simulations. We design test structures to target specific interconnect and fill configurations, simulate these structures using TCAD, relate the results to known reliability concerns and provide guidelines to designers for BEOL stress optimization.

In Section V.A, we provide test structures to analyze the impact of CMP and via fills on stress. We provide our simulation methodology in Section V.B. We provide experimental results in Section V.C, provide design guidelines in Section V.D and conclude the chapter in Section V.E.

V.A Test Structures to Analyze Impact of CMP and Via Fill on Stress

We design different classes of test structures to analyze the impact of fills.

V.A.1 Evaluation of Fill Pattern on Intralayer Dielectric Stress

The first class of test structures helps analyze the impact due to fill pattern for intralayer parallel lines. To evaluate the impact of intralayer fill pattern on stress, we use the layouts shown in Figures V.1 (a)-(c) for traditional, two-pass and staggered fill patterns. Test structure $p0$ does not include any fills and has the same interconnect pitch as the other test structures. We use minimum widths for the interconnects. Fifty percent dense horizontal lines are used on neighboring layers. The minimum fill width is chosen to be the same as the interconnect minimum width. We design all patterns to have an equal density in a given window.

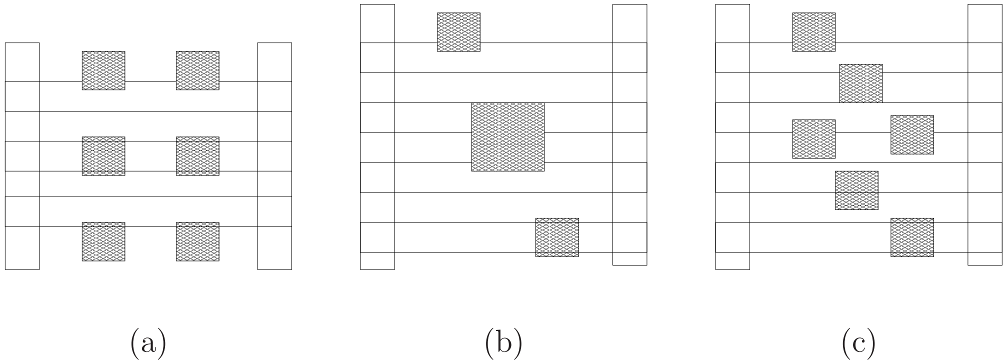


Figure V.1: Fill pattern test structures. Top view is shown. Fill is indicated by hatched area. Horizontal overlapping interconnects are on metal layers $M1$ and $M3$; vertical interconnects and fills are on $M2$. Test structure $p0$ has no fills in between interconnects. (a) Test structure $p1$. Traditional fill pattern. (b) Test structure $p2$. Two-pass fill pattern. (c) Test structure $p3$. Staggered fill pattern.

V.A.2 Impact of Fill Between First Neighboring Layer Orthogonal Lines

Most designs have two orthogonal lines in neighboring layers with or without a fill in between. To evaluate the impact of a fill presence between neighboring orthogonal lines, we use the structures as shown in Figures V.2 and V.3. We change the fill location and monitor the impact on stress. In Figure V.3 (b), we also introduce a via fill to electrically connect the fill to the neighboring layer interconnect. The fill width is chosen to be twice as wide as the minimum interconnect width.[1]

The purpose of test structure $t2$ is to analyze if insertion of a fill close to these lines improve or degrade the local dielectric and metal reliability with respect to test structure $t0$, which contains no fills. Test structures $t1$ and $t3$ are used to observe whether shifting the fill along one of the interconnects improves the local stress distribution. Finally, test structure $t4$ is used to analyze whether the inclusion of a via fill is justifiable in terms of stress distributions. The via fill connects the CMP fill to an interconnect (or another fill) on the neighboring layer. For test structures $t5$ to $t7$ in Figure V.4, we also introduce a via into our test structure, as vias are especially important for a stress migration analysis.

We hypothesize that the fill insertion will increase the delamination possibility as the fills contain larger tensile stresses as compared to dielectric. A fill close to the interconnect will increase the normal stress components on the interconnect surface. Introduction of a via fill should reduce the delamination possibility.

V.A.3 Impact of Fill Between Second Neighboring Layer Parallel Lines

To evaluate the impact of a fill between second neighboring parallel lines, we use the setups as shown in Figures V.5 **(a)-(c)**. We introduce a fill and then change the location of the fill, which belongs to the common neighboring layer.

The purpose of test structure $s1$ is to analyze if insertion of a fill between the two overlapping lines residing two layers apart improves or degrades the local dielectric and metal reliability with respect to test structure $s0$, which contains no fills. We include test structure $s2$ to analyze whether shifting the fill away from the interconnects improves the local stress distribution.

[1] To evaluate worst-case conditions, we have chosen the fill to interconnect spacing as the minimum interconnect to interconnect spacing; although for most technologies, this distance is set to at least twice the minimum interconnect to interconnect spacing per layer.

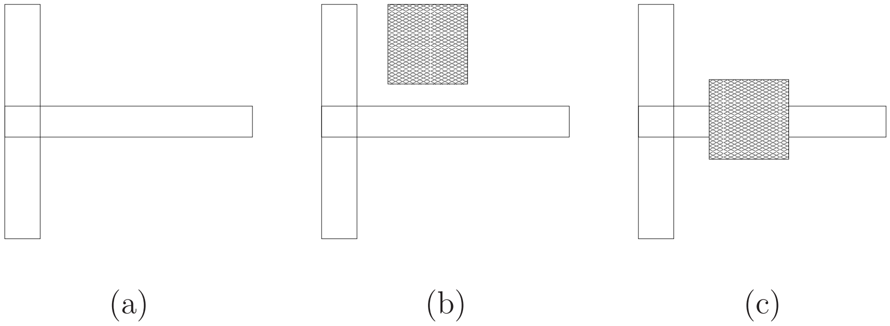


Figure V.2: Test structures for CMP fill between neighboring layer orthogonal lines. Top view is shown. Fill is indicated by hatched area. Horizontal interconnect is on metal layer $M1$; vertical interconnect and fill are on $M2$. (a) Test structure $t0$. No fill, reference structure. (b) Test structure $t1$. Fill is shifted up by fill width. (c) Test structure $t2$. A fill is inserted overlapping the underlying line.

Our hypothesis is that the fill insertion will increase the delamination possibility as the fills contain larger tensile stresses as compared to dielectric. A fill close to the interconnect will increase the normal stress components as compared to a dielectric.

V.A.4 Impact of Via Fill

To evaluate the impact of via fill, we use the layout as shown in Figures V.6 **(a)-(b)**. The setup consists of two layers consisting of four fills each in traditional orientation overlapping the fills on the neighboring layer. We simulate the structure with and without via fills between the fills of neighboring layers.

The purpose of the via fill test structure is to justify whether the via fills improve delamination and dielectric reliability. Our hypothesis is that the via fill should improve the delamination by reducing stress gradients which would otherwise be present across interconnects and vast dielectric regions.

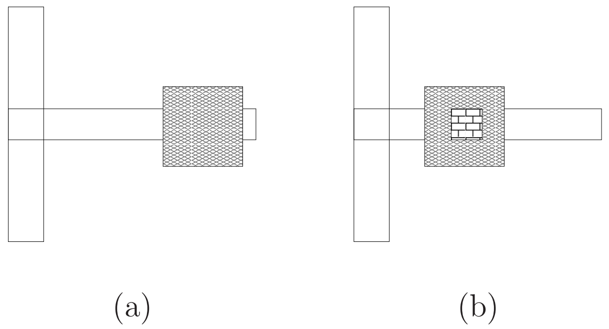


Figure V.3: Test structures for CMP fill between neighboring layer orthogonal lines. Top view is shown. Fill is indicated by hatched area. Horizontal interconnect is on metal layer $M1$; vertical interconnect and fill are on $M2$. Via fill is indicated by the brick-pattern rectangle and connects the fill and the layer $M1$ interconnect. (a) Test structure $t3$. Fill shifted away from the interconnect in the same layer. (b) Test structure $t4$. A via fill is inserted between the fill and the neighboring layer interconnect.

V.A.5 Impact of Fill Size

We design the test structures in Figure V.7 to evaluate the impact of fill size on the traditional fill pattern test structure $p1$. We increase the fill sizes 50% per side in test structure $p1a$. In $p1b$, we further add via fills to $p1a$.

V.A.6 BRI Test Structures

We hypothesize that the BEOL reliability will improve if floating fills are connected to interconnects using via fills. In order to evaluate this hypothesis, we design the test structures in Figure V.8. Via fills are used in $b1$ to connect an interconnect to neighboring layer floating fills. Although we use long rectangular fills in our experiments, other sizes and shapes are possible.

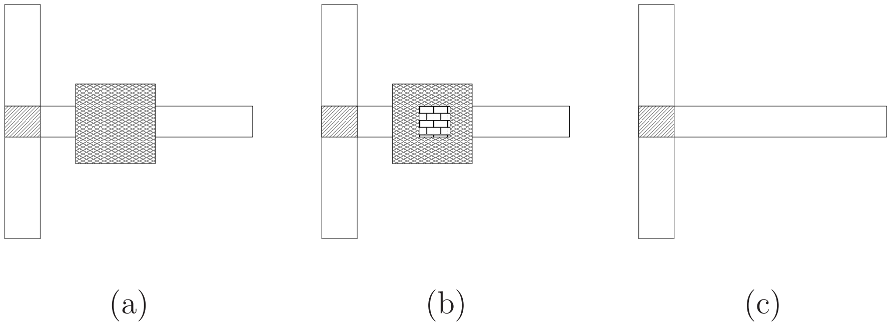


Figure V.4: Test structures for CMP fill between neighboring layer orthogonal lines. Top view is shown. Fill is indicated by hatched area. Via fill is indicated by the brick-pattern rectangle and connects the fill and the layer $M1$ interconnect. Via is indicated by the gray region. Horizontal interconnect is on metal layer $M1$; vertical interconnect and fill are on $M2$. (a) Test structure $t5$. A via and a fill are present. (b) Test structure $t6$. Via and via fill are present. (c) Test structure $t7$. Via only is present.

V.A.7 Impact of Process Parameters

A very important factor in the final stress value can be the choice of process parameters. Although process parameters are determined taking multiple considerations such as electrical performance and manufacturability into account, BEOL stress should be one of the targets to improve reliability.

To analyze the impact of process parameters on stress, we conduct experiments and observe the sensitivity of stress to process parameters. The process parameters that we focus on are temperature, interconnect aspect ratio and via height. We use the test structures $p1$ but alter the process steps to assess the sensitivity of stress to a process parameter.

V.B Methodology

In this section, we present the methodology we follow in our study.

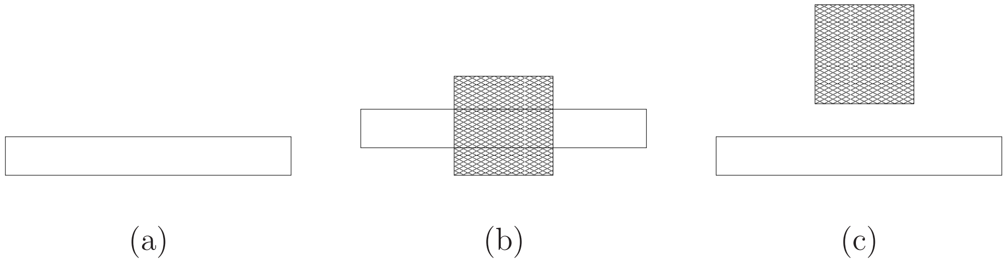


Figure V.5: Test structures for CMP fill between second neighboring layer parallel lines. Top view is shown. Fill is indicated by hatched area. Overlapping interconnects are on metal layers $M1$ and $M3$; fill is on $M2$. (a) Test structure $s0$. No fill, reference structure. (b) Test structure $s1$. A fill is inserted between the parallel lines. (c) Fill is shifted up by fill width.

