Physical Design Solutions to Tackle FEOL/BEOL Degradation in Gate-level Monolithic 3D ICs

Bon Woong Ku[†], Peter Debacker[§], Dragomir Milojevic[§], Praveen Raghavan[§],
Diederik Verkest[§], Aaron Thean[§], and Sung Kyu Lim[†]
[†]School of ECE, Georgia Institute of Technology, Atlanta, GA
[§]IMEC, Leuven, Belgium
bwku@gatech.edu, sklim@ece.gatech.edu

ABSTRACT

In this paper, we develop physical design tools and methodologies to tackle the inter-tier performance variations caused by low temperature manufacturing in 2-tier gate-level monolithic 3D ICs (M3D). First, we model the top tier front-end-of-line (FEOL) device mobility degradation and its impact on cell delay/power values. Next, we quantify the impact of tungsten interconnect and cost-driven metal layer saving in the back-end-of-line (BEOL) of the bottom tier. These device and interconnect degradation models are used in our new full-chip M3D physical design flow named Derated 2D. This flow overcomes the well-known drawback of the state-of-the-art Shrunk 2D that requires shrinking of layout objects and RC parasitics. Also, Derated 2D performs low-temperature process-aware tier partitioning to effectively keep timing-critical components in the bottom tier. Moreover, Derated 2D conducts timing-driven monolithic inter-tier via (MIV) planning to cope with the resistivity increase in tungsten BEOL. Lastly, Derated 2D offers an effective timing closure solution through a post-route optimization. Experiments based on a foundry-grade 7nm FinFET process design kit (PDK) show that Derated 2D achieves up to 36% performance improvement and 10% energy saving compared with Shrunk 2D. Using a post-route optimization, Derated 2D further improves timing under various FEOL/BEOL degradation settings at a minimum energy overhead.

CCS Concepts

•Hardware → 3D integrated circuits; Physical design (EDA); Methodologies for EDA; Process variations;

Keywords

Gate-level Monolithic 3D IC; Low Temperature Manufacturing Process; FEOL/BEOL Degradation;

1. INTRODUCTION

One of the most critical problems with monolithic 3D integration is the limited thermal budget for the top tier fabrication process. Once the bottom tier devices are implemented with the normal process, they suffer from additional thermal exposure during dopant activation step of the top tier ($T > 1000\,°C$). In the meantime, integrating Copper (Cu) interconnects in the bottom tier implies that thermal budget for the top tier has to be under 450 °C, because Cu diffuses away into the Low-K regions at such a high temperature. In order to preserve the device performance and the integrity of the back-end-of-line (BEOL) of the bottom tier, recent studies [1, 3] introduce molecular bonding and solid phase epitaxial regrowth (SPER) dopant activation process ($T \sim 450\,°C$) for the top tier device manufacturing based on planar FDSOI device. In the industrial environment, however, implementing FinFET or nano-wire devices requires conformal in-situ doping in S/D region due to the 3D structure of the device, leading to high temperature annealing processes ($T > 1100\,°C$). Therefore, the limited thermal budget for the top tier manufacturing is expected to bring serious performance loss on the device. Tungsten (W) interconnects in the bottom tier is an alternative to offer more thermal budget for the top tier manufacturing, but the high resistivity of W degrades the overall performance.

An earlier work [6] addresses inter-tier performance variations in monolithic 3D ICs (M3D). However, this work is based on block-level M3D, where the design seriously under-utilizes MIVs and is thus not practical. The authors of [5] present a design methodology for 2-tier gate-level M3D, so-called Shrunk 2D flow. The drawback of this flow is that it requires the same RC parasitic despite shrinking geometry of layout objects. In the advanced node, parasitics are changing non-linear along with metal geometry. Therefore, it is possible to exaggerate the parasitic value, leading to low quality M3D designs. Recently, M3D benefits have been studied for a predictive 7nm FinFET technology [2], but this work does not consider inter-tier variation caused by limited thermal budget.

In this paper we propose new physical design solutions for gate-level M3D that tackle the inter-tier performance variations caused by low temperature manufacturing. The key contributions of this paper are as follows: (1) Using a 7nm bulk FinFET from a foundry-grade process design kit (PDK), we model the top tier device mobility degradation caused by the low thermal budget process, and show the impact on 2-tier gate-level full-chip M3D designs with Shrunk 2D flow. (2) We quantify the impact of both tungsten BEOL and cost-driven metal layer saving in the bottom tier on M3D design performance. (3) Using these transistor corners and interconnect models, we propose Derated 2D flow, where we do not alter the geometry in technology files, but only derate the RC parasitic corner. (4) We develop a tier partitioning algorithm, named Cell-Slack Sorting, that tries to assign timing critical elements into the bottom tier to address the top tier cell degradation. (5) We present a timing-driven MIV planning method and a post-route optimization flow that minimize the routing congestion and performance loss

ISLPED '16, August 08-10, 2016, San Francisco Airport, CA, USA
© 2016 ACM. ISBN 978-1-4503-4185-1/16/08…$15.00
DOI: http://dx.doi.org/10.1145/2934583.2934622

Table 1: Nomenclature list used in this paper.

