Heterogeneous Monolithic 3D ICs: EDA Solutions, and Power, Performance, Cost Tradeoffs

Sai Surya Kiran Pentapati
School of Electrical and Computer Engineering
Georgia Institute of Technology
Email: sai.pentapati@gatech.edu

Sung Kyu Lim
School of Electrical and Computer Engineering
Georgia Institute of Technology

***Abstract*—In this paper, we present a novel heterogeneous design of Monolithic 3D ICs along with crucial design flow enhancements and better partitioning methods. The heterogeneous M3D ICs are designed with a combination of low-cost, low-power, and low-performance cells on one die and a higher-cost, power, and performance technology variant on the tier, for heterogeneity. These heterogeneous designs out-perform most 2D, 3D variants in Power-Delay Product and Cost metrics. Using 4 different netlists, we see up-to 23% improvement in Performance per Cost, and 16% improvement of Power Delay Product with heterogeneous M3D compared to the best 2D designs.**

I. Introduction

Monolithic 3D (M3D) IC design is based on a fabrication methodology that sequentially constructs two dies on top of each other [1] that are connected using dense Monolithic Inter-tier Vias (MIVs). To evaluate the benefits of monolithic 3D design, several RTL-to-GDS "pseudo-3D flows" that make use of commercial EDA tools to varying extents have been proposed. Most of these flows improve the Power-Performance-Area (PPA) in homogeneous M3D ICs by exploiting improvements in 3D wirelength, placement, routing, or a specific 3D stacking. Similarly, heterogeneous integration (where each tier of an M3D IC utilizes a different technology node) is an exclusive benefit of M3D fabrication but is not well studied.

Only the most recent pseudo-3D flow Pin-3D [2], support heterogeneous 3D IC optimization at the dense gate-level partitioning. Heterogeneity is a major focus in 2.5D ICs and is also recently fabricated with microbump bonded 3D ICs [3]. These implementations use a coarse level heterogeneity at chip-let level in 2.5D ICs, and at the block level in bonded 3D ICs. M3D ICs, with their dense pitch, can support technology heterogeneity at a finer level of partitioning. So, a single netlist can be divided into two tiers at a gate-level and manufactured with different technologies.

In our work, we use an enhanced Pin-3D flow to study the behavior of heterogeneous 3D, and PPAC (PPA, Cost) impact using 4 different RTLs: AES (cell-dominant design), Netcard (large design, slightly wire dominant), LDPC encoder and decoder circuit (extremely wire dominant), CPU design from a commercially available core (large, general purpose design with memory blocks). Each of these four netlists is designed in 5 technology and design configurations as shown in Fig. 1. Overall, we see that heterogeneous design achieves better Power-Delay Product (Energy Efficiency) and PPC (Performance/(Power*Cost)).

978-1-6654-3274-0/21/$31.00 ©2021 IEEE

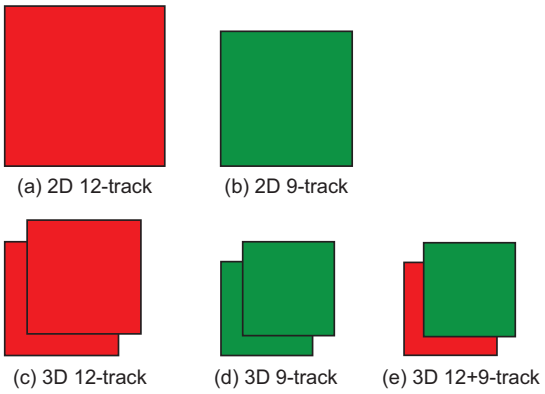


Fig. 1. 5 different configurations (to scale, assuming equal number of cells) of 2D and 3D using 9-track and 12-track cells studied in this work. We use commercial 28nm libraries.

II. Technology Setup

Heterogeneous technology can be used to target different aspects of the PPAC metrics. For example, the authors in [3] use a low-power $22\,\mathrm{nm}$ and a high-performance $10\,\mathrm{nm}$ technologies to design different blocks of an SoC with dissimilar PPA requirements such as the compute core and periphery modules. In our work, we cost reduction along with power reduction at a high operational frequency (Section IV-A defines our interpretation of a 'high' frequency). By studying the PPC trends of previous technology nodes, we pick a mixture of libraries from the set of foundry provided $28\,\mathrm{nm}$ libraries that meet our PPAC requirements.