V.B.1 BEOL Process Steps

In our study, we use a low temperature BEOL flow as given in Table V.1. In Step 2, CVD stands for chemical vapor deposition. To make the representation easier, we provide the steps per a single via and trench. In our simulations, we use up to 3 metal layers, all having the same minimum width, nominal metal height and via sizes. We use the physical dimensions as given in Table V.2. The simulator we use accounts for the stress changes due to deposition, etch and thermal mismatches. We did not simulate stress impact due to CMP. We take a deposition temperature of 250^oC. If the process has a higher deposition temperature, the stress change due to the thermal mismatch will be higher.

V.B.2 Stress Measurements and Reliability Correlation

Figure V.9 shows an interconnect trench. Using ellipses, we indicate a number of locations of interest to understand the impact due to stress. Multiple materials meet in these locations, making them significant for reliability. The local stresses can impact these points and hence degrade the reliability of a circuit. In particular, the top boundary is the most critical. Hence, the impact of stress in

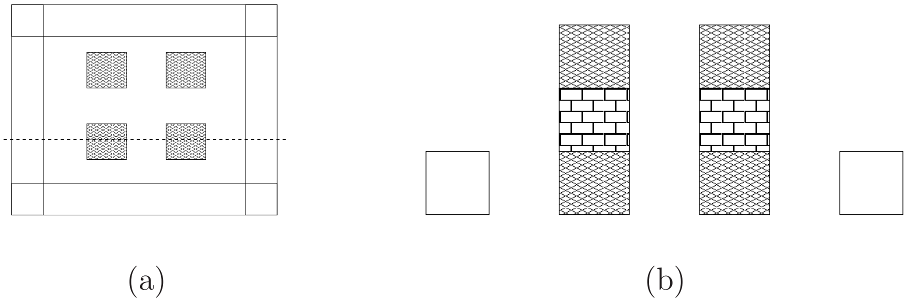


Figure V.6: Via fill test structures. CMP fills are indicated by hatched and via fills are indicated by bricked pattern. Vertical lines and four fills are on $M1$; horizontal lines and four fills are on $M2$. Fills on $M2$ overlap fills on $M1$. (a) Test structure $v0$. No via fill present. Top view is shown. (b) Test structure $v1$. Via fills are added between overlapping CMP fills. Side view is shown for the cross-section indicated by dashed line in the figure to the left.

this region should be monitored.[2]

We show the measurement points for our test structures in Figures V.10-V.12. Points indicated by o's are the points measured at the top boundary of the interconnects. The measurements inside the interconnect, i.e., the interconnect center and edge measurements are conducted $10nm$ from the closest boundary at the half length of the interconnects. We monitor the stress at the locations:

[2] If there is a via underneath, the bottom boundary also becomes as important.

Table V.1: BEOL Process Steps Per Each Via and Trench.

1.Nitride Cap Deposition	Deposit $0.03\mu m$ cap layer at 250^oC.
2.Via Dielectric Deposition	Deposit $0.13\mu m$ dielectric at 250^oC using CVD.
3.Trench Dielectric Deposition	Deposit $0.16\mu m$ dielectric at 250^oC using CVD.
4.Trench Etch	Etch the trench dielectric by $0.16\mu m$ at 25^oC.
5.Metal Deposition	Deposit copper at 250^oC.
6.CMP	CMP excess metal.

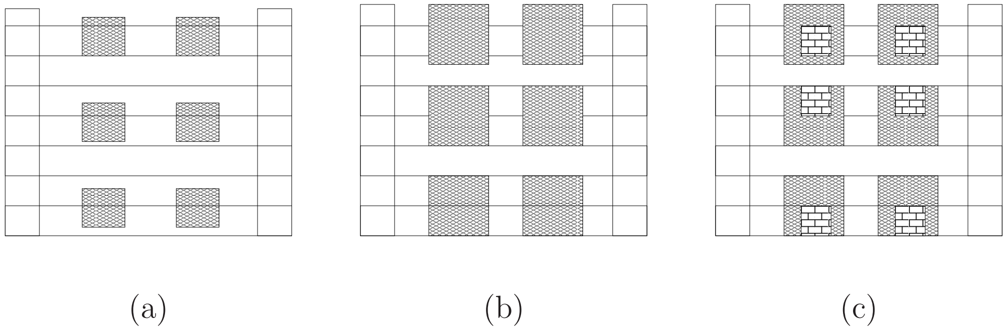


Figure V.7: Fill size test structures. Top view is shown. Fills are indicated by hatched areas. Horizontal overlapping interconnects are on metal layers $M1$ and $M3$; vertical interconnects and fills are on $M2$. (a) Test structure $p1$. Traditional fill pattern. (b) Test structure $p1a$. Fills are 50% larger. (c) Test structure $p1a$. Fills are larger and via fills are added to neighboring layer fills (or interconnects).

Table V.2: Process Physical Parameters.

Trench Height	Via Height	Trench Width
0.16 μm	0.16 μm	0.1 μm
Trench Spacing	Via Size	Cap Layer Thickness
0.1 μm	0.1 μm	0.03 μm

1. $m1$: at the center of the interconnect close to the top boundary,

2. $m2$: at the center of the interconnect close to the top boundary shifted $0.1\mu m$ along the interconnect,

3. $m3$: at the edge of the interconnect close to the top boundary,

4. $m4$: at half width away from the interconnect in the dielectric and

5. $m5$: above the center of the interconnect in the dielectric close to the top boundary.

These points are also enumerated in Figure V.10 (a). We measure the points $10nm$ away from the interconnect and dielectric boundaries at each of these

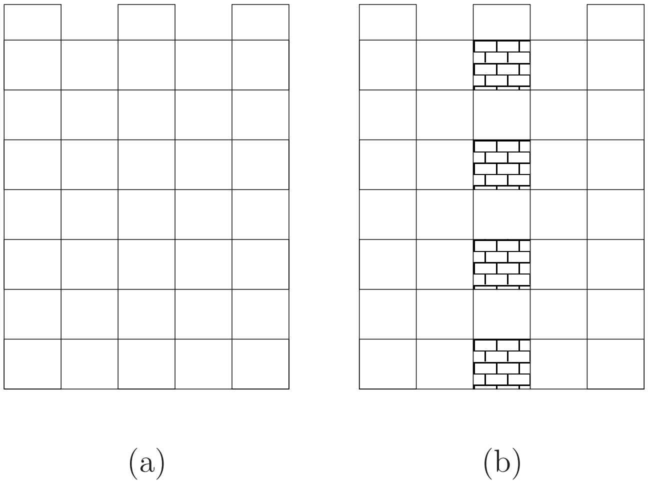


Figure V.8: BRI (BEOL reliability improvement) test structures. Top view is shown. Horizontal overlapping interconnects (or rectangular fills) are on metal layers $M1$ and $M3$; vertical interconnects and fills are on $M2$. (a) Test structure $b0$. Middle vertical line is the reference. (b) Test structure $b1$. Via fills are added.

measurement points. $m2$ is not shown in Figure V.10 **(a)**, as it would be $0.1\mu m$ into the plane where $m1$ is located.

- Average stress at $m1$ [95] and the stress gradient between $m1$ and $m2$ should provide the stress migration impact. Larger stress gradients would increase stress-induced voiding, especially if they are close to a via.

- Average stress at $m4$ should be helpful to monitor the dielectric breakdown reliability. Higher stress would be detrimental for the dielectric.

- The normal components at $m1$ and $m5$ should give information about the delamination and crack formation possibilities as normal stress would try to separate the metal from the dielectric.

- Von Mises stress at $m3$ should indicate Cu^+ diffusion possibility. A higher von Mises stress indicates that the metal is likely to deform, which may result in Cu^+ diffusion into the dielectric. Normal components at $m1$ and $m5$ may also provide similar information for Cu^+ diffusion into the dielectric similar

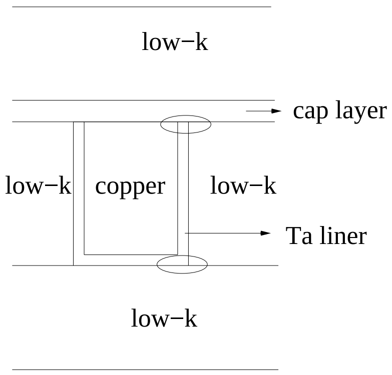


Figure V.9: Important regions to check for the impact of stress indicated on a trench.

to delamination, although the trend of these measurements may differ from the von Mises stress.

For electromigration, stress and stress gradients may amplify or counteract the electron flow [126], hence the result should depend on the particular layout configuration, i.e., where the via is located, what the current flow direction is, the length of the wire among the other process parameters such as the trench depth and wire width. Keeping all these parameters in account, the stress gradient between $m1$ and $m2$ may be used to determine if there will be any electromigration improvement. In this chapter, we focus particularly on the reliability concerns identified above, and not electromigration.

V.C Experimental Results

We use a finite element stress simulator, FAMMOS 2006.12 [88], in our simulations. Through our methodology, we have obtained the results in Table V.3. The first column gives the structure name. The results are in MPa. In the table, compressive and tensile stress values are shown with negative and positive signs, respectively. σ_{vm} is the von Mises stress and the σ_{av} is the volume average hydrostatic stress, the computations of which are given in Chapter I. The columns

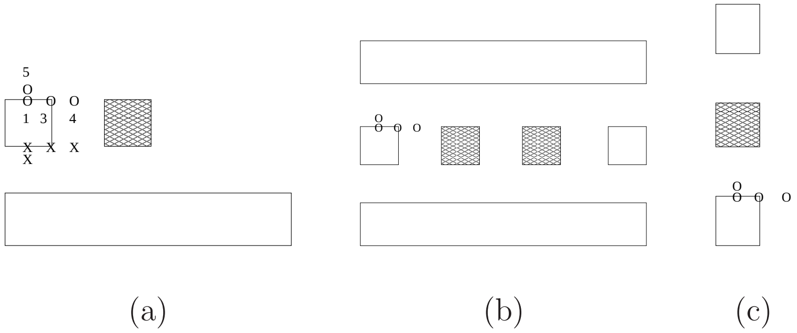


Figure V.10: Stress measurement points for test structures. Side view is shown. (a) Measurements for orthogonal interconnect test structures $t0$-$t7$. (b) Measurements for fill patterns test structures $p0$-$p3$. (c) Measurements of second-neighboring layer parallel line test structures $s0$-$s2$.

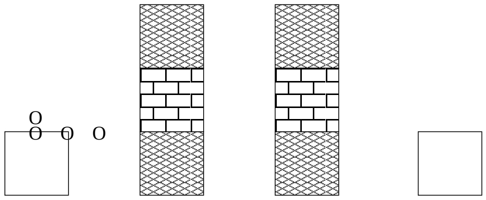


Figure V.11: Stress measurement points for via and no via fill structures. Side view. Top boundary measurements for via and no via fill test structures $v0$-$v1$.

of the table correspond to the measurement points $m1$, $m2$, $m3$, $m4$ and $m5$, respectively.

