Monolithic 3D IC Design: Power, Performance, and Area Impact at 7nm

Kartik Acharya[1], Kyungwook Chang[1], Bon Woong Ku[1], Shreepad Panth[1],
Saurabh Sinha[2], Brian Cline[2], Greg Yeric[2], and Sung Kyu Lim[1]
[1]School of ECE, Georgia Institute of Technology, Atlanta, GA
[2]ARM Inc., Austin, TX
[1]E-mail: kartik.acharya@gatech.edu, limsk@ece.gatech.edu

Abstract

In this paper, we present a comprehensive study of full-chip power, performance, and area metric for monolithic 3D (M3D) IC designs at the 7nm technology node. We investigate the benefits of M3D designs using our predictive 7nm FinFET libraries. This paper outlines detailed iso-performance power comparisons between M3D and 2D full-chip GDSII designs using both 7nm high performance (HP) and low stand-by power (LSTP) library cells. We achieve significant wire-length and buffer reduction with 7nm HP M3D designs over 2D counterparts, thus more power saving at high iso-performance frequency. In addition, this power saving is also realized in 7nm LSTP M3D designs running at low iso-performance frequencies. We also study the impact of clock tree design on the clock power consumption in M3D designs. Lastly, we demonstrate the impact of clock tree partitioning on the total power of full-chip M3D designs. Our experiments show that 7nm HP and LSTP M3D designs outperform its 2D counterparts by 12% and 10% on average, respectively.

Keywords

M3D, monolithic, 3DIC, FinFET, 7nm

1. Introduction

Monolithic 3D (M3D) is an emerging integration technology that can extend the semiconductor roadmap beyond the traditional 2D scaling trajectory predicted by Moore's Law. In M3D technology, two or more tiers of devices are fabricated sequentially one above the other, with nano-scale *monolithic inter-tier vias* (MIVs) for connections across tiers. Unlike other 3D integration technologies such as *through-silicon-vias* (TSVs) where pre-fabricated dies are bonded together, M3D eliminates the need for aligning tiers, by fabricating the top tier devices and interconnects on top of the bottom tier with a low temperature transistor process. Moreover, thanks to the extremely small size of inter-tier vias, M3D offers orders of magnitude higher integration density with increased vertical connectivity. This ultra-high integration density provides reduced silicon area and cost, with considerably reduced MIV parasitics that improve the power performance benefit of M3D technology.

There are three types of M3D implementations: transistor-level, gate-level, and block-level. In transistor-level M3D, the NMOS and PMOS transistors itself are partitioned into separate tiers with MIVs for intra-cell and inter-cell connections. In gate-level, which is the focus of this paper, cells are split into two tiers where the MIVs are used only for inter-cell connections. For block-level M3D, higher level functional blocks are floorplanned into separate tiers instead with the lowest granularity of MIV connections. This paper uses the gate-level M3D implementation because it allows sufficiently high integration density using standard cells to obtain significant power and area benefits.

FinFET technology offers faster switching times with lower leakage currents and variability, to realize the potential benefit of technology scaling. In this technology advance, M3D implementations in FinFET technology have not been widely explored. Recently, a study on the benefits of transistor-level M3D on a 7nm library has been investigated in [5]. However, their work focused on transistor-level metrics for few low drive strength cells. It is important to consider standard cell design and model wire parasitics, in a full-chip M3D implementation to get a complete and accurate estimate of the impact of M3D design using FinFETs. Figure 1 shows an illustration of monolithic 3D IC structure based on FinFET technology.

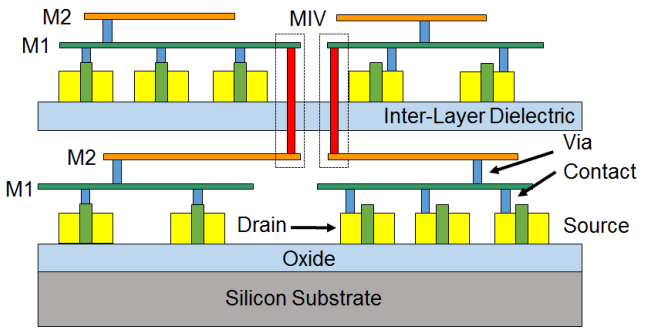

Figure 1: An illustration of monolithic 3D IC based on FinFET and monolithic inter-tier via (MIV) technologies. The tier-to-tier distance is typically 100nm, and the diameter of MIV is 50nm [2].

In this paper, we perform a comprehensive study of the power, performance and area benefits of M3D designs at 7nm technology node. We develop a predictive 7nm standard cell library of 122 high-performance (HP) and low stand-by power (LSTP) cells, using FinFET transistors. Using this library, we implement both 2D and gate-level M3D full-chip GDSII layouts of benchmark designs and perform a detailed iso-performance comparison of design metrics to understand the impact of M3D on 7nm FinFET designs.

