# How Much Cost Reduction Justifies the Adoption of Monolithic 3D ICs at 7nm Node?

Bon Woong Ku†, Peter Debacker§, Dragomir Milojevic§, Praveen Raghavan§, 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 study power, performance, and cost (PPC) tradeoffs for 2-tier, gate-level, full-chip GDS monolithic 3D ICs (M3D) built using a foundry-grade 7nm bulk FinFET technology. We first develop highly-accurate wafer and die cost models for 2D and M3D to study PPC tradeoffs. In our study, both 2D and M3D designs are optimized in terms of the number of BEOL metal layers used for routing to obtain the best possible PPC values. We develop a new CAD methodology for 2-tier gate-level M3D, named Projected 2D Flow, that allows us to accurately compare RC parasitics of equivalent nets in both 2D and M3D designs. Our experiments based on two different circuit types (BEOL-dominant vs. FEOL-dominant) confirm that M3D designs indeed offer a significant footprint saving. However, to our surprise, the PPC quality of M3D turns out to be worse than that of 2D by 34% due to the high wafer cost of M3D. Our study also reveals that M3D wafer yield should be as high as 90% of 2D wafer yield, and the M3D device manufacturing cost should be less than 33% of that of 2D to justify the adoption of M3D technology at the 7nm era. Lastly, and counter-intuitively, our study shows that FEOL-dominant circuit shows more PPC benefits from M3D technology than BEOL-dominant circuit.

## CCS Concepts

**•Hardware → 3D integrated circuits; Physical design (EDA); Methodologies for EDA; Yield and cost modeling;**

## Keywords

Monolithic 3D IC; Gate-level; Cost Modeling;

## 1. INTRODUCTION

As 2D device scaling faces toward physical limitation, considerate efforts for 3D integration have been made to extend technology scaling benefits. Over the last few years, monolithic 3D (M3D) technology, where active layers are implemented on top of the bottom tier sequentially, has emerged as a promising solution for the massive vertical interconnection. While through-silicon-via (TSV) based 3D integration requires chip-level or wafer-level alignment process with $\mu m$-scale precision, M3D integration manipulates litho-scale alignment enabling extremely small size of the monolithic inter-tier vias (MIVs). These tiny MIVs not only have minimized area overhead, but also offer the inter-tier vertical connections in orders of magnitude. Therefore, effectively inserted MIVs significantly reduce wirelength of 3D nets resulting in huge power-delay benefits [1].

Depending on granularity of tier partitioning, M3D technology is categorized into transistor-level, gate-level, and block-level [2, 3, 4]. Out of these 3 types, gate-level M3D allows to harness dense vertical interconnections more than block-level M3D, resulting in sufficient wire length decrease in global routing. While transistor-level M3D requires additional efforts for new layout and characterization of standard cells due to the split of PMOS and NMOS into different tiers, gate-level M3D allows to reuse existing 2D standard cell libraries for the full-chip GDS M3D ICs [3]. In this paper, we study power, performance, and cost (PPC) tradeoffs for 2-tier, gate-level, full-chip GDS M3D IC built on a foundry-grade 7nm bulk FinFET technology.

Most of the earlier works on gate-level M3D have focused on power, performance, and area improvement in 2-tier design given the same routing resources as 2D IC, and no silicon area overhead. For example, if a 2D IC has 5 metal layers and $100mm^2$ footprint, then 2-tier M3D IC has 5 metal layers and $50mm^2$ footprint on top and bottom tiers each. Based on those assumptions, [5, 6, 7] shows that gate-level M3D ICs indeed offer huge iso-performance power saving compared with 2D ICs. Simply and ideally thinking, 50% footprint saving in M3D ICs results in 29.3% wire length reduction ($1/\sqrt{2}$ wire length scaling) if the design aspect ratio is assumed to be the same, and also if tier partitioning is done by placement-driven partitioning [8]. This wire length saving not only decreases wire capacitance (switching power saving) but also provides path timing margin to reduce buffer counts (internal power saving). Therefore, if the type of a design is wire capacitance dominant circuit, power saving in gate-level M3D is expected to be more.

However, since the footprint of wire capacitance dominant circuit could be determined by routability of limited routing resources, the design quality of this type of circuit would be easily improved when we add more routing layers. While M3D design needs to have the number of metal layers as few as possible for the process cost reduction, adding more metal layers and optimizing Back-End-Of-Line (BEOL) metal stack in 2D IC can be easily achieved with reasonable cost overhead [9]. Therefore, it leads us to the next questions on how to set the proper 2D reference design for the fair PPC comparison with M3D design, and how much M3D has PPC-competitiveness to make us move toward M3D era. This paper addresses above questions.

*ICCAD '16, November 07-10, 2016, Austin, TX, USA*
© 2016 ACM. ISBN 978-1-4503-4466-1/16/11…$15.00
DOI: http://dx.doi.org/10.1145/2966986.2967044

**Table 1: Nomenclature.**

| | | | |
|---|---|---|---|
| $C_{W_{FEOL}}$ | Manufacturing cost for FEOL | | |
| $C_{M_i}$ | Normalized manufacturing cost for metal layer $M_i$ | | |
| $C_{W_{BEOL,N}}$ | Manufacturing cost for $N$ BEOL layers | | |
| $A_{W\|D}$ | Wafer \| Die area | $Y_{W\|D}$ | Wafer \| Die yield |
| $D_W$ | Wafer defect density | $DPW$ | # dies per wafer |
| $C_{W\|D_N}$ | Wafer \| Die cost for 2D IC with $N$ BEOL layers | | |
| $C_{W\|D_{N,M}}$ | Wafer \| Die cost for M3D IC with $N$ (top) and $M$ (bottom) BEOL layers | | |
| $\alpha$ | Variable for M3D top tier manufacturing & bonding | | |
| $\beta$ | Variable for M3D wafer yield degradation | | |

Our contributions include the followings: (1) We develop highly-accurate full-chip, GDS-based wafer and die cost model for 2D and M3D. Based on these cost modeling, we optimize the number of routing metal layers to obtain the best possible PPC values in 2D IC of two widely different circuit types (BEOL-dominant vs. FEOL-dominant). (2) We propose a new design solution for 2-tier gate-level M3D, named Projected 2D Flow, that offers more than 50% footprint saving compared with that of 2D with minimum effort. Through Projected 2D Flow, we also compare RC parasitics of equivalent nets in both 2D and M3D designs, enabling for accurate understanding of from where the benefits of M3D design derive. (3) We study PPC tradeoffs for 2-tier, gate-level full-chip GDS M3D ICs built using a foundry-grade 7nm bulk FinFET technology. Our experiments based on two widely different circuit types reveal by how much cost should be further reduced to justify the adoption of M3D technology at the 7nm era.

