A 14nm Finfet Transistor-Level 3D Partitioning Design to Enable High-Performance and Low-Cost Monolithic 3D IC

Jiajun Shi[1,3], Deepak Nayak[1], Srinivasa Banna[1], Robert Fox[2], Srikanth Samavedam[2], Sandeep Samal[4] and Sung Kyu Lim[4]

[1]Technology Research, GLOBALFOUNDRIES, Santa Clara, CA, USA
[2]Technology Development, GLOBALFOUNDRIES, Malta, NY, USA
[3]Department of ECE, University of Massachusetts, Amherst, MA, USA
[4]School of ECE, Georgia Institute of Technology, Atlanta, GA, USA
jiajun@umass.edu, deepak.nayak@globalfoundries.com

Abstract

Monolithic 3D IC (M3D) shows degradation in performance compared to 2D IC due to the restricted thermal budget during fabrication of sequential device layers. A transistor-level (TR-L) partitioning design is used in M3D to mitigate this degradation. Silicon validated 14nm FinFET data and models are used in a device-to-system evaluation to compare the TR-L partitioned M3D's (TR-L M3D) performance against the conventional gate-level (G-L) partitioned M3D's performance as well as standard 2D IC. Extensive cell-level and system-level evaluation, including various device and interconnect process options, shows that the TR-L M3D provides up to 20% improved performance while still maintaining around 30% power saving compared to standard 2D IC. Additionally, the TR-L partitioning design enables M3D with a simplified process flow that leads to 23% lower cost compared to that of G-L partitioning scheme.

Introduction

With increased challenges in scaling CMOS technology below 22nm node, a lot of research work is focusing on monolithic 3D IC (M3D) which shows power savings over 2D IC. M3D is enabled by sequential integration of multiple device layers [1]. However, this sequential integration has its own challenges. In order to preserve the circuits in the bottom-tier (bot-tier), the top-tier is processed with a Low-Temperature (LT) process below 650°C [1]. In another option, the Bot-tier uses Tungsten (WB) for wiring [2] while top-tier is processed with normal thermal budget. In the conventional M3D [4] that uses gate-level (G-L) partitioning scheme (G-L M3D), LT and WB process options degrade transistors in top-tier and cell-to-cell interconnects in bot-tier, respectively, leading to timing issue [2-3].

In this work, we explore a transistor-level (TR-L) partitioning based M3D (TR-L M3D) in a 14nm technology node and investigate its advantages over the G-L M3D. The TR-L M3D achieves interconnect saving by using 3D standard cells. The 3D cells are designed by splitting PMOS and NMOS transistors into two tiers within a standard cell, and the monolithic inter-layer vias (MIVs) are used for intra-cell vertical interconnection between pull-down network (PDN) and pull-up network (PUN) (Fig. 1c). PMOS (in PUN) is placed in the bot-tier and NMOS (in PDN) is placed in the top-tier with one layer of inter-layer dielectric (ILD) for isolation (Fig. 1a). This way, the intra-cell capacitance is significantly reduced due to the elimination of coupling between PUN and PDN, which improves cell performance and thus nullifies the negative impact of low performance transistors in LT process option. Also, in TR-L M3D, the cell-to-cell interconnects use the full metal stack (Copper) in top-tier (Fig. 1a-b), and only two metal layers are used in bot-tier for intra-cell wiring. Therefore, the cell-to-cell interconnect is not affected by the tungsten wiring in the bot-tier if WB process option is used. In addition, the use of only two metal layers in bot-tier and single of type of transistor in each tier contributes to lower cost than G-L M3D.

In the paper, we investigate the optimum dimensions of MIV, taking into account 3D cell footprint restrictions, ease of manufacturability, and electrostatic coupling between top- and bot-tier. The 3D cells are designed following our 14nm Finfet design rules and the MIV dimensions. The 3D cell RC is extracted precisely by using CalibrexACT and Sentaurus Interconnect. We performed cell performance evaluation in various LT cases and quantified intra-cell capacitance reduction in 3D cells. We then extensively benchmarked system-level circuits to investigate both WB and LT process's impacts on TR-L M3D's system timing and the timing-associated impact on power. Modeling of cost per die, considering practical die size and critical mask layers, is carried out to demonstrate the cost saving in TR-L M3D compared to that of G-L M3D.