A. Cost Trends

The equation *Cost per Element* = *Wafer Cost per* mm^2 × *Area of Element* is a useful to understand the cost trends with scaling. In general, the increase in wafer cost at each generation is accompanied by a significant increase in transistor density, resulting in cost per element reduction for advanced technologies [4]. In such cases, it is more expensive to manufacture a chip using an older technology node. Wafer cost has been growing at an increasingly higher pace only in the recent technology nodes, and cost per chip has only increased slightly at the $5\,\mathrm{nm}$ node [5]. Technology scaling also improves power and performance, and advanced technology nodes always achieve better PPC. So, such technology heterogeneity is not the best candidate to achieve PPC improvement without specially designed libraries [3].

TABLE I
COST MODEL PARAMETERS AND ASSUMPTIONS [6]

Baseline wafer cost (FEOL+8 metals)	C'
Wafer FEOL cost	$0.3 \times C'$
Wafer BEOL cost (up-to 6 metals)	$0.66 \times C'$
3D integration cost (α)	$0.05 \times C'$
Wafer Diameter	300 mm
Defect Density (D_w)	$0.2\,\mathrm{mm}^2$
Wafer yield (κ)	0.95
3D Yield Degradation (β)	0.95
2D Wafer Cost (C_{2D})	$0.96 \times C'$
3D Wafer Cost (C_{3D})	$1.97 \times C'$

TABLE II
"QUALITATIVE" COMPARISONS OF EXPECTED PPAC BEHAVIOR OF THE 5 TECHNOLOGY AND DESIGN CONFIGURATIONS AT THEIR EXPECTED MAXIMUM FREQUENCIES. 1 MEANS THE WORST, AND 5 THE BEST

	9 Track (slow & small)		12 Track (fast & large)		9+12 Tracks (combined)	
	2D	3D	2D	3D	2D	3D
Frequency	1	2	4	5	-	3
Power	4	5	1	2	-	3
Power/Freq	3	4	1	2	-	5
Footprint	4	5	1	2	-	3
Si Area	5	5	1	1	-	3
Die Cost	5	4	2	1	-	3

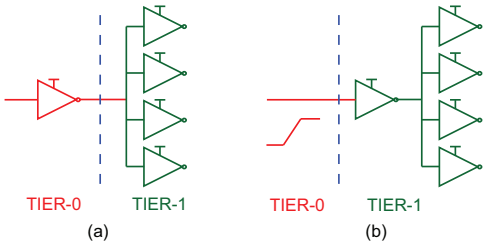


Fig. 2. The two types of boundary conditions due to heterogeneity in a FO-4 inverter. (a) Heterogeneity at driver output, (b) Heterogeneity at driver input.

TABLE III
IMPACT OF HETEROGENEOUS TECHNOLOGY WHEN INPUT TO DRIVER OF AN FO4 IS FROM DIFFERENT TIER (SEE FIGURE 2(B)). TIME IS IN ns, POWER IS IN µW, VOLTAGE IS IN V

	Case-I	Case-II	Δ%	Case-III	Case-IV	Δ%
Tier-0	fast	slow	–	slow	fast	–
Tier-1	fast	fast	–	slow	slow	–
Driver V_G	0.9	0.81	10.0	0.81	0.9	11.1
Rise Slew	15.6	16.9	8.1	14.6	13.1	-9.9
Fall Slew	18.2	19.4	6.6	19.1	17.6	-8.1
Rise Del.	12.5	13.0	3.4	23.6	22.4	-5.3
Fall Del.	16.4	17.1	4.1	26.2	24.8	-5.1
Lkg. Pow.	0.093	0.330	250	0.003	0.002	-44.9
Total Pow.	3.86	4.21	9.2	2.00	1.99	-0.6
% Input Pins	31	15	–	33	21	–

Apart from technology node scaling, using cells with a fewer tracks (shorter cell height), the cell area decreases without incurring additional wafer costs as the design rules and the mask layers complexity remain the same [7]. The smaller cells with fewer tracks have better power, but worse timing and larger cells with more tracks have worse power but better timing. So, we cannot optimize all the PPAC metrics at once using just a single cell track type. In our work, we use a 12-track (highest available) and a 9-track (smallest available) libraries from the commercial foundry 28 nm node in heterogeneous 3D ICs to extract benefits from the both track types simultaneously without incurring significant drawback. The two track variants have remarkably similar BEOL (Back-End-Of-Line), and so the routing layers in the three 3D cases are similar to each other. To analyze the die cost and PPC metrics, we use a cost model previously used for M3D ICs in [6] assuming it will hold in our 28 nm heterogeneous case. Table I shows the list of parameters used. The BEOL cost is split between the metal layers based on the routing pitch.