# 2. COST MODELING

Previous works [10, 11] on cost modeling for 3D IC are based on estimation of design parameters. Those studies use empirical constant for the area of standard cells, and expected wirelength distribution to predict total die area, and the number of required BEOL layers. In this paper, we develop cost models based on real full-chip GDS design result. By using physical design flow proposed in this paper, routing and placement utilization is maximized in M3D design, resulting in accurate footprint and practical number of metal layers required to meet design specification.

## 2.1 Wafer Cost Model

Wafer cost has been increasing as dimensional scaling advances toward aggressive pitch node. On the manufacturing side, complex lithography, such as multiple patterning, is one of the main reasons for high wafer cost. Even though introduction of Extreme Ultraviolet Lithography (EUVL) is expected to reduce manufacturing cost, complicated device structure and the growth of system design complexity require more efforts on process control challenges for sufficient die yield [12, 9].

Through the cost analysis framework from our industry partner, we develop simple but self-contained wafer cost models for 2D and M3D technology. Considering prescribed sequence of 7nm bulk FinFET process flow, and based on Cost-of-Ownership (CoO) where a database framework considers throughput of fab tools, material, labor, repair, utility and overhead expenses due to the equipment operation [13, 12], we set the ratio between FEOL and BEOL manufacturing cost as 30%:70%. 2D BEOL metal stack configuration used in this paper is in accordance with International Technology Roadmap for Semiconductor (ITRS) guidelines for 7nm technology node. Since our foundry-grade 7nm bulk FinFET device technology is assumed to have Middle-End-Of-Line (MEOL), MINT layer is included in the metal stack, but it is only used for intra cell routing.

**Table 2: Assumed patterning option and manufacturing cost per metal layer.**

| Layer | Patterning | Pitch | Width | Thickness | Normalized Cost ($C_{M_i}$) |
|---|---|---|---|---|---|
| MINT (M0) | SAQP | 32nm | 21nm | 24nm | 2.8 |
| M1 | LELE | 42nm | 24nm | 24nm | 1.7 |
| Mx | SAQP | 32nm | 24nm | 24nm | 2.8 |
| My | LELE | 48nm | 24nm | 48nm | 1.5 |
| Mz | LE | 80nm | 40nm | 80nm | 1.0 |

Table 2 shows the assumed patterning option and manufacturing cost per metal layer ($C_{M_i}$) obtained from industry partner. Manufacturing costs for MEOL and intermediate interconnect layers are normalized with the cost for global interconnect layer (Mz). With this Table and proposed ratio between FEOL and BEOL manufacturing cost, we set the reference design as 2D IC with 8 BEOL metal layers, and calculate the normalized wafer cost for another designs with different metal stack as shown below.

$$C_{W_{FEOL}} = 0.3 \times C_{W_8}, C_{W_{BEOL,8}} = 0.7 \times C_{W_8}$$
$$C_{W_{BEOL,N}} / C_{w_{BEOL,8}} = \sum_{i=0}^{i=N} C_{M_i} / \sum_{i=0}^{i=8} C_{M_i}$$
$$C_{W_N} / C_{W_8} = (C_{W_{FEOL}} + C_{W_{BEOL,N}}) / C_{W_8}$$

**2D Wafer Cost Model**: For $N$ BEOL metal layers,

$$C_{W_N} / C_{W_8} = 0.3 + 0.7 \times \sum_{i=0}^{i=N} C_{M_i} / \sum_{i=0}^{i=8} C_{M_i} \qquad (1)$$

In literature, no work has previously studied cost estimation for M3D integration. Cost for sequential integration is not fully known yet, and top tier manufacturing should be limited due to the FEOL and BEOL integrity on the bottom tier. Therefore, we assume that the FEOL cost for both tiers are the same as default, and include a variable to take into account the different device manufacturing cost in each tier, and bonding cost ($\alpha$). M3D BEOL cost is calculated by the sum of BEOL cost for each tier.

**M3D Wafer Cost Model**: For $N$ (top) and $M$ (bottom) BEOL metal layers,

$$C_{W_{N,M}} / C_{W_8} = 0.6 + \alpha + 0.7 \times (\sum_{i=0}^{i=N} C_{M_i} + \sum_{i=0}^{i=M} C_{M_i}) / \sum_{i=0}^{i=8} C_{M_i} \qquad (2)$$

## 2.2 Die Cost Model

We assume that considerations for the cost of I/O pins, packaging, testing, and cooling are out of the scope in this paper. We also assume that edge clearance, and notch height of the wafer are ignorable. Based on those assumptions, our die manufacturing cost takes into account the number of dies per wafer, die yield, and die area. For M3D die yield, we multiply sensitivity variable $\beta$ with 2D wafer yield, so that it leads to evaluating how much M3D wafer yield should be improved to guarantee the M3D benefits compared with 2D. Finally,

**2D Die Cost Model**: For $N$ BEOL metal layers,

$$DPW_N = A_W/A_{D_N} - \sqrt{2\pi A_W/A_{D_N}} \tag{3}$$

$$Y_{D_N} = Y_W \times (1 + A_{D_N} D_W/2)^{-2} \tag{4}$$

$$C_{D_N}/C_{D_8} = \frac{C_{W_N}}{C_{W_8}} \times \left( \frac{DPW_8 \times Y_{D_8}}{DPW_N \times Y_{D_N}} \right) \tag{5}$$

**M3D Die Cost Model**: For $N$ (top) and $M$ (bottom) BEOL metal layers,

$$DPW_{N,M} = A_W/A_{D_{N,M}} - \sqrt{2\pi A_W/A_{D_{N,M}}} \tag{6}$$

$$Y_{D_{N,M}} = \beta \times Y_W \times (1 + A_{D_{N,M}} D_W/2)^{-2} \tag{7}$$

$$C_{D_{N,M}}/C_{D_8} = \frac{C_{W_{N,M}}}{C_{W_8}} \times \left( \frac{DPW_8 \times Y_{D_8}}{DPW_{N,M} \times Y_{D_{N,M}}} \right) \tag{8}$$

Experiments are done with $300mm$ of wafer diameter, and $0.2mm^{-2}$ of $D_W$, and 0.95 of $Y_W$.

