Power Benefit Study for Ultra-High Density Transistor-Level Monolithic 3D ICs

ABSTRACT

The nano-scale 3D interconnects available in monolithic 3D IC technology enable ultra-high density device integration at the individual transistor-level. In this paper we demonstrate the power benefits of transistor-level monolithic 3D designs. We first build a cell library that consists of 3D gates and model their timing/power characteristics. Next, we build timing-closed, full-chip GDSII layouts and perform sign-off iso-performance power comparisons with 2D IC designs. We also study the characteristics of benchmark circuits that maximize the power benefits in monolithic 3D designs. Lastly, our study is extended to predict the power benefits of monolithic 3D designs built with future devices.

1. INTRODUCTION

To better exploit the benefits from 3D die stacking, monolithic 3D technology is currently being investigated as a next generation technology. In a monolithic 3D IC, the device layers are fabricated sequentially. When the top layer is attached to the bottom layer, the top layer is a blank silicon. Alignment precision is determined by lithography stepper accuracy, which is around 10nm today. Also, the top layer can be made very thin, around 30nm [1]. Thus, monolithic vias (MVIA) for vertical connections are very small—about two orders of magnitude smaller than through-silicon-via (TSV)—with almost negligible RC parasitics. With these small MVIA, designers can truly exploit the benefit of vertical dimension.

The early works for monolithic 3D ICs were technology-driven [6, 4, 9]. Recently, logic design methodologies for monolithic 3D ICs were demonstrated [2, 8, 7]. In these works, the authors presented various comparisons among monolithic 3D ICs and TSV-based 3D ICs and conventional 2D ICs in terms of footprint, timing, and power. However, timing was not closed in these works, which make the studies not practical. In addition, all these works assume that the timing and power characteristics of 3D monolithic gates are the same as 2D gates and did not demonstrate why that is a reasonable assumption. The authors also did not provide in-depth analyses and discussions on why monolithic 3D technology reduces power consumption and what factors affect the power reduction margin. This knowledge is crucial to maximize the benefit and justify on-going and future research on fabrication and design technologies for monolithic 3D ICs.

As discussed in [2, 8], monolithic 3D technology enables a very fine-grained 3D circuit partitioning. We can divide standard cells into PMOS and NMOS parts, place them in different layers, and connect them using MVIA, which we call transistor-level monolithic 3D integration (T-MI) in this paper. Or, as in TSV-based 3D ICs, we may place planar cells in different layers and connect them using MVIA, which is named gate-level monolithic 3D integration (G-MI). In this paper we focus on transistor-level integration that allows the highest integration density possible. The T-MI designs are different from G-MI: (1) Most of the 3D interconnects are embedded in the cells, (2) PMOS and NMOS transistors are on different layers, thus manufacturing processes can be optimized separately. (3) Physical layout (placement, routing, optimization, etc.) can be performed using existing 2D electronic design automation (EDA) tools, with modifications.

In this paper, we study the power benefit of T-MI based on timing-closed, detailed routing completed GDSII-level layouts and sign-off analysis on timing and power. Our comprehensive work encompasses device and interconnect-level study, gate-level modeling and optimization, and full-chip layout constructions, optimization, and timing/power analysis for the current and future technology nodes. With our layout-based simulations and in-depth analyses, we demonstrate how to maximize the power benefit of T-MI technology. For fair comparisons between 3D and 2D designs, timing is closed on all designs (iso-performance), and power consumption is compared. We also investigate the circuit characteristics that affect the power benefit of monolithic 3D ICs.

Our major contributions are as follows: (1) To the best of our knowledge, this is the first work to characterize the timing and power of the individual transistor-level monolithic 3D cells. We extract the internal RC parasitics of our T-MI cells and characterize their timing and power. We then compare T-MI cells with 2D counterparts. (2) We study the design aspects that significantly affect the power benefit of monolithic 3D ICs. We discuss what kind of logic circuits are suitable for power reduction in monolithic 3D ICs. In addition, we demonstrate that the power reduction rate also depends on the target clock period. (3) We build the libraries and full-chip layouts for monolithic 3D ICs implemented using 7nm devices. The goal is to predict the future trend of power saving with monolithic 3D technology and study how the smaller dimensions and varying RC parasitics affect the power benefit.

2. DESIGN AND ANALYSIS FLOW

One of the major benefits of T-MI is that existing 2D EDA tools can be used, with simple modifications if needed. We extensively use commercial EDA tools in this study. Our design and analysis flow, summarized in Fig. 1, consists of four parts: (1) library preparations, (2) synthesis, (3) layout, and (4) analysis. In the li-
brary preparation part, we prepare T-MI-specific library files. We synthesize the RTL codes of benchmark circuits using Synopsys Design Compiler. In the layout part, we perform placement, routing, and optimizations using Cadence Encounter (v10.12). Finally, we perform static timing analysis and statistical power analysis.

Our major efforts for T-MI design flow are spent on T-MI cell library construction and characterization, T-MI interconnect structure modeling, and T-MI wire load modeling. We modify the technology files and design rules to account for additional layers on the bottom tier as well as additional metal layers on the top tier (see Section 3.3). Using Cadence Virtuoso, we create our T-MI cells by modifying existing 2D cells. The cells are then abstracted to create the T-MI physical cell library. We also build interconnect RC libraries using Cadence capTable generator and QRC Techgen. For synthesis, we create the T-MI wire load models (see Section 3.4) that guide synthesis optimizations.

