Metal Stack and Partitioning Exploration for Monolithic 3D ICs

Metal Stack and Partitioning Exploration for Monolithic 3D ICs

Nikolaos K. Sketopoulos*, Christos P. Sotiriou* and Vasilis F. Pavlidis†‡
*Dept. of Electrical & Computer Engineering, University of Thessaly, Volos, Greece
sketopou@eece.uth.gr, chsotiriou@eece.uth.gr
†School of Computer Science, The University of Manchester, Manchester, UK
‡Dept. of Electrical & Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece
†pavlidis@cs.man.ac.uk, ‡vpavlid@ece.auth.gr

***Abstract*—In this work, we investigate the effect of metal stack and tier 3D IC partitioning methodologies on the Quality of Results (QoR) of monolithic 3D circuits compared to their 2D counterparts. Two interconnect options are considered. For the interconnect option, termed Single, a single metal stack is used where cell pins lie on two lower metal layers. A Face to Face (F2F) interconnect option is also considered where cell pins lie on symmetrical lower and upper metal layers in two different tiers. In addition, two 3D circuit partitioning methodologies are investigated including the Greedy Bin-Based Fidducia-Mattheyses (GBBFM) and a Displacement-based 3D Legaliser (3DLG). For both the 2D and 3D circuits, a multi-pass timing-driven In-Place Optimisation (IPO) is performed with an industrial P&R tool to extract the best QoR. For the 3D circuits, the IPO is applied after tier partitioning. The 45 nm nanoCAS library, a 3D library based on NANGATE 45 nm, is utilized and three typical benchmark circuits are analysed. The performance of 3D circuits is improved between 1% to 3%, total wirelength is significantly reduced, via usage is increased, yet the estimated power and cell area do not necessarily decrease. Single metal stack overall demonstrates better QoR than F2F integration.**

I. Introduction

The IC industry has entered the so called "More-than-Moore" era. Obtaining "More-than-Moore" scaling with considerable performance and power gains requires new design paradigms. 3D integration is a promising technology that aims to improve circuit packing density, reduce the area spanned by global interconnects, as well as global Wire Length (WL). The area spanned by 3D circuits is typically less than the respective 2D circuits, with the former composed of a number of vertically stacked transistor tiers. 3D wafer-level packaging, 2.5D and 3D interposer-based integration, 3D stacked ICs, monolithic 3D ICs, *etc.* represent different types of 3D integration. In this work, we focus on monolithic ICs. Rather than stacking multiple wafers or dies, monolithic 3D ICs follow a sequential process, where each device and interconnect layers are formed on top of the other. Several published works compare the benefits of monolithic ICs against conventional 2D ICs. As noted in [1], [2], the most commonly used metrics for 2D vs 3D, are the area footprint and WL. The area depends on the number of tiers while the WL on several parameters, such as circuit connectivity, number of tiers, *etc.* In general, WL reduction enables higher performance, as well as lower power. Thus, in theory, by merely partitioning the circuit into several tiers and connecting these tiers together, 3D circuits can exhibit superior Power, Performance and Area (PPA). However, there is a lack of 3D Electronic Design Automation (EDA) tools to support 3D ICs, and properly exploit these benefits [2]. Prior art (*e.g.* [1], [2], [3], [4], [5]) utilizes conventional 2D EDA tools, within a 3D flow, for implementing 3D ICs. Based on these tools, these works place and route (P&R) the circuits in several tiers (typically two) and aim to optimise Power, Performance, Area (PPA), connectivity between tiers, and thermal issues.

3D flows usually begin with tier partitioning. Post-partitioning, timing-driven optimisations must then be performed on the partitioned circuit. Several 3D circuits partition techniques commence with a Shrunk 2D placement and perform post Shrunk 2D partitioning across typically two tiers [6], [7], [3], thereby being 2D placement aware. Shrunk 2D placement places shrunk (minimum width) cells in half of the original 2D area, *i.e.* a multiplication factor of $\sqrt{2}$ for width and height. Post Shrunk 2D placement, Greedy Bin-Based FM (GBBFM) [6], [8] or 3D Legalisation (3DLG) [3] may be used for partition. After the partition step, timing driven optimisation is performed by running IPO, with the existing 2D tools, to improve Worst Negative Slack (WNS), Total Negative Slack (TNS), and recover area. The effect of the metal stack used is key, as it affects WL, number of vias and wire *RC* delay. Two commonly used metal stack styles include Single, where the Back-End-Of-the-Line (BEOL) is fabricated on the top tier and top tier cell pins are at a local metal layer, *e.g.* Metal3, and F2F, where two, flipped metal stacks (BEOL) are symmetrically fabricated on top of each device layer.