TT	typical transistor corner (= no I_{on} degradation)
LT10p	10% I_{on} degradation in the top tier device
LT20p	20% I_{on} degradation in the top tier device
SVT	standard threshold voltage cell
LVT	low threshold voltage cell
Cu5	5 layers of copper BEOL used in the bottom tier
Cu3	3 layers of copper BEOL used in the bottom tier
W5	5 layers of tungsten BEOL used in the bottom tier
W3	3 layers of tungsten BEOL used in the bottom tier

Table 2: Our benchmark circuits, where the metrics are from 2D IC designs. All designs are implemented with a foundry-grade 7nm bulk FinFET technology.

	DES3	LDPC
Cell Count	44,978	59,297
Wire Cap : Pin Cap	28:72	64:36
Avg Net Length (um/net)	2.54	10.02
Avg Net Wire Cap (fF/net)	0.41	1.77
Avg Net Pin Cap (fF/net)	1.03	0.97
Circuit Type	FEOL-dominant	BEOL-dominant

from the BEOL degradation. Experiments show that our design solutions allow only 3% performance degradation compared with 2D under harsh FEOL/BEOL variation settings.

2. FEOL/BEOL VARIATION IMPACT

Table 1 shows the nomenclature used in this paper. We choose Triple Data Encryption Standard Cipher (DES3) and Low-Density Parity Check Decoder (LDPC) from OpenCore benchmark suites to cover different types of circuit. With regard to capacitance composition in Table 2, LDPC is a BEOL-dominant, and DES3 is a FEOL-dominant circuit. Figure 1 shows GDS layouts of their 2D implementation. All designs are implemented with foundry-grade 7nm bulk FinFET PDK. Using these two benchmarks, we factorize the impact of inter-tier variations caused by low temperature process on the performance of full-chip 2-tier gate-level M3D design. The diameter of MIV is assumed to be the width of top metal layer in the bottom tier (36nm for 5 metal, 24nm for 3 metal) with resistance of 16Ω and capacitance of $0.01fF$.

2.1 Top Tier Device Degradation

The exact correlation between the low thermal budget process and the degree of top tier device degradation in the advanced node is not fully known. However, one of the main factors affected by low temperature process is expected to be the mobility of top tier devices. We illustrate the degree of top tier degradation for two scenarios, where 10% and 20% I_{on} decrease by mobility reduction. We refer to these degraded transistors as LT10p, LT20p corner, respectively. In order to evaluate the impact of mobility degradation on a device, we measure I_{on}, I_{off} of the device by sweeping mobility parameter. We use 7nm bulk FinFET Standard V_{TH} (SVT), and Low V_{TH} (LVT) compact model from a foundry-grade PDK with the nominal supply voltage 0.65V. From simulation results, we observe that 18.2% and 33.4% mobility degradation cause 10% and 20% I_{on} decrease, and 18.5% and 34.0% I_{off} reduction, respectively.

To analyze cell-level performance degradation, we characterize standard cell libraries for LT10p and LT20p corners with Cu local interconnect using Virtuoso Liberate. Using these cell models, we measure output slew and gate delay of cells assuming 10ps input slew and FO3 inverters with 300 Contacted Poly-Pitch (CPP $= 42nm$)-length M2 wire loading using Synopsys PrimeTime. Table 3 shows that LT10p and LT20p corners result in 10.0%, 22.7% of cell performance degradation, respectively.

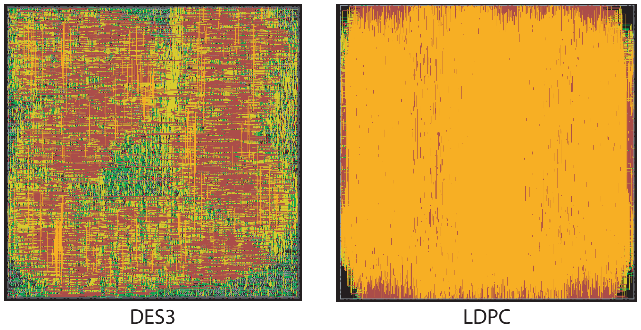

Figure 1: GDS layouts of 2D designs of our benchmark.

Table 3: Impact of mobility degradation on cell performance. We show the average output slew and delay in (ps) among INVx1, ND2x1, XNR2x1, AOI22x1, and DFF Clk-Q. Copper local interconnects are used.

