# How to Cope with Slow Transistors in the Top-tier of Monolithic 3D ICs: Design Studies and CAD Solutions

Sandeep Kumar Samal[†], Deepak Nayak[§], Motoi Ichihashi[§],
Srinivasa Banna[§], and Sung Kyu Lim[†]
[†]School of ECE, Georgia Institute of Technology, Atlanta, GA
[§]Technology Research, GLOBALFOUNDRIES, Santa Clara, CA
sandeep.samal@gatech.edu, limsk@ece.gatech.edu

## ABSTRACT

In this paper we study the impact of low thermal budget process on design quality in monolithic 3D ICs (M3D). Specifically, we quantify how much the tier-to-tier transistor performance difference affects full-chip power and performance metrics in a foundry 14nm FinFET technology. Our study first shows that 5%, 10%, and 15% top-tier device degradation in a wire-dominated, timing-closed monolithic 3D IC design leads to 7%, 12%, and 18% full-chip timing violation, respectively. Next, we address this impact with our CAD solution named Tier-Aware M3D (TA-M3D) flow that identifies potential timing-critical paths and partitions them into the faster (bottom) tier to minimize the top-tier degradation impact. One unique challenge in timing closure in this case, is how to conduct buffering and sizing on the paths that lie entirely in the top or bottom-tier as well as those that span both tiers. Our approach handles all 3 types of paths carefully and closes timing under the given top-tier degradation assumption, while minimizing the total power consumption. Our enhanced monolithic 3D IC designs, even with 5%, 10%, and 15% slower transistors in the top-tier, still offers 26%, 24%, and 5% power savings over 2D IC, respectively. Our study also covers other types of circuits.

## CCS Concepts

**•Hardware → 3D integrated circuits; Electronic design automation; Methodologies for EDA;**

## Keywords

Monolithic 3D IC; Inter-tier Variation; Tier-Aware 3D Design

## 1. INTRODUCTION

Monolithic 3D ICs (M3D) is enabled by the sequential integration of device layers in the vertical direction [1]. Figure 1 shows the process of sequential integration. First, the bottom-tier is fabricated with conventional 2D IC process resulting in regular quality transistors. An empty wafer with $H^+$ ions implanted below the silicon surface, and with thermal oxide grown over it, is bonded on to the first tier using low temperature molecular bonding. The empty silicon wafer is then sheared along the $H^+$ line and polished. This is followed by the fabrication of monolithic inter-tier vias (MIVs), new transistors and interconnect layers for the second tier. Due to the presence of FEOL and BEOL in the first tier, the thermal budget of fabricating the sequential device layers is constrained. As a consequence, matching the device performance to that of a regular CMOS process becomes a major challenge. It also necessitates the requirement of proper design methods to handle this issue and evaluate M3D technology.

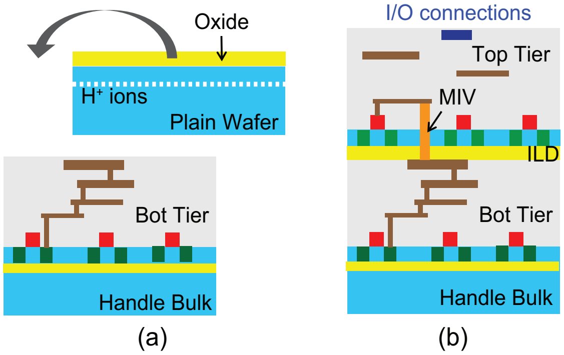


**Figure 1: Sequential integration [1] in monolithic 3D ICs.**

Low thermal budget of sequential device layers mainly affects the dopant activation process, leading to reduced mobility. Researchers have tried Solid Phase Epitaxy at 625$^o$C [2] and laser annealing with low in-depth thermal diffusion [3] for dopant activation. However, there is a performance reduction in the resulting transistors.For M3D design, Billiont *et al.* propose a simple design flow using 2D IC tools [4]. This method folds a 2D placement result along an edge to get 3D designs and does not utilize the true potential of high density MIVs. Chan *et al.* [5] have used Shrunk2D design methodology [6] for their modeling and estimation study. However, the design flow in [6] assumes equal transistor performance in both device layers which can lead to performance failure and optimistic power benefits. Panth *et al.* provide a very detailed analysis of various kinds of inter-tier variations for block level floorplanning and design [7]. In their work, they synthesize and layout all possible scenarios for all the blocks. This incurs a significant design-time overhead and the design approach is conservative. Moreover, they do not address gate-level M3D design folding.

