Match-making for Monolithic 3D IC: Finding the Right Technology Node

Kyungwook Chang[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
k.chang@gatech.edu, limsk@ece.gatech.edu

ABSTRACT

Monolithic 3D IC (M3D) has the potential to provide a breakthrough in the power and performance scaling challenges. We, for the first time, present a comprehensive study of M3D on a commercial design across multiple technology nodes. The performance and power impact of M3D is investigated using a commercial, in-order, 32-bit application processor, implemented on foundry 28nm and 14/16nm process nodes, as well as a predictive 7nm node. We study the factors across the technology nodes that affect the efficiency of M3D, and propose a roadmap for optimum technology and design interaction that will enable the full entitlement of M3D.

1. INTRODUCTION

Monolithic 3D IC (M3D) is a promising technology that can overcome 2D scaling challenges involving physical limits of channel length scaling, increasing contact resistivity, lithography limitations, increased wire resistance and increasing higher manufacturing costs. In M3D, transistors are fabricated into two tiers; bottom tier and top tier. Unlike through-silicon via (TSV)-based 3D ICs, which drill cavities for TSVs that connect between tiers, in M3D, after fabricating transistors and wires on the bottom tier, transistors on the top tier are fabricated and connected by fine-pitch monolithic inter-tier vias (MIVs), sequentially. Due to the difference in fabrication methodology, it is possible to utilize MIVs with much higher density and lower parasitics compared to TSVs. Recent improvements in manufacturing technology such as higher alignment precision and the ability to process thinner dies are the stepping stones to real-world enablement of M3D.

Depending on the granularity of vertical integration, M3D technology can be categorized into three flavors; transistor-level, gate-level, and block-level M3D design. Among them, this paper focuses on gate-level M3D, utilizing conventional 2D cells to split them across two different tiers, which are then connected using MIVs. The benefit of gate-level M3D first comes from reducing wire-length in a design. Additional benefits such as reducing cell strength or decreasing the buffer insertion count (e.g., for driving long wires) are also possible.

Previous studies have explored different aspects of M3D design and shown improvement in power and performance [1]. However, most, if not all of these studies have used open source academic benchmarks and libraries to demonstrate the effectiveness of M3D technology. In spite of being a valuable research tool, academic benchmarks and libraries have limited validity and accuracy when compared to commercial designs. Recent works using commercial technology nodes include a 28nm study using open source microcontrollers as benchmarks [2] to explore important aspects of M3D design such as clock tree design and power delivery networks.

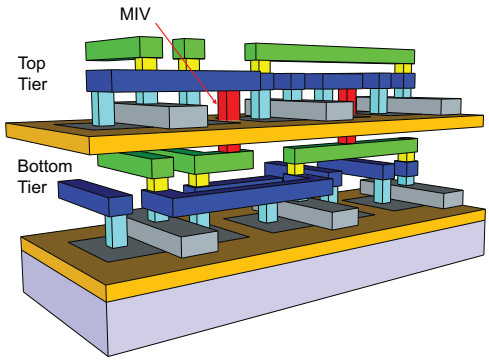

Figure 1: A schematic showing gate-level monolithic 3D IC cross-section.

The industry transitioned from planar transistors to 3D FinFETs at the 14/16nm node to combat worsening electrostatics and degraded short channel effects due to channel length scaling. Improved transistor characteristics in 3D FinFETs are achieved at the cost of higher parasitic capacitances associated with the 3D fins and the introduction of the local interconnects that are needed to contact the devices to metal routing layers. Due to limited viable transistor options beyond FinFETs and the increasing cost and complexity of lithography strategies to print sub-7nm node features, traditional Moore's law scaling is slowing down. These limitations create a technology inflection point for "More than Moore" technologies [3] such as M3D to bring value and be adopted into mainstream designs.

In order to be deployed in real-word designs, M3D needs to be cost-effective and deliver power or performance improvement of the order of magnitude similar to that obtained by "Moore's law" technology scaling. Evaluating cost-effectiveness is non-trivial as M3D is still under active research and development. Hence, this work focuses on evaluating the power/performance improvement of M3D in a real-world design - an in-order, 32-bit applicaton processor - and assessing whether or not that improvement is independent of the underlying technology node.

DAC '16, June 05-09, 2016, Austin, TX, USA

DOI: http://dx.doi.org/10.1145/2897937.2898043

Table 1: Key metrics for foundry 28nm, 14/16nm and the predictive 7nm technology node used in this study. MIV stands more monolithic inter-tier via.