TR-L M3D Cell Design and RC Extraction

The design of each TR-L M3D cell contains three parts: the PUN in bot-tier, PDN in top-tier, and MIVs that connect input/output ports between PUN and PDN (Fig. 1c). Each component in PUN and PDN is designed using a standard 2D layout design flow with 14nm FinFET rules. Fig. 2b shows the top-view of our proposed 3D cell. The PDN is placed in the top-tier exactly aligning with PUN in bot-tier. For both PUN and PDN design, the power rail uses 1 metal track and the active device region uses 3 metals tracks (=4 fin pitch, Fig. 2a-b). The MIVs are placed in the 5th track with the spacing equal to minimum M1 distance from the active device region. This is set to avoid M1 routing violation inside the cell. The total cell height of the 3D cell takes 5 metal tracks (vs. 9 metal tracks in 2D cell). The width of MIV is determined primarily by its impact on 3D cell footprint saving. Fig. 3a shows the ratio of 3D and 2D footprint (3D/2D) as MIV width varies. Each curve has an inflection point where the 3D cell would lose severely its footprint saving if MIV width exceeds 50nm. We also want the MIV width to be as large as possible, which can reduce MIV aspect ratio for easy manufacturability. Therefore, we set the width to be 50nm in 3D cells, resulting in a 45% footprint saving against 2D cells (Fig. 3a) which significantly contributes to interconnect reduction in systems. The MIV height is equal to the sum of device dielectric thickness and ILD (Fig. 2c). The ILD thickness is determined by considering the electrostatic coupling between the top-most metal in the bot-tier and NMOS in the top-tier which is simulated through Sentaurus TCAD (STCAD) simulation. When 0.8V is applied to the metal line in the bot-tier, it changes the Electron Quantum-potential [5] and Electrostatic Potential of the NMOS channel in top-tier due to electrostatic coupling [5] (Fig. 3c-d). This leads up-to 150mV Vth variation in the top-tier NMOS (Fig. 3b). We then determine the ILD thickness to be 110nm which controls the Vth variation to be within 5% (Fig. 3b). The MIV height is 310nm with an aspect ratio of 6 which is acceptable for manufacturability.

The intra-cell RC is composed of four parts: RC inside

978-1-5090-3902-9/16/$31.00 ©2016 IEEE

PUN, RC inside PDN, RC of MIVs, and coupling capacitance between PUN and PDN (Fig. 4a). The layouts of PUN and PDN are independently prepared and their RC is extracted using normal extraction flow (assuming no coupling between PUN and PDN). The vertical coupling between PUN and PDN and the RC of MIVs are extracted from the actual 3D cell structure (Fig. 4b-c) built in Sentaurus Interconnect (Sinterconnect). The key parameters in the 3D structure such as dielectric constant, doping concentration and gate metal work function are defined based on the foundry data. The dimensions of each component are set according to the technology and the 3D cell design. Table I and II show the extracted capacitance values. The vertical coupling capacitance between PUN and PDN is negligible, which contributes to the intra-cell capacitance reduction (see next section). Each MIV also has very small coupling capacitance to the adjacent MIVs (~30-50aF) and a resistance of 5.5Ω.

Cell-level Evaluation

In M3D with LT process, the transistors in top-tier are processed with limited thermal budget (<650^{o}C) which specifically increases the sheet resistance of S/D (Fig. 5a) due to activation with low-temperature [1][6]. The WB process option has negligible impact on cell-level performance. In this work, we evaluate the 3D cell with top-tier NMOS in various LT options: 400^{o}C (LT400), 500^{o}C (LT500), 600^{o}C (LT600) and compare with the cells in G-L M3D top-tier using the same LT options (cells in bot-tier use regular process). Based on the experimentally demonstrated S/D resistance increase factor in [6], we build LT device models (Fig. 5b) and use them in cell- and system-level evaluation.