The main contributions of this paper are: (1) We quantify the system level impact of low performance transistors in the top-tier of monolithic 3D ICs. (2) We propose a new Tier-Aware gate-level monolithic 3D IC design flow (TA-M3D) for optimization

*ISLPED '16, August 08-10, 2016, San Francisco Airport, CA, USA*
© 2016 ACM. ISBN 978-1-4503-4185-1/16/08...$15.00
DOI: http://dx.doi.org/10.1145/2934583.2934643

**Table 1: Cell delay comparison with slower transistors (-5%, -10%, -15%) in the top-tier vs. regular transistors (0%) in the bottom-tier.**

| Cell type | Degradation | | | |
|---|---|---|---|---|
| | 0% | -5% | -10% | -15% |
| INVX1 | 1.00 | 1.04 | 1.08 | 1.14 |
| NANDX1 | 1.00 | 1.04 | 1.10 | 1.15 |
| DFFX1 (clk->Q) | 1.00 | 1.08 | 1.17 | 1.24 |

**Table 2: Benchmarks used in our study.**

| Benchmark | Frequency | #Cells | #FFs | # I/Os | Type |
|---|---|---|---|---|---|
| LDPC | 1.87GHz | 66K | 2,048 | 4,100 | wire-dominated |
| AES | 4.10GHz | 166K | 10,769 | 389 | gate-dominated |
| T2 core | 1.90GHz | 207K | 46,732 | 477 | CPU core |

with low performance (slower) transistors in the top-tier. (3) Using our flow, we successfully demonstrate optimized M3D designs with high power savings and matching system performance under practical settings (i.e. low performance transistors in the top-tier). Our studies are based on full RTL-GDSII layouts of the benchmarks using a silicon-validated foundry 14nm FinFET PDK. To the best of our knowledge, this is the first work to address the issue of inter-tier performance difference in gate-level monolithic 3D ICs and the first monolithic 3D IC research work with actual foundry FinFET PDK.

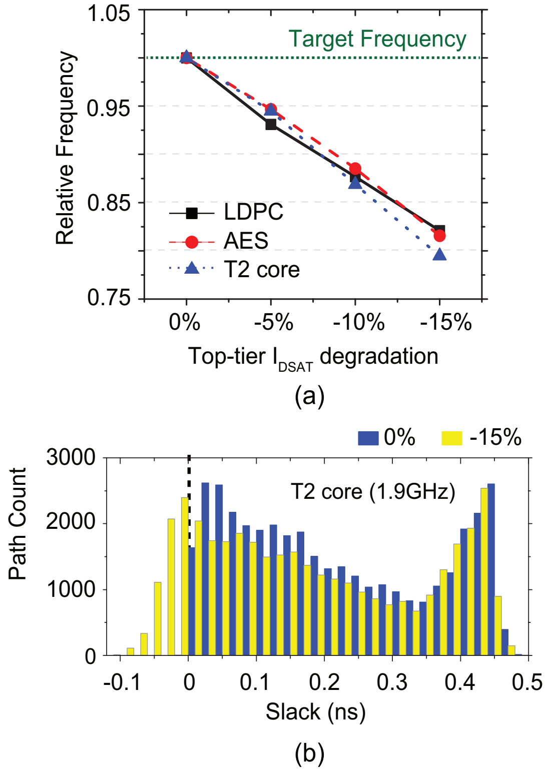
# 2. SLOW-TIER IMPACT STUDY

## 2.1 Full-Chip Design Settings

In our study, we use 3 cases (-5%, -10% and -15%) of performance degradation in the top-tier over the regular bottom-tier transistors. This covers a broad spectrum including the advances in low temperature fabrication and helps us assess tolerable limits of degradation. We use different standard cell timing and power libraries characterized with the respective transistor models having different ON-currents ($I_{DSAT}$). Table 1 shows the relative cell performance of some basic cells.

We use a wire-dominated low-density parity check (LDPC), a gate-dominated advanced encryption standard (AES) and a cpu core (OpenSPARC T2 single core) benchmark to cover the different design categories. Table 2 shows the details of the benchmarks used in our study. All the designs use 14nm FinFET PDK at the typical PVT corner and are designed for the same high frequency in all implementations of the given benchmark.

For comparison purpose, we use the state-of-the-art Shrunk2D design flow [6] as baseline. These designs are optimistic because they assume the same transistor performance in both the tiers. Here, the physical dimensions of all cells, interconnects and chip dimension are scaled by $1/\sqrt{2}$ to capture the 50% footprint scaling in the final M3D design. Then, a regular 2D-like design and optimization is followed by localized partitioning into two tiers (to maintain x-y locations of cells), expanding cells back to the original size, monolithic inter-tier via (MIV) planning, and tier-by-tier routing. The individual tier netlists, wire and MIV parasitics, and top-level 3D netlist are used for timing and power analysis. For MIV planning, we use a 3D metal stack with cell-pins defined in the appropriate metal layers followed by full routing. The locations of the vias going from the top metal of the bottom-tier to the bottom metal (metal1) of the top-tier give the optimized MIV locations [4, 6]. MIV size ($50nm$) and parasitics ($[10\Omega, 0.2fF]$) are determined based on the foundry 14nm PDK via-sizes and the via aspect ratio. The baseline design is 0% degradation (equal-tiers), while the other three cases are named as -5%, -10% and -15% to represent the performance degradation in the top-tier.