	TT, Cu	LT10p, Cu	LT20p, Cu
Avg. SVT output slew	22.72 (1.00)	25.16 (1.11)	28.37 (1.25)
Avg. SVT cell delay	86.29 (1.00)	94.61 (1.10)	105.05 (1.22)
Avg. LVT output slew	16.62 (1.00)	18.27 (1.10)	20.32 (1.22)
Avg. LVT cell delay	57.28 (1.00)	62.90 (1.10)	69.81 (1.22)

Starting from ideal scenario where there is no degradation on top tier devices and equivalent 5 metal layers of Cu BEOL in both tiers, Figure 2 shows the impact of top tier device degradation on the maximum performance of full-chip 2-tier gate-level M3D design. We use Shrunk 2D flow [5, 7] based on foundry-grade 7nm bulk FinFET PDK. Since Shrunk 2D flow does not handle the inter-tier variation, we modify only the last tier-by-tier routing stage to consider the inter-tier performance variation as described in Section 3.3. Assuming 20% I_{on} reduction on the top tier, the performance of DES3, and LDPC is degraded by 21%, and 10% respectively. DES3 is a FEOL-dominant circuit, and 99% of the longest path delay consists of cell delays. Therefore, the performance of DES3 is sensitive to the FEOL degradation. On the other hand, LDPC has net delays of 21% out of the longest path delay, so the impact of cell delay increase is less than that of DES3. The simulation result identifies top tier device degradation as one of the critical obstacles to meet the timing of M3D design.

2.2 Bottom Tier Interconnect Degradation

In order to provide more thermal budget for the top tier manufacturing, integrating W BEOL in the bottom tier is an alternative. To quantify the impact of W interconnect, we first modify the process file (ICT) from foundry-grade 7nm PDK assuming W conducting layers and TiN liners with same geometry as Cu BEOL, and generate QRC technology file (TCH) using Cadence Techgen. Next, we characterize standard cell libraries based on W local interconnect using Virtuoso Liberate. Since wirelength of local interconnect is very short in the cell layout, we observe only 2% slew degradation and 1% output delay increase in both SVT and LVT cells. Based on these interconnect and cell models, Figure 3 shows the impact of tungsten interconnect in the bottom tier on the maximum performance of full-chip 2-tier M3D designs. We assume no device degradation on the top tier. We observe that M2, M3 layers have 2.20 times as high as Cu resistance (ohm/um), and 2.46 times

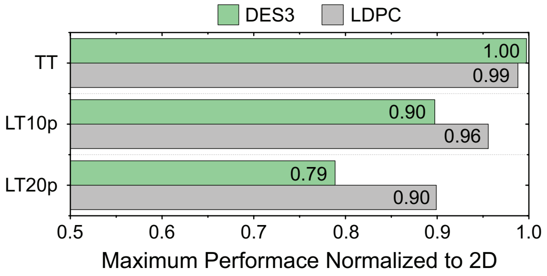

Figure 2: Impact of top tier device degradation on full-chip 2-tier M3D performance. We use 5 layers of Cu BEOL in both tiers. DES, our FEOL-dominant circuit, is more sensitive to the degradation.

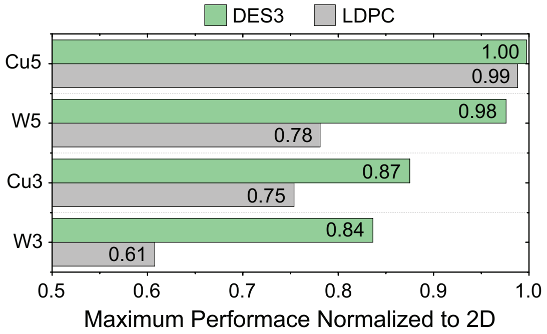

Figure 3: Full-chip impact of tungsten BEOL and metal layer saving in the bottom tier. LDPC, our BEOL-dominant circuit, is more sensitive to the changes.

for M4, M5 layers. When we compare the maximum performance under 5 layers of Cu BEOL (Cu5) with the result under 5 layers of W BEOL (W5) case, LDPC has 21% performance degradation while the performance of DES3 is decreased by only 2%. This is because net delays of BEOL-dominant LDPC are significantly increased due to the highly resistive W interconnect. DES3 has minor performance degradation since most of the path timing consists of cell delays.

Another interesting perspective on the bottom tier BEOL is to reduce the number of metal layers. BEOL cost increases significantly from N28 to N7 nodes due to dimensional scaling and multiple patterning processes [4]. Therefore, reducing metal stack must be taken into account to make M3D cost-effective. We consider 2 metal layers saving from the bottom tier in Figure 3. When we compare Cu3 result with Cu5 case, both DES3 and LDPC have significant performance loss. Reduced routing resources cause huge routing congestion in the bottom tier, resulting in 14% capacitance increase and 17% resistance increase on average. Under the worst scenario where we use W BEOL and reduce 2 metal layers from the bottom tier, 16% and 39% of maximum performance is degraded in DES3 and LDPC, respectively.