Fig. 5c shows the evaluation results of INVx1. The INVx1 of TR-L M3D shows 4% (LT600) to 15% (LT400) degradation compared with 2D baseline while the INVx1 in GL-M3D (top-tier) shows 17% (LT600) to 32% (LT400) degradation. Moreover, for the NOR2x1, the TR-L 3D cell shows 10% improved performance over 2D cell (Fig. 5d). This is because the critical path of NOR2x1 cell (PUN) is in the bot-tier that uses normal devices. But the NOR2x1 in G-L M3D still shows severe degradation compared with 2D, due to the degraded devices in the top-tier. The significantly reduced degradation in INVx1 and improved performance in NOR2x1 are both contributed by the intra-cell capacitance reduction (Fig. 5f). Fig. 6a shows the effective intra-cell capacitance of various 3D cells vs. 2D cells, which are measured using HSPICE simulations. Our TR-L 3D cells show up to 23% total intra-cell capacitance reduction where the reduction of cell device capacitance (due to elimination of coupling between NMOS and PMOS) makes the major contribution (Fig. 6b). Additionally, as driving strength increases, the TR-L M3D cell shows decreased degradation against 2D cell (Fig. 5e) since larger cell has more capacitance reduction (Fig. 6a).

System-level Evaluation

In system-level evaluation, we focus on LT and WB process's impact on system timing and power. The LT cases include LT400, LT500, and LT600. The WB cases include R_{bot}/R_{copper}=2 (WBx2), R_{bot}/R_{copper}=3 (WBx3), R_{bot}/R_{copper}=4 (WBx4). The actual bulk resistivity of Tungsten is 3.3x of Copper. We developed a system-level evaluation flow for TR-L M3D (Fig. 7a). In this flow, multiple .lib files of 3D cells are generated to match the LT and WB cases. Since the cell-to-cell interconnects use the full Copper metal stack in top-tier and follows conventional routing style, the layout of TR-L M3D (Fig. 7b) benchmark circuit can be generated and analyzed using current commercial EDA tools (Fig. 7a). G-L M3D is benchmarked with our 14nm FinFET technology following the methodology in [2-3] (Fig. 7b). 2D circuits are benchmarked as baseline using the same technology.

Fig. 9a-d show the system performance comparison between TR-L M3D and G-L M3D in Ideal, LT and WB process cases. The LT and WB cases of G-L M3D show severely degraded performance and timing violations (compared to the 2D IC, the dash lines in Fig. 9a-d). The degradation in turn leads to associated penalty on circuit footprint and power. Since this degradation leads to the usage of larger cells and more buffers to fix the timing, the G-L M3D circuit footprint increases as the top-tier process temperature goes down (Fig. 8a) or the bot-tier resistivity increases (Fig. 8b). Since the TR-L M3D has reduced intra-cell capacitance and avoids bot-tier for inter-cell routing, it has minimal impact from each LT or WB process options. So the circuit footprint of TR-L M3D increases at much slower rate than G-L M3D in the LT options (Fig. 8a) and remains constant in the WB options (Fig 8b). Moreover, the use of larger cells leads to increased system energy and power. For LT options, the G-L M3D has 1.9x faster increase in energy (Fig. 9a) with lowering of temperature in the interconnect-dominated low-density parity check (LDPC) core and 2.5x faster increase in energy (Fig. 9b) in the gate-dominated advanced encryption standard (AES) core compared to TR-L. For WB options (Fig. 9c-d), the G-L M3D has much faster increase in energy as the bot-tier resistivity increases, but the TR-L M3D energy remains flat as bot-tier resistivity goes up. Overall, TR-L M3D shows up to 20% performance benefit while maintaining up to 32% power saving against 2D (Fig. 9e-f). G-L M3D shows degraded performance and up to 7% power saving over 2D (Fig. 9e-f).

Modeling and Comparison of Cost

The TR-L M3D has only two metal layers in bot-tier while the G-L M3D uses at least 4 metal layers in bot-tier [4]. Additionally, the TR-L M3D uses single type of transistor in each tier. The number of critical mask layers in TR-L M3D's process is thus significantly reduced. This leads to reduced cost per wafer and higher yield which contribute to reduced cost per die (see calculation in Fig. 10a). Under our metal use assumption (top-tier 10 metal layers in both options), the TR-L M3D shows 23% lower cost compared to G-L (Fig. 10b).