One general expectation we have is that the stress due to thermal mismatch will be higher at the top boundary. The reason is that the coefficient of thermal expansion (CTE) for copper is 17.7$ppm/^oC$ and for CVD low-k 20$ppm/^oC$ [95]. On the other hand, the cap layer CTE is 3.9$ppm/^oC$ [111]. The CTE difference between the copper and nitride is more than that between the copper and low-k. As the former corresponds to the top interconnect boundary, whereas the latter corresponds to the bottom interconnect boundary, it is expected that the stress due to thermal mismatches will be higher in the top boundary.[3]

[3] As our simulation methodology includes both etch, deposition and thermal mismatch stresses,

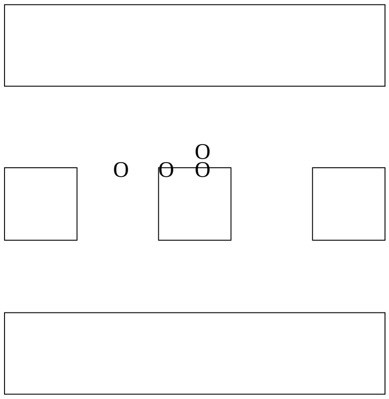


Figure V.12: Stress measurement points for BRI fill structures. Side view. Top boundary measurements for $b0$-$b1$.

In Table V.4, we show results computed using the measurements that are used to generate Table V.3. The second column is the σ_{zz}, i.e., normal component, stress difference between $m1$ and $m5$. The third column (difference along wire at the center) is the difference between the average stress at $m1$ and $m2$. The last column is the difference between the average stress at $m1$ and $m3$. Although we do not use this term directly, it may be another helpful criterion to decide on the stress migration impact.
Using our simulations, we come up with the following observations.

- The cap layer usually introduces a compressive stress, whereas the interconnects and dielectrics have mostly tensile stresses.

- The stress in the dielectrics are lower as compared to the stress in the interconnects.

Our observations on the individual test structures are given below.

it is not possible to determine the stress due to thermal mismatch only using our simulation results without conducting an additional sensitivity analysis.

V.C.1 Impact Due to Fill Pattern

We can observe from Table V.3 and Table V.4 that the two-pass, staggered and traditional fill patterns result in lowest average $m1$ stress, in that order. The $m1$ and $m5$ normal component difference follows the same order. For the von Mises stress at $m3$, the lowest stress occurs with two-pass, traditional and staggered fill patterns in that order. As can be seen from Table V.4 comparing $p0$ and $p1$-$p3$, insertion of fills between intralayer parallel lines increases the normal stress component difference between $m1$ and $m5$.

V.C.2 Impact of Fill On Neighboring Layer Orthogonal Lines

Using our test structures $t0$ to $t2$, we have seen that inserting a fill near first neighboring layer orthogonal lines reduces the average stress at $m1$ by 13.17%. The dielectric stress at $m4$ on the other hand increases by 28.57%. Shifting the fills away as in structures $t1$ or $t3$ increases the average stress at $m1$ by at least 4.74% and the von Mises stress at $m3$ by at least 3.61%. Insertion of a via fill reduces the difference of normal components between $m1$ and $m5$ by 86.36%.

V.C.3 Impact of Fill On Second Neighboring Layer Parallel Lines

Using our test structures $s0$ to $s2$, we observe from Table V.3 that introducing a fill between second neighboring parallel lines reduces the average stress at $m1$ by 11.16%. However, von Mises stress at $m3$ increases by 4.97%. Based on Table V.4, the normal stress component difference at $m1$ and $m5$ reduces by 60%.

If the fill is shifted away from the overlap of the interconnects in the common neighboring layer, the normal stress difference increases by 37.5%. Yet this improves the von Mises stress at $m3$ instead by 9.11%.

V.C.4 Impact of Via Fill

We can observe from Table V.4 that introducing via fills to structure $v0$ in $v1$ reduces the normal component difference by 62.96%.

V.C.5 Impact of Process

Table V.5 shows the impact of process parameters on the traditional fill pattern test structure $p1$. Increasing the process temperature from 250^oC (row 1) to 400^oC (row2) increases the average stress at $m1$ by 67.12%. The von Mises stress at $m3$ increases by 66.80%. Similarly, the dielectric average stress at $m4$ increases by 66.00%. Table V.6 shows additional computations based on the same data as in Table V.5. The stress difference between the normal components of $m1$ and $m5$ increases by 65.22% when temperature is increased to 400^oC.

Increasing the metal aspect ratio (AR) by 25% by keeping the metal width constant decreases the average stress at $m1$ by 3.65%, yet the dielectric average stress increases by 2.00 %. Von Mises stress at $m3$ reduces by 9.04%. The difference in normal stress components increases by 23.91% when the aspect ratio is increased by 25%.

Increasing the via height by 25% while keeping other parameters constant decreases the average stress at $m1$ by 5.79%, while the stress at $m4$ increases by 2.00%. Von Mises stress at $m3$ decreases by 2.53%. The difference in normal stress components increases by 6.52%.

In summary, we have seen that the process temperature has the most effect on final stress values. A 60% increase in temperature may bring similar increase in stress. To reduce stress, a low temperature process needs to be chosen.[4] Although aspect ratio and via heights have a nonnegligible effect on the final stress, their effect on stress is less than 10% for most cases for a change of 25% in height.

[4] As process temperature can alter the nominal stress values considerably, we summarize percentage changes rather than nominal changes during our analyses.

V.C.6 Impact of Fill Size

We can observe from Table V.7 that increasing the fill size in a traditional fill pattern by 50% on each side increases the average stress at $m1$ by 4.11%. The dielectric stress at $m4$ increases by 10.00%. Von Mises stress at $m3$ reduces by a mere 1.17%. Based on Table V.8, the normal component differences increase by 8.70%.

Adding via fills (with respect to increased fill sizes only) to all the fills so that the floating fills are connected to the neighboring layer interconnects (or rectangular fills) reduces the average stress at $m1$ by 3.51%. Dielectric stress at $m4$ increases by 3.63%. From Table V.8, the normal component differences decrease by 60%.

V.C.7 Analysis of BRI Test Structures

We can observe from Table V.9 that inserting via fills on an interconnect may slightly increase the stress components. However based on Table V.10, the normal component differences decrease by 42.86%.

V.D Design Guidelines for Reliability Improvement

Based on the analyses from the previous section, we have come up with the following design and process guidelines. We first itemize the top priority ones, which cause improvement over 20%[56].

- Reduce process temperature to improve reliability.
- Insert via fills to reduce delamination and crack starts.

[5] Notice that the improvements will be less if the temperature, aspect ratio and via height changes are less than what is studied here.

[6] We have assumed a linear relationship between stress and reliability improvement. However, this may not be the case and this relationship may vary based on the reliability concern as well as the technology used.

- Insert a fill between second neighboring parallel lines to decrease the delamination and crack formation possibility.

Next, we itemize design and process guidelines with secondary improvements (8%-20%).

- Insert fill near orthogonal or parallel lines on neighboring layers to reduce stress migration.
- Increase aspect ratio to decrease Cu^+ diffusion possibility.
- Decrease aspect ratio to decrease the delamination and crack formation possibility.
- Reduce fill size to reduce delamination.

V.E Conclusions

Using a TCAD-based analysis and specific test structures we have designed, we have evaluated the impact of CMP fills on first and second neighboring layer interconnects, impact of CMP fill pattern on intralayer parallel lines and impact of via fill. Our test structures target replication of common layout configurations in the presence of fills. We have provided stress measurement points to tie the stress values to known reliability issues. We have provided design guidelines and structures to improve BEOL reliability. In particular, we have identified that via fills can reduce delamination possibility in neighboring layer interconnects significantly.

V.F Acknowledgments

We would like to thank Kok-Yong Yiang, Charlie Jun Zhai and Melida Chin of AMD for the productive discussions. This chapter is based on or in part a reprint of the following publication.

- A. B. Kahng and R. O. Topaloglu, "A TCAD-based study of fill pattern and via fill impact on low-k dielectric stress," **Invited Paper**, *Proc. International Chemical-Mechanical Planarization for ULSI Multilevel Interconnection Conference (CMP-MIC)*, 2007, pp. 337-346.

Table V.3: Top Boundary Stress Measurements (in MPa) for the Test Structures.

	top boundary									
	int. ctr.		ctr. shft.		int. edge		diel. ctr.		diel. edge	
str.	σ_{vm}	σ_{av}	σ_{vm}	σ_{av}	σ_{vm}	σ_{av}	σ_{vm}	σ_{av}	σ_{vm}	σ_{av}
$p0$	488	207	472	201	524	142	69	41	493	-245
$p1$	483	219	491	213	512	177	68	50	414	-239
$p2$	479	202	473	214	508	159	70	44	426	-249
$p3$	470	207	471	208	532	162	69	50	416	-238
$t0$	501	243	462	221	536	178	75	42	604	-265
$t1$	477	221	446	219	545	190	72	45	526	-255
$t2$	450	211	476	231	526	203	72	54	509	-250
$t3$	464	233	465	225	563	189	75	43	563	-254
$t4$	490	239	466	227	463	218	74	55	590	-258
$t5$	516	221	445	215	523	192	70	54	531	-243
$t6$	454	208	451	230	536	196	73	56	530	-247
$t7$	474	206	469	214	553	176	71	42	447	-221
$s0$	477	224	508	207	523	174	69	43	522	-240
$s1$	504	199	515	186	549	153	71	43	492	-255
$s2$	509	217	480	204	499	171	67	43	561	-253
$v0$	505	238	511	249	518	196	68	53	463	174
$v1$	667	295	468	217	475	207	63	52	418	165

Table V.4: Top Boundary Stress Computations (in MPa) for the Test Structures.

str.	σ_{zz} diff.	diff. along wire at ctr.	ctr. vs. edge diff.
$p0$	7	7	65
$p1$	46	6	42
$p2$	39	-12	43
$p3$	42	-1	45
$t0$	7	21	65
$t1$	20	2	31
$t2$	22	-20	9
$t3$	21	9	44
$t4$	-3	13	21
$t5$	-19	6	29
$t6$	7	-22	12
$t7$	15	-9	30
$s0$	20	17	50
$s1$	8	12	46
$s2$	-11	14	47
$v0$	27	-11	42
$v1$	10	78	88

Table V.5: Top Boundary Stress Measurements (in MPa) for Process Sensitivity of Traditional Pattern Test Structure ($p1$).

	top boundary									
	int. ctr.		ctr. sftd.		int. edge		diel. ctr.		diel. edge	
str.	σ_{vm}	σ_{av}	σ_{vm}	σ_{av}	σ_{vm}	σ_{av}	σ_{vm}	σ_{av}	σ_{vm}	σ_{av}
250^oC	483	219	491	213	512	177	68	50	414	-239
400^oC	806	366	816	355	854	295	111	83	692	-399
1.25x AR	451	211	460	199	505	161	67	51	432	-249
1.25x Via	463	207	483	212	499	161	68	51	412	-242

Table V.6: Top Boundary Stress Computations (in MPa) for Process Sensitivity of Traditional Pattern Test Structure ($p1$).

str.	σ_{zz} diff.	diff. along wire at ctr.	ctr. vs. edge diff.
250^oC	46	6	42
400^oC	76	11	70
1.25x AR	57	13	50
1.25x Via	49	-6	46