3. PHYSICAL DESIGN SOLUTIONS

To tackle the inter-tier performance variations caused by top tier low temperature manufacturing, we propose new full-chip M3D physical design flow named Derated 2D. Four CAD methodologies are proposed in Derated 2D flow as follows: (1) We do not modify the layout objects but complete the 2D IC design with derated RC parasitic corner, named Derated 2D design. Then we project the placement result of Derated 2D design into the final footprint of M3D design. (2) For the low-temperature process-aware tier partitioning, we use cell-slack as a metric for the timing criticality, and assign the timing critical elements into the bottom tier to address top tier cell degradation. (3) Timing-driven MIV planning deals with resistive W interconnect and reduced metal stack in the bottom tier. (4) A post-route optimization flow compensates the performance degradation under various FEOL/BEOL degradation settings at a minimum energy overhead. Overall design methodologies for Derated 2D flow is shown in Figure 4.

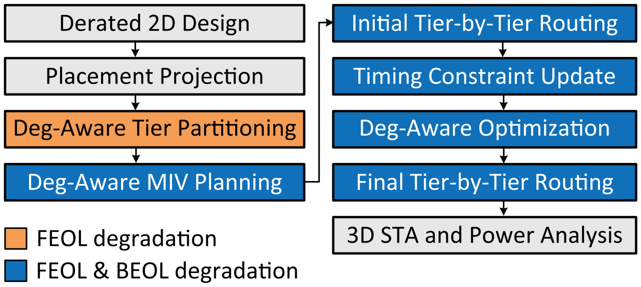
Derated 2D Design
Placement Projection
Deg-Aware Tier Partitioning
Deg-Aware MIV Planning
Initial Tier-by-Tier Routing
Timing Constraint Update
Deg-Aware Optimization
Final Tier-by-Tier Routing
3D STA and Power Analysis
🟧 FEOL degradation
🟦 FEOL & BEOL degradation

Figure 4: Derated 2D, our FEOL/BEOL degradation-aware physical design flow for gate-level M3D. Our tier partitioning step tackles FEOL degradation, while the subsequent steps address both FEOL and BEOL degradation.

Table 4: Comparison between our Derated 2D flow and state-of-the-art Shrunk 2D flow [5].

	Derated 2D	Shrunk 2D
Shrink chip footprint?	No	Yes
Shrink cell layout?	No	Yes
Shrink metal dimension?	No	Yes
Scale unit-length RC parasitics?	Yes	Yes
Consider FEOL degradation?	Yes	No
Consider BEOL degradation?	Yes	No
Bottom tier cells use top tier metal?	Yes	No
Post-route optimization supported?	Yes	No

3.1 Derated 2D Design and Projection

Unlike Shrunk 2D flow [5] that requires shrinking of layout objects and RC parasitic scaling, Derated 2D uses original layout objects. However, Derated 2D is also possible to have overestimated wire load and redundant buffers, unless we consider the wirelength saving from reduced footprint of M3D design. Assuming no silicon area overhead, 2-tier M3D design has half footprint of that of 2D. In order to optimize Derated 2D design with same RC parasitics of M3D design, we first create an RC corner which is derated by $1/\sqrt{2}$ for total R and total C while not scaling coupling capacitance due to the same routing pitch in Derated 2D and M3D. Then we project the whole placement result of Derated 2D design into the footprint of final M3D design. Since every manhattan distance between each macro is scaled by $1/\sqrt{2}$ as a result of placement projection, RC parasitic of Derated 2D design is expected to be the same as that of M3D design. Table 4 and Figure 5 show comparison between our Derated 2D flow and state-of-the-art Shrunk 2D flow.

3.2 Tier Partitioning and MIV Planning

Cell slack is a metric to measure how long each cell may delay without compromising the timing of paths propagating through the cell. We extract the slack value of timing constrained cells on the Derated 2D design using Synopsys PrimeTime, and use it as a met-

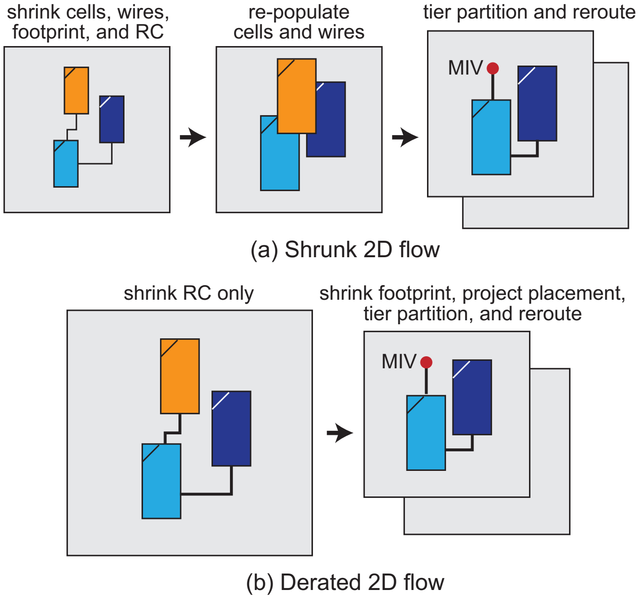
shrink cells, wires, footprint, and RC
re-populate cells and wires
tier partition and reroute
MIV
(a) Shrunk 2D flow
shrink RC only
shrink footprint, project placement, tier partition, and reroute
MIV
(b) Derated 2D flow

