Tier Partitioning Strategy to Mitigate BEOL Degradation and Cost Issues in Monolithic 3D ICs

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 develop tier partitioning strategy to mitigate back-end-of-line (BEOL) interconnect delay degradation and cost issues in monolithic 3D ICs (M3D). First, we study the routing overhead and delay degradation caused by tungsten BEOL interconnect in the bottom-tier of M3D. Our study shows that tungsten BEOL reduces performance by up to 30% at 4X resistance increase of bottom-tier interconnect. In addition, the bottom-tier BEOL adds a routing overhead to 3D nets, which is neglected in the state-of-the-art flow. Next, we develop two partitioning methods targeted specifically towards BEOL impact reduction. Our path-based approach identifies critical timing paths and places their cells in the top-tier to reduce the impact of delay degradation and routing overhead. Our net-based partitioning methodology confines the nets with long 2D wirelength into the top-tier to reduce the overall routing demand, and hence the metal layer usage in the bottom-tier. This in turn results in BEOL cost savings. Using a foundry 22nm FDSOI technology and full-chip GDS designs, we achieve tolerance of up to 4X increase in the bottom-tier BEOL resistance using our partitioning strategy. In addition, we save up to 3 metal layers in the bottom-tier of our M3D designs with up to 32% power savings over 2D IC for an interconnect dominated benchmark.

CCS Concepts

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

Keywords

Monolithic 3D IC; BEOL Degradation; Tier-Aware Partitioning

1. INTRODUCTION

3D ICs exhibit good power, performance and system scaling benefits over 2D ICs. With increased challenges in scaling device technology, a lot of research works are focusing on monolithic 3D ICs (M3D), enabled by sequential integration of device layers in the vertical direction [1, 2]. Monolithic Inter-Tier Vias (MIVs) are similar to metal-to-metal vias in dimensions and parasitics, enabling very fine-grained 3D partitioning in both the gate level as well as intra-gate level i.e. transistor level. Minuscule 3D vias are necessary in advanced technology nodes for fine-grained partitioning.

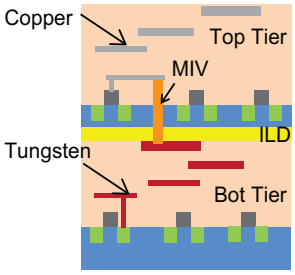

Figure 1: Vertical structure of 2-tier monolithic 3D IC. Tungsten is used for the interconnects in the bottom-tier to withstand high temperature during the top-tier device fabrication.

As with any new technology, M3D has multiple bottlenecks in the technology and EDA fields. Figure 1 shows the vertical layers of two-tier monolithic 3D IC. One major challenge in the successful fabrication of sequential device layers is to obtain high performance transistors with low thermal budget. Batude *et al.* have successfully demonstrated low temperature process (<650^oC) for transistor fabrication with the measured performance very similar to that of regular high temperature process transistors (~1050^oC) [2, 3]. Therefore, that particular challenge has been addressed. However, there is the requirement to use tungsten for back-end- of-line (BEOL) in bottom-tier, since copper cannot withstand temperatures close to 650^oC. The bulk resistivity of tungsten (56 Ω.nm) is 3.3X higher than that of copper (17 Ω.nm), and hence has significant impact on performance of the design. Billiont *et al.* developed an EDA methodology to use standard 2D IC tools to obtain M3D designs [4]. They studied the impact of tungsten, and observed that even after optimization, performance is reduced by up to 11% along with increase in power. However, their methodology just folds the 2D IC placement results along an edge with very few 3D connections. They do no utilize the tremendous potential of high density MIVs. As a result, in their work, the wirelength and power of M3D designs are higher than even 2D IC designs.

Panth *et al.* proposed a more attractive CAD methodology, the Shrunk2D approach, to obtain significant power reduction in M3D over 2D ICs [5]. Chan *et al.* have used this Shrunk2D design methodology as the "Golden 3D IC implementation" for their modeling and estimation work on monolithic 3D ICs [6]. The Shrunk2D approach allows the use of more MIVs to achieve significant wirelength savings (20-30%) and hence power savings. But it completely ignores the impact of either lower performance devices in

ICCAD '16, November 07-10, 2016, Austin, TX, USA

DOI: http://dx.doi.org/10.1145/2966986.2967080

the top-tier or tungsten interconnect in the bottom-tier. Both factors cannot be ignored simultaneously for practical M3D technology. With the device issue solved, handling the tungsten impact is the major challenge. While over-designing is the simplest approach, it results in power overhead over 2D ICs [4], therefore nullifying one of the primary benefits of M3D.