During layout construction, we first run Encounter placer. The tool recognizes T-MI cells as the cells with pins on multiple layers. For routing, we set up Encounter to utilize the additional metal layers on bottom and top tiers. Since our T-MI cells contain routing blockages on the MVIA layer, the router avoids 3D routing through the top tier part of the cells using MVIA. Using our T-MI interconnect library that reflects the T-MI metal layer structures and materials, we perform RC extraction on all the nets in the layout. Our full-chip timing/power optimizations and analyses for T-MI and 2D are the same, because the entire T-MI design (top/bottom tiers) is captured in a single Encounter session. We perform statistical power analysis with the switching activity of the primary inputs and sequential cell outputs at 0.2 and 0.1, respectively.

3. 45NM TECHNOLOGY SETUP

3.1 Monolithic 3D Cell Design

We design our T-MI 3D cells using the (2D) standard cells in Nangate 45nm library [10] as our baseline. As shown in Fig. 2, we fold the 2D standard cells into 3D and create T-MI 3D cells. The thicknesses of top/bottom tier silicon substrate and inter-layer dielectric (ILD) are 30nm and 110nm, respectively. The diameter of MVIA is 70nm. Note that by folding, each input/output pin is on both tiers. We prefer to place the PMOS transistors on the bottom tier and the NMOS on the top tier. In Nangate 45nm library, P/NMOS transistors show hole/electron mobility skew. To compensate the difference, in Nangate 45nm library, a PMOS is larger than the corresponding NMOS. Since extra silicon space on the top tier is required for MVIA (not on the bottom tier—see Fig. 2(b)), placing PMOS transistors on the bottom tier balances top/bottom silicon area usage. However, we should also consider manufacturing aspects in deciding the P/NMOS layer assignment.

After folding the cell, VDD and VSS strips are overlapping, as shown in Fig. 2. The power to VDD on the bottom tier can be delivered down through arrays of MVIA, placed apart from the VSS strip. We may need extra space for these VDD MVIA. Yet, power delivery network design and IR-drop analysis are outside our scope. Also, since VDD and VSS strips are overlapping, it may act as a small decoupling capacitor. However, in the extracted cell internal RC data for our inverter cell, the coupling capacitance (or cap) between VDD and VSS strips is around 0.01 fF, which is small compared with other cell internal parasitic capacitances.

The transistor model in Nangate 45nm library is PTM 45nm with bulk silicon technology [11]. In monolithic 3D technology, because of the structure, top tier transistors are similar to silicon-on-insulator (SOI) devices [1]. However, in this study we assume the same transistor model for T-MI and 2D cells, because (1) the original Nangate 45nm library is based on bulk silicon technology, and (2) if we assume both devices and interconnect structures in T-MI are different from 2D, it becomes harder to understand which factor contributes to power reduction, by how much.

3.2 Comparison with 2D Cells

Our T-MI cells preserve the same transistor sizes as in the original 2D cells. The T-MI cell height is 0.84 m, which is 40% smaller than the original 2D cell height (1.4 m). Thus, cell footprint reduces by 40%. The reasons why it is not 50% are (1) P/NMOS size mismatch incurs extra space on NMOS side, and (2) MVIA require extra space on the top tier.

When designing T-MI cells, care should be taken to reduce cell internal RC parasitics. As shown in Fig. 2(b), the connection from the PMOS on the bottom tier to the NMOS on the top tier needs to go through CTB, MB1, MVIA, CT, M1, then CT to diffusion. This 3D path may become larger than the original 2D path and may increase cell internal parasitic RC. Similarly, the path from the PB on the bottom tier to the P on the top tier goes through multiple layers. To reduce cell internal RC parasitics, it is important to minimize the lengths of 3D paths. To achieve shorter 3D paths, we should place MVIA close to the connecting transistors. We also need to utilize direct source/drain (S/D) contacts (see Fig. 6(c) in the supplement). The direct S/D contacts reduce the detour in the 3D paths and unnecessary RC parasitics.

We examine the cell internal RC parasitics of 3D and 2D cells and the impact on timing/power. In previous works [2, 8, 7], the authors assumed that the delay and power of 3D cells are the same as 2D cells and used 2D timing/power library. In [1], the authors fabricated a transistor-level monolithic 3D IC and measured the top/bottom transistor performances. They reported that the differences between 3D transistors and baseline 2D transistors were negligible. Yet, the delay and power of cells are also affected by cell internal RC parasitics. From Fig. 2(b), we can conjecture that there

1Our benchmark circuits and the synthesis results are shown in Section S2.
2The impact of switching activity is shown in Section S8.
3In sub-32nm nodes, thanks to advanced channel engineering techniques, the hole/electron mobility is about the same.
4GDSII layouts of some of our T-MI cells are shown in Fig. 6 in the supplement.
are coupling capacitances among PB, CTB, MB1, MVIA, CT, and M1. Using Mentor Graphics Calibre XRC with EM-simulation-based extraction rules, we extract these capacitance values as well as resistances and transistors from our T-MI cell layout. Then, we generate a SPICE netlist of the cell that consists of transistors and parasitic RC components.

Since Calibre XRC is designed for 2D ICs, it can only model one diffusion layer. Due to this tool limitation, top tier diffusion layer can be modeled as either dielectric or conductor. Even though the top tier silicon is doped (low resistivity) and the bodies of top tier transistors are tied to the ground, we expect that some amount of electric field may penetrate the top tier silicon and coupling among top and bottom tier objects (M1, MB1, P, PB, etc.) may exist. When we assume that the top tier silicon is dielectric, the coupling between top and bottom tier objects would be overestimated; when it is conductor, the coupling would be underestimated. The real case would be between these two extreme cases.