Parameters	28nm	14/16nm	7nm
Transistor type	Planar	FinFET	FinFET
Supply Voltage	0.9V	0.8V	0.7V
Contacted Poly-pitch	110-120nm	78-90nm	50nm
Metal1 Pitch	90nm	64nm	36nm
MIV cross-section	80x80nm	40x40nm	32x32nm
MIV height	140nm	170nm	170nm

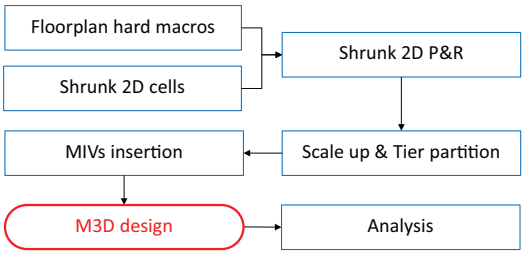

Figure 2: The CAD methodology flow for generating monolithic 3D IC (M3D) designs from commercial 2D IC designs [9].

The main contributions of this work are as follows: (1) We implement and analyze M3D designs across multiple technology nodes, namely, 28nm, 14/16nm and 7nm process nodes; (2) To the best of our knowledge, this is the first work on M3D technology that utilizes a commercial microprocessor as the benchmark and libraries and memory blocks designed for foundry processes; and (3) We present an extensive set of results explaining the factors that impact power savings and/or performance improvement in M3D and propose guidelines for optimum technology and design interaction that would enable the full entitlement of M3D.

2. EXPERIMENTAL SETUP

2.1 Process Nodes and Design Libraries

Table 1 shows the representative metrics for each process technology used in our study, based on previous publications [4, 5, 6, 7, 8]. The 28nm process is planar transistor based while 14/16nm is the first generation foundry FinFET process. For these nodes, we have used production level standard cell libraries containing over 1,000 cells and memory macros that were designed, verified and characterized using foundry process design kits (PDK).

Since the 7nm technology node parameters are still under development by foundries, we utilized a predictive PDK to generate the required views for this study. We have developed the predictive 7nm PDK containing electrical models (BSIM-CMG), DRC, LVS, extraction and technology library exchange format (LEF) files. The transistor models incorporate scaled channel lengths and fin-pitches and increased fin-heights compared to previous technology nodes in order to improve performance at lower supply voltages. Multiple threshold voltages (V_T) and variation corners are supported in the predictive 7nm PDK. Process metrics such as gate pitch and metal pitches are linearly scaled from previous technology nodes and design rules are created considering lithography challenges associated with printing these pitches. The interconnect stack is modeled based on similar scaling assumptions. A 7nm standard cell library and memory macros are designed and characterized using this PDK.

The M3D design requires six metal layers on both top and bottom tiers. The MIVs connect M6 of the bottom tier with M1 of the top tier. We limit the size of the MIVs to be 2x the minimum via size allowed in the technology node to reduce MIV resistance. The MIV heights take into account the fact that the MIVs need to traverse through inter-tier dielectrics and transistor substrates to contact to M1 on the top tier. The MIV height increases from 28nm to 14/16nm and 7nm technology nodes because of the introduction of local interconnect middle-of-line (MOL) layer in the sub-20nm nodes.

Since M3D fabrication is done sequentially, high temperature front-end device processing of the top tier can adversely affect the interconnects in the bottom tier while low temperature processing will result in inferior top tier transistors. Recent work reporting low temperature processes that achieve similar device behavior across both tiers have been presented [10] and hence, all our implementation studies are done with the assumption of similar device characteristics in both the tiers.

2.2 Monolithic 3D Design Flow

The methodology described in [9] to implement M3D designs involves using a shrunk 2D design to transform the original 2D design into a 2-tier, gate-level M3D design using commercial EDA tools. The design flow is shown in Figure 2. The first step involves determining the floorplan of hard macros like memory blocks for M3D design. They can be placed in either or both the tiers. Next, we create the floorplan for shrunk 2D design which includes creating cell-blockage on locations where both tiers have hard macros and partial blockage if one of the tier contains hard macros. In the final M3D design, the cells in the partial blockage area will be moved to the tier which does not contain the hard macros.