In this work, we implement and investigate the 3D QoR benefits, for these two metal stack styles, by utilising two tier partitioning methodologies. Moreover, we show that a multi-pass, tier by tier, IPO is necessary, and effectively improves circuit PPA results, as a single IPO does not always produce best results, and may even worsen specific design metrics.

A. 3DIC Tier Partitioning

Several algorithms for graph partitioning exist. Recent 3D partition methods [3], [8] use the Fidducia-Mattheyses (FM) algorithm [9], for assigning components onto the circuit tiers. As FM does not scale gracefully with circuit size and is slow even for small circuits (>200k cells), a Bin-Based FM (BBFM) approach is used [8]. Furthermore, considering the global net connectivity, when optimising the bin cutsize (all tier to tier connections) using BBFM is also prohibitively expensive. However, considering only local bin connections (greedily) for FM cutsize optimisation, partitioning is performed significantly faster. Therefore, most approaches are based on the Greedy BBFM (GBBFM) approach. An alternative 3D partition approach has been proposed in [3], [10]. A 3D Legalisation Algorithm (3DLG) is utilized, where the cutsize may or may not be constrained. 3DLG assigns components across the multiple available tiers, mainly aiming to minimise displacement from the initial placement positions. The unconstrained version does not consider cutsize at all and solely minimises displacement.

In this work, we contrast the GBBFM and Unconstrained 3DLG partitioning approaches as we target Monolithic 3D (M3D) circuits. Consequently, the cutsize is proportional to the number of Monolithic Interlayer Vias (MIVs). In monolithic 3D, MIVs are as small as regular vias, 50 nm [4], thus their number does not affect PPA results.

B. 3D Metal Stacks

In M3D, multiple silicon device tiers are grown on top of each other. M3D integration offer two orders of magnitude, higher vertical interconnect density at a much lower footprint area. Therefore, the same number of pins as with a 2D circuit must now be routed within half the wiring area. This situation considerably increases routing congestion, particularly, for the lower layers of the BEOL process. The severity of routing congestion is also a function of the interconnect metal stack (or technology) utilised in the manufacturing process.

To investigate the correlation between congestion and interconnect stack, we use the open-source 45nm nanoCAS library [11]. This is a M3D library, providing cells for two tier vertical integration. Figure 1 illustrates the library metal stack and routing layers. We call this metal stack Single, as the same metal layers are used for both device tiers.

The grey rectangles represent the individual metal layers, *e.g. m1*, *m2*, ..., *m12*, with the rectangle height drawn proportionally to the respective layer resistance, *i.e.* less resistive metals are taller. The blue cylinders illustrate layer to layer vias, with cylinder height again inversely proportional to via resistance. The red spots, indicate the layer where the cell pin connects. Thus, in the 45 nm nanoCAS library [11], bottom tier cells pins are allocated to the first metal layer, whereas the top tier cell pins to the third metal layer. This layer pin assignment may lead to high routing congestion areas, as the two pin assignment metal layers are rather adjacent to each other. This assignment will force the router to over-utilise the lower metal layers as these are used both for intra- and inter-tier connections. In order to measure and compare the effect of congestion in the lower metal layers, we have created a new, alternative pin layer assignment for the 45 nm library, presented in Figure 1. Our alternative metal stack pin assignment contains the same number of layers but tiers are integrated Face-to-Face (F2F), meaning that bottom tier cell pins are at metal 1, top tier cell pin at metal 12.

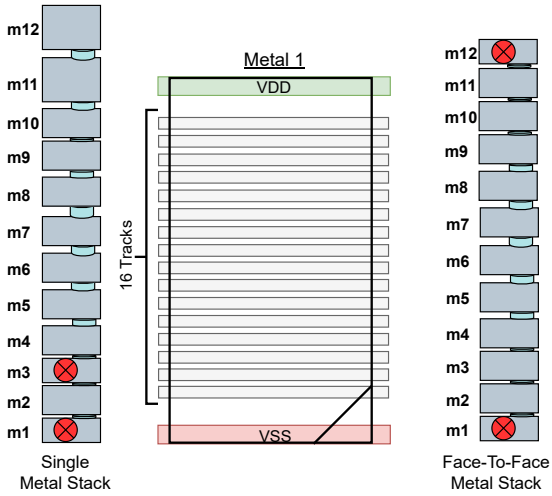


Fig. 1: Metal Stacks and Tracks

II. 3DIC Design Flow

In this section, we present the design flows used to compare the two different interconnect options, while utilising two 3D partition methodologies, the Greedy Bin-Based FM partitioning (GBBFM) and the 3D Legaliser (3DLG). We use the 2D benchamrk circuits as baseline with the 2D circuits undergoing the exact same process for timing optimisation.