# 3. DESIGN METHODOLOGIES

In [3], authors present state-of-the-art Shrunk 2D flow for full-chip GDS gate-level monolithic 3D IC. The idea of this design flow is to manipulate the powerful optimization capability of the commercial tool built for 2D ICs at pseudo-3D design environment where shrunk layout objects are placed and routed in the floorplan with the same dimension as final M3D. For example, assuming 2-tier, gate-level M3D design with zero silicon area overhead, the footprint of each tier should become 50% of 2D design footprint. For the Shrunk 2D flow, we first fix the floorplan size same as the footprint of final M3D, and shrink the geometric dimension of original 2D layout objects to scale by $\sqrt{2}$. Then the area of standard cells become 50% of original cell area, and also the pitch and width of interconnects become 70.7% of the original. Now, we also need to scale unit-length RC parasitic to let commercial router use the original parasitic of interconnects in optimization stages. This scaling procedure is necessary to remove the overlap between standard cells in the shrunk chip footprint, and to obtain reasonable timing optimization by commercial tool.

However, shrinking layout objects is subject to Design-Rule-Violations (DRV) in the complicated standard cell layouts in the advanced technology nodes. Also, scaling RC parasitics of shrunk interconnects to match the parasitics same as the original either is incorrect due to the exaggerated extrapolation of parasitics with internal algorithm in the commercial tool, or requires many efforts to modify the geometric and electrical characteristics in the interconnect files. Furthermore, those layout objects are not reusable in the design with more than 2-tiers. Lastly, Shrunk 2D Flow possibly maximizes the placement utilization of each tier in M3D design, but it does not fully optimize the design in terms of routing utilization since reduced footprint and effectively routed nets in M3D decreases total wirelength. Therefore, we propose new physical design solution for gate-level M3D design flow named Projected 2D. The main idea of this flow is to use 2D design itself as a starting point for implementation of M3D design.

## 3.1 Projected 2D Flow

Projected 2D does not require shrinking of layout objects, and scaling RC parasitics unlike Shrunk 2D flow. The beauty of Projected 2D flow is as follows: (1) After we implement 2D design that already closes design specification with normal Process-Design-Kit (PDK), Projected 2D uses final netlist and placement result of 2D design to implement M3D design. Since there is no difference between the netlist of 2D and that of M3D, it is possible to directly compare the routing result of equivalent nets. Analyzing RC parasitics of the nets through comparison with 2D, it allows us to confirm the wirelength saving from M3D, or to improve tier partitioning result for better M3D design quality. (2) Projected 2D maximizes either placement or routing utilization by projection of 2D placement result. Modulating the projection factor, we easily reduce final M3D design footprint by more than 50% if there is enough routing usage saving. (3) Projected 2D enables multi-tier, gate-level M3D design without any efforts for modification of geometric information in input design files.

**Table 3: Comparison between Projected 2D and state-of-the-art Shrunk 2D flow [3].**

| | Projected 2D | Shrunk 2D |
|---|---|---|
| Shrink macro layout? | No | Yes |
| Shrink interconnect dimension? | No | Yes |
| Scale unit-length RC parasitics? | No | Yes |
| Consider buffer saving in M3D? | No | Yes |
| Have same netlist as 2D? | Yes | No |
| Maximize routing utilization? | Yes | No |
| LDPC M3D result, 7nm bulk FinFET, M5 (top) / M5 (bottom) | | |
| Chip Area ($\mu m^2$) | 4499 | 5408 |
| Maximum routing utilization | 0.762 | 0.666 |
| Total buffer count | 16163 | 15980 |
| Total power (mW) | 32.76 | 32.41 |
| WNS (ns) | 0.057 | -0.015 |

However, Projected 2D overestimates wire loads, and inserts redundant buffers during 2D optimization by commercial tool. Table 3 shows qualitative, and quantitative comparison between Projected 2D and Shrunk 2D. Assuming 2-tier LDPC M3D design with a foundry-grade 7nm bulk FinFET PDK and 5 metal layers in both tiers, design result of Projected 2D flow has more buffers resulting in larger positive slack than that of Shrunk 2D. On the other hand, due to the reduced footprint of Projected 2D design, it has more wirelength saving and consequent switching power saving to compromise increase in internal power due to redundant buffer counts.

## 3.2 Tier Partitioning and MIV planning

Based on projected placement location of macros and netlist, we use placement-driven min-cut partitioning for tier partitioning [8]. This partitioning scheme divides the whole design in regular fashion for the balanced local area skew, so-called partitioning bin, and do Fiduccia Mattheyses (FM) min-cut partitioning inside each of partitioning bins. Therefore, the number of inter-tier connections depends on the size of partitioning bins. In [6], it is shown that there is an optimization point for the minimum power consumption along with the inter-tier connections. This is because too many 3D connections cause routing congestion and redundant snaking between each tiers, while few 3D connections leads to small wirelength savings. Therefore, we sweep the size of partitioning bin, and find the best partitioning bin size per benchmark for the maximum power saving.