Cost is another important factor driving technology progress. The return on investment of using advanced fabrication techniques, multiple patterning etc. is diminishing with technology scaling. To provide a strong contention to an alternative technology or extension to current technology, M3D requires good cost savings along with power savings. Also, M3D fabrication has higher cycle time of fabrication due to multiple layers, even though the footprint is smaller. In addition. robust EDA machinery is necessary to successfully handle the impact of technology and fabrication requirements. Nayak *et al.* carried out cost modeling calculations and comparison for 2D ICs and 3D ICs [7]. Their results show that M3D can provide almost one node power, performance and cost (PPC) advantage over 2D IC at the same design node, but with the assumption of reduced metal layers (two to four) in the bottom-tier to cut on BEOL cost. To achieve this, without compromising on power, it is imperative to develop new EDA methodologies.

The contributions of this paper are as follows: (1) We study and discuss the adverse impact of BEOL in the bottom-tier of monolithic 3D ICs using different kind of circuits. (2) We develop an effective path-based partitioning methodology to mitigate the performance degradation due to tungsten without any additional design optimization. (3) We develop a net-based tier-partitioning methodology to reduce congestion and metal layer requirement in bottom-tier of monolithic 3D IC and hence save on BEOL cost by saving up to three metal layers.

Our studies are based on full RTL-GDSII layouts of the design benchmarks using a silicon-validated foundry 22nm FDSOI Process Design Kit (PDK). To the best of our knowledge, this is the first work on monolithic 3D ICs to develop partitioning strategy to handle BEOL impact on performance. In addition, our work is the first to address cost benefits in monolithic 3D ICs and develop CAD strategy to potentially reduce the cost.

2. FULL CHIP DESIGN SETTINGS

2.1 Reference M3D Design Settings

For the reference M3D design optimization with high quality commercial tools, we have two options, (1) Shrunk2D flow [5] followed by partitioning and (2) Regular 2D IC design with pre-fixed pins, followed by folding along an edge. Shrunk2D flow is shown to have higher power savings [5, 6] and it provides more freedom to implement different partitioning schemes to mitigate technology issues without any over optimization. The second edge-folding technique does not have this flexibility because the partitioning has to be done along edge and cannot be modified. In addition, wirelength and power savings are absent in this approach [4].

We use the Shrunk2D optimization approach for our reference designs. These designs are optimistic because they use same device performance and copper BEOL in both tiers, both of which are not simultaneously practical. In this method, the physical dimensions of cells, wire pitches/widths and chip dimensions are scaled down to create a virtual next node. However, the electrical properties of cells (.lib file) and interconnects (RC per unit length) is kept the same as original technology. These tricks enable the correct mapping of interconnect savings obtained by 3D IC designs during 2D-like 3D IC timing optimization. Since RC per unit length (.tch file) is same, the parasitics are correctly evaluated as in a two-tier 3D IC designs but with both the tiers overlapped. The end result is the design with cells in both tiers of monolithic 3D IC projected onto a single 2D plane and well optimized in terms of timing and cell sizing. This is treated as an ideal 3D IC design with zero vertical interconnect overhead.

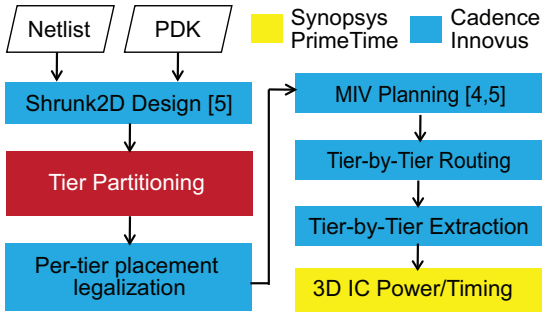

Figure 2: Monolithic 3D IC Design Flow. Our work focuses on the 3D IC Tier Partitioning step.

Table 1: Benchmarks used in our work. Design metrics are based on 2D IC GDSII layouts using a foundry 22nm FDSOI PDK and commercial CAD tools.

		LDPC	SIMD	AES
		wire dominated	medium	cell dominated
Frequency	(GHz)	1.64	2.36	3.27
Footprint	(μm)	320×320	320×320	300×300
Density	$(\%)$	49.1	70.0	77.6
# Total cells		97,466	150,655	175,548
# Total nets		100,357	154,630	175,808
Metals used	(# layers)	6	6	6
Wirelength	(m)	2.29	2.08	1.19
Wire Power	(mW)	107.4	48.7	31.7
Cell Power	(mW)	117.4	84.8	93.9
Total Power	(mW)	224.8	133.5	125.6
Wire Power %		**48%**	**37%**	**25%**

The next step is division of the 2D placement into a rectangular grid with multiple bins and then partitioning the cells into two tiers using local mincut in each such bin while maintaining decent area balance and similar x-y location as obtained after optimization. MIV planning is carried out by using 3D metal stack of both tiers together with cell pins defined in appropriate locations [4, 5]. The vias running from top-metal of the bottom-tier to bottom-metal of the top-tier give the optimized MIV locations for the provided partitioning solution. Finally, the individual tiers are routed followed by 3D power analysis. MIV size (50nm) and parasitics (10Ω, 0.2fF) are fixed as per foundry 22nm PDK metal pitches, via-sizes, and via aspect ratio. The overall design flow is summarized in Figure 2, with our work focusing on the 3D tier partitioning techniques to mitigate BEOL impact and reduce cost in two-tier monolithic 3D ICs. We assume equal transistor performance in both-tiers but with tungsten interconnect in the bottom-tier. We have not included power delivery network (PDN) impact study, 3D IC thermal analysis, PVT variation analysis and yield calculations in this study.