Then, the x-y dimensions of all cells are scaled down by a factor of $(1/\sqrt{2})$, so that all cells are placed in half the area in the original 2D design. With shrunk cells and shrunk 2D floorplan, all the design stages including placement, post-placement optimization, clock-tree-synthesis (CTS), routing, and post-route optimization are performed using Cadence® Innovus™, creating a shrunk version of the 2D design.

The cell instances in the resulting shrunk 2D design are then scaled up to the original size, creating overlaps. Then, an area-balanced min-cut bi-partitioning algorithm is run on the layout to determine the location of cell instances, so that the area of the bottom and top tier is balanced, and the number of connection (MIVs) between two tiers is minimized.

The BEOL metal stack is duplicated to account for six metal layers in each tier. Additionally, the cells are annotated with their respective tiers. Every cell has pins on metal layers of their corresponding tier. Depending on the location and tier of the cells, they are placed on the corresponding placement layer. The design is routed using the two metal stacks using Cadence® Innovus™. The location of MIVs are determined by the location of vias between M6 of the bottom tier and M1 of the top tier.

Then, using the netlists of each tier and MIV location, we perform trial-route, and the initial routed design is fed to Synopsys PrimeTime® to obtain timing constraints for each tier. Once the timing constraints are determined, we run timing-driven routing, and the results are analyzed with Synopsys PrimeTime®.

2.3 Implementation Methodology

The standard cell libraries and memory macros for the 28nm, 14/16nm and 7nm technology nodes are used to synthesize, place and route the full-chip design. 2D and M3D designs of the application processor are implemented sweeping the target frequency from 500MHz to 1.2GHz in 100MHz increments across the three

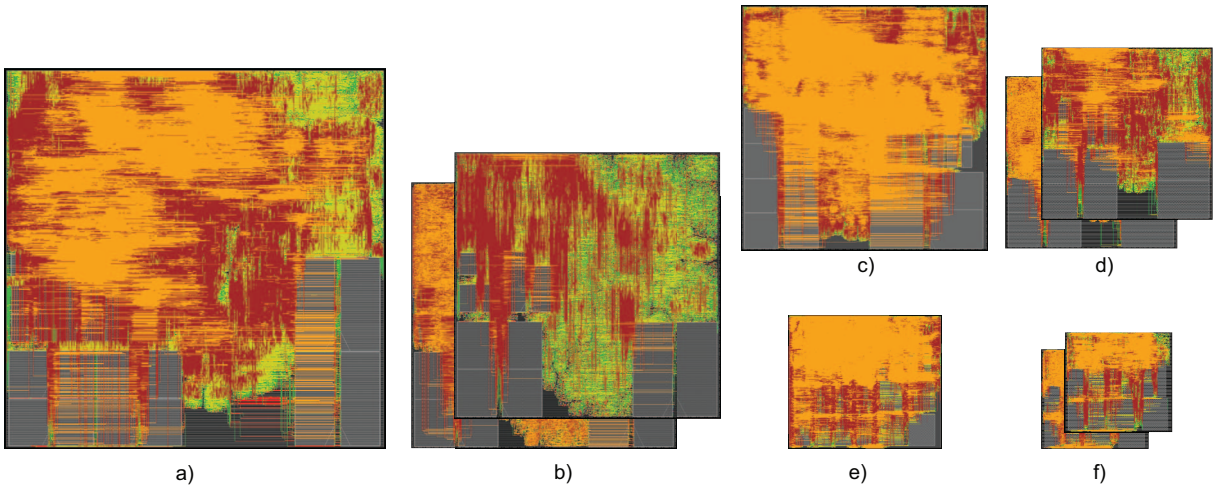

Figure 3: GDS layouts of a) 28nm 2D, b) 28nm M3D, c) 14/16nm 2D, d) 14/16nm M3D, e) 7nm 2D and f) 7nm M3D of the application processor at 1.1GHz. We use foundry PDKs except 7nm and a commercial microprocessor benchmark in our designs.

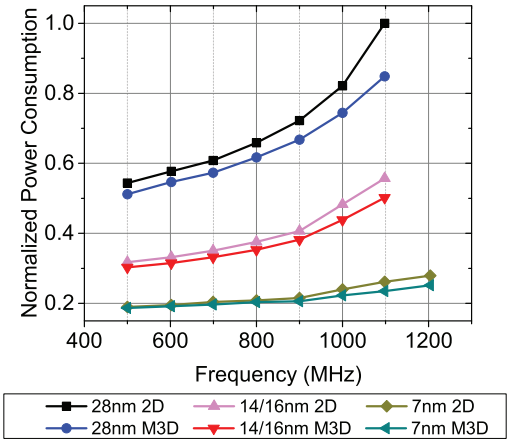

Figure 4: Normalized total power consumption of 2D and M3D designs across technology nodes.