Figure 2 shows the steps of our 3D IC flow in detail. These consist of three main parts, the initialisation step, where the 2D design is prepared for 3D tier partitioning, the tier assignment step where cells are assigned onto the bottom or top tier, and the timing optimisation step where the multiple tier netlist is PPA-optimised. We describe these steps in detail below.

A. Initialisation Step

To perform a fair comparison for 2D and 3D circuits, the 3D flow starts from the same netlist as the 2D flow. A modified LEF file with shrunk cell dimensions is used. This LEF contains the same cells as in 2D flow, but the width and height of the cells are modified to be equal to the minimum allowable placement dimensions of the LEF, *i.e.* the site width and height, respectively. This is necessary to allow the conventional, industrial P&R tool to place components into half the original 2D footprint area, otherwise utilisation would be greater than 100%. Thus, by using the shrunk cells LEF, we can apply a standard industrial P&R flow into the half area of the 2D Flow. After loading the netlist and the LEF file, we perform floorplanning, placement and IPO into half of the original 2D area. For timing optimisation, we use a multiple IPO strategy. Instead of performing a single IPO

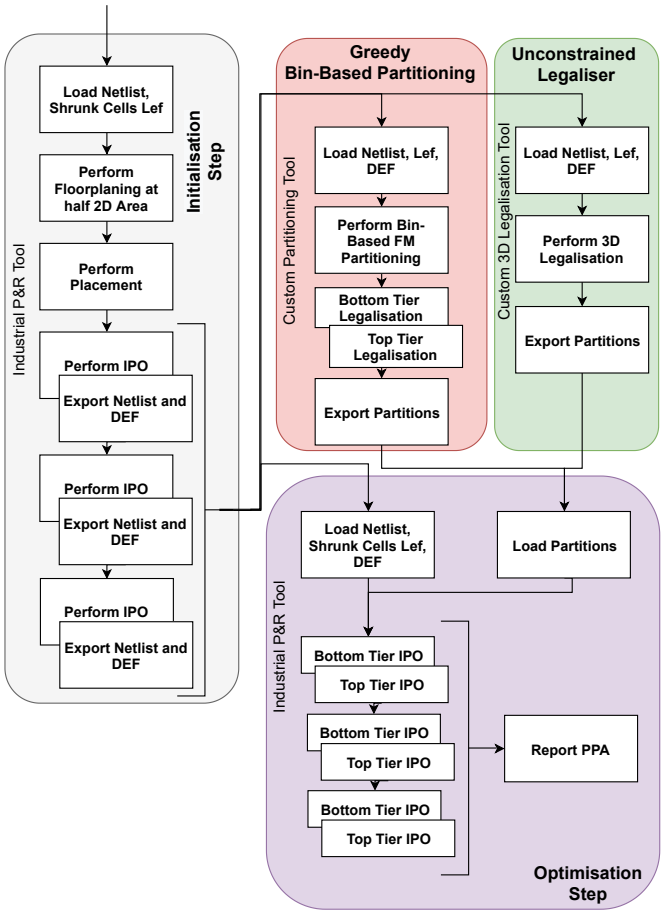


Fig. 2: 3D Experimental Flow

run, we perform three IPO iterations, as shown in Figure 3, with the goal to determine and use the best overall IPO result. The reason behind the practice is the fact that the IPO process is path based and exhibits randomness, meaning that fixing the timing of some critical paths, disturbs the timing of others, or even increases the cell area. In addition, the IPO algorithm sometimes terminates prematurely, returning an arbitrary solution, which may be worse than the solution prior to the IPO step. These points justify running multiple and successive IPOs. Having optimised the Shrunk 2D design, we extract the new timing optimised netlists, *i.e.* three of them, are written out as a DEF file, for the next step, tier partitioning.