2. 7nm Library Generation

Two 7nm standard cell libraries - one for high performance (HP) and one for low stand-by power (LSTP) are characterized with a total of 122 cells each using *Synopsys SiliconSmart* as outlined in [7]. In this library

generation methodology, the authors incorporate the effects of FinFET technology into the 7nm libraries by scaling down the dimensions from the NanGate 45nm library with commercial-grade EDA tools. Then, in the 7nm generated netlist, the planar MOSFETs are converted to their equivalent FinFET models by replacing the planar widths with integer numbers of fins. Taking the dummy fins into account [3] and the maximum number of fins in a cell, the planar widths are divided by the fin pitch in the extracted 7nm netlists, and the new netlists which utilize FinFETs are generated.

Using the new netlists and ASU PTM-MG FinFET transistor models [1] for the equivalent FinFET models, both the HP and LSTP 7nm libraries are generated and used for full-chip M3D design implementations. Due to the decrease in cell delay, reduced V_{DD}, and smaller input capacitance caused by reduced dimension, the internal PDP of the 7nm library cells are reduced significantly from previous technology nodes. The 7nm LSTP library has longer cell delay compared to 7nm HP library because of lower leakage transistors, but the internal PDP of LSTP cells is lower than HP cells. Both libraries also include the *Tch* and *CapTbl* files that are used for RC model extraction during placement and routing as well as final timing closure.

3. Full-Chip M3D Design Flow

The full-chip implementation for M3D designs using each library follows the CAD methodology outlined in [6]. In this CAD methodology, the authors exploit the fact that in monolithic 3D ICs the z dimension is negligible, which enables the use of commercial 2D tools to perform place and route for M3D. This allows a 2D placer to be used to place all the gates in a *monolithic 3D* IC footprint that is half the footprint area of a 2D IC. Placing all the cells in half the area is accomplished by shrinking the area of each standard cell by $1/\sqrt{2}$ (0.707), including the location of all pins within the cell. The chip width and height are scaled by 0.707 to reduce the 2D footprint area by half, which becomes the footprint area for each tier in the final M3D design. The overall CAD methodology flow to generate M3D design is shown in Figure 2.

All the design processing steps of placement, post-placement optimization, clock tree synthesis (CTS), routing, and post-route optimizations are performed on this shrunk 2D design in *Cadence Encounter*. In order to accurately represent the routing in monolithic 3D, the metal width and pitch is shrunk by 0.707, but the RC per unit length is unchanged. This allows the extracted RC values from the tool to represent the final M3D routing, using the original metal geometries, once the shrunk 2D design is scaled back up to the original size.

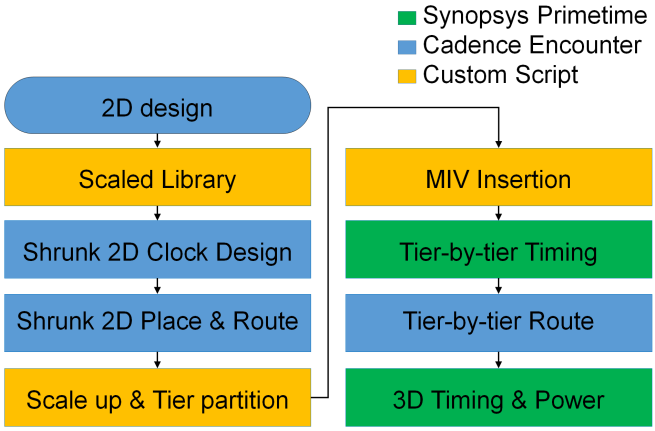

Figure 2: The CAD methodology flow to generate M3D design from 2D design used in [6].

Once an optimized shrunk 2D design is obtained with half the footprint area as a 2D IC, the cells in the resulting shrunk 2D design are then scaled up to the original size causing overlaps in the 2D design. This overlapped 2D design is split into two tiers using a modified Fiduccia-Mattheyses [4] min-cut partitioner, so that half of the cells are located in tier 0, and the other half in tier 1. During partitioning the chip is tiled into partition bins and the area balance is maintained within each partition bin, instead of area balance in the whole chip. Changing the bin size changes the partitioner constraints, and hence, the number of MIVs. In addition, during the splitting of tiers, we ensure that an adequate clock tree is built.

During MIV insertion, we utilize a 2D router that can route pins on multiple metal layers. First, all metal layers within the cells are duplicated, thereby generating a new 3D LEF. Then, we define two different types of cells, one for each tier. Pins of each cell type are mapped onto different layers depending on the tier (e.g. tier 0 type of a cell utilizes the original metal layers, tier 1 type the duplicated metal layers). All the cells in tier 0 and tier 1 are mapped on the corresponding type, and forced onto the same placement layer. This structure is fed into *Cadence Encounter*, and routed. Then, the locations of MIVs are determined, and the separate designs for each tier are generated. In our experiments, MIV diameter is 25*nm*, MIV resistance is 16Ω, and MIV capacitance is 0.1*fF*. Sample MIV maps for the same region designed with two different MIV counts are shown in Figure 3.

Once the MIV locations are determined, each tier is trial routed and estimated parasitics are generated for each tier. In addition, a top-level netlist and parasitic file is created that contains the MIV connectivity. The netlists for each tier, along with its parasitics are fed into *Synopsys PrimeTime* to perform an initial timing analysis and generate tier-by-tier timing constraints. These constraints are used to run tier-by-tier timing-driven routing, and the extracted parasitics are fed back into *PrimeTime* for final timing and power numbers.