After tier partitioning, we plan the proper MIV location by using commercial tool built for 2D ICs as proposed in [8]. The main idea of this methodology is to let commercial router treat MIVs same as normal vias while there are routing blockages on the area of macros on the top tier to avoid overlap between MIVs and top macros during routing stage. The limitation of this flow is that the direction of metal layers should not be the same between adjacent layers, and that the number of interconnect layers on the bottom tier should be an even number. Since our foundry-grade 7nm bulk FinFET standard cell layout contains MINT layer for internal routing, we

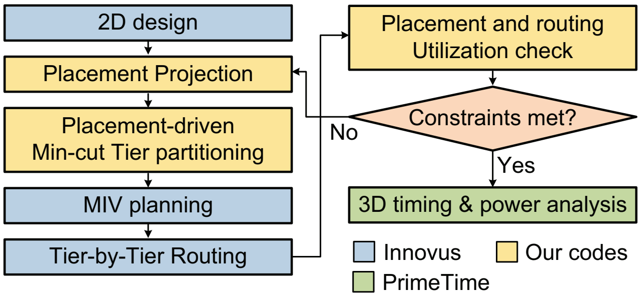
2D design
Placement Projection
Placement-driven Min-cut Tier partitioning
MIV planning
Tier-by-Tier Routing
Placement and routing Utilization check
Constraints met?
No
Yes
3D timing & power analysis
Innovus
Our codes
PrimeTime

**Figure 1: Projected 2D design flow.**

assume an odd number of interconnect layers on the bottom 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. Then we route them initially, 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.

## 3.3 Footprint Resizing

Once initial M3D design is done, we check if the maximum placement or routing utilization on each tier is over 70%. Since M3D placement utilization on each tier is same as 2D placement utilization considering balanced area skew from placement-driven min-cut partitioning, meeting the sufficient placement utilization is guaranteed from 2D design result. However, if a circuit is BEOL-dominant type, then 2D placement utilization is possible to be lower than 70% because insufficient routing resources requires large die area. In that case, even though 2D routing utilization is over 70% in certain metal layers, routing utilization in M3D could be lower than 70% due to the wirelength reduction. To maximize the utilization of die area, we estimate proper footprint as $A'_D = A_D \times U_r/0.7$, where $U_r$ is the maximum routing utilization out of all metal layers, $A_D$ is the current footprint area, and $A'_D$ is the updated footprint area. We project the 2D placement into the updated footprint, and iterate the design flow shown in Figure 1 until the $U_r$ is over 70%.

# 4. EXPERIMENTAL RESULTS

We choose Triple-Data-Encryption-Standard cipher (DES3) and Low-Density Parity-Check decoder (LDPC) from OpenCore benchmark suites to cover two widely different circuit types. 2D Design of these two benchmarks built using a foundry-grade 7nm bulk FinFET PDK shows that 72% of total capacitance in DES3 is pin capacitance while 64% of total capacitance in LDPC is wire capacitance. Also, average net length of LDPC is 3.94 times longer than that of DES3. Therefore, we define LDPC as a BEOL-dominant circuit, and DES3 as a FEOL-dominant circuit. The diameter and pitch of an MIV in the experiments is assumed to be 24nm and 48nm with resistance of 16Ω and capacitance of $0.01fF$.

## 4.1 2D Design Results

Table 4 shows the impact of changing metal stack configuration on the design result of FEOL-dominant circuit DES3, and BEOL-dominant circuit LDPC. Designs for each benchmark are constrained with the same clock period, (0.5ns for DES3, 1.0ns for LDPC). Total power in the Table 4 is iso-performance power number, and the maximum performance is calculated by reversing the sum of clock period and the worst timing slack. PPC is calculated as follows:

$$PPC = \frac{Max\_Performance}{Total\_Power \times Die\_Cost} \tag{9}$$

Since wafer and die cost is normalized with that of 8 BEOL metal stack (M8 in Table 4) design, PPC is also normalized with the PPC value of M8 design.

### *4.1.1 FEOL-Dominant Circuit Type*

Starting from M8 design, reducing metal layers in FEOL-dominant circuit has little impact on routing utilization overhead. Since most of nets in DES3 is locally routed, maximum routing utilization is only 20.7% in M3 layer even though there are 8 BEOL metal layers for routing. The placement utilization and die area are also unchanged along with metal stack reduction since M3 design already has sufficient routing resources. All designs close the timing, and small change in iso-performance power along with metal stack reduction is caused by slightly increased routing congestion. Even though the total power in M3 design is increased by 4% compared to M8 design, wafer and die costs are reduced by 26%. Therefore, overall PPC saving of M3 design is 31% more than the saving of M8 design, and we define M3 design as the most optimized design for DES3 in terms of PPC.

### *4.1.2 BEOL-Dominant Circuit Type*

BEOL-dominant circuit LDPC shows interesting results in Table 4. In M5 design, the die area is determined by the maximum routing utilization in M4 layer. The lack of routing resources increase chip size even though placement utilization is only 35.9%. The large footprint not only increases die cost, but also makes overall wirelength longer and leads to higher wire capacitance. Therefore, adding only one more metal layer significantly improves the design quality of BEOL-dominant circuit. Compared to M5 design result, M6 design has total power saving by 15%, area reduction by 39%, lower die cost by 36%, and PPC improvement by 85%.

Once we have enough metal stack for the routing in LDPC, the die area needs to be determined by both placement and routing utilization. Therefore, area saving and the impact of adding more interconnect layers become saturated as shown in M8 design. As a result, reduced power saving and additional cost for more metal layer have a tradeoff relationship.

## 4.2 Impact of Metal Stack Optimization

Optimizing dielectric constant, and conductivity in the metal stack by changing material composition is one of the cheapest solutions to improve design quality. We assume that the dielectric constant of global interconnect layers (from M6 to M10) has been reduced by 14%, and generate new technology file (TCH) using Cadence Techgen. Scaling dielectric constant reduces 12% of total capacitance per unit length for the global interconnect metal layers, and we call this metal stack configuration as Low-K metal stack. We also consider the wafer cost change for the Low-K metal stack. Based on the wafer cost model in Section 2, we increase the BEOL cost from 0.70 to 0.71 and takes it into account for the PPC calculation.