Table V.7: Top Boundary Stress Measurements (in MPa) for Layout Sensitivity of Traditional Pattern Test Structure ($p1$).

	top boundary									
	int. ctr.		ctr. sftd.		int. edge		diel. ctr.		diel. edge	
str.	σ_{vm}	σ_{av}	σ_{vm}	σ_{av}	σ_{vm}	σ_{av}	σ_{vm}	σ_{av}	σ_{vm}	σ_{av}
$p1$	483	219	491	213	512	177	68	50	414	-239
$p1a$	483	228	469	212	506	176	66	55	403	-232
$p1b$	482	220	457	214	504	168	72	57	469	-240

Table V.8: Top Boundary Stress Computations (in MPa) for Layout Sensitivity of Traditional Pattern Test Structure ($p1$).

str.	σ_{zz} diff.	diff. along wire at ctr.	ctr. vs. edge diff.
$p1$	46	6	42
$p1a$	50	16	52
$p1b$	20	6	52

Table V.9: Top Boundary Stress Measurements (in MPa) for BRI Test Structures.

	top boundary									
	int. ctr.		ctr. sftd.		int. edge		diel. ctr.		diel. edge	
str.	σ_{vm}	σ_{av}	σ_{vm}	σ_{av}	σ_{vm}	σ_{av}	σ_{vm}	σ_{av}	σ_{vm}	σ_{av}
$b0$	475	204	491	218	521	194	64	50	556	-234
$b1$	485	213	485	212	525	210	66	53	442	-235

Table V.10: Top Boundary Stress Computations (in MPa) for BRI Test Structures.

str.	σ_{zz} diff.	diff. along wire at ctr.	ctr. vs. edge diff.
$b0$	-35	-14	10
$b1$	20	0	2

VI

Fill Thermal Impact Analysis

Device temperature needs to be reduced to improve its performance and reduce leakage. Device temperature may be reduced by efficiently utilizing fill structures and optimizing layout. In this chapter, we conduct a TCAD-based analysis on the thermal effects of fills. In particular, we evaluate whether device temperature can be reduced using CMP, via and contact fills. Fills can help improve the device temperature if properly used. Based on our analysis, we provide design guidelines to reduce temperatures through layout optimizations.

In Section VI.A, we introduce the test structure layouts for our fill thermal analysis study. In Section VI.B, we present our TCAD-based simulation setup. In Section VI.C, we present and analyze our results. In Section VI.D, we provide guidelines for the designers to reduce device temperature through fills. We then conclude the chapter.

VI.A Test Structures to Evaluate the Impact on Temperature

In this section, we provide the test structure layouts and descriptions we use in our analysis. We select each addition as an incremental change with respect to the reference or one of the structures below so that the sensitivity with respect to the added option can be traced from the simulation results. We keep the design rules such as via to channel spacing, via to via spacing, etc. constant across the structures for a fair evaluation. There are two main reference structures, one with regular active area and one with larger active area. The purpose of each test structure is given below:

Reference: reference structure. The structure has two contacts (CAs) on each side connecting the diffusion layers to $M1$ interconnects. This structure is given in Figures VI.1 and VI.2.

In the figures, P stands for polysilicon, ACT stands for active region (or diffusion), BOX stands for the buried oxide, STI stands for shallow trench isolation, CA stands for contact, $M1$ and $M2$ are first and second layer interconnects, respectively. In layout views, metal lines are shaded. Dotted lines indicate underlying structures. $V1$ and $V2$ are contacts over interconnect layers $M1$ and $M2$, respectively.

Sparse CA's: single CA's, instead of two on each side, connect drain and source diffusions to $M1$ interconnects, which are parallel to the polysilicon gate. This structure evaluates the impact of sparse CA's with respect to the reference.

$M2$ **Route:** includes $M2$ interconnects. This test structure evaluates whether routing $M2$ interconnects on top of a transistor alters the device temperature with respect to the reference. This structure is given in Figures VI.3 and VI.4.

$M2$ **Via Fill:** evaluates the impact of having $M2$ fills which are connected to $M1$ interconnects through via fills with respect to the reference structure. This structure is given in Figures VI.5 and VI.6. The side view corresponds to the cross-section indicated by the thin dashed line in Figure VI.5.

CA **Fills Over Poly:** CA fills (dummy CAs) over polysilicon are added with respect to the reference. This structure is given in Figures VI.7 and VI.8.[1]

Large Active Reference: with respect to the reference, the diffusion area is enlarged past the CA's. This structure is given in Figures VI.9 and VI.10.

[1] Such CA fills can only be placed if the polysilicon is large enough or when body-tied devices are used in SOI.

CA **Fills:** evaluates the impact of having CA fills over the active region with respect to the large active reference. This structure is given in Figures VI.11 and VI.12.

CA **Fills with** $M1$ **Connected:** CA fills are connected through the $M1$ layer interconnects with respect to the previous structure. This structure is given in Figures VI.13 and VI.14.

Sparse CA **Fills with** $M1$ **Connected:** sparse active CA fills with $M1$ connected with respect to the previous structure. This structure is given in Figures VI.15 and VI.16.

Sparse CA **Fills with** $M1$ **Fill:** a larger fill is used in $M1$ layer with respect to the previous structure. This structure is given in Figures VI.17 and VI.18.

Poly Fills: evaluates the impact of poly fills over the active region with respect to the large reference structure. This structure is given in Figures VI.19 and VI.20.

Poly Fills with CA **Fills:** CA fills are added on the poly fills with respect to the previous structure. This structure is given in Figures VI.21 and VI.22.

Poly Fills with CA **Fills and** $M1$ **Route:** CA fills are connected in the $M1$ level with respect to the previous structure. This structure is given in Figures VI.23 and VI.24.

Poly Fills with CA **Fills and** $M1$ **Fill:** CA fills are connected using a large $M1$ fill with respect to the previous structure. This structure is given in Figures VI.25 and VI.26.

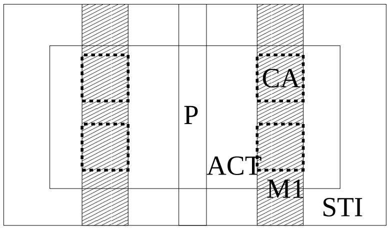


Figure VI.1: Reference test structure. CA is contact. P is polysilicon. ACT is active region.

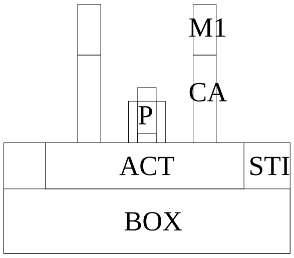


Figure VI.2: Reference test structure. Side view. CA is contact. P is polysilicon. ACT is active region.

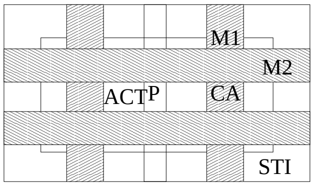


Figure VI.3: $M2$ test structure.

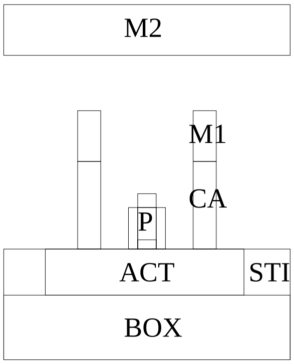


Figure VI.4: $M2$ test structure. Side view.

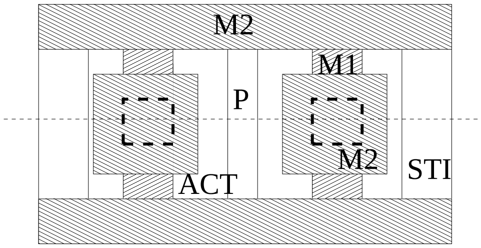


Figure VI.5: $M2$ via fill test structure.

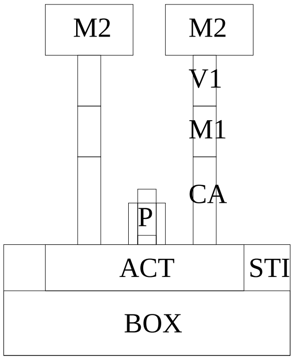


Figure VI.6: $M2$ via fill test structure. Side view.

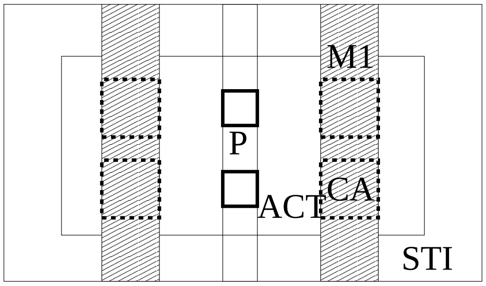


Figure VI.7: Poly CA fills test structure.

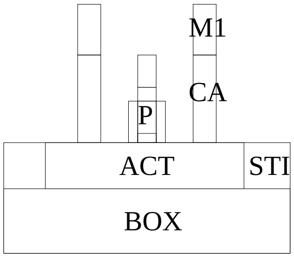


Figure VI.8: Poly CA fills test structure. Side view.

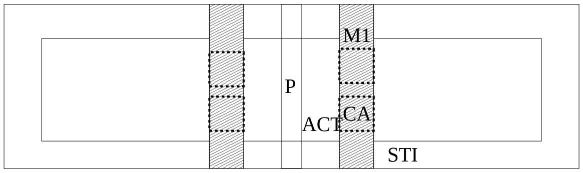


Figure VI.9: Large active test structure.

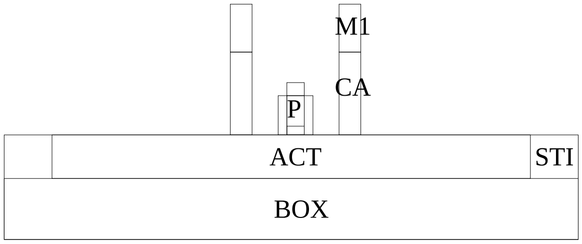


Figure VI.10: Large active test structure. Side view.

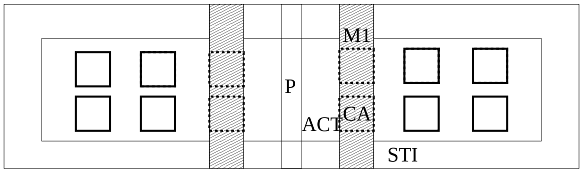


Figure VI.11: *CA* fills test structure.

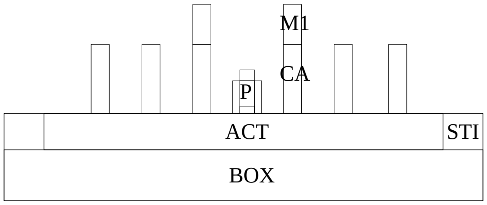


Figure VI.12: CA fills test structure. Side view.

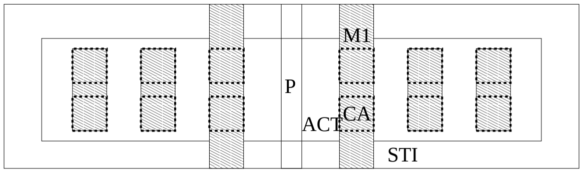


Figure VI.13: CA fills with $M1$ connected test structure.