Conclusion

The TR-L M3D is evaluated from device-to-system using silicon validated 14nm FinFET technology data and models. The use of TR-L design enables low-cost M3D ICs and mitigates the performance degradation issue in G-L M3D while achieving significant power saving against 2D IC.

Reference

[1] P. Batude, et al, "3-D Sequential Integration: A Key Enabling Technology for Heterogeneous CoIntegration of New Function With CMOS," JESTCS, 2012.

[2] S. Samal, et al.,"Tier Partitioning Stragety to Mitigate BEOL Degradation and Cost Issues in Monolithic 3D ICs," ICCAD, 2016, *in press*

[3] S. Samal, et al., "How to Cope with Slow Transistors in the Top-tier of Monolithic 3D ICs: Design Studies and CAD Solutions," ISLPED, 2016, *pp. 320-325*

[4] O. Billoint, et al., "A Comprehensive Study of Monolithic 3D Cell on Cell Design Using Commercial 2D Tool," DATE, 2015, pp. 1192–1196.

[5] R. Chen, et.al, "A quantum corrected energy-transport model for nanoscale semiconductor devices" JCP, 2005, pp. 131-156

[6] P. Batude, et al., " Low Temperature FDSOI Devices, a Key Enabling Technology for 3D Sequantial Integration," VLSI-TSA, 2013, pp.1 - 4

[7] T. Okabe, et.al, "Analysis on yield of integrated circuits and a new expression for the yield" Elec. Eng. in Japan, 1992, pp. 135-14

Introduction

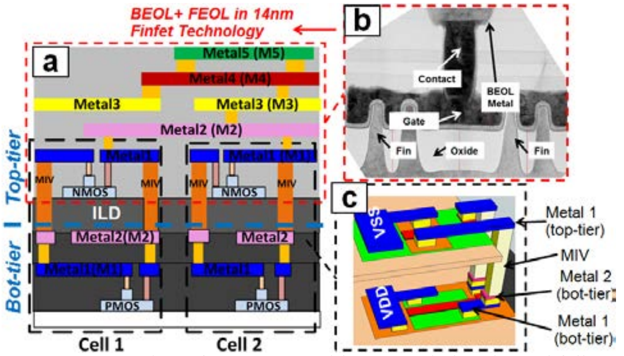


Figure. 1 (a) Overview of TR-L M3D with M1-M2 in bot-tier and full metal stack (BEOL) in top-tier. **(b)** The BEOL and FEOL design rules and RC parameters of standard 14nm Finfet technology are used in the top-tier's interconnect. **(c)** Schematic of 3D INV cell in TR-L M3D.

3D Cell Design and RC Extraction

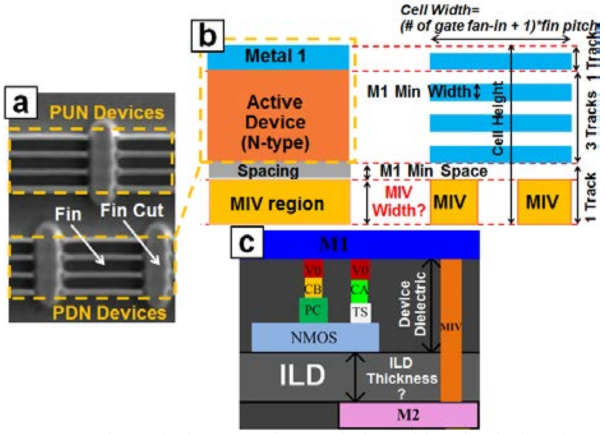


Figure. 2 (a) The typical PDN and PUN design with 4-fin devices in 14nm Finfet technology. **(b)** Top view of feature size design of TR-L M3D cell: cell height is 5 M1-tracks; cell width= (# of gate fan-in+1)* poly pitch. **(c)** Side view of the inter-layer between top- and bot-tier: MIV connects top-most metal (M2) of bot-tier and top-tier M1.