2.2 Benchmarks and Metrics

All the designs in our study use foundry 22nm FDSOI PDK at the typical PVT corner and are designed for the same high frequency in both 2D and 3D implementations of the respective benchmarks. We used Cadence Innovus for standard place and route optimization and Synopsys Primetime for power and timing analy-

Table 2: 3D IC design metrics using Shrunk2D flow [5].

		LDPC	SIMD	AES
Frequency	(GHz)	1.64	2.36	3.27
Footrpint	(μm)	198×198	220×220	210×210
Density	$(\%)$	46.2	70.6	76.0
# Total cells		71,355	147,760	174,229
# Total nets		74,779	151,458	174,489

sis. We used a wire-dominated low-density parity check (LDPC), a commercial single instruction multiple data (SIMD) engine and a gate-dominated advanced encryption standard (AES) benchmark to cover different kinds of designs categories. Table 1 shows the 2D design details of the benchmarks used in our study. Table 2 shows the high-level design details after 2D-like 3D optimization (Shrunk2D) of the benchmarks. Depending on contribution of wire-power in 2D IC and the wirelength savings in 3D IC, buffering and cell-upsizing is also reduced in 3D IC implementation, leading to cell-power savings as well. The 2D and 3D designs are optimized to have similar final cell-placement density as reported in Table 1 and 2. Figure 3 shows the GDSII level final layouts of the benchmarks. LDPC has very long nets with high wire-power contribution, while AES has shorter nets and localized clusters of dense connections. These factors influence the final power savings observed in the M3D designs, which are discussed in later sections.

3. M3D BOTTOM-TIER BEOL ISSUES

3.1 Impact of Tungsten Resistance

Figure 4 shows the delay degradation due to the increase in the resistance of bottom-tier BEOL in the three different benchmarks. The bulk resistance of tungsten is 3.3X higher than copper. We evaluate the degradation at the different points up to 4X worse resistance. Further increase in the resistance will lead to further degradation in timing. In this analysis, we use the reference design flow while maintaining similar design density in 2D IC and 3D IC designs. For the baseline case, we carry out grid-based regular partitioning on the optimized 3D designs using Fiduccia Mattheyses (FM) algorithm [8]. During partitioning, the only constraint is to maintain an area-balance with area skew of <10% across both tiers. Since the 2D-like 3D designs are optimized with copper as interconnects, the impact of tungsten in bottom-tier BEOL is not accounted for during optimization. This leads to significant negative slack in the timing paths which pass through the bottom-tier. The increase in resistance worsens timing in terms of resulting negative slack. However, the relative degradation depends on the original path delay. The degree of degradation also depends on how many timing paths and how much portion of each path crosses the bottom-tier. LDPC is a heavily interconnect-dominated benchmark. The timing paths have longer wires compared to AES benchmark. Therefore, the resulting magnitude of negative slack, due to increase in resistance, is highest for LDPC, followed by SIMD and then AES. Simple partitioning schemes have no control on the distribution of these paths. They only consider the connectivity graph and area-balance during partitioning.

3.2 Accurate Routing Overhead Modeling

Shrunk2D [5] methodology carries out design and optimization in a single 2D plane. However, on splitting the cells into two tiers, the vertical connections between cells have to cross through the entire BEOL stack of bottom-tier before reaching the MIVs in the top-tier. Figure 5 shows the two device layers and the 3D routing in cells across different tiers which has to go through bottom-tier