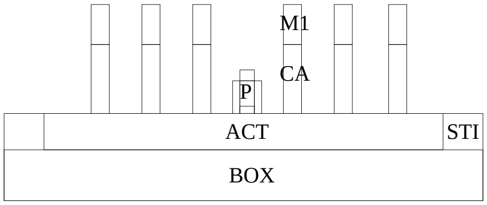


Figure VI.14: CA fills with $M1$ connected test structure. Side view.

VI.B Experimental Setup

To conduct our analysis, we use the physical parameter heights shown in Table VI.1. We use the design rules as given in Table VI.2. In terms of CTC's, we use the values in Table VI.3. We use the field solver tool Raphael [115] for the simulations.

We assume that a $30nm$ region underneath the channel dissipates 0.25W/μm^3. Another $30nm$ region at the bottom of the substrate is assumed to be at a fixed temperature of 25^oC to simulate a heat sink connected to the substrate. We use reflective boundary conditions. We simulate the interconnect stack up to and including half height of the dielectric over $M2$. We select the substrate height as $0.5\mu m$.

Table VI.1: Process Physical Parameter Heights in μm.

$M1$	V1	CA over diffusion	CA over poly	silicide
0.15	0.1	0.25	0.12	0.03
$M2$	V2	SOI	active	spacer
0.15	0.15	0.075	0.075	0.03

Table VI.2: Design Rules in μm.

$M1$ width	$M2$ width	CA width	Poly width
0.1	0.1	0.1	0.65
Poly-CA spacing	Min CA-CA spacing	Min fill width	Min pitch
0.1	0.1	0.3	1.87

Table VI.3: Material CTC's in W/mK.

copper	polysilicon	diffusion	dielectric	silicide	spacer
385	110	110	1	20	20

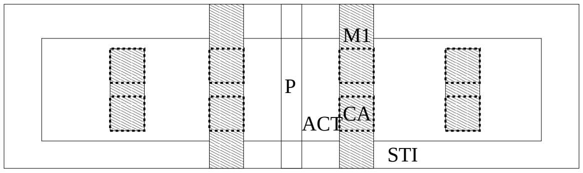


Figure VI.15: Sparse CA fills with $M1$ connected test structure.

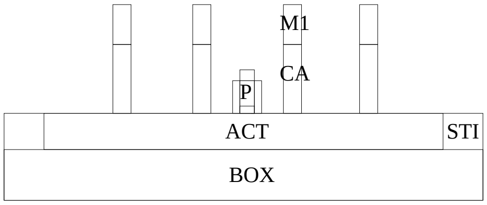


Figure VI.16: Sparse CA fills with $M1$ connected test structure. Side view.

VI.C Experimental Results and Analysis

The results of our experiments are summarized in Table VI.4. We observe that making the contacts sparse increases the temperature by 2.86^oC, which roughly corresponds to a 6% increase in leakage. Dummy poly CAs do not result in any improvement. Routing $M2$ layer interconnect without physical connection to lower layers results in 2.66^oC increase. Introducing via fills connecting $M2$ fills to $M1$ results in 4.55^oC reduction in temperature.

We see that extending the source and drain diffusion past the CA's closest to the channel, 35.42^oC reduction in temperature is possible with respect to the reference device. If CA fills are present, introducing $M1$ fills so that they are connected to diffusion through the CA fills results in 2.85^oC reduction in temperature. We explain the main reason of such a temperature reduction as the same

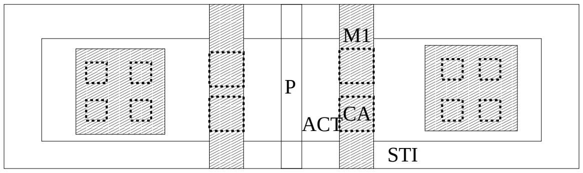


Figure VI.17: CA fills with $M1$ fill test structure.

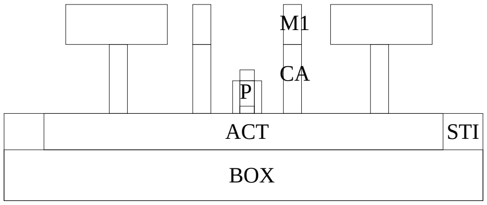


Figure VI.18: CA fills with $M1$ fill test structure. Side view.

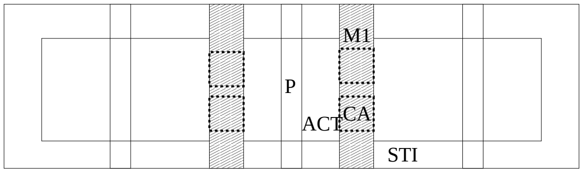


Figure VI.19: Poly fills test structure.

heat source having to warm up a larger volume of material. We see that poly fills do not improve temperature, but if they are present, CA fills result in 2.85^oC reduction.

VI.D Proposed Design Guidelines

Prior to providing the design guidelines for lower leakage through reduced temperature, it might be beneficial to visit some of the layout design guidelines for standard cells. Poly fills are advantageous for multiple reasons, such as reducing lithographic variations and variations due to stress. CA fills may be used to provide lithographic printability and reduce contact resistance variations. Extending the source and drain diffusion regions can benefit from reducing the STI impact. On

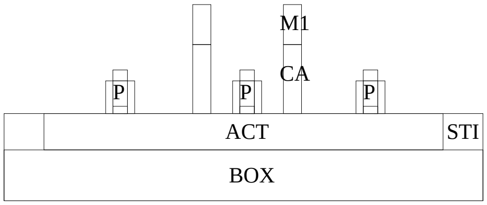


Figure VI.20: Poly fills test structure. Side view.

Table VI.4: Temperature Measurements in ^oC.

Structure Name	Channel Temp.
Reference layout	97.64
Sparse CA's	100.5
$M2$ route	100.3
$M2$ via fill	95.75
CA fills over poly	100.3
Large active reference	62.22
CA fills	64.66
CA fills with $M1$ connected	64.62
Sparse CA fills with $M1$ connected	61.75
CA fills with $M1$ fill	61.81
Poly fills	64.5
Poly fills with CA fills	61.65
Poly fills with CA fills and $M1$ route	61.91
Poly fills with CA fills and $M1$ fill	62.24

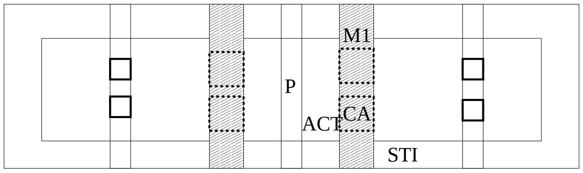


Figure VI.21: Poly fills with CA fills test structure.

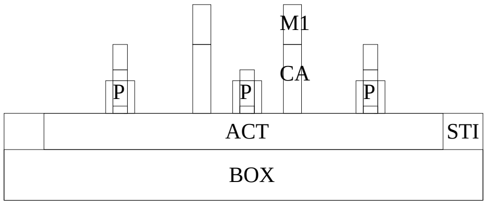


Figure VI.22: Poly fills with CA fills test structure. Side view.

the other hand, larger diffusion regions may increase the interconnect to diffusion coupling capacitances. Using our analysis, we come up with the following design guidelines to reduce leakage through lower temperature.

- Extend source and drain diffusion regions past the closest vias whenever possible. This change results in negligible capacitance increase but 30^oC temperature reduction.
- Make contacts dense to get a 6% reduction in leakage.
- When there are poly fills, insert CA fills over the poly fills.
- When CA fills over diffusion are present, insert $M1$ fills.
- Via fills connecting fills to lower layers can reduce the leakage.

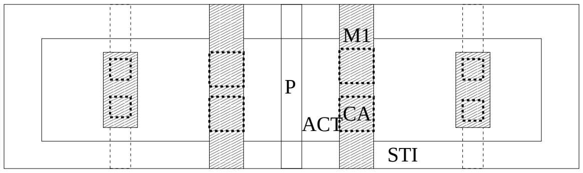


Figure VI.23: Poly fills with CA fills and $M1$ route test structure.

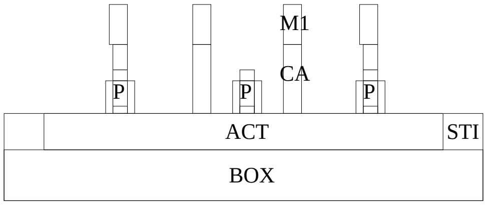


Figure VI.24: Poly fills with CA fills and $M1$ route test structure. Side view.

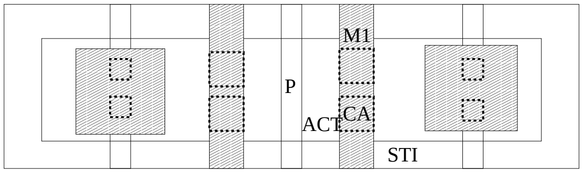


Figure VI.25: Poly fills with CA fills with $M1$ fill test structure.

- Routing metal lines over transistors without physical connections will increase temperature.

VI.E Discussion and Conclusions

Through TCAD simulations, we have analyzed the impact of temperature as a result of transistor-level layout modifications such as addition of via fills, contact fills and diffusion region extensions. We have found that making contacts denser will improve leakage by 6% through a reduction in temperature. Via fills in lower metal layers can result in lower leakage through reduced channel temperatures. The temperature reduction can go above 30^oC by enlarging the active region area past the vias closest to the channel. We have provided design guidelines to take advantage of these observations. These guidelines can result in temperature

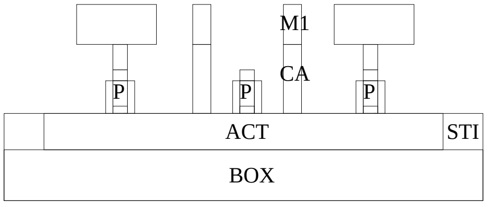


Figure VI.26: Poly fills with CA fills with $M1$ fill test structure. Side view.

reductions with negligible cost. Analysis of current cell libraries and custom circuit blocks have revealed potential leakage reduction possibilities through temperature-aware device design. Our ongoing work targets optimization of larger layouts and using proper subsets of the proposed guidelines. Using the guidelines provided in this chapter, it will be possible to improve leakage in standard cell and custom circuit designs at no area and performance cost in many cases. Using available area and design margins, it is possible to significantly boost the leakage savings through reduced temperature.

VI.F Acknowledgments

We would like to thank Frank Geelhaar of AMD for very helpful discussions of methodology to simulate thermal effects.

VII

Conclusions

In this thesis, we have provided solutions for accurate floating fill extraction and optimization, FEOL stress modeling and optimization, BEOL stress analysis and guidelines for reliability-aware fill insertion and FEOL thermal analysis in the presence of fills and thermal improvements using fills. In the thesis, we have followed a methodology where we (i) design test structures, (ii) conduct TCAD simulations, (iii) develop models, (iv) provide design guidelines, (v) optimize circuits, and (vi) automate the optimization process.

In Chapter II, we have provided a DOE-based methodology and analyses based on TCAD simulations for accurate extraction of fill impact on capacitances. We have provided a compact DOE set. We have provided an integration methodology based on normalization to be used with industrial extraction flows. We have analyzed the impact of floating fills on capacitances and compared different fill algorithms and configurations. We have provided this framework to designers for them to conduct such experiments on their technology and designs.

In Chapter III, we have provided a fill synthesis framework based on fill insertion guidelines to reduce capacitances and improve circuit timing. We have implemented our framework on energy minimization principles using a grid model consisting of bond energies with adjustable models. We have enabled interlayer- and power-aware fill options. We have applied our technique on large testcases to improve performance in terms of timing and power.