(a)

(b)

**Figure 2: Full-chip impact of slower transistors in the top-tier of monolithic 3D ICs. (a) Full-chip frequency degradation, (b) Slack distribution of all timing paths in T2 core.**

In this work, we use copper interconnects in both tiers. The focus of our work is to study the impact of tier-performance and develop design optimization methods for the same. We have not included power delivery network (PDN) impact study, 3D IC thermal analysis, PVT variation analysis and yield calculations in this paper.

## 2.2 Full-Chip Timing Impact

We implement the baseline M3D designs for all the three benchmarks. Baseline (0%) designs, with equal transistors in both tiers, meet the performance requirement. But in practical cases, the same design will experience slower transistors in the top-tier leading to an impact on the overall performance of the design.

Figure 2a shows the relative degradation in performance of the three design benchmarks using the baseline design approach. We observe that the system frequency can reduce by 18-21% if the top-tier transistors are 15% slower. Even for 5% slower transistors, the system performance reduces by 6-7%. Figure 2b shows the path timing slack distribution for T2 core design. While the baseline design satisfies timing constraints (blue distribution), the presence of slower transistors in the top-tier leads to a negative slack for many paths (yellow distribution) for the same design, therefore violating the full-chip timing constraints by 21%. For wire-dominated LDPC design, 5%, 10%, and 15% top-tier device degradation leads to 7%, 12%, and 18% full-chip timing violation, respectively. Therefore, designs with prior methodologies, which assume equal quality transistors in both tiers, will fail under practical scenarios. It is imperative to have a robust tier-aware M3D design approach to meet the required performance under practical conditions.

# 3. TIER-AWARE M3D DESIGN FLOW

## 3.1 Overview of Our Methodology

Timing optimization in commercial tools is carried out in three steps of preCTS, postCTS and postRoute optimization. Majority of path-optimization, buffer addition and cell sizing is carried out during the preCTS optimization stage which includes the parasitics of global routing impact but assumes an ideal clock at all clock-sinks. After full clock tree synthesis (CTS), postCTS timing optimization fixes the resulting clock skew impact. This is followed by detailed routing and parasitic extraction. The residual timing violation cases after full routing are fixed in the postRoute stage.

Figure 3 shows the detailed flow of our proposed Tier-Aware Monolithic 3D IC (TA-M3D) design approach. We leverage the fact that most of the timing optimization is achieved in the preCTS stage, and rest of the optimization stages are for fixing any resulting impact of clock/routing etc. In our approach, we provide the optimization tool with information of the new impact of inter-tier performance difference only after the preCTS stage. Not many timing paths in a design will be violated even with 50% of the cells (top-tier) becoming slower (see Figure 2b). Therefore, it is not necessary to over-design the entire top-tier to meet full-chip performance requirement. Hence, the overhead associated with fixing of the affected paths is minimized by handling them at an intermediate step and not from the beginning.

We use scaled physical dimensions of the cells and interconnects to capture the impact of reduced footprint and wirelength in evaluating the full-chip parasitic impact, while accommodating all the cells in 50% footprint. Therefore, commercial 2D IC optimization tools can be used to carry out 2D-like M3D optimization. The key to our approach is (1) We use only regular (bottom-tier) cells in the preCTS stage. (2) We use critical-path based tier partitioning to control the impact on timing distribution, therefore reducing the additional optimization effort. (3) We use only the top-tier (slower) cell library for CTS and to fix the newly created timing violations after adding the tier information. Figure 4 shows the layouts for T2 single core 2-tier M3D design.

## 3.2 Critical-Path based Partitioning

The use of only regular (bottom-tier) cells during preCTS optimization stage ensures that none of the paths are over-designed, though some will fail timing after introducing tier information. The output at this stage is the detailed placement and timing slack information with newly added/upsized cells and the design is well-optimized, though not 100%. To minimize the optimization effort and runtime in later stages which include the impact of low performance cells, we use critical-path based tier partitioning right after the preCTS stage. In this scheme, we partition the placement such that some of the most critical paths (paths with least positive slack) are intentionally confined in the bottom-tier with minimal change in 2D (x-y) location of all the cells in both the tiers. Though there will still be new violated paths appearing due to the top-tier performance impact, it helps in ensuring that the potentially worst paths are avoided from the top-tier. To maintain area balance in both tiers (<10% area skew), we only confine 10% of the total cells in the bottom-tier based on decreasing order of critical path delays. We use localized FM partitioning [8] for every 5$\mu m$ x 5$\mu m$ placement grid in the whole design, with some cells fixed in bottom-tier determined by path criticality. Since MIVs have negligible area overhead, we can tolerate a large total cutsize in this fine grid structure to maintain the optimized x-y placement locations in both the tiers.

The partitioned cells are then marked as per their tier location and this information is provided to the commercial optimization tool which treats them accordingly. We continue to use scaled dimensions to accommodate all the cells in half the footprint. The design now represents a placement projection of both the tiers on a single plane but is aware of which cell lies in which tier. The timing optimization tool can then fix the new set of violating paths only. For subsequent stages after preCTS, we only use the slower (top-tier) cell library. This enables the newly added/upsized cells to satisfy the timing constraints irrespective of which tier they are finally placed on. For most designs, only few paths need to be fixed. and hence, the number of new cell additions is relatively low.