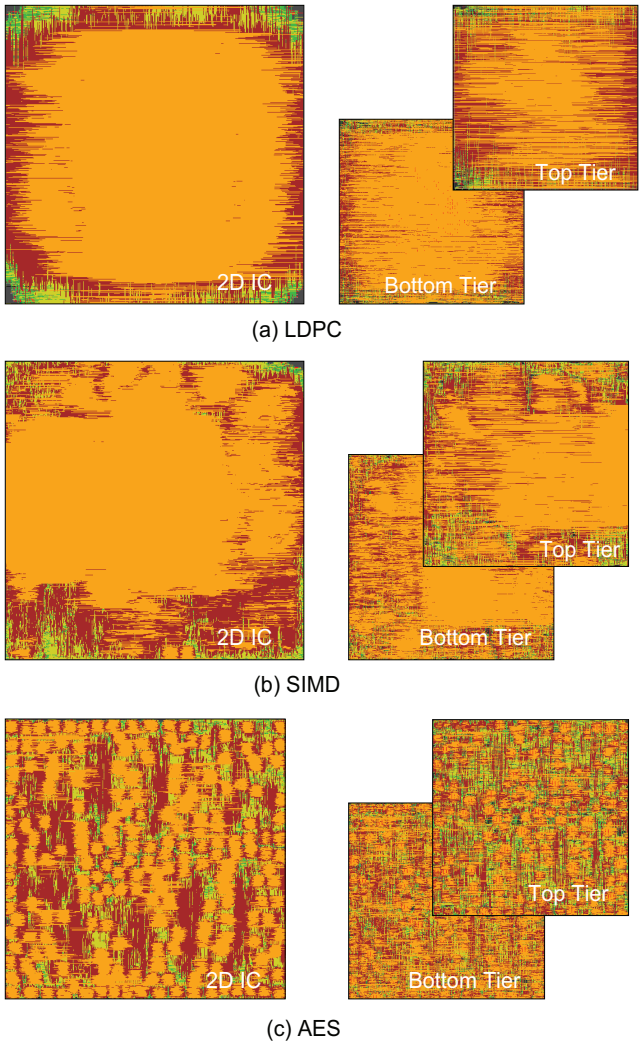

Figure 3: 2D IC and Monolithic 3D IC GDSII layouts. Metal6 (topmost metal) is amber color, and Metal5 is maroon. (a) LDPC: very long nets and global spread (b) SIMD: long nets (c) AES: short nets but locally dense. All layouts are to scale.

BEOL. With gate to gate 2D distance scaling to sub-micron values in advanced technologies, the 3D routing cannot be considered negligible anymore. This extra routing for 3D nets has two consequences: (1) The timing optimization carried out in the Shrunk2D phase does not account for the additional routing. (2) Extra routing means extra interconnect capacitance which results in additional wire power. Both of these issues aggravate further if tungsten is used in bottom-tier. Therefore, reducing metal layers from bottom-tier not only helps in reducing cost, it also helps in reducing negative impact on power and timing. However, directly reducing metal layers without any design consideration will lead to heavy congestion, long detours and a possible routing failure with multiple errors.

The use of tungsten needs to be accounted for during optimization. Also, the 3D routing across the bottom-tier BEOL cannot be completely avoided. However, we can reduce these adverse impacts to a significant extent by using clever partitioning strategy as discussed in the following sections. To keep things simple, we discuss the two proposed partitioning methodologies independently

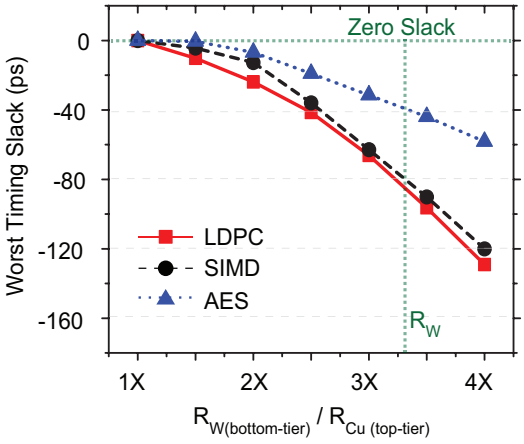

Figure 4: Full-chip timing degradation with respect to increase in the bottom-tier tungsten BEOL resistance. Higher interconnect component in timing paths results in more degradation.

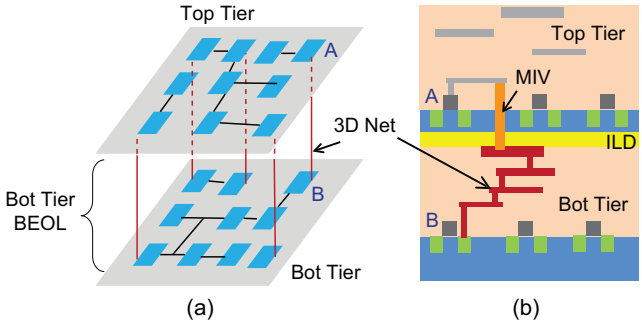

Figure 5: 3D interconnect overhead in monolithic 3D ICs. (a) simplified model of two-tiers with 3D nets, (b) vertical structure showing the 3D routing in bottom-tier. Shrunk2D [5] ignores this overhead

without combining one with the other.

4. PATH-BASED TIER PARTITIONING

4.1 Motivation

The slack distribution of all timing paths of the SIMD benchmark is shown in Figure 6. The slack information is obtained after placement, routing and timing optimization of the 3D designs as discussed in Section 2.1. The key observation is that the distribution is wide and many paths have very high positive slack. Not all paths are equally critical. In this particular example, almost 50% of the paths have a positive slack of >50ps. Therefore, these paths can tolerate an additional delay of up to 50ps and still satisfy the system timing constraints. Our idea uses this fact by confining some of the most critical timing paths in the top-tier of M3D, so that they remain protected from the adverse impact of tungsten interconnect. The tolerance achieved by confining them in the top-tier is influenced by the actual distribution of the timing paths and the degree of connectivity in the netlist.