In Chapter IV, we have provided models for STI width stress effect for CMOS devices. We have also provided a similar analysis and guidelines for dual stress liner technology. We have presented a standard-cell optimization method and used it to improve performance of circuits.

In Chapter V, we have conducted TCAD analyses on BEOL stress and related the impact of fills on BEOL reliability. We have also provided corresponding design of test structures. We have identified stress measurement points to relate stress to low-k BEOL reliability concerns such as stress migration and delamination. We have then provided design guidelines based on our observations. In particular, we have identified that via fills can be used to improve BEOL reliability.

In Chapter VI, we have analyzed the impact of fills on thermal performance of devices and provided test structures and guidelines for fill utilization for thermal performance improvement. In particular, we have identified that via and contact fills can be utilized to reduce device temperature.

With the help of the proposed techniques, frameworks, analyses and guidelines, it will be possible to manage and optimize fill structures and stress, which seem to be mandatory for semiconductor manufacturing for upcoming technology generations. With our contributions, design pessimism and underutilized resources will reduce, and design performance will improve considerably.

Looking forwards, at the $22nm$ technology node, new or improved device types may be introduced. Possible candidates are finFETs, germanium-channel MOSFETs [127][128] or III-V devices incorporating, e.g., GaAS and InAs [129][130]. Stress to improve performance will still be necessary both for finFETs [131] and germanium channel devices [132]. Hence, the stress analysis, modeling and optimization topics will still be prevalent.

Future BEOL technology will contain ultra low-k dielectrics with dielectric constants lower than 2.0 [133], as well as airgaps [134][135][136][137]. Furthermore, mechanical stress analyses will be necessary to be able to design circuits with such new materials and structures. With airgaps, fills will need to be analyzed both in terms of electrical performance as well as mechanical stress stability and reliability. Reliability concerns such as stress induced voiding [138] and stress migration [139] will still be issues. Efficient test structures will be needed to analyze these effects [140][141].

Device and chip heating will be more of an issue in future integrated-circuit designs [142][143]. There will be rectangular contacts, possibly using copper, and contact routes. The thermal analysis methods that we have used, along with test structures, will be required to analyze new device types in the presence of new contact and BEOL modules. Such analyses may be extended to heat pipe design with possible introduction of highly thermally conductive yet electrically insulative materials.

Finally, packaging will continue to be of increased interest [144][145], particularly as the lithography cost roadmap becomes increasingly daunting. With the usage of stacked chips and TSVs (through silicon vias) [146], stress, thermal and electrical performance co-analyses will be increasingly needed. The test structure-based TCAD simulation and automated optimization methodologies that we have developed will constitute the basics to help analyze and optimally design semiconductor integrated circuits.

Based on this projection, we believe that the methods proposed in this thesis are sufficient for 45nm technology and can be applied down to 22nm technology with appropriate enhancements. We hope that this thesis will enable a faster transition to such advanced technologies by providing the necessary tools and framework for analysis and optimization of performance and reliability, in addition to providing increased performance and reliability in current technologies.

Bibliography

[1] A. B. Kahng, K. Samadi and P. Sharma, "Study of floating fill impact on interconnect capacitance," *Proc. IEEE International Symposium on Quality Electronic Design*, 2006, pp. 691-696.

[2] R. O. Topaloglu, "Energy minimization model for fill synthesis," *Proc. IEEE International Symposium on Quality Electronic Design*, 2007, pp. 444-451.

[3] Y. Kim, S. Sridhar and A. Chatterjee, "Trench isolation step-induced (TRISI) narrow width effect on MOSFET," *IEEE Electron Device Letters*, 23(10), 2002, pp. 600-602.

[4] A. B. Kahng and K. Samadi, "CMP fill synthesis: a survey of recent studies," *IEEE Trans. on Computer-Aided Design*, 27(1), 2008, pp. 3-19.

[5] B. Stine, D. Boning and J. Chung, "Modeling pattern dependent variation in CMP processes: Process, consumable, and tool comparison methodologies, metal fill patterning practices, and other issues," *CMP User's Group (CMPUG)*, 1996.

[6] S. Lakshminarayanan, P. J. Wright and J. Pallinti, "Electrical characterization of the copper CMP process and derivation of metal layout rules, *IEEE Trans. on Semiconductor Manufacturing*, 16(4), 2003, pp. 668-676.

[7] P. Zarkesh-Ha, S. Lakshminarayann, K. Doniger, W. Loh and P. Wright, "Impact of interconnect pattern density information on a 90 nm technology ASIC design flow," *Proc. IEEE International Symposium on Quality Electronic Design*, 2003, pp. 405-409.

[8] O. Cueto, F. Charlet and A. Farcy, "An efficient algorithm for 3D interconnect capacitance extraction considering floating conductors," *Proc. International Conference on Simulation of Semiconductor Processes and Devices*, 2002, pp. 107-110.

[9] W. Yu, M. Zhang and Z. Wang, "Efficient 3-D extraction of interconnect capacitance considering floating metal fills with boundary element method," *IEEE Trans. on Computer-Aided Design*, 25(1), 2006, pp. 12-18.

[10] X. Wang, C. C. Chiang, J. Kawa, Q. Su, "A min-variance iterative method for fast smart dummy feature density assignment in chemical-mechanical polishing," *Proc. IEEE International Symposium on Quality Electronic Design*, 2005, pp. 258-263.

[11] M. Nelson, B. Williams, C. Belisle, S. Aytes, D. Beasterfield, J. Liu, S. Donaldson and J. Prasad, "Optimizing pattern fill for planarity and parasitic capacitance," *International Semiconductor Device Research Symposium*, 2003, pp. 428-429.

[12] S. Batterywala, R. Ananthakrishna, Y. Luo and A. Gyure, "A statistical method for fast and accurate capacitance extraction in the presence of floating dummy fills," *Proc. International Conference on VLSI Design*, 2006, pp. 129-134.

[13] A. Kurokawa, T. Kanamoto, A. Kasebe, Y. Inoue and H. Masuda, "Efficient capacitance extraction method for interconnects with dummy fills," *Proc. IEEE Custom Integrated Circuits Conference*, 2004, pp. 485-488.

[14] K.-H. Lee, J.-K. Park, Y.-N. Yoon, D.-H. Jung, J.-P. Shin, Y.-K. Park and J.-T. Kong, "Analyzing the effects of floating dummy-fills: from feature scale analysis to full-chip RC extraction," *Proc. IEEE International Electron Devices Meeting*, 2001, pp. 31.3.1-31.3.4.

[15] A. Kurokawa, T. Kanamoto, T. Ibe, A. Kasebe, C. W. Fong, T. Kage, Y. Inoue and H. Masuda, "Dummy filling methods for reducing interconnect capacitance and number of fills," *Proc. IEEE International Symposium on Quality Electronic Design*, 2005, pp. 586-591.

[16] J.-K. Park, K.-H. Lee, J.-H. Lee, Y.-K. Park and J.-T. Kong, "An exhaustive method for characterizing the interconnect capacitance considering the floating dummy-fills by employing an efficient field solving algorithm," *Proc. International Conference on Simulation of Semiconductor Processes and Devices*, 2000, pp. 98-101.

[17] Y. W. Chang, H. W. Chang, T. C. Lu, Y. King, W. Ting, J. Ku and C. Y. Lu, "A novel CBCM method free from charge injection induced errors: investigation into the impact of floating dummy-fills on interconnect capacitance," *Proc. International Conference on Microelectronic Test Structures*, 2005, pp. 235-238.

[18] W.-S. Lee, K.-H. Lee, J.-K. Park, T.-K. Kim, Y.-K. Park and J.-T. Kong, "Investigation of the capacitance deviation due to metal-fills and the effective interconnect geometry modeling," *Proc. IEEE International Symposium on Quality Electronic Design*, 2003, pp. 373-376.

[19] Y. Kim, D. Petranovic and D. Sylvester, "Simple and accurate models for capacitance increment due to metal fill insertion," *Proc. Asia and South Pacific Design Automation Conference*, 2007, pp. 456-461.

[20] "Quickcap," http://www.magma-da.com/products-solutions/verification/quickcap.aspx

[21] "Raphael NXT," http://www.synopsys.com/products/tcad/raphael_nxt_ds.html

[22] P. R. Findley, "Methods and apparatus for extracting parasitic capacitance values from a physical design of an integrated circuits," *U.S. Patent No. 6,243,653*.

[23] Y. Chen, A. B. Kahng, G. Robins and A. Zelikovsky, "Area fill synthesis for uniform layout density," *IEEE Trans. on Computer-Aided Design* 21(10), 2002, pp. 1132-1147.

[24] B. E. Stine et al., "The physical and electrical effects of metal-fill patterning practices for oxide chemical-mechanical polishing processes," *IEEE Trans. on Electron Devices*, 45(3), 1998, pp. 665-679.

[25] T. Tugbawa et al., "A mathematical model of pattern dependencies in Cu CMP processes," *Proc. Electrochemical Society Meeting*, 1999.

[26] A. B. Kahng, G. Robins, A. Singh, H. Wang and A. Zelikovsky, "Filling algorithms and analyses for layout density control," *IEEE Trans. on Computer-Aided Design* 18(4), 1999, pp. 445-462.

[27] Y. Chen, P. Gupta and A. B. Kahng, "Performance-impact limited area fill synthesis," *Proc. ACM/EDAC/IEEE Design Automation Conference*, 2003, pp. 22-27.

[28] A. B. Kahng, G. Robins, A. Singh, H. Wang and A. Zelikovsky, "Filling and slotting: analysis and algorithms," *Proc. ACM/IEEE International Symposium on Physical Design*, 1998, pp. 95-102.

[29] H. Xiang, L. Deng, R. Puri, K.-Y. Chao and M. D. F. Wong, "Dummy fill density analysis with coupling constraints," *Proc. ACM/IEEE International Symposium on Physical Design*, 2007, pp. 3-9.

[30] H. Xiang, K.-Y. Chao, R. Puri and M. D. F. Wong, "Is your layout density verification exact? - a fast exact algorithm for density calculation," *Proc. ACM/IEEE International Symposium on Physical Design*, 2007, pp. 19-26.

[31] A. B. Kahng and R. O. Topaloglu "A DOE set for normalization-based extraction of fill impact on capacitances, *Proc. IEEE International Symposium on Quality Electronic Design*, 2007, pp. 467-474.

[32] T. Chan, J. Cong and K. Sze, "Multilevel generalized force-directed method for circuit placement," *Proc. ACM/IEEE International Symposium on Physical Design*, 2005, pp. 185-192.

[33] P. G. Paulin and J. P. Knight, "Force-directed scheduling for the behavioral synthesis of ASICs," *IEEE Trans. on Computer-Aided Design*, 8(6), 1989, pp. 661-678.

[34] T. Zhang, Y. Zhan and S. S. Sapatnekar, "Temperature-aware routing in 3D ICs," *Proc. Asia and South Pacific Design Automation Conference*, 2006, pp. 309-314.

[35] J. Cong, G. Luo, J. Wei and Y. Zhang, "Thermal-aware 3D IC placement via transformation," *Proc. Asia and South Pacific Design Automation Conference*, 2007, pp. 780-785.

[36] C.-H. Tsai and S.-M. S. Kang, "Standard cell placement for even on-chip thermal distribution," *Proc. ACM/IEEE International Symposium on Physical Design*, 1999, pp. 179-184.