B. Tier Assignment Step

In this step, cells are assigned to tiers, using either the GBBFM or the 3DLG tier assignment approach. Also, cell sizes are restored, *i.e.* the original LEF, not the shrunk size is used. Each partitioning approach optimises different metrics. GBBFM minimises grid cutsize, *i.e.* number of MIVs, by assigning strongly connected components of the same grid into the same tier. On the other hand, 3DLG minimises component displacement from their original placement position. In our flow, the initial positions of the components, for the assignment step (*i.e.* the partitioning), correspond to the post IPO positions. In this way, cells are legalised as close as possible to their original, considered optimal, positions and the 3DLG assigns them into the tier, where the minimum perturbation occurs. Both GBBFM and 3DLG utilise the original cell dimensions in order to utilise with the actual cell area. This is because both algorithms fully support 3D tiers and are cell area aware.

C. Optimisation Step

After assigning cells into tiers, timing-driven optimisation must be performed to each tier, considering the tier assignment changes. Thus, the optimised Shrunk 2D netlist and the shrunk cell LEF are utilised again in this step. Cell tier assignment is used to perform IPO on a tier by tier basis, *i.e.* keeping other tier cells soft fixed (SOFTFIX attribute). This allows the cells of one tier to be fully optimised with the cells of the other tier allowed to only be upsized or downsized. In this way, in our flow, there is no need to add any virtual I/O pins for tier by tier optimisation, as in [8]. We have concluded, based on our experiments that three IPOs are sufficient and further IPO runs do not make significant QoR differences. Figure 3 shows the multi-stage IPO process for the 3D designs in detail.

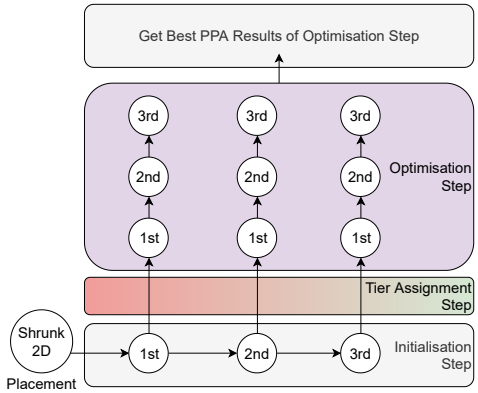


Fig. 3: 3D Multiple IPOs Flow

III. Experimental Setup and Results

In this section, we present key characteristics of the 45nm nanoCAS/NANGATE Library, used for the experimental results. Then we present congestion analysis reports for single and F2F metal stacks, and finally we compare PPA results of the examined metal stacks against 2D.

A. 45nm nanoCAS/NANGATE Library Analysis

Figure 4 contrasts the *RC* wire delay over FO4 (Fanout of 4) gate delay for the 1.1 V, 45 nm nanoCAS library used for the 2D and 3D flows. This library has been generated based on the NANGATE 45 nm library, and has identical metal parameters (LEF file). A 70 μm wire has 2 ps delay, whereas the FO4 delay is equal to the delay of metal wire with 200 μm length. This is a strong indication that wire *RC* delay is significantly smaller than gate delay, thus gate delay dominates. Consequently, Figure 4 illustrates that wire length does not have a significant impact to timing for the

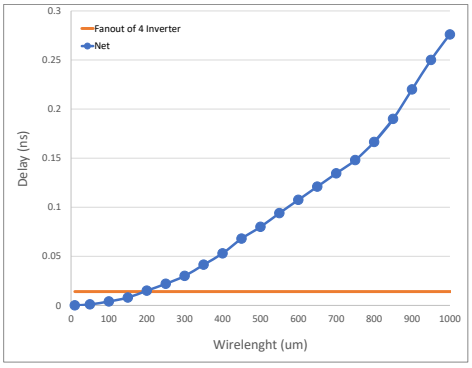


Fig. 4: ASU nanoCAS Library Wire Delay

specific metal stack parameters that this library uses, compared to gate delay. This result is important in the sense that a library with low interconnect delay is expected to produce lower gains for a 3D circuits. In addition, we have also investigated more advanced node libraries, *e.g.* 28 nm and the wire delay does increase but not significantly. Thus, only at 20 nm and below, we can expect a more pronounced wire *RC* delay effect, or at lower voltages. In terms of wire capacity as shown in Figure 1, the standard-cell track height for this library is 16, whereas for 20 nm and below industrial libraries, track heights are typically 12 tracks for high-performance and 9 tracks for low-power. The 16 track height stems from the larger transistor size and reduces wire congestion.