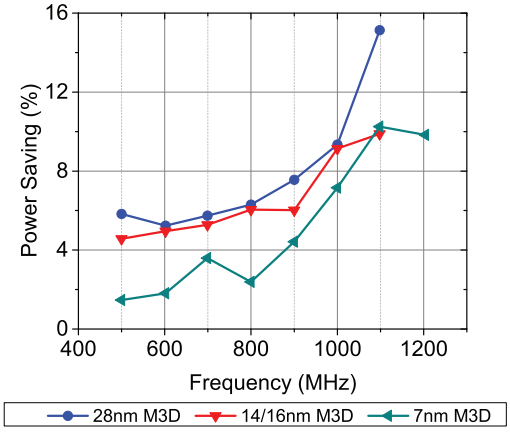

Figure 5: M3D power saving over 2D in foundry 28nm, 14/16nm, and a predictive 7nm technology nodes.

technology nodes. Full-chip timing is met at the appropriate corners, i.e., slow corner for setup and fast corner for hold. Power is reported at the typical corner. The floorplan of the design is customized for each technology node to meet timing but kept constant during frequency sweeps. Multiple iterations of the 2D and M3D floorplan are required at each node to ensure that the design meets timing. The chip area is fixed such that the final cell utilization is similar across technology nodes.

3. RESULTS AND ANALYSIS

Figure 3 shows the die images of 2D and M3D implementations of the application processor after completion of routing at 1.1GHz. The implementation tools are unable to meet timing at 1.2GHz target frequency for the 28nm and 14/16nm designs, hence we report their results up to 1.1GHz, and 7nm results are reported up to 1.2GHz.

3.1 M3D Power Saving Trend

The normalized total power consumption of the 2D and M3D designs across technologies are shown in Figure 4. Both 2D and M3D total power increases with increasing frequency across technology nodes which is expected. The power saving with M3D design over 2D is shown in Figure 5. There are two important trends in Figure 5: 1) 28nm node shows the maximum power savings in M3D design across all frequencies and 2) M3D power savings increase with increasing target frequency of the designs.

To interpret and analyze the results, we use Equation 1 which describes the components of dynamic power in a 2D design.

$$\begin{aligned} P_{dyn} &= P_{INT} + \alpha \cdot (C_{pin} + C_{wire}) \cdot {V_{DD}}^2 \cdot f_{clk} \\ &= P_{INT} + \alpha \cdot (r_{p2w} \cdot C_{wire} + C_{wire}) \cdot {V_{DD}}^2 \cdot f_{clk} \end{aligned} \quad (1)$$

The first term P_{INT}, is internal power of the gates and the second term describes switching power where C_{pin} is the pin capacitance of the gates, C_{wire} is the wire capacitance in the design, α is the activity factor, f_{clk} is the design clock frequency and r_{p2w} is the ratio of the pin capacitance to the wire capacitance. The primary advantage of M3D design comes from wire-length reduction resulting in reduced wire-switching power dissipation. With the reduction in wires, the synthesis, place and route (SP&R) tools can also reduce the drive strengths of the gates and buffers

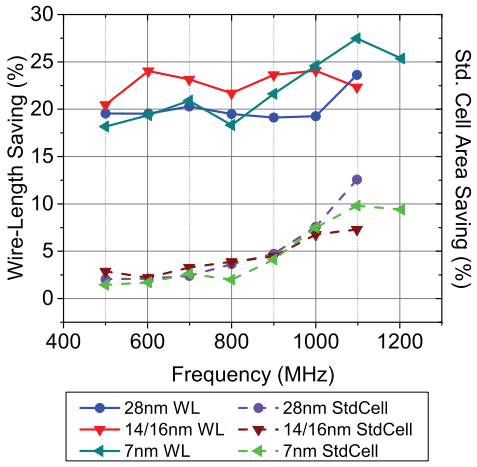

Figure 6: Impact of M3D technology on wire-length (solid lines) and standard cell area (dotted lines) savings across technology nodes.

used to meet the design targets leading to reduced internal power (P_{INT}) and pin capacitance switching component as well. The total power reduction in an M3D design depends on wire-length reduction, gate size/number reduction, the ratio of gate-cap versus wire-cap or switching power versus internal power in the 2D design.