4.2 Algorithm and Complexity Analysis

Algorithm 1 explains our path-based tier partitioning algorithm. The key idea is to try and confine as many worst critical paths in the top-tier as possible without violating area skew constraint and

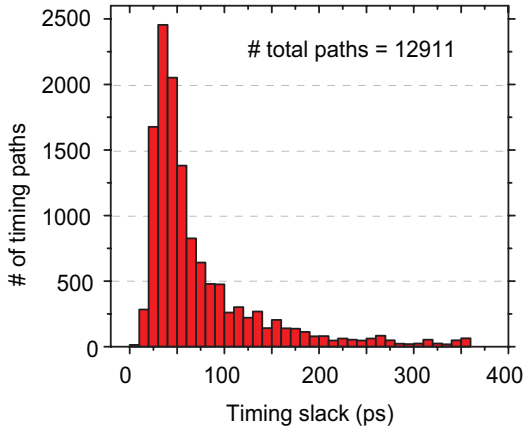

Figure 6: Timing path distribution of optimized 3D design (SIMD benchmark) before partitioning. The wide distribution offers good room of positive slack to tolerate additional interconnect delay.

without increasing the cutsize drastically. MIVs are minuscule and occupy negligible area. Therefore, it is possible to increase the cutsize to achieve our goal. However, too much increase leads to congestion issues in the bottom-metal (metal1) of the top-tier since all the MIVs need to be routed. The inputs to our algorithm are already present in the design database after the 3D placement optimization using the Shrunk2D approach. The detailed timing information of all paths is directly obtained from the Cadence Innovus after optimization. Therefore, there is no additional runtime overhead of evaluating timing in our approach. The *GetDesignDatabase()* function reads the input information into partitioning engine.

The main function *PathBased()* takes in a high max path count as input, and carries out FM partitioning with these critical paths pre-partitioned in the top-tier. Too many paths being fixed in the top-tier may result in high area-skew across tiers. In such a case, the max paths count is reduced by 20 paths and the function is run recursively until the final partitioning result is obtained. The *GridBasedFMPartition()* function divides the layout in a grid structure and carries out mincut FM bi-partitioning ($O(n)$ complexity) in each bin of the grid while maintaining <10% local area skew across tiers. The major difference from the baseline partitioning (Section 3.1) is that many cells are pre-partitioned into the top-tier before starting the FM algorithm. The overall runtime depends on the number of recursions occurring to obtain the final results. The recursions can be limited by choosing an aggressive, yet judicious max path count after observing the timing slack distribution.

4.3 Experimental Results

Table 3 summarizes the design results for the different benchmarks using our path-based partitioning algorithm. For comparison, the baseline partitioning results are also presented. One of the very important information provided is the cell-area ratio across tiers. With our approach, we maintain very good area balance. High area-skew will result in one-tier requiring more silicon. Due to sequential fabrication process, the other tier has to be of same area, even though major part of it will be whitespace. Therefore, good area-balance is very critical in maintaining footprint reduction benefits in M3D.

Using our algorithm, we can tolerate the impact of bottom-tier interconnect resistance to significant extent, without any further timing optimization. The second-last column shows the tolerable

Table 3: Results with our path-based partitioning. We can tolerate up to 2.2X to 4X BEOL resistance degradation in the bottom-tier without compromising full-chip M3D power and area balance across tiers. The runtime includes tier partitioning step only.

Method	Bot:Top Cell-Area	#MIVs	Wirelength (m)			Congestion		WireCap (pF)	Power (mW)			**Tolerable Bot-tier R**	Runtime (sec)
			Top-Tier	Bot-Tier	Total	Hor (X)	Ver (Y)		Wire	Cell	Total		
LDPC (1.64 GHz)													
baseline	49:51	19,931	0.72	0.76	1.48	0.01	0.01	244.8	65.4	78.2	143.9	**1X**	45
path-based	53:47	21,803	0.69	0.82	1.51	0.03	0.09	253.8	67.5	78.2	146.0	**2.2X**	66
SIMD (2.36 GHz)													
baseline	50:50	40,558	0.93	0.89	1.82	0.01	0.04	318.6	42.4	80.7	123.1	**1X**	173
path-based	51:49	46,213	0.92	0.95	1.87	0.01	0.08	332.3	44.9	80.7	125.6	**2.9X**	245
AES (3.27 GHz)													
baseline	51:49	35,797	0.54	0.57	1.11	0	0.04	166.5	26.4	89.8	116.2	**1X**	348
path-based	50:50	42,270	0.55	0.65	1.20	0.01	0.11	170.8	27.7	89.8	117.5	**4X**	489

Algorithm 1: Our path-based tier partitioning algorithm

Data: Shrunk2D design (Placement and Timing Slack)
Result: Tier-location of all cells

Function `PathBased` (*path_max*)
  `GetDesignDatabase();`
  path = worst_path;
  while *path_count* <= *path_max* **do**
    `AssignPathToTier` ($path, top$);
    path_count++;
    path = next_worst_path;
  `GridBasedFMPartition();`
  if $(complete == 0)$ **then**
    new_path_max = path_max-20;
    `PathBased` (new_path_max);