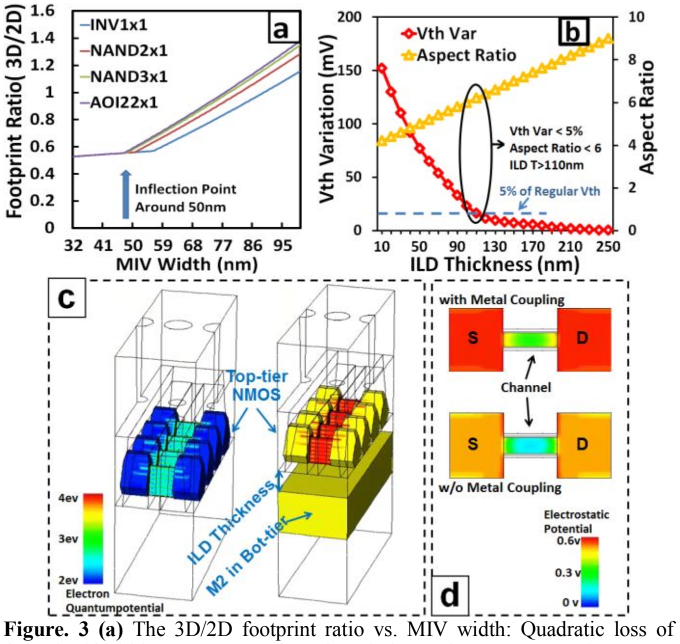


Figure. 3 (a) The 3D/2D footprint ratio vs. MIV width: Quadratic loss of footprint if width >50nm. **(b)** The Vth variation and MIV aspect ratio vs. ILD thickness: assume width=50nm. **(c)** Electron quantum potential simulation in STCAD: using 14nm FinFET data; **evaluates electrostatic filed and coupling**; metal is charged from 0V to 0.8V; Simulated at ILD T=50nm. **(d)** potential in NMOS channel with added filed from metal.

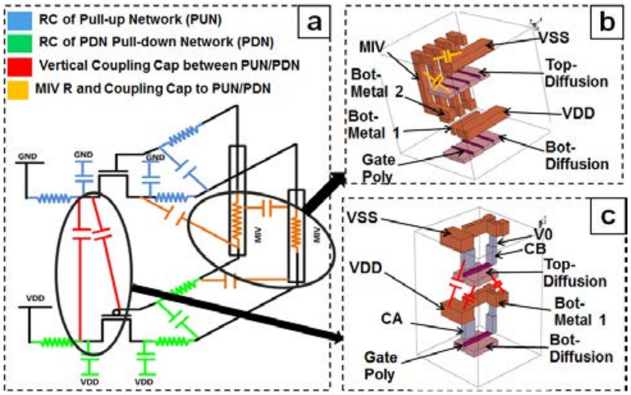


Figure. 4 (a) The composition of RCs in the 3D cell. **(b)** 3D NAND3 Cell structure built in Sinterconnect for MIV RC extraction. **(c)** PUN and PDN of 3D INV built in Sinterconnect for extracting vertical coupling capacitance.

Table I: Extracted Parasitic Capasitance of MIV

	Top-tier Diffusion	VSS	Bot-tier Diffusion	VDD	MIV
MIV	18aF	5.5aF	1aF	2.1aF	14aF

Table II: Extracted Coupling Capasitance between PUN and PDN

	Top-tier Drain	Top-tier Source	Top-tier Gate	Top-tier VSS	Top-tier M1
Bot-tier VDD	2aF	2aF	0.7aF	NA	0.009aF
Bot-tier M1	3aF	3aF	1aF	0.02aF	0.01aF