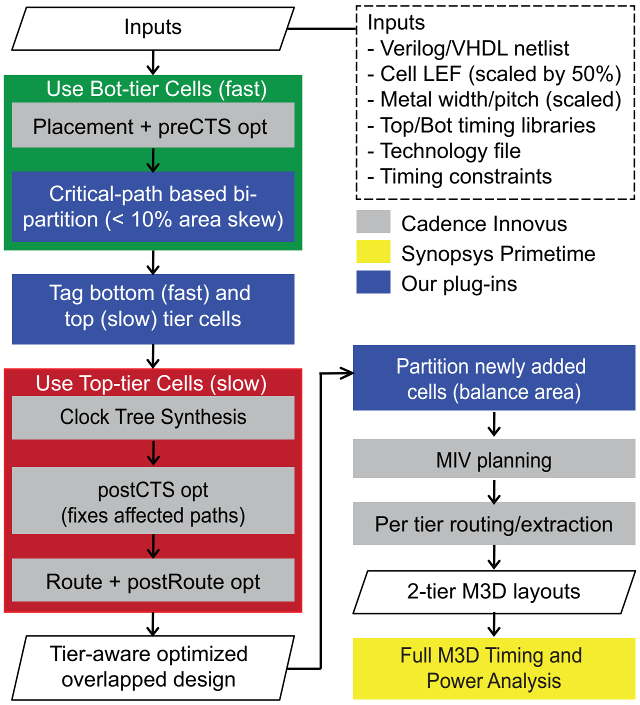

Inputs
Use Bot-tier Cells (fast)
Placement + preCTS opt
Critical-path based bi-partition (< 10% area skew)
Tag bottom (fast) and top (slow) tier cells
Use Top-tier Cells (slow)
Clock Tree Synthesis
postCTS opt (fixes affected paths)
Route + postRoute opt
Tier-aware optimized overlapped design
Inputs
- Verilog/VHDL netlist
- Cell LEF (scaled by 50%)
- Metal width/pitch (scaled)
- Top/Bot timing libraries
- Technology file
- Timing constraints
Cadence Innovus
Synopsys Primetime
Our plug-ins
Partition newly added cells (balance area)
MIV planning
Per tier routing/extraction
2-tier M3D layouts
Full M3D Timing and Power Analysis

**Figure 3: Our proposed Tier-Aware Monolithic 3D IC (TA-M3D) design and optimization flow to address slower transistors in the top-tier.**

## 3.3 M3D Clock Tree Design

Clock tree design is a critical step in any IC design flow and minimizing clock skew across the entire design is an important requirement. Therefore clock tree design is important in M3D as well. Since all the optimization and design is carried out in a 2D-like environment (scaled dimensions) involving cells in both the tiers, we fix the clock tree network in one tier only to maintain the original optimized tree structure. We plan the MIV connections for clock sinks of flip-flops/register-files in the other tier. Figure 5a shows a simplified version of our clock tree design method.

In M3D designs, I/O access in only through the top-tier which is low performance tier (Figure 1). It is best to fix the clock tree in the tier closest to the I/O i.e. the top-tier. This ensures that there is no additional global clock skew caused by the tree traveling back and forth across the tiers. The final M3D clock tree closely matches the one designed during 2D-like 3D optimization. The clock tree uses only slower cells as clock buffers since it is to be fixed in the top-tier. However, there will be a minor local skew impact at the end points of the tree after partitioning and tier-by-tier routing because the clock-net has to cross the bottom-tier BEOL stack (Figure 1) to reach the bottom-tier flip-flop from the clock MIV.