[37] J. Cong and Y. Zhang, "Thermal via planning for 3-D ICs," *Proc. IEEE/ACM International Conference on Computer-Aided Design*, 2005, pp. 745-752.

[38] B. Goplen and S. Sapatnekar, "Efficient thermal placement of standard cells in 3D ICs using a force directed approach," *Proc. IEEE/ACM International Conference on Computer-Aided Design*, 2003, pp. 86-89.

[39] C.-H. Tsai and S.-M. S. Kang, "Macrocell placement with temperature profile optimization," *Proc. IEEE International Symposium on Circuits and Systems*, 1999, pp. 390-393.

[40] W.-L. Hung, G. M. Link, Y. Xie, N. Vijaykrishnan and M. J. Irwin, "Interconnect and thermal-aware floorplanning for 3D microprocessors," *Proc. IEEE International Symposium on Quality Electronic Design*, 2006, pp. 98-104.

[41] W-L. Hung et al., "Thermal-aware floorplanning using genetic algorithms," *Proc. IEEE International Symposium on Quality Electronic Design*, 2005, pp. 634-639.

[42] K. K. Lee, E. J. Paradise and S. K. Lim, "Thermal-driven circuit partitioning and floorplanning with power optimization," *Georgia Institute of Technology CERCS Technical Report*, 2003.

[43] M. Cho, S. Ahmed and D. Z. Pan, "TACO: temperature aware clock-tree optimization," *Proc. IEEE/ACM International Conference on Computer-Aided Design*, 2005, pp. 582-587.

[44] B. Li, D. Harmon, J. Gill, F. Chen and T. Sullivan, "Thermal and electromigration challenges for advanced interconnects," *Proc. International Reliability Workshop*, 2004, pp. 46-51.

[45] M. R. Casu, M. Graziano, G. Masera, G. Piccinini, M. Zamboni, "Power supply wire sizing considering self-heating in bulk-to-SOI migrated designs," *Proc. International Workshop on Power and Timing Modeling, Optimization and Simulation*, 2001, pp. 8.3.1-8.3.10.

[46] K. Banerjee, "Thermal effects in deep submicron VLSI interconnects," *Tutorial at IEEE International Symposium on Quality Electronic Design Conference*, 2000.

[47] T. Y. Chiang, K. Banerjee and K. C. Saraswat, "Compact modeling and SPICE-based simulation for electrothermal analysis of multilevel ULSI interconnects," *IEEE/ACM International Conference on Computer-Aided Design*, 2001, pp. 165-172.

[48] K. Roy, S. Mukhopadhyay and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proc. the IEEE*, 91(2), 2003, pp. 305-327.

[49] T.-Y. Wang and C. C.-P. Chen, "3D Thermal-ADI : a linear-time chip level transient thermal simulator," *IEEE Trans. Computer-Aided Design*, 2002, pp. 1434-1445.

[50] P. Li, L.T. Pileggi, M. Asheghi and R. Chandra, "Efficient full-chip thermal modeling and analysis," *Proc. IEEE/ACM International Conference on Computer-Aided Design*, 2004, pp. 319-326.

[51] Y. Zhan and S. S. Sapatnekar, "Fast computation of the temperature distribution in VLSI chips using the discrete cosine transform and table look-up," *Proc. Asia and South Pacific Design Automation Conference*, 2005, pp. 87-92.

[52] S. Zimin, L. Litian and L. Zhijian, "Self-heating effect in SOI MOSFETs," *Proc. International Conference on Solid State and Integrated Circuit Technology*, 1998, pp. 572-574.

[53] K. Oshima, S. Cristoloveanu, B. Guillaumot, H. Iwai and S. Deleonibus, "Advanced SOI MOSFETs with buried alumina and ground plane: self-heating and short-channel effects, *International Journal of Solid-State Electronics*, 48, 2004, pp. 907-914.

[54] W. Jin, S. K. H. Fung, W. Liu, P. C. H. Chan and C. Hu, "Self-heating characterization for SOI MOSFET based on AC output conductance," *Proc. IEEE International Electron Devices Meeting*, 1999, pp. 175-178.

[55] J. Olsson, "Self-heating effects in SOI bipolar transistors," *Microelectronic Engineering Journal*, 56, 2001, pp. 339-352.

[56] T. Grasser, R. Quay, V. Palankovski and S. Selberherr, "A global self-heating model for device simulation," *Proc. European Solid-State Device Research Conference*, 2000, pp. 324-327.

[57] B. M. Tenbroek, M. S. L. Lee, W. Redman-White, R. J. T. Bunyan and M. J. Uren, "Impact of self-heating and thermal coupling on analog circuits in SOI CMOS," *IEEE Journal of Solid-State Circuits*, 33(7), 1998, pp. 1037-1046.

[58] G. O. Workman, J. G. Fossum, S. Krishnan and M. M. Pelella Jr., "Physical modeling of temperature dependences of SOI CMOS devices and circuits including self-heating," *IEEE Trans. on Electron Devices*, 45(1), 1998, pp. 125-133.

[59] O. Semenov, A. Vassighi and M. Sachdev, "Impact of self-heating effect on long-term reliability and performance degradation in CMOS circuits," *IEEE Trans. on Device and Materials Reliability*, 6(1), 2006, pp. 17-27.

[60] T. Zheng, J. Luo and X. Zhang, "On pure self-heating effect of MOSFET in SOI," *Proc. International Conference on Solid-State and Integrated-Circuit Technology*, 2001, pp. 665-668.

[61] M.-C. Hu and S. L. Jang, "An analytical fully-depleted SOI MOSFET model considering the effects of self-heating and source/drain resistance," *IEEE Trans. on Electron Devices*, 45(4), 1998, pp. 797-801.

[62] L. T. Su, J. E. Chung, D. A. Antoniadis, K. E. Goodson and M. I. Flik, "Measurement and modeling of self-heating in SOI nMOSFET's," *IEEE Trans. on Electron Devices*, 41(1), 1994, pp. 69-75.

[63] H. A. Rueda, "Modeling of mechanical stress in silicon isolation technology and its influence on device characteristics," *University of Florida Ph.D. Thesis*, 1999.

[64] C. S. Smith, "Piezoresistance effect in germanium and silicon," *Physical Review*, 94(1), 1953, pp. 42-49.

[65] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics Magazine*, 38(8), 1965.

[66] J.-P. Han et al., "Novel enhanced stressor with graded embedded SiGe source/drain for high performance CMOS devices," *Proc. IEEE International Electron Devices Meeting*, 2006, pp. 1-4.

[67] Q. Ouyang et al., "Characteristics of high performance PFETs with embedded SiGe source/drain and $< 100 >$ channels on 45^o rotated wafers," *Proc. International Symposium on VLSI Technology*, 2005, pp. 27-28.

[68] Q. Ouyang et al., "Investigation of CMOS devices with embedded SiGe source/drain on hybrid orientation substrates," *Proc. Symposium on VLSI Technology*, 2005, pp. 28-29.

[69] H. S. Yang et al., "Dual stress liner for high performance sub-45nm gate length SOI CMOS manufacturing," *Proc. IEEE International Electron Devices Meeting*, 2004, pp. 1075-1077.

[70] W.-H. Lee et al., "High performance 65 nm SOI technology with enhanced transistor strain and advanced-low-k BEOL," *Proc. IEEE International Electron Devices Meeting*, 2005.

[71] C. Ortolland, "Stress memorization technique (SMT) optimization for 45nm CMOS," *Proc. Symposium on VLSI Technology*, 2006, pp. 78-79.

[72] C.-H. Chen et al., "Stress memorization technique (SMT) by selectively strained-nitride capping for sub-65nm high-performance strained-Si device application," *Proc. Symposium on VLSI Technology*, 2004, pp. 56-57.

[73] M. Yang et al., "Hybrid-orientation technology (HOT): opportunities and challenges," *IEEE Trans. on Electron Devices*, 53(5), 2006, pp. 965-978.

[74] Y. Tateshita et al., "High-performance and low-power CMOS device technologies featuring metal/high-k gate stacks with uniaxial strained silicon channels on (100) and (110) substrates," *Proc. IEEE International Electron Devices Meeting*, 2006, pp. 1-4.

[75] C. Gallon et al., "Electrical analysis of mechanical stress induced by STI in short MOSFETs using externally applied stress," *IEEE Trans. on Electron Devices*, 51(8), 2004, pp. 1254-1261.

[76] A. T. Bradley, R. C. Jaeger, J. C. Suhling and K. J. O'Connor, "Piezoresistive characteristics of short-channel MOSFETs on (100) silicon," *IEEE Trans. on Electron Devices*, 48(9), 2001, pp. 2009-2015.

[77] Y.-M. Sheu et al., "Modeling well edge proximity effect on highly-scaled MOSFETs," *Proc. IEEE Custom Integrated Circuits Conference*, 2005, pp. 831-834.

[78] Y.-M. Sheu et al., "Modeling mechanical stress effect on dopant diffusion in scaled MOSFETs," *IEEE Trans. on Electron Devices*, 52(1), 2005, pp. 30-38.

[79] K.-W. Su et al., "A scaleable model for STI mechanical stress effect on layout dependence of mOS electrical characteristics," *Proc. IEEE Custom Integrated Circuits Conference*, 2003, pp. 245-248.

[80] M. Miyamoto, H. Ohta, Y. Kumagai, Y. Sonobe, K. Ishibashi and Y. Tainaka, "Impact of reducing STI-induced stress on layout dependence of MOSFET characteristics", *IEEE Trans. on Electron Devices*, 51(3), 2004, pp. 440-443.

[81] H. Tsuno et al., "Advanced analysis and modeling of MOSFET characteristic fluctuation caused by layout variation," *Proc. Symposium on VLSI Technology*, 2007, pp. 204-205.

[82] N. Elbel, Z. Gabric, W. Langheinrich and B. Neureither, "A new STI process based on selective oxide deposition," *Proc. Symposium on VLSI Technology Digest of Technical Papers*, 1998, pp. 208-209.

[83] H. S. Lee et al., "An Optimized Densification of the filled oxide for quarter micron shallow trench isolation (STI)," *Proc. Symposium on VLSI Technology Digest of Technical Papers*, 1996, pp. 158-159.

[84] V. Moroz et al., "The impact of layout on stress-enhanced transistor performance," *Proc. International Conference on Simulation of Semiconductor Processes and Devices*, 2005, pp. 143-146.

[85] V. Moroz, L. Smith, X.-W. Lin, D. Pramanik and G. Rollins, "Stress-aware design methodology," *Proc. IEEE International Symposium on Quality Electronic Design*, 2006, pp. 807-812.

[86] L. Smith, "TCAD modeling of strain-engineered MOSFETs," *Material Research Society Symposium Proceedings*, 913, 2006.

[87] H. Cai et al., "Coherent Chip-Scale Modeling for Copper CMP Pattern Dependence," *Material Research Society Symposium Proceedings*, 816, 2004.

[88] "FAMMOS," http://www.synopsys.com/products/tcad/pa/pa_fammos.html.

[89] X. Xu and V. Moroz, "Three dimensional interconnect stress modeling for back end process," *Material Research Society Symposium Proceedings*, 812, 2004.

[90] J.-M. Paik, H. Park and Y.-C. Joo, "Effect of low-k dielectric on stress and stress-induced damage in Cu interconnects," *Microelectronic Engineering*, 71, 2004, pp. 348-357.