The total cell internal RC values, extracted from the original 2D cells and our 3D (T-MI) cells, are shown in Table 1. For 3D case, the results with top tier silicon as both dielectric (3D) and conductor (3D-c) are shown. From the results, we observe the followings: (1) For INV, NAND2, and MUX2, the R values of 3D are noticeably smaller than 2D counterparts, because we reduce the length of poly and metal lines inside the cells, using 3D interconnects. (2) The C values of 3D are comparable with those of 2D – the 2D value is between 3D and 3D-c. (3) For DFF, both R and C of 3D are larger than 2D counterparts. Due to the complex internal connections, we could not create a 3D cell that match RC parasitics of 2D. In summary, depending on the cell layout complexity, the internal RC ratio between 3D and 2D may vary.

Yet, the delay and power of the cells are more important metrics. We perform cell timing/power characterizations using commercial softwares. The SPICE netlists obtained from the previous extractions are fed into Cadence Encounter Library Characterizer, which runs SPICE simulations to characterize delay and power of cells under various input slew and load capacitance conditions. The delay/power of 3D and 2D cells are shown in Table 2. The values are obtained from the data tables in the characterized Liberty library. The delay is the cell internal delay including load effect, and the power is the dynamic power consumed within cell boundary (including short circuit power and power for gate/parasitic capacitances). We observe that for INV, NAND2, and MUX2, the delay and power of 3D are slightly better than 2D, whereas for DFF, they are a little worse. In addition, as the input slew and load capacitance condition changes from fast to slow case, the difference between T-MI and 2D becomes smaller. Note that depending on cell design quality and manufacturing technology, the results may change.

We believe that with proper cell designs, the delay and power of 3D cells could be similar to 2D counterparts.

### 3.3 Monolithic Interconnect Setup

Our T-MI interconnect structure is an extension of the Nangate (2D) 45nm library. As shown in Table 3, we use 8 out of 10 metal layers in the Nangate 45nm. For T-MI, we make two modifications: We add (1) a new metal layer on the bottom tier (MB1), and (2) three local metal layers on the top tier (M4-6).\(^5\)

With T-MI cell folding, the cells become 40% smaller than 2D (see Section 3.2). This results in about 40-50% smaller core footprint area. As a result, the cell pin density in T-MI becomes about 1.7-2X larger than in 2D, leading to a higher routing demand per unit area (or routing tile). To satisfy the high routing demand, we need to increase the routing capacity (#routing tracks per routing tile). The most area-efficient way is to add local metal layers, because of the small pitch. We found that adding 3 local metal layers increases routing capacity sufficiently.

Due to manufacturing issues (low thermal budget), in [2] the authors suggest tungsten is suitable for bottom tier metal. However, in this work we assume copper, because a copper-based manufacturing process may be developed. Besides, MB1 is mostly used for short interconnects such as within cells or short nets\(^6\). In our benchmark circuit M256 (see Table 11), the wirelength of MB1 (for net routing) is only 0.3% of the total wirelength. Thus, the impact of MB1 material on the timing and power of a whole circuit is minimal. When tungsten is used, IR-drop on the VDD strips could be an issue, which is outside our scope.

### 3.4 Monolithic 3D Wire Load Model

In T-MI designs, the wires are about 20-30% shorter than in 2D designs (see Table 4). We feed this information to the synthesis part by modifying wire load models (WLM). A WLM defines the statistical average of unit length resistance, capacitance, and area of wires, as well as the fanout vs. wirelength table. For each net, according to the fanout, the synthesis engine finds the corresponding wirelength and the capacitance/resistance/area from the WLM. We reflect the reduced wirelength of T-MI designs in the fanout

---

\(^5\) Our 2D and T-MI metal layers are shown in Fig. 9 in the supplement.

\(^6\) The impact of using MB1 for routing on optimization quality is discussed in Section S3.
vs. wirelength. From preliminary layout simulations, per each circuit we extract a WLM for T-MI as well as 2D. With these WLMs, the synthesized netlists for 2D and T-MI are different. The fanout vs. wirelength trends for our benchmark circuits are shown in Fig. 3. Note that the curves of circuits are distinct, which is related to the circuit characteristics (discussed in Section 4.3).7

4. 45NM RESULTS

4.1 Design Analysis Results

The layout simulation results for 45nm node are summarized in Table 4.5 With T-MI, the footprint reduces by 40.9-43.4%, which is larger than the cell footprint reduction rate, 40%. With T-MI, timing is better because of shorter wirelengths, and the optimizer may downsize cells and use less number of buffers while still meeting the target clock period. Thus, the footprint of the whole T-MI design could be further reduced than the individual cell footprint reduction rate. With T-MI, total wirelength reduces by 21.5-33.6%. Depending on the circuit characteristics, the wirelength reduction rate varies. We observe that the circuit with a larger wirelength reduction rate tends to show a larger power reduction rate. All designs met the timing. The power reduction was the largest in LDPC, 32.1%, whereas in DES, only 4.1%. In LDPC, the net power is much larger than the cell power, thus a large net power reduction with T-MI leads to a large total power reduction. We also observe that with T-MI, not only net power but also cell power reduces; with a better timing, cells are downsized and less number of buffers are used, to reduce cell power.

4.2 Comparison with Existing Works

Our results and the results from previous works ([2],[7]) are summarized in Table 5.9 All three works use Nangate 45nm library as baseline 2D. The footprint reduction rate of 3D over 2D in this work, [2], and [7] are about 42.3%, 30%, and 40%, respectively. This footprint reduction rate mostly affects overall design quality of 3D designs, because the timing and power reduction in the monolithic 3D designs is from reduced footprint and wirelength. Our results show larger wirelength reduction than these previous works. In [2, 7], they intentionally chose small target clock periods, thus timing was not closed. Note that power values in different works vary by much. For AES and LDPC, our results show larger power reduction rate than previous works. Interestingly, in all three works, the power reduction rates for DES circuit are low (only 2-4%).