Table 5 shows the impact of Low-K metal stack on the BEOL-dominant LDPC 2D designs. When we compare M5 and M5 + Low-K designs, reduced wire capacitance by using Low-K metal stack further improves total power due to the switching power saving. Also, decreased routing congestion from the reduced number and drive strength of buffers make room for die area saving. Since we assume that BEOL cost for Low-K metal stack is different from the normal metal stack, it shows different tradeoff between power saving and wafer cost increase. Even though M9 + Low-K design has more power saving than M8 + Low-K design, the PPC value of

**Table 4: 2D IC PPC analysis and comparisons. Our PPC is defined in Equation 9.**

| Circuit Type | Metal Stack | Tot. Power ($mW$) | Max. Perf ($GHz$) | Placement Utilization | Max. Routing Utilization | Wafer Cost | Area ($\mu m^2$) | DPW (1e+6) | Die Yield | Die Cost | PPC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FEOL dominant DES3 | M3 | 37.70 | 2.00 | 0.719 | M2, 0.287 | 0.739 | 6048 | 11.679 | 0.949 | 0.739 | **1.306 (best)** |
| | M4 | 36.96 | 2.00 | 0.718 | M3, 0.242 | 0.804 | | | | 0.804 | 1.224 |
| | M5 | 36.52 | 1.99 | 0.716 | M3, 0.215 | 0.870 | | | | 0.870 | 1.140 |
| | M6 | 36.69 | 2.00 | 0.716 | M3, 0.213 | 0.913 | | | | 0.913 | 1.086 |
| | M7 | 36.39 | 1.99 | 0.716 | M3, 0.214 | 0.957 | | | | 0.957 | 1.040 |
| | M8 | 36.21 | 1.99 | 0.715 | M3, 0.207 | 1.000 | | | | 1.000 | 1.000 |
| BEOL dominant LDPC | M5 | 39.28 | 0.99 | 0.359 | M4, 0.824 | 0.870 | 10816 | 6.529 | 0.948 | 1.720 | 0.433 |
| | M6 | 33.45 | 0.99 | 0.581 | M6, 0.807 | 0.913 | 6561 | 10.765 | 0.949 | 1.094 | 0.799 |
| | M7 | 31.49 | 0.99 | 0.686 | M6, 0.790 | 0.957 | 5476 | 12.899 | 0.949 | 0.957 | 0.972 |
| | M8 | 29.28 | 0.99 | 0.794 | M8, 0.613 | 1.000 | 5476 | 12.899 | 0.949 | 1.000 | 1.000 |
| | M9 | 28.39 | 0.99 | 0.787 | M8, 0.678 | 1.043 | 4692 | 15.055 | 0.949 | 0.894 | 1.154 |
| | M10 | 27.48 | 1.00 | 0.789 | M4, 0.535 | 1.087 | 4692 | 15.055 | 0.949 | 0.931 | **1.156 (best)** |

**Table 5: Impact of Low-K metal stack on BEOL-dominant LDPC 2D designs.**

| Metal Stack | Tot. Power ($mW$) | Max. Perf ($GHz$) | Wafer Cost | Area ($\mu m^2$) | Die Cost | PPC |
|---|---|---|---|---|---|---|
| M5 | 39.28 | 0.99 | 0.870 | 10816 | 1.720 | 0.433 |
| M5 + Low-K | 37.27 | 0.99 | 0.878 | 8190 | 1.314 | 0.598 |
| M6 | 33.45 | 0.99 | 0.913 | 6561 | 1.094 | 0.799 |
| M6 + Low-K | 32.4 | 0.99 | 0.922 | 6561 | 1.105 | 0.818 |
| M7 | 31.49 | 0.99 | 0.957 | 5476 | 0.957 | 0.972 |
| M7 + Low-K | 30.72 | 0.99 | 0.966 | 5476 | 0.966 | 0.987 |
| M8 | 29.28 | 0.99 | 1.000 | 5476 | 1.000 | 1.000 |
| M8 + Low-K | 28.35 | 0.99 | 1.010 | 4692 | 0.865 | **1.194** |
| M9 | 28.39 | 0.99 | 1.043 | 4692 | 0.894 | 1.154 |
| M9 + Low-K | 27.56 | 1.00 | 1.054 | 4692 | 0.903 | 1.188 |
| M10 | 27.48 | 1.00 | 1.087 | 4692 | 0.931 | 1.156 |

M8 + Low-K is higher than M9 + Low-K due to the BEOL cost. Overall, we define M8 + Low-K design as the most optimized design for LDPC with regard to PPC.

For the FEOL-dominant DES3 design, the impact of reducing wire capacitance on PPC by using Low-K metal stack is negative since it has little power saving with increased die cost.

## 4.3 M3D Design Results

Table 6 shows the M3D design results using normal metal stack of various combinations. 2D design in Table 6 is the best design with regard to PPC, defined as the reference for the comparison with M3D design. In this section, we assume the variable for the sequential integration and bonding cost for the top tier ($\alpha$) as 0.1, and M3D wafer yield ($\beta$) as 90% of 2D wafer yield.

### *4.3.1 FEOL-Dominant Circuit Type*

While 2D DES3 design with only M3 metal stack already has enough resources to finish the routing, M3D DES3 design should have M5 metal stack in the bottom tier. This is because if M3 metal stack is used in the bottom tier, part of routing resource in M3 layer will be dedicated to inter-tier connection (MIV planning), resulting in compromise of routability, and many DRVs. Also, due to the limitation of MIV planning scheme using commercial 2D router, we need to have odd number of BEOL metal layers on the bottom tier so that top metal layer of the bottom tier and MINT layer of the top tier has routing direction orthogonal to each other. Therefore, we set the 5 metal layers as the minimum metal stack on the bottom tier, and evaluate the PPC benefit of M3D design.

Since the die area of FEOL-dominant circuit is determined by placement utilization, 2-tier M3D DES3 design indeed has 50% of footprint saving compared to 2D design. However, high wafer cost of M3D integration, and the assumptions on reduced M3D wafer yield increase die cost for M3D. In addition, total power saving in M3D is not significantly large, since DES3 is FEOL-dominant and most of routing in DES3 are done locally. Performance loss in DES3 M3D design is worth to notice. Because M3D design keeps the same nets as 2D design through Projected 2D flow, we compare the worst resistance net in M3D design with the equivalent net in 2D design as shown in Table 7.