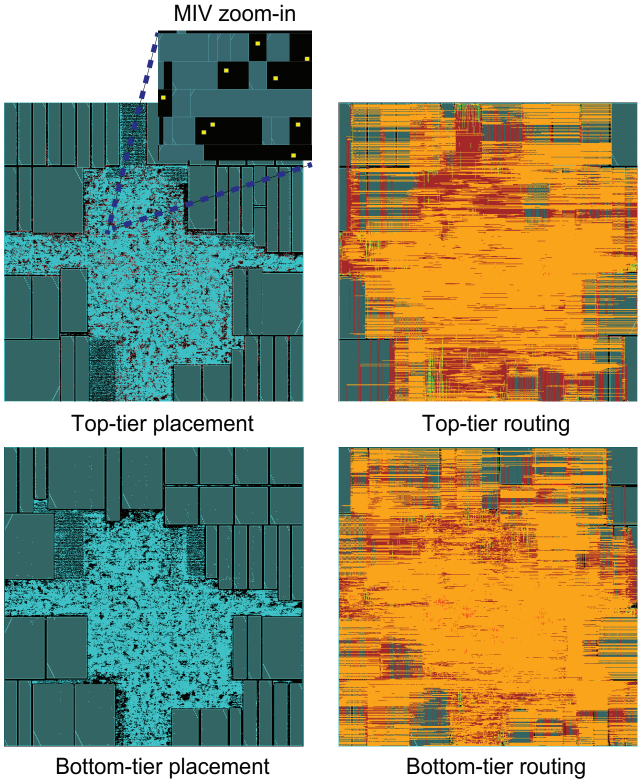
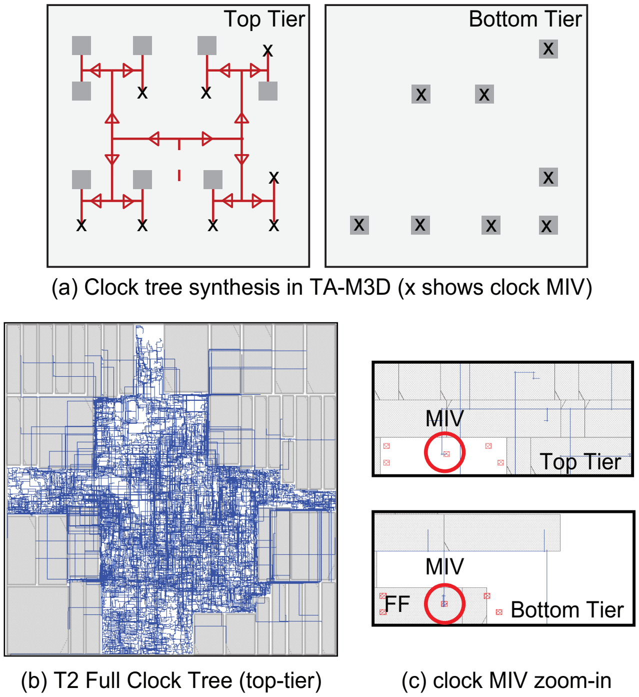


**Figure 4: Full-chip monolithic 3D IC layouts of OpenSPARC T2 core using a foundry 14nm FinFET PDK. The footprint is 415x415um. Zoom-in shows MIVs (yellow) and cells (cyan).**



(a) Clock tree synthesis in TA-M3D (x shows clock MIV)

(b) T2 Full Clock Tree (top-tier)

(c) clock MIV zoom-in

**Figure 5: Our clock tree synthesis method. (a) FFs in the top-tier are connected with a single tree, and FFs in the bottom are connected only using clock MIVs, (b) Full-chip top-tier clock tree in T2 core, (c) zoom in of a single clock MIV.**

Figure 5b shows the full clock tree for T2 core. While the clock sinks are spread across both the tiers, the primary tree (blue lines) is fixed in the top-tier. There is some clock routing in the bottom-tier due to the clock MIV to flip-flop connections as explained earlier.

## 3.4 Rest of the Design Flow

The inclusion of tier specific cell information affects some paths lying partially or fully in the top-tier. For the bottom-tier cells, the tool only needs to fix clock skew and net routing impact similar to 2D ICs. Since, we only use the slow (top-tier) cell library for later stages, the timing optimization fixes all kinds of paths (confined to one tier or crossing tiers) using only slower cells. Hence, after a full 2D-like tier-aware 3D optimization, we include another step of partitioning the newly added cells. The original faster cells are fixed in the bottom-tier only. The idea is to have area balance and maximize interconnect savings by allowing high 3D cutsize while maintaining the optimized tier-aware 3D placement results. This is followed by scaling up of the cells back to their original size, tier-by-tier placement legalization, MIV planning [4], tier-by-tier routing and power/timing analysis.

# 4. RESULTS AND DISCUSSIONS

## 4.1 Power Saving Challenge with FinFET

The total power in a digital design (excluding I/O pads) can be divided into cell-internal, switching and leakage power. Cell-internal power is the power dissipated inside a logic gate (excluding cell pins) due to the switching of internal nodes and short-circuit power. Switching power is further divided into wire switching and cell-pin switching. Wire switching comes from the interconnect capacitance while cell-pin switching is due to the switching of input gate capacitance of the logic gates.

FinFETs have high input gate capacitance and high cell power. For FinFETs, cell-power becomes relatively more dominant compared to planar technologies. Table 3 shows the detailed results for all designs implementations for the three different benchmarks. Comparing 2D IC and M3D designs, the key observations are (1) Relative contribution of wire power to total power is lesser for designs using FinFET technology. (2) Significant savings in wirelength and wire power in M3D designs. (3) Due to high cell power contribution, total power saving in M3D is modest for AES and T2 core. LDPC is wire-dominated and hence shows significant power savings of up to 28% in M3D. This includes the cell power savings obtained by reducing buffers and cell-sizes.