4.3 Circuit Characteristics Study

As shown in Table 4, LDPC and DES showed much different power reduction rate with T-MI. By contrasting these two designs, we explain for what kind of circuits T-MI provides large power benefit. With T-MI, the buffer count reduces by 48.6% (in LDPC) vs. 3.2% (in DES), total wirelength reduces by 33.6% vs. 21.5%, total power reduces by 32.1% vs. 4.1%, cell power reduces by 12.8% vs. 1.6%, and net power reduces by 39.2% vs. 7.7%. Compared with LDPC, the buffer count reduction for DES is very small, which leads to very small cell power reduction. Although the wirelength reduction in DES is not so small, the net power reduction rate is significantly smaller than LDPC. The net capacitance/power consists quality of ours to the previous works; due to various reasons (floorplan setup, design and analysis flow, optimization methods, target clock period, switching activity factors, etc.), it is not possible to provide fair comparisons.

Table 5: Summary of design results in our work and previous works. The [2]-3D means their INTRACEL method with timing driven + IPO, which corresponds to transistor-level monolithic 3D design. The [7]-3D means their 3TM setup.

<table>
<thead>
<tr>
<th>circuit name</th>
<th>design type</th>
<th>total wirelength (m)</th>
<th>longest path delay (ns)</th>
<th>total power (mW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AES</td>
<td>ours-2D</td>
<td>0.260</td>
<td>0.770</td>
<td>13.69</td>
</tr>
<tr>
<td></td>
<td>ours-3D</td>
<td>0.199 (-23.5%)</td>
<td>0.775</td>
<td>12.20 (-10.9%)</td>
</tr>
<tr>
<td></td>
<td>[7]-2D</td>
<td>0.271</td>
<td>1.310</td>
<td>13.7</td>
</tr>
<tr>
<td></td>
<td>[7]-3D</td>
<td>0.214 (-21.0%)</td>
<td>1.165</td>
<td>12.8 (-6.6%)</td>
</tr>
<tr>
<td>LDPC</td>
<td>ours-2D</td>
<td>3.806</td>
<td>2.400</td>
<td>54.79</td>
</tr>
<tr>
<td></td>
<td>ours-3D</td>
<td>2.528 (-33.6%)</td>
<td>2.388</td>
<td>37.22 (-32.1%)</td>
</tr>
<tr>
<td></td>
<td>[2]-2D</td>
<td>1.83</td>
<td>2.461</td>
<td>1.554</td>
</tr>
<tr>
<td></td>
<td>[2]-3D</td>
<td>1.60 (-12.6%)</td>
<td>2.421</td>
<td>1.461 (-6.0%)</td>
</tr>
<tr>
<td>DES</td>
<td>ours-2D</td>
<td>0.611</td>
<td>0.976</td>
<td>63.88</td>
</tr>
<tr>
<td></td>
<td>ours-3D</td>
<td>0.479 (-21.6%)</td>
<td>0.968</td>
<td>61.24 (-4.1%)</td>
</tr>
<tr>
<td></td>
<td>[2]-2D</td>
<td>0.671</td>
<td>1.132</td>
<td>620.2</td>
</tr>
<tr>
<td></td>
<td>[2]-3D</td>
<td>0.581 (-13.4%)</td>
<td>0.971</td>
<td>608.2 (-19.9%)</td>
</tr>
<tr>
<td></td>
<td>[7]-2D</td>
<td>0.849</td>
<td>1.086</td>
<td>134.9</td>
</tr>
<tr>
<td></td>
<td>[7]-3D</td>
<td>0.682 (-19.7%)</td>
<td>0.923</td>
<td>130.7 (-3.1%)</td>
</tr>
</tbody>
</table>

Figure 4: Snapshots of routing results for LDPC and DES.

Table 4: Summary of layout results for 45nm node. The values represent the percentage difference of T-MI over 2D.

<table>
<thead>
<tr>
<th>circuit name</th>
<th>footprint (um)</th>
<th>total wirelength (um)</th>
<th>total cell footprint (um)</th>
<th>total net footprint (um)</th>
<th>total leakage footprint (um)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPU</td>
<td>-41.7%</td>
<td>-26.3%</td>
<td>-14.5%</td>
<td>-9.4%</td>
<td>-19.5%</td>
</tr>
<tr>
<td>AES</td>
<td>-42.4%</td>
<td>-23.6%</td>
<td>-10.9%</td>
<td>-7.6%</td>
<td>-13.9%</td>
</tr>
<tr>
<td>LDPC</td>
<td>-43.2%</td>
<td>-33.6%</td>
<td>-32.1%</td>
<td>-12.8%</td>
<td>-39.2%</td>
</tr>
<tr>
<td>DES</td>
<td>-40.9%</td>
<td>-21.5%</td>
<td>-4.1%</td>
<td>-1.6%</td>
<td>-7.7%</td>
</tr>
<tr>
<td>M256</td>
<td>-43.4%</td>
<td>-28.4%</td>
<td>-17.5%</td>
<td>-10.7%</td>
<td>-22.2%</td>
</tr>
</tbody>
</table>
of wire and (cell input) pin parts.\textsuperscript{10} For most nets in DES, wires are very short\textsuperscript{11}. This difference is also observed in Fig. 4. In DES layout, there are many small regions where cells are tightly connected inside but not so much to outside. For these short nets, pin capacitances dominate wire capacitances, thus reducing wirelength does not reduce net power as much. Although these two circuits are similar in size (\#cells, nets) and average fanout, because of the inherent difference in circuit characteristics, the power benefit of T-MI differs by much.

### 4.4 Impact of Target Clock Period

The power benefit of T-MI also depends on the target clock period. For AES and M256, we vary the target clock period and perform full designs, from synthesis to layout optimizations. The power reduction rate is shown in Fig. 5. The trend is clear; when the target clock is faster, the power benefit of T-MI becomes larger. This is because at faster clock speeds, the timing of the 2D design becomes harder to meet than T-MI, because of longer wires. The optimization engine uses more buffers and larger cells, leading to steep increase in cell power. Thus, the cell power reduction rate increases noticeably as clock becomes faster. With faster clock speeds, core footprint and wirelengths also become larger, leading to larger net power reduction rate with T-MI.