Extending Equation 1, as internal power and pin capacitance depends on standard cell area, and wire-length affects wire capacitance, we can come up with an expression to denote M3D power savings as

$$\begin{aligned}\Delta P_{dyn} =& \Delta_{cell} \cdot (P_{INT} + \alpha \cdot r_{p2w} \cdot C_{wire} \cdot {V_{DD}}^2 \cdot f_{clk}) \\ &+ \Delta_{wire} \cdot \alpha \cdot C_{wire} \cdot {V_{DD}}^2 \cdot f_{clk}\end{aligned} \quad (2)$$

where Δ_{cell} denotes the standard cell area saving from M3D to 2D, and Δ_{wire} denotes wire-length saving in the M3D design. This simple linear model gives useful insight in explaining the power saving trends across technology nodes and frequencies.

3.2 Analysis of Results

As can be seen from Figure 6, at a given frequency, wire-length saving (Δ_{wire}) as well as standard cell area saving (Δ_{cell}) is nearly the same across all the three technology nodes.

As the clock frequency is swept, wire-length saving (Δ_{wire}) does not vary by a large magnitude, ranging between 20 to 25% as shown in Figure 6. However, with increasing clock frequency, 2D designs utilize more buffers and higher drive strength cells to meet timing whereas M3D designs can meet timing with lesser number of buffers and lower drive strength cells because of wire-length saving. Hence standard cell area saving (Δ_{cell}) increases from 2% up to 10-12% with increasing frequency. With these observations, we modify Equation 2 to denote Δ_{cell} as a function of f_{clk} in order to reflect the impact of frequency on standard cell area savings.

$$\begin{aligned}\Delta P_{dyn} =& \Delta_{cell}(f_{clk}) \cdot (P_{INT} + \alpha \cdot r_{p2w} \cdot C_{wire} \cdot {V_{DD}}^2 \cdot f_{clk}) \\ &+ \Delta_{wire} \cdot \alpha \cdot C_{wire} \cdot {V_{DD}}^2 \cdot f_{clk}\end{aligned} \quad (3)$$

3.2.1 M3D Power Saving at Low Frequency

At low frequencies (500MHz), Δ_{cell} is small (2.5%) while Δ_{wire} is much higher. Hence most of the power saving in M3D design comes from reduction in wire-switching power. Figure 7 a) shows normalized power components of the 2D and M3D design across technology nodes at the minimum (500MHz) and maximum (1.1/1.2GHz) frequencies. This figure clearly shows that internal power is the dominant portion of the total power in our design accounting for nearly 50% of the total across all frequencies and technology nodes. The rest is split between pin capacitance switching and wire capacitance switching with leakage power taking up the smallest portion. It is important to note that the pin capacitance switching power and internal power are both related to the number and size of gates used in the design. Power saving due to reduction in wire-switching power is determined by r_{p2w} in the design. Hence, even with 20-25% wire length reduction, the total power saving at low frequencies ranges between 1.5% for 7nm node to 6% for the 28nm node because wire-switching power is a small portion of the total power. The 28nm M3D design has better power saving at 500MHz because it has a larger wire capacitance to pin capacitance ratio as shown in Figure 8.

This difference in pin capacitance versus wire capacitance from 28nm to 14/16nm node can be attributed to the difference in gate capacitance associated with planar and FinFET transistors. FinFET based technologies have higher gate capacitance due to the 3D fin structure and the introduction of local interconnect middle-of-line (MOL) layers that contact the device terminals to M1. This observation that planar transistor based designs are more likely benefit from M3D compared to FinFET based designs at advanced nodes, is a key finding of this study.

Another point to note is that with technology scaling wire RC, especially resistance, increases per unit length. Improving drive strength of transistors at advanced nodes like 7nm is extremely challenging. As the ratio of transistor drive versus wire load decreases at scaled nodes, implementation tools end up using larger cells to drive the same wire-length, hence, effectively increasing r_{p2w}. Hence, technologies with larger transistor fan-outs will benefit more from M3D designs.

3.2.2 M3D Power Saving at High Frequency

At high operating frequencies, as Δ_{cell} increases it affects both pin capacitance switching power and internal power. As evident from Figure 7, internal power and pin capacitance switching power can contribute up to 70% of the total power of the 2D design at high frequencies. Hence the total power savings at maximum frequencies approach 10% or more as M3D designs benefit from reduction in all power components, predominantly internal power and pin capacitance switching power.