Table II shows the expected full-chip behavior of the technology and design configurations. In the following subsection, we discuss the various quirks of using such heterogeneous cells in a commercial 2D tool.

B. *Quirks of Heterogeneity*

To achieve a heterogeneous design with significant power and area benefits without incurring a marked performance loss, we have to choose a suitable process and voltage corner for the different libraries. Here, we choose a slower process (higher threshold voltage) and a lower voltage (0.81 V) for the 9-track libraries, and a faster process (lower threshold voltage) along with higher voltage (0.90 V) for the 12-track libraries to separate the libraries further in terms of their achievable PPA. When the two libraries have similar characteristics, heterogeneous 3D cannot be much different or better than a homogeneous 3D. The voltage difference (0.09 V) is considerably smaller than the PMOS threshold voltage $V_{thp} \sim 0.3$ V of the cells in the 0.90 V domain. This ensures that the pull-up network turns-off without the need for voltage shifters that impede dense connectivity benefits of M3D ICs.

An FO-4 inverter with the two extreme heterogeneous configurations is shown in Fig. 2. The cells connected to other tier are referred to as 'boundary cells'. In the '**heterogeneity at driver output**' case, the heterogeneity affects the load pin capacitances and the output slew of driver. The liberty (LIB) model files capture the pin capacitance variation due to the heterogeneous cells, and the slews are accurately calculated. Spice analysis also shows that when the driver is in a separate tier from the load, the slew changes by at-most $\pm 15\%$. So, as long as the libraries have a significant overlap in characterized slew ranges these variations lie within the characterized range that usually span 2-3 orders of magnitude.

The next type of variation at boundary cells occurs when the entire FO-4 setup is on a single tier, but the driver input comes from the other. This case is referred to as '**heterogeneity at input**' and the LIB files used here do not model differences in input logic-high level and the power-domain level. The impact of this type of boundary heterogeneity is shown in Table III. With only $\sim \pm 5\%$ difference in cell delays, the impact on total path timing is negligible. In a heterogeneous 3D design,

TABLE IV
IMPROVEMENTS OBTAINED WITH OUR HETEROGENEOUS VERSION OF PIN-3D FLOW [2] FOR THE COMMERCIAL CPU DESIGN

		Pin-3D [2]	Hetero-Pin-3D
Frequency	GHz	1.200	1.200
WL	m	3.22	3.07
WNS	ns	-0.489	-0.055
Total Power	mW	224.1	188.0

the number of boundary input pins of the two heterogeneous cases (Cases II and IV from Table III) differ at most by 1. So, even with multiple MIVs (Monolithic Inter-tier Vias) on a path that goes back and forth between the two tiers, the estimated timing using the LIB files only differ by $\sim 5\%$ of cell delay.

The main discrepancy for 'input heterogeneity' is in the leakage power, which increases $2.5\times$ when logic-high is at $0.81\,\mathrm{V}$, and pull-up network connected to $0.90\,\mathrm{V}$ (Case-II in Table III). This increase is caused by the exponential nature of I_{DS} vs. V_{GS} in the cut-off region. At full-chip level, the leakage power is just $\sim$1% of the total power estimate in our heterogeneous designs. On average only 15% of the input pins over all the four heterogeneous implementations fall in Case-II. Assuming each input pin has an equal contribution to leakage and considering a $2.5\times$ increase in individual leakage value, the leakage power increases by 37.5 and total power by just 0.375%. Even after assuming the worst-case leakage increase of 37.5%, heterogeneous designs are still better than using only fast cells in a 2D or 3D configuration as the slow cells have a $\simeq 10-20\times$ smaller leakage than fast cells.

III. HETEROGENEOUS 3D IC DESIGN FLOW

A. Enhancing the Pin-3D Flow

To design a timing optimized heterogeneous 3D IC in [2], the authors make several assumptions regarding technology, clock tree, voltage levels, etc. Using Pin-3D in its proposed form for heterogeneous 3D, the timing and total power become significantly worse due to the unconventional clock-tree and critical path behavior of heterogeneous 3D as seen in Table IV. These peculiarities are revisited using full-chip analysis in Section IV-C. As the post-CTS optimized pseudo-3D input to Pin-3D is done in a 2D fashion, it is based on homogeneous technology, and differs significantly from a clock tree that is optimized for a heterogeneous 3D. Moreover, the inability of the area-balanced min-cut partitioning to account for timing difference between the dies in heterogeneous circuits exacerbates the timing issues. Several enhancements are proposed to the Pin-3D flow to alleviate aforementioned issues and achieve the results in Table IV.