Function `AssignPathToTier`(*path, tier_no*)
  $\forall cell \in path : cell \rightarrow$tier = tier_no;

Function `GridBasedFMPartition()`
  if $(Bin\_Area\_skew > 10\%)$ **then**
    complete = 0;
    return;
  else
    FM partition per bin;
    complete = 1;
    return;

limit of worse resistance in the bottom-tier. The baseline designs cannot tolerate any degradation in interconnects. The degree of tolerance depends on the role of interconnect in the benchmarks. LDPC is heavily interconnect dominated and hence interconnect degradation is more critical than in other designs. Even then, we can handle up to 2.2X resistance increase in the bottom-tier interconnects. Further degradation in resistance requires new optimization techniques or new interconnect material innovations. On the other hand, AES has short nets and hence interconnect impact is relatively lower. With our approach, we can handle full 4X resistance increase in the bottom-tier interconnect.

The side-affects of achieving our goals is the increase in cutsize (hence MIV count) and congestion which leads to more wirelength with minor increase (<2%) in total power. The congestion in both horizontal and vertical direction is also shown in the Table 3. The runtime for path-based tier-partitioning is higher because the partitioning process goes through a few recursions, depending on choice of initial max path count to be fixed on top-tier. However, the overall runtime for partitioning is a few minutes only and is negligible compared to the total design runtime of few hours (includes MIV planning, tier-by-tier routing and parasitic extraction). Therefore, our path-based partitioning algorithm proves highly beneficial to reduce or completely remove the optimization overhead of handling tungsten interconnects.

5. NET-BASED TIER PARTITIONING

5.1 Motivation

While CAD methodology and power reduction for monolithic 3D ICs has been extensively studied [4, 5, 6], there are no prior works targeted towards saving cost in M3D to push it further as an alternative to technology scaling or extension of current technology node. Nayak *et al.* [7] have modeled power-performance-cost (PPC) benefits of M3D and TSV-based 3D ICs, but only with estimated cost benefits and no actual designs. In this section, we focus on reducing the cost in M3D by proposing a net-based partitioning algorithm with the objective of reducing metal layer usage in the bottom-tier. M3D gives significant savings in wirelength. Therefore, for the same design, we can easily reduce the usage of metal layers in one or both the tiers and hence reduce overall fabrication cost of an IC, by reducing the number of masks and cycle time. The reduction is significant especially in advanced nodes where tight minimum pitch and multiple patterning increase the back end of line (BEOL) cost by a considerable amount.

5.2 Algorithm and Complexity Analysis

Algorithm 2 describes our net-based tier partitioning algorithm for gate-level M3D. The target is to reduce the metal usage in bottom-tier as much as possible without affecting the cell-area balance between the two tiers and while minimizing power overhead. Initial overlapped 2D placement results give us a clear idea of the x-y locations of the cells but not the tier of placement. Figure 7 shows the HPWL distribution of all nets in the LDPC benchmark after 3D design, but before any partitioning. The longer nets, though relatively lesser in count, account for 50% of total wirelength.

Our idea is to choose the longest nets from these 2D placement results and force such nets in top-tier only. The other shorter nets and the resulting 3D nets use the bottom-tier metal with reduced demand of routing resources. Note that by placing the long nets in top-tier, we are not adding any extra wires as the placement locations of cells are already determined. In fact, it actually helps in reducing the 3D routing overhead by avoiding unnecessary snaking of wires across two tiers for longer nets, which may be cut during simple area-balance partitioning only. The routing resource demand in the bottom-tier is also reduced as relatively shorter nets are routed in the bottom-tier.

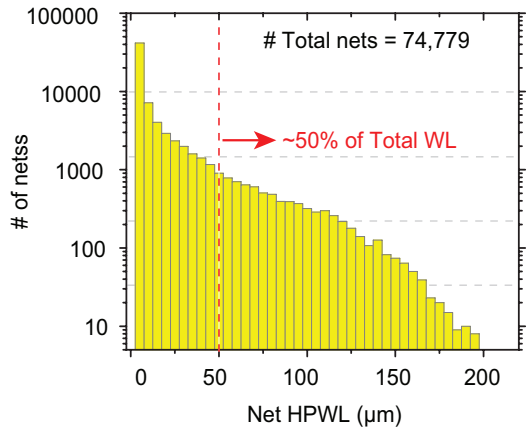

Figure 7: HPWL distribution of all nets in optimized LDPC M3D design before partitioning. Longer nets add up to 50% of total HPWL, although their count is lower than shorter nets (y-axis is in log scale).

There are two good ways to determine the length of a net. They are net's cell-count i.e. number of cells in the net and the 2D half perimeter wirelength (HPWL) which is the Manhattan distance between the placement location of the extreme cells in the net. A net with many cells may not be necessarily spread across a larger area with longer wirelength, depending on how the design was optimized and the cells placed. Therefore, it is not the best metric to assess the length of a net. Therefore, we focus on the HPWL data obtained from the Shrunk2D placement results. Our partitioning tries to maintain the 2D locations of the cells. Therefore, HPWL gives a practical estimate of the net wirelength excluding the 3D routing overhead. The calculation and sorting of HPWL of all nets is a one-time requirement just after Shrunk2D placement. Since the cell 2D placement information is already available, time complexity is $O(nlogn)$, where n is the total number of nets.