In order to understand the impact of frequency on M3D power savings, we can consider a hypothetical scenario when, with increasing clock frequency, $\Delta_{cell}(f_{clk}) = \Delta_{wire}$. At this frequency point, Equation 3 can be modified to the following expression

$$\Delta P_{dyn} = \Delta_{wire} \cdot (P_{INT} + \alpha \cdot C_{tot} \cdot {V_{DD}}^2 \cdot f_{clk}) \quad (4)$$

where C_{tot} is total capacitance ($C_{pin} + C_{wire}$) of the 2D design. At this clock frequency, the M3D power saving does not depend on r_{p2w}. Moreover, as discussed previously, P_{INT} being the dominant component of total power, M3D power saving depends more on Δ_{cell} and the ratio of internal power versus switching power than Δ_{wire} or r_{p2w}.

Figure 7 b) shows power breakdown according to the type of cells. As the number of hard macros (e.g. memory blocks) and sequential cells are fixed, power consumed by these cells did not change in M3D designs. On the other hand, power consumed by combinational cells and clock signal can be reduced effectively in M3D designs utilizing lower number of buffers and using low

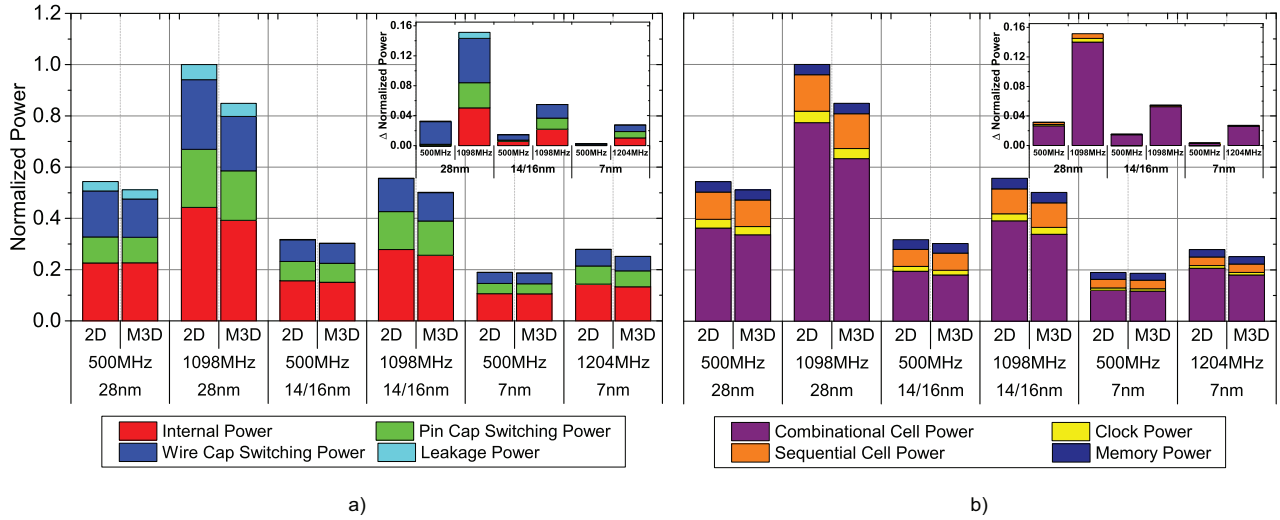

Figure 7: Power breakdown into a) internal power, pin cap switching power, wire cap switching power and leakage power, b) combinational cell power, clock power, sequential cell power and memory power at the minimum and maximum frequency of each technology nodes. The inset plots shows the power reduction of M3D for each power component.

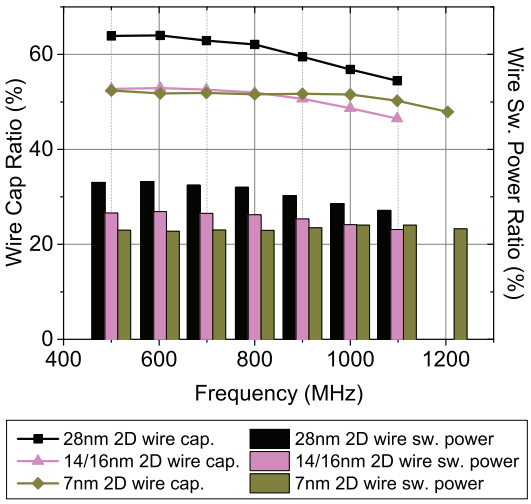

Figure 8: Wire capacitance out of total capacitance ratio and the resulting switching power out of total power ratio in 2D implementations across technology nodes.