B. Congestion Analysis

Figure 5 depicts congestion analysis, reported by the Cadence Innovus tool [12], of the *ldpc* OpenCores benchmark [13], for the 2D version and the two 3D metal stacks under investigation. We illustrate this benchmark as its high connectivity leads to higher routing congestion. As a result, 2D presents 0.07% over congested layers, while single metal stack presents 3.89% and F2F 0.56%, respectively. The low routing congestion of F2F is because same tier connections use their nearest metals, the upper or lower, and do not interfere much with tier to tier connections. However, for the F2F metal stack, routing interconnects become longer, leading to greater WL. Figure 6 shows the metal stack layout of the *ldpc* benchmark for a single metal stack that is more congested.

C. 2D vs. 3D Benchmarks Analysis

We now present our experimental results and findings, using the 3D flow with the Single and F2F metal stack configurations. We use the flow detailed in Section II. Two partition methodologies are explored in the flow, Greedy Bin-based FM (BBFM) [8] and 3D Unconstrained LG (3DLG) [3]. Compared to [3], we performed multiple IPO (In-Place Optimisation) steps, in the industrial P&R tool, so as to further improve IPO QOR. IPO is a key step for timing optimisation, whereby negative slack paths are optimised by gate resizing (upsizing or downsizing), cell moving for reducing wirelength, buffering to reducing wire delay, or even local logic re-synthesis, *i.e.* changing gate polarities and inversions, preserving Boolean functionality. As for IPO for the 3D circuits, this was achieved by running IPO, in the industrial P&R tool, on a tier by tier basis, *i.e.* bottom tier IPO followed by top tier IPO. During IPO of one tier, the components of the other tiers are soft fixed.

We used a set of three open-source benchmarks from [13] to compare the 2D with 3D flow, including an LDPC (Low-Density Parity Check) controller, a pipelined FFT (Fast Fourier Transform), and an JPEG image encoder. Multiple IPOs are absolutely necessary to satisfy aggressive timing targets. This is illustrated in Table I, where single and multiple IPO 3D flow results are compared.

The circuit with the highest performance gains is *ldpc*, which has a high wire to gate ratio and a large number of top-Level I/Os. The flow with multiple IPOs manages to increase performance, reduce power, and minimise cell area.

Table II contrasts the 2D vs. the 3D flow results, for the different partition algorithms, Greedy Bin-Based FM, 3D Unconstrained Legaliser, and different metal stack types, Single or F2F stacks. The row labelled 'Pre-Partitioning' corresponds to the 2D design with shrunk cells, *i.e.* before tier partitioning. We report this scenario as a baseline to illustrate that despite all cells are shrunk and on one tier, as the same metal track is used for all shrunk cells I/O pins, this may indeed yield worse performance and power results. We do not show cell area results for the 2D circuit with shrunk cells, as it is inconsistent due to the cell shrinking. A key point illustrated by these results is that the timing improvement of 3D circuits illustrates that the tier by tier IPO does achieve effective results.

Table II reports that the design with the most chaotic behaviour is *ldpc*. TNS rather oscillates for different partitioning methodologies and different metal stacks with the sweetspot being a single metal stack and Greedy Bin-Based FM, which demonstrates the highest performance gain of 3.7%. The best performance gains of *fft* and *jpeg* are 2.5% and 2.8%, respectively. These results show that the rather large TWL gains, listed in the respective column do not directly yield performance benefits.

The last two columns of Table II report the number of GCELLS which are over congested, *i.e.* routing grid cells where the router must add tracks (zig-zag), as direct connections produce violations, as well as the average number of vias per net. Despite lower congestion and a lower average number of vias per net typically indicate a superior design, there are cases, *e.g. ldpc*, using 3D Unconstrained Legaliser and Single metal stack, which perform better at TNS. We believe that this behavior is because all critical paths have been fixed by the multiple IPOs and congestion and vias do not impact them.

Table III reports an in-depth Critical Path analysis for the *ldpc* circuit. The top 5 unique, *i.e.* different startpoint, endpoint, critical paths are analysed with respect to their WL, Slack and number of Logic Levels for the 2D and different 3D configurations. Critical path analysis confirms that the large total WL reduction seen in Table II is not that pronounced for critical paths. Moreover, the number of path logic levels

Layer	OverCon #Gcell (1)	OverCon #Gcell (3)	#Gcell OverCon
metal1	0(0.00%)	0(0.00%)	(0.00%)
metal2	2(0.00%)	0(0.00%)	(0.00%)
metal3	2(0.00%)	1(0.00%)	(0.01%)
metal4	0(0.00%)	0(0.00%)	(0.00%)
metal5	1(0.00%)	0(0.00%)	(0.00%)
metal6	2(0.00%)	0(0.00%)	(0.00%)
metal7	0(0.00%)	0(0.00%)	(0.00%)
metal8	1(0.00%)	0(0.00%)	(0.00%)
metal9	171(0.36%)	1(0.00%)	(0.36%)
metal10	99(0.21%)	0(0.00%)	(0.21%)
metal11	0(0.00%)	0(0.00%)	(0.00%)
metal12	0(0.00%)	0(0.00%)	(0.00%)
Total	278(0.07%)	2(0.00%)	(0.07%)