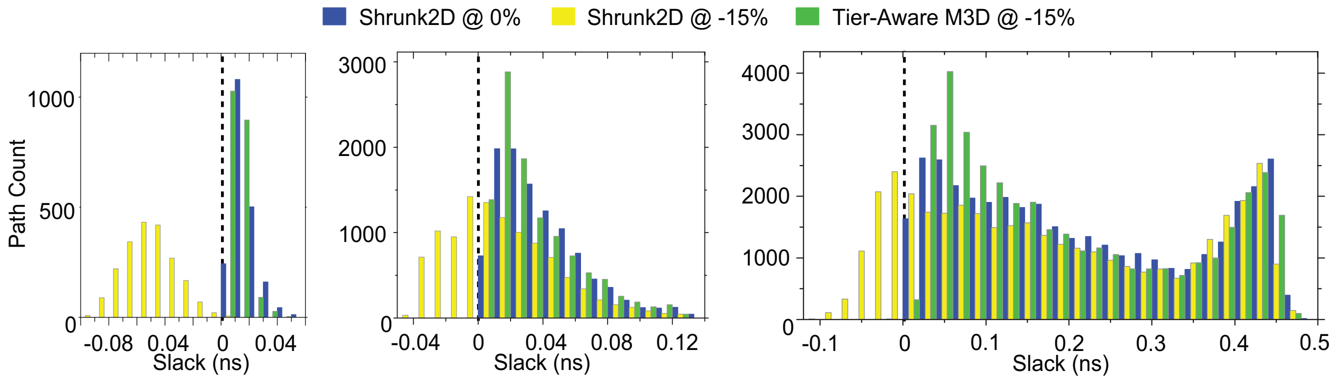
## 4.2 Power vs. Performance Trade-off

Our TA-M3D design flow guarantees timing closed design using high quality commercial tools even with slow transistors in the top-tier, as shown in Figure 6. But to recover the lost performance, there is an additional usage of timing buffers and larger cells. While our critical-path based partitioning reduces this usage, too much degradation can lead to a heavy power overhead.

Table 3 summarizes the power-performance trade-off of handling slow transistors in the top-tier of monolithic 3D ICs. Figure 7 shows the effectiveness of our partitioning method in confining most part of the worst paths in the bottom-tier only. The relative presence of worst paths (red lines) in the top-tier of our optimized designs is much lower than that in the baseline case. The cell area ratio is also reported in Table 3 to highlight that we maintain area balance. Note that for similar cell layouts, lower $I_{DSAT}$

**Table 3: Design comparisons under 0%, 5%, 10%, and 15% top-tier performance degradation. Results with Shrunk2D [6] (which ignores top-tier degradation) are shown for comparison. Leakage power is very small (< 1%) at typical PVT corner and hence not reported separately. Power values are reported in mW. Power saving shows the total power saving w.r.t. 2D results.**

| Design case | | Freq. (GHz) | Footprint ($\mu m \times \mu m$) | #Cells (x1000) | Bot:Top Cell Area | #MIVs | WL (m) | Wire Power | Cell-Pin Power | Cell-Internal Power | Total Power | Power Saving |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **LDPC, wire-dominated** | | | | | | | | | | | | |
| 2D IC | | 1.87 | 290×290 | 65.6 | - | - | 1.76 | 127.1 | 73.8 | 103.8 | 305.0 | - |
| Shrunk2D | 0% | 1.87 | 205×205 | 58.0 | 51:49 | 22,162 | 1.22 | 87.8 | 54.9 | 78.0 | 220.9 | **27.6%** |
| Our Flow | -5% | 1.87 | 205×205 | 62.0 | 52:48 | 22,647 | 1.22 | 88.1 | 57.1 | 81.8 | 227.1 | **25.5%** |
| | -10% | | | 62.5 | 54:46 | 22,628 | 1.23 | 88.5 | 59.2 | 83.5 | 231.4 | **24.1%** |
| | -15% | | | 70.9 | 53:47 | 22,646 | 1.35 | 102.1 | 83.3 | 103.6 | 289.1 | **5%** |
| **AES, gate-dominated** | | | | | | | | | | | | |
| 2D IC | | 4.10 | 320×320 | 166.4 | - | - | 1.22 | 67.4 | 132.9 | 181.8 | 382.7 | - |
| Shrunk2D | 0% | 4.10 | 225×225 | 163.0 | 51:49 | 51,051 | 0.92 | 59.3 | 119.5 | 170.9 | 350.2 | **8.5%** |
| Our Flow | -5% | 4.10 | 225×225 | 164.3 | 52:48 | 51,098 | 0.96 | 61.1 | 129.6 | 173.2 | 365.2 | **4.6%** |
| | -10% | | | 167.2 | 52:48 | 50,921 | 0.96 | 62.1 | 131.5 | 176.3 | 369.2 | **3.5%** |
| | -15% | | | 170.6 | 52:48 | 51,104 | 0.97 | 62.4 | 136.5 | 174.8 | 374.0 | **2.3%** |
| **T2 core, processor core** | | | | | | | | | | | | |
| 2D IC | | 1.90 | 585×585 | 206.7 | - | - | 4.18 | 144.3 | 89.9 | 247.2 | 483.1 | - |
| Shrunk2D | 0% | 1.90 | 415×415 | 204.7 | 51:49 | 68,022 | 3.32 | 124.7 | 88.9 | 231.4 | 446.6 | **7.6%** |
| Our Flow | -5% | 1.90 | 415×415 | 204.9 | 51:49 | 68,139 | 3.33 | 125.1 | 89.9 | 229.4 | 445.8 | **7.7%** |
| | -10% | | | 205.0 | 51:49 | 68,399 | 3.33 | 125.1 | 90.0 | 229.3 | 445.8 | **7.7%** |
| | -15% | | | 205.6 | 51:49 | 71,978 | 3.33 | 125.8 | 90.3 | 229.8 | 447.3 | **7.4%** |