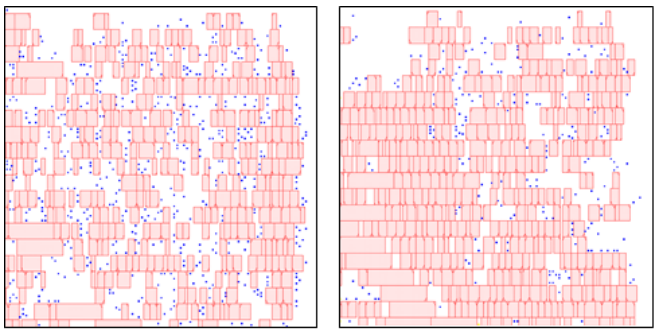

Figure 3: Zoom-in shots of 7nm HP M3D implementation of AES with two different MIV (=blue dots) counts (a) Total #MIVs = 99.3K and (b) Total #MIVs = 35.8K.

4. 7nm M3D Design Analysis

4.1. Benchmark Designs

For our benchmark 2D and M3D implementations, we chose five designs with varying gate counts. These are LDPC decoder, AES Encryption Module, RCA array, FFT core, JPEG Encoder which contain increasing gate counts, with JPEG having the highest gate count. For iso performance comparison, we first design the 2D implementation for each benchmark design to determine the best target frequency and use the same frequency target to implement the corresponding M3D design. For each benchmark design, we implement two flavors of both 2D and M3D implementations, one using the 7nm HP library and the other using the 7nm LSTP library. Table 1 shows the benchmark designs we use for this study with their 2D cell and net counts with the target frequency for iso-performance comparison.

Table 1: Our benchmark designs.

7nm HP					
	LDPC	AES	RCA	FFT	JPEG
cells	40,864	140,566	149,445	359,657	409,484
nets	45,598	142,367	159,609	361,045	452,602
target clock (*ns*)	0.850	0.200	0.200	0.250	1.000
7nm LSTP					
	LDPC	AES	RCA	FFT	JPEG
cells	41,024	135,086	114,752	362,026	406,202
nets	45,725	136,836	127,837	363,581	450,656
target clock (*ns*)	2.000	0.400	0.400	0.500	1.500

In our M3D CAD implementation, since the 2D footprint is scaled by 0.707, and then split into two tiers, the final M3D footprint is reduced by almost 50%, leading to significant wire-length and footprint area savings over its 2D counterpart. GDSII die shots of AES implemented in the HP library are shown in Figure 4. As seen from this figure, even sub-regions of designs without a lot of global interconnects can benefit from M3D, with reduced wiring.

4.2. Design Tradeoff Study

The metrics used in this study can be broadly classified into two categories - first, the design metrics that include silicon area, wire-length, and buffer/cell counts, second the power metrics that report the various components of power consumption, across both the 2D and M3D implementations. All the metrics reported are after final placement and routing for both 2D and M3D designs. Table 2 shows the design metrics for 2D and M3D implementations of all the benchmark designs in 7nm HP and 7nm LSTP library. Table 3 shows the comparison of the power metrics for 2D and M3D implementations in 7nm HP and 7nm LSTP. We summarize our findings as follows:

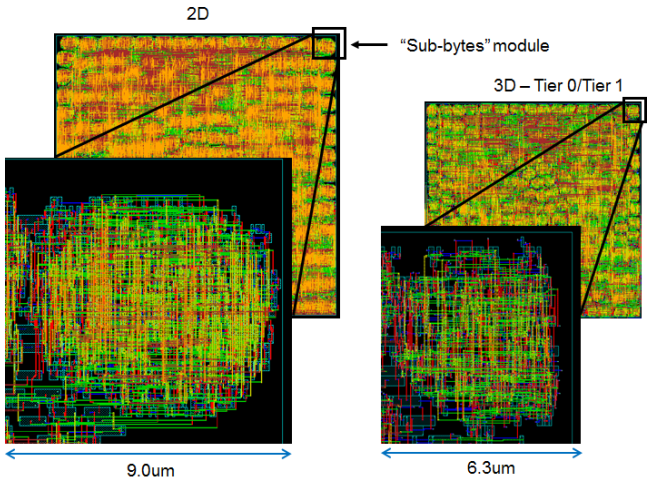

Figure 4: 2D and 3D implementations of AES benchmark in the HP library. Due to the small size of MIVs, even small locally connected modules such as the "sub-bytes" in AES can be folded, and reduced wiring density is observed.

Area and Wire-length: Our M3D designs show smaller footprint area compared to its 2D counterparts by 50% on average for both HP and LSTP designs. For almost the same silicon area, this reduction in footprint and the MIV insertion for M3D results in significant wire-length reduction by 20% on average for both HP and LSTP, across all the benchmark designs. The shorter wire-length also reduces the wire parasitics that is a big portion of the total power in 7nm designs. This helps meeting timing closure at the target frequency much easier in M3D, compared to 2D designs.