1) Timing-based Partitioning: When the input pseudo-3D stage for the heterogeneous 3D is designed with the faster library, a placement-driven and cut-size driven partitioning can assign cells on critical paths to the slower die to improve the cut-size. In such cases, the critical paths show an increase in worst slack, and optimization with this partitioning solution may not result in timing closure. In our experiments, we see that a path-based partitioning solution [8] cannot be applicable here. In path-based timing queries, the paths are differentiated by the begin-end pair of registers, and there can be several different paths between a same pair. The different paths between a specific register pair naturally have a large intersection in terms of the cells on the path, and the coverage increases at a very slow rate as we increase the number of paths per pair. In our experiments, even after considering up-to 1024 paths per register pair, we were only able to cover 60% of the total constrained cells in the design and eventually, the run-time to just calculate cell coverage increased to up-to an hour with only 60-65% unique cells traversed.

Therefore, we use a cell-based partitioning where each cell is visited exactly once and the slack value of the worst timing path through the cell is assigned as 'cell-slack'. If the cell is not constrained, slack is assumed to be a large positive value. When partitioning is based only on timing, 3D placement becomes highly clustered, and the legalization significantly alters the input placement. So, only a 20-30% of the total cell area is assigned to the faster die. Bin-based FM min-cut completes the partitioning to obtain a tradeoff between the timing and placement based partitioning solutions. The area threshold depends on the design and their architecture, based on how many critical paths need to be fixed to fast die.

2) Supporting Heterogeneous Clock Tree: The clock tree engine within the commercial EDA tool, Innovus, doesn't treat the zero-sized overlaps as zero-area cells in Pin-3D, and causes a density overflow during the clock design. To remedy this, we use a cleaner and equally efficient approach of transparent cells using the COVER cell construct provided in LEF. This class of cells is considered by the tool to have no active area and does not break the clock engine. As the tool still uses the LIB files for accurate PPA estimations, COVER cells are a useful construct for the zero-sized cells. Based on the placement and timing optimization in Pin-3D, the clock tree optimization is done in two stages for the top and bottom tiers.

B. Re-partitioning Using ECO

Area-balancing between the tiers in a 3D design is crucial for efficient usage of silicon area in 3D. A heavily skewed utilization leads to placement and routing congestion within the high utilization die and worsens full-chip timing and power. In heterogeneous designs, the timing-based partitioning is not always sufficient to ensure area balance as not all the critical cells in 3D are identified after pseudo-3D. Additionally, a sub-par area threshold value can end up with some critical cells on the slower die. So, we introduce incremental re-partitioning 1 based on 3D timing, unbalance to alleviate area unbalance after any 3D optimization stage.

IV. EXPERIMENTAL RESULTS

As discussed in Section II-A, we use a 9-track library at $0.81\,\mathrm{V}$ and a 12-track library at $0.90\,\mathrm{V}$. In heterogeneous 3D designs, the bottom tier has the faster 12-track tech, and the top tier has the slower 9-track tech. Six metal layers are used per tier in 3D, and six metal layers are used in total in 2D to complete signal routing. As the foundry provided BEOL file is

Algorithm 1: Re-partitioning algorithm

$unbalance \leftarrow (area_{slow} - area_{fast})/area_{total}$
$d_k \leftarrow d_0; n_p \leftarrow n_0$
▷ Initialize delay threshold factor, number of critical paths considered per loop **while** $unbalance > unbalance_{th}$ **do**
  $d_{th} \leftarrow d_k \times$ (avg. cell delay of n_p critical paths)
  foreach $cell \in$ *the* n_p *critical paths* **do**
    $d_{cell} \leftarrow$ delay of $cell$ on its critical path
    if $d_{cell} > d_{th}$ **then**
      $all\_crit+ = 1$
      if $cell \in$ *slow die* **then**
        $slow\_crit+ = 1$
        append $cell$ to $move\_list$
      end
    end
  end
  if $slow\_crit/all\_crit < crit_{th}$ **then**
    break; ▷ slow and fast tier cells have similar delay
  end
  Move all cells in $move\_list$ to the fast die
  update timing
  if $Delta_{WNS} < W_{th}$ || $Delta_{TNS} < T_{th}$ **then**
    Undo the cell moves done
    $d_k* = \alpha$ ▷ $\alpha < 1$, update delay threshold
  end
  update $unbalance$
end

virtually identical for both the track variations, the MIV layer is the same in all the three M3D configurations. The MIV is similar to a routing via at $140\,\mathrm{nm}$ tall, $70\,\mathrm{nm}$ wide square face, and resistance and capacitance of $\sim 2\,\Omega$ and $\sim 0.015\,\mathrm{fF}$. The BEOL parasitics, including the MIV layer parasitics, are generated using Quantus from Cadence®using the material properties and physical dimensions of the layers.