Figure 5: Illustration of Shrunk 2D [5] and Derated 2D flow.

ric to represent timing criticality of a cell. For the clock network cells, we keep them to the bottom tier. The simplest partitioning scheme using cell slack is to sort them in decreasing order, and if the slack of a cell is less than median, then to place it on the bottom tier. In most cases, however, these timing critical cells are usually placed close to each other, resulting in local area skew between each tier. With unbalanced area skew, cell slack sorting does not guarantee minimum performance degradation since the original location of a cell that is already optimized at Derated 2D design must be changed during placement legalization. Therefore, we divide the design footprint by small size partitioning bin in the regular fashion, and sort cells within each bin by the slack value in decreasing order to meet the local area balance.

For the MIV planning, Shrunk 2D flow introduces CAD methodology to manipulate commercial engine built for 2D ICs [7]. Our timing-driven MIV planning in Derated 2D flow also uses the basic idea of MIV planning scheme in Shrunk 2D flow but the differences are as follows: (1) In the same way that we create a 3D LEF to define two macro flavors for each standard cell - one for each tier, we create a 3D LIB that defines two timing flavors to consider inter-tier device performance variation. Thus, timing model for each cell in 3D space is mapped to its appropriate transistor corner. We import the 3D LIB into commercial router (Candence Innovus) to create delay corner for timing-driven routing. (2) Since each tier is possible to have different routing material and number of metal layers, we create a process file (ICT) for full 3D metal stack and generate 3D TCH file using Cadence Techgen. This 3D TCH file contains the RC parasitic information for every routing layers in M3D design. Then, we create a parasitic corner with this 3D TCH file in the commercial route. With timing constraint same as 2D design, we do timing-driven routing to insert MIVs. Using full 3D metal stack makes it possible to share the routing resources from all tiers. If we reduce the metal stack on the bottom tier, then the router uses top tier metal layers to route bottom tier nets in order to minimize routing congestion. If W BEOL is used in the bottom tier, the tool tries to use low resistive Cu BEOL on the top tier to minimize timing degradation. Figure 6 shows the differences of MIV planning scheme between Shrunk 2D and Derated 2D flow.

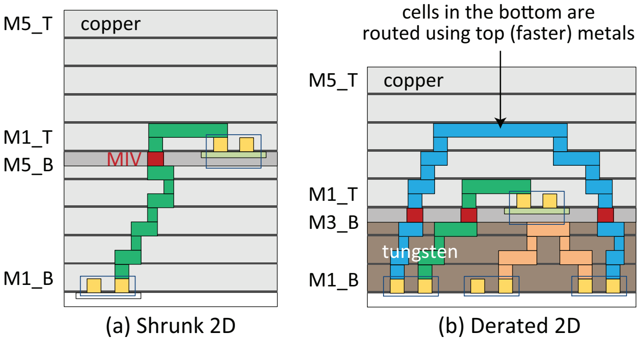

Figure 6: Metal stack comparison. (a) Shrunk 2D [5] with 5 Cu metal layers in both tiers, (b) Derated 2D flow with 5 layers of Cu in the top, and 3 tungsten in the bottom. Top cells contain MIV routing obstacle underneath.

3.3 Post-Route Optimization and Routing

Since the initial Derated 2D design only involves normal transistor corner and Cu BEOL, it is clear that there exists limitation for timing closure under inter-tier variations in M3D design. Since we create delay and parasitic corners at timing-driven MIV planning stage, it is also possible to use a post-route optimization flow to update initial Derated 2D design for timing closure at a mimimum energy overhead. We update the timing constraint for a post-route optimization in consideration of cell legalization during tier-by-tier routing. Then, we change the size of macros in 3D LEF into the size of placement site, which is the smallest dimension that a macro can have. By using unit size 3D macros, we remove the placement overlap and again minimize the cell legalization during post-route optimization. We keep the location of initial top tier cells, and allow the commercial tool to optimize bottom tier cells by resizing and VT swapping for timing closure. The reason why we do not play with the top tier cells is that the MIV routing blockages are initially fixed under the placement result of top tier cells. After post-route optimization, we proceed final tier-by-tier routing to create separate GDS for each tier.