Buffer and Sizing: Though the signal buffer count is not reduced significantly in proportion to the wire-length reduction, the standard cell area utilized by M3D is lower than in 2D. In some cases, for LDPC, which is a wire-cap dominated design, the buffer count actually increases. This is because, the M3D designs, even when they use more buffers, tend to use smaller drive-strength cells compared to it 2D counterparts. With a reduction in wire-length, the M3D designs can meet their timing constraints, and drive the load capacitance with smaller drive-strength cells. It is evident that M3D design uses smaller cell sizes which have

lesser internal and leakage power. M3D designs have more cells with X1 drive-strength, and cell usage reduces significantly in M3D designs as we go from X4 to X32 variants. Both the observations are supported by the reduction in leakage and cell internal power in the benchmarks.

Power Consumption: The 7nm HP benchmarks shows on average a total power savings of 12%, while the 7nm LSTP benchmarks shows on average a total power savings of 10%. We represent the net switching power with two components - gate cap power and wire cap power, where the gate cap power is determined as the portion of net switching power that can be attributed to the total pin capacitance in the design, while the wire cap power comes from the total wire capacitance. Since switching power is directly proportional to capacitance, assuming uniform switching activity, we use the ratios of pin and wire capacitance to the total capacitance to determine the equivalent distribution of wire-length, with the gate cap power reducing in a smaller proportion. The total power is further split into cell internal power and leakage power. All the benchmarks show a consistent reduction in wire cap power of 18% on average for both HP and LSTP. The changes in design metrics also affect power consumption of the designs, as presented in Table 3. As the wire-length of designs is reduced, the net switching power of M3D designs is reduced as well. The overall cell and buffer counts remain unaffected or increase in some cases for the M3D implementation versus their 2D counterparts, but the reduction in wire-length translates to a significant reduction of the wire cap power.

We notice that comparing M3D designs between the 7nm HP and LSTP cells, both the implementations have total power savings. For the LSTP M3D designs, we see similar wire-length reduction, with an overall power saving, as the standard cells used are inherently low power and slower than the HP cells, leading to slower operating frequencies. The 7nm M3D designs are using both same-sized HP and LSTP cells, but on average the standard cell area utilized by M3D is lower than in 2D.

The LDPC benchmark displays an interesting phenomenon - it being a wire-cap dominated design, both the wire-length and clock buffer count reduction contribute to a significant reduction in both net switching power and internal power. These metrics show that M3D technology can be a good solution for power reduction in advanced technology nodes.

Clock Power Impact: The clock buffer count does have an impact on the total power consumption of the M3D designs. Table 4 shows the benchmark clock metrics of clock buffer count, clock wire-length, and total clock power for both HP and LSTP designs. In our baseline benchmark M3D designs, the clock network is first built on the shrunk 2D design and then partitioned across both tiers.

Table 4 shows total clock power savings of 10% on average for 7nm HP M3D implementations and 7% on average for 7nm LSTP implementations. The clock wire-length reduces by 28% on average for both HP and LSTP, while the clock buffer count reduces by an average of 27% for HP and 15% for LSTP. The reduction in wire-length translates to greater savings in net switching power. For benchmark designs that have a comparable proportion of clock power to the total power consumption, the clock buffer count reduction in turn gives an overall power reduction in the M3D implementation. For example, the JPEG benchmark design in 7nm HP, shows 42% reduction in clock buffer count for M3D, which results in a significant reduction in clock internal power. Since in this design, the clock power is more than 50% of the total power consumption, the overall power is reduced by 16.9%.

4.3. 3D Clock Tree Partitioning Impact

In this section, we explore the power benefit of different clock partitioning techniques in our 3D clock tree synthesis methodology shown in Figure 2. As outlined in our CAD methodology, we build the clock tree during the shrunk 2D design phase, once we scale up the design for partitioning into two tiers, we split the clock tree similar to the signal nets, allowing the flip-flops and clock buffers to be partitioned across both tiers depending on area balance within the partition bins. During the partitioning phase, the clock tree is traced from the source clock pin to the sink flip-flops and we perform a topological sort to determine the number of levels in the tree starting from the sink flip-flops. We label the sink flip-flops as *Level 0*, traversing down to the root clock pin in incremental *Levels 1, 2,* and so on. This tracing function is embedded into our custom 3D placement engine that uses the modified FM min-cut algorithm [4] to partition the clock tree into the two tiers. In order to obtain a fine grained control of the clock tree partitioning, we introduce a mechanism to fix varying levels of the clock tree onto a single tier and study their impact on the clock power. Figure 5 shows the different levels in the clock tree.

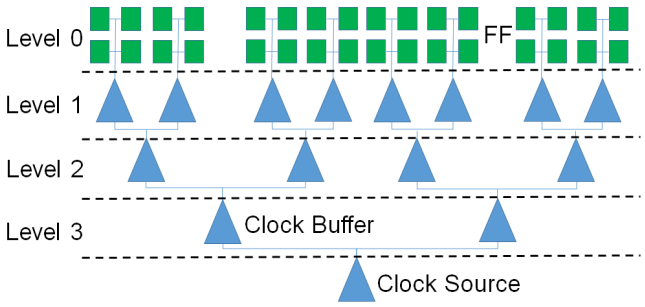

Figure 5: M3D clock tree partitioning, where varying levels of the clock tree are fixed on a single tier.