A. Methodology

The first step in RTL-to-GDS conversion is netlist synthesis. The footprint is fixed to be a square and the area is determined using a provided target utilization value and the synthesized netlist. To allow for any noise in the PNR optimization, a worse negative slack of up-to $\sim 5-7\%$ of clock period is considered as timing-met. AES can achieve a rather small clock periods, and the slack threshold of 10% the clock period is used for determining the timing-met condition. For the iso-performance comparison among various technology configurations, a 'high' frequency is used. This is defined as the value where slow 2D does not meet timing, and fast 2D shows negative timing slacks at the initial target utilization.

Homogeneous 3D ICs follow the same design flow as in [2] using the 'high' frequency from 2D implementations, and the are from its corresponding (12-track or 9-track) 2D design. Heterogeneous 3D IC design starts with pseudo-3D stage at fast-node using same footprint area as its fast 2D analogue, and the timing-based partitioning enhancement gives the partitioning solution. As the 9-track cells are 25% smaller, more than 50% ($\sim 60\%$) of the 12-track cell area should be converted to 9-track tier, to achieve area-balanced partitioning. So, the foot-print is further shrunk by $\sim 13.5\%$ to maintain chip utilization. Finally, the enhanced Pin-3D flow is used to create the heterogeneous 3D GDS.

TABLE V
PPAC RESULTS OF OUR 3D HETEROGENEOUS DESIGNS (RAW DATA BASED ON A COMMERCIAL FOUNDRY 28 nm TECHNOLOGY)

	Units	netcard	aes	ldpc	cpu
Frequency	GHz	1.750	3.000	1.125	1.200
Area	mm^2	0.384	0.126	0.216	0.390
Chip Width	µm	438	251	329	442
Density	%	82	86	64	88
WL	m	6.560	1.022	5.500	3.073
# MIVs	k	153	62	83	98
Total Power	mW	550	138	339	188
WNS	ns	-0.037	-0.028	-0.026	-0.055
TNS	ns	-0.41	-4.827	-0.902	-15.54
Effective Delay	ns	0.608	0.361	0.597	0.888
PDP	pJ	334.5	49.8	310.1	167
Die Cost	$10^{-6}C'$	6.16	1.97	3.41	6.26
PPC	$\frac{\mathrm{GHz}}{\mathrm{mW}\times 10^{-6}C'}$	0.517	11.06	0.946	1.02

B. Full-Chip PPAC

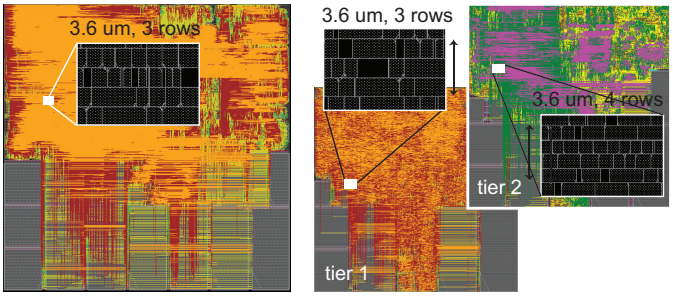
1) Heterogeneous 3D IC Results: Table V shows the raw values of the 4 RTL implementations with heterogeneous 3D. The density value reported for 3D designs is the average of the two tiers. Being heavily wire dominant, LDPC cannot achieve high placement densities without causing routing congestion. 3D routing in LDPC significantly changes its critical paths, and up-to 20% area-unbalance occurs after the first optimization stage. Re-partitioning fixes this issue and brings down the final unbalance to within 5%. The other three designs are not as wire dominant, and timing-based partitioning alone limits the area unbalance to ¡5%. The Worst Negative Slack (WNS) is within $\sim 7\%$ of the clock period as a result of the heterogeneous enhancements for all the four different RTLs. Effective delay (=clock period - worst slack) value is used for PDP calculation to factor in the small WNS imperfections. Finally, PPC is calculated based on achieved frequency, power consumption, and die cost.