(a) 2D

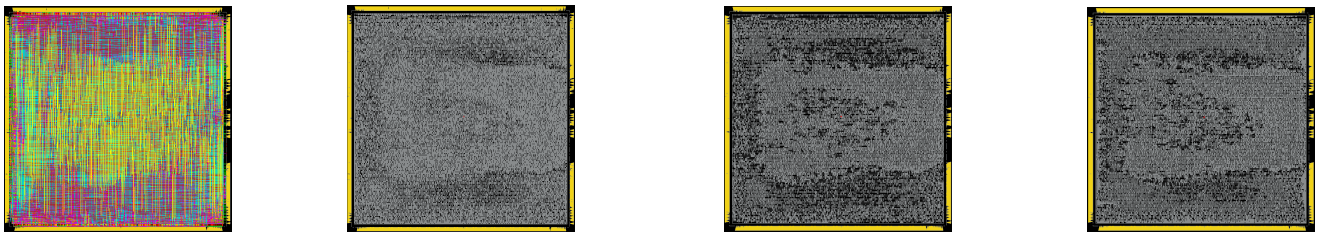
Layer	OverCon #Gcell (1-2)	OverCon #Gcell (3-5)	OverCon #Gcell (6-8)	OverCon #Gcell (9-11)	#Gcell OverCon
metal1	12(1.06%)	0(0.00%)	0(0.00%)	0(0.00%)	(1.06%)
metal2	84(5.38%)	31(1.99%)	4(0.26%)	4(0.26%)	(7.88%)
metal3	369(34.2%)	67(6.21%)	0(0.00%)	0(0.00%)	(40.4%)
metal4	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal5	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal6	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal7	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal8	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal9	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal10	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal11	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal12	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
Total	465(3.16%)	98(0.67%)	4(0.03%)	4(0.03%)	(3.89%)

(b) Single Metal Stack

Layer	OverCon #Gcell (1-4)	OverCon #Gcell (5-8)	OverCon #Gcell (9-12)	OverCon #Gcell (13-16)	#Gcell OverCon
metal1	7(0.62%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.62%)
metal2	52(3.33%)	14(0.90%)	2(0.13%)	2(0.13%)	(4.49%)
metal3	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal4	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal5	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal6	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal7	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal8	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal9	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal10	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal11	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	(0.00%)
metal12	25(1.84%)	0(0.00%)	0(0.00%)	0(0.00%)	(1.84%)
Total	84(0.46%)	14(0.08%)	2(0.01%)	2(0.01%)	(0.56%)

(c) F2F Double Metal Stack

Fig. 5: ldpc Benchmark Congestion Analysis



(a) 3D Design Routing (b) 3D Layout (c) Bottom Tier Layout (d) Top Tier Layout

Fig. 6: *ldpc* Benchmark Routing and 3D Tiers Layout using Single Metal Stack

TABLE I: Single vs. Multiple IPO Impact to 3D Flows - Best Single/Multiple IPO Results

Benchmark	Partitioning Methodology	IPO	Power (mW)	Circuit delay (ns)	Violating Paths	Total Components Area (um^2)
ldpc #Cells = 96361 #Nets = 98412 #IOs = 4100	Greedy BBFM	Single	1064.763	2.0165	2046	267206.082
		Multiple	797.023	1.6275	0	266638.970
	3DLG	Single	798.605	1.6595	367	267206.082
		Multiple	790.288	1.6275	0	266874.670
fft #Cells = 194175 #Nets = 194219 #IOs = 90	Greedy BBFM	Single	630.063	1.481	16	614178.838
		Multiple	624.466	1.395	0	613528.728
	3DLG	Single	638.795	1.395	0	615576.253
		Multiple	634.988	1.395	0	615426.188
jpeg #Cells = 576366 #Nets = 576394 #IOs = 67	Greedy BBFM	Single	2258.869	1.5264	53	1730458.782
		Multiple	2257.626	1.4694	0	1730142.602
	3DLG	Single	2249.683	1.4954	46	1735985.287
		Multiple	2244.926	1.4744	4	1735112.334

is high, 30 gates, confirming that critical paths are gate dominated. WL variation across critical paths depends on the configuration, yet 2D circuit exhibits higher WL as well as numbers of logic levels. Another observation is that, as F2F metal stack reduces the routing congestion, the router utilizes fewer routing vias compared to the single metal stack. This is illustrated by the average vias per net of Table II. Thus, a path based partition method combining the two algorithms may reduce the mismatch between global TWL and critical path WL, as well as total WL gains and TNS gains.