We study the impact of fixing varying degrees of the clock tree levels on a single tier, say tier 0, thereby changing the number of clock MIVs in the final M3D designs. In our baseline runs, a single MIV is inserted into each individual clock net and the entire clock backbone is partitioned. First, we fix *all levels* of the clock tree on tier 0, with no MIVs being used for clock routing. Second, we fix only the clock cells on tier 0, with the FFs free to be partitioned i.e. *source-to-L1* is fixed on tier 0 and lastly, we fix all the clock cells from source to *Level 2* on tier 0 i.e. *source-to-L2*. Table 5 shows the impact of three different settings for fixing the clock tree on tier 0.

Table 2: Comparison of 2D and M3D **design metrics** in 7nm HP and LSTP libraries. The percentage values in M3D designs are with respect to their 2D counterparts.

Design	Metric	7nm HP 2D	7nm HP M3D (Δ% 7nm HP 2D)		7nm LSTP 2D	7nm LSTP M3D (Δ% 7nm LSTP 2D)	
LDPC	footprint (μm^2)	120x120	85x85	(-49.8%)	120x120	85x85	(-49.8%)
	silicon area (μm^2)	14,400	14,450	(0.3%)	14,400	14,450	(0.3%)
	cell count (no buffer)	33,067	33,067	(0.0%)	33,448	33,410	(-0.1%)
	signal buffer count	**6,621**	**7,640**	**(15.4%)**	**5,580**	**6,554**	**(17.5%)**
	wire-length (μm)	**738,169**	**510,285**	**(-30.9%)**	**721,456**	**550,987**	**(-23.6%)**
	MIVs	-	21,476		-	20,523	
AES	footprint (μm^2)	145x145	103x103	(-49.5%)	145x145	103x103	(-49.5%)
	silicon area (μm^2)	21,025	21,218	(0.9%)	21,025	21,218	(0.9%)
	cell count (no buffer)	111,823	112,146	(0.3%)	107,899	107,896	(0.0%)
	signal buffer count	**25,433**	**26,180**	**(2.9%)**	**24,238**	**24,187**	**(-0.2%)**
	wire-length (μm)	**507,291**	**398,913**	**(-21.4%)**	**504,528**	**394,967**	**(-21.7%)**
	MIVs	-	46,911		-	47,554	
RCA	footprint (μm^2)	136x136	96x96	(-50.2%)	136x136	96x96	(-50.2%)
	silicon area (μm^2)	18,496	18,432	(-0.3%)	18,496	18,432	(-0.3%)
	cell count (no buffer)	96,722	95,475	(-1.3%)	101,089	101,092	(0.0%)
	signal buffer count	**51,862**	**51,309**	**(-1.1%)**	**10,774**	**12,525**	**(16.3%)**
	wire-length (μm)	**367,148**	**318,509**	**(-13.2%)**	**302,613**	**252,458**	**(-16.6%)**
	MIVs	-	34,426		-	28,695	
FFT	footprint (μm^2)	250x250	175x175	(-51.0%)	250x250	175x175	(-51.0%)
	silicon area (μm^2)	62,500	61,250	(-2.0%)	62,500	61,250	(-2.0%)
	cell count (no buffer)	240,251	242,023	(0.7%)	213,135	211,997	(-0.5%)
	signal buffer count	**85,382**	**87,668**	**(2.7%)**	**88,854**	**87,043**	**(-2.0%)**
	wire-length (μm)	**1,281,805**	**1,061,463**	**(-17.2%)**	**1,252,813**	**1,041,298**	**(-16.9%)**
	MIVs	-	85,391		-	87,613	
JPEG	footprint (μm^2)	291x291	200x200	(-52.8%)	291x291	200x200	(-52.8%)
	silicon area (μm^2)	84,681	80,000	(-5.5%)	84,681	80,000	(-5.5%)
	cell count (no buffer)	269,986	269,929	(0.0%)	288,853	282,386	(-2.2%)
	signal buffer count	**98,867**	**99,832**	**(1.0%)**	**92,640**	**92,890**	**(0.3%)**
	wire-length (μm)	**1,541,322**	**1,217,661**	**(-21.0%)**	**1,920,913**	**1,496,580**	**(-22.1%)**
	MIVs	-	125,961		-	130,371	

When the entire clock tree is fixed on tier 0, it gives the lowest clock power and clock skew, with zero clock MIVs being used. The tradeoff is that the standard cell area is now unbalanced between tier 0 and tier 1. This is because the larger size flip-flops skew the area balance for the M3D designs with this setting. This forces all the sequential logic to be 2D, while the combinational logic is partitioned into two tiers, leading to long routes for the signal nets and possibly more wire-length. The Table 5 shows that as fewer clock cells are fixed, the clock MIV count goes up and has a degrading effect on clock power. For our benchmarks, the best option is to only partition the FFs and fix all the rest of the cells on tier 0.