[91] A. B. Kahng and R. O. Topaloglu, "Performance-aware CMP fill pattern optimization," Invited Paper, *Proc. International VLSI/ULSI Multilevel Interconnection (VMIC) Conference*, 2007, pp. 135-144.

[92] J. Srinivasan, S.V. Adve, P. Bose and J.A. Rivers, "The impact of technology scaling on lifetime reliability," *Proc. International Conference on Dependable Systems and Networks*, 2004, pp. 177-186.

[93] J. Thurn et al., "Stress hysteresis and mechanical properties of plasma-enhanced chemical vapor deposited dielectric films," *Journal of Applied Physics*, 95(3), 2004, pp. 1988-1992.

[94] D. Degryse, B. Vandevelde, S. Stoukatch and E. Beyne, "Mechanical behavior of BEOL structures containing lowk dielectrics during bonding process," *Proc. Electronics Packaging Technology Conference*, 2003, pp. 815-820.

[95] P. R. Besser and C. J. Zhai, "Modeling and measurement of stress and strain evolution in Cu interconnects," *Proc. International Workshop on Stress-Induced Phenomena in Metallization*, 2004.

[96] S. Kakinuma, M. Kodera, K. Nishikata, J. Aoyama, Y. Saijo and G. Pezzotti, "Local stress assessment in patterned interlayer dielectric films using cathodoluminescence spectroscopy," *Microscopy and Microanalysis*, 2006, pp. 1508-1509.

[97] R. C. J. Wang, L. D. Chen, P. C. Yen, S. R. Lin, C. C. Chiu, K. Wu and K. S. Chang-Liao, "Interfacial stress characterization for stress-induced voiding in Cu/low-k interconnects," *Proc. International Conference on Physical and Failure Analysis*, 2005, pp. 96-99.

[98] G. B. Alers, J. Sukamto, P. Woytowitz, X. Lu, S. Kailasam and J. Reid, "Stress migration and the mechanical properties of copper," *Proc. Reliability Physics Symposium*, 2005, pp. 36-40.

[99] D. N. Bentz, M. O. Bloomfield, H. Huang, J.-Q Lu, R. J. Gutmann and T. S. Cale, "Grain based modeling of stress induced copper migration for 3D-IC interwafer vias," *Proc. International Conference on Simulation of Semiconductor Processes and Devices*, 2006, pp. 345-348.

[100] C. J. Zhai, H. W. Yao, A. P. Marathe, P. R. Besser and R. C. Blish, "Simulation and experiments of stress migration for Cu/low-k BEOL," *IEEE Trans. on Device and Materials Reliability*, 4(3), 2004, pp. 523-529.

[101] L. T. Shi and K. N. Tu, "Finite-element modeling of stress distribution and migration in interconnecting studs of a three-dimensional multilevel device structure," *Applied Physics Letters*, 65(12), 1994, pp. 1516-1518.

[102] Y.-B. Park and I.-S. Jeon, "Effects of mechanical stress at no current stressed area on electromigration reliability of multilevel interconnects," *Microelectronic Engineering*, 71(1), 2004, pp. 76-89.

[103] W. Wu, X. Duan and J. S. Yuan, "Modeling of time-dependent dielectric breakdown in copper metallization," *IEEE Trans. on Device and Materials Reliability*, 3(2), 2003, pp. 26-30.

[104] N. Hwang et al., "Investigation of intrinsic dielectric breakdown mechanism in Cu/low-k interconnect system," *IEEE Electron Device Letters*, 27(4), 2006, pp. 234-236.

[105] K. Y. Yiang, W. J. YooImpact, A. Krishnamoorthy and L. J. Tang, "Impact of buried capping layer on TDDB physics of advanced interconnects," *Proc. International Reliability Physics Symposium*, 2005, pp. 490-494.

[106] W. Wu, X. Duan and J. S. Yuun, "A physical model of time-dependent dielectric breakdown in copper metallization," *Proc. Annual International Reliability Physics Symposium*, 2003, pp. 282-286.

[107] H. Miura, S. Ikeda and N. Suzuki, "Effect of mechanical stress on reliability of gate-oxide film in MOS transistors," *Proc. International Electron Devices Meeting*, 1996, pp. 743-746.

[108] X. H. Liu, M. W. Lane, T. M. Shaw and E. Simonyi, "Delamination in patterned films," *International Journal of Solids and Structures*, 44(6), 2007, pp. 1706-1718.

[109] K. C. Aw, N. T. Salim, W. Gao and K. Prince, "Study of copper diffusion in low-k thin film using sims," *International Journal of Modern Physics B*, (20)25-27, 2006, pp. 4165-4170.

[110] K.-I. Takeda, D. Ryuzaki, T. Mine, K. Hinode and R. Yoneyama, "Copper-induced dielectric breakdown in silicon oxide deposited by plasma-enhanced chemical vapor deposition using trimethoxysilane," *Journal of Applied Physics*, 94(4), 2003, pp. 2572-2578.

[111] N. Cherault, J. Besson, C. Goldberg, N. Casanova and M.-H. Berger, "Finite element simulation of thermomechanical stress evolution in Cu/low-k interconnects during manufacturing and subsequent thermal cycling," *Proc. European Solid-State Device Research Conference*, 2005, pp. 493-496.

[112] B. K. Lim et al., "Bias-temperature stress analysis of Cu/ultrathin Ta/SiO2/Si interconnect structure," *Journal of Vacuum Science Technology B*, 22(5), 2005, pp. 2286-2290.

[113] S.-H. Rhee, C. E. Murrayand P. R. Besser, "Effects of BEOL stack on thermal mechanical stress of Cu lines," *Material Research Societcy Symposium Proc.*, 914, 2006.

[114] S. M. Alam, "Design tool and methodologies for interconnect reliability analysis in integrated circuits," *Massachusetts Institute of Technology Ph.D. Thesis*, 2004.

[115] "Synopsys Raphael," http://www.synopsys.com/products/tcad/raphael_ds.html.

[116] R. Arghavani et al., "Stress management in sub-90nm transistor architecture," *IEEE Trans. on Electron Devices*, 51(10), 2004, pp. 1740-1744.

[117] S. W. Chung et al., "Novel shallow trench isolation process using flowable oxide CVD for sub-100nm DRAM," *Proc. IEEE International Electron Devices Meeting*, 2002, pp. 233-236.

[118] "Cadence SoC Encounter," http://www.cadence.com/products/digital_ic/soc_encounter/index.aspx.

[119] "Synopsys Star RCXT," http://www.synopsys.com/products/starxct/star_rcxt_ds.html.

[120] "Blaze IF," http://www.blaze-dfm.com/products/products2.html.

[121] "Mentor Calibre," http://www.mentor.com/products/.

[122] "Synopsys PrimeTime," http://www.synopsys.com/products/analysis/primetime_ds.html.

[123] "OpenAccess API," http://openeda.si2.org

[124] "Synopsys Design Compiler," http://www.synopsys.com/products/logic/design_compiler.html.

[125] "Synopsys HSPICE," http://www.synopsys.com/products/mixedsignal/hspice/hspice.html.

[126] H. Haznedar et al., "Impact of stress-induced backflow on full-chip electromigration risk assessment," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 25(6), 2006, pp. 1038-1046.

[127] T. Yamamoto et al., "High performance 60 nm gate length germanium p-MOSFETs with Ni germanide metal source/drain," *Proc. IEEE Electron Devices Meeting*, 2007, pp. 1041-1043.

[128] D. Kuzum et al., "Interface-engineered Ge (100) and (111), N- and P-FETs with high mobility," *Proc. IEEE Electron Devices Meeting*, 2007, pp. 723-726.

[129] K. D. Cantley et al., "Performance analysis of III-V materials in a double-gate nano-MOSFET," *Electron Devices Meeting*, 2007, pp. 113-116.

[130] M. Passlack et al., "High mobility III-V MOSFETs for RF and digital applications," *Proc. IEEE Electron Devices Meeting*, 2007. pp. 621-624.

[131] K. Shin, C.O. Chui and T.-J. King, "Dual stress capping layer enhancement study for hybrid orientation finFET CMOS technology," *Proc. IEEE Electron Devices Meeting*, 2005, pp. 988-991.

[132] O. Weber et al., Examination of additive mobility enhancements for uniaxial stress combined with biaxially strained Si, biaxially strained SiGe and Ge channel MOSFETs *Proc. IEEE Electron Devices Meeting*, 2007, pp. 719-722.

[133] T. Watanebe et al., "Robust BEOL process integration with ultra low-k (k=2.0) dielectric and self-formed MnOx barrier technology for 32 nm-node and beyond, *Proc. International Interconnect Technology Conference*, 2008, to appear in Session 10.7.

[134] H. W. Chen et al., "A self-aligned air gap interconnect process," *Proc. International Interconnect Technology Conference*, 2008, to appear in Session 3.4.

[135] S. Nitta, "A multi-level Cu/low-K/airgap BEOL technology," *Proc. International Interconnect Technology Conference*, 2008, to appear in Session 10.1.

[136] N. Nakamura, N. Matsunaga, T. Kaminatsui, K. Watanabe and H. Shibata, "Cost-effective air-gap interconnects by all-in-one post-removing process," *Proc. International Interconnect Technology Conference*, 2008, to appear in Session 10.2.

[137] R. Gras et al., "300 mm multi level air gap integration for edge interconnect technologies and specific high performance applications," *Proc. International Interconnect Technology Conference*, 2008, to appear in Session 10.3.

[138] H. Matsuyama et al., “New degradation phenomena of stress-induced voiding inside via in copper interconnects,” *Proc. IEEE International Reliability physics symposium*, 2007. pp. 638-639.

[139] Y. K. Lim et al., “Design for manufacturability and its role in enhancing stress migration reliability of porous ultra low-k copper interconnects,” *Proc. IEEE International Reliability Physics Symposium*, 2007, pp. 134-140.

[140] T. L. Tan et al., “Test structure design for precise understanding of Cu/low-k dielectric reliability,” *Proc. IEEE International Reliability Physics Symposium*, 2007, pp. 632-633.

[141] Z. Tokei, “Low-k dielectric reliability: impact of test structure choice, copper and integrated dielectric quality,” *Proc. International Interconnect Technology Conference*, 2008, to appear in Session 6.1.

[142] E. Pop, R. Dutton and K. Goodson, “Thermal analysis of ultra-thin body device scaling [SOI and FinFet devices],” *Proc. IEEE Electron Devices Meeting*, 2003, pp. 36.6.1-36.6.4.

[143] S. Kolluri, K. Endo, E. Suzuki and K. Banerjee, “Modeling and analysis of self-heating in FinFET devices for improved circuit and EOS/ESD performance,” *Proc. IEEE Electron Devices Meeting*, 2007, pp. 177-180.

[144] X. Zhang et al., “Impact of process induced stresses and chip-packaging interaction on reliability of air-gap interconnects,” *Proc. International Interconnect Technology Conference*, 2008, to appear in Session 7.5.

[145] C. J. Uchibori, X. Zhang, P. S. Ho and T. Nakamura, “Investigation of interconnect design on chip package interaction and mechanical reliability of Cu/low-k multi-layer interconnects in flip chip package,” *Proc. International Interconnect Technology Conference*, 2008, to appear in Session 7.10.

[146] P. Ramm, “Through silicon via technologies for extreme miniaturized 3D integrated wireless sensor systems,” *Proc. International Interconnect Technology Conference*, 2008, to appear in Session 2.1.