Cell-level Evaluation

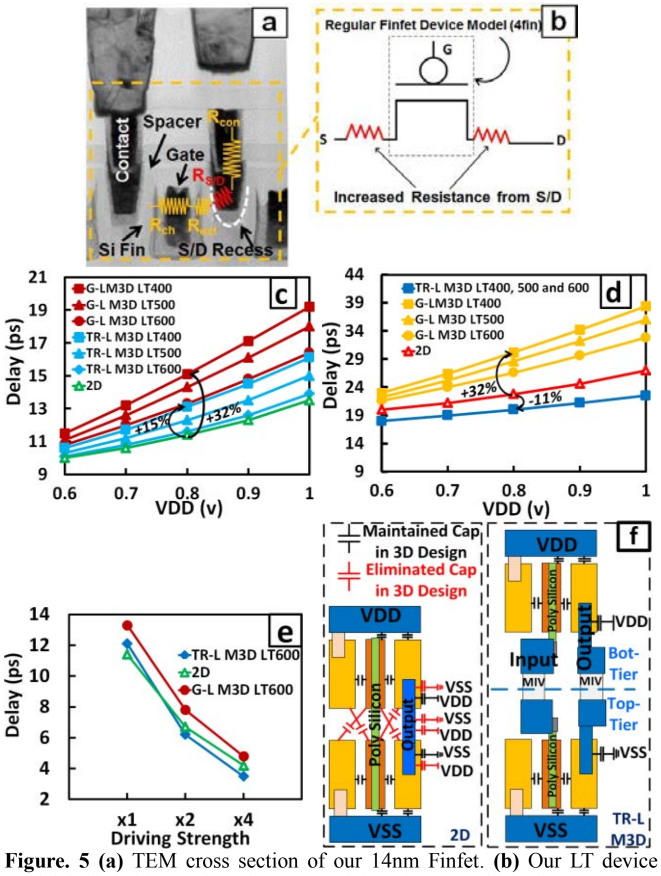


Figure. 5 (a) TEM cross section of our 14nm Finfet. **(b)** Our LT device model built with regular FinFET model and extra S/D resistance. **(c-d)** Delay of INVx1 and NOR2x1 in multiple LT options: TR-L vs. G-L; use 2D as baseline. **(e)** Delay of INVx1, x2 and x4 at LT600**: LT600 is the option close to the practical process option[1]**. **(f)** Capacitance reduction in TR-L 3D cell vs. 2D cell: Vertical splinting of PUN and PDN leads to both intra-cell interconnect and device coupling capacitance reduction.

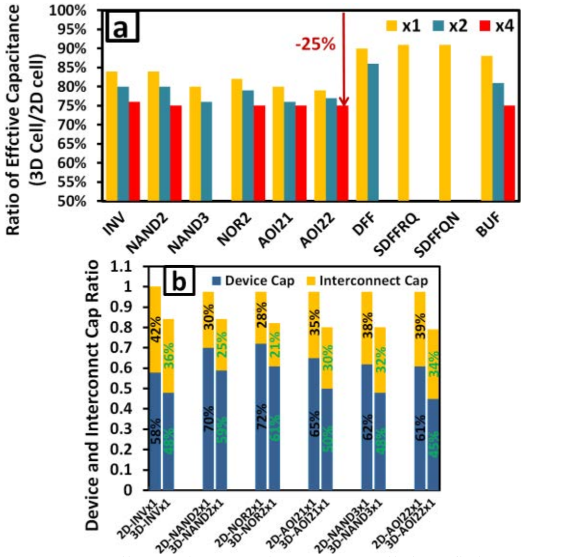


Figure. 6 (a) 3D cell capacitance ratio over 2D cell: benefit increases as the driving strength goes up; 3D cell achieves 10% to 25% capacitance saving against 2D cell. **(b)** The ratio of intra-cell device and interconnect capacitance over total intra-cell capacitance of 2D cell: 3D design shows up to 15% device cap reduction and 7% interconnection reduction