It shows detailed wirelength distribution and net resistance of the equivalent net in 2D M5 design and M3D M5 / M5 design. The 2D net has long wirelength, but most of routing are done in M5 layer. However, although the M3D net has 22% total wirelength saving, total net resistance is reduced by 9% only. Unit-length resistance of the M3D net is 17% higher than that of the 2D net. Based on the net comparison, we observe that when locally placed and routed cells in 2D design are split into different dies through tier partitioning, routing utilizations for intermediate interconnect layers are increased. Since part of the top metal layer in the bottom tier should be dedicated to MIV planning, commercial router is not able to fully use the top metal routing resource in the bottom tier. Instead, it uses more intermediate interconnect layers. Besides, wires should go through the whole metal stack in the bottom tier to route top tier cells. Therefore, it is likely to increase the routing congestion, and redundant wire capacitance.

Furthermore, top tier routing also uses intermediate interconnect layers since only local routing remains. The resistance of M2,M3 layer is 2.46 times higher that of M4,M5 layer. Therefore, locally routed FEOL-dominant circuit requires more effective tier partitioning, otherwise the timing of the critical path worsens. With regard to PPC value, we define M3D design with M4 / M5 metal stack as the most optimized M3D design for FEOL-dominant DES3.

### *4.3.2 BEOL-Dominant Circuit Type*

If we compare M3D M5 / M5 design with 2D M5 design, BEOL-dominant LDPC M3D design indeed has increases of power saving by 17% and die area saving by 58%. However, in Section 4.2, we set 2D M8 + Low-K design as the reference design for the fair comparison with M3D designs. Since placement and routing utilization of our 2D reference design is highly optimized, die area of 2D design is small enough to provide cheap die cost. Due to the die area as small as 57% of 2D M5 design, huge wirelength and buffer saving result in M3D-compatible power consumption.

Therefore, unlike FEOL-dominant DES3 M3D design, LDPC M3D M5 / M7 design is the best M3D design out of given metal stack combinations with regard to PPC value though it only has 25% area saving compared with 2D reference. Table 8 shows the impact of Low-K metal stack on LDPC M3D design. By using

Table 6: M3D PPC analysis and comparison. Our PPC is defined in Equation 9.

| Circuit Type | Design Flavor | Metal Stack (top / bottom) | Tot. Power ($mW$) | Max Perf ($GHz$) | Placement Utilization (top / bottom) | Maximum Routing Utilization (top / bottom) | Wafer Cost | Area ($\mu m^2$) | DPW (1e+6) | Die Yield | Die Cost | PPC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FEOL dominant DES3 | 2D | M3 | 37.7 | 2 | 0.719 | M2, 0.287 | 0.739 | 6048 | 11.679 | 0.949 | 0.739 | 1.306 |
| | M3D | M3 / M5 | 37.38 | 1.776 | 0.744 / 0.718 | M2, 0.278 / M4, 0.215 | 1.826 | 3041 | 23.232 | 0.854 | 1.019 | 0.848 |
| | | M4 / M5 | 36.83 | 1.901 | 0.745 / 0.718 | M3, 0.203 / M4, 0.215 | 1.872 | | | 0.854 | 1.045 | **0.899 (best)** |
| | | M5 / M5 | 36.74 | 1.901 | 0.745 / 0.718 | M3, 0.187 / M4, 0.215 | 1.917 | | | 0.854 | 1.070 | 0.880 |
| | | M6 / M5 | 36.74 | 1.898 | 0.745 / 0.718 | M3, 0.187 / M4, 0.215 | 1.948 | | | 0.854 | 1.087 | 0.864 |
| BEOL dominant LDPC | 2D | M8 + Low-K | 28.35 | 0.99 | 0.794 | M8 0.713 | 1.010 | 4692 | 15.055 | 0.949 | 0.865 | 1.194 |
| | M3D | M5 / M5 | 32.76 | 1.060 | 0.481 / 0.425 | M4, 0.762 / M4, 0.639 | 1.917 | 4499 | 15.702 | 0.854 | 1.750 | 0.547 |
| | | M6 / M5 | 32.55 | 1.050 | 0.563 / 0.491 | M6, 0.666 / M4, 0.679 | 1.948 | 3894 | 18.142 | 0.854 | 1.538 | 0.620 |
| | | M7 / M5 | 32.37 | 1.018 | 0.563 / 0.491 | M6, 0.631 / M4, 0.694 | 1.978 | 3894 | 18.142 | 0.854 | 1.562 | 0.596 |
| | | M5 / M7 | 28.5 | 1.035 | 0.606 / 0.528 | M4, 0.756 / M4, 0.545 | 1.978 | 3504 | 20.162 | 0.854 | 1.406 | **0.764 (best)** |

**Table 7: Equivalent net comparison between M3D and 2D design. The worst resistance net in DES3 M3D design is analyzed.**

| Wirelength distribution (um) | 2D | M3D (top/bottom) |
|---|---|---|
| M5 | 122.35 | 0.00 / 74.90 |
| M4 | 67.09 | 7.42 / 51.74 |
| M3 | 0.32 | 5.87 / 3.78 |
| M2 | 0.27 | 2.46 / 0.19 |
| M1 | 0.46 | 1.50 / 0.46 |
| Net Total Wirelength ($\mu m$) | 190.50 | 148.05 |
| Net Total Resistance ($\Omega$) | 11187 | 10206 |
| Unit-length Resistance ($\Omega/\mu m$) | 58.72 | 68.94 |