Once the MIV locations are determined by MIV planning, we create a DEF file for each tier containing the location of MIVs as primary I/O. Using original macro LEF, we repopulate the cell size and legalize the placement overlap. We route them initially with appropriate LIB (TT,LT10p,LT20p) and TCH (Cu,W) to the specific FEOL/BEOL degradation scenario, and create the timing context of each tier to optimize the routing quality. After routing under the timing context, we extract the parasitic, and proceed to 3D timing and power analysis using Synopsys PrimeTime.

4. EXPERIMENTAL RESULTS

4.1 Impact of Tier Partitioning

Figure 7 shows the impact of cell-slack sorting tier partitioning on the design performance compared with Fiduccia-Mattheyes (FM) min-cut partitioning algorithm [5]. To be an equal comparison, we use Derated 2D design for both of the partitioning algorithms, and assume 5 layers of Cu BEOL in both tiers. Even under 20% I_{ON} degradation on the top tier, cell-slack sorting partitioning allows only 5% of performance degradation in both benchmarks. Table 5 shows detailed statistics of M3D designs from different partitioning algorithms. Min-cut partitioning tries to minimize the connections between each tier inside the partitioning bin. There-

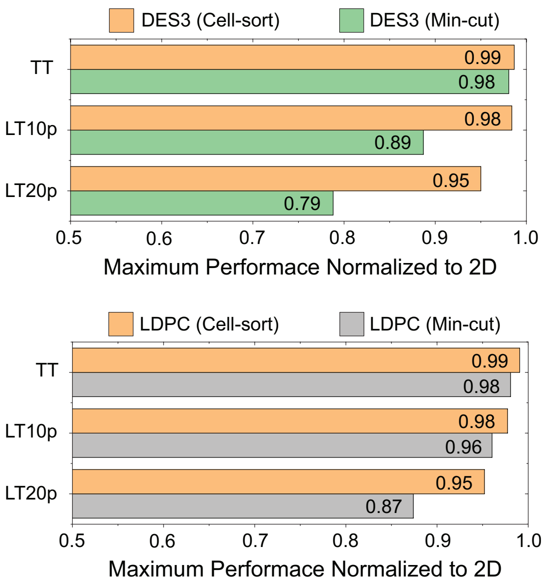

Figure 7: Tier partitioning impact on performance under FEOL degradation. Our cell sorting-based method withstands the degradation better than min-cut for both circuits.

Table 5: Comparison between cell-slack sorting vs. min-cut tier partitioning. We use LT20p transistor corner in the top tier, and 5 layers of Cu BEOL in both tiers.

	LDPC		DES3	
	min-cut	cell-sort	min-cut	cell-sort
Cell Count	57451		44805	
Net Count	59696		45036	
2D Net (top tier) Count	23171	16284	16677	10289
2D Net (bot tier) Count	24118	21718	23063	13701
3D Net Count	12407	21694	5296	21046
MIV Count	25958	37189	6772	25500
Avg. MIV# of 3D Net	2.09	1.71	1.28	1.21
Avg. WL of 2D Net (um/net)	3.29	2.83	2.18	1.69
Avg. R of 2D Net (ohm/net)	416.81	372.45	342.47	268.52
Avg. WL of 3D Net (um/net)	23.42	14.72	5.60	3.91
Avg. R of 3D Net (ohm/net)	2692.18	1743.30	867.53	650.95
Target Clock (ns)	1.0	1.0	0.5	0.5
WNS (s)	-0.16	-0.07	-0.18	-0.06
TNS (s)	-68.59	-2.86	-52.46	-3.58
TPS (s)	19.85	90.87	2605.71	2618.53
Runtime (sec)	24	111	12	44

Table 6: Comparison between MIV planning in Shrunk 2D [5] vs. our Derated 2D. We assume no FEOL degradation and use 3 tungsten BEOL layers in the bottom tier in LDPC benchmark. Derated 2D encourages more routing in the top tier (= faster Cu BEOL).

	metric	Shrunk 2D	Derated 2D
net stats	top placed, top routed	17,432	17,410
	top placed, top/bot routed	0	22
	bot placed, bot routed	22,280	19,072
	bot placed, top/bot routed	0	3,208
	top/bot placed, top/bot routed	19,984	19,984
top tier	Avg. Net Length (um/net)	5.40	**6.85**
	Avg. Net Cap (ff/net)	2.70	**2.92**
	Avg. Net Wire Cap (ff/net)	0.92	**1.24**
	Avg. Net Res (Ohm/net)	601.81	**758.12**
bot tier	Avg. Net Length (um/net)	**3.50**	2.64
	Avg. Net Cap (ff/net)	**2.62**	2.45
	Avg. Net Wire Cap (ff/net)	**0.78**	0.52
	Avg. Net Res (Ohm/net)	**1192.32**	916.06
	Avg. MIV# per 3D net	2.1	1.6
	Max. Performance (GHz)	**0.68**	**0.75**
	Power-Delay Product (pJ)	32.59	32.22

fore, 2D nets on each tier get longer and congested, leading to further longer 3D nets. However, cell slack sorting partitioning uses as many MIVs as necessary in order to assign the timing critical cells to the bottom tier. While minimizing the impact of top tier cell delay increase, these many and short 3D connections also effectively reduce net delay, resulting in significant timing saving. The incremental gain update makes FM min-cut heuristic run in $O(C)$, where C is the number of cells. Cell-slack sorting runs in $O(C \log C)$ by sorting algorithm.