### 5. 7NM TECHNOLOGY SETUP

Another major aspect that affects the power benefit of T-MI is the technology node. As the technology advances, devices and wires shrink at different speed, affecting timing/power of the circuit and changing power benefit of T-MI. According to the latest ITRS 2011 roadmap\textsuperscript{5}, 7nm node is near the end of the roadmap.\textsuperscript{12} In ITRS projection for 7nm node, devices become dramatically efficient, however wires do not. The copper effective resistivity in 7nm is 3.7X larger than in 45nm, due to various reasons (edge scattering, barrier thickness, etc.).

We now predict how the power benefit of T-MI changes in the future 7nm node. The comparison between our 45nm and 7nm setup is shown in Table 6. Since there is no real 7nm node data available today, we scale down our 45nm library data as well as use data from ITRS projection. As a transistor model, we use ASU PTM-MG HP 7nm model\textsuperscript{[11]}. The interconnect dimensions are scaled down to (7/45)X = 0.156X, and the interconnect RC libraries are rebuilt, with a lower dielectric k (=2.2). We scale down the physical shapes of cells to 0.156X. Based on preliminary SPICE simulations,\textsuperscript{13} we also scale down cell input capacitance to 0.179X, cell delay to 0.471X, output slew to 0.420X, cell power to 0.084X, and cell leakage power to 0.678X. We apply these scaling factors to the 45nm Library and create our 7nm Library. Since the transistors in 7nm node are not planar but multi-gate (e.g. FinFET), the coupling between top/bottom tier transistors would be much smaller. Thus, we can reduce ILD thickness to keep the aspect ratio of MVIA reasonable.\textsuperscript{14}

The interconnect RC characteristics for 45nm and 7nm are obtained from the capTable built with Cadence Encounter, which runs EM simulations. The unit length resistances (\Omega/m) of 45nm and 7nm nodes for a local metal layer (M2) are 3.57 and 638, respectively, whereas for a global metal layer (M8), 0.188 and 2.650, respectively. The unit length capacitances (fF/m) of 45nm and 7nm nodes for M2 are 0.106 and 0.153, respectively, whereas for M8, 0.100 and 0.095, respectively. We observe that in 7nm node, the local metal layers become very resistive, due to the larger copper effective resistivity and the smaller metal width/thickness. Yet, in 7nm node, the wirelengths of the nets on local metal layers become shorter, thus the resistances of the net wires do not increase as dramatically. The capacitance per unit length increases for local metal layers, even though the dielectric k becomes smaller.

### 6. 7NM RESULTS

The layout simulation results for 7nm node are summarized in Table 7.\textsuperscript{15} Compared with the results in Table 4, we see that the footprint reduction rate is larger, especially for AES where 62\% footprint reduction was achieved. In the AES case, the target clock period is very small, 0.27ns. For the 2D design, Encounter performed high-effort optimization techniques to meet the timing, while for T-MI design it did not. As a result, the buffer count of the T-MI design is 84.5\% smaller. We also observed similar optimization differences for FPU. Wirelength reduction is 21.9-47.8\%. In the FPU case, total power reduction is the largest, 37.3\%. For DES, the power reduction is the smallest, 3.4\%.

\textsuperscript{10}We provide wire vs. pin power breakdown in Section S6.
\textsuperscript{11}The average wirelengths of DES-2D and LDPC-2D are 10.5\mu m and 72.0\mu m, respectively.
\textsuperscript{12}A summary of 45nm and 7nm node device and interconnect characteristics from ITRS projections are shown in Table 15 in the supplement.

---

Table 6: Comparison of our 45nm and 7nm node setup.

<table>
<thead>
<tr>
<th>Circuit</th>
<th>45nm</th>
<th>7nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transistor</td>
<td>Planar</td>
<td>Multi-gate</td>
</tr>
<tr>
<td>VDD (V)</td>
<td>1.1</td>
<td>0.7</td>
</tr>
<tr>
<td>Transistor length (drawn, nm)</td>
<td>50</td>
<td>11</td>
</tr>
<tr>
<td>Transistor width</td>
<td>Varies</td>
<td>Fixed</td>
</tr>
<tr>
<td>Back-end-of-line ILD k</td>
<td>2.5</td>
<td>2.2</td>
</tr>
<tr>
<td>M2 width (nm)</td>
<td>70</td>
<td>10.8</td>
</tr>
<tr>
<td>MVIA diameter (nm)</td>
<td>70</td>
<td>10.8</td>
</tr>
<tr>
<td>ILD thickness (nm)</td>
<td>110</td>
<td>50</td>
</tr>
<tr>
<td>Standard cell height (um)</td>
<td>1.4</td>
<td>0.218</td>
</tr>
</tbody>
</table>

Table 7: Summary of layout results for 7nm node.