**Table 8: Impact of Low-K metal stack on BEOL-dominant LDPC M3D designs.**

| Metal Stack (top / bottom) | Tot. Power (mW) | Max. Perf (GHz) | Wafer Cost | Area (um) | Die Cost | PPC |
|---|---|---|---|---|---|---|
| 2D | | | | | | |
| M8 + Low-K | 28.35 | 0.99 | 1.010 | 4692 | 0.865 | 1.194 |
| M3D | | | | | | |
| M5 / M5 | 32.76 | 1.060 | 1.917 | 4499 | 1.750 | 0.547 |
| M5 / M5 + Low-K | 32.12 | 1.074 | 1.929 | 4499 | 1.760 | 0.562 |
| M6 / M5 | 32.55 | 1.050 | 1.948 | 3894 | 1.538 | 0.620 |
| M6 / M5 + Low-K | 31.9 | 1.071 | 1.960 | 3894 | 1.548 | 0.641 |
| M7 / M5 | 32.37 | 1.018 | 1.978 | 3894 | 1.562 | 0.596 |
| M7 / M5 + Low-K | 31.72 | 1.031 | 1.991 | 3894 | 1.572 | 0.611 |
| M5 / M7 | 28.5 | 1.035 | 1.978 | 3504 | 1.406 | 0.764 |
| M5 / M7 + Low-K | 27.91 | 1.050 | 1.991 | 3504 | 1.414 | **0.787** |

Low-K metal stack, M5 / M7 + Low-K design finally beats 2D reference in terms of both total power and maximum performance. Even though it is clear that using Low-K metal stack and adding routing resources are very effective solutions to improve M3D design quality, too much expensive metal stack for BEOL-dominant circuit increases the wafer cost almost 2 times higher than 2D reference, resulting in lower PPC of M3D than that of 2D.

## 5. 7NM M3D COST AND YIELD STUDY

In Section 4.3.1 and 4.3.2, assuming M3D wafer yield ($\beta$) as 90% of 2D wafer yield, and additional cost for top tier device implementation ($\alpha$) is 10% of wafer cost for 2D M8 design, we observe that PPC of FEOL-dominant DES3 M3D design is worse by 31% and BEOL-dominant LDPC M3D design by 34% compared to 2D reference. Then the next question is how much M3D wafer yield and additional cost for M3D integration should be further reduced for the cheap M3D die cost to justify the adoption of M3D technology. In Figure 2, red surface of each plot shows the valid region along with $\alpha$, and $\beta$ where the best M3D design defined in the previous Sections beats PPC of the 2D reference. Z-axis of these plots is calculated by the ratio of PPC value between M3D and 2D design. We observe that for the adoption of gate-level M3D integration, M3D wafer yield needs to be higher than 90% of 2D wafer yield, and the device manufacturing cost of M3D design should be limited by less than 33% of 2D device manufacturing cost.

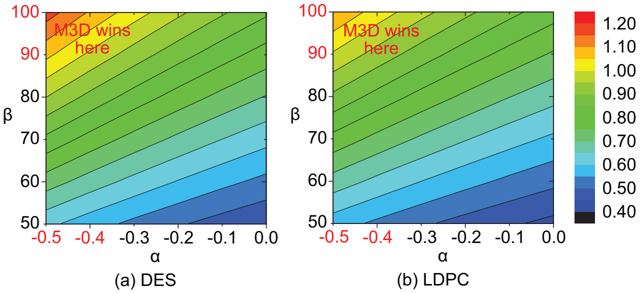

**Figure 2: M3D cost vs. yield vs. PPC sensitivity analysis. $\alpha$ denotes cost variable for top-tier devices fabriacation and bonding in M3D, e.g., $\alpha$ = -0.4 means that FEOL manufacturing cost for M3D (0.6) should be 67% lower (0.6 + $\alpha$ = 0.2). $\beta$ denotes M3D wafer yield (percentage w.r.t. 2D wafer yield). Z-axis denotes PPC ratio of M3D over 2D, e.g., 1.2 means M3D PPC is 20% better.**

Moreover, the experiment result show that FEOL-dominant circuit type has more room for the adoption of M3D, and benefits more from M3D integration than BEOL-dominant circuit type in terms of PPC. This is because the impact of metal stack optimization and giving more routing resources to BEOL-dominant type circuit drastically reduce both power and die area of 2D design compatible to M3D counterpart. The differences in total power and die area between LDPC 2D reference (M8 + Low-K design) and M3D design with the best PPC (M5 / M7 + Low-K design) are only 2% and 25%. However, since the die area of FEOL-dominant circuit type is determined by placement utilization, 50% of footprint saving from M3D technology is guaranteed, resulting in more spaces in terms of die cost for adoption of M3D technology.

Two benchmarks for the previous experiments, DES3 and LDPC, are logic circuits where the number of standard cells in the full-chip 2D design is less than 60$k$ based on foundry-grade 7nm bulk FinFET. The chip area of these two small circuits is less than 0.01$mm^2$. Since the 2D die yield of those extremely small benchmarks is already sufficient, it explains why the huge footprint saving and die cost benefit from M3D technology does not show up. Therefore, we evaluate the impact of die area of logic-only design on the die cost of M3D and 2D design based on the cost models proposed in

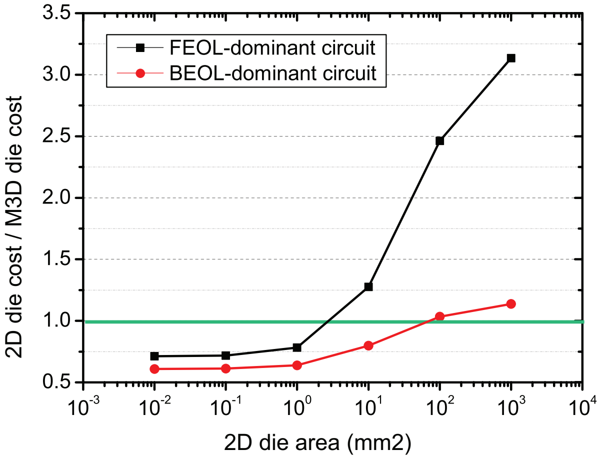

**Figure 3: Die size impact on the die cost ratio between 2D and M3D. Two different circuit type (FEOL-dominant and BEOL-dominant) are investigated. The region above the green line indicates where the M3D die cost is cheaper than 2D die cost.**

Section 2. Since the 2D die area of BEOL-dominant circuit is effectively reduced when we use more routing resources, footprint saving of gate-level 2-tier M3D design is only 25% as shown in Section 4.3.2. When the die area is determined by placement utilization like FEOL-dominant circuit, 50% of M3D area saving is guaranteed as analyzed in Section 4.3.1.