4.2 Impact of MIV Planning

Based on cell slack sorting tier partitioning, Figure 8 shows the impact of our timing-driven MIV planning compared with Shrunk 2D flow. Under no top tier device degradation, W BEOL and 2 metal layer reduction in the bottom tier leads to 23% and 36% performance degradation in DES3 and LDPC with MIV planning in Shrunk 2D flow. Our timing-driven MIV planing, however, allows only 13% and 20% of the performance degradation in DES3 and LDPC respectively. In Table 6, we analyze net distribution and parasitics of LDPC design. Since 2D nets are possible to become 3D nets with our timing-driven MIV planning to close the timing, nets in the resistive bottom tier are routed by Cu BEOL in the top tier. Also, sharing routing resources between each tier decreases average net length and RC parasitics on the bottom tier and balances the routing congestion caused by metal layer reduction. Therefore, net delay degradation caused by W BEOL and routing congestion on the bottom tier are minimized, resulting in performance saving. In addition, top tier device degradation has a minor impact on design performance when bottom nets are in the worst scenario as a result of cell-slack sorting tier partitioning.

4.3 Comparison with Shrunk 2D Flow

Based on a foundry-grade 7nm FinFET PDK, we compare the design results of our Derated 2D flow with results of Shrunk 2D flow under all inter-tier variation scenarios in Table 7. Under the worst scenario, where there is 20% I_{on} reduction in the top tier while saving 2 metal layers of W BEOL in the bottom tier, Derated 2D result of LDPC achieves 36% of performance improvement, and 10% of energy saving compared with Shrunk 2D result without post-route optimization.

When we compare the design results between LDPC and DES3, we observe that energy saving from M3D depends on the circuit type. LDPC, which is a BEOL-dominant circuit, has energy saving of 22% in ideal scenario, and still has 16% saving under the worst scenario without post-route optimization. This is because major source of energy saving from M3D design is wirelength reduction. In 2-tier M3D design with Derated 2D flow, expected maximum total wirelength saving is 29.3% considering 50% footprint saving from M3D design. Although the routing congestion in each tier is possible to degrade the wirelength saving depending on the circuit, this huge wirelength saving leads to around 30% of wire capacitance saving, and if the design is BEOL-dominant such as LDPC that 64% of total capacitance is wire capacitance, total capacitance saving becomes 22%. This capacitance saving is directly converted into switching power saving of 22%. However, since the total power consists of switching power, internal power, and leakage, the final power saving become 16%. In the case of DES3, wire

Table 7: Performance and power-delay product (= energy) comparison under various FEOL and BEOL degradation settings. Our Derated 2D consistently outperforms Shrunk 2D [5] in terms of both performance and energy, even in the worst-case scenario (20% slow device, 3 layers of tungsten routing). Our post-route optimizer further improves performance at the expense of energy increase.

FEOL/BEOL setting		Maximum performance normalized to 2D			Post-route Optimization impact on TNS		Power-Delay Product normalized to 2D		
top tier	bot tier	Shrunk 2D	Derated 2D	D2D+PostOpt	Derated 2D	D2D+PostOpt	Shrunk 2D	Derated 2D	D2D+PostOpt
LDPC									
TT, Cu5	TT, Cu5	0.98	1.01	1.01	-0.01	-0.01	0.84	0.78	0.78
TT, Cu5	TT, Cu3	0.78	0.91	0.99	-13.06	-0.27	0.92	0.84	0.85
LT10p, Cu5	TT, Cu3	0.77	0.89	0.98	-16.05	-0.33	0.92	0.84	0.85
LT20p, Cu5	TT, Cu3	0.75	0.84	0.98	-38.12	-0.28	0.92	0.84	0.86
TT, Cu5	TT, W3	0.61	0.80	0.98	-137.90	-0.12	0.93	0.85	0.90
LT10p, Cu5	TT, W3	0.60	0.79	0.98	-159.11	-0.33	0.93	0.85	0.90
LT20p, Cu5	**TT, W3**	**0.58**	**0.79**	**0.97**	-186.12	-0.19	**0.93**	**0.84**	**0.92**
DES3									
TT, Cu5	TT, Cu5	1.00	0.97	1.03	-0.82	-1.08	0.98	0.98	0.98
TT, Cu5	TT, Cu3	0.87	0.90	1.02	-4.16	-1.78	1.01	0.99	1.00
LT10p, Cu5	TT, Cu3	0.86	0.90	1.03	-4.42	-1.43	1.01	0.99	1.00
LT20p, Cu5	TT, Cu3	0.79	0.90	1.03	-7.15	-2.32	1.01	0.99	1.01
TT, Cu5	TT, W3	0.84	0.87	1.01	-8.97	-2.77	1.02	1.00	1.01
LT10p, Cu5	TT, W3	0.82	0.87	1.03	-8.80	-2.44	1.02	1.00	1.01
LT20p, Cu5	**TT, W3**	**0.78**	**0.87**	**1.02**	-11.33	-1.96	**1.02**	**1.00**	**1.01**