(a) LDPC (b) AES (c) T2 core

**Figure 6: Path delay distributions under 0% and 15% top-tier degradation. All paths to the left of dotted line (= negative slack region) are violating timing constraint. We use Shrunk2D flow [6] for comparison. The paths in our design satisfy timing even under 15% degradation. (a) LDPC with 2,048 paths, (b) AES with 10,767 paths, (c) T2 core with 38,082 paths.**

implies lesser cell-internal power. However cell-pin power depends on input gate-capacitance of a cell layout and hence remains similar in all cases. Therefore, the overall full-chip power impact is dependent on the total power and path distribution for the baseline (0% ) case, and the number of affected paths with a slower top-tier. Figure 8 summarizes the overall power (w.r.t. 2D IC) for the benchmarks under different cases of top-tier performance degradation.

## 4.3 Analysis of Results

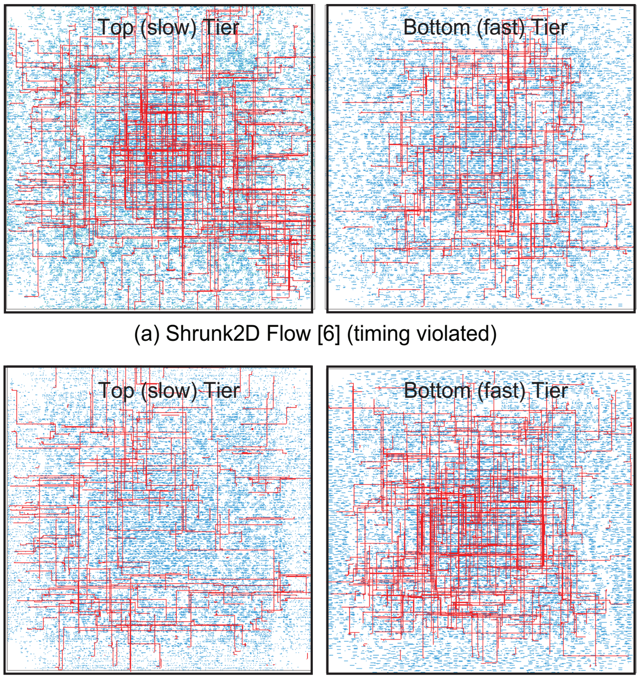
LDPC is a heavily wire-dominated circuit and has only 2,048 timing paths. With -15% degradation in the top-tier, all paths fail to meet timing (Figure 6a) in the baseline design and hence the overhead in fixing them is significant (Table 3). However, for the other cases, the overall power savings is still very high (24-26%) by using our critical path based partitioning. The transition in going from top-tier degradation of 10% to 15% is very sharp due to 14% extra cell usage and the additional wires.

For AES, cell power dominates the total power. Addition of buffers and larger cell sizes used to meet the performance requirement increase this cell power further. Though cell-internal power for cells in top-tier reduce due to slower transistors, cell-count and cell-pin power increase is higher.

T2 core has two key design features. (1) It has ~47K flip-flops (Table 2) and ~205K cells. Flip-flops have very high cell power and their count does not change for a given RTL. (2) The timing paths are widely distributed and hence a slower top-tier affects ~15% paths (Figure 6c). Therefore, after fixing the paths using our TA-M3D flow, the relative addition of cells is low, and above that, the top-tier flip-flop power is lower due to slower transistors. Hence the overall power savings remain similar (~7.5%) in all cases.

## 4.4 Clock Tree Metrics

Table 4 shows details of the clock buffers and clock power for the different design cases for all benchmarks. LDPC has very few flip-flops (=clock sinks) and hence uses less clock buffers and clock power. T2 core is a cpu core with ~47K flip-flops and 79 register-file modules. Therefore, the clock tree is large (Figure 5b) with many clock buffers. However, even with the clock tree using only



(a) Shrunk2D Flow [6] (timing violated)

(b) Our Tier-Aware Flow (timing closed)

**Figure 7: 100 worst timing paths (red lines) in LDPC design under 10% degradation. (a) Shrunk2D Flow [6], timing not closed, (b) Our TA-M3D Flow, timing closed. In our design, fewer critical paths are placed in the top (= slow) tier. In addition, we did not utilize excessive buffers and sizing to optimize the slow (= top) tier.**

slower cells, the overall change in power is negligible since the cell-internal power reduction compensates for the extra clock buffer usage. The clock routing required in the bottom-tier is very less compared to the full clock tree and has similar impact for all cases.