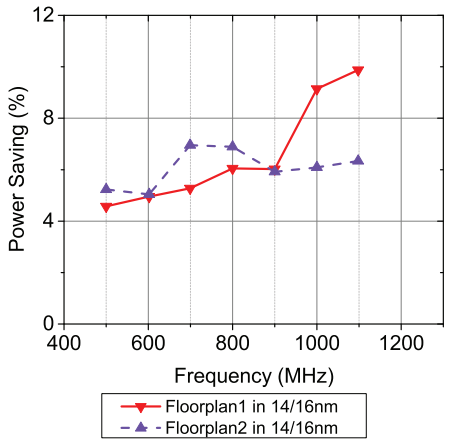

Figure 9: Impact of memory floorplanning on M3D power saving on 14/16nm technology node.

drive-strength cells.

Table 2 shows all the important design metrics of both 2D and M3D designs across foundry 28nm, 14/16nm and the predictive 7nm technology nodes at 1.1GHz. Since 1.1GHz is the maximum frequency for 28nm and 14/16nm implementation, and the second highest for 7nm design, we clearly see the significant standard cell area saving as well as wire-length saving with M3D designs.

Since the operating clock frequency is high, M3D designs save standard cell area by 9.9% on average for the three implementations, resulting in internal power and pin switching power savings. Although the ratios of internal power and pin-switching power on 2D versus M3D (11.4% and 14.9% in 28nm design) is smaller than wire-switching power ratio (21.8%), since those components account for more than 70% of the total power, they have a bigger contribution to the total power savings.

3.2.3 Impact of Floorplanning

In addition to the transistor technology and design characteristics of gate drive versus wire load, we found that M3D power savings is highly dependent on the design floorplan. Figure 9 shows M3D power savings for two different floorplans of the design at 14/16nm technology node versus clock frequency. Foorplan1 is highly optimized where tightly coupled memory blocks are located in the same location across two tiers to reduce wire-length, whereas in Floorplan2 the memory blocks are not well-aligned. As seen in the figure, Floorplan1 gives up to 10% power savings at 1.1GHz, while Floorplan2 achieves only upto 6%. This highlights the extremely high sensitivity of M3D power savings to the design floorplan.

3.3 Observations and Guidelines

The main observations and design guidelines from this study are summarized as follows.

Table 2: Normalized iso-performance comparison of 2D implementations and their M3D counterparts of the application processor across technology nodes at 1.1GHz. All values are normalized to corresponding 28nm 2D parameters. Capacitance and power values are normalized to 28nm 2D total capacitance and 28nm 2D total power, respectively.

	Normalized 2D			M3D percentage change from 2D		
Parameters	28nm	14/16nm	7nm	28nm	14/16nm	7nm
Footprint	1x1	0.64x0.64	0.41x0.35	-51.1 %	-50 %	-54.7 %
Density	1	0.899	0.803	-10.9 %	-8.9 %	-12.3 %
Cell count	1	1.029	1.251	-7.8 %	-7.3 %	-9.5 %
Std. cell area	1	0.32	0.085	-12.6 %	-7.3 %	-9.8 %
Wire-length	1	0.649	0.437	-23.6 %	-22.3 %	-27.5 %
Wire cap	0.544	0.328	0.207	-23.3 %	-13.1 %	-13.2 %
Pin cap	0.456	0.378	0.205	-16.5 %	-9.1 %	-12 %
Total cap	**1**	**0.706**	**0.412**	**-20.2** %	**-11** %	**-12.6** %
Internal power	0.443	0.278	0.136	-11.4 %	-7.9 %	-8.6 %
Wire switching power	0.271	0.129	0.063	-21.8 %	-14 %	-12.7 %
Pin switching power	0.227	0.148	0.062	-14.9 %	-10.1 %	-11.5 %
Leakage power	0.059	0.001	0.001	-13.4 %	-5 %	-3.2 %
Total power	**1**	**0.557**	**0.262**	**-15.1 %**	**-9.9 %**	**-10.3 %**