Table IV lists the respective power analysis breakdown for the *ldpc* circuit. The leakage power component is small and net switching power and internal cell power are of the same order of magnitude. The circuit with the least area and total WL, *i.e.* Greedy Bin-Based, Single does not exhibit the lowest power. We suspect that this is due to a large number of its critical paths having higher WL.

IV. Conclusions and Future Work

We have presented a complete 3DIC flow with two different tier partition methodologies, including a multi-pass, timing-driven IPO strategy, focusing on achieving maximum performance. We have shown that our multi-pass IPO strategy achieves significant QoR improvements over a single IPO run, as the latter does not always improve QoR in a single pass. The 45 nm library used, yields gate-dominated path delays. The resulting 3D circuits indeed achieve higher performance, in the range of 1% to 3%, for the same multi-pass IPO flow. We have shown that 3D critical paths do have less WL and logic levels, yet if wire *RC* delay is not that significant, the speedup is not greater. From the two tier partition approaches used, GBBFM yields better QoR than 3DLG, on average. Also, a Single metal stack, despite the higher congestion, for this specific library, where cells are 16 tracks high, achieves better QoR than the F2F stack. Based on the obtained results and

TABLE II: 2D vs. 3D Experimental Results for different 3D Flow configurations

Benchmark	Flow	Partitioning Methodology	Metal Stack Type	Power (mW)	Circuit Delay (ns)	TNS	Violating Paths	Total Components Area (um^2)	WL	Total #Gcell OverCon	Average Vias per Net
ldpc	2D	-	-	841.530	1.6885	-47.426	1926	284988.6	3.41E+06	0.07%	7.79
	Pre-Partitioning		Single	737.18	1.6275	0	0		2.29E+06	0.64%	7.99
			F2F	755.584	2.2335	-0.606	148		2.33E+06	0.32%	8.06
	3D	Greedy BBFM	Single	797.023	1.6275	0	0	266638.970	2.36E+06	3.98%	9.14
			F2F	876.126	1.7055	-75.97	2024	277590.456	2.48E+06	0.56%	9.11
		3DLG	Single	790.288	1.7055	0	0	269874.670	2.38E+06	15.16%	9.03
			F2F	889.458	1.7275	-95.186	2026	272690.964	2.47E+06	7.83%	9.02
fft	2D	-	-	643.201	1.43	-0.175	15	605736.872	2.70E+06	0%	7.61
	Pre-Partitioning		Single	625.36	1.554	-0.159	15		1.96E+06	1.48%	8.15
			F2F	625.85	1.405	-0.01	1		1.97E+06	1.22%	8.15
	3D	Greedy BBFM	Single	624.466	1.395	0	0	613528.728	1.95E+06	0.73%	9.50
			F2F	637.117	1.396	-0.001	1	615376.294	2.03E+06	0.49%	7.78
		3DLG	Single	634.988	1.395	0	0	615426.188	2.23E+06	0.68%	9.50
			F2F	646.056	1.395	0	0	612395.498	2.27E+06	0.58%	7.90
jpeg	2D	-	-	2351.182	1.5104	-2.603	188	1736969.2	8.54E+06	0%	6.90
	Pre-Partitioning		Single	2247.128	2.0384	-0.569	50		6.16E+06	0.17%	7.13
			F2F	2292.492	1.5124	-0.043	15		6.18E+06	0.21%	7.99
	3D	Greedy BBFM	Single	2257.626	1.4694	0	0	1730142.602	6.22E+06	0.38%	8.89
			F2F	2327.227	1.4694	0	0	1758207.256	6.23E+06	0.02%	6.87
		3DLG	Single	2244.926	1.4744	-0.009	4	1735112.334	6.20E+06	0.14%	9.10
			F2F	2344.746	1.4724	-0.006	4	1756345.864	6.30E+06	0.01%	7.07