2) Heterogeneous 3D vs. 9-track 2D: In Table VI, the change of each metric in heterogeneous 3D (referred to as HT-3D) w.r.t. the four other design configurations is shown. In the first set of four columns, we see that when compared to 2D 9-track (9T-2D), HT-3D shows a significant total area benefit as 9T-2D designs require a strict timing target and a larger area to meet the iso-performance targets. The large CPU, netcard design still cannot meet timing targets even with higher area usage. While HT-3D contains some large 12-track cells on the faster tier, it utilizes them efficiently to meet timing targets without the need for a larger area. We see that this area increase in 9T-2D is not unwarranted as the densities at this point are only 3-4% smaller than the $\sim 85\%$ density of HT-3D. Overall we see that the 'high' frequency is either

TABLE VI

PPAC PERCENTAGE DELTA (= (3D HETERO - CONFIG) / CONFIG × 100) OF 3D HETEROGENEOUS DESIGN W.R.T. DIFFERENT HOMOGENEOUS CONFIGURATIONS. A -VE (+VE FOR PPC) VALUE IMPLIES THAT HETEROGENEOUS IMPLEMENTATION OUTPERFORMS THE PARTICULAR CONFIGURATION.

CONFIG →	2D 9-Track				2D 12-Track				M3D 9-Track				M3D 12-Track			
	netcard	aes	ldpc	cpu	netcard	aes	ldpc	cpu	netcard	aes	ldpc	cpu	netcard	aes	ldpc	cpu
Si Area	-28.6	-27.6	-13.9	-23.1	-12.3	-13.7	-13.9	-7.8	-28.6	-27.6	-13.6	-23.2	-12.3	-14.9	-13.6	-8.0
Density	3.2	2.9	-12.4	4.1	-1.8	14.5	2.2	5.4	8.7	6.9	-10.0	4.1	4.2	14.8	27.9	5.4
WL	-33.0	-30.9	-14.4	-34.8	-25.3	-17.0	-31.1	-33.5	7.8	-10.4	18.6	-8.9	-2.2	0.0	-8.7	-2.1
Total Power	-12.7	-15.0	-8.1	-21.1	-10.5	-8.0	-29.7	-14.6	-4.2	-8.6	5.4	-15.9	-6.0	-3.4	-5.6	-10.1
Eff. Delay	-14.7	-1.6	-0.8	-12.9	1.0	4.0	-2.6	-6.2	-19.3	0.0	1.2	-12.9	6.1	6.2	2.2	4.2
PDP	-25.6	-16.3	-8.8	-31.2	-9.6	-4.2	-31.5	-9.3	-22.6	-8.5	6.6	-26.7	-0.2	2.7	-3.5	-6.3
Die Cost	-27.9	-23.3	-9.5	-21.8	-9.7	-8.4	-9.5	-4.7	-29.7	-27.8	-13.9	-24.1	-12.7	-15.1	-13.9	-8.3
PPC	59.1	53.5	20.3	61.9	23.7	18.6	57.2	23.0	48.6	51.6	10.2	56.7	21.9	21.9	23.0	21.4
Width (μm)	733	417	501	712	662	382	501	650	518	295	353	504	468	272	353	460
WNS (ns)	-0.14	-0.03	-0.03	-0.19	-0.03	-0.01	-0.05	-0.01	-0.18	-0.03	-0.02	-0.19	-0.00	-0.01	-0.01	-0.02
TNS (ns)	-450	-3.61	-5.13	-357	-2.14	-0.44	-7.60	-0.03	-832	-1.95	-0.04	-316	-0.00	-0.03	-0.01	-0.59



(a) 2D 12-track cells (b) 3D heterogenous

Fig. 3. Routing and zoomed placement GDS layouts of our commercial CPU. (a) 2D 12-track, (b) 3D heterogeneous, where tier 1 is using 12-track cells and tier 2 9 track cells.

unattainable or the 9T-2D fail to show any of their expected benefits from Table II at the frequency target.

3) Heterogeneous 3D vs. 12-track 2D: 12-track technology is the faster technology used, and from Table VI we see that the 12T-2D are near their max-frequency limit as the worst slack is negative, showing the limits of timing optimization. They also have a high utilization at around 75% for AES, and >80% for the others. The effective delay of HT-3D is very close to the 12T-2D designs showing that HT-3D can truly achieve the 'high' frequency of fast 2D. By virtue of using 9-track cells, HT-3D is consistently the best in terms of PDP and PPC, with $18.6-57.2\%$ improvement in the PPC. In the CPU design, the memory cell technology is left unchanged within the two dies. So, we only see a 7.8% area reduction rather than the $12-13\%$ reduction seen in other netlists. Even without this additional cost and power benefits due to memory macros, the CPU design still shows a 23% PPC improvement.