System-level Evaluation

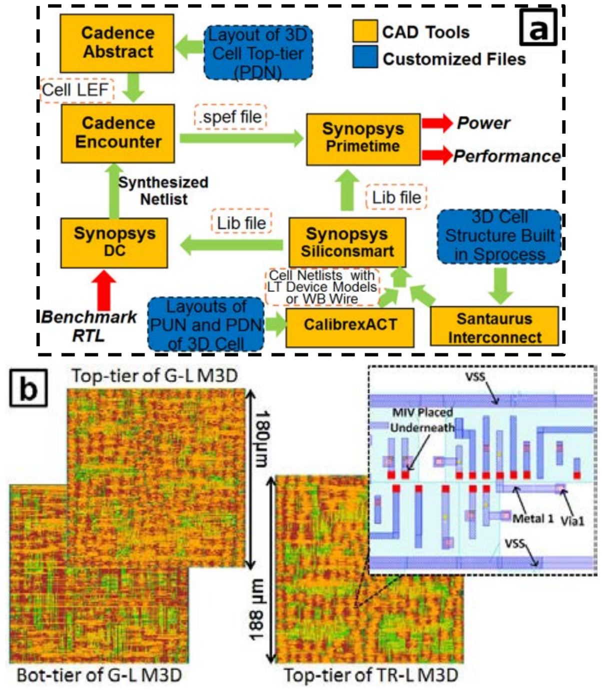


Figure. 7 (a) Our device-to-system flow for TR-L M3D evaluation. **(b)** The routed layouts of AES cores in G-L and TR-L M3D including inter-cell routing, power rails, inserted MIVs and clock tree (power delivery network design is not included in both TR-L M3D and G-L M3D).

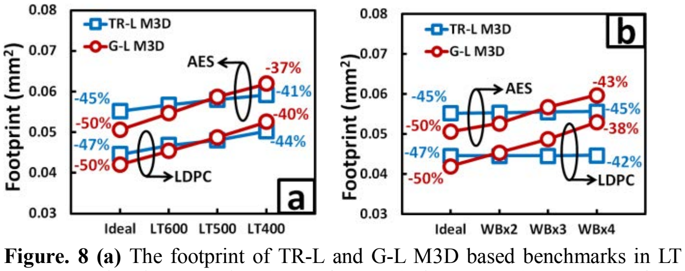


Figure. 8 (a) The footprint of TR-L and G-L M3D based benchmarks in LT cases: G-L has better saving (-50% from 2D) than TR-L (-45 to -47% from 2D) in ideal process case; **G-L's footprint increases fast due to usage of large cells to fix timing**. **(b)** The footprint of TR-L vs. G-L MD in WB cases.

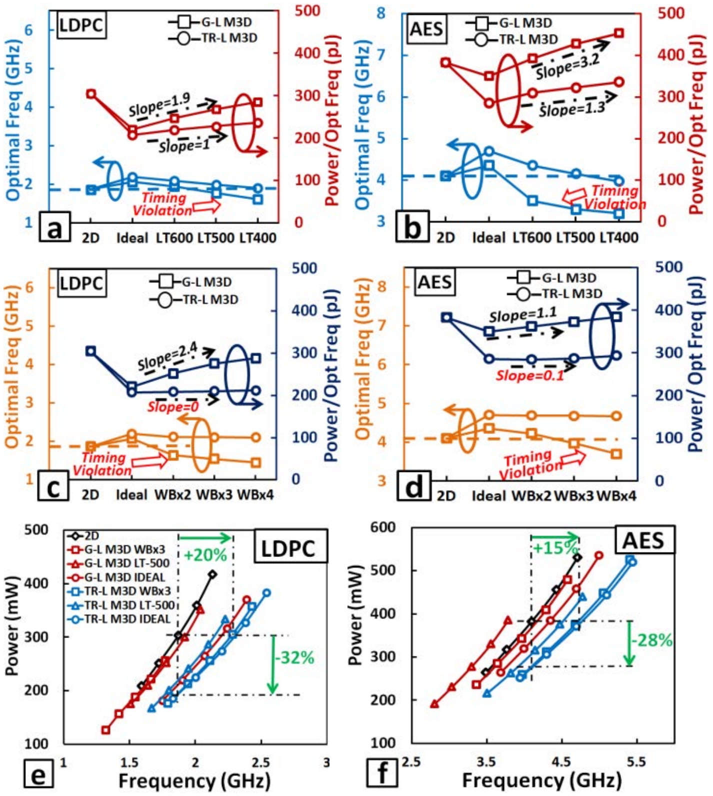
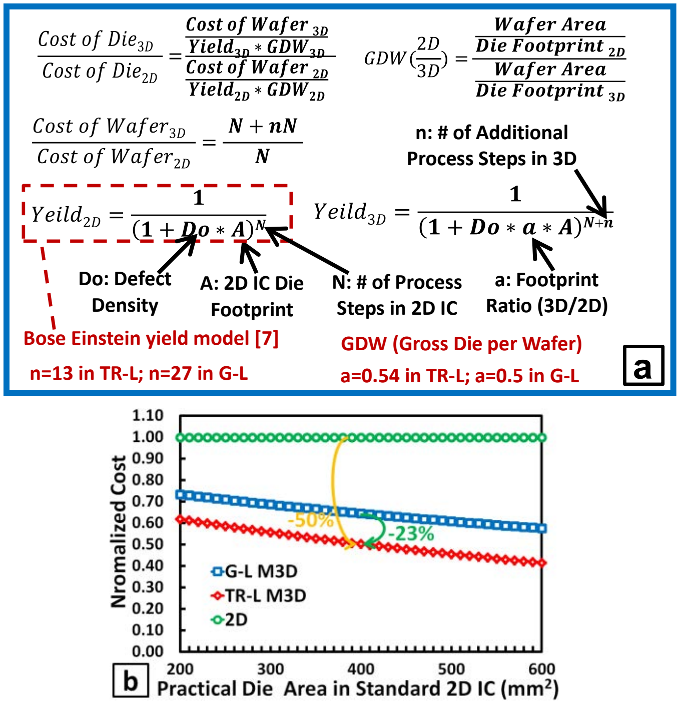