The main function *NetBased()* picks all nets with their HPWL greater than a threshold HPWL value ($hpwl\_max$), and fixes the cells in these nets in top-tier. We recursively reduce $hpwlmax$ until an area-balanced partitioning solution is achieved. The final cell and net count in each tier is similar to that in normal area-balanced partitioning but the longer nets get confined to the top-tier. The runtime overhead is negligible compared to the runtime of entire process of monolithic 3D IC design (Figure 2) which involves much more time intensive steps of design optimization and routing.

5.3 Experimental Results

Using our net-based partitioning algorithm, we study monolithic 3D IC metal layer savings for the three different benchmarks. Table 4 shows the detailed results in terms of per-tier wirelength, congestion, wire capacitance, wire-power and total power with different metal layer usage for different design implementations. Cost savings by reduction of metal layers cannot be disclosed due to foundry confidentiality requirements. One key advantage of our net-based algorithm is that, we can skew the wirelength per-tier significantly without affecting the area-skew. This is because we pick the long nets and fix them on the top-tier leading to controlled wirelength skew across tiers. As a result, we are able to reduce the routing demand in the bottom-tier significantly, resulting in up to three metal layers reduction. The number of metal layers used across the various implementations are shown in the second column of Table 4. All the 2D IC and baseline M3D implementations

Algorithm 2: Our net-based tier partitioning algorithm

Data: Shrunk2D design (Placement and Nets' HPWL)
Result: Tier-location of all cells

```
Function NetBased(hpwl_max)
    GetDesignDatabase();
    ∀n ∈ Nets such that n →hpwl >= hpwl_max
    AssignNetToTier(n, top);
    GridBasedFMPartition();
    if (complete == 0) then
        new_hpwl_max = hpwl_max-5;
        NetBased (new_hpwl_max);

Function AssignNetToTier(net, tier_no)
    ∀cell ∈ net : cell →tier = tier_no;

Function GridBasedFMPartition()
    if (Bin_Area_skew > 10%) then
        complete = 0;
        return;
    else
        FM partition per bin;
        complete = 1;
        return;
```

of the three different benchmarks use 6 metal layers (both tiers of M3D). We also show the cell-area ratio in top and bottom-tiers to stress on the fact that we have good area-balance across the tiers to maintain footprint savings. The cutsize increases with higher wirelength skew across tiers.

As a consequence of reducing routing resources, the overall relative demand of routing resources increase, which results in more congestion and higher total wire capacitance due to increased proximity of signal wires. Simply reducing the metal layer limit in the bottom-tier for the same baseline partitioning, (i.e. without any wirelength skew) makes the bottom-tier unroutable. This is because, the routing demand in bottom-tier remains the same as the baseline case but with much lesser resources. However, our approach intentionally creates the wirelength skew, while maintaining area-balance. Therefore, we are able to reduce the metal layers and hence cost with minor power overhead. Also, the routing in the top-tier metals becomes denser resulting in more wire capacitance. Depending on the design characteristics, we achieve varying results in terms of reducing number of metal layers in the bottom-tier vs. power increase. We observe that using only two metal layers in the bottom-tier leads to very heavy congestion and incomplete routing. Moreover, power delivery requires the use of some intermediate metal layers. Hence, we evaluate our partitioning methodology down to usage of three metal layers in the bottom-tier. The normalized power saving comparison is shown in Figure 8. All values are normalized to the 2D IC power of the respective benchmark.

LDPC is an interconnect dominated benchmark with long nets having global spread. With our approach, we successfully reduce three metal layers in the bottom-tier. The congestion increases from 1% in the baseline case to 14% in this case. However, the power benefits over 2D IC is still a significant 32% compared to the 36% in the baseline case. While wire savings reduce with more congestion, cell-power savings remain almost constant across all implementations, resulting in relatively lower impact on total power savings. For the SIMD and AES, benchmarks, the nets are relatively shorter and localized as was shown in Figure 3. Cell power savings

Table 4: Results with our net-based partitioning and metal layer saving in the bottom-tier. Top-tier uses six metal layers in all cases. We save 3 metal layers in all cases with minimal impact on full-chip M3D power and area balance across tiers. The runtime includes tier partitioning step only.