5. Key Findings

In our study we show the benefits of monolithic 3D ICs (M3D) at 7nm in terms of area, wire-length, power consumption, and frequency. We summarize our key findings as follows:

Major source of 7nm M3D power saving: Looking at the different components of the total power, such as net switching power, cell internal power, and leakage power, we observe that the wire-length reduction results in net switching power reduction, while buffer count reduction affects the internal power and leakage power reduction in M3D. The drive-strength usage distribution of the cells used in M3D implementation also has an impact on the leakage power reduction.

Clock power saving in high performance applications: We observe that for designs where the clock power is a substantial contributor to the total power consumption, especially in advanced technologies, the monolithic 3D implementation shows significant reduction in clock buffer count along with wire-length reduction. This enables higher power savings with a more fine-tuned clock tree design. We show that with a careful consideration of clock network design in monolithic 3D IC, the clock buffer count and clock wire-length can be improved, which in turn results in

Table 3: Comparison of 2D and M3D **power metrics** in 7nm HP and LSTP libraries. The percentage values in M3D designs are with respect to their 2D counterparts.

Design	Metric	7nm HP 2D	7nm HP M3D (Δ% 7nm HP 2D)		7nm LSTP 2D	7nm LSTP M3D (Δ% 7nm LSTP 2D)	
LDPC	frequency (*MHz*)	1,176	1,176	(0.0%)	500	500	(0.0%)
	wire cap power (*mW*)	14.168	9.290	(-34.4%)	6.062	4.352	(-28.2%)
	gate cap power (*mW*)	3.132	3.110	(-0.7%)	1.145	1.117	(-2.4%)
	internal power (*mW*)	5.158	3.918	(-24.0%)	0.857	0.819	(-4.4%)
	leakage power (*mW*)	0.295	0.283	(-4.1%)	0.001	0.001	(0.0%)
	total power (*mW*)	**22.800**	**16.600**	**(-27.2%)**	**8.065**	**6.289**	**(-22.0%)**
AES	frequency (*MHz*)	5,000	5,000	(0.0%)	2,500	2,500	(0.0%)
	wire cap power (*mW*)	8.334	6.832	(-18.0%)	4.942	4.004	(-19.0%)
	gate cap power (*mW*)	14.466	13.868	(-4.1%)	4.374	4.022	(-8.0%)
	internal power (*mW*)	20.100	19.400	(-3.5%)	7.037	6.913	(-1.8%)
	leakage power (*mW*)	1.728	1.679	(-2.8%)	0.003	0.003	(0.0%)
	total power (*mW*)	**44.600**	**41.800**	**(-6.3%)**	**16.400**	**14.900**	**(-9.1%)**
RCA	frequency (*MHz*)	5,000	5,000	(0.0%)	2,500	2,500	(0.0%)
	wire cap power (*mW*)	3.419	3.002	(-12.2%)	1.512	1.237	(-18.2%)
	gate cap power (*mW*)	5.568	4.888	(-12.2%)	2.014	1.838	(-8.7%)
	internal power (*mW*)	7.171	6.868	(-4.2%)	2.373	2.332	(-1.7%)
	leakage power (*mW*)	1.311	1.248	(-4.8%)	0.005	0.005	(0.0%)
	total power (*mW*)	**17.500**	**16.000**	**(-8.6%)**	**5.904**	**5.412**	**(-8.3%)**
FFT	frequency (*MHz*)	4,000	4,000	(0.0%)	2,000	2,000	(0.0%)
	wire cap power (*mW*)	33.624	30.404	(-9.6%)	19.101	17.589	(-7.9%)
	gate cap power (*mW*)	46.776	46.796	(0.0%)	18.799	18.811	(0.1%)
	internal power (*mW*)	90.500	90.200	(-0.3%)	48.900	46.600	(-4.7%)
	leakage power (*mW*)	4.590	4.458	(-2.9%)	0.014	0.013	(-7.1%)
	total power (*mW*)	**175.400**	**171.900**	**(-2.0%)**	**86.900**	**83.000**	**(-4.5%)**
JPEG	frequency (*MHz*)	1,000	1,000	(0.0%)	666	666	(0.0%)
	wire cap power (*mW*)	12.981	10.913	(-15.9%)	10.068	8.091	(-19.6%)
	gate cap power (*mW*)	13.819	13.287	(-3.8%)	6.832	6.709	(-1.8%)
	internal power (*mW*)	30.800	23.200	(-24.7%)	11.400	10.900	(-4.4%)
	leakage power (*mW*)	3.952	3.755	(-5.0%)	0.013	0.012	(7.7%)
	total power (*mW*)	**61.500**	**51.100**	**(-16.9%)**	**28.400**	**25.800**	**(-9.2%)**

both internal power and net switching power reduction. As the monolithic 3D IC tiers are stacked on top of each other with MIV connections across tiers, this inherently enables a lower wire-length solution and with our shrunk 2D based CAD methodology, the final M3D footprint is half of the 2D area. Adding an efficient clock network design with reduced clock power, takes the M3D power benefit to another level in the 7nm implementations.