Figure. 9 (a) LDPC in LT case: G-L shows 1.9x faster increase in energy against TR-L and timing violation in LT500, 400. **(b)** AES in LT case: G-L shows 2.5x faster increase in energy against TR-L and timing violation in LT400. **(c)** LDPC in WB case: G-L shows much faster increase in energy against TR-L and timing violation in WBx2, x3 and x4. **(d)** AES in WB case: G-L shows 11x faster increase in energy against TR-L and timing violation in WBx3 and x4. **(e)** Power vs. Performance in LDPC **(±10% VDD)**: **WBx3 and LT600 are close to practical process option**; G-L has degraded performance and about 7% power saving against 2D. **(f)** Power vs. Performance in AES **(±10% VDD)**: G-L has degraded power and performance against 2D.

Cost Evaluation



$$\frac{Cost\ of\ Die_{3D}}{Cost\ of\ Die_{2D}}=\frac{\frac{\boldsymbol{Cost\ of\ Wafer}_{3D}}{\boldsymbol{Yield}_{3D}*\boldsymbol{GDW}_{3D}}}{\frac{\boldsymbol{Cost\ of\ Wafer}_{2D}}{\boldsymbol{Yield}_{2D}*\boldsymbol{GDW}_{2D}}}\qquad GDW\left(\frac{2D}{3D}\right)=\frac{\frac{\boldsymbol{Wafer\ Area}}{\boldsymbol{Die\ Footprint}_{2D}}}{\frac{\boldsymbol{Wafer\ Area}}{\boldsymbol{Die\ Footprint}_{3D}}}$$

$$\frac{Cost\ of\ Wafer_{3D}}{Cost\ of\ Wafer_{2D}}=\frac{\boldsymbol{N+nN}}{\boldsymbol{N}}$$

$$Yeild_{2D}=\frac{\boldsymbol{1}}{(\boldsymbol{1}+\boldsymbol{Do}*\boldsymbol{A})^{\boldsymbol{N}}}\qquad Yeild_{3D}=\frac{\boldsymbol{1}}{(\boldsymbol{1}+\boldsymbol{Do}*\boldsymbol{a}*\boldsymbol{A})^{\boldsymbol{N+n}}}$$

n: # of Additional Process Steps in 3D
Do: Defect Density
A: 2D IC Die Footprint
N: # of Process Steps in 2D IC
a: Footprint Ratio (3D/2D)
Bose Einstein yield model [7]
n=13 in TR-L; n=27 in G-L
GDW (Gross Die per Wafer)
a=0.54 in TR-L; a=0.5 in G-L
a
Nromalized Cost
1.10
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-50%
-23%
G-L M3D
TR-L M3D
2D
200
300
400
500
600
b
Practical Die Area in Standard 2D IC (mm^2)

Figure. 10 (a) Equations used for calculating yield and cost. **(b)** TR-L shows 23% and 50% lower cost compared to G-L and 2D respectively (at the 400 mm^2 die size).