First, M3D technology offers the best power benefit on the foundry 28nm process node among the three technology nodes studied in this paper. The higher power benefit in 28nm is mainly achieved as a result of the lower pin capacitance versus wire capacitance ratio at low frequencies, due to the underlying planar transistors versus FinFET transistors used for 14/16nm and 7nm nodes. Hence this observation can be generalized to claim that designs using planar transistors are more likely to gain from M3D technology compared to FinFET transistors. Another point to note is that with technology scaling at the sub-10nm nodes, wire RC gets worse, and larger gates are required to drive increasing wire parasitics. Hence pin capacitance to wire capacitance ratio is likely to increase with technology scaling. M3D designs will see less benefit from just wire-length savings at advanced technology nodes. However, novel transistor technologies such as 2D devices (e.g. Transition metal dichalcogenides) might mitigate the high pin capacitance issue and take advantage of M3D designs.

Second, although power benefit from wire switching power is steady across clock frequency, it is lower compared to that achieved from standard cell area reduction. This is because wire switching power is a small portion of the total power consumption in our benchmark design. This observation can be further generalized to claim that designs with very large wire-loads or higher wire capacitance portions will benefit more from M3D technology.

Third, in our benchmark design, internal power and pin-cap switching power are the dominant components of total power. For such designs, M3D technology provides maximum benefit at higher frequencies where it has the potential to reduce buffer count and drive strengths of the cells, resulting in internal power and pin capacitance switching power reduction. For these designs, M3D provides nominal power improvement at low frequencies.

Fourth, M3D power savings is highly sensitive to the design floorplan. Multiple iterations, especially with fixed hard macros are required to get the power savings entitlement.

Fifth, the maximum power reduction observed in this study is 15% for the 28nm node at 1.1GHz clock frequency. To provide a full technology node benefit, M3D needs to deliver 30% or more power savings compared to their 2D counterparts. However, it should be noted that the results of this study are limited by 2D implementation tools pushed by "plug-in" flows to generate 3D designs. The full entitlement of M3D power benefit can only be realized by advancements in commercial EDA tools to support M3D designs, optimized gate/cell/block partitioning algorithms tuned to target designs and frequencies, M3D optimized memory designs and M3D-aware micro-architectures.

4. CONCLUSIONS

In this paper, for the first time, we presented a comprehensive study investigating the performance and power impact of M3D using a commercial in-order 32 bit application processor as the benchmark, implemented on foundry 28nm, foundry 14/16nm and predictive 7nm process nodes. We found that M3D provides maximum power savings at the 28nm technology node. The benefits improve at higher clock frequencies with the reduction of standard cell area in addition to wire-length savings. We support our observations with an in-depth analysis of the results and guidelines for M3D designs. This work demonstrates the potential of M3D and paves the way for future research that will enable it as a complement to traditional Moore's law scaling.

5. REFERENCES

[1] S. Bobba *et al.*, "CELONCEL: Effective Design Technique for 3-D Monolithic Integration targeting High Performance Integrated Circuits," in *Proc. Asia and South Pacific Design Automation Conf.*, 2011.

[2] O. Billoint *et al.*, "A Comprehensive Study of Monolithic 3D Cell on Cell Design Using Commercial 2D Tool," in *Proc. ACM Design Automation Conf.*, 2015.

[3] W. Arden *et al.*, "Morethan-moore white paper."

[4] Inside the iPhone 5s, "https://www.chipworks.com/about-chipworks/overview/blog/inside-the-iphone-5s".

[5] S. Yang *et al.*, "28nm Metal-gate High-K CMOS SoC Technology for High-Performance Mobile Applications," in *Proc. IEEE Custom Integrated Circuits Conf.*, 2011.

[6] S.-Y. Wu *et al.*, "A 16nm FinFET CMOS Technology for Mobile SoC and Computing Applications," in *Proc. IEEE Int. Electron Devices Meeting*, 2013.

[7] T. Song *et al.*, "A 14nm FinFET 128Mb 6T SRAM with VMIN-Enhancement Techniques for Low-Power Applications," in *IEEE Int. Solid-State Circuits Conference Digest of Technical Papers*, 2014.

[8] K.-I. Seo *et al.*, "A 10nm platform technology for low power and high performance application featuring FINFET devices with multi workfunction gate stack on bulk and SOI," in *Symposium on VLSI Technology Digest of Technical Papers*, 2014.

[9] 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.

[10] P. Batude *et al.*, "3DVLSI with CoolCube process: An alternative path to scaling," in *IEEE Int. Symposium on VLSI Technology, Systems, and Applications*, 2015.