4) Heterogeneous 3D vs. 9-track 3D: Similar to 9T-2D, 9T-M3D designs have a worse area, timing, power, and cost in every case except the LDPC design. The 9T-3D of the wire-dominant LDPC has a smaller wirelength with the help of the 3D placement and routing space, and greatly benefits the power and even delay to a lesser extent. But the required area increase to meet the timing is significant drawback and PPC is improved by 10% in HT-3D. Like its 2D implementation, the CPU design still cannot meet the timing, but power and PDP in 9T-3D are slightly better than the 9T-2D case.

TABLE VII

CLOCK NETWORK, CRITICAL PATH, MEMORY INTERCONNECT ANALYSES OF THE COMMERCIAL CPU DESIGN

		12T 2D	12T 3D	Hetero 3D
Memory Interconnects				
RMS Input Net Latency	ps	25.0	16.1	15.5
RMS Output Net Latency	ps	37.5	33.2	28.7
Net Switching Power	mW	5.47	4.02	3.41
Clock Network				
Buffer Count		1593	1502	1330
Buffer Area	μm^2	1277	1175	982
Wirelength	m	0.114	0.108	0.107
Max Latency	ns	0.234	0.292	0.713
Max Skew	ns	0.058	0.142	0.344
100 Path Avg. Skew	ns	-0.008	0.000	-0.011
Critical Path				
Slack	ns	-0.003	-0.012	-0.055
Clock Skew	ns	-0.014	0.013	0.045
Setup Time	ns	0.004	0.008	0.001
Path Delay	ns	0.845	0.831	0.845
# MIVs		–	9	6
Bottom Cells		45	20	25
Bottom Cell Delay	ns	0.830	0.375	0.482
Avg. Bottom Delay	ns	0.019	0.019	0.019
Top Cells		–	23	8
Top Cell Delay	ns	–	0.447	0.343
Avg. Top Delay	ns	–	0.019	0.043

5) Heterogeneous 3D vs. 12-track 3D: 12-track 3D designs are expected to perform better than their 2D equivalent in terms of power, performance, and this is what we see in most of the cases as the power, timing benefit in HT-3D is smaller w.r.t 12T-3D than 12T-2D. But the 12T-3D has a worse die cost due to 3D cost overhead, and the overall PPC benefit is still $\sim 20\%$ for HT-3D compared to 12T-3D.

C. Analysis of clock, critical path, and memory connections

Using the commercial CPU core, we analyze the clock network, critical paths, and the connections to and from

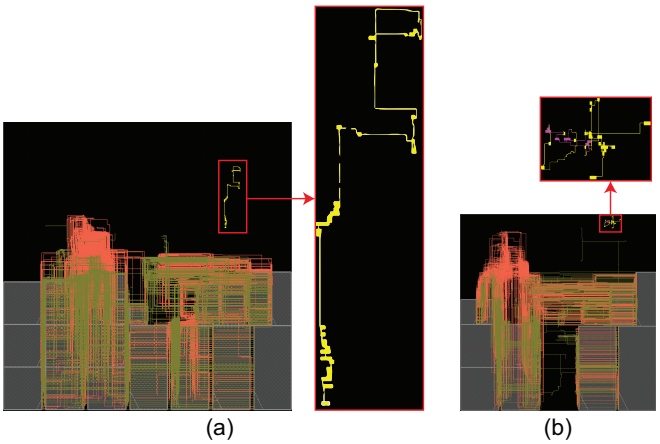


Fig. 4. Timing critical paths and memory nets of (a) 12-track 2D, (b) Heterogeneous 3D implementations of the CPU design. Yellow: 12-Track tier wires and cells, Magenta: 9-Track tier wires and cells, Dark Red: memory output nets, and Dark Green: memory input nets.

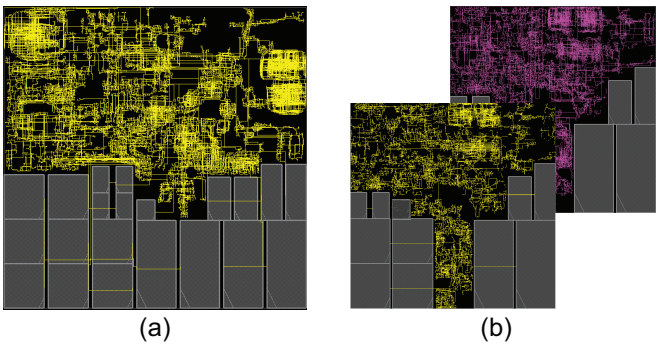


Fig. 5. Clock tree layouts of (a) 12-track 2D, (b) Heterogeneous 3D implementations of the CPU design. Yellow: 12-Track tier clock wires, Magenta: 9-Track tier clock wires

memories in 12-track 2D, 12-track 3D, and heterogeneous 3D to understand the various physical design aspects of heterogeneous M3D. These results are tabulated in Table VII.