### 4.5 Runtime Analysis

From a design runtime point of view, the only additional design step we have, compared to the prior works [4, 6, 9], is the second partitioning step to determine the tier location of the newly added cells. This additional step has an average runtime of only 63s for LDPC, 463s for AES and 242s for T2 core benchmark. The other runtime increase is the extra time taken by the tool to optimize the newly degraded paths after addition of information about slower (top-tier) cells. However, the overall design cycle for very large designs, even for prior monolithic design implementation flows [4, 6], is few hours (includes 2D-like 3D placement, optimization, routing, partitioning, MIV planning, etc.). Therefore, the addition of a few minutes using our TA-M3D flow is insignificant, especially when the end result is a fully optimized timing-closed design with slower transistors in the top-tier.

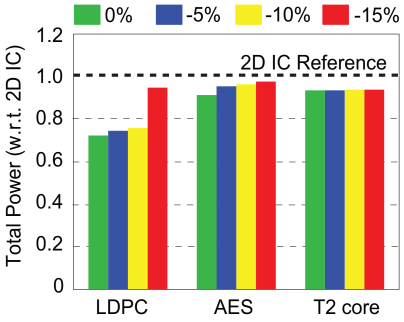


**Figure 8: Power consumption (w.r.t 2D ICs) under various top-tier transistor degradation.**

**Table 4: Clock buffer count and clock power results. Note that slower transistors in the top-tier show little impact on power due to the small number of buffers added.**

| | 2D | M3D | | | |
|---|---|---|---|---|---|
| | | Shrunk2D | -5% | -10% | -15% |
| **LDPC (1.87GHz, 60K cells)** | | | | | |
| Clock Buffers | 73 | 77 | 75 | 75 | 79 |
| Clock Power (mW) | 2.6 | 2.4 | 2.4 | 2.4 | 2.4 |
| **AES (4.1GHz, 165K cells)** | | | | | |
| Clock Buffers | 550 | 356 | 431 | 495 | 504 |
| Clock Power (mW) | 25.7 | 25.4 | 25.2 | 25.3 | 25.9 |
| **T2 core (1.9GHz, 205K cells)** | | | | | |
| Clock Buffers | 1,553 | 1,475 | 1,514 | 1,515 | 1,515 |
| Clock Power (mW) | 66.9 | 66.0 | 66.4 | 66.4 | 66.3 |

## 5. CONCLUSION

We studied and quantified the full-chip impact of slower transistors in the top-tier of monolithic 3D ICs when designed using previously studied design flows. Not considering the performance degradation during design process can result in full-chip timing failure. We then proposed a new critical-path based Tier-Aware M3D (TA-M3D) design flow to handle such slow transistors in the top-tier using industry-quality tools. Our critical-path based partitioning and design approach ensures minimal design time and power overhead. We demonstrate up to 26% power savings in M3D for wire-dominated benchmarks with slower transistors in the top-tier. While our design flow is robust and ensures timing closure even with a 15% degradation of the top-tier, our studies show that for the 14nm FinFET PDK, up to 10% degradation in the top-tier is well tolerable to maintain similar power savings as a no-variation design case.

## 6. REFERENCES

[1] P. Batude, *et al.*, “3D Sequential Integration Opportunities and Technology Optimization,” in *IITC/AMC*, May 2014, pp. 373–376.

[2] P. Batude, *et al.*, “Low temperature FDSOI devices, a key enabling technology for 3D sequential integration,” in *VLSI-TSA*, April 2013, pp. 1–4.

[3] B. Rajendran, *et al.*, “Low Thermal Budget Processing for Sequential 3-D IC Fabrication,” *IEEE Transactions on Electron Devices*, vol. 54, no. 4, pp. 707–714, April 2007.

[4] O. Billoint, *et al.*, “A Comprehensive Study of Monolithic 3D Cell on Cell Design Using Commercial 2D Tool,” in *DATE*, March 2015, pp. 1192–1196.

[5] W.-T. J. Chan, *et al.*, “3DIC Benefit Estimation and Implementation Guidance from 2DIC Implementation,” in *DAC*, June 2015, pp. 30:1–30:6.

[6] S. Panth, *et al.*, “Design and CAD Methodologies for Low Power Gate-Level Monolithic 3D ICs,” in *ISLPED*, Aug 2014, pp. 171–176.

[7] S. Panth, *et al.*, “Power-performance study of block-level monolithic 3D-ICs considering inter-tier performance variations,” in *DAC*, June 2014, pp. 1–6.

[8] C. M. Fiduccia and R. M. Mattheyses, “A Linear-Time Heuristic for Improving Network Partitions,” in *DAC*, June 1982, pp. 175–181.

[9] K. Chang, *et al.*, “Power benefit study of monolithic 3D IC at the 7nm technology node,” in *ISLPED*, July 2015, pp. 201–206.