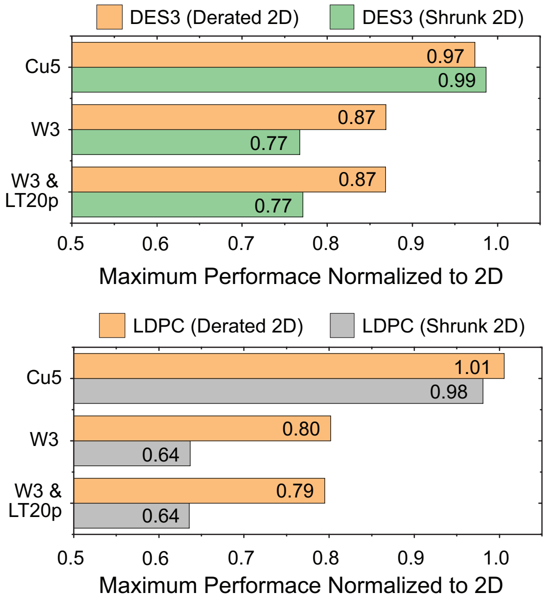

Figure 8: Impact of MIV planning in Derated 2D vs. Shrunk 2D [5]. Our Derated 2D withstands the FEOL and BEOL degradation better than Shrunk 2D.

capacitance is only 28% of total capacitance. Therefore, switching power saving from 30% of wirelength reduction in M3D is degraded into only 8%, and the final power saving degraded by internal power ratio become 2%.

We also tabulate the impact of a post-route optimization flow on timing and power-delay product. Under any scenarios, the post-route optimization makes it possible to restore the performance degradation of M3D design up to minimum 97% of 2D performance. Under the worst scenario where 20% I_{on} degradation on the top tier and W BEOL with 2 metal layer saving in the bottom tier, we recover the M3D performance of LDPC from 79% to 97% of 2D performance at the expense of 8% of energy. In case of DES3, we observe that although FEOL-dominant circuit has less energy saving, since it is less affected by resistive W interconnect and bottom routing congestion than BEOL-dominant circuit, it requires only 1% of 2D energy to restore the performance degradation.

5. CONCLUSION

In this paper we proposed CAD methodologies for gate-level monolithic 3D ICs (M3D) that tackle the FEOL/BEOL inter-tier variations caused by low temperature manufacturing. To address the top tier device degradation, we presented a cell-slack sorting-based tier partitioning algorithm that assigns timing critical elements into the bottom tier. To deal with the BEOL impact, we developed a timing-driven MIV planning flow and a post-route optimization flow to compensate for the reduced routing layers and increased resistance of tungsten interconnect. Experiments along with 7nm bulk FinFET from a foundry-grade PDK demonstrated the effectiveness of our approach.

6. REFERENCES

[1] P. Batude et al. 3D sequential integration opportunities and technology optimization. In *Proc. IEEE Int. Interconnect Technology Conference*, pages 373–376, 2014.

[2] K. Chang et al. Power benefit study of monolithic 3D IC at the 7nm technology node. In *Proc. Int. Symp. on Low Power Electronics and Design*, pages 201–206, 2015.

[3] F. Luce et al. Methodology for thermal budget reduction of SPER down to 450C for 3D sequential integration. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 370:14 – 18, 2016.

[4] A. Mallik et al. The economic impact of EUV lithography on critical process modules. In *Proc. SPIE*, volume 9048, pages 90481R–90481R–12, 2014.

[5] S. Panth et al. Design and CAD methodologies for low power gate-level monolithic 3D ICs. In *Proc. Int. Symp. on Low Power Electronics and Design*, pages 171–176, 2014.

[6] S. Panth et al. Power-performance study of block-level monolithic 3D-ICs considering inter-tier performance variations. In *Proc. ACM Design Automation Conf.*, pages 1–6, 2014.

[7] S. Panth et al. Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs. *IEEE Trans. on Computer-Aided Design of Int. Circuits and Systems*, 34(4):540–553, 2015.