<table>
<thead>
<tr>
<th>Circuit name</th>
<th>Footprint</th>
<th>Total wirelen.</th>
<th>Total cell</th>
<th>Net</th>
<th>Leakage</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPU</td>
<td>-47.0%</td>
<td>-34.2%</td>
<td>-37.3%</td>
<td>-44.4%</td>
<td>-21.0%</td>
</tr>
<tr>
<td>AES</td>
<td>-62.0%</td>
<td>-47.8%</td>
<td>-19.8%</td>
<td>-10.3%</td>
<td>-28.4%</td>
</tr>
<tr>
<td>LDPC</td>
<td>-42.9%</td>
<td>-27.7%</td>
<td>-19.1%</td>
<td>-3.7%</td>
<td>-26.0%</td>
</tr>
<tr>
<td>DES</td>
<td>-40.8%</td>
<td>-21.9%</td>
<td>-3.4%</td>
<td>-1.3%</td>
<td>-7.3%</td>
</tr>
<tr>
<td>M256</td>
<td>-44.6%</td>
<td>-23.0%</td>
<td>-17.8%</td>
<td>-14.1%</td>
<td>-23.0%</td>
</tr>
</tbody>
</table>

\textsuperscript{13}Our 7nm cell characterizations are presented in Section S1.
\textsuperscript{14}Note that our 7nm library setup is just one of many possibilities; there is a limitation in the prediction accuracy.
\textsuperscript{15}Our detailed layout results for 7nm node are presented in Section S4.
resistivity is not too high. We choose M256 as the test circuit, be-
ediate layers by 50%. The resistivity of global metal layers is not
of nets in 2D more than in 3D.
longer in 2D designs and the lower resistivity would reduce delay
may be lower than expected. In this scenario, we may expect that
ribbon) and manufacturing process, the resistivity of interconnect
6.2 Impact of Lower Metal Resistivity
reduction rate with T-MI becomes smaller.
As pin cap reduces, the net power reduces. Then, the
power benefit of T-MI does not increase with larger pin cap reduc-
power benefit may become larger or smaller.

6.1 Impact of Pin Cap Reduction Rate
As mentioned in Section 5, when we compare 7nm node with
45nm node, the cell pin cap reduces by 82.1%, which is smaller
than the wirelength reduction rate of designs, about 85% (com-
pare total wirelength of designs in Table 12 and 13). Thus, in
7nm node, the (pin cap)/(wire cap) ratio may become larger than
in 45nm node. Then, the wire cap reduction with T-MI reduces the
net total cap by a smaller percentage in 7nm node. However,
depending on the materials and manufacturing technology, the pin
cap of cells may reduce at faster rate than our projection. Thus,
we explore how the power benefit of T-MI changes when pin cap
reduces more.
For this study, we choose DES as the test circuit, because it
showed the largest (pin cap)/(wire cap) ratio among our circuits.
Thus, we expect to see larger impact with various pin cap settings.
Our simulation results are summarized in Table 8. Surprisingly,
the power benefit of T-MI does not increase with larger pin cap reduc-
tion rate. As pin cap reduces, the net power reduces. Then,
the cell power becomes more dominating factor, because cell power
does not decrease so much with smaller pin caps. Thus, the power
reduction rate with T-MI becomes smaller.
6.2 Impact of Lower Metal Resistivity
As discussed in Section 5, in 7nm node, the effective resistivity
of copper becomes very high. However, in the future, thanks to
better interconnect materials (e.g. carbon nanotube, graphene nano
ribbon) and manufacturing process, the resistivity of interconnect
may be lower than expected. In this scenario, we may expect that
the timing benefit of 3D may become smaller, because the nets are
longer in 2D designs and the lower resistivity would reduce delay
of nets in 2D more than in 3D.
As a case study, we reduce the resistivity of local and interme-
diate layers by 50%. The resistivity of global metal layers is not
changed, because the wires on the global layers are large and the
resistivity is not too high. We choose M256 as the test circuit, be-
cause it is the largest circuit among our benchmark circuits and more affected by net delay change.
The impact of the reduced metal resistivity is shown in Table 9. All designs met the timing. With lower resistivity, the power
consumption reduces, because with better timing smaller cells are
used. However, there is not much difference in wirelength and total
power reduction percentage. The cell and net power reduction rate
drove down a little, however the leakage power reduction rate went
up. From this result, we conclude that the lower metal resistivity
does not necessarily lead to smaller power reductions in monolithic
3D ICs.
7. CONCLUSIONS
In transistor-level monolithic 3D ICs, reduced footprints lead
to shorter wirelengths, better performances, and lower power consump-
tions. With carefully designed T-MI 3D cells, we performed
layout simulations for the benchmark circuits and demonstrated
up to 32.1% and 37.3% total power reductions in 45nm and 7nm
nodes. In addition, we discussed other factors that affect the power
benefit of T-MI, such as circuit characteristics and target clock
periods. We expect to see larger power benefits with T-MI in future
technology nodes, where wires become serious problems.
8. REFERENCES
Monolithic Integration targeting High Performance Integrated
Conf. on 3D System Integration, pages 1–4, 2010.
Edition.
SRAM Technology with the smallest S3 (Stacked Single-crystal Si)
Cell, 0.16 um, and SSTFT (Stacked Single-crystal Thin Film
Transistor) for Ultra High Density SRAM. In Proc. Symposium on
over 90nm 9 layer Cu CMOS. In Proc. Symposium on VLSI

Table 8: Impact of lower cell pin cap in 7nm node. The ‘-p’
suffix means the cell pin cap reduction rate (p20 means 20%
reduced pin cap).