Physical design methods for M3D: In our M3D physical design methodology, we shrink the 2D design by 0.707 by simply scaling all the cell dimensions. This results in theoretical HPWL reduction of 29.3%. Our final M3D footprint matches closely with this theoretical estimate, which shows that the shrunk 2D design is a good measure to determine the final design metrics of our M3D designs for this fine-grained technology. Optimizing the shrunk 2D design for timing, provides a good basis for both performance and power benefit in the final M3D design.

Monolithic inter-tier via (MIV) impact: Given the footprint area reduction, the overall wire-length reduction using MIVs provides a significant power saving across both HP and LSTP applications. We also investigate the impact of numbers of MIVs used in the final M3D implementation on total power. Table 6 shows different MIV counts in 7nm HP AES implementation, and the corresponding total power metric. The numbers show a direct correlation between total power and MIV counts, where after a certain value, the total power degrades as number of MIVs increase in the M3D implementation.

Table 6: Impact of number of MIVs on total power in 7nm HP AES M3D implementation.

MIVs	11,438	20,660	**35,837**	46,726	99,375
total power (*mW*)	41.5	41.5	**41.2**	41.4	42.3

Table 4: Comparison of 2D and M3D **clock metrics** in 7nm HP and LSTP libraries. The percentage values in M3D designs are with respect to their 2D counterparts.

Design	Metric	7nm HP 2D	7nm HP M3D (Δ% 7nm HP 2D)		7nm LSTP 2D	7nm LSTP M3D (Δ% 7nm LSTP 2D)	
LDPC	clock buffer count	1,176	528	(-55.1%)	1,996	1,308	(-34.5%)
	clock FF count	2,048	2,048	(0.0%)	2,048	2,048	(0.0%)
	clock wire-length (μm)	5,305	3,956	(-25.4)	5,427	4,300	(-20.8%)
	clock net switching power (mW)	0.598	0.519	(-13.2%)	0.238	0.208	(-12.6%)
	clock internal power (mW)	0.463	0.469	(1.3%)	0.153	0.147	(-3.9%)
	clock leakage power (mW)	0.003	0.003	(0.0%)	0.000	0.000	(0.0%)
	clock total power (mW)	**1.065**	**0.991**	**(-6.9%)**	0.392	**0.354**	**(-9.7%)**
AES	clock buffer count	3,310	2,961	(-10.5%)	2,949	2,940	(-0.3%)
	clock FF count	10,688	10,688	(0.0%)	10,688	10,688	(0.0%)
	clock wire-length (μm)	17,421	13,040	(-25.1%)	19,603	12,539	(-36.0%)
	clock net switching power (mW)	11.000	10.200	(-7.3%)	5.127	4.507	(-12.1%)
	clock internal power (mW)	12.500	12.100	(-3.2%)	4.792	4.722	(-1.5%)
	clock leakage power (mW)	0.035	0.031	(-11.4%)	0.000	0.000	(0.0%)
	clock total power (mW)	**23.600**	**22.300**	**(-5.5%)**	**9.919**	**9.229**	**(-7.0%)**
RCA	clock buffer count	872	671	(-23.1%)	2,889	2,465	(-14.7%)
	clock FF count	20,480	20,480	(0.0%)	20,480	20,480	(0.0%)
	clock wire-length (μm)	8,408	5,732	(-31.8%)	8,184	5,703	(-30.3%)
	clock net switching power (mW)	3.221	2.700	(-16.2%)	1.560	1.300	(-16.7%)
	clock internal power (mW)	2.077	1.895	(-8.8%)	1.224	1.184	(-3.3%)
	clock leakage power (mW)	0.007	0.006	(-14.3%)	0.000	0.000	(0.0%)
	clock total power (mW)	**5.305**	**4.602**	**(-13.3%)**	**2.784**	**2.484**	**(-10.8%)**
FFT	clock buffer count	33,934	32,305	(-4.8%)	60,037	54,963	(-20.5%)
	clock FF count	75,555	75,555	(0.0%)	75,555	75,555	(0.0%)
	clock wire-length (μm)	116,515	86,899	(-25.4%)	116,999	89,666	(-23.4%)
	clock net switching power (mW)	65.700	63.900	(-2.7%)	32.300	31.300	(-3.1%)
	clock internal power (mW)	70.400	70.400	(0.0%)	41.800	39.600	(-5.3%)
	clock leakage power (mW)	0.352	0.329	(-6.5%)	0.002	0.001	(-50.0%)
	clock total power (mW)	**136.500**	**134.700**	**(-1.3%)**	**74.100**	**70.900**	**(-4.3%)**
JPEG	clock buffer count	40,631	23,565	(-42.0%)	24,709	19,633	(-20.5%)
	clock FF count	56,481	56,481	(0.0%)	56,481	56,481	(0.0%)
	clock wire-length (μm)	106,471	72,563	(-31.8%)	86,860	61,544	(-29.1%)
	clock net switching power (mW)	14.200	12.200	(-14.1%)	7.312	6.752	(-7.7%)
	clock internal power (mW)	21.700	14.200	(-34.6%)	6.499	6.109	(-6.0%)
	clock leakage power (mW)	0.436	0.246	(-43.6%)	0.001	0.000	(-100.0%)
	clock total power (mW)	**36.300**	**26.600**	**(-26.7%)**	**13.800**	**12.900**	**(-6.5%)**