1) Memory Interconnects: Root Mean Square average of the memory net latencies is used in Table VII as it is more affected by the larger delay values. Memory latency is better in 3D in general due to its smaller footprint and better macro placement that achieves dense wiring as shown in Fig. 4. In heterogeneous 3D, the smaller floorplan and pin-capacitance reduction from 9-track cells contributes to the additional latency reduction. Some of the nets connected to memory blocks in HT-3D are driven by lower V_{DD} and this contributes to additional switching power reduction.

2) Clock Network: The clock latency varies widely in HT-3D, resulting in large max latency as well as clock skew in heterogeneous 3D design due to the presence of clock network on both slow and fast dies as seen in Fig. 5(b). Even though the maximum skew and latency are worse, we see the benefit of our clock tree design in the average clock skew on the first 100 critical paths. Critical path clock skew is very important as they have a significant impact on critical path slack (Hetero-3D has same critical path delay as 12T-3D but a worse slack due to the clock skew).

3) Critical Path: The breakdown of the critical path is the most revealing aspect of the heterogeneity. The clock period is the same ($= 0.833$ ns) for the three configurations, and path delay is mostly ($\sim 97\%$) made up of cell delay. We see that the critical path in HT-3D is shorter in length with only a few cells on the slow die, unlike the 12T-3D path whose cells are nearly evenly split between the dies. We can see from the average cell delays that without such skewed critical path partitioning, the timing would worsen in HT-3D. This skewing is achieved with timing-based partitioning. Critical path layout in Fig. 4 also show the benefit of 3D placement as its critical path has a significantly smaller bounding box than 2D.

V. Conclusion

In this paper, we have presented a novel arrangement for gate-level monolithic 3D ICs using heterogeneous technology integration, along with various enhancements in partitioning, clock tree, and a new re-partitioning stage to support such 3D integration. We saw that using different cell tracks on the two tiers work the best for heterogeneous 3D ICs in terms of voltage and BEOL compatibility required for densely connected M3D ICs. Overall, the 3D hetero designs achieve high frequencies very close to fast-2D, and provide a PPC improvement of $18.6 - 57.2\%$ compared to timing-met 2D designs, and $10.2 - 51.6\%$ compared to the timing-met 3D.

Acknowledgement

This research is partially funded by the DARPA ERI 3DSOC Program under Award HR001118C0096, the Semiconductor Research Corporation under Task 2929, the National Research Foundation of Korea under NRF-2020M3F3A2A02082445, and Samsung Semiconductor, Inc.

References

[1] L. Brunet and C. Fenouillet-Beranger et al. Breakthroughs in 3D Sequential technology. In *2018 IEEE International Electron Devices Meeting (IEDM)*, Dec 2018.

[2] Sai Pentapati et al. Pin-3D: A Physical Synthesis and Post-Layout Optimization Flow for Heterogeneous Monolithic 3D ICs. In *Proceedings of the International Conference on Computer-Aided Design*, 2020.

[3] W. Gomes and S. Khushu et al. 8.1 Lakefield and Mobility Compute: A 3D Stacked 10nm and 22FFL Hybrid Processor System in 12×12mm2, 1mm Package-on-Package. In *IEEE International Solid- State Circuits Conference - (ISSCC)*, 2020.

[4] Kenneth Flamm. Measuring Moore's Law: Evidence from Price, Cost, and Quality Indexes. Technical report, National Bureau of Economic Research, April 2018.

[5] S. Khan and A. Mann. AI Chips: What They Are and Why They Matter. cset.georgetown.edu/research/ai-chips-what-they-are-and-why-they-matter/, April 2020.

[6] Ku, Bon Woong and Debacker, Peter et al. How much cost reduction justifies the adoption of monolithic 3d ics at 7nm node? In *Proceedings of the International Conference on Computer-Aided Design*, 2016.

[7] Peter Debacker et al. Low track height standard-cells enable high-placement density and low-BEOL cost (Conference Presentation). In *Design-Process-Technology Co-optimization for Manufacturability XI*, 2017.

[8] S. K. Samal and D. Nayak et al. Tier partitioning strategy to mitigate BEOL degradation and cost issues in monolithic 3D ICs. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2016.