<table>
<thead>
<tr>
<th>design</th>
<th>total WL (mm)</th>
<th>total power (mW)</th>
<th>cell (mW)</th>
<th>net (mW)</th>
<th>leak (mW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DES-2D</td>
<td>81.2</td>
<td>13.11</td>
<td>9.49</td>
<td>3.65</td>
<td>0.60</td>
</tr>
<tr>
<td>DES-3D</td>
<td>63.5 (-21.9%)</td>
<td>14.60 (-3.4%)</td>
<td>9.36</td>
<td>4.67</td>
<td>0.58</td>
</tr>
<tr>
<td>DES-2D-p20</td>
<td>81.3</td>
<td>14.38</td>
<td>9.48</td>
<td>4.30</td>
<td>0.60</td>
</tr>
<tr>
<td>DES-3D-p20</td>
<td>63.5 (-21.9%)</td>
<td>14.12 (-1.8%)</td>
<td>9.42</td>
<td>4.09</td>
<td>0.60</td>
</tr>
<tr>
<td>DES-2D-p40</td>
<td>81.2</td>
<td>13.54</td>
<td>9.39</td>
<td>3.56</td>
<td>0.59</td>
</tr>
<tr>
<td>DES-3D-p40</td>
<td>63.2 (-21.8%)</td>
<td>13.17 (-2.7%)</td>
<td>9.31</td>
<td>3.27</td>
<td>0.59</td>
</tr>
<tr>
<td>DES-2D-p60</td>
<td>81.3</td>
<td>12.74</td>
<td>9.35</td>
<td>2.81</td>
<td>0.59</td>
</tr>
<tr>
<td>DES-3D-p60</td>
<td>63.5 (-21.9%)</td>
<td>12.45 (-2.3%)</td>
<td>9.32</td>
<td>2.55</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Table 9: Impact of the lower metal resistivity in 7nm node for
M256. The ‘-m’ suffix means reduced metal resistivity.

<table>
<thead>
<tr>
<th>design</th>
<th>total WL (mm)</th>
<th>total power (mW)</th>
<th>cell (mW)</th>
<th>net (mW)</th>
<th>leak (mW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M256-2D-m</td>
<td>795</td>
<td>25.12 (-17.8%)</td>
<td>11.39</td>
<td>11.71</td>
<td>1.80</td>
</tr>
<tr>
<td>M256-3D-m</td>
<td>612 (-23.0%)</td>
<td>22.67 (-17.8%)</td>
<td>10.42</td>
<td>10.69</td>
<td>1.57</td>
</tr>
</tbody>
</table>
SUPPLEMENT

S1 Scaling Factors of 7nm Standard Cells

To obtain the scaling trends of 7nm cell characteristics, we first create SPICE netlists of 7nm cells. From the SPICE netlists of Nangate 45nm cells, the transistor models are replaced by ASU PTM-MG HP 7nm model [11]. The transistor fin height, width, and length of the ASU model are 18, 7, and 11 nm, respectively. We assume the number of fins per MOS transistor is 1, because the original cells are of X1 strength; the results may change if we use multiple fins. We also scale the cell internal parasitic R and C components in the original SPICE netlists by 7.7X and 0.156X, respectively, because: (1) The resistance of metal interconnect is not change much. And the length of cell internal interconnects become 0.156X. Thus, the R components become 0.156X of the original. (2) The unit length capacitance does not change much. And the length of cell internal interconnects become 7.7X, because M1 thickness, M2 thickness, and barrier thickness. Both the length (L) and width (W) of cell internal interconnects become 0.156X of the original. (3) The unit length capacitance does not change much. And the length of cell internal interconnects become 7.7X, because M1 thickness, M2 thickness, and barrier thickness. Both the length (L) and width (W) of cell internal interconnects become 0.156X. Thus, the R components become 0.156X of the original.

With the SPICE netlists of 7nm cells, we run Cadence Encounter Library Characterizer (ELC) to obtain Liberty timing and power library. The ELC runs SPICE simulations for various input slew and load capacitance conditions and builds a library with timing and power data. The characterization results are shown in Table S1.

Table S1: Scaling Factors of 7nm Standard Cells

<table>
<thead>
<tr>
<th>Component</th>
<th>45nm</th>
<th>7nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>input cap (fF)</td>
<td>0.463</td>
<td>0.125</td>
</tr>
<tr>
<td>cell delay (ps)</td>
<td>44.27</td>
<td>25.56</td>
</tr>
<tr>
<td>output slew (ps)</td>
<td>31.35</td>
<td>15.13</td>
</tr>
<tr>
<td>cell power (fJ)</td>
<td>0.446</td>
<td>0.020</td>
</tr>
<tr>
<td>leakage (pW)</td>
<td>2.844</td>
<td>2.583</td>
</tr>
</tbody>
</table>

Table 11: Benchmark circuits and synthesis results.

<table>
<thead>
<tr>
<th>Circuit</th>
<th>FPU</th>
<th>AES</th>
<th>LDPC</th>
<th>DES</th>
<th>M256</th>
</tr>
</thead>
<tbody>
<tr>
<td>#cells</td>
<td>1.8</td>
<td>0.8</td>
<td>2.4</td>
<td>1.0</td>
<td>2.4</td>
</tr>
<tr>
<td>cell area</td>
<td>9694</td>
<td>13,891</td>
<td>38,289</td>
<td>51,162</td>
<td>202,877</td>
</tr>
<tr>
<td>#nets</td>
<td>19,123</td>
<td>16,756</td>
<td>60,590</td>
<td>85,526</td>
<td>293,636</td>
</tr>
<tr>
<td>fanout</td>
<td>11,345</td>
<td>14,218</td>
<td>44,153</td>
<td>54,724</td>
<td>222,569</td>
</tr>
<tr>
<td>fanout</td>
<td>2.35</td>
<td>2.40</td>
<td>2.38</td>
<td>2.33</td>
<td>2.23</td>
</tr>
</tbody>
</table>

Figure 7: A zoom-in shot of T-MI design for AES. Skyblue rectangles are standard cells. For clarity, only MB1, M1, and MVIA layers are shown.

S2 Benchmark Circuits and Synthesis Results

Our benchmark circuits and synthesis results for 45nm and 7nm nodes are summarized in Table 11. The FPU is a double precision floating point unit. The AES and the DES are encryption engines. The LDPC is a low-density parity-check engine for the IEEE 802.3an standard. And the M256 is a simple partial-sum-add-based 256bit integer multiplier. The circuits are in different sizes. Note that target clock periods for 7nm node are smaller than those for 45nm node. We use Synopsys Design Compiler (ver. F-2011.09) for synthesis. The synthesis results are from 2D results. All synthesized designs (2D, T-MI, in 45nm, 7nm) met target clock periods.