Opportunities in low power applications: Both our 7nm HP and LSTP applications show significant wire-length and clock buffer count improvement. But, since the LSTP library has much slower cells, the operating frequency is much lower, compared to 7nm HP. The HP designs also meet their timing budget much easier than the LSTP designs. We observe that for the LSTP applications, we need more accurate timing delay values for lower output load capacitance in our 2-dimensional delay model for given input slew. This allows the *PrimeTime* tool to accurately interpolate/extrapolate the delay calculation for our LSTP designs. Moreover, since the LSTP designs are dominated by net switching power, the reduction in clock buffer count has a limited impact on the total power savings.

Benchmark characteristics for power saving: For a wire-cap dominated design such as LDPC, the net switching power dominates the total power and the increased wire-length reduction in the M3D implementation contributes more to the total power savings, than the reduction in buffer count. Overall, for both high-performance and low standby power applications in 7nm, M3D technology offers significant power and area benefit over 2D designs.

6. Conclusion

In this paper, for the first time, we demonstrated the power, performance and area impact of gate-level monolithic 3D IC designs, using our predictive 7nm standard cell libraries.

Table 5: Impact of M3D clock tree partitioning. Figure 5 shows the illustration of various options.

		HP			LSTP		
Design	Metric	all levels in one die	src-to-L1 in one die	src-to-L2 in one die	all levels in one die	src-to-L1 in one die	src-to-L2 in one die
LDPC	fixed clock cells (tier 0)	2,346	298	70	2,314	259	60
	clock wire-length (μm)	3,295	3,850	4,007	3,661	4,300	4,316
	clock MIVs	0	228	264	0	223	224
	clock skew (ns)	0.030	0.035	0.032	0.095	0.080	0.064
	total clock power (mW)	0.933	0.986	0.997	0.335	0.356	0.355
AES	fixed clock cells (tier 0)	13,541	2,853	725	13,588	2,731	292
	clock wire-length (μm)	12,239	12,359	12,529	11,958	12,045	12,811
	clock MIVs	0	1,884	1,938	0	2,193	2,126
	clock skew (ns)	0.023	0.027	0.028	0.041	0.035	0.041
	total clock power (mW)	21.600	22.100	22.100	8.935	9.199	9.250
RCA	fixed clock cells (tier 0)	21,068	570	129	21,708	1,228	402
	clock wire-length (μm)	5,285	5,697	5,725	5,524	5,608	5,621
	clock MIVs	0	357	358	0	375	423
	clock skew (ns)	0.014	0.013	0.014	0.052	0.056	0.061
	total clock power (mW)	4.391	4.620	4.623	2.398	2.490	9.250
FFT	fixed clock cells (tier 0)	107,245	31,647	5,211	129,556	53,915	17,560
	clock wire-length (μm)	75,512	81,364	83,506	86,736	86,784	87,606
	clock MIVs	0	21,236	19,450	0	26,657	22,111
	clock skew (ns)	0.033	0.032	0.034	0.067	0.061	0.065
	total clock power (mW)	125.000	132.200	132.200	67.700	71.300	70.900
JPEG	fixed clock cells (tier 0)	79,692	23,191	3,459	74,620	18,114	3,769
	clock wire-length (μm)	80,396	67,522	70,226	64,078	60,839	61,597
	clock MIVs	0	15,081	13,018	0	12,367	12,135
	clock skew (ns)	0.108	0.084	0.078	0.086	0.068	0.059
	total clock power (mW)	25.200	25.600	25.600	12.400	12.900	12.900

We built full-chip GDSII layouts of five benchmark designs using both 7nm high-performance and low standby power library cells. We analyzed the various design metrics such as footprint area, buffer count, wire-length, clock and total power to understand the full impact of M3D technology at advanced technology nodes. We presented a 3D clock tree partitioning methodology to efficiently split the clock tree between two tiers for best clock power savings. The simulation studies show that monolithic 3D IC offers significant power and area benefits over traditional 2D for future FinFET technologies.

7. References

[1] ASU-PTM. http://ptm.asu.edu/

[2] P. Batude et al. Advances in 3D CMOS Sequential Integration. In *proc. IEEE Int. Electron Device Meeting,* pages 1-4, 2009.

[3] M. Choi, V. Moroz, L. Smith, and O. Penzin. 14nm FinFET Stress Engineering with Epitaxial SiGeo Source/Drain. In *"Int. Silicon-Germanium Technology and Device Meeting",* 2012.

[4] C. M. Fiduccia and R. M. Mattheyses. A linear-time heuristic for improving network partitions. In *Proc. ACM Design Automation Conf.,* 1982.

[5] 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.

[6] S. A. 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.

[7] K. Chang et al. Power Benefit Study of Monolithic 3D IC at the 7nm Technology Node. *IEEE Int. Symp. On Lower Power Electronics and Design,* 2015.