We assume that the ratio of the die area of 2D design and that of M3D design is fixed in each circuit type, and calculate die cost for each design scheme considering die yield. Figure 3 shows that M3D die cost becomes cheaper than 2D die cost along with the increase in die size. M3D design of FEOL-dominant circuit has significant die cost saving compared to 2D design starting from $2mm^2$ while M3D design of BEOL-dominant circuit becomes cheaper from $70mm^2$ as well. In addition, with the same die size of design for two circuit types, the gap for the ratio between 2D and M3D die cost of FEOL-dominant and BEOL-dominant circuit becomes wider along with die size increase. Assuming $100mm^2$ of 2D die size, FEOL-dominant circuit has 2.5 times more cost competitiveness from M3D technology than BEOL-dominant circuit. The result indicates FEOL-dominant circuit benefits sooner and more from M3D technology in terms of cost than BEOL-dominant circuit.

## 6. CONCLUSIONS

In this paper we study power, performance, and cost (PPC) tradeoffs with full-chip GDS based cost modeling for 2-tier, gate-level, full-chip GDS monolithic 3D ICs (M3D) built using a foundry-grade 7nm bulk FinFET technology. We propose normalized wafer and die cost models based on the number of metal stacks and die area for 2D and M3D. In our PPC tradeoff study with the simple but self-contained cost models, both 2D and M3D designs are optimized in terms of the number of BEOL metal layers used for routing to obtain the best possible PPC values for the fair comparison. Also, we develop a new CAD methodology for 2-tier gate-level M3D, named Projected 2D Flow, that maximizes the placement and routing utilization of M3D design by reducing its footprint by more than 50% compared with that of 2D. Furthermore, this flow allows us to accurately compare RC parasitics of equivalent nets in both 2D and M3D designs since final netlists of these two design flavors are the same.

Based on the experiments with two widely different circuit types (BEOL-dominant vs. FEOL-dominant), we confirm that while M3D has indeed a great footprint saving, the PPC quality of M3D is actually worse than that of optimized 2D reference by 34% due to high M3D wafer cost. Our study also shows that, for the adoption of M3D technology at the 7nm era, M3D wafer yield needs to be higher than 90% of 2D wafer yield, and the 2-tier device manufacturing cost of M3D design needs to be limited by less than 33% of 2D device manufacturing cost, and lastly the die area should be large enough ($100mm^2$-scale) to have fruitful die cost reduction from huge M3D footprint saving. Lastly, and counter-intuitively, this study shows that FEOL-dominant type circuit has PPC benefits from M3D technology more and sooner than BEOL-dominant type circuit.

## 7. ACKNOWLEDGEMENT

The authors would like to thank Arindam Mallik at IMEC, Belgium for helpful comments on the cost models presented in this paper.

## 8. REFERENCES

[1] W.-T. J. Chan, S. Nath, A. B. Kahng, Y. Du, and K. Samadi, “3DIC Benefit Estimation and Implementation Guidance from 2DIC Implementation,” in *Proc. ACM Design Automation Conf.*, 2015.

[2] Y.-J. Lee, D. Limbrick, and S. K. Lim, “Power benefit study for ultra-high density transistor-level monolithic 3D ICs,” in *Proc. ACM Design Automation Conf.*, 2013, pp. 1–10.

[3] S. Panth, K. Samadi, Y. Du, and S. K. Lim, “Design and CAD methodologies for low power gate-level monolithic 3D ICs,” in *Proc. Int. Symp. on Low Power Electronics and Design*, 2014, pp. 171–176.

[4] ——, “Power-performance study of block-level monolithic 3D-ICs considering inter-tier performance variations,” in *Proc. ACM Design Automation Conf.*, 2014, pp. 1–6.

[5] K. Chang, K. Acharya, S. Sinha, B. Cline, G. Yeric, and S. K. Lim, “Power benefit study of monolithic 3D IC at the 7nm technology node,” in *Proc. Int. Symp. on Low Power Electronics and Design*, 2015, pp. 201–206.

[6] D. K. Nayak, S. Banna, S. K. Samal, and S. K. Lim, “Power, performance, and cost comparisons of monolithic 3D ICs and TSV-based 3D ICs,” in *SOI-3D-Subthreshold Microelectronics Technology Unified Conference*, 2015, pp. 1–2.

[7] N. L. Penmetsa, C. Sotiriou, and S. K. Lim, “Low Power Monolithic 3D IC Design of Asynchronous AES Core,” 2015, pp. 93–99.

[8] S. Panth, K. Samadi, Y. Du, and S. K. Lim, “Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs,” *IEEE Trans. on Computer-Aided Design of Int. Circuits and Systems*, vol. 34, no. 4, pp. 540–553, 2015.

[9] A. Mallik, J. Ryckaert, A. Mercha, D. Verkest, K. Ronse, and A. Thean, “Maintaining Moore's law: enabling cost-friendly dimensional scaling,” vol. 9422, 2015, pp. 94 221N–94 221N–12.

[10] X. Dong, J. Zhao, and Y. Xie, “Fabrication Cost Analysis and Cost-Aware Design Space Exploration for 3-D ICs,” vol. 29, no. 12, pp. 1959–1972, 2010.

[11] Q. Zou, J. Xie, and Y. Xie, “Cost-driven 3D design optimization with metal layer reduction technique,” in *Proc. Int. Symp. on Quality Electronic Design*, 2013, pp. 294–299.

[12] A. Mallik, N. Horiguchi, J. BÃűmmels, A. Thean, K. Barla, G. Vandenberghe, K. Ronse, J. Ryckaert, A. Mercha, L. Altimime, D. Verkest, and A. Steegen, “The economic impact of EUV lithography on critical process modules,” in *Proc. SPIE*, vol. 9048, 2014, pp. 90 481R–90 481R–12.

[13] A. Mallik, W. Vansumere, J. Ryckaert, A. Mercha, N. Horiguchi, S. Demuynck, J. BÃűmmels, T. Zsolt, G. Vandenberghe, K. Ronse, A. Thean, D. Verkest, H. Lebon, and A. Steegen, “The need for EUV lithography at advanced technology for sustainable wafer cost,” vol. 8679, 2013, pp. 86 792Y–86 792Y–10.