S3 Concerns in Layout Optimizations

In the post-route optimization step, the Encounter optimization engine tries to preserve routed wires. In T-MI designs, the MB1 wires and the routing MVIA blocks the cell placement, thus the optimizer cannot place cells at (nor move cells to) such places. For example, in Fig. 7, the white spaces (dotted boxes) cannot be used for optimization such as buffering or gate sizing.

To see whether these MVIA/MB1 blockages cause design quality degradation, we perform a layout simulation. For this case study, we use AES as the target circuit, because it showed a high placement utilization with lots of densely packed placement re-
regions. From layout simulations, we observe that there are negligible differences in design quality, in terms of wirelength (+0.1%), timing (WNS = +25ps in original vs. +21ps without MB1 and MVIA), and total power (-0.1%). Thus, we conclude that under our settings (placement, routing, optimization options, final utilization, etc.), the routings on MB1 and MVIA do not degrade design quality noticeably. Note that the utilization of the above AES design is around 80%; we may see problems caused by the MVIA/MB1 blockages when utilization is very high. However, in general, it is customary not to exceed the 80% utilization, due to various reasons (placement and routing quality, optimization quality, decap area, etc).

### S4 Detailed Layout Results

The detailed layout simulation results for 45nm node are shown in Table 12. We set the target utilization to around 80%, which is common in industry designs. Since we observed severe wire congestions in LDPC (see Fig. 4(a)), the target utilization was lowered to about 33%; the 2D design was barely routable with this setting. We also observed significant wire congestions in M256, thus the target utilization was lowered to 68%. All designs met the timing (WNS≥0).

The detailed layout simulation results for 7nm node are shown in Table 13. We set similar target utilizations as for 45nm node. All designs met timing.

### S5 Impact of T-MI Wire Load Model

As mentioned in Section 3.4, we create custom WLMs for T-MI designs. There have been debates on whether WLM is helpful or not to the final layout results [3]. Since our target circuits are small to medium sized, we may expect that WLM is helpful to some extent. To see the impact of the custom WLMs on design quality, we perform the synthesis for T-MI designs with our T-MI WLMs but the 2D WLMs. As a result, the synthesized netlists for T-MI and 2D become similar. The layout results with/without custom WLM for T-MI designs are shown in Table 14. For FPU, AES, and DES, the design quality difference is negligible. However, for LDPC and M256, we observe significant increase in wirelength and total power without T-MI WLM. Thus, we conclude that for some designs, T-MI WLM models are helpful for obtaining larger power benefits with T-MI.

### S6 Breakdown of Net Power

We break net power into wire and pin power components (net = wire + pin). Wire means metal wires and vias used for routing outside cells, and pin means input pins of cells. As shown in Table 16, in LDPC, wire cap is much larger than pin cap, and so is wire power. Most of the net power reduction is from reduced wirelengths, as seen by the wire power reduction. In contrast, in DES, pin cap is much larger than wire cap. Thus, reduced wirelengths and wire power only reduces a small portion of the net power. In fact, most of the nets in DES are short, whereas most are long in LDPC; the average wirelength of LDPC-2D and DES-2D are 72.0μm and 10.5μm, respectively.

### S7 Impact of the Metal Layer Setup

To see the impact of the metal layer setup on power benefit of T-MI, we modify the metal layer stack of T-MI. Instead of adding 3 local metal layers on the top tier, we add 2 to local and 2 to intermediate metal layers. The original and modified metal stacks are shown in Fig. 9. We use LDPC and M256 for this case study. The results are summarized in Table 17. With the modified metal layer structure, compared with our T-MI results, total wirelength of the design with modified metal layers decreases by 1.6% for LDPC and increases by 1.0% for M256. The cell power, net power, and leakage power reduces, and the total power of LDPC and M256 reduces by 2.4% and 2.8%, respectively. Thus, we conclude that the metal layer structure of T-MI affects power benefit and should
switching activity factors to the primary input ports and the outputs of sequential cells (e.g. flipflop). Our default settings for primary inputs and sequential cell outputs are 0.2 and 0.1, respectively. Then, the given switching activity values are propagated to the rest of the circuit, based on the netlist connectivity and the functionality of cells. Since the switching activities of primary inputs affects until the first sequential cells and these paths are usually short, changing the switching activity factor of primary inputs affects the power by a small amount.

### S8 Impact of Switching Activity Factor

Another major factor that affects the power consumption is the switching activity factor. The switching activity factor is defined as the number of signal transitions (0-1 or 1-0) per a given clock period. The power values of cells and nets are linearly proportional to the related switching activities. Depending on various factors (architecture, usage scenario, etc.), the actual switching activity values may vary. For statistical power analyses, we provide switching activity factors to the primary input ports and the outputs of sequential cells (e.g. flipflop). Our default settings for primary inputs and sequential cell outputs are 0.2 and 0.1, respectively. Then, the given switching activity values are propagated to the rest of the circuit, based on the netlist connectivity and the functionality of cells.

Since the switching activities of primary inputs affects until the first sequential cells and these paths are usually short, changing the switching activity factor of primary inputs affects the power by a small amount.
small amount. In this case study, we vary the switching activity factors of the sequential cell outputs only. The total power of 2D and 3D designs for M256 under various switching activity factors are shown in Fig. 11(a). Although the total power increases with a larger switching activity factor, the power reduction rate does not change much, as shown in Fig. 11(b). The other circuits also show negligible differences in power reduction rate under various switching activity factors. Thus, we conclude that the power benefit of T-MI is not largely affected by the switching activity level.