Method	# Metals (Bot-Tier)	Bot:Top Cell-Area	#MIVs	Wirelength (m)			Congestion		WireCap (pF)	Power (mW)			Runtime (sec)
				Top-Tier	Bot-Tier	Total	Hor (X)	Ver (Y)		Wire	Cell	Total	
LDPC (1.64 GHz)													
baseline	**6**	49:51	19,931	0.72	0.76	1.48	0.01	0.01	244.8	65.4	78.5	143.9	45
net-based	**5**	51:49	20,110	0.87	0.61	1.48	0.01	0.01	250.3	66.7	78.5	145.3	45
	4	52:48	21,871	0.97	0.52	1.49	0.01	0.06	265.3	71.2	78.6	149.8	51
	3	55:45	23,301	1.09	0.43	1.52	0.14	0.07	278.3	74.8	78.6	153.4	81
SIMD (2.36 GHz)													
baseline	**6**	50:50	40,558	0.93	0.89	1.82	0.01	0.04	318.6	42.4	80.7	123.1	173
net-based	**5**	50:50	45,386	0.99	0.78	1.78	0.06	0.07	320.8	42.8	80.7	123.5	210
	4	52:48	52,178	1.13	0.68	1.81	0.07	0.19	343.7	46.1	80.8	126.9	225
	3	54:46	50,274	1.28	0.66	1.94	0.29	0.26	385.5	52.2	81.1	133.3	275
AES (3.27 GHz)													
baseline	**6**	51:49	36,069	0.54	0.57	1.11	0	0.04	166.5	26.4	89.8	116.2	348
net-based	**5**	49:51	42,270	0.56	0.56	1.12	0.01	0.06	170.2	27.2	89.8	117.0	442
	4	55:45	38,993	0.69	0.53	1.22	0.17	0.18	184.2	32.6	90.3	122.9	482
	3	55:45	44,893	0.82	0.47	1.29	0.27	0.28	198.6	34.8	90.5	125.3	528

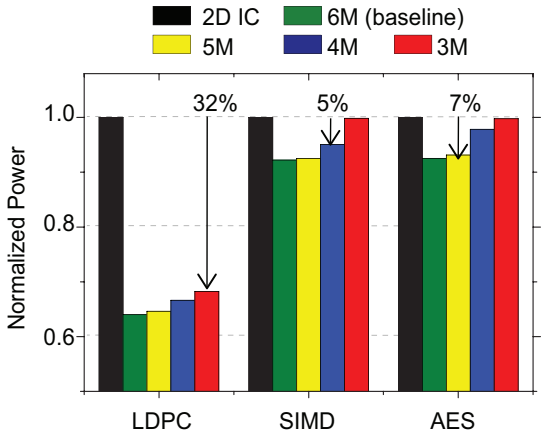

Figure 8: Normalized power comparison of 2D IC, baseline 3D IC and net-based partitioned 3D IC with reduced metal layers in the bottom-tier. Top-tier has six metal layers in all cases.

are comparatively much lower in these designs. Therefore, any impact on wire-power savings reflects heavily on total power savings. The localized congestion spots results in relatively higher power increase for the design implementation with three metal layers in the bottom-tier. The overall power savings are modest 5-9% for the different cases of bottom-tier metal layers reduction. The partitioning runtime becomes higher to obtain larger wirelength skew because the starting $hpwl\_max$ is lower and multiple recursions occur before obtaining an area-balanced partitioning result. However, as discussed earlier, this increase is negligible compared to the total design runtime of few hours which includes MIV planning, tier-by-tier routing and parasitic extraction.

In general, usage of more metal layers helps in increasing the power savings due to the relaxed routing conditions. Our net-based tier partitioning algorithm helps in keeping power in check, while reducing the number of metal layers in the bottom-tier. With our net-aware tier partitioning methodology, we successfully achieve wirelength skew with proper area balance. Therefore, the congestion in bottom-tier is reduced significantly and this helps in error-free routing with low wire capacitance. Though the power-only savings are higher while using more metal layers, the combined savings including cost of reduced masks is higher with reduced metal layers in bottom-tier.

6. CONCLUSION

In this work, we addressed the critical issues of BEOL impact on the performance of gate-level monolithic 3D ICs. We proposed a path-based tier partitioning algorithm to handle the impact of increased resistance of the bottom-tier interconnects with negligible design overhead. We demonstrate tolerance of up to 4X resistance increase in the bottom-tier interconnect, without any additional timing optimization. We then focused on reducing the cost of monolithic 3D ICs by reducing number of metal layers in the bottom-tier. Our net-based tier partitioning algorithm helps in creating wirelength skew without area skew. This helps in reducing the number of metal layers in bottom-tier without any routing congestion, therefore enabling cost reduction. Using our algorithm for two-tier monolithic 3D IC for an interconnect dominated benchmark, we reduced three metal layers in the bottom-tier while saving 32% power compared to 2D IC.

7. REFERENCES

[1] P. Batude, *et al*, "3-D Sequential Integration: A Key Enabling Technology for Heterogeneous Co-Integration of New Function With CMOS," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, no. 4, pp. 714–722, 2012.

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

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

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

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

[7] D. K. Nayak, *et al.*, "Power, performance, and cost comparisons of monolithic 3D ICs and TSV-based 3D ICs," in *SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), 2015 IEEE*, Oct 2015, pp. 1–2.

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