TABLE III: *ldpc* Benchmark Critical Paths Analysis

Critical Path	2D			Greedy BBFM Single Metal Stack			Greedy BBFM F2F Metal Stack			3DLG Single Metal Stack			3DLG F2F Metal Stack		
	WL	Slack	Levels	WL	Slack	Levels	WL	Slack	Levels	WL	Slack	Levels	WL	Slack	Levels
1	1834.375	-0.061	34	1223.893	0	30	1139.522	-0.078	33	833.4115	0	28	1248.524	-0.1	36
2	1623.165	-0.061	37	1118.964	0	36	1131.616	-0.07	35	1167.025	0	21	1050.561	-0.098	33
3	1505.872	-0.054	30	1231.764	0	34	1094.505	-0.069	32	1157.744	0	22	1090.111	-0.093	29
4	1649.029	-0.05	32	1055.661	0	26	1245.489	-0.067	34	1189.290	0	29	1282.279	-0.087	33
5	1427.631	-0.047	34	1155.783	0	32	1262.69	-0.065	34	921.174	0	32	997.490	-0.085	39

TABLE IV: *ldpc* Benchmark Power Analysis

Flow	Metal Stack Type	Total	Leakage	Switching	Internal
2D	-	829.54	21.264	393.093	425.23
Greedy BBFM	Single	797.023	20.523	364.24	412.26
	F2F	876.126	22.123	401.29	452.71
3DLG	Single	790.288	20.956	367.88	401.45
	F2F	889.458	22.308	409.85	457.3

having also investigated 28 nm wire *RC* delays which are small with respect to gate delay, we believe that higher 3D gains may only be exhibited for 16 nm and below libraries where gate to wire delay ratio and cell track height is smaller or at low voltages.

ACKNOWLEDGMENTS

The authors would like to thank to Prof. Emre Salman from Stony Brook University for valuable discussions and providing the 3D nanoCAS library.

REFERENCES

[1] W.-T. J. Chan, A. B. Kahng, and J. Li, "Revisiting 3DIC benefit with multiple tiers," *Integration*, vol. 58, pp. 226–235, 2017.

[2] S. S. K. Pentapati, D. E. Shim, and S. K. Lim, "Logic Monolithic 3D ICs: PPA Benefits and EDA Tools Necessary," in *Proceedings of the 2019 on Great Lakes Symposium on VLSI*, 2019, pp. 445–450.

[3] N. Sketopoulos, C. P. Sotiriou, and V. Samaras, "Investigation and Trade-offs in 3DIC Partitioning Methodologies," in *Proceedings of the 2019 on Great Lakes Symposium on VLSI*, 2019, pp. 451–455.

[4] S. K. Samal, D. Nayak, M. Ichihashi, S. Banna, and S. K. Lim, "Monolithic 3D IC vs. TSV-based 3D IC in 14nm FinFET technology," in *2016 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*. IEEE, 2016, pp. 1–2.

[5] S. Bobba, A. Chakraborty, O. Thomas, P. Batude, V. F. Pavlidis, and G. De Micheli, "Performance analysis of 3-D monolithic integrated circuits," in *2010 IEEE International 3D Systems Integration Conference (3DIC)*. IEEE, 2010, pp. 1–4.

[6] S. Panth, K. Samadi, Y. Du, and S. Kyu Lim, "Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs," in *Proceedings of the 2014 International Symposium on Physical Design (ISPD)*, 2014.

[7] B. W. Ku, P. Debacker, D. Milojevic, P. Raghavan, D. Verkest, A. Thean, and S. K. Lim, "Physical design solutions to tackle FEOL/BEOL degradation in gate-level monolithic 3D ICs," in *Proceedings of the 2016 International Symposium on Low Power Electronics and Design*. ACM, 2016, pp. 76–81.

[8] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Placement-driven partitioning for congestion mitigation in monolithic 3D IC designs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 4, pp. 540–553, 2015.

[9] C. M. Fiduccia and R. M. Mattheyses, "A linear-time heuristic for improving network partitions," in *19th Design Automation Conference*. IEEE, 1982, pp. 175–181.

[10] N. Sketopoulos, C. Sotiriou, and S. Simoglou, "Abax: 2D/3D legaliser supporting Look-ahead Legalisation and Blockage strategies," in *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2018, pp. 1469–1472.

[11] C. Yan and E. Salman, "Mono3D: Open source cell library for monolithic 3-D integrated circuits," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 3, pp. 1075–1085, 2017.

[12] Cadence, "Cadence Innovus," 2019, https://cadence.com/.

[13] OpenCores, "OpenCores Open IP Community," 2019